




M E A S U R I N G

ALCOHOL
OUTLET
DENSITY

A TOOLKIT FOR STATE
AND LOCAL SURVEILLANCE



U.S. Department of
Health and Human Services
Centers for Disease
Control and Prevention



Mike Dolan Fliss, PhD, MPS, MSW

University of North Carolina, Injury Prevention Research Center
North Carolina Division of Public Health, Injury Prevention Branch

Jessica B. Mesnick, MPH

Marissa B. Esser, PhD, MPH

National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention

Acknowledgements

The authors recognize the contributions of the following people to this toolkit: Valerie Goodson and Mia Israel of the Council of State and Territorial Epidemiologists assisted with project management for development of the toolkit. Elle Law reviewed early drafts and authored the first draft of Appendix 2. Peggy Dana led the layout and design of the toolkit. Alison Gatherum contributed to the scientific illustrations. Robert Brewer, Mary Beth Cox, Matthew Garnett, Kendall Knuth, Mandy Stahre, Pamela Trangenstein, and other members of the Council of State and Territorial Epidemiologists' Alcohol Epidemiology Subcommittee provided feedback on early drafts or provided suggestions during presentations on the contents of the toolkit. Teams from Colorado, Minnesota, Michigan, Maryland, New Mexico, and Utah provided alcohol licensure data for use in the toolkit or to guide its contents.

Development of this toolkit, analysis of indicators, and related technical assistance was supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award totaling \$150,000, with 100% funded by CDC/HHS to the Council of State and Territorial Epidemiologists (CSTE) (Cooperative Agreement No. NU38OT000297-03).

Suggested Citation

Fliss MD, Mesnick JB, Esser MB. *Measuring Alcohol Outlet Density: A Toolkit for State and Local Surveillance*. Centers for Disease Control and Prevention, US Dept of Health and Human Services; 2021.

This publication is available online at <https://stacks.cdc.gov/view/cdc/150909>. Two supporting appendices are available at www.cdc.gov/alcohol/php/alcohol-outlet-density-tools/index.html.

Website addresses of nonfederal organizations are provided solely as a service to our readers. Provision of an address does not constitute an endorsement by CDC or the federal government, and none should be inferred. CDC is not responsible for the content of other organizations' web pages.



CONTENTS

Executive Summary	3
Background	5
Toolkit Organization	9
Steps for Measuring Alcohol Outlet Density	10
Step 1. Build a measurement team.	10
Step 2. Define the purpose and indicators.	13
Step 3. Obtain and validate license data.	16
Step 4. Filter and classify outlets by license type.	22
Step 5. Spatially locate outlets.	26
Step 6. Calculate indicators.	32
Step 7. Visualize, report, and communicate.	38
Additional Concepts Not Covered in This Toolkit	45
Conclusions	45
References	46





Executive Summary

The [Community Preventive Services Task Force](#) recommends several strategies, including the regulation of alcohol outlet density, for reducing the availability and accessibility of alcohol to reduce excessive alcohol consumption and associated harms.^{1,2} However, standardized alcohol outlet density surveillance indicators had not previously been developed. To address this gap, the Centers for Disease Control and Prevention (CDC), in partnership with the Council of State and Territorial Epidemiologists, convened a group of experts on alcohol outlet density in September 2019 in Atlanta, Georgia. The group discussed how to support state and local health departments in the measurement and surveillance of alcohol outlet density and identified alcohol outlet density measurement indicators useful for public health practice.

The *Measuring Alcohol Outlet Density: A Toolkit for State and Local Surveillance* (hereafter called *Measuring Alcohol Outlet Density Toolkit*) provides steps for using these alcohol outlet density indicators for surveillance in states and local jurisdictions. It is a companion to CDC's *Guide for Measuring Alcohol Outlet Density*,³ published in 2017. This guide covers key concepts, high-level steps, and underlying measurement theory. The *Measuring Alcohol Outlet Density Toolkit* provides code, screenshots, and guiding questions to help you accomplish the six steps outlined in the *Guide for Measuring Alcohol Outlet Density*. It also adds a seventh step on visualization, reporting, and communication.

This toolkit is specifically designed for teams looking for practical instructions on how to measure alcohol outlet density for surveillance. Teams may include people with a range of expertise, including public health researchers, geographers, policy makers, or law enforcement personnel.

This toolkit also includes steps for preparing to conduct surveillance on alcohol outlet density and sample code and Geographic Information System (GIS) screenshots for calculating the following four indicators of alcohol outlet density:

A. Count-Based Indicators

- Count or rate of alcohol outlets per square land mile.
- Count or rate of alcohol outlets per 10,000 people.

B. Distance-Based Indicators

- Average distance from alcohol outlet to its nearest outlet (outlet to outlet).
- Average distance from a person to their nearest alcohol outlet (person to outlet).

It is supported by two appendices: Appendix 1. Calculating and Visualizing Four Indicators of Alcohol Outlet Density Using R and Appendix 2. Calculating and Visualizing Four Indicators of Alcohol Outlet Density Using QGIS. Both are available online at www.cdc.gov/alcohol/php/alcohol-outlet-density-tools/index.html.

LIQUOR

ROSE WINE

QUALITY FLOREST



Background

Who is this toolkit for?

This toolkit guides teams through the steps of measuring alcohol outlet density, which is the number and concentration of alcohol outlets in a region. It is designed to complement the *Guide for Measuring Alcohol Outlet Density*³, which describes the main steps at a higher level. This toolkit specifically focuses on helping team members who will be gathering data, calculating alcohol outlet density indicators, and communicating those results. Those team members may include a combination of state or local epidemiologists, academic researchers supporting a public health project, statisticians with spatial experience, GIS professionals or geographers from other departments (e.g., planning and zoning), and students.

What steps are covered in this toolkit?

This toolkit follows steps 1–6 from the *Guide for Measuring Alcohol Outlet Density* and adds a seventh step on visualization, reporting, and communication (Figure 1). Completing these steps in order will help ensure that your team establishes a common language and agrees on decisions for the alcohol outlet density analysis and plans for communicating the results. If you are more familiar with certain steps, you may choose to refer to individual sections only.

Figure 1. Seven steps in the *Measuring Alcohol Outlet Density Toolkit*



What indicators will be calculated?

This toolkit includes steps for preparing to conduct surveillance on alcohol outlet density and calculating two broad types of indicators (count-based and distance-based). Sample codes and GIS screenshots for calculating the following four indicators of alcohol outlet density are provided:

C. Count-Based Indicators

- Count or rate of alcohol outlets per square land mile.
- Count or rate of alcohol outlets per 10,000 people.

D. Distance-Based Indicators

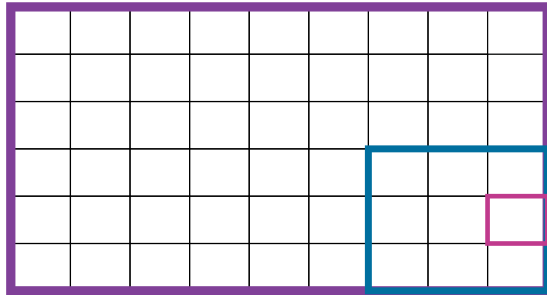
- Average distance from alcohol outlet to its nearest outlet (outlet to outlet).
- Average distance from a person to their nearest alcohol outlet (person to outlet).



What three area types are needed?

The three area types needed are the study zone, the study regions within study zones, and the small population units. As shown in Figure 2, the entire **study zone** includes all the smaller **study regions** for which alcohol outlet density will be calculated. **Smallest population units** are used as a proxy for a person's distance from their house to the nearest alcohol outlet when calculating a region's average distance to the nearest outlet.

Figure 2. Study zone, study region, and smallest population units



1) Entire study zone

The outer borders of the study. Valid outlets must be inside it.

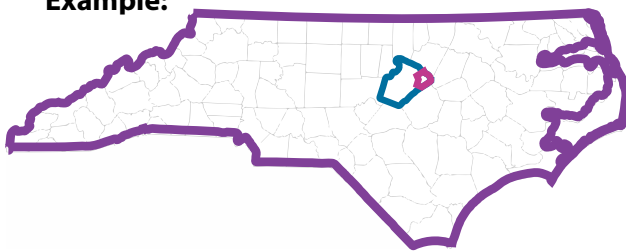
2) Study regions within study zone

1 or more regions, the size of the entire study zone or smaller, for which alcohol outlet density indicators will be calculated.

3) Smallest population units

The smallest areas for which population data are available. Used to represent local residents' distance to their nearest outlet.

Example:



1) Entire study zone

State of North Carolina

2) Study regions within study zone

Counties of North Carolina, like Wake County, NC.

State-wide or sub-county indicators could also be calculated.

3) Smallest population units

US Census block groups, like this one in Wake County, represent residential distance to nearest outlet.

What data will my team need?

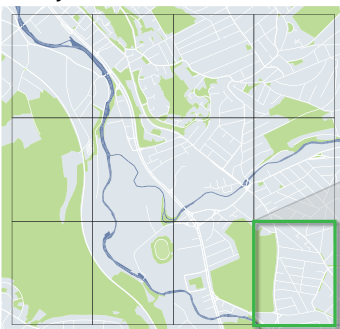
Your team will need the (A) shape of your study zone and your study regions, (B) population of those regions, (C) small population units used to represent people's distance to outlets, and (D) alcohol outlet locations. Figure 3 shows the data required for a study.

Details about these data appear after Figure 3.

Figure 3. Data needed to calculate alcohol outlet density indicators

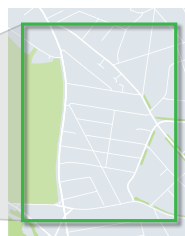
Gather Data for Each Study Region in Study Zone

Study Zone



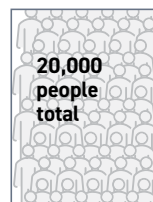
Study Regions & Region Borders

State, county, district, tract, etc.



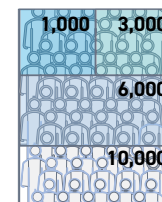
Population of Regions

Total number of people in each region



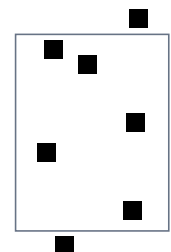
Small Population Units

Borders and counts of people in blocks, block groups, and households



Alcohol Outlet Locations

Location of outlets in and, when available, near the study zone



Core Alcohol Licensure Data: Your team will need the physical locations of the alcohol outlets in your jurisdiction(s) of interest (e.g., state or county), often stored with their alcohol license information. These data points may be in various formats, including the following:

1. A nonspatial, tabular form (e.g., an Excel spreadsheet).
2. A database with multiple tables related to each other.
3. A spatial data file with attribute and location information.

Your data may or may not require geocoding (e.g., physical addresses converted into point locations such as latitude or longitude points). It may, particularly if you plan to calculate the distance-based indicators. Your data may also require some transformation. For example, your data may include both active and inactive licenses (requiring filtering) or have multiple licenses for a single outlet (requiring combining).

Other Data: To calculate these indicators, your team will need some additional data. For rates based on land area, your team will need area information (e.g., size in square miles) or the geometric borders of jurisdictions of interest. Population data for your desired units are required to calculate the rate per population. You will need small population unit data, such as from census block groups or blocks, to estimate a person's distance to their nearest alcohol outlet.

Your team may want to use license information to group outlet licenses by their type (e.g., restaurant, bar, liquor store, gas station), but this is not required. Other useful data may include demographic characteristics of the jurisdiction, such as on race/ethnicity or socioeconomic position (for assessing disparities), or other contextual data (e.g., other health indicators, crime, or local points of interest) to help others understand and interpret the data presented in maps.

What pre-calculations or gathering of information is required?

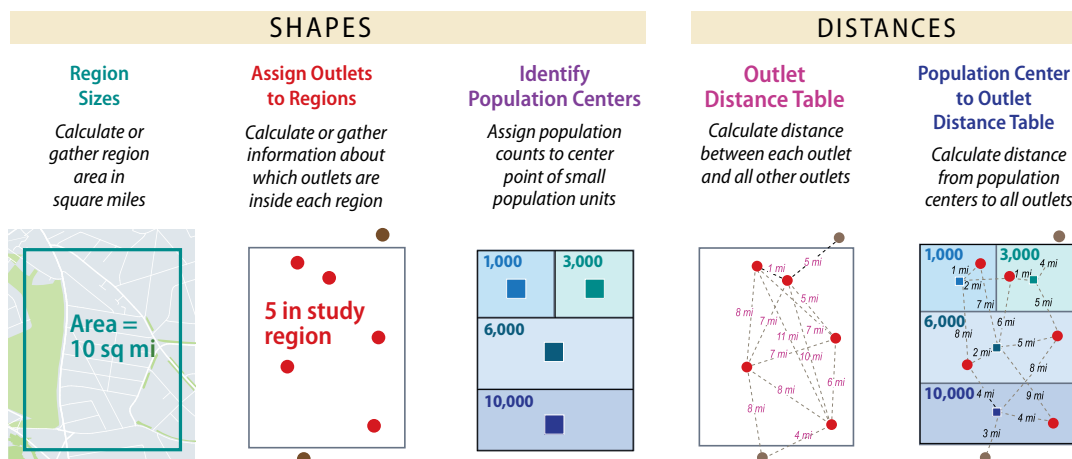
Calculating alcohol outlet density requires pre-calculating or gathering the following five shape and distance variables (Figure 4):

1. Determining region sizes
2. Assigning outlets to regions.
3. Identifying the center of small population unit shapes.
4. Calculating outlet to outlet distances.
5. Calculating small unit population center to outlet distances.

These pre-calculations are explained in Step 6.

Figure 4. Shape and distance calculations to complete before calculating alcohol outlet density indicators

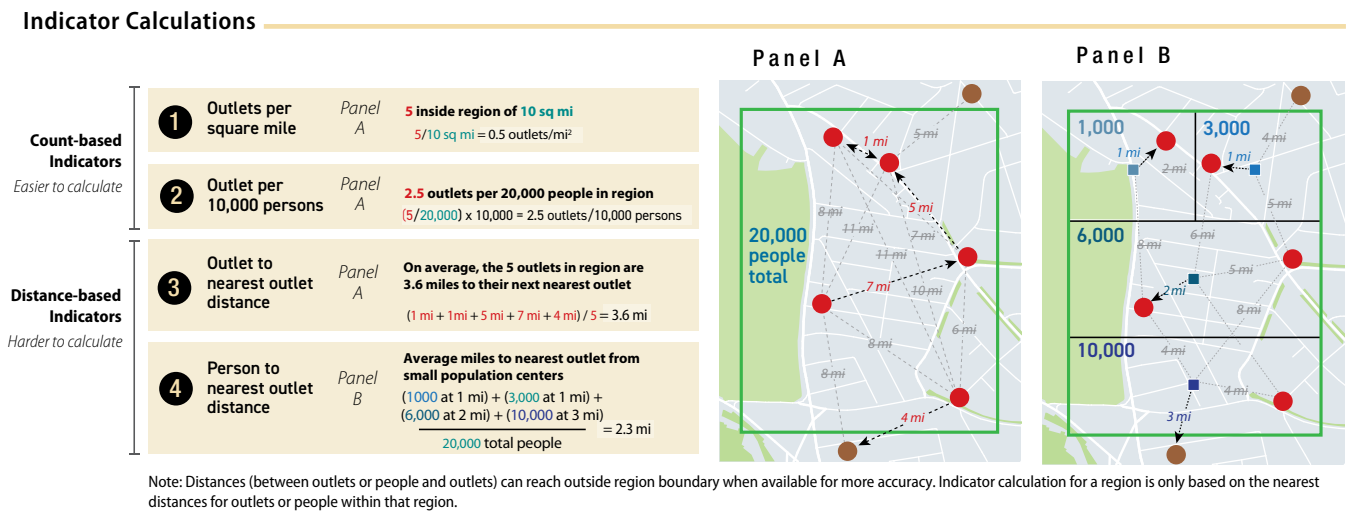
Shape and Distance Calculations



How are alcohol outlet density indicators calculated?

To illustrate the indicator math, Figure 5 is a simplified, rectangular spatial neighborhood of 10 square miles with 5 alcohol outlets inside it and 2 outside its boundary but nearby.

Figure 5. Calculations for four indicators of alcohol outlet density



There are 5 alcohol outlets in the 10 square mile region, or 0.5 alcohol outlets per square mile (Indicator 1). There are 20,000 total people living in this region with those 5 outlets, so there are 2.5 outlets per 10,000 people (Indicator 2). The 5 outlets in the region are the following distances from their nearest outlet: 1 mile, 1 mile, 5 miles, 7 miles, and 4 miles, respectively. The nearest neighbor to one of the outlets is just outside of the region boundary, and this is included in the distance, but only in the average of the nearest outlet distances for the 5 outlets within the region, yielding an average of 3.6 miles from an outlet to its nearest neighbor (Indicator 3). The region also has 4 population centers. The average of the population-weighted distances to the nearest outlet indicates that the average distance from a person to their nearest outlet is 2.3 miles (Indicator 4). Population centers are used as a best estimate or proxy for where people reside because it is rare to have access to household-level population data. Step 6 demonstrates in more detail how to calculate these indicators.

How were these indicators determined?

Many indicators are used by practitioners and researchers to measure alcohol outlet density.⁴ The *Guide for Measuring Alcohol Outlet Density* groups these into three major approaches.³ However, standardized alcohol outlet density surveillance indicators had not previously been developed. To address this gap, CDC partnered with the Council of State and Territorial Epidemiologists to convene a group of experts on alcohol outlet density in September 2019 in Atlanta, Georgia. The group discussed how to support state and local

health departments in the measurement and surveillance of alcohol outlet density and identified alcohol outlet density measurement indicators useful for public health practice.

Why were these indicators chosen, and is there a limit to what they can be used for?

The group chose these indicators because the results can be communicated to nonscientific audiences, such as community groups and policy makers, and can be useful in public health practice. The two count-based indicators do not require geocoding and require less expertise to calculate than the two distance-based indicators. However, all four indicators require less specialization than other more technical methods, such as cluster and spatial scan statistics. In public health practice, these indicators of alcohol outlet density can be used for conducting surveillance, assessing comparisons over time within a jurisdiction or comparisons between regions, and developing policies pertaining to alcohol outlet density (e.g., zoning or licensing regulations). However, the indicators may not be as useful for determining the association between alcohol outlet density and specific outcomes.

What software programs will my team need?

Software: You can often calculate the count-based density indicators (count or rate of outlets per square land mile and per 10,000 people) in a spreadsheet software tool such as Microsoft Excel or Google Sheets. However, to maximize your use of the methods in this toolkit, you may want to have a

team member who has access to and training in one or both of the following: (1) a statistical software with spatial capabilities or (2) a GIS. Examples in this toolkit use R and QGIS because both are powerful, open source, and free software programs for calculating spatial statistics and building maps. Both run on either PC or Mac operating systems. If your team has access to other statistical software like SAS or Stata, or GIS systems like ArcGIS, you can use the information in this toolkit and adapt it for use with your software.

Toolkit Organization

This toolkit introduces each of the seven measurement steps, followed by step-specific summary questions to serve as a quick reference for what is needed to complete the step. The details for executing the step are described as objectives, which are further divided into specific tasks.

There are four types of objectives:

1. Team conversations and decisions.
2. Data gathering.
3. Spatial analyses.
4. Visualization of results.

Each objective type is denoted by a corresponding icon (Table 1). The tasks for each objective provide step-by-step instructions for calculating the indicators. General tips and troubleshooting advice are also provided.

[Appendices 1 and 2](#) provide step-by-step instructions for calculating each alcohol outlet density indicator using both R and QGIS.

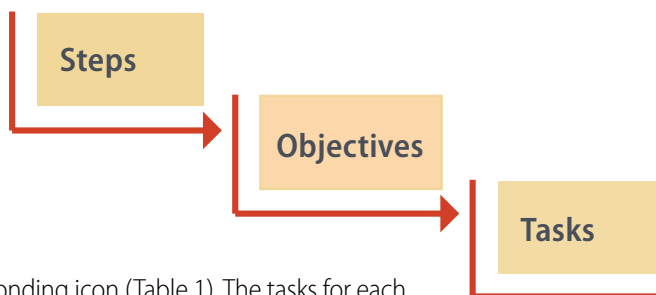
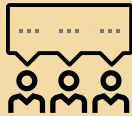





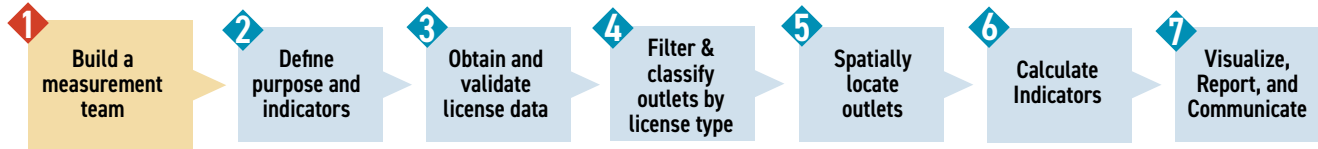


Table 1. Icons used to help the reader find components in the toolkit

Icon	Meaning
	Team Conversation or Decision Objective Hold one or more team meetings, discuss options, and document decisions.
	Data Gathering Objective Identify and collect data needed for analysis.
	Spatial and Data Analysis Objective Clean, manipulate, and perform tabular and spatial calculations with your data.
	Spatial and Data Visualization Objective Visualize results of analysis and calculations.
	Checklist See an overview of the objectives for each step.
	Summary Questions Questions for your teams to consider at each objective.

Steps for Measuring Alcohol Outlet Density

Step 1. Build a measurement team.



Objectives

- A. Gather the team and assign roles.
- B. Establish project timeline and goals.
- C. Choose and acquire software.
- D. Get training.

? Step 1 Summary Questions

- Who is on your team, and what are their roles? Is the team missing anyone? Is more training required?
- What is your project timeline? How often will your team meet?
- What are your project goals?
- What software and tools will you use?
- Does your project have or require funding for software, staff time, or contracts with other partners?

Step 1. Objective A: Gather the team and assign roles.

Tasks

- A1. Identify what skills you will need for the project.
- A2. Reach out to complete team.
- A3. Decide who is responsible for what role.



Determine the skills needed for team members and identify people with appropriate skills. Successful teams need a mix of subject matter expertise and spatial analysis skills. One person may perform many roles. A team might consider including people for the following roles:

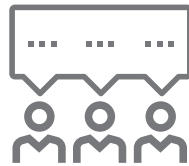
- a. **Data Provider:** Knows how to obtain the alcohol outlet density data.
- b. **Spatial Analyst:** Runs spatial analyses and understands underlying statistics and data flow.
- c. **Map Builder:** Takes output of spatial analysis and builds maps and reports.
- d. **Alcohol License Expert:** Knowledgeable about license structure and history.
- e. **Local Context Expert:** Knows the local environment and identifies characteristics that may influence alcohol outlet density, such as how population is distributed and neighborhood contextual factors.
- f. **Epidemiologist, Scientist, or Researcher:** Can conduct scientific or epidemiologic research and guide the team in applying relevant alcohol epidemiology principles and can consider how to interpret the data within the broader alcohol environment or public health context.
- g. **Administrative Coordinator:** Can coordinate meeting logistics and track progress towards goals.
- h. **Other:** Skilled in areas such as public speaking, graphic design, or policy research.

University researchers or student partners may help on teams. Public health, geography, data science, or public policy departments may have practicum requirements for students, which means these students could contribute to alcohol outlet density projects.

Step 1. Objective B: Establish project timeline and goals.

Tasks

- B1. Set initial timelines and goals.
- B2. Make backup plans.



B1. Set initial timelines and goals.

If you do not yet have a complete team, estimate how long it will take to gather the additional members needed. Once your team is formed, decide on meeting frequency, meeting structure (e.g., subcommittees), and how to address questions that arise. Establish a timeline for gathering data, conducting analyses, and making maps, and identify milestones that team members will be expected to accomplish and update the group on. While establishing a timeline, consider that the coding and categorization of license types can take time, and decide on the level of precision of your analyses according to your project goals. Consider the time needed to clean data, perform analyses, make maps, build graphs, write reports, interpret the results, ask additional analytic questions, and review and revise at various steps. The seven steps provided in this toolkit may serve as a useful foundation for setting a timeline.

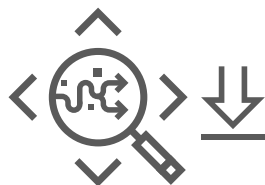
B2. Make backup plans.

Establish clear steps for handling unexpected events, such as missed deadlines or team staffing changes.

Step 1. Objective C: Choose and acquire software.

Tasks

- C1. Choose your software.
- C2. Acquire your software.



C1. Choose your software.

Choose the software your team will use according to your existing or available resources. In this toolkit, examples and sample code are based on R and QGIS, which are free and open source. The techniques can be applied to other software programs (e.g., SAS, Stata) and GIS systems (ArcGIS) that team members might already have access to.

The analysis can be performed entirely in either R or QGIS. It is not necessary to use both, but they can be used together. For example, data cleaning, combining, spatial and statistical analysis, and drafting maps might be easier for some people to do using R, and final shapefiles can then be exported and used in QGIS (or another GIS) for more attractive maps and reporting. The [Appendices](#) show examples using both software packages.

C2. Acquire your software.

Acquiring your software may involve downloading free and open-source options or purchasing licenses before installations.

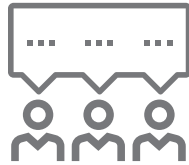
R is available for free [online](#). Teams interested in using R would install both R (the language) and RStudio (the graphical user interface). You will need additional packages, including [tidyverse](#),⁵ [sf](#),⁶ and [tidycensus](#),⁷ which are high quality, commonly used, peer-reviewed, and well-maintained. To install these packages, open a new R script and run the `install.packages()` command one time for each package, calling the `library()` subsequently to include those package's functions in your workspace.

Teams interested in using QGIS may download it for free [online](#). Install the **Group Stats** plugin, which will be useful to summarize measurements in one column according to the category of another column. To do this, select Plugins, then select Manage and Install Plugins from the program menu. Then, in the search bar, type in “Group Stats,” then “Install plugin” for the result, and click Close. Now when you select Vector on the QGIS menu bar, Group Stats should appear as an option. Plugins, often developed by third parties, are a valuable resource for making QGIS meet your needs. The **Quick Map Services** plugin might also be helpful, which you can install the same way. It allows you to search for and add base maps within QGIS.

Step 1. Objective D: Get Training

Tasks

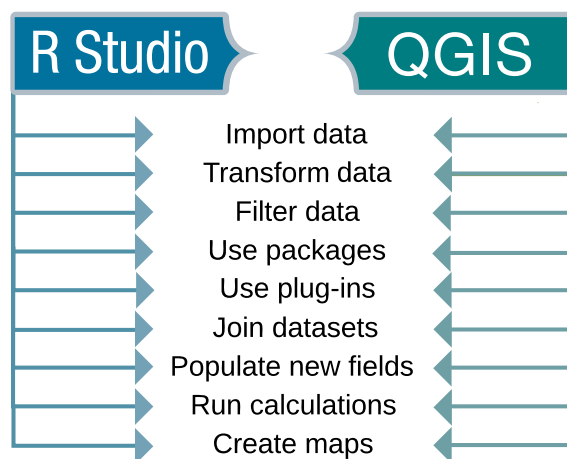
- D1. Identify specific skills for training.
- D2. Plan your training.



D1. Identify specific skills for training.

Figure 6 shows a description of the available software and skillsets you will need to follow the analyses described in this toolkit. R and QGIS both have strong online communities if your team is interested in free training on the skills shown in Figure 6.

Figure 6. Skills for conducting alcohol outlet density spatial analyses, by software program



D2. Plan your training.

Assess whether your team will have the skillsets for conducting these analyses or whether trainings are needed. If you or a member of your team is affiliated with an educational institution, they may already have software or spatial analysis experts who can consult with your team. It might be possible to find assistance online, in software documentation, or from an expert who is able to provide technical assistance. If you are using R or QGIS, you can find several excellent training resources available free online, such as the following:

R Resources

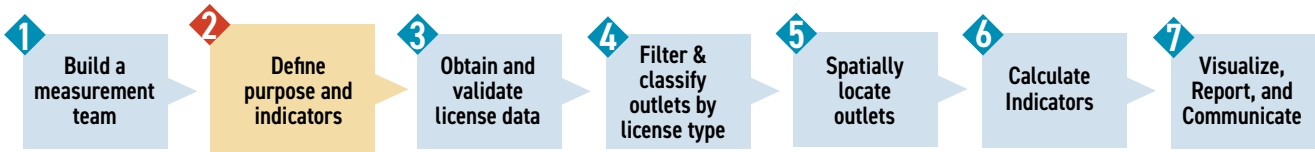
- [R for Data Science](#)⁸ Hadley Wickham and Garrett Grolemund.
- [Simple Features for R](#),⁶ Edzer Pebesma.
- [Geocomputation with R](#),⁹ Robin Lovelace, Jakub Nowosad, and Jannes Muenchow.

QGIS Resources

- [QGIS Official Training Manual](#), QGIS team.
- [QGIS Tutorials and Tips](#), Ujaval Gandhi.

Both R and QGIS are well covered in the [Stack Exchange](#), a universe of websites dedicated to community peer-to-peer assistance. The [GIS Stack Exchange](#) contains answers and troubleshooting for many of the steps in this toolkit. Course materials for a class on R for public health, including a unit on spatial analysis, are freely available [online](#) through the University of North Carolina – Chapel Hill.

Step 2. Define the purpose and indicators.



Objectives

- A. Prioritize alcohol outlet density measurement surveillance questions.
- B. Choose indicators of alcohol outlet density to calculate.
- C. Choose study zone and regions.
- D. Gather region shape and population data.

? Step 2 Summary Questions

- What is the main goal of your alcohol outlet density measuring project?
- What questions does your team plan to prioritize?
- What questions are of interest but may take longer to complete?
- How does data availability and team capacity affect your questions and tasks?
- Who will you share your results with and why?

Step 2. Objective A: Prioritize alcohol outlet density surveillance questions.

Task

A1. Determine alcohol outlet density surveillance questions.



A1. Determine alcohol outlet density surveillance questions.

The *Guide for Measuring Alcohol Outlet Density* lists many example purposes for a project that assesses alcohol outlet density. To perform an analysis, teams must further hone ideas into specific questions that can be analyzed, visualized, and communicated. Developing well-defined alcohol outlet density surveillance questions can be challenging. Two tiers of sample questions are provided below. The Tier 1 questions can be answered using the indicators described in this toolkit. Tier 1 questions may lead teams to develop an interest in other questions (Tier 2), which might require other datasets or analytic techniques that are not included in this toolkit.

The following Tier 1 questions are examples of surveillance questions that can be answered using measurement techniques described in this toolkit. Teams may have other surveillance questions as well.

Tier 1 Questions

- What are the spatial locations (mapped) and count of alcohol outlets within your regions of interest?
- What is the overall alcohol outlet density in your study zone, expressed as a count per square mile, count per population, outlet to nearest outlet distance, or person to nearest outlet distance?
- Does alcohol outlet density differ by county, census tracts, or neighborhoods? Where is it higher and where is it lower?
- How many alcohol outlets operate by license type? How many operate by license type groups (e.g., on- or off-premises outlets, breweries)?

- If you can obtain historical data, how has alcohol outlet density changed over time in your study zone and regions? Where has it increased and where has it decreased? If this is the first alcohol outlet density analysis, data obtained for this analysis can serve as a baseline for future analyses.

The following Tier 2 questions are examples of potentially meaningful questions related to alcohol outlet density. Some of the Tier 2 questions would require use of other datasets (e.g., health, crime) to answer, some questions may extend beyond the scope of public health surveillance into research, while others may benefit from initial research before public health surveillance is possible.



Tier 2 Questions

- How are alcohol-related health data (e.g., injury, chronic disease outcomes, or years of life lost due to alcohol^{10,11}) spatially associated with alcohol outlet density?
- How are overall health and disease indicators (e.g., all-cause mortality, life expectancy) related to alcohol outlet density?
- Are residents of one racial or ethnic group exposed to higher alcohol outlet density than a group of residents of another racial or ethnic group in a region?
- How are tobacco outlets, marijuana outlets (if applicable), and alcohol outlet densities related?
- What regions have alcohol outlet density above a certain threshold? What thresholds, if any, are relevant to your community?
- Where are schools, parks, and healthy food options located? How does that compare to regions with high alcohol outlet density?
- How are historical zoning decisions associated with alcohol outlet density? For example, if your study zone has red-lining maps (communities systematically denied housing loans and infrastructure investment along racial lines), how are those related?
- How are socioeconomic position indicators (e.g., median household income, percentage under federal poverty level) associated with alcohol outlet density? How are specific license groups associated?
- Do regions with higher alcohol outlet density have higher crime or police-reported incidents?
- How is alcohol outlet density associated with any-cause or alcohol-related motor vehicle crashes?
- How does the alcohol outlet density in your study zone compare to other jurisdictions or states? How are issues of rurality and urbanicity accounted for?
- If alcohol outlet density regulations changed, how would health, crime, or motor vehicle crash outcomes of interest be affected?
- How are local alcohol-related ordinances and practices like open container laws, buying or consumption limits, and sale hours associated with alcohol outlet density?
- What strategies exist to address alcohol outlet density in your study zone, and how would the use of a particular strategy affect different indicators of alcohol outlet density?

Step 2. Objective B: Choose indicators of alcohol outlet density to calculate.

Task

B1. B1. Determine indicators to calculate.



B1. Determine indicators to calculate.

Teams do not need to calculate all four indicators of alcohol outlet density described in this toolkit. However, you should calculate more than one, if possible. The various indicators capture different aspects of alcohol outlet density and may each provide different information to communicate with various audiences. The first two **count-based** indicators are simpler to calculate. The second two **distance-based** indicators may be more challenging to calculate (Table 2). Consider your anticipated audience and your available resources (e.g., time, team members, skillsets, data availability), then determine which indicators your team plans to calculate.

Table 2 provides examples for interpreting results for each of these indicators. Note that these examples could also be applied to types of outlets or groups of regions by comparing within those groups, such as for off-premises outlets only or for rural regions only. See Step 6: Calculate Indicators for diagrams visualizing their calculation.

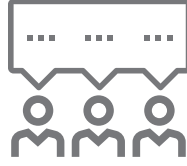
Table 2. Advantages and disadvantages of the four indicators

Type of indicator	Indicator	Advantages	Disadvantages	Examples of indicator use in sentence format
Count-based	Number of alcohol outlets per area (in square land miles)	Simple to calculate. Simple to describe in policy.	Can create fractional rates. Does not capture clustering.	The outlet density rate is 3 outlets per mi ² in Region A. Region A has 2 more outlets per mi ² than Region B (at 1/mi ²), and twice as many outlets as Region C (1.5/mi ²). Within sub-regions of Region A (e.g., tracts within counties), alcohol outlet density varies between 0.5 outlets/mi ² and 10/mi ² .
	Number of alcohol outlets per population (10,000 people)	Simple to calculate. Adjusts for population density.	Can create fractional rates. Does not capture clustering.	The number of outlets per 10,000 people in region D is 2.5. There are 2 more outlets per 10,000 people in Region D than in region E (0.5 outlets/10,000). This represents 5 times as many outlets per 10,000 people in region D than in Region E (0.5 x 5 = 2.5).
Distance-based	Average distance from outlet to its next nearest outlet	Captures some cluster dynamics.	Harder to calculate, requires geocoding. Harder to describe in policies.	The average distance between an alcohol outlet and its next nearest outlet is 1 mile in Region F. Even though the total number of outlets in Region G is the same, they are closer to each other (0.5 miles on average) than those in Region F are.
	Average distance from a person to their nearest outlet	Person-centered indicator.	Harder to calculate, requires geocoding. Harder to describe in policies.	The total average distance between a person and their nearest outlet is 1 mile in Region H. However, there are differences by demographics in that region. White non-Hispanic people live on average 1.5 miles to their nearest outlet, while Black people live on average only 0.5 miles to their nearest outlet.

Step 2. Objective C: Choose study zone and regions.

Task

C1. Choose your study zone and regions.

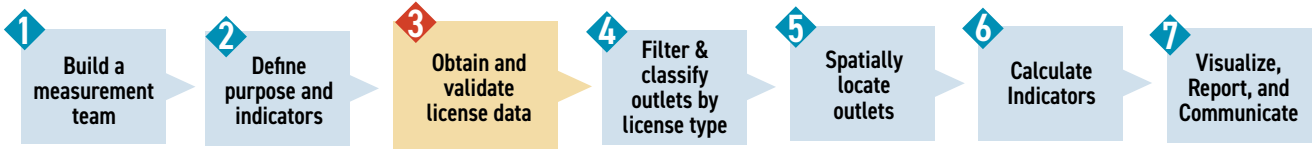


C1. Choose your study zone and regions.

Alcohol outlet density analyses require both a study zone and one or more regions within it (Figures 2 and 3). The study zone is the outer geographical bounds of the analysis (like a state), while the study regions (e.g., counties in a state or tracts in a county) are the smaller shapes within the study zone for which you will calculate the alcohol outlet density indicators. Your team's goals and the availability of contextual data at that level of analysis will determine what study zone and regions you choose. Consider the level of jurisdiction that your team is interested in, such as states, counties, zip codes, census tracts, or census blocks. For distance-based calculations, smaller population units (e.g., census blocks or block groups) will represent people's homes, and those smaller population unit calculations are averaged within the regions of interest.

When choosing a study zone and regions, your team might be making comparisons between them or assessing changes over time. Census regions, like counties and census tracts, tend to have accessible and stable boundaries and population data, making them a good unit of analysis for initial calculations. Custom regions such as legislative districts, local neighborhoods, historically meaningful boundaries, and zoning districts might be harder to calculate as population data can be challenging to gather or estimate for non-standard, non-census geographies. However, the results might be consequential for some audiences.

Step 3. Obtain and validate license data.



Objectives

- A. Collect outlet and license data.
- B. Understand license data structure.
- C. Perform joining, cleaning, recoding, and filtering.

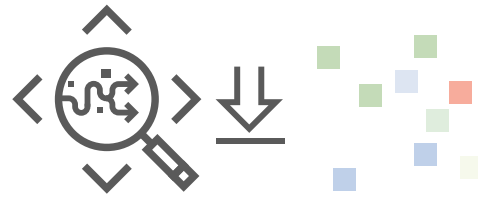
Step 3 Summary Questions

- What is the relationship of outlets and licenses in your jurisdiction?
- What outlet or license tables exist? If more than one table exists, how are the tables related?
- What fields are available in these tables? Which are useful? Which must be cleaned?
- What recoding, filtering, formatting, and structural changes need to be made to the original data?

Step 3. Objective A: Collect outlet and license data.

Tasks

- A1. Identify and request data from regulatory outlet and license data sources.
- A2. Explore data collection alternatives if no regulatory sources exist.



A1. Identify and request data from regulatory outlet and license data sources.

Jurisdictions with a centralized licensing system might have databases containing alcohol license information and outlet locations. It might be possible to download these data or to request them.

Other administrative datasets may be useful if centralized licensing and outlet data are not available. Sales and tax data may be attached to address information that can be used to identify outlets. You may be able to use proportional tax types (e.g., food versus liquor) to classify outlets as wholesale or transportation businesses (to exclude from density calculations), on-premises restaurants, or off-premises stores.

A2. Explore data collection alternatives if no regulatory sources exist.

If alcohol licensing and outlet data are not available from centralized systems or entities that collect alcohol sales and tax data, your team might be able to obtain data from community-based or coalition data collection efforts, for-profit organizations or businesses, novel online survey methods, commercial business lists, or geocoding Application Programming Interface (API) services.

Step 3. Objective B: Learn license data structure.

Tasks

- B1. Identify rows, columns, and table sizes.
- B2. Understand table and field concepts.
- B3. Document concept-level relationships, diagramming if necessary.
- B4. Document field-level table relationships, diagramming if necessary.
- B5. Review and obtain missing metadata.
- B6. Validate data and record data quality.

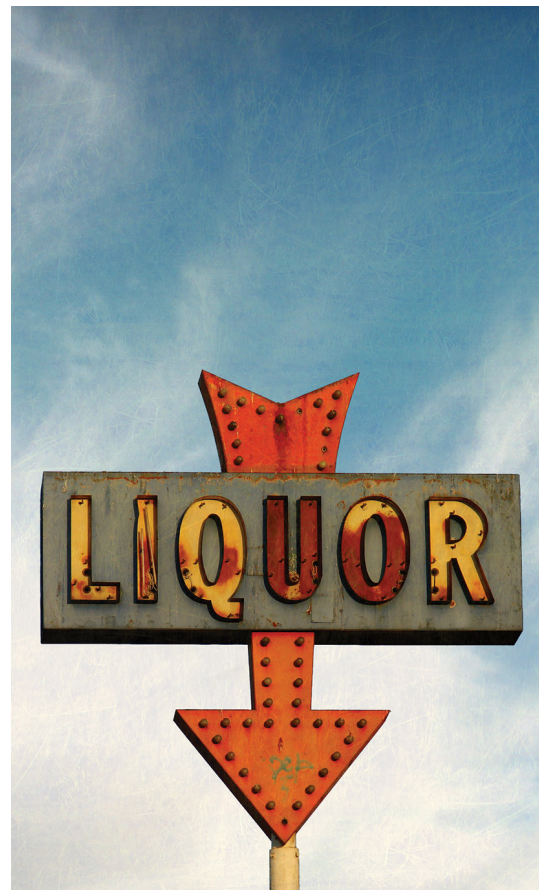


B1. Identify rows, columns, and table sizes.

First, determine what each row represents in the original dataset table or tables. For example, if outlets have multiple licenses and you receive a single “flat” table, that table may have multiple rows per outlet, with each row representing a license. In other cases, the dataset may only have one row per license or one row per outlet. In the case of multiple tables, determine what one row (record) of each table is meant to uniquely identify.

Second, determine what variable each column represents. Identify whether the variable is expected to be key for analysis tasks like filtering, grouping, geolocating, or linking tables together. Identify variables not expected to be used in the analysis. Variables that are not needed for analyses can be removed in subsequent steps.

Lastly, describe your data’s dimensions by counting the rows and columns in each table. This will help for understanding the table and field relationships in the next tasks.



B2. Understand table and field concepts.

Outlet and license data from databases might be stored in separate tables and connected through explicit or implicit linked relationships. Your team may either receive the data in these separate table formats (e.g., as separate spreadsheet worksheets in workbooks or separate flat tabular files) or in a pre-combined and “flattened” single table format. If separate, carefully combine the tables to produce a flat analysis table of open businesses with active licenses and then filter into active alcohol outlets. If pre-combined, it will still be useful for your team to understand the implied relationships between concepts, such as businesses (of which physical outlets are a subset) and licenses in the flat table.

There are many ways to specify the relationship between tables in a relational database.¹²

Tables contain rows of data, called **records**, with one or more **fields** that make up the columns. These fields each have a **data type**, which may be explicit or implied, such as a string of text characters, a number, or a date. One or more fields in a table may serve as that table’s **primary key**, a unique way to identify a single record in that table. The table may also have **foreign keys** that identify a single unique row in another table, used to link those two tables together. The relationships are

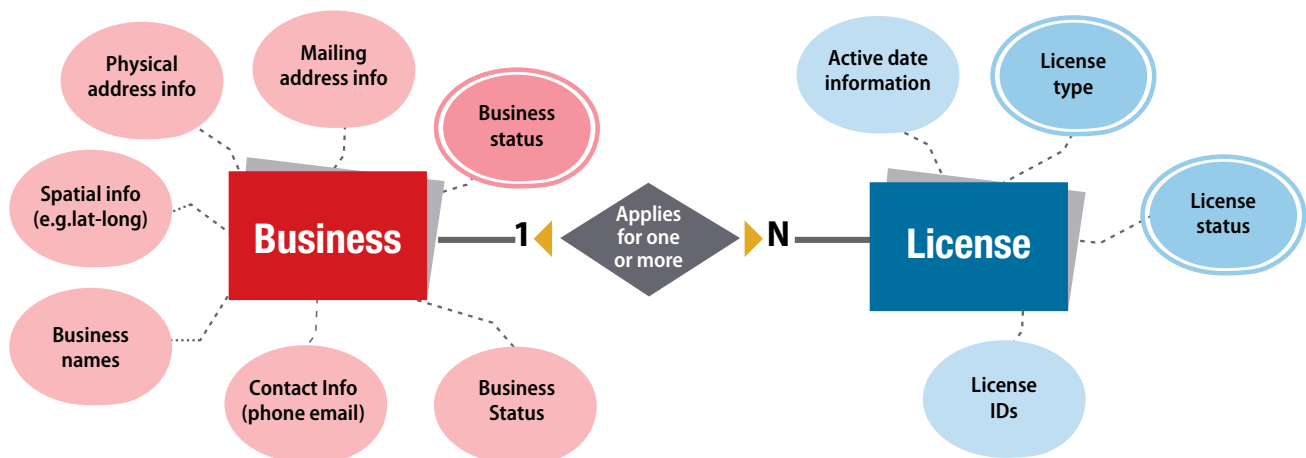
defined by **cardinality** information, sometimes also called **multiplicity** (e.g., one-to-many, one-to-one, many-to-many) and whether the relationship must have at least one link for every record in each table.

B3. Document concept-level relationships, diagramming if necessary.

You might find it helpful to diagram the table and database concepts to understand the tables, variable fields, and table-field relationships, which is important for subsequent analytic steps.

Figure 7 shows an example of a given business that is issued one or more licenses. Business records and licenses are stored in separate tables. This business can have more than one license, but a unique license cannot be granted to more than one business, so the license table has more data rows than the business table. Each table has data fields, represented by ovals in the diagram. Both business and license records include a unique identifier (underlined variable name) and other types of data shown in the diagram (e.g., status of active or inactive). Unlike some other text fields, business status must be chosen from a selection of valid choices. It can be represented as its own simple table of choices or as a categorical variable (with double borders) as in Figure 7.

Figure 7. Example of an entity relationship schema diagram for alcohol license and outlet database concepts

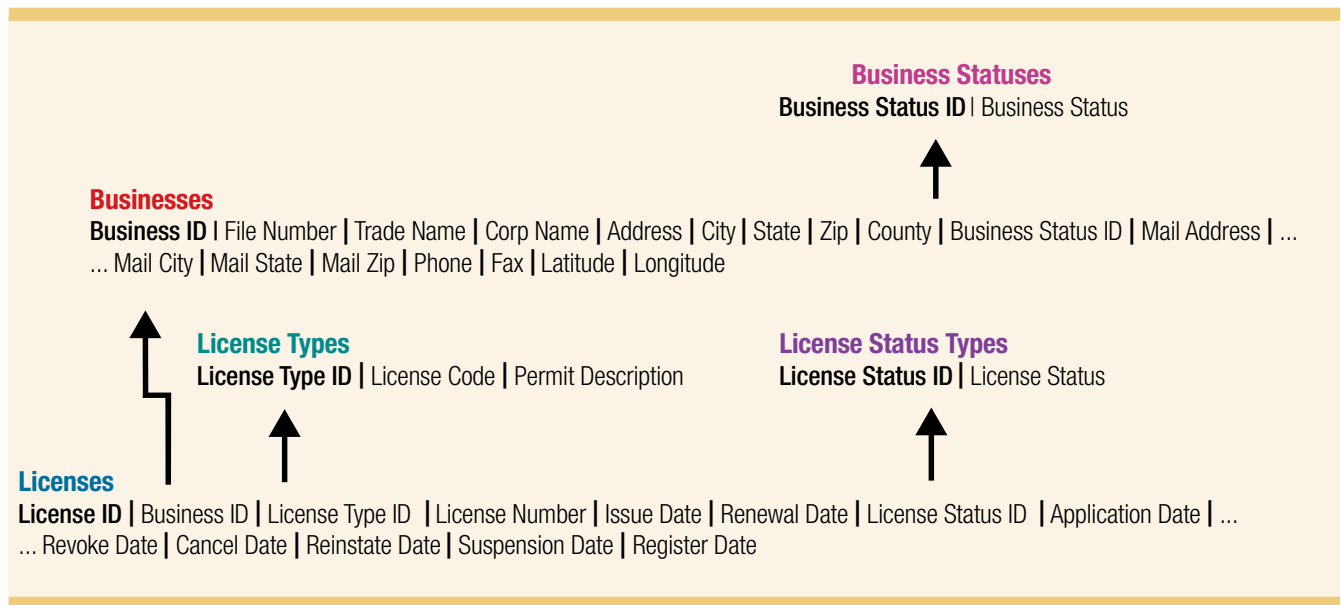


B4. Document field-level table relationships, diagramming if necessary.

To join tables and conduct statistical analyses, your team will need to understand table details, including each table’s field names, how primary and foreign keys fields are used to join tables, and information about field types (optional). Below are two ways to document these field-level relationships.

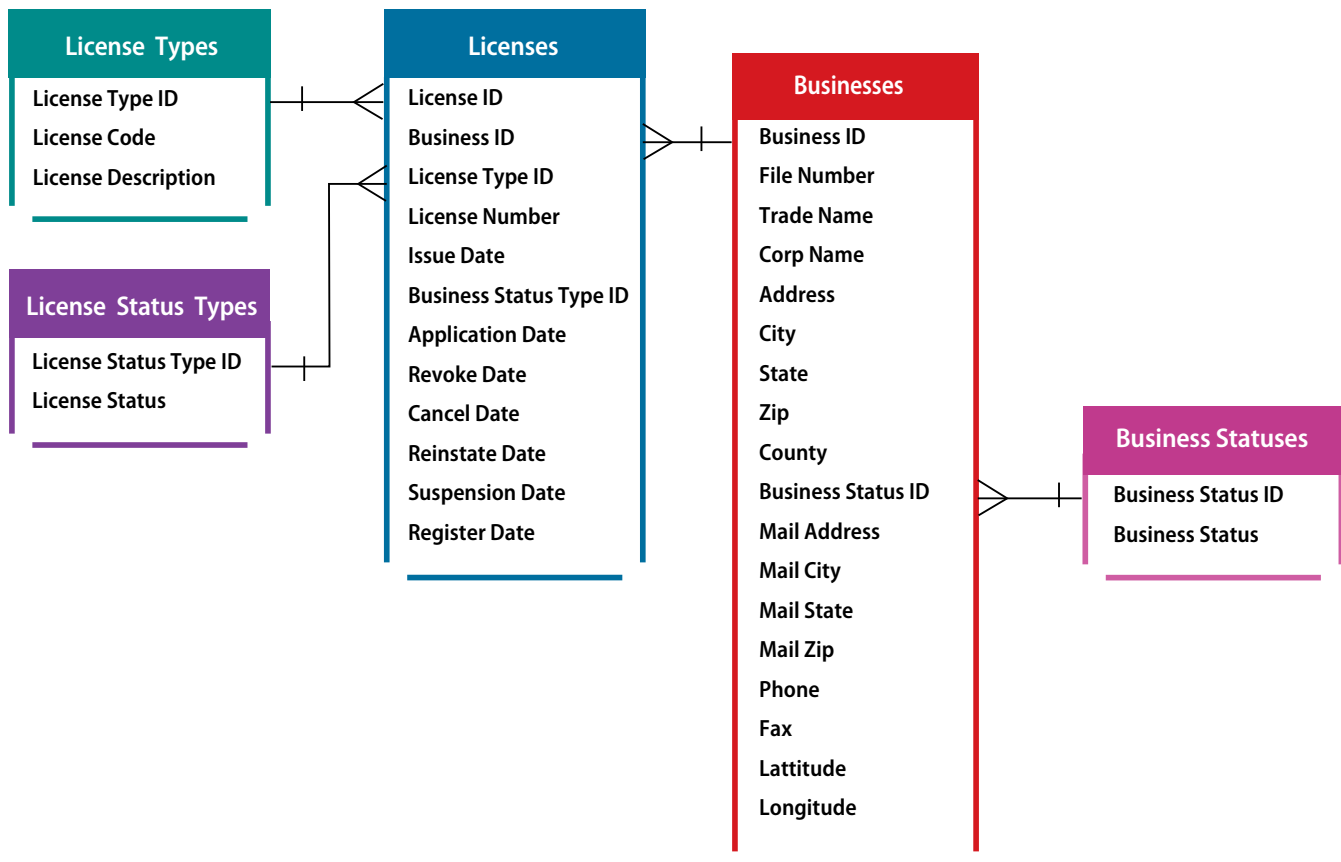
A text-based relational schema is a space-efficient way to document specific variable names in each table. As shown in Figure 8, underlined variable names are that table’s primary key, uniquely identifying each row of data. Foreign keys in other tables point to these primary keys, visually indicating the relationship of these key fields. In this example detailed view, business status, license status, and license types are documented as related tables of few variables, and each record is an allowable value of that categorical variable.

Figure 8. Example of a text-based relational schema for an alcohol license and outlet database



A table-based relationship schema, as shown in Figure 9, might be easier to understand, and it provides more data than just a list of field names. This type of diagram can be generated in many database programs. Symbols or text can be included to clarify data relationships (e.g., one-to-many) and whether linked data are required. These diagrams can include additional table columns to document the field variable type (e.g., string, integer, date) and a short description of the variable.

Figure 9. Example of a table-relationship field-detail schema for an alcohol license and outlet database



B5. Review and obtain missing metadata.

Review the detailed data you have on the definitions and relationships for each field in your dataset. To review and determine if any data are missing, consider the following questions:

- Are there underlying paper or online forms that produce the data?
- Consider the data origin for each field; who determines each field's value?
- What data cleaning or validation, if any, is done on key fields (e.g., are errors or typos possible in business name or address fields)?
- Can the physical and mailing addresses be reliably distinguished?
- Is your dataset complete, or are certain alcohol outlets not included (e.g., are state-controlled ABC stores retained in a different database than licensed private outlets)?
- Are all outlets in operation, or are some outlets closed? Are all license types in active use, or are some no longer offered to businesses?

Your team may need to contact the source of your license data to answer these questions or identify someone to ask for clarifications.

B6. Validate data and record data quality.

You will need to record whether the data were of good quality. You will want to consider, for example, validating a data sample by comparing to actual alcohol outlets.

After determining key fields of interest, report the number and percentage of records missing data in those fields and make note of how missing data may bias the results. This will be the record of how complete the data are. Step 5 covers issues of missing data in the address and geocoded outlet location data.

To validate complete data, your team may want to go to a small sample of outlets, by address or geocoded point location, to confirm they are present, open for business, and generally match the license and business types implied by the database. You can also use online maps that provide street-level views to help with this validation.

Data from other sources can be used to validate the use of license types in practice, including sales data or in-store surveys, to validate the proportion of sales of alcohol by type. Enforcement personnel may know information about certain outlets and their licenses to add real-world context that is essential for interpreting the database.



Step 3. Objective C: Perform joining, cleaning, re-coding, and filtering.

Tasks

- C1. Join tables to flatten and combine data.
- C2. Clean and recode fields used in data preparation and analysis.
- C3. Filter to active, unduplicated outlets and licenses.



C1. Join tables to flatten and combine data.

Original outlet and license data structure varies widely between jurisdictions, so the transformation process from original data to analysis-ready data will be different for all teams. If your team's data arrived in a single table of active outlet locations, you may not need to change its underlying data structure any further. If your data has multiple tables, these need to be rejoined into a single flat table of deduplicated outlets for easier analysis.

In the example outlet and license database discussed in objective B, outlets (businesses) and licenses are separate tables. Small helper tables document the categorical options for certain key variables. To combine and flatten this data into a single table, you will need to join tables together using their primary and foreign keys. When binding data, each dataset must have one or more matching columns (often identically named). For columns not in both datasets, the values of that column would be missing in the new table when not available in the original table but present when the column was available.

If similar data come in multiple tables (as with separate tables from each county in a state, or state liquor store data separated from licensed private outlets), the data will need to be row bound (also referred to as appended) together. When binding data, columns might need to be renamed to be identical in one or both tables. When binding data, consider whether two columns of similar concepts in different tables truly have the same meaning. If they do not, one or both may need to be harmonized to form identical concepts, such as address data divided into parts in one dataset and another containing all address data in a single field, or if multiple years of license data are combined but the meaning has changed for licenses with identical names.

After completing these steps, you will have a single, combined table where each row is an outlet-license combination. For outlets with multiple licenses, a single outlet may have multiple rows in this dataset, which will be covered in later steps of this toolkit.

Remember that analysis projects are often iterative. After completing all steps in an analysis, a team member or other partner may identify small errors that require rerunning the analysis steps. Scripting these steps allows them to be easily repeated if necessary. Avoid direct editing of original data whenever possible. Direct editing doesn't document the process to transform your original data to analysis-ready data, and it is hard to repeat. If scripting languages are not available, carefully document point-and-click steps.

C2. Clean and recode fields used in data preparation and analysis.

Prepare the variables of interest by considering some issues that are common to many alcohol outlet density measurement analyses.

To prepare for filtering data that includes closed outlets or expired licenses, consider the relevant date or available active or closed status fields and an appropriate way to determine whether an alcohol outlet is active. Prepare these date and categorical fields for filtering in the next step. For example, if date fields were stored as text strings, convert them to date variable types so logical filtering questions (e.g., drop all records with expired dates before January 1 of this year) can be implemented. It is also important to check that dates are formatted in a readable way and reflect actual possible dates (e.g., not indicating a 13th month or start date years of 1900).

If you want, you can reformat or modify field names (for example, you might standardize capitalization or punctuation), but you should avoid special characters or spaces. Some programming languages and software have rules requiring exact variable matching and capitalization. You should check whether the programming language or software requires exact variable matching and capitalization handled a specific way.

Review the dataset to verify each outlet has a unique ID, or create one if each outlet does not already have one.

C3. Filter to active, unduplicated outlets and licenses.

If there are many ways to identify an open outlet with an active license (e.g., a license status field, a business status field, and multiple license event date fields such as issue date, application date, revocation date, cancellation date, reinstatement date, suspension date, and registration date), teams can examine and discuss the relationships among these fields with database experts.

After examining the data and discussing with experts, filter the single business-license table to only those open businesses with active licenses.

Step 4. Filter and classify outlets by license type.



Objectives

- A. Prepare license-type classification table.
- B. Join classification table to outlet data.
- C. Perform filtering and grouping.



Step 4 Summary Questions

- What outlet or license types exist? Which will you include in your surveillance efforts? Which will you drop?
- How many outlets are there of each type?
- Are there sub-groups of outlet types of specific interest? How many of each group are there?

Step 4. Objective A: Prepare license-type classification table.

Tasks

- A1. Understand license type options in your jurisdiction.
- A2. Create frequency table of license types.
- A3. Make and record team decisions on license type filtering and grouping.



A1. Understand license type options in your jurisdiction.

License rules vary greatly among jurisdictions but typically include the following: license types, recent type additions, and groups; whether licenses are centrally or locally controlled; whether a single outlet can have one or many licenses; and the process for renewal, oversight, denial, and complaint for these licenses. You should know these license rules because they are key to both obtaining and correctly analyzing data.

Although most alcohol outlets (e.g., bars, restaurants, liquor stores) may have simple license types, some may have more unexpected or complex combinations of license types. Outlets without storefronts (e.g., wholesalers, wineries, caterers, airports) are typically identified and filtered out of alcohol outlet density analyses because they operate in circumstances that differ from those of other alcohol retailers.

While the total density of all physical alcohol outlets is meaningful, teams may also want to group alcohol outlets by their type and calculate group-specific densities. One common grouping is into on-premises retailers, where patrons drink alcohol at the site (e.g., restaurants, bars, and nightclubs), and off-premises retailers, where alcohol is purchased for consumption elsewhere (e.g., liquor stores, gas stations, and some grocery stores). Teams may also have other outlet groups of interest specific to their localities, such as beer gardens or breweries that allow on-premises consumption. Some license types may no longer be in use. Orient yourself with the licensure structure and procedures in your jurisdiction in advance of analyzing your data by speaking with content experts and conducting online research.

A2. Create frequency table of license types.

Classifying all license types at once using a look-up table of groups and indicating whether each license type will be included in your analysis can help you create a frequency table of license types. Some license types (or license combinations, when looking at a multi-license system) can be flagged to be filtered out, and the others can be categorized to useful sub-groups. That license type table, sorted by frequency, is a useful tool for recording your team’s decisions about how to group license inclusion. To build that table, identify the license type field or fields in your data. License types may be represented as a number, an alphanumeric license code of some kind, a human-readable license description, or some combination. The example frequency table (Table 3) shows 10 rows of license types. You can build and export frequency tables using statistical languages in a spreadsheet program or a dedicated GIS tool.

If your jurisdiction allows multiple licenses at a single outlet, the frequency table will be based on all actual combinations of license types. This is not covered in our code in the Appendices. You can accomplish this by combining multiple “rows” of license data for an outlet into a single row, with multiple licenses combined into a single text field, then considered as a group.

Table 3. Example frequency table for licenses—top 10 rows of license types

Code	Description	Count
AJ	Malt Beverage On Premise	28655
AK	Malt Beverage Off Premise	22456
AL	Unfortified Wine On Premise	21616
BA	Salesman	19803
AM	Unfortified Wine Off Premise	19446
AO	Unfortified Wine Off Premise	13682
AY	Mixed Beverages Restaurant	10658
AN	Fortified Wine On Premise	9846
AZ	Mixed Beverages Private Club	3837
BH	Mixed Beverages Private Club	3107

A3. Make and record team decisions on license type filtering and grouping.

Use the frequency table to discuss license types, including which to include in your alcohol outlet density calculations and maps and how to code license type groups. To make decisions about classifying outlets, you may want to solicit expert advice on the license types, their intent, and their use in practice. Grouping license types can be challenging (e.g., by on-premises or off-premises outlets) so you may want to consider how to group licenses to best accomplish the goals of the project or perform calculations on multiple groupings if there are varying perspectives.

To record these decisions, save a copy of the frequency table with new variable columns, such as a *study_include* variable for which licenses to include or filter out as well as a *permit_group1* variable for grouping decisions (e.g., Table 4). The fields that will be included in analyses should be coded with “machine readable” data.

Table 4. A classification table for licenses

ID	Permit-code	Permit-description	Permit-count	Study-include	Permit-code 1	Study-note
7	AJ	Malt beverage on premises	28655	Yes	On	include
8	AK	Malt beverage off premises	22456	Yes	Off	include
9	AL	Unfortified wine on premises	21616	Yes	On	include
24	BA	Salesman	19803	Yes	Drop	Wholesale only
10	AM	Unfortified wine off premises	19446	Yes	Off	include
12	AO	Fortified wine off premises	13682	Yes	Off	include
22	AY	Mixed beverages restaurant	10658	Yes	On	include
11	AN	Fortified wine on premises	9846	Yes	On	include
23	AZ	Mixed beverages private club	3837	Yes	On	include

Step 4. Objective B: Join classification table to outlet data.

Tasks

B1. Read classification table.

B2. Join classification table to outlet table.



B1. Read classification table.

Read in your classification table to your GIS or statistical software. You may need to save your table in comma separated value (CSV) format as they are widely accepted and do not have spreadsheet program-specific additional formatting. In QGIS, read your file in as a text delimited layer. In R, read your file using any tabular format type using the correct function (e.g., *read_csv* or *read_excel* from the *readxl* package).

B2. Join classification table to outlet table.

Your outlet data file can be joined to your spatial outlet table using their shared license type variables in your GIS. At the end of this task, each outlet will have its filter status (to drop or keep), as well as whatever groups you defined in the table, saved in its record. See [Appendices 1 and 2](#) for more information.



Step 4. Objective C: Perform filtering and grouping.

Tasks

- C1. Apply filters.
- C2. Apply additional filters and deduplicate if necessary.
- C3. Optional: Apply groupings.



C1. Apply filters.

Use the variable created to indicate whether a particular license type or group of types will be included to remove outlet records irrelevant for your study. If you created the frequency table as a means of collecting all unique license types, you will have a record of the number of outlets removed by license type.

C2. Apply additional filters and deduplicate if necessary.

You may also want to use other unique filters besides outlet type. For instance, if your outlet records are not yet deduplicated, your team may need to review alcohol outlets that are indicated as being at the same location, filtering out all but one unless it is possible for two outlets to be at the same place (as with multiple outlets in the same high-rise building). Your team may also need to filter inactive licenses if the list includes both active and inactive ones. It is good practice to record the filters applied and the number of outlets dropped with each one.

C3. Optional: Apply groupings.

If your team wants to use characteristics of grouped outlets (e.g., on-premises and off-premises outlets) to calculate group-specific indicators of alcohol outlet density, you may prepare the database at this step by filtering your database to the group of interest. You can perform the remaining steps and calculations using that subset of the data.

Although this toolkit does not demonstrate more advanced techniques to calculate outlet density indicators for many groups at once, two high-level suggestions follow:

- Statistical programming tools like R can use list and nested table objects with *purrr::map** functions to calculate many group-specific variations simultaneously.
- QGIS and other point-and-click GIS tools can create and save automated workflows for batch processing.

These two techniques can be more efficient than creating separate, duplicated code for each grouping.

Step 5. Spatially locate outlets.



Objectives

- A. Gather shape data, population data, and accessory data.
- B. Geocode outlets.
- C. Tabulate, review, and improve geocoded results.
- D. Spatially project and filter to region.

Step 5 Summary Questions

- How will your team address outlets that are missing spatial data?
- What proportion of outlets did not geocode correctly? Is that proportion acceptable?
- How many outlets were removed for not meeting criteria?

Step 5. Objective A: Gather shape data, population data, and accessory data.

Tasks

- A1. Determine data needed and storage plans.
- A2. Acquire shape data for regions and small population units in the study zone.
- A3. Understand census population data.
- A4. Acquire population data.
- A5. Decide on additional datasets.



A1. Determine data needed and storage plans.

As you gather data on populations, region boundaries, outlets, and license types, it is best to plan a consistent and clear way to store, label, and share your data, folders, analyses, and reports early in your project. Directory names like *input* and *output* can help differentiate what you use from what you create, and *data*, *maps*, *code*, *shapes*, *projects*, and reports may also be useful. Decide how your team will organize, store, back up, and pass on your work in the future.

A2. Acquire shape data for regions and small population units in the study zone.

The US Census shares borders of a region, like a state, county, or census tract, in a commonly used format for spatial data called a “shapefile.” Shapefiles store data on borders with other key information about the geography like its population, land area, and how the boundaries fit on the earth. Your team will need border shape data of at least two kinds in your study zone:

1. the borders of your regions of interest and
2. small population units (e.g., blocks or block groups) to represent people’s homes in person-to-outlet distance calculations. Non-spatial data from other datasets can be linked to data in the shapefile if they have shared information (e.g., name of a jurisdiction).

To download US Census shapefiles, search the Internet for “United States census tiger shapefile,” and click the page result that specifies [US Census TIGER/Line Shapefiles](#) on the [US Census](#) website. Under Download, click Web interface. Select the year and boundary type that matches your needs to download state, counties, tract, or block groups data. Click Submit, then click Download. The resulting dataset may include the entire geographic United States, so you may want to subset the data to your study zone of interest, such as by using a Federal Information Processing Series (FIPS) code or state name to make the dataset size more manageable.

If you use R, you can use the [tidycensus](#) package to download these files with their population data using the US Census API.

A3. Understand US Census population data.

Consider the most relevant population data needed according to your team’s surveillance questions. Two sources are the US Census and the American Community Survey (ACS). The US Census is conducted every 10 years. The ACS provides census estimates each year (with a 1-year lag time until data are available), calculated using data from the 5 most recent years.

Block group, group, and tract boundaries are redrawn according to population changes every 10 years. Thus, the 2010 US Census will be different from the 2020 US Census, and the 2019 ACS will be different from the 2020 US Census. Consider how your analysis is influenced by potential boundary changes over time.

A4. Acquire population data.

R users can use the Census API and the [tidycensus](#) package to download population data and shapes directly within R. See [Appendix 1](#) for details.

QGIS users can first download datasets and shapefiles onto their computer from the US Census website, and then import the files into QGIS. See [Appendix 2](#) for details.

Population data are available [online](#). On the homepage, type “acs” into the “Explore Census Data” search field. Click on “DP05: ACS Demographic and Housing Estimates.”

A new page will open with a data table. Click on the Geos icon, then select your region (e.g., county, state, census block). If you select counties, first specify the state, then all counties. If your team is interested in calculating distance-based indicators, the smaller population units (e.g., block groups) will also be used in that calculation even if they are not your region of focus.

Click on the Download icon and select the data years to download. Typically, you will select the ACS 5-year estimates data profiles for which the most recent date year is your analysis year of interest. If you are interested in data for a decennial census year (e.g., 2010 or 2020), you may want to use US Census data instead. You may receive two files: a metadata file with variable descriptions and a data file with population counts.

Clean the data CSV file before analyzing it by performing the following actions:

- Remove the top-most row and make sure your column headings are named appropriately. QGIS and R will automatically read in the first row as column headings when adding your dataset.
- If you only need total population per region, remove other columns.
- Simplify the geographic region “Name” column, if overly complex. For example, you may want to edit “Alamance County, North Carolina” to simply read “Alamance.”

A5. Decide on additional datasets.

Only data on alcohol outlet locations, population, land area, and spatial boundaries are needed for calculating the indicators in this toolkit and answering the Tier 1 questions. However, your team may have other questions to explore that might require additional datasets. Consider appropriate timeframes or years of data sources, as well as possible implications of using a dataset with a different timeframe if data availability is an issue.

Step 5. Objective B: Geocode outlets.

Tasks

- B1. Choose geocoding approach, if required.
- B2. Prepare data for geocoding.
- B3. Geocode addresses and save results.



B1. Choose geocoding approach, if required.

If the outlet dataset has a variable describing what region (e.g., county, zip code) it is contained in, then teams calculating the count-based indicators can calculate them without geocoding outlet addresses to a point at all. The outlet's region variable can simply be tabulated, counting the total outlets for each region without spatial analysis. This total count will be the numerator (top of the rate fraction) for both the population-based and area-based denominators. This can be the fastest and simplest method to derive alcohol outlet density indicators, and many teams may choose to start with this tabular method. Even in this case, you may still choose to geocode your outlets so they can include specific alcohol outlet point locations on their final maps.

If you want to calculate the distance-based indicators, or if one or more region types of interest (e.g., cities, census tracts) are not already assigned in the licenses database, you will need to geocode outlet addresses to calculate point-to-point distances and assign outlets to regions spatially. Geocoding is this process of transforming a text representation of a place, most commonly its address, to its spatial coordinates that can be mapped, such as its latitude and longitude or coordinates on a local planner projection. Geocoding can be a hurdle for many teams, and this means these teams may choose not to calculate the distance-based indicators. But there are ways to make geocoding less of a hurdle.

Teams with appropriate internal geocoding tools will not find geocoding to be a hurdle, and in fact are ready to begin geocoding right away. State- or local-specific geodatabases (sometimes called "master address tables") of valid addresses within a study zone, if available, are often of very high, and addresses coded against a jurisdiction-specific database may be less likely to incorrectly geocode to a location states or counties away. These databases are often maintained by special units in planning or other departments that routinely work with geocoding processes. They may be able to assist with geocoding of address locations.

If a high-quality local address table and geocoder is not available, or if that geocoder is outdated or struggles with unconventional address formatting, a free or commercial geocoder may be useful. For a small set of addresses, this process can be done online by hand. However, this approach

might not be practical for longer lists of alcohol outlets. Free or commercially available APIs, point-and-click tools in GIS, and online tools to submit address batches for geocoding can assist in these batch processes for geocoding.

Many kinds of free and commercial geocoders are available (e.g., Google, US Census TIGER, ArcGIS, Geocodio, TomTom, MapQuest, SmartyStreets, and OpenStreetMap). However, choosing a geocoder can be difficult. No two geocoders are alike, and only some fit the unique needs of a project like the one described in this toolkit.

Teams may want to consider the cost of systems. Some are free (e.g., US Census and OpenStreetMap geocoders), and some require a payment method on file for per-geocode request or monthly total geocoding. Many of these systems have useful cost-protection and security tools, like capping the total number of requests in a period to avoid accidentally submitting too many requests and limiting the use of the API to certain IP network addresses to prevent misuse by anyone not on the project team.

Geocoders vary in their geocoding quality, including the level of accuracy of their results, their stored knowledge of address locations, and how flexibly or strictly they can interpret outlet addresses. Free geocoders may not perform as well as commercial geocoders. For example, a free geocoder may return geocoding results for 70% of a dataset of moderately well formatted addresses, and the remainder would need to be manually geocoded or submitted to a commercial geocoding tool.

Be sure to review the geocoder terms of service carefully. There may be limits on whether geocoded results can be stored and whether results can be plotted on maps that do not have that service's logo. You might also learn whether publicly available elements can be scraped and stored through web scraping without repercussion when doing so does not damage the service, as well as whether key elements not constructed by a private entity (like an outlet's name, address, phone number, and location) may be uniquely unprotected from scraping and storage.

When cost is an issue, teams may use a process of tiered geocoding by first attempting to geocode using a free tool, then, for those tools that do not geocode or geocode

inaccurately, taking subsequent passes with commercial tools. When using multiple geocoders, it can be useful to sample some addresses to be geocoded multiple ways so you can assess the quality of each geocoder. Commercial geocoders may have price plans that allow free geocoding up to a point. You can use these geocoders to develop results and then compare the quality of the geocoder to others you have tested.

Think through whether additional potential geocoding projects that your team may perform will require privacy protections. For example, privacy and security issues can arise when geocoding personal address information. If you think this will come up in any of your projects, consider a HIPAA-compliant commercial geocoder API or agree to a shared statistical strategy to preserve address privacy. More information on geocoding public health surveillance data with higher levels of privacy protections is available in the [Harvard Public Health Disparities Geocoding Project](#).

Although this toolkit does not endorse any specific geocoder, Geocodio was selected for geocoding outlets in the example analyses.

B2. Prepare data for geocoding.

GIS-based geocoders and online APIs expect to receive address information in specific formats. The documentation will explain the expected input address format.

In this example of 100 addresses, the geocoder (Geocodio) expects a single address string as the input. To create this table of full addresses, combine the street address, city, state,

and zip code with commas and combine the new full address column with the column of unique outlet IDs from your system. While it is possible for multiple outlets to have the same address, this is relatively rare in most jurisdictions, so it is efficient to attach addresses to outlet IDs.

B3. Geocode addresses and save results.

Once addresses are formatted for and submitted to the geocoder, the amount of time it takes for the geocoded results to be returned varies across geocoding tools. Consider geocoding a small sample of addresses first so you can estimate the length of time required for a full batch. Geocoding tools often have the option to choose simplified or full results. The full results can help you assess accuracy.

To reduce time and resources spent on geocoding, teams might not want to re-geocode the same address using the same system unless you think the results have changed. When saving geocoding results, save the unique ID and address information with the data returned by the geocoder. It may be useful to create and save other fields to help users later (e.g., geocoding tool name, geocoding date).

Geocoded tables can be saved as CSV tables or integrated into a central database if the alcohol outlet address and license information came from a central database.

If a geocoded database does not exist, geocode all alcohol outlet addresses. If a geocoded database exists, look up addresses in a local table, and geocode only the addresses that are new or better cleaned or those that failed to geocode.

Step 5. Objective C: Tabulate, review and iteratively improve geocoded results.

Tasks

- C1. Assess failed geocodes and geocoding accuracy.
- C2. Review, replace, and re-geocode addresses.
- C3. Consider optional review strategies.



C1. Assess failed geocodes and geocoding accuracy.

Some addresses may fail to geocode. Identify the number and percentage of successfully and unsuccessfully geocoded alcohol outlets, regardless of accuracy. Separate the alcohol outlets that failed to geocode, or a sample if they cannot all be reviewed individually, and review to determine the geocoding accuracy. You can calculate the percentage accuracy of the geocoding by comparing known addresses and point locations (e.g., a submitted address was a 95% match compared to a known address). You can also assess accuracy types according to the precision category, which describes the map element used in matching the alcohol outlet (e.g., where “residential rooftop” and “commercial building” are excellent matches, “street interpolation” a reasonable match, and “zip or county centroid” would usually not be precise enough).

Perfection in geocoding accuracy of alcohol outlets is not required for calculating outlet density but consider your study design as you make decisions about the accuracy types and inclusion of certain alcohol outlets. For instance, if you are calculating count-based indicators by county or census tracts, it would be important for the geocoding to successfully place addresses in the correct county or census tract. Geocoded addresses to the zip code or state level might be sufficient for some count-based surveillance of alcohol outlet density at those levels. However, such accuracy types might not be sufficiently accurate for distance-based indicators. Removing alcohol outlets entirely might have a greater effect on count-based indicators than on distance-based indicators.

Separate the addresses that geocoded less accurately and should be further reviewed (Figure 10).

Figure 10: Examples of geocode accuracy tables, by accuracy type (left) and accuracy percentage

Accuracy Type	N	Geocode Accuracy Percent	N	
Nearest rooftop match	1	Nearest rooftop match	0.33	1
Place	1	Place	0.60	1
Range interpolation	15	Range interpolation	0.78	1
Rooftop	83	Rooftop	0.80	1
			0.87	2
			0.90	8
			1.00	86

C2. Review, replace, and re-geocode addresses.

The accuracy of alcohol outlet density calculations depends on accurate representation of alcohol outlets in space (for distance calculations) or within correct regions (e.g., counties, zip codes, census tracts). To improve the geocoding accuracy, teams may want to iteratively review, replace, and re-geocode the addresses that need further review by manually or programmatically cleaning. With **manual review** and cleaning, teams might be able to correct some errors (e.g., typos, P.O. Box information) to retry geocoding. With **programmatically cleaning**, statistical programs (e.g., R, SAS) can be used to improve address formats, or formulas in spreadsheet programs can often use tools like regular expressions (sometimes called “regex”) to efficiently clean addresses represented as strings.

Rather than directly replace the original database address, create a separate table of outlet IDs, original addresses, and corrected addresses to document these edits. Edit types (e.g., manual review, programmatic cleaning or regex), reviewer name, and date may be other useful documentation fields. Improved addresses that successfully geocode can also be integrated with the original data source.

C3. Consider optional review strategies.

Your team may consider other review strategies besides those described above.

Longitude and latitude points outside of your study zone’s **spatial bounding box** can be identified quickly by mapping or filtering. For example, searching for “range of latitude and longitude for North Carolina” shows the North Carolina bounding box result of “34° N to 36° 21’ N” and “75° 30’ W to 84° 15’ W.” You can use those ranges to check whether outlet geocode results in North Carolina are within that box. Outlet latitude coordinates should be approximately between 34 and 36, and longitude coordinates between -75 to -84.

Manual search and **reverse lookup** can also be useful for a few spot checks. Choose a pair of geocoded coordinates and place into an online mapping tool to see if the location matches the original address by using road or satellite overlays. Recall that successful geocode results may represent central business locations instead of an outlet’s physical location.

Step 5. Objective D: Remove outlets that do not meet study criteria.

Tasks

- D1. Decide on spatial inclusion criteria.
- D2. Filter outlets by spatial fields and by spatial relation.
- D3. abulate results.



D1. Decide on spatial inclusion criteria.

Questions might arise about spatial inclusion criteria after addresses are geocoded, such as if an outlet is geocoded within a boundary of interest but the text address places it across the border. For this task, decide the spatial bounds for outlets to keep, determine relevant spatial fields and how to apply appropriate filters, and decide how to approach outlets on boundary edges with discrepant information. To make informed decisions, it might help to map the geocoded alcohol outlets.

D2. Filter outlets by spatial fields and by spatial relation.

Some alcohol outlet databases include fields with relevant spatial data that can be used to filter the outlets. However, the data might not align perfectly with study goals (e.g., zip codes are provided, but the study goals pertain to differences by county, and the zip codes might overlap more than one county). Another option is to filter outlets by spatial relation using a GIS tool or spatial capable programming language (e.g., R). For spatial filtering, usually all geocoded outlets should be required to be within or touching the study zone (e.g., the entire state if doing a study of counties within a state or the entire county if doing all census tracts within a county).

GIS tools provide a “spatial clipping” function to help filter outlets by spatial fields and by spatial relation. The QGIS Spatial Query plugin will enable the clipping function dialogue box.

In addition, R provides many ways to restrict your outlets to those in your study zone. The two most straightforward methods are the spatial row subsetting grammar and the *st_intersects()* function from the *sf* package. Examples are provided in the [Appendices](#).

Because consistent alcohol licensing procedures are not used across states, a nationwide alcohol outlet database is not publicly available. It can be challenging to assess alcohol outlets across state borders if alcohol license data from a bordering state are not publicly available, making distance-based indicators of alcohol outlet density less precise. You can state that this is a limitation, and your team may want to consider specific implications for your surveillance. Cross-state collaborations may strengthen the ability for teams to calculate distance-based indicators of alcohol outlet density.

D3. Tabulate results

Once you have filtered your outlets spatially, tabularly, or both, count the number of outlets that were removed and those that remain. To do this, count the number of outlets in the dataset before and after filtering, or by tabulating the frequency of the *is_in_study_zone* variable by whether it is true or false.

After filtering by both type of outlet and the outlet location, your team will know the final number of alcohol outlets in your study zone for the purpose of calculating alcohol outlet density.



Step 6. Calculate indicators.



Objectives

- A. Import data into software.
- B. Check projection of spatial data.
- C. Count outlets in each region.
- D. Join supplementary data by region.
- E. Calculate indicators.

? Step 6 Summary Questions

- Do any of your chosen analyses require a spatial analysis tool? If so, which ones?
- What projected coordinate reference system (CRS) will your project use? If calculating distance indicators, will you calculate them using the same CRS as the one you use to visualize maps?

Step 6. Objective A: Import data into software.

Tasks

- A1. Import each dataset.
- A2. Verify import.



A1. Import each dataset.

If your team plans to calculate all four indicators of alcohol outlet density presented in this toolkit, you will use the following four datasets:

1. **Regions** (e.g., all the counties in a state in a spatial format).
2. **Alcohol outlets** in your study zone (e.g., the outlet/license table you prepared in Steps 3–5, either as a flat table with geocoded information or a spatial file type).
3. **Population data** for your regions.
4. **Small area population centers** to use as a proxy for where people live in distance calculations.

If your data are not already imported into a spatial tool, confirm they are correctly formatted for import.



Importing tips:

- Data in tables or attached to spatial attributes should match in format.
- Avoid special characters and ensure any text naming or content follows the same capitalization scheme.
- Import your data into your software.
- When adding multiple datasets, group them by format to import so you can quickly identify potential dataset formatting errors.
- If you have longitude and latitude columns for your outlets in a CSV file, use a special procedure for QGIS to recognize it as spatial data (see [Appendix 2](#)).

As you import each layer, review the imported file to ensure it was imported correctly. If using QGIS, tables must be in CSV format to be added as a layer. Refer to [Appendices 1 and 2](#) to see how to read different formats into each software.

You may receive spatial data as an ESRI shapefile, which is a collection of files with the same name but different ending extensions (e.g., .dbf, .prj, .shp, and .shx). Before you import a shapefile, ensure that its supporting files are in the same folder on your computer where you will save it.

A2. Verify import.

After importing, view the content of each dataset to verify it was added successfully. For a table, open the content and check that the columns and rows preserved their names and structures. When importing datasets, consider re-naming columns so that they are generally consistent across CSV files. For example, you may find that one CSV file may list county under “County,” but another uses “County Name.” Making column names match for the same information across datasets may prevent confusion and avoid mistakes when joining data. For spatial data, create a map display of your spatial data as a visual check that your outlet points and regions are plotted and overlaid as you would expect. If your spatial data seems invisible or strangely aligned, you may need to adjust your coordinate reference system, which is discussed in the next objective.

When you first import data, it is helpful to have a written record of dataset characteristics. It will serve as a reference when troubleshooting or verifying that commands have executed as expected. For your outlet dataset, record the number of alcohol outlet density outlets (rows). Review the data fields available and make sure their contents are clearly labeled.

If the software reacts slowly when you try to view your data, your region shapefile may be too large and at a higher resolution than needed for your analysis. Consider simplifying the high-resolution boundary data and reimporting to improve the speed. Options to simplify shapefiles include the online [Mapshaper](#) tool,¹³ the “Simplify geometries” tool in QGIS, and the `st_simplify()` function in R.

Step 6. Objective B: Check projection of spatial data.

Tasks

- B1. Check CRS of each separate layer.
- B2. Select project CRS.
- B3. Transform CRS if needed.



B1. Check CRS of each separate layer.

The CRS specifies the projection used to map your spatial dataset accurately to locations on earth. Without any coordinate system or projection, the software will not plot the points correctly, leading to an error message or an empty or incorrect map. If you force software to plot spatial datasets on different projections or incorrectly specify the data’s projection, they will not overlay correctly and may appear strangely warped.

Write down the name of each spatial dataset that you have imported and examine the properties of each to see what CRS it is using. It can be helpful to record the initial projection and CRS of each dataset before it is reprojected or transformed to your selected projection. If a spatial layer does not have an initial CRS assigned to it, correctly assign it a CRS according to information from its source.

B2. Select project CRS.

For map-making, all datasets should be projected into the same CRS. If your spatial data are not appearing or not overlaid as expected, your data may not be projected correctly. Even if it is projected correctly on the map, your team may want to calculate distance indicators in a different CRS, such as measuring miles or meters between outlets or people to their nearest outlet.

Select a project CRS and projection that will be accurate for measuring distances in your study zone. Use this for all your spatial datasets, even if you transform using another projected CRS for creating maps. If your data are stored in latitude and longitude pairs, your origin CRS is likely the WGS84 coordinate system (EPSG code: 4978), and your final CRS/projection is likely the NAD 83 system centered on your US state. For instance, when mapping within North Carolina, the NAD 83 (EPSG code: 2264) projection is a good choice. The choice of your projection and CRS also implies a certain unit of distance (e.g., meters, feet, miles), which will determine the units that will be used when your distance indicators are returned. There are free tutorials on spatial CRS and projections available on the Internet, and projections can be searched on the spatialreference.org¹⁴ website. There are several ways to uniquely identify a CRS and projection, including its unique EPSG code, its “well known text” (WKT) string, and the proj4string.

B3. Transform CRS if needed.

If any of your spatial datasets are not in the correct projection, reproject them using a GIS tool or spatial-capable statistical software (e.g., R or QGIS). If using QGIS, save the reprojected layer as a permanent layer. QGIS attempts to automatically transform subsequently imported projected data into the first layer’s projection, so teams will want to make note of these changes and choose their project projections according to the considerations above in Task B2.

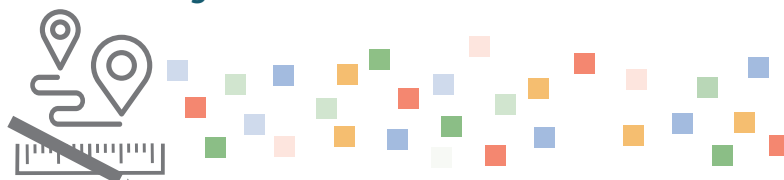
If you have transformed your data, plot your spatial datasets again and examine them visually to verify they have projected correctly.

Step 6. Objective C: Count outlets in each region.

Tasks

C1. Calculate outlets in each region.

C1. Calculate outlets in each region.



If each outlet has a field describing the study unit containing it, such as county or census block, you can count the number of outlets with the same category using **tabular methods** in a non-spatial, spreadsheet tool. To use tabular methods, create a frequency table of the number of outlets for each spatial unit, similar to the process in Step 5 for outlets per license type. This can be done in QGIS, but it might be easier in a spreadsheet or statistical program (e.g., in R, use the `count()` function).

If there is no study region field, use **spatial methods** to count the number of outlet points within each region. Even without a study region field, QGIS and R both can determine the region that a point is in and count outlets spatially without that field. The next two paragraphs describe how.

In QGIS, use the *Vector* → *Analysis Tools* → *Count Points in Polygon* function, select the regions as the polygon layer, select outlets as the point layer, and run the function. QGIS will create a new region shapefile with an additional field for the number of outlets in each region.

In R, the `st_intersects()` function returns a list of all intersections between each outlet and the regions. The `length()` function returns the length of each element of that region-named list, which corresponds to the number of outlets per region, which can be saved as an attribute of that spatial object.

Step 6. Objective D: Join supplementary data by region.

Tasks

- D1. Check data formatting.
- D2. Add population data to region shapefile.
- D3. Add or calculate land area of each region.
- D4. Add any other data needed.



D1. Check data formatting.

Calculating count-based indicators of alcohol outlet density requires supplementary data on population and land area. If your team aims to map other data alongside alcohol outlet density, those data can also be joined in. Before combining datasets, make sure your data are prepared. Record the column names in each dataset that you will be using to merge the datasets and check that the columns to join have matching formatting, including consistent capitalization.

D2. Add population data to region shapefile.

Review the population data downloaded in Step 5. Familiarize yourself with the dataset, such as by identifying the regions with the largest and smallest populations. When sourcing population data from large datasets with sub-population breakdowns, take care to avoid importing sub-population data instead of total population data per region. Use the `left_join()` function in R or equivalent tool in QGIS to join the population data to the regions so population totals become a new column available for calculations.

D3. Add or calculate land area of each region.

Your region shapefile may already have total, land, and water area fields if gathered from the US Census TIGER files. If not, either join this data into your dataset from a flat non-spatial file using `left_join()` instructions or calculate area in your GIS tool.

In QGIS, use the field calculator `area()` function. If your data are not currently transformed to a planar CRS where area calculations make sense, use the `transform()` function and the current and desired EPSG codes to create a new area column (refer to Step 6, Objective B for details on CRS projections). In R, use the `st_area()` function and the `st_transform()` function, if necessary.

D4. Add any other data needed.

Review the full data linked to your region shape data's attributes. If you need any further information to answer your team's surveillance questions, such as demographic characteristics or socioeconomic data to consider disparities, add the information from those datasets at this time.

Step 6. Objective E: Calculate indicators.

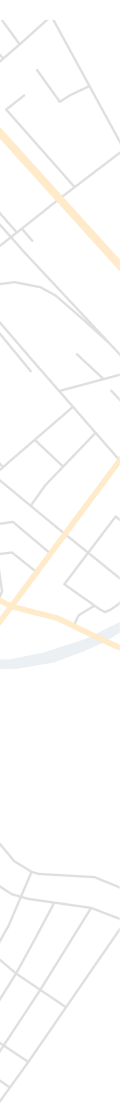
Tasks

- E1. Understand indicator math.
- E2. Calculate outlets/square mile.
- E3. Calculate outlets/population.
- E4. Calculate average distance between outlets and their nearest neighbor outlet.
- E5. Calculate average distance from person to their nearest outlet.



E1. Overview of indicator requirements

The data and skills required to calculate density vary by the alcohol outlet density indicator. Once you have completed adding or calculating outlet count, population, and geographic area for each region, you are ready to calculate the first two count-based indicators: outlets per square mile and outlets per 10,000 people. These calculations may not require spatial analysis if outlets are already assigned to regions and region areas and populations are known.



Distance-based indicators (indicators 3 and 4) require spatial analysis distance calculations at the outlet or small population unit level, which will then be summarized into your study's spatial unit of choice. Straight-line (Euclidian) distance calculations are presented below. Actual alcohol outlet density policies often are based on square miles of an area or the population size, which limits the usefulness for public health practice of assessing alcohol outlets by roadway miles or drive-time. Calculations involving alcohol outlets per roadway miles or by drive-time are more complex than calculations involving straight-line distances and are highly correlated in many regions. However, teams with high levels of spatial analytic capacity may want to calculate distance-based indicators using straight-line distance methods, as well as by roadway miles or drive-time.

The indicators in this toolkit are communicable and are relatively simple to calculate. Teams are therefore encouraged to consider whether other, more complex indicators for measuring alcohol outlet density would be advantageous to assist with translating the findings into practice or to accomplish their project goals.

E2. Calculate outlets/square mile.

To calculate Indicator 1, teams should have data for each region and should verify that the data are based on land area, not water area. The units of the region's area should be in square miles. If area is provided in another scale, like in square meters or acres, convert that data to square miles in a new column. Divide the number of outlets by the number of square land miles and save as a new attribute. These are the Indicator 1 results.

E3. Calculate outlets/population.

Using population data and the pre-calculated outlet count for each region, calculate the number of outlets per 10,000 people in each region. Divide the count of outlets in each region by the population of that region and multiply by 10,000. These are the Indicator 2 results. Variations may aid teams in communicating results for very small regions, such as using another denominator (e.g., divide by 1,000 people instead of 10,000 people) or calculating population per outlet (e.g., 20,000 people per 1 outlet).

E4. Calculate average distance between outlets and their nearest neighbor outlet.

Calculating the average distance between outlets and their nearest neighbor is a two-step process. First, calculate the distance from an outlet to its nearest neighbor outlet for all outlets in the study zone.

In QGIS, use the *Vector* → *Analysis Tools* → *Distance Matrix* tool. Set both the input and target point layers to the alcohol outlet layer, set the unique ID fields, and output a linear (N x k x 3) distance matrix using only the k=1 nearest point. Save this matrix as *min_distances.csv*. This will export a flat file of the minimum distances for each outlet, but it will also include ties. Use a spreadsheet tool or QGIS to save only one record for each unique ID and rejoin it to your outlet spatial file.

In R, the *st_distance()* function returns the distance matrix, and the *units::set_units("mi")* function can be used to convert distance to miles. Since the distance from an outlet to itself (zero) is not of interest, you can set the diagonal of this distance matrix to not applicable (NA) to ease calculations. The minimum of each row can be calculated using *apply()* over rows with the *min()* function and *na.rm=T* parameter to ignore the diagonal.

After completing the first step, aggregate the outlet's minimum distances. Since you previously spatially assigned each outlet to its study unit, these distances can be averaged. In QGIS, use the *Processing Toolbox* → *Statistics by Categories* tool to export a grouped average to a CSV file to be rejoined to the spatial unit file. In R, use the *group_by()* and *summarize()* functions, then rejoin the average distances to the region layer.

If duplicates were not previously removed, zero distances can also happen (i.e., some outlets will appear to be exactly zero miles to their nearest neighbor). Zero distances can also happen in dense regions where multiple outlets may be in the same building or in rural regions with approximate geocodes. It is sometimes possible to use spatial jittering to account for issues like these, rather than have very small distances between outlets be quantified as a zero distance.

E5. Calculate average distance from person to their nearest outlet.

Measuring the average distance from person to their nearest alcohol outlet is a multi-step calculation. First, if you have not already done so, decide the small population unit to use for calculating centroids and download the data. It is best to use the smallest available proxy for residential households, and census block and block group data are the smallest commonly available population units. Center points and the population sizes of these small population units will represent residential households.

You may need to separately find population data for your small population units, then combine it with its shapes. The US Census website's set of TIGER/Line Shapefiles are pre-joined with population data in geodatabase and shapefile format. In an Internet search engine, search for "TIGER/Line with Selected Demographic and Economic Data" and click on "Download these files from the FTP archive" to enter the database. Navigate to the year and then locate the sub-region and state geodatabase that you need.

In R, use the *tidycensus* package as described in Step 5 for the smaller population units.

Once you have loaded the small population units, transform the shapes into center points to represent the average location of people within that unit. This can be done with the *Vector* → *Geometry Tools* → *Centroids* tool in QGIS or the *st_centroid()* function in R.

Next, calculate the minimum distance from these population centers to their nearest alcohol outlet, like the average distance between an outlet and its nearest neighbor calculated in E3. Create a distance matrix, subset to the minimum distance, remove duplicate minimum distances (if necessary), and rejoin the summary data of minimum distances into the population data.

Lastly, summarize the small population unit data by their larger region. Multiply each small unit center's population by its distance, sum those numbers for each larger region, then divide by the total population in the larger region. This yields the average, population-weighted distance for each person in the region to their nearest outlet.



Step 7. Visualize, report, and communicate.



Objectives



- A. Determine plans to communicate findings.
- B. Calculate initial summary statistics.
- C. Build graphs to visualize data relationships.
- D. Make maps.
- E. Design reports.

Step 7 Summary Questions

- Who makes up your intended audience?
- What are your main messages and results to communicate to that audience?
- What is your larger public health framework for addressing alcohol and public health issues? How does alcohol outlet density fit in?
- What visualizations (e.g., maps, graphs) can most effectively communicate these results?
- If other contextual indicators are of interest, what graphs best show the relationship between alcohol outlet density and those other indicators?

Step 7. Objective A: Determine plans to communicate findings.

Tasks

- A1. Identify your target audience, key findings, and main messages.
- A2. Communicate key concepts carefully.



A1. Identify your target audience, key findings, and main messages.

After calculating one or more indicators of alcohol outlet density, communicate your findings. Teams might have several different audiences and should therefore identify a target audience and consider the key findings that will be most relevant to that audience group. Teams will also want to consider how messages will be conveyed and heard and discuss how to strategically present key findings. The audience groups might have varying levels of understanding about fundamental concepts, including public health principles or the alcohol license structure in your jurisdiction, so it might help to establish a common understanding before presenting the alcohol outlet density results. Conveying main messages to an audience group might be an interactive process, as audiences could provide feedback, request other visualizations, or propose other considerations to inform the interpretation of the results. With some audiences, it might be appropriate to discuss what work on measuring alcohol outlet density will come next so you create a feedback loop that can potentially improve the quality and usefulness of the analyses.

A2. Communicate key concepts carefully.

Carefully consider how to avoid unintentional messages when presenting your findings. For example, displaying the lowest density regions in green on a map or using a labeling of “low” in a legend may imply a level of alcohol outlet density that is not concerning. Despite this apparent implication, those regions might still benefit from practices to reduce the density of alcohol outlets. Therefore, it might help to use more neutral colors in a map or to use objective terms to imply relative density (e.g., “lower” instead of “low”).

Consider that audiences can confuse correlation with causation. As an example, consider that alcohol outlet density measurement is often included with other variables to assess correlations. Recall that this is different than assessing causation. Think carefully about the interpretation of findings in the context of how alcohol outlet density is associated with the population's characteristics and spatial patterns that arise for other variables (e.g., violence, crime rates).

Step 7. Objective B: Calculate initial summary statistics.

Tasks

- B1. Report overall study zone and region indicators.
- B2. Identify outliers, zero indicators, and skewness.



B1. Report overall study zone and region indicators.

First, report overall counts and alcohol outlet density indicator calculations by the full study zone, such as by the state or by a single county (if that region was also your full study zone). This can help set a baseline for interpreting the variation in smaller regions and helps document the total data that went into the analysis.

Second, report counts and alcohol outlet density indicators by smaller regions. If your study zone was the state, the smaller regions might be counties, zip codes, or census tracts. If your study zone was a single county, the smaller regions might be census tracts or neighborhoods. You can do this in a tabular form initially.

B2. Identify outliers, zero indicators, and skewness.

Understanding the skewness and zeros in the alcohol outlet density indicators can help you make design choices on maps and graphs. This information varies by the alcohol outlet density indicator, but the following examples demonstrate how this understanding is important for interpreting the results:

- When there are zero outlets or people living in a region.
- When land area is small, but outlet counts are relatively high (e.g., potentially in some beach towns and islands that are popular tourist destinations).
- When land area is large, but outlet counts are relatively small (e.g., potentially in rural areas or certain industrial regions).
- When regions in the study zone have starkly different rural and urban characteristics.

You can identify these dynamics multiple ways. You can sort the table of results looking for smallest regions by size (area in square miles), regions with the smallest number of outlets, and regions with smallest population. You can compare these regions to each other and to the overall study zone results, looking for alcohol outlet density indicators that are extreme (e.g., high, low, or zero). After identifying the outlying data points, consider how to visualize them appropriately on graphs and maps.

Step 7. Objective C: Build graphs to visualize data relationships.

Tasks

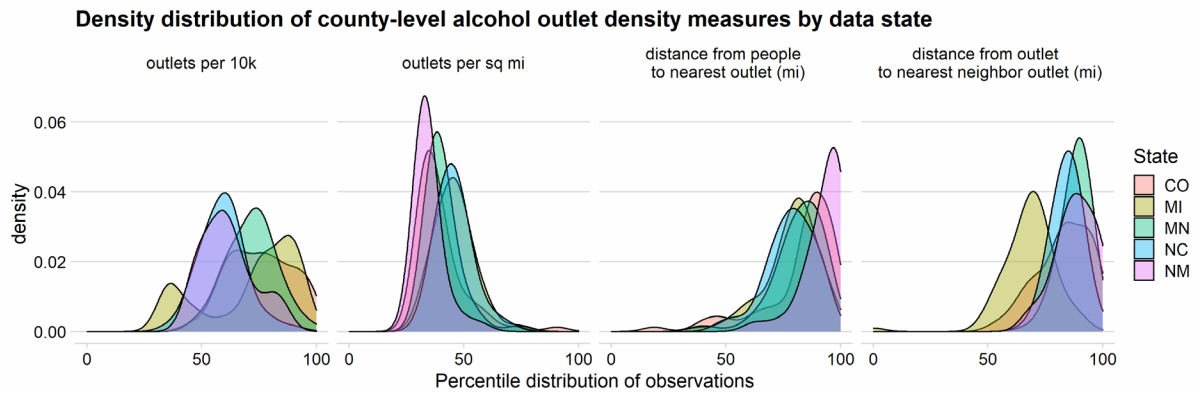
- C1. Create univariate graphs to explore data quality, distributions, and outliers.
- C2. Create graphs to explore the relationship between multiple variables.
- C3. Create brief text summaries of major findings.
- C4. Optional: Consider relationships to other variables carefully.



C1. Create univariate graphs to explore data quality, distributions, and outliers.

Maps are uniquely suited to communicating results and variation across space. However, you may be interested in non-spatial results and visualizations as well. Univariate graphs (of only alcohol outlet density) can demonstrate statistical results like its range and its skewness. Graphs can be simple, such as a histogram of county density scores, or more complex, such as Figure 11, which presents the density distribution of four alcohol outlet density indicators for counties in five states.

Figure 11. Skewness graph from preliminary analysis of states that participated in alcohol training and technical assistance in 2019 and 2020



C2. Create graphs to explore the relationship between multiple variables.

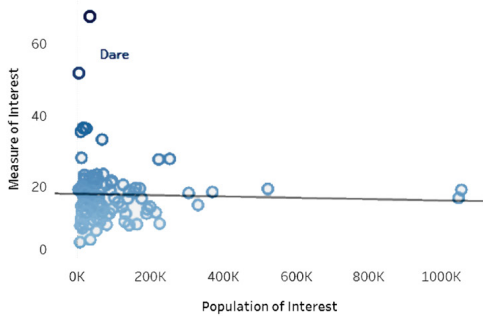
Graphs can also help teams explore, identify, and communicate relationships between multiple variables, such as between outlet density and health outcomes, population demographics, or comparisons of rural and urban regions.

Figure 12. Alcohol outlet density (outlets/10,000 population) versus county population

The choropleth map (right) shows outlets per 10,000 population for North Carolina counties. It can identify outlier counties by that single outlet density metric. The other graph (left) shows outlets per 10,000 versus the total population for North Carolina counties. The rightmost county on the map, Dare County, includes beach towns with many visitors and tourists and relatively fewer long-term residents. It has the highest number of outlets per population in the state (6.8 outlets per 10,000 residents).

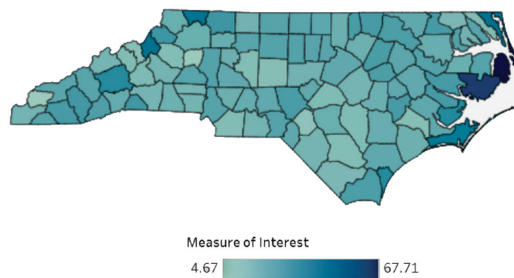
Alcohol outlet density vs. Population

(A2) Num outlets / 10k pop (#) vs Total Population for NC county units



Alcohol outlet density in North Carolina

(A2) Num outlets / 10k pop (#) by county



C3. Create brief text summaries of major findings.

Brief text summaries of findings can be essential for communicating results, significance, and implications to your audience. Maps and graphs do not fit in all communication settings or media, and communicating the main findings of alcohol outlet density analyses can provide useful talking points for other people to use. The four indicators of alcohol outlet density in this toolkit can be turned into talking points to contextualize outliers and summarize overall relationships in many ways.

For example, consider Figure 12 of North Carolina county-level alcohol outlet density (outlets per population). In that figure, outlet density per population decreased slightly overall as population increased, hovering close to around 2 outlets per 10,000 population on average. For every additional 100,000 people in a county, on average the outlets per population decreased very slightly (by 0.002 outlets).

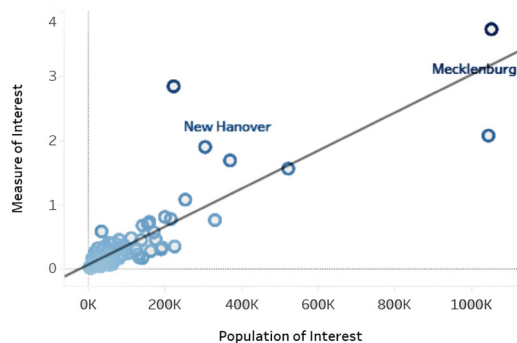
In contrast, Figure 13 shows outlet density measured by outlets per square mile against county population. Mecklenburg County is North Carolina's most populous county at over one million people. It also has the highest number of outlets per square mile in North Carolina, at 3.7. As shown in the left side of Figure 13, you can see that outlets per square mile generally increase as the population increases. On average, for every additional 100,000 people in a county, the outlets per square mile increases by 0.3. Sometimes the equivalent reverse association is more intuitive: on average, for every additional 240,000 people in a county, there is 1 additional outlet per square mile.

This average relationship is not true of all counties. Some counties have higher outlets per square mile when compared to counties of similar population size. For example, New Hanover County (at the southern border of North Carolina) contains Wilmington, a large city with beach tourism. It has the second highest outlets per square mile (2.8) even though its resident population is only around 200,000. This is more than 3 times the outlet density (per square mile) of Gaston County, which has 0.8 outlets per square mile and 216,000 people and borders Mecklenburg County. While beach town alcohol outlets may serve a larger commuting and tourist population, residents of these counties experience higher alcohol outlet density all year long.

Figure 13. Alcohol outlet density (outlets/mi²) versus county population

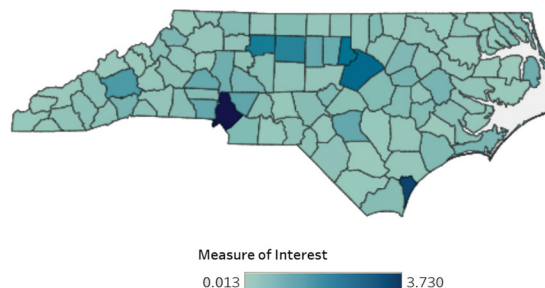
Alcohol outlet density vs. Population

(A1) Num outlets / sqmi (#) vs Total Population for NC county units



Alcohol outlet density in North Carolina

(A1) Num outlets / sqmi (#) by county



You may want to share associations in many forms. The shared average (slope) associations of two variables are included here and compared outliers against average densities within (population) groups. The above examples used both additive and multiplicative comparisons: for every 100,000 people, outlet density per square mile decreases by 0.3, and New Hanover has over 3 times the outlets per square mile as Gaston County, which has a comparable population. Other indicators of association and comparison may be more appropriate for your chosen analysis questions.

Teams may be asked to describe associations of other variables besides total population. Alcohol outlet density associations with population demographics can be used to assess disparities in exposure to dense alcohol outlet environments. These associations may be partly due to other factors if demographic proportions like race or ethnicity or income indicators are associated with population density. The associations may have many historical causes. However, whatever the causes and whatever other collinear associations, reporting associations like these can still be useful in conversations about differences in exposure.

C4. Optional: Consider relationships to other variables carefully.

After summarizing the relationship of alcohol outlet density across a study zone, you may want to report relationships to alcohol-related outcomes in other datasets (e.g., health, motor vehicle crash, or crime data) and social determinants (e.g., population demographic characteristics, community wealth, or food deserts). Communicating associations and relationships can be tricky and is beyond the scope of this toolkit. Associations can be false if confounded by other variables, and not everyone has experience distinguishing correlation from causation. However, carefully presented associations can prompt conversations about upstream causes and confounders. These conversations, in turn, can lead to understanding interpretations and limitations and to future analyses. Collaborations with other epidemiologists trained in causal inference and those in academic settings may help ensure these analyses and communication points are accurate and unbiased.

Step 7. Objective D: Make maps.

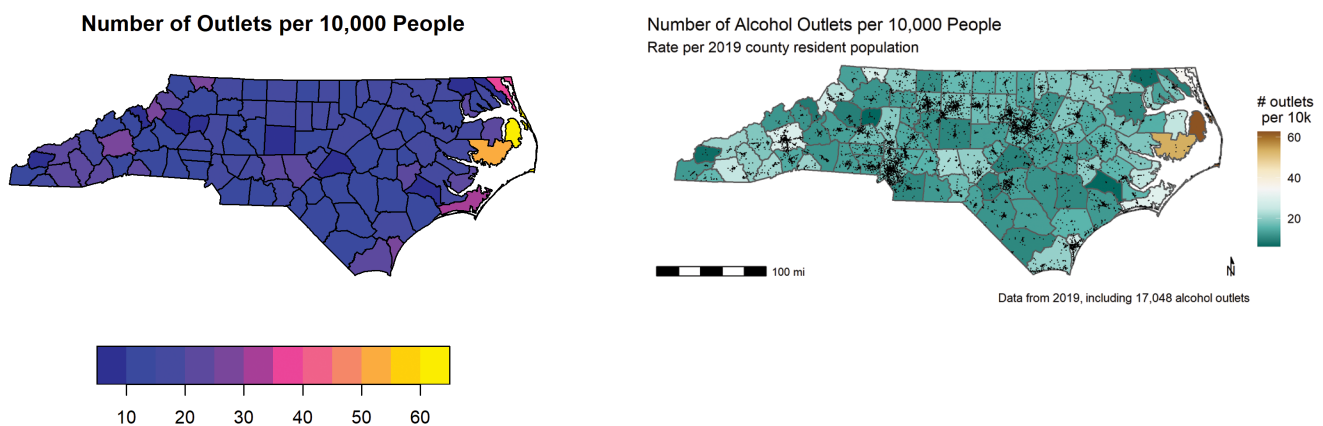
Tasks

- D1. Choose the level of effort and detail.
- D2. Choose legend colors and density groupings.
- D3. Consider additional layers that provide context.

D1. Choose the level of effort and detail.

Static maps are a common way to visualize spatial patterns of alcohol outlet density. Some maps can be produced relatively quickly (as on the left side in Figure 14), and others can be more complex and time consuming to build (as in the right side in Figure 14). The types of maps produced will likely depend on the skillsets of the team members and their intended use.

Figure 14. Two maps made in R with different levels of effort



D2. Choose legend colors and density groupings.

Some alcohol outlet density indicators may have a distribution with many zeros or few larger numbers, such as when calculating outlets per square mile or 10,000 population for residential regions with no outlets. If that is the case, consider the cut-points and color choices for your maps. Logarithmic transformations can be useful to address outliers and skew in more technical settings, but such analytic techniques might make it harder for your audience to interpret the findings. Your team may have set thresholds of interest to use for establishing cut-points. For instance, if an alcohol outlet density of more than 1.5 outlets per square mile is meaningful for a team, the team might set at least one of the choropleth cut points at that threshold.

There is not a widely accepted alcohol outlet density threshold that can be used across all contexts. Therefore, without clear thresholds or cut points of interest, you can determine color groups or choose to use a continuous color band. Percentiles and "Jenks" (mathematically determined cut points based on

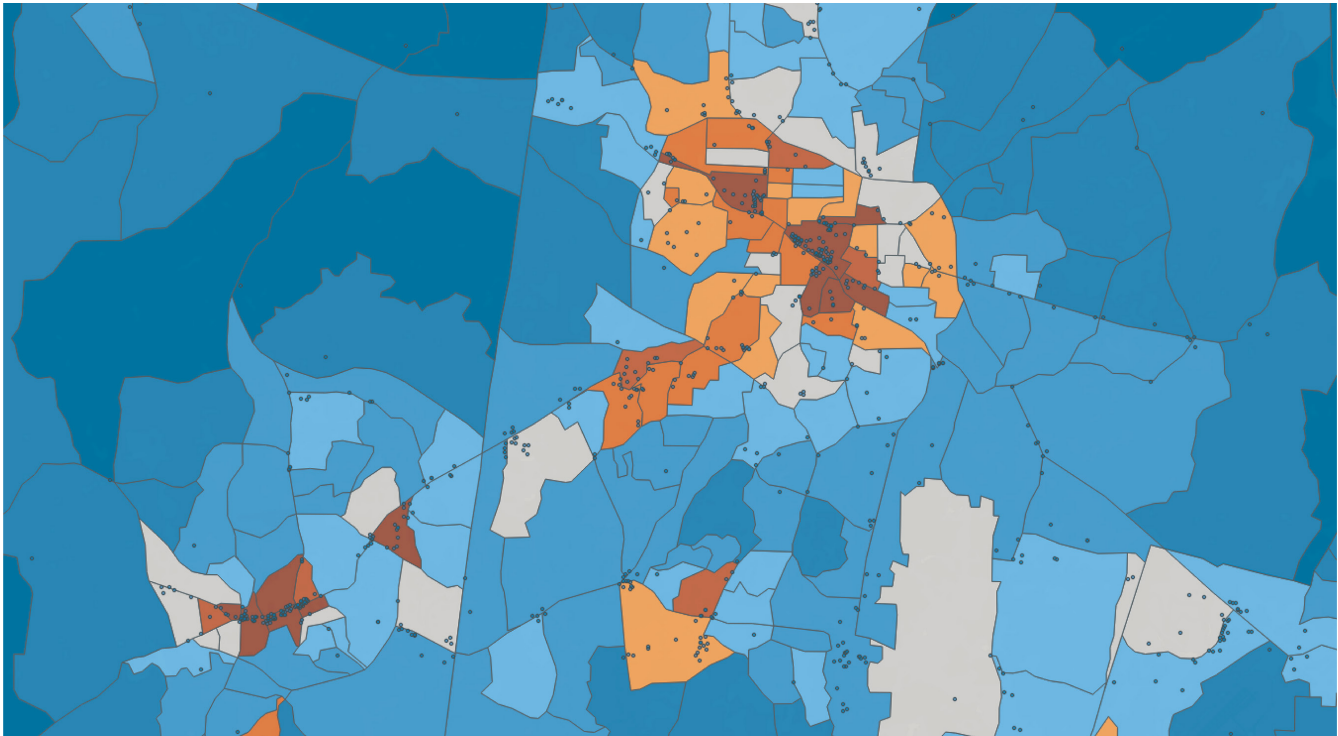
sharp changes in the distribution of values) may be helpful for summarizing the findings and allow teams to explain some regions that are in the top percentile of one or more indicators of alcohol outlet density. Geographers have researched what colors work well for maps: the freely available [Color Brewer](#)¹⁵ online tool and associated map palettes can help select accessible and interpretable palettes.

D3. Consider additional layers that provide context.

Other spatial data can help give context to high and low alcohol outlet density regions on your maps.

Point Data: Figure 15 shows a choropleth of outlets per square mile by census tract in a three-county region of North Carolina. The pink triangles, which represent the alcohol outlets, are overlaid on the blue palette regions. Other point data of interest may include schools, grocery stores, or other social amenities.

Figure 15. Alcohol outlet density as a choropleth with alcohol outlet point overlay



Shape Data: Other region shapes may be useful to show at the same time as your outlet density choropleth. Semi-transparent overlays (such as buffers around schools, neighborhoods, or historical regions of interest) can help identify unique regions of interest and higher alcohol outlet density. Alternatively, teams could produce two side-by-side graphs, augmented with a graph that demonstrates an indicator's association. Bivariate choropleths (maps with two main color axes) can help with this but can be complex to interpret and create. Interactive maps could include other contextual points in mouseover popups or clickable layers that can be stacked together.

Step 7. Objective F: Design reports.

Tasks

E1. Decide on one or more report formats and approaches.

E2. Design static and interactive reports.



E1. Decide on one or more report formats and approaches.

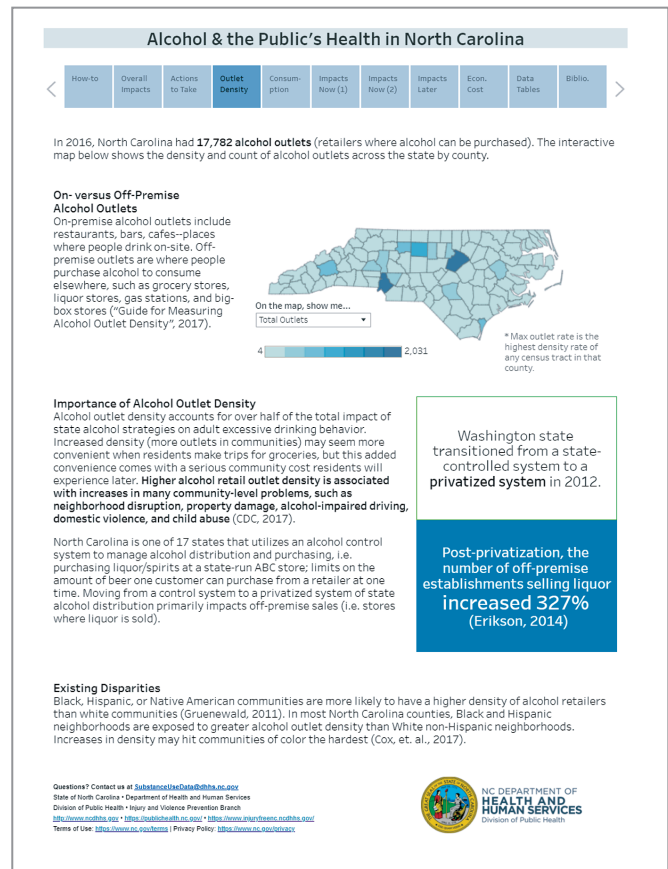
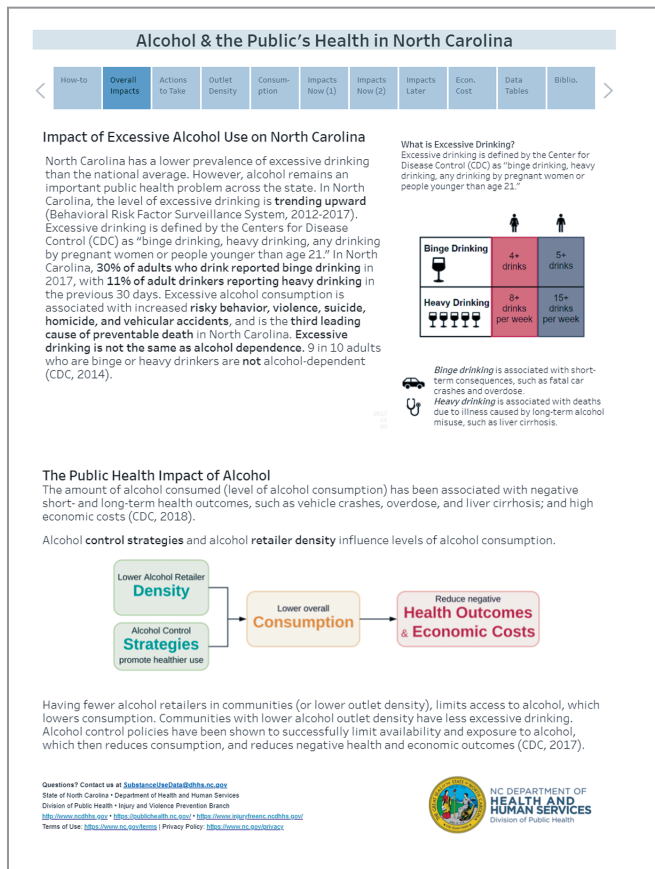
Teams can present alcohol outlet density in many ways to aid in dissemination. Static reports (e.g., written reports, fact sheets, slide decks, memos, academic manuscripts), videos, or interactive data visualization tools can aid teams in communicating density results to various audiences.

E2. Design static and interactive reports.

Your team may want to use multiple report formats. In addition to more traditional static reports, interactive tools can allow users to explore maps and customize graphs and reports (e.g., Figure 16).



Figure 16. Example of Tableau dashboard pages from North Carolina Alcohol and Public Health Tableau story



Additional Concepts Not Covered in This Toolkit

This toolkit does not cover several concepts in public health and geography that relate to alcohol outlet density. However, some familiarity with the following concepts could improve the planning and communication of alcohol outlet density measurements:

- The “Four P’s of Marketing”: Place, Product, Promotion, and Price.¹⁶ Teams may consider a broader approach for assessing an alcohol environment, including data from enforcement and violation records, alcohol sales data, advertising practices, and in-store surveys.
- Other related datasets on alcohol use, alcohol-related harms, and alcohol control measures. These include public health surveillance surveys (e.g., Behavioral Risk Factor Surveillance System, Youth Risk Behavior Survey), alcohol-related hospitalizations or emergency department visits, motor vehicle crash data, alcohol-related mortality, data from law enforcement agencies or administrative sources (e.g., alcohol-involved emergency services calls, violent incidents, crimes, arrest, jail and prison populations).
- Epidemiology and causal inference methods¹⁷ used to adjust for factors that may influence the relationship between alcohol use, alcohol outlet density, and particular outcomes.
- Alcohol outlet locations are often driven by licensing, zoning, and development practices. These practices can be influenced by historical and structural factors, including community-level disparities by race or ethnicity and socioeconomic position. Frameworks addressing disparities might help guide analytic methods and map-making decisions, providing important context for teams assessing outlet density and health inequities. Public health social epidemiology frameworks¹⁸ and multi-level models¹⁹⁻²¹ may be useful. In the case of racial or ethnic disparities, racism has been considered a fundamental cause of health disparities.²² Critical anti-racism frameworks (e.g., the Public Health Critical Race Praxis^{23,24}) and critical examination of the meanings of race and ethnicity in epidemiology analyses²⁵ may be useful for considering the complex relationship between race and development beyond skin phenotype alone.
- Best practices for conducting spatial analyses in a graphical GIS tool, such as code and data version control, analysis repeatability, code review to minimize errors, and strong documentation. These topics are beyond the focus of this toolkit, but analysts are encouraged to think about these questions of documentation, backups, and sustainability as they complete spatial analyses.

Conclusions

The steps in this toolkit provide information to measure alcohol outlet density for surveillance in their states and local jurisdictions. It includes four complementary indicators: two count-based indicators (rate of alcohol outlets per square land mile and rate of alcohol outlets per 10,000 people) and two distance-based indicators (outlet-to-outlet distance, which is the average distance from alcohol outlet to its nearest outlet, and person-to-outlet distance, which is the average distance from a person to their nearest alcohol outlet).

Although teams do not need to calculate all four indicators of alcohol outlet density, results based on more than one indicator will help to assess different aspects of alcohol outlet density. This approach may enhance what you can communicate with key audiences. Measurement of alcohol outlet density can promote an understanding of changes in alcohol outlet density over time or across jurisdictions. It can also be used to assess the association between alcohol outlet density and particular outcomes. Measurement and surveillance of alcohol outlet density in communities, counties, and states can help those who write regulations or other strategies designed to reduce excessive drinking and improve public health.

References

1. The Task Force on Community Preventive Services. Recommendations for reducing excessive alcohol consumption and alcohol-related harms by limiting alcohol outlet density. *Am J Prev Med.* 2009;37(6):570–571. doi:10.1016/j.amepre.2009.09.021
2. Campbell CA, Hahn RA, Elder R, et al. The effectiveness of limiting alcohol outlet density as a means of reducing excessive alcohol consumption and alcohol-related harms. *Am J Prev Med.* 2009;37(6):556–569. doi:10.1016/j.amepre.2009.09.028
3. Centers for Disease Control and Prevention. *Guide for Measuring Alcohol Outlet Density.* US Dept of Health and Human Services; 2017. <https://www.cdc.gov/alcohol/pdfs/CDC-Guide-for-Measuring-Alcohol-Outlet-Density.pdf>
4. Sacks JJ, Brewer RD, Mesnick J, et al. Measuring alcohol outlet density: an overview of strategies for public health practitioners. *J Public Health Manag Pract.* 2020;26(5):481–488. doi:10.1097/PHH.0000000000001023
5. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw.* 2019;4(43):1686. doi:10.21105/joss.01686
6. Pebesma E. Simple Features for R: Standardized support for spatial vector data. *R J.* 2018;10(1):439–446. doi:10.32614/RJ-2018-009
7. Walker K. Tidycensus: Load US Census Boundary and Attribute Data as “tidyverse” and ‘Sf’-Ready Data Frames. Accessed September 20, 2021. <https://CRAN.R-project.org/package=tidycensus>
8. Wickham H, Grolemund G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* First edition. O’Reilly; 2016.
9. Lovelace R, Nowosad J, Muenchow J. *Geocomputation with R (The R Series).* CRC PRESS; 2020.
10. Esser MB, Sherk A, Liu Y, et al. Deaths and years of potential life lost from excessive alcohol use — United States, 2011–2015. *MMWR Morb Mortal Wkly Rep.* 2020;69:1428–1433.
11. Stahre M, Roeber J, Kanny D, Brewer RD, Zhang X. Contribution of excessive alcohol consumption to deaths and years of potential life lost in the United States. *Prev Chronic Dis.* 2014;11:130293. doi:10.5888/pcd11.130293
12. Elmasri R, Navathe S. *Fundamentals of Database Systems.* 7th edition. Pearson; 2016.
13. Bloch M. Mapshaper: An Editor for Map Data. Accessed April 9, 2021. <https://mapshaper.org/>
14. Spatial Reference. Accessed April 9, 2021. <https://spatialreference.org/>
15. Harrower M, Brewer CA. ColorBrewer.org: an online tool for selecting colour schemes for maps. *Cartogr J.* 2003;40(1):27–37. doi:10.1179/000870403235002042
16. Greisen C, Grossman ER, Siegel M, Sager M. Public health and the four P’s of marketing: alcohol as a fundamental example. *J Law Med Ethics.* 2019;47(suppl 2):51–54. doi: 10.1177/1073110519857317
17. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology.* 3rd edition. Wolters Kluwer Health, Lippincott Williams & Wilkins; 2008.
18. Krieger N. A glossary for social epidemiology. *J Epidemiol Community Health.* 2001;55(10):693–700. <http://dx.doi.org/10.1136/jech.55.10.693>

19. Kim R, Subramanian SV. What's wrong with understanding variation using a single-geographic scale? A multilevel geographic assessment of life expectancy in the United States. *Procedia Environ Sci.* 2016;36:4–11. doi:10.1016/j.proenv.2016.09.002
20. Subramanian SV, Jones K, Kaddour A, Krieger N. Revisiting Robinson: the perils of individualistic and ecologic fallacy. *Int J Epidemiol.* 2009;38(2):342–360. doi:10.1093/ije/dyn359
21. Diez R. A glossary for multilevel analysis. *J Epidemiol Community Health.* 2002;56(8):588. doi: 10.1136/jech.56.8.588
22. Phelan JC, Link BG. Is racism a fundamental cause of inequalities in health? *Annu Rev Sociol.* 2015;41(1):311–330. doi:10.1146/annurev-soc-073014-112305
23. Ford CL, Airhihenbuwa CO. Commentary: just what is critical race theory and what's it doing in a progressive field like public health? *Ethn Dis.* 2018;28(suppl 1):223–230. doi: 10.18865/ed.28.S1.223
24. Ford CL, Airhihenbuwa CO. The public health critical race methodology: praxis for antiracism research. *Soc Sci Med.* 2010;71(8):1390–1398. doi:10.1016/j.socscimed.2010.07.030
25. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology.* 2014;25(4):473–484. doi:10.1097/EDE.0000000000000105

COCKTAILS
& SPIRITS

BOTTLE
SHOP

DRINK

DRINK



24 HR
LIQUOR

Restaurant
& BAR

