*Put the title page, table of contents, etc. here.*

**CHAPTER 2**

**LITERATURE REVIEW**

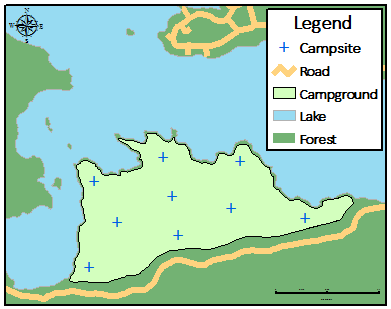
Neighborhoods need information about the people that live in these neighborhoods and about the houses that they live in.

*What information can we get from the Bureau of the Census?Information about people and the places they live in. (Mostly ACS). Then talk about geographic organization of this information. Most of the information about R goes in Methods.*

**Mapping Files:** Geographical data, which is spatial data recorded into a file format, is contained in GIS mapping files. The U.S. Census Bureau's geographic spatial data is represented by TIGER, which stands for Topologically Integrated Geographic Encoding and Referencing system.

Shapefiles (Partnership, TIGER/Line) are some examples.

**Shapefile**- Maps are made using shapefiles, which are digital representations of geographic features including lakes, landmarks, highways, and borders. The non-topological geometry and attribute data for the spatial features of a data collection are stored in a shapefile. Environmental Systems Research Institute\* (Esri) shapefile format is what the Census Bureau uses to produce shapefile layers.



**TIGER/LINE SHAPEFILES:** The Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER) System of the Census Bureau is where the TIGER/Line Shapefiles are extracted (MTS). The fifty states, the District of Columbia, Puerto Rico, and the Island regions (American Samoa, the Commonwealth of the Northern Mariana Islands, Guam, and the United States Virgin Islands) are all represented in the shapefiles. The shapefiles comprise point features, linear features like highways and hydrography, and polygon borders of geographic regions and features.

The TIGER/Line Files were originally made available in 1989. The American Standard Code for Information Interchange (ASCII) format fixed tables or record types contained in these files provide the first comprehensive street centerline coverage of the United States, Puerto Rico, and the Island Areas. The Census Bureau initially created the TIGER Database using the 1980 Geographic Base Files/Dual Independent Map Encoding (GBF/DIME Files), the U.S. Geological Survey (USGS) 1:100,000-scale Digital Line Graph (DLG), USGS 1:24,000-scale quadrangles, and a variety of other maps for selected areas outside the contiguous 48 states (predecessor to the current MTS). Throughout the 1990s and 2000s, the Census Bureau periodically issued ASCII 2-4 versions of the TIGER/Line Files. starting with the 2007 update, the shapefile file format replaced ASCII as the default for the TIGER/Line Files.

With the use of partner data, aerial photography, and fieldwork, the Census Bureau continuously updates and improves the MTS. Through a variety of partner initiatives, the Census Bureau receives updates on borders, features, and addresses from partners in the federal, state, and municipal governments. To increase the geographic accuracy of the road network, the Census Bureau performed a significant reconfiguration of the MTS in the 2000s. The Census Bureau has since imposed quality requirements for data sources used to update the MTS.

In addition, shapefiles for the community districts and neighborhoods in the Kansas City metropolitan area. The two geographies, community district, and neighborhood have slightly different sources. The neighborhood geography was originally developed by the Kansas City Missouri planning department in the 1980s as part of the user-defined geography initiative of the US Census Department. These neighborhood boundaries, though contested at the margin in some places, were based on the pre-existing (pre-1980) social geography. The Wyandotte County neighborhoods have a similar history. The North Kansas City, Independence, and Raytown neighborhoods were developed as part of the KC Health CORE initiative. Slight changes were made to the North Kansas City shapefiles to make them continuous (minimal gaps, no overlaps).

The Community District geography was developed as part of the Center for Economic Information's neighborhood and urban development work. Following the work of the Chicago School of Urban Sociology (Burgess, McKenzie, etc.), CEI recognized the need for mid-level geography, between the neighborhood and city level, that was rooted in the living patterns of residents. The construction of the community district geography is discussed at length in Dr Bowles’s dissertation and a slightly modified version of the community district geography is used by the KCMO planning department.

**R-**R is a language and environment for visual design and statistical computation. It is a GNU project(GNU is an operating software which is free software and it respects user’s freedom) that is comparable to the S language and environment that John Chambers and colleagues created at Bell Laboratories (previously AT&T, now Lucent Technologies). R might be thought of as an alternative S implementation. Although there are some significant changes, much of the code created for S works well in R. R offers a wide range of graphical and statistical tools, including time-series analysis, classification, clustering, and linear and nonlinear modelling. It is also very extendable. R offers an Open-Source alternative for those interested in participating in statistical methods research, which frequently uses the S language as its preferred vehicle.

The simplicity with which well-designed charts of publication quality may be created using R, complete with mathematical symbols and equations when necessary is one of its strengths. The user still has complete power despite careful consideration being given to the visuals' minor design decisions' defaults. R is accessible as Free Software under the conditions of the GNU General Public License from the Free Software Foundation. On a wide range of UNIX platforms and related systems (including Linux), Windows, and MacOS, it builds and executes. R is frequently used as a statistics tool by users. It is considered a setting in which statistical methods are applied. Through packages, R can (simply) be expanded. The R distribution comes with roughly eight packages, and the CRAN family of websites offers many more that cover a fairly broad spectrum of contemporary statistics.

**R Studio-**R studio now Posit. An integrated development environment (IDE) for R and Python is called RStudio. A console, a syntax-highlighted editor with direct code execution support, and tools for graphing, history, debugging, and workspace management are all included. RStudio runs on desktop computers and is offered in both open-source and paid editions (Windows, Mac, and Linux). The following are some of the features that RStudio offers: local RStudio access, Code completion, clever indentation, syntax highlighting, and immediate running of R code from the source editor. Go directly to the definition of the function. Using Projects, you can easily manage many working folders. R help and documentation are integrated. View Python data, render and publish Python content, and share Python objects with R. An interactive debugger is used to find and correct faults. Tools for package creation in plenty. Publishing applications and reports is simple.

**Packages:**

core tidyverse: It is likely to utilize these packages in regular data analysis because they are part of the core tidyverse. The following packages are part of the core tidyverse as of version 1.3.0:

ggplot2 - The Grammar of Visuals is the foundation for ggplot2, a framework for declaratively constructing graphics. We should provide ggplot2 the data, instruct it on how to map variables to looks and what graphical primitives to employ, and it handles the rest.

dplyr- It offers a language for data manipulation, offering a standardized collection of verbs to address the most typical problems with data manipulation.

tidyr- We may access tidy data with the use of a number of functions provided by tidyr. Data that is organized and uniform in form is called tidy data.

readr-A quick and easy technique to read rectangular data is provided by readr (like csv, tsv, and fwf). It is built to flexibly analyze a variety of data kinds that are seen in the field, but also cleanly failing when data changes unexpectedly.

tibble-The data frame is modernized in tibble tibble by preserving what time has shown to work and discarding what has not. Tibbles are data frames that are slack and grumpy; they do less and moan more, requiring you to deal with issues sooner, usually resulting in cleaner, more expressive code.

stringr-Working with strings is made as simple as possible with the help of stringr, which offers a unified collection of functions. It is developed on top of stringi, which offers quick, accurate implementations of common string manipulations using the ICU C library. Visit Docs.

tidycensus-With the help of the R package tidycensus, users may interact with a subset of the US Census Bureau's data APIs and receive tidyverse-ready data frames with the choice of basic feature geometry.

An R package called tidycensus (K. Walker and Herman 2021), which was originally made available in 2017, was created to make it easier to get and use US Census Bureau population data in the R environment. The bundle has two different objectives. In order to facilitate the process of extracting insights from US Census data, tidycensus intends to first make Census data accessible to R users in a tidyverse-friendly manner. Second, the program is intended to make it easier for geographic Census data analysts to wrangle data. To aid with mapping and geographical analysis, tidycensus allows R users to request geometry along with characteristics for their Census data. Through its APIs and other data download sites, the US Census Bureau offers a wide variety of datasets to the user community.

sf- In late October 2016, the sf1 package was initially made available on CRAN. It implements the "Simple Features" formal standard, which outlines a paradigm for the storing and retrieval of spatial geometries like points, lines, and polygons. When a feature geometry satisfies specific requirements, it is said to be simple. Simple polygons, for instance, cannot have spikes or hanging vertices, nor can they self-intersect. Simple Characteristics lack explicit knowledge of their neighbors or other spatially related features and are independent. Both web standards like GeoJSON and spatial databases like PostGIS have largely accepted this standard. A basic data frame with a particular column that holds the details for the geometry coordinates is how spatial objects are stored in sf. A list the same size as the number of rows in the data frame makes up that special column.

*Start a methods section*

*It has the following sections:*

*+ Measurements (where did you get the TIGER files and the neighborhood boundaries)*

*+ Procedure (simple features (sf) and tidycensus)*

*Start to build a bibliography*

*Use APA format.*

*Start outlining what goes in the appendix (well documented programs and data dictionaries)*