GEOGRAPHY 5303
Exercise 1 – 100 points
Due: February 1, 2020, by 5 PM
Graphical Summarization of Spatial Data

Purpose of Exercise

This exercise focuses on exploratory data analysis (EDA), making use of a variety of standard and spatial techniques that summarize data so they can be more easily interpreted and searched for patterns that can ultimately help us come up with explanations or further research questions. The tasks you are directed to complete will provide you will useful tools and methods that should be part of every researcher's skill set. Because of the diverse nature of the methods employed on this exercise, two different datasets are used.

The first dataset contains variables that measure numerous geographic, demographic, socioeconomic status (SES), and political indicators in the state of Oklahoma using county, school district, and tract-level data *circa* 2015 (exceptions are noted in the Exercise Metadata). This is therefore a cross-sectional dataset and will be used on Part A of all exercises, so a thorough EDA is useful for both putting summary measures into practice as well as learning about the data before you perform advanced statistical analyses on them. Tasks undertaken for this exercise on this dataset will involve simple numerical and graphical summary methods so as to compare different variables at varying scales and across regions of the state. The variables include social, economic, and political traits in the state and their interrelationships.

A second, separate dataset will be used for Part B on most exercises. This dataset is for Milwaukee home sales in 2012 that is provided by, and frequently discussed in, the Rogerson textbook. Besides consisting of points instead of polygons, this dataset also features many attributes that are either binomial (presence/absence) or are counts of discrete features (bedrooms, stories, etc.) that may require somewhat different methods of analysis. While this is also a cross-sectional dataset of sorts (all home sales in 2012), the issues of scale explored in Part A are not pertinent here. However, sales are coded by month of sale and by city alderman district, so on future exercises you will explore the dataset in *temporal* and *spatial* subunits. Just saying.

In the end, the overall goal is to give you, the student, practice in performing various important, but often overlooked or ignored, methods of EDA that are valuable preliminaries to conducting valid statistical analysis. You will be expected to conduct many of the analyses here on your project data set as part of Benchmark 3 and thus form part of the Data and/or Appendix section of the final report itself.

A. Cross-sectional Social and Demographic Analysis of Oklahoma

Characterizing the Dataset: Consider the dataset described above, further information provided in the Data Appendix, and by reviewing the actual data in the Excel/SPSS file. Before commencing analysis, the two questions below require you to evaluate the various traits/characteristics of the dataset and its various variables.

1. With reference to Section 1.2 in supplemental (BBR) readings (as well as Figure 1-4), define and describe the various *traits* of the overall dataset in a hierarchical manner, starting at the top (Archival or "To be collected" {ick}) and working your way down through various dichotomies described in the book. Your description should consist of a paragraph per trait (not per variable) explaining why you chose that trait (and not its counterpart), and any qualms you had in being forced to choose one over the other. Also, note that there are actually three different components to this dataset, one geographical (columns D-G, AH), one socioeconomic (columns H-AE), and one political (AF-AG), so you *may* need to make distinctions between these different components in your answers for a given trait.

2. Here you will be examining the following fields/columns (variables): D (scale), E (area), F (latitude), your five assigned variables for this exercise (see the Student Variable Assignment Excel file in Canvas), AF (percent republican), and AH (region).

   Characterize each of your 10 assigned fields (variables) in the Excel file according to the following *dimensions*:

a.	Individual	or	Spatially Aggregated
b.	Sample	or	Population Data
c.	Implicitly	or	Explicitly Spatial
d.	Discrete	or	Continuous
e.	Qualitative	or	Quantitative
f.	Nominal or Ordinal	or	Interval or Ratio

   In addition to explicitly stating which dimension (a.-f.) applies to which variable, write a sentence explaining your rationale for each.

   Since there are 10 total variables to assess, you are encouraged to group variables and write about them in aggregate whenever possible. For example, the seven racial category variables are all subdivisions of a single variable type (race) so just describe them together instead of writing out identical answers each time, once for each variable.

   Fair warning: not all the answers will be cut-and-dry, not all students will necessarily agree if they find themselves debating (as opposed to collaborating on) these answers, but either answer could be correct for some dimensions/variables depending on how they are defended by the student. You are strongly encouraged to review the SES Metadata for Oklahoma document on Canvas for this.

Summary Graphical Methods: SPSS has many graphical and descriptive tools that produce a large amount of summary information (tables and graphs) with just a few clicks.

3. Perform a basic, comprehensive analysis on the **five** variables you have been assigned (see Student Variable Assignment Excel file in Canvas):

   Click on “Analyze”, “Descriptive Statistics”, “Explore”. Put your five assigned variables (together) in the “Dependent List”, put **Area Name** in the “Label Cases by”space, and put **Geographic Scale of Data** in the “Factor List” space. Next, open the “Plots” window, check “Histograms” and “Normality plots with tests”, and uncheck “Stem-and-leaf plots”. Leave all other defaults alone. Click “OK”. This will produce tables/plots for all variables separately at each geographic scale except the boxplots, which are plotted on the same chart for all three scales (Factor List).

   Write a narrative describing the overall characteristics of each variable based on the Descriptives table and Tests of Normality table at the beginning of the output, as well as the histograms and boxplots (insert in Word). Organize your answer by variable, that is, discuss your first variable, then the second, the third, the fourth, and finally the fifth.

   For each variable, there are two key areas of interest in this discussion: (a) the overall distribution of each variable (shape, dispersion, symmetry, normality), especially comparing across spatial scales, and (b) the identification of outliers. The boxplots especially facilitate this discussion because SPSS plots all three scales on the same chart, and you can compare their median values, the sizes of their inter-quartile ranges, and their whiskers/outliers. The histograms will supplement the shapes revealed by the boxplots, and the skewness and kurtosis values provide the numerical measures.

4. Assess the degree to which your assigned traits vary regionally (by county only):

   Go to “Data”, “Select Cases”, choose “If condition is satisfied” and click the “If...” button. In the top window type in **Scale = “C”** and hit “Continue”. This selects the 77 county data rows (note how SPSS now “marks out” the tracts and the schools).

   Click “Graphs”, “Legacy Dialogs”, “Boxplot...”, and “Define”. Put one of your variables in the “Variable” window, then put the regionalization scheme (**Region**) in the “Category Axis” window and **Area Name** in “Label Cases by”; click “OK”. This will produce a panel of side-by-side boxplots of your variable, one for each region. Repeat this process for each of your other variables assigned in Question 3, thus producing one boxplot panel for each of your five variables.

   Summarize the five panels of boxplots for your variables (insert boxplots in Word). Within each variable, review how much variation exists between the 6 regions and discuss the degree of heterogeneity present. Then, compare between variables and discuss which variables have the most and least spatial variability and why you think this occurred. Be sure to note outliers in your discussion.

Summarization through Mapping: Mapping classified data can provide further means of assessing a dataset. This will require some basic GIS or computer cartography skills.

5. Map one variable for the purpose of reviewing the spatial variation of this trait across the state, at the county level. Choroplethic mapping is a common technique for socioeconomic data, though it has both benefits and pitfalls.

   Select one of the five variables you were assigned in Question 3 (excluding median age, per capita income, or median home value) and classify the data into *k* categories:

   a. Determine an optimum *number* of categories *k* into which to classify your variable for mapping purposes. Explain how you came by this choice.
   b. Determine a logical *method* for classifying your data into *k* categories. Briefly explain the methodology and why you chose this method over other alternatives.
   c. Explain addressing how well your system adheres to general "rules" of classification. Your goal is to produce a sound and pleasing classification system that follows as many of the rules as possible, constrained by the traits of your chosen variable (you should refer back to your discussion in Question 3 when pondering which rules are more important to follow, should a conflict arise).
   d. Submit a map of your classified counties, following proper mapping and presentation conventions concerning maps.
   e. Discuss the spatial patterns apparent on your map, focusing on areas at the extremes. Offer an explanation for what you observe based on your knowledge of the state.

B. Home Sales in Milwaukee, 2012

This part of the exercise conducts a basic examination of a point dataset consisting of typical attributes of homes that are both measurable/countable as well as influencing sale price. A data dictionary is given both in the textbook (page 21) as well as the City of Milwaukee website given in the textbook. However, the data dictionary is not as complete as we might like, as you will discover. One important note – Fireplace is actually the *number* of fireplaces, not a binary as the data dictionary indicates. An examination of the dataset reveals two homes sold in 2012 that had 4 fireplaces, two more with 3 fireplaces, and thirty-five with 2 fireplaces, and the Milwaukee Sales History website form lists the category as "Fireplaces" with numbers, so that seems definitive.

1. For each of the 16 attributes listed in the textbook and/or website, indicate and explain why you think each attribute is (a) nominal, interval, ordinal, or ratio; and (b) discrete or continuous. As in Part A, if several attributes are identical in their traits, you can group them for discussion. Also, three attributes are not absolutely, clearly defined -- #9 (age), #15 (x-coordinate), and #16 (y-coordinate), so determine their most likely units and explain your logic.

2. Perform a basic, comprehensive analysis on all analyzable variables (exclude Record, Sale Date, Alderman District, and the coordinates):

   > Click on "Analyze", "Descriptive Statistics", "Explore". Put all the variables (together) in the "Dependent List". Next, open the "Plots" window, check "Histograms", and uncheck "Stem-and-leaf plots". Leave all other defaults alone. Click "OK".

   Briefly review the distribution of each variable, focusing mainly on the boxplots and histograms. Some boxplots are quite weird; extemporize.

3. Construct a simple scatterplot of X and Y to make a "map" of the study area (click on "Graphs", "Legacy Dialogs", "Scatter/Dot...", "Simple Scatter"). Next, go to http://assessments.milwaukee.gov/mainsales.html and view a basic map of Milwaukee (you can also play around here and see where the data come from, and see data for different years). Why doesn't the shape of your scatterplot seem to resemble the outline of this map of Milwaukee? Can you do something to your scatterplot to make it conform better to reality? Do so, explaining your method and including your refined scatterplot in your Word write-up.

What you should upload to the Canvas dropbox:

- Word document answering all numbered questions above (don't forget to label lettered subparts and to answer all subparts, labeled or otherwise) with SPSS charts and one map inserted directly where discussed.