

Assignment #2

Descriptive Statistics & Exploratory Data Analysis

This week's assignment is to explore the distributional characteristics of a subset of attributes measured using the Detroit Census tract database. You will perform several tasks. First, you will calculate several descriptive statistics using functions within EXCEL (e.g., mean, median, standard deviation, CV). Second, you will graphically summarize the selected attributes by creating histograms and boxplots using GeoDa. Your overall objective is to use these quantitative measures and graphical displays along with the linking and brushing features within GeoDa to summarize the main characteristics of each selected variable. In other words, how would you describe these selected attributes and their associated spatial patterns. There is no central question – this is an exploratory analysis.

Part A. Descriptive Statistics in EXCEL

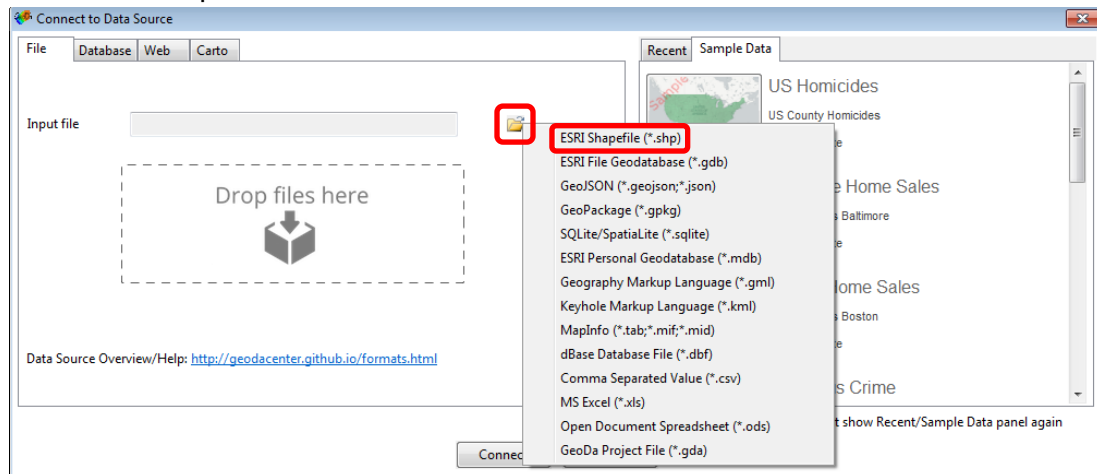
1. Open the EXCEL spreadsheet created in Assignment #1 and save a copy of this file using the name: Detroit_Assign2.xlsx
2. You will calculate a series of descriptive statistics for a subset of attributes: **WPOP**, **BPOP**, **PCINC**, **OHU**, **VHU**, and **PER_POV**. Variable descriptions can be found in Attribute_Descriptions.pdf. You can delete all other attributes from the EXCEL spreadsheet.
 - You will summarize each attribute by calculating the mean (average), median, mean deviation, standard deviation, coefficient of variation, and skewness.
3. To calculate your descriptive statistics, you will use the functions **AVERAGE**, **MEDIAN**, **AVEDEV**, **STDEV.P**, and **SKEW** in EXCEL.


As an example, to calculate the mean (or average) for a series of values you would use the command =average(A2:A310), where A2:A310 represents the range of data values for the attribute of interest (e.g., column A, rows 2-310). Use this equation as an example.





- Place your calculated statistic values immediately below the last observation value in each column.
- Format all cells so that the descriptive statistics are reported using a single decimal place.
- To calculate the CV for each attribute, use the appropriate descriptive statistics and the formula provided in class (see Session 3, video 4).
- **Create a table in a Word doc that lists:** a) variable name, b) mean, c) median, d) mean deviation, e) standard deviation, and f) coefficient of variation, and g) skewness.

Part B. Exploratory Spatial Data Analysis in GeoDa

1. You will utilize a new geographic data analysis software program to complete this analysis → GeoDa, a free, open source software program developed by Luc Anselin (<http://geodacenter.github.io/download.html>). GeoDa contains a variety of exploratory spatial data analysis tools for use with point and areal data.
2. Open GeoDa and select File → New Project.
3. In the Connect to Data source window, click on the folder icon and select ESRI Shapefile. Navigate to the folder where you saved Detroit2015_CTTracts.shp, select the file and click Open.



4. You will use GeoDa, and its linking and brushing capabilities, to (a) describe the statistical distribution of each variable using graphics, (b) identify any outlying observations, and (c) describe the key features of the spatial pattern of each variable. Outliers might be defined as “surprisingly high maximums or surprisingly low minimums”.
5. Create a histogram for each variable (variables of interest: **WPOP**, **BPOP**, **PCINC**, **OHU**, **VHU**, and **PER_POV**) using the histogram tool ().
 - Hover your cursor over a histogram bin to view basic information, including the values that define the bin range and the number and percentage of Census tracts that fall within that bin.
 - You can alter the number of bins used to classify the data (note: default is 7) by right-clicking within the histogram window and selecting Choose Intervals. Select the number of intervals / bins you feel best represents the distributional characteristics of each variable.
 - You must include a histogram for each variable in your Word doc. You can copy the histogram graphic by right-clicking within the histogram window and selecting Copy Image to Clipboard. You can then paste the image within your Word doc.
 - When describing each histogram (see Part C), use terms like: right skewed, left skewed or symmetric; unimodal or bimodal; or whether outlying observations may be present. Of course, these descriptions have implications for the kinds of statistical tests that can be run.

6. Next, use the linking and brushing functionality within GeoDa to select low or high values on the histogram and determine the locations of these Census tracts within the city of Detroit. Think about whether high or low values are concentrated in particular parts of the city for a particular variable.
7. Create a boxplot for each variable using the boxplot tool ().
 - Boxplots are great tools for identifying potential outliers.
 - As a reminder, the lines extending from the box (“whiskers”) define the interquartile range ± 1.5 or $3.0 \times \text{IQR}$. Potential outliers are plotted as points falling beyond these whiskers. Observations falling beyond $\pm 1.5 \times \text{IQR}$ are often considered potential outliers, while observations falling beyond $\pm 3.0 \times \text{IQR}$ are often considered probable outliers. You decide whether the whiskers define 1.5 or 3.0 times the IQR. This is set by right-clicking within the boxplot window and selecting Hinge.
 - When describing each boxplot, use terms like right skewed, left skewed or symmetric; compare each boxplot to its corresponding histogram – do they both suggest similar distributional patterns?; identify whether outlying observations exist.
 - Again, you must include a boxplot for each variable in your Word doc. Copy the boxplot by right-clicking within the boxplot window and selecting Copy Image to Clipboard. You can then paste the image within your Word doc.
8. Again, use the linking and brushing functionality within GeoDa to determine the locations of potential outlying Census tracts.
 - Do you see similarity in the location of outlying observations across variables?
9. If you would like to view the complete set of attribute data, click the Table icon (). When linking and brushing is active, the selected Census tracts will also be highlighted in the table. Note: You can sort the table according to a specific variable by double-clicking on the header for that variable; e.g., double-click on VHU to sort this variable from smallest to largest values.
10. While you are only required to include histograms and boxplots created for each variable in your report, I would encourage you to explore the scatterplot and PCP tools as well.
 - Use the scatterplot tool () to examine the bivariate relationships among the six selected variables. Are they weak or strong based on your visual assessment? What is the direction of each relationship (i.e., positive or negative)? Be sure to explore any “bivariate outliers” – points fall far away from the best fit line for two variables that are fairly well related; i.e., the value for one variable doesn't conform to what would be predicted using the best fit line based on the other variable. Determine where these tracts are located in the map. Can you offer any explanation for any of these cases?
 - Multivariate relationships can also be investigated using parallel coordinate plots (PCP, ). Can outliers be found in these graphs? Note: to select multiple individual lines in the PCP, use the Shift key. Unfortunately, PCPs can be difficult to interpret when there are a large number of observations, as is in our case.

Part C. Summarizing Your Findings

Based on your calculated descriptive statistics, summary graphics, and exploratory analysis, provide a thoughtful summary of the key characteristics of each attribute. Focus on the data characteristics you believe are of greatest interest. In writing your response, think “*what are the most interesting characteristics of these variables and what are the most important insights that should be conveyed?*” Be sure to comment on the spatial characteristics of each variable as well. Include your write-up in your Word doc. Limit your response to one-to-two paragraphs per variable, or no more than 2 pages of text.

Submission:

Your submission will consist of two files:

1. PDF file containing your table listing the descriptive statistics for each attribute, histogram for each attribute, boxplot for each variable, and written descriptive of the key characteristics of each attribute (LastName_Assignment2.pdf)
2. EXCEL file (Detroit_Assign2.xlsx) containing all calculated descriptive statistics

Submit all files to the Assignment 2 folder on HuskyCT. Your submission is due Monday, February 8th by 11:59pm.