Exercise 2 - Introduction to Data Science

Exercise 2.1

This exercise relates to analyze a dataset containing various types of information about average households across all of the zip codes in the United States. Use the analytical and visualization techniques covered in Module 3 to analyze this data and make conclusions about the different regions of the United States.

a. Load the file **zeta.csv** of income data into R.
b. Change the column names of your data frame so that zcta becomes zipCode and meanhouseholdincome becomes income.
c. Analyze the summary of your data. What are the mean and median average incomes?
d. Plot a scatter plot of the data. Although this graph is not too informative, do you see any outlier values? If so, what are they?
e. In order to omit outliers, create a subset of the data so that: $7,000 < income < $200,000, What's your new mean?
f. Create a simple box plot of your data. Be sure to add a title and label the axes.
g. In the box plot you created, notice that all of the income data is pushed towards the bottom of the graph because most average incomes tend to be low. Create a new box plot where the y-axis uses a log scale. Be sure to add a title and label the axes.
h. Use the ggplot library in R, which enables you to create graphs with several different types of plots layered over each other. Be sure to read the documentation for ggplot and load the library ggplot2 (you may have to install this package into R).
i. Make a ggplot that consists of just a scatter plot using the function geom_point() with position = "jitter" so that the data points are grouped by zip code. Be sure to use ggplot's function for taking the log10 of the y-axis data. (Hint: for geom_point, have alpha=0.2).
j. Create a new ggplot by adding a box plot layer to your previous graph. To do this, add the ggplot function geom_boxplot(). Also, add color to the scatter plot so that data points between different zip codes are different colors. Be sure to label the axes and add a title to the graph. (Hint: for geom_boxplot, have alpha=0.1 and outlier.size=0).
k. What can you conclude from this data analysis/visualization?

Exercise 2.2

This exercise relates to the Household electricity usage data set for 50 U.S. states including Washington D.C. and Puerto Rico **(income_elec_state)**. You have been asked to cluster all U.S. states by mean household income and mean household electricity usage. You have decided to use a k-means clustering algorithm.

a. Cluster the data and plot all 52 data points, along with the centroids. Mark all data points and centroids belonging to a given cluster with their own color. Here, let k=10.
b. Repeat step (a) several times. What can change each time you cluster the data? Why? How do you prevent these changes from occurring?

c. Once you've accounted for the issues in the previous step, determine a reasonable value of k. Why would you suggest this value of k?
d. Convert the mean household income and mean electricity usage to a log10 scale and cluster this transformed dataset. How has the clustering changed? Why?
e. Reevaluate your choice of k. Would you now choose k differently? Why or why not?
f. Have you observed an outlier in the data? Remove the outlier and, once again, reevaluate your choice of k.
g. Color a map of the U.S. according to the clustering you obtained. To simplify this task, use the "maps" package and color only the 48 contiguous states and Washington D.C.