Data Science Practical Task

Requirements:

1. Make sure to download the three files (income_elec_state.csv, zeta.csv, zipIncome.txt) that contains the needed dataset.
2. You are asked to solve the two sections of the task (Review of big data and Advanced analytics using K-means).
3. Make a team of 3 members max (you are free to choose from different sections and different departments).
4. Implement the task using R environment.
5. The code must be well documented with clear comments within the code itself.
6. You are required to answers all the needed explanation and observation questions in a word document. This document should contain at the beginning, in the first page, all the team members' names, ID and department.
7. Your deliverables:
    a. R file that contains your well documented code.
    b. The word document (or pdf) that contains your names and your answers on the given questions.
8. The project delivery will be during the practical exam's week.

Wish you all Best of Luck!

Review of Big Data Analytic Methods

In the following steps you are going 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 the course to analyze this data and make conclusions about the different regions of the United States.

Step 1: Retrieve and Clean Up Data using R

1. Analyze the zeta table (zeta.csv), which has data on households in different zip codes. Look at the column descriptions and record the column names.
2. How many rows of data are there in the zeta table?
3. Are there any duplicate rows of data in the zeta table? If so, how can you tell?
4. If there are duplicates, make a new table called zeta_nodupes that has no duplicates. Now are there any duplicate rows of data? How can you tell?
5. Save the table in a file named "zeta_nodupes.csv"

Step 2: Data Analysis in R

1. Load the text file of income data (zipIncome.txt) into R.
2. Change the column names of your data frame so that zcta becomes zipCode and meanhouseholdincome becomes income.
3. Analyze the summary of your data. What are the mean and median average incomes?
4. 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?

5. In order to omit outliers, create a subset of the data so that:

   $7,000 < income < $200,000

6. What's your new mean?

Step 3: Visualize your data

1. Create a simple box plot of your data. Be sure to add a title and label the axes.

2. 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.

3. What can you conclude from this data analysis/visualization?

Advanced Analytics/Methods (K-means)

Now you have learned how to retrieve, clean, investigate and visualize your data from the previous steps. Now it's time for applying advanced analytics. Use a k-means clustering algorithm to cluster all 50 U.S. states, including Washington D.C. and Puerto Rico, by mean household income and mean household electricity usage using following steps.

1. Access the census data saved as 'income_elect_state.csv' provided to you. Create a table with three columns: state, mean household income, and mean electricity usage.

2. Cluster the data using k-means function 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.

3. Determine a reasonable value of k using the "elbow" of the plot of the within-cluster sum of squares.

4. Convert the mean household income and mean electricity usage to a log10 scale and cluster this transformed dataset. How has the clustering changed? Why?

5. Reevaluate your choice of k. Would you now choose k differently? Why or why not?

6. Have you observed an outlier in the data? Remove the outlier and, once again, reevaluate your choice of k.