**Team Members:**

|  |  |  |
| --- | --- | --- |
| **ID** | **Dep** | **Name** |
| 2018170434 | CS | ندى محمد عبد الحميد عبد المحسن |
| 2018170001 | CS | ابانوب اسعد |
| 2018170 | CS | كريم شريف فتحي |

**Review of Big Data Analytic Methods:**

|  |  |
| --- | --- |
| **Step** | **Actions** |
| 1 | **Retrieve and clean up data:**   1. **Analyze the zeta table:**         1. **How many rows of data are there in the zeta table?**  * We can get the number of rows from the **data section** or   by executing the following command: **nrow (Zeta)**.     1. **Are there any duplicate rows of data in the zeta table? If so, how can you tell?**  * No, we know by executing the following commands:   **duplicated (Zeta)** >> return true in duplicated rows  **sum (duplicated (Zeta))** >> return number of duplicated rows. |
| 2 | **Data Analysis:**   1. **Load the text file of income data (zipIncome.txt):**  * Execute this command   **zipIncome <- read. table ("zipIncome.txt", sep = ",", header = TRUE)**   1. **Change the column names of your data frame:**  * Execute this command:   **colnames (Zeta) [colnames (Zeta) %in% c("zcta","meanhouseholdincome")] <- c("zipCode","income")**   1. **Analyze the summary of your data:**  * After executing this command **summary (Zeta$income):**      1. **Plot a scatter plot of the data:**   Do you see any outlier values? If so, what are they?  Yes, there are outliers, values greater than 200,000.     1. **omit outliers, create a subset of the data so that: $7,000 < income < $200,000:**  * execute this command:Zeta<- subset (Zeta,Zeta$income<200000 & Zeta$income>7000)  1. **what is your new mean?** |
| 3 | **Visualize your data:**   1. **Create a simple box plot of your data:**      1. **Create a new box plot where the y-axis uses a log scale:**      1. **What can you conclude from this data analysis/visualization?** |

**Advanced Analytics/Methods (K-means)**

|  |  |
| --- | --- |
| **Step** | **Actions** |

|  |  |
| --- | --- |
| **1** |  |
| **2** | **Cluster the data using k-means at K=10:** |
| **3** | **Determine a reasonable value of k on Data:**    **Suitable K = 3 or 2** |
| **4** | **K-means of Log10(Data) With K=10:** |
| **5** | **Reevaluate your choice of k on Log10(Data):**    **Suitable K = 3 or 4** |
| **6** | **Have you observed an outlier in the data? Remove the outlier**          **Suitable K =2 or 3** |