Data Analytics for Business

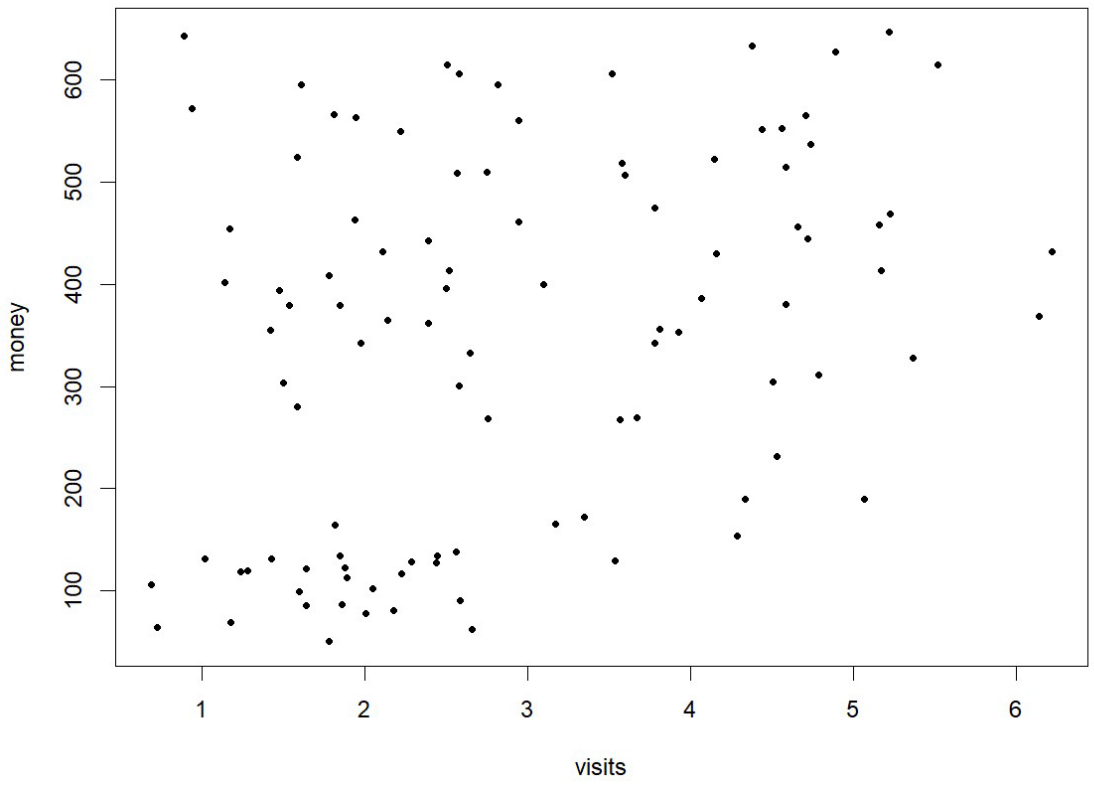
Homework 4 – Part 2

Cluster Analysis

In this homework, we will implement both K-means clustering and hierarchical clustering. We will use a simplified data set on shoppers' store visit and spending information. The data set contains 100 shoppers' visit and spending records for a department store. For the ease of visualization, we will only focus on two key variables: (i) each shopper's average number of store visits per quarter, and (ii) each shopper's average amount of spending per quarter. Similar analysis can be naturally extended to include additional variables if needed in practice. Below are the data description and a scatter plot of all data points.

Data Description

Column Name	Variable Description
visits	Average number of store visits per quarter
money	Average amount of spending per quarter



Download the data file "ShoppingVisits.csv" and complete the following tasks.

Task 1: K-means Clustering and Scaling

1. Perform K-means clustering with 3 clusters on the original data without scaling the variables. Set the number of random initial starts to be 20. Show the clustering output.

2. Scale the variables, and then perform K-means clustering with 3 clusters on the scaled data (with 20 random initial starts). Show the clustering output.

3. Create a scatter plot of the **original** data points with points belonging to different clusters marked in different colors. Create such a plot for both the clustering outcomes with and without scaling, and display the two plots side by side for comparison.

   Note: Data scaling is only for the cluster analysis (i.e., to obtain the cluster assignments); when drawing the scatter plot, we should still draw the data points in their original scales.

4. Discuss the effects of variable scaling on the outcomes of cluster analysis.

Task 2: K-means with Different Number of Clusters

5. Perform K-means clustering with 2 clusters, 4 clusters, and 5 clusters, respectively, on the scaled data (with 20 random initial starts). Save all the cluster outcomes.

6. Display in a 2x2 panel the four scatter plots for the clustering outcomes with 2, 3, 4, and 5 clusters, respectively. Same as previously, mark the **original** data points belonging to different clusters in different colors. Label the plots properly.

7. Based on these outputs, intuitively, which number of clusters do you think is the most appropriate for this data set?

Task 3: Choosing *k*

8. In order to choose the most reasonable value for *k*, let's create the so-called "elbow chart," that is, a line chart depicting the total within-cluster sum of squares for the clustering outcomes under different number of clusters (*k*). Plot the elbow chart for $k=1,\ldots,5$.

9. Which number of clusters would you choose, and why?

Task 4: Hierarchical Clustering and Dendrogram

10. Now perform the hierarchical clustering on the same scaled data. Use the usual Euclidean distance. Apply the average linkage first. Plot the dendrogram and properly label it.

11. If we want to have three clusters, what is a proper height where we can cut the tree? Mark this height onto the dendrogram as a dashed horizontal line.

Task 5: K-means vs. Hierarchical

12. Form **three** clusters from the hierarchical clustering output. Again, create a scatter plot of the **original** data points with their cluster membership marked in different colors. Display this plot alongside the scatter plot from K-means clustering with the same number of clusters (i.e., 3 clusters).

13. Discuss whether or not the clustering outcomes from the two clustering methods are "***robust***," which means whether the outcomes remain consistent across different methods.

Task 6: Hierarchical Clustering with Different Linkages

14. Now perform the same hierarchical clustering using different linkages. In addition to the average linkage that has been applied, now use complete, single, and centroid linkages, respectively. Create a scatter plot (with colored cluster membership) for each case, and display all four of them in a 2x2 panel. Label them properly.

15. Are the clustering outcomes robust across different linkages? Which linkage has the least reasonable result, and why?

Task 7: Interpret the clusters

16. Based on your choice of clustering method and number of clusters, how would you interpret each cluster? Try to describe each cluster of customers with a label.

17. **Think**: if you were to direct marketing resources to one of the clusters, which one would you target, and why?