Report due date: 11th December, 2024 - 11.59PM
Class presentations: 4th December, 2024
Any other information will be duly communicated to you. Do not hesitate to forward all questions to me and schedule zoom meetings if you need any help. You are free to divide and conquer within each group.

In this project, you are expected to perform a complete analysis of a dataset. You will submit a written exposition that thoroughly describes your process. This is an opportunity for you to make use of all your knowledge in the class so far and apply it to a real problem. An important part of modeling is clearly communicating it to others in writing and the written project is how you disseminate your work.

Part 1 - Cluster Analysis

Use clustering techniques—specifically k-means and hierarchical clustering—to analyze the Mall Customers dataset and identify distinct customer segments. Based on these segments, provide insights and recommendations for targeted marketing strategies.

Dataset Overview:

The Mall Customers dataset offers insights into customer demographics and spending habits at a retail shopping mall. This dataset is commonly used for customer segmentation, a method that enables businesses to categorize their customer base into distinct groups based on purchasing behavior and demographics. Analyzing this data allows organizations to create targeted marketing strategies, enhance customer satisfaction, and ultimately boost revenue.

The dataset includes 200 entries with the following predictors:

- **Customer ID**: A unique identifier for each customer.
- **Gender**: The gender of the customer.
- **Age**: The age of the customer.
- **Annual Income (k$)**: The customer's annual income in thousands of dollars.
- **Spending Score (1-100)**: A score assigned by the mall based on customer spending and behavior, with higher scores indicating more active spending patterns.

Instructions

1. Initial Data Exploration

- Begin by looking at summaries of each variable to understand the range, distribution, and general patterns within the data.

2. Exploring Relationships Before Clustering

- Explore and comment on relationships between pairs of variables using scatter plots.
- **Question**: What trends do you observe in the relationships between the variables?

3. Clustering Analysis

- **K-Means Clustering**
  - Implement k-means clustering using an optimal number of clusters and visualize the clusters.
- **Hierarchical Clustering**
  - Implement hierarchical clustering and visualize using a dendogram. Choose an appropriate number of clusters.
  - And visualize clusters for comparison.
- Compare and comment on your clusters from K-Means and Hierarchical clustering.

4. Post-Clustering Analysis - Use the results from K-means clustering

- After forming clusters, reuse the scatter plots to examine relationships within each cluster.
- Use box plots to compare the clusters across each variable to identify distinct differences between clusters.

5. Summary and Recommendations

- Summarize all insights from the initial exploration, clustering results, and post-clustering analysis.

- Based on these insights, provide recommendations for potential marketing strategies. Describe how different customer segments could be targeted to enhance engagement and sales.

Part 2

Project Objective The objective of this project is to predict medical insurance costs for individuals based on demographic and lifestyle factors using different regression techniques. By comparing various regression models, you will gain insights into the best predictive model for this dataset, which can help in understanding the impact of each predictor on medical costs.

Dataset Overview

The **Medical Cost Personal Dataset** contains information on medical insurance charges for individuals based on factors such as age, BMI, smoking status, and region. This dataset is well-suited for regression analysis as it includes both numerical and categorical predictors that influence the target variable, **charges**.

Dataset Features:

- **Age**: Age of the individual.
- **Sex**: Gender of the individual (male/female).
- **BMI**: Body mass index, a measure of body fat based on height and weight.
- **Children**: Number of children covered by the insurance.
- **Smoker**: Smoking status (yes/no).
- **Region**: Region of residence (e.g., northeast, northwest, southeast, southwest).
- **Charges**: Medical insurance cost (target variable).

Step 1: Data Exploration and Preprocessing

- **Load and Explore the Dataset**: Load the dataset and display the first few rows to understand its structure.
- **Data Summaries**: Summarize each variable (mean, median, min, max) and visualize distributions for numerical variables (age, BMI, children, charges).

- **Handle Categorical Variables**: Encode categorical variables appropriately. Use one-hot encoding for **Region** and **Sex** to create binary features.
- **Scaling and Standardization**: Standardize all numerical features (age, BMI, children) to prepare for modeling.
- **Exploratory Analysis Questions**:
  - How are insurance costs distributed across different values of BMI and age?
  - Does smoking status appear to have a strong effect on medical costs?
  - Are there correlations between any of the predictors?

Step 2: Regression Models

Using the data from Step 1, implement k-fold cross-validation for each regression model to evaluate its performance. Use metrics such as **Root Mean Square Error (RMSE)** and **Mean Absolute Error (MAE)**. Which model performed best according to RMSE and MAE?

- **Linear Regression**: Fit a baseline linear regression model to predict **charges** and interpret the coefficients.
- **Ridge Regression**: Use ridge regression to manage multicollinearity and compare results with the linear regression model.
- **Lasso Regression**: Apply lasso regression to perform feature selection. Identify which predictors remain in the model after regularization.
- **Elastic Net**: Fit an elastic net model, which combines the penalties from ridge and lasso regression, for a balanced approach.
- **Subset Selection**: Perform feature selection on the least squares regression model using either forward, backward, or stepwise selection to identify the most influential predictors. Fit the final model
- **Principal Components Regression (PCR)**: Apply principal components regression to reduce dimensionality, then fit a linear regression model on the transformed components.

Step 3: Summary and Insights

- **Interpretation of Results**: Summarize the main findings for each model, discussing the impact of significant predictors on insurance costs.

Some guidelines for the project

- As you see fit, incorporate the R output and graphics into your report. Every graphic produced should be relevant to some part of your discussion, hence do not incorporate too many graphics, especially if they are unnecessary. You should include phrases like "based on Figure 2,..." if you are including a particular figure or table. You can summarize your findings from your data exploration without attaching too many graphs and or tables.
- Your document should be double spaced. Also, divide your project onto sections with titles.
- Aim for a report length of **at most** 10 pages (excluding the appendix), including figures, tables. There is no minimum number of pages you should attain.
- Use a professional and legible font, such as Times New Roman, size 12.
- Maintain consistent 1-inch margins on all sides.
- All figures and tables should have clear labels and captions. Once again, ensure they are referenced in the text.
- Your report should be error-free. Make sure to edit it as many times as necessary, at least 5 times before you submit your final report.
- Do not plagiarize anybody's work.
- Be creative.
- Submit your written report and your R file you used.

Evaluation Criteria

Your report will be assessed based on:

- **Content Depth:** Comprehensive coverage of your topic and its application to the dataset.
- **Analysis Quality:** Rigor and correctness of the data analysis process.
- **Interpretation and Insight:** Ability to draw meaningful conclusions and insights from the results.
- **Presentation and Clarity:** Organization, structure, and clarity of writing and oral presentation.