Data Analysis Report: Clustering, Trends, and Imputation Logic

Dataset Overview

The dataset consists of individual records including variables such as age, gender, and salary. Before any modelling or cluster analysis could be performed, a data cleaning process was necessary to handle inconsistencies and ensure the accuracy of insights derived from the data.

1. Data Cleaning and Imputation

The following issues were found in the raw dataset:

- Gender Standardization: The gender column had inconsistent labels such as "female", "Female", “male” and “Female”. These were standardized to “F” for all females and “M” for all males.
- Salary Cleaning: Entries such as "Not sure" or negative salaries were considered invalid and converted to missing values (NaN). Replacing these values with NaN ensures that the data retains its integrity by indicating that these entries are invalid or unknown, rather than introducing misleading values (like 0).
- Age Cleaning: Some values were recorded in words (e.g. "forty") or included text(e.g. “I am thirty”). These were handled accordingly by changing them to numeric values.

2. Clustering Analysis (KMeans)

In this analysis, KMeans clustering with k=2 was applied to categorize individuals based on their salary and age. The objective was to identify distinct groups within the dataset and explore trends associated with these clusters. Additionally, various data imputation strategies were used to handle missing or inconsistent values in the dataset.

Heterogeneous Groups Identification

Through the application of KMeans clustering, two distinct groups were identified:

- Cluster 0 (Younger, Lower Salary): This cluster comprises individuals who tend to be younger and have lower salaries.
- Cluster 1 (Older, Higher Salary): This group includes individuals with higher salaries and older ages.

These groups reflect a clear correlation between age and salary, where older individuals typically earn higher salaries, while younger individuals tend to have lower earnings.

Cluster Visualization

A scatter plot of salary vs. age provided a visual confirmation of the clustering, clearly distinguishing the two groups as seen in Figure 1. The first group, consisting of younger individuals, had lower salary levels, while the second group, composed of older individuals, had higher salary levels.

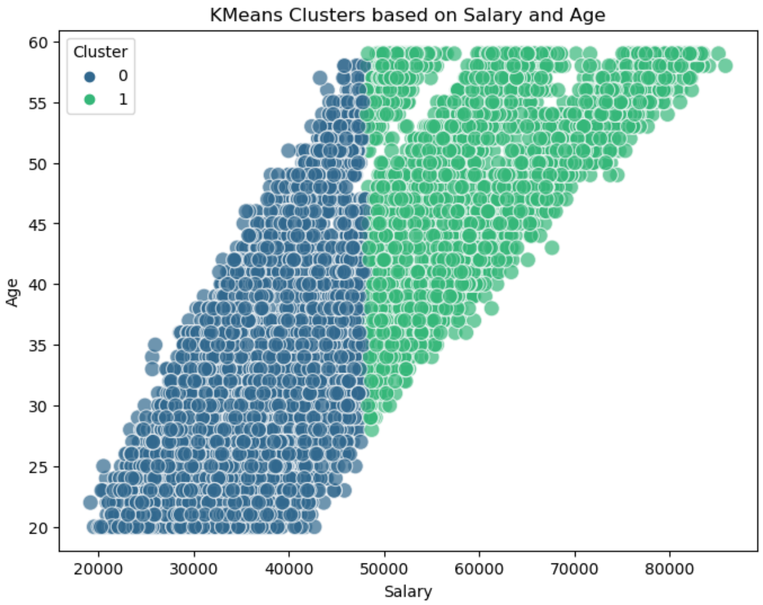

Figure 1: Scatter plot visualizing KMeans clustering of individuals based on salary and age. Cluster 0 includes individuals with lower salaries and varying ages, while Cluster 1 consists of individuals with higher salaries and older ages.

Trends

- Cluster 0: This group, with younger individuals and lower salaries, likely represents individuals at the beginning of their careers.
- Cluster 1: This group consists of older individuals with higher salaries, indicating a trend of increased earnings associated with experience and seniority.

These trends illustrate a natural relationship between age and salary, where experience, typically correlated with age, plays a significant role in salary determination.

Imputation Logic

To ensure the accuracy and consistency of the analysis, the following imputation steps were implemented:
Negative or Invalid Salary Entries: These entries were treated as missing (NaN) and subsequently imputed with the mean salary value from the dataset.

3. Predictive Modelling

Two types of linear regression models were developed:

- Standard Regression Model: Predicts salary using an individual's age and encoded gender. The salary of customer with id 1571 was determined to be R61786.90
- Group-Based Prediction: Predicts an individual's salary using the attributes of another person from the same cluster or group. This is particularly useful when salary is missing or unreliable for some individuals. The salary of customer with id 1571 was determined to be R57105.14