**Student Id: 23024452**

**Student Name: Rizwan Hussain Shah**

**GitHub Link:**

**Introduction:**

The purpose of this report is to apply fitting and clustering methods to the given dataset. The dataset has a target variable that indicates a binary result along with a variety of individual demographic and financial characteristics. We will examine the data, apply K-Means clustering, fit a line to some variables, and produce a number of illuminating visualizations.

**Data Summery and statistics:**

The summery statistics provide an overview of the dataset including measures of central tendency and dispersion for numerical columns. Use describe function for this.

**Data Preprocessing:**

Before performing clustering and fitting, we need to preprocess the data. This include handling missing values, encoding categorical variables and scaling numerical features.

**Data Visualization**

**Histogram for age distribution:**

This Histogram shows the age distribution in the dataset. This help us understand the demographics of the individuals of study.

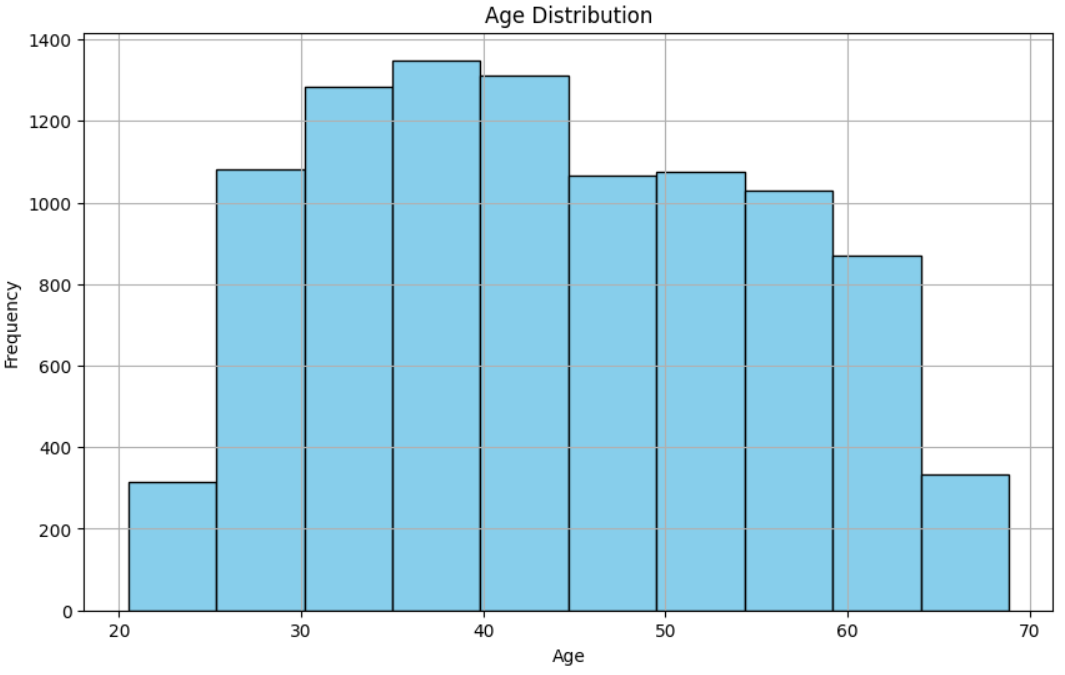


Figure 1: Histogram of Age Distribution

**Bar Chart Showing the count of Target variable:**

This bar chart showing the count of target variable from the Dataset which is the column of this dataset. It shows the distribution of binary outcomes.

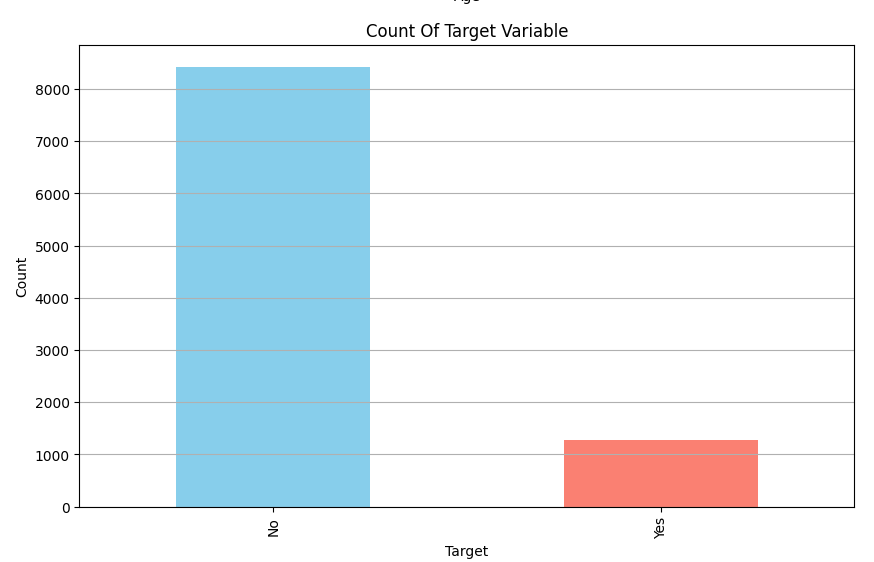


Figure 2: Count of Target Variable

**Scatter Plot Income vs. Age:**

This scatter plot shows the income according to the age form the dataset. This plot actually shows the relationship between income and age.

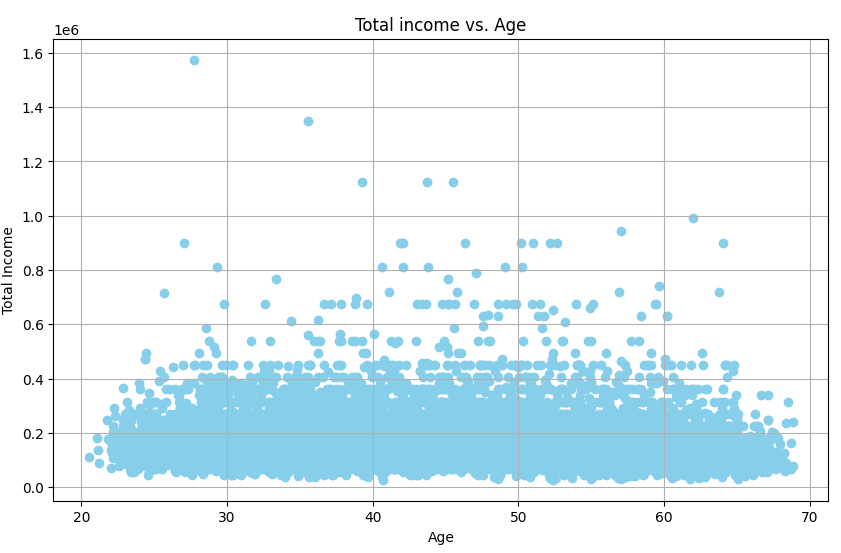


Figure 3: Income vs. Age

**Box Plot for Total income by education type:**

This box plot compares the total income across different education types. It shows the spread and median of income values for each education type.

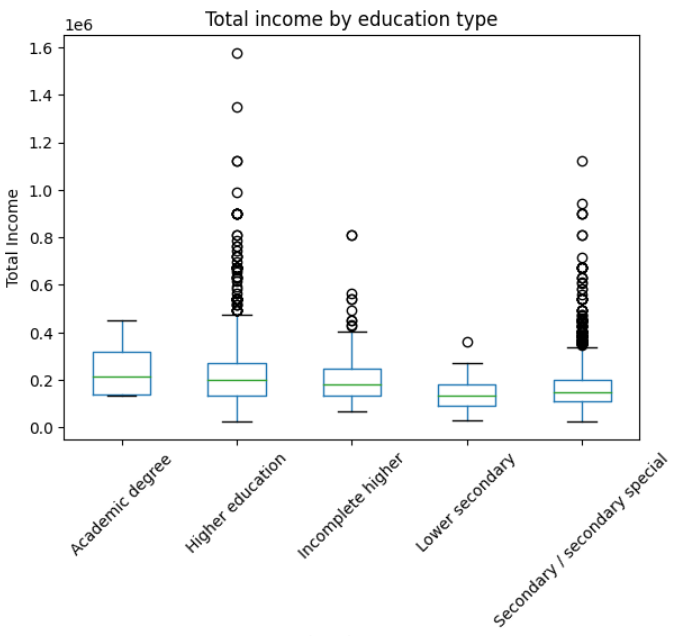


Figure 4: Box plot for total income by education

**Correlation Matrix Heatmap:**

This heatmap visualizes the correlation matrix, showing the relationship between all variables. Strong positive and negative correlations are highlighted, making it easy which variables are related.

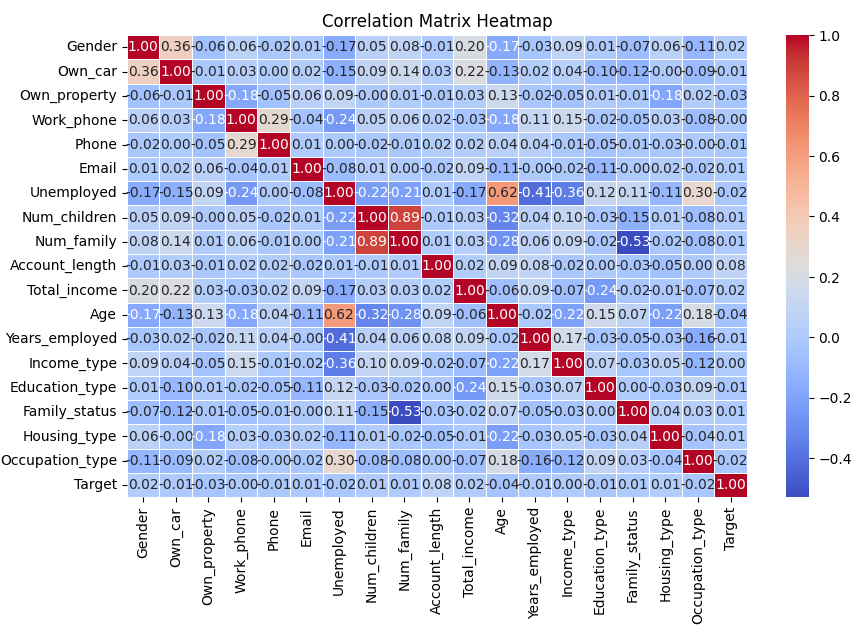


Figure 5: Heatmap of Dataset

**Elbow plot for K-Mean Clustering:**

The elbow plot helps in determining the optimal number of clusters for k-mean Clustering. The ‘elbow’ point indicates the number of clusters where the sum of squared errors starts to diminish at a slower rate.

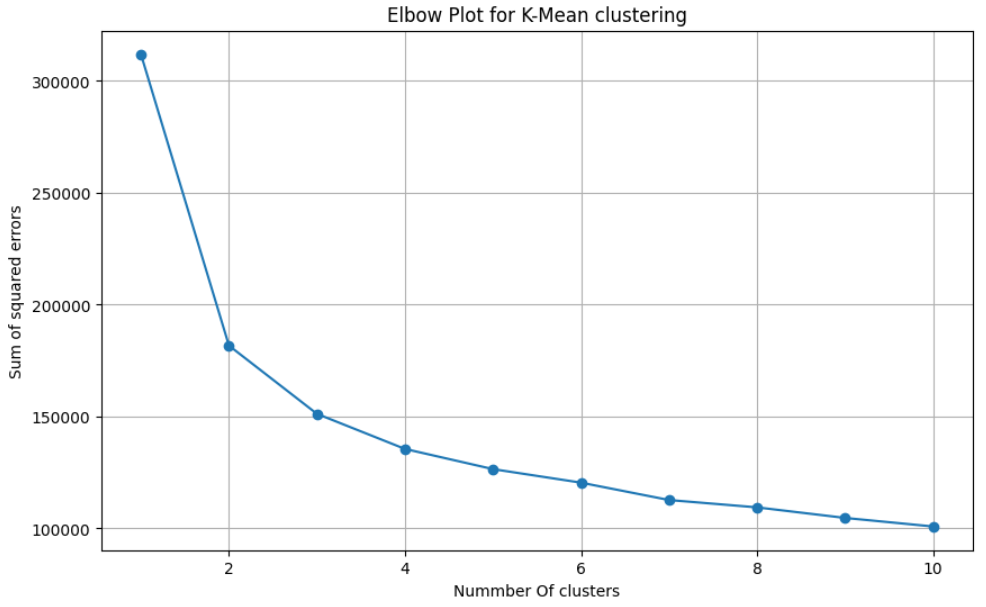


Figure 6: Elbow Plot for K Mean Clustering

**Line fitting:**

This plot shows linear fit for total income vs. age. The red line represents the best-fit line, indicating the relationship between age and total income.

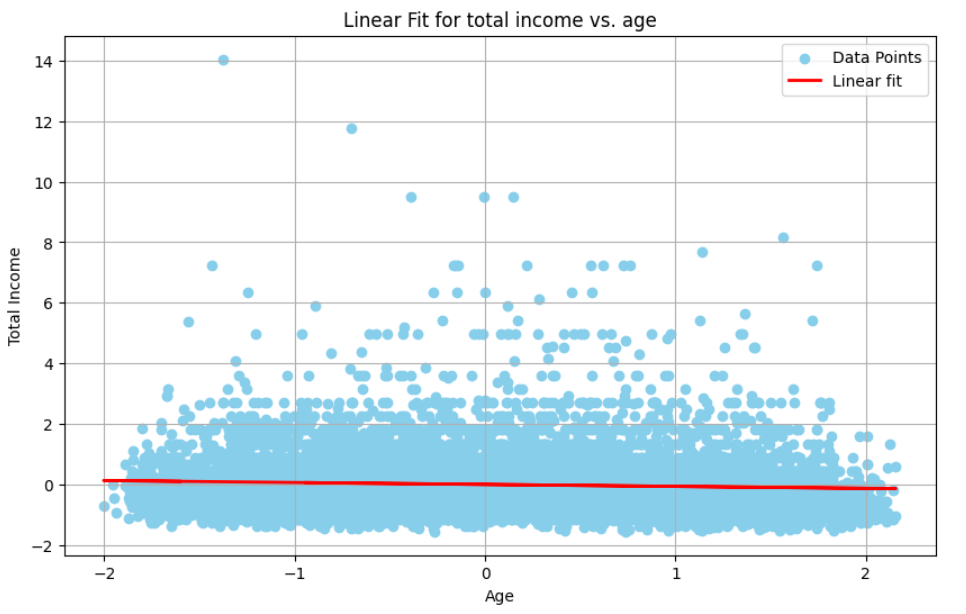


Figure 7:Line fitting for total income vs. age

**Conclusion:**

The research fits a line to look at the link between age and total income, and it offers insights into the financial and demographic aspects of individuals. Understanding the underlying trends in the data is aided by the visualisations and clustering algorithms.

Additional grouping strategies, more intricate regression models, and a greater understanding of the features influencing the target variable could all be investigated further.