PREDICTING RISK LEVEL AND RETURN EARNED

GR Sangavi

21PW28

INTRODUCTION

Project Topic

- The project focuses on predicting risk level and return earned based on demographic and investment-related factors.

Dataset

- The analysis is based on a dataset containing information on investors' age, income, education level, investment knowledge, and more.

Goal

- The goal of the analysis is to develop a predictive model that can accurately estimate the risk level and potential return of an investment based on the given factors.

OVERVIEW

Dataset Summary

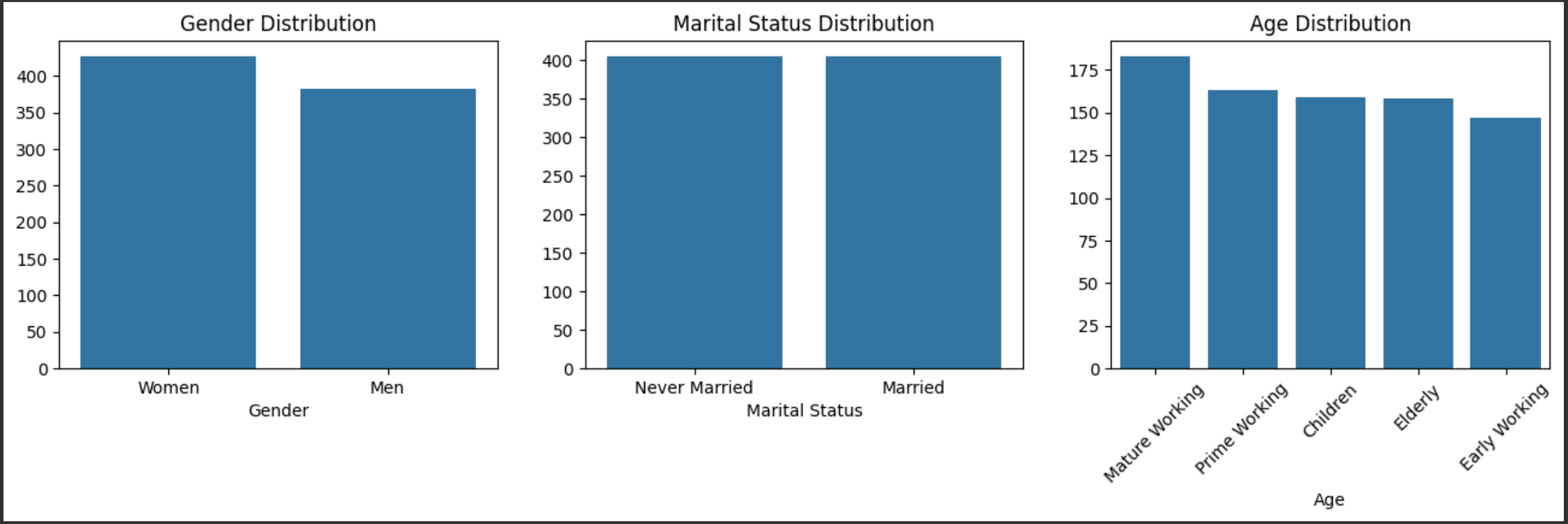
- The dataset contains a total of 810 records and 19 columns.
- The data includes demographic information, such as age, gender, and income, as well as investment-related factors, such as risk level and return earned.

Key Variables

- Age: The age of the individual.
- Gender: The gender of the individual.
- Income: The income level of the individual.
- Risk Level: The risk level associated with the investment.
- Return Earned: The return earned from the investment.

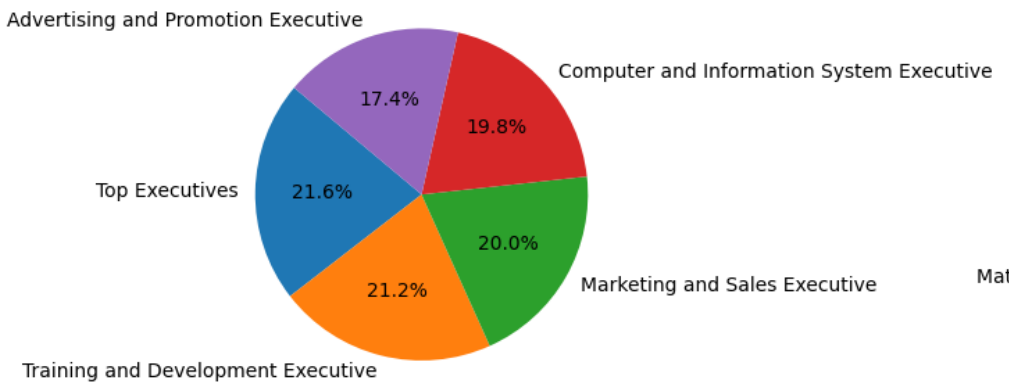
EXPLORATORY DATA ANALYSIS (EDA)

DEMOGRAPHIC DISTRIBUTIONS

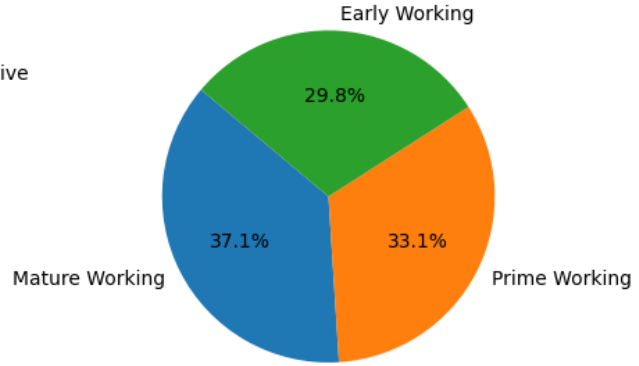


EMPLOYMENT DETAILS

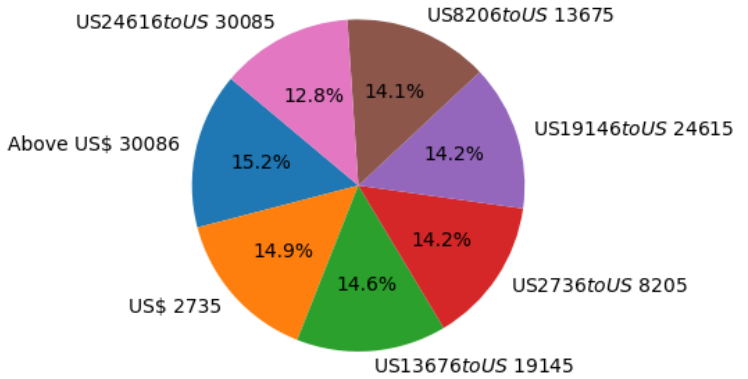
Distribution of Roles



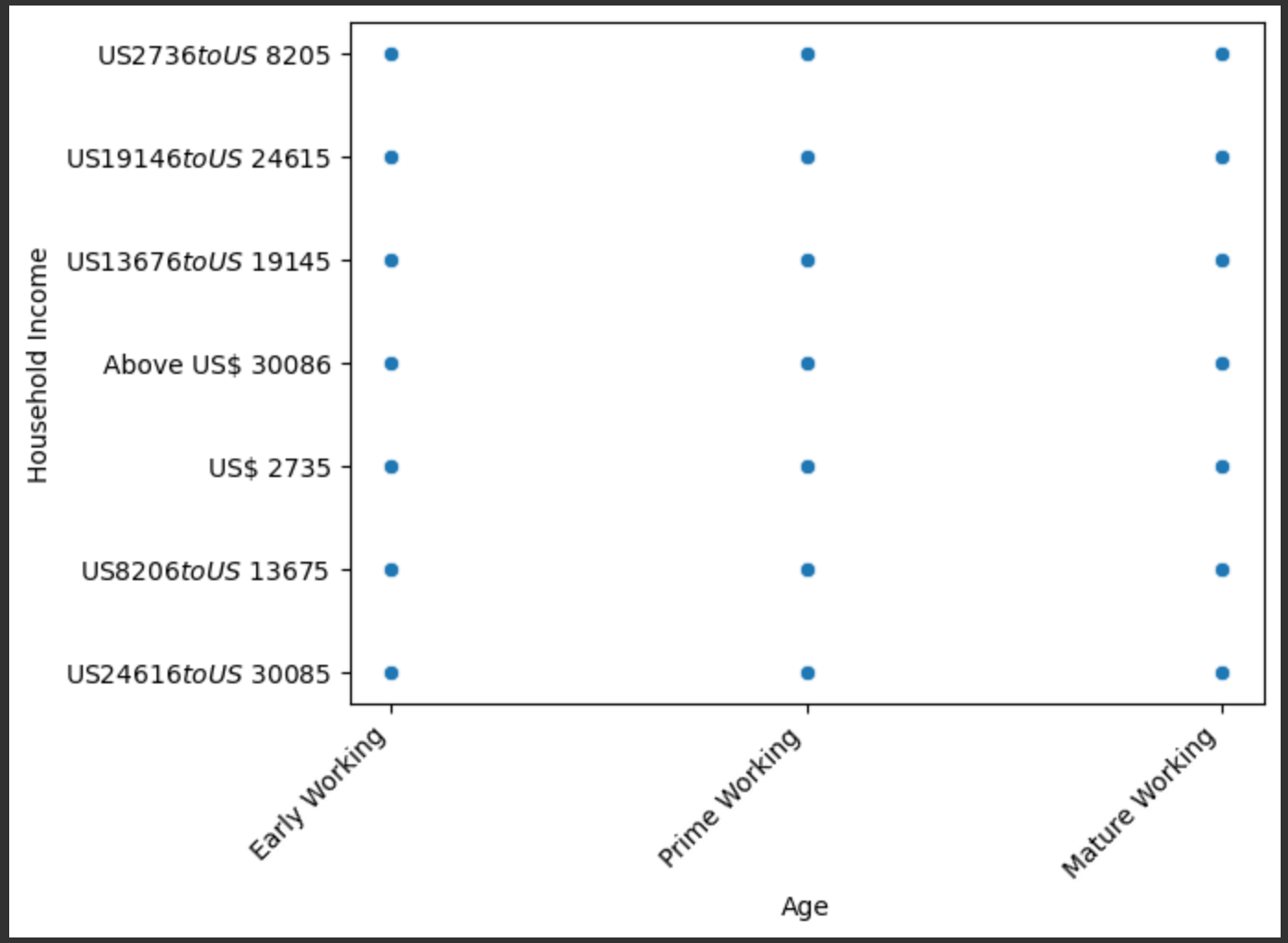
Distribution of Career Stages



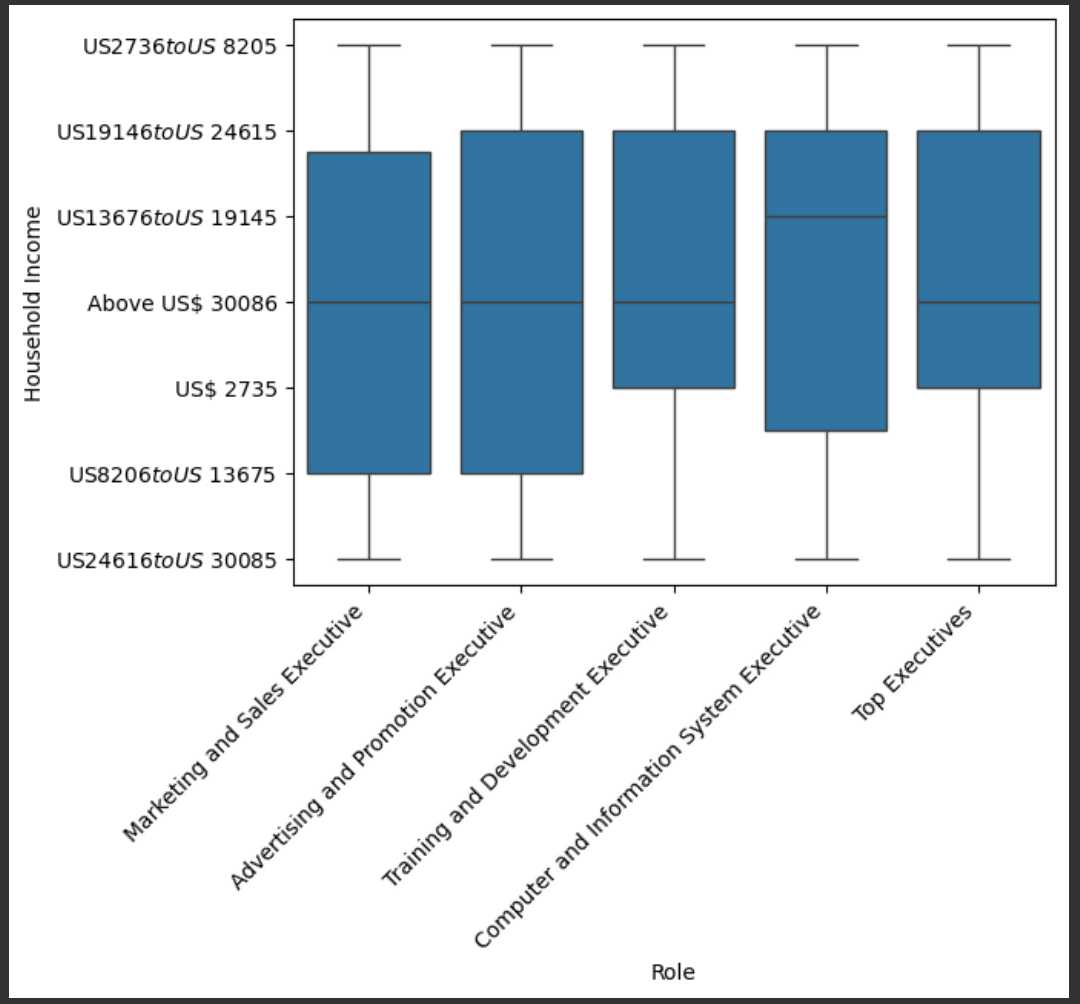
Distribution of Household Income Brackets



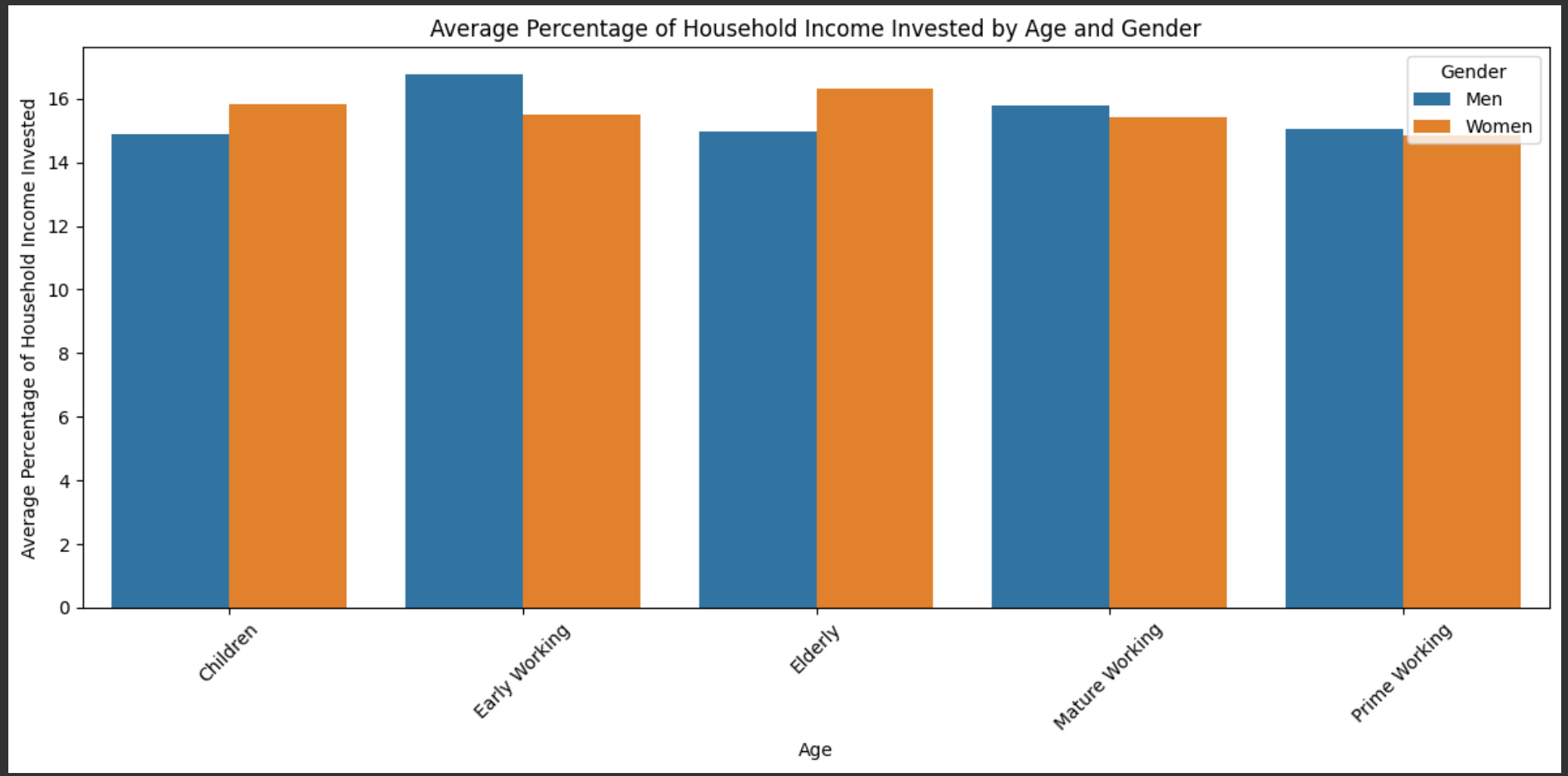
INCOME VS AGE



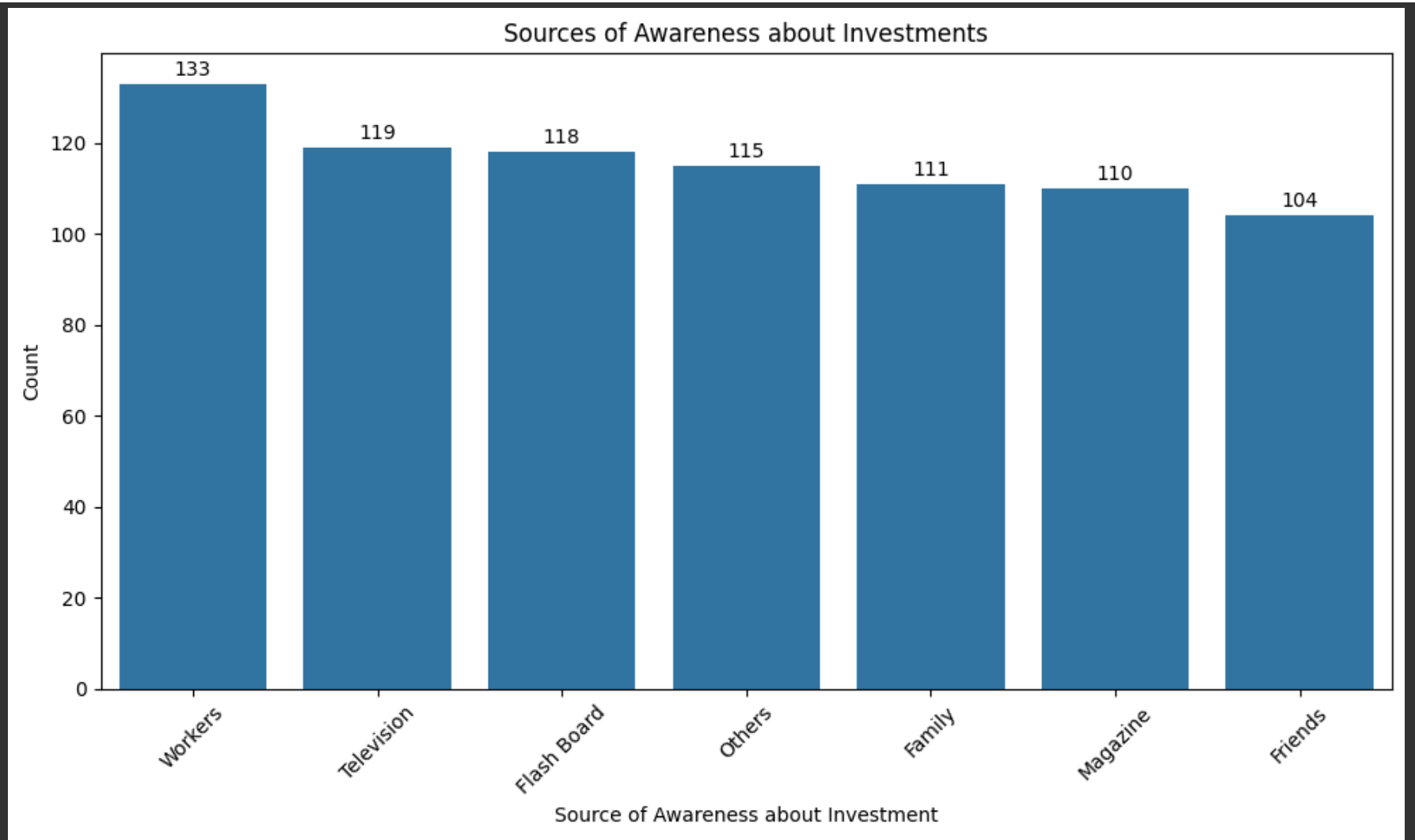
INCOME VS ROLE



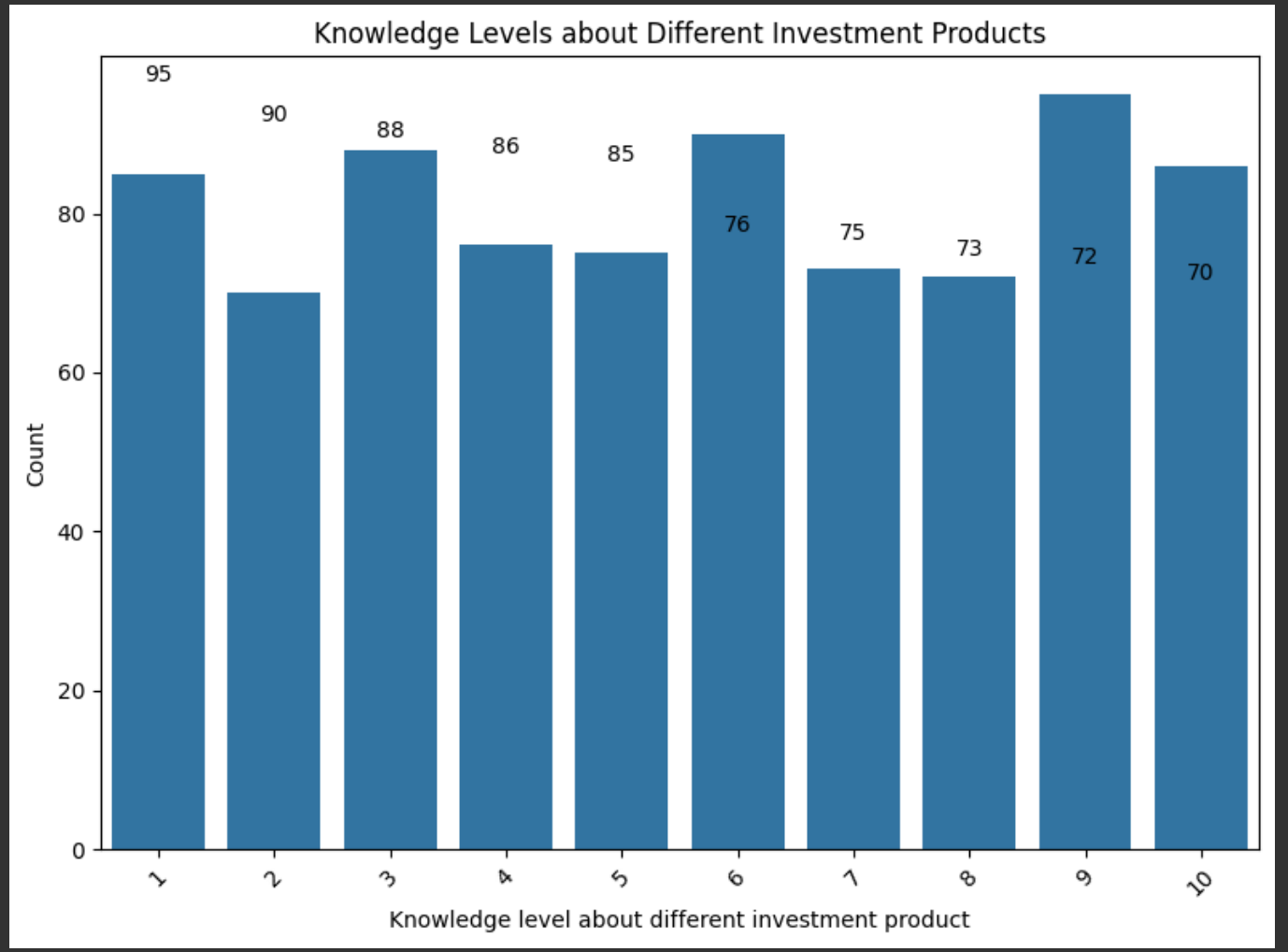
AVERAGE PERCENTAGE OF HOUSEHOLD INCOME INVESTED BY AGE AND GENDER



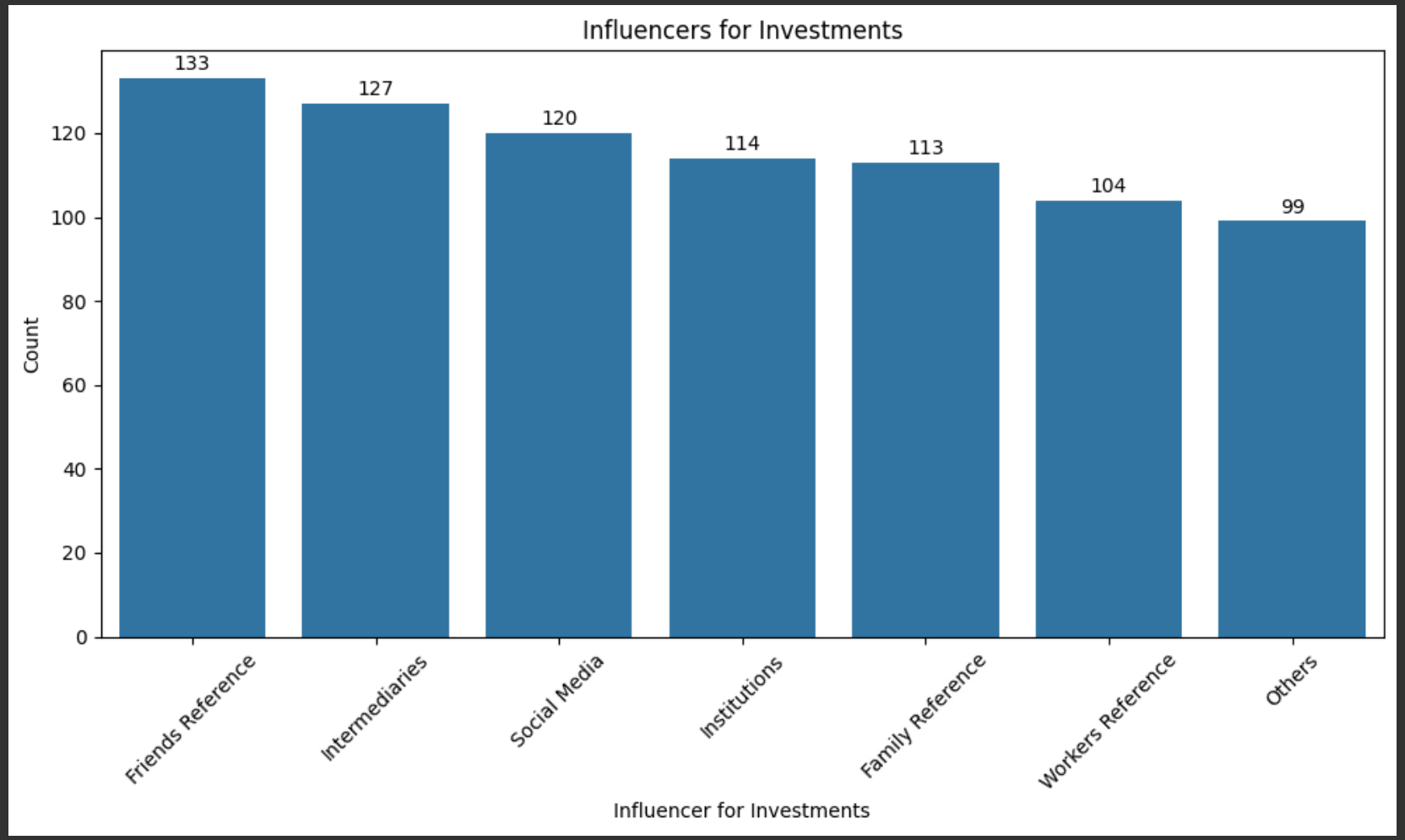
SOURCES OF AWARENESS ABOUT INVESTMENTS



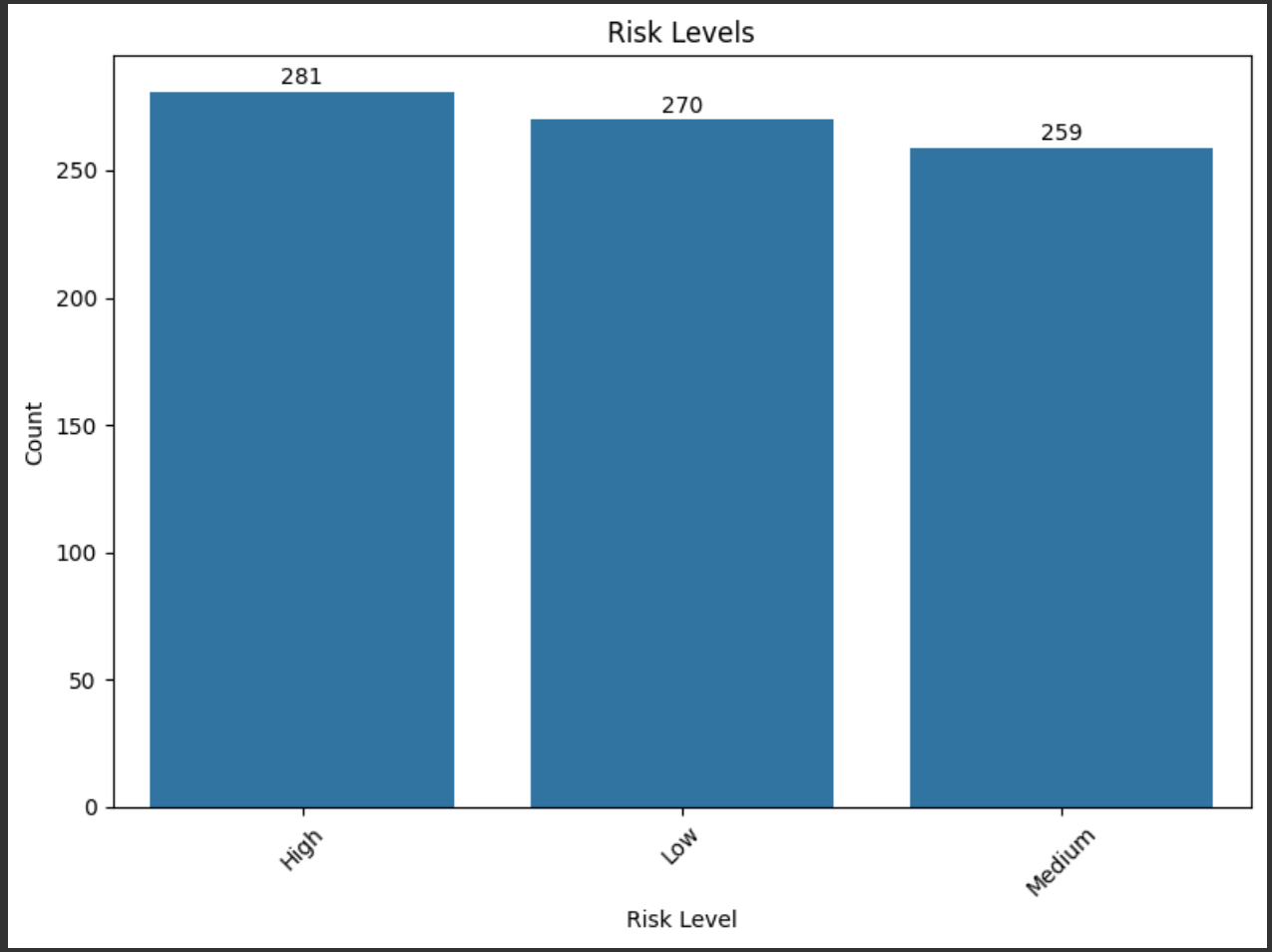
KNOWLEDGE LEVELS ABOUT DIFFERENT INVESTMENT PRODUCTS



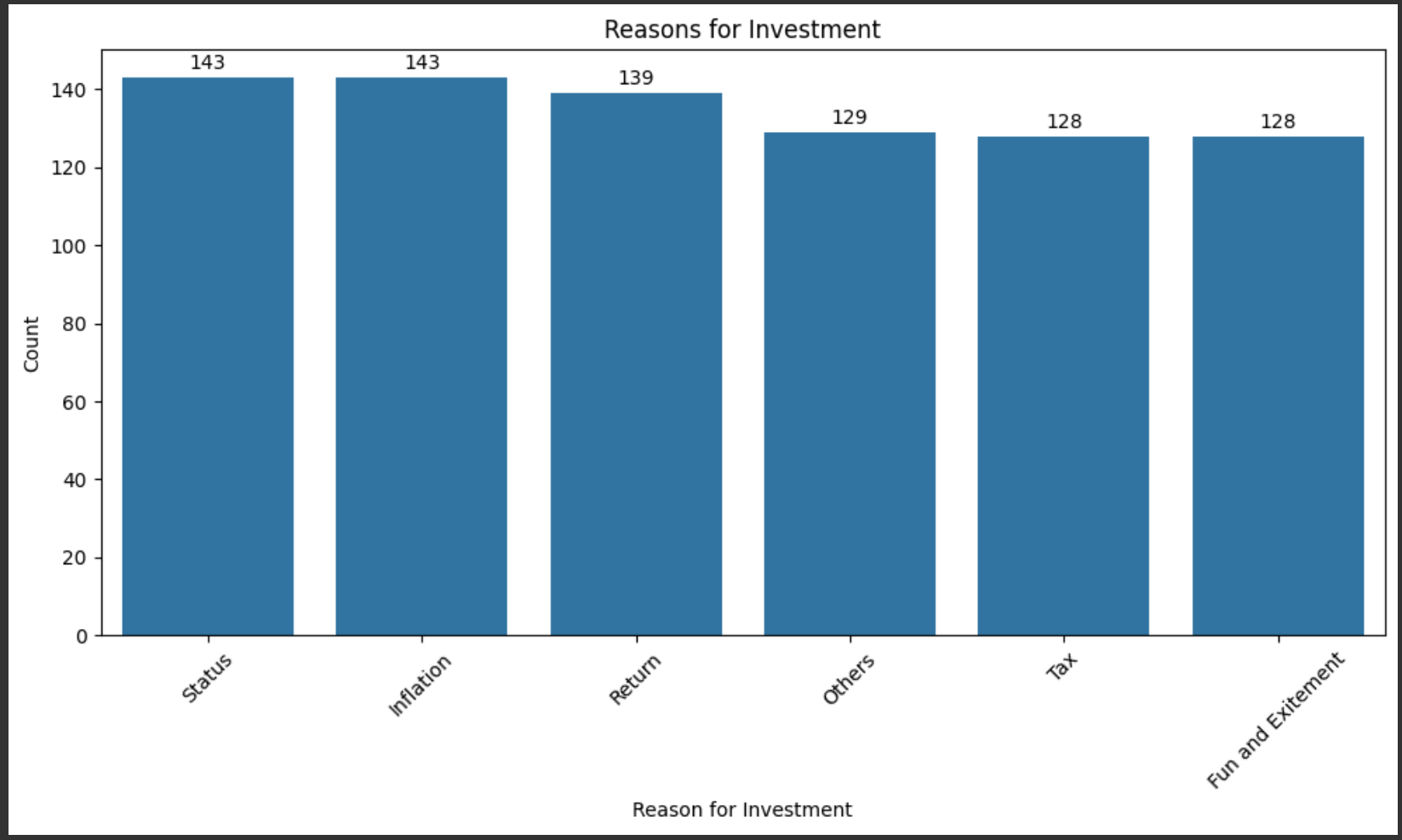
INFLUENCERS FOR INVESTMENTS



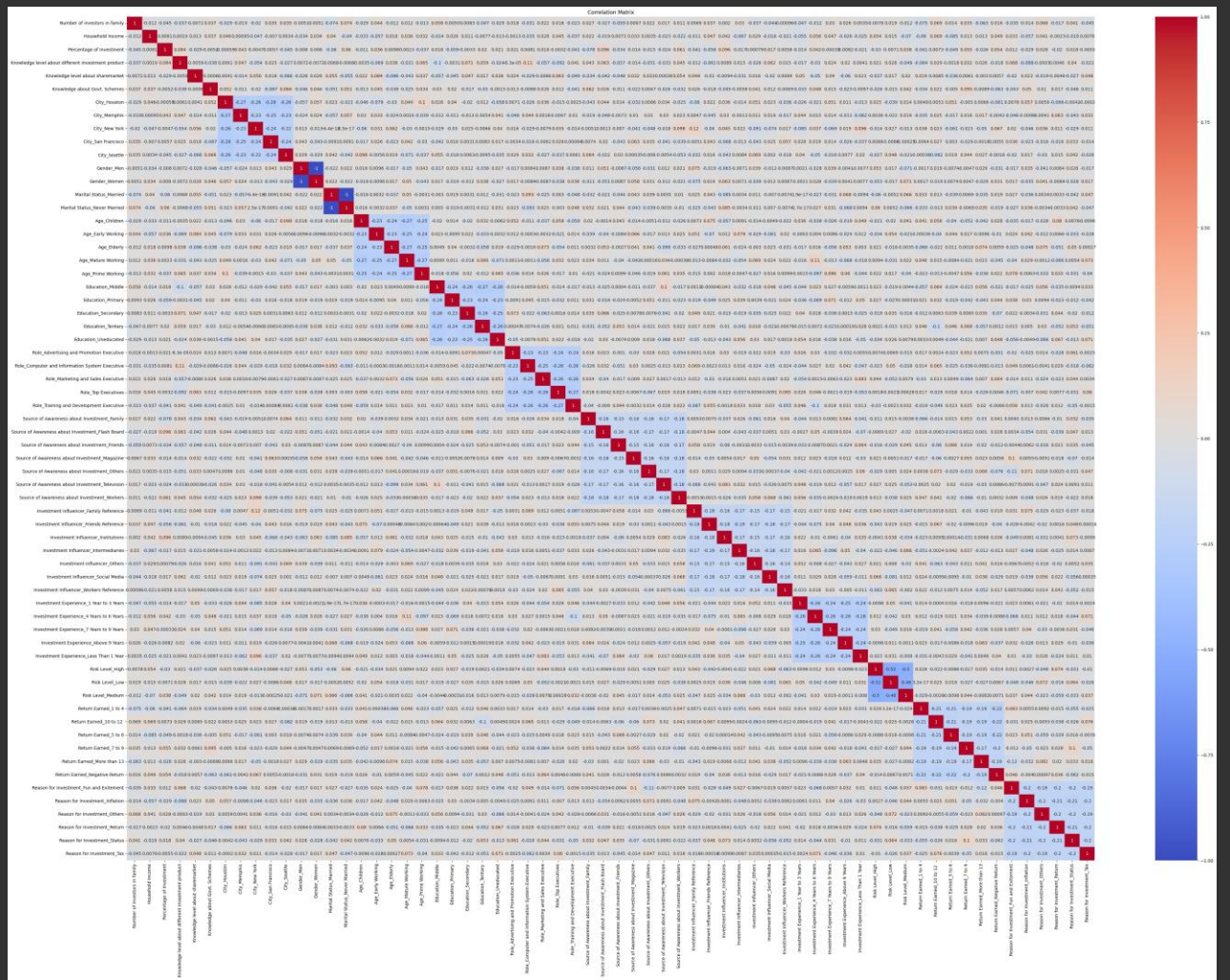
RISK LEVELS


Risk Levels
281
270
259
250
200
150
100
50
0
Count
High
Low
Medium
Risk Level

REASONS FOR INVESTMENT



CORRELATION MATRIX



DATA PREPROCESSING

1. **Handling Missing Values:**

- Identified missing values in the dataset and decided on an appropriate strategy to handle them.
- Used techniques such as mean imputation, mode imputation, or deletion of rows/columns with missing values.

2. **Encoding Categorical Variables:**

- Converted categorical variables into numerical representations that can be used by machine learning algorithms.
- Used techniques such as one-hot encoding, label encoding, or ordinal encoding.

3. **Scaling Numerical Variables:**

Numerical variables are scaled using StandardScaler to bring them to the same scale.

DATA PREPROCESSING

In our dataset, the 'Household Income' column originally contained values that represented income bracket ranges, such as 'Above 50,000 US$', 'Upto 30,000 US$', and specific ranges like '30,000 to 50,000 US$'. To convert these values to float, we implemented a custom function that extracted the numerical values from these strings. For example, 'Above 50,000 US$' would be converted to 50000.0, 'Upto 30,000 US$' to 30000.0, and '30,000 to 50,000 US$' to the average of the two values, which is 40000.0. This conversion allows us to work with these values as numerical data in our analysis.

Similarly, the 'Percentage of Investment' column contained values such as '10%', '20%', 'Upto 30%', and 'Above 50%'. We applied a custom function to extract the numerical percentage values from these strings. For example, '10%' would be converted to 10.0, '20%' to 20.0, 'Upto 30%' to 30.0, and 'Above 50%' to 50.0. This conversion enables us to use these values for calculations and analysis in our model.

FEATURE SELECTION

Random Forest

- Random Forest is a machine learning algorithm that can be used for feature selection. It works by creating an ensemble of decision trees and then aggregating their predictions.

Top Features

After applying Random Forest for feature selection, the top features selected were:

'Knowledge level about sharemarket', 'Knowledge level about different investment product', 'Percentage of Investment', 'Household Income', 'Knowledge about Govt. Schemes', 'Number of investors in family', 'Gender_Women', 'Marital Status_Never Married', 'Age_Mature Working', 'Education_Secondary'

MODEL TRAINING AND EVALUATION

Train-Test Split

- The dataset was split into a training set and a testing set. The training set was used to train the Random Forest model, while the testing set was used to evaluate the model's performance.

Model Performance Metrics

- The performance of the Random Forest model was evaluated using various metrics, including accuracy, precision, recall, and F1 score.

Classification Reports

- Classification reports were generated to provide a detailed analysis of the model's performance for each class. These reports include metrics such as precision, recall, and F1 score for each class.

MODEL TRAINING AND EVALUATION

```
Overall Accuracy - 0.2500

Classification Report for Risk Level:
              precision    recall  f1-score   support

           0       0.27      0.32      0.29        53
           1       0.32      0.29      0.30        55
           2       0.43      0.39      0.41        54

    accuracy                           0.33       162
   macro avg       0.34      0.33      0.34       162
weighted avg       0.34      0.33      0.34       162


Classification Report for Return Earned:
              precision    recall  f1-score   support

           0       0.25      0.23      0.24        31
           1       0.16      0.17      0.17        29
           2       0.16      0.13      0.15        30
           3       0.07      0.08      0.08        26
           4       0.14      0.16      0.15        19
           5       0.21      0.22      0.21        27

    accuracy                           0.17       162
   macro avg       0.16      0.16      0.16       162
weighted avg       0.17      0.17      0.17       162
```

RECOMMENDATION SYSTEM

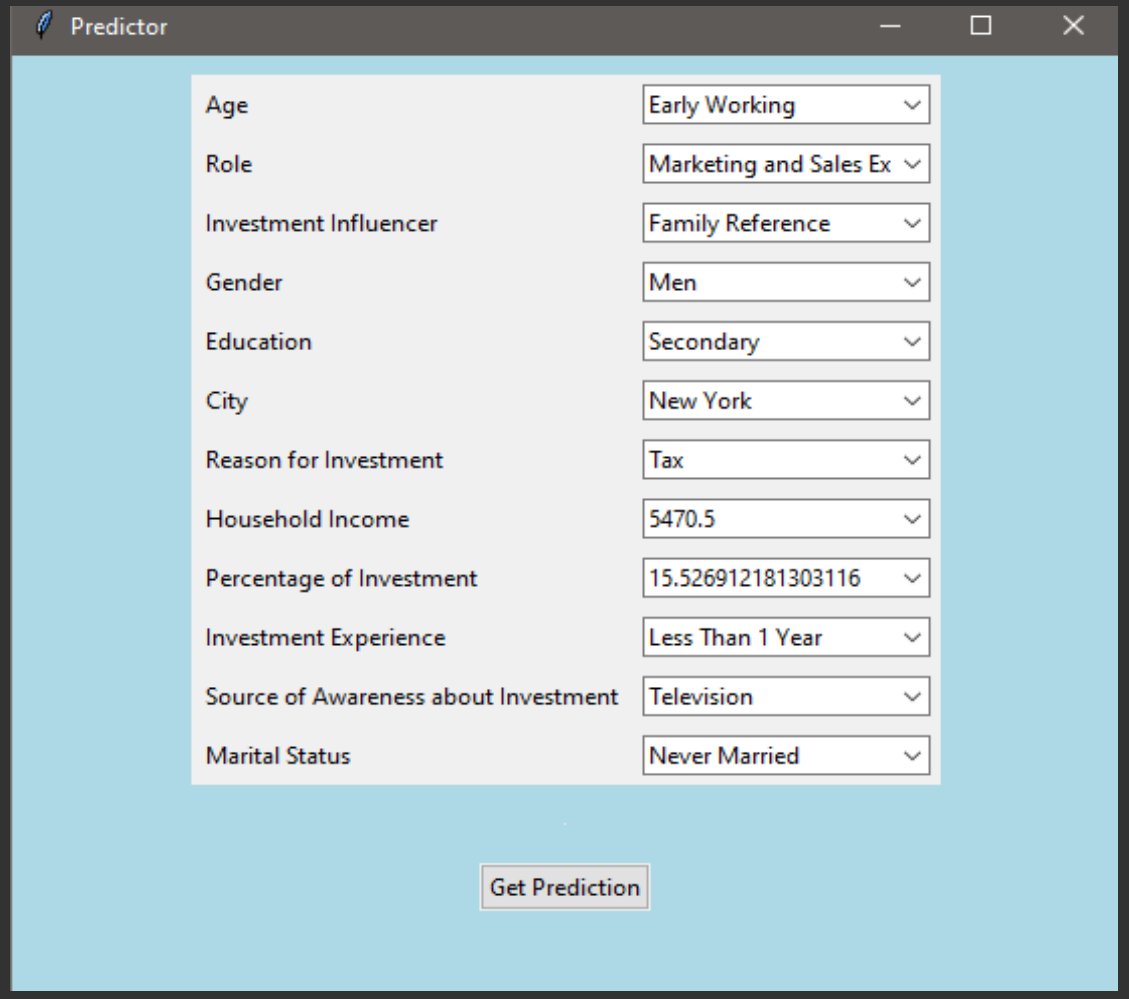
Input Features

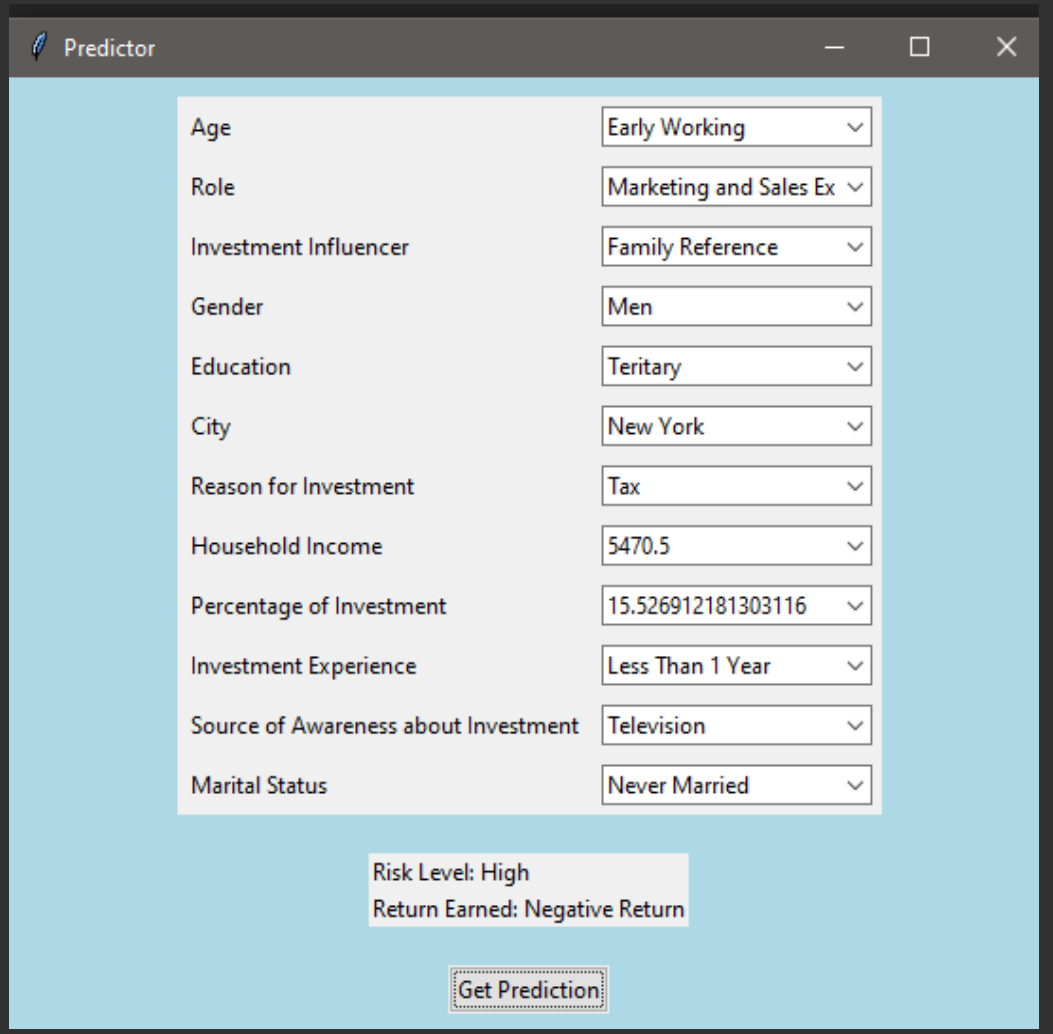
The recommendation system takes into account various demographic and investment-related factors to generate personalized recommendations for users.

Prediction Process

Based on the input features, the recommendation system uses machine learning algorithm to analyze historical data and make predictions on the risk level and potential return of different investment options.

RECOMMENDATION SYSTEM





CONCLUSION

- Understanding the demographic factors that influence investment risk can help investors make more informed decisions.
- The identified investment-related factors can guide investors in selecting investment options that align with their risk tolerance and financial goals.