**Buckman Hackathon Documentation**

21PD03- Benny Abhishek M

Note: 1)to run the program run: **streamlit run app.py** from the directory containing app.py file.

2) All plots are done using Plotly and saved as image and displayed here for documentation pupose. The plots in the streamlit app run in realtime.

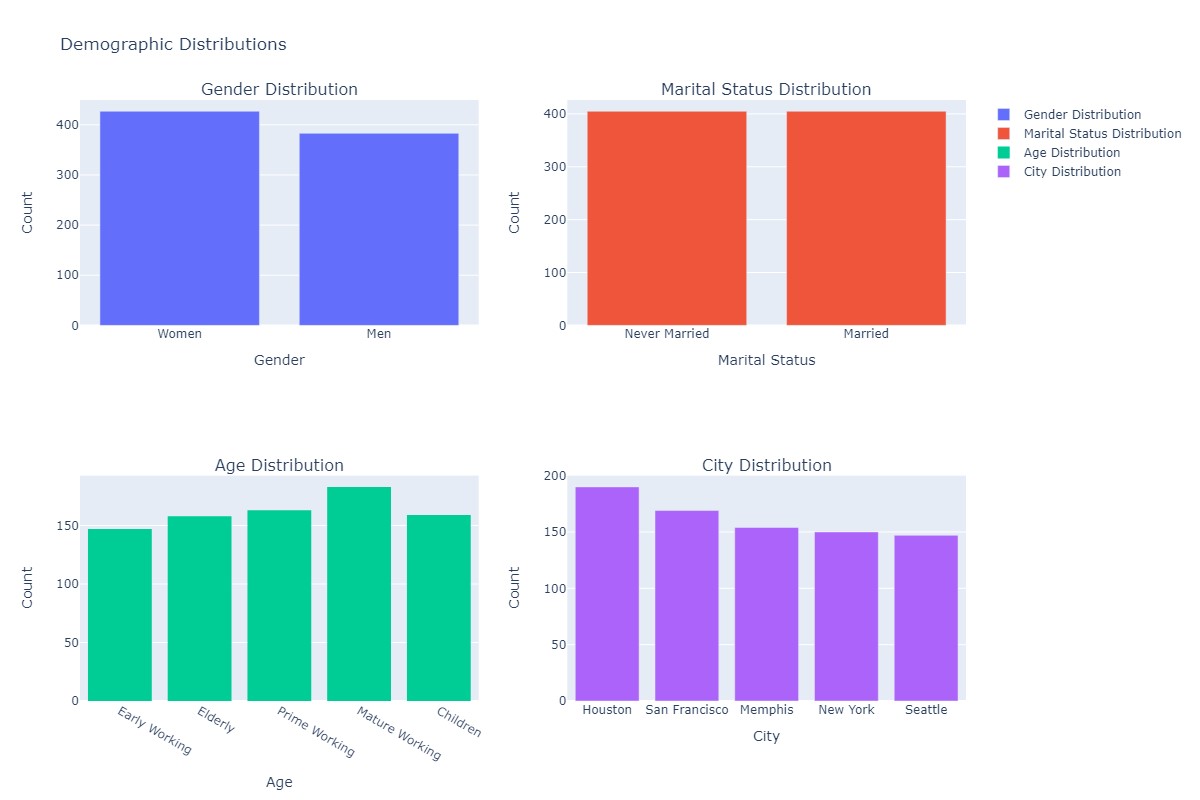
3) The ppt contains all output screenshots of the app.

Libraries required – Pandas, streamlit, sklearn, plotly,scipy

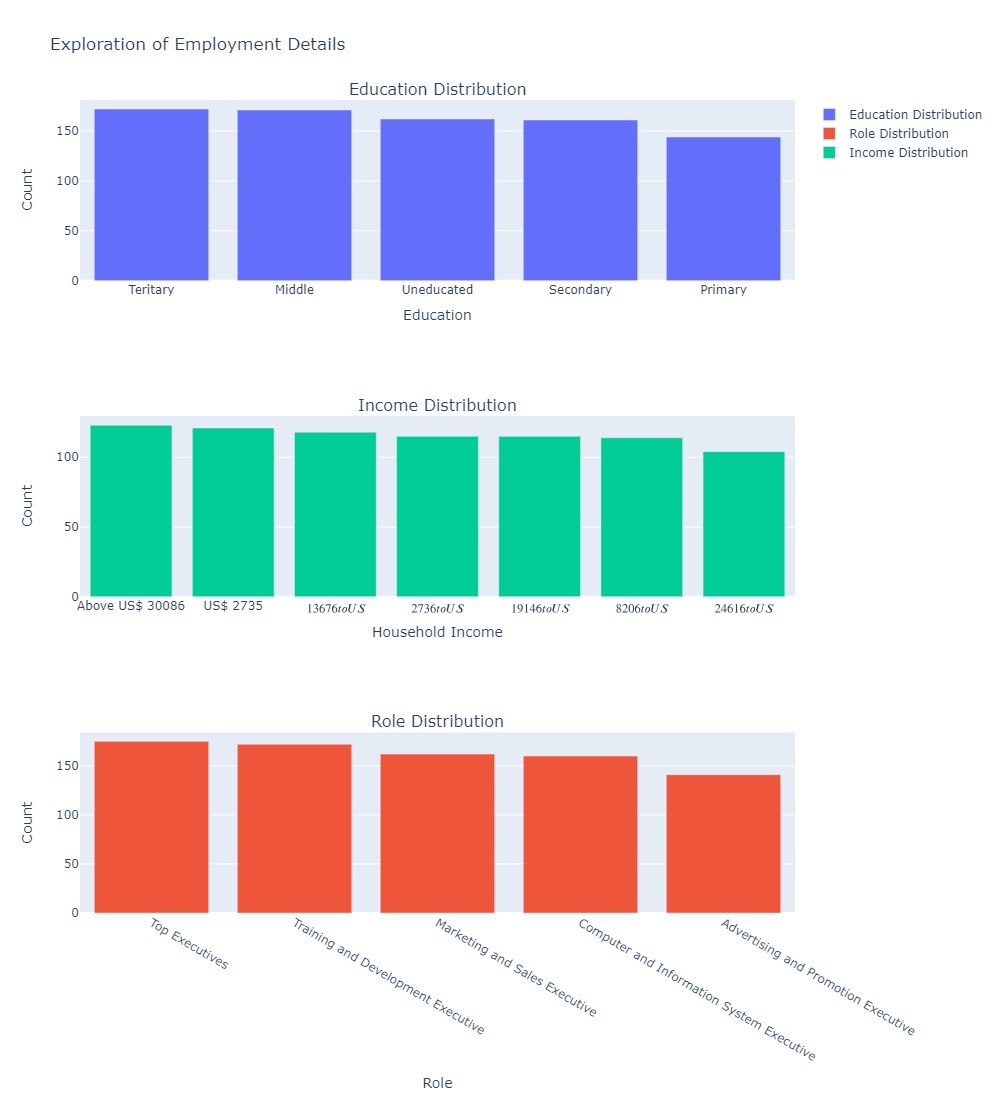
Data Preprocessing:

1. The data is pre-processed by finding null values and outliers if any.
2. The “Return Earned” column of the dataset is converted to a numeric score range from 0-4 and -1 for negative returns. This makes it easier for the upcoming tasks.

Data Exploration:

1. Plots for analysing demographic distribution are done for Gender, Marital status, Age, City.

2)Plots for employment details are done using Plotly.

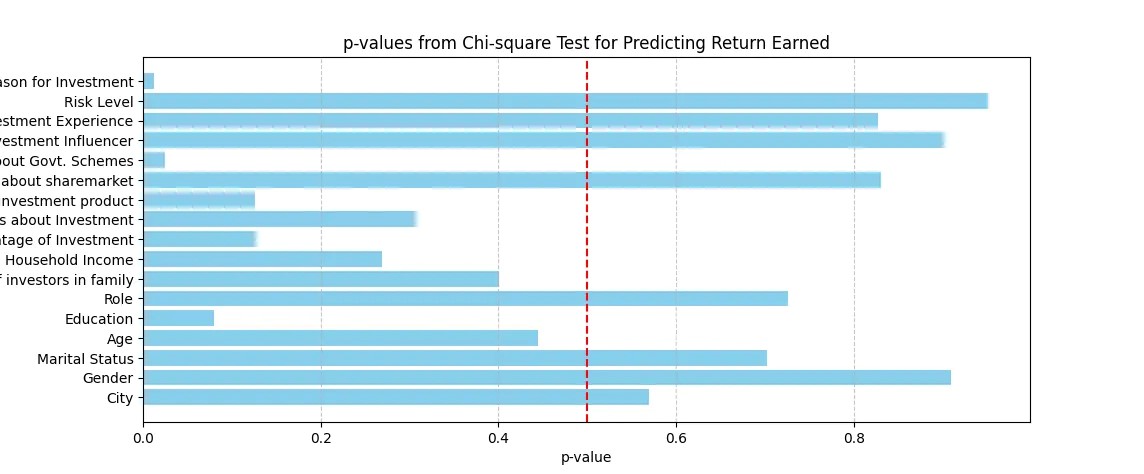
3) Plots for investment behaviour insights including the percentage of household income invested, sources of awareness about investments, knowledge levels, influencers, risk levels, and reasons for investment are done.

4) Also other simple EDA has been done.

Best Investment Decision Identification

1. Since all the columns are categorical, The

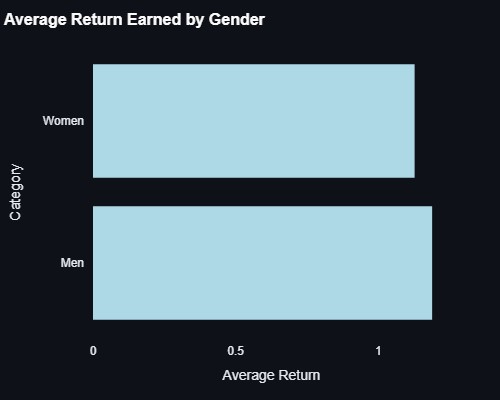
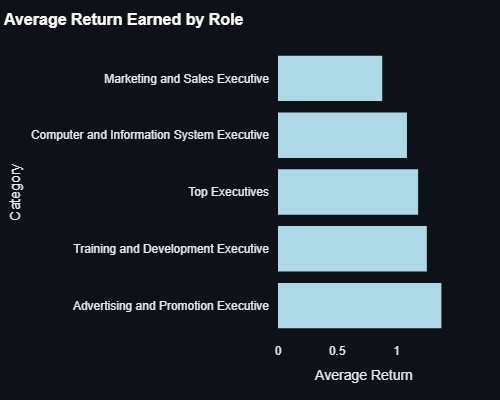
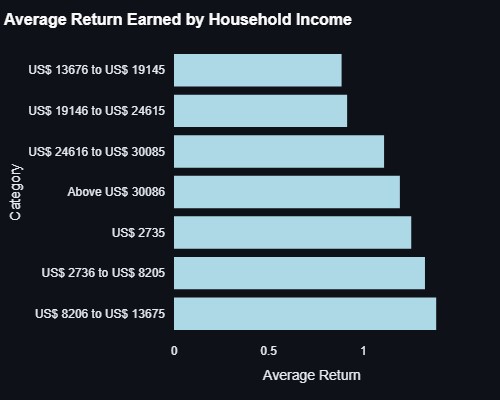
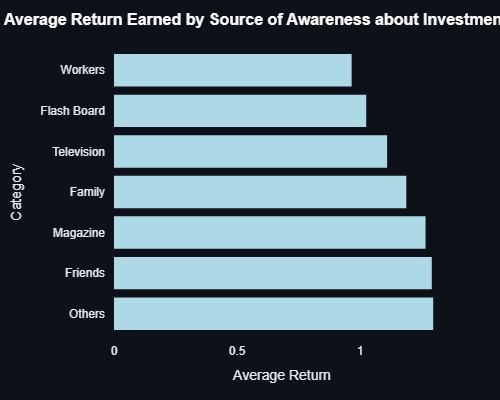
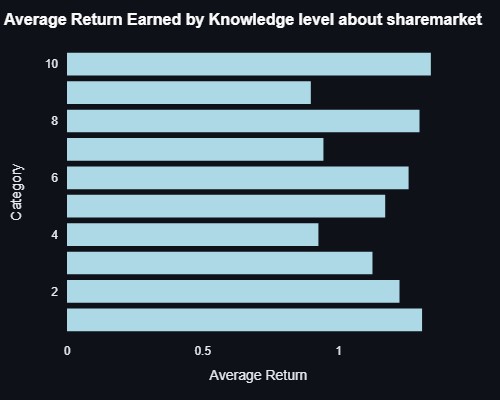
**Chi-Square test** is used to find the factor variables contributing to the best investment results. Here we have taken the significance level as 0.5



Then the demographic, employment, and behavioral characteristics which correlate successful investment outcomes is found by calculating the average returns for each categorical value under every column.

(The average can be found because the Return Earned column is transformed to a Numerical Score)

1. These results are also plotted for easy visualization



Model Selection & Evaluation

1)Different models like DecisionTree, Random Forest, Multi-layer perceptron(Neural network) and Support Vector Machine was applied to the categorical data and the performance was measure using performance metrics like Accuracy, precision, recall and F1-Score and it was found to give a reasonable performance even for a small dataset.(Output screenshot in PPT)

2)Finally after tuning and validation, the best model is chosen to predict the returns based on inputs from the user for all the input features.