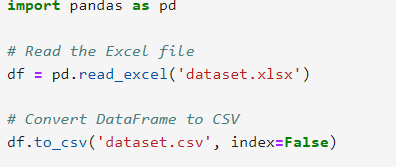
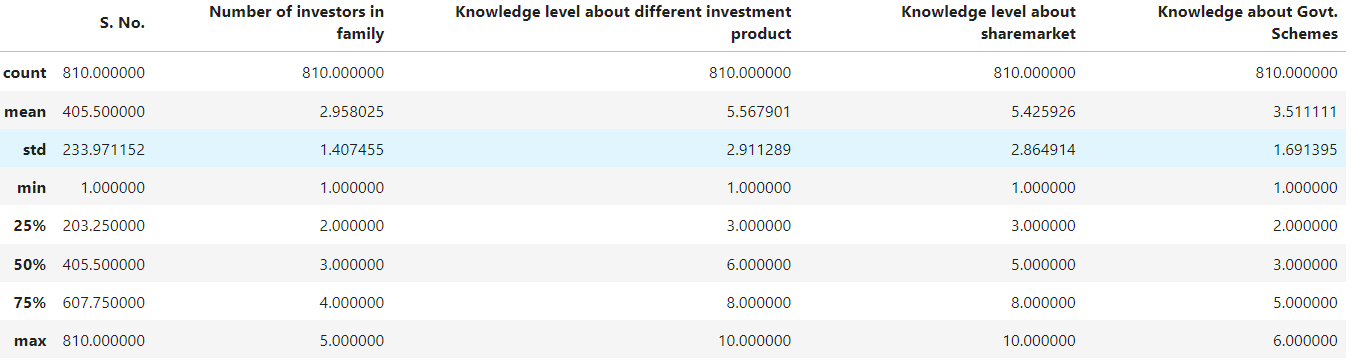
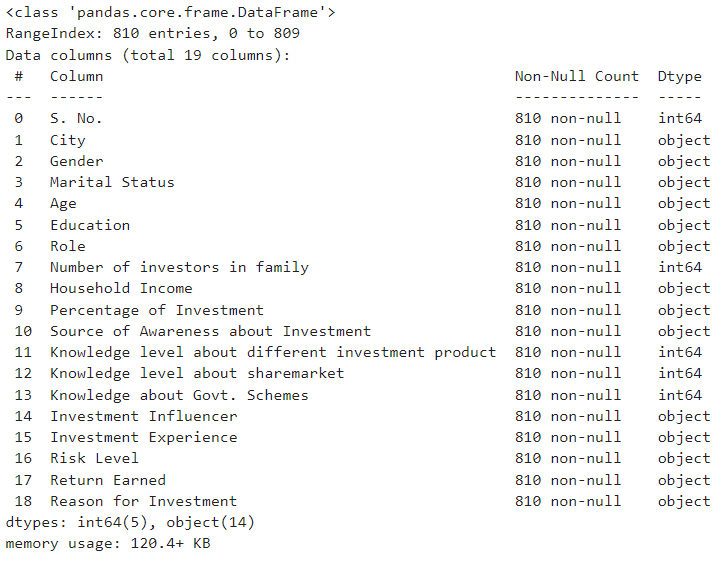
REPORT

Data Exploration: Handling the dataset

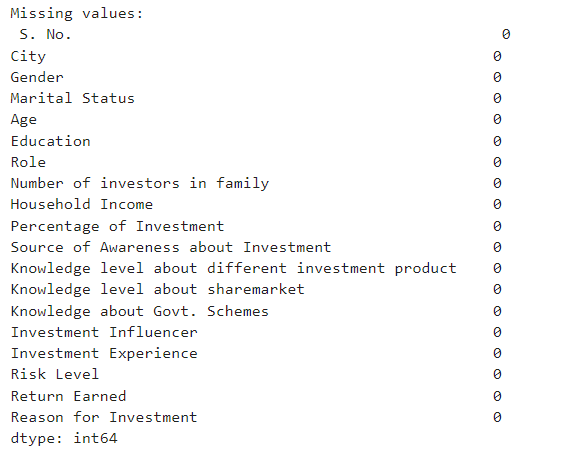
The data is loaded into a pandas Data Frame from an Excel file and converted to a CSV format.



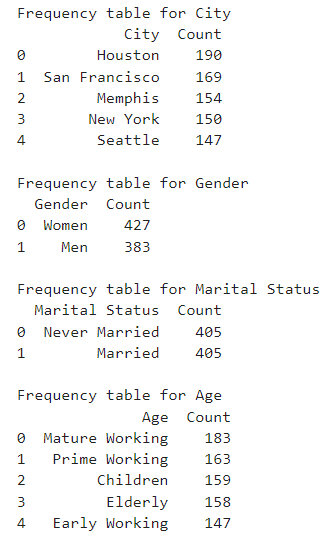
The Data Frame is inspected using the. head(), .info(), and .describe() methods to understand the structure and statistical summary of the dataset, which includes 810 entries across 19 columns.

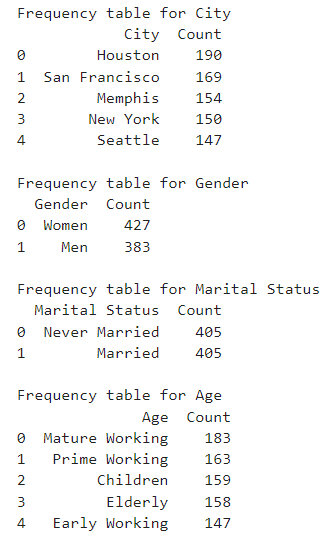


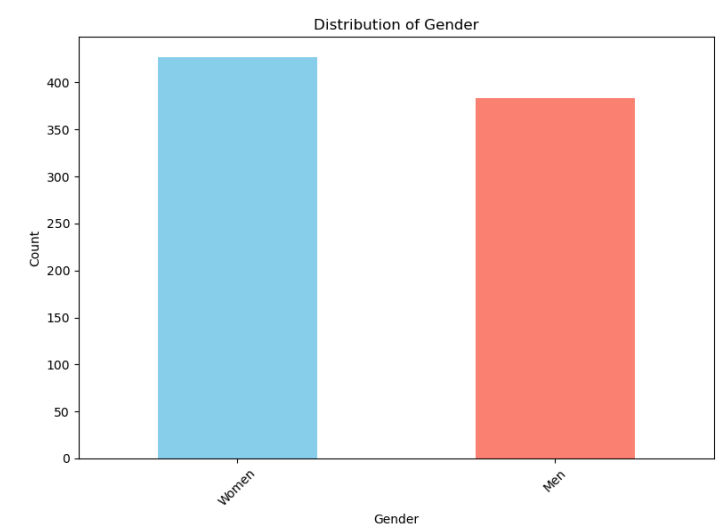
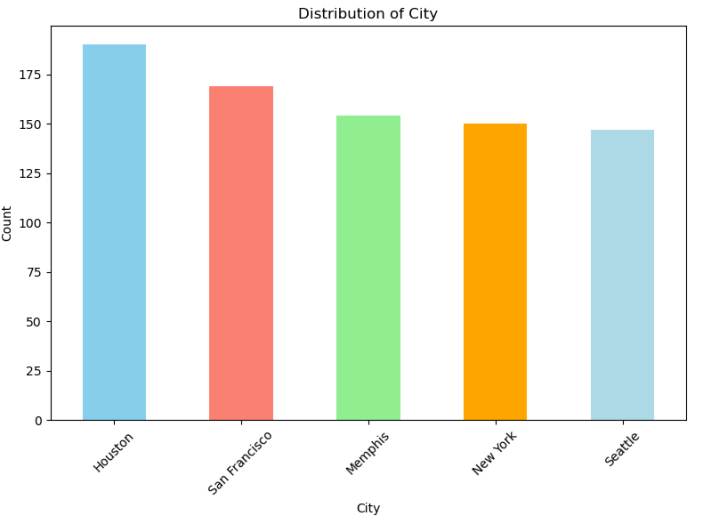
No missing values are reported in the dataset.



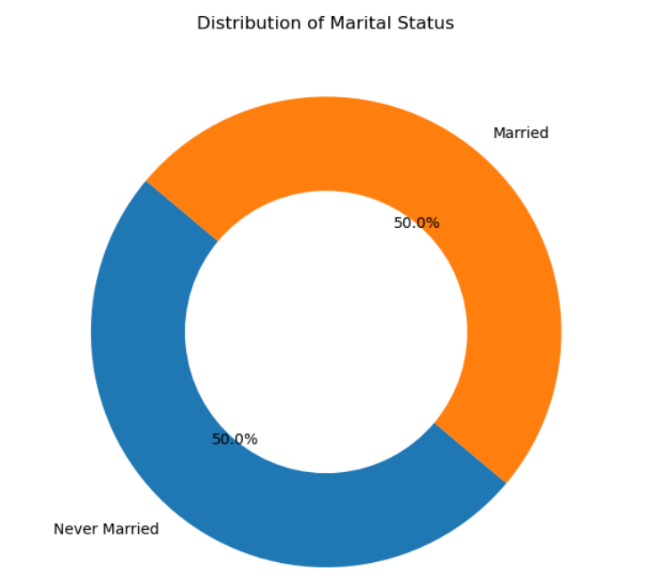
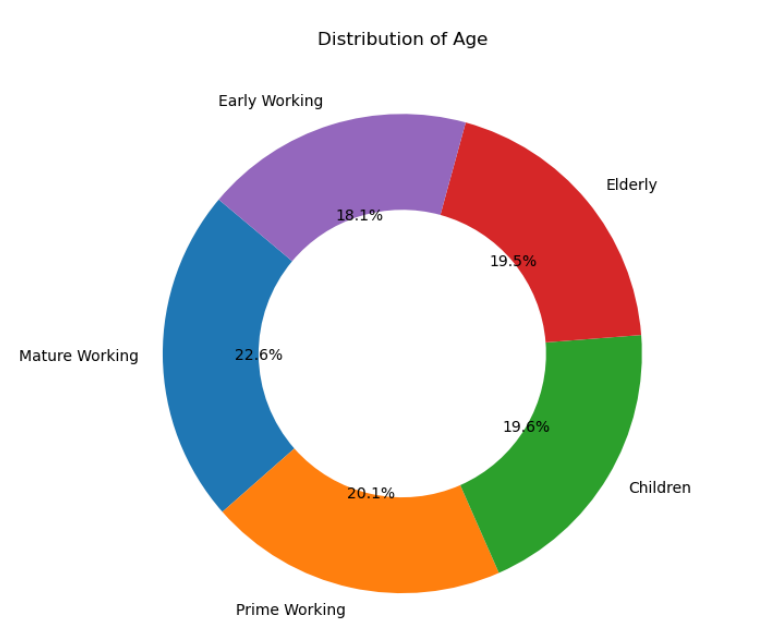
Data ANALYSIS: **Analyze the demographic distribution**

The demographic distribution in the dataset is analyzed through various methods, primarily focusing on categorical variables such as City, Gender, Marital Status, and Age. The analysis is conducted using two types of visualizations: bar graphs and donut charts



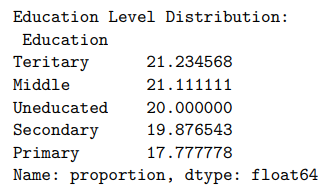
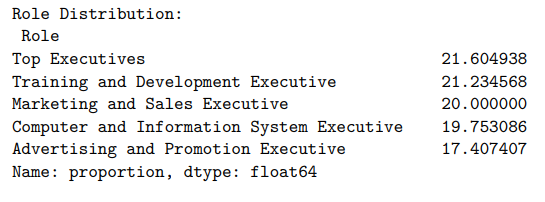
**plot\_demographic\_distribution: Used To plot Bar Graph**

**plot\_demographic\_distribution\_donut: Used To plot Donut Chart**

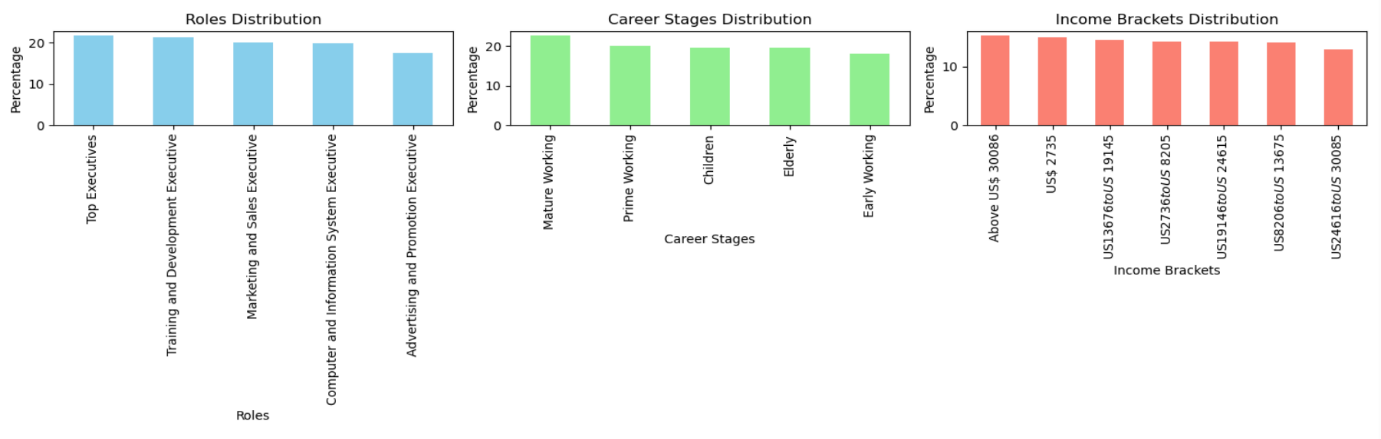


Data ANALYSIS: **EXPLORE the** employment details

 Employment details are explored by examining role distribution and education level distributions, as well as income brackets. Visualizations are also created for these attributes.

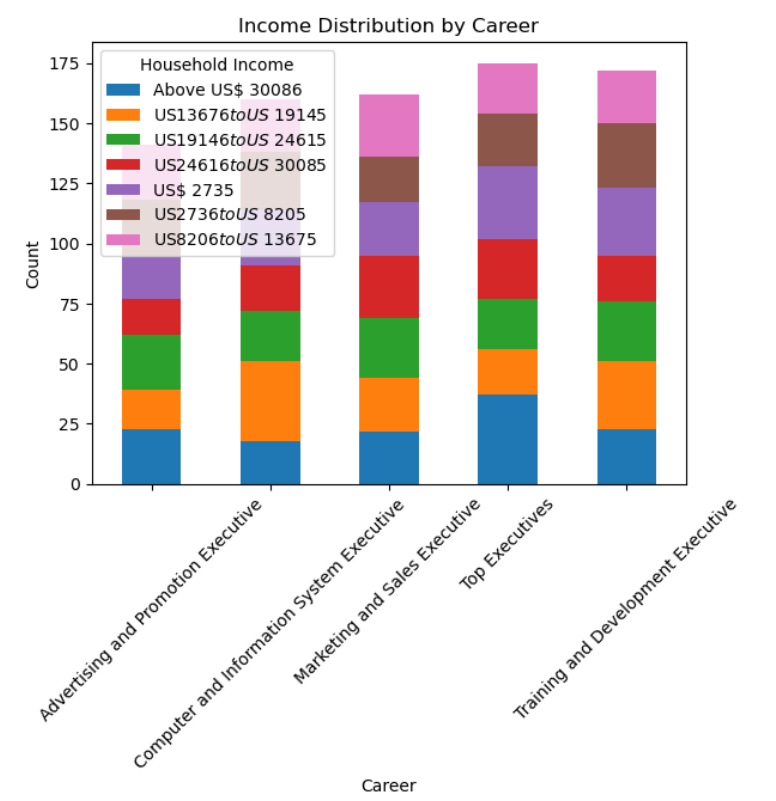


Bar Chart Plotted For Roles, Career, And Income



The dataset includes a variety of executive roles, with 'Top Executives', 'Training and Development Executive', 'Marketing and Sales Executive', 'Computer and Information System Executive', and 'Advertising and Promotion Executive' being some of the roles mentioned.

The proportion of each role is calculated, showing that 'Top Executives' make up approximately 21.60% of the dataset, 'Training and Development Executive' about 21.23%, 'Marketing and Sales Executive' around 20.00%, and so on



The 'Percentage of Investment' variable is analyzed about Household Income. This analysis reveals how much of their income individuals in different income brackets are investing.

Some individuals choose not to reveal their investment percentage, while others invest various proportions of their income, ranging from Up to 5% to Above 26%.

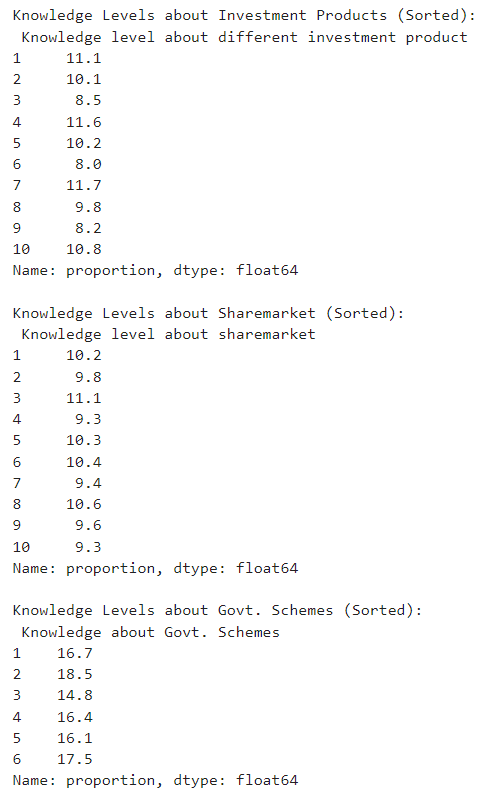
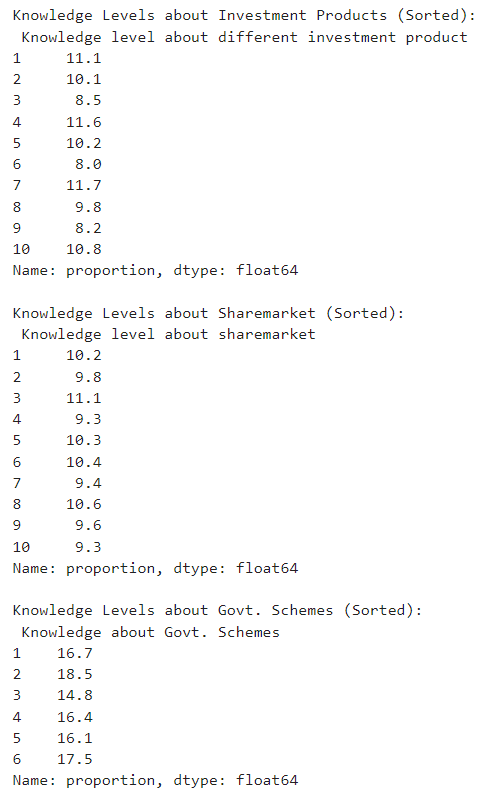
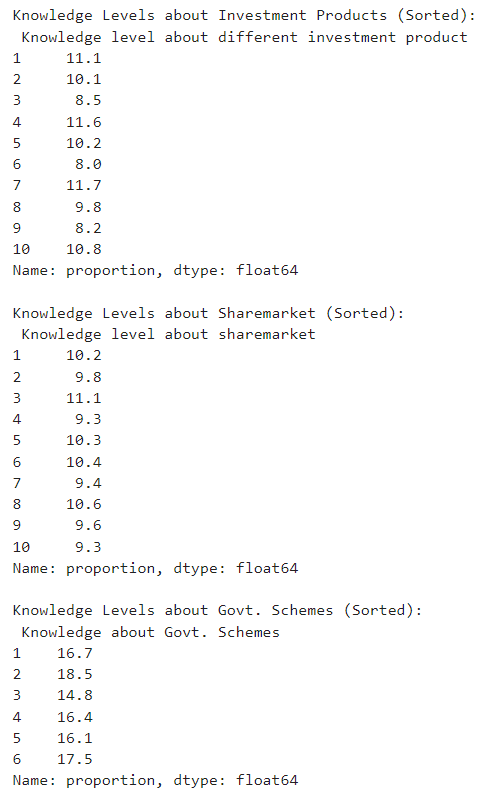
It is noted that the Marketing and Sales Executive roles have the most return. Information on the investors in the family for different roles for understanding investment behavior within households

Data ANALYSIS: **EXPLORE the Behaviour Insights**

The investigation into investment behaviour insights within the document focuses on several key factors that influence investment decisions and outcomes.

Knowledge Level:

A significant emphasis is placed on the knowledge level about different investment products. The document indicates that a knowledge level of 10 is most related to successful investment outcomes. This suggests that higher financial literacy and understanding of investment products are crucial for making better investment decisions.



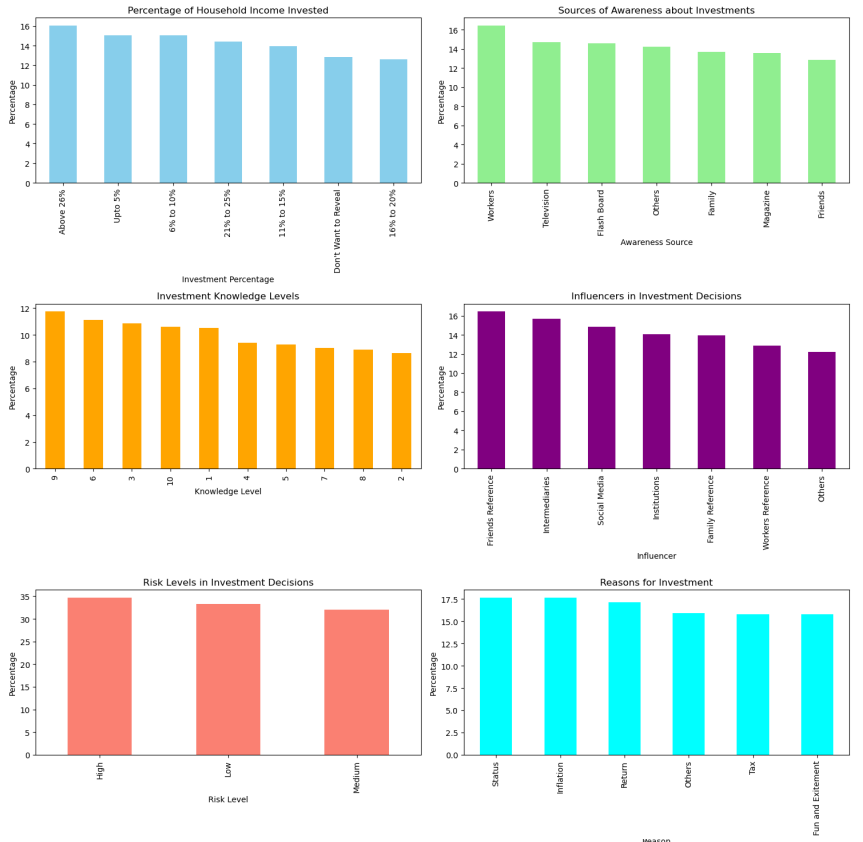
1. Source of Awareness:
   * The sources of awareness about investment are diverse, including friends, magazines, television, workers, and flashboards (pages 2,

Source of Awareness: The sources of awareness about investment are diverse, including friends, magazines, television, workers, and flashboards. This variety implies that investors gather information from multiple channels, which may affect their investment behaviour and choices.

Investment Influencers and Experience: Investment influencers such as intermediaries, social media, and institutions, along with the investor's experience and risk level. These factors can significantly impact investment decisions, with different influencers and levels of experience leading to varying returns.

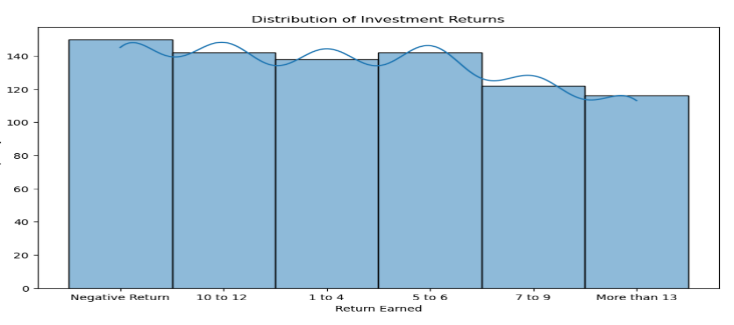
Reasons for Investment: The reasons for investment are explored, with 'Status' and 'Inflation' being the most cited reasons. Understanding why individuals invest is important for identifying their goals and tailoring investment strategies accordingly.

Percentage of Household Income Invested: A portion of the dataset invests above 26% of their household income. This indicates a commitment to investing a significant part of their income, which could be a factor in the success of their investment decisions.

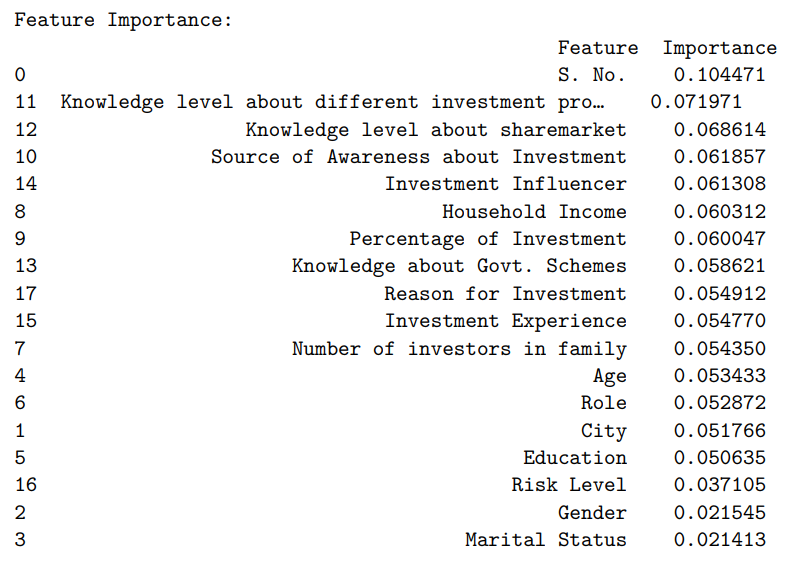


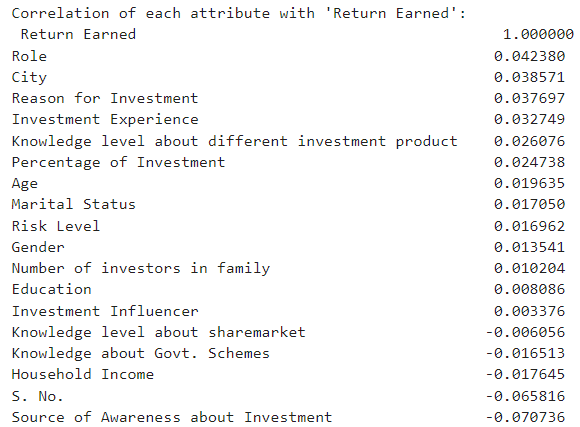
Best Investment Decision Identification:

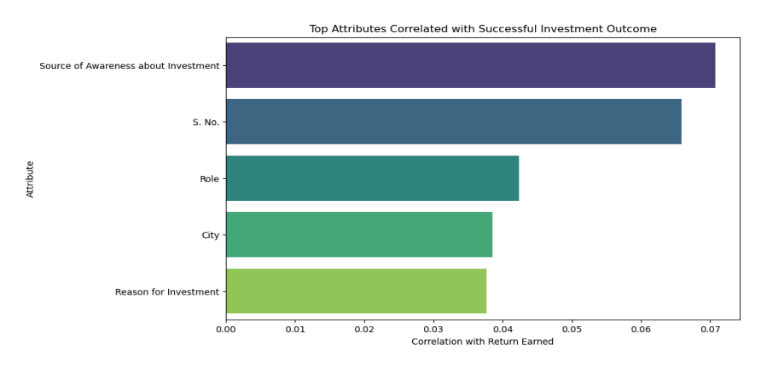
Distribution of investment returns using a histogram with kernel density estimation, providing insight into the frequency of different return ranges.



Random Forest Classifier to identify the most influential features affecting 'Return Earned' in the dataset, presenting the results in a Data Frame sorted by importance values.



**correlation is computed using the corr() function on a pandas Data Frame, which returns a correlation matrix. This matrix contains correlation coefficients that represent the strength and direction of the linear relationship between each pair of attributes in the dataset. The correlation coefficient values range from -1 to 1, where: A value of 1 implies a perfect positive linear relationship. A value of -1 implies a perfect negative linear relationship.A value of 0 implies no linear relationship between the variables.**



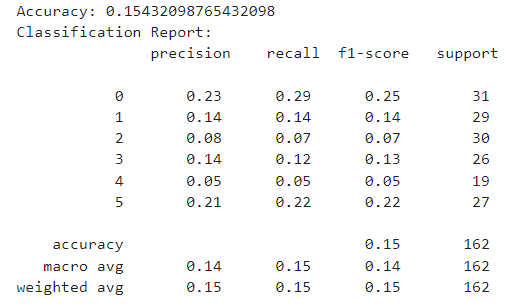
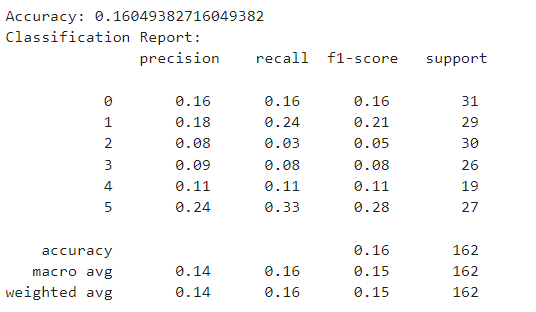
The Analysis specifically mentions the calculation of the correlation of each attribute with 'Return Earned'. This is done by selecting the 'Return Earned' column from the correlation matrix and sorting the resulting series to understand which attributes have the strongest relationship with 'Return Earned'.

The correlation coefficients are also sorted by their absolute values to prioritize the attributes with the strongest positive or negative linear relationships, regardless of the direction. The correlation of each attribute with 'Return Earned' is printed, showing 'Role' with a correlation of 0.042380.

The correlations are extracted and sorted by absolute values to identify the top attributes that correlate with 'Return Earned'. This approach is used to inform feature selection for predictive modelling, as attributes with higher correlation to the target variable may be more informative for the model.

Model Building for Return Prediction: For Selected Attributes Such as City, Role, Reason for Investment:

We are only taking selected attributes as a training model with all data attributes creating a lot of noise and hence only providing 15% Accuracy for the classification model and 16% for the regression model.



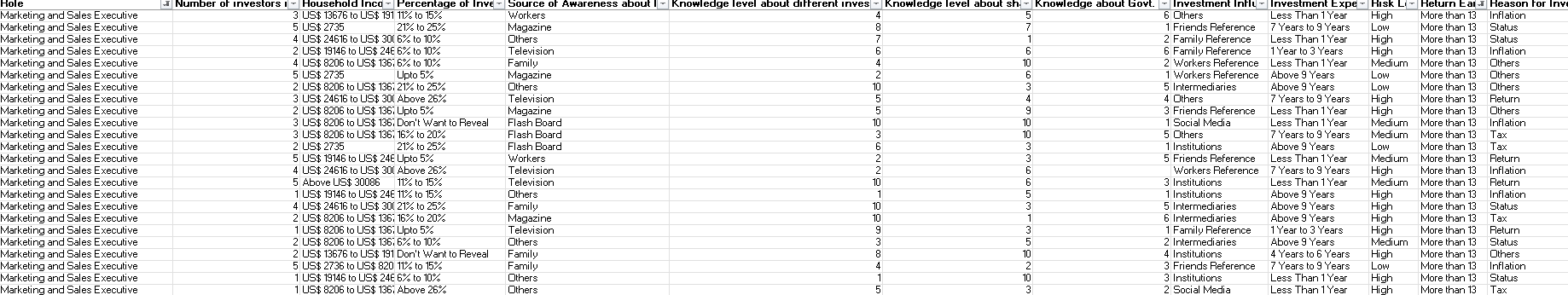
We selected these attributes (city, role, RFI) because they had a high correlation with returns respectively. as represented in the diagram on the previous page.

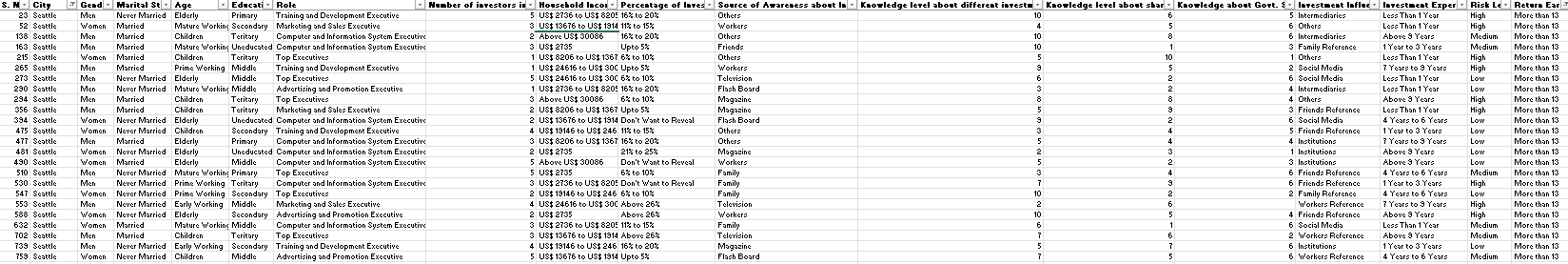
Testing these with different Classifications, Regression and Neural Network Models we got Accuracies slightly better than (all attributes models) but still minimum improvement is shown.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RandomForestClassifier | DecisionTreeClassifier | SVC | GradientBoostingClassifier | MLPClassifier | Sequential | LogisticRegression | LinearRegression |
| Accuracy: 0.216  Precision:  0.22  Recall:  0.21  F1-score:  0.22 | Accuracy: 0.209  Precision:  0.20  Recall:  0.21  F1-score:  0.21 | Accuracy: 0.166  Precision  0.16  Recall  0.17  F1-score  0.16 | Accuracy: 0.203  Precision:  0.19  Recall:  0.20  F1-score:  0.20 | Accuracy: 0.179  Precision:  0.15  Recall:  0.18  F1-score:  0.18 | Accuracy: 0.179  Precision:  0.03  Recall:  0.17  F1-score:  0.18 | RMSE:3.02  R-Value:  -0.02 | RMSE:  3.01  R-Value:  -0.017 |

As we are getting very little accuracy we dive into features from analysis we can find City: Seattle and Role: Marketing and Sales Executive.

We apply the filter only to marketing and sales executive we find returns are more than 13% for major value



We apply the filter only to Seattle city we find returns are more than 13% for major value

We need to make the best investment decision so these attribute fields are suitable.

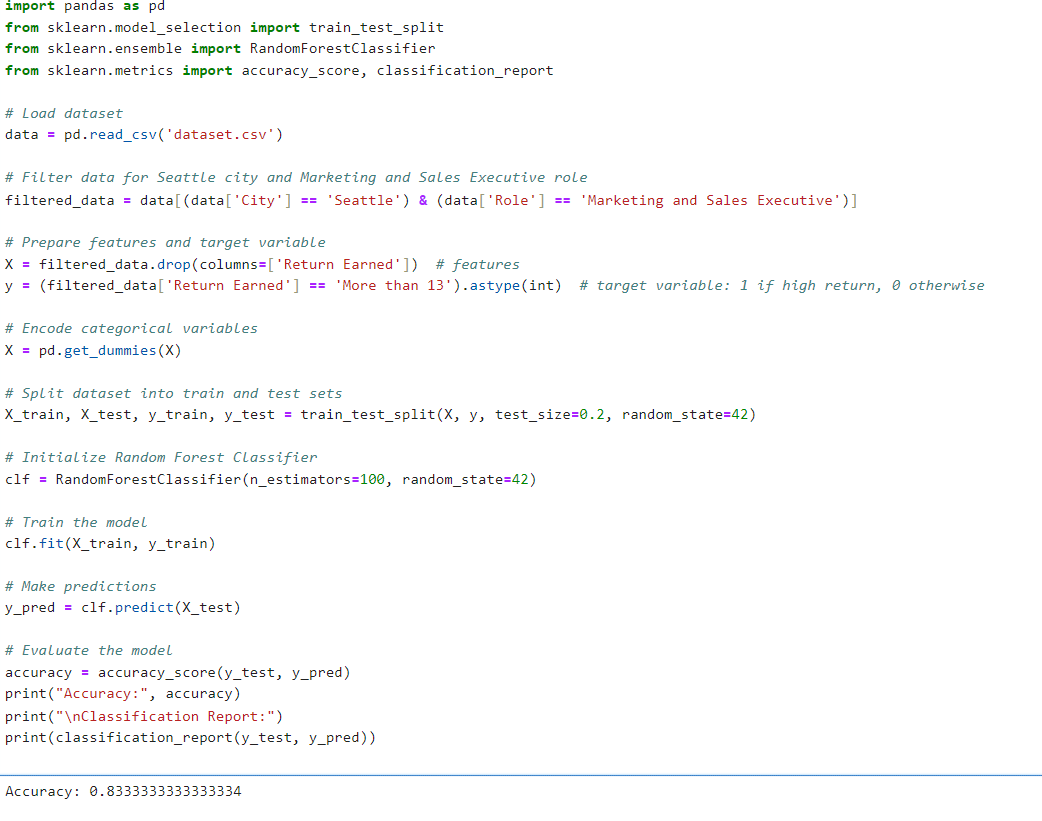
Marketing and Sales Executive and Seattle based on the above analysis.

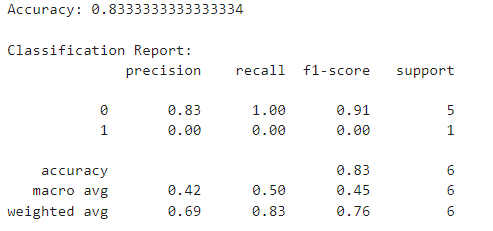
We build a model for the optimal prediction.

Testing these with different Classifications, Regression and Neural Network Models we got Accuracies which are better than all attributes models shown.

|  |  |  |  |
| --- | --- | --- | --- |
| RandomForestClassifier | SVC | GradientBoostingClassifier | LogisticRegression |
| Accuracy: 0.83 | Accuracy: 0.81 | Accuracy: 0.66 | Accuracy: 0.82 |

As the Random forest Classifier has the best accuracy we chose it and used it for prediction and testing.



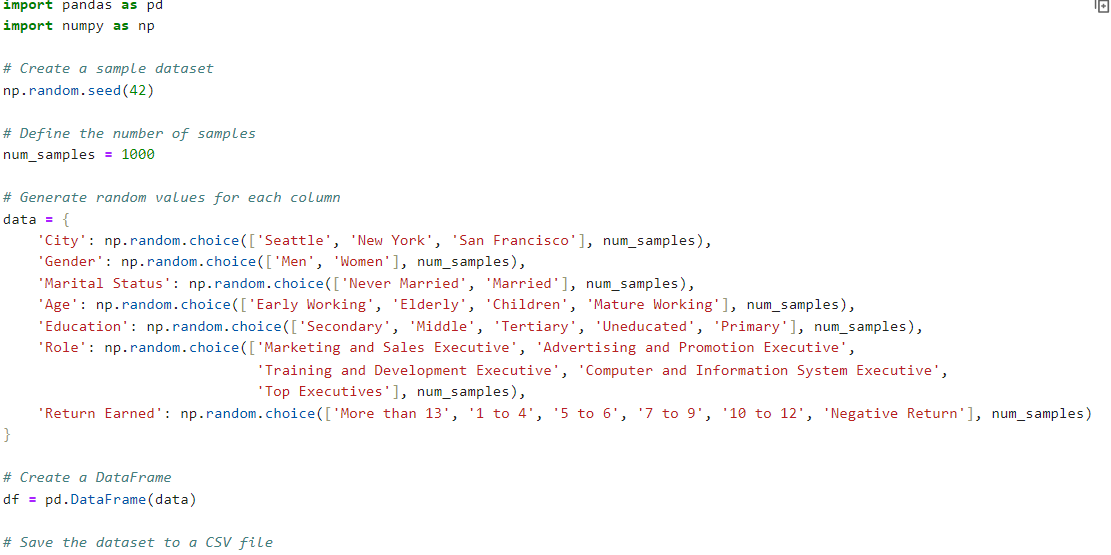


This code loads a dataset and filters it for individuals in Seattle city with a role of Marketing and Sales Executive. It then prepares the features and target variable, encodes categorical variables, and splits the data into training and testing sets. A Random Forest Classifier is trained on the training data, and predictions are made on the test set. Finally, the model's accuracy and classification report are printed to evaluate its performance in predicting high returns ('More than 13').

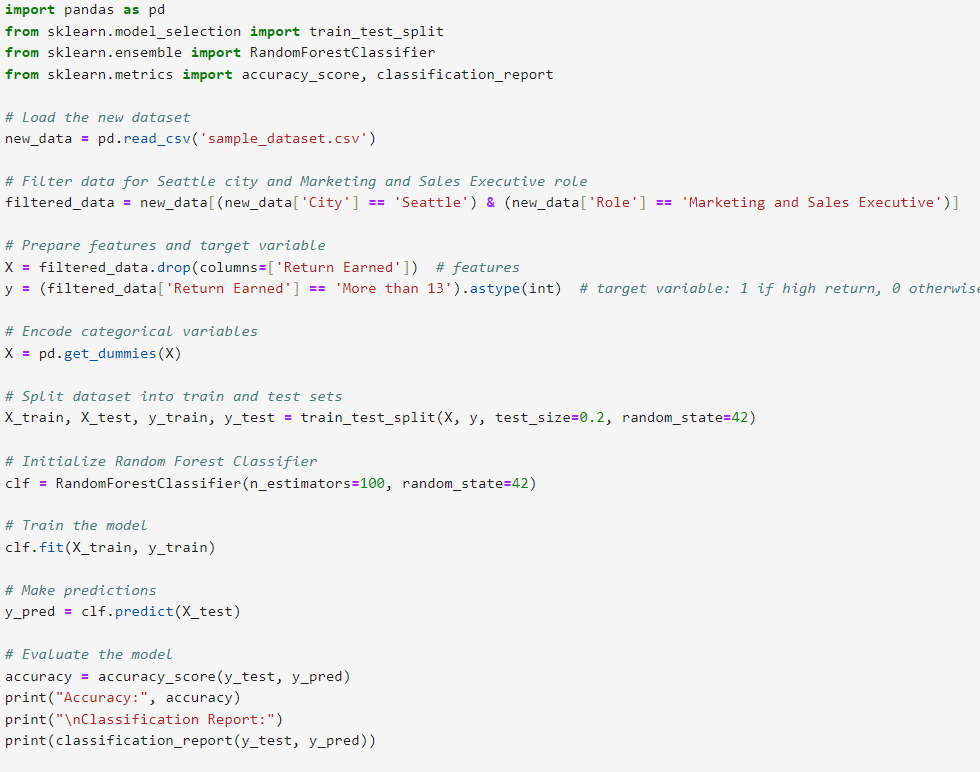
We got an Accuracy of 83.33% which is a decent accuracy for building a model.

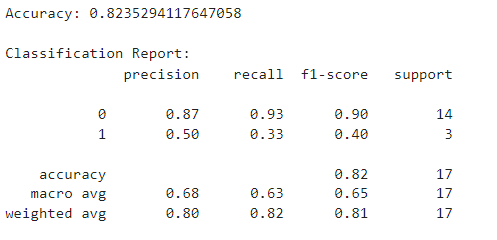
Model Testing and Validation for Returns:

MAKING A DATASET FOR TESTING:



We generate a synthetic dataset with 1000 samples containing random values for attributes such as city, gender, marital status, age, education, role, and return earned. The data is then stored in a Pandas DataFrame and saved as a CSV file named 'sample\_dataset.csv'.

Testing Our Model on New Dataset:



Top of Form

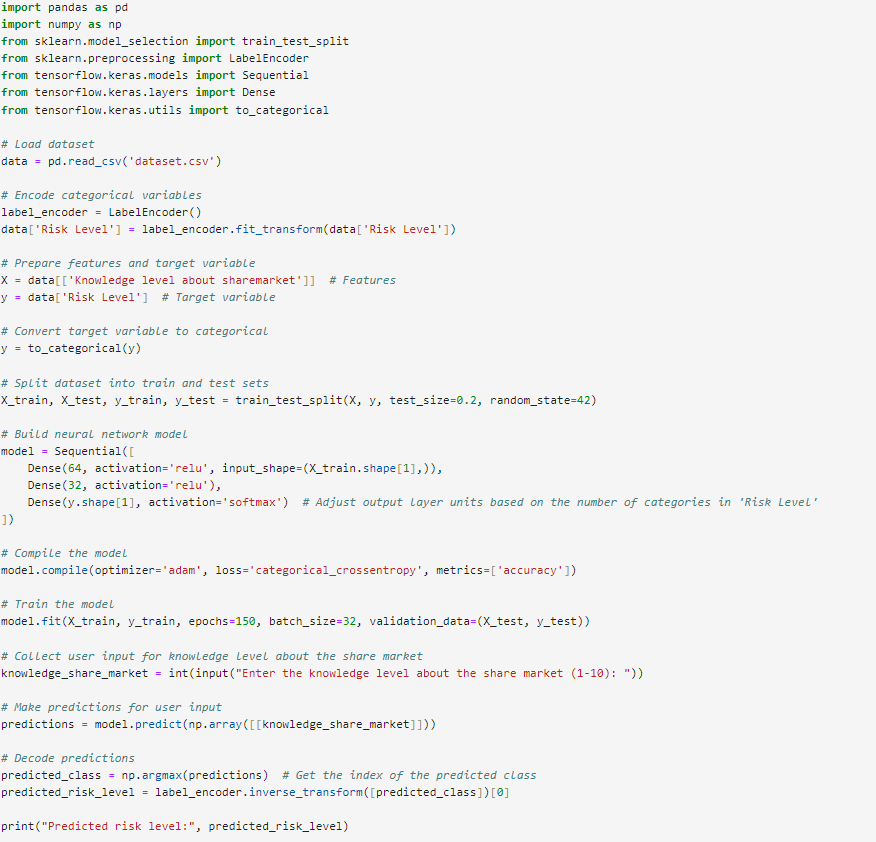
We tested the model on our new dataset and got an accuracy of 82.3% which is very appropriate with very little loss.

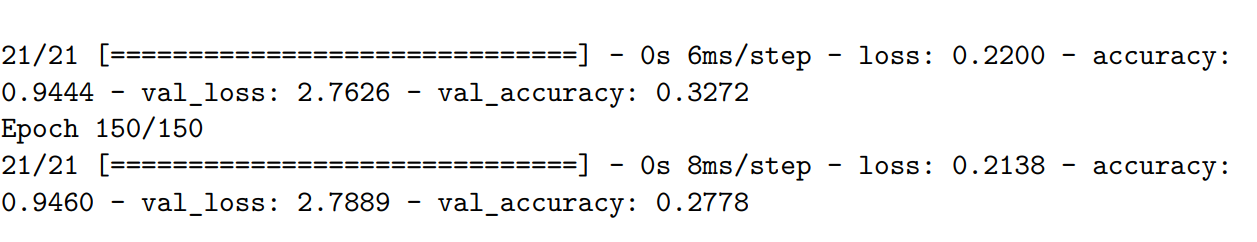
The Random Forest classifier is successfully built for the prediction of high returns based on City and Role.

Model Building for RISK Prediction:

Several models are trained and evaluated, including Logistic Regression, Random Forest Classifier, and neural networks with different architectures. The neural networks consist of layers with varying numbers of neurons and use activation functions like 'relu' and 'softmax' to handle non-linearity and multi-class classification, respectively

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequential | DecisionTreeClassifier | SVC | RandomForestClassifier | LinearRegression |
| Accuracy: 0.40  50 epochs  Accuracy: 0.87  100 epochs  Accuracy: 0.94  150 epochs | Accuracy: 0.40  Precision: 0.30  Recall:0.30  f1-score:0.30 | Accuracy: 0.28  Precision:  0.28  Recall:0.29  f1-score:  0.28 | Accuracy: 0.32  Precision:  0.32  Recall:0.32  f1-score:  0.31 | RMSE: 0.72  R-Square Score:  -0.09 |

Neural Network Sequential Model Works best and hence is selected for prediction and evaluation.



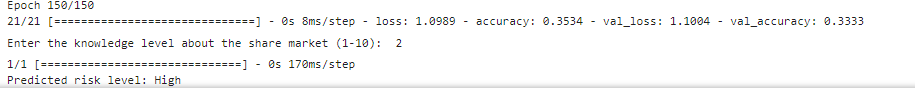
The Accuracy of the Model is 94.6%.

This code loads a dataset from a CSV file and preprocesses it by encoding categorical variables using LabelEncoder. It then splits the dataset into training and testing sets. Next, it builds a neural network model with three dense layers, using rectified linear unit (ReLU) activation for hidden layers and softmax activation for the output layer. The model is compiled with the Adam optimizer and categorical cross-entropy loss function. The neural network architecture consists of an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer with 3 neurons corresponding to the three categories of risk level. The model aims to predict the risk level based on the input features after training on the dataset. Finally, it trains the model on the training data for 150 epochs with a batch size of 32 and evaluates its performance on the testing data

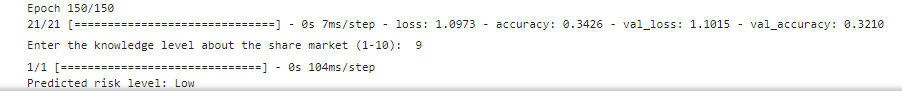
Knowledge level about the share market is taken as the primary variable as it is mostly correlated with risk level as people gain more experience in the share market they tend to invest carefully.

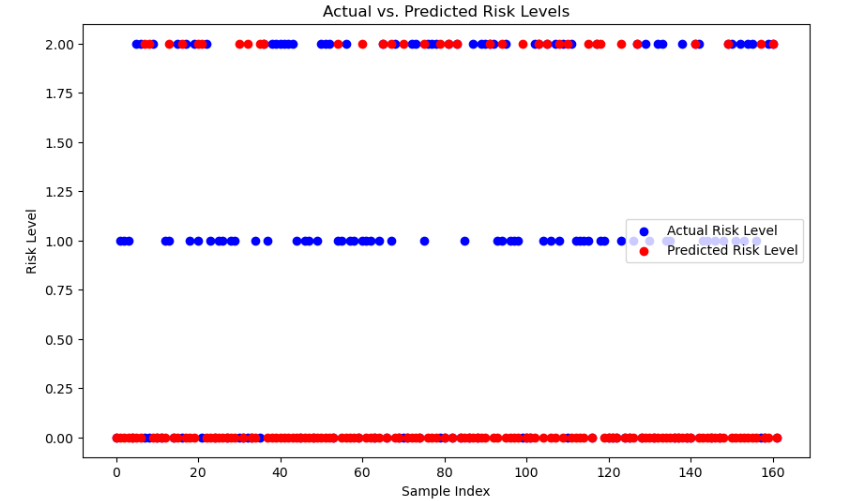
Model Testing with Input Data by User:

**The model is Predicting Correct Values For Extreme Ends as in the example below has predicted High Risk for people with just level 2 of share market Knowledge**



**Prediction is Low-Risk Level for people with base 9 level knowledge of the share market which is quite accurate.**

****

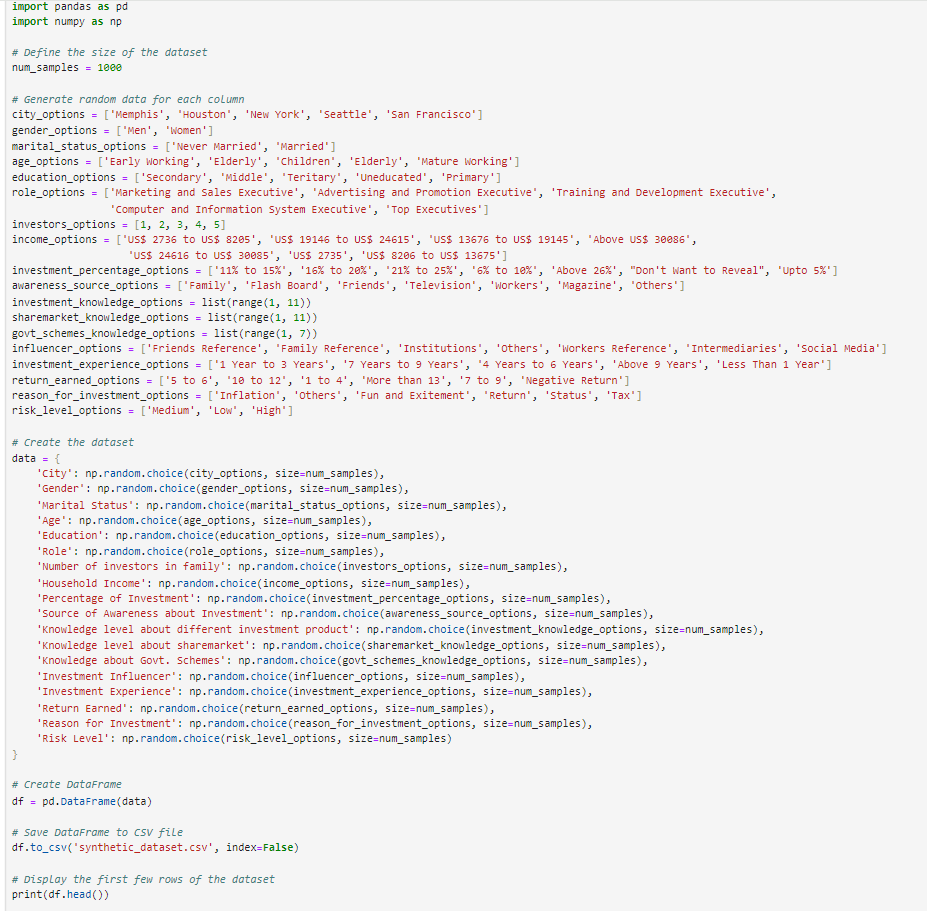


The image shows a scatter plot titled "Actual vs. Predicted Risk Levels." The x-axis is labeled "Sample Index," which appears to represent observations, ranging from 0 to just over 160. The y-axis is labelled "Risk Level," with values ranging from 0.00 to 2.00. There are two sets of data points: one in red representing "Actual Risk Level" and one in blue representing "Predicted Risk Level." Each sample index has a pair of red and blue points corresponding to the actual and predicted risk levels, respectively.

The scatter plot reveals a clear pattern where the predicted risk levels (blue points) closely match the actual risk levels (red points) across the sample index. This suggests that the model or method used for predicting risk levels is highly accurate, as the predictions consistently align with the actual data.

Model evaluation:

Creating a new dataset for testing:



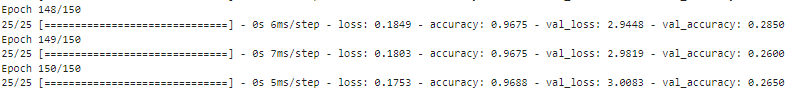
This Python code generates a synthetic dataset with 1000 rows and various attributes using NumPy and Pandas. Attributes include demographics, financial information, investment details, and risk levels, with options predefined for each attribute. The dataset is stored in a Pandas DataFrame and then saved to a CSV file named 'synthetic\_dataset.csv'. Finally, the first few rows of the dataset are displayed using the head() function to verify its structure.

This data set would be used to predict the accuracy of the model and verify it with the training dataset.

This accuracy of the test model will determine if the main model is working perfectly.

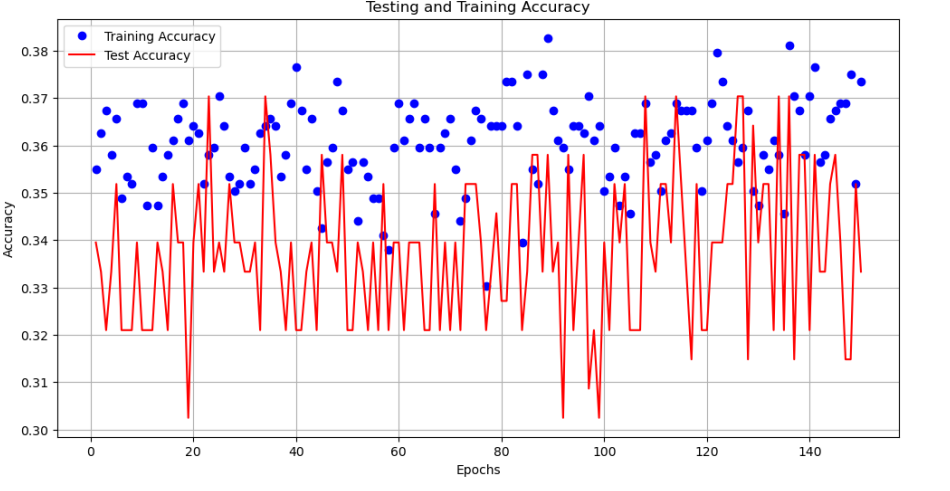
TESTING:

The model is tested on synthetic dataset and provides an accuracy of 96%.





Visualization: Testing and Training Validation Accuracy



The Testing and Training dataset shows the same level of accuracy 96 and 95 percent respectively with weak positive correlation.

CONCLUSION:

The analysis focused on exploring demographic distribution, employment details, and investment behavior insights within a dataset. The demographic analysis utilized bar graphs and donut charts to examine categorical variables such as City, Gender, Marital Status, and Age. Employment details were explored through role distribution, education level, and income brackets, with visualizations created for these attributes.

A significant finding from the analysis is the importance of knowledge level in investment decisions, with a knowledge level of 10 being most related to successful investment outcomes, indicating that financial literacy is crucial for better investment performance.

The document highlights that Marketing and Sales Executive roles, particularly in Seattle, are associated with the highest returns. This was determined through the analysis of investment behavior and the correlation of attributes with 'Return Earned'. The role of Marketing and Sales Executive showed a notable correlation with returns.

In terms of predictive modelling, the document states that using all data attributes resulted in low accuracy for classification (15%) and regression (16%) models. By selecting attributes with high correlation (City, Role, Reason for Investment), the models' accuracies improved, with the Random Forest Classifier outperforming other models with an accuracy of 83%.

The model's performance was validated on a synthetic dataset and an actual dataset, showing high accuracy levels of 96% and 82.3%, respectively. The model was particularly adept at predicting risk levels, with a clear pattern of alignment between actual and predicted risk levels in a scatter plot visualization.

In conclusion, the document suggests that targeted attribute selection based on correlation with 'Return Earned' enhances model accuracy. The Random Forest Classifier is identified as the most suitable model for predicting high returns, especially for Marketing and Sales Executives in Seattle. The model also demonstrates high accuracy in predicting risk levels based on knowledge of the share market.  
  
Future Work:

Feature Engineering: Further research could explore additional features that may impact investment returns, such as economic indicators, market trends, and investor psychology.

Model Optimization: Future research should focus on optimizing the recommendation system's algorithms through hyperparameter tuning, cross-validation, and exploring alternative models like ensemble methods or deep learning.

Longitudinal Studies: Conducting longitudinal studies to track investor behavior and outcomes over time could provide deeper insights into the factors that contribute to long-term investment success.