Investment Decision Recommendation System Project - Documentation

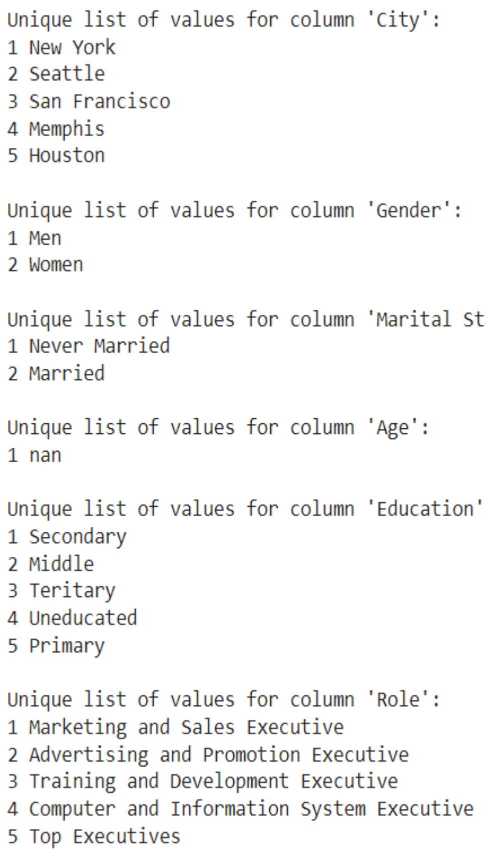
- 21pc03-Akhil.S.M

Checkpoint 1:

The demographic distribution like gender, marital status, age among individuals in the dataset is analysed. 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 been investigated.

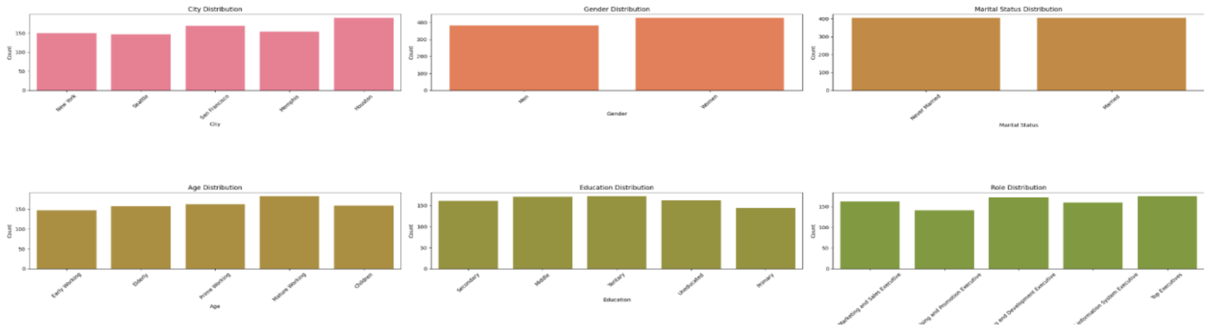
The necessary preprocessing of the given dataset like cleaning is done. The unique values in each attribute is been printed. The libraries like pandas , numpy , matplot are used for visualization of distribution of each and every attribute.

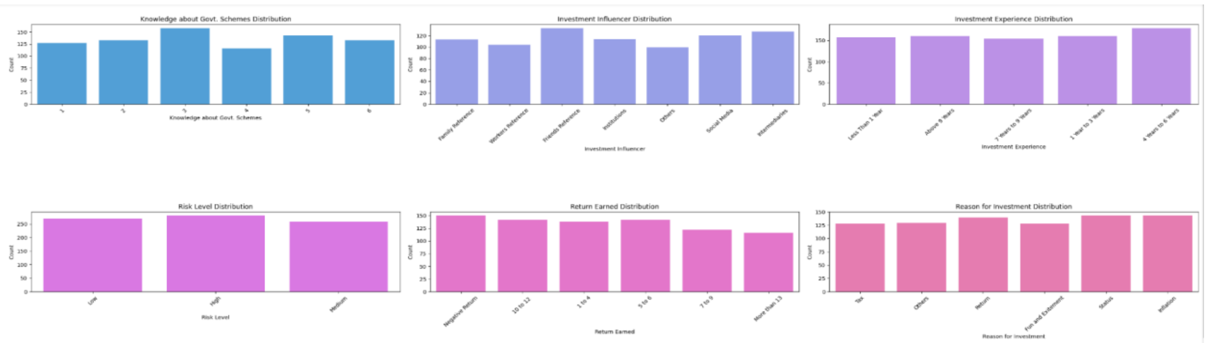
Example output:

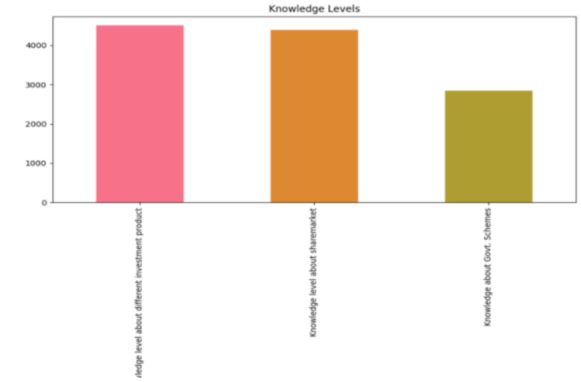


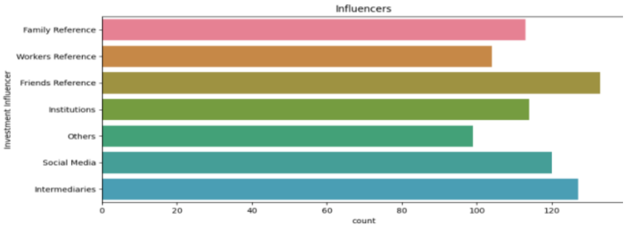
```
Unique list of values for column 'City':
1 New York
2 Seattle
3 San Francisco
4 Memphis
5 Houston

Unique list of values for column 'Gender':
1 Men
2 Women

Unique list of values for column 'Marital Status':
1 Never Married
2 Married

Unique list of values for column 'Age':
1 nan

Unique list of values for column 'Education':
1 Secondary
2 Middle
3 Teritary
4 Uneducated
5 Primary

Unique list of values for column 'Role':
1 Marketing and Sales Executive
2 Advertising and Promotion Executive
3 Training and Development Executive
4 Computer and Information System Executive
5 Top Executives
```

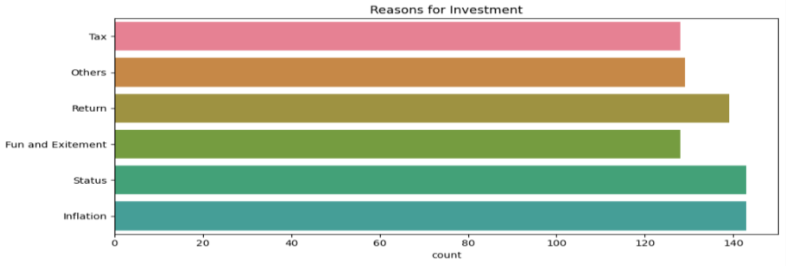
```
S.No       ColumnName                     Datatype
1          S. No.                         int64
2          City                           object
3          Gender                         object
4          Marital Status                 object
5          Age                            object
6          Education                      object
7          Role                           object
8          Number of investors in family int64
9          Household Income               object
10         Percentage of Investment       object
11         Source of Awareness about Investmentobject
12         Knowledge level about different investment productint64
13         Knowledge level about sharemarketint64
14         Knowledge about Govt. Schemes int64
15         Investment Influencer          object
16         Investment Experience          object
17         Risk Level                     object
18         Return Earned                  object
19         Reason for Investment          object
```











Checkpoint 2:

The factors that contribute for making best investment decisions are been identified. The attributes like Roles, city, Investment influencer and reason for investment are been encoded using one-hot encoder and all other attributes are encoded with label encoder. Then the correlation matrix is been built and then the logistic regression model is been built and the coefficients of the attributes influencing the returns earned is plotted in a bar graph, with which we can infer that,

-The people who has the reason to invest by others has the most positive influence

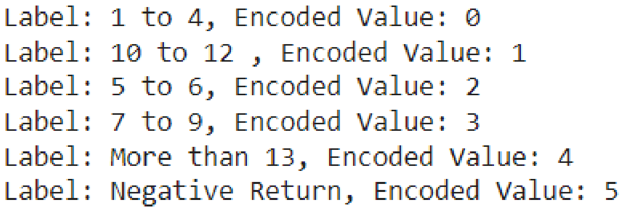
-The percentage of investment doesn't have an impact.

-The people who invest for fun and excitement has the most negative influence

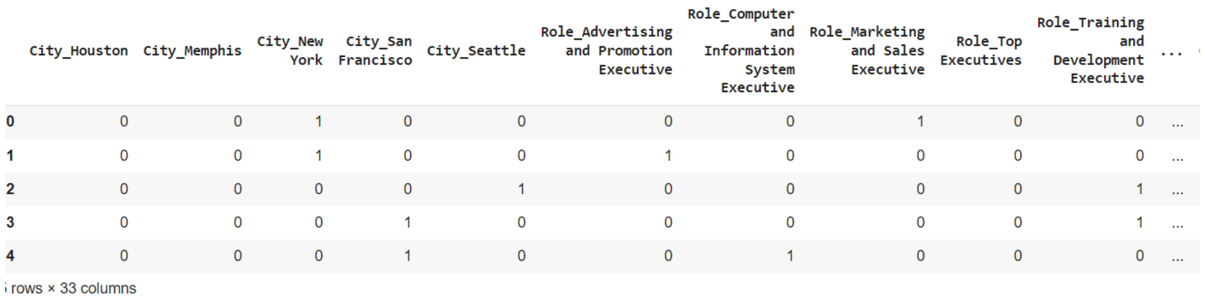
-The people influenced by social media has the second most negative influence

And many more inferences ,

Example Output:

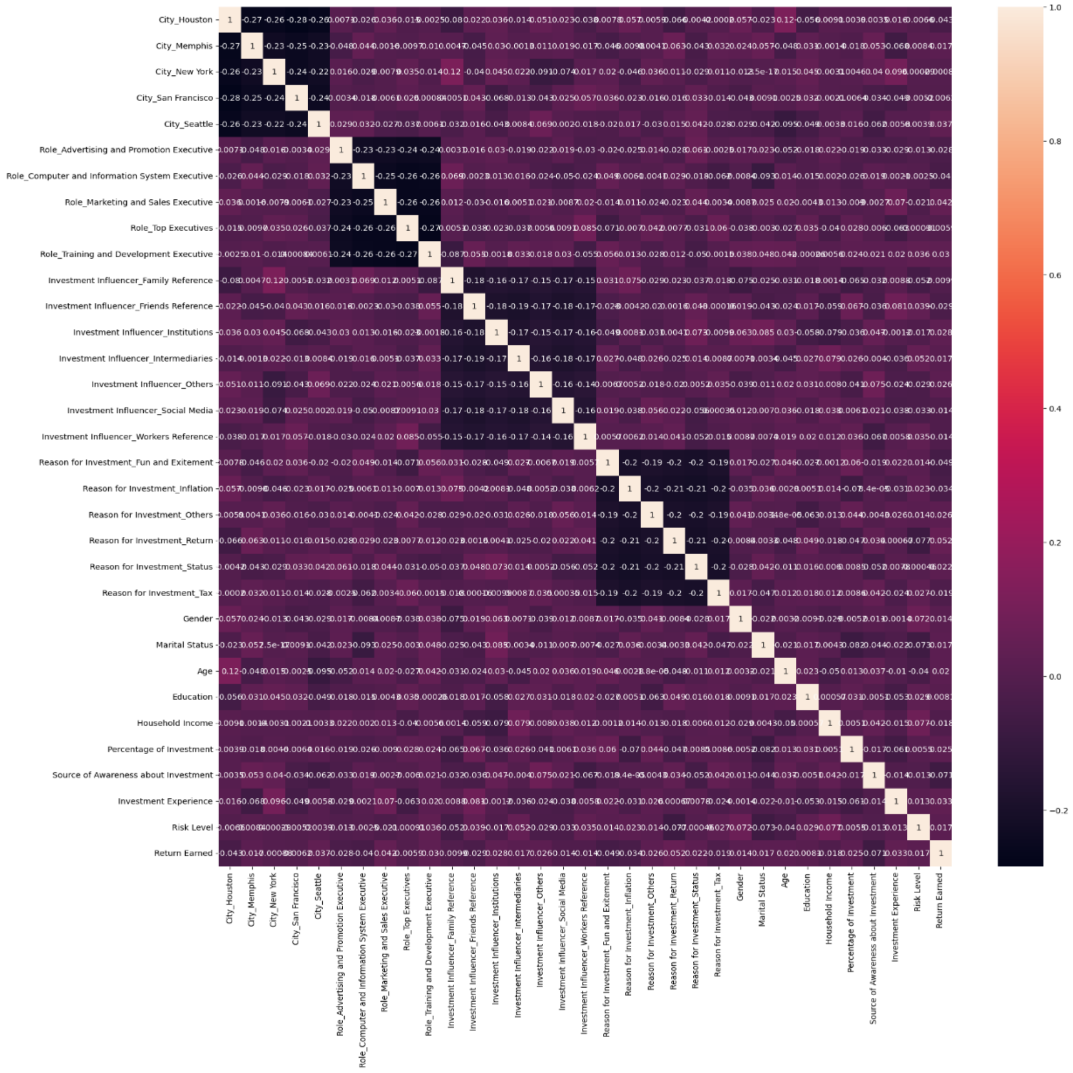


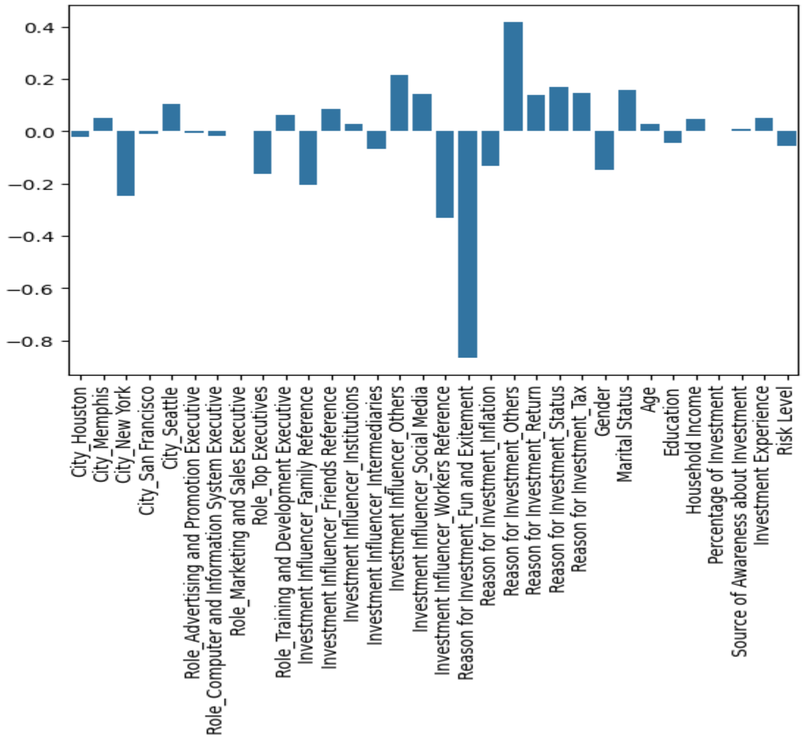
```
Label: 1 to 4, Encoded Value: 0
Label: 10 to 12 , Encoded Value: 1
Label: 5 to 6, Encoded Value: 2
Label: 7 to 9, Encoded Value: 3
Label: More than 13, Encoded Value: 4
Label: Negative Return, Encoded Value: 5
```



	City_Houston	City_Memphis	City_New York	City_San Francisco	City_Seattle	Role_Advertising and Promotion Executive	Role_Computer and Information System Executive	Role_Marketing and Sales Executive	Role_Top Executives	Role_Training and Development Executive	...
0	0	0	1	0	0	0	0	1	0	0	...
1	0	0	1	0	0	1	0	0	0	0	...
2	0	0	0	0	1	0	0	0	0	1	...
3	0	0	0	1	0	0	0	0	0	1	...
4	0	0	0	1	0	0	1	0	0	0	...

5 rows × 33 columns

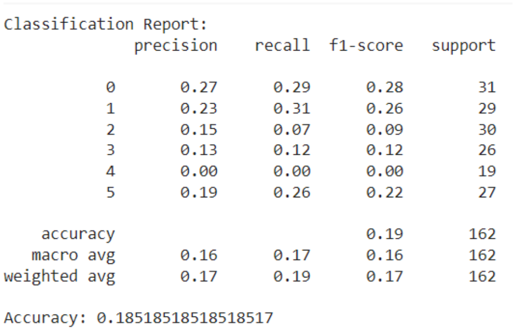

City_Houston
City_Memphis
City_New York
City_San Francisco
City_Seattle
Role_Advertising and Promotion Executive
Role_Computer and Information System Executive
Role_Marketing and Sales Executive
Role_Top Executives
Role_Training and Development Executive
Investment Influencer_Family Reference
Investment Influencer_Friends Reference
Investment Influencer_Institutions
Investment Influencer_Intermediaries
Investment Influencer_Others
Investment Influencer_Social Media
Investment Influencer_Workers Reference
Reason for Investment_Fun and Exitement
Reason for Investment_Inflation
Reason for Investment_Others
Reason for Investment_Return
Reason for Investment_Status
Reason for Investment_Tax
Gender
Marital Status
Age
Education
Household Income
Percentage of Investment
Source of Awareness about Investment
Investment Experience
Risk Level
Return Earned
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
City_Houston
City_Memphis
City_New York
City_San Francisco
City_Seattle
Role_Advertising and Promotion Executive
Role_Computer and Information System Executive
Role_Marketing and Sales Executive
Role_Top Executives
Role_Training and Development Executive
Investment Influencer_Family Reference
Investment Influencer_Friends Reference
Investment Influencer_Institutions
Investment Influencer_Intermediaries
Investment Influencer_Others
Investment Influencer_Social Media
Investment Influencer_Workers Reference
Reason for Investment_Fun and Exitement
Reason for Investment_Inflation
Reason for Investment_Others
Reason for Investment_Return
Reason for Investment_Status
Reason for Investment_Tax
Gender
Marital Status
Age
Education
Household Income
Percentage of Investment
Source of Awareness about Investment
Investment Experience
Risk Level

Checkpoint 3:

The Random forest ML model is built and prediction is done. The evaluation process is also done for the given dataset, where the attribute “Return Earned” is taken as Y and all other encoded attributes are taken as X.

Example output:



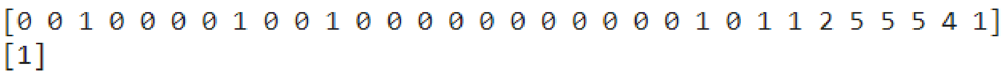
```
Classification Report:
              precision    recall  f1-score   support

           0       0.27      0.29      0.28        31
           1       0.23      0.31      0.26        29
           2       0.15      0.07      0.09        30
           3       0.13      0.12      0.12        26
           4       0.00      0.00      0.00        19
           5       0.19      0.26      0.22        27

    accuracy                           0.19       162
   macro avg       0.16      0.17      0.16       162
weighted avg       0.17      0.19      0.17       162

Accuracy: 0.18518518518518517
```

The accuracy was too low, so I tried out with SVM and also tried with ensemble, voting method

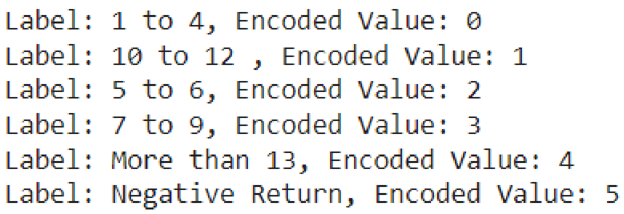
Checkpoint 4:

I implemented the same random forest and generated the output for a given list of features in the encoded format and the output is returned as predicted returns earned in the encoded format, for which we should map with the respective values (The second part of this checkpoint is still on process).

Example output:



```
[0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 2 5 5 5 4 1]
[1]
```



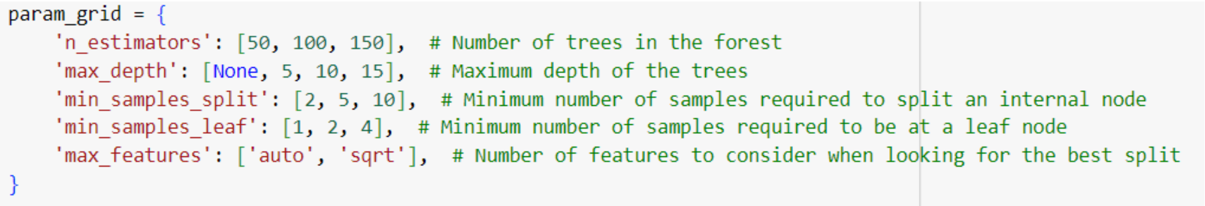
```
Label: 1 to 4, Encoded Value: 0
Label: 10 to 12 , Encoded Value: 1
Label: 5 to 6, Encoded Value: 2
Label: 7 to 9, Encoded Value: 3
Label: More than 13, Encoded Value: 4
Label: Negative Return, Encoded Value: 5
```

Which means for the given value it is predicted as the returns will be 10 to 12.

Checkpoint 5:

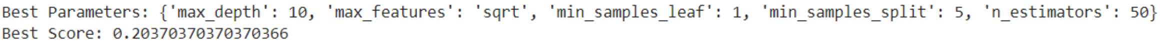
The fine tuning of the model is done by gridsearch CV method and I tried to improve the accuracy .

Example output:



```
param_grid = {
    'n_estimators': [50, 100, 150],  # Number of trees in the forest
    'max_depth': [None, 5, 10, 15],  # Maximum depth of the trees
    'min_samples_split': [2, 5, 10],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 2, 4],  # Minimum number of samples required to be at a leaf node
    'max_features': ['auto', 'sqrt'],  # Number of features to consider when looking for the best split
}
```

These are the different combinations of parameter that is been tried.

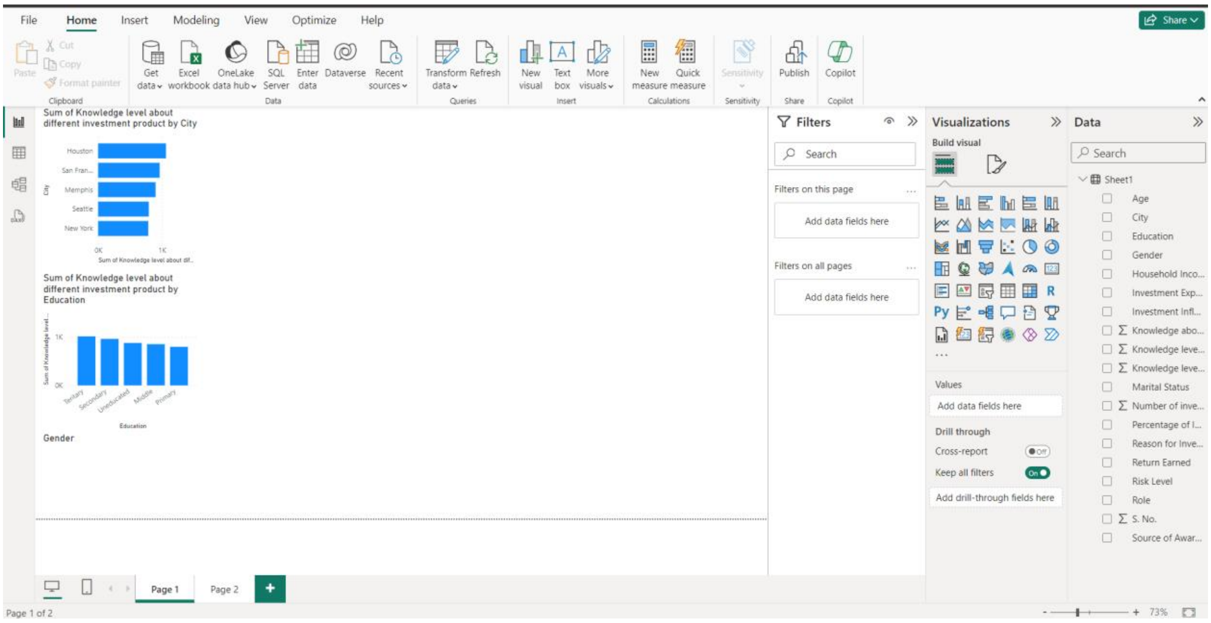


```
Best Parameters: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50}
Best Score: 0.20370370370370366
```

(But even then the accuracy is very low, but by giving minimum sample test as 1 we can get an accuracy of 0.33)

Checkpoint 6:

The power BI dashboard is created and the visualizations for different combinations is done!!



Many more visualizations can be done and its really very interesting to use this tool!!