
Buckman Hackathon

— Mahizha 21PW13 —

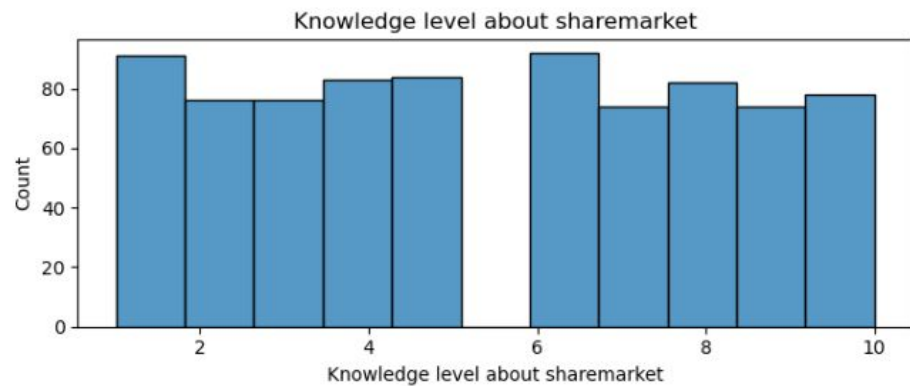
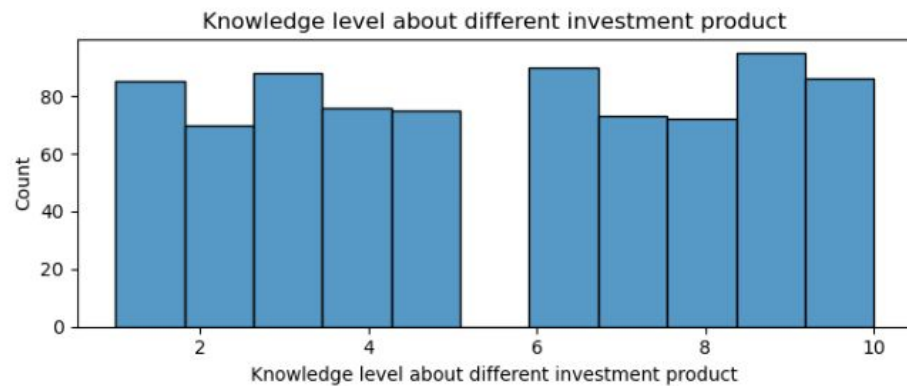
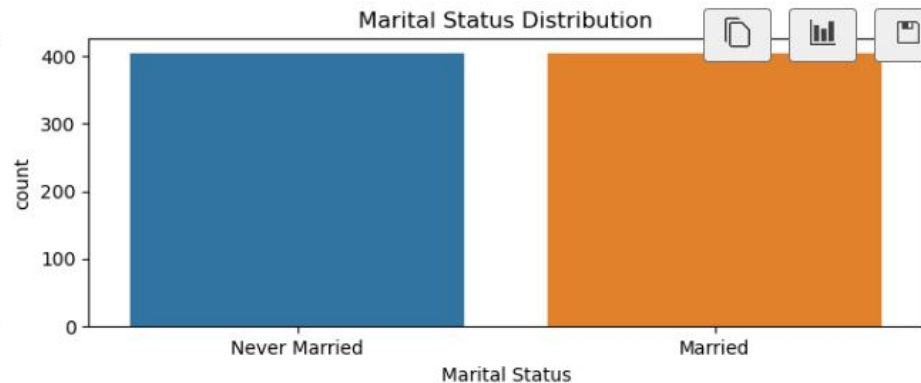
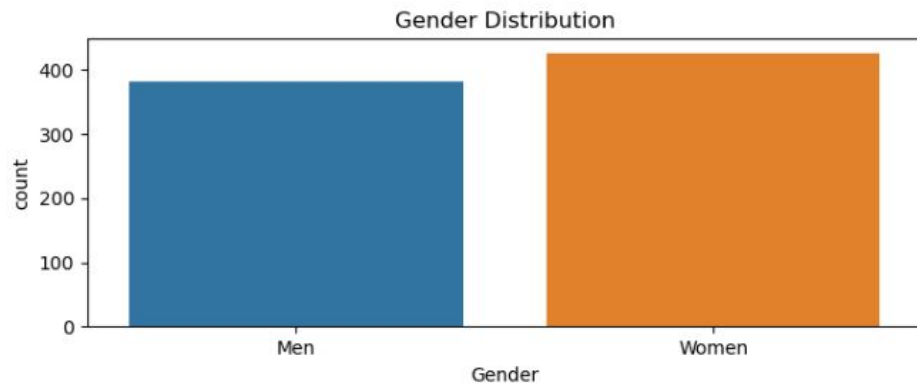
Investment Decision Recommendation System Project

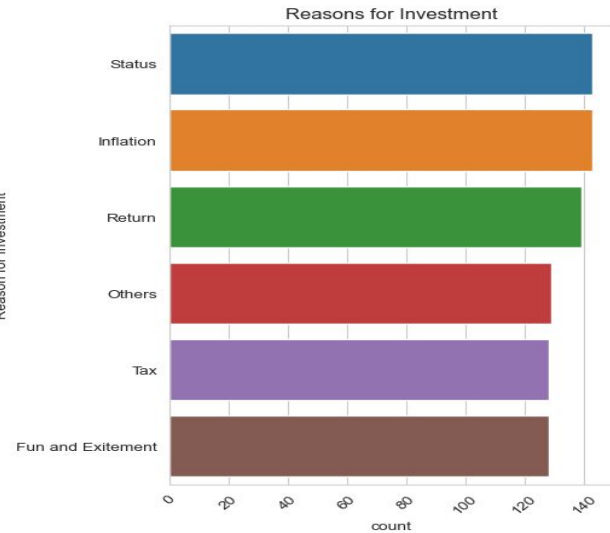
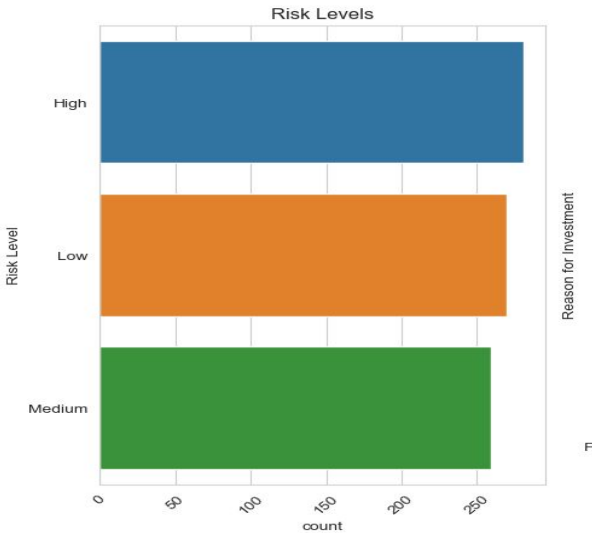
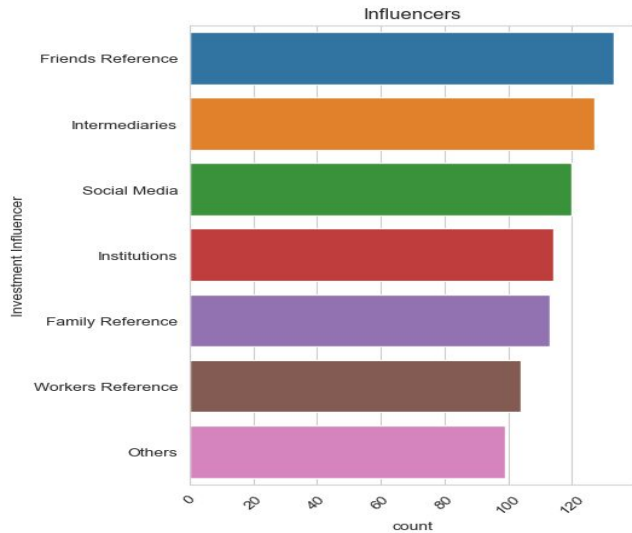
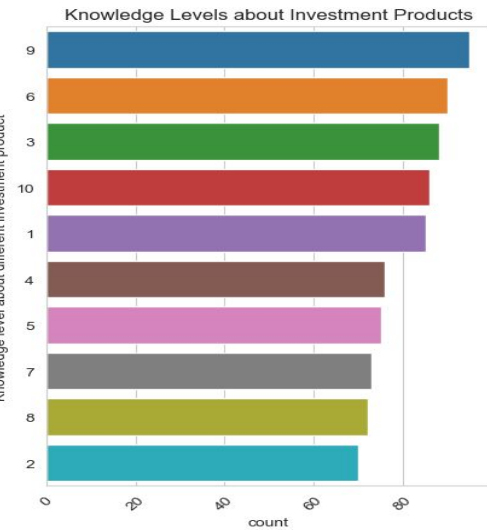
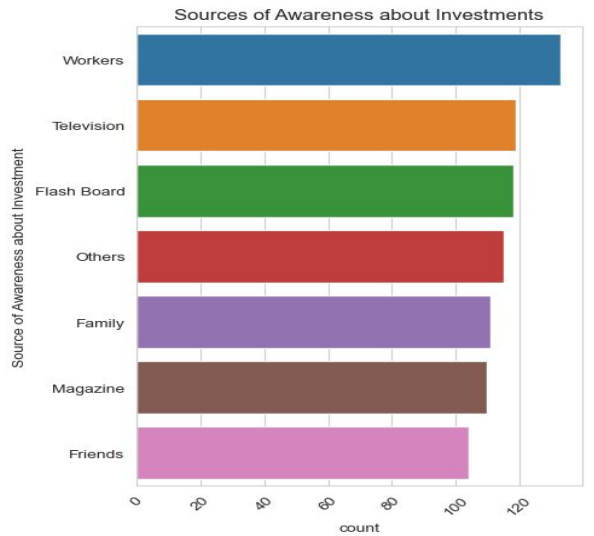
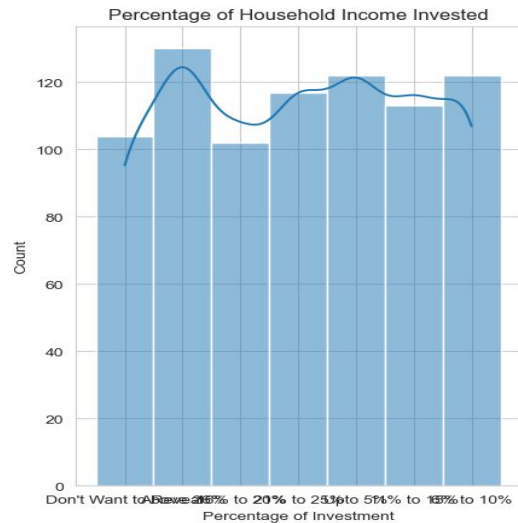
- We were provided with a data set containing information about individuals and their investment behavior.
- The task is to build a recommendation system that can predict the best investment decision for new data based on various factors available in the dataset.

Here I will guide through my the insights I observed and how i re-learned my approach. (approach varies by person's perspective)

1)Data Exploration:

- Initial step is to go through the given data set.
- Data Cleaning and Understanding:
 - Check for null values
 - And to make it understandable , i have converted it to integral values
- Python Visualization to understand the segregation
- Performed the investment behavioural insights including the percentage of household income invested, sources of awareness about investments, knowledge levels, influencers, risk levels, and reasons for investment.





2)Best Investment Decision Identification:

- Converted categorical data to numerical data , to train the model.
- Done using LabelEncoder and OneHotEncoder
- The correlation is performed for each behavioural features , to find the best feature which influences the recommendation system.
- Chi-Square Test is used for feature selection

Initially while reading the data set , i thought the main objective was to make the client get maximum profit , and used “Returns Earned” feature as the target variable.

From both Chi-Square test and Random Forest model learning, the Top features are learnt. ie) these features have much more value in contributing to the "Return Earned"

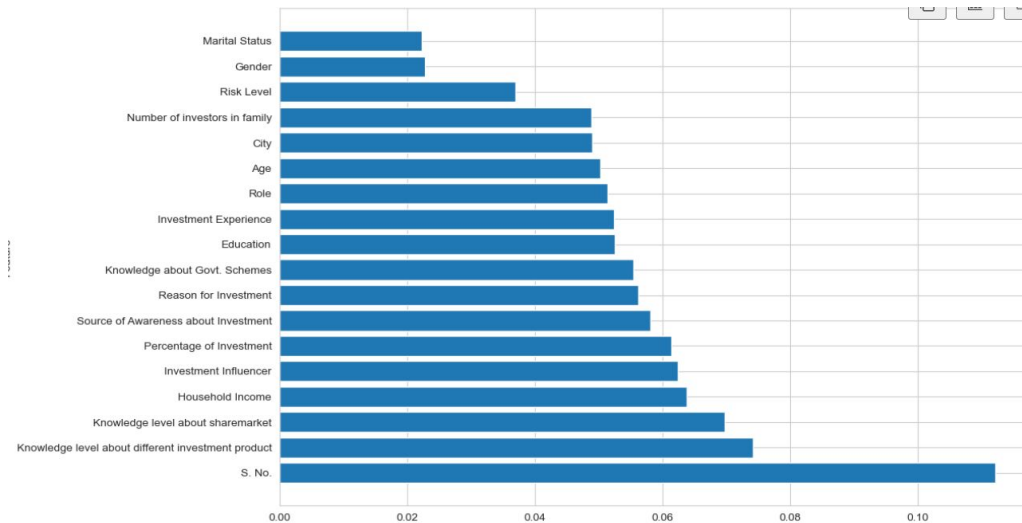
Top-Features being :

- * 'Knowledge level about sharemarket'
- * 'Knowledge level about different investment product'
- * 'Investment Influencer'
- * 'Percentage of Investment'

and

Least-contributing features being :

- * Gender
- * Martial Status



re-learning

- But on re analysing the data set and the problem statement, the main outcome is the recommendation system , so I have kept the Amount to invest as needed and , given the clients behaviour as my recommended value.
- “Percentage to Investment” as the target value, and proceeded my findings based on that.

2. Determine which demographic, employment, and behavioral characteristics correlate with

successful investment outcomes

To approach this:

Chi-Square Test: Conducted a chi-square test for each categorical variable in X to determine if there is a significant association between the variable and the Returns.

Chi-Square Test Results:

Feature Importances:

| | |
|-------------|----------------------|
| Feature 1: | 0.10367665087519233 |
| Feature 2: | 0.04982687924779722 |
| Feature 3: | 0.021180054872153957 |
| Feature 4: | 0.022431000516132112 |
| Feature 5: | 0.053308353174029904 |
| Feature 6: | 0.05175693173476706 |
| Feature 7: | 0.05047909275126268 |
| Feature 8: | 0.051547706010960266 |
| Feature 9: | 0.06298851188055753 |
| Feature 10: | 0.06027558841039122 |
| Feature 11: | 0.06161514934700321 |
| Feature 12: | 0.06853318407282565 |
| Feature 13: | 0.07141957497515575 |
| Feature 14: | 0.0569521882574521 |
| Feature 15: | 0.061905715609347337 |
| Feature 16: | 0.05378338911848524 |
| Feature 17: | 0.03889181966296834 |
| Feature 18: | 0.059428209483518166 |

3)Recommendation System Development:

- Here i have checked the presence of outliers, using first quartile and third quartile ,And the results was found to be that no presence of outliers, which is great , to train the model.
- I have found the accuracy, f-1 score for the 3 models - decisionTree, RandomForest,Neural Network.
- The one with best score is chosen.
- Note: I did observe that relatively all the models score range were small.

Selected features based on insights - “Percentage of Investment”

Selected features:

```
Index(['S. No.', 'Role', 'Knowledge level about different investment product',  
      'Investment Influencer', 'Return Earned'],  
      dtype='object')
```

Feature Selection Information:

```
{'selected_features': ['S. No.', 'Role', 'Knowledge level about different investment product',  
  'Investment Influencer', 'Return Earned'], 'scores': [389.8491633948299, 4.047448551211681,  
  0.44031412066083203, 4.4469568117173, 4.0967606945080846, 7.431002344646254, 10.98712004532755,  
  3.8557528106983856, 1.3793511863568446, 3.495605388744059, 13.374459533885819, 4.295986891602299,  
  0.614220097693017, 13.407256857158444, 9.039224928213692, 4.761813268440632, 11.319975434378211,  
  6.786373955195057]}
```

4)Recommendation System Development:

```
#Decision tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from joblib import dump
from sklearn.neural_network import MLPClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report

dt_clf = DecisionTreeClassifier(random_state=42)
dt_clf.fit(X_train, y_train)
y_pred_dt = dt_clf.predict(X_test)

#random forest

rf_clf = RandomForestClassifier(random_state=42)
rf_clf.fit(X_train, y_train)
y_pred_rf = rf_clf.predict(X_test)

# Assuming rf_clf is your trained Random Forest model
dump(rf_clf, 'random_forest_model.joblib')
```

- Used Random Forest Model and created the Model called -
- **Random_forest_model.joblib**
- I have dumped the learned model, so can be used for later and outside the environment.

```
from joblib import load

# Load the saved Random Forest model
loaded_rf_clf = load('random_forest_model.joblib')
```

✓ 0.0s

5)Implementation:

- 1.Implement the recommendation system using a programming language of your choice - Used Python Environment.
- 2. Ensure the system is user-friendly and can accept input data for prediction.
- For a user-friendly input i have used Flask which renders the HTML page and gets the user form input ie) clients behaviour, and recommends him/her the percentage of investment he should put , so that he gets good return back.

```
# Get user inputs from the terminal
knowledge_sharemarket = int(input("Enter Knowledge level about sharemarket(0-10): "))
knowledge_investment = int(input("Enter Knowledge level about different investment product (0-10): "))
awareness_source = int(input("Enter Source of Awareness about Investment (Television/Workers/Family/Magazine/Other): "))
household_income = int(input("Enter Household Income ('US$ 2736 to US$ 8205', 'US$ 19146 to US$ 24615', 'US$ 1367 to US$ 19145'): "))

# Make a prediction using the trained model
prediction = loaded_rf_clf.predict([[knowledge_sharemarket, knowledge_investment, awareness_source, household_income]])

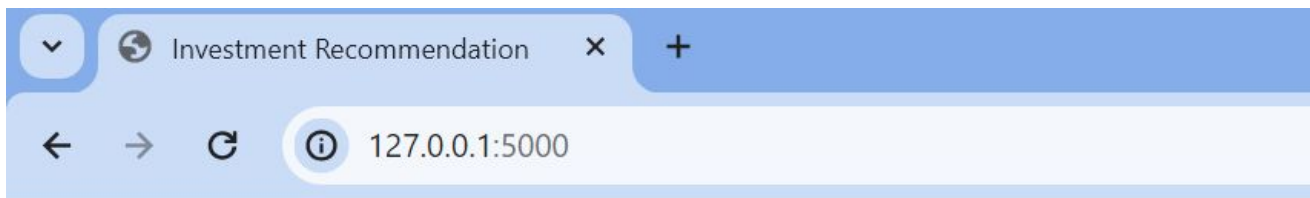
# Print the prediction
print("Prediction:", prediction)
print("My Recommendation is that you invest in range:", mapping[int(prediction)])
```

✓ 7.4s

Python

Prediction: [6]

My Recommendation is that you invest in range: Upto 5%



Investment Recommendation

Knowledge level about sharemarket (0-10):

Knowledge level about different investment product (0-10):

Source of Awareness about Investment:

Household Income:

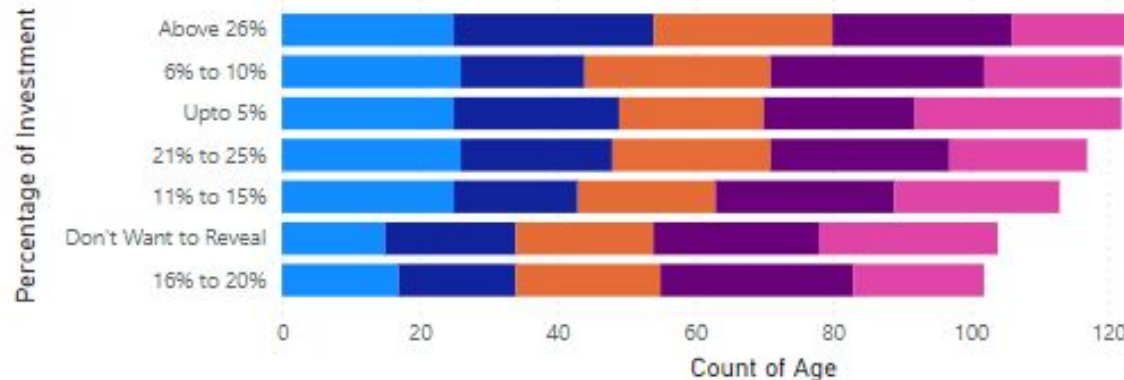
My Recommendation is that you invest in range: Above 26%

6)Analysis Using POWERBI

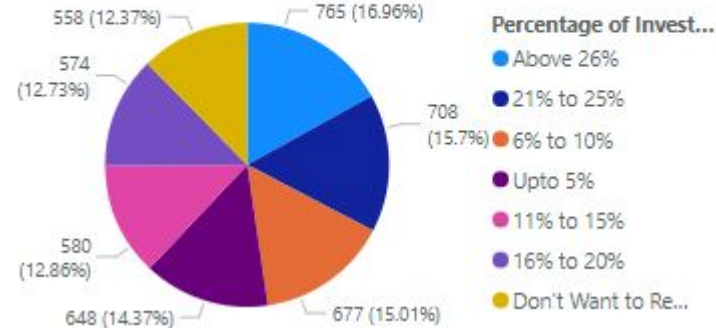
- Additionally i have used POWERBI for visualization and insights gathering.
- Following will be the visual elements i have performed.

Count of Age by Percentage of Investment and Age

Age ● Children ● Early Working ● Elderly ● Mature Working ● Prime Working



Sum of Knowledge level about different investment product by Percentage of Investment



11% to 15%

First Percentage of Investment

City

- ☒ Houston
- ☐ Memphis
- ☐ New York
- ☐ San Francisco
- ☐ Seattle

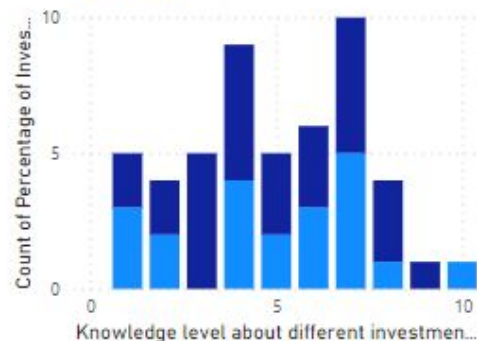
Age

- ☐ Children
- ☐ Early Working
- ☐ Elderly
- ☒ Mature Worki...
- ☐ Prime Working

INVESTMENT RECOMMENDATION SYSTEM

Count of Percentage of Investment by Knowledge level about different investment product and Gender

Gender ● Men ● Women



City and Percentage of Investment

Percentage of Inv... ● 11% to 1... ● 16% to 2... ● 21% to 2... ● 6% to 1...



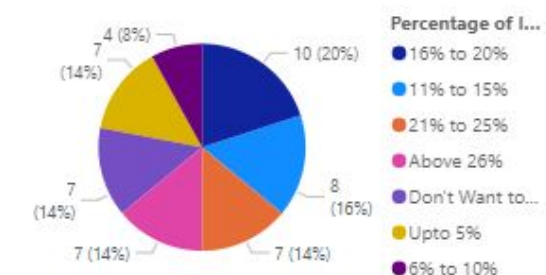
Count of Role by Household Income



Count of Percentage of Investment by Knowledge about Govt. Schemes



Count of Household Income by Percentage of Investment



11% to 15%

City

☐ Houston

First Percentage of Investment

☐ Memphis

☒ New York

☐ San Francisco

☐ Seattle

Age

☐ Children

☐ Early Working

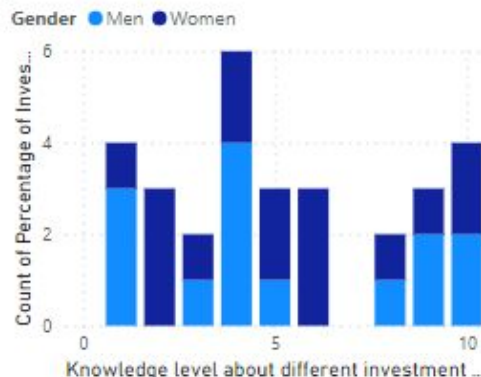
☐ Elderly

☐ Mature Worki...

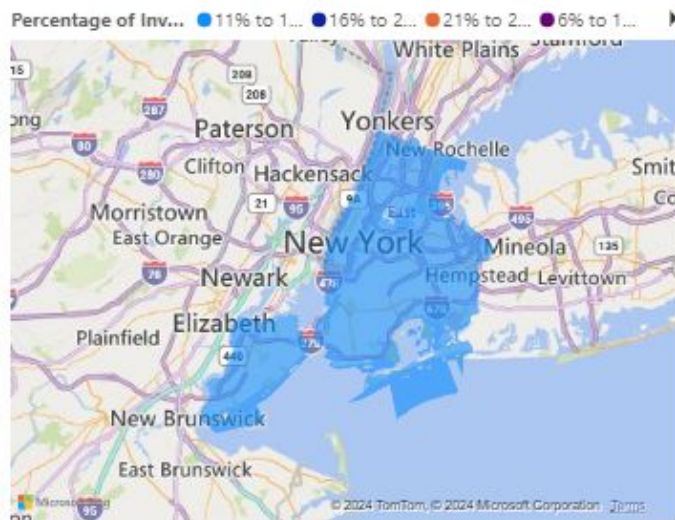
☒ Prime Working

INVESTMENT RECOMMENDATION SYSTEM

Count of Percentage of Investment by Knowledge level about different investment product and Gender



City and Percentage of Investment



Count of Role by Household Income



Count of Percentage of Investment by Knowledge about Govt. Schemes



Count of Household Income by Percentage of Investment

