Report Author: Khushi Sali

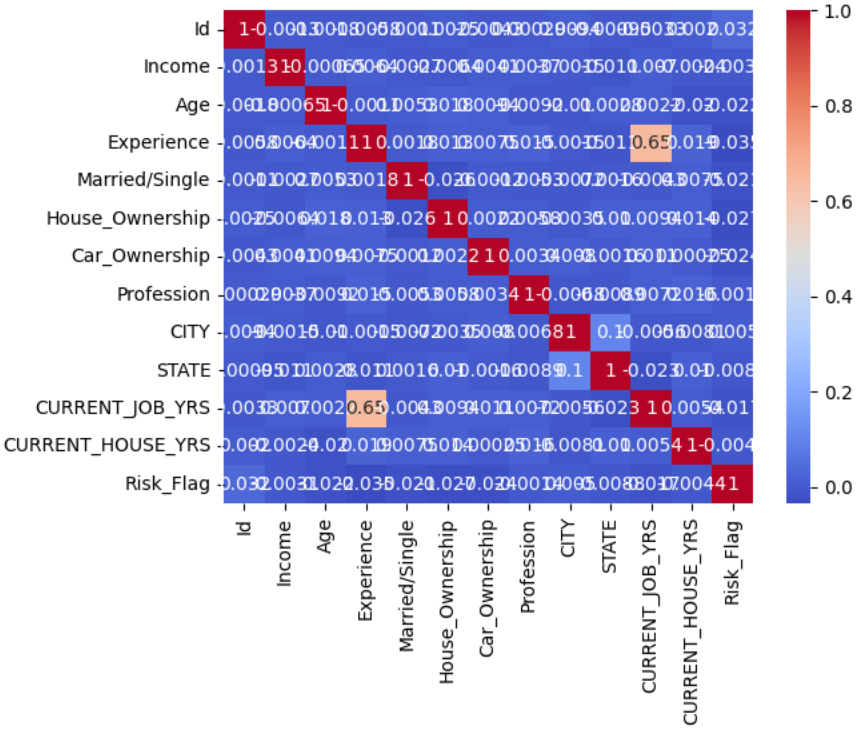
Report on

LOAN RISK PREDICTION

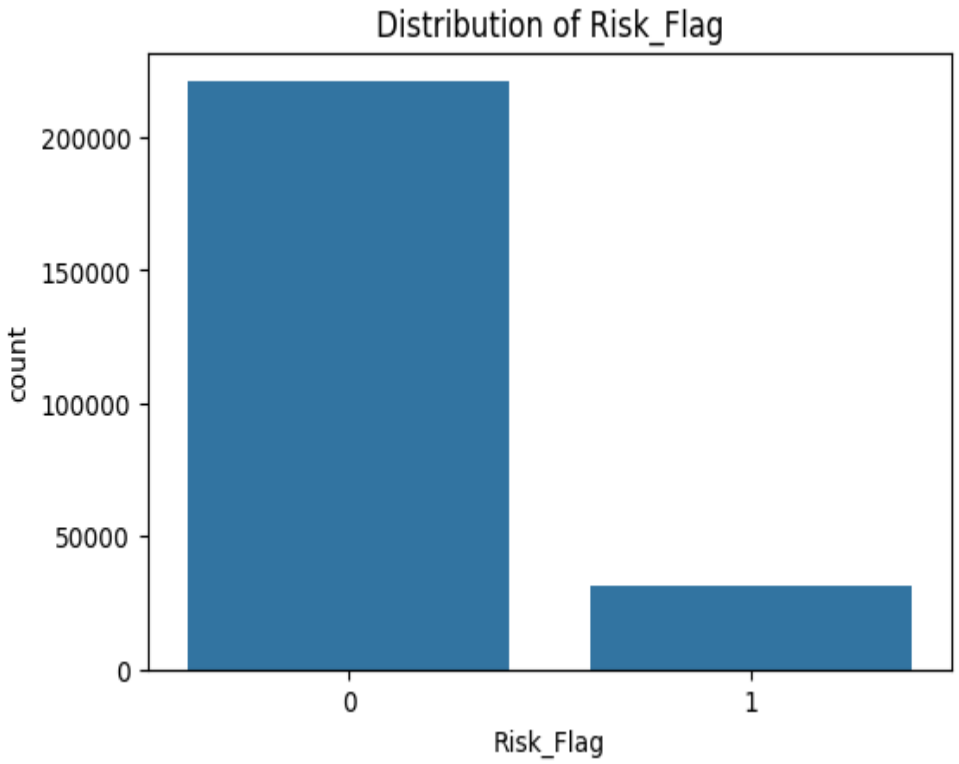
DATA VISUALIZATION

Different types of graphs have been generated to study how the features relate to each other and how they affect the deciding factor i.e if a person will be at a high risk or not.

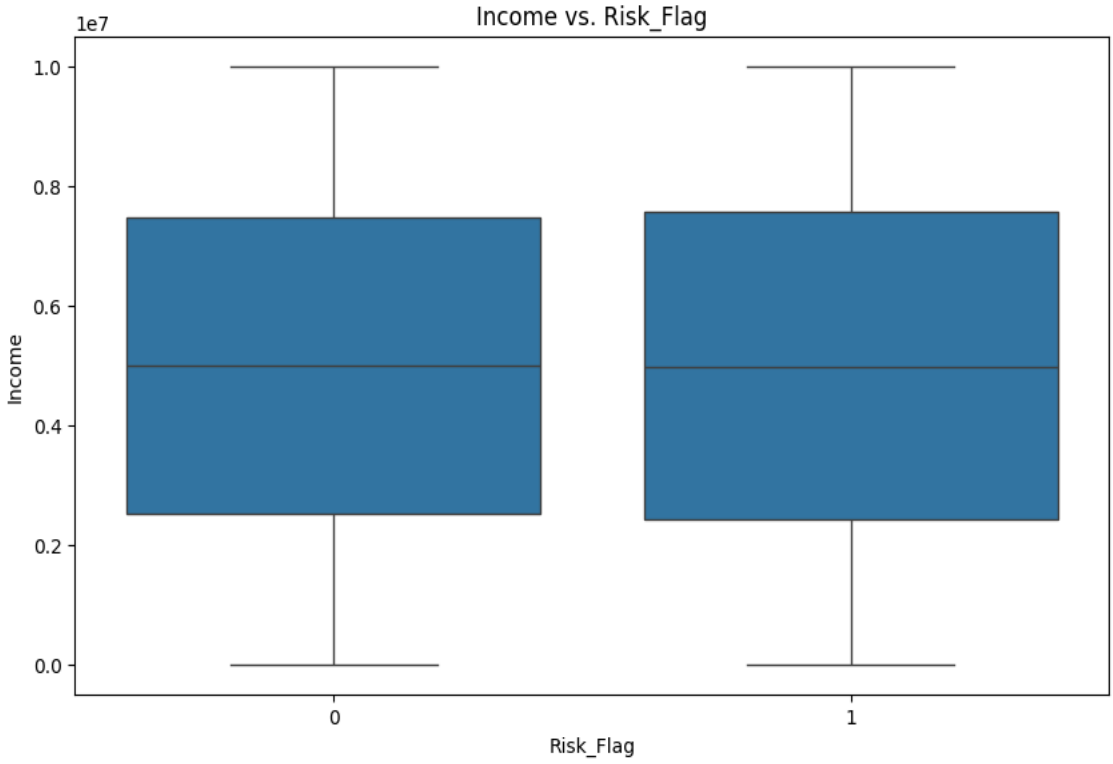
- Correlation Map



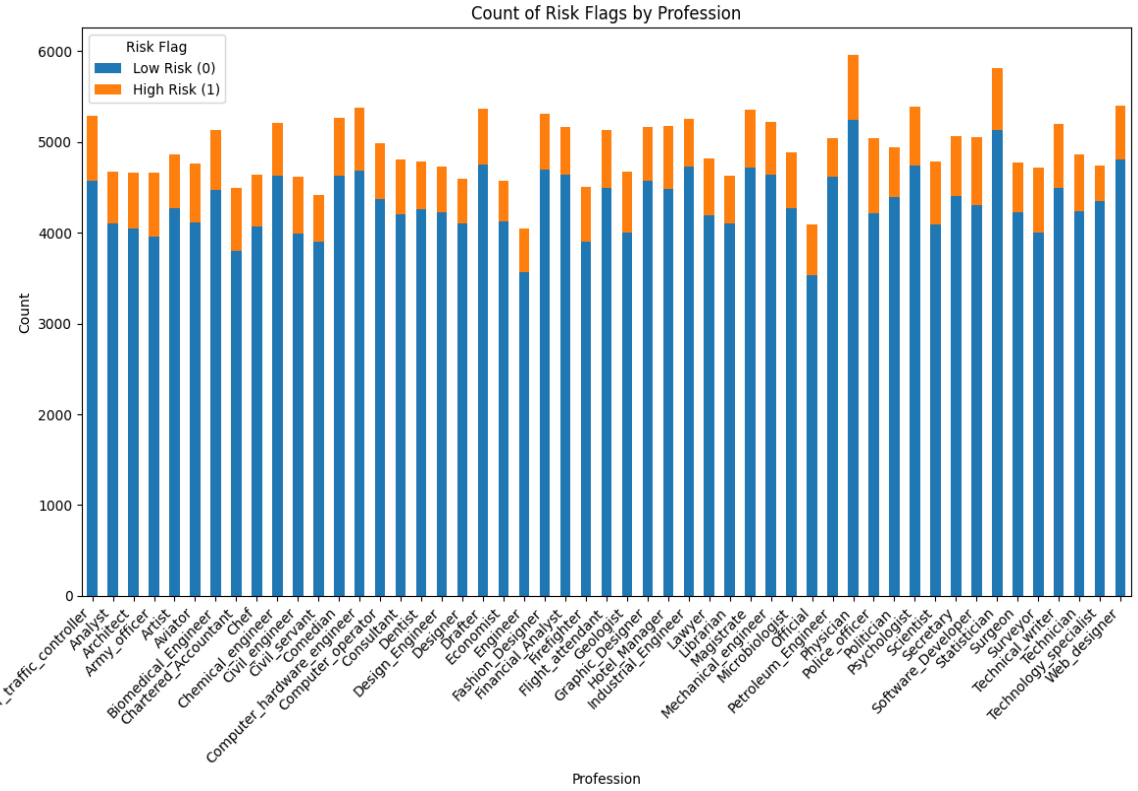
- Distribution of Risk Flag



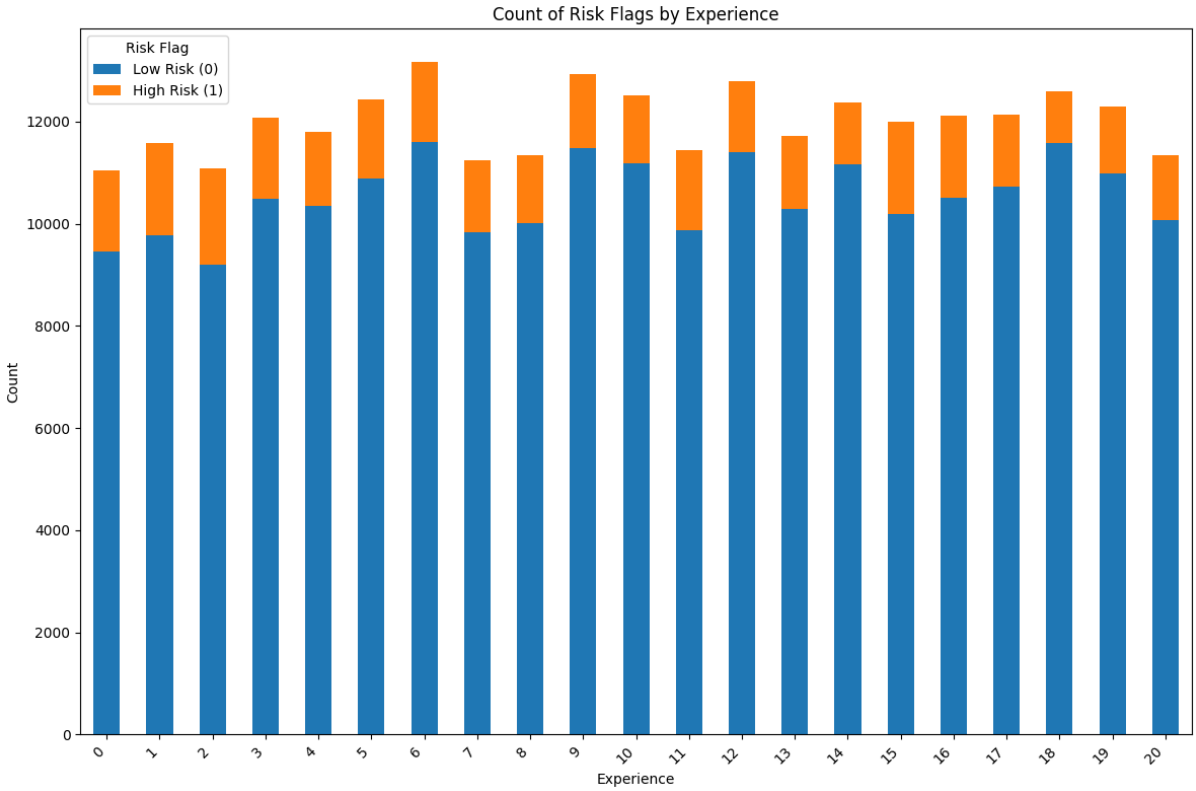
- Income VS Risk Flag



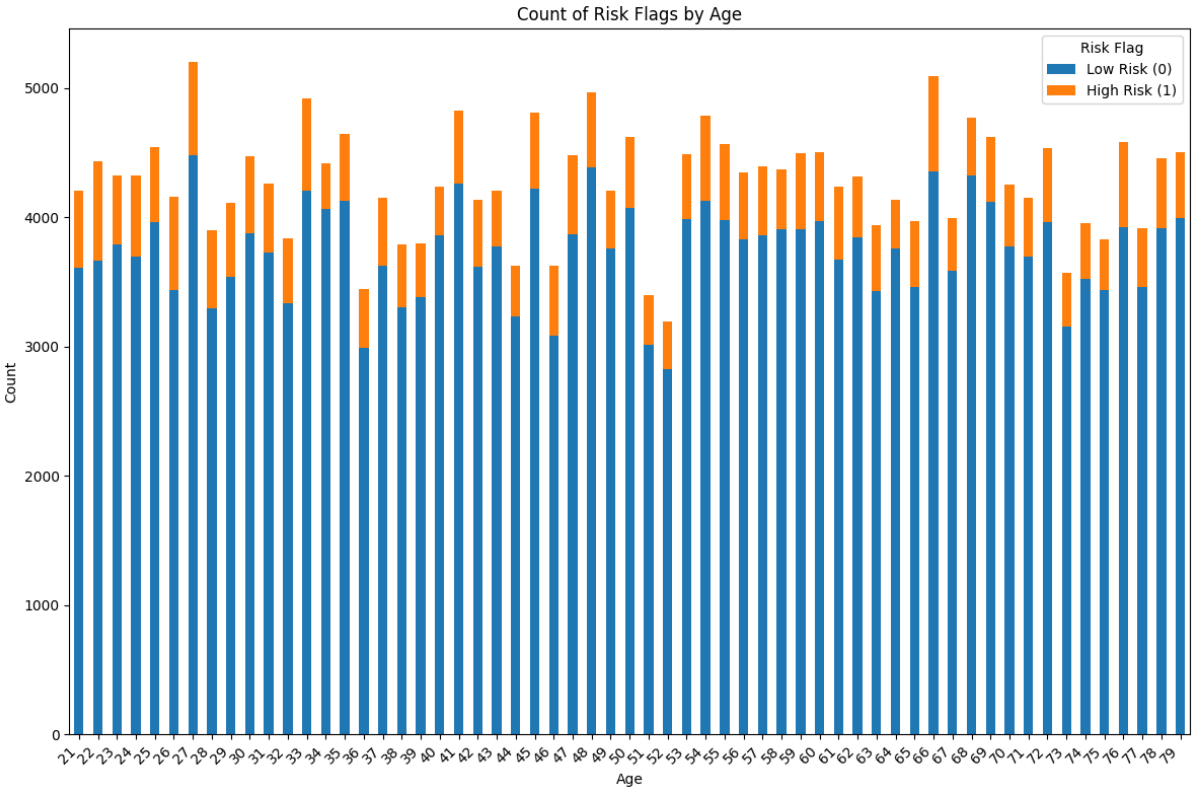
- Count of risk flag by profession

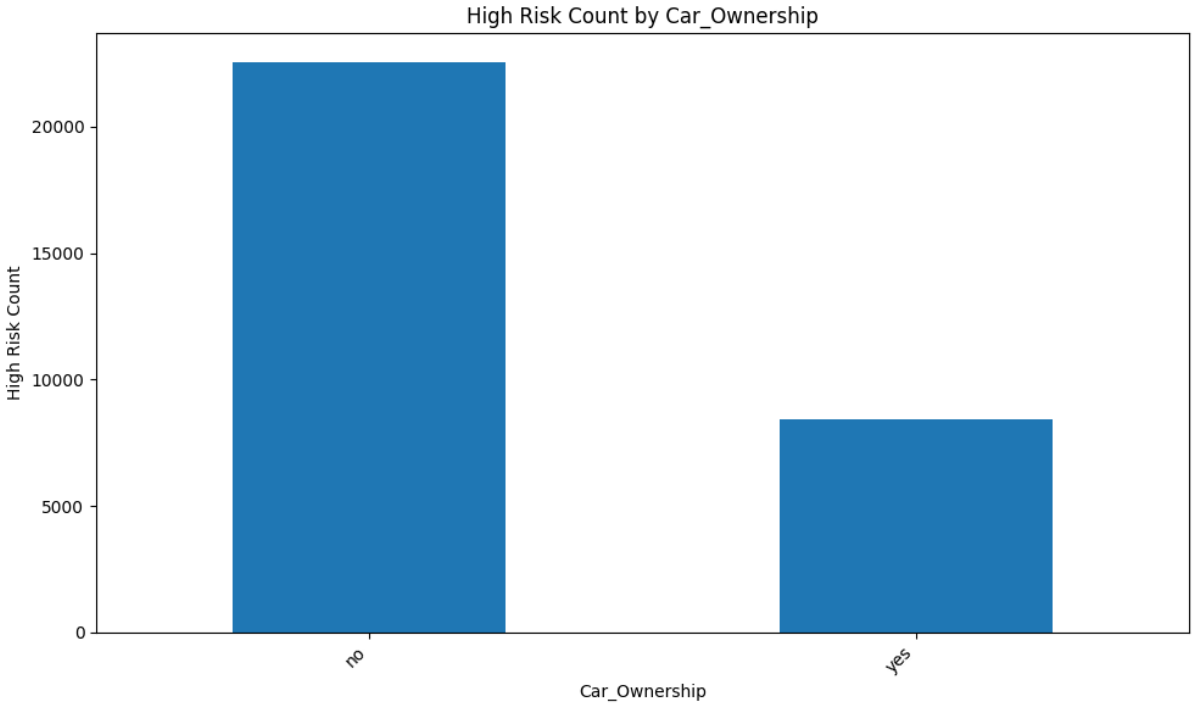


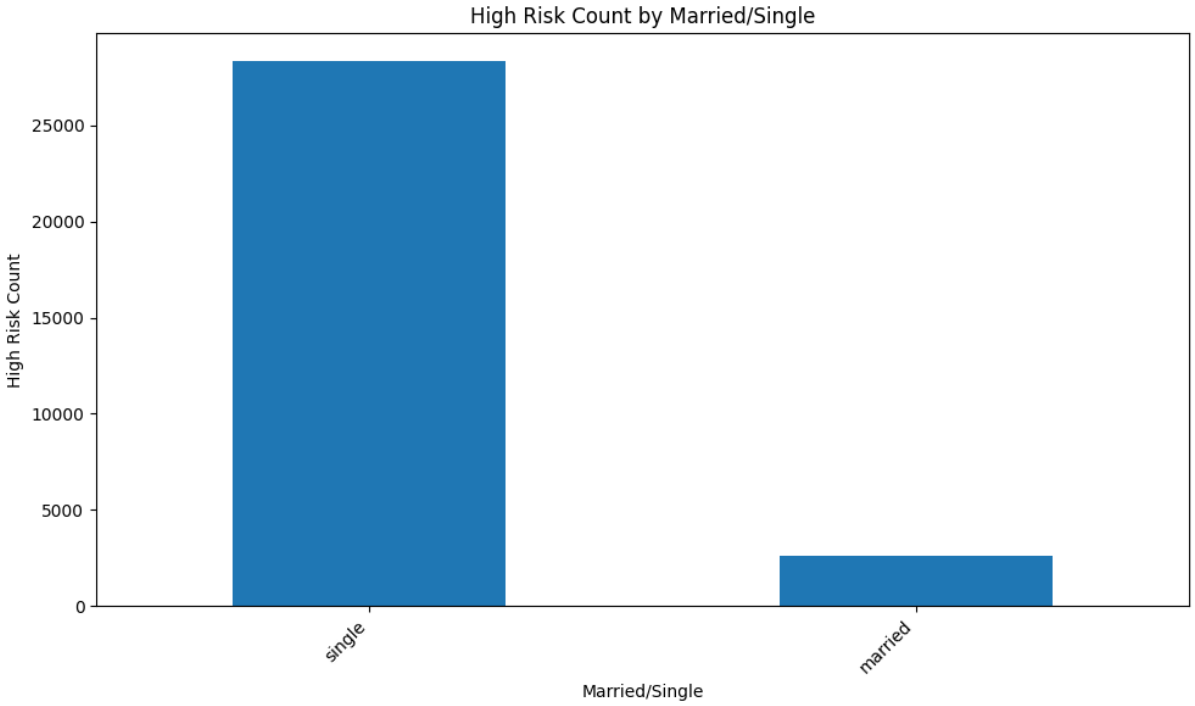
- Count of Risk Flags by Experience

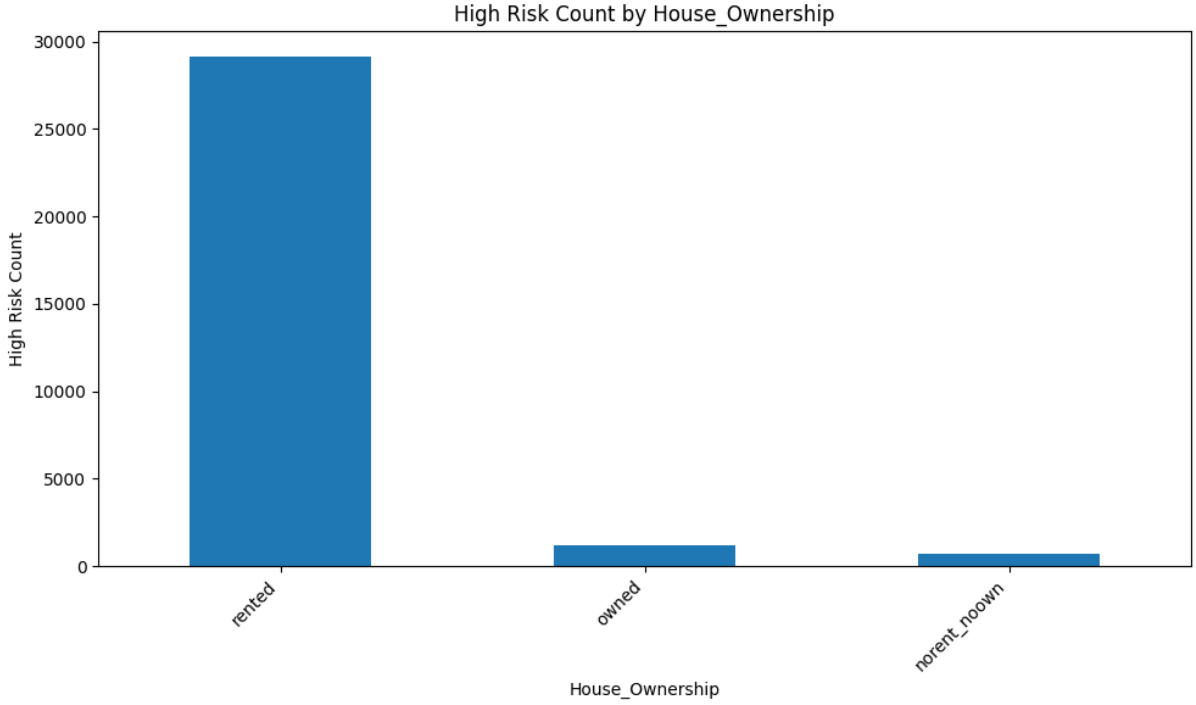


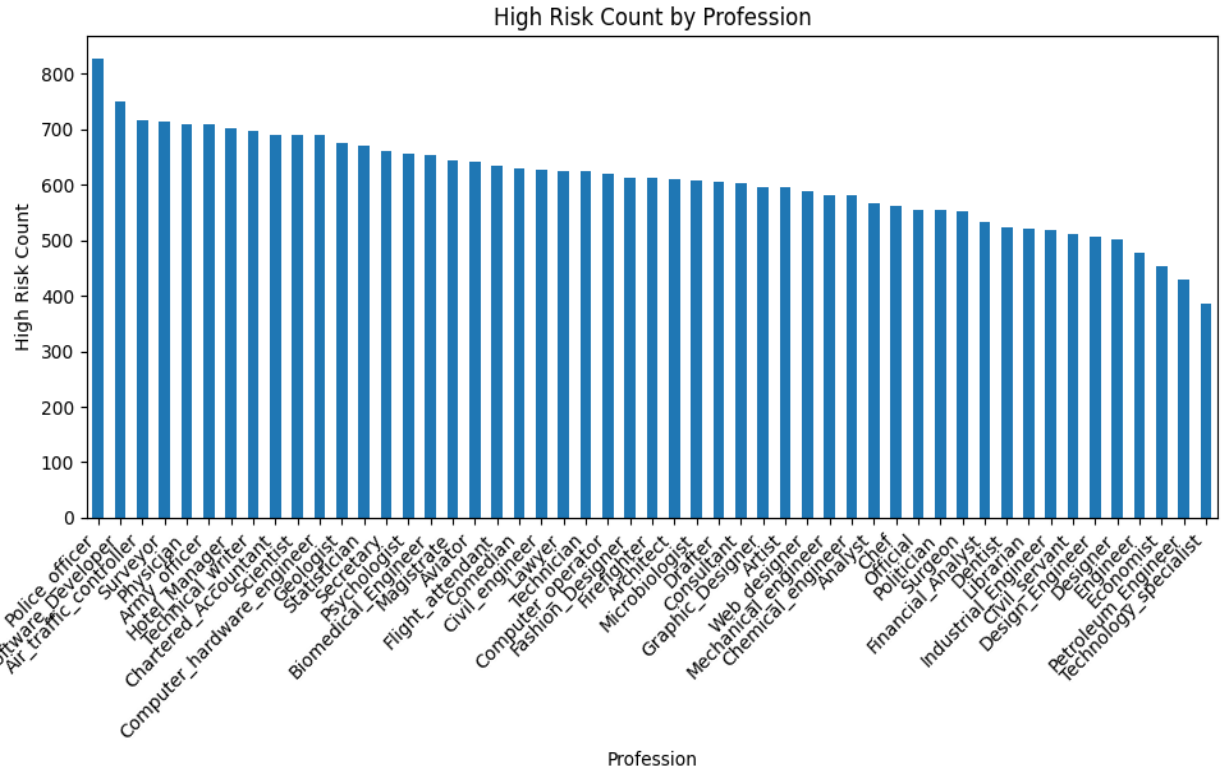
- Count of Risks Flags by Age






High Risk Count by Married/Single
High Risk Count
25000
20000
15000
10000
5000
0
single
married
Married/Single


High Risk Count by House_Ownership
High Risk Count
30000
25000
20000
15000
10000
5000
0
rented
owned
norent_noown
House_Ownership



DATA EXPLORATION INSIGHTS

Numerical values have been printed to understand how many people turn out to be at high risk according to the category they fall in.

```
Counts for Married/Single:
                Low Risk (0)  High Risk (1)
Married/Single
married                23092           2636
single                197912          28360
```

```
Counts for House_Ownership:
                 Low Risk (0)  High Risk (1)
House_Ownership
norent_noown             6469            715
owned                   11758           1160
rented                 202777          29121
```

```
Counts for Car_Ownership:
               Low Risk (0)  High Risk (1)
Car_Ownership
no                   153439          22561
yes                   67565           8435
```

```
Counts for Profession:
                           Low Risk (0)  High Risk (1)
Profession
Air_traffic_controller             4566            715
Analyst                            4101            567
Architect                          4046            611
Army_officer                       3952            709
Artist                             4265            596
Aviator                            4116            642
Biomedical_Engineer                4473            654
Chartered_Accountant               3803            690
Chef                               4072            563
Chemical_engineer                  4624            581
Civil_engineer                     3989            627
Civil_servant                      3902            511
Comedian                           4630            629
Computer_hardware_engineer         4682            690
Computer_operator                  4371            619
Consultant                         4206            602
Dentist                            4258            524
Design_Engineer                    4223            506
Designer                           4096            502
Drafter                            4754            605
Economist                          4119            454
Engineer                           3570            478
...
Technical_writer                   4498            697
Technician                         4240            624
Technology_specialist              4351            386
Web_designer                       4808            589
```

```
High Risk Counts for Married/Single:
 Married/Single
single     28360
married     2636
```

```
High Risk Counts for House_Ownership:
 House_Ownership
rented          29121
owned            1160
norent_noown      715
```

```
High Risk Counts for Car_Ownership:
 Car_Ownership
no     22561
yes     8435
```

MODEL PERFORMANCE

After training multiple machine learning model, it was found that Random Forest gives the most satisfying results.

	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting
Accuracy	0.877368	0.844299	0.906931	0.877526
Precision	0.438684	0.645382	0.822198	0.876263
Recall	0.500000	0.655037	0.684096	0.500740

AdaBoost	Bagging	SVM
0.877368	0.887844	0.877368
0.438684	0.786055	0.438684
0.500000	0.575978	0.500000

K-Nearest Neighbors	Naive Bayes	Multilayer Perceptron
0.859907	0.514802	0.126508
0.551776	0.507417	0.472137
0.514685	0.517227	0.498451

UNDERSTAND MAIN DECIDING FACTORS ASSOCIATED WITH RISK

After careful analysis of the data, it can be concluded that the features ‘Id’, ‘State’, ‘City’ can be dropped from the dataframe and the rest of the features can be used to train the Machine Learning model. After training multiple machine learning model, it was found that Random Forest gives the most satisfying results.