Midterm Project Loan Approval Prediction

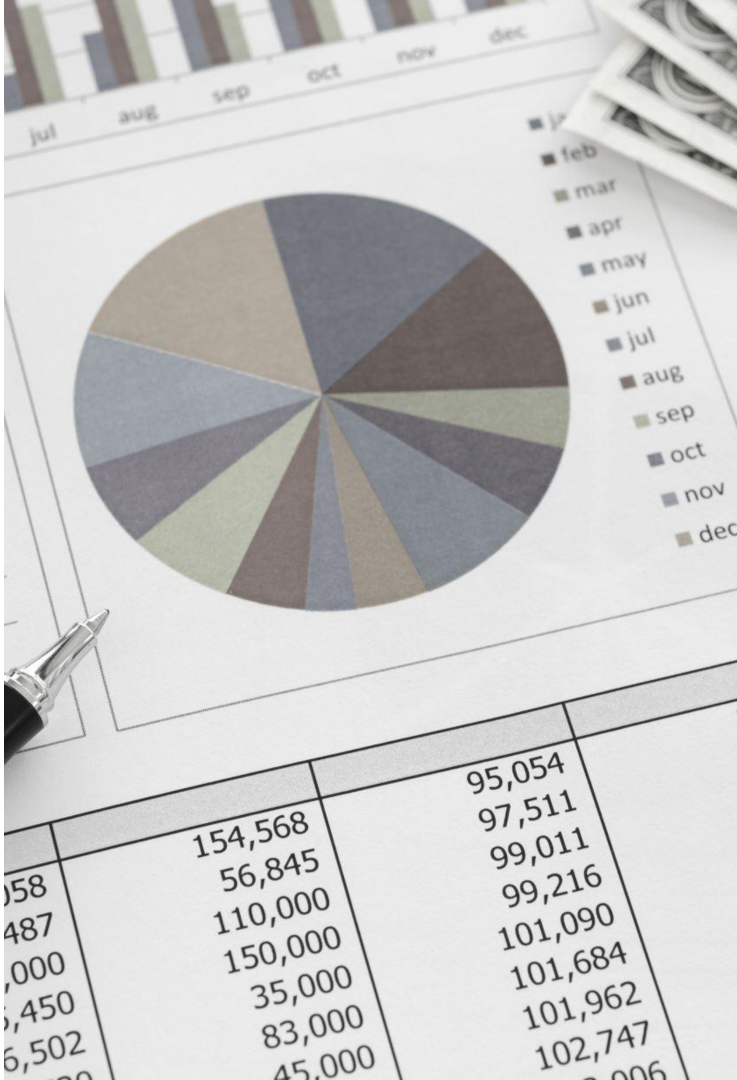
by
Ethan Lam
Cynthia Okaja
Yuchen He

11th August 2023

LIGHTHOUSE LABS

Project Goals

To develop a model that will predict the likelihood of loan approval based on customer details such as gender, marital status, education, number of dependents, income, loan amount, credit history and others.



Project Execution

- Data Collection
- Data Preprocessing
- EDA
- Data Visualization
- Hypothesis Testing
- Modelling

Getting Started

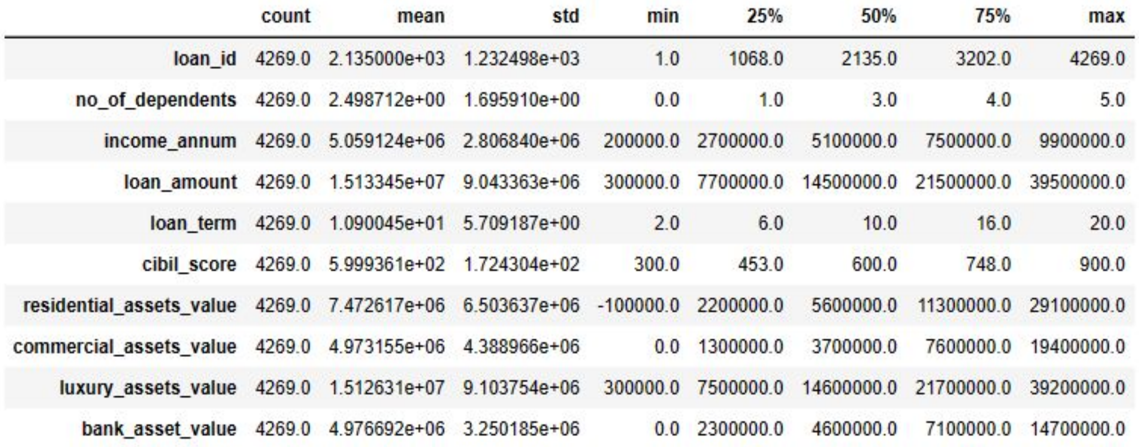
Data Collection – Kaggle

https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset

Data preprocessing

Check data info, null values, data types and column names

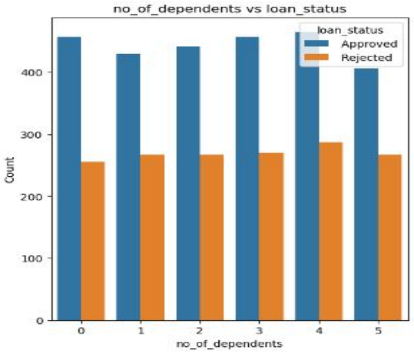
Describe ()

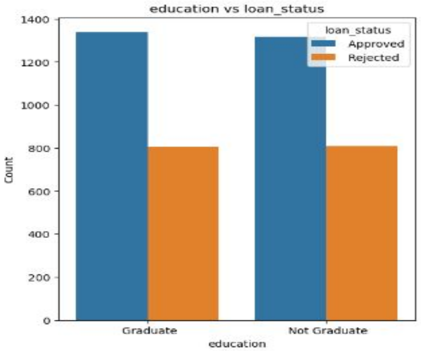


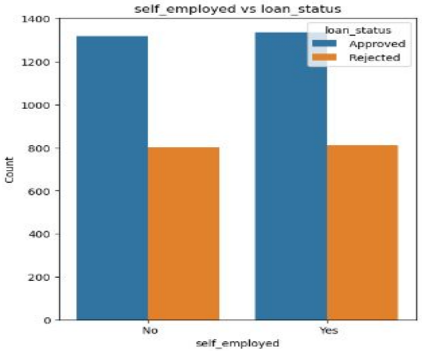
	count	mean	std	min	25%	50%	75%	max
loan_id	4269.0	2.135000e+03	1.232498e+03	1.0	1068.0	2135.0	3202.0	4269.0
no_of_dependents	4269.0	2.498712e+00	1.695910e+00	0.0	1.0	3.0	4.0	5.0
income_annum	4269.0	5.059124e+06	2.806840e+06	200000.0	2700000.0	5100000.0	7500000.0	9900000.0
loan_amount	4269.0	1.513345e+07	9.043363e+06	300000.0	7700000.0	14500000.0	21500000.0	39500000.0
loan_term	4269.0	1.090045e+01	5.709187e+00	2.0	6.0	10.0	16.0	20.0
cibil_score	4269.0	5.999361e+02	1.724304e+02	300.0	453.0	600.0	748.0	900.0
residential_assets_value	4269.0	7.472617e+06	6.503637e+06	-100000.0	2200000.0	5600000.0	11300000.0	29100000.0
commercial_assets_value	4269.0	4.973155e+06	4.388966e+06	0.0	1300000.0	3700000.0	7600000.0	19400000.0
luxury_assets_value	4269.0	1.512631e+07	9.103754e+06	300000.0	7500000.0	14600000.0	21700000.0	39200000.0
bank_asset_value	4269.0	4.976692e+06	3.250185e+06	0.0	2300000.0	4600000.0	7100000.0	14700000.0

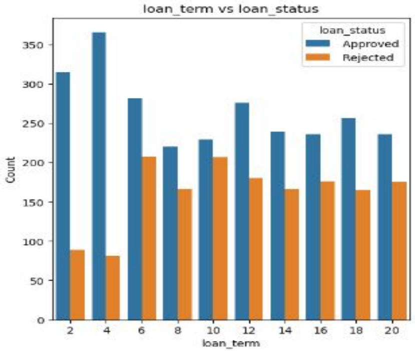
EDA

Shape - 4269 -rows, 13 - columns

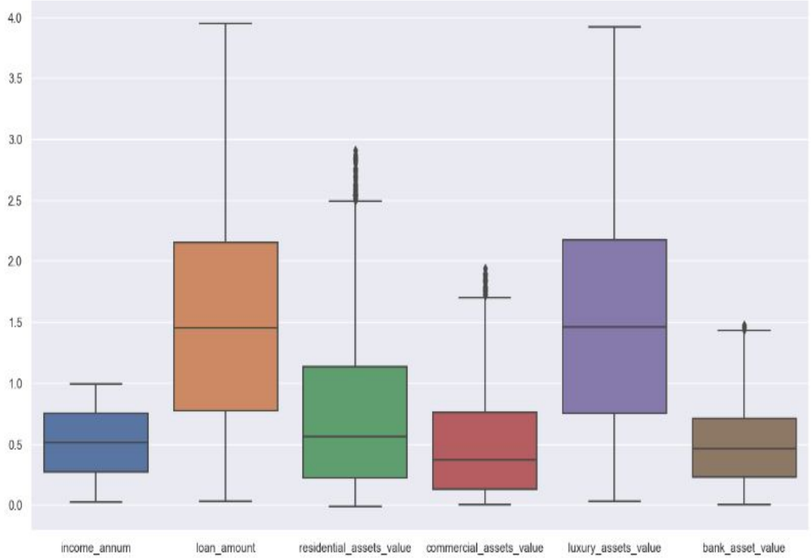




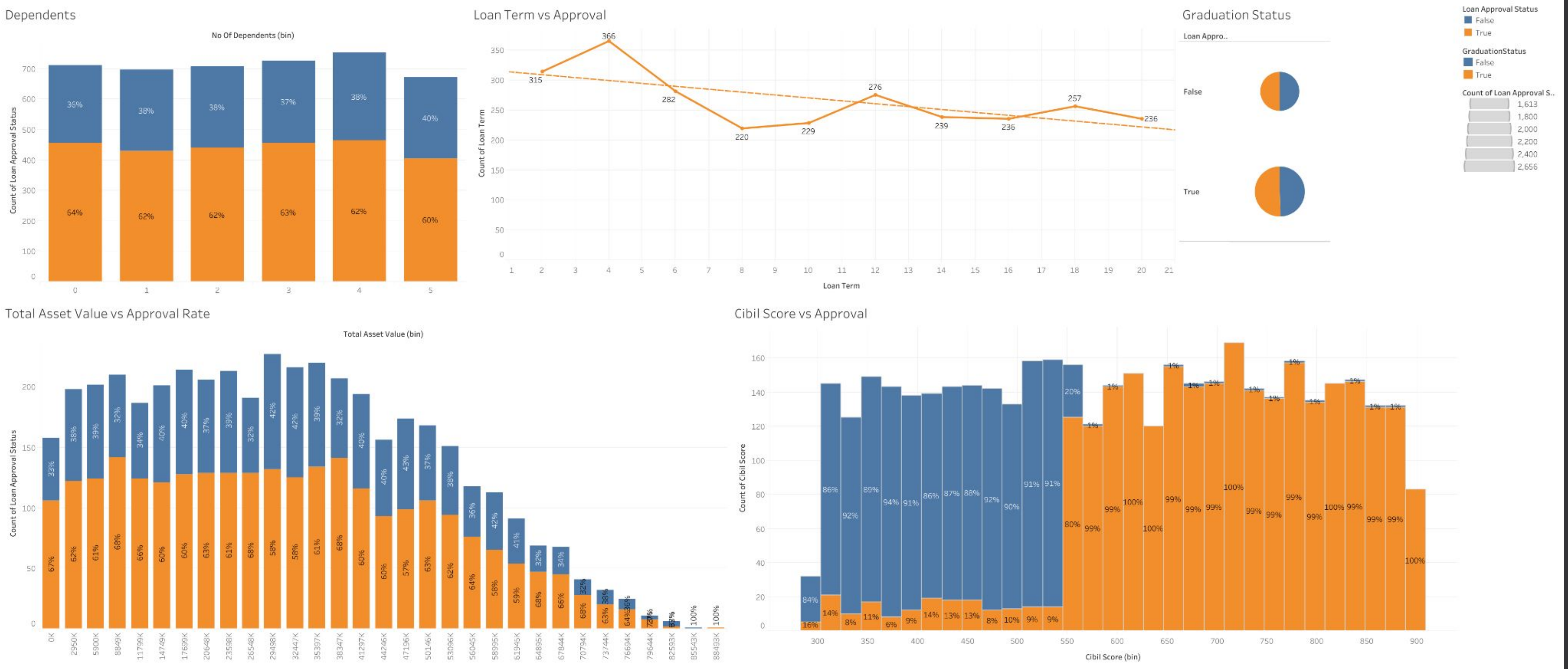




Identifying outliers



Visualization using Tableau



Dependents

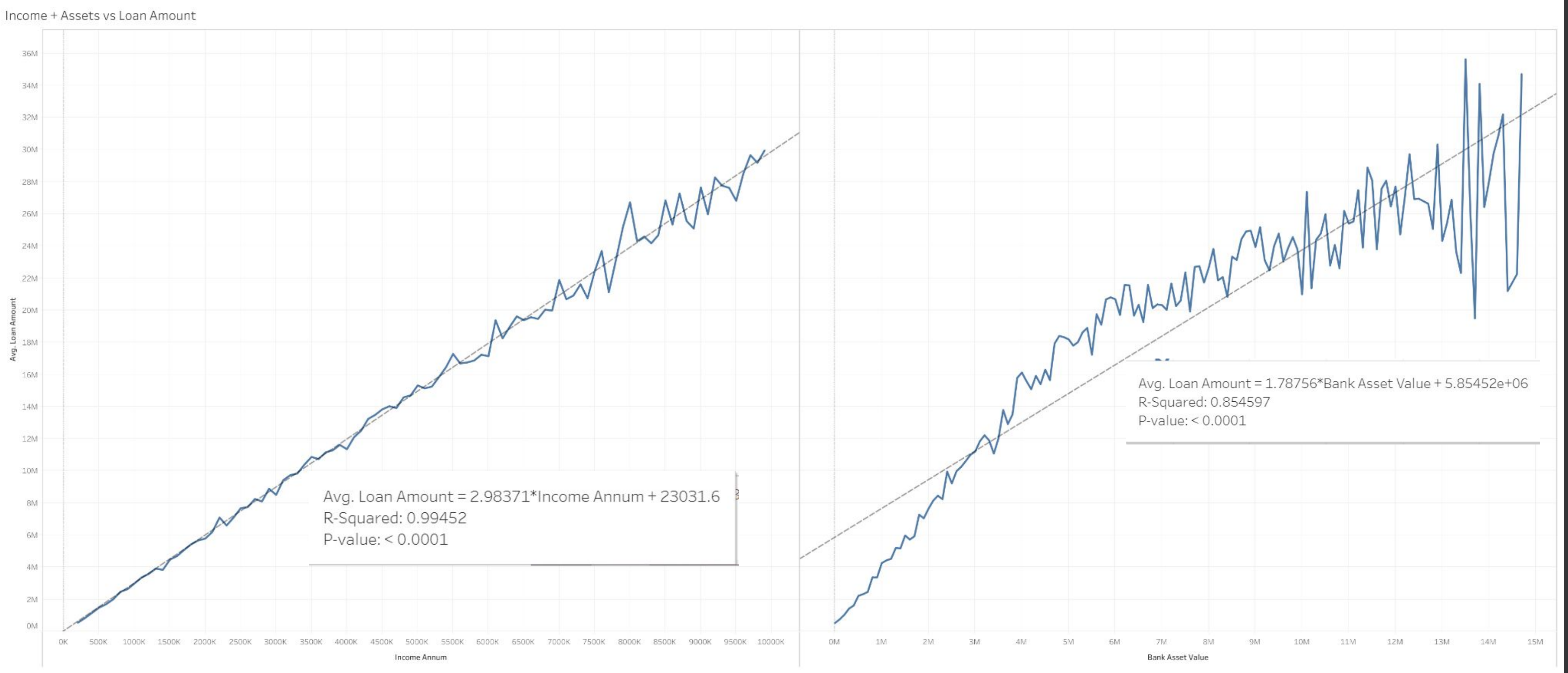
Loan Term vs Approval

Graduation Status

Total Asset Value vs Approval Rate

Cibil Score vs Approval

Visualization using Tableau



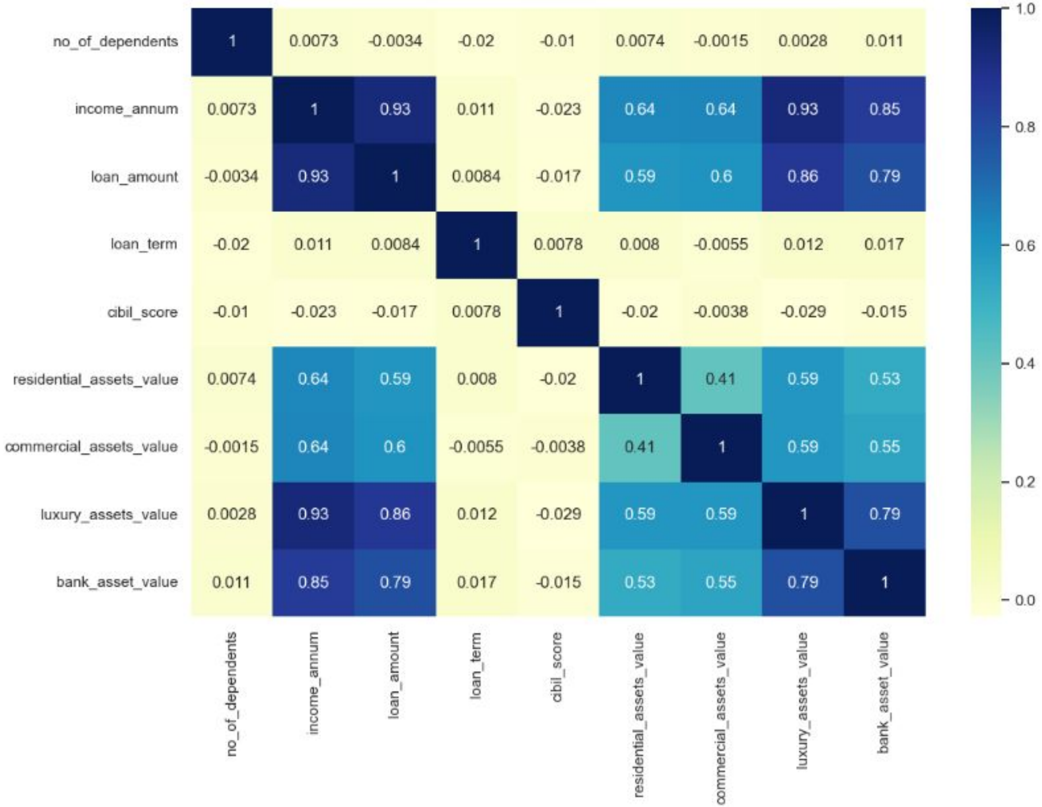
Income + Assets vs Loan Amount

Avg. Loan Amount = 2.98371*Income Annum + 23031.6
R-Squared: 0.99452
P-value: < 0.0001

Avg. Loan Amount = 1.78756*Bank Asset Value + 5.85452e+06
R-Squared: 0.854597
P-value: < 0.0001

EDA

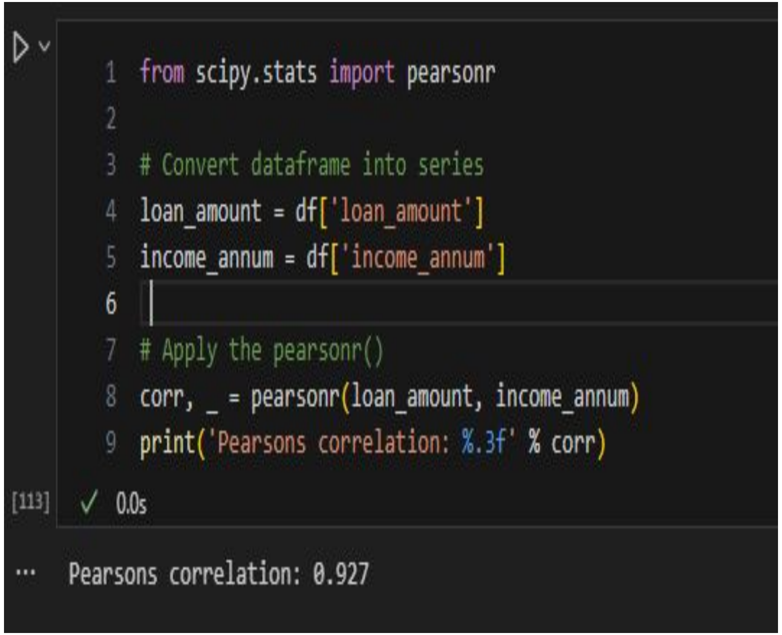
Using heat map to check the correlation between features



	no_of_dependents	income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value
no_of_dependents	1	0.0073	-0.0034	-0.02	-0.01	0.0074	-0.0015	0.0028	0.011
income_annum	0.0073	1	0.93	0.011	-0.023	0.64	0.64	0.93	0.85
loan_amount	-0.0034	0.93	1	0.0084	-0.017	0.59	0.6	0.86	0.79
loan_term	-0.02	0.011	0.0084	1	0.0078	0.008	-0.0055	0.012	0.017
cibil_score	-0.01	-0.023	-0.017	0.0078	1	-0.02	-0.0038	-0.029	-0.015
residential_assets_value	0.0074	0.64	0.59	0.008	-0.02	1	0.41	0.59	0.53
commercial_assets_value	-0.0015	0.64	0.6	-0.0055	-0.0038	0.41	1	0.59	0.55
luxury_assets_value	0.0028	0.93	0.86	0.012	-0.029	0.59	0.59	1	0.79
bank_asset_value	0.011	0.85	0.79	0.017	-0.015	0.53	0.55	0.79	1

Correlation test

Question: does annual income has correlation with loan amount?

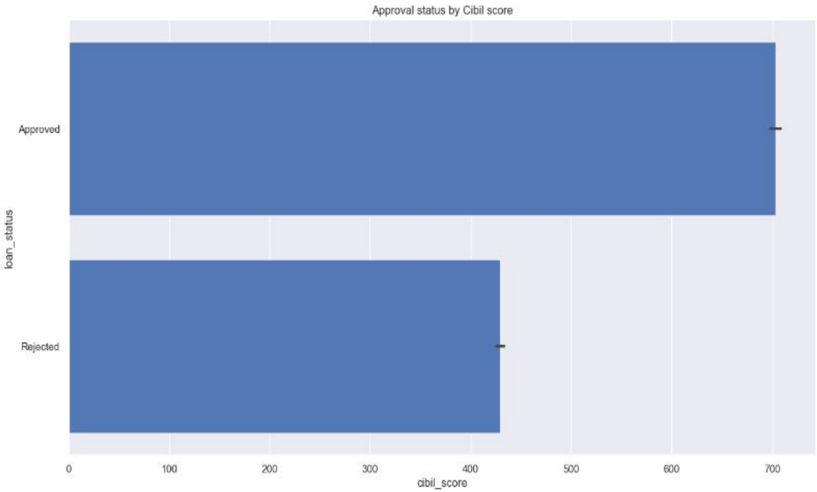


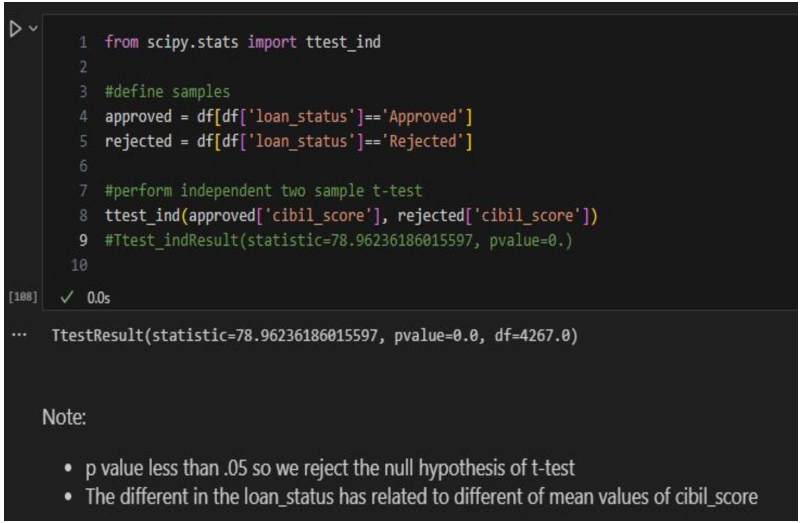
```python
from scipy.stats import pearsonr

# Convert dataframe into series
loan_amount = df['loan_amount']
income_annum = df['income_annum']

# Apply the pearsonr()
corr, _ = pearsonr(loan_amount, income_annum)
print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.927
```

Hypothesis Testing

- T-Test





```python
from scipy.stats import ttest_ind

#define samples
approved = df[df['loan_status']=='Approved']
rejected = df[df['loan_status']=='Rejected']

#perform independent two sample t-test
ttest_ind(approved['cibil_score'], rejected['cibil_score'])
#Ttest_indResult(statistic=78.96236186015597, pvalue=0.)

```

[108] ✓ 0.0s

```
TtestResult(statistic=78.96236186015597, pvalue=0.0, df=4267.0)
```

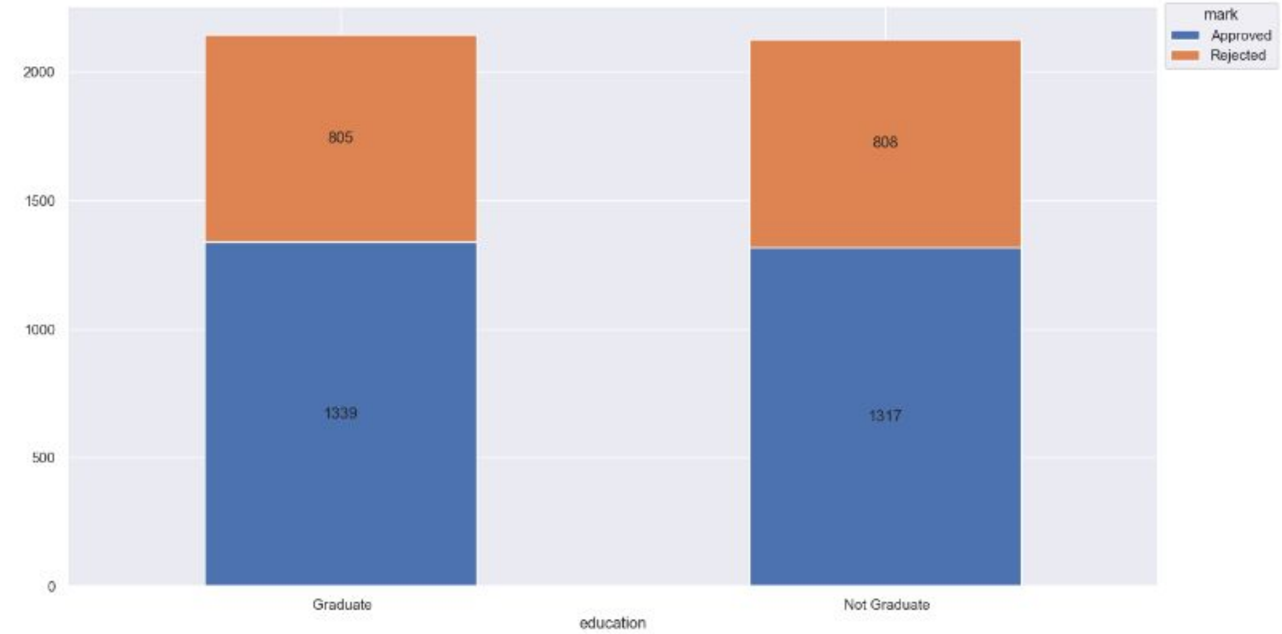
Note:

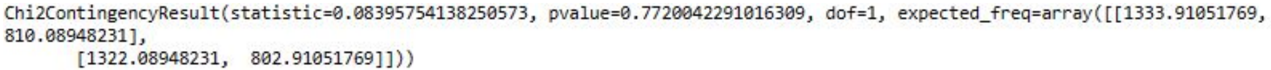
- p value less than .05 so we reject the null hypothesis of t-test
- The different in the loan_status has related to different of mean values of cibil_score

Cibil Score has statistic significantly impact to the loan_status

Hypothesis Testing

- Chi-Squared Test: testing the relationship between two categorical variables



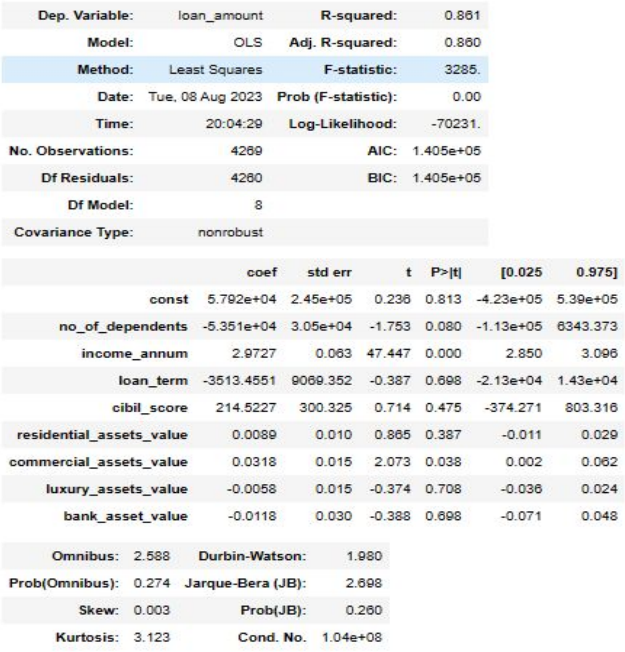


```
Chi2ContingencyResult(statistic=0.08395754138250573, pvalue=0.7720042291016309, dof=1, expected_freq=array([[1333.91051769,
810.08948231],
       [1322.08948231,  802.91051769]]))
```

Since the pvalue is 0.77 so we can't reject the Null Hypothesis which means that there is a relationship between education and loan_status

Modelling

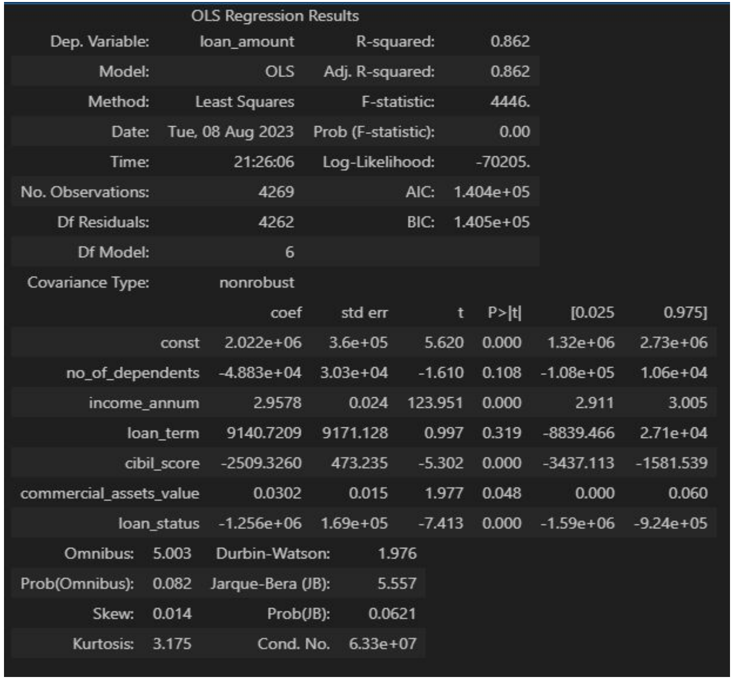
Multiple Linear Regression – Backward Selection



Dep. Variable:	loan_amount	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.860
Method:	Least Squares	F-statistic:	3285.
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	20:04:29	Log-Likelihood:	-70231.
No. Observations:	4269	AIC:	1.405e+05
Df Residuals:	4260	BIC:	1.405e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.792e+04	2.45e+05	0.236	0.813	-4.23e+05	5.39e+05
no_of_dependents	-5.351e+04	3.05e+04	-1.753	0.080	-1.13e+05	6343.373
income_annum	2.9727	0.063	47.447	0.000	2.850	3.096
loan_term	-3513.4551	9069.352	-0.387	0.698	-2.13e+04	1.43e+04
cibil_score	214.5227	300.325	0.714	0.475	-374.271	803.316
residential_assets_value	0.0089	0.010	0.865	0.387	-0.011	0.029
commercial_assets_value	0.0318	0.015	2.073	0.038	0.002	0.062
luxury_assets_value	-0.0058	0.015	-0.374	0.708	-0.036	0.024
bank_asset_value	-0.0118	0.030	-0.388	0.698	-0.071	0.048

Omnibus:	2.588	Durbin-Watson:	1.980
Prob(Omnibus):	0.274	Jarque-Bera (JB):	2.698
Skew:	0.003	Prob(JB):	0.260
Kurtosis:	3.123	Cond. No.	1.04e+08



OLS Regression Results

Dep. Variable:	loan_amount	R-squared:	0.862
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	4446.
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	21:26:06	Log-Likelihood:	-70205.
No. Observations:	4269	AIC:	1.404e+05
Df Residuals:	4262	BIC:	1.405e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.022e+06	3.6e+05	5.620	0.000	1.32e+06	2.73e+06
no_of_dependents	-4.883e+04	3.03e+04	-1.610	0.108	-1.08e+05	1.06e+04
income_annum	2.9578	0.024	123.951	0.000	2.911	3.005
loan_term	9140.7209	9171.128	0.997	0.319	-8839.466	2.71e+04
cibil_score	-2509.3260	473.235	-5.302	0.000	-3437.113	-1581.539
commercial_assets_value	0.0302	0.015	1.977	0.048	0.000	0.060
loan_status	-1.256e+06	1.69e+05	-7.413	0.000	-1.59e+06	-9.24e+05

Omnibus:	5.003	Durbin-Watson:	1.976
Prob(Omnibus):	0.082	Jarque-Bera (JB):	5.557
Skew:	0.014	Prob(JB):	0.0621
Kurtosis:	3.175	Cond. No.	6.33e+07

Model accuracy - 86%

Interpretation

R-squared and Adjusted R-squared values are close to 0.860, 86%

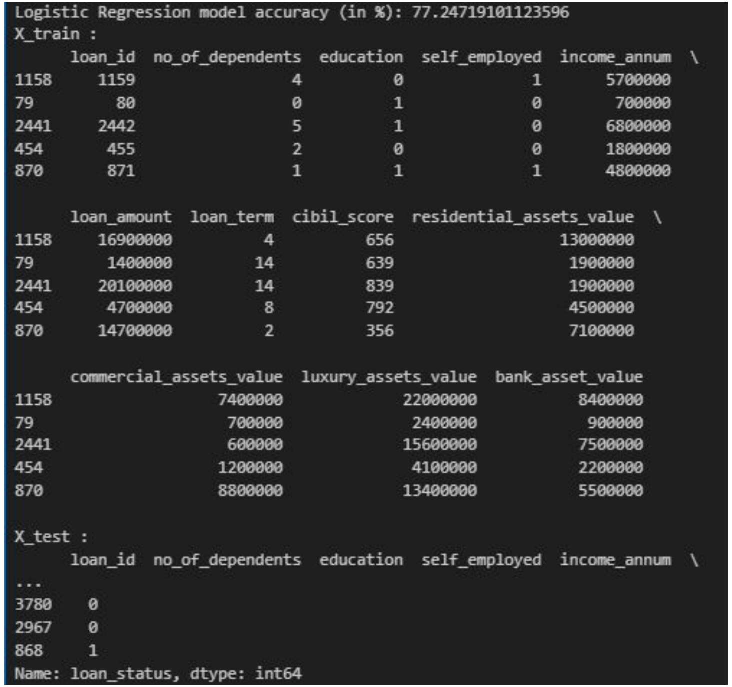
R-squared values in last model with fewer variables might be preferred due to simplicity and fewer non-significant variables.

'income_annum' appears to be statistically significant and positively associated with 'loan_amount’

Both models have similar R-squared values, the last model with fewer variables might be preferred due to simplicity and fewer non-significant variables.

Modelling

Logistics Regression for loan_status



```
Logistic Regression model accuracy (in %): 77.24719101123596
X_train :
      loan_id  no_of_dependents  education  self_employed  income_annum  \
1158     1159                 4          0              1       5700000
79         80                 0          1              0        700000
2441     2442                 5          1              0       6800000
454       455                 2          0              0       1800000
870       871                 1          1              1       4800000

      loan_amount  loan_term  cibil_score  residential_assets_value  \
1158     16900000          4          656                  13000000
79        1400000         14          639                   1900000
2441     20100000         14          839                   1900000
454       4700000          8          792                   4500000
870      14700000          2          356                   7100000

      commercial_assets_value  luxury_assets_value  bank_asset_value
1158                  7400000             22000000           8400000
79                     700000              2400000            900000
2441                   600000             15600000           7500000
454                   1200000              4100000           2200000
870                   8800000             13400000           5500000

X_test :
      loan_id  no_of_dependents  education  self_employed  income_annum  \
...
3780    0
2967    0
868     1
Name: loan_status, dtype: int64
```

Model accuracy 77%

Conclusion

- Produced two models with 86% and 77% accuracy of predicting loan approval status
- Loan term, cibil score have major effects on the approval

Challenges & Future Goals

Challenges

- Trying to optimize models
  - Selecting the “perfect” combination of features
- Using advanced features of Tableau
  - Forecasting: not available since the dataset does not contain temporal data
  - Casting data to correct data types

Future Goals

- Stretch: determining whether the loan can be paid back in time
- Getting a more detailed dataset

Thank you!