Loan Application

Classify applicants : Approval | Rejection

Work Flow

- Hypothesis Generation
- Data Exploration
- Data Cleaning
- Modelling
- Deployment

Hypothesis generation

Problem Statement

"Predict whether a loan application gets approved or rejected based on the information provided by applicant"

Hypothesis

- Good Credit standing
- Good Salary
- Assets
- Less debt
- Low loan amount
- Old customers to bank
- less dependents
- Educated
- Provide all details to the bank

Hypothesis

Good Credit standing

Good Salary

Assets

Less debt

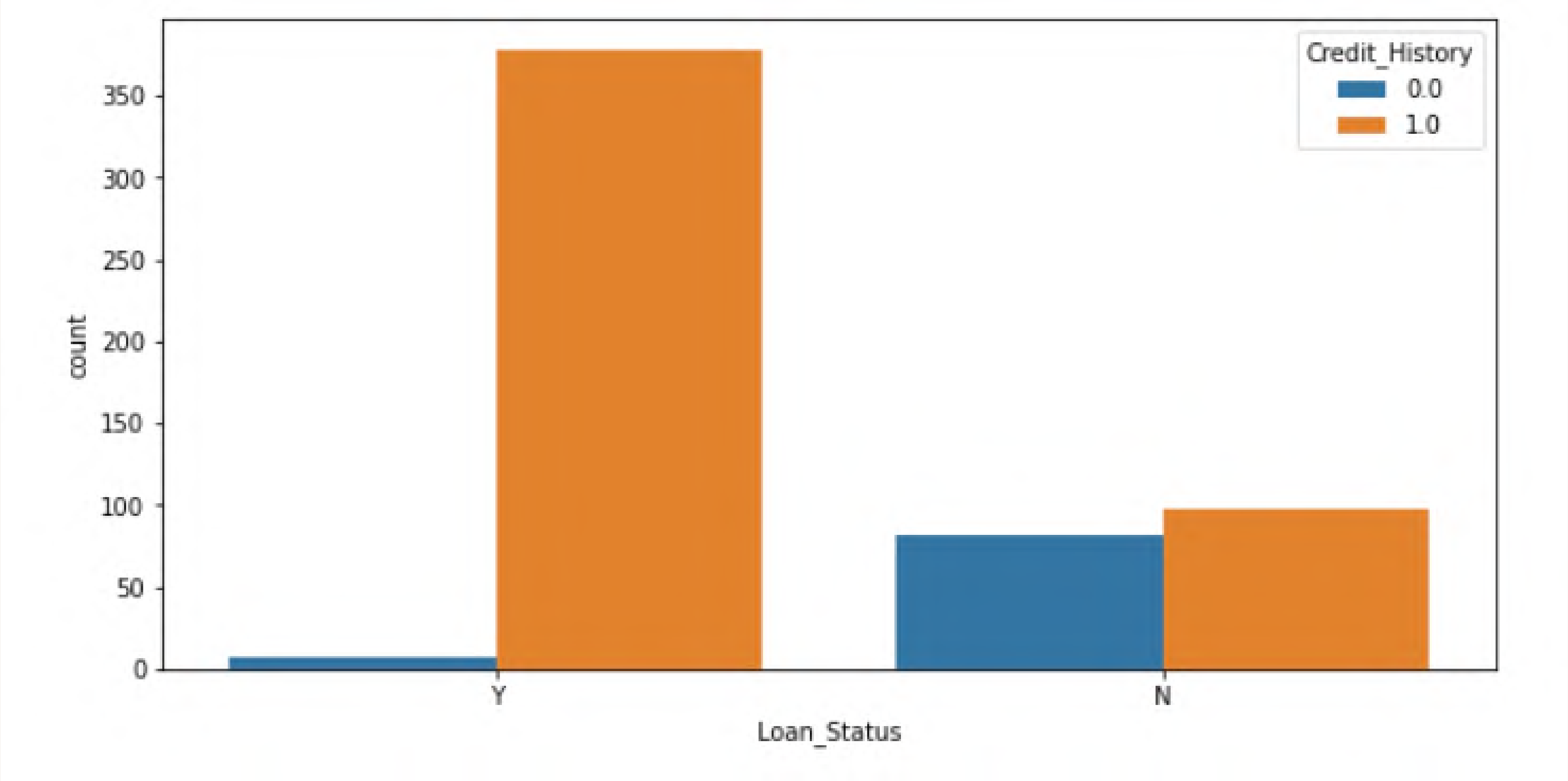
Low loan amount

Old customers to bank

less dependents

Educated

Provide all details to the bank



Hypothesis

Good Credit standing

Good Salary

Assets

Less debt

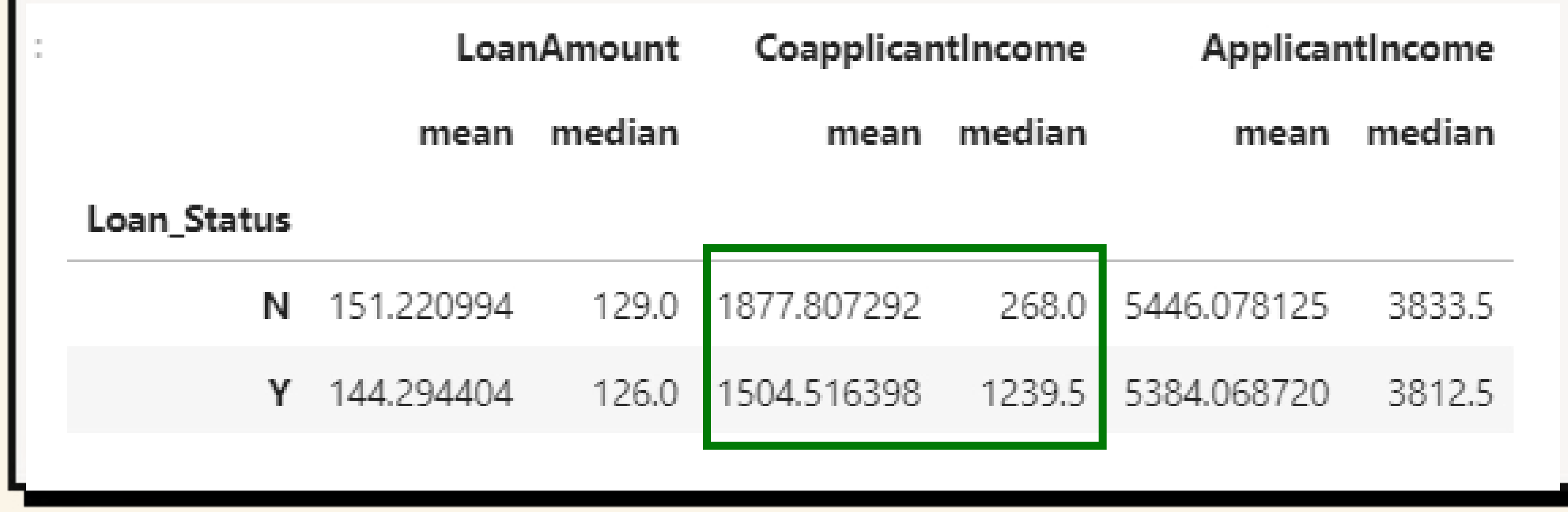
Low loan amount

Old customers to bank

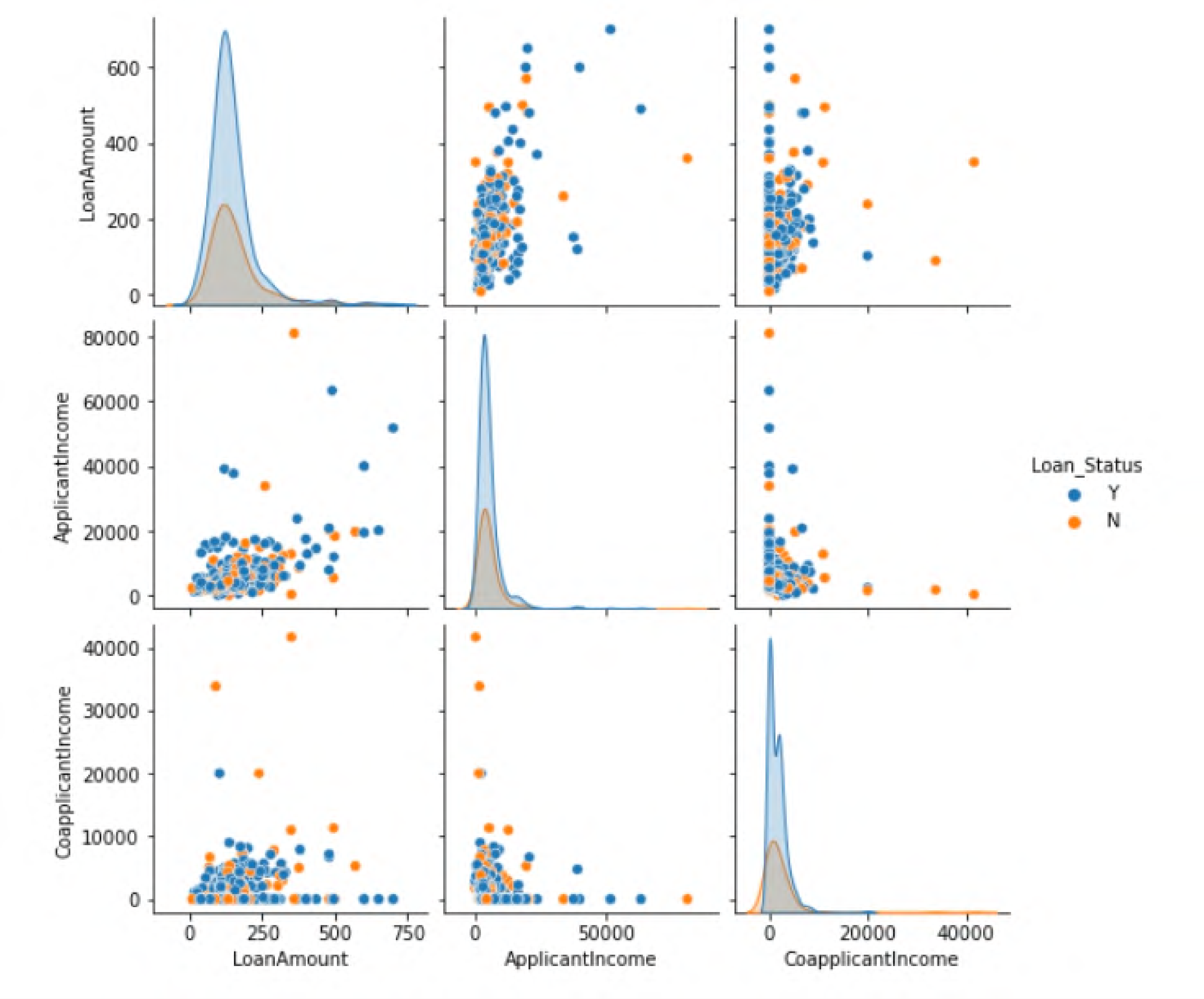
less dependents

Educated

Provide all details to the bank



	LoanAmount		CoapplicantIncome		ApplicantIncome	
	mean	median	mean	median	mean	median
Loan_Status						
N	151.220994	129.0	1877.807292	268.0	5446.078125	3833.5
Y	144.294404	126.0	1504.516398	1239.5	5384.068720	3812.5



Hypothesis

Good Credit standing

Good Salary

Assets - Property Area

Less debt

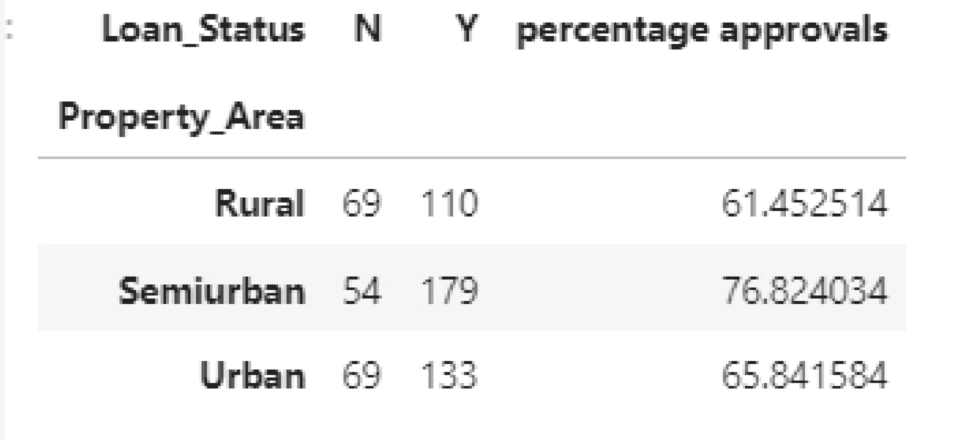
Low loan amount

Old customers to bank

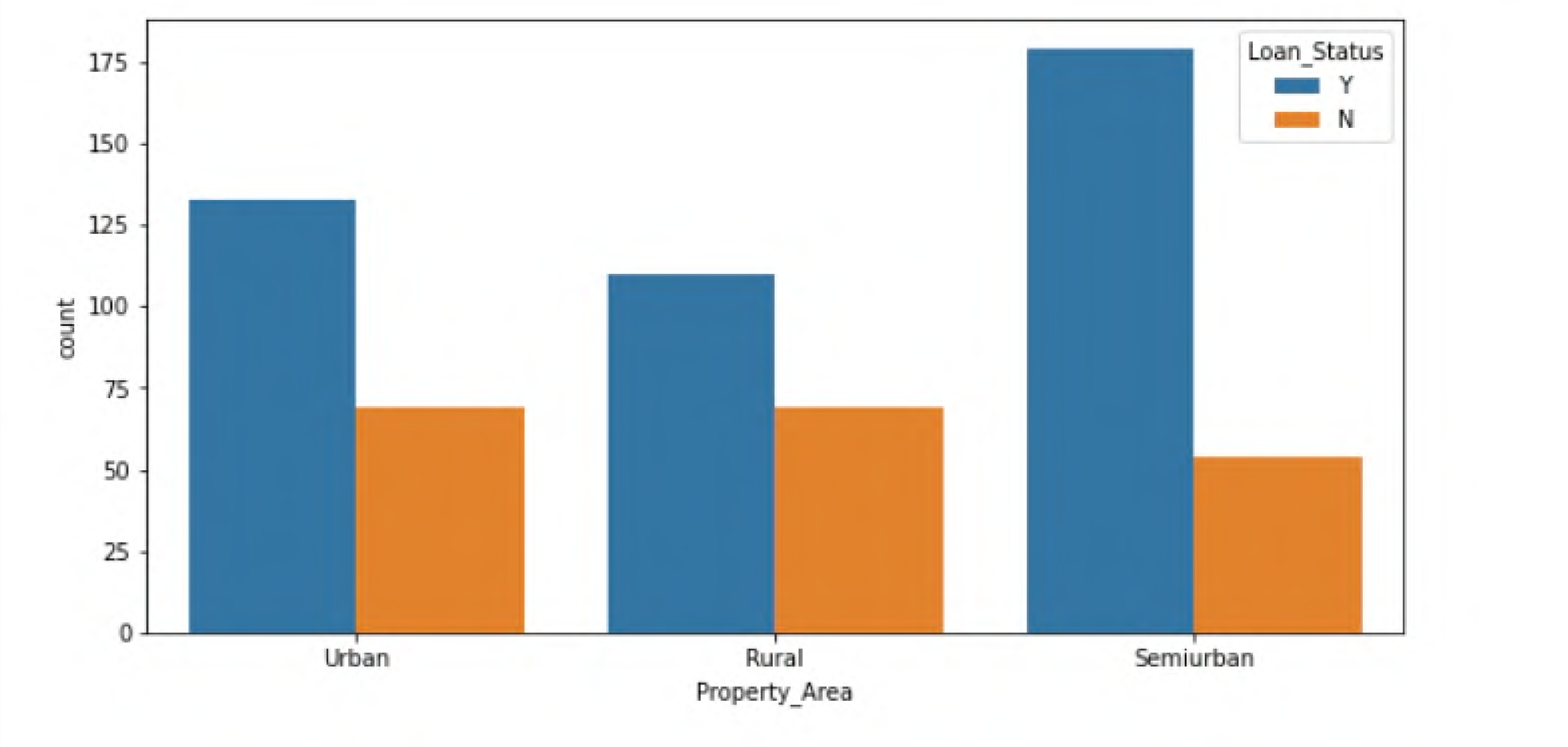
less dependents

Educated

Provide all details to the bank



Loan_Status	N	Y	percentage approvals
Property_Area			
Rural	69	110	61.452514
Semiurban	54	179	76.824034
Urban	69	133	65.841584



Hypothesis

Good Credit standing

Good Salary

~~Assets~~ - **Property Area**

Less debt

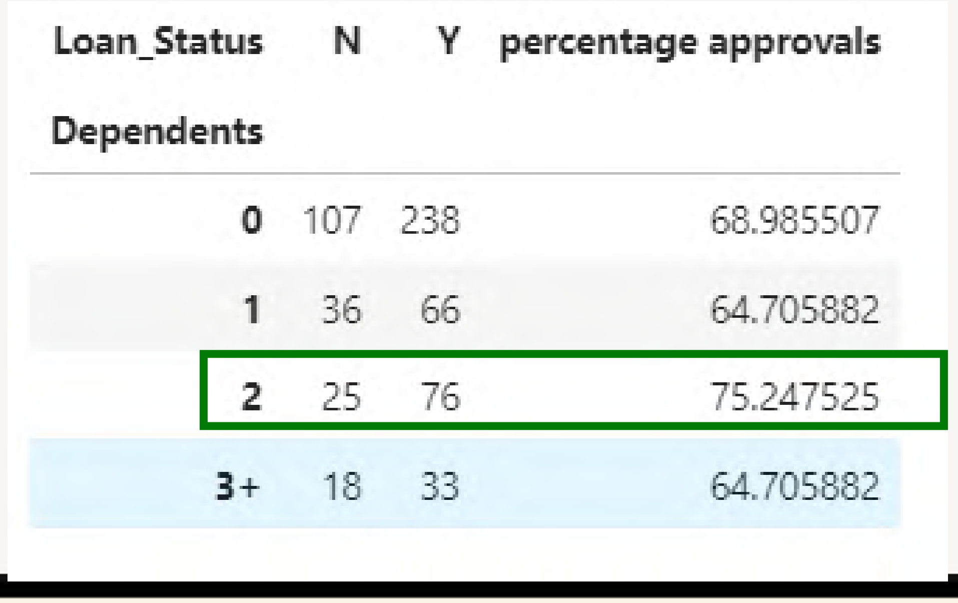
Low loan amount

Old customers to bank

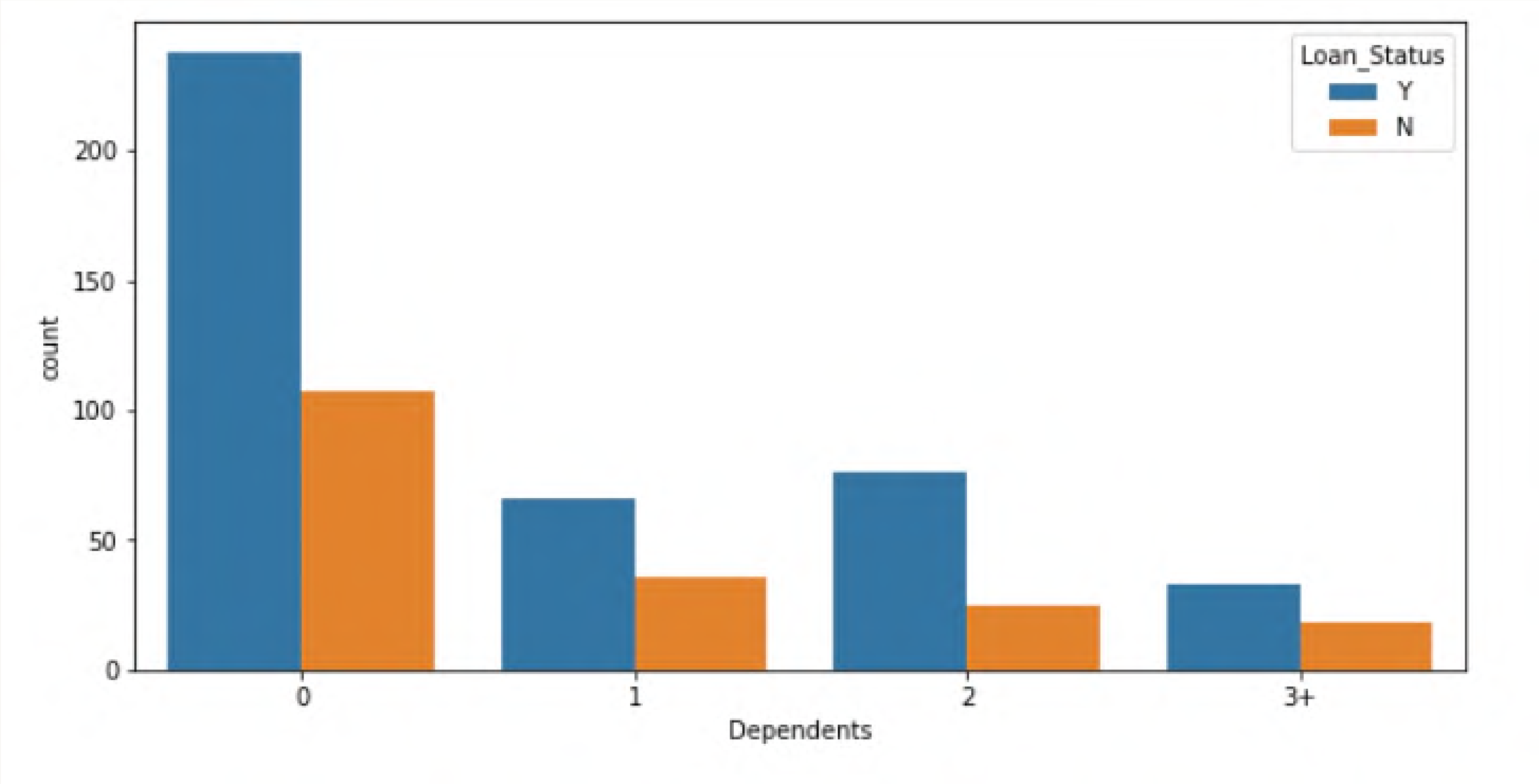
<u>less dependents</u>

Educated

Provide all details to the bank



Loan_Status	N	Y	percentage approvals
Dependents			
0	107	238	68.985507
1	36	66	64.705882
2	25	76	75.247525
3+	18	33	64.705882



Hypothesis

Good Credit standing

Good Salary

~~Assets~~ - **Property Area**

Less debt

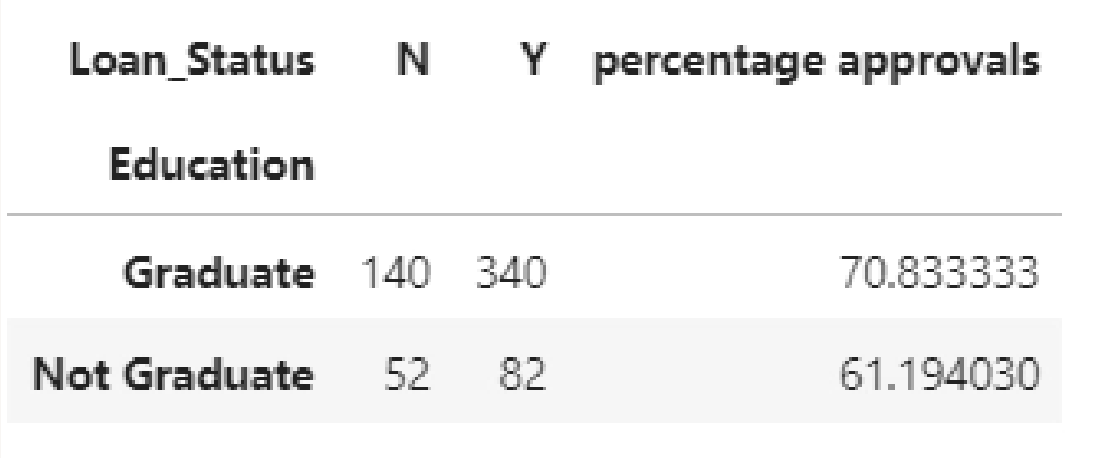
Low loan amount

Old customers to bank

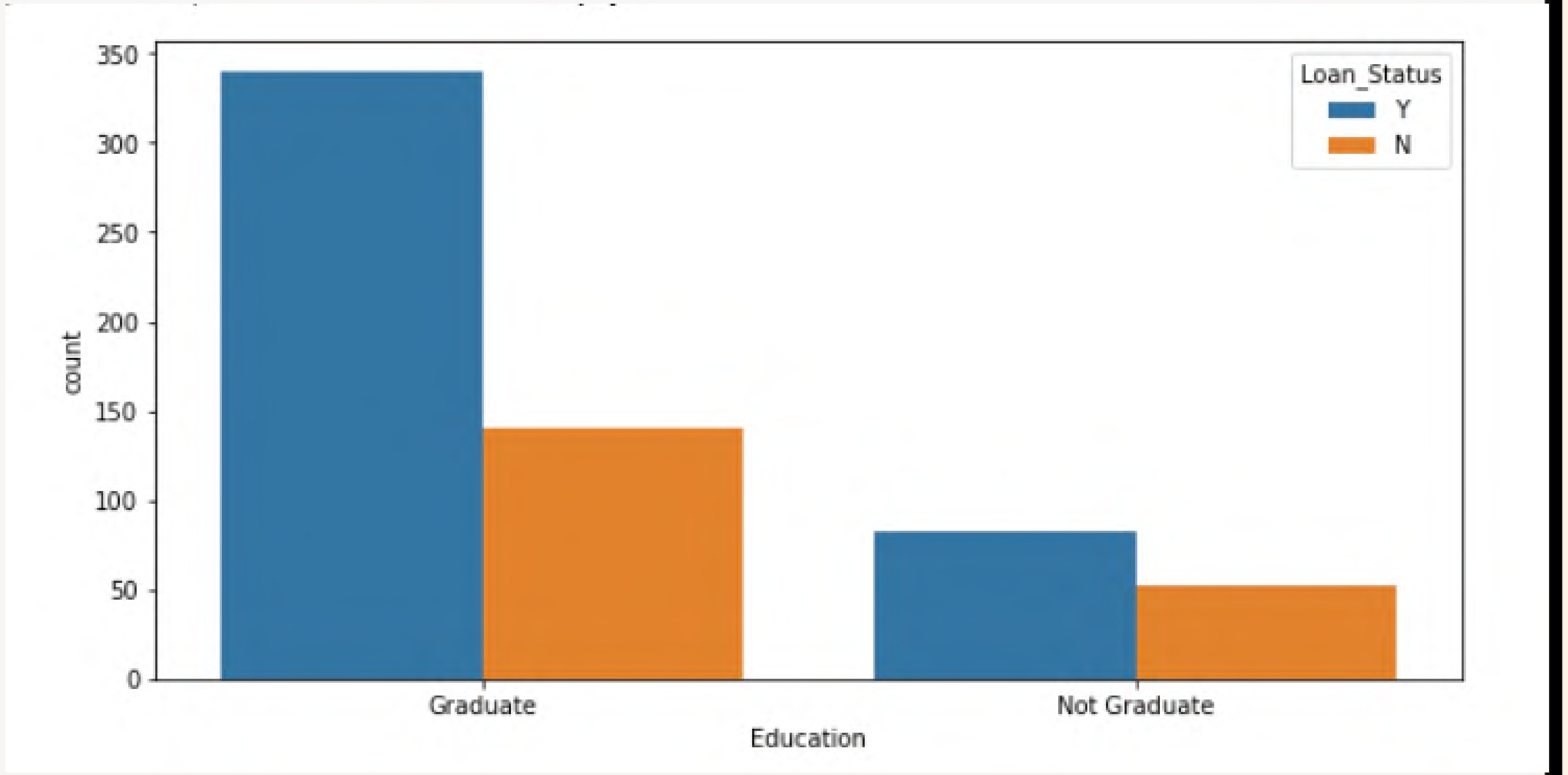
less dependents

Educated

Provide all details to the bank



Loan_Status	N	Y	percentage approvals
Education			
Graduate	140	340	70.833333
Not Graduate	52	82	61.194030



Hypothesis

Good Credit standing

Good Salary

~~Assets~~ - Property Area

Less debt

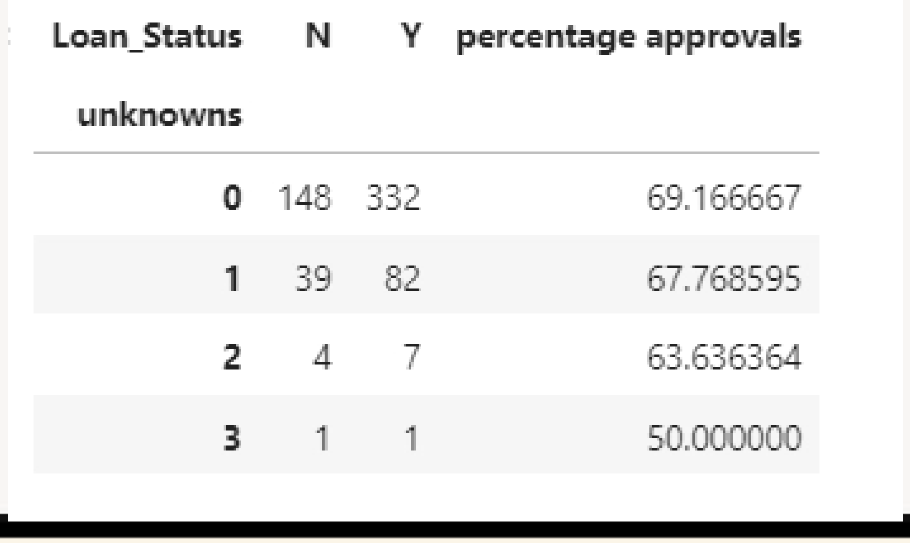
Low loan amount

Old customers to bank

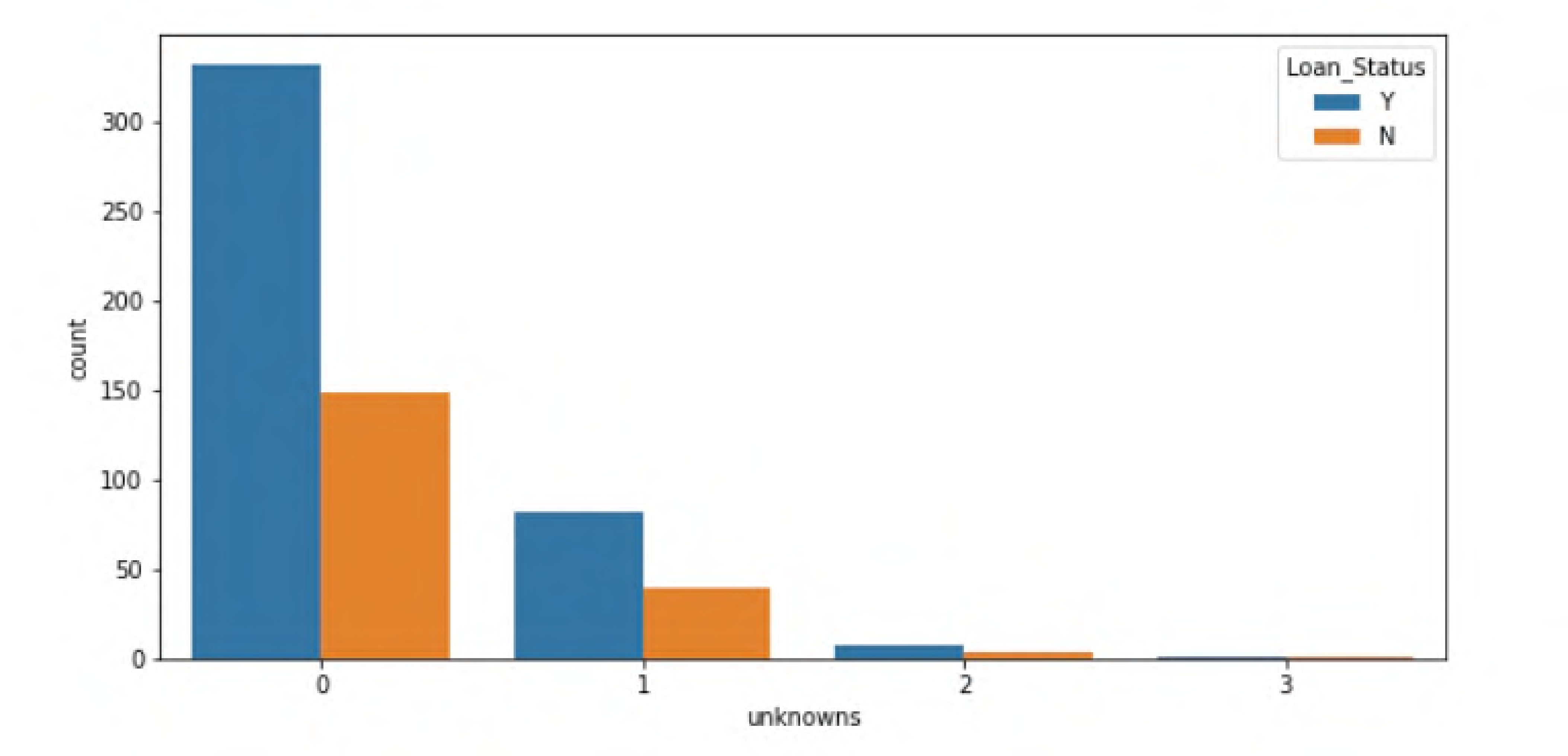
less dependents

Educated

Provide all details to the bank



Loan_Status	N	Y	percentage approvals
unknowns			
0	148	332	69.166667
1	39	82	67.768595
2	4	7	63.636364
3	1	1	50.000000



Hypothesis

Good Credit standing

Good Salary

~~Assets~~ - Property Area

Less debt

Low loan amount

Old customers to bank

less dependents

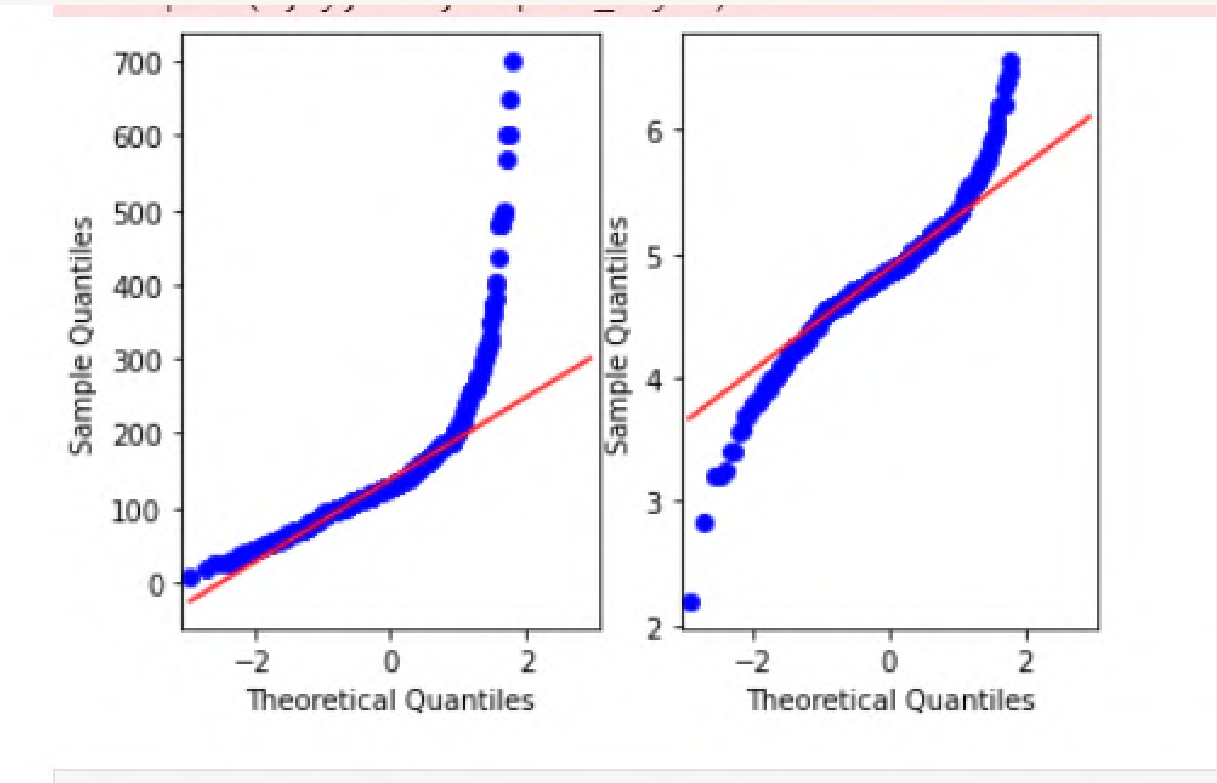
Educated

Provide all details to the bank

Data Cleaning

Missing values

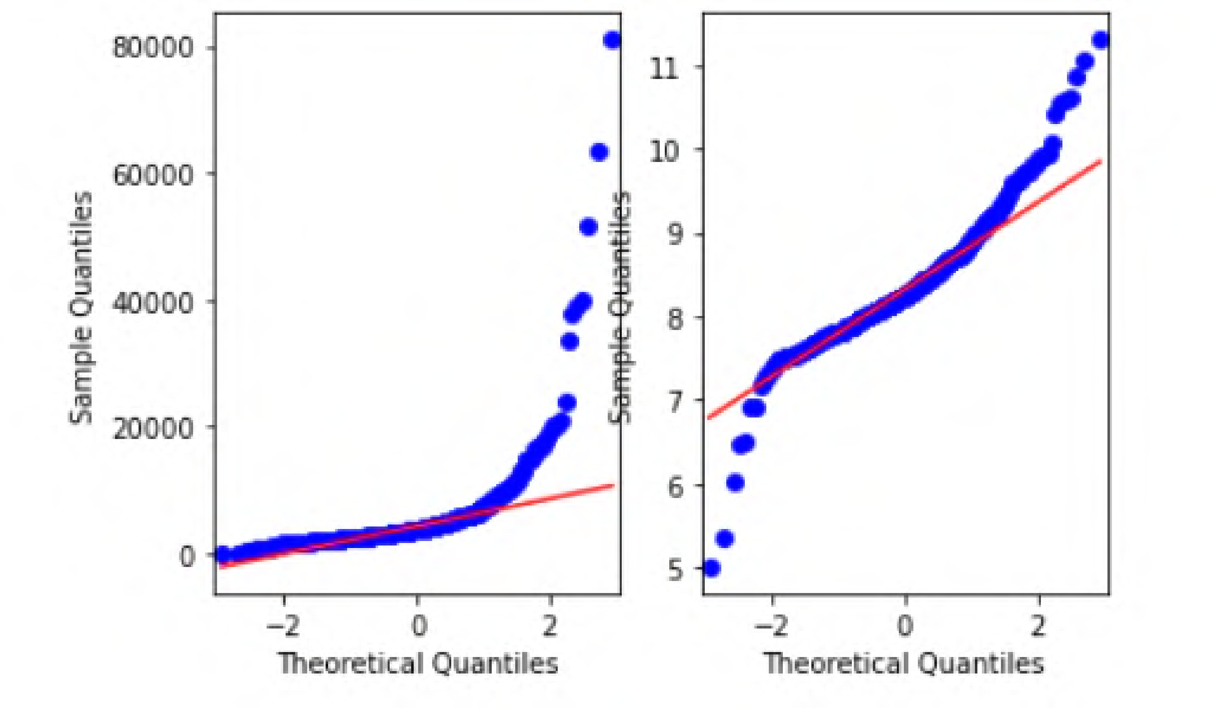
- Missing values have been imputed using **KNNImputer**
  - Data is grouped into clusters and missing values are copied from nearest neighbor



~~Outliers~~

Feature Engineering

- **Transformed**: Income and Loan Amounts were not normally distributed
  - **Log transformation**
- **Added**:
  - Total income = applicant+ co-applicant income
  - Details not provided: 'unknowns': number of fields left empty
- **Dropped**: LoanId, Loan amount term, co-applicant income

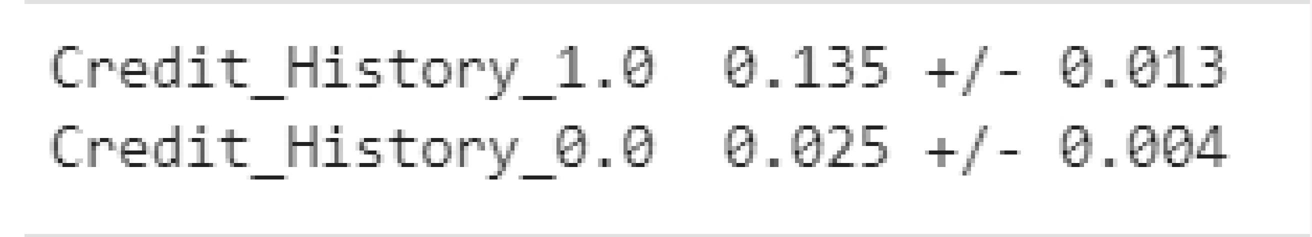


Modelling

Algorithm	f1_score	Recall
Logistic Regression	0.869 (test) 0.874 (train)	0.988 (test) 0.984 (train)
Random Forest Classifier	0.870 (test) 0.872 (train)	0.982 (test) 0.968 (train)
XGBoost	0.863 (test) 0.871 (train)	0.94 (test) 0.93 (train)

Reduce False Negatives: We should not make wrong predictions regarding Loan rejection - We might lose customer.

Important features



```
Credit_History_1.0  0.135 +/- 0.013
Credit_History_0.0  0.025 +/- 0.004
```

Permutation Importance : random forest

Hypothesis

Good Credit standing

Good Salary

~~Assets~~ - Property Area

Less debt

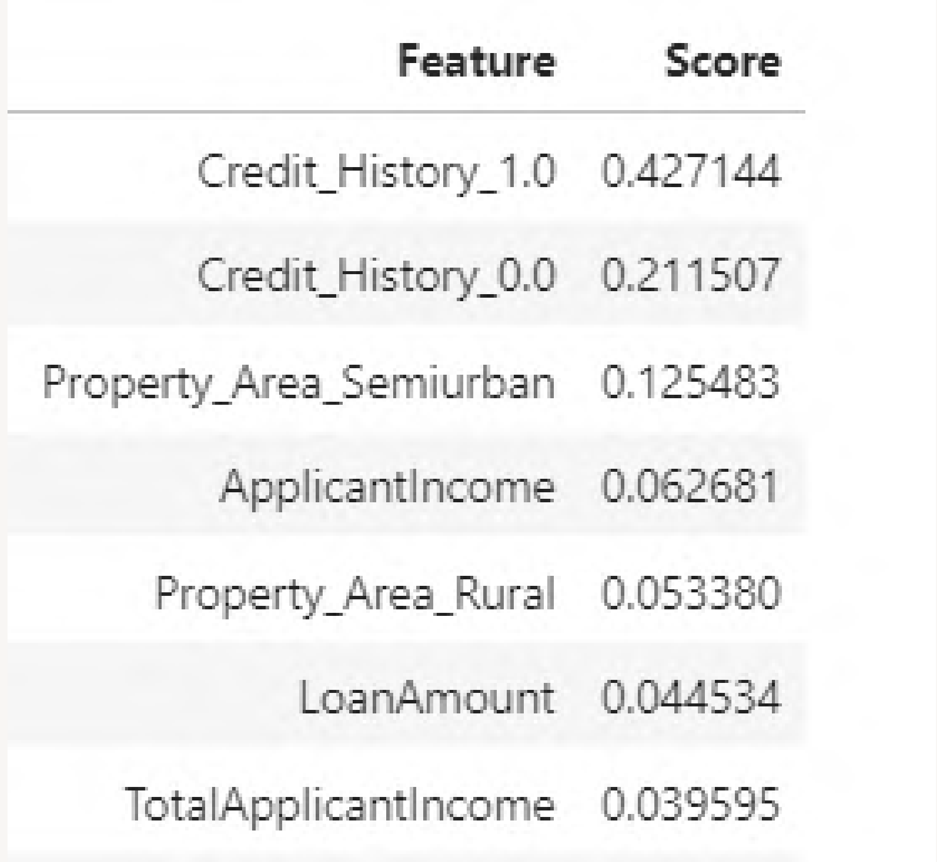
Low loan amount

Old customers to bank

less dependents

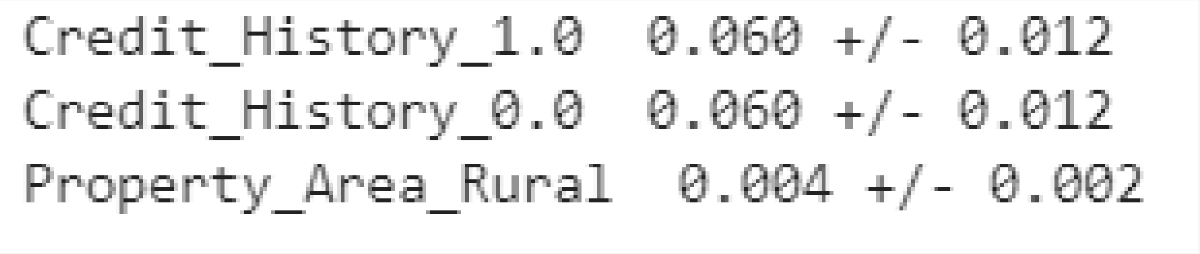
Educated

~~Provide all details to the bank~~



Feature	Score
Credit_History_1.0	0.427144
Credit_History_0.0	0.211507
Property_Area_Semiurban	0.125483
ApplicantIncome	0.062681
Property_Area_Rural	0.053380
LoanAmount	0.044534
TotalApplicantIncome	0.039595

Random Forest feature importance
Based on mean decrease in impurity



```
Credit_History_1.0  0.060 +/- 0.012
Credit_History_0.0  0.060 +/- 0.012
Property_Area_Rural  0.004 +/- 0.002
```

Permutation Importance : Logistic Regression

Deployment

- Deployed on Amazon EC2 instance with Flask RestfulAPI serving as both backend and frontend of the server
- [DEMO IN POSTMAN]

Challenges and Setbacks

- Could not make the **form submission** work
  - My model is crashing with:
    - Form submission
    - Even with random data sample converted to dict and back to a dataframe
- **Inbuilt transformers**:
  - Almost all the inbuilt transformers strip away the index and column data which makes it hard to ensure the smooth flow of data in pipe after featureunions and column transforms
- **Dealing with imbalanced data AND missing feature (due to less data samples) in pipelines: (train_test_split() outside pipe and before feeding it into pipeline)**
  - SMOTE AFTER dummy generation (almost final step in pipeline)
  - Dummy generation: missing categories from a feature results in mismatch of columns between test and train