# ANALYSIS ON LOAN STATUS DATASET

IDS Project – Sem 3


| Raunak Sengupta | Saioni Chatterjee | Aishwarya Pramod |
|---|---|---|
| PES1201700072 | PES1201700118 | PES1201700770 |
| Department of CSE | Department of CSE | Department of CSE |
| PES University | PES University | PES University |
| Bengaluru, India | Bengaluru, India | Bengaluru, India |
| raunaks42@gmail.com | saionichatterjee1999@gmail.com | aishwarya.1999@gmail.com |

# BRIEF DESCRIPTION

The dataset presents a model of credit-worth analysis of bank loan applicants. The study of the analysis emphasizes the possibility of using accounting information when analysing the reliability of loan applicants. The essence, prerequisites and the factors that influence credit-worth are defined. The analysis is based on preliminary research into the factors and the prerequisites that can influence the repayment of the credit and the loans in due course.

# INTRODUCTION & DATA SET

The lender or the bank needs certain documents like proof of assets, proof of income, etc. in order to study the financial and economic stability of the customer before approving the personal loan amount. The borrower must have enough assets or income to repay the loan.

The actual dataset consists of 100000 rows and 18 columns describing the financial details of the loan applicants. A random sample of 25000 rows is generated for the detailed analysis.

The credit-worth of the loan applicants is thoroughly studied through the dataset. The columns represents the eligibility criteria for the customers such as :

1. Credit Score-The lender's credit check reveals the credit score. This is an important factor in determining the personal loan eligibility and interest rate.
2. Current income and expenses-Lender's estimate the monthly debt based on the annual income, term (whether short or long) and the monthly expenditure.
3. Delinquency-Credit issuers actively seek information about the last delinquency committed by the customer and the current credit problems if any or bankruptcies in the recent years.
4. Employment history- Lenders want to see establish proof of ongoing income and employment stability. Self-employed applicants receive closer scrutiny by lenders.
5. Tax liens-The customer must be in a position to write a property to secure the payment of the loans including penalty in case he's not able to repay back .
6. Purpose-The lender must be aware of the purpose of the loan in order to decide the interest rate.
7. Repayment history-Unpaid debts can linger on the credit score and affect loan eligibility.

The source of our dataset is Kaggle - https://www.kaggle.com/zaurbegiev/my-dataset

# PRE PROCESSING

Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

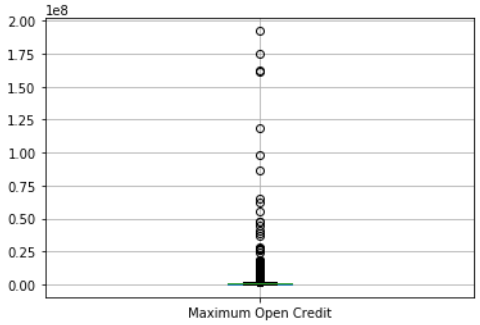
For the column containing credit score and annual income, median for that particular column is calculated and the empty cells are modified with it. For the column stating details about years in current job, the empty row is filled with '0 years' since the

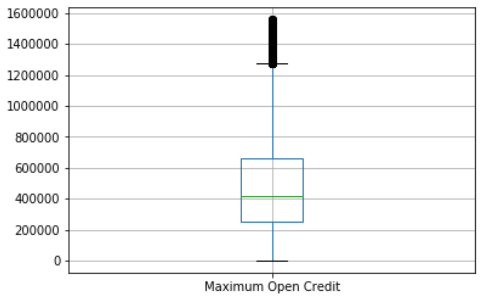
applicant has not filled in the details, so it's considered as the applicant must have not been working. For the column of purpose, 'other' is replaced with 'Other' for uniformity. Using pandas, a clean sample of 25000 is generated and the outliers are removed from this sample. For better comparisons removal of outliers code is applied again to generate another cleaned dataset with even fewer outliers. Data visualization is carried out for this cleaned dataset.

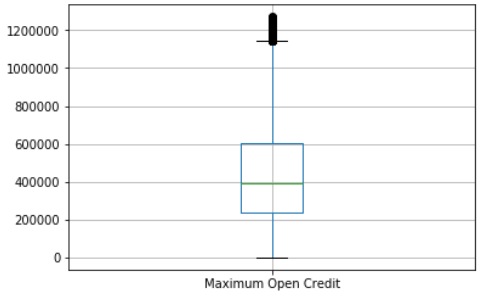
A second round of cleaning is conducted for filling empty values of Bankruptcies, Tax Liens with '0.0' and filling 'HaveMortgage' as 'Home Mortgage' for uniformity.

## Box Plot Comparison for Maximum Open Credit:

Maximum Open Credit - Maximum amount a credit card company allows its browser to spend on a single card.




The first figure represents a boxplot for the sample set containing all outliers.

The second figure represents a boxplot for the cleaned set after removing the outliers once. The remaining outliers now range from 12300-16000 approx. which is lesser as compared to the boxplot drawn from considering the sample set.

The third figure is generated by removing the outliers again and the outliers now range from 11500-12500 approx. which is lesser as compared to the previous bar plot and the skewness is reduced.

# DESCRIPTIVE ANALYSIS

A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of a collection of information.

They are usually classified under two kinds:

- Measures of Central Tendency
- Measures of Spread

Measures of Central Tendency are:

- Mean
- Median
- Mode
- Trimmed Mean

Measures of Spread are:

- Range
- Inter-Quartile Range
- Standard Deviation
- Variance

# Measures of Central Tendency and Spread:

## MEAN

| | |
|---|---|
| Current Loan Amount | 2.573060e+05 |
| Credit Score | 7.235938e+02 |
| Annual Income | 1.087248e+06 |
| Monthly Debt | 1.486348e+04 |
| Years of Credit History | 1.670675e+01 |
| Months since last delinquent | 1.774736e+02 |
| Number of Open Accounts | 1.017528e+01 |
| Number of Credit Problems | 1.842318e-01 |
| Current Credit Balance | 2.013556e+05 |
| Maximum Open Credit | 4.431181e+05 |
| Bankruptcies | 1.320297e-01 |
| Tax Liens | 2.904503e-02 |

## STANDARD DEVIATION

| | |
|---|---|
| Current Loan Amount | 141136.778305 |
| Credit Score | 14.295337 |
| Annual Income | 321989.959884 |
| Monthly Debt | 7914.319750 |
| Years of Credit History | 5.590276 |
| Months since last delinquent | 132.562814 |
| Number of Open Accounts | 4.046443 |
| Number of Credit Problems | 0.503269 |
| Current Credit Balance | 134643.692852 |
| Maximum Open Credit | 263742.115405 |
| Bankruptcies | 0.366683 |
| Tax Liens | 0.270902 |

## QUARTILES

| | Current Loan Amount | Credit Score | Annual Income | Monthly Debt | Years of Credit History | Months since last delinquent | Number of Open Accounts | Number of Credit Problems | Current Credit Balance | Maximum Open Credit | Bankruptcies | Tax Liens |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0.25** | 150403.0 | 717.0 | 864965.5 | 8961.540 | 12.7 | 36.0 | 7.0 | 0.0 | 98762.0 | 240526.0 | 0.0 | 0.0 |
| **0.50** | 224268.0 | 724.0 | 1174162.0 | 13970.510 | 16.0 | 300.0 | 10.0 | 0.0 | 173831.0 | 391402.0 | 0.0 | 0.0 |
| **0.75** | 338041.0 | 733.0 | 1179539.0 | 19912.855 | 20.2 | 300.0 | 13.0 | 0.0 | 281827.0 | 600809.0 | 0.0 | 0.0 |

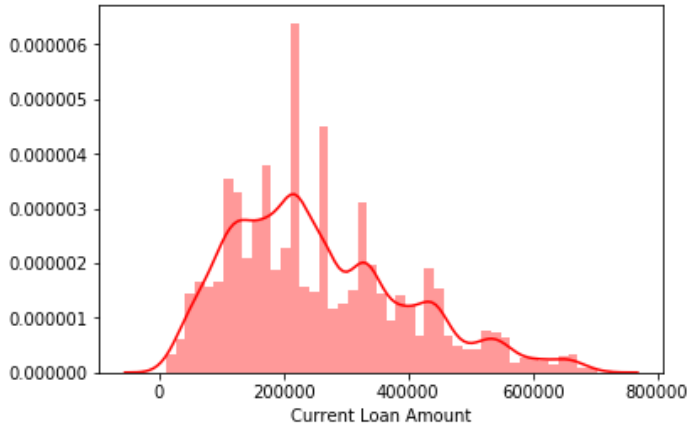
## MEDIAN

| | |
|---|---|
| Current Loan Amount | 224268.00 |
| Credit Score | 724.00 |
| Annual Income | 1174162.00 |
| Monthly Debt | 13970.51 |
| Years of Credit History | 16.00 |
| Months since last delinquent | 300.00 |
| Number of Open Accounts | 10.00 |
| Number of Credit Problems | 0.00 |
| Current Credit Balance | 173831.00 |
| Maximum Open Credit | 391402.00 |
| Bankruptcies | 0.00 |

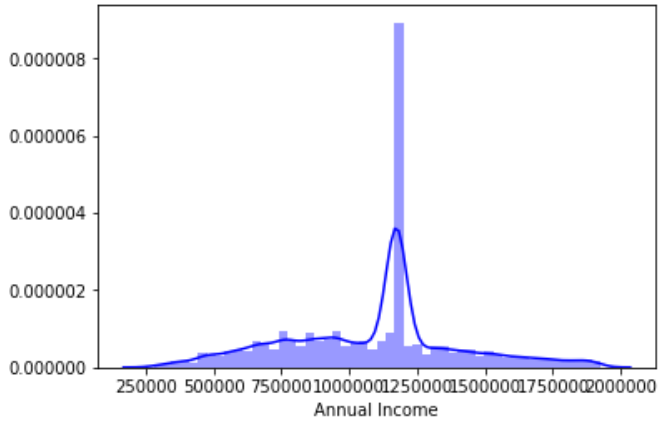
## VARIANCE

| | |
|---|---|
| Current Loan Amount | 1.991959e+10 |
| Credit Score | 2.043567e+02 |
| Annual Income | 1.036775e+11 |
| Monthly Debt | 6.263646e+07 |
| Years of Credit History | 3.125118e+01 |
| Months since last delinquent | 1.757290e+04 |
| Number of Open Accounts | 1.637370e+01 |
| Number of Credit Problems | 2.532793e-01 |
| Current Credit Balance | 1.812892e+10 |
| Maximum Open Credit | 6.955990e+10 |
| Bankruptcies | 1.344563e-01 |
| Tax Liens | 7.338797e-02 |

## Histogram representations:

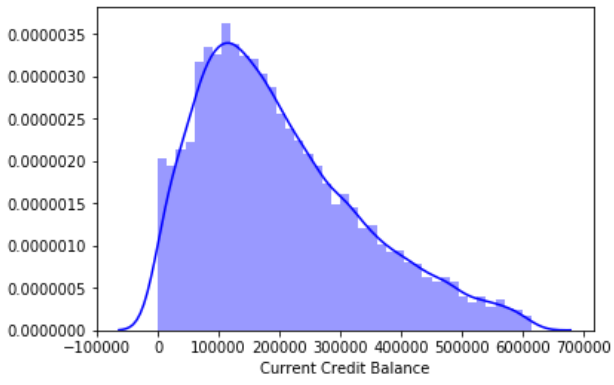



The histogram is right skewed and it depicts that not many people take higher loans.

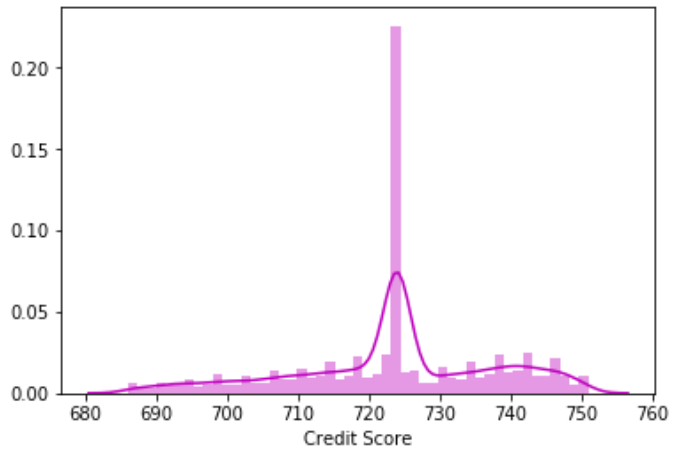



The annual income for most people

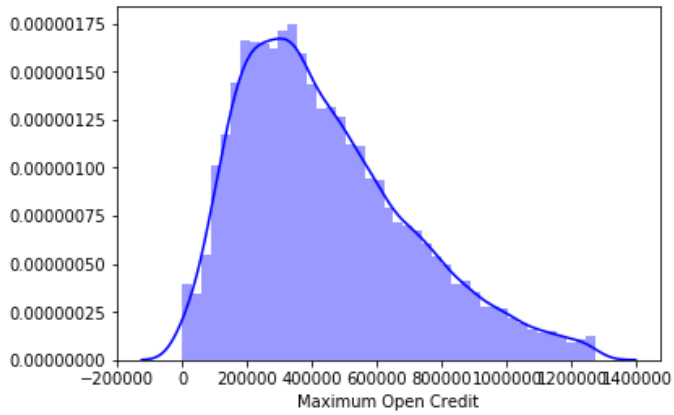
ranges from 1.1 crore-1.25 crores



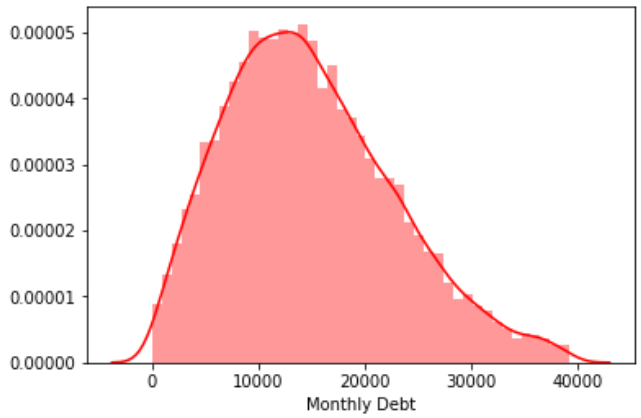

The histogram is slightly right skewed and maximum value ranges from 100000-150000. From the graph we can infer that majority of the people do not have a high current credit



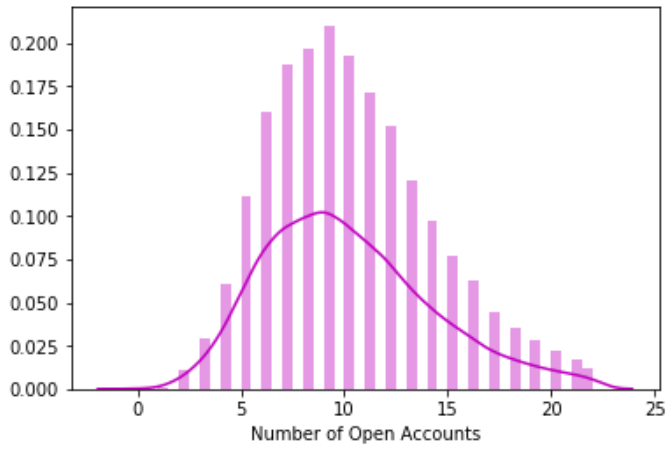

Credit scores above 700 are considered to be a good credit score. Majority of the customers have their credit score in the range 700-750.



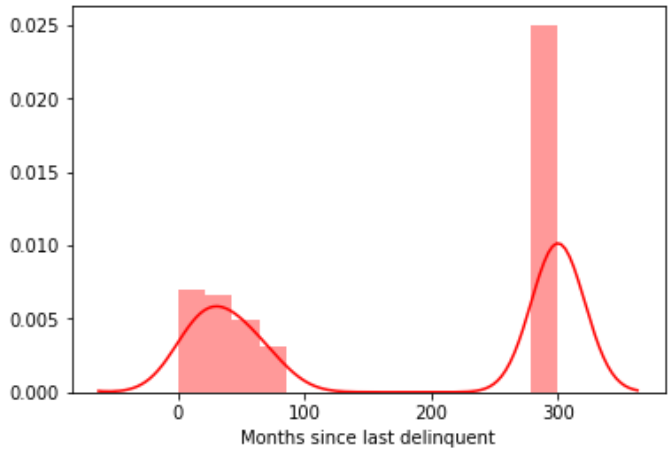

This histogram represents a right skewed graph where the maximum value ranges from 2000000-3000000.



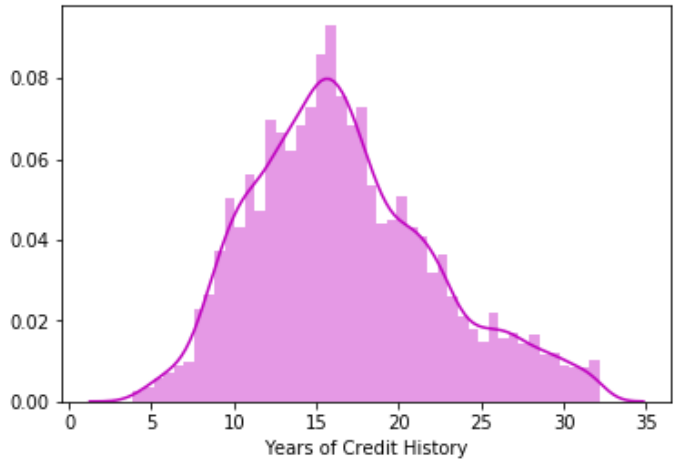

The histogram depicts a normally distributed graph.




The plotted histogram represents normal distribution of the total number of credit accounts opened by the customer.



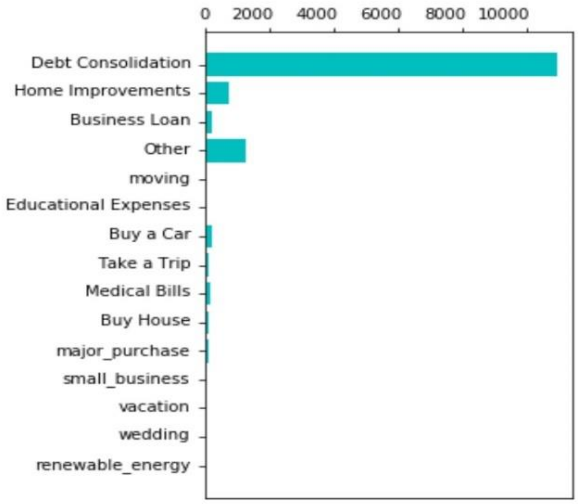

The Histogram shows that few of the customers have faced problems in the past 90 months and the rest has not faced for a long time.



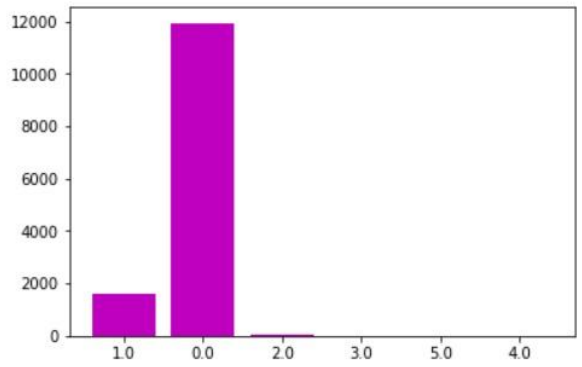

A normally distributed histogram is drawn from the information obtained.

## Bar Graph Representations:

Purpose: The graph is to convey the various motives of the customers to take a loan. Larger part of the people require a loan to pay their debt consolidation.

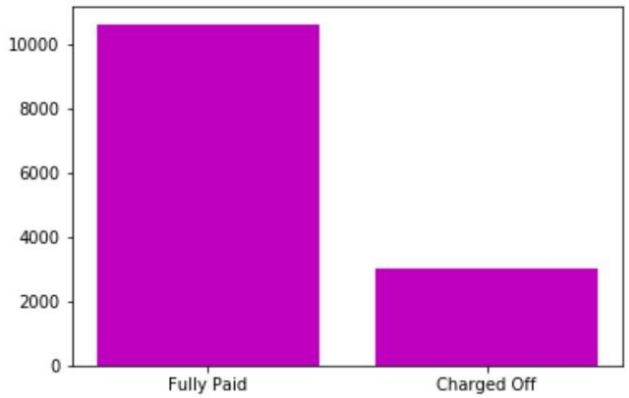



Bankruptcies: Vast number of people have no bankruptcy.

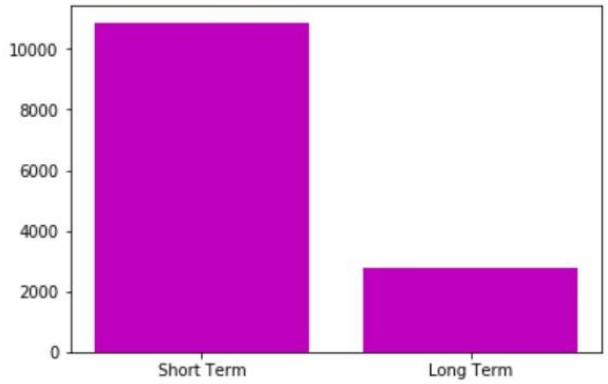



Home Ownership: Many customers who've taken a loan, live on rent.

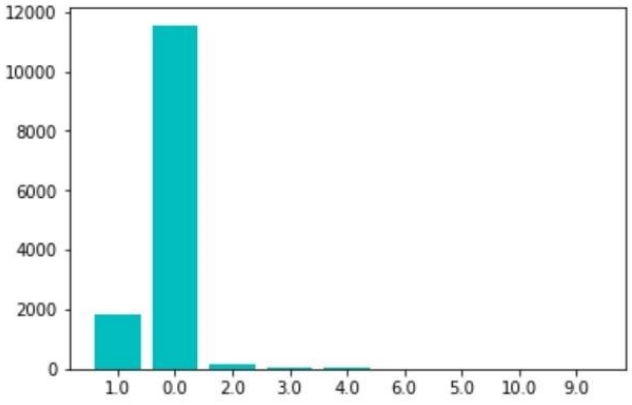
Loan Status: The graph consists of people where most of them have paid off their loans.



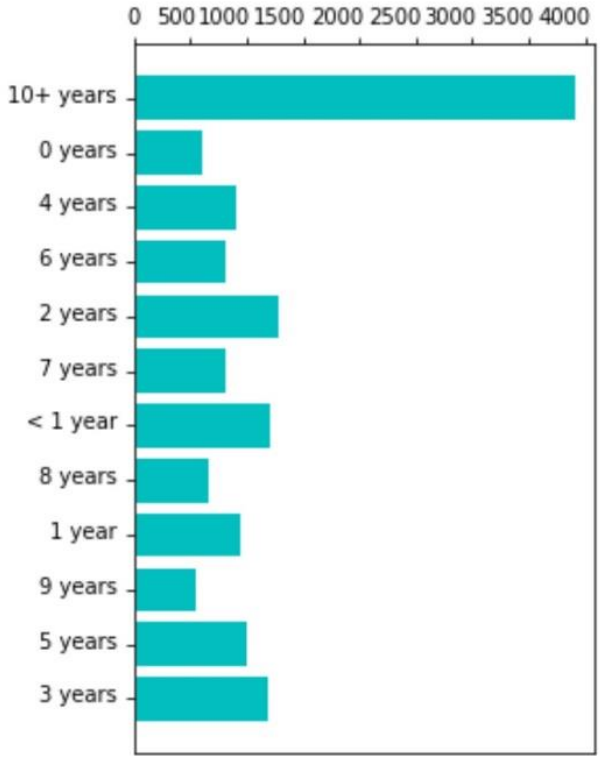

Term: Most of the customers have taken a loan for a short period of time.



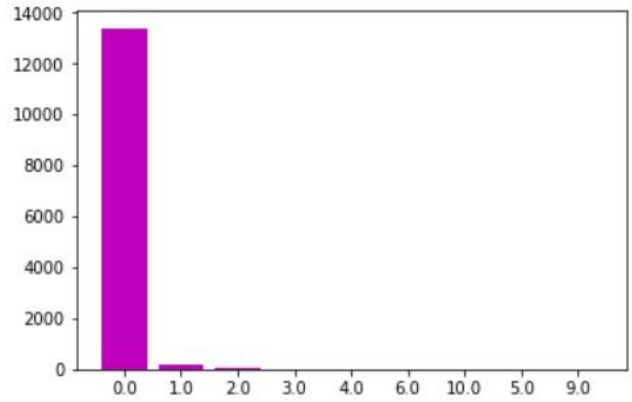

Number of Credit Loans: Majority of the people have not taken credit loans.




Years in current job: The below graph depicts the number of years the customers have been working and most of the customers have worked for more than 10 years.



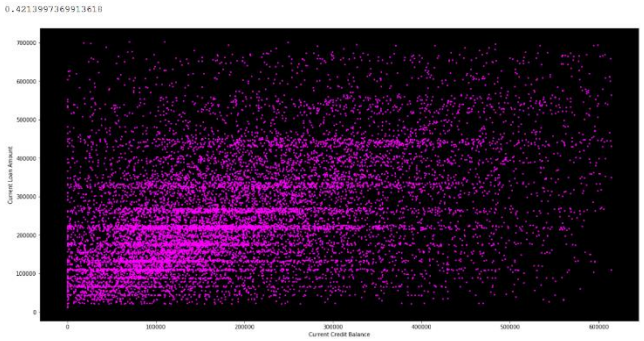

Tax Liens: From the graph it can be inferred that majority of them have no tax liens.



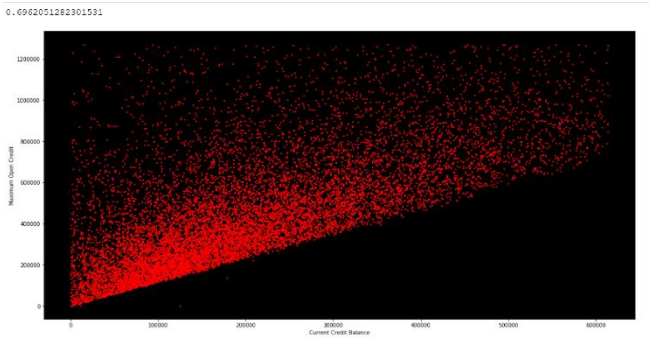
# STATISTICAL ANALYSIS

## Pearson's Correlation Test:

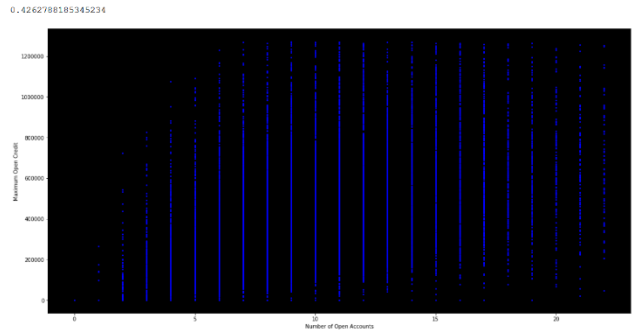
The test returns a correlational coefficient between two columns that indicates the increase/decrease of one column with the increase of the other. A scatterplot may be used to visualise this.




For Current Loan Amount vs Current Credit Balance, the Pearson Correlational Coefficient is 0.4214, indicating a strong correlation between the columns.




For Maximum Open Credit vs Current Credit Balance, the PCC is 0.6962, indicating a very strong correlation.



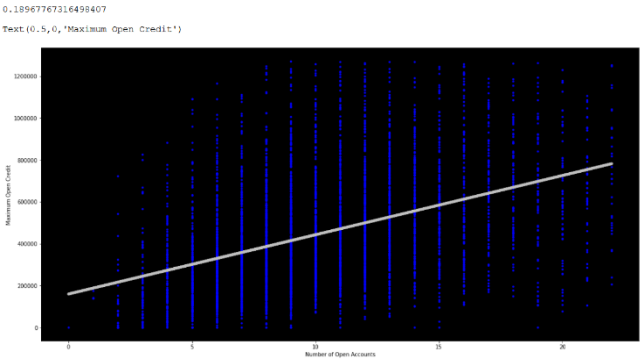

For Number of Open Accounts vs Maximum Open Credit, PCC is 0.4263, indicating a strong correlation.

## Multivariate Linear Regression:

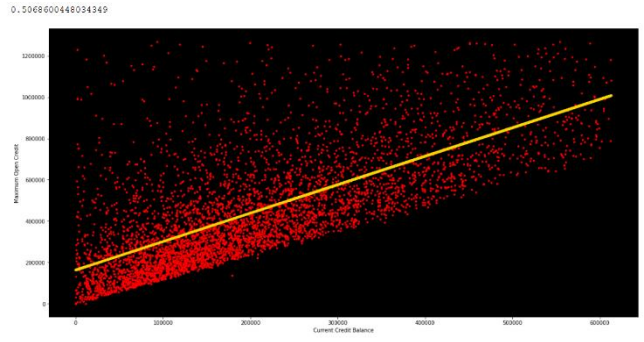
Linear Regression Models are used to predict a future value of a dependant variable using the given values of multiple independent variables by plotting a regression line to best fit the scatter plot of our data.

We shall conduct a multivariate regression model for Current Credit Balance and Number of Open Accounts to predict Maximum Open Credit.

The fitted Simple Linear Regression models look like:

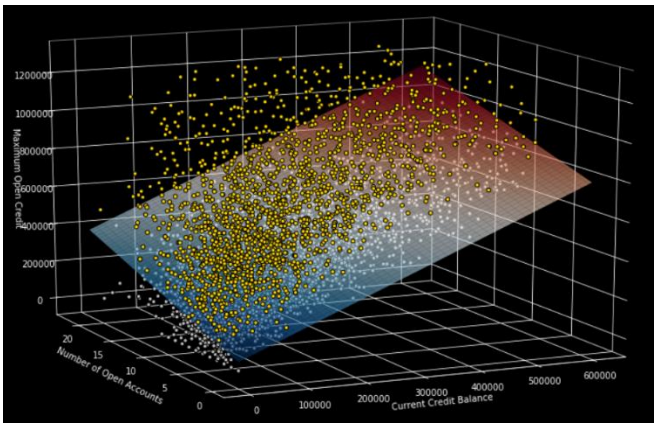



For Number of Open Accounts vs Maximum Open Credit with 0.1897 goodness of fit.




For Current Credit Balance vs Maximum Open Credit with 0.5069 goodness of fit.

The Multivariate Regression Model fits a plane to a 3d scatter plot to implement the model, it looks like:




A python program can be used to learn from the database to form a model, and can take inputs of Number of Open Accounts and Current Credit Balance to predict Maximum Open Credit.

## Chi Square Test:

The Chi Square Independence Test is used to test the dependence between two categorical variables. It is implemented as a hypothesis test with a certain probability set for Independence.

A contingency table is plotted for a pair of categorical variables and the result is printed.

For our study, we would like to know which variables Loan Status depends upon.

First, For Loan Status and Home Ownership

```
Home Mortgage {'Fully Paid': 4749, 'Charged Off': 1228}
Rent {'Fully Paid': 4841, 'Charged Off': 1545}
Own Home {'Fully Paid': 1015, 'Charged Off': 257}
dof=2
[[4648.77777778 1328.22222222]
 [4966.88888889 1419.11111111]
 [ 989.33333333  282.66666667]]
probability=0.999, critical=13.816, stat=27.078
Dependent (reject H0 as stat>=critical)
significance=0.001, p=0.000
Dependent (reject H0 as p<=alpha)
```

The observed and expected contingency table are printed, and the result is: They are dependant even at a 99.9% probability of Independence.

Then, for Loan Status and Bankruptcies

```
1.0 {'Fully Paid': 1292, 'Charged Off': 338}
0.0 {'Fully Paid': 9255, 'Charged Off': 2681}
2.0 {'Fully Paid': 42, 'Charged Off': 7}
3.0 {'Fully Paid': 12, 'Charged Off': 2}
5.0 {'Fully Paid': 1, 'Charged Off': 1}
4.0 {'Fully Paid': 3, 'Charged Off': 1}
dof=5
[[1.26777778e+03 3.62222222e+02]
 [9.28355556e+03 2.65244444e+03]
 [3.81111111e+01 1.08888889e+01]
 [1.08888889e+01 3.11111111e+00]
 [1.55555556e+00 4.44444444e-01]
 [3.11111111e+00 8.88888889e-01]]
probability=0.850, critical=8.115, stat=5.684
Independent (fail to reject H0)
significance=0.150, p=0.338
Independent (fail to reject H0)
```

The result, even after lowering probability of Independence to 85%, is that they are independent.

For Loan Status and Term

```
Short Term {'Fully Paid': 8660, 'Charged Off': 2216}
Long Term {'Fully Paid': 1945, 'Charged Off': 814}
dof=1
[[8459.11111111 2416.88888889]
 [2145.88888889  613.11111111]]
probability=0.999, critical=10.828, stat=105.570
Dependent (reject H0 as stat>=critical)
significance=0.001, p=0.000
Dependent (reject H0 as p<=alpha)
```

The result is the same as that of Home ownership test, they are dependant even at 99.9% probability.

For Tax Liens, for 95% probability, they are independent

```
0.0 {'Fully Paid': 10410, 'Charged Off': 2980}
1.0 {'Fully Paid': 132, 'Charged Off': 33}
2.0 {'Fully Paid': 42, 'Charged Off': 5}
3.0 {'Fully Paid': 8, 'Charged Off': 7}
4.0 {'Fully Paid': 7, 'Charged Off': 3}
6.0 {'Fully Paid': 3, 'Charged Off': 0}
10.0 {'Fully Paid': 1, 'Charged Off': 0}
5.0 {'Fully Paid': 2, 'Charged Off': 1}
9.0 {'Fully Paid': 0, 'Charged Off': 1}
dof=8
[[1.04144444e+04 2.97555556e+03]
 [1.28333333e+02 3.66666667e+01]
 [3.65555556e+01 1.04444444e+01]
 [1.16666667e+01 3.33333333e+00]
 [7.77777778e+00 2.22222222e+00]
 [2.33333333e+00 6.66666667e-01]
 [7.77777778e-01 2.22222222e-01]
 [2.33333333e+00 6.66666667e-01]
 [7.77777778e-01 2.22222222e-01]]
probability=0.950, critical=15.507, stat=14.522
Independent (fail to reject H0)
significance=0.050, p=0.069
Independent (fail to reject H0)
```

But for 92.5% probability, they are dependant

```
0.0 {'Fully Paid': 10410, 'Charged Off': 2980}
1.0 {'Fully Paid': 132, 'Charged Off': 33}
2.0 {'Fully Paid': 42, 'Charged Off': 5}
3.0 {'Fully Paid': 8, 'Charged Off': 7}
4.0 {'Fully Paid': 7, 'Charged Off': 3}
6.0 {'Fully Paid': 3, 'Charged Off': 0}
10.0 {'Fully Paid': 1, 'Charged Off': 0}
5.0 {'Fully Paid': 2, 'Charged Off': 1}
9.0 {'Fully Paid': 0, 'Charged Off': 1}
dof=8
[[1.04144444e+04 2.97555556e+03]
 [1.28333333e+02 3.66666667e+01]
 [3.65555556e+01 1.04444444e+01]
 [1.16666667e+01 3.33333333e+00]
 [7.77777778e+00 2.22222222e+00]
 [2.33333333e+00 6.66666667e-01]
 [7.77777778e-01 2.22222222e-01]
 [2.33333333e+00 6.66666667e-01]
 [7.77777778e-01 2.22222222e-01]]
probability=0.925, critical=14.270, stat=14.522
Dependent (reject H0 as stat>=critical)
significance=0.075, p=0.069
Dependent (reject H0 as p<=alpha)
```

But since we wish for our test to be more accurate, we consider the first result and consider them Independent.

Thus, Loan Status is dependant on Term and Home Ownership.

# HYPOTHESIS TESTING & CONFIDENCE INTERVAL

1) The hypothesis that is made from the data set considering the column annual income and loan status is that

$H_a$ : The bank employee claims that the customers who have fully paid their loan have more than average annual income.

$H_o$ : According to null hypothesis it is not necessary that those who have fully paid their status loan should have a higher average annual income.

Confidence interval calculation

The lower limit is : 1135077.62638 (LL)

The upper limit is: 1206352.79362 (UL)

Margin of Error = $Z_{\alpha/2} * \sigma/\sqrt{n} = (UL-LL)*2$

From this $Z_{\alpha/2} = 3.21$

Therefore the confidence level is 99.91 %

```
population sales mean= 1144530.10736541
standard deviation= 388628.1695394347
7696
1135077.62638
1206352.79362
```

2) The hypothesis that is made from the data set considering the column term is that

$H_a$ : The bank employee claims that the customers who have short term has a slightly lesser credit score.

$H_o$ : According to null hypothesis it is not necessary that the customer should have a lesser credit score though he has a short term.

Confidence interval calculation for credit score and current credit balance

The lower limit is : 725.76264486 (LL)

The upper limit is: 727.86535514 (UL)

Margin of Error = $Z_{\alpha/2} * \sigma/\sqrt{n} = (UL-LL)*2$

From this $Z_{\alpha/2} = 3.12$

Therefore the confidence level is 99.87 %

```
population sales mean= 222596.38175503115
standard deviation= 158666.36375614215
8639
725.76264486
727.86535514
```

# CONCLUSION

From the above data analysis and data visualisation we can strongly conclude:

1) The Loan Status (Fully Paid or Charged Off) depends on Annual Income, ie, will be more likely to be fully paid for higher annual income, at 99.91% confidence.
2) The Current Credit Balance and Maximum Open Credit depend positively with each other, with 0.69 Correlation Coefficient.
3) The Loan Status has a dependence on Term (chi square statistic=105.57) and Home Ownership (chi square statistic= 27.08)
4) The Multivariate Linear Regression Model is able to approximately determine the Maximum Open Credit, given values of Current Credit Balance and Number of Open Accounts.