Risk of Loan DefAult

**Prepared by**

Anurag Sharma

Aman Panwar

Dinesh Poudel

**Prepared for**

Dr. David J. Lucus, Professor

October 12, 2019

**Contents**

[Introduction & Analysis of Data 2](#_Toc21804056)

[Flow Chart 2](#_Toc21804057)

[Data Source 2](#_Toc21804058)

[Data Selection 3](#_Toc21804059)

[Primary Analysis 3](#_Toc21804060)

[Data Processing 3](#_Toc21804061)

[Data Visualization 3](#_Toc21804062)

[Data Merging 3](#_Toc21804063)

[Common Elements 3](#_Toc21804064)

[Multilevel data 4](#_Toc21804065)

[Combined Values of data 4](#_Toc21804066)

[Data Cleaning 4](#_Toc21804067)

[Primary Cleaning of Data 4](#_Toc21804068)

[Quality of Data 5](#_Toc21804069)

[Missing Data 5](#_Toc21804070)

[Analysis of Visualization 6](#_Toc21804071)

[Principles of good visualization 9](#_Toc21804072)

[Conclusion 9](#_Toc21804073)

[Instruction for code 10](#_Toc21804074)

[Members 11](#_Toc21804075)

[Name of Datasets 11](#_Toc21804076)

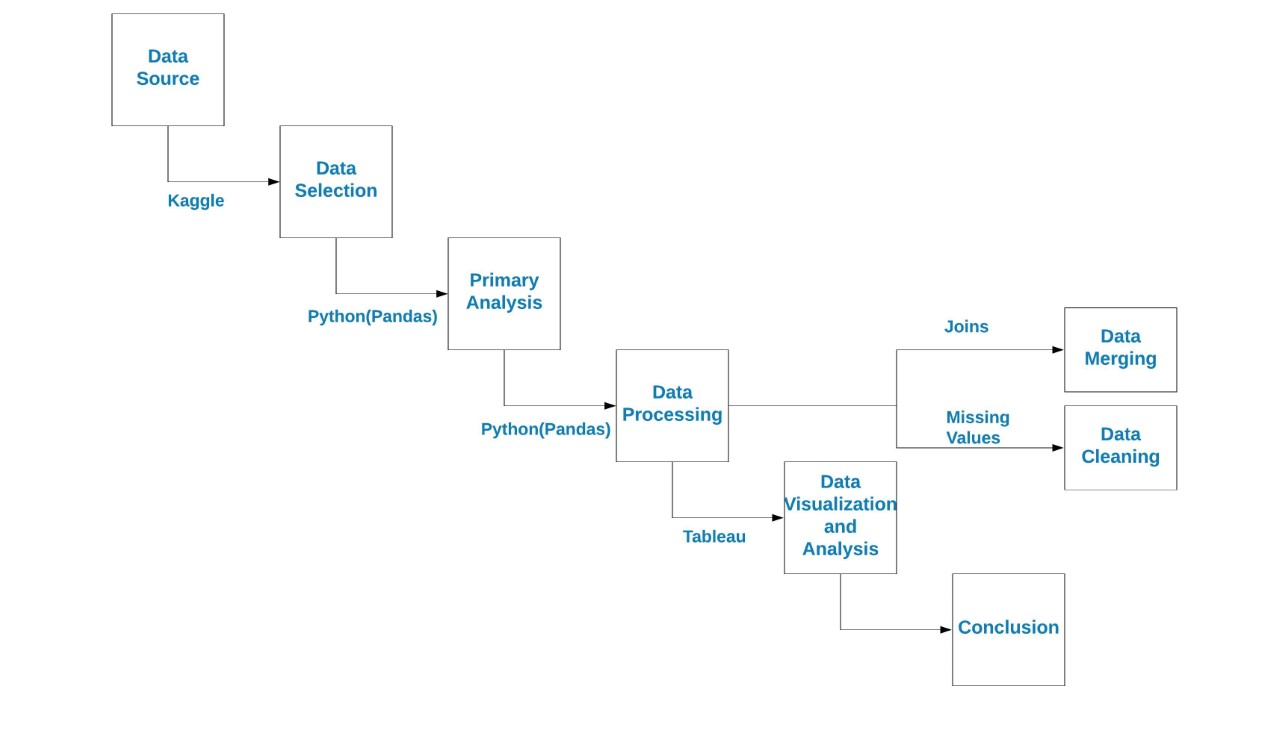
[List of Categorical and Running Variables 11](#_Toc21804077)

[Correlation Matrix 12](#_Toc21804078)

# Introduction & Analysis of Data

Loan delinquency and loan default are becoming a growing problem for the banking and other financial institutions. Failure to recover loans on time puts these loan lending institutions at financial loss. This eventually makes it difficult for people who genuinely need loan but never had any banking history. While it is very difficult to accurately claim who will pay back the loan, it is possible to estimate the probability of person paying back the loan. In order to help answer this question, we obtained data posted by Home Credit on Kaggle. Home credit, a non-bank financial institution focuses primarily on lending loans to people with little or no credit history. The main objective of this project is to prepare a dataset that has necessary variables to predict how likely a person would repay a loan. Correlation studies was also performed between some of the variables in final dataset to check if there existed any significant relation between those variables as this may help with selection of variables for machine learning model. The results from this study can also be used to identify the factors that are common to people who face a difficulty in paying back the loan.

## Flow Chart

****

This flow chart explains the life cycle of our project from data gathering to data visualization. There are total 6 stages in this process.

## Data Source

Data was downloaded from online website “Kaggle” which is a platform for data science and data analytics competitions.

## Data Selection

It is the process of data collection. We did some research on what type different datasets when collected and combine produce some business value. We decided to go with financial data provided by Home Credit as combined data would help predict the loan pay back ability of the loan applicant.

## Primary Analysis

This is the initial analysis process where the collected data is analyzed, and the required data is selected from the raw data. This is the step where we decided which columns of data to be used for our future analysis.

## Data Processing

It is the process of data manipulation where we removed, added and operated on data as per our project needs. In this case the data processing was done in two stages: Merging and Cleaning. **Merging** is the process of combing different datasets which we have initially analyzed, and **Cleaning** is the step which is performed before or after merging, which includes handling missing values adding and removing the data as required. We used “pandas” library for the processing, merging and cleaning.

## Data Visualization

It is the graphical representation of the data which shows the relationship between dependent and independent variables, independent and independent variables. We found the correlation between different variables, and plotted different graphs representing them. We used Tableau platform for the graphical representation of the data.

# Data Merging

Data merging is the process of combining two or more datasets into a single dataset. It is most commonly used when we have multiple raw datasets, and we need to analyze all of them using single dataset. It also helps to combine all the distributed information through datasets, and merge them to fulfill the requirements of the project. In our project, we choose 4 datasets out of 6 datasets which we merged according to our requirements.

## Common Elements

Firstly, we merged all the datasets with an inner join method using a common column named “SK\_ID\_CURR”, which combined all the common data from all 4 datasets. Secondly, we combined all the 4 datasets using join method (left join) on loan applicant application We used this approach because there were some loan application data missing when we merged the datasets using inner join. So, we used left join method to combine the datasets which gave us all the loan application information. However, it was missing information from bureau, previous application and credit balance information who had not previous banking history.

## Multilevel data

There were some multilevel data in our datasets where the “SK\_ID\_CURR” had multiple values in some datasets: Previous application data, Credit bureau data and Credit balance data. There were multiple records of each customer. For example, in Credit bureau dataset we had “Active credit” column which had different categorical data. We splitted the unique values of “Active credit” and created a column for each category, and while grouping on “SK\_ID\_CURR” counted the values of category and added them. This was our primary tool for solving multilevel data issues.

## Combined Values of data

In our datasets, there were columns like Amt balance, Amt credit limit actual, Amt credit and Amt goods price which had more value when combined into one dataset compared to single dataset. This was signified with the positive correlation between Amt balance and Amt credit limit actual, Amt credit and Amt goods. After merging the datasets, we found certain relationships between data that made the analysis more promising and relevant. For example, when there was an increase in the amount of goods for products for which the loan was applied for, there was also an increase in the amount of loan applied by the customers.

# Data Cleaning

Data Cleaning is the process of detecting the flaws and errors in the data and clean or correct it. Data cleaning plays critical role in every project where data is the key element. Clean data enhances the analysis accuracy.

If our dataset contains incorrect, incomplete, inaccurate and irrelevant parts, we can resolve it by modifying, correcting or deleting the data.

Initially we had 6 datasets from the source but based on our primary analysis we decided to select 4 datasets.

* Current applications data
* Previous application data
* Credit bureau data
* Credit balance data

## Primary Cleaning of Data

The primary cleaning includes the steps like slicing and extracting the more relevant data from the available data so that we can perform further steps on data munging.

The initial data which we got from Home Credit contains:

* Current applications data (122 columns)
* Previous application data (37 columns)
* Credit bureau data (17 columns)
* Credit balance data (23 columns)

There were some unwanted and repeated columns which we determined to be irrelevant to the requirements of our project. Some columns were dropped and only a portion of data was selected for further processing.

The datasets after primary cleaning:

* Current applications data (21 columns)
* Previous application data (5 columns)
* Credit bureau data (5 columns)
* Credit balance data (3 columns)

## Quality of Data

As our data was from a competitive source, it was in the raw form as provided by the Home Credit. The quality of data determines how satisfactory the data is relation to the requirements of the project. So, data from Home Credit was relevant and reliable as it was providing the current applications data, previous application data, credit bureau data and credit balance data. The final dataset after merging the 4 datasets incudes both numerical and categorical data. It contains 22 numeric and 12 categorical data columns.

## Missing Data

The missing data is created when data is not included in its initial data collection process, and when we combine datasets. It is a very common occurring problem and has a great impact on analysis results. In some cases, it may lead to derive a false conclusion. In our final dataset, there were many missing values which we treated with different techniques.

In our final dataset, following columns were missing values:

* AMT\_GOODS\_PRICE
* OCCUPATION\_TYPE
* CNT\_FAM\_MEMBERS
* AMT\_REQ\_CREDIT\_BUREAU\_YEAR

Occupation type column had categorical data for which we cannot use the operations which we can use in other 3 numerical type columns. Missing values in this column were not initially present in the dataset but these this information play an important role in our dataset value, hence we decided not to delete these records.

**Solution**

We replaced all the missing values in this column with text “unknown” because there were no records present and manipulating data to fill these missing values would introduce new error.

Two columns:“AMT\_GOODS\_PRICE” and “CNT\_FAM\_MEMBERS” which represents the Actual amount of the goods for which the loan was taken for and number of family members for each customer respectively.

**Solution:** We used the mean value of the columns to fill these missing values because other techniques of missing values were not giving as promising results in comparison to mean value.

Column “AMT\_REQ\_CREDIT\_BUREAU\_YEAR” is the number of credit inquiries for each customer in a year. The missing values in this column state that there were some new customers who did not have any previous credit history.

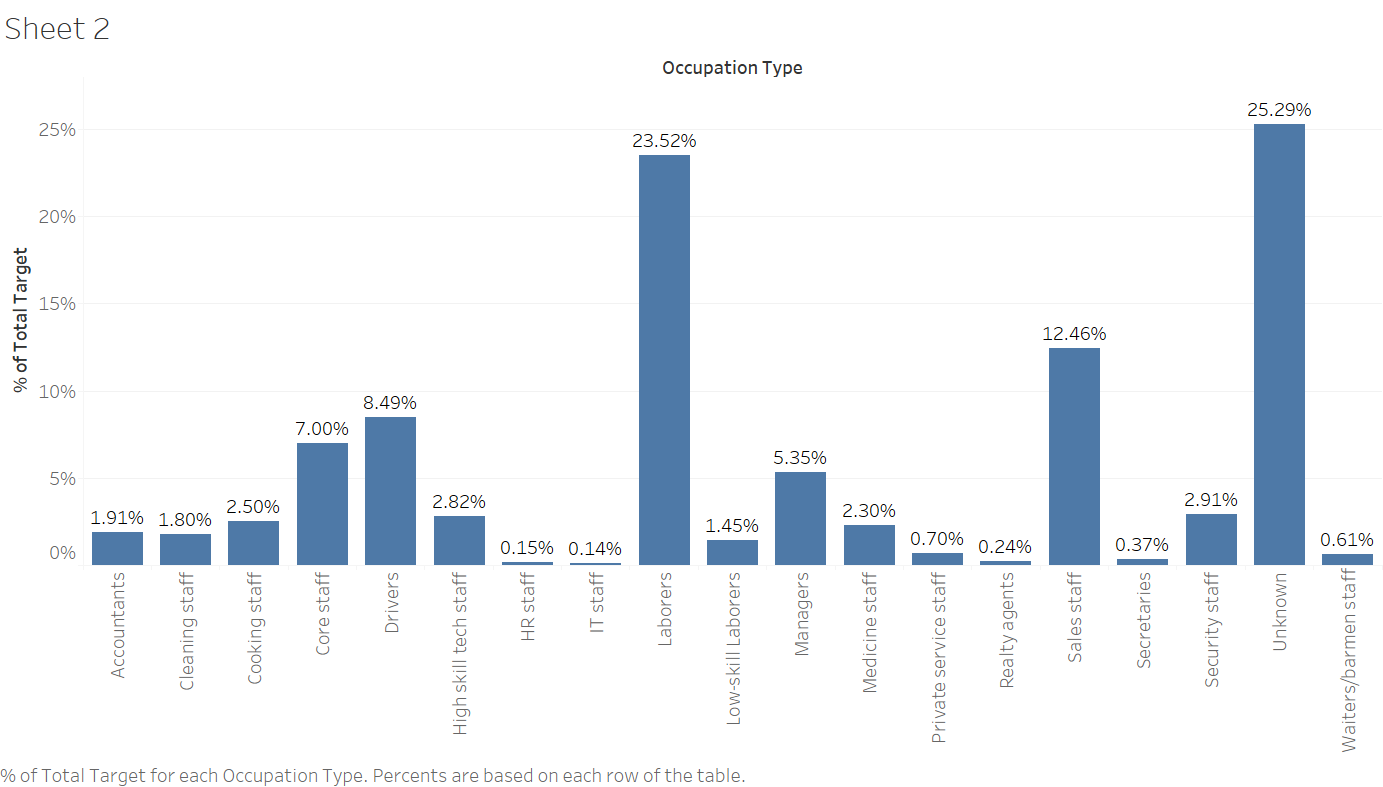
**Solution**

It is possible that some people do not have previous credit history and for this reason we decided to replace all the nan values with “0”.

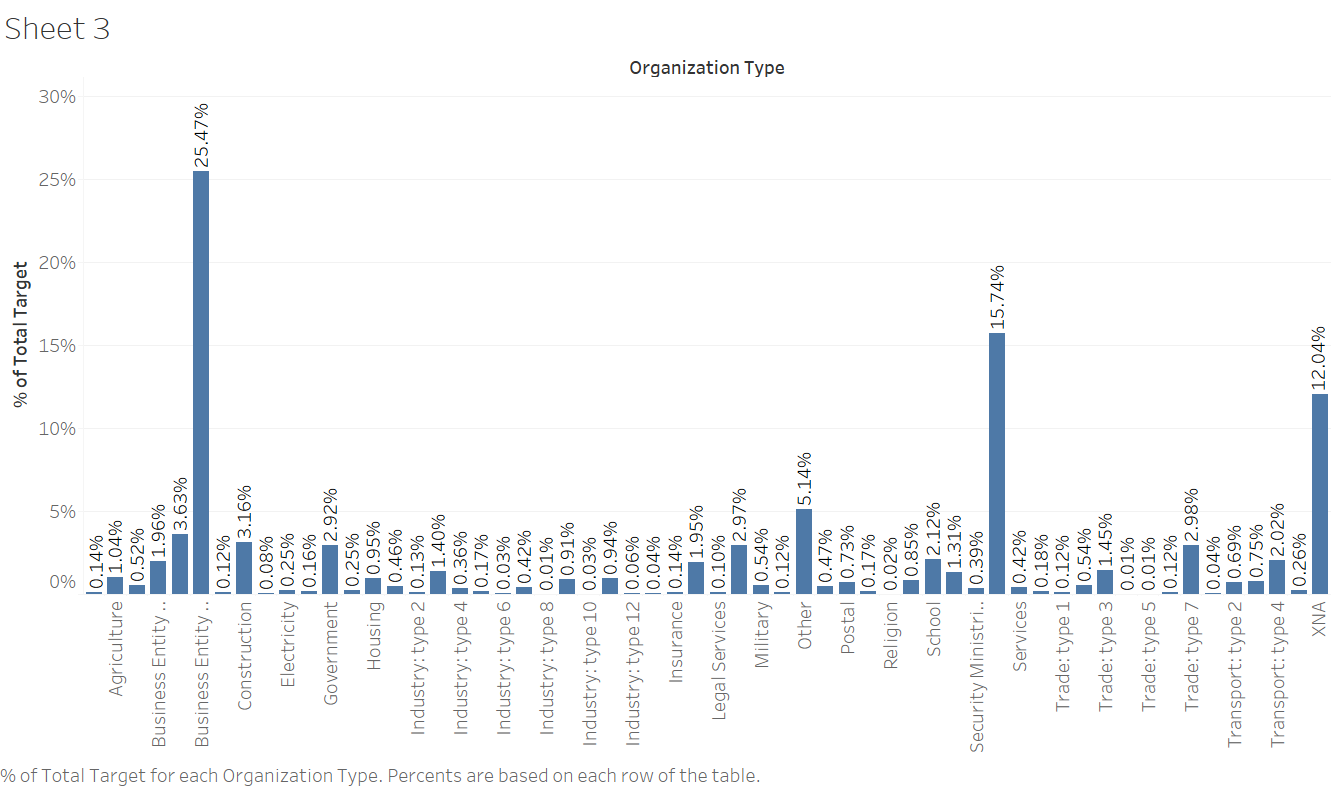
# Analysis of Visualization

“Target” is a dependent variable in our dataset. It takes two values 0(non-risky for loan) and 1(risky to repay loan). Bar plots and scatter plots were used for visualizing categorical and continuous variables respectively in Tableau. While not all plots were created, some of the more interesting ones are plotted below. “% of total target” represents proportion of people who had difficulty paying back loan.

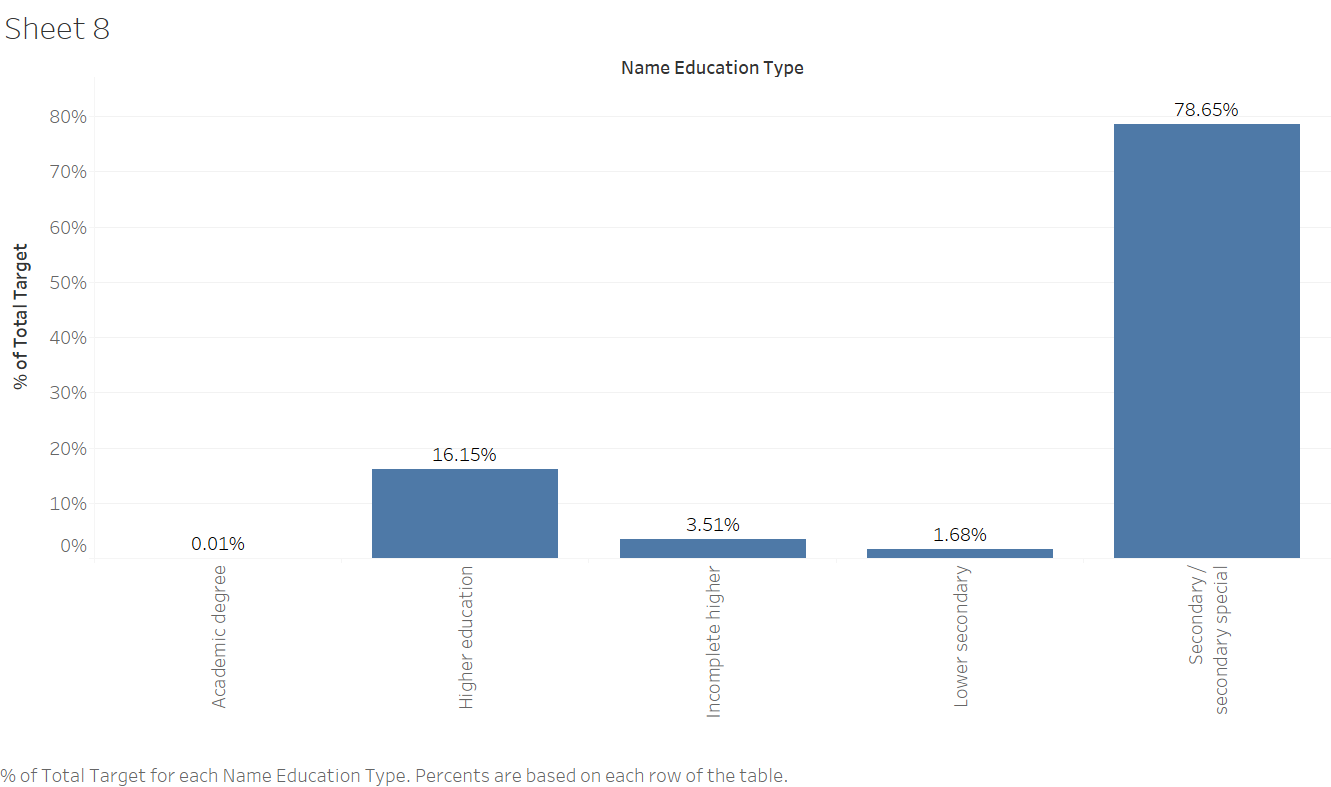
Proportion of total target by occupation type shows that laborers and “unknown” occupation type had the most difficulty paying back the loan. “Unknown” occupation type refers to all the people who had missing values in their occupation type column. We believe that laborers are unskilled people who do manual work for wages and therefore have a difficulty paying back the loan.



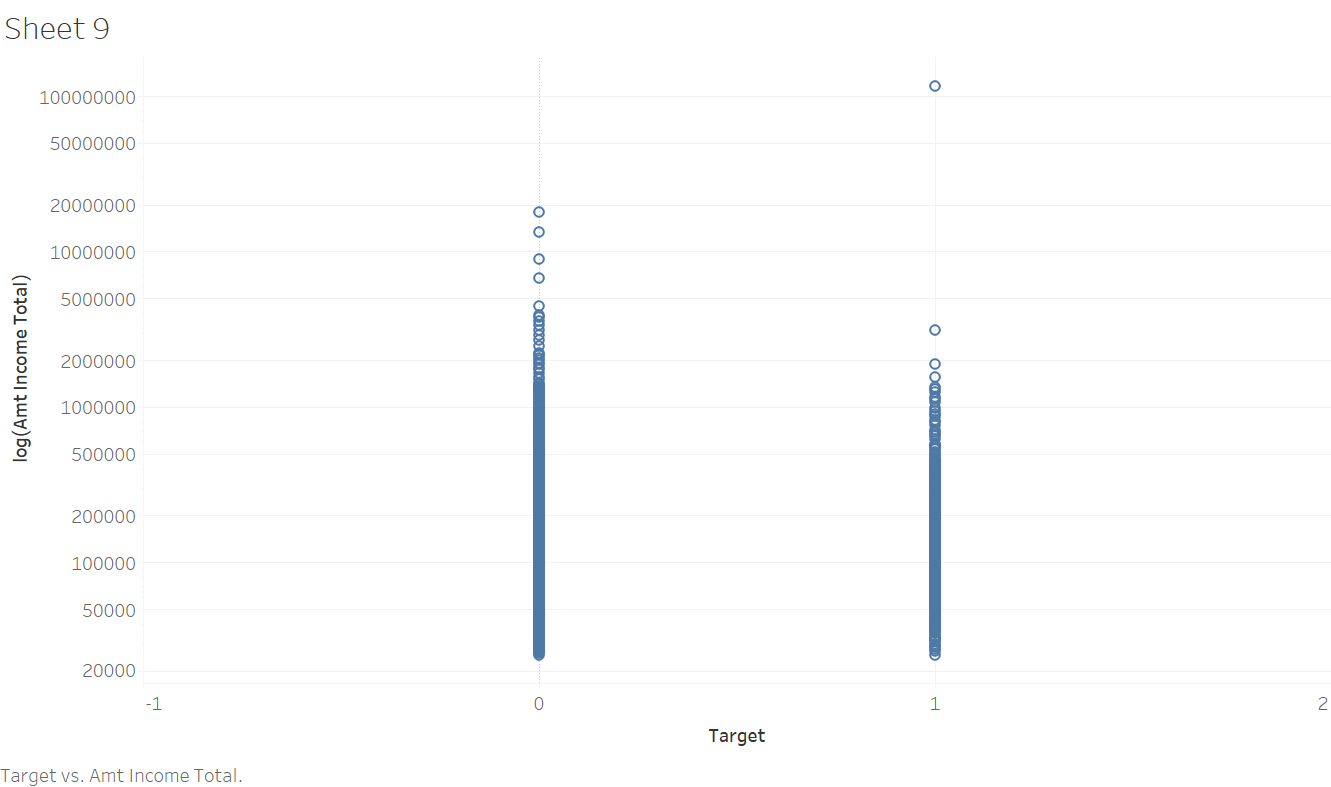
Among all the people who had difficulty paying back the loan, people belonging to business organization type comprised a major portion because there is relatively more uncertainty involved with the outcome of a business and thus people may find it hard to pay loan timely.



Proportion of people by education type shows that people with secondary education had most difficulty paying back the loan.

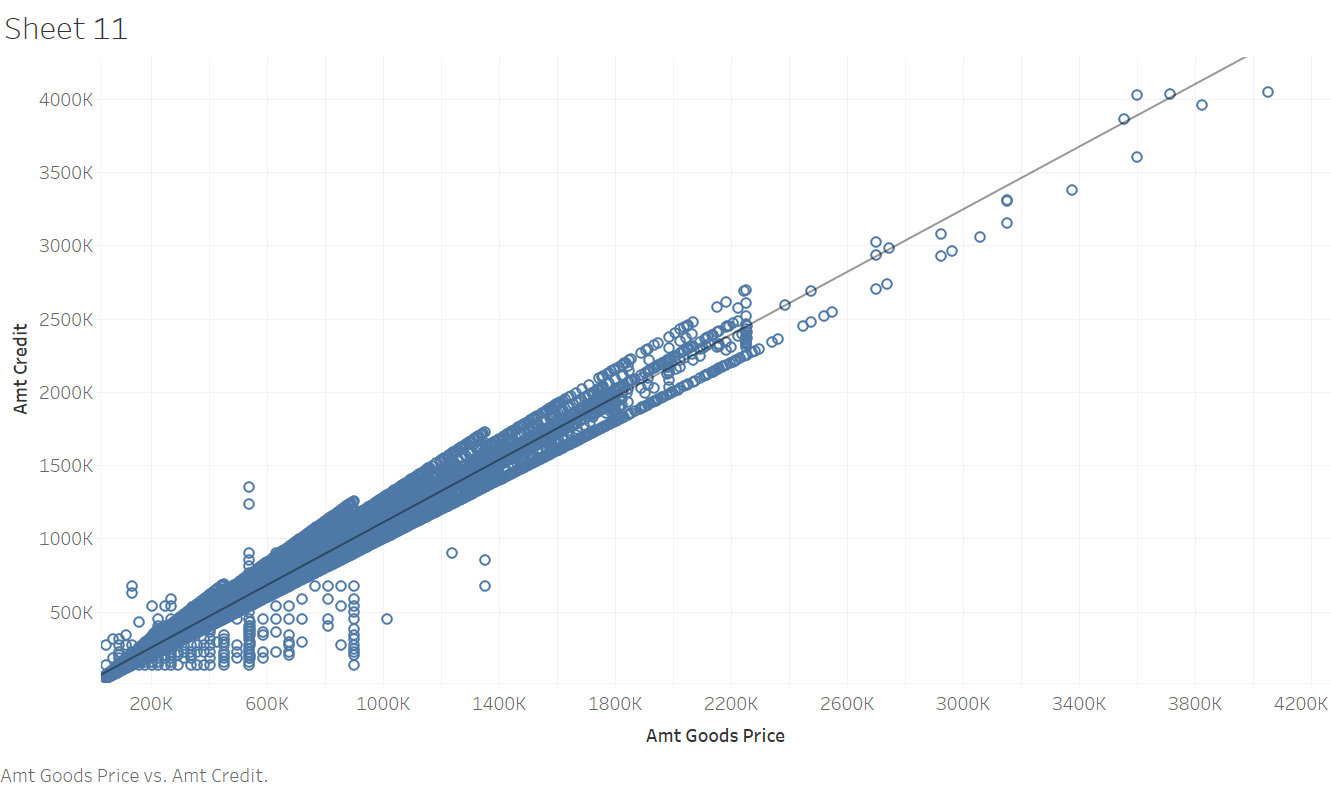


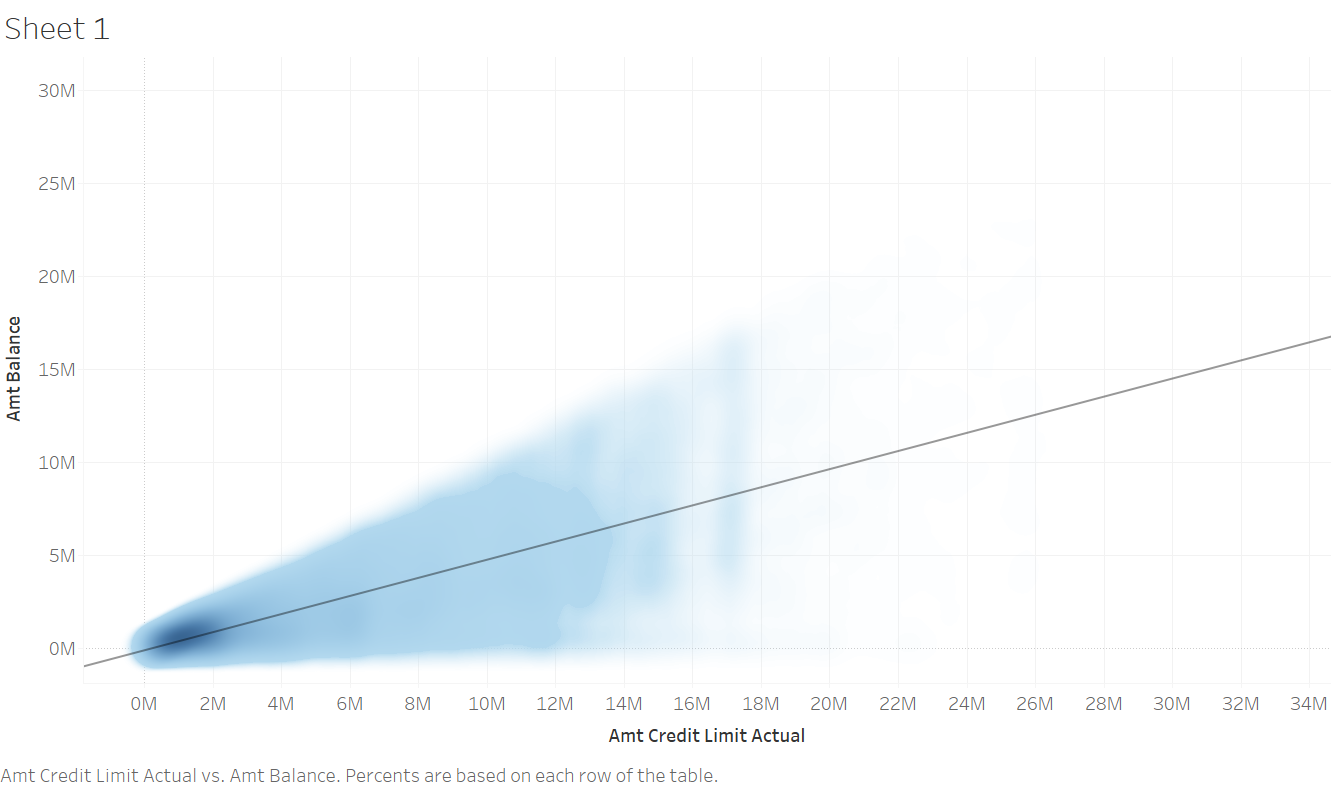
People with lower income were more likely to default the payment of loan as compared to higher income group.



Correlation studies was performed between “TARGET” variable and some of the other variables from the dataset and the r values for these set of variables are tabulated below.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable 1 | Variable type | Correlation type | r- value |
| Occupation type | Categorical | Cramer’s V | 0.08 |
| Organization type | Categorical | Cramer’s V | 0.07 |
| Name Education type | Categorical | Cramer’s V | 0.06 |
| Amt Income Total | Continuous | Point Biserial | -0.004 |





We also wanted to see if there are any highly correlated variables except the target variable in the processed dataset. This would help in dimensionality reduction, a method used to remove variables that do not add too much value in prediction process, for a machine learning engineer. Out of all the variables in dataset, pair of “Amt Credit” and “Amt Goods Price” and pair of “Amt Balance” and “Amt Credit Limit Actual” have high r value of 0.98 and 0.78 respectively and thus only one variable out of one of pair can be used in the machine learning model.

## Principles of good visualization

We were interested in finding out the characteristics that make someone more likely to default a loan and therefore we used “% of Total Target”, a proportion of people who had difficulty paying back loan, as our dependent variable. While plotting the continuous variable for different values of target, we took logarithm of continuous variable to show the distribution more clearly.

# Conclusion

Based on our results, we would recommend not to use linear regression model for building machine learning model and remove one variable out of pair of highly correlated variables. This would make model simpler and efficient. Additional screening can be performed for the loan cases with high risk factors. For example, people with laborers as education type and secondary school as education type may need an additional verification before approving the loan.

# Instruction for code

These are the instructions to follow before running the code.

1. Download the all 4 data set from this link.

Link: <https://www.kaggle.com/c/home-credit-default-risk/data>

1. Value of variable “path” should be changed to location where all 4 datasets are downloaded.
2. Runs the code named “ISQS6339-001-2019-GROUP-3-Final Project Code” on python platform.

# Members

Aman Panwar

Anurag Sharma

Dinesh Poudel

# Name of Datasets

Current applications data

Previous application data

Credit bureau data

Credit balance data

# List of Categorical and Running Variables

|  |
| --- |
| **Continuous** |
| AMT\_INCOME\_TOTAL |
| AMT\_CREDIT | **Categorical** |
| AMT\_GOODS\_PRICE | ORGANIZATION\_TYPE |
| REGION\_POPULATION\_RELATIVE | NAME\_INCOME\_TYPE |
| DAYS\_BIRTH | NAME\_FAMILY\_STATUS |
| DAYS\_EMPLOYED | OCCUPATION\_TYPE |
| AMT\_BALANCE | FLAG\_OWN\_CAR |
| AMT\_CREDIT\_LIMIT\_ACTUAL | CODE\_GENDER |
| Canceled\_1\_CONTRACT\_STATUS | FLAG\_CONT\_MOBILE |
| CNT\_FAM\_MEMBERS | NAME\_CONTRACT\_TYPE |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | FLAG\_OWN\_REALTY |
| CREDIT\_1\_Closed | NAME\_EDUCATION\_TYPE |
| CREDIT\_1\_Bad debt |
| Refused\_1\_CONTRACT\_STATUS |
| REGION\_RATING\_CLIENT |
| CREDIT\_1\_Active |
| Approved\_1\_CONTRACT\_STATUS |
| CREDIT\_1\_Sold |
| Unused offer\_1\_CONTRACT\_STATUS |

# Correlation Matrix

