Impact of Credit Utilization on Loan Interest Rates

Abstract

In this study, we seeks to investigate the connection between credit utilisation and loan interest rates using a loan dataset. The interest rates on credit cards can appear absurd because they are far higher than those on mortgages or vehicle loans, with some of them exceeding 20% APR. In finance, we may typically expect a better potential reward if we take on more risk. Credit cards pose a significant risk to banks and other card issuers since many users make late or non-payments. As a result, issuers impose hefty interest rates to offset that risk. Principal component analysis (PCA), Pearson correlation coefficient, and linear regression are the three machine learning techniques used in the study to process and analyse the data. The PCA results show a clear association between the variables, the first PCA is essentially orthogonal to the and both variables. With a statistically significant p-value, the Pearson correlation coefficient shows a positive association between credit utilisation and loan interest rates. The low R-squared value of 0.2005, obtained from the linear regression, shows that the model only explains 20.05% of the variation in the response variable. According to the study's findings, credit utilisation does affect loan interest rates and can assist both consumers and lenders make informed decisions about their credit policies.

Introduction

Do consumers who have high credit use pay higher loan interest rates? Along with many other advantages, credit cards offer the chance to establish credit history and obtain a credit score. However, if a user has a high credit usage rate on their cards, they may experience worse credit ratings, difficulty paying larger monthly payments, and a higher interest rate on their cards should they ever make a late payment (What is a Credit Utilization Rate?, 2023). Credit usage measures how much of a user's available credit is being used to pay down existing credit card bills. It calculates how much of the user's available credit is being used. Lenders may view high credit use as a risk factor and charge clients higher interest rates on loans as a result. Using a loan dataset, this study seeks to investigate the relationship between credit utilisation and loan interest rates. The results of this study will provide insight on the factors that influence loan interest rates and could be helpful to both borrowers and lenders in deciding on credit policies (Irby, 2022).

Data

Massive amounts of data are being obtained, and machine learning tools like Python, R studio, and Excel are being used to collect data information across the raw data process. To filter, alter, organise, and manage data in order to provide relevant information on the impact of credit use on loan interest rates, this study makes use of the characteristics of three machine learning algorithms.

The author ITSSURU provided the website Kaggle with the data that was used in this investigation (2020). The loan data file was included, and it was downloaded in "csv" format. This dataset was first released to the general public by LendingClub.com. Lending Club matches lenders with borrowers, who are those who need money (investors). The information comes from lending records between 2007 and 2010, and it aims to categorise and forecast whether or not a borrower will repay their loan in full.

The dataset consists of 9,578 rows with 14 variables consist of 7 integer (binary) variables 6 number variables and 1 character variable . The definitions of all the attributes will be illustrated in appendices for detailed knowledge. Data cleaning is the act of locating, preparing, and validating data (Hillier, 2021). However, as we can see from the graph produced by the DataExplorer function, this dataset does not include any missing data.

Before moving on to the method, this study has scaled the data and divided it into a test and training set in an 8:2 ratio to ensure that the model can generalise effectively to brand-new, unexplored data. This ensures that every variable is scaled equally, which is essential for the effective operation of many machine learning algorithms and make it possible to compare variables fairly and prevents bias against variables with larger magnitudes.

Exploratory Analysis

On the correlation matrix, we can see that based on this dataset, *revol.util* and *int.rate* has the highest correlation in this dataset of 0.46. After removing outliers, the boxplot displays the distribution of the data for each "purpose" category and the relationship between the two variables, *revol.util* and *int.rate* within those categories. It gives an idea of the median, quartiles, minimum, and maximum values of the data for each category. The graph shows the wider range between the minimum and maximum numbers further demonstrates the more dispersed nature of the interest rate distribution for "small business" loans. The loans for "debt consolidation" also seem to have had the second highest median interest rate but the narrowest spread between the minimum and maximum values.

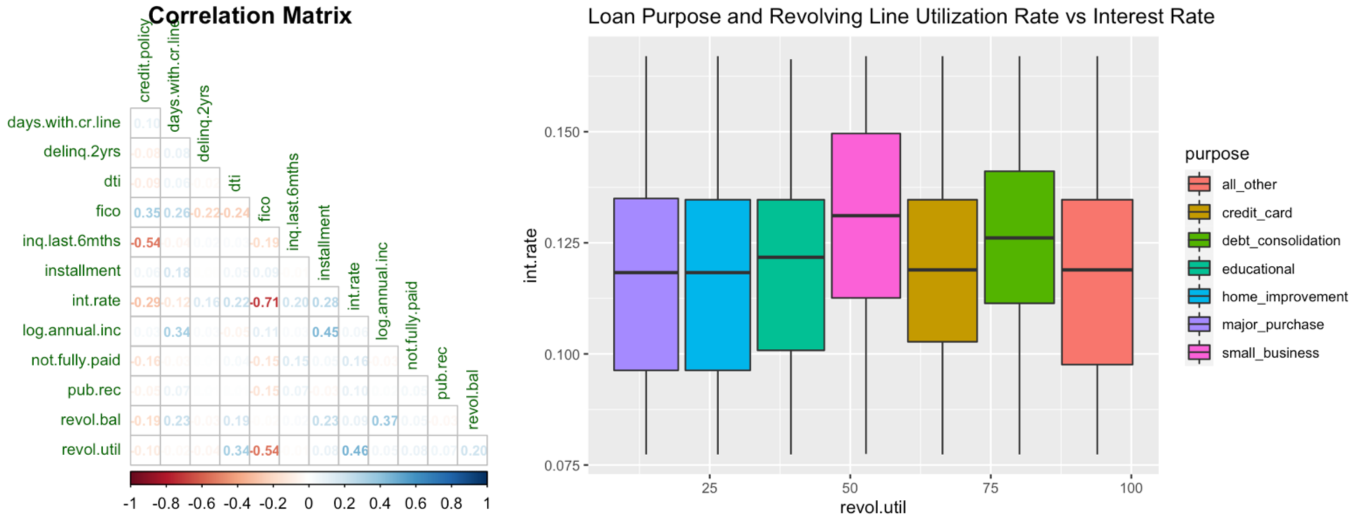


Figure 1.0: Correlation Matrix (on left), Box-plot of Loan Purpose and Revolving Line Utilization Rate vs Interest Rate (on right)

Methods

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

In this study, data processing, statistical analysis, and data visualisation are all done using RStudio. In order to generate conditional expressions, iterative operations, insight directions, and built-in technicians for estimations using arrays and matrices, it provides a toolset for both basic and complicated statistical procedures. The research uses a number of distinct approaches to establish its goal.

Three different approaches will be used in this study to determine the study's goal. A method known as principal component analysis (PCA) reduces the number of dimensions in big data sets by condensing a vast collection of variables into a smaller set while retaining the majority of the large set's information. In addition, the Pearson correlation coefficient runs from -1 to 1, where -1 denotes a perfect deficiency in correlation, 0 denotes a lack thereof, and 1 denotes a perfect surplus in correlation. The last type of method is liner regression, which is typically employed to shed light on the relationship between a continuous dependent variable and one independent variable. The appendices contain information on all procedures and codes.

Results and Discussion

The findings of this research will be demonstrated in this section using a variety of machine learning techniques to spot trends and predict likely outcomes.

Principal Component Analysis (PCA)

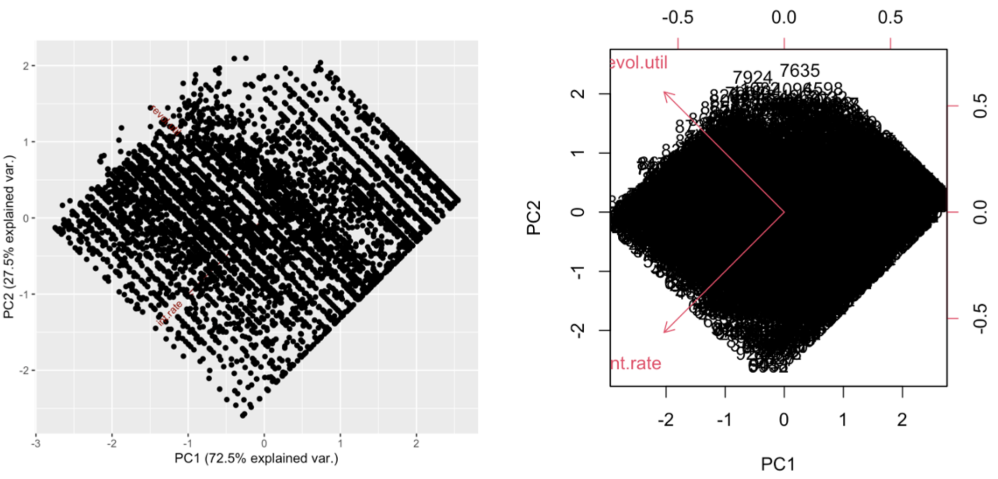
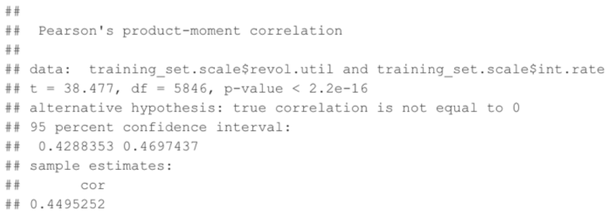


Figure 2.0: PCA ggbiplot and bi-plot evaluating of revol.util and int.rate (PC2 vs PC1)

The relationship between the variables is made clearer by graphing them. In Figure 1.0, PCA1 accounts for 72.5% of the entire variance, PCA2 for 27.5%, and both PCA together account for 100% of the variance. The graph indicates that the of *revol.util* and *int.rate* variables point in the same direction as PCA2. The first PCA is essentially orthogonal to the and both variables.

Pearson correlation coefficient

The Pearson correlation coefficient, which measures a linear relationship, is the most widely used approach. A number between -1 and 1 is used to describe the strength and direction of the relationship between two variables. Assuming a linear relationship between the two variables, this correlation coefficient is appropriate for usage in this dataset (Turney, 2022).



```
##
##  Pearson's product-moment correlation
##
## data:  training_set.scale$revol.util and training_set.scale$int.rate
## t = 38.477, df = 5846, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4288353 0.4697437
## sample estimates:
##       cor
## 0.4495252
```

Figure 3.0: Result of Pearson correlation coefficient

Based on this result above the correlation coefficient in this instance is 0.4495252, indicating a positive correlation, which means that as the *revol.util* rises, so does the *int.rate*. Since the correlation has a p-value that is less than 0.05, it is statistically significant. This is also supported by the 95% confidence interval, which has a range of 0.4288353 to 0.4697437

Linear Regression

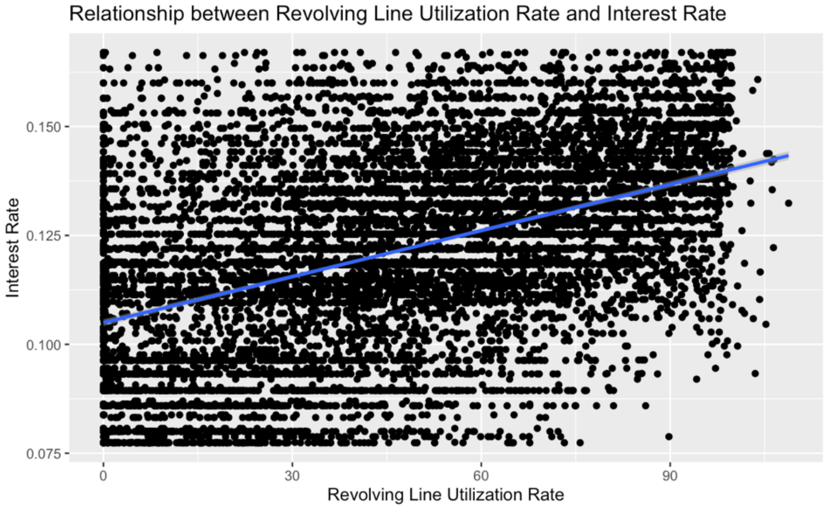
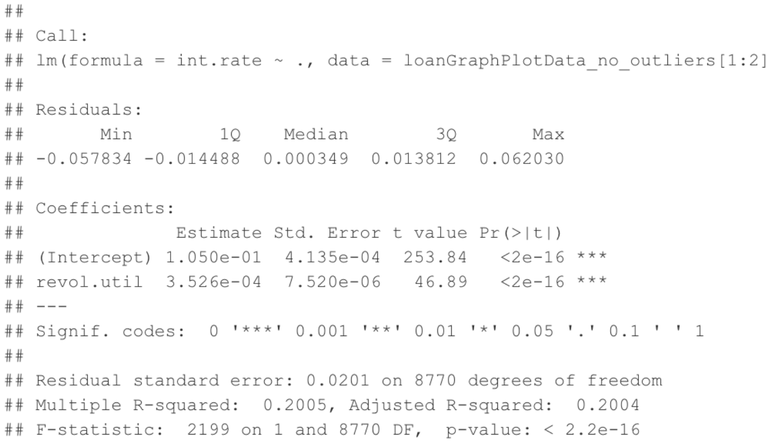


Figure 4.0: Relationship between Revolving Line Utilization Rate and Interest Rate

As mentioned before, *int.rate* is utilised as the dependent variable to assess the importance of the independent variables. Therefore, in Figure 3.0, the y-axis is *int.rate*, while the x-axis is *revol.util*. There is a rising trend that is fairly densely clustered. This implies that the interest rate on the loan rises in conjunction with the borrower's revolving line utilisation rate.



```
##
## Call:
## lm(formula = int.rate ~ ., data = loanGraphPlotData_no_outliers[1:2]
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057834 -0.014488  0.000349  0.013812  0.062030
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.050e-01  4.135e-04  253.84   <2e-16 ***
## revol.util  3.526e-04  7.520e-06   46.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0201 on 8770 degrees of freedom
## Multiple R-squared:  0.2005, Adjusted R-squared:  0.2004
## F-statistic:  2199 on 1 and 8770 DF,  p-value: < 2.2e-16
```

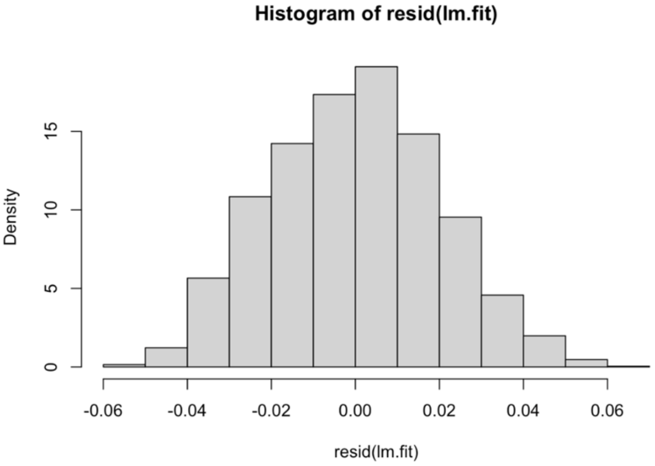


Figure 4.1: Linear Regression Output (on left), Histogram of Linear Model (on right)

A regression model should be assessed for a variety of important factors, including R2, significance F, variable coefficients, and p-value, as seen in the output above. The coefficient of 1.050e-01 indicates that we anticipate an increase in *int.rate* of 3.526e-04 for every unit increase in *revol.util*.

Given that the residuals appear to be fairly regularly distributed based on the statistics for the residuals that are presented in the output. The residuals have a median value of 0.000349, a range of -0.014488 to 0.062030, and a minimum value of -0.057834. Furthermore, the residual standard error is 0.0201 on 8770 degrees of freedom, which is negligibly low and suggests that the residuals are reasonably near to the fitted values. Since the residuals are roughly symmetrical, the residuals' histogram appears to be somewhat bell-shaped.

Consider the idea that "the higher the F value, the better the model." We can see from this result that the F-statistic yields a significant large value of 2199 and a p-value of almost zero. This result allows we to reject the null hypothesis while accepting the alternative hypothesis.

Conclusions

As a result, the study's findings reveal a significant relationship between the *int.rate* and the *revol.util*. The p-value of an almost zero result and the correlation coefficient of 0.4495252, determined using Pearson's product-moment correlation, both point to a significant correlation between the two variables. Furthermore, the linear regression model generated an adjusted R-squared value of 0.2005, showing that variations in the *revol.util* can account for 20.1% of the variation in the *int.rate*. These results imply that lenders may view borrowers with greater revolving line utilisation rates as higher risk clients and may therefore charge them higher interest rates for loans. With this knowledge, borrowers can better understand how their credit utilisation may affect the interest rate they are offered on a loan. Lastly, this analysis has shown that liner regression is the most effective technique for precisely forecasting the relationship between the borrower's revolving line utilisation rate and the loan's interest rate. The F-statistic produces a significant large number and a low p-value, which lends support to the null hypothesis being rejected and accepting the alternative hypothesis.

References

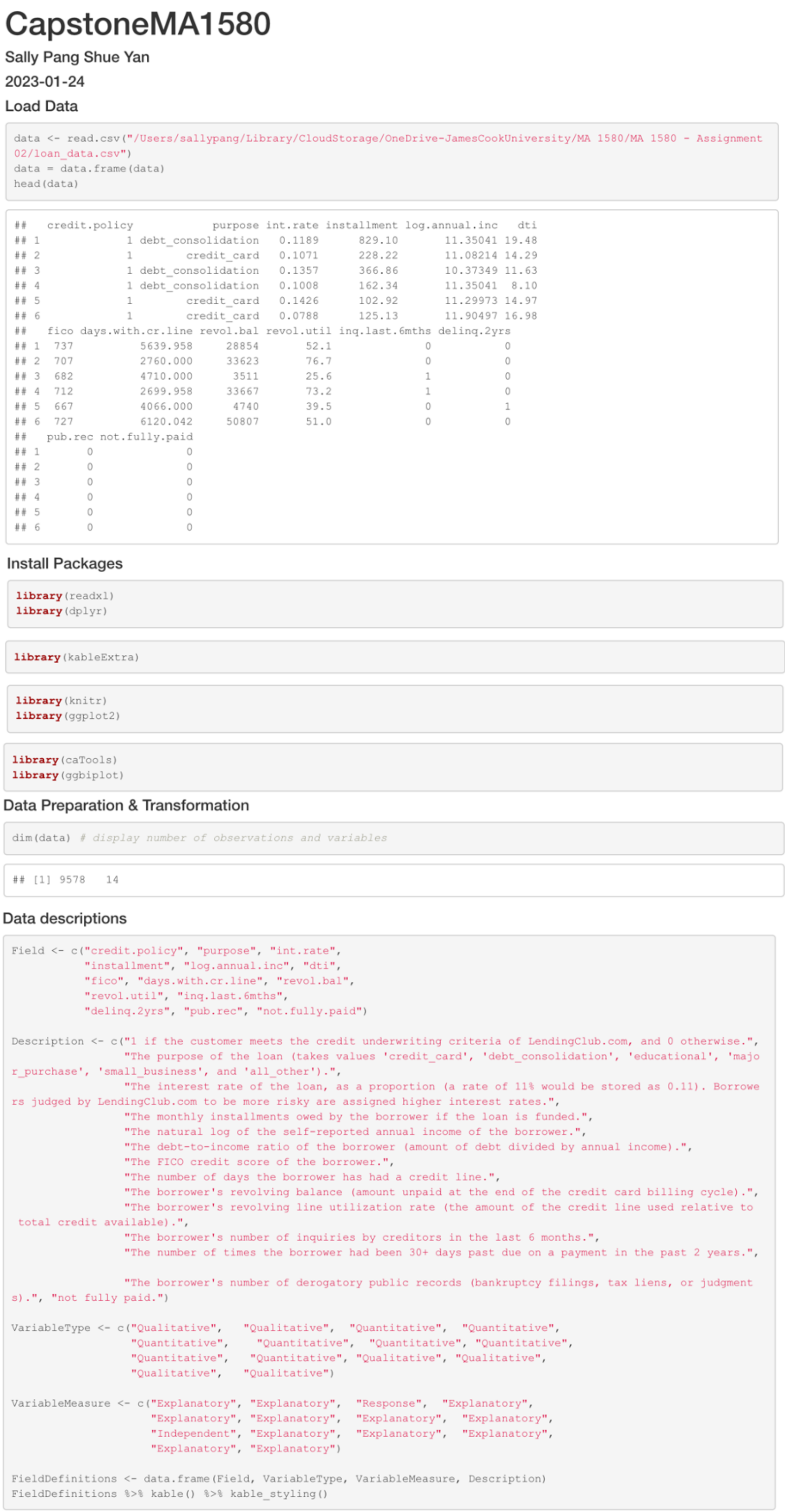
What is a Credit Utilization Rate? (2023). Retrieved from experian: https://www.experian.com/blogs/ask-experian/credit-education/score-basics/credit-utilization-rate/

Irby, L. (20 February, 2022). *Understanding Credit Utilization*. Retrieved from the balance: https://www.thebalancemoney.com/understanding-credit-utilization-960451

Hillier, W. (12 November, 2021). *What Is Data Cleaning and Why Does It Matter?* Retrieved from careerfoundry: https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/

Turney, S. (13 May, 2022). *Pearson Correlation Coefficient (r) | Guide & Examples*. Retrieved from scribbr: https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Appendices



CapstoneMA1580

Sally Pang Shue Yan

2023-01-24

Load Data

```
data <- read.csv("/Users/sallypang/Library/CloudStorage/OneDrive-JamesCookUniversity/MA 1580/MA 1580 - Assignment
02/loan_data.csv")
data = data.frame(data)
head(data)
```

```
##   credit.policy            purpose int.rate installment log.annual.inc   dti
## 1             1 debt_consolidation   0.1189      829.10       11.35041 19.48
## 2             1        credit_card   0.1071      228.22       11.08214 14.29
## 3             1 debt_consolidation   0.1357      366.86       10.37349 11.63
## 4             1 debt_consolidation   0.1008      162.34       11.35041  8.10
## 5             1        credit_card   0.1426      102.92       11.29973 14.97
## 6             1        credit_card   0.0788      125.13       11.90497 16.98
##   fico days.with.cr.line revol.bal revol.util inq.last.6mths delinq.2yrs
## 1  737          5639.958     28854       52.1              0           0
## 2  707          2760.000     33623       76.7              0           0
## 3  682          4710.000      3511       25.6              1           0
## 4  712          2699.958     33667       73.2              1           0
## 5  667          4066.000      4740       39.5              0           1
## 6  727          6120.042     50807       51.0              0           0
##   pub.rec not.fully.paid
## 1       0              0
## 2       0              0
## 3       0              0
## 4       0              0
## 5       0              0
## 6       0              0
```

Install Packages

```
library(readxl)
library(dplyr)
```

```
library(kableExtra)
```

```
library(knitr)
library(ggplot2)
```

```
library(caTools)
library(ggbiplot)
```

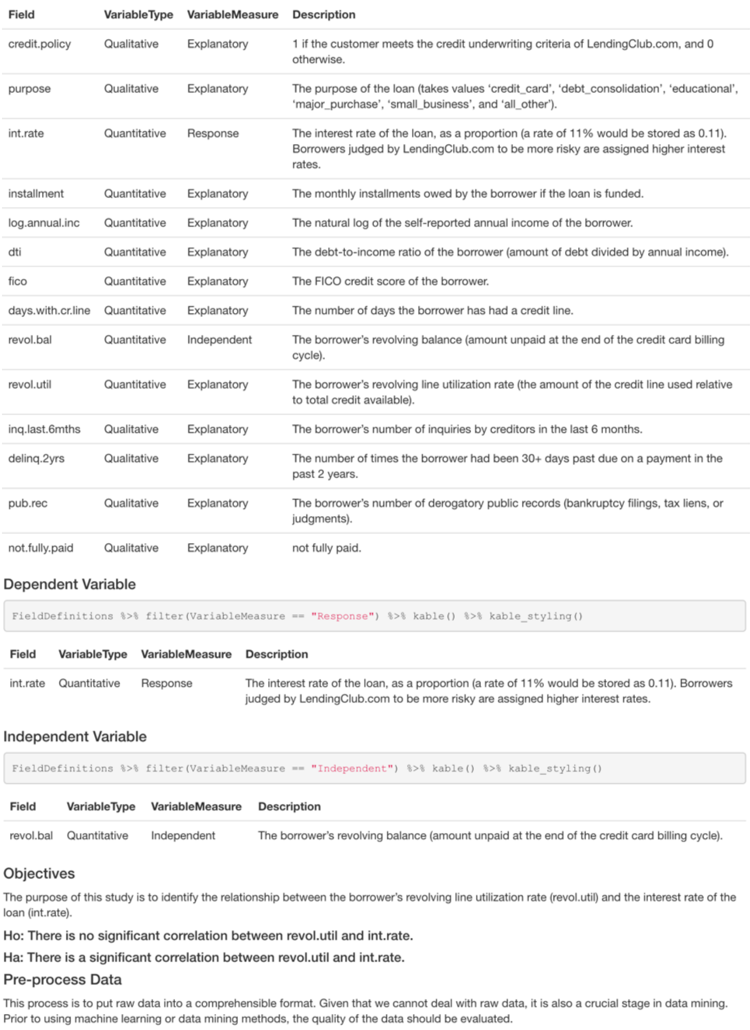
Data Preparation & Transformation

```
dim(data) # display number of observations and variables
```

```
## [1] 9578   14
```

Data descriptions

```
Field <- c("credit.policy", "purpose", "int.rate",
           "installment", "log.annual.inc", "dti",
           "fico", "days.with.cr.line", "revol.bal",
           "revol.util", "inq.last.6mths",
           "delinq.2yrs", "pub.rec", "not.fully.paid")

Description <- c("1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.",
                 "The purpose of the loan (takes values 'credit_card', 'debt_consolidation', 'educational', 'majo
r_purchase', 'small_business', and 'all_other').",
                 "The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowe
rs judged by LendingClub.com to be more risky are assigned higher interest rates.",
                 "The monthly installments owed by the borrower if the loan is funded.",
                 "The natural log of the self-reported annual income of the borrower.",
                 "The debt-to-income ratio of the borrower (amount of debt divided by annual income).",
                 "The FICO credit score of the borrower.",
                 "The number of days the borrower has had a credit line.",
                 "The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).",
                 "The borrower's revolving line utilization rate (the amount of the credit line used relative to
 total credit available).",
                 "The borrower's number of inquiries by creditors in the last 6 months.",
                 "The number of times the borrower had been 30+ days past due on a payment in the past 2 years.",

                 "The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgment
s).", "not fully paid.")

VariableType <- c("Qualitative",   "Qualitative",  "Quantitative",  "Quantitative",
                  "Quantitative",    "Quantitative",  "Quantitative", "Quantitative",
                  "Quantitative",   "Quantitative", "Qualitative", "Qualitative",
                  "Qualitative",   "Qualitative")

VariableMeasure <- c("Explanatory", "Explanatory",  "Response",  "Explanatory",
                     "Explanatory", "Explanatory",  "Explanatory",  "Explanatory",
                     "Independent", "Explanatory",  "Explanatory",  "Explanatory",
                     "Explanatory", "Explanatory")

FieldDefinitions <- data.frame(Field, VariableType, VariableMeasure, Description)
FieldDefinitions %>% kable() %>% kable_styling()
```



Field	VariableType	VariableMeasure	Description
credit.policy	Qualitative	Explanatory	1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
purpose	Qualitative	Explanatory	The purpose of the loan (takes values 'credit_card', 'debt_consolidation', 'educational', 'major_purchase', 'small_business', and 'all_other').
int.rate	Quantitative	Response	The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
installment	Quantitative	Explanatory	The monthly installments owed by the borrower if the loan is funded.
log.annual.inc	Quantitative	Explanatory	The natural log of the self-reported annual income of the borrower.
dti	Quantitative	Explanatory	The debt-to-income ratio of the borrower (amount of debt divided by annual income).
fico	Quantitative	Explanatory	The FICO credit score of the borrower.
days.with.cr.line	Quantitative	Explanatory	The number of days the borrower has had a credit line.
revol.bal	Quantitative	Independent	The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
revol.util	Quantitative	Explanatory	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
inq.last.6mths	Qualitative	Explanatory	The borrower's number of inquiries by creditors in the last 6 months.
delinq.2yrs	Qualitative	Explanatory	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub.rec	Qualitative	Explanatory	The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).
not.fully.paid	Qualitative	Explanatory	not fully paid.

Dependent Variable

```
FieldDefinitions %>% filter(VariableMeasure == "Response") %>% kable() %>% kable_styling()
```

Field	VariableType	VariableMeasure	Description
int.rate	Quantitative	Response	The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.

Independent Variable

```
FieldDefinitions %>% filter(VariableMeasure == "Independent") %>% kable() %>% kable_styling()
```

Field	VariableType	VariableMeasure	Description
revol.bal	Quantitative	Independent	The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).

Objectives

The purpose of this study is to identify the relationship between the borrower's revolving line utilization rate (revol.util) and the interest rate of the loan (int.rate).

Ho: There is no significant correlation between revol.util and int.rate.

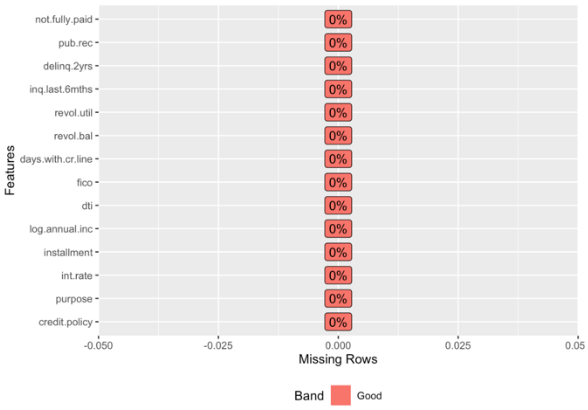
Ha: There is a significant correlation between revol.util and int.rate.

Pre-process Data

This process is to put raw data into a comprehensible format. Given that we cannot deal with raw data, it is also a crucial stage in data mining. Prior to using machine learning or data mining methods, the quality of the data should be evaluated.

Clean Data

```
plot_missing(data)
```



The plot's results demonstrate that the dataset contains no missing data.

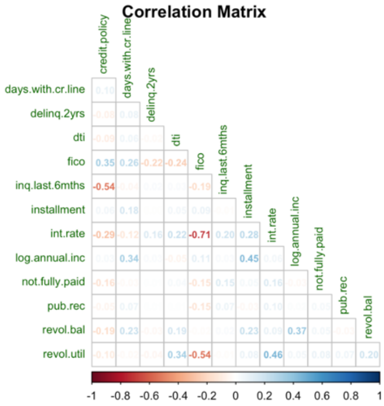
Exploratory Analysis

This section helps to explore the dataset to understand the dataset before making any assumptions.

Correlation Diagram

A correlation matrix is a table showing the correlation coefficients between multiple variables. Each cell in the table represents the correlation between two variables. The correlation coefficient is a value between -1 and 1 that measures the strength and direction of the linear relationship between two variables. A correlation matrix can be used to identify which variables are highly correlated with one another.

```
cor_mat <-
  data %>%
  select(where(is.numeric), -c(purpose)) %>%
  cor(use = "pairwise.complete.obs")

corrplot(
  title = "\n\nCorrelation Matrix",
  cor_mat,
  method = "number",
  order = "alphabet",
  type = "lower",
  diag = FALSE,
  number.cex = 0.7,
  tl.cex = 0.8,
  tl.col = "darkgreen",
  addgrid.col = "gray")
```



Type Coversion

```
data$purpose<-as.factor(data$purpose)
data$int.rate<-as.numeric(data$int.rate)
data$revol.util<-as.numeric(data$revol.util)
# install.packages("magrittr")
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.2
```

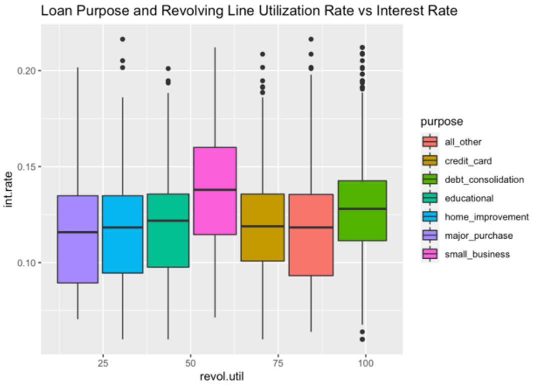
```
library(dplyr)
loanGraphPlotData <-data %>% select(revol.util,int.rate, purpose)
head(loanGraphPlotData)
```

```
##   revol.util int.rate            purpose
## 1       52.1   0.1189 debt_consolidation
## 2       76.7   0.1071        credit_card
## 3       25.6   0.1357 debt_consolidation
## 4       73.2   0.1008 debt_consolidation
## 5       39.5   0.1426        credit_card
## 6       51.0   0.0788        credit_card
```

Box-plot

The boxplot displays the distribution of the data for each "purpose" category and the relationship between the two variables, "revol.util" and "int.rate" within those categories. It gives an idea of the median, quartiles, minimum, and maximum values of the data for each category.

```
ggplot(data=loanGraphPlotData, aes(x=revol.util, y=int.rate, fill=purpose)) +
  geom_boxplot() +
  ggtitle("Loan Purpose and Revolving Line Utilization Rate vs Interest Rate")
```

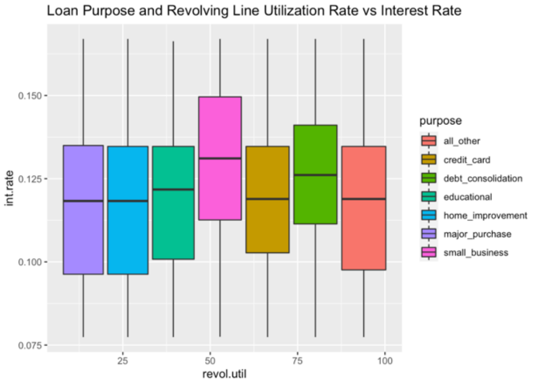


These data points are known as outliers because they are much higher or lower than the average. They offer significant insight into the data's distribution and may flag any anomalies or outliers in the dataset.

Removing outliers

```
lower_limit <- quantile(loanGraphPlotData$int.rate, 0.05)
upper_limit <- quantile(loanGraphPlotData$int.rate, 0.95)

loanGraphPlotData_no_outliers <- subset(loanGraphPlotData, int.rate >= lower_limit & int.rate <= upper_limit)
```

```
ggplot(data=loanGraphPlotData_no_outliers, aes(x=revol.util, y=int.rate, fill=purpose)) +
  geom_boxplot() +
  ggtitle("Loan Purpose and Revolving Line Utilization Rate vs Interest Rate")
```



According to the results, all purposes in each category have median values that are generally similar, with the exception of small business, which has a somewhat higher median value than the rest.

Summary of dataset

```
summary(loanGraphPlotData_no_outliers)
```

```
##    revol.util       int.rate                       purpose    
##  Min.   :  0.0   Min.   :0.0774   all_other         :2096  
##  1st Qu.: 23.8   1st Qu.:0.1062   credit_card       :1184  
##  Median : 46.7   Median :0.1221   debt_consolidation:3704  
##  Mean   : 47.0   Mean   :0.1215   educational       : 318  
##  3rd Qu.: 70.5   3rd Qu.:0.1380   home_improvement  : 577  
##  Max.   :108.8   Max.   :0.1670   major_purchase    : 388  
##                                   small_business    : 505
```

```
dim(loanGraphPlotData_no_outliers) # display number of observations and variables
```

```
## [1] 8772    3
```

The minimum revolving line utilisation rate is 0, the first quartile (25th percentile) is 23.8, the median (50th percentile) is 46.7, the mean is 47, the third quartile (75th percentile) is 70.5, and the maximum value is 108.8. The summary also highlights the distribution of loans' intended uses, with debt consolidation ranking first among them, followed by credit cards and all other. Furthermore, the minimum interest rate is 0.0774, the first quartile is 0.1062, the median is 0.1221, the mean is 0.1215, the third quartile is 0.138, and the highest is 0.167.

Split dataset into the Training set and Test set

Divide a dataset into train and test sets to see how effectively our machine learning model works.

```
split <- sample.split(loanGraphPlotData_no_outliers, SplitRatio = 0.8)  # split data into ratio of 8:2
training_set <- subset(loanGraphPlotData_no_outliers, split == "TRUE")
test_set <- subset(loanGraphPlotData_no_outliers, split == "FALSE")
```

Scale data

Scaling the data is one of the pre-processing steps used in machine learning algorithms on the data set, which makes it easier for the model to understand and learn about the problem.

```
training_set.scale <- scale(select(training_set, -purpose))
test_set.scale <- scale(select(test_set, -purpose))
```

Descriptive Analysis

Using historical data, analytical modeling, data mining techniques, and machine learning, predictive analytics is a subset of new insights that forecasts probable outcomes.

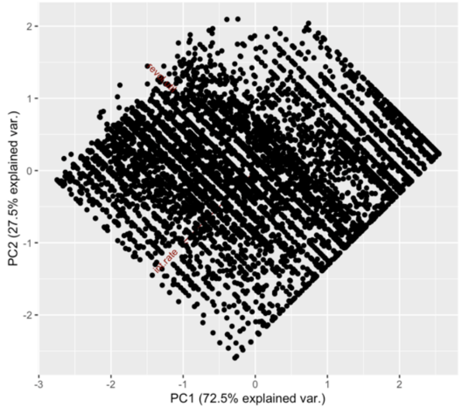
1. Principal component analyse (PCA)

A type of unsupervised statistical learning is this algorithm. It offers data visualization, dimension reduction techniques, and, most importantly, it offers data pre-processing techniques before applying another methodology.

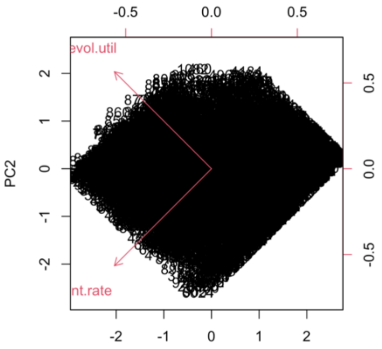
```
data.table <- round(cor(training_set.scale), 3)
head(data.table)
```

```
##            revol.util int.rate
## revol.util       1.00     0.45
## int.rate         0.45     1.00
```

```
data.pca <- prcomp(training_set.scale, center = TRUE,scale. = TRUE)

ggbiplot(data.pca, scale=0)
```



```
biplot(data.pca, scale = 0)
```



2. Pearson correlation coefficient

The Pearson correlation coefficient ranges from -1 to 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation.

```
training_set.scale <- as.data.frame(training_set.scale)
cor.test(training_set.scale$revol.util, training_set.scale$int.rate, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  training_set.scale$revol.util and training_set.scale$int.rate
## t = 38.477, df = 5846, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4288353 0.4697437
## sample estimates:
##       cor
## 0.4495252
```

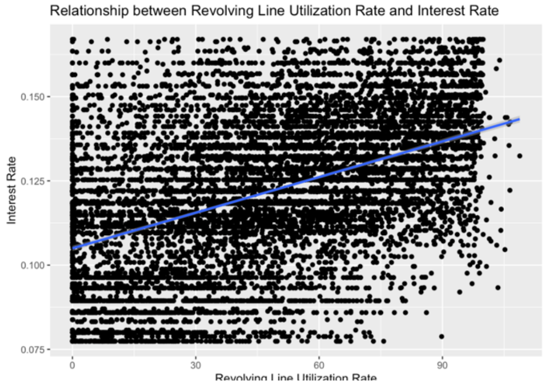
The two variables have a moderately positive correlation, as indicated by the correlation coefficient (cor), which is at 0.4458614. This association is statistically significant at a very high level, according to the t-value of 38.085 and the p-value of 0.00000000000000022. This is also supported by the 95% confidence interval, which has a range of 0.4269395 to 0.4679330.

3. Linear Regression

This method is a statistical approach that forecasts the result of a dependent variable using one independent variable. Using this method, we may calculate the model's variance as well as the proportional contributions of independent variable to the overall variance.

```
ggplot(data=loanGraphPlotData_no_outliers, aes(x=revol.util, y=int.rate)) + geom_point() + geom_smooth(method =
"lm") + ggtitle("Relationship between Revolving Line Utilization Rate and Interest Rate") + xlab("Revolving Line
 Utilization Rate") + ylab("Interest Rate")
```

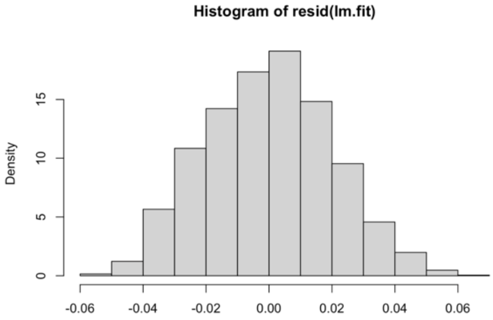
```
## `geom_smooth()` using formula 'y ~ x'
```



```
lm.fit =lm(int.rate~.,data= loanGraphPlotData_no_outliers[1:2])
summary(lm.fit)
```

```
##
## Call:
## lm(formula = int.rate ~ ., data = loanGraphPlotData_no_outliers[1:2])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057834 -0.014488  0.000349  0.013812  0.062030
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.050e-01  4.135e-04  253.84   <2e-16 ***
## revol.util  3.526e-04  7.520e-06   46.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0201 on 8770 degrees of freedom
## Multiple R-squared:  0.2005, Adjusted R-squared:  0.2004
## F-statistic:  2199 on 1 and 8770 DF,  p-value: < 2.2e-16
```

```
hist(resid(lm.fit), prob = TRUE)
```



A measurement of how well the model fits the data is the multiple R-squared value. With an R-squared value of 0.2005, the model can account for 20% of the variability in interest rates. The overall significance of the model is evaluated using the F-statistic and its corresponding p-value. The model fits the data well, as shown by the p-value of less than 0.05. However, the residuals appear to be fairly regularly distributed based on the statistics for the residuals that are presented in the output. The residuals have a median value of 0.000349, a range of -0.014488 to 0.062030, and a minimum value of -0.057834. Furthermore, the residual standard error is 0.0201 on 8770 degrees of freedom, which is negligibly low and suggests that the residuals are reasonably near to the fitted values. Since the residuals are roughly symmetrical, the residuals' histogram appears to be somewhat bell-shaped.

Conclusion

The residual standard error being small and the R-squared value being high are both signs of a good model fit. The low p-value which is almost 0 for the F-statistic also suggests that the model is a good fit for the data. However, the relatively low adjusted R-squared value suggests that the model does not explain a large proportion of the variance in the response variable.

Ho: There is no significant correlation between revol.util and int.rate. (reject)

Ha: There is a significant correlation between revol.util and int.rate.