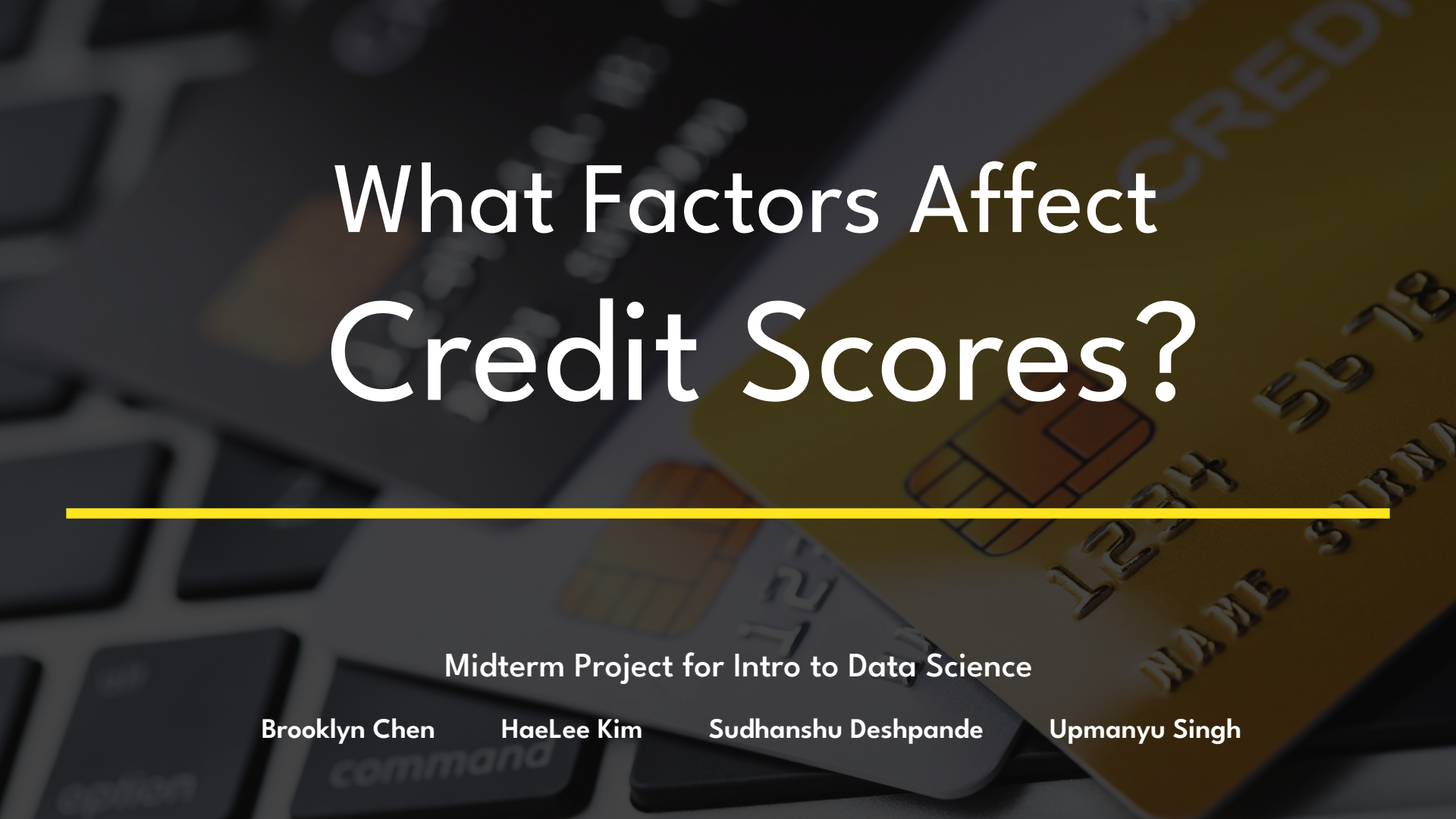


What Factors Affect Credit Scores?

Midterm Project for Intro to Data Science

Brooklyn Chen HaeLee Kim Sudhanshu Deshpande Upmanyu Singh

SMART Question:

What Factors Affect the Credit Score?

Age?

Credit-Mix?

Number of Delayed Payment?

Total EMI per Month?

About the Dataset

a. Source : Credit Classification in Kaggle
https://www.kaggle.com/datasets/parisrohan/credit-score-classification

a. 12500 Observations of 28 Variables
   - Credit Score
   - Credit Mix
   - Number of Delayed Payment
   - Total EMI
   - Age
   - ID, Name, Number of Loan, Annual Income, Number of Bank Account, Number of Credit Card, Interest Rate, Monthly Balance, etc.

Exploratory Data Analysis

To Prepare EDA

- Dropped nulls and outliers for all variables (used ezids outlierKD2)
- Converted the variable of credit-score into factor
- Dropped observations with age below 18 and over 100
- Tests used:
  - Chi-square Test ○ ANOVA ○ Post-hoc Tukey
- Graphs used:
  - Histogram ○ QQ-plot ○ Boxplot

Ready to EDA

Summary of the Refined Dataset

	Age	Num_of_Delayed_Payment	Credit_Mix	Total_EMI_per_month	Credit_Score
X	Min. :18.0	Min. : 0.00	Min. :1.000	Min. : 0.00	Poor :2338
X.1	1st Qu.:26.0	1st Qu.: 9.00	1st Qu.:2.000	1st Qu.: 26.73	Standard:4689
X.2	Median :34.0	Median :13.00	Median :2.000	Median : 60.33	Good :1414
X.3	Mean :34.3	Mean :13.13	Mean :2.099	Mean : 84.89	NA
X.4	3rd Qu.:42.0	3rd Qu.:18.00	3rd Qu.:3.000	3rd Qu.:125.01	NA
X.5	Max. :55.0	Max. :28.00	Max. :3.000	Max. :328.33	NA
X.6	NA's :1	NA's :72	NA	NA's :529	NA

Credit Mix

H0: Credit Mix and Credit Score are independent

H1: Credit Mix and Credit Score are not independent

What’s Credit Mix ?

A credit mix refers to the multiple types of loan accounts you hold, such as credit cards, student loans, mortgages, and car loans.

Credit Mix - Data Type

int

Credit_Mix
2
3
2
2
2
2
1
3
3
2

1 = Bad Credit Mix
2 = Standard Credit Mix
3 = Good Credit Mix

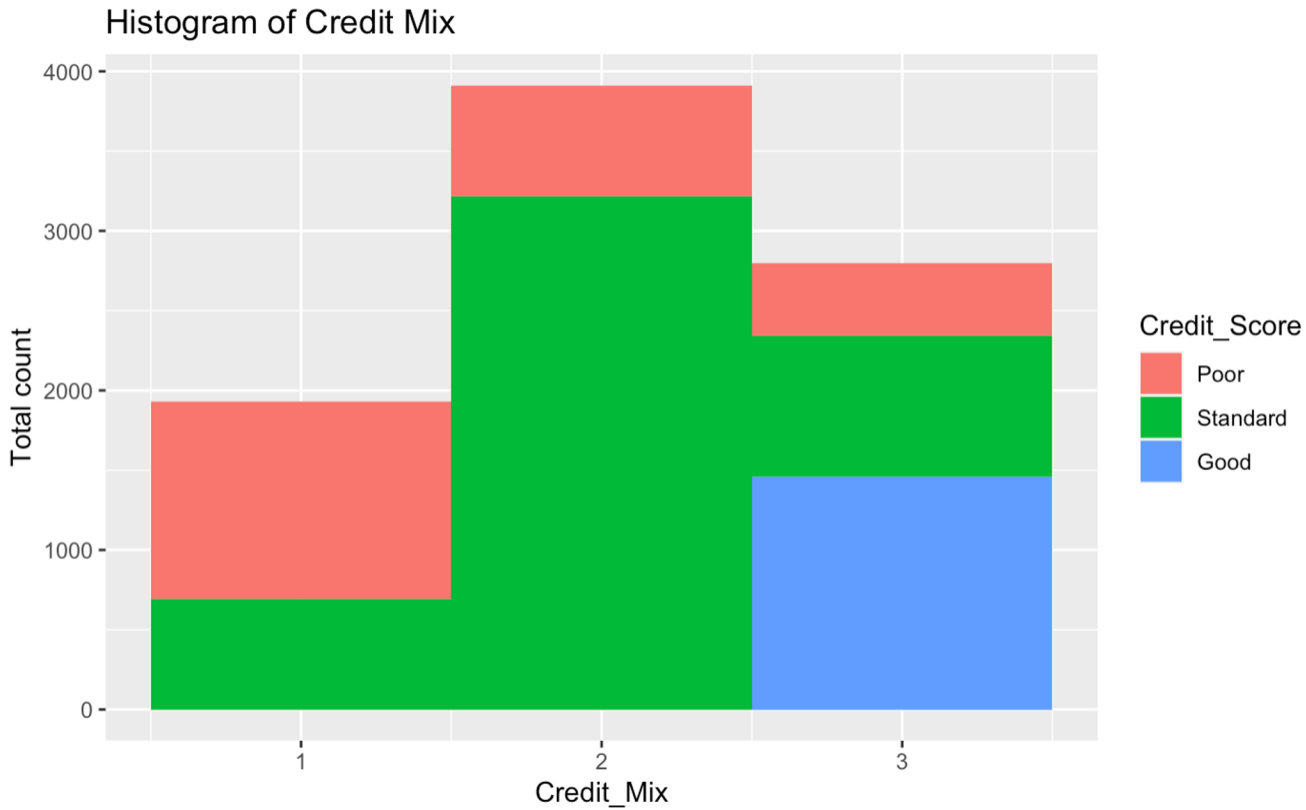
chr

Credit_Score
Standard
Standard
Standard
Poor
Standard
Standard
Standard
Good
Good
Standard

Factor

Credit_Score
Standard
Standard
Standard
Poor
Standard
Standard
Standard
Good
Good
Standard

Credit Mix - Histogram



Credit Mix - Chi-square Test

Whether Credit_Mix and Credit_Score are likely to be related or not ?

Contingency table for Credit_Mix vs Credit_Score

	Poor	Standard	Good
Bad	1213	678	0
Standard	683	3138	0
Good	442	873	1414

Cross table for the expected frequencies between Credit_Mix and Credit_Score

	Poor	Standard	Good
Bad	524	1050	317
Standard	1058	2123	640
Good	756	1516	457

```
        Pearson's Chi-squared test

data:  mix
X-squared = 5021, df = 4, p-value <2e-16
```

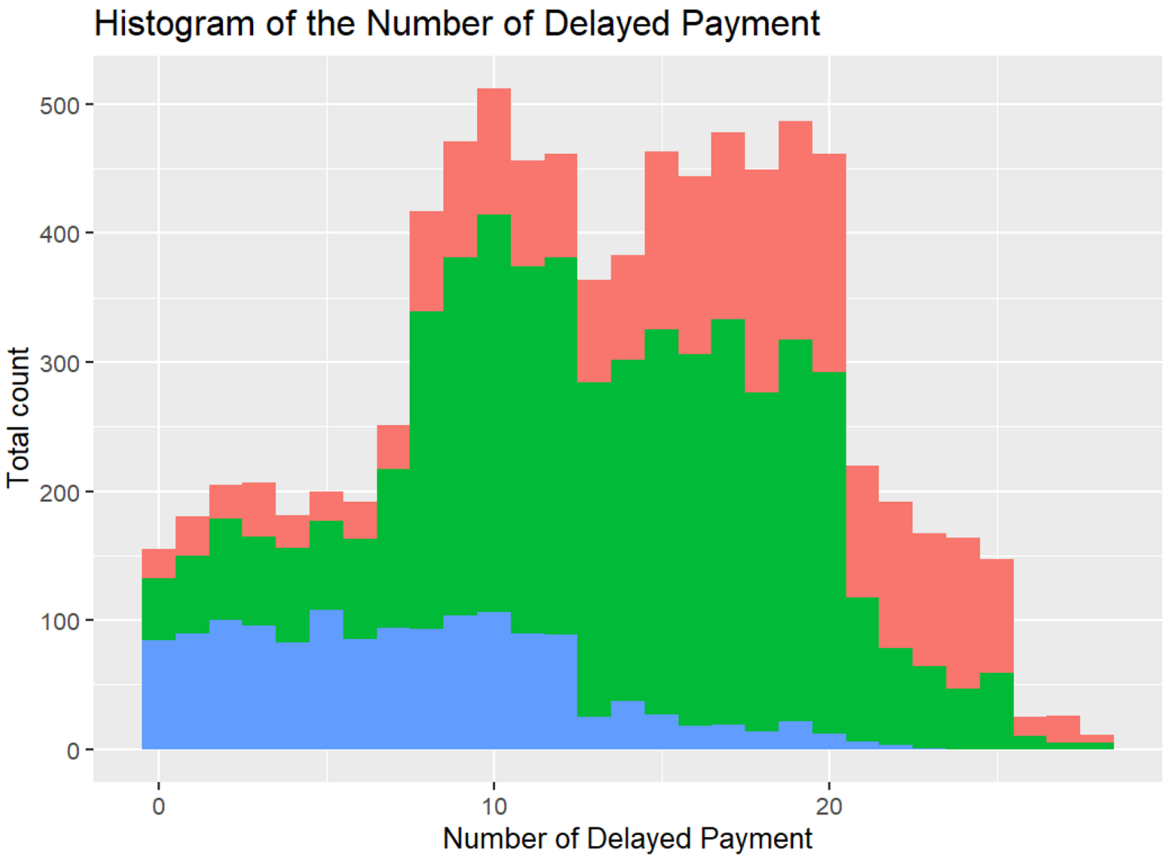
Reject H0

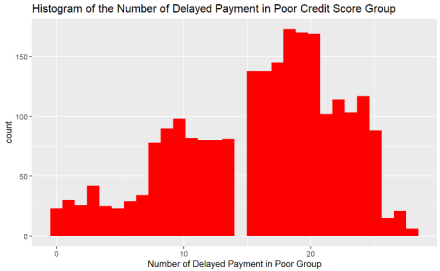
Number of Delayed Payment

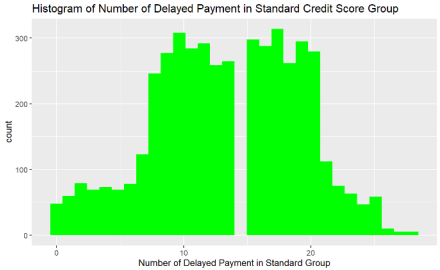
H0: The average of delayed payment in poor, standard, good groups are same.

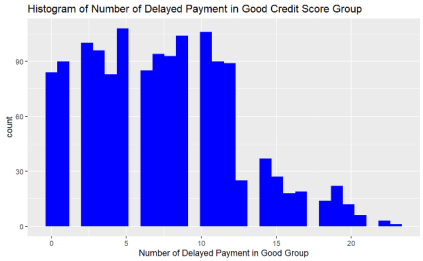
H1: The average of delayed payment in poor, standard, good groups are not same.

Number of Delayed Payment Continued.

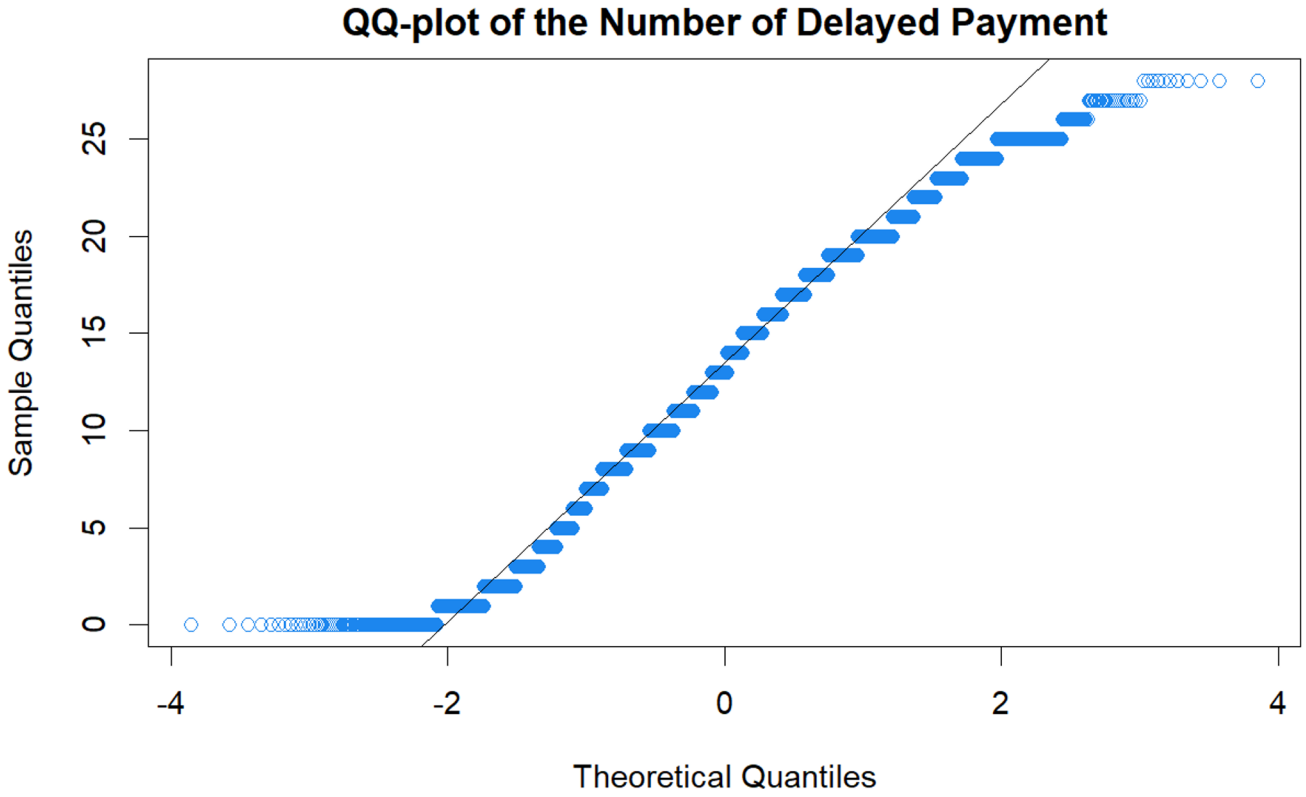


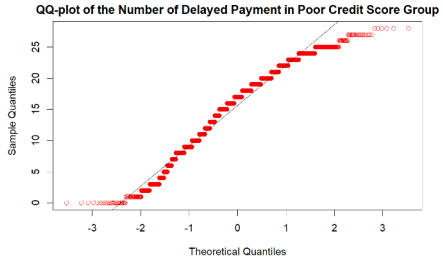


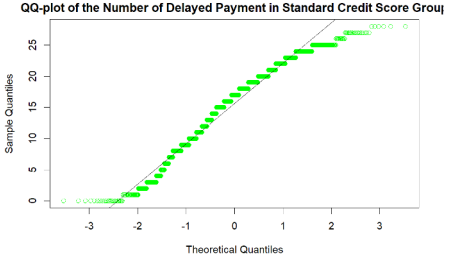


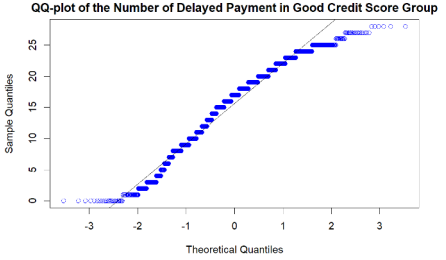


Number of Delayed Payment Continued.

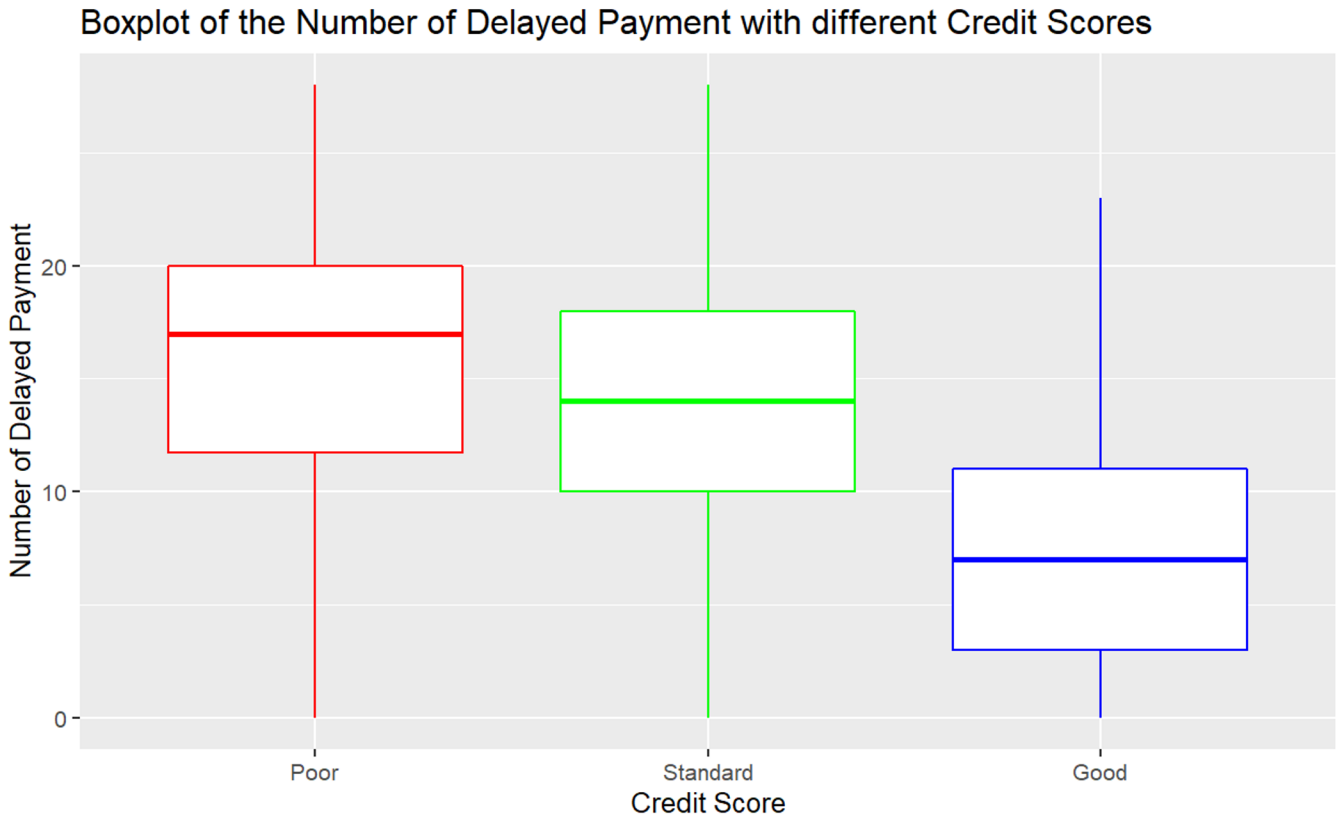








Number of Delayed Payment Continued.




```
Num_of_Delayed_Payment
Min.    : 0.00
1st Qu.:11.75
Median :17.00
Mean   :15.91
3rd Qu.:20.00
Max.   :28.00
```

```
Num_of_Delayed_Payment
Min.    : 0.00
1st Qu.:10.00
Median :14.00
Mean   :13.47
3rd Qu.:18.00
Max.   :28.00
```

```
Num_of_Delayed_Payment
Min.    : 0.000
1st Qu.: 3.000
Median : 7.000
Mean   : 7.397
3rd Qu.:11.000
Max.   :23.000
```

Number of Delayed Payment Continued.

- **ANOVA result summary the Number of Delayed Payment between Credit Score**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Credit_Score	2	64654.99	32327.4926	1009.328	0
Residuals	8366	267952.21	32.0287	NA	NA

- **Post-hoc Tukey result summary**

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Num_of_Delayed_Payment ~ Credit_Score, data = df)

$Credit_Score
                    diff       lwr       upr p adj
Standard-Poor -2.432203 -2.769494 -2.094913     0
Good-Poor     -8.510026 -8.958393 -8.061659     0
Good-Standard -6.077823 -6.481653 -5.673992     0
```

Total EMI per Month

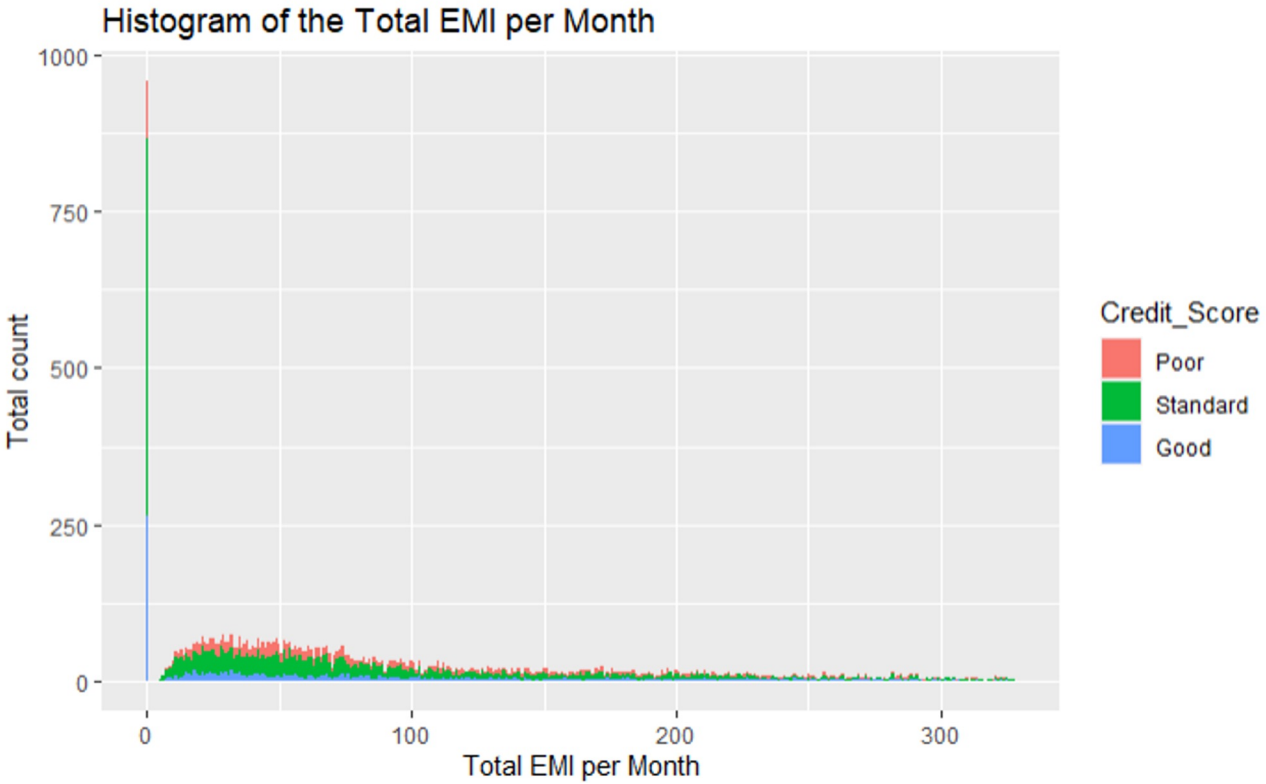
H0: The average of monthly EMI in poor, standard, good groups are same.

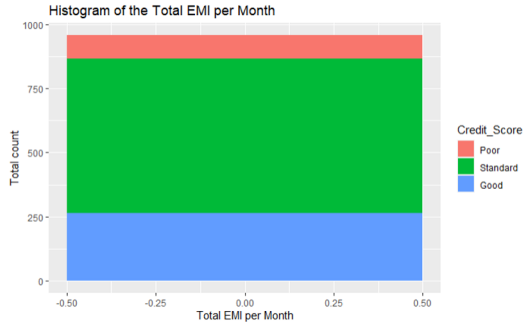
H1: The average of monthly EMI in poor, standard, good groups are not same.

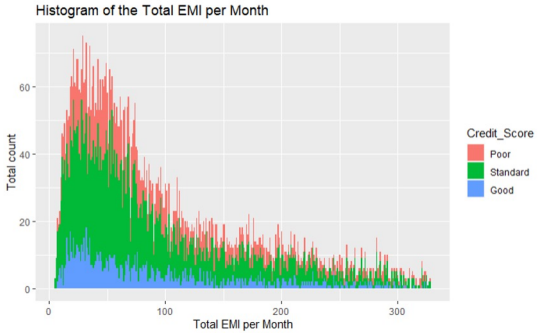
What's total EMI ?

Total EMI refers to an **E**quated **M**onthly **I**nstalment (EMI) is a set monthly payment provided by a borrower to a creditor on a set day, each month. EMIs apply to both interest and principal each month, and the loan is paid off in full over some years.

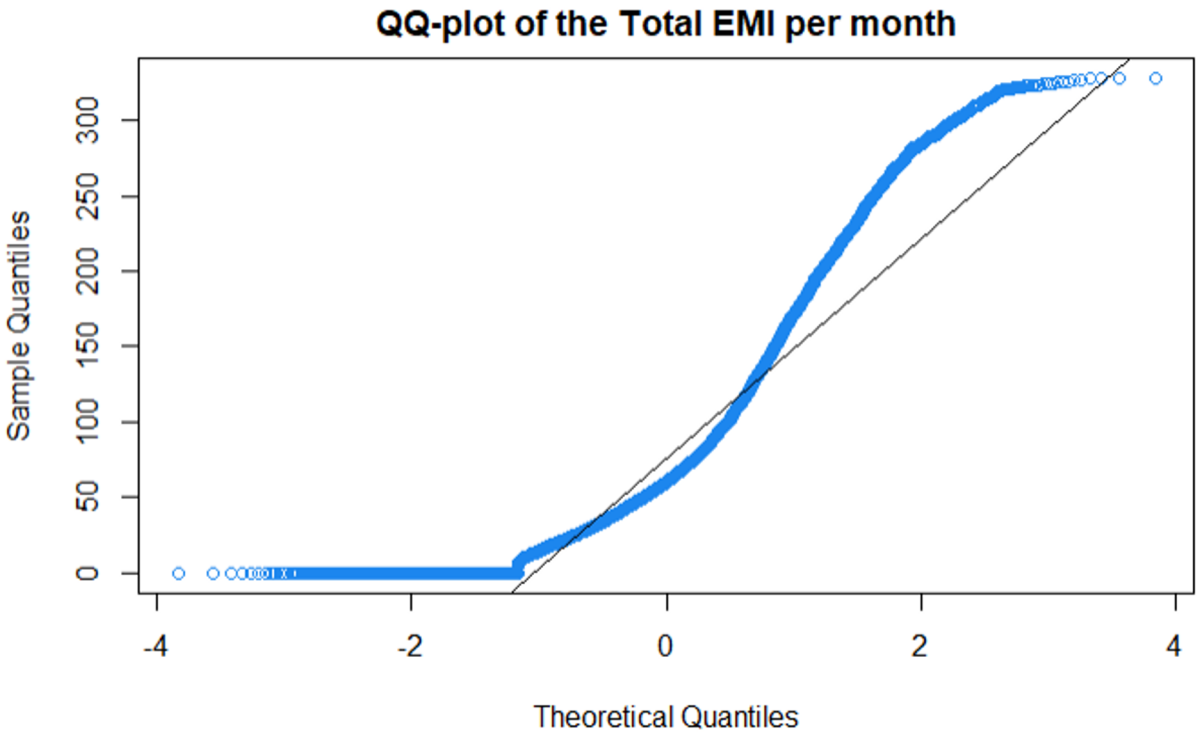
Total EMI per Month Continued.

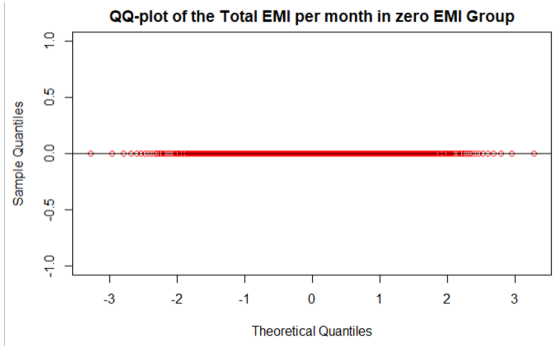


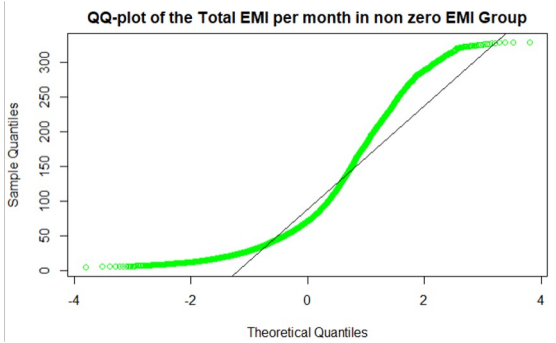




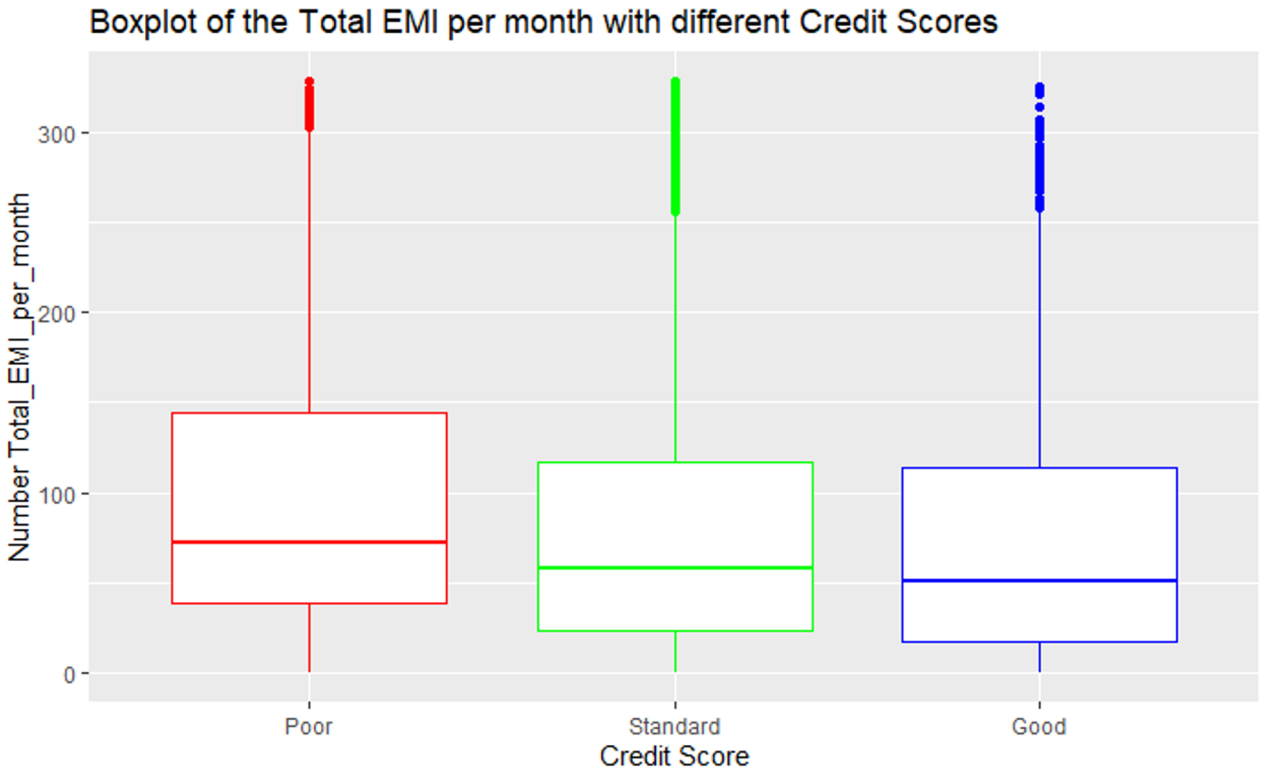
Total EMI per Month Continued.







Total EMI per Month Continued.




```
Total_EMI_per_month
Min.    :  0.00
1st Qu.: 38.10
Median : 71.84
Mean   : 98.27
3rd Qu.:144.00
Max.   :328.33
```

```
Total_EMI_per_month
Min.    :  0.00
1st Qu.: 23.75
Median : 57.51
Mean   : 81.00
3rd Qu.:116.86
Max.   :328.24
NA's   :285
```

```
Total_EMI_per_month
Min.    :  0.00
1st Qu.: 17.57
Median : 50.78
Mean   : 75.97
3rd Qu.:113.98
Max.   :325.28
NA's   :76
```

Total EMI per Month Continued.

- **ANOVA result summary the Total EMI per Month between Credit Score**

ANOVA result summary total EMI per month between Credit Score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Credit_Score	2	561219.7	280609.855	46.8888	0
Residuals	7909	47332086.9	5984.586	NA	NA

- **Post-hoc Tukey result summary**

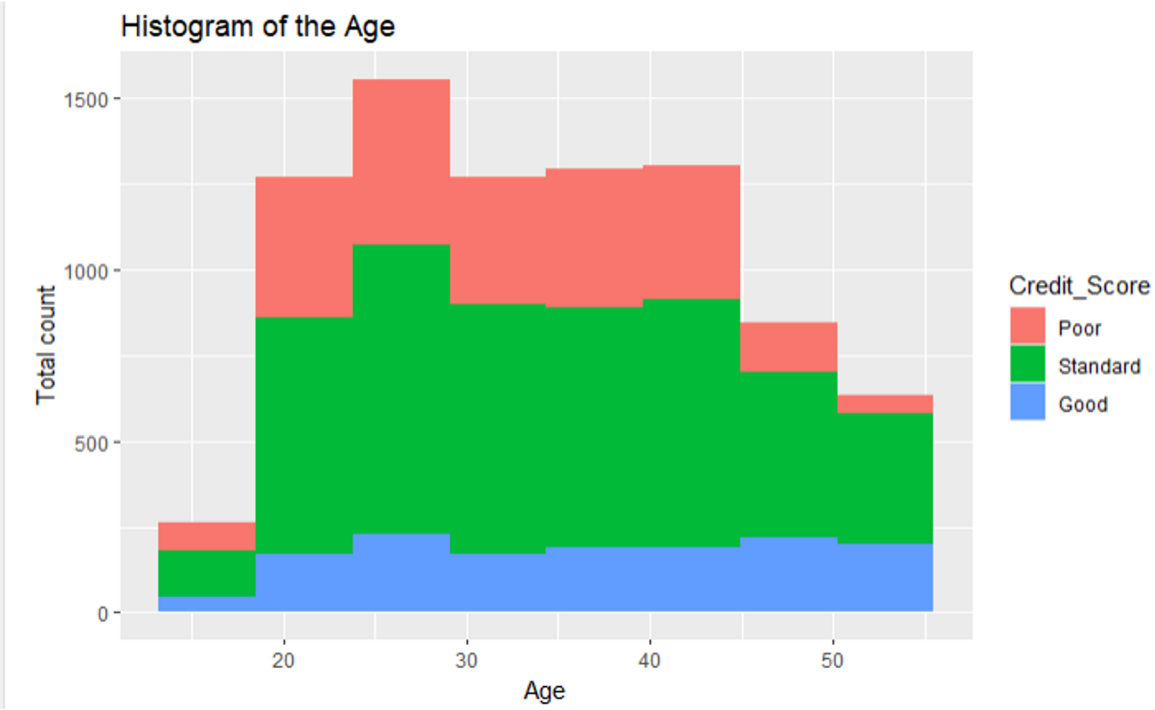
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Total_EMI_per_month ~ Credit_Score, data = df)

$Credit_Score
                    diff       lwr          upr     p adj
Standard-Poor -17.263376 -22.01960 -12.5071503 0.0000000
Good-Poor     -22.294214 -28.59759 -15.9908411 0.0000000
Good-Standard  -5.030838 -10.69168   0.6300036 0.0933578
```

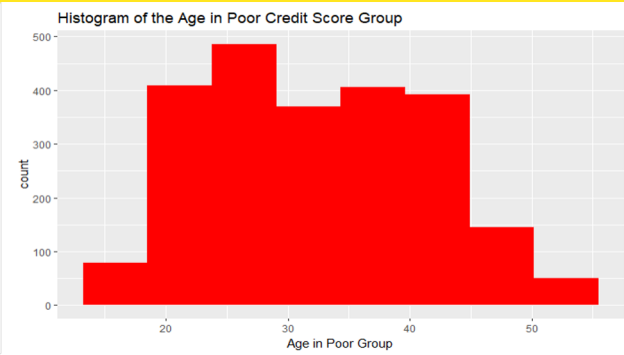
Age

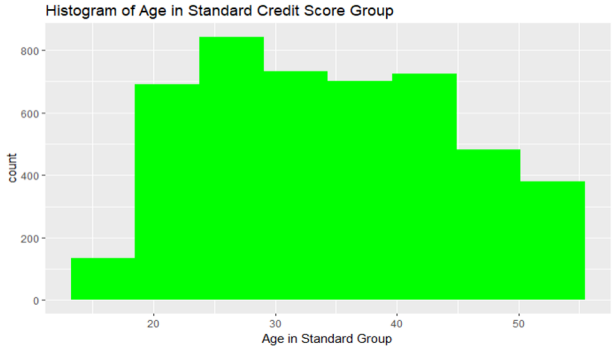
H0: The average of age in poor, standard, good groups are same.

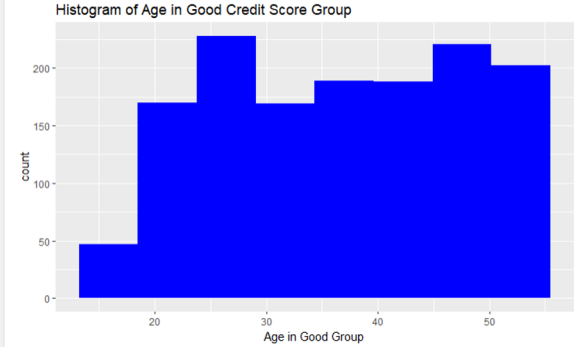
H1: The average of age in poor, standard, good groups are not same.

Age Continued.

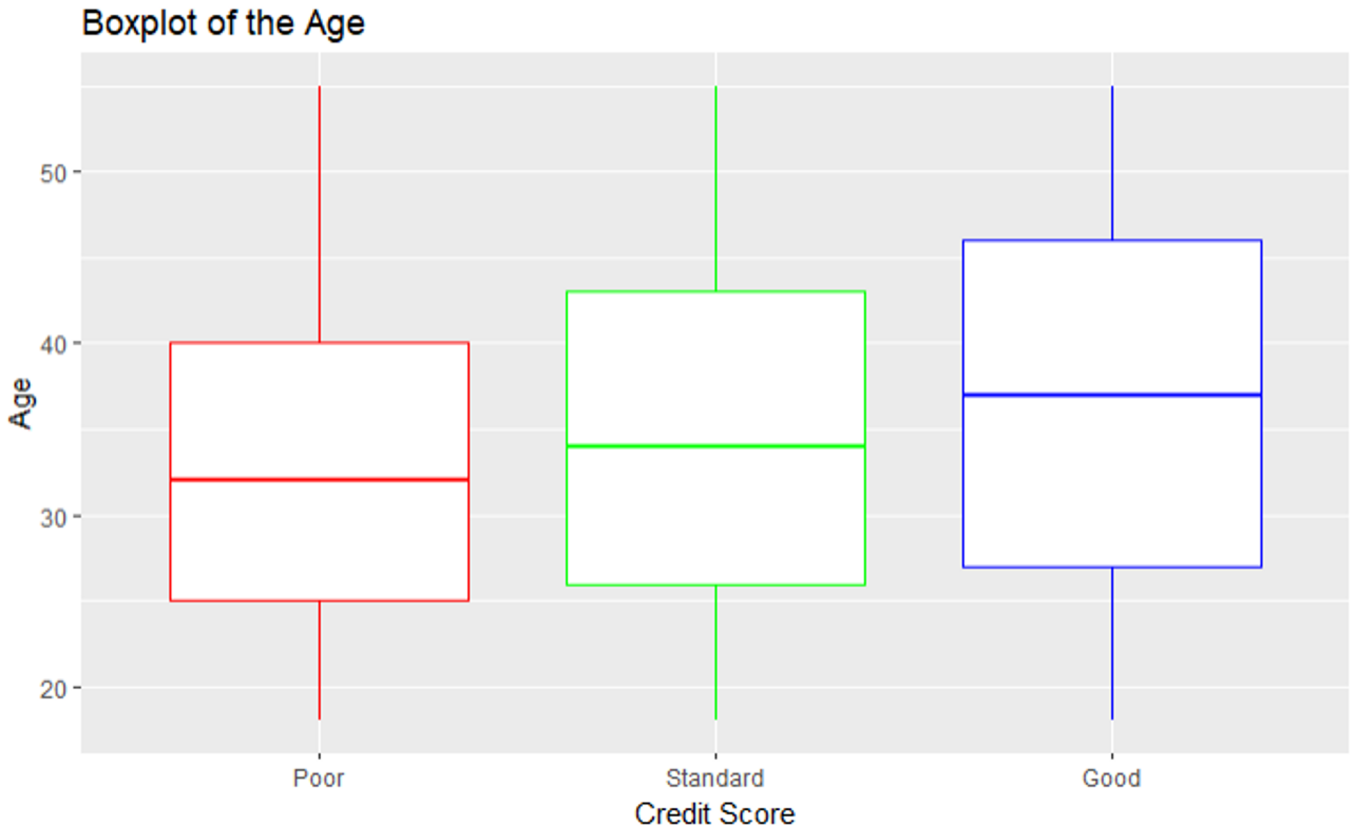








Age Continued.




```
      Age
 Min.   :18.00
 1st Qu.:25.00
 Median :32.00
 Mean   :32.32
 3rd Qu.:40.00
 Max.   :55.00
```

```
      Age
 Min.   :18.00
 1st Qu.:26.00
 Median :34.00
 Mean   :34.55
 3rd Qu.:43.00
 Max.   :55.00
```

```
      Age
 Min.   :18.00
 1st Qu.:27.00
 Median :37.00
 Mean   :36.73
 3rd Qu.:46.00
 Max.   :55.00
```

Age Continued.

- **ANOVA result summary the Age between Credit Score**

ANOVA result summary the Age between Credit Score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Credit_Score	2	17817.37	8908.6840	89.0123	0
Residuals	8437	844406.21	100.0837	NA	NA

- **Post-hoc Tukey result summary**

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Num_of_Delayed_Payment ~ Credit_Score, data = df)

$Credit_Score
                    diff       lwr       upr p adj
Standard-Poor -2.432203 -2.769494 -2.094913     0
Good-Poor     -8.510026 -8.958393 -8.061659     0
Good-Standard -6.077823 -6.481653 -5.673992     0
```

We could developed a sketch of the answer to our question,

What Factors Affect the Credit Score?

	Test	Result
Credit Mix	Chi-square	Rejected H0
Number of Delayed Payment	ANOVA & Post-hoc Tukey	Rejected H0
Total TMI per month	ANOVA & Post-hoc Tukey	Rejected H0
Age	ANOVA & Post-hoc Tukey	Rejected H0

→ We can have a sketch that these factors affect the credit score.

Questions?

Thank You!