# Loan Status Analysis part 2

by

Tae Hyon Lee

For

Feng Tang
Monjaco

5/31/2017

# Content

# Introduction

**Background**

SourceData.csv and DataFix.csv files are given.
SourceData.csv file contains the loan status and loan pplication data of the borrower.
There are total of 12,068 loans in SourceData.csv

SourceData.csv has the following variables:

1. **ID**: unique loan identifier
2. **loan_status**: either current (the loan is paying on time) or late (the borrower is late on the most recent payment)
3. **loan_amnt**: total amount of the loan borrowed
4. **term**: number of months the loan is scheduled to be paid back
5. **installment**: amount in dollars of the monthly payment of the loan including principal and interest
6. **emp_length**: length of emploment history as reported by the borrower
7. **annual_inc**: annual income as reported by the borrower
8. **is_inc_v**: whether or not the income of the borrower has been verified to be correct or has not been verified.
9. **fico**: FICO Score
10. **dti**: debt to income ratio, ratio of monthly debt payments to monthly income
11. **revol_util**: revolving utilization in percentage
12. **mort_acc:** number of mortgage accounts
13. **open_acc_6m:** number of account opened in the last 6 months
14. **inq_last_6mths:** number of credit inquiries in the last 6 months
15. **inq_last_12m:** number of credit inquiries in the last 12 months
16. **mths_sin_rcnt_bc:** number of months since the most recennt bank card opening
17. **mths_sin_rcnt_inq**: number of months since the most recent credit inquiry

# Loan Status Analysis

**Background**

There are 2370 data in DataFix.csv file.
DataFix.csv has the following variables:
1. ID: loan identifier
2. column_name: name of variable with missing or corrupted value in the DataSource.csv file
3. column_vale: original value of the variable

**Purposes**

The goal of this analysis is to predict loan_status based on other variables in the data set.

**Assumptions**

- The loan status analysis is based on 12, 068 loans.
- The analysis assumes that the data was collected at the time loan was lent out because that is when lenders check borrower's credit report. So DTI is not influenced by the current installment.
- This analysis assumes that the data provided by the borrowers and the credit bureau are correct and accurate unless the data is noticeably out of place (ex. 999,999 for income).
- This analysis also assumes that the cost of living is the same and the value of annual income is same for all the loans (ex. a borrower who lives in a rural area with an annual income of 50K might have a lot of money to spare but a borrower who lives in San Francisco with an annual income of 50K might be struggling to pay off other expenses because the cost of living in San Francisco is high).
- Lastly, this analysis assumes that the provided variables is enough to predict the loan status. The amount of money available for borrower, down payment, and whether or not the borrower is paying off other bills on time can all be a big factor in determining whether or not the loan payment is going to be late.

# Loan Status Analysis

## Analysis

### Sample distribution

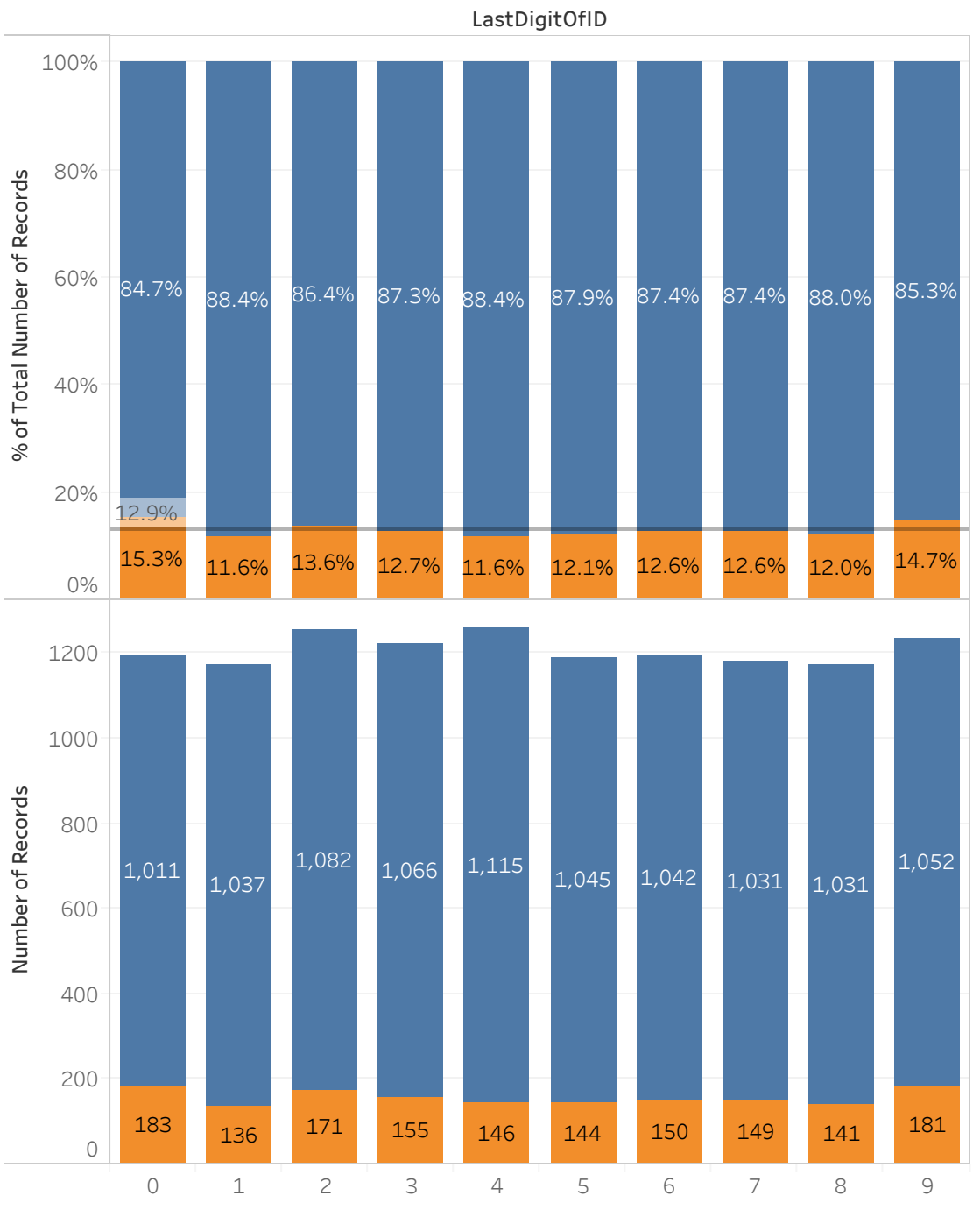
### Loan Status

| Loan Status | % of Total Number of Recor.. | Number of Records |
|---|---|---|
| Current | 87.11% | 10,512 |
| Late | 12.89% | 1,556 |
| Grand Total | 100.00% | 12,068 |

Loan Status Current Late

- 12.89% of loans in the given data have a ”Late” loan status.

### Sample Distribution by Last Digit of ID

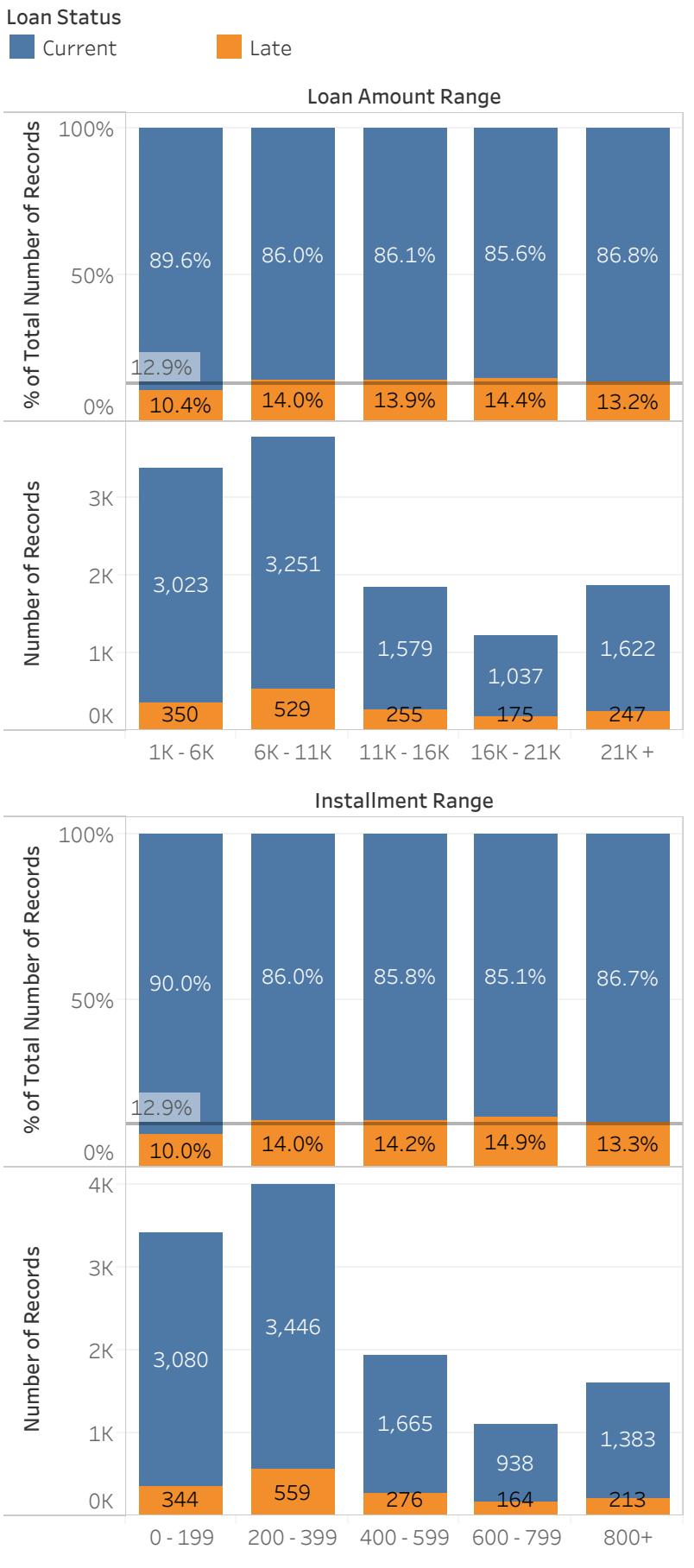



**Validating my approach/data**

- The sample distribution is very uniform across all groups.
- Although there are some fluctuations, every loan group seperated by the last digit of the loan ID has around 13% late loan status, which is close to the percentage of late loan status in the whole sample.
- Plus or minus fluctuations from 12.89% is very little and therefore insignificant.

# Loan Status Analysis

## Loan Amount and Installment




### Loan Amount

### Anomalies

- There is clearly a percentage difference between loan amounts of 1K to 6K and other loan amounts.
- When loan amount is between 1,000 to 5,999, only 10.4% of the loans has a late status while for loan amount of 6,000 to 10,999, 14% of the loans has a late loan status.
- The percentage of late loan status stays high for loan amounts that are larger than or equal to 11K.
- After running Chi-square test, we get a p value of <.0001, which means that this anomaly is statistically significant.
- We can confidently say loan amounts that are equal to or more than 6,000 dollars are more likely to have a late loan status than the ones that are lower than 6,000 dollars.

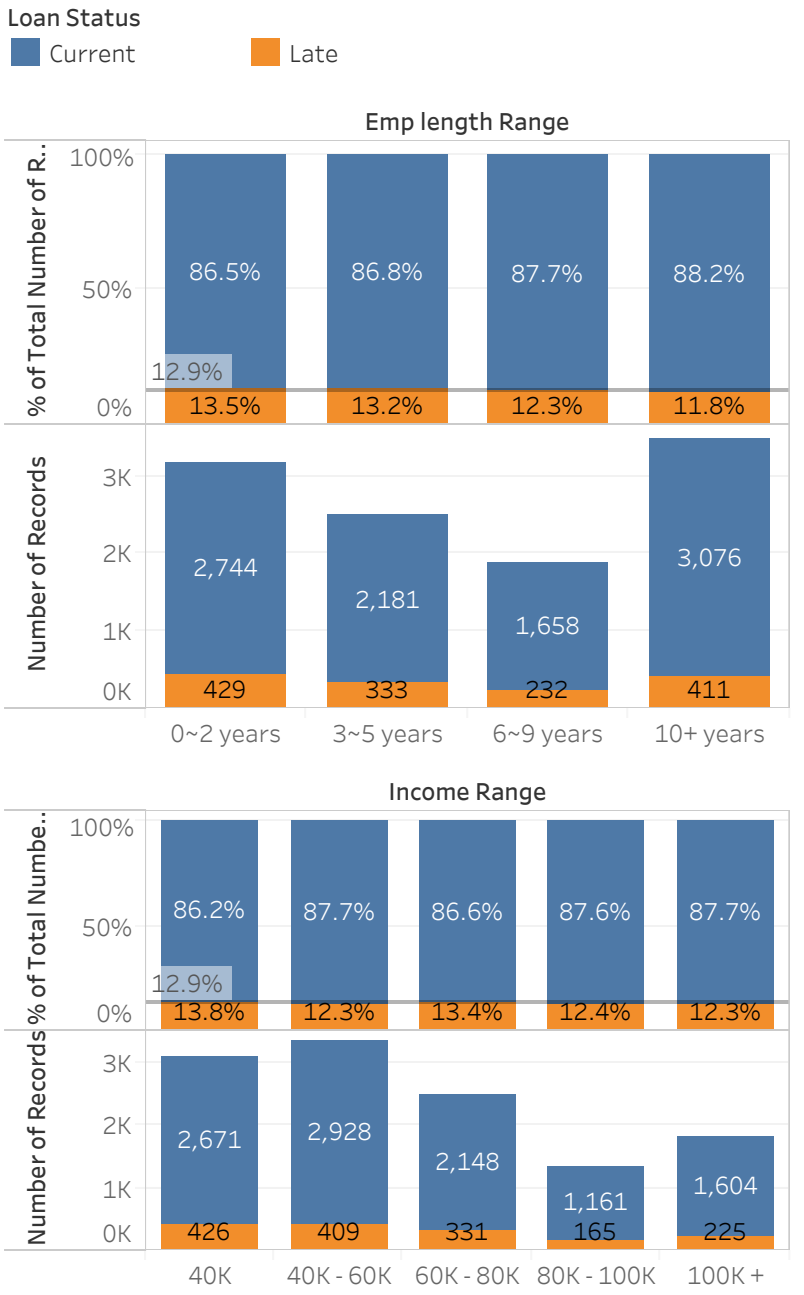



### Anomalies

- We get a very similar result to that of above, probably because monthly payment is correlated to loan amount.
- Loans with monthly payment of 0 to 199 dollars have only 10% chance that the loan will have a late status.
- But if the installment goes up to 200 or above, it stays very close to 14% which is above 12.9%, the percantage of late loan status of overall sample.
- And again, we get p value of <.0001 from the Chi-square test, which suggests that this anomaly is definitely statistically significant and something to pay attention to.
- We can confidently say installment of more than or equal to 200 dollars are more likely to have a late status than the installments of less than 200 dollars.

# Loan Status Analysis

## Employment and Income




### Employment Length

Anomalies

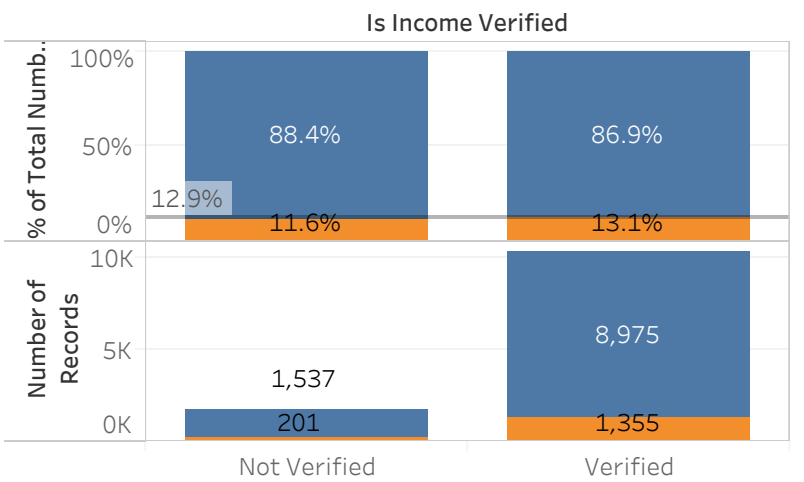
- There is a fairly small downward trend for the late loan status by employment length.
- However, after running Chi-square test, we get a p value of .1369, which means that this finding is statistically insignificant
- Although the categorical test shows us smaller percentage of late status loan for 10 + years employment length than that of less employment length, this finding is not statistically significant enough to be considered an anomaly




### Income

Anomalies

- There is also a somewhat downward trend for the late loan status by income but it does not look very significant
- After running Chi-square test, we get a p value of .339, which means that the percentage of late loan status is evenly distributed among all income ranges and that there is no anomaly



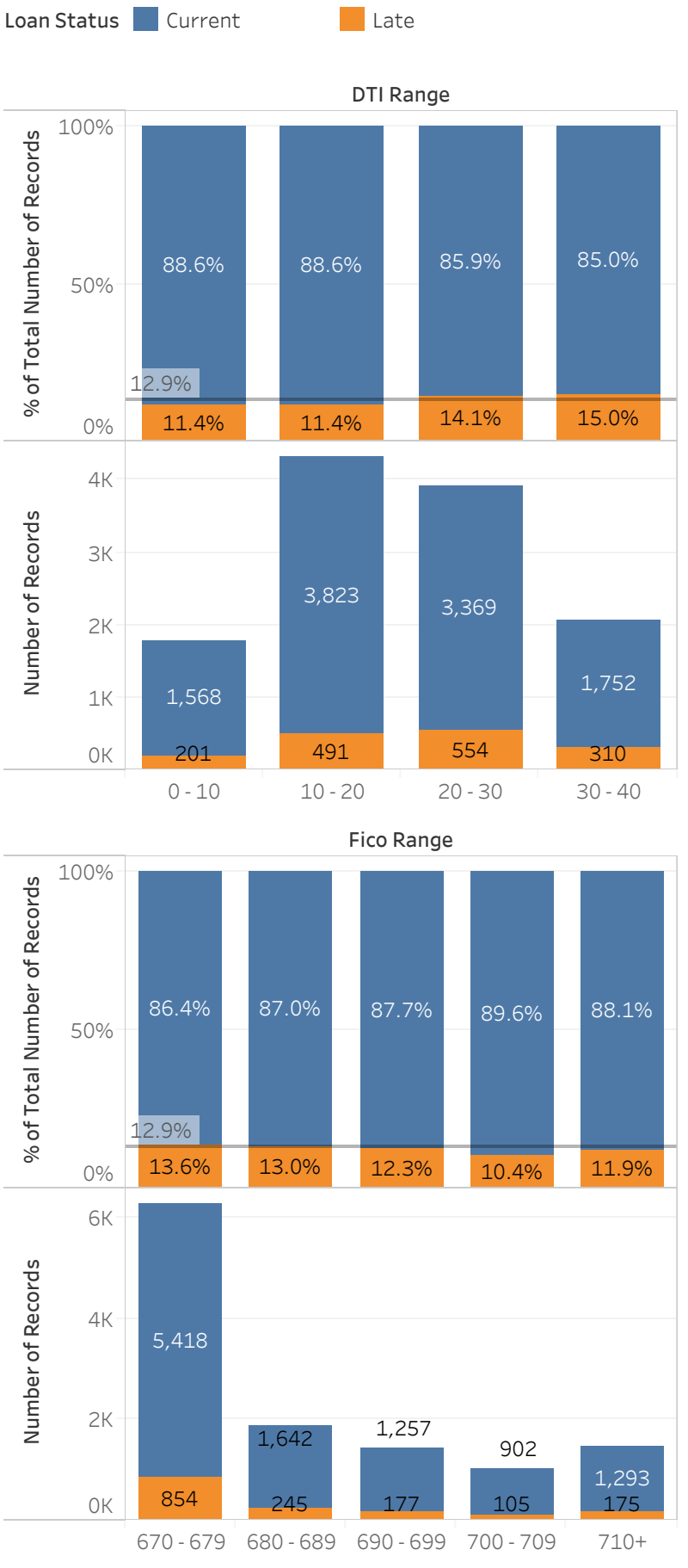

### Income Source Verified

Anomalies

- Although there is some difference between the percentages of late status loans for not verified and verified, it does not look very significant as the sample size of Not Verified is very small compared to that of Verified.
- After running Chi-square test, we get a p value of .074 and come up a conclusion that there is no significant difference between the two groups.

# Loan Status Analysis

## DTI and FICO Score




### Debt to Income (DTI)

Anomalies

- For DTI Range between 0 to 20, only 11.4% of the loans have a late status but for DTI Range between 20 and 30 and DTI Range 30 to 40, 14.1% and 15.0%, respectively, of the loans have a late status.
- There seems to be a fairly significant discrepancy between the first two groups and the last two groups.
- After running Chi-square test, we get a p value of <.0001, which means that this anomaly is statistically significant.
- We can confidently say loans with DTI that are equal to or more than 20% are more likely to have a late loan status than the ones that are lower than 20%.

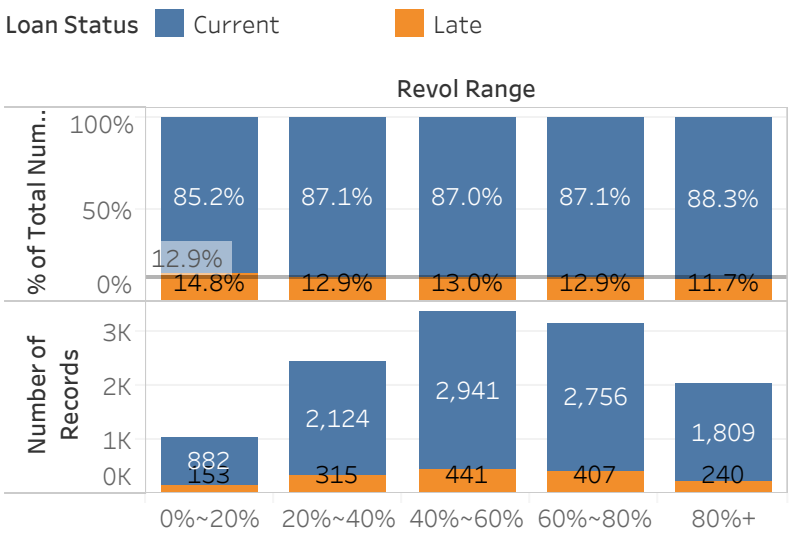
### FICO Score

Anomalies

- The test shows that the loans with high FICO score is less likely to have a late loan status.
- There seems to be a discrepancy between the late loan status percentages for 670-679 FICO score and FICO score 700-709.
- After running Chi-square test, we get a p value of .0403, which means that this anomaly is statistically significant.
- Although not as significant as the ones with p value with < 0.001, We can say loans with high FICO score are less likely to have a late loan status.

# Loan Status Analysis

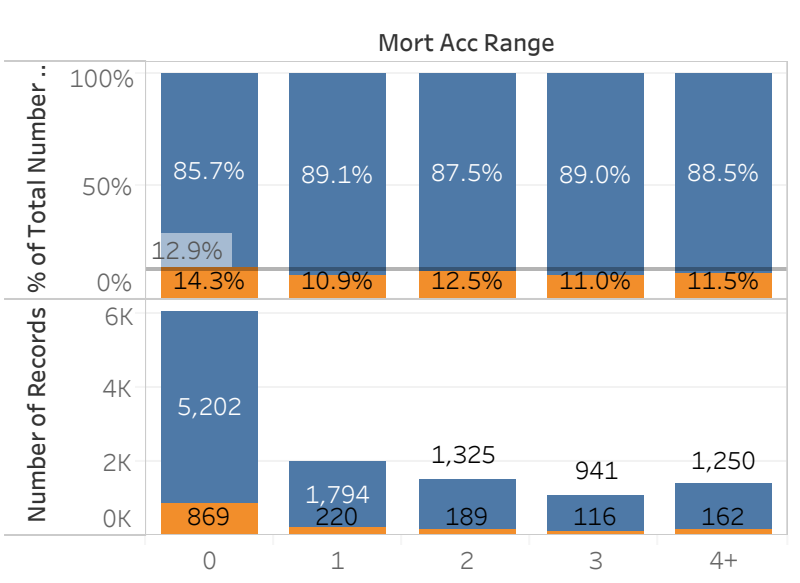
## FICO Score Factors 1




## Revolving Utilization

Anomalies

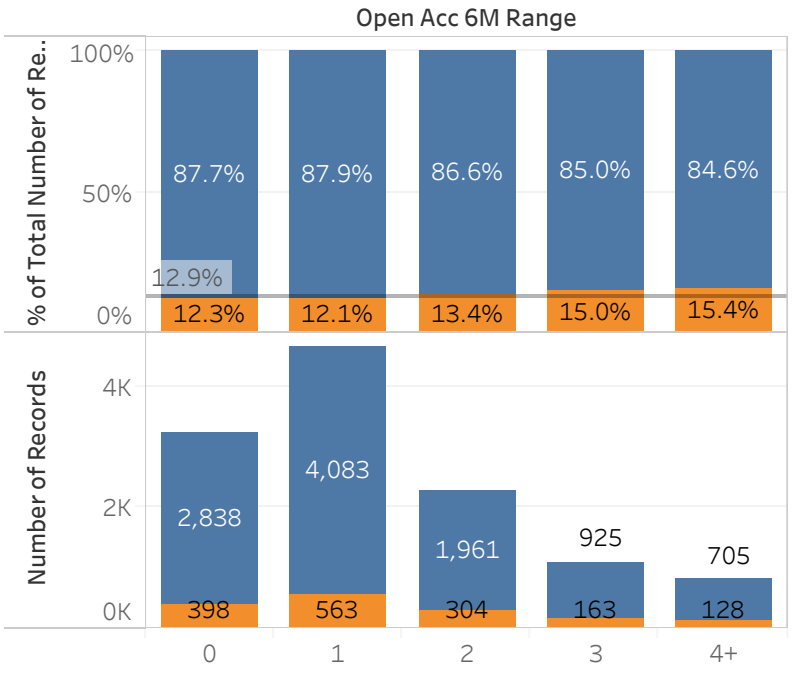
- There is also a downward trend for the late loan status percentage by revolving utilization percentage but it does not look very significant.
- After running Chi-square test, we get a p value of 0.2067, which means that there is no significant difference in the percentage of late loan status among revolving utilization percentage ranges.




## Debt to Income (DTI)

Anomalies

- For DTI Range between 0 to 20, only 11.4% of the loans have a late status but for DTI Range between 20 and 30 and DTI Range 30 to 40, 14.1% and 15.0%, respectively, of the loans have a late status.
- There seems to be a fairly significant discrepancy between the first two groups and the last two groups.
- After running Chi-square test, we get a p value o..



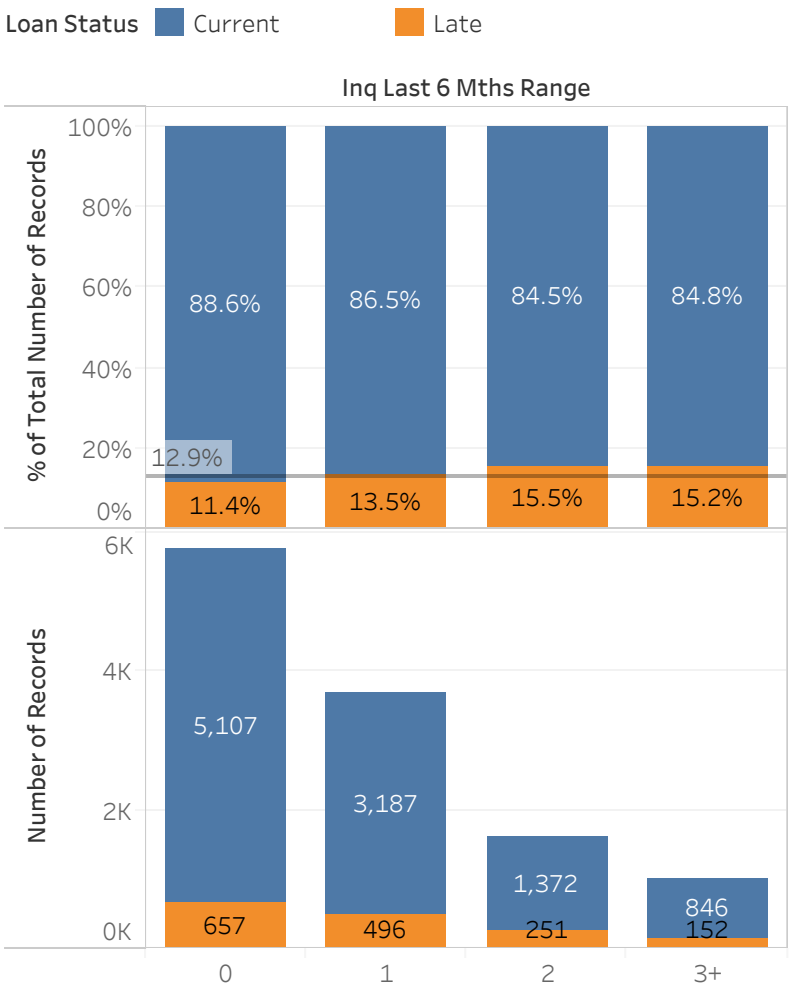

## Debt to Income (DTI)

Anomalies

- For DTI Range between 0 to 20, only 11.4% of the loans have a late status but for DTI Range between 20 and 30 and DTI Range 30 to 40, 14.1% and 15.0%, respectively, of the loans have a late status.
- There seems to be a fairly significant discrepancy between the first two groups and the last two groups.
- After running Chi-square test, we get a p value of <.0001, which means that this anomaly is statistically significant.
- We can confidently say loans with DTI that are e..

# Loan Status Analysis

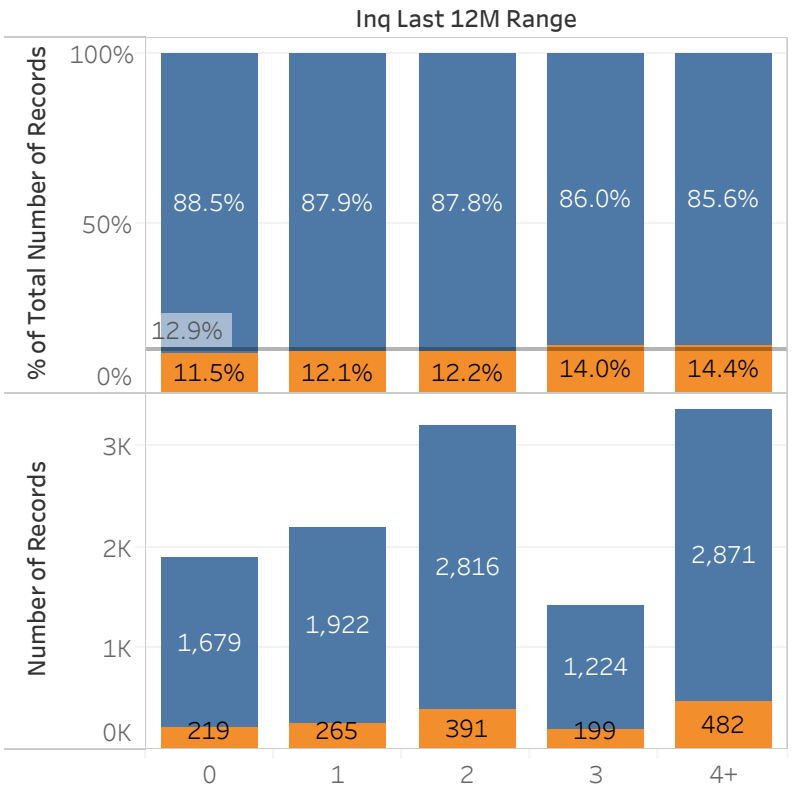
## FICO Score factors 2




### Number of credit inquiries in the last 6 months

#### Anomalies

- The loans with no credit inquiries in the last 6 months have lower percentage of late loan status than the loans with one or more inquiries
- Loans with two or more credit inquiries have a really high percentage of late loan status (15.5% for 2 inquiries and 15.2% for 3+ inquiries).
- After running Chi-square test, we get a p value of <.0001, which means that this anomaly is statistically significant.
- We can confidently say that the loans with 0 inquiries in the lat 6 months are less likely to have a late loan status than the loans with one or more inquiries.



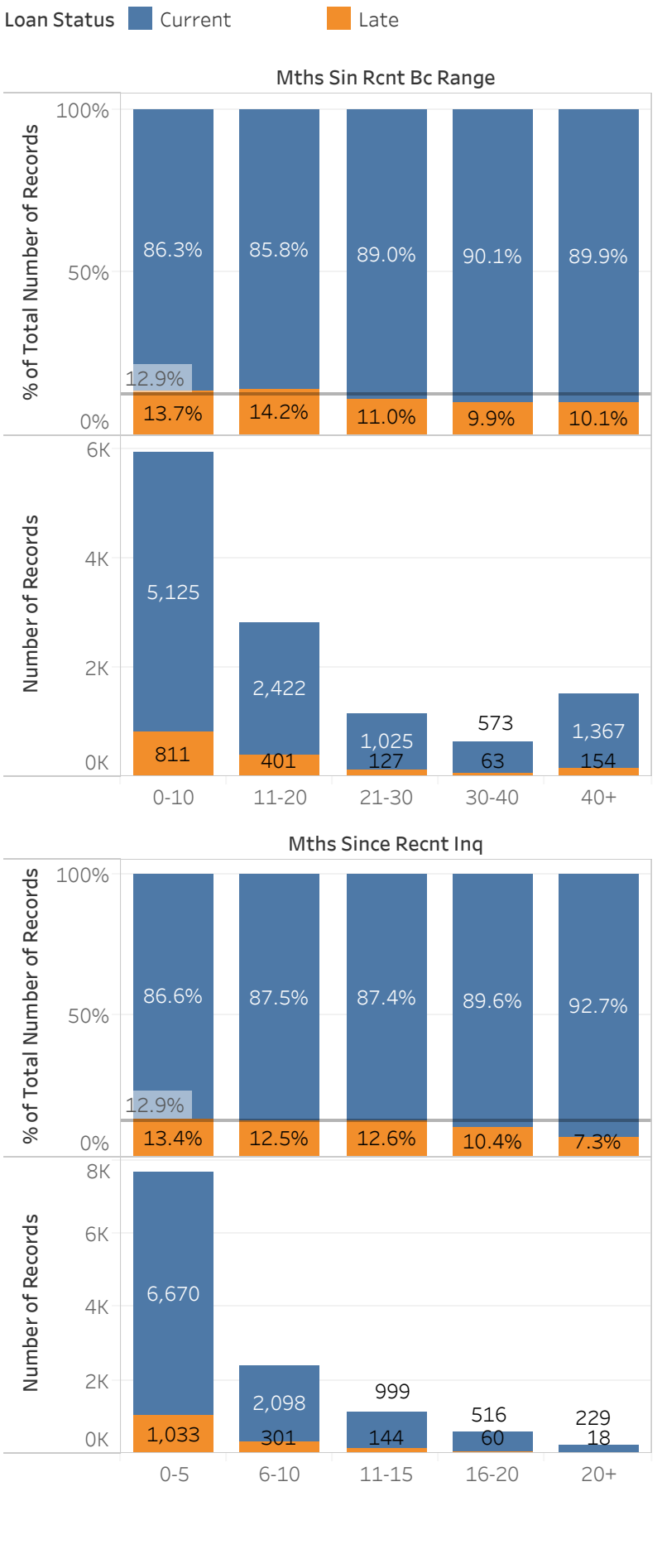

### Number of credit inquiries in the last 12 months

#### Anomalies

- The loans with no credit inquiries in the last 12 months have lower percentage of late loan status than the loans with one or more inquiries
- Loans with three or more credit inquiries have a really high percentage of late loan status (14% for 3 inquiries and 14.4% for 4+ inquiries).
- After running Chi-square test, we get a p value of 0.0082, which means that this anomaly is statistically significant.
- We can confidently say that the loans with 0 , 1, or 2 inquiries in the lat 6 months are less likely to have a late loan status than the loans with three or more inquiries.

# Loan Status Analysis

## FICO Score factors 3




### Number of months since the most recent bank card opening

#### Anomalies

- As the number of months since the most recent bank card opening increases, percentage of late loan status decreases.
- This is especially noticeable if you compare the data between the 0-10 months group and 30-40 months group (3.8% (13.7 - 9.9)difference).
- After running Chi-square test, we get a p value of <.0001, which means that this anomaly is statistically significant.
- We can confidently say that the loans with 21 months or more since the most recent bank card opening are less likely to have a late loan status than the loans with less than 21 months since recent bank card opening.

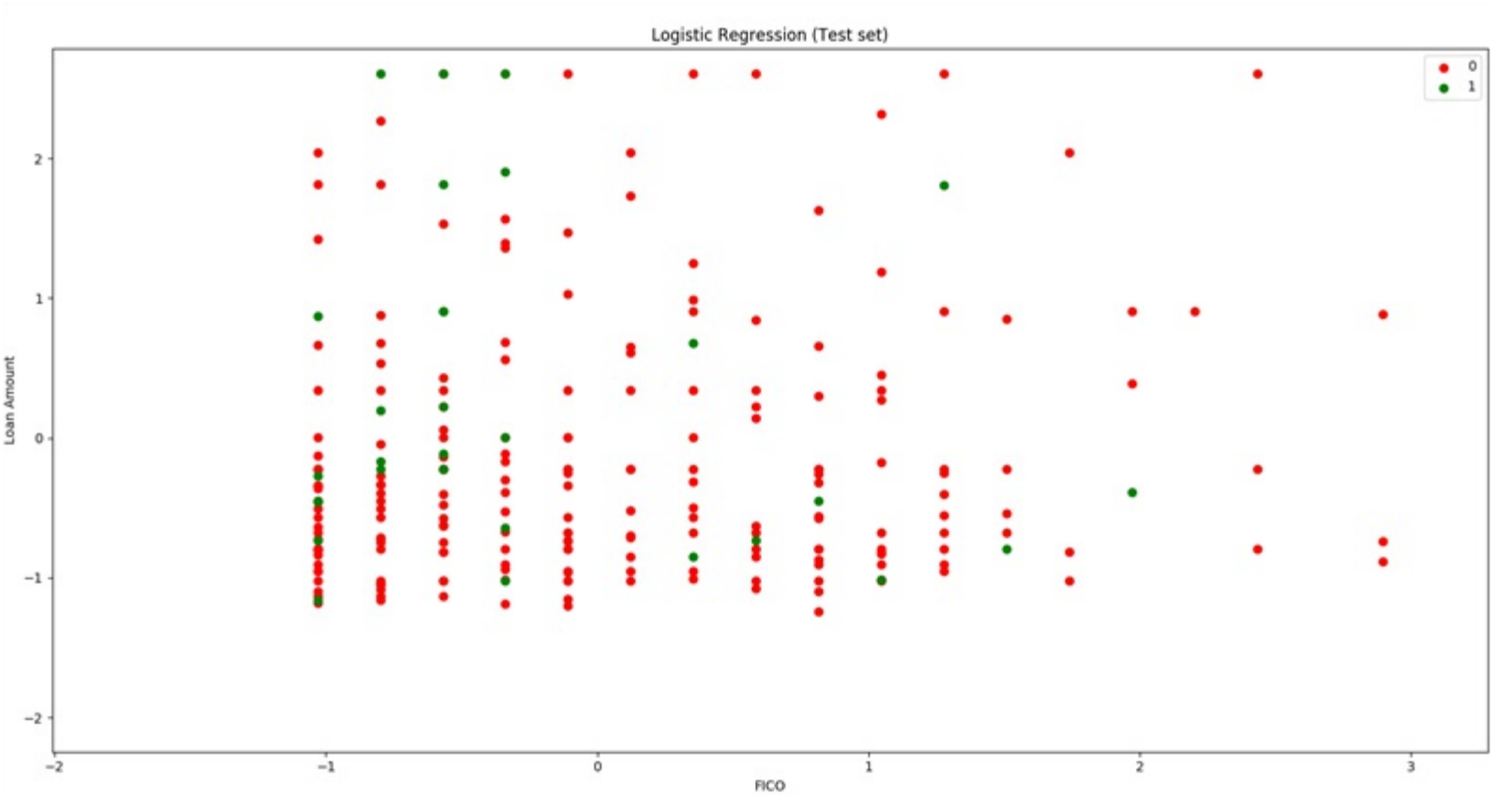
### Numer of months since the most recent credit inquiry

#### Anomalies

- As the number of months since the most credit inquiry increases, percentage of late loan status decreases.
- This is especially noticeable if you compare the data between the 0-5 months group and 20+ months group (6.1% (13.4- 7.3) difference).
- After running Chi-square test, we get a p value of 0.0157, which means that this anomaly is statistically significant.
- We can confidently say that the loans with 16 months or more and especially 20 months or more since the most recent credit inquiry are less likely to have a late loan status than the loans with less ..

# Loan Status Analysis

## Logistic Regression for Multivariables



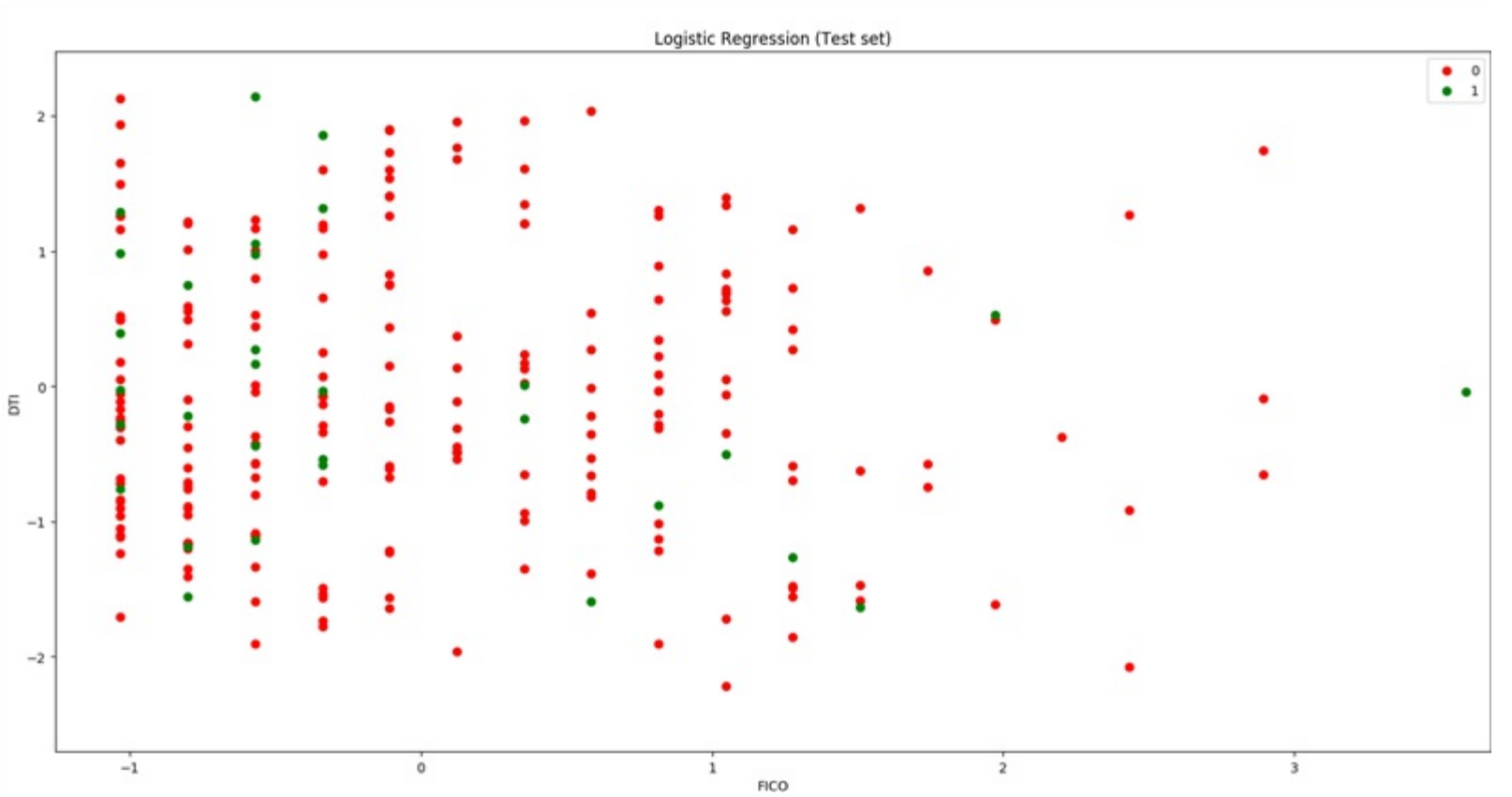

Prediction Model
0 = Current
1 = Late

- When Loan Amount is high and FICO is low, there is a high percentage of late status
- When Loan Amount is low and FICO is low, there is a decent amount of late status.
- When Loan Amount is High and FICO is high, there is barely any late status.
- When Loan Amount in High and FICO is low, there is a few late status but still barely any compared to the amount of current status present.

*No colinearity

# Loan Status Analysis

## Logistic Regression for Multivariables




**Prediction Model**
0 = Current
1 = Late

- In area where FICO is low  and DTI is high, there is quite a lot of Late compared to the number of Current
- In area where FICO Is high and DTI is high, there is barely any Late status seen
- In area where FICO is High and DTI is low, there are few late status.
- In area where FICO is low and DTI is low, it does not seem that unusual as there is a quite a bit of Current as well as Late.

*No colinearity

# Summary

## Findings

**Single Variable**

After running A/B Test and A/B/n test and validating my approach through Chi-square test, I was able to come up with some insights:

● When the loan amount was less than $6,000, only 10.4% was late to payment but when the loan amount is more than or equal to $6,000, the probability rose to around 14%.

● Higher FICO scores had lower percentage of late loan status than the lower ones. The percentage of late loan status for FICO scores that is higher than or equal to 690 was lower than the average but the percetage of late loan status for scores that were lower than 690 was above average.

● FICO score factors such as number of accounts opened and inquiries in the last 6 and 12 months, and number of months since the most recent bank card opening and credit inquiry all had similar result to that of FICO score in that fewer inquiries and accounts opened, and more months since bank card opening and credit inquiries increases FICO score and for this case, have lower percentage of late loan status than the average.

● One other factor, revolving utilization showed very strange trend in that the lower revolving utilization showed higher percentage of late loan status than the ones with high revolving utilization.

● From our intuition, the result should show the opposite because high revolving utilization negatively affects FICO. This perhaps suggests that those who borrow money more and pay back are more financially responsible (since using a credit card is technically temporarily borrowing money from the bank). However, it did not pass the Chi-square test and it was marked as statistically insignificant.

● Also looking at number of mortgage accounts opened, the more mortage accounts borrowers have the more likely they are going to pay on time, reinforcing the theory that the more money one borrows and pays back, the more likely he/she is going to be more financially responsible.

● Surprisingly, income, employment length and whether or not source of income was verified did not have much influence on whether or not loan was paid on time.

● If DTI, which is correlated to the income, is less than 20%, only 11.4% of the loans have a late status but if its bigger than or equal to 20%, 14.1% to 15% of the loans have a late status.

● Lastly, 10% of the loans that have installments, which is direct result of Loan amount, FICO score and DTI and indirectly related to all of other variable, had a big anomaly as well. Less than $200 was late to payment while around 14% for the loans that have installments that are higher than or equal to $200.

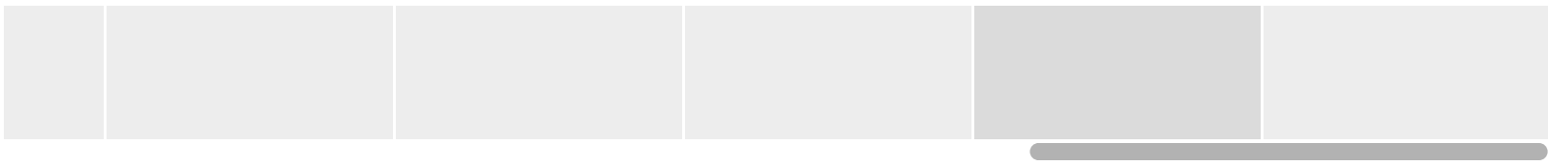
● This result is very similar to that of loan amount probably because loan amount directly influences installment.

● Also FICO score and DTI are correlated to installment since higher FICO score and lower DTI lowers installment. Lower percentage of late loan status for high FICO score and low DTI contribute to the low percentage of late loan status for small installment.

● The three variables I would choose to make a prediction model are FICO, DTI and Loan Amount as they are not correlated to each other and have good influence on whether or not the payment was paid on time.

● One other variable I would consider checking out is number of mortgage accounts. As discussed earlier, the number of mortgage account shows the ability to borrow a lot of money and pay it back. It also shows that previous lenders trusted the borrower to be able to pay back the money.

# Loan Status Analysis



## Findings

Multiple Variables

**Variables: FICO and Loan Amount**

● FICO seems to matter more as when FICO  was high, there is fewer percentage of late status compared to when FICO was low.
● If FICO is low and Loan Amount is high, almost 40% of the loans have a late status.
● Even though Loan Amount is high, if FICO is high, there seems to very low percentage of late loan.
● What is more surprising is that there is higher percentage of late status loan when DTI is low and FICO is high than when DTI is high and FICO is high, which suggests that if the borrower has high FICO score and DTI is also high, he/she is more likely to pay on time than the borrower who has same FICO score but lower DTI.

**Variables: FICO and DTI**

● Here also, FICO seems to matter more than DTI as there is a very low percentage of late status  when FICO is high.
● There is a high percentage of late loan when DTI is high and FICO is low.
● There is higher percentage of late status loan  when DTI low and FICO is high than when FICO is high and DTI is high, which is interesting.

# Loan Status Analysis

## Methodologies

Languages and Tools Used: Java, Python, R, Eclipse, Spyder, RStudio, Tableau

**Step 1) Data Cleaning**

● First, the files given were examined using Notepad++ to check the delimiter and separator.
● Imported DataSource.csv file to RStudio and opened the file to check the data was structured correctly.
● Created a script file that imported libraries ggplot2 and dplyr, and ran table, summary and histogram for all of the variables in the data to check missing data and outliers.
● Observed that quite a lot of data was missing.
● Ran Eclipse Java. imported csv reader and wrote a script file that read DataSource.csv and DataFix.csv, combined the two data files if the ID was matching using a while loop (while next line from reader.readNext() is not null), append to StringBuilder to create a new file with delimiter and separator \n, and exported the combined and fixed data as a new csv file (PrintWriter.write(StringBuilder.toString());).
● Imported to RStudio again and ran the previous test and observed that some of the data were still missing.
● The data that has some missing variables could carry other important variables so decided to replace the missing values with median from the respective column instead of mean because there are some outliers and other values in the same column are whole numbers.
● Tested for outliers and although there were some obvious extreme outliers (greater than 3rd quartile + 3 * IQR), they did not seem to be incorrectly entered so all outliers were kept.
● After the cleaning process, all 12,068 observations were kept and had no missing data.

**Step 2) Data Analysis**

● Performed A/B Testing or A/B/n Testing (an extension of A/B Testing where you test out multiple versions for one variable, in this case up to 5 versions, rather than just version A and B) on each variable except for "term", which was a constant value for all loans.
● Ran Chi squared test for independnce to test and see if the results were due to a chance or not.
● Used Python on Spyder to convert loan status to 0 if "Current" and 1 if "Late"
● Ran Logistic Regression for multivariate testing on Spyder by:
1. importing libraries numpy, matplotlib.pyplot. pandas.
2. importing the dataset.
3. declaring X (x1, x2), independent variables, and y (loan status), dependent variable.
4.Splitting the dataset into the training set and test set.
5. Applied feature scaling for accurate prediction.
6. Fitted Logistic Regression to the training set
7. Predicted the test set results
8. Made the confusion Matrix
9. Visualized the training set result and test set results
● Created visualizations with explanations and imported graph images files created from other tools into Tableau for visual interpretations.
● Organized the analysis into a presentable PDF format using Tableau.