# FINAL REPORT

# P2P Loan Lending Transaction Dataset

# Capstone Group 7

# Group Mentor:

## Mr. Jatinder Bedi

Chinmay Vaishnav

Harish Krishna Bhyravajyosula

P.Pranay

Raj Keswani

Padma

# **Table of Content**

1. **Introduction and business understanding**
2. **Problem Statement**
3. **Important Features**
4. **Data Findings and Insights**
5. **Process Overview**
6. **Data Pre-processing**
7. **Problem Solving Methodology**
8. **Model Techniques**
9. **Regression-Interest Rate Prediction**
10. **Classification- Paid off / Charged off**
11. **Conclusion**

# 

# Introduction

## **BUSINESS UNDERSTANDING**

### **Business Description**

Lending Club is an online market place where investors and borrowers meet. It is the biggest peer-to-peer lending service where the borrowers can apply for a loan within few minutes. The procedure starts by registering online for both, investors and borrowers. Lending Club employees check if investors and borrowers meet their criteria. The verification procedure takes about seven days.

After successful verification process, the borrower’s loan will appear on the platform and will be valid for 15 days. On the other hand, investors can browse available loans and decide where to invest and how to build their investment portfolio. The minimum investment is 25 USD per loan. They can view many different information before deciding where to invest. This information includes loan information (amount, purpose, grade, interest rate, length, monthly payment), borrower information (job title, income, state), and credit information.

Borrower’s personal details such as first and second name, address, personal number are not available to investors. Lending Club has been in the business since 2007, and has issued over 26 billion dollars in loans to more than 1.5 million borrowers.

The given dataset and the problem statement deal with Loan lending and the risks involved in it that the investment could go bad. Risk has always been a part of finance and investment. But we’ll see how accurately can risk be predicted in order to identify the high risk investments from the low risk ones, the performing from the non-performing ones. During this analysis we try to make investment more transparent for the investors or the lenders so that they can judge better before investing.

**[](https://cdn.crowdfundinsider.com/wp-content/uploads/2019/05/LendingClub-Business-Model.png)**

# PROBLEM STATEMENT

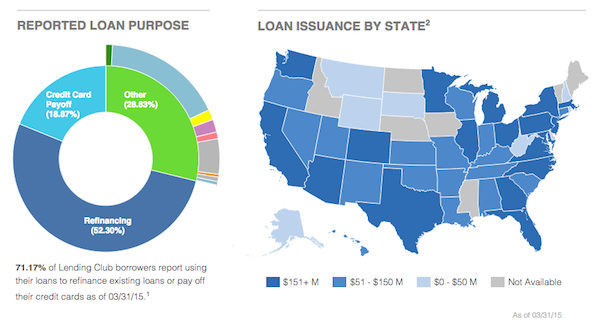
The objective is to analyse P2P Lending loan transaction data from one of major US market player for over a decade and build model for identification and recommendation of borrowers/loan for an investor.

The main aim is to predict the below:

* Predict the interest rate of potential lenders
* Predict the probability of default for a potential loan.

**Objective**:

The exercise is to build model for potential (individual) investors, to diversify and advise for best possible loan funding options. This will enable them to efficiently leverage the P2P Lending platform for optimal ROI.



# Important Features:

* **Debt to income ratio**:

A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.

* **Employment length:**

Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

* **Funded amount:**

The total amount committed to that loan at that point in time.

* **Loan amount:**

The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.

* **Delinquency Amount:**

The past-due amount owed for the accounts on which the borrower is now delinquent.

* **Grade:**

LC assigned loan grade

* **Installment:**

The monthly payment owed by the borrower if the loan originates.

* **Interest rate:**

Interest Rate on the loan

* **Settlement Date:**

The date that the borrower agrees to the settlement plan

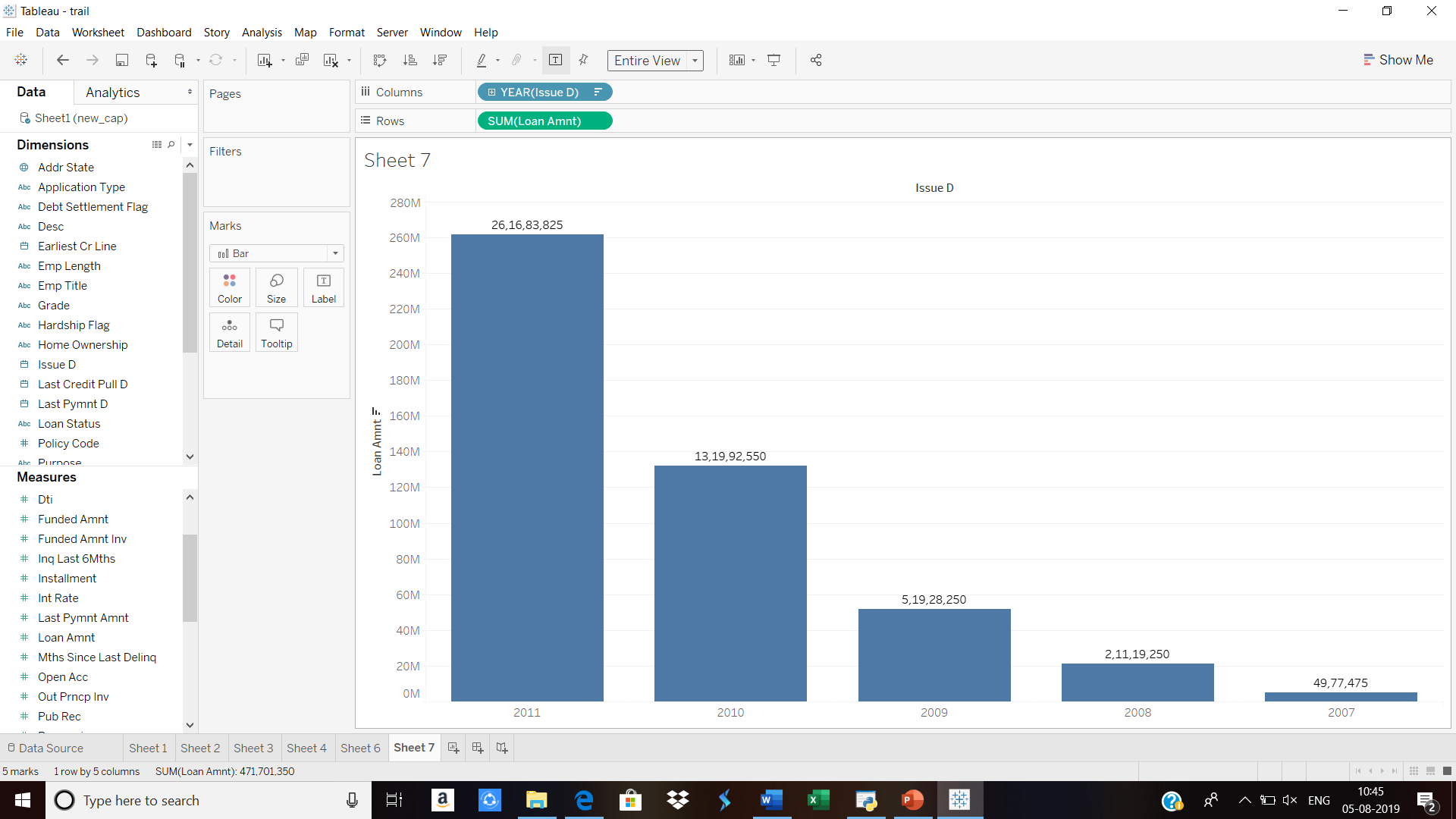
* **Total payment:**

Payments received to date for total amount funded.

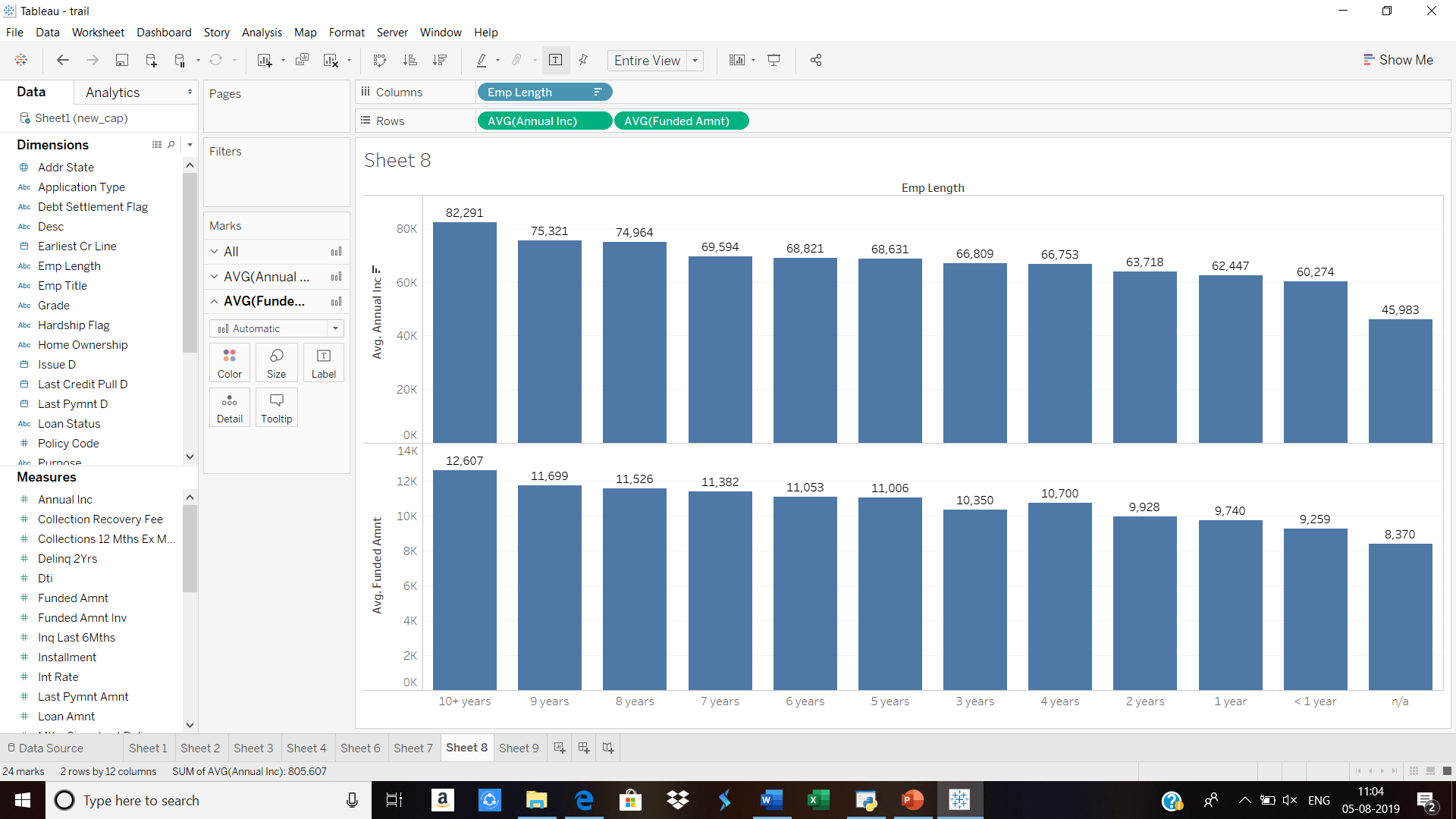
* **Total\_pymnt\_inv:**

Payments received to date for portion of total amount funded by investors

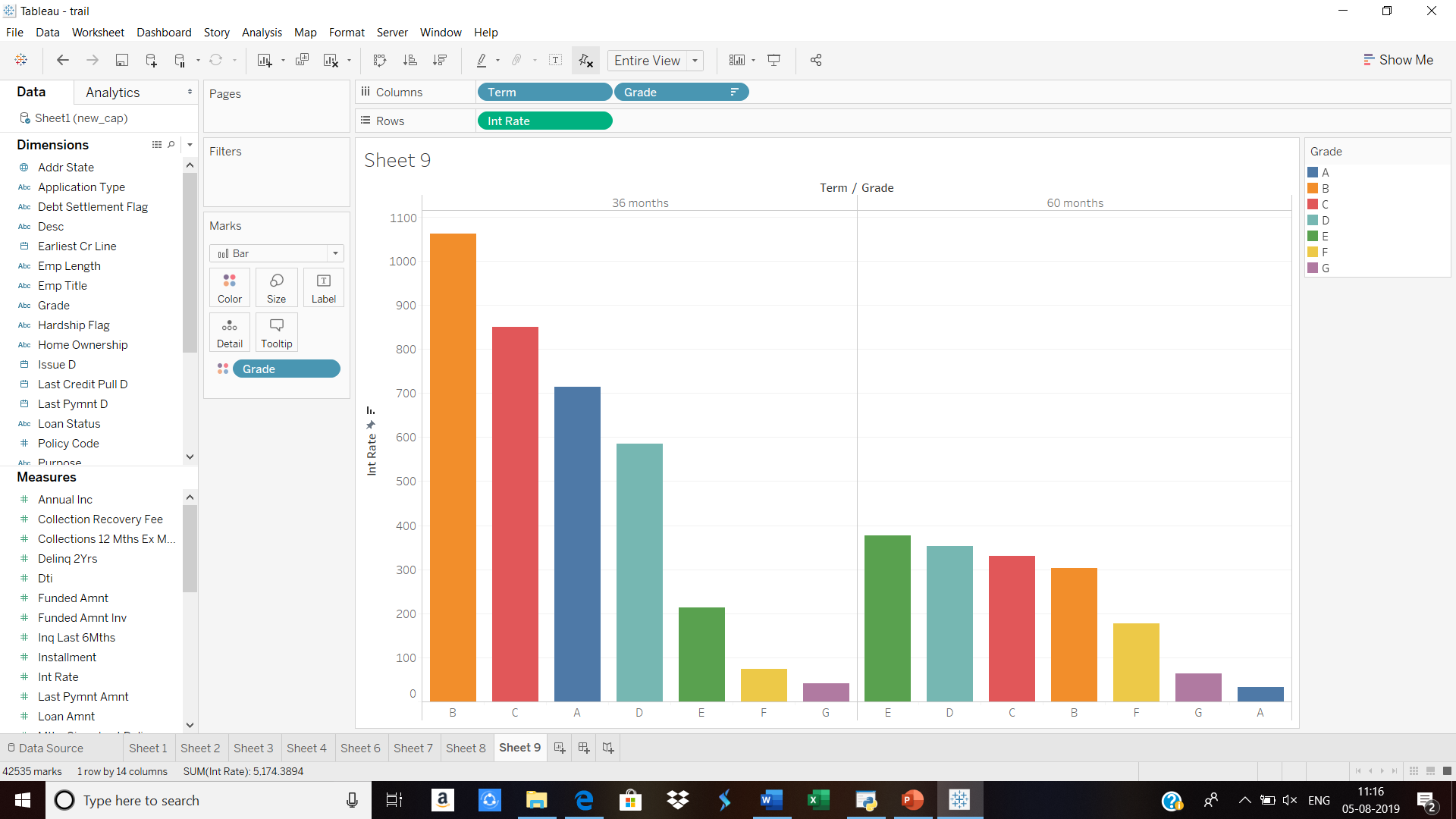
# Data Findings



The amount of loan dispersed in the 2011 is the highest, with an increasing trend. We can infer that the loan amount for provided by the lender’s ranges from 10000-15000$.

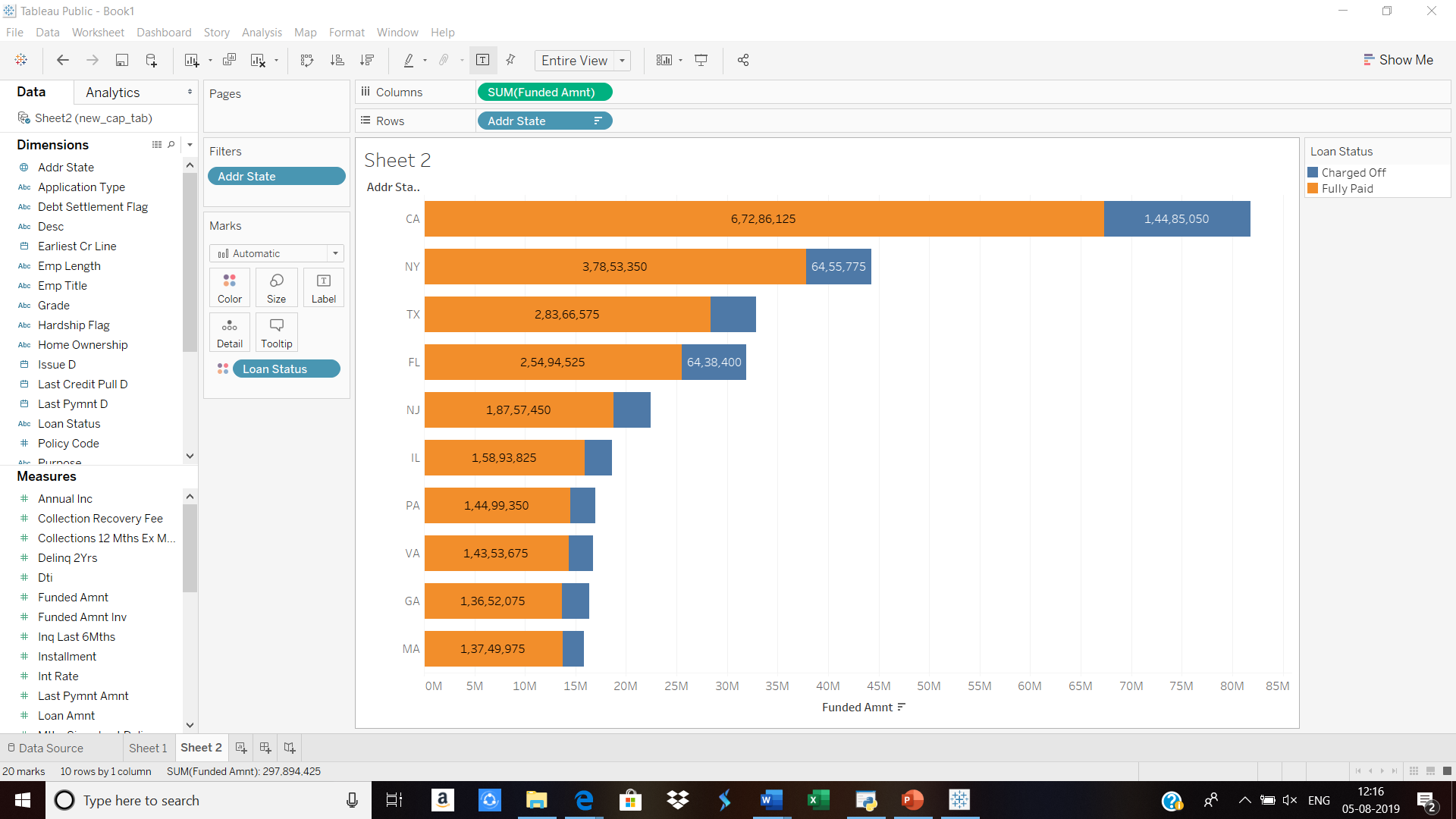


It’s observed that the average loan amount funded was towards the category of borrowers whose annual income is above 82291$ and are holding 10years + experience. The average amount funded is 12607$. Above is the visualization showing the different groups of borrowers



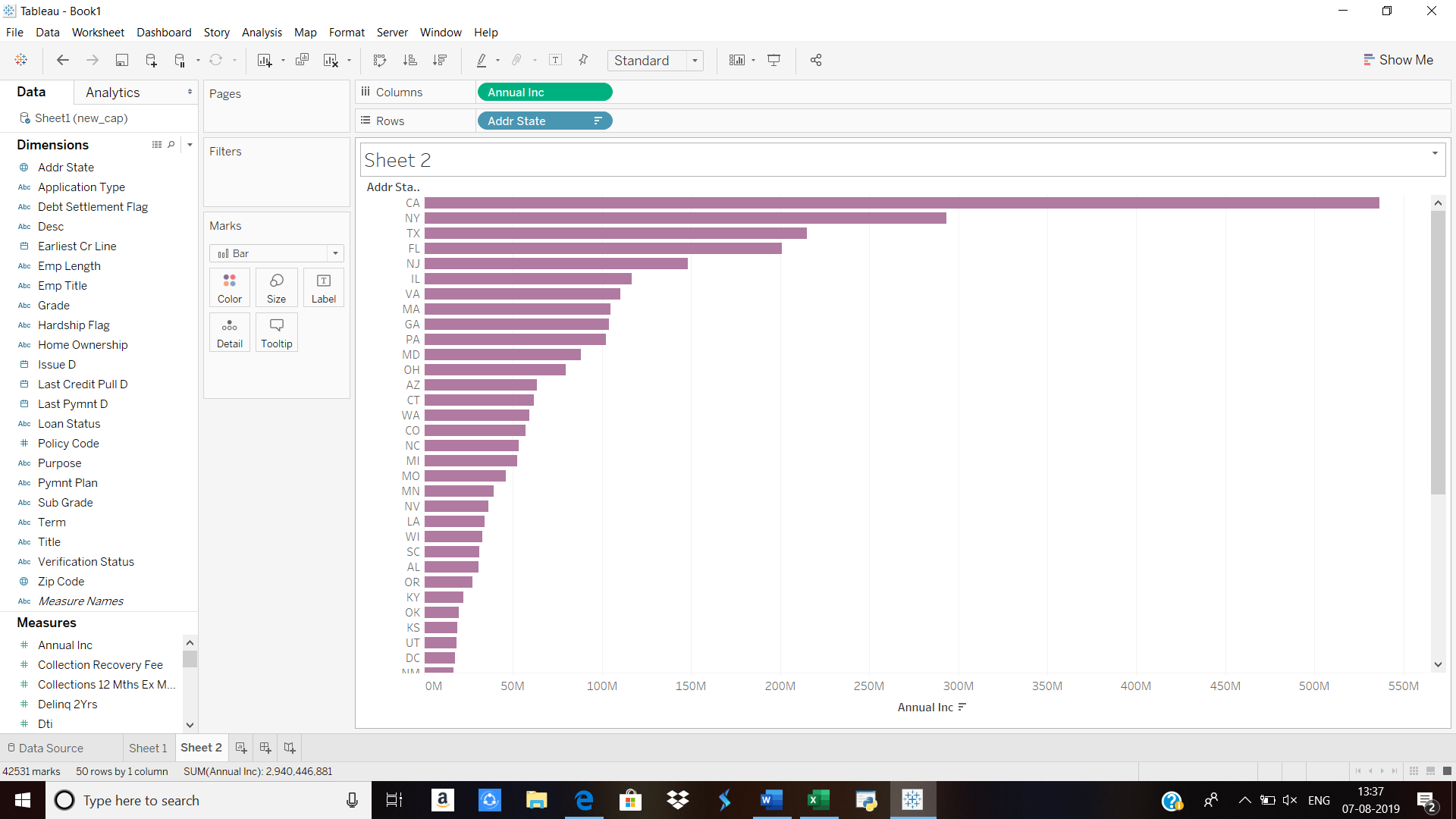
**Finding:** Majority of the borrowers have terms of 36 months who are in grade B.

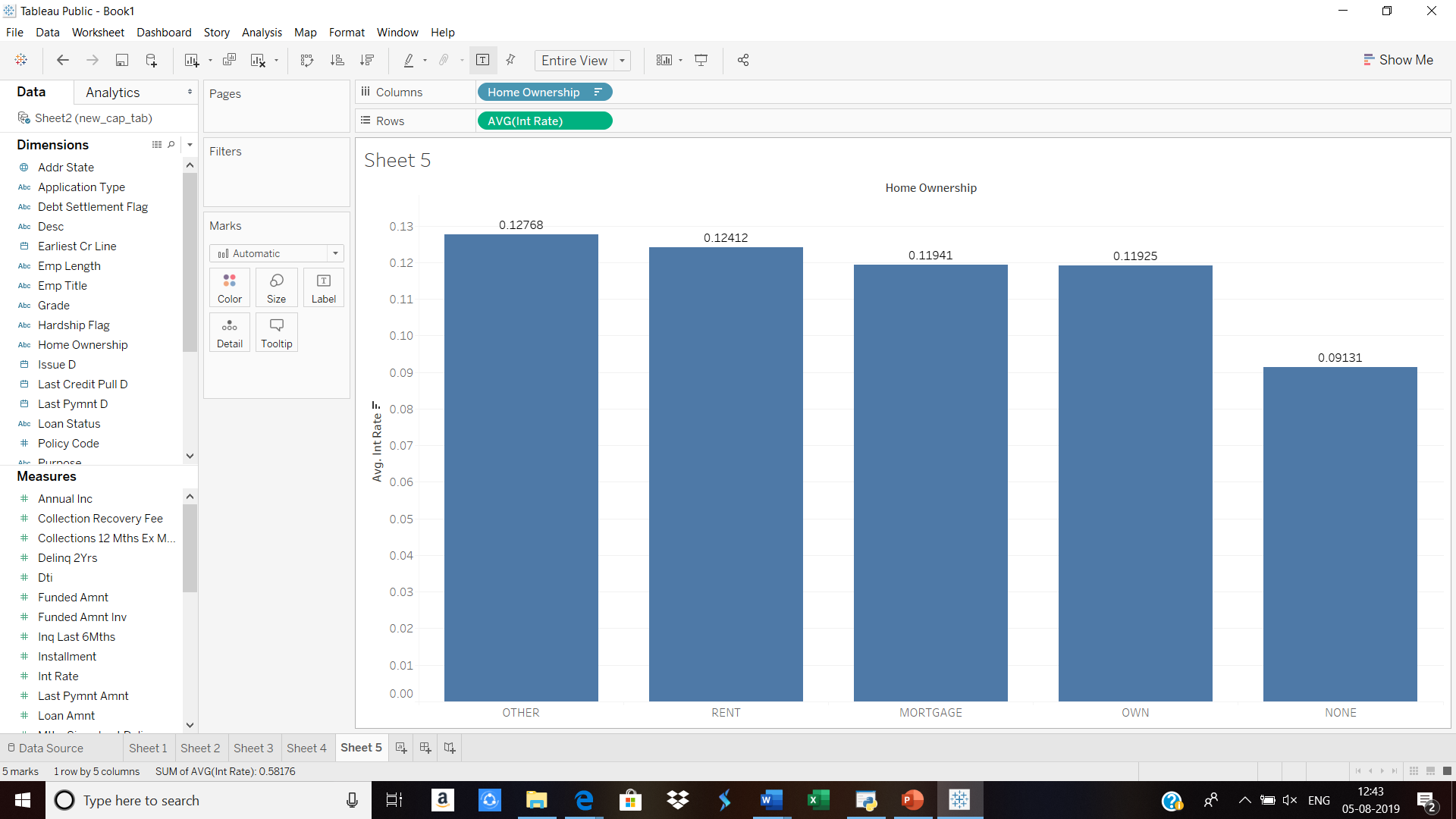
**Insight:** Borrowers corresponding to loans with 60-month terms deliver higher returns compared to those with 36-month terms. Keep in mind investing in longer term notes adds both risk and return exposure, borrowers may wish to invest in both terms to diversify.



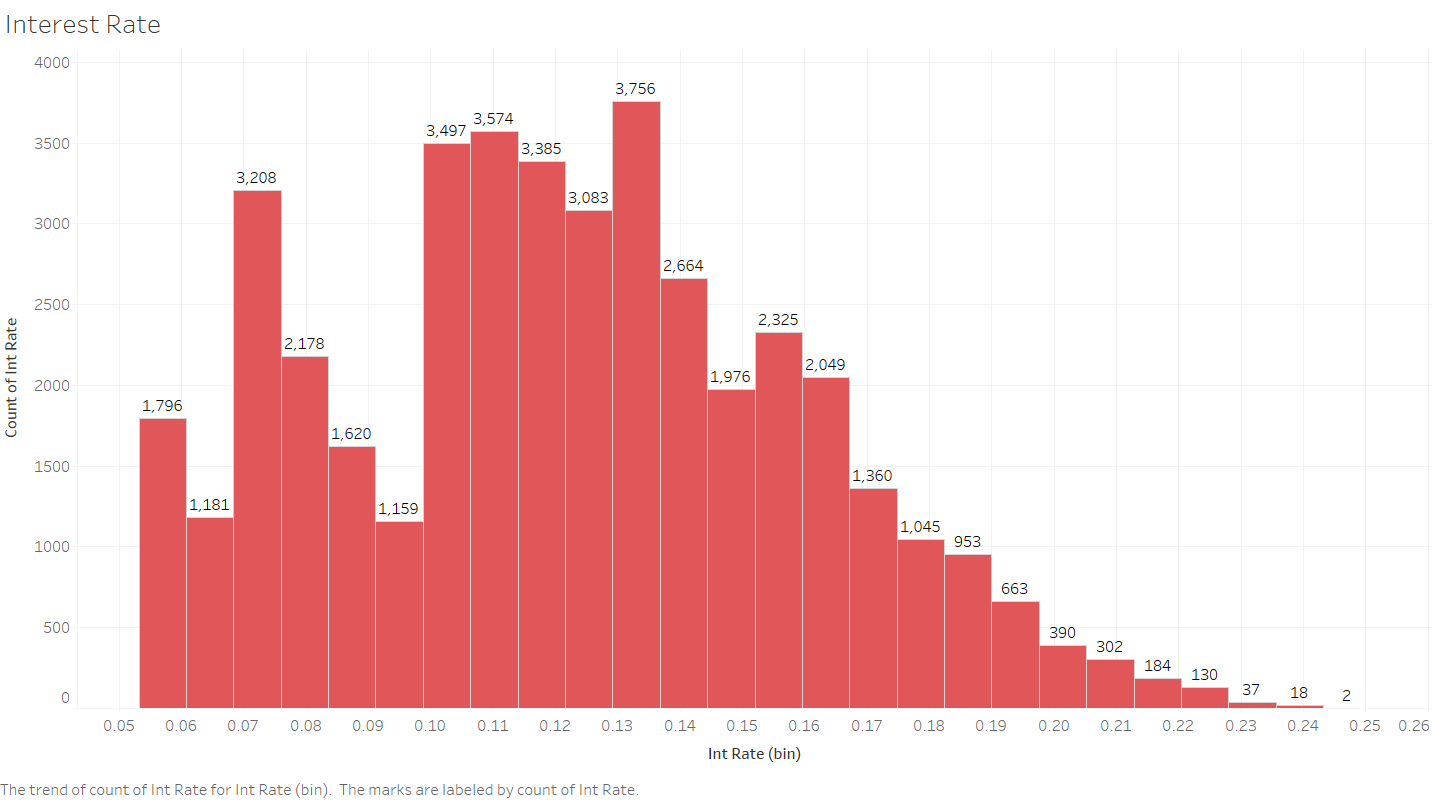
Above is the visulaization of top 10 states in US with the loan amount funded and also amount that was paid back versus charged off. The below visulaization shows the average annual income with the states, it tells that the loan funding have used **Average annual income as a key feature** to understand where there is potential market for lending club.California, Texas, New York and Florida are the states in which the highest amount of loans was issued.

**Insight:** Interesting enough, all four states have a approximate **interest rate of 12.19%** which is at the same level of the average interest rate for all states (12.17%) California, Texas and New York are **all above the average annual income**, this might give possible indication why most loans are issued in these states.

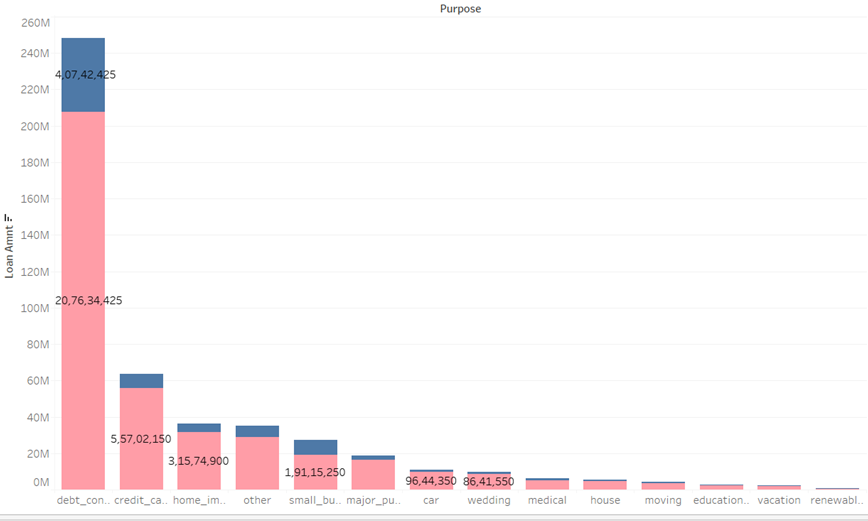




The interest rate offered based on the home ownership as below shows that borrowers who are staying own property, mortgage have been offered 11.9%. borrowers who are staying in rent and other have been offered a rate of 12.4%.



Distribution of interest Rate (Target variable): image is right-skewed, we can infer that majority of the borrowers are paying an interest rate between 10-16%.

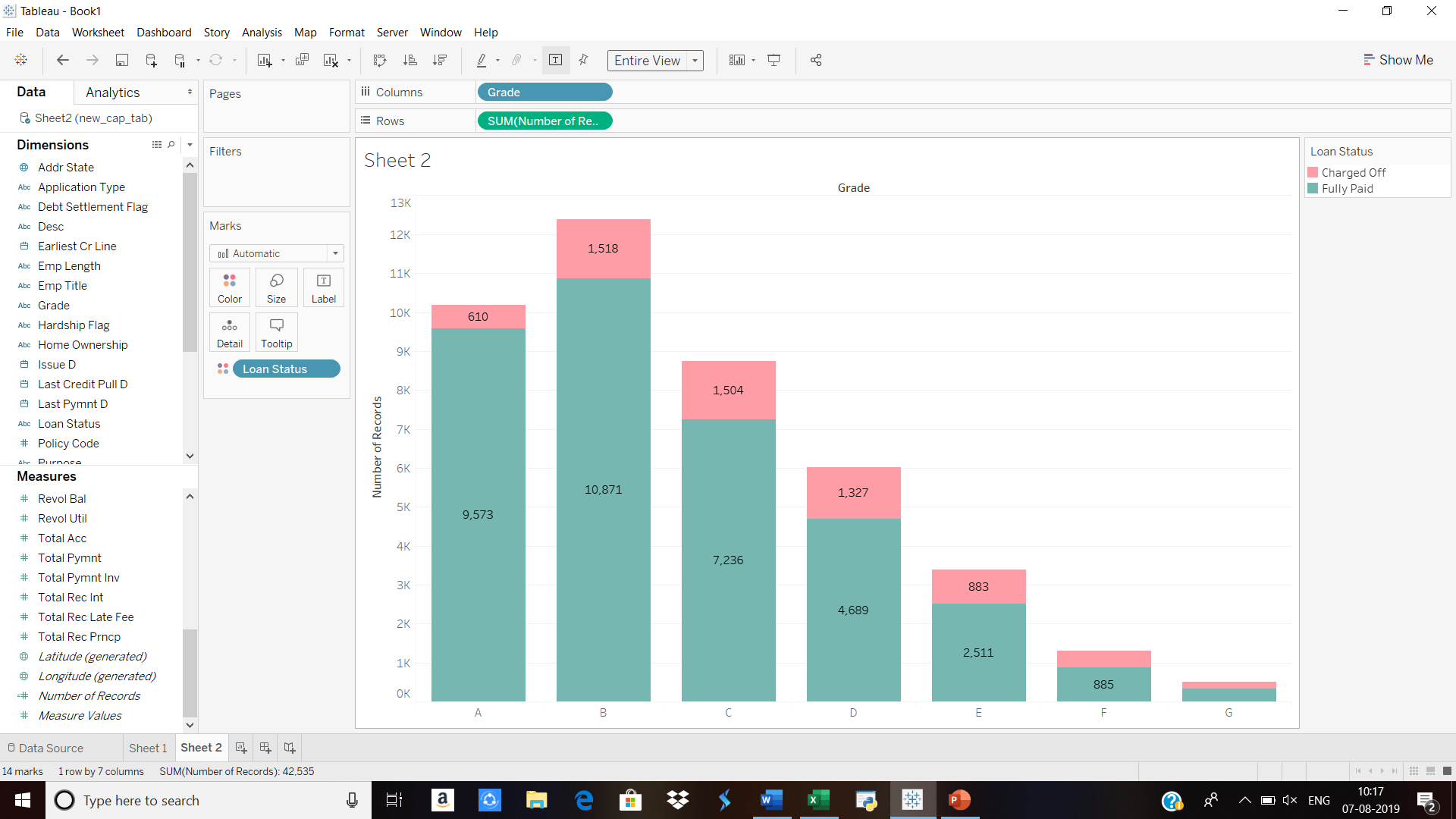


The next interesting analysis is to check distribution of loans by purpose. There are 14 different factors for this variable. Above visualization shows most common purposes of taking loans by borrowers at Lending Club. We can see that more than half of loans are taken due to debt consolidation, while around 20% for repaying credit card debts.

On the other hand, we can see that the riskiest loans are for small businesses as 30% of these loans are not fully paid. However, these loans make only 1.4% of all loans.

We can conclude that data make sense since this kind of loans are mostly created for refinancing previous debts and debts on credit cards since credit cards have higher interest rates. In addition, we see that house credits make only 0.5% of all credits which makes sense since banks offer much better loans for houses. Therefore, we can ensure investors that data is valid.

Distribution of loans by grading scores and loan status



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Grades** | **A** | **B** | **C** | **D** | **E** | **F** | **G** |
| **Fully-Paid** | 9573.00 | 10871.00 | 7236.00 | 4689.00 | 2511.00 | 885.00 | 173.00 |
| **Charged-off** | 610.00 | 1518.00 | 1504.00 | 1327.00 | 883.00 | 416.00 | 339.00 |
| **% of fully-paid** | 94.01 | 87.75 | 82.79 | 77.94 | 73.98 | 68.02 | 33.79 |
| **% of charged off** | 5.99 | 12.25 | 17.21 | 22.06 | 26.02 | 31.98 | 66.21 |
| **Total Grade%** | 24.06 | 29.27 | 20.65 | 14.21 | 8.02 | 3.07 | 1.21 |
| **Total** | 10183.00 | 12389.00 | 8740.00 | 6016.00 | 3394.00 | 1301.00 | 512.00 |
| **Average Interest rate** | 7.34 | 11.01 | 13.5 | 15.58 | 17.38 | 19.27 | 20.45 |

From the above visualization and table, we can see that 24% of all loans belongs to best grading score A, and that only 5.9% of A loans are not paid.

Moreover, we can see that for last two grades F and G percentage of unpaid loans are 31.9% and 66.2%. However, F loans make only 3% of all loans, while G loans make 1.2% of all loans. Therefore, we can conclude that loans grading is correct as the percentage of unpaid loans increase as grade score quality decreases.

Also, investors can be ensured that Lending Club takes seriously credit approval since only 4.2% of all loans belongs two lowest scoring grades. In addition, we used only two outcomes of response variable (Fully Paid and Charged Off) so we are not aware if only few instalments are not paid for these poor loans. We can include other outcomes of response variable in our analysis, but that is not necessary as only 4.2% of all loans are classified in F and G.





**Insight:**

What is diversification?

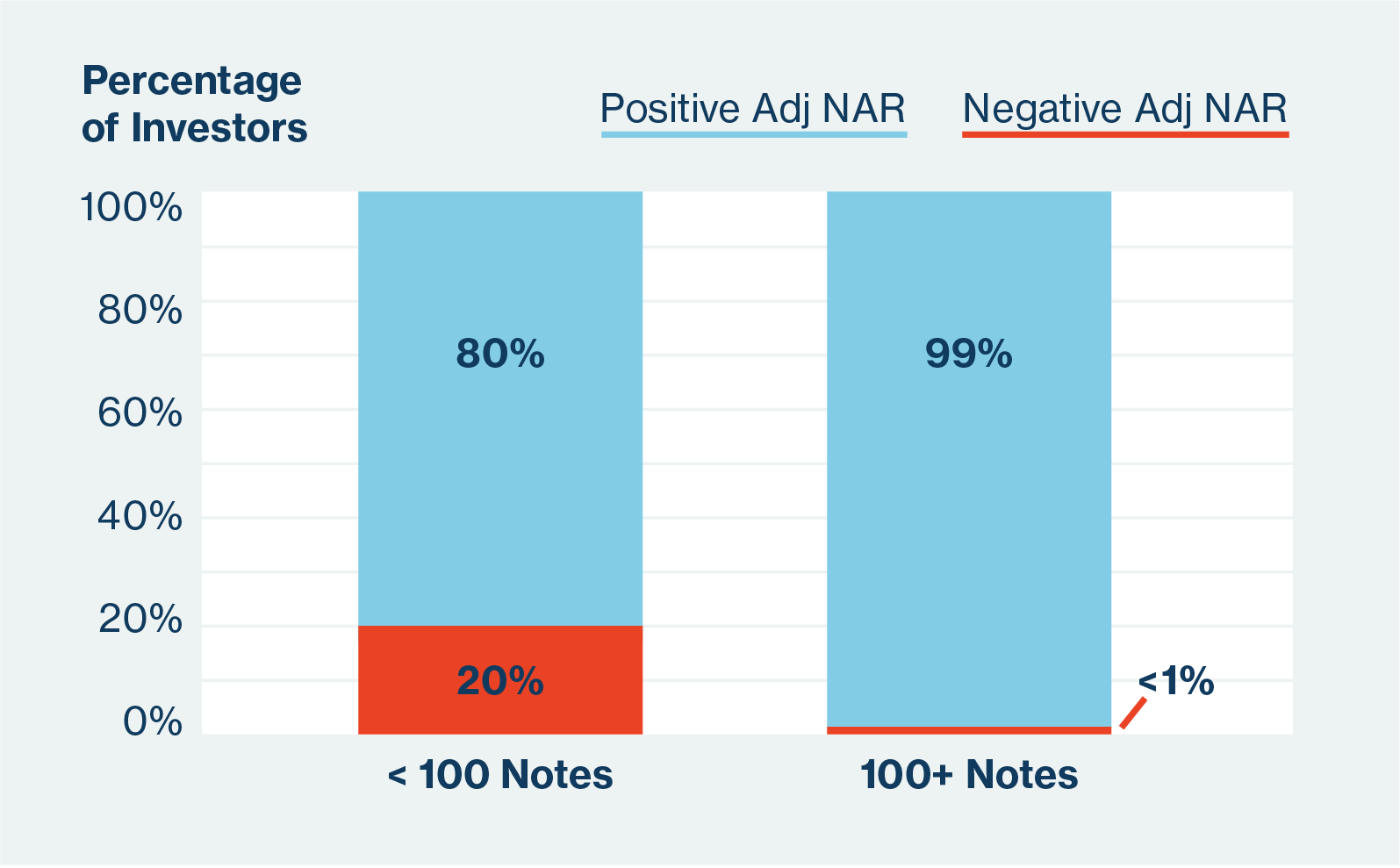
Diversification is a way to manage investment risk by spreading your dollars across many different investments to reduce the exposure to and the risk of a single investment. Investing in a combination of assets that are not correlated can lead to a return with lower volatility and less unique risk.

### Example of the power of diversification

Diversification helps to limit the impact of any single charge off by spreading your money across many different Notes. For example, say you have $2,500 to invest in Lending Club Notes. You could invest:

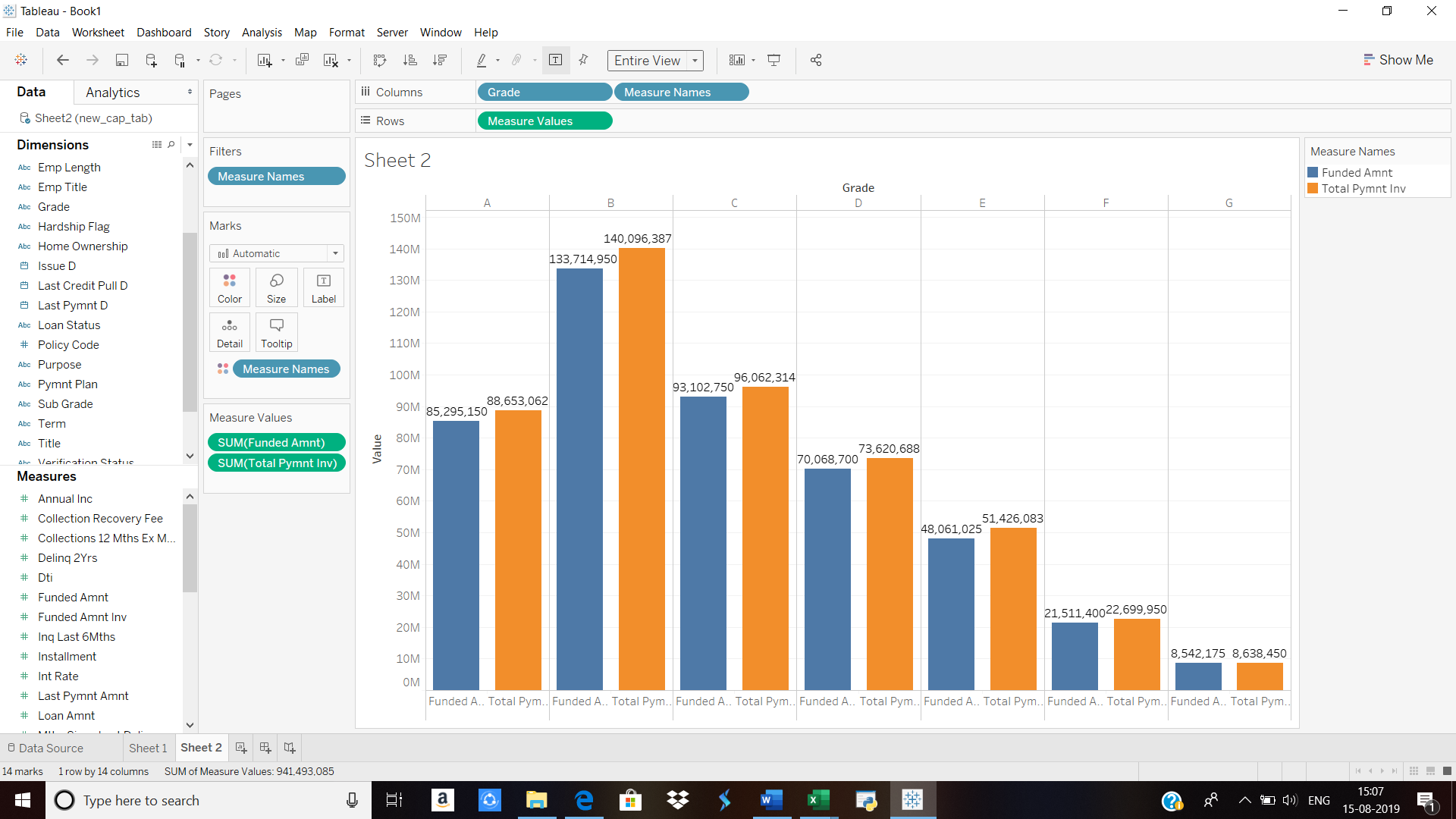
* $2,500 in one borrower; or
* $25 in 100 different borrowers.

If you invest in one borrower and that borrower becomes late and the loan eventually charges off, you could potentially lose 100% of your total investment amount. If you invested a relatively equal amount in 100 different borrowers and that same borrower becomes late, your potential loss on that particular Note would be limited to 1% of your total investment amount.



**Automated investing:** You set your investment criteria and Automated Investing will place orders for Notes as matching inventory becomes available.

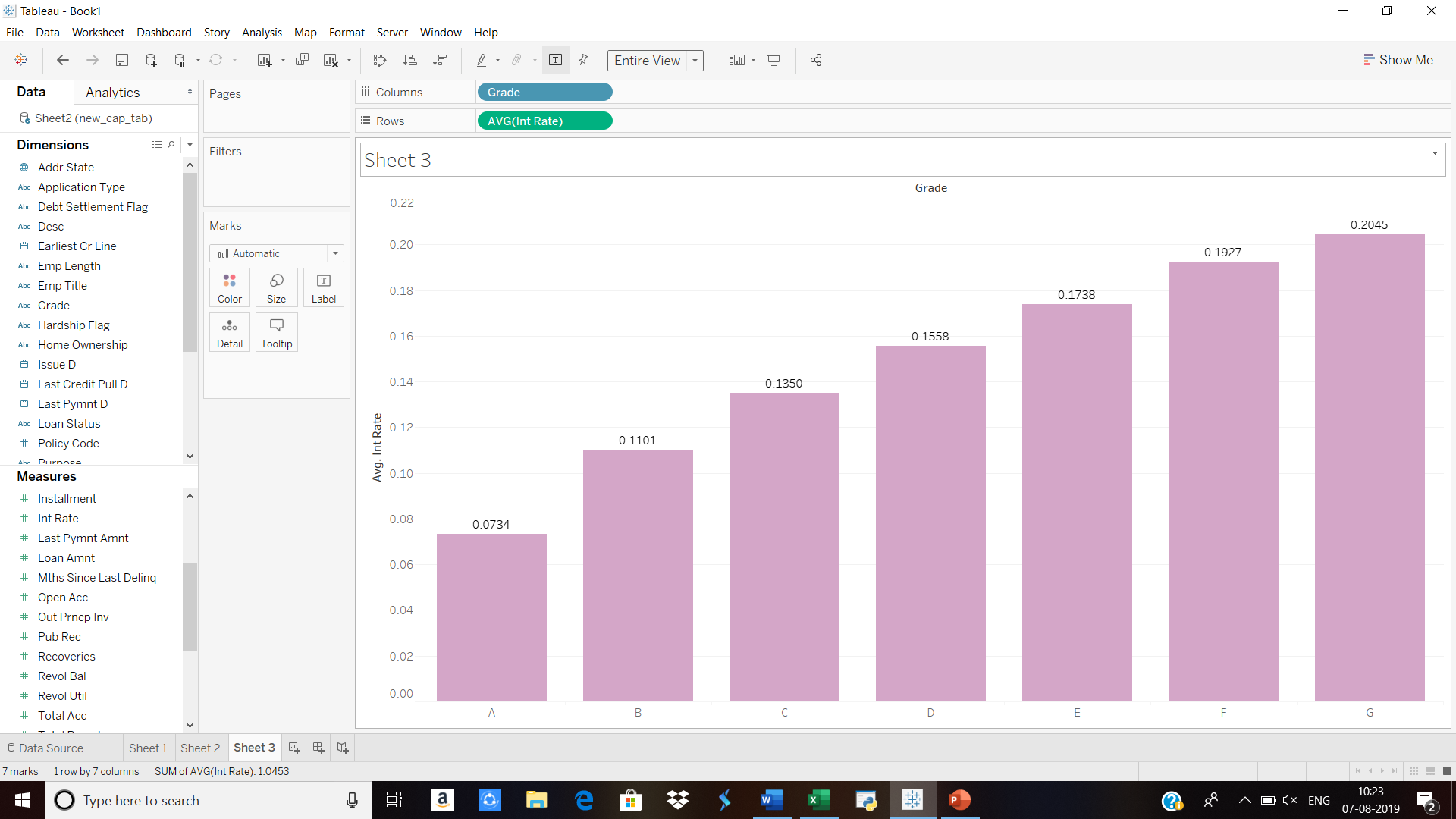
**Browse loans:** Browse the loans listed on our platform and manually pick the borrowers you want to invest in. You will find credit and loan information in each listing.

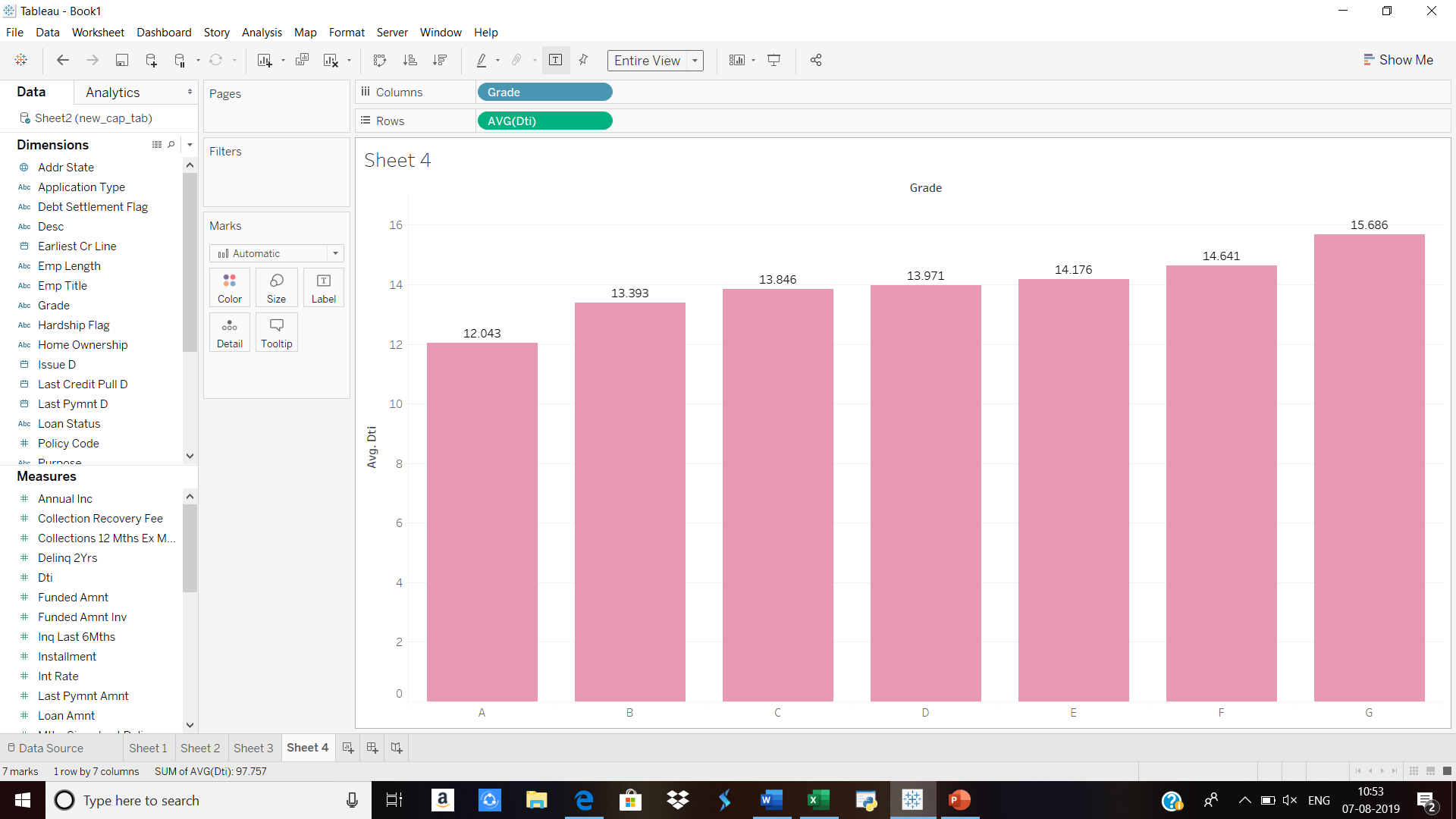


|  |  |  |  |
| --- | --- | --- | --- |
| **Grade** | **Measure Names** | **Measure Values** | **% of returns** |
| **A** | Funded Amnt | 8,52,95,150 |  |
|  | Total Pymnt Inv | 8,86,53,061.55 | 103.9368142 |
| **B** | Funded Amnt | 13,37,14,950 |  |
|  | Total Pymnt Inv | 14,00,96,387.45 | 104.7724188 |
| **C** | Funded Amnt | 9,31,02,750 |  |
|  | Total Pymnt Inv | 9,60,62,313.52 | 103.1788143 |
| **D** | Funded Amnt | 7,00,68,700 |  |
|  | Total Pymnt Inv | 7,36,20,688.40 | 105.069294 |
| **E** | Funded Amnt | 4,80,61,025 |  |
|  | Total Pymnt Inv | 5,14,26,083.42 | 107.0016368 |
| **F** | Funded Amnt | 2,15,11,400 |  |
|  | Total Pymnt Inv | 2,26,99,950.29 | 105.5252112 |
| **G** | Funded Amnt | 85,42,175 |  |
|  | Total Pymnt Inv | 86,38,450.04 | 101.1270553 |

**Insight:** The bigger the risk, the bigger the potential reward. At Lending Club, this means that riskier Notes such as grades D & E, have historically delivered a range of returns with higher highs and lower lows compared to grades A and B. To balance risk and return, investors can create Lending Club portfolios with a mixture of Note grades.

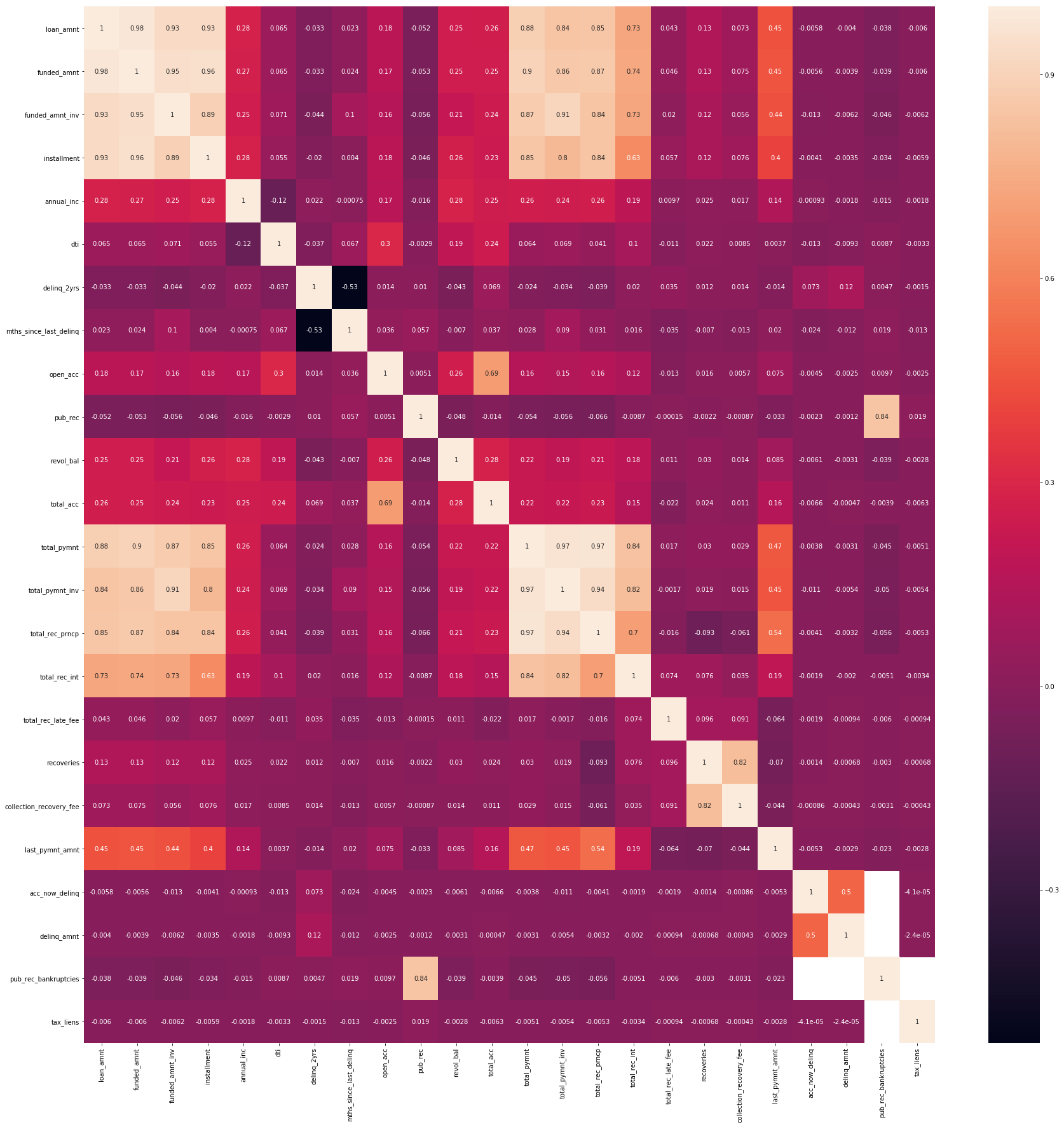
Distribution of interest rate by grading scores

Grade wise interest rates: Better grades should have lower interest rates, while poorer grades will have higher interest rates.



Grade-wise DTI ratios: The DTI ratio is one of the metrics that lenders, including mortgage lenders, use to measure an individual’s ability to manage monthly payments and repay debts. The average DTI ratio ranges between 28% - 35 %. More than this range is considered as risky. DTI is highest for Grade G. The reason is due to the high expenses, However DTI ratio is less than 28 %, so it can be considered is as safe. The income level of Grade G is high, because of this reason the DTI for Grade G is less than threshold of 28%

*Formula to calculate DTI= Total of monthly debt payments/Gross monthly income*



Used co-relation matrix to Identify the relationship between the variables, find the above visualisation for the same.

# Overview of the final process

**Problem solving methodology:**

* Used tableau for visualizations along excel and python for inferring insights from data.
* We would be trying linear regression as our base model for predicting interest rates and further build on it.
* For classification we would use logistic regression as the base model and build further model on it.
* Ensemble methods would be used on top of the base models to improve our accuracy of the models.

**Pre-processing steps:**

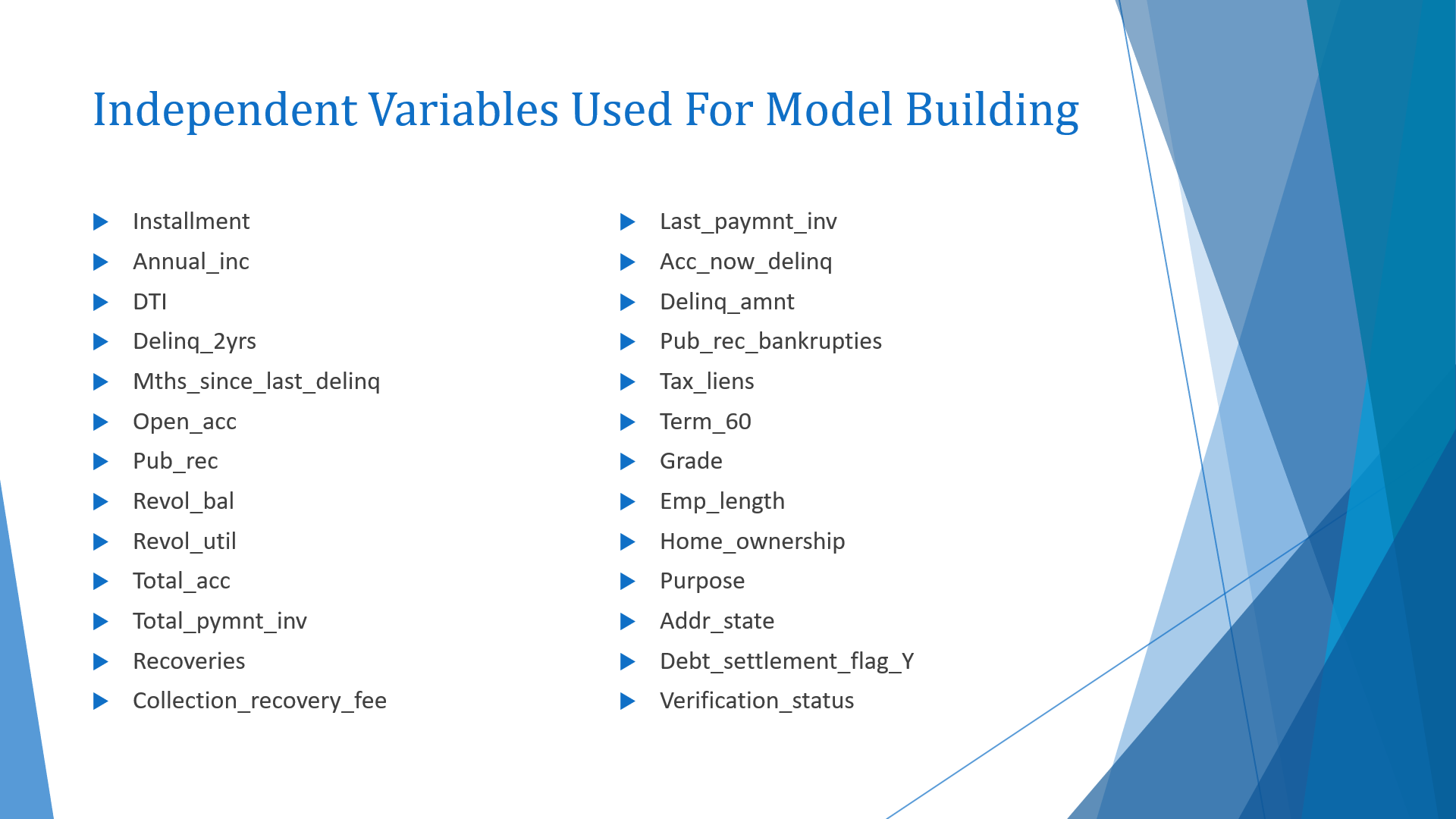
* Removed 120 features as they had more than 90% of null values, certain columns such us id, payment plan, etc that are important has been dropped.
* For categorical features such as terms, grade, purpose, etc were converted using pd.get\_dummies.
* For feature state were categorized region-wise, like west, mid-west and north-east.
* Null values of annual income were imputed with median values.
* Column Emp\_length had missing values and was imputed with 0-2 years of experience.
* The target column for regression is Interest Rate and for classification is loan status.

# Modelling Techniques

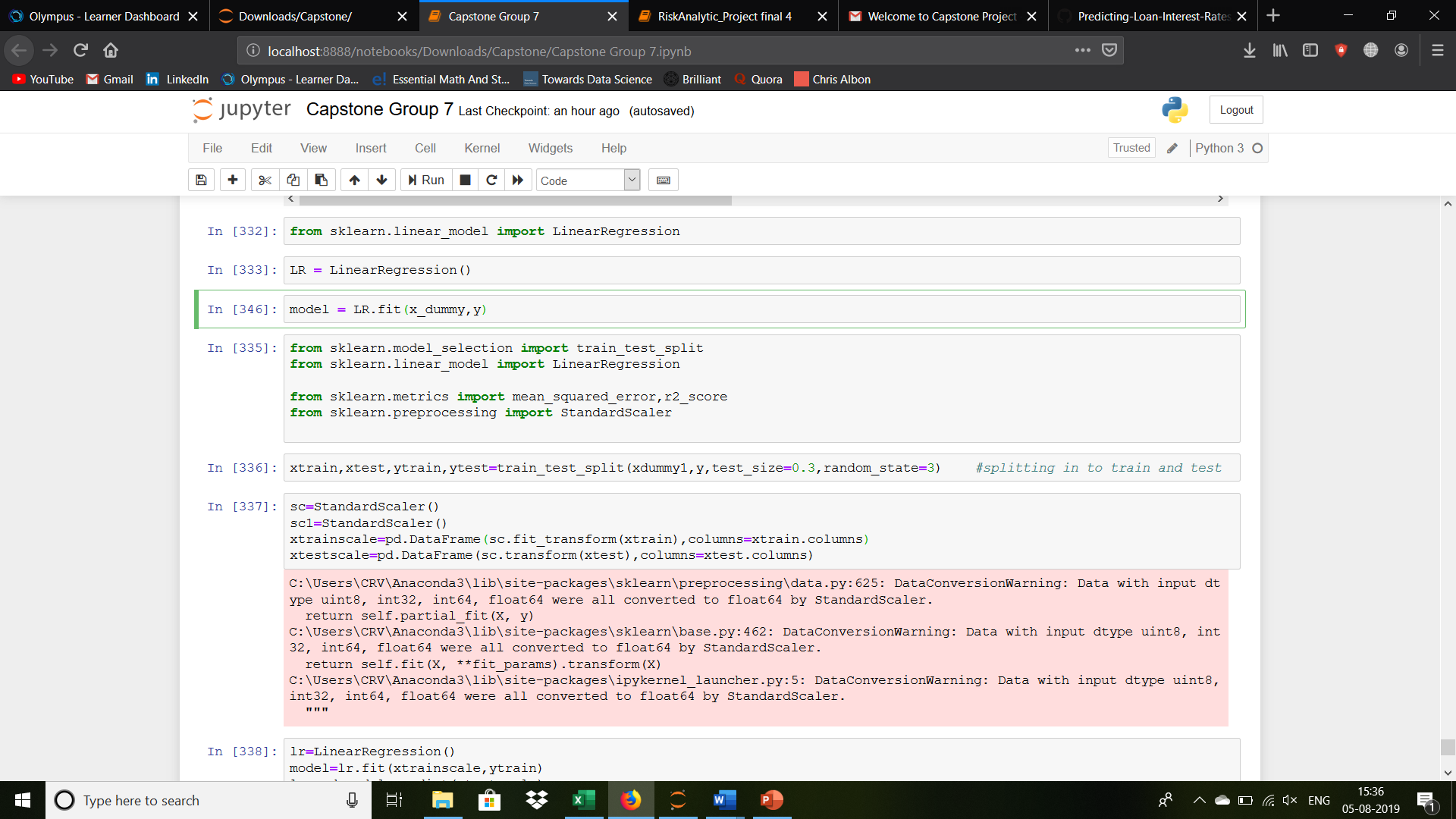
**Regression- Predicting interest rates.**

We started off building a Linear Regression model for predicting the interest rates for the potential lenders based on 2007-12-year data.

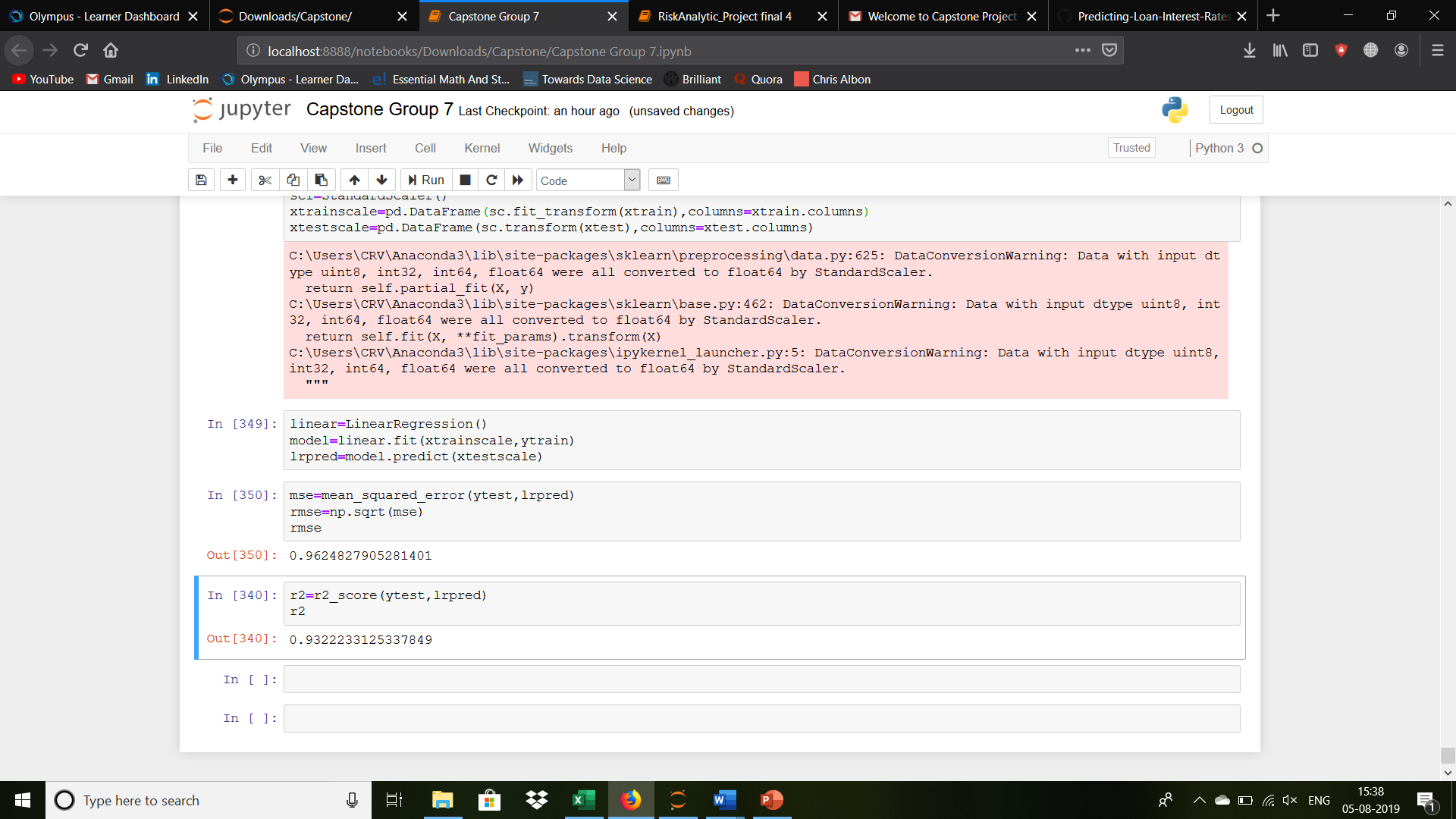
Features used for building a regression model are:



* Split data into train and test (70,30)
* Scaling of data using preprocessing.StandardScaler() to bring all the attributes to same unit.



* Fit model using Scaled data.
* Evaluating the model.
* Compile the RMSE square.

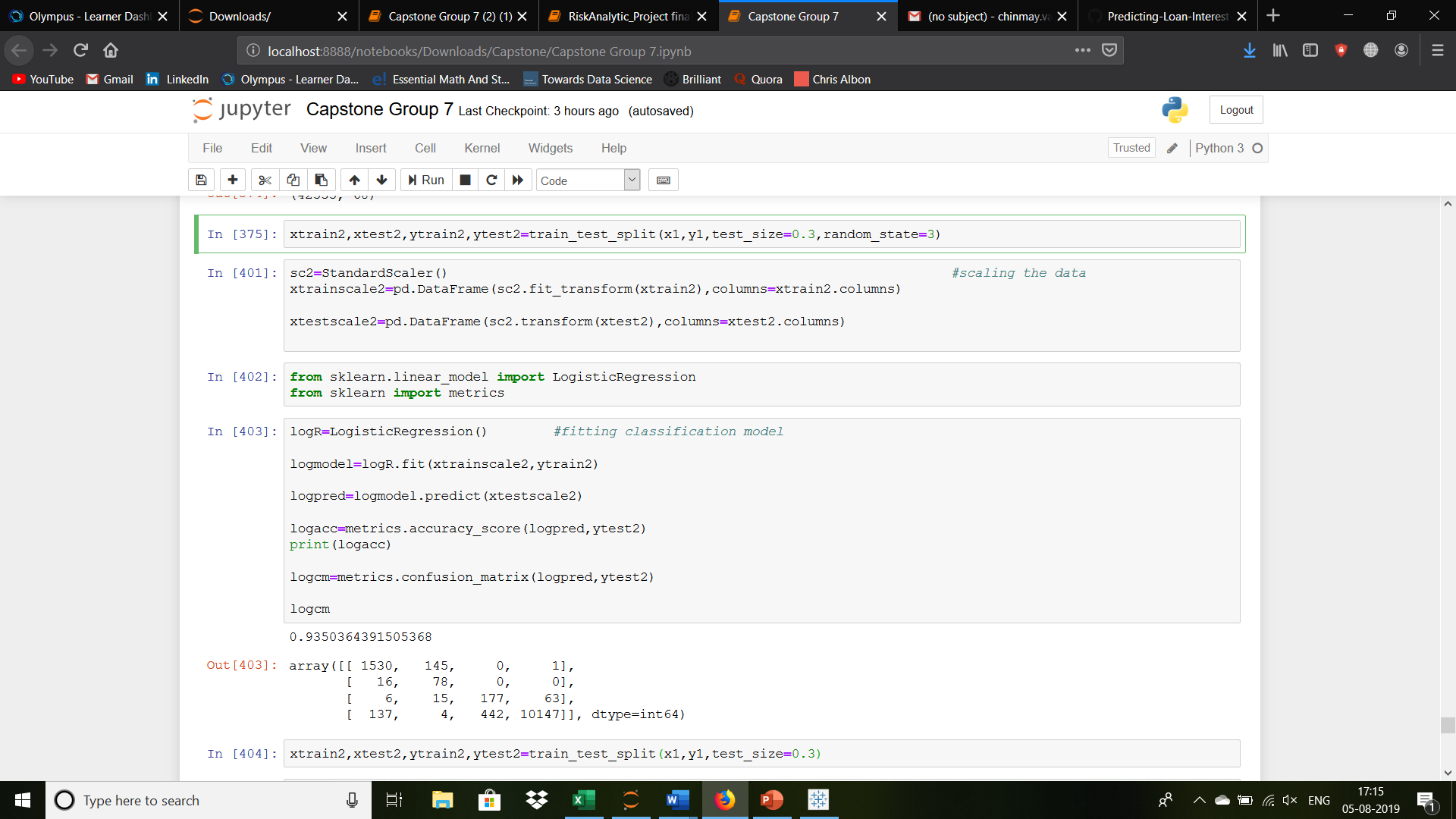


**Classification- Predicting the probability of default for a potential loan.**

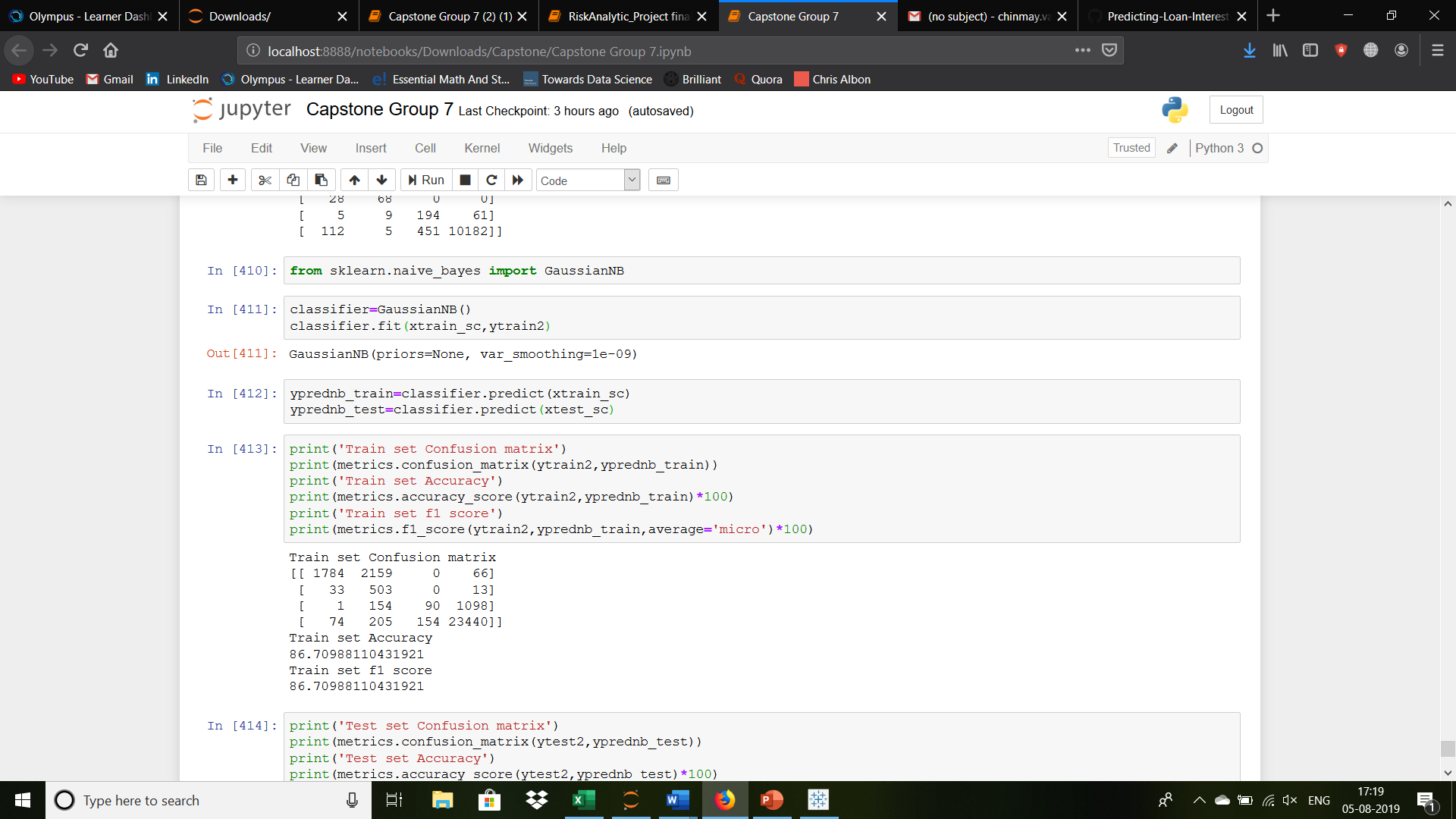
For classification the target variable is the ‘Loan Status’.

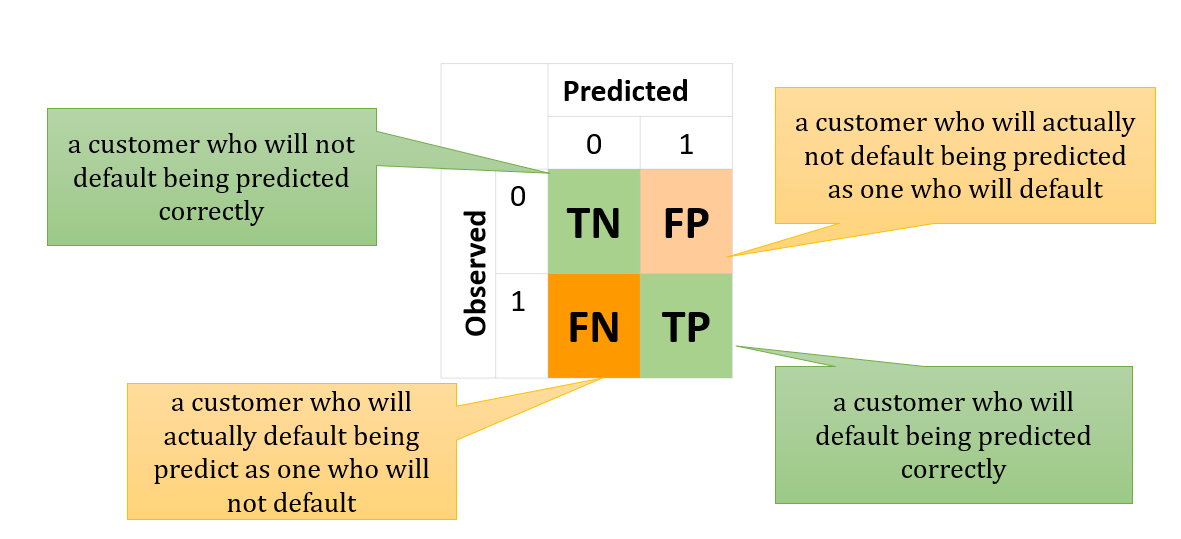
We take the same variables that we took earlier for predicting the interest rates.

* Again, we split the data into new train and test variables:



* Here, we fit a logistic regression as our baseline model.
* We also trained a Naïve Bayes model.



[](https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=2ahUKEwiplq7qj-zjAhXClOYKHaRUCz4QjRx6BAgBEAU&url=https%3A%2F%2Ftowardsdatascience.com%2Fmodel-performance-cost-functions-for-classification-models-a7b1b00ba60&psig=AOvVaw12DUJSbM9OK7MJqX3TSSzw&ust=1565108005789794)

* These models were applied to get an idea of how they perform on this data.
* Which will later be refined by applying Ensemble techniques such as bagging and boosting.
* Although here, the models perform well, but these scores are not reliable as the scores may differ for different sets of train and test.

# Conclusion

## **How investors can benefit**

* Investing in the consumer credit asset class through Lending Club provides a portfolio diversification opportunity and the potential to earn competitive returns.
* 99% of portfolios with 100+ Borrowers see positive returns. Investors may avoid the ups and downs associated with the stock market because Lending Club Notes have low correlation to the stock market.
* 12 Monthly payments of principal and interest: Investors receive principal and interest monthly as borrowers make payments on their loans.

It is clear that from an investor’s point of view, [Lending Club](http://www.anrdoezrs.net/click-7130697-10913071-1443562058000) is a great new service that allows them to automate investments and earn steady returns. Keep in mind that when it comes to investing online, it is important to do due diligence – figure out which investments suit your profile and which ones you should stay away from.

It is good that investors can decide the percentage of bad rate that they are willing to accept in their portfolio, and based on that they can decide very easy the percentage of the new loans that they want to finance.

We were able to identify the important features and fit a regression as well as classification model to it, But due to inconsistency in data the results may not be reliable for which we plan apply Cross Validation to the data. Also, several classification models will be tried on it, to identify the defaulters from the performers. Tuning methods like GridSearchCV will also follow.