Arsadur Rahman

AIT 580-04

Fall 2018

Big Data Analytics Project – Final part

On

All Lending Club Loan Data

December 15, 2018

The Dataset

There are two datasets, I will be exploring to complete this Big Data Analytics Project. Titled "All lending club loan data – Accepted & rejected from 2007 through the 2018 current quarter has two separate datasets where as one the accepted_2007 to 2018 _Q2 and the other is rejected_2007 to 2018_Q2. The dataset contains information of clients and originated loan from the time 2007 - 2018_Q2. It contains loan amount, outstanding balance, interest rate etc.   The datasets originated from the peer to peer lending company call Lending Club and it in Business/Finance domain. Both Datasets are over 1.3 GB and 2.7 Gb combined.

The accepted data set has more than 2 million observations and 151 variables and the rejected dataset has more than 22 Million observations and 9 variables.

The datasets variables are represented in these four major Data types

Data Type	Variables
Nominal	Id, member id, employee title, home ownership, description, purpose, tittle, zip code, state, policy code, loan tittle, policy code etc.
Ordinal	Grade, sub grade, loan status, etc.
Interval	Interest rate etc.
Ratio	Loan amount, funded amount, funded amount inventory, term, installment, empl oyee length, amount requested, debt to income ratio, employment length, etc

Data Type major observations: 1) Accepted 2007 to 2018 Q2 (Dataset 1)

```
Console  Terminal
C:/Users/rahma/OneDrive/Desktop/
> glimpse(accepted_2007_to_2017)
Observations: 2,004,091
Variables: 151
$ id                            <int> 38098114, 36805548, 37842129...
$ member_id                     <chr> NA, NA, NA, NA, NA, NA, NA, ...
$ loan_amnt                     <dbl> 15000, 10400, 21425, 12800, ...
$ funded_amnt                   <dbl> 15000, 10400, 21425, 12800, ...
$ funded_amnt_inv               <dbl> 15000, 10400, 21425, 12800, ...
$ term                          <chr> "60 months", "36 months", "6...
$ int_rate                      <dbl> 12.39, 6.99, 15.59, 17.14, 1...
$ installment                   <dbl> 336.64, 321.08, 516.36, 319....
$ grade                         <chr> "C", "A", "D", "D", "C", "C"...
$ sub_grade                     <chr> "C1", "A3", "D1", "D4", "C3"...
$ emp_title                     <chr> "MANAGEMENT", "Truck Driver ...
$ emp_length                    <chr> "10+ years", "8 years", "6 y...
$ home_ownership                <chr> "RENT", "MORTGAGE", "RENT", ...
$ annual_inc                    <dbl> 78000, 58000, 63800, 125000,...
$ verification_status           <chr> "Source Verified", "Not Veri...
$ issue_d                       <chr> "Dec-2014", "Dec-2014", "Dec...
$ loan_status                   <chr> "Fully Paid", "Charged Off",...
$ pymnt_plan                    <chr> "n", "n", "n", "n", "n", "n"...
$ url                           <chr> "https://lendingclub.com/bro...
$ desc                          <chr> NA, NA, NA, NA, NA, NA, NA, ...
$ purpose                       <chr> "debt_consolidation", "credi...
$ title                         <chr> "Debt consolidation", "Credi...
$ zip_code                      <chr> "235xx", "937xx", "658xx", "...
$ addr_state                    <chr> "VA", "CA", "MO", "CA", "AZ"...
$ dti                           <dbl> 12.03, 14.92, 18.49, 8.31, 3...
$ delinq_2yrs                   <dbl> 0, 0, 0, 1, 0, 0, 0, 1, 1, 0...
$ earliest_cr_line              <chr> "Aug-1994", "Sep-1989", "Aug...
$ fico_range_low                <dbl> 750, 710, 685, 665, 685, 680...
$ fico_range_high               <dbl> 754, 714, 689, 669, 689, 684...
$ inq_last_6mths                <dbl> 0, 2, 0, 0, 1, 0, 1, 0, 0, 0...
$ mths_since_last_delinq        <dbl> NA, 42, 60, 17, NA, NA, 55, ...
$ mths_since_last_record        <dbl> NA, NA, NA, NA, NA, NA, NA, ...
$ open_acc                      <dbl> 6, 17, 10, 8, 11, 12, 9, 7, ...
$ pub_rec                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ revol_bal                     <dbl> 138008, 6133, 16374, 5753, 1...
$ revol_util                    <dbl> 29.0, 31.6, 76.2, 100.9, 91....
```

Data Source Title in Kaggle: All Lending Club Loan Data – 2007 through current lending club accepted and rejected loan data (All lending Club, 2018).

Data Type major observations: 1) Rejected 2007 to 2018 Q2 (Dataset 2)

```
> glimpse(rejected_2007_to_2017)
Observations: 22,469,074
Variables: 9
$ `Amount Requested`     <dbl> 4000, 20000, 1000, 15000, 8500, 5000, 20000, 500...
$ `Application Date`     <date> 2016-07-01, 2016-07-01, 2016-07-01, 2016-07-01,...
$ `Loan Title`           <chr> "major_purchase", "debt_consolidation", "renewab...
$ Risk_Score             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ `Debt-To-Income Ratio` <chr> "4.21%", "0.39%", "42.38%", "5.98%", "29.14%", "...
$ `Zip Code`             <chr> "750xx", "930xx", "923xx", "910xx", "210xx", "28...
$ State                  <chr> "TX", "CA", "CA", "CA", "MD", "NC", "CA", "KY", ...
$ `Employment Length`    <chr> "5 years", "< 1 year", "5 years", "5 years", "< ...
$ `Policy Code`          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
>
```

Data Source Title in Kaggle: All Lending Club Loan Data – 2007 through current lending club accepted and rejected loan data (All lending Club, 2018).

The Source of the Data (Organizations):

I downloaded the datasets from Kaggle website, but Kaggle sourced the Data from Lending Club Database. Kaggle is a platform to find and publish data for anyone who is interested to work with different kinds of data. It is a hub for world's data scientist to come together and solve big data and machine learning problems. Kaggle has more than 540,000 active users from 194 countries and receives more than 150,000 submissions per months (What is Kaggle, 2018). Basically, Kaggle is a platform where complex dataset problems get poured and talented data scientists take the challenge, solve the problems and share their work with worlds in the form of findings and code snippets. On the other hand, the original owner of the datasets is Lending Club a peer to peer lending company based in San Francisco, California. Lending club was founded in 2006 and since then the company have originated more than $15 of loans (Personal Loans, 2018). The Data was accumulated over the years from their customer transactions, each borrower ID has created multiple types of data to work with. Lending Club aims to offer unsecured loans from $1000 - $40000 to its eligible customers throughout the US (AvantCredit, 2018).

Reasons the Data was collected:

The data was collected to understand the borrower's behavior and mostly to assess the risk of lending. It is almost required to keep track the information on an originated loan to determine the loan payment, aggregated interest, determining borrower's eligibility, understanding company's ability to lend etc. Collecting the data make it easy to track all the information necessary to navigate a loan and determine the future risk of lending to a future potential client. Initially the data was small, but with the time and increasing amount of borrower the data kept pilling up. Every loan ID generates enormous amount of valuable data by providing personal information and payment information. The borrowers consented in collecting and restoring the data to the All lending club.

Major Questions that can be answered analyzing these two datasets:

- How the loan amount is distributed among its clients?
- Does Interest rate of the loan vary for the different grade of loan?
- How the interest rate fluctuated over the time to keep up with demand and supply of loan?
- Which state as most charged off /bad loan history? And wants the percentage of charged off loan and paid off loan?
- How debt to income ratio is an important factor in deciding a good borrower?
- What is the average annual income of the different kind of borrowers and what Is the minimum to maximum range of FICO score?

Software used to explore the Datasets

- R/ R Studio version 3.5.1 for exploration and clean the Data sets
- Python analysis
- Tableau to visualize the datasets
- SQL to create a Database

Hardware of the computer the data set was processed

The both datasets were performed in a laptop with the following requirements:

Item name	Description
Processor	Intel(R) Core(TM) i7-8750H CPU
Processor Speed	2.20GHz
Number of logical processors	12
L2 Cache (per score)	1.5MB
L3 Cache	9.0MB
RAM	16 GB
System	64-bit OS
Cores	6

Problems with The Dataset

➢ Privacy

- This Datasets are accessible to anyone with internet connection, anyone can download the datasets, however the two datasets do not include any personal sensitive (customer) information such as Name, Address. Social Security Number, Individual income etc. which are very important to a person's identity. The data provides the information which will help to do research and understand the pattern of customers based on how each client handle the loan from its origination to pay off.

➢ Quality

- From the initial exploration I have found a good amount of missing data in the first data set which I will be elaborating in the data preparation and analysis section below. Almost 58 variables will be dropped out of 151variables from the data set because those variables have missing data, and this makes the dataset incomplete, so this is a quality issue with the data set.
- Assuming the dataset was created through a computer-generated database which can contain wrong information and inaccurate information, which may bring irrelevant results at the end.
- The features have a very short-cut names which makes it hard for the reader to understand which column what is and may lead to a misleading assumption.

Loan Data Exploration (accepted 2007 – 2018 Q2)

To import the data on python using pandas it will be easier to clean and analyse the data, the snippet below shows the importing process.

```
sns.set(style='whitegrid')
```

```
In [14]: loans = pd.read_csv('C:/Users/rahma/accepted_2007_to_2018Q2.csv.gz', compression='gzip', low_memory=True)

C:\ProgramData\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2785: DtypeWarning: Columns (0,19,49,59,118,129,13
0,131,134,135,136,139,145,146,147) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

```
In [15]: loans.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2004091 entries, 0 to 2004090
Columns: 151 entries, id to settlement_term
dtypes: float64(113), object(38)
memory usage: 2.3+ GB
```

Since we have a large data set with numerous features and variables, my first step is to clean the data as much as I can to prepare for the visualization. In the cleaning process first, I would remove the missing and unnecessary features and variables from the data set.

jupyter Big Data Loan Last Checkpoint: 2 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

```
In [22]: loans['loan_status'].value_counts(normalize=True, dropna=False)

Out[22]: Fully Paid     0.797588
         Charged Off    0.202412
         Name: loan_status, dtype: float64
```

```
In [24]: missing_data = loans.isnull().sum().sort_values(ascending= False)
         drop_columns = list(missing_data[missing_data > loans.shape[0] *0.1].index)
         loans = loans.drop(drop_columns, axis = 1)

         missing_data

Out[24]: member_id                                   843934
         next_pymnt_d                                843934
         orig_projected_additional_accrued_interest  843832
         hardship_start_date                         843341
         hardship_length                             843341
         hardship_type                               843341
         hardship_reason                             843341
         hardship_status                             843341
         deferral_term                               843341
         hardship_amount                             843341
         hardship_end_date                           843341
         payment_plan_start_date                     843341
         hardship_dpd                                843341
         hardship_loan_status                        843341
         hardship_payoff_balance_amount              843341
         hardship_last_payment_amount                843341
         sec_app_mths_since_last_major_derog         842464
         sec_app_revol_util                          839720
         sec_app_chargeoff_within_12_mths            839641
         sec_app_fico_range_low                      839641
         sec_app_open_acc                            839641
         sec_app_mort_acc                            839641
         sec_app_inq_last_6mths                      839641
         sec_app_num_rev_accts                       839641
         sec_app_open_act_il                         839641
         sec_app_collections_12_mths_ex_med          839641
```

then narrow down the variables to use for visualization. In the below snippet we can see how many missing values we have for each variable in the data set

From the snippet it is visible that we have missing data problem, in the next step, the top missing values would be dropped from the dataset to make it ready for visualization.

The below snippet is the primary descriptive statistics of loan amount disbursed from 2007 to 2018 Q2.

```
In [65]: loans['loan_amnt'].describe()

Out[65]: count    843934.000000
         mean      14322.610506
         std        8560.537426
         min         500.000000
         25%        8000.000000
         50%       12000.000000
         75%       20000.000000
         max       40000.000000
         Name: loan_amnt, dtype: float64
```

```
loans['loan_status'].value_counts(dropna=False)

Current                                                 1108697
Fully Paid                                               673112
Charged Off                                              170822
Late (31-120 days)                                        27678
In Grace Period                                           13775
Late (16-30 days)                                          7157
Does not meet the credit policy. Status:Fully Paid         1988
Does not meet the credit policy. Status:Charged Off         761
Default                                                      72
NaN                                                          29
Name: loan_status, dtype: int64
```

Visualization

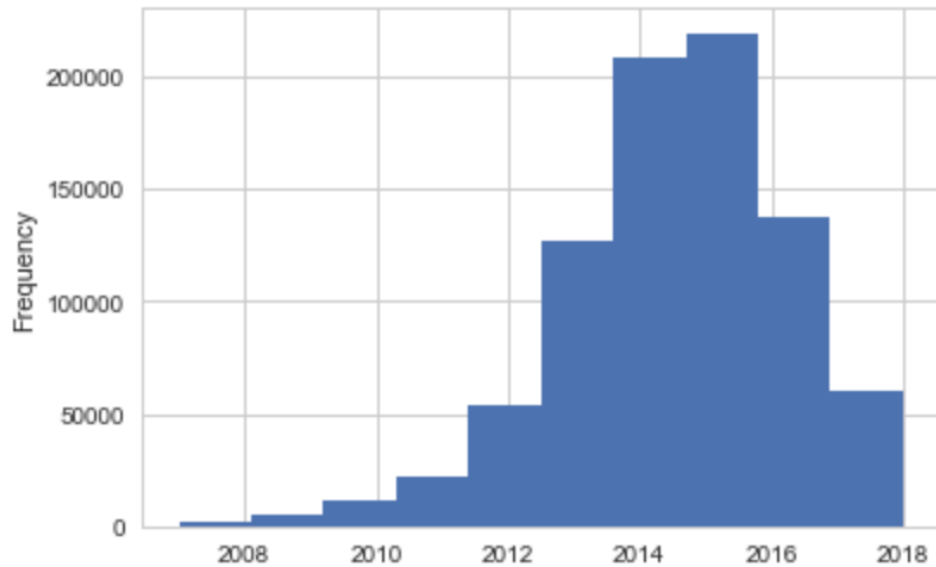

Figure 1: Number of loan disbursement from the year 2007 – 2018

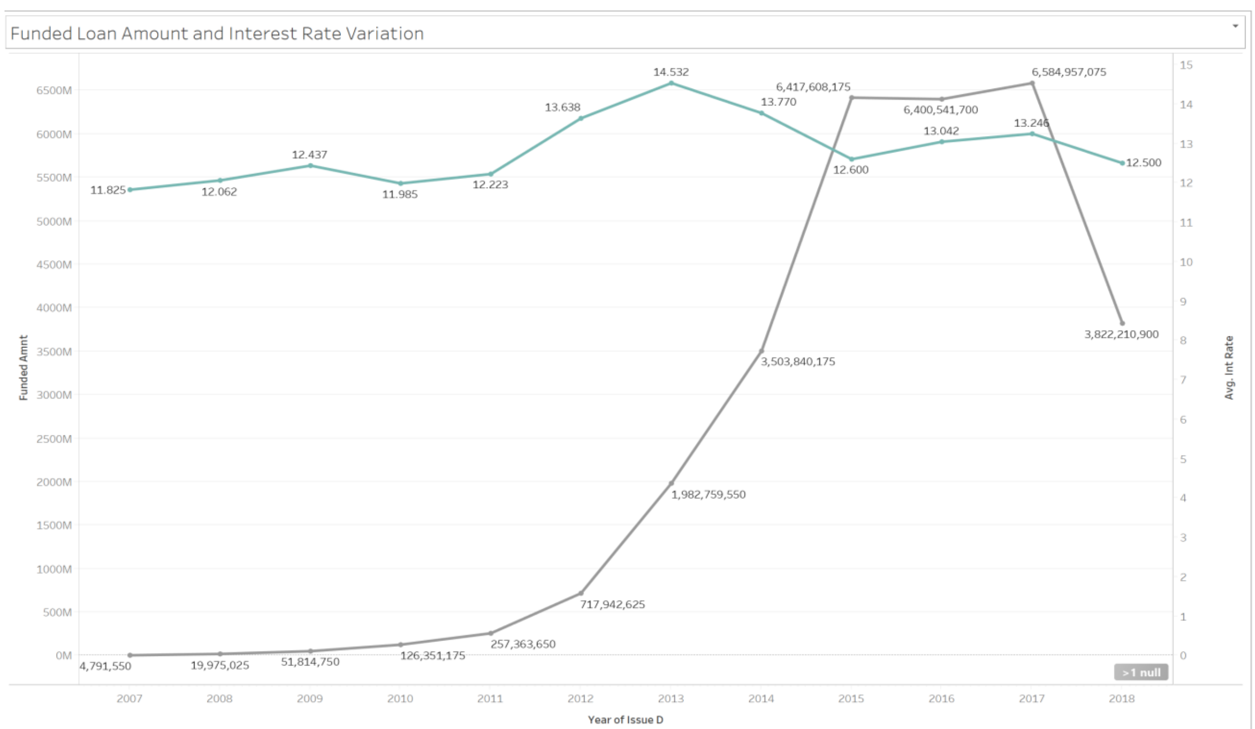

Figure2: Interest rate and funded loan amount variation

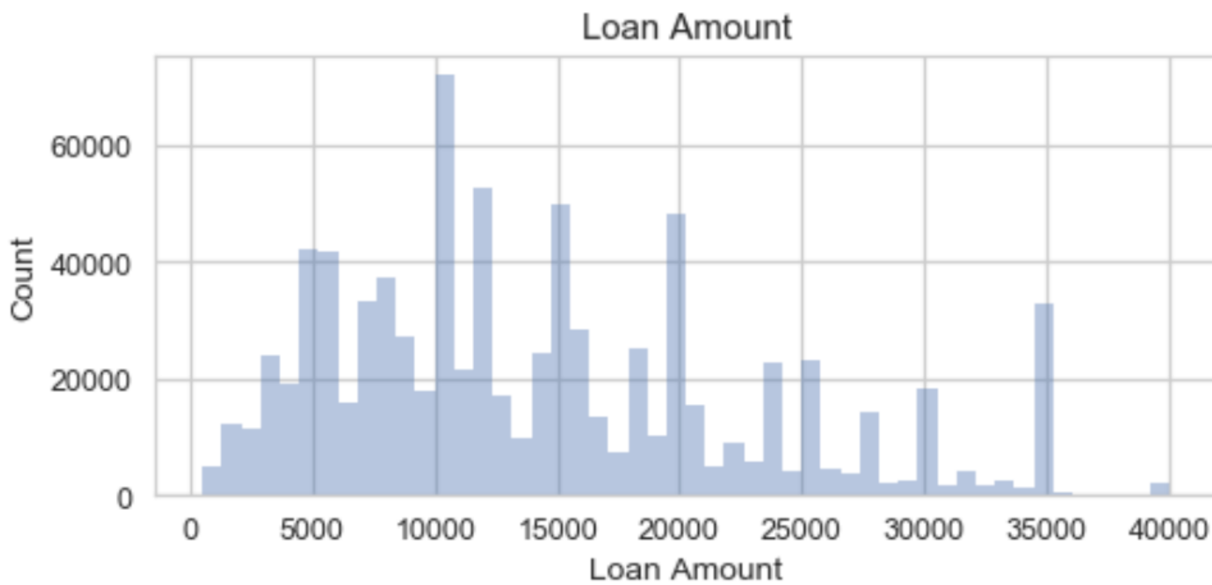

Figure 3: Loan amount range distribution

Figure 4: Loan Status in the Data Set

Figure 5: Range of Debt to income Ratio in receiving loan

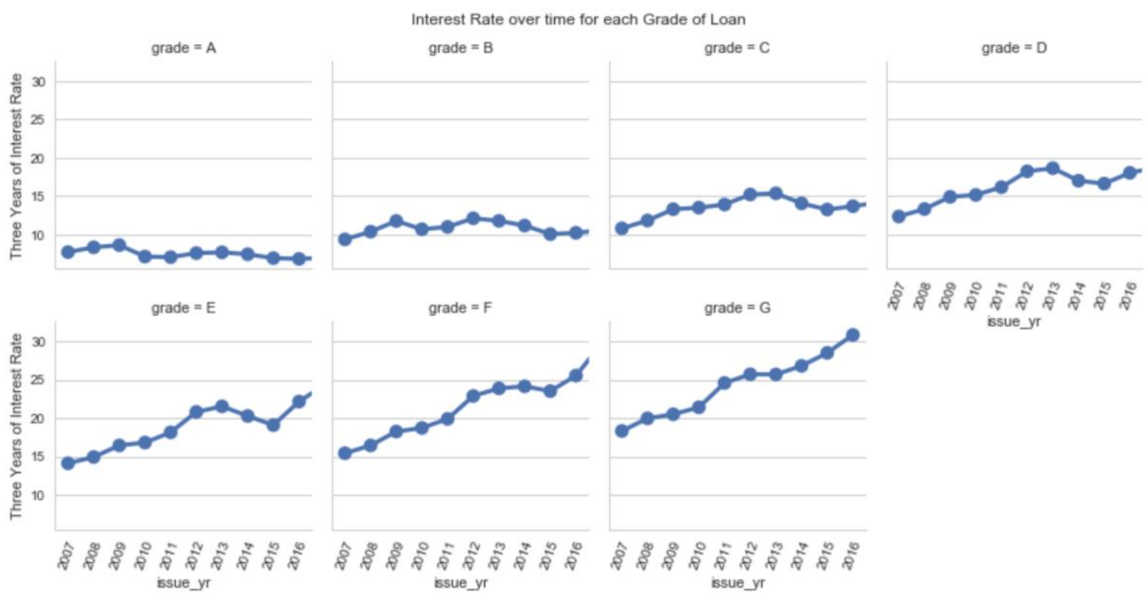

Figure 6: Interest rate time line for each grade

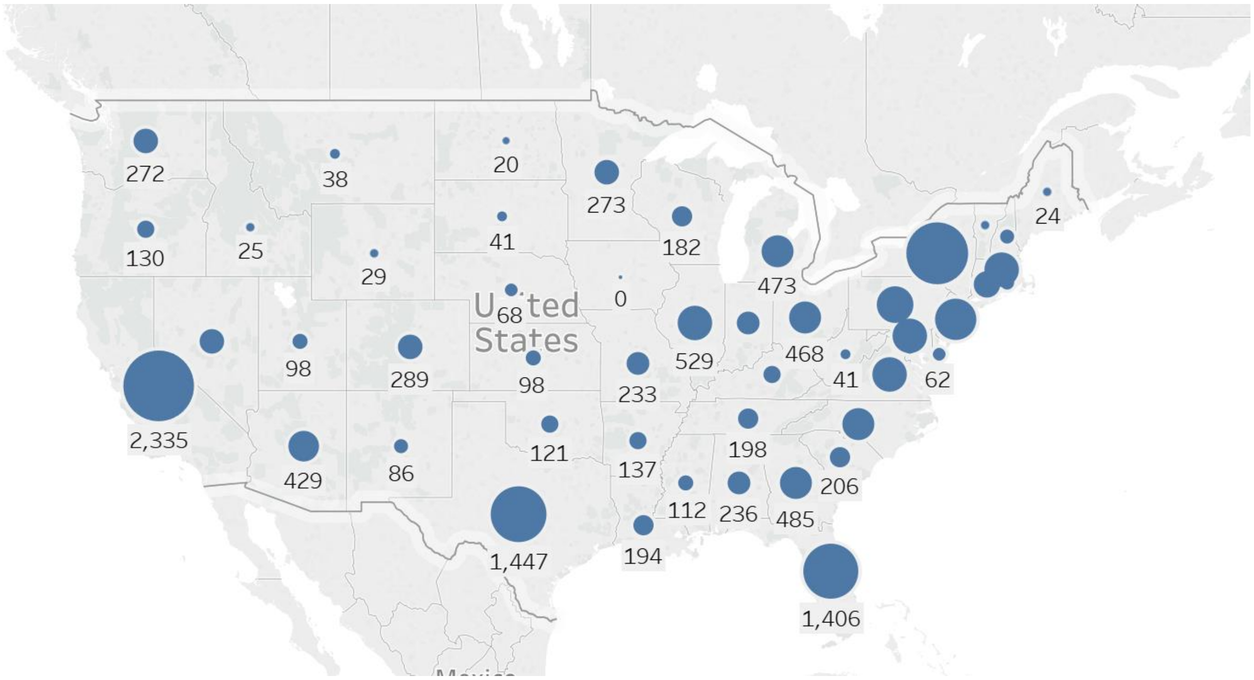

Figure 7: Number of loan charged off within 12 months in each state

Figure 8: Average Income of different home owners and the lowest rang Avg FICO score

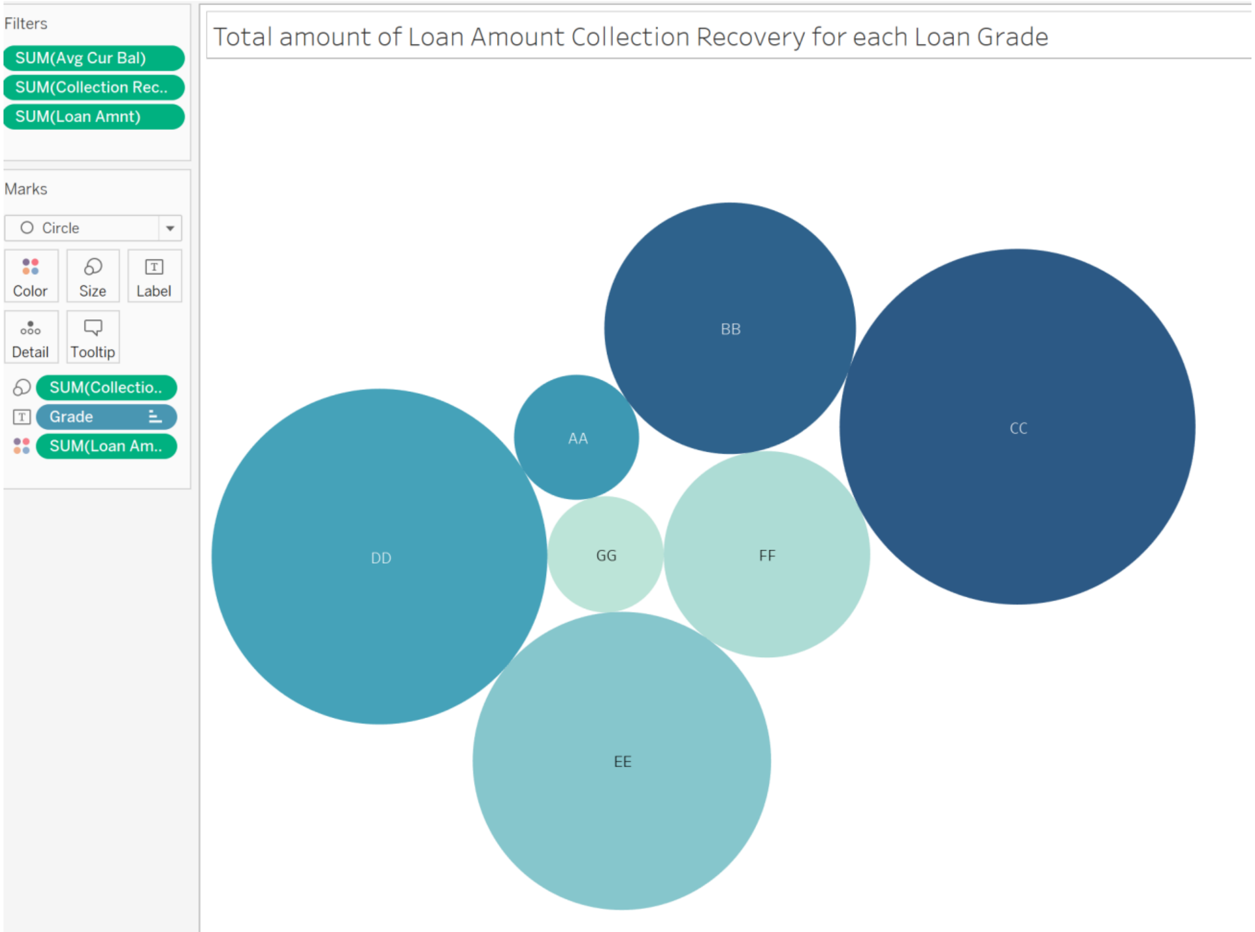

Figure 9: Done In tableau when keep the cursor, we can see the amount of loan each grade have generated and amount of fees each grade collected.

Meta data Definition:

Loan ID: Nominal data type. Number data, the identifier has eight digits, each entry has a unique number and mostly signifies it the most valuable information and links to all other attributes. Example:1077501, 1077430, 1077175, 1076863.

Loan Amount: Ratio data type, the value in each loan amount describes how much loan was disbursed and connected to loan ID. The value has true zero Example:2500, 2400, 10000, 3000, 40000.

Funded Amount: Ratio data type, has true zero, most values are same as loan amount, this is the amount was originally given to the borrower. Example:5000, 2500.

Term: Ordinal data type and has two levels one is 36 months loan" and the other one is 60 months loan term.

Interest rate: Interval: interest rate of originated loan to borrowers, example: 0.7, 15.3, 16, 13.5, 12.7. This data is in percentage. Interest is charged to the borrowers with the principal amount.

Installment: Ratio type. This is will be how many payments a borrower would make to pay off the loan. Example: 162 59 84 339 67

Grade: Ordinal data, has 7 levels, these are the loan grade based on loan type, and each loan grade has different criteria. Example: A, B, C, D

Sub Grade: Ordinal type, includes 35 levels, the sub groups are distributed over loan grade. For A it has A1, A2, A3 for B it has B1, B2, B3 and so on.

Employee Title: Nominal data type, borrower's employment title. Example: Driver, Consultant, teacher, It Consultant etc.

Employee length: Ordinal data. Number years a borrower employed in career. The range is from 1 year to 12 years in the datasheet.

Home ownership: Nominal data type with 6 levels of identifier. Example, any", Mortgage, Rent, Owner etc.

Annual Income: Ratio data type. Yearly income of the borrower. Such as: 24000 30000 12252 49200 80000 etc.
Verification Status: Ordinal data type, defines whether borrowers provided information is Verified or not or in which status it belongs.

Issue date: Data type data, the month and year the loan was issued. Example: Apr-2008, Apr-2009.

Loan Status: Ordinal Data, in ten levels, this data defines the present status of the loan. Since the datasheet stretches from 2007 to 2017 some of the loan is paid off or still on the way to be paid off.

Description: Nominal data type, overall string description of the data. What the loan was used for or a general client note.

Loan Title: Nominal data, describes what kind of loan is, this is line of product based on the customers demand. Example: Credit Card, personal loan, auto loan, mortgage Etc.
Zip Code: Nominal, in which zip code of a state the loan was disbursed.

Results and Graphics Interpretation

In the first descriptive statistics analysis, we can see the average amount of loan originated was $14322, maximum loan was $40000 and the minimum loan was $500 also 50% of the loan was below $12000 range. This tells a lot about the company, it is more of small loan generator and targets to the clients who needs small loan need.

From graphs 1 , the years of 2014 and 2016 the All lending club have generated more loan than any other years in company's lifetime, and there is an upward trend in loan disbarment.

In Graph 2, we can observe the very flat interest rate that all lending club was charging to its customers, which is from 11% to 12% and there was no big change in interest rate.

In graph 3, using python I visualized the descriptive statistics which I described in the beginning. Seeing what the range of loan amount are.

In graph 4: The box plot shows what percentile of the loan is fully paid and charged off, I found there was almost 70% of the loan was fully paid and 30% of the loan was charged off. The table below will show how the paid off loan and charged off loans are compared to their status

	count	mean	std	min	25%	50%	75%	max
loan_status								
Charged Off	168084.0	15384.975964	8606.476332	900.0	9000.0	14000.0	20000.0	40000.0
Fully Paid	646902.0	14037.566316	8449.740674	500.0	7500.0	12000.0	20000.0	40000.0

In graph 5: The debt to income ratio is very much centered towards 10-20, a lot of customer have 1 10-20 dti against their loan, this tells us the company does not give out loan an applicant who has higher dti.

In graph 6: The Grade A category has the lowest interest rate which is around 10% on the other hand grade G category loans has the highest interest rate which hovers from 10% to 30%, it seems like the category D, E, F, and G brings highest amount of interest revenue for the company. However, these types of loan category hold high interest.

In graph 7: Which state has highest number of charged off loan within the 12 months of loan origination. State of California, Texas, Florida, New York are the among the top five in the list. Which means these states had more defaulter than any other states within 12 months of loan origination.

In graph 8: This graph shows how the average annual income differs with the homeowner status, customers who has mortgage has higher annual income better FICO score than the customers who rent and answered other in their application. These are two most important factors to determine to give out loan.

And the last graph, graph 9 shows which category offers the highest recovery fees for the company. Category D, C, and E earn more collection recovery fees than the other categories. So these categories are profitable for the company.

The questions that were answered through the big data project are below:

- How the loan was distributed?

The median loan amount in $12000, most of the loans are small loans, the standard deviation s high which means the amount of loan disbursed was spread over in various range. Whoever applies for a medium range of loan can consider the lending company to apply for a loan.

- Does Interest rate of the loan vary for the different grade?

Yes, interest rate is different for the different grade of loan, the G category has the highest interest rate over the years, and the category A has the lowest interest rate over the years.

- How the interest rate fluctuated over the time to keep up with demand and supply of loan?

The interest rate remains steady from 2007 to 2018, however the when demand for the loan was all-time high the interest rate was below the average. This shows that there is correlation, and effect of loan amount to the interest rate.

- Which state as most charged off /bad loan history? And wants the percentage of charged off loan and paid off loan?

The state of California has most default loan, followed by the state of Texas. The number of default loan in the state of California was 2335 making it as highest risky state to offer loan. 70 % loan was paid off and nearly 30% was default.

- How debt to income ratio is an important factor in deciding a good borrower?

The debt to income ratio is very important to decide who is going to receive the loan and why is not. The lower the ratio the higher the chance to be approved for the loan. A high debt to income ratio is a risky applicant to be considered. From the graph it is learned the if an applicant has 15% DTI he or she more likely to be approved for the loan.

- What is the average annual income of the different kind of borrowers and what Is the minimum to maximum range of FICO score?

The average annual income of the borrowers with a mortgage and home is $87000 having the highest income, on the other hand borrowers who answered none in their homeownership earn $62000 and have an average FICO score of 698. If the FICO score is around 700 an applicant is most likely to get approved for a loan.

- What are most predictors of charge off loans?

Using python and determining correlation and F statistics, I found out 10 most predictors of of a charge off loan. Using the sklearn feature on python and training two data set below the table shows the top predictors of a future default loan.

	variable	pearson_corr	F	p_value
0	int_rate	0.256691	51570.855220	0.0
1	term	0.188604	26965.820737	0.0
2	fico_score	-0.139928	14600.790875	0.0
3	dti	0.131005	12766.403302	0.0
4	mort_acc	-0.076676	4044.121007	0.0
5	log_annual_inc	-0.074252	4053.225270	0.0
6	sub_grade_A4	-0.069089	3506.566794	0.0
7	revol_util	0.066569	3252.333239	0.0
8	sub_grade_A5	-0.066226	3220.711982	0.0
9	loan_amnt	0.063598	2969.118688	0.0
10	home_ownership_RENT	0.063576	2967.107649	0.0

The most correlated items on the list are interest rate, term of the loan, FICO credit score, Debt to income ratio, Mortgage account, annual income of the borrower etc. The highly positive correlated and highly negative correlated are the part of predictors, and top ten times has the highest F score. These times can help decide which loan will be a default in the future.

SQL schema code:
#Using MySQl version 8.0 the dataase or the collected was created.

```
CREATE SCHEMA `lendin_club_loan_database` ;


CREATE TABLE `lendin_club_loan_database`.`customer` (
  `loan_id` INT(20) NOT NULL,
  `member_id` INT(25) NULL,
  `cust_address` VARCHAR(45) NULL,
  `cust_phone` INT(20) NULL,
  PRIMARY KEY (`loan_id`));



CREATE TABLE `lendin_club_loan_database`.`loan_info` (
  `loan_amt` INT NOT NULL,
  `loan_terrm` VARCHAR(15) NULL,
  `int_rate` INT(10) NULL,
  `grade` VARCHAR(45) NOT NULL,
  `subgrade` VARCHAR(45) NULL,
  PRIMARY KEY (`grade`));


CREATE TABLE `lendin_club_loan_database`.`eligibility` (
  `cust_income` INT NOT NULL,
  `cust_emp_title` VARCHAR(45) NULL,
  `fic_score` INT(10) NOT NULL,
  `home_ownership` VARCHAR(45) NULL,
  `purpose` VARCHAR(45) NULL,
  `total_acc` INT(10) NULL,
  PRIMARY KEY (`fic_score`));


CREATE TABLE `lendin_club_loan_database`.`pays` (
  `payment_amt` INT NOT NULL,
  `loan_status` VARCHAR(45) NULL,
  `fees` INT(20) NULL,
  `total_paid` INT(20) NULL,
  `total_due` INT(20) NULL,
  PRIMARY KEY (`payment_amt`));
```

```sql
CREATE TABLE `lendin_club_loan_database`.`payment_plan` (
  `plan_id` INT NOT NULL,
  `loan_type` VARCHAR(45) NULL,
  `plan_type` VARCHAR(45) NULL,
  `installment_amt` INT(20) NULL,
  `application_type` VARCHAR(45) NULL,
  PRIMARY KEY (`plan_id`),
  INDEX `loan_type_idx` (`loan_type` ASC) VISIBLE,
  CONSTRAINT `loan_type`
    FOREIGN KEY (`loan_type`)
    REFERENCES `lendin_club_loan_database`.`loan_info` (`grade`)
    ON DELETE RESTRICT
    ON UPDATE RESTRICT);


CREATE TABLE `lendin_club_loan_database`.`settles` (
  `settlment_id` INT NOT NULL,
  `settlment_amt` INT(20) NULL,
  `settlement_date` DATE NULL,
  `settlment_percentage` INT(10) NULL,
  `settlement_term` VARCHAR(45) NULL,
  PRIMARY KEY (`settlment_id`));


ALTER TABLE `lendin_club_loan_database`.`customer`
ADD INDEX `member_id_idx` (`member_id` ASC) VISIBLE;
;
ALTER TABLE `lendin_club_loan_database`.`customer`
ADD CONSTRAINT `member_id`
  FOREIGN KEY (`member_id`)
  REFERENCES `lendin_club_loan_database`.`customer` (`loan_id`)
  ON DELETE RESTRICT
  ON UPDATE RESTRICT;
ALTER TABLE `lendin_club_loan_database`.`loan_info`
ADD INDEX `subgrade_idx` (`subgrade` ASC) VISIBLE;
;
ALTER TABLE `lendin_club_loan_database`.`loan_info`
ADD CONSTRAINT `subgrade`
  FOREIGN KEY (`subgrade`)
  REFERENCES `lendin_club_loan_database`.`customer` (`cust_address`)
  ON DELETE RESTRICT
  ON UPDATE RESTRICT;
```

Bibliography

All Lending Club loan data. (2018). Kaggle.com. Retrieved 19 November 2018, from https://www.kaggle.com/wordsforthewise/lending-club

Personal Loans Borrow up to $40, f., Institutions See the options for banks, a., Loans, P., Loans, B., Refinancing, A., & Solutions, P. (2018). How Peer Lending Works | LendingClub. Lendingclub.com. Retrieved 19 November 2018, from https://www.lendingclub.com/public/how-peer-lending-works.action

What is Kaggle, Why I Participate, What is the Impact?. (2018). Kaggle.com. Retrieved 10 November 2018, from https://www.kaggle.com/getting-started/44916

AvantCredit Raises $225 Million From Tiger Global, Peter Thiel. (2018). Business Insider. Retrieved 19 November 2018, from https://www.businessinsider.com/r-avantcredit-raises-225-million-from-tiger-global-peter-thiel-2014-12

Technical Terms:

Debt to Income Ratio (DTI): How much debt an applicant carries against his/her income. If the DTI is 20 that means that person has $20 of debt against $100 of his income, which is 20% of his/her overall income.

Loan Origination: When a loan is approved, and the loan is distributed to a customer.

FICO Score: FICO is credit core, usually it ranges from 500 to 850 700 is considered to be a very good credit score where 500/600 s bad credit score.

Interest rate: Cost of borrowing. When someone borrows money from a financial institution, he/she must pay back the borrowed amount with the interest. Usually the lender decides the rate of interest based on different factors.

Collection Recovery fees: This is a fee a lender collects when liabilities are unpaid for more than 90 days. It is a fee that brings revenue for any lender.

Fully paid: Fully paid loan is when a borrower pays of all principal and interest amount to the lender.

Charged off: Charged off are type of loan, which is default loans, when a borrower cannot pay his/her loan and declares bankrupt or deceased without paying then it is a charge off loan.

Mortgage: Type of loan, usually pays for buying a new home and pays over 15-20 years.