Advanced Programming Technique DATA 608

Muthukumar Srinivasan

Data Set:

(1) SF Salaries (Explore San Francisco City Employee Salaries)

(2) Lending Club Loan

Reference:

(1) https://www.kaggle.com/kaggle/sf-salaries

(2) https://www.kaggle.com/wendykan/lending-club-loan-data

(1) Scope of the Analysis

The data set that has been chosen is San Franciscan Salaries. I am hoping to identify the

a. Average base pay, mean , median and Mode in Base Pay
b. Average Base Pay, mean , median and Mode in Overtime, other pay and benefits
c. Average in Total Pay
d. Year Vs Pay analysis
e. Job title Vs Total Pay
f. Various graphs

(2) Data Cleaning

Data Cleaning may not be needed. I am planning to load this in CSV file and also do the same in Excel to provide both Excel and R Programming. Some of the rows which has Not Applicable in numeric column has been removed to achieve right and perfect results. We could write program to check each column for non-numeric value and remove them. But for now, we have done this manually.

(3) What Type of Visualization?

There are several ways the data visualization can be done. I have categorized as

1. Structured Programming way – Like Python, D3, Java Script, PHP , Programming R
2. Non Structured Data Driven Way – Like Excel, Talend, other Data visualization tool etc.

Structural Programming Way – R Programming

We can do lot of graphs and map using the R Programming. The output of what I have done loading the SF San Francisco Salary has been attached herewith.

Non Structured Data Driven Way

There are several tools available in the market by which the CSV file can be loaded and analyzed. One of them is Excel. Other famous tools are talend, jasper reports etc. I am unable to use these because these requires Server Infrastructure. Analyzing huge set of Data requires server infrastructure and good resources in the computer. So I have chosen Excel and have tried to put it together.

(4) Data Sets and Details

Data Set:

1. SF Salaries (Explore San Francisco City Employee Salaries)
2. Lending Club Loan

Reference:

1. https://www.kaggle.com/kaggle/sf-salaries
2. https://www.kaggle.com/wendykan/lending-club-loan-data

San Francisco Salaries

One way to understand how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014.

Loan Data:

These files contain complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

(4.1)

Data Set : SF Salaries

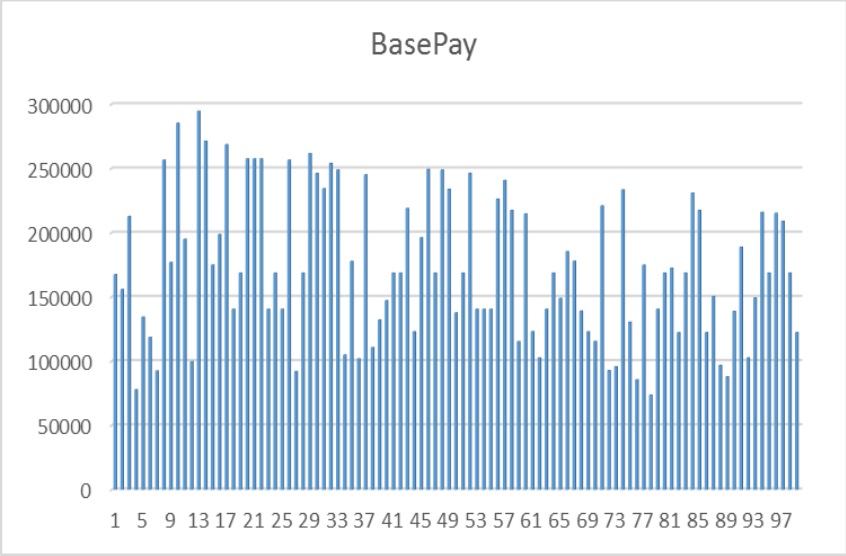
(Explore San Francisco City Employee Salaries)

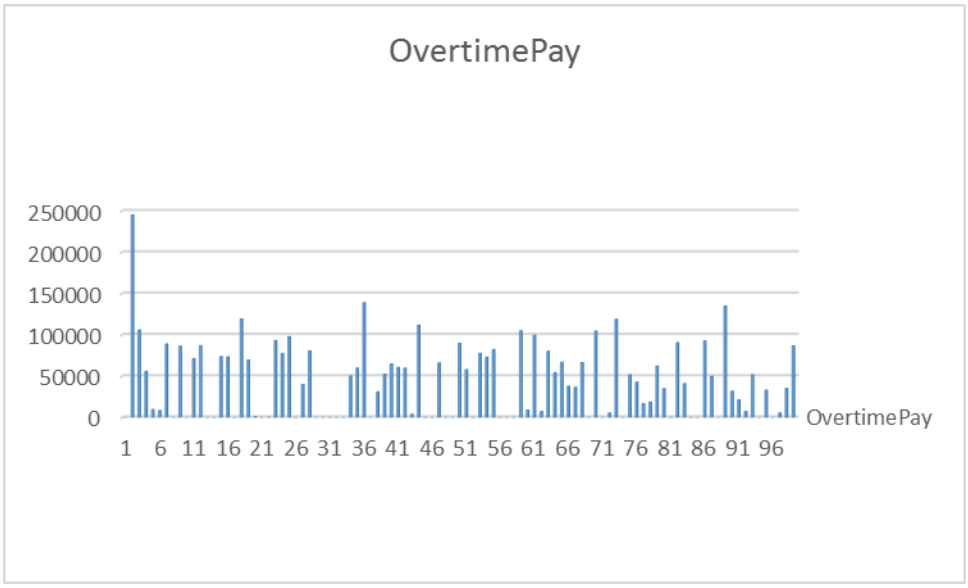
This is downloaded from Kaggle.com. It is about San Francisco city employee's annual base salary.

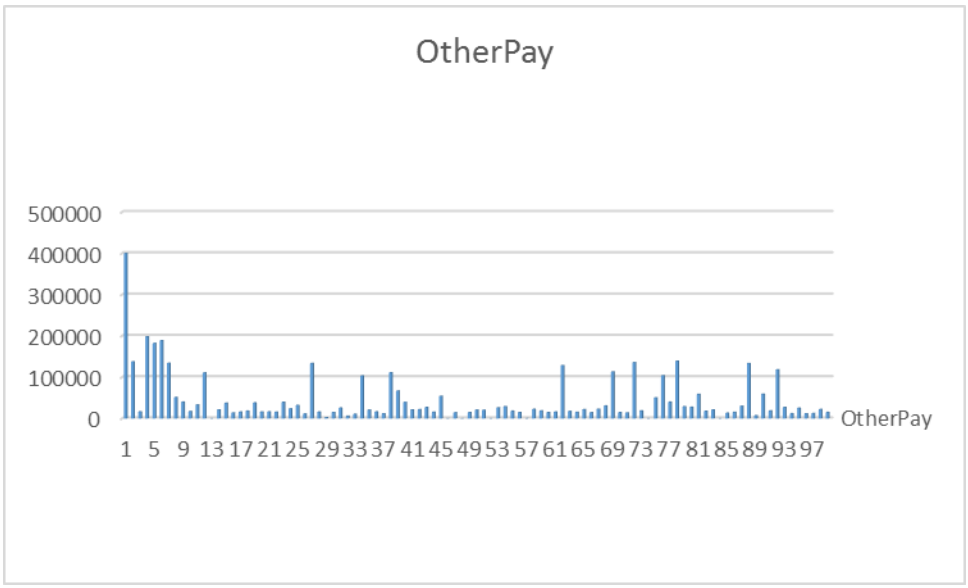
Zip File content attached with this write up

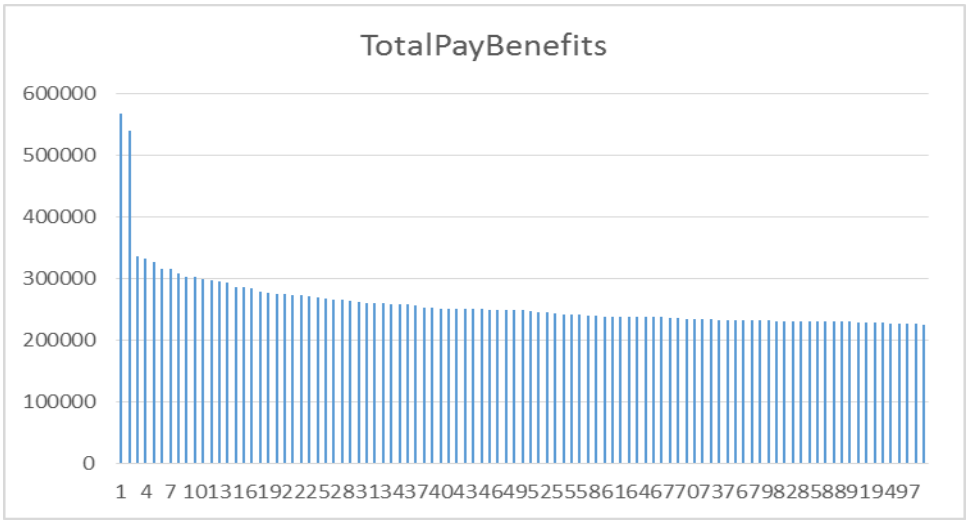
Salaries.csv	File download from kaggle.com
SFSalaries.xls	Loaded the csv file in excel and have developed some chart using excel
SFSalaries-Cleaned.csv	Removed few rows at the end which has empty values in the cell to cleanup
SFSalaries-sliced100Rows.csv	100 rows out of the main salaries.csv has been pulled to reduce the run time. 148547 rows is too high to process using my laptop. So reduced it to 100 rows and took sample.
SFSalaries-sliced100Rows.xls	Converted as XLS and created some charts too
Project-IS-608-MSrinivasan-D3JS	D3JS program to load the csv file
Project-IS-608-MSrinivasan-D3JSBarChart	For Bar Chart

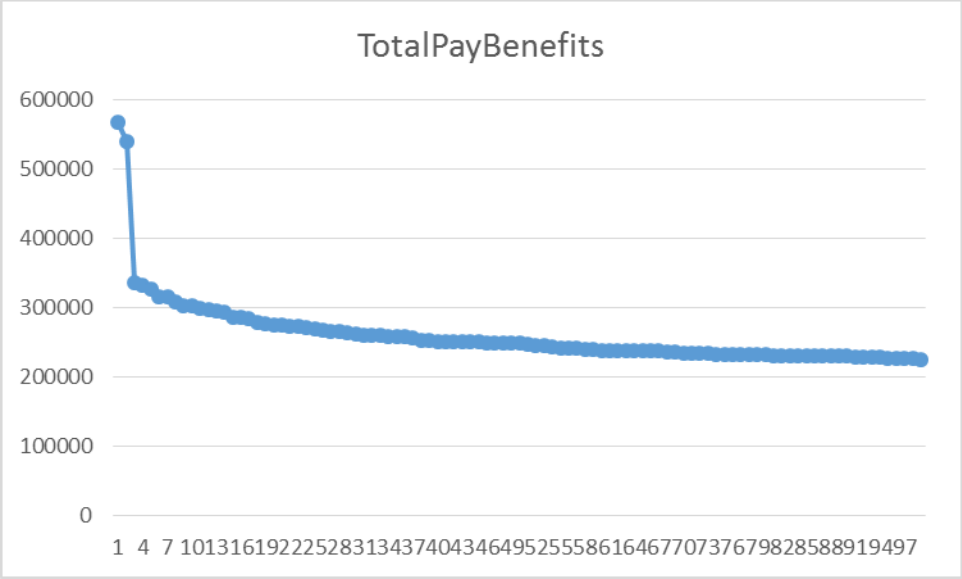
Sample 100 Rows Charts using Excel

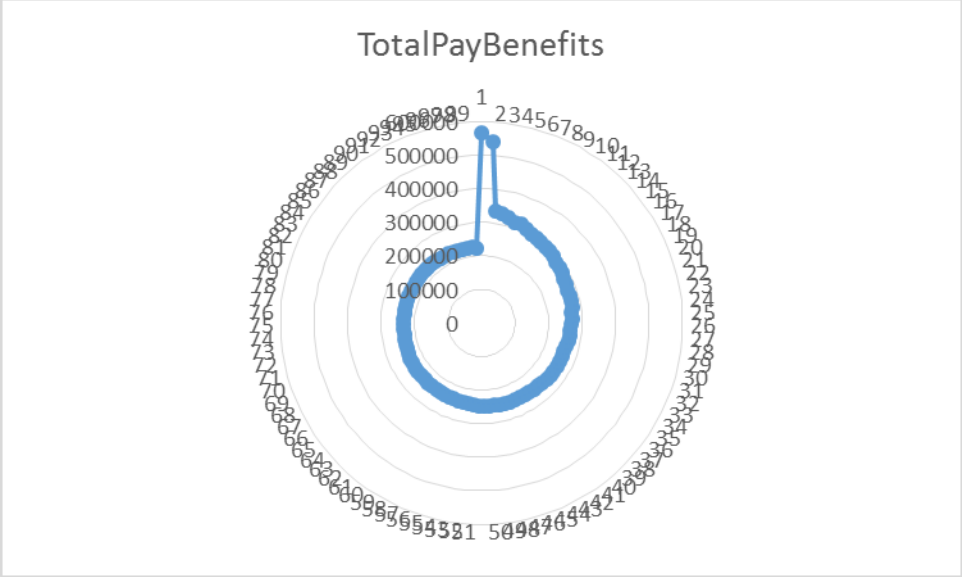


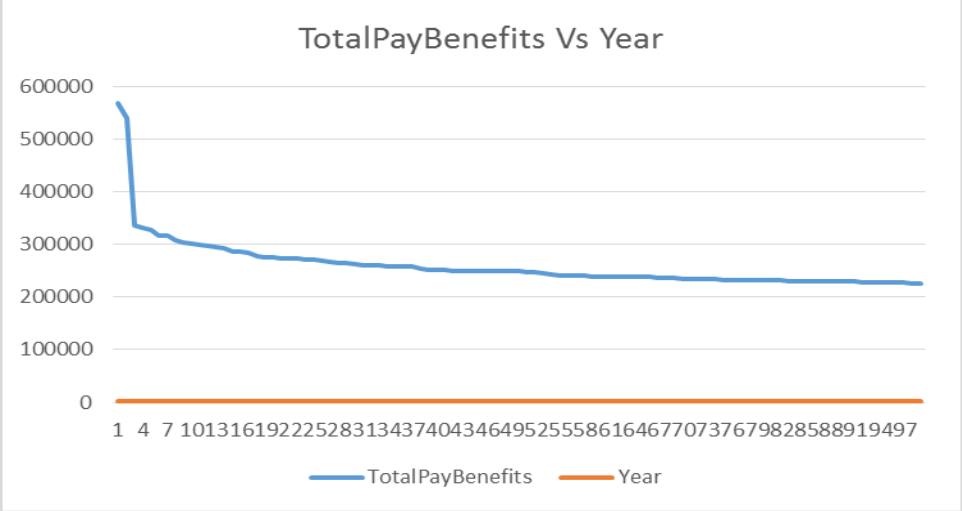

OvertimePay
250000
200000
150000
100000
50000
0
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96
OvertimePay


OtherPay
500000
400000
300000
200000
100000
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
OtherPay


TotalPayBenefits
600000
500000
400000
300000
200000
100000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97

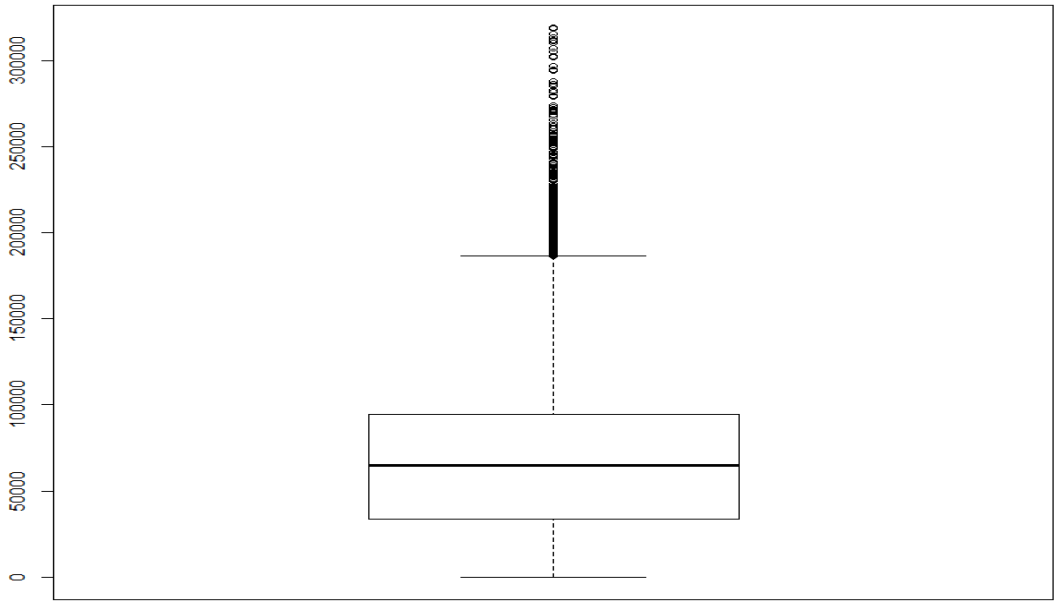

TotalPayBenefits
600000
500000
400000
300000
200000
100000
0
1 4 7 10131619222528313437404346495255586164677073767982858891949 7


TotalPayBenefits
500000
400000
300000
200000
100000
0

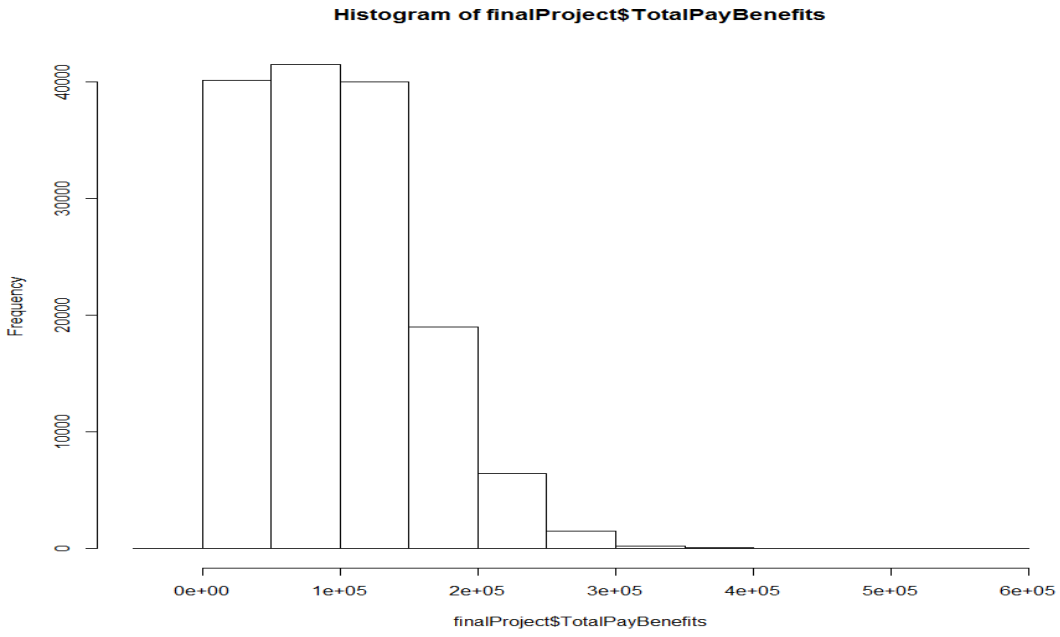

TotalPayBenefits Vs Year
600000
500000
400000
300000
200000
100000
0
1 4 7 10131619222528313437404346495255586164677073767982858891949 7
TotalPayBenefits
Year

SF Salaries – All Rows Analysis through Programming R

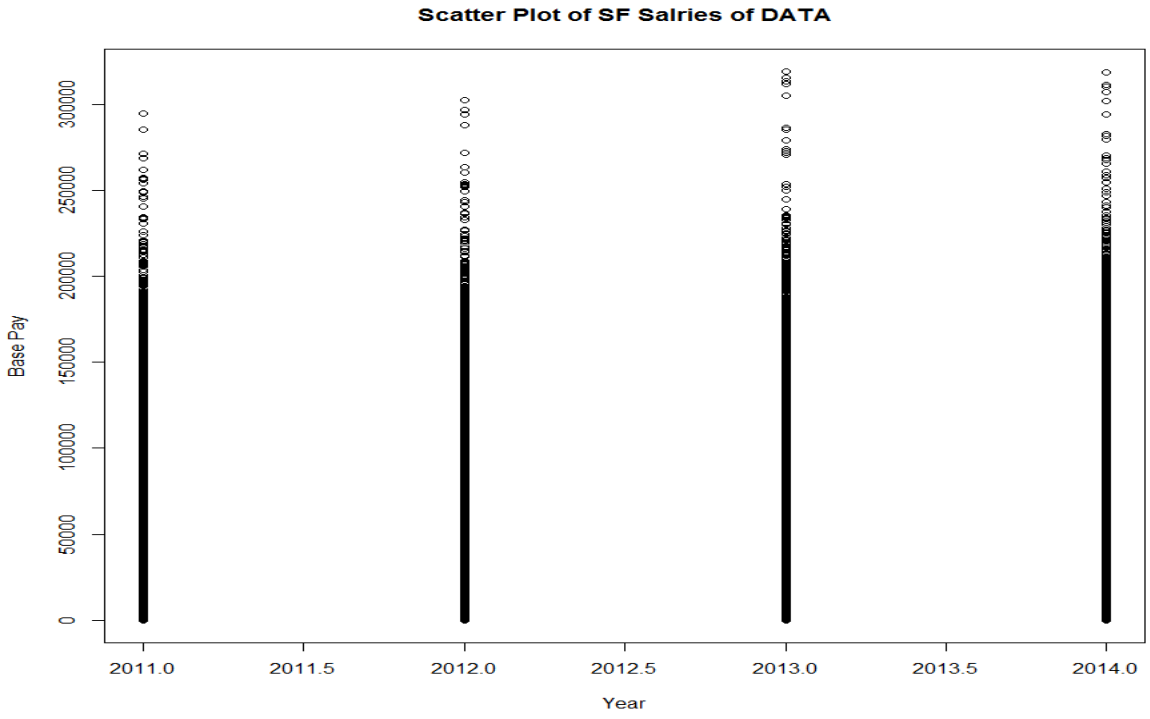
Box Plot (Base Pay)



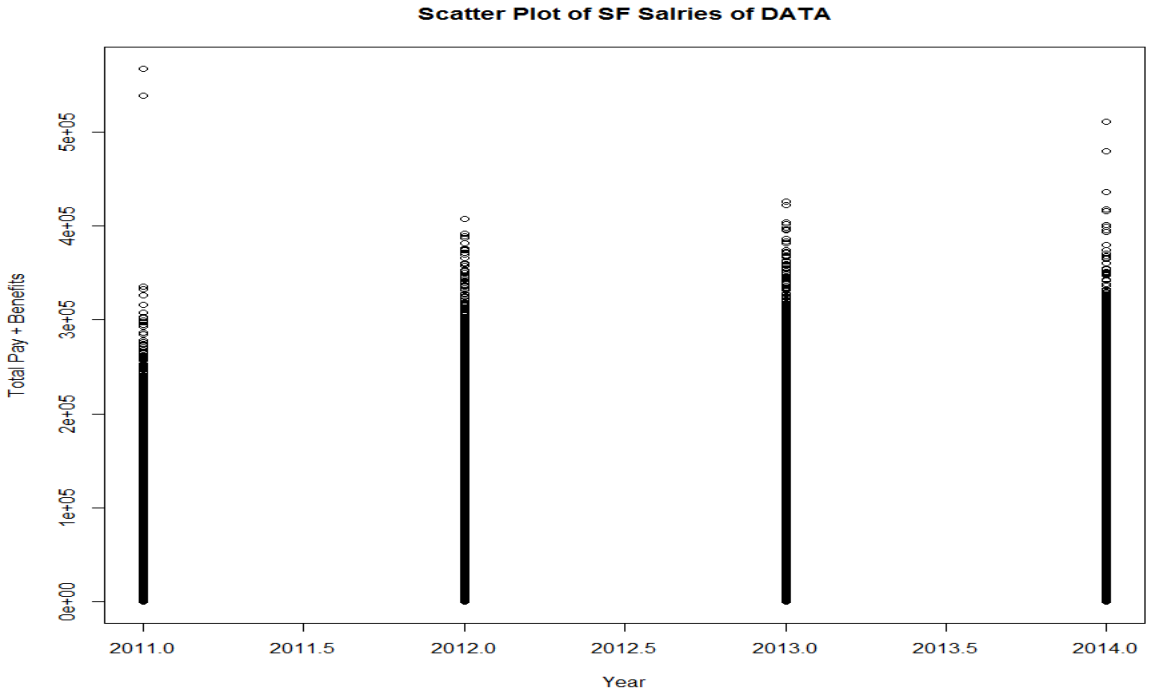
Histogramm of TotalPay Benefits



Scatter Plot of Year Vs Base Pay



Scatter Plot of Year Vs Total Pay + Benefits



R Programming commands for the above Charts

```
> finalProject<-read.csv('c:/temp/SFSalries-Cleaned.csv',header=TRUE)
> boxplot(finalProject$BasePay)
>
> boxplot(finalProject$TotalPayBenefits)
> hist(finalProject$TotalPayBenefits)
> x1<-finalProject$Year
> y1<-finalProject$BasePay
> plot(x1,y1, main="Scatter Plot of SF Salries of DATA", xlab="Year", ylab="Base Pay",)
> y1<-finalProject$TotalPayBenefits
>
> plot(x1,y1, main="Scatter Plot of SF Salries of DATA", xlab="Year", ylab="Total Pay + Benefits",)


>
```

R Programming Additional commands for other charts

boxplot(finalProject$loan_amnt)

hist(finalProject$loan_amnt)

boxplot(finalProject$int_rate)

hist(finalProject$int_rate)

boxplot(finalProject$funded_amnt)

hist(finalProject$funded_amnt)

boxplot(finalProject$total_amnt)

hist(finalProject$total_amnt)

boxplot(finalProject$last_pymnt_amnt)

hist(finalProject$last_pymnt_amnt)

(4.2)

Data Set : Loans

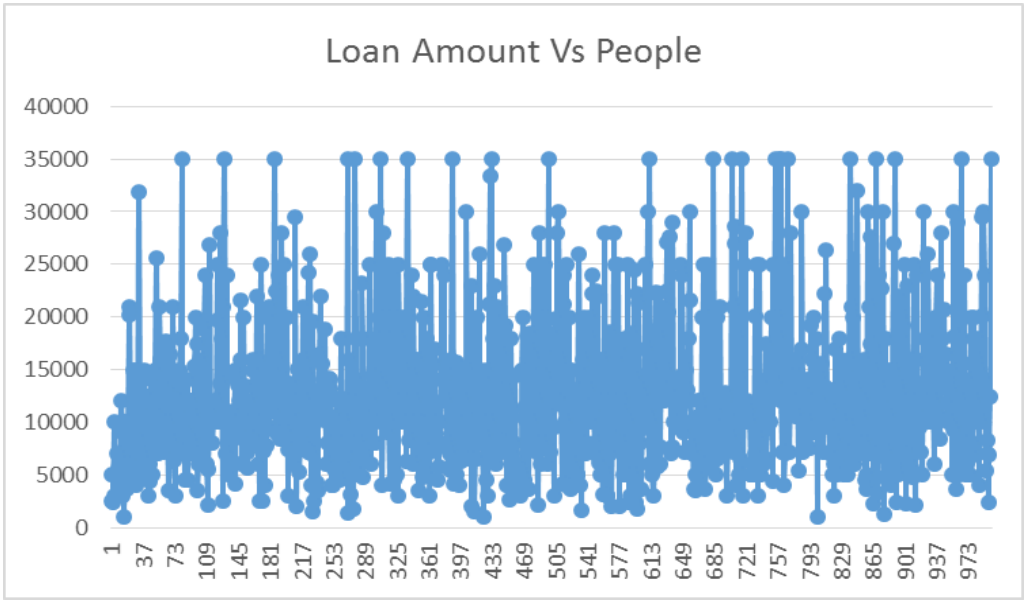
(Loan data for all loans issued through the 2007-2015)

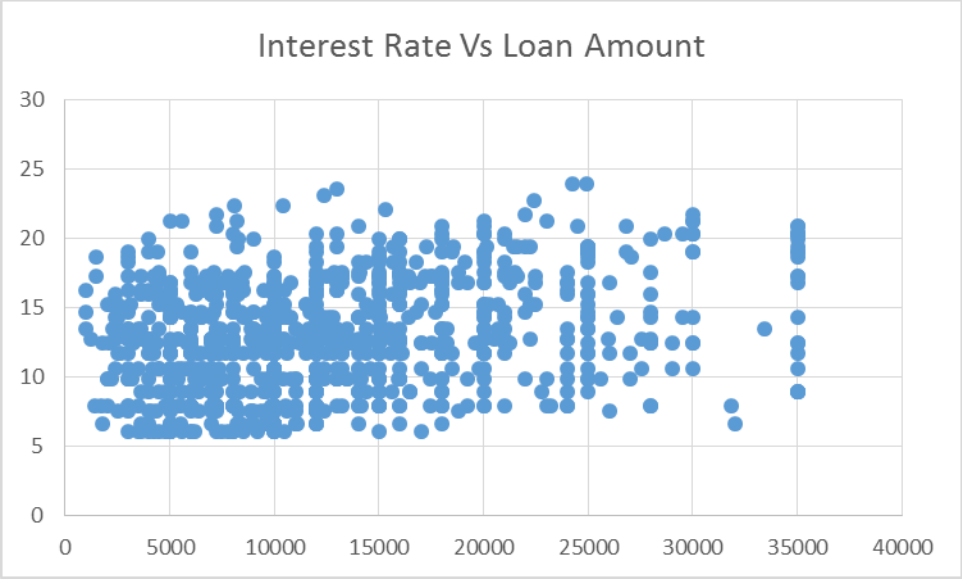
Zip File content attached with this write up

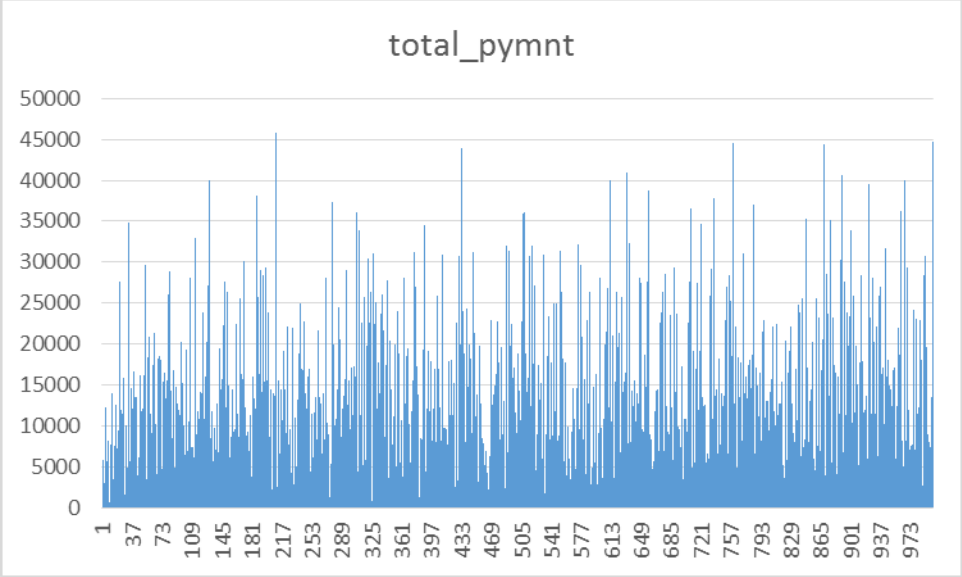
loans.csv	File download from kaggle.com
Loans.xls	Loaded the csv file in excel and have developed some chart using excel
Loans-1000rows.csv	sample 1000 rows csv file
Loans-1000rows.xls	Sample 1000 rows xls file

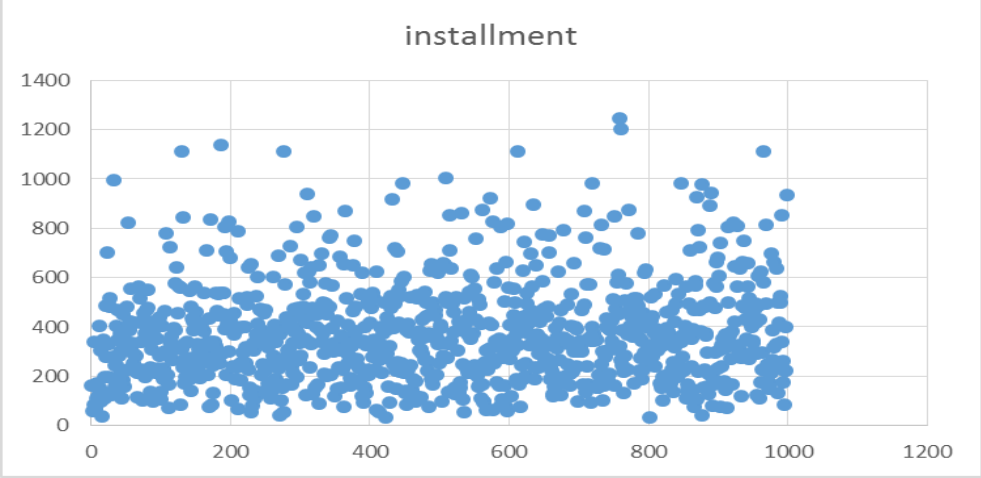
This file has over 480,000 rows. In order to work with Excel, this is too much row for Excel to handle it. so I have taken part of it . 500-1000 rows and making sample charts out

Charts Based out of Excel is given below



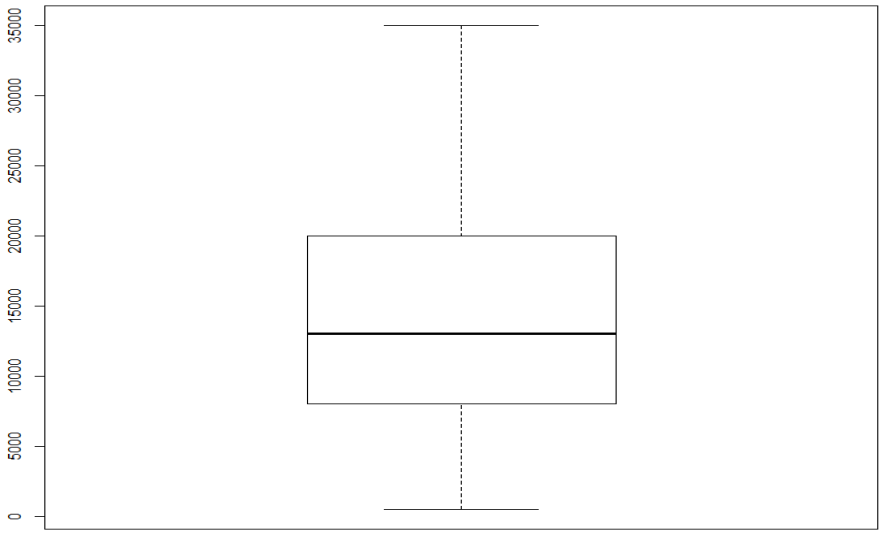

Interest Rate Vs Loan Amount
30
25
20
15
10
5
0
0
5000
10000
15000
20000
25000
30000
35000
40000

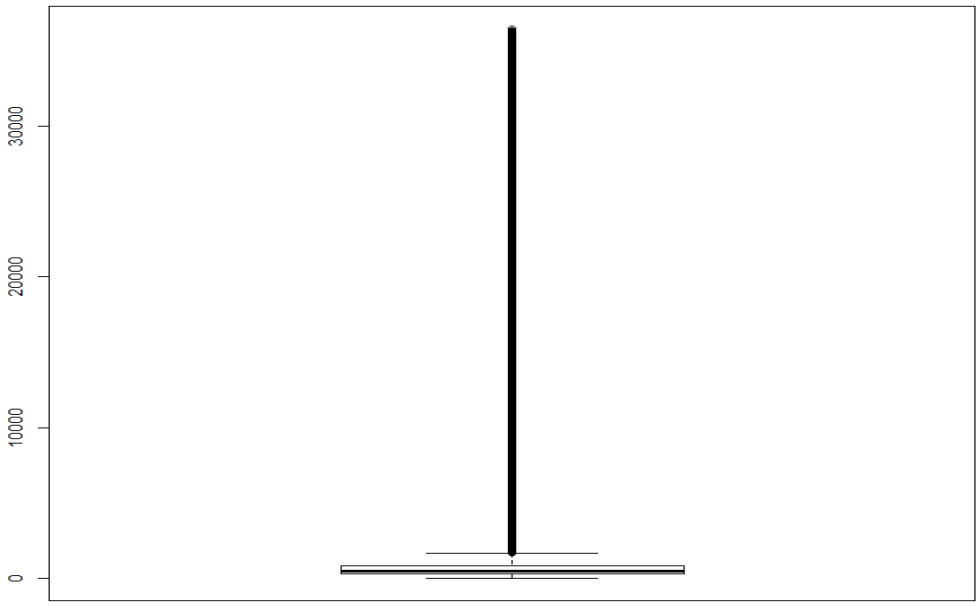

total_pymnt
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1
37
73
109
145
181
217
253
289
325
361
397
433
469
505
541
577
613
649
685
721
757
793
829
865
901
937
973

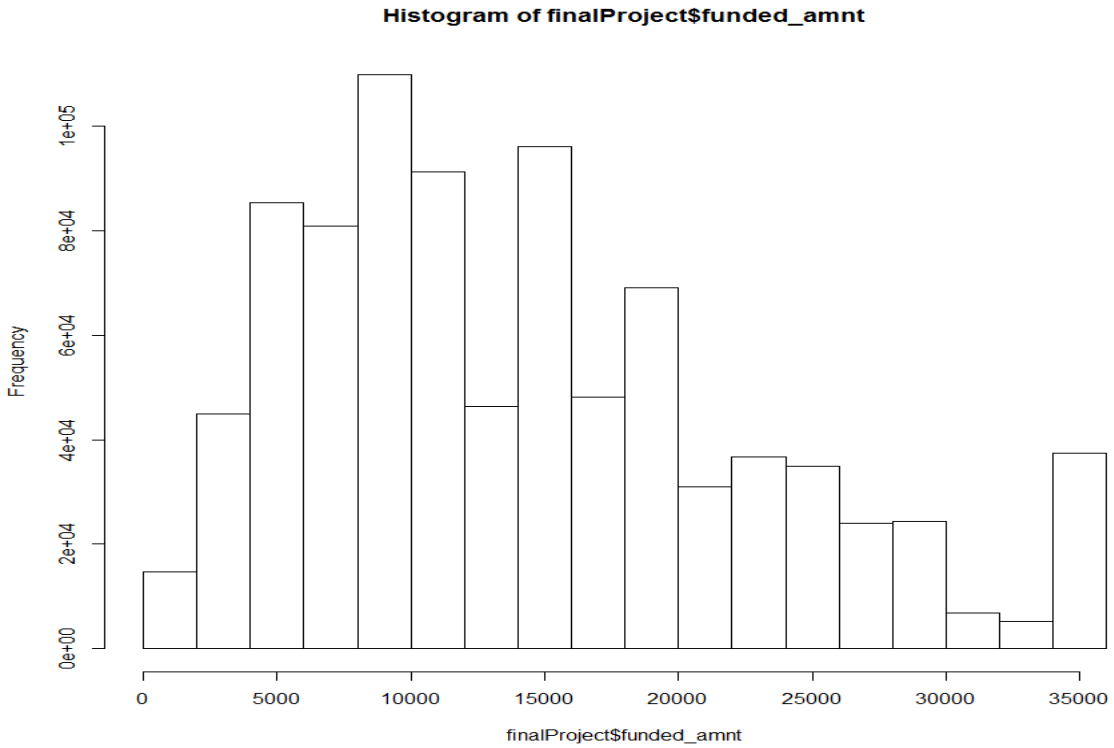

installment
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200

Charts Based on Programming R

```{r}

#### Plot loan amount, interest rate, funded amount , total amount and last payment amount

timeLine <- c(0 , +500000)

plot(finalProject$loan_amnt, type="b", xlim=timeLine)

timeLine <- c(0 , +500000)

plot(finalProject$int_rate, type="b", xlim=timeLine)

timeLine <- c(0 , +500000)

plot(finalProject$funded_amnt, type="b", xlim=timeLine)

timeLine <- c(0 , +500000)

plot(finalProject$total_amnt, type="b", xlim=timeLine)

timeLine <- c(0 , +500000)

plot(finalProject$last_pymnt_amnt, type="b", xlim=timeLine)
```








```
> finalProject<-read.csv('loan.csv',header=TRUE)
> timeLine <- c(0 , +500000)
> plot(finalProject$loan_amnt, type="b", xlim=timeLine)
> plot(finalProject$total_amnt, type="b", xlim=timeLine)
> boxplot(finalProject$loan_amnt)
```

```
> boxplot(finalProject$last_pymnt_amnt)
> hist(finalProject$funded_amnt)
```

Final Conclusion

There are several charts available. I have even loaded this in Talend Open Stuido which is Big Data Analytics tool. Over all the charts performed in Excel or Programming in R is all same representation and derive the right output.

Disclaimer

I am unable to load all 480,000 Rows or 180,000 rows in R Studio or Excel due the limitations in Laptop resources. Some cases, I have taken 100 or 1000 rows to analyze.