2019

Business Intelligence in the Finance Industry

EXPLORING LENDING CLUB LOAN DATASET

MAOYI SONG

Exploratory Data Analysis

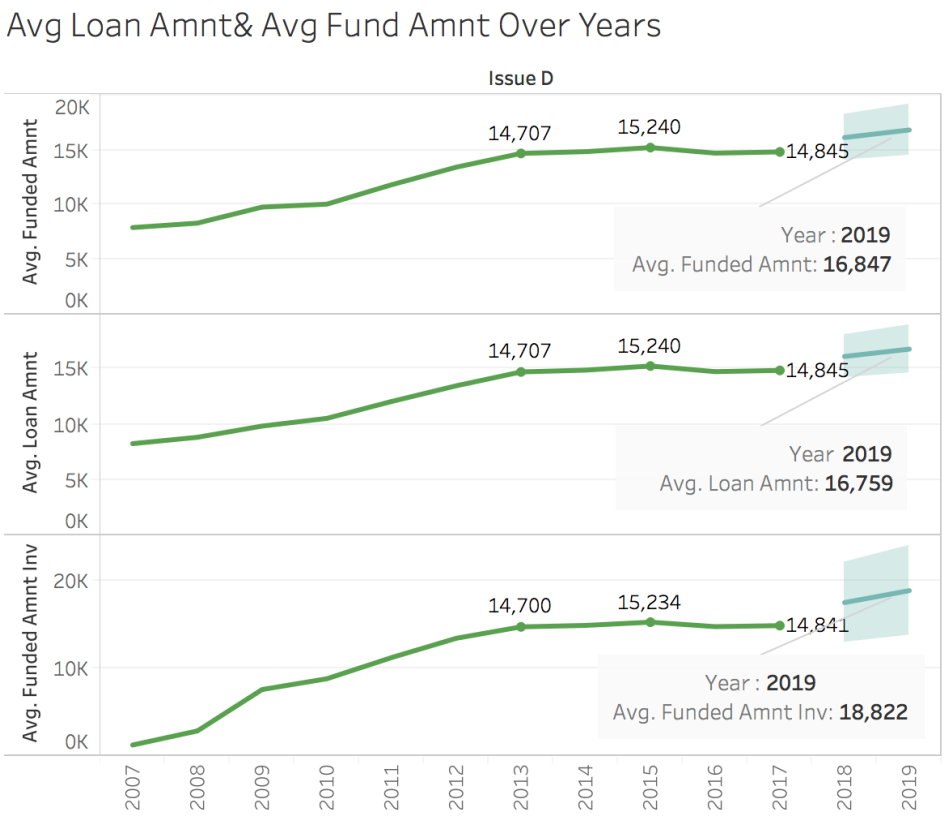
Data Preparation (Tableau)

- Change data type of **Issue D** to **Date.**
- Split the column of **Term** “36 months”, “60 months” to number **36 and 60.**
- Group “Charged off, Late (31-120 days), Late (16-30 days), In Grace Period, Does not meet the credit late, Default” to **Bad Loan,** “Does not meet the credit Full Paid, Full Paid, Current” to **Good Loan.**

Feature Distribution

What we should know:

- **Which year** we issued most loan and fund.
- **Prediction** for the next year (2019).
- What **amount mostly issued**.



Summary:

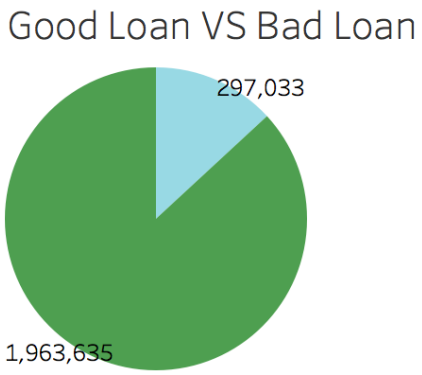
- The mostly issued loan amount is between **100,000 to 200,000.**
- Based on the history data, the prediction for average loan amount is 16,759 average funded amount is 16,847, and average of investor requested fund amount is 18,822 **in 2019**.
- **The year of 2015** was the year were most loans were issued.
- The distribution for these three graphs are similar, which means **most qualified borrowers are going to get the loan they had applied for.**

Types of Loans

What we should know:

- Combine the loan statues into two different groups, **good loan and bad loan**.
- “Charged off, Late (31-120 days), Late (16-30 days), In Grace Period, Does not meet the credit late, Default” to Bad Loan. Others to Good Loan.

Loan Status	
Fully Paid	1,041,952
Current	919,695
Charged Off	261,655
Late (31-120 days)	21,897
In Grace Period	8,952
Late (16-30 days)	3,737
Does not meet the credit ..	1,988
Does not meet the credit ..	761
Default	31



Summary:

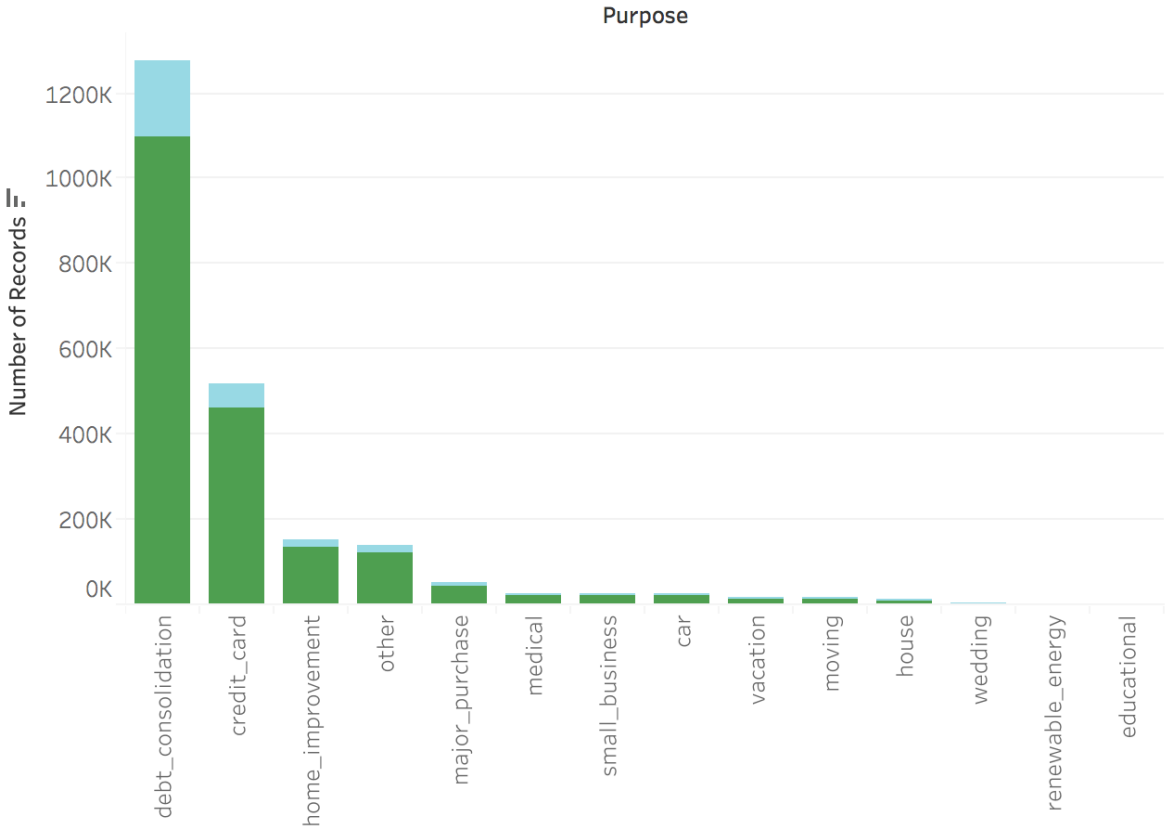
- Most of the records are **Fully paid** and **Current active**.
- **Good Loan** are **larger** than Bad Loan.

Good Loan VS Bad Loan

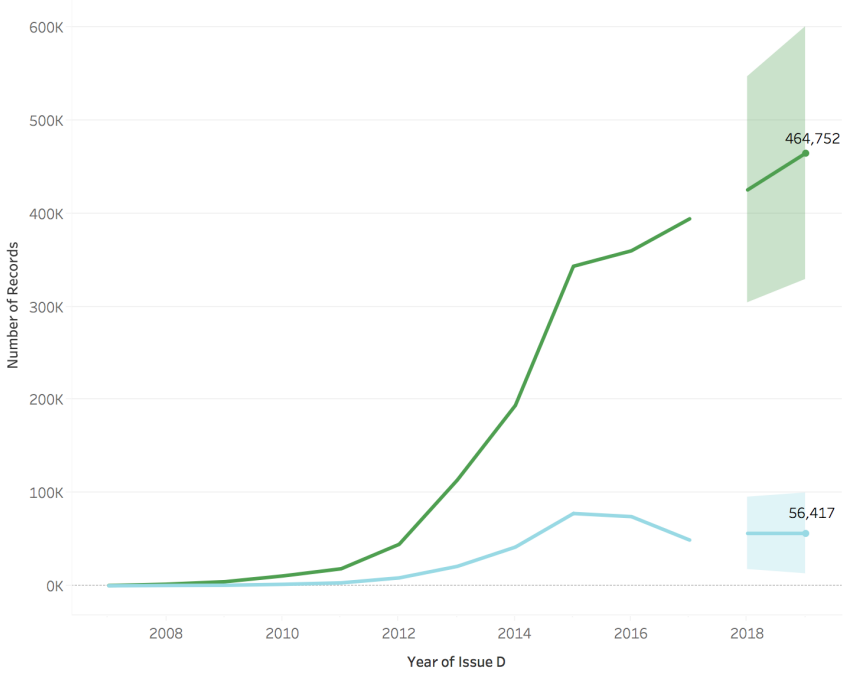
What we should know:

- How good loan and bad loan performed **over the years**.
- **Prediction** for the next year(2019).
- **Comparison** of the average interest rate between **good loan and bad loan**.
- More information about **bad loan**.

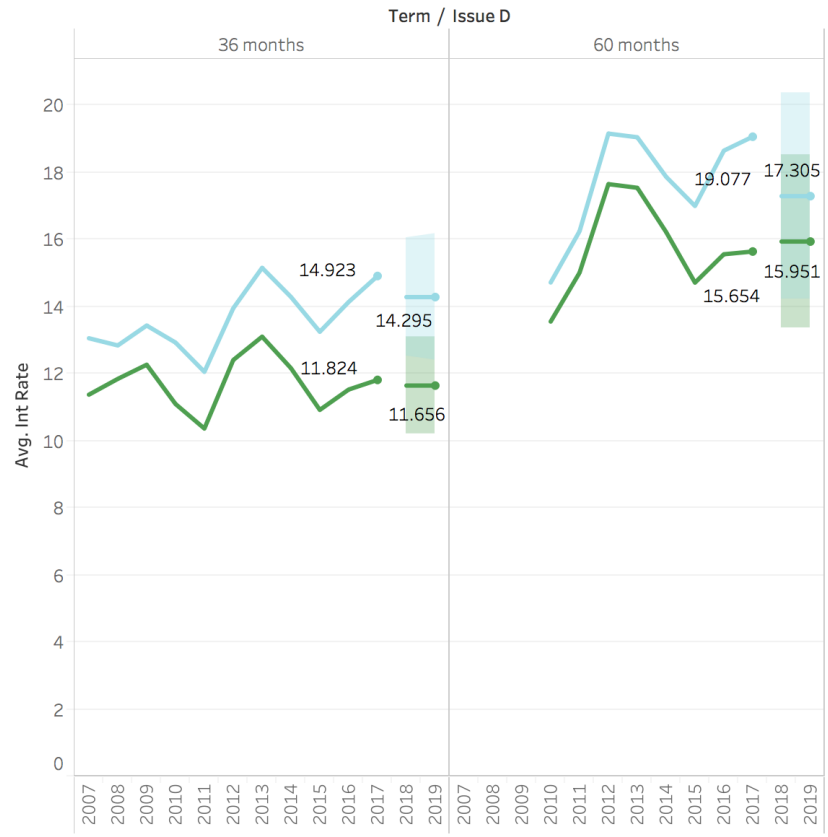
Types of Loans by Purpose



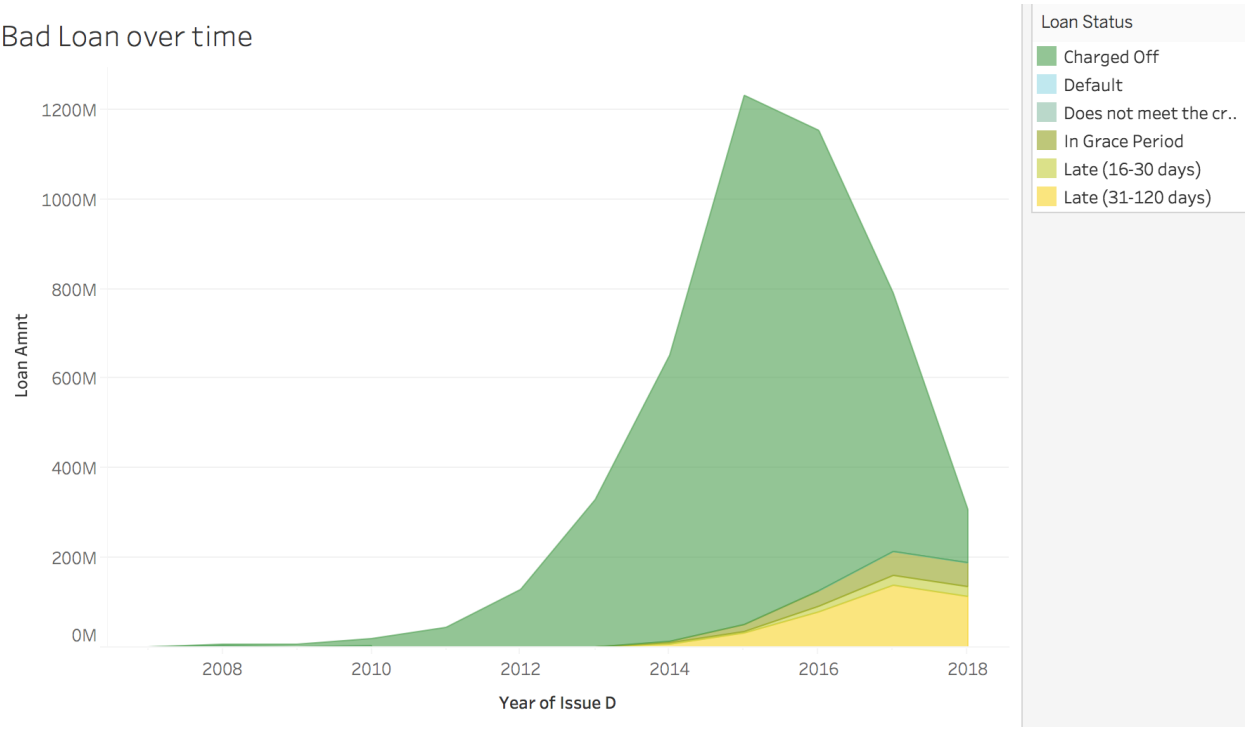
Timeline



Avg Interest Rate by Loan Condition And Term



Bad Loan over time



Summary:

- Most of borrower who issued the loan have the purpose of **debt consolidation or credit card.**
- **Good loan** trend to **increasing** fast in the **next year** at amount of 464,752. And, Bad loan will decrease in the next year.
- Most people choose to issued the loan in **60 terms** rather than 30 terms and both bad loans interest rates are higher than good loans.
- For the next year, the good loan's interest rate should be **decreased** to **17** for 60 terms and **14** for 30 terms. However, the bad loan's interest rate should stay the same or increase a little bit.

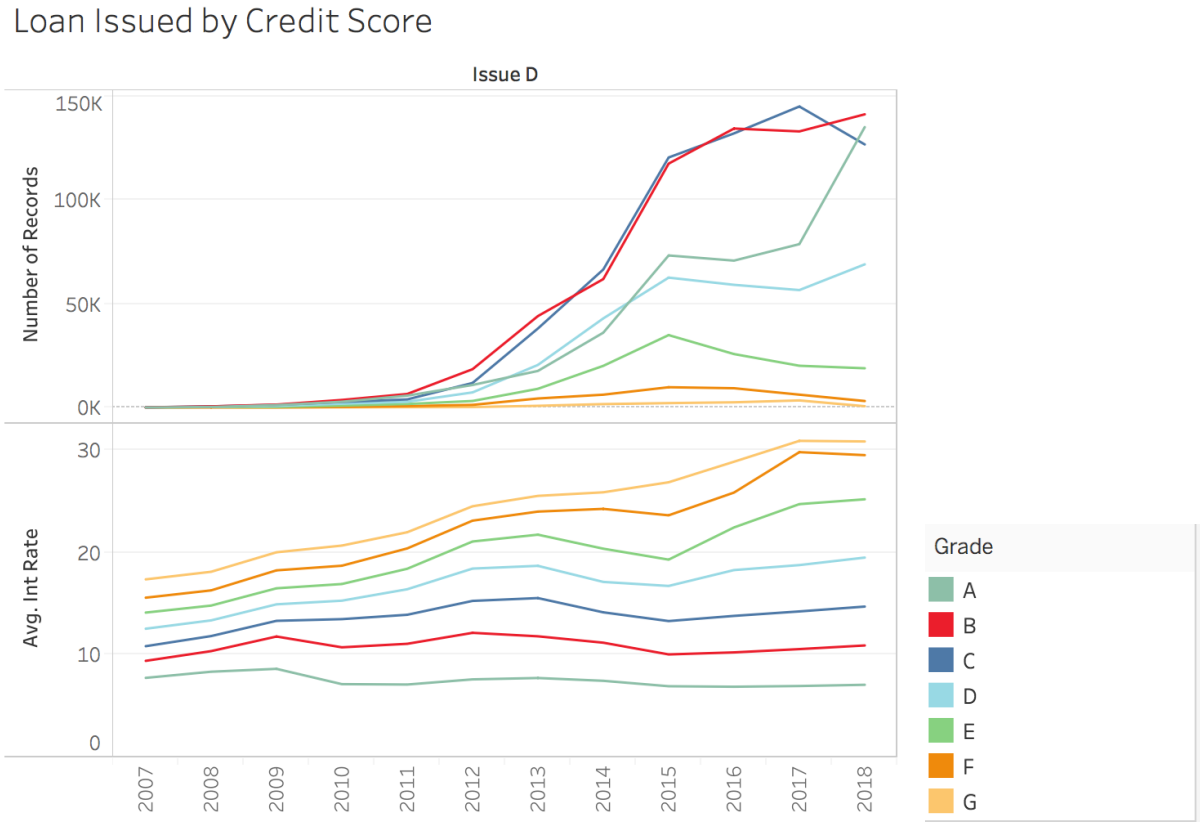
- Most of bad loans are **charged off in 2015.**

Credit Score and Risk

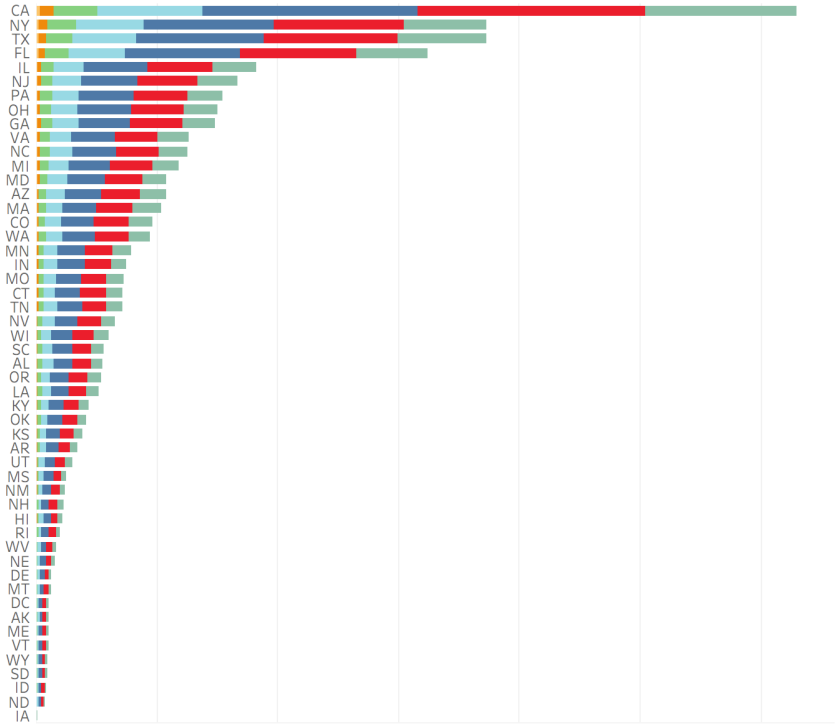
What we should know:

- There are 7 different grade which indicated different levels of risk.

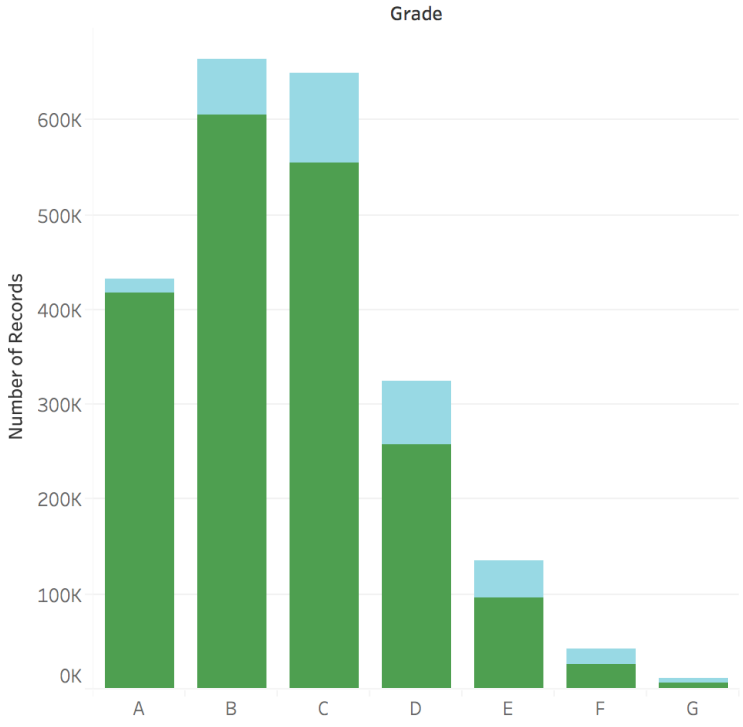
- Which grade has the highest number of issued loan.
- How it performed after we grouped in two different category loan.
- Grades among different regions.



Total Loan Issued by Region



Types of Loans by Grade



Summary:

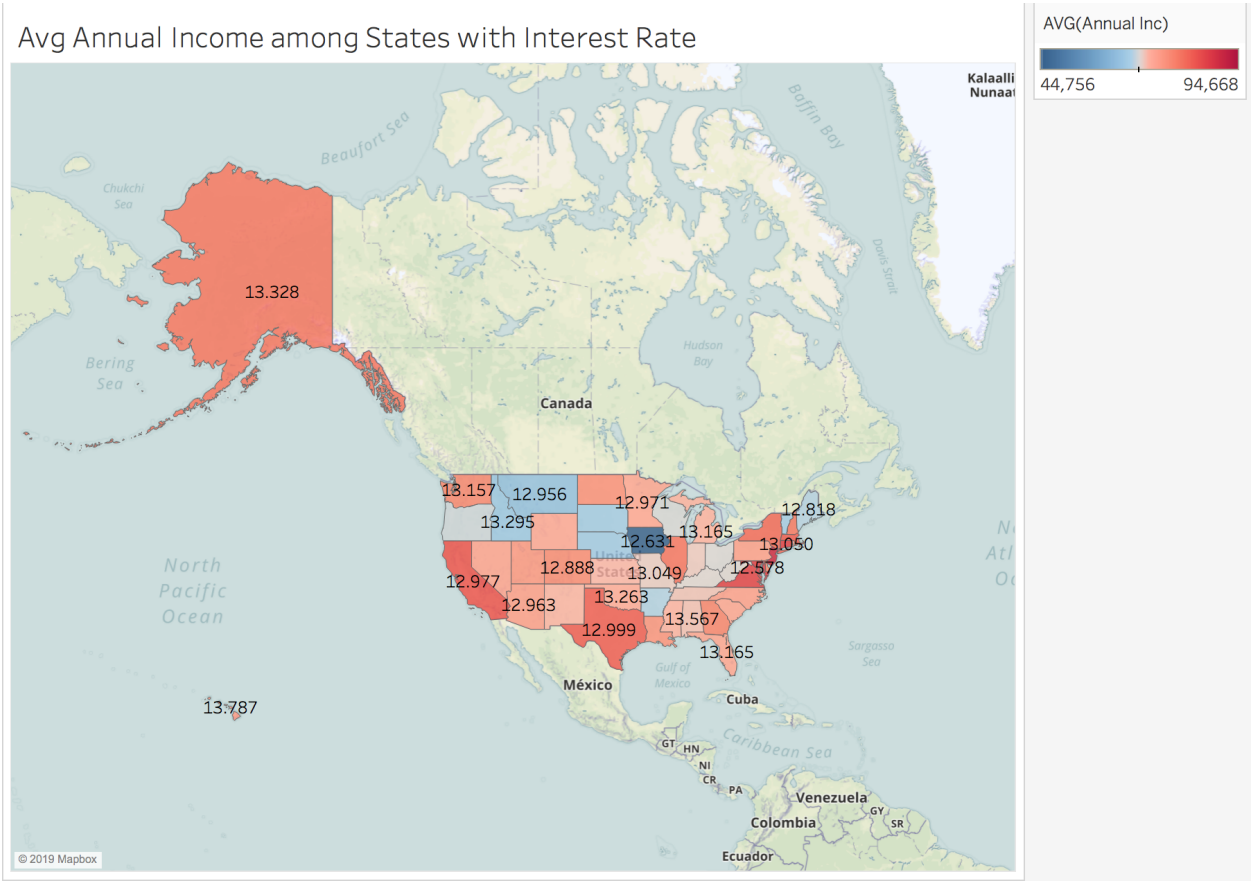
- The **gap** of **average interest rate** between **D, E, and F** is **huge** in **2017-2018**. And **E and F** interest rate has **increase**d fast from 2015.
- **Most people** issued the loan in grade **B and C**, especially in the state of **CA and NY**.
- People who issued the loan at **A** grade has a big **increase** from 2017.
- People who issued the loan at **E, F and G** grades has **decreased** from 2017.

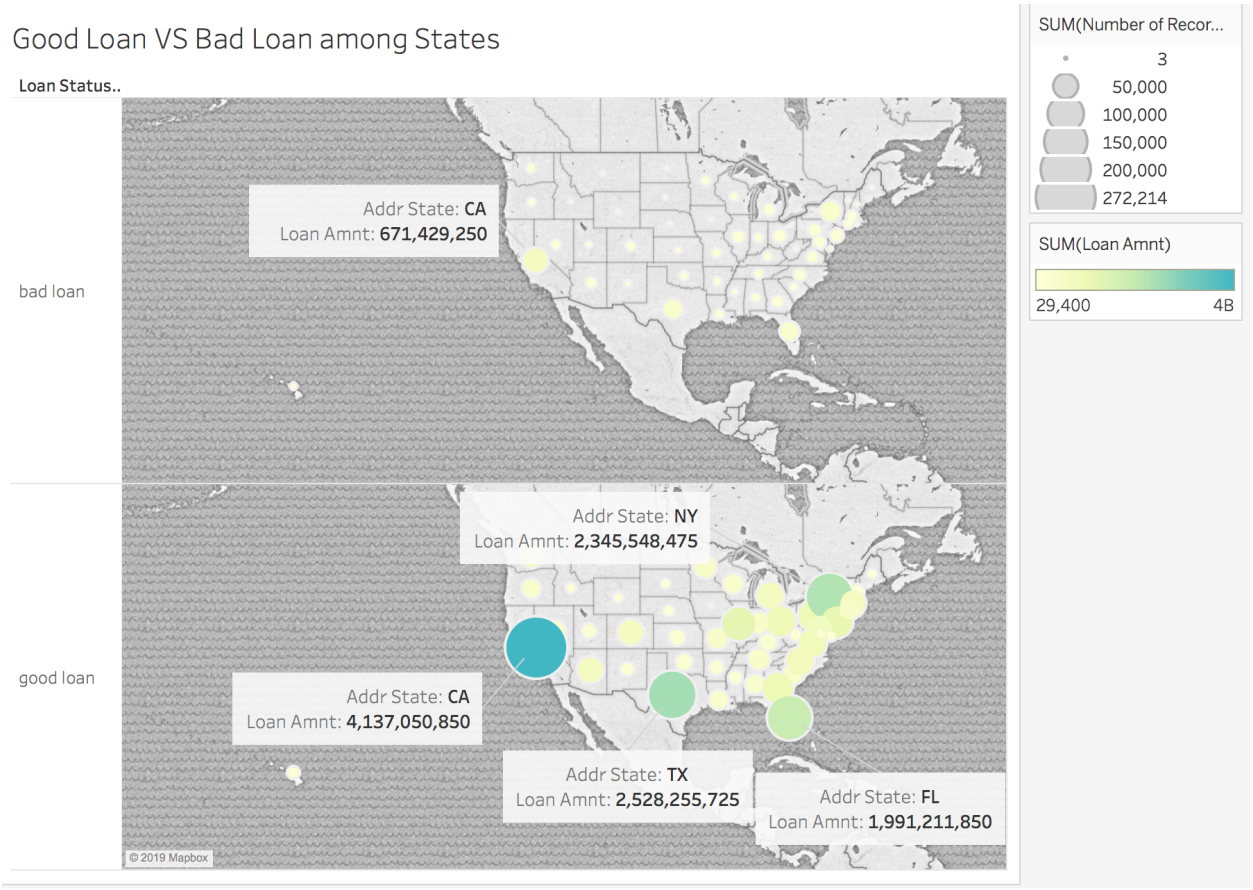
- Most of loan are in good condition, and **grade C** has the **highest bad loan percentage** compare to others.

Region Contribution

What we should know:

- Which state has the highest average annual income
- What average interest rate for each state
- Bad loan vs good loan among the states





Summary

- **New York** and **California** has the highest average annual income.
- Each state has different interest rate, in **Seattle**, people receive the interest rate over **18**.
- **CA** has the highest amount of good loan, and followed by **NY, TX** and **FL.**

Feature Selection and Modeling

1. **Data Mining (Python):**

In Tableau:

- Delete the descriptive, categorical, and discontinuous columns.
- Export data to .cvs file.

In Excel:

- Change A - G to 1-7 in both column of **Grade** and **Sub_Grade**.

In python:

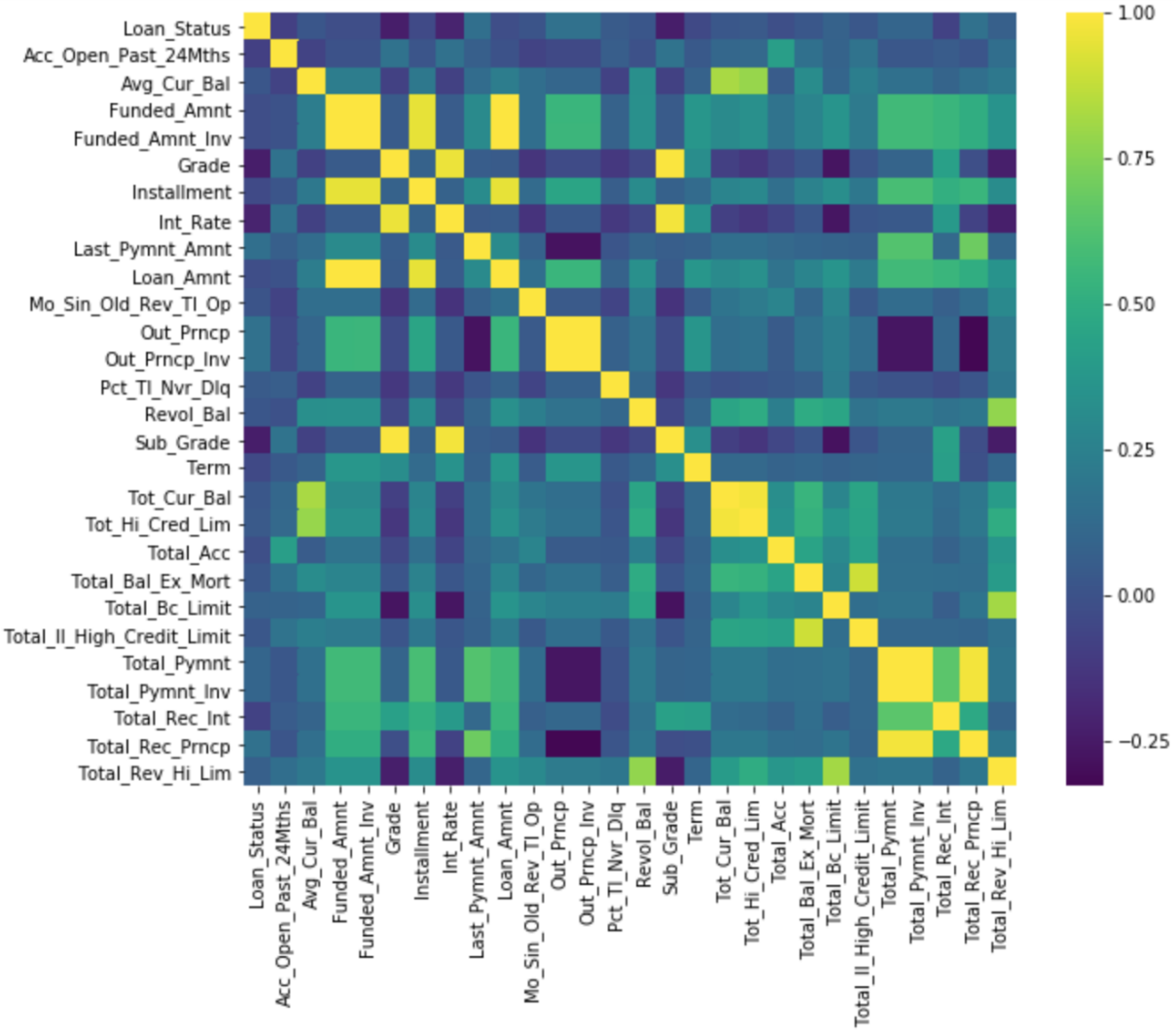
- Drop **Nan** Values in the columns.
- Randomly select **5000** records.

	Loan_Status	Acc_Open_Past_24Mths	All_Util	Annual_Inc	Avg_Cur_Bal	Bc_Open_To_Buy	Dti
count	5000.000000	5000.000000	4216.000000	4.956000e+03	5000.000000	4946.000000	4991.000000
mean	0.886600	4.534800	56.748102	7.943830e+04	13280.702400	12199.822887	19.005700
std	0.317113	3.181826	20.921117	1.196619e+05	16121.850318	17152.065700	18.329072
min	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	1.000000	2.000000	43.000000	4.700000e+04	2954.750000	2049.500000	11.755000
50%	1.000000	4.000000	58.000000	6.500000e+04	6994.500000	6025.500000	17.750000
75%	1.000000	6.000000	72.000000	9.500000e+04	18339.000000	14936.500000	24.475000
max	1.000000	26.000000	129.000000	7.600000e+06	235434.000000	175120.000000	999.000000

2. Correlation between Features

Summary:

- Most of the features are correlated with **Loan_Status**
- Some features are correlated with each other, such as **Grade and Sub_Grade**.



3. Feature Selection

Summary:

- Best score 0.55 when choose 19 features.
- Drop the irrelevant features.

```
Num Features: 19
Selected Features: [ True False False False False  True  True  True  True  True  True False
  True False False  True  True  True  True False  True  True  True False
 False  True False False False False  True  True False  True False]
Feature Ranking: [ 1  3 17 12  8  1  1  1  1  1  1  5  1 10  2  1  1  1  1  7  1  1  1 14
 13  1 16 15 11  9  1  1  4  1  6]
[(1, 'Bc_Open_To_Buy'), (1, 'Dti'), (1, 'Funded_Amnt'), (1, 'Funded_Amnt_Inv'), (1, 'Grade'), (1, 'Installment'), (1,
'Last_Pymnt_Amnt'), (1, 'Loan_Status'), (1, 'Mo_Sin_Old_Rev_Tl_Op'), (1, 'Out_Prncp'), (1, 'Out_Prncp_Inv'), (1, 'Pct
_Tl_Nvr_Dlq'), (1, 'Revol_Bal'), (1, 'Revol_Util'), (1, 'Sub_Grade'), (1, 'Tot_Hi_Cred_Lim'), (1, 'Total_Il_High_Cred
it_Limit'), (1, 'Total_Pymnt'), (1, 'Total_Rec_Int'), (2, 'Max_Bal_Bc'), (3, 'Acc_Open_Past_24Mths'), (4, 'Total_Pymn
t_Inv'), (5, 'Int_Rate'), (6, 'Total_Rec_Prncp'), (7, 'Percent_Bc_Gt_75'), (8, 'Avg_Cur_Bal'), (9, 'Total_Bc_Limit'),
(10, 'Loan_Amnt'), (11, 'Total_Bal_Il'), (12, 'Annual_Inc'), (13, 'Tot_Cur_Bal'), (14, 'Term'), (15, 'Total_Bal_Ex_Mo
rt'), (16, 'Total_Acc'), (17, 'All_Util')]
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:578: DataConversionWarning: A column-vector y was
passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

#number features=19, score=0.55

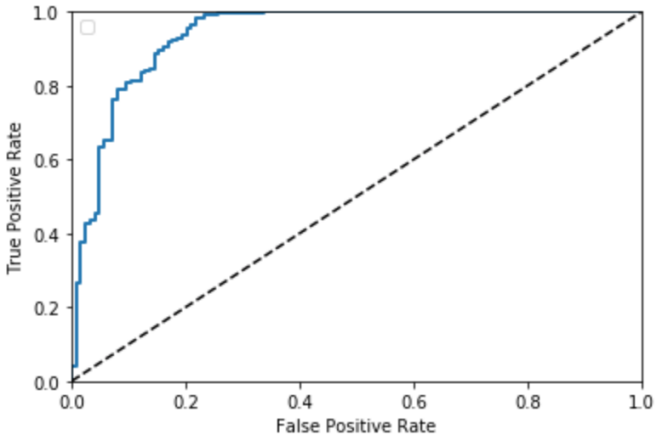
4. Random Forest Modeling & ROC Score

Summary:

- Highest score with **Random Forest Model** compare with others.
- Best score with Random Forest Model with **Max depth=15.**
- **Best model** when cv=10, accuracy score=0.958, precision= 0.955 and recall=0.998.
- ROC AUC=0.848, model has around **85%** of accuracy.

```
Kernel SVC: 0.887400 (0.014589)
Decision Tree: 0.942400 (0.009113)
Logistic Regression: 0.963000 (0.006277)
BOOSR: 0.940000 (0.006450)
RF15: 0.966400 (0.009583)
SVM linear: 0.915800 (0.095881)
KNN: 0.901400 (0.012651)
```

```
RF4: 0.899000 (0.011841)
RF5: 0.926800 (0.012007)
RF6: 0.945400 (0.012587)
RF7: 0.953400 (0.012101)
RF8: 0.957600 (0.011723)
RF15: 0.966400 (0.009583)
RF20: 0.966200 (0.009734)
RF25: 0.966400 (0.009583)
RF30: 0.966400 (0.009583)
RF35: 0.966400 (0.009583)
```



Recommendations

Based on the EDA, we can see the most qualified borrowers are going to get the loan they had applied for. The prediction for average loan amount is 16,759 average funded amount is 16,847, and average of investor requested fund amount is 18,822 in 2019. The total loan amount and funded amount will increase in the next year. In order to increase the percentage of total good loans or the people who issued loan, we can:

- Pay attention to the people who issued the loan on the purpose of debt consolidation, has been charged off or marked as late, choose the loan on C grade in 60 term.
- Decrease the interest rate for the loan in grade of D.
- Decrease the interest rate of good loan in the next year for both 36 and 60 terms.
- Re-assign the average interest rate over states. For example, decrease the average interest rate in Settle in order to increase the total amount of issued loans.
- Build the random forest model to predict future loan status for each portfolio.