Bank Loan's dataset investigation

Investigation Overview

> I am going to draw 9 pictures in Univariate,bivariate and multivariate chapter.

> For univariate, I am going to find some characteristics of dataset, including distribution of credit rank, interest rate and borrower's location on the map.

> For bivariate visualisation, I am going to dig deeper based on the result above. Such as the different rank's interest rate, the relationship between loan price and interest rate and average loan USD in different states.

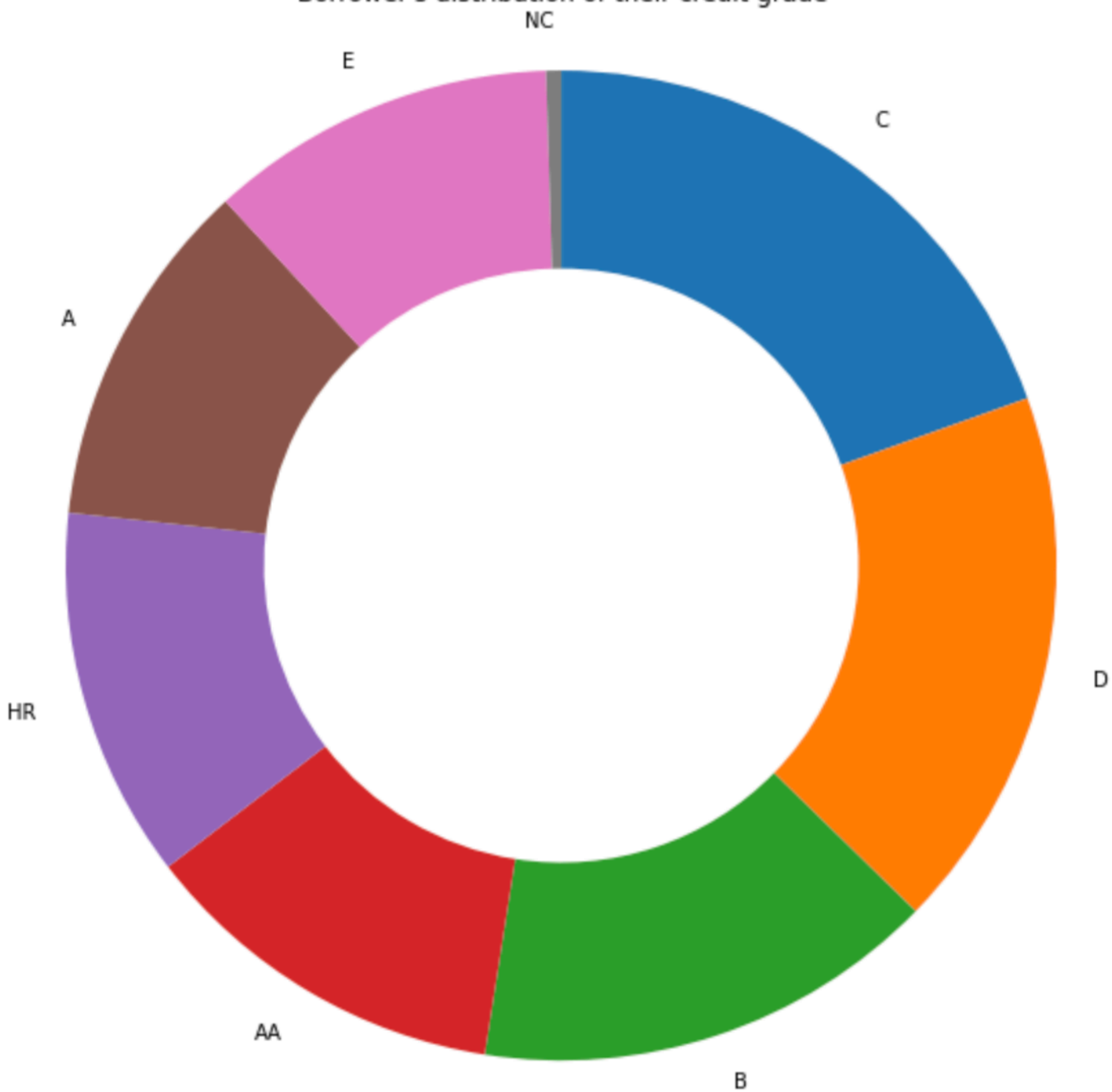
> For multivariate visualisation, I am going to add more dimension on bivariate visualisation result.

Dataset Overview

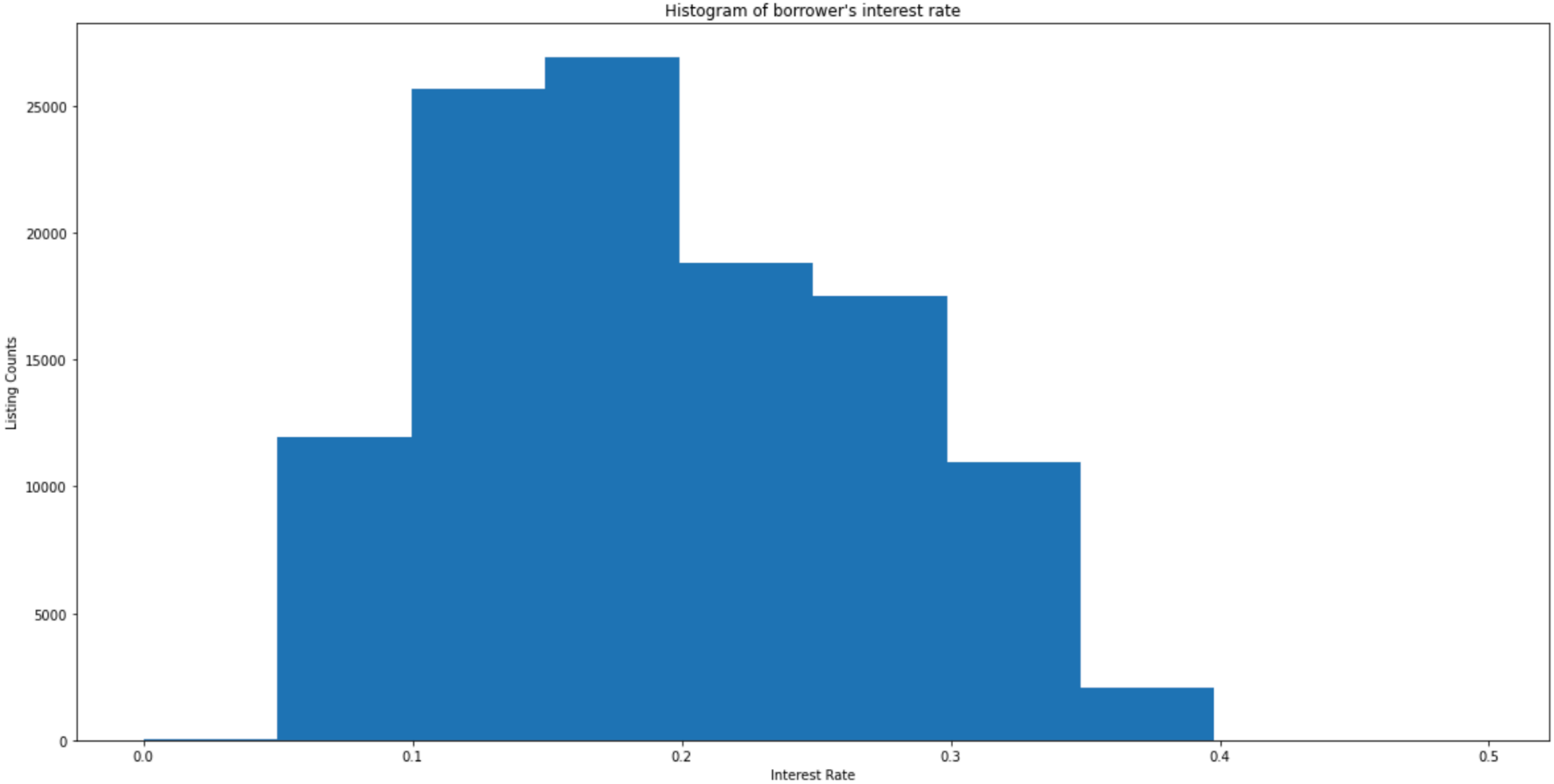
> 81 columns, 113,937 rows, no duplicate value, some of columns needs wrangle.

Visualization 1

> Here, we found the distribution of different credit rank. However, from the pictures below we cannot find very obvious information. If we add more variable, we may find some valuable information from mass dataset.

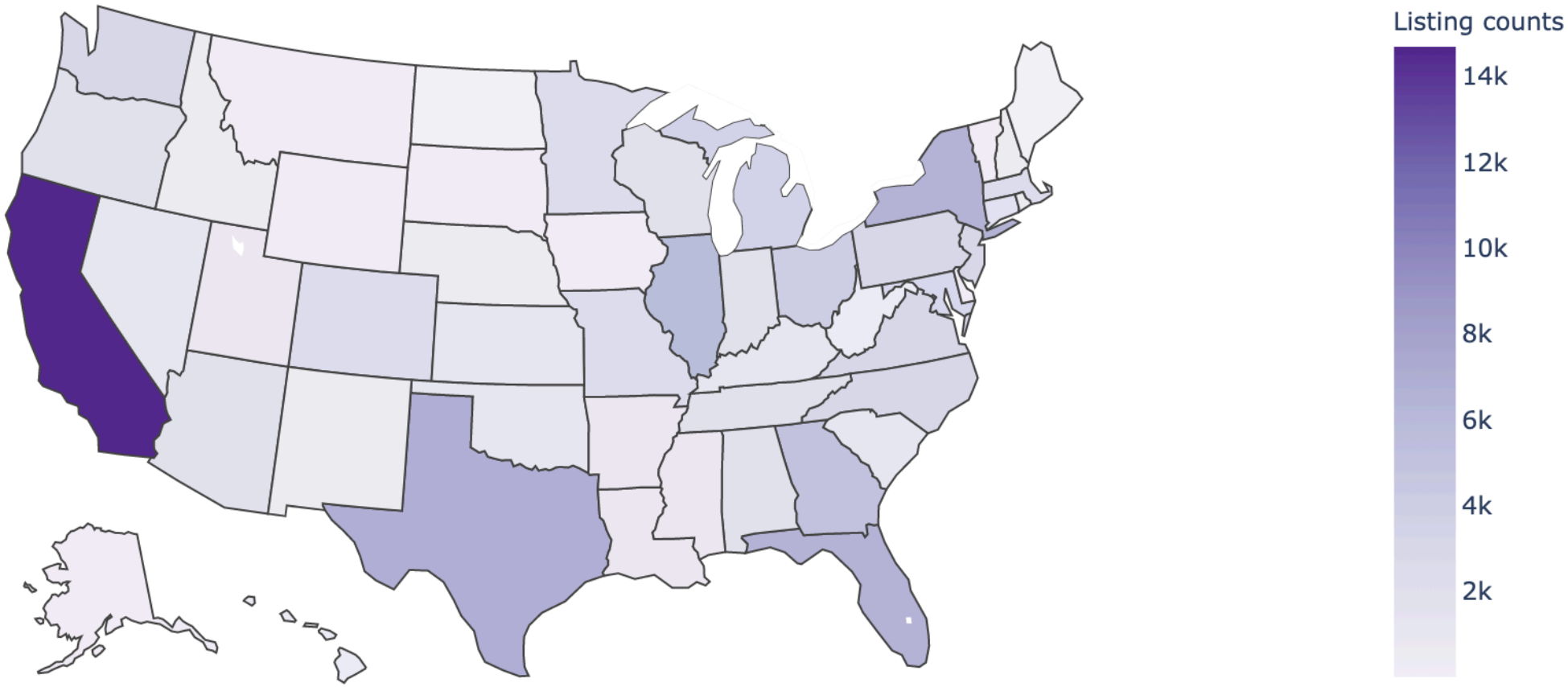

Borrower's distribution of their credit grade
NC
E
C
A
D
HR
AA
B

> From the picture below we know that most of borrower's rate is between 10% to 20%. For the next step, we can try to analyse the characteristics of extremely low(lower than 5%) and high(higher than 35%). Do they have higher/lower credit rank? Where are they living? What is their occupation? We may find the reason why they can borrow money cheaper/more expensive.


Histogram of borrower's interest rate
Listing Counts
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5
Interest Rate

> The first impression of the visualisation below are brilliant. However, I found the listing distribution is perfectly match the population of the United States. It weaken the result of the map below. So, if I am going to dive deeper, I am going to use some other variables to find the relationship between the listing in different states.

Distribution of borrower in the US nationalwide.



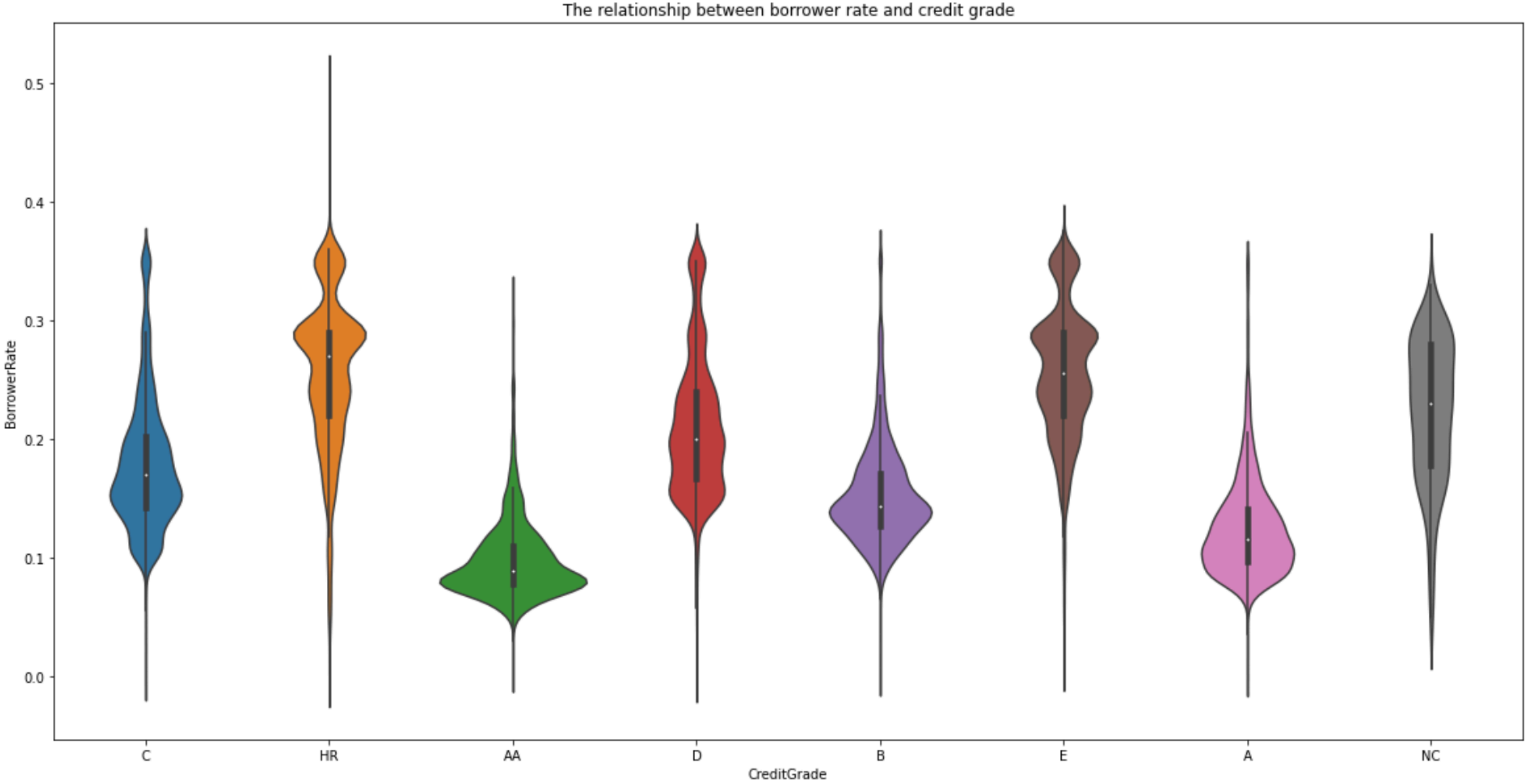
Visualization 2

> There are some unusual points in interests, we can see some incredible low and high interests in the distribution histogram. We may find more about what kind of people can borrow money with cheaper/more expensive price.

> Firstly, We found there are not enough information in the first pie chart. Credit rank information needs to combine with other variables to find useful information. For geographic information, it gives some useful information, but those information can be reasoning from people's common sense. If we combine this and some other variables, it may gives us better insight(such as average loan, interests, etc.)

> Here, combine the borrower rate and credit rate, we can find the relationship between borrower rate and their credit. In the most of situations, if people have better credit grade, they can borrow money with cheaper price. such as people with AA grade's obvious peak is lower than 0.1, which means less than 10 per cent. But in grade A cluster, the peak of violin is settled higher than 0.1, approximately 11-12 per cent.

> In the next chapter I am going to use facet grid to dive deeper based on the result we have to find more information group by credit grade.


The relationship between borrower rate and credit grade
BorrowerRate
0.5
0.4
0.3
0.2
0.1
0.0
C
HR
AA
D
B
E
A
NC
CreditGrade

> From the scatter plot below, we can see approximate shape of the dots are triangle skewed to the left. This means for those people can borrow money in cheaper price, in most of situation they can borrow more money than people in higher interest rate cluster.

> Another interesting result is from the chart, most of people choose to borrow whole number of money. We can see some clear line in 10000, 15000, 20000, 25000 and 30000.

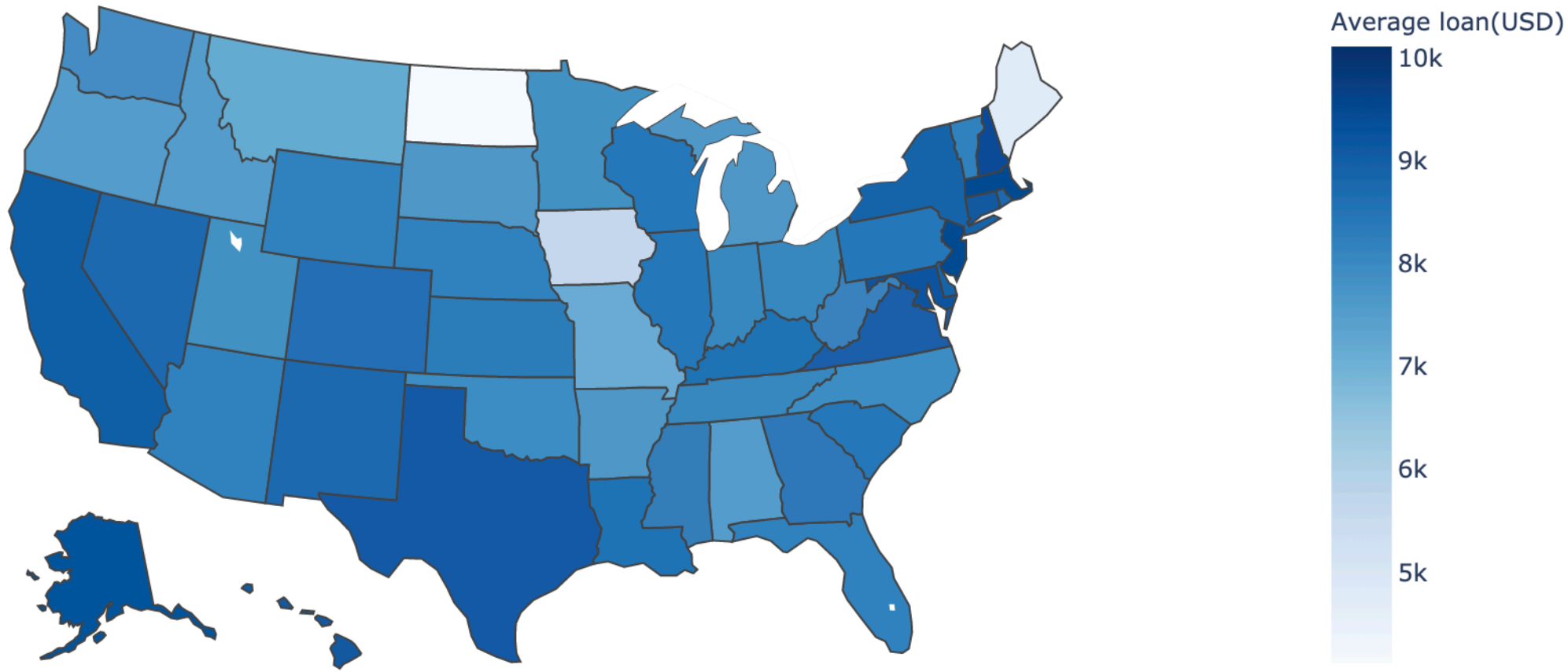
> We are going to find the difference between estate owner and non estate owner's loan status in the next part.


The relationship between borrower rate and loan original amount
35000
30000
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5

Comparing with the the first choropleth, this one gives us more information. We can see east/west coast's average loan is higher than the rest of country, and there are some clusters on the map, such as new England and California-Nevada has higher average loan, and bible belt have lower average loan. However, there are three special outliers on the map. They are North Dakota, Iowa and Maine. their average loan is extremely lower than the most of states in the US.

Maybe we can add more variables to find more information about the difference between each states, and luckily we may find the reason of those outlier.

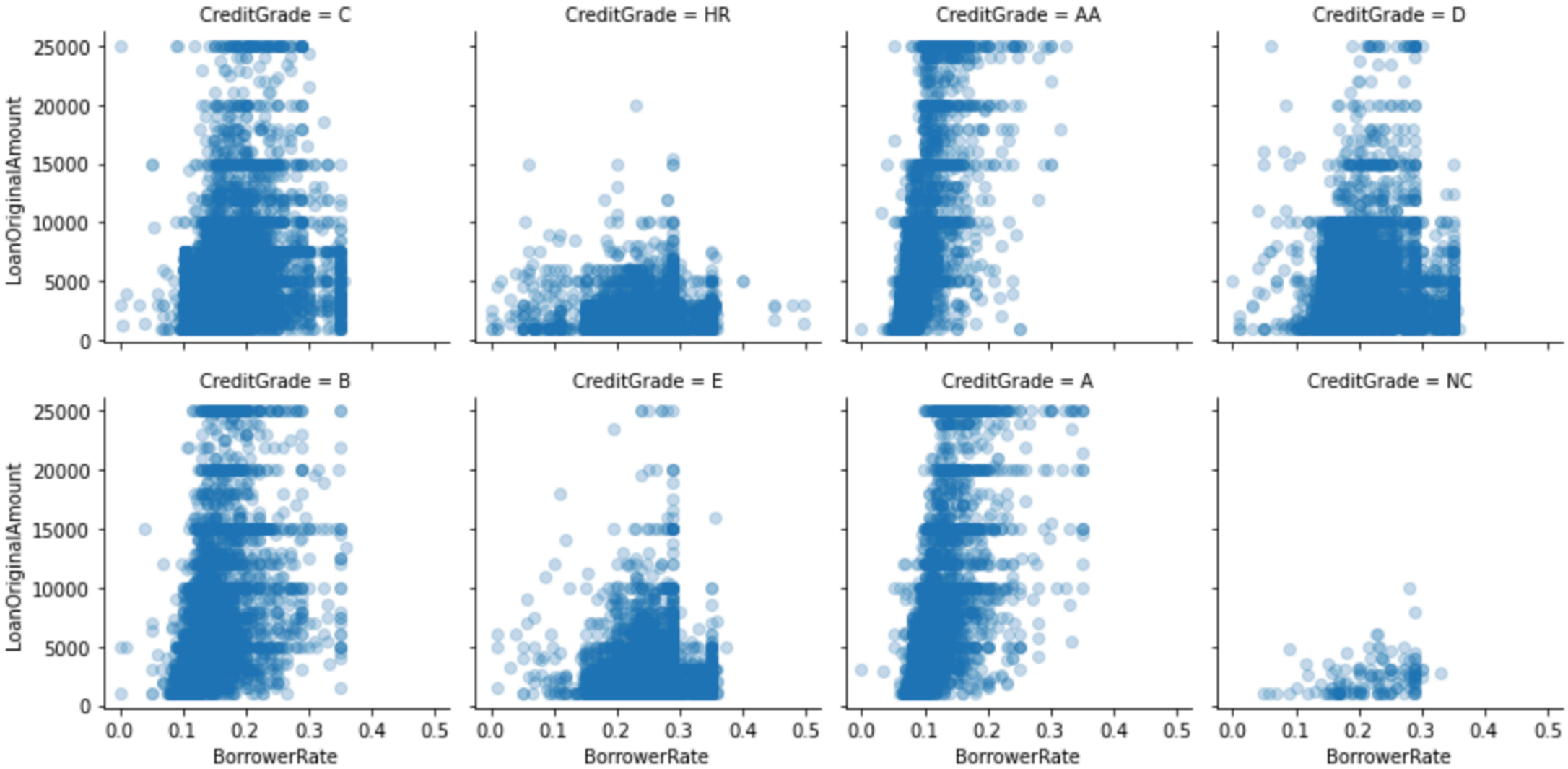
Distribution of loan original amount in the US nationalwide.



Visualization 3

> Now, I am going to add more dimension on bivariate visualisation result.

> The following result shows in different credit grade's people's loan status. Overall, the shape of dot cluster looks like a square. The difference between high credit grade's square and lower's one is the square's location and shape. People with higher credit grade such as AA and A's shape is very tall and thin, which means they are very easy to borrow a lot of money in very low interest. and the situation begin to change from grade B. the right side of the square begin to move far right, and the top of square's dot are more and more sparse. This means for those people in lower credit grade, even their interest are higher, they cannot borrow too much money. For people with low and unknown credit record such as HR, NC and E, it is almost impossible to borrow money more than 10000.

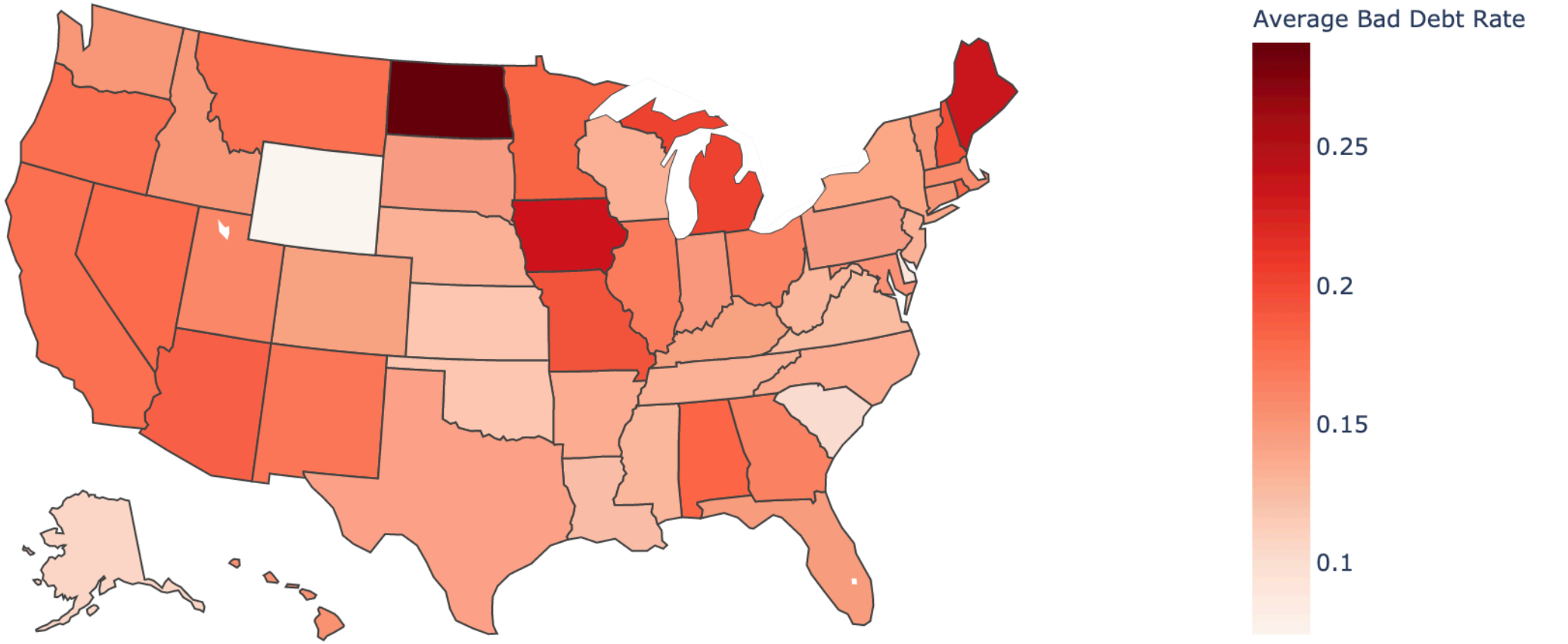

CreditGrade = C
CreditGrade = HR
CreditGrade = AA
CreditGrade = D
CreditGrade = B
CreditGrade = E
CreditGrade = A
CreditGrade = NC
LoanOriginalAmount
BorrowerRate
0
5000
10000
15000
20000
25000
0.0
0.1
0.2
0.3
0.4
0.5

> Compare with the The relationship between borrower rate and loan original amount, it is obvious that people with real estate are more convenient to borrow money from financial department. We can see people is getting harder and harder to borrow more than 10000 money if they have no real estate.


Comparison of home owner and no home owner's loan status.
LoanOriginalAmount
35000
30000
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5
BorrowerRate
IsBorrowerHomeowner
Yes
No

> Although east/west coast has the highest average loan, their bad debt rate is not very high when comparing with middle part of the United States. Another interesting result from North Dakota, Iowa and Maine‘s low average loan is, their average bad debt rate is obviously top 3 nationwide. On the countrary, Wyoming, which is the lowest bad debt rate, its average loan is higher than its neighborhood.

Average Bad Debt Rate in Different States



Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

> If a person has better credit rank, this means he or she can borrow money with lower interests. People with low interests have to loan less money in higher interest rates. Ownership of a property can also helps. People without property almost cannot borrow money more than 25000 in low interest rate.

Were there any interesting or surprising interactions between features?

> I was trying to find the relationship between geographic location and loan status. We can find some interesting result. East/west coast people loan more money than the rest of country, but their credit rank and average bad debt rate are not that low. And we can see some regular patterns on the map, such as some states have very serious bad debt situation and their average loan price is also lower than their neighborhood. We cannot say which one comes first, but we know they must have some relationship.