Part1 Report

Group 6

1. Data download and pre-processing

To programmatically download data, use RoboBrowser, to wrap and interact with form in website. Put account information in form and submit form to get cookie.

Send requests with cookie to download data from the website.

Using download data to create summary files for both origination and performance data.

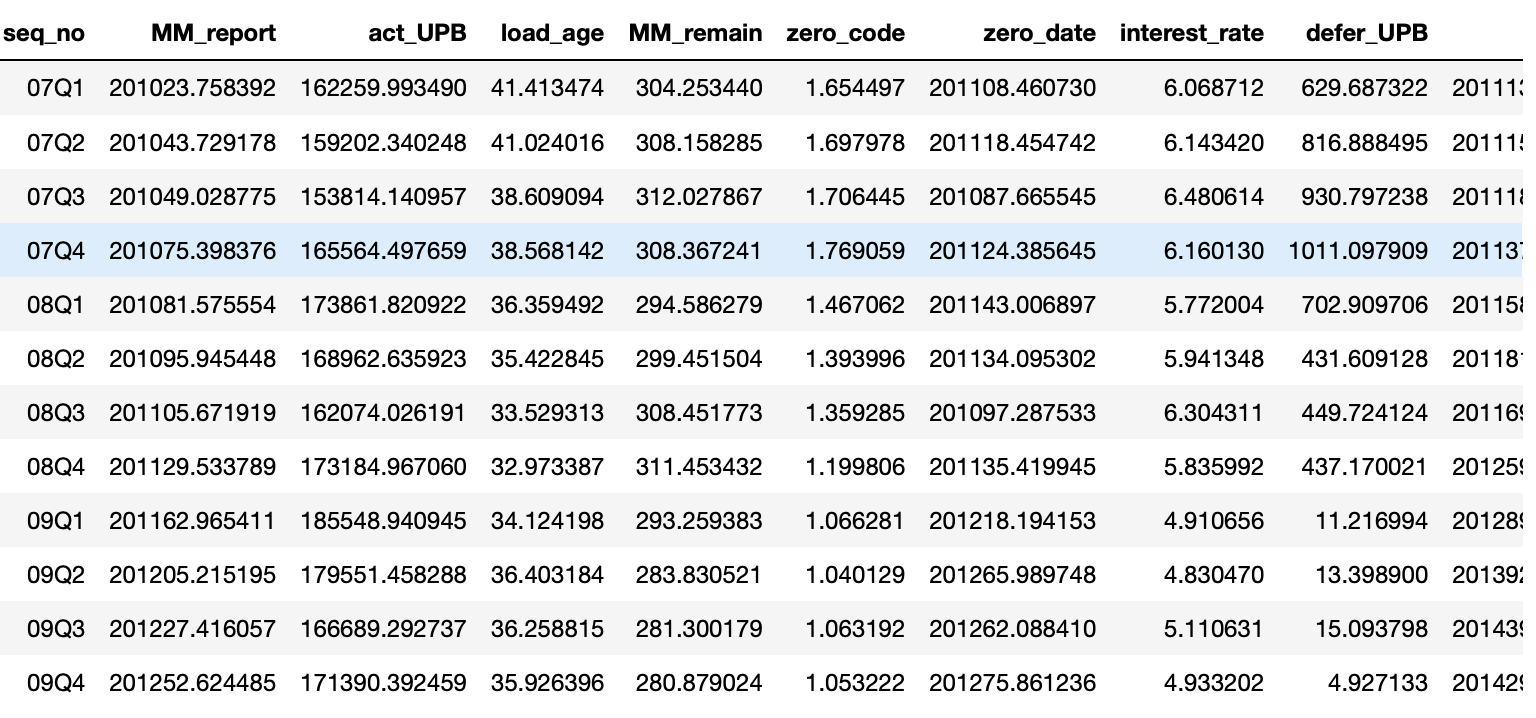
1. Exploratory Data analysis

* Numerical parameters

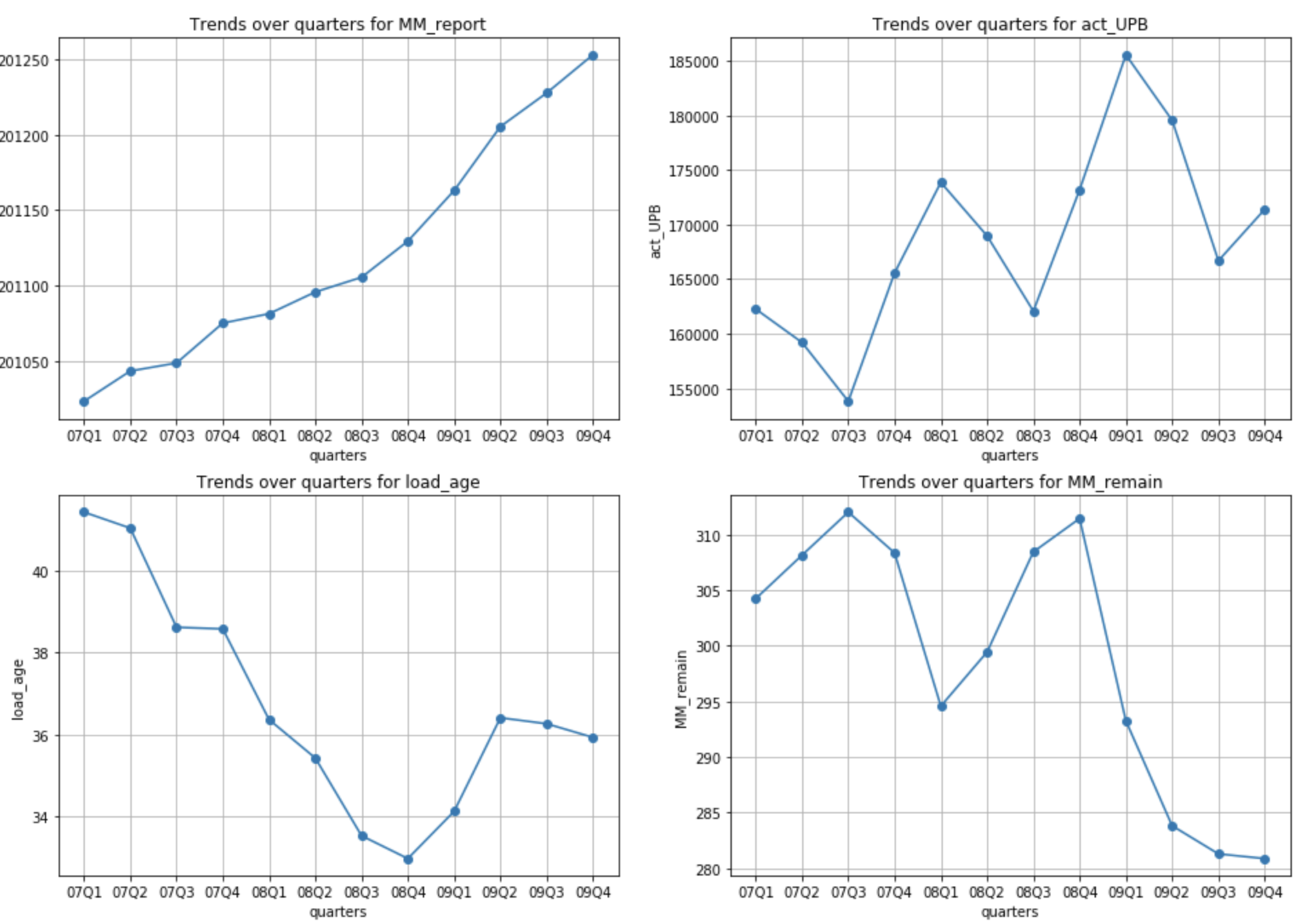
By using sequence number in both origination and performance data, divide each year 2007, 2008, 2009 data to four quarters and group them. Store numerical data summary files in summary directory.

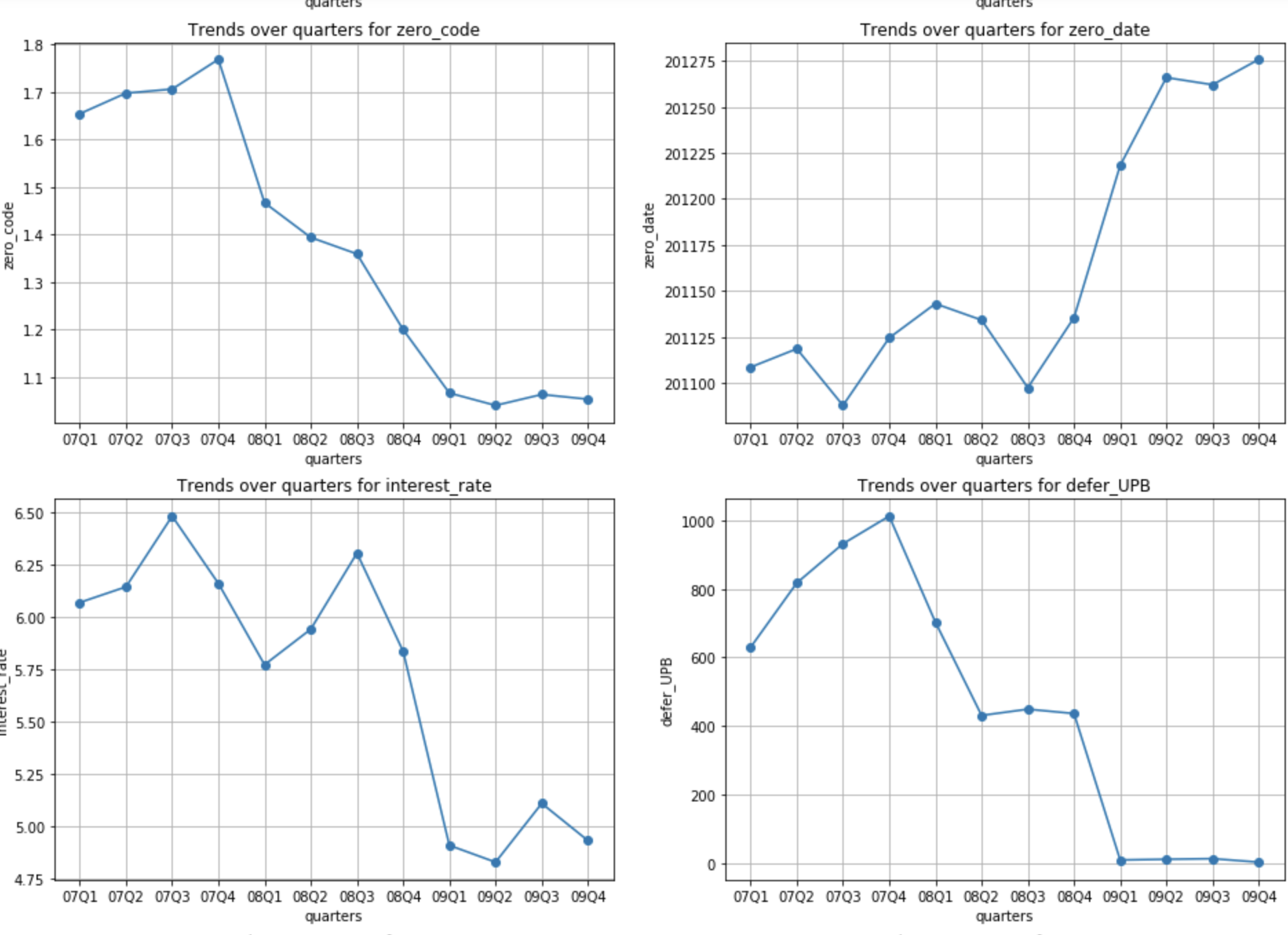
Use average amount to represent condition in each quarter.

**Performance summary**

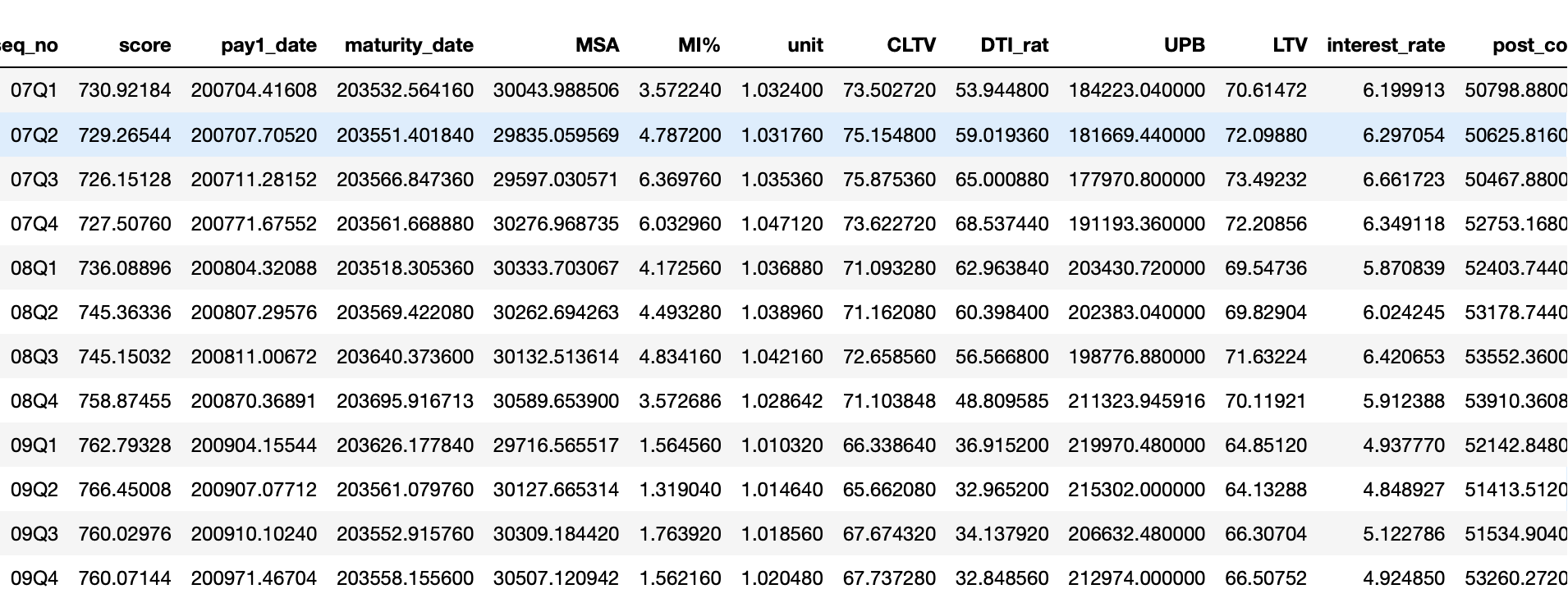


Visualizations are like that

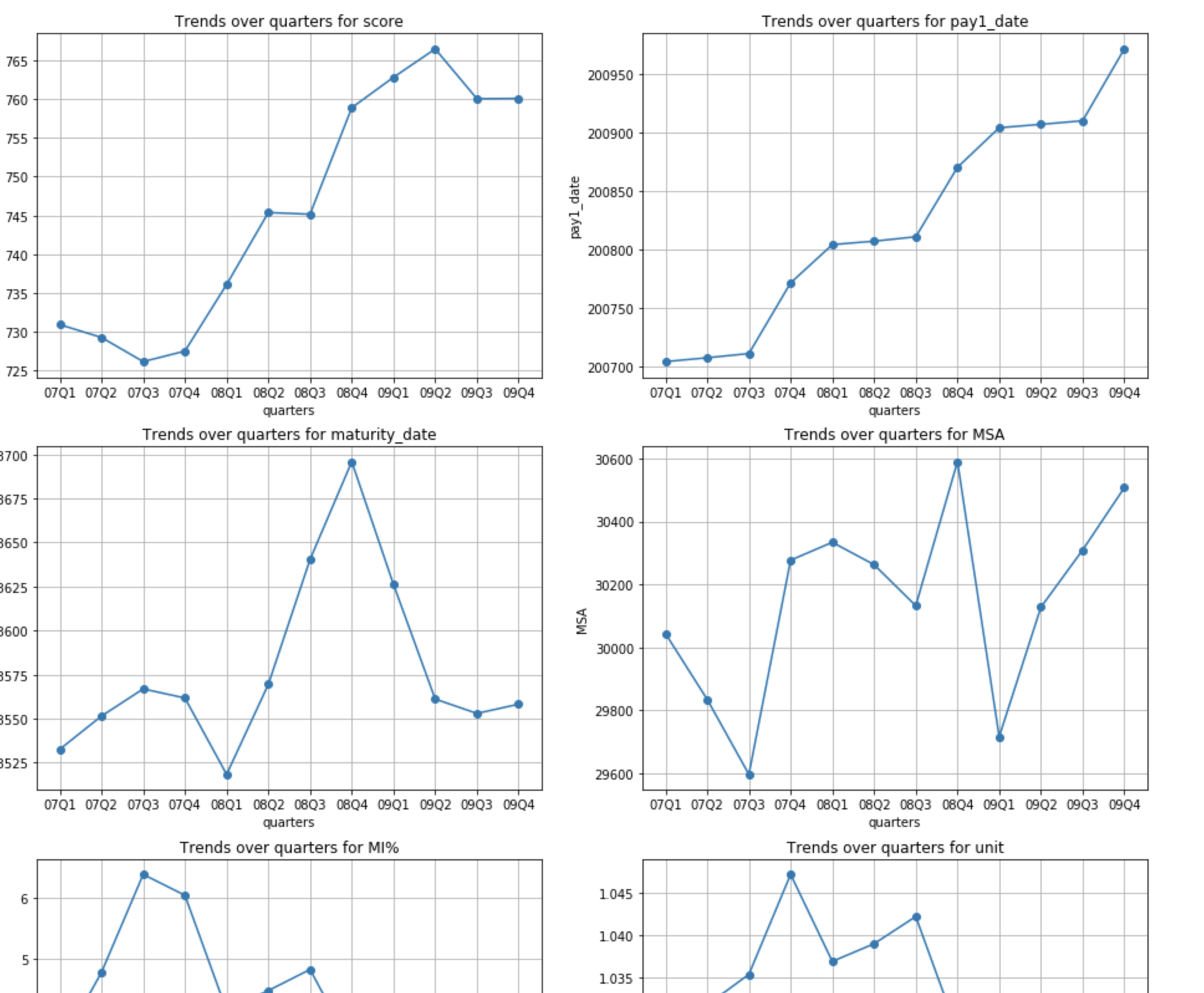




**Origination summary**



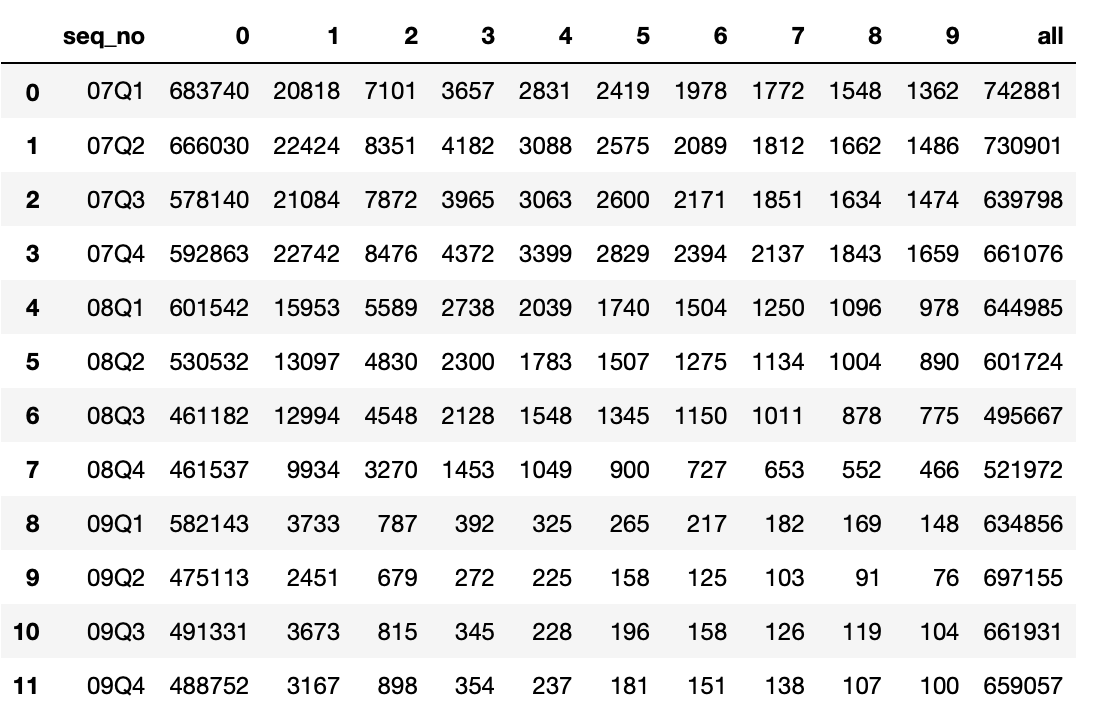
Visualization



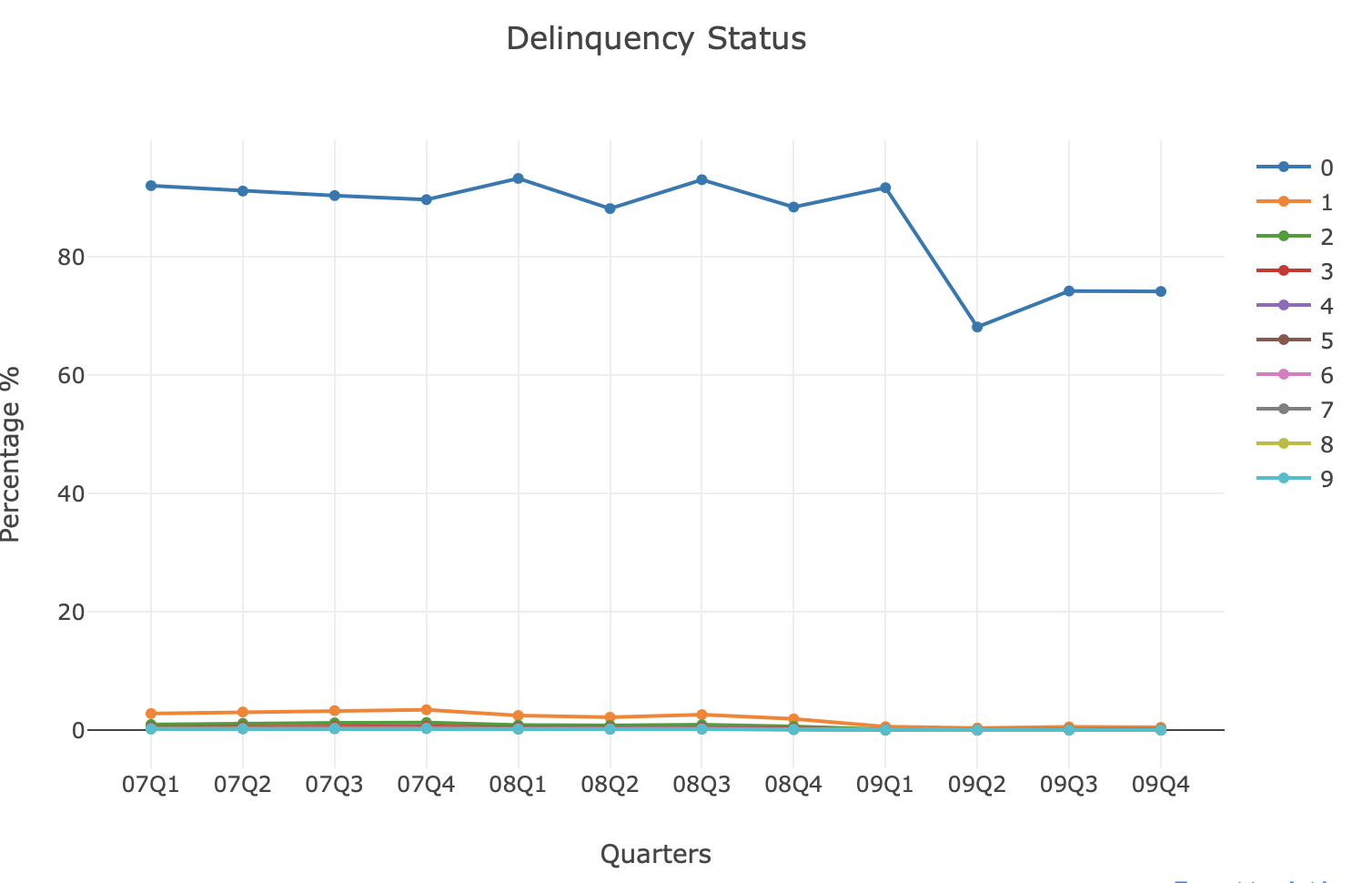
Through summary visualizations, it is easy to fine trend of each numerical through time series.

* Categorical parameters

**Delinquency Status**



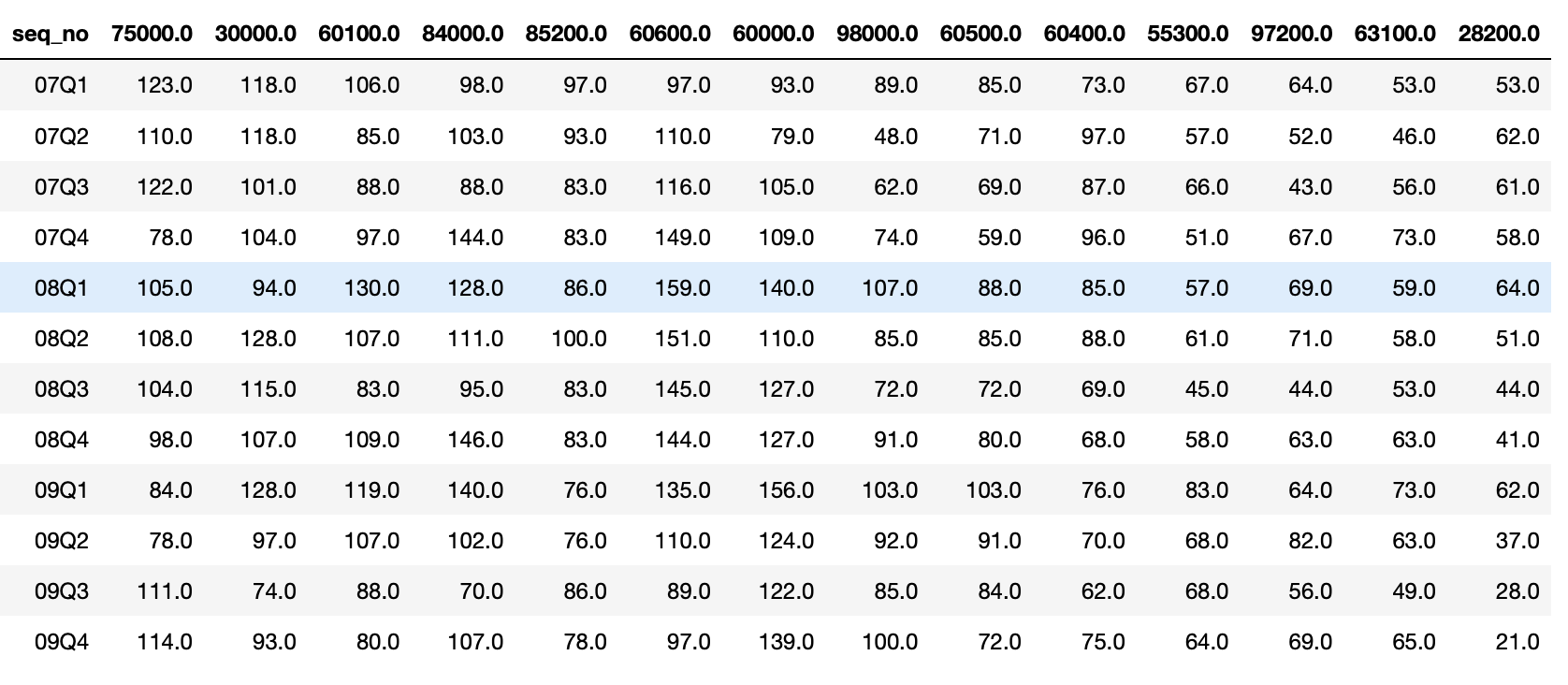
Use top 10 status to compare, and “all” columns is the total number of records in each quarter



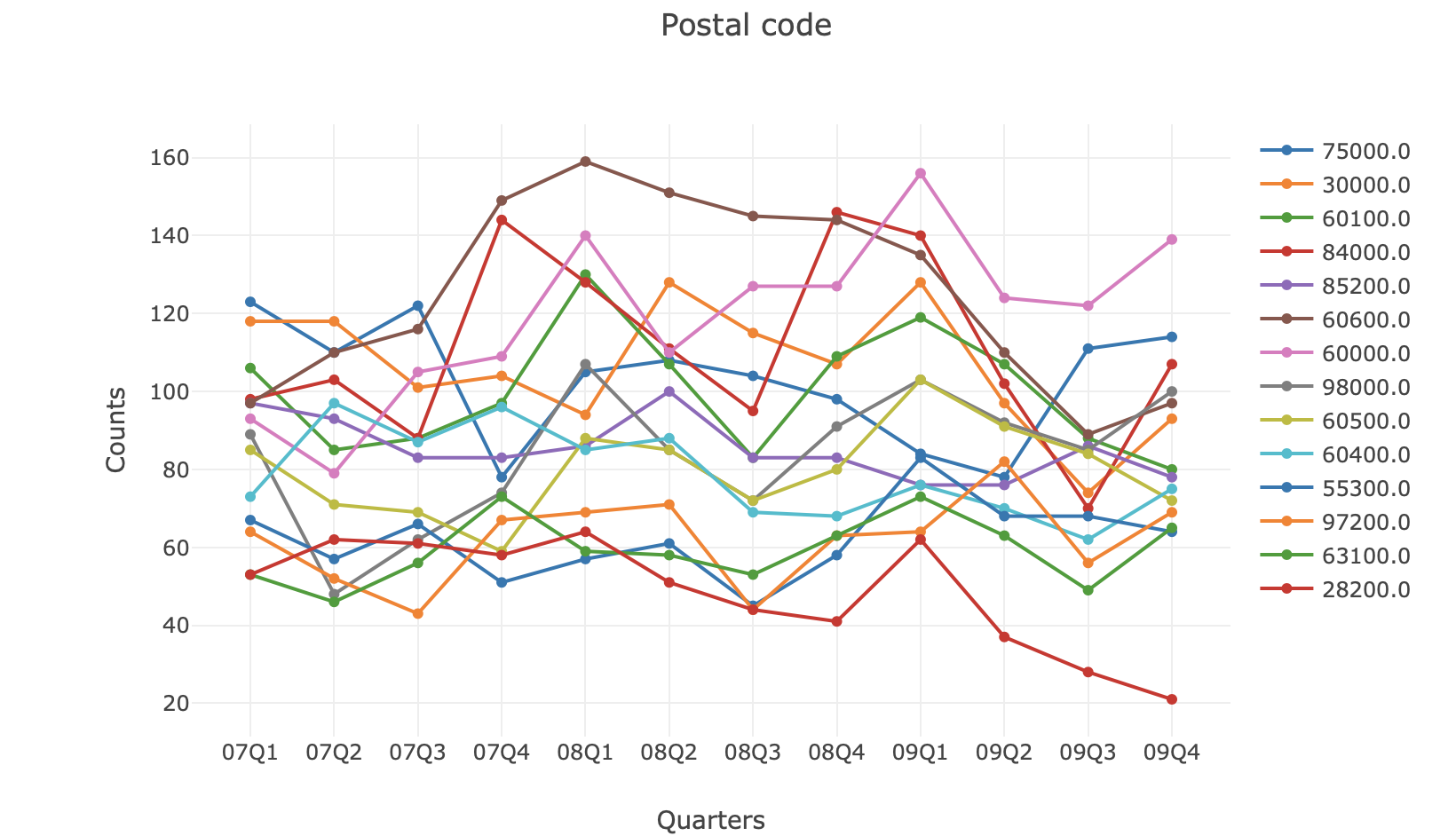
Visualize the percentage of each status through time series. Delinquency status 0, means current, is the most common status in the dataset, nearly involve all of the dataset. And through time series, all delinquency status percentage decrease a little bit.

**Postal Code**

Since there are lots of postal code, we choose top frequent postal code, and record the number of each code in different quarters.

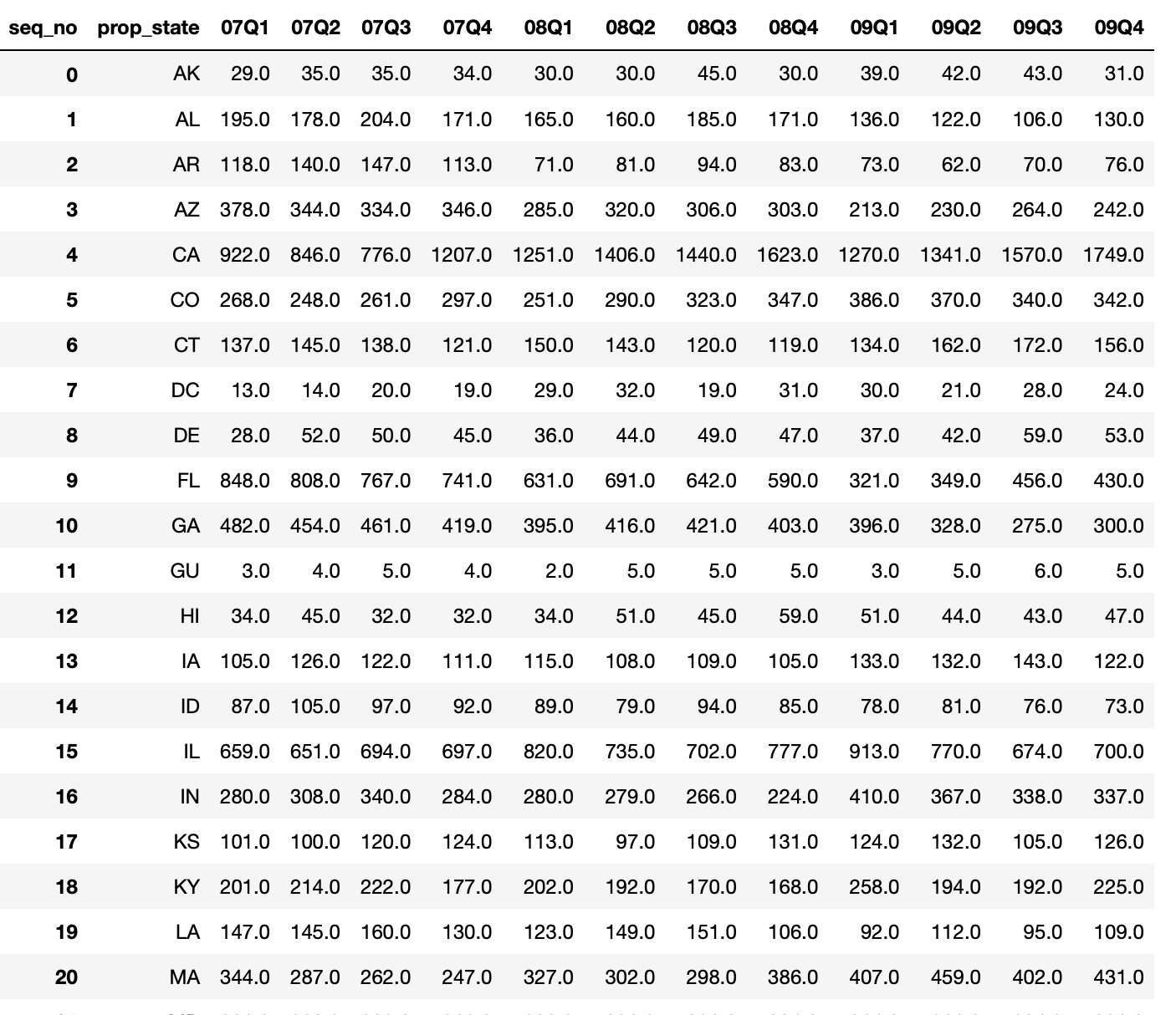


And the visualization is like that, from the graph, we can find the code numbers fluent but don’t show an apparent up or down trend through time series.

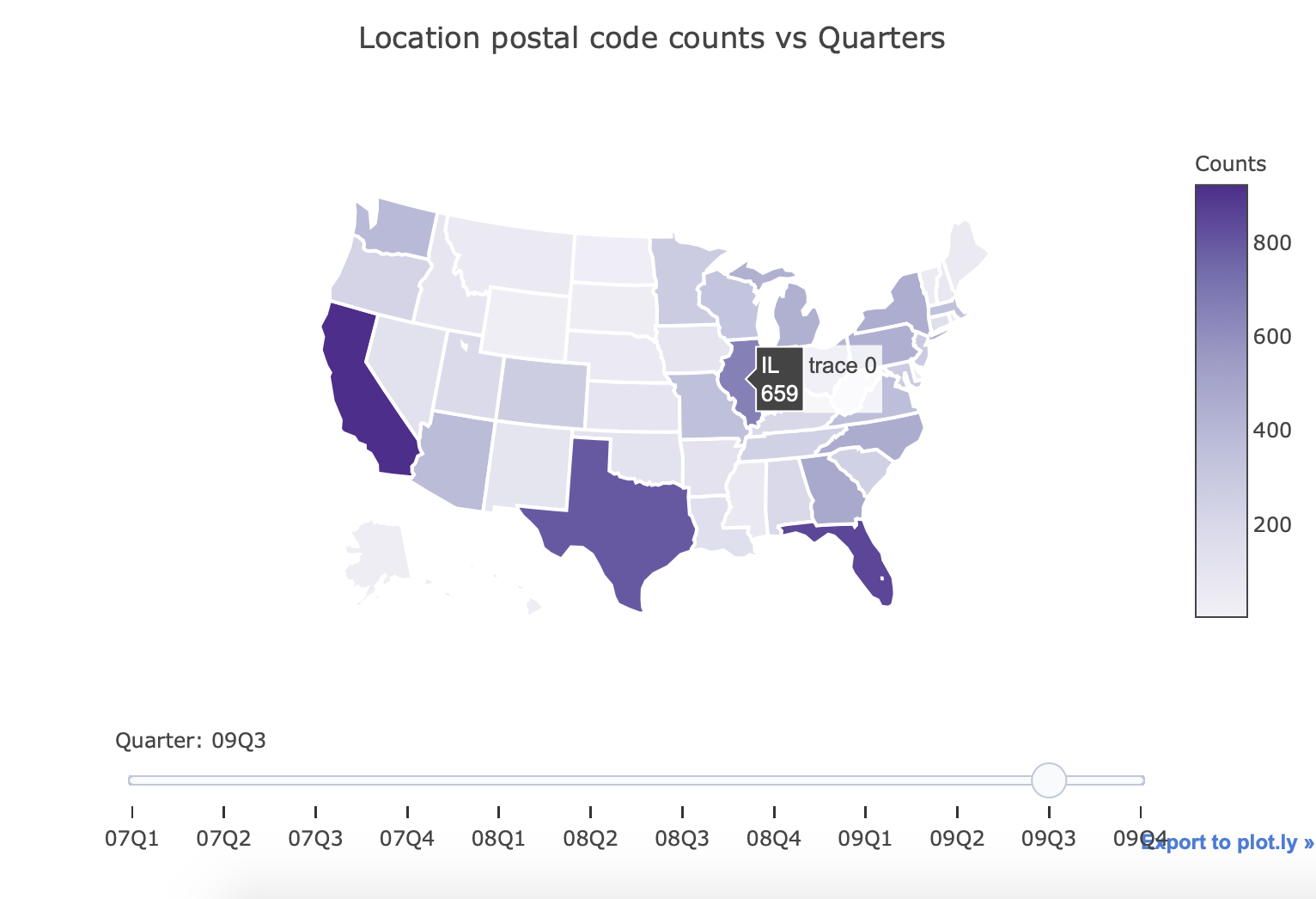


**Property state**

Same with previous, group by quarter and state code, count number of loans.



Visualization is like that, use plotly choropleth graph, to show data on USA map.

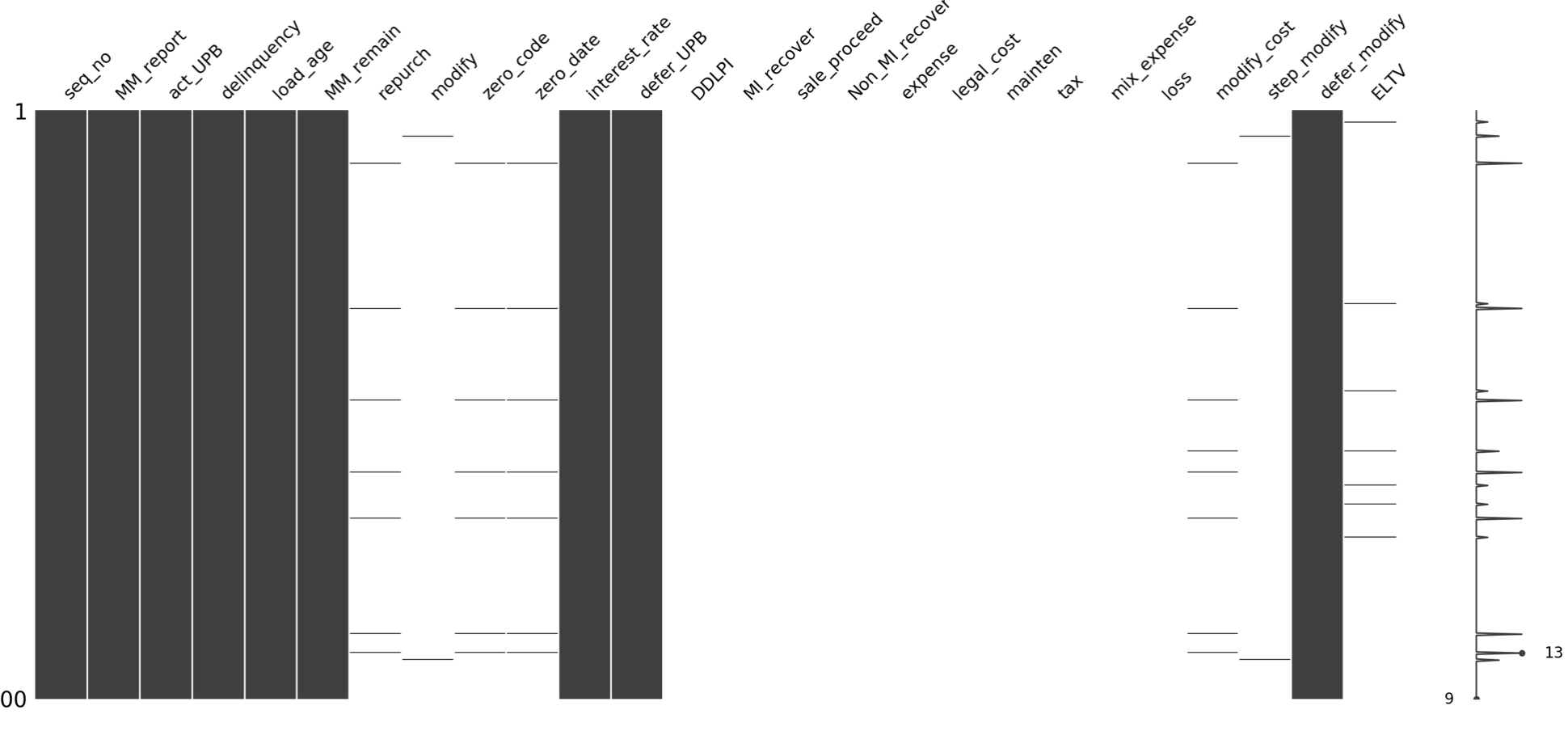


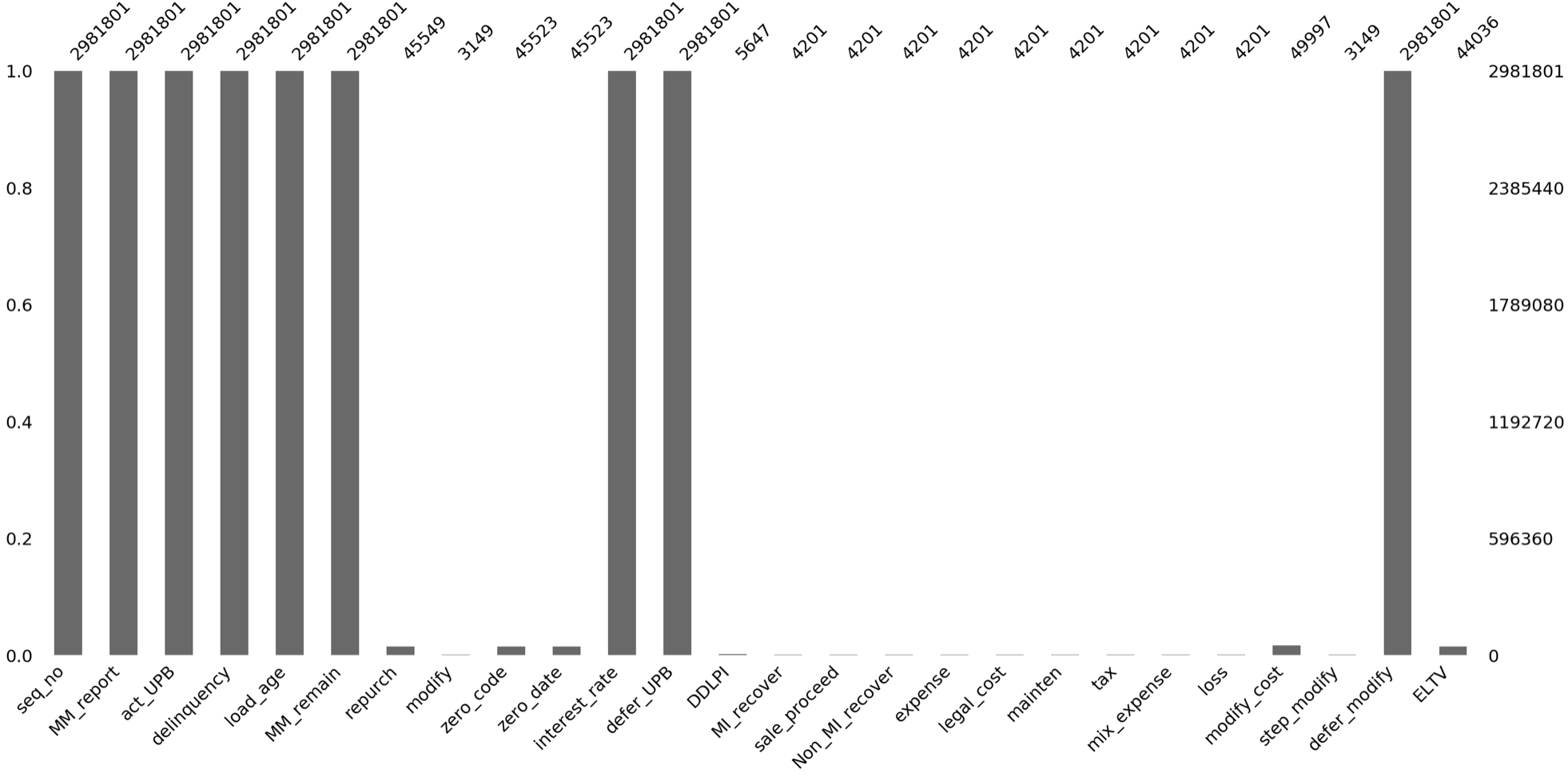
* Single year comparison

Instead of compare between multiple years, find change in each year is also important.

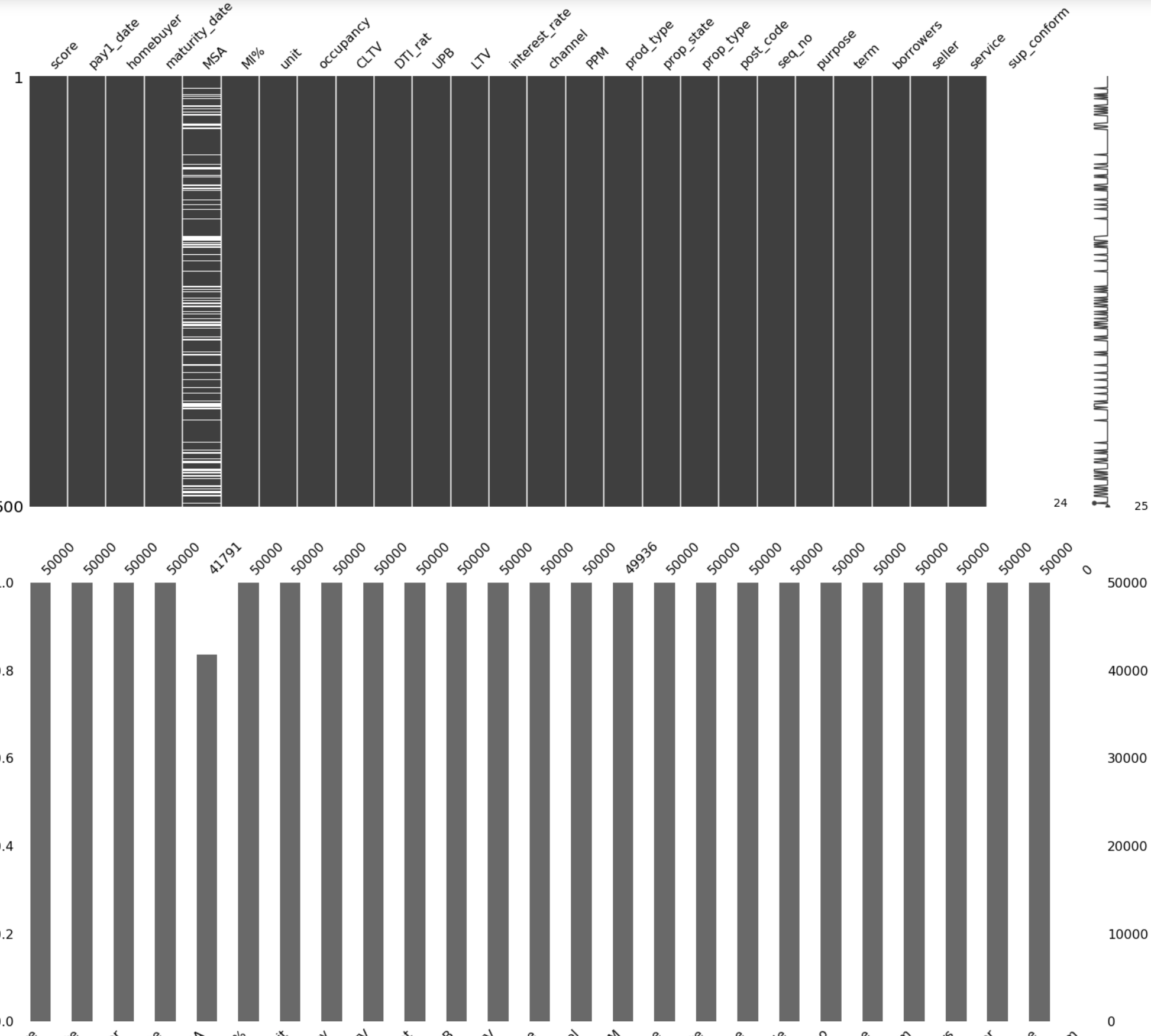
Missing value visualization

**Performance data**

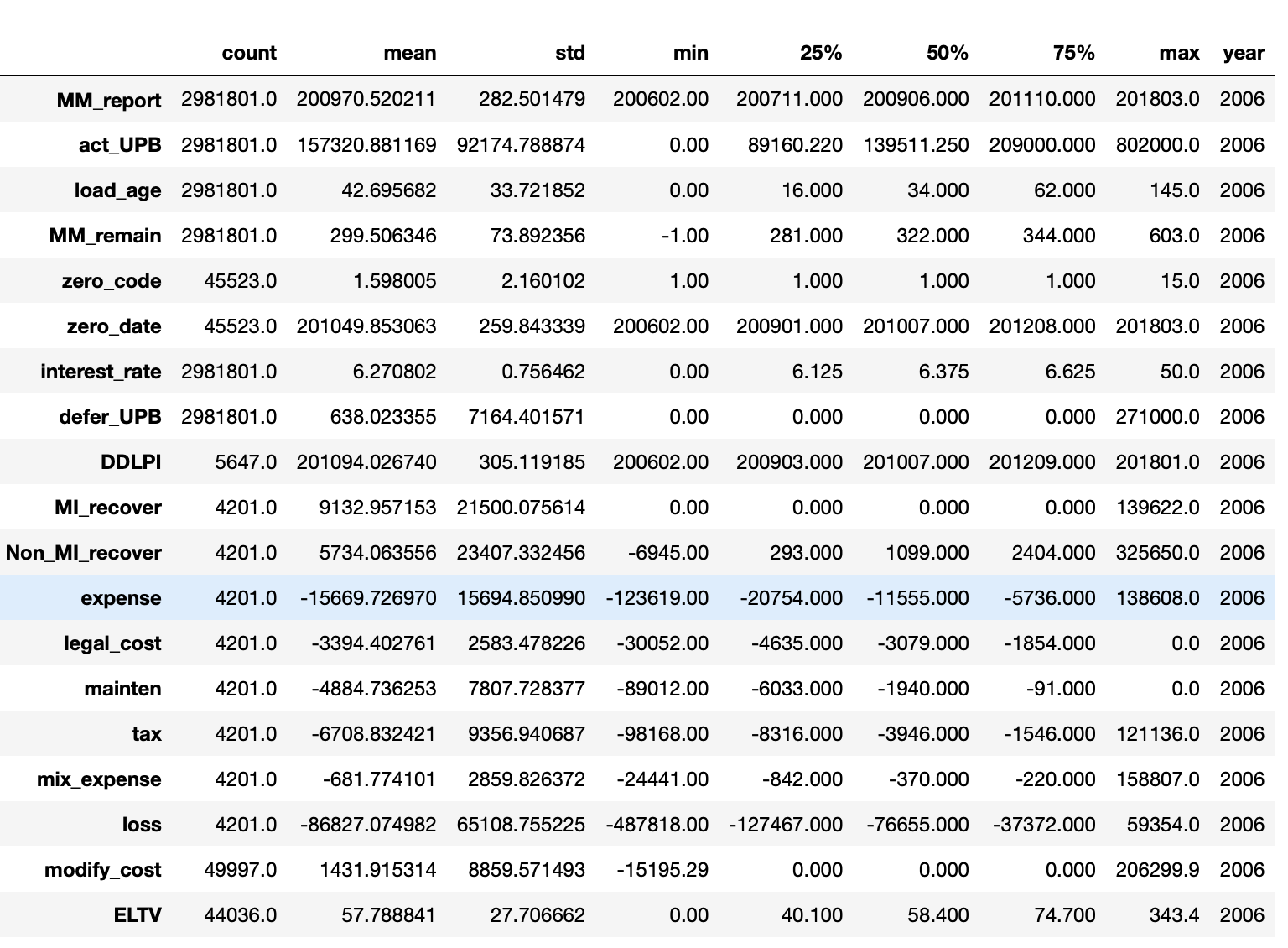




**Origination data**

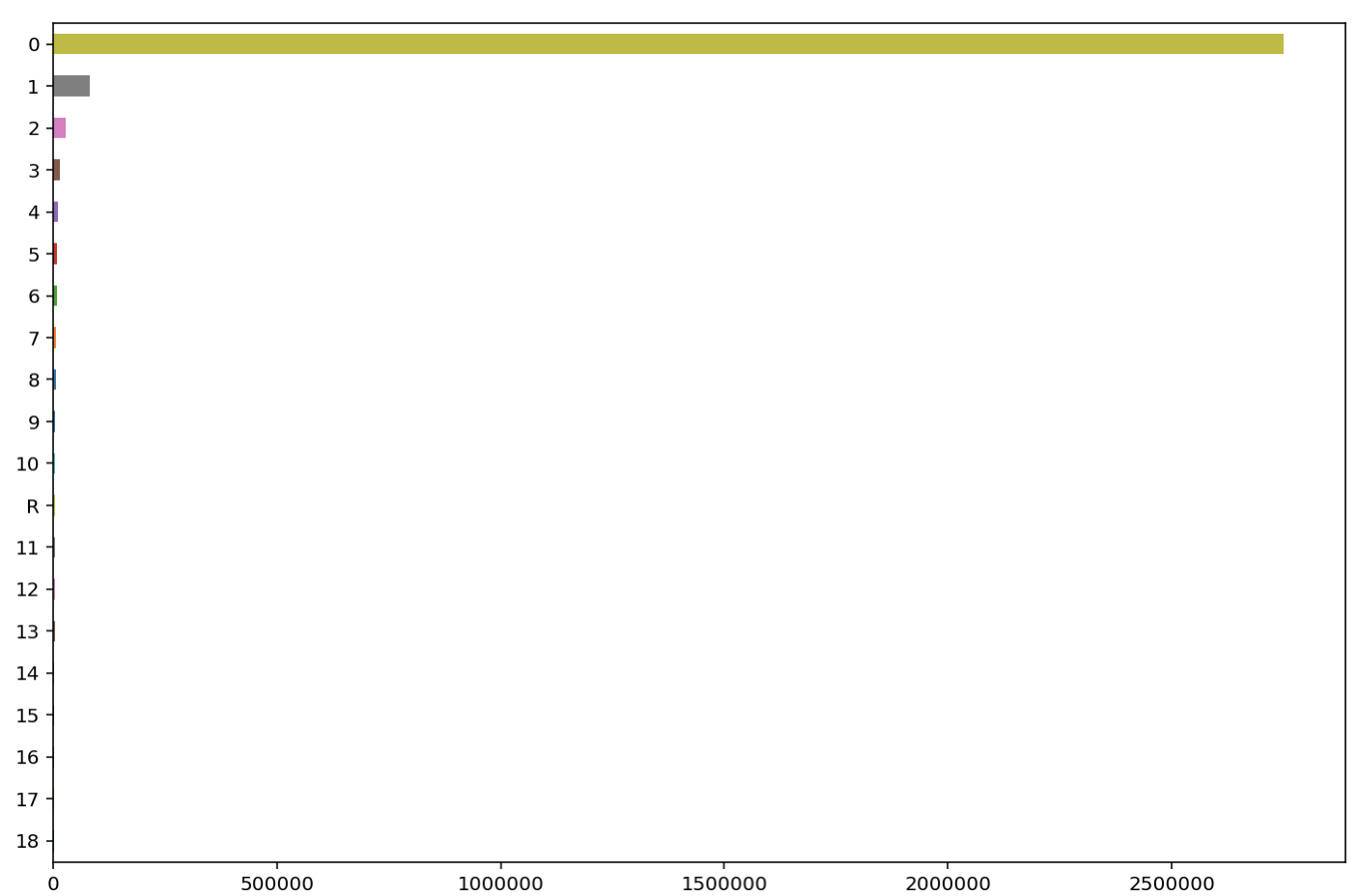
****

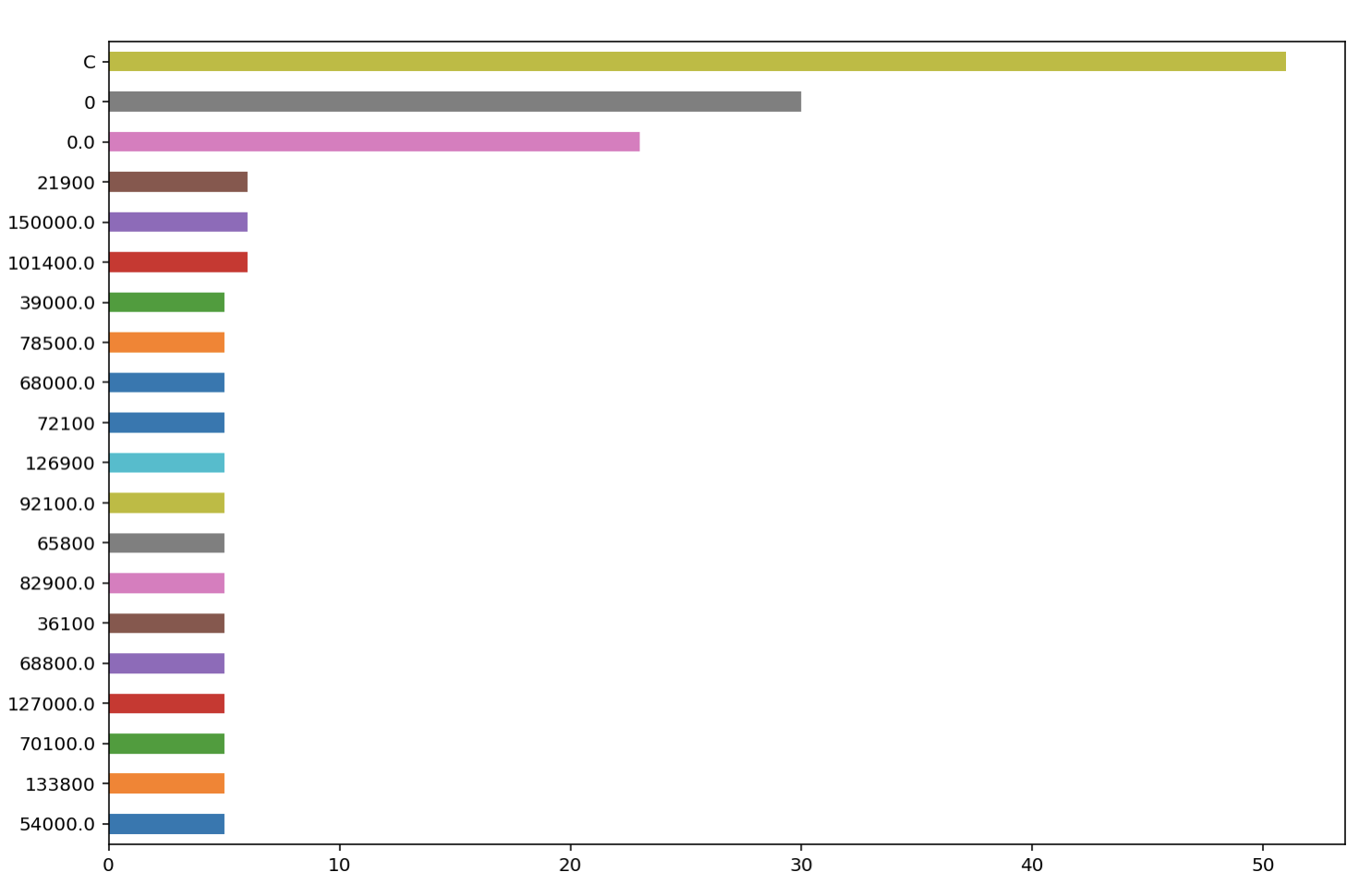
**Numerical Parameters**

****

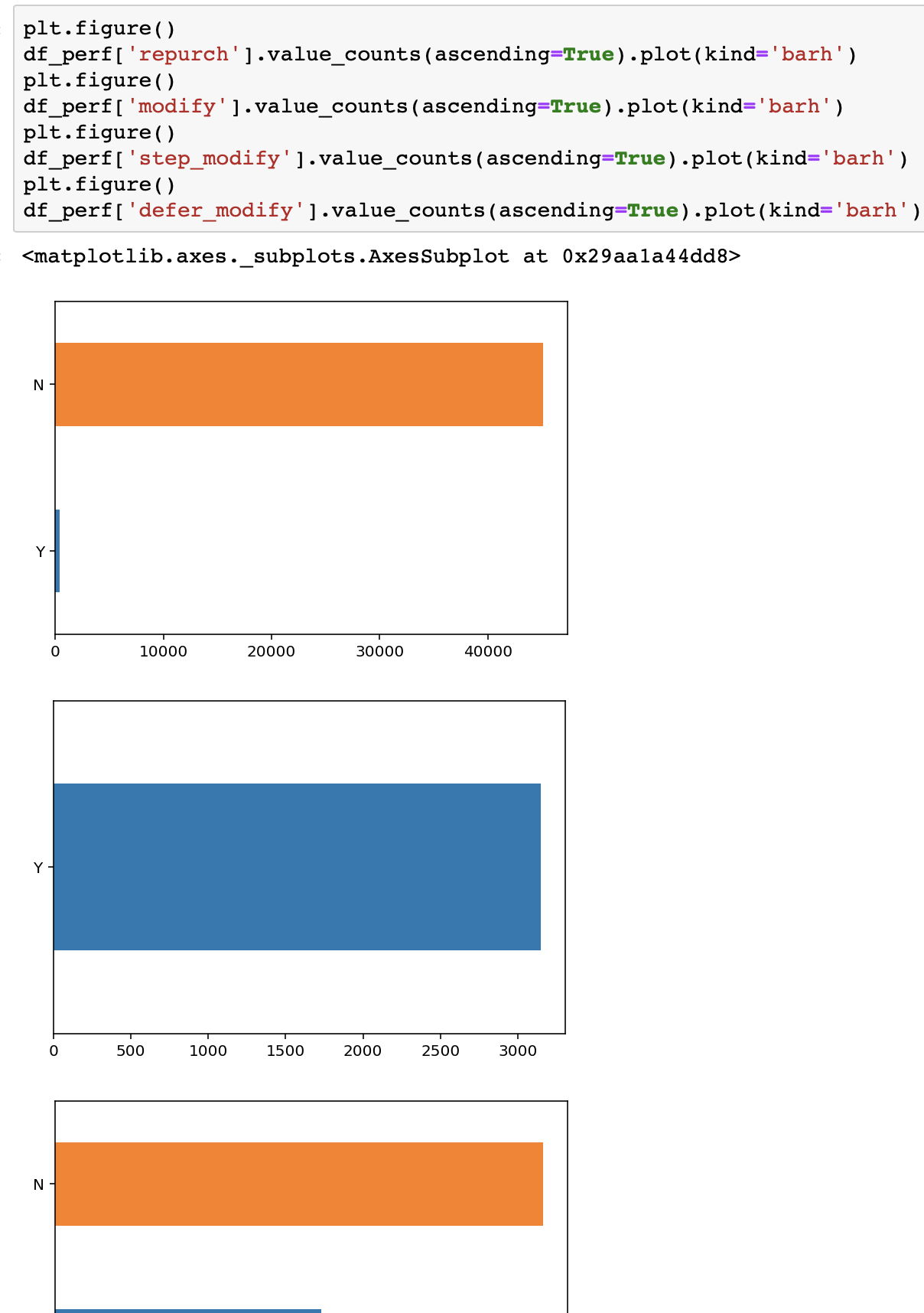
**Categorical parameters**

Delinquency status distribution in each year

****

Sale proceeddistribution in each year****

And many other categorical parameters.

****