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.

2. Exploratory Data analysis

- Numerical parameters

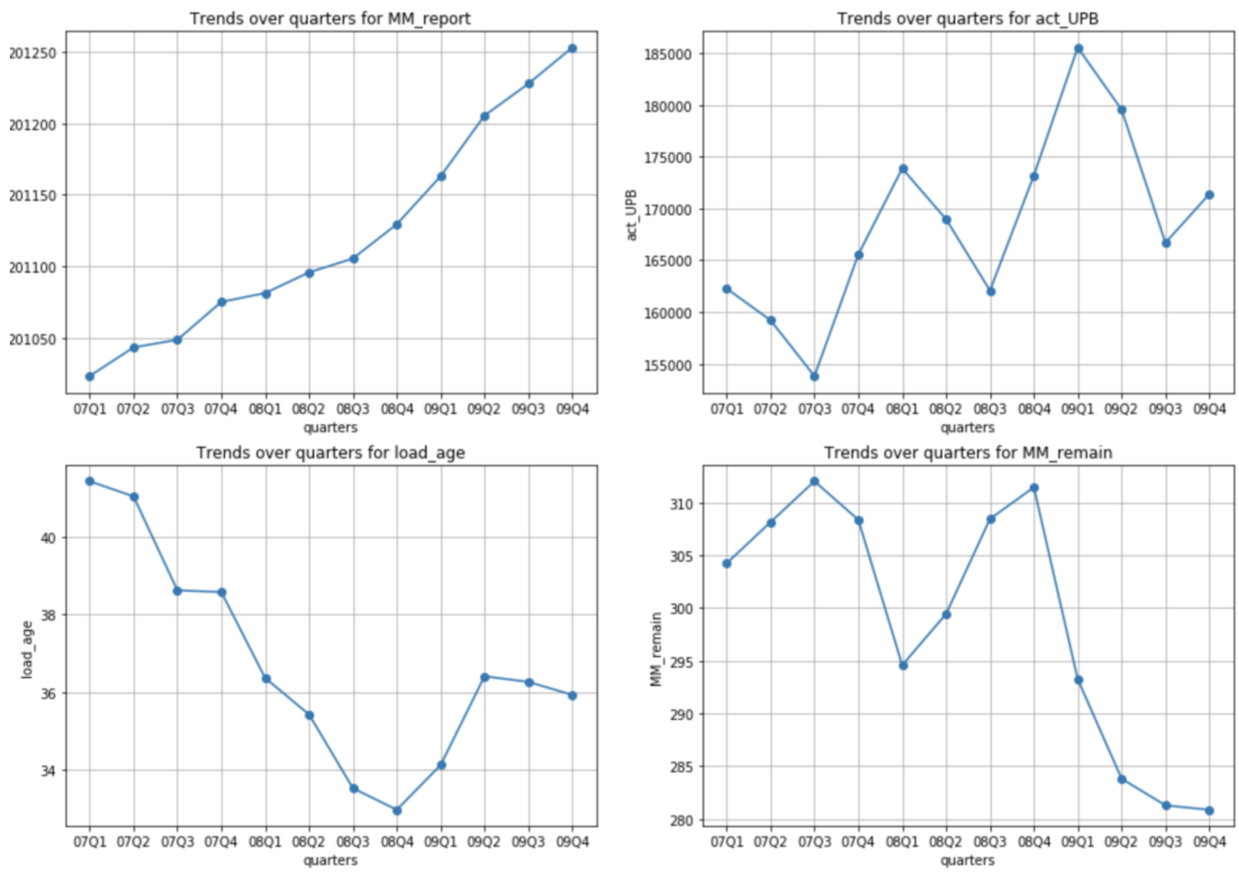
By using sequence number in both origination and performance data, divide each year 2007, 2008, 2009 data to four quarters and group them. Store numerical data summary files in summary directory.

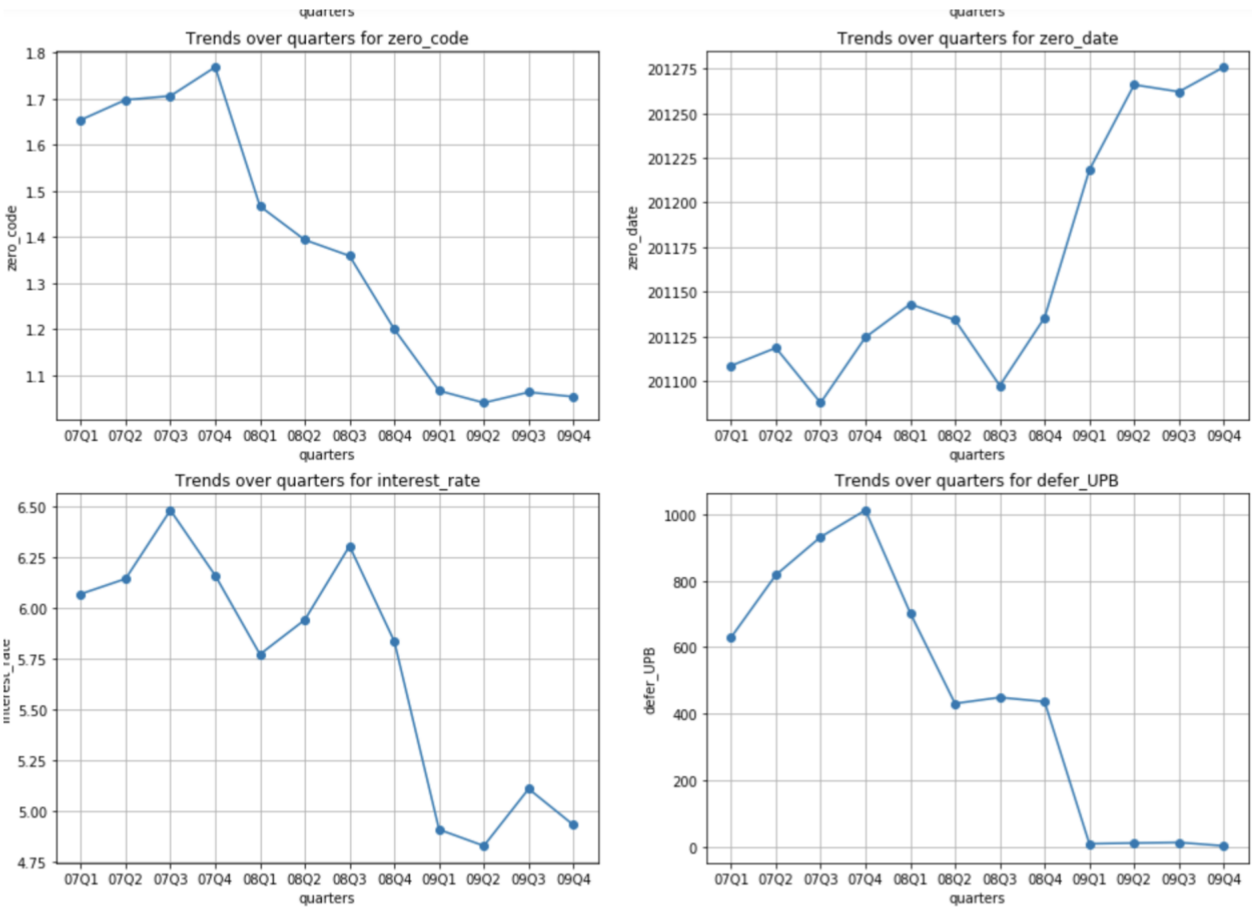
Use average amount to represent condition in each quarter.

Performance summary

seq_no	MM_report	act_UPB	load_age	MM_remain	zero_code	zero_date	interest_rate	defer_UPB	
07Q1	201023.758392	162259.993490	41.413474	304.253440	1.654497	201108.460730	6.068712	629.687322	20111
07Q2	201043.729178	159202.340248	41.024016	308.158285	1.697978	201118.454742	6.143420	816.888495	20111
07Q3	201049.028775	153814.140957	38.609094	312.027867	1.706445	201087.665545	6.480614	930.797238	20111
07Q4	201075.398376	165564.497659	38.568142	308.367241	1.769059	201124.385645	6.160130	1011.097909	20113
08Q1	201081.575554	173861.820922	36.359492	294.586279	1.467062	201143.006897	5.772004	702.909706	20115
08Q2	201095.945448	168962.635923	35.422845	299.451504	1.393996	201134.095302	5.941348	431.609128	20118
08Q3	201105.671919	162074.026191	33.529313	308.451773	1.359285	201097.287533	6.304311	449.724124	20116
08Q4	201129.533789	173184.967060	32.973387	311.453432	1.199806	201135.419945	5.835992	437.170021	20125
09Q1	201162.965411	185548.940945	34.124198	293.259383	1.066281	201218.194153	4.910656	11.216994	20128
09Q2	201205.215195	179551.458288	36.403184	283.830521	1.040129	201265.989748	4.830470	13.398900	20139
09Q3	201227.416057	166689.292737	36.258815	281.300179	1.063192	201262.088410	5.110631	15.093798	20143
09Q4	201252.624485	171390.392459	35.926396	280.879024	1.053222	201275.861236	4.933202	4.927133	20142

Visualizations are like that

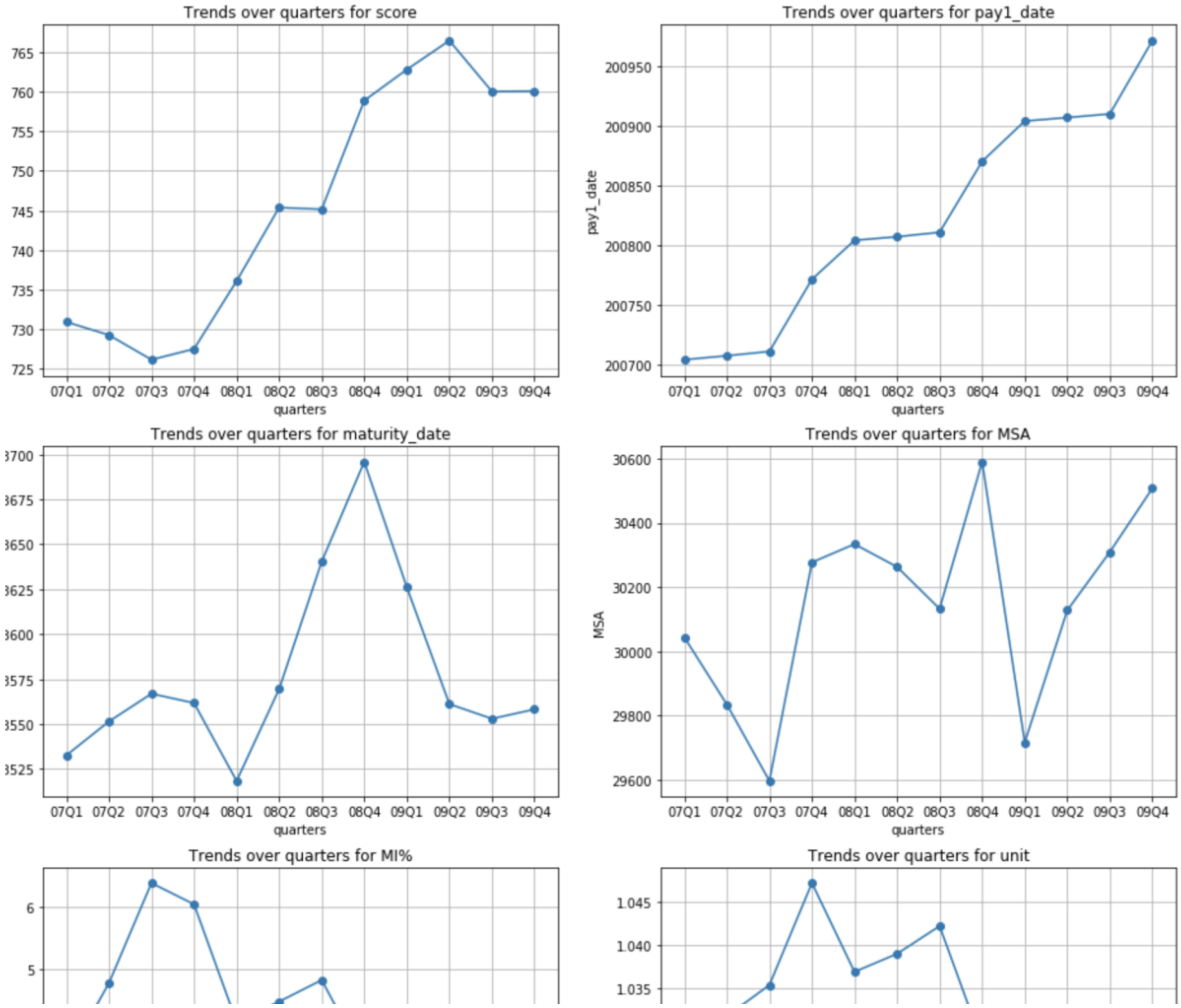

Trends over quarters for MM_report
Trends over quarters for act_UPB
act_UPB
Trends over quarters for load_age
load_age
Trends over quarters for MM_remain
MM_remain
quarters
07Q1 07Q2 07Q3 07Q4 08Q1 08Q2 08Q3 08Q4 09Q1 09Q2 09Q3 09Q4



Origination summary

eq_no	score	pay1_date	maturity_date	MSA	MI%	unit	CLTV	DTI_rat	UPB	LTV	interest_rate	post_co
07Q1	730.92184	200704.41608	203532.564160	30043.988506	3.572240	1.032400	73.502720	53.944800	184223.040000	70.61472	6.199913	50798.8800
07Q2	729.26544	200707.70520	203551.401840	29835.059569	4.787200	1.031760	75.154800	59.019360	181669.440000	72.09880	6.297054	50625.8160
07Q3	726.15128	200711.28152	203566.847360	29597.030571	6.369760	1.035360	75.875360	65.000880	177970.800000	73.49232	6.661723	50467.8800
07Q4	727.50760	200771.67552	203561.668880	30276.968735	6.032960	1.047120	73.622720	68.537440	191193.360000	72.20856	6.349118	52753.1680
08Q1	736.08896	200804.32088	203518.305360	30333.703067	4.172560	1.036880	71.093280	62.963840	203430.720000	69.54736	5.870839	52403.7440
08Q2	745.36336	200807.29576	203569.422080	30262.694263	4.493280	1.038960	71.162080	60.398400	202383.040000	69.82904	6.024245	53178.7440
08Q3	745.15032	200811.00672	203640.373600	30132.513614	4.834160	1.042160	72.658560	56.566800	198776.880000	71.63224	6.420653	53552.3600
08Q4	758.87455	200870.36891	203695.916713	30589.653900	3.572686	1.028642	71.103848	48.809585	211323.945916	70.11921	5.912388	53910.3608
09Q1	762.79328	200904.15544	203626.177840	29716.565517	1.564560	1.010320	66.338640	36.915200	219970.480000	64.85120	4.937770	52142.8480
09Q2	766.45008	200907.07712	203561.079760	30127.665314	1.319040	1.014640	65.662080	32.965200	215302.000000	64.13288	4.848927	51413.5120
09Q3	760.02976	200910.10240	203552.915760	30309.184420	1.763920	1.018560	67.674320	34.137920	206632.480000	66.30704	5.122786	51534.9040
09Q4	760.07144	200971.46704	203558.155600	30507.120942	1.562160	1.020480	67.737280	32.848560	212974.000000	66.50752	4.924850	53260.2720

Visualization



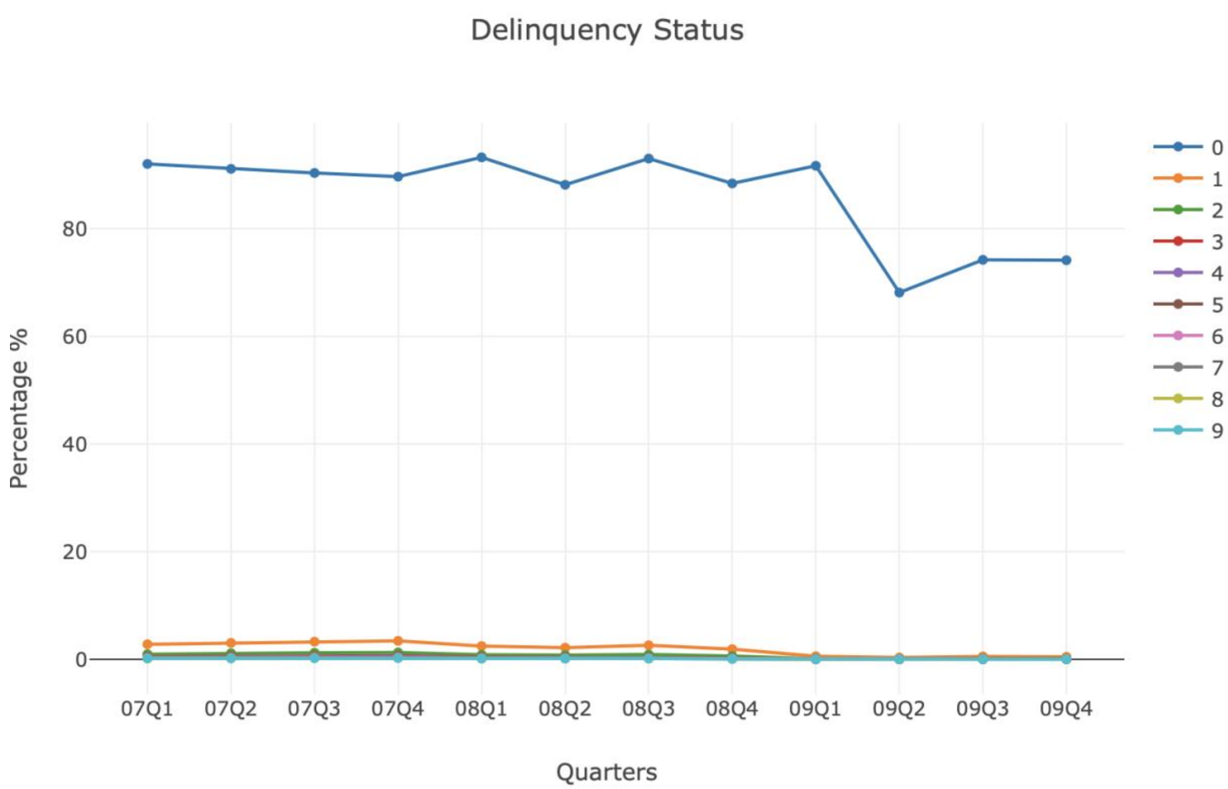
Through summary visualizations, it is easy to fine trend of each numerical through time series.

- Categorical parameters

Delinquency Status

	seq_no	0	1	2	3	4	5	6	7	8	9	all
0	07Q1	683740	20818	7101	3657	2831	2419	1978	1772	1548	1362	742881
1	07Q2	666030	22424	8351	4182	3088	2575	2089	1812	1662	1486	730901
2	07Q3	578140	21084	7872	3965	3063	2600	2171	1851	1634	1474	639798
3	07Q4	592863	22742	8476	4372	3399	2829	2394	2137	1843	1659	661076
4	08Q1	601542	15953	5589	2738	2039	1740	1504	1250	1096	978	644985
5	08Q2	530532	13097	4830	2300	1783	1507	1275	1134	1004	890	601724
6	08Q3	461182	12994	4548	2128	1548	1345	1150	1011	878	775	495667
7	08Q4	461537	9934	3270	1453	1049	900	727	653	552	466	521972
8	09Q1	582143	3733	787	392	325	265	217	182	169	148	634856
9	09Q2	475113	2451	679	272	225	158	125	103	91	76	697155
10	09Q3	491331	3673	815	345	228	196	158	126	119	104	661931
11	09Q4	488752	3167	898	354	237	181	151	138	107	100	659057

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



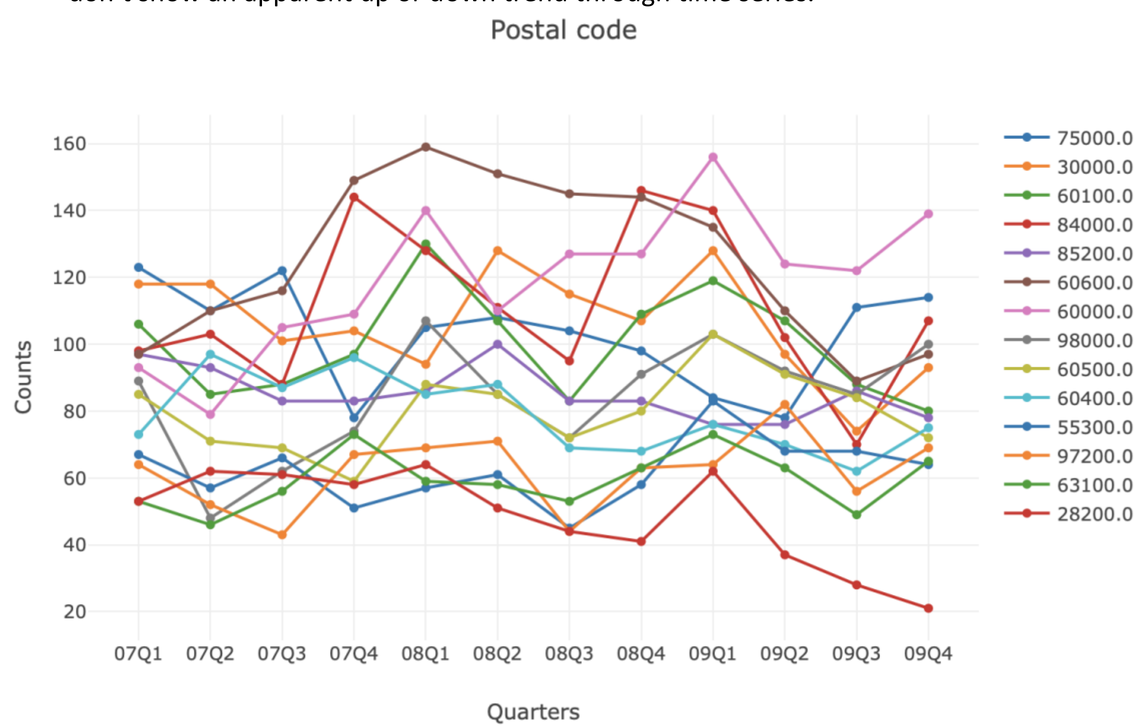
Visualize the percentage of each status through time series. Delinquency status 0, means current, is the most common status in the dataset, nearly involve all of the dataset. And through time series, all delinquency status percentage decrease a little bit.

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.

seq_no	75000.0	30000.0	60100.0	84000.0	85200.0	60600.0	60000.0	98000.0	60500.0	60400.0	55300.0	97200.0	63100.0	28200.0
07Q1	123.0	118.0	106.0	98.0	97.0	97.0	93.0	89.0	85.0	73.0	67.0	64.0	53.0	53.0
07Q2	110.0	118.0	85.0	103.0	93.0	110.0	79.0	48.0	71.0	97.0	57.0	52.0	46.0	62.0
07Q3	122.0	101.0	88.0	88.0	83.0	116.0	105.0	62.0	69.0	87.0	66.0	43.0	56.0	61.0
07Q4	78.0	104.0	97.0	144.0	83.0	149.0	109.0	74.0	59.0	96.0	51.0	67.0	73.0	58.0
08Q1	105.0	94.0	130.0	128.0	86.0	159.0	140.0	107.0	88.0	85.0	57.0	69.0	59.0	64.0
08Q2	108.0	128.0	107.0	111.0	100.0	151.0	110.0	85.0	85.0	88.0	61.0	71.0	58.0	51.0
08Q3	104.0	115.0	83.0	95.0	83.0	145.0	127.0	72.0	72.0	69.0	45.0	44.0	53.0	44.0
08Q4	98.0	107.0	109.0	146.0	83.0	144.0	127.0	91.0	80.0	68.0	58.0	63.0	63.0	41.0
09Q1	84.0	128.0	119.0	140.0	76.0	135.0	156.0	103.0	103.0	76.0	83.0	64.0	73.0	62.0
09Q2	78.0	97.0	107.0	102.0	76.0	110.0	124.0	92.0	91.0	70.0	68.0	82.0	63.0	37.0
09Q3	111.0	74.0	88.0	70.0	86.0	89.0	122.0	85.0	84.0	62.0	68.0	56.0	49.0	28.0
09Q4	114.0	93.0	80.0	107.0	78.0	97.0	139.0	100.0	72.0	75.0	64.0	69.0	65.0	21.0

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.

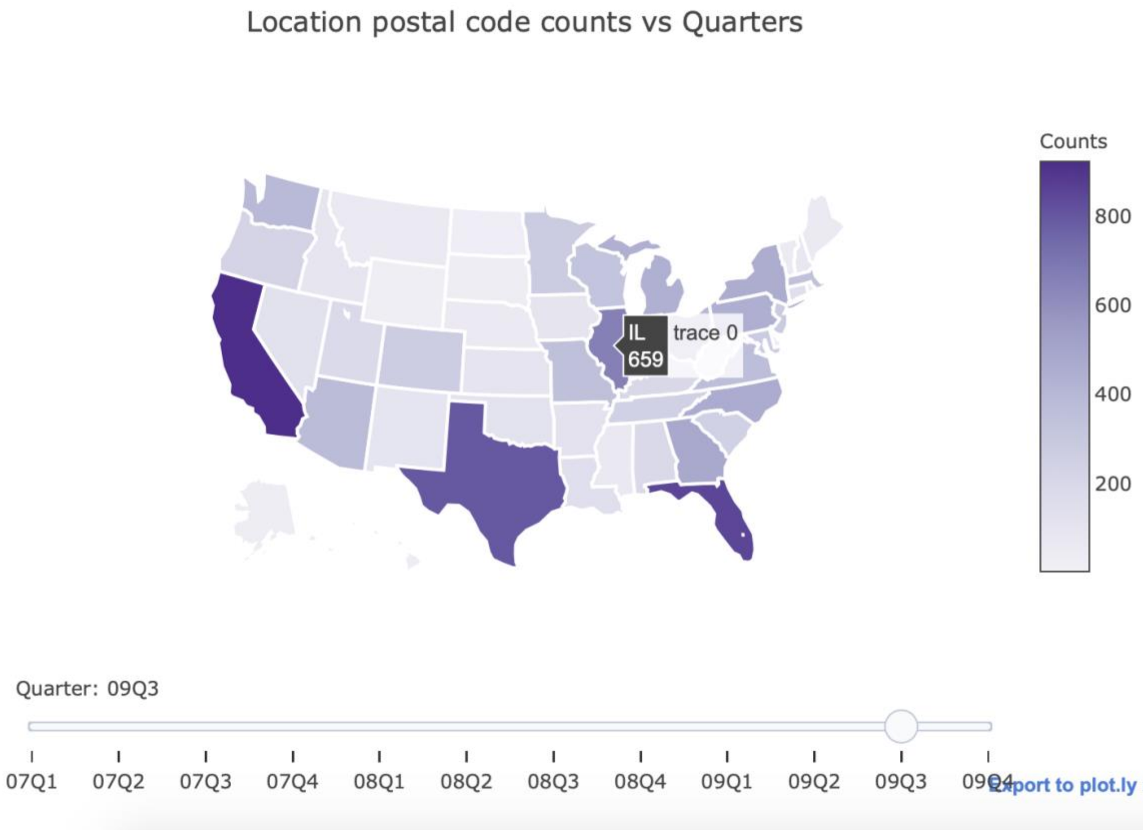


Property state

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

seq_no	prop_state	07Q1	07Q2	07Q3	07Q4	08Q1	08Q2	08Q3	08Q4	09Q1	09Q2	09Q3	09Q4
0	AK	29.0	35.0	35.0	34.0	30.0	30.0	45.0	30.0	39.0	42.0	43.0	31.0
1	AL	195.0	178.0	204.0	171.0	165.0	160.0	185.0	171.0	136.0	122.0	106.0	130.0
2	AR	118.0	140.0	147.0	113.0	71.0	81.0	94.0	83.0	73.0	62.0	70.0	76.0
3	AZ	378.0	344.0	334.0	346.0	285.0	320.0	306.0	303.0	213.0	230.0	264.0	242.0
4	CA	922.0	846.0	776.0	1207.0	1251.0	1406.0	1440.0	1623.0	1270.0	1341.0	1570.0	1749.0
5	CO	268.0	248.0	261.0	297.0	251.0	290.0	323.0	347.0	386.0	370.0	340.0	342.0
6	CT	137.0	145.0	138.0	121.0	150.0	143.0	120.0	119.0	134.0	162.0	172.0	156.0
7	DC	13.0	14.0	20.0	19.0	29.0	32.0	19.0	31.0	30.0	21.0	28.0	24.0
8	DE	28.0	52.0	50.0	45.0	36.0	44.0	49.0	47.0	37.0	42.0	59.0	53.0
9	FL	848.0	808.0	767.0	741.0	631.0	691.0	642.0	590.0	321.0	349.0	456.0	430.0
10	GA	482.0	454.0	461.0	419.0	395.0	416.0	421.0	403.0	396.0	328.0	275.0	300.0
11	GU	3.0	4.0	5.0	4.0	2.0	5.0	5.0	5.0	3.0	5.0	6.0	5.0
12	HI	34.0	45.0	32.0	32.0	34.0	51.0	45.0	59.0	51.0	44.0	43.0	47.0
13	IA	105.0	126.0	122.0	111.0	115.0	108.0	109.0	105.0	133.0	132.0	143.0	122.0
14	ID	87.0	105.0	97.0	92.0	89.0	79.0	94.0	85.0	78.0	81.0	76.0	73.0
15	IL	659.0	651.0	694.0	697.0	820.0	735.0	702.0	777.0	913.0	770.0	674.0	700.0
16	IN	280.0	308.0	340.0	284.0	280.0	279.0	266.0	224.0	410.0	367.0	338.0	337.0
17	KS	101.0	100.0	120.0	124.0	113.0	97.0	109.0	131.0	124.0	132.0	105.0	126.0
18	KY	201.0	214.0	222.0	177.0	202.0	192.0	170.0	168.0	258.0	194.0	192.0	225.0
19	LA	147.0	145.0	160.0	130.0	123.0	149.0	151.0	106.0	92.0	112.0	95.0	109.0
20	MA	344.0	287.0	262.0	247.0	327.0	302.0	298.0	386.0	407.0	459.0	402.0	431.0

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

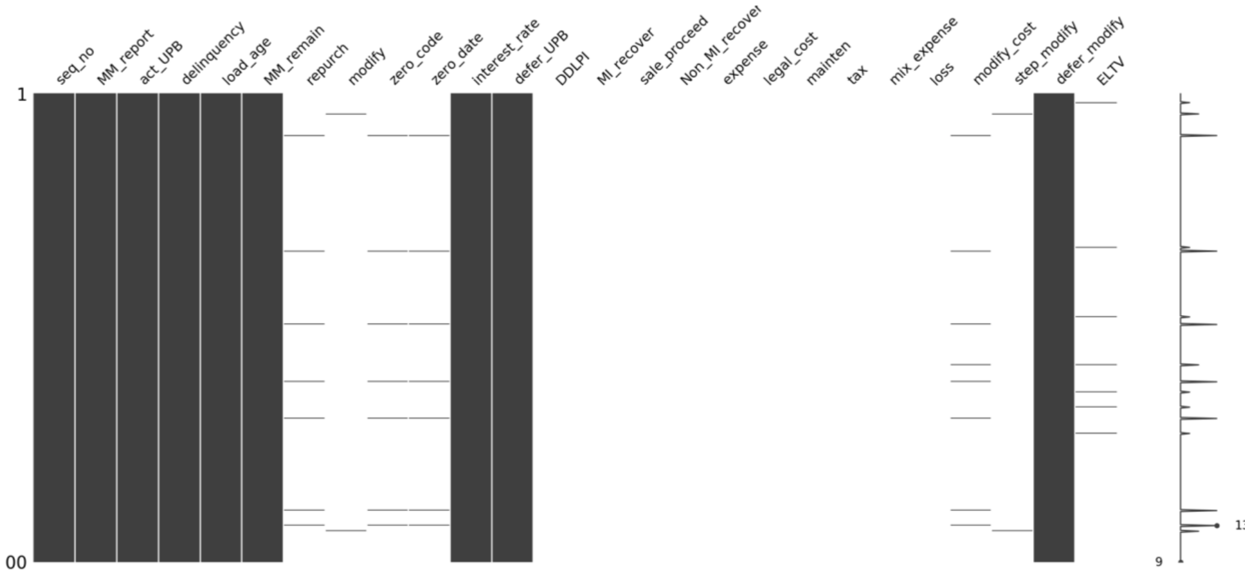


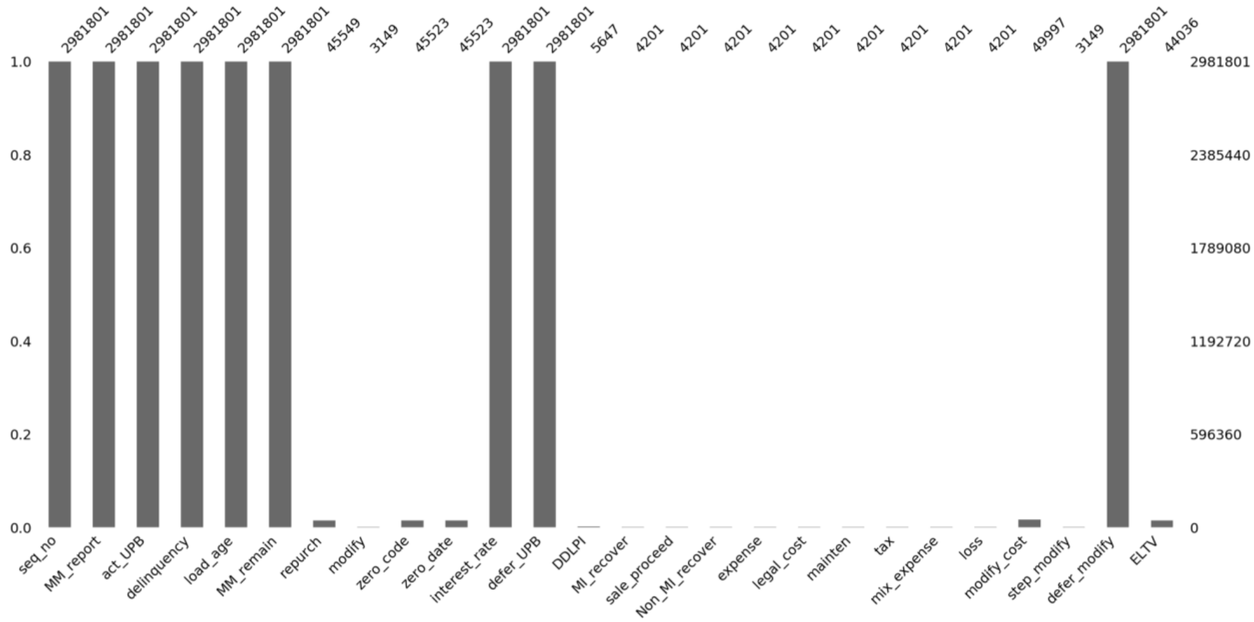
- Single year comparison

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

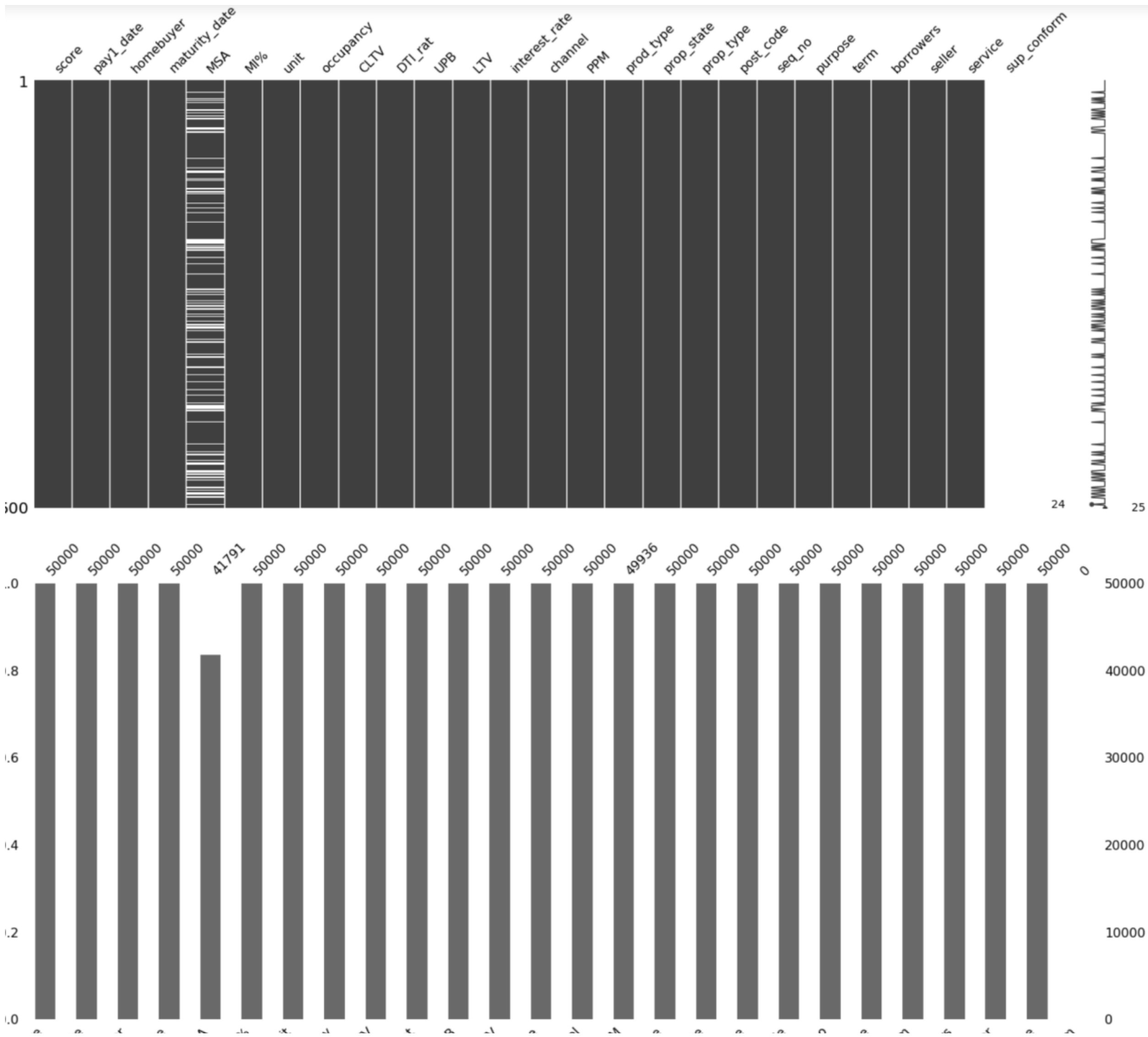
Missing value visualization

Performance data





Origination data

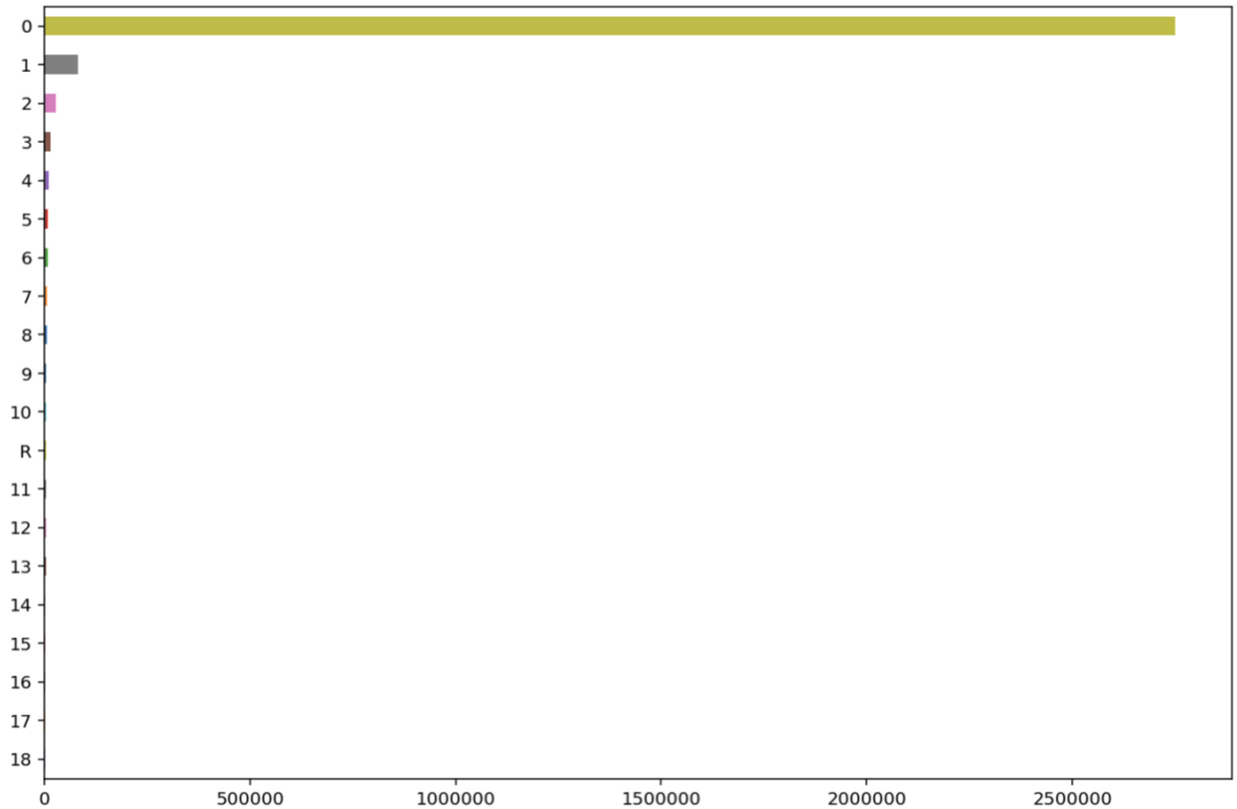


Numerical Parameters

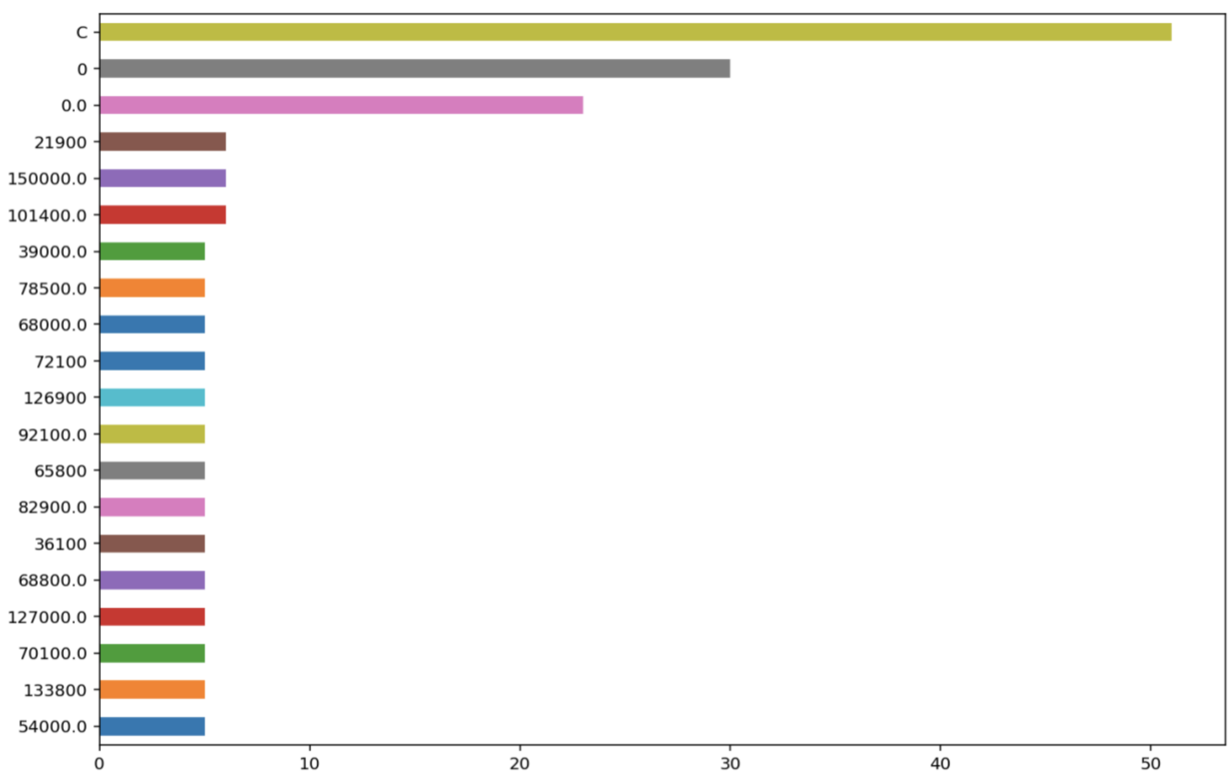
	count	mean	std	min	25%	50%	75%	max	year
MM_report	2981801.0	200970.520211	282.501479	200602.00	200711.000	200906.000	201110.000	201803.0	2006
act_UPB	2981801.0	157320.881169	92174.788874	0.00	89160.220	139511.250	209000.000	802000.0	2006
load_age	2981801.0	42.695682	33.721852	0.00	16.000	34.000	62.000	145.0	2006
MM_remain	2981801.0	299.506346	73.892356	-1.00	281.000	322.000	344.000	603.0	2006
zero_code	45523.0	1.598005	2.160102	1.00	1.000	1.000	1.000	15.0	2006
zero_date	45523.0	201049.853063	259.843339	200602.00	200901.000	201007.000	201208.000	201803.0	2006
interest_rate	2981801.0	6.270802	0.756462	0.00	6.125	6.375	6.625	50.0	2006
defer_UPB	2981801.0	638.023355	7164.401571	0.00	0.000	0.000	0.000	271000.0	2006
DDLPI	5647.0	201094.026740	305.119185	200602.00	200903.000	201007.000	201209.000	201801.0	2006
MI_recover	4201.0	9132.957153	21500.075614	0.00	0.000	0.000	0.000	139622.0	2006
Non_MI_recover	4201.0	5734.063556	23407.332456	-6945.00	293.000	1099.000	2404.000	325650.0	2006
expense	4201.0	-15669.726970	15694.850990	-123619.00	-20754.000	-11555.000	-5736.000	138608.0	2006
legal_cost	4201.0	-3394.402761	2583.478226	-30052.00	-4635.000	-3079.000	-1854.000	0.0	2006
mainten	4201.0	-4884.736253	7807.728377	-89012.00	-6033.000	-1940.000	-91.000	0.0	2006
tax	4201.0	-6708.832421	9356.940687	-98168.00	-8316.000	-3946.000	-1546.000	121136.0	2006
mix_expense	4201.0	-681.774101	2859.826372	-24441.00	-842.000	-370.000	-220.000	158807.0	2006
loss	4201.0	-86827.074982	65108.755225	-487818.00	-127467.000	-76655.000	-37372.000	59354.0	2006
modify_cost	49997.0	1431.915314	8859.571493	-15195.29	0.000	0.000	0.000	206299.9	2006
ELTV	44036.0	57.788841	27.706662	0.00	40.100	58.400	74.700	343.4	2006

Categorical parameters

Delinquency status distribution in each year



Sale proceed distribution in each year



And many other categorical parameters.

```
plt.figure()
df_perf['repurch'].value_counts(ascending=True).plot(kind='barh')
plt.figure()
df_perf['modify'].value_counts(ascending=True).plot(kind='barh')
plt.figure()
df_perf['step_modify'].value_counts(ascending=True).plot(kind='barh')
plt.figure()
df_perf['defer_modify'].value_counts(ascending=True).plot(kind='barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x29aa1a44dd8>
```