Over View:

HMDA data is given for the years between 2012 to 2014 for the states of Washington DC, Delaware, Virginia, West Virginia, Maryland. The purpose of report is to identify market segments and give data driven solutions for Change financial to enter home loans market. Loan Data is given on home loans originated in states where change financial operates.

Institution Data is data for originating institutions. The procedure for analysis is as follows.

1. Data Munging
2. Quality check
3. Visualizations
4. Future study

1. Data Munging:

The loan data and Institution data are merged using pandas' library on 'Respondent_ID', 'As_of_Year', 'Agency_Code'. Function hmda_init () is created to merge loan and institution data into single data frame. Loan Amount in the merged data is divided into buckets based on quantiles and new column "Loan_Amount_Size" is added.

hmda_to_json (data, states=None, conventional_conforming=None) function with optional parameters states and conventional conforming is created to filter data based on group parameters passed and create JSON files. It returns True if json files for the given list of states are created or error "Export to JSON unsuccessful" is popped.

Format: hmda_to_json (full_data, states = [' ' , ' '], conventional_conforming = True)

2. Quality check:

Quality assessment on Loan_Amount_000:

CODE: full_data['Loan_Amount_000']. describe ()

count	1321158
mean	290.23
std	965.77
min	1
25%	153
50%	235
75%	347
max	9962.5

Rules for entering of Loan Amount given by guide of HMDA 2013:

1. For reporting submissions, automated form loan amount must be entered with leading zeros. ($55000 should be entered as 00055) Not required for paper forms.
2. Loan amount must be entered in thousands. ($200000 must be entered as 200.)
3. Loans below $500 should not be reported.
4. Enter principal loan amount originated home purchase loan, for purchased home loan enter unpaid principal balance loan.
5. For home improvement loan, enter total amount even add any unpaid finance charges.
6. For declined loan applications enter the amount applied for which it got rejected.

Quality assessment on Respondent_Name:

count	1321158
unique	1602
frequency	137060
Name: Respondent_Name_TS	dtype: object

Various names are given to same respondent because they report to different agencies and each agency reported in different format.

Type error is also a problem. Duplicate values are also present.

Respondent must provide unique ID and Name to all agencies working. Any update in name must be done with change in previous records in data base or query must be generated to match the data with new name.

Rules for quality improvement and monitoring:

1) Before adding data or updating data organizations must ensure they have identified the degree to which the previous data is improved without poor quality. It is better to have no data than data with poor quality.
2) Business should define data quality rules and measurements. Common information like primary keys, dates, countries, states (geographical information), ID and addresses should be of customized format.
3) These formats must be made into rules and conditions to make reference models.
4) These models must be used to design and develop or update data environment.
5) These standardized rules should be published for users working on various projects.
6) Quality improvement process to remove flaws in data updating must be used.
7) Validating data should be followed by monitoring and cleansing.

Other columns to monitor:

1. Respondent_ZIP_Code data can be converted into standard format for each state. Some zip codes are in format of (00000) - (0000). This can be converted to 5-digit zip codes which will be useful for cluster analysis to study geo special data in market analysis
2. Conforming limit data which is conforming limit for the county has 837 missing values.

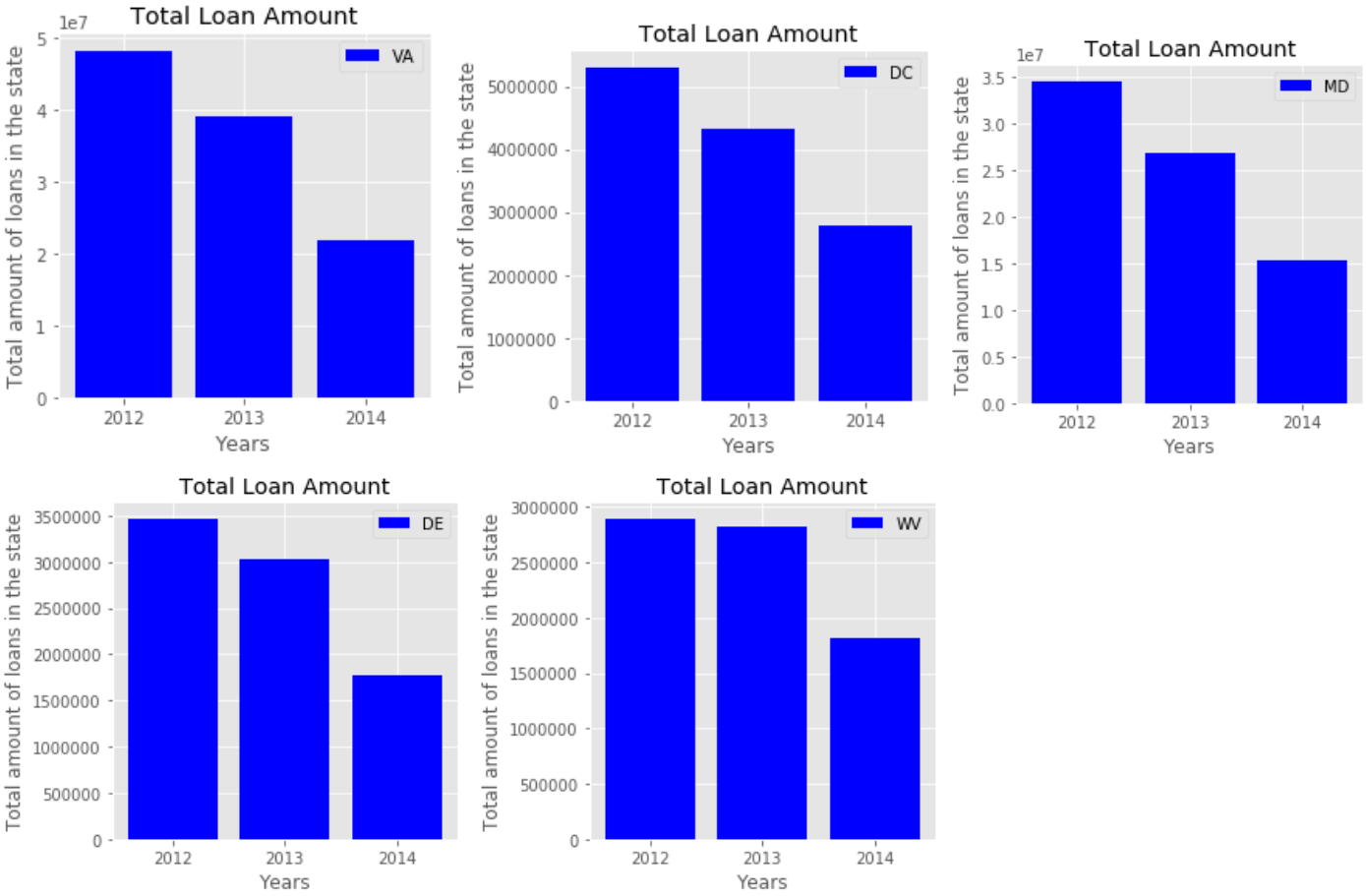
3. Tract_to_MSA_MD_Income_Pct has 900 null values. While studying market analysis at metropolitan areas in certain state this column can be useful in assessing loan application capability.

3. Visualizations & Analysis of HMDA Data

Change Financials decision to enter a new geography in home loans market?

While entering a new market segment it should consider most affordable segment rather than investing more. The loan purpose in the given states is Refinance or Purchase. The hypothesis for Change Financial is, that the refinance loans will take less amounts and to be at low risk and purchase loans require high lending initially. The main question to look is, can Change Financial enter home loan market by studying information on loan metrics, types of loans and top market competitors in the geography.

Total Loan Amount by state and year:



From 2012 to 2014 there is a decline in total loan amount every year. The amount of loan spent in each state will be different and based on various constraints. In the year 2012 51% of total loan amount is originated from

Virginia, ~37% from Maryland. 2013 In 2013 they hold 51%, ~35% and 50%, ~35% in year of 2014. Washington DC, West Virginia, Delaware shows the same trend but the percent of loan amount concentration is almost stable. Let's further analyze market capacity by loan volume and further break down into counties, top companies.

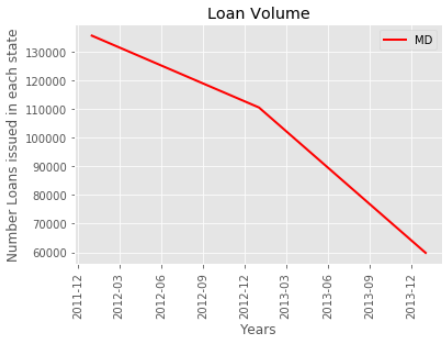
As_of_Year	State	Total_loan_amount	percent
2012	DC	5302490	5.61
2012	DE	3468746	3.67
2012	MD	34553610	36.57
2012	VA	48256768	51.08
2012	WV	2897320	3.07

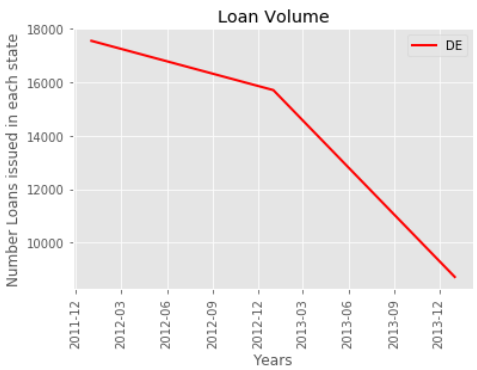
As_of_Year	State	Total_loan_amount	percent
2013	DC	4332573	5.69
2013	DE	3035340	3.99
2013	MD	26870750	35.29
2013	VA	39092849	51.34
2013	WV	2817404	3.7

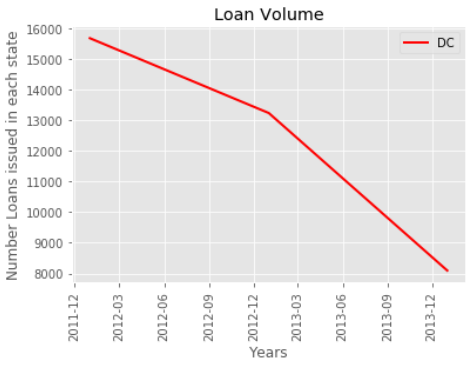
As_of_Year	State	Total_loan_amount	percent
2014	DC	2786543	6.39
2014	DE	1771439	4.06
2014	MD	15273937	35.01
2014	VA	21972546	50.37
2014	WV	1820078	4.17

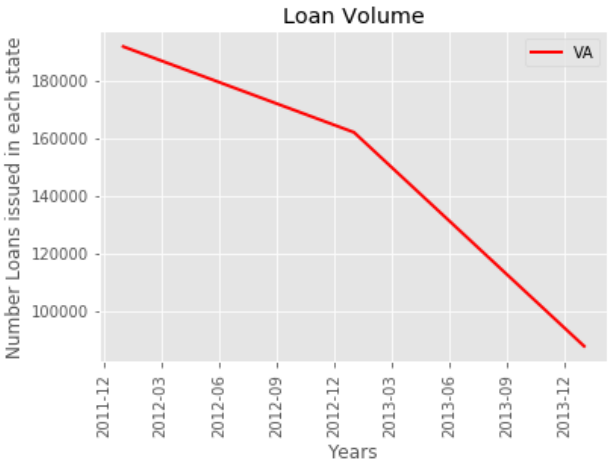
Total Loan Volume in each state and year

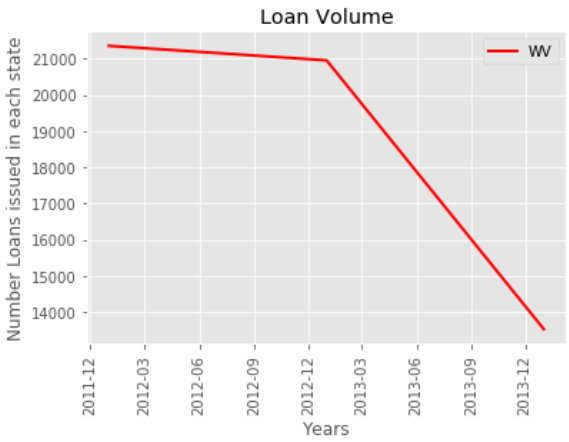
Note: years are considered at the end of previous year for analysis (2012- 2011-12)











From the graphs above there is a decline in total loan volume from 2012 to 2014, i.e., total number of confirming loans in Washington DC, Delaware, Virginia, West Virginia and Maryland. Every state has different number of loans originated but VA holds about 50%, MD holds ~36% of total loans originated in 2012. In 2013 they hold 50%, ~35% and 49%, ~34% in 2014 respectively. It shows that major share of loans is in areas of Maryland and Virginia. (graphs for loan volume). From the data of amount of loans and volume of loans the market share of VA has consistently stayed above ~50% and Maryland above ~35%. Now we can consider major concentration of counties and areas in the states and how top contenders are performing. For the state of DC, WV, DE the loan volume is stable.

As_of_Year	State	Total_Confirmed_loans	percent
2012-01-01 00:00:00	DC	15676	4.1
2012-01-01 00:00:00	DE	17555	4.6
2012-01-01 00:00:00	MD	135611	35.51
2012-01-01 00:00:00	VA	191696	50.2
2012-01-01 00:00:00	WV	21355	5.59

As_of_Year	State	Total_Confirmed_loans	percent
2013-01-01 00:00:00	DC	13234	4.1
2013-01-01 00:00:00	DE	15712	4.87
2013-01-01 00:00:00	MD	110548	34.29
2013-01-01 00:00:00	VA	161944	50.23
2013-01-01 00:00:00	WV	20954	6.5

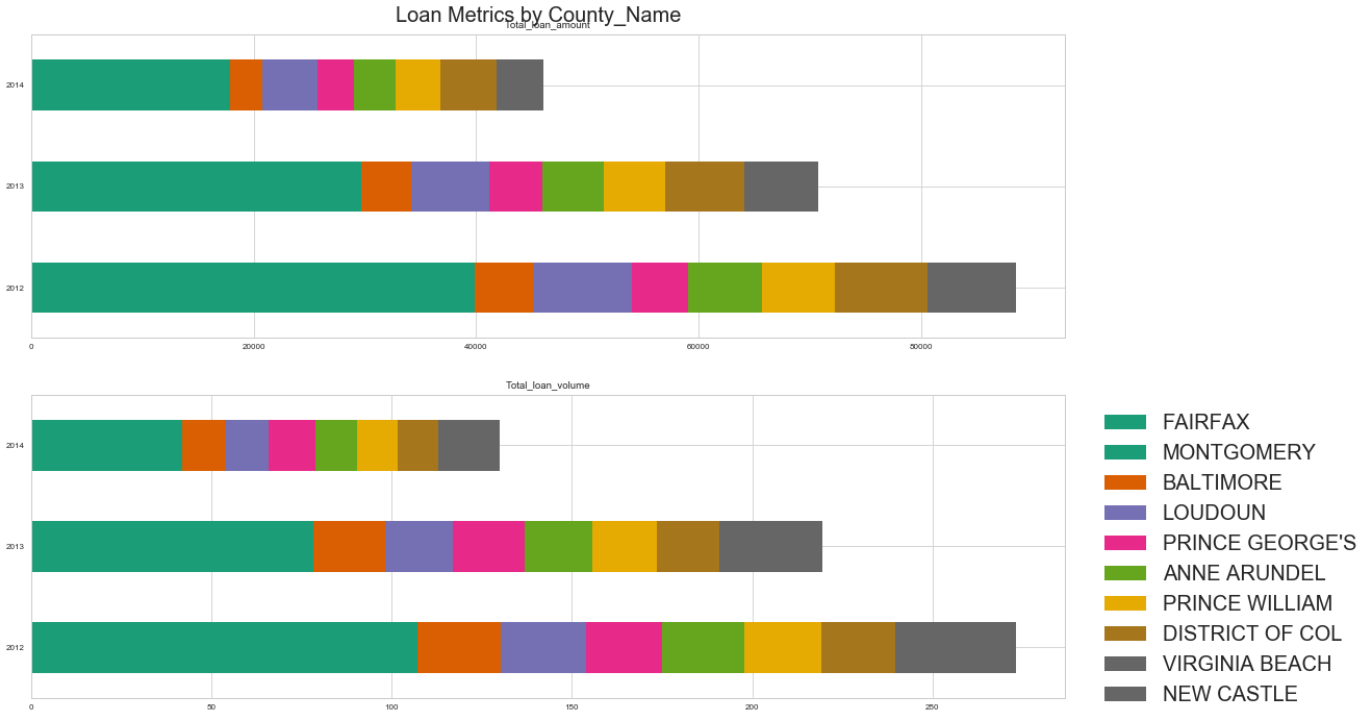
As_of_Year	State	Total_Confirmed_loans	percent
2014-01-01 00:00:00	DC	8096	4.55
2014-01-01 00:00:00	DE	8721	4.9
2014-01-01 00:00:00	MD	59820	33.62
2014-01-01 00:00:00	VA	87749	49.32
2014-01-01 00:00:00	WV	13535	7.61

County wise Analysis:

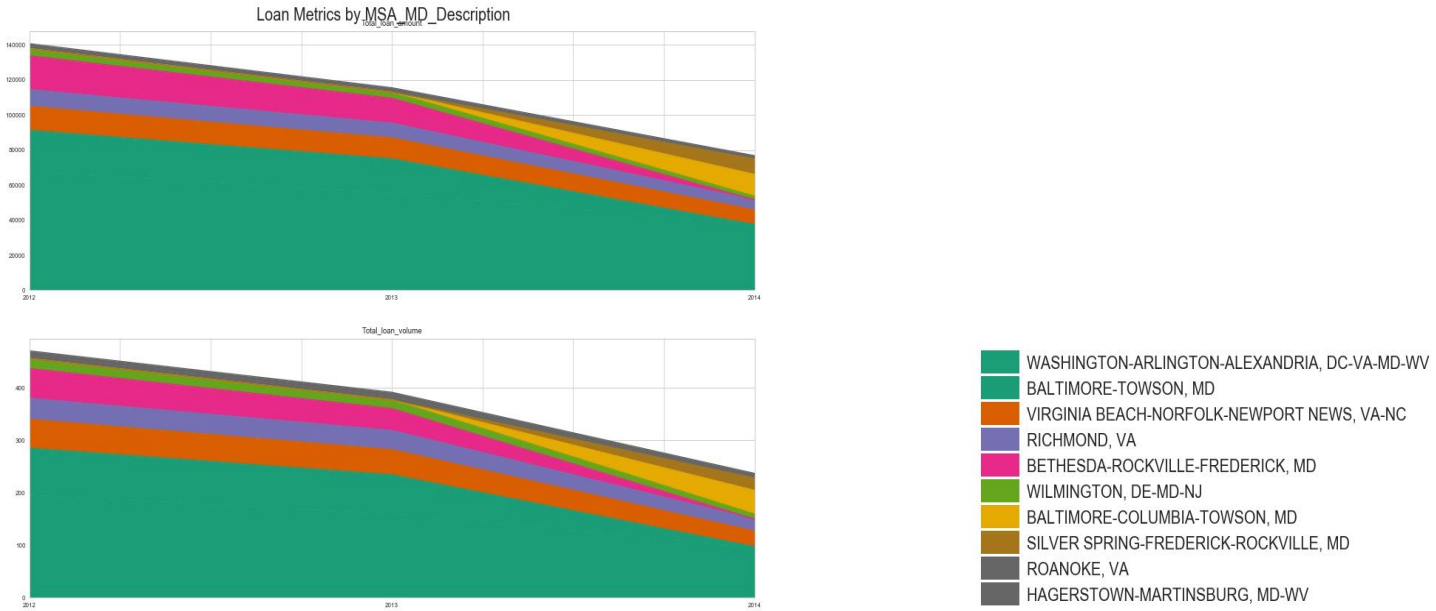
Tables for loan amount and loan volume metrics.

	FAIRFAX	MONTGOMERY	BALTIMORE	LOUDOUN	PRINCE GEORGE'S	ANNE ARUNDEL	PRINCE WILLIAM	DISTRICT OF COL	VIRGINIA BEACH	NEW CASTLE
2012	25.8	19.21	6.0	9.98	5.72	7.58	7.35	9.45	5.1	3.81
2013	24.46	17.46	6.41	9.94	6.66	7.85	7.8	10.03	5.25	4.16
2014	22.52	16.37	6.26	10.7	7.23	8.11	8.61	10.95	5.39	3.86

	FAIRFAX	MONTGOMERY	BALTIMORE	LOUDOUN	PRINCE GEORGE'S	ANNE ARUNDEL	PRINCE WILLIAM	DISTRICT OF COL	VIRGINIA BEACH	NEW CASTLE
2012	21.62	17.63	8.43	8.72	7.7	8.38	7.75	7.59	6.38	5.81
2013	19.84	15.75	9.08	8.56	9.08	8.57	8.11	7.92	6.68	6.4
2014	17.55	14.78	9.3	9.08	9.93	9.0	8.57	8.73	6.8	6.25



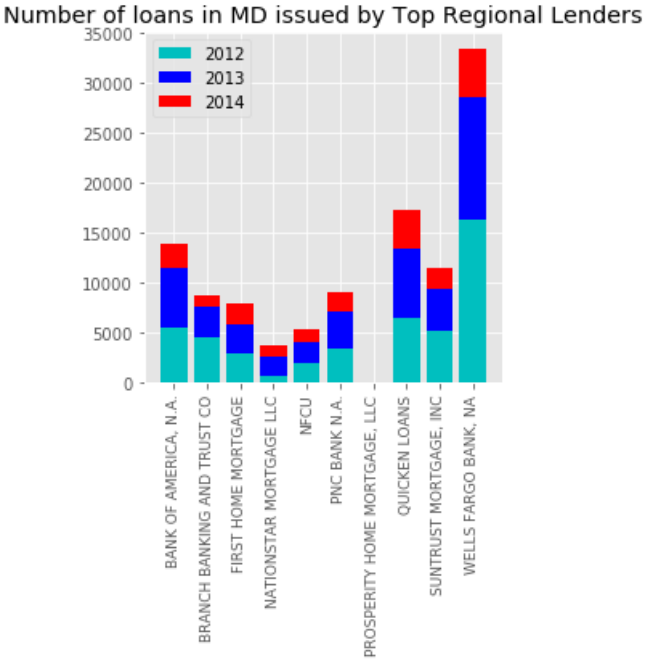
From the graph, and metrics shows the grouped top ten counties by the total loan amount and loan volume. Most of the market share is in Fairfax and Montgomery but by observing the data, the counties originated are from Virginia and Maryland states. The following graph of MSA_MD confirms the major share of VA, MD in home loan market. Most of metropolitan statistical areas fall under VA, MD.

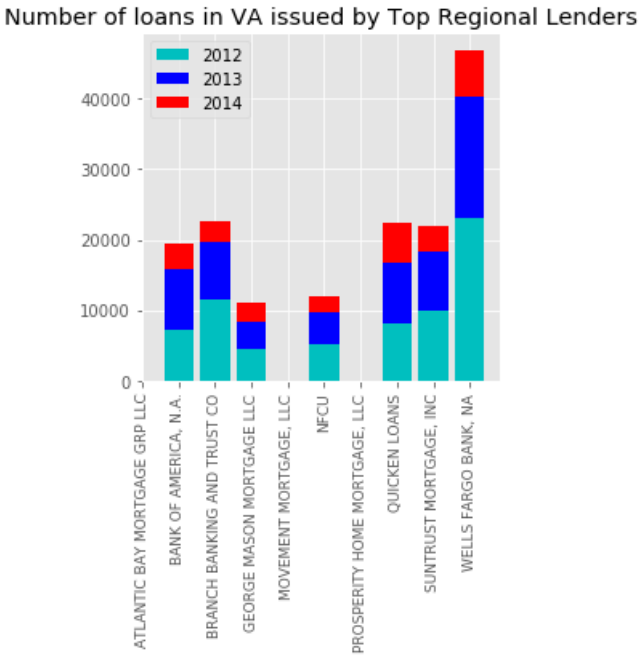


As we found out the major geography of loan market, let's consider the types of loan and top companies' preferences in the regions.

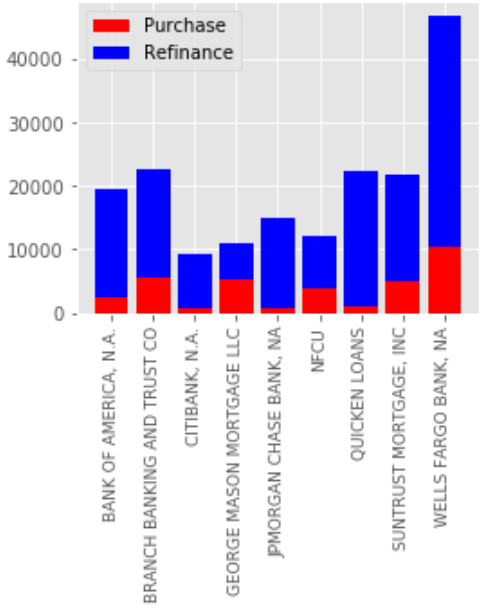
Loan Analysis based on purpose of loan:

The following graphs represents the information with top companies with highest market share in the regions. The graphs strongly show the Metrics for volume of loans for each type of loans processed by of top lenders. The companies like JP Morgan Chase, wells Fargo, Bank of America are major market leaders in all the states and the concentration of investment is more in Refinance loans. The major types of loan are Refinance loans rather than new loans.

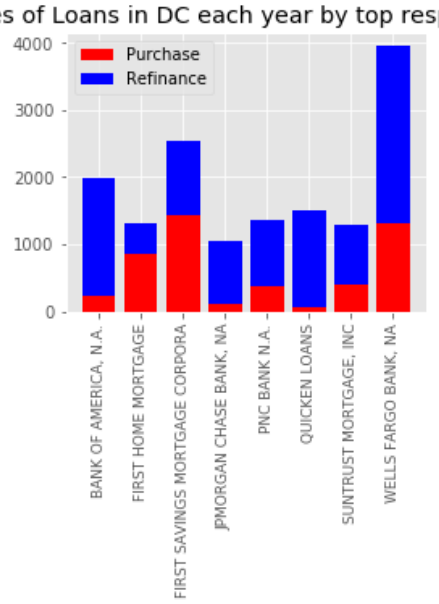




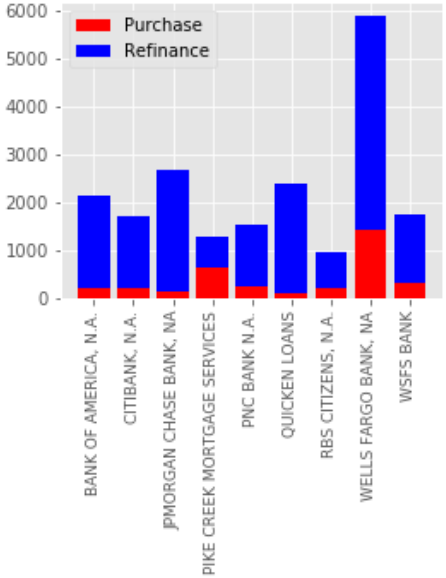
Types of Loans in VA each year by top respondents



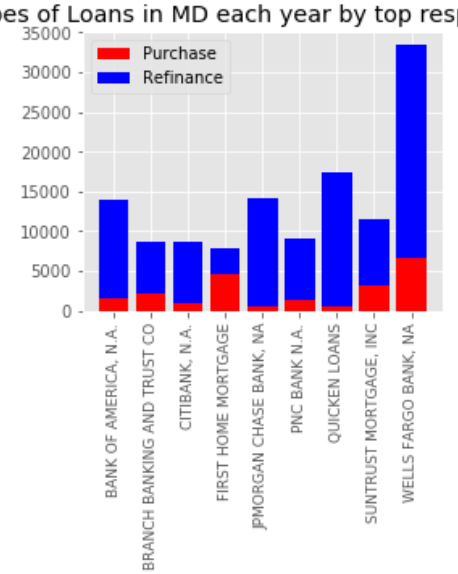
Types of Loans in DC each year by top respondents



Types of Loans in DE each year by top respondents

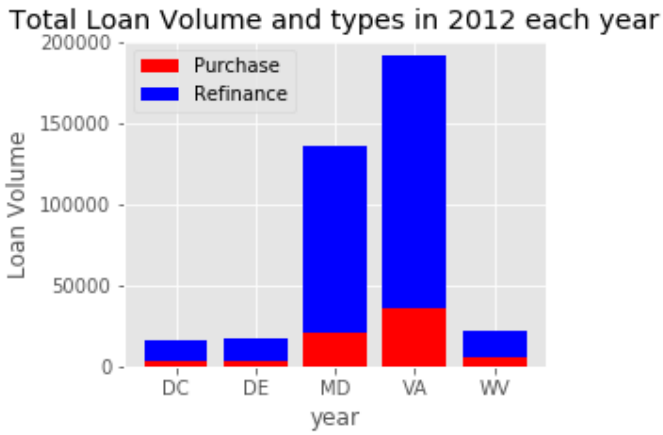


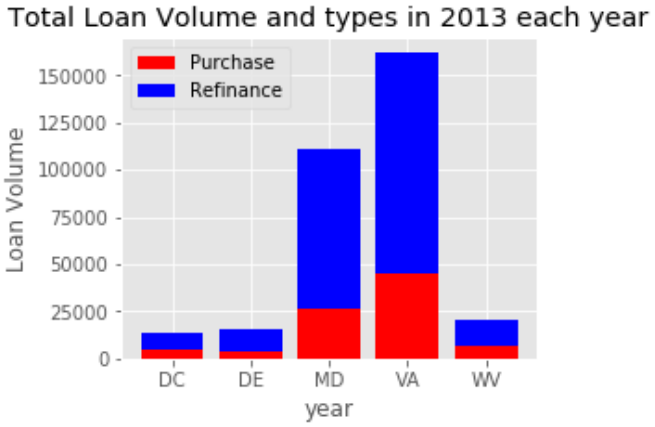
Types of Loans in MD each year by top respondents

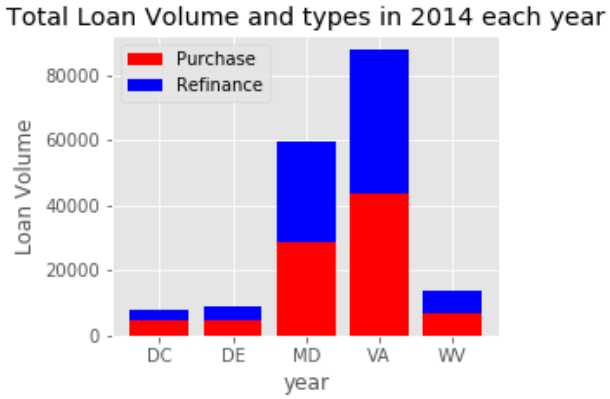


Let's consider yearly changes in types of loans originated.

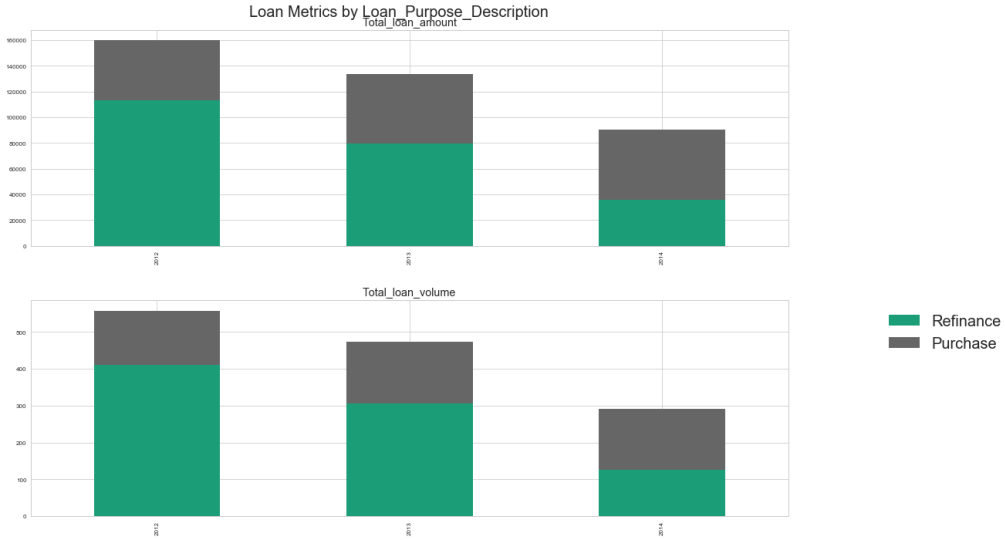
The bar graphs of years 2012, 2013, 2014 shows the total volume of conventional conforming loans grouped by purpose of loan. Purchase loan which is for buying new house and refinancing loan. Most of the loans given by the states are refinance loans in year 2012. But we can see a decrease in number of refinance loans each year in top markets like Virginia and Maryland. Other states it is constant.







Total loan amount and loan volume based on loan purchase description:



The graphs give clear trend in total loan amount and total volume of loans based on loan purpose description. From the year 2012 to 2014 the refinance loans are decreasing from 73.59% to 43.27%.
The purchase loans market is stable and slightly increasing every year from 26.41% in 2012 to 35.11% in 2013 and 56.73 % in 2014.

	Refinance	Purchase
2012	73.59	26.41
2013	64.89	35.11
2014	43.27	56.73

As_of_Year	Loan_Purpose_Description	Total_loan_amount	Total_loan_volume
2012	Purchase	17929385	69116
2012	Refinance	76549549	312777
2013	Purchase	23325203	87400
2013	Refinance	52823713	234992
2014	Purchase	23689096	88246
2014	Refinance	19935447	89675

Number of organiastion based on loan purchase description:

As_of_Year	Loan_Purpose_Description	Total_loan_volume
2012	Purchase	69116
2012	Refinance	312777
2013	Purchase	87400
2013	Refinance	234992
2014	Purchase	88246
2014	Refinance	89675

The number of companies processing Refinance loans are decreasing each year. There is almost 28% drop in number of companies. But number of companies dealing with purchase type loans are stable.

This information clears out the hypothesis; the refinancing loans would be better choice for change financial if it wants to enter in low market share states like WV, DC, DE. In the states of Virginia and Maryland the major shares in home loan market is for Purchase loans and is stable. The decreasing trend in Refinancing loans is not a good sign to enter in huge competition markets. Entering the market to attract the customers may depend upon various factors like low upfront fees, high prevailing market rate, Low interest's rates than existing or before loan offerings. (This data is not given but can be used in future for analysis).

From the analysis, we saw that the major geography areas to concentrate are Virginia and Maryland, the top regional lenders originate largest market share in refinance loans during years 2012 and 2013. But, studying total loan amount, total volume trends of loan types should the diminishing effect of refinance loans. Visualizations on percentage of market shares by lenders and loan types has given a significant approach in understanding the segments of loan purpose types. With more data regarding trend, risk, interests and legal entities we could strongly improvise on hypothesis. Even though, it is obvious that the Change Financial could enter the big market geography areas by lending purchase loans which looks more stable.

4. INTERACTIVE QUALITY:

To keep the data clean and optimize the memory, the outliers and duplicates must be removed. Two functions are developed in HMDA_Main.py file to address these issues.

1) data_outlier (df, column_df, threshold=n)
   Functionality: Pass the data, column and threshold value at which you want to filter the data. This can be z Score value or any limit per data. This checks whether the given column observations are in indicated standard deviations from mean.

2) update_duplicate (data, name)

   Functionality: pass the data set and pass other parameter name indicating loan_data or institutions_data
   If loans data file is getting updated Agency_Code', 'Respondent_ID', 'As_of_Year' are the unique columns or primary keys to identify unique rows in data. If institutions data is getting updated 'Agency_Code', 'Respondent_ID', 'As_of_Year', 'Sequence_Number' are main columns.

   This checks for similar rows and returns duplicate records.

Clusters of lending patterns and volume trends across states and county are coded in HMDA_Main.py file.

Future Analysis:
Historical data from hmda website can be collected for past years. A forecasting model can be built on loan volume and loan amounts using stats model package in python. Trends and patterns can be studied from the data. Auto correlation and Partial auto correlation plots can be used to build ARIMA (Auto regressive moving average models) or STL model can be built to forecast trends in upcoming year.
From data link given in Data_Challenge_metadata tried to develop model but python crashes due to huge data size (~ GB)

Basic interaction user interface:

Tkinter is used for developing basic GUI. Interactive menu bar is created to get function to execute.