**Executive Summary:**

The purpose of this report is to provide a detailed analysis on HDMA Data. The main objective of the analysis is to make suggestions to Change Financial about the home loan market so that they can make a decision on entering the market. The first step of the process is Data Munging where we imported the data files, created functions to read and join the files. Next would be Quality check in which I have removed data which is irrelevant, outliers and duplicates based upon my assumptions. After obtaining the clean data I created Data Visualizations to show state wise statistics of the data. Finally, I concluded the report with my hypothesis based on the analysis.

**Data Munging:**

Below I have given a step wise explanation of all the work which was conducted in ipython Notebook.

1. Imported libraries such as numpy, pandas for data munging and matplotlib.pyplot for data visualizations.
2. Loaded both the Institutions and loans data CSV files into Institutions\_file and Loans\_file.
3. Implemented as mentioned a hmda\_init() function which returns the data frame containing inner join of Institutions data and loans data.
4. This data frame is saved as combined\_data.
5. Created a hmda\_to\_json function which takes data frame, state and flag variables as parameters.
6. Hmda\_to\_json function logic:

* This function first checks the length of the parameters passed while calling the function.
* If the length is zero then it will convert the data frame passed into json called hmda\_data.JSON.
* Else it creates two lists which are named as states and flags. State and Conforming\_flag (flag here) are the columns in the data frame.
* If length of the parameter is 1 then it indicates a flag value and it will be appended to flags list or else, it adds it to the states list.
* If the list states[ ] has values, then a new data frame is created wherein it returns only the records whose state value matches with those values in states [ ]. Same is repeated with Flag [ ] values.
* So the final data frame is saved in final\_df after filtering the rows according to the parameters, which is then converted into JSON.

1. For values that are not available for Applicant Income and FFIEC Median Income such as NA are converted into Strings for future steps.
2. We considered records where Conforming status is ‘Y’ which means the Loan taken falls under the state limit as well as loan is of less risk.
3. Since our main task was to analyze if the Change Financial should enter Home loans market, it would be more reasonable to consider loans of less risk and fall under the loan limit.

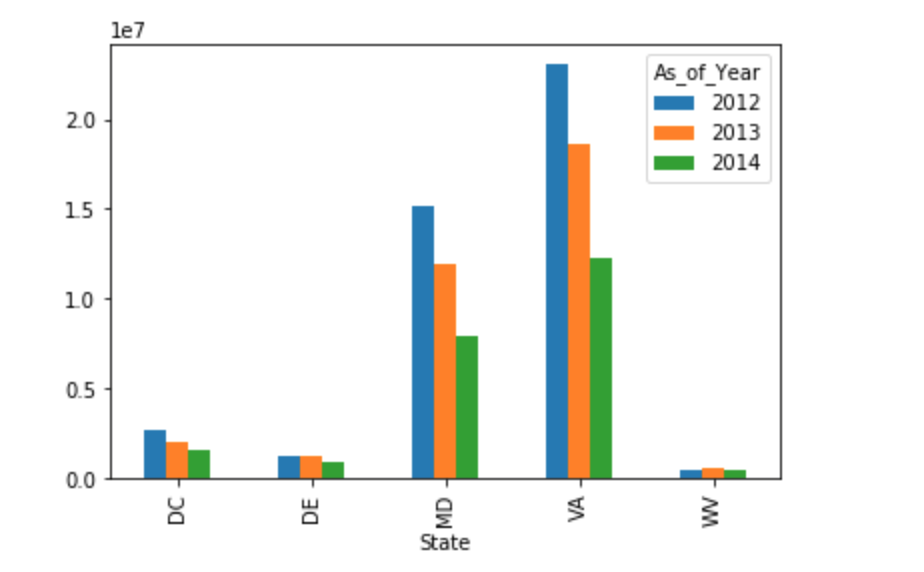
**Data Quality Check:**

Below you will find the quality checks I have performed based on few assumptions.

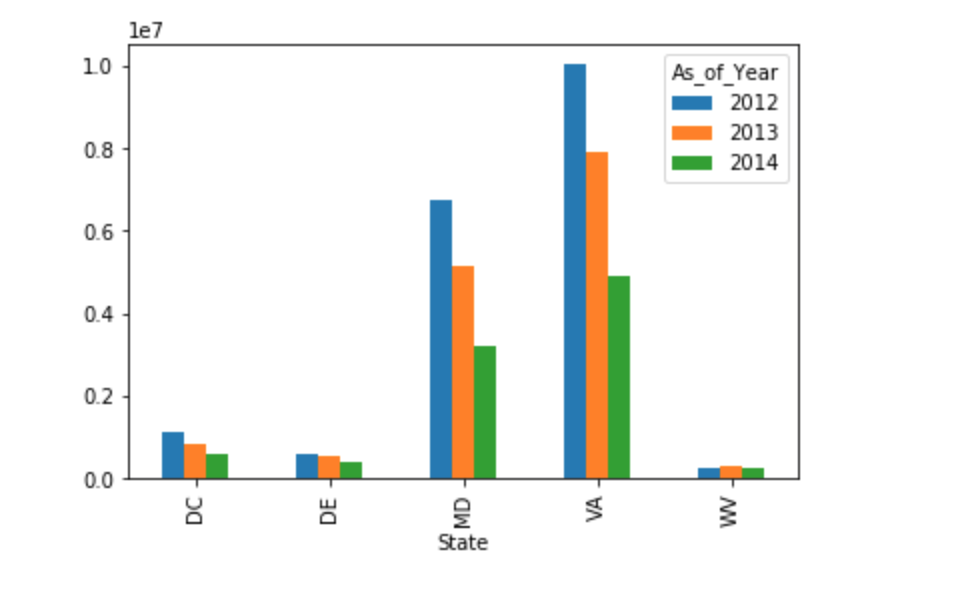
1. First quality check I am performing is to write a formula which gives us a range of applicants who fall under category affordable to pay off the loans. I took this range because these records have consistent data and are in the proximity of the average income. This data can be analyzed to derive insights.
2. Second assumption is that I am assuming loans which are less than 10(‘000s) are more likely to be irrelevant since those loan amounts are very minimal and doesn’t provide any quality information towards deciding the entry into home loan market
3. Another assumption is removing records where Applicant Income <10 and Loan amount >100, as applicants having such low income won’t be receiving such huge loans or won’t be able to pay off such high loans. Hence its considered to be bad data.
4. Removed records which have duplicate state and MSA\_MD\_Description values.
5. Also, I have removed records which have duplicate Respondent Names and displayed the number of duplicate records.

**Data Visualizations:**

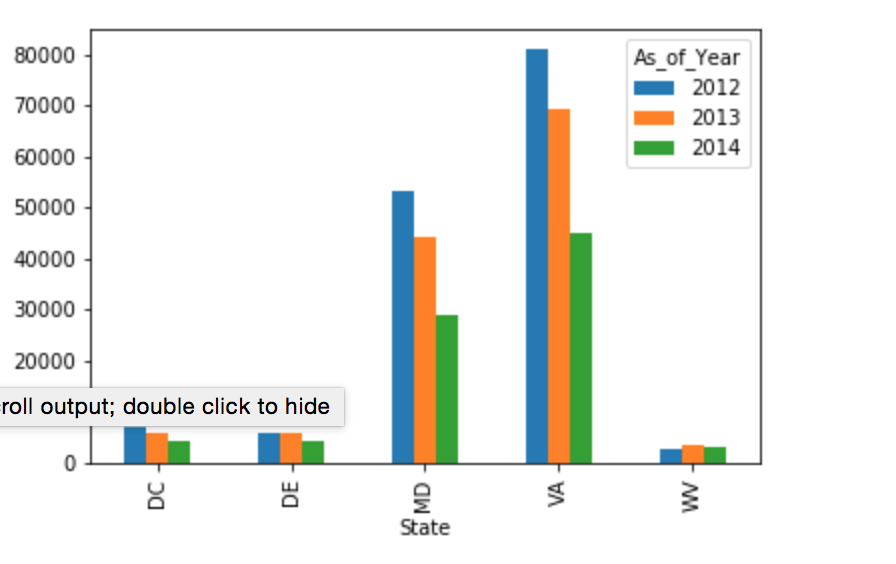
1. Below figure shows state wide Total Loan amount taken over the years.



1. This graph shows us the state wide Total Applicant Income of people over the years.



1. Below bar chart shows the state wide Total Applicants over the years.



Based on the above visualizations, we can see that both Maryland, Virginia has high number of applicants and higher incomes when compared to other three states. Although, over the years’ applicants have been gradually decreasing but still both states have healthy applicants. From above, we can infer that entering Home Loan market in Virginia land Maryland states would be safe as more people with good income are inclined towards taking loans.

**Insights and Future scope**

* Its always a challenge to work on real time data. what particularly intrigued me is that the number of applicants gradually decrease over the years. We can run better analysis when we have quality data. The cleaner the data is better are the outputs. To clean the outliers and obtain more quality data we can use excel miner, python, R etc. we can also run predicting modeling to check the effectiveness of different variables. We can also do clustering using XL miner to check if certain areas fall under the same cluster and perform further analysis from the clusters obtained.
* We can use Tableau for developing interactive systems where in we use filtering and create dashboards to visualize different scenarios. I don’t think its too hard to get market insights, it can be obtained if we can access more data from the past and compare it over the years.
* From the historical data we can check for any trend, seasonality or other external factors effecting the data. Using these insights, we can predict future market more precisely. For example, there is a linear trend for states Virginia and Maryland going downwards for the number of applicants over the 3 years from the data. Also, region wise analysis can be done to check if certain regions have more number of loans to others.