Data Analysis for Capital One

The first step is to import all the required packages required for analysis. Below is not the complete list of packages used, A few packages have been installed as required in the code cells below.

In [72]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import numpy as np
import warnings
import csv
import json
warnings.filterwarnings('ignore')
```

TASK 1: Part A,B

The first task was to design a non-parameterized function called 'hmda_init()' which reads the two datasets from disk and merges them into a single pandas data frame, the function outputs three data frames in total the two original datasets that were loaded from the disk and the merged dataset.

The 'hmda_init()' function calls two more helper functions 'leadingzeros' and 'merge' the first function 'leadinzeros' cleans the two datasets by removing leading zeros in a few columns and the second helper function 'merge' merges the data on the keys 'Respondent_ID','Agency_Code','As_of_Year'. The main function 'hmda_init()' just loads the datasets and calls these two helper functions.

##Note: Give the local file path to the data in the hmda_init() function, in order to run the function successfully.

In [73]:
```
def hmda_init():
    df= csv.DictReader(open('2012_to_2014_loans_data.csv'), delimiter=',')#give th
    dictn = {}
    for row in df:
        for column, value in row.iteritems():
            dictn.setdefault(column, []).append(value)
    loans_data=pd.DataFrame(dictn)
    institutions_data=pd.read_csv('2012_to_2014_institutions_data.csv')
    loans_data,institutions_data=leadingzeros(loans_data,institutions_data)
    return (loans_data,institutions_data,merge(loans_data,institutions_data))
```

In [74]:
```
def leadingzeros(loans_data,institutions_data):
    institutions_data.loc[:,'Respondent_ID']=institutions_data.loc[:,'Respondent_I
    leadingzerolist=[2,4,9,11,18,19,20,23]
    for i in leadingzerolist:
        loans_data.iloc[:,i]=loans_data.iloc[:,i].str.lstrip('0')
    return (loans_data,institutions_data)
```

In [75]:
```
def merge(loans_data,institutions_data):
    loans_data.loc[:,['As_of_Year','Agency_Code']]=loans_data.loc[:,['As_of_Year',
    fulldf=pd.merge(loans_data, institutions_data, on=['Respondent_ID','Agency_Cod
    return (fulldf)
```

Running the hmda_init function and collecting the results into respective dataframes.

In [76]:
```
loans_data,institutions_data,mergeddf=hmda_init()
```

the 'clean()' function written below handles the blank values by converting them into NaN's and also assigns the appropriate data type to each column of the merged dataframe.

In [77]:
```
def clean(loans_data,institutions_data,mergeddf):
    intigerlist=loans_data.iloc[:,[2,4,5,9,11,13,16,18,20,22,23]].columns
    intlist=institutions_data.iloc[:,[6,10,11]].columns
    keys=loans_data.iloc[:,[0,3,19]].columns
    categrylist=mergeddf.columns-(intigerlist+intlist+keys)
mergeddf.loc[:,(intigerlist+intlist)]=mergeddf.loc[:,(intigerlist+intlist)].ap
mergeddf=mergeddf.applymap(lambda x: np.nan if x =='' else x)
    mergeddf[categrylist]=mergeddf[categrylist].apply(lambda x:x.astype('category'
    return (mergeddf)
```

```
In [78]: mergeddf=clean(loans_data,institutions_data,mergeddf)
```

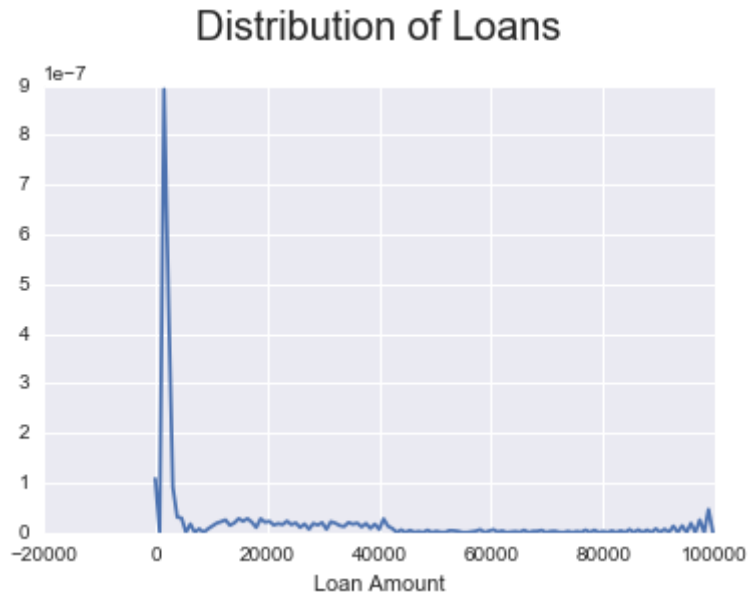
Task 3: Visualization

The question of interest here for change financials is, whether or not it is a good idea to enter the home loans market. To answer the question let us take help of visualization, which is a very effective way of finding and conveying insights.

First visualizing the distribution of Loan_Amount_000, the density plot shows that the distribution is heavily right skewed and a lot of values fall in the lower bin indicating a lot of loans are under the conforming limit and the observations on the right tale indicate jumbo status loans.

```
In [87]: sns.distplot(mergeddf.Loan_Amount_000,hist=False)
         plt.title("Distribution of Loans",y=1.08,size=20)
         plt.xlabel('Loan Amount')
```

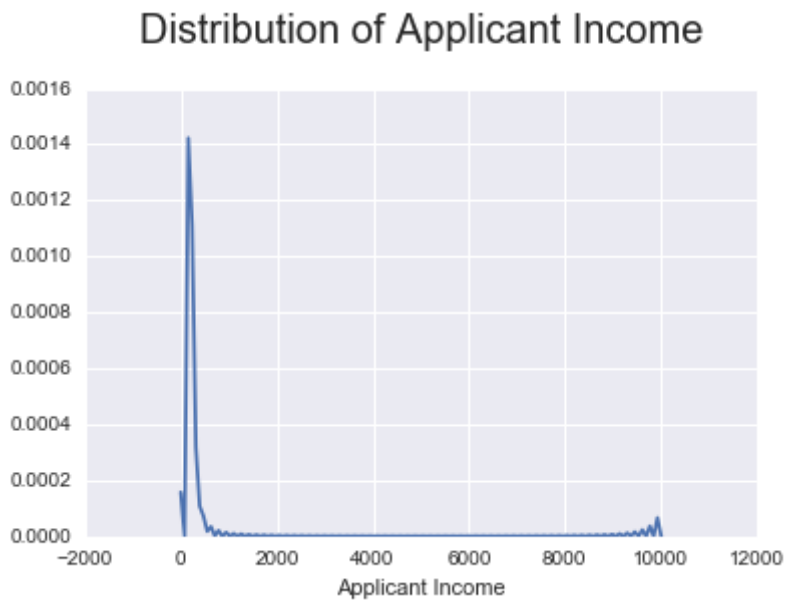
```
Out[87]: <matplotlib.text.Text at 0x1437ccdd8>
```



The density plot of Applicant income is also similar to the density plot of loans this is as expected because applicants with low income generally apply for loan amounts which are less or close to the confirming limit.

```
In [88]: sns.distplot(mergeddf.Applicant_Income_000,hist=False)
         plt.title("Distribution of Applicant Income",y=1.08,size=20)
         plt.xlabel('Applicant Income')
```

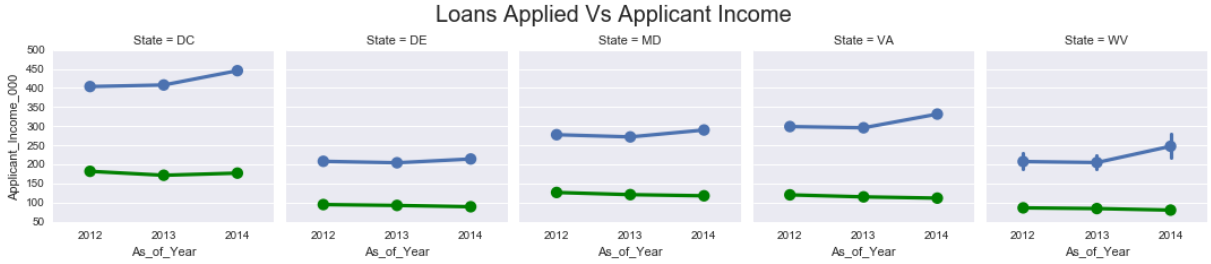
```
Out[88]: <matplotlib.text.Text at 0x966e9eb8>
```



To get a better understanding of how these two columns loan and applicant income and how they interplay, consider the below plot which visualizes the loans and applicant income averages together in the same plot. In the below plot the blue color lines plot on the top is the average of loans applied through the years, one can notice that in year 2014 the average loans increased, and the green color lines plot is the applicant average income, and as we can see the loans applied are close to twice the average of applicants incomes.

In [89]:
```
ax=sns.FacetGrid(mergeddf,col='State',col_wrap=5)
ax.map(sns.pointplot,'As_of_Year','Loan_Amount_000',labels='Loan Amount' )
ax.map(sns.pointplot,'As_of_Year','Applicant_Income_000',color='g',labels='Income
# handles, labels = ax.
ax.add_legend()
plt.suptitle('Loans Applied Vs Applicant Income',y=1.08,size=20)
# plt.legend([Loans Applied, Applicant Income],[])
```

Out[89]: <matplotlib.text.Text at 0x16dfc0d30>



Note : 'Blue' : average loans amount , 'Green' : average applicant income.

Let us now plot a pair plot to check and compare if there is any pairwise linear relationship amongst the various integer variables of interest present in the data, namely 'Loan_Amount_000', 'FFIEC_Median_Family_Income', 'Conforming_Limit_000', 'Assets_000_Panel' and 'Applicant_Income_000'. The plots throughout the diagonal are just histogram distribution of the variables of interest

In [90]:
```
sns.pairplot(data=mergeddf.dropna(),vars=['Loan_Amount_000','FFIEC_Median_Family_I
'Applicant_Income_000'],dropna=True,hue='Loan_Typ                    plt.suptitle("Pairwise
Plots",y=1.08,size=20)
```

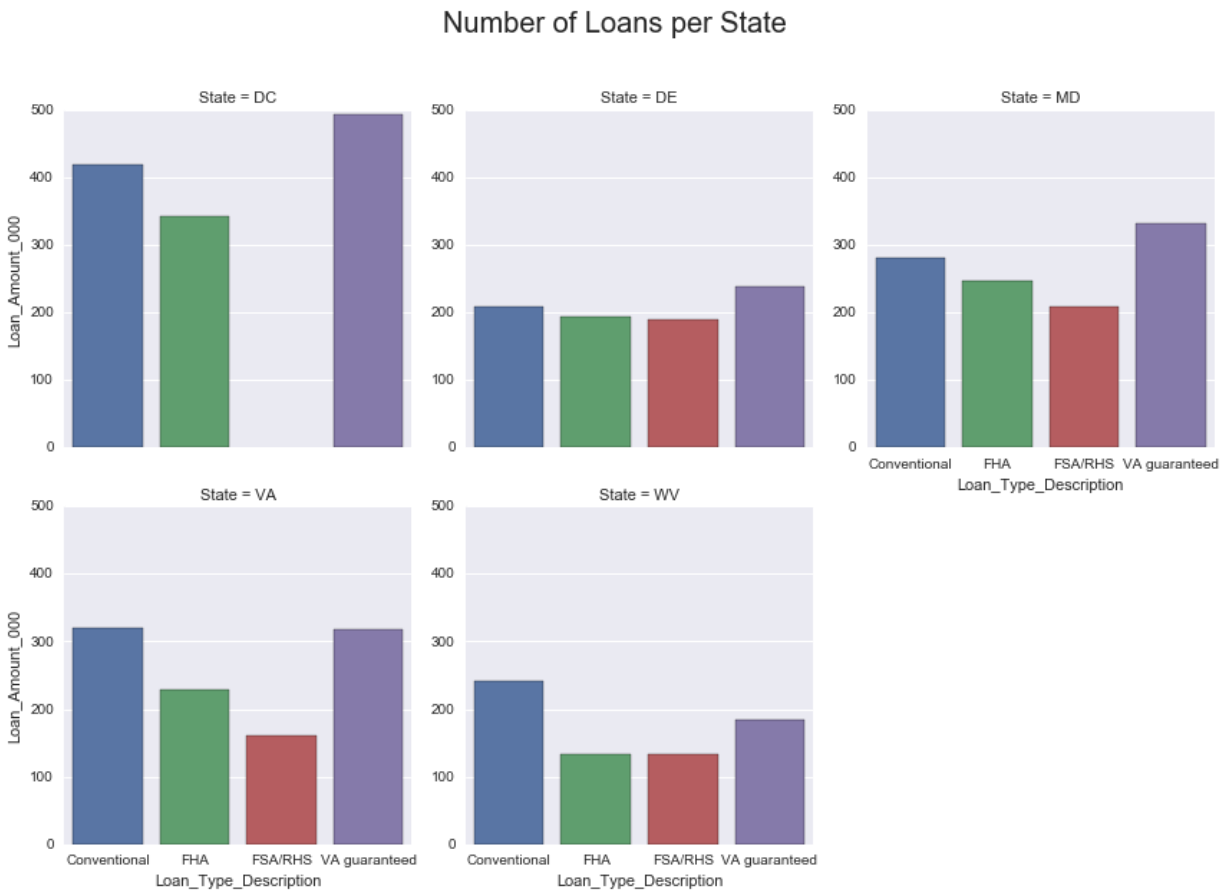
Out[90]: <matplotlib.text.Text at 0x8d8ad128>

There is no significant linear relationships amongst the variables of interest as we can see from the above plot, but there are random Patterns in a few plots that can be noticed above, this implies additional exploration is required to derive meaningful relations.

Let us now try to see the distribution of loans by loan types for each state, and for doing such kind of visulizations, factor plot is a good option.

In [91]:
```
ax=sns.factorplot(x="Loan_Type_Description", y="Loan_Amount_000", col="State",data
                  ci=None,sharey=False,col_wrap=3)
ax.set_xticklabels(["Conventional","FHA","FSA/RHS","VA guaranteed"])
ax.set(ylim=(0, 500)) plt.suptitle('Number of Loans per
State',y=1.08,size=20)
```

Out[91]: <matplotlib.text.Text at 0xa4d705f8>



The above factor plot shows that the number of loans are higher in the state 'DC' than the other states so we can expect a lot of market share for loans from the state of DC. Let us see if the inference we made from the above plot holds true

Plotting influence on market share of the applied loans by factors like state and year.

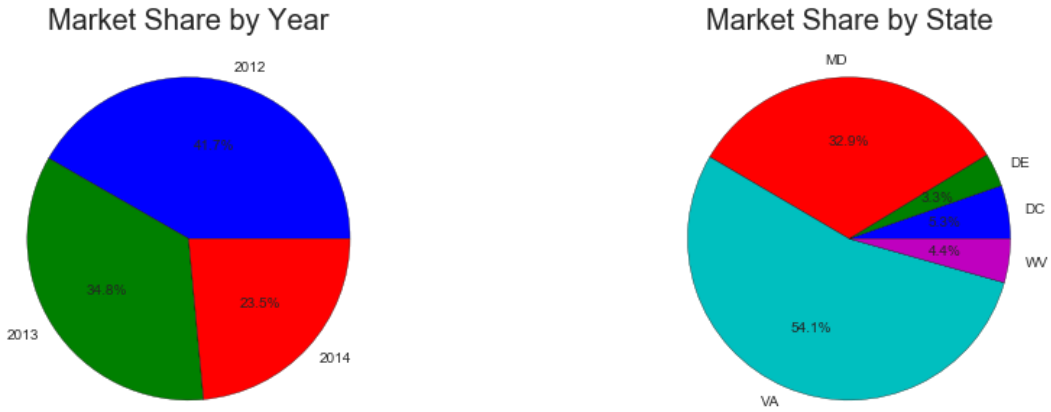
From the first plot below which visualizes market share by year, we can see that the market share for loans is decreasing year by year at a rate close to 10%, but we cannot conclude by this that the market is degrading in performance year after year as the data at hand is the loans application data, and we have no data about how many of these applied loans were sanctioned and how many were rejected. Which makes it difficult to make an inference right away.

The second pie chart represents the loan market by state, and as we can see 'VA' and 'MD' are the states with major market for loans which is contradicting the inference made earlier that the state 'DC' might have higher market share,as we can see the remaining other states including 'DC' contribute less than 5% each in the total aggregate of loans applied. So if Change Financial is planning to enter the market they should be focusing extensively on these states.

In [92]:
```
from matplotlib.pyplot import pie, axis, show
plt.figure(figsize=(15,5))


ax1  = plt.subplot(1,2,1)
agg = mergeddf.Loan_Amount_000.groupby(mergeddf.As_of_Year).sum()
axis('equal');
plt.pie(agg, labels=agg.index,autopct="%1.1f%%");
plt.title("Market Share by Year", fontsize = 20)


ax2  = plt.subplot(1,2,2)
agg = mergeddf.Loan_Amount_000.groupby(mergeddf.State).sum()
axis('equal');
plt.pie(agg, labels=agg.index,autopct="%1.1f%%");
plt.title("Market Share by State", fontsize = 20)
plt.show()
```

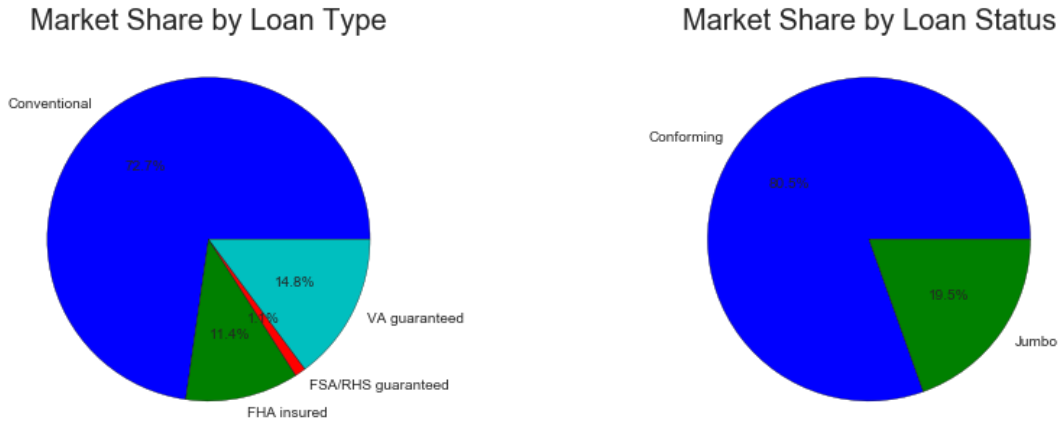


The below two pie charts are visualization of market share by Loan type and loan status, as we can see 70% of market share is in conventional loan types.

In the second pie chart the conforming status loans take the major share of 80% and jumbo loans constitute to 20% which is visibly large even though the total number of jumbo loans are not more than 5% of all the loans, this implies jumbo loans are crucial for business because of their magnitude. So more exploration needs to be done on the jumbo loan staus.

In [93]:
```
plt.figure(figsize=(15,5))
ax1  = plt.subplot(1,2,1)
agg = mergeddf.Loan_Amount_000.groupby(mergeddf.Loan_Type_Description).sum()
axis('equal');
plt.pie(agg, labels=agg.index,autopct="%1.1f%%");
plt.title("Market Share by Loan Type", fontsize = 20)


ax2  = plt.subplot(1,2,2)
agg = mergeddf.Loan_Amount_000.groupby(mergeddf.Conforming_Status).sum()
axis('equal');
plt.pie(agg, labels=agg.index,autopct="%1.1f%%");
plt.title("Market Share by Loan Status", fontsize = 20)
plt.show()
```



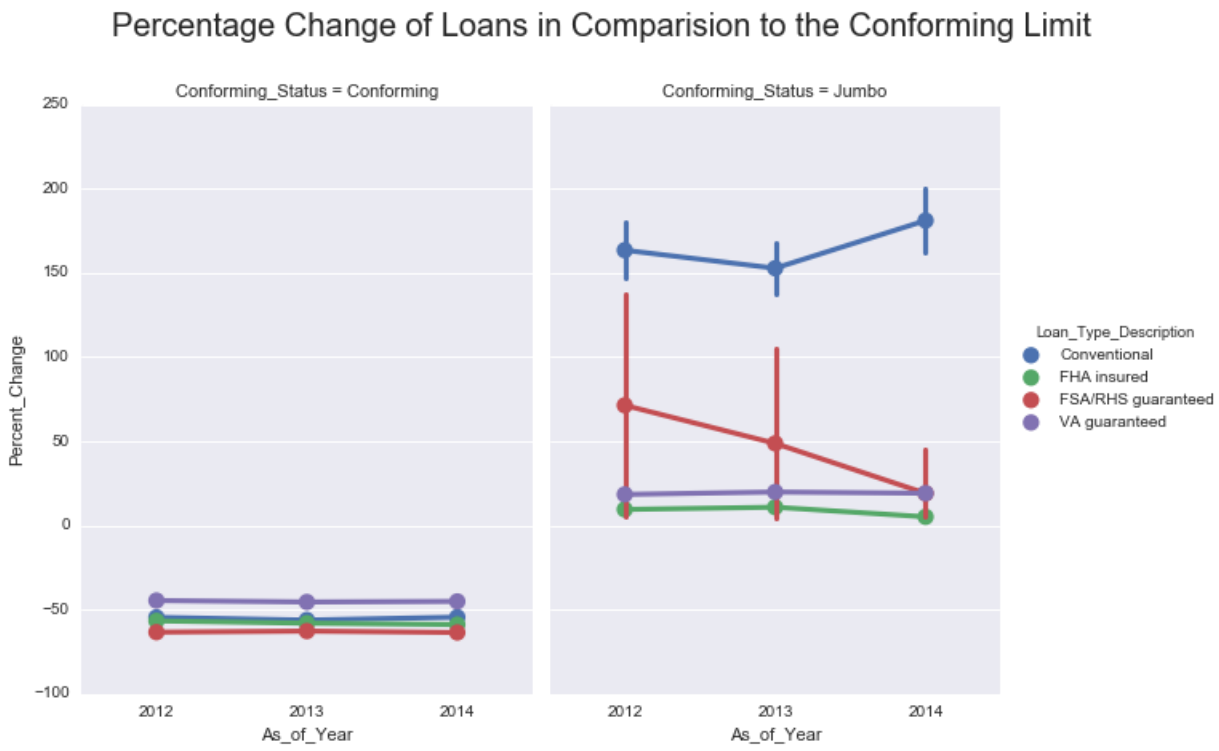
As part of digging deeper into the jumbo types of loans, creating a new column 'Jumbo_Percent' in the dataframe, which will have the information of how much percent are the jumbo loans higher than the conforming limit

In [94]:
```
mergeddf['Percent_Change']=(mergeddf.Loan_Amount_000-mergeddf.Conforming_Limit_000
#mergeddf.columns
```

In the below plot we are trying to visualize how huge are the jumbo loans using the column created above. In the below plot we can consider the x axis to be the conforming limit, and the below plots visualize the average of each loan types through the years. The plot on the left describes what percentage less are the confriming loans than the limit year over year and it seems to be constant showing market stability. The second plot on the right visualizes how big the jumbo loans are in terms of percentage than the conforming limit. As we can see the jumbo loans for conventional type of loans starts at 150% higher than the conforming limit and the trend seems to be only increasing after a slight drop in 2013.

In [95]:
```
ax=sns.factorplot(x='As_of_Year',y='Percent_Change',hue='Loan_Type_Description', c
                  capsize=.2, size=6, aspect=.75) plt.suptitle('Percentage Change
of Loans in Comparision to the Conforming Limit',y
```

Out[95]: <matplotlib.text.Text at 0xb674b3c8>



Note: X axis at zero is the conforming limit.

From the above plot we can confirm that the market performance is stable over the years for conforming type of loans and the jumbo type of loans the average percentage is only increasing which is a strong sign of good market performance.

TASK 4:

Recommendations:-

The below points are major recommendations from the above visual analysis.

- The loan amount and applicant income rates have a similar kind of distribution which entrusts confidence in prompt repayment of loans by the applicant, and also 80% of loans fall under first lien this ensures safety for the sanctioned loans.
- The market share pie charts give away a lot of information and clarity regarding the loans, The "Market Share by Year" plot shows that the loans aggregate is decreasing year over year by 10 % which is not a very good sign for entering into the business.

- The "Market Share by State" plot gives clarity about majority of market share in the states 'VA' and 'MD' on which 'Change Financial' has to focus more rather than on 'DC' which has highest number of loans recorded in the available data.
- The "Market Share by Loan Status" pie chart displays the importance of jumbo loans even though the jumbo loans constitute to less than 5% of the total data. The 'Percentage Change of Loans in Comparison to the Conforming Limit' plot emphasis on the magnitude of these jumbo loans and how important it is for business, the plot also shows

that the loans less than the conforming limit have also performed with stability throughout the years.

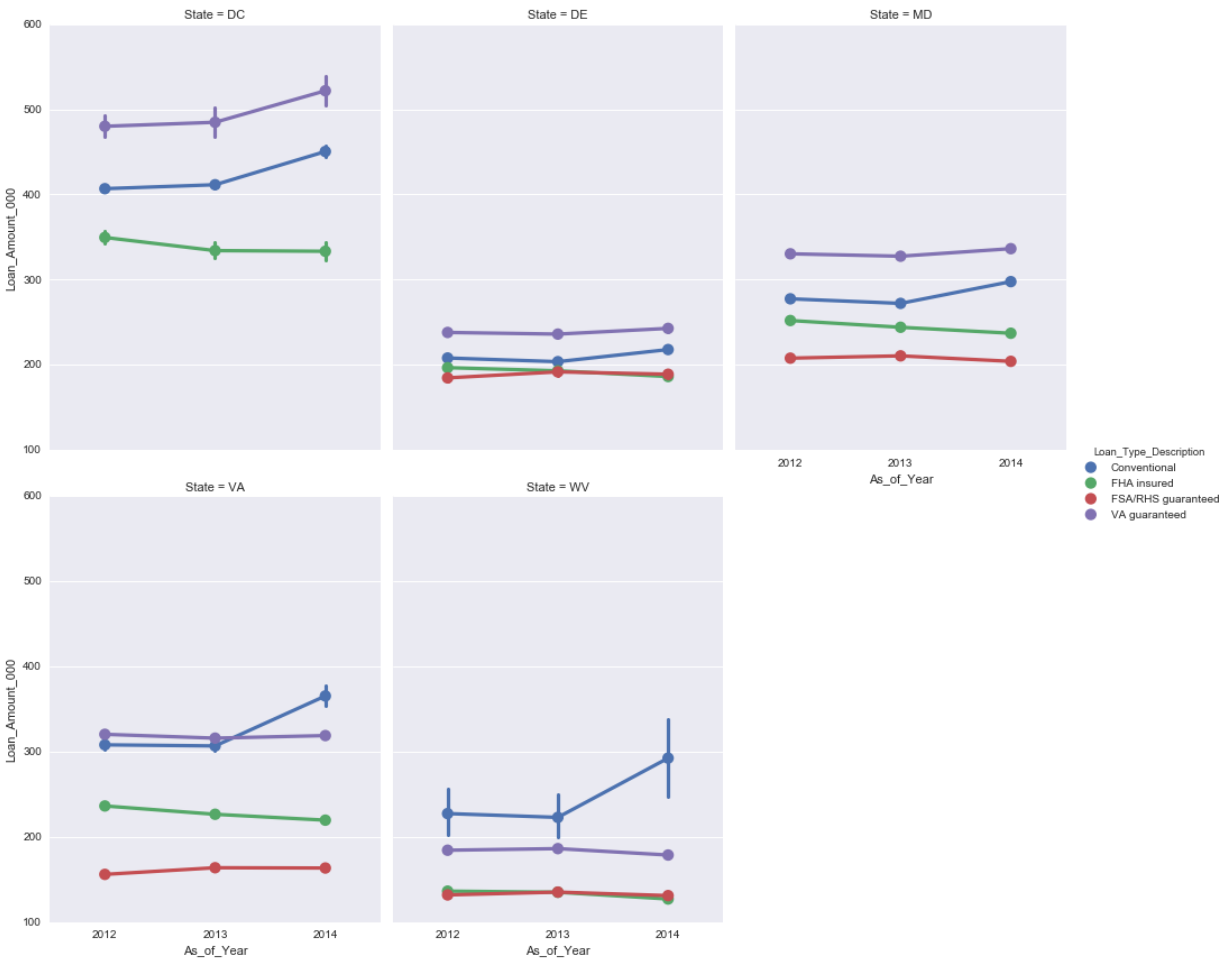
Final Conclusion :-

I think it is hard to come to a formidable conclusion based on the data available, as the data at hand is the loans application data and there is no information of how many loans were approved of all the applied loans. The results from the above visualizations are contradictory. The majority of plots show strong increase in the average of loan amounts through the years, but the market share for loans is decreasing at a rate of 10% each year as indicated by the 'Market Share by Year' plot, this cast doubts in making a clear and concise decision. A formidable decision can be taken, if additional data about approved loans is provided.

Additional Analysis:

The below plot gives the details about which loan type 'Change Financial' should focus on majorly in each state.

In [96]:
```
ax=sns.factorplot(x='As_of_Year',y='Loan_Amount_000',hue='Loan_Type_Description',
                  capsize=.2, size=6, aspect=.75)
```



From the above plot it is clear that the focus should be more on conventional types as they are significant in almost all the states and very important in the states of 'VA' and 'WV' also the pie chart in the task 3 shows that the 70% of amount comes from Conventional loan types.

The VA Guaranteed loan types are also very crucial in all the states and are also leading in terms of average in the states "DC","DE" and "MD".

Also Change Financial should lay special focus on Conventional Jumbo loans and have promotional offers on that category as the magnitude of jumbo status loans is huge for conventional loan types.