Mortgage Loans Analytics

- Loan Type: to indicate the loan if fixed rate, or balloon loan , or ARM, Loan program type: to indicate conforming loan, FHA/VA loan, Jumbo loan or sub-prime loan
- Current Interest rate:
- FICO Score: the updated fico score
- LTV: the current loan to value ratio
- Loan Size: the loan amount of the loan
- Loan origination location (City & Zip)
- Unit Types (Types of property)

Predicting mortgage demand using machine learning techniques

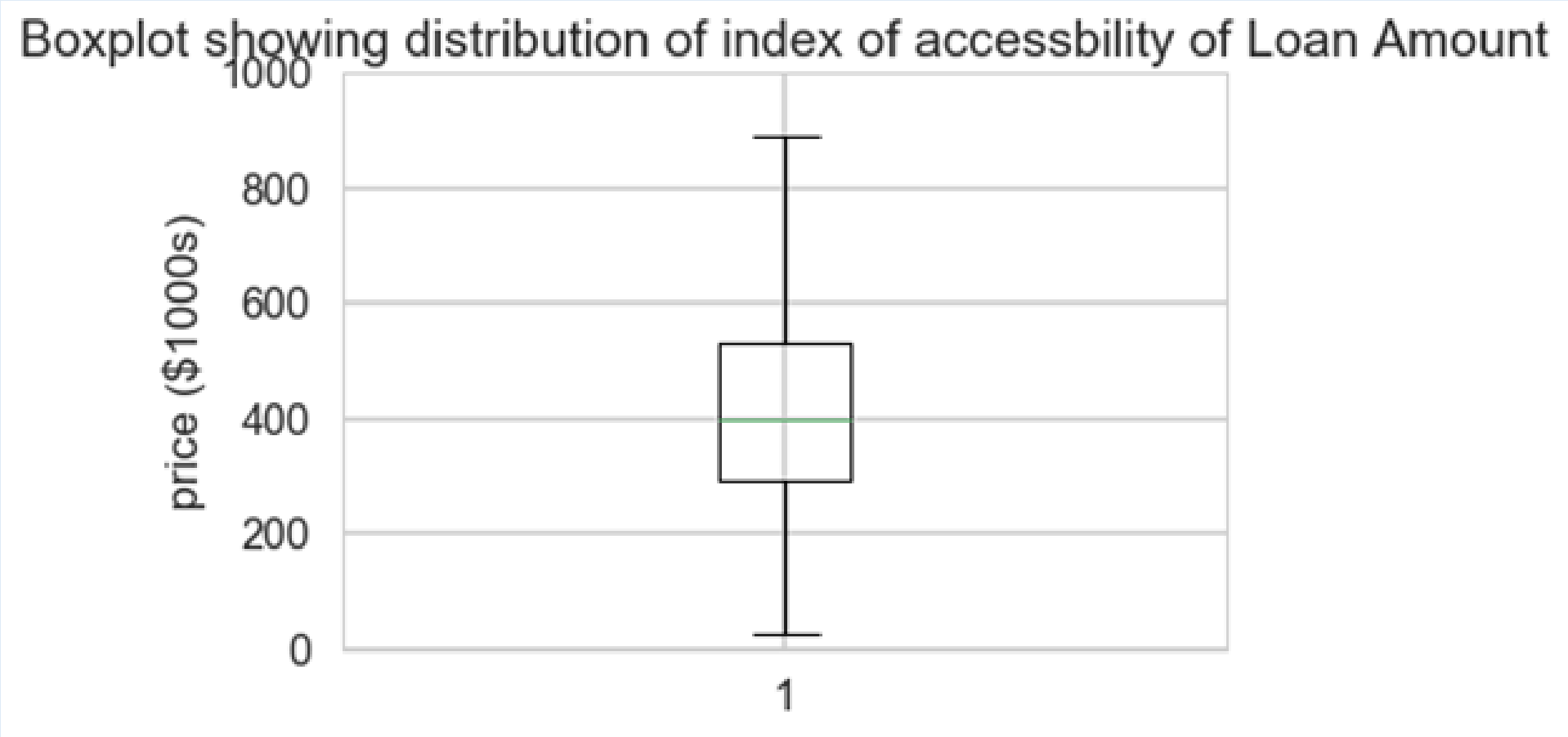
- ***Linear Regression,***
- ***Logistic Regression***
- ***Random Forests (RF)***
- ***Support Vector Regression (SVR)***
- ***Support Vector Machine (SVM)***
- ***k-Nearest Neighbors***
- ***Decision Tree Classifier***

Data Preparation

- The goal of the Data Preparation stage is to transform and enrich the dataset so that it can be fed into the models. After the data is collected and explored, it can be pre-processed so that it can be used

- directly in our predictive model. With the pre-processed data one can perform feature engineering

Distribution of Loan Amount

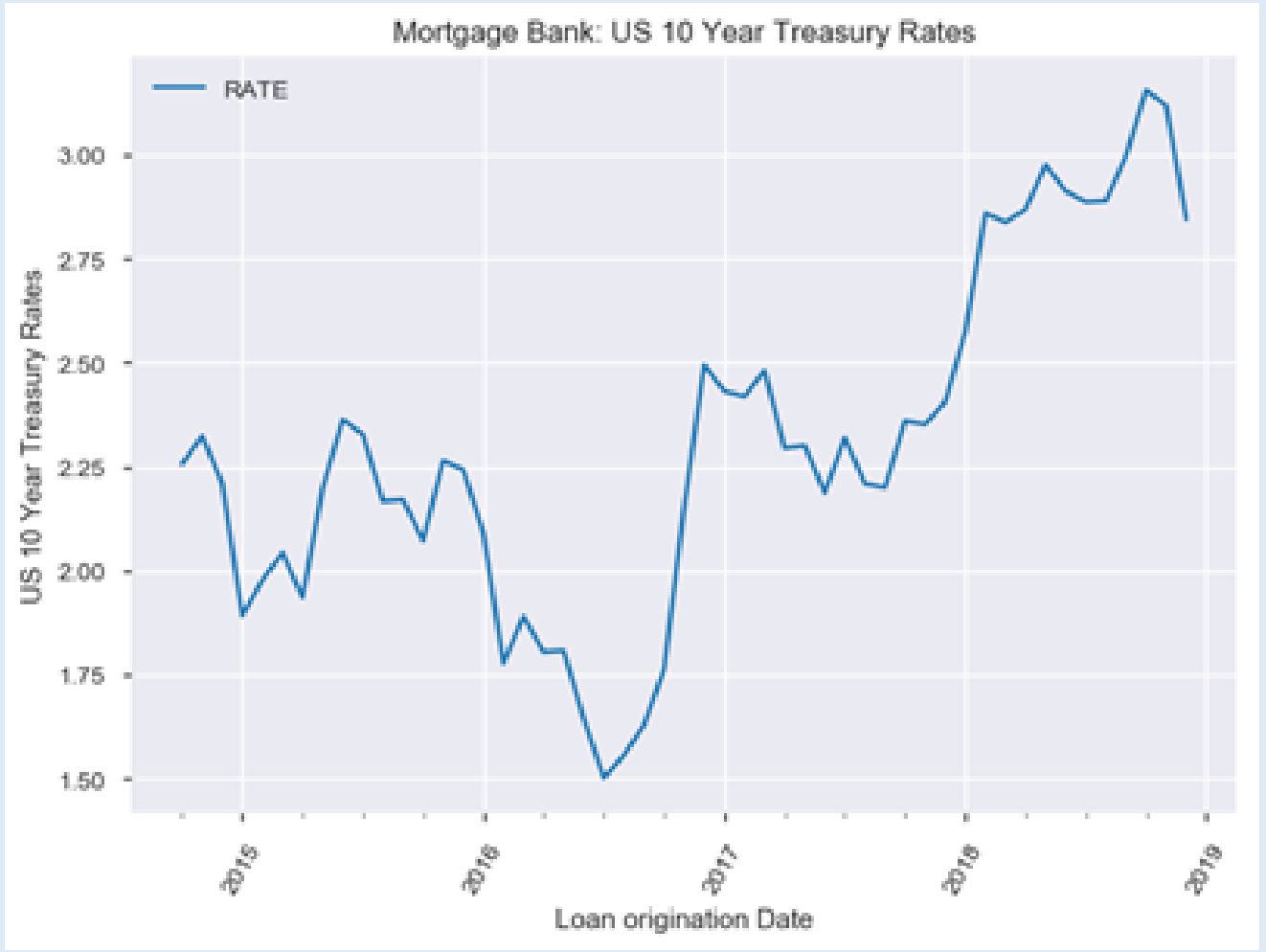


Mortgage interest rates

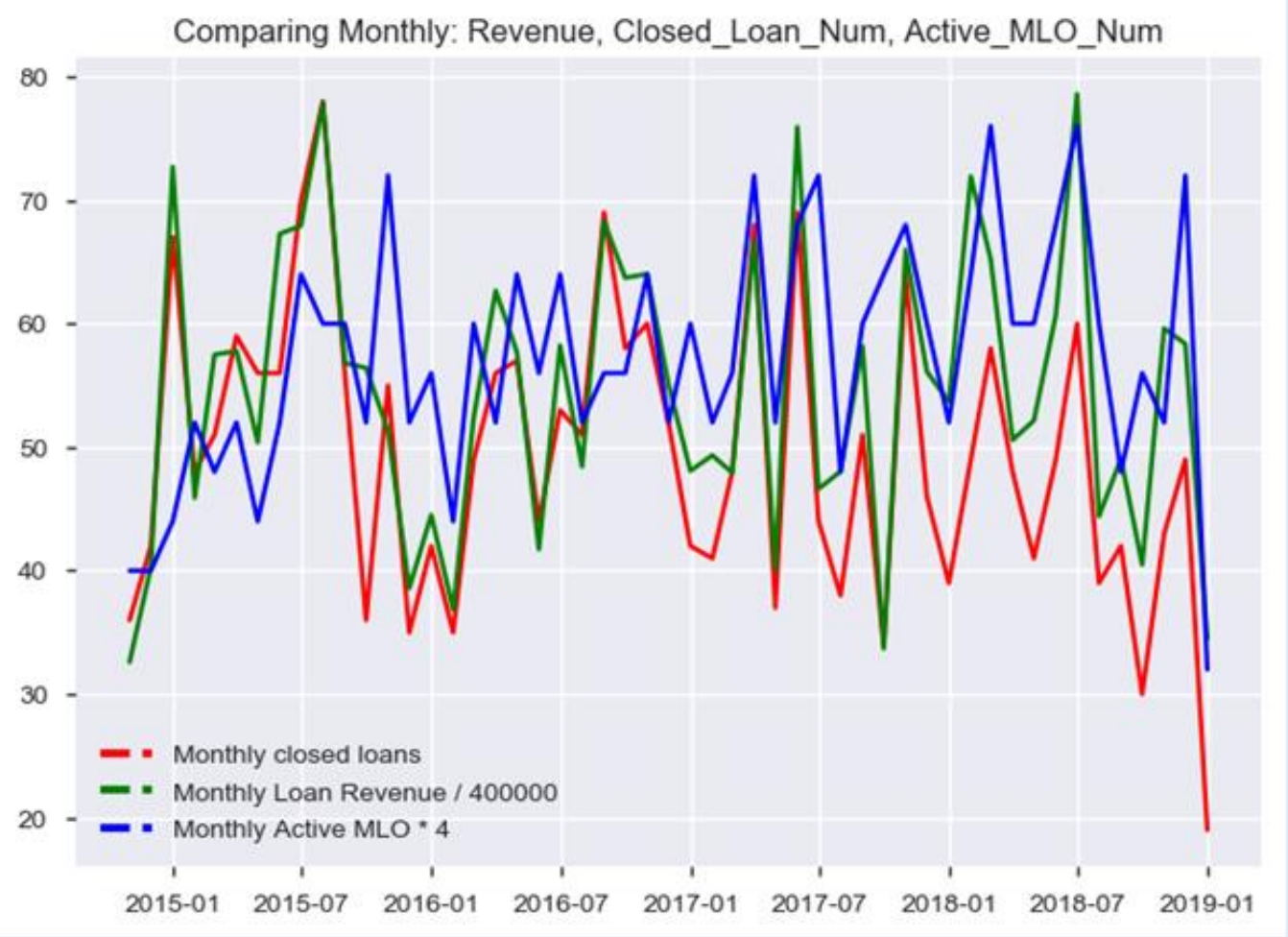
- Mortgage interest rates have a significant impact on the amount of mortgage applications. If the interest rates are low, the mortgages are relatively cheaper for the borrower as they have to pay less interest, which leads to an increased amount of mortgage applications.
- A high mortgage interest rate means the mortgage borrower pays a high amount of interest to the lender, which makes the mortgage less attractive for the borrower.

Interest Rate :

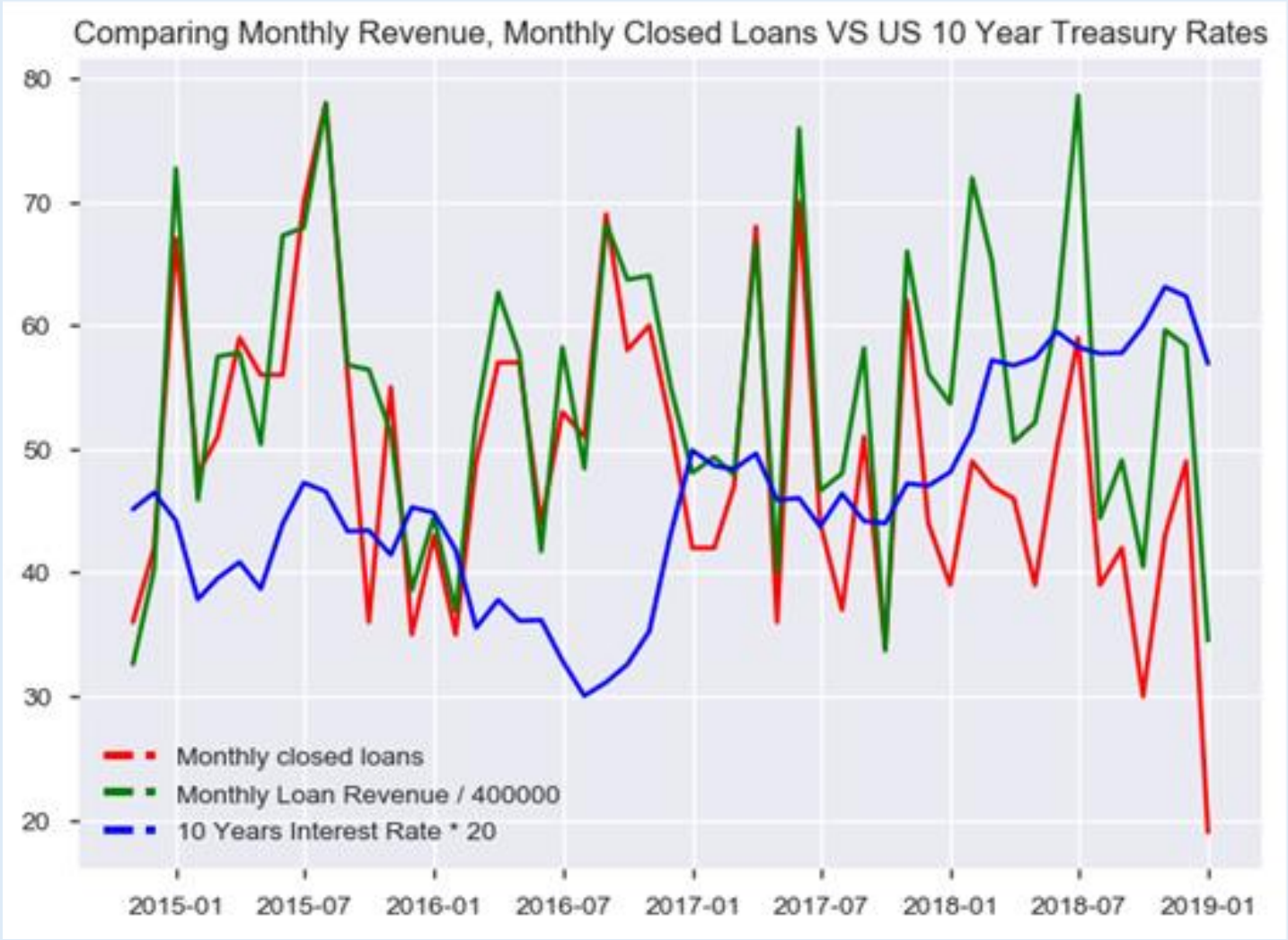
In the short run rate may go ups and down but in the long run rate will go up. As housing price goes up, interest rate will go up to control the housing price.



Monthly data of closed loan numbers , monthly revenue and active MLO numbers per months, all moving at the same direction.



Correlation coefficient between Closed Loans & loan Rev: .83
Correlation coefficient between Interest & loans Closed Data: -.33



Random Walk

Are Interest Rates or Monthly Loan Returns Prices a Random Walk?

- **Most returns prices follow a random walk (perhaps with a drift). We will look at a time series of Monthly Sales Revenue, and run the 'Augmented Dickey-Fuller Test' from the statsmodels library to show that it does indeed follow a random walk.**

- ***The p-value of the test on Loan Revenue is: 0.0001969041976389 6495***

According to this test, p-value is very low (lower than 0.05). We **reject** the hypothesis that monthly_loan_rev_data follow a random walk. '''

- ***The p-value of the test on Monthly Interest Rate is: 0.5839314748568241***
- According to this test, **p-value is very is higher than 0.05**. We **cannot reject** the hypothesis that Monthly **Interest Rate prices follow a random walk**.

Are Interest Rates Auto correlated?

When we look at daily changes in interest rates, the autocorrelation is close to zero. However, if we resample the data and look at annual changes, the autocorrelation is negative.

- ***The autocorrelation of annual interest rate changes is -0.97***

- We have notice how the daily autocorrelation is small but the annual autocorrelation is large and negative.

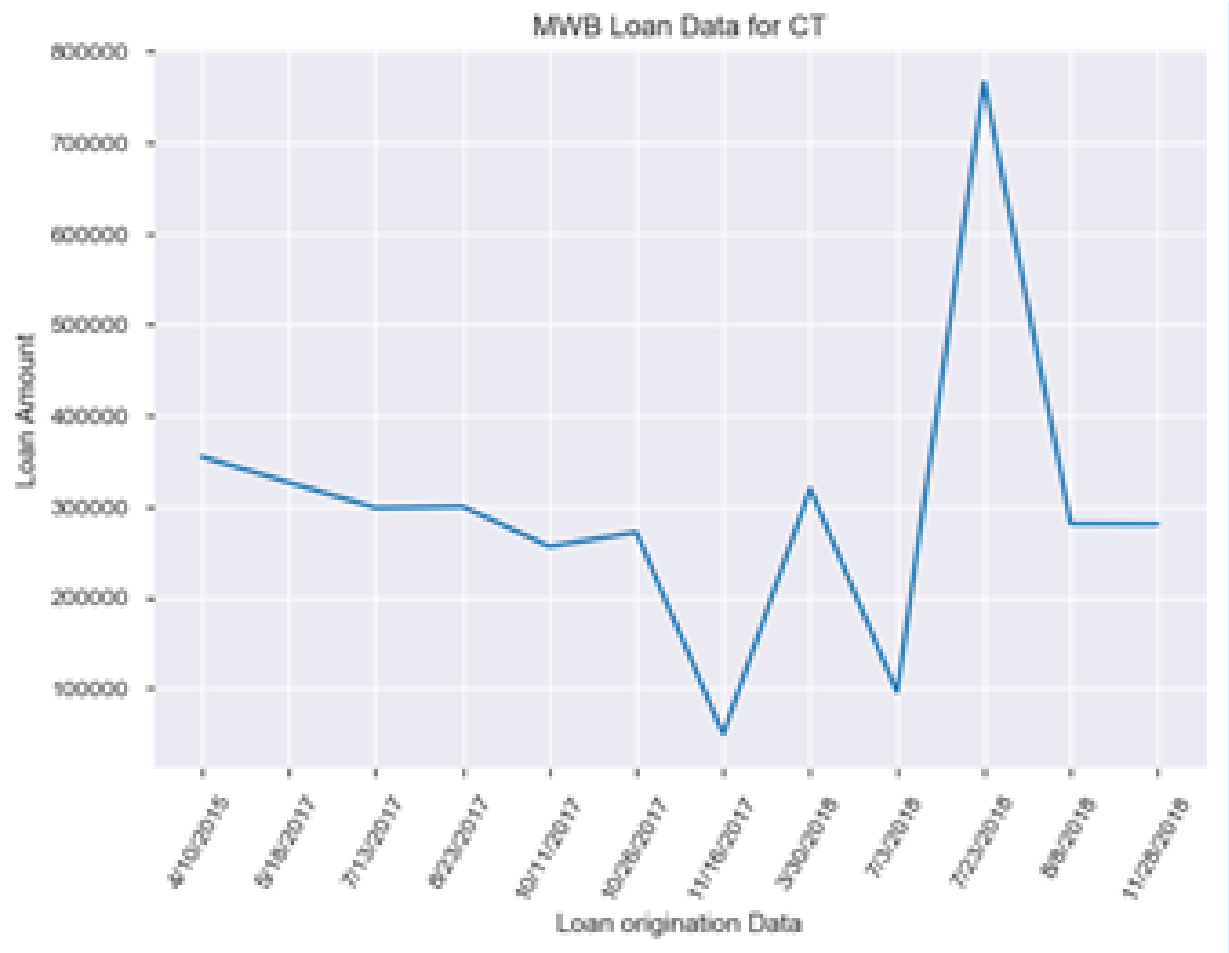
DATA COLLECTION

- Besides Mortgage Application DataSet, we have joined two separate (**10 Years US Treasury Rate, Home Supply Index)** with our existing Mortgage Application DataSet to enhance predictive power of our model.

- **Join two DataFrames model_data1 & US10Y save the results in model_data**
- **Integrated Monthly housing supply index data and merging with current dataset**

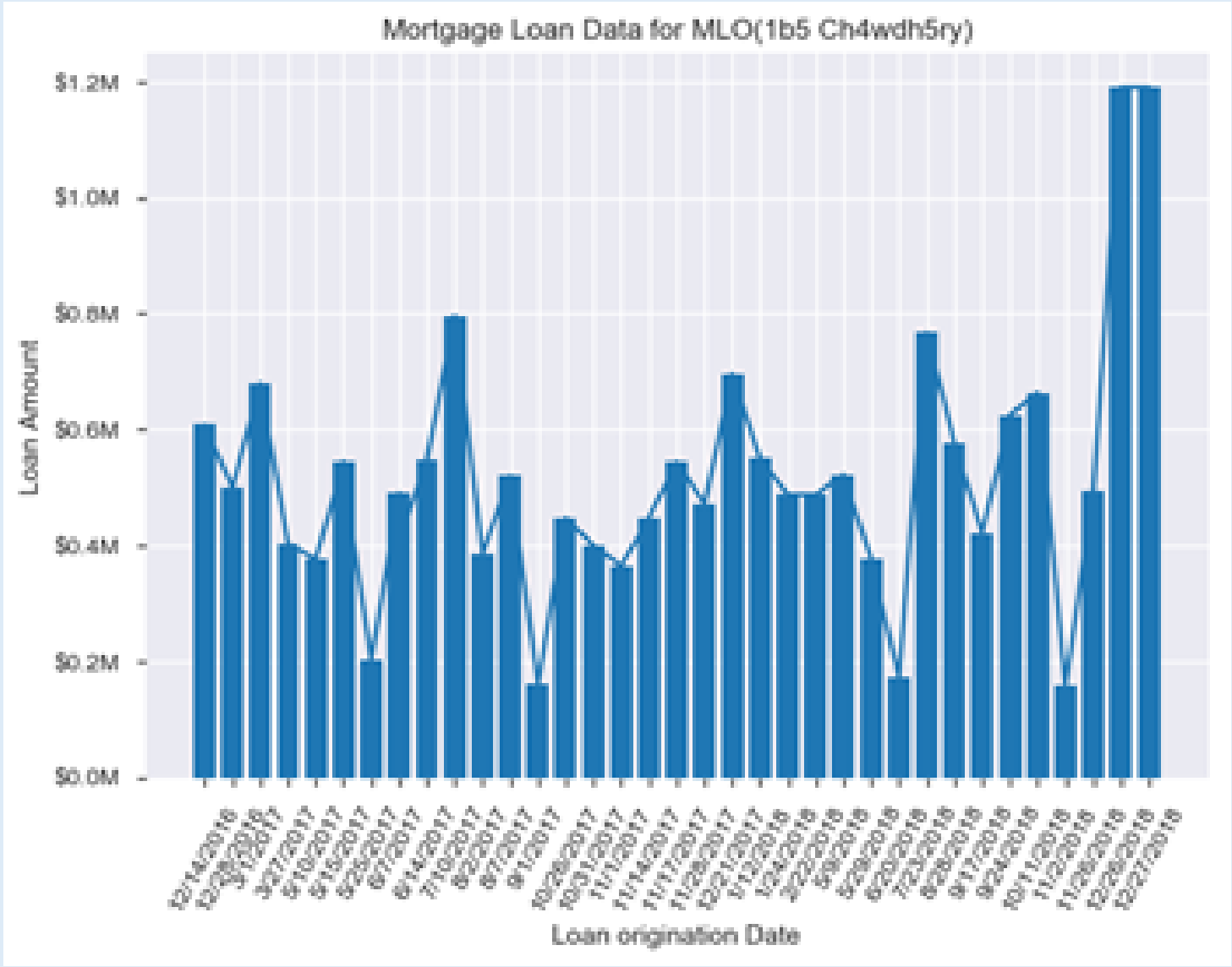
DATA EXPLORATION

Since our dataset can be grouped per day to create meaningful visualizations. The dataset contains data from October 2014 until December 2018.

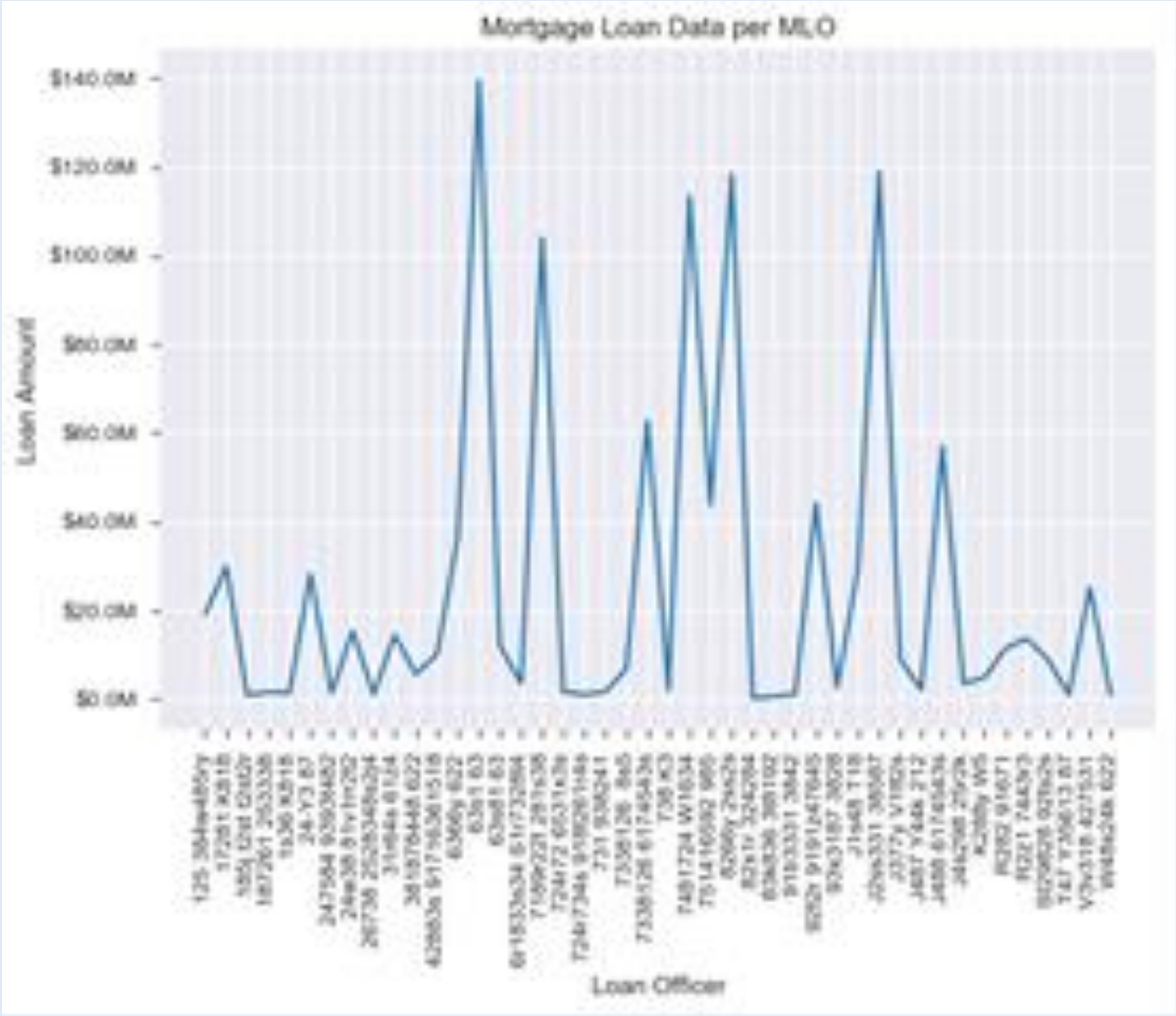


We can see that New York and New Jersey is the major Loan Origination market for the Bank.

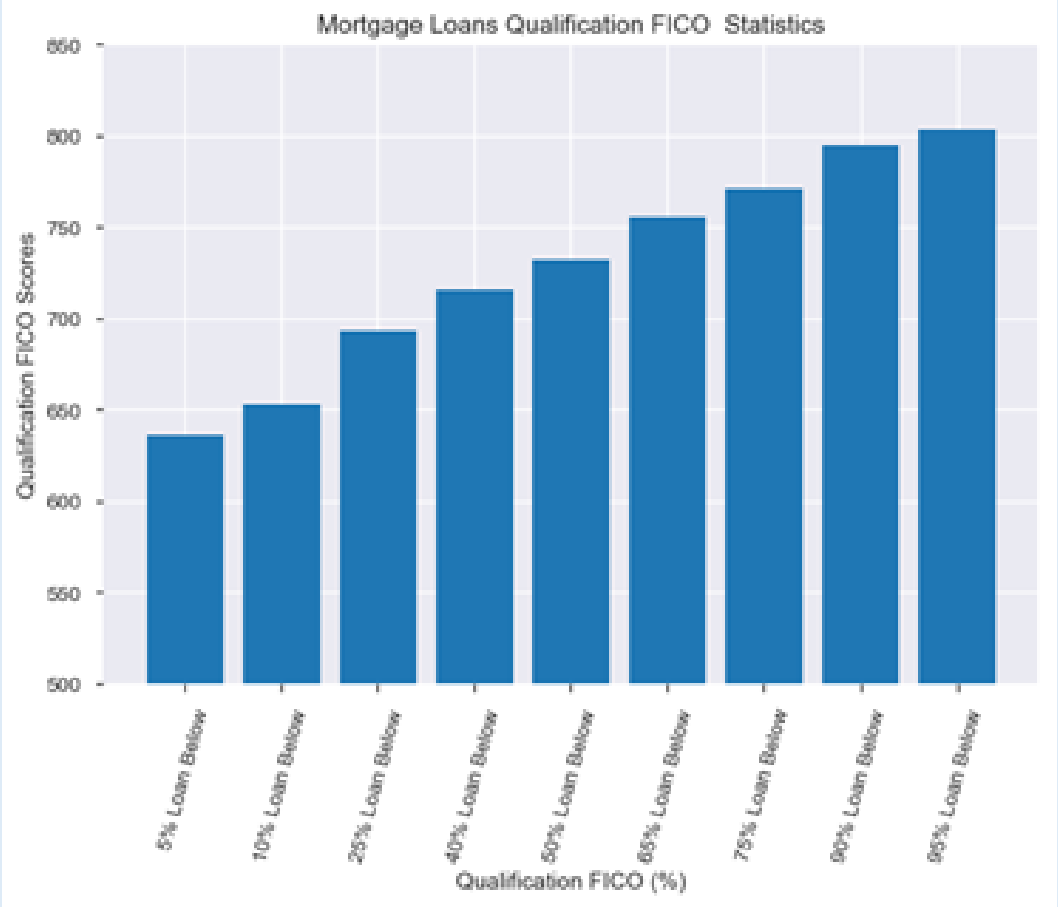
We can easily explore the Loan Origination history for particular MLO



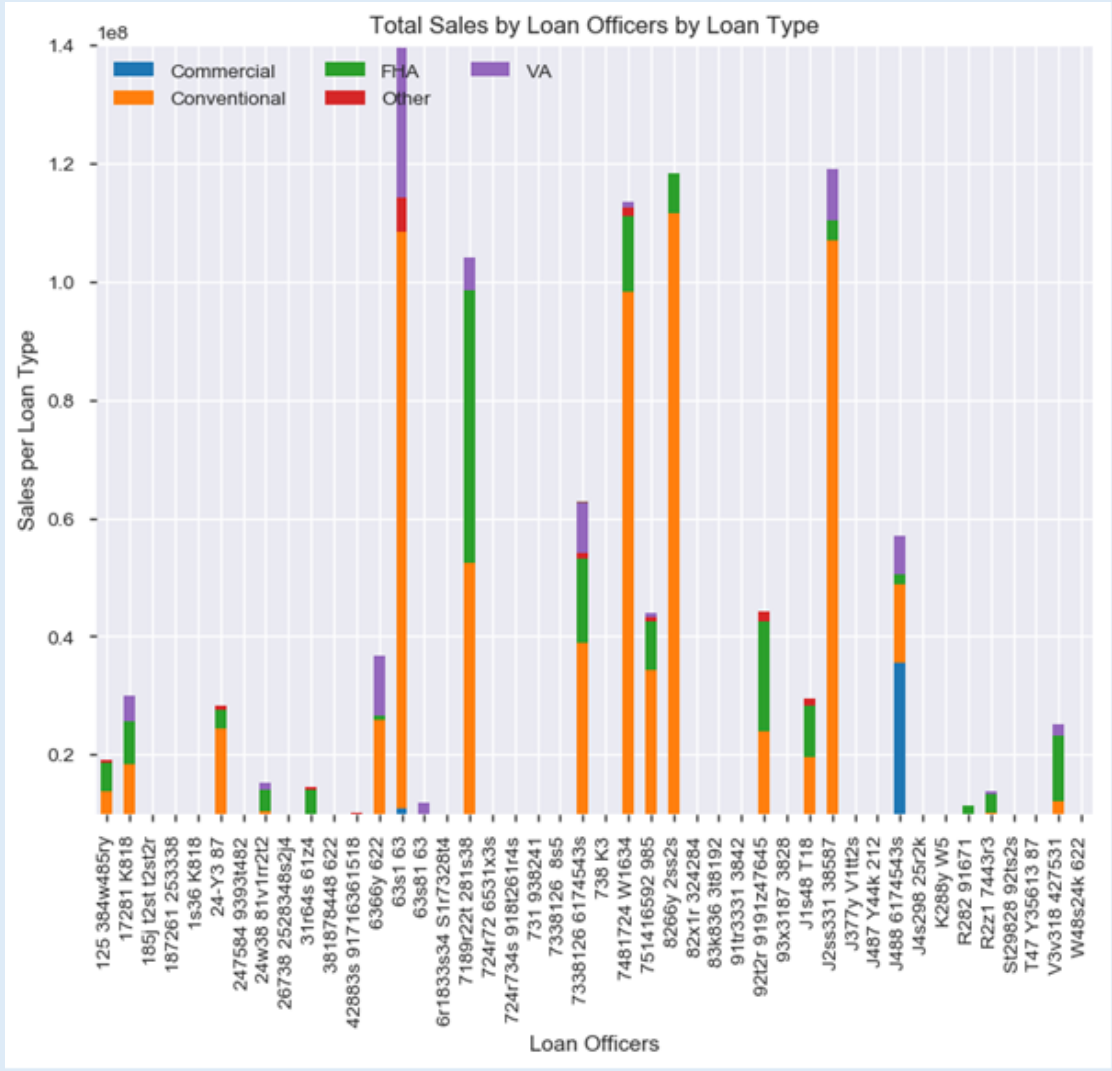
MLO Performance: Few MLO generates major portions of the sales revenue for the Bank, and many of the MLO is performing way below company standards.



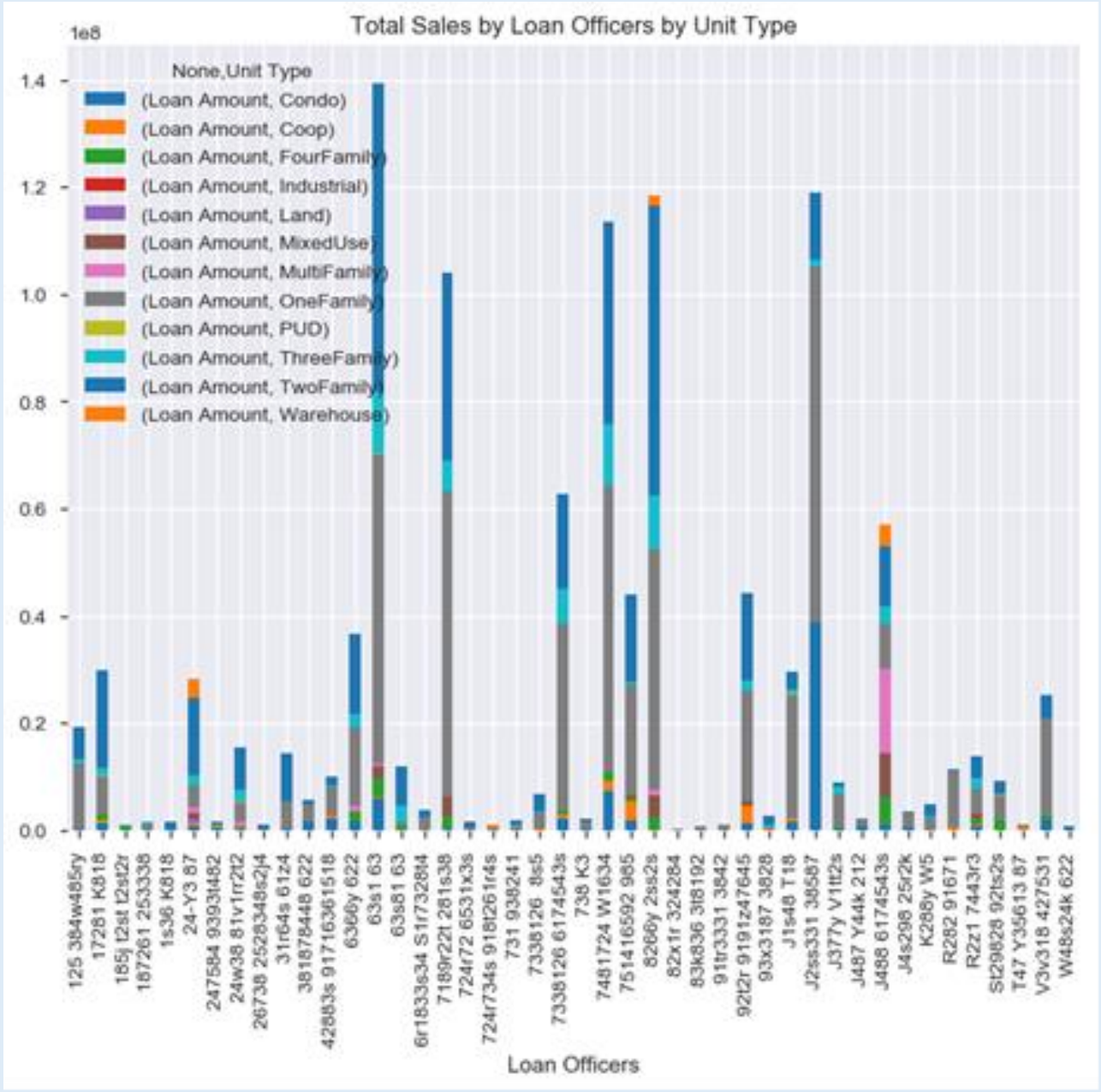
1. 90% of the FICO score is over 650. FHA Loan type could be only option with LOW FICO score;
2. Many bank uses cut-off points for FICO Score (640-680) for conventional mortgages.
3. FHA accepts FICO score below 600.



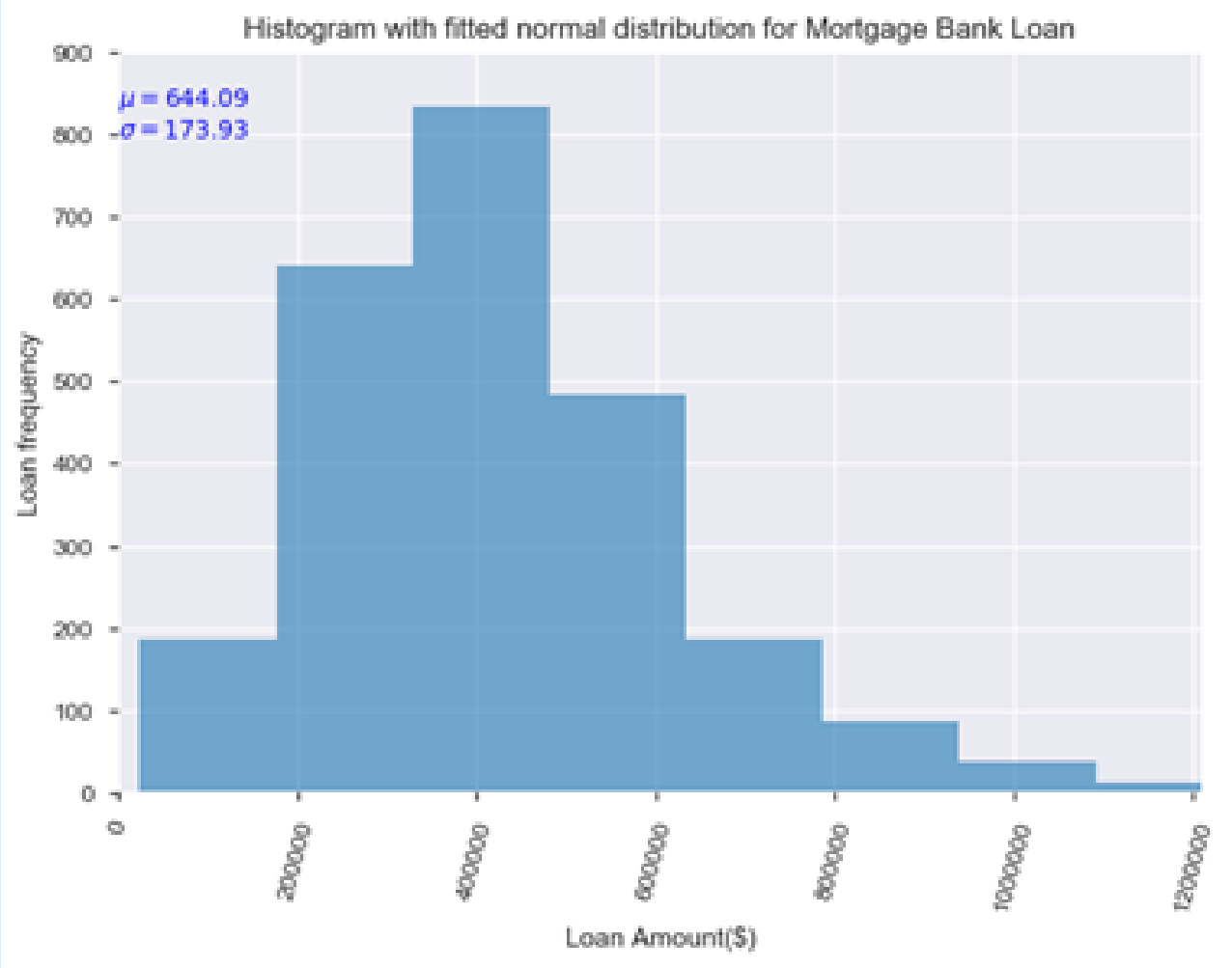
Let’s Analyze Mortgage Loan Officer’s Sales Volume per Loan Type. Processing different loan types need different expertise.



Similarly, we can explore MLO's Sales Volume per Unit Type.

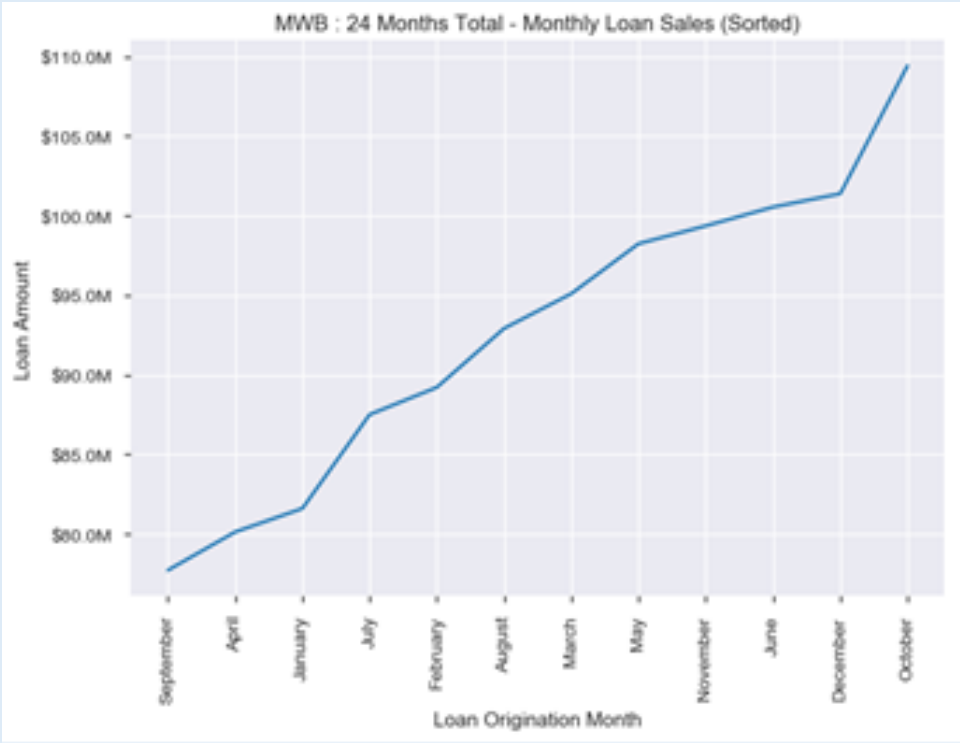


A histogram allows us to group purchases together so we can see how big the customer transactions are.

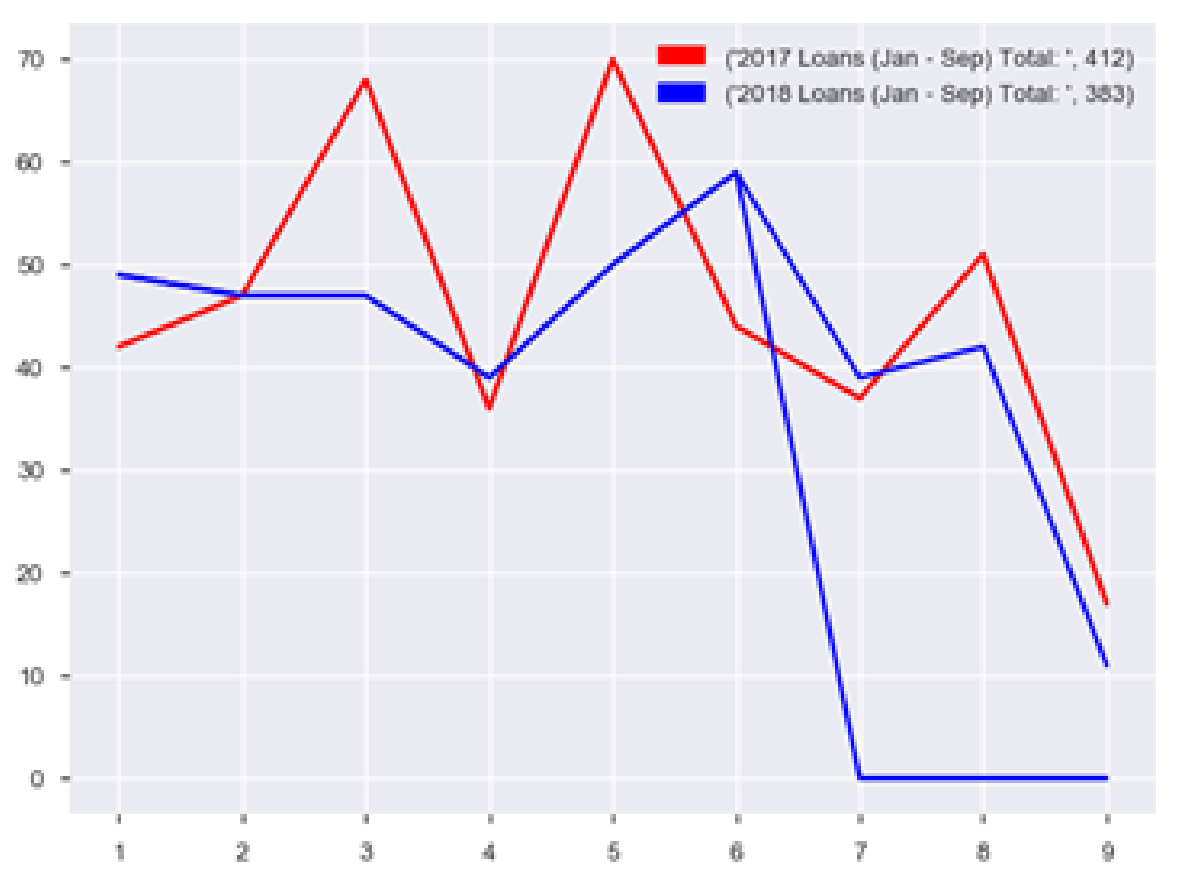


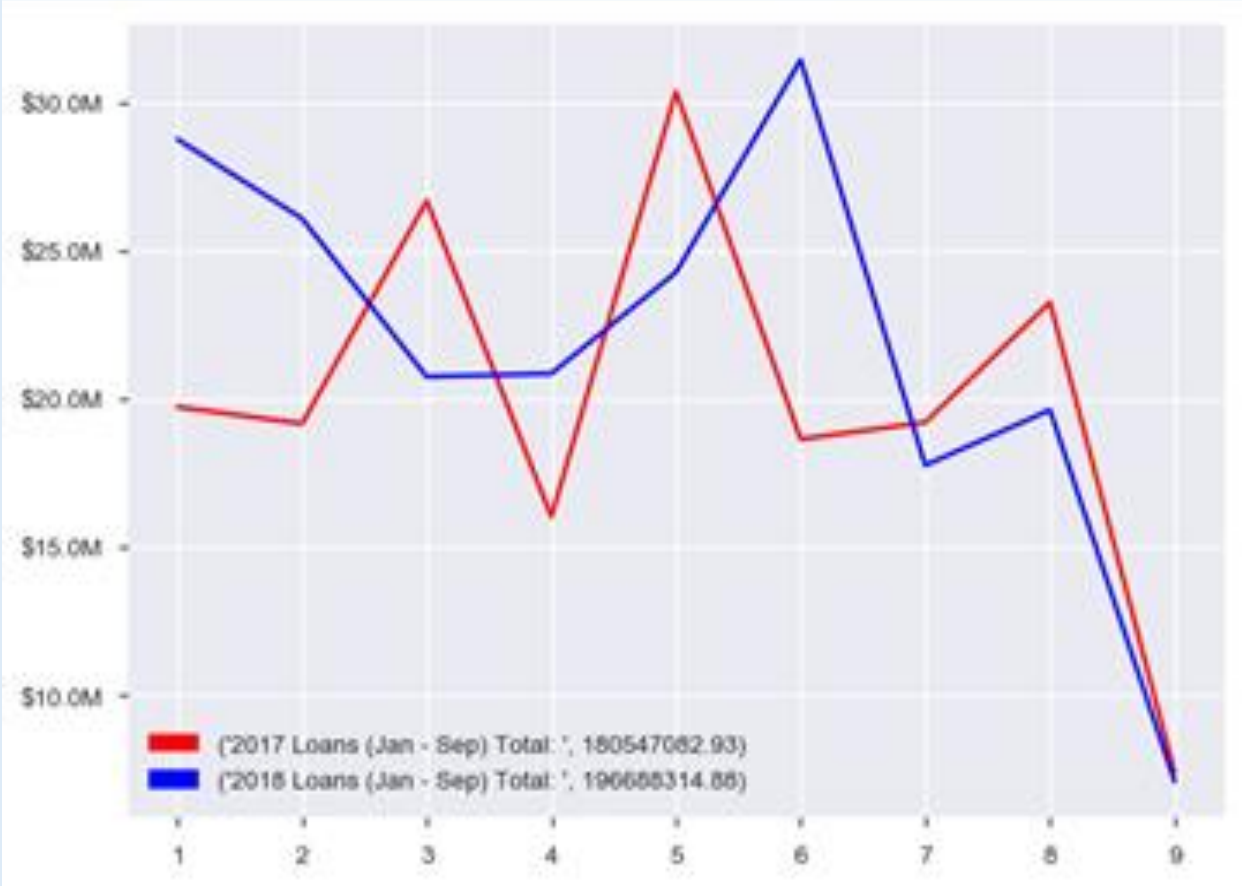
- **mean : 644.087**
- **var : 30252.443**
- **skew : 1.6231**
- **kurt : 10.20**

We can see that monthly mortgage loan sales volume varies between $15M and $32M. Another interesting find is Loan sales are at the peak during summer seasons.



Comparing year-over-year Mortgage Bank Loan performance (Loan Application & Revenue) :





Monthly Loan numbers and Monthly Sales volumes are strongly positively correlated.

- **Pearson correlation coefficient between**

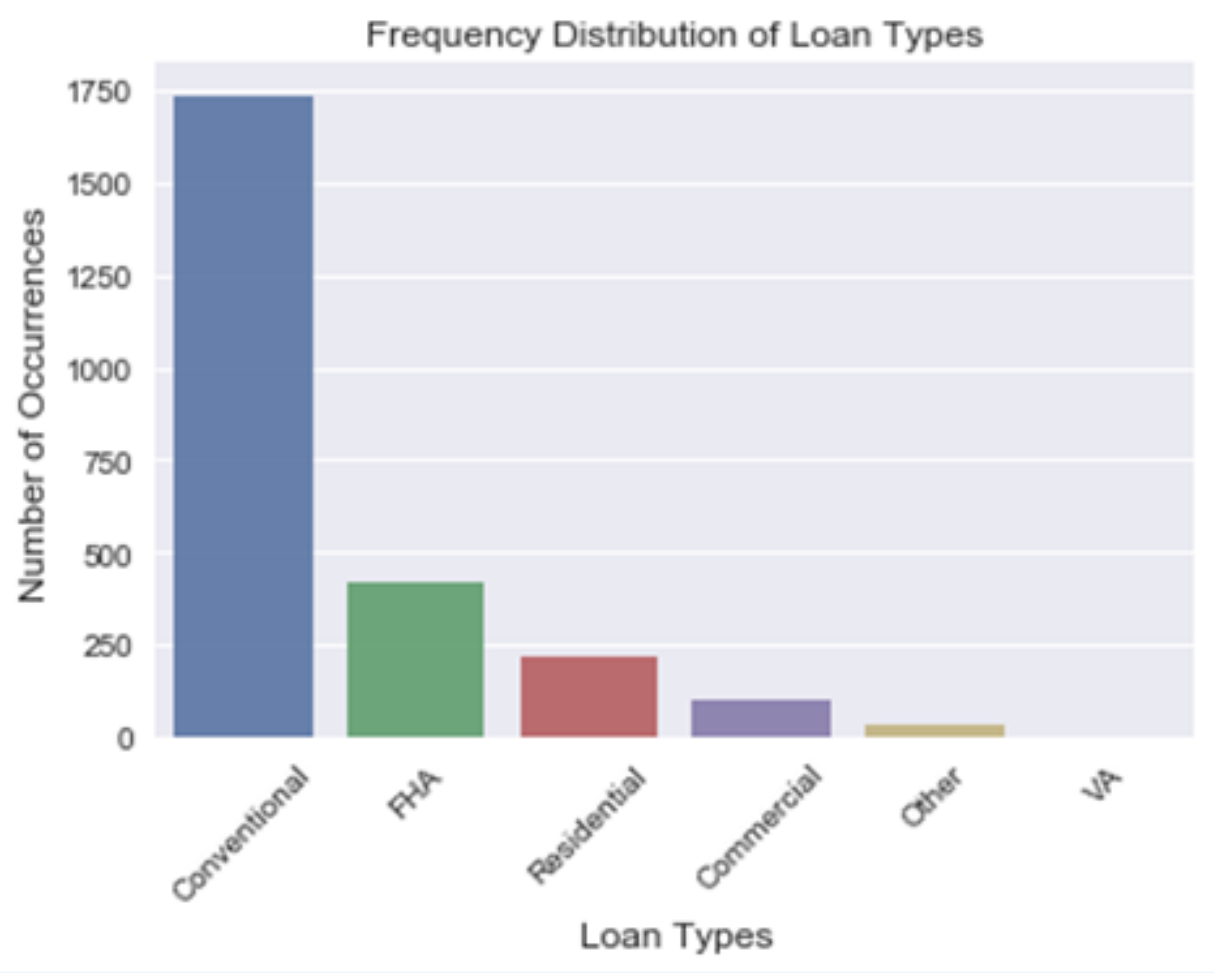
  Monthly Closed Loans

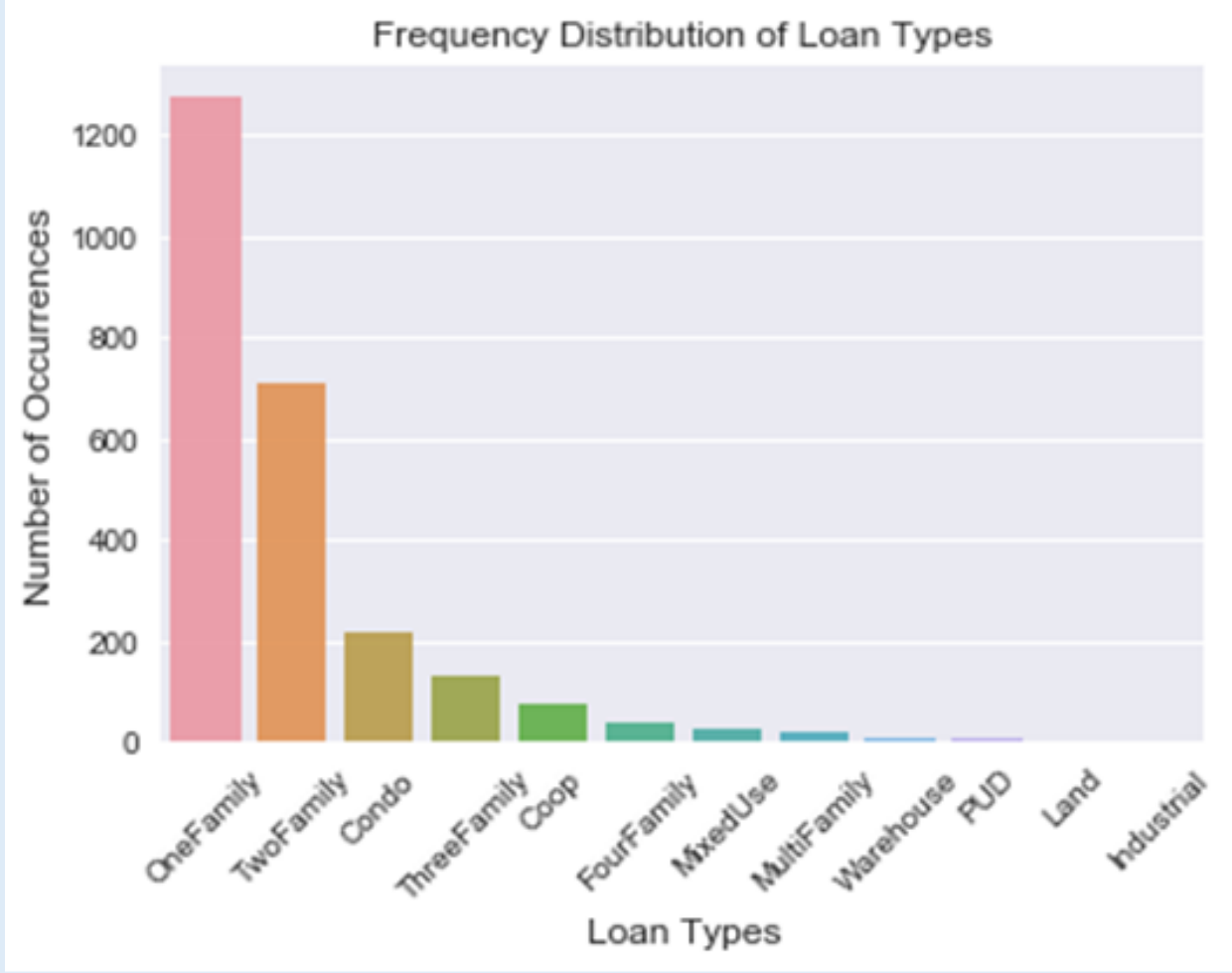
  and

  Monthly loan evenue is:

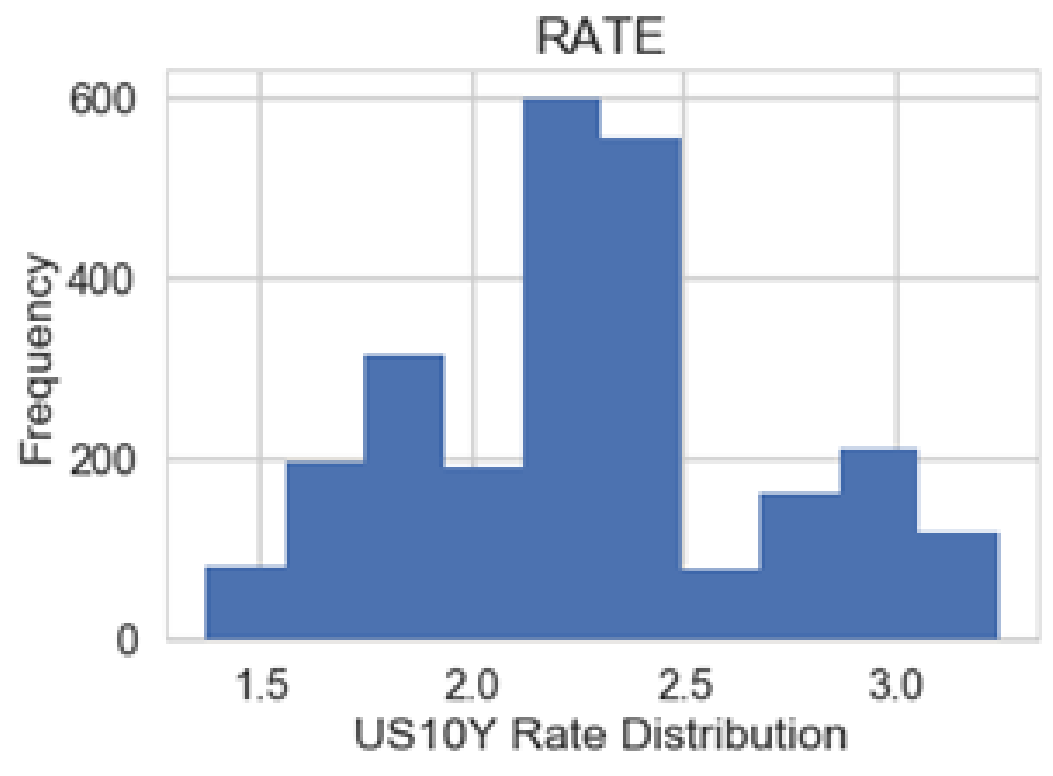
  0.83

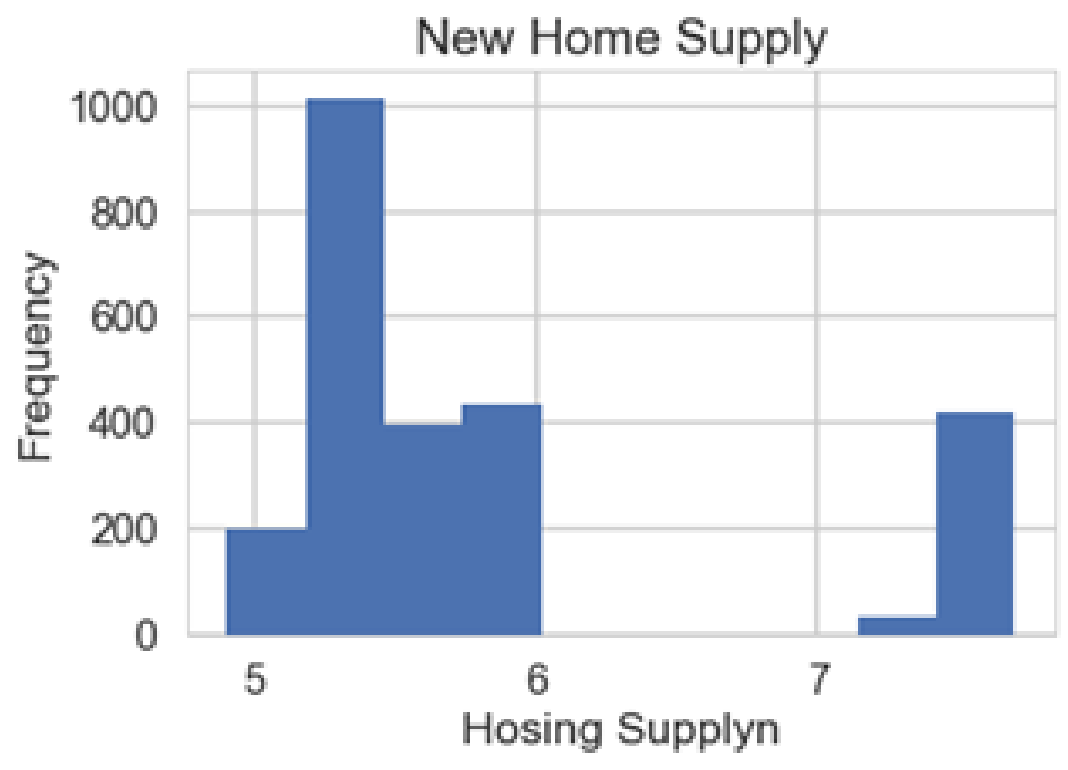
Visual exploration is the most effective way to extract information between variables



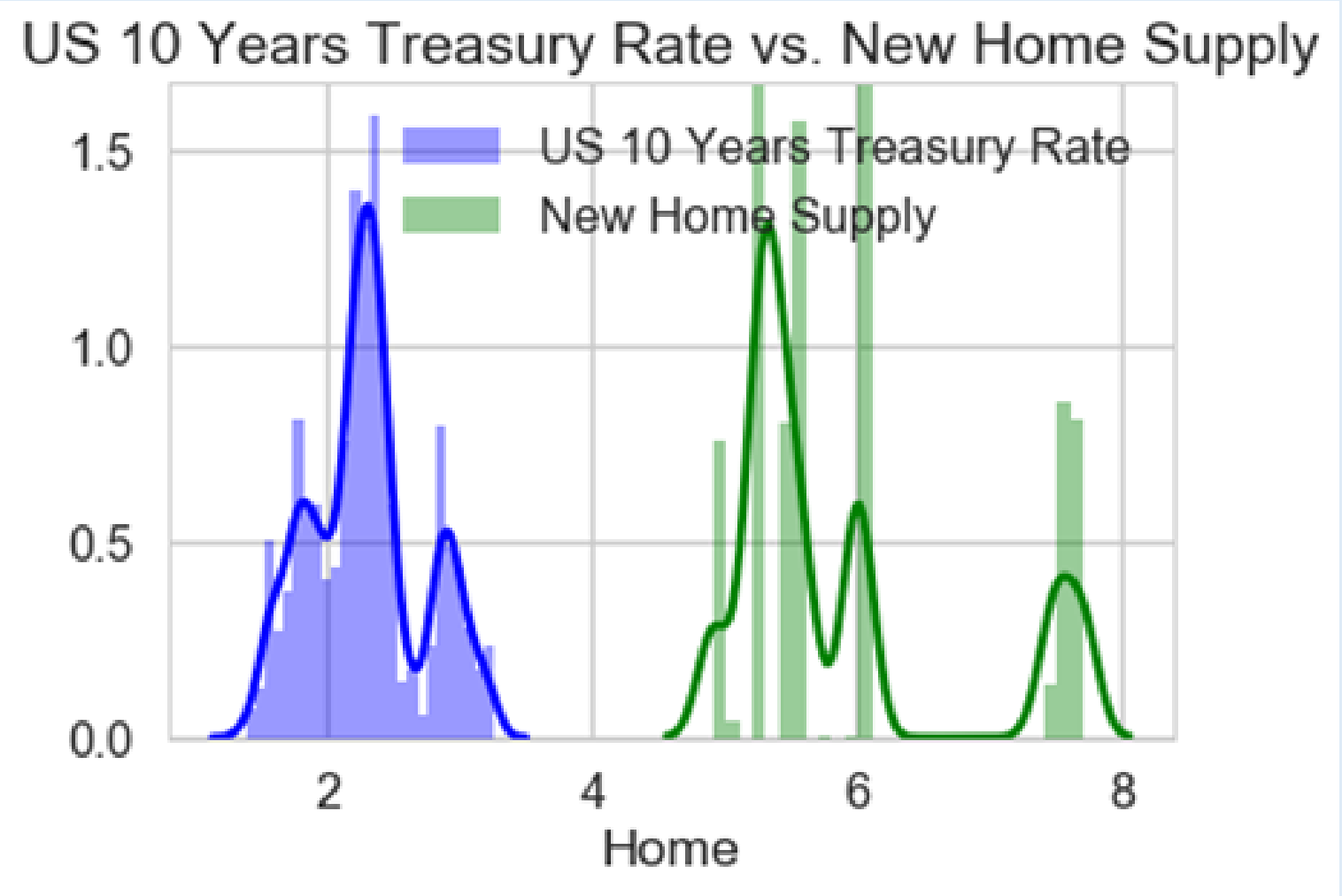


The distribution plot comparing US 10 Years Treasury Rate & New Home Supply shows that US 10Y RATE is normally distributed





Home Supply goes up, Interest Rate goes up
New Home Supply is skewed to the right.



DATA PREPARATION

The Data Preparation stage contains three elements: data pre-processing, feature engineering and feature selection.

Encoding Categorical Data

There are different techniques to encode the categorical features to numeric quantities.

The techniques are as following:

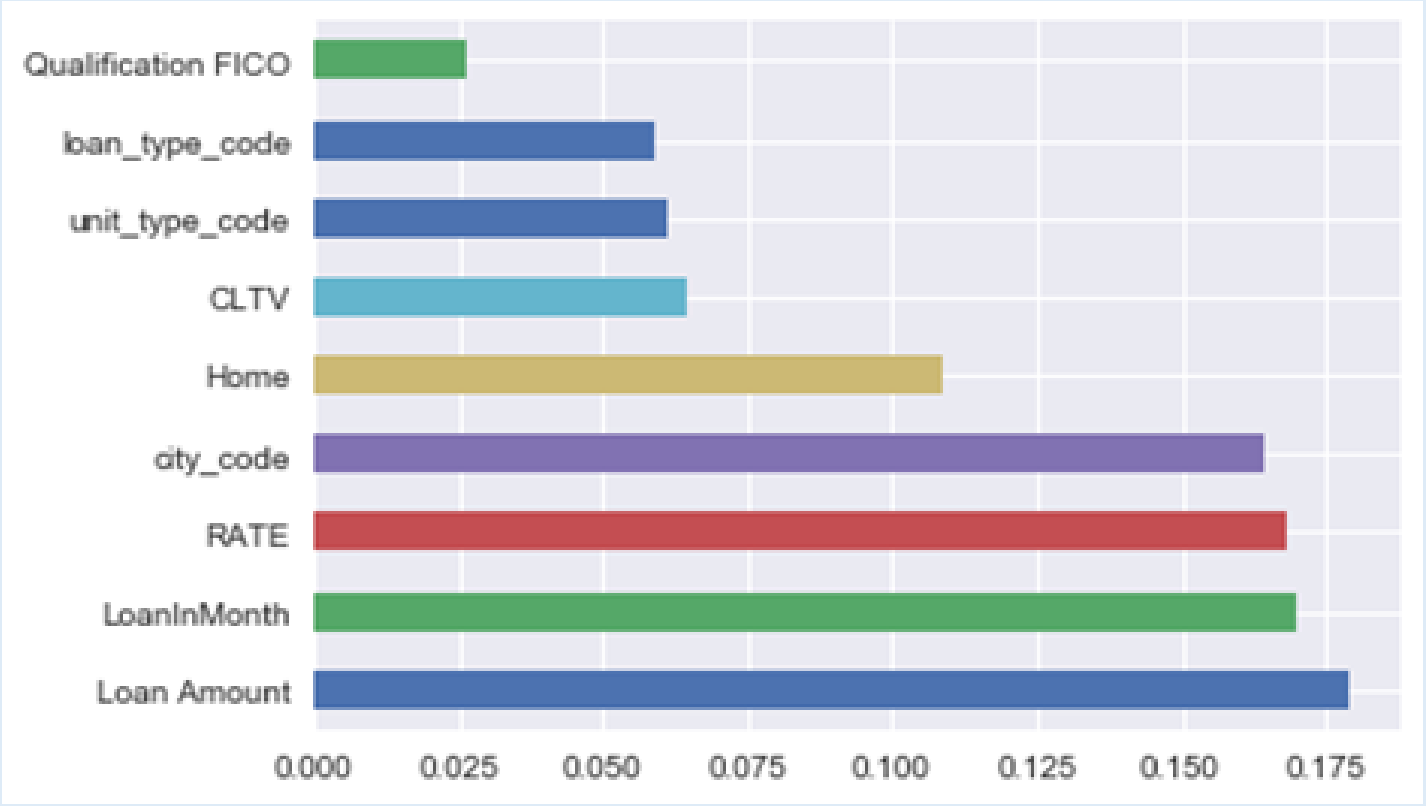
- **Replacing values**
- **Encoding labels**
- **One-Hot encoding**

DATA PRE-PROCESSING

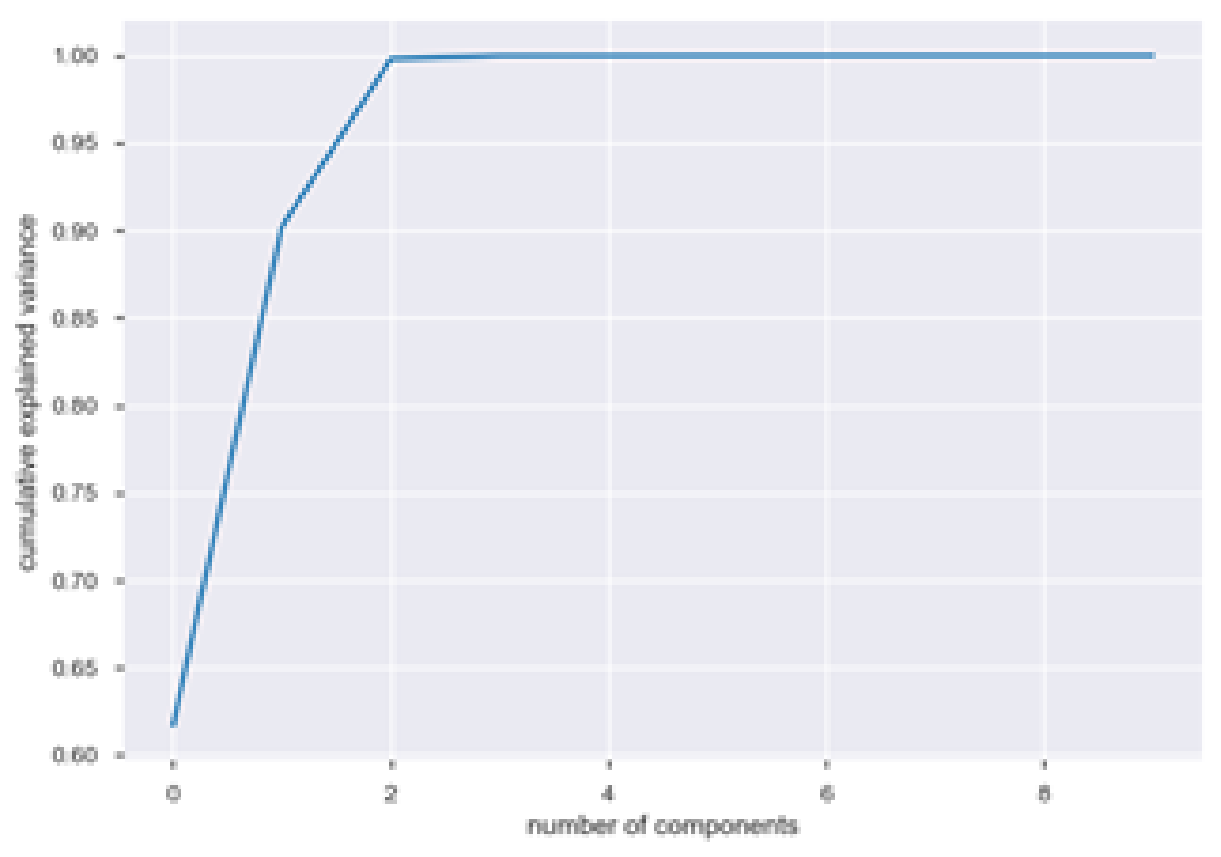
In the data pre-processing phase we pre-process our data to a suitable format for our predictive model.

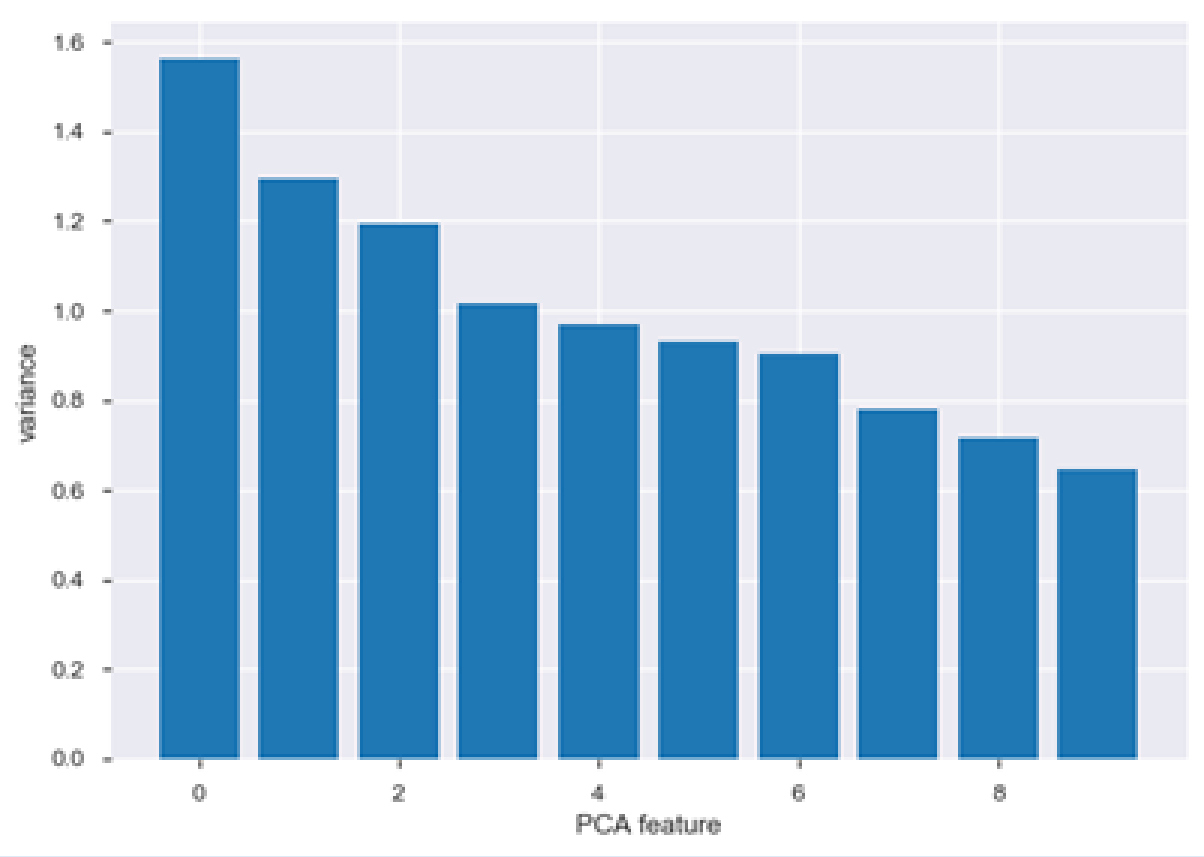
This phase consists out of the following steps:

- **Handling Categorical Data**
- **Preprocessing: scaling**
- **FEATURE ENGINEERING**
- FEATURE SELECTION
  - **Univariate Selection**
  - **Feature Importance**
  - **Correlation Matrix with Heatmap**

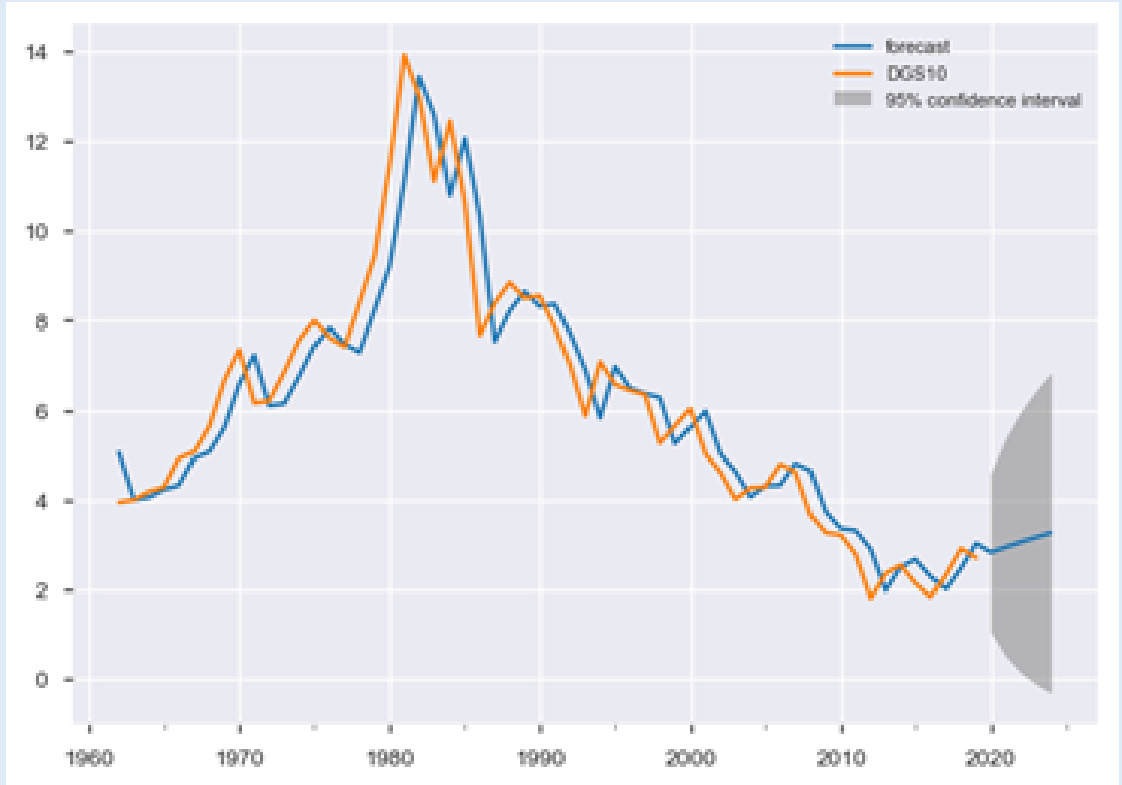


Using PCA for dimensionality reduction involves zeroing out one or more of the smallest principal components, resulting in a lower-dimensional projection of the data that preserves the maximal data variance.

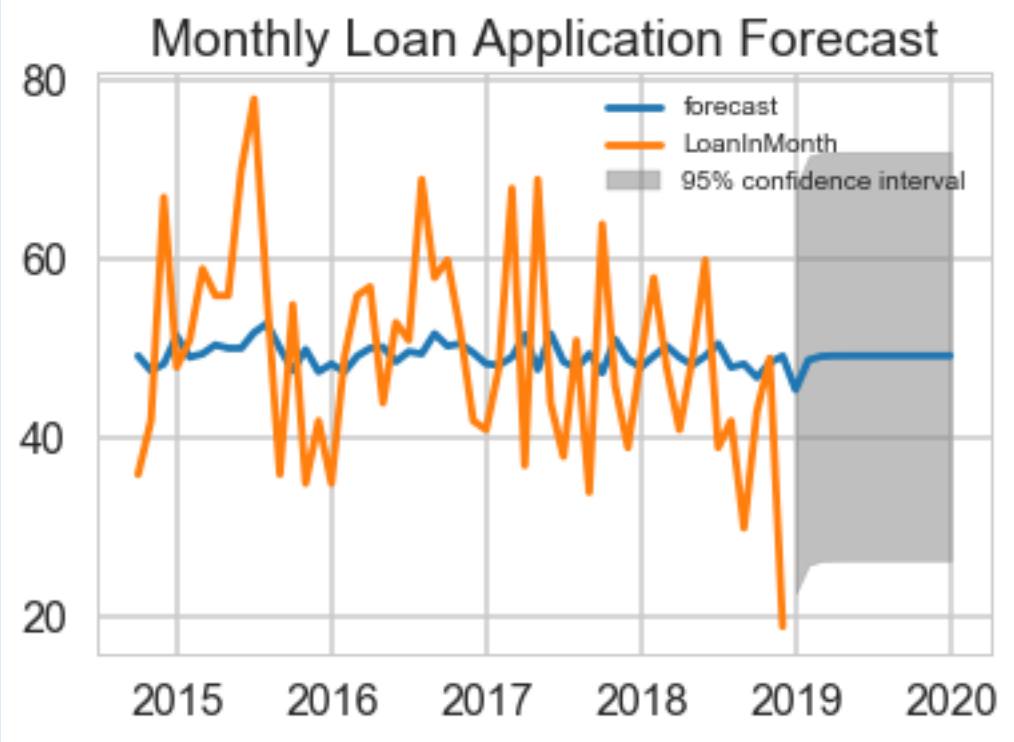


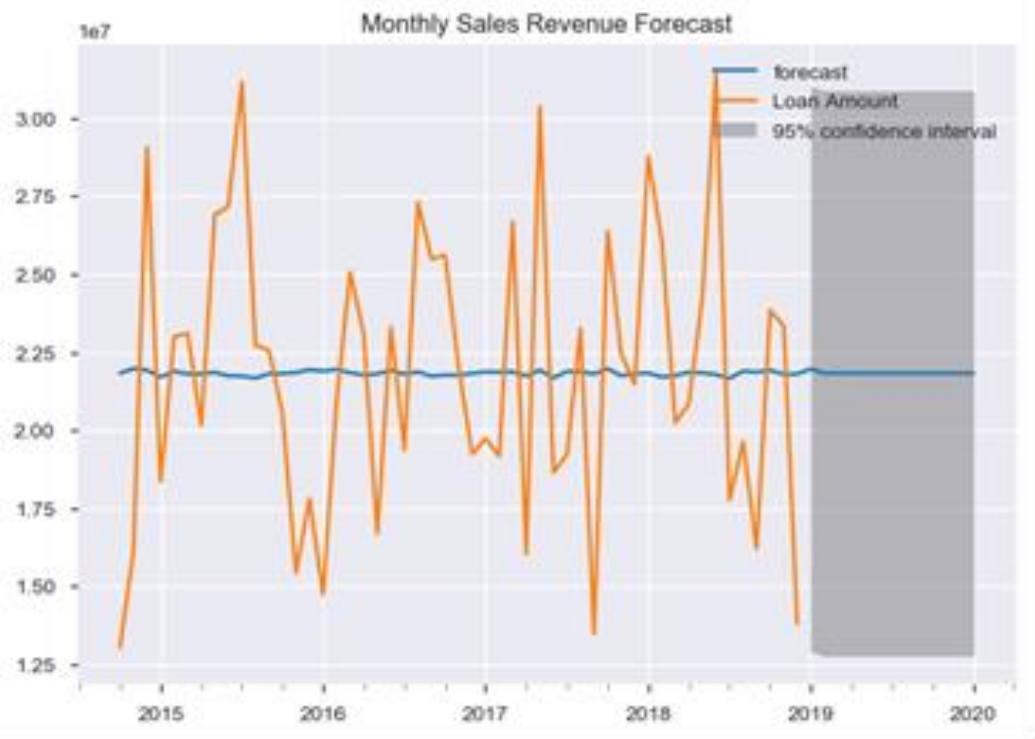


There is some mean reversion in interest rates over long horizons. When interest rates are high, they tend to drop and when they are low, they tend to rise over time. Currently they are below long-term rates, so they are expected to rise, but an AR model attempts to quantify how much they are expected to rise.

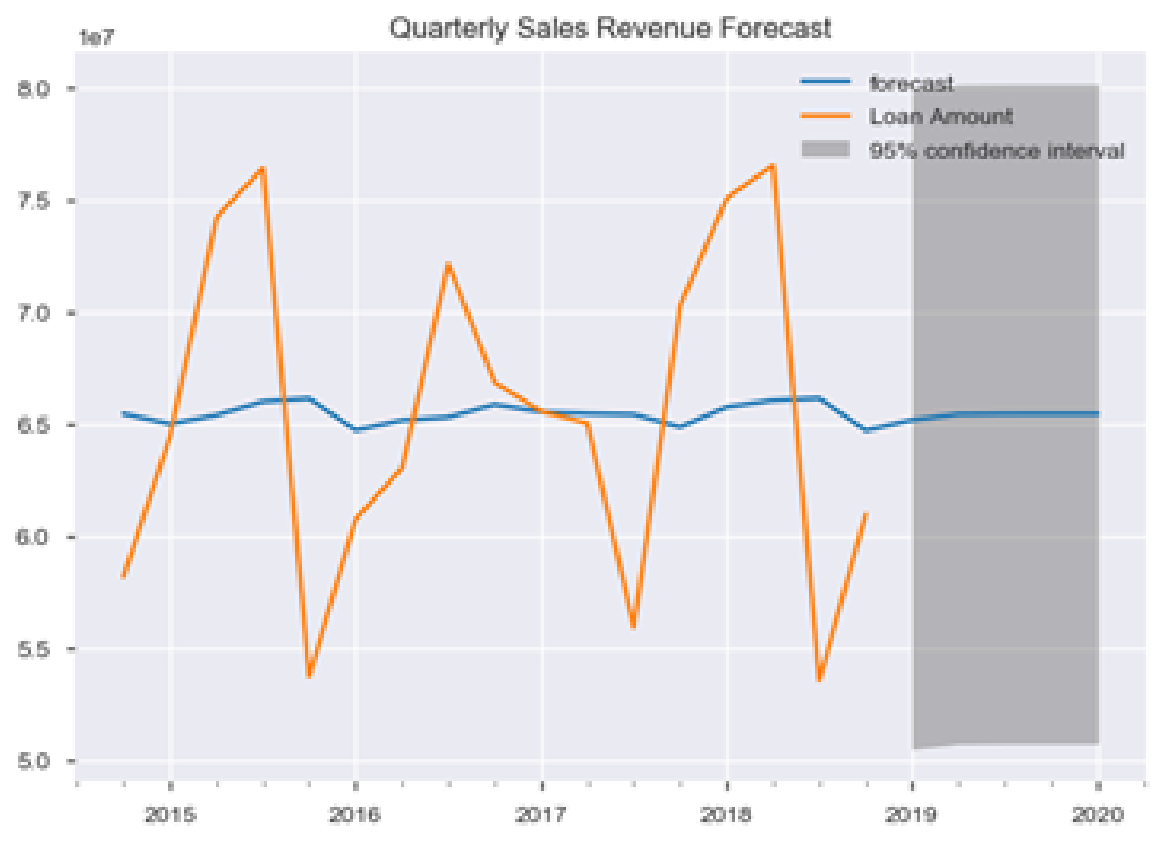


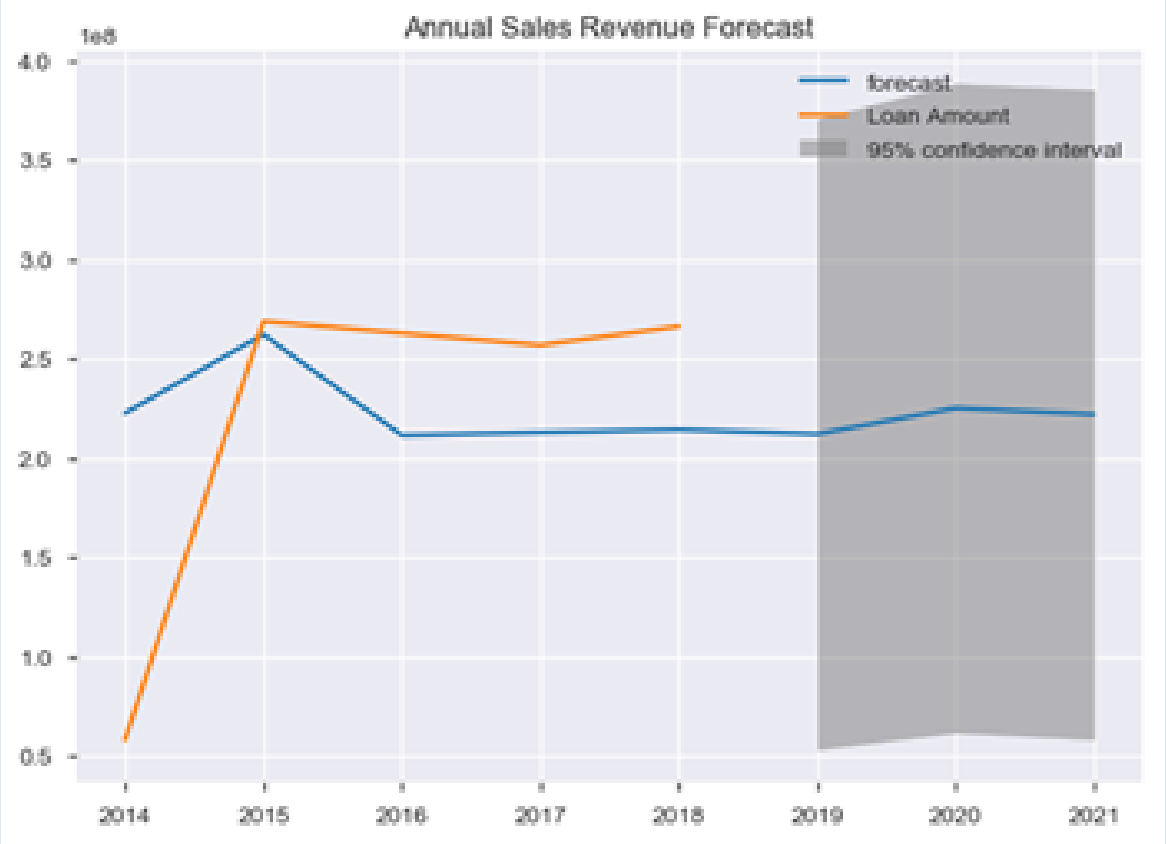
Forecast Using ARIMA Model: Monthly Loan Application vs. Loan Revenue





By end of 2021, with 95% confidence interval, Expected Revenue per Years will be around 220M. Low end is 70M & High End is 360M.
With more annual data, we can do better annual revenue prediction.

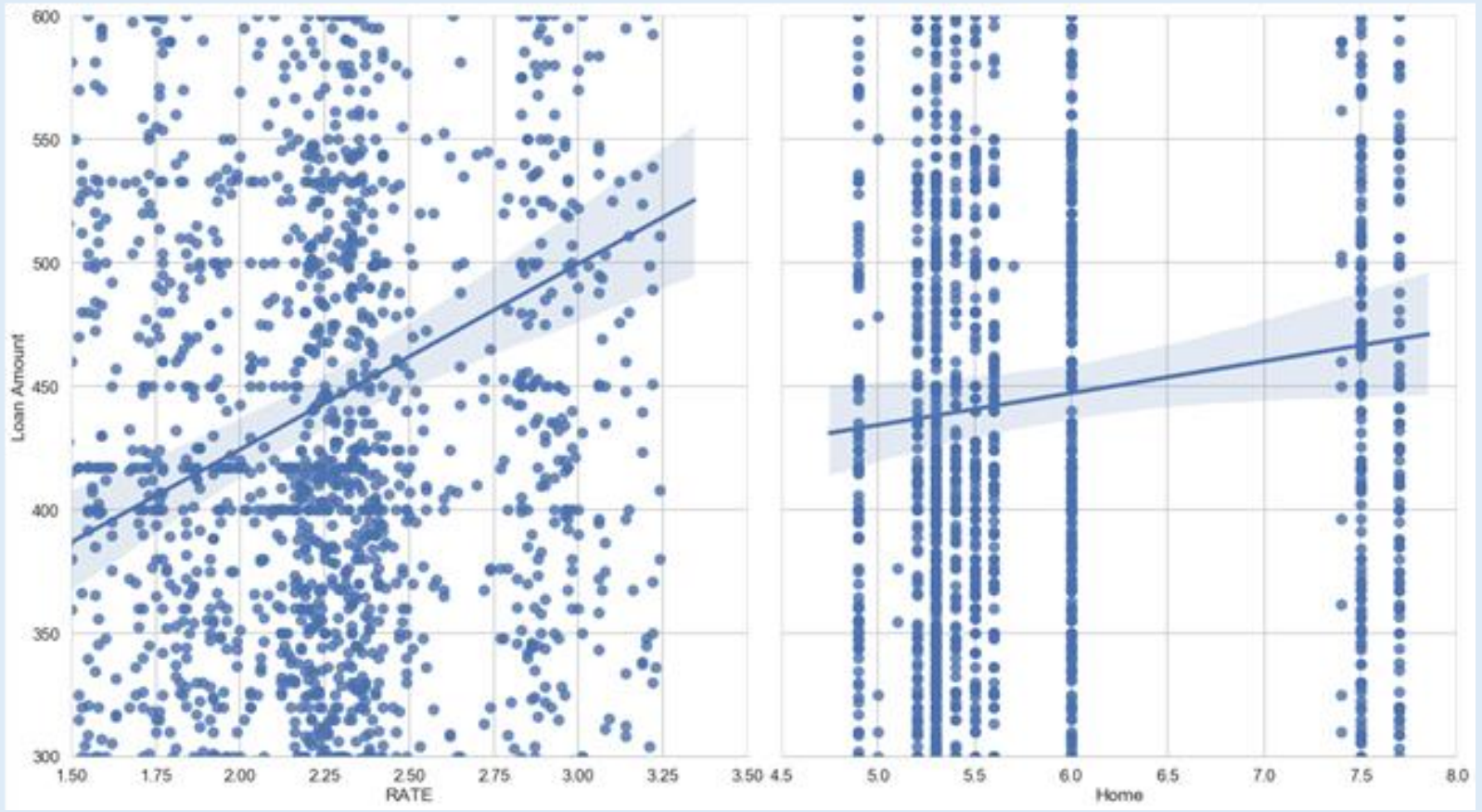




Linear regression

- Linear Regression - Root Mean Square Error : 0.298
- Lasso model - Root Mean Square Error : 0.308
- Ridge Regression - Root Mean Square Error : 0.298
- Elastic Net - Root Mean Square Error : 0.308
- Stochastic Gradient Descent - Root Mean Square Error : 0.298
- **Fitting Linear Regression using statsmodels**
- **Fitting Linear Regression using sklearn**

sns.pairplot(model_data, x_vars=['RATE', 'Home'], y_vars='Loan Amount', size=11, aspect=0.9, kind='reg')

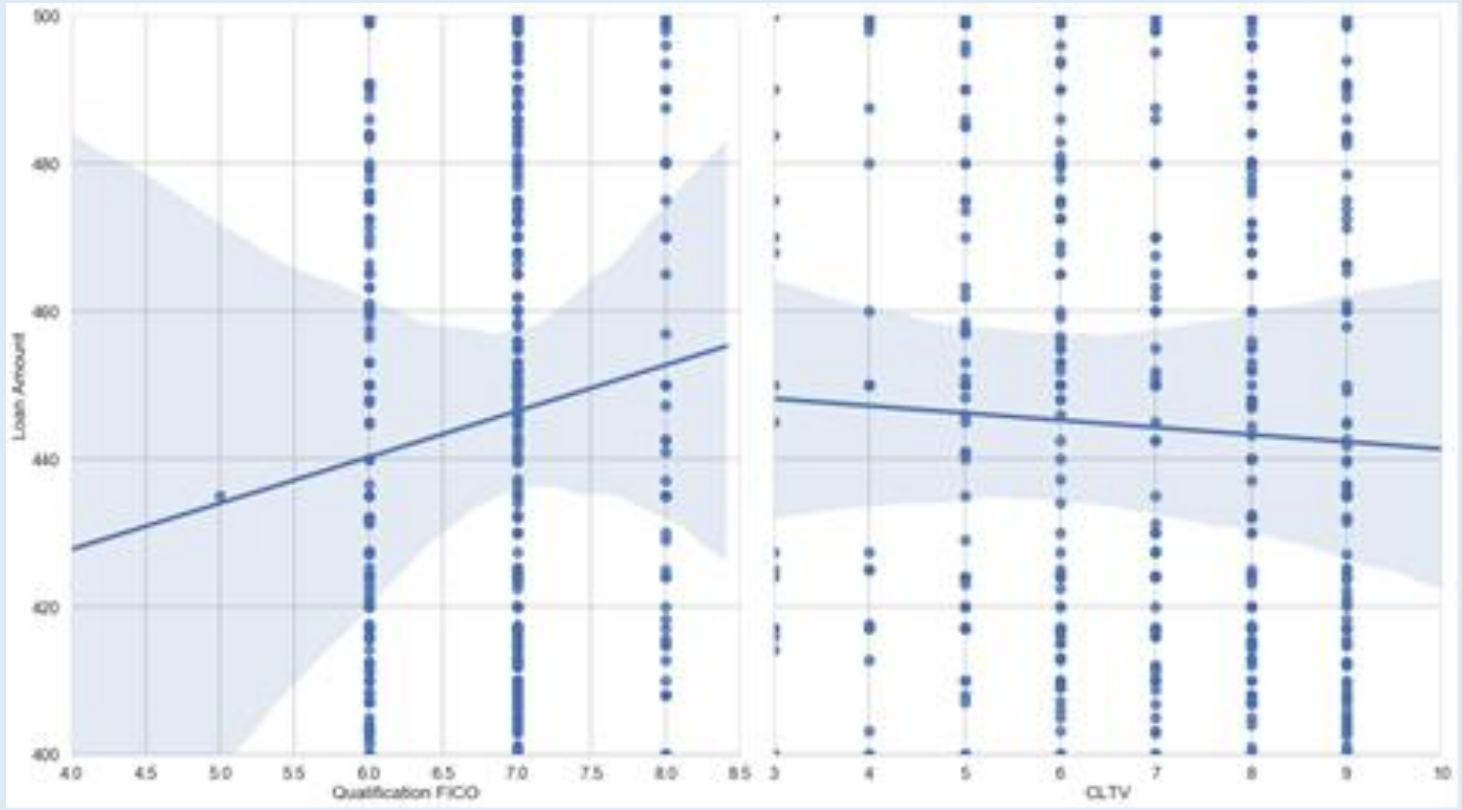


US10Y goes up loan amount slightly increases (Loan Amounts in 1000s)

Home Supply increases, loan amount also increases (Loan Amounts in 1000s)

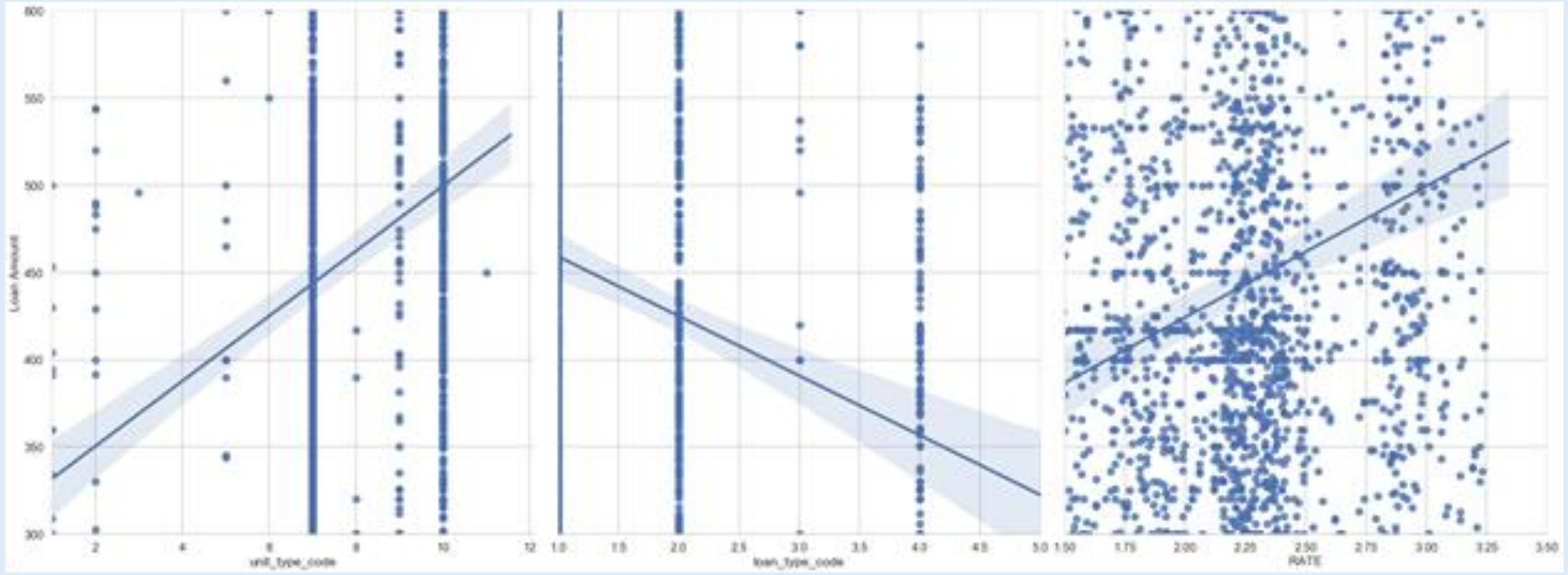
Interest Rate Chage has bigger impact on Loan Amount compare to Home Supply

sns.pairplot(model_data, x_vars=['Qualification FICO', 'CLTV'], y_vars='Loan Amount‘]



- ***Majority of the FICO scores between 600 and 820 (Loan Amounts in 1000s)***
- ***Majority of the CLTV scores between 30 and 100 (Loan Amounts in 1000s)***
- ***FICO goes up, Loan Amount goes up***
- ***CLTV goes up, Loan Amount Goes down***

```
sns.pairplot(model_data, x_vars=['unit_type_code',
    'loan_type_code', 'RATE'], y_vars='Loan Amount',
size=11, aspect=0.9, kind='reg')
```

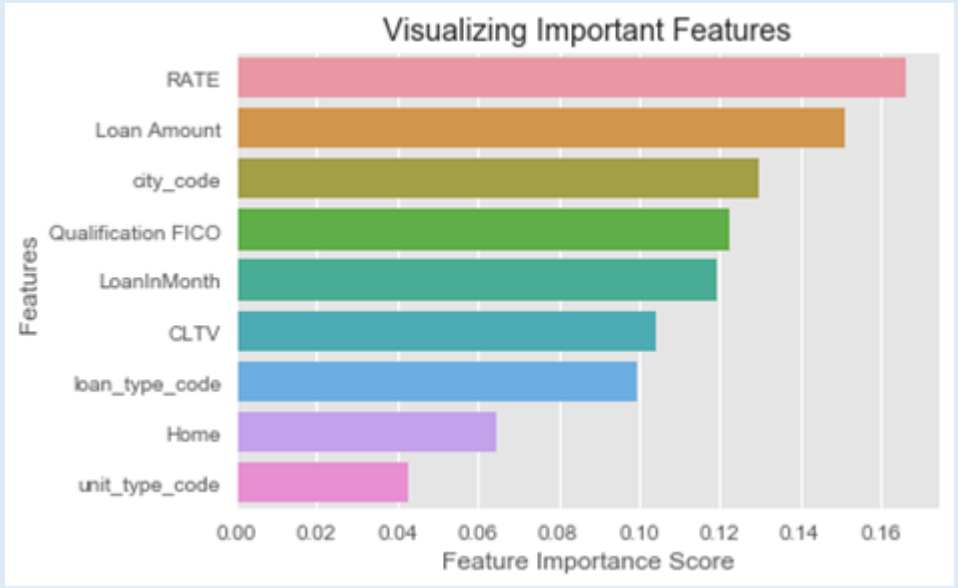


- *As num of unit decreases ave Loan Amount also decreases*
- *loan_type_code: Conventional is high volume but Slightly low average loan amount*
- *Two Family (Code = 10) has higher Loan Amount that three Family (Code = 9)*
- *Conventional Mortgage has Higher Loan Amount FHA*
- *Loan Amount increases with 10 Years Treasury Rate.*

Random forests (RF): is a supervised *learning algorithm*.

- ***Unscalled: Random Forest Classifier Accuracy : 0.78***
- ***Scalled: Random Forest Classifier Accuracy : 0.80***

- **Finding Important Features in Scikit-learn**



Logistic Regression & KNN

- ***Logistic Regression Accuracy: 0.69 or 69%***
- **Centering, Scaling and Logistic Regression**
- The performance of **logistic regression did not improve** with data scaling
- **After Scaling Synthesized Data**
- ***logistic regression score for test set: 0.925000 or 92.5%***

- *With Scale and Synthesize the data we can see huge improvement on the* ***KNN model accuracy.. 28% to 93%***

Support Vector Regression (SVR)

- **SUPPORT VECTOR REGRESSION. SVR is a bit different from SVM.**
- **SVR is an regression algorithm, use SVR for working with continuous Values instead of Classification which is SVM.**
- ***SVR accuracy: -0.0027 which is unacceptable for our Mortgage Loan Data Sets***
- **A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.**
- ***With* Tunning Parameter for SVM - *Support Vector Classifier Accuracy : 0.6866267465069861 or 69%***

KNN

- ***KNeighborsClassifier Accuracy for Loan Types: 0.6826347305389222 or 68%***
- **Preprocessing: scaling**
- ***k-NN score for training set: 0.770000 or 77%***

Using Scikit-learn, optimization of decision tree classifier performed by only pre-pruning. Maximum depth of the tree can be used as a control variable for pre-pruning. In the following the example, we can plot a decision tree on the same data with max_depth=4.

- ***Decision Tree Classifier Accuracy : 0.839 or 83.4%***