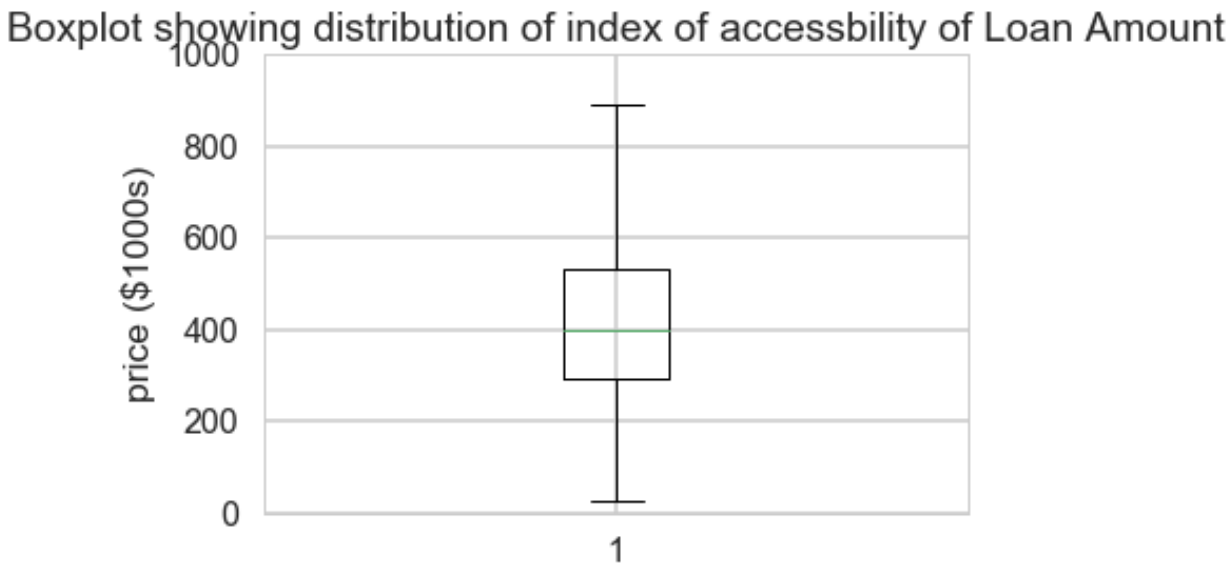
Capstone Project – EDA & Statistics

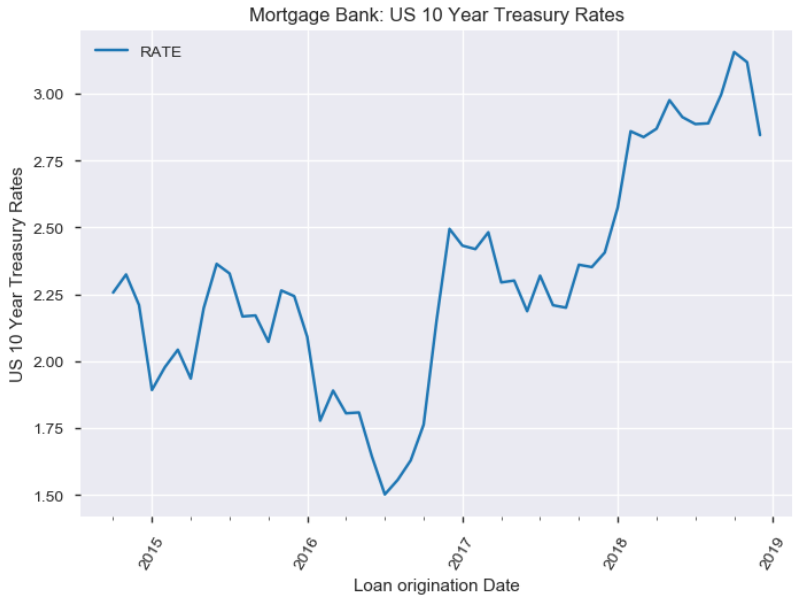
Table of Contents

The goal of the Data Preparation stage is to transform and enrich the dataset so that it can be fed into the models. We use box plot for outlier detection.

Outlier detections using Boxplot:

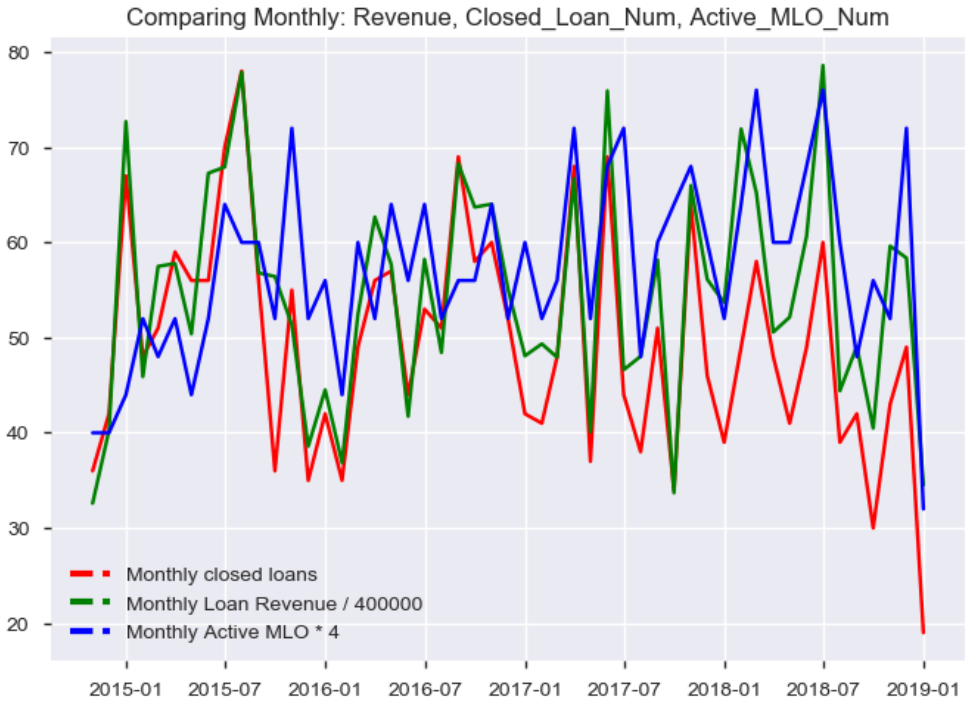


Generally, when US 10 Years Treasury Rate fluctuates, that leads lenders to adjust their internal bank rates accordingly.



Interest Rate is currently historical low. 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.

Now we will compare Monthly Revenue, Monthly Closed Loan Number and Active Mortgage Loan Originators. We will count number of MLO actively closing loans on any given month.

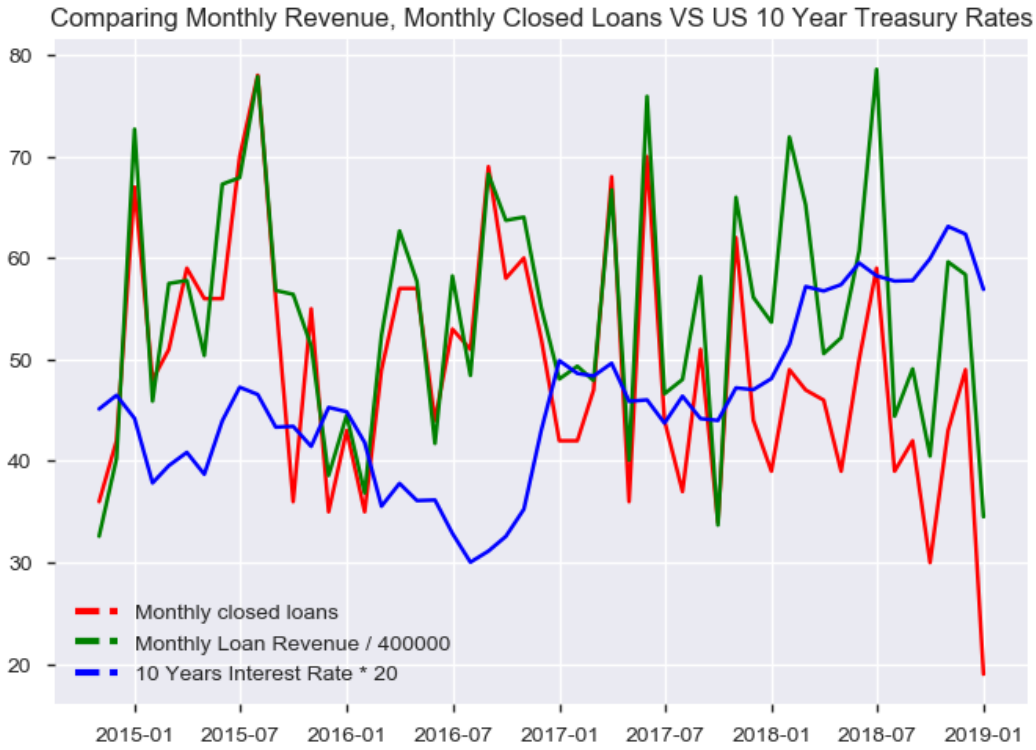


As number of active MLO goes up, which will directly and positively impact numbers of loan closed per month, eventually mortgage revenue will go up. On the other hand, once number of active MLO goes

down, mortgage revenue and number of loan per month goes down as well. By visualizing the graphs, we can see that monthly data of closed loan numbers , monthly revenue and active MLO numbers ber months, all moving at the same direction.

Let's find out interest rate effect on Monthly Closed Loans and Monthly Revenue.

Pearson correlation coefficient between Monthly Closed Loans & Monthly loan Rev: 0.8295841987344892
Pearson correlation coefficient between Monthly Interest & Monthly loans Closed Data: -0.3343475255276081



We can see the strong positive correlation between Monthly Closed Loans and Monthly Revenue. This graph also suggest that, as interest rates goes down, banks monthly revenue and numbers of loans increases, and when the Rates goes up, both Monthly Closed Loans and Monthly Revenue for the Mortgage bank decline. Pearson correlation coefficient between Monthly Interest & Monthly loans Closed Data is **-0.334**, which clearly proves that Monthly Interest Rates & Monthly loans Closed Data is **negatively correlated**.

Acquire 1000 **pairs bootstrap replicates of the Pearson correlation coefficient using the draw_bs_pairs()** function you wrote in the previous exercise for **CLTV data VS Qualification FICO** Data and **Monthly Loan_num_data VS. Monthly_loan_rev**. Compute the **95% confidence** interval for both using your bootstrap replicates. We have created a NumPy array of percentiles to compute. These are the 2.5th, and 97.5th. By creating a list and convert the list to a NumPy array using np.array(). For example, np.array([2.5, 97.5]) would create an array consisting of the **2.5th and 97.5th percentiles**.

CLTV data VS Qualification FICO Data : -0.037060429592168175 [-0.08 0.]
Monthly Loan_num_data VS. Monthly_loan_rev : 0.8295841987344892 [0.73 0.9]

''' It shows that there is **statistically significant relationship** between number of loans closed and loan revenue'''

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. With the ADF test, the "null hypothesis" (the hypothesis that we either reject or fail to reject) is that the series follows a random walk. Therefore, **a low p-value** (say less than 5%) means we can reject the null hypothesis that the series is a random walk.

(-4.49967715626648, 0.00019690419763896495, 10, 40, {'1%': -3.6055648906249997, '5%': -2.937069375, '10%': -2.606985625}, 1307.285137861741)
The p-value of the test on loan_rev is: 0.00019690419763896495

'''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. '''

Let's try same for Monthly Loan Data:

(-5.7659481612690495, 5.532460937310067e-07, 0, 50, {'1%': -3.568485864, '5%': -2.92135992, '10%': -2.5986616}, 300.2408550906095)
The p-value of the test on loan_num is: 5.532460937310067e-07

According to this test, p-value is very low (lower than 0.05). We reject the hypothesis that monthly_loan_num_data follow a random walk.
Let's try same for Interest Rate Data:

(-1.396544767435691, 0.5839314748568241, 1, 49, {'1%': -3.5714715250448363, '5%': -2.922629480573571, '10%': -2.5993358475635153}, -34.97121939430423)
The p-value of the test on monthly_rate_data 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 monthly or annual changes, the autocorrelation is negative. This implies

that while short term changes in interest rates may be uncorrelated, long term changes in interest rates are negatively auto correlated.

The autocorrelation of daily interest rate changes is -0.06

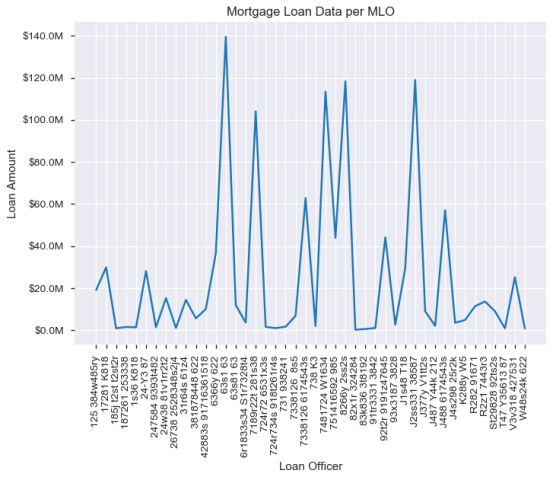
The autocorrelation of annual interest rate changes is -0.97

'''Notice how the daily autocorrelation is small but the annual autocorrelation is large and negative'''

5.2 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.
Let's explore Loan Origination Data with Unit Types:



Some MLO generates major portions of the sales revenue for the Bank, many of the MLO is performing way below company standards. Additional research required for the root causes behind the performance. Let's explore the Loan Statistics data for the Mortgage Bank:

*********** Loan Statistics for Mortgage Bank **********

Average Loan Amount is	***: $ 445099.91***
Median Loan Amount is	***: $ 400000.00***
Standard deviation of is	***: $ 288803.96***
Minimum Loan Amount is	***: $ 23600.00***
Maximum Loan Amount is	***: $ 6125000.00***
Total of Loan Amount is	***: $ 1113194869.34***
10% of Loan Amount is below	***: $ 200000.00***
25% of Loan Amount is below	***: $ 292500.00***
50% of Loan Amount is below	***: $ 417000.00***
75% of Loan Amount is below	***: $ 533000.00***
90% of Loan Amount is below	***: $ 696500.00***

FICO Score is one of the critical information for processing mortgage loan application.

********** Qualification FICO Statistics for Mortgage Bank **********

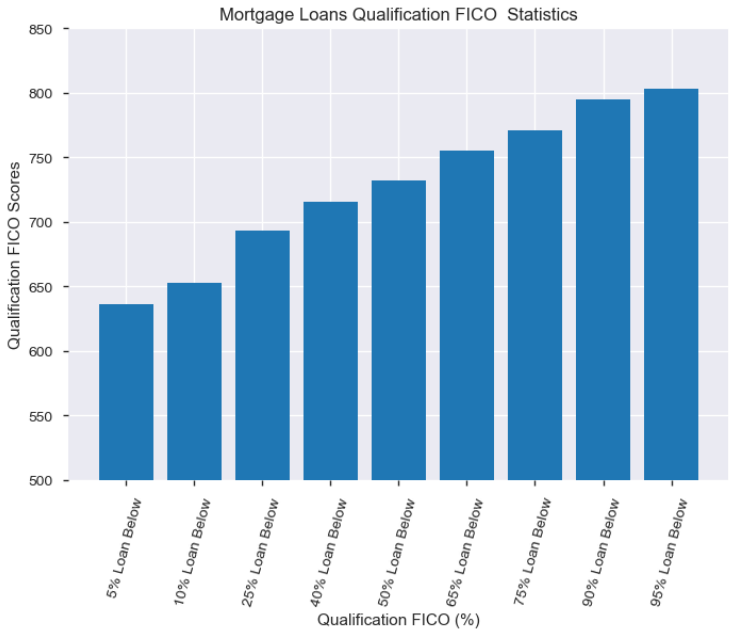
Average FICO is	***: $ 727.81***
Median FICO is	***: $ 732.00***
Standard deviation is	***: $ 56.61***
Minimum FICO is	***: $ 0.00***
Maximum FICO is	***: $ 825.00***
5% of FICO is below	***: $ 636.00***
10% of FICO is below	***: $ 653.00***
25% of FICO is below	***: $ 693.00***
40% of FICO is below	***: $ 716.00***
50% of FICO is below	***: $ 732.00***
65% of FICO is below	***: $ 755.00***
75% of FICO is below	***: $ 771.00***
90% of FICO is below	***: $ 795.00***
95% of FICO is below	***: $ 803.00***

==========

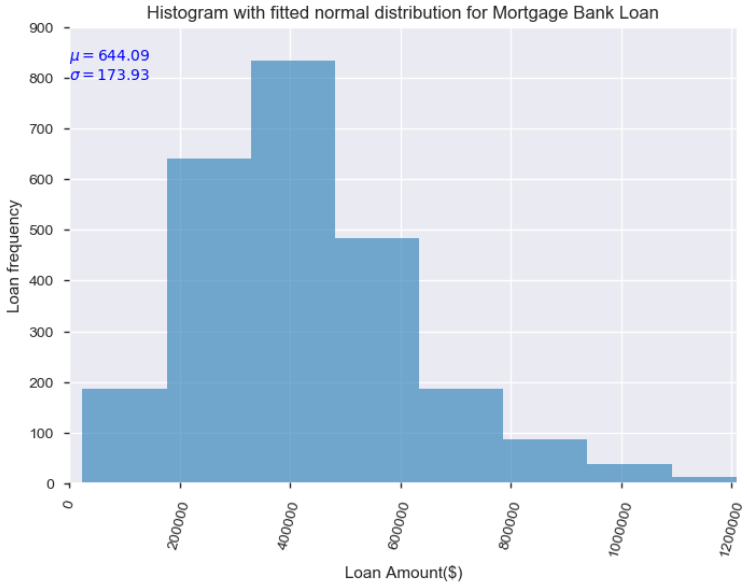
We can see that 50% of the consumers FICO score is between 693 and 771. Let's visualize FICO statistics

fico_score = [fico_q7, fico_q0, fico_q1, fico_q5, fico_q2, fico_q6, fico_q3, fico_q4, fico_q8]

fico_pct=['5% Loan Below', '10% Loan Below', '25% Loan Below', '40% Loan Below', '50% Loan Below', '65% Loan Below', '75% Loan Below','90% Loan Below', '95% Loan Below']

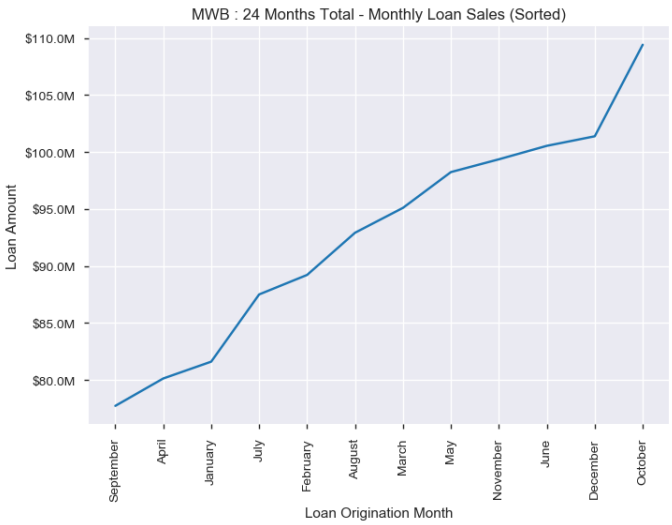


We can see that 90% of the consumers FICO score is over 650. Most of cases applicant with lower FICO score will not qualify for Conventional mortgages. In many cases, FHA Loan type could be only option remaining for the applicants with FICO score; many bank uses cut-off points for FICO Score (640-680) for conventional mortgages. FHA accepts FICO score below 600.



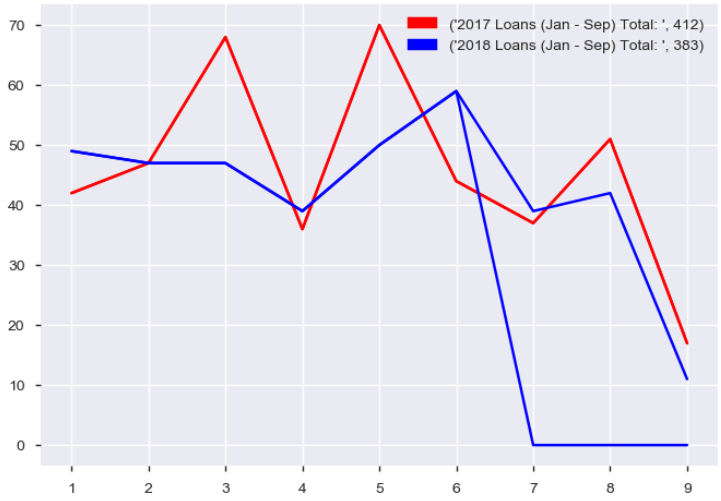
excess kurtosis of normal distribution (should be 0): 10.20423896299732
skewness of normal distribution (should be 0): 1.6231675369974472
mean : 644.086535155362
var : 30252.443004451427
skew : 1.6231675369974472
kurt : 10.20423896299732

The fancy text to show us what the parameters of the distribution are (mean and standard deviation). We can create a histogram with 20 bins to show the distribution of purchasing patterns. Fit a normal distribution to the data then plot the histogram.



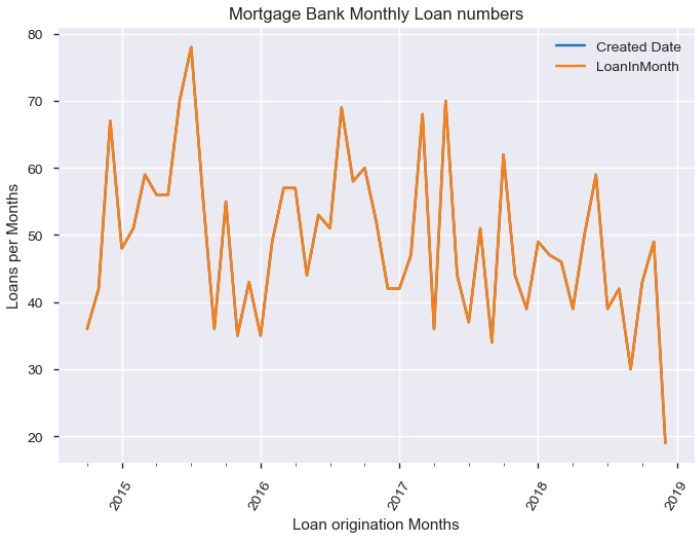
Monthly Worst Sales: January, April and September and Best is June, December and October.

Comparing year-over-year performance:



In 2017, Mortgage Bank has closed more loan compare to 2018

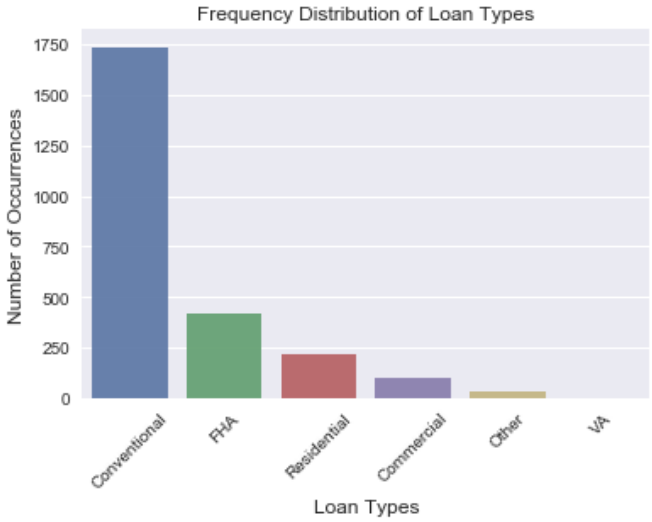
In 2017, Mortgage Bank sales revenue was (M_amount_17.sum()) = $180M
For 2018, (M_amount_18.sum())= $197M

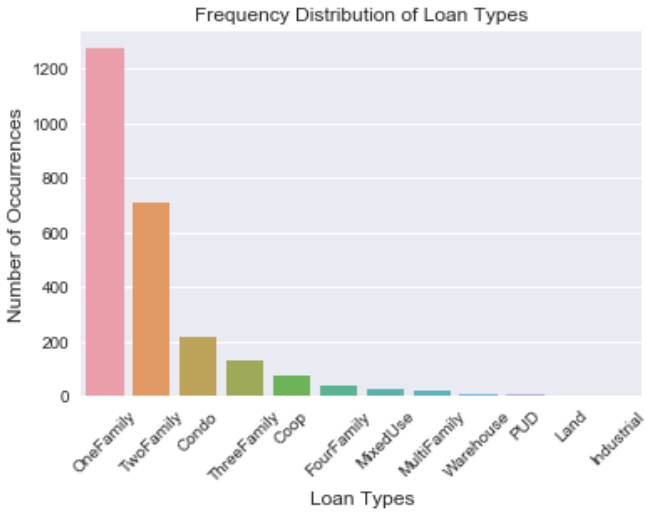


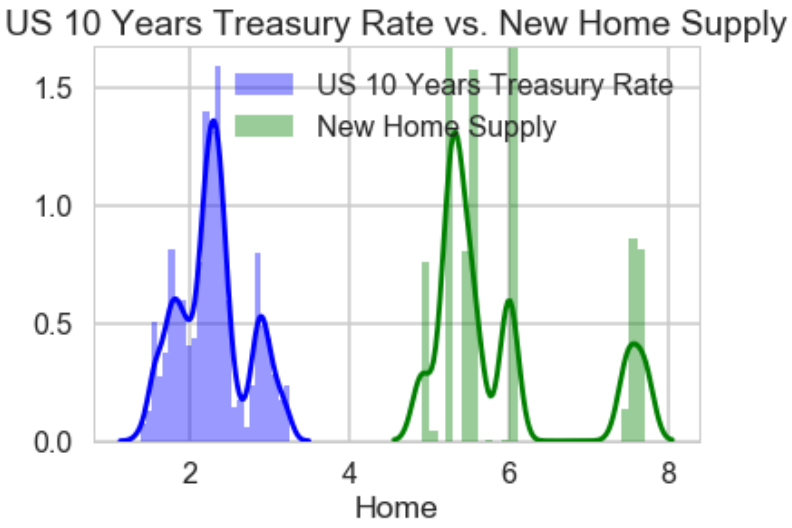
Pearson correlation coefficient between Monthly Closed Loans & Monthly loan Rev: 0.8295841987344892

Monthly Loan numbers and Monthly Sales volumes are strongly positively correlated. Once we analyze '''Visual exploration is the most effective way to extract information between variables.

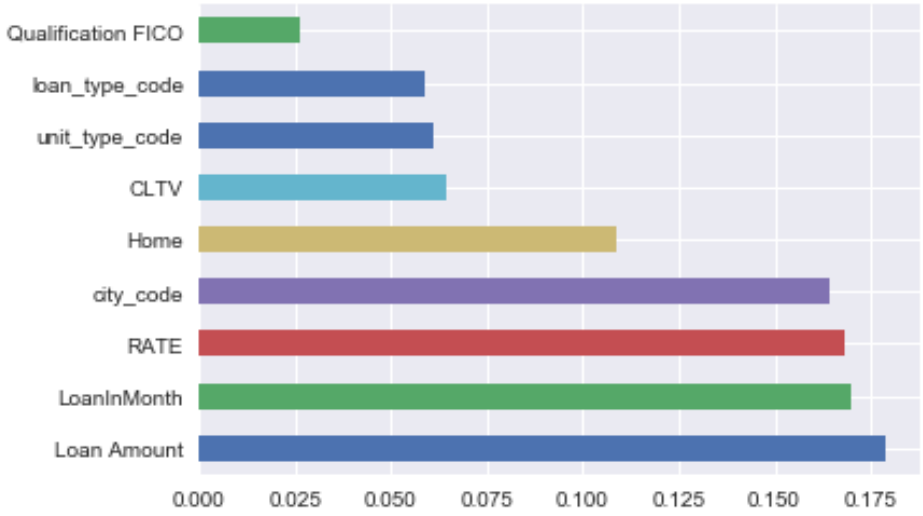
We can plot a barplot of the frequency distribution of a categorical feature using the seaborn package, which shows the frequency distribution of the mortgage dataset column.





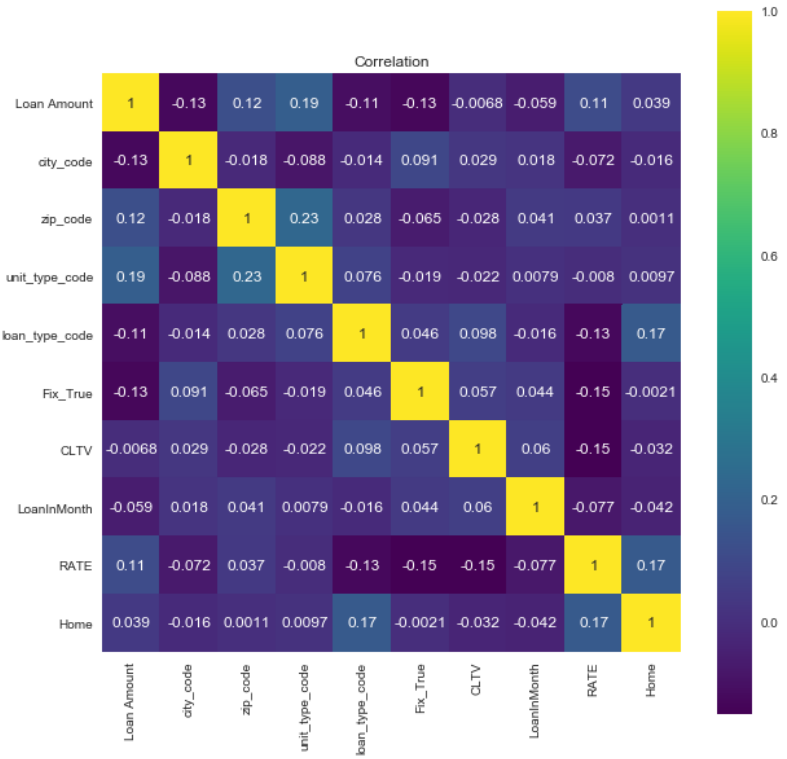


''' Home Supply goes up, RATE goes up ''' '''



Correlation Matrix with Heatmap

Correlation states how the features are related to each other or the target variable.
Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable)
Heatmap makes it easy to identify which features are most related to the target variable, we will plot heatmap of correlated features using the seaborn library.



ARMA Model Results

==========

Dep. Variable:	**y**	**No. Observations:**	**51**
Model:	**ARMA(1, 0)**	**Log Likelihood**	**23.792**
Method:	**css-mle**	**S.D. of innovations**	**0.149**
Date:	**Thu, 31 Jan 2019**	**AIC**	**-41.584**
Time:	**11:49:00**	**BIC**	**-35.788**
Sample:	**0**	**HQIC**	**-39.369**

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	**2.3886**	**0.247**	**9.653**	**0.000**	**1.904**	**2.874**
ar.L1.y	**0.9310**	**0.045**	**20.707**	**0.000**	**0.843**	**1.019**

Roots

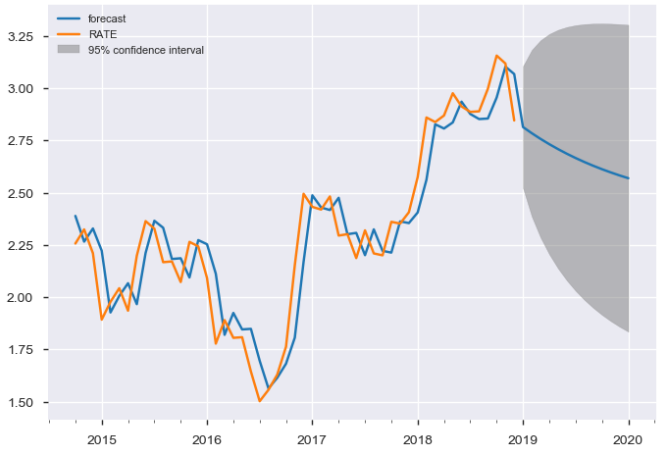
	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	**1.0741**	**+0.0000j**	**1.0741**	**0.0000**

When the true phi=0.9, the estimate of phi (and the constant) are:[2.39 0.93]

Forecasting with an AR Model

In addition to estimating the parameters of a model, we can also do forecasting using statsmodels. These forecasts can be made using either the predict() method if we want the forecasts in the form of a series of data, or using the plot_predict()

)



Since we have only used only 4 years of Monthly Interest Rate Data, we can see the short term downward momentum on the rate. A daily move up or down in interest rates is unlikely to tell us anything about interest rates tomorrow, but a move in interest rates over a year can tell us something about where interest rates are going over the next year. The DataFrame daily_data contains daily data of **10-year interest rates from 1962 to 2017**.

With 95% confidence the autocorrelation of daily interest rate changes is 0.07
With 95% confidence the autocorrelation of annual interest rate changes is -0.22

Fitting Linear Regression using statsmodels

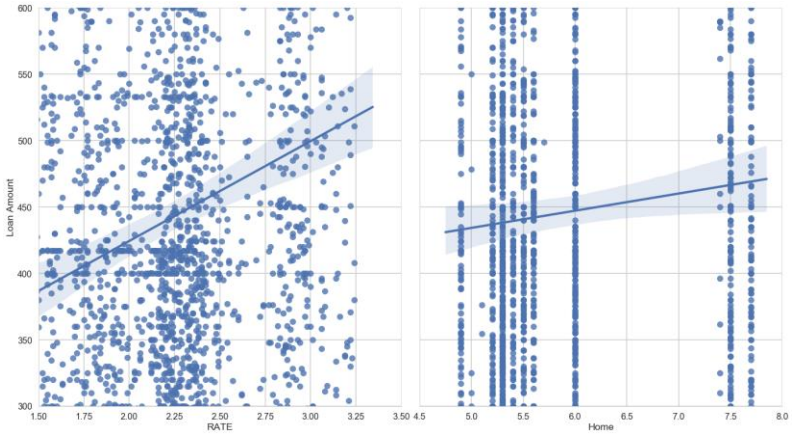
Statsmodels is a great Python library for a lot of basic and **inferential statistics**. It also provides basic regression functions using an R-like syntax, so it's commonly used by statisticians. The version of least-squares we will use in statsmodels is called ordinary least-squares (OLS). There are many other versions of least-squares such as partial least squares (PLS) and weighted least squares (WLS).

OLS Regression Results

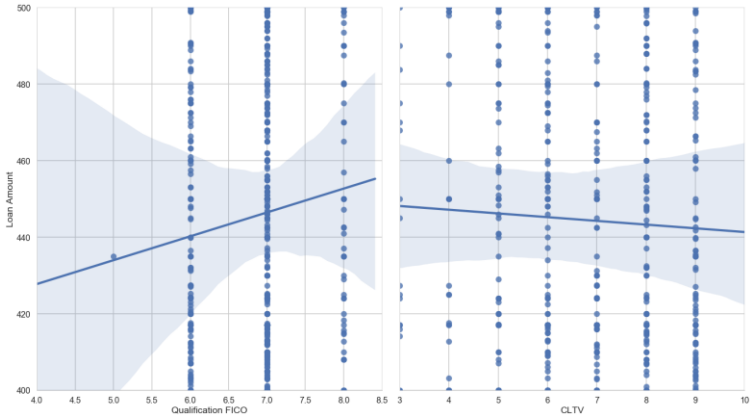
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.026
Model:                            OLS   Adj. R-squared:                  0.025
Method:                 Least Squares   F-statistic:                     16.97
Date:                Fri, 01 Feb 2019   Prob (F-statistic):           9.82e-14
Time:                        03:57:19   Log-Likelihood:                -542.85
No. Observations:                2501   AIC:                             1096.
Df Residuals:                    2496   BIC:                             1125.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0463      0.054     19.536      0.000       0.941       1.151
LoanInMonth    0.0006      0.000      1.602      0.109      -0.000       0.001
RATE          -0.1074      0.015     -7.335      0.000      -0.136      -0.079
Home           0.0088      0.007      1.260      0.208      -0.005       0.023
CLTV           0.0045      0.002      1.913      0.056      -0.000       0.009
==============================================================================
Omnibus:                     1214.830   Durbin-Watson:                   1.884
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4810.993
Skew:                          -2.509   Prob(JB):                         0.00
Kurtosis:                       7.581   Cond. No.                         291.
==============================================================================
```

Fitting Linear Regression using sklearn

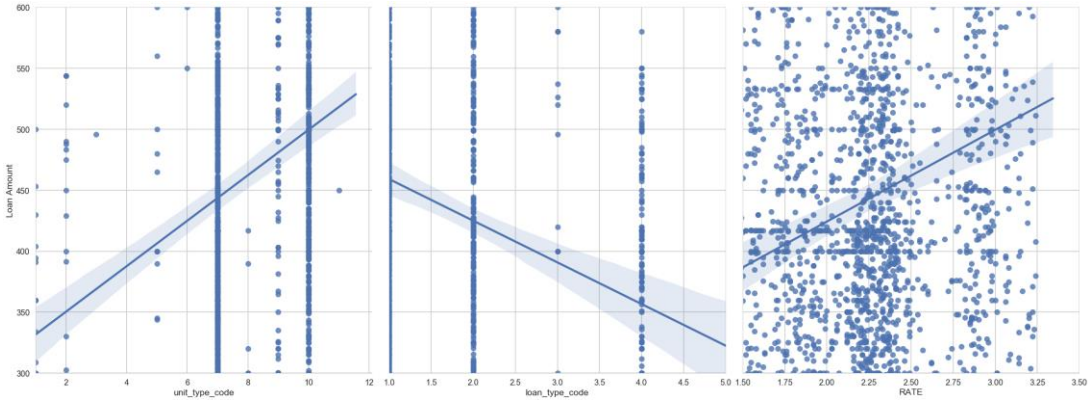
np.mean((lm.predict(X) - y) ** 2))
Mean squared error (Fix_True Rate): 0.09
Mean squared error (loan_type_code): 0.80
Mean squared error (Loan Amount): 75543.58
Mean squared error (unit_type_code): 7.76



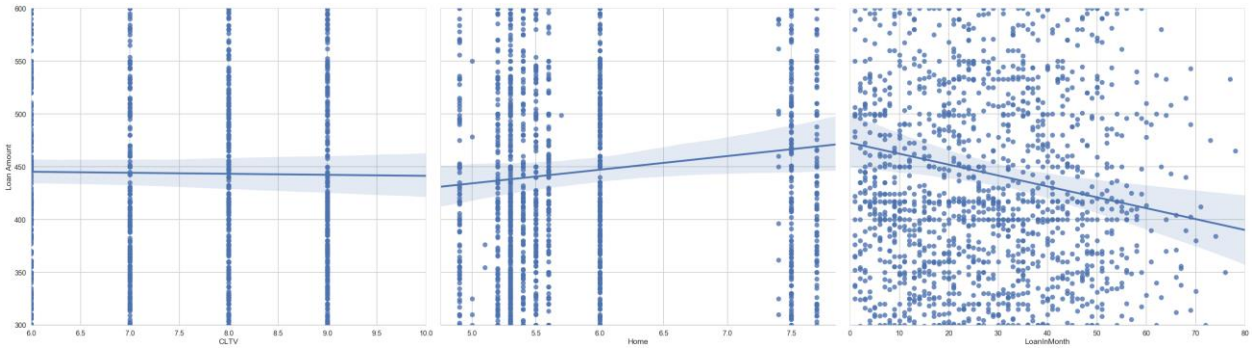
US10Y goes up loan amount slightly increases (Loan Amounts in 1000s)
Home Supply increases, loan amount also increases (Loan Amounts in 1000s)
Interest Rate Chage has bigger impact on Loan Amount compare to Home Supply



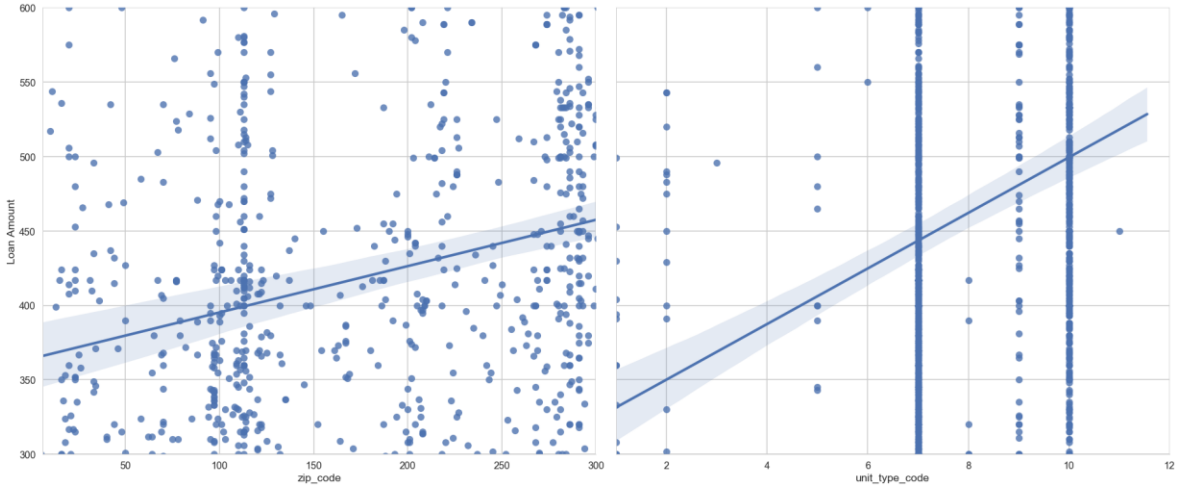
Majority of the FICO scores between 600 and 820 (Loan Amounts in 1000s)
Majority of the CLTV scores between 30 and 100 (Loan Amounts in 1000s)
FICO goes up, Loan Amount goes up
CLTV goes up, Loan Amount Goes down



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.



CLTV does not have much impact on Loan Amount
As number of loans per month increases, loan amount decreases
As Home Supply increases, loan amount also increases



NYC NJ (zip_code 1st digit starting with 7 & 11) seems to be closing more loans and generation more revenues for the Mortgage Bank. City shows similar results. Population density plays bigger role on Loan Amount. As number of loans per month increases, loan amount decreases