Forecasting LendingClub Loans in New York

In peer-to-peer (P2P) lending platforms, consumers borrow from other consumers. The typical process is as follows: Consumers who are in need of borrowing money make a request by entering their personal information, including the SSN number, and the amount of money requested. If a request passes the initial checks, LendingClub's algorithm assigns a grade to the request, which translates into an interest rate (the higher the grade, the lower the interest rate). Other consumers who would like to invest into personal loans lend the money. For the most part, the lending is automated, so the P2P lending model is different from crowdfunding models.

In this project, I have developed a number of time series models to understand the loans issued by LendingClub. This includes visualizing data and developing statistical models to help LendingClub management understand better the changes in the characteristics of loans issued in New York over time. I have developed time series models for the total dollar value of loans per capita, to guide LendingClub in its attempts to increase its market share in NY.

Datasets

lendingClub.csv: Data for all the loans issued in the platform from June, 2007 to March, 2017. The data is aggregated to the state-month level.

nyEcon.csv: Some economic indicators for NY for the same timeframe (from June, 2007 to March, 2017).

Census.csv: 2010 U.S. Census data for the population of each state (at the month level)

Data Dictionaries

lendingClub.csv (All averages are the values averaged over the # of loans per state per month)

Variable	Definition
date	Monthly date
state	State abbreviation
Loans (avg and total)	The amount of loan issued in dollars
term (average)	The period in which the number of payments made are calculated (months)
intRate (average)	Interest rate on the loan (in percentages)
grade (average)	Loan grade assigned by the algorithm (A=1, B=2, C=3, D=4, E=5, F=6)
empLength (average)	Employment length of the borrower (in years)
annualInc (average)	The self-reported annual income provided by the borrower during registration

verifStatus (average)	Indicates if the income is verified by LendingClub (Verified=1, Not Verified=0)
homeOwner (average)	The home ownership status provided by the borrower during registration or obtained from the credit report (OWN=1, RENT OR OTHERWISE=0)
openAcc (average)	The number of open credit lines in the borrower's credit file
revolBal (average)	Total credit revolving balance
revolUtil (average)	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
totalAcc (average)	The total number of credit lines currently open in the borrower's credit file
countOfLoans	The number of loans per month per state *(tally taken during aggregation)*

nyEcon.csv

Variable	**Definition**
Date	Monthly date
NYCPI	Consumer price index in New York
NYUnemployment	Unemployment rate in New York -Seasonally adjusted
NYCondoPriceIdx	Condo price index in New York -Seasonally adjusted
NYSnapBenefits	Number of SNAP benefits recipients in New York

census.csv

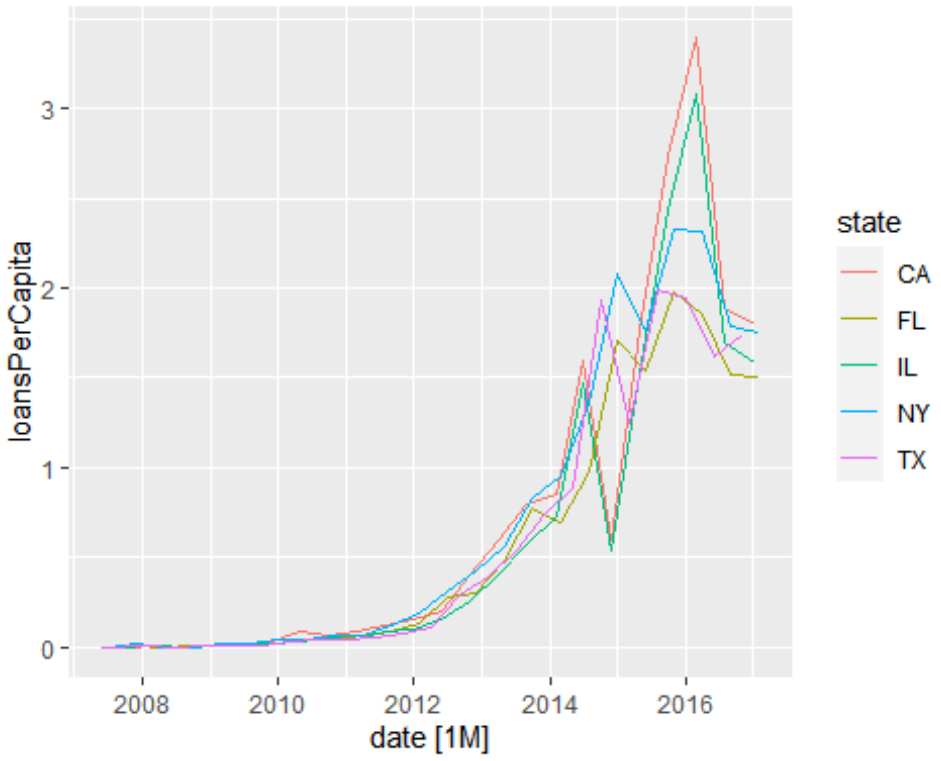
Variable	**Definition**
State name	State
Census2010Pop	Per state population in 2010

1. Data Processing

LendingClub loans data is joined with economic indicators data for the state of New York and with the 2010 census population data to obtain the loans per capita.

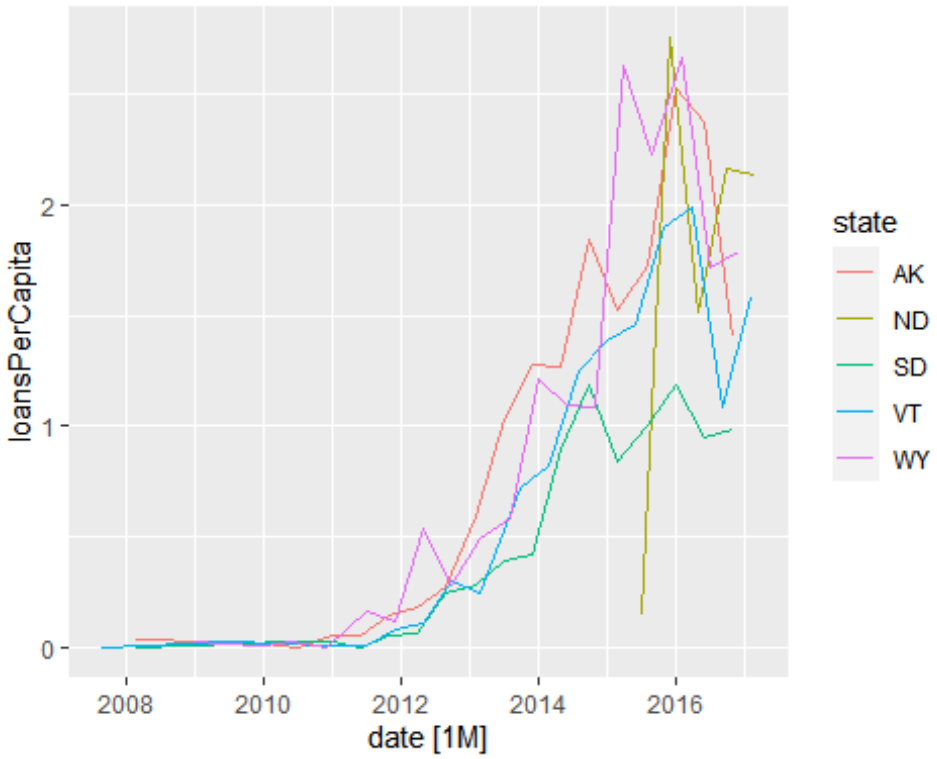
2. Exploratory Analysis

a. Plotted the loans per capita for the states within the top 10th percentile and bottom 10th percentile in terms of population.



The states within the top 10th percentile in terms of population are CA, FL, IL, NY, TX.

If we look at the plot of the states in the top 10th percentile, there is low variance in loansPerCapita until beginning of 2014, after which the loansPerCapita CA and IL has fallen at the end of 2014 and increased substantially as compared to the other states in 2016.

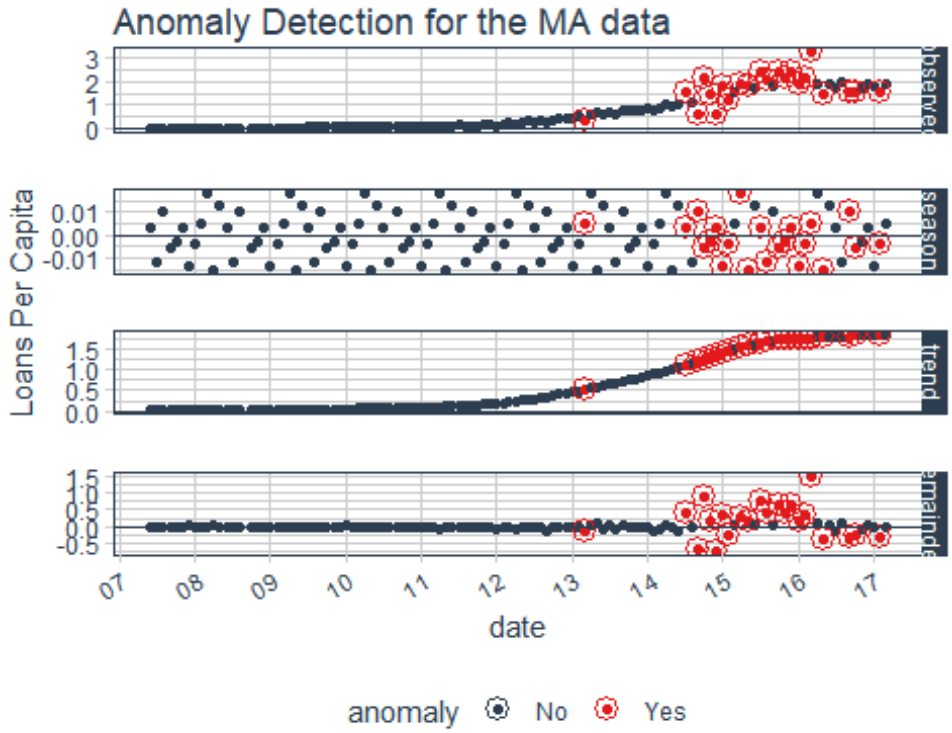


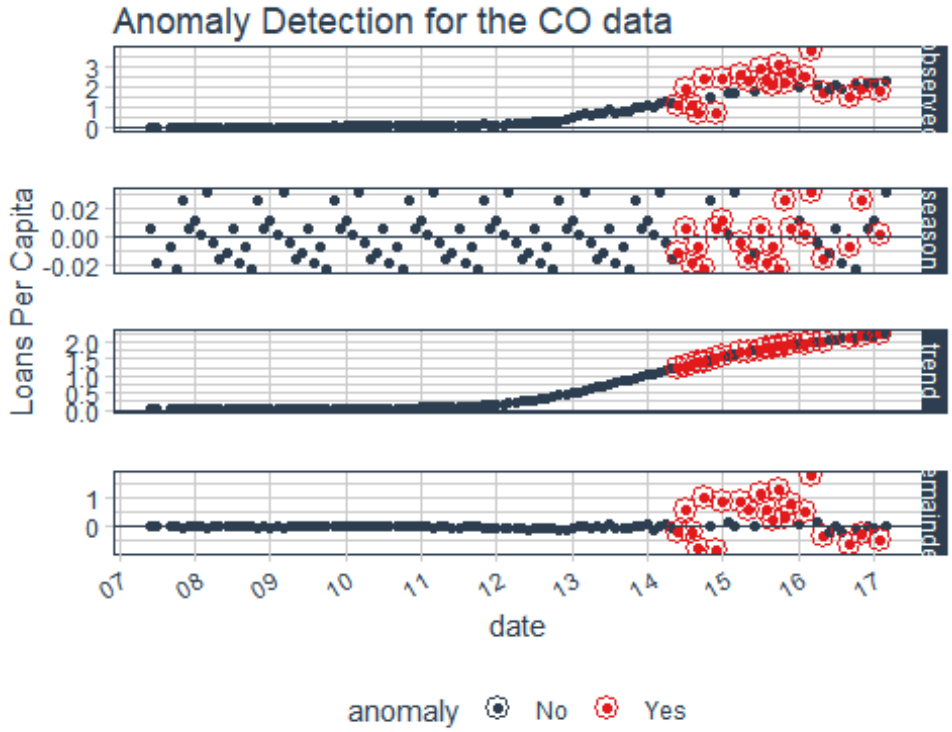
The states within the bottom 10th percentile in terms of population are AK, ND, SD, VT, WY.

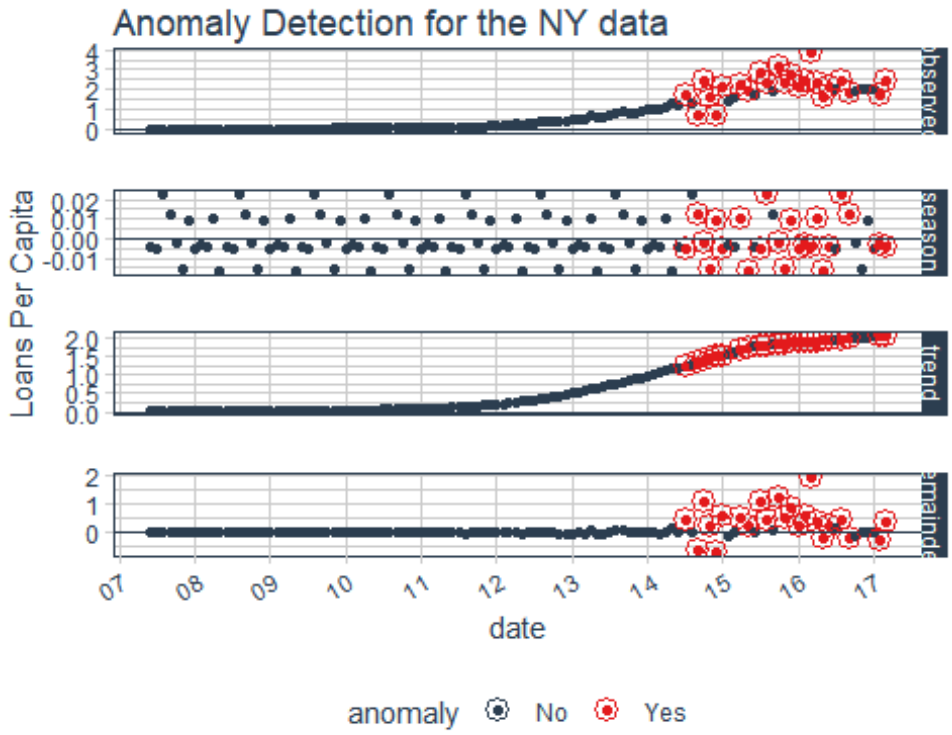
If we look at the plot of the states in the bottom 10th percentile, there is little variance in loansPerCapita until 2011 and there is a lot of variance after that.

The variance is higher in the bottom 10th percentile states as compared to the top 10th percentile states. This could be because of the very small population of these states as compared to the ones in the top 10th percentile.

b. Created anomaly plots to compare the NY data with Massachusetts and Colorado. Used the STL decomposition and interquartile range to mark the anomalies.

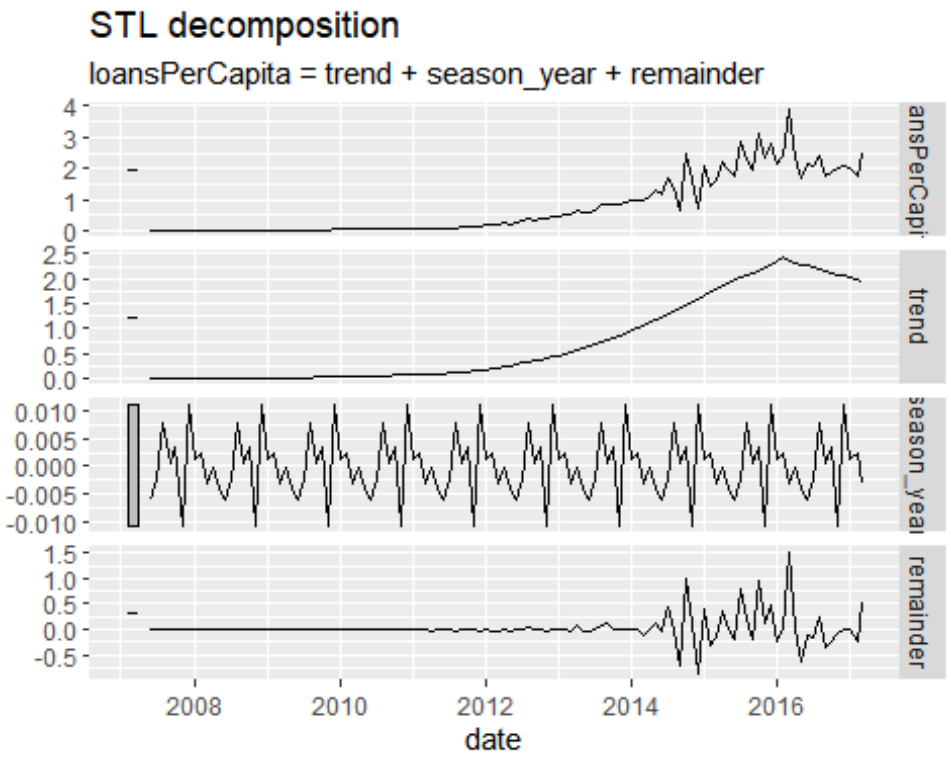






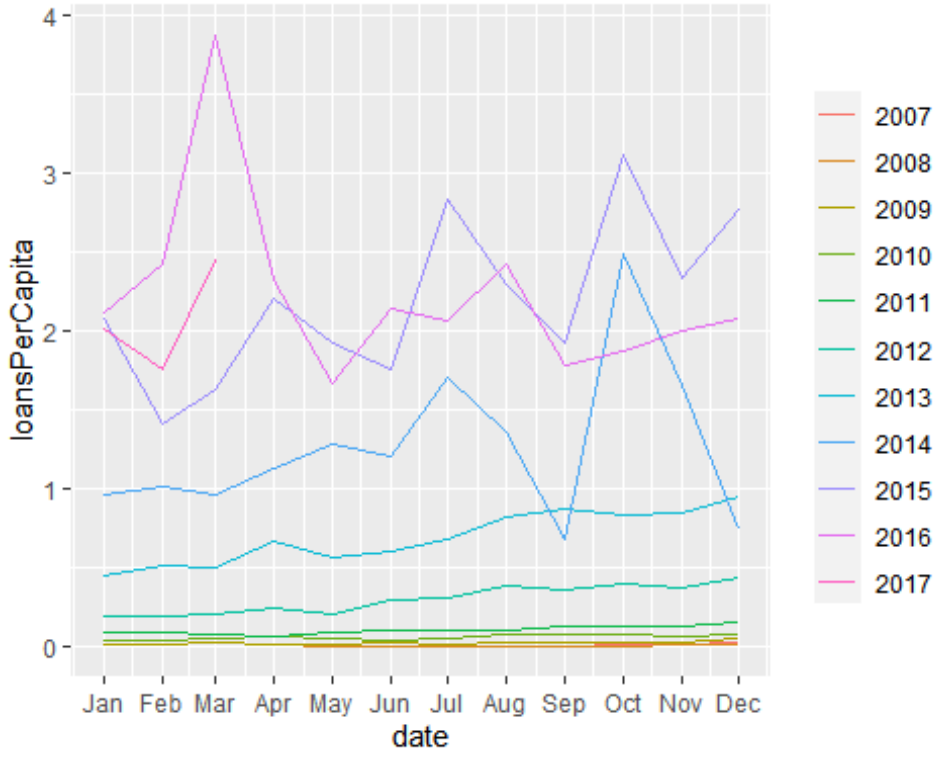
The all 3 states have similar number of anomalies and they occur between 2014 - 2017. Massachusetts has one anomaly prior to this period in 2013. New York has slightly more anomalies during the downtrend in 2016. This could be because of higher loansPerCapita in New York and more variance in the data.

c. STL decomposition to the loan per capita in NY.



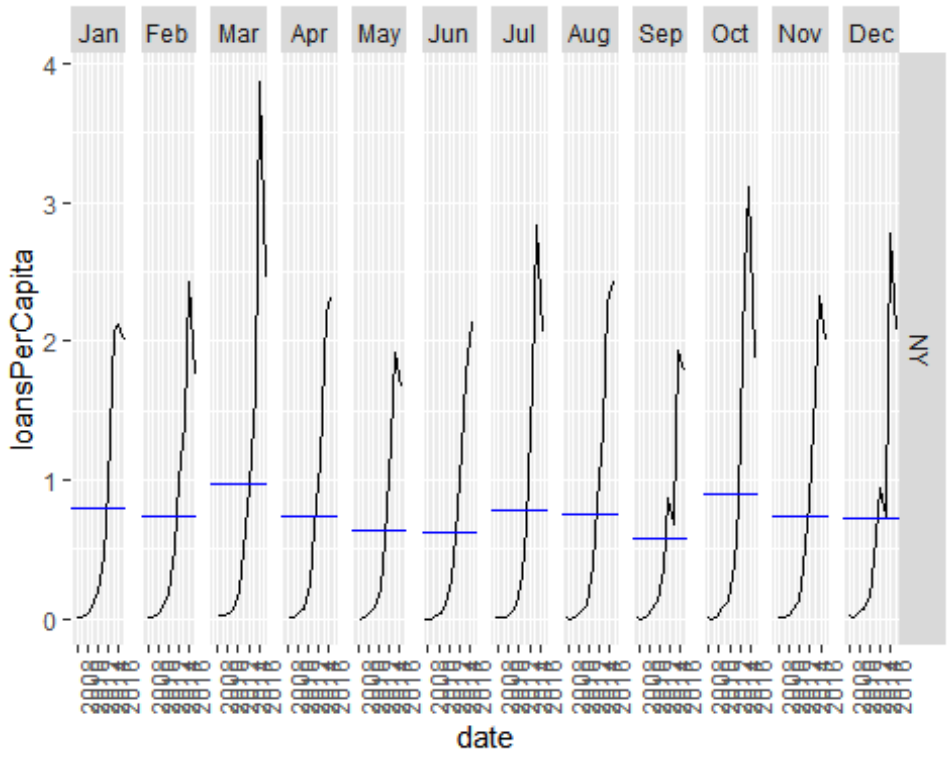
For the issued loans, the trend reverses around 2016 January. This could be because of an interest rate hike for the first time since the 2008 financial crisis when the rate had been cut down to almost 0.

d. Created a seasonal plot and a seasonal subseries plot for NY.



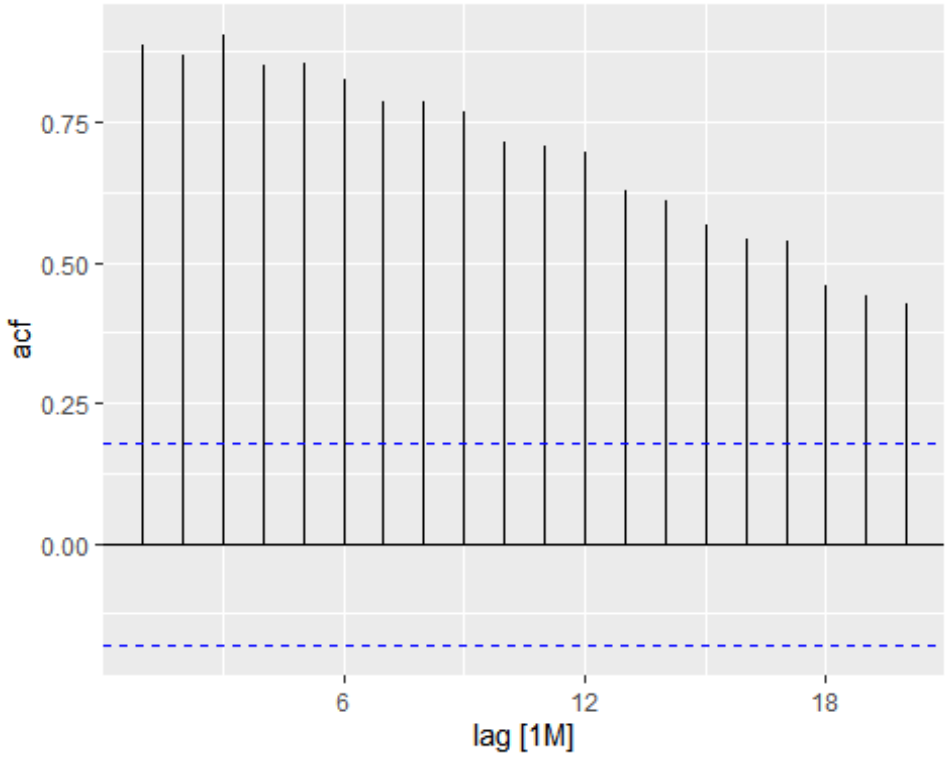
If we look at the seasonal plot for NY, we can see that the loansPerCapita is very less and negligible seasonality from 2007-2011, this could be because of the 2008 financial crisis. And it starts to go up from 2012 onwards. There is slight seasonality in 2012 and 2013. 2014 onwards there is seasonality but it is irregular. There seems to be an unusually high loanpercapita for March 2017.

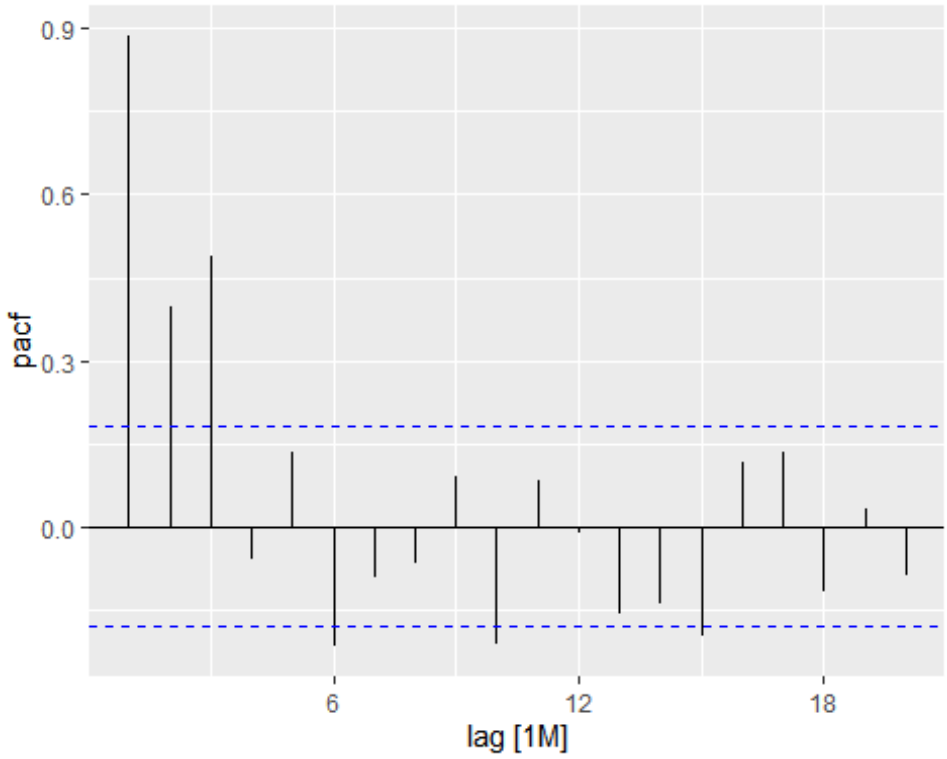
Even if we just consider the last 3 years, there does not seem to be a consistent seasonality. The peaks seem to be in March, July and October.



The subseries plot also reports similar results as compared to the seasonal plot, where in there is one major spike in March 2017. Also the mean loanpercapita is higher for the months of March, July and October.

e. Plotted the autocorrelation function and partial autocorrelation function results for NY.

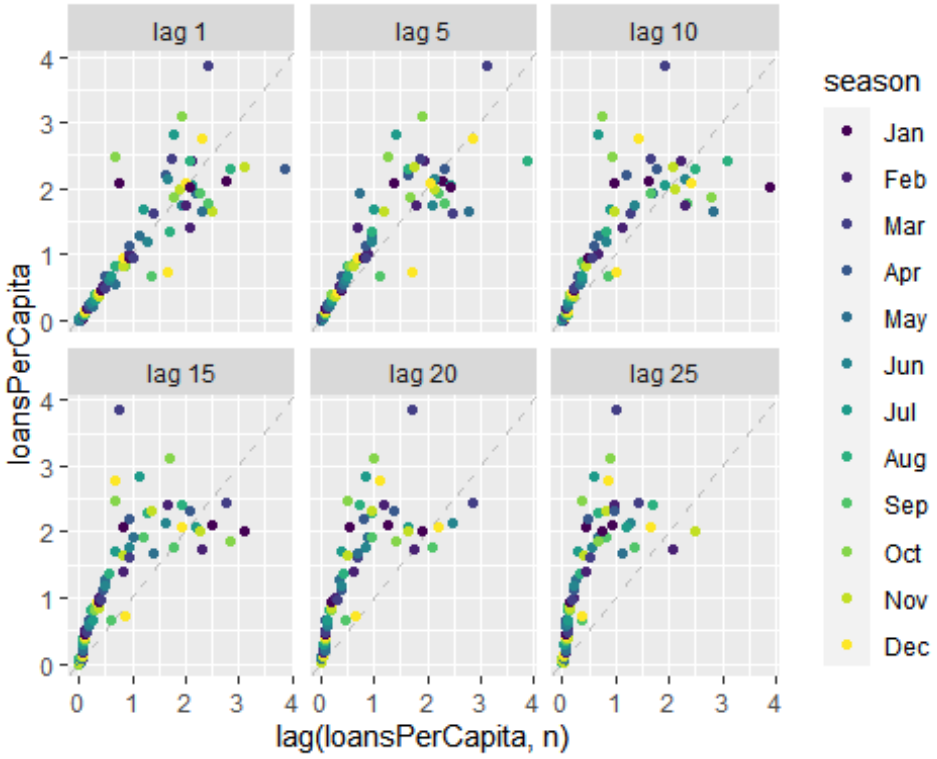




The ACF plot tells that there is high positive autocorrelation in loanPerCapita data. There are significant spikes at lags as far as 18. In the PACF plot which shows the autocorrelation at different lags after removing the effect of intermediate lags, we see that the autocorrelation

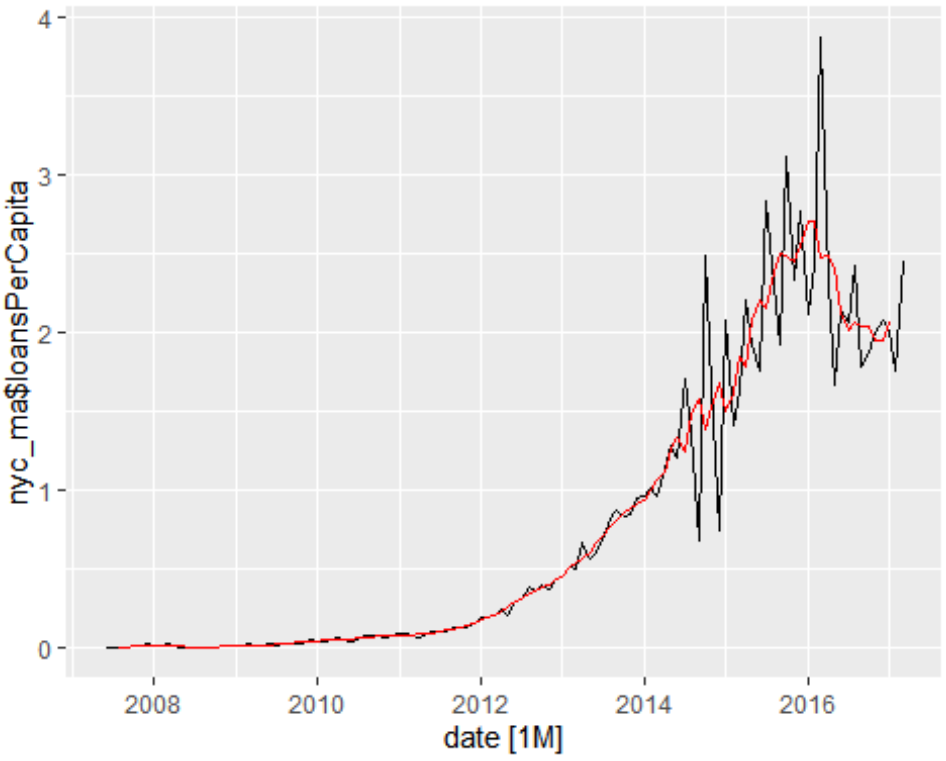
quickly falls to an insignificant value at lag 4 and last highly significant spike is at lag 3. This suggests that the loansPerCapita data might follow an ARIMA(3,d,0) model.

f. Created a lag plot for NY for the lags 1, 5, 10, 15, 20, 25.



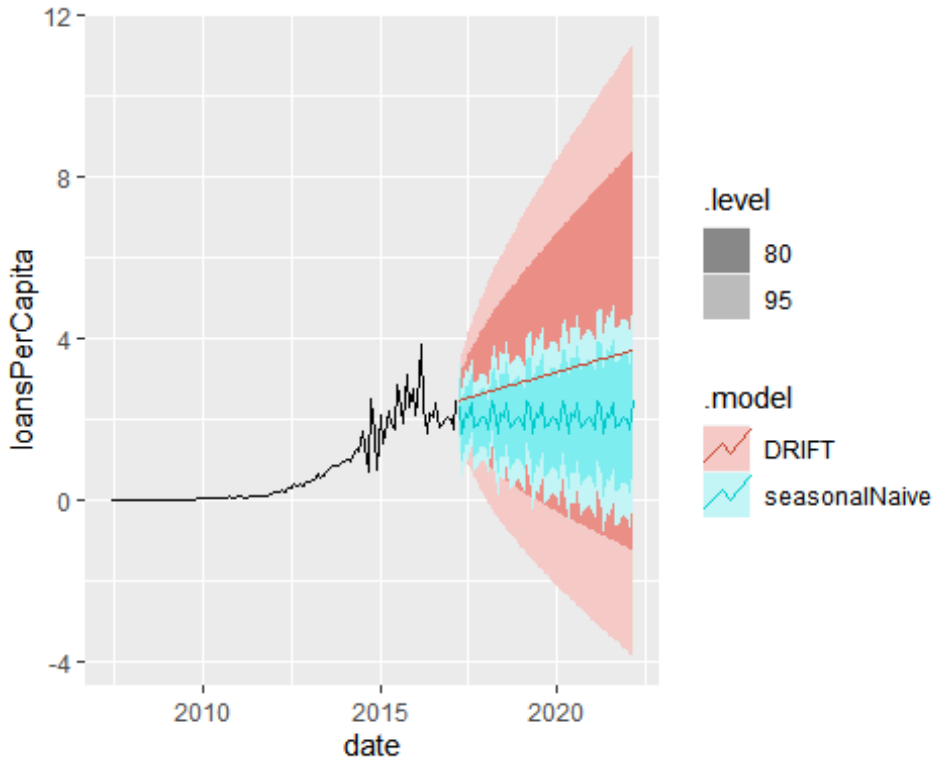
There is positive autocorrelation at all the given lags. The autocorrelation is the strongest at lag 1 and gradually decreases as the lag number increases (the points move away from the correlation = +1 line). The tightly clustered points are from the 2007-2011 data when the loanPerCapita is very less and there is negligible variance in the data.

g. Plotted the loans per capita in NY over time. Then, created a fifth order moving average smoothing and plotted the smoothed values on the actual loan data.



3. Modeling the loans issued in NY

a. Created a seasonal naive and drift forecast for NY data five years into the future.



If we take a look at the seasonal naïve forecast, it can be seen that it follows a similar pattern as compared to the original data, this is because in the seasonal naïve method each forecast is set to be equal to the last observed value from the same season.
Whereas in the drift forecast the trend seems to be increasing, this is because a variable called drift is added to the forecast which is the average change seen in the historical data.
The prediction intervals are much wider in the drift method and increase steeply as compared to the seasonal naïve method.

I think the seasonal naïve forecast is able to better capture the change in the amount of loanpercapita, because the width of the shaded regions which represent the 80% and 95% prediction interval is less for the seasonal naïve forecast, which suggests that it captures the change more accurately. Also, the seasonal naïve forecast captures the change in loans per capita across the seasons, showing seasonality. Whereas the drift method has aggressive forecasts and does not capture the seasonality.

b. Built a time series regression using both the time trend and seasons, as well as other variables that can be used to explain the loan issued per capita.

I have created a Time series regression using trend, seasons and the following variables: avgTerm, avgIntRate, avgGrade, avgEmpLength, avgAnnualInc, avgVerifStatus, avgHomeOwner , avgRevolBal , countOfLoans, NYCPI, NYUnemployment, NYCondoPriceIdx, NYSnapBenefits.

The regression model resulted in a R Squared value of 99.87 and an adjusted R Squared value of 99.84, which suggests that the above-mentioned dependent variables explain most of the variation of the loansPerCapita.

The variables with statistically significant coefficients are:

1) **avgIntRate**: This is the average interest rate of the loan, it is logical that this is a significant variable, because the interest rate of the loan can be a very important factor on whether a loan is taken or not, and it definitely can explain the variation in loansPerCapita. loansPerCapita is negatively related to the average interest rate.

2) **avgRevolBal**: The revolving balance can be considered as an important factor, because people with high revolving balance have high credit requirement and tend to take loans to pay off their balances, so it makes sense that it is a statistically significant variable.

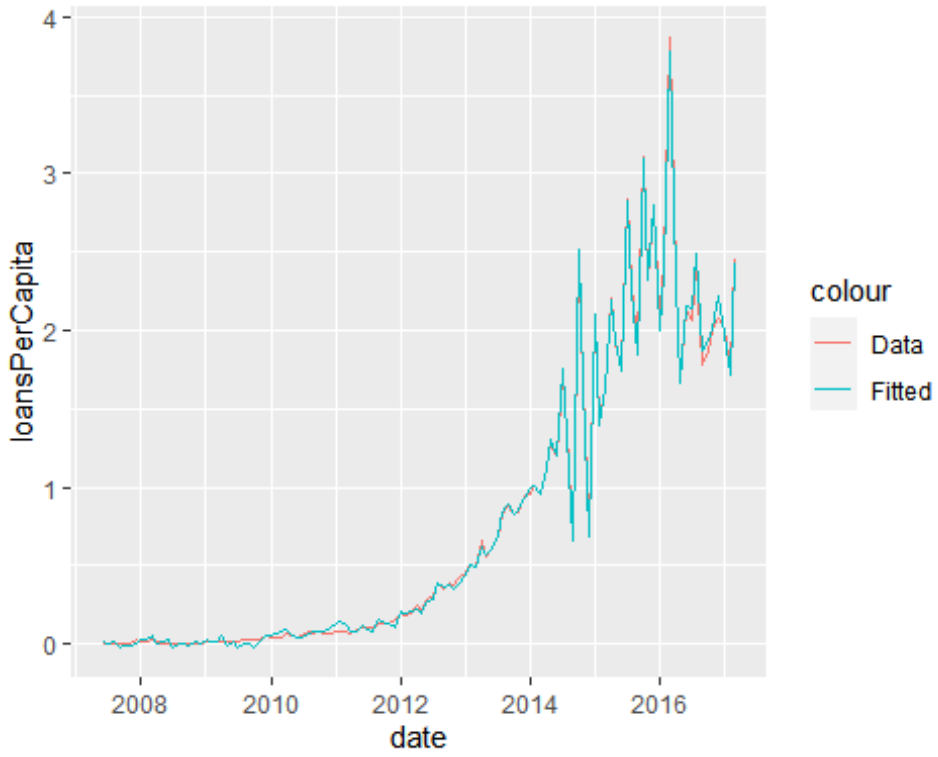
3) **CountOfLoans**: The number of loans granted in the state per month can definitely explain the variation in loansPerCapita. While holding all other variables constant, the loans per capita increases with increase in number of loans.

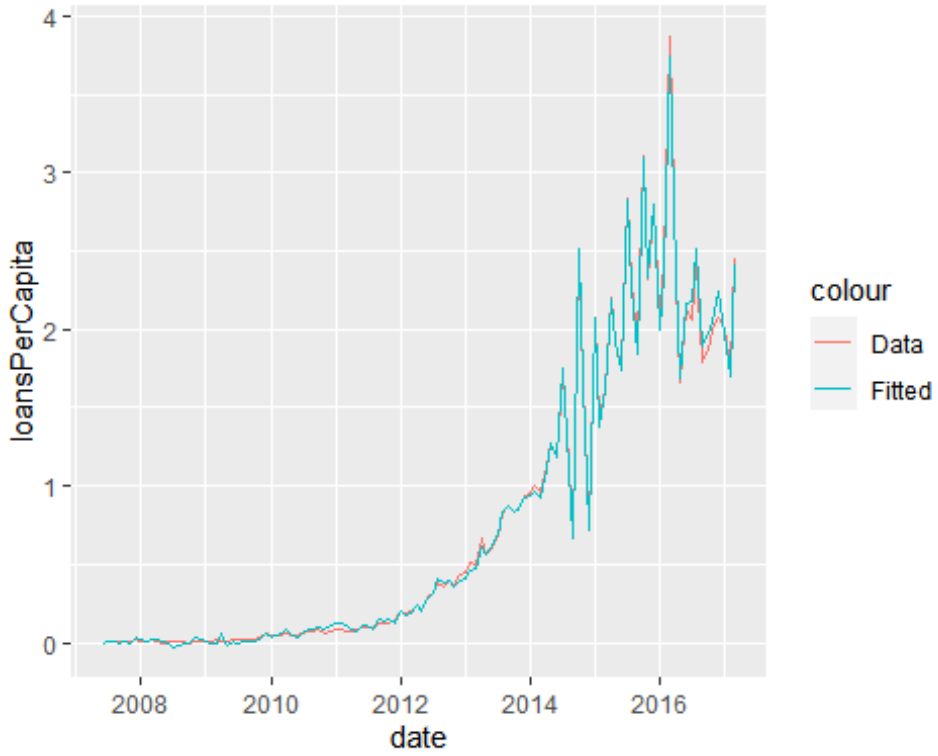
4) **NYSnapBenefits**: Increase in the number of recipients of social schemes like Snap Benefits indicates that more people tend to take loans for their needs. Hence, loansPerCapita increases.

5) **Trend**: Trend is a highly significant variable, this is because, if we take a look at the plot of the decision variable(loansPerCapita), a clear trend can be identified.

6) **Seasonality**: Similar to trend, if we plot the loansPerCapita, seasonality is pretty evident.

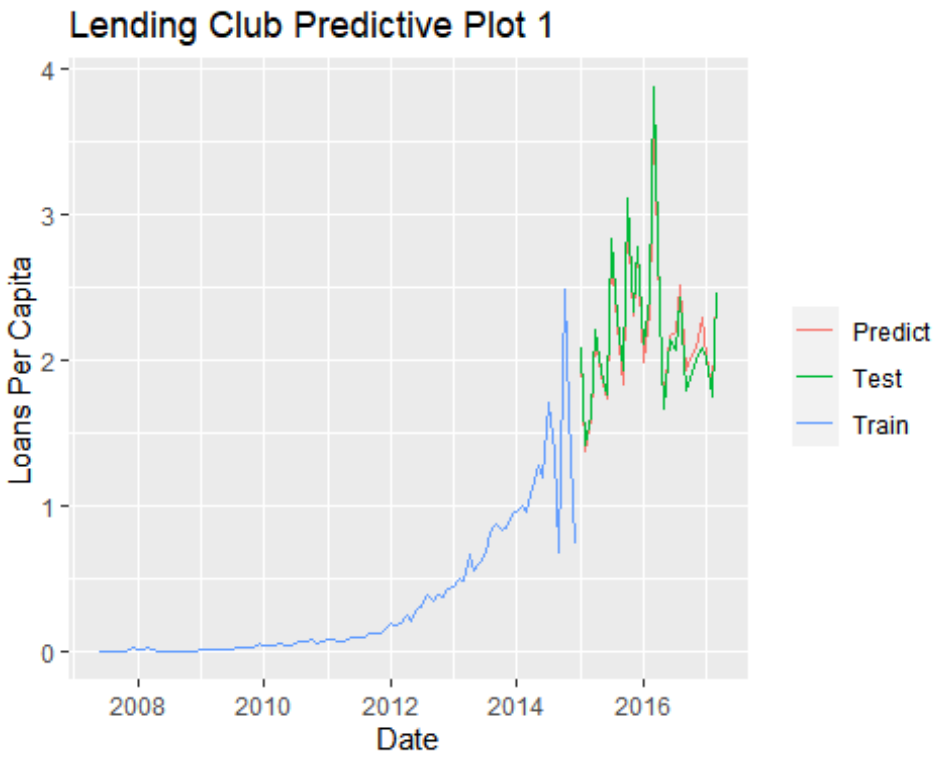
c. Plotted the fitted values from the model above and an alternative model excluding the time trend and seasons.

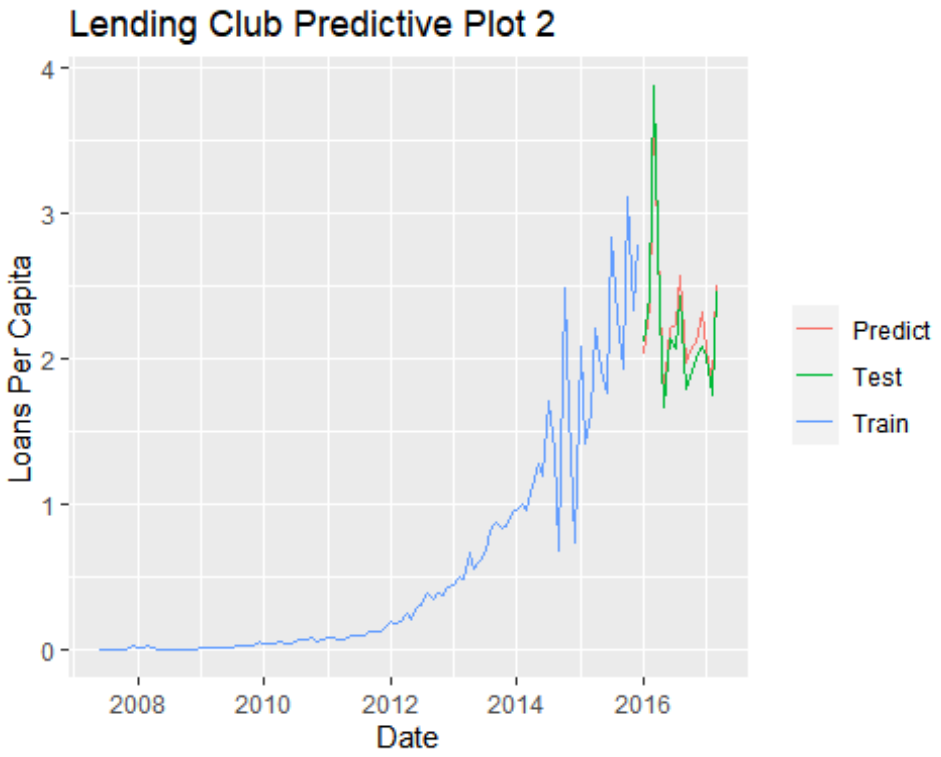

4
3
2
1
0
loansPerCapita
2008
2010
2012
2014
2016
date
colour
Data
Fitted


4
3
2
1
0
loansPerCapita
2008
2010
2012
2014
2016
date
colour
Data
Fitted

Both the models, the one from 3-b and the model which excluded the trend and seasons, the plot between fitted values and the actual values are pretty similar. Both the plots suggest that there is not much difference between the fitted values and the actual values. This is because of the high R-Squared values in the models, which suggests that the independent variables explain most of the variation in the dependent variable.

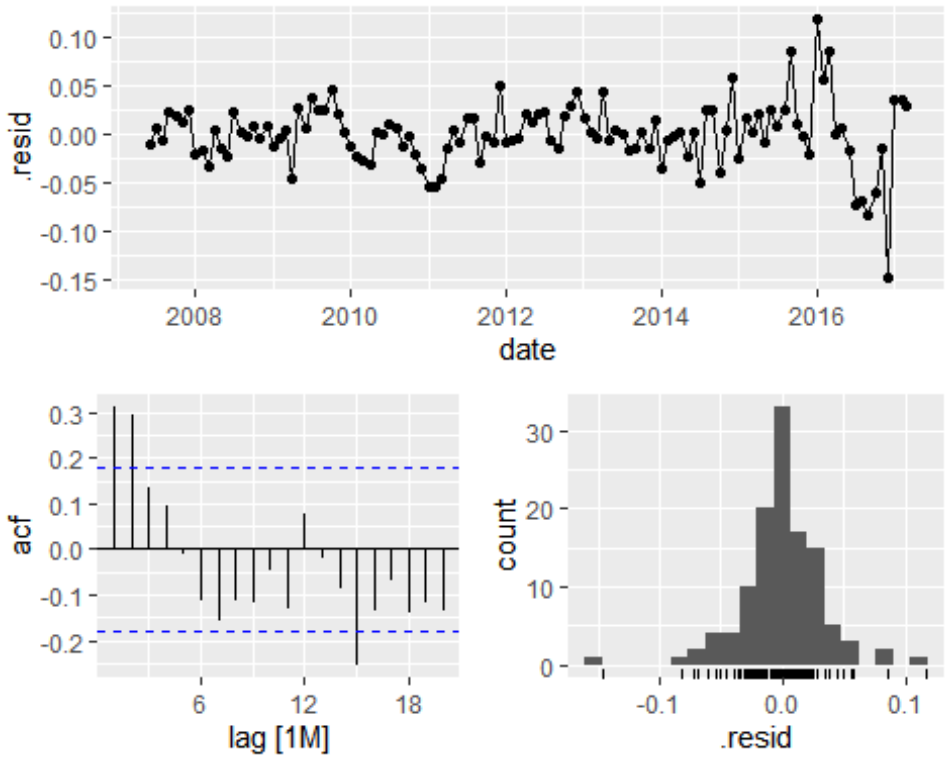
d. Created a predictive modeling plot using the model from (b) using two train/test splits. In the first split, use the data from 2014 and before for training, and in the second split, use the data from 2015 and before for training.





If we look at the predictive modeling plots of both the models, both the models predict the loansPerCapita pretty well and are very similar. In the first split, the predictions in 2015 are very accurate. When the trend reverses in 2016 (April, May) and sees a sharp fall, the predictions are little higher than the test data. In the second split, we see similar predictions as in the first split. The training data in the second and the first splits did not include the sharp down trend in 2016. Hence, the predictions are higher than the actual data.

e. **Examined the residual diagnostics for the model.**



The residual diagnostics look fine, because:

1) The residuals seem to be normally distributed.
2) The mean of the residuals seems to be very close to 0.
3) After analyzing the residuals plot, it can be said that the residuals have a constant variance, until the trend reversal in 2016.
4) The autocorrelation of residuals quickly falls to below the significance level to 0.

The above mentioned points suggest that the above time series model is a good one.

f. Built an ARIMA model using the same variables from (b) and using a grid search.

The ARIMA model using the same variables from 3-b and using a grid search has reported a (p,d,q) values of (1,0,2). When I ran the same model by including the parameters stepwise = FALSE, approximation = FALSE, to conduct a more through search, it has reported a (p,d,q) values of (4,00)).

Model with (p,d,q) = (4,0,0) is better since it has a lower AIC value of -464.65. The (1,0,2) model has an AIC value of -461.21.

The statistically significant variables (in the (4,0,0) model) are: avgIntRate, NYCPI and NYCondoPriceIdx. avgIntRate is the only variable, which is common in the above model, the other 2 variables are not reported as statistically significant in 3-b.

***g.* Checked the differencing suggested by the KPSS test.**

After checking for differencing by running the KPSS test, it has reported a d value of 1, which is different from the one obtained by the one generated by the previous ARIMA model.

h. Compared the new model performance with the model from (f).

Based on the AIC value, the new model performance (AIC = -453.32) is lower than the performance of the model generated in 3-f (AIC = -464.65). The new model has a higher AIC as compared to the previous model.

The only remedy that the KPSS test gives for addressing the non-stationarity issue is differencing the series. However, differencing is not always the best solution. Other solutions like demeaning a series might be more appropriate to address the non-stationarity issue. This could be the reason why the ARIMA model without differencing is a better model.

4. Predictive modeling of the loans issued in NY

a. Split the data into training (earlier than March, 2016) and test sets (on and after March, 2016). Built and compared the performance of the following models.

i. Time series regression with only trend and season
ii. Time series regression built in 3(b)
iii. ARIMA grid search model without any parameters
iv. ARIMA grid search model built in 3(f)

After comparing the performance of all the above models, it can be said that the Time Series regression built in 3-b is the best forecasting model, as it has the lowest RMSE value.

b. Split the data differently this time: training set (before April, 2016) and test set (on and after April, 2016). Built and compared the performance of the same models.

After comparing the performance of all the above models, it can be said that the ARIMA model built in 3-f is the best forecasting model, as it has the lowest RMSE value.

c. The only difference between the two sets of models (a) vs. (b) is that the second one uses one more month of data for training. Analyzed these two models.

There is a peak in March 2016 followed by a sharp decline and trend reversal in April, May 2016. In the second train set, the March 2016 peak is included. When the Time series

regression model sees the peak in March 2016, it immediately updates the model aggressively. This leads to higher predictions than the test set. This is evident from the negative mean error.

Whereas in the ARIMA model (4,0,0) the predictions are a weighted average of past 4 values. The March 2016 peak is smoothed into the model and does not aggressively influence the predictions. The predictions are higher than the test set (negative mean error), but not as high as the predictions with the time series regression model. ARIMA has low mean error than regression.