Problem Statement

Management team from the Bank of America would like to know from a Data Analyst, for all of their bank products, how many potential consumer complaints there could be in 2022

Data

Our data set contains the list of consumer complaints received by the Consumer Finance Protection Bureau (CFPB) about the financial product services offered by the Bank of America across the United States. The data set contains complaints received by the CFPB for (Bank of America) and it was retrieved from the www.consumerfinance.gov site (*https://www.consumerfinance.gov/data-research/consumer-complaints/search/).* I have filtered the date range from December 2011 till January 2021 and company name as 'Bank of America, National Association'.

The original data contains 98,021 observations of 18 variables. We have seen that (Date.recieved) was a (character) variable instead of date so, we need transform this variable to its appropriate type. I would like to perform time series on yearly data so I will extract Year, Month & Day and then later take monthly sums of the complaints across all the years. Finally, I have transformed the data to a timeseries so I can begin the time series analysis. I will slice the data such that it starts from December 2011 and ends in January 2021

Planning

I will perform a time series analysis from the chosen data set and thereby, constructing a model. Steps are given below:
Exploratory data analysis | Decomposition of data | Test the stationarity | Fit into a model | Calculate forecasts

Exploratory & Visualizations/Graphical Analysis

From **Figure 01**, 2012 starts off at a lower dip that has a spike for few months then shows a series of downward trend, sudden rise at the beginning of 2013 with a high spike and then again falling downward. This pattern is repeated throughout the time span. The overall data seems to have a weak seasonality & without a clear trend. The sharp spike in early 2013, it's clear that most consumers filed more complaints compared to the other years. This data does not seem to have autocovariance but we will run an additional test to verify our intuition.

Boxplots (in **Figure 02**) show complaints at an increasing drift and rising as months progress to mid-year then begins to fall towards the end of the year. Disassembling the data shows seasonality a bit clearer but still no clear trend.

Finally, the seasonal plot (**Figure 03**) shows the seasonal patterns a bit more clearly and makes it easy to spot that 2012 began with a very low minimum and early 2013 had a sharp spike.

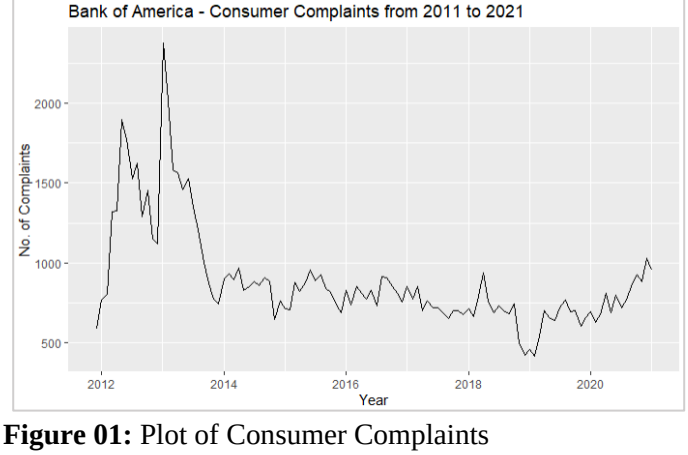

Figure 01: Plot of Consumer Complaints (Bank of America) from 2011 to 2021

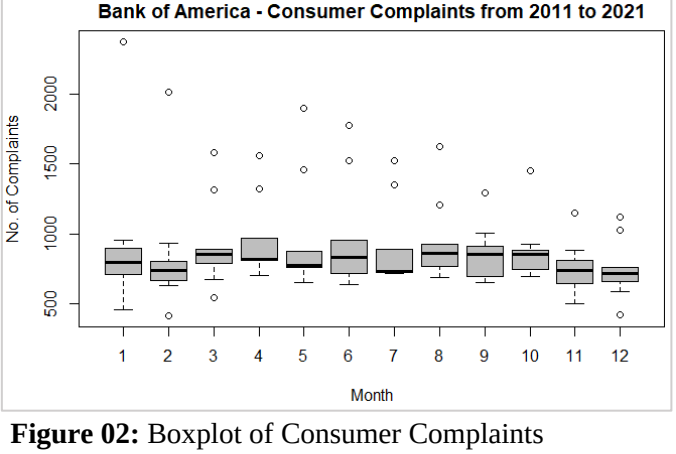

Figure 02: Boxplot of Consumer Complaints (Bank of America) from 2011 to 2021

Comparing Seasons – it does not provide any such specific reasons why most consumers complained in 2013 than rest of the years within this time span shown in **Figure 05** (Appendix).

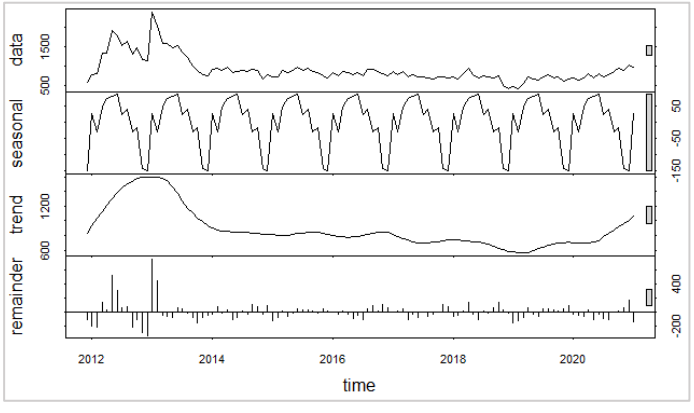

Figure 03: Seasonal Plot of Consumer Complaints (Bank of America) from 2011 to 2021

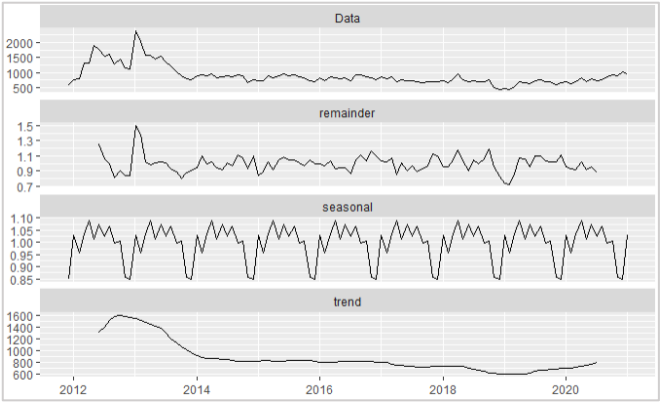

Figure 04: Decomposition plot of Consumer Complaints (Bank of America) from 2011 to 2021

Decomposition of Consumer Complaints Data

We have decomposed the data using multiplicative model. In these decomposed plots shown in **Figure 04**, it is clear that the time series is non-stationary (has random walks) because of seasonal effects and a trend (linear trend) and the estimation of the random component depicted under the "remainder".

Test Stationarity of the Time Series with (ADF Test) and (Autocorrelation)

To test the stationarity of the time series, let's run the **Augmented Dickey-Fuller** (ADF) Test. First, set hypothesis test: The null hypothesis $\mathbf{H_0}$: that the time series is non stationary. The alternative hypothesis $\mathbf{H_A}$: that time series is stationary. The output: (Appendix **Table 01**) p-value (0.4089) greater than 0.05 therefore, the data are not stationary. Test Stationarity of the Time Series (Autocorrelation): The analysis (in **Figure 06** and **Figure 07**) has shown the maximum at lag 1 or 12 months, indicates a positive relationship with the 12-month cycle. These tests tell us the correlation between points separated by various time lags. So, if we had to manually assess, which ARIMA models to use, a good point to start would be with p = 2 & q = 1 values or p = 1 and q = 0. The ACF residuals at 0 (**Figure 08-09**). The residual plots of ARIMA model (**Figure 10** and **Figure 11**) also centred around 0 as noise, with no pattern.

Model Building

It is recommended to fit a wide range of models and work out with a single model that's best performing based on some specific diagnostic parameters. STL + ETS model (**Figure 17**) forecasts seem better fitting the data. Structural models (**Figure 18**) seem to have linear trend. A naive model (**Figure 19**), which forecasts a flat line is also included in this analysis. We can also see that ARIMA models (**Figure 12**) forecasted well but data were not stationary. So, we have taken the 'log' value to make data stationary and build the (log) ARIMA model to predict the model.

Forecast and Model Accuracy

Predicted values are in 'log' hence converted it and then, plot the original and forecasted values (in **Figure 20**). We have split the original data set into 2 and fit the ARIMA model into the first one. Then compared both the values, and found that the predicted values are very close to the original values (in **Table 03**). Thus, our model is justifiable.

Conclusion

We have used statistical method for time series forecasting using the ARIMA model. The model's predicted potential complaints from consumers in 2022 has been shown to the Bank Management Team from the (**Table 03 - Appendix**).

Appendix:

Bibliographic Reference

Kimnewzealand (September 25, 2017) *Time Series Analysis and Modeling with the Air Passengers Dataset* Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/311446_08b00d63cc794e158b1f4763eb70d43a.html

Raut, N. (2020) *Air Passengers Forecast* Retrieved from https://rpubs.com/neharaut05/TimeSeries_AirPassangerForecast

Variawa, R. (October 31, 2019) *Time Series Analysis* Retrieved from https://rpubs.com/Ryder/555527

```
        Augmented Dickey-Fuller Test

data:  cc
Dickey-Fuller = -2.4042, Lag order = 4, p-value = 0.4089
alternative hypothesis: stationary
```

Table 01: Augmented Dickey-Fuller (ADF) Test Results

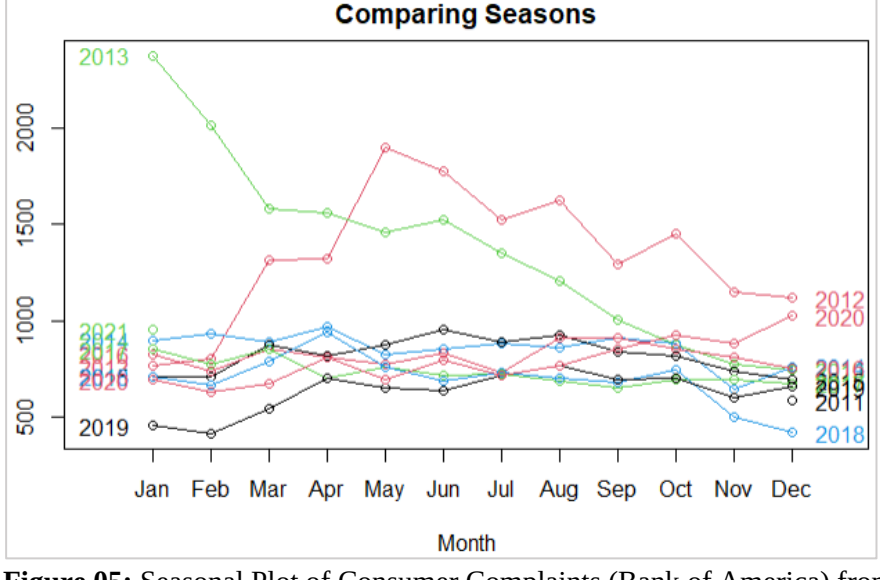

Figure 05: Seasonal Plot of Consumer Complaints (Bank of America) from 2011 to 2021

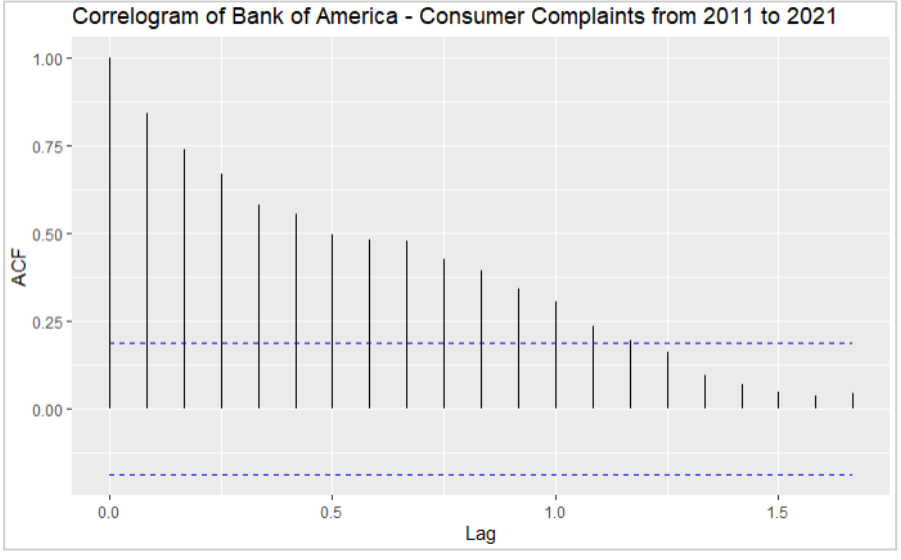

Figure 06: ACF Test - Correlogram of Bank of America - Consumer Complaints from 2011 to 2021

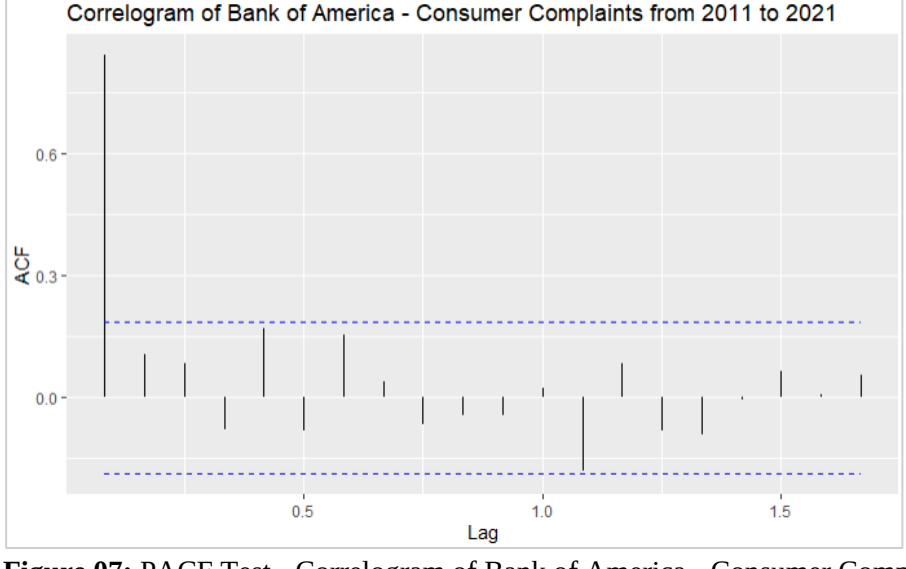

Figure 07: PACF Test - Correlogram of Bank of America - Consumer Complaints from 2011 to 2021

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011												NA
2012	NA	NA	NA	NA	NA	1.2583106	1.0586151	0.9980258	0.8189896	0.8994813	0.8436491	0.8402316
2013	1.4918005	1.3822767	1.0209378	0.9825959	1.0124665	1.0238167	1.0027395	0.9367191	0.8910485	0.8048154	0.8806443	0.9021797
2014	0.9478912	1.0950519	0.9906300	1.0232043	0.9426805	0.9225495	1.0076315	0.9634838	1.1076053	1.0762637	0.9257176	1.0954450
2015	0.8398973	0.8936484	1.0225968	0.9155747	1.0507228	1.0808486	1.0513921	1.0435510	1.0130765	0.9733090	1.0443658	0.9983117
2016	1.0012627	0.9674787	1.0286210	0.9280071	0.9410109	0.9493075	0.8672069	1.0433350	1.1099114	1.0390487	1.1658559	1.1003930
2017	1.0318939	1.0232036	1.0669836	0.8513943	1.0107417	0.9056308	0.9660599	0.8984957	0.9261154	0.9673273	1.1245594	1.0960536
2018	0.9556451	0.9545668	1.0520777	1.1793703	1.0394318	0.9078702	1.0436783	0.9920213	1.0632606	1.1865322	0.9585499	0.8250852
2019	0.7419327	0.7226437	0.8674214	1.0681837	1.0606930	0.9607542	1.0930486	1.0942837	1.0399081	1.0231378	1.0265734	1.1122157
2020	0.9595919	0.9310456	0.9206472	1.0215851	0.9121681	0.9570668	0.8757830	NA	NA	NA	NA	NA
2021	NA											

Table 02: Random time series for any missing values - Bank of America - Consumer Complaints from 2011 to 2021

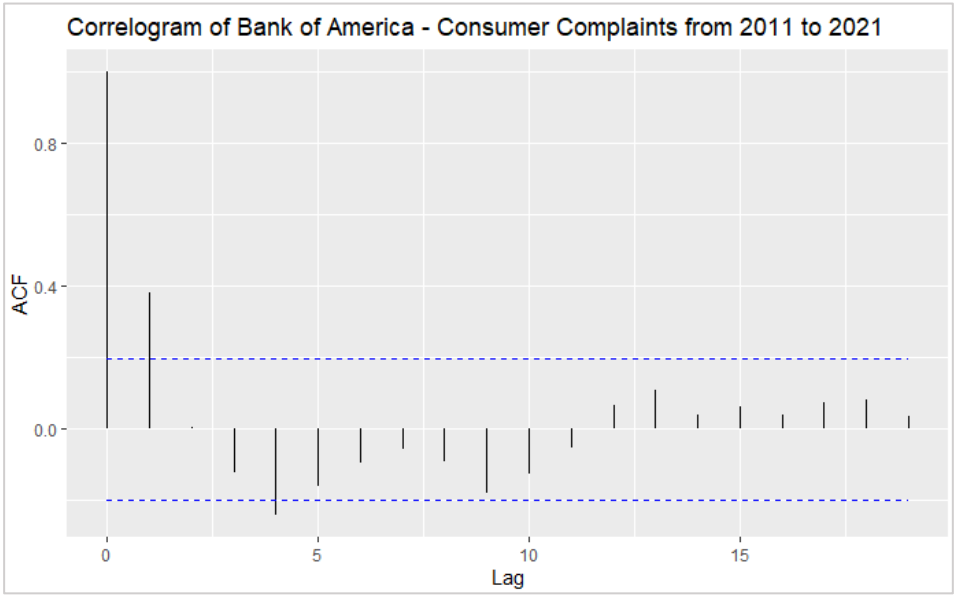

Figure 08: Autoplot the random Time Series (ACF) - Correlogram of Bank of America - Consumer Complaints from 2011 to 2021

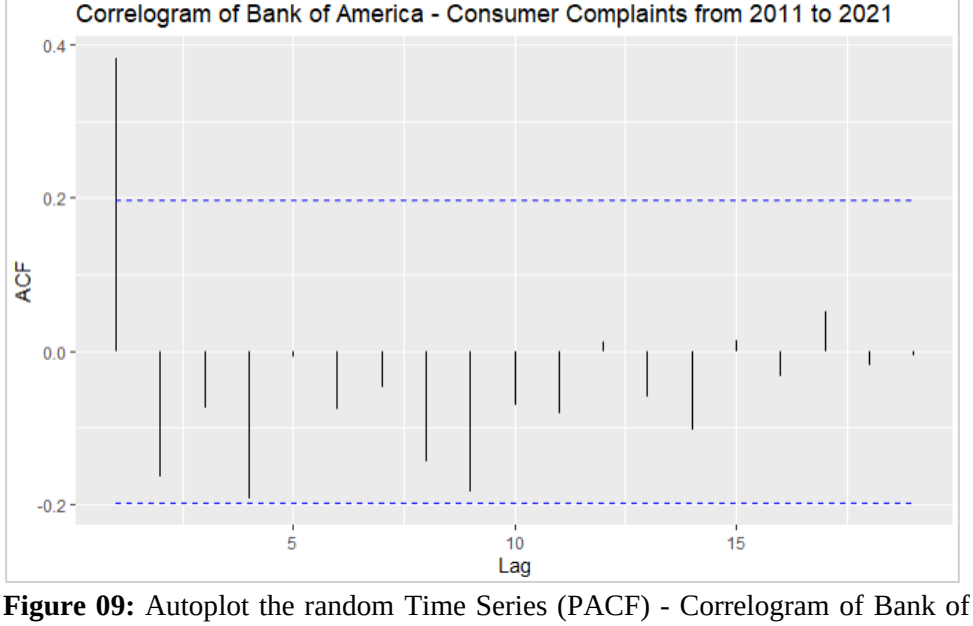

Figure 09: Autoplot the random Time Series (PACF) - Correlogram of Bank of America - Consumer Complaints from 2011 to 2021

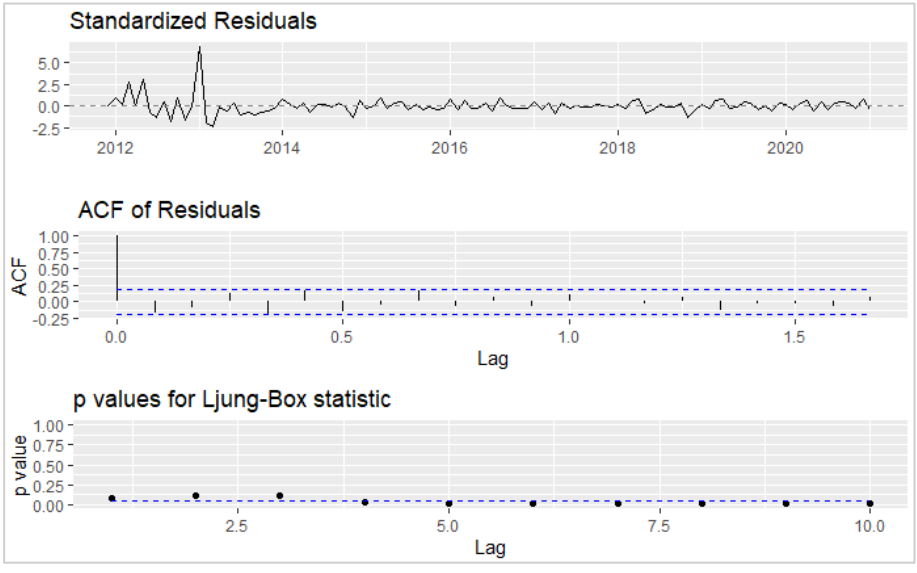

Figure 10: Residual plots of ARIMA model

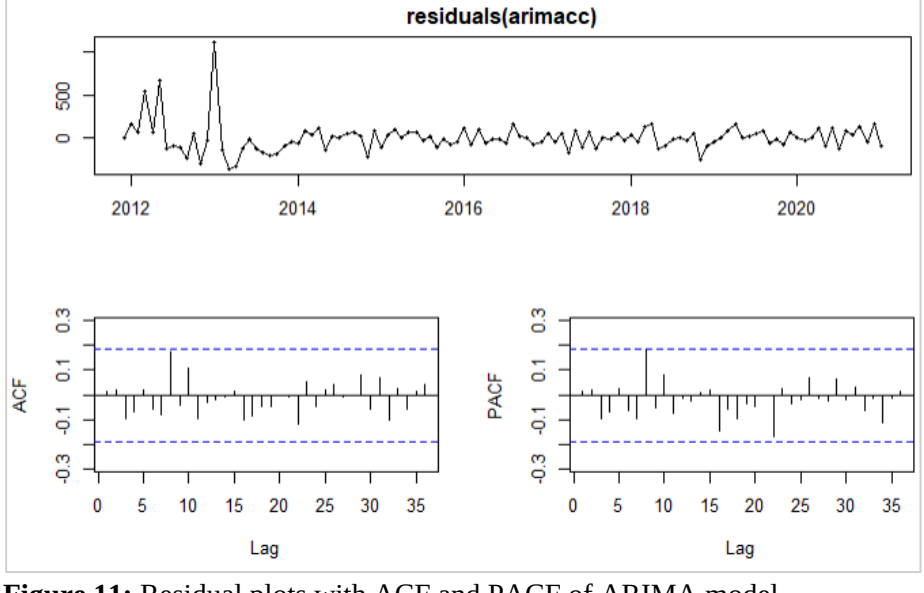

Figure 11: Residual plots with ACF and PACF of ARIMA model

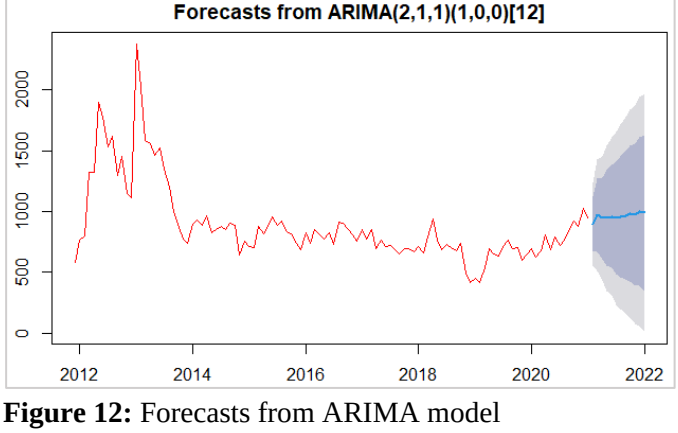

Figure 12: Forecasts from ARIMA model

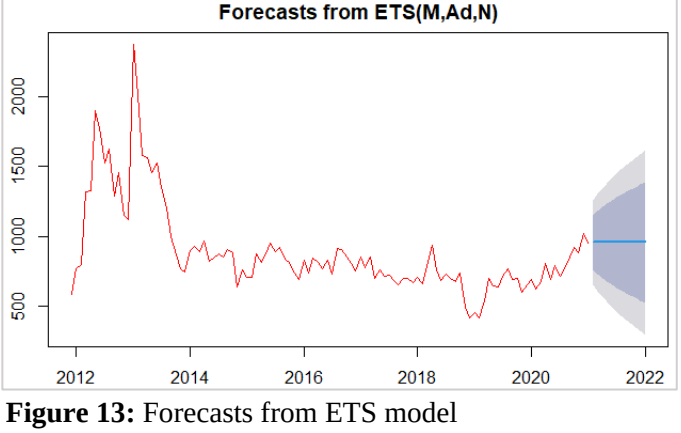

Figure 13: Forecasts from ETS model

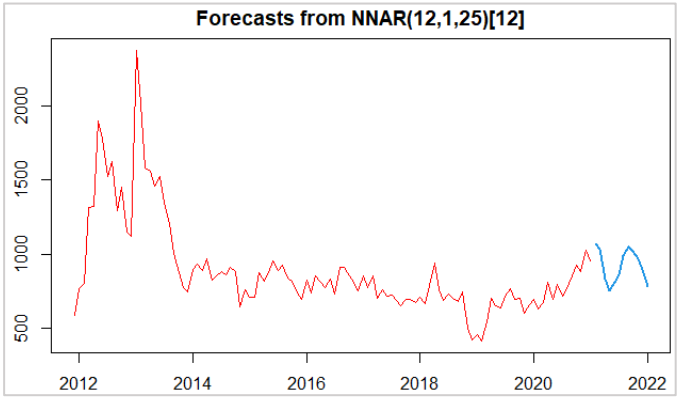

Figure 14: Forecasts from ARIMA model

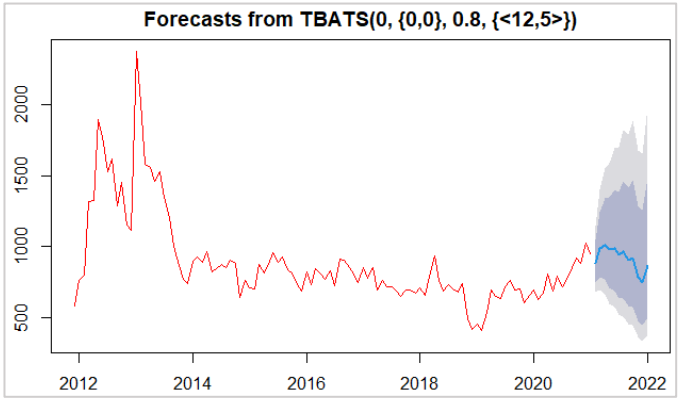

Figure 15: Forecasts from ETS model

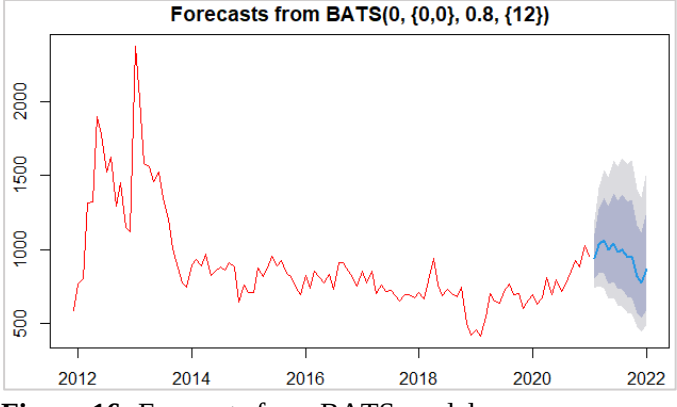

Figure 16: Forecasts from BATS model

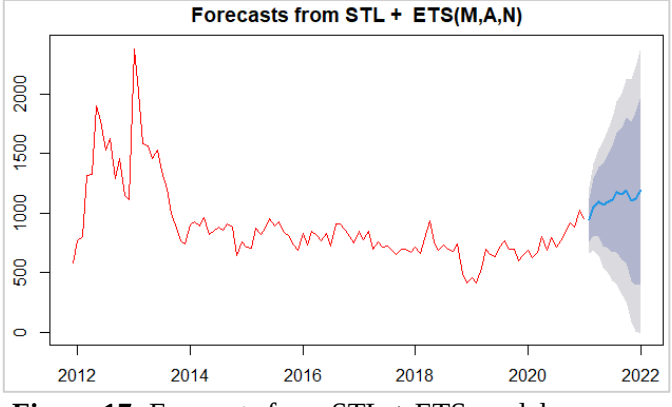

Figure 17: Forecasts from STL + ETS model

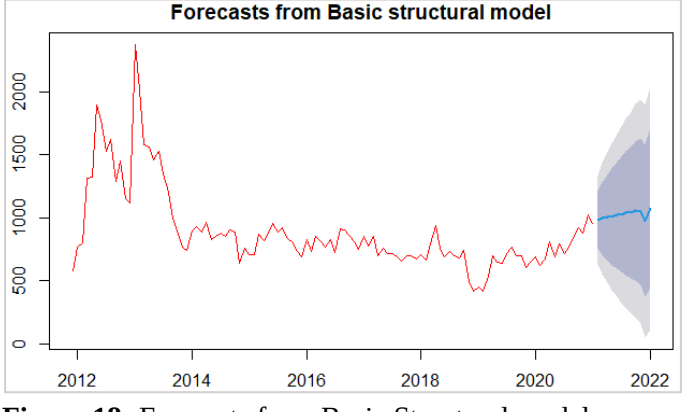

Figure 18: Forecasts from Basic Structural model

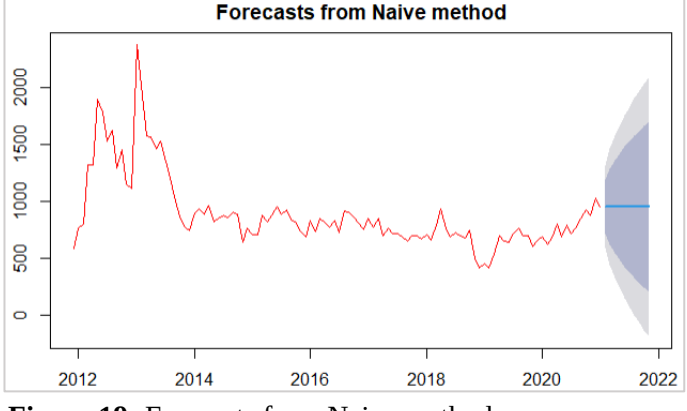

Figure 19: Forecasts from Naive method

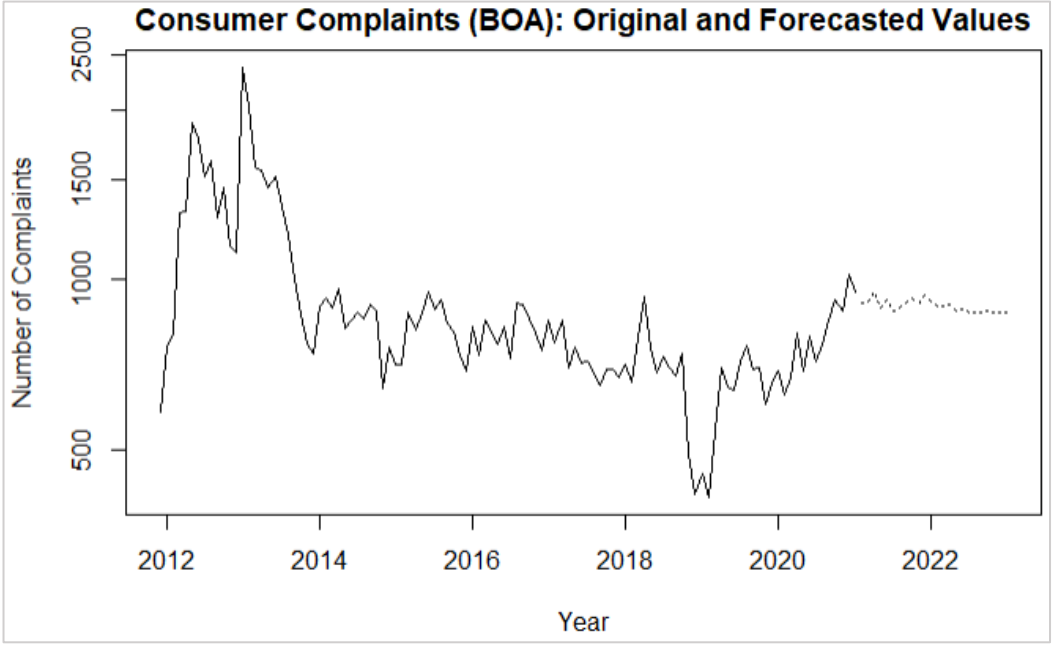

Figure 20: Consumer Complaints (BOA): Original and Forecasted Values

```
#Predicted values
pred_2022 <- round(predicted_values_test_converted, digits=0)
pred_2022

```

```
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2021     912 916 951 896 922 885 894 913 927 907 943
2022 917 900 897 903 884 888 875 875 878 880 872 880
```

```
#Original values
orig_2022 <- tail(cc, 12)
orig_2022

```

```
     Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
2020      626  675  806  690  794  718  769  851  923  880 1025
2021  952
```

Table 03: Original and Predicted values of Consumer Complaints – Bank of America