Individual Assignment 6, Dion Chang - 20812576

1) Problem statement and data used: The public consumer complaints data used for analysis had over 2 million rows of complaints ranging from the December 01, 2011 up to March 14, 2021. The objective was to determine how many potential complaints we would get from consumers in 2022 using time series analysis.

2) Planning: The data was aggregated by having one column to be the date (year-month) and the second column to be the total number of complaints for that date. The last 12 observations showed a sudden increase in complaints after March 2020. Therefore, in order to show a clear upward trend of the time plot as shown in Figure 1 those observations were filtered out. Figure 1 clearly shows an increasing and seasonal trend as the complaints go up over time and a dip can be observed near the end of each year. The positive trend and seasonality was further confirmed after performing a decomposing the data with the multiplicative model (graphs not shown in this report) as the appropriate choice.

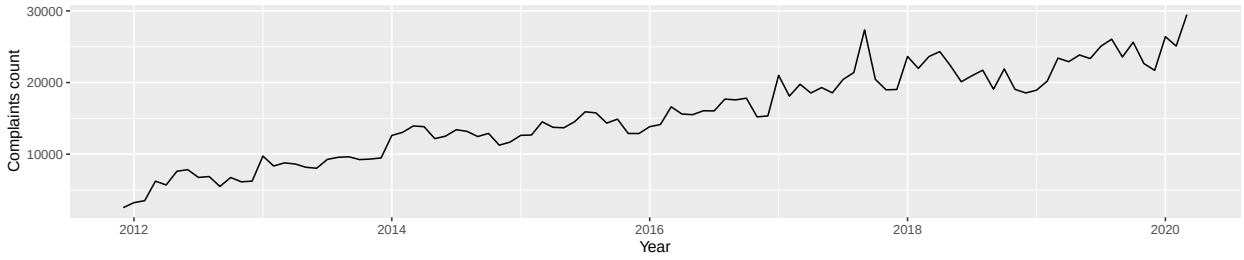

Figure 1. Original consumer complaints data from December 2011 to March 2020.

Since time series is required to be stationary in order to fit ARIMA models, it was necessary to convert the complaints data since it was initially non-stationary. Log (to remove unequal variances) and differentiation (to get constant mean) functions were applied in R to perform these tasks. Figure 2 shows the plot of the data after making it stationary, which confirmed that the data did not have a trend and was constant over time.

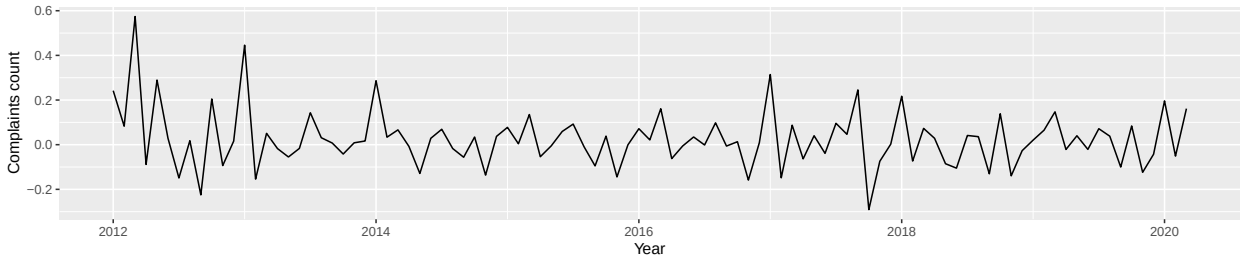

Figure 2. Stationarized consumer complaints data from December 2011 to March 2020.

We need to fit an ARIMA model in order to make a forecast of the consumer complaints. Therefore, autocorrelation plots (ACF and PACF) and the Dickey-fuller test are used to test the stationarity of the time series.

3) Analysis: From the **Dickey-Fuller test,** it was confirmed that the original data was not stationary (lag order = 12, p-value = .26). The lag length was determined since the our data is recorded monthly. Therefore, the time series data was converted to stationary as shown in Figure 2.

ACF and PACF plots for non-stationary and stationary data: The ACF plot (Figure 3) was used to determine the q (moving average) value while the PACF plot (Figure 4) was used to determine the p (auto regressive lags) value. The d (order of differentiation) was 1, which represented the number of differencing required to make the consumer complaints time series data stationary. It was determined that q = 1, and p = 0.

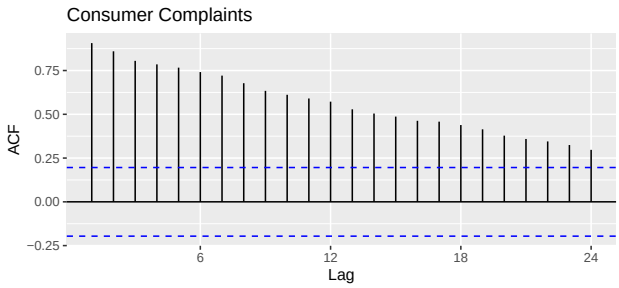

Figure 3. ACF Plot of the consumer complaints time series data.

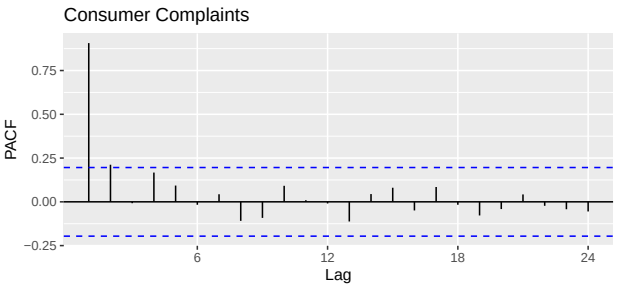

Figure 4. PACF Plot of the consumer complaints time series data.

The p and q values from the ACF and PACF plots was then used to help fit the ARIMA model. The model and forecast are plotted as shown in Figure 5:

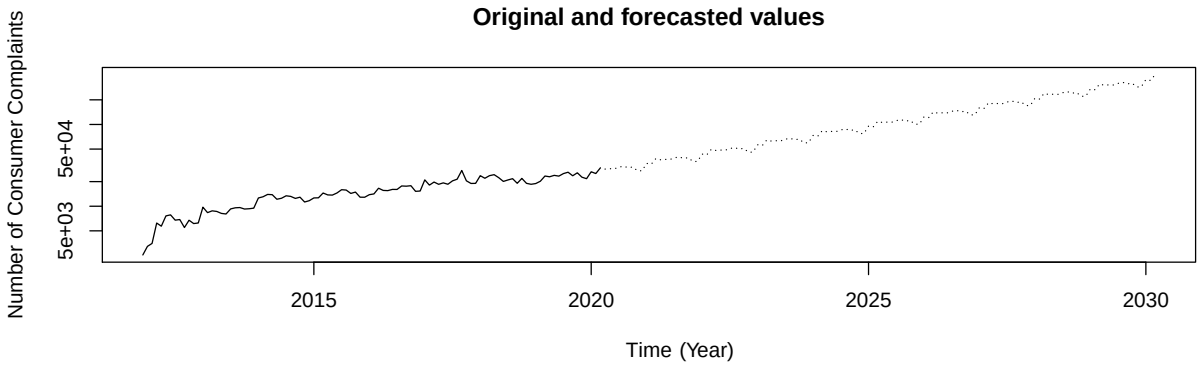

Figure 5. Fitted ARIMA model and forecast of consumer complaints time series data.

Complaints forecast: The forecast in Figure 5 shows a seasonal pattern with and upward trend. The total sum of consumer complaints from January 2022 to December 2022 was calculated to be 576444. To test the validity of the model, the predicted values and original values of the year 2019 were compared as shown in Table 1:

Table 1: Testing the ARIMA model of the consumer complaints predicted vs. original values for year 2019.

Date	Predicted	Original
Jan 2019	23431	18932
Feb 2019	22409	20204
Mar 2019	25797	23410
Apr 2019	24975	22913
May 2019	24815	23848
Jun 2019	24457	23350
Jul 2019	25629	25083
Aug 2019	26503	26057
Sep 2019	25878	23573
Oct 2019	25969	25635
Nov 2019	23226	22650
Dec 2019	23281	21704

4) Conclusion: We have successfully fit a time series model that can forecast the number of consumer complaints. The model predicted the total number of potential complaints we would get in 2022 would be **576444**. This model is justifiable as shown in Table 1, we can see that the predictions are very close to the original values.