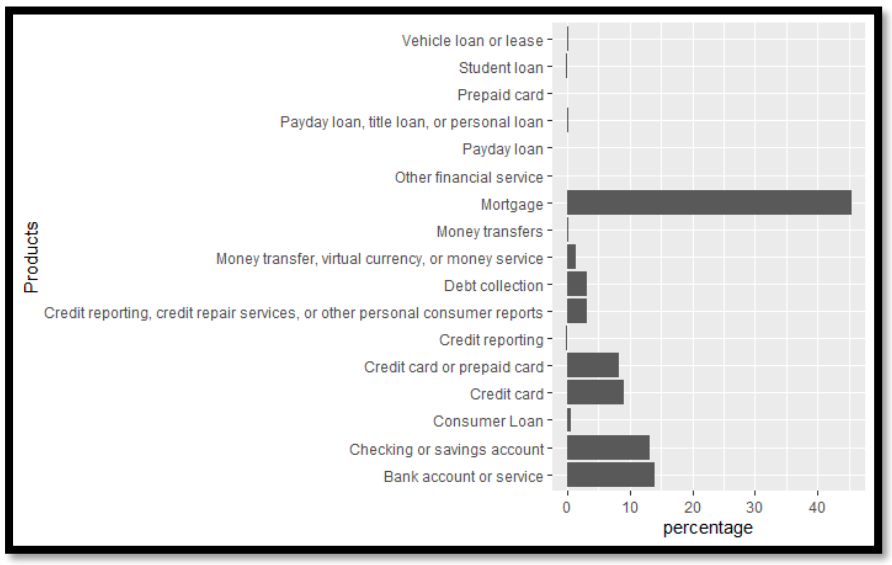
1. Problem Statement and Data Summarization:

- **Problem Statement**: Prediction of potential complaints received from consumers received by Bank of America in year 2022.
- **Dataset summary:** This dataset is taken from source https://www.consumerfinance.gov/data-research/consumer-complaints/ which reflects the total consumer complaints received from consumers at Bank of America from December 2011 to March 2021. There are total 99215 observations and 18 variables in this dataset. There are no null values in this dataset. Thus, this dataset is tidy and can be used for further Exploratory Data Analysis. The subject-wise distribution of total complaints registered at BOA from December 2011 to March 2021 is as follows:

Figure 1: Percentage of complaints registered under each Product



It can be visualize that maximum number of consumer complaints (44985) are registered under the subject "mortgage" at Bank of America. Also, least number of complaints (14) are registered under "payday loan".

2. Planning:

This dataset is filtered to further conduct the analysis as per the problem statement. The number of observations remains the same i.e. 99215. Out of 18, only 2 variables are used for further analysis. They are as follows:

- Date.recieved: The date (mm/dd/yyyy) on which the complaint is received at BOA.
- Product: The subject of complaint.

Using lubridate function, the variable Date.recieved is further bifurcated into Day, Month and Year. As an outcome, the number of variables increases to 4. With the help of times series function, the total number of complaints registered from December 2011 to March 2021 are segregated month wise. Thus, the final dataset ready for analysis has total 112 observations and 3 variables (Month, Year, Total_complaints).

Figure 2: Year vs Total complaints at BOA between 2011-2021

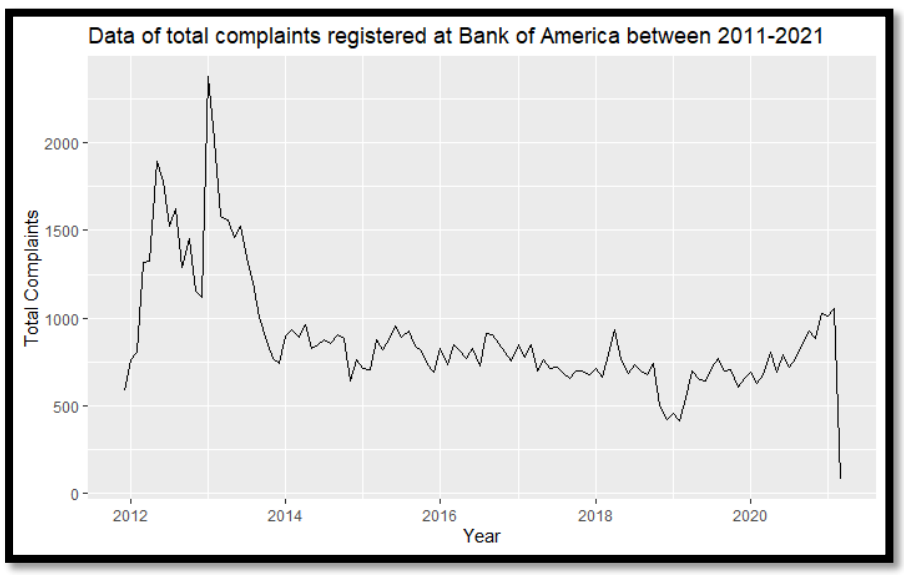
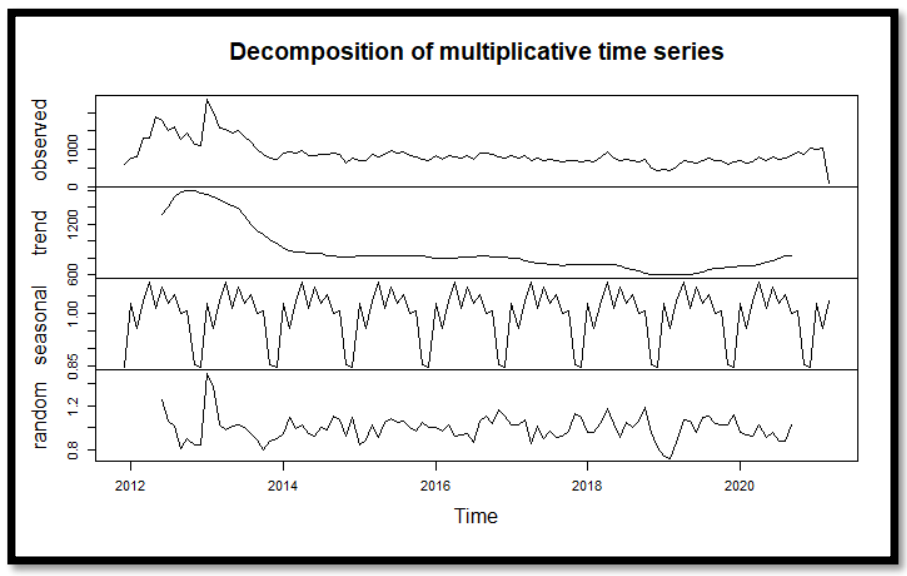


Figure 3: Decomposition of Multiplicative time series



From Figure 2, it can be visualised that the total consumer complaints show an increasing trend from year 2012 to 2012 mid-year. After mid-year 2012 to 2013, the complaints start to decrease. From 2013 to 2014, the data of total complaints shows a huge spike which means that the complaints were increased at the starting of 2013 and then decreased by the end of year 2014. From year 2014 to 2020, the total complaints registered yearly remains between 1000 to 500. After 2020, the data shows a sudden decreasing spike indicating that the number of complaints registered in year 2021 (until March) is very less.

From figure 2 and figure 3, we can visualize that the seasonal trend variation is decreasing in size over time and seasonal fluctuations depend on the level of time series. Thus, a multiplicative decomposition method is best fitted for the model.

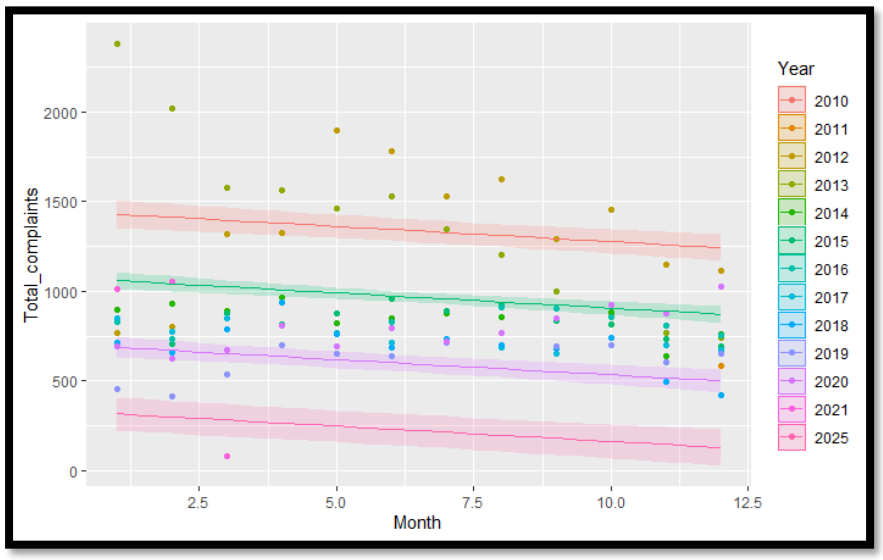
3. Analysis:

A multiple linear regression model is to be designed to predict the total consumer complaints in year 2022. In the dataset of 112 observations and 3 variables, there are two predictor variables: Month and Year. Also, there is one outcome variable: Total_complaints. The assumptions for multiple linear regression model are as follows:

- The predictor variables "Year" and "Month" are categorical and outcome variable "Total_complaints" is quantitative, continuous and unbounded. Thus, 1st assumption is fulfilled.
- All the 3 variables have non-zero variances. Thus, 2nd assumption is fulfilled.

- There is no perfect multicollinearity between predictor variables Year and Month as the correlation coefficient is -0.08870047 which is not significant. Thus, 3rd assumption is fulfilled.
- Predictors are uncorrelated with external variables. Mostly for interpretation and validity of the predicted values this can be a problem. So, for now it is assumed that there aren't any external variables. Thus, 4th assumption is fulfilled.
- Residuals are homoscedastic (constant variance), independent (test with Durbin-Watson) and Normal. This can be computed after the model is developed.
- Linearity: Below visualization in figure 3 shows that outcome variable: Total_complaints show a linear relationship.

Figure 4: Month vs Total Complaints for year 2011-2021



From the above plot, it can be predicted that total number of consumer complaints registered at BOA are decreasing from year 2010 to 2025. Thus, it can be further predicted that the total number of consumer complaints registered in year 2022 might be less compared to the past years. This can be justified using multiple linear regression model.

Thus, all assumptions for a multiple regression model are fulfilled. The formula for multiple regression model can be defined as B0 + B1*X1 + B2*X2; where B0 is the intercept, B1 and B2 are coefficients of predictor variables X1 = Month and X2 = Year respectively.

- **Null hypothesis H0:** The slope between predictor variables (B1, B2) and outcome variable is equal to zero. (B1 = 0, B2 = 0).
- **Alternative hypothesis H1:** The slope between predictor variables (B1, B2) and outcome variable is not equal to zero. (B1 ≠ 0, B2 ≠ 0).

If the relationship between Month, Year and Total_complaints is significant then the slope will not equal zero. The intercept and model coefficient values obtained are as follows:

Table 1: Intercept and Coefficients of variables

Coefficient	Name of variable	Value
B0	Intercept	150675.34
B1	Month	-16.97
B2	Year	-74.24

The negative coefficients indicates that with the increase in years, the total consumer complaints registered at BOA is decreasing. This prediction is supported by figure 2 and figure 3.
The summary of the model is as mentioned below:

Table 2: Summary of Model

Function	Value
p-value	1.102e-11
Residual Standard Error (RSE)	269.9 on 109 degrees of freedom
F-statistics	32.09 on 2 and 109 degrees of freedom
The 95% confidence intervals of coefficient B1 for "Month"	-15.18859 to 20.86715
The 95% confidence intervals of coefficient B2 for "Year"	-56.96306 to -14.56608.

Here, it can be interpreted that p-value for this multiple linear regression model is closer to zero value which is a significant value. Thus, it can be said that this model is better fitted. The F-statistics indicates that the slope between predictor variables and outcome variable in not equal to zero and hence there is a relation between predictor variables and outcome variable. Thus, null hypothesis is rejected here.

4. Conclusion:

- All the assumptions are fulfilled for the multiple linear regression model. The Alternative hypothesis is true in case of multiple linear regression model which means that there is a linear relationship between predictor variables (Month, Year) and outcome variable (Total_complaints).
- From figure 4, it can be predicted that total number of consumer complaints registered at BOA are decreasing from year 2010 to 2025. Thus, it can be further predicted that the total number of consumer complaints registered in year 2022 might be less compared to the past years.
- From the dataset it is observed that in case of month March 2021, the complaints registered for first 3 days (79 complaints) are reflected. Thus, from the multiple regression model, it can be predicted that the number of complaints registered in the month of March 2021 might be close to 579.45.
- On the basis of the model designed using multiple linear regression, the total number of consumer complaints registered in year 2022 is predicted to be 5349.685 rounding of gives value 5350 which is less than the total complaints in year 2020 i.e., 9450.
- As predicted from figure 4, the total complaints registered in year 2022 is less than the predicted total complaints registered in year 2021 i.e. 6240.6.
- On the basis of p-value, multiple regression model is better fitted to answer the problem statement.