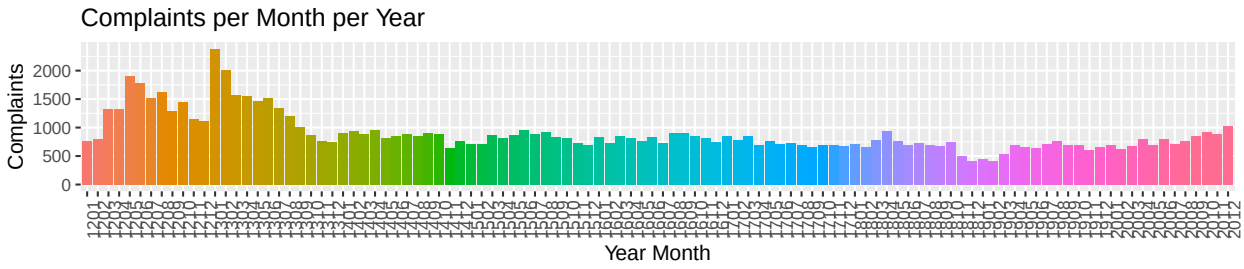
Individual Assignment 3

Data

Dataset "Complaints" is used from official website of Consumer Financial Protection Bureau,United States Government.Complaint data for various companies of United States are available on the website, from which data for Bank of America was filtered and downloaded to be used for this assignment. The dataset used consist of 99,160 observations and 18 features(For feature description refer Part 1 Appendix). Out of 18, 17 features are of character type and 1 feature "Complaint ID" is integer type in the dataset.
From the 18 features, only 2 feature Date.Received and Product were selected for predicting the number of complaints to be received by Bank Of America in year 2022.Feature description is as follows:
1. Date.Received: Date of complaint received from year 2011 to 2021
2. Product: Product category of Bank of America in which the complaint falls

Planning

Because total number of complaint received in year 2022 by Bank of America is to be predicted, feature date received is used for further calculation and analysis.I created dataset with 2 features "number of complaints" and "yearmonth" from the selected feature date received.For year 2011 only December month data is available and for year 2021 data for months January, February and March, so these values were dropped as we need to observe the year round trend in complaints for predicting the number of complaints to be received in year 2022.The below graph shows the number of complaints received by Bank of America per month from year 2012 to 2020.



Before conducting linear regression on the data we need to check the assumptions of linear regression.
1. All predictor variables must be quantitative or categorical, and outcome variable must be quantitative, continuous, and unbounded. The predictor variable "year" is categorical and outcome variable "number of complains/month per year" is quantitative.
2. Variance should be non -zero: The test shows that variance of data is non-zero(For output of variance refer Part 2 Appendix).

3. No perfect multicollinearity: There's only one predictor variable, so no need to check for multi-collinearity.
4. Predictors are uncorrelated with external variables: It is assumed there aren't any external variables. There are few assumptions that should to be satisfied by residuals, which we will check after conducting linear regression.

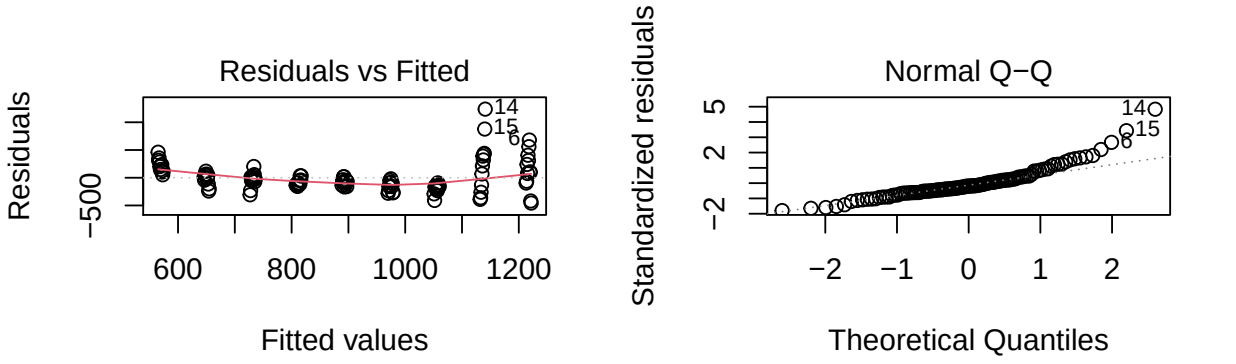
Analysis

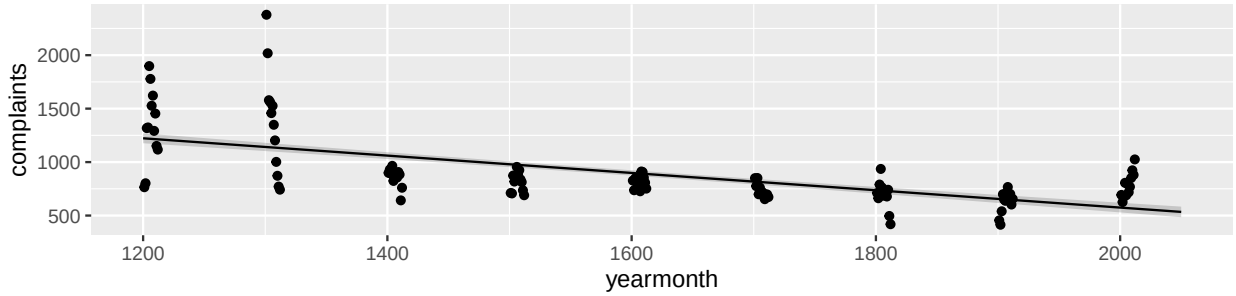
The coefficients(For Linear regression results refer Part 3 Appendix) obtained after applying linear regression corresponds to formula $\texttt{complaints}_i = \beta_0 + \beta_{\texttt{yearmonth}}\texttt{yearmonth}_i + \epsilon_i$. Where 'intercept' value 2194.65 is beta0 and 'yearmonth' value -0.81 is beta(yearmonth).

The summary (For summary of model refer Part 4 Appendix) of the model gives various parameters from

which we can interpret the performance of the model.
1. The R_ squared value(0.3997),which is 39.97%, indicates month and year account for 39.97% of variation in complaints received.In other words, if we are trying to explain why some product has more complaints than others, we can look at the complaints obtained by Bank of America for each product.
2. F-statistic(70.57), which is significant with p value <0.001. Which means there is less than 0.1% chance that model would have performed better if mean value would have been used to predict the outcome. Therefore, we can conclude that our regression model results in significantly better prediction.

Now, let's check the assumptions of the model to be satisfied based on the residuals.Below are assumptions to residuals to be satisfied: **1. The Durbin Watson test**(For results of Durbin watson test refer Part 5 Appendix) gives the value of d = 0.551 with p=0 showing that the result is significant, which means that first order correlation exist.In our case the autocorrealtion is due to time series variable month and year, indicating that system depends in prior data.It adds degrees of uncertainty, which means that we can predict the future values well but 100% accuracy can't be achieve.Since, I need to apply simple linear regression in the model this is the most I can do in this assignment.
2. Residuals are homoscedastic (constant variance): From the residual vs fitted graph below it can be concluded that the residuals are follows homoscedasticity and linearity.
3. Normal: The QQ plot indicates that residuals follow normality.





Conclusion
The above graph shows that there is decrease in the number of complaints received by Bank of America over the years.
From all the above assumptions it can be concluded that we have a model which is significantly useful.But there is one exception of auto correlation between variables due to use of time series, which can be further analyzed using time series analysis.Thus, it can be concluded with certain degree of confidence that the total number of complaints to be received in year 2022 by Bank Of America is 4890 approximately(For result of prediction refer Part 6 Appendix).