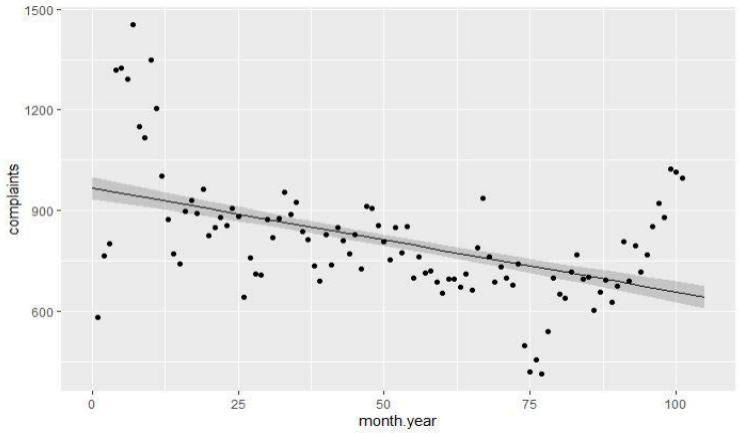
Individual Assignment 3

Ali Rafieepouralavialavijeh, 20900871

Problem statement and data used:

The Consumer Complaint Database is a collection of complaints about consumer financial products and services that are sent to companies for response. This database contains more than 2 million rows of data collected from 2011 to 2021. Each row corresponding to each single complaint, includes data about the request, type of the issue, date received and responded to, ID, etc. However, for this analysis the focus of the study would be on the complaints associated with the Bank of America which as a result consist 99,159 cases from 2011 to 2021. Since the research question of this report is "***how many potential complaints would we get from consumers in 2022?***" we would first remove columns of data that are not useful for this study and then group data based on the number of occurrence of complaints. As such "*monthly*" and "*annually*" granularity was checked, and it is found that "*monthly*" granularity has a more R2 value which is more descriptive of the variance of the original model. In addition to this, F-statistic of the "*annually*" model shows that in 5% significance level we have enough evidence to say our new model does no better than simply using the means. Thus, we proceed with "*monthly*" model [please find the details of the output of this step in appendix 1]. The Figure 1 below shows the distribution of the resulting data:




```
Call:
lm(formula = complaints ~ month.year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-649.20 -140.77  -56.62   91.28 1227.13

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1238.4604    51.6046  23.999  < 2e-16 ***
month.year    -6.2567     0.7927  -7.892 2.38e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 271.2 on 110 degrees of freedom
Multiple R-squared:  0.3615,    Adjusted R-squared:  0.3557
F-statistic: 62.29 on 1 and 110 DF,  p-value: 2.381e-12
```

Planning:

Based on the specifics of this data set, which is a collected data through the time we should now investigate the data to see if we need to make some changes. Given the fact that the last data point is associated with March 2021 and that is collected in the middle of the month it seems to be too out of scale in comparison to other data points. As such, we would remove that to make a more consistent data set. Since for the purpose of this study "*Regression*" is to be used, we should analyze certain assumptions to produce a more reliable model. First, we should check all predictor variables and they must be quantitative or categorical, and outcome must be quantitative, continuous, and unbounded. This assumption holds as our predictor variable is quantitative and our outcome is also quantitative, continuous, and can be considered unbounded. Second, the variance should be non-zero. This assumption also holds since the data points vary clearly through the time (variance = 107596.8). Third, no perfect multicollinearity (predictor variables should not correlate highly). This assumption holds since we only have one predictor (the number of the month in the sequence of all months). Forth, predictors should be uncorrelated with external variables. While there could be many factors involving in the number of the complaints from a company, here we can assume that we have no other external variable affecting the number of complaints. Fifth, the residuals should be normal, homoscedastic, and independent. In the Analysis part the model is built and these assumptions are tested in details.

Analysis:

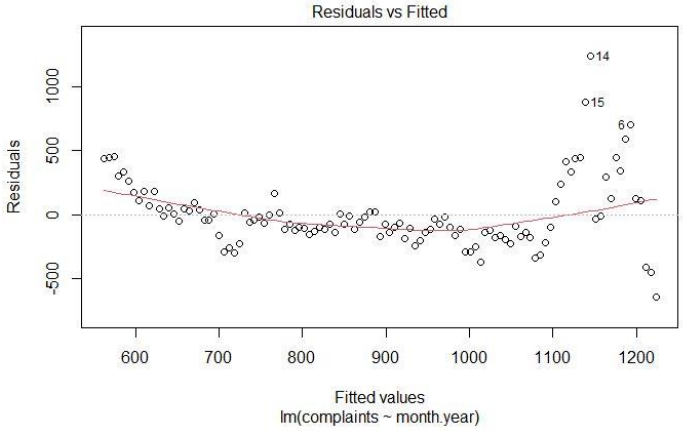
For this purpose, we would first build a regression model. Figure 2, above, is the output of the model that is built based on our data. All predictor variables have an influence on complaints at the 5% level of significance: (*month.year t=-7.892 p-value=2.38e-12*). The intercept is significantly different from 0 (*t = 23.999, p=2e-16*). R^2 is 0.3615, and adjusted R^2 is 0.3557. Coefficients and 95% confidence intervals are listed below:

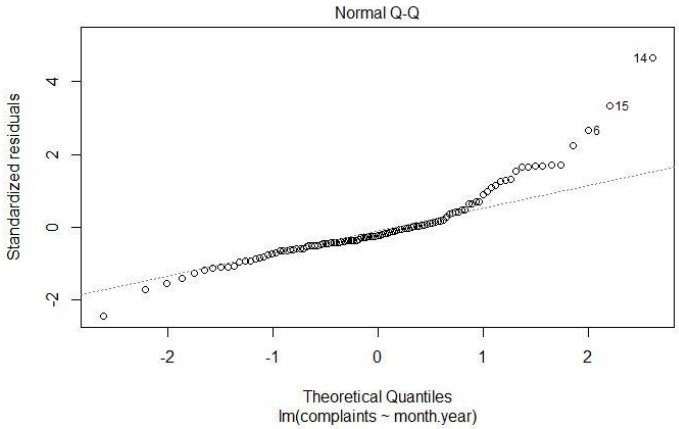
```
                2.5 %      97.5 %
(Intercept) 911.083715 1047.013742
month.year   -4.589998   -2.298626
```

The F-statistic=62.29 and its corresponding p-value=2.381e-12 gives us enough evidence to make sure that our model outperforms the simple mean. Following graphs are as a result generated for further analysis of these assumptions.

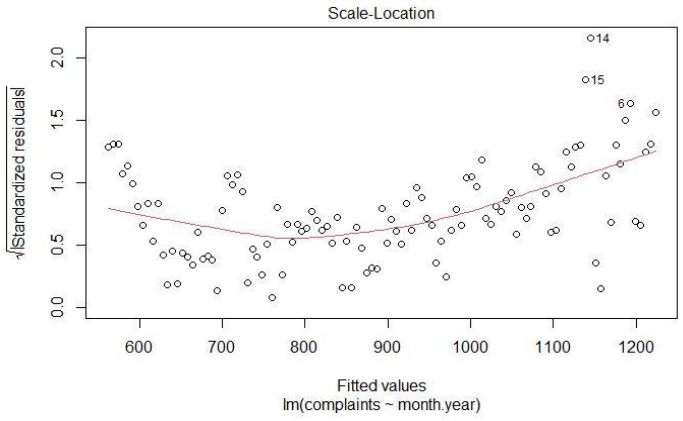
```
 lag Autocorrelation D-W Statistic p-value
   1       0.7274346     0.4686392       0
Alternative hypothesis: rho != 0
```

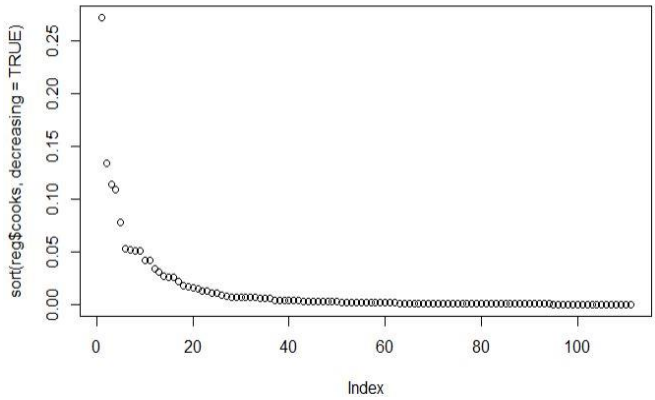
Output above is the result of Durbin-Watson test of independence. It appears that for 5% significant level, we have enough evidence to reject the hypothesis of independency of the residuals. Figure 3 illustrates the distribution of residuals and the fitted values. Even though it does not perfectly show a linear relationship but still this result is not too bad, and we could possibly continue checking other assumptions. Figure 4 shows the QQ plot of our residuals. Based on this figure and Shapiro-Wilk test of normality we could confidently conclude that our assumption of normality is violated.





If we were to polish our model so as to make it generalize better we could check for the outliers and influential points and possibly remove them.





Here (in figure 5), we found 5 residuals are above or below 1.96 standard deviations. As this represents 4% of the observations, if the residuals were normal (5% of data is expected to be outside of 1.96 standard deviations), we do not consider any of these observations as outliers and continued with all 115 observations included in the model. In addition, to investigate influential cases, Cook's Distance on the developed model is calculated. Cook's distance (Figure 6) was a maximum of 0.25, far below the chosen cutoff value of 1. We thus conclude that there are no influential cases.

Conclusion:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
589.8415	586.397	582.952	579.508	576.064	572.619	569.175	565.731	562.287	558.842	555.398	551.954

Table above shows the number of the complaints in each month of 2022. Overall, given the downward trend of the number of complaints through the time. It is anticipated that the number of complaints keep going down for 2022. However, this result might not be reliable enough because our model has violated some of the assumptions of regression analysis and as such would not generalize very well to the data that are not in our sample.

Appendix 1:

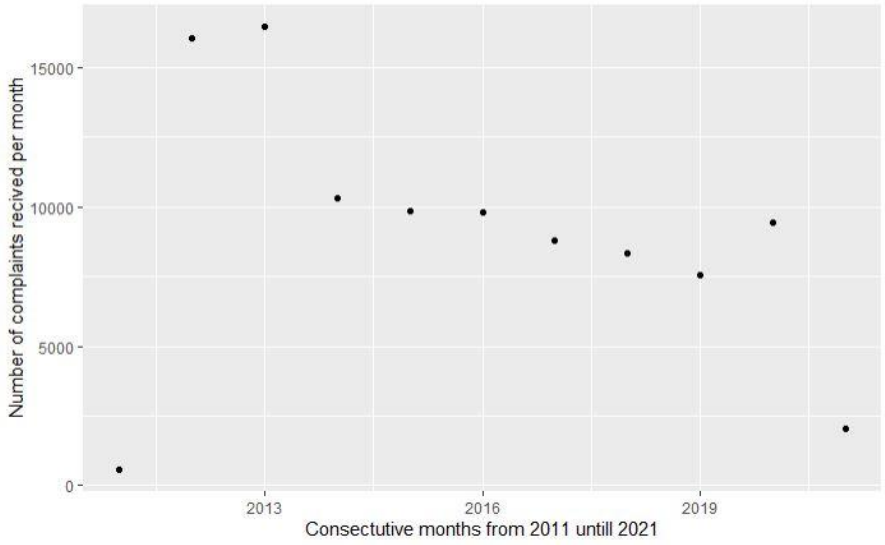


Figure above shows the distribution of complaints annually.

```
Call:
lm(formula = X0 ~ Week.Year, data = data_annual)

Residuals:
     Min       1Q   Median       3Q      Max
-10739.0     68.8    351.9   1531.2   6061.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 941025.6   921489.7   1.021    0.334
Week.Year     -462.3      457.1  -1.011    0.338

Residual standard error: 4794 on 9 degrees of freedom
Multiple R-squared:  0.1021,    Adjusted R-squared:  0.002292
F-statistic: 1.023 on 1 and 9 DF,  p-value: 0.3382
```

Results above show that using this dataset and fitting a regression model on it loses its significance because the mean does better (p-value=0.3382)