Econ 144 Project 1

Ju Won Chung

I. **Introduction**

Background

The dataset I am using is Retail Sales: Electronic Shopping and Mail-order Houses from *fred.st.louisfed.org*. This subset industry within U.S retail is defined by “establishments primarily engaged in retailing all types of merchandise using nonstore means” which includes catalogs, electronic media etc. The data is observed monthly from January 1, 1992 to November 1, 2019 for a total of 335 observations. Retail sales in e-shopping and mail-order houses is measured in millions of dollars.

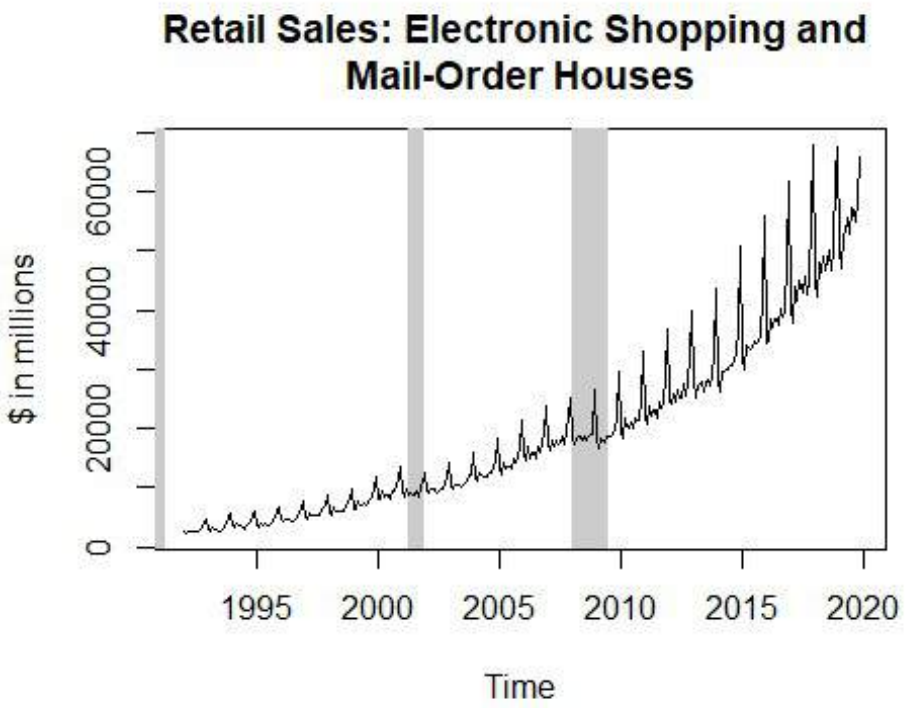
Goal

I want to take into account factors such as seasonal consumer behavior and overall changes in trend due to technological progress in order to better understand the success and inner workings of these 2 aspects of convenient retail evolutions and forecast future values to gain insight in potential prospects of this ever growing industry.

II. **Results**

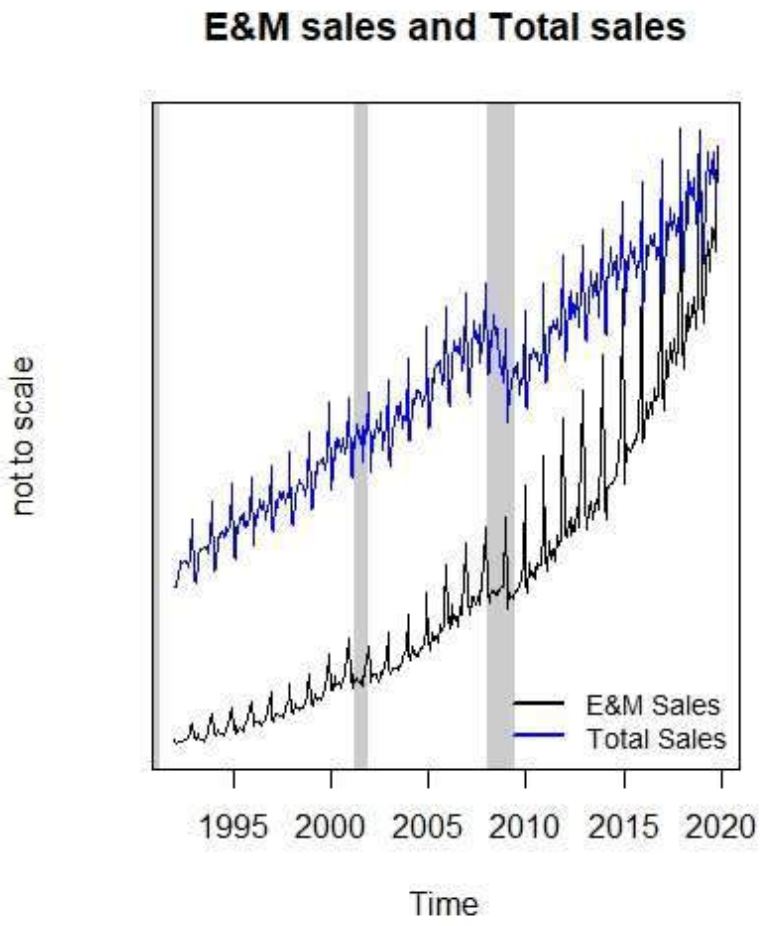
1. Modeling and Forecasting Trend

a) Time Series Plot



b)

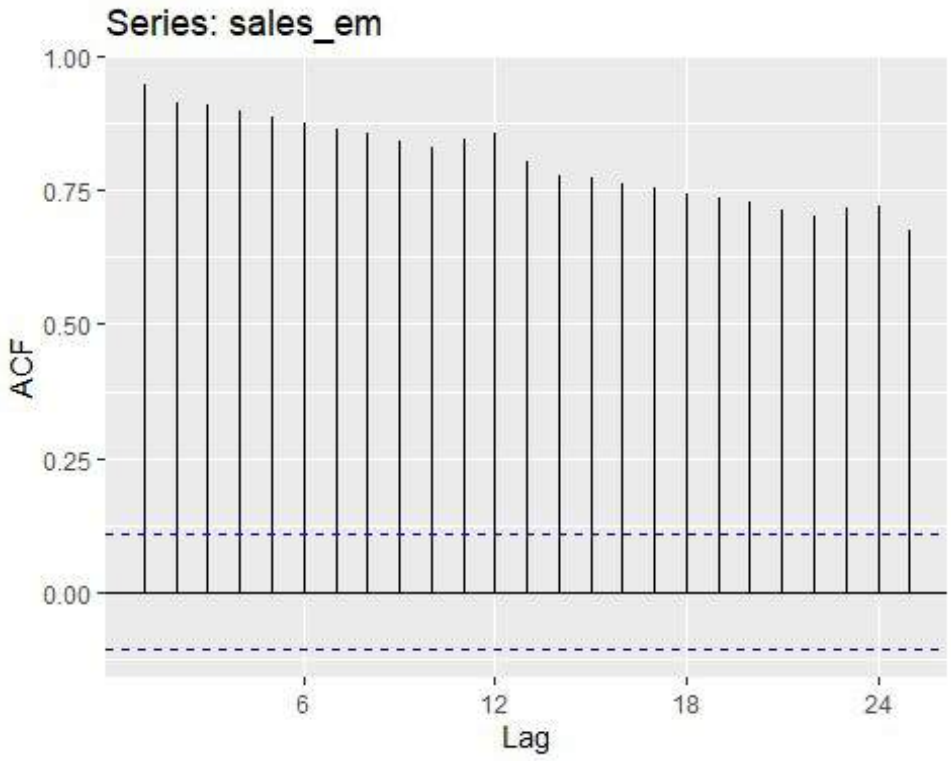
The plot suggests that this time series is nonstationary rather than covariance stationary. This data has a clear upward trend with no stable expected mean throughout time while also showing increasing variance.

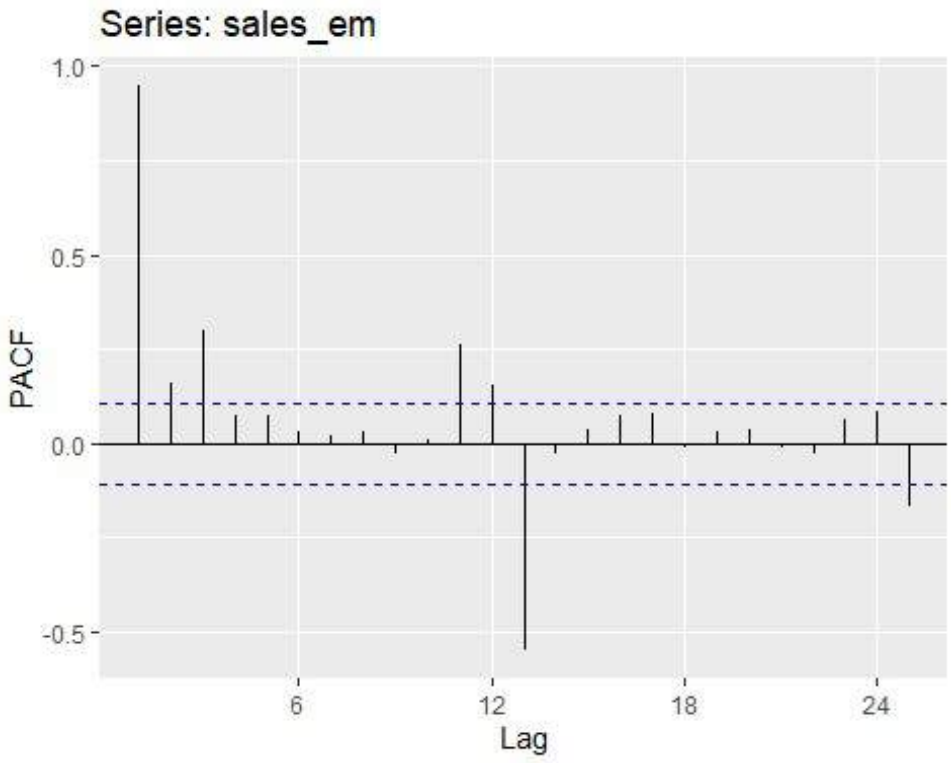


Following the characteristics of Total Retail sales, E-Shopping and Mail-Orders sales also gradually increase when the U.S economy is strong and falls off during recessions. However, it is not impacted as much and the declines in sales are not as drastic, as can be visualized. Another key difference is that sales during Christmas season have volatility with much higher peaks as time passes as opposed to the fairly stable variance in Total Sales. This indicates that the market for online retail is constantly increasing and may have even more growth opportunities as consumers move toward convenient shopping.

- **This change in variance over time may complicate the forecasting process and will likely require smoothing.**

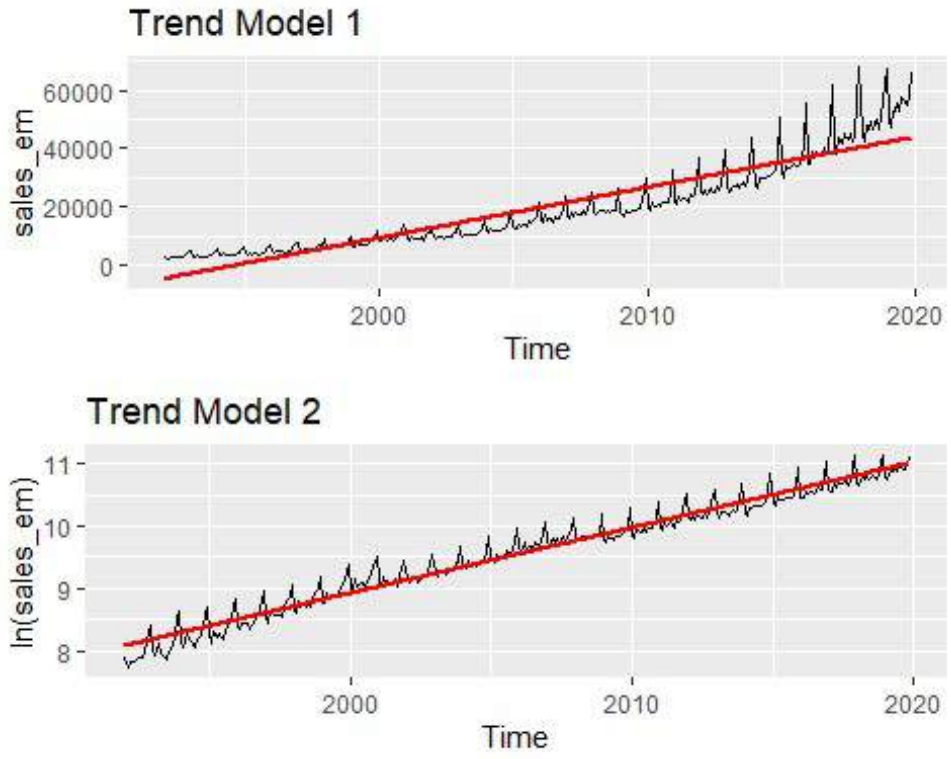
c) Autocorrelations of “sales_em” (Sales from E-Shopping & Mail-order)





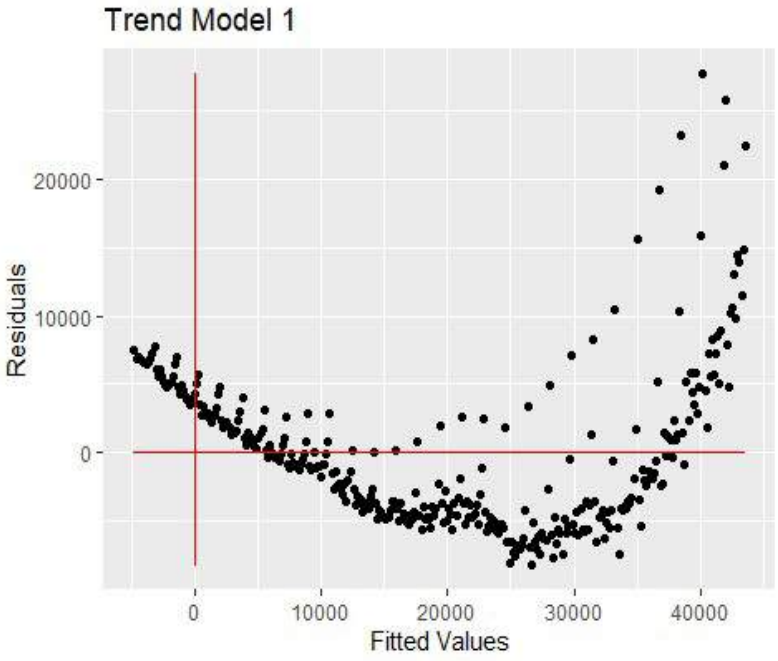
The ACF shows slowly decreasing lags, indicating that autocorrelation lasts for a long period of time (persistence). The PACF reveals that there are a multitude of significant lags beyond the initial 3. The second longest spike at Lag 13 seems to be due to the increasing variance over time (~ every year) possibly explained by consumer behavior leaning towards online shopping or simply an increasing U.S population. Together, these measurements strongly suggest an S-AR (seasonal autoregressive) model as the plot also reflects.

d) Linear Model and Natural Log Model Fits

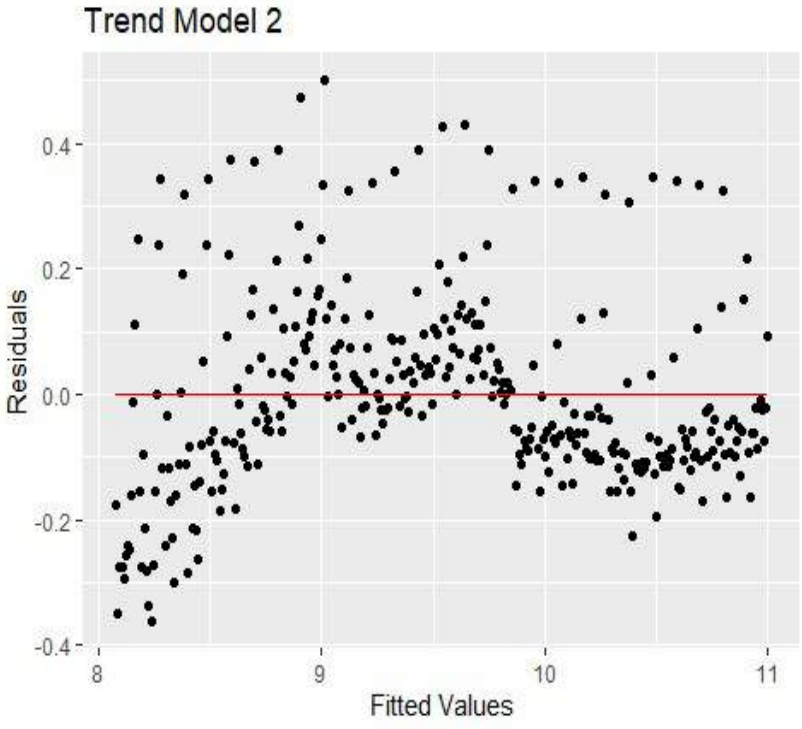


- Here, the original data is logged to stabilize the variance as discussed before

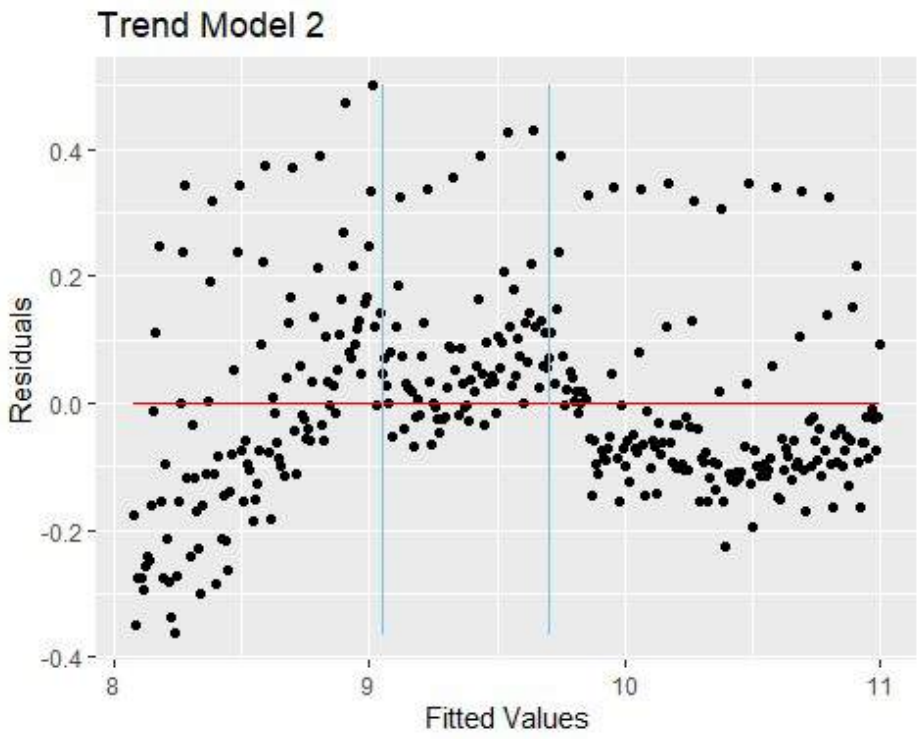
e) Residuals vs. Fitted Values

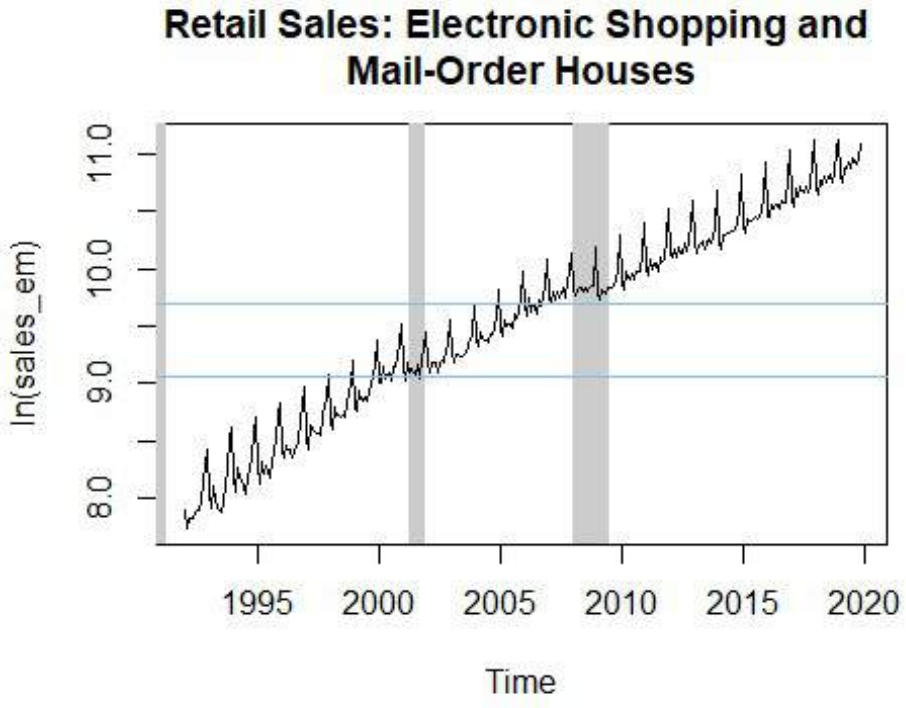


In the linear plot (model 1), the bulk of the residuals seem to reflect overestimated values as shown by the points below line y = 0. However, the tail ends are underestimated (above y = 0), though in the right tail the increasing in values increase faster in residuals. This is likely due to the linear model's inability in capturing the increasing volatility over time.



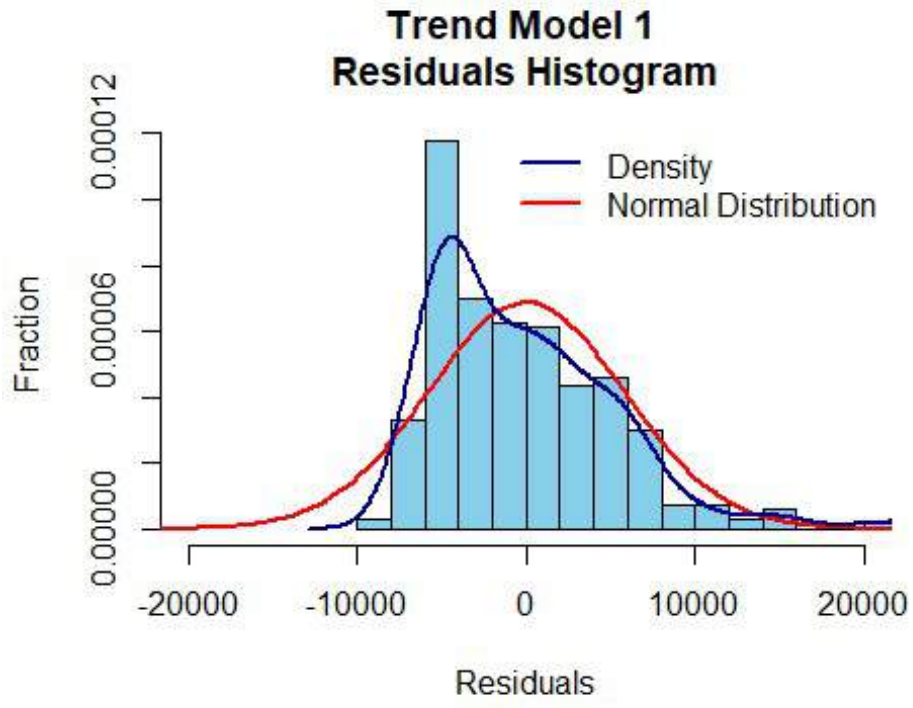
In the natural log plot (model 2), the residuals seem more randomly scattered with a few exceptions. The middle area (x = ~ 9, 9.7) has both fewer and smaller magnitude values of overestimation than both the horizontal left and right tails. Looking at the ln plot of sales_em, the gray recession bands seemingly reflect this lack of overestimation and abundance of underestimation of sales.



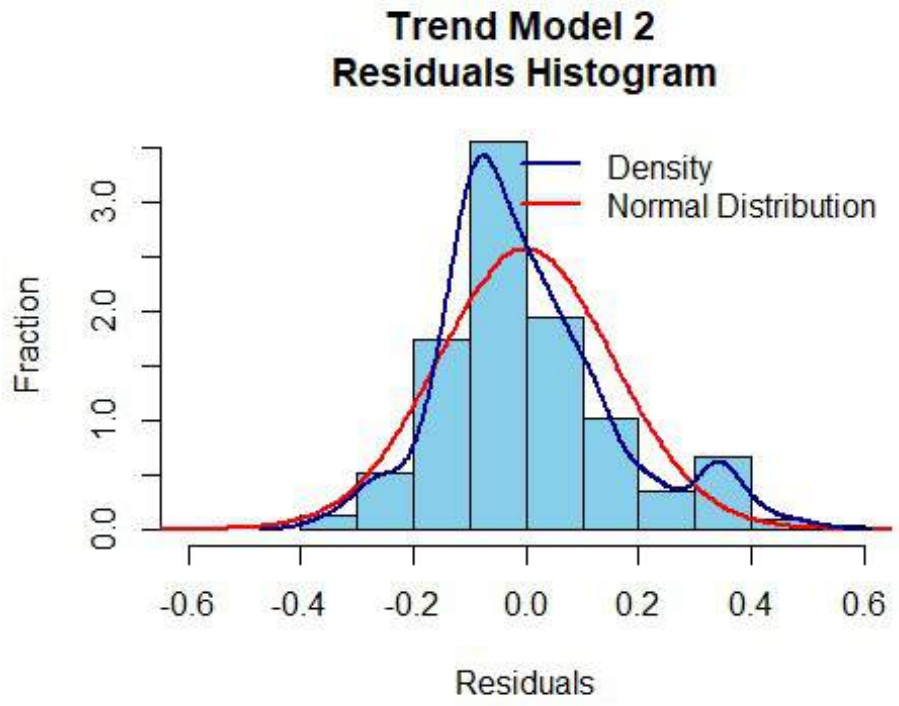


The light blue lines indicate the decrease in sales during recessions (gray bands) which correspond to the high underestimation of many values (in Trend Model 2) along with nearly no overestimation during this time period.

f) Distribution of Residuals



The residuals of model 1 looks somewhat like a normal distribution but is not symmetric and displays a right skewness. The bulk of the model 1 residuals due to some overestimation are around -50,000 as shown by both the histogram and previous residual vs fit plot.



The residuals of model 2 seems to have a fair resemblance to a normal distribution. The distribution is slightly taller and thinner, showing that the density of residuals around 0 is proportionately even greater than the normal distribution. Overall, model 2 seems a closer to normal than model 1.

g) Summary Statistics

```
##
## Call:
## tslm(formula = sales_em ~ trend)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8197  -4406  -1027   2993  27713
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4915.943    636.467  -7.724 1.34e-13 ***
## trend         144.643      3.283  44.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5812 on 333 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8531
## F-statistic:  1941 on 1 and 333 DF,  p-value: < 2.2e-16
```

Model 1 has a strong R^2 of 0.8535 which shows that it explains the data very well. The t-stats (absolute value) of both the intercept and t are significant, with corresponding low p-values that

pass significance tests well past 1%. The high f-stat and its low p-value reflect the model as significant as a whole.

```
##
## Call:
## tslm(formula = lsales_em ~ trend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36225 -0.09789 -0.02546  0.07562  0.50134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.067e+00  1.704e-02  473.36   <2e-16 ***
## trend       8.762e-03  8.792e-05   99.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1556 on 333 degrees of freedom
## Multiple R-squared:  0.9676, Adjusted R-squared:  0.9675
## F-statistic:  9933 on 1 and 333 DF,  p-value: < 2.2e-16
```

Model 2 has an even stronger R^2 at .9676 which is extremely closed fit. The t-stats and p-values of the intecept and t are also significant at very low levels. The same can be said for the whole model, which is supported by a much higher f-stat and low p-value.

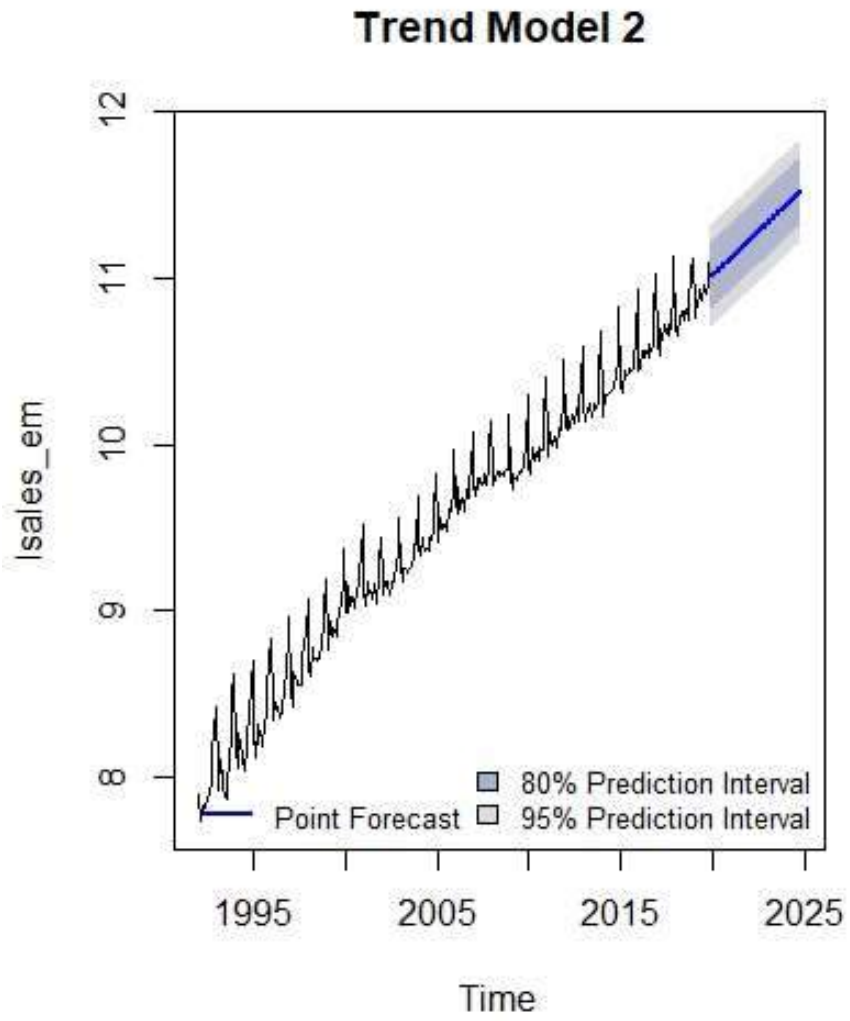
h) Model Selection Criteria

```
##        df       AIC
## t_mod1  3 6761.9820
## t_mod2  3 -291.7754

##        df      BIC
## t_mod1  3 6773.424
## t_mod2  3 -280.333
```

Model 2 has a lower AIC and BIC than Model 1 in both actual and absolute value, revealing a clear agreement that Model 2 is a better fit according to these model selection criteria.

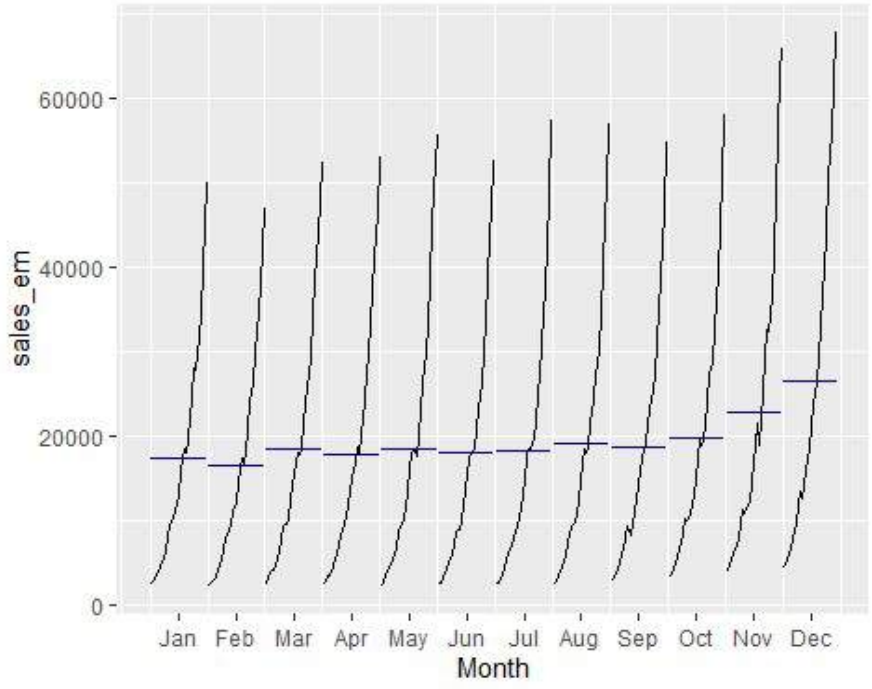
i) 5 Year Ahead (h = 60) Trend Forecast

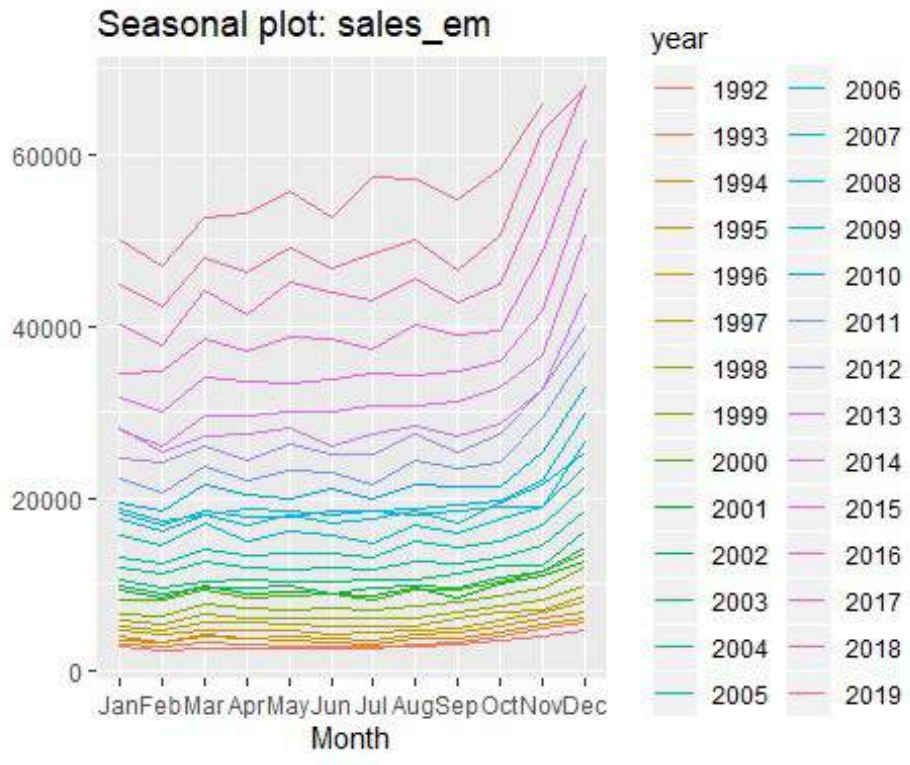
- lsales_em = ln(sales_em)



2. Implementing a Seasonality model

a) Seasonal Patterns

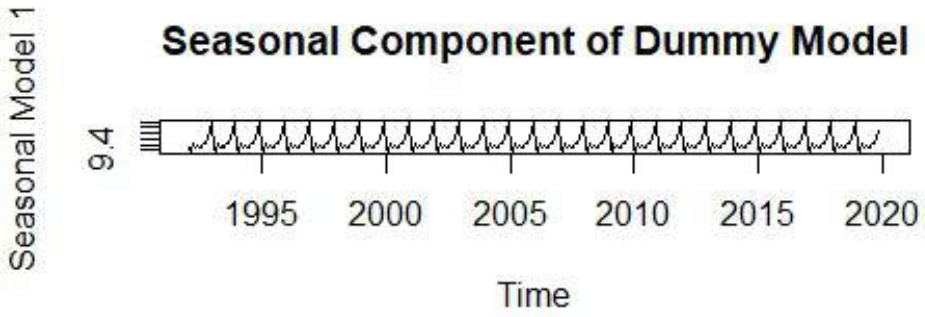


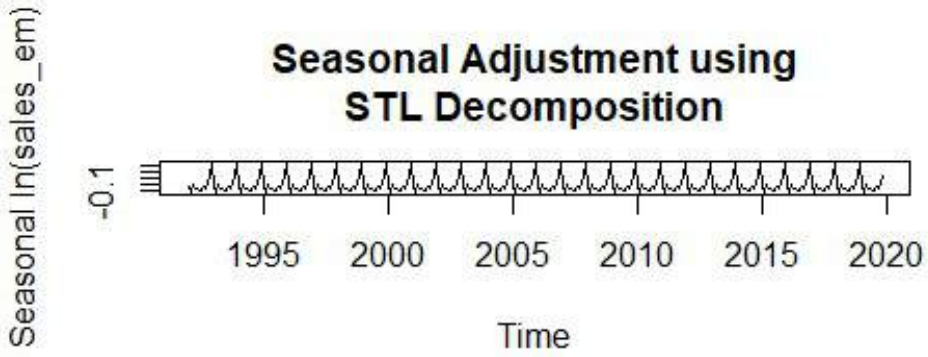


- The seasonal trend of increase in sales near month-end in the original dataset is confirmed here.

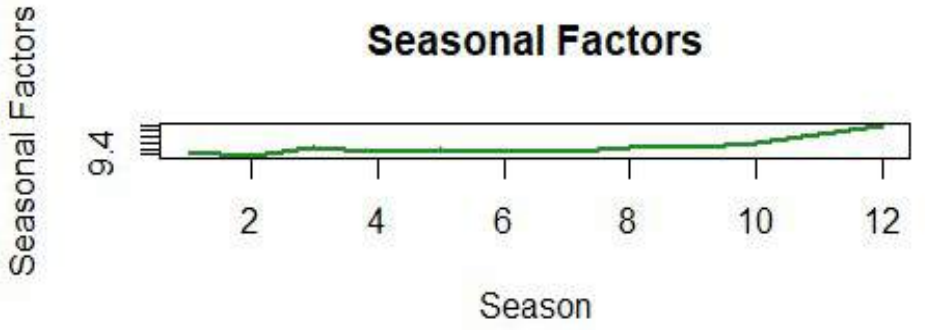
Seasonal Dummy Model based on lsales_em

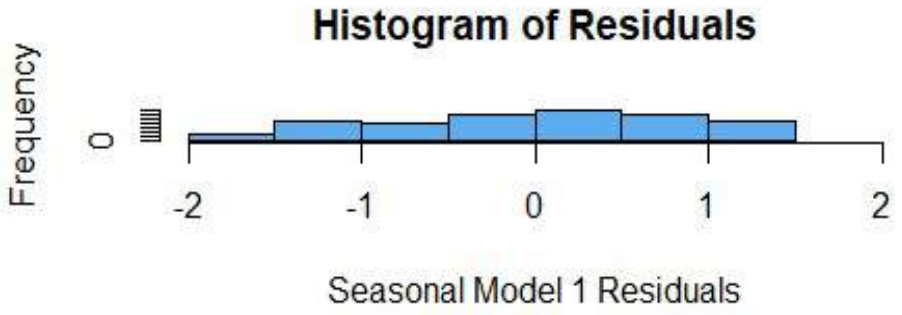
```
##
## Call:
## tslm(formula = lsales_em ~ season + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69358 -0.68070  0.09469  0.69678  1.49889
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## season1    9.4559     0.1637   57.75   <2e-16 ***
## season2    9.3803     0.1637   57.29   <2e-16 ***
## season3    9.5112     0.1637   58.09   <2e-16 ***
## season4    9.4647     0.1637   57.80   <2e-16 ***
## season5    9.4811     0.1637   57.90   <2e-16 ***
## season6    9.4626     0.1637   57.79   <2e-16 ***
## season7    9.4586     0.1637   57.76   <2e-16 ***
## season8    9.5151     0.1637   58.11   <2e-16 ***
## season9    9.5221     0.1637   58.15   <2e-16 ***
## season10   9.6040     0.1637   58.65   <2e-16 ***
## season11   9.7332     0.1637   59.44   <2e-16 ***
## season12   9.8932     0.1668   59.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8665 on 323 degrees of freedom
## Multiple R-squared:  0.9921, Adjusted R-squared:  0.9918
## F-statistic:  3384 on 12 and 323 DF,  p-value: < 2.2e-16
```


Seasonal Component of Dummy Model
Seasonal Model 1
9.4
1995
2000
2005
2010
2015
2020
Time


Seasonal Adjustment using
STL Decomposition
Seasonal ln(sales_em)
-0.1
1995
2000
2005
2010
2015
2020
Time

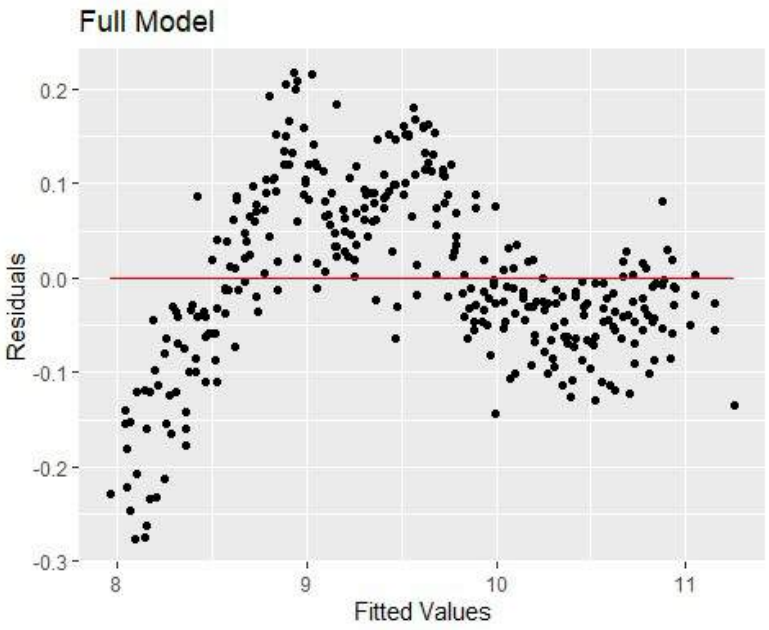
b)


Seasonal Factors
Seasonal Factors
9.4
2
4
6
8
10
12
Season

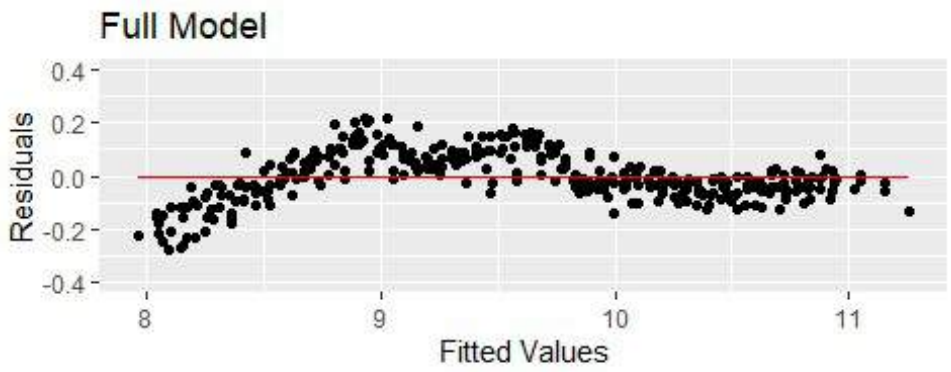

Histogram of Residuals
Frequency
0
-2
-1
0
1
2
Seasonal Model 1 Residuals

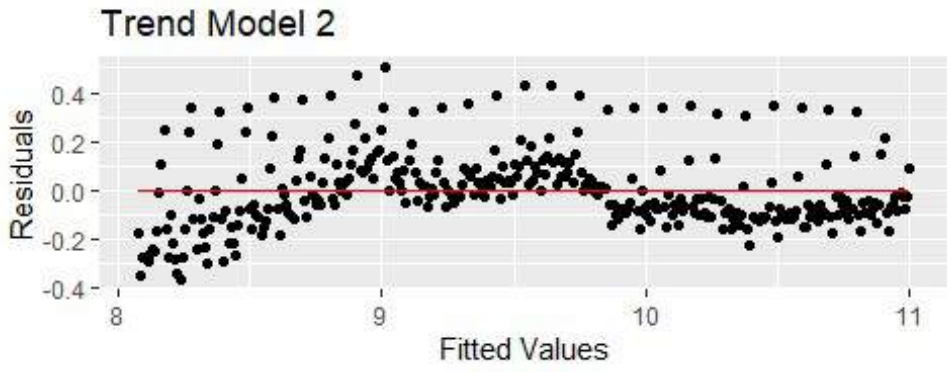
The significance testing of the all the seasons (including Season 1) results in Factors that are all significant at very low levels and increase in magnitude as Season increases (especially at 9 to 12).

c) Residuals vs. Fitted Values



Comparison to Trend Model 2 (lsales_em)





The residual vs fit plot of the Full Model is very similar in structure to the residual vs fit plot of Model 2 (ln(sales_em)). However, in this case, the variance of the residuals is much more even than previous models. It seems that the drastic underestimated outlier values have disappeared. There still remains some slight pattern which may reflect the full model lacking some significant aspect of the data.

d) Summary Statistics and Error Metrics

```
##
## Call:
## tslm(formula = lsales_em ~ trend + season)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27619 -0.05530 -0.01056  0.07222  0.21737
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.030e+00  1.996e-02 402.361  < 2e-16 ***
## trend        8.749e-03  5.365e-05 163.070  < 2e-16 ***
## season2     -8.440e-02  2.537e-02  -3.327  0.00098 ***
## season3      3.778e-02  2.537e-02   1.489  0.13737
## season4     -1.746e-02  2.537e-02  -0.688  0.49187
## season5     -9.855e-03  2.537e-02  -0.388  0.69794
## season6     -3.710e-02  2.537e-02  -1.462  0.14463
## season7     -4.978e-02  2.537e-02  -1.962  0.05063 .
## season8     -2.074e-03  2.537e-02  -0.082  0.93489
## season9     -3.796e-03  2.537e-02  -0.150  0.88116
## season10     6.935e-02  2.537e-02   2.733  0.00662 **
## season11     1.898e-01  2.537e-02   7.478  7.2e-13 ***
## season12     3.936e-01  2.560e-02  15.372  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09492 on 322 degrees of freedom
## Multiple R-squared:  0.9883, Adjusted R-squared:  0.9879
## F-statistic:  2272 on 12 and 322 DF,  p-value: < 2.2e-16
```

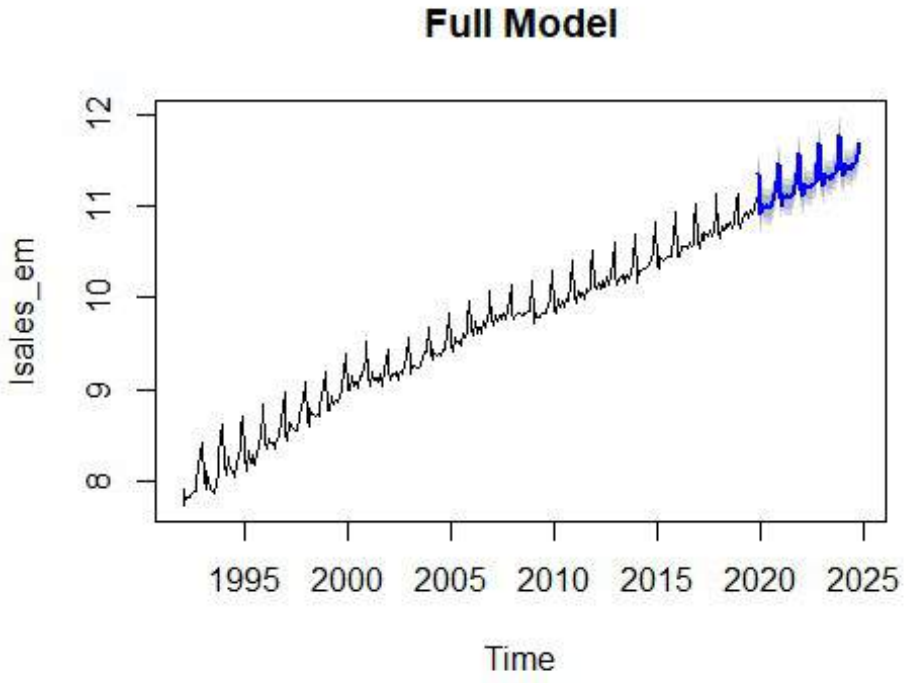
The summary statistics of the Full Model shows significant fits of the intercept, trend fit, and seasonal fit. The R^2 is at 0.9883 which reflects a very accurate model and a high f stat with a very low p-value which support the model as significant.

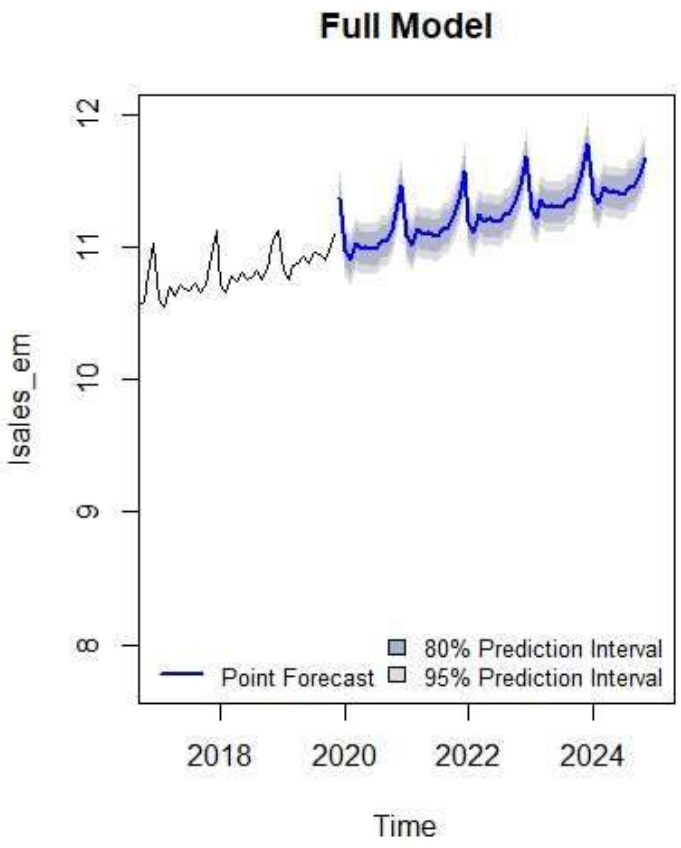
```
##          df       AIC
## t_mod1    3 6761.9820
## t_mod2    3 -291.7754
## s_mod1   13  868.4316
## full_mod 14 -612.2278

##          df       BIC
## t_mod1    3 6773.4244
## t_mod2    3 -280.3330
## s_mod1   13  918.0153
## full_mod 14 -558.8299
```

```
accuracy(full_mod)

##                     ME      RMSE       MAE         MPE      MAPE      MASE
## Training set 1.067055e-17 0.09306079 0.07380441 -0.01594407 0.8040179 0.6473993
##                    ACF1
## Training set 0.8431088

accuracy(t_mod2)

##                     ME      RMSE       MAE         MPE     MAPE     MASE
## Training set -5.36055e-18 0.1551474 0.1188303 -0.03347399 1.273471 1.042359
##                    ACF1
## Training set 0.5381467
```

AIC and BIC shows that the Full Model is a better model than both Model 1 and Model 2. The error metrics (comparing full_mod with the next best model: t_mod2) which show much lower errors in comparison confirm this.

e) 5 Year Ahead (h = 60) Full Forecast




Full Model
lsales_em
8
9
10
11
12
2018
2020
2022
2024
Time
Point Forecast
80% Prediction Interval
95% Prediction Interval

III. **Conclusions and Future Work**

Though initially the data plot seemed to present a potential problem in modelling and forecasting due to the volatility of the variance over time, logging the dataset proved to stabilize the variance and smooth the trendline.

The final model which makes use of both seasonal and trend dummies has a very close looking fit to the natural log of the sales data. Overall, the forecast seems very plausible due to the simplistic nature of the increasing values at a stable pace.

In the end, the leftover residual patterns (possibly cyclical volatility clustering) were not accounted for. The model could also potentially be improved if the seasonal component could be captured even more accurately (perhaps using different methods of decomposition e.g X11, SEATS etc). Finally, given the somewhat lacking number of observations due to the nature of the data, more time and thus more data could further improve the model in accuracy and predictive power.

IV. **References**

- *Federal Reserve Bank of St. Louis (Retail Sales: E-Shopping & Mail-Order Data)*: https://fred.stlouisfed.org/series/MRTSSM4541USN?utm_source=series_page&utm_medium=related_content&utm_term=other_formats&utm_campaign=other_format
- *U.S Census Bureau (Definitions)*: https://www.census.gov/cgi-bin/sssd/naics/naicsrch?code=454110&search=2017%20NAICS%20Search
- https://www.census.gov/retail/definitions.html
- *U.S Census Bureau (Total Retail Trade Data)*: https://www.census.gov/econ/currentdata/dbsearch?program=MRTS&startYear=1992&endYear=2019&categories=44X72&dataType=SM&geoLevel=US&notAdjusted=1&submit=GET+DATA&releaseScheduleId=