# Introduction

This project is aimed to perform decomposition and address correlation issues using ARIMA on the datasets to prepare the data for producing time series forecasts in R programming and making inferences about the forecasted results. The project engages to use “ffp2” package that includes all the functions required for forecasting and graphical analysis.

This report explains the process and subsequent findings derived from decomposition, removal of seasonality and addressing the problem of auto correlation using ARIMA on Retail and Food Service Sales data [1] (US Census Bureau, 2019). The sales dataset provides the retail and food service sales for US in “million dollars” from January 2010 to April 2019. The data is non stationary and seasonal which will be adjusted using R programming to predict the sales for next two years.

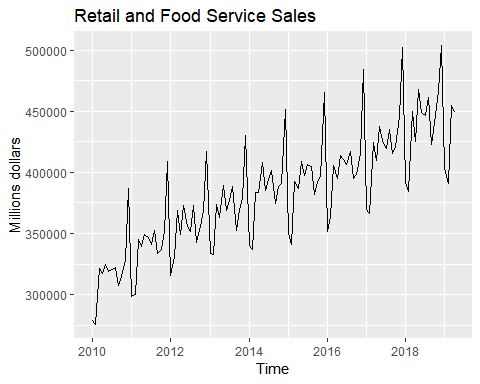
This dataset also requires to be adjusted for inflation over the last nine years. For this, purpose we have used Consumer Price Index data [2] (Bureau of Labor Statistics, 2019) to deflate and convert the sales number to dollar value in April 2019. This adjustment will remove the inflation factor and forecast the real sales based on the dollar value in April 2019.

The real sales data have been used for decomposition and ARIMA modelling. This report outlines and describes the codes used for running our analysis, their subsequent outputs and graphical representation of time series dataset with our synopsis and interpretation of the results derived in the process.

# Analysis

The retail and food services sales dataset extracted from US Census Bureau is recorded from January 2010 to April 2019. The visual representation of the sales overtime is given in figure 1.

**Figure 1: Retail and Food Service Sales from Jan 2010 to April 2019**



As it can be seen in the above graph, the sales data has an upward trend, which means the data is not stationary. Also, it has a seasonality over the period as similar upward and downward trends are observed in each time interval. Before catering to the problem of non-stationarity and seasonality, it is important to analyze and hidden or added trends in the data.

We can see that sales have increased over time at increasing rate, but one factor that is missing here, is the relative increase in sales values due to inflation. The relative values of sales would have been higher if they were adjusted to the current value of dollar. Therefore, running a time series analysis on an unadjusted data would give us the results that forecast sales with inflation. The forecast results would not be comparable with the current dollar value. Thus, we need to deflate our sales using the consumer price index data that has been extracted from Bureau of Labor Statistics. We have used the following code to adjust the historical sales with the dollar value in April 2019:

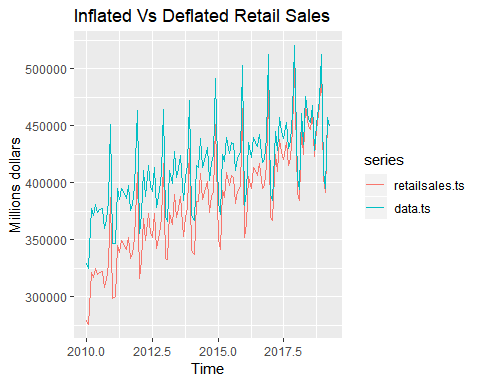
#adjusting the inflation factor  
data$index <- data$CPIAUCNS/(data[nrow(data),3])  
data$real\_sales <- data$RETAILSMNSA/data$index

As a result, we get the inflation adjusted retail and food services sales which are defined as a time series data in R console using the code given below. A comparison of Inflated and Deflated retail sales in given in Figure 2.

#defining the timeseries data  
data.ts <- ts(data$real\_sales,start = c(2010,1),end = c(2019,4),frequency = 12)

Here, we have defined the start and end dates of our dataset by assigning the frequency at 12 since our data is monthly.

**Figure 2: Comparison of Inflated and Deflated Retail and Food Service Sales**



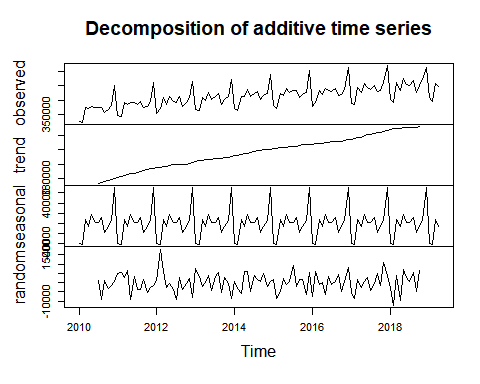
In the above graph, the red line represents “Inflated Sales”, however, the blue line represents “Deflated Sales”. As it can be seen, this CPI adjustment has increased the value of retail sales by adjusting them to the level of inflation and dollar value in April 2019, the two trends are merging at the end. Our data is now prepared to be used for further decomposition and ARIMA time series analysis.

## **Decomposition of Seasonality – Retail and Food Services**

In this section, we will be working on decomposing the adjusted sales to trend, seasonality and random residuals. We will be using the following code to segregate the three parameters and visually analyzing it as given in the figure 3:

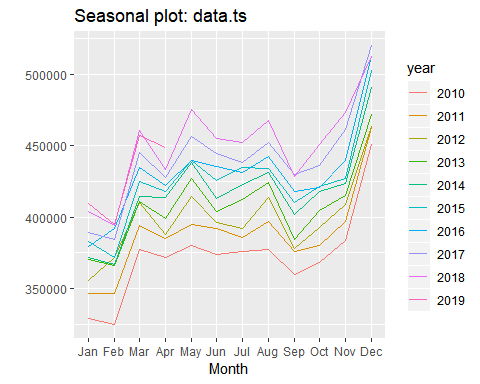
#decomposing the trend and seasonality  
data.decomp <- decompose(data.ts)  
plot(data.decomp)

**Figure 3: Decomposition of Adjusted Sales**



As seen in the above graph, there is an upward trend in our data with seasonality in every interval, hence proving the claim made above. Another visual representation that confirms the existence of seasonality in our data is given in Figure 4.

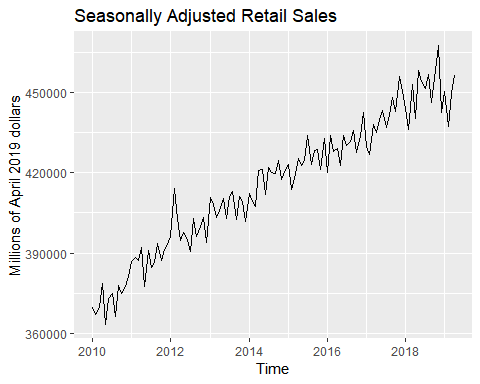
**Figure 4: Seasonality Plot – Retail and Food Service Sales**



The above graph plots all the 10 years of data as 10 separate trends which tells us that there is an increase in sales in December and a fall in January. This confirms the existence of seasonality in our data. To adjust the data for its seasonality, we can subtract the seasonality index out of our real data using the code given below. The seasonally adjusted sales after running this code are given in figure 5.

#removing seasonality from data  
data.decomp.adjus.season <- data.decomp$x - data.decomp$seasonal

**Figure 5: Seasonally Adjusted Retail and Food Service Sales**



As seen in the above graph, with the adjustment of seasonality the sales depict an upward trend without any similar ups and downs within the intervals.

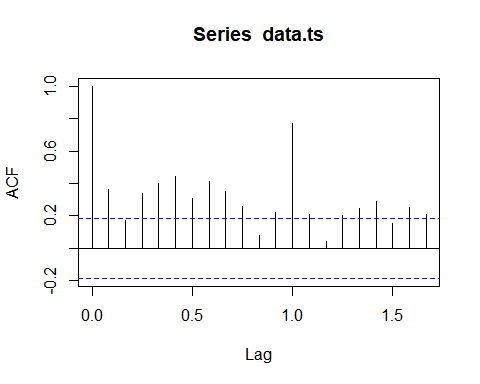
## **Addressing the problem of Autocorrelation with ARIMA**

While building the model, the residual term gives a random shock to the model and if that residual follows a trend, it means that it is not randomly generated in the model which causes the problem of autocorrelation. The function of ACF and Partial ACF in R programming determines the existence of autocorrelation in the data. The following code has been used to determine autocorrelation and the output is given in Figure 6 and 7:

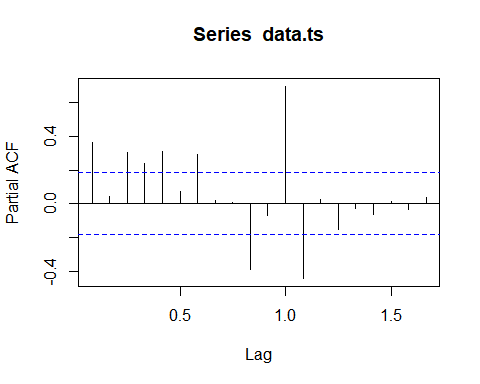
#checking auto correlation in data  
acf(data.ts)

pacf(data.ts)

**Figure 6: Autocorrelation Function**



**Figure 7: Partial Autocorrelation Function**

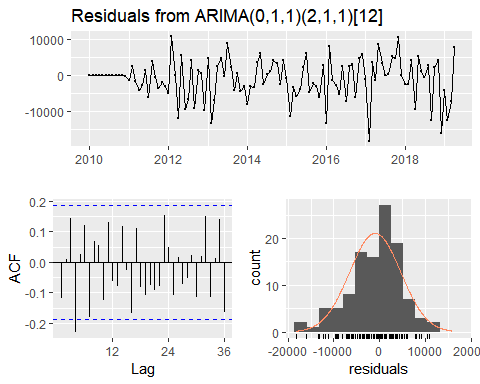


As seen in the above graph, the lags of our residual terms cross the blue line of 95% confidence level in the ACF and PACF plots, thereby, suggesting the existence of autocorrelation. The ARIMA model best addresses the issue of autocorrelation by identifying the most suitable order for the model. We have used the following code to build ARIMA model:

#building ARIMA model #d = 1 for differencing the data to make it stationary  
fit.arima <- auto.arima(data.ts,d=1,D=1,approximation = FALSE, stepwise = FALSE,trace = TRUE)

Here, the d and D are set at 1 to define the seasonality and trend in the model. The order for most suitable ARIMA model is (0,1,1) (2,1,1) [12] and residual results from this model are given in Figure 8.

**Figure 8: ARIMA Model: Residuals – Retail and Food Service Sales**



As seen in the above figure, the plot of residuals from our model do not have any particular trend which suggests the non-existence of seasonality factor. Also, most of the lags in ACF plot do not cross the blue line of 95% confidence interval, thereby suggesting the same.

## **Forecasting Retail and Food Service Sales**

Addressing the problems of seasonality, trend and autocorrelation, we can now forecast the data for next two years using the following code:

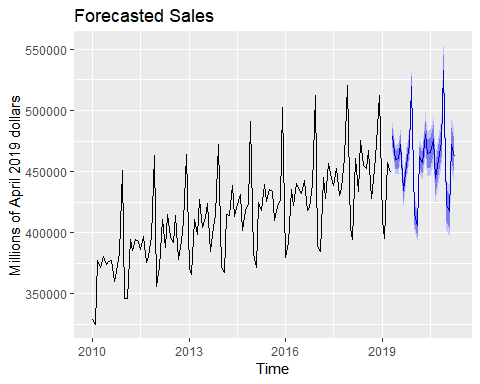
#prediction  
forecastsales <- forecast(fit.arima, h = 24)  
autoplot(forecastsales) + ggtitle("Forecasted Sales") + ylab("Millions of April 2019 dollars")

Tabular and Visual representation of the forecast is given in Table 1 and Figure 9.

**Table 1: Retail and Food Service Sales Forecast from May 2019 to April 2021**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Month-Year** | **Forecast** | **Lo 80** | **Hi 80** | **Lo 95** | **Hi 95** |
| **May-2019** | 478424.7 | 470534.3 | 486315.2 | 466357.3 | 490492.1 |
| **Jun-2019** | 459520.5 | 451528 | 467513 | 447297.1 | 471744 |
| **Jul-2019** | 460420.3 | 452327 | 468513.5 | 448042.8 | 472797.8 |
| **Aug-2019** | 472300.6 | 464107.8 | 480493.3 | 459770.9 | 484830.2 |
| **Sep-2019** | 434163.4 | 425872.4 | 442454.5 | 421483.4 | 446843.5 |
| **Oct-2019** | 455684.8 | 447296.7 | 464073 | 442856.2 | 468513.4 |
| **Nov-2019** | 472294.5 | 463810.2 | 480778.7 | 459319 | 485269.9 |
| **Dec-2019** | 519983.8 | 511404.6 | 528563 | 506863.1 | 533104.6 |
| **Jan-2020** | 415583.5 | 406910.5 | 424256.4 | 402319.3 | 428847.6 |
| **Feb-2020** | 406029.3 | 397263.4 | 414795.2 | 392623.1 | 419435.5 |
| **Mar-2020** | 462023.3 | 453165.5 | 470881.1 | 448476.5 | 475570.1 |
| **Apr-2020** | 457160.4 | 448211.6 | 466109.2 | 443474.4 | 470846.3 |
| **May-2020** | 480335.8 | 468548.1 | 492123.5 | 462308.1 | 498363.5 |
| **Jun-2020** | 464608.1 | 452597.2 | 476619.1 | 446239 | 482977.3 |
| **Jul-2020** | 465872.7 | 453642.5 | 478102.8 | 447168.3 | 484577.1 |
| **Aug-2020** | 476178.1 | 463732.6 | 488623.6 | 457144.4 | 495211.9 |
| **Sep-2020** | 444198.9 | 431541.7 | 456856.1 | 424841.4 | 463556.4 |
| **Oct-2020** | 459326.3 | 446460.9 | 472191.7 | 439650.4 | 479002.2 |
| **Nov-2020** | 474612.6 | 461542.3 | 487682.8 | 454623.3 | 494601.8 |
| **Dec-2020** | 532822.4 | 519550.4 | 546094.4 | 512524.6 | 553120.2 |
| **Jan-2021** | 422880.7 | 409409.8 | 436351.6 | 402278.7 | 443482.7 |
| **Feb-2021** | 417408.5 | 403741.8 | 431075.1 | 396507.1 | 438309.8 |
| **Mar-2021** | 472083.1 | 458223.4 | 485942.7 | 450886.6 | 493279.5 |
| **Apr-2021** | 462290.4 | 448240.4 | 476340.3 | 440802.9 | 483777.9 |

**Figure 9: Retail and Food Service Sales Forecast from Jan 2010 to April 2021**



As seen in the above graph, our forecast also follows the same trend and seasonality as our historical data while addressing the problem of autocorrelation.

# Conclusion

From this project, I have learnt that decomposition of a time series data is very important for initiating the modelling process. Understanding the trend, seasonality and variations in the random term helps in selecting the right model for the data. As in this case, knowing that data is autocorrelated, non-stationary and seasonal helped in determining that ARIMA could be one of the suitable models for this type of data. And as the result suggests, it performed well with these causalities in the data by developing a model that addressed the issue of autocorrelation while taking into account the trend and seasonality of the model. auto.arima function of R helped us in analyzing various moving average and lag order for the model while selecting the best model that has lowest AIC, BIC, RMSE, MAPE and other model accuracy parameters. The model of order (0,1,1) (2,1,1) [12] had a MAE of 4241.32 suggesting that our forecast might have an error of 4241.32 million dollars in the retail sales. MAPE for our model is 1.02% which means that our forecasts are 98.8% accurate considering the low and high value of our confidence interval.

# References

1. US Census Bureau, Retail Trade and Food Services Sales: U.S. Total , 2019, Retrieved from: <https://www.census.gov/econ/currentdata/dbsearch?program=MRTS&startYear=2010&endYear=2019&categories=44X72&dataType=SM&geoLevel=US&notAdjusted=1&submit=GET+DATA&releaseScheduleId=>
2. Bureau of Labor Statistics, CPI-All Urban Consumers (Current Series), 2019, Retrieved from: <https://data.bls.gov/pdq/SurveyOutputServlet>