**A picture containing diagram

Description automatically generated**

**Table of Contents**

[**OVERVIEW OF BELGIUM AIR PASSENGER DATASET** 3](#_Toc102254470)

[**INTRODUCTION** 3](#_Toc102254471)

[**METHODOLOGY** 6](#_Toc102254472)

[**Seasonal Naïve Method** 6](#_Toc102254473)

[**Seasonal and Trend decomposition using Loess (STL)** 7](#_Toc102254474)

[**Error, Trend, Seasonality (ETS)** 11](#_Toc102254475)

[**ARIMA Model** 13](#_Toc102254476)

[**Comparing ARIMA and Auto ARIMA Models** 14](#_Toc102254477)

[**Best Model Result** 16](#_Toc102254478)

[**REPORT OVERVIEW FOR STOCK MARKET ANALYSIS** 18](#_Toc102254479)

[**INTRODUCTION** 18](#_Toc102254480)

[**METHODOLOGY** 20](#_Toc102254481)

[**Autoregressive Integrating Moving Average (ARIMA)** 20](#_Toc102254482)

[**Generalized Autoregressive Conditional Heteroskedasticity Model (GARCH)** 26](#_Toc102254483)

[**Prophet Model** 29](#_Toc102254484)

[**Knn Regression For Time-Series Forecasting** 30](#_Toc102254485)

[**Feed-Forward Neural Networks** 31](#_Toc102254486)

[**CONCLUSION** 32](#_Toc102254487)

[**REFERENCES** 33](#_Toc102254488)

## **OVERVIEW OF BELGIUM AIR PASSENGER DATASET**

The main objective for this research was to create a forecasting model for the air passenger dataset containing a total yearly count of international intra-EU airline passengers from January 2003 to October 2021 between Belgium and EU countries. I used R Studio to generate commands that would analyze the data and build models that would be evaluated and explained in its own section.

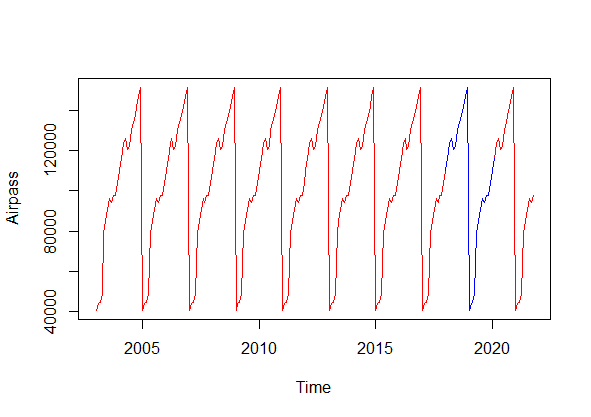
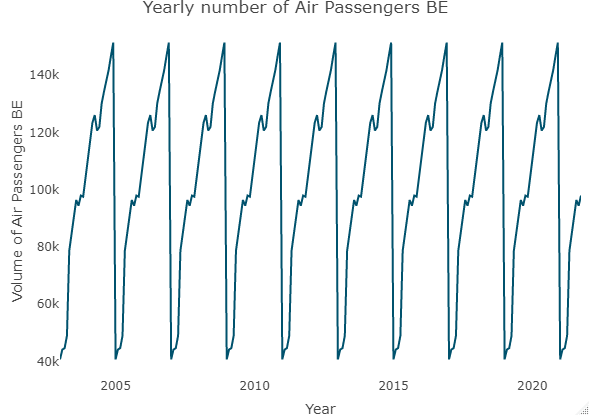
Knowing the emerging trends of passengers traveling via air transportation is critical in today's world (Adrangi et al., 2001). It is essential to accurately forecast the number of airline passengers because such predictions can be used in a variety of contexts ranging from simple planning phase to complex business decisions (Carson et al., 2011).

Air traffic forecasts are a critical feature into an airline's fleet scheduling and route network development, and the readiness of the airline's annual operational framework (Coshall, 2006). Furthermore, analyzing and forecasting air travel demand can help an airline reduce risk by providing an objective analysis of the demand side of the airline business (Cho, 2003; Cuhadar, 2014; Kulendran & Witt, 2003)

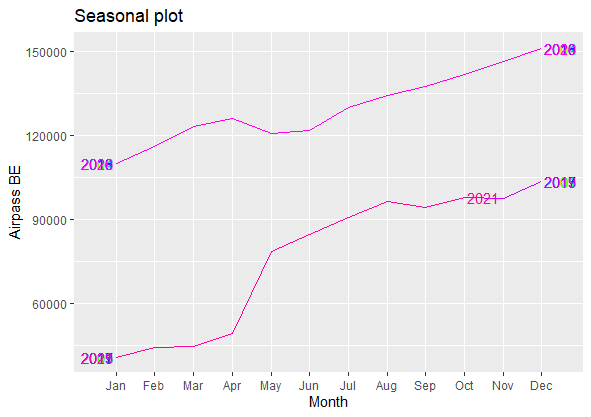
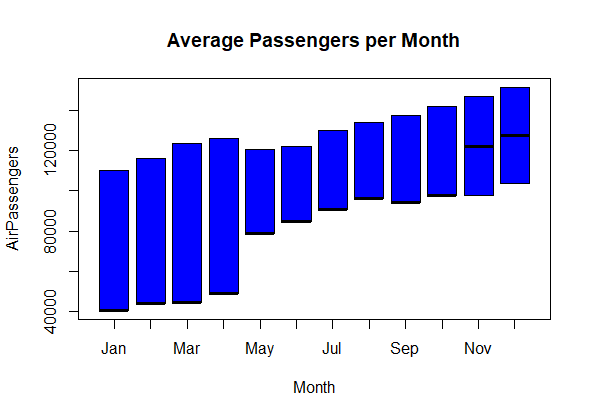
## **INTRODUCTION**

First step is to check the dataset with a time series plot, seasonal and sub-seasonal plot and PACF plot to discover the trends and seasonal effect on the dataset.

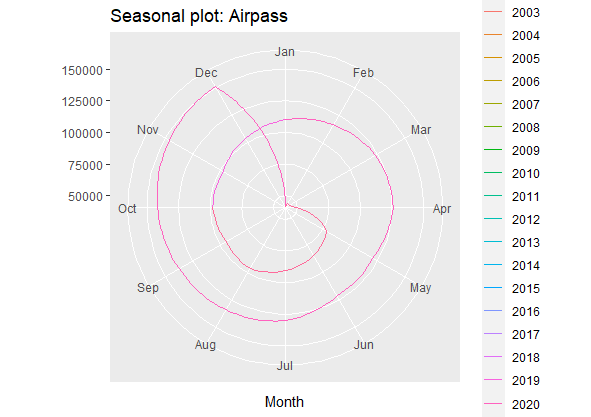
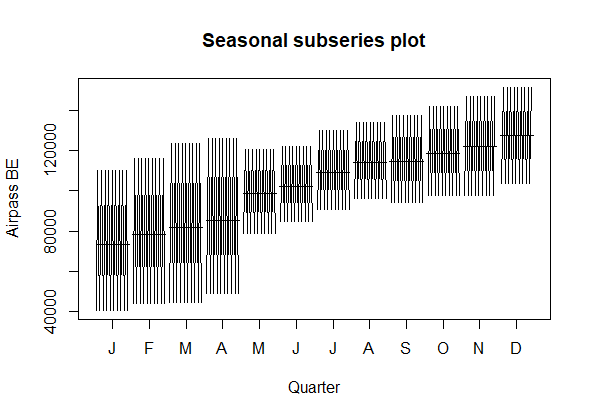
There are three things that stand out right away. For instance, the number of passengers tends to rise over time. Second, there are recurring patterns that occur each year: periodic behavior. Lastly, the variation appears to be increasing over time: the 'swings', in specific, become relatively large. The below graph shows a definite positive trend and seasonality pattern with a clear decrease and increase in the number of passengers which could either be either linear or quadratic in nature. According to the graph, the number of international airline passengers appears to peak during the summer through the holiday season, then decline near the end of each year.



To analyze further the pattern from the above graph, a seasonal plot was constructed to take a closer look into the monthly data trend to see if there is a connectivity between both. The graphs below show that the year after year increase in passenger arrivals has an increasing variance as there is an increase in the number of passengers. The variance and mean values constant from November to December compared to other months is significantly greater. Although the average value varies by month, the deviation is small. As a result, we have a strong seasonal impact with a cycle of twelve months or less.

The trend of the seasonality follows a multiplicative model than an additive model. Using both analysis below, there are always fewer passengers in November than in December, whereas the difference between July and August can be either way. In the beginning, January had fewer passengers than February, but this was reversed later. When compared to the decomposition plot, it also highlights the increase in the summer months.

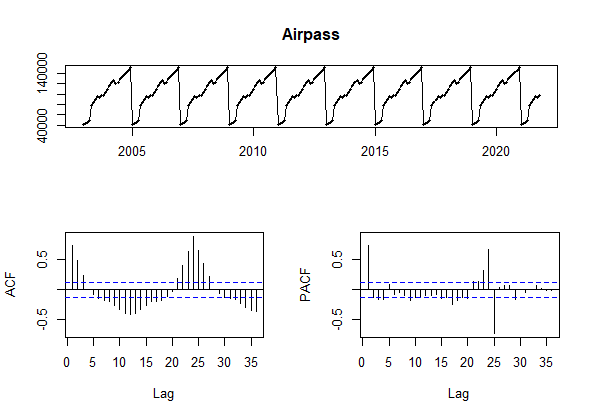


**ACF Plots**

It can also be used to detect trends, cycles, and seasonality. If there is no trend or cycle in the data, the ACF plot will experience a rapid decay, if there is no repeated pattern, there is no seasonality. As a result, the ACF plots can be used to look for infringements of weak stationarity.

Looking at the first ACF plot, the remaining portion is seasonal, indicating that the decomposition did not completely isolate the seasonal pattern. A further potential outcome is that it is an AR(k) process with k>1 and interchanging signs. The R function decompose() assumes that seasonality is not time-varying, which is most likely what is causing this problem. One simple solution is to split the remaining portion. We do so and plot it.

On the second PACF plot, as the variance rises, this series violates weak stationarity, but the ACF plot found it challenging to see violations. Thereby, ACF plots can be used to rule out variables are stationary but do not prove it. The ACF plot of the remaining portion is a useful way to consider. I will check it to show no trend/cycle or seasonality.

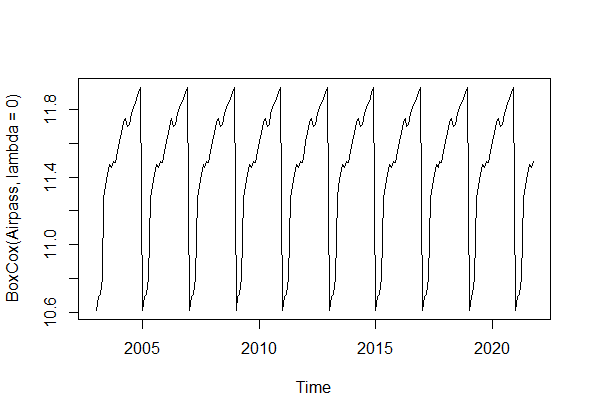


Box Cox Transformation is the conversion of non-normal predictor variable into normal patterns. Normality is an important factor for many statistical methods; if  the data is really not normal, using a Box-Cox means more tests can be applied.

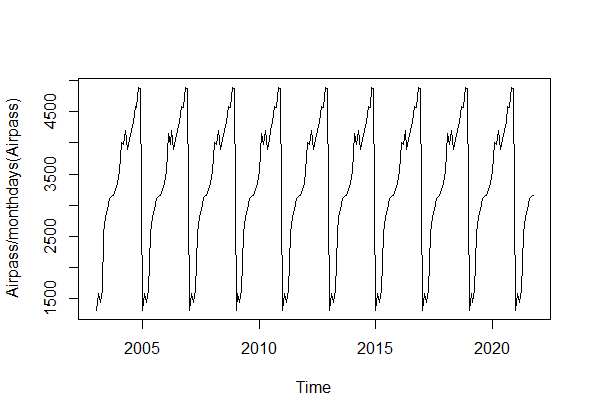
A λ value of 1 for the Box-Cox transformation is equivalent to using the original data. As a result, if the confidence interval for the optima λ includes 1, no transformation is required.



The resulting graph below a major improvement when the lambda = 0 , it gives estimation for the trend component of the original dataset. There is a cumulative number of passengers was increasing following a relatively stable trend year by year.



The below graph gives the total number of air passengers per month days, because the pattern is the same irrespective of the year, it suggests that passenger behaviour and attitude does not change over the years. There seems to be a positive trend, and the variation grows over time. The series features the expected pattern and, predictably, resembles the airline passenger procedure.



## **METHODOLOGY**

### **Seasonal Naïve Method**

Seasonal data in this case, the forecast is set to the very same value observed in the same period the previous year. The correlogram shows that the sample autocorrelation for the in-sample forecast errors at lag 9 exceeds the significance bounds. By chance, one in every twenty-four autocorrelations for the first twenty lags would be expected to exceed the 95 percent significance bounds. The test statistic is Q = 1003.5, and the p-value is 2.202.2e-16, which is significantly less than 0.05. As a result, we reject the null hypothesis of the test and conclude that the data values differ, indicating that no evidence of non-zero autocorrelations exist in the in-sample forecast errors at lags 1-24.

To forecast errors, have constant variance over time, it can be accomplished by creating a time plot of forecast errors as well as a histogram of forecast error distribution with an overlaid normal distribution curve.

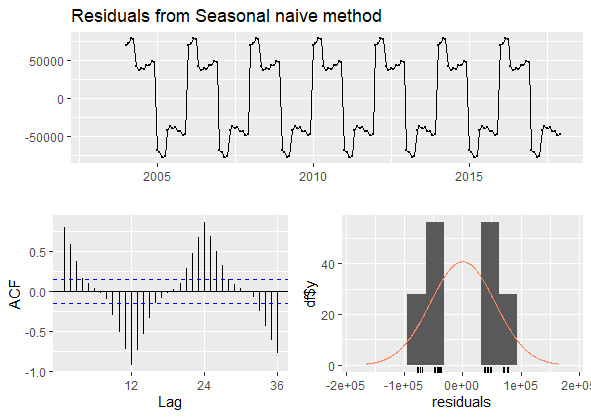
Forecast errors have a roughly constant variance over time, according to the time plot of forecast errors. The forecast error histogram indicates that the forecast errors are likely to be normally distributed with a mean of zero and a constant variance.

As a result, the Ljung-Box test reveals that there is no evidence of autocorrelations in forecast errors, while the time series shows same thing.

The plot and histogram of forecast errors indicate that the forecast errors are most likely normally distributed.

The average is zero, and the variance is constant. As a result, we can conclude that Loess decomposition is adequate.

Predictive model for skirt hem diameters, which is unlikely to be improved. Furthermore, it implies that the assumptions upon which the 80 percent and 95 percent prediction intervals were based are most likely accurate.

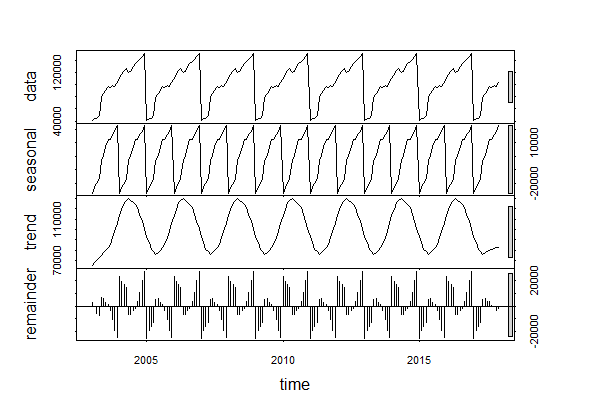


### **Seasonal and Trend decomposition using Loess (STL)**

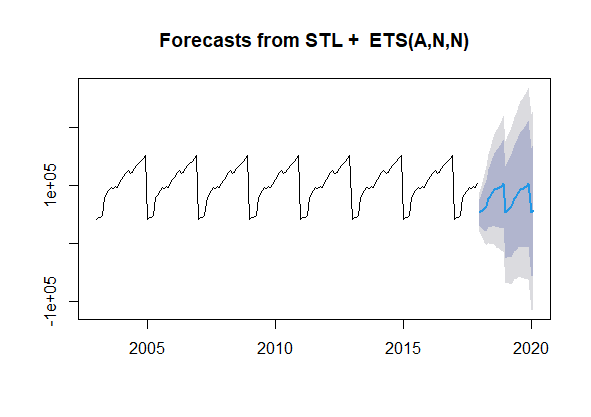
STL is a flexible and reliable time series decomposition method. The acronym STL stands for "Seasonal and Trend decomposition using Loess," and Loess is a technique for assessing interdependences.

As  seen in the result, in the upper panel, we could recognize the bar to be a large unit of variation. The seasonal panel's bar is only significantly larger than the data panel's, indicating that the seasonal signal is large in comparison to the variation in the data. If the seasonal panel is shrunk to the same size as the box in the data panel, the range of variation on the shrunk seasonal panel would not be similar to but slightly larger than the range of variation on the data panel. All the bars have relative sizes on the plot below showing a significant variation in the seasonal scale.

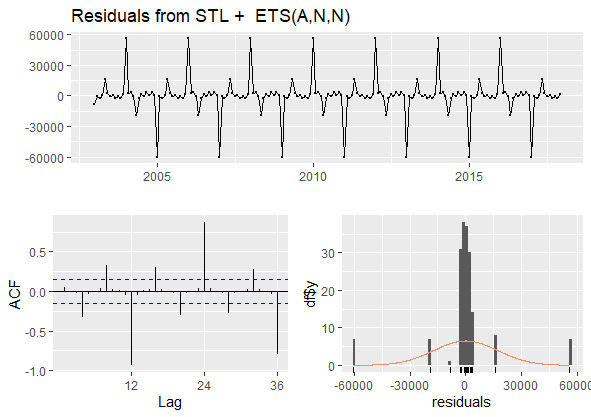
Considering the trend panel; the grey box is now much larger than the data panel, indicating that the variation attributed to the trend is much smaller than the seasonal component, and thus only a small part of the variation in the data series. The trend-related variation is smaller than the stochastic component (the remainders). As a result, we can conclude that these data does show a slight trend.



The forecasting intervals illustrated in this plot are built in the same manner as the point anticipates. That is, the upper and lower limits of the prediction intervals on seasonally adjusted data are "reseasonalised" by including the seasonal component forecasts. The seasonal component's forecast volatility was completely disregarded in this estimation. The explanation behind this decision is that the seasonal component's unpredictability is much lower than that of the seasonally adjusted data, so it is a fair estimate to reject it.

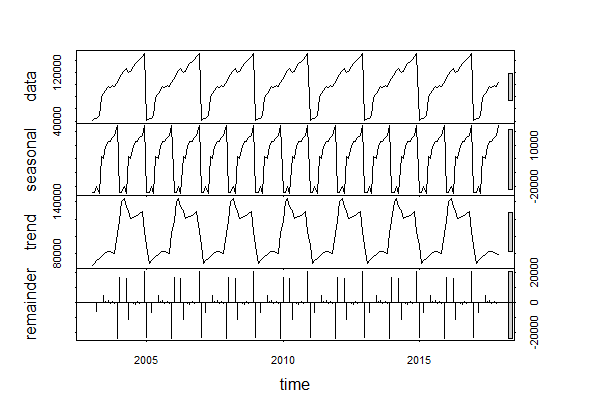


The residual plots emerge to be noise centered around 0 with no pattern. The model is a good fit.  There is a normal distribution around the residual plot and lag = 3 has crossed the significant point in the ACF plot.

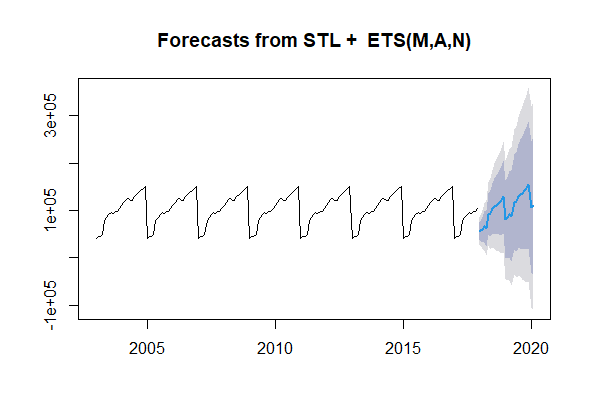


The result below shows that the data panel has a smaller unit of variation. The seasonal panel's bar is only significantly larger than the data and residual panels, indicating that both signals are large in comparison to the variation in the data.

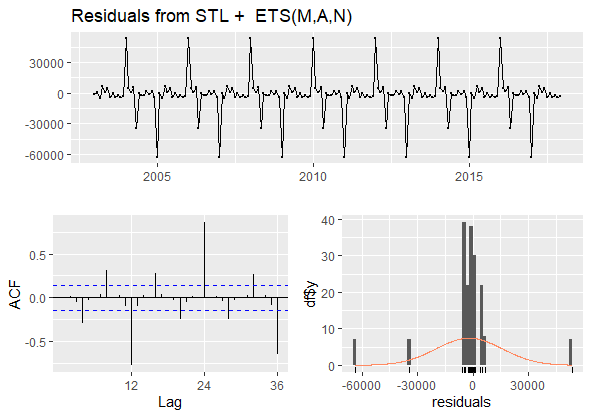
In the trend panel, the grey box is now much larger than the data panel, indicating that the variation attributed to the trend is much smaller than the seasonal component. The trend-related variation is smaller than the stochastic remainders.



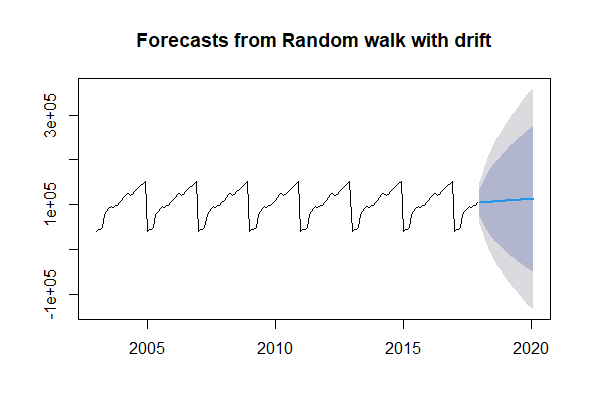
The observed seasonality above is not a typical calendrical pattern, but it is clearly associated with the business cycle, and more specifically, with market. The resulting plot below shows the predictability result of the original data set and the forecasted value shown in the blue line.



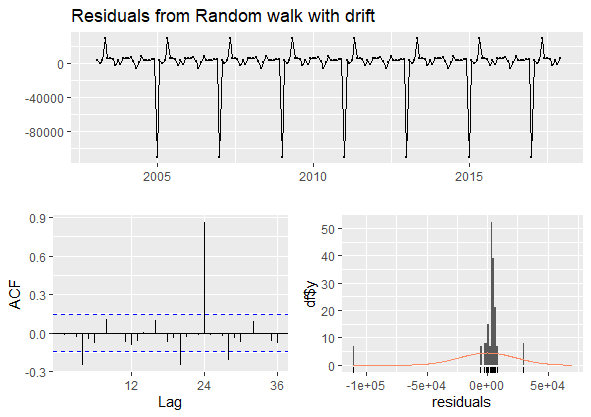
The ACF plot below shows that when lag = 3 the significant boundary is crossed, and it can also be seen that all values are well centred around zero. There is also a fine residual graph showing the normal distribution of the residual value.



Using Random walk with Drift, the graph plot result below shows no sign of seasonality in the forecast and model is not a good fit.



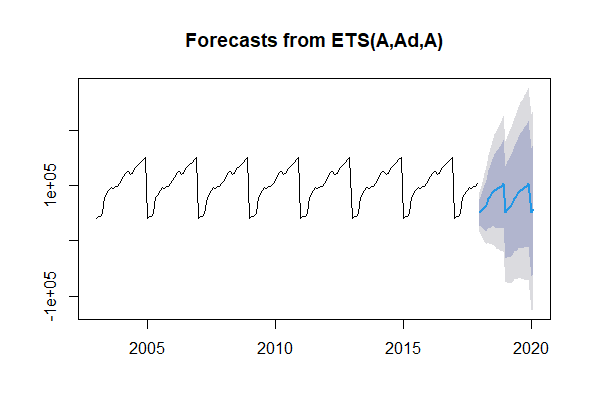
The graph below does not provide sufficient details to further gain insights in the interpretation of the model. Residual shows a definite patten, but the ACF plot with all values around zero and is also showing a lag = 3, this is also seen in the histogram distribution curve.



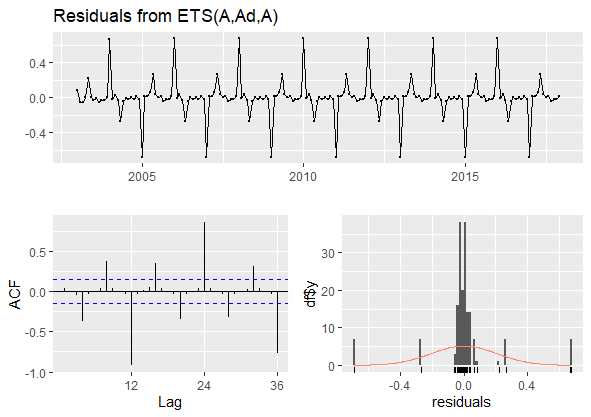
### **Error, Trend, Seasonality (ETS)**

Each model incorporates an observation equation and transition equations, one for each state (level, trend, and seasonal), i.e., state space approaches. The medians of the forecast distributions are used to calculate ETS point forecasts. Forecast distributions for models with only additive components are normal, so the medians and means are equal. The point forecasts for ETS models with multiplicative errors or multiplicative seasonality would not be equal to the means of the forecast distribution functions.

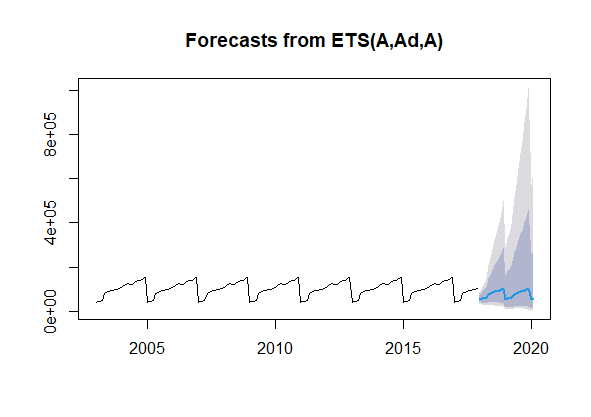
The forecast below shows a multiplicative error and seasonal components detected in the original data and it is good for prediction of time series.



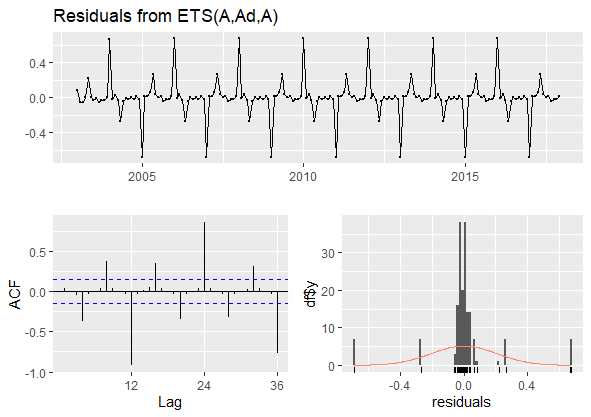
All results below show good results compared to previous ACF plot and normal distribution plot above.



The forecasting result is good for the model of the original dataset, it can be seen the predictability power is better compared to some other results as it follows the seasonal trend

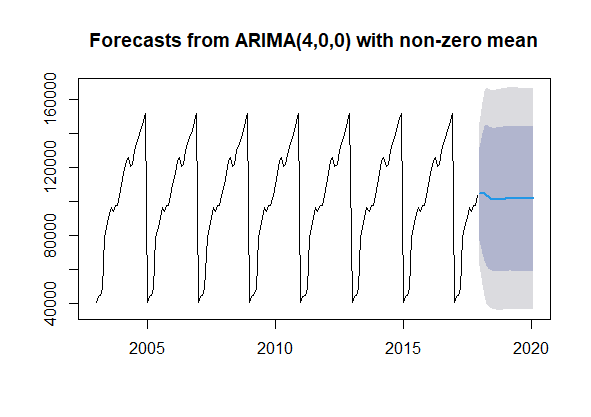


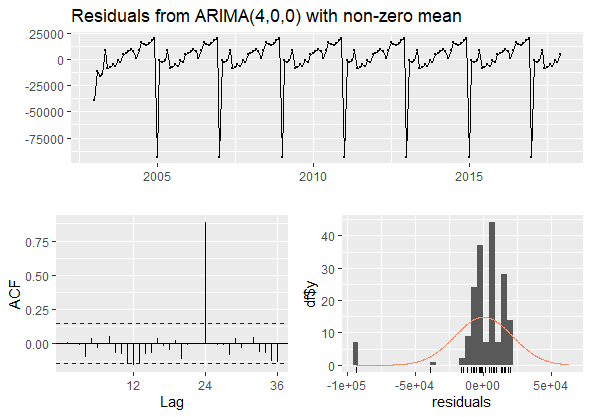
The below residual distribution, ACF plot and normal distribution follows the trend of other models with lag = 3, showing a zero centered values of significance and residual trends being of similar patterns to other models.



### **ARIMA Model**

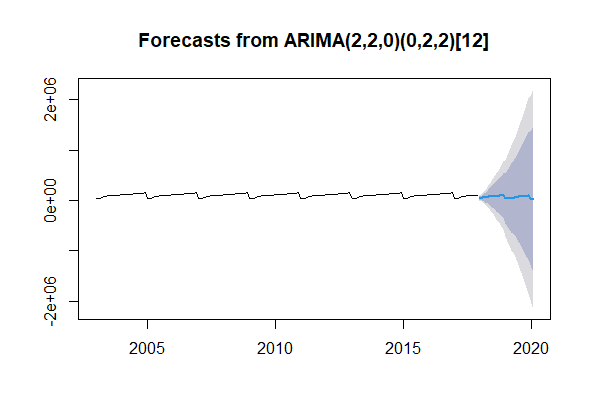
This is the plot of the ARIMA model with no seasonal trend detected in the forecasting result and it does not cover the initial prediction in the original dataset. Looking at the residual plot, it gives enough reason as to the model not being a good fit for prediction interval s it underestimates the time series in the data.



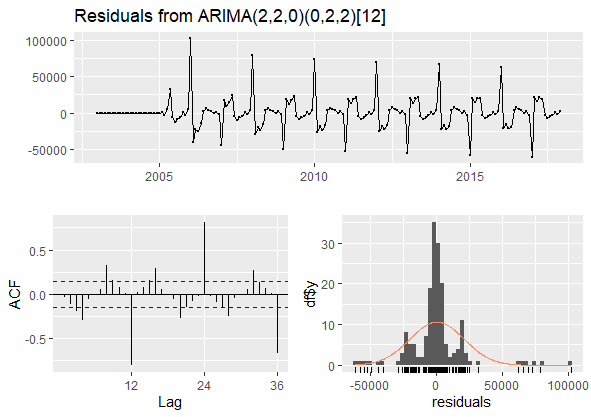


### **Comparing ARIMA and Auto ARIMA Models**

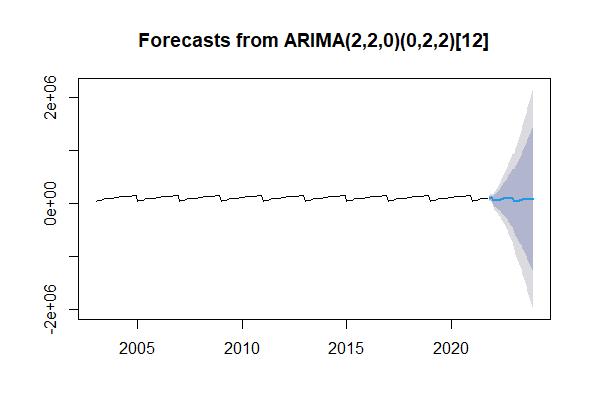
The model shows that there will be a consecutive steady trend in the following years. The accuracy results can be found be found below with low MASE error in both the test and training set and being the most suitable model and also in the residual plot we can get some clear information as the model’s performance:



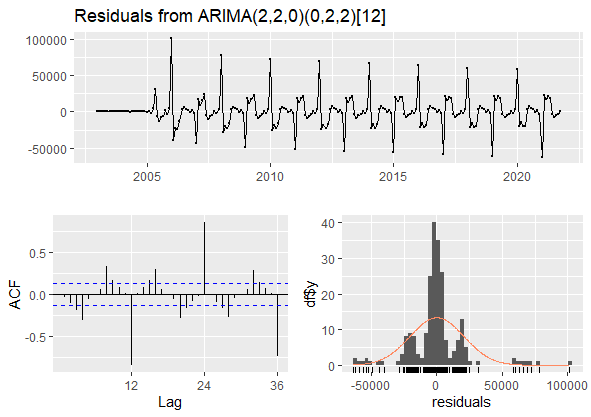




### **Best Model Result**



It can be observed from the above graph that the predicted series for the model is in line with the original data and indeed shows signs of seasonality in the trend. There will indeed be similar increase in passengers in particular periods of the year as compared to some other months seen through out the forecasting analytical trends which will cause the increase in the graph plot as the year winds up and same trend is seen when there is a reduction in the number of passengers in some given period. This information is backed up by these plots below showing significant level of points close to zero and a lag =3 causing a spike to cross given threshold in the model.

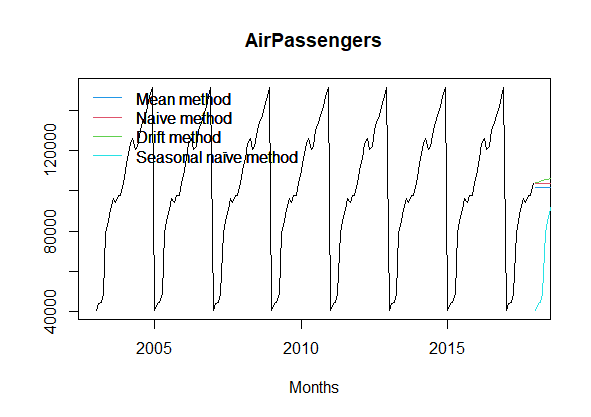




The results above show that the Mean absolute error has the lowest percentage in the model

**Comparing Models**

This graph below shows a comparison for the different models during the Covid era, all models seem to have similar trends except for the Seasonal naïve method which gives least predictability trend for the given period.



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model/Dataset | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil's U |
| Mean Training | 5.819523e-12 | 32696.90 | 26618.01 | -16.62528 | 36.51033 | 0.5012407 | 0.7395625 | NA |
| Mean Test | 1.085650e+03 | 35479.39 | 30618.70 | -18.01072 | 41.34418 | 0.5765772 | 0.7233634 | 2.536626 |
| Naive Training | 351.9721 | 23220.79 | 9638.575 | -5.516159 | 16.30789 | 0.1815029 | -0.01822578 | NA |
| Naive Test | -983.2000 | 35476.40 | 30618.700 | -20.412573 | 42.18565 | 0.5765772 | 0.72336344 | 2.617543 |
| Drift Training | -4.552681e-12 | 23218.12 | 9404.583 | -5.516159 | 16.30789 | 0.1770966 | -0.01822578 | NA |
| Drift Test | -1.383018e+04 | 39101.46 | 32378.560 | -35.749938 | 49.21454 | 0.6097169 | 0.73594121 | 3.236109 |
| Seasonal Naive Training | 0.0 | 55313.03 | 53104.25 | -22.19573 | 63.94806 | 1.0000000 | 0.8017336 | NA |
| Seasonal Naive Test | 24798.8 | 36558.94 | 24798.80 | -20.412573 | 18.80235 | 18.80235 | 0.7536242 | 1.133402 |

## **REPORT OVERVIEW FOR STOCK MARKET ANALYSIS**

This project encapsulates various forecasting approaches for a Microsoft stock price analysis. Stock prices are influenced by a variety of metrics such as supply and demand, performance of the company, and investor sentiment. The stock market is a marketplace that facilitates the buying and selling of corporate stock. The dataset contains the daily close price of the "S&P500 index" stock from 1st January 2015 to 30th April 2022 which will be used to predict trends for the next 60 days and help reduce the risk of loss and increase profit. The main aim of this project is to apply forecasting prediction models on price movement over time in the stock market, then these predictive  models will be validated, analyzed, and evaluated by comparing with the project objectives.

## **INTRODUCTION**

Forecasting techniques are data-driven procedures that aim to forecast future results based on historic data. This historical information is extracted and organized to forecast future values for a dataset variable. In the implementation of this project, I will be concentrating on quantitative forecasting using the variable to predict future close price, descriptive statistics, and complex concepts applied to past data. Good forecasts encapsulate the true patterns and relationships that appear in historical data, but they do not try to recreate past events that will never take place afterwards. Stevenson (2012) summarized several forecasting characteristics in common:

(1) Forecasting techniques assume that the trend, cyclic, and seasonal components are stable and that previous patterns will be repeated.

(2) Forecast errors are unavoidable due to the random nature of time series.

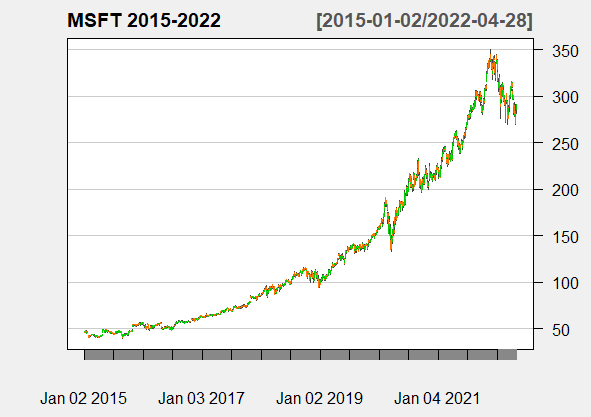
(3) A forecast can be defined as a guesstimate of the average. Forecasts for a group are more accurate than individual forecasts for the group.

(4) Longer-term forecasts are less accurate than short-term forecasts because the former contains more uncertainties.

**STEPS FOR STOCK VISUALIZATION AND ANALYSIS:**

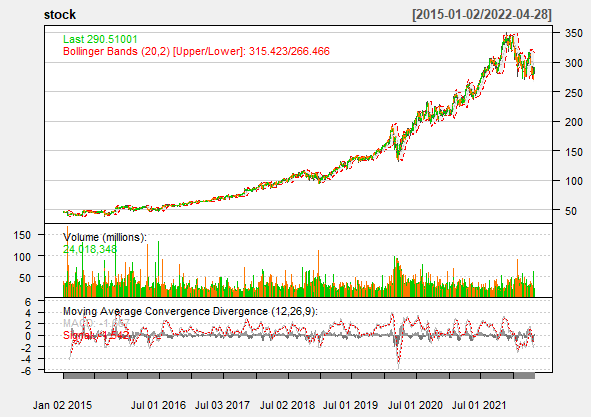
* Load the libraries into the R environment
* Explore the data using head() function
* Investigate the historic dataset with a chartSeries() function to see the closing price behaviour.

The graph below shows an upward trend as well as seasonal patterns in the closing price data from 2015 – 2022 which will aid in investigating patterns to help make better investment decisions.



The Bollinger Bands chart below shows a percentage change, Moving Average Convergence Divergence and Volumed Traded of the previous graph. Moving Average Convergence Divergence is a momentum indicator that follows a general pattern and displays the relationship between two trendlines of a stock's returns. It indicates the entry and exit windows for trading, and it basically calculates the qualitative distinction between the 26 days and 12- day Exponential Moving Average (EMA). The MACD histogram is a beautiful graphic illustration of the MACD's difference from its nine-day EMA. When the MACD has gone beyond its nine-day EMA, the histogram is positive; when it is below its nine-day EMA, the histogram is negative. If prices are rising, the histogram expands as the rate of price movement accelerates and contracts as the rate of price movement slows. When prices are falling, the same concept applies. As such, many traders use the MACD indicator to measure the strength of a price move rather than the direction of a trend because it responds to the speed of price movement, the MACD histogram is the sole indicator for traders to rely on to measure momentum.

The Exponential Moving Average (EMA) is a type of moving average (MA) that gives much more current data points more weight and significance. The exponential moving average, also known as the exponentially weighted moving average, is a type of moving average.



## **METHODOLOGY**

### **Autoregressive Integrating Moving Average (ARIMA)**

ARIMA is a statistical analytical method that employs time series data to either better understand the data or forecast future events. ARIMA is a type of regression analysis that measures the strength of one dependent variable in relation to other variables that change. It smooths time series data using lagged moving averages and can be used in technical analysis to forecast future security prices. In certain market conditions, such as financial crises or periods of rapid technological change, they can be inaccurate. That is why ARIMA model is commonly used in financial services and economics because it is known to be dependable, efficient, and capable of predicting short-term share market movements. ARIMA requires that time series be made stationary before any kind of analysis. To be stationary, a time series' statistical properties (mean, variance, etc.) must be the same throughout the series, regardless of the time at which they are observed. A stationary time series appears to lack long-term predictable patterns such as trends and seasonality. There are three components of this method:

* **Autoregression (AR):** a model in which a changing variable regresses on its own lag, or preceding values. This is indicated by the model's "p" value.
* **Integrated (I) or Differencing:** the differencing of raw observations that enables the time series to become stationary (that is the data values will be replaced by the difference between the data values and the previous values). The "d" value in the model indicates this. If d = 1, it examines the difference between two time series entries; if d = 2, it examines the differences of the differences obtained at d = 1.
* **Moving average (MA):** a moving average model applied to lagged observations that integrates the dependency between an observation and a residual error. This is represented by q values.

**STEPS FOR BUILDING ARIMA MODEL**

1. **Testing and Assurance of Stationarity:** A time series is said to be stationary if being modelled using the Box-Jenkins method, it has no trend and has a constant mean and variance over time, making it simple to predict values.

Testing for Stationarity – We implement Augmented Dickey-Fuller(ADF) unit root test to check for stationarity. If a time series is to be stationary, the p-value from the ADF test must be less than 0.05 or 5%. If the p-value is greater than 0.05 or 5%, the time series has a unit root, indicating that it is a non-stationary process.

Differencing – The differencing method is used to transform a non-stationary process to a stationary process. Finding the differences between consecutive values of a time series data is referred to as differencing a time series. The differenced values form a new time series dataset that can be tested for new correlations or other statistical properties.

We can use the differencing method repeatedly, resulting in "first differences," "second order differences," and so on. To make a time series stationary, we use the right differencing order (d).

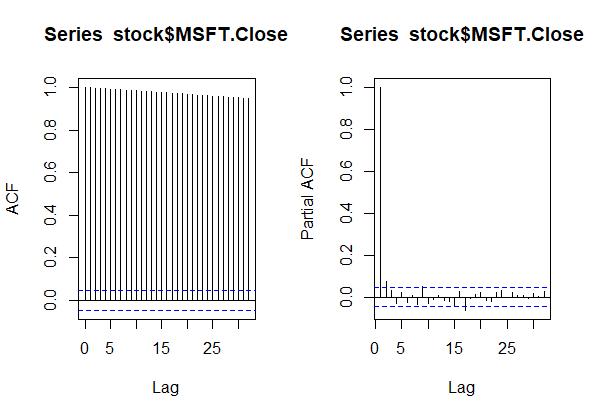
1. **Identification of p and q:** We use the Autocorrelation function (ACF) and the Partial Autocorrelation function (PACF) to find the appropriate order of Autoregressive (AR) and Moving average (MA) processes (PACF).

First thing is finding the p order of an AR model for the AR models, the ACF will dampen exponentially, and the PACF will be used to find the order (p) of the AR model. If the PACF has one significant spike at lag 1, we have an AR model of order 1, i.e., AR (1). If we have significant spikes on the PACF at lags 1, 2, and 3, we have an AR model of order 3.

We start by defining the q order of the MA model f or MA models, the PACF will dampen exponentially, and the ACF plot will be used to determine the pattern of the MA process. If we have one significant spike on the ACF at lag 1, we have an MA model of order 1, i.e., MA (1). If the ACF has significant spikes at lags 1, 2, and 3, we have an MA model of order MA (3).

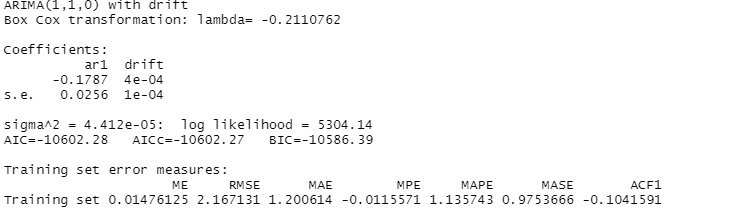
1. **Estimation and Forecasting:** After the parameters (p, d, q) have been ascertained, we can estimate the ARIMA model's accuracy on a training data set but use the fitted model for forecasting the values of the test data set using a forecasting function. Finally, we check to see if our forecasted values match the actual values.

After performing an ADF test on the close price set, we apply the ACF (Autocorrelation function) and PACF (Partial autocorrelation function) functions to the dataset.

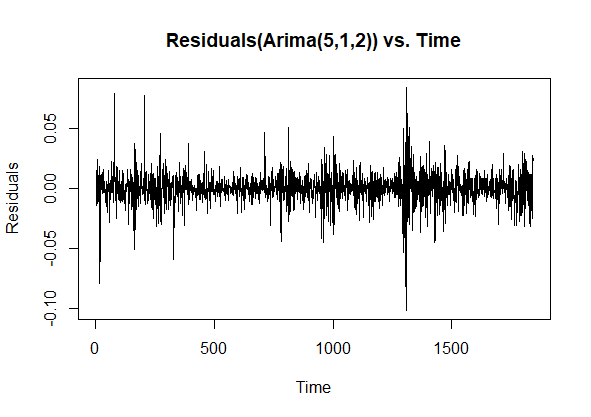


The autocorrelation, which is the correlation of a time series with its past values, is significant for many lags, but it is possible that the autocorrelation at posterior lags is simply due to the propagation of autocorrelation at the first lags.

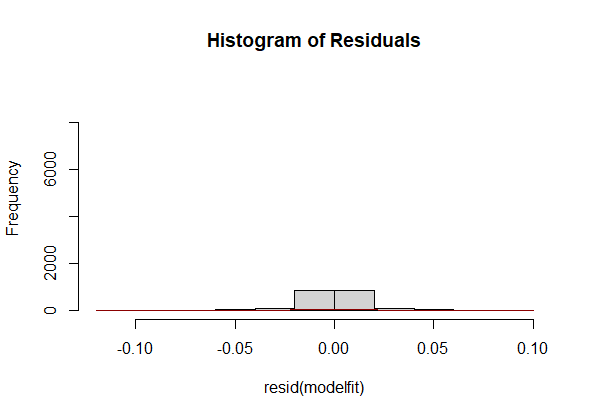
Following that, use the ACF and PACF plots to identify the (q) order, and the PACF will dampen exponentially. This results in a significant spike only at first lags, implying that all higher order autocorrelation is effectively explained by first lag autocorrelation. The model is then fitted to the price data. Then the model summary is examined using the residuals of the ARIMA model.



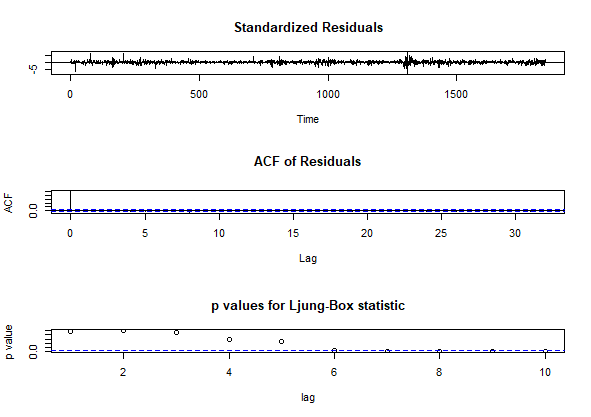
From the below residual plot after fitting the model, this is basically the difference between the observations and the fitted values. This can later be observed with a normal curve to examine the residuals.



The Residual Histogram is used to find whether the variance is normally distributed. A symmetric bell-shaped histogram with an evenly distributed distribution around zero shows that the normality assumption is most probably correct. If the histogram shows that random error is not normally distributed, the model's underlying assumptions may have been violated. The residuals plot has a normal curve adjustment, which gives a useful starting point to proceed this review. We can make the final residuals plot, which will provide us with the standardized residuals, ACF of residuals, and p-values for Ljung-Box statistic plots.



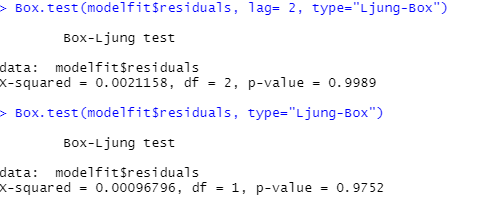
Ljung-Box test is a method for testing for the lack of serial autocorrelation up to a given lag k. The test evaluates whether the errors are iid (i.e., white noise) or if there is something else behind them, if the autocorrelations for the errors or residuals are non-zero. It is principally a lack of fit test: if the autocorrelations of the residuals are very small, we say that the model does not exhibit 'significant lack of fit.' The Ljung-Box p-values are the focus of these three graphs. Our null hypothesis for the Ljung-Box test is as follows:



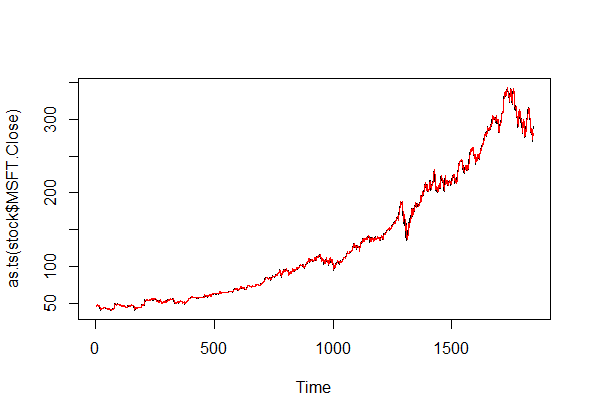
A significant p-value greater than 0.05 does not reject the fact that the dataset points are not correlated with this null hypothesis. This means the residuals are independent in the time series model which is also the reason for the model creation.

A lag of 1, would show there is autocorrelation with each lag. The test statistic is Q = 0.0021158, and the p-value is 0.9989, that is much significantly larger than 0.05. As a result, we fail to reject the test's null hypothesis and conclude that the data values are independent. At a lag of 2, test statistics is Q = 0.00096796 and the p-value is 0.9752, that is much significantly larger than 0.05 as well, which also means we fail to reject the null hypothesis.

Based on this visual observation, we proceed to conduct an independent test to investigate the lag. The results from the Box plot shows that we do not reject the null hypothesis and still further observe this.

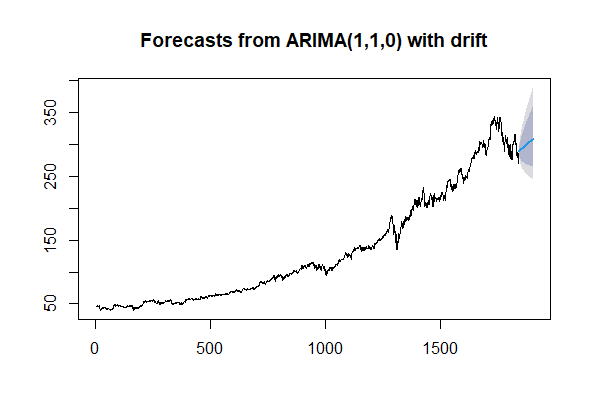


The  model prediction will be plotted in a red line over the real train set stock close price after implementing and evaluating the new ARIMA model.

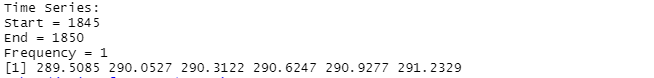


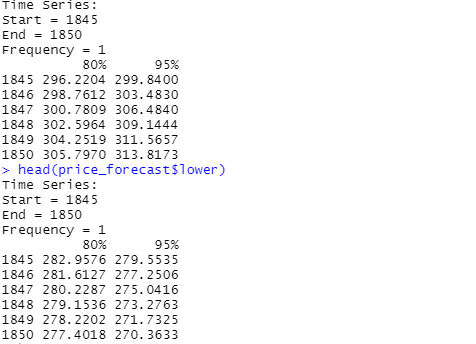
**ARIMA Results Overview**

With the model in place, I already can forecast  daily close price values into the upcoming months.  We are primarily concerned with forecasting the close stock price for the next 60 days.The mean of our prediction is represented by a blue line.

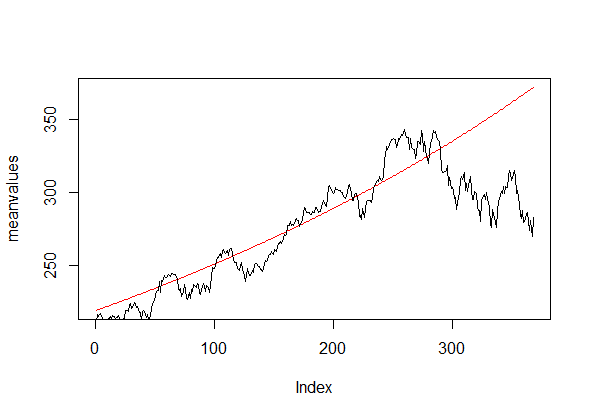


With blue line described, it can be observed that the darker and lighter areas represent 80 percent and 95 percent confidence intervals in lower and upper scenarios, respectively. It can be seen that the first set of result shows the average forecasted price. While the other two categories are the upper and lower scenario with the confidence level for the forecasted price.





For the ARIMA model to be built, I will perform a 70% - 30% train-test split method on the close price dataset. The prediction is then implemented to the train set, and the mean tendency of the forecasting over the test set close price move is plotted.



The red line depicts our average forecasting prediction tendency over the stock's actual close price. The trend demonstrates a good method for forecasting the future direction of the close price.

### **Generalized Autoregressive Conditional Heteroskedasticity Model (GARCH)**

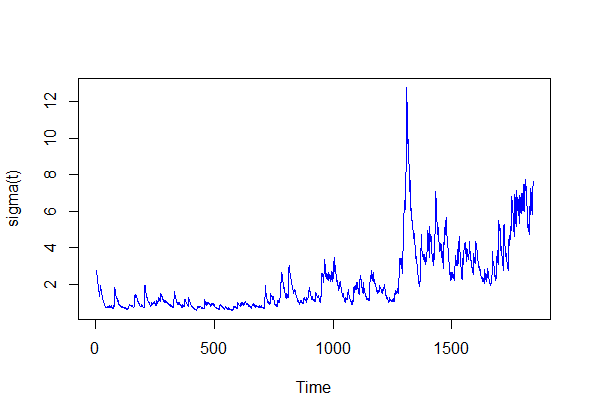
The earlier ARIMA model produced excellent results, but the biases are primarily explained by the volatile analyses in our dataset and financial market series. When it comes to predicting new values, this circumstance is reason for concern. GARCH models are frequently used to measure the volatility of stock, currency, index, and cryptocurrency returns. This tool is used by market participants to price assets and determine which asset will foreseeably provide the highest return in their investment. They can use this tool to fine-tune their asset allocation and risk management. As explained previously, volatility clustering is based on periods of relative calm and periods of high volatility, which is very typical in financial stock market data, and the GARCH model is a very good approach to minimizing the volatility effect. We will take the normal residuals and square them after evaluating the GARCH model implementation. Any volatile values will be visible in the residual plots. We compare a standard GARCH (1, 1) model to an ARMA (5,2) model to see if we can improve our accuracy and model parameters.

**Autoregressive Conditional Heteroscedasticity (ARCH)**

Autoregressive Conditional Heteroscedasticity (ARCH) is a time series statistical theory that represents the variance of the existing error term or innovation as a component of the actual sizes of the previous time periods' error terms; regularly, the variance is related to the squares of the previous innovations. When the error variance in a time series follows an autoregressive (AR) model, the ARCH model is appropriate; if an autoregressive moving average (ARMA) model is assumed for the error variance, the model is a Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model.

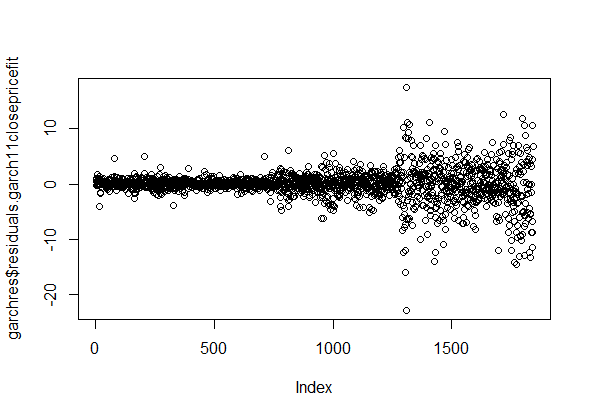
**Autoregressive Fractionally Integrated Moving Average (ARFIMA or FARIMA)**

Autoregressive fractionally integrated moving average models are time series models that extend ARIMA (autoregressive integrated moving average) models by allowing non-integer differencing parameter values. These models are useful for modeling time series with long memory, or deviations from the long run mean that decay more slowly than an exponential decay. It is  frequently utilized and also common to simply extend the ARIMA (p,d,q) notation for models by allowing the order of differencing, d, to take fractional values.



The series appears to vibrate around y = 0 and shows volatility clustering. The coefficient estimates have p values close to zero, indicating that the estimates are significant. The calculated conditional variances show high volatility through time = 1400, then low volatility through time = 2000, proposing a stochastic model for conditional volatility.

Apply the previously defined model to the close price dataset to find ARFIMA (autoregressive fractionally integrated moving average) parameters. With the variables gathered, and choose ARFIMA (3,0,2) then integrate them into a GARCH model. Fit the  model to create a volatility plot. From this plot, the last years of the data have higher gains, that can be associated to market economic insecurity in recent years.



The Akaike information criterion (AIC) is a measure of the out-of-sample prediction error and, as a result, the relative quality of statistical models for a given set of data. AIC is a single-number score which can be used to evaluate that also of several models is most likely to be the best fit for a given dataset. It forecasts models in a relative manner, which means that AIC scores are only useful when compared to other AIC scores for the same dataset. A lower AIC score is preferable. Furthermore, the Akaike (AIC), Bayes (BIC), Hannan-Quinn, and Shibata criterion values are lower than those seen in the other model setting. It demonstrates the Akaike (AIC), Bayes (BIC), Hannan-Quinn, and Shibata maximum likelihood estimation criteria. The lower these values, the better the fit of the model. It can be seen our Akaike and other model information below:

Akaike 3.831490

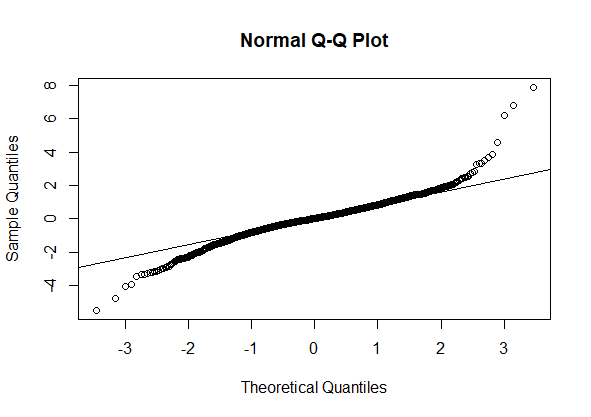
Bayes 3.858430

Shibata 3.831443

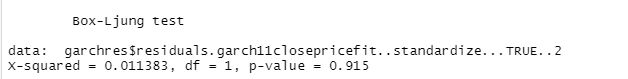
Hannan-Quinn 3.841422

A Q-Q plot, which stands for "quantile-quantile" plot, is a kind of plot which could be used to evaluate whether a set of data came from a theoretical distribution. Quantiles are points in a dataset where a certain percentage of the data falls. Q-Q plots describe the quantiles in your sample data and plot them against the theoretical distribution's quantiles.

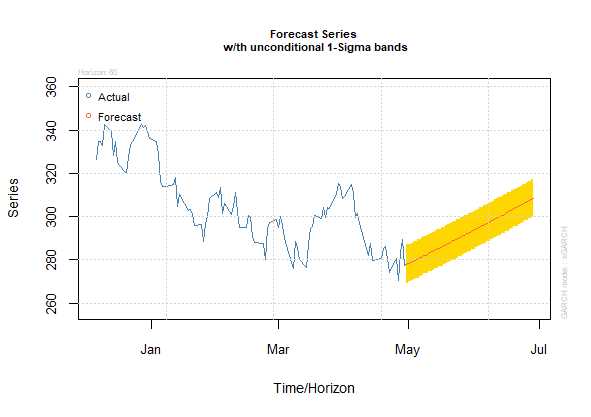
In this Q-Q plot, there is a clear centering of data points around the straight line, indicating that this dataset most likely does not follow a normal distribution, because there is a deviation particularly near the tails.



With our standardized residuals' normality plots in hand, we perform a Ljung Box test on the squared standardized residuals. It is seen from the Ljung Box test that our standardized squared residuals do not reject the null hypothesis, confirming that there really is no autocorrelation between them. We can now forecast our next 60 days based on our volatility and residuals behavior and make a comparison to other modeling techniques.

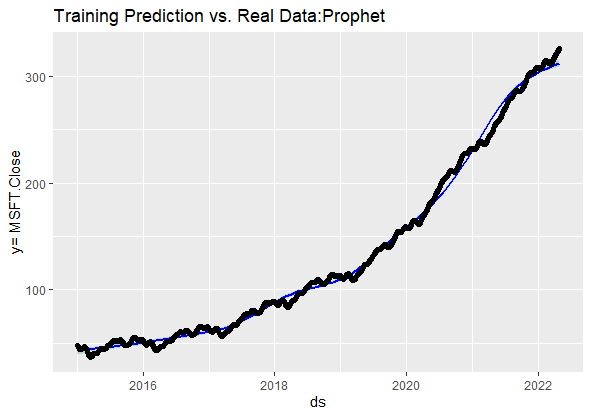


From our fitted model, the visual plot below shows the data prediction. We will not thoroughly evaluate this model because we will be using GARCH to rectify our ARIMA predictions.



### **Prophet Model**

Prophet's mission is to "make it easier for experts and non-experts to produce high-quality forecasts that keep up with demand." It can produce robust and reliable forecasts (often outperforming other common forecasting techniques) with minimal manual effort, while also allowing for the application of domain knowledge via easily interpretable variables.



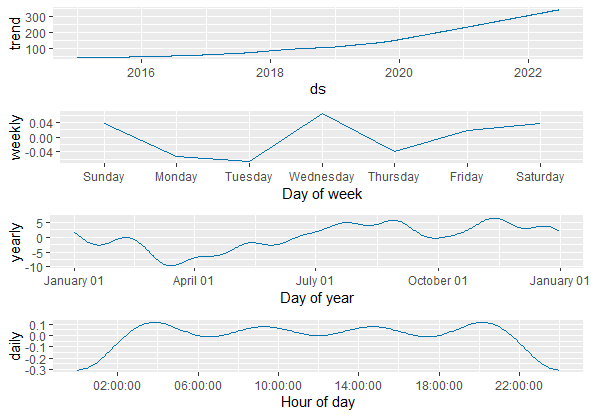
To create a dataframe of future dates, Prophet includes a helper function called make\_future\_dataframe. The make\_future\_dataframe function allows users to specify the frequency and number of periods into the future that you want to forecast. The frequency is set to days by default. Because I am using daily periodicity data in this example, freq has its default periods argument to 365, so I indicated that I want to forecast 60 days in advance.

To make accurate predictions for each row in the future dataframe, use the predict() function. Prophet will have introduced a new dataframe assigned to the forecast parameter at this point, which consists of the forecasted attributes for future dates in a column called yhat, and unpredictability intervals and forecast components. Using Prophet's built-in plot helper function, we can visualize the forecast.

Proceed by calculating the model performance after applying the model and plotting the forecast. Because this is a new model application, I will be using the accuracy function to make comparisons of the real values to the train set's parameter estimates. In Prophet, the correct approach is to create a cross-validation process and analyze the model performance metrics by attempting to make comparisons between the ARIMA vs the other models using the same method. Decide the accuracy after showing the dataset, compare the real data to the predicted values.



Prophet was able to precisely model the fundamental trend in the data, but also weekly and yearly seasonality, as evidenced by the forecast and component visualizations (e.g., lower order volume on weekend and holidays).



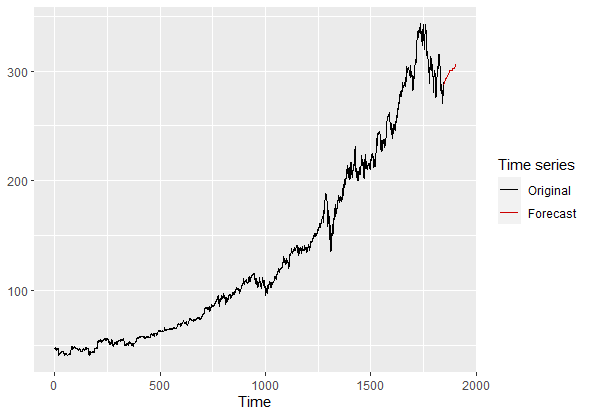
### **Knn Regression For Time-Series Forecasting**

The KNN model applies to both classification and regression problems. The most common application is for classification problems. The KNN can be applied to almost any regression problem. The focus of this research is to show various forecasting techniques, make a comparison them, and analyze prediction behaviour patterns. Regarding our KNN exploration, I recommended that it be used for both classification and regression tasks. The model employs 'feature similarity' to predict the values of new data points, assigning a new point to a value based on how closely it represents the points in the training set. The primary aim is to use a KNN exploratory procedure to forecast new values of the stock price.

Because I am using a heuristic approach to find the best k value,  k = 70 as is used as an experimental value. KNN techniques necessitate tuning experimental studies, but also because the focus of this research is to illustrate the various forecasting techniques, I will concentrate more on the implementation than the model tuning to achieve the best accuracy.



In the rolling origin function, I used the model and the time series attributed with it to analyze the model's forecasting accuracy. After studying the model, I  plotted my  predictions in the graph below.



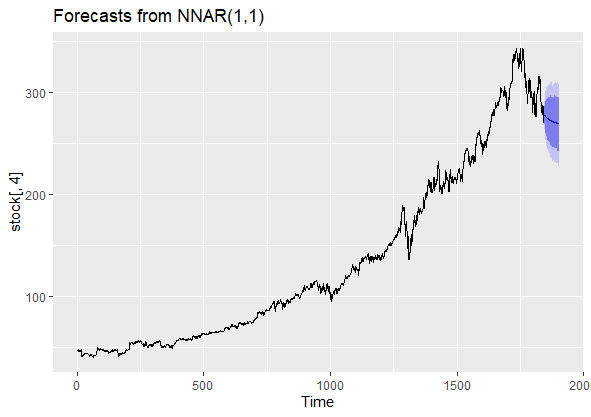
With this new KNN predictions graph, a comparison between it to and the other models. Prior to making comparisons of these predictions, I will quickly introduce a fourth final approach to forecasting using neural networks.

### **Feed-Forward Neural Networks**

The most basic type of neural network is a single hidden layer neural network. There is only one layer of input nodes in this single hidden layer form, which transmits weighted inputs to a successive layer of receiving nodes. A time series is fitted with a single hidden layer neural network model. The function model approach employs lagged time series attributes as input data, resulting in a non-linear autoregressive model.



Forecasting univariate time series using feed-forward neural networks with a single hidden layer and lagged inputs. The nnetar function in the forecast package for R fits a neural network model to a time series using the time series' log transformation as inputs. To ensure that the residuals are roughly homoscedastic, I used a Box Cox lambda. With the neural net fitted, we forecast the next 60 values. Then applied the nnetarfunction with the parameterlambda that were chosen.



## **CONCLUSION**

In this study, we concentrated on the implementation of specific models, learning how to use them with the objective of forecasting new price values. As seen in the results, the models performed similarly in terms of future tendency forecasting. All of the models predicted a price increase in the next 60 days. In  conclusion,  the ARIMA and KNN models performed admirably within their respective prediction intervals and precision metrics. The other approaches did not perform as well as ARIMA or KNN models because they are new in this forecasting approach and the primary objective is to apply them in an intuitive form. Perhaps Prophet and Neural Networks require more fine-tuning to produce more accurate results.

Another important point we neglected to mention is that auto-regressive models, which are using past data to predict future values, have an asymptotic prediction in long-term forecasts. Finally, we conclude that ARIMA and KNN are the better predicting models in this scenario, with very interesting results when we added GARCH to our ARIMA approach. The other models used did not perform and also ARIMA and KNN under our performance measures, and this could be attributable to the fact that they demand more tuning processes and training, testing approaches, or they are not as efficient as the other models due to their key application use in classification rather than forecasting.

## **REFERENCES**

https://core.ac.uk/download/pdf/323054843.pdf

Forecasting/Exam\_2.R at master · Usmanmajid-IESEG/Forecasting. (2022). <https://github.com/Usmanmajid-IESEG/Forecasting/blob/master/Exam_2.R>

Forecasting: Principles and Practice (2nd ed). (2022). <https://otexts.com/fpp2/>

How to Trade the MACD Divergence. (2022). <https://www.investopedia.com/articles/forex/05/macddiverge.asp>

Nour, M. (2022). Time Series Analysis. http://ropatics.com/BI/BI\_-\_time\_series\_analysis.html

plot.stl?, I., & Simpson, G. (2022). Interpreting range bars in R's plot.stl? <https://stats.stackexchange.com/questions/7876/interpreting-range-bars-in-rs-plot-stl>

RPubs - Time series analysis - accuracy in forecasting. (2022). https://rpubs.com/RatherBit/90267

<https://buildmedia.readthedocs.org/media/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf>

https://koalatea.io/r-ljung-box-test/

Stevenson, J. W. (2012). Operations Management (11th Edition). McGraw-Hill.

Forecasting Air Passenger Data - Maple Help. (2022). <https://www.maplesoft.com/support/help/maple/view.aspx?path=examples/AirPassengers>

GitHub - Stat-Wizards/Forcasting-A-Time-Series-Stock-Market-Data: A Comparative Analysis of a Real Close-price Stock Market Data with The Classical Time Series, Machine Learning and a Single Layered Neural Network Model. (2022). <https://github.com/Stat-Wizards/Forcasting-A-Time-Series-Stock-Market-Data>

Visualizing Time Series in R - Boostedml. (2022). <https://boostedml.com/2020/05/visualizing-time-series-in-r.html>

https://www.investopedia.com/articles/forex/05/macddiverge.asp

https://www.mssqltips.com/sqlservertip/6778/time-series-forecasting-methods-with-r-examples/

https://www.originlab.com/doc/origin-help/residual-plot-analysis

https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical,values%20based%20on%20past%20values.

https://www.statology.org/ljung-box-test/

https://uk.mathworks.com/help/econ/perform-garch-model-residual-diagnostics-using-econometric-modeler.html

https://www.statology.org/q-q-plot-r/

https://www.idrisstsafack.com/post/garch-models-with-r-programming-a-practical-example-with-tesla-stock

https://mode.com/example-gallery/forecasting\_prophet\_r\_cookbook/

http://uc-r.github.io/feedforward\_DNN

https://rpubs.com/mr148/313595

https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/

https://towardsdatascience.com/forecasting-time-series-data-stock-price-analysis-324bcc520af5

https://medium.com/swlh/seasonality-analysis-and-forecast-in-time-series-b8fbba820327

https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/control-charts/how-to/box-cox-transformation/interpret-the-results/key-results/#:~:text=For%20the%20Box%2DCox%20transformation,1%2C%20a%20transformation%20is%20appropriate.

https://www.statisticshowto.com/box-cox-transformation/