APPLE STOCK PRICE FORECASTING



Gunabhiram billa(11484513)

Bhanu Shruthi Nayudu (11480187)

**TABLES OF CONTENTS**

1. **INTRODUCTION**
2. **LITERATURE REVIEW**
3. **RESEARCH HYPOTHESIS**
4. **METHODOLGY**
5. **CODE**
6. **CONCLUSION AND IMPLICATIONS FOR FUTURE**
7. **REFERENCES**

**1.INTRODUCTION**

Forecasting is required in many situations: deciding whether to build another power generation plant in the next five years requires forecasts of future demand; scheduling staff in a call center next week requires forecasts of call volumes; stocking an inventory requires forecasts of stock requirements. Forecasts can be required several years in advance (for the case of capital investments), or only a few minutes beforehand (for telecommunication routing). Whatever the circumstances or time horizons involved, forecasting is an important aid to effective and efficient planning. Stock market process is full of uncertainty; hence stock prices forecasting very important in finance and business. For stockbrokers, understanding trends and supported by prediction software for forecasting is very important for decision making. Stock markets are where individual and institutional investors come together to buy and sell shares in a public venue. Nowadays these exchanges exist as electronic marketplaces. That supply and demand help determine the price for each security or the levels at which stock market participants — investors and traders — are willing to buy or sell. For a new investor general research which is associated with the stock or share market is not enough to make the decision. The common trend towards the stock market among the society is highly risky for investment so most of the people are not able to make decisions based on common trends. The seasonal variance and steady flow of any index will help both existing and new investors to understand and make a decision to invest in the share market.

A popular and widely used statistical method for time series forecasting is the ARIMA model. It is one of the most popular models to predict linear time series data. This model has been used extensively in the field of finance and economics as it is known to be robust, efficient, and has a strong potential for short-term share market prediction. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting and provide complementary approaches to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the auto-correlation(Autocorrelation is the degree of similarity between a given time series and a lagged version of itself over successive time intervals) in the data.

In this project we analyze data of Apple stock from Kaggle. This data has been collect from years 2014 – 2019 in weekly pattern. This data helps in predicting the apple stock price in future days in Delhi stock exchange. In this project we will be using TIME SERIES REGRESSION, ARIMA, AND BOX-JERKINS models to predict the stock values.

**2.LITERATURE REVIEW**

There are two main schools of thought in the financial markets, technical analysis and fundamental analysis. Fundamental analysis attempts to determine a stock’s value by focusing on underlying factors that affect a company’s actual business and its future prospects. Fundamental analysis can be performed on industries or the economy as a whole. Technical analysis, on the other hand, looks at the price movement of a stock and uses this data to predict its future price movements.

The basic theory regarding stock price forecasting is the Efficient Market Hypothesis (EMH), which asserts that the price of a stock reflects all information available and everyone has some degree of access to the information. The implication of EMH is that the market reacts instantaneously to news and no one can outperform the market in the long run. However, the degree of market efficiency is controversial and many believe that one can beat the market in a short period of time1. Time series analysis covers a large number of forecasting methods. Researchers have developed numerous modifications to the basic ARIMA model and found considerable success in these methods. The modifications include clustering time series from ARMA models. Almost all these studies suggest that additional factors should be taken into account on top of the basic or unmodified model.

In Devi, Sundar and Alli’s paper (2013), “An effective time series analysis for stock trend prediction using ARIMA model for Nifty Midcap-50”, the authors claimed that time series analysis will be the best technique in forecasting stock price and predicting the trend and this technique helped naïve investors make rational decisions through considering seasonal variance and steady flow in the stock market. On the one hand, they believed that the general trend of stock market in society had highly risky in investment and not appropriate for trade; on the other hand, past investors and professionals relied on a wealth of knowledge about share price movement forecasting without proper forecast tools; however, for beginners, they are helpless as they lack of excellent financial knowledge and experiences.

**3.RESEARCH HYPOTHESIS**

Null Hypothesis (Ho): ARIMA model is best suitable for the dataset.

Alternative Hypothesis (Ha): ARIMA model is not the best suitable for the data set.

**4.METHODOLOGY**

In this project we will be performing different ARIMA models and box jerkins models on the data to find out different values such as lag, ACF, PACF and depending on these values we will select a best model which has better accuracy at predicting the time series for the dataset.

In this data set, we have 4 attributes DATE, STATE, COMMODITY AND PRICE, but we use only price and date attributes. Price is dependent variable and date is independent variable for our forecasting. We have a total of 262 rows of data. We will be focusing on day seasonal cycle, price value for projecting time series.

The p-value is the most important concept of statistical significance and therefore an applicability of the results of a (trend) test. The p-value represents the probability of an error when considering the real value of estimated parameter differs from the computed (or static) one, e.g., that the zero hypothesis holds although we considered the alternative one. Usually, if the p-value is under 5%, we accept the alternative hypothesis, because the risk of its invalidity is relatively low.

For F-statistics,

1. Formulate the test statistic for the F-test a.k.a. the **F-statistic**.
2. Identify the **P**robability **D**ensity **F**unction of the random variable that the F-statistic represents *under the assumption that the null hypothesis is true*.
3. Plug in the values into the formula for the F-statistic and calculate the corresponding probability value using the **P**robability **D**ensity **F**unction found in step 2. This is the probability of observing the F-statistic value *assuming that the null hypothesis is true*.
4. If the probability found in step 3 is less than the error threshold such as 0.05, reject the null hypothesis and accept the alternate hypothesis at a confidence level of (1.0 — error threshold), for e.g. 1–0.05 = 0.95 (i.e. 95% confidence level). Otherwise, accept the null hypothesis with a probability of error equal to the threshold error, for e.g. at 0.05 or 5%.

**RANDOM WALK**

A random walk is different from a list of random numbers because the next value in the sequence is a modification of the previous value in the sequence. The process used to generate the series forces dependence from one-time step to the next. This dependence provides some consistency from step-to-step rather than the large jumps that a series of independent, random numbers provides.

It is this dependency that gives the process its name as a “random walk” or a “drunkard’s walk”.

A simple model of a random walk is as follows:

1. Start with a random number of either -1 or 1.
2. Randomly select a -1 or 1 and add it to the observation from the previous time step.
3. Repeat step 2 for as long as you like.

y(t) = B0 + B1\*X(t-1) + e(t)

Where *y(t)* is the next value in the series. *B0* is a coefficient that if set to a value other than zero adds a constant drift to the random walk. *B1* is a coefficient to weight the previous time step and is set to 1.0. *X(t-1)* is the observation at the previous time step. *e(t)* is the white noise or random fluctuation at that time.

**DECOMPOSITION**

* Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components.
* Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

**MOVING AVERAGE**

Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.



where εt is white noise. We refer to this as an **MA(**q**) model**, a moving average model of order qq. Of course, we do not *observe* the values of εt, so it is not really a regression in the usual sense.

Notice that each value of yt can be thought of as a weighted moving average of the past few forecast errors. However, moving average *models* should not be confused with the moving average *smoothing*. A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values.

**HOLTS EXPONENTIAL SMOOTHING**

Holt's two-parameter model, also known as linear exponential smoothing, is a popular smoothing model for forecasting data with trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for last period's trend. The trend itself is updated over time through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast. Holt's model uses two parameters, one for the overall smoothing and the other for the trend smoothing equation. The method is also called double exponential smoothing or trend-enhanced exponential smoothing.

**ARIMA**

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.

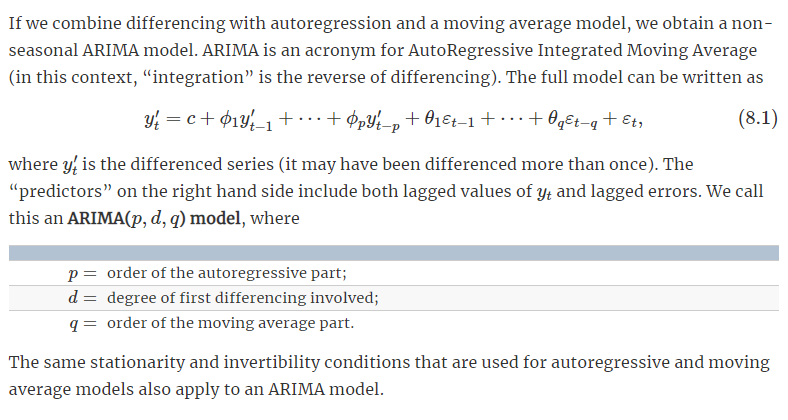
Some cases can be confusing — a time series with cyclic behaviour (but with no trend or seasonality) is stationary. This is because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.

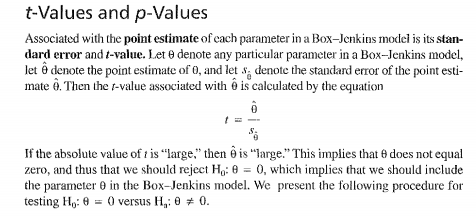
In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behaviour is possible), with constant variance.

**NON-SEASONAL BOX-JENKINS MODEL**

*Checking for Stationarity*

For ARIMA, time series has to be made **stationary**for further analysis**.**For a time series to be stationary, its statistical properties(mean, variance, etc) will be the same throughout the series, irrespective of the time at which you observe them. A stationary time series will have no long-term predictable patterns such as trends or seasonality. Time plots will show the series to roughly have a horizontal trend with the constant variance.

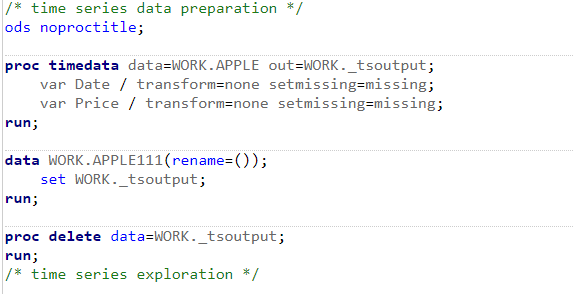




**5.CODES**

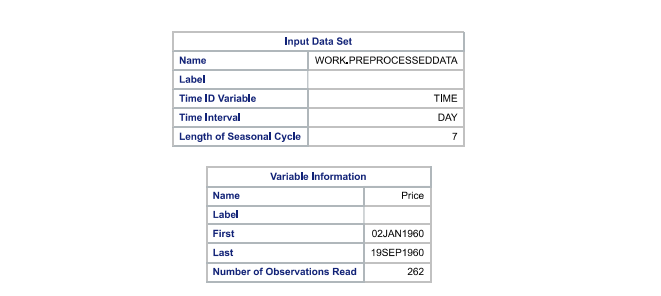
SAS CODE:

**DATA PREPARATION:**



RESULT:

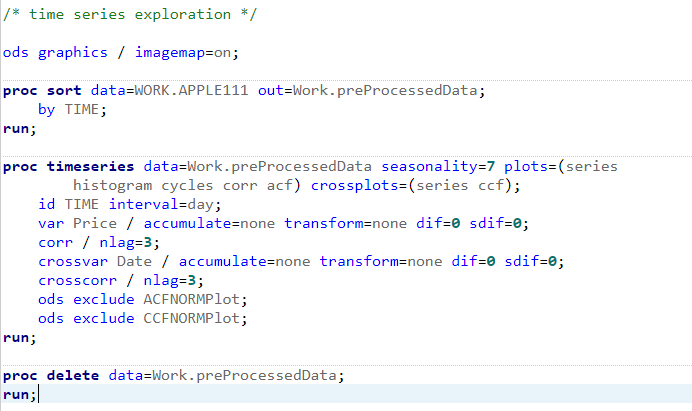
We insert data set into the preparation code and we add time id to the dataset for using it in the futher code.

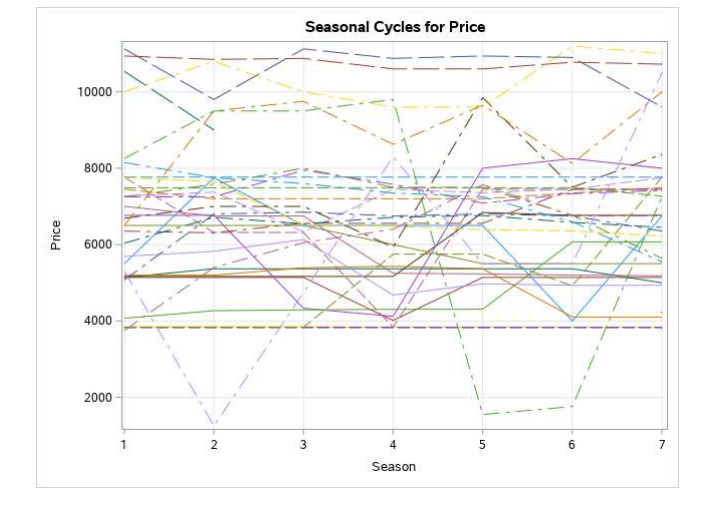
Next we do data exploration to plot graphs for the given values

time id variable is TIME, time interval is day and length of seasonal cycle is 7.

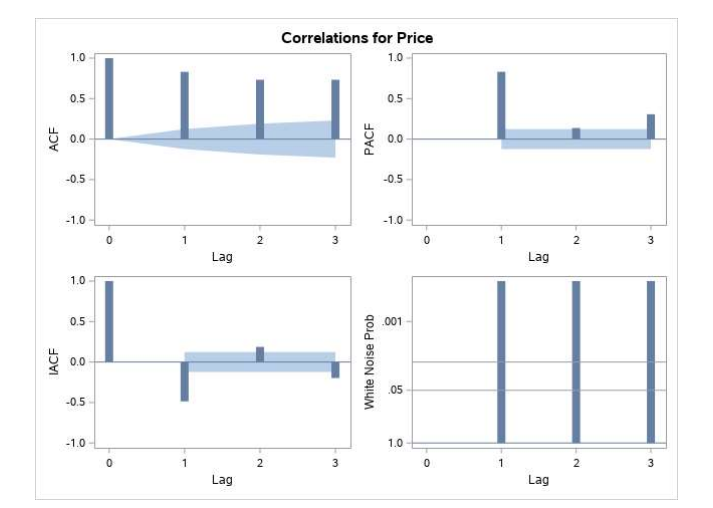


**DATA EXPLORATION:**

In this step we plot correlation, auto correlation, cross-correlation plots for the data sets by ploting between ACF, PACF verse lag to find the sationarity and trends. 

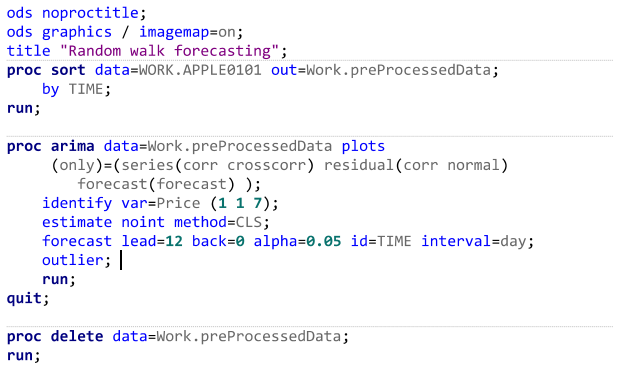
Correlation plot:



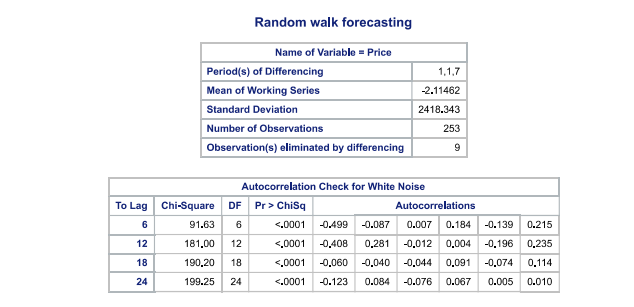
FORCASTING MODELS:

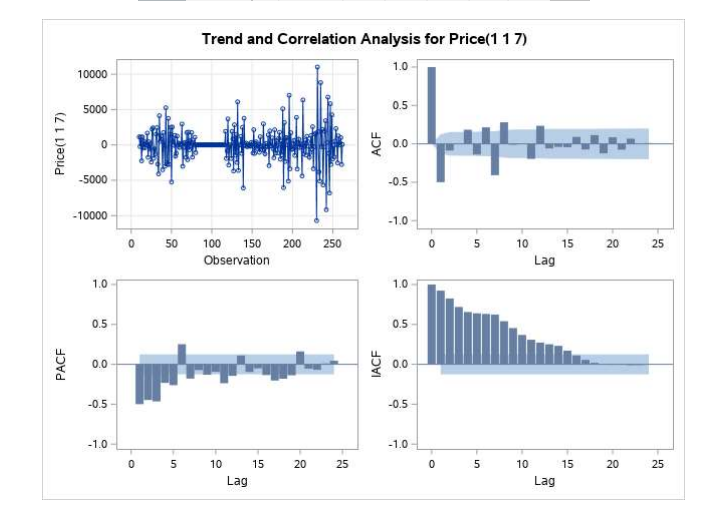
In this model, in the first step we find Trends and seasonal cycles in the data set. The main point is the data should be stationary with mean and variance, for the data set we provided the data set is non-seasonal which means it is stationary.

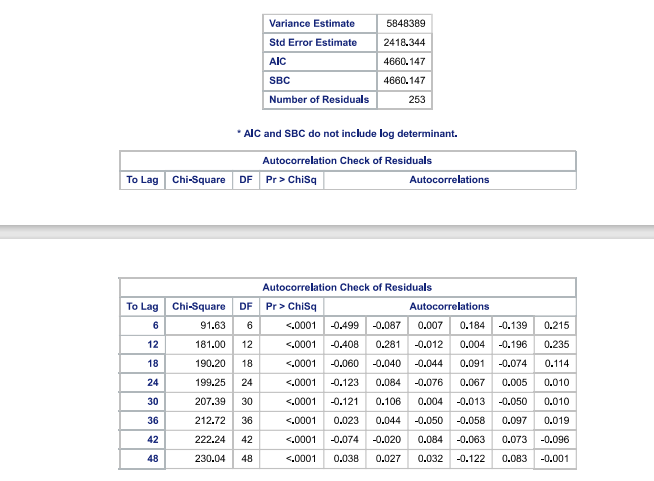
01.Random walk:

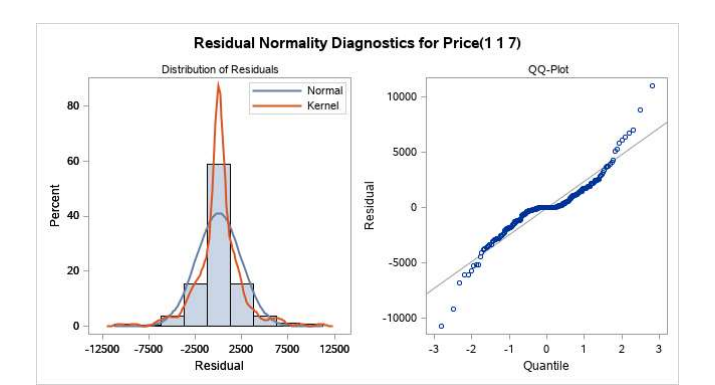
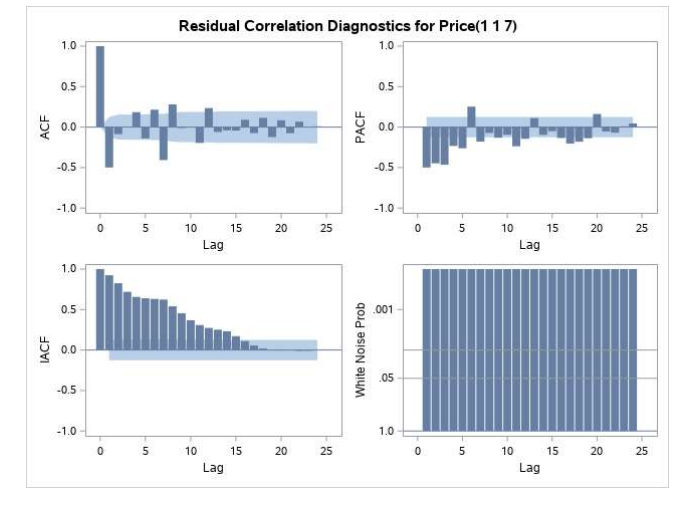
This model assumes that in each period the variable takes a random step away from its previous value, and the steps are independently and identically distributed in size. 

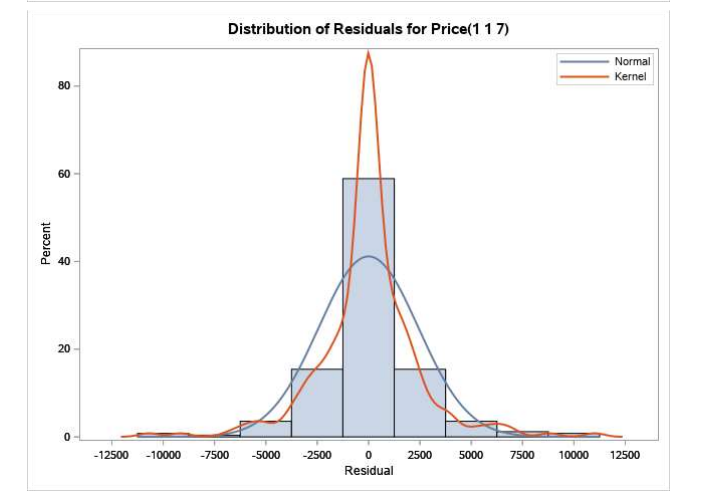
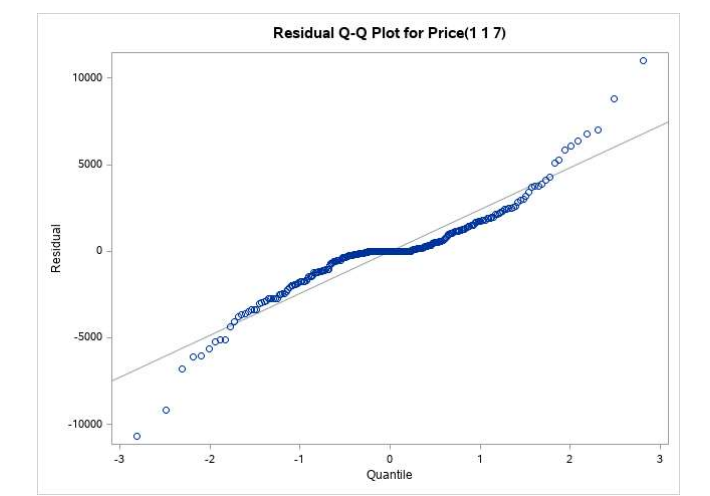
Output:

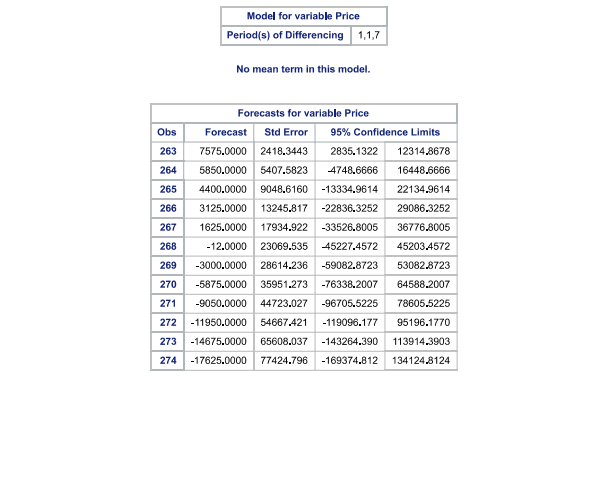


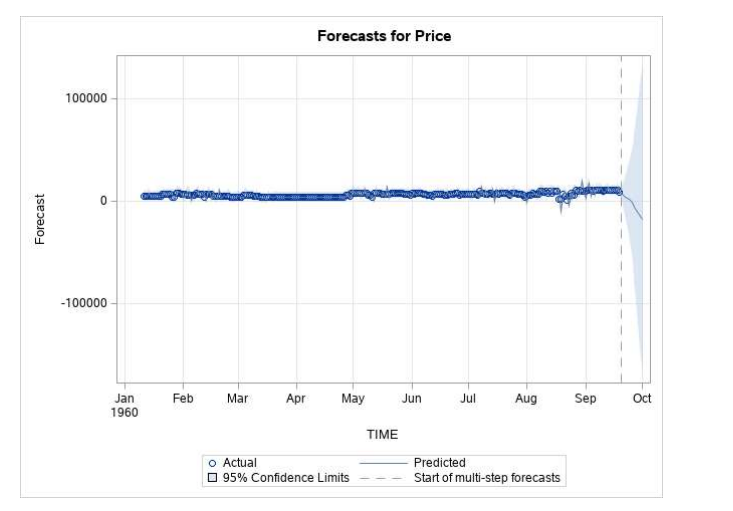


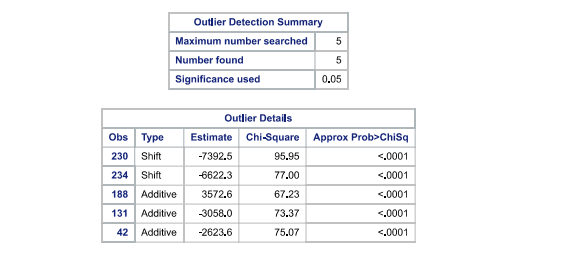




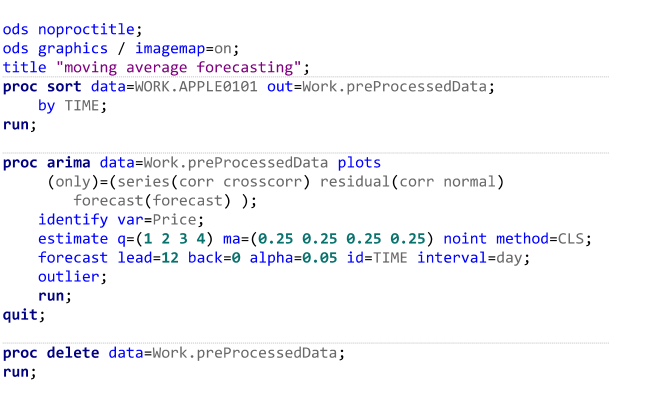




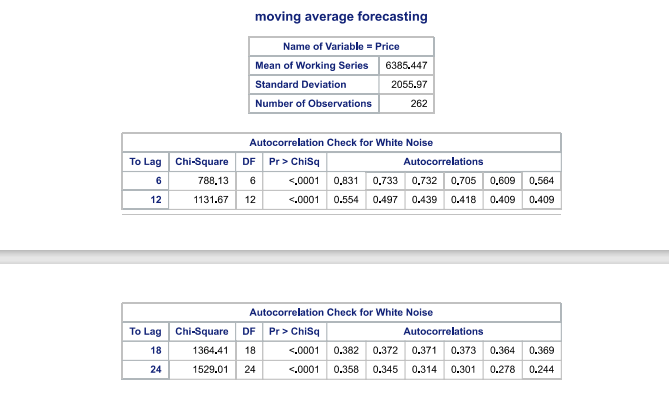


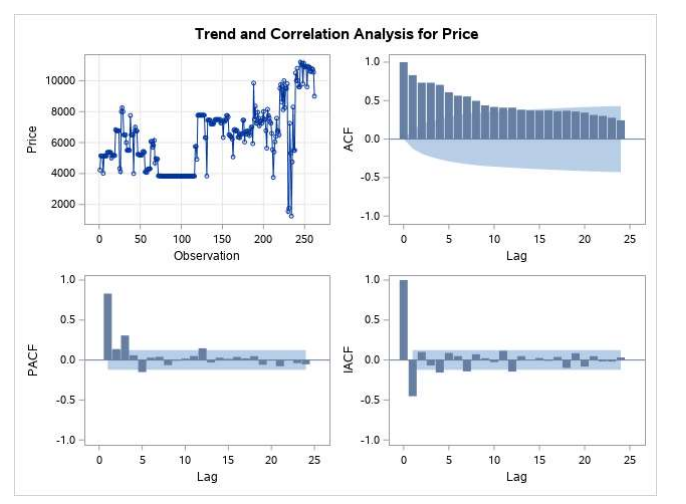
02. Moving average:

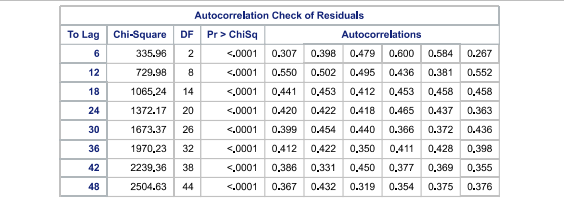
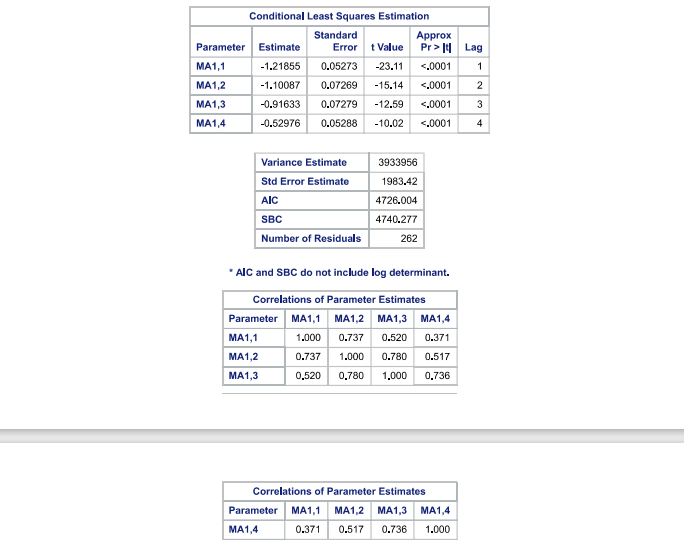
Code;

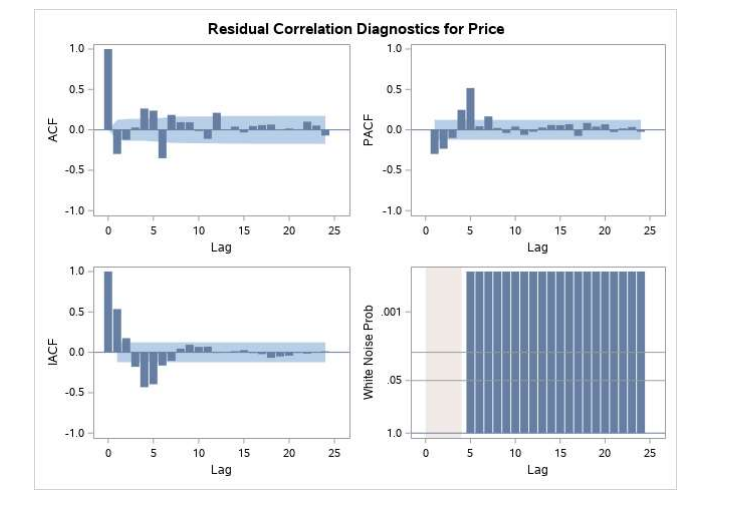


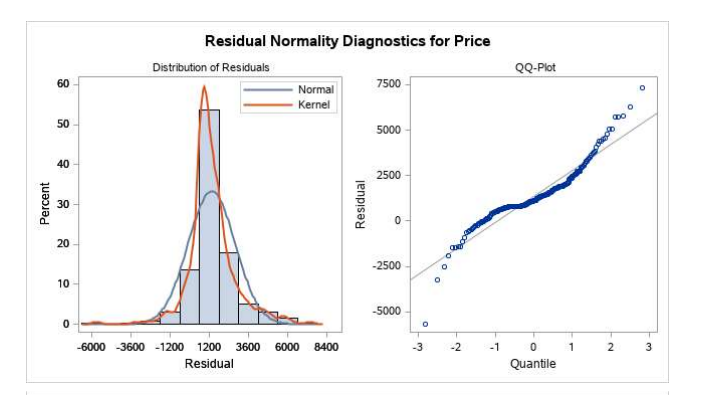
Result:

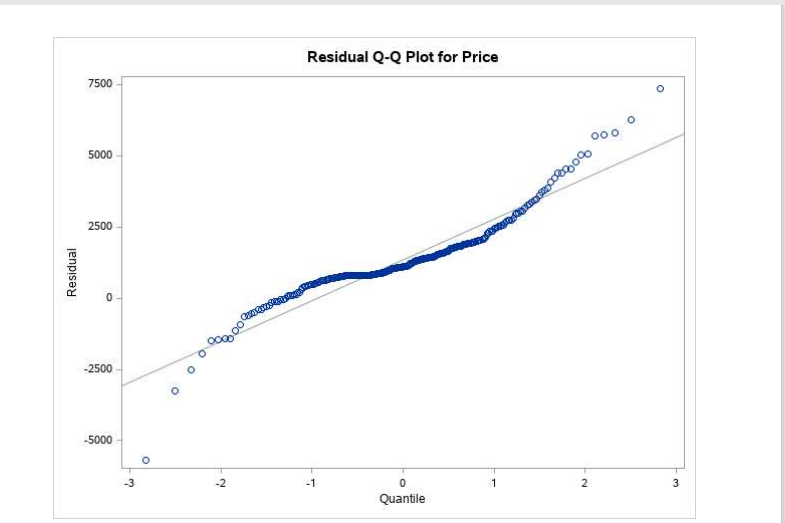


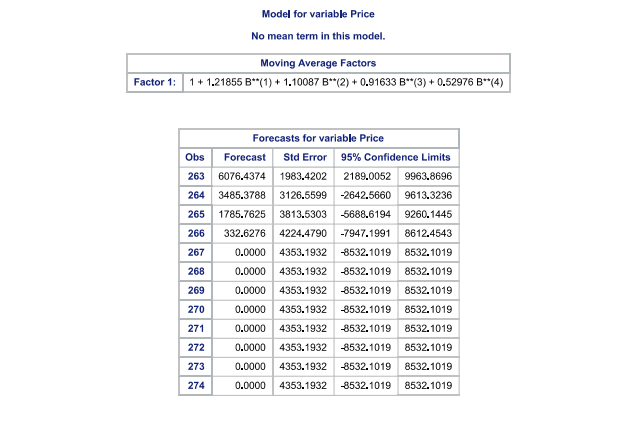


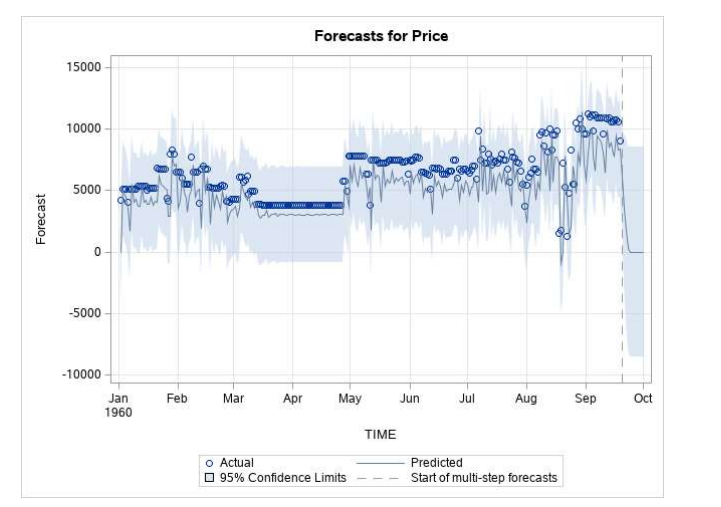


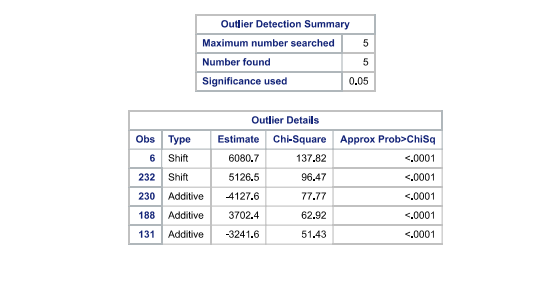






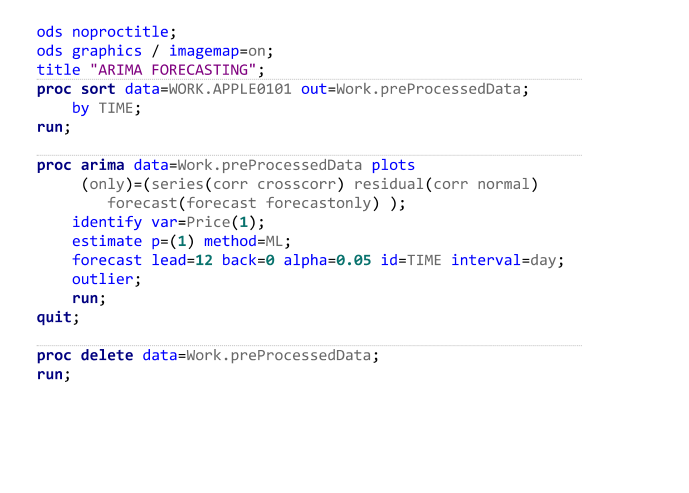




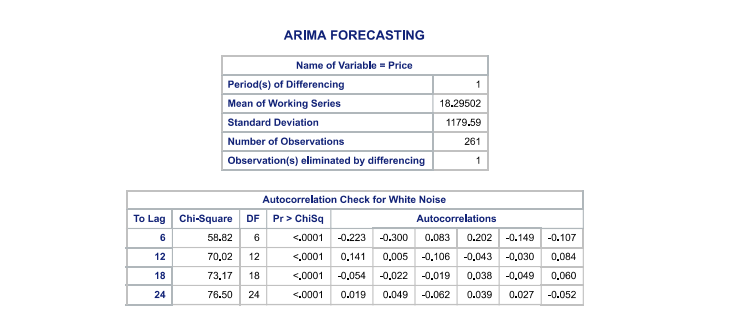


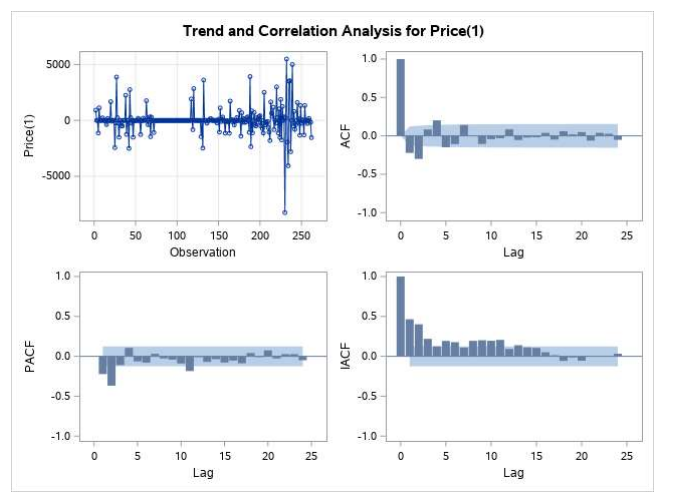
03.ARIMA MODEL:

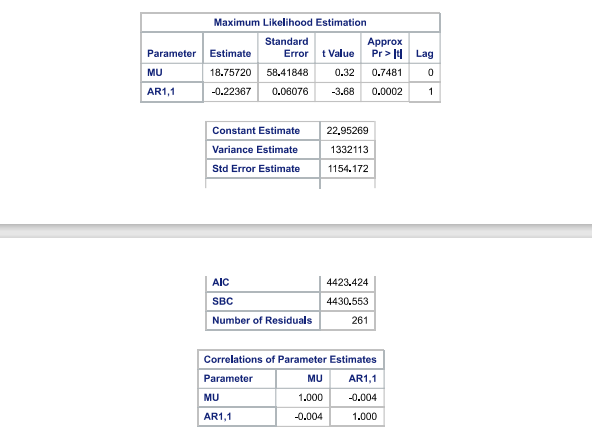
Code:

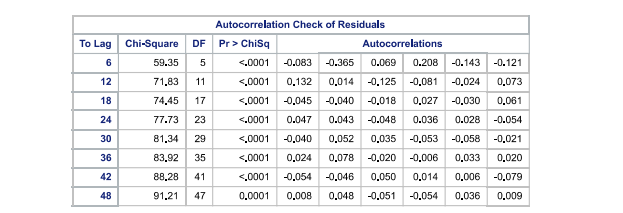


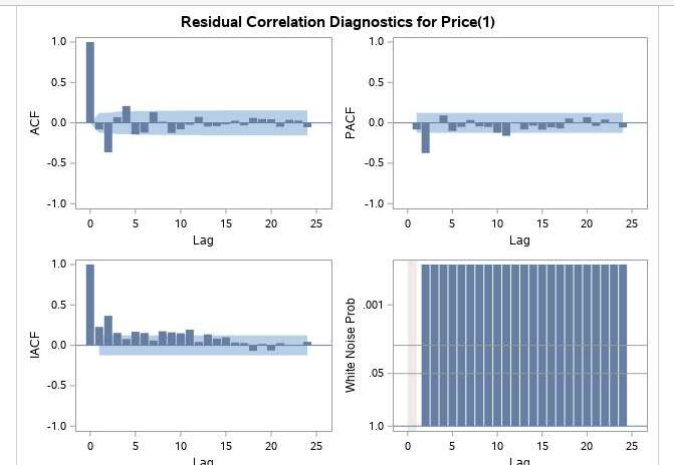
Output:

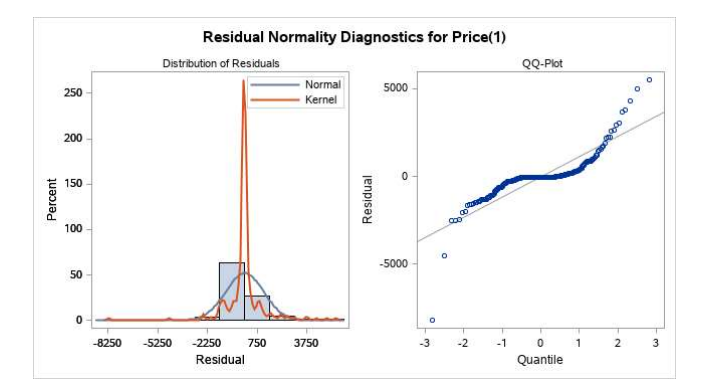


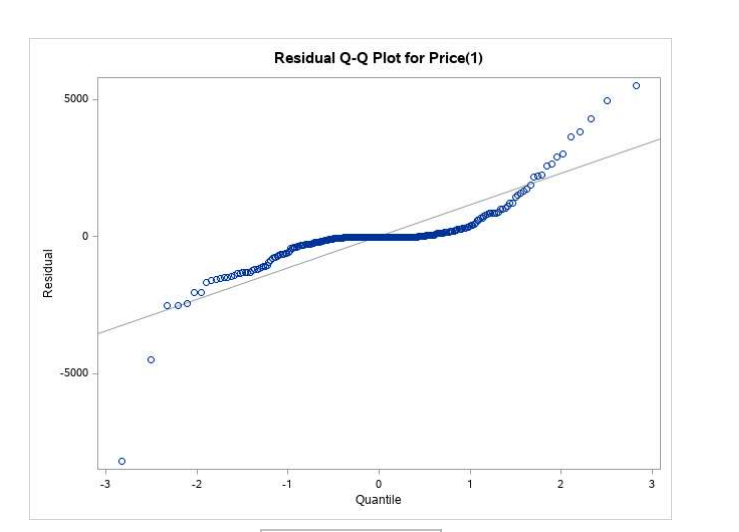


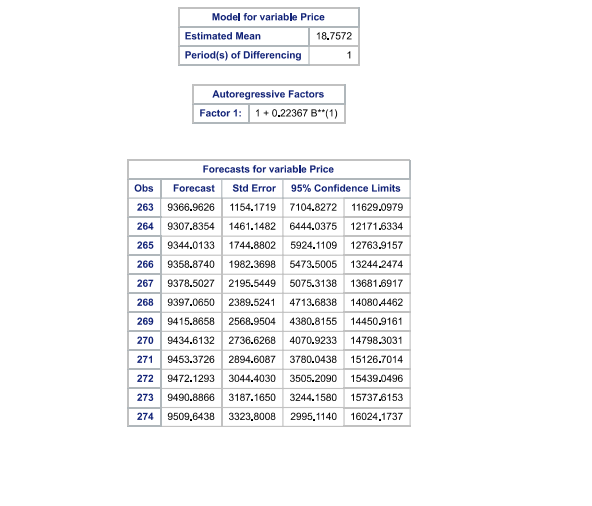




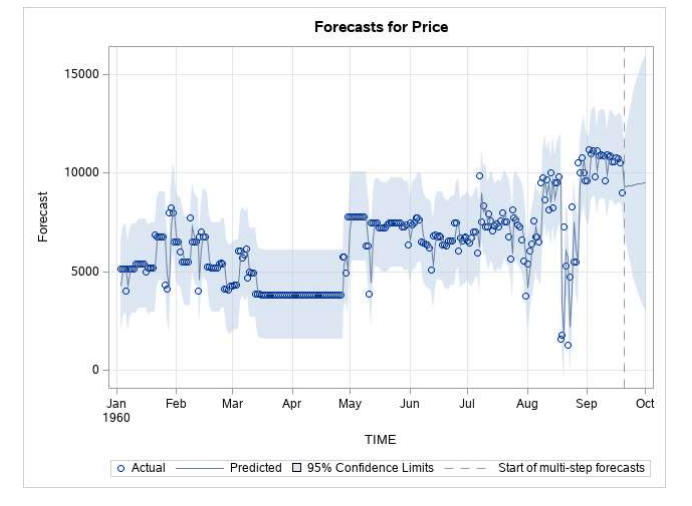


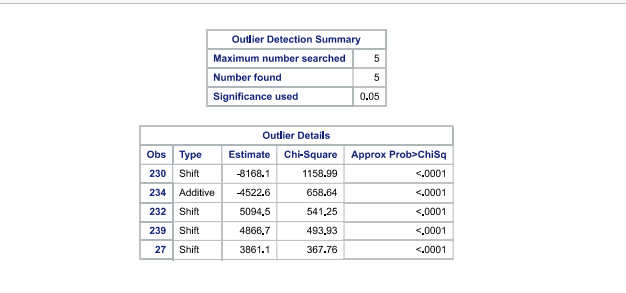












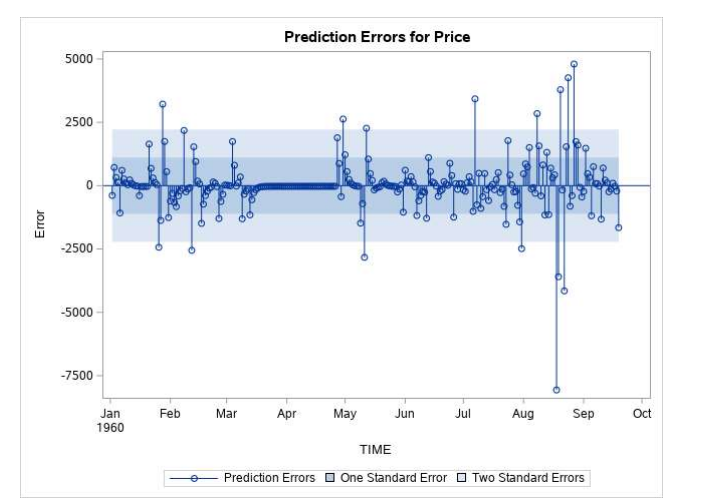
04.HOLT EXPONENTIAL SMOOTHING:

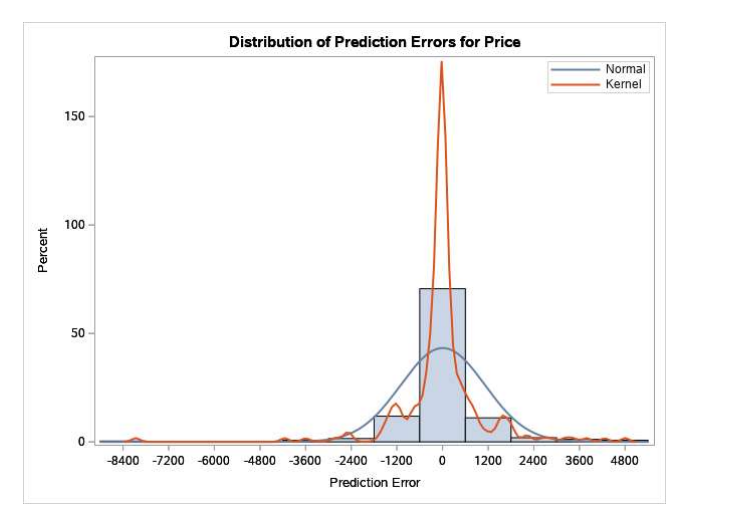
Code:

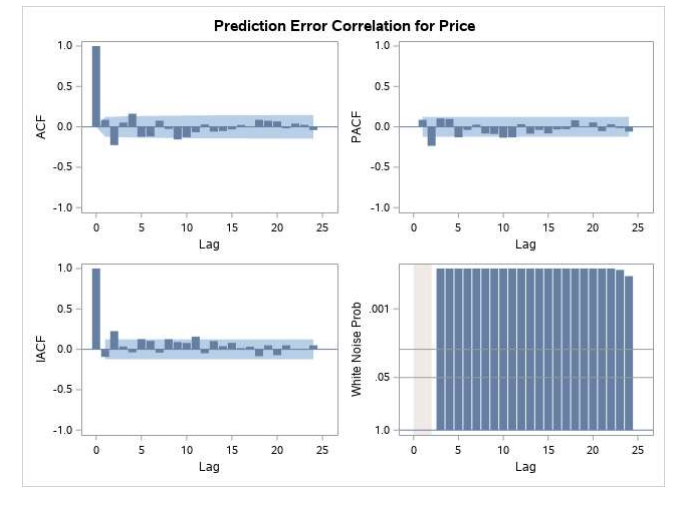


Result:

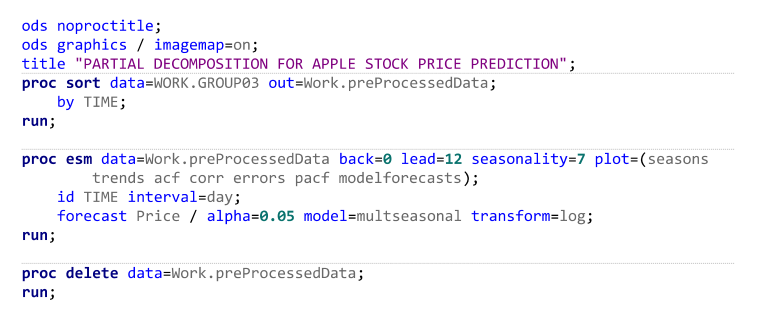




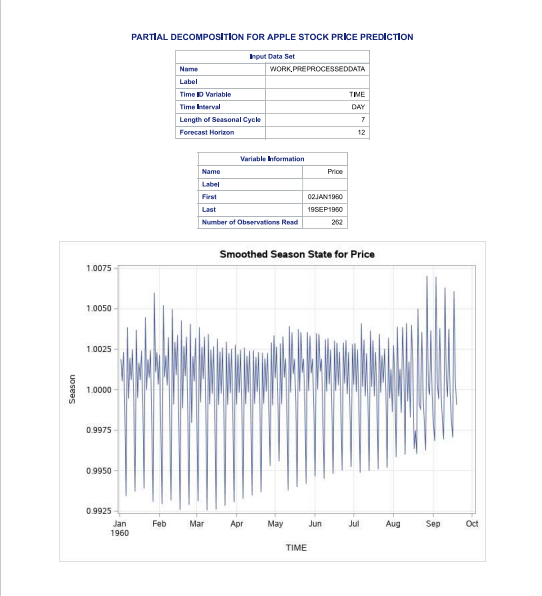


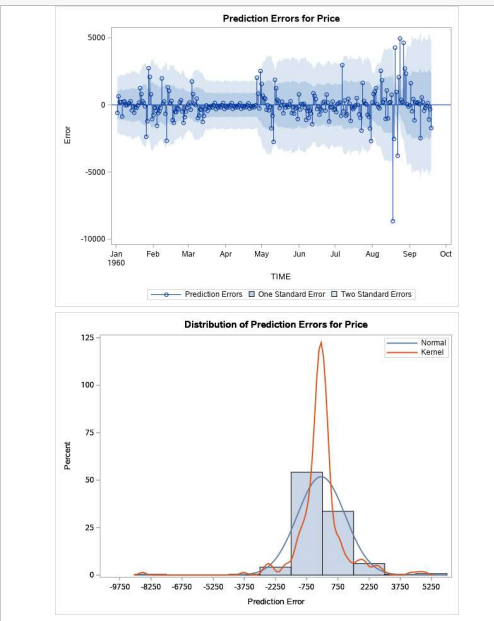


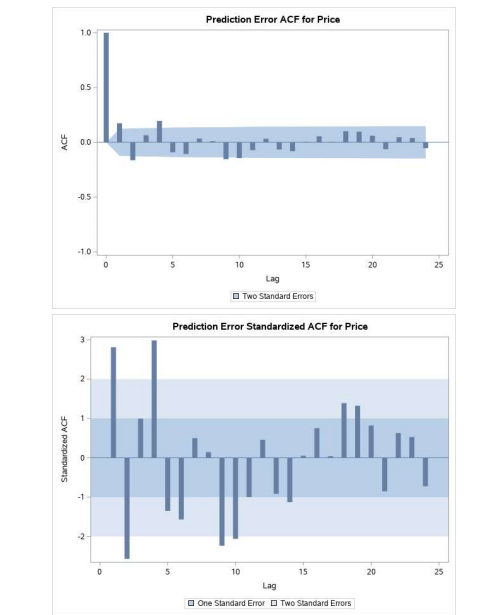
05. PARTIAL DECOMPOSITION:

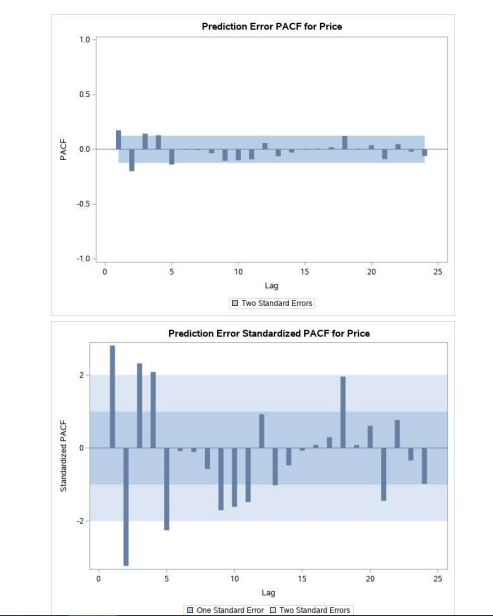


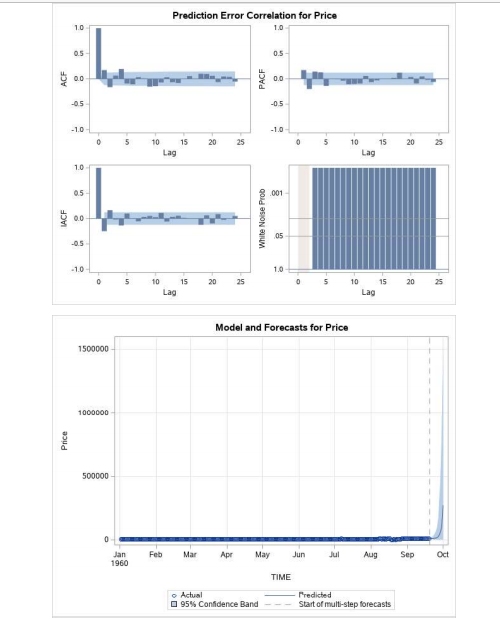
RESULT:





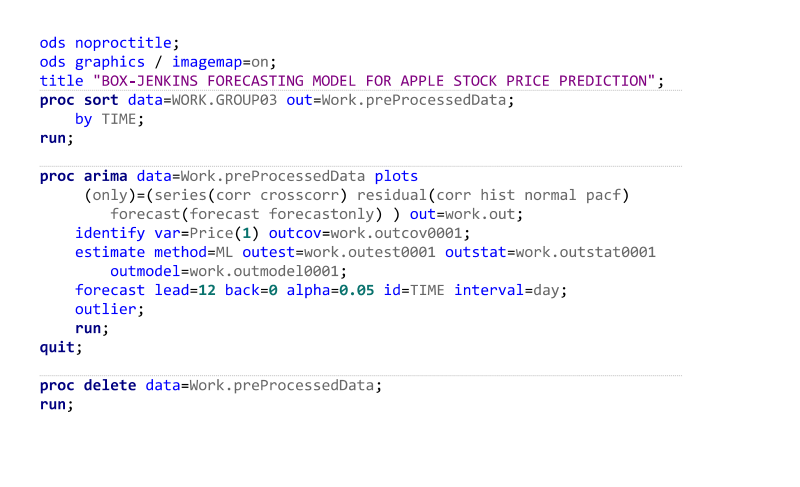




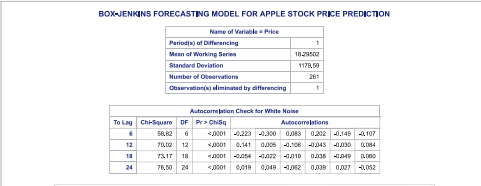


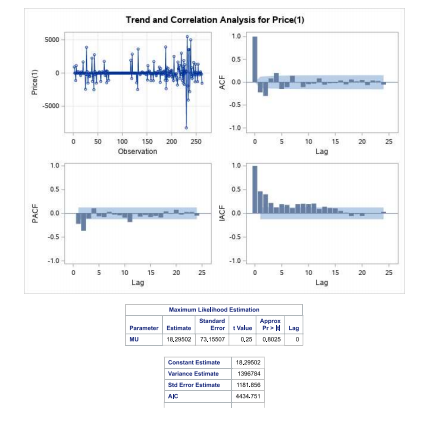
06. BOX-JENKINS MODEL:

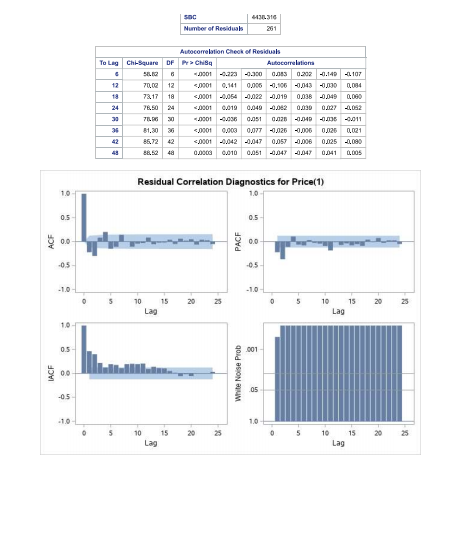
CODE:

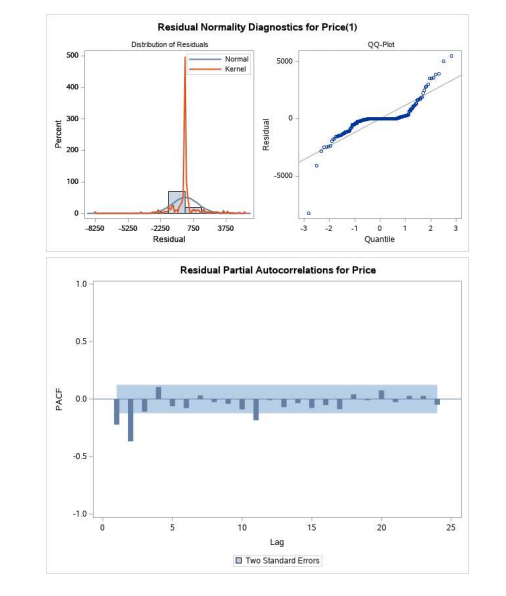


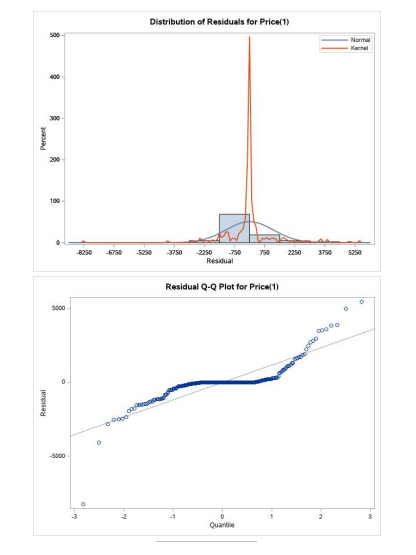
RESULTS:

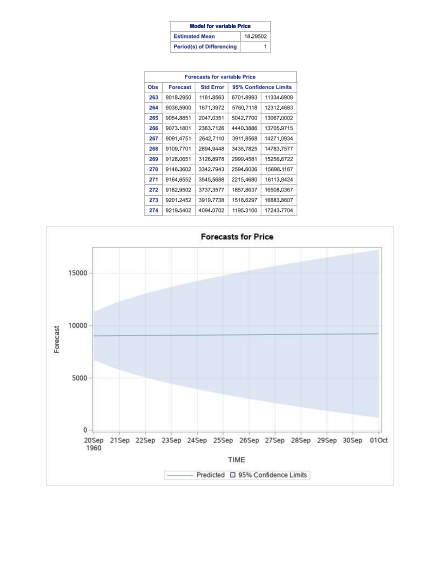


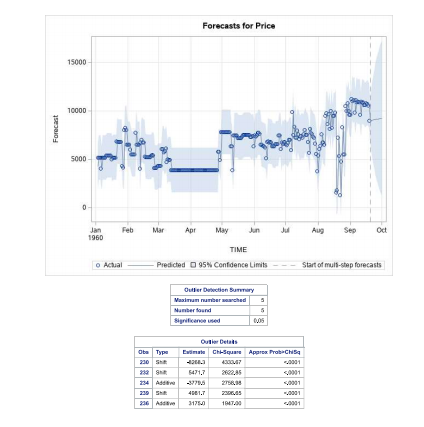












**6.CONCLUSION AND IMPLICATIONS FOR FUTURE**

On the basis of t value and p value of maximum likelihood estimation, standard error, AIC and SBC we can say ARIMA model is best suitable for forecasting and we don’t reject the null hypothesis. A seasonal data would have been better for understanding the concepts behind the predictive analytics in time series, further data collection would help in determining the nature of series.

**7.REFERENCES**

<https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf>

<https://www.genasis.cz/time-series/index.php?pg=home--trend-analysis>

<https://otexts.com/fpp3/index.html>

<https://en.wikipedia.org/wiki/Forecasting>

<https://zhuanlan.zhihu.com/p/372590298>