Fruit and vegetable price prediction using Auto regression model

Yogesh Thangamuthu - 210270974

Jan - 2023

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Data Analytics

Supervisor - Dr. Philip Trevelyan

Abstract

The COVID-19 pandemic has had a significant impact on the production and supply of fruits and vegetables, as well as on the prices of these and other goods. The pandemic has messed up global supply chains and the way markets work, which has caused shortages and higher prices for some goods. Prices have also been affected by the fact that people are stockpiling certain goods and buying more goods online because of the pandemic. Governments and organizations all over the world have taken different steps to try to lessen the effects of the pandemic on the economy and make sure that essential goods, like fruits and vegetables, are available to consumers at reasonable prices. The usefulness of forecasting the price of fruits and vegetables is important. This motivates the need to determine the most suitable model to predict the weekly price of fruits and vegetables at the wholesale markets in Birmingham, Bristol, Manchester, and London (New Spital Fields and Western International, respectively). This dissertation gives a comparative study of seasonal and nonseasonal models. Auto Regressive Integrated Moving Average is a non-seasonal model, and Seasonal Auto Regressive Integrated Moving Average is a seasonal model in which the function and results of the system are compared.

January 6, 2023

Contents

1 Introduction

1.1 Background

The pandemic has created catastrophic disruptions in social and economic systems all across the world, including the world's longest and most severe recession since the Great Depression. This recession has been triggered by the lockdown to prevent the spread of disease. A movement in the people's concern regarding the diet plan may be traced back to the beginning of the two or three years that have passed since the outbreak. Because of this significant change in the diets of people all over the world, there is a greater need for fruit and vegetables, which drives up the demand in the market. As a result, the price of fruit and vegetables has increased. Because of how the system works, the prices of fresh fruits and vegetables can go up or down. Disruptions in the supply chain were the main reason for widespread supply shortages, the most important of which was a lack of food. This disruption also has a larger impact on the cost of fruits and vegetables, and while the pandemic is still going on, there has been some fluctuation in the prices of fruits and vegetables.

A time series is a collection of data that shows how the values of a variable have changed over a specific period of time. This data is part of a time series. Instead of being in a straight line, they are influenced by the timestamps. The fundamental purpose of a time series model is to conduct an analysis of the time series data set in order to create projections on the future values of the time series for the period that is still to come in the future. Over the course of history, a great number of different time series models have been developed in order to enhance the effectiveness and precision of time series forecasting. [Abewickrama, 2022]ARIMA, which stands for Auto Regressive Integrated Moving Algorithm, is one of the models of time series forecasting that is recognized and utilized by the broadest range of individuals. When it comes to forecasting time series, ARIMA is both the simplest and most accurate predictive model. SARIMA, which stands for Seasonal Auto-Regressive Integrated Moving Algorithm, is yet another time series forecasting method that has gained widespread popularity and recognition. The SARIMA model is constructed by taking into account the cyclical patterns of the data. A variability in pricing happens after some predetermined amount of time has passed is the seasonality of the data. Forecasting the price of fruits and vegetables has a rolling mean and variances from time to time, so the prices are not linear, and to analyze and predict the non stationary data, we used a non-stationary predictive model to predict the future. The non stationary predictive models we used are Auto Regressive Integrated Moving Object (ARIMA) and Seasonal Auto Regressive Integrated Moving Object (SARIMA)[**?**].

1.2 Dissertation objective:

The consumption of food is a fundamental component of human existence, and it is also an important factor in maintaining good health and warding off illness. Many people believe that fruits and vegetables, in particular, are among the best places to have their vitamin and mineral needs met.

In general, the prices of fresh fruits and vegetables fluctuate on a regular basis, either in accordance with a predefined pattern or in a way that is not compatible with that pattern. It is possible for this to happen in either direction. The price of purchasing fresh fruits and vegetables will be analyzed, and price forecasts will be generated based on the findings of the inquiry that will be included in this dissertation paper. An analysis of the data that has been time stamped will be carried out with the assistance of the ARIMA and SARIMA algorithms in the very first stage of this procedure. the second step is to utilise the ARIMA and SARIMA forecasting algorithms in order to formulate an estimate of the cost of locally grown fruits and vegetables.

1.3 Gaps in current studies

1.4 Data source

- The weekly vegetables and fruits in United Kingdoms wholesale market price are retrieved from government website dataset
- The data set includes an average price by wholesaler in the Birmingham, Bristol, Manchester and London markets.
- The dataset used has a combined price for four categories of fruits and vegetables. The enclosed data set contains 54 varieties of fruits and vegetables.

1.5 literature review

In recent years, fluctuations in the market price of fruits and vegetables have had a significant impact on economies all over the world. There have been a few studies done to anticipate and analyze the price volatility of fresh fruits and vegetables. Research was conducted to investigate the effect that the COVID-19 pandemic had on the costs of basic food products. The study found that these prices rose by a sizable amount during the pandemic compared to the time before the epidemic began. An unanticipated tragedy can have a differential effect on the prices of critical food items in a nation that is still growing. To add insult to injury, the COVID-19 epidemic caused a surge in the cost of major dietary staples on the market. [**?**] The price of fruits and vegetables may be determined through the use of a model that employs image classification and linear regression. Image processing techniques are used to categorize the kinds of fruits and vegetables, and linear regression is used to forecast the prices of image-categorized fruits and vegetables. The results of this study contribute to the linear rise in the price of fresh fruits and vegetables. Recently, the prices have been undergoing nonlinear transformations, which conflicts with the accuracy of prediction models.[Wijekoon et al., 2021] In yet another piece of research, the cost of fresh produce and its relation to time stamp data is investigated. A time series model is used to make comparisons with these timestamps. In the process of crop price forecasting, a comparison analysis of past market prices is performed. Autoregressive integrated moving average (ARIMA), the partial least square (PLS), and the artificial neural network (ANN).

When there are more fruits and vegetables involved, it gets more difficult to prepare them all. The obstacle posed by time-stamped data sets can be overcome by using partial least squares and an artificial neural network.[Peng et al., 2015] Another study establishes a connection between the price prediction and the Long Short-Term Memory (LSTM) model. The LSTM model is designed to help people remember information for extended periods of time. a method for forecasting stock prices in frontier markets by using the LSTM network model, with the "training parameters," "number of prior data," and "time period" of the model being the only variables that need to be modified in order to get the desired results. In this comparison of two LSTM models (LSTM-0 and LSTM-1), it is found that the LSTM-0 model has a higher level of accuracy than the other LSTM model. [Rahman et al., 2021] An investigation of the impact that the predictive capacity of the SARIMA model has on the cost of vegetables in India. The model takes into account the seasonal changes that occur in the cost of fresh fruits and vegetables. The seasonal backshift in data establishes a comparative link between the most recent price value in the data and the value that was backshifted in time. The model is accurate when it comes to predicting the price over an extended period of time. [Dharavath and Khosla, 2019]

1.6 Dissertation outline

This analysis and prediction of fruit and vegetable prices is split into 4 chapters. Chapter 1 gives the fundamental literature, motivation for the study, and methods to perform this analysis. The second chapter expands on the insights and methodology for analyzing fruit and vegetable prices. In chapter 3, we discussed the findings. Chapter 4 summarizes the study and discusses the policy implications, the limitations, and the implications for future research. In the following chapters, these models are explained and fitted.

2 Methodology

2.1 Data understanding

This analysis of prices makes use of data that is collected on a weekly basis beginning in November 2017 and continuing all the way through to October 2022. The information is arranged in two distinct groups: training and testing. The training data spans the years November 2017 to June 2022, while the testing data spans the years June 2022 to October 2022. The precision of the model is evaluated with the use of the mean absolute error, Mean absolute percentage error Scores for mean absolute scaled error, root-mean-square deviation, and using the concept of forecasting, we can speculate on what the cost of fresh fruits and vegetables will be over the course of the following two months.

2.1.1 Sampling model

The main problem with the analysis method is that the data set has both values that aren't there and values that are outside of what they should be. When examining a data collection, it is usual for problems to arise, such as missing numbers and values that are outside of their boundaries. The missing value could be dependent on or independent of the model. It is recommended that the values that are missing or out of bounds be eliminated from the data set since this will bias the system. Every week, the prices of different fruits and vegetables are written down and added to the data collection. Each entry also has a time stamp.

Variable name	description
category	fruit, vegetable, cutflowers, potplants
item	55 kind of fruits and vegetables
variety	77 kind of variant of fruits and vegetables in 55 items
date	weekly price stamped date
price	weekly price of fruits and vegetables
unit	quantity of the product

2.1.2 Features of time series analysing:

2.2 Data preparation

In this stage, the dataset was processed to fit into the statistical and analytic model. The data set is in csv (comma-separated values) format. The dataset contains missing values which are usually denoted by NA. For any kind of machine learning modeling, NA values are not suitable or feasible for the model. There are some NA values in the price column that are not suitable for the regression model. To regularize the data, the price of fruits and vegetables is formulated into a standard unit, so that the price of all the varieties of fruits and vegetables is listed in kilograms.

2.3 Modelling

2.3.1 Forecasting using Linear Regression

A straightforward prediction model that is constructed based on the relationship that exists between a dependent variable and an independent variable is called linear regression. This model produces a single line that is the best match for the whole collection of values in the data. The $x - axis$ value represents the independent variable, and the value of the dependent variable is determined by the calculation based on the independent variable. The linear regression equation is an universal equation that can be applied to any straight line, and it consists of two variables: m, which represents the gradient of the line (how steep the line is), and c, which represents the $y - intercept$ (the point in which the line crosses the$y - axis$).

$$y = mx + C$$

where

- y is predited value(dependant variable).
- m is the gradient.
- x is value for independent variable.
- C is the intercept.

2.3.2 Forecasting using ARIMA model

2.3.3 Data preparation for analysis

- All of the products' units are converted to the standard kilogram format.
- labeling the column of items and variety in to numerical format
- Making the date and time an index and removing all the other columns except price data
- Normalizing the data, the lowest value is 0 and the highest value is 1.

Auto-regressive Integrated moving algorithm models are generally used for forecasting and analyzing time series data sets. The data set for the ARIMA model relates to the previous data set to predict the future data set.[McDonald et al., 2013]

- AR:Auto Regressive: allows you to use the future values to predict the present value.
- I:Integrated:difference from current to previous time stamp.
- MA:Moving Average:change in price which are updated using this variable.

It is feasible, through the utilisation of functional analysis, to unearth the concealed patterns and correlations contained within the data, which, in turn, makes it possible to recognise notable occurrences as they take place.The average value, trend, seasonality, and residual are the four key components that may be extracted from any data set that forms a time series. They are universal, appearing in all time periods. Time series data may or may not contain observable trends or seasonal patterns.[Kamil and Razali, 2015]

"Trend analysis" refers to the process of identifying whether or not there is steady movement in a particular direction over time. There are two sorts of trends: deterministic, in which the underlying rationale can be determined, and stochastic, in which the outcomes are random and cannot be predicted. In deterministic trends, the underlying explanation can be determined. In the actual world, we can see examples of both of these types of trends.

These shifts are referred to as "seasonal variation," which is also the name of the term that is used to describe them since they take place during the course of a year at predictable intervals throughout the course of the year. Things vary with time, such as a year, month, week, or day. Serial dependence is the term given to the phenomenon that takes place when data points that are spaced out over a relatively short period of time have a tendency to be coupled with one another. Error, residual/irregular activity that cannot be explained by trend or seasonal value. The model type parameter might be additive or multiplicative, depending on whether the seasonality of your data is level- mean-dependent. If the amplitude of the seasonality is independent of the level, use the additive model; if the amplitude of the seasonality is dependent on the level, use the multiplicative model.

Additive model

$$A = T + S + I$$

- A Additive model
- T is trend
- S is seasonal
- I is irregular

Additive model is the sum of trend ,seasonal and irregular function.The time series is made up of its constituent parts when put together. When there is an upward trend, the amplitude of the peaks and troughs over the time series are, for the most part, equivalent to one another. When the absolute value rises yet the changes continue to be relative, as seen above, this occurs frequently with indexed data sets.

multiplicative model

$$M = T \times S \times I$$

- M is Multiplicative model
- T is trend

- S is seasonal
- I is irregular

Multiplicative model is the product of trend,seasonal and irregular function.The time series is created by multiplying the components. If the trend is ascending, seasonal activity will increase in magnitude. This is advantageous as seasonal variance grows over time.

2.3.4 Validating the data

Auto regression is a model which uses non stationary data to predict the future.Two different techniques are used to evaluate the property of the data. The techniques are

- rolling mean and variance graph
- Augmented Dickey Fuller Test

Rolling mean and variance Rolling mean is an collection of series of average calculated in the data set.The first average value is is obtained by taking the average of first data with the fixed number subset,rest of the values are obtained by shifting forward the data set values.The value is calculated on taking the average of the previous data value with next data value.This is the continuous value update in the data set

$$RA = \frac{x_{i-1} + x_{i+1}}{2}$$

where
RA is Rolling Average
x_{i-1} is previous value of current data
x_{i+1} is the next value of current data.

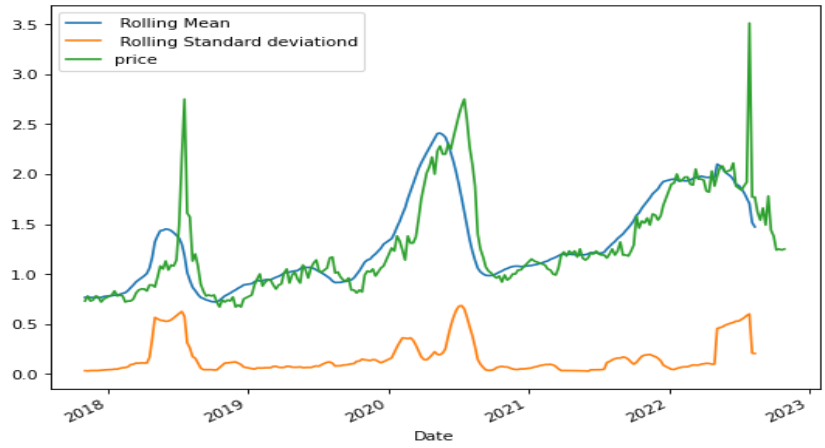

Figure 1: Rolling mean and variance

The above Rolling mean and variance graph visualizes the spot trends and seasonal variation of the data set.

Augmented Dickey Fuller Test: ADFT is an commonly used statistical technique to determine whether the data is stationary or not.

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \gamma y_{t-1} + e_t$$

where

- α =coefficient of the first lag on Y
- y_t value of data at time t
- y_{t-1} value of data at time t-1.
- e_t exogenous variable

The hypothesis depends on the p The null hypothesis (H0) cannot be rejected since the p-value is greater than 0.05 and the data are non-stationary and have a unit root.A statistical hypothesis is known as a "null hypothesis," and it is a type that suggests that there is no statistical significance present in a particular set of facts. The reliability of a hypothesis can be evaluated through the process of hypothesis testing, which makes use of the original time series dataset. If the p-value is less than 0.05, the null hypothesis (H0) is unlikely to be true because the data are stationary and do not have a unit root.The unit root test is a type of test that determines whether or not a time series is stationary. Stationarity can be said to exist in a time series if moving forward or backward in time does not result in a change in the shape of the distribution. Non-stationarity can be caused by the presence of unit roots.

2.3.5 Parameters

ARIMA(p,d,q)

- P:Number of autoregressive terms
- d:Number of nonseasonal differences
- q:Number of lagged forecast errors in the prediction equation

These parameters of ARIMA model is labelled based on the kind of data in which they vary from data to data.In parallel with creating the ARIMA model ,the data for the model is to be as stationary data.As offen the time series data are in non stationary format(with varying mean),this data are converted to an stationary data by implementing the differencing technique.Differencing is performed by subtracting original time series data to the lag of same time series data.

$$y't = y_t - y_{t-l}$$

where:

- $y't$ is differenced data with time lag of t.
- y_t is the original time series data
- y_{t-l} is the laged time series data
- l is the lag of the differencing
- t is the time stamp of data

2.3.6 Estimating the values for parameter

- **Auto Correlation function:** The ACF, also known as the "auto correlation function," is a correlation function that links two different time series data sets. In this case, the first data set is the real data set, while the second data set is the lagged successive version of the time series data set. It is a representation of the similarity degree time series data and the lagged successive time series data. The auto correlation function is the variance between the covariance between data and lagged data to the multiplying of standard deviation of time series data with the standard deviation of h time lagged time series data set. This results in the auto correlation function.

$$ACF = \frac{\text{Covariance } (x_t, x_{t-h})}{\text{Std.Dev. } (x_t)\,\text{Std}\,.Dev.\,(x_{t-h})} = \frac{\text{Covariance } (x_t, x_{t-h})}{\text{Variance}\,(x_t)}$$

  - h is the h time lag for the data set.

- x_i is time series date set,
- x_{i-h} is the h time lagged time series data set.

Covariance: The direction of the relationship between the returns on two assets is measured by covariance. If the covariance is positive, the returns on the assets move together. If the covariance is negative, the returns move in the opposite direction.

Covariance is found by looking at at-return surprises (standard deviations from the expected return) or by multiplying the correlation between two random variables by the standard deviation of each variable. The auto correlation graph is modeled with lags are on $x - axis$ and auto correlation corresponds to the lags are on $yaxis$.

Positive Covariance A positive covariance between two variables indicates that they tend to increase or decrease together. Positive covariance between stocks one and two happens when both stocks are above average at the same times, and vice versa. When represented on a two-dimensional graph, the data points will generally have an upward slope.

Negative Covariance When the predicted covariance is above zero, the relationship between the two variables is reversed. In other words, a stock one value that is below average is typically associated with a stock two value that is above average, and vice versa.

.

$$\text{Covariance}(x_t, x_{t-h}) = E[(x_t - Ex_t)(x_{t-h} - Ex_{t-h})] = E[x_t x_{t-h}] - (Ex_t)(Ex_{t-h})$$

- h=lag difference
- x_t=data set with time
- x_{t-h}=data set with h time lagged data

standard deviation The relationship of the data to the set's mean is defined by the standard deviation. Data with a small standard deviation are clustered closely around the mean.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N - 1}}$$

- x_i is the i th component of dataest.
- μ is the mean of the data set.
- N is number of data points.

Partial Auto correletion Function

Similarly to the autocorrelation function, the partial autocorrelation function illustrates the association between two observations for shorter lags between those observations.

$$PA = \frac{\text{C}\ \left(\left[T_i \mid T_{(i-1)}, T_{(i-2)\ldots T_{(i-k+1)}}\right]\ \left[T_{(i-k)} \mid T_{(i-1)}, T_{(i-2)} \ldots T_{(i-k+1)}\right]\right)}{\sigma_{\left[T_i | T_{(i-1)}, T_{(i-2)} \ldots T_{(i-k+1)}\right]} \times \sigma_{\left[T_{(i-k)} | T_{(i-1)}, T_{(i-2)} \ldots T_{(i-k+1)}\right]})}$$

- PA is Partial Auto Correlation.
- C is covariance
- σ is standard deviation.

- **d-value** The technique of differencing is used to convert a non-stationary time series to a stationary time series. The difference in each consecutive data set distinguishes this type of analysis from others.

  The variance of a time series can be stabilized with the use of transformation techniques, for example, logarithms. The process of differencing can help stabilize the mean of a time series by getting rid of changes in the level of the time series. This gets rid of (or at least greatly reduces) trend and seasonality.

- **p-value** When determining the p value with the help of partial autocorrelation, a plot is generated with the help of the time series data set. A connection is established by the partial autocorrelation function between the initial time series data and the lag-based time series data. The graph is modeled with the log values of time series data and the number of logs in the logged time series. Calculations are done to determine the total average value of the lag. The p value is determined by looking at the first correlation value that is closer to the overall average value of the lag boundary.

$$y_t = a_0 + \sum_{n=1}^{p} a_n y_{t-n} + \epsilon_t$$

  - ϵ_t=white noise process
  - a_0 and a_n=estimated values.

  values for the P

  - white noise:value chosen at random from a normal distribution that has a mean of zero and a variance of one.
  - P=0 : Then there is no auto regressive term,So the time series is an white noise.
  - P=1: Means previous timestamp is adjusted by multiplier with white noise is added.
  - Other p values:Means adding more time stamp by own multiplier.

- **q-value** The non stationary data is converted to stationary-data by using differencing method .Auto autocorrelation function technique is implemented in the stationary data (differenced time series data set). The total outbound values in the autocorrelation are counted and that value is represented as the q value.

  Akaike's Information Criterion (AIC) The Akaike Information Criterion is a metric that is used in the process of establishing how accurate a model is. It is a measuring instrument

that is utilised in this process. It does this by establishing a relationship between the actual quality of the created model and the data collecting that was done. It makes an estimate of the proportionate amount of information that is lost in the model by using the data set that you provide.

$$\text{AIC} = -2\log(L) + 2(p + q + k + con)$$

where

- AIC is Akaike's Information Criterion
- L is likelihood of the function
- p, q is the parameters in the ARIMA model
- k the penalty per parameter.
- con is the constant term with number of parameters in the model.

Likelihood The maximum likelihood estimating approach is a method that may be utilised to calculate the appropriate values to provide a model's parameters. This strategy was developed to maximise the possibility of obtaining accurate results. The purpose of this approach is to arrive at values for the model's parameters in such a way as to maximise the likelihood that the process represented by the model was responsible for producing the data that were actually observed. This goal can be expressed as finding values for the model's parameters in such a way as to maximise the likelihood that the process represented by the model.

2.3.7 Forecasting using SARIMA model

SARIMA is known as "Seasonal Auto Regressive Intergrated Moving Average." Like the ARIMA model, the SARIMA model was developed to address seasonal variation in time series model prediction. SARIMA models are made to override the seasonality parameter of the ARIMA model with the seasonal parameter, which deflects with the time series data set. This forecasting model is defined for the data set that has a seasonal trend or a deflection in values that Proportion to the change in time period. A repeated or patterned change of data in the time series data set is called seasonality. Seasonality pattern of the time series data are in time of day, day of the month, week of the year, month of the year, and year of the period of years. An backward shifting technique is used to shift the data in backward of forward direction. The model is designed with the parameter of (p,d,q)(P,D,Q)s in which the [Chang and Liao, 2010]

- p is Non seasonal order of Auto regression
- d is Non seasonal differencing

- q is non seasonal order of moving average
- P is Seasonal order of Auto regression
- D is Seasonal differencing
- Q is non seasonal order of Autoregressive

$$\Phi\left(B^{s}\right)\varphi(B)\left(1-B^{s}\right)^{D}\left(1-B\right)^{d}Y_{t}=\ Theta\left(B^{s}\right)theta(B)\varepsilon_{t}$$

- Φ is seasonal Auto regression
- Θ is seasonal Moving Average
- φ is Non seasonal auto regression
- θ is Non seasonal Moving Average
- B is Backward shift
- $(1-B^{s})^{D}$ is the seasonal difference
- $(1-B)^{d}$ non seasonal difference

Backward shifting Backward shifting is a technique used to shift the data over a period of time. Backward shifting works as an lag function with the time series data.

$$By_{t}=y_{tT}$$

where

- By_t is the backward shifted data
- y_{tT} is shifting of data with time period
- T is the shifting time

Non Seasonal equation p,d,q is a non seasonal values, which are calculated using the autocorrelation and partial autocorrelation functions.

Seasonal equation P,D,Q are a seasonal values ,which are calculated using auto correlation and partial auto correlation functions.The seasonal order (s) is the backward shift of the original time series data set.The backward shift value depends on the time stamp value of recorded time series data set.The s value is 24 for hour of day ,7 for days of the week,12 for months of the year and 52 for weeks in the year.

Standard residual The ratio of the count that was actually taken to the count that was predicted is known as a standardised residual.The standard deviation of the expected count in chi-square testing. **Histogram density plot** The explanation of the distribution of the residuals is presented through the use of the graphic. The measured distribution is depicted for us by the histogram; the smoothed version of this histogram is represented by the orange line, and the normal distribution is depicted by the green line in the histogram. If the model is accurate, then these two lines should be identical to one another. In this regard, there are marginal distinctions between them, which serve as evidence that our model is performing well. **Normal Q-Q plot** A comparison between the distribution of the residuals and the normal distribution is shown by the Q-Q plot. If the residuals follow a normal distribution, then all of the points should be located along the red line, with the exception of a few values at the end of the graph. **Correlogram plot** The ACF plot of the residuals is what the correlogram plot represents rather than the data itself. There should not be a meaningful relationship between lags higher than zero and 95% of the correlations (within the blue shades). It indicates that there is information in the data that was not captured by the model if there is a substantial correlation in the residuals. **Seasonal differencing** The difference between the original time series data and the back-shifted value of the time series data is what we mean when we talk about seasonal differencing.The backward shifting is done with the multiple of season s.

$(1 - B^s)\, x_t = x_t - x_{t-s}$

2.3.8 Model validation

Mean Absolute Error Mean absolute error is a method that is utilised in the process of determining how accurate or effective a model is. The formula for calculating the mean absolute error is the sum of the differences between the original value and the projected value, divided by the total number of data sets that were recorded. The observed value of the mean absolute error provides an overall deflection of value for the machine learning model that is being implemented.Mean absolute error neglect the negative deflection in the model,which gives an absolute mean value deflection of the predicted to the actual model.despite the fact that they are simple to compute, due to the scale dependency, they cannot be utilised in order to compare various series.[Dharavath and Khosla, 2019]

$$MAE = \frac{1}{n} \sum_{i=0}^{n-1} (|y_i - \hat{y}_i|)$$

where

- MAE is the Mean Absolute Error.
- y_i is actual data set.
- $\hat{y}_i$ is the predicted data set.

- n is the total number of records in the data set.

Mean Absolute Percentage Error Mean Absolute Percentage Error (also known as mean absolute percentage deviation, or MAPD) is a statistic that is utilised to determine how accurate a forecasting model is. MAPD validator is calculated with out regards in the sign of the value. The Mean Absolute Percentage Error is the fraction of the sum of the fraction of the difference between the actual value and the predicted value to the actual value that is divided by the total number of entries in the data collection. Due to the fact that they are not affected by scale, they may be utilised to compare several distinct series. However, you can't utilise them if there are any zeros in the data you're working with.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

where

- M is the mean absolute percentage deviation or Mean Absolute Percentage Error.
- y_i is actual data set.
- $\hat{y}_i$ is the predicted data set.
- n is the total number of records in the data set.

Naive forecasting The calculation for naive forecasting involves comparing the most recent value of the forecast to the result of the most recent time stamp.Due to the fact that it is defined by the ratio of mistakes in the forecast, MASE is unaffected by the magnitude of the error in the prediction. This indicates that MASE values will be comparable in the event that are anticipating highly valued time series. naïve forecasting may be broken down into seasonal and non-seasonal forecasting; Non seasonal naive forecasting focuses on the previous predicted data set. The mean absolute error is the proportion of the total of the differences between an actual data set and a previously predicted data set, as well as the number of data sets that have one fewer value.

$$MAE_{naive} = \frac{1}{n-1}\sum_{i=0}^{n-1}\left(|y_i - \hat{y}_{i-1}|\right)$$

where

- n is total number of data sets.
- y_i is actual data set value.
- $\hat{y}_{i-1}$ is predicted value of previous dataset.

The seasonal forecasting process involves establishing a link between the currently available data set and the seasonal data set from the prior season.

$$MAE_{naive} = \frac{1}{n-s} \sum_{i=s+1}^{n} (|y_i - \hat{y}_{i-s}|)$$

where

- n is number of elements in the data set.
- s is seasonal value for the the model.
- y_i is current value.
- $\hat{y}_{i-s}$ is seasonal value of the predicted data set.

Mean Absolute Scaled Error Mean Absolute Scaled Error is the fraction of naive mean absolute error to the mean absolute error.

$$MASE = \frac{MAE}{MAE_{naive}}$$

where

- $MASE$ is the Mean Absolute Scaled Error.
- MAE is the Mean Absolute Error.
- MAE_{naive} is the naive of mean absolute error.

Root mean square error The root mean square error is calculated by taking the root of the percentage of the sum of squared differences between the original value and the value that was predicted and dividing it by the total number of elements in the data set. In the event that we are anticipating highly valued time series, the RMSE values will be equivalent to one another. RMSE are appropriate for scalar variable data set .A set containing a constant number of data sets is referred to as a scalar variable data set[**?**].

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y_i} - y_i)^2}$$

where

- $RMSE$ is root mean square error.
- $\hat{y_i}$ is predicted value.

- y_i is the actual value.
- n is number of recorded data sets.

3 Result

3.1 Linear regression

Apple bramleys seedling The equation that represented the model was as follows: Price is labled in pence per kilogram = -0.00387547X + 1.77981204 time (weeks). This indicates that a rise in average costs is 1.77981 pence per kilogram occurs for every one unit increment in time (week). Furthermore, the findings of modified R-squared indicate that only 31.5% of the data can be explained by the model, indicating that the regression model is not the best match.

Tomatoes round The equation that represented the model was as follows: Price (Kg) = 0.0014744X + 0.934726642 time (weeks). This indicates that a rise in average costs is 0.934726642 pence per kilogram occurs for every one unit increment in time (week). Furthermore, the findings of modified R-squared indicate that only 5.1% of the data can be explained by the model, indicating that the regression model is not the best match.

Top washed carrot The equation that represented the model was as follows: Price are labled in pence per kilogram = 0.0017X + 0.813251 time (weeks). This indicates that a rise in average costs of kg 0.813251 occurs for every one unit increment in time (week). Furthermore, the findings of modified R-squared indicate that only 16.3% of the data can be explained by the model, indicating that the regression model is not a good match.

3.2 ARIMA model

This section show the predictive model of fruits and vegetables using Auto Regressive Integrated Moving Average.The price Apple Bramley,tomato round,cucumber original prices are analysed and predicted.

3.2.1 predicting the price of apples (varient is bramleys seedling)

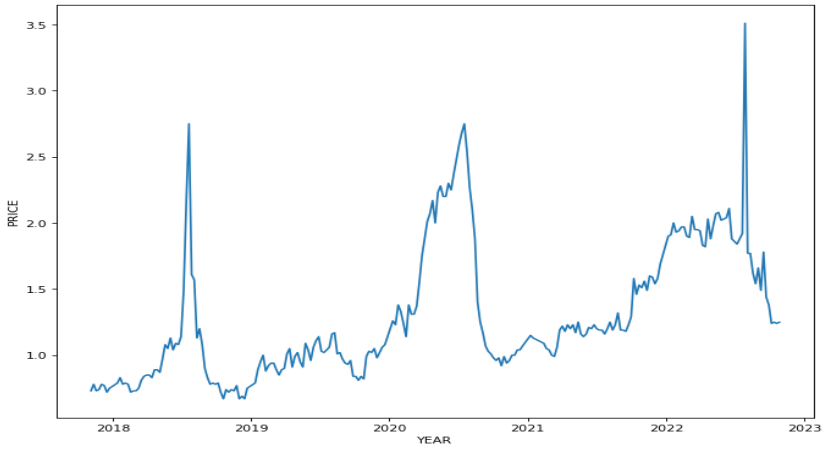

Figure 2: weekly price of apples bramleys seedling varient

During the time period that is depicted in the graph that is located above, there was a considerable shift in the price of fresh fruits and vegetables. The price of apples bramleys seedling is currently subject to a broad volatility throughout the market. This price volatility can be attributed to a number of variables, including seasonality, the level of demand, and the influence of nature. The middle of the year is often when apple bramley prices are at their greatest, while the beginning of the year and the end of the year are when they are at their lowest. When presented with such a variety of variables, the linear regression cannot be utilised to provide accurate forecasts. An examination of the data pertaining to apple bramley seedlings reveals that the price of the seedlings decreases after the middle of the year and gradually increases until it reaches the middle of the following year. This pattern repeats itself every year. The gathering of bramley apples begins in the month of December and continues all the way through the month of March. Although the price was considerably higher just before the beginning of the yielding period, it is currently at its all-time low and will continue to fall during the duration of the yielding period.

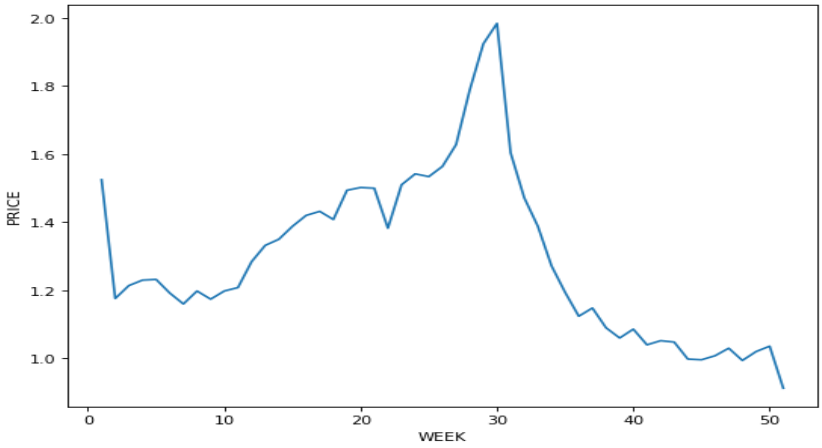

Figure 3: Average weekly price from 2015 to 2022

Data framing Within the data collection that was utilised for the purpose of conducting this study, weekly reports on the costs of fresh fruits and vegetables were included. The datasets have been reorganised into a descending order, with the earlier dates showing at the top of the list and the most recent or most recent dates appearing at the bottom of the list, respectively. Using the test train split methodology, the majority of the time, the test train split of the dataset is performed in a random manner. This strategy is used. After that, the datasets are segmented differently according to the ratio of test data to training data. When performing a test on time series, the data for the train are typically partitioned according to the time boundary. After being educated with the aid of the train data set, the model is subsequently tested with the assistance of the test data set. The dates associated with the first 190 week records contained in the train data set range from "2017-11-03" to "2021-08-13." The range of values that are included in the Testing data set is from "2021-08-20" to"2022-10-28."

Estimating the values for parameter

- **Augmented Dickey Fuller Test** utilising the Augmented Dickey Fuller Test as a guiding principle for analysing the data. The p value for the data is lower than 0.05, which is considered significant. The value of p is 0.02374, and this indicates that 95 percent of the data fall below the threshold of 0.02374. The values that are below the statistically significant threshold are extremely close to zero. It is referred to be the null hypothesis when the value is over the significant threshold.

- **P value** The Partial auto correlation function is utilised in order to do research on the significance level of p. According to the graph, the partial autocorrelation provides a relation between the overall lag associated border value and the lag related data. The first two lags in the graph are outside of the acceptable range; the lag value that is closest to the average autocorrelation is the second one, which means that the value for the p factor is 1. A substantial threshold is represented by the thin blue line in the graph.The graph's meaningful threshold is represented

by the narrow blue line. The value that is below the considerable threshold is deemed to be quite near to the value of zero.

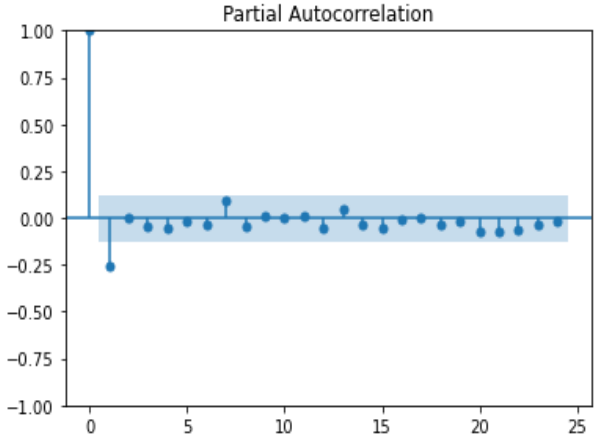

Figure 4: Partial Auto-correlation

- **d value** Through differentiation of the data, the value of d can be determined. When one compares the graph to each of the differences, one may determine the value. Plotting the normal data, making a differentiation of the original data, and plotting the second order differentiation are all part of the process of fitting time series data to a differencing technique. A comparison between the graph of the original data, the graph of the first order differentiation of the data, and the graph of the second order differentiation of the data The random fluctuations in the data are arranged in-an ascending order for each and every difference. A differencing graph has been plotted for the data set described above. The value d is defined by comparing all three graphs, and the comparison reveals that the white noise in the first order differencing is low than the original data, leading to the conclusion that the value of d for apple bramley is 1.

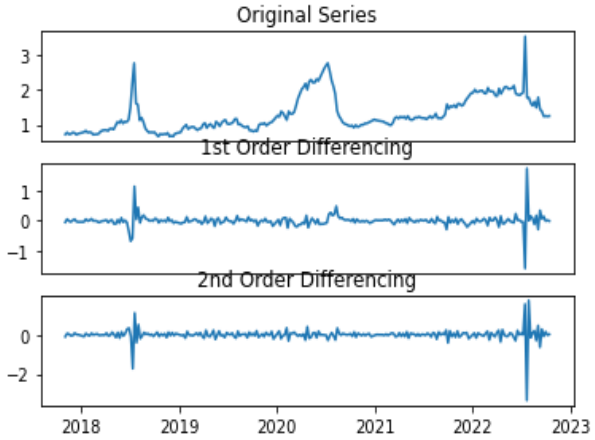

Figure 5: First and second order differentiation of time series data set

- **q value** The q value of this data set can be calculated by first doing a first-order differentiation

and then carrying out an autocorrelation. The q value can be determined by first finding all of the points that peak out from the significant threshold, and then adding those points together. When one looks at the graph, there is one point that stands out as being outside of the normal range. These values, which are known as the q values and have a value of 0, are considered to be equal.

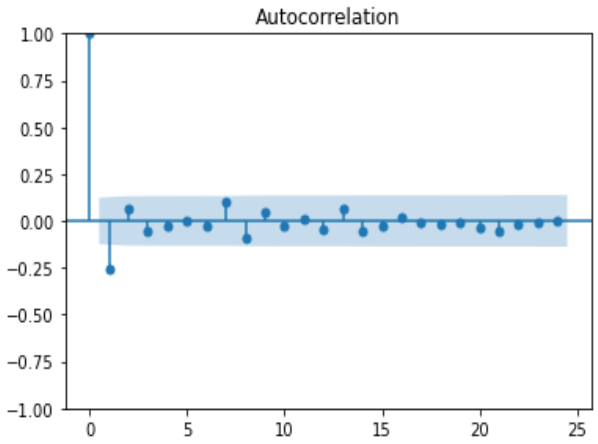

Figure 6: Auto correlation of time series data set

Analysing the time series data set with the defined parameters

In most cases, the PACF and ACF techniques are utilised in order to do an analysis on the ARIMA model's parameter. ALL of the parametrical values are compared to the theoretical values of the parameters, and a comparison is performed between the two.

Table 1: parameters for ARIMA model

PARAMETER(p,q,d)	AIC
(1,1,1)	-100.296
(0,1,0)	-87.711
(1,1,0)	-102.290
(0,1,1)	-101.579
(2,0,1)	-89.684
(2,1,0)	-100.294
(2,1,1)	inf

The table 1 provides a relationship between the ARIMA model and all of the parameters that are applicable. The lowest possible number of AIC values indicates that the model has a very low rate of error overall. Estimated utilising the underlying theory, the values of the parameters for the Auto Regressive Integrated Moving Average are as follows:

Table 2: Parameteric value for ARIMA equation

Parameter	coefficient	standard error
Auto Regression(1)	-0.2532	0.017
Moving Average(1)	0.0379	0.001
Sigma(0)	0	0

- p=1
- d=1
- q=0

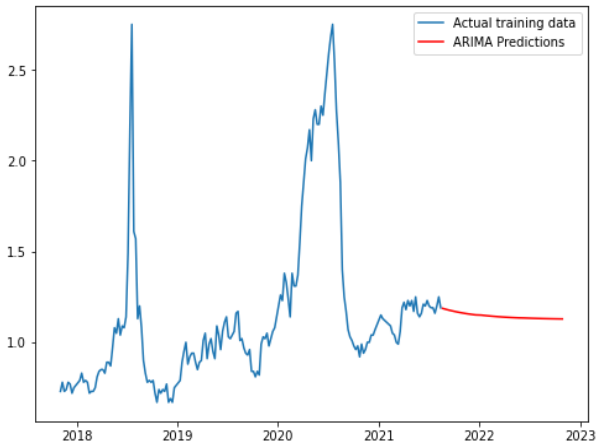

Figure 7: Apple bramley tested and predicted values

The table 2 is determined with the value of test and train of the data .The data set are fitted in the model with best parametric value for p,d,q.

Table 3: Accuracy for ARIMA:

Validation method	accuracy
MAE	90.29%
MAPE	88.8%
MASE	88.9%
RMSE	80.03%

The table 3 relates the accuracy of ARIMA model for apple bramley.The score for the model are mean absolute error is 90.29 percent,mean absolute percentage error is 88.8 percent,mean absolute square error is 88.9 percent and root mean square error is 80.03 percent.

3.2.2 predicting the price of tomato (round)

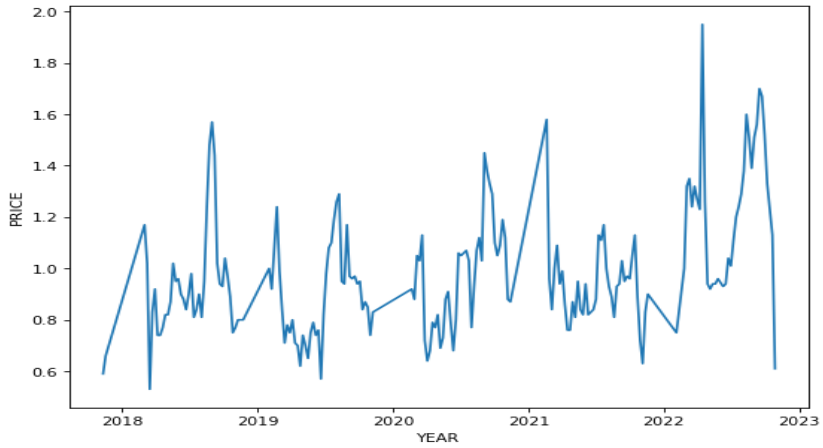

Figure 8: Weekly price of round tomatoes

Figure 8 visualizes the price fluctuation of round tomatoes occurs over a period of time. ARIMA Model is to estimate the approaching pricing since this is a typical time series problem that has evolved as a result of the fact that the price has gone growing over the course of the previous many years.

Table 4: Parametetric value for ARIMA model

PARAMETER(p,q,d)	AIC —
(1,0,1)	46.55
(0,0,0)	525.97
(1,0,0)	44.77
(0,0,1)	351.75
(2,0,0)	46.53
(2,0,1)	47.81
(1,0,0)	12.55
(3,0,0)	-2.626
(3,0,1)	-1.043
(2,0,1)	2.197
(3,0,0)	48.095

In table 4 the model is evaluated with all the possible parameters value, the associated AIC value is determined. ARIMA (3,0,0) yields

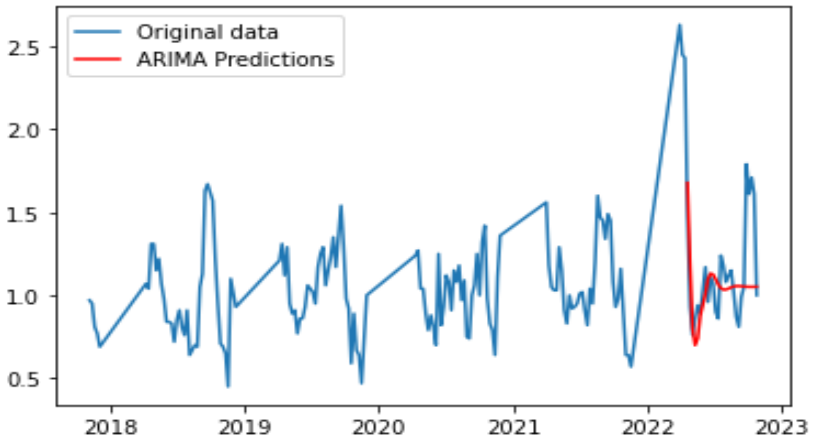

Figure 9: Predictiveweekly price of round tomatoes

In figure 9,The ARIMA (3, 0, 0) parameter was used to make a prediction, and the graph displays the difference between the value of the original data and the value of the data that was forecasted using that parameter. The model, which covers the years 2018 through 2022, is trained with the use of training data. The data set allows for a projection that falls anywhere between the years 2022 and 2023. The value that was predicted for the model is rather close to the value that it actually is.

Table 5: coefficient value for ARIMA equation

Parameter	coefficient	standard error
constant	1.0518	0.054
Auto Regression(1)	0.7121	0.099
Auto Regression(2)	0.0730	0.137
Auto Regression(3)	-0.2901	0.137
sigma(2)	0.0565	0.004

In table 5 ,the coefficient value for the equation of ARIMA(3,0,0)is , autoregressive value is found to be 0.7121,0.0730,-0.2901. The value of sigma is 0.0565, while the value of the constant is 1.0518. The accuracy for the ARIMA(3,0,0) for the Round tomatoes

Table 6: Accuracy score for ARIMA model

Validation method	accuracy
MAE	93.29%
MAPE	91.8%
MASE	90.9%
RMSE	83.03%

In table 6,The ARIMA model with parameters 3, 0, and 0 has an accuracy of 93.29 percent in terms of its mean absolute value error, 91.8 percent in terms of its mean absolute percentage error, 90.9 percent in terms of its mean absolute scalar error, and 83.03 percent in terms of its root mean square error.

3.2.3 predicting the price of carrot top washed

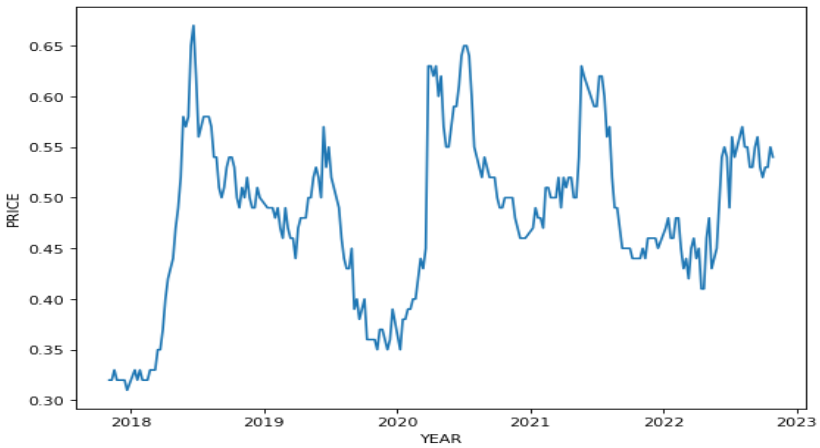

Figure 10: weekly price of top washed carrot

The above graph gives an overview of the distribution of price of top washed carrot over a period of time. This clears that the price of are varied with time period. It is possible to harvest carrots in the United Kingdom for almost the whole year by taking use of the varying natural conditions found around the country and employing a variety of harvesting methods. Carrots for the early season are often seeded in the winter and very early spring .The price of top washed carrot are fitted to all the possible parameters in the ARIMA model.

Table 7: Parametric value for ARIMA model

PARAMETER(p,q,d)	AIC —
(1,0,1)	-1103.575
(0,0,0)	349.222
(1,0,0)	inf
(0,0,1)	29.56
(2,0,0)	inf
(2,0,1)	-1101.502
(1,1,0)	-1111.271
(2,1,0)	-1109.66
(2,0,2)	-1099.41
(2,0,1)	-1107.67

The table 7 is defined with every conceivable parameter along with its associated AIC value. The parameter that provides the best estimate for the ARIMA model's prediction of (1,0,0) for the top washed carrot.

- p:1
- q:0
- d:0

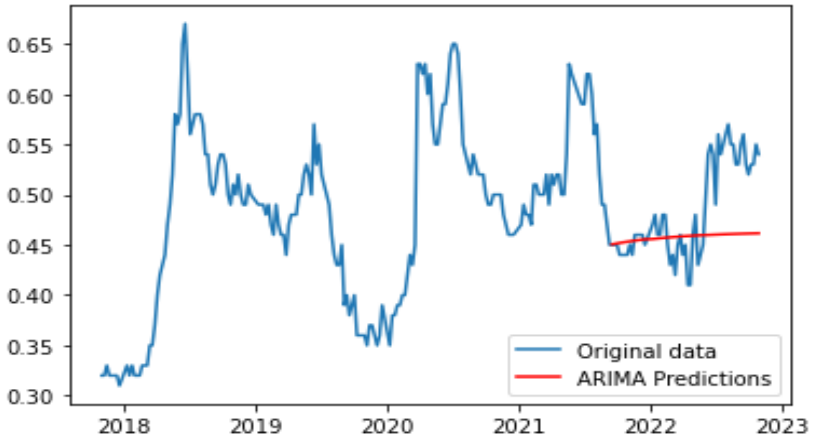

Figure 11: Price prediction of top washed carrot

Figure 11 visualizes relates the original price top washed carrot to the predicted models price. The model is trained with 4 years data and predicted with the one year data . The original price varies highly ,but the predicted price are in linear increase.

Table 8: Coefficient value for ARIMA equation

Parameter	coefficient	standard error
intercept	0.0268	0.009
Auto Regression(1)	0.9438	0.019
sigma(2)	0.0006	2.35e-05

In table 8 ARIMA model with parameters p = 1, d = 0, and q = 0 for the top washed data has an intercept of 0.0268, an auto regression coefficient value of 0.9438, and a sigma value of 0.0006 in its equation.

Table 9: Accuracy score for the ARIMA

Validation method	accuracy
MAE	88.79%
MAPE	85.77%
MASE	82.99%
RMSE	80.82%

Table 9 compares the ARIMA model with parameters 1, 0, and 0 has an accuracy of 88.79 percent in terms of its mean absolute value error, 85.77 percent in terms of its mean absolute percentage error, 82.99 percent in terms of its mean absolute scalar error, and 80.82 percent in terms of its root mean square error.

3.3 SARIMA model

3.3.1 predicting the price of apples (varient is bramleys seedling)

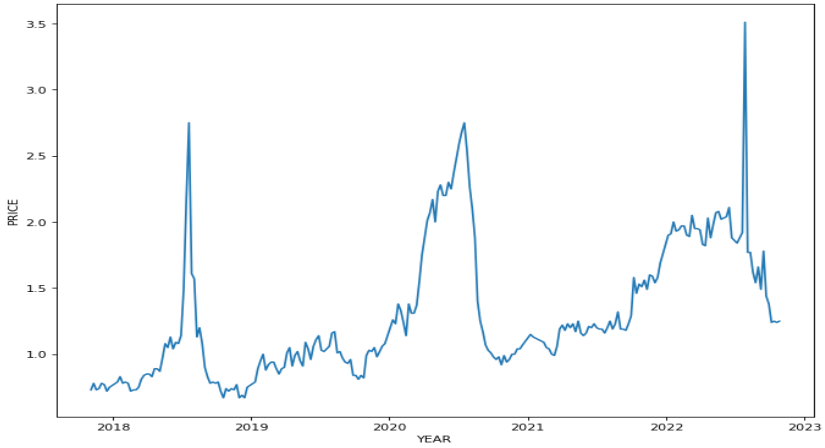

Figure 12: weekly price of apples bramleys seedling varient

The price's movement through time is depicted in greater detail in the graph that was just presented. Figure 12 illustrates how the cost of apple bramleys shifts over the course of a certain amount of time. Every week there is a new range of prices for fruit to be found. Because this change in price is not presented in a suggestion, it is impossible to analyse these sets of data by employing a straightforward linear function. So Auto regression model is used to analyse the data. Seasonal changes are causing these kinds of statistics to fluctuate.

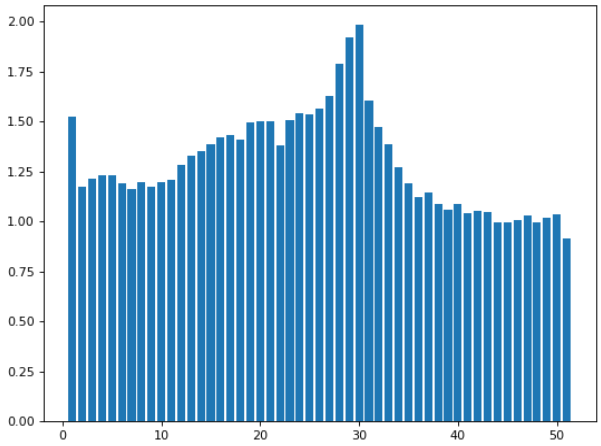

Figure 13: Weekly price of overall dataset

Figure 13 says that the data are seasonally getting varied .In which the , the price of weekly price depends on the previous years same week data .This above graph is the overall mean of the price of apple.

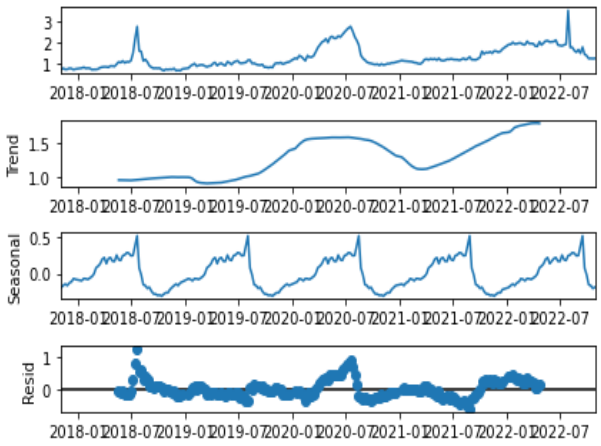

Figure 14: Seasonality,trend and resid of apple bramely

Figure 14 provides a more in-depth illustration of the seasonal fluctuation of the data set that is being utilised and can be found here. These statistics demonstrate seasonality in addition to a trend and a pattern that remains consistent over the course of time. [Case in point:] An observation is made of a stochastic trend in which the pattern of the trend is not stated. The observation is made with regard to the trend. The graph labelled "Seasonality" depicts how the price tends to go through periodic cycles of rising and falling prices over a set period of time. The fact that this is the case allows us to draw the conclusion that the data set is sensitive to seasonal variability. The Seasonal Auto Regression Model is utilised in order to analyse and estimate the value that is included in a data set that incorporates seasonal information. This is done by analysing the data set.

The SARIMA model is built using two types of time-stamped data: seasonal parameters and nonseasonal parameters. Both types of parameters are time-stamped.

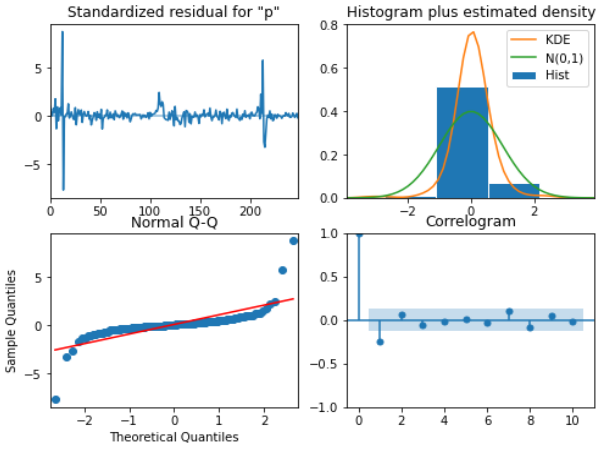

Figure 15: Diagnostic plot of Apple bramley data values

Standard residual for p In figure 15, Top left graph is an standardised residual for Apple dataset. There is no obvious pattern observed in the graph. **Histogram density plot** Orange line is

the smoothed version of histogram and Green line is the normal distribution. **Normal distribution** The normal distribution, also known as the Gaussian distribution, is a type of probability distribution that is symmetric around the mean. This means that it demonstrates that data that are closer to the mean are more likely to occur than data that are further away from the mean. **normal q-q plot** The data set is in normally distributed, in which all the data points are around red line and end of plot the values are scattered far away from red line. **Correlogram plot** Correlogram plot is an Auto Correlation function plot. In which 95% of the correlations for lag greater than zero should not be significant.

Methodologies such as auto correlation and partial auto correlation are utilised in the process of calculating seasonal and nonseasonal parameters.

In the case of the seasonal parameter, the data set for the parameter is either backshift or delayed in relation to time.

The lag time, also known as the back shift time, is characterised by the pattern of change in the data over a certain amount of time.

Seasonal parameter The recorded apple bramley variety is a weekly data set that records the seasonal backshift. P,D,Q are the seasonal parameters identified after the seasonal backshift of the data. By applying the PACF to the Apple bramleys data set. Using the plot of PACF the pattern of change in the data is 6, which refers to every 52 log difference where the value becomes repeated. The data is backshift with a lag of 52 and the acf and pacf are identified. Every year, the data are cyclic, and the data are season for every 52 data entries. The values for the parameters are:

- P are equal to 1. (one significant positive spikes in ACF and PACF plots)
- D are equal to 0. (first difference and seasonal difference)
- Q are equal to 0. (significant negative spikes, PACF decay is more gradual)
- s equivalent to 52 (yearly seasonal component)

non seasonal parameter The non seasonal parameters are the ones that are derived using the original time series data set through the use of ACF and PACF. The values of the ARIMA model's parameters and the non-seasonal parameters have a similar relationship. The following are the values for the Non seasonal models

- p are equal to 1. (one significant positive spikes in ACF and PACF plots).
- d are equal to 0. (first difference and seasonal difference).
- q are equal to 0. (significant negative spikes, PACF decay is more gradual).

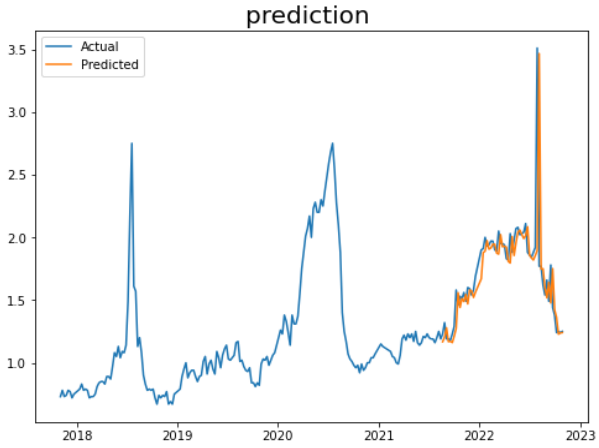

Figure 16: Sarima model apple prediction

In figure 16 is the price of apple bramley is predicted using SARIMA model. The is trained with data set of 2018 to 2022 and tested up with 2022 to 2023. The data of the model is seasonal in which the price is high in august for the consecutive years.

Table 10: Coefficient value for SARIMA model

Parameter	coefficient	standard error
Auto Regression(1)	0.987	0.006
Auto Regression(1) seasonal(52)	0.0532	0.240
sigma(2)	0.0402	0.001

In table 10 compares the coefficient value for the equation of SARIMA(1,0,0)(1,0,0,52)is, autoregression value is found to be 0.987. The seasonal auto regression value is 0.0532, while the sigma value is 0.0402.

Table 11: Accuracy score for SARIMA model with apple data set

Validation method	accuracy
MAE	95.29%
MAPE	94.8%
MASE	92.9%
RMSE	90.03%

The correctness of the model is demonstrated by the table 11. When measured against the first batch of data, the price of the SARIMA model's anticipated value is found wanting. Several distinct methods of measurement are utilised in order to evaluate the correctness of the models. The method and their score are as follows: the means absolute error is 95.29 percent, the means absolute percentage

error is 94.8 percent, the means absolute square error is 92.9 percent, and the means root mean square error is 90.03 percent.

3.3.2 predicting the price of Tomatoes (varient is round)

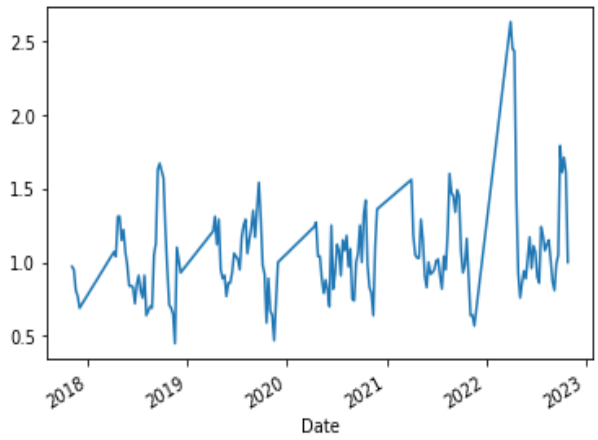

Figure 17: Weekly price of Round tomatoes

The accompanying figure 17 depicts the price's evolution over time more clearly. The price of round tomatoes has been plotted on this graph to show how it has changed over time. The pricing range for fruit changes every week. This price shift is not shown in a proposal, therefore, a linear function cannot be used to analyze these data sets. So The data is analyzed with an auto regression model. Such numbers are erratic because of seasonal variations.

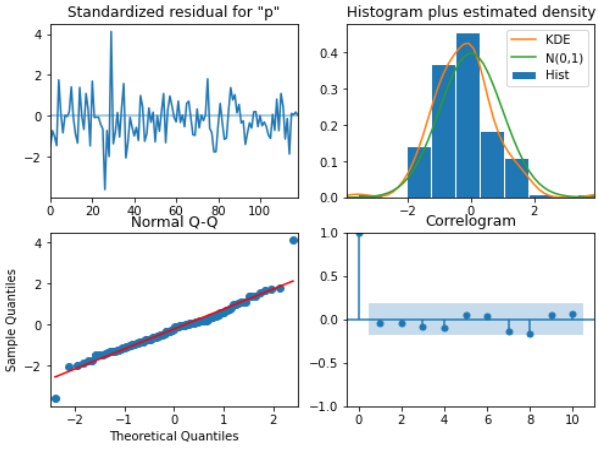

Figure 18: plot diagnostics of Round tomatoes data values.

Standard residual for p In figure 18,Top left graph is an standardised residual for Round tomatoes data set values. There is no obvious pattern observed in the graph. **Histogram density plot** Orange line is the smoothed version of histogram and Green line is the normal distribution. **normal q-q plot** The data set is in normally distributed, in which all the data points are around red line and

end of plot the values are scattered far away from red line. **Correlogram plot** Correlogram plot is an Auto Correlation function plot. In which 95% of the correlations for lag greater than zero should not be significant.

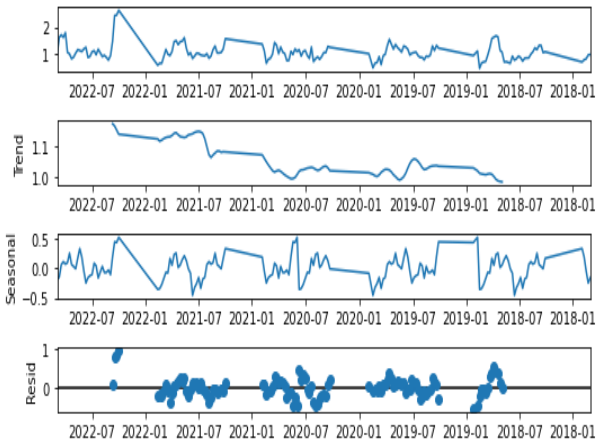

Figure 19: Seasonality,trend and resid of round tomatoes

Figure 19 illustrates the data set's seasonal volatility. These statistics show seasonality, a trend, and a constant pattern. An unstated stochastic tendency is seen. Trend observation. "Seasonality" shows how prices cycle up and down over time. This suggests that the data set is subject to seasonal variation. The seasonal auto regression model is used to analyze and estimate seasonal data.

Table 12: Coefficient for the SARIMA equation

Parameter	coefficient	standard error
Auto Regression(1)	0.7202	0.100
Auto Regression(2)	0.0530	0.146
Auto Regression(3)	-0.3016	0.137
Auto Regression(1) seasonal(52)	-0.8326	0.240
Auto Regression(1) seasonal(104)	-0.4441	1.25e+04
Auto Regression(1) seasonal(156)	0.0004	1.5e+04
sigma(2)	0.0573	0.661

Table 12 defines the coefficient value for the equation of SARIMA (3, 0, 0)x(3, 1, 0, 52) is, autoregression value is found to be 0.7202, 0.0530, and -0.3016. The value of sigma is 0.0573, while the value of the seasonal Auto regression is -0.8326, -0.4441, 0.0004.

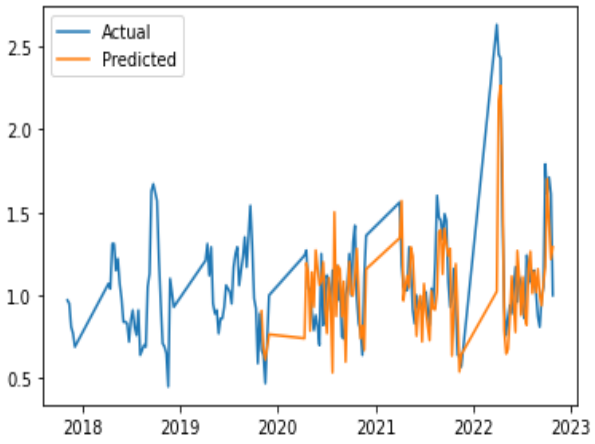

Figure 20: price prediction of Round tomatoes using SARIMA model

Figure 20 forecasts the prices using SARIMA. The model was trained on data from 2018 to 2023 and tested on data from 2022 to 2023. The model's seasonal data shows low prices at the year's conclusion.

Table 13: Score for predicted model

Validation method	accuracy
MAE	93.44%
MAPE	92.7%
MASE	91.74%
RMSE	97.93%

Table 13 explains the accuracy score for the fitted model. It is decided how well the expected model fits the data that was collected at the beginning. From 2018 until 2023, the model is refined using round tomatoes, and from 2022 until 2023, it is put to the test. There are four different methods that are used to determine how accurate the model is. When it comes to accuracy, the scores are as follows: the mean average error is 93.44 percent, the mean absolute percentage error is 92.7 percent, the mean absolute square error is 91.74 percent, and the root mean square error is 97.93 percent.

4 Discussion and Conclusion

Consuming an appropriate amount of fruits and vegetables each day is associated with a reduced risk of developing non-communicable illnesses. Fruits and vegetables are an important source of the micronutrients that are found in diets. The World Health Organization (WHO) suggests that individuals consume a daily total of 400 grams of fruit and vegetables. However, the intake of fruits and vegetables across the world frequently falls well short of that aim. The epidemic has caused catastrophic disruptions in social and economic systems all across the world, including the world's longest and most severe recession since the Great Depression. This recession has been caused by the pandemic. The lockdown that were implemented in an effort to stop the spread of the disease are to blame for this recession. The primary cause of widespread supply shortages, the most significant of which was a lack of food, was disruptions in the supply chain.This distruption trends the price of fruits and vegetable price.A research has proposed to predict the price of fruits and vegetables using machine learning technique. Autoregressive integrated moving average and the seasonal auto regressive integrated moving average are the methodologies used to analyze and forecast the price of fruits and vegetables. The data set are from United Kingdom weekly wholesaler in Birmingham, Bristol, Manchester and London market.The percentage of yield and the amount of demand for fruits and vegetables in the market are used to formulate an estimate for the price of fruits and vegetables. Because of the seasonal nature of their production, fruits and vegetables have prices that change from week-to-week. The five-year average weekly price of fruits and vegetables was the data set that we analyzed for this study. After being cleaned (by getting rid of null values, zero values, and outbound values), the data sets are then separated into training and testing data sets. Both the ARIMA and SARIMA models require training data sets in order to be constructed, and testing data sets in order to be validated. Mean absolute error, mean absolute percentage error, mean absolute scaled error, and root mean square error are the metrics that are utilized in the process of determining the model's level of precision. The price of fruit changes with the seasons and goes up and down over time. The SARIMA model provides more accurate predictions than the ARIMA model does. Fruits and vegetables price are estimated not only by the previous value,the price of fruits and vegetables are depends on other calamities, those factors are seasonal change in nature, economic crisis, rainfall, temperature. These are the dimension which changes the price of fruits and vegetables. In future These are estimated with the price forecasting model using seasonal auto regression integrated model.

References

[Abewickrama, 2022] Abewickrama, G. (2022). Analyzing and predicting the market price fruits and vegetables.

[Chang and Liao, 2010] Chang, Y.-W. and Liao, M.-Y. (2010). A seasonal arima model of tourism forecasting: The case of taiwan. *Asia Pacific Journal of Tourism Research*, 15(2):215–221.

[Dharavath and Khosla, 2019] Dharavath, R. and Khosla, E. (2019). Seasonal arima to forecast fruits and vegetable agricultural prices. In *2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, pages 47–52.

[Kamil and Razali, 2015] Kamil, M. S. and Razali, A. M. (2015). Time series sarima models for average monthly solar radiation in malaysia. In *2015 International Conference on Research and Education in Mathematics (ICREM7)*, pages 256–261.

[McDonald et al., 2013] McDonald, S., Coleman, S., McGinnity, T. M., and Li, Y. (2013). A hybrid forecasting approach using arima models and self-organising fuzzy neural networks for capital markets. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

[Peng et al., 2015] Peng, Y.-H., Hsu, C.-S., and Huang, P.-C. (2015). Developing crop price forecasting service using open data from taiwan markets. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 172–175.

[Rahman et al., 2021] Rahman, M. H., Nahid, S. I., Al Fahad, I. H., Nahid, F. M., and Khan, M. M. (2021). Price prediction using lstm based machine learning models. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0453–0459.

[Wijekoon et al., 2021] Wijekoon, W., Wijewardana, L. W., Wattegedara, S., Kumara, W., Sasini, W., and Abeygunawardhana, P. K. (2021). Iot based classification and price prediction of organically and inorganically grown vegetables and fruits. In *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, pages 1–5.

Appendix:

Data sets and the python code are contained in the given google drive. google drive weblink.

The file named vege_price are in csv format used for implementation models.