Onion Agro Supply chain analysis and optimization

A PROJECT REPORT

Submitted by

Kanishkan B -17BME031

Hari Prasad R -17BME052

Hrishikesh N M -17BME117

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

MECHANICAL ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY

COIMBATORE 641049.

(An Autonomous Institution Affiliated to Anna University, Chennai)

DECEMBER 2020

KUMARAGURU COLLEGE OF TECHNOLOGY, COIMBATORE

BONAFIDE CERTIFICATE

Certified that this project report **“Onion Agro Supply chain analysis and optimization"** is the bonafide work of **“KANISHKAN B, HARI PRASAD R, HRISHIKESH N M”,** who carried out the project work under my supervision during the year 2020 - 2021.

PROJECT GUIDE **HEAD OF THE DEPARTMENT**

DECLARATION

We affirm that the project work titled "Onion Agro Supply chain analysis and optimization" being submitted in partial fulfillment for the award of **MECHANICAL ENGINEERING** (B.E) is the original work carried out by us. It has not formed the part of any other work submitted for award of any degree or diploma, either in this or any other University.

Signature of candidate	Signature of candidate	Signature of candidate
(HRISHIKESH N M)	(KANISHKAN B)	(HARI PRASAD R)

I certify that the declaration made above by these candidates is true.

Signature of project guide

(Dr. M. Balaji)

ABSTRACT:

Agriculture is the primary source of livelihood for about 58% of India's population. India is the third-largest producer of vegetables and the second-largest producer of onion. Indian onions are famous for their pungency and are available round the year. Indian onions have two crop cycles, first harvesting starts in November to January, and the second harvesting from January to May.

Having said the significance of onion cultivation in India, this project mainly deals with the cost forecasting and the supply chain optimization of the Indian onion agro supply chain. The main purpose of this project is to develop a mathematical model to forecast and analyze the cost of onion in the Indian market starting from 2004 to 2020. We specifically took the onion price data of Chennai Market as the acquired data was relevant and accurate. The monthly wholesale prices and arrivals data for the study collected from the National Horticultural Research and Development Foundation. The three main machine learning models used in the project are the Mean Model, Linear Model, and the Random walk Model. We have also visualized the forecasted data through a line graph comparing the factors: cost and date. The accuracy of proportion among the forecasted and actual price value of the onion was found in between _ to _ percent. We found that the linear model was more accurate comparing the other three mathematical models.

Keywords: Supply chain optimization, Mean Model, Linear Model, Random walk Model.

ACKNOWLEDGEMENT

We are wholeheartedly thank our Chairman **Dr. B.K KRISHNARAJ VANAVARAYAR**; our Correspondent **Thiru. M.BALASUBRAMANIAM;** our Joint Correspondent **Thiru.SHANKAR VANAVARAYAR,** our advisor **Dr.M.GURUSAMY** for providing us the required infrastructure at **KUMARAGURU COLLEGE OF TECHNOLOGY.**

We express our gratitude to our beloved Principal **Dr. J. SENTHIL ,** for his invaluable support, motivation and guidance, and also for providing us all the necessary facilities required for carrying out this project work.

We are very grateful to our respected Head of the Department, Mechanical Engineering, **DR.C.VELMURUGAN** for his constant and continuous motivation, review and cooperation throughout this project work.

We wish to record our profound happiness and gratitude to our Project Coordinator **DR.SATHYABALAN** and Project Guide **Dr. M. BALAJI** Our sincere and hearty thanks to all the faculty members and staff of Mechanical Engineering Department for their well wishes, timely help and support rendered to us for doing this final year design work.

We are very greatly indebted to our family, relatives and our all friends without whom our life would not have been shaped to this level.

TABLEOF CONTENTS

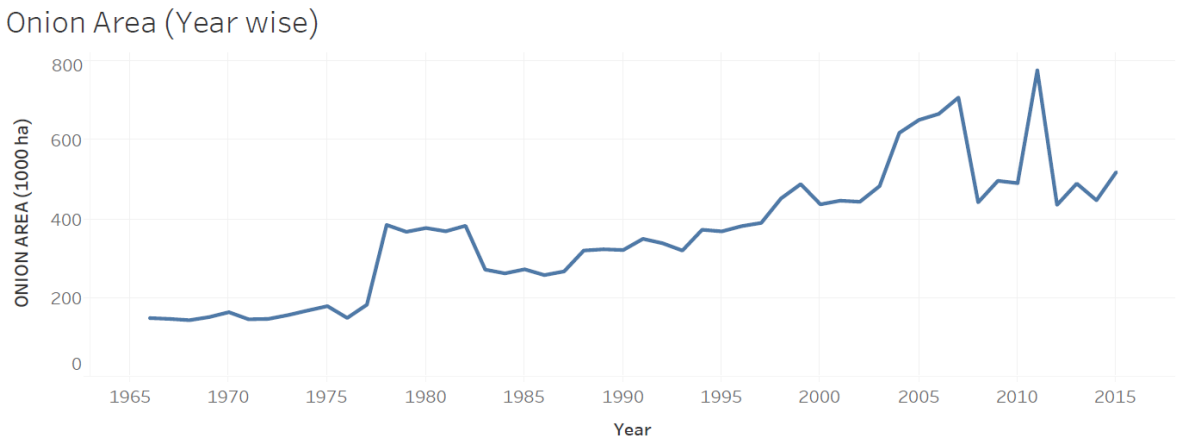
CHAPTER 1

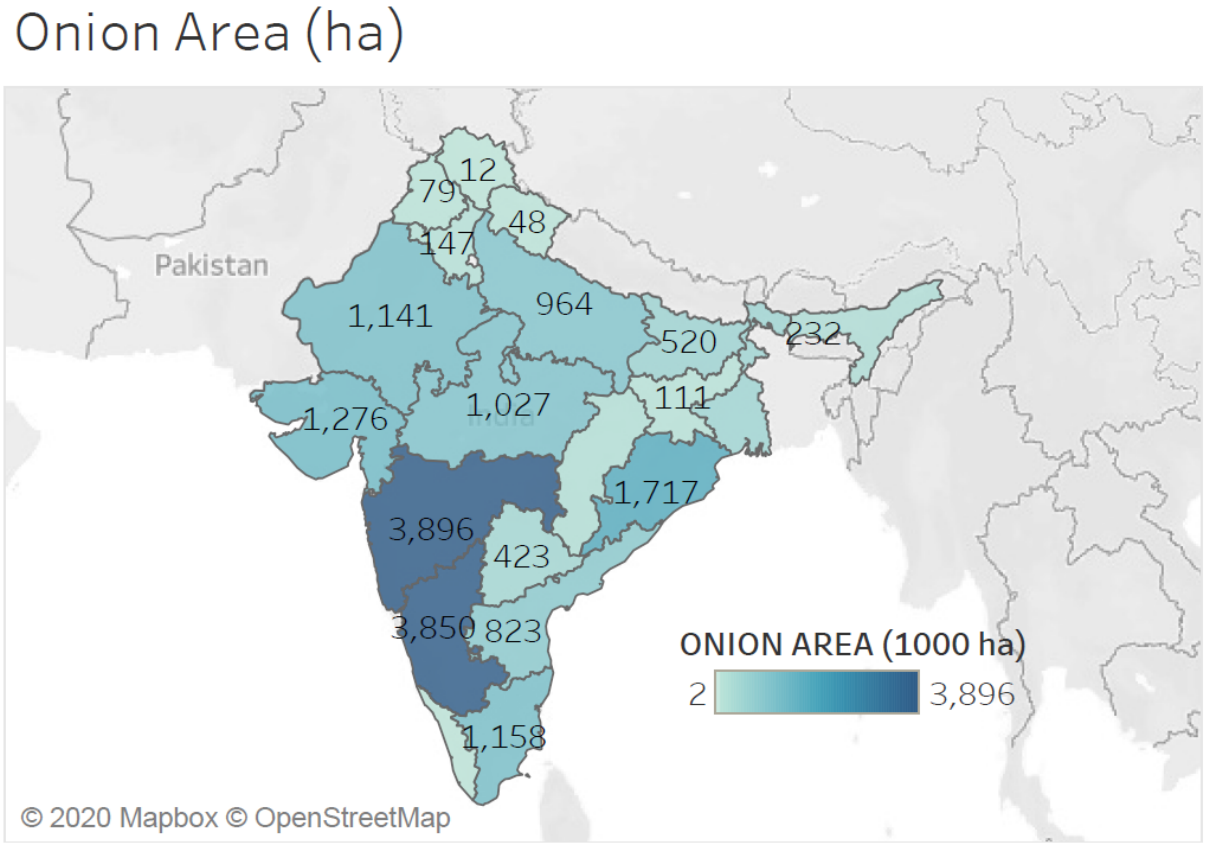
INTRODUCTION:

Onion is one of the second most important commercial crops of India which is next to Potato. In the world, Onion crop is grown in about 5.30 million hectare area with an annual production of 88.48 million tons with a productivity of 16.70 tons per hectare. China stands first in Onion production (22.61 million tons from an area of 1.03 million hectares area) in the world with a productivity of 21.85 tons per hectare followed by India. In India, the Onion crop is grown in about 1.20 million hectares area with an annual production of 19.40 million tons with a productivity of 16.12 tons per hectare. The quantity of Onion 2415.75 thousand tons is exported from India which outputs the value of 3, 10,650.09 Rs. lakhs.

It is also one of the most important commercial vegetables grown widely by large, small, and marginal farmers in different parts of the country. Present production is enough to meet the domestic requirement with a limited quantity of onion export to gulf countries and need to focus on export for European countries. The self-sufficiency of onion production is achieved due to proper planning of onion production in three seasons namely Kharif, late Kharif, and rabi seasons, handling of post-harvest techniques like topping, curing and storage, etc. Still, it is necessary to make efforts for enhancement of productivity and minimizing post-harvest losses to meet increasing demands of both domestic as well as export markets. So we have come up with a solution to better manage and optimize the popular commodity “Onion”.

This project deals with the visualization and forecasting of onion prices in Chennai. The data visualization is done through Tableau and python. This project helps stakeholders and participants of the onion agro supply chain to take a data-driven decision and supports them to react proactively to the varying market demands.





1.4. TERMINOLOGY:

Time series forecasting

Time series forecasting is the use of a model to predict future values based on previously observed

values. Time series are widely used for non-stationary data, like economic, weather, stock price, and retail sales in this post. We will demonstrate different approaches for forecasting retail sales time series.

$$X_t = X_{t-1} + e_t \ ,$$

Mean (constant) model

For purposes of statistical forecasting, the simplest non-trivial kind of time series is one that is stationary and completely random.

The natural forecast to use for all future values is therefore the sample mean of the past data. The forecasting equation for the mean model is thus:

where the estimated constant (alpha) is the sample mean of Y

LINEAR TREND MODEL $Y(t) = \alpha$

It is a model that models or fits the data into a straight line. It provides the line of best fit that can be used to represent the behavioral aspects of the data to determine if there is any particular pattern. A trend line used on the scatter plot determines the type of relationship between two variables. Linear trend model expresses the data as a linear function

The forecasting equation for the linear trend model is:

where t is the time index. The parameters alpha and beta (the "intercept" and "slope" of the trend line) are usually estimated via a simple regression in which Y is the dependent variable and the time index t is the independent variable.

$$Y(t) = \alpha + \beta t$$

Random walk model

A time series said to follow a random walk if the first differences (difference from one observation to the next observation) are random.A random walk model for a time series Xt can be written as

where Xt is the value in time period t, Xt−1 is the value in time period t−1 plus a random shock set (value of error term in time period t).

Log Transformation

Log transformation is a data transformation method in which it replaces each variable x with a log(x). The choice of the logarithm base is usually left up to the analyst and it would depend on the purposes of statistical modeling

Machine learning RMSE

Each machine learning model is trying to solve a problem with a different objective using a different dataset . RMSE (Root Mean Square Error) represents the sample standard deviation of the differences between predicted values and observed values (called residuals). Mathematically, it is calculated using this formula:

Machine learning model accuracy

Evaluating your machine learning algorithm is an essential part of any project. There are many different ways to evaluate the accuracy. .

Classification Accuracy:

$$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{All Samples}}$$

Logarithmic Loss:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Mean Absolute Error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, x_i - x \,|$$

Mean Squared Error:

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

CHAPTER 2

2.1. LITERATURE REVIEW :

A literature survey has been conducted to support our project work. Crop yield prediction in agriculture is significant because it supports better crop management and planning marketing activities. The use of ML algorithms in providing real-time analytic insights for pro-active data-driven decision-making is very helpful [1]. The rise of data analytics in farming is commonly referred to as precision agriculture (PA) [2]. Data Management Resources should be considered a key building block in supply chain planning [3].

Machine learning models can be leveraged to accurately predict supply chain risks. Decision tree models can reveal correlations that influence Supply Chain Risk Management (SCRM) decision-making [4]. Strategies such as the adoption of industry 4.0 technologies, supply chain collaboration, and shared responsibility are identified for a sustainable future [5]. A stochastic model for such a process explains in probabilistic terms how a series arises [6]. Investigated the applicability of advanced machine learning techniques, including neural networks, recurrent neural networks, and support vector machines to forecasting distorted demand at the end of a supply chain [7].

The mainstream of modern agriculture made a pathway where the future of postmodern agriculture can be transformed by the principles of eco-agriculture. Sustainability of ecological agricultural practices and typical alternate agriculture can be achieved using supply chain analytics [8]. Disruptive information helps to find out the solutions for problems like productivity and yield improvement, water conservation, ensuring soil and plant health, and enhancing environmental stewardship. Data analytics hold the key to ensure future food security, food safety, and ecological sustainability [9]. Supply chain risks such as technical risk, information risk, quality and coordination of organizational risk, and security risk can be evaluated with the help of AHP and fuzzy comprehensive evaluation methods. The agricultural supply chain has a structural complexity and market uncertainty, market power imbalance and vulnerability, and so on [10].

CHAPTER 3:

3.1. PROBLEM IDENTIFICATION:

- India is the world's third-largest producer of food grains, the second-largest producer of fruits and vegetables, and the largest producer of milk; it also has the largest number of livestock. It is estimated that global food production must be increased by 60–110% to feed 9–10 billion of the population by 2050.

- To satisfy this huge demand and to proactively react to evolving changes a well-optimized and well-connected data-driven supply chain is needed for the agricultural industry.

CHAPTER 4

4.1. METHODOLOGY:

Data collection:

After identifying the objective behind our analysis, the next step is to collect the necessary data required by us to draw appropriate conclusions. There are various methods by which we can collect data. Some of which are:

- API or Web Scraping — If the data needed is available in a particular website(s), then we can use the website's API (if available) or Web Scraping techniques to collect, and store data in our local storage/ databases. Often, data collected from the Internet is stored in a JSON format, and further processing is needed to convert JSON to the commonly used ".csv" format.
- Databases — If the data required is available in our companies databases, then we can easily use SQL queries to extract the data needed from them.
- Sites like kaggle.com store data sets in appropriate formats to be downloaded by the members for practice/ competitions.

For this project, we have used the NHRDF website. This is the website of the National Horticultural Research & Development Foundation and maintains a database on Market Arrivals and Price, Area and Production and Export Data for three commodities - Garlic, Onion, Tomato, and Potatoes. It also has data from 2004 onwards and the data is exported in tabular format.

SAMPLE DATA

	Market	Month Name	Year	Arrival (q)	Price Minimum (Rs/q)	Price Maximum (Rs/q)	Modal Price (Rs/q)
0	CHENNAI	January	2004	103400	798	1019	910
1	CHENNAI	January	2005	120500	430	638	533
2	CHENNAI	January	2006	111900	428	621	524
3	CHENNAI	January	2007	84800	900	1370	1129
4	CHENNAI	January	2008	127400	588	1000	797
5	CHENNAI	January	2009	111320	1428	2028	1762
6	CHENNAI	January	2010	110000	1639	2259	1980
7	CHENNAI	January	2011	102000	3583	4583	4083
8	CHENNAI	January	2012	126000	771	1013	892
9	CHENNAI	January	2013	116700	1786	2132	1964

Data cleaning:

Data cleaning is the process of detecting and correcting missing, or inaccurate records from a data set. In

this process, data present in the "raw" form (having missing, or inaccurate values) are cleaned appropriately so that the output data is void of missing and inaccurate values. Since no two data sets are the same, therefore the method of tackling missing and inaccurate values vary greatly between data sets, but most of the time, we either fill up the missing values or remove the feature which cannot be worked upon. Redundant data is removed and date is extracted from the year and month column.

Finally, one crucially important element of data preparation not to overlook is to make sure that our data and our project are compliant with data privacy regulations. Personal data privacy and protection are becoming a priority for users, organizations, and legislators alike and it should be one for us from the very start of our data journey. In order to execute privacy-compliant projects, we'll need to centralize all our data efforts, sources, and datasets into one place or tool to facilitate governance. Then, we'll need to clearly tag datasets and projects that contain personal and/or sensitive data and therefore would need to be treated differently.

CLEANED DATA

	market	month	year	quantity	priceMin	priceMax	priceMod	date
0	CHENNAI	January	2004	103400	798	1019	910	2004-01-01
1	CHENNAI	January	2005	120500	430	638	533	2005-01-01
2	CHENNAI	January	2006	111900	428	621	524	2006-01-01
3	CHENNAI	January	2007	84800	900	1370	1129	2007-01-01
4	CHENNAI	January	2008	127400	588	1000	797	2008-01-01
5	CHENNAI	January	2009	111320	1428	2028	1762	2009-01-01
6	CHENNAI	January	2010	110000	1639	2259	1980	2010-01-01
7	CHENNAI	January	2011	102000	3583	4583	4083	2011-01-01
8	CHENNAI	January	2012	126000	771	1013	892	2012-01-01
9	CHENNAI	January	2013	116700	1786	2132	1964	2013-01-01

Exploratory Data Analysis:

Once the data is collected, cleaned, and processed, it is ready for Analysis. As we manipulate data, we may find we have the exact information we need, or we might need to collect more data. During this phase, we can use data analysis tools and software (in this project Tableau and Python is being used) which will help us to understand, interpret, and derive conclusions based on the requirements.

INFO ABOUT THE DATA

	year	quantity	priceMin	priceMax	priceMod
count	204	204	204	204	204
mean	2012	110667	1443	1791	1622
std	5	15078	1155	1324	1237
min	2004	63900	304	456	384
25%	2008	101550	753	1000	879
50%	2012	110450	1089	1446	1263
75%	2016	120275	1784	2138	1954
max	2020	150400	8696	11130	9876

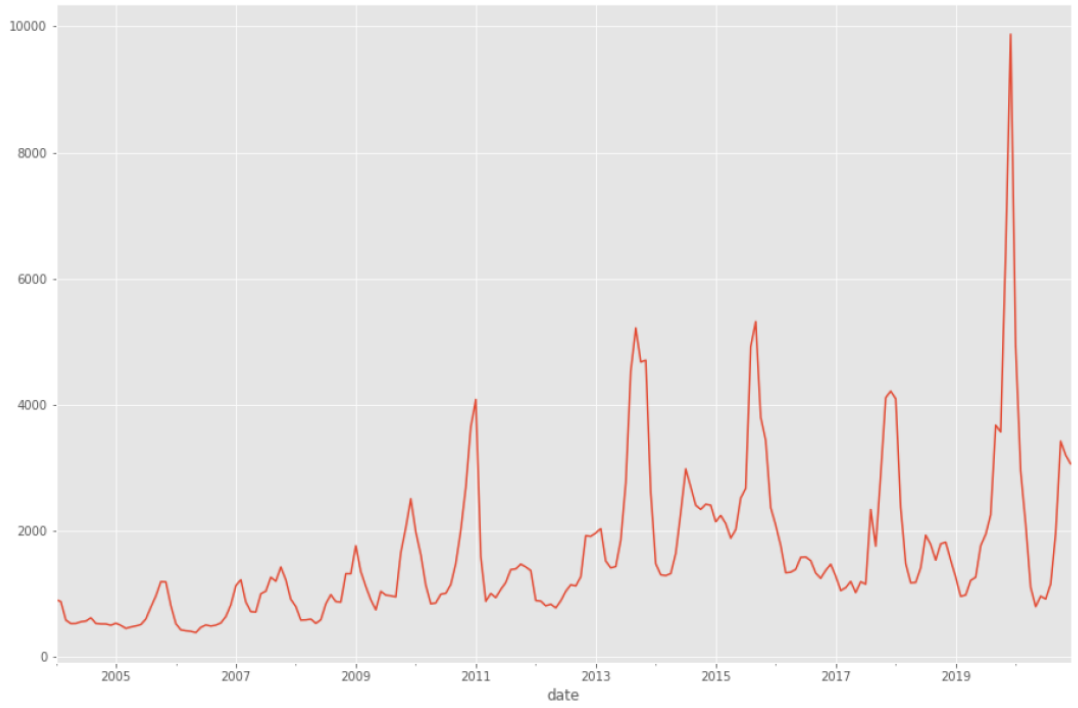
Data visualization:

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

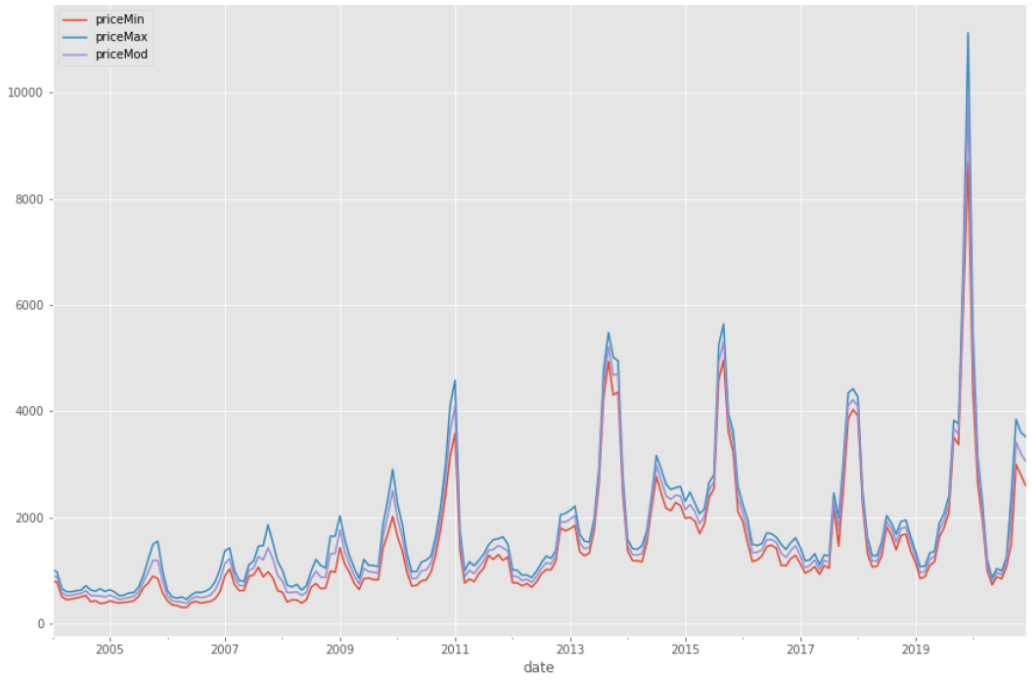
In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Our eyes are drawn to colors and patterns. We can quickly identify red from blue, square from the circle. Our culture is visual, including everything from art and advertisements to TV and movies.

Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If we've ever stared at a massive spreadsheet of data and couldn't see a trend, we know how much more effective visualization can be.

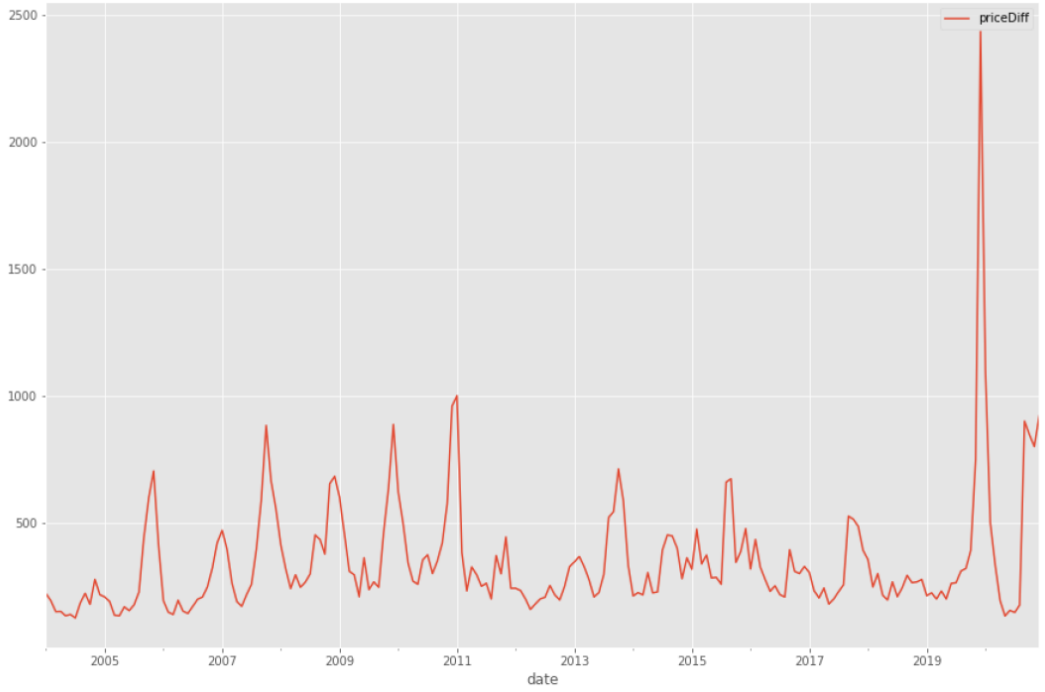
MOD PRICE vs DATE



PRICES vs DATE



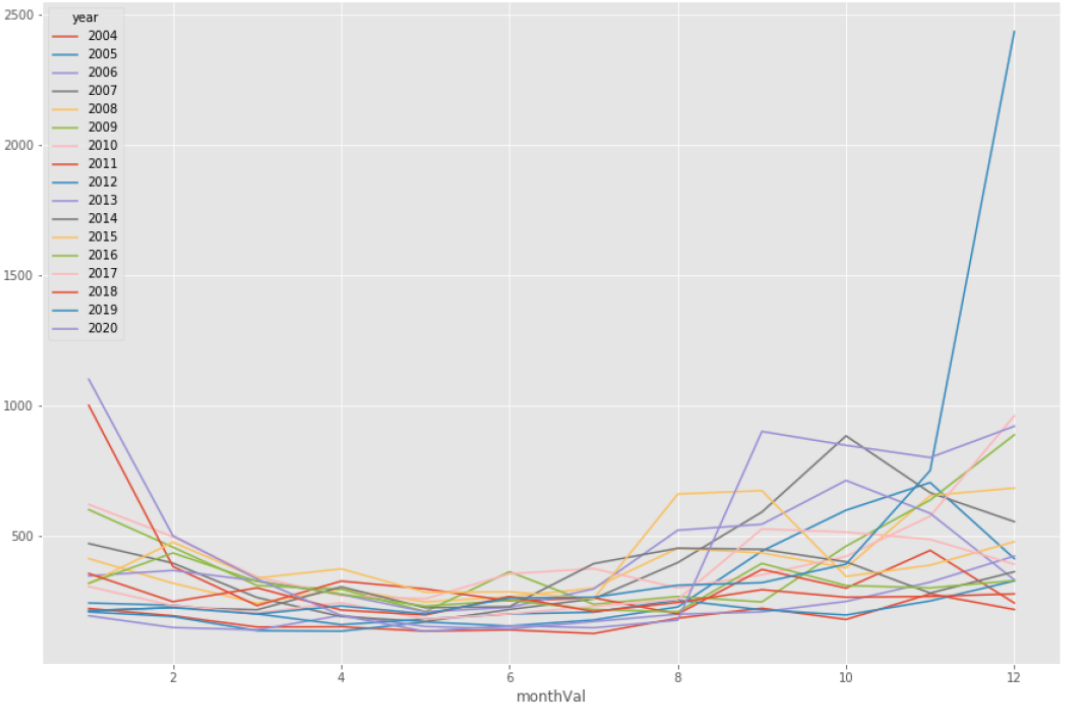
PRICE DIFFERENCE vs DATE



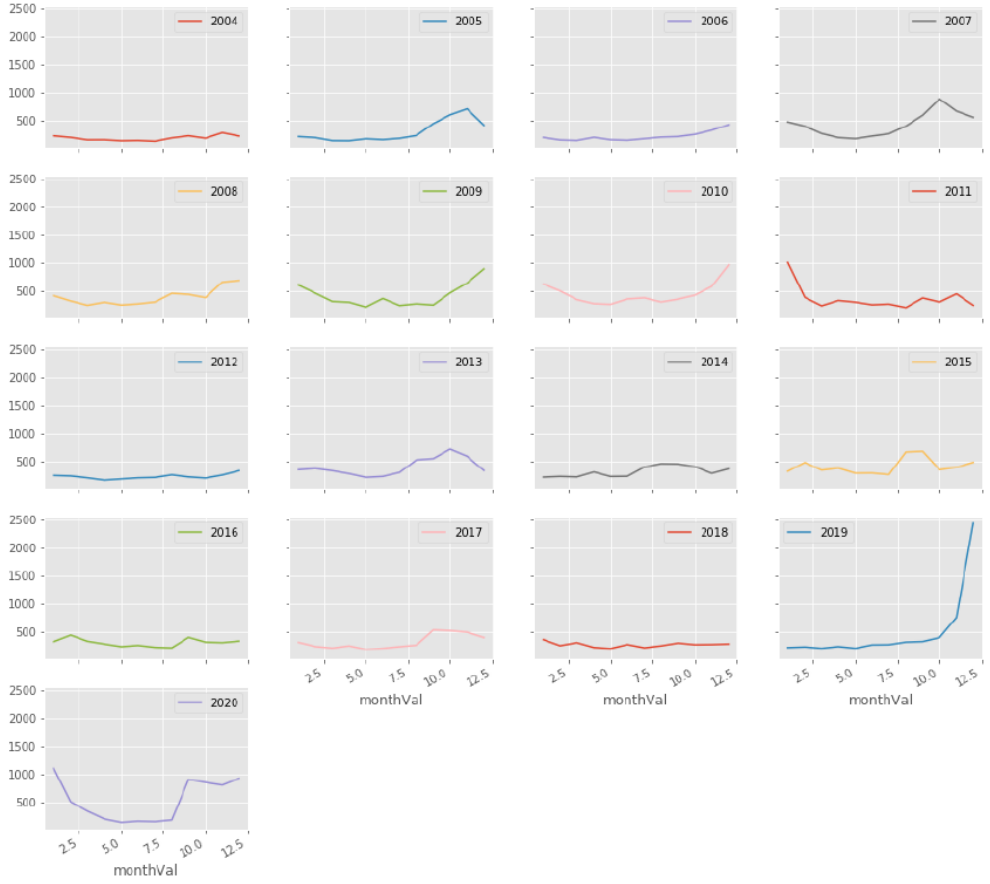
PIVOT TABLE MONTH-WISE

year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	201
monthVal														
1	221	208	193	470	412	600	620	1000	242	346	212	317	318	30
2	193	190	148	395	318	456	496	382	233	367	225	475	434	23
3	150	136	139	263	241	309	342	232	200	327	216	338	325	20
4	151	134	195	190	295	295	270	326	159	275	304	373	276	24
5	134	169	152	171	246	209	258	296	180	208	224	283	230	18
6	139	154	143	217	266	362	354	250	200	225	228	285	252	20
7	125	177	171	258	299	237	374	262	207	296	393	258	217	22
8	184	227	199	397	452	267	300	200	253	521	452	660	208	25
9	222	441	208	591	434	246	350	371	216	544	448	673	394	52
10	179	598	248	883	377	460	421	299	196	712	400	344	309	51
11	277	704	322	665	654	637	575	444	250	586	280	387	300	48
12	217	412	422	554	683	887	960	242	327	331	362	477	328	39

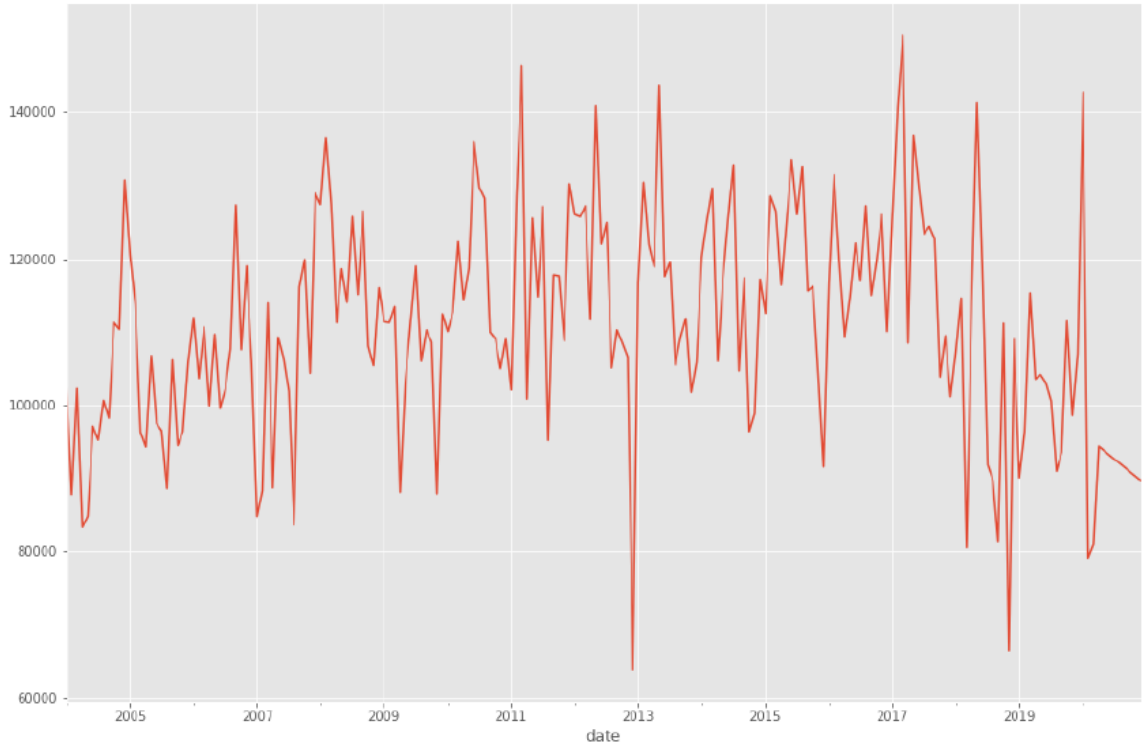
MONTH vs PRICE DIFFERENCE



PRICE TREND EVERY YEAR



DATE vs QUANTITY



Time series forecasting:

Time series analysis comprises methods for analyzing time-series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series are widely used for non-stationary data, like economics, weather, stock price, and retail sales in this post.

What measures can we check to see if the model is good?

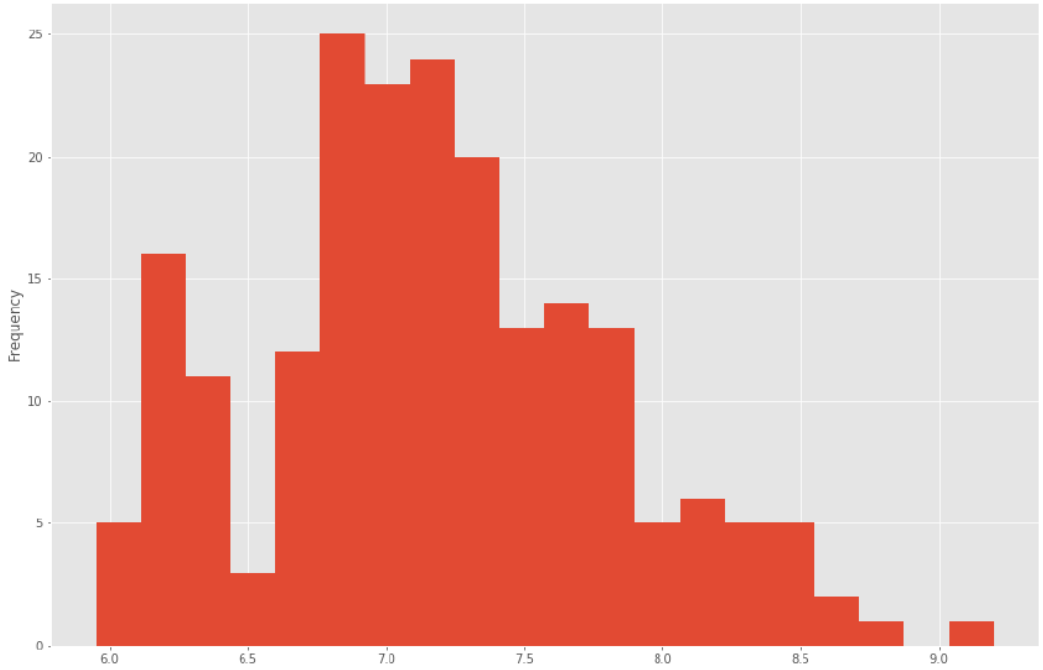
It is seen here (and also evident on the regression line plot, if you look closely) that the linear trend model has a tendency to make an error of the same sign for many periods in a row. This tendency is measured in statistical terms by the lag-1 autocorrelation and Durbin-Watson statistic. If there is no time pattern, the lag-1 autocorrelation should be very close to zero, and the Durbin-Watson statistic ought to be very close to 2, which is not the case here. If the model has succeeded in extracting all the "signal" from the data, there should be no pattern at all in the errors: the error in the next period should not be correlated with any previous errors. The linear trend model obviously fails the autocorrelation test in this case.

- Durbin Watson statistic is a test for autocorrelation in a data set.
- The DW statistic always has a value between zero and 4.0.
- A value of 2.0 means there is no autocorrelation detected in the sample. Values from zero to 2.0 indicate positive autocorrelation and values from 2.0 to 4.0 indicate negative autocorrelation.

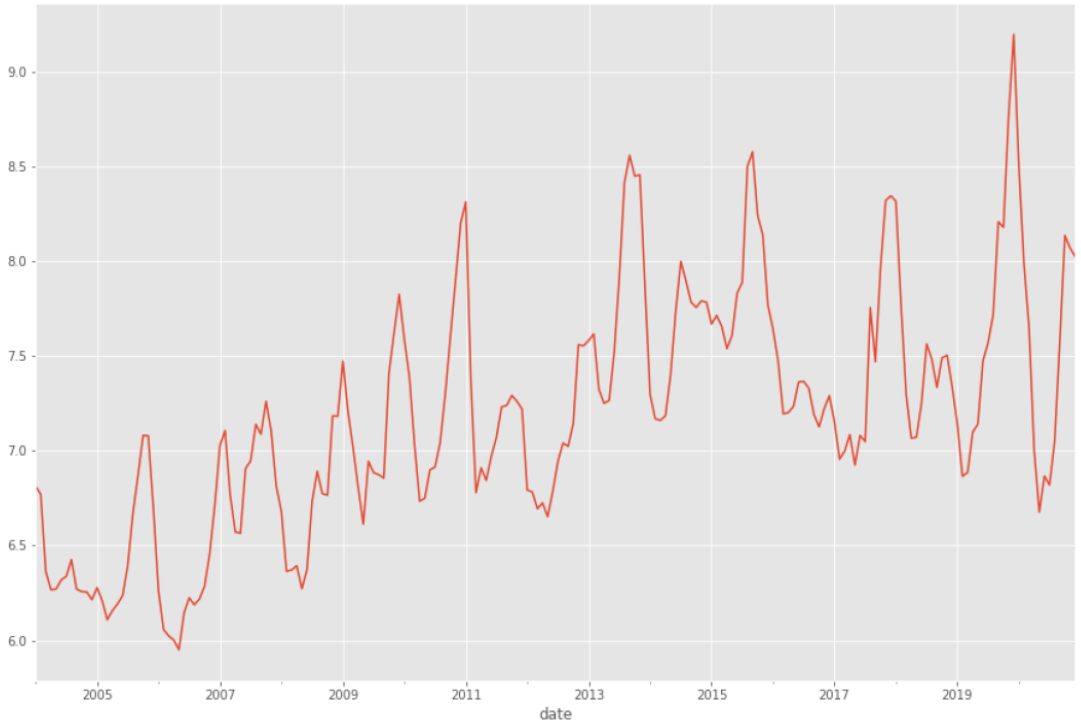
A stock price displaying positive autocorrelation would indicate that the price yesterday has a positive

correlation on the price today—so if the stock fell yesterday, it is also likely that it falls today. A security that has a negative autocorrelation, on the other hand, has a negative influence on itself over time—so that if it fell yesterday, there is a greater likelihood it will rise today.

LOGGED PRICE DISTRIBUTION



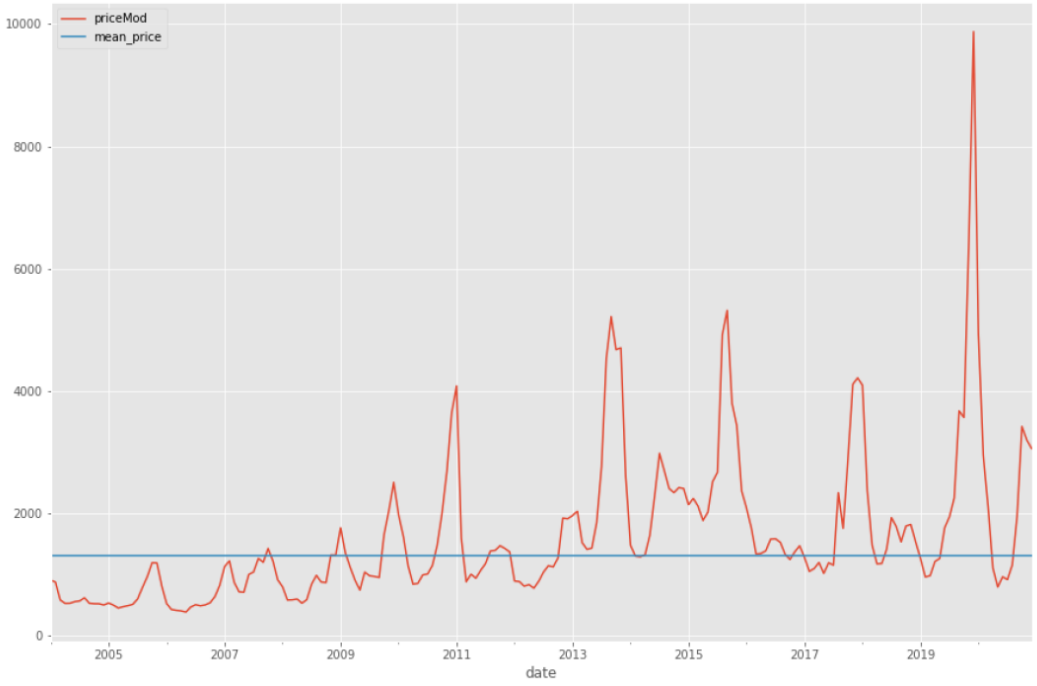
LOGGED PRICE vs DATE



CHAPTER 5

MODEL OUTCOMES

MEAN CONSTANT MODEL



OLS REGRESSION RESULTS

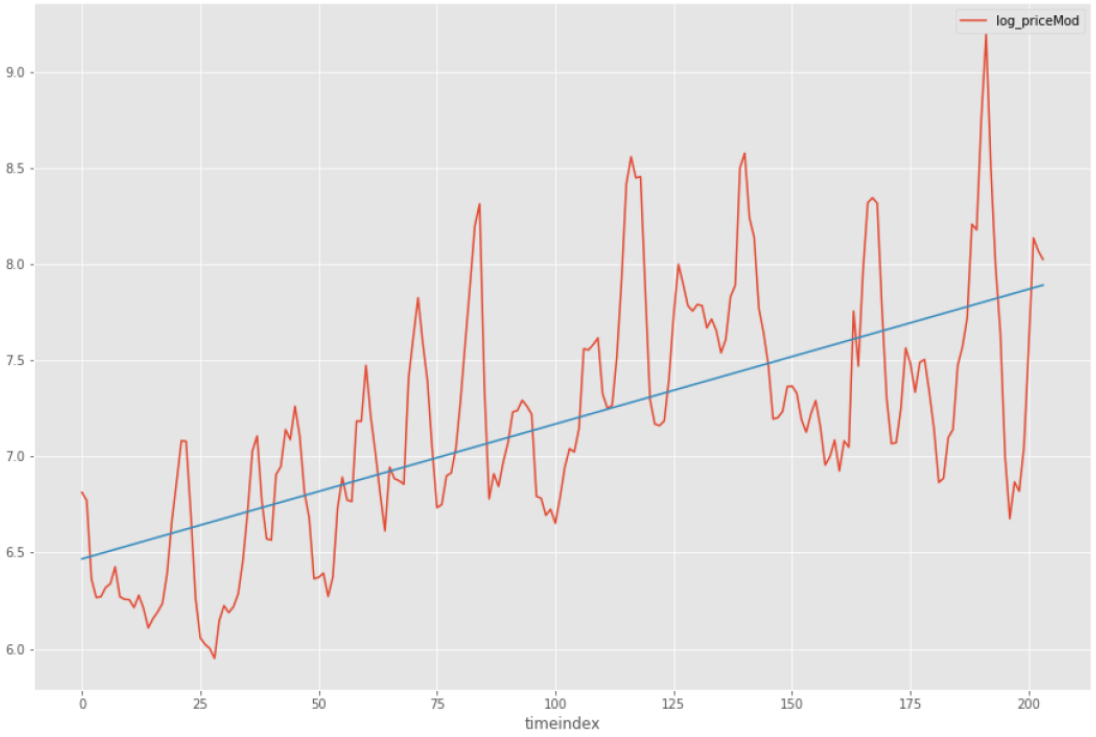
OLS Regression Results

Dep. Variable:	log_priceMod	**R-squared:**	0.431
Model:	OLS	**Adj. R-squared:**	0.428
Method:	Least Squares	**F-statistic:**	153.2
Date:	Sun, 20 Dec 2020	**Prob (F-statistic):**	1.50e-26
Time:	16:55:15	**Log-Likelihood:**	-137.36
No. Observations:	204	**AIC:**	278.7
Df Residuals:	202	**BIC:**	285.4
Df Model:	1		
Covariance Type:	nonrobust		

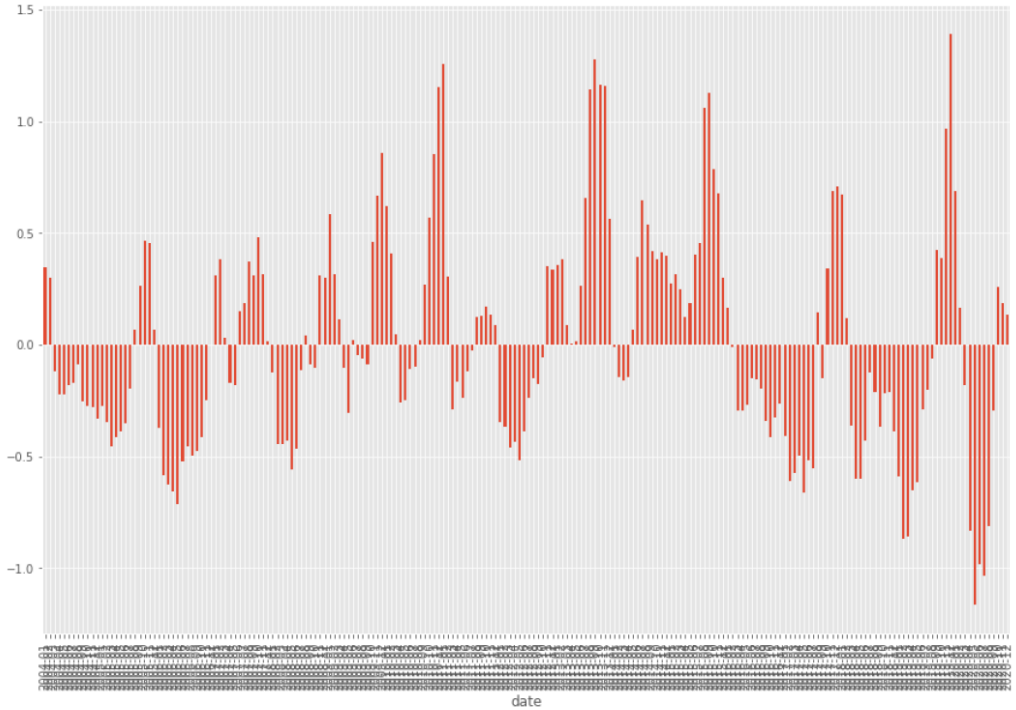
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.4679	0.067	97.232	0.000	6.337	6.599
timeindex	0.0070	0.001	12.376	0.000	0.006	0.008

Omnibus:	9.357	**Durbin-Watson:**	0.276
Prob(Omnibus):	0.009	**Jarque-Bera (JB):**	9.379
Skew:	0.512	**Prob(JB):**	0.00919
Kurtosis:	3.232	**Cond. No.**	234.

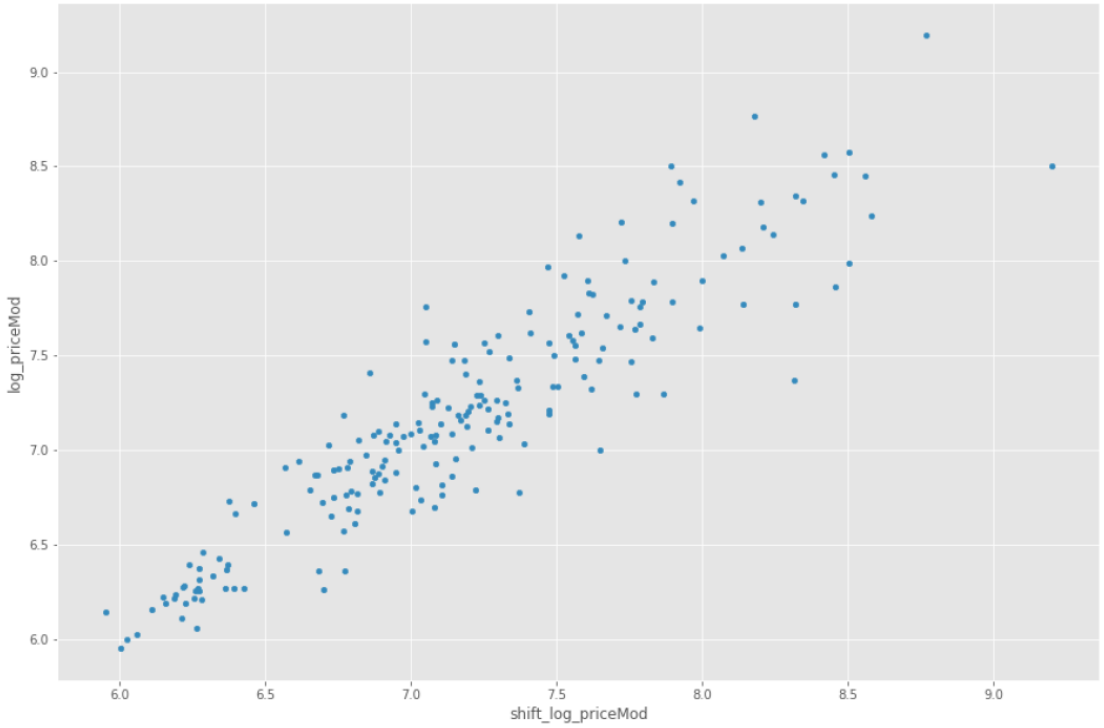
LINEAR TREND MODEL



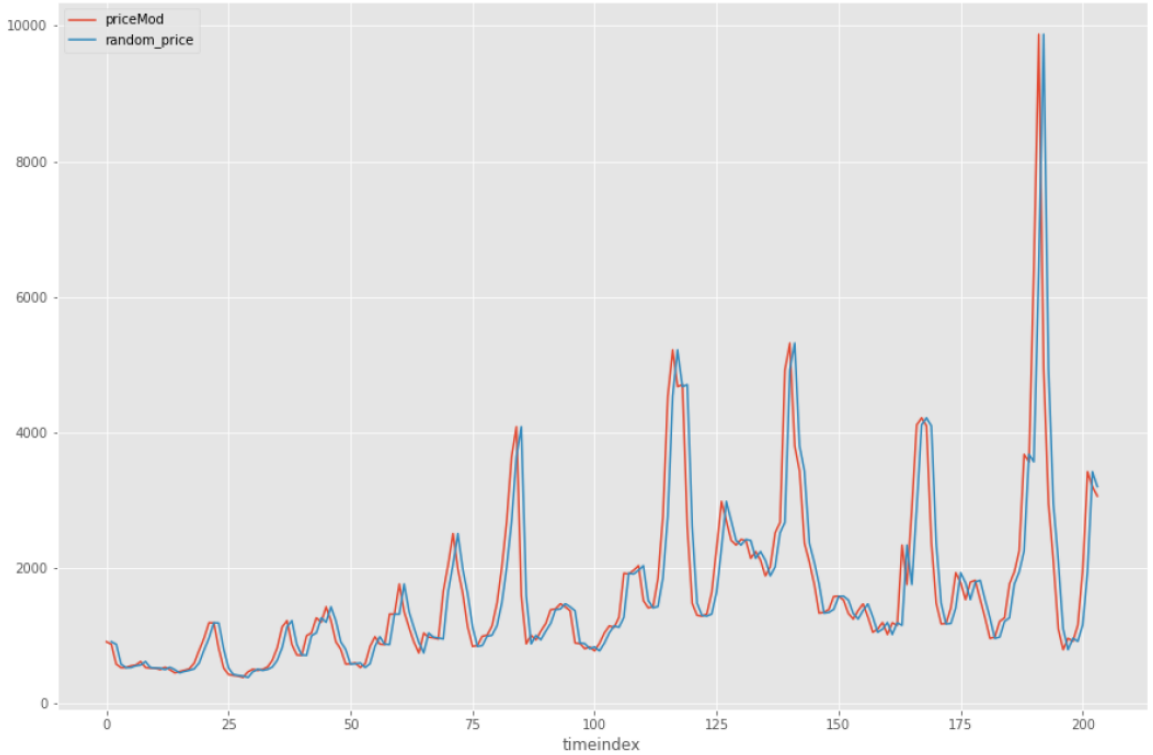
RESIDUAL PLOT



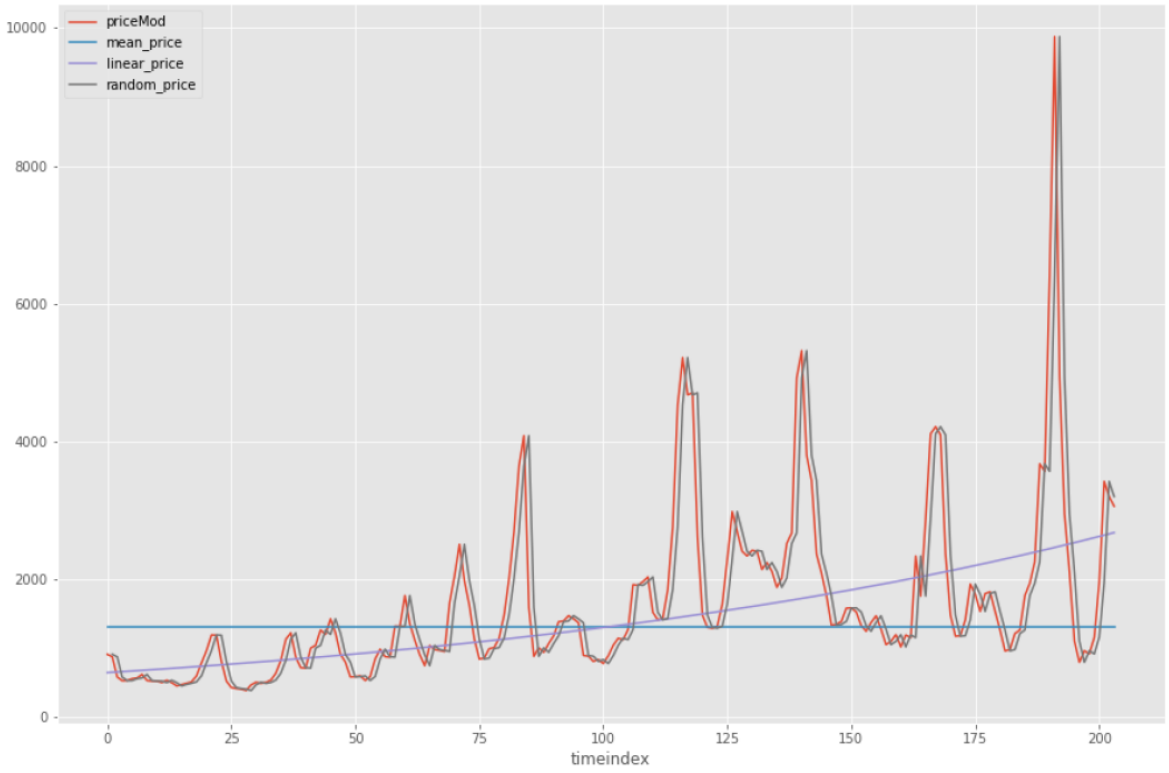
LOGGED vs SHIFT LOGGED PRICE



RANDOM WALK MODEL



RESULTS OF EVERY 3 MODELS



FORECASTED OUTPUT

S.No.	Model	Actual Price (Rs/q)	Predicted Price (Rs/q)
1	Mean Model	3060	1312.94
2	Linear Model	3060	3074.81
3	Random walk Model	3060	3200

CHAPTER 6

6.1. NEXT PHASE:

- Develop a supply chain model with data collected and analyzed in phase 1.
- Optimize the supply chain by finding correlations between cost and climatic conditions
- Identify the reason behind the constant fluctuation in the Onion price
- Performance measurement using Supply chain measurement indexes.
- Strategies to improve and make the current supply chain efficient.

CHAPTER 7

7.1. REFERENCE:

1. PRIYANKA KUMBHAR, NAMRATA GAIDOLE and DR. HEMANT SHARMA, "Price forecasting and Seasonality of Soybean in Amravati District of Maharashtra India", Current Agriculture Research Journal, 2019, pg. 417-423.
2. Xuan Pham, Martin Stack, "How data analytics is transforming agriculture", Business Horizons, 2018, 125-133.
3. Bongsug (Kevin) Chaea , David Olsonb and Chwen Sheua, "The impact of supply chain analytics on operational performance: a resource-based view" International Journal of Production Research, 2013.
4. George Baryannisa, Samir Dani, Grigoris Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability", Future Generation Computer Systems, 2019, 993-1004.
5. Rohit Sharma, Anjali Shishodia,Sachin Kamble, "Agriculture supply chain risks and COVID-19: mitigation strategies and implications for the practitioners", International Journal of Logistics Research and Applications, 2020.
6. D. J. Bartholomew, "Time Series Analysis Forecasting and Control", Journal of the Operational Research Society, 2017, 199-201.
7. Real Carbonneau, Kevin Laframboise, Rustam Vahidov, "Application of machine learning techniques for supply chain demand forecasting", European Journal of Operational Research, 2008, 1140-1154.
8. Tan Xuewen Du Zhixiong , "Postmodern Agriculture: From an Approach of Sustainable Food Supply Chain [J]" , Journal of China Agricultural University (Social Sciences Edition) 1, 2010
9. Rohit Sharma, Sachin S Kamble, Angappa Gunasekaran, Vikas Kumar, Anil Kumar, " A systematic literature review on machine learning applications for sustainable agriculture supply chain performance",Computers & Operations Research, 104926, 2020
10. Liu Qiao, Shen Xin, Sun Xu , "Research on risk evaluation of agriculture product supply chain [J]",Journal of Agricultural Mechanization Research 9, 2011