2020

Natural Gas Price Forecasting For 2020 & 2021

United States

Suvam Bit

B.SC-VI

DEPARTMENT OF STATISTICS

SIKSHA BHAVANA

VISVA BHARATI UNIVERSITY

CONTENT

ACKNOWLEDGEMENT

In the accomplishment of this project successfully, many people have best owned upon me their blessings and the heart pledged support, this time I’m utilizing to thank all the people who have been concerned with this project.

Primarily I would like to thank God for being able to complete this project with success. Then I would like to thank Prof. Saran Ishika Maiti whose valuable guidance has been the ones that helped me patch this project and make it full proof success. Her suggestions and instructions have served as the major contributor towards the completion of the project. Moreover, I would like to thank Mr. Krish Naik whose YouTube videos helped me a lot to proceed through this project work.

Then I would like to thank my parents and friends who have helped me with their valuable suggestions and guidance has been very helpful in various phases of completion of the project.

Last but not the least I would like to thank my classmates who have helped me a lot.

INTRODUCTION

WHAT IS NATURAL GAS?

Natural gas is a **fossil energy source** that formed deep beneath the earth's surface. Natural gas contains many different compounds. The largest component of natural gas is **methane**, a compound with one carbon atom and four hydrogen atoms (CH4). Natural gas also contains smaller amounts of natural gas liquids (NGL, which are also **hydrocarbon gas liquids**), and nonhydrocarbon gases, such as **carbon dioxide** and **water vapor**. We use natural gas as a **fuel** and to make materials and chemicals.

HOW DID NATURAL GAS FORM?

Millions to hundreds of millions of years ago and over long periods of time, the remains of plants and animals (such as diatoms) built up in **thick layers** on the **earth's surface** and **ocean floors**, sometimes mixed with sand, silt, and calcium carbonate. Over time, these layers were buried under sand, silt, and rock. **Pressure and heat** changed some of this carbon and hydrogen-rich material into coal, some into oil (petroleum), and some into **natural gas**.

WHERE IS NATURAL GAS FOUND?

In some places, natural gas moved into **large cracks and spaces between layers** of overlying rock. The natural gas found in these types of formations is sometimes called **conventional natural gas**. In other places, natural gas occurs in the tiny pores (spaces) within some formations of **shale**, **sandstone**, and other types of **sedimentary rock**. This natural gas is referred to as **shale gas** or **tight gas**, and it is sometimes called **unconventional natural gas**. Natural gas also occurs with deposits of crude oil, and this natural gas is called **associated natural gas**. Natural gas deposits are found on land, and some are offshore and deep under the ocean floor. A type of natural gas found in coal deposits is called **coalbed methane**.

HOW DO WE FIND NATURAL GAS?

The search for natural gas begins with **geologists** who study the structure and processes of the earth. They locate the types of geologic formations that are likely to contain natural gas deposits.

Geologists often use **seismic surveys** on land and in the ocean to find the right places to drill natural gas and oil wells. Seismic surveys create and measure seismic waves in the earth to get information on the geology of rock formations. Seismic surveys on land may use a **thumper truck**, which has a vibrating pad that pounds the ground to create seismic waves in the underlying rock. Sometimes small amounts of **explosives** are used. Seismic surveys conducted in the ocean use blasts of sound that create sonic waves to explore the geology beneath the ocean floor.

If the results of seismic surveys indicate that a site has potential for producing natural gas, an exploratory well is drilled and tested. The results of the test provide information on the quality and quantity of natural gas available in the resource.

DRILLING NATURAL GAS WELLS AND PRODUCING NATURAL GAS

If the results from a test well show that a geologic formation has enough natural gas to produce and make a profit, one or more production (or development) wells are drilled. Natural gas **wells can be drilled vertically and horizontally into natural gas-bearing formations**. In conventional natural gas deposits, the natural gas generally flows easily up through wells to the surface.

In the United States and in a few other countries, natural gas is produced from shale and other types of sedimentary rock formations by **forcing water**, **chemicals**, and **sand** down a well under **high pressure**. This process, called **hydraulic fracturing or fracking**, and sometimes referred to as unconventional production, breaks up the formation, releases the natural gas from the rock, and allows the natural gas to flow to and up wells to the surface. At the top of the well on the surface, natural gas is put into **gathering pipelines** and sent to natural gas processing plants.

NATURAL GAS IS PROCESSED FOR SALE AND CONSUMPTION

Natural gas withdrawn from natural gas or crude oil wells is called **wet natural gas** because, along with methane, it usually contains **NGL—ethane, propane, butanes, and pentanes—and water vapor**. Wellhead natural gas may also contain nonhydrocarbons such as **sulphur, helium**, **nitrogen**, **hydrogen sulphide**, and **carbon dioxide**, most of which must be removed from natural gas before it is sold to consumers.

From the wellhead, natural gas is sent to processing plants where water vapor and nonhydrocarbon compounds are removed and NGL are separated from the wet gas and sold separately. Some ethane is often left in the processed natural gas. The separated NGL are called **natural gas plant liquids (NGPL)**, and the processed natural gas is called **dry**, **consumer-grade**, **or pipeline quality natural gas**. Some wellhead natural gas is sufficiently dry and satisfies pipeline transportation standards without processing. Chemicals called **odorants** are added to natural gas so that leaks in natural gas pipelines can be detected. Dry natural gas is sent through pipelines to underground storage fields or to distribution companies and then to consumers.

In places where natural gas pipelines are not available to take away associated natural gas produced from oil wells, the natural gas may be reinjected into the **oil-bearing formation**, or it may be **vented** or **burned** (flared). Reinjecting unmarketable natural gas can help to maintain pressure in oil wells to improve oil production.

Coalbed methane can be extracted from coal deposits before or during coal mining, and it can be added to natural gas pipelines without any special treatment.

Most of the natural gas consumed in the United States is produced in the United States. Some natural gas is imported from **Canada** and **Mexico** in pipelines. A small amount of natural gas is also imported as liquefied natural gas.

WHERE NATURAL GAS IS USED

NATURAL GAS USED BY U.S. CONSUMING SECTORS BY AMOUNT AND SHARE OF TOTAL U.S. NATURAL GAS CONSUMPTION IN 2018

The United States used about **30 trillion cubic feet (Tcf)** of natural gas in 2018 which is 31% of total U.S. primary energy consumption.

35%
10.59 Tcf
Electric Power

33%
10.06 Tcf
Industrial

17%
4.99 Tcf
Residential

12%
3.52 Tcf
Commercial

3%
0.91 Tcf
Transportation

THE FIVE LARGEST NATURAL GAS CONSUMING STATES AND THEIR SHARE OF TOTAL U.S. NATURAL GAS CONSUMPTION IN 2018

Natural gas is used through out the United States, but **five states** accounted for about **37%** of total U.S. natural gas consumption in 2018.

14.7%
Texas

7.1%
California

5.8%
Louisiana

4.9%
Florida

4.8%
Pennsylvania

PROJECT OBJECTIVES

The main objective of this project is to **propose a time series model** and thereby **forecast the prices of natural gas in 2020 and 2021 in United States**. We'll do the forecasting with the reference of previous 23 years' prices of natural gas in United States. We analyse the time series data with various statistical methods and tools of Time Series Analysis. All are done by **python programming**.

IMPORTANCE

It is important for policy purposes to forecast natural gas prices accurately. Many policy decisions such as **regulatory action in technologies** (i.e., appliance efficiency standards), **use of public lands** (i.e., mineral exploration) and **foreign policy** may all be influenced by changing expectations about natural gas prices. For example, **AEO (Authorized Economic Operator)** natural gas price forecasts have consistently **overestimated** the price every year since 1982, never falling below 68%. The inaccuracy of these forecasts has encouraged economists and energy policy makers to seek **alternative methods to forecast natural gas prices**.

The natural gas futures market provides an alternative forecast of natural gas prices determined by the interaction of numerous **buyers** and **sellers** in the natural gas market. Natural gas futures markets were set up in 1989 to help insure buyers against the risk of energy price fluctuations. In theory, futures market prices summarize privately available information about natural gas supply and demand. As a result, some economists believe that prices determined in the futures markets provide accurate price forecasts.

Policy makers can turn to two primary sources of information about future natural gas prices – **price forecasts based upon economic models of energy supply and demand** and **price forecasts derived from the natural gas futures market**. Historically, policy makers have relied almost exclusively upon forecasts based upon economic models.

THE DATA

The data shown here is a time series of major Natural Gas Prices including US Henry Hub. Data is collected from **U.S. Energy Information Administration (EIA)**. Dataset contains **monthly prices** of natural gas, starting **from January 1997 to December 2019**. Prices are in nominal dollars. The price unit is in terms of **dollars per million British thermal unit (MMBtu)**.

Source: https://datahub.io/core/natural-gas/eia.gov

NATURAL GAS PRICES IN UNITED STATES

Month	Price
1997-01	3.45
1997-02	2.15
1997-03	1.89
1997-04	2.03
1997-05	2.25
1997-06	2.2
1997-07	2.19
1997-08	2.49
1997-09	2.88
1997-10	3.07
1997-11	3.01
1997-12	2.35
1998-01	2.09
1998-02	2.23
1998-03	2.24
1998-04	2.43
1998-05	2.14
1998-06	2.17
1998-07	2.17
1998-08	1.85
1998-09	2.02
1998-10	1.91
1998-11	2.12
1998-12	1.72
1999-01	1.85
1999-02	1.77
1999-03	1.79
1999-04	2.15
1999-05	2.26
1999-06	2.3
1999-07	2.31
1999-08	2.8
1999-09	2.55
1999-10	2.73
1999-11	2.37
1999-12	2.36
2000-01	2.42
2000-02	2.66
2000-03	2.79
2000-04	3.04
2000-05	3.59
2000-06	4.29
2000-07	3.99
2000-08	4.43
2000-09	5.06
2000-10	5.02
2000-11	5.52
2000-12	8.9
2001-01	8.17
2001-02	5.61
2001-03	5.23
2001-04	5.19
2001-05	4.19
2001-06	3.72
2001-07	3.11
2001-08	2.97
2001-09	2.19
2001-10	2.46

Month	Price
2001-11	2.34
2001-12	2.3
2002-01	2.32
2002-02	2.32
2002-03	3.03
2002-04	3.43
2002-05	3.5
2002-06	3.26
2002-07	2.99
2002-08	3.09
2002-09	3.55
2002-10	4.13
2002-11	4.04
2002-12	4.74
2003-01	5.43
2003-02	7.71
2003-03	5.93
2003-04	5.26
2003-05	5.81
2003-06	5.82
2003-07	5.03
2003-08	4.99
2003-09	4.62
2003-10	4.63
2003-11	4.47
2003-12	6.13
2004-01	6.14
2004-02	5.37
2004-03	5.39
2004-04	5.71
2004-05	6.33
2004-06	6.27
2004-07	5.93
2004-08	5.41
2004-09	5.15
2004-10	6.35
2004-11	6.17
2004-12	6.58
2005-01	6.15
2005-02	6.14
2005-03	6.96
2005-04	7.16
2005-05	6.47
2005-06	7.18
2005-07	7.63
2005-08	9.53
2005-09	11.75
2005-10	13.42
2005-11	10.3
2005-12	13.05
2006-01	8.69
2006-02	7.54
2006-03	6.89
2006-04	7.16
2006-05	6.25
2006-06	6.21
2006-07	6.17
2006-08	7.14
2006-09	4.9
2006-10	5.85
2006-11	7.41
2006-12	6.73
2007-01	6.55
2007-02	8
2007-03	7.11
2007-04	7.6
2007-05	7.64
2007-06	7.35
2007-07	6.22
2007-08	6.22
2007-09	6.08
2007-10	6.74
2007-11	7.1
2007-12	7.11
2008-01	7.99
2008-02	8.54
2008-03	9.41
2008-04	10.18
2008-05	11.27
2008-06	12.69
2008-07	11.09
2008-08	8.26
2008-09	7.67
2008-10	6.74
2008-11	6.68
2008-12	5.82
2009-01	5.24
2009-02	4.52

Month	Price		Month	Price
2009-03	3.96		2012-11	3.54
2009-04	3.5		2012-12	3.34
2009-05	3.83		2013-01	3.33
2009-06	3.8		2013-02	3.33
2009-07	3.38		2013-03	3.81
2009-08	3.14		2013-04	4.17
2009-09	2.99		2013-05	4.04
2009-10	4.01		2013-06	3.83
2009-11	3.66		2013-07	3.62
2009-12	5.35		2013-08	3.43
2010-01	5.83		2013-09	3.62
2010-02	5.32		2013-10	3.68
2010-03	4.29		2013-11	3.64
2010-04	4.03		2013-12	4.24
2010-05	4.14		2014-01	4.71
2010-06	4.8		2014-02	6
2010-07	4.63		2014-03	4.9
2010-08	4.32		2014-04	4.66
2010-09	3.89		2014-05	4.58
2010-10	3.43		2014-06	4.59
2010-11	3.71		2014-07	4.05
2010-12	4.25		2014-08	3.91
2011-01	4.49		2014-09	3.92
2011-02	4.09		2014-10	3.78
2011-03	3.97		2014-11	4.12
2011-04	4.24		2014-12	3.48
2011-05	4.31		2015-01	2.99
2011-06	4.54		2015-02	2.87
2011-07	4.42		2015-03	2.83
2011-08	4.06		2015-04	2.61
2011-09	3.9		2015-05	2.85
2011-10	3.57		2015-06	2.78
2011-11	3.24		2015-07	2.84
2011-12	3.17		2015-08	2.77
2012-01	2.67		2015-09	2.66
2012-02	2.51		2015-10	2.34
2012-03	2.17		2015-11	2.09
2012-04	1.95		2015-12	1.93
2012-05	2.43		2016-01	2.28
2012-06	2.46		2016-02	1.99
2012-07	2.95		2016-03	1.73
2012-08	2.84		2016-04	1.92
2012-09	2.85		2016-05	1.92
2012-10	3.32		2016-06	2.59

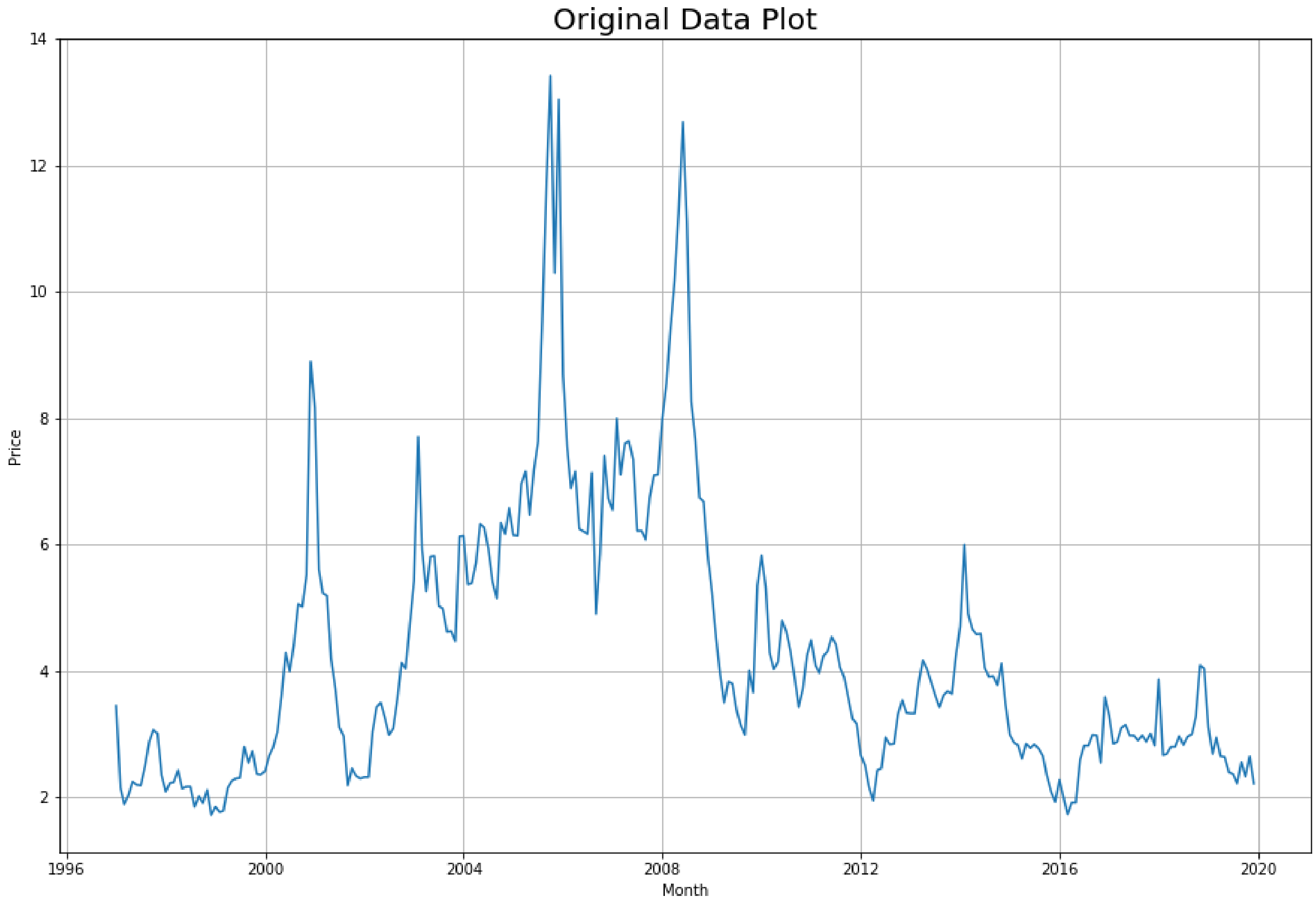
Month	Price
2016-07	2.82
2016-08	2.82
2016-09	2.99
2016-10	2.98
2016-11	2.55
2016-12	3.59
2017-01	3.3
2017-02	2.85
2017-03	2.88
2017-04	3.1
2017-05	3.15
2017-06	2.98
2017-07	2.98
2017-08	2.9
2017-09	2.98
2017-10	2.88
2017-11	3.01
2017-12	2.82
2018-01	3.87
2018-02	2.67
2018-03	2.69
2018-04	2.8
2018-05	2.8
2018-06	2.97
2018-07	2.83
2018-08	2.96
2018-09	3
2018-10	3.28
2018-11	4.09
2018-12	4.04
2019-01	3.11
2019-02	2.69
2019-03	2.95
2019-04	2.65
2019-05	2.64
2019-06	2.4
2019-07	2.37
2019-08	2.22
2019-09	2.56
2019-10	2.33
2019-11	2.65
2019-12	2.22

GRAPHICAL REPRESENTATION OF THE DATA

The data, we have in hand is a clean data i.e. there's no missing value present in the data. So we don't need to do data cleansing.

Therefore, the first job before analysing a time series data is to plot the raw data in the graph and draw some initial interpretations viewing the original data plot.

NATURAL GAS PRICES DATA PLOT



INITIAL INTERPRETATIONS

From the graph of the data we can make following initial interpretations

- The prices are suddenly shot up in 2000-2001 due the energy crisis in US.
- The prices in 2005 had a historical rise due to the Hurricanes Rita and Katrina.
- The Prices in 2008 had risen significantly due the financial crisis in US.
- As per as outliers is concerned, we see there is no outliers present in our data.
- We've to look for the sudden shifts in the graph and we see no such prominent sudden shift occurred in the data.
- By viewing from naked eye there's an upward trend at the middle of the graph and then it goes down. However, we'll do trend analysis later on.
- Out of rough conception we see there may be a slight seasonality present in the data, though it's not that much prominent.

DATA ANALYSIS

Time series analysis comprises methods for analysing 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.

Now there two types of model to express a time series observation which are

I. Additive Model
II. Multiplicative Model

In this project we adopt the **multiplicative model** to express the time series variable i.e. Price. We propose multiplicative model because the seasonal pattern increases or decreases as the trend increases or decreases.

MULTIPLICATIVE MODEL

If we've reasons to assume that the various components of a time series operate proportionally to the general level of the series, the traditional or classical model is appropriate. According to multiplicative model,

$$Y_t = T_t \times S_t \times C_t \times I_t$$

Where T_t, S_t, C_t and R_t are the trend component, seasonal component, cyclical component and irregular components respectively.

STATISTICAL TOOLS AND METHODS

Now we'll use a number of statistical tools and methods to analyse our data. The tools we're going to use are as follows,

I. Augmented Dickey Fuller Test
II. Moving Average Trend
III. Decomposition of the Data
IV. Holt-Winters Exponential Smoothing
V. Autocorrelation and Partial Autocorrelation Function
VI. Fitting Model
VII. Forecasting

I. AUGMENTED DICKEY-FULLER TEST

In statistics, an augmented Dickey-Fuller test (ADF) is used to **check the stationarity** of a time series. Here, the **null hypothesis** is that a unit root is present in a time series sample i.e. the time series is **non-stationary**. The **alternative hypothesis** is the time series is **stationary**. It is an augmented version of Dickey-Fuller test for a larger and more complicated set of time series models.

What is Stationarity?

Stationarity means that the statistical properties of a process generating a time series do not change over time.

The augmented Dickey-Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

METHOD

The testing procedure for the ADF test is the same as for the Dickey-Fuller test but it is applied to the model,

$$\Delta y_t = \alpha + \beta t + \gamma y_t + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta\, y_{t-p+1}$$

Where α is a constant, β the coefficient on time trend and p the lag order of the autoregressive process. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraint $\beta = 0$ corresponds to modeling a random walk with a drift.

By including lags of the order p, the ADF formulation allows for higher-order autoregressive processes. This means that the lag length p has to be determined when applying the test. One possible approach is to test down from high orders and examine the t-values on coefficients. An alternative approach is to examine information criteria such as the Akaike Information Criterion, Bayesian Information Criterion or the Hannan-Quinn Information Criterion.

The unit root test is then carried out under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. Once a value for the test statistic,

$$DF_\tau = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey–Fuller test. As this test is asymmetrical, we are only concerned with negative values of our test statistic $DF\tau$. If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of $\gamma = 0$ is rejected and no unit root is present.

ANALYSIS

Performing the augmented Dickey-Fuller test on our data we got the following results,

- ADF statistic: -2.063891
- p-value: 0.259280
- Critical value at 5% level of significance: -2.872

Since at 5% level of significance the ADF statistic is greater than the critical value, we fail to reject our null hypothesis.
Therefore, our time series data is **non-stationary**.

INTERPRETATION

Since our data is not stationary, it is necessary to make the data stationary for further analysis.

MAKING THE DATA STATIONARY

Now it's important to make the time series data stationary because it helps to identify the driving factors: when we detect a change in a time series, we may be able to infer a correlation. But we need the time series to be stationary, otherwise the correlation we find will be misleading.

To make our time series data stationary we take the first forward differences of prices as follows,

$$\Delta y_t = y_t - y_{t-1}$$

Now we'll again perform augmented Dickey-Fuller test on the observations of first forward differences of prices.

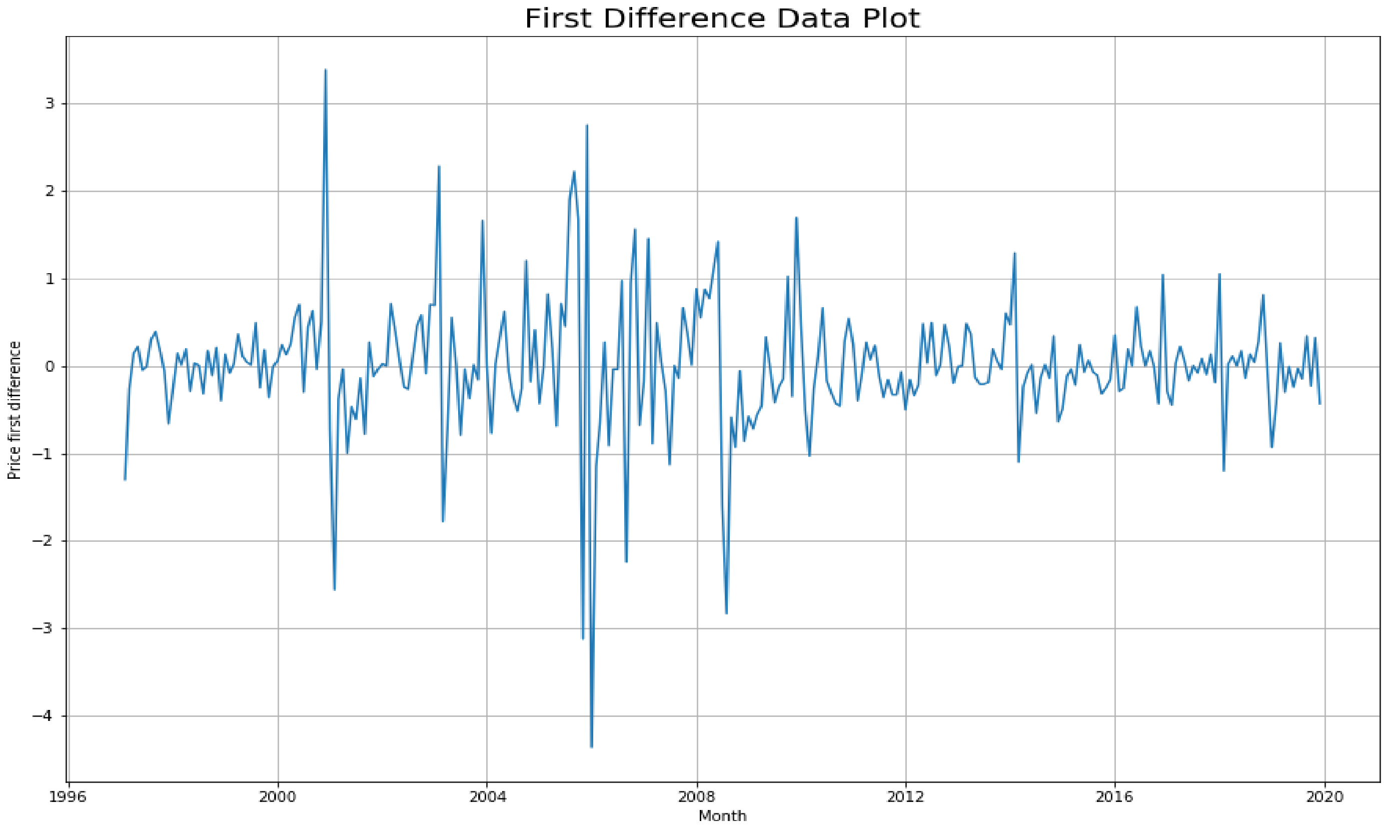
ANALYSIS

Performing the augmented Dickey-Fuller test on the observations of first forward differences we got the following results,

- ADF statistic: -7.457893
- p-value: 0.000000
- Critical value at 5% level of significance: -2.872

Since at 5% level of significance the ADF statistic is very less than the critical value, we reject our null hypothesis.
Therefore, the first forward difference of prices is **stationary**.

GRAPHICAL REPRESENTATION OF STATIONARY DATA



INTERPRETATION

So, we can use this stationary data for determining correlations in model fitting. From the graph we can also see that our data becomes **free from upward or downward trend**.

II. MOVING AVERAGE TREND

It consists of measurement of trend by smoothing out the fluctuations of the data by means of moving average. Moving averages smooth the price data to form a trend following indicator. They do not predict price direction, but rather define the current direction, though they lag due to being based on past prices. Despite this, moving averages smooth price action and filter out the noise. The two popular types of moving averages are the Simple Moving Average (SMA) and the Exponential Moving average (EMA).

To analyse the trend of our data we're going to use the Simple Moving Average (SMA).

METHOD

Since our price data is a monthly time series data, we'll perform a 12-point moving average on our data.

Moving average of period 12 (uncentred) is a series of successive averages of 12 terms at a time, starting with 1st, 2nd, 3rd term etc. Thus, the first average is the mean of the 1st 12 terms; the 2nd is the mean of the 12 terms from 2nd to 13th term, the 3rd is the mean of the 12 terms from 3rd to 14th term and so on.

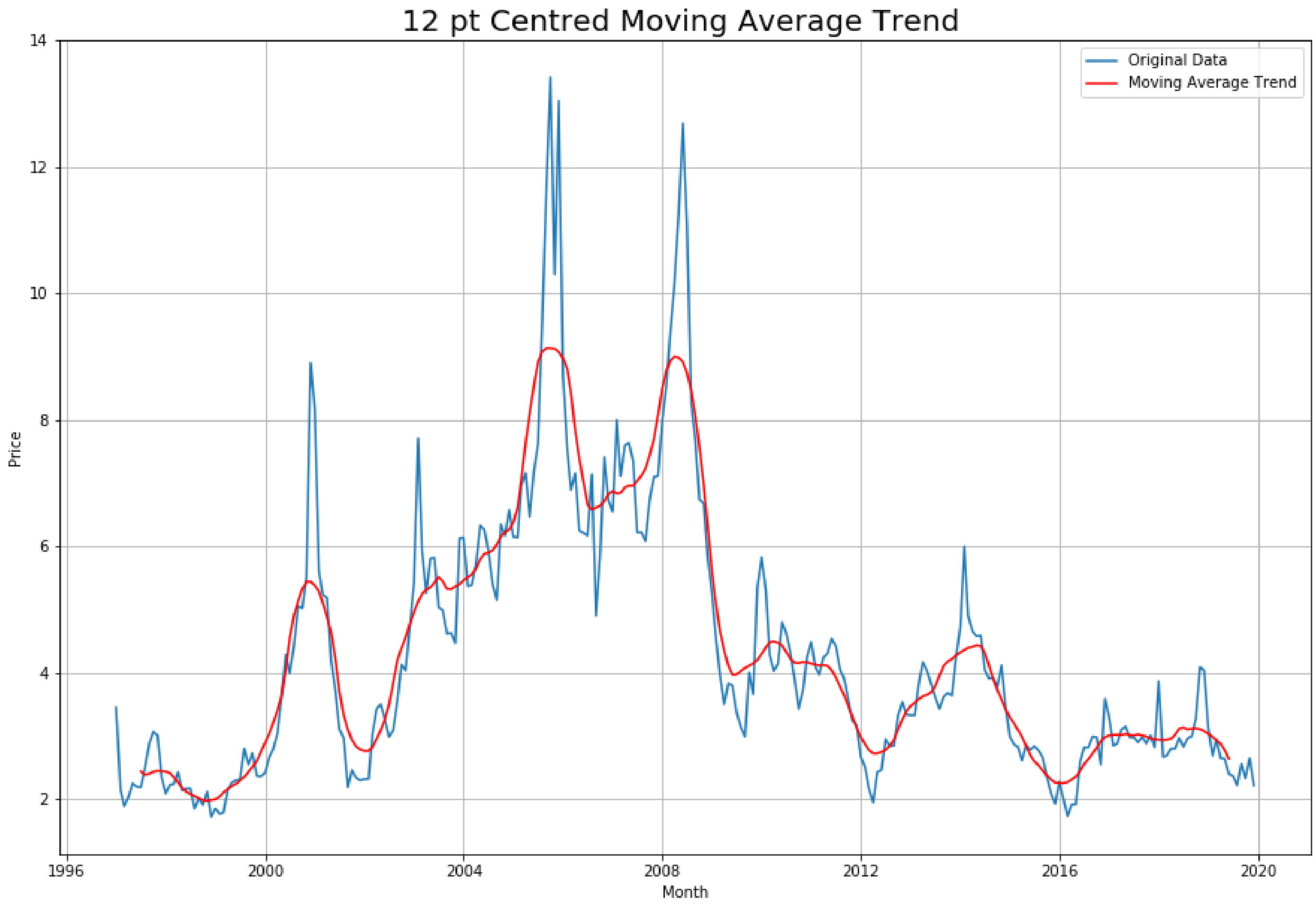
$$y^*_{ma_{t+5}} = \frac{y_t + y_{t+1} + y_{t+2} + \cdots + y_{t+11}}{12}$$

Moving average of period 12 (centred) is a series of successive averages of 2 terms at a time, starting with 1st, 2nd, 3rd term etc. Thus, the first average is the mean of the 1st and 2nd term; the 2nd is the mean of the 2nd and 3rd term, the 3rd is the mean of the 3rd and 4th term and so on.

$$y_{ma_{t+2}} = \frac{y^*_{ma_{t+1}} + y^*_{ma_{t+2}}}{2}$$

ANALYSIS

Performing 12-point moving average on our price data and we got the following trend line below:



INTERPRETATION

From the above moving average trend graph we can see that the first half of the data (till 2009) has a general upward trend and the second half of the data (from 2009), a general downward trend is observed. Due to moving average trend, the irregular components are also washed out.

III. DECOMPOSITION OF THE DATA

Time series data exhibit a variety of patters, and it is often helpful to split a time series into several components, each representing an underlying pattern category.

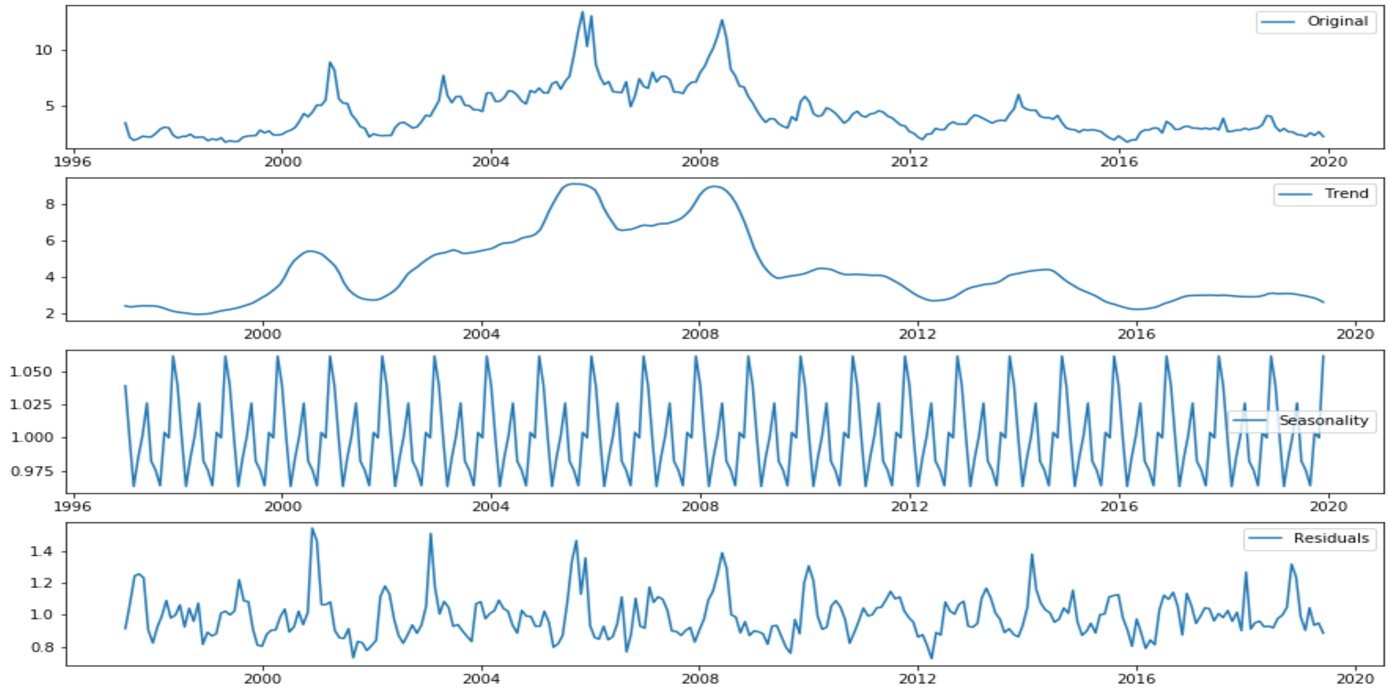
We discussed earlier that there're three types of time series patterns: trend (T_t), seasonality (S_t) and cycles (C_t). When we decompose a time series into components, we usually combine the trend and cycle into a single trend-cycle component (sometimes called the trend for simplicity). Thus, we think of a time series as comprising three components: a trend-cycle component, a seasonal component and a remainder component (containing anything else in the time series, I_t). Our multiplicative model of time series may look like as earlier

$$Y_t = T_t \times S_t \times C_t \times I_t$$

Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

ANALYSIS

After decomposition of our data we got the following results:



INTERPRETATION

- The trend line, which is decomposed in the 2nd figure, is alike the trend line we got from moving averages.
- After decomposition, we can see the seasonal components in the 3rd figure and we can say that our data has seasonality in high frequency.
- Due to decomposition, the irregular components in our data are evident in the 4th figure.

IV. HOLT-WINTERS EXPONENTIAL SMOOTHING

Holt-Winters forecasting is an exponential smoothing technique used to predict the behaviour of a sequence of values over time-a time series. Holt-Winters is **one of the most popular forecasting techniques** for time series.

Forecasting always requires a model three aspects of the time series: **a typical value (average)**, **a slope (trend) over time** and **a cyclical pattern (seasonality)**. Holt-Winters uses exponential smoothing to encode lots of values from past and use them to predict "typical" values for present and future.

METHOD

Suppose we have a sequence of observations $\{x_t\}$, beginning at time $t = 0$ with a cycle of seasonal change of length L.

The method calculates a trend line for the data as well as seasonal indices that weight the values in the trend line based on where that time point falls in the cycle of length L.

$\{s_t\}$ represents the smoothed value of the constant part for time t. $\{b_t\}$ represents the sequence of best estimates of the linear trend that are superimposed on the seasonal changes. $\{c_t\}$ is the sequence of seasonal correction factors. c_t is the expected proportion of the predicted trend at any time t *mode* L in the cycle that the observations take on. As a rule of thumb, a minimum of two full seasons (or $2L$ period) of historical data is needed to initialize a set of seasonal factors.

The algorithm is again written as F_{t+m}, an estimate of the value of x at time t+m, m>0 based on the raw data up to time t. Triple exponential smoothing with multiplicative seasonality is given by the formulas:

$$s_0 = x_0$$

$$s_t = \alpha \frac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$$

$$c_t = \gamma \frac{x_t}{s_t} + (1-\gamma)c_{t-L}$$

$$F_{t+m} = (s_t + mb_t)c_{t-L+1+(m-1)}$$

Where α is the data smoothing factor, 0 < α < 1, β is the trend smoothing factor, 0 < β < 1, and γ is the seasonal change smoothing factor, 0 < γ < 1.

The general formula for the initial trend estimate b_0 is:

$$b_0 = \frac{1}{L}\left(\frac{x_{L+1} - x_1}{L} + \frac{x_{L+2} - x_2}{L} + \cdots + \frac{x_{L+L} - x_L}{L}\right)$$

ANALYSIS

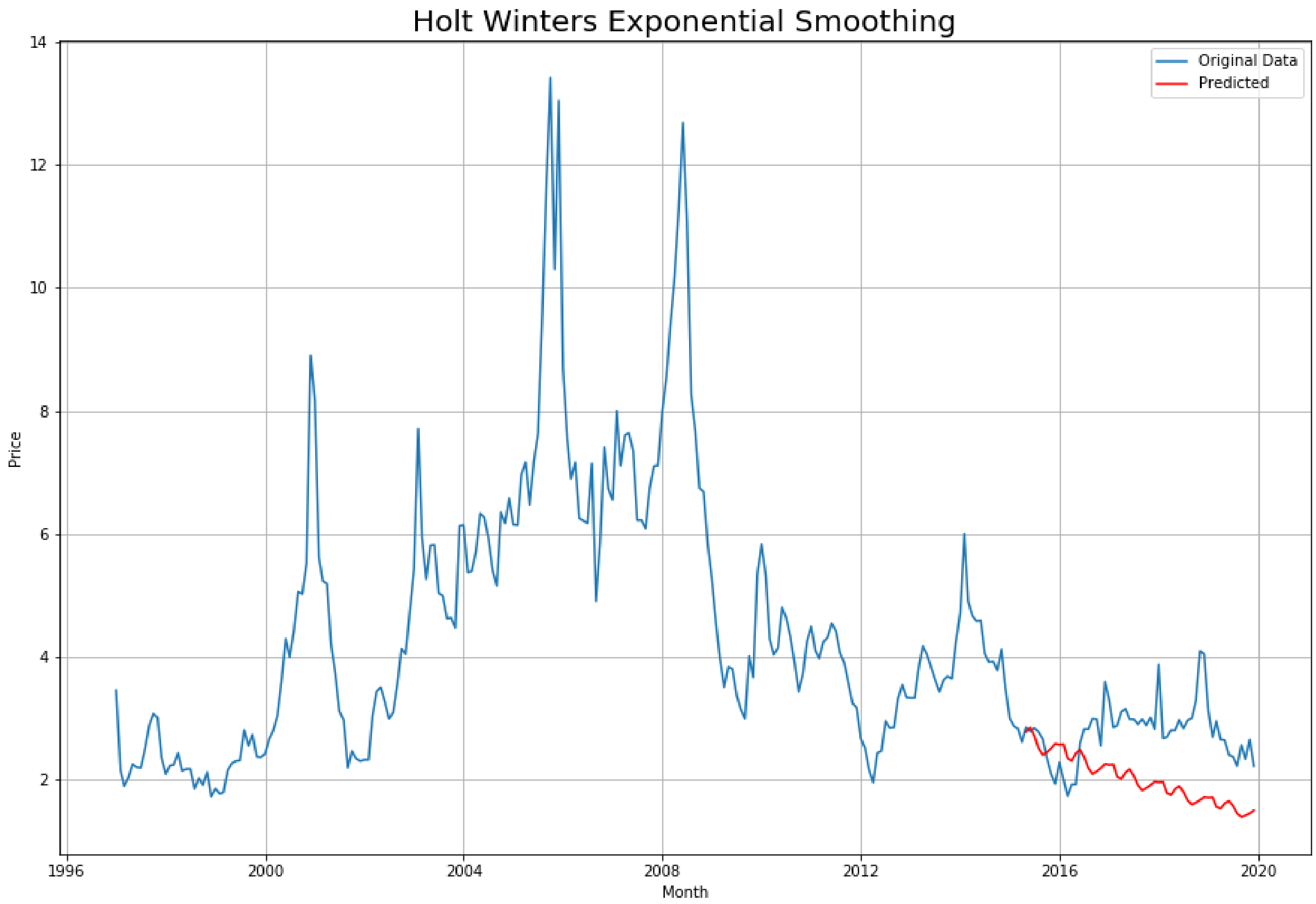
Before performing the Holt-Winters exponential smoothing on the data, we split our data into two parts such as **training data** and **testing data**. We use the training data to train the Holt-Winters model and fit the model. Next, we apply the fitted Holt-Winters model to the testing data and predict the present values of the testing data. After that we calculate the **RMSE (Root Mean Squared Error)** of the predicted values with respect to the observed values in the testing data. From RMSE we can understand how good the fit is.

RMSE is formulated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(y_{obs} - y_{pred}\right)^2}{n}}$$

Now, the split ratio will be 80:20 i.e. we consider the first 220 observations of the data (80% of the data) as the training data and remaining 56 observations (20% of the data) as the testing data.

Therefore, we train our Holt-Winters model through the training data and predict the values of the testing data. The results are as follows.



INTERPRETATION

In the graph above we see the prediction of the testing data along with the observed values and we can say that the prediction of our testing data is not that much perfect which depicts that the fit of Holt-Winters model on the data is **not good enough**. However, the **RMSE is 1.0126**.

V. AUTOCORRELATION FUNCTION (ACF) AND PARTIAL AUTOCORRELATION FUNCTION (PACF)

AUTOCORRELATION FUNCTION (ACF)

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). The ACF is most useful for identifying the **order of a moving average model (MA model)**.

PARTIAL AUTOCORRELATION FUNCTION (PACF)

A partial autocorrelation is a summary of the relation between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. The PACF is most useful for identifying the **order of an autoregressive model (AR model)**.

METHOD

AUTOCORRELATION FUNCTION (ACF)

The autocorrelation function (ACF) at lag k, denoted ρ_k, of a stationary stochastic process is defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_k}$$

Where $\gamma_k = cov(y_i, y_{i+k})$ for any i. Note that γ_0 is the variance of the stochastic process.

PARTIAL AUTOCORRELATION FUNCTION (PACF)

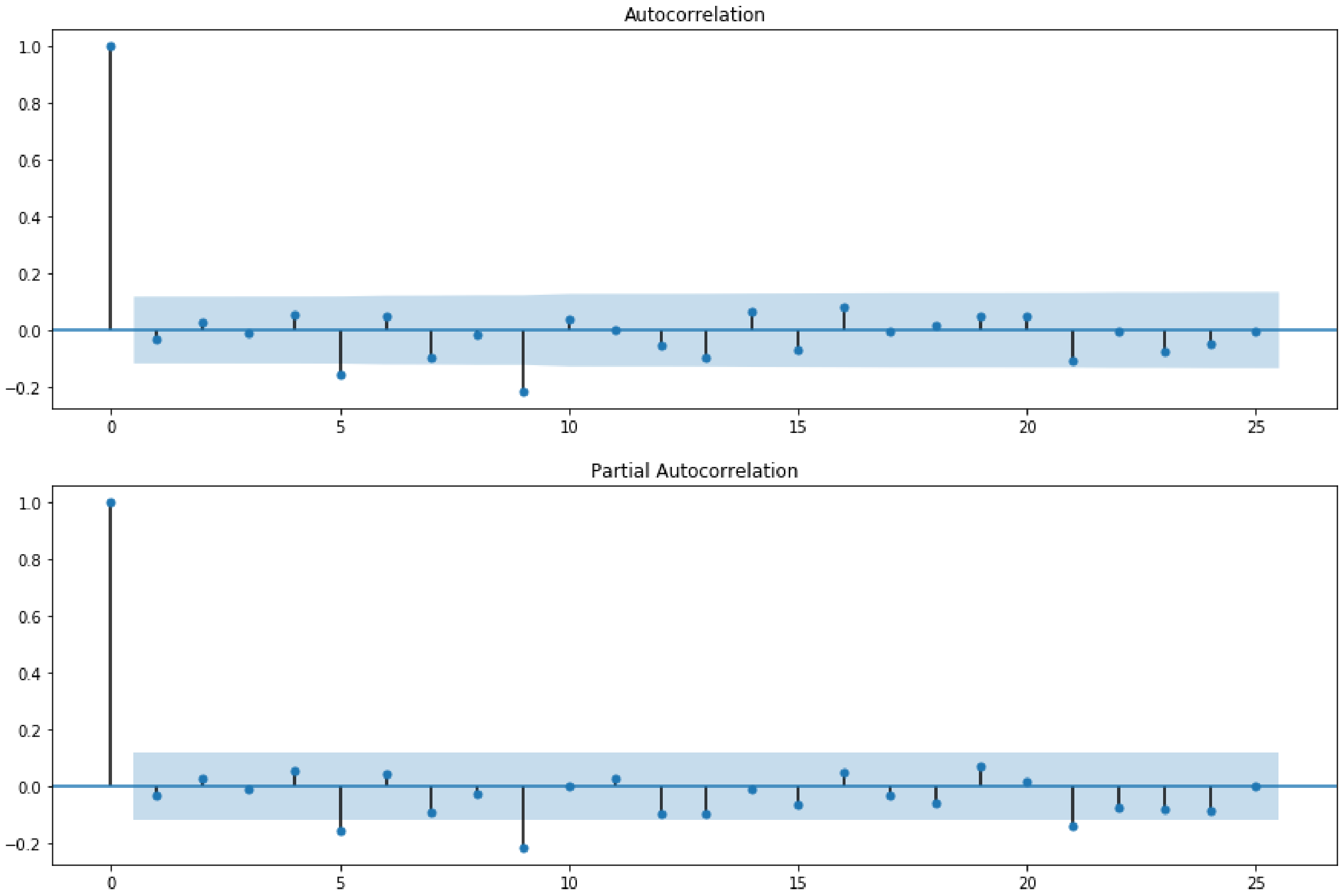
The partial autocorrelation function (PACF) of order k, denoted p_k, of a time series, is defined as the last element in the following matrix divided by r_0.

$$R_k^{-1} C_k$$

Here R_k is the $k \times k$ matrix $R_k = [s_{ij}]$ where $s_{ij} = r_{|i-j|}$ and C_k is the $k \times 1$ column vector $[r_i]$. We also define $p_0 = 1$ and p_{ik} to be the i^{th} element in the matrix $R_k^{-1} C_k$ and so $p_k = p_{kk}$.

ANALYSIS

Now we’ll draw the ACF and PACF graph of the first forward difference of the price data as the first forward difference data is a stationary data. Therefore, the graph of ACF and PACF of the stationary data is as follows:



INTERPRETATION

We see both the plots coming out to be more or less same except some slight differences. Both the autocorrelation and partial auto correlation at lag 1 are suddenly dampened out.

Since the autocorrelation at lag 1 is negative, we consider the order of the moving average model as 1. And for the same reason we consider the order of the autoregressive model as 1.

VI. FITTING MODEL

Autoregressive Integrated Moving Average, or **ARIMA**, is one of the **most widely used forecasting methods** for univariate time series data forecasting.

Although the method can handle data with trend, it does not support time series with a seasonal component.

An **extension to ARIMA** that supports the direct modelling of seasonal component of the series is called **Seasonal ARIMA** or **SARIMA**.

Since our data is a seasonal time series data, we would go for **SARIMA model** for forecasting.

METHOD

The SARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is

$$\text{ARIMA } (p, d, q) \times (P, D, Q)S$$

With p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

Without differencing operations, the model could be written more formally as:

$$\Phi(B^S)\phi(B)(x_t - \mu) = \Theta(B^S)\theta(B)w_t$$

The non-seasonal components are:

- AR: $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$
- MA: $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$

The seasonal components are:

- Seasonal AR: $\Phi(B^S) = 1 - \Phi_1 B^S - \cdots - \Phi_P B^{PS}$
- Seasonal MA: $\Theta(B^S) = 1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS}$

ANALYSIS

Before fitting the SARIMA model on the data, we split our data into two parts such as **training data** and **testing data**. We use the training data to train our SARIMA model and fit the model. Next, we apply the fitted SARIMA model to the testing data and predict the present values of the testing data. After that we calculate the **RMSE (Root Mean Squared Error)** of the predicted values with respect to the observed values in the testing data. From RMSE we can understand how good the fit is.

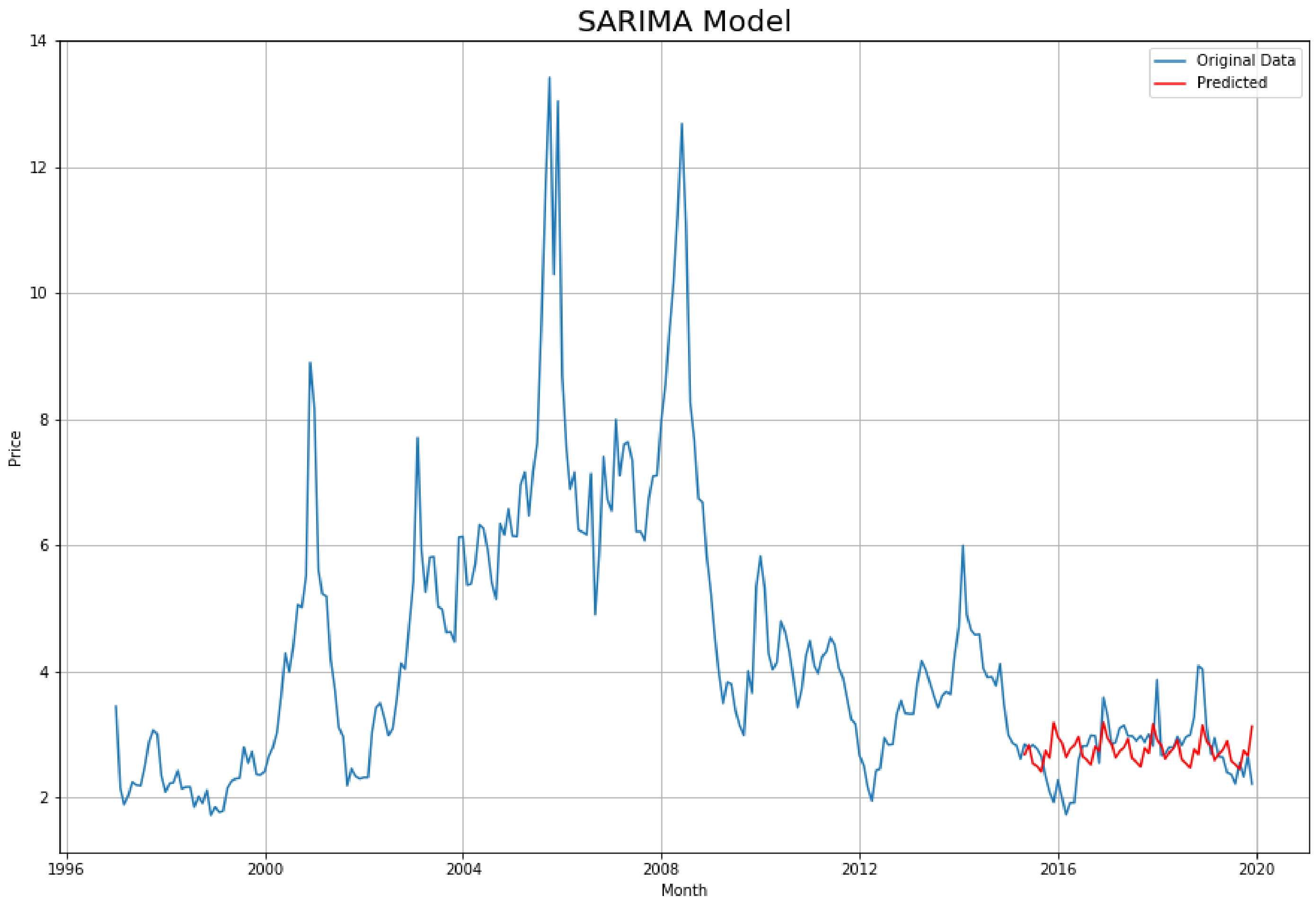
RMSE is formulated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(y_{obs} - y_{pred}\right)^2}{n}}$$

Now, the split ratio will be 80:20 i.e. we consider the first 220 observations of the data (80% of the data) as the training data and remaining 56 observations (20% of the data) as the testing data.

Observing the ACF plot we take the non-seasonal and seasonal MA orders i.e. **q and Q as 1**. Observing the PACF plot we take the non-seasonal and seasonal AR orders i.e. **p and P as 1**. Now, since our stationary data is obtained by the first forward differences of the prices, we take the non-seasonal and seasonal differencing i.e. **d and D as 1**. Next, we take the seasonal time span i.e. **S as 12** as our data is a monthly data.

Therefore, we train our SARIMA model through the training data and predict the values of the testing data. The results are as follows.



INTERPRETATION

In the graph above we can see the prediction of the testing data along with the observed values and we can say that prediction of our data is pretty well predicted which depicts that the fit of SARIMA model on the data is **good enough** and the **RMSE is 0.496**.

VII. FORECASTING

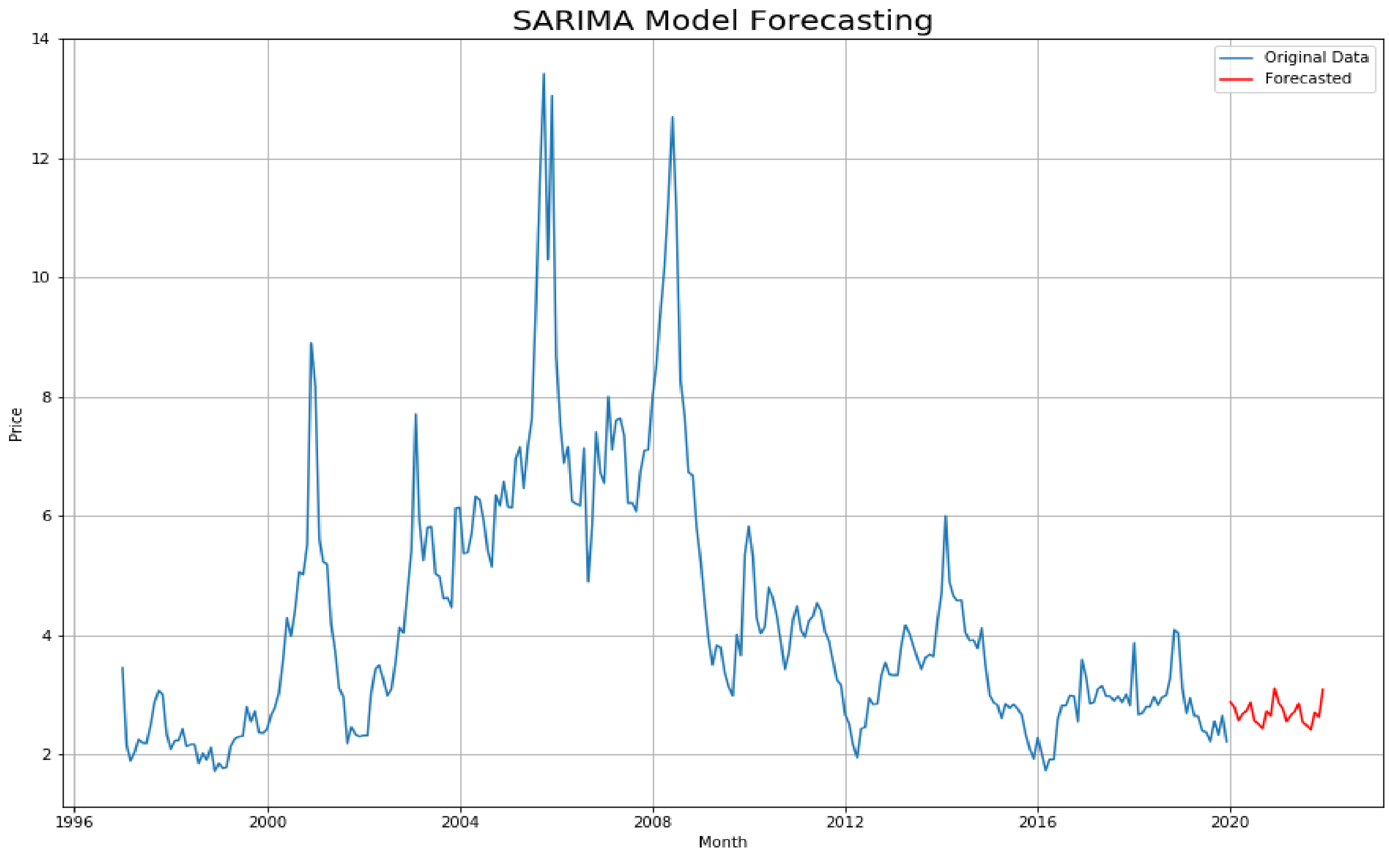
Now we are going forecast the prices in 2020 and 2021. Before forecasting, we have to choose the **appropriate model** for forecasting.

Earlier we have fit two models, Holt-Winters and SARIMA model on our data and tested both the models that how good they can predict values. Now, to compare which model is better at predicting we would like to **compare the RMSEs** of the models. We know that **less value of RMSE implies better prediction**.

As calculated earlier, RMSE of Holt-Winters model is 1.0126 and RMSE of SARIMA model is 0.496. Comparing both the RMSEs, we can say that SARIMA model with parameters (1,1,1)×(1,1,1,12) is better in predicting than the Holt-Winters model as the RMSE of SARIMA model is less than that of Holt-Winters. So, we'll incorporate **SARIMA model for forecasting** the future prices.

ANALYSIS

The forecasting of prices in 2020 and 2021 are shown in the graph below along with the original data.



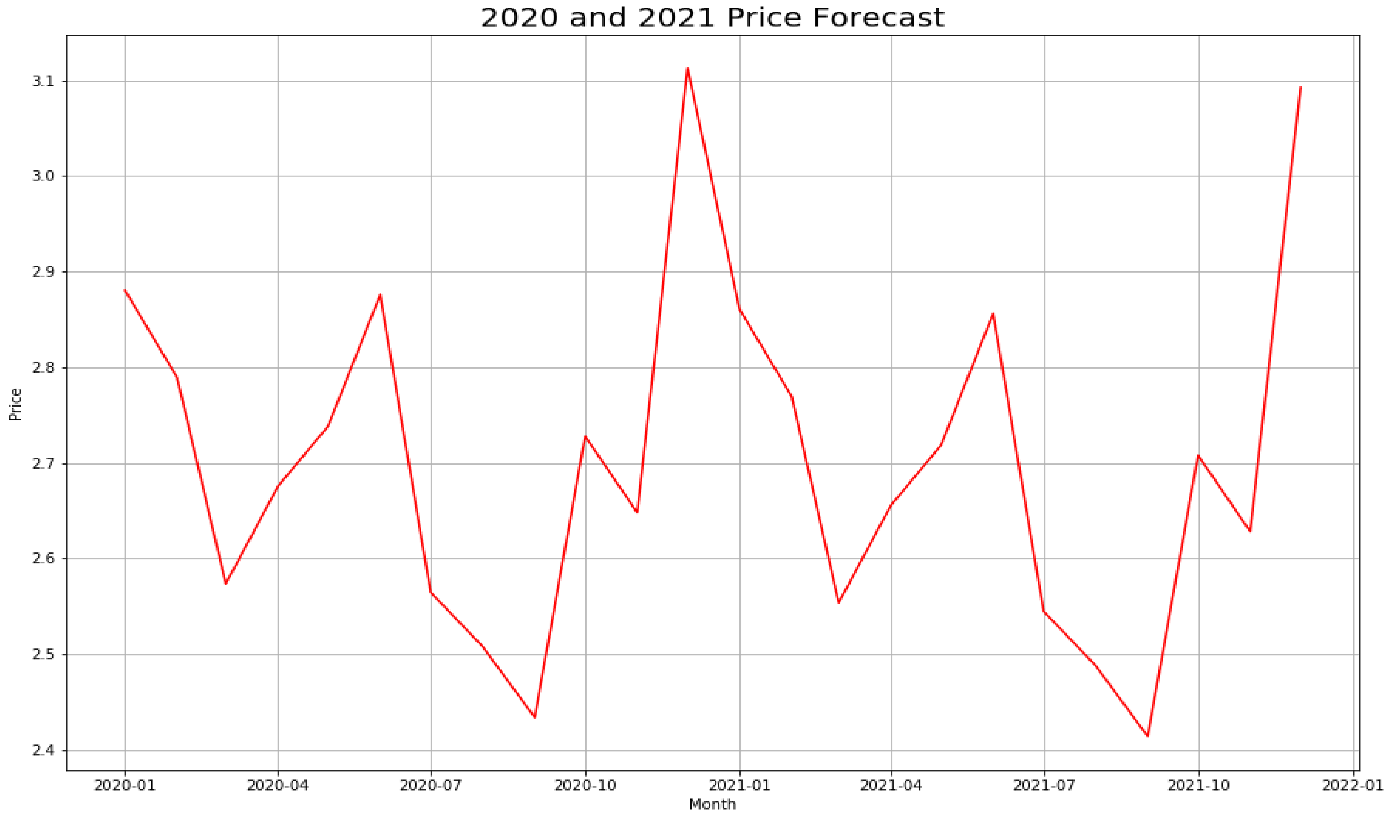
FINAL INTERPRETATION

After forecasting the prices, we got the following expected prices:

2020	
Month	**Price ($/MMBtu)**
Jan	2.880208
Feb	2.789136
Mar	2.573309
April	2.675023
May	2.738397
Jun	2.875911
Jul	2.564475
Aug	2.507681
Sept	2.433567
Oct	2.727618
Nov	2.647892
Dec	3.112230

2021	
Month	**Price ($/MMBtu)**
Jan	2.860322
Feb	2.769250
Mar	2.553423
April	2.655136
May	2.718510
Jun	2.856024
Jul	2.544589
Aug	2.487795
Sept	2.413680
Oct	2.707732
Nov	2.628005
Dec	3.092343

The Forecasted price graph may look like:



From the forecasted price graph, we can make following interpretations:

- The average value of the natural gas prices will be **$2.7/MMBtu**.
- The average **upper** and **lower limit** of the price will be around **$3.1/MMBtu** and **$2.425/MMBtu**. The prices are expected to vary these limits throughout the two years.
- There will be a bit of **seasonal impact** on the prices.
- The natural gas price may meet the **lower average price during the September month** in both the years. Due the mild cool temperature in rainy season, the household electricity consumption will be less.
- The natural gas prices may rise to **the upper average price during the December month** in both the years. During cold months, natural gas demand for heating by residential and commercial consumers generally increases overall natural gas demand and can put upward pressure on prices.
- The natural gas prices will **increase slightly** to $2.85/MMBtu in the **summer season**. Hot weather tends to increase demand for air conditioning in homes and buildings, which generally increases the power sector's demand for natural gas. As a result, the prices may rise.
- We can expect the natural gas prices will be **moderate** and **around the average price** $2.7/MMBtu during the **Spring season**.
- There will **not** be any upward or downward **trend** throughout the two years.

CONCLUSION

Now this project work was based on the data **before the recent outbreak of COVID-19 pandemic**. So, as we have seen earlier natural gas prices were affected by sudden irregularities such the **energy crisis, Hurricanes, financial crisis** etc. Therefore, we can expect that the natural gas prices may be **affected by this pandemic** as the government declared nationwide lockdown.

On May 12, the **Energy Information Administration (EIA)** releases its latest **Short-Term Energy Outlook (STEO)**. This report reflects the first deep dive into how the energy demand collapse brough on by the **COVID-19 pandemic is impacting energy market projections**.

According to the STEO, natural gas prices will naturally **rise through the rest of 2020** as US Production declines. EIA forecasts that Henry Hub natural gas spot prices will average **$2.14/MMBtu in 2020** and **the increase in 2021**, reaching an annual average **$2.89/MMBtu**. So, as compared to our project work forecast, the prices will fall through out 2020 and will shoot up in 2021 due the pandemic situation.

President Donald Trump on April 21 ordered **Energy Secretory Dan Brouillette** and **Treasury Secretory Steven Mnuchin** to put together a **plan to get funding** to the struggling US oil and gas industry as historic sell-off in crude continued.

"We will never let the great US Oil & Gas Industry down. I have instructed the secretory of Energy and Secretory of the Treasury to formulate a plan which will make funds available so that these very important companies and jobs will be secured long into future!" Trump tweeted on April 21 morning.

APPENDIX

PYTHON CODES

READ THE DATA

```
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
%matplotlib inline

data = pd.read_csv("natural_gas_prices.csv")
data["Month"] = pd.to_datetime(data["Month"], infer_datetime_format = True)
data = data.set_index("Month")
```

GRAPHICAL REPRESENTATION OF DATA

```
plt.figure(figsize=(15,10))
plt.grid(True)
plt.xlabel("Month")
plt.ylabel("Price")
plt.title("Original Data Plot", size = 20)
plt.plot(data["Price"])
```

AUGMENTED DICKEY-FULLER TEST

```
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller

result = adfuller(data["Price"])
print("ADF Statistic: %f" %result[0])
print("p-value: %f" %result[1])
print("Critical Values:")
for key, value in result[4].items():
    print("\t%s: %.3f" %(key, value))

if result[0] < result[4]["5%"]:
    print("At 5% level of significance - Reject H0 - Time series is stationary")
else:
    print("At 5% level of significance - Failed to reject H0 - Time series is non-
stationary")
```

MAKING THE DATA STATIONARY

```
data["Price first difference"] = data["Price"] - data["Price"].shift(1)
data["Price seasonal difference"] = data["Price"] - data["Price"].shift(12)

#Dickey Fuller Test
result1 = adfuller(data["Price first difference"].dropna())
print("ADF Statistic: %f" %result1[0])
print("p-value: %f" %result1[1])
print("Critical Values:")
for key, value in result1[4].items():
    print("\t%s: %.3f" %(key, value))

if result1[0] < result1[4]["5%"]:
    print("At 5% level of significance - Reject H0 - Time series is stationary")
else:
    print("At 5% level of significance - Failed to reject H0 - Time series is non-
stationary")
```

GRAPHICAL REPRESENTATION OF THE STATIONARY DATA

```
plt.figure(figsize=(15,10))
plt.grid(True)
plt.xlabel("Month")
plt.ylabel("Price first difference")
plt.title("First Difference Data Plot", size = 20)
plt.plot(data["Price first difference"])
```

MOVING AVERAGE TREND

*** computations of 12 pt. centered moving averages are done in MS Excel and it is saved in 'natural_gas_prices_2.csv' ***

```
data2 = pd.read_csv("natural_gas_prices_2.csv")
data2["Month"] = pd.to_datetime(data2["Month"], infer_datetime_format = True)
data2 = data2.set_index("Month")

plt.figure(figsize=(15,10))
plt.grid(True)
plt.xlabel("Month")
plt.ylabel("Price")
plt.title("12 pt Centred Moving Average Trend ",size=20)
plt.plot(data2["Price"], label = "Original Data")
plt.plot(data2["12_MA_cen"], color = "Red", label = "Moving Average Trend")
plt.legend(loc="best")
```

DECOMPOSITION OF THE DATA

```
from statsmodels.tsa.seasonal import seasonal_decompose
decomposition = seasonal_decompose(data["Price"], model = "multiplicative")
plt.figure(figsize = (15,10))

trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

plt.subplot(411)
plt.plot(data["Price"], label = "Original")
plt.legend(loc = "best")
plt.subplot(412)
plt.plot(trend, label = "Trend")
plt.legend(loc = "best")

plt.subplot(413)
plt.plot(seasonal, label = "Seasonality")
plt.legend(loc = "best")

plt.subplot(414)
plt.plot(residual, label = "Residuals")
plt.legend(loc = "best")
```

TRAIN TEST SPLIT

```
data_train = data[:220]
data_test = data[220:]
```

HOLT-WINTERS EXPONENTIAL SMOOTHING

```
from statsmodels.tsa.holtwinters import ExponentialSmoothing
model = ExponentialSmoothing(data_train["Price"], trend = "mul", seasonal = "mul",
seasonal_periods = 12)
fit = model.fit()
pred = fit.predict(start = 220, end = 275)

plt.figure(figsize=(15,10))
plt.plot(data["Price"], label = "Original Data")
plt.plot(pred, label = "Predicted", color = "red")
plt.xlabel("Month")
plt.ylabel("Price")
plt.title("Holt Winters Exponential Smoothing", size = 20)
plt.legend(loc="best")
plt.grid(True)
plt.show()

from sklearn.metrics import mean_squared_error
np.sqrt(mean_squared_error(data_test["Price"],pred))
```

ACF AND PACF PLOT

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

fig = plt.figure(figsize = (15,10))

ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(data["Price first difference"].dropna(), ax = ax1)

ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(data["Price first difference"].dropna(), ax = ax2)
```

SARIMA MODEL

```
import statsmodels.api as sm

model1 = sm.tsa.statespace.SARIMAX(data_train["Price"], order = (1,1,1), seasonal_order
= (1,1,1,12))
fit1 = model1.fit()
pred1 = fit1.predict(start = 220, end = 275)

plt.figure(figsize=(15,10))
plt.plot(data["Price"], label = "Original Data")
plt.plot(pred1, label = "Predicted", color = "red")
plt.xlabel("Month")
plt.ylabel("Price")
plt.title("SARIMA Model", size = 20)
plt.legend(loc="best")
plt.grid(True)
plt.show()

from sklearn.metrics import mean_squared_error
np.sqrt(mean_squared_error(data_test["Price"],pred1))
```

FORECASTING

```
forecast = fit1.predict(start = 276, end = 299)

plt.figure(figsize=(15,10))
plt.plot(data["Price"], label = "Original Data")
plt.plot(forecast, label = "Forecasted", color = "red")
plt.xlabel("Month")
plt.ylabel("Price")
plt.title("SARIMA Model Forecasting", size = 20)
plt.legend(loc="best")
plt.grid(True)
plt.show()

plt.figure(figsize = (15,10))
plt.plot(forecast, color = "red")
plt.title("2020 and 2021 Price Forecast", size = 20)
plt.xlabel("Month")
plt.ylabel("Price")
plt.grid(True)


forecast_20_21 = pd.DataFrame(forecast, columns = ["Forecasted Price"])
forecast_20_21

forecast_20_21.describe()
```