Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 21 Issue 1 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Assessing the Price Relationship and Weather Impact on Selected Pairs of Closely Related Commodities

By Adebanjo Adeniyi, Franklyn Ogbeide Okogun & Olaniyo Opiribo
WorldQuant University

Abstract- As indicated by various works of literature, climate change has a significant impact on agricultural commodities resulting in variation between demand and supply. The research study adopted quantitative analysis for comparative analysis of price relationships for three pairs of agricultural commodities against closely related products and how weather impacts them. As an interesting comparison, we also selected a pair of non-agricultural commodities for analysis. Downloaded data for the analysis were daily historical price data for the commodities, and daily summary of weather data for precipitation and temperature for the regions were the selected commodities are most produced. Using programming languages like Python and R, we carried out exploratory data analysis using the following statistics, such as graphs, scatter plots of returns, QQ plots for normality, time series diagnostics (AC, PAC) ARIMA, correlation. An exciting part of our work is our model selection, where we used SARIMAX for regressing endogenous data, i.e., commodity prices and exogenous data weather data.

Keywords: commodity, weather, python, Q-Q, ARIMA, AC, PAC, SARIMAX, correlation, data.

GJCST-C Classification: G.1

ASSESSINGTHEPRICERELATIONSHIPANDWEATHERIMPACTONSELECTEDPAIRSOFCLOSELYRELATEDCOMMODITIES

Strictly as per the compliance and regulations of:

Assessing the Price Relationship and Weather Impact on Selected Pairs of Closely Related Commodities

Adebanjo Adeniyi [α], Franklyn Ogbeide Okogun [σ] & Olaniyo Opiribo [ρ]

Abstract- As indicated by various works of literature, climate change has a significant impact on agricultural commodities resulting in variation between demand and supply. The research study adopted quantitative analysis for comparative analysis of price relationships for three pairs of agricultural commodities against closely related products and how weather impacts them. As an interesting comparison, we also selected a pair of non-agricultural commodities for analysis. Downloaded data for the analysis were daily historical price data for the commodities, and daily summary of weather data for precipitation and temperature for the regions were the selected commodities are most produced. Using programming languages like Python and R, we carried out exploratory data analysis using the following statistics, such as graphs, scatter plots of returns, QQ plots for normality, time series diagnostics (AC, PAC) ARIMA, correlation. An exciting part of our work is our model selection, where we used SARIMAX for regressing endogenous data, i.e., commodity prices and exogenous data weather data.

Keywords: commodity, weather, python, Q-Q, ARIMA, AC, PAC, SARIMAX, correlation, data.

I. Introduction

a) Background Study

Agriculture is an activity that involves the "rearing of livestock and cultivation of crops for human need and commercial activity." Agriculture relevance is evident in the economy of a country, primarily through commodity trading [9].

Commodity markets avail traders to buy and sell commodities, which include raw materials or primary agricultural products, which is as a result of what farmers and industry produce or extract. It has a similarity to the equity market. However, in the equity market, investors buy and sell shares.

We can categorize commodities into soft commodities and hard commodities. The soft commodities comprise coffee, cocoa, and heat, while gold, silver, and oil make up the hard commodity. We can further break down Commodity market into four categories; Energy (heating oil, crude oil, natural gas), metals (silver, gold, platinum, zinc), Livestock and meat (poultry eggs, cattle, lean hogs) and, Agricultural (rice, wheat, corn, and soybeans).

Author α σ ρ: WorldQuant University. e-mails: nextbrand16@gmail.com, okogun.franklyn@gmail.com, opriolani@gmail.com

The commodity market can influence the cost of commodity products and also determines the price for some products. Nonetheless, weather can also have a positive or negative effect on the yield of an agricultural product.

The agricultural commodities market is subjected to unavoidable change in prices as a result of seasonal transition due to climate change which give rise to underlying extreme events like heat stress, droughts, floods, hail, frost, pest and disease outbreaks, rising carbon dioxide, which could give rise to adverse effect on agricultural commodity availability. On the other hand, a notable significant effect of weather changes could give rise to critical factors, which include the concentration of carbon dioxide (CO_2), which increases light intensity, soil moisture, water availability, soil nutrients, and temperature.

Previous work by Masters had emphasized on some agricultural commodities in specific regions and their relationship to climate change. The work emphasized that "Without doubt, climate change is occurring and is already having a dramatic impact on climatic variability, global temperatures, and sea level. Climate change will have significant impacts on agriculture, reflecting the close link between climate (temperature and precipitation in particular) and productivity, and these effects are likely to have the greatest effect in the least developed countries of the tropical zones where productivity will decrease" [36].

b) Problem statement

Demand and supply, an economic concept, is a conventional fundamental analysis in a market where prices are not regulated. Product availability is substantially controlled by consumption and production at various periods in a calendar year. The agricultural commodity market is not an exception where there are numbers of production and consumption impacting factors. These tend to have demand variation among items of similar class and also causing wide swings in commodity prices.

Majorly, macroeconomic factors ranging from inflation, foreign reserve, and exchange rate are known factors that can cause variation in agricultural products. Nevertheless, seasonal transition as a result of climate change tends to cause a more significant impact, and

Global Journal of Computer Science and Technology (C) Volume XXI Issue I Version I Year 2021

this leaves investors with the choice of whether to buy or sell at a given period, mainly due to the weather impact on commodity products.

Climate change has adverse effects on the world and has become a significant barrier to economies; the effect of climate change on agricultural commodities causes volatility, and this causes the commodity price to fluctuate in extreme weather events. Moreover, variation in temperature, precipitation, and the frequency and intensity of extreme weather could have significant impacts on crop yields.

This study seeks to investigate the price relationships between three categories of agricultural commodities and one non-agricultural commodity against closely related products, i.e., Corn/Oat, Soybean/Wheat, Coffee/Cocoa, and Gold/Silver as the non-agricultural commodities.

Furthermore, the study seeks to analyze the impact of weather on Agricultural commodities. It also examines the implications of extreme weather conditions on commodities prices, and to this end, SARIMAX will be used to further check for seasonality in the prices, and the effect temperature and precipitation trend have on the different commodities.

c) *Goals and Objectives*

i. *Goal*

The goal of this research project is to assess the price relationship for three pairs of closely related agricultural commodities and a pair of closely related non- agricultural commodities and also the effect of weather on the commodities.

Corn/Oat, Soybean/Wheat, Coffee/Cocoa, and Gold/Silver

ii. *Objectives*

- To analyze returns to see correlation peak across different differencing intervals: daily, weekly, and monthly.
- To investigate the seasonality in prices of Corn/Oat, Soybean/Wheat, Coffee/Cocoa, and Gold/Silver.
- To compare the trend in temperature and precipitation with price variation of the chosen commodities.
- To determine a suitable model to regress weather data with the commodity data.

d) *Significance of the study*

The research work, when completed, will be useful to institutions, policy-makers, educators, strategists, and researchers with interest in the impact of weather on agricultural commodities. The study, therefore, bridges the research gap with an insight into the quantitative analysis of price variation as a result of weather impact on agricultural commodities.

II. Literature Review

a) *Introduction*

As outlined in our problem statement, the objective of this study is to analyze daily data of some agricultural commodities against closely related products in certain regions, investigating the seasonality of these prices and the correlation peak across different differencing intervals relative to temperature and precipitation.

This chapter comprehensively summarizes previous research work of literature on methods used over time to measure the effect of climate change on agricultural commodities and the merits and demerits of these methods. The reviews include a survey of scholarly articles, books, and other sources relevant to the impact of weather on commodity products.

b) *Theoretical review*

Masters in their working paper had emphasized on some agricultural commodities in specific regions and their relationship to climate change. The work emphasized that "Without doubt, climate change is occurring and is already having a dramatic impact on climatic variability, global temperatures, and sea level. Climate change has significant impacts on agriculture, and this change cannot be overemphasized on agriculture, considering the correlation between temperature, precipitation, and productivity resulting in a noticeable effect on less developed countries" [36].

Master's work also highlighted how "climate as a condition could give rise to underlying extreme events like heat stress, droughts, and floods, pest and disease outbreaks, rising carbon dioxide levels, which could have both detrimental and beneficial, on crop yields in specific cases." In all, these extreme conditions will hurt the production of agricultural commodities leading to food shortages and food insecurity [36].

Eva et al. in their work pointed out that "the only certainty about climate change on agriculture are increasing uncertainty and variability and an increase in frequency and severity of extreme events (storms, hurricanes, droughts)" They also identified some developed countries with extreme scenarios where production declines severely [8].

Jasmien et al. estimate the consequences of exogenous shifts in global agricultural commodity prices on real GDP for a panel of 75 industrialized and developing countries. In their working paper, "they discovered that increases in global agricultural commodity prices that are caused by unfavorable harvest shocks in some regions of the world significantly curtail domestic economic activity." Jasmien's overall findings imply that the consequences of climate change on advanced economies are likely more significant than previously thought [33].

Munasinghe et al. developed a metric called "record equivalent draws" (RED) based on record-high (low) temperature observations by assessing the impact of climate changes, especially during very high and low temperatures, estimating the frequency of extreme temperatures in the 19th century. The simulated result for the period shows that mean temperature is positively correlated with RED's high temperature while negatively correlated with RED's low temperature. This metric model proved to serve as a precise instrumentation of global warming and cooling [41].

Addison et al. carried out a study on nine African countries that are dependent on a commodity that has a significant effect on their income. "This paper used a quantitative method to measure the effect of commodity price surge using a structural non-linear dynamic model." The paper addresses whether the response of GDP per capita for the selected countries is different from unexpected increases in agricultural commodity prices as opposed to decreases in prices. Hence, it finalized that there is very little evidence that an unanticipated price increase (decrease) will lead to a significantly different response in per capita incomes [1].

Maria et al. considered models to measure the effect of climate change on agriculture. They emphasized that selection of model should consider various aspects which include "specific object of the analysis, the temporal and geographical scales, the specific forms of climate change (climate warming, weather fluctuations or extreme climatic events), the magnitude of the effects expressed according to the agricultural dimensions (biological, social or economic)." Emphasis was laid on the choice of model to implement as one of the vital steps in the assessment. Their work also "considered the lack of information by offering to researchers a useful tool with which to identify all the possible alternatives of models analyzing the effects of climate change on agriculture." Different models where consider, which include the Crop Simulation Models, the Production-Function Model, the Ricardian Model, the Mathematical Programming, the General Equilibrium Model, and the Integrated Assessment Models [35].

c) *Models used to measure the effect of Climate Change*

i. *The crop simulation method*

The crop simulation method focuses on “crop physiological responses to ascertain the potential impacts of climate change on agriculture. It is one of the most popular methods for assessing the impact of climate change on agriculture. The crop simulation approach begins with controlled experiments in laboratories and other controlled settings to describe and model the bio-physical reactions of different crops to changing environmental conditions”. In these controlled experiments, researchers attempt to isolate the influence of the various inputs on the actual magnitude of outputs. They attempt to identify the influence of climate, changes in carbon dioxide content in the atmosphere, soil, and management practices on yields of various crops.

These models use the “best available knowledge on plant physiology, agronomy, soil science, and meteorology in order to predict how a plant will respond under specific environmental conditions” [34]. The crop simulation models are calibrated to the selected location for selected crops given a particular management practice. From these experiments, “the yield changes are then extrapolated to the real world and speculate what the experimental results imply for the agricultural systems across the given region. Some examples of crop simulation models include CERES-Maize and CERES-Wheat”. The methods are based on detailed experiments to find out the response of specific crops and crop varieties to different climatic and other conditions [42].

A study carried out by Iglesias et al. “estimated the impact of climate change across spatial scales in significant wheat-growing sites of Spain. They used CERES-Wheat, a dynamic process crop growth model for examining wheat growth. Using the model, the authors further examined the response of irrigation, temperature, precipitation, and CO2 concentration on wheat yield. The results from the spatial analysis revealed similar results to the CERES-Wheat crop growth model”. The important conclusion from the empirical results is that water (both precipitation and irrigation) and temperature during the farming season significantly affect the variability of simulated crop yield [45].

Schneider et al. used an “Erosion Productivity Impact Calculator (EPIC) crop simulation model to see how farmers respond to natural variability to climate changes in the US Great Plains. They used the EPIC model, under a doubling of CO2 scenario, to calculate changes in crop yields for three groups of farmers in terms of adaptation practices: no adaptation, perfect adaptation, and 20-year lagged adaptation”. The 20-year lagged adaptation group is used to mimic the masking effects of natural variability on their ability to notice changes in climate. Adaptation options tested in the EPIC crop model included: varying planting dates, changing crop varieties, and regulating crop growth period. Their findings suggest that the warmer temperatures enabled farmers to plant early in the spring to avoid the risk of damage from high heat levels in critical reproductive periods in mid- summer. Besides, with a longer growing period, farmers were able to attain higher yields by choosing to grow lengthy maturity varieties with more extended grain- filling periods. The results from the EPIC crop model show that adaptation improves crop yields and support findings from other studies that adaptation serves to reduce potential adverse effects from changes in climate. There are "some critical limitations of crop simulation models. These limitations mostly relate to adaptation. The crop simulation model does not endogenize farmer behavior, and the model does not predict how farmers are likely to change their behavior as climate changes. The weakness of this approach is its inability to modeling the intricate farmer responses to the environment change". The management practice of the farmer is assumed to be exogenous or fixed. If "farmers continue to behave as they did when they calibrated the model, the results are accurate" [48].

Furthermore, crop simulation models have been calibrated only in a limited number of places. The model is associated with a very high cost, and this makes poor and developing countries should rely on experiments conducted in a developed country. If these locations are not representative of all farms, using such approaches in aggregate studies can provide misleading predictions [38].

ii. *Empirical Yield Method*

The empirical yield methods measure the sensitivity of yields to climate by measuring how yields vary under different climate conditions through actual observations. "The basic idea of this approach is that the growth of agricultural production depends on water, soil, economic inputs, and climate variables that the model uses as explanatory variables in estimating the production function for specific crops" [31].

From the empirical production Function, "one can isolate the effects of climate from other factors influencing yields. For example, one can construct cross- sectional studies of actual yields across different climate zones. Another way to empirically measure the sensitivity of yields is to examine the effect of weather on yields over time [40]. The first study in this area relied on a unique weather condition called the 'dustbowl' in the middle of the USA in the 1930s". For a brief period, temperatures were higher and precipitation slightly lower, leading to unusually dry soil conditions in this region. The study measures the reduction in yields of selected grains in this period compared to periods with typical weather across the region.

Poudel et al. attempted to investigate the effects of rapid change in climate patterns driven by global warming on agricultural production in Nepal with a focus on whether the impacts vary across seasons, altitudes, and the types of crops. Their work empirically identified the "changes in climate condition and its effect on agricultural production from the data of rice, wheat, and climate variables in Nepal." They employed a stochastic production function approach by controlling a novel set of season-wise climatic and geographical variables. They found that an increase in the variance of both temperature and rainfall has adverse effects on crop yields in general. Furthermore, the impact of the difference in the average rainfall and temperature found beneficial or harmful was related to the altitudes and the kinds of crops. The findings project that adaptation strategies should be adopted in "Nepalese farming activities, owing to altitudes, growing season, and the types of crops." [47]

The empirical yield function approach has some of the same limitations as the crop simulation approach. The main weakness of the production–function model is that it focuses on a specific crop or limited set of crops. It endorses the so-called 'dumb farmer' hypothesis, and farmers are assumed to continue growing the same crop, with the same technology regardless of the change in the climate. The model excludes from the analysis of the plausible farmer strategies that replace crops that are more sensitive to others that are less so. The model does not pay due attention to the social and economic dimensions of agriculture. This model, coupled with other models, will be relevant to treat the economic dimension better.

iii. *Cross-sectional (Ricardian) Analysis*

Mendelsohn, Nordhaus, and Shaw introduced a cross-sectional approach that examines how farmland value varies across a set of exogenous variables such as economic, climatic, soils, and environmental factors. It is called the "Ricardian Method" after the 19th-century classical economist David Ricardo (1772-1823), who observed that land values would reflect land profitability within a perfectly competitive market. The approach is a hedonic model of farmland pricing that assumes the value of a tract of land equals the discounted value of the stream of future rents or profits derivable from the land. The "cross-sectional Ricardian approach is a direct method of measuring climate sensitivity across locations." The technique estimates the net productivity of farmland as a function of climate, soils, and other control variables. The method stands on the theoretical foundation that one can measure the impact of the climatic variable of interest on the value or net revenue of the land by examining the relationship between climatic variables and land value [37]. The technique that relies on a cross-sectional sample of farms that span a range of climates and agricultural systems in different climate zones are observed to see how the systems respond to being indifferent climate settings [39].

As with all empirical methods, the more accurate the measurements of the variables, the better-uncontrolled variables are accounted for, the more variation in the desired variables (climate), and more extensive the sample size, the more accurate the results. The method is based on the idea that farmland value contains the value of climate as well as all other attributes that determine land productivity. By regressing farmland value (or net farm revenue) per hectare on a set of climate variables (for example, rainfall and temperature measured either in annual or seasonal basis) environmental characteristics (for example soil), socio-economic and other control variables, “one can determine the marginal contribution of each of these factors to farm income capitalized in land value (or net farm income)." The economic impact of climate change is captured by the difference in land values (or net revenue) across different climatic conditions. This approach estimates of farm performance across different climate conditions that can be used to infer the consequences of future climate change [40].

The model considered that farmers, given limiting factors, that they cannot control, choose a set of outputs and inputs to maximize profits. The Ricardian method implicitly captures adaptation by including decision making changes that farmers would make to tailor their operations to a changing climate. A notable example of armer adaptation strategies is crop choice, where a particular crop will become the optimal choice depending on the effects of a warmer climate. Optimal crop switching is, therefore, an essential component of measuring the agricultural impact of climate.

The “advantage of the cross-sectional approach is that it fully incorporates farmer adaptations. The underlying assumption of the model is that farmers will automatically make adjustments in their management practices and respond to changes in climate; the approach does not suffer from the ad hoc adaptation adjustments of all the other approaches”. The assumption of implicit farmer adaptations frees the analyst from the burden of including adaptation while estimating the impacts of climate change.

Moreover, “it is assumed that because farmer adaptations will be reflected in land values, the costs and benefits of adaptation are embedded in those values." However, the adjustments are not explicitly modeled; the technique treats adaptation as a "black box." It does not reveal the precise adjustments made by individual farmers to suit the local conditions they face. Since the Ricardian approach implicitly captures the adaptations, it becomes possible to make a comparative assessment of climate change impact on agriculture with and without adaptation and provides valuable insight to know how adaptation measures reduce the impact of climate change [37].

According to the IPCC AR4 chapter addressing climate change impacts on food production deals almost exclusively with estimates of effects of changes in the long-run means of temperature and precipitation on crop yields and livestock productivity [6].

Extreme events may lower long-term yields by directly damaging crops at specific developmental stages, such as temperature thresholds during flowering, or by making the timing of field applications more difficult, thus reducing the efficiency of farm inputs [46].

“Several simulation studies have developed specific aspects of increased climate variability within climate change scenarios." Rosenzweig et al. computed that, under scenarios of increased heavy precipitation, production losses due to excessive soil moisture would double in the USA by 2030 to $3bn per year [44].

The reviewed work had so far established the relationship between agricultural commodities and the impact of climate on their production. More so, the majority of the work had addressed the subject matter from a qualitative point of view. Also, the world forum focus had been towards creating a framework to address climate change issues. The majority of countries had adopted policies that address climate change. Nevertheless, these changes impacted the production of an agricultural product positively?

This research work poses to quantitatively investigate the impact of climate change on the selected agricultural and non-agricultural commodities in a specific region.

d) Competitor Analysis

In this section, we took a look at three of the world top five producers of the selected commodities and compared most under the following criteria:

1. Production
2. Export
3. Domestic Consumption
4. Growth rate

Corn production (1000MT) by country shows the United States to be in the lead with 347,006, followed by China with 254,000 and Brazil with 101,00 [10]. The United States has a production growth rate of -5.26%, China, with -1.29% and Brazil with 0.00% [11]. Export for the United States is 46,992 and China with 20 and Brazil with 36,000 [12]. As for domestic consumption, the United States with 306,466 seems to consume most of what they produce, while China domestic consumption stood at 277,000 and Brazil at 66,000 [13].

Oats production (1000MT) by country, EU-27 is number one with 7,920 with Russia in second with 4,300 and the third place going to Canada with 4,000 [14]. Norway, on the other hand, leads the growth rate for Oats at 108.33%, with Russia at -8.80% and Canada with 16.41% [15]. Oats export from EU-27 is at 125 and Russia at 90, while Canada is at 1800 [16]. Domestic consumption in the EU-27 stands at 7,750 while in Russia it is 4,200 and Canada with 2,000 [17]

Soybean production (1000 MT) by country shows China as the leading producer with 66,924, followed by the United States with 44,904, while Brazil is in 3rd with 33,950. [18]. China has a growth rate of -0.59%, the United States with 1.14%, and Brazil with 1.80 % [19]. China export is at 900, and the United States has 12,111 and Brazil with 15,200 [20]. China domestic consumption is 66,074, United States is 33,249 while Brazil is 18,950 [21].

Global Journal of Computer Science and Technology (C) Volume XXI Issue I Version I Year 2021

Wheat production (1000 MT) by country indicates EU-27 is number one with 153,000, followed by China at 132,000, and India being 3rd in the world has 102,190 [22]. EU-27, though, is the number one producer, but they only have a growth rate of 11.79%, while China has 0.43%, and India has 2.32% [23]. EU-27 export is at 29,000, which is 2nd in the world, China is 1,300, making them 10th in the world, and India is a distance 19th with 500 [24]. Domestic Consumption for EU-27 is 127,500, China is 128,000 and India is 98,000 [25]

Coffee production by country (1000 60 KG Bags) has Brazil on top with 59,300, followed by Viet Nam with 30,500 and then Colombia with 14,300 [26]. Interestingly, Brazil has a growth rate of -8.49% and Viet Nam 0.33% and Colombia with 0% [27]. Brazil maintains the highest export at 36, 820, with Viet Nam having 28,300 and Colombia 13,400 [28]. Brazil consumes 23,530 domestically, and Viet Nam consumes 3,400 and Colombia with 2,050 [29].

Cocoa production by country (1000MT) has Cote d'Ivoire on top with 1,449, followed by Ghana with 836 and Indonesia with 778. [51]

China is the top Gold producing nation with 399.7 tons, followed by Australia with 312.2 tons, with production up 6 percent in 2018. Russia, with a production of 281.5 tons' accounts for a massive 83 percent of European gold, which has been increasing its production every year since 2010 with output growth of 11 tons in 2018, or about 2 percent. [50]

The number one silver-producing country in the world is Mexico, with 5,600 metric tons of the metal, followed by Peru with a significant jump that took its silver production to 4,500 metric tons of silver in 2017. China, which produced 2,500 metric tons of silver, is on the 3rd. [50]

In conclusion, one can observe that the world's highest producers of a given commodity do not necessarily have the world's highest growth rate. Brazil, for example, being the world's highest producer of coffee, has a negative growth rate of -8.49%, and this might not be farfetched from climate-related events.

III. Research Methodology

a) Introduction

In this chapter, we introduce the method we used to carry out time series analysis in detail. We started with identifying the source for our data and the method for collection – where and how we got these data, and also, we show the background knowledge about our statistic method. Finally, we present the research criteria – validity and reliability. In other to achieve our goal as stated in chapter one the following methodology was used; Figure 3.1 below shows the methodology used in the study from gathering the data to drawing conclusions.

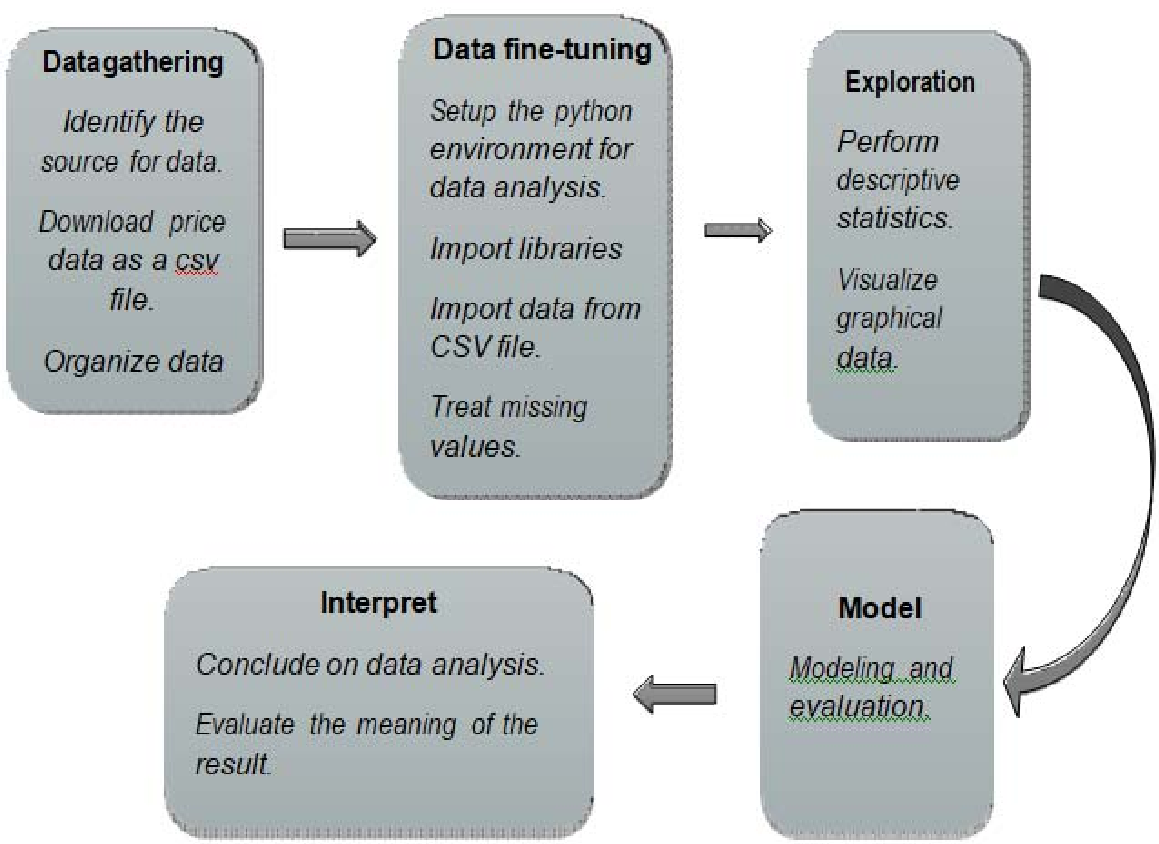


Figure 3.1: Taxonomy of the methodology

b) Data gathering

We address the research question first by identifying the data source, which is evidence to study. The approach to collect evidence depends on the research strategy and research question itself. “Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes” [5].

As illustrated by Ellen et al., their work emphasizes the importance of data collection. It is a critical part of time series analysis and about the vital part of research work. Hence, "Collect your data as if your life depends on it!”. [7]

For this research work, data gathering and method form an integral part of the study. According to Adi, there are mainly two methods to collect data, which are the primary methods of data collection and secondary methods of data collection. [2]

The primary data source is direct evidence of the originator, and it is not used in the past. The data gathered by primary data collection methods are specific to the motive of the research and highly authentic and accurate. We can further break down the primary data collection method into two categories: quantitative methods and qualitative methods. [2]

Secondary data is the data that has been used in the past and can be obtained from sources such as internal; Organization's health and safety records, Mission and vision statements, Financial statements Magazines, Sales reports, CRM software executive summaries and external sources of secondary data: Government reports, Press releases, Business Journals, Libraries, Internet. [2]

For our research work, our data collection and source fit both categories of the data collection method. However, its relevance can be seen in the secondary data collection, which cascaded down to a quantitative technique where statistical methods are highly reliable as the element of subjectivity is minimum in these methods. A vital tool in this method is the time series analysis, which can accommodate smoothing techniques to eliminate a random variation from the historical data.

Sourcing the right data was a critical part of our capstone project, and care was taken to ensure our data was sources from a reliable site.

Two sets of data were sourced for the project as follows:

- Commodity data
- Weather data

Data were sourced for the following eight commodities Oats, Corn, Wheat, Soybean, Coffee, Cocoa, Gold and Silver from the following sites:

- Macrotrends - https://www.macrotrends.net/ (Oats, Corn, Wheat, Soybean, and Coffee)
- Quandl - https://www.quandl.com/ (Cocoa, Gold, and Silver)

The number of years of data collected varies from one commodity to another, and Table 3.1 below is a summary of the range of the data collected.

Table 3.1: Price data

Commodity	Start Date	End Date	Number of Years
Cocoa	1/5/1970	11/15/2019	49
Coffee	8/20/1973	11/11/2019	46
Corn	7/1/1959	11/11/2019	60
Oats	1/5/1970	11/11/2019	49
Wheat	7/1/1959	11/11/2019	60
Soybean	12/5/1968	11/11/2019	49
Gold	1/2/1968	11/14/2019	51
Silver	1/2/1968	11/14/2019	51

Data was also collected for the weather as it relates to areas/regions where the selected commodities are most produced. This weather data is to investigate the impact or non-impact of weather on the price of the commodities.

The period of weather data collected varies for each region and where taken from specific weather station within the production area from the site https://www.ncdc.noaa.gov/cdo-web/.

Table 3.2: Summary of Weather Stations used for data collection

COMMODITIES	STATION	NAME	LATITUDE	LONGITUDE	ELEVATION
Oats	CA005021695	MARQUETTE, MB CA	50.0167	-97.8	244
Corn	AR000087374	PARANA AERO, AR	-31.783	-60.483	74
Wheat	RSM00021946	CHOKURDAH, RS	70.6167	147.8831	44
Soybean	ARM00087582	AEROPARQUE JORGE NEWBERY, AR	-34.559	-58.416	5.5

Coffee	BR00E3-0520	SAO PAULO AEROPORT, BR	-23.617	-46.65	802
Cocoa	IVM00065578	ABIDJAN FELIX HOUPHOUET BOIGN, IV	5.261	-3.926	6.4
Gold	ASN00066037	SYDNEY AIRPORT AMO, AS	-33.9465	151.1731	6
Silver	MXM00076525	ZACATECAS ZAC. LA BUFA ZAC, MX	22.77492	-102.5869	2,673

c) Research Method

"Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics" [3].

In essence, our research method is a quantitative study with a time series analysis. We choose historical price data and precipitation/ temperature data as our objectives to study the impact of weather on agricultural commodities. We collect a daily sequence of commodity historical price data from macrotrends website; "www.macrotrends.net" and "quandl" website; "www.quandl.com" websites. Similarly, we also downloaded daily summary of weather data (temperature and precipitation) for top producing countries of the chosen commodities from the National Centers for Environmental Information website "www.ncdc.noaa.gov/cdo-web/."

These data were analyzed thoroughly with statistic tools with an emphasis on comparative technique. Since data collected were from different periods, there was a need for us to carry out some form of data clean up to bring the data to a usable form. To archive this, we used Python 3.6 to write the code to carry out the data clean up. Steps followed in the time series analysis are listed below;

i. *Programming tool*

- We used the Python programming language (version 3.6) to carry out time- series analysis on the downloaded data. Python allows us to perform manipulation on time and date based data, visualize time series data, identify which models are suitable for a given dataset, create models for time series data. It also contains libraries that are suitable for time series analysis.

Imported libraries for our analysis include but not limited to, the following: matplotlib, numpy, pandas, sklearn, csv, scipy, stasmodels, seaborn. With the use of Jupyter notebook, required libraries were loaded using mostly the "import" statement

ii. *Price data*

- Preliminary analysis of price data: In this step, based on our objectives, we imported our data using the "CSV" library in python through jupyter notebook. Due to the anomalies that are found present in our imported dataset, we perform cleaning of our data to fit the analysis that will be performed subsequently. The data cleaning mainly encompasses using python programing language to carry out treating missing values and selecting/grouping of data according to findings in the preliminary analysis,
- Calculate Daily and Weekly Returns of grouped data: Daily and weekly returns were calculated on grouped data using the "pct_change()" function.
- Calculate spread and percentage change in the spread: Spread and percentage change in the spread were also calculated on grouped data.
- Calculate Ratio of Product Pair: Ratio was calculated among closes related product pairs.
- Plotting of product pairs: Product pairs were plotted in three categories; these include; plotting of raw data pair, plotting returns of product pairs, and plotting the spread/ratio of product pairs.
- A normality test: This test was carried out using the Q-Q test in python. This test was carried out on each of the returns of the commodity price data.
- Compute correlation matrix: This was carried out on daily and weekly returns to find out correlation among pairs.
- Skewness and Kurtosis: This test was carried out on the daily and weekly return of commodity data.
- Time Series Analysis (Serial correlation, ARIMA, ADF test): Based on the assumption that the time series are stationary for time series models, it is significant to validate it. The Augmented Dickey-Fuller test, which is a type of statistical test called a unit root test, was used to test for stationarity of our data. The Null-hypothesis for the test is that the time series is not stationary. So if the test statistic is less than the critical value, we reject the null hypothesis and say that the series is stationary. This test was carried out on the daily returns of the commodity data.

Next, after confirmation of stationarity, to select the relevant time series model, we carried out an autocorrelation plot to determine the value of q and p for SARIMAX.

iii. *Weather data*

- Loading and Cleanup of weather data using python pandas
- Interpolate to replace missing values
- Plotting raw precipitation and average temperature data for different production areas

Global Journal of Computer Science and Technology (C) Volume XXI Issue I Version I Year 2021

d) Data Analysis

In this research, we mainly use comparative analysis as our statistical tools to analyze our closely related product data.

A comparative analysis is mainly used to investigate the relationships between different variables; it provides a way for an investigator to explore a specific quantitative causal effect between these variables. For a long time, comparative techniques have become a central tool for multi-factor data analysis in the economic statistics field.

In our approach for comparative, we considered:

- Loading and cleaning data
- Treat missing value
- We regress different commodities data against weather data to assess the impact

e) Model Justification

SARIMAX model stands for Seasonal Auto Regressive Integrated Moving Average. It is a general time series model, "and is used to analyze and forecast data which have an additional seasonal component. We derive values for p, d, and q in order to make the time series stationary. A stationary series has a constant mean and variance." A general explanation of a SARIMAX model is illustrated in figure3.1 below; [43]

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Non-seasonal part}\\ \text{of the model}\end{array}\right)}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Seasonal part}\\ \text{of the model}\end{array}\right)}}$$

Figure 3.1: SARIMA model [43]

The model is usually of the form; *SARIMAX (p, d, q) x (P, D, Q) m,* which contains the non-seasonal and seasonal parts, as shown in figure 3.1.

The interpretation of SARIMAX (p,d,q)(P, D, Q)m is as follows:

p - auto-regressive order.
d - differencing order.
q - moving average order.
P - seasonal auto-regressive order.
D - seasonal differencing order.
Q - seasonal moving average order.
m – seasonality period (e.g., 24, 7, 52, 12)

The model accommodates an exogenous variable that is independent of the states of other variables in series. Factors outside our model determine its value. For our research, temperature and precipitation is the exogenous variable, as the occurrence of high temperature could give rise to the negative effect of agricultural commodity thereby affect demand and supply of an agricultural product. Our reliability high as the implementation of this work is dependent on different Python libraries. The majority of the algorithm to use is already implemented in the python modules. Hence usability, expressivity, and readability of the programming language structure are enhanced.

IV. Data Analysis, Results and Interpretation

a) Dataset analysis

We must analyze our data set and its form. So in this section, we will further detail our analysis of the data set and its properties. This analysis will cover the preliminary stage to the final stage.

b) Sample selection

We choose the price data for four pairs of closely related commodities and also weather database on the locations where the commodities are most produced. The data were pulled from 1953 to 2019, with the frequency being daily.

Table 4.0: Summary of data for analysis

Date	Oats	Corn	Wheat	Soybean	Coffee	Cocoa	Gold	Silver
1973-08-20	1.272	3.115	5.085	8.40	0.6735	1310.0	107.25	2.673
1973-08-21	1.212	3.015	5.285	8.14	0.6710	1265.0	103.00	2.694
1973-08-22	1.152	2.915	5.085	7.74	0.6580	1247.0	103.50	2.590
1973-08-23	1.092	2.815	5.005	7.40	0.6675	1291.0	102.50	2.556
1973-08-24	1.150	2.900	5.180	7.80	0.6660	1271.0	100.50	2.587

Date	Oats	Corn	Wheat	Soybean	Coffee	Cocoa	Gold	Silver
2019-11-05	3.0075	3.8175	5.1525	9.3425	1.0580	2508.0	1504.60	18.045
2019-11-06	3.0700	3.7875	5.1675	9.2750	1.0800	2480.0	1488.55	17.540
2019-11-07	3.0525	3.7525	5.1250	9.3650	1.0910	2458.0	1484.10	17.530
2019-11-08	3.0425	3.7725	5.1025	9.3100	1.0945	2507.0	1466.85	16.810
2019-11-11	3.1225	3.7325	5.0575	9.1700	1.0600	2524.0	1465.50	16.880

c) Statistical description

As shown in Table 4.0 above, the table represents the dataset for the selected data, which includes all the commodity data for analysis. The commodities are characterized by variation in price data. Each of the commodities was selected within and included 1973 and 2019. We already treated missing values in the data set by interpolating missing values. Interpolation is a mathematical method, adjusts a function to data and uses the function to extrapolate the missing data.

We performed descriptive statistics of the dataset to analyze the features of each commodity.

Table 4.1 below is the summary of the descriptive coefficients of the dataset, which we achieved by using the "describe ()" method on the Data Frame.

Table 4.1: Descriptive statistic of the datasets

	Oats	Corn	Wheat	Soybean	Coffee	Cocoa	Gold	Silver
count	11893.000000	11893.000000	11893.000000	11893.000000	11893.000000	11893.000000	11893.000000	11893.000000
mean	1.943889	3.129648	4.181742	7.586832	1.258932	1963.916421	591.260424	10.168358
std	0.751662	1.209386	1.456574	2.630739	0.481101	735.023371	432.911233	7.619002
min	0.945000	1.427500	2.147500	4.100000	0.425000	736.000000	89.750000	2.522000
25%	1.400000	2.325000	3.192500	5.715000	0.928000	1377.000000	319.050000	4.919000
50%	1.700000	2.742500	3.762500	6.675000	1.234500	1876.000000	389.800000	6.317500
75%	2.287500	3.610000	4.807500	8.895000	1.466000	2446.000000	842.500000	14.640000
max	4.960500	8.312500	12.825000	17.682500	3.356300	4508.000000	1896.500000	49.450000

As shown in Table 4.1 above, each of the variables has an equal number of observations, i.e., N = 11893 observations denote the sample size from 1973 to 2019. Also, in the result, the mean value for each commodity is displayed. i.e., Coffee has the lowest mean price of approximately 1.25, while cocoa has the highest mean price of approximately 1964.

In this report, the standard deviation for Oats is 0.7517. With normal data, most of the observations are spread with "3" standard deviations on each side of the mean. Base on the standard deviation and the mean, wheat and soybean appears to have normal data.

The three values (25%, 50%, and 75%) indicate quartiles at different levels. The 1st quartile is at 25% (Q1), the 2nd quartile at 50% (Q2 or median), and the 3rd quartile at 75% (Q3) that divide a sample of ordered data into four parts. For oats, i.e., an ordered data, the Q1 is 1.4, which implies that 25% of the data are less than or equal to $1.4. Also, Cocoa has Q1 = 1377, which implies that 25% of cocoa price data are less than or equal to $1377

Also going by closely related pairs of commodities, the following can be deduced:

For Oats/Corn, standard Deviation for Corn is greater than the Standard Deviation of Oats. This means that Corn is more volatile than Oat. Mean for Corn is greater than the mean for Oats; this shows the return for Corn is greater than the return for Oats (for our sample date range).

For Wheat/Soybean, standard Deviation for Soybean is greater than the Standard Deviation of Wheat. This means Soybean is more volatile than Wheat. Mean for Soybean is greater than the mean for Wheat, this shows the return for Soybean is greater than the return for Wheat (for our sample date range).

For Coffee/Cocoa, standard Deviation for Cocoa is greater than the Standard Deviation of Coffee. This means that Cocoa is more volatile than Coffee. Mean for Cocoa is greater than the mean for Coffee; this shows the return for Cocoa is greater than the return for Coffee (for our sample date range).

Lastly, for Gold/Silver, Standard Deviation for Gold is greater than the Standard Deviation of Silver. This means Gold is more volatile than Silver. Mean for Gold is greater than the mean for Silver; this shows the return for Gold is greater than the return for Silver (for our sample date range).

The final descriptive analysis performed on the raw data is a histogram of the data overlaid with a normal curve to examine the normality of the price data. Figure 4.0 shows the plot for the selected raw dataset.

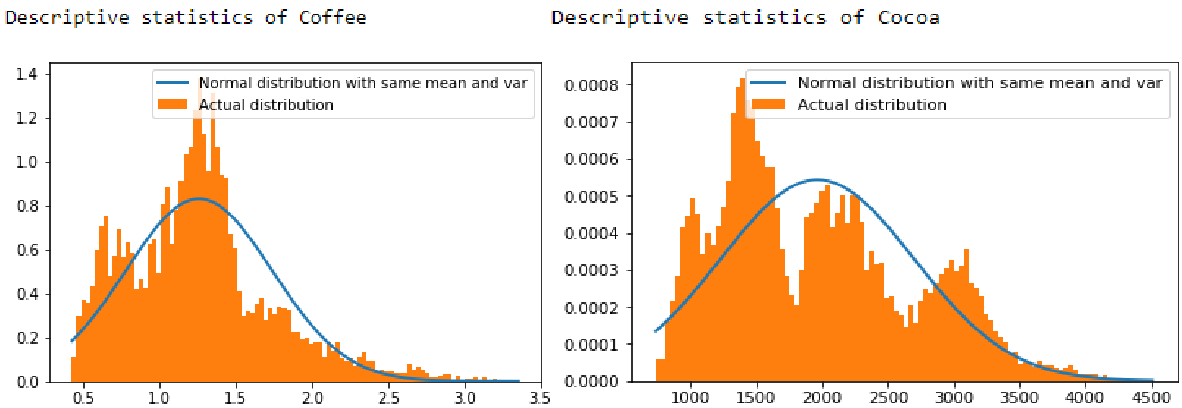


Figure 4.0: Normality test

As illustrated in the plots, the data appears to be a poor fit. A normal distribution would be a distribution that is symmetric and bell-shaped. For all the price data considered in our dataset, they appeared to be poor fit hence cannot conclude that they are normally distributed.

The next step involves calculating daily returns and weekly returns for each of the commodities. The figure below illustrates the result of descriptive statistics for the daily return.

Table 4.2a: Descriptive statistic for daily return

	Oats	Corn	Wheat	Soybean	Coffee \
count	11892.000000	11892.000000	11892.000000	11892.000000	11892.000000
mean	0.000247	0.000121	0.000141	0.000112	0.000265
std	0.018542	0.014547	0.016823	0.014485	0.021356
min	-0.112518	-0.076232	-0.097874	-0.071429	-0.140410
25%	-0.010152	-0.007595	-0.009616	-0.007739	-0.010770
50%	0.000000	0.000000	0.000000	0.000390	0.000000
75%	0.010737	0.007796	0.009386	0.008175	0.010992
max	0.117417	0.094257	0.126984	0.068760	0.271298

	Cocoa	Gold	Silver
count	11892.000000	11892.000000	11892.000000
mean	0.000206	0.000297	0.000424
std	0.017357	0.012426	0.023461
min	-0.143293	-0.148100	-0.467213
25%	-0.009757	-0.005047	-0.008795
50%	0.000000	0.000000	0.000000
75%	0.009909	0.005481	0.009662
max	0.099598	0.133539	0.815650

As illustrated in Table 4.2a above, there is about 11892 sample size considered for each of the commodities. Silver, having the highest return, had a maximum value of 0.82 and a minimum value of -0.47. However, its Q1 is – 0.009, which implies that 25% of the daily return for silver is less than or equal to – 0 .09. Going by our first pairs of commodities, the standard deviation suggests that Oats is more volatile than Corn, Wheat is more volatile that Soybean, Coffee is more volatile than Cocoa, and Silver is more volatile than Gold.

Going by the weekly returns, 2412 sample size was considered for the weekly return of each commodity. As earlier noticed on daily return, Silver

Global Journal of Computer Science and Technology (C) Volume XXI Issue I Version I Year 2021

maintained the highest value of 0.8915 and a minimum of -0.4352 when compared to other commodities.

Going by closely related pairs of commodities, the following can be inferred;

For Oats/Corn, the mean value for the daily returns of Oats is slighter greater than Corn's value. Also, the standard deviation for the daily returns of Oats is slightly higher than that of its closely related product, which implies that corn is less volatile when compared to its closely related product. This occurrence is also replicated in the weekly and monthly returns of Oats and Corn.

Wheat is slightly more volatile when compared to Soybeans. This is also replicated in the weekly and monthly returns for the closely related products.

In other closely related products, Coffee is moderately volatile than Cocoa; also, Gold is less volatile when compared to Silver. This description is also replicated in the weekly and monthly returns of each pair. This can be seen in Table 4.2b and 4.2c in appendix B.

Table 4.3a: Skewness and Kurtosis for daily returns

SKEWNESS and KURTOSIS RESULT ON DAILY RETURNS

Commodities	Skewness	Kurtosis
Oats	0.0626773	1.77503
Corn	0.0426126	2.58415
Wheat	0.211654	2.62211
Soybean	-0.116768	1.74379
Coffee	0.510605	8.60062
Cocoa	0.0272938	2.0969
Gold	0.357106	13.9622
Silver	3.31957	150.634

Skewness tells the amount and direction of skew. For the daily return, the skewness values are within the range of -0.5 and 0.5. This implies that the return distribution for the commodities is approximately symmetric except for silver with skewness of 3.32, which implies that its distribution is highly skewed.

We further investigated the unusually high skewness value for silver return data to understand this anomaly. From the graphs below, Figure 4.0a is the plot of the Silver data we used in our project, when compared with a similar graph of Silver data in Figure 4.0b sourced from "https://silverprice.org/silver-price-history.html" we noticed they are precisely the same which proof that our data are correct.

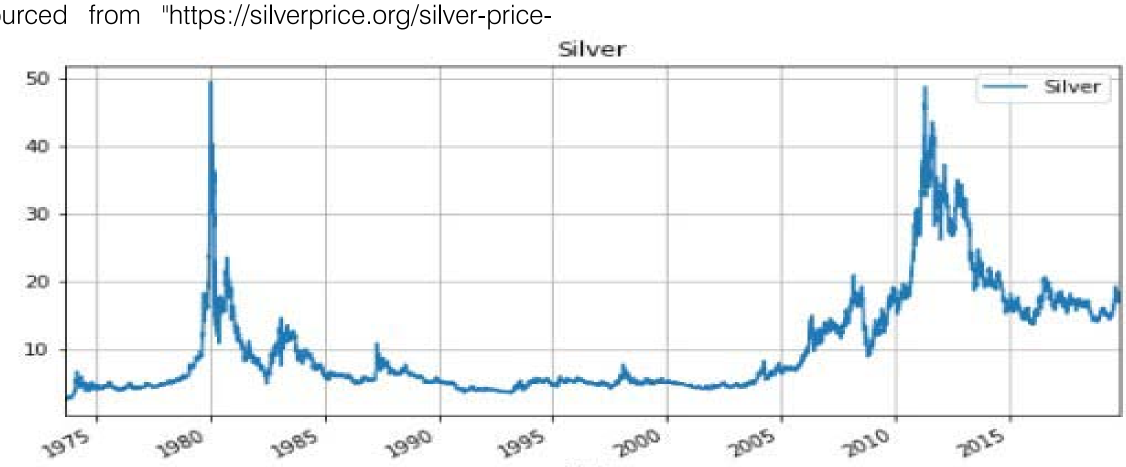


Figure 4.0a: Plot for Silver using our project data

Figure 4.0b: Plot for Silver [49]

To understand this further, we decided to break down our data into five years period and analyze each period separately. Table 4.3b below shows the skewness and Kurtosis values calculated for each five years period, and we discovered similar high skewness numbers in the period 1978 – 1983, which corresponded to the period silver price dropped to under $11 from its high of $48.70 [4].

Table 4.3b: Silver Skewness and Kurtosis for five years' period

Periods	Daily returns Skewness	Daily Returns Kurtosis
period_1 = ['1973-8-21': '1978-8-21']	0.13276071	4.760545859
period_2 = ['1978-8-22': '1983-8-22']	3.896131243	84.35246364
period_3 = ['1983-8-23': '1988-8-23']	-0.408416305	17.51717774
period_4 = ['1988-8-24': '1993-8-24']	0.320304353	4.827690375
period_5 = ['1993-8-25': '1998-8-25']	0.347668785	4.067214547
period_6 = ['1998-8-26': '2003-8-26']	0.115281399	1.835099532
period_7 = ['2003-8-27': '2008-8-27']	-1.041440941	5.62224845
period_8 = ['2008-8-28': '2013-8-28']	0.10090149	8.160369753
period_9 = ['2013-8-29':'2019-11-11']	0.071765419	2.488109109

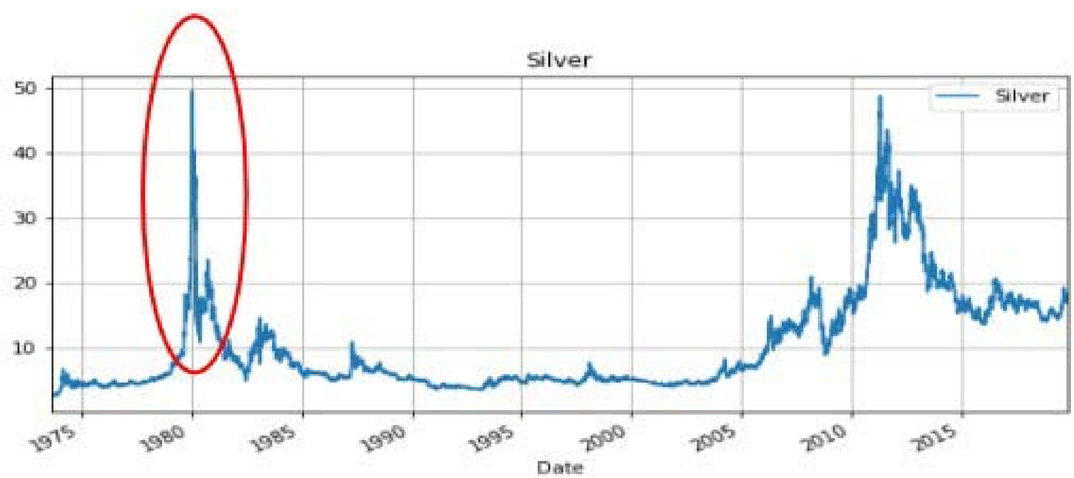


Figure 4.0c: showing the period when Silver Thursday occurred

The phenomenon of 1980 is what is commonly referred to as "Silver Thursday," which was the "dramatic drop in the price of silver and the panic that ensued in the commodities market on Thursday, March 27, 1980. The sharp fall occurred because of the failed attempt of two brothers, Nelson Bunker Hunt and William Herbert Hunt, to corner the silver market." [32]

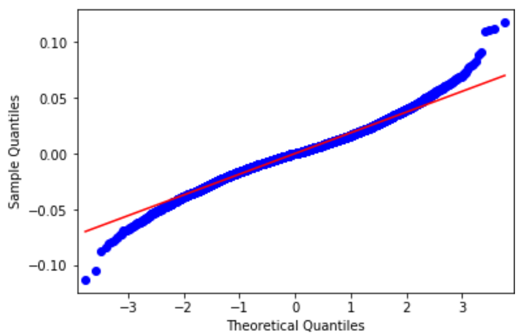
Global Journal of Computer Science and Technology (C) Volume XXI Issue I Version I Year 2021

For the Kurtosis, as seen in table 4.3a, the majority of the commodity appears to be Platykurtic which, when compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader. However, Coffee, Gold, and Silver appear Leptokurtic which, when compared to a normal distribution, has its tails longer and fatter, and often its central peak is higher and sharper. Excess kurtosis for Coffee, Gold, and Silver is reported for the return series and implies non-normality of distribution. This is also seen in the Kurtosis report for weekly and monthly returns of table 4.3c and 4.3d of appendix C.

d) Normality test

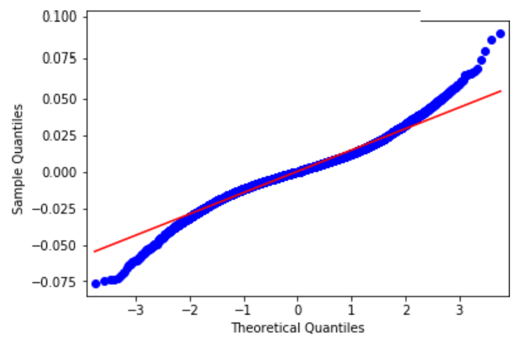
We went further to carry out a normality test to determine how well a normal distribution models our return series. A quantile-quantile (Q-Q) plot was used to show the distribution of the return series against the expected normal distribution.

```
=============
QQ Plot
Oats
=============
```



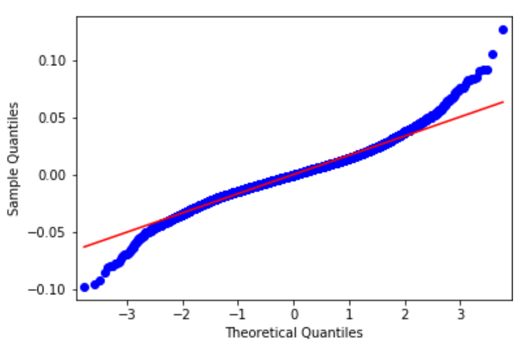

```
normality test for Oats
Statisitic: 43.757
15.000: 0.576,data looks normal(fail to reject H0)
10.000: 0.656,data looks normal(fail to reject H0)
5.000: 0.787,data looks normal(fail to reject H0)
2.500: 0.918,data looks normal(fail to reject H0)
1.000: 1.092,data looks normal(fail to reject H0)
.............................................
```

```
QQ Plot
Corn
=============
```



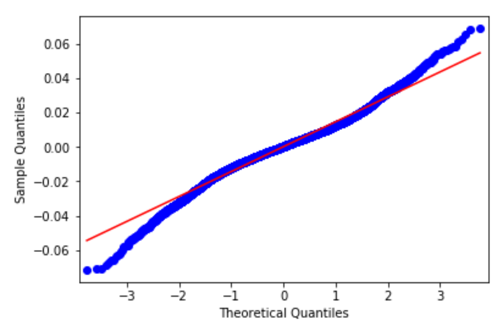

```
normality test for Corn
Statisitic: 80.317
15.000: 0.576,data looks normal(fail to reject H0)
10.000: 0.656,data looks normal(fail to reject H0)
5.000: 0.787,data looks normal(fail to reject H0)
2.500: 0.918,data looks normal(fail to reject H0)
1.000: 1.092,data looks normal(fail to reject H0)
............................................
```

```
=============
QQ Plot
Wheat
=============
```



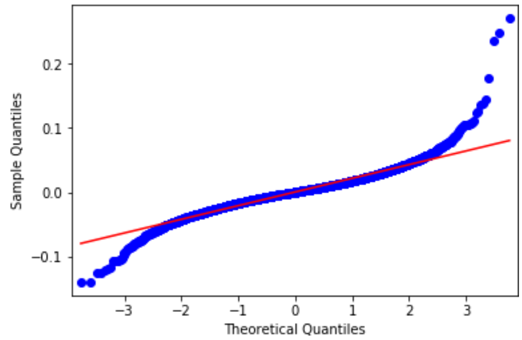

```
normality test for Wheat
Statisitic: 52.897
15.000: 0.576,data looks normal(fail to reject H0)
10.000: 0.656,data looks normal(fail to reject H0)
5.000: 0.787,data looks normal(fail to reject H0)
2.500: 0.918,data looks normal(fail to reject H0)
1.000: 1.092,data looks normal(fail to reject H0)
.............................................
```

```
=============
QQ Plot
Soybean
=============
```



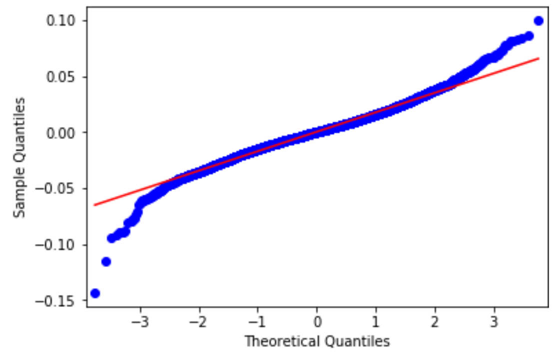

```
normality test for Soybean
Statisitic: 64.909
15.000: 0.576,data looks normal(fail to reject H0)
10.000: 0.656,data looks normal(fail to reject H0)
5.000: 0.787,data looks normal(fail to reject H0)
2.500: 0.918,data looks normal(fail to reject H0)
1.000: 1.092,data looks normal(fail to reject H0)
............................................
```

```
=============
QQ Plot
Coffee
=============
```



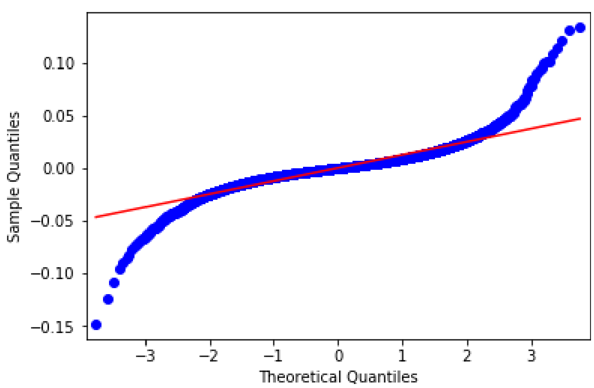

```
normality test for Coffee
Statisitic: 99.208
15.000: 0.576,data looks normal(fail to reject HO)
10.000: 0.656,data looks normal(fail to reject HO)
5.000: 0.787,data looks normal(fail to reject HO)
2.500: 0.918,data looks normal(fail to reject HO)
1.000: 1.092,data looks normal(fail to reject HO)
...............................................
```

```
=============
QQ Plot
Cocoa
=============
```



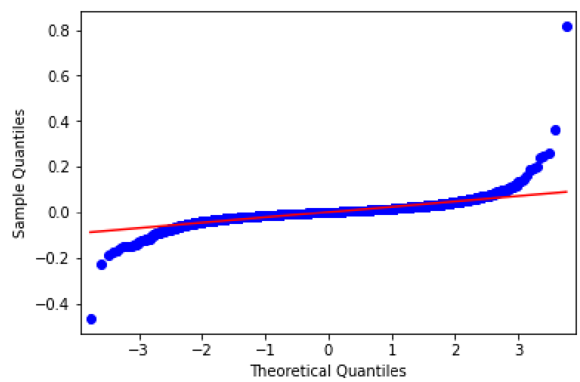

```
normality test for Cocoa
Statisitic: 37.967
15.000: 0.576,data looks normal(fail to reject HO)
10.000: 0.656,data looks normal(fail to reject HO)
5.000: 0.787,data looks normal(fail to reject HO)
2.500: 0.918,data looks normal(fail to reject HO)
1.000: 1.092,data looks normal(fail to reject HO)
...............................................
```

```
=============
QQ Plot
Gold
=============
```




```
normality test for Gold
Statisitic: 263.804
15.000: 0.576,data looks normal(fail to reject HO)
10.000: 0.656,data looks normal(fail to reject HO)
5.000: 0.787,data looks normal(fail to reject HO)
2.500: 0.918,data looks normal(fail to reject HO)
1.000: 1.092,data looks normal(fail to reject HO)
...............................................
```

```
=============
QQ Plot
Silver
=============
```




```
normality test for Silver
Statisitic: 369.479
15.000: 0.576,data looks normal(fail to reject HO)
10.000: 0.656,data looks normal(fail to reject HO)
5.000: 0.787,data looks normal(fail to reject HO)
2.500: 0.918,data looks normal(fail to reject HO)
1.000: 1.092,data looks normal(fail to reject HO)
...............................................
```

Figure 4.1: Q-Q plots for returns

As in q-q plot for various daily commodity returns, while our skewness says otherwise for some commodity returns, our distribution looks normally distributed.

We can go with the report of the q-q plot as deviations from the straight line are minimal.

e) Correlation among product pair

We conducted correlation at different frequencies among closely related pairs to show how or whether chosen pairs are related. While we confirmed that some pairs are closely correlated, performing the test lead to which pairs are the strongest in terms of correlation. Figure 4.2a below shows the matrix for the correlation coefficient among pairs.

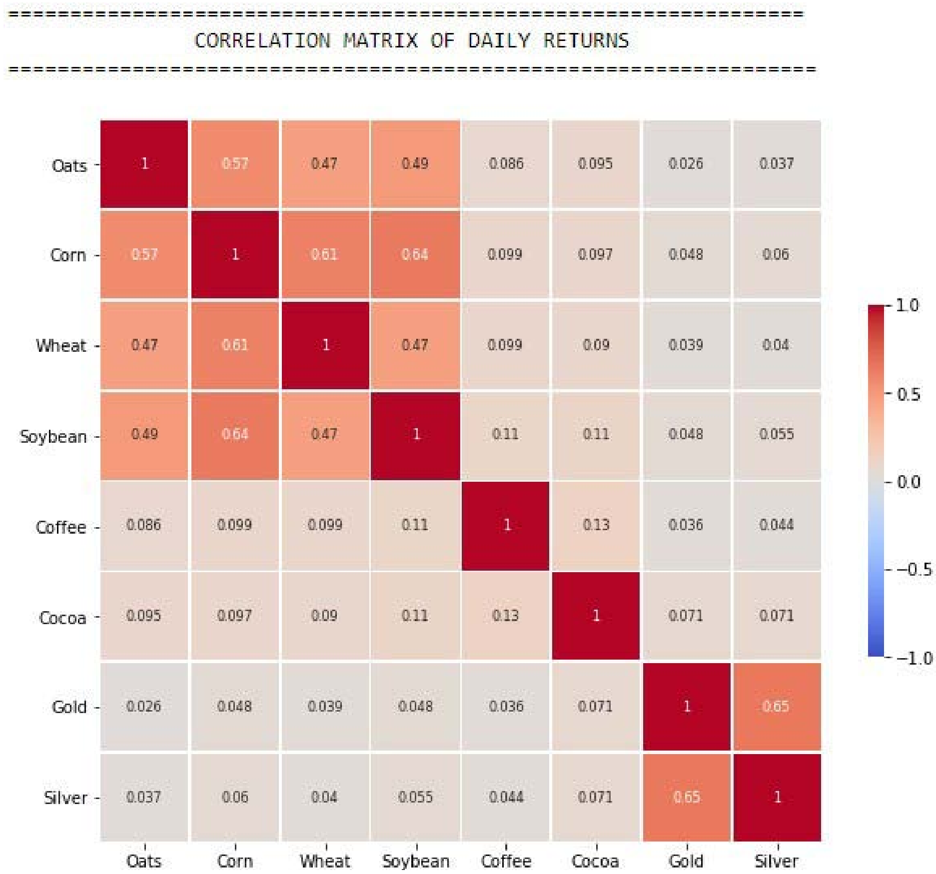


Figure 4.2a: Correlation matrix for daily returns

As shown in figure 4.2a above, Gold/Silver are strongly correlated and are the most correlated among the pairs considered. This observation is also confirmed in the weekly and monthly correlation matrix found in Figures 4.2b and 4.2c of the appendix D. Corn and Oats have a moderate rate correlation of their daily returns which is also repeated in the weekly returns matrix but they, however, have a strong correlation in their monthly returns.

Wheat and soybeans have a moderate correlation of 0.47 in their daily return, which is also replicated in their weekly and monthly return correlation matrix.

Coffee and cocoa have a very weak correlation of 0.13 in their daily returns, which can also be seen in their weekly and monthly return correlation matrix.

f) Model Selection

In other to guide us on our model selection, we carried out a statistical test called a unit root test on the daily returns of each commodity. The Augmented Dickey- Fuller (ADF test) was used to carry out stationarity of the return series.

Augmented Dickey-Fuller Test on "Oats"

Null Hypothesis: Data has a unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -48.7205

No. Lags Chosen = 4

Critical value 1% = -3.431

Critical value 5% = -2.862

Critical value 10% = -2.567

=> P-Value = 0.0. Rejecting Null Hypothesis.

=> Series is Stationary.

Augmented Dickey-Fuller Test on "Corn"

Null Hypothesis: Data has a unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -19.6698

No. Lags Chosen = 28

Critical value 1% = -3.431

Critical value 5% = -2.862

Critical value 10% = -2.567

=> P-Value = 0.0. Rejecting Null Hypothesis.

=> Series is Stationary.

Figure 4.3: ADF test for Oats and Corn return series

The result of the report in figure 4.3 indicated that Oats and Corn are stationary. This is also shown in the report of the ADF test for the other pairs, which is available in figure 4.3a of appendix E.

Going by the test carried out so far, it is imminent that SARIMAX will be a better choice to fit our model. The model incorporates endogenous and exogenous variables. Our commodity data formed the endogenous variables, while precipitation and temperature make up the exogenous variables.

Before analyzing the model fit, we have first analyzed precipitation/temperature data for each commodity in specific regions.

g) Weather impact

Weather data (precipitation/temperature) was downloaded for regions where each of the commodities has high production. The weather data is saved in the "CSV" format, and the necessary data were loaded on python notebook using the "pandas" module. The weather data was considered from 1973 to 2019 to match the date covered for each commodity price data. The precipitation value and the average temperature value were used to carry out the analysis. The next step was followed by treating missing values using the interpolate method, which resulted in plotting weather data against commodity prices.

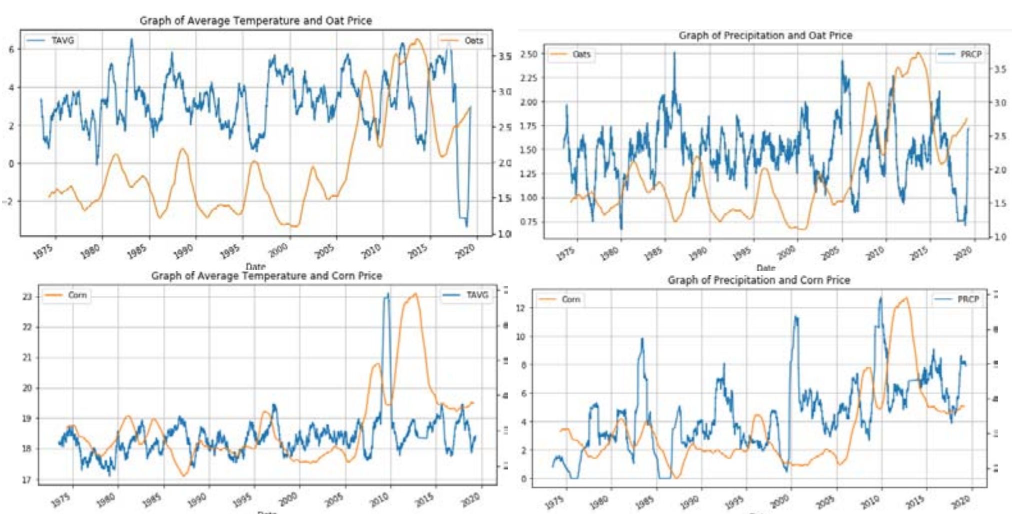


Figure 4.4a: Plot illustrating weather data and Oats/Corn prices

Weather data characterize oats and corn prices. As illustrated in Figure 4.4a, there are moments when a spike in temperature resulted in price reduction for the pair. This is also evident in the precipitation plot where a spike in its value resulted in a deep in commodity price. While 2010 – 2015 is characterized by a period of the most price for oat and corn, a sharp fall in the price of oats and corn is evident in 2016 – 2017, which is evident to be a period of high temperature in history. Also, the pair which exhibited most volatility in the last decade of the year under review (figure 4.4 in appendix F) recorded their highest price thou with occasions of spikes in temperature leading to a reduced price.

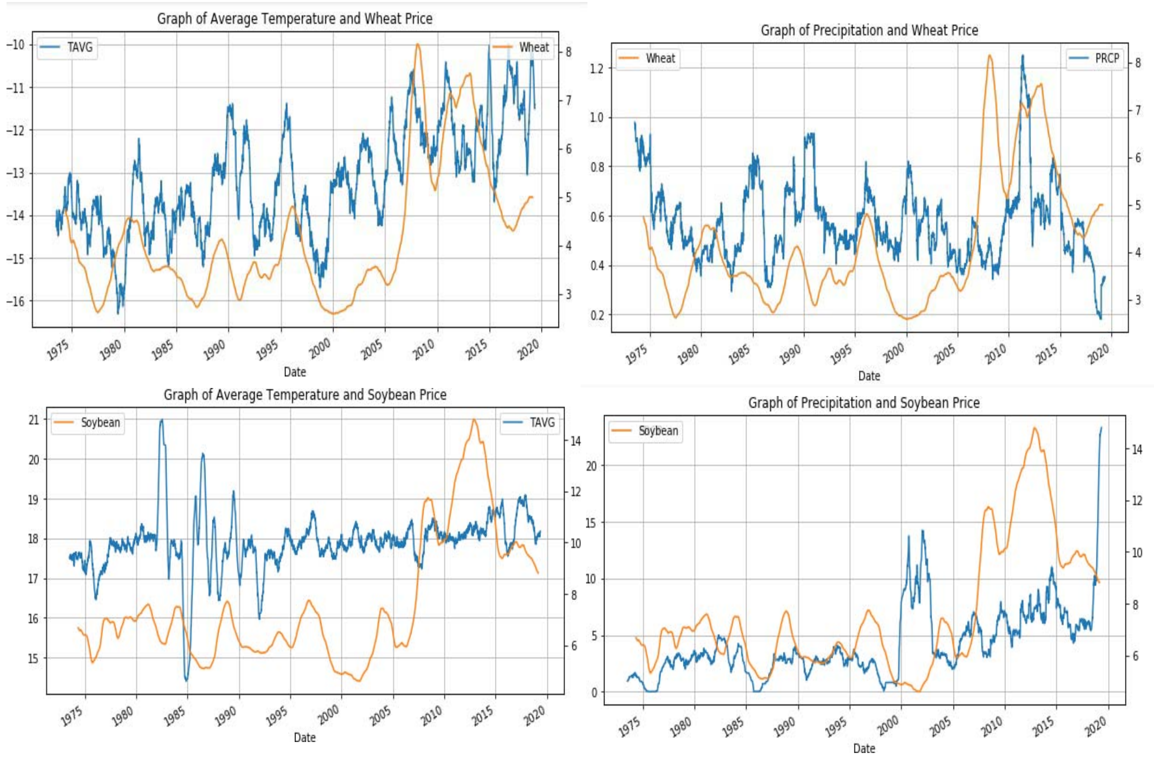


Figure 4.4b: Plot illustrating weather data and Wheat/Soybean prices

Weather data also characterize wheat and Soybean. For most of the features of the plot, Wheat and Soybean had maintained a low price where average temperature and precipitation had a spike in their value. A 2017/2018 study for wheat shows a deep in the price characterized by high temperature and precipitation.

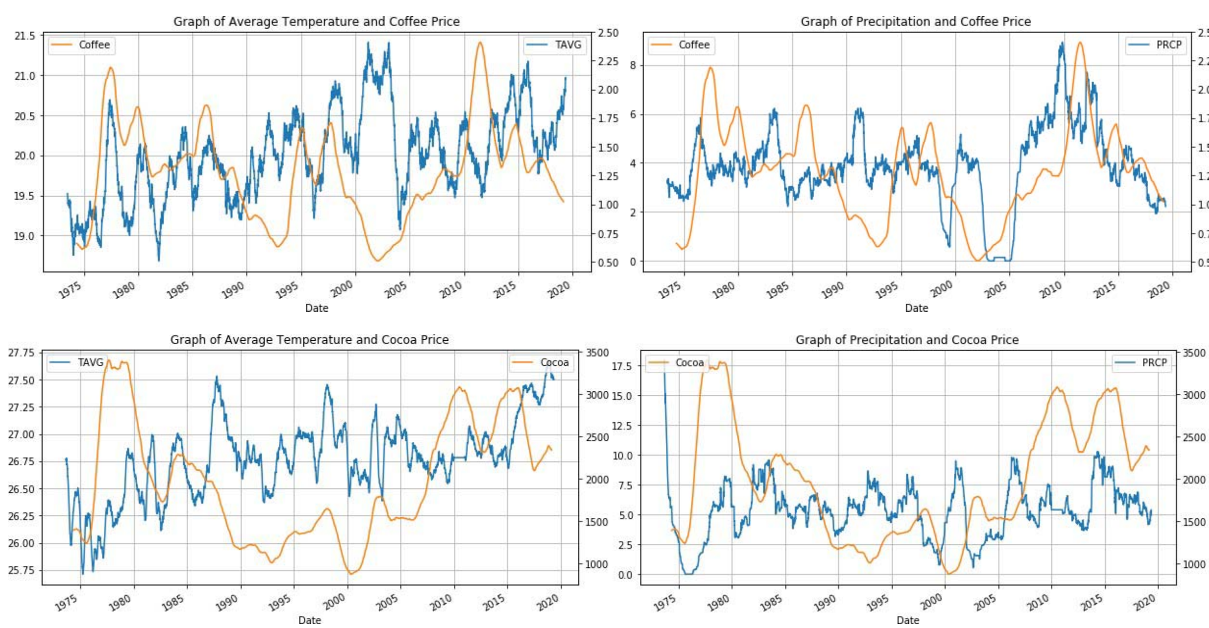


Figure 4.4c: Plot illustrating weather data and Coffee/Cocoa prices

Coffee and Cocoa weather chart has an interesting characteristic. The price data exhibit an inverse relationship with the average temperature, mostly during the last decade of the years under review. Cocoa, when compared to coffee, has a seasonal trend. The price data for the two commodities also exhibit an inverse relationship with the precipitation value. The closely related pairs had experience declined volatility (Figure 4.4 in appendix F), as shown in the latter part of the year under review.

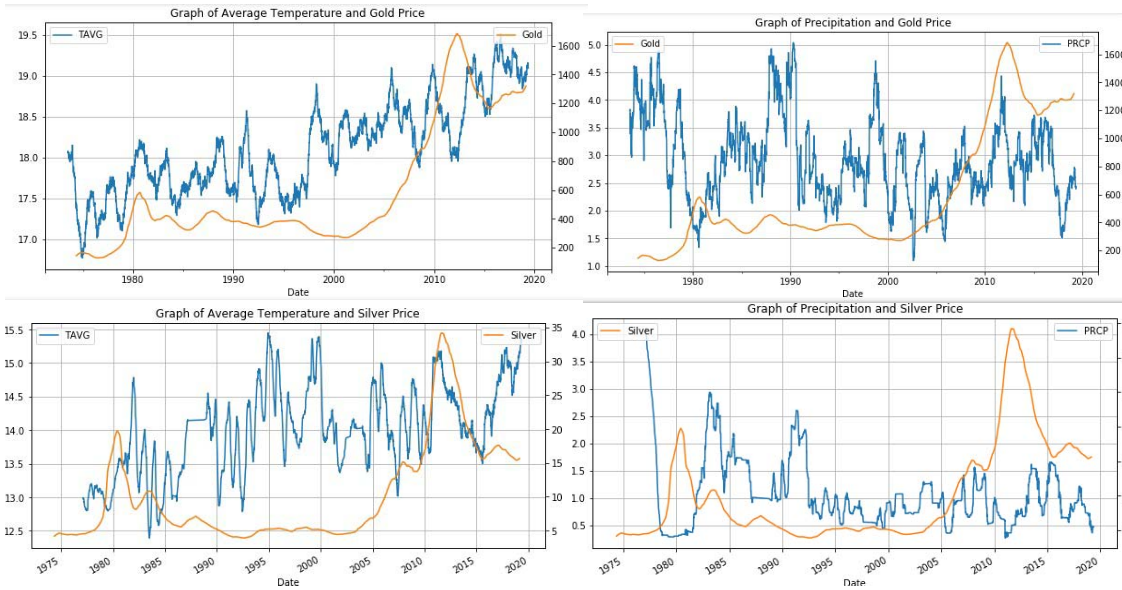


Figure 4.4d: Plot illustrating weather data and Gold/Silver prices

Gold and Silver constitute the only closely related non-agricultural commodity. Going by the plot of average temperature and precipitation for each commodity, the two pairs seem to exhibit no relationship with weather data. We can infer from these findings that weather data does not correlate with the price of the closely related product.

h) Model analysis

SARIMAX was selected to perform a regression analysis of commodity data and weather data. This model allows for a situation that requires a dependent variable (commodity price data) to be regressed with an exogenous variable (weather data). SARIMAX model was carried out on commodity data with their respective exogenous variable. Optimal parameters for our models was carried out by selecting optimal parameter values systematically using the grid search (hyperparameter optimization) method. It iteratively explores different combinations of the parameters, and for each discovery of parameters, we fitted a new seasonal ARIMA model with the SARIMAX() function and assessed its best value. The values for (p, d, q) (P, D, Q) in the SARIMAX model were selected to choose a combination with the lowest AIC (a more parsimonious model) while "m," seasonality period of 12 was used. The first part of the model (p, d, q) accommodates auto-regressive order, differencing order, and moving average order. While the second part of the model included a seasonal effect (P, D, Q), which is essential.

i. *Oats/corn and weather data*

As shown in figure 4.6a and 4.6b below, from the standardized residual plot of both commodities, the residual errors seem to fluctuate around a mean of zero and have a uniform variance. The residuals over time do not display any apparent seasonality and appear to be white noise.

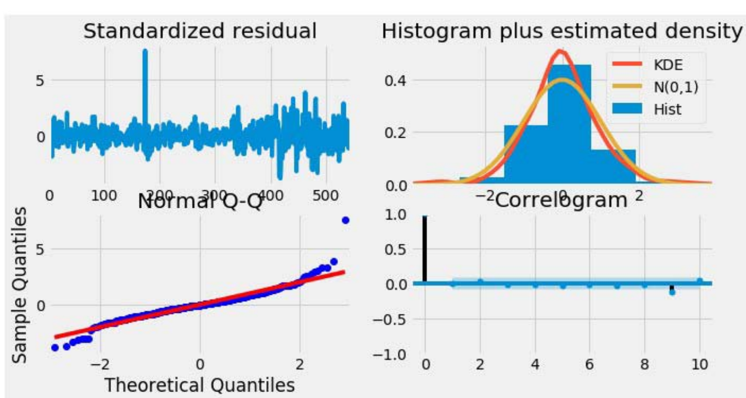


Figure 4.6a: Arima diagnostic plot for oats and temperature data

The density plot suggests normal distribution with a mean of zero. KDE line follows closely with the N (0,1) line. Where N (0,1) is the standard notation for a normal distribution with mean 0 and a standard deviation of 1. This implies that the residuals are normally distributed.

On the Q-Q plot, almost all the dots fell perfectly in line with the red line suggesting a normal distribution.

The Correlogram shows the residual errors are not autocorrelated for both corn and oats since there was no visible pattern, and the residuals have low correlation with lagged versions of itself.

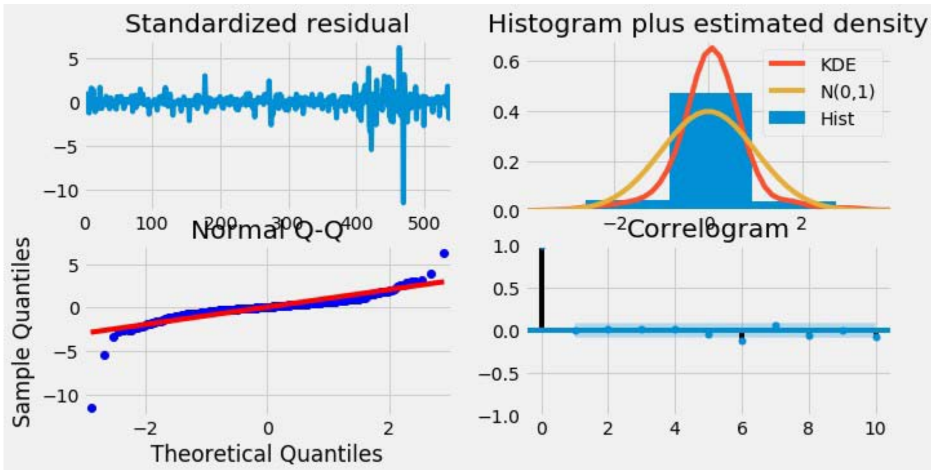


Figure 4.6b: Arima diagnostic plot for corn and temperature data

The log-likelihood of corn is -15.789 (result in appendix H), which is much lower in absolute value than that of oat, which had a log-likelihood of 250.425. That means that the regression of corn is a better fit for the data as compared to oats. The model has estimated that the AIC and the P values (<< 0.05) of the coefficients look significant. In summary, it seems to be a good fit. Also, oats and corn exhibit a negative coefficient in the temperature and precipitation coefficient, suggesting that as the independent variable increases, the dependent variable tends to decrease.

ii. *Wheat/Soybean and weather data*

As illustrated in figure 4.6c and 4.6d, the residual errors seem to fluctuate around a mean of zero and have a uniform variance. The density plot suggests normal distribution with a mean of zero.

On the Q-Q plot, almost all the dots fell perfectly in line with the red line suggesting a normal distribution.

The diagnostic tests report for wheat suggests that our residuals do not appear to be white noise - as such, we can reject at the 5% level the null hypotheses of serial independence (Ljung-Box test), Hetero-skedasticity test, and normality test.

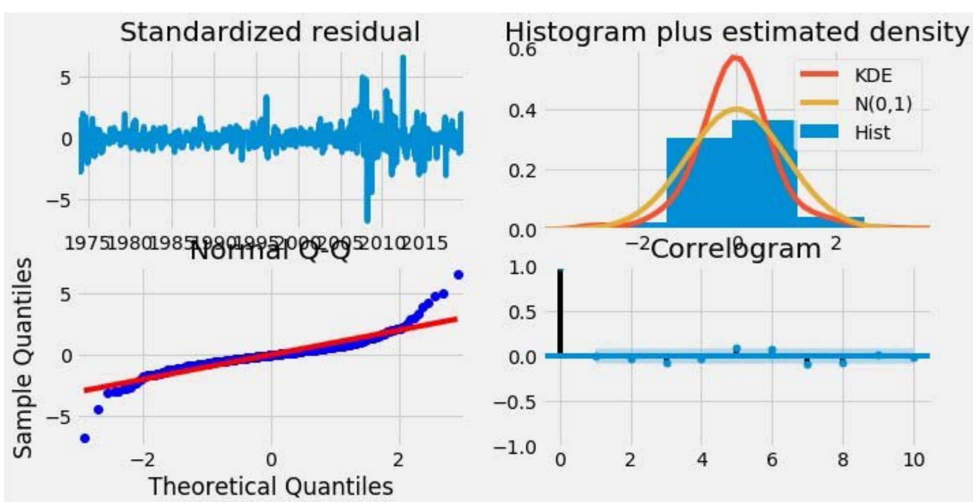


Figure 4.6c: Arima diagnostic plot for wheat and temperature data

The Correlogram shows the residual errors are not auto correlated for both wheat and soybean since there was no visible pattern.

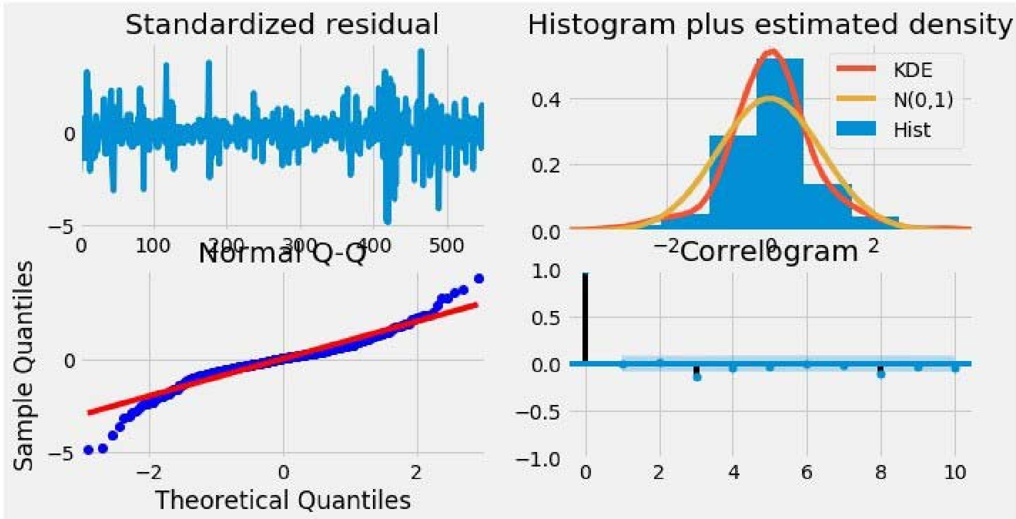


Figure 4.6d: Arima diagnostic plot for Soybean and temperature data

The log-likelihood of wheat is -146.335, which is much lower in absolute value than that of soybean, which had a log-likelihood of -375.201. That means that the regression of wheat is a better fit for the data as compared to the two commodities. The model has estimated that P values (<< 0.05) of the coefficients of wheat look more significant as compared to soybean. In summary, Wheat seems to be a better fit when compared to soybean.

iii. *Coffee/Cocoa and weather data*

For Coffee/Cocoa and weather data, the "coef" column of the model result (Appendix H) indicates the importance of each feature and how they contribute to the dataset. The P>|z| column informs us of the significance of each feature weight. As shown, coffee has some of the features with a p-value close to 0, while the exogenous data are not. With this, we may not satisfactorily conclude that the features should make up our model.

A model diagnostic was illustrated to make an informed assumption about the model further. The residual errors indicate that there may trend information not included by the model.

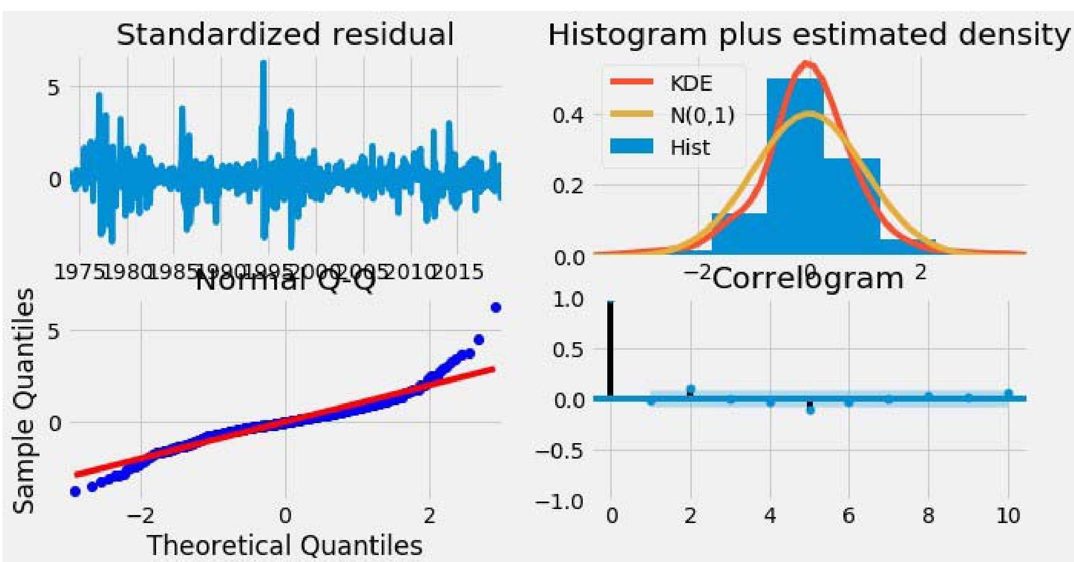


Figure 4.6e: Arima diagnostic plot for Coffee and temperature data

The QQ plot has most of its dot falling on the red line without any indication of a break. This is an indication that the residuals are normally distributed. The correlogram plot for the commodities shows that the residuals have a low correlation with lagged versions of itself. This implies that our model produces a satisfactory fit that could help us understand our dataset.

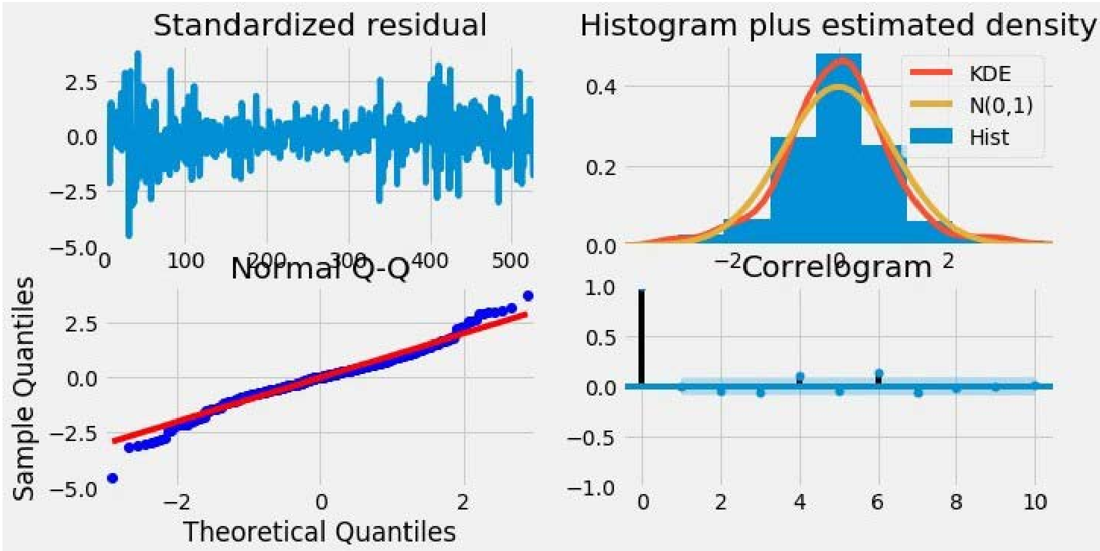


Figure 4.6f: Arima diagnostic plot for Cocoa and temperature data

Coffee has log-likelihood of the regression as 410.29, while Cocoa has it is to be - 3329.24. This is an indication that the regression of Coffee is a better fit for the data as compared to Cocoa.

iv. *Gold/Silver and weather data*

As shown in figure 4.6g and 4.6h below, the standardized residual plot for gold is characterized by spike at the beginning and end of the year, however, most of the time, the residual errors fluctuate around a mean of zero and have a uniform variance. For silver, residual error of the standard plot is characterized by a spike at the time of "silver Thursday" but maintained residual errors fluctuating around mean of zero and having a uniform variance.

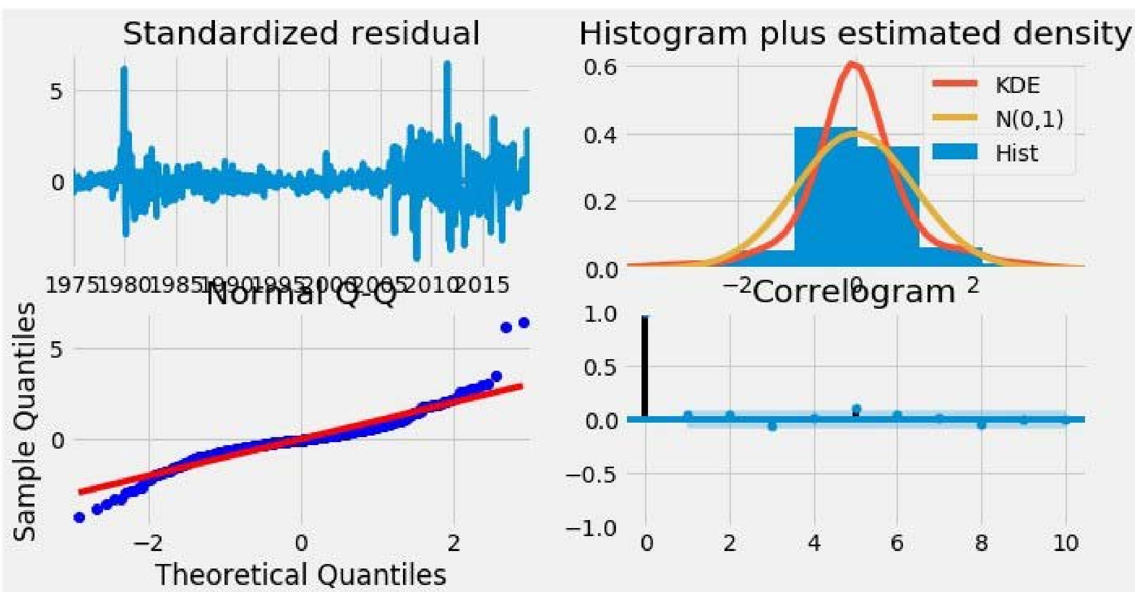


Figure 4.6g: Arima diagnostic plot for Gold and temperature data

For Gold/Silver, the density plot does not ultimately suggest a normal distribution as sample sizes of residuals are generally small (<50), so the histogram may not be the best choice for judging the distribution of the residuals.

A more sensitive graph is the Q-Q plot. Both plots are characterized by few departures from the red line, which is a normal probability plot are common. There are no visible breaks near the middle of this plot, and all dots seem to fall on the red line; hence, suggest normality in their residual distribution.

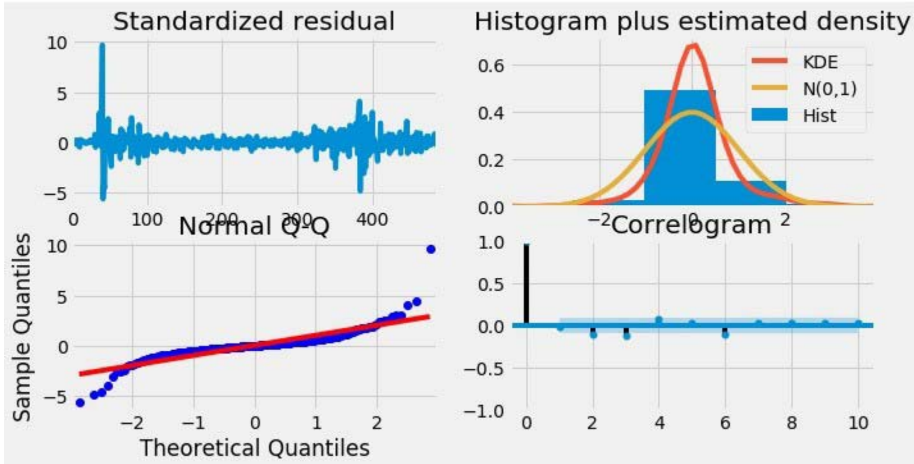


Figure 4.6h: Arima diagnostic plot for Silver and temperature data

In the Correlogram plot for both commodities, the correlations are very low (the y- axis goes from +1.0 to -1.0) and do not seem to have a pattern.

The log-likelihood of the regression for Gold is -2617.25, which is much higher in absolute value than that of Silver, which had a log-likelihood of -862.753. This means that the regression of Silver is a better fit for the data as compared to Gold.

V. Conclusion and Recommendation

This Capstone project was an Exploratory Data Analysis (EDA) that looked at daily historical price data of selected commodities and closely related products to understand their price relationships and the impact of weather on price variation, if any.

Time series data values are obtained at a sequential time interval. In other words, the values are characterized with or without increasing or decreasing trend and seasonality. A comparative analysis is a required time series analysis method to describe and extract information from time-descriptive data, and an informed decision could be made about the datasets.

A quantitative analysis using a time series model was used to check for the effect of weather on commodity dataset. We first came up with a plot of our time series dataset for the commodities to have an idea of the visual trend and seasonality of the series. We then use a descriptive statistic to check the raw data and its returns for the type of distribution.

The correlation test at different frequencies among commodity pairs was carried out on the returns to show how or whether chosen pairs are related. While we confirmed that some pairs are closely correlated, performing the test lead to which pairs are the strongest in terms of correlation. This justifies our selected closely related pairs for comparison.

Before regressing the price dataset with weather data, an essential technique of finding the value of SARIMAX (p, d, q) (P, D, Q) m was carried out to implement the model that best optimize our metric of interest. Optimal parameters for our model were carried out by selecting optimal parameter values systematically using the grid search (hyperparameter optimization) method.

While our model appreciably showed the relationship between the regression of commodity price data and temperature/precipitation. This is evident as the coefficients for the selected agricultural commodity price data tends to zero. However, the result provided for the non-agricultural commodity indicated that the temperature and precipitation data for these commodities (Gold and Silver) are highly insignificant, which can be seen in their respective p-values for the coefficients.

It is essential to point out at this junction that the project was never intended to forecast future commodities prices or weather patterns, but there is always room for future study in these areas

Furthermore, though there was some form of correlation between the temperature values and the agricultural commodities, the study is not sufficient to conclude whether or not the weather has a direct impact on the prices of these commodities. This is an area that is open to further study in the future by taking into consideration other commodity variables like yield and growth rate to investigate any indirect relationship to prices.

References Références Referencias

1. Addison, T., Ghoshray, A. and Stamatogiannis, MP 2016, 'Agricultural commodity price shocks and their effect on growth in sub-Saharan Africa,' Journal of Agricultural Economics, vol. 67, no. 1, pp. 47-61. [Online]. Available: https://doi.org/10.1111/1477-9552.12129
2. Adi Bhat. “Data collection methods: definition, examples, and sources." Available at: https://www.questionpro.com/blog/data-collection-methods/
3. Alan Hylands, 2019. "WTF Is Data Wrangling?". Available at; https://simpleanalytical.com/wtf-is-data-wrangling

4. Andrew Beattie, Investopedia (2019), Silver Thursday: How two wealthy traders cornered the Market, Online available at https://www.investopedia.com/articles/optioninvestor/09/silver-thursday-hunt-brothers.asp
5. Data collection. Available at; https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html
6. Easterling, W. E., P.K. Aggarwal, P. Batima, K.M. Brander, L. Erda, S.M. Howden, A. Kirilenko, J. Morton, J.-F. Soussana, J. Schmidhuber, and F.N. Tubiello (2007) 'Food, fiber and forest products,' in M.L Parry et al. (eds.) (2007). [Online] Available: https://pubs.giss.nasa.gov/abs/ea01000b.html
7. Ellen Friedman, Ted Dunning. "Time Series Databases: New Ways to Store and Access Data." [Online] Available at: https://www.oreilly.com/library/view/time-series-databases/9781491920909/ch01.html
8. Eva Ludi, Christopher Stevens, Leo Peskett, and Lídia Cabral Overseas Development Institute. 2007. Climate change and agriculture: Agricultural trade, markets, and investment. [Online] Available at: https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications- opinion-files/1884.pdf
9. Ghisellini, Patrizia & Setti, Marco & Ulgiati, Sergio. (2015). Energy and land use in worldwide agriculture: an application of life cycle energy and cluster analysis. Environment, Development, and Sustainability. 18. 10.1007/s10668-015-9678-2.
10. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=beef-and-veal- meat&graph=production.
11. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=beef-and-veal- meat&graph=production-growth-rate
12. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=beef-and-veal-meat&graph=exports.
13. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=beef-and-veal-meat&graph=domestic-consumption.
14. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=chicken-meat&graph=production
15. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=chicken-meat&graph=production- growth-rate.
16. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=chicken-meat&graph=exports
17. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=chicken-meat&graph=domestic- consumption.
18. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=palm-oil&graph=production.
19. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=palm-oil&graph=production-growth- rate.
20. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=palm-oil&graph=exports
21. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=palm-oil&graph=domestic- consumption.
22. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=peanut-oil&graph=production
23. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=peanut-oil&graph=production- growth-rate.
24. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=peanut-oil&graph=exports
25. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=peanut-oil&graph=domestic- consumption).
26. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=green-coffee&graph=production
27. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=green-coffee&graph=production- growth-rate.
28. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=green-coffee&graph=exports.
29. IndexMundi 2019. [Online]. Available: https://www.indexmundi.com/agriculture/?commodity=green-coffee&graph=domestic- consumption.
30. Investopedia (2019), 10 Countries that Produces the Most Silver. Online available at https://www.investopedia.com/articles/markets-economy/083116/10-countries-produce- most-silver.asp.
31. J. P. Lhomme, R. Mougou, Mohsen Mansour (2009). "Potential impact of climate change on durum wheat cropping in Tunisia" Available at: https://www.researchgate.net/publication/225320259_Potential_impact_of_climate_change_ on_durum_wheat_cropping_in_Tunisia.
32. James Chen, Investopedia (2018), Silver Thursday, Online available at https://www.investopedia.com/terms/s/silver_thursday.asp.
33. Jasmien De Winne, Gert Peersman 2018. Agricultural Price Shocks and Business Cycles; A Global Warning for Advanced Economies. CESifo Working Papers, ISSN 2364-1428. [Online] Available at: https://www.cesifo.org/DocDL/cesifo1_wp7037.pdf
34. Jeffrey W. White, Gerrit Hoogenboom (2009) "Crop Response to Climate: Ecophysiological Models" Available at: https://link.springer.com/chapter/10.1007%2F978-90-481-2953-9_4.

35. Maria De Salvo, Diego Begalli, Giovanni Signorello.(2013). "Measuring the effect of climate change on agriculture." Available at: https://www.researchgate.net/publication/258437340
36. Masters, G.; Baker, P.; Flood, J. 2010. Climate change and agricultural commodities. CABI Working Paper 2, 38 pp. [Online]. Available: https://www.cabi.org/Uploads/CABI/expertise/invasive-alien-species-working-paper.pdf
37. Mendelsohn, R., Nordhaus, W. D., & Shaw, D. (1994). The Impact of Global Warming on Agriculture: A Ricardian Analysis. American Economic Review, 84(4), 753-771. [Online]. Available: https://www.jstor.org/stable/2118029?seq=1#page_scan_tab_contents
38. Mendelsohn, R., & Dinar, A. (2003). Climate, water, and agriculture. Land Economics, 79(3), 328-341. [Online] Available at: https://www.jstor.org/stable/3147020?seq=1#page_scan_tab_contents
39. Mendelsohn, R. (2007). Measuring climate impacts with cross-sectional analysis. Climatic Change, 81, 1-7.
40. Mendelsohn, R., & Dinar, A. (2009). Climate Change and Agriculture: An Economic Analysis of Global Impacts, Adaptation, and Distributional Effects. Cheltenham UK Northampton, MA, USA: Edward Elgar.
41. Munasinghe, Lalith, Tackseung Jun, and David H. Rind (2012) "Climate change: A new metric to measure changes in the frequency of extreme temperatures using record data," Climatic Change, Vol. 113, pp. 1001–1024.
42. Ritchie, J., Singh, U., Goodwin, D., & Hunt, L. (1989). A User's Guide to CERES Maize 2.10. Muscle Shoals, AL: Michigan State University-IFDC-IBSNAT. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-0862.2002.tb00104.x
43. Robert R.F. DeFilippi, (2018). SARIMA Modelling for Car Sharing — Basic Data Pipelines Applications with Python Pt. 1. Available at; https://medium.com/@rrfd/sarima- modelling-for-car-sharing-basic-data-pipelines-applications-with-python-pt-1-75de4677c0cd R
44. Rosenzweig, C., Iglesias, A., Fischer, G. et al. Environmental Modeling & Assessment (1999) 4: 115. [Online]. Available at: https://doi.org/10.1023/A:1019008116251
45. Rosenzweig, C., F.N. Tubiello, R.A. Goldberg, E. Mills, and J. Bloomfield (2002) 'Increased crop damage in the US from excess precipitation under climate change,' Global Environmental Change 12(2002): 197–202. [Online]. Available: https://pubs.giss.nasa.gov/abs/ro02200i.html
46. Porter, J.R. and MA Semenov (2005) 'Crop responses to climatic variation,' Philosophical Transactions of the Royal Society B 360(1463): 2021–35. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569569/
47. Poudel, Santosh. (2013). Climatic impacts on crop yield and its variability in Nepal: Do they vary across seasons and altitudes?. Climatic Change. 116. 327-355. 10.1007/s10584-012- 0491-8. Available: https://www.researchgate.net/publication/257547916_Climatic_impacts_on_crop_yield_and_its_variability_in_Nepal_Do_they_vary_across_seasons_and_altitudes
48. Schneider, S. H., Easterling, W. E., & Mearns, L. O. (2000). Adaptation Sensitivity to Natural Variability, Agent Assumption, and Dynamic Climatic Changes. Climatic Change 45, 203-221. [Online]. Available at: https://link.springer.com/article/10.1023/A:1005657421149.
49. Silverprice (2002). "Online resources for Silver data." Available at: https://silverprice.org/silver-price-history.html
50. U.S. Global Investors (2019), Top 10 Gold Producing Countries. Online available at http://www.usfunds.com/investor-library/frank-talk/top-10-gold-producing-countries/
51. WorldAtlas (2019), Top 10 Cocoa Producing Countries. Online available at https://www.worldatlas.com/articles/top-10-cocoa-producing-countries.html).

Appendix A

Descriptive statistic of raw data

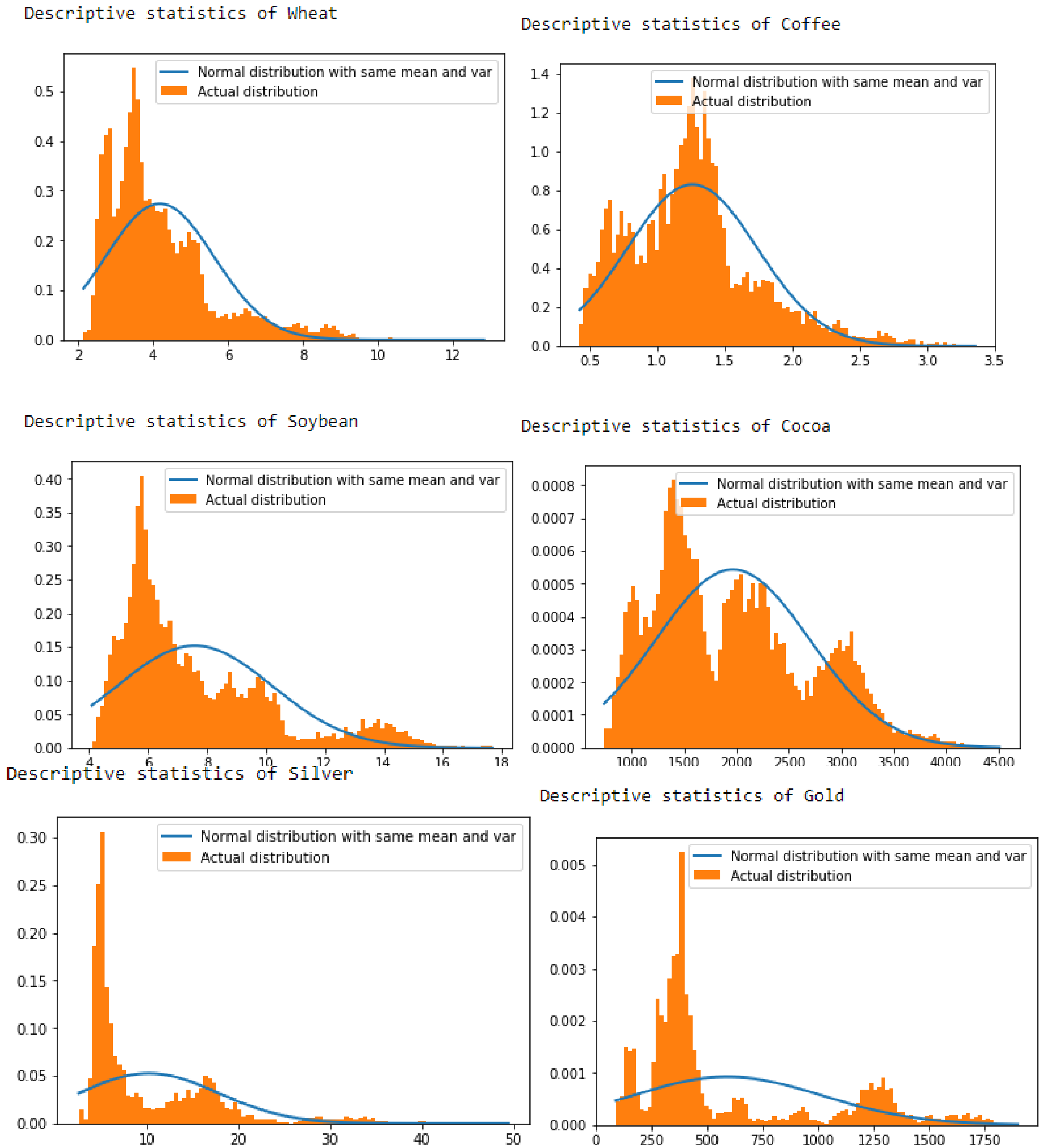


Figure 4.0: Normal distribution test for commodity price data

Appendix B

Descriptive statistic of return

Table 4.2b: Descriptive statistics of weekly return

	Oats_W	Corn_W	Wheat_W	Soybean_W	Coffee_W
count	2412.000000	2412.000000	2412.000000	2412.000000	2412.000000
mean	0.001370	0.000707	0.000678	0.000653	0.001336
std	0.043824	0.034695	0.037228	0.034132	0.048383
min	-0.186851	-0.163904	-0.171213	-0.162380	-0.178421
25%	-0.022706	-0.017162	-0.022785	-0.017109	-0.025575
50%	0.000000	0.001026	-0.001024	0.001568	0.000367
75%	0.025742	0.018060	0.020859	0.019969	0.026161
max	0.323155	0.207761	0.197068	0.164715	0.549424

	Cocoa_W	Gold_W	Silver_W
count	2412.000000	2412.000000	2412.000000
mean	0.001034	0.001460	0.001902
std	0.038731	0.026592	0.048207
min	-0.204117	-0.165214	-0.435206
25%	-0.021614	-0.011838	-0.019098
50%	-0.000450	0.001151	0.000235
75%	0.022711	0.013608	0.021359
max	0.196182	0.353279	0.891512

Table 4.2c: Descriptive statistics of weekly return

	Oats_M	Corn_M	Wheat_M	Soybean_M	Coffee_M	Cocoa_M
count	555.000000	555.000000	555.000000	555.000000	555.000000	555.000000
mean	0.005802	0.003445	0.002984	0.003387	0.006115	0.004689
std	0.093528	0.074636	0.077403	0.076944	0.105188	0.082740
min	-0.281658	-0.225996	-0.248192	-0.279766	-0.361795	-0.318365
25%	-0.046174	-0.035670	-0.048558	-0.035782	-0.059641	-0.051973
50%	-0.000771	-0.000357	0.001394	0.001362	-0.004656	-0.000929
75%	0.050303	0.046855	0.046292	0.042049	0.053132	0.053326
max	0.943014	0.496544	0.386356	0.570313	0.533759	0.334545

	Gold_M	Silver_M
count	555.000000	555.000000
mean	0.006338	0.007836
std	0.056593	0.094331
min	-0.213672	-0.620101
25%	-0.025116	-0.042227
50%	0.000260	-0.002205
75%	0.034325	0.050445
max	0.274809	0.694737

Appendix C

Skewness and kurtosis of returns

Table 4.3c: Skewness and Kurtosis for weekly returns

SKEWNESS and KURTOSIS RESULT ON WEEKLY RETURNS

Commodities	Skewness	Kurtosis
Oats	0.269816	2.7101
Corn	0.0771866	2.77241
Wheat	0.397143	2.04088
Soybean	-0.176502	2.31248
Coffee	1.0185	9.4295
Cocoa	0.109242	1.45824
Gold	0.983068	16.5317
Silver	2.41927	55.997

Table 4.3d: Skewness and Kurtosis for monthly returns

SKEWNESS and KURTOSIS RESULT ON MONTHLY RETURNS

Commodities	Skewness	Kurtosis
Oats	2.04217	18.3864
Corn	0.632904	3.96188
Wheat	0.390539	1.47566
Soybean	0.616611	5.98835
Coffee	1.15048	3.82862
Cocoa	0.459498	1.16766
Gold	0.738831	4.12935
Silver	0.850449	11.2526

Appendix D

Correlation matrix of return

CORRELATION MATRIX OF WEEKLY RETURNS

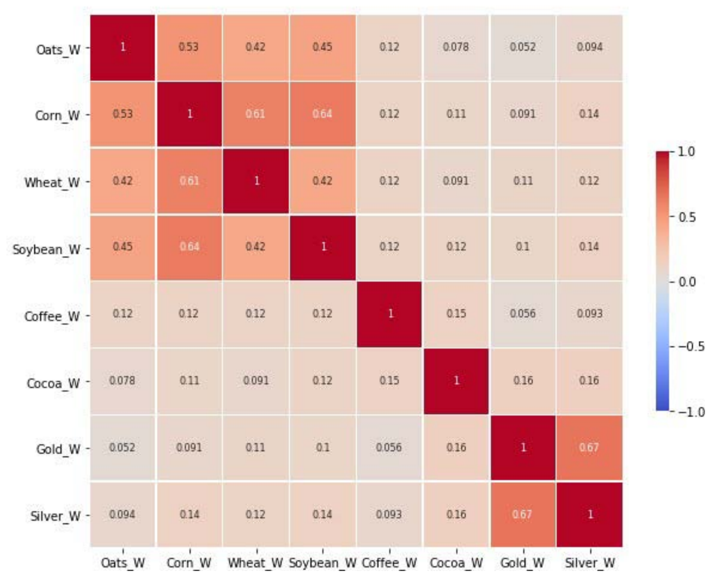


Figure 4.2b: Correlation matrix for daily returns

CORRELATION MATRIX OF MONTHLY RETURNS

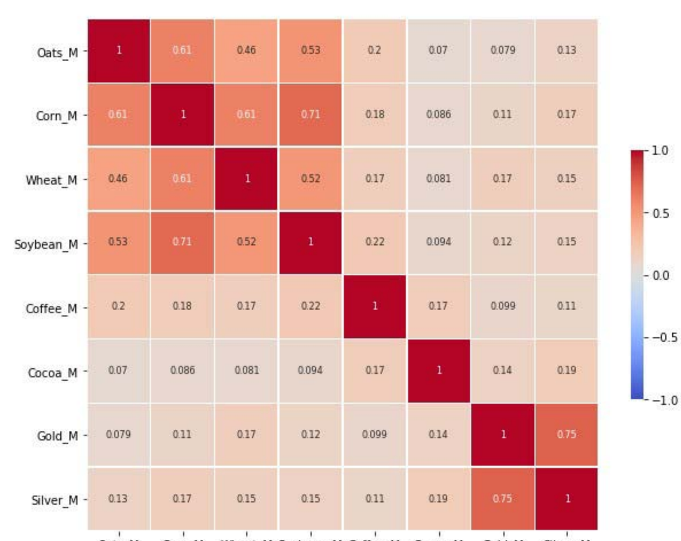


Figure 4.2b: Correlation matrix for daily returns

Appendix E

Augmented Dickey-Fuller test

```
    Augmented Dickey-Fuller Test o
n "Wheat"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -78.4073
 No. Lags Chosen        = 1
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

```
    Augmented Dickey-Fuller Test o
n "Cocoa"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -45.1571
 No. Lags Chosen        = 5
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

```
 Augmented Dickey-Fuller Test on "
Soybean"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -19.8719
 No. Lags Chosen        = 28
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

```
 Augmented Dickey-Fuller Test on "
Gold"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -16.6773
 No. Lags Chosen        = 40
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

```
    Augmented Dickey-Fuller Test o
n "Coffee"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -24.7959
 No. Lags Chosen        = 22
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

```
 Augmented Dickey-Fuller Test on "
Silver"
    -------------------------------
-----------------
 Null Hypothesis: Data has unit ro
ot. Non-Stationary.
 Significance Level     = 0.05
 Test Statistic         = -16.6123
 No. Lags Chosen        = 38
 Critical value 1%      = -3.431
 Critical value 5%      = -2.862
 Critical value 10%     = -2.567
 => P-Value = 0.0. Rejecting Null
Hypothesis.
 => Series is Stationary.
```

Figure 4.3a: ADF test report

Appendix F

Graph of product pair

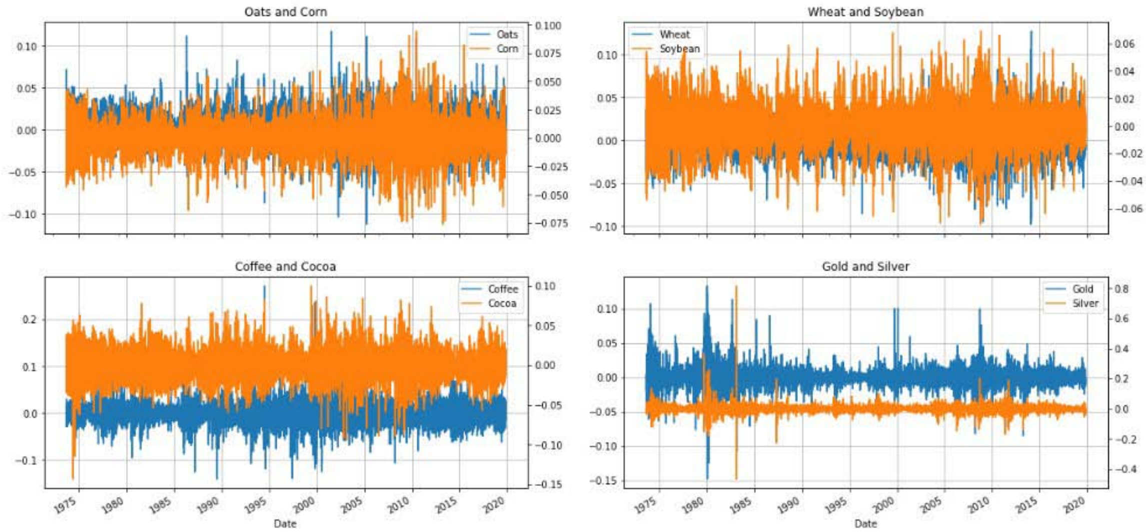


Figure 4.4: Plot of product pair returns

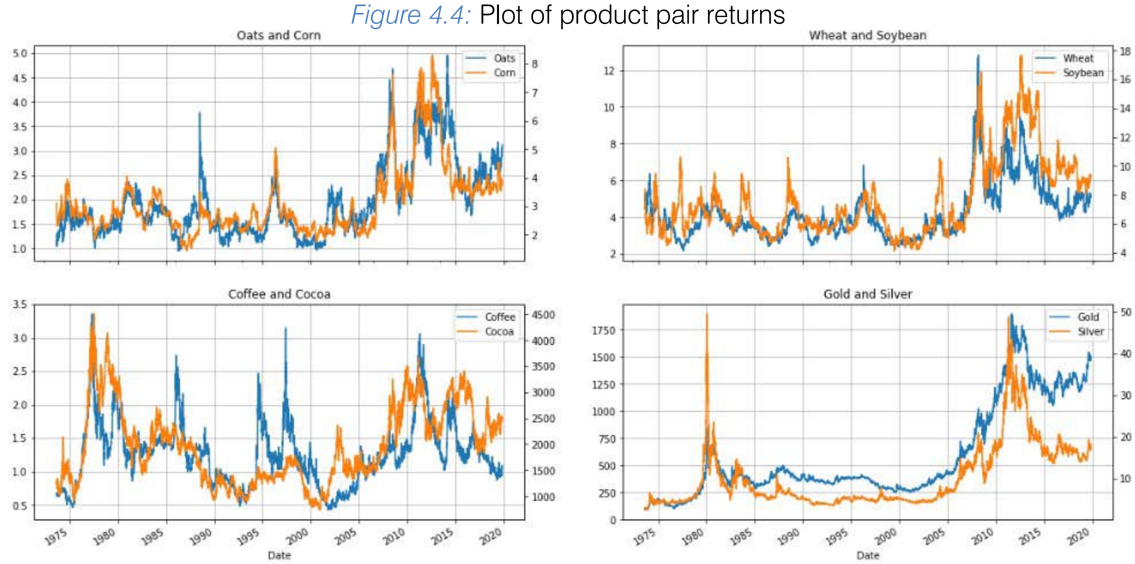


Figure 4.4e: Pair plot for a closely related commodity (price data)

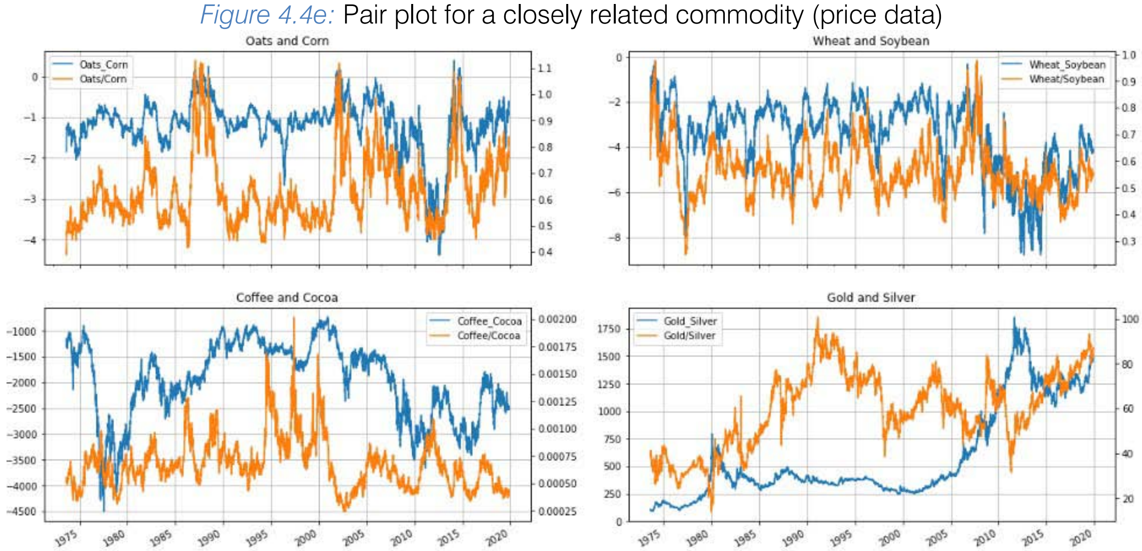
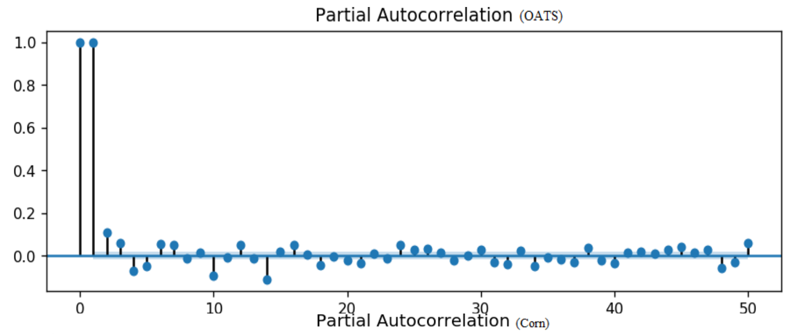
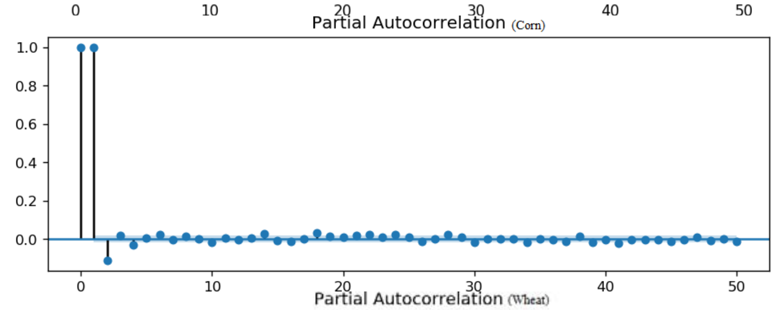


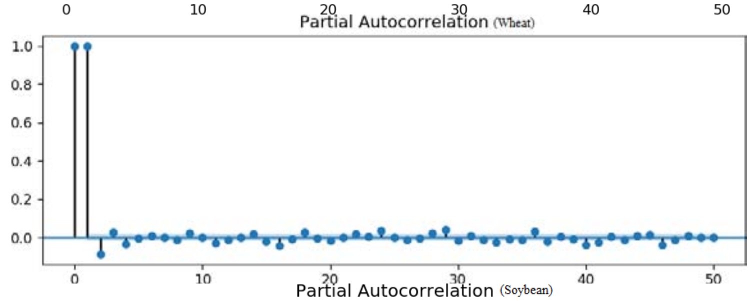
Figure 4.4f: Plot of Spread and closely related product ratio

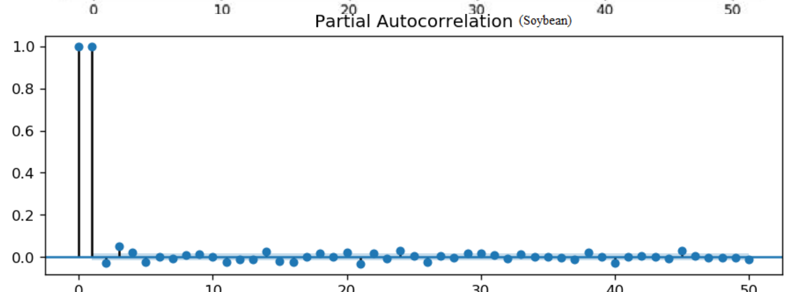
Appendix G

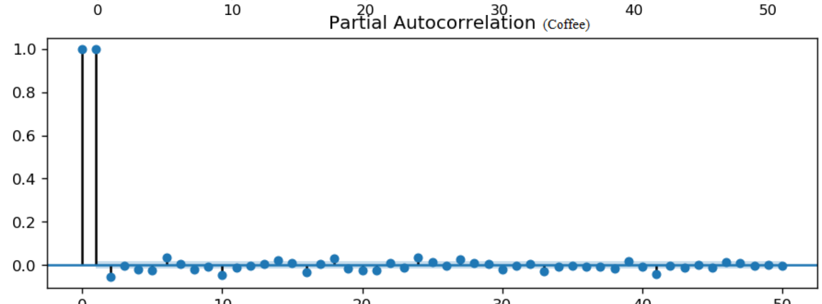
PAC plot

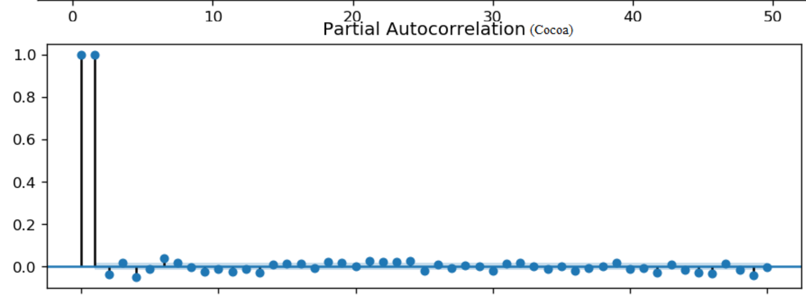












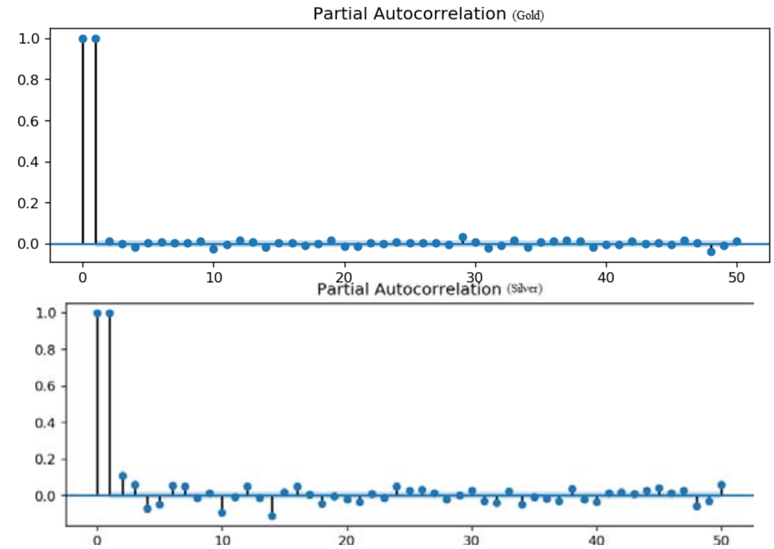


Figure 4.5f: Partial autocorrelation plot for commodity price data

Appendix H

Model report using SARIMAX

Dep. Variable:	Oats	**No. Observations:**	547
Model:	SARIMAX(1, 1, 0)	**Log Likelihood**	250.425
Date:	Sun, 05 Jan 2020	**AIC**	-492.850
Time:	23:41:12	**BIC**	-475.639
Sample:	0	**HQIC**	-486.122
	- 547		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
TAVG	-0.0009	0.001	-0.798	0.425	-0.003	0.001
PRCP	-0.0059	0.005	-1.299	0.194	-0.015	0.003
ar.L1	0.1350	0.035	3.900	0.000	0.067	0.203
sigma2	0.0234	0.001	33.277	0.000	0.022	0.025

Ljung-Box (Q):	48.63	**Jarque-Bera (JB):**	1219.95
Prob(Q):	0.16	**Prob(JB):**	0.00
Heteroskedasticity (H):	2.07	**Skew:**	0.78
Prob(H) (two-sided):	0.00	**Kurtosis:**	10.15

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7a: Arima model for oat and weather data

Dep. Variable:	Corn	No. Observations:	540
Model:	SARIMAX(1, 1, 0)	Log Likelihood	-15.785
Date:	Sun, 05 Jan 2020	AIC	39.571
Time:	22:50:25	BIC	56.730
Sample:	0	HQIC	46.282
	- 540		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	-0.0022	0.005	-0.461	0.645	-0.011	0.007
PRCP	0.0001	0.002	0.054	0.957	-0.004	0.005
ar.L1	0.1970	0.023	8.428	0.000	0.151	0.243
sigma2	0.0621	0.001	53.206	0.000	0.060	0.064

Ljung-Box (Q):	54.45	Jarque-Bera (JB):	32837.67
Prob(Q):	0.06	Prob(JB):	0.00
Heteroskedasticity (H):	6.20	Skew:	-2.70
Prob(H) (two-sided):	0.00	Kurtosis:	40.85

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7b: Arima model for corn and weather data

Dep. Variable:	Wheat	**No. Observations:**	556
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	**Log Likelihood**	-146.335
Date:	Sun, 05 Jan 2020	**AIC**	306.669
Time:	22:57:02	**BIC**	336.902
Sample:	08-01-1973	**HQIC**	318.479
	- 11-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	-0.0016	0.002	-0.916	0.360	-0.005	0.002
PRCP	-0.0265	0.029	-0.923	0.356	-0.083	0.030
ar.L1	-0.3010	0.099	-3.028	0.002	-0.496	-0.106
ma.L1	0.5204	0.082	6.383	0.000	0.361	0.680
ar.S.L12	0.2041	0.528	0.387	0.699	-0.831	1.239
ma.S.L12	-0.2748	0.517	-0.531	0.595	-1.289	0.739
sigma2	0.0992	0.003	33.796	0.000	0.093	0.105

Ljung-Box (Q):	68.26	**Jarque-Bera (JB):**	2465.06
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	4.25	**Skew:**	0.43
Prob(H) (two-sided):	0.00	**Kurtosis:**	13.29

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7c: Arima model for wheat and weather data

Dep. Variable:	Soybean	**No. Observations:**	555
Model:	SARIMAX(1, 1, 1)x(0, 0, 1, 12)	**Log Likelihood**	-375.201
Date:	Sun, 05 Jan 2020	**AIC**	762.402
Time:	22:59:59	**BIC**	788.305
Sample:	0	**HQIC**	772.521
	- 555		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
TAVG	-0.0039	0.009	-0.435	0.663	-0.021	0.014
PRCP	-0.0024	0.003	-0.785	0.433	-0.008	0.004
ar.L1	0.2415	0.078	3.086	0.002	0.088	0.395
ma.L1	0.1032	0.081	1.277	0.202	-0.055	0.262
ma.S.L12	0.0113	0.036	0.312	0.755	-0.060	0.082
sigma2	0.2268	0.008	27.864	0.000	0.211	0.243

Ljung-Box (Q):	45.34	**Jarque-Bera (JB):**	368.75
Prob(Q):	0.26	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.73	**Skew:**	-0.31
Prob(H) (two-sided):	0.00	**Kurtosis:**	6.95

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7d: Arima model for Soybean and weather data

Dep. Variable:	Coffee	**No. Observations:**	556
Model:	SARIMAX(1, 1, 0)	**Log Likelihood**	410.294
Date:	Sun, 05 Jan 2020	**AIC**	-812.588
Time:	23:48:46	**BIC**	-795.312
Sample:	08-01-1973	**HQIC**	-805.839
	- 11-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	0.0016	0.003	0.539	0.590	-0.004	0.007
PRCP	-0.0017	0.001	-1.543	0.123	-0.004	0.000
ar.L1	0.2118	0.024	8.951	0.000	0.165	0.258
sigma2	0.0133	0.000	31.030	0.000	0.013	0.014

Ljung-Box (Q):	64.24	**Jarque-Bera (JB):**	634.37
Prob(Q):	0.01	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.50	**Skew:**	0.58
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.11

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7e: Arima model for Coffee and weather data

Dep. Variable:	Cocoa	No. Observations:	544
Model:	SARIMAX(0, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-3329.215
Date:	Sun, 05 Jan 2020	AIC	6670.430
Time:	23:15:45	BIC	6696.078
Sample:	0	HQIC	6680.468
	- 544		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	3.2803	7.101	0.462	0.644	-10.637	17.198
PRCP	1.1317	0.814	1.390	0.165	-0.464	2.728
ma.L1	0.2544	0.035	7.227	0.000	0.185	0.323
ar.S.L12	-0.0718	0.042	-1.713	0.087	-0.154	0.010
ma.S.L12	-1.0000	22.805	-0.044	0.965	-45.698	43.698
sigma2	1.496e+04	3.41e+05	0.044	0.965	-6.54e+05	6.84e+05

Ljung-Box (Q):	49.56	Jarque-Bera (JB):	80.73
Prob(Q):	0.14	Prob(JB):	0.00
Heteroskedasticity (H):	0.79	Skew:	-0.10
Prob(H) (two-sided):	0.11	Kurtosis:	4.90

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7f: Arima model for Cocoa and weather data

Dep. Variable:	Gold	No. Observations:	556
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-2617.251
Date:	Sun, 05 Jan 2020	AIC	5246.501
Time:	23:19:52	BIC	5272.284
Sample:	08-01-1973	HQIC	5256.582
	- 11-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	1.8066	1.137	1.589	0.112	-0.422	4.035
PRCP	-0.0452	0.344	-0.132	0.895	-0.719	0.628
ar.L1	-0.5906	0.067	-8.811	0.000	-0.722	-0.459
ma.L1	0.7974	0.051	15.773	0.000	0.698	0.897
ma.S.L12	-1.0282	0.031	-33.629	0.000	-1.088	-0.968
sigma2	798.7640	41.556	19.222	0.000	717.316	880.212

Ljung-Box (Q):	51.90	Jarque-Bera (JB):	1125.66
Prob(Q):	0.10	Prob(JB):	0.00
Heteroskedasticity (H):	2.88	Skew:	0.70
Prob(H) (two-sided):	0.00	Kurtosis:	9.91

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7g: Arima model for Gold and weather data

Dep. Variable:	Silver	No. Observations:	499
Model:	SARIMAX(1, 0, 1)x(0, 1, 1, 12)	Log Likelihood	-862.753
Date:	Sun, 05 Jan 2020	AIC	1737.507
Time:	23:23:04	BIC	1762.636
Sample:	0	HQIC	1747.378
	- 499		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TAVG	-0.0735	0.030	-2.457	0.014	-0.132	-0.015
PRCP	-0.0153	0.033	-0.461	0.644	-0.080	0.050
ar.L1	0.9677	0.006	165.937	0.000	0.956	0.979
ma.L1	0.4313	0.016	27.289	0.000	0.400	0.462
ma.S.L12	-0.9516	0.030	-31.738	0.000	-1.010	-0.893
sigma2	1.9045	0.058	32.853	0.000	1.791	2.018

Ljung-Box (Q):	43.15	Jarque-Bera (JB):	10833.71
Prob(Q):	0.34	Prob(JB):	0.00
Heteroskedasticity (H):	0.77	Skew:	1.34
Prob(H) (two-sided):	0.10	Kurtosis:	25.95

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 4.7h: Arima model for Silver and weather data