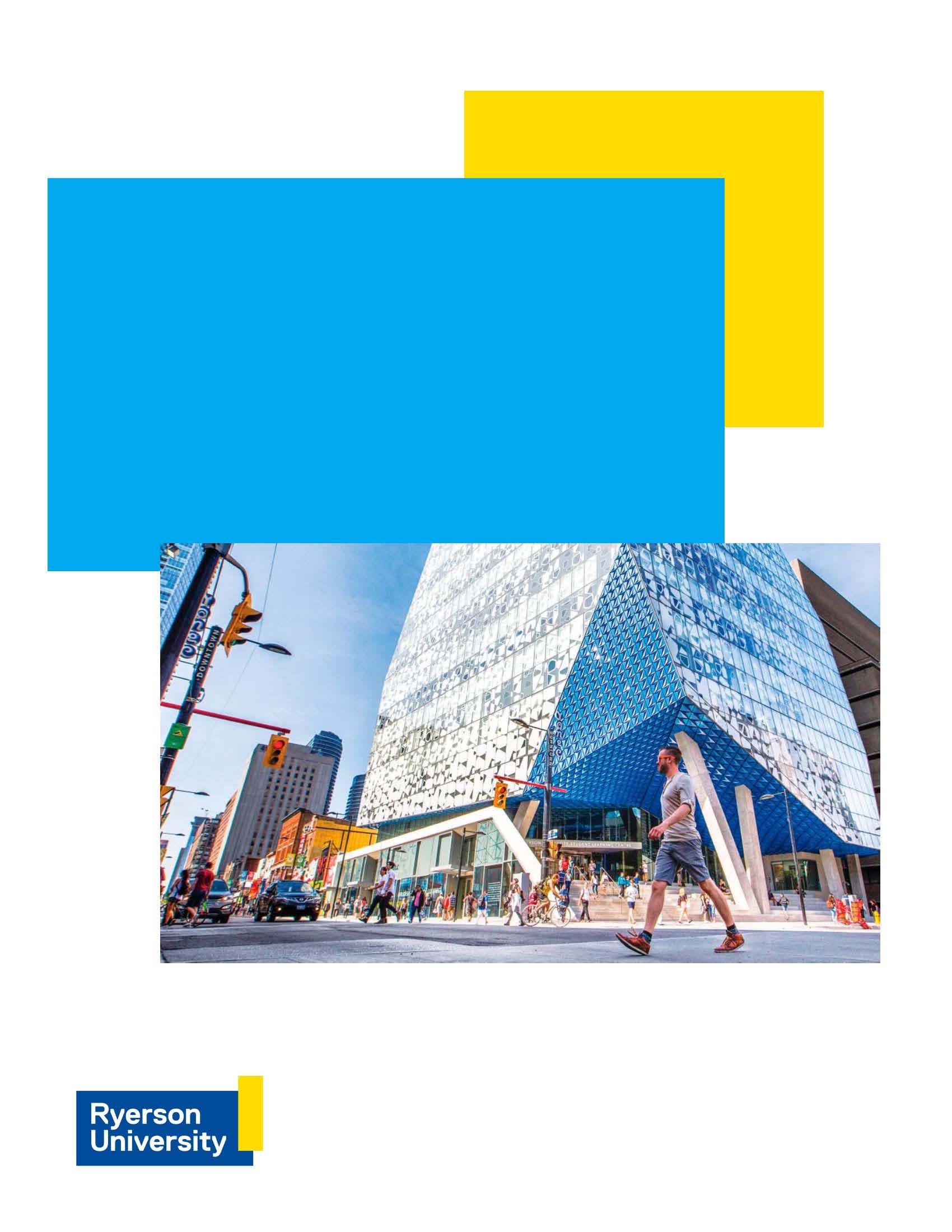
EDA and Prediction of International Coffee Prices with a Variety of Algorithms of Machine

Simon Atehortua. ID 501094927

Supervisor: Ceni Babaoglu, Ph.D.

October 24, 2022

**Table of Contents**

[**Links**](#_slt2itrqwyyv) **3**

[GitHub Link](#_dgyazsjvl1o7) 3

[Data Set Link](#_s6opd5o5b83o) 3

[**Abstract**](#_6oh3ihv0hrll) **4**

[**Literature Review**](#_qkgrg984vyjh) **6**

[**Overall Methodology**](#_712qeenzdvi4) **9**

[Diagram of an overall methodology](#_viqhwta4jp17) 9

[**Data Acquisition**](#_i1ehs3eb1sfx) **10**

[Columns Metadata](#_41qsuesb0q00) 12

[**Multivariate Analysis**](#_frdmfh4yapve) **14**

[Summary Statistics](#_k3emm2hpzj2r) 14

[Min, Max, and Outliers](#_4zcds9dhk07e) 16

[Boxplot or candle sticks](#_5lho099wl8qz) 17

[Frequency Distribution](#_w008zfatfbjg) 18

[Correlation and Heatmap](#_hb9yutoevbx0) 20

[**Exploratory Data Analysis**](#_qtg3z7vt87ud) **23**

[**Features Selection and Backward Pairwise Relations**](#_n9q6wxyj5cf3) **26**

[**Normalization**](#_nbslnk7e86d2) **30**

[**References**](#_ivsfcs3c3ew8) **33**

# Links

## GitHub Link

<https://github.com/montwa/Ryerson>

## Data Set Link

<https://federaciondecafeteros.org/app/uploads/2020/01/Precios-%C3%A1rea-y-producci%C3%B3n-de-caf%C3%A9-3.xlsx>

**Excel Sheet:**  6. Precio OIC Mensual

## 

# 

# Abstract

Coffee is one of the greatest drinks that humanity could have ever discovered. The prices of coffee are expressed in US cents per pound of green coffee and many factors affect the prices. The ICO (International Coffee Organization), will collect daily prices from the New York, Germany, and France futures exchanges and will set up a price based on these factors and a variety of conditions.

The idea of this project and after doing an exploratory analysis of the data is to predict the prices of the ICO market given the prices in Europe, more exactly in the countries of Germany and France, as well as in the United States of America of a different variety of Coffee that is grouped as follows: Colombian Mild Arabicas (Colombian Excelso UGQ screen size 14, Colombian Excelso European preparation screen size 15), Other Mild Arabicas (Costa Rica hard bean, Mexico Prime washed, Honduras high grown, Guatemala prime washed, El Salvador Strictly High grown, Guatemala hard bean, Honduras High grown European preparation ) Brazilian (Brazil Santos ¾ screen size 14/16, Brazil Santos ⅔ screen size 17/18, Brasil Santos ¾ screen size 14/16 ) and Robustas (Vietnam grade 2, Indonesia EK grade 4, Uganda Standard, Côte d’Ivoire grade 2).

All data for this project comes from Federación Nacional De Cafeteros de Colombia (The Colombian National Coffee Growers), which is the only authorized entity by the government of Colombia, that is responsible for exporting, buying, and setting the prices of coffee, the website (<https://federaciondecafeteros.org/wp/coffee-statistics/?lang=en>), section Coffee prices, area, and production have an excel file with the data required during this project. The data will be cleaned and checked for any inconsistencies or errors that can alter our results, after that, there will be an Exploratory Data Analysis to check the behavior of the time series data, and to possibly discover new information that could help find important and relevant information about our study. Through a variety of different Machine Learning regression algorithms like linear regression, SVM (Support Vector Machine), KNN regressor (K-Nearest Neighbors), Decision Tree regressor, and Neural Networks, I will intend to predict the future price of the coffee set by the International Coffee Organization.

A variety of tools will be used for this project: Jupyter Notebook, Python, R studio, Microsoft Excel, Github, and Google Documents.

During this project, a variety of questions have to be formulated in order to obtain the best answer to our results. An example of these questions, among many others, are: which is the most accurate algorithm to predict the ICO prices of coffee and why the selection of this algorithm? What will be the future price in the short term given that the model has been adequately trained and formulated? Why is this investigation relevant to the student and how it can help to develop his knowledge not only in exploratory data analysis but as well in machine learning algorithms?

*Keywords:* coffee, exploratory data analysis, regression analysis, prediction, machine learning algorithms

# Literature Review

Coffee has been the most important agricultural product in Colombia, providing livelihood for around two million Colombians (Sanz Uribe et al., 2021, 35). Linear regression has many applications in life, one of them, and the one that I am going to be applying is predicting the price of coffee based on different inputs or variables, this has been already tried by a variety of people using different linear models, Neural Networks, and Support Vector Machines, like in a case study made in India, to forecast Indian green coffee, a study made in Brazil to predict as well the value of the grain using Neural Networks, (Deina et al., 2021, 2)

But why would we want to predict the coffee price? What would be the advantage of doing this? Well, just imagine if coffee growers had been prepared by 2020 when the price of coffee was the highest on the New York stock exchange at US$1.10 per pound instead of selling the coffee in the harvest of 2019 when prices were below US$ 1 per pound, (Diario el Mundo, 2021), this fluctuation in prices could have been a great opportunity to growers, and perhaps if is seen from the client or purchaser side, it could be argued that they would want to buy at the lowest prices, creating perhaps a balance in the economy as both ends won’t be able to take advantage from each other.

One important factor to be analyzed would be the market size and the demographic impact on different societies around the world, with emerging markets like China, India, and Brazil that constitute 36% of the global GDP (Li et al., 2022, 1) is in these countries where tea consumption is superior to coffee, with the exception of Brazil, and is in these countries where customers would pay a premium price for the symbolic western experience at stores like Starbucks (Li et al., 2022, 2) just for a cup of coffee.

Just imagine for a moment, if the regression analysis to predict the prices is combined as well with an algorithm for coffee selection based on its quality using different techniques like k-Neighbours, Decision Tree, SVM Support Vector Machine, Logistic regression, or Neural Networks, and having these results or classification of superior quality as independent variables in the price regression, and not depending solely on other factors that are not quality, luckily there are new studies that are already doing the first step, like Suarez Pena, on his thesis, to classify based on quality, the results having a prediction of 83% in classification, and Mean Absolute Error (MAE) of 14,61% (Suarez Pena, 2019,100), this would be a big tool not only for coffee tasters but growers and purchasers as well as people who really enjoy drinking quality coffee.

‘Professional investors favour two dominant schools of thought on investing which are Fundamental Analysis and Technical Analysis, and Machine learning techniques are one of the Technical Analysis, (Siew et al., 2012,2), and some of these algorithms are the ones that I will be using for the project, using examples and getting help from some of the papers published in the scientific community (Sreehari et al., 2018,1) where the author uses Multiple Linear Regression to predict climate, another very interesting article where they use regression tree to predict the Indonesian stock price during the Covid-19 era (Hindrayani, et al.,2020,1), as well as a comparative analysis on linear regression and Support Vector regression (SVR)(Kavitha S, et al.,2016,1) where they analyze a time series data in order to have better prediction and accuracy on their data. Which is the most accurate algorithm to predict the future price of the OIC\_prices?

The need to choose a different algorithm varies according to the needs of each investigation (Gil Serna, 2012, 82), however the final decision, in this case, will be made once the results of the analysis have concluded.

The project is very important for the student as it opens new fields of thinking in the machine learning and Data Science world, and encourages him to do his own research about topics that the student has never seen or taken and is because of research that new things are discovered.

# Overall Methodology

## Diagram of an overall methodology



After acquiring the data and preparing it, will be used to extract a summary of Statistics and do an Exploratory Data Analysis in order to answer possible questions that might arise, then the data will be separated into Training Set and Testing Set, and Algorithms like Linear Regression, Decision tree Regression, Lasso Regressor, SVM Regressor, and Neural Network will be used to answer the questions that were formulated in the abstract and to see how our regressors are behaving

When data needs to be normalized, it will be normalized via MinMax Normalizer and Standard Normalizer before applying the algorithms previously mentioned above, however, some of the algorithms will be applied without normalization.

# Data Acquisition

The source of the dataset, can be downloaded from the website of the Federación Nacional de Cafeteros de Colombia (National Federation of Coffee Growers of Colombia) here:

<https://federaciondecafeteros.org/app/uploads/2020/01/Precios-%C3%A1rea-y-producci%C3%B3n-de-caf%C3%A9.xlsx>

Even if it contains a variety of sheets, please use the sheet called: ***6. Precio OIC Mensual***

It will look like this:

Table, calendar

Description automatically generated

This data contains 4 sub-divisions of Coffee types:

**Name in Spanish Name in English**

Suaves Colombianos (Arábigo) Colombian Mild (Arabicas)

Otros Suaves (Arábigo) Other Mild (Arabicas)

Naturales de Brazil (Arabigo) Brazilian (Arabicas)

Robustas Robustas

There are 14 Columns, the names have been translated or changed as per the following table.

| **Subdivision Coffee Type** | **Column Original name** | **Column New name** |
| --- | --- | --- |
|  | Mes | Date |
|  | Precio del indicador compuesto OIC | OIC\_price |
| **Suaves Colombianos (Arábigos)** | Nueva York | Colombia\_ny |
| Europa | Colombia\_europe |
| Promedio ponderado | Colombia\_average |
| **Otros suaves (Arábigo)** | Nueva York | Other\_ny |
| Europa | Other\_europe |
| Promedio ponderado | Other\_average |
| **Naturales de Brazil (Arabigo)** | Nueva York | Brazil\_ny |
| Europa | Brazil\_europe |
| Promedio ponderado | Brazil\_average |
| **Robustas** | Nueva York | Robustas\_ny |
| Europa | Robustas\_europe |
| Promedio ponderado | Robustas\_average |

## Columns Metadata

**Date:** Column expressing the date monthly beginning January of 2000, all prices will have a reference for this date.

**OIC\_price:** Is the average price of the International Coffee Organization for the month and year shown, measured in US cents/lb

**Colombia\_ny:** Are the average price of Colombian Mild Arabicas for the month and year shown and is expressed in US cents/lb, in the US market

**Colombia\_europe:** Is the average price of Colombian Mild Arabicas for the month and year shown and is expressed in US cents/lb, in Germany and France

**Colombia\_average:** Is the weighted average Colombian Mild Arabicas Composite Indicator Price, as per "Section 4 "of the Indicator Prices SC-106/21 for the month and year shown and is expressed in US cents/lb

**Other\_ny:** Is the average price of Other Mild Arabicas for the month and year shown and is expressed in US cents/lb, in the US market

**Other\_europe:** Is the average price of Other Mild Arabicas for the month and year shown and is expressed in US cents/lb, in Germany and France

**Other\_average:** Is the average price of Other Mild Arabicas Composite Indicator Price, as per "Section 4 "of the Indicator Prices SC-106/21 for the month and year shown and is expressed in US cents/lb

**Brazil\_ny:** Is the average price of Brazilian Naturals for the month and year shown and is expressed in US cents/lb, in the US market

**Brazil\_europe:** Is the average price of Brazilian Naturals for the month and year shown and is expressed in US cents/lb, in Germany and France

**Brazil\_average:** Is the average price of Brazilian Naturals Composite Indicator Price, as per "Section 4 "of the Indicator Prices SC-106/21 for the month and year shown and is expressed in US cents/lb

**Robustas\_ny:** Is the average price of Robustas for the month and year shown and is expressed in US cents/lb, in the US market

**Robustas\_europe:** Is the average price of Robustas for the month and year shown and is expressed in US cents/lb, in Germany and France

**Robustas\_average:** Is the average price of Robustas Composite Indicator Price, as per "Section 4 "of the Indicator Prices SC-106/21 for the month and year shown and is expressed in US cents/lb

# Multivariate Analysis

## Summary Statistics

There are 272 rows

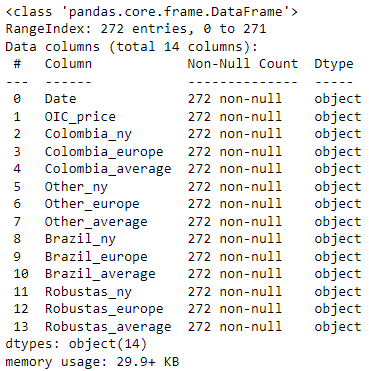
The data does not have NA’s or empty cells

Measures of central tendency and measures of dispersion table:

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OIC\_price** | 272 | 113.633237 | 43.687941 | 41.17 | 88.5475 | 113.155682 | 133.130252 | 231.24 |
| **Colombia\_ny** | 272 | 153.333888 | 62.437782 | 58.92 | 112.94 | 144.413636 | 182.577237 | 319.63375 |
| **Colombia\_europe** | 272 | 148.409523 | 59.732402 | 57.72 | 111.6975 | 141.123636 | 178.760455 | 311.45 |
| **Colombia\_average** | 272 | 151.222932 | 61.156478 | 58.1 | 112.67 | 143.529552 | 179.0225 | 312.95 |
| **Other\_ny** | 272 | 145.099616 | 58.41441 | 51.95 | 108.72 | 141.896818 | 169.232857 | 303.59 |
| **Other\_europe** | 272 | 143.098234 | 55.916915 | 55.76 | 110.295 | 138.104348 | 165.942045 | 297.22 |
| **Other\_average** | 272 | 143.96633 | 56.927377 | 54.28 | 109.7125 | 140.704773 | 166.686126 | 300.12 |
| **Brazil\_ny** | 272 | 118.365112 | 49.732689 | 37.67 | 94.405 | 111.98 | 132.726023 | 271.39 |
| **Brazil\_europe** | 272 | 123.408718 | 51.170812 | 38.71 | 96.056883 | 117.983409 | 143.199599 | 273.43 |
| **Brazil\_average** | 272 | 121.959193 | 50.984818 | 38.63 | 95.605714 | 116.833333 | 140.72888 | 273.4 |
| **Robustas\_ny** | 272 | 79.670588 | 28.448193 | 21.25 | 57.895 | 84.680554 | 103.530147 | 126.3 |
| **Robustas\_europe** | 272 | 74.456376 | 26.5673 | 22.79 | 54.765 | 78.150682 | 97.329432 | 121.3 |
| **Robustas\_average** | 272 | 75.376692 | 26.767165 | 22.81 | 55.3475 | 79.203636 | 98.367841 | 121.98 |

Note that Colombian\_average maximum is almost 3 times the price of the Robustas\_average which is predominantly from Vietnam, the reason is that “Robustas has a high caffeine content (2% to 4%), so the flavor is not as pure as Arabica” (Roldan Perez et al., 2009,32), “quality of the Robusta produced is uneven because of processing technology, drying equipment and post-harvest technological problems. These cause the coffee beans to have a high humidity level, and not meet the required standard of color, quality, and so on. This is the reason that Vietnam’s coffee price is lower than the world price.” (Roldan Perez et al., 2009,32).

The 3rd Quantil (75%) of Robustas\_average is below the 1st Quantil (25%) of Colombian\_average and Other\_average, and almost the same for the 1st Quantil (25%) of Brazil\_average.



There are a total of 14 Columns, each with 272 rows, and the type of data per column is the type Object that is like a String type, which will have to be converted to float type and time series for the column Date

Checking for duplicates on each column [duplicated index]:

Colombia\_europe [20 - 40] In the years 2001-09 and 2003-05, it was exactly the same price coffee.

Brazil\_ny[75 - 89] In the years 2006-04 and 2007-06, it was exactly the same price coffee.

Other\_ny [158 - 159] In the years 2013-03 and 2013-04, it was exactly the same price as the coffee

Other\_europe [29 - 41] In the years 2002-06 and 2003-06, it was exactly the same price as the coffee

Robustas\_europe [118 - 119 - 130 - 183] On years 2010-11 and 2020-04, it was exactly the same price of the coffee

Robustas\_ny [35 - 39] On years 2002-12 and 2003-04, it was exactly the same price of the coffee

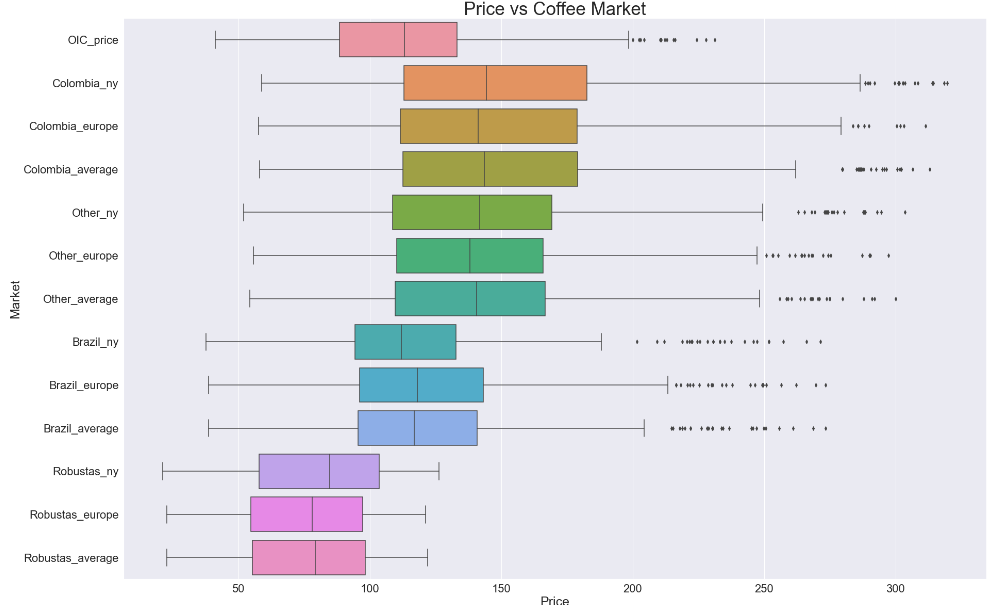
## Min, Max, and Outliers

There are no outliers in the Robustas columns, but the other columns do have outliers

| **Column** | **Max Outlier** | **Min Outlier** | **# of Outliers** |
| --- | --- | --- | --- |
| OIC\_price | 231.24 | 200 | 13 |
| Colombia\_ny | 319.63 | 288.43 | 16 |
| Colombia\_europe | 311.45 | 283.74 | 9 |
| Colombia\_average | 312.95 | 279.55 | 19 |
| Other\_ny | 303.59 | 262.94 | 19 |
| Other\_europe | 297.22 | 250.75 | 19 |
| Other\_average | 300.12 | 255.9 | 19 |
| Brazil\_ny | 271.39 | 201.6 | 25 |
| Brazil\_europe | 273.43 | 216.46 | 23 |
| Brazil\_average | 273.4 | 214.8 | 24 |

## Boxplot or candle sticks

Creating a boxplot for the columns, which can be seen the outliers graphically, we can appreciate that the variety of coffee Robustas, doesn’t have outliers, but the rest of the varieties have outliers



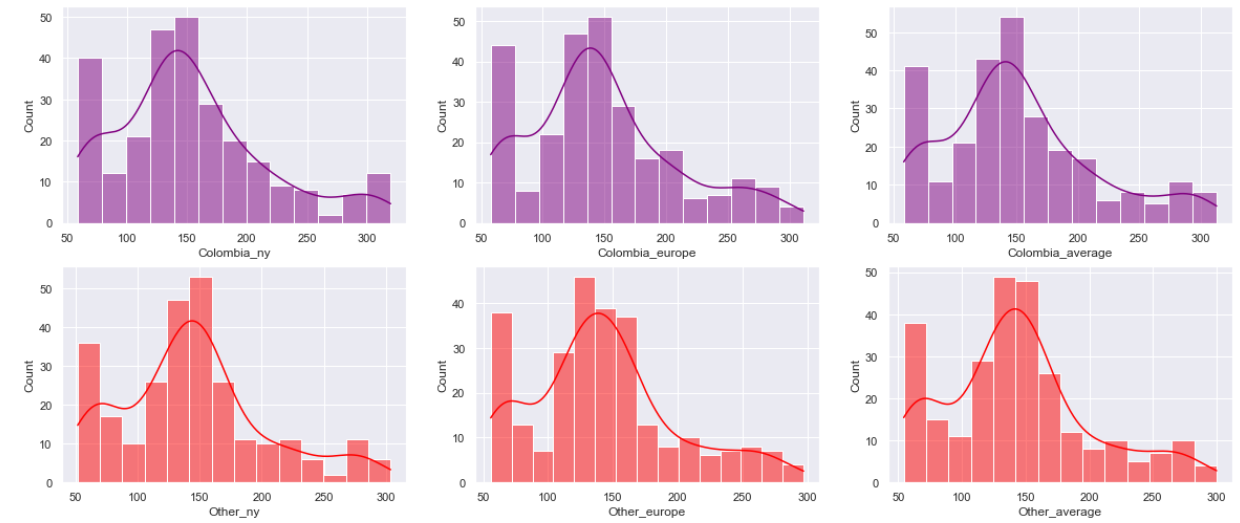
The median of OIC\_price is basically where the lower quartile Q1 of all “Colombian” columns start as well as the “Other” columns

The median of “Robustas” columns is smaller than the lower quartile Q1 of OIC\_price column

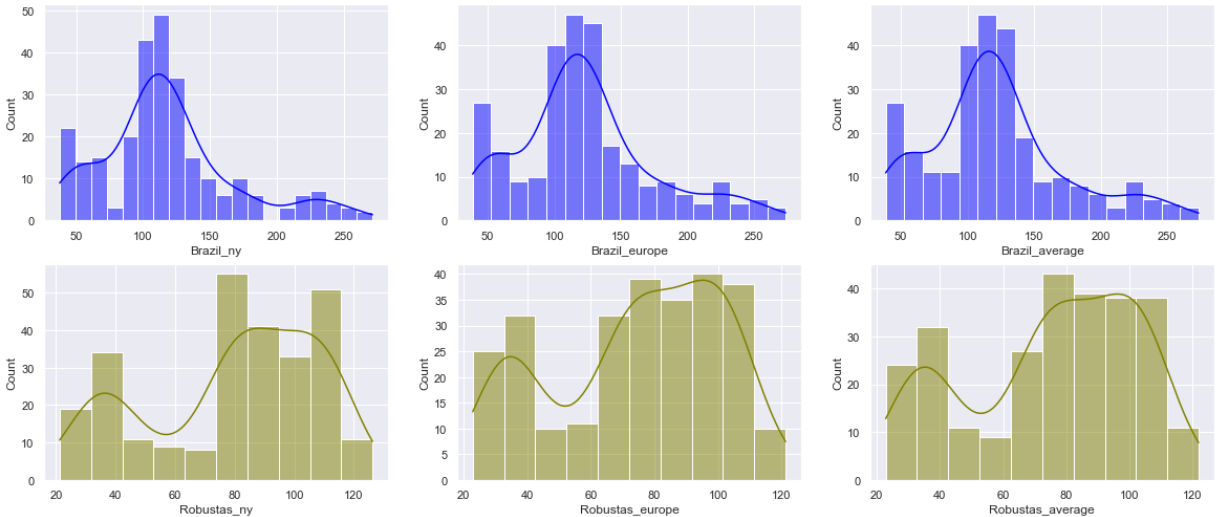
The upper quartile or Q3 of “Robustas” is smaller than the lower quartile Q1 of all the columns with the exception of OIC\_price and “Robustas\_ny”

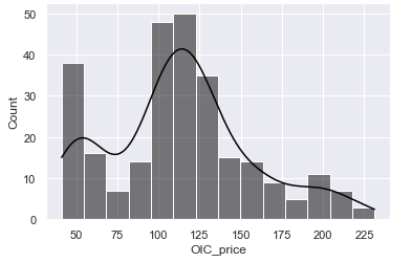
Brazil\_ny has a small IQR, while all “Colombia” columns have a big IQR

## Frequency Distribution

Creating a histogram with each average variable: 

All columns follow almost the same distribution pattern, however, Robustas\_average is quite different from the others, and the data is not that “normally distributed”. Please note that Robustas type coffee follows a different distribution that looks like a bimodal distribution, it has two local maxima points that are notable in the graph.





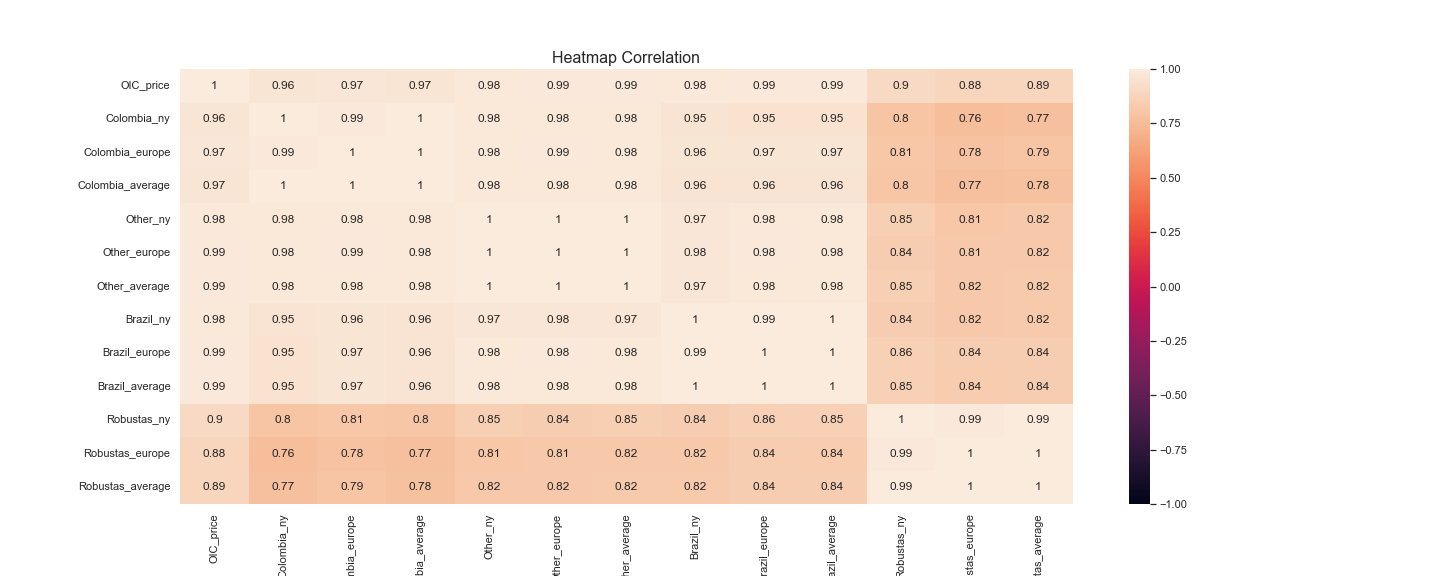
The variable OIC\_price, here in color black, follows almost the same distribution of all the other columns, with a slight peak at around US $55 dollars.

## 

## Correlation and Heatmap

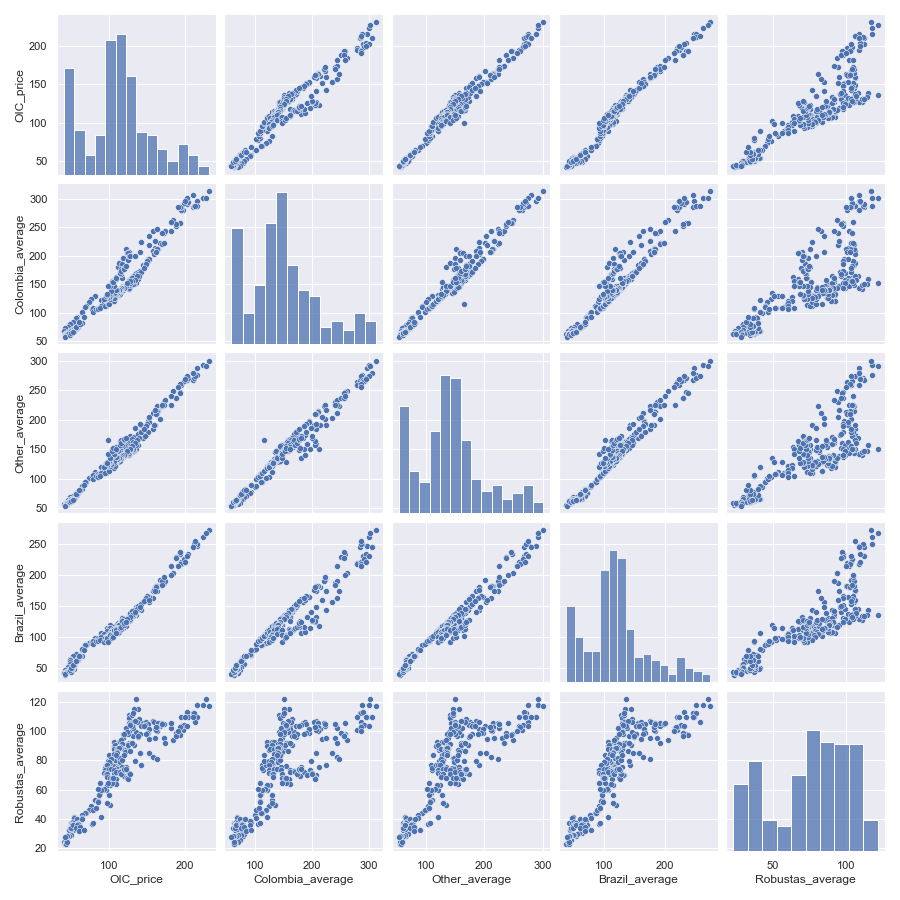
Creating a correlation matrix between all the columns, can be seen that there is a high correlation between all the averages, not that much with Robustas\_average, but there is still a correlation, the lowest correlation is between Robustas\_average and Colombia\_ny with a value of 0.76, and with our target column which is OIC\_price, the lowest correlation is Robustas\_europe with 0.88

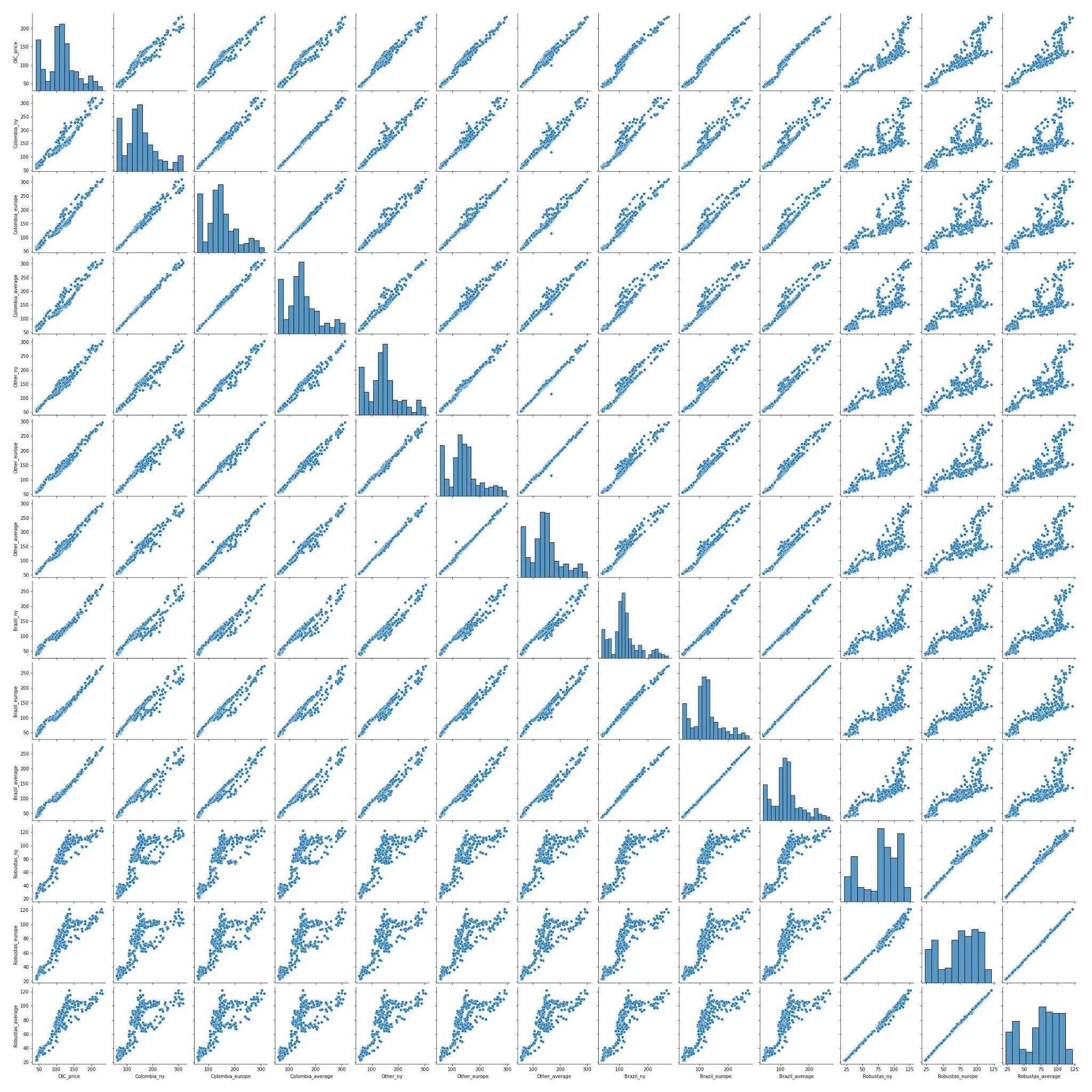
Overall, all columns have a high correlation between all of them, and it can be seen as well that the average columns have a correlation of 1 with the same group of coffee that they are averaging.

****

**Scatterplot Matrix**

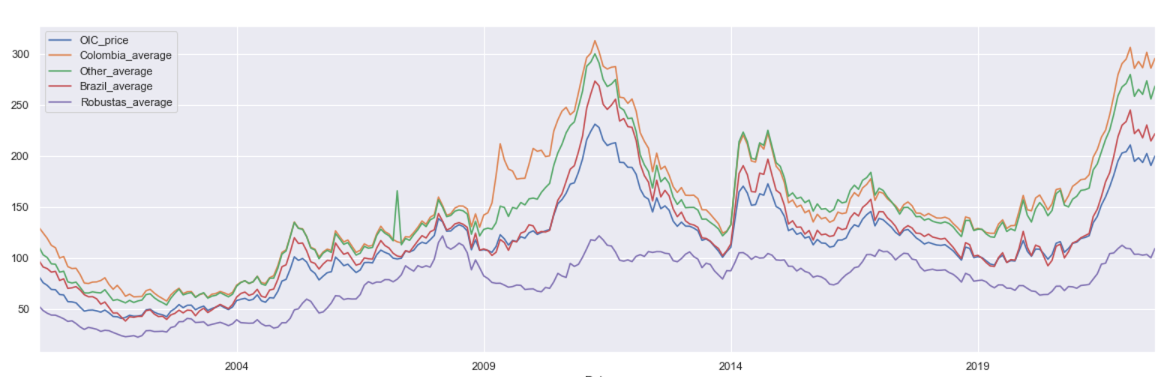
The scatterplot between all average columns and OIC\_price confirms the correlations seen above on the matrix.

There is a linear relationship between the variables and the target, which is confirmed by the scatterplot.

There is a linear relationship between almost all columns, but this can be seen better in a heat map.

# Exploratory Data Analysis

**Coffee averages price distribution**

****

The date column is on a monthly basis for a period of 10 years and behaves as a queue, where it has a front and rear, end every time a new month is added to the end of the queue, the month on the front is withdrawn from the queue.

On the Graph distribution between the averages over time can be seen that there some coffees reached up to $300 dollars twice during the decade and had lots of peaks, while the robustas had a more flat-like behavior.

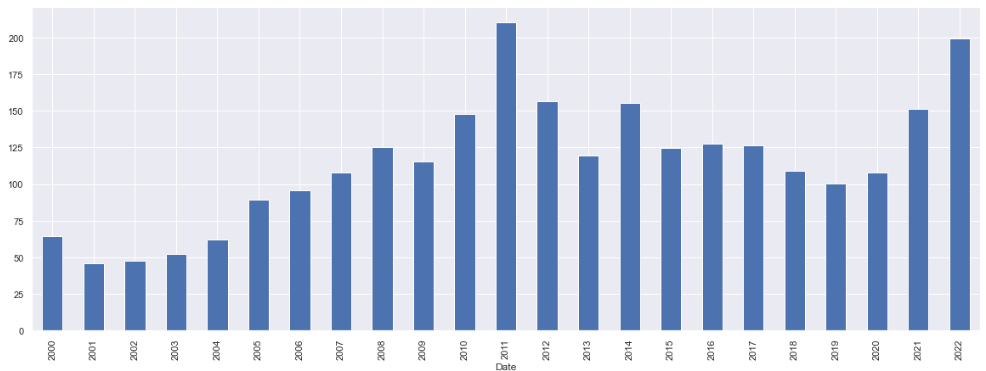
****

Note, that “the daily price increased by over 50% between 30 January and 10 March, as the ongoing drought in Brazil and uncertainty over the 2014/15 crop put upward pressure on prices” (ICO.ORG, 2014), this is particularly important due that January 2014 was one of the hottest months in “Brazil, which produces nearly 40% of the world’s coffee” (Wile, 2014), and lots of the crops were lost on this year

Average price per year

| **Date** | **OIC\_ price** | **Colombia\_ ny** | **Colombia\_ europe** | **Colombia\_ average** | **Other\_ ny** | **Other\_ europe** | **Other\_ average** | **Brazil\_ ny** | **Brazil\_ europe** | **Brazil\_ average** | **Robustas\_ ny** | **Robustas\_ europe** | **Robustas\_ average** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2000** | 64.25 | 102.60 | 99.80 | 102.60 | 85.09 | 92.89 | 87.08 | 79.86 | 83.67 | 79.86 | 42.12 | 40.36 | 41.41 |
| **2001** | 45.59 | 72.21 | 68.24 | 72.05 | 61.94 | 63.14 | 62.28 | 50.52 | 52.42 | 50.70 | 27.30 | 27.49 | 27.54 |
| **2002** | 47.76 | 65.27 | 64.78 | 64.90 | 60.44 | 62.35 | 61.55 | 45.10 | 45.92 | 45.23 | 30.84 | 29.78 | 30.03 |
| **2003** | 51.90 | 67.31 | 64.34 | 65.33 | 64.09 | 64.30 | 64.20 | 50.82 | 50.16 | 50.31 | 38.39 | 36.50 | 36.94 |
| **2004** | 62.15 | 83.85 | 79.49 | 81.44 | 80.15 | 80.64 | 80.47 | 68.18 | 69.11 | 68.97 | 37.28 | 35.66 | 35.98 |
| **2005** | 89.34 | 117.00 | 114.67 | 115.73 | 114.29 | 114.83 | 114.83 | 101.33 | 102.49 | 102.29 | 53.38 | 49.86 | 50.51 |
| **2006** | 95.75 | 117.92 | 115.70 | 116.80 | 113.95 | 114.80 | 114.40 | 102.88 | 104.19 | 103.92 | 70.28 | 66.98 | 67.56 |
| **2007** | 107.68 | 126.74 | 124.70 | 125.57 | 123.16 | 123.81 | 127.83 | 110.69 | 112.06 | 111.79 | 88.26 | 86.30 | 86.37 |
| **2008** | 125.47 | 146.08 | 144.27 | 145.37 | 139.63 | 141.99 | 141.21 | 124.47 | 129.48 | 128.35 | 107.66 | 106.32 | 106.56 |
| **2009** | 115.67 | 180.87 | 174.58 | 177.43 | 141.65 | 145.48 | 143.84 | 111.39 | 116.55 | 115.33 | 77.16 | 74.02 | 74.58 |
| **2010** | 148.16 | 224.62 | 227.08 | 226.33 | 195.44 | 197.62 | 196.97 | 146.68 | 156.92 | 154.66 | 85.07 | 78.46 | 79.55 |
| **2011** | 210.39 | 283.82 | 283.67 | 283.84 | 273.20 | 269.55 | 271.07 | 243.67 | 248.72 | 247.62 | 115.99 | 107.91 | 109.21 |
| **2012** | 156.36 | 203.95 | 200.53 | 202.15 | 187.59 | 185.76 | 186.53 | 171.37 | 176.13 | 175.03 | 110.58 | 101.30 | 102.76 |
| **2013** | 119.51 | 148.25 | 147.53 | 147.87 | 141.08 | 138.42 | 139.53 | 117.95 | 123.56 | 122.23 | 100.50 | 92.95 | 94.16 |
| **2014** | 155.26 | 198.09 | 198.16 | 197.95 | 202.85 | 199.08 | 200.39 | 161.30 | 175.29 | 171.59 | 105.60 | 99.47 | 100.43 |
| **2015** | 124.67 | 149.88 | 154.02 | 151.80 | 160.53 | 159.54 | 159.94 | 123.11 | 135.72 | 132.45 | 94.20 | 86.84 | 88.05 |
| **2016** | 127.38 | 155.58 | 155.37 | 155.38 | 164.63 | 163.49 | 163.88 | 124.18 | 142.72 | 137.86 | 94.28 | 87.47 | 88.63 |
| **2017** | 126.68 | 154.07 | 150.41 | 152.37 | 152.41 | 149.50 | 150.73 | 126.55 | 133.78 | 131.91 | 104.09 | 100.28 | 100.95 |
| **2018** | 109.04 | 139.59 | 133.26 | 136.70 | 137.40 | 129.10 | 132.73 | 109.62 | 115.10 | 113.65 | 88.34 | 84.03 | 84.80 |
| **2019** | 100.52 | 137.07 | 129.19 | 133.60 | 137.46 | 125.52 | 130.66 | 100.07 | 101.99 | 101.53 | 80.06 | 72.15 | 73.56 |
| **2020** | 107.94 | 166.28 | 146.10 | 157.67 | 156.81 | 146.77 | 150.72 | 102.46 | 107.86 | 106.42 | 78.20 | 66.68 | 68.75 |
| **2021** | 151.29 | 228.42 | 206.53 | 219.05 | 209.88 | 201.42 | 204.62 | 158.23 | 163.01 | 161.70 | 99.81 | 87.74 | 89.87 |
| **2022** | 199.43 | 309.18 | 272.31 | 293.69 | 277.86 | 260.35 | 266.66 | 228.77 | 225.65 | 226.49 | 114.73 | 103.73 | 105.49 |

Plotting the average price per year as a distribution graph, shows the Average Maximum Price of OIC has been US $210.38 in the year 2011, while the Average Minimum Price of IC has been US $ 45.59 in the year 2001.

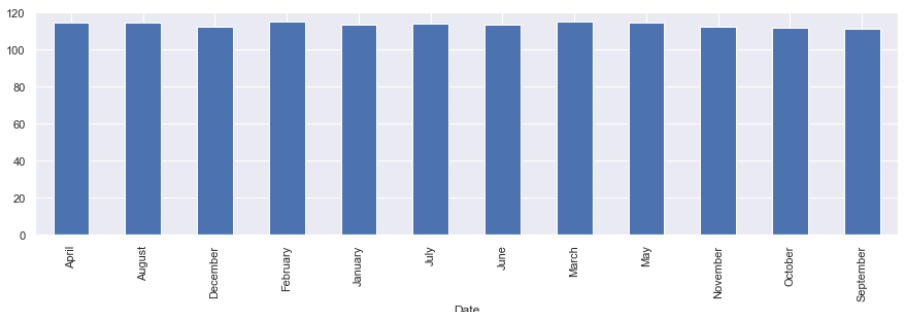


The months of February and March, are when the coffee has the highest prices, however, there is not a significant increase in price, only 3%

Minimum Month price: 111.48

Maximum Month price: 115.20

The graph shows the average price per month for OIC\_price

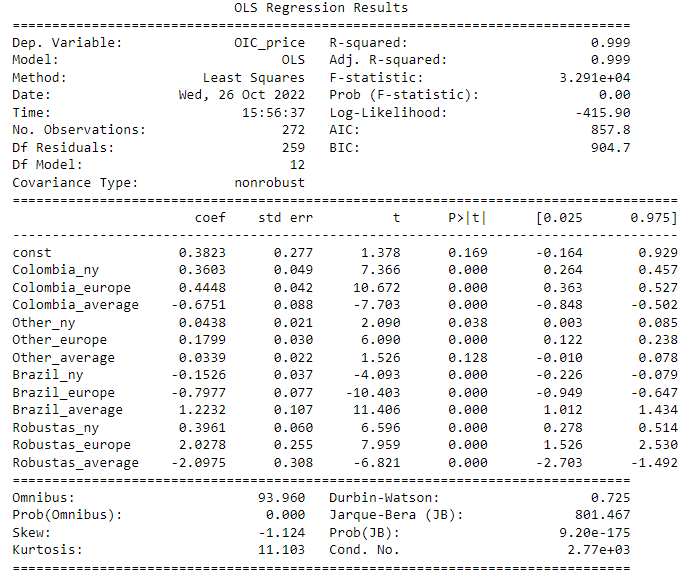


# Features Selection and Backward Pairwise Relations

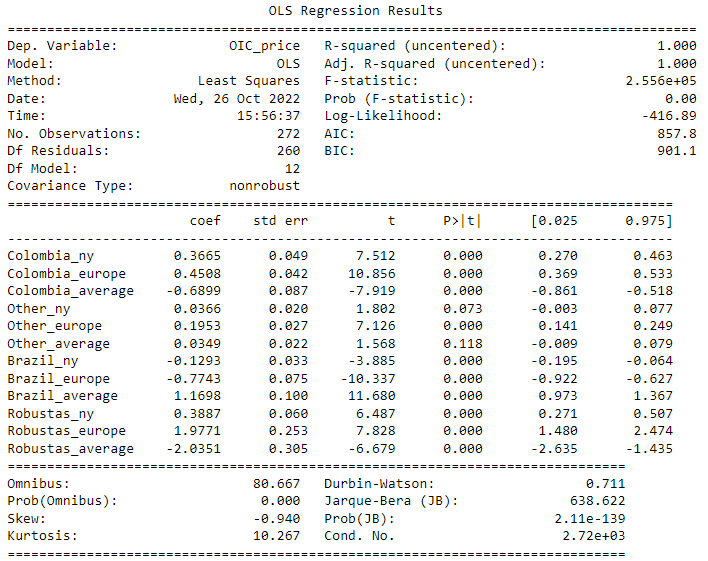
As there are 14 columns in our data set, in order to build a proper model to do regression, there will be a limit of SL = 0.05 and obtain the p-values for each of the columns, please note that the independent column is OIC\_price, and there won’t be an analysis for this column, and the same will be done with the column Date.

In order to obtain an equation that satisfies the model of multiple linear regression 

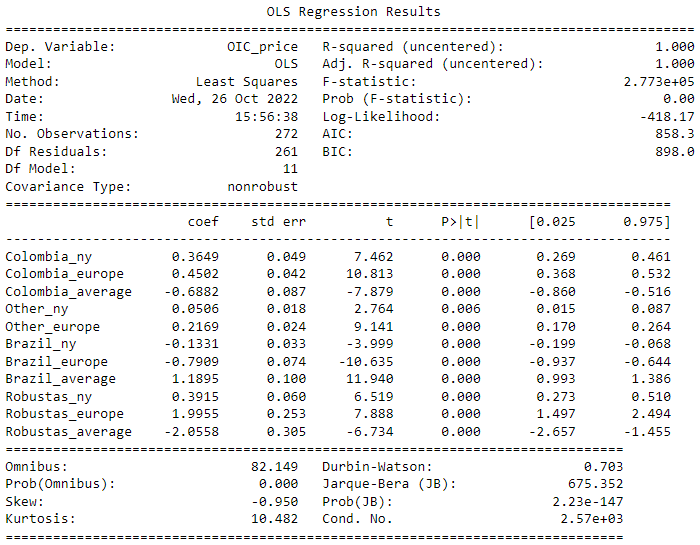
I’ll obtain the p-value, using a backward stepwise regression, and if the p-values for the columns are > SL, then it will be discarded.



There are two columns with a high p-value, but only one will be removed in this step, in this case, the column with the constant value, which has a p-value of 0.169



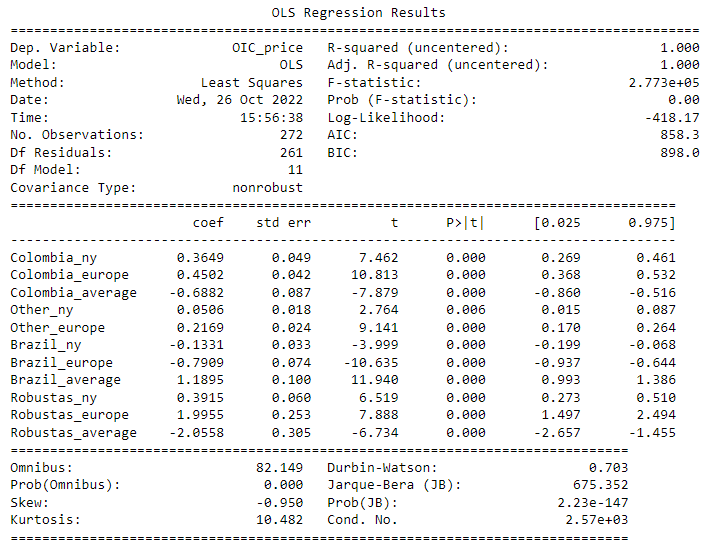
Then after removing the constant column, there are more columns again with different value, and once again, I’ll remove the one with the highest p-value, in this case, is column number 6 called Other\_average, with a p-value of 0.118



Now is the turn to remove the column Other\_ny as it has a p-value = 0.06 and is above my limit of 0.05

The model is completed, keeping all the columns with the exemption of Columns: Other\_average and Other\_ny.

The coefficients b in the equation can be seen in the column **coef**, and the p-values can be seen in the column **p>|t|** and they all tend to zero, without being zero themselves.



**The model:**

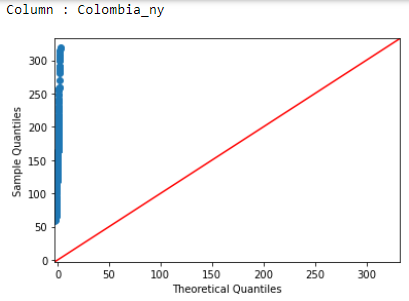
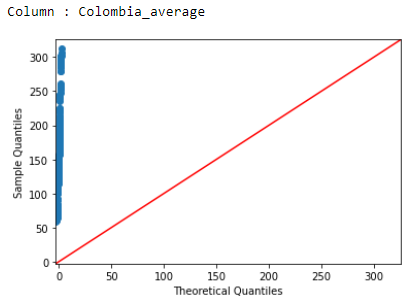
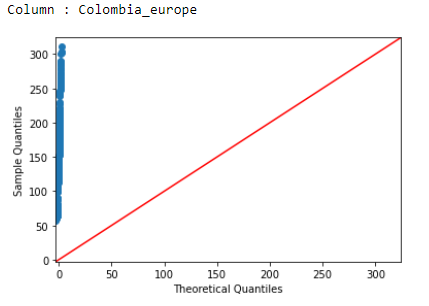
# 

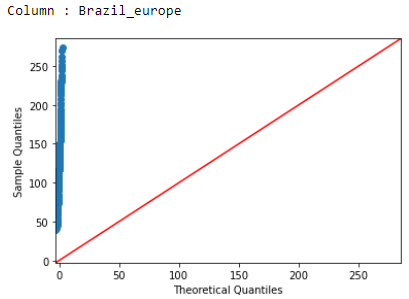
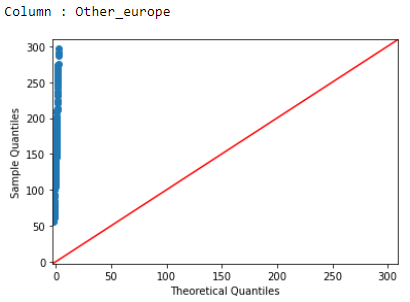
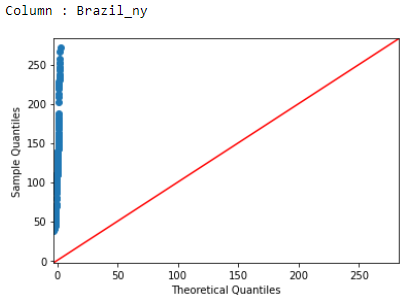
# Normalization

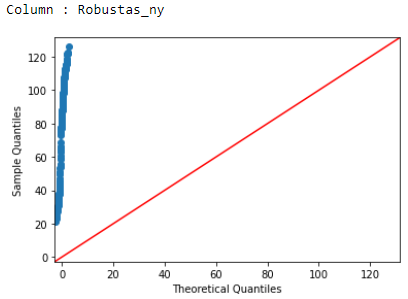
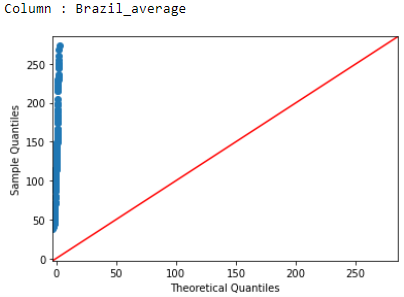
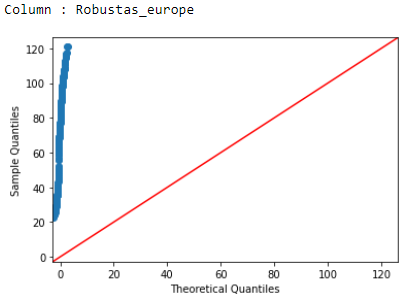
This work will be done without Normalization and with Normalization, and at the end compare all the errors and values in a data frame, this is done for academic purposes and as a learning tool.

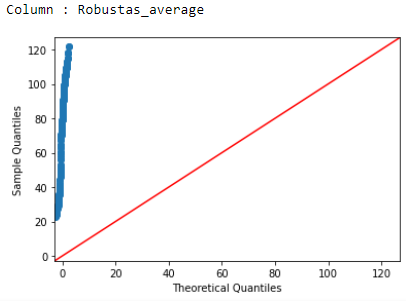
To normalize the data, I am going to create a data frame without the ['Date'] column, named **test**, then I’ll select the **X** which contains the columns of the model, and **y** which contains the column with the independent variable “Price\_OIC”.

To confirm that normality has to be done, I’ll use a Q-Q plot where can be seen that the scatterplot doesn’t follow the 45-degree line and needs to be normalized.







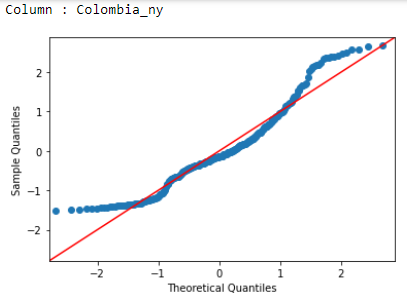
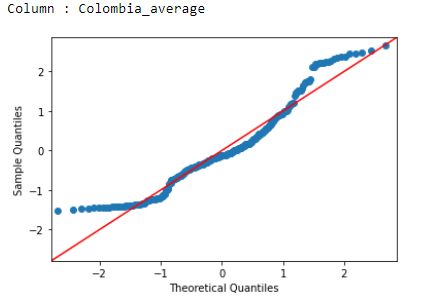
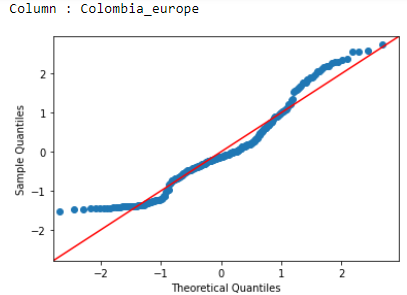


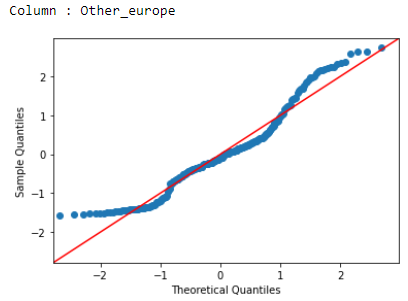
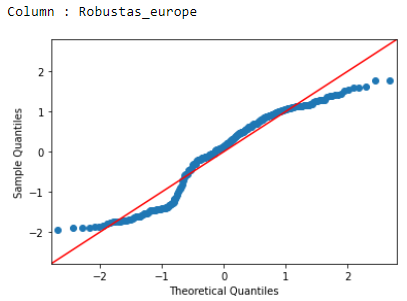
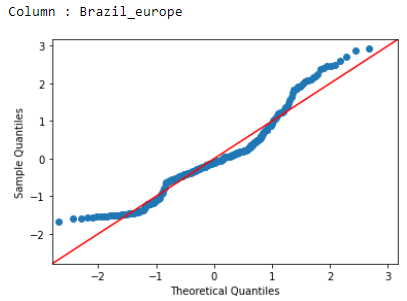
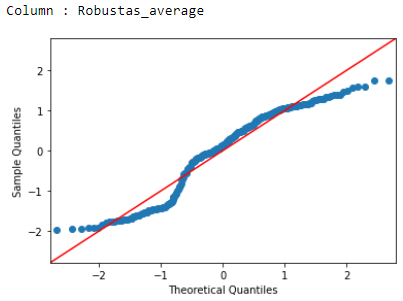
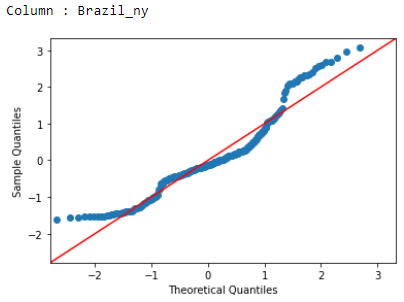
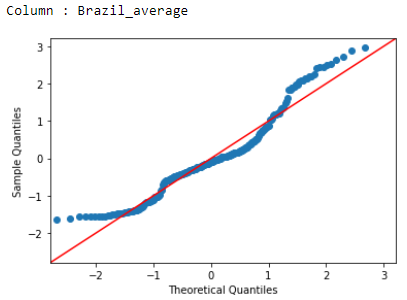
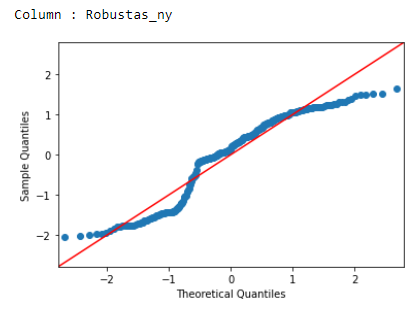
The function **StandardScaler()** will take care of the normalization:

**X** **=** [ 'Colombia\_ny', 'Colombia\_europe', 'Colombia\_average', 'Other\_europe', 'Brazil\_ny',

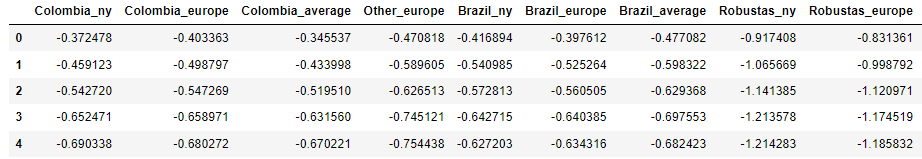
'Brazil\_europe', 'Brazil\_average', 'Robustas\_ny', 'Robustas\_europe', 'Robustas\_average' ]

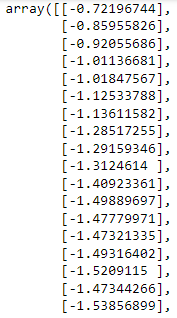
**y =** [ 'Price\_OIC' ]

After normalizing, the same function to draw a Q-Q plot is applied to the columns and can be seen that now it follows normality.



Screenshot of **X** normalized, not including all the columns as it won’t fit in the screenshot



Screenshot of **y** Normalized, not including all the values as it won’t fit in the screenshot

# References

# 

Deina, C., do Amaral Prates, M. H., Rodrigues Alves, C. H., Ribeiro Martins, M. S., Trojan, F., Stevan Jr, S. L., & Valares Siqueira, H. (2022, September 27). A methodology for coffee price forecasting based on extreme learning machines. *Information Processing in Agriculture*, *9*(3), 10. <https://www.sciencedirect.com/science/article/pii/S2214317321000597>

Diario el Mundo. (2021, February 15). ¿Cuál es el panorama mundial de la producción y del precio del café en 2021? *Diario El Mundo*. <https://diario.elmundo.sv/Econom%C3%ADa/cual-es-el-panorama-mundial-de-la-produccion-y-del-precio-del-cafe-en-2021>

E. Sreehari and S. Srivastava, "Prediction of Climate Variable using Multiple Linear Regression," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777452.

Federación Nacional de Cafeteros de Colombia. (2021). La cosecha asistida de café y su impacto en la economía de la recolección en finca. Ensayos sobre Economía Cafetera, 34(1), 35-50

Federación Nacional de Cafeteros de Colombia. (2022, septiembre 16). Precios, área y producción del café. Bogota; <https://federaciondecafeteros.org/app/uploads/2020/01/Precios-%C3%A1rea-y-producci%C3%B3n-de-caf%C3%A9-3.xlsx>.

Gil Serna, J. G. (2012). *ESTIMACIÓN DE UN PRONÓSTICO DE EXPORTACIONES DE CAFÉ SUAVE COLOMBIANO: REDES NEURONALES ARTIFICIALES Y ARDL*.

H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), 2012, pp. 1-5, doi: 10.1109/ICSSBE.2012.6396535.

ICO.ORG. (2014, March 11). MONTHLY COFFEE MARKET REPORT - February 2014. *International Coffee Organization*. <http://www.ico.org/documents/cy2013-14/cmr-0214-e.pdf>

Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear regression and support vector regression," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.

K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 344-347, doi: 10.1109/ISRITI51436.2020.9315484.

Li, R., Laroche, M., Richard, M. O., & Cui, X. (2022, January 24). More than a mere cup of coffee: When perceived luxuriousness triggers Chinese customers’ perceptions of quality and self-congruity. *Journal of Retailing and Consumer Services*, *64*(102759). <https://www.sciencedirect.com/science/article/abs/pii/S0969698921003258>

Sanz Uribe, J. r., Duque Orrego, H., Gaitan Bustamante, A. Federación Nacional de Cafeteros de Colombia. (2021, June 25). La cosecha asistida de café y su impacto en la economía de la recolección en finca. *Ensayos sobre Economía Cafetera*, 35-50.

Statistics Committee. (2021, March 29). sc-106e-rules-indicator-prices.pdf. London; International Coffee Organization.

Suarez Pena, J. A. (2019, December). *MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA CALIDAD DEL CAFÉ.* [Proyecto de Grado para obtener el título de Magister en Ingeniería Industrial]. Repositorio Institucional Universidad Distrital - RIUD. <https://repository.udistrital.edu.co/bitstream/handle/11349/23560/SuarezPe%C3%B1aJavierAndres2019.pdf?sequence=1&isAllowed=y>

Wile, R. (2014, February 24). *Why Coffee Prices Are Surging*. Business Insider. Retrieved October 23, 2022, from <https://www.businessinsider.com/why-coffee-prices-are-surging-2014-2?r=US&IR=T>