Contents

[INTRODUCTION 2](#_Toc165079380)

[1.1 Data and Machine learning? 2](#_Toc165079381)

[1.2 About Dataset 3](#_Toc165079382)

[1.3 Bitcoin stock price between 2010-2014 3](#_Toc165079383)

[1.4 Problem understanding 3](#_Toc165079384)

[1.5 Objectives 3](#_Toc165079385)

[2.1 Data exploration and Preparation 3](#_Toc165079386)

[2.3 Statistical exploration 4](#_Toc165079387)

[Box-Plots: 4](#_Toc165079388)

[3.1 K-Means and DBSCAN: 7](#_Toc165079389)

[3.2 Implementation 8](#_Toc165079390)

[2.3 Hyperparameter tunning: 9](#_Toc165079391)

[2.4 PCA and scaling: 10](#_Toc165079392)

[2.5 Applying optimal cluster: 10](#_Toc165079393)

[3.1 DBSCAN 11](#_Toc165079394)

[3.2 Implementation: 11](#_Toc165079395)

[3.3 Best cluster with DBSCAN 13](#_Toc165079396)

[3.3 Evaluation: 13](#_Toc165079397)

[4.1 Bitcoin time-series 14](#_Toc165079398)

[EDA and Data preparation 14](#_Toc165079399)

[4.3 Differencing 14](#_Toc165079400)

[References: 18](#_Toc165079401)

# INTRODUCTION

## 1.1 Data and Machine learning?

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on building computer systems that learn from data (Linda, 2021). While Ml has 2 broad aspect this project will focuses on the clustering and Time series analysis of artificial intelligence.  Clustering is the branch of unsupervised model that operates by dividing the objects into clusters that are similar between them and are dissimilar to the objects belonging to another cluster (Banoula, 2023).  Time series analysis generally involve analysing observation obtained over a period usually regular interval (Binhuraib, 2021).

Using CRISP-DM Methodology (Matsumoto and Carrinho, 2023), I will be reporting my methodology and result in this report. I have included in my Jupiter notebook a more detail structure of project implementation.

Diagram of a diagram of data

Description automatically generated

Figure 1: CRISP-DM Life Cycle

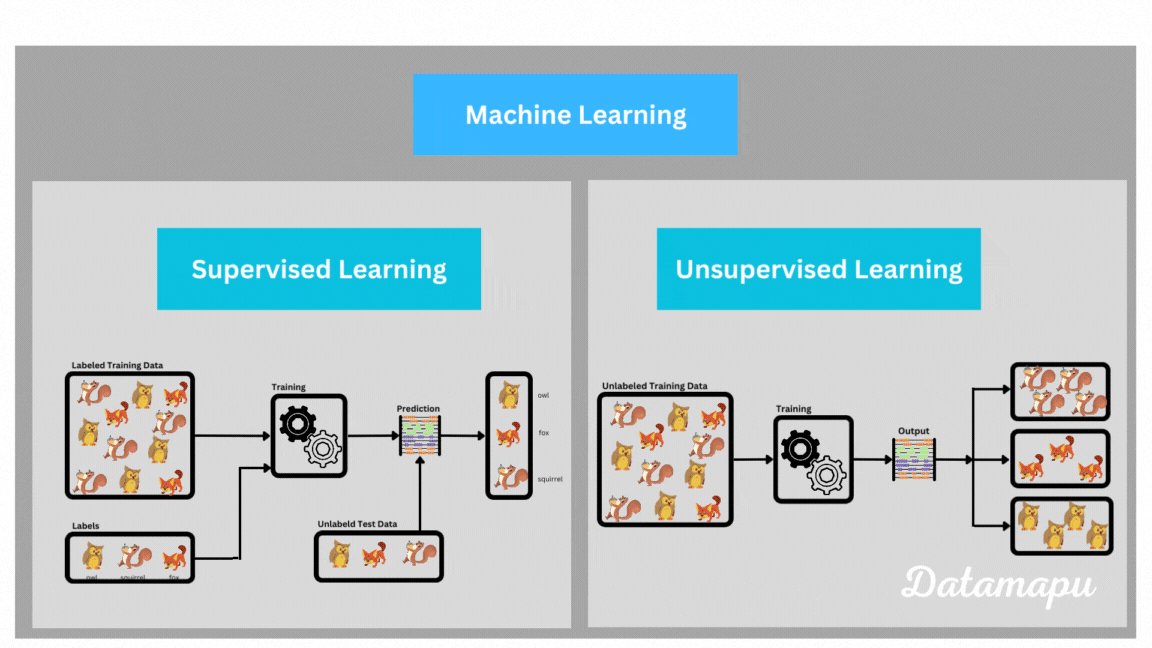


Figure 2: Supervised and Unsupervised Learning are different types of Machine Learning (datamapu, 2023).

### 1.2 About Dataset

Forty Soybean Cultivars from Subsequent Harvests dataset (UCI (Repository), 2023)

Data was obtained from forty soybean cultivars planted in 2 seasons. Each cultivar group of have were divided into with four replications. Data contain 11 variables and 320 rows.

### 1.3 Bitcoin stock price between 2010-2014

Data was obtained from Yahoo finance. It contains the stock market prices of bitcoin between 2010 and 2024. The data is a monthly data, with result in 120 observations. 5 numerical variables and 1 object are present.

### 1.4 Problem understanding

My research seeks to explore the relationship between soybeans plant height and grain yield and cluster observation to clusters of distinct characteristics using clustering algorithm (de Oliveira et al 2024). Motivated by the lack of sufficient ML experiment in this area, project aim at uncovering statistical insights in dataset and paving way for more experimentation subsequently in soybeans characteristics evaluation.

### 1.5 Objectives

* Group the soydata into clusters of distinct characteristics using appropriate clustering models and given criteria, and compare results obtained from them.
* Use appropriate hyperparameters like Elbow and silhouette to get the best parameter for K-Means algorithm and interpret cluster results of the relationship between plant height and grain yield.
* Get a stock dataset from yahoo finance and apply ARIMA time series model and interpret results.

## 2.1 Data exploration and Preparation

A screenshot of a computer code

Description automatically generated

Figure 3: EDA liabraries.

 Csv file was read, no null-values and duplicated values present.

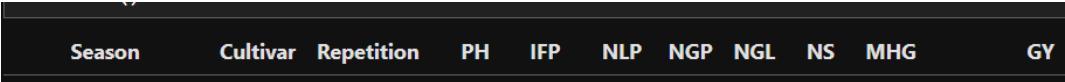
In other to accurately interpret the result that would be gotten from different phases of this analysis, it is imperative to know what variable mean (Niklas Donges, 2018). 

Figure 4: Screenshot of the variable names.

#### 2.2 Feature Engineering

Variable names were renamed using data dictionary  *fig.5*.

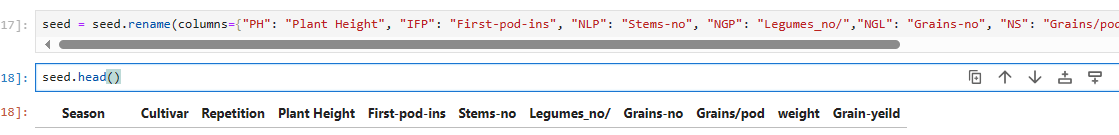


Figure 5: Renaming variables.

### 2.3 Statistical exploration

Statistics summary gives a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is gaussian distributed or left/right skewed (Mahadevan, 2022).

320 observations counted, comparing Mean and median indicate left skewness in ('Stems-no', 'Legumes\_no', 'Grain-yield') and right skewness in ('weight' and 'Grains-no'). Standard deviation suggests observations are close to the mean because all the numerical variables have a significantly small STD relative to the mean.

### Box-Plots:

I plotted the box plots specifically to check for outliers in the variables (Mcleod, 2019).

A diagram of a rectangular object with a blue label

Description automatically generated

Figure 6: Box plot diagram.

Repetition and season are not statistically significant. Significant numerical variables show outliers *fig.7.*

Histogram plot shows distribution under curve. Result shows Gy and Ph are gaussian distribution, with the rest slightly skewed.

A green and black graph

Description automatically generatedA diagram of a plant height distribution plot

Description automatically generated

A diagram of a plant distribution

Description automatically generatedA green and black graph

Description automatically generatedFig.7

**Relationship Plot**

A diagram of different colored dots

Description automatically generated

Figure 7: Relationship plot.

Scatterplot is a good relationship evaluation plot (Khan Academy, 2016). correlation coefficients show Ph and Gy have weak correlation. Ph of ~60-80cm and Gy of 3000-4000kg/ha are areas of high density in the plot compared to the other datapoints.

## 3.1 K-Means and DBSCAN:

Both K-means and DBSCAN are unsupervised ML algorithms used for partitioning datapoints into distinct clusters (Pandey, Lutins, 2020). DBSCAN is a density-based while K-means is medoids-based clustering algorithm (Pandey, Lutins, 2020), DBSCAN focuses on separating clusters of high density from clusters of low density (Lutins, 2020). K-means clusters data by trying to separate samples in n groups of equal variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares (scikit-Learn, 2010). I chose K-means because its simplicity generalization and scalability were important for the context of my dataset (D, 2020), and DBSCAN because of its unique ability to filter out outliers which is present in my data (Lutins, 2020).

A math equation with black text

Description automatically generated with medium confidence Figure 8: K-Means formula (scikit-Learn, 2010).

3.2 Implementation:

I plotted Ph vs Gy using seaborn and set the values of x as all the numerical variables using indexing.

A screenshot of a computer code

Description automatically generated

Figure 9: Model production and fitting.

Random K and iteration are 2 and 10 respectively, default ‘Max\_iter’, random state of 38 was set to ensure model reproducibility. Model ‘y\_kmeans’ was then fitted with the x values.

Fig.10 shows Ph and Gy grouped into 2 clusters. The yellow points are the centroids while group 1 and 2 are 2 distinct clusters K-Means have identified. We can see group 1 are all ph that their yields are >= 3500 up to 5000kg/ha, while group 2 are ph and grain yield that are lesser than 3500kg/ha

A diagram of a plant growth

Description automatically generated

Figure 10: 2 K result

 Group 1, 2,3 are 3 distinct clusters K-Means have identified between Ph and Gy. Group 1 are ph between (50-90) cm with yields (1500- 3000) kg/ha, while group 2 are ph between (10- >90) cm with yields (~3200- 3600) kg/ha, group 3 are ph between (55- 90) cm with yields between (~3600- >4500) kg/ha. Datapoint sizes were reduced using the ‘s’, the smaller size makes the within cluster density more insightful compared to fig.10, additionally, increasing K makes the noise present in the data more visible as these values are separated from most of the observation that are within 2300 and 4000kg/ha.

A diagram of a plant growth

Description automatically generated

Figure 11: 3 K result.

### 2.3 Hyperparameter tunning:

Elbow and silhouette methods were used to evaluate the best K value of which both returned 3 as the optimal K value. The silhouette score for 3 and 4 were approximately the same 0.49, and the same was noticed on the elbow curve as highlighted *fig.12*.

A graph showing the number of clusters

Description automatically generated

Figure 12: Elbow curve result.

### 2.4 PCA and scaling:

I implemented PCA and scaling of observations because raw data contain noise in it. Results after pca and scaling show that the within cluster distance was significantly reduced, the clusters were visually denser, but the high Davies and low silhouette score suggest model is not better performing. Pca also had limitation of interpretability, it was impossible to know what variable PH and GY is, therefore couldn’t tell if there is any improvement specific to my goal.A diagram of a plant growth

Description automatically generated

Figure 13: Ideal 3 K cluster of PH and Gy.

### 2.5 Applying optimal cluster:

Graph show the extreme values on the top and below i.e. above 4000kg/ha and below 3000kg/ha stayed the same as of the 3 clustering. However, the densest arear of the distribution 3000 - 4000kg/ha were now cluster into two with group 1 falling between 300-3500kg/ha and group 2 falling between 3500-4000kg/ha *fig.13*. I did 4 clusters because of domain knowledge that says observation were obtain from 4 repetitions, I wanted to check if it makes a difference on cluster results (UCI Repository, 2023).

### 3.1 DBSCAN

DBSCAN works bottom-up by picking a point and looking for more points within a given distance. It further expands the cluster by repeating this process for new points until the cluster cannot be further expanded (ML\_note, 2024). It does this by tunning epsilon (eps) which is how far to search for near points, and (minPts) which is how many points minimum should be in each cluster (ML\_note, 2024)

### 3.2 Implementation:

A screenshot of a computer

Description automatically generated

Figure 14: DBSCAN first trial using default values.

Initial model trail was fitted with default values of eps and minPts. This was to establish a ground truth for model performance.

The labels of the based model were printed, and 320 data point where ladled –1 which means noise. This was because the radius of eps is very small, and less data points are considered neighbours. I tried to improve this by increasing eps, no changes were made to result until eps was set at 10. At 10 eps and 6 minPts DBSCAN identified 10 clusters of which 250 where noise, which is still very significant.

A diagram of a grain cluster

Description automatically generated

Figure 15: 4 cluster.

At 40 eps and 6 minpts, eps identified 6 clusters, of which Gy between 3000-3800kg/ha where denser area and assigned 0 and 56 are now noise (-1) which is a significant improvement. I left the minpts at 6 because I wanted the model to have room to take the least noise data point.

### 3.3 Best cluster with DBSCAN

A diagram with many colored dots

Description automatically generated

Figure 16: 4 cluster result.

### 3.3 Evaluation:

Relative to K-means clustering, evaluated the performance of k (3&4) models using silhouette score(ss) and Davies-Bouldin index (DBI). In DBI the lower the score, the higher the density of that model meaning clusters separated well from each other, our result shows that K-mean model 3 is well separated than model 4, while in SS the higher the score, the better performance of the model, consistent with DBI model 3 had a higher score than model 4 which indicate the clusters in 3 are better separated from each other than model 4 (Sklearn, 2019). In DBSCAN, DBI and ss score show model trained with 'eps=40, min\_samples = 6' is well separated than model trained with 'eps=40, min\_samples = 10'.

### 4.1 Bitcoin time-series

### EDA and Data preparation

With ARIMA (Autoregressive Integrated Moving Average) model I aim to forecast bitcoin stock price closing data between January 2014 - April 2024. Data preparation involved checking data characteristics, no NAN and duplicate is present in data. Data column was set as the index as it helps to identify the period observation was collected when plotting.

A diagram of a mathematical equation

Description automatically generated with medium confidence

Figure 17: ARIMA model equation (Forecasting Home page, 2023).

**p** is the number of autoregressive terms,

* **d** is the number of nonseasonal differences needed for stationarity, and
* **q** is the number of lagged forecast errors in the prediction equation.

This is a univariate time series forecasting analysis as only previous closing values were used in forecasting (Forecasting Home page, 2023)

Initial visualisation of the data pattern, ACF and PACF, shows the data is not stationary, 'nlags' was set to 12 because the data is a monthly data, each lag represent the number of time unit between each observation in this case month (Wang, 2023). Prove of pattern presence existed. PACF shows the correlation between observations at specific lags while controlling for the influence of all other shorter lags (HEX, 2024). Quantile-quantile plot shows the distribution of bitcoin data significantly deviated from expected normal distribution as the data point a large not in line with the best fit line.

### 4.3 Differencing

Differencing of shift 1 was performed on the data to make it stationary, Augmented Dickey-Fuller Test was used to confirm stationarity before and after differencing. After the first trial of differencing sufficient evidence existed to reject null hypothesis of non-stationarity.

A graph of a bitcoin

Description automatically generated

Figure 18: Time series after differencing.

After differencing all lags are now within the zero mark which show stationarity. To choose the best pdq combination for forecasting, I used Akaike Information Criterion. AIC is a mathematical method used to evaluate how well a model fits the data it was generated from (Zach, 2021). Result shows ‘[0, 2, 2, 2220.105304163832]’ are the best combination to fit ARIMA model for forecast.

Performance evaluation:

A graph with red and blue lines

Description automatically generated

Figure 19: Time series forecasting after differencing.

 Normal Q-Q Plot compares the distribution of residuals to a normal distribution. In comparison to the first plot,  qqplot result shows a closer to normal distribution, because the theoretical quantile and the sample quantile are now closer. Histogram shows the distribution of the residuals. The distribution result largely follows a normal distribution as the median and mean of 0 are very close.  Correlogram (ACF and PACF of Residuals) shows residual ACF (Zach, 2021), while PACF accounts for the correlation explained by earlier lags (Zach, 2020). Result shows no significant spikes in the ACF and PACF plots which as earlier explained suggest stationarity.

A graph with a line and a red line

Description automatically generated

Figure 20: Interactive Time series forecasting.

Overall evaluation metrics shows that model perform well fitting data as fitness measures are largely met. In other to further investigated model performance I will visualise both the actual and predicted close price to visually evaluate performance.

4.4 Conclusion:

Arima model with 022 pdq performed well fitting and forecasting the stock price of bitcoin. Graph show the forecast was very close to the actual, interactive visualisation aid more comparison between actual closing values and forecasted.

Both K-means and DBSCA show 3 clusters as the ideal group than could be identified between GY and PH of soybeans data.

Word: 1300

## References:

A G. (2019). Silhouette Algorithm to determine the optimal value of k. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/silhouette-algorithm-to-determine-the-optimal-value-of-k/>.

Banoula, M. (2023). Supervised and Unsupervised Learning in (Machine Learning). [online] Simplilearn.com. Available at: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning#what_is_supervised_learning>.

Bartz-Beielstein, T. (2024). Hyperparameter Tuning Cookbook. [online] sequential-parameter-optimization.github.io. Available at: <https://sequential-parameter-optimization.github.io/Hyperparameter-Tuning-Cookbook/> [Accessed 26 Apr. 2024].

Binhuraib, T. (2021). An Introduction to Time Series Analysis with ARIMA. [online] Medium. Available at: <https://towardsdatascience.com/an-introduction-to-time-series-analysis-with-arima-a8b9c9a961fb?gi=64a317707b41> [Accessed 26 Apr. 2024].

D, V.A. (2020). K-Means Clustering Algorithm from Scratch. [online] Machine Learning Plus. Available at: <https://www.machinelearningplus.com/predictive-modeling/k-means-clustering/?utm_content=cmp-true> [Accessed 26 Apr. 2024].

datamapu (2023). Supervised versus Unsupervised Learning - Explained. [online] datamapu.com. Available at: <https://datamapu.com/posts/ml_concepts/supervised_unsupervised/> [Accessed 17 Apr. 2024].

Davies, D.L. and Bouldin, D.W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, [online] PAMI-1(2), pp.224–227. doi:<https://doi.org/10.1109/TPAMI.1979.4766909>.

de Oliveira, B.R. (2024). Vista do Dataset: Forty soybean cultivars from subsequent harvests. [online] Editorapantanal.com.br. Available at: <https://editorapantanal.com.br/journal/index.php/taes/article/view/8/5>.

Forcasting Home page (2023). *Introduction to ARIMA models*. [online] Duke.edu. Available at: <https://people.duke.edu/%7ernau/411arim.htm>.

Gupta, A. (2019). Elbow Method for optimal value of k in KMeans. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>.

Hector Cuesta (2013) Practical Data Analysis. Birmingham, UK: Packt Publishing. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=654543&site=eds-live&scope=site> (Accessed: 18 October 2023).

HEX (2024). *Unveiling Patterns Using Time Series Visualization | Hex*. [online] hex.techpatterns-using-time-series-visualization. Available at: <https://hex.tech/blog/patterns-using-time-series-visualization/>.

Hu, X., Chu, L., Pei, J., Bian, J. and Liu, W. (2021). Deep Learning Model Complexity: Concepts and Approaches. [online] Available at: <https://lingyangchu.github.io/External_Webpages/SDM2021_tutorial_DLcomplexity.pdf> [Accessed 26 Apr. 2024].

Khan Academy (2016). Scatterplots and correlation review. [online] Khan Academy. Available at: <https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/a/scatterplots-and-correlation-review>.

Kim, S. and SenGupta, A. (2018). Clustering Methods for Spherical Data: An Overview and a New Generalization. ICSA book series in statistics, pp.155–164. doi:<https://doi.org/10.1007/978-981-10-8168-2_11>.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics, 54(1-3), 159–178. [DOI]

Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python -. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/>.

Mahmood, M.S. (2021). *Understand Q-Q plot using simple python*. [online] Medium. Available at: <https://towardsdatascience.com/understand-q-q-plot-using-simple-python-4f83d5b89f8f>.

McKinney, W. (2017). Pythonic Data Cleaning With Pandas and NumPy. 2nd Edition ed. pp.17–18

Pandey, A.K. (2020). A Simple Explanation of K-Means Clustering and its Adavantages. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/>.

Petrović, S. (n.d.). A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. [online] Available at: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b>.

Prabhakaran, S. (2019). *Augmented Dickey Fuller Test (ADF Test)*. [online] Machine Learning Plus. Available at: <https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>.

Ray Tune (2022). Tune: Scalable Hyperparameter Tuning — Ray 1.11.0. [online] docs.ray.io. Available at: <https://docs.ray.io/en/latest/tune/index.html>.

Safwan, H. (2020). What is K Medoid Clustering: Why and How? [online] Analytics Vidhya. Available at: <https://medium.com/analytics-vidhya/what-is-k-medoid-clustering-why-and-how-b4e518c49e10> [Accessed 26 Apr. 2024].

Saurav, K. (2016). Clustering | Different Methods, and Applications (Updated 2024). [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#Different_Types_of_Clustering_Algorith> [Accessed 21 Apr. 2024].

scikit-Learn (2010). 2.3. Clustering — scikit-learn 0.20.3 documentation. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/clustering.html#k-means>.

Simplilearn (2023). Data Preprocessing in Machine Learning: A Beginner’s Guide. [online] Simplilearn.com. Available at: <https://www.simplilearn.com/data-preprocessing-in-machine-learning-article#:~:text=Data%20preprocessing%20is%20the%20process>.

SKlearn (2019). sklearn.metrics.silhouette\_score — scikit-learn 0.21.3 documentation. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>.

Topre, G. (2022). Bank Customer Churn Dataset. [online] www.kaggle.com. Available at: <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>.

Torres-Reyna, O. (2008). Data Preparation & Descriptive Statistics. [online] Available at: <https://www.princeton.edu/~otorres/DataPrep101.pdf> [Accessed 14 May 2022].

UCI Repository (2023). UCI Machine Learning Repository. [online] Uci.edu. Available at: <https://archive.ics.uci.edu/dataset/913/forty+soybean+cultivars+from+subsequent+harvests>.

UCI Repository (2023). *UCI Machine Learning Repository*. [online] Uci.edu. Available at:<https://archive.ics.uci.edu/dataset/913/forty+soybean+cultivars+from+subsequent+harves>.

Wang, Y. (2023). Oil Price Forecasting based on Improved SARIMA Model. *Highlights in Science, Engineering and Technology*, 39, pp.238–245. Doi :<https://doi.org/10.54097/hset.v39i.6535>.

Wei, H. (2020). How to measure clustering performances when there are no ground truth? [online] Medium. Available at: <https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>.

Zach (2020). *What Are Standardized Residuals?* [online] Statology. Available at: <https://www.statology.org/standardized-residuals/>.

Zach (2021). *How to Calculate AIC in R (Including Examples)*. [online] Statology. Available at: <https://www.statology.org/aic-in-r/>.