CA1 Machine Learning for Business

**Performing Clustering Algorithm for Education Student Performance and Investigate how Covid 19 pandemic affect the stock market of hotel in 2020 and 2021**

*By*

*Rachel Mae Lademora – 2023396*

Higher Diploma in Science in Data Analytics for Business

Machine Learning for Business

**Dr. Muhammad Iqbal**

**CCT College**

**Dublin, Ireland**

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning for Business |
| **Assessment Title:** | CA1 |
| **Lecturer Name:** | **Dr. Muhammad Iqbal** |
| **Student Full Name:** | Rachel Mae Lademora |
| **Student Number:** | 2023396@student.cct.ie |
| **Assessment Due Date:** | April 24, 2023 |
| **Date of Submission:** | April 24,2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of Contents

Introduction **4**

Project Objectives5

Motivation5

Description of the Problem Domain **5**

Rationale for chosen dataset 5

Task II5

Task III…………………………………………………………………………………………………………………………………………… 10

Conclusion……………………………………………………………………………………………………………………………………….16

References……………………………………………………………………………………………………………………………………….17

**Introduction**

In data science, clustering analysis is a basic method for organizing related data points into clusters. Measuring how similar or distinct various clusters are to one another is a crucial component of clustering analysis. This assessment is essential for assessing the quality of clustering solutions and figuring out the data's underlying structure.

These days, education is the primary factor in shaping us for a brighter future. An important way to determine whether a student does well in school, particularly in class, is to examine the student's performance in multiple areas. How well children do on each test they take in school will be determined by their test scores for each topic. (Rasul and Bukhsh, 2011). This will make it easier for the teachers to assign the student correctly in the event that their test results show a lower or higher score.

Deciding how stock prices affect the hotel segment is one of the most important topics in stocks. This may be connected to the increase in stock prices toward the end of 2019. The COVID-19 pandemic caused significant issues for people all around the world. It significantly reduces the hotel industry's revenue. (Carter et al., 2021)

**Motivation:**

Effectively quantifying the cluster's similarity and dissimilarity is essential when clustering is being used. I analyse the clustering algorithm and ascertain its strength when the clustering is completed. Have in-depth understanding of stock analysis; this will help you predict the future value of the stocks by showing you which month and year they have the best growth and drop fluctuation.

**Description of the Problem Domain:**

Implementing cluster analysis is needed to have a deep understanding of the data and to observe patterns if the dataset is properly segmented. One of the major problems with the real-world dataset is the outliers. The outliers have a bad effect upon executing the clustering because it will cause noise or improperly cluster the X label numerical variables. (Hassan, 2023). In performing the ARIMA time series, it is necessary to determine residuals because this can also cause noise or outliers to the forecasting value that can affect the evaluation results, such as r2 scores and the root mean square error.

**Project objectives:**

This project primary objectives are to evaluate and identify the clustering technique that best captures the similarity and dissimilarity between various clusters. Using the time series autoregressive integrated moving average (ARIMA) model, one may predict future values by identifying trends, residuals, and seasonality in the stocks market dataset. The goal of this research is to assess the most effective clustering algorithms for segmentation as well as the precise values of p, d, and q for the ARIMA model.

**Secondary Objective will address these questions:**

* Use the clustering criteria to implement and determine the distinction between the Davies-Bouldin index and silhouette score.
* From the student performance dataset, compare the outcomes of the Davies-Bouldin index and silhouette score clustering algorithms.
* Determine the stock market dataset seasonality, trend, and residuals.

**Rationale for chosen dataset:**

In order to gain insight into the pattern for accurately segmenting the points in each cluster that I will execute; I have chosen to use the education dataset student performance for my data analysis of clustering in this project. In addition, I utilize the hotel stock market dataset to calculate the extent of the COVID-19 pandemic impact on inflation. It will determine which month sees a significant drop in hotel stock, as some hotels are forced to close as a result of the COVID-19 epidemic.

Task II

1. **Which clustering algorithms would you consider for segmentation, and why?**

The clustering algorithm that I analysed to be considered for segmentation is DBSCAN clustering. The DBSCAN clustering gives me a well-informed, detailed cluster of the student performance dataset using the X labels as numerical variables. In order for me to achieve this accurate circle type of clustering point, I used to convert the columns from integers to strings using the preprocessing function shown in Figure 1. The parameters of the DBSCAN that I used to execute to determine which has the highest number of epsilons to find out the circumference of the neighbours in the points. This is shown in figure 2. The DBSCAN is setting an epsilon value of 40 and a minimum sample score of 5, as clearly shown in figure 3. Implementing DBSCAN with these parameters resulted in proper segmentation of the dataset, even in the presence of noise or outliers (visualized in Figure 4). Density-based spatial clustering of applications with noise (DBSCAN) ability to identify dense regions of data points makes it particularly effective for segmentation tasks where the underlying structure of the data may be irregular or non-linear.

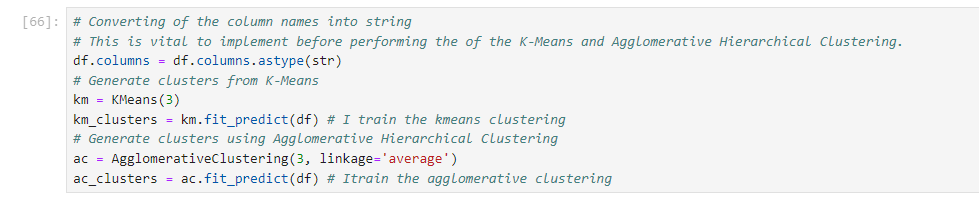


Figure 1. Code function for converting the integer column to string.

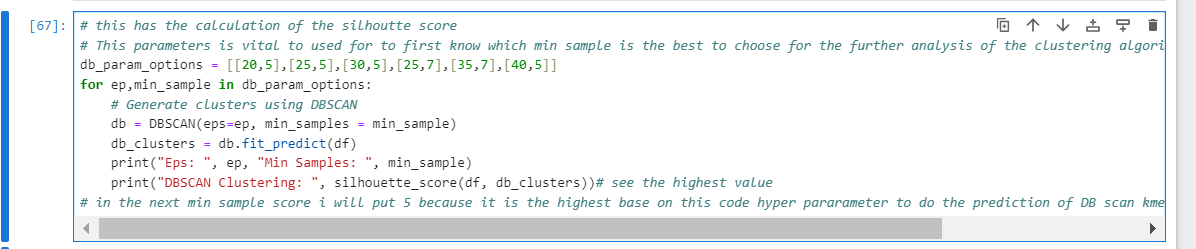


Figure 2. Using parameters in DBSCAN.

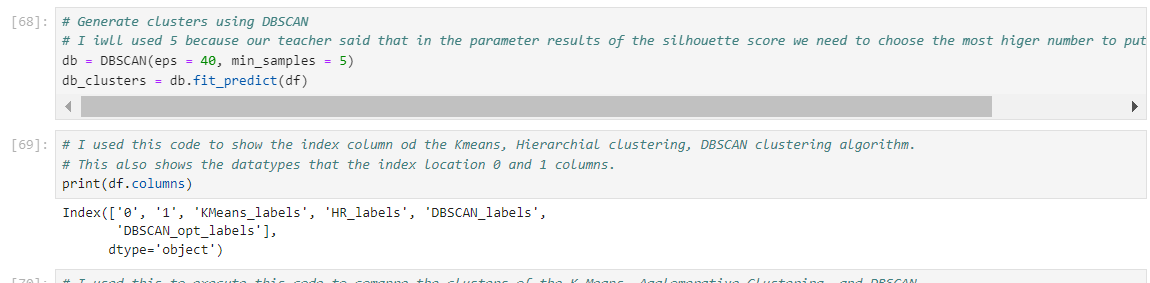


Figure 3. Setting the epsilon value and min sample for DBSCAN prediction.

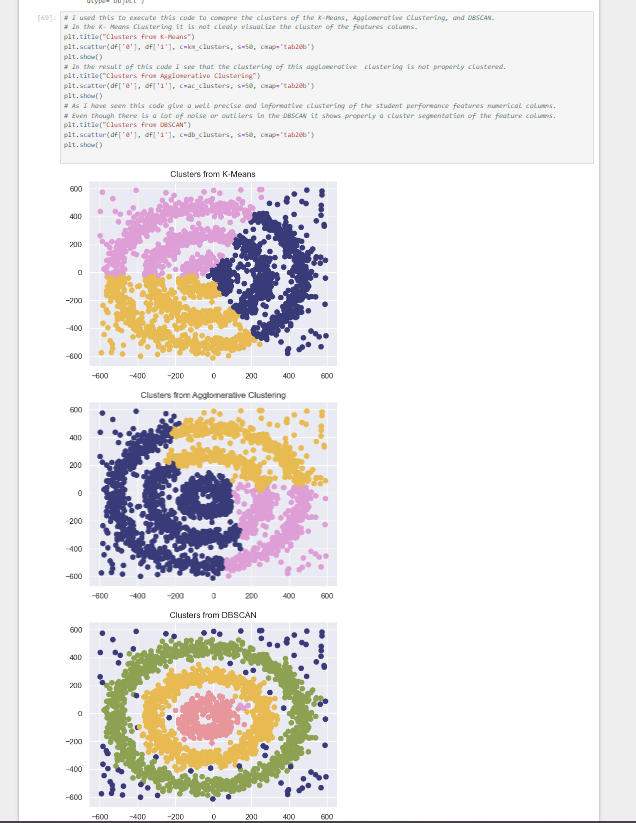


Figure 4. Visualizing the most informative clustering segmentation.

1. **Explain the differences between silhouette score and Davies-Bouldin index in the context of clustering.**

The silhouette score construe in any pattern which measure the similarity of its own cluster compared to the other cluster. It has a range of -1 and +1. The +1 range it is a high and well indicate a better clustering but it will be having a not good neighbouring clustering. While the silhouette score is O it will results to imbricated clustering resulting to have a negative value that will lead to a negative clustering. (ximnet, 2021). As I do my data analysis silhouette score on the first Kmeans that I have implemented it shows that the graph gets more convex flatten on the value if 0.35 this is shown in Figure 5. While I execute the silhouette score using the DBSCAN parameters for clustering it improves and increase the score in gives the results of 0.36. The Davies-Bouldin index measure moderate or the median similarity in each of the clusters. It has the ability to calculate the average of the overall cluster’s similarity. The lower the value of the Davies-Bouldin index meaning it will provide an accurate clustering. If it has a lower value this will indicate a well disconnected cluster. (GeeksforGeeks, 2023).

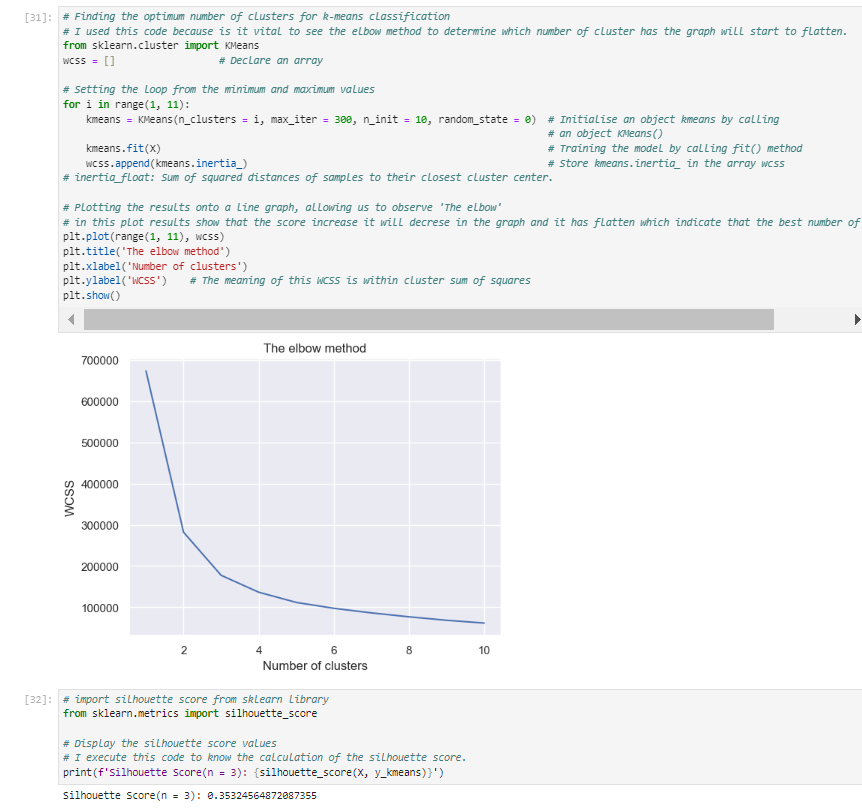


Figure 5. First K means Clustering silhouette score using the X labels numerical columns.

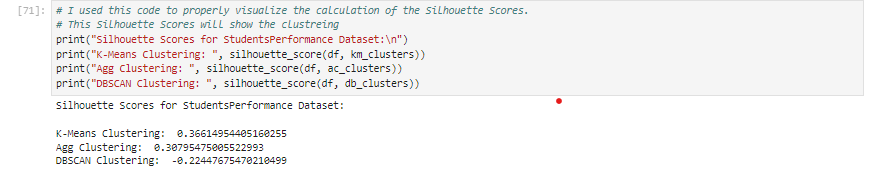


Figure 6. Silhouette score results using the DBSCAN parameter.

**C. Compare the results obtained from any two clustering algorithms from the chosen dataset.**

The results of the two-clustering algorithm that I used to do my analysis is having a significant plot of the elbow method. It is because the Davies-Bouldin index decrease the index which that can have an accurate performance of the clustering disengagement amid the clusters. One of the reasons why it forms a convex elbow plot because it has a greater convex clustering aside from the other density predicated cluster. (Wei, 2020). The reason why the silhouette score has the concave graph plot it is because of the cluster range that shown in figure 7. It set the cluster range in (2, 11) this is the calinski harabasz index. This index helps to properly separate the the clusters to each other. (Habib, 2021). The Davies-Bouldin index decrease the index has a detailed visualization of the plot compare to the silhouette score, It clearly show the Davies-Bouldin index appears to start to flat at the value of 3. While the silhouette score appears to be higher concave plot at the value of 3.

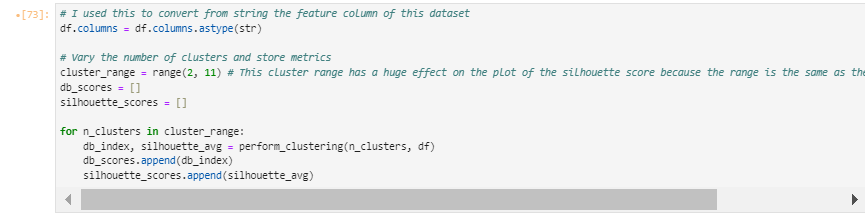


Figure 7: Cluster range that cause the silhouette score and the Davies-Bouldin index to have a different graph.

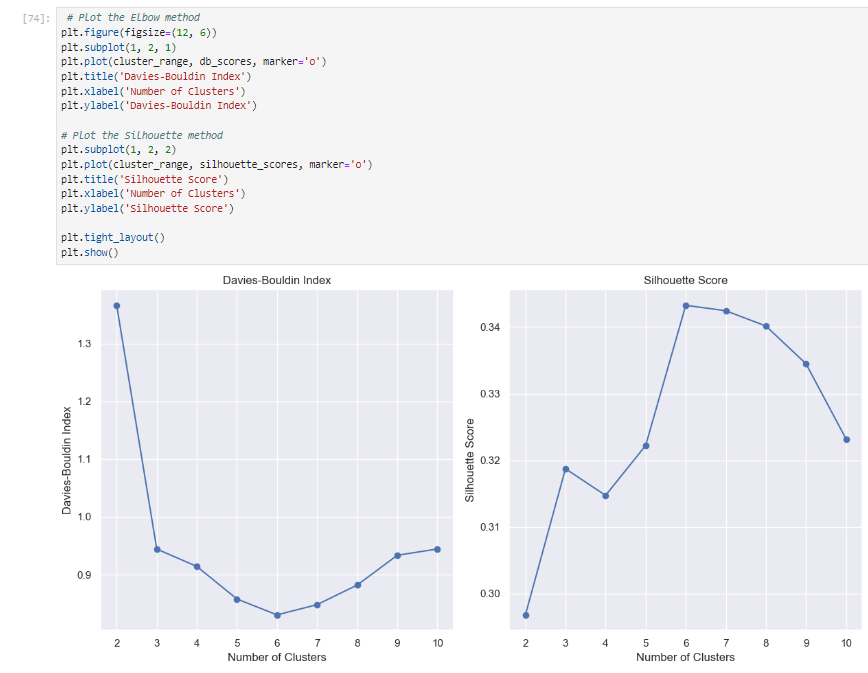


Figure 8: Comparing the results obtain from Silhouette Score and the Davies-Bouldin index.

Task III

**What insights can you derive from the initial exploration of the time series data based on the provided topics? Describe any trends, seasonality, or anomalies observed.**

In this time series analysis, I used the Hilton Grand Vacations Inc. (HGV) stocks dataset that I get from the yahoo finance. In June 01 2020 there is a huge decrease of stock when the lock down starts on that month it showed in Figure 9. Since all of the hotels are close it start to have vast of wide range of inflation because there are no more guest coming to check in in the hotel. Before implementing the augmented dickey-fuller test it vital to analysed first the trend , seasonal and the residual plots this will serve as a guide to do the step by step way of properly implementing the Seasonal ARIMA model it is clearly shown in figure 10. In order determine if there is animalities in the dataset close column it need to calculate the p value should be lower than 0.5 to reject the null hypothesis and make the time series trend to make the time series close column to be stationary. I did twice to test the p value before I get the null hypothesis to be rejected.it is shown in Figure 11 that the trend become a stationary.

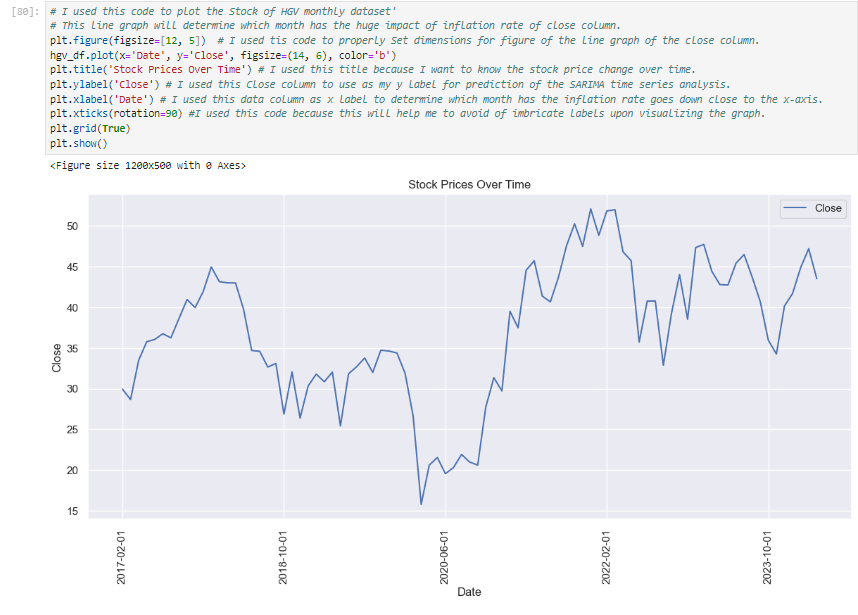
Figure 9. Showing the huge impact of inflation in the year 2020-2021

Figure 10. Showing the trend seasonal and residual of the HGV dataset.

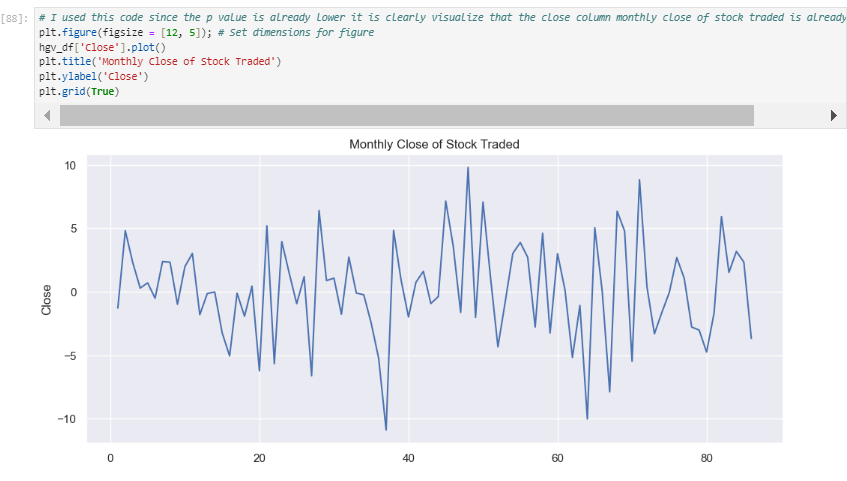


Figure 11. Visualizing the stationary plot monthly close of stock traded.

**How did you determine the appropriate parameters (p, d, q) for the ARIMA model.**

I determine the approach of the parameters by implementing the code that will show in Figure 12. In order to get the appropriate parameter to be used I need to pick the very lowest number base on that code array numerical values results. The next step of it is to perform and implement the array values to this code that will be shown in Figure 13. This is very vital to implement because this shown the SARIMAX result it has shown the if the dataset has a skewness. It will also show the coefficient value and the standard error.

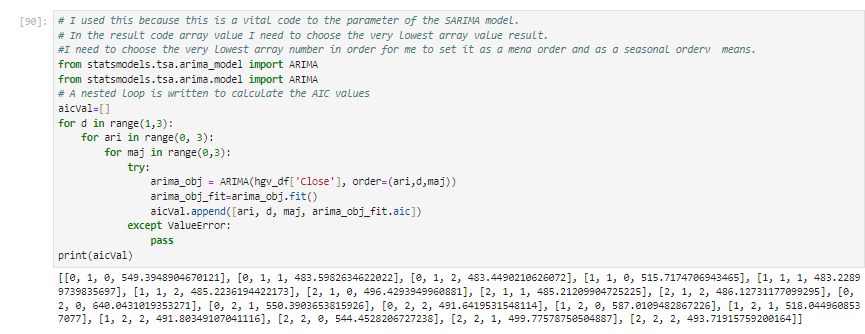


Figure 12. Parameter code for the p,d,q for ARIMA model.

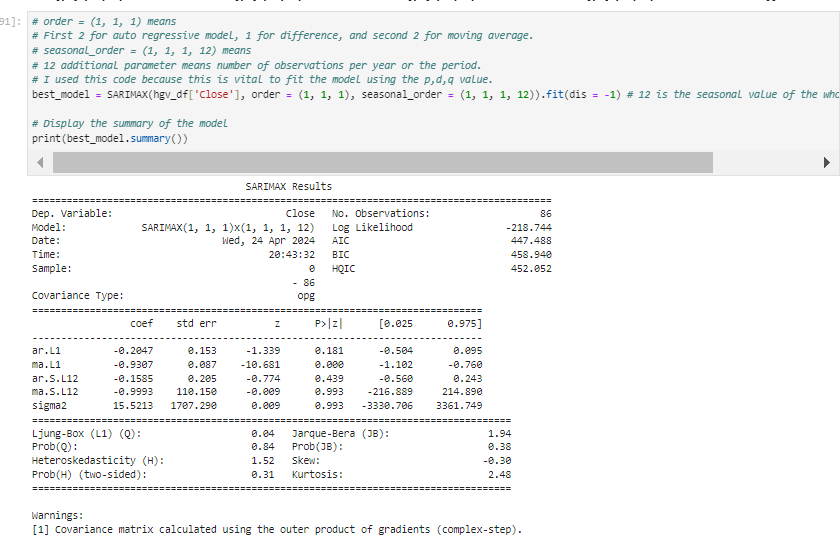


Figure 13. SARIMAX Results.

**Evaluate the performance of the ARIMA model in forecasting future values, highlighting any strengths and limitations based on your chosen dataset.**

I evaluate the ARIMA model for forecasting using the five years which equivalent to 60 months. It shows that the forecast values for five years are stationary it is clearly shown in Figure 14. Additionally, It is vital to highlight that this HGV dataset has a good value of the r2 score of -0.94272 and the value of root mean squared error is 4.141536 meaning there is a less error in the SARIMA time series forecast.



Figure 14. Visualizing the forecast for 5 years.

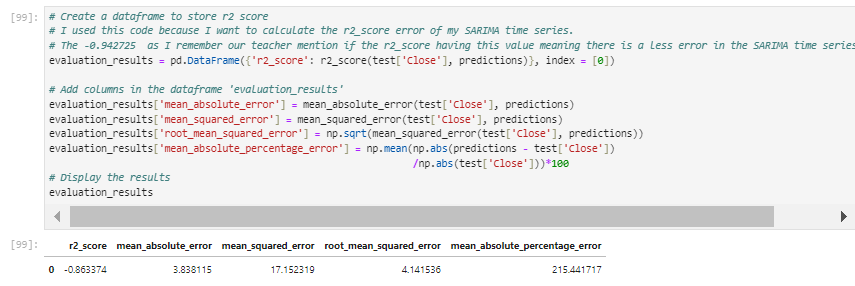


Figure 15. Visualizing the r2 score and the root mean squared error.

**Conclusion:**

Overall, I can say that the this CA1 of machine learning for business give me a deep understanding of the clustering on how I can properly know the clusters is well segmented. I can say I learned more in this CA because I understand the code that I execute to do my data analysis for the time series forecast. I face many challenges especially in doing the code because I have so many times, I encounter some errors but I try my best to overcome to solve it.

**References**

Carter, D., Mazumder, S., Simkins, B. and Sisneros, E. (2021). The Stock Price Reaction of the COVID-19 Pandemic on the Airline, Hotel, and Tourism Industries. *Finance Research Letters*, 44, p.102047. doi:https://doi.org/10.1016/j.frl.2021.102047.

GeeksforGeeks. (2023). *Davies-Bouldin Index*. [online] Available at: <https://www.geeksforgeeks.org/davies-bouldin-index/>.

Habib, A.B. (2021). *Elbow Method vs Silhouette Co-efficient in Determining the Number of Clusters*. [online] Medium. Available at: https://adria708.medium.com/elbow-method-vs-silhouette-co-efficient-in-determining-the-number-of-clusters-33baff2fbeee [Accessed 24 Apr. 2024].

Hassan, A. (2023). *What Is Cluster Analysis? (Examples + Applications) | Built In*. [online] builtin.com. Available at: https://builtin.com/data-science/cluster-analysis.

Rasul, S. and Bukhsh, Q. (2011). A study of factors affecting students’ performance in examination at university level. *Procedia - Social and Behavioral Sciences*, 15, pp.2042–2047. doi:https://doi.org/10.1016/j.sbspro.2011.04.050.

ximnet (2021). *Understanding your performance metrics for clustering*. [online] Medium. Available at: <https://ximnet.medium.com/understanding-your-performance-metrics-for-clustering-e7ec01762252>.

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.