**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning for Business and Data Visualisation Techniques |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | James Garza/Sam Weiss |
| **Student Full Name:** | Danton Escobar Prestes |
| **Student Number:** | Sbs02386 |
| **Assessment Due Date:** | 26/11/2023 |
| **Date of Submission:** | 01/12/2023 |
| **GITHUB REP:** | https://github.com/dantonprestes/CA\_Integrated.git |

**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. |

**" DATA"**

Data Visualization Techniques

Machine Learning for Business

November, 2023

Danton Prestes

**Table of Content:**

1. Introduction.....................................................................................................5
2. Argument........................................................................................................6
3. Question 1…………………………………………………………………………..7
4. Question 2…………………………………………………………………………11
5. Question 3…………………………………………………………………………17
6. Conclusion…………………………………………………………………………22
7. References………………………………………………………………………...23

**Table of Figures:**

*Figure 1: Head() function, showing first 5 rows of the dataset.......................................................6*

*Figure 2: Dataset dimensions showing column and rows……………………………………………………………7*

*Figure 3: Arima Model………………………………………………………………………7*

*Figure 3: Forecast Errors……………………………………………………………….…..8*

1. **Introduction**

The dataset used for the analysis and studies where activities proposed on the Continuous Assignment was pulled from Kaggle. The dataset was elaborated based on a ficticious jewelry company and aimed to know the best selling products thus it contains transactions that happened between 01.12.2018 and 01.12.2021.

The fictitious company mimics a jewellery company that sells pendants, necklaces, earrings, rings, bracelets, clocks and souvernirs with different colours and brands.



*Figure 1: Head() function, showing first 5 rows of the dataset.*

The dataset was chosen in order to use this data for a forecast which would indicate a clear behaviour of consumers with insights on figures that are mostly spent with products.

**Data Attributes**

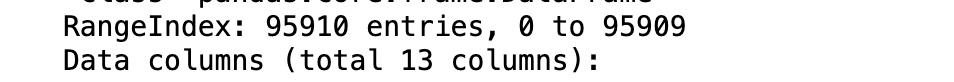
* Date of the order (order\_date)
* ID of the order (order\_id)
* ID of the user (user\_id)
* Number of purchases) purchase
* Category ID (category\_id)
* Category of the product (category)
* Brand of the product brand
* Price of the product (price)
* Price ID (price\_id)
* Genre of the client (buyer)
* Colour of the product (colour1)
* Secondary colour of the product (colour2)
* Type of product (type)

In the next section, more details about how the Machine Learning as well as Data Visualisation can be elaborated and will perform with the dataset chosen.

1. **Argument**

Initially, the dataset chosen has to be treated which means the Exploratory Data Analysis (EDA) was applied to have a clear and facilitate the implementation of techniques for Machine Learning as well as provide structure to the development of the code.

Initial analysis presents dataset with 95910 rows and 13 columns.



*Figure 2 : Info() function, showing number of rows and number of columns of the dataset.*

The main purpose of the analysis is to be able to identify trending on sales and forecast based on observations as well as use Box-Jenkins model to provide a time series analysis output. Since time series comprises a systematically arranged set of observations, each corresponding to specific time intervals we would need to consider the importance of having a stationary dataset and perform analysis based in an attribute that would not pose challenges in terms of data handling such as gaps or non-linear data (NaN or null values).

Time series is a regular event in real life and it is more common than we can notice being part of our day to day activities, where we can read variations across time for example, in energy bill which can shows higher or lower consumption in Kwa within periods of time (hour, days, months). Other examples are fuel consumption of vehicles which measures how far (in kilometres or miles) a car can run with 1 litre of diesel/petrol or 1 gallon.

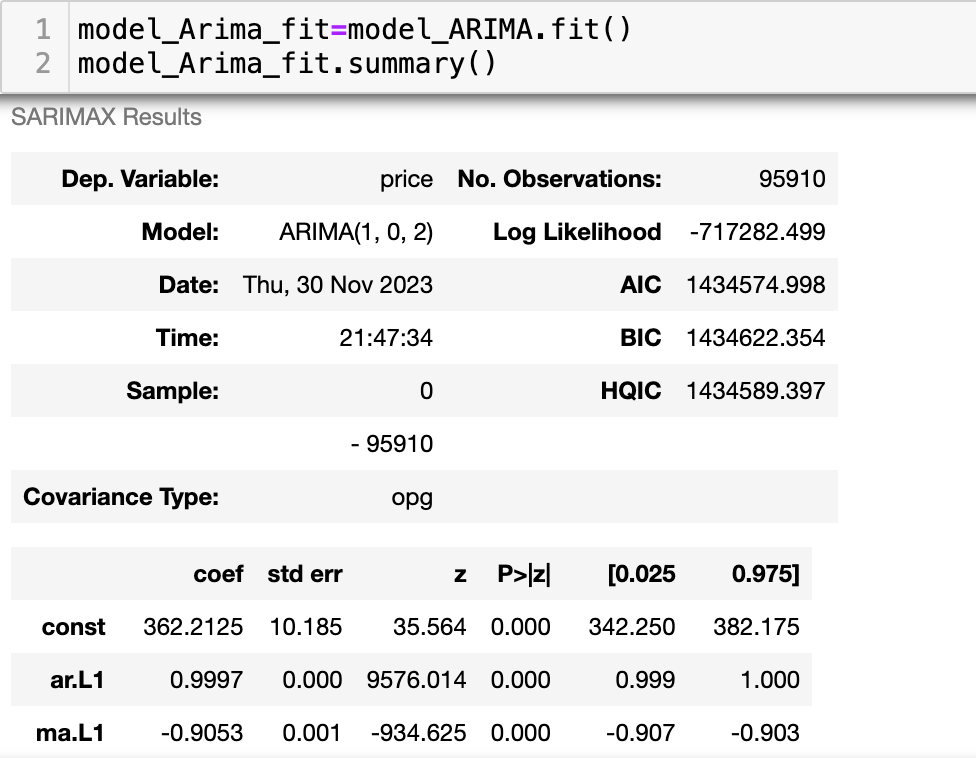
This are examples for simple events that can be assessed with time series analysis.

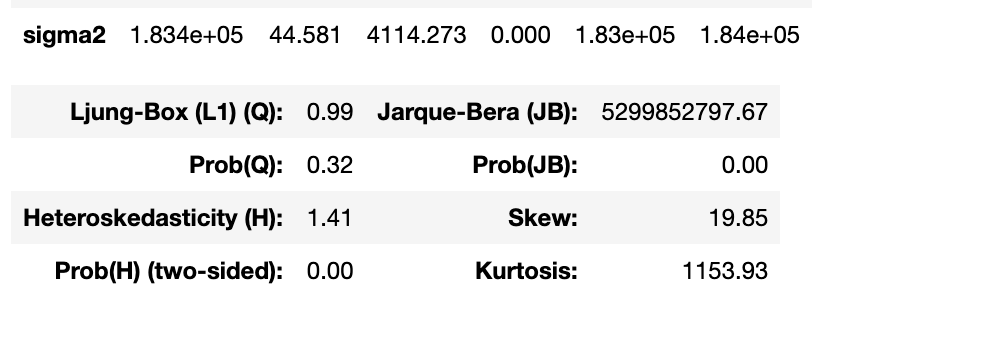
Additionally, the Augmented Dickey-Fuller test (ADF Test) is a prevalent statistical method employed to assess the stationarity of a given time series. It stands out as one of the most frequently utilized tests in the analysis of stationarity. Since the ARIMA model is typically applied to stationary datasets, the Dickey-Fuller Test serves as a crucial step in verifying the stationarity of the dataset before applying the model.

1. **Question 1  
     
   Arima Model**

In order to analyze the performance of the Machine Learning, the Arima (autoregressive integrated moving average) model was chosen to identify the statistic and econometric events which occur in a timeframe. “*The ARIMA model is a generalization of the ARMA model (AutoRegressive Moving Average model), suitable for handling non-stationary time series*.”  
So from what literature explains, the models are implemented to help to forecast/predict events based in previous observations. This means to the logical thought that events occurred in past will affect the prediction, however there are factors that may imply higher impact or lower impact due to the characteristics of the time series. Considering that events that are not random during a sales process as it has a recurrency as well as a pattern (even it can present some seasonal influence) the process assessed to be used was towards the concept of univariate which only uses previous values in the time series to predict future values.

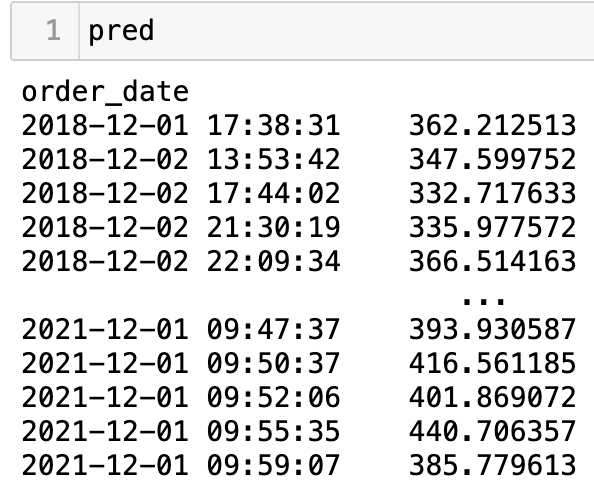
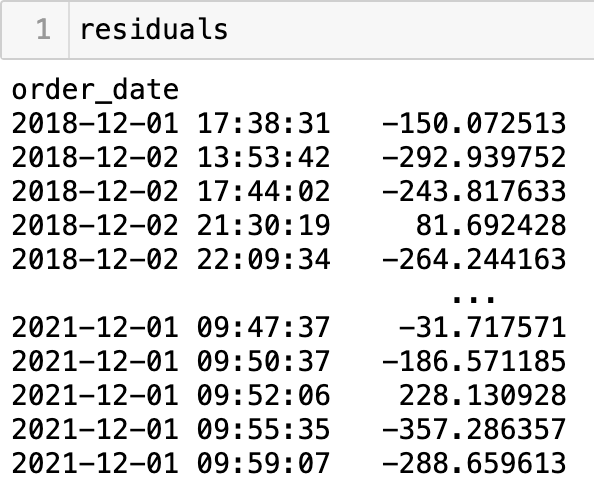
**a)Arima model applied:**

****

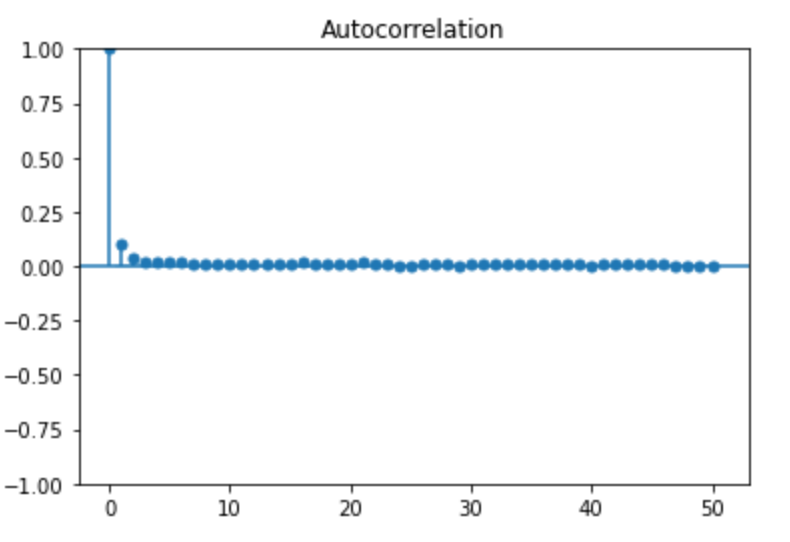
****

**Figure 2: Arima model.**

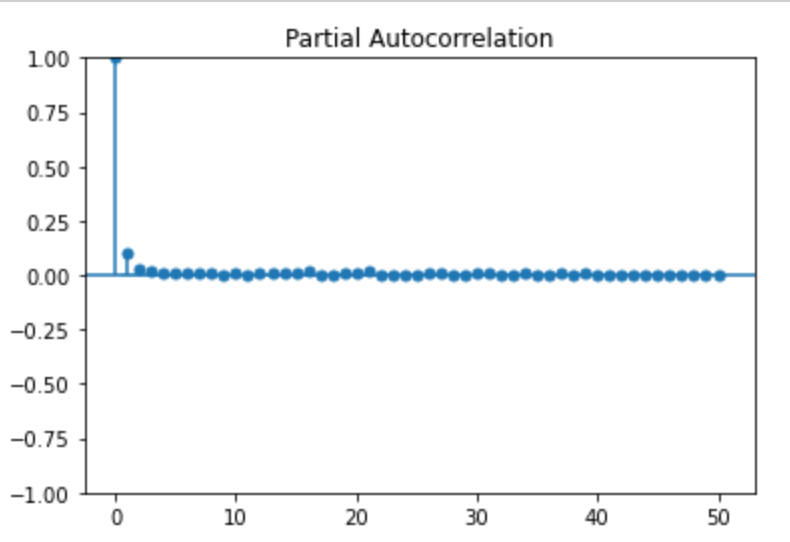
**b) Forecast errors.**

**  
**

**c) ACF:**

****

**PACF: .**

****

**\*\*ACF:\*\* Direct and indirect effect of values in previous time lags. In the graph we can see the strongest correlation at early points of the time series and the following lags are stabilized with weak correlation between lags (but not negative).**

**\*\*PACF\*\*: Plot shows the most common scenarion for PACF which is a significant spike on the initial lags and drops to zero on subsequent lags.**

1. **Question 2**

**Text analytics**

The dataset used for Text Analysis using analytics tools, was also collected from Kaggle. The main idea was to

1. **References**

[**MDPI**](file:///Users/dan/Documents/CCT%20-%20Data%20Science/Semester%202/CA2%20-%20Integrated%20ML%20and%20Visualisation/MDPI)**. 2023 [online] Available at: <** **https://www.mdpi.com/1999-5903/15/8/255> [Accessed 30 November 2023] A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks https://www.mdpi.com/2413114**