<https://www.kaggle.com/code/khansahil128/online-shopping-dataset> from her

Kate O Dwyer (sbs23025)

CA2 (***Integrated***)

**Data Visualisation and Machine Learning for Business**

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Visualisation and Machine Learning for Business |
| **Assessment Title:** | CA2 (Integrated) |
| **Lecturer Name:** | James Garza and Sam Weiss |
| **Student Full Name:** | Kate O'Dwyer |
| **Student Number:** | sbs23025 |
| **Assessment Due Date:** | 26/11/2023 |
| **Date of Submission:** | 26/11/2023 |

**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**

Contents

[1. Data Visualisation 4](#_Toc151922916)

[Preprocessing and Cleaning 4](#_Toc151922917)

[Wireframe 4](#_Toc151922918)

[Plots Created 6](#_Toc151922919)

[Map Plot 6](#_Toc151922920)

[Stacked Bar Chart 6](#_Toc151922921)

[Line Plot 7](#_Toc151922922)

[Pie Chart 8](#_Toc151922923)

[Other Panel Sections 9](#_Toc151922924)

[Heading 9](#_Toc151922925)

[Tables 10](#_Toc151922926)

[Text 11](#_Toc151922927)

[2. Machine Learning for Business 12](#_Toc151922928)

[Question 1 12](#_Toc151922929)

[Introduction 12](#_Toc151922930)

[Argument 12](#_Toc151922931)

[Conclusion 13](#_Toc151922932)

# 

# Data Visualisation

## Preprocessing and Cleaning

The data was preprocessed and cleaned in the Machine Learning for Business section of this assignment. The dataset features (Table 1) contained missing and repeated values, as well as different data types.

**Table 1.** Data set prior to data preparation and cleaning.

|  |  |
| --- | --- |
| **Features (Columns)** | 18 |
| **Observations (Rows)** | 9880 |
| **NaN Values** | Yes |
| **Data types** | Object(15), int64(1), and float64(2) |

The data was cleaned and prepared to perform exploratory data analysis. This included:

* Removing ‘Row ID’ which contained the index information, and was not necessary
* Removing null values (NaNs)
* Removing one duplicate row.
* Order Date was transformed to datetime64 to have in a date series format
* The data set was sorted by order date
* Average sales was added
* Year was added

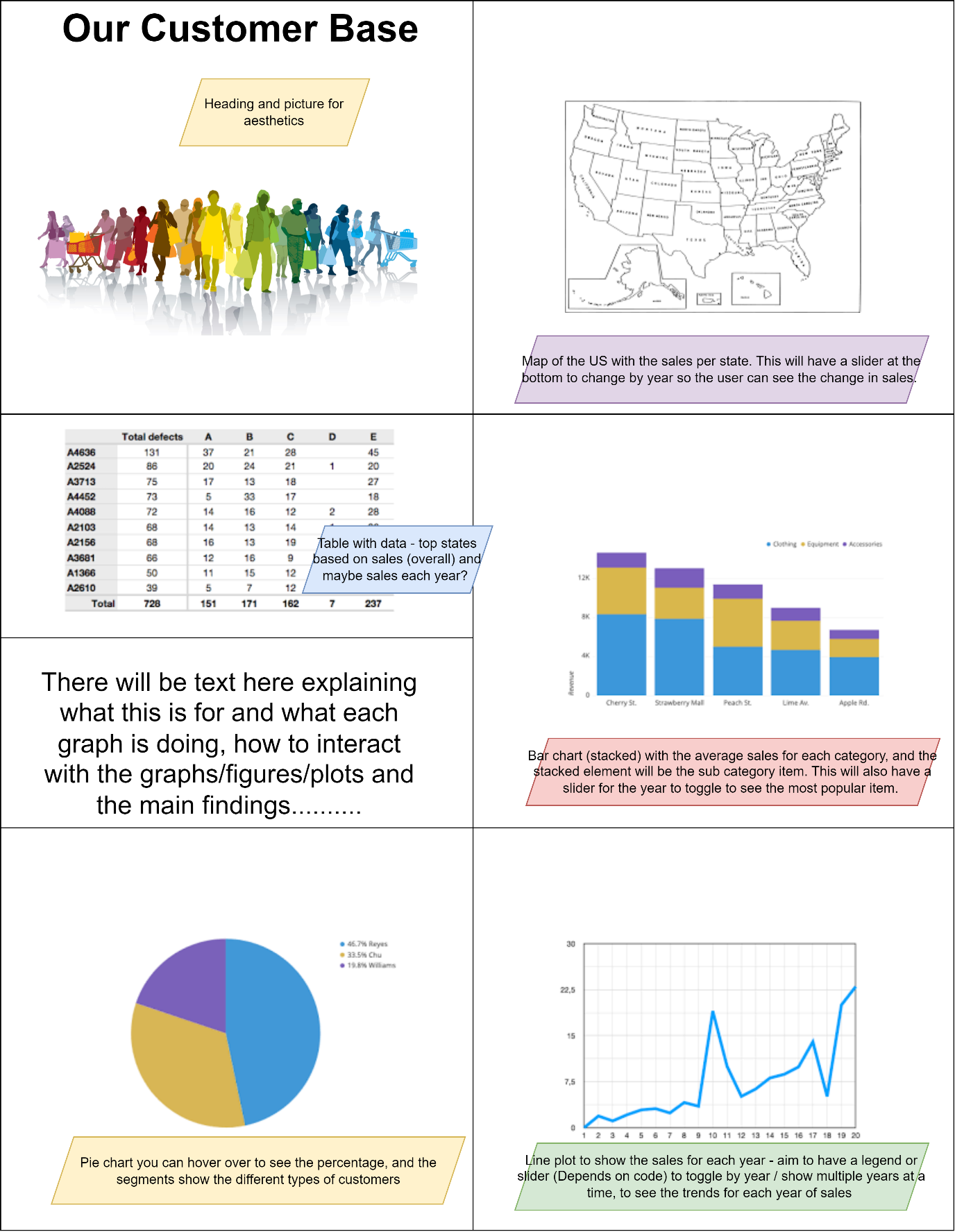
The new data set (Table 2) will be used for this report as the data preparation was deemed sufficient.

**Table 2.** Data set prior to data preparation, cleaning, and column addition.

|  |  |
| --- | --- |
| Features (Columns) | 18 |
| Observations (Rows) | 9,789 |
| NaN Values | No |
| Data types |  |

## Wireframe

Before creating a dashboard for this data set, a wireframe was created using draw.io to brainstorm ideas for how the dashboard should look (Figure 1)

******

***Figure 1.*** *WireFrame for dashboard precreation.*

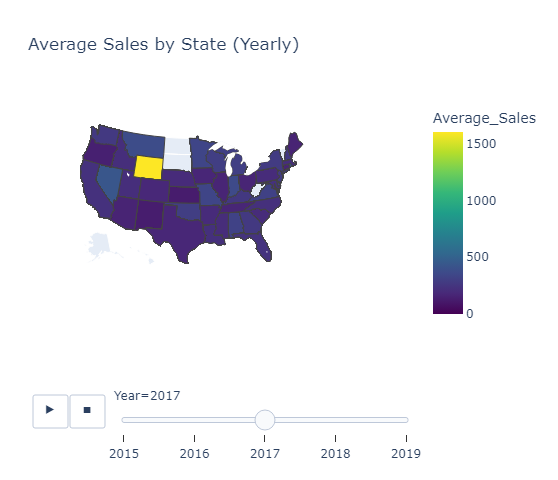
The wireframe shows the layout of the dashboard before creation, and has explanations on each segment, and why these would be beneficial for the user.

## Plots Created

All plots were created using px.plot, which creates interactive plots on Jupyter notebook. This data frame consisted of many categorical variables, and I wanted to show as much data as possible with the plots

### Map Plot

A map plot (Figure 2) was created to show the average sales per state and slider to change the year. This plot is interactive, and when the user hovers over a state it shows the state (as two letters, for example New York is NY) and the average sales. The slider can be used to change the year. This plot is beneficial at also showing what states did not have any sales for a given year, which are the greyed-out states.



***Figure 2.*** *Map Plot of the United States use in the top right hand segment of the Panel dashboard.*

This plot was created to understand the customers from a geographic perspective, i.e., which states spend the most and which states spend the least. From using this plot, it was clear that most of the states have similar sales, mostly less than 500. In 2017 however, Wyoming was the highest state with 1200, this was insightful and can be something the company can investigate to determine why Wyoming had such a big spend that year. A lot of states did not have purchases for 2019 this was because the data set does not go past January 2019.

The colour used is the default, however I think that it is very helpful in showing the differences as the yellow and blue are so different but there is a gradient to show slight differences.

### Stacked Bar Chart

A stacked bar chart (Figure 3) was created to visualise the top categories and the subcategories within these, based on average sales, with the year on a slider similar to Figure 2.

A screenshot of a graph

Description automatically generated

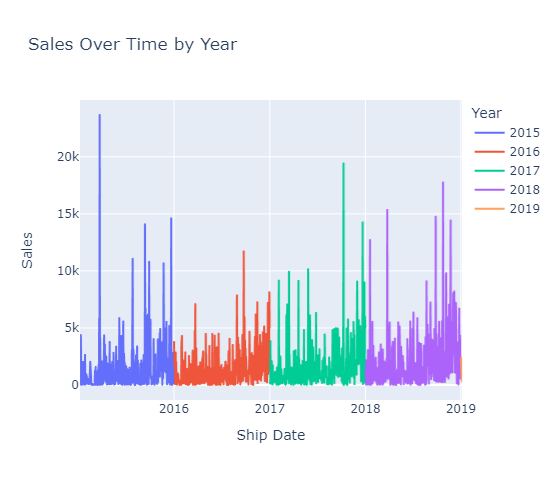
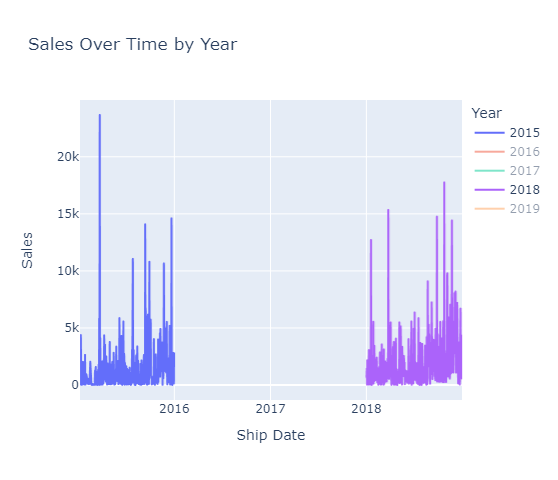
***Figure 3.*** *Stacked Bar Chart for Subcategories and Categories.*

There were three categories: Furniture, Office Supplies, and Technologies. There were many subcategories within these, and when the user hovers over the bar chart, the top five (there may be less than five subcategories within a category) subcategories are shown. This graph is very helpful in determining which category had the highest sales each year, and within that category, it is easy to determine which subcategory sold the most. Although the colours do overlap, (tables and copiers for example) when the use hovers over the bar chart, it is highlighted what the subcategory is.

The top category each year was technology, and depending on the year the top subcategory within technology was copiers and machines.

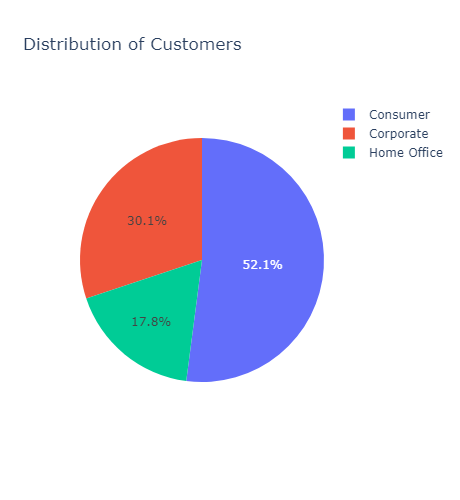
### Line Plot

A line plot (Figure 4) was created to visualise the sales per day for each year. A slider was not used here and instead a legend used where the user can pick which years to show.

***Figure 4.*** *Line plot for the sales each year. Left hand side (LHS) showcases all data. Right hand side (RHS) is an example of the interactive element of the plot, where the user can click the legend to remove or add years.*

The sales per year increased every year (this is in the table on the LHS of the dashboard) which isn’t necessarily evident from the graphs plotted in Figure 4. However it is interesting that each year there is the same rough trends in the sales visually. There was only a small amount of data for 2019 hence 2019 cannot be seen well in the images provided. Different colours were used by default and they are different enough that they did not need to be changed to enhance usability.

### Pie Chart

A pie chart (Figure 5) was created to visualise the percentage of customer types: consumer, corporate, and home office.

***Figure 5.*** *Pie Chart for Customer Types.*

This pie chart is interactive; the user can hover over the segments to see which customer type the segment represents and the associated percentage. The ‘consumer’ section is the highest percentage (52.1 %) and home office is the lowest (17.8 %). This suggests that the company could increase sales by targeted home office items in discount and could also reach out to companies to increase the corporate section.

## Other Panel Sections

Other panel sections (Figure 6) were created.

A group of people walking in a line

Description automatically generated

***Figure 6.*** *Heading, tables, and text section of the panel.*

### Heading

The heading (Figure 7) is the same as that in the wire frame (Figure 1) and the aim of this was to make the panel look attractive to the user, and to make it as colourful as possible.

A group of people walking with shopping cart

Description automatically generated

***Figure 7.*** *Heading and picture used in the panel dashboard.*

### Tables

Two tables outlining the sales per year and top sales per state were added using the tab command for the panel, where the user can click on the tab and show the desired table (Figure 8).

A screenshot of a computer

Description automatically generated

***Figure 8.*** *Tables on the Panel, activated by clicking on the tabs.*

This section is to give a brief overview of important information that the plots don’t capture in a clear and specific manner.

### Text

A group of people walking in a line

Description automatically generatedThe text section explains each section of the panel dashboard, and the key findings from the visualisations performed (Figure 9). These have been explained in the ‘Plots’ section above.

***Figure 9.*** *Text section of the Panel Dashboard.*

# Machine Learning for Business

## Question 1

### Introduction

Time series analysis is the act of recording data at consistent time intervals over a set time period, however, it is not just recording the data points over time. What is important in time series analysis is how the variables change over time. (Tableau, 2023) The data set chosen is for online transactions for a company, spanning from 2015 – 2019. The variable of interest for this analysis was the Sales.

The Box Jenkins ARIMA model are univariate models used to better understand a single variable. To use these models, the data must be stationary. The Augmented Dickey-Filler (ADF) test is a statistical test to determine if the time series data is stationary (Santra 2023). This test is more complex than the Dickey Fuller test as it takes into account more complex data sets and larger data sets.

Based on the model chosen, the order must be chosen (known as p, d, and q variables). In order to do this the Akaike Information Criterion (AIC) is a measure of goodness of fit for the ARIMA model, in the case of this report, the AIC was found for different values for the order to determine the best fit. (Armstrong, 2001).

The autocorrelation function (ACF) is used to measure the linear relationship between a data point at time *t* and the data at the *t -1*  and so on, i.e. the ACF is the correlation between the two values. This aids in determining if there is a linear relationship. The partial autocorrelation function (PACF) is used to find the order of a model. Both of these can be plotted for the time series data and used to determine the order of the model and if there is any autocorrelation. (PennState, 2018)

### Argument

Firstly, the ADF test was performed. The null hypothesis states that the time series is nonstationary, and the alternative hypothesis stated that the time series was stationary. The results yielded a p-value of 0.0, and at a confidence level of 0.05 the null hypothesis can be rejected; hence this indicates that the time series was stationary.

Because this time series was stationary the ARMA model was used. To determine the order, the ‘p’ and ‘q’ values were investigated for a range of 0 to 5 for each an the AIC value found. This was done in order to find the best statistical fit for the model, as the lower the AIC value, the better the fit. The differencing is equal to 0 as the ARMA model is being used, hence there is no differencing as the data does not need to be turned to stationary.

The order was found to be (0,0,0) and this yielded a high AIC of 153,830.794. This suggest that this model is not the best fit, however upon looking at the data it does not appear to be seasonal. From this, the ARMA model was used to predict the last ten time observations and to see if these predictions were correct. The forecast errors (Table 1) and the actual versus predicted plot (Figure 10) show that the model was not successful in predicting the Sales data.

**Table 1.** Forecast error for the last ten predicted time points.

|  |  |
| --- | --- |
| **Time Point** | **Forecast Error** |
| 9779 | -128.93 |
| 9780 | -217.53 |
| 9781 | -161.59 |
| 9782 | -20.75 |
| 9783 | -209.33 |
| 9784 | 93.082 |
| 9785 | -177.28 |
| 9786 | -139.12 |
| 9787 | -216.15 |
| 9788 | -227.03 |

A graph with blue lines and orange lines

Description automatically generated

***Figure 10.*** *Actual and Predicted Sales Values for the last ten time series points.*

The time series plot (Figure 11) shows the fill data set and the ACF and PACF plots (Figure 12) were graphed.

A graph showing a graph of sales

Description automatically generated

***Figure 11.*** *Time series plot of the sales for online transactions.*

**A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated*Figure 12.*** *ACF (Left) and PACF (Right) Plot for the data.*

There is no strong correlation at lag = 1 for either the ACF or PACF plot, meaning that there is no strong autocorrelation for the sales time series.

### Conclusion

The ADF test was used to find that the data was stationary, and thus the ARMA model was used. The order of the ARMA model was found by finding multiple AIC values, and the order yielding the lowest AIC values was chosen. This AIC value was still quite high ( > 150,000) indicating that that ARMA model may not have been a good fit for the data set. The ARMA model was then used to predict the last ten time points, which it did poorly (please see the errors in Table 1). Finally, the ACF and PACF plots were graphed and indicated that there is not a strong correlation for the sales data. Overall, the model (ARMA) did not perform well for this time series data, as it did not appropriately predict the sales and had a high AIC value.

## Question 2

### Introduction

Text Analytics has many different branches to it. Text categorisation can include sentiment analysis, language detection, intent detection, and topic labelling. This is essentially taking in text and trying to add a label to it in some way. The data set chosen for this report was tweets relating to COVID-19. Sentiment analysis can be used to predict of a tweet will be positive or negative, for example. However, in order to use machine learning, there must already be a sentiment found on the data set. Topic modelling is a natural language processing (NLP) technique that can identify common themes in large text data sets. In the case of the COVID-10 data set, topic modelling could be used to find the main theme when people tweet about COVID-19 (For example, lockdown) (Peddireddi, 2023)

### Argument

Word clouds were created to get a better sense of the COVID-19 tweet data set (Figure 13-15)

***A close-up of words

Description automatically generatedA close up of words

Description automatically generatedFigure 13.*** *Word Cloud for HashTags in Tweets.*

A close up of words

Description automatically generated***Figure 14.*** *Word Cloud for the Tweets*

***Figure 15.*** *Word Cloud for Location*

The word clouds are useful at determining if there is any missing data (in Figure 15, for example, NaN comes up a lot for location) and can be good at initial check for topic in tweets. These tweets are learning about COVID-19 from Figure 14, but we can also see other topics, adjacent to COVID-19 are mentioned such as: Donald Trump, voting, and mental health.

In order to predict sentiment analysis a data set was taken where sentiments were already added to the tweets, specifically tweets from India. (Kumar, 2020). Vectorisation is used to map words to a vector of real numbers, which is then in turn used to predict words – this was used for sentiment analysis (Sieg, 2018) Count Vectorization was used as it is good for simple data sets, it can however be unsuitable when there is a lot of stop words but they were removed in this case.

For the machine learning aspect, a simple logistic regression model was used. The C values were tuned using GridSearchCV and this was cross validated using a repeated stratified k-fold validation. An accuracy of 67.5 % was found for C = 1.0 to predict the sentiment of the tweets. Although this is not a favourable result, there are many other machine learning models that could be tested to predict the sentiment. As well as this, this is a small data set particular to one country, and more data would make this more accurate.

### Conclusion

Sentiment analysis was performed for COVID-19 tweets from India. Using a Linear Regression model and Count Vectorization to turn the words in numbers, it was found that the model was 67 % accurate at predicting sentiments. Different Count Vectorization models, such as Bag of Words, or Transformers such as Bidirectional Encoder Representations from Transformers (BERT) could be used to improve the input going into the Machine learning model.

# References

Armstrong, J.S. (2001). ‘Principles of Forecasting: A handbook for Researchers and Practitioners’ New York: Springer.

Kumar, S. (2020) ‘Covid 19 Indian Sentiments on Covid 10 and Lockdown’ *Kaggle* [online] Available at: <https://www.kaggle.com/datasets/surajkum1198/twitterdata/data> Accessed 15th November 2023.

Peddireddi, Y. (2023) ‘Topic Modelling in Natural Language Processing’ *Analytics Vidhya* [online] Available at: <https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/#:~:text=Topic%20modeling%20is%20a%20natural,Latent%20Dirichlet%20Allocation%20(LDA)>. Accessed 26th November 2023.

PennState (2018). ’10.2 – Autocorrelation and Time Series Methods’ *Applied Regression Analysis*. [online] Available at: <https://online.stat.psu.edu/stat462/node/188/#:~:text=Autocorrelation%20and%20Partial%20Autocorrelation,%2C%202%20%2C%20.%20.%20.%20>. Accessed 26th November 2023.

Santra, H. (2023). ‘Tests for Stationarity in Time Series – Dickey Fuller Test & Augmented Dickey Fuller (ADF) Test’. *Medium* [online] Available at: <https://medium.com/@ritusantra/tests-for-stationarity-in-time-series-dickey-fuller-test-augmented-dickey-fuller-adf-test-d2e92e214360> Accessed 26th November 2023.

Sieg, A. (2018). ‘Text Similarities: Estimate the degree of similarity between two texts’ *Medium* [online] Available at: <https://medium.com/@adriensieg/text-similarities-da019229c894> Accessed 26th November 2023.

[Tableau (2023). ‘Time Series Analysis: Definition, Types, Techniques, and When It’s Used. *Tableau.* [online] Available at: *https*://www.tableau.com/learn/articles/time-series-analysis#:~:text=What%20is%20time%20series%20analysis,data%20points%20intermittently%20or%20randomly](https://www.tableau.com/learn/articles/time-series-analysis#:~:text=What%20is%20time%20series%20analysis,data%20points%20intermittently%20or%20randomly). Accessed 26th November 2023.