Contents

[Introduction 1](#_Toc151721903)

[Datasets summary 1](#_Toc151721904)

[Machine Learning for Business 2](#_Toc151721905)

[Concept of Time Series Analysis and applicability. 2](#_Toc151721906)

[The Augmented Dickey-Fuller test in time series 2](#_Toc151721907)

[Box-Jenkins Models 2](#_Toc151721908)

[Weekly ADR Plot 3](#_Toc151721909)

[Stationarity of the series 3](#_Toc151721910)

[Checking for Stationarity 4](#_Toc151721911)

[Time Series Plot of the Data 4](#_Toc151721912)

[Series Autocorrelation and Partial Autocorrelation Plots 5](#_Toc151721913)

[One-step-ahead forecasts of the last 10 observations 6](#_Toc151721914)

[Forecast Errors 6](#_Toc151721915)

[Concept and application of Text Analytics 7](#_Toc151721916)

[Political Social Media Posts Plot 8](#_Toc151721917)

[LDA (Latent Dirichlet Allocation) 9](#_Toc151721918)

[World Cloud 9](#_Toc151721919)

[Predicted Topics Results 9](#_Toc151721920)

[Sentiment Analysis 10](#_Toc151721921)

[Conclusions 10](#_Toc151721922)

[Data Visualisation Techniques 11](#_Toc151721923)

[Wireframe 11](#_Toc151721924)

[Dashboard Summary 12](#_Toc151721925)

[Countries the Guests come from 12](#_Toc151721926)

[Line chart showing trends that change over time across the ADR. 13](#_Toc151721927)

[Target variable across different categories. 14](#_Toc151721928)

[Histogram for Market Segment & Customer Type 15](#_Toc151721929)

[References 16](#_Toc151721930)

[Machine Learning for Business 16](#_Toc151721931)

[Data Visualisation Techniques 17](#_Toc151721932)

<https://github.com/ArielGoldman89/CA2-MLB-and-DV-Integrated>

Number of Words: 2216 (excluding titles, diagrams, contents, and HARVARD References).

# Introduction

The objective of this project is to perform Time Series Analysis and Text Analytics for Machine Learning for Business and create an interactive Dashboard for Data Visualization Techniques. Through both analyses, we aim to gain valuable insights into two different real-world datasets: Political Social Media Posts and Hotel Booking.

On one hand, we will use the dataset called Political Social Media Posts to apply Text Analytics where we will analyse 5000 messages from politicians’ social media accounts, along with human judgments about the purpose, partisanship, and audience of the messages posted on Twitter and Facebook.

On the other hand, we will work on the dataset called Hotel Booking to apply Time Series Analysis and use the same dataset to create an interactive Dashboard to effectively communicate the key insights derived from the exploratory data analysis. In this particular dataset, we can find information of two different hotels: the City Hotel and Resort Hotel, offering a comprehensive look into one key factor influencing the hospitality industry: seasonality, where there is a repeated behaviour. This behaviour is related to peak seasons (high demand) and off-peak seasons (low demand).

# Datasets summary

The Hotel Booking dataset contains over 100,000 observations, with the target variable ‘is\_canceled’ column indicating whether the bookings were cancelled or not.

There are key features to use in the interactive Dashboard.

* Type of Hotel
* Country: Represents where the guests’ country origin
* Average Daily Rate
* Arrival Date Month: Represents the number of arrivals per month.

For the Time Series Analysis, the key features include:

* Reservation Status Date: Represents the date where the guests booked a room.
* Average Daily Rate: Total revenue generated by all the occupied rooms.

The other dataset, Political Social Media Posts, provides 5000 messages, where the sources of the messages are equally distributed: 2500 are from Twitter and 2500 from Facebook posts.

The key features for Text Analytics are:

* Unique ID: Identifies each unique message.
* Judgement timestamp: Contains the date and time when the message was posted.
* Text: is the message itself sent by the politicians.

# Machine Learning for Business

In this phase of the project, our objective is to focus on two topics: predicting trends over the time by applying Time Series Analysis and using Text Analytics to gain deeper insights about the messages from the politicians.

## Concept of Time Series Analysis and applicability.

The concept of Time Series Analysis involves data collecting over time to identify patterns, trends, and make predictions about future values in constant time intervals, such as daily, weekly monthly, quarterly, or yearly. It also includes forecasting, which means predicting unknown values based on the collection of historical data and then estimating future values based on patterns learned from historical data. For example, one interesting aspect of this concept is the applicability across different fields, such as finance to predict stock process or in weather forecasting, where it helps anticipate climate patterns, and in economic analysis, where it contributes to assessing the gross domestic product (GDP) for a particular country over the years.

## The Augmented Dickey-Fuller test in time series

The purpose of the Augmented Dickey-Fuller test in Time Series is used to test whether a given Time Series is stationary or not. In the context of Time Series Analysis, stationary is important because it makes it easier to understand and predict future values. In other words, Time Series are stationary if they do not have trend or season effects.

# Box-Jenkins Models

We need to apply statistical methods used in Time Series Analysis. If our Time Series model is stationary, we can use the AutoRegressive Moving Average (ARMA) otherwise, we have to apply a more complex model called AutoRegressive Moving Average (ARIMA).

## Weekly ADR Plot

The below plot shows the weekly fluctuations in Average Daily Rate over the years. It is evident that the highest Average Daily Rate, indicating peak season, occurs during the summer months of July and August, while the lowest Average Daily Rate indicates off-peak season takes place in the winter month of January and February. In the next section, we will make this plot stationary.

A graph showing a line of a graph

Description automatically generated with medium confidence

## Stationarity of the series

In order to check whether the above plot is stationary or not we need to apply The [Augmented Dickey-Fuller test](https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test) which takes as its null hypothesis that the time series has a unit root - a characteristic of non-stationary time series. Conversely, the alternative hypothesis (under which the null hypothesis is rejected) is that the series is stationary.

* Null Hypothesis (HO): The series is not stationary or has a unit root.
* Alternative hypothesis (HA): The series is stationary with no unit root.

Since the null hypothesis assumes the presence of a unit root, the p-value obtained should be less than a specified significance level, often set at 0.05, to reject this hypothesis. (“ARIMA and SARIMAX Models with Python”)

## Checking for Stationarity

* After performing The [Augmented Dickey-Fuller Test](https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test) based on the values in the ADR feature, it is evident that these values are not stationary (p-value = 0.144171). Therefore, we reject the Null Hypothesis, given that the p-value is greater than the specified significance level of 0.05.
* Consequently, the data was split into training set 80% and test set 20% to evaluate stationarity. Even after applying The [Augmented Dickey-Fuller Test](https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test) to the training set, it remains non-stationarity (p-value = 0.150541).
* As the Time Series Analysis is still not stationary, it needs to be stationarised through differencing. Therefore, we will perform the AutoRegressive Moving Average (ARIMA).   
  We had to set 1 to differentiate the series to obtain a p-value less than the significance as a result we have a p-value of 8.044706e-22, significantly lower than the 0.05 significance level. Therefore, we reject the null hypothesis and consider the series as stationary.

## Time Series Plot of the Data

In this context, ***d*** represents the degree of differencing, indicating the number of times that past values have been subtracted from the data. Therefore, the most appropriate selection for the ARIMA parameter ***d*** is 1.

A green line graph with white text

Description automatically generated

## Series Autocorrelation and Partial Autocorrelation Plots

There seems to be a correlation in the time of the year. The autocorrelations seem to die down fairly after lag 0, then remain constantly lower , and decrease further after lag 25 . There seems to be some repetition: 1 up, 2 down or 1 up, 1 down. There seems to be some seasonality every 3 months, and there are small spikes in repetition.

The partial autocorrelations seem to be small after the first lag, so we decide to fit an ARIMA between 0 and 1. Here, there seems to be fluctuations in seasonality every 8 months, 6 months, 3 months, and so on.

Based on the autocorrelation function, the optimal value for parameter***p*** is 0. However, we will assign a value of 1 to provide an autoregressive component to the model. Regarding the ***q***component, the partial autocorrelation function suggests a value of 1.

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

# One-step-ahead forecasts of the last 10 observations

This step is crucial to evaluate the performance of the forecasting model by comparing the predicted values with the actual values. It is clear that from the plot, after 100 lags, there is very minimal effect on the response variable.

A graph showing a graph

Description automatically generated with medium confidence

## Forecast Errors

We performed the Mean Square Error and Mean Absolute Error to assess forecast errors and evaluate our model.

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values. (Acharya)

From the plot above, there is a gap between predicted and actual values. The R-Squared Score of   
-0.0362 suggests that the model is not capturing the variance effectively and the MAE of 38.2958 implies the model is not accurate in forecasting the actual values.

# Concept and application of Text Analytics

Text analysis is the process by which information is a­utomatically extracted and classified from text data such as, survey responses, emails, support tickets, call centre notes, product reviews and social media posts (Facebook, Instagram, Reddit and so on).

The two most widely use techniques in text analysis are:

* Sentiment Analysis: this technique helps identify the underlying sentiment (positive, neutral, and negative) of text responses.
* Topic detection/categorisation: this technique groups similar themes that are relevant to the business or industry.

Business Applications:

Text analysis is used in several different businesses such as, customer experience (increasing loyalty, preventing churn), employee experience (Employee Attrition, Employee Well Being, Work Life Balance) and Product experience (New Product Launch, product usage)

Another interesting aspect of Text Analytics is topic modeling, which refers to grouping similar concepts or themes in the text responses. This process transforms different topics into a single, understandable structure. For example, in a utility company a customer might say, “The dual tariff is expensive” while another may argue “The dual pricing package is expensive”. Although the words they are using are different (‘tariff’ vs ‘pricing package’), they are both referring to the same topic. (“Text Analysis: The Definitive Guide”)

## Political Social Media Posts Plot

The plot depicts the messages sent by politicians on Twitter and Facebook at five-minute intervals. It is interesting to note that the peak of the messages occurs by 9pm. Afterward, there are small or null fluctuations until 4am when the frequency of the messages starts increasing again, reaching an even higher peak than before. The frequency then decreases by 6am, with a null response until around 1pm.

A graph showing a number of times

Description automatically generated with medium confidence

## LDA (Latent Dirichlet Allocation)

We will underline the top-30 most salient terms using a probabilistic model called Latent Dirichlet Allocation, which is used in natural language processing for topic modelling.

## World Cloud

The world cloud shows the most frequent commented terms on Twitter and Facebook posts by United States politicians. We removed unwanted words/characters through the use of Stop Words and some data exploration, and also removed punctuations, which could otherwise affect our analysis.

A close-up of words

Description automatically generated

## Predicted Topics Results

The predicted topic for each document suggests the main theme or subject that the LDA model has identified within the text data, while the predicted score indicates how strongly the document is associated with a particular theme. The document in Topic 3 has the strongest association with a high predicted score of 94%.

A screenshot of a computer

Description automatically generated

## Sentiment Analysis

The compound score for the most positive tweet is 0.98, indicating a very positive sentiment for this particular Tweet. In contrast, the compound score for the most negative Tweet is -0.92, indicating a very negative sentiment.

# A screenshot of a computer Description automatically generated A screenshot of a computer Description automatically generated

# Conclusions

Both Time Series Analysis and LDA can be used to obtain meaningful insights from raw data. Time Series Analysis is focused on analysing patterns or trends in the data collected over time, while LDA identifies the themes or topics in a large collection of text documents.

Machine Learning Algorithms can automatically identify patterns and trends in Time Series Analysis and classify text documents based on their themes or topics.

# Data Visualisation Techniques

The data was transformed to show key insights related to hotel industry. Personal details were removed from the dataset to ensure data protection regulations, and outliers were also removed. The months were sorted chronologically to present accurate visualisations and results to the viewers.

## Wireframe

The below Wireframe propose the design of the dashboard before implementation. In the next section, we will explain the rationale behind the sections.

A screenshot of a graph

Description automatically generated

## Dashboard Summary

The following Dashboard is a summary of the key insights derived from the exploratory data analysis. All the plots are interactive using Plotly Library.

### Countries the Guests come from

The choropleth map was chosen to represent the Monthly Arrivals by country for both hotels. This map also provides an easy way to visualise the geographic areas (countries or continents) from which the guests come from, using different colours depending on the number of arrivals.

For business and marketing purposes, the Pie Chart shows the top 10 countries where the guest come from. It is evident that people from all over the world are staying in these two hotels, where the largest proportion come from Portugal, followed by the UK and mostly other countries in Europe.

A map of the world

Description automatically generatedA pie chart with numbers and text

Description automatically generated

### Line chart showing trends that change over time across the ADR.

Both line charts are useful to track changes or visualise trends over the months to compare the ADR and Monthly Arrivals for both hotels.

The Monthly Average ADR comparison illustrates that the ADR Resort Hotel is much higher during the summer, followed by a sharp decrease. Afterwards, the ADR increases by 18% in December suggesting higher demand due to Christmas and New Year. In contrast, the ADR City Hotel remains steady during Spring and Autumn where the ADR is significantly higher.

The Monthly Arrivals shows that the City Hotel has more arrivals through the year, with a peak in August, then starts to decrease in the following months. In contrast, the number of arrivals for the Resort Hotel remains steady during Winter and Spring; however, the arrivals are higher in summer than the rest of the year.

A graph with green and purple lines

Description automatically generatedA graph with a line graph and numbers

Description automatically generated with medium confidence

### Target variable across different categories.

The boxplots are useful for visualising the distribution of multiple features against the Target Variable as they are helpful to quickly identify the median and detect outliers.

In this case, the blue colour represents non-canceled bookings, while red colour represents canceled bookings.  
All the features have the same median and the outliers are similar except for Lead Time (the time taken when a guest makes a reservation and the actual arrival date). It is clear that the number of canceled bookings is much higher than non canceled bookings.

A group of graphs with different colored lines

Description automatically generated with medium confidence

The below boxplots clearly show the lowest, the median and the highest ADR per room type for both hotels.

A graph with green and blue squares

Description automatically generated

### Histogram for Market Segment & Customer Type

The histogram is helpful for understanding the distribution of both Market Segment and Customer Type over the months against the Target Variable. The height of the bars represents the data frequency in each month, and the colours were designed to maintain consistency between the plots.  
Both plots are animated for the viewer to interact with over the months.

A screenshot of a graph

Description automatically generated A graph of a bar chart

Description automatically generated with medium confidence

# References

## Machine Learning for Business

“Political Social Media Posts.” Www.kaggle.com, [www.kaggle.com/datasets/crowdflower/political-social-media-posts](http://www.kaggle.com/datasets/crowdflower/political-social-media-posts).

“Hotel Booking Demand.” *Www.kaggle.com*, [www.kaggle.com/datasets/jessemostipak/hotel-booking-demand](http://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).

All, Moez. “Time Series Forecasting Tutorial.” Www.datacamp.com, 1 Feb. 2020, [www.datacamp.com/tutorial/tutorial-time-series-forecasting](http://www.datacamp.com/tutorial/tutorial-time-series-forecasting).

Jason Brownlee. “A Gentle Introduction to the Box-Jenkins Method for Time Series Forecasting.” *Machine Learning Mastery*, 21 Feb. 2017, machinelearningmastery.com/gentle-introduction-box-jenkins-method-time-series-forecasting/.

“Using Machine Learning for Time Series Forecasting Project.” CodeIT, codeit.us/blog/machine-learning-time-series-forecasting.

“ARIMA and SARIMAX Models with Python.” Cienciadedatos.net, cienciadedatos.net/documentos/py51-arima-sarimax-models-python.html.

<https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/The_Box-Jenkins_Method.pdf>

Kumar, Vijay. “Statistical Tests to Check Stationarity in Time Series.” Analytics Vidhya, 16 June 2021, [www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/](http://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/).

Acharya, Shwetha. “What Are RMSE and MAE?” Medium, 14 May 2021, towardsdatascience.com/what-are-rmse-and-mae-e405ce230383#:~:text=Root%20Mean%20Squared%20Error%20(RMSE.

Gramener. “Text Analysis: An Overview Guide on Concept, Techniques & Use Cases.” Medium, 27 Dec. 2021, gramener.medium.com/text-analysis-overview-339a529d4910.

“Text Analysis: The Definitive Guide.” Qualtrics, www.qualtrics.com/uk/experience-management/research/text-analysis/.

## Data Visualisation Techniques

“Hotel Booking Demand.” *Www.kaggle.com*, [www.kaggle.com/datasets/jessemostipak/hotel-booking-demand](http://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).

Rategain. “Impact of Seasonality on Revenue: Strategies to Navigate through the Peak and Off-Peak Seasons.” *RateGain*, 8 June 2023, rategain.com/blog/the-impact-of-seasonality-on-hotel-revenue/#:~:text=Seasonality%20is%20a%20crucial%20factor. Accessed 12 Nov. 2023.

“Sort a Pandas Dataframe Series by Month Name.” Stack Overflow, stackoverflow.com/questions/48042915/sort-a-pandas-dataframe-series-by-month-name. Accessed 12 Nov. 2023.

“Choropleth Maps.” Plotly.com, plotly.com/python/choropleth-maps/.

“Pie Charts.” Plotly.com, plotly.com/python/pie-charts/.

plotlygraphs. “Line Charts.” Plotly.com, 3 July 2019, plotly.com/python/line-charts/.

plotlygraphs. “Histograms.” Plotly.com, 3 July 2019, plotly.com/python/histograms/.

“Box Plots.” Plotly.com, plotly.com/python/box-plots/.

“Plotly.subplots.make\_subplots — 5.10.0 Documentation.” Plotly.com, plotly.com/python-api-reference/generated/plotly.subplots.make\_subplots.html.

plotlygraphs. “Subplots.” Plotly.com, 3 July 2019, plotly.com/python/subplots/.