**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | MSc in Data Analytics |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | David McQuinn  David Gonzalez  Marina Iantorno  Muhammad Iqbal |
| **Student Full Name:** | Ana Carla Ferreira de Oliveira |
| **Student Number:** | sba23438 |
| **Assessment Due Date:** | 05th January 2024 |
| **Date of Submission:** | 05th January 2024 |

**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. |

**CA2 | MSC in Data Analytics**

***1º S E M E S T E R***

*Ireland’s Aviation Trends in Focus Alongside European Counterparts*

***https://github.com/Oliveiranac/CA2\_MScDataAnalytics.git***

**S t u d e n t:**

|  |  |
| --- | --- |
| **sba43** | Ana Carla Ferreira de Oliveira |



**L e c t u r e r s:**

|  |  |
| --- | --- |
| **David McQuaid** | Data Preparation & Visualization |
| **David Gonzalez** | Python Programming |
| **Marina Iantorno** | Statistics |
| **Muhammed Iqbal** | Machine Learning |

**J A N U A R Y 2 0 2 4**

Contents

[1. INTRODUCTION 4](#_Toc155549896)

[1.1 OBJECTIVES 4](#_Toc155549897)

[2. BUSINESS UNDERSTANDING 4](#_Toc155549898)

[2.1 RESEARCH QUESTIONS 4](#_Toc155549899)

[3. DATA UNDERSTANDING 6](#_Toc155549900)

[3.1 DATA 1 6](#_Toc155549901)

[2.1.1 Data Source 6](#_Toc155549902)

[2.1.2 Data Description 7](#_Toc155549903)

[2.1.3 Data Usage 8](#_Toc155549904)

[3.2 DATA 2 8](#_Toc155549905)

[2.1.1 Data Source 8](#_Toc155549906)

[2.1.2 Data Description 8](#_Toc155549907)

[2.1.3 Data Usage 9](#_Toc155549908)

[1. MODELLING 9](#_Toc155549909)

[2. CONCLUSION 9](#_Toc155549910)

[3. REFERENCES 9](#_Toc155549911)

List of Figures

# **INTRODUCTION**

Air travel is an integral component of global connectivity, linking nations and fostering economic and cultural exchange. Understanding the patterns of passenger movement at airports is crucial for various stakeholders, from aviation authorities to travel agencies.

This project embarks on a comprehensive analysis of airport transport in Europe, focusing on Ireland and three comparable European countries with proportional population sizes. The objective is to unravel the trends and relationships between key indicators across these nations. The datasets encompass a diverse range, including passenger statistics, freight movements, commercial flight details, and even the sentiments expressed by passengers during their journeys. The temporal scope spans from 2020 to 2023, providing a robust foundation for longitudinal insights.

The research employs a multistage process, including data collection, pre-processing, visualization, statistical analysis, machine learning, and even sentimental analysis. Python programming language is applied in the analysis, facilitating a robust and scalable examination of the datasets.

The significance of this project extends beyond the world of academics, aiming to illuminate industry stakeholders, policymakers, and the general public with practical insights. As we explore the data, our goal is to find useful patterns that can help make smart decisions. This information can be valuable for those dealing with the ever-changing world of the aviation industry.

## OBJECTIVES

The main goal is to conduct a comprehensive analysis of Ireland's air transport, comparing it with other European countries for the period 2020 to 2023. The study will specifically delve into passengers, freight, and commercial flights, contributing to a clearer picture of air travel trends in the selected countries.

# **BUSINESS UNDERSTANDING**

In this project, several research questions are taken into consideration to help in gaining insights into the data.

## RESEARCH QUESTIONS

1. Are there differences in mean values for airports in Ireland?

* Hypothesis 1:
* Null hypothesis(H0): there is no significant difference in the mean values of VALUE across different Airports in Ireland.
* Alternative hypothesis(H1): there is a significant difference in the mean values of VALUE across different Airports in Ireland.

1. Is there a difference between the distribution of mean VALUE between Ireland and Denmark?

* Hypothesis 2:
* H0: There is no significant difference in the distributions of VALUE between Ireland and Denmark.
* H1: There is a significant difference in the distributions of VALUE between Ireland and Denmark.

1. Does the Year and VALUE variables have a correlation between each other?

* Hypothesis 3:
* H0: There is no correlation between VALUE and Year (correlation coefficient equals 0).
* H1: There is a significant correlation between VALUE and Year (correlation coefficient is not equal to 0).

1. Is there a difference in distribution across different months of the year?

* Hypothesis 4:
* H0: There is no significant difference in the distribution of VALUE across different months.
* H1: There is a significant difference in the distribution of VALUE across different months.

1. Are mean values of scheduled flights different from unscheduled flights?

* Hypothesis 5:
* H0: There is no significant difference in the mean values of VALUE between Scheduled and Unscheduled flights.
* H1: There is a significant difference in the mean values of VALUE between Scheduled and Unscheduled flights.

1. Is there a correlation between flight type and direction?

* Hypothesis 6:
* H0: There is no correlation between direction and flight type.
* H1: Flight\_Type and Direction have a significant correlation.

1. Does the distribution of VALUE for scheduled flights alter according to arrival and departure?

* Hypothesis 7:
* H0: For scheduled flights, there is no discernible change in the VALUE distribution between the points of arrival and departure.
* H1: For scheduled flights, there is no discernible difference in the distribution of VALUE between the points of arrival and departure.

# **DATA UNDERSTANDING**

This section offers a thorough look at the data used to analyze air travel in Ireland and other European nations. The study leverages insights from two datasets, each providing a unique perspective on the dynamics of aviation.

The combination of these datasets enables a holistic examination of the aviation landscape, blending quantitative metrics with qualitative insights. The ensuing exploratory analysis aims to uncover patterns, trends, and relationships within the data, laying the foundation for subsequent statistical and machine-learning analyses.

## DATA 1

TAM07.20231217131259.CSV: This dataset provides a thorough overview of Ireland's aviation trends, enabling in-depth examination and research of passenger, freight, and commercial flight patterns throughout time.

### 3.1.1 Data Source

Data.gov is the source of Ireland's passenger, freight, and commercial flight data. The Pandas package in Python can be used to access the dataset, which offers insightful statistics about air travel.

### 3.1.2 Data Description

With 29,160 entries and 9 columns, the collection contains data on several topics. The following are the columns and the corresponding data types for them:

* Statistical\_Label: This column represents the type of statistic being recorded. Divided into 'Freight,' 'Passengers,' and 'Commercial Flights.'
* Month: The data is recorded per month.
* Airports in Ireland: A variety of airports are located around Ireland, such as 'Cork, 'Dublin', 'Kerry', 'Knock', and 'Shannon.'
* Country: The nations taking part in the transit.
* Direction: 'All directions', 'Arrival', or 'Departure.'

**Note:**

|  |  |
| --- | --- |
| Arrivals: | This direction indicates statistics related to passengers arriving at the specified airports during the given period. |
| Departures: | Indicates statistics related to passengers departing from the specified airports during the given period. |
| All directions: | There could be a category for transits, connecting flights, or other specific aspects of air travel. |
| Understanding the "Direction" column is crucial for analyzing air travel patterns and understanding the dynamics of passenger movements at airports. | |

* Flight\_Type: Divided into three categories: scheduled, unscheduled, and all flights.

**Note:**

|  |  |
| --- | --- |
| Scheduled Flights: | These are regular flights that operate according to a published schedule. Airlines plan and advertise these flights well in advance. |
| Unscheduled Flights: | Also known as charter flights, these are not part of a regular schedule and are often booked for a specific purpose or by a particular group of travelers. |
| Understanding the Flight Type can provide insights into the nature of air travel activities. Scheduled flights are usually associated with routine passenger travel, while unscheduled flights might involve special events, group travel, or cargo transport based on specific needs. | |

* UNIT: This column represents the unit of measurement for the passenger count. In this dataset, it is labelled as “Thousand”, indicating that the passenger counts are in thousands.
* VALUE: The number corresponding to the particular category, measured in thousands.
* Year: The year the data was first recorded.
* Float values that indicate different metrics pertaining to passengers, freight, and commercial flights are found in the 'VALUE' column.

### 3.1.3 Data Usage

This dataset holds significance as it carries the potential to offer valuable information, benefiting a diverse range of stakeholders including the general public, lawmakers, and business experts.

## DATA 2

It utilizes the Twitter Airline Sentiment dataset to perform an in-depth analysis of sentiments directed towards diverse air companies. By employing sentiment analysis, our goal is to reveal patterns, sentiments, and trends that provide insights into the ever-evolving landscape of public opinion within the airline industry.

### 3.2.1 Data Source

The Twitter Airline Sentiment dataset, available on Kaggle, provides a diverse collection of tweets related to airline experiences, making it a valuable resource for conducting sentiment analysis research.

### 3.2.2 Data Description

With 14,485 entries and 15 columns, this dataset delves into various aspects of tweets associated with airline sentiments. Key columns include:

* tweet\_id: A distinct number assigned to every tweet.
* airline\_sentiment: Labels for sentiment that can be classified as "positive," "neutral," or "negative."
* airline\_sentiment\_confidence: The sentiment label's confidence score.
* negativereason: The cause of the unfavorable attitude.
* airline: The airline that the tweet is connected to.
* retweet\_count: The tweet's total number of retweets.
* text: The tweet's contents.
* tweet\_created: The creation time stamp of a tweet.
* user\_timezone: The user's time zone.

### 3.2.3 Data Usage

This dataset is a useful tool for sentiment analysis, enabling an investigation of public sentiment towards different airlines on Twitter. The sentiment labels and associated features provide comprehensive insights into user perspectives and opinions regarding their airline experiences.

# **DATA PREPARATION**

The data preparation process for this analysis adheres to Tuftes' principles of clarity, simplicity, and accuracy, ensuring that the information presented is both insightful and easily comprehensible.

## DATA IMPORT OF CSV AND SQLITE FILES

# **MODELLING**

# **CONCLUSION**