**MSc in Data Analytics (SB+) - Sept 2023 - 2023 - YR1**

Author: Jose Maria Rico Leal

e-mail: sba23021@student.cct.ie

Student ID: sba23021

GitHub Link:

Irish transport sector

## **Abstract**

*Empty*

[**Abstract** 1](#_Toc152942890)

[**1.** **Introduction** 3](#_Toc152942891)

[**2.** **Programming** 4](#_Toc152942892)

[2.1. Programming 4](#_Toc152942893)

[2.2. Data structures 4](#_Toc152942894)

[2.3. Documentation 4](#_Toc152942895)

[2.4. Testing and optimisation 4](#_Toc152942896)

[2.5. Data manipulation 4](#_Toc152942897)

[**3.** **Statistical Analysis** 4](#_Toc152942898)

[3.1. Descriptive statistics 4](#_Toc152942899)

[*3.1.1.* *Dataset for Confidence interval* 4](#_Toc152942900)

[*3.1.2.* *Dataset for Hypothesis Test Two Populations* 4](#_Toc152942901)

[*3.1.3.* *Dataset for Anova one-way* 4](#_Toc152942902)

[*3.1.4.* *Dataset for Chi-squared test* 4](#_Toc152942903)

[*3.1.5.* *Dataset for Kruskal-Wallis* 4](#_Toc152942904)

[*3.1.6.* *Dataset for U-Mann Whitney* 4](#_Toc152942905)

[3.2. Confidence interval 4](#_Toc152942906)

[3.3. Inferential statistics 5](#_Toc152942907)

[*3.3.1.* *Parametric* 5](#_Toc152942908)

[*3.3.1.1.* *T-test* 5](#_Toc152942909)

[*3.3.1.2.* *Anova* 5](#_Toc152942910)

[*3.3.1.3.* *Chi-squared test* 5](#_Toc152942911)

[*3.3.2.* *Non-parametric* 5](#_Toc152942912)

[*3.3.2.1.* *Kruskal-Wallis* 5](#_Toc152942913)

[*3.3.2.2.* *U-Mann Whitman* 5](#_Toc152942914)

[3.4. Outcome and challenges faced 5](#_Toc152942915)

[**4.** **ML** 5](#_Toc152942916)

[4.1. ML supervised learning 5](#_Toc152942917)

[4.2. Sentiment analysis 5](#_Toc152942918)

[4.3. Comparing Supervised, Unsupervised and semi-supervised ML models. 5](#_Toc152942919)

[4.4. Table and conclusions 5](#_Toc152942920)

[**5.** **Data Preparation and Visualisation** 5](#_Toc152942921)

[5.1. Data acquistion 5](#_Toc152942922)

[5.2. EDA methodology 5](#_Toc152942923)

[5.3. Visualisations 5](#_Toc152942924)

[5.4. Dashboard 5](#_Toc152942925)

[**6.** **Conclusion** 5](#_Toc152942926)

[**7.** **References** 5](#_Toc152942927)

# **Introduction**

# **Programming**

# Programming

# Data structures

# Documentation

# Testing and optimisation

# Data manipulation

# **Statistical Analysis**

# Descriptive Statistics

# *Dataset for Confidence Interval*

Dataset used is *“tii03-passenger-journeys-by-luas”.* We want to estimate the range of potential values for the parameter *“LUAS average passenger number”.*Here we have the descriptive statistics total LUAS passenger numbers (green and red lines) for years 2019, 2020, 2021 and 2022:

A table with numbers and a number on it

Description automatically generated

We will be looking at the mean values to stablish the confidence intervals.  
Interesting to note that both lines are quite balanced in terms of usage:

A group of pie charts with numbers

Description automatically generated

In section *3.2. Confidence Interval,* an in-depth analysis will be conducted, for this dataset.

# *Dataset for Hypothesis Test Two Populations*

In this section we are comparing Ireland with some of the European countries, we formulate hypothesis to assess if there are statistic significance differences for the *“percentage average passenger-kilometres based on type of transport (Bus, Car and Train”.* As we will be using Ireland mean against other countries mean, this plot will be helping us constructing the hypothesis:

A group of colorful bars

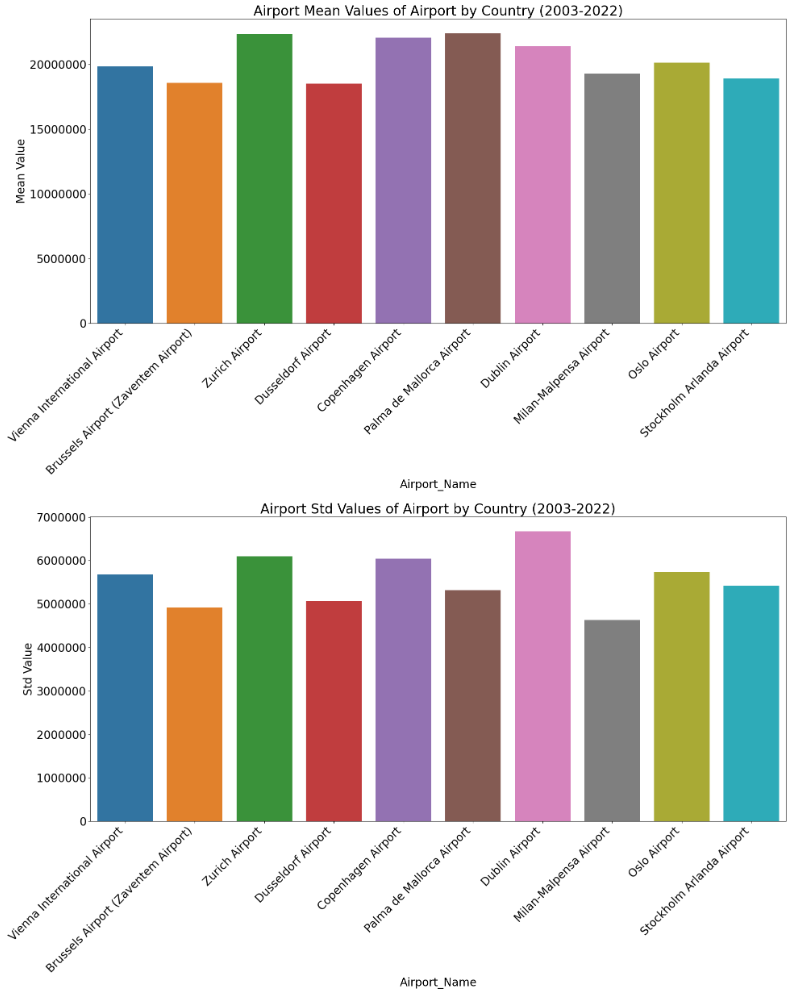
Description automatically generated with medium confidence

Dataset used *“Modal split of inland passenger transport”* Eurostat code: *“TRAN\_HV\_PSMOD”.*

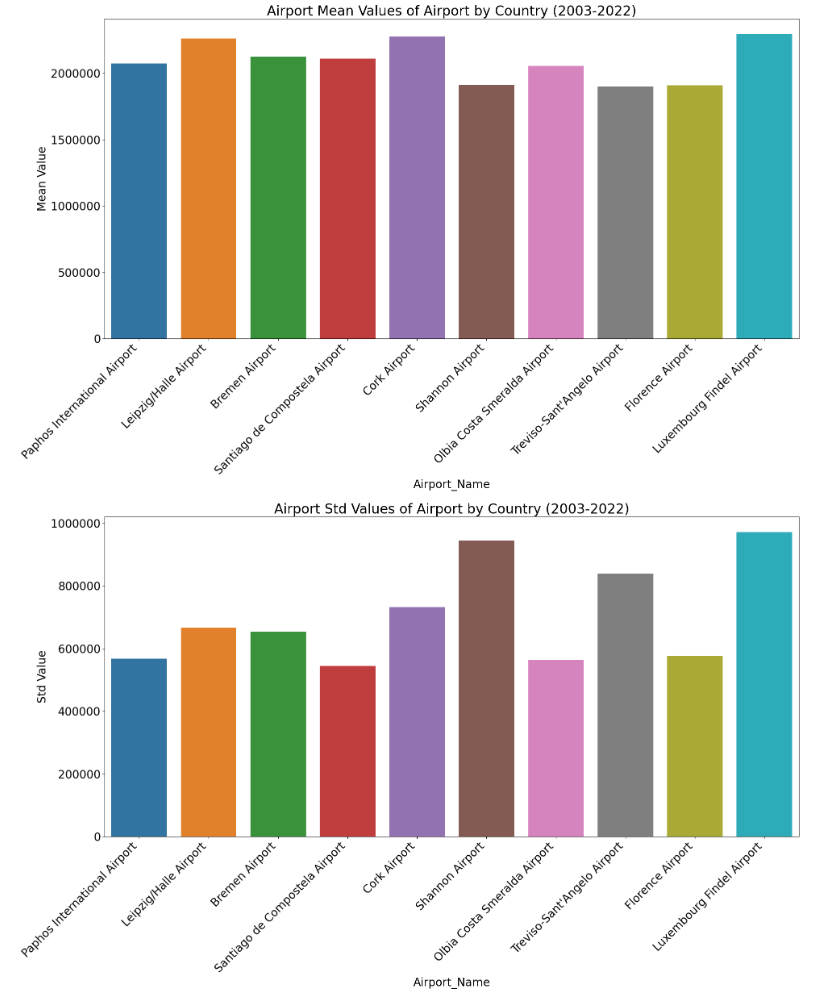
# *Dataset for ANOVA One-Way*

Having Irish airports as a reference, ANOVA will be carried out test whether there are any statistically significant differences in the means with other European airports.  
Dataset used *“Air passenger transport by main airports in each reporting country”* Eurostat code *“AVIA\_PAOA”.*

First scenario Dublin Airport, below a graph of subset airports selected for this case:



Second scenario Shannon Airport, below a graph of subset airports selected for this case:



# *Dataset for Chi-Squared Test*

This test will be performed to examine the association between *“Motor\_energy\_type”* categorical variables for Ireland and Austria. Dataset used *“New passenger cars by type of motor energy”*, Eurostat code *“road\_eqr\_carpda”.*

This model is very sensitive to the frequency of the variables, I stated two cases, one where we reject H0 and the second one where I manually changed values to accept H0. I am going to show how the categorical variable number will look like for each case, more analysis to follow in *“3.3.1.3. Chi-squared test”.*

Scenario 1 rejecting H0:

A graph of a number of cars

Description automatically generated with medium confidence

Scenario 2 accepting H0:

A graph of different types of cars

Description automatically generated

# *Dataset for Kruskal-Wallis*

For this test I used the same dataset as ANOVA, some of the airports were violating the assumptions of normality required for ANOVA, good think is Kruskal-Wallis do not require normality to perform the test. I am going to have two scenarios one to accept H0 and one to reject H0.

Scenario 1 accepting H0:

A screenshot of a graph

Description automatically generated

Scenario 2 rejecting H0:

A screenshot of a graph

Description automatically generated

# *Dataset for U-Mann Whitney*

New dataset for this test, *“Passengers transported (Railway transport)”,* Eurostat code *“rail\_pa\_total”*. I choose this dataset because data was not following normal distribution and that helped me to demonstrate that this can handle non normally distributed data.

Scenario 1, accepting H0:

A number on a white background

Description automatically generated

Scenario 1, rejecting H0:

A number on a white background

Description automatically generated

# Confidence Interval

# Inferential statistics

# *Parametric*

# *T-test*

# *Anova*

# *Chi-squared test*

# *Non-parametric*

# *Kruskal-Wallis*

# *U-Mann Whitman*

# Outcome and challenges faced

# **ML**

# ML supervised learning

# Sentiment analysis

# Comparing Supervised, Unsupervised and semi-supervised ML models.

# Table and conclusions

# **Data Preparation and Visualisation**

# Data acquistion

# EDA methodology

# Visualisations

# Dashboard

# **Conclusion**

# **References**

Statistics: