**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: https://github.com/JoseRicoCct/CA2\_Integrated\_Assesment\_MSc\_Data\_Analytics\_CCT\_Semester\_1.git

Irish transport sector

## **Abstract**

*Empty*

[**Abstract** 1](#_Toc152999956)

[**1.** **Introduction** 3](#_Toc152999957)

[**2.** **Programming** 4](#_Toc152999958)

[2.1. Programming 4](#_Toc152999959)

[2.2. Data structures 4](#_Toc152999960)

[2.3. Documentation 4](#_Toc152999961)

[2.4. Testing and optimisation 4](#_Toc152999962)

[2.5. Data manipulation 4](#_Toc152999963)

[**3.** **Statistical Analysis** 4](#_Toc152999964)

[3.1. Descriptive Statistics 4](#_Toc152999965)

[*3.1.1.* *Dataset for Confidence Interval* 4](#_Toc152999966)

[*3.1.2.* *Dataset for Hypothesis Test Two Populations* 5](#_Toc152999967)

[*3.1.3.* *Dataset for ANOVA One-Way* 7](#_Toc152999968)

[*3.1.4.* *Dataset for Chi-Squared Test* 8](#_Toc152999969)

[*3.1.5.* *Dataset for Kruskal-Wallis* 9](#_Toc152999970)

[*3.1.6.* *Dataset for U-Mann Whitney* 11](#_Toc152999971)

[3.2. Confidence Interval 12](#_Toc152999972)

[3.3. Inferential statistics 13](#_Toc152999973)

[*3.3.1.* *Parametric* 13](#_Toc152999974)

[*3.3.1.1.* *T-test* 13](#_Toc152999975)

[*3.3.1.2.* *Anova* 13](#_Toc152999976)

[*3.3.1.3.* *Chi-squared test* 13](#_Toc152999977)

[*3.3.2.* *Non-parametric* 13](#_Toc152999978)

[*3.3.2.1.* *Kruskal-Wallis* 13](#_Toc152999979)

[*3.3.2.2.* *U-Mann Whitman* 13](#_Toc152999980)

[3.4. Further research and challenges faced 13](#_Toc152999981)

[**4.** **ML** 13](#_Toc152999982)

[4.1. ML supervised learning 13](#_Toc152999983)

[4.2. Sentiment analysis 13](#_Toc152999984)

[4.3. Comparing Supervised, Unsupervised and semi-supervised ML models. 13](#_Toc152999985)

[4.4. Table and conclusions 13](#_Toc152999986)

[**5.** **Data Preparation and Visualisation** 13](#_Toc152999987)

[5.1. Data acquistion 13](#_Toc152999988)

[5.2. EDA methodology 13](#_Toc152999989)

[5.3. Visualisations 13](#_Toc152999990)

[5.4. Dashboard 13](#_Toc152999991)

[**6.** **Conclusion** 13](#_Toc152999992)

[**7.** **References** 13](#_Toc152999993)

# **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

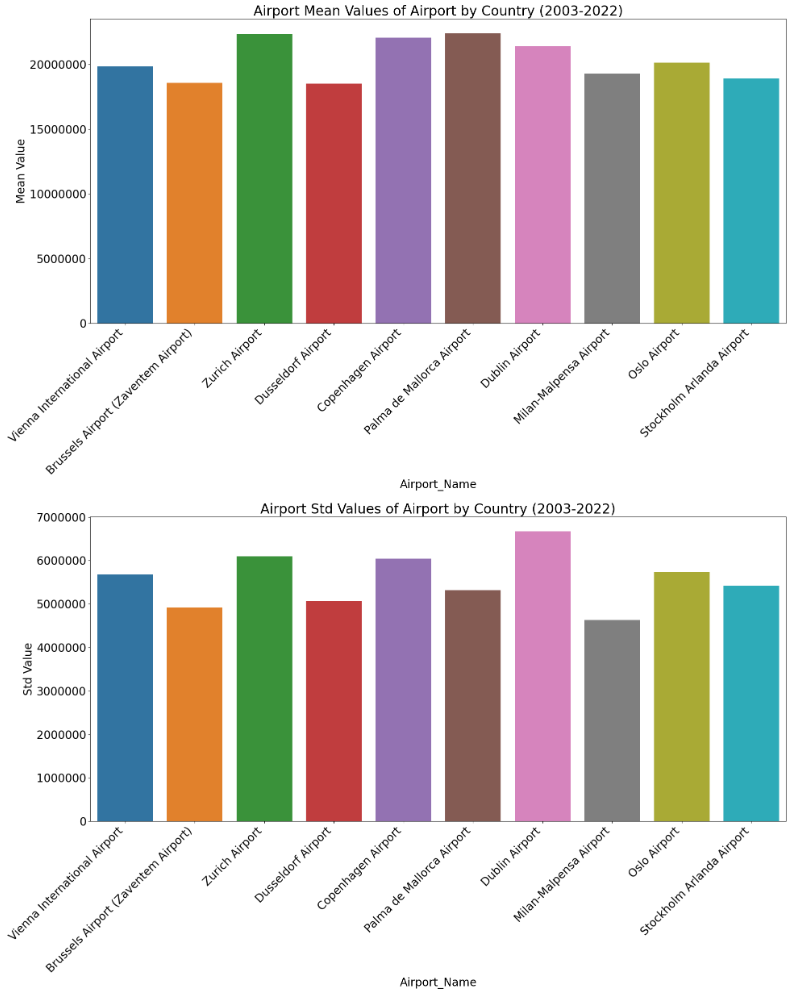
Description automatically generated with medium confidence

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

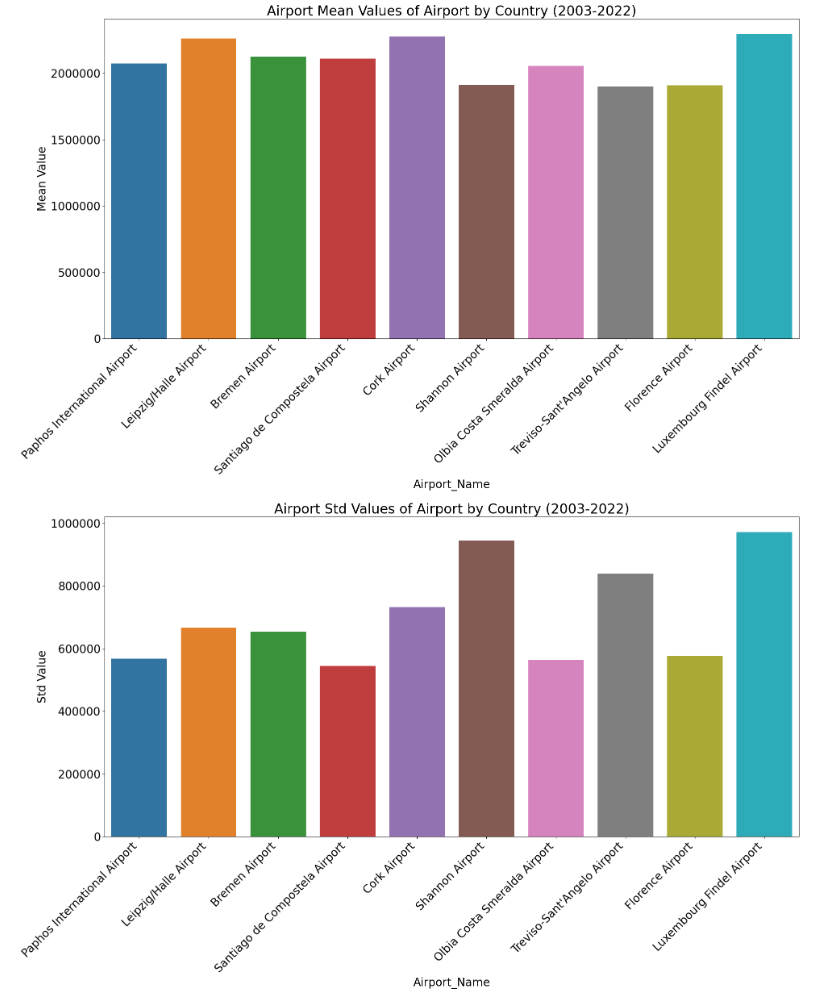
# *Dataset for ANOVA One-Way*

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

The task to accomplish is, to find out the weekly LUAS average for the total number of passengers in years 2019, 2020, 2021 and 2022. Please note we have both LUAS lines in scope, red and green. We will use 90% confidence, here are the results:

A table with numbers and a few black text

Description automatically generated

E.g. At a 90% confidence level, for 2019 the weekly LUAS number of passengers average is between 453K and 472K. Subsequently the same formulation for the rest of the years.

Plotting the confidence intervals:

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

Having performed confidence intervals, the next natural step is to verify if the weekly average is the same for red and green LUAS lines. We will use a t test:

A table with numbers and letters

Description automatically generated

Hypothesis:

H0: μ green line = μ red line.  
H1: μ green line != μ red line.

Results:

A group of graphs with numbers and symbols

Description automatically generated with medium confidence

A table with numbers and text

Description automatically generated

At a 5% significance level we accept the Null Hypothesis, there not enough evidence to state that weekly mean values for LUAS green line is different from the red one.

# Inferential statistics

# *Parametric*

# *T-test Two Populations*

To perform this test, we will compare Ireland with three different countries also having three different ways of transportation method. First scenario Ireland against Slovenia for car, second Ireland Denmark for bus and third Ireland and Slovenia for train.

Hypothesis:

H0: μ Ireland = μ EU Country.

H1: μ Ireland != μ EU Country.

Results:

A graph of a normal distribution

Description automatically generated

As p\_value > alpha, then we accept H0, there is not enough evidence to state that there is a significant difference between the percentage average on passenger-kilometres for Vehicle BUS in Ireland and Slovenia.

A graph of a normal distribution

Description automatically generated

As p\_value < alpha, then we reject H0, there is not enough evidence to state that there is a significant difference between the percentage average on passenger-kilometres for Vehicle Car in Ireland and Denmark.

A graph of a normal distribution

Description automatically generated

As p\_value < alpha, then we reject H0, there is not enough evidence to state that there is a significant difference between the percentage average on passenger-kilometres for Vehicle Train in Ireland and Slovenia.

A screenshot of a calculator

Description automatically generated

# *ANOVA One-Way*

First scenario Dublin, Zurich and Copenhagen airports, we want to verify if yearly average passenger numbers for the period 2003 and 2022 are the same or not.

To perform ANOVA we need to have Shapiro-Wilk and Levene tests with a p-value greater than 5% alpha:

A screenshot of a test results

Description automatically generated

Now we can perform ANOVA:

H0: μ IE\_EIDW = μ CH\_LSZH = μ DK\_EKCH.

H1: there are at least 2 μ that are different one to another.

Result p-value = 0.889

A graph of a flight

Description automatically generated

There is no reason to reject the Null Hypothesis; therefore, we can state that with a 5% alpha the mean of annual passengers carried (2003-2022) for Dublin, Zurich and Copenhagen airports is the same.

Second scenario Shannon, Billund and Treviso airports, we want to verify if yearly average passenger numbers for the period 2003 and 2022 are the same or not.

Shapiro-Wilk and Levene tests:

A screenshot of a test

Description automatically generated

ANOVA hypothesis:

H0: μ IE\_EINN= μ IT\_LIPH= μ DK\_EKBI.

H1: there are at least 2 μ that are different one to another.

Result p-value = 0.04

A graph of a flight

Description automatically generated

We fail to accept the Null Hypothesis; therefore we can state that with a 5% alpha the mean of annual passengers carried (2003-2022) for Shannon, Treviso-Sant'Angelo and Billund airports is different.

# *Chi-squared test*

First Scenario:

Hypothesis:

H0: There is no significant difference between the observed and expected frequencies. Ireland and Austria are independent and there is no association or relationship between them.

A screenshot of a number of numbers

Description automatically generated

H1: There is a significant difference between the observed and expected frequencies. Ireland and Austria are not independent and there is association or relationship between them.

Result p-value = 1.64e-14, we fail to accept H0.

A graph of a car

Description automatically generated

Second scenario, I manually changed values for Austria to approximate frequencies to Ireland, this model is very sensitive to big differences between categorical variables:

A screenshot of a number

Description automatically generated

After applying Chi-Square, we have a p-value of 0.59, we can accept H0:

A graph of a car

Description automatically generated

# *Non-parametric*

# *Kruskal-Wallis*

Scenario 1:

We don’t have normality for our samples:

A screenshot of a computer

Description automatically generated

Hypothesis:

H0: μ IE\_EIKN= μ FR\_LFBP= μ SE\_ESGP.

H1: there are at least 2 μ that are different one to another.

Result, p-value = 0.23511.

A graph of a line

Description automatically generated

We accept H0, the mean of annual passengers carried (2003-2022) for Ireland West Knock, Pau Pyrenees and Goteborg airports is the same.

Scenario 2:

We don’t have normality for our samples:

A screenshot of a computer

Description automatically generated

Hypothesis:

H0: μ IE\_EIKY= μ DE\_EDSB= μ PL\_EPRZ.

H1: there are at least 2 μ that are different one to another.

Result, p-value = 8.48e-08

A graph of a passenger carrier

Description automatically generated

We fail to accept H0, the mean of annual passengers carried (2003-2022) for Kerry, Karlsruhe/Baden and Rzeszow-Jasionka airports is different.

# *U-Mann Whitman*

Scenario 1:

Checking normality:

**A white rectangular object with black text

Description automatically generated**

Hypothesis:

H0: μ Ireland = μ Croatia.

H1: μ Ireland != μ Croatia.

Result p-value = 0.3068

A graph of a train passenger

Description automatically generated

We accept H0 as p-value is greater than alpha, there is no difference between the average number of train passengers between Ireland and Croatia.

Scenario 2:

Checking normality:

A white background with black text

Description automatically generated

Hypothesis:

H0: μ Ireland = μ Slovakia.

H1: μ Ireland != μ Slovakia.

Result p-value = 3.45e-06

A graph of a train passenger

Description automatically generated

We reject H0 as p-value is lower than alpha, there is a significant difference between the average number of train passengers between Ireland and Slovakia.

# Further research 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: