**MSc in Data Analytics (SB+) - Sept 2023 - 2024 - 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

Analysis of the Irish Transport Sector: A Comparative Study with EU Countries

## **Abstract**

*This paper compares some key areas of the Republic of Ireland transportation sector with some European Countries. The study employs, statistical analysis, machine learning and data visualisation. We completed the research through programming, testing, optimization, and sentiment analysis. Results highlight the sector's performance, challenges faced, and offer insights for future research*. *The combination of rigorous statistical techniques and machine learning enhances the understanding of Ireland's transport landscape, contributing valuable perspectives for policy and industry stakeholders.*

[**Abstract** 1](#_Toc155548332)

[**1.** **Introduction** 3](#_Toc155548333)

[**2.** **Programming** 5](#_Toc155548334)

[2.1. Programming 5](#_Toc155548335)

[2.2. Data Structures 5](#_Toc155548336)

[2.3. Documentation 8](#_Toc155548337)

[2.4. Testing and Optimisation 8](#_Toc155548338)

[2.4.1. Testing 8](#_Toc155548339)

[2.4.2. Optimization 11](#_Toc155548340)

[2.5. Data Manipulation 12](#_Toc155548341)

[**3.** **Statistical Analysis** 14](#_Toc155548342)

[3.1. Descriptive Statistics 14](#_Toc155548343)

[*3.1.1.* *Dataset for Confidence Interval* 14](#_Toc155548344)

[*3.1.2.* *Dataset for Hypothesis Test Two Populations* 15](#_Toc155548345)

[*3.1.3.* *Dataset for ANOVA One-Way* 17](#_Toc155548346)

[*3.1.4.* *Dataset for Chi-Squared Test* 19](#_Toc155548347)

[*3.1.5.* *Dataset for Kruskal-Wallis* 20](#_Toc155548348)

[*3.1.6.* *Dataset for U-Mann Whitney* 22](#_Toc155548349)

[3.2. Confidence Interval 23](#_Toc155548350)

[3.3. Inferential Statistics 25](#_Toc155548351)

[*3.3.1.* *Parametric* 25](#_Toc155548352)

[*3.3.1.1.* *T-test Two Populations* 25](#_Toc155548353)

[*3.3.1.2.* *ANOVA One-Way* 28](#_Toc155548354)

[*3.3.1.3.* *Chi-squared test* 30](#_Toc155548355)

[*3.3.2.* *Non-parametric* 32](#_Toc155548356)

[*3.3.2.1.* *Kruskal-Wallis* 32](#_Toc155548357)

[*3.3.2.2.* *U-Mann Whitman* 35](#_Toc155548358)

[3.4. Further Research and Challenges Faced 36](#_Toc155548359)

[**4.** **ML** 38](#_Toc155548360)

[4.1. Supervised Learning. 38](#_Toc155548361)

[4.2. Cross Validation and Feature Reduction. 40](#_Toc155548362)

[4.3. Unsupervised Learning. 41](#_Toc155548363)

[4.4. Sentiment Analysis. 41](#_Toc155548364)

[4.5. Table and Conclusions 44](#_Toc155548365)

[*4.5.1.* *Supervised Learning* 44](#_Toc155548366)

[*4.5.2.* *Cross Validation and Feature Reduction* 47](#_Toc155548367)

[*4.5.3.* *Unsupervised Learning* 49](#_Toc155548368)

[**5.** **Data Preparation and Visualisation** 54](#_Toc155548369)

[5.1. Data acquisition 54](#_Toc155548370)

[5.2. EDA methodology 54](#_Toc155548371)

[5.3. EDA ML implementation 55](#_Toc155548372)

[5.4. Dashboard 59](#_Toc155548373)

[**6.** **Conclusion** 62](#_Toc155548374)

[**7.** **References** 63](#_Toc155548375)

[**8.** **Annex** 64](#_Toc155548376)

[8.1. Programming 64](#_Toc155548377)

[*8.1.1.* *MySQL Setup* 64](#_Toc155548378)

[*8.1.2.* *Testing and Optimisation* 71](#_Toc155548379)

[8.2. Statistics 74](#_Toc155548380)

[8.3. Data Visualisation 88](#_Toc155548381)

[*8.3.1.* *EDA Methodology* 88](#_Toc155548382)

[*8.3.2.* *EDA ML implementation* 93](#_Toc155548383)

[*8.3.3.* *Dashboard* 103](#_Toc155548384)

# **Introduction**

In this assignment, we deep dive into some aspects of the Irish transport sector, conducting a multifaceted analysis that extends to various key dimensions. With a specific focus on programming, statistical analysis, machine learning, and data visualization, our examination goes beyond the surface, comparing Ireland's transport landscape with selected EU nations. By employing a diverse array of methodologies, we aim to provide clarity in some of the aspects of the sector's strengths, challenges, and opportunities for improvement. This introduction sets the stage for a thorough exploration, offering valuable insights into the complex interplay of factors shaping the Irish transport sector within the broader European context.

# **Programming**

# Programming

I organized the project into five Jupyter Notebooks:  
*CA2\_ML\_Code.ipynb*, *CA2\_Programming\_Code.ipynb*, *CA2\_Statistics\_Code.ipynb* and *CA2\_Data\_Visualisation.ipynb*. The project is programmatically explored using Python, emphasizing modularity and narrative-style programming.

# Data Structures

For this project I gathered and process data from:

CSV file:

A screenshot of a computer

Description automatically generated

Figure 1

Web API in CSV format:

A screenshot of a computer screen

Description automatically generated

Figure 2

Eurostat module:

A screenshot of a computer

Description automatically generated

Figure 3

Also, I explored gathering data from a:

Web API in JSON format:



Figure 4

Entire parsing process is documented in *“CA2\_Programming\_Code\_sba23021.ipynb”* from cell 8 to 17 (Goel, 2020).

MySQL database:

A screenshot of a computer

Description automatically generated

Figure 5

As I could not find an open MySQL database to connect, I will be demonstrating how to extract data from a local MySQL database. I created a table named *“tran\_hv\_psmod”* within a schema called *“eurostat”* and imported the previously fetched file from Eurostat, *“TRAN\_HV\_PSMOD.csv”*:

A screenshot of a computer

Description automatically generated

Figure 6

Once I had the records in the table, I was able to retrieve them using the Python package *“sqlalchemy”* (Oyama, 2022):

A screenshot of a computer

Description automatically generated

Figure 7

Complete instructions on creating the table, pushing, and fetching records are provided in the annex (*8.1. MySQL Setup*) (dev.mysql.com, n.d.).

# Documentation

In terms of documentation, I endeavoured to provide comments explaining the rationale behind each line of code. This includes details about the purpose of each line and the workflow for data manipulation and visualization.

I maintained code quality standards by adhering to the *“PEP 8”* styling guidelines (van Rossum, Warsaw and Coghlan, 2001). Given the project's nature, which primarily involves data manipulation and visualization, there was no need for the introduction of complex programming constructs or advanced *“OOP”* principles. My objective was to ensure clean and modularized code.

# Testing and Optimisation

# Testing

I have conducted extensive testing in the statistical section, examining each scenario to ensure the statistical model fits appropriately. Furthermore, I refined the testing process when accepting the null hypothesis. All statistical models in *“CA2\_Statistics\_Code\_sba23021.ipynb”* are implemented after thorough testing.  
Now, we are going to explore a testing example for a hypothesis test involving two populations.

This table will help us formulate the hypothesis:

A screenshot of a graph

Description automatically generated

Figure 8

H0: mu IE\_BUS = mu EU\_Country\_BUS; There is no significant difference between the percentage average on passenger-kilometres for Vehicle BUS in Ireland and EU\_Country.

H1: mu IE\_BUS != mu EU\_Country\_BUS; There is a significant difference between the percentage average on passenger-kilometres for Vehicle BUS in Ireland and EU\_Country.

Let us start the testing and choose Italy to determine whether we accept or reject the null hypothesis:

A white rectangular object with black text

Description automatically generated

Figure 9

A graph of a bus

Description automatically generated

Figure 10

We can clearly see that H0 is rejected therefore there is a significant difference in the average percentage of BUS passengers between Ireland and Italy.

Now we are going to select Slovenia as its mu is closer to the Irish one:

A white rectangular box with black text

Description automatically generated

Figure 11

A graph of a normal distribution

Description automatically generated

Figure 12

We accept H0 and conclude that there is not enough evidence to say that there is a significant difference between the percentage average on passenger-kilometres for vehicle BUS in Ireland and Slovenia.

There is another testing scenario in annex 8.1.2.

# Optimization

Good examples of optimization can be found in *“CA2\_ML\_Code\_sba23021.ipynb”*,where I enriched the datasets to achieve better accuracy in machine learning models.

I observed a low accuracy of 69.23% for SVC and the best score at 90.88% (C: 1000, Gamma=0.01). I identified that the model was not fitting properly at Gamma = 0.0001.

After enriching the dataset, I achieved an accuracy of 87.01% for SVC, with the best score at 91.88% (C: 1000, Gamma=0.001). The results indicate a less overfitted model as the best score is obtained at a higher Gamma. Additionally, the graph at Gamma=0.0001 shows that the test and train sets are more fitted after enrichment.  
In the annex 8.1.2 we can find other two scenarios for optimization.

A graph of different values

Description automatically generated

Figure 13

# Data Manipulation

In analysing different data sources, I utilized various libraries and techniques for both processing and aggregating data (GeeksforGeeks, 2020). The comparisons and contrasts for each data source are summarized in Figure X, and the detailed implementation and explanation of the code can be found in *“CA2\_Programming\_Code\_sba23021.ipynb”.*

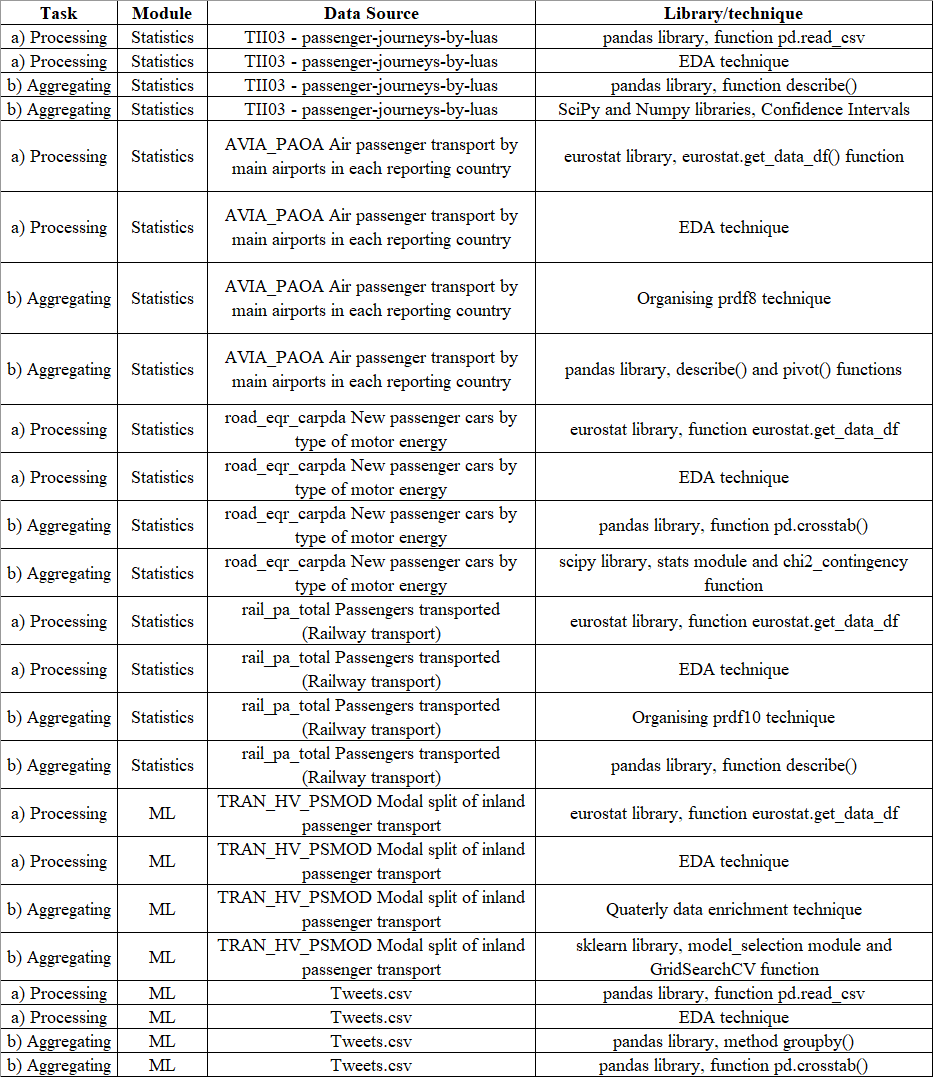


Figure 14

# **Statistical Analysis**

# Descriptive Statistics

# *Dataset for Confidence Interval*

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

A table with numbers and a number on it

Description automatically generated

Figure 15

We will be examining the mean values to establish the confidence intervals.  
It is interesting to note that both lines are quite balanced in terms of usage.

A group of pie charts with numbers

Description automatically generated

Figure 16

In Section 3.2, a detailed analysis of confidence intervals will be conducted for this dataset.

# *Dataset for Hypothesis Test Two Populations*

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

A group of colorful bars

Description automatically generated with medium confidence

Figure 17

The dataset used is *“Modal split of inland passenger transport”* withthe Eurostat code *“TRAN\_HV\_PSMOD”.*

# *Dataset for ANOVA One-Way*

Using Irish airports as a reference, an ANOVA will be conducted to test whether there are any statistically significant differences in the means compared to other European airports. The dataset used is *“Air passenger transport by main airports in each reporting country”* with the Eurostat code *“AVIA\_PAOA”.*

In the first scenario Dublin Airport is considered, and below a graph of the subset of airports selected for this case:

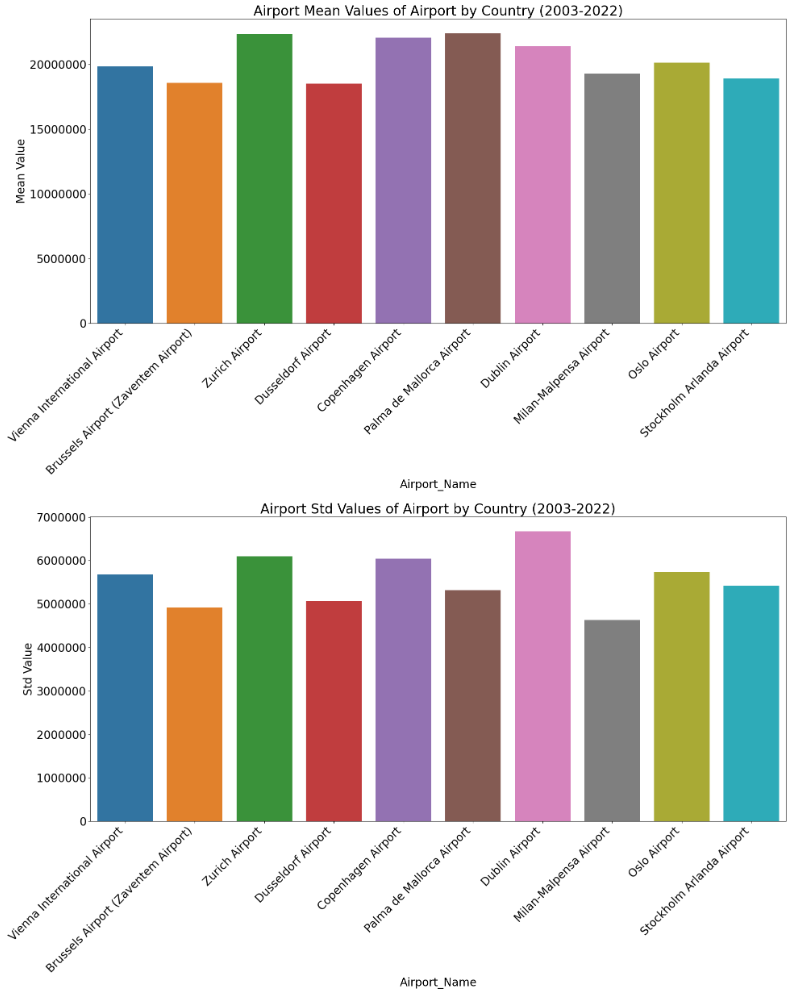


Figure 18

In the second scenario, Shannon Airport is considered, and below is a graph showing the subset of airports selected for this case:

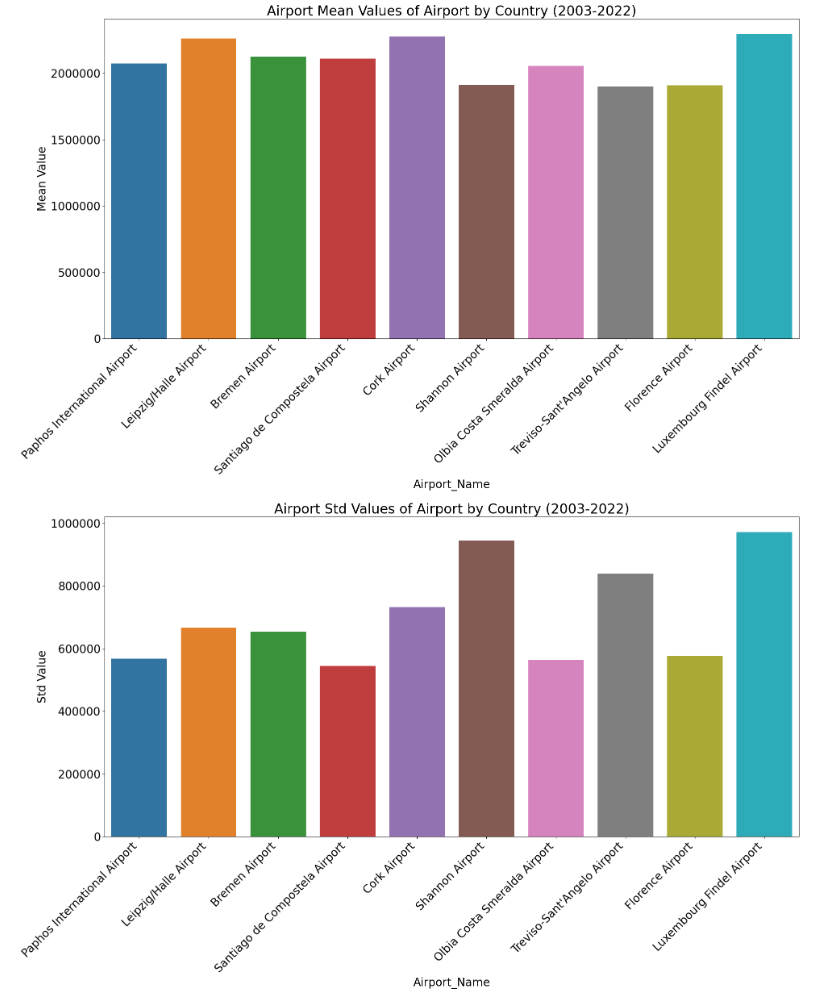


Figure 19

# *Dataset for Chi-Squared Test*

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

This model is highly sensitive to the frequency of the variables. I have presented two cases: one where we reject H0, and the second one where I manually changed values to accept H0. I will now illustrate how the categorical variable numbers will look for each case, with further 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

Figure 20

Scenario 2: Accepting H0.

A graph of different types of cars

Description automatically generated

Figure 21

# *Dataset for Kruskal-Wallis*

For this test, I utilized the same dataset as for ANOVA. Some of the airports violated the assumptions of normality required for ANOVA. The advantage of using the Kruskal-Wallis test is that it does not require normality to perform the test. I will present two scenarios: one to accept H0 and another to reject H0.

Scenario 1: Accepting H0.

A screenshot of a graph

Description automatically generated

Figure 22

Scenario 2: Rejecting H0.

A screenshot of a graph

Description automatically generated

Figure 23

# *Dataset for U-Mann Whitney*

I used a new dataset for this test, *“Passengers transported (Railway transport)”,* with the Eurostat code *“rail\_pa\_total”* because the data did not follow a normal distribution. This choice allowed me to demonstrate the test's capability to handle non-normally distributed data.

Scenario 1: Accepting H0.

A number on a white background

Description automatically generated

Figure 24

Scenario 1: Rejecting H0.

A number on a white background

Description automatically generated

Figure 25

# Confidence Interval

The task is to determine the weekly LUAS average for the total number of passengers in the years 2019, 2020, 2021, and 2022. It is important to note that both LUAS lines (red and green) are in scope. The analysis will be conducted with a 90% confidence level, and here are the results:

A table with numbers and a few black text

Description automatically generated

Figure 26

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

Figure 27

After performing confidence intervals, the next natural step is to verify if the weekly averages are the same for both the red and green LUAS lines. We will use a t-test (www.statisticslectures.com, n.d.).

A table with numbers and letters

Description automatically generated

Figure 28

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

Figure 29

A table with numbers and text

Description automatically generated

Figure 30

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

# Inferential Statistics

# *Parametric*

# *T-test Two Populations*

To perform this test, we will compare Ireland with three different countries, each having a distinct transportation method. The first scenario involves comparing Ireland with Slovenia for cars, the second compares Ireland with Denmark for buses, and the third compares Ireland with Slovenia again, but this time for trains.

Hypothesis:

H0: μ Ireland = μ EU Country.

H1: μ Ireland != μ EU Country.

Results:

A graph of a normal distribution

Description automatically generated

Figure 31

Since the p-value is greater than alpha, we accept the null hypothesis (H0). There is not enough evidence to conclude that there is a significant difference between the percentage average of passenger-kilometres for the BUS vehicle in Ireland and Slovenia.

A graph of a normal distribution

Description automatically generated

Figure 32

As the p-value is less than alpha, we reject the null hypothesis (H0), providing sufficient evidence to conclude a significant difference in the percentage average of passenger-kilometres for the Car vehicle between Ireland and Denmark.

A graph of a normal distribution

Description automatically generated

Figure 33

As the p-value is less than alpha, we reject the null hypothesis (H0), indicating sufficient evidence to conclude a significant difference in the percentage average of passenger-kilometres for the Train vehicle between Ireland and Slovenia.

A screenshot of a calculator

Description automatically generated

Figure 34

# *ANOVA One-Way*

In the first scenario, we examine Dublin, Zurich, and Copenhagen airports to verify whether the yearly average passenger numbers for the period from 2003 to 2022 are the same or not (Laerd Statistics, 2018).

To perform ANOVA, we need both Shapiro-Wilk and Levene tests to have p-values greater than 5% alpha.

A screenshot of a test results

Description automatically generated

Figure 35

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

Figure 36

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

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

Shapiro-Wilk and Levene tests:

A screenshot of a test

Description automatically generated

Figure 37

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

Figure 38

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 with no association or relationship.

A screenshot of a number of numbers

Description automatically generated

Figure 39

H1: There is a significant difference between the observed and expected frequencies, indicating a non-independent relationship between Ireland and Austria

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

A graph of a car

Description automatically generated

Figure 40

In the second scenario, I manually adjusted values for Austria to align frequencies more closely with those of Ireland; this model is highly sensitive to substantial differences between categorical variables.

A screenshot of a number

Description automatically generated

Figure 41

After applying the Chi-Square test, we obtained a p-value of 0.59, leading us to accept H0.

A graph of a car

Description automatically generated

Figure 42

# *Non-parametric*

# *Kruskal-Wallis*

Scenario 1:

We do not have normality for our samples (Xia, 2020):

A screenshot of a computer

Description automatically generated

Figure 43

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

Figure 44

We accept H0; the means of annual passengers carried (2003-2022) for Ireland West Knock, Pau Pyrenees, and Goteborg airports are quite similar.

Scenario 2:

We do not have normality for our samples:

A screenshot of a computer

Description automatically generated

Figure 45

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

Figure 46

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

# *U-Mann Whitman*

Scenario 1:

Checking normality:

**A white rectangular object with black text

Description automatically generated**

Figure 47

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

Figure 48

As the p-value is greater than alpha, we accept H0, indicating that there is no significant difference between the average number of train passengers in Ireland and Croatia.

Scenario 2:

Checking normality:

A white background with black text

Description automatically generated

Figure 49

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

Figure 50

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

The tests I have conducted above reveal interesting findings:

*“Number of public transport journeys at highest level since the beginning of the pandemic”* (BreakingNews.ie, 2022), based on the confidence intervals, we can confirm that the number of passengers is recovering:

A graph of a number of people

Description automatically generated with medium confidence

Figure 51

*“Paris Charles De Gaulle recorded the highest number of air passengers”* (ec.europa.eu, n.d.). *“Dublin Airport Was EU’s 11th Largest Airport in 2018”* (DublinAirport, n.d.):

A graph of different colored bars

Description automatically generated

Figure 52

Analising means in the ANOVA section we can see that those headings are highly correlated with *“FR\_LFPG”* and *“IE\_EIDW”.*

The challenges faced included gathering the data and establishing the scenarios to perform the tests.

# **ML**

The methodology for ML part can be seen as follows (docs.aws.amazon.com, n.d.):

A diagram of a data processing process

Description automatically generated

Figure 53

# Supervised Learning.

The choice of the dataset *“TRAN\_HV\_PSMOD”* and the selection of supervised ML models (Decision Tree, Random Forest, K-Nearest Neighbours, and Support Vector Machine) are purely matters of modelling. After numerous attempts, this combination has proven effective (Thanh Noi and Kappas, 2017).

We will be modelling the dataset to see how models react having these classes:

A screenshot of a computer

Description automatically generated

Figure 54

I paired Ireland with Hungary because it showed the best performance association for selected ML models, here are the results:

A screenshot of a computer

Description automatically generated

Figure 55

Overall model is performing well however for Gamma = 0.0001, models seem to be overfitted (Kilic, 2023) as this score is lower compared to Gamma 0.01 and 0.001:

A graph of a graph with numbers

Description automatically generated with medium confidence

Figure 56

To address the issue of overfitting (Ghosh, 2023), I enriched the dataset by quarterly weighting yearly values:

A close-up of a number

Description automatically generated

Figure 57

A screenshot of a graph

Description automatically generated

Figure 58

The results are as follows:

A screenshot of a computer

Description automatically generated

Figure 59

A graph of a number of numbers

Description automatically generated with medium confidence

Figure 60

Now we have a better fit, with higher scores observed at Gamma = 0.0001.

# Cross Validation and Feature Reduction.

Ireland will be our target variable for each transportation method. In the first attempt with yearly data, we obtained the following results:

A screenshot of a number of features

Description automatically generated

Figure 61

After quarterly enrichment, increasing the number of rows from 32 to 128, we obtained the following results:

A screenshot of a computer

Description automatically generated

Figure 62

Linear regression estimation has performed much better with enriched data.

# Unsupervised Learning.

Dataset it is split into each mode of transportation having Ireland as a reference.

First attempt with quarterly data we get the following results:

A screenshot of a computer

Description automatically generated

Figure 63

A screenshot of a computer

Description automatically generated

Figure 64

To improve the results, we are going to enhance the dataset by breaking down each year into monthly values. The results are as follows:

A screenshot of a computer

Description automatically generated

Figure 65

A screen shot of a number

Description automatically generated

Figure 66

*KMeans* and *PCA* have performed better with more data (L, 2020).

# Sentiment Analysis.

We are going to conduct sentiment analysis using Ryanair reviews and a set of tweets related to USA airlines. Results:

A comparison of a number of squares

Description automatically generated with medium confidence

Figure 67

*Tweets* dataset is larger than *Ryanair* one that explains more correctly classified inputs:

A screenshot of a graph

Description automatically generated

Figure 68

A screenshot of a computer

Description automatically generated

Figure 69

Model accuracy for *Ryanair* is 73% and 75% for *USA Airlines.* ROC results as it follows:

A graph of a curve

Description automatically generated

Figure 70

*USA Airlines* seem to be classifying bad inputs better, while the other two are closer to each other. Let us test the classifier by adding reviews:

A screenshot of a computer program

Description automatically generated

Figure 71

Result:

A graph of different sizes and colors

Description automatically generated with medium confidence

Figure 72

*Ryanair's* sentiment analysis performs well even though it has fewer inputs than *USA Airlines.* This difference could be attributed to the collection of reviews from *Tripadvisor,* where I ensured capturing opinions across good, bad, and neutral categories.

# Table and Conclusions

# *Supervised Learning*

Let us compare each model after enriching the dataset.

Decision Trees CM:

A screenshot of a diagram

Description automatically generated

Figure 73

Random Forrest:

A screenshot of a graph

Description automatically generated

Figure 74

KNN:

A graph of different colored lines

Description automatically generated with medium confidence

Figure 75

GridSearchCV CM:

A screenshot of a graph

Description automatically generated

Figure 76

Gamma Accuracy:

A graph of different values

Description automatically generated

Figure 77

Accuracy Table:

A table with numbers and text

Description automatically generated

Figure 78

By adding more values, models have performed better, with the exception of RF. However, we solved overfitting in the hyperparameter tuning phase, achieving higher accuracy at a lower level of gamma.

# *Cross Validation and Feature Reduction*

Same approach as we followed previously, we increased dataset row count by breaking down years into quarters, here the results:

A group of graphs showing the value of a number of patients

Description automatically generated with medium confidence

Figure 79

Summary table:

A screenshot of a number

Description automatically generated

Figure 80

Optimal feature selection and improved accuracy after enriching the dataset, undoubtedly, this method requires a larger dataset to perform well.

# *Unsupervised Learning*

PCA Variance:

A collage of blue and red graph

Description automatically generated

Figure 81

PCA Scatter Plot:

A group of colored dots

Description automatically generatedFigure 82

PCA Heatmap:

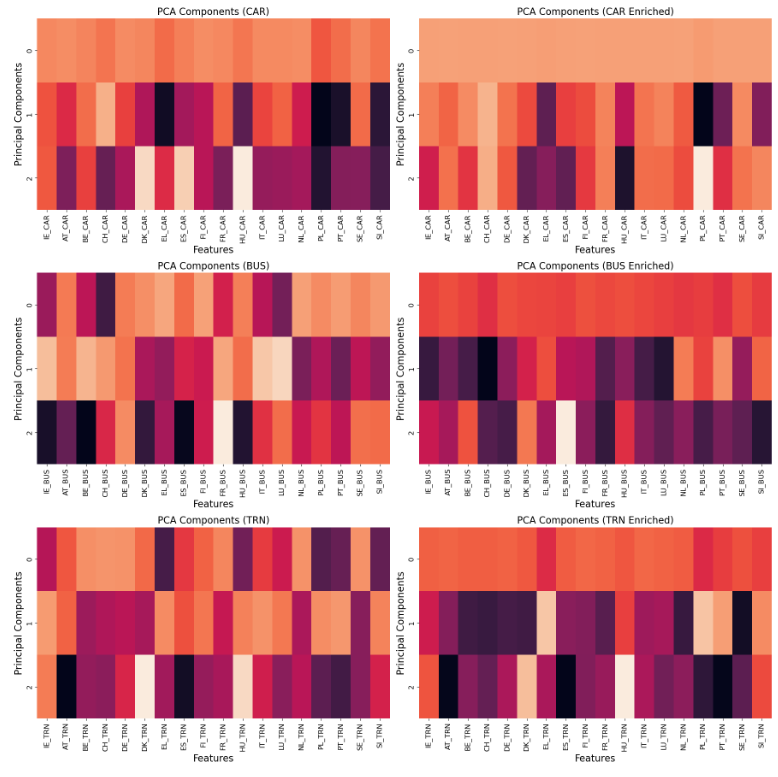


Figure 83

Elbow Method:

A graph of a number of different numbers

Description automatically generated with medium confidence

Figure 84

Silhouette Score:

A group of graphs showing different sizes of clusters

Description automatically generated

Figure 85

Tables PCA and Silhouette Score:

A screenshot of a computer screen

Description automatically generated

Figure 86

A screenshot of a graph

Description automatically generated

Figure 87

After enriching the data, the first principal component explains almost 97% of the variability for the CAR and BUS datasets, while the TRN dataset exhibits 76.86% variability on the first component. This demonstrates a clear improvement.  
The silhouette score has improved, indicating that the clusters are now closer to being well-defined (closer to +1). Additionally, all plots demonstrate a clear improvement after enlarging the dataset.

# **Data Preparation and Visualisation**

# Data acquisition

Undoubtedly, this was the most challenging part of the assignment. However, I discovered the *Eurostat* website, which provides extensive content on data transportation. Additionally, the *CSO* offers interesting datasets related to the Irish transportation sector. Other platforms utilized included *GitHub*, *Kaggle*, and *TripAdvisor.*

A positive aspect of this research is that I discovered platforms such as *Eurostat*, which even has a *Python* package to download its datasets. On the negative side, the length of transportation datasets I found was a challenge for the machine learning part. There was not enough data to create accurate and consistent models. However, I found a solution to overcome this issue.

In terms of licenses, we are mostly covered by *Creative Commons*, allowing us to use these datasets. Only for *TripAdvisor Ryanair* reviews, I do not have licenses. After thorough research, I believe I am not infringing any law, as I am not collecting any personal data. I am compliant with GDPR (Your Europe, 2019), and *TripAdvisor* does not state that you cannot collect their reviews for research purposes.

A white text on a black background

Description automatically generated

Figure 88

# EDA methodology

One of the issues was that Eurostat uses its nomenclature, and each dataset needed to be crosschecked to obtain accurate numbers. However, the main challenge was adapting datasets for the statistical models. EDA for this section can be found in *“CA2\_Data\_Visualisation\_Code\_sba23021.ipynb”* and some explanation in the annex (8.3.1). Nevertheless, EDA is also prevalent throughout the entire assignment. Here is a summary of the EDA performed:

A white sheet with black text

Description automatically generated

Figure 89

# EDA ML implementation

For ML I utilized *“TRAN\_HV\_PSMOD”* due to the versatility this dataset provides for modelling both *supervised* and *unsupervised* learning, as well as for *cross-validation* and *feature reduction*. Additionally, the dataset enabled a comparison of Ireland's modal split transport with that of other European countries.

The challenge lay in improving accuracy. The solution involved breaking down years into quarters for both *supervised* and *cross-validation/feature reduction*. For *unsupervised* learning, I subdivided years into months to leverage the increased data availability. Undoubtedly, the models performed better with the additional data.

**Quarterly enrichment:**

A screenshot of a computer code

Description automatically generated

Figure 90

A graph of a number of numbers and a number of numbers

Description automatically generated with medium confidence

Figure 91

A screenshot of a computer code

Description automatically generated

Figure 92

A graph of a number of people

Description automatically generated with medium confidence

Figure 93

**Monthly enrichment:**

**A screenshot of a computer code

Description automatically generated**

Figure 94

**A comparison of a graph

Description automatically generated**

Figure 95

All EDA for ML can be seen in *“CA2\_ML\_Code\_sba23021.ipynb”*. I also created a more concise version *“CA2\_Data\_Visualisation\_Code\_sba23021.ipynb”*, which emphasizes the relevant aspects of the EDA.Additional information is available in the annex (8.3.2).

# Dashboard

The rationale and visualization choices can be seen in the table below, as follows:

A screenshot of a computer

Description automatically generated

Figure 96

As mentioned above I developed a dashboard in *“CA2\_Data\_Visualisation\_Code\_sba23021.ipynb”* outlining the relevant results for ML models. To comply with point *“Modern Transport planning has a great dependence… web based”* I adapted the dashboard to a *streamlit* (Parker, 2023)web dashboard. You can visit it at: <https://datavisualisationmoderntransportdashstudentsba23021.streamlit.app/>.

*“CA2\_Data\_Visualisation\_Code\_sba23021.ipynb”* dashboard:

A screenshot of a graph

Description automatically generated

Figure 97

*Streamlit* dashboard:

A screenshot of a graph

Description automatically generated

Figure 98

Complete dashboards can also be viewed in the annex (8.3.3).

# **Conclusion**

In conclusion, the assignment most formidable phase involved the collection and adaptation of data for diverse methodologies. While my findings may not offer a panacea for the challenges faced by the Irish Transportation sector, I am confident that the statistical tests and Machine Learning models employed are presented in an accessible and pragmatic manner. This journey underscored the intricacies of data integration and analysis, highlighting the importance of methodological clarity. As I reflect on the complexities encountered, I recognize that this study contributes to a nuanced understanding of the sector. While not providing all-encompassing solutions, it lays the groundwork for informed discussions and potential pathways toward addressing the intricacies of the Irish Transportation landscape. I believe that the transparent modelling and interpretability fostered in this study pave the way for future endeavours in unravelling the challenges inherent in the field of transportation.

Word Count: 3,297 words

# **References**

Programming:

Goel, A. (2020). *How to parse JSON data with Python Pandas?* [online] Medium. Available at: https://towardsdatascience.com/how-to-parse-json-data-with-python-pandas-f84fbd0b1025 [Accessed 23 Dec. 2023].

Oyama, F.K. (2022). *How to Connect MySQL Database in Jupyter Notebook.* [online] Faith Kinkema Oyama. Available at: https://kema.hashnode.dev/how-to-connect-mysql-database-in-jupyter-notebook [Accessed 7 Jan. 2024].

dev.mysql.com. (n.d.). MySQL :: *Getting Started with MySQL.* [online] Available at: <https://dev.mysql.com/doc/mysql-getting-started/en/>.

van Rossum, G., Warsaw, B. and Coghlan, N. (2001). *PEP 8 – Style Guide for Python Code | peps.python.org.* [online] peps.python.org.Available at: <https://peps.python.org/pep-0008/>.

GeeksforGeeks. (2020). *Data Manipulation in Python using Pandas.* [online] Available at: <https://www.geeksforgeeks.org/data-manipulattion-in-python-using-pandas/>.

www.statisticslectures.com. (n.d.). *Confidence Intervals for Independent Samples t-Test.* [online] Available at: http://www.statisticslectures.com/topics/ciindependentsamplest/.

‌ Laerd Statistics (2018). *One-way ANOVA - An introduction to when you should run this test and the test hypothesis*. [online] Laerd.com. Available at: <https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php>.

Xia, Y. (2020). *Kruskal Wallis Test - an overview | ScienceDirect Topics.* [online] www.sciencedirect.com. Available at: https://www.sciencedirect.com/topics/medicine-and-dentistry/kruskal-wallis-test.

BreakingNews.ie. (2022). *Number of public transport journeys at highest level since the beginning of the pandemic.* [online] Available at: https://www.breakingnews.ie/ireland/number-of-public-transport-journeys-at-highest-level-since-the-beginning-of-the-pandemic-1304260.html [Accessed 11 Dec. 2023].

ec.europa.eu. (n.d.). *Air transport statistics.* [online] Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Air\_transport\_statistics#:~:text=In%202022%2C%20820%20million%20people%20in%20the%20EU%20travelled%20by%20air.&text=In%202022%2C%20Paris%20Charles%20De [Accessed 11 Dec. 2023].

DublinAirport. (n.d.). *Dublin Airport Was EU’s 11th Largest Airport in 2018*. [online] Available at: <https://www.dublinairport.com/latest-news/2019/05/31/dublin-airport-was-eu-s-11th-largest-airport-in-2018>.

docs.aws.amazon.com. (n.d.). *Machine learning lifecycle - Machine Learning Best Practices in Healthcare and Life Sciences.* [online] Available at: https://docs.aws.amazon.com/whitepapers/latest/ml-best-practices-healthcare-life-sciences/machine-learning-lifecycle.html [Accessed 7 Jan. 2024].

Thanh Noi, P. and Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland),* [online] 18(1), p.18. doi:https://doi.org/10.3390/s18010018.

Kilic, D.K. (2023). *Tuning Hyperparameters in Machine Learning Models 101*. [online] Medium. Available at: https://medium.com/@deniz.kenan.kilic/changing-hyperparameters-in-machine-learning-models-101-d969f51fe414 [Accessed 7 Jan. 2024].

Ghosh, B. (2023). *What Matters More — Data Size or Model Size.* [online] Medium. Available at: <https://medium.com/@bijit211987/what-matters-more-data-size-or-model-size-31cb004d7209>.

L, P. (2020). *Intro to PCA, K-means Clustering | Unsupervised Learning Algorithms.* [online] Medium. Available at: <https://towardsdatascience.com/into-to-pca-k-means-clustering-unsupervised-learning-algorithms-5cc5acea274d>.

Your Europe (2019). *Data protection.* [online] Your Europe - Business. Available at: <https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm>.

Parker, K. (2023). *Creating a Streamlit-Powered Data Dashboard: A Step-by-Step Guide.* [online] Medium. Available at: https://medium.com/@parktwin2/creating-a-streamlit-powered-data-dashboard-a-step-by-step-guide-e24fa016eadd [Accessed 7 Jan. 2024].

# **Annex**

# Programming

# *MySQL Setup*

I used *MySQL Workbench* to simulate the process of gathering data from a MySQL database.

Steps:

1. For schema creation, click on the database icon, name it *“eurostat”* in this case, and then click on *Apply*:

A screenshot of a computer

Description automatically generated

Figure 99

1. Click *Apply:*

A screenshot of a computer

Description automatically generated

Figure 100

1. Click *finish:*

A screenshot of a computer

Description automatically generated

Figure 101

1. Our schema *eurostat* will appear:

A screenshot of a computer

Description automatically generated

Figure 102

5) To create the table within the schema for inserting *“TRAN\_HV\_PSMOD.csv”,* load the *“TRAN\_HV\_PSMOD\_table\_creation.sql”* file.  
Click on File and select Open SQL Script:

A screenshot of a computer program

Description automatically generated

Figure 103

6) Execute the code from *“TRAN\_HV\_PSMOD\_table\_creation.sql”:*

A screenshot of a computer

Description automatically generated

Figure 104

7) After successful execution, the table will appear under the *“eurostat”* schema. Please refresh to view the table:

A screenshot of a computer

Description automatically generated

Figure 105

8) Hover over the table, right-click, and select Table Data Import Wizard:

A screenshot of a computer

Description automatically generated

Figure 106

9) Browse and locate *“TRAN\_HV\_PSMOD.csv”,* the click on open and next:

A screenshot of a computer

Description automatically generated

Figure 107

10) *Next:*

A screenshot of a computer

Description automatically generated

Figure 108

11) *Next:*

A screenshot of a computer

Description automatically generated

Figure 109

12) *Next:*

A screenshot of a computer

Description automatically generated

Figure 110

13) *Next:*

A screenshot of a computer

Description automatically generated

Figure 111

14) *Finish:*

A screenshot of a computer

Description automatically generated

Figure 112

15) Checking the results, we should have same number of columns and row count:

A screenshot of a computer

Description automatically generated

Figure 113

A screenshot of a computer

Description automatically generated

Figure 114

16) Connecting to MySQL and fetching the table *“tran\_hv\_psmod”:*

A screenshot of a computer

Description automatically generated

Figure 115

This is how we can connect to a MySQL database and pull data from there.

# *Testing and Optimisation*

**Testing:**

The choice of selecting Hungary, along with Ireland, for the Supervised Learning models was not a random decision. I tested the models with each possible combination involving Ireland and another country, and the pair of Ireland and Hungary offered the best outcome. In *“CA2\_Programming\_Code\_sba23021.ipynb”* I implemented the models for Ireland and the Netherlands here the results:

Results for Ireland and the Netherlands:

A screenshot of a graph

Description automatically generated

Figure 116

Results for Ireland and Hungary:

A table with numbers and text

Description automatically generated

Figure 117

**Optimisation:**

For cross validation, feature reduction and unsupervised ML models I enlarged the datasets to optimize accuracy scores.

Initial data frame for cross validation and feature reduction:

A screenshot of a computer code

Description automatically generated

Figure 118

After data enrichment:

A screenshot of a computer code

Description automatically generated

Figure 119

We went from 32 rows to 128, that helped ML performance.

Initial data frame for unsupervised ML models:

A screenshot of a computer screen

Description automatically generated

Figure 120

After data enrichment:

A screenshot of a computer code

Description automatically generated

Figure 121

We went from 32 rows to 384, that helped ML performance.

Full implementation of this code can be found in *“CA2\_ML\_Code\_sba23021.ipynb”.* And a concise implementation of quarterly and monthly functions for data enrichment can be found in *“CA2\_Programming\_Code\_sba23021.ipynb”.*

# Statistics

A summary of all statistical models and their results can be seen as it follows:

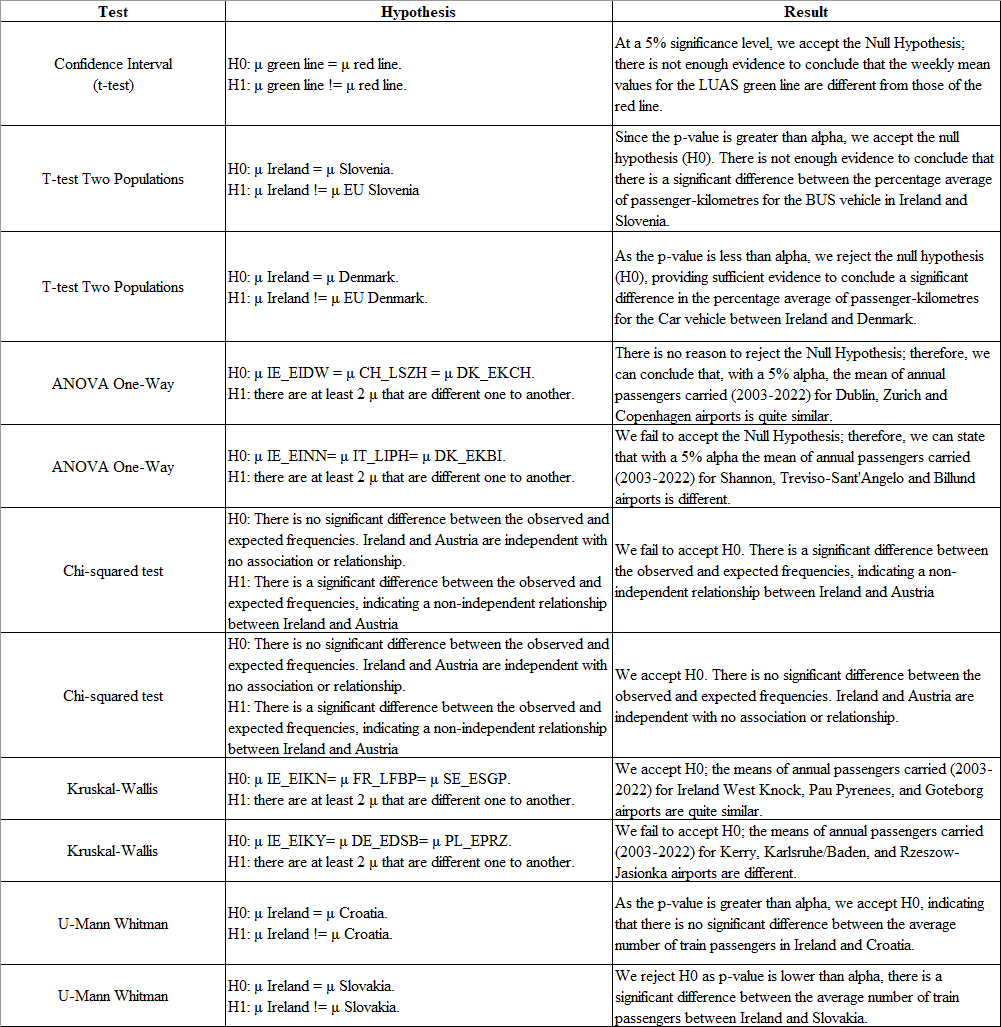


Figure 122

The Python calculations for each of the statistical tests can be seen as follows:

**Confidence Intervals:**

Year 2019:

**A screenshot of a computer code

Description automatically generated**

Figure 123

Year 2020:

A screenshot of a computer code

Description automatically generated

Figure 124

Year 2021:

A math equation with numbers and letters

Description automatically generated with medium confidence

Figure 125

Year 2022:

A math equation with numbers and letters

Description automatically generated with medium confidence

Figure 126

Summary:

A screenshot of a graph

Description automatically generated

Figure 127

A group of graphs showing different sizes of lines

Description automatically generated with medium confidence

Figure 128

**Hypothesis Test Confidence Intervals:**

Year 2019:

A close-up of a number

Description automatically generated

Figure 129

Year 2020:

A white rectangular object with text

Description automatically generated

Figure 130

Year 2021:

A close-up of a computer screen

Description automatically generated

Figure 131

Year 2022:

A white rectangular object with numbers and letters

Description automatically generated

Figure 132

Summary:

A table with numbers and text

Description automatically generated

Figure 133

A group of graphs with numbers and symbols

Description automatically generated with medium confidence

Figure 134

**Inferential statistics**

**Parametric tests**

**Hypothesis Test Two Populations:**

HT Test Ireland & Slovenia BUS % average on passenger-kilometres:

A white background with black text and numbers

Description automatically generated

Figure 135

HT Test Ireland & Denmark CAR % average on passenger-kilometres:

A white rectangular object with black text

Description automatically generated

Figure 136

HT Test Ireland & Slovenia TRN % average on passenger-kilometres:

A white background with black and green text

Description automatically generated

Figure 137

Summary:

**A number of numbers on a white background

Description automatically generated**

Figure 138

**Anova one-way:**

First Scenario Dublin Airport Anova test:

1. Normality plot:

A screen shot of a graph

Description automatically generated

Figure 139

1. Shapiro Wilk test:

A screenshot of a computer

Description automatically generated

Figure 140

1. Levene test:

A screenshot of a computer

Description automatically generated

Figure 141

Summary before Anova:

A screenshot of a test results

Description automatically generated

Figure 142

1. Anova:

A screenshot of a computer code

Description automatically generated

Figure 143

We accept H0.

Second Scenario Shannon Airport Anova test:

1. Normality plot:

A screen shot of a graph

Description automatically generated

Figure 144

1. Shapiro Wilk test:

A screenshot of a computer program

Description automatically generated

Figure 145

1. Levene test:

A screenshot of a computer

Description automatically generated

Figure 146

Summary before Anova:

A screenshot of a test

Description automatically generated

Figure 147

1. Anova:

A screenshot of a computer code

Description automatically generated

Figure 148

We reject H0.

**Chi-squared test:**

Scenario 1:

A close-up of a computer screen

Description automatically generated

Figure 149

We reject H0.

Scenario 2:

A white rectangular object with black numbers

Description automatically generated

Figure 150

We accept H0.

**Non-parametric tests**

**Kruskal-Wallis:**

Scenario 1:

Step 1, Shapiro Wilk test:

A screenshot of a computer

Description automatically generated

Figure 151

Summary:

A screenshot of a table

Description automatically generated

Figure 152

Step 2, Kruskal-Wallis test:

A screenshot of a computer test

Description automatically generated

Figure 153

We accept H0.

Scenario 2:

Step 1, Shapiro Wilk test:

A screenshot of a computer program

Description automatically generated

Figure 154

Summary:

A screenshot of a graph

Description automatically generated

Figure 155

Step 2, Kruskal-Wallis test:

A screenshot of a computer code

Description automatically generated

Figure 156

We reject H0.

**U Mann-Whitney:**

Scenario 1:

Step 1, checking normality:

A screenshot of a computer

Description automatically generated

Figure 157

Summary:

A white background with black text

Description automatically generated

Figure 158

Step 2, U Mann-Whitney test:

A screenshot of a computer

Description automatically generated

Figure 159

We accept H0.

Scenario 2:

Step 1, checking normality:

A white rectangular object with black text

Description automatically generated

Figure 160

Summary:

A white background with black text

Description automatically generated

Figure 161

Step 2, U Mann-Whitney test:

A screenshot of a computer

Description automatically generated

Figure 162

We reject H0.

# Data Visualisation

# *EDA Methodology*

**EDA *Confidence Intervals.***

We are dropping *“All Luas lines”* except for the red and green lines for modelling purposes*:*

*A graph of different sizes and colors

Description automatically generated with medium confidence*

Figure 163

We are also dropping the year 2023 due to missing data; modelling under such conditions is not feasible.

A graph of blue bars

Description automatically generated

Figure 164

**EDA *Hypothesis Test for Two Populations***

We have many unnecessary columns in this dataset for modelling purposes:

A graph of a column

Description automatically generated

Figure 165

Organising numbers for each modal split transport to fit the scenarios:

A graph of blue and white bars

Description automatically generated with medium confidence

Figure 166

**EDA *Anova one-way* and *Kruskal-Wallis***

Quite large dataset that required dropping a significant number of columns:

*A blue column with white text

Description automatically generated*

Figure 167

**EDA *Chi-Squared test***

We had to drop some columns for this dataset:

*A graph showing a comparison of columns

Description automatically generated*

Figure 168

**EDA *U-Mann Whitney test***

There are some columns in this dataset that are not needed.:

*A graph with blue rectangular bars

Description automatically generated with medium confidence*

Figure 169

# *EDA ML implementation*

**Quarterly enrichment *Supervised Learning:***

A screenshot of a computer code

Description automatically generated

Figure 170

A graph of a number of numbers and a number of numbers

Description automatically generated with medium confidence

Figure 171

After enrichment, the second confusion matrix shows a significant improvement in classification:

A screenshot of a grid

Description automatically generated

Figure 172

Gamma=0.0001 is more fitted after enlarging the dataset hyperparameters have changed from 90.88% (C: 1000, gamma: 0.01) to 91.04% (C: 1000, gamma: 0.001):

A graph of different values

Description automatically generated

Figure 173

**Quarterly enrichment *Cross Validation and Feature Reduction:***

A screenshot of a computer code

Description automatically generated

Figure 174

A graph of a number of people

Description automatically generated with medium confidence

Figure 175

After enrichment, the second confusion matrix shows a significant improvement in classification:

A group of graphs showing the value of a number of patients

Description automatically generated with medium confidence

Figure 176

**Monthly enrichment *Unsupervised Learning:***

**A screenshot of a computer code

Description automatically generated**

Figure 177

**A comparison of a graph

Description automatically generated**

Figure 178

Visually, we can observe improved results after breaking down the data monthly:

**PCA Variance:**

A collage of blue and red graph

Description automatically generated

Figure 179

**PCA Scatter Plot:**

A group of colored dots

Description automatically generated

Figure 180

**PCA Heatmap:**

A group of colorful squares

Description automatically generated with medium confidence

Figure 181

**Elbow Method:**

A graph of a number of different numbers

Description automatically generated with medium confidence

Figure 182

**Silhouette Score:**

A group of graphs showing different sizes of clusters

Description automatically generated

Figure 183

# *Dashboard*

First, I created a dashboard in the Jupyter notebook *“CA2\_Data\_Visualisation\_Code\_sba23021.ipynb”* let us explore all the different graphs it contains. You can choose from the dropdown menu among various options, as follows:

A screenshot of a computer

Description automatically generated

Figure 184

Supervised Learning:

A graph of a number of people

Description automatically generated with medium confidence

Figure 185

Cross Validation and Feature Reduction:

A screenshot of a graph

Description automatically generated

Figure 186

Unsupervised Learning Silhouette Score:

A graph with red and blue squares

Description automatically generated

Figure 187

Unsupervised Learning PCA Variance:

A graph of different colored bars

Description automatically generated

Figure 188

Sentiment Analysis:

A screenshot of a graph

Description automatically generated

Figure 189

We also have a *Streamlit* version that can be shared online, where we can select options from the dropdown menu:

A screenshot of a computer

Description automatically generated

Figure 190

Supervised Learning:

A screenshot of a graph

Description automatically generated

Figure 191

Cross Validation and Feature Reduction:

A screenshot of a graph

Description automatically generated

Figure 192

Unsupervised Learning Silhouette Score:

A screenshot of a graph

Description automatically generated

Figure 193

Unsupervised Learning PCA Variance:

A graph of different colored bars

Description automatically generated

Figure 194

Sentiment Analysis:

A screenshot of a graph

Description automatically generated

Figure 195

The above dashboard can be visited at:

https://datavisualisationmoderntransportdashstudentsba23021.streamlit.app/