**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](#_Toc154490471)

[**1.** **Introduction** 3](#_Toc154490472)

[**2.** **Programming** 4](#_Toc154490473)

[2.1. Programming 4](#_Toc154490474)

[2.2. Data Structures 4](#_Toc154490475)

[2.3. Documentation 4](#_Toc154490476)

[2.4. Testing and Optimisation 4](#_Toc154490477)

[2.5. Data Manipulation 4](#_Toc154490478)

[**3.** **Statistical Analysis** 4](#_Toc154490479)

[3.1. Descriptive Statistics 4](#_Toc154490480)

[*3.1.1.* *Dataset for Confidence Interval* 4](#_Toc154490481)

[*3.1.2.* *Dataset for Hypothesis Test Two Populations* 5](#_Toc154490482)

[*3.1.3.* *Dataset for ANOVA One-Way* 7](#_Toc154490483)

[*3.1.4.* *Dataset for Chi-Squared Test* 8](#_Toc154490484)

[*3.1.5.* *Dataset for Kruskal-Wallis* 9](#_Toc154490485)

[*3.1.6.* *Dataset for U-Mann Whitney* 11](#_Toc154490486)

[3.2. Confidence Interval 12](#_Toc154490487)

[3.3. Inferential Statistics 14](#_Toc154490488)

[*3.3.1.* *Parametric* 14](#_Toc154490489)

[*3.3.1.1.* *T-test Two Populations* 14](#_Toc154490490)

[*3.3.1.2.* *ANOVA One-Way* 16](#_Toc154490491)

[*3.3.1.3.* *Chi-squared test* 18](#_Toc154490492)

[*3.3.2.* *Non-parametric* 19](#_Toc154490493)

[*3.3.2.1.* *Kruskal-Wallis* 19](#_Toc154490494)

[*3.3.2.2.* *U-Mann Whitman* 21](#_Toc154490495)

[3.4. Further Research and Challenges Faced 23](#_Toc154490496)

[**4.** **ML** 24](#_Toc154490497)

[4.1. Supervised Learning, GridSearchCV Hyperparameter Tunning. 25](#_Toc154490498)

[4.2. Cross Validation and Feature Reduction. 26](#_Toc154490499)

[4.3. Unsupervised Learning. 27](#_Toc154490500)

[4.4. Sentiment Analysis. 27](#_Toc154490501)

[4.5. Table and Conclusions 30](#_Toc154490502)

[*4.5.1.* *Supervised Learning* 30](#_Toc154490503)

[*4.5.2.* *Cross Validation and Feature Reduction* 33](#_Toc154490504)

[*4.5.3.* *Unsupervised Learning* 34](#_Toc154490505)

[**5.** **Data Preparation and Visualisation** 35](#_Toc154490506)

[5.1. Data acquistion 35](#_Toc154490507)

[5.2. EDA methodology 35](#_Toc154490508)

[5.3. Visualisations 35](#_Toc154490509)

[5.4. Dashboard 35](#_Toc154490510)

[**6.** **Conclusion** 35](#_Toc154490511)

[**7.** **References** 35](#_Toc154490512)

# **Introduction**

# **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

# 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 statistical 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

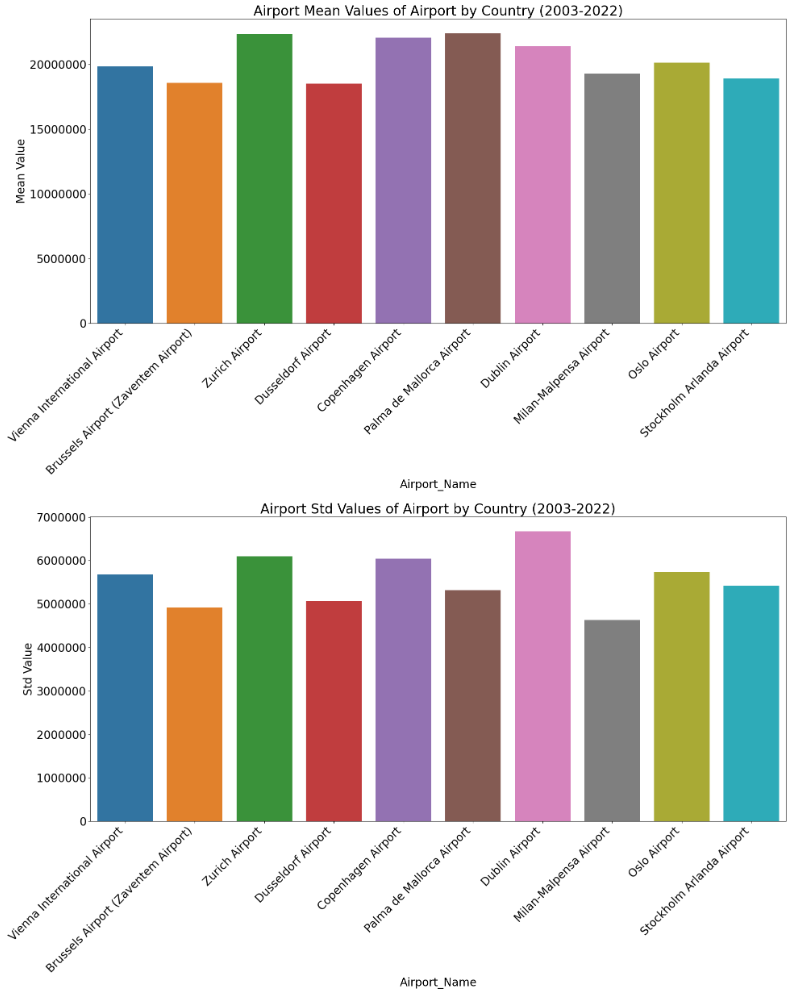
Description automatically generated with medium confidence

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

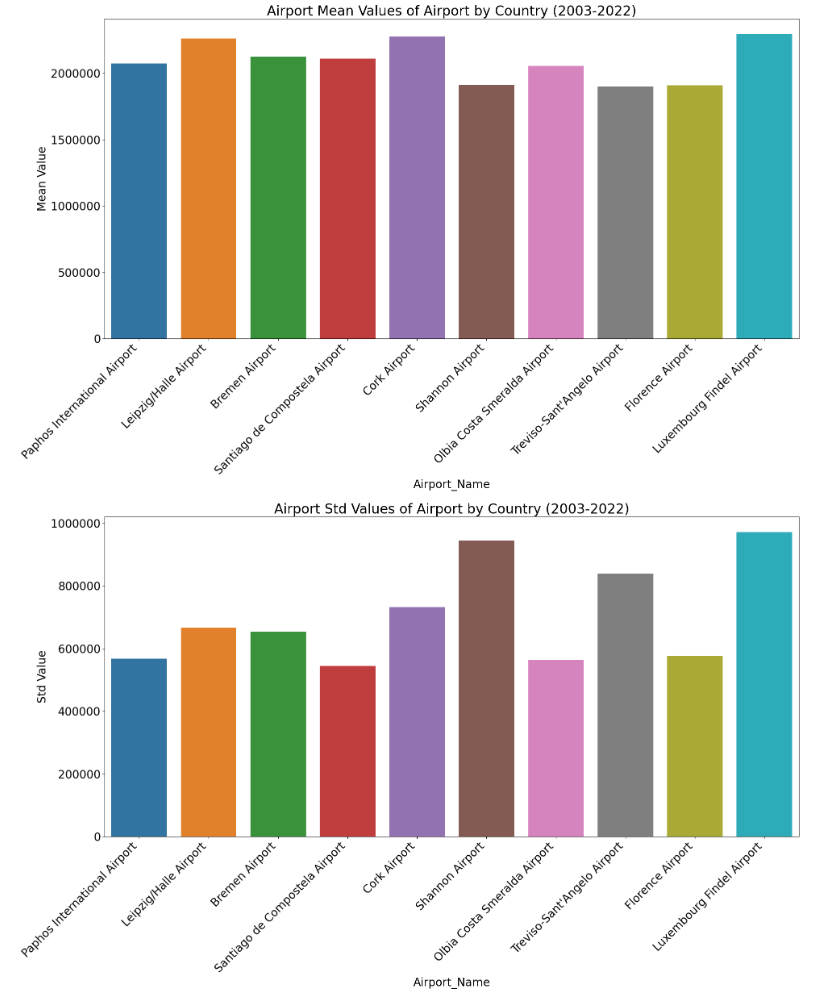
# *Dataset for ANOVA One-Way*

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

From the tests I have performed above we can find interesting facts:

*“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

*“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

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:

A diagram of a data processing process

Description automatically generated

# Supervised Learning, GridSearchCV Hyperparameter Tunning.

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.

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

A screenshot of a computer

Description automatically generated

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

Overall model is performing well however for Gamma = 0.0001, models seem to be overfitted 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

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

A close-up of a number

Description automatically generated

A screenshot of a graph

Description automatically generated

The results are as follows:

A screenshot of a computer

Description automatically generated

A graph of a number of numbers

Description automatically generated with medium confidence

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

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

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

A screenshot of a computer

Description automatically generated

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

A screen shot of a number

Description automatically generated

*KMeans* and *PCA* have performed better with more data.

# 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

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

A screenshot of a graph

Description automatically generated

Ryanair figure X

A screenshot of a computer

Description automatically generated

USA Airlines

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

A graph of a curve

Description automatically generated

*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

Result:

A graph of different sizes and colors

Description automatically generated with medium confidence

*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

Random Forrest:

A screenshot of a graph

Description automatically generated

KNN:

A graph of different colored lines

Description automatically generated with medium confidence

GridSearchCV CM:

A screenshot of a graph

Description automatically generated

Gamma Accuracy:

A graph of different values

Description automatically generated

Accuracy Table:

A table with numbers and text

Description automatically generated

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 objects

Description automatically generated with medium confidence

Summary table:

A screen shot of a computer

Description automatically generated

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

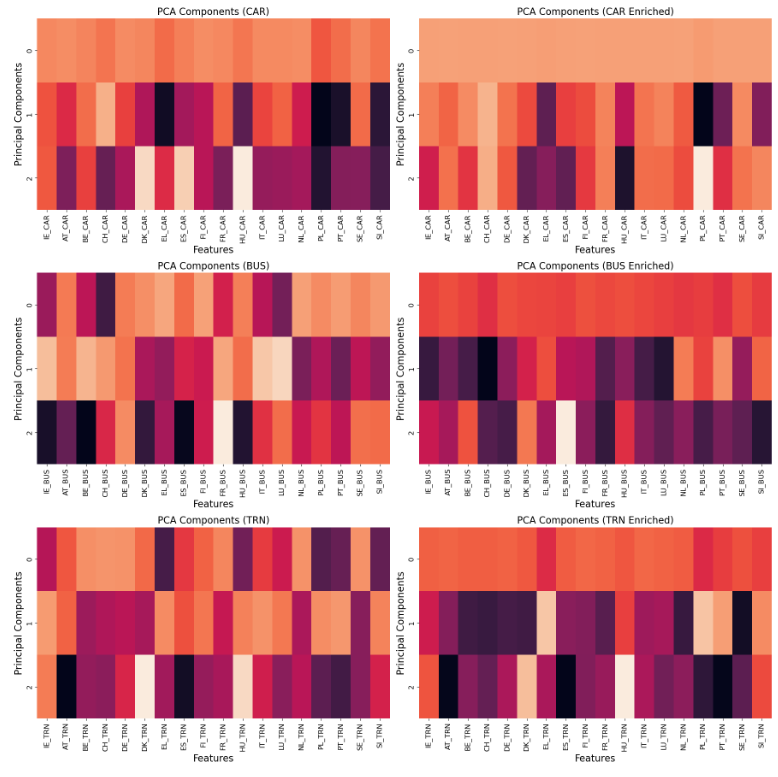
Description automatically generated

PCA Scatter Plot:

A group of colored dots

Description automatically generated

PCA Heatmap:



Elbow Method:

A graph of a number of different numbers

Description automatically generated with medium confidence

Silhouette Score:

A group of graphs showing different sizes of clusters

Description automatically generated

Tables PCA and Silhouette Score:

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

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 acquistion

# EDA methodology

# Visualisations

# Dashboard

# **Conclusion**

# **References**

Statistics:

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.

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.