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:

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:

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

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:

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

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:

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:

Scenario 2 accepting H0:

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:

Scenario 2 rejecting H0:

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:

Scenario 1, rejecting H0:

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:

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:

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:

Hypothesis:

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

Results:

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.

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:

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.

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.

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.

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:

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

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:

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

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.

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.

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.

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

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

Scenario 1:

We don’t have normality for our samples:

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.

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:

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

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

Scenario 1:

Checking normality:

Hypothesis:

H0: μ Ireland = μ Croatia.

H1: μ Ireland != μ Croatia.

Result p-value = 0.3068

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:

Hypothesis:

H0: μ Ireland = μ Slovakia.

H1: μ Ireland != μ Slovakia.

Result p-value = 3.45e-06

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.

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”* based on the confidence intervals, we can confirm that the number of passengers is recovering:

*“Paris Charles De Gaulle recorded the highest number of air passengers” “Dublin Airport Was EU’s 11th Largest Airport in 2018”*

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

The challenges I faced was gathering the data and stablishing the scenarios to perform the tests.

The methodology for ML part can be seen as follows:

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:

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

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:

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

The results are as follows:

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

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

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

Linear regression estimation has performed much better with enriched data.

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

First attempt with quarterly data we get the following results:

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

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

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

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

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

*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:

Result:

*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.

Let us compare each model after enriching the dataset.

Decision Trees CM:

Random Forrest:

KNN:

GridSearchCV CM:

Gamma Accuracy:

Accuracy Table:

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.

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

Summary table:

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

PCA Variance:

PCA Scatter Plot:

PCA Heatmap:

Elbow Method:

Silhouette Score:

Tables PCA and Silhouette Score:

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.