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.

For this project I gathered and process data from:

CSV file:

Web API in CSV format:

Eurostat module:

Also, I explored gathering data from a:

Web API in JSON format:

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

MySQL database:

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”*:

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

Complete instructions on creating the table, pushing, and fetching records are provided in the annex.

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. 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.

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:

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:

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:

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.

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 analyzing different data sources, I've utilized various libraries and techniques for both processing and aggregating data. 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”.*

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:

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.

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

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:

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

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:

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

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.

Scenario 2: Accepting H0.

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.

Scenario 2: Rejecting H0.

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.

Scenario 1: Rejecting H0.

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:

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:

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.

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 is not enough evidence to conclude that the weekly mean values for the LUAS green line are different from those of the red line.

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:

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.

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.

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.

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.

To perform ANOVA, we need both Shapiro-Wilk and Levene tests to have p-values 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 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:

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 with no association or relationship.

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.

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.

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

Scenario 1:

We do not 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 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:

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 means of annual passengers carried (2003-2022) for Kerry, Karlsruhe/Baden, and Rzeszow-Jasionka airports are different.

Checking normality:

Hypothesis:

H0: μ Ireland = μ Croatia.

H1: μ Ireland != μ Croatia.

Result p-value = 0.3068

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:

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.

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:

*“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.):

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.

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:

*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.

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, and *TripAdvisor* does not state that you cannot collect their reviews for research purposes.

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:

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:**

**Monthly enrichment:**

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).