# Group ID - MSc in Data Analytics

* Author: J. Doyle
* e-mail: sba23437@student.cct.ie
* Student ID: sba23437

# CA2: Integrated Assessment: Transport in Ireland

## Introduction

This is an investigation into Irish transport and specifically commuting times in Ireland. We have used datasets published by the Irish Central Statistics Office, the European group Eurostat and the global organisation OECD which provide us with total population estimates with demographics. We’ve selected related datasets all related to commuting local, regionally and internationally.

Employing best practices to explore, prepare, transform and utilise large data repositories, we have programmatically developed a step-by-step python solution which describes for us key insights from the data.

A Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was employed in the execution of the project as a cross-industry standardised and proven methodology (Chapman et al, 2000). This gives us a standard process and framework for the typical phases and tasks needed for a modelling implementation.

Our first steps in the project were to apply descriptive and inferential statistics to understand our data as well as to help us plan how to investigate it further in our Machine Learning and Visualisation sections.

We have taken the data and applied sentiment analysis with machine learning classification methodologies in python helping us to further examine these fascinating insights and to understand potential insights for our populations.

The work was accomplished using the programming language Python as an open source, quick to deploy and one of the most popular languages available for this type of undertaking (Liebowitz 2013, p261). It’s Open source and “ideal for computationally-intensive applications and general purpose systems” (McKinney, W., 2013). The Jupyter Notebook environment was employed for our development environment. It too is open source and compatible with Python.

Finally, the output from the project is an interactive and simple to use prototype dashboard using the python Dash library. This type of dashboard has the potential to inform users as they plan their lives. Recently a comprehensive Irish Transportation dashboard was published by the Irish CSO on their public website, indicating how infographics and interactive dashboarding can be used by scientific and statistical groups to help citizens plan, make policy and investigate by using reliable, valid and accessible data (CSO, 2023).

## Data Exploration and Preparation

In the CRISP-DM framework, is an iterative framework for data mining, modelling and deployment endorsed across many business areas. Stepping through the main phases of this planning framework became the template for this project, with specific focus on the iterative aspects needed in developing this final submission. In the following visual, we can see the process starts with business understanding, however, while this informs the next phases of Data Understanding, Data Preparation, and Modelling these phases still iteratively impact the business understanding and each other. The benefits of approaching data projects like this are that it provides an agile methodology that lends itself to prototyping and learning as you go while discovering new and unexpected data and environmental constrains. The downside to an agile approach like this however is that it can lead to scope creep, an unending loop with incremental improvements for diminishing gains. Careful planning and project planning can offset these risks throughout an implementation however.

A diagram of data mining life cycle

Description automatically generated

Looking into the development required for this solution, the first stage was to setup the required libraries with the necessary functions for the analysis, visualisation, statistics and machine learning functions:

## Statistics

1. Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. **[0-20]**
2. Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with public transport, you could find a confidence interval for the population proportion of users commuting to Dublin by train). **[0-20]**
3. Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. **[0-40]**
4. Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. **[0-20]**

# Discussion & Reflections for CA

## Programming Discussion

### Programming Section 1, 2 and 3

For programming tasks 1, 2 and 3, please refer to the 3 Jupyter notebooks attached to this submission.

For task 2, please note the use of the use of CSV formats in the main data acquisition notebook called ‘CA2\_Data\_Acquisition&Stats’ whereas APIs in JSON format are used in the notebook called ‘CA2\_DUBLIN\_TRANSPORT\_REDDIT\_Sentiment\_Analysis&ML Techniques’.

### Programming tasks 4 and 5

**Testing & Optimisation:** You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas. **[0-20]**

**Data manipulation:** For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques. **[0-20]**

## Machine Learning Discussion

Use of multiple models (at least two) to compare and contrast results and insights gained.

* Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

**[0 - 30]**

* Collect and develop a dataset based on the transport topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate transport topic (e.g., public transport, freight movement etc…) for producers and consumers point of view in Ireland.

**[0 - 25]**

* You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

**[0 - 30]**

* A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

**[0 - 15]**

## Data Preparation & Visualisation Discussion

* Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data. **[0-15]**
* Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.**[0-25]**
* Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.**[0-30**]
* Modern Transport planning has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern Transport planning, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development. **Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).[0-30]**

# References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide,* SPSS Inc.

Tufte, E. (2001) *The Visual Display of Quantitative Information, (2nd ed.)*, Graphics Press

Liebowitz, J. (2013) *Big Data and Business Analytics,* CRC Press

McKinney, W., (2013) Python for Data Analysis. Sebastopol: O’Reilly Media, Inc. pp. 1-3.

CSO (2023) The Transport Statistics Hub: <https://lnkd.in/eHJNzJky>

[Choosing the right estimator — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html)

[1.4. Support Vector Machines — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/modules/svm.html#tips-on-practical-use)

[Plot different SVM classifiers in the iris dataset — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html#sphx-glr-auto-examples-svm-plot-iris-svc-py)

NTA118 - Distribution of journeys (Trrans method with age)

<https://data.gov.ie/dataset/nta118-distribution-of-journeys>

[E6028 - Population Aged 15 Years and Over at Work, Usually Resident and Present in the State 2011 to 2016 (cso.ie)](https://data.cso.ie/)

E6028

Last Updated: 10/07/2020 11:00:00

Population Aged 15 Years and Over at Work, Usually Resident and Present in the State 2011 to 2016

[Time spent commuting by men and women in OECD countries 2016 | Statista](https://www.statista.com/statistics/521886/travel-time-spent-work-study-countries/)

Published by [Statista Research Department](https://www.statista.com/aboutus/our-research-commitment), Mar 7, 2016

Average minutes per day spent travelling to paid work or study by men and women in OECD countries plus China, India and South Africa, as of 2016

[Majority commuted less than 30 minutes in 2019 - Products Eurostat News - Eurostat (europa.eu)](https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20201021-2)

Average European commuting Times: 25 Mins

5 Inferential Stat Tests

T-test AND U-Mann Whitman (2 Means)

ANOVA (1-way) AND ANOVA (2-way) AND Kruskal-Wallis (3 Means)