***Scenario***

***Transport in Ireland***

*“Smartcard ticketing systems currently available in many cities’ public transport systems enable a vast amount of data.*

*These datasets reflect how people behave, which in turn means it is possible to assess their transport needs and provide an accurate picture of their habits, either as groups (based on fare types, such as student, elderly and so on) or at an individual level.*

*Thanks to big data analytics, we can adapt public transport to these needs, plan new services, minimise walking time and so on.”—* *https://www.engineersireland.ie/Brexit/big-data-enables-better-urban-transport-networks*

**You have been tasked with analysing Ireland's Transport data and comparing the Irish Transport sector with other countries (of your choosing) worldwide.** This analysis should also include forecasting, sentiment analysis and evidence-based recommendations for the sector as well as a complete rationale of the entire process used to discover your findings. Your Research could include vehicle traffic, type of vehicles used, freight transport, air traffic, transport infrastructure etc. (or any other relevant topic EXCEPT Climate change) with Ireland as your base line.

**Note:**

* **While topical, Construction impact on Climate Change SHOULD NOT be chosen as an area of research for this assessment.**

You must source appropriate data sets from any available repository (eg. But not restricted to https://data.gov.ie/organization/transport-infrastructure-ireland) to inform your research (all datasets MUST be referenced and the relevant licence/permissions detailed).

***Criteria of Analysis***

Discuss the choice of project management framework you have deemed suitable for this project.

It is expected that you use some type of version control software eg: GitHub, Gitlab, BitBucket etc with regular commits of code and report versions. **Please include the address of your version control repository in your report.**

**Programming for DA Tasks [0-100]**

1. **Programming:** The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook.
2. **Data structures:** You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format.
3. **Documentation:** The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied.
4. **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 (e.g. computing, time etc). Note any trade-offs that you've made in these areas.
5. **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.

**Statistics for Data Analytics Tasks**

* Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models.
* 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).
* 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.
* Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process.

*Note: All your calculations and reasoning behind your models must be documented in the report and/or the appendix.*

**Machine Learning Tasks**

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.
* 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.
* 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.
* 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.

**Data Preparation & Visualisation Tasks**

* 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.
* 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.
* 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.
* 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).**

**All documentation, code, examples, and any other files MUST be evidenced in your Version Control repository. (Git or Similar) using your CCT email address ONLY. Your repository MUST show your continual development of the project throughout its lifecycle and include notes detailing progress at each commit.**

**You may be called to a Viva to defend your work.**

**Additional notes :**

All:

* Your documentation should present your approach to the project, including elements of project planning ( timelines).
* Ensure that your documentation follows a logical sequence through the planning / research / justification / implementation phases of the project.
* Ensure that your final upload contains a **maximum of 1 jupyter notebook per module task.**
* Please ensure that additional resources are placed and linked to a logical file structure eg, Scripts, Images, Report, Data etc…
* Ensure that you include your raw and structured datasets in your submission
* 3000(+/- 10%) words in report (not including code, code comments, titles, references, or citations)
* Your Word count MUST be included

**Plagiarism and/or inclusion of AI (CHATGPT, BARD etc…) is absolutely, wrong.**

All assessment submissions must:

* Jupyter Notebook (.ipynb), Word Document ONLY, Dashboard, and version control address
* Use [Harvard Referencing](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) when citing third party material