**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. **[0-20] – DONE**
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.**[0-20] – CSV file will be downloaded from the site, for the sentimental analysis a web API will be used – Twitter**
3. **Documentation:** The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied. **[0-20] – DONE WITHIN WORKING**
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 (eg computing, time etc). Note any trade-offs that you've made in these areas. **[0-20] – NOT SURE IF THIS SHOULD BE UNIT TESTING – ASK**
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. **[0-20] – COMPARISON OF LIBRARIES**

**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. **[0-20] – WILL BE EASILY DONE**

●        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] – FOR THIS CHECK LECTURES – ESPECIALLY 25th of NOV**

●        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] - FOR THIS CHECK LECTURES – ESPECIALLY 25th of NOV AND BEFORE THAT**

●        Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. **[0-20] – CHECK THIS**

*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. **– ML MODELS**

**[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. – **TWITTER**

**[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. - **ML MODELS**

**[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. – **COMPARISON OF THE ML MODELS**

**[0 - 15]**

**Data Preparation & Visualisation Tasks**

●        Discuss in detail the process of acquiring your [raw data](https://moodle.cct.ie/mod/resource/view.php?id=143484), 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] – TWITTER Dev and official sites**

●        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](https://moodle.cct.ie/mod/resource/view.php?id=143484) 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] – EDA AS USUAL**

●        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**] – **EDA AGAIN**

●        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] – INTERACTIVE DASHBOARD**

Gathering data:

<https://www.cso.ie/en/releasesandpublications/ep/p-tranom/transportomnibus2021/data/>

ML:

I can apply the time series to the data I have with months for the Luas traffic – code can be found in the lecture on Saturday