## Appendix 1: Examination Criteria

Programming for DA Tasks

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.

The entirety of my Jupyter notebooks can show evidence of this.

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.

I have gathered a processed data from both CSV, GeoJSON and JSON format.

3) Documentation: The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied.

Documentation is provided in my report and also within my notebooks.

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.

I have often discussed processing and computing time within my Jupyter notebook, see section 3.7 and modelling section 2.8 and 2.9.

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.

I’ve often compared techniques such as interactive and non-interactive plotting,

Statistics for Data Analytics Tasks:

1) Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models.

Number visualisations were used to help justify my choice of Regression model, ultimately leading me to favour the Random Forrest regression model.

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).

Please see section 2.7.

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.

I have used over 5 statistical tests throughout my project.

4) Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process.

I was able to use the outcomes of the tests I performed in Section 3: UN data to research specific cities which were comparable to Dublin. Unfortunately data availability was a big issue here.

Machine Learning Tasks

1) 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.

Please see sections 2.8-2.10 and Sections 4.

2) 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.

See section 4.

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.

See sections 2.8-2.10.

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.

See sections 2.8-2.9. Please see commentary in Notebooks also.

Data Preparation & Visualisation Tasks

1) 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.

Discussed throughout the report, discussed issues when trying to get LinkedIn data.

2) 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.

Discussed and performed throughout report and notebook.

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.

See section 4 for data preparation for sentiment analysis. Also present in other sections.

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).

Made use of Dash library and made number interactive maps. See section 1 and section 3.13.