# Group ID - MSc in Data Analytics

Author: Conor McKenna

e-mail: [sbs23014@student.cct.ie](mailto:sbs23014@student.cct.ie)

Student ID: sbs23014

Table of Contents

[Group ID - MSc in Data Analytics 1](#_Toc135127163)

[Abstract 2](#_Toc135127164)

[Introduction 2](#_Toc135127165)

[1. Data Preparation & Visualisation Tasks 3](#_Toc135127166)

[Acquiring Raw Data 3](#_Toc135127167)

[Exploratory Data Analysis 3](#_Toc135127168)

[Data Cleansing and Preparation 3](#_Toc135127169)

[Interactive Dashboard 4](#_Toc135127170)

[2. Statistical Analysis 4](#_Toc135127171)

[Descriptive Statistics 4](#_Toc135127172)

[Analysis of Variables 4](#_Toc135127173)

[Inferential statistical techniques for comparison 4](#_Toc135127174)

[Outcome to deepen research 4](#_Toc135127175)

[3. Machine Learning Tasks 5](#_Toc135127176)

[Rationality of choice of machine learning models 5](#_Toc135127177)

[Webscraping and sentiment analysis 5](#_Toc135127178)

[Testing models that were developed 5](#_Toc135127179)

[4. Programming for DA Tasks [0-100] 5](#_Toc135127180)

[Programming 5](#_Toc135127181)

[Data structures: 6](#_Toc135127182)

[Documentation: 6](#_Toc135127183)

[Testing & Optimisation: 6](#_Toc135127184)

[Data manipulation: 6](#_Toc135127185)

# Abstract

*There is a lot of pressure on Ireland’s construction sector, There are a lot of published datasets surrounding the construction industry including numerous*

# Introduction

This project uses a dataset published by The Department of Housing which contains applications for planning permission in Ireland between 1987 and 2023. The goal of the project is to assess whether statistical and machine learning methods can be applied to predict whether an application for a “one off house” was granted or refused. The dataset itself comes in the form of a CSV file, and requires exploratory data and statistical analysis and visualisation to help gain insight in data quality and usability.

# Data Preparation & Visualisation Tasks

## Acquiring Raw Data

*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]*

Eurostat is an online resources which offers datasets and insights into EU countries.

## Exploratory Data Analysis

*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]***

## Data Cleansing and Preparation

*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]*

## Interactive Dashboard

*Modern construction 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 construction industries, 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]***

# Statistical Analysis

## Descriptive Statistics

*Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. [0-20]*

## Analysis of Variables

*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 international home building, you could find a confidence interval for the population proportion of yearly apartment builds out of all home builds).* ***[0-20]***

## Inferential statistical techniques for comparison

*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]***

## Outcome to deepen research

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

# Machine Learning Tasks

## Rationality of choice of machine learning models

*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]***

## Webscraping and sentiment analysis

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

***[0 - 25]***

Sentiment analysis done but I need to build out the output a bit more. Maybe try to incorporate Bigrams and identify themes in the comments that way to find the comments most relevant to my topic.

## Testing models that were developed

*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]***

***Graphics and Table that demonstrates the ML outcomes***

*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]***

# Programming for DA Tasks [0-100]

## Programming

***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]***

## 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]***

## Documentation:

*The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied.* ***[0-20]***

## 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]***