**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *MSc in Data Analytics*   * *Programming for DA* * *Statistics for Data Analytics* * *Machine Learning for Data Analysis* * *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_Integr\_CA2\_Sem1*  **Scenario:** Irish Construction sector |
| **Lecturer Name:** | *Sam Weiss*  *John O’Sullivan*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Jose Luis Alvarado Carranza |
| **Student Number:** | 2020092 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

# Group ID - MSc in Data Analytics

Author: Jose Luis Alvarado Carranza

e-mail: [2020092@student.cct.ie](mailto:2020092@student.cct.ie)

Student ID: 2020092

Introduction

Construction is the process of creating a plan and adding structure to real property or construction of buildings. In the modern industrialized world, construction usually involves the translation of designs into reality. However, understanding and estimating construction costs can be challenging. Luckily for us, there are a lot of data sets from all over the world to pull out data from.

The field of construction is extremely vast and there are many variables to consider when developing a project. Therefore, the focus of our project will revolve around Residential Property Price Index (RPPI). By adhering to this topic many of the other topic get considered and funnelled into a single variable, making it more comprehensible to the reader.

We begin by defining what cost index is and how it is calculated to better interpret our data.

BACKGROUND

The Residential Property Price Index (RPPI) is designed to measure the change in the average level of prices paid by households for residential properties sold in Ireland. The RPPI specifically excludes non-household purchases, non-market purchases and self-builds (i.e. where the land is purchased separately). The index is mix-adjusted to allow for the fact that different types of property are sold in different months. (CSO, 2023)

Now that we have defined our variable, let us find out how to calculate it.

The RPPI can be used to estimate the updated value of an individual dwelling provided that a prior value is known subsequent to January 2005. Simply multiply the sale or valuation price by the current relevant index and divide by the index at the date of sale/valuation.

For example, consider a house sold in South Dublin in June 2010 for €220,000.

      South Dublin house price index June 2010:                           88.9

      South Dublin house price index currently:                            121.5

      Estimated current value:                                                      €220,000 x (121.5 / 88.9)   =   €300,675

This estimate provides an approximate value only, based on aggregate price movements of all dwellings. It presupposes that there has been no material change in either the dwelling or its neighbourhood in the period concerned. The CSO takes no responsibility for calculation of value based on the RPPI. (CSO, 2023)

Before delving into processing the data, a methodology has been put in to place to provide structure to the overall project. We have decided to choose SEMMA project management because we don’t need to have a lot of knowledge on construction to draw our own conclusions, after the data has been processed, to gain an advantage in the market.

As it has been established before, the topic of construction has many aspects to it. It can be the cost of construction from the commercial, residential, industrial, a country’s infrastructure, etc. We are going to consider the residential aspect of construction. According to the brief in the dataset “ The House Building Cost Index monitors labour costs in the construction industry and the cost of building materials. It does not include items such as overheads, profit, interest charges or land development. The labour costs include insurance cover and the building material costs include V.A.T. Coverage:  
The type of construction covered is a typical 3 bed-roomed, 2 level local authority house and the index is applied on a national basis”. (Government, 2021).

Stage 1: Sample

We first begin by gathering data from the recommended website.

After the proper EDA has been performed, we have gathered a useful insight into our topic.

It is worth mentioning that we have chosen the year 2015 as it is the latest usable information available in this data set.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure Histogram of the Cost of construction in 2015

This is the histogram of the RPPI for the year 2015 in Ireland. From this figure we can gather that:

* The cost of construction in Ireland increased and reached the highest index of 207.6.
* The cost of construction in Ireland had the lowest index of 206.0.
* There is a variation in the index of 1.6 points within a year.

A picture containing screenshot, text, rectangle, square

Description automatically generated

Figure Boxplot of the Index in 2015

This is the boxplot of the RPPI for the year 2015 in Ireland. From this figure we can gather that:

* There are no outliers.
* The median value is of 207.2.
* The value of 206.8 represents Q1 of the data and the value of 207.4 Q3 while Q2 is equal to 207.2.

A picture containing text, screenshot, line, parallel

Description automatically generated

Programming notes

Task 2: data found in .csv files is easy to manipulate using python but data using an API in JSON format is tricky to manipulate because of the limited number of requests and the the time it takes to establish a connection

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.

We have been able to get plenty of data

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).

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.

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

Develop an interactive dashboard tailored to modern farmers, 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.