**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_CA2* |
| **Lecturer Name:** | *Sam Weiss*  *Marina Iantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Liam Croash |
| **Student Number:** | sbs23036 |
| **Assessment Due Date:** | 26/May/2023 |
| **Date of Submission:** | 26/May/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: Liam Croash

e-mail: [sbs23036@student.cct.ie](mailto:sbs23036@student.cct.ie)

Student ID: sbs23036

## Abstract

*This report examines…..*

## Introduction

This report examines …

## Methodology

### 1. Data preparation and Visualization

#### Data Preparation:

The data ingestion phase was concerned with sourcing, acquiring, ingesting and cleansing of multiple datasets relating to the construction industry. Data was required to perform analyses to compare the rate of new homes being built in Ireland (IE), England, Scotland, Wales and Northern Ireland. *(add citations)****.*** In discovering information for IE, there were a number of datasets available providing new build volumes at various levels of aggregation including County, Rural or Urban, Type of House & with Seasonal and non-Seasonal adjustments made to align with Calendar quarters and all are available under a Creative Commons Attribution licence *(add citation)* requiring a citation and reference to allow for any use.For the purposes of this project, files were downloaded and processed in CSV format.

Data for new builds in England, NI, Wales & Scotland was more difficult to acquire in a raw format. The difficulty lay in tha fact that the initial data that was sourced was normalised into a index rather that in the raw count format that was desired to allow for comparison to IE. I could not find any quantitative information to detail the process of how the index was calculated so this data was discarded. Data was located that counts the number of completed houses in each of the UK countries in total, by housing type and broken down by volume started per quarter vs volume completed by quarter. This data was downloaded and processed in an Excel spreadsheet (XLSX) format under the UK Open Government Licence v3 *(add citation)* which necessitates citation and has additional exemptions related to personal data or others.

Given that the underlying theme of the analysis is understanding the Housing Crisis, additional sources of data for population were sourced.

For IE, this was sourced from the World Bank using their Total Population estimates. This was downloaded and processed in XML format and is available under the Creative Commons Attribution licence described above. Unfortunately, the World Bank did not provide a country level breakdown of the UK, so this data was sourced from the ONS.giv website *(add citation)* under the Open Government Licence v3. UK data was based on yearly estimates in June (Q2) of each year and broke down totals by country and overall.

The final set of data required was text data to interrogate in order to understand the sentiment of the public towards the Housing Crisis. The initial approach was to leverage the free Twitter API to search for #housingcrisis. However, due to changes in Twitter API licencing and cost, it was prohibitive to use the available API. Following attempts to bypass using external company APIs, web scraping libraries and even attempting to build a web scraper using Beautiful Soup, there was one blocker that could not be bypassed which was that Twitter uses an scrolling load mechanism rather than pages. This put web scraping outside the scope of this project. The solution was to use a Chrome extension Instant Data Scraper *(add citation)* to recover the data. Then the text was manually assembled using pandas and the CSS tags within the HTML that was scraped. Twitter offers a Developer Agreement and Policy (*add citation)* that allows for unrestricted use except in certain cases outlined in except in certain cases and also affords Twitter the right to scrape the content of licensees anywhere Twitter data is used.

#### Exploratory Data Analysis

For IE housing data, the data was complete with no missing values, significant outliers that implied errors in data. There were, however, a number of duplicate columns in the data.

#### Feature Preparation & Engineering

### 2. Statistical Analysis

### 3. Machine Learning

### 4. Programming

## References

Creative commons <https://creativecommons.org/licenses/by/4.0/>

Open Government Licence <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>

IE 1 # <https://data.gov.ie/dataset/ndq01-new-dwelling-completions/resource/8fdc8f87-88ca-43f7-aa16-dbc3a52cacde>

IE 2 **#** [**https://data.gov.ie/dataset/ndq03-new-dwelling-completions/resource/45701f7f-1192-406f-980f-a8ee23862147**](https://data.gov.ie/dataset/ndq03-new-dwelling-completions/resource/45701f7f-1192-406f-980f-a8ee23862147)

**IE 3 #** [**https://data.gov.ie/dataset/ndq05-new-dwelling-completion/resource/108d82d0-43f1-4574-9e47-bb54f078fd84**](https://data.gov.ie/dataset/ndq05-new-dwelling-completion/resource/108d82d0-43f1-4574-9e47-bb54f078fd84)

IE 4 # <https://data.gov.ie/dataset/ndq06-new-dwelling-completion/resource/0b5041bb-9f52-4b54-8c8d-0000306e74e2>

UK: # <https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/ukhousebuildingpermanentdwellingsstartedandcompleted>

Pop: # <https://data.worldbank.org/indicator/SP.POP.TOTL>

UK Pop <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatestimeseriesdataset>

Twitter: [www.twitter.com](http://www.twitter.com)  
Licence: <https://developer.twitter.com/en/developer-terms/agreement-and-policy>

Twitter scraping <https://aitechtrend.com/how-to-scrape-twitter-data-without-using-twitters-api/>

<https://os2int.com/toolbox/scraping-twitter-without-an-api-or-user-account-with-tweeds/>

Instant Data Scraper <https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah>