**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| Module Title: | MSc in Data Analytics |
| Assessment Title: | CA2 |
| Lecturer Name: | Sam Weiss  Marina Iantorno  Muhammad Iqbal  David McQuaid |
| Student Full Name: | Glenn Tarpey |
| Student Number: | SBS23066 |
| Assessment Due Date: | 09th June 2023 |
| Date of Submission: | 09/06/2023 |

**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. |

## Abstract

## Sections

[Abstract 2](#_Toc136613756)

[Sections 3](#_Toc136613757)

[Figures 4](#_Toc136613758)

[Introduction 5](#_Toc136613759)

## Figures

**No table of figures entries found.**

## Introduction

In 2021, the size of the building market in Ireland was $31.8 Billion. In the coming years 2023–2026 the market is expected to grow at an Average Annual Growth Rate of at least 4%. This will be supported by investments in energy, transportation, and housing. The government intends to invest in projects related to energy and transportation infrastructure through the National Development Plan. (Global Data, 2022)

As compared to the overall European construction market which is currently valued at approximately $3.01 Trillion and expected to increase to 3.73 Trillion by 2028 it is clear that the construction industry is a rapidly expanding market with increasing scopes of work. In recent years the construction industry have started adopting business intelligence practices to further identify opportunities, cost savings and to maintain a strategic advantage over competitors. (researchandmarkets.com, 2023) (datashapa.com, 2022)

AI and Machine Learning are increasingly utilized in large scale construction projects for tasks such as project design, installation, data collection, and analysis, all with the aim of optimizing construction operations.

This report aims to explore this market, both nationally in Ireland and also in Europe in the goal of identifying trends, predictions and recommendations through the use of data analytics.

## Data Sources

The study and analysis of the construction sector, primarily focusing on Ireland as the primary case study alongside other countries in Europe, involves easily collecting data from government data sources and market research data. However, it is important to justify the choice of data and understand how crucial it is to comprehend the construction sector, observe patterns, trends, and make accurate forecasts.

In this report data from Eurostat will be used to analyse various aspects of the industry, including construction output, labour input, and gross value added from construction. These data sources will help to provide a good overview of the European construction industry. Eurostat is well regarded as a provider of trusted construction data as well as other industries. (Eurostat, 2021)

The data types being used in this report are:

* Construction Output (Building and Civil Engineering) indicates the value of production.
* Building Permit.
* Labor Input (Wages and Salary, Hours worked).
* Structural business statistics, including the Number of enterprises involved in construction, Multi-year Subcontracting, and Turnover.
* Gross Value Added from Construction and GDP from Construction, mainly encompassing Gross Fixed Capital Formation (to comprehend investments and returns).

## Data Explanation

Data was primarily sourced from Eurostat, categorized into two main categories:

1. Short-term business statistics (STS) provide information on various economic activities, including production, building permits indicators, producer prices in construction, and labour input indicators.
2. Structural business statistics (SBS) offer detailed insights into the structure, conduct, and performance of economic activities across several sectors. Key data include business demographics, output-related and input-related variables.

This data was collected in excel/csv format. It's used to analyse several aspects of the construction sector such as GDP share, producing values, turnover, challenges, costs, and exporting services.

Gross fixed capital formation (GFCF) data was considered to understand the economic impact of the construction sector. It refers to the investments made in fixed assets by resident producers over a specific period, minus disposals. This data, in Million Euros, is sourced from Eurostat's National Accounts.

The production value index, which represents the output and activity of the construction sector, was taken from Eurostat's short-term business statistics. The index measures changes in the volume of output on a monthly basis and includes building construction and civil engineering.

Building permits and dwelling data were also considered. A building permit authorizes the start of a building project, and an index based on these permits provides an indication of future workload for the building industry. Dwelling data refers to the rooms or suite of rooms intended for private habitation.

Labour inputs include the total number of people working in the unit (employment size), total hours worked, and total wages and salaries. These aim to show the volume of work done and the development of the wage and salaries bill.

Finally, Structural Business Statistics provided information on turnover value added, the number of enterprises, employment size, and subcontracting.

## ETL Process – Construction Data

A total of 12 construction datasets were used in this assignment. A variety of functions and techniques were performed to extract, manipulate and save these datasets to a new folder. To summarize this process an overall summary of the data cleaning process will be given and a summary of the functions used will be provided.

Overall steps:

1. Extract the data into variables to store them
2. Data Summary to check for missing value, info, shape, and summary statistics.
3. Transformation of the extracted data through a function to remove unnecessary values, change datatypes, and convert country codes to the country name.
4. Load the data into a folder named Transformed Data.

Functions:

* Data\_summary: This function accepts a file name, it reads the data from this file into a data frame and prints a variety of information about the data such as columns, shape and null values. This function was used to understand the datasets and identify any issues with them early on.
* Get\_country\_name This function takes the alpha numeric shorthand country name used in Eurostats datasets and converts it into it’s full English spelling e.g. IE to Ireland. This allowed the dataset to be easier to read and understand.
* Clean\_and\_transform: This function was the most complex in this process. It takes custom parameters such as columns to drop (irrelevant columns) and a value name to replace the common “OBS\_Value” in the Eurostat datasets with a more meaningful name, making the datasets easier to work with and understand.

## ETL Process – Sentiment Analysis

As part of this report data was scraped using the Reddit API to gain insights into positive and negative viewpoints of construction topics. Like the construction data, this data also needed to be transformed.