# Group ID - MSc in Data Analytics-CA2

Author: Osuade Adereti

e-mail: [sbs23074@student.cct.ie](mailto:sbs23074@student.cct.ie)

Student ID: 23074

Address of repository: https://github.com/osuade

**REPORT ON DATA ANALYTICS OF IRISH CONSTRUCTION MANAGEMENT IN COMPARISON WITH OTHER COUNTRIES WORLDWIDE**

## Abstract

The Plans of Action of this report are basically:

-Collection, processing, analysing and interpretation of dataset gotten from data.gov.ie on CSO website, ons.gov.uk and Kaggle dataset from US Census Bureau on construction data to identify issues/problems at present and make predictions/classification to compare Irish construction data with other countries worldwide using Python Programming Language to analyse the data and interpret the results and building Machine Learning (ML) Algorithms for the dataset.

-To compare the correlation between the construction data collected in Ireland with other countries worldwide (UK and USA) to deduce if there is high correlation or strong relationships among them.

-To conduct forecasting analysis and sentimental analysis and evidence-based recommendations for construction sector in Ireland, UK and USA

-To conduct research on building productions in construction industry in Ireland and other countries like UK and USA.

## TABLE OF CONTENT AND FIGURES

Abstract 2

Table of content and figures 3

Section 1 - Introduction 4

Section 2 - Materials and Methods 5

Section 3 - Criteria of Analysis 6

3.1 – Programming 7

3.2 – Statistics 8

3.3 - Machine Learning 8

3.4 - Data Preparation & Visualisation 9

Section 4 - Results and Interpretations 10

Section 5 – Conclusion 12

Section 6 – Acknowledgement 12

Section 7 – References 13

## Section 1- Introduction

The datasets for this project work were collected from Central Statistics Office (CSO) Ireland, UK and US Census Bureau. The construction data of Ireland were accessed on [www.data.cso.ie](http://www.data.cso.ie) while that of UK were accessed on [www.ons.gov.uk](http://www.ons.gov.uk) and USA are accessed on Kaggle dataset. The dataset on construction output in Northern Ireland are accessed on www.ons.gov.uk. This dataset showed the number of houses constructed privately and publicly over a number of years.

The Irish data showed the cost unit value of various constructions done from year 2000 to year 2022 with year 2015 as a base year consisting of five types of building which are as follows:

-All building and construction

-Building (excluding civil engineering works)

-Residential building

-Non-Residential

-Civil Engineering

The dataset is made up of 460 rows and 5 columns. They consist of numerical numbers, alphabets, integers strings and categorical data.

The UK data consists of 53 rows and 10 columns.

The layout of this report is as follows:

-Introduction of the datasets on construction using Python Programming Language.

This project was conducted with suitable Python tools (code and libraries) which are documented on Jupyter notebook.

-Statistical computations for data analytics are done using appropriate inferential statistics (As per attached Jupyter notebook) to gain insights on the values of the building in Ireland, UK and USA over the years to deepen the research work and justify the chosen models.

-Machine Learning models are developed for the predictions, classifications, clustering, sentimental analysis, and time series analysis of the construction datasets collected from Central Statistics Office Ireland, UK and US Census Bureau.

-Data Preparation and Visualisation - Appropriate Exploratory Data Analysis (EDA) was carried out on the dataset by inspecting the data to understand its characteristics better, uncover the patterns, relationships and insights, and identify potential issues or abnormalities before building reliable Machine Learning (ML) models and making sound decisions after cleaning the data and checking for missing data and null values. Python, Seaborn and pandas are used to perform EDA on the dataset.

The dataset are visualised to see the relationship among the variables using charts and graphs. It was done on jupyter notebook and by examining the data file structure after importing the Numpy, Pandas, Seaborn and matplotlib libraries taking into consideration the task required in the machine learning section to support the methodology with the visualisations done.

## Section 2 – Materials and Methods

The world has generated data in every field giving rise to big data. There is plenty of data everywhere and data storage is becoming critical at present. The importance of big data does not revolve around the availability of data, but the purpose of data.

Analytics is the science of examining raw data with the purpose of drawing conclusions about the information and using it for decision making.

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model by cleaning the dataset and checking for null values and see if it requires feature engineering to make it suitable to be used for machine learning. It is the first and crucial step while creating a machine learning model. To build and develop Machine Learning models, relevant dataset must first be acquired. This dataset will be composed of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. There is need to know which data type that will be dealing with to choose the right visualization method. The two main categories of data are numerical and categorical data.

Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labelled data. Even though the data needs to be labelled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances. Learning under supervision directly translates to being under guidance and learning from an entity that oversees providing feedback through this process. When training a machine, supervised learning refers to a category data learning to fit mapping between examples of input features with their associated labels. Where we train a machine learning algorithm using data while guiding the algorithm model with labels associated with the data.

The Construction Industry plays a significant role in the prosperity of the economy, contributing approximately $10 trillion to the global domestic product (GDP; Barbosa et al., 2017). In the U.S., the construction industry’s contribution reached more than $892 billion in 2020, which represented about 5% of the total U.S. GDP that year (BEA, 2020).

## Section 3 – Discussion on Criteria of Analysis

Project Management Framework for Data Mining Using CRISP-DM

CRISP-DM means Cross-Industry Standard Process for Data Mining. It was developed in 1996 in IBM for data mining tasks and it fits in general problem-solving business projects/research work.

This framework has the same cycle nature as both KDD and SEMMA. The key difference in the structure is that the transactions between stages can be reversed.

Data mining projects follow iterative, adaptive life cycle consisting of six stages as follows:

1. Business/Research Understanding Stage - It is crucial to understand the application of the product to be developed by:

-Determine business, background, and business success criteria.

-Assess the situation-thru requirements, assumptions, and constraints, risk and contingencies, terminology, cost, and benefit

-Determine Goals - Data mining goals and success criteria

-Produce project plan-Project plan and tools

B) Data Understanding - This consists collating and exploring the input dataset.

Stages involved are:

-Collect initial data

-Describe data

-Explore data

-Verify data quality

C) Data Preparation- Bad input leads to bad output. The activities involved here are:

select data, clean data, construct data, integrate data and format the

data into datasets and give some data description.

D) Modelling- This is an execution of all your findings from the previous stages by:

- Select modelling techniques

- Generate test design - Test design

- Build model - Parameter settings, model etc

- Assess model - Thru model assessment and revised parameter setting.

E) Evaluation: This stage is aimed at the evaluation of results obtained

F) Deployment: This is to put your model into an existing pipeline, create your own

or deploy to cloud computing services. Some of the activities are:

-Play deployment

-Plan monitoring and maintenance

-Review project

**3.1** Programming for Data Analytics Task

1.  Python is a popular and versatile programming language which has become a popular choice among Data Scientists for its ease of use, extensive libraries, and flexibility. Python provides an efficient and streamlined approach to handing complex data structure and extracts insights as demonstrated in the attached Jupyter Notebook.

2. The building construction dataset project are carried out using Python Programming Language and documented in the attached Jupyter Notebook with necessary explanations in this report for sound justification with the use of libraries and necessary commands to analyse the dataset for proper understanding and results.

3. The project documentation on Jupyter Notebook involves by importing firstly some libraries like Pandas, Seaborn, Numpy, Matplotlib. And then read the CSV file name to see the shape, head, tail, info before describing the dataset to include objects using df.shape, df.head(), df.tail(), df.info, df.describe(), and df.isnull() respectively for explanation of code choices, the missing values are checked to replace with zero or median value and skewness.

Pandas is a data manipulation package in Python for tabular data. That is, data in the form of rows and columns, which is also known as data frames. Other popular Python data services packages are:

-Numpy for numerical computing

-Matplotlib, Seaborn, Plotly and other data visualisation packages

-Scikit learn for machine learning

4. Testing & Optimisation: These are carried out for the analysis done on construction datasets by ensuring errors are removed from raw data collected, execute every line of code at least once (code coverage), and data coverage (input data and programme generated data) with defensive programming to ensure no errors in any section of the codes and validating the inputs. The GridsearchCV was implemented in Machine Learning model used to optimise the analysis carried out in addition to memory usage, dftypes function and info() to monitor the speed of the notebooks if they are optimal.

Timing codes and scrapper are also used for testing and optimisation along with profiler which test the performance of codes used for this analysis.

The trade-off here is the time spent to prevent errors and validating the inputs and the efforts expended to eliminate the bugs/errors from conception to specification, to design and to the codes written.

5. Data manipulation: The two main libraries used here are NUMPY and PANDAS.

While Numpy is used for numerical computing, Pandas is a powerful data manipulation tool and is used throughout the data analysis workflow for:

-Import datasets from data bases, spreadsheets, commas separated values (csv) files, etc

-Clean datasets like dealing with missing values if any.

-Tidy datasets by reshaping their structure into a suitable format for analysis.

-Aggregate data by calculating summary statistics such as the means of columns, correlation between them and more.

-Visualise datasets and uncover insights.

Pandas also contains functionality for time series analysis and analysing text data. Pandas supports an extensive set of operations from exploratory data analysis, dealing with missing values, calculating statistics, visualizing univariate, and bivariate data and much more unlike other libraries.

It works with large data of different types to solve any problems.

Three ways of combining dataframes are as follows:

merge () for combining data on common columns or indices.

join () for combining data on a key column or an index.

concat () for combining Data Frames across rows or columns.

The techniques involved in this process must take into consideration if the data is indexed by inner, outer, left, or right alignment.

**3.2** Statistics:

The datasets available for statistical testing were taken from

[www.ons.gov.uk](http://www.ons.gov.uk) (“outputukhousebuilding.xls”) and

[www.data.gov.ie](http://www.data.gov.ie) (“national\_house\_construction\_cost\_index\_0.csv”).

A JSON format dataset was also used. It was taken from [www.data.gov.ie](http://www.data.gov.ie) (bcms\_commencement\_notices\_2019.json). Incidentally, a commencement notice is an official document submitted by a builder or construction company to a planning authority. They are a good indicator of how many houses developments are actually being built.

The following tests are used in this assessment.

Pearson’s correlation – This is the standard test to check if two variables are associated.

Levene’s tests for homogeneity of variance – To check if the mean variances are equal, that is, the samples are drawn from the same population.

Wilcoxon Ranked test – This test investigates if the medians of the samples are equal. Levene’s test and Wilcoxon’s test are similar.

One way Anova. This test was used to compare the variances across multiple groups.

Welch’s t-test – This is a non-parametric method to see if the means of the samples are equal.

The findings of these tests are explained in the table below under the Section 4a of Results and Interpretations.

**3.3** Machine Learning

The rationale and justification of the models used is the basis of comparing and contrasting the results and insights gained on the construction datasets used.

Multiple regression was used for predicting the target variable with the other variables. while Time Series Analysis (TSA) was carried out for forecasting future results based on prior information.

The logistic regression was used to examine the likelihood of one outcome over another. Both Logistic regression and Decision Trees are classification models .

Finally, TSA was found to be very effective at forecasting future events. Also, a simple linear regression was very precise and accurate in its predictions.

The table below in Section 4b is used to illustrate the similarities and contrast of Machine Learning outcomes with all the necessary observations.

**3.4** Data Preparation and Visualization

Data are raw facts and figures or set of observations/evidences which can be examined to provide information or it can be analysed to provide evidence.

A variable may be either of the following:

-Qualitative or Quantitative

-Categorical or Numerical

-Discrete or Continuous

The main level of measurement of data in a research project is a determining factor to the type of descriptive and inferential statistical procedures that can and cannot be used to summarize and analyse variable in a data set.

The process of acquiring data is basic methods/ways of sourcing data which are:

-A study-specific data collection instrument (like questionnaire)

-To facilitate consistent collection of participant data by the Researcher/Investigator

-The basis of the research/domin electronic database

-The repository of source data for reference during data editing and analysis.

The negative aspect/limitations of data collection for the research work are:

-How to ensure consistency in data collection

-What instruments will be used to collect them

-What variable will be required for the research work.

The method used to prepare the dataset for Machine Learning (ML) was to use Pandas to load the dataset. Pandas is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool through data frames using the read\_csv function. The head and tail of the dataset were examined to know the data types and check for null values before describing the data frame to know the count, mean, standard deviation, minimum, percentiles and maximum values of the dataset. Other methods of preparing data for Machine Learning (ML) are feature engineering, log transformation and feature scaling or normalisation. This is taking available information in the domain to turn into numbers which can be used to build feature matrix. There are features for representing categorical data, features for representing text and features for representing images. Another method is Vectorisation which is the process of using derived features for increasing model complexity and imputation of missing data by encoding categorical data to numerical values. Vectorisation involves converting arbitrary data into well-behaved vectors.

Appropriate visualizations were used by plotting some graphs and charts (see notebook(s)) to investigate the relationships between the variables in the dataset. Pandas with plotly was utilised for the interactive dashboard. The scatter plot showed the correlations while boxplots demonstrated max-min values, outliers, etc while histogram shows frequency distribution ranges of any the variable in the dataset for easy comparison immediately.

All design and implementation of the visualizations are documented are in the attached Notebook on Jupyter for easy reference.

## Section 4(a) – Results and Interpretations – Statistical tests

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test(s) 1-5 | H0: | Results | Interpretation | Remarks | Suggestions for further study |
| Pearson’s correlation  (parametric test) | The hypothesis is that there is a strong correlation between the number of dwellings started and completed in the U.K. | r = 0.924 | H0 was supported by the outcome of the test | Over 92% of houses are completed in UK | It may perhaps be helpful to investigate why 8% of houses are **not** completed in the U.K. |
| Levene’s test  (non-parametric test). | H0 states that the variances are equal. Here we are using data from housing construction cost index dataset. | P. < 0.05 (alpha) | The data was **not** drawn from samples with equal means/variances. We reject H0. | Many external factors influence the unit cost of construction which may explain the findings. | It would be prudent to check if, using data analysis, the cost of materials should be subsidised for builders to stabilise construction costs of buildings. |
| Wilcoxon ranked test  (non-parametric) | H0 states that the variances in our sample are unequal. | P. < 0.05 (alpha) | We accept H0. The variances are unequal. | Levene’s test also showed that the samples had unequal variances. Both tests support one another. | It's not surprising that H0 was supported as many unrelated factors contribute to construction costs. |
| One way ANOVA  (non-parametric)  Welch’s t-test (non-parametric) | H1 states that the samples do not originate from the same distribution.  H0 states that the mean averages are the same between houses completed privately and houses completed publicly | P. < 0.05. (F statistic is 200.11)  P. < 0.05 (alpha) | We reject H0 and accept H1. The samples means are significantly different. They do not come from the same distribution.  We reject H0. There is a significant difference between the average no. of homes completed by the local authority and the private sector. | There are 26 counties in the Republic of Ireland. We expected that the samples would have different means.  As Local authorities and the private sector have different budgets, expectations and roles in housing provision, this result is not unexpected | Further analysis of the data from Urban areas (like Dublin, Galway and Cork) may be undertaken to see if their distributions are similar.  Further studies of the impact of Private and Public funding upon housing construction should be considered. |

## Section 4(b) – Results and Interpretations – ML Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Model** | **Accuracy** | **Speed** | **Memory usage** | **Observations** |
| Time Series | Over 99.7% | Very fast processing time | As the dataset was very small, RAM was not a problem. | A very satisfactory model. Going forward, it would be of great benefit when construction companies are preparing their budgets |
| Logistic Regression  Decision Tree model was more accurate, it was observed | Accuracy was low, approximately 27%  Accuracy was 50% | Very fast processing time  Very fast processing time | As the dataset was very small, RAM was not a problem.  As the dataset was very small, RAM was not a problem. | The dataset available were not very suitable for log-regression.  This study showed that Decision Trees algorithm was more precise. Over 50% accuracy. |
| Simple Regression | Accuracy was very good. R-squared was greater than 86%). | Very fast processing time as only 2 variables were compared. | As the dataset was very small, RAM was not a problem. | The study suggests that perhaps a much larger dataset was needed. |

## Section 5 – Conclusion

Based on Pearson’s Correlation carried out on the construction datasets of both Ireland and UK, we accept Ho outcome. The hypothesis is that there is a strong correlation between the number of dwellings started and completed in the U.K compared with Ireland.

Predictions were made with TSA to forecast the average cost index price of construction. This has far reaching consequences for what the consumer might have to pay in the future.

A logistic regression was used to determine the type of construction based on its unit value costs. This model was not as effective enough.

Finally, TSA was found to be very effective at forecasting future events. Also, a simple linear regression was very precise and accurate in its predictions which has implications for the effective deployment of manpower, money and materials in building and construction.

This is an interesting area of study where further research can be conducted through Neural Networks or Deep Learning.

## Section 6 – Acknowledgements

Everyone has been a major source of inspiration, motivation, and support for this course.

Warm appreciation to my family and friends who encouraged me to do this programme to enhance my knowledge by taking up this new challenge.

This project has provided me with a good baseline understanding of Data Analytics and key concepts in Machine Learning and I am ready to tackle the challenging subject in earnest by building on the research work.

Finally, I would like to express my gratitude to Springboard and CCT Management for providing the learning platform to upskill my knowledge in this field of Data Analytics with my fellow colleagues.

## Section 7 – References

1. Andreas, C. Müller and Sarah Guido, O'Reilly Media, ‘Introduction to Machine Learning with Python’, Inc. October 2016
2. Bailey, M. et al. (2019) ‘House Price Beliefs and Mortgage Leverage Choice’, Review of Economic Studies, 86(6), pp.
3. Barbosa, F., Woetzel, J., Mischke, J., Ribeirinho, M.J., Sridhar, M., Parsons, M., Bertram, N. & Brown, S. (2017, Feb. 17). Reinventing construction through a productivity revolution. McKinsey Global Institute
4. Bill Schmarzo and Dr. Kirk Borne (2020) *The Economics of Data, Analytics, and Digital Transformation: The Theorems, Laws, and Empowerments to Guide Your Organization’s Digital Transformation*. Birmingham: Packt Publishing.
5. Bureau of Economic Analysis (BEA). (2020). Industry economic account data; components of value added by industry; GDP by industry-Academic Journal
6. Gopal Sakarkar (2021) *Machine Learning Algorithms Using Python Programming*. New York: Nova (Internet of Things and Machine Learning).
7. Keras, and TensorFlow, Aurélien Géron, O'Reilly Media, ‘Hands-On Machine Learning with Scikit-Learn’, 2nd Edition, September 2019, ISBN: 9781492032649
8. Severance Charles R. Python for Everybody (Exploring Data Using Python 3)’ Editorial support by Elliott Hauser, Sue Blumberg
9. Sneha Kumari; K. K. Tripathy; Vidya Kumbhar. Bingley : ‘Application of Big Data and Business Analytics’ Emerald Publishing Limited. 2021
10. Theobald, Oliver, ‘Machine Learning for Absolute beginners’, 2nd Edition, 1997
11. Tufte, Edward, ‘Data Preparation-Visual Display of Quantitative Information’
12. Weiss Neil A, ‘Introductory Statistics’ Global 10th Edition 2017 Pearson Education Limited.

**Total words count (excluding acknowledgement and references) is 3285**