**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

Word count: 5208 including Abstract, Introduction, Citations and References.

## Abstract

*This report examines the process and choices made in completing each step of a data project lifecycle from data sourcing through to surfacing data in a human usable format.*

*Using data related to the construction industry, supporting data extracted from IE gov, UK gov and the World Bank, analysis, modelling and predictions were made to forecast the volume of new housing builds in IE in the near future. Supplementary to this was data extracted from Twitter to determine trends in sentiment for #housingcrisis. The end result of the analysis being an understanding of the public perception of housing in IE and whether there are enough houses to meet demand.*

*Completed in 3 sections, the first is a Jupyter notebook containing data processing, modelling and analysis procedures carried out using the Python programming language. The second is a standalone Jupyter notebook that houses a dashboard built using Dash and Plot.ly. The third is this report which will carry out a critical analysis of the analytics lifecycle, methodologies employed, technologies used and a comparison of the results of predictions made.*

## Introduction

This report examines the process of data collection, data preparation & visualisation, statistical analysis and the application of machine learning models to make predictions using Time Series & text based data and evaluates a number of approaches that can be applied to these problems. The time series data used to support this analysis was collected from Irish & UK government websites for data relating to the construction industry i.e. new house production. Time series data relating to population estimates was collected from UK government and World Bank websites and text based data to support sentiment analysis was collected via web scraping from Twitter.

The project consists of three portions, the first being a Jupyter notebook, the second a standalone dashboard in a Jupyter notebook and the third being this report. The Jupyter notebook contains the source code for the ingestion, processing and analysis of the above data with all work carried out using the Python programming language. The goal of this portion of the project was concerned with applying cleansing and pre-processing techniques to time series and text data. Followed by the application of statistical, machine learning and deep learning models to infer, forecast and predict. Each application was compared and assessed for accuracy.

The second portion consists of a standalone dashboard that compliments the machine learning portion of the project. In this instance, the dashboard is a time series forecasting tool to display historical new build volumes and to forecast based on user controlled inputs.

The goal of the third portion of the of the project was concerned with providing a critical analysis of the methodologies chosen, justifications for the decisions made and an interpretation of the results of the data modelling and analysis processes. This is detailed below in sections 1 – 4.

## Methodology

The overall structure of this project is based on a CRISP-DM methodology (Quantum, 2021). When selecting the data under review, it was crucial that the data could be useful to construction sector professionals. In this case the question is “Can we predict the expected number of new builds we need to keep up with demand in the next year?”. This question has real business implications in terms of budget requirements, labour requirements, supply chain dependencies, land availability and tooling.

### 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. (data.gov.ie. n.d.).***.*** 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 (Creative Commons, 2016) 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 the fact that the initial data that was sourced was normalised into an 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. This was broken down by volume started per quarter vs volume completed by quarter. The data was downloaded and processed in an Excel spreadsheet format (Office for National Statistics, 2023) under the UK Open Government Licence v3 (The National Archives, 2019)which necessitates citation and has additional exclusions 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 (‌World Bank, 2021). Data 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 data for the UK was sourced from the ONS.gov website (‌www.ons.gov.uk, n.d.) 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. Ultimately, the scrolling load mechanism that Twitter uses to serve content was too much of a technical challenge for me to overcome. This put web scraping outside the scope of this project. The solution was to use a Chrome extension Instant Data Scraper (‌chrome.google.com, n.d.) to recover the data. The cleansed text was manually assembled using pandas and the CSS tags within the HTML that was scraped. Twitter offers a Developer Agreement and Policy (‌developer.twitter.com, n.d.) 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, Feature Preparation & Engineering

All visualisations provided below are designed to align with Tufte’s principles by enforcing the use of zeroed axes, minimising additional chart junk i.e. gridlines, excessive tick marks and labels. Colour was used minimally and only to clearly distinguish between comparable chart entities.(Tufte, E. R., 2001)The use of pre-attentive attributes (Barrera-Leon, Corno and De Russis, 2020) to increase visualisation impact were also applied where possible.

For IE housing data, the data was complete with no missing values or significant outliers that implied errors in data. There were, however, a number of duplicate columns in the data. These columns were not duplicates but served as a pseudonym for other more descriptive features. To avoid correlation bias in later regression exercises, these were removed. (Tolosi and Lengauer, 2011)*.* IE data was in the form of Time Series, line plots as the standard display mode were chosen (Fang, Xu and Jiang, 2020). At the quarterly level of aggregation, visually identifying seasonality (fig. 1) was difficult. With lower levels, e.g. daily, more seasonality could be determined and more complex charts e.g. spiral used. The distribution of builds per quarter used a histogram to display the right skewed distribution. (fig. 2)

A picture containing text, diagram, plot, line

Description automatically generatedA picture containing diagram, screenshot, plot, rectangle

Description automatically generated

fig. 1 fig. 2

Box, Swarm & Violin plots were used to detail the distribution of quarterly builds. (fig. 3) This allowed to leverage the general understanding of boxplots with the addition of distribution shape from the violin plot and density of point by the swam plot. (Hintze and Nelson, 1998).*.*

A picture containing diagram, screenshot, origami

Description automatically generated A picture containing screenshot, text, diagram, plot

Description automatically generated

fig. 3 fig. 4

Similar activities were carried out for the other IE time series with bar charts included to assess.

In assessing the various IE build datasets, there was a minor difference in the aggregates of all chosen data sets. For 3 of the 4 chosen, the total builds from Q1 2011 to Q1 2023 was 169997. For NDQ03, this was 169985. A difference of 12. As this could not be mapped directly to another field, the time was not invested to track down the difference. At 0.000071%, this was an acceptable margin of error.

### UK data was aggregated by year and was not directly comparable with the quarterly IE data. There were no missing values in the data, however, there were missing values for Wales due to a lack of collection. The plan was to impute these values by using the ARIMA (as was used later in the analysis) but this was not included due to more time being invested elsewhere. The approach taken to derive quarterly values for the UK data was to create blank rows for Q1, Q3 & Q4 (data was generated in Q2). Then both Forward & Backfill (to catch missing values not accounted for in the other) and linear interpolation.

### fig. 5 shows the build trend for Scotland including gaps. Linear interpolation provides a smoother gradient than forward filling so this approach was used.

### A picture containing line, plot, diagram, font Description automatically generated A picture containing text, screenshot, line, plot Description automatically generated

fig. 5 fig. 6 fig. 7

Twitter data was extracted, using a web scraping extension for Google Chrome. This provided a very broad table view of data split by CSS tags from the Twitter webpage. Clean-up was focused on identifying tweet text/ hashtags and mentions vs URLs & web page text. This involved dropping columns exclusively used for URLs and selecting only columns that contained a specific DIV tag. These comments were join into a single string and then hashtags and @mentions were extracted to separate features. As dates were recorded in text form these were parsed. Where year was missing, this was current year, otherwise values were extracted and mapped to dictionaries of values. Fig. 8 shows the trend in number of tweets containing the hashtag #housingcrisis.

### A picture containing text, diagram, line, plot Description automatically generated

### fig. 8

Basic text pre-processing (Camacho-Collados, 2017b)was carried out on the tweet data. The removal of rare and common words was avoided. Rare words in this case resembled mentions, names and words missing characters. Common words were housing crisis related.

In order to predict sentiment for unlabelled data, a sentiment score was required to be calculated. 3 approaches were attempted. Using Valence Aware Dictionary for Sentiment Reasoning (VADER) and the TextBlob library (Abiola *et al.*, 2023b) and attempting to use pre-trained Deep Learning models (HuggingFace) via the transformers library. Attempt 3 was not possible as I could not get the necessary libraries to cooperate. Approaches 1 & 2 were carried out on the full tweet where the resulting polarity score was bucketed in Negative < 0.45, Neutral 0.45 to 0.55 and Positive > 0.55. Following this, TfidVectorizer package was used to vectorize the lemmatized terms. This was chosen rather than CountVectorizer in order to take into account term importance rather than just a count of the terms (Kaplan, 2022b)*.*The final step was to carry out Principal Component Analysis on the vectorizer data in order to reduce the object size and complexity before passing to machine learning models. Fig. 9 shows the number of principal components to achieve 60, 70 & 80% coverage of feature variance.

A picture containing text, line, plot, diagram

Description automatically generated

Fig. 9

The dashboard provided is concerned with providing construction professional with a means to forecast expected volumes of new builds in each county in Ireland. This was built using dash and plot.ly to display the visualisation and input controls. The model behind the same auto-ARIMA model that was developed in the machine learning portion of the project. Designed using the same design principles as the visualisations through the rest of the analysis e.g. low ink ratio, low chart junk, minimal colouring, use of pre-attentive attributes and reading left to right and top to bottom. The layout and display was designed to predict a window into the future and to provide feedback on model quality. There is scope for improvements in models offered, pre-processing in the backend and more robust checks to ensure that high quality data is being presented.

### 2. Statistical Analysis

The dataset under investigation is *df\_con\_and\_pop.*  This data represents new builds per quarter, population and builds per 10k people for IE, England, Scotland, Wales & NI. Observations recorded from Q1 2011 to Q1 2021..

The dataset consists of 44 observations across 18 columns, 15 numerical, 3 string.

**IE:** Sample mean new builds per year: 3033 with a standard deviation of 1796.   
IE population and new builds show a linear relationship with a very strong positive correlation efficient of 0.91 (fig. 10 & 11). Both features have a non-normal distribution with a right skew. A Shapiro-Wilk test confirms non-normality with a p-value 1.2e-3. Following a t-test, we can be 95% confident that true population mean for new builds per quarter lies between 2.5k & 3.5k.

A picture containing diagram, line, plot, screenshot

Description automatically generated A screenshot of a graph

Description automatically generated with low confidence

fig. 10 fig. 11

**England:** Sample mean new builds per year: 35651 with a standard deviation of 7.8k. EN population and new builds show a linear relationship with a strong positive correlation coefficient of 0.72 (fig. 12 & 13). New builds shows an approx. normal distribution and population showing non-normal with a left skew. A Shapiro-Wilk test confirms normality with a p-value 0.07. Following a t-test, we can be 95% confident that true population mean for new builds per quarter lies between 33.3k & 38k.

A picture containing diagram, screenshot, text, plot

Description automatically generated A screenshot of a graph

Description automatically generated with low confidence

fig. 12 fig. 13

**Scotland:** Sample mean new builds per year: 4372 with a standard deviation of 928. Scotland population and new builds show a linear relationship with a medium positive correlation coefficient of 0.53 (fig. 14 & 15). New builds shows a approx. normal distribution with population showing non-normal with a left skew. A Shapiro-Wilk test confirms non-normality with a p-value 0.009. Following a t-test, we can be 95% confident that true population mean for new builds per quarter lies between 4079 & 4620.

A picture containing diagram, screenshot, text, line

Description automatically generated A screenshot of a graph

Description automatically generated with low confidence

fig. 14 fig. 15

**Wales:** Sample mean new builds per year: 1283 with a standard deviation of 607.   
Wales population and new builds show a non-linear relationship with weak negative correlation coefficient of -0.16 (fig. 16 & 17). New builds shows a approx. normal distribution with population also showing as approx. normal. A Shapiro-Wilk test confirms non-normality of new builds with a p-value 9.2e-7. Following a t-test, we can be 95% confident that true population mean for new builds per quarter lies between 1112 & 1451.

A picture containing diagram, screenshot, plot, line

Description automatically generated A screenshot of a graph

Description automatically generated with low confidence

fig. 16 fig. 17

**Northern Ireland:** Sample mean new builds per year: 1588 with a standard deviation of 310. NI population and new builds show a non-linear relationship with a medium positive correlation coefficient of 0.6 (fig. 18 & 19). New builds shows a approx. normal distribution with population showing non-normal almost uniform distribution. A Shapiro-Wilk test confirms normality with a p-value 0.5. Following a t-test, we can be 95% confident that true population mean for new builds per quarter lies between 1495 & 1683.

A picture containing diagram, screenshot, text, line

Description automatically generated A screenshot of a graph

Description automatically generated with low confidence

fig. 18 fig. 19

Based on visualisations from the data, a number of questions were raised that required additional research:

1. *Is there a significant difference in the distribution of number of houses built in IE and Wales?*

As neither IE or Wales follow a normal distribution, a comparison of the distributions requires the use of a non-parametric test. In this case, the Mann-Whitney U rank test was applied.

H0: There is no difference between the number of builds in IE & Wales.  
H1: There is a difference between the number of builds in IE & Wales.

The confidence interval suggests that, at a 95% confidence level, the true difference in medians between the two samples is likely to be between -2664 and -558.95. This interval does not include zero, indicating that there is evidence of a significant difference in medians, supplemented with a p value of 0.0002(< 0.05)

This allows us to reject the null hypothesis and supports the conclusion that there is a significant difference in the distribution of the two groups.

1. *Is there a significant difference in the variance of builds per 10k people between England and Northern Ireland?*

England & NI both have approximately normal distributions, we can use a parametric test. In this case, we apply the Levene test for equal variances. As both samples are normally distributed a parametric test is suitable rather than non-parametric alternative – Fligner test.

H0: There is no significant difference in variance in builds per 10k between England and NI.

H1: There is a significant difference in variance in builds per 10k between England and NI.

The large p-value of 0.43 (> alpha) indicates evidence supporting the null hypothesis and provides support for the conclusion that there is no statistically significant difference between variance of builds per 10k between England and NI, therefore we fail to reject the null hypothesis and cannot say with certainty that there is a significant difference in the variance of builds per 10 between England and NI.

1. *Is there a significant difference in the mean number of builds per 10k people between England and Northern Ireland?*

We apply a t-test to determine whether there is a difference in the mean number of builds per 10k people in England and NI. As both samples are normally distributed a parametric test is suitable.

H0: There is no significant difference between the mean number of builds per 10k in England and NI.

H1: There is a significant difference between the mean number of builds per 10k in England and NI.

The confidence interval suggests that, at a 95% confidence level, the true difference in medians between the two samples is likely to be between -4.1 and -0.02. This interval does not include zero, indicating that there is evidence of a significant difference in means of the samples.

The small p-value of 2.36e-09 (< 0.05) allows us to reject the null hypothesis and provides support for the conclusion that there is a significant difference in the mean builds per 10k between England and NI.

1. *Is there a significant difference between median number of builds in IE, Scotland & Wales?*

We apply a non-parametric Kruskal-Wallis test to examine if there is a significant difference between the median values of new builds in IE, Scotland and Wales as none of the sample distributions are normally distributed.

H0: There is no significant difference between the median number of new builds in IE, Scotland and Wales.

H1: There is a significant difference between the median number of new builds in IE, Scotland and Wales.

The confidence interval for the Kruskal-Wallis statistic, at a 95% confidence level, is between 50.8 and 75.9. The interval does not include zero, which indicates evidence of a statistically significant difference the medians of new builds between IE, Wales & Scotland. The small p-value of 1.77e-14 (< 0.05) allows us to reject the null hypothesis and provides support for the conclusion that there is a significant difference in the median of new builds between IE, Wales & Scotland.

1. *Does the number of builds in IE follow an upward trend?*

As IE new builds is non-normal, we use a non-parametric test, Pymann-Kendall, to evaluate whether new builds in IE are following an upward trend.

H0: There is no monotonic trend in this sample.

H1: There is a monotonic trend in this sample.

The small p-value of 2.0e-1 (< 0.05) provides evidence against the null hypothesis and provides support for the conclusion that there is a monotonic trend for new builds per quarter in IE.

1. *Is there a significant correlation between population and new builds in IE?*

As IE new builds is non-normal, we use a non-parametric test, Spearmans R, rather than the parametric Pearsons R test.

H0: There is no significant correlation between IE new builds and IE population.  
H1: There is a significant correlation between IE new builds and IE population.  
The small p-value of 7.95e-15 (< 0.05) allows us to reject the null hypothesis and provides support for the conclusion that there is a significant correlation between IE new builds and IE population.

This analysis has shown that there is a strong relationship between population growth and new builds of houses. Further areas of research would dig deeper into all markets and investigate the change in proportion of housing across time e.g. moving from terraced housing to semi-detached and apartments. Reviewing the sentiment analysis performed in the ML portion, suggests that the housing crisis is being made more extreme by the increases in population from non-nationals and asylum seekers. Additional data on the volume and location as well as the emigration rate would supply additional context and support for forecasting housing requirements in the future.

### 3. Machine Learning

This analysis required the use of a number of difference Machine Learning and statistical models. For the sentiment analysis portion, 2 approaches were used. The first was in using a Feed-forward Neural Network (FNN). In this case, an FNN was used rather than another form of NN such as an Recurrent NN mainly due to limit memory use and to increase the training speed but also RNN is more suited to sequential and temporal inputs where the order of data is important. The other choice was Logistic Regression in order to use a simple and interpretable model. Logistic regression can work well with lower numbers of observations and can help with understanding the probability of observations relating to a certain class.

The time-series portion leveraged ARIMA forecasting and XG Boost based models. ARIMA was used as this is a flexible model which can handle trends, seasonality and other temporal patterns in the data. ARIMA provides pdq parameters to help understand the nature of time series and it is a well understood approach with solid theoretical foundations.

XGB was used as this is an extremely flexible model providing the ability to handle complex, non-linear relationships and missing values. There is an enormous amount of social proof as XGB is regularly used in Kaggle winning models. Additionally, XGB provides a vast number of hyperparameters to allow for finely tuned modelling and offers a feature importance metric.

Fig. 20 shows the results of model testing to predict the class of sentiment calculated by the TextBlob and Vader libraries. There was no discernible difference in accuracy, training rate or convergence rate for the FNN between the different levels of PCA features included. FNN scored higher on TextBlob classes vs Vader. This may be due to the fact that VADER is tuned to look for certain words, punctuation etc that appear on social media. These were removed in pre-processing and may have removed the nuance that FNN could have recognized. There was also no hyper-parameter tuning performed with the FNN and the data set was only 10 records. Fig. 21 shows that after 4 epochs the model had peaked. More intense hyper-parameter tuning would likely improve this.

Logistic Regression was remarkably consistent across all PCA levels and for both sentiment types. Unlike the FNN, LogReg had extensive hyper-parameter tuning. This was more supportable using the compute resources available. With additional compute, the range of tuning could have been wider and improved the classification rate. Additional experiments to adjust the probability levels required to classify the LogReg classifier would also improve the result.

Sentiment Analysis:

A picture containing text, receipt, number

Description automatically generated

fig. 20

A picture containing text, screenshot, line, plot

Description automatically generated

Fig. 21 shows the results of testing of 3 forecasting models.

Manual ARIMA was by far the worst performing of the 3. This was likely down to variation in the time series data that was not optimised for during model training. The manual training used a rudimentary approach in calculation and was limited to pdq parameters. Auto-ARIMA was the highest performing and demonstrates the importance of deep hyper-parameter tuning. The Auto-ARIMA model was tuned across a greater set of hyper-parameters and in a more performant way. Both models could have benefited from additional processing of the original data. One diff was performed to ensure stationarity but no actions were taken to address the variance and linearity of the data.

A picture containing text, screenshot, number, line

Description automatically generated

fig. 21

XGB was initially trained to predict the time series using highly correlated variables which produced a 100% test accuracy. A change in approach to create a set of lag features was used and produced a high accuracy result. This model also suffered from a lack of hyper-parameter tuning. Additional work should be done for tuning and in identifying any patterns in importance of the lag features.

### 4. Programming

The testing and optimisation strategy employed throughout this project consisted on focusing on data validation. During the data ingestion and cleansing processes, duplicate data was removed in the form of removing columns had a one to one relationship e.g. TLIST(Q1) and Quarter for df\_ie\_ndq1. At each stage of completion of data cleansing for the IE data frames, there was an assertion to ensure that the sum of values from the original data was equal to the post processing values. Finally for IE data, each data frame which was a disaggregation of the total value for each quarter was summed and validated in comparison to the others. This highlighted a discrepancy (referenced in section 1). Cross Validation was used in training of the Logistic Regression models for sentiment analysis. This approach maximised the accuracy score and helped to mitigate model overfitting. The difference in prediction quality of the models that used CV and those that did not highlights the effect on model quality and effectiveness.

In order to optimise the operation of the notebook, steps were taken to reduce the overhead needed to recalculate all fields. The main approach here was to write the PCA variables to disk rather than recalculating each time. As the values did not change, there was no effect on model performance.

For all tabular data sources, pandas and SQLAlchemy were 2 possibly options for data manipulation. Pandas has a lower overhead, integrates well with ML models and is supported by the Python ecosystem. This provides a wide array of functions and libraries for advanced data manipulation. The limiting factor is that pandas works in memory so is well suited to smaller datasets. Larger data sets are not suitable. SQLAlchemy on the other hand allows for the use of a relational database built to efficiently store and query tabular data. SQLAlchemy allows not only for SQL queries to be used but for the SQL database and the query planner to optimise operations.

For text based data, two options are pandas and networkx. Pandas for text has the same advantages as for numerical. A large collection of supporting libraries, wide functionality and a low barrier to entry. Networkx uses graph based structures so can help to identify relationships within text and entities using the text. From the point of view of manipulation of text data, pandas is more suited to cleansing and pre-processing where network is better suited to analysis and text modelling.

## References

Data Sources:

data.gov.ie. (n.d.). *NDQ01 - New Dwelling Completions - New Dwelling Completions - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/ndq01-new-dwelling-completions/resource/8fdc8f87-88ca-43f7-aa16-dbc3a52cacde.

‌

data.gov.ie. (n.d.). *NDQ03 - New Dwelling Completions - New Dwelling Completions - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/ndq03-new-dwelling-completions/resource/45701f7f-1192-406f-980f-a8ee23862147.

‌

data.gov.ie. (n.d.). *NDQ05 - New Dwelling Completion - New Dwelling Completion - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/ndq05-new-dwelling-completion/resource/108d82d0-43f1-4574-9e47-bb54f078fd84.

‌

data.gov.ie. (n.d.). *NDQ06 - New Dwelling Completion - New Dwelling Completion - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/ndq06-new-dwelling-completion/resource/0b5041bb-9f52-4b54-8c8d-0000306e74e2.

‌

Office for National Statistics (2023). *UK house building: permanent dwellings started and completed - Office for National Statistics*. [online] Ons.gov.uk. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/ukhousebuildingpermanentdwellingsstartedandcompleted.

‌www.ons.gov.uk. (n.d.). *Population estimates time series dataset - Office for National Statistics*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatestimeseriesdataset.

‌World Bank (2021). *Population, Total | Data*. [online] Worldbank.org. Available at: https://data.worldbank.org/indicator/SP.POP.TOTL.

Creative Commons (2016). *Creative Commons — Attribution 4.0 International — CC BY 4.0*. [online] Creativecommons.org. Available at: https://creativecommons.org/licenses/by/4.0/.

The National Archives (2019). *Open Government Licence*. [online] Nationalarchives.gov.uk. Available at: https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

‌developer.twitter.com. (n.d.). *Developer Agreement and Policy – Twitter Developers*. [online] Available at: https://developer.twitter.com/en/developer-terms/agreement-and-policy.

‌Anon, (2023). *How to Scrape Twitter Data without Using Twitter’s API - AITechTrend*. [online] Available at: https://aitechtrend.com/how-to-scrape-twitter-data-without-using-twitters-api/ [Accessed 26 May 2023].

‌

OS2INT (n.d.). *Scraping Twitter without an API or user account with Tweeds | OS2INT*. [online] os2int.com. Available at: https://os2int.com/toolbox/scraping-twitter-without-an-api-or-user-account-with-tweeds/ [Accessed 26 May 2023].

‌

M, A.J. (2023). *Tweeds - Twitter Scraping Tool*. [online] GitHub. Available at: https://github.com/achyuthjoism/tweeds [Accessed 26 May 2023].

White, B. (2020). *Sentiment Analysis: VADER or TextBlob?* [online] Medium. Available at: https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540 [Accessed 26 May 2023].

‌chrome.google.com. (n.d.). *Instant Data Scraper*. [online] Available at: https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah.

‌Tufte, E. R. (2001), *The Visual Display of Quantitative Information* , Graphics Press , Cheshire, CT .

Knaflic, C.N. (2019) *Storytelling with Data: Let’s Practice!* John Wiley & Sons.

Barrera-Leon, L., Corno, F. and de Russis, L. (2020). *Systematic Variation of Preattentive Attributes to Highlight Relevant Data in Information Visualization*. [online] IEEE Xplore. doi:https://doi.org/10.1109/IV51561.2020.00022.

‌

Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, [online] 27(14), pp.1986–1994. doi:https://doi.org/10.1093/bioinformatics/btr300.

‌

Fang, Y., Xu, H. and Jiang, J. (2020). A Survey of Time Series Data Visualization Research. *IOP Conference Series: Materials Science and Engineering*, 782(2), p.022013. doi:https://doi.org/10.1088/1757-899x/782/2/022013.

‌

Hintze, J.L. and Nelson, R.D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52(2), p.181. doi:https://doi.org/10.2307/2685478.

Nguyen, Minh & He, Tong & An, Lijun & Alexander, Daniel & Feng, Jiashi & Yeo, B.T. Thomas. (2019). Predicting Alzheimer's disease progression using deep recurrent neural networks. 10.1101/755058.

Camacho-Collados, J. (2017) *On the Role of Text Pre-processing in Neural Network*

*Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis*.

Available at: https://arxiv.org/abs/1707.01780.

Abiola, O. *et al.* (2023) “Sentiment analysis of COVID-19 tweets from selected

hashtags in Nigeria using VADER and Text Blob analyser,” *Journal of Electrical*

*Systems and Information Technology*, 10(1). Available at:

https://doi.org/10.1186/s43067-023-00070-9.

Kaplan, D. (2022) “Machine Learning 101: CountVectorizer Vs TFIDFVectorizer »

EML,” *EML*, 18 October. Available at:

https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer/.

Quantum (2021) *Data Science project management methodologies - DataDrivenInvestor*. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb.