5/26/2023

Name: Rachel Hurley

Student No.: SBS23033

Word Count: Approx. 3,200 words

GITHUB Link: [SBS23033\_CA2\_GitHub\_Link](https://github.com/rachelhurleycct/CA2)

CA2

MSc Data Analytics CCT College Dublin

Abstract

This assignment analyzed various sectors of the construction industry using data based on Ireland and other countries. Some areas looked at were construction employment numbers, various construction indices and worker hours worked/hourly earnings. To complete this analysis of the industry the data was cleaned, adjusted, explored and statistically evaluated. Machine learning models were then used for prediction, classification and regression. Sentiment analysis was complete using the Reddit API and models of average accuracy were produced.

From the statistical analysis it was deduced that the average earnings of clerical workers in the construction workers is significantly less than that of the male clerical workers which identifies a pay gap. Ireland was compared to New Zealand in terms of number of construction workers and using inferential statistics it was determined the numbers are very different between the countries. This may possibly be due to immigration of construction workers from Ireland to New Zealand, Canada and Australia. An interactive dashboard was created using Dash and included various graphical representations produced throughout the notebook. The dashboard contained 3 pages and focused on displaying important and appealing visualizations.

Table of Contents

[1. Introduction 5](#_Toc136012716)

[2. Data and Libraries 5](#_Toc136012717)

[3. Exploratory Data Analysis 6](#_Toc136012718)

[4. Statistical Analysis 6](#_Toc136012719)

[4.1 Descriptive Statistics 6](#_Toc136012720)

[4.2 Inferential Statistics 9](#_Toc136012721)

[5. Machine Learning Models 11](#_Toc136012722)

[5.1 Classification 11](#_Toc136012723)

[5.2 Regression 12](#_Toc136012724)

[5.3 Clustering and PCA 14](#_Toc136012725)

[5.4 Time Series Analysis 18](#_Toc136012726)

[5.5 Sentiment Analysis 20](#_Toc136012727)

[6. Optimisation 22](#_Toc136012728)

[6.1 Efficiency 22](#_Toc136012729)

[6.2 Data Manipulation 22](#_Toc136012730)

[7. Interactive Dashboard 22](#_Toc136012731)

[8. References 23](#_Toc136012732)

# Introduction

The construction industry has recently received a lot of attention in mainstream media. Ireland is currently in a housing crisis due to a combination of a growing population and lack of affordable housing. There is said to be a lack of understanding surrounding data involving the construction industry which is making it difficult to effectively make decisions surrounding the planning and development of housing. The construction industry has often been volatile with many years highlighted particularly during the Celtic Tiger.

This assessment looks at various data sets on areas such as construction employment, construction cost indices, property price indices and more for Ireland as well as data for other countries. The data is manipulated and analysed statistically and by the relevant machine learning models in order to gain insight.

# Data and Libraries

|  |  |
| --- | --- |
| **Name:** | National House Construction Cost Index |
| **Ref:** | (DATA.GOV.IE, 2016) |

|  |  |
| --- | --- |
| **Name:** | Indices of Total Production in Building and Construction Sector |
| **Ref:** | (DATA.GOV.IE, 2022) |

|  |  |
| --- | --- |
| **Name:** | Earnings and Hours Worked in Building and Construction Industry |
| **Ref:** | (DATA.GOV.IE, 2020) |

|  |  |
| --- | --- |
| **Name:** | Labour Input in Construction Europe |
| **Ref:** | (eurostat, 2023) |

|  |  |
| --- | --- |
| **Name:** | Construction Production Index Europe |
| **Ref:** | (eurostat, 2023) |

Different data types were used including CSV and JSON. The main difference between CSV and JSON is that JSON files contain nested structures. CSV is said to have a faster processing speed, take up less space and be more secure so is generally better for handling large datasets (IndustryTrends, 2022). To import the JSON file various methods could be used such as Ujson module, Orjson library, Pyjstat or read\_json from Pandas. Pyjstat library was chosen as it has functionality to read json-stat formats which is a commonly used format in statistical organisations.

Web/data scraping was conducted for sentiment analysis and converted to a json file as a dataset was not available. This is an efficient means of collecting data from public sources (Khan, 2021). The data was obtained from Reddit. Typically, twitter would have been a more useful platform for web scraping but has recently altered its permissions so was not a viable option. A paid subscription is required to avail of the Twitter API. Reddit has a similar API to Twitter. It requires special authentication for commercial use of the data and has rules around how you can use the data. The PRAW (Python Reddit API Wrapper) module was implemented. This allows simplified access to the Reddit API from python.

# Exploratory Data Analysis

First steps of EDA are to look at the structure and properties of the data. This was done by using various commands such as .info, .shape, dtypes and .describe. By using these it was possible to confirm the number of rows/columns, non-null values, number of unique values etc. This then lay the foundation for the alterations required to the data. Columns had to be converted to the appropriate data types using either the .astype or pandas pd.to\_numeric, numerical columns being converted to float/ int and categorical columns to objects. In certain datasets there were letters and special characters present in the numeric columns so these had to be removed before changing the type.

Null values are always a big part of data cleaning and EDA. It is often difficult to determine how best to deal with these. For sparse missing values the issue can often be resolved by back/forward filling or replacing with the mean. When data is missing for a certain section and the data is sporadic, it is not logical to replace nans with the mean and not possible to back/forward fill as there are too many missing values. This issue was encountered with the construction\_cost\_index and property\_price\_index merged df. A number of values were missing for construction\_cost\_index so the impact of dropping the rows which contained nan on the distribution of the property\_price\_index had the be assessed. Distribution of the data was looked at before and after dropping the values. It was then concluded that this had no significant effect, and the values were permanently dropped.

Boxplots were used to determine whether datasets contained many outliers. Density plots were used to visualise the distribution of the data and get an idea of the normality of the data. Correlation plots were used for determining relationships between different variables.

When combining construction\_cost\_index and property\_price\_index a left merge was implemented based on the common date columns. This was chosen as the dataset had common dates and different index types, so this allowed an additional column being added to the right of the dataframe.

# Statistical Analysis

## Descriptive Statistics

Descriptive statistics summarizes the characteristics of a dataset and mainly uses central tendency, distribution, measure of variability and correlation (Bhandari, 2020). Central tendencies are based on the central limit theorem and include mean, median and mode. The central limit theorem works off the principle that a sample population has a mean of μ and standard deviation σ. Measure of variability are based on variance which include deviation, variance, kurtosis and skewness. Kurtosis represents the ‘tailedness’ of the data. Skewness represents the symmetry of the data. Negative skew means the tail is to the left said of the distribution, positive means to the right side. The closer the skewness value is to 0, the more symmetrical the data is.

Boxplots are useful visualizations for checking for outliers and visualizing the spread of data between the different quartiles. Figure 1 below shows the boxplot for average weekly hours worked and average hourly earnings in the construction industry. From the plot there are a small number of outliers present for the hourly earnings and no outliers present for the average hours worked. The mean of the hourly earnings is located towards the first quartile more than the centre of the box. The whiskers of the box are also significantly longer for the hourly earnings than the hours worked indicating a larger spread of data which is supported by the presence of outliers.

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 2 Boxplot of hours worked/ hourly earnings

The displot shows us the distribution of the data from which we can conclude the data is approximately normal. From the plot it can be said the average earnings per hour is slightly positively skewed. This is confirmed by the values in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Median | Skew | Kurtosis |
| Average earnings per hour | 8.83 | 7.57 | 0.69 | -0.20 |
| Average hours worked | 43.82 | 44.70 | -1.00 | 0.04 |

A picture containing text, screenshot, diagram

Description automatically generated

Figure 1 Density plot of average hours worked/ average earnings

## Inferential Statistics

Inferential statistics uses analytical tools. It is useful for drawing conclusions about the population. For inferential statistics to be meaningful the sample should be representative of the entire population. Several inferential statistics tests were used to compare Ireland with other countries:

* T-test is a parametric test which is used to compare the means of two groups and is frequently used in hypothesis testing to gauge whether groups are similar (Bevans, 2022). There are different variations of the t-test such as paired t-test, one sample t-test and two sample t-test. It was decided that a paired t-test would be the best option as samples are from the same population. Two sample t-test is conducted by setting ‘equal\_variance=True’. Bartletts test was used to determine whether variances were equal in a population (McClenaghan, 2023). .
* Kruskal-Wallis H-test.
* Mann Whitney U-test.
* Wilcoxon Signed-Rank.
* ANOVA is a parametric test.

It was decided to consider the number of construction workers in Ireland against that of the number of New Zealand. This comparison was chosen as both countries have similar populations. Labour Input index was compared between Ireland and Finland. Finland has a population almost ¼ of Ireland but was looked at for the purpose of investigation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Construction Employ- Inferential Statistic Test** | **Statistic Value** | **P value** | **Accept/Reject H0** |
| Paired T-test | -14.311 | 0.0001 | Reject |
| Kruskal-Wallis H-test | 12.500 | 0.009 | Reject |
| Welches T-Test | -11.000 | 3.95e-05 | Reject |
| Mann Whitney U test | 0.000 | 0.008 | Reject |
| Wilcoxon Signed-Rank | 0.000 | 0.062 | Accept |
| ANOVA | 121.000 | 4.18e-06 | Reject |

|  |  |  |  |
| --- | --- | --- | --- |
| **Labour input Index - Inferential Statistic Test** | **Statistic Value** | **P value** | **Accept/Reject H0** |
| Paired T-test | 2.75 | 0.006 | Reject |
| Kruskal-Wallis H-test | 2.165 | 0.141 | Accept |
| Welches T-Test | 1.29 | 0.197 | Accept |
| Mann Whitney U test | 38833 | 0.141 | Accept |
| Wilcoxon Signed-Rank | 14326.5 | 0.018 | Reject |
| ANOVA | 1.667 | 0.197 | Accept |

The findings from the inferential tests can be seen above. It is interesting as different results were obtained using different tests. Wilcoxon Signed-Rank test is the only one to accept H0 considering employment in Ireland to New Zealand. This could be attributed to the fact this test compares sample distributions whereas the t-test, ANOVA and Kruskal -Wallis involve the means of the populations. The same tests were completed to determine similarities between Ireland and Finland for labour input index. The only tests that reject H0 are paired t-test and Wilcoxon. Overall, we can say Ireland does not have similar data to New Zealand for construction workers and does have similar labour input to Finland.

# Machine Learning Models

Different machine learning models were trialled on the data. Supervised and unsupervised models were used. Supervised models included classification, regression and time series analysis along with unsupervised models such as clustering and sentiment analysis were used. The datasets used to complete the modelling are construction cost index/property price index and construction employment. For the purpose of sentiment analysis, data was obtained by web scraping from Reddit using PRAW as discussed in section 2.

## Classification

Different classification algorithms were applied to the data to determine which algorithm gave the best result. The data was split using sklearn and a test/train split of 20/80 as this is a standard value frequently used (Stojiljkovic, 2023). Cross validation (CV) was applied to the different classification algorithms. CV is used to gain insight on the skill of the machine learning model (Brownlee, 2020). CV utilises a parameter ‘k’ which represents the number of groups the sample will be split into.

* **KNN**. This classifies the data based on the distance between the training/test set (Christopher, 2021). To get optimal accuracy a number of k values between 0-30 were trialled. From the below plot the highest accuracy seems to be achieved using an n\_neighbours value= 9.

A picture containing text, diagram, line, plot

Description automatically generated

Figure 6 Graph of KNN accuracy with varying n\_neighbours or train and test sets

* **Decision Tree.** Each node represents a feature, the branches represent rules (Naviani, 2023). There is a built-in classifier in scikit-learn which is simple yet efficient.
* **SVM Classification.**  This works by finding a hyperplane to classify the data points (Gandhi, 2018). Parameters requiring specification are kernel and c. RBF and poly kernels were used to assess which was more accurate. C parameter has a significant effect on the model. If c is too small, then it may result in misclassification of the training data. A value of 10 was decided for the model.
* **Hyper parameter tuning**. GridSearch CV was used for SVM.

The best model is achieved using the decision tree classification as it achieves a high accuracy and high CV score. Although SVM gives better accuracy the cv score is the lowest of all models meaning the model does not generalise well which could result due to overfitting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **CV Score** |
| KNN | 0.93 | 0.94 | 0.93 | 0.93 | 0.87 |
| Decision Tree | 0.95 | 0.96 | 0.95 | 0.94 | 0.95 |
| Support Vector Machine (rbf) | 0.96 | 0.96 | 0.97 | 0.96 | 0.61 |
| Support Vector Machine (poly) | 0.93 | 0.92 | 0.91 | 0.91 | 0.84 |

## Regression

Regression is used to investigate relationships between independent and dependant variables (Castillo, 2021). It is a fundamental base for prediction and forecasting. Regression plots a line of best fit and aims to produce a model in which the distance between the data and the line is minimized as much as possible.

* **Linear Regression**: Many points are significant distance from the line of best fit. It can be anticipated the model will not be very accurate.

A picture containing screenshot, text, line, plot

Description automatically generated

Figure 7 Linear Regression Plot

* **RANSAC:** This is most useful when the dataset has outliers as it handles the outliers within the model rather than having to do it manually beforehand. Fig X shows that RANSAC did not detect any outliers for the data. This is contrary to what was found during EDA where a small number of outliers were detected.

A picture containing screenshot, text, plot, line

Description automatically generated

Figure 8 RANSAC Regression Outliers/Inliers plot

* **Ridge Regression:** Like linear regression but uses ‘ridge estimator’ instead of ordinary least squares. Training/test data are extremely similar.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 9 Ridge regression predicted values

* **Decision Tree Regression:** Works by fitting a sine curve.

A picture containing screenshot, line, plot, diagram

Description automatically generated

Figure 10 Decision Tree

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | R Score | | Mean Square Error |
| R2 Train | R2 Test |
| Linear | 0.36 | 0.41 | - |
| Lasso | 0.36 | 0.42 | 33.96 |
| Ridge | 0.36 | 0.42 | - |
| Elastic Net | 0.02 | 0.31 | 39.52 |
| Decision Tree | 0.93 | 0.95 | - |
| RANSAC | 0.36 | 0.42 | 25.43 |

The best regression model for this data was found to be the decision tree. Overall, the accuracy scores of the data were poor which could be an indication of a lack of significance between variables.

## Clustering and PCA

Clustering and PCA are unsupervised machine learning algorithms and are often used for gaining insight to patterns in data. (Brownlee, 2020). The scikit-learn library provides multiple different clustering algorithms.

* **Principal component analysis** (PCA) is used in order to determine dependencies among different variables and reduce dimensionality without altering the characteristics of the data (Sugiyarto, 2021,). PCA is particularly useful for data which contains large amounts of dependent variables.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 11 Cumulative Variance

* **K-means clustering:** This is one of the most simplistic algorithms which works by separating the data by equal variance and is based on 3 basic steps, assign points to the nearest centroid, create new points using the mean assigned to each point for the previous centroid and get the difference between the new and existing centroids (scikit-learn.org, 2023). Elbow method was used to determine the optimal k value.

A graph with a line

Description automatically generated with low confidence

Figure 12 Elbow Point graph

* **Gaussian Mixture model** is a probabilistic model that assumes various Guassian distributions present, each of which are a cluster (Singh, 2019). It is the fastest prediction algorithm and uses all components present in a dataset. The only parameter required to be specified is number of components. Number of components can be assessed on a density basis using Akaike information criterion (AIC) or Bayesian information criterion (BIC) (VanderPlas, 2016). This is a measure of effectiveness as a density estimator but can often be a good indication for clustering efficiency. The value was determined by Figure 13 to be 3 as this is the point at which both BIC and AIC begins to increase after dropping.

A picture containing diagram, text, plot, screenshot

Description automatically generated

Figure 13 Number of Components graph

A picture containing text, screenshot, map, diagram

Description automatically generated

Figure 14 Clustering - Gaussian Mixture Method

* **DBSCAN** is a density-based clustering model. The parameters required for DBSCAN are eps and minimum points. Eps defines the distance between data points for them to be considered neighbours. If the Eps value is too large it will mean most points will be in the same cluster, if the value is too small then the model will determine a large number of points to be outliers (Dey, 2023).

An appropriate eps value can be determined by. Min\_points is the fewest number of points to form a cluster. It is generally recommended to be at least three and should increase with an increasing sample size. If the min\_points value is set to 1 it will result in each point representing its own cluster (Sefidian, 2023). For this model an arbitrary value of 12 was chosen for n\_neighbours. Knn distance was then found by using the elbow/knee plot. ‘KneeLocator’ was used to confirm the location so to avoid any error from reading the point from the graph. A value of 8.61 was determined.

A picture containing text, screenshot, plot, line

Description automatically generated

Figure 15 Elbow/knee graph for KNN distance

Optimal min\_samples was determined by passing a range 1 to 20 minimum samples with the optimum eps value. Min\_samples was then plotted against number of samples to determine the optimal value which returned a value of 7.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 16 DBSCAN cluster plot

Silhouette score is a means of determining the efficiency of a clustering techniques and its values range from -1 to 1. A value of 1 indicated are well spread and clearly defined. A value of 0 means the clusters are not significant and a value of -1 means clusters are assigned incorrectly (Bhardwaj, 2020).

|  |  |
| --- | --- |
| **Clustering Algorithm Used** | **Silhouette Score** |
| K-means | 0.47 |
| Gaussian Mixture model | 0.55 |
| DBCAN | 0.41 |

By analysing the silhouette scores obtained from the clustering techniques it can be said that the clusters are defined and assigned correctly but not entirely significant. I believe further investigation on the number of neighbours ideally should be done as this massively effects the data. Overall, it can be said that the data present is not ideal for being clustered. If the data was of higher dimensionality and contained more groups/unlabelled data, the effectiveness of the clustering model would be far greater.

## Time Series Analysis

Time series analysis provides insights to features of a dataset which change over time. The only assumption of TSA is that “the origin of time does not affect the properties of the process under statistical factor” (Pandian, 2023). The main steps of TSA are determining stationarity, model building and extracting insights. There are different ways of determining stationarity including ADF and KPPS. For the purpose of the assignment both the KPSS test and ADF were investigated. KPPS test for stationarity on a deterministic trend meaning the slope of the given trend generally stays the same (Prabhakaran, 2019). ADF assumes H0 = series is not stationary and KPPS assumes H0= series is stationary therefore interpretation of the results for the two are opposite. Both tests determined the series to be stationary.

|  |  |  |
| --- | --- | --- |
| Test | p-value | Conclusion |
| KPPS | 0.03 | Series is stationary |
| ADF | 0.30 | Series is stationary |

A picture containing text, line, plot, diagram

Description automatically generated

Figure 17 Trend/ Season and Residual TSA plots

A train/test split was used of approximately 25%. The year 2000 was specified for the split. Training data used was values less than 2000 and testing data used was values greater than 2000.

A graph showing the growth of construction employment index

Description automatically generated with low confidence

Figure 18 Train/ test split for TSA

There are numerous models which can be used for forecasting including AR, MA, ARMA and ARIMA. For this assignment the ARIMA was chosen as it is described as a combination of AR and MA while also includes an integral operator (Shetty, 2020). It is said to perform well on short term forecasts. As we can see from Figure 18 the forecasted value differs from the actual value, but it is directionally correct and within the 95% confidence interval so can said to be acceptable. This model could be further improved by adding a constant in order to reduce AIC and give a more accurate model (Prabhakaran, 2021).

A picture containing plot, diagram, line, text

Description automatically generated

Figure 19 ARIMA Forecast plot

## Sentiment Analysis

Sentiment Analysis (SA) is a branch of Natural Language Processing (NLP) which classifies strings/words into categorical sentiments (Arora, 2022). SA typically involves steps such as tokenization, feature extraction, classification, post-processing and evaluation. There are several Python libraries available for SA including text blob, VADAR, Bag of Words, LSTM and Transformer based models. Data first needed to be obtained by Web Scraping Reddit, (refer to section 2). Typically, Twitter would have been a better option as the data already had sentiment applied but due to configuration constraints this was not an option. Initially a model was run using VADER, a rule-based SA specifically for sentiments in social media which scores the degree of positivity/negativity (geeksforgeeks.org, 2023). VADAR is open-source under the MIT License making it extremely useful. VADER has a built in Sentiment intensity analyser from NLTK. Several Sub-Reddits were specified along with key word combinations relevant to the construction industry and run through the sentiment analyser.

A picture containing screenshot, rectangle, line, square

Description automatically generated

Figure 20 VADER results

VADER accuracy was then tested on a subset of the JSON data which had been annotated manually based on the title perception. The manually annotated feature and VADER feature are used in conjunction (DeLancey, 2020).

Additional models were created. These included a model using NBClassifier in NLTK to train the data, TfidfVectorizer using MultinomialNB classifier to train the data and another tfidf using logistic-regression. TfidfVectorizer –Linear Reg gave the most inaccurate result. The best model to use would be the VADER or Bag of Words. The accuracy could be further improved with a larger sample size, more key words included and advanced cleaning of the data.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| VADER | 0.60 |
| Naïve Bayes | 0.44 |
| Bag of Words – Count Vect | 0.56 |
| TfidfVectorizer – MultiNom | 0.48 |
| TfidfVectorizer –Linear Reg | 0.26 |

# Optimisation

## Efficiency

Where possible ‘pipinstall’ was used instead of Conda. Pip is built into python. Conda is an open-source package management system (Brar, 2020). The use of Conda for packages was often slow to install. Redefining variables was avoided where possible as this was taking up memory. Loops were frequently used to avoid reptation, for example plots.

## Data Manipulation

Different techniques were required for the different data structures. Data was obtained by web scraping for the sentiment analysis section. To achieve this on python either ‘PRAW’ or ‘Request’ could be used. PRAW (Python Reddit API Wrapper) it is used in conjunction with pandas. There are many built in simple aggregations using the Pandas and Numpy library in Python. Some of these aggregations can be seen in the table below (VanderPlas, 2016). Another aggregation available is ‘GroupBy:’ which allows splitting and grouping of data based on specific conditions. The specified group can be return as either a Series or DataFrame. This groupby is particularly useful when using .describe for a dataset with different categorical variables contained in the rows. ‘aggregate’ is like ‘groupby’ and can use a string/ list and calculate numerous values at once. Filtering is also accurate to reduce the data set based on desired values and returns a Boolean value.

Pandas is said to be slower on smaller datasets and indexing than Numpy (Pathak, 2022). Numpy arrays are said to be faster than Panadas data frames. Numpy is also said to use less memory as Pandas often stores datatypes as object which can result in additional memory consumption. This makes it important to avoid redefining where possible and using parameters eg, ‘inplace=True’ to make permanent modification to the existing dataframe. Pandas is more powerful at reading in external data as it can read various file formats while Numpy is restricted to text and CSV files and not compatible with JSON files.

A screenshot of a computer

Description automatically generated with low confidence

# Interactive Dashboard

Python offers different packages to create dashboards including Dash, Mercury and Voila. Dash was chosen as the package. Dash is an open-source framework that’s core revolves around flask for server functionality, React.js for interface to the web and plotly.js for graphics. When initially attempting to use Mercury, it proved difficult to render the dashboard from the notebook without running in an external command prompt. Jupyter Dash extension for Dash allows seamless deployment directly from the notebook. An added benefit of Dash is the readily available pre-built components from libraries such as dash bootstrap components. Dash also allows conditional variables to be set through its clever use of call-back functions. This functionality was vital for the interactive aspect of the dashboard.

# References

Arora, S., 2022. *Sentiment Analysis Using Python.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/  
[Accessed April 2023].

Bevans, R., 2022. *An Introduction to t Tests | Definitions, Formula and Examples.* [Online]   
Available at: https://www.scribbr.com/statistics/t-test/#:~:text=A%20t%20test%20is%20a,are%20different%20from%20one%20another.  
[Accessed April 2023].

Bhandari, P., 2020. *Inferential Statistics | An Easy Introduction & Examples.* [Online]   
Available at: https://www.scribbr.com/statistics/inferential-statistics/  
[Accessed April 2023].

Bhardwaj, A., 2020. *Silhouette Coefficient Validating clustering techniques.* [Online]   
Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished.  
[Accessed April 2023].

Brar, K., 2020. *Understand Conda and Pip.* [Online]   
Available at: https://medium.com/analytics-vidhya/understand-conda-and-pip-9e5c67da47cc#:~:text=Pip%20installs%20Python%20packages%20whereas,the%20packages%20installed%20in%20them.  
[Accessed April 2023].

Brownlee, J., 2020. *10 Clustering Algorithms With Python.* [Online]   
Available at: https://machinelearningmastery.com/clustering-algorithms-with-python/  
[Accessed April 2023].

Castillo, D., 2021. *Machine Learning Regression Explained.* [Online]   
Available at: https://www.seldon.io/machine-learning-regression-explained#:~:text=Machine%20Learning%20Regression%20is%20a,used%20to%20predict%20continuous%20outcomes.  
[Accessed April 2023].

Christopher, A., 2021. *K-Nearest Neighbor.* [Online]   
Available at: https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4  
[Accessed April 2023].

DATA.GOV.IE, 2016. *Annual New Property Prices.* [Online]   
Available at: https://data.gov.ie/dataset/annual-new-property-prices?package\_type=dataset  
[Accessed March 2023].

DATA.GOV.IE, 2016. *National house Construction Cost Index.* [Online]   
Available at: https://data.gov.ie/dataset/national-house-construction-cost-index?package\_type=dataset  
[Accessed April 2023].

DATA.GOV.IE, 2020. *BDA01 - Earnings and Hours Worked in Building and Construction Industry.* [Online]   
Available at: https://data.gov.ie/dataset/bda01-earnings-and-hours-worked-in-building-and-construction-industry-1969-to-2008?package\_type=dataset  
[Accessed April 2023].

DATA.GOV.IE, 2022. *BEA04 - Indices of Total Production in Building and Construction Sector (Base 2015=100).* [Online]   
Available at: https://data.gov.ie/dataset/bea04-indices-of-total-production-in-building-and-construction-sector-base-2015100/resource/5a790502-2001-452b-978c-82ecf6f5b9e0  
[Accessed April 2023].

DATA.GOV.IE, 2022. *WPA15 - Wholesale Price Index (Excl VAT) for Building and Construction Materials.* [Online]   
Available at: https://data.gov.ie/dataset/wpa15-wholesale-price-index-excl-vat-for-building-and-construction-materials?package\_type=dataset  
[Accessed April 2023].

DeLancey, J., 2020. *Improving NLTK Sentiment Analysis with Data Annotation.* [Online]   
Available at: https://www.codeproject.com/Articles/5269453/Improving-NLTK-Sentiment-Analysis-with-Data-Annota  
[Accessed April 2023].

Dey, D., 2023. *DBSCAN Clustering in ML | Density based clustering.* [Online]   
Available at: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/  
[Accessed April 2023].

eurostat, 2023. *Labour input in construction - annual data.* [Online]   
Available at: https://ec.europa.eu/eurostat/databrowser/view/sts\_colb\_a/default/table?lang=en  
[Accessed April 2023].

eurostat, 2023. *Short term business statistic.* [Online]   
Available at: https://ec.europa.eu/eurostat/web/short-term-business-statistics/data/database  
[Accessed April 2023].

Gandhi, R., 2018. *Support Vector Machine — Introduction to Machine Learning Algorithms.* [Online]   
Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47  
[Accessed April 2023].

geeksforgeeks.org, 2023. *Python | Sentiment Analysis using VADER.* [Online]   
Available at: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/  
[Accessed April 2023].

IndustryTrends, 2022. *CSV or JSON, which Format is better for your AI training data?.* [Online]   
Available at: https://www.analyticsinsight.net/csv-or-json-which-format-is-better-for-your-ai-training-data/  
[Accessed April 2023].

Khan, S., 2021. *Web scraping social media giants: Facebook, Twitter and Reddit.* [Online]   
Available at: https://www.computertechreviews.com/web-scraping-social-media-giants/  
[Accessed April 2023].

McClenaghan, E., 2023. *The Wilcoxon Signed-Rank Test.* [Online]   
Available at: https://www.technologynetworks.com/informatics/articles/the-wilcoxon-signed-rank-test-370384  
[Accessed April 2023].

Naviani, A., 2023. *Decision Tree Classification in Python Tutorial.* [Online]   
Available at: https://www.datacamp.com/tutorial/decision-tree-classification-python  
[Accessed April 2023].

NoteableInc., 2023. *Python JSON Loads – 4 Faster Alternatives to json.loads in Python | Ojrson Ujson Pandas Msgspec.* [Online]   
Available at: https://noteable.io/blog/fast-json-loads-in-python/#:~:text=orjson%20is%20a%20fast%20and,in%20C%20for%20maximum%20speed.  
[Accessed April 2023].

Pandian, S., 2023. *Time Series Analysis and Forecasting | Data-Driven Insights (Updated 2023).* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/  
[Accessed April 2023].

Pathak, A., 2022. *Pandas vs NumPy in Data Science: Top 15 Differences.* [Online]   
Available at: https://www.knowledgehut.com/blog/data-science/pandas-vs-numpy  
[Accessed April 2023].

Prabhakaran, S., 2019. *KPP Test for Stationarity.* [Online]   
Available at: https://www.machinelearningplus.com/time-series/kpss-test-for-stationarity/  
[Accessed April 2023].

Prabhakaran, S., 2021. *ARIMA Model – Complete Guide to Time Series Forecasting in Python.* [Online]   
Available at: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/  
[Accessed April 2023].

scikit-learn.org, 2023. *2.3. Clustering.* [Online]   
Available at: https://scikit-learn.org/stable/modules/clustering.html  
[Accessed April 2023].

Shetty, C., 2020. *Time Series Models.* [Online]   
Available at: https://towardsdatascience.com/time-series-models-d9266f8ac7b0  
[Accessed April 2023].

Singh, A., 2019. *Build Better and Accurate Clusters with Gaussian Mixture Models.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/  
[Accessed April 2022].

Sefidian, A., 2023. *How to determine epsilon and MinPts parameters of DBSCAN clustering§.* [Online]   
Available at: http://www.sefidian.com/2022/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/  
[Accessed April 2023].

Stojiljkovic, M., 2023. *Split Your Dataset With scikit-learn's train\_test\_split().* [Online]   
Available at: https://realpython.com/train-test-split-python-data/  
[Accessed April 2023].

Sugiyarto, 2021,. *Clustering with Principal Component Analysis and Fuzzy Subtractive Clustering Using Membership Function Exponential and Hamming Distance.* [Online]   
Available at: https://iopscience.iop.org/article/10.1088/1757-899X/1077/1/012019/pdf  
[Accessed April 2023].

VanderPlas, J., 2016. *Python Data Science Handbook.* s.l.:O'Reilly Media, Inc..