**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **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*  *John O’Sullivan*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Clarissa dos Santos Cardoso* |
| **Student Number:** | *2020274* |
| **Assessment Due Date:** | *26/05/2023* |
| **Date of Submission:** |  |

**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. |

**MSc in Data Analytics FT - Feb 2023**

Author: Clarissa Cardoso

e-mail: [2020274@student.cct.ie](mailto:2020274@student.cct.ie)

Student ID: 2020274

# Abstract

*This report presents and overview of determined variables to investigate the number of new building permits over the period from 1977 to 2022 for the Irish State and followed by a broader analysis to compare Ireland with four other countries in the Euro Zone in terms of employment in construction, GDP per capita and investment in dwellings. Several techniques of data exploration, cleaning, preparation and visualisation were used to find the adequate statistical distribution patterns for application of Hypothesis Testing and Machine Learning models, including sentimental analysis, for the availability of construction workers to provide insights of future increase in workforce demands in construction sector considering Ireland as a starting point.*

Table of Contents

[Abstract 2](#_Toc136205508)

[List of Tables 4](#_Toc136205509)

[List of Figures 4](#_Toc136205510)

[Introduction 4](#_Toc136205511)

[1. Part I - Planning Permissions for New Houses in Ireland 5](#_Toc136205512)

[Overview of dataset 5](#_Toc136205513)

[1.1 Data Cleaning 6](#_Toc136205514)

[1.2 Exploratory Data Analysis and Data Visualisation 7](#_Toc136205515)

[1.3 Hypothesis testing: 12](#_Toc136205516)

[2. Part II - Comparing Ireland to other countries in the Euro Area 13](#_Toc136205517)

[2.1 Data Cleaning 14](#_Toc136205518)

[2.2 Exploratory Data Analysis and Data Visualisation 15](#_Toc136205519)

[2.3 Hypothesis Testing 20](#_Toc136205520)

[3. Data Preparation for Machine Learning 21](#_Toc136205521)

[4. Results for Machine Learning Models 26](#_Toc136205522)

[5. Sentimental Analysis: “Is there really a shortage of skilled tradesmen?” 29](#_Toc136205523)

[Conclusion 30](#_Toc136205524)

[References 31](#_Toc136205525)

[Appendices 38](#_Toc136205526)

[Appendix 1: GitHub 38](#_Toc136205527)

[Appendix 2: Project Management applying CRISP-DM (Luna, 2021) 39](#_Toc136205528)

[Appendix 3: Datasets available in CSO database 40](#_Toc136205529)

[Appendix 4: Outputs of .info, .shape and .isnull.sum 40](#_Toc136205530)

[Appendix 5: Comparing Seaborn and Plotly libraries for Visualisation. 41](#_Toc136205531)

[Appendix 6: Test results for checking Normal Distribution 44](#_Toc136205532)

# List of Tables

[Table 1 Summary statistics for dataframe analysed (Elaborated by author) 8](#_Toc136205533)

[Table 2 Hypotheis testing results 13](#_Toc136205534)

[Table 3 New csv files with national figures 15](#_Toc136205535)

[Table 4 Results of Parametric and Nonparametric Tests 21](#_Toc136205536)

[Table 5 Machine Learning Models - Comparison of Results 26](#_Toc136205537)

[Table 6 Classification Matrix for Sentimental Analysis 30](#_Toc136205538)

# List of Figures

[Figure 1 Output of head () function of data 7](#_Toc136205539)

[Figure 2 New heading for dataframe 8](#_Toc136205540)

[Figure 3 Boxplot of first column: Planning Permission Granted 9](#_Toc136205541)

[Figure 4 Pairplot of dataframe to check distribution pattern 10](#_Toc136205542)

[Figure 5 Histogram of planning permissions to check distribution 10](#_Toc136205543)

[6 Histogram of Planning permissions from 1977 to 2022 12](#_Toc136205544)

[7 Histogram created using Pyplot 12](#_Toc136205545)

[Figure 8 GDP Per Capita in the Euro Area 2022 15](#_Toc136205546)

[Figure 9 OECD's datframes to check headings 16](#_Toc136205547)

[Figure 10 Dwellings line chart 17](#_Toc136205548)

[Figure 11 Dwellings in Europe 18](#_Toc136205549)

[Figure 12 Line chart for GDP values 18](#_Toc136205550)

[Figure 13 Choropleth map for GDP per capita evolution 19](#_Toc136205551)

[Figure 14 People employed in construction 19](#_Toc136205552)

[Figure 15 Map of Europe comparing peopel working in construction 20](#_Toc136205553)

[Figure 16 Comparing two variables 20](#_Toc136205554)

[Figure 17 Distribution of Employment in Construction per country 21](#_Toc136205555)

[Figure 18 dataframe for ML modeling 23](#_Toc136205556)

[Figure 19 Negative Mean Absolute Error and alpha for Ridge Regression 24](#_Toc136205557)

[Figure 20 Negative Mean Absolute Error and alpha for Lasso Regression 25](#_Toc136205558)

[Figure 21 Ireland Values df 26](#_Toc136205559)

[Figure 22 Predictions from ML Model-Linear Regression 28](#_Toc136205560)

[Figure 23 Random Forest Regressor - Tree 29](#_Toc136205561)

[Figure 24 Decision Tree for merged\_df modelling scores 30](#_Toc136205562)

[Figure 25 ‘r/Ireland’ Reddit community: Post used for sentimental analysis 31](#_Toc136205563)

[Figure 26 Diagram of CRISP DM applied for the project created with draw.io based on Hotz diagram (Hotz, 2023) 40](#_Toc136205564)

# Introduction

The construction sector in Ireland have been showing positive signs of growth, especially in the post-covid recovery path. The most recent country profile for Ireland from the European Construction Sector Observatory (ECSO, 2022) shows that the sector is partially one of the drivers of the jump in the country’s GDP in the last decade, from EUR 188.8 billion in 2010 to EUR 353.8 billion in 2020 (ECSO, 2022, p. 2) , (OECD, 2023). Despite the overall positive macroeconomics trends, in reality the construction sector in Ireland, and other European nations, face a series of challenges to maintain production at high level while working with significant increases in inflation rates in general, but also the effects of economic downturn caused by the pandemic and the supply chain access such as the blockage in Suez Canal in 2021 and the current war in Ukraine as stated by a KPMG report on the “Drivers of Cost and Availability of Finance for Residential Development” published in 2022 are major contributors for the rise in cost of construction materials (KPMG, 2022). Aligned with the rise in costs of production, the costs of labour have also increased, caused mainly for the shortage of skilled workers for the sector, which saw an increase of 80% in job vacancies in the narrow construction sub-sector between 2010 and 2020 (ECSO, 2022, p. 18), showing the ongoing increase in demand and the estimation is that the sector will require “an additional 40,000 to 50,000 new construction workers until 2027” (ECSO, 2022, p. 2) .

Among the consequences of the costs of production/labour and the lack of trained professionals, the number of planning permissions issued for construction in Ireland have also showed a slight decrease since fewer projects are approved by the government as published by the Central Statistics Office (CSO), in which is one of the components of concerns related to the housing crisis in the country and the availability of new developments for the population in general. In general aspects, not enough people working in construction results in less projects being granted permission, therefore less houses being built (Burke-Kennedy, 2023) (KPMG, 2022)[[1]](#footnote-1).

This report is divided in two main sections to investigate data considering the relation between planning permissions for new houses and employment in construction. The project was developed using a Jupyter Notebook with use of Python programming language and to ensure the progress for a successful version control storage within GitHub (link for repository available in appendix 1). The first part analyses the evolution of planning permissions for new houses in Ireland from 1975 to 2022 with a dataset released by the CSO, where the material is freely available to share and adapt as stated by the Creative Commons 4.0 License (Creative Commons, 2023). Techniques of exploratory data analysis and data cleaning are performed in the dataset to enable understanding of distribution of the data present to perform hypothesis testing and create useful visual representations of it.

The second part aimed to compare Ireland to other countries in the Euro Area in three areas: GDP per capita, Employment in construction and Investment in assets by Dwellings, using the GDP per capita as a starting point to filter out the five countries with highest values in the last year to perform the required techniques for cleaning, visualisation, hypothesis testing and machine learning algorithms. The datasets used were obtained from Organisation for Economic Co-operation and Development (OECD) public database, in which all member countries operate in data compliance according to the 2008 System of National Accounts (SNA) (OECD, 2018), defined by the United Nations Statistical Commission (UNSC) as “the latest version of the international statistical standard for the national accounts” (UNSC, 2023). The database is available for extraction, distributed and shared for any purpose as long as the user give appropriate credit and referenced properly with citation available for all datasets which are also available under CC license (OECD, 2023).

In terms of project management frameworks decided to apply for the project, CRISP\_DM was selected for its popularity among the data science community and well defined structure to manage its life-cycle through the phases to obtain effective results. More details on how each phase was covered on appendix 2 (Hotz, 2023).

# Part I - Planning Permissions for New Houses in Ireland

### Overview of dataset

The CSO database has an extensive data regarding the planning permissions from the earliest of 1975 to date values spread in different categories (the full list of datasets available is in appendix 3). There is a overwhelmingly amount of data available for different types of permits (such as for residential, non-residential building, or type of dwelling), and the first challenge in this project was to filter the information and chose a single path to investigate. The dataset chosen for this report was the labelled “BHQ12” containing data of planning permissions for new houses and apartments from 1975 to 2022, and there was an option to filter the information between the type of dweling (new houses and/or apartments) so for this analysis the type houses was selected to result in a more cleared dataset to reflect on. According to Citizens Information(2023)the waiting time between requesting and obtaining a planning permission is on average 90 days, until the application is revised there is a set of requirements contractors must follow to be granted a permission.

## Data Cleaning

After importing the importing csv file containing Planning Permissions for New Houses and Apartments - Ireland (1975 to 2022) collected from the CSO database under the variable name “data” and displaying the heading (figure 1) to see main features present, a couple of first insights can be made: the dataframe contains 5 columns where the first shows years of data available and the others contain the total amount of permissions granted, the units for which permissions were granted with discrete numerical values and the last two contain continuous numeric values of area in squared meters for total floor area and the average floor area per unit, where each observation brings a different year. To better understand the nature of the dataset, the functions ‘.info( )’(names of columns, sum of non-null values, data types), ‘.shape’(49 rows and 5 columns) and ‘.isnull().sum()’ (to confirm sum of null values) were used on the same line to optimisation and to interpret their outputs (appendix 4) at the same time (Patil, 2018).

Figure Output of head () function of data

A picture containing text, screenshot, number, font

Description automatically generated

As stated above, it was expected the dataset displayed all numerical values, but the .info ( ) output displayed a majority of ‘objects’ as datatype for the first four columns, creating the need to clean the data according to this steps:

* Replace the commas for an empty space;
* Rename columns to shorten labels to improve readability
* Set the the column containing years as the new index, to display an organise dataframe
* After concluding the amount of NaN values was significantly small and would not interfere in major calculations, it was decided to drop the first 3 rows instead of filling the missing values with the mean values as suggested by the Pandas documentation (Pandas, 2023).
* Once the NaN values were dropped, all the objects were converted to numeric data type using the ‘pd.to\_numeric’ function to obtain a higher level success in the conversion than selecting ‘astype’ (Pandas, 2023).

The result of cleaning the data is displayed in figure 2 with the new heading as well as using the .info once again to confirm the new dtypes were converted correctly for integers.

Figure New heading for dataframe

A picture containing text, receipt, screenshot, font

Description automatically generated

## Exploratory Data Analysis and Data Visualisation

Once the data cleaning is done, it possible to obtain summary of the statistics for the dataframe within its 45 rows and 4 features. From the ‘.describe( )’ pandas function the mean of planning permissions granted is 10934, with a standard deviation of 5826.388 and the majority of datapoints on the second quadrantile is around the 14096 mark, illustrated in the boxplot in figure 3 and the table 1 (Zach, 2023).

Table Summary statistics for dataframe analysed (Elaborated by author)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Summary Statistics** | **Planning Permissions Granted** | **Total Units of Permission Granted** | **Permissions Granted (sq m)** | **Avg Area per Unit (sq m)** |
|  |  |  |  |
| **Count** | 45 | 45 | 45 | 45 |
| **Mean** | 10934 | 28359 | 3921.778 | 143.180 |
| **Standard Deviation** | 5826.388 | 18910.501 | 2714.805 | 28.940 |
| **Min value** | 3103 | 5389 | 1110 | 73.8 |
| **Q1 (25%)** | 6396 | 16719 | 2258 | 119 |
| **Median (50%)** | 9126 | 19964 | 2981 | 139.5 |
| **Q2 (75%)** | 14096 | 33473 | 4217 | 167 |
| **Max value** | 25751 | 75650 | 11278 | 206 |

* The mean value for Planning Permissions Granted is 10,934 with minimum value recorded is 3,103 for the year 2013 and maximum is 25,751 corresponding to 2004.
* The sum of Units that were granted planning permissions has a mean value of 28,359 and a standard deviation of 18910.501, with a minimum in 2012 (5,389) and a maximum value recorded in 2005.

Figure Boxplot of first column: Planning Permission Granted

A picture containing text, screenshot, diagram, line

Description automatically generated

The boxplot for each variable was created with the Ploytly function for dropdown menu, after a few attempts the result is in appendix 5, where ployly by default created a scale to reduce the decimals for the y-axis which improves comprehension of values (ie. 200k instead of 200,000). In addition, to display all variables in the same figure to check their overall characteristics, the different scales the data was originally collected affected the visualisation negatively, since the numbers for ‘Avg Area per Unit (sq m)’ are in a significant smaller scale than  for example ‘Total Units of Permission Granted’, for that reason a standard scale was applied to display all values in the same level (also available in appendix 5) showing outliers for each column (Galarnyk, 2022).

To better understand how the data is distributed, using Seaborn's pairplot function with the kernel density estimation to visualise the distribution of each variable as well as their overall shape, peaks and skewness of the data and possible relationship between various columns (Seaborn, 2022). This also allows to see a compact representation of the distribution to compare different variables at the same time and identify any patterns between them with less noise than a default pairplot, since it estimates the density based on the observed data points (fig4). From the shape of the bell curve is possible to infer the data does not follow a normal distribution which is a common aspect faced when working with real work data, which in this scenario shows a slight skew to the right as the details in fig 5 with the probability plot of the data in the first column with a parameter ‘dist=norm’ where the further from the red line the points are, the less probable is that the data is normally distributed (Frost, 2022) (Radečić, 2020).

Figure Pairplot of dataframe to check distribution pattern

A picture containing text, diagram, screenshot

Description automatically generated

Figure Histogram of planning permissions to check distribution

A picture containing screenshot, plot, line, text

Description automatically generated

Also, it was created a series of plots for featuring EDA stored in appendix 5, to see how the Planning Permissions changed over time through the use of different libraries for visualisation for histograms in bar charts and line charts for one or more variables, such as Seaborn and Pyplot with added interactive features (Plotly, 2023) (Yi, 2019) (Yi, 2021). Based on Muth’s (2018) work heavily inspired by Tufte’s principles (2001), it was considered using a more diverse colour palette beyond the default “RGB” by looking for complementary colour set and using less hues and colour neighbours to make visualisations less polluted and more appealing to the viewers. Besides, the colours in plots throughout this report were chosen while trying to avoid bright, saturated colours to transmit a more sober tones and highlight the important features while maintaining a proportional presentation of multiple variables in a clear way (Muth, 2018) (Plotly, 2023).

To illustrate the progression in planning permissions from 1977 to 2022 (considering the new dataframe without missing values), figure 6 displays a histogram as a bar chart using Plotly hover function to shows values of each year as the cursor moves over the bars (which could have been placed with a bigger spacing between them), followed by a line chart containing a histogram of all features, where we can see a sharp decrease in permits granted 2008 and 2010 with a slow recovery from 2016 to date. The following plot (fig7) contains a multivariable line chart also created with Plotly, illustrating the proportional progression of the dataset where we can better identify the values for their respective year following the whitegrid . The reduction in the overall permissions granted in recent years, as mentioned by Burke-Kennedy (2023), raises concern for the Government’s plan of achieving to build 33,000 new homes by 2030 under the “Housing for All” strategy (Burke-Kennedy, 2023) (KPMG, 2022).

Histogram of Planning permissions from 1977 to 2022

A picture containing text, screenshot, plot, diagram

Description automatically generated

Histogram created using Pyplot

A picture containing text, plot, diagram, screenshot

Description automatically generated

After standardising the dataset, a heatmap was created to check correlation between variables as illustraded in figure 8 using the ‘coolwarm’ continuous scale with highly correlated values are darker on the red scale (from -1 to 1) with Planning Permissions Granted is mostly correlated with Total Units of Permission Granted and Permission Granted (sq m) (Matplotlib, 2023):

Heatmap Correlation between variables in ireland\_permits[[2]](#footnote-2) dataset

A screenshot of a graph

Description automatically generated with low confidence

## Hypothesis testing:

As an attempt to reduce the skewness (fig x), the logaritimic transformation was applied to the dataset as well as the application of Shapiro-Wilk tests (for before and after obtaining log values with results in the table in appendix 6) (SciPy , 2023). However, according to Feng, et al (2014)*,* the log method cannot be solely used to transform distribution as it can have the opposite effect in some cases due to its limitations to perform hypothesis testing too far for the original data which is why the majority of hypothesis testing will be performed with non-parametric tests since they do not require to confirm a normal distribution to be applied (Bray, 2019). In addition, a Kolmogorov-Smirnov test for Planning Permissions Granted feature was also applied to check its distribution once again which at first resulted in a p-value = 0, so according to a Statology (2018) tutorial on the matter it suggested to use mu and sigma values to perform the test (Zach, 2018). After standardising the parameters, new p-value is 0.16 which is *>0.05,* meaning that there is sufficient evidence to reject 𝐻0 where the sample is consistent to a normal dist (Bray, 2019).

For the hypostasis testing 2 main tests were applied for the first dataset: a parametric two-sample t-test with the *log\_values* (obtained from the features [ *Planning Permissions Granted index=0 ]and [Permissions(sq m) index=1]* considering their mean values) and a non-parametrical Mann-Whitney U test for the original data without normal dist. For the t-test it was first considered thar the variances are equal for the two mean and second where there are not expected to be euqual (Aandahl, 2022). The results are in the following table (2), where we can see all p-values are located very close to 0 and smaller than the alpha, hence the result is that there is a strong evidence to reject the null-hypothesis that the there is a substantial difference between their distributions (Frost, 2017).

Table Hypotheis testing results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis Testing** | | | | |
| **Features** | **Parametrical** | | **Non-parametrical** | |
| **Two-sample t-test** | | **Mann-Whitney U** | |
| **stat** | ***p-value*** | **stat** | ***p-value*** |
| ireland\_permits\_log [ 0 ] (equal\_var)  ireland\_permits\_log [ 1 ] (equal\_var) | 8.803 | 1.04E-13 | - | - |
| ireland\_permits\_log [ 0 ]  ireland\_permits\_log [ 1 ] | 8.803 | 1.14E-13 | - | - |
| ireland\_permits[ 0 ]  ireland\_permits [ 1 ] | - | - | 1814.5 | 1.64E-12 |

# Part II - Comparing Ireland to other countries in the Euro Area

Initially, the objective was to analyse data from 2002 to 2022 for the top five countries with hieghst GDPs to have a more complete picture of the changes in patterns, however after considering the drastic changes in the first dataset containing planning permissions after the 2008 economic downturn, it was decided to collect data from 2012 to 2022 and consider only the more recent data with the aftereffects during the global recovery. The goal is to understand the current stage of the information to prepare the prediction models based on recent figures. To select the countries that would be compared with Ireland’s scenario, the GDP per capita measured in US dollars is the variable chosen to sort the values with the final ranking illustrated in figure 8. The datasets for this section were obtained from OECD’s database applying filter for period desired for the analysis and selecting only countries in the Euro Area which were stored on a same directory (“/Data”).

Figure GDP Per Capita in the Euro Area 2022

A picture containing text, screenshot, font, plot

Description automatically generated

From the ranking the analysis will be filtered for Luxemburg, **Ireland**, Netherlands, Austria and Belgium.

## Data Cleaning

After checking for any missing values, which are not present in the new dataframes, a few techniques for cleaning were used once all the headings were compared. First was to drop columns with values that would not be relevant for this analysis as displayed in fig 9, since the raw data was collected by the same organisation, the measurements are consistent and the only variables needed was the ‘LOCATION’, ‘TIME’ and ‘Value’. The ‘MEASURE’ column was kept at first but once the value columns were renamed incuding the measure for reference, it was also dropped.

Figure OECD's datframes to check headings

A screenshot of a computer

Description automatically generated with low confidence

There are a total of four datasets in the respective directory, where each was stored under a distinct variable name for data manipulation as stated on the table 3 below, where the ‘top\_five’ countries are presented for the same time frame, and have the same amount of observations to facilitate merging them into a single dataframe:

Table New csv files with national figures

|  |  |  |
| --- | --- | --- |
| **File name** | **Original Dataframe name in Jupyter Notebook + Shape** | **Variable name and shape() after cleaning and sorting values** |
| ['gdp(per\_capita)\_2022.csv', | gdp\_2022 (19, 8) | *Not applicable* |
| 'Investment\_by\_asset:Dwellings.csv', | Invest (187, 6) | top\_five\_invest (55,6) |
| 'gdp(per\_capita)\_euro.csv', | gdp\_euro (209, 6) | top\_five\_gdp (55.6) |
| 'Employment\_by\_activity:  Construction.csv'] | construct\_employ (186,6) | top\_five\_employ (55,6) |

## Exploratory Data Analysis and Data Visualisation

Once we obtained the data from Luxemburg, Ireland, Netherlands, Austria and Belgium, they were sorted by default in alphabetical order in the ‘top\_five’ dataframes for each indicator: Dwelling (% of GFCF), GDP per capita (in US$) and Employment in construction (in thousands of people) and a few line plots to compare each country were created with Plotly and its interactive function to show values with a cusor hover as well as a few choropleth maps to display the evolution of each variable over time among the European continent (Plotly, 2023), (Yi, 2019), (The Data Frog, 2021).

At the end of the Jupyter Notebook elaborated for this project available on the GitHub repository linked in appendix 1, some of these visualisations are available in the form of a dashboard created with the Dash framework combined with Plotly to enhance the data experience while creating customised and flexible solutions at a more marketable front, as stated by Schmitt (2023). Among other tools that could be used for creating a dashboard there is Panel, which is considered to be better integrated with Jupyter Notebook and Python in general whereas if a project were to be developed in R language, Shiny library would be the appropriate choice since it was developed specifically for it (Schmitt, 2023) , (Vu, 2022). Due to time limitations to produce important steps for other sections in the report, not all plots are present in the dashboard, as well as the plots produced with the Panel in earlier stages were removed from the notebook since the Dash tutorial for creating “A Minimal Dash App” worked more effectively once all the necessary plots were renamed, added to the *app* and displayed results in line (Dash , 2023).

* Plots for Investment by asset: Dwellings (% of GFCF), Euro Area from 2012 to 2022:

Figure Dwellings line chart

A picture containing line, plot, diagram, screenshot

Description automatically generated

Despite Ireland having the second highest GDP last year, the country in position in last within the level of investment in residential properties with an average of 7.11, a direct impact of the housing crisis and a reflection from the reduction of planning permissions granted in the past decade as seen on the first part of this report. And the opposite happens with Belgium, where the investment in dwelings is the highest, with an mean value of 25.303. the other countries follow a similar pattern, with Nederlands having a high oscilation of values until 2016 when it showed steady growth afterwards. The color scale in the maps below shows Ireland on the lighter shade reflecting the values.

Figure Dwellings in Europe

A picture containing text, screenshot

Description automatically generated

* Plots for Comparison of GDP (in US dollars per capita) - 2012 to 2022:

Figure Line chart for GDP values

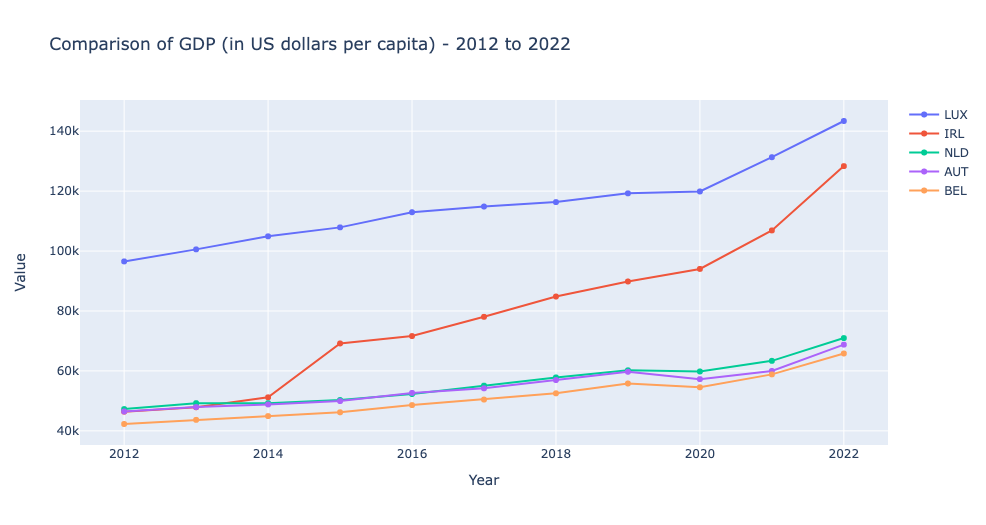
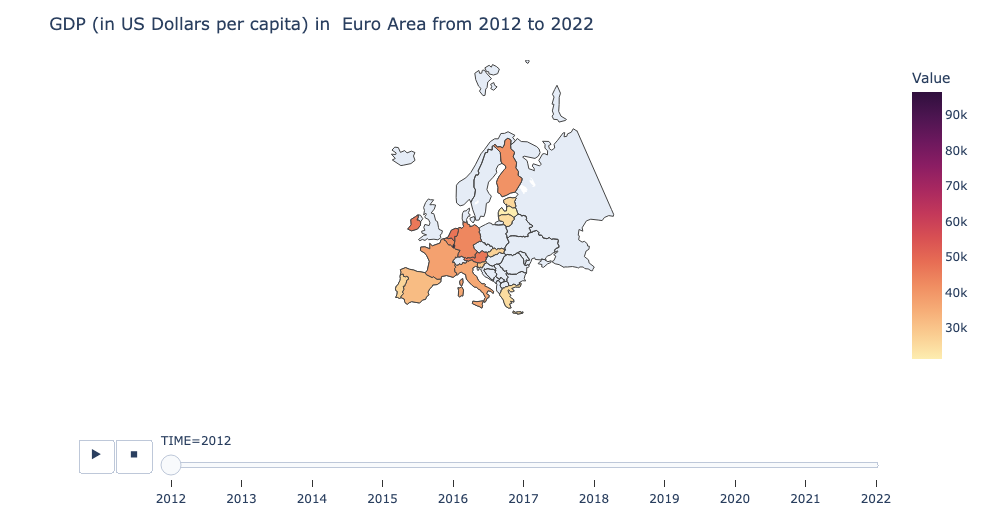


Figure Choropleth map for GDP per capita evolution



* Plots for Employment rates by activity: Construction (in thousands of persons):

In terms of people employed in construction activities we can also identify Ireland has a lower average than Belgium, Netherlands and Austria, with a maximum value of 165k of workforce in 2022, reflecting the rising demand for engaging people in the sector (Fox, 2022). Although a more in depth analysis would require to also investigate the employment rate based on total population for each country, which can be added in further projects that can bring more insights for understanding the context of employment.

Figure People employed in construction

A picture containing text, screenshot, line, plot

Description automatically generated

Figure Map of Europe comparing peopel working in construction

A map of europe with different colored squares

Description automatically generated with low confidence

Combining all variables into the same dataset was an extra challenge until finding and tuning the right features to be dropped, pivoted and finally merged, the goal was to set the time as an index but since it follows a time series structure it was not possible. Also, to plot visualisations, it became a polluted bar char to display all values at once, so only GDP and Employment in Construction were considered for this section and the following stages of analysis. With a dropdown menu it was possible to make it more interactive and select each country to compare their patterns over the years with multiple y-axis in different respective scales (an example of the output in fig 16) (Plotly, 2023).

Figure Comparing two variables

A graph with red and blue lines

Description automatically generated with low confidence

## Hypothesis Testing

The hypothesis testing for this part was done based on the Employment in Construction in Euro Area dataframe after pivoting the values to display each country in separate columns to check their distribution as fig 17 demonstrates:

Figure Distribution of Employment in Construction per country

A picture containing diagram, line, plot, text

Description automatically generated

A Shapiro-Wilk test was also applied to confirm or reject the null hypotesis that the dataframe follows a normal distribution, with the independent results in below where the p-values > 0.05 highlighted are an indication that parametric tests can be applied for two countries: Belgium and Ireland since we fail to reject the H0 that they would be normally distributed (Frost, 2022).

Variable Test Statistic P-value

0 AUT 0.937240 0.488612

1 BEL 0.956662 0.729475

2 IRL 0.944284 0.572085

3 LUX 0.852163 0.045541

4 NLD 0.808691 0.012242

A two-sample t-test was applied for the highlithed countries while for the other countries, nonparametric tests can be used since they do not depend on assumptions about the shape of the data (Aandahl, 2022) (Frost, 2017). In this case a Kruskal-Wallis test was applied to compare the means between the countries where the p-value obtained was of *1.79e-10* significantly close to zero, where we reject 𝐻0, since the mean values are distant from one another as the results can be compared in table 4.

Table Results of Parametric and Nonparametric Tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis Testing** | | | | |
| **Features** | **Parametrical** | | **Non-parametrical** | |
| **Two-sample t-test** | | **Mann-Whitney U** | |
| **stat** | ***p-value*** | **stat** | ***p-value*** |
| BEL, IRL | -23.531 | 1.75E-11 | - | - |
| AUT, BEL, IRL, LUX, NLD | - | - | 51.448 | 1.80E-10 |

# Data Preparation for Machine Learning

* 1. – Data preparation using ‘*top\_five\_employ’ df*

For this section, the data preparation for Machine Learning models and its application will first take into consideration the same dataset used for Hypothesis Testing: "Employment in Construction in Euro Area" under the variable “*top\_five\_employ*” with data from the 5 countries selected based on highest GDP’s rates. Based on the dataset of construction employment (in thousands of people) this project will apply Regression models to attempt a prediction of many workers each country will require based on the data available from 2012 to 2022. The goal is to predict if the growth in amount of workers will be sufficient to cover the expected demand as published by ECSO (ECSO, 2022).

Prior to applying machine learning algorithms, we must consider the nature of the data available. Using the same dataset from the previous section containing data of employment in construction for the 5 countries analysed, there are a few considerations from the way the features are correlated, as illustrated in the heatmap 2 where a yellow-blue colour scale was applied where the higher correlation (the more close to 1), the darker it is, as opposed to the light – yellow tones regarding variables less correlated, in this dataset, NLD and AUT possesses the bigger correlation of 0.65 while IRL has a median correlation to LUX in terms of absolute values.

Heatmap Correlation between countries and total of people woking in construction

A picture containing text, screenshot, square, number

Description automatically generated

However, for the applying the models the non-pivoted data to model along the entire dataset as sample shows in fig 18 where the third column with ‘Construction(thnd\_per)’ was selected as independent (Y) variable for applying regression models while ‘Country’ and ‘Year” are the dependent (X) variables for splitting the data for testing and training.

Figure dataframe for ML modeling

dataframe selected for modelling


Once the data is split accordingly into dependent and independent, dummy variables had to be created to encode the ‘Country’ labels for each country code (AUT, IRL, etc) and then concatenating both datasets to scale the dataframe as a numpy array and storing them into a new dataframe and splitting them using 70% for training and 30% for testing with a random state of 100 (Numpy, 2023). The correlation between variables was checked once again to prevent any overfitting caused by a multi-correlated features which in this case are considered not to be overfitted.

* 1. – Model Building and Evaluation

**Ridge Regression :**

For the Ridge Regression a GridSeacrchCV was implemented to increase the accuracy of the model by tuning the parameters over 5 levels of cross validation using a negative mean absolute error as scoring for the regressor where the train/test score displayed on fig 19 are significantly close across the 15 alphas selected as parameters (Scikit Learn, 2023). Using an alpha of 200, the Ridge coefficient is stated on the following array:

array([ 0.13289318, 8.17034341, 5.87119464, -9.89167585,

-16.89881758, 12.74895537])

Figure Negative Mean Absolute Error and alpha for Ridge Regression

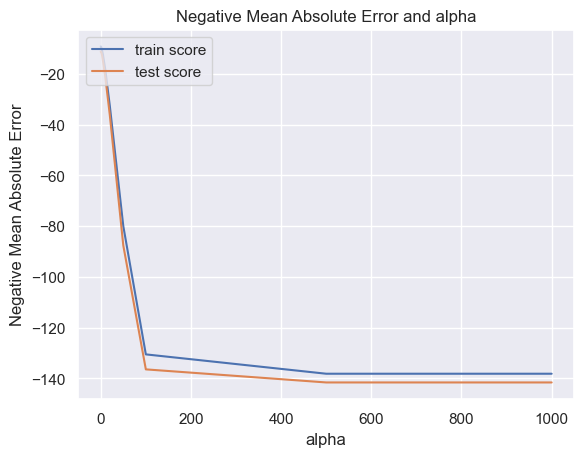
A picture containing text, screenshot, line, plot

Description automatically generated

**Lasso Regression:**

For the Lasso Regression, similar steps were followed, using the cross validation method to apply the model over 5 distinct folds of testing and training the data to obtain better results, comparing the alphas with mean test score and mean train score considering the Negative Mean Absolute Error as an estimator as displayed in figure 20:

Figure Negative Mean Absolute Error and alpha for Lasso Regression



With an alpha = 10, the Lasso model was fit into train and test data after it was cross validated and the coefficient array resulted in:

array([ 0. , 3.91577449, 0. , -75.79225065,

-114.80190022, 26.78865466])

However, when the alpha was changed (as it was tested with values 50, 100, 200), the coefficient also changed significantly, for example, an tunning a Lasso (alpha=0.01) resulted in different coefficients: array([16.6782117 , 9.79659484, -1.26083238, 0. ])

* 1. – Data Preparation for Regression Models u sing *'merged\_df'*

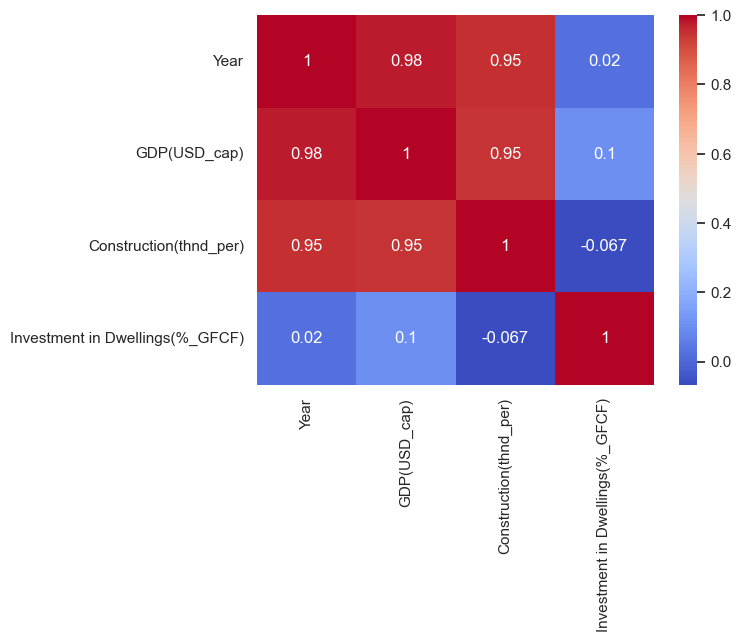
Reassessing the dataframe merged with the three variables analysed (employment in construction, GDP per capita and investment in Dwellings) for the five same five countries selected, the second part of the modelling was performed considering only the Ireland (IRL) category and created a new df containg these values, as figure 21 demonstrates, followed by the display of correlation between their columns (heatmap 3).

Figure Ireland Values df

IRL data

Description automatically generated with low confidence

Heatmap Correlation between IRL values



In this dataset it was also required to encode ‘IRL’ values to obtain only numeric features before applying the model, in using dummy variables where IRL=1 and added to original dataframe as a substitute column.

# Results for Machine Learning Models

Cross validation was implemented to improve performance of the models and provide authenticity for their, or a less biased outcome. Methods for reducing dimensionality of data were not applied since the datasets used were reasonably small, once the cleaning techniques were applied, by dropping features not relevant for the analysis and filtering specific categories to apply the models .

A comparative of all models applies using a root mean squared error scores is stated on table 5, including cross validation scores:

Table Machine Learning Models - Comparison of Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **ML Model** | **Results** | |
| **top\_five\_employ** | **Ridge Regressor** | Cross-Validation R-squared scores: | [-1.16727071e+02 -1.40414173e+02 -3.50282841e+01 -4.55848880e+04 -2.87678634e+02] |
| Mean Cross-Validation R-squared: | -9232.947234896355 |
| **Lasso Regressor** | Cross-Validation R-squared scores: | [-9.69461308e+01 -1.26770594e+02 -3.98523304e+01 -5.17313857e+04 -6.40345881e+02] |
| Mean Cross-Validation R-squared: | -10527.060120483513 |
| **Linear Regression** | R-squared of the model in training set | 0.991734815 |
| R-squared of the model in test set | 0.986218461 |
| Root mean squared error of the prediction | 17.63681065 |
| Mean absolute percentage error of the prediction | 7.508346239 |
| **merged\_df ['IRL']** | **Ridge Regressor** | Ridge Coef with alpha= 10 | 7.68604245, 7.792089 , -1.1795494 , 0. |
| Ridge Regression MSE | 101.6357898 |
| **Lasso Regressor** | Lasso coef with alpha= 2.5 | 14.40670374, 8.86914461, -0.17010673, 0. |
| Lasso Regression MSE | 103.5952937 |
| **Linear Regression** | R-squared of the model in training set | **0.978759249** |
| R-squared of the model in test set | 0.795690819 |
| Root mean squared error of the prediction | **10.25771518** |
| Mean absolute percentage error of the prediction | 6.110056351 |
| **Decision**  **Tree**  **Regressor** | R-squared of the model in training set | 1 |
| R-squared of the model in test set | 0.35907624 |
| Root mean squared error of the prediction | 19.23491292 |
| Mean absolute percentage error of the prediction | **12.56160558** |
| **Decision Tree Regressor with Hyperparameter Tuning** | R-squared | 0.35907624 |
| Mean Squared Error | 369.981875 |
| Root Mean Squared Error | 19.23491292 |
| Mean Absolute Error | 16.34166667 |
| **Random Forest** | R-squared of the model in training set | 0.947088387 |
| R-squared of the model in test set | **0.916486459** |
| Root mean squared error of the prediction | 6.943292918 |
| Mean absolute percentage error of the prediction | **6.403872979** |

For the supervised models applied in the second dataset, it also possible to predict the values for total amount of people employed in construction with a 5 year range, from 2023 to 2027 based on the original data combined with the Linear Regression results, as is illustrated in figure 22, where it was concluded a pattern of growth for Ireland’s scenario as an increasing demand for skilled professionals. The prediction shows a maximum of 200.000 people will be working in the sector, aligned with ECSO’s expectations mentioned in this report. This prediction is considered to have nearly 80% of accuracy considering the r-square of the model in the training set, which is quite low with a 10.25 percentage of error (Statology, 2021).

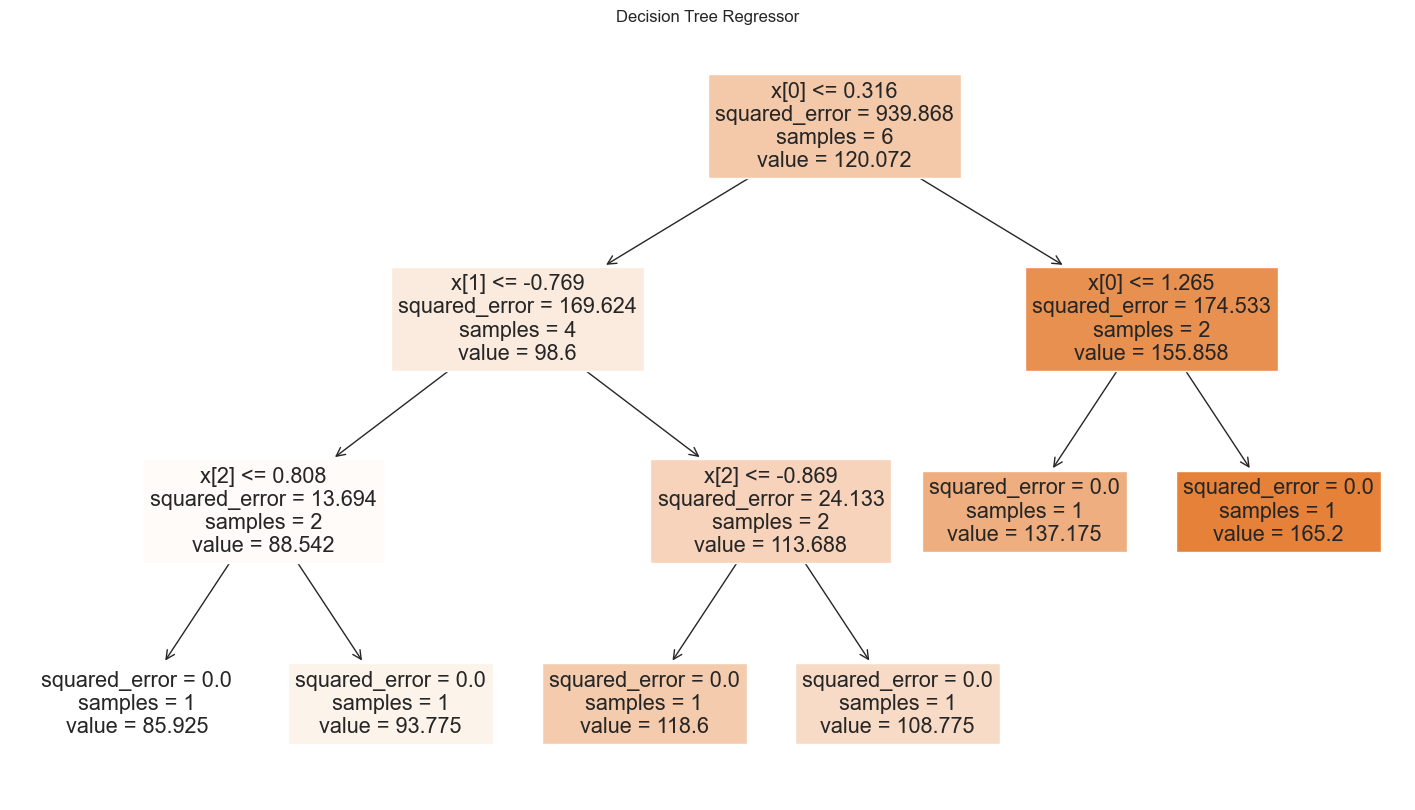
Figure Predictions from ML Model-Linear Regression

A picture containing text, plot, diagram, line

Description automatically generated

In addition to Lasso, Ridge and Linear regressor, for the merged\_df sorted with IRL values it was also applied a Decision Tree Regressor and Random Forrest Regressor using Sklearn’s package data pre-processing to standardise scales of distinct variables before separating them into testing (30%) and training (70%). While if analysed in more detail, the best results were obtained with the Random Forest where the mean percentage of error of 6.40 and root squared test above 90%. The graphical demonstration of the random forest with 4 levels until a 0.0 squared\_error was tested is present in figure 23 (Müller & Guido, 2017) (Scikit-learn , 2023).

Figure Random Forest Regressor - Tree



As for the Decision Tree model, the results obtained (fig 24) are as follow, and after including a hyperparameter tuning combined with GridSearchCV, it was concluded the best maximum depth for the tree is 3 levels for the using the GDP as a starting point.

Figure Decision Tree for merged\_df modelling scores

A picture containing text, screenshot, font, line

Description automatically generated

# Sentimental Analysis: “Is there really a shortage of skilled tradesmen?”

Using Reddit API it was possible to apply machine learning models to perform sentimental analysis with Natural Language Processing (NLP) models such as bag of words (BW) and feature generation (FG) following class tutorials given on the subject. The post (fig25) chosen brings comments on the lack of skilled labour in the Irish Construction sector, contained 13 main comments but the subcomments were not possible to retrieve, so the models were applied to a significant small sample which may have affected its accuracy (50% for BW and 75% for FG).

Once the credentials for API use were loaded, the comments were storage on a ‘json’ file for text manipulation (‘reddit\_response.json’), which can facilitate to extract only the data desired for the modelling, in this case to use only the direct comments, without user names or other information from reddit posts. After a few attempts to understand how to use this structure, data cleaning techniques were performed with the text through tokenizing, removing stopwords, and removing words with similar meaning which were the classified between ‘negative’, ‘positive’ or ‘neutral’ levels of sentiment and once Count Vectorizer method was applied to variables and the data was split into training and testing for modelling as the results are displayed on table 6 containing the classification reports. The overall sentiment based on the comments as stated below, is neutral with 9 comments out of 13 classified with neutrality of sentiment and 4 classified as negative sentiment after the model determined the sentiment based on the count of positive and negative words.

Name: Sentiment, dtype: object

0 Neutral

1 Negative

2 Neutral

3 Neutral

4 Neutral

5 Negative

6 Neutral

7 Negative

8 Neutral

9 Neutral

10 Negative

11 Neutral

12 Neutral

Figure ‘r/Ireland’ Reddit community: Post used for sentimental analysis

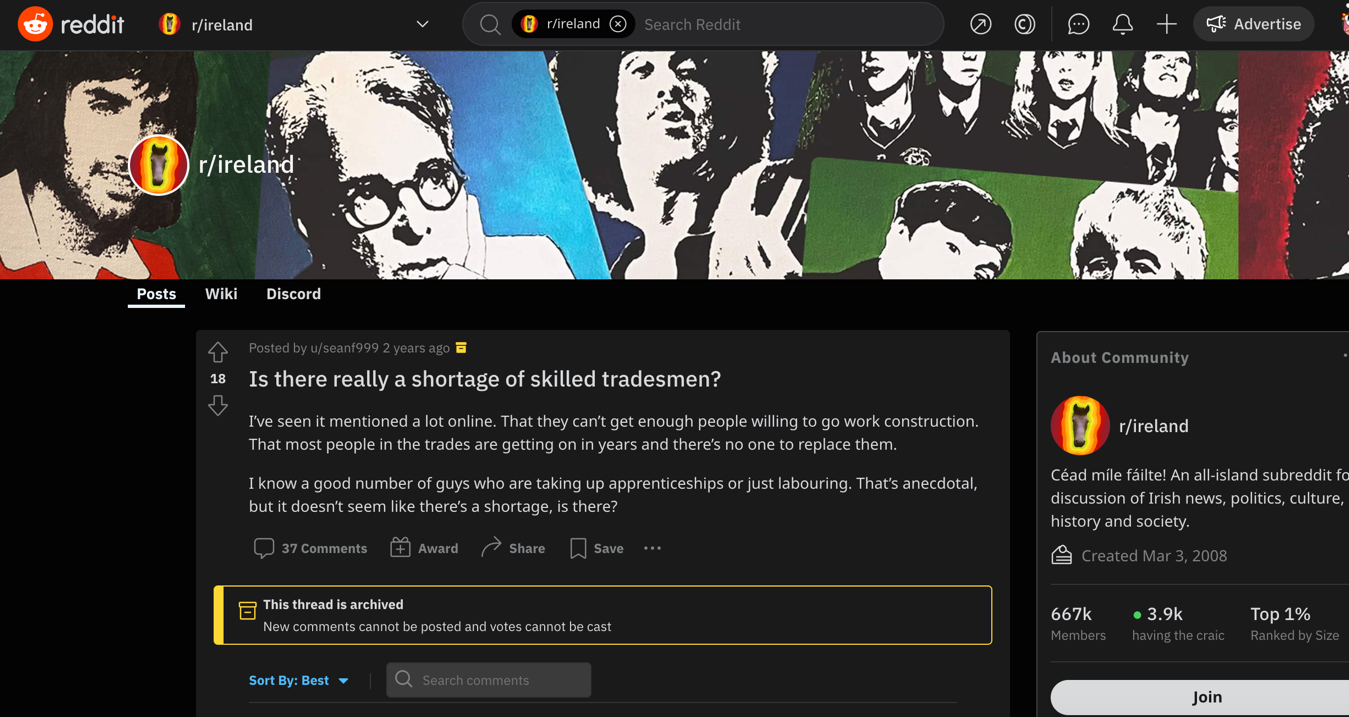


Table Classification Matrix for Sentimental Analysis



# Conclusion

The Construction sector is an important factor for a country’s socioeconomic development and for that reason is of critical responsibility for governments and enterprises to combine efforts and maintain growth levels progressing as a drive for recovery after a period of downturn caused by the global pandemic of 2020. Within this report, comparing nations GDP per capita and their employment level in the sector with investment in Dwellings included as an additional variable to be compared. In the specific for Ireland, with the second highest GDP over the period analysed has the lowest percentage of investment in houses and the second to last in terms of population working in Construction, which is also a reflection of reduction of planning permissions granted in recent years and the increasing demand for new housing option for the general population.

In fact, the shortage of skilled professionals in the area has exponentially increased staff costs up to an unsustainable levels for companies to keep up with, causing effects across even infrastructure developments as well as the housing market. Initially it was also considered to include absolute population for each of the five countries to do an in-depth analysis on the relationship between values to perform additional hypothesis testing and even create better predictions in the Machine Learning section, however, due to time limitations to collect, process and undersand the additional data, this logic was left to be included in future research and reassessment of the models as part of the deployment phase of the project.

Future research could investigate in a more deeper level the employment by specific occupations within construction, to better understand the demands in labour for the sector and compare with index of employment throughout the years. The CSO only shows index of employment in construction from 1975 to 2008, but since the focus of the report is to display the changes from the past decade this information was left out of the research, and can be added to the report on further reviews.

# References

Aandahl, E. M., 2022. *Parametric versus nonparametric tests.* [Online]   
Available at: https://ledidi.com/academy/parametric-versus-nonparametric-tests  
[Accessed 20 May 2023].

Bray, D., 2019. *Sonalake: An introduction to hypothesis testing.* [Online]   
Available at: https://sonalake.com/latest/an-introduction-to-hypothesis-testing/  
[Accessed 20 May 2023].

Burke-Kennedy, E., 2023. *Irish TimesNumber of planning permissions for new homes drops sharply.* [Online]   
Available at: https://www.irishtimes.com/business/economy/2023/03/10/number-of-planning-permissions-for-new-homes-drops-sharply/  
[Accessed 10 May 2023].

Chen, C.-h., Härdle, W. & Unwin, A., 2008. *Handbook of Data Visualization.* s.l.:Springer.

Chris, K., 2022. *freeCodeCamp: Python Functions – How to Define and Call a Function.* [Online]   
Available at: https://www.freecodecamp.org/news/python-functions-define-and-call-a-function/  
[Accessed 2 April 2023].

Citizens Information , 2023. *Citizens Information - Planning Permissions.* [Online]   
Available at: https://www.citizensinformation.ie/en/housing/planning-permission/planning-permission-general/  
[Accessed 02 May 2023].

Creative Commons, 2023. *Attribution 4.0 International (CC BY 4.0).* [Online]   
Available at: https://creativecommons.org/licenses/by/4.0/  
[Accessed 2023 April 2023].

CSO, 2022. *Planning Permissions Quarter 3 2022.* [Online]   
Available at: https://www.cso.ie/en/releasesandpublications/ep/p-pp/planningpermissionsquarter32022/data/  
[Accessed 19 May 2023].

Curran, I., 2022. *Irish Times: ‘Large decline’ in planning permissions for new homes amid costs squeeze.* [Online]   
Available at: https://www.irishtimes.com/business/2022/12/09/large-decline-in-planning-permissions-for-new-homes-amid-costs-squeeze/  
[Accessed 10 May 2023].

Dash , 2023. *Dash - Plotly: A Minimal Dash App.* [Online]   
Available at: https://dash.plotly.com/minimal-app  
[Accessed 25 May 2023].

ECSO, 2022. *European Construction Sector Observatory - Ireland - country fact sheet,* s.l.: European Construction Sector Observatory.

Feng, C., Wang, H., Lu, N. & Tian Chen, H. H. Y. L. X. M. T., 2014. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry.*, 26 Apr, p. 105–109..

Fox, H., 2022. *63% OF IRISH CONSTRUCTION COMPANIES STRUGGLE TO FIND SKILLED WORKERS – NEW RESEARCH.* [Online]   
Available at: https://irishtechnews.ie/63-of-irish-construction-companies-struggle-to-find/  
[Accessed 20 May 2023].

Frost, J., 2017. *Nonparametric Tests vs. Parametric Tests.* [Online]   
Available at: https://statisticsbyjim.com/hypothesis-testing/nonparametric-parametric-tests/  
[Accessed 24 May 2023].

Frost, J., 2022. *Skewed Distribution: Definition & Examples.* [Online]   
Available at: https://statisticsbyjim.com/basica/skewed-distribution/  
[Accessed 12 April 2023].

Galarnyk, M., 2022. *Built in - Understanding Boxplots.* [Online]   
Available at: https://builtin.com/data-science/boxplot  
[Accessed 12 April 2023].

Gupta, A., 2023. *Mean Squared Error : Overview, Examples, Concepts and More.* [Online]   
Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error#what\_is\_mean\_squared\_error  
[Accessed 14 April 2023].

Hotz, N., 2023. *Data Science Process Alliance - What is CRISP DM?.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-2/  
[Accessed 12 April 2023].

IBM, 2023. *Why linear regression is important.* [Online]   
Available at: https://www.ibm.com/topics/linear-regression#:~:text=Linear%2Dregression%20models%20have%20become,can%20be%20trained%20very%20quickly.  
[Accessed 11 April 2023].

KPMG, 2022. *Project Emerald - Report on the Drivers of Cost and Availability of Finance for Residential Development,* s.l.: KPMG.

Luna, Z., 2021. *Understanding CRISP-DM and its importance in Data Science projects.* [Online]   
Available at: https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b  
[Accessed 12 April 2023].

LZP, J., 2022. *Geek Culture - Simple Way To Extract Reddit Comments In Python.* [Online]   
Available at: https://medium.com/geekculture/simple-way-to-extract-reddit-comments-in-python-c8cb2afe2fce  
[Accessed 25 May 2023].

Müller, A. C. & Guido, S., 2017. *Introduction to Machine Learning with Python.* First Edition ed. s.l.:O’Reilly Media, Inc.,.

Matplotlib, 2023. [Online]   
Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.boxplot.html  
[Accessed 10 April 2023].

Matplotlib, 2023. *Matplotlib - Colors.* [Online]   
Available at: https://matplotlib.org/stable/tutorials/colors/colormaps.html  
[Accessed 12 April 2023].

McDonald, A., 2022. *Seaborn Pairplot: Enhance Your Data Understanding with a Single Plot.* [Online]   
Available at: https://towardsdatascience.com/seaborn-pairplot-enhance-your-data-understanding-with-a-single-plot-bf2f44524b22  
[Accessed 13 April 2023].

MplCursors, 2023. *mplcursors 0.5.2 documentation.* [Online]   
Available at: https://mplcursors.readthedocs.io/en/stable/  
[Accessed 17 May 2023].

Mulani, S., 2022. *Digital Ocean - // Tutorial // Using StandardScaler() Function to Standardize Python Data.* [Online]   
Available at: https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python  
[Accessed 21 MAy 2023].

Muth, L. C., 2017. *How to choose a color palette for choropleth maps.* [Online]   
Available at: https://blog.datawrapper.de/how-to-choose-a-color-palette-for-choropleth-maps/  
[Accessed 20 May 2023].

Muth, L. C., 2018. *What to consider when choosing colors for data visualization.* [Online]   
Available at: https://blog.datawrapper.de/colors/  
[Accessed 20 MAy 2023].

Numpy, 2023. *Numpy Documentation.* [Online]   
Available at: https://numpy.org/doc/stable/user/index.html#user  
[Accessed 10 April 2023].

OECD, 2018. *Terms and Conditions.* [Online]   
Available at: https://www.oecd.org/termsandconditions/  
[Accessed 17 May 2023].

OECD, 2023. *Investment by asset.* [Online]   
Available at: https://data.oecd.org/gdp/investment-by-asset.htm#indicator-chart  
[Accessed 10 May 2023].

OECD, 2023. *OECD - Gross domestic product (GDP) (indicator).* [Online]   
Available at: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm#indicator-chart  
[Accessed 10 May 2023].

Panchotia, R., 2020. *Predictive Modelling Using Linear Regression.* [Online]   
Available at: https://medium.com/swlh/predictive-modelling-using-linear-regression-e0e399dc4745  
[Accessed 11 April 2023].

Pandas, 2023. *Pandas Documentation: Essential basic functionality.* [Online]   
Available at: https://pandas.pydata.org/docs/user\_guide/basics.html  
[Accessed 10 April 2023].

Patil, P., 2018. *What is Exploratory Data Analysis?.* [Online]   
Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15  
[Accessed 20 May 2023].

Pete Chapman, J. C. R. K. T. K. T. R. C. S. R. W., 2000. *CRISP-DM 1.0.* [Online]   
Available at: https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf  
[Accessed 14 April 2023].

Plotly, 2023. *Built-in Continuous Color Scales in Python.* [Online]   
Available at: https://plotly.com/python/builtin-colorscales/  
[Accessed 20 May 2023].

Plotly, 2023. *Choropleth Maps.* [Online]   
Available at: https://plotly.com/python/choropleth-maps/  
[Accessed 12 May 2023].

Plotly, 2023. *Line Charts in Python.* [Online]   
Available at: https://plotly.com/python/line-charts/  
[Accessed 19 May 2023].

Python , n.d. *Python Documentation - Built-in Types.* [Online]   
Available at: https://docs.python.org/3/library/stdtypes.html  
[Accessed 9 April 2023].

Radečić, D., 2020. *Top 3 Methods for Handling Skewed Data.* [Online]   
Available at: https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45  
[Accessed 20 May 2023].

Saltz, J., 2022. *Data Science Process Aliance - CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-still-most-popular/  
[Accessed 08 April 2023].

Schmitt, M., 2023. *Streamlit vs. Dash vs. Shiny vs. Voila vs. Flask vs. Jupyter - Comparing data dashboarding tools and frameworks.* [Online]   
Available at: https://www.datarevenue.com/en-blog/data-dashboarding-streamlit-vs-dash-vs-shiny-vs-voila#:~:text=Dash%20and%20Panel%20are%20both,based%20only%20on%20Python%20syntax.  
[Accessed 25 May 2023].

Scikit Learn, 2023. *Scikit Learn - Sklearn Linear Regression.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html  
[Accessed 11 April 2023].

Scikit-learn , 2023. *sklearn.ensemble.RandomForestRegressor.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn-ensemble-randomforestregressor  
[Accessed 13 April 2023].

SciPy , 2023. *scipy.stats.shapiro.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html  
[Accessed 14 April 2023].

Seaborn, 2022. *Overview of seaborn plotting functions.* [Online]   
Available at: https://seaborn.pydata.org/tutorial/function\_overview.html  
[Accessed 02 April 2023].

Sklearn, 2023. *sklearn.linear\_model.LinearRegression.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html  
[Accessed 12 April 2023].

Stack Overflow, 2015. *Stack Overflow - Check if value is zero or not null in python.* [Online]   
Available at: https://stackoverflow.com/questions/28210060/check-if-value-is-zero-or-not-null-in-python  
[Accessed 11 April 2023].

Statology, 2021. *How to Make Predictions with Linear Regression - Zach.* [Online]   
Available at: https://www.statology.org/predictions-regression/  
[Accessed 11 April 2023].

Summerfield, M., 2010. *Programming in Python 3: A Complete Introduction to the Python Language.* Second Edition ed. s.l.:Pearson Education, Inc.

The Data Frog, 2021. *Choropleth Maps in Python (2021).* [Online]   
Available at: https://thedatafrog.com/en/articles/choropleth-maps-python/  
[Accessed 18 May 2023].

Trevisan, V., 2022. *Towards Data Science - Comparing sample distributions with the Kolmogorov-Smirnov (KS) test.* [Online]   
Available at: https://towardsdatascience.com/comparing-sample-distributions-with-the-kolmogorov-smirnov-ks-test-a2292ad6fee5  
[Accessed 21 MAy 2023].

Tufte, E. R., 2001. *The Visual Display of Quantitative Information.,.* Second ed. Cheshire: Graphics Press.

UNSC, 2023. *System of National Accounts 2008 - 2008 SNA.* [Online]   
Available at: https://unstats.un.org/unsd/nationalaccount/sna2008.asp  
[Accessed 21 May 2023].

Víquez, S., 2023. *91% of ML Models Degrade in Time by.* [Online]   
Available at: https://www.nannyml.com/blog/91-of-ml-perfomance-degrade-in-time  
[Accessed 14 April 2023].

Vu, T., 2022. *How to Create a Beautiful Python Visualization Dashboard With Panel/Hvplot.* [Online]   
Available at: https://www.youtube.com/watch?v=uhxiXOTKzfs  
[Accessed 23 May 2023].

Yi, M., 2019. *Nightingale - How to Choose Colors for Your Data Visualizations.* [Online]   
Available at: https://nightingaledvs.com/how-to-choose-colors-for-your-data-visualizations/  
[Accessed 20 May 2023].

Yi, M., 2021. *A Complete Guide to Line Charts.* [Online]   
Available at: https://chartio.com/learn/charts/line-chart-complete-guide/  
[Accessed 10 April 2023].

Zach, 2018. *Statology - Here is How to Interpret a P-Value of 0.000.* [Online]   
Available at: https://www.statology.org/here-is-how-to-interpret-a-p-value-of-0-000/  
[Accessed 20 May 2023].

Zach, 2023. *Statology - Pandas: How to Use describe() and Suppress Scientific Notation.* [Online]   
Available at: https://www.statology.org/pandas-describe-no-scientific-notation/  
[Accessed 9 April 2023].

# Appendices

## Appendix 1: GitHub

GitHub repository for version control:

<https://github.com/clarissa2020274/CA2_version_control.git>

## Appendix 2: Project Management applying CRISP-DM (Luna, 2021)

Figure Diagram of CRISP DM applied for the project created with draw.io based on Hotz diagram (Hotz, 2023)

A picture containing text, screenshot, font, design

Description automatically generated

**Business Understanding and definition of goals:** the goal of ML models will be to investigate the growing demand for skilled labour in Construction in the coming years as the construction sector in Ireland in special shows signs of consistent increase, despite the inflation rates as pointed by the ECSO report. The shortage of skilled labour is one of the factors in the increase cost of production of new buildings, along with high cost of materials and supply chain that also reflects on the high rate of late conclusion of projects and even the reduction of projects being granted planning permission to be initialised, as it was elaborated on the first part of the report. Also, the period delimited for project from 2012 to 2022 was to investigate how the countries started to recover after the financial crisis of 2007/2008 and had to face another downturn with the effects of the global pandemic and now attempt to maintain financial growth to a pre-pandemic level.

**Data understanding:** a series of distinct datasets across construction subjects were collected from three main databases: OECD, CSO and Eurostat. Within CSO’s database for Irish data were cost of production in construction in Ireland, index of employment (but only data available was between 2005 and 2007, not part of the period analysed), purchases in construction (to understand demand for raw materials).

Also in this phase, there also the need for understanding the timeframe required to deploy each phase until projects deadline, considering the time to perform research for libraries that could be used for each requirement and analysing different datasets to select the variables that would actually be used for the modelling and performing the necessary cleaning for any missing values, mislabelled observations to verify the quality and perform exploratory tasks to understand data distribution and insightful visualisation to complement and enrich the project.

**Data preparation** – Applied techniques to standardise scales, dropping irrelevant columns, renaming country labels and matching each country with its proportional values for GDP, employment in Construction and Investment in Dwelling to apply chosen supervised Models. Also important phase of the project was to check any missing values or duplicates in each dataset as well as understanding their distribution for hypothesis testing. Merginng data in the same dataframe to better investigate their relationship and pivoted features to investigate how the data could change according to each country and see patterns of behaviour.

**Modelling** – based on the nature of the data in this project, it was decided to apply a range of regressions models to predict the pattern of the employment rates in Ireland, as it has been documented there is a lack of professionals in the sector, so models were applied to infer future values.

**Evaluation** – Based on squared mean errors compared all models results on a table and find better parameters for the tunning necessary.

**Deployment** – Review project and displaying results as this part of thisreport. Consider what other data could be included in further revisions and submit for lecturers.

## Appendix 3: Datasets available in CSO database

Table of Datasets related to CSO Planning Permissions (CSO, 2022) release on Number of dwelling units approved decreased by 41% in Quarter 3 2022:

|  |
| --- |
| **Datasets used for CSO release** |
| BHQ03 Planning Permissions Granted |
| BHQ04 Planning Permissions Granted for Communal Dwellings |
| BHQ06 Planning Permissions Granted for Civil Engineering Projects |
| BHQ07 Planning Permissions Granted for Non Residential Buildings |
| BHQ12 Planning Permissions Granted for New Houses and Apartments |
| BHQ13 Planning Permissions Granted |
| BHQ14 Total Floored Area for which Permission Granted in New Construction and Extensions |
| BHQ15 Planning permissions granted for apartment, multi-development and all house units |

## Appendix 4: Outputs of .info, .shape and .isnull.sum

*#print summarised information from dataframe*

*data.info()*

*print(data.shape)*

*print(data.isnull().sum())*

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 49 entries, 0 to 48

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 49 non-null object

1 Planning Permissions Granted 45 non-null object

2 Units for which Permission Granted 48 non-null object

3 Total Floor Area for which Permission Granted 45 non-null object

4 Average Floor Area per Unit 45 non-null float64

dtypes: float64(1), object(4)

memory usage: 2.0+ KB

(49, 5)

Unnamed: 0 0

Planning Permissions Granted 4

Units for which Permission Granted 1

Total Floor Area for which Permission Granted 4

Average Floor Area per Unit 4

dtype: int64

## Appendix 5: Comparing Seaborn and Plotly libraries for Visualisation.

* Boxplots for each variable using plotly’s dropdown menu:

A picture containing text, screenshot, line, diagram

Description automatically generated

* A screenshot of a diagram

  Description automatically generated with low confidence

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

* Boxplot of all variables before standardising:

**To interpret this boxplot considering all features, its comprehension is debilitated since they were originally measured in different units. For example: Total Units of Permission Granted measured in discrete whole number (count by thousands of units, 1000, 2000, 3000...), while the average area per sq m is in squared meters(105, 203...), hence the nearly invisible boxplot at the corner left representing this feature. So the solution found was to use a standard scaler to display all values on the same scale and facilitate visualisation and understanding since using** different scales of the data may negatively affect the modelling of a dataset resulting in biased predictions and misclassification errors and accuracy rates (Mulani, 2022).

A picture containing text, screenshot, diagram, plot

Description automatically generated

* Boxplot after Standardised Scaler applied:

A picture containing text, diagram, screenshot, line

Description automatically generated

## Appendix 6: Test results for checking Normal Distribution

-Shapiro-Wilk:

|  |  |  |
| --- | --- | --- |
| **Tests of Normality** | | |
| **Features** | **Shapiro-Wilk** | |
| **stat** | ***p-value*** |
| *ireland\_permits(df)* | 0.655 | 4.06E-21 |
| *Planning Permissions Granted* | 0.925 | 0.0065 |
| *Total Units of Permission Granted* | 0.867 | 0.0001 |
| *Permissions Granted (sq m)* | 0.806 | 2.38E-06 |
| *Avg Area per Unit (sq m)* | 0.967 | 0.2268 |
| *Planning Permissions Granted\_log* | 0.967 | 0.226 |
| *Total Units of Permission Granted\_log* | 0.967 | 0.226 |
| *Permissions Granted (sq m)\_log* | 0.967 | 0.226 |
| *Avg Area per Unit (sq m)\_log* | 0.967 | 0.226 |

* Kolmogorov-Smirnov test

|  |  |  |
| --- | --- | --- |
| **Tests of Normality** | | |
| **Features** | **Kolmogorov-Smirnov test** | |
| **stat** | ***p-value*** |
| Planning Permissions Granted | 1.00 | 0.00E+00 |
| Planning Permissions Granted (*xbar,s)* | 0.163 | 0.162 |

1. Access to funding and mortgage rates are also important factors in this context of the housing crisis but are not the focus point in this report at this stage of the research. [↑](#footnote-ref-1)
2. Dataset originally stored under the variable “data” was renamed as “Ireland\_permits” to better code practice and avoid confusion (Summerfield, 2010). [↑](#footnote-ref-2)