**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | CA2 50% Integrated Assessment |
| **Assessment Title:** | Ireland Construction Data Compared to Other Countries |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Dornel Duarte Rotta |
| **Student Number:** | 2023066 |
| **Assessment Due Date:** | Friday, 14 April 2023 |
| **Date of Submission:** | Friday, 14 April 2023 |

****

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Abstract**

*Homelessness is an issue experienced by people who lack a place to live that is supportive, affordable, decent, and secure. (Peter McVerry Trust, 2020).*

*We accessed official data published by the Government of Ireland (Data GOV IE, 2023) and performed an analyze and investigate of these data sets, summarized their main characteristics, applied statistical data, graphical presentation, and modeled through machine learning.*

*Through this study it was possible to identify the strong increase in the number of people in this situation in recent years, with a large majority of cases among adult men between 25 and 44 years old, where Dublin has the highest per capita rate.*

*By crossing homeless and inflation data, we were also able to verify a strong correlation between the beginning of the cost-of-living crisis after the covid pandemic and the growth of the homeless population.*

*With the addition of machine learning, we used the linear regression algorithm to predict how the situation will be if the crisis scenario does not change.*

**Introduction**

The purpose of this study was to analyze construction data from Ireland and compare it with other countries.

For this purpose, Data Analytics concepts were used, such as: Data Business - stage where the business need is evaluated, and the objectives are defined. Data Understanding - phase where the data is analyzed for a complete understanding of how to prepare them, Data Preparation - moment in which the data is selected, structured, and cleaned so that it can be used. Data Visualization - Phase where graphs and dashboards are presented with the desired information. Machine Learning - Step where algorithms are applied to deliver results based on learning, such as predictions, classification, among others.

All this work was prepared using the CRISP-DM project management framework as a basis. Python and its libraries were used as a programming language and tool to carry out the Data Analytics steps.

**Business Understanding**

The Business Understanding phase focuses on understanding the objectives and requirements of the project. This phase consists basically in determine the business objectives, its necessary first “thoroughly understand, from a business perspective, what the customer really wants to accomplish.” (Data Science Process Alliance, 2023).

The business perspective of this study was analyzing Ireland's Construction data and compare the Irish Construction sector with other countries worldwide.

In this analysis it was expected the inclusion of forecasting, sentiment analysis and evidence-based recommendations for the sector as well as a complete rationale of the entire process used to discover your findings.

It is known which Ireland has been suffering with the cost and lack of house in the last years, based on this topic, it was decided to carry on a work comparing Ireland’s data from total house construction growth and the total house construction cost growth along the time, crossing this data within a depth analysis on the employment rate per total population and total active population.

The main idea was to identify correlations between the increase or decrease in the total number of houses built and its cost with the employment rate in the construction sector as well as with the population growth, not only the absolute total population of a country, but the active (16 years old to 65 years old)) total population. Was the increase or decrease in the total active population a key factor to see an increase in the total cost of houses? The increase or decrease in the employment rate has some impact on the total of houses built?

The countries which were chosen to be compared with Ireland was Italy and Germany. Italy has an older population compared to Ireland and this can bring some interesting comparations about the house data and available work force, on the other hand, Germany has a more productive and active production sector delivering a very good range of different realities to be compared.

**Data Understanding**

This phase of CRISP methodology drives the focus to identify, collect, and analyze the datasets that can help to accomplish the project goals. (Data Science Process Alliance, 2023).

In this step it was collected initial data to be analyzed, this data was described, it means, it was examined and documented its properties like data type, data format, number of records, field identities, and others. Some sort of exploration was done to have a deeper look into the data, printing some recordings and analyzing them, and get some initial answers: How structured was our data available? How clean/dirty was it? Was it necessary to prepare/transform this data? (Data Science Process Alliance, 2023).

## Collect Initial Data

As already mentioned, the data defined in the business understanding phase were:

* Total House Cost Growth along time
* Total House Construction Growth along time
* Construction Employment Rate along time
* Total Population along time
* Total active population along time

In this step it was collected and printed a first look on the chosen datasets:

**Total House Construction Cost**

**Ireland**

A picture containing text, screenshot, font, number

Description automatically generated

|  |  |
| --- | --- |
| **Germany** | **Italy** |
| A screenshot of a computer  Description automatically generated with low confidence | A picture containing text, screenshot, font, number  Description automatically generated |

After collecting and carrying out a first superficial analysis of the datasets with data on construction house cost, it was already possible to have some first understandings:

- The datasets have different amounts of features (columns).

- Datasets are based on percentage growth with a base year index (info captured on the original source).

- This base year is different in each dataset.

- The start and end dates are different.

- Ireland has annual data, Germany quarterly and Italy monthly data.

**Total House Construction Total**

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
| A picture containing text, screenshot, font, number  Description automatically generated | A screenshot of a computer  Description automatically generated with low confidence |  |

After collecting and carrying out a first superficial analysis of the datasets with data on total construction house, it was already possible to have some first understandings:

- The datasets have the same number of features (columns).

- Datasets are based on percentage growth with a base year index (info captured on the original source).

- This base year is different in each dataset.

- The start and end dates are different.

- Ireland has quarterly data, Germany monthly data and Italy quarterly data.

**Construction Employment**

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

After collecting and carrying out a first superficial analysis of the datasets with data on construction employment, it was already possible to have some first understandings:

- The datasets have the same number of features (columns).

- The start dates are different.

- Both datasets have quarterly data.

**Total Population**

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

After collecting and carrying out a first superficial analysis of the datasets with data on total population, it was already possible to have some first understandings:

- The datasets have the same number of features (columns).

- The start and end dates are the same.

- Both datasets have annual data.

**Total Active Population**

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

After collecting and carrying out a first superficial analysis of the datasets with data on total active population, it was already possible to have some first understandings:

- The datasets have the same number of features (columns).

- The start dates are different.

- Both datasets have quarterly data.

## Describe Data

After collecting and taking a first look at the data, in this phase the data was examined and its properties like data format, data type, and number of records were documented. (Data Science Process Alliance, 2023).

**Total House Construction Cost**

It was executed other commands to have a better understanding and to be possible to classify the dataset.

Describe: This python function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

Info: This method prints useful information which contains the number of columns, column labels, column data types, number of cells in each column (non-null values).

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

With these printed values it was possible to document the features of the dataset:

**Ireland**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Statistic | Categorical | N/A | Object | Non-null |
| STATISTIC Label | Categorical | N/A | Object | Non-null |
| TLIST(A1) | Quantitative | Discrete | Int64 | Non-null |
| Year | Quantitative | Discrete | Int64 | Non-null |
| C02196V02652 | N/A | N/A | Object | Non-null |
| State | Categorical | N/A | Object | Non-null |
| Unit | Categorical | N/A | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

**Germany and Italy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Date | Quantitative | Discrete | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

**Total House Construction**

It was executed other commands to have a better understanding and to be possible to classify the dataset.

Describe: This python function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

Info: This method prints useful information which contains the number of columns, column labels, column data types, number of cells in each column (non-null values).

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

With these printed values it was possible to document the features of the dataset:

**Ireland, Germany, and Italy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Date | Quantitative | Discrete | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

**Construction Employment**

It was executed other commands to have a better understanding and to be possible to classify the dataset.

Describe: This python function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

Info: This method prints useful information which contains the number of columns, column labels, column data types, number of cells in each column (non-null values).

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

With these printed values it was possible to document the features of the dataset:

**Ireland, Germany, and Italy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Date | Quantitative | Discrete | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

**Total Population**

It was executed other commands to have a better understanding and to be possible to classify the dataset.

Describe: This python function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

Info: This method prints useful information which contains the number of columns, column labels, column data types, number of cells in each column (non-null values).

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

With these printed values it was possible to document the features of the dataset:

**Ireland, Germany, and Italy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Date | Quantitative | Discrete | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

**Total Active Population**

It was executed other commands to have a better understanding and to be possible to classify the dataset.

Describe: This python function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

Info: This method prints useful information which contains the number of columns, column labels, column data types, number of cells in each column (non-null values).

|  |  |  |
| --- | --- | --- |
| **Ireland** | **Germany** | **Italy** |
|  |  |  |

With these printed values it was possible to document the features of the dataset:

**Ireland, Germany, and Italy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Type | Data Type | Null value |
| Date | Quantitative | Discrete | Object | Non-null |
| Value | Quantitative | Discrete | Float64 | Non-null |

## Explore Data

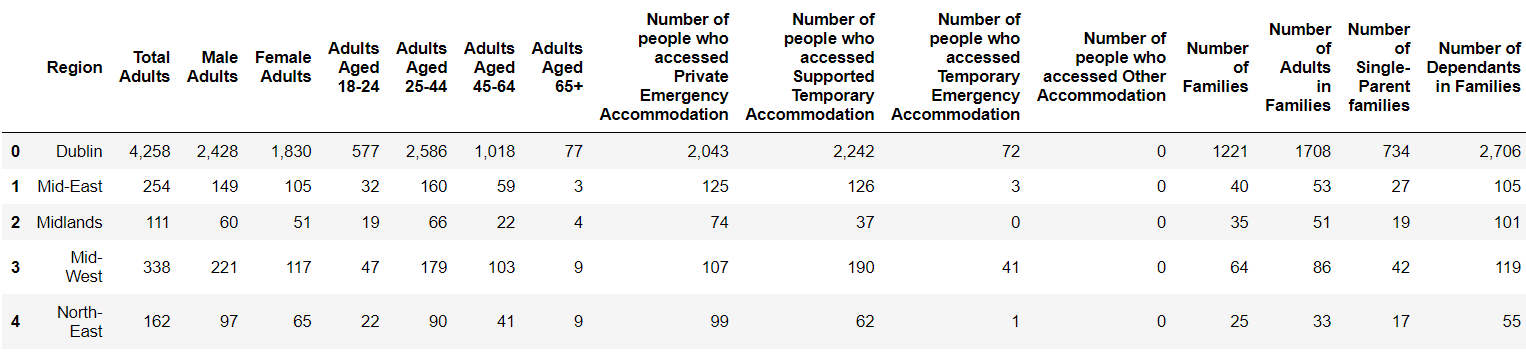
After data was examined and its properties documented, in this phase it was done a dig deeper into the data. Query it, visualize it, and identify relationships among the data. (Data Science Process Alliance, 2023).

## Verify Data Quality

After data depth explored, in this last phase of the Data Understanding, it was the moment to summarize the data quality, what is needed to be cleaned? What needs to be transformed? (Data Science Process Alliance, 2023).

Head() python function prints the first 5 observations (rows) of our dataset and shows all the features (columns). Head() function is important because we can start to classify our data, understanding if our features are discrete or continuous, categorical, or quantitative.

Below we have printed the first 5 observations of our dataset. this dataset contains data from January 2019:



Taking a first look into the dataset, it was possible to notice an interesting range of sixteen features where fifteen are quantitative and only one is categorical.

All quantitative variables are totals, that is, they are the total number of homeless people represented per month. We do not have data representing each of these people. This means that, despite numerical and continuous values, we will have difficulties in making some kind of classification or clustering based on machine learning. On the other hand, we have a massive and complete total of 16 features for linear regression machine learning testing.

|  |  |  |
| --- | --- | --- |
| Feature | Type | Type |
| Region | Categorical | N/A |
| Total Adults | Quantitative | Discrete |
| Male Adults | Quantitative | Discrete |
| Female Adults | Quantitative | Discrete |
| Adults Aged 18-24 | Quantitative | Discrete |
| Adults Aged 25-44 | Quantitative | Discrete |
| Adults Aged 45-64 | Quantitative | Discrete |
| Adults Aged 65+ | Quantitative | Discrete |
| Number of people who accessed Private Emergency Accommodation | Quantitative | Discrete |
| Number of people who accessed Supported Temporary Accommodation | Quantitative | Discrete |
| Number of people who accessed Temporary Emergency Accommodation | Quantitative | Discrete |
| Number of people who accessed Other Accommodation | Quantitative | Discrete |
| Number of Families | Quantitative | Discrete |
| Number of Adults in Families | Quantitative | Discrete |
| Number of Single-Parent families | Quantitative | Discrete |
| Number of Dependants in Families | Quantitative | Discrete |

## Function: Describe()

Describe() function calculates a range of statistical values from our numerical dataset columns. It gives the mean, standard and interquartile values.

Graphical user interface

Description automatically generated with medium confidence

The main result of the Describe() function was that not all numeric values were calculated, we have fifteen numeric features, but only seven were shown. This can happen for several reasons, we can have blank values, we can have string values mixed with numerical values, we can have values with different formatting, among other possibilities.

Analyzing deeply into our dataset, we were able to find some issues:

Shape

Description automatically generated with low confidence A picture containing shape

Description automatically generated

As we can see, we have different formatation depending of the feature.

Shape

Description automatically generated with medium confidence

These thousand commas can induce python to interpret this value as a string instead of a numeric.

Fortunately, we do not have blank or null values on our dataset.

## Command: Shape

This python command returns the dimensions of the dataset. This is important to understand if we have enough data to do the work.

Shape: (9, 16)

Although we have a more than satisfactory number of features (16), our total of observations is insufficient for a quality analysis (nine lines only).

This dataset only represents one month of data (January 2019). Therefore, it is necessary to evaluate the import of more datasets.

Knowing that we have available datasets up to December 2022, this December 2022 dataset was analyzed to verify compatibility and, consequently, the feasibility of using all monthly samples from January 2019 to December 2022, totaling 48 datasets.

Once we already know that our dataset has only nine rows, let’s look on the entire December 2022 dataset:



The first analysis on this dataset is that we have more features, a total of 19 features, three more than the first dataset (January 2019).

Despite having these extra features, all other columns have the same name and format, which demonstrates that we can use all these 48 datasets for a richer analysis.

We can still see differences in the formatting of thousand values, but this is an easily solved problem during the data preparation phase.

**Data Preparation**

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data, then exploring and visualizing it. (Amazon, 2023).

## Data Collection

The first step was to collect all data to be cleaned and labeled. As we noticed during the analysis phase, we have each dataset containing data of only one month, for this reason we have imported 48 datasets and concatenated in just one. Just concatenate was possible because the main structure of our datasets was the same, join or merge wasn’t necessary.

The year and month date values were present only in the CSV file name, so during this data collecting phase, the month and year values were extracted from the file name, as well as a complete datetime value.



New features extracted from file name:

* Year: Only the correspondent year.
* Month: Only the correspondent month.
* Date: A complete date time type value.

Year and month were purposely separated into different features to facilitate the use for graph plotting and machine learning.

## Data Cleaning

The first step of our data cleaning was to remove all features which were added over the months. As we saw during the data analysis phase, four features were added:

* Number of Child Dependants in Families;
* Number of people with citizenship Irish;
* Number of people with citizenship EEA/Uk;
* Number of people with citizenship Non-EEA;

Probably the government has decided to include this data along the years to enrich data available, but for us, this lack of data will distort our results. For example, it’s only possible to find the feature “Number of people with citizenship Non-EEA” on the last seven datasets.

The second step was to decide which features will be more valuable to understand homeless situation in Ireland.

Analyzing the dataset, we have decided to exclude extra features about types of accommodation and family group types. We kept focus on Ireland regions, gender, and ages groups which we have classified as more important on this work.

Removed features:

* Number of people who accessed Private Emergency Accommodation
* Number of people who accessed Supported Temporary Accommodation
* Number of people who accessed Temporary Emergency
* Number of people who accessed Other Accommodation
* Number of Families
* Number of Adults in Families
* Number of Single-Parent families
* Number of Dependants in Families

The last data cleaning step was to format the thousand values to have the same pattern and convert values to be represented as a numeric value or datetime value.

For the thousand, we have decided to remove the comma separator. This will avoid having some Python error interpretation.

All the numeric features were converted to Int data type, and the date feature was converted to datetime data type.

After these three data cleaning steps, the resulted data frame was:

A picture containing table

Description automatically generated

Now we have 432 observations representing data for 48 months: January 2019 to December 2022.

**Data Visualization**

Data visualization is the practice of designing and creating easy-to-communicate and easy-to-understand graphics or visual representations of an amount of complex quantitative and qualitative data (Wikipedia, 2023).

On this work we have plotted graphs to have a more deep understand about homeless situation in Ireland based on government data. We have crossed data per region, per gender, per age and per different viewpoints, for example, comparing the homeless situation with cost of life increasing (represented by inflation) and population.

## Total Homeless in Ireland

Chart, line chart

Description automatically generated

Our first graphic plot was to present data about homeless people in Ireland from 2019 to 2022. In the first two years, the situation of homeless has remained relatively stable at 6500 cases between January 2019 and March 2020, after this period, it dropped by around 500 cases, 8% reduction, remaining stable at 6000 cases between June 2020 and July 2021. However, from August 2021, we noticed an almost uninterrupted increase (apart from December 2021) to a peak of more than 8000 cases in December 2022, an increase of almost 40% in the rate.

We have two main factors which can be the reason for the decreasing and increasing patterns during this 4-year period.

1. In March 2020, the covid pandemic started in Ireland, government applied restrictions to go outside as a social distance isolation rules, additionally, extra subsistence allowances were delivered to help people. This can explain the reduction in the number of homeless during this specific time.
2. After pandemic times the world started to face a cost of live economic crisis, which was aggravated by Russian-Ukrainian war. This can explain the 40% increasing in the homeless situation in Ireland.

To have a better view of this cost of live economic crisis, we crossed this homeless data with inflation (Central Statistics Office, 2023):

Chart, line chart

Description automatically generated

Inflation rate is a strong number to analysis the cost-of-living crisis impacts, and as we predicted, we can see a strong correlation between the scaling of prices and the homeless issue.

The escalation in prices and homeless don't exactly start together, Prices started increasing in October 2020 and the homeless started increasing in July 2021, eight months apart. This behavior can be explained by the fact that it takes a period for the crisis to start to have a profound impact on the population.

The same behavior maybe will be possible to see at the end of the graph, where inflation has started to stabilize (5 months), but homeless continues to increase.

It is possible that in another 3 months (totalizing the same 8 months) the homeless situation will start to stabilize, which would “prove” the correlation between inflation and homelessness.

## Total Homeless in Ireland: per groups

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Visualizing data per age and gender, as we could expected, men are more vulnerable to homeless situation, representing almost 65% of total homeless. Homeless in Ireland was associated with adult males because men are more exposed to addiction or alcoholism issues (Wikipedia, 2023).

Analyzed the visualization per age, it was noticed that young adults (25-44 years old) are by far the majority representing almost 55% of the total. The second group, youth adults (18-25 years old) represent 25%, less the half of young adults.

It is interesting to note that elderly people represented less than 1.5% of the total, which could be explained by access to pension incomes.

## Total Homeless in Ireland: per region

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The graphs including and excluding Dublin were intentionally separated because Dublin has almost 10 times more homeless than the second position in this rank (South-West region). On the same plot, we would have difficulties in visualizing the values of the other regions (they would be overlapping).

The Dublin metropolitan area has a much larger population than the others, this makes it obvious that it would be positioned very far from the others in the first place.

To have a visualization that provides a better quality of comparison between the regions, we crossed the data of total homeless with the total population of each region (Government of Ireland, 2023), so we could make a visual analysis per capita of the situation between the regions.

## Total Homeless in Ireland per region: per 1000

**Chart

Description automatically generated**

Chart, line chart

Description automatically generated

If before the Dublin metropolitan region had ten times more homeless than the South-West region in absolute numbers, this time analyzing the incidence of homelessness per 1000 inhabitants, the number dropped to five times more homeless than the new second place region (Mid-West).

Comparing the differences in the homeless rankings excluding Dublin:

|  |  |  |
| --- | --- | --- |
| **Position** | **Absolute Number** | **Per 1000** |
| 1  2  3  4  5  6  7  8 | South-East  Mid-East  Mid-West  West  South-East  Midlands  North-East  North-West | Mid-West  South-West  West  Mid-East  South-East  North-East  Midlands  North-West |

The positions change as how the data is treated, a value per 1000 brings rich and assertive information in the face of the reality of each region, for example, based on numbers per 1000, governments can better allocate their funds to combat the problem.

## Total Homeless: Most recent data in Ireland

The most recent data from homeless situation in Ireland is December 2022.

Recent situation:

Chart

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated A picture containing graphical user interface

Description automatically generated Chart, bar chart

Description automatically generated

Analyzing a smaller range of data, in this case a specific month, we can use bar chart graphs for better distribution and comparison.

In this latest wave of visualizations, we could also compare the differences between homeless crossing age and gender by region. We could see that despite small differences, the highest incidence of men and young adults (25-44 years old) occurs in all regions of Ireland.

**Statistical Analysis**

For the statistical analysis in this study, we used the Poisson distribution algorithm.

A Poisson distribution is a discrete probability distribution, meaning that it gives the probability of a discrete (i.e., countable) outcome. For Poisson Distribution, the discrete outcome is the number of times an event occurs. It’s possible to use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space (Scribbr, 2023).

The events in this case were the probability of homeless elderly people in Dublin.

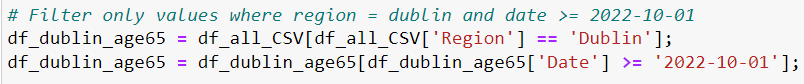
Below are the boundaries for the analysis:

* Seniors (65+ years old)
* Region: Dublin
* Base period: last three months (Oct/22 to Dec/22)

Based on the frequency of elderly people in the last three months, we calculated using poison the probability of the number of elderly people for the month of January/2023

## Poisson: Single Probability

First step was to filter registers to have only observations from Dublin and months from October 2022 to December 2022:



Once we had the appropriate records in our data frame, we could calculate the lambda value. Lambda is the mean number of events occurring within a given interval of time or space, in this case occurrence of adults aged 65+.

We interacted with our dataset (row per row) adding the values and then dividing by the number of records and obtaining the mean (lambda):

Text

Description automatically generated

With the lambda in hands, it was time to define an array of values to be predicted.

For this scenario, we added a range of 20 numbers to each side, that means with a lambda of 88, we got a range from 68 to 108 to calculate the probability, in better words, we calculated the probability to have from 68 to 108 elderly people in homeless situation in January 2023.

Text

Description automatically generated

The last step was to calculate the probability of our range of values, to do this we used the function PMF from Python.

PMF calculates the probability of each value of the array based on the number of events (lambda):

Text

Description automatically generated

Result:

A picture containing text

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Each bar line in the graph represents the probability of the event (elderly people in homeless situation). For example, the probability of having 80 elderly people in a homeless situation in January 2023 is approximately 3%.

## Poisson: Cumulative Probability

Generally, for these cases knowing the probability of just one value does not give us a valuable answer.

For example, knowing the probability of occurrence of terms between 78 and 98 elderly people who are homeless can provide us with more useful information.

For this type of calculation, Python brings us the CDF function.

This function calculates the cumulative probability of an event occurring starting from zero, for example, if we want to calculate the probability using lambda 78, the function will return the cumulative probability from 0 to 78.

Since we want to calculate the cumulative probability between 78 and 98 events, we need:

* First calculate the probability from 0 to 98
* Second calculate the probability from 0 to 78
* Finally, subtract the cumulative odds (0 to 98 - 0 to 78)

Text, letter

Description automatically generated

Text

Description automatically generated

**Modeling and Evaluation - Machine Learning**

Machine learning is a field of inquiry devoted to understanding and building methods that "learn" – that is, methods that leverage data to improve performance on some set of tasks.

Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so (Wikipedia, 2023).

## Selecting Model Techniques

In this study, it was initially defined to analyze the data collected through at least three of the following algorithm options:

* Prediction:
* Classification:
* Clustering:

After the Data Understanding, Preparation, and Visualization phases, it was possible to have a better understanding of the types of features that our dataset provides. In our case, we have a variety of data through features of the discrete type that totalize the number of records monthly for 4 years, the algorithm that has the best fit is the supervised linear regression algorithm for prediction.

Classification algorithms would not have much benefit for this study since all our features are totals. We have the total number of homeless young adult men in Dublin in January 2019, but we do not have features that detail this total, such as values for each of the records, detailing whether the homeless person has problems with alcohol, drugs, mental problems, diseases, etc. If we had this data, we could use a decision tree algorithm to classify whether a person would be homeless if these features indicated it.

The same for clustering algorithms, it doesn't make sense to cluster our dataset, separating it into clusters of totals will not help us to have answers on top of our data.

To maintain the value of this study even without these last two algorithms, four options were explored for the algorithm that best suits our dataset, Linear Regression.

## Linear Regression

Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables (Analytics Vidhya, 2022).

We have a wide range of algorithms options to use linear regression, in this work we have tried four different options and we have evaluated the differences:

* Time Step Feature.
* Lag Feature.
* Two Variables (one dependent, one independent).
* Multi Featured (one dependent, n independent).

In both types we have used the same approach that consisted in predict the Total Adults in homeless situation in Ireland based on the past data:

Chart, line chart

Description automatically generated

## Time Step Feature

The basic object of forecasting is the time series, which is a set of observations recorded over time. In forecasting applications, the observations are typically recorded with a regular frequency, like daily or monthly (Kaggle, 2021).

Time-step features are features we can derive directly from the time index. The most basic time-step feature is the time dummy, which counts off time steps in the series from beginning to end (Kaggle, 2021).

In our case, once we have our data per month, our dummy value was calculated based on our date value:

Text

Description automatically generated

Next step was separate the prepared data to use in our predictive algorithm, our dummy value represented by “Time” was our independent feature and Total (Total adults Homeless in Ireland) was our target (dependent feature):

A picture containing text

Description automatically generated

Our time-step value represented by XT:

Text

Description automatically generated with medium confidence

After separate the training data, was time to train our model and get the results:

Text

Description automatically generated

Plot the results:

Chart, line chart

Description automatically generated

The graph shows the predictive line (in blue) vs the real data.

Model Coefficient: 22.40186713

* The coefficient signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x (The BMJ, 2023).

Model Intercept: 6027.431122448979

* The intercept value shows the point where the estimated regression line crosses the 𝑦 axis. It's the value of the estimated response 𝑓(𝑥) for 𝑥 = 0. This value determines the slope of the estimated regression line (The BMJ, 2023).

Based on the results of our model, we could predict the future values:

Text

Description automatically generated with medium confidence

How to predict using time-step?

We have 48 time series in our data frame representing dates from January 2019 to December 2022, any number from 49 will predict new values, for example:

49 = 01/2023

50 = 02/2023

56 = 08/2023

60 = 12/2023

Sample:





Analyzing our results, we can verify the following points:

* A modest coefficient indicates a worst fit and means that the model can't better explain the variation of the output with different inputs (Machine Learning Mastery, 2019).
* We can see this when our line of predicted values didn't fit with real values on the graph (Machine Learning Mastery, 2019).
* Perhaps separating the values into a smaller range, for example a specific case over time, such as the post-covid crisis, can bring us better predictive information.

**Time Step Feature (reduced data)**

Here, the same algorithm as the previous one was executed, but with a reduction in the amount of data. The main idea was to use data from the post covid cost of living crisis.

So, we could try to predict how the homeless situation in Ireland will be if the crisis is not controlled in the future.



Plotting the results:

Chart, line chart

Description automatically generated

The graph shows the predictive line (in blue) vs the real data.

Model Coefficient: 130.07368421

Model Intercept: 5797.284210526315

Based on the results of our model, we can predict the future values.

We have 19 time series in our data frame representing dates from 06/2021 to 12/2022, any number from 20 will predict new values, for example:

20 = 01/2023

21 = 02/2023

27 = 08/2023

31 = 12/2023

Samples:

Text

Description automatically generated

Text

Description automatically generated

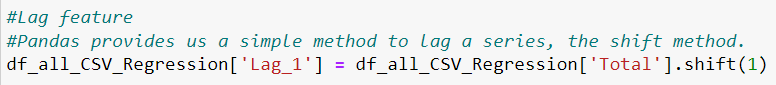
Analyzing our results, we can verify the following points:

* A larger coefficient (130) indicates a better fit and means that the model can better explain the variation of the output with different inputs (Machine Learning Mastery, 2019).
* We can see this when our line of predicted values fit with real values on the graph.

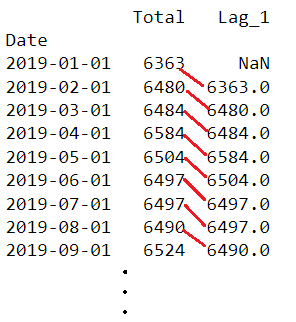
## Lag Feature

In the lag feature type, we shift the observations of the target series, in our case the Total homeless, so that they appear to have occurred later in time. In this example we've created a 1-step lag feature.

The first step, as previously mentioned, we need to shift the value of our target feature, now our data frame has a Total and previous Total value added as a new column “Lag\_1”:



Data frame:



The previous Total value was shifted as a new feature, but a null value is found because for the first row we do not have a previous value, so we need to remove this line using the dropna function and align dropping the correspondent values in target:

Company name

Description automatically generated with medium confidence

And training the model:

Text, letter

Description automatically generated

Plot the result:

Chart, scatter chart

Description automatically generated

Model Coefficient: 1.06087426

Model Intercept: -357.97086688965555

Analyzing our results, we can verify the following points:

* A tiny coefficient (1.06087426) indicates a worst fit and means that the model can’t better explain the variation of the output with different inputs (Machine Learning Mastery, 2019).

## Train Test Split: Two Variables

In this linear regression option, we could split data into training data and test data.

Unlike previous algorithms where we had only one dependent variable (target) and one independent variable, in this case we used two independent variables:

Text

Description automatically generated with medium confidence

Month and Year were separated and used as an independent variable.

Total (Total homeless) still being used as dependent (target).

The *train\_test\_split* function is used to separate a test and training dataset for the algorithm. The function takes our prepared data X and y and returns the input test and training data in separate variables.

The function has some important parameter features, which are:

* Test\_size: Determines how the training and test data will be divided, for example, when passing 0.20, the function will divide 20% of the data for testing and 80% for training (Scikit Learn, 2023).
* Random\_state: Controls the shuffling applied to the data before applying the split. Pass an int for reproducible output across multiple function calls (Scikit Learn, 2023).
* Suffle: Whether or not to shuffle the data before splitting (Scikit Learn, 2023).

The use of function model.Fit also differs in this case, where we pass the training arrays that were previously separated by the split function:

Text

Description automatically generated

After that we can predict based on trained data and the test data:

A picture containing text

Description automatically generated

R2 score of our model:

A picture containing text, orange, close

Description automatically generated

Model evaluation:

Text

Description automatically generated

Ploting the linear regression between predicted and actual values:

A picture containing diagram

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Results:

Text

Description automatically generated

Our model accuracy was not good, probably because our range of data is similar to the previous examples.

## Train Test Split: Two Variables (Reduced Data)

In this case, we used the same previous example, but with the data reduction that we have already applied before, where we want to predict how the increase in homeless will behave if the government is unable to intervene in the cost-of-living crisis.

Data Range:

Chart, line chart

Description automatically generated

Plotting the linear regression between predicted and actual values:

Chart, line chart

Description automatically generated

Results:

Text

Description automatically generated

Despite the amount of data being low - for less than 400 data points linear regression can be inaccurate - we achieved an excellent result for accuracy (more than 80), showing that we were on the right track for predictions.

Testing future predictions:

Text

Description automatically generated with medium confidence

Text

Description automatically generated

## Train Test Split: Multi Featured (Reduced Data)

In this algorithm we used the same previous system, but with the difference that we used more than two independent variables.

We used the inflation data and thus tried to predict what the homeless scenario would be like in the future with variations in inflation.

As we have already seen in past tests that our prediction is based on the period of post-pandemic cost of living crisis, we went directly to the scenario with a reduced date.

Three independent variables being used:

A picture containing chart

Description automatically generated

Predictions:

A picture containing text

Description automatically generated

Text

Description automatically generated

Results:

Carrying out prediction tests, we realized that by reducing the predicted inflation, the predicted number of homeless is also reduced. This is the expected behavior; however, the reduction is timid and does not reflect the expected correlation between inflation and homelessness.

Probably the month and year variables have a greater influence on the predicted results than inflation.

Another important factor is that we have a dataset with small amount of data, which makes it difficult for this split training data algorithm to have good results.

**Python vs Power BI for Data Analytics**

## What is Python?

Python is a high-level, general-purpose programming language, dynamically typed and garbage collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming (Wikipedia, 2023).

Python has a wide range of frameworks dedicated to analysis, including Pandas. For the most part, data analytics libraries in Python are at least somewhat derived from the NumPy library, which includes hundreds of mathematical calculations, operations, and functions (Wikipedia, 2023).

For machine learning purposes, Python offers libraries like Sklearn which provides analytics capabilities with machine learning.

Although Python is widely used, there are other languages that are also powerful for use with Data Analytics and one of them is Power BI.

## What is Power BI?

Power BI is an interactive (GUI: Graphical User Interface) data visualization software product developed with a primary focus on business intelligence. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. It offers data warehouse capabilities including data preparation, data discovery, and interactive dashboards (Wikipedia, 2023)

Power BI offers the Data Analysis Expressions (DAX) which is a programming language that is used for creating calculated columns, measures, and custom tables. It is a collection of functions, operators, and constants that can be used in a formula, or expression, to calculate and return values. DAX can be used to solve several calculations and data analysis problems (Microsoft, 2022).

## Comparison between Python and Power BI

Both Python and Power BI are powerful tools for Data Analytics, however they are different and have advantages and disadvantages at different levels. Let’s look on the main differences.

**Licensing**

The first big difference between the two is that Python and its libraries are Open Source and free, it is not necessary to acquire licenses for their use. Although Power BI has a free license, this can only be for personal use, and we can’t share content with others. For the corporate use of Power BI, it is necessary to purchase licenses.

**Graphical User Interface (GUI)**

With Power BI, it is possible to carry out all data analysis without executing lines of code, that is, without having prior knowledge of a programming language. Although DAX is considered a programming language by many, it is an "expression language" based on formulas, in a superficial comparison, it would be the evolution of Excel formulas for Data Analysis.

With Power BI it is possible to connect to various data sources such as CSV files, databases, APIs, also just using a graphical interface, the same for data cleaning and preparation and graph plotting without a line of code.

With Python, as seen throughout this study, it is necessary to have prior programming knowledge. Data manipulation involves mastering the manipulation of data frames, arrays, lists, functions, and other programming concepts.

**Programming vs Tool**

Python is 100% programming oriented and, as already mentioned, it is necessary to have prior knowledge of the programming language to use it in Data Analytics. However, Python is considered one of the easiest and fastest languages to learn, in a few months it is possible to already have the necessary knowledge to start with Data Analytics.

Power BI, on the other hand, is considered a graphical tool, despite having access to the use of programming languages (it is even possible to use Python), the focus is to give the user the ability to generate analyzes and reports in a 100% graphical way.

**Deploy / Publication**

As part of the licensed service, Power BI has a publishing service called Power BI embedded analytics, which allows you to embed your Power BI such as reports, dashboards, and tiles, in a web application or on a website.

The service already has access profile control, menu structuring and other systemic facilities without the need to build a web site. With a few clicks, it is possible to publish Power Bi as a web page and provide access control according to access profiles.

With Python it is necessary to create the site structure to accommodate what was developed, for example, in this study, if the desire was to publish the graphics for continuous use, it would be necessary to create a portal using a programming language for this purpose.

There are some pre-made Python frameworks that speed up the creation of portal structures (menu structure, access control, etc.) but it is still necessary to adjust them for a later deployment.

**Capabilities and limitations**

Big data and/or large data can be a problem using Power BI, using large amounts of data even on a computer or server with good hardware is a challenge, for example, working with more than 200,000 lines can lead to performance problems and crashes.

Python at the same time is lightweight and has excellent performance, so it is possible to handle large volumes of data without major performance compromises.

Python has more advanced data cleaning and preparation features as it is a complete advanced programming language. Power BI provides basic and limited resources, for advanced data cleaning and preparation, it is necessary to use programming languages such as Python itself within Power BI.

## Comparison between Python and Power BI

Both are powerful languages/tools and can deliver great results in Data Analytics, but Power BI can be considered a more user-friendly tool for use by advanced users who do not have prior knowledge of a programming language.

The Python language, on the other hand, has more advanced features and can be used for complex scenarios that may involve large masses of data or advanced features such as Machine Learning.

**Conclusion**

Homeless situation has been increasing in Ireland quickly and the cost-of-living crisis that was generated by the pandemic and the war in Ukraine proved to be direct correlating factors.

The situation was relatively stable until mid-January 2019 where, due to policies aimed at combating the covid pandemic, we had a reduction in cases by about 8%.

The situation then begins to deteriorate from August 2021 onwards, with an uninterrupted strong growth in cases of around 40% in total.

All ages and genders are affected, but the situation is even more serious among adult men aged 25 to 44 years, representing about 55% of all cases.

Among the regions, Dublin has the worst situation, with a rate of around 4.35 inhabitants per 1000, almost five times more than the second most affected region (Mid-West). On the other hand, North-West has the best situation, with a rate of just 0.3 per 1000.

The increase in inflation, which is a key indicator when we have a cost-of-living crisis, showed a strong correlation with the increase in homelessness. As the indicator increased, a period of eight months was perceived for the same proportion of growth to be seen for homeless people.

Inflation showed a reduction in the last five months of data available for analysis, as new data becomes available, we will be able to verify if we will also have a decrease in homelessness.

With the use of machine learning, we were able to predict what the homeless situation will be like in the future if the cost-of-living crisis continues. We were able to show a correlation between inflation and time to make forecasts, but the correlation was timid.

**References**

Data.Gov.IE (2023) HSA09 - House Construction Cost Index. Available at: <https://data.gov.ie/dataset/hsa09-house-construction-cost-index> (Accessed: 28 April 2023).

*Data.Gov.IE (2023) National House Construction Cost Index. Available at:* [*https://data.gov.ie/dataset/national-house-construction-cost-index?package\_type=dataset*](https://data.gov.ie/dataset/national-house-construction-cost-index?package_type=dataset) *(Accessed: 28 April 2023).*

FRED Economic Data (2023) Employment by Economic Activity: Construction: All Persons for Ireland (LFEACNTTIEQ647S). Available at: <https://fred.stlouisfed.org/series/LFEACNTTIEQ647S> (Accessed: 28 April 2023).

FRED Economic Data (2023) Total Construction for Ireland (PRCNTO01IEQ659S). Available at: <https://fred.stlouisfed.org/series/PRCNTO01IEQ659S> (Accessed: 02 May 2023).

FRED Economic Data (2023) Production: Construction: Total Construction: Total for Germany (DEUPROCONMISMEI). Available at: <https://fred.stlouisfed.org/series/DEUPROCONMISMEI> (Accessed: 28 April 2023).

FRED Economic Data (2023) Employment - by Economic Activity: Construction: All Persons for Germany (LFEACNTTDEQ647S). Available at: <https://fred.stlouisfed.org/series/LFEACNTTDEQ647S> (Accessed: 28 April 2023).

FRED Economic Data (2023) Other Prices: Cost of Construction: Residential: Total for Germany (OPCNRE01DEQ661N). Available at: <https://fred.stlouisfed.org/series/OPCNRE01DEQ661N> (Accessed: 28 April 2023).

FRED Economic Data (2023) Employment by Economic Activity: Construction: All Persons for Italy (LFEACNTTITQ647S). Available at: <https://fred.stlouisfed.org/series/LFEACNTTITQ647S> (Accessed: 02 May 2023).

FRED Economic Data (2023) Total Cost of Residential Construction for Italy (OPCNRE01ITM661N). Available at: <https://fred.stlouisfed.org/series/OPCNRE01ITM661N> (Accessed: 02 May 2023).

FRED Economic Data (2023) Production of Total Construction in Italy (ITAPROCONQISMEI). Available at: <https://fred.stlouisfed.org/series/ITAPROCONQISMEI> (Accessed: 02 May 2023).

FRED Economic Data (2023) Population, Total for Italy (POPTOTITA647NWDB). Available at: <https://fred.stlouisfed.org/series/POPTOTITA647NWDB> (Accessed: 04 May 2023).

FRED Economic Data (2023) Active Population: Aged 15-64: All Persons for Italy (LFAC64TTITQ647S). Available at: <https://fred.stlouisfed.org/series/LFAC64TTITQ647S> (Accessed: 04 May 2023).

FRED Economic Data (2023) Population, Total for Germany (POPTOTDEA647NWDB). Available at: <https://fred.stlouisfed.org/series/POPTOTDEA647NWDB> (Accessed: 04 May 2023).

FRED Economic Data (2023) Active Population: Aged 15-64: All Persons for Germany (LFAC64TTDEQ647S). Available at: <https://fred.stlouisfed.org/series/LFAC64TTDEQ647S> (Accessed: 04 May 2023).

FRED Economic Data (2023) Population, Total for Ireland (POPTOTIEA647NWDB). Available at: <https://fred.stlouisfed.org/series/POPTOTIEA647NWDB> (Accessed: 04 May 2023).

FRED Economic Data (2023) Active Population: Aged 15-64: All Persons for Ireland (LFAC64TTIEQ647S). Available at: <https://fred.stlouisfed.org/series/LFAC64TTIEQ647S> (Accessed: 04 May 2023).

Reddit (2023) Property prices Ireland: Exodus of landlords behind 40pc of house sales as property prices set for massive slowdown. Available at: <https://www.reddit.com/r/ireland/comments/10ao287/property_prices_ireland_exodus_of_landlords/> (Accessed: 10 May 2023).

Data Science Process Alliance (2023) What is CRISP DM? Available at: <https://www.datascience-pm.com/crisp-dm-2/> (Accessed: 13 May 2023).