**Homeless Ireland 2019 to 2021**

Student Name: Denise Martins

Student Number: sbs24065

Machine Learning

Lecturer: Dr. Muhammad Iqbal

https://github.com/Denisbs24065/MachineLearning

CCT College, Dublin

March, 2024

**Table of content:**

[Introduction](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250006) 3

Problem Definition and [Objectives](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250005) 3

[Scope](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250003) 3

Figure 1: The data mining life cycle 4

Business Understanding 4

Data Understanding 4

Data Preparation 4

Modeling 4

Evaluation 4

Deployment 5

[Potential data for the project](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250002) 5

Data sources 5

[Ethical considerations](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250001) 5

[References](file:///C:\Users\denis\Downloads\CA%201%20Example%202.docx#_TOC_250000) 6

### Introduction

### Scope

I will analysis this data, about occurrence of homeless people in the years between 2019 and 2021, including region of occurrence, total of adults, and with year for example, to analyse and insights that can drive some strategy and decrease the occurrence of homeless and understand the cases of this numbers.

I load a CSV file named Homeless\_Ireland\_2019\_to\_2021.csv and apply .head() function to understand the size of the dataset and with information I have for this analyse.

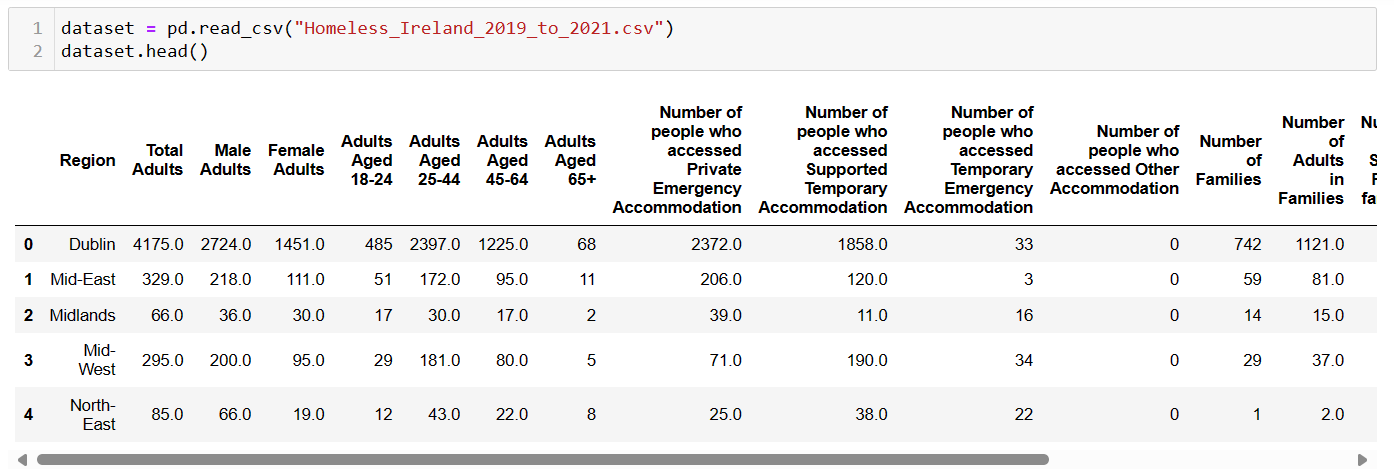


Figure 1

I use two search functions, searching for Non-Null values or missing data. The function isnull().sum() checks for missing values in each column of the dataset and sums up the count of missing values for each column. In answer I had one column that had Non-Null values, but I drop this column that was with same content as another column.

And the function .info() provides a concise summary of the DataFrame's information including the number of entries (288 rows) and the data type of each column. It also shows the total memory usage of the DataFrame.

Using the option replace(), I replace qualitative value for quantitative in two important columns that I will be using in my analysis.

According with Geeks for Geeks “Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If [feature scaling](https://www.geeksforgeeks.org/python-how-and-where-to-apply-feature-scaling/) is not done, then a [machine learning](https://www.geeksforgeeks.org/machine-learning/) algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values”.

In my analyse I applied MinMaxScaler, Normalization and Standardization to improve the results.

I split my dataset in test and training, X\_train, X\_test, y\_train, y\_test using cross validation techniques for the variation in the accuracy across using three trainings splits (20%, 25% and 30%).

I decide to apply Linear Regression because, I the goal of this analyse is understand the relationship between a dependent variable (Total of Adults) and one more independent variable (in my case year), and this relationship can be approximated by a linear equation.

In the Classification It is provided the class of the dataset based on the independent input variable, which although it is a good idea to analyse the occurrence in certain regions, it is this also important to know the total per year. Regression, on the other hand, predicts the continuous output variables based on the independent input variable, the total number of adults per year.

In Linear Regression I plot comparing the actual values (y\_test) with the predicted values (y\_pred) for different trainings as was mentioned before. Bellow all three results of the plot:

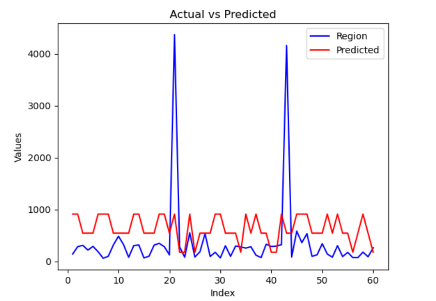
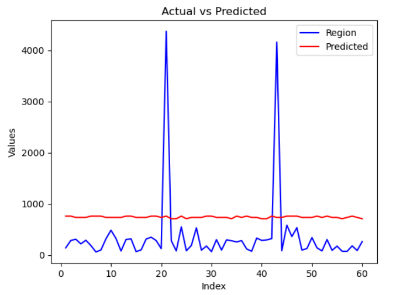
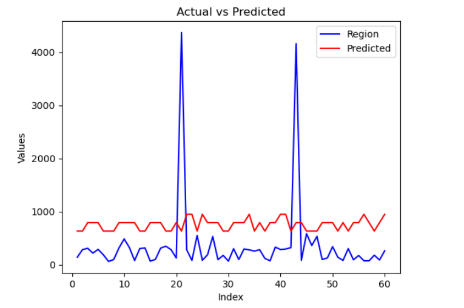
  

Figure 2 (20%) Figure 3 (25%) Figure 3 (30%)

Here in this figures we can check the blue line represents the actual values (y\_test) and the red line (y\_pred), represents the predicted values, both corresponds to a specific index, but the red line its associated predicted values. By comparing the blue and red lines on the chart, we can visually assess how well your predictive model is performing. Ideally, the red line (predicted values) should closely follow the blue line (actual values). However, as we can see the discrepancies between the two lines, they indicate areas where the model's predictions deviate from the actual values. This visualization helps to visually evaluate the performance of the predictive model, which despite the aforementioned discrepancy still presents a good visualization.

2 - Cross Validation

3 - Hiperparameter

### Conclusion

### References