**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | MSC\_DA\_CA1 |
| **Lecturer Name:** | Sam Weiss/ David Gonzalez  Bharathi Chakravarthi/ Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Aaron Bowman |
| **Student Number:** | sba23336 |
| **Assessment Due Date:** | 10/11/2023 |
| **Date of Submission:** | 11/11/2023 |

**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. |

Contents

[Abstract 2](#_Toc150714539)

[Introduction 3](#_Toc150714540)

[Data Gathering and Exploration 3](#_Toc150714541)

[Data Gathering 3](#_Toc150714542)

[Exploratory Data Analysis 3](#_Toc150714543)

[Data Cleaning 3](#_Toc150714544)

[Data Visualisation 4](#_Toc150714545)

[Data Visualisation Choice and Justification 8](#_Toc150714546)

[Conclusions from Data Exploration 9](#_Toc150714547)

[Statistical Analysis 9](#_Toc150714548)

[Descriptive Statistics 9](#_Toc150714549)

[***Working Population*** 9](#_Toc150714550)

[***Population*** 9](#_Toc150714551)

[***Earnings Euro*** 9](#_Toc150714552)

[Machine Learning 10](#_Toc150714553)

[Outline 10](#_Toc150714554)

[CRISP-DM Framework 10](#_Toc150714555)

[Linear Regressor Analysis 11](#_Toc150714556)

[Support Vector Regressor Analysis 11](#_Toc150714557)

[Comparison 11](#_Toc150714558)

[Analysis 12](#_Toc150714559)

[Methodology and Program Design 12](#_Toc150714560)

[Machine Learning Program Design 12](#_Toc150714561)

[EDA Program Design 13](#_Toc150714562)

[Conclusions 13](#_Toc150714563)

[References 13](#_Toc150714564)

# Abstract

*Through the use of Statistical Analysis and Machine Learning, we can see that there are a number of features that affect the Total Household Income of a given County in the Republic of Ireland. However, due to the presence of outliers and tight grouping of the existing independent variables, clearly more features are needed in order to efficiently model this relationship in a way that could be used in a predictive manner.*

# Introduction

There are a variety of Income levels across the various Countries in the Republic of Ireland. Throughout the course of this analysis we aim to investigate the relationship between Income Levels without a County, said Counties population and the number of people working in major Industry groups within those group to determine which of these has the largest effect on predicting the income of households.

# Data Gathering and Exploration

## Data Gathering

All data for this research was gathered from the Central Statistics Office (CSO) website. All data contains numerous features. But most important, a year in which it was collected, and a County & State feature describing the geographic area that the data refers to.

Additional each of the of the datasets included at least one feature that I aimed to use for analysis, be it population, population working is a specific area of Industry, of a Income value for households within that County.

Three datasets were selected for final use:

1. CNA27.20231016T221051.csv – Which represents employment numbers by Country, Year, and Industry
2. RAA02.20231109T231100.csv – Which represents a number of different incomes measurements by different categories by income type.
3. FY001.20231016T201055.csv – Which represents the populations of each county and the State as a whole across different time periods.

## Exploratory Data Analysis

These datasets each contains numerous features, with some commonality across them due to the nature of how the CSO website produces CSV files. Three common features across all datasets were the Statistical Label (discrete variable labels the type of data), UNIT, which denotes the what unit of measurement is being used for each observation and VALUE which stores the value for each given observation.

All observations values are numeric, either as floats or as integers. For the purposes of this work, it does not matter which.

All of the County values are loaded as strings, albeit with some differences in the naming conventions for County and Region names. These will be addressed in the data cleaning stage.

Only in our population there are there NAN values, specifically due to the different categories of information captured in pervious census collections (as the observations go back to the 1800s).

## Data Cleaning

The first step we undertake in this process is to fill all NAN values with numpy’s nan so that they can easily be distinguished. It was through this step that determined that our datasets individually had no nan values, but they would have a number if the data was joined improperly.

There are many unneeded observations and variables across the data. Particularly, there are hundreds of observations in the observation dataset that we have no other data to join with due to how far back it goes.

Additionally, our population dataset contains a breakdown of the population by both genders, and the aggerate. Because none of our other datasets contain these dimensions, we will simply take the aggerate value.

The employment dataset contained a number of County observations that were not consistent with the other datasets. This appears to be due to the changes that were made over time to the geographic areas used to describe the data; specifically the Waterford, Limerick and Tipperary areas. On this basis, we exclude any observations that had NAN values included in them.

Finally, our income dataset contained 15 different Statistic Labels measuring different types of income across Counties and Years. Some of these include “*Disposable Income per Person”* or “*Net Interest and Dividends*”. For the purposes of this analysis, we used “*Total Household Income*” by County. This value is represented in hundreds of millions, and is the aggregate Total Household Income for every Household in a given County.

The VALUE columns from each dataset were renamed in accordance with what it represented in that dataset. Further columns were renamed such a CensusYear to Year in order to make joins easier.

The joins between the datasets took two steps. First we joined the employment data to the population data on the Year and County columns. The in order to facilitate a full-left join on that now joined dataset, we looped through the joined dataset and matched relevant values from the income dataset based on Year and County. Finally, using the WorkingPopulation and Population variables by County, we worked out the percentage of people working in a given county Industry by County.

This result in a final dataset included seven columns and 1728 rows of data. This consisted of the Year, Industry, County, WorkingPopulation Population, EarningsEuro and pctIndustryEmployement. This dataset contains no NAN values.

## Data Visualisation

In an effort to understand the relationships between the various variable in the now merged data, we used various data visualisations to plot them against one another.

A graph of a number of people

Description automatically generated with medium confidence

Figure 1: Earnings (€) by County and Year

As can be in Figure 1: Earnings (€) by County and Year there is a steady increase year over year, with sudden larger increases for each of the major population centres; Dublin, Galway and Cork. In those areas we see major increases.

A graph of different colored lines

Description automatically generated

Figure 2 Earnings (€) by Percentage Employed in Industry by Year

As can be seen in *Figure: 2 Earnings (€) by Percentage Employed in Industry by Year* there are two to three major spikes per year, each representing the three major population centres (Dublin, Cork, Galway). We can also see there are a number of years where the total percentage of people in employment caps out at a lower level that others. 2006 saw the highest peak while 2011 saw the lowest peak.

A graph of a line

Description automatically generated with medium confidence

Figure 3: Population by Earnings (€) by Year

As seen in Figure*:* 3 Population by Earnings (€) by Yearthere is a relationship between the population Mean Earnings of given County and the Population of said County. This relationship appears to shift further along the Earnings axis over successive years.

A white background with many small colored dots

Description automatically generated

Figure 4: Percentage Employed in all Industries by Earnings(€) by Year and County

As shown in *Figure 4: Percentage Employed in all Industries by Earnings(€) by Year and County,* the vast majority of datapoints exist below the €10,000 threshold, irrespective of the pctIndustryEmployement value. This aligns with our other analysis that shows that most Counties values are clustered together, with the major population centres acting as outliers.

A graph with numbers and lines

Description automatically generated

Figure 5: Population Boxplots by Year

As can be seen in *Figure: 5 Population Boxplots by Year,* there are the same number of outliers each year, albeit with the values of those outliers increasingly steadily. These outliers are each from the major population centres (Dublin, Cork, Galway). The remaining Counties are grouped tightly around the mean.

A graph with numbers and lines

Description automatically generated

Figure 6 Earnings (€) by Year

A similar pattern is visible with regards to outliers in *Figure: 6 Earnings (€) by Year* as was in *Figure: 5 Population Boxplots by Year.* In both cases, the major population centres act as the outliers, while the remaining observation’s values group tightly around the mean.

## Data Visualisation Choice and Justification

For the data visualization and exploration we chose to use three types of charts in order to properly understand the data being used:

1. Line Charts
2. Scatter Plots
3. Box Plots

These were chosen because they effectively illustrate the relationship between the chosen subsets of our data, be it the categorical relationship between County’s and Earnings, or the relationship between pctIndustryEmployement in Industry and earnings.

As the continued, it became clear that a number of outliers existed in the data, with *Figure 4: Percentage Employed in all Industries by Earnings(€) by Year and County* illustrating best the fact that a small number of counties act as substantial outliers in all areas.

Therefore we used box plots to analyse these outliers in more detail, seeing more clearly in *Figure 5: Population Boxplots by Year* and *Figure 6 Earnings (€) by Year* the number of outliers was consistent from year to year.

## Conclusions from Data Exploration

From viewing a variety of visuals that display a number of combinations of features within our data we can see that

1. There is a connection between the numeric features that were selected for analysis.
2. Those connections are present across all years.
3. The outliers are consistently centred on observations of the major population centres.
4. All other observations are highly similar.

# Statistical Analysis

## Descriptive Statistics

Exploring our joined dataset (the same set used for the visual analysis), we can note a number of interesting statistical features that can further our understanding of the dataset and how the features relate to one another.

All of our joined columns have the same number of values within them , 1728 observations. When we exclude those observations, we get only 1664 observations. This second dataset is the one that we will use for most of our analysis.

Discounting Year (a our only numeric discrete variable), we have four numeric continuous variables

### ***Working Population***

Our WorkingPopulation variable, broken out by County and Year has a minimum value of 32, maximum of 614,776 (representing a observation that is the aggerate for the whole state) and a mean of 8876.96. The large standard deviation; 31,988, which is 19.21 times the mean. This difference is due to the incredibly small numbers of workers in the *“Mining, quarrying and turf production”* sectors in a number of Couties. This plus Dublin and other major population centres status as outliers as show in Figure 4: Population Boxplots by Yearcreates a high standard deviation.

### ***Population***

The Population variable by contrast displayed a much smaller Standard Deviation (232,151 people), only 1.37 times the Mean value of 168,338. This can be attributed to the fact that the variable is less discrete, only being divided by Year and County, instead of Year, County and Industry like the WorkingPopulation or pctIndustryEmployement variables.

### ***Earnings Euro***

EarningsEuro describes the mean household income for a give County and Year. It is heavily skewed by the major population centres, with its maximum value of €45,235.86 being over 100 times greater than its minimum value of €445.61. The fact that the maximum value is also 12 times greater than the 75th percentile value demonstrate how much of an outlier this value is. This tracks with our other outlier values, as the County of Dublin has a consistently higher values across the board

# Machine Learning

## Outline

Our objective is to create test a number of machine learning models in order to predict the EarningsEuro value of a County by the input of Population, WorkingPopulation, pctIndustryEmployementand County.

Note that for a prediction mode of this type we would exclude the Year as a input variables this would make it impossible to predict into the future as this data would be out of scope of the model.

The second purpose of our regression model is to understand what of these input features is the best predictor of a County’s’ EarningsEuro value.

A number of different regression models will be tested in order to determine which best predicts the chosen output.

All values given below are a mean of the cross validation version of the evaluation metric listed. That means all five values, one from each cross validation were averaged together in order to give a single representative value for each evaluation metric.

## CRISP-DM Framework

CRISP-DM or Cross Industry Standard Process for Data Mining is an approach to managing a data science project in a standardised manner (Shearer, 2000). It broadly consist of:

1. Business understanding – What does the business need?

2. Data understanding – What data do we have / need? Is it clean?

3. Data preparation – How do we organize the data for modelling?

4. Modelling – What modelling techniques should we apply?

5. Evaluation – Which model best meets the business objectives?

6. Deployment – How do stakeholders access the results?

In our case, we need to identify the Counties with the lowest Household Earnings and the relationship between that and our Employment data. Our data from the CSO is not clean, hence the steps outlined in the *Data Cleaning* were undertaken.

Once the data was joined, preprocessing was needed in order to encode and scale the data appropriately for our modelling approach.

Three regression models have been chosen in order to determine that which is most accurate, a Linear Regressor, A Decision Tree Regressor, and a Support Vector Regressor.

Of those models, the Linear Regressor proved to be the most effeitive, neither over nor underfitting, but it would appear that that based on our testing (elaborated on bellow) that there are still some features that should be added in order to improve our understanding of the relationship between our independent and dependent variables.

## Linear Regressor Analysis

Our first model tested was a sklearn’s Linear Regression model. This model on its baseline fit (no additional features excluded from the training data bar the target variable) had an R-Squared Percentage score of 99.42%, even as it had a cross validation with a k=5. This would indicate a very that most variables were accounted for in the fitting of the model and would give a strong indication that the model would prove acceptable for prediction provided input values were available.

Notable, even when select features were removed from the training dataset (no more than a single feature per iteration), the R-Squared score did not drop below 95%. The lowest R-Squared score that was achieved by this model was when the Population feature was removed, resulting in a R-Squared score of only 95.54%. This feature being removed also led to the highest being removed did not however lead to the highest Mean Square Error (MSE) of the test. That came when County was removed (MSE of 1069314.74). That MSE was 3.8 times the baseline score.

**Decision Tree Regressor Analysis**

A Decision Tree Regressor, fitted to the same data as the baseline Linear Regression data was tested next. This model was evaluated using the same metrics as the Linear Regression model to facilitate easy comparison between the two models.

This achieved a R-Squared Score of 100%, with MSE and MSE Percent scores all reporting zero. This result appears unlikely, and can be attributed to the nature of decision trees to overfit in order to correctly maximise the R-Squared score.

Only when the Population feature was removed from the training data did the R-Squared score drop bellow 100% (to 93.04%). There was a small uptick in the MSE however when the County feature was removed, rising from 0 to 90.28.

## Support Vector Regressor Analysis

The Support Vector Regressor preformed worst of the three chosen Regressors, achieving a baseline R-Squared Score of -5.296%. This poor performance persisted as different features were removed in line with our testing methodology, becoming worst when Population feature was removed. This version of the model only achieved a R-Squared Score of -6.053%, notably worse than any of its counterparts.

It should be noted that Support Vector models are also some of the slowest to build due to the methodology applied by them behind the scene. This poor performance and longer than average build time makes for a poor choice of model for our analysis.

## Comparison

The table below represents the results of the baseline models of each type. That is the model that has no features removed from its training set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R-Squared Percentage** | **Mean Square Error** | **Mean Square Absolute Percent** |
| Linear Regression | 99.42% | 281033.11 | 0.09% |
| Decision Tree Regressor | 100% | 0.0 | 0.0% |
| Support Vector Regressor | -5.296% | 51543541.83 | 0.68% |

As can be seen, the Decision Tree Regressor preforms the best out of all three of our model types, with a best R-Square Score of 100%. However, it appears highly unlikely that this model is the best choice for future prediction as opposed to past analysis. Decisions trees are highly susceptible to small variables in input parameters and subject to overfitting.

Comparatively, our Linear Regression Model preformed slightly worse, with an mean R-Squared Score of 99.42%, but a notably larger MSE score. While this is a small drop in R-Squared, the increase in MSE would indicate that it is not perfectly fit, which makes its predictive potential more valuable.

Finally, our Support Vector Regressor achieved a negative R-Squared Score, making it by far the worst model that we tested.

## Analysis

From the analysis of the baseline models, and our tests with removing certain features from those models and refitting them we draw a number of conclusions.

There is a relationship between our features of sufficient strength that we can predict the Total Household Earnings of a County based on the features used.

Of the features we used to create these models, Population of a given County at a given time was the most important. With all models seeing a drop in their performance across all evaluation metrics when this was removed. In the case of our Liner Regression model, County also appeared to have a larger influence on the accuracy of the model.

Finally, we can conclude that our optimal choice for a Regression model is a Linear Regression model, with the alternatives either appearing to overfit (as is the case of our Decision Tree Regressor) or fail to fit with a satisfactory level of accuracy (our Support Vector Regressor).

# Methodology and Program Design

## Machine Learning Program Design

In an effort to make the process of iterating through different combinations of features and models more efficient, we create a single function that accepted a input dataset, a target feature and a regressor. This standardised function let us loop through a variety of options quickly

This function used sklearn’s pipeline functionality in order to easily allow iteration through different types of Scalers, Encoders and Regressors. Through the process of experimentation, we determined that the StandardScaler and the OneHotEncoders were the most successful for our dataset.

To facilitate a wide range of testing on our data, lists were created of both the discrete and continuous features. These were then edited as needed by the parameters passed to the function, and the target variable removed. Then, these lists were fed to the Column Transformer part of the pipeline, ensuring that the correct variables were acted on by the corresponding encoder or scaler.

We also facilitated a system whereby select features could be excluded from the training data, in order to assess what features if any were most important for our task.

In order to limit the possibility of our linear regression model overfitting, k-fold cross-validation was used, where k=5 was used. We then evaluate the mean scores of the cross-validation process to arrive at the final scoring metrics for our regression models.

Overall, this process allowed for the rapid testing and evaluation of 18 distinct combinations of models and features, letting us extract significant insight into the performance of different models, and the importance of a wide range of features.

## EDA Program Design

In comparison to the design of the Machine Learning Program, our EDA Program design was simple. A simple script where each cell executed a single task, designed to let us explore the data in depth.

The data join that was needed to let us visualise and model the dataset was preformed and two csv’s were produced; one with the Year feature included, and one without.

Our plotting was done using pythons Matplotlib, with datasets being filtered, split and reworked as needed for each individual plot. This was a simpler approach that allowed for iterations on each specific plot without impacting the outputs of existing plots.

# Conclusions

In conclusion, we can say that our objective of analysing the factors that drive differences in was only partially achieved. While we did find that Population and to a lesser extent, County were features that when removed had impacts on overall model accuracy, they both their effect saw less than a 5% reduction in the overall R-Squared Score.

This is not significant compared to the overall high performance of the Linear Regression model, and suggests that there are further features that may better explain the differences between Counties with regards to their Total Household Incomes.

Further research adding more features and ideally, adding more time periods for study would be valuable to identify what these features are and by extension what the extend of their effect on the differences between Counties are.

# References

Shearer, C., 2000. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing,* Volume 5, pp. 13-22.

[Figure 1: Earnings (€) by County and Year 5](#_Toc150714565)

[Figure 2 Earnings (€) by Percentage Employed in Industry by Year 5](#_Toc150714566)

[Figure 3: Population by Earnings (€) by Year 6](#_Toc150714567)

[Figure 4: Percentage Employed in all Industries by Earnings(€) by Year and County 6](#_Toc150714568)

[Figure 5: Population Boxplots by Year 7](#_Toc150714569)

[Figure 6 Earnings (€) by Year 8](#_Toc150714570)

**GitHub Link**

https://github.com/cct-abowman/CA1.git