National College of Ireland

BSHTM4

Data Analysis

2020/2021

Owen Thompson

X17516666

X17516666@student.ncirl.ie

The Correlation Between Suicide and Economic Inequality

Technical Report

Contents

Executive Summary

The purpose of this report is to conduct a statistical analysis on suicide data in Ireland in the year of 2016 (most recent available), and data on factors which could potentially influence it such as average household income per county and average monthly rent rates per county.

Throughout this report, I will be testing several aspects regarding the data and determining if it is suitable as a means of determining a correlation between suicide per county and various levels of economic inequality based on the factors mentioned above. I will outline all aspects of this project, from its inception to its conclusion below. These aspects include the methodologies necessary in order to carry out this project, the analysis stage of the project, the evaluation and interpretation of the results, and a conclusion detailing my final thoughts on the project following the evaluation process. I will also be outlining any future potential which this project/concept could entail and will be attaching any relevant documents to the appendix section of this report.

1.0 Introduction

This section will address the background of this concept, the aims of the project, the technology used in order to effectively execute it and the overall structure of the report.

1.1. Background

I undertook this project as I believe that there are several socioeconomic factors which have an influence on the rate of suicide in Ireland. Through carrying out a statistical analysis on the data gathered on these factors I hope to prove this assumption using empirical statistical evidence, and if I cannot, determining why, and analyse the alternative results.

As part of my final year curriculum, I have completed modules for both business data analysis and advanced business data analysis. During this time I identified a number of statistical test which I determined would be effective in carrying out the tasks outlined in my software project as correlation is a common factor in the modules.

1.2. Aims

The aim of this project is to achieve the following:

1. Successfully implement the KDD data mining methodology.
2. Determine if the variables within the linear models created are themselves linearly related.
3. Finding the relevant metadata relating to both the simple linear regression model created and the multiple linear regression model created and interpreting them in a manner that will allow me to create a detailed report on each aspect of the metadata.
4. Acquire new R programming skills which will allow me to effectively carry out my analysis. I am to acquire this knowledge through online video tutorials and peer

reviewed books and studies relating to data mining and data analysis and referencing them appropriately in the Harvard referencing format.

5. Challenging assumptions on the output of the tests with particular emphasis on assumption relating to the linear models used, and the correlation coefficients found.
6. Finding both the skewness and kurtosis of the variables within the data as they are important indicators as to how the data is distributed (leptokurtic/platykurtic).
7. Evaluating my overall analysis findings in order to reach a conclusion as to the correlations between the variables involved, the legitimacy of the correlations based on factors which must be addressed beforehand (assumptions, linearity, distribution)
8. Outlining the conclusions derived from my findings in the evaluation and results stage of the project. In this section I will determine the advantages and disadvantages of various aspects of the process including the methodology and the data used.

1.3. Technology

The technology used in this project are as follows:

- R (data programming language): This will be used as the sole programming language within the data mining process of this project.
- RStudio: An R programming integrated development environment. This is where R code will be written as part of the data mining and analysis stages of the project.
- RMarkdown: An R script writing file type used for outputting and displaying project artefacts such as code in a HTML or PDF format.
- Microsoft Excel (comma-separated value files).

1.4. Structure

This report will cover several key areas regarding the various elements of this project. I will describe:

- The data, how it was acquired, explored and extracted.
- The methodology applied and how it will be implemented throughout the project.
- The analysis component of the project, i.e., the practical aspect of the project, covering all the coding necessary in order to acquire the necessary results.
- The evaluation and results of the analysis once it has been conducted. In this section of the report, I will be stating the findings of my analysis in detail.
- The conclusions section will focus on the interpretations of our results and what they mean for the data being used.

2.0 Data

This section will outline where I sourced my data and the variables contained within the dataset that I have used.

2.1. Data Source

My data was sourced from the Irish Central Statistics Office's (CSO) online database. The CSO provides citizens with statistics and records of Ireland's people, economy, and society. The data they provide can be used to inform decisions in areas relating to health, the environment, and the economy. The data which I will be using falls under the categories of people and society, labour market and earnings, and economy. The CSO database allows you to create a dataset based on parameters you specify, meaning that I was able to select the factors relevant to my project and they were converted into a csv format which I could then read into R.

Rent data was gathered from articles recording average monthly rent rates per county, with a breakdown of different Dublin constituencies in which the mean value was determined in order to be applicable to the data set and be the same format as the other values within the variable.

(Weston, 2017)

(CSO, 2016)

2.2. Variables

The following variables used in this project are as follows:

- Annual suicides per county (dependant variable).
- Average monthly rent rate per county (independent variable).
- Average annual household income per county (independent variable).
- Population per county (independent variable).

3.0 Methodology

This section of the report will focus on the transition between acquiring my data and analysing it.

3.1. KDD

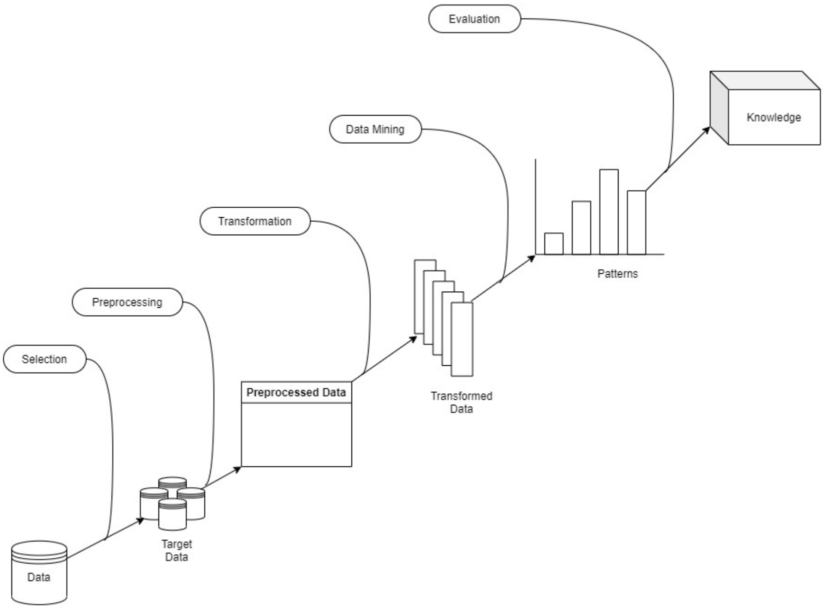
I will begin this section by first outlining the data mining framework in which I will be using throughout this project. The framework I have chosen is the KDD (knowledge discovery in databases) framework.

What is KDD?:

The abbreviation KDD refers to the application of the scientific method of carrying out data mining techniques. Whilst this framework is used as a means of carrying out data mining techniques, KDD is also a methodology which is used to extract and prepare data in as well as carrying out decision making steps in relation to actions to be taken once the data mining process has been carried out.

Steps of KDD:

The KDD framework consists of several important stages in the data mining process.



(Azevedo & Santos, 2008)

1. **Goal Identification:** To correctly carry out this initial process, there needs to be an understanding of the data being considered. Determining the goals of the project requires a concise statement and hypothesis, ones which clearly state the aims and intentions of the process. Contained in this statement should be the problem which this report aims to address, the tools used, project cost, and delivery date.
2. **Creating a Target Data Set:** From the chosen domain, a target data set must be extracted. Data used in this project was extracted via the CSO's database selection tools.
3. **Data Pre-processing:** Data must be optimised so that it is eligible to undergo transformation. Accounting for certain missing values, process outliers or attribute values which are incorrect, will make the processes in the later stages of the framework far less time-consuming. R and RStudio will be used as the data pre-processing tools for my project and the above actions will be carried out through these means.
4. **Data Transformation:** When carrying out this stage of the KDD framework, I will be using the data transformation process as a means of both adding and removing and instances and attributes which I deem either necessary or unnecessary to my data. I will carry out the normalisation of this by changing certain non-numeric values which are not essential to the statistical test which will be carried out during the analysis stage of this project.
5. **Data Mining:** This is the stage of my project in which I will conduct several statistical tests in order to determine the quality of the data in relation to how well it can determine a correlation between suicides and average annual household income as well as average rent per month.
6. **Interpretation and Evaluation:** This is the step in which the linear model is deemed either fit or unfit to be used as a means of testing the data correlation. If it is successful, it must be then translated into terms which can be interpreted by the users.
   This stage also refers to the overall results of each test, i.e., a table displaying the results for the skewness and kurtosis of each of the four variables in which the data is comprised of.
7. **Taking Action/Knowledge:** This is the last phase of the KDD framework and consists of applying knowledge which has been gathered throughout the process. This is the point in which a return on investment is expected. Following the successful application of

knowledge gained form the discovery process, several results are made possible, and in the case of my project I will be writing a detailed report documenting the process from start to finish.

(Skrubbletrang & Vachon, 2012)

(McCue, 2015)

(Roiger, 2017)

(Aiguo & Jiang Lanling, 2011)

3.2. Statistical Tests

Before determining the correlation coefficients of our data, there are a number of statistical assumptions which need to be checked before going any further. This will test id the data being used meets then requirements to be tested using the Pearson's correlation coefficient formula.

3.2.1. Tests Overview

- **Pearson Correlation Coefficient**
  This will be used to measure the level/significance in the association between the variables used. PCC uses the statistic *"r"* which ranges between -1 to +1. If *r* is zero, it implies that there is no linear relation between the two variables. The closer *r* gets to zero, the weaker the linear association becomes. Equally, the farther *r* is from zero, the stronger the linear association becomes. A perfect positive correlation occurs when the correlation coefficient is 1. A perfect negative correlation occurs when the correlation coefficient is exactly – 1.
  When interpreting the output of this test, I will also be using visual representations of the test output as part of my analysis.

- **Regression**
  This test will focus on how the variables involved are or are not interrelated.
  I will be testing two types of regression, simple linear regression, and multiple linear regression. The type of regression analysis that I will be conducting through these means is a *causal analysis*. Causal analysis considers the independent variable to be the cause of the dependant variable.
  The **simple linear regression** aspects of this analysis will constitute a dependant variable (Y) and an independent variable (X). These variables will be linearly related. This form of regression will be used to help describe the linear dependence between two variables in my data.
  The **multiple linear regression** aspects of this analysis will constitute adding a 3rd variable to the existing linear regression model.

- **Confidence interval**
  The confidence interval of the linear models will represent the likelihood that a parameter will occur between a pair of values close to the mean value. They will

measure degrees of either certainty or uncertainty within the data. The confidence level applied to my data will be 95%. For example, if the 95% confidence interval was between correlation coefficient values of 0.6884426 and 0.8854712 and the actual correlation coefficient value outputted was 0.7922459, it can be determined that the predicted confidence interval is accurate as the correlation coefficient falls between its two values.

(Pyle, 2003)
(Hayes, 2021)

- **Kurtosis**

  Kurtosis is used to measure the distribution's peak level. The larger a distribution's kurtosis, the more peaked the distribution is. Kurtosis's calculation can be reported as a value either relative or absolute. If the kurtosis is absolute, it will likely be a positive number. In terms of normal distribution, the absolute kurtosis is the value which is used as a means at which relative kurtosis is computed. As a result, the relative kurtosis of the distribution is determined between the absolute kurtosis and 3.

  When negative, kurtosis is known as being platykurtic, meaning that it likely has a flatter distribution rather than a normal one. When positive, kurtosis is said to be leptokurtic, meaning that its distribution is peaked rather than normal.

- **Skewness**

  Skewness is used as a means of measuring the degree of asymmetry of a distribution. If the distribution is skewed to the right-hand side, it is said to be positively skewed. With positively skewed distributions, the mean will be to the right of the median which will be to the right of the mode. If distribution is skewed to the left-hand side, it is said to be negatively skewed. With negatively skewed distributions, the mean sits left of the median which sits left of the mode (a mirror opposite to the positively skewed distribution).

  Skewness's calculation is represented by a number which can be positive, negative, or zero. If the skewness is zero, this indicates that the distribution is symmetrical.

- **Checking assumptions**

  Simple/Multiple Linear Regression:

  - The Y and X are linearly related, e.g., do the *x* and *y* values of the model increase/decrease in a linear fashion at the same time.
  - The number of observations must be greater than the amount parameters which are to be estimated.

  Correlation Coefficient:

  - There must be independent observations.
  - The population contains normally distributed bivariate.
  - The p value is presumed to be zero.

Model assumptions can be proven or disproven when plotting the linear model in R. In the analysis and evaluation sections of this report, will be breaking down the output of these plots and what they mean for the linear models.

- **Quantile-Quantile (Q-Q) Plots**
  These are plots used to convey data density using plotting the quantiles of the data. These plots will be a good indicator as to how it is distributed i.e., are they linear or not?

(Chambers, et al., 2018)

(Hayes, 2021)

(Akinkunmi A, 2019)

(Salem Press, 2014)

4.0 Analysis

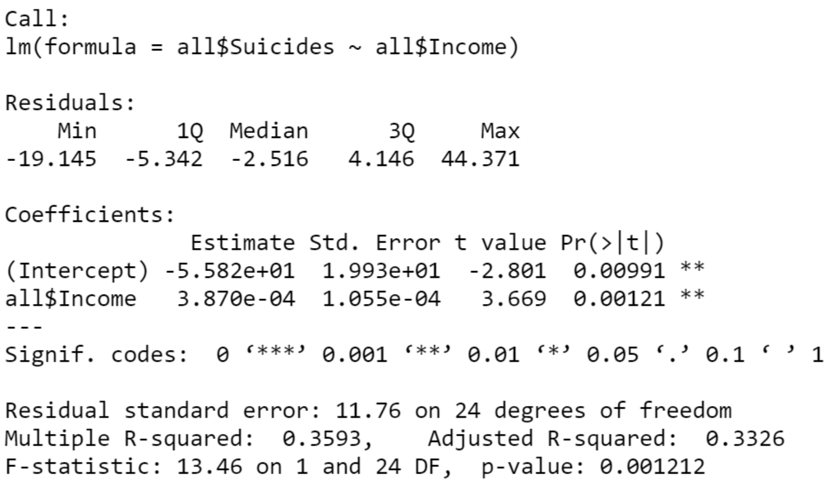
I chose to use the Pearson correlation coefficient method of determining the correlation of my data as it gives me several different parameters which help me to determine the strength of the correlation between two variables being tested such as, t &p values, confidence intervals, and the correlation coefficient figure itself.

I began my analysis by loading my data onto RStudio using the 'read.csv' function. Once the data had been successfully loaded, I began the preprocessing stage of the project. Variables were renamed to simplify the process going forward.

4.1. Regression Analysis

4.1.1 Simple Linear Regression Model

The first test carried out was for linearity of the data distribution. This test was conducted by using the linear model function to create a model containing suicides (the dependant variable) and income. The summary is then printed in R:



```
Call:
lm(formula = all$Suicides ~ all$Income)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.145  -5.342  -2.516   4.146  44.371 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -5.582e+01  1.993e+01  -2.801  0.00991 **
all$Income   3.870e-04  1.055e-04   3.669  0.00121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.76 on 24 degrees of freedom
Multiple R-squared:  0.3593,	Adjusted R-squared:  0.3326 
F-statistic: 13.46 on 1 and 24 DF,  p-value: 0.001212
```

As we can observe by the model summary, the measure residuals do not appear strongly symmetrical, as the figures are uneven either side of the median.

The coefficients are the expected values of our linear model. The intercept represents of the number of suicides expected when considering the average income of all counites in the dataset within the dataset.

The second row of the coefficients represents the slope, or in terms of my project, the effect which income has on suicide. The slope shows that for every increase of 1 suicide, the income level increased by: € 3.87.

The standard error of the coefficient measures the mean amount which the coefficient estimates can vary between the actual mean value of the income variable. The standard error will be used to determine an estimate value of the predicted difference, should the model be run again, i.e., the expected variation of income and be as much as €1.05.

The t-value measures the number of standard deviations the coefficient estimate is from zero. My t-value figure is shown to be 3.669, meaning that because it is not zero, I can reject the null hypothesis that t = 0.

Pr(>|t|) represents the likelihood that a value greater than or equal to *t* will be observed. As can be seen from the figure above, the probability is low at 0.00121.

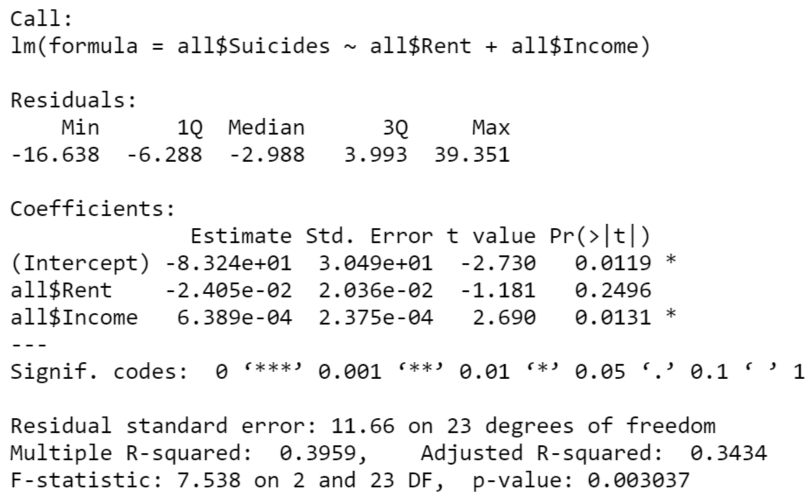
The residual standard error is used as a measurement of how well the linear regression model fits the data. The mean level of suicides is 16.81. As my residual standard error figure is 11.76, the percentage error is 0.69 or 69%.

The multiple r-squared (r2r2) output is used as a means of providing a measurement of how well the model fits the data used. The r2r2 figure outputted in the model summary is 0.3593, or just under 36%.

Finally, the F statistic outputted in the summary, is used as an indicator of the relationship between the predictor and the income variables. The further the outputted F-statistic is from 1, the better suited the data is to the model. The F-statistic given is 13.46, which is larger than 1 based on the size of the data used.

4.1.2. Multiple Linear Regression Model

This regression model will contain an additional variable: "Rent", as a new independent variable. When called upon in R, the multiple linear model can be summarised as seen bellow:



```
Call:
lm(formula = all$Suicides ~ all$Rent + all$Income)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.638  -6.288  -2.988   3.993  39.351 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -8.324e+01  3.049e+01  -2.730   0.0119 *
all$Rent    -2.405e-02  2.036e-02  -1.181   0.2496  
all$Income   6.389e-04  2.375e-04   2.690   0.0131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.66 on 23 degrees of freedom
Multiple R-squared:  0.3959,	Adjusted R-squared:  0.3434 
F-statistic: 7.538 on 2 and 23 DF,  p-value: 0.003037
```

As can be observed from the model summary, just like with the simple linear model, the residuals appear to be asymmetrical since the figures given are uneven on both sides of the median figure.

The coefficient figures show that with multiple linear regression in effect, the slope shows that for every increase of 1 suicide, the income level increased by: €6.38.

The expected variation based on the standard error coefficient is now €2.03 for rent and €2.37 for income.

The null hypothesis for income can be rejected as its t-value is greater than zero. The t-value for rent however, is -1.181, meaning that the sample mean is less than the hypothesised mean.

Probability levels of greater values than the predicted *t* value is low for income at just 0.0131 but significantly higher for rent at 0.2496.

In terms of the residual standard error, it remains virtually the same as the previous model at 11.66, with 0.10 less than when it was simple linear regression.

R2R2 is slightly improved over the previous model. While the overall percentage is still quite low to be determined as a good model fit, it has risen to 0.3959, almost 40%.

The F-statistic for the multiple linear regression model is approximately half of what it was under the simple linear regression model as there are now two variables across 23 degrees of freedom.

4.2 Skewness & Kurtosis

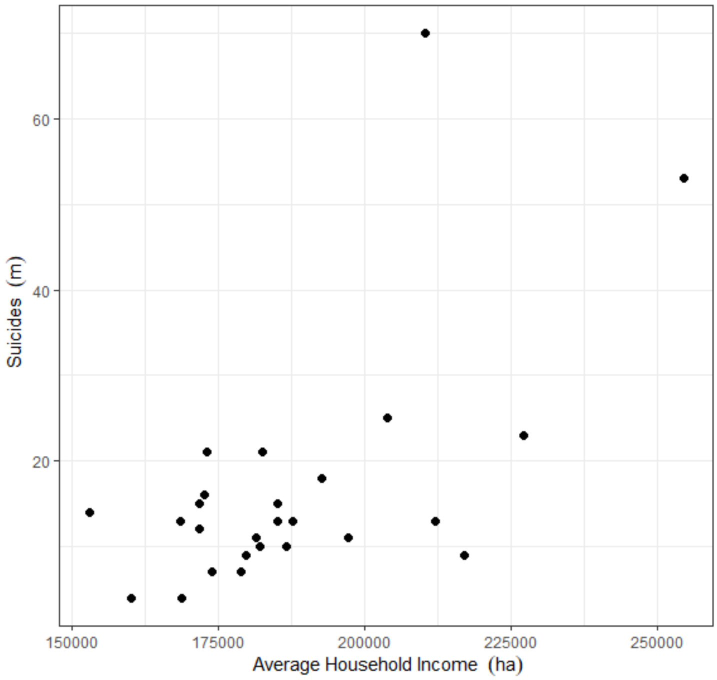
The skewness function in R is simply denoted as “skewness”, and the same can be applied to the kurtosis function too. This stage of the process will look at the skewness and kurtosis of the variables used based on the function output in R.

Variable	Mean	Median	Standard Deviation	Skewness	Kurtosis
Suicides per County	16.81	13	14.3945	2.455525	5.793593
Avg. Household Income	187651	182364	22294.12	1.118698	1.103928
Average Monthly Rent	825	749.5	260.0474	1.803463	3.313499
Population per County	183149	123851	257806.7	3.604246	13.16105

4.4. Correlations

This section of the report will focus on the relationship between the dependant variable and each of the independent variables in order to determine if there is a correlation between them and if so, to what extent they are correlated.

4.4.1. Suicides and Average Household Income per County

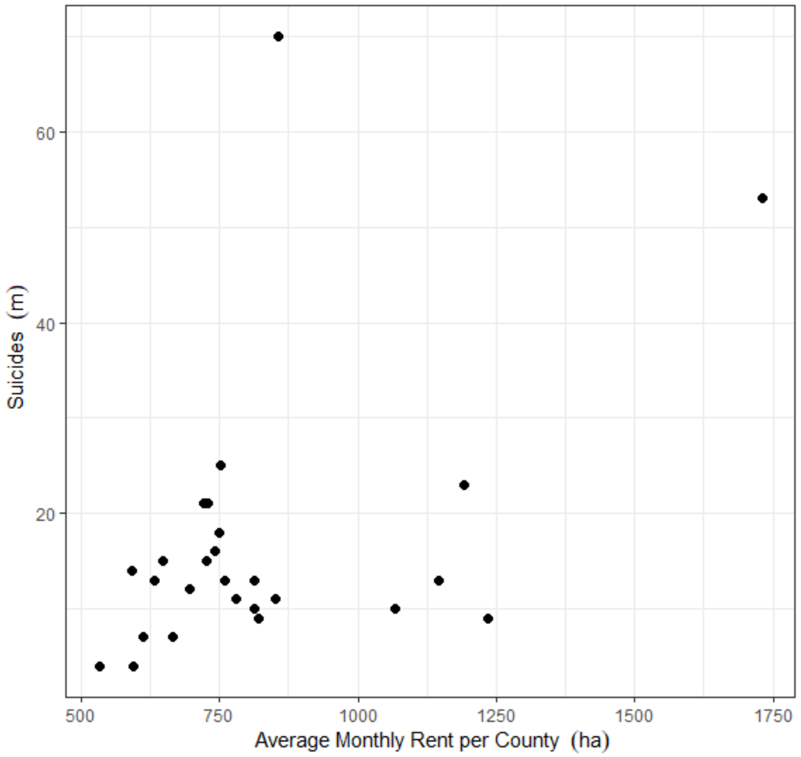


```
        Pearson's product-moment correlation

data:  all$Income and all$Suicide
t = 3.6686, df = 24, p-value = 0.001212
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2761903 0.8008273
sample estimates:
      cor 
0.5994151 
```

When plotting and summarizing the correlation between suicides per county and average household income per county, several observations can be made:

- There are 24 degrees of freedom across the data analysed. This was the case for all correlation tests conducted.
- The p-value of the correlation is 0.001212. As this is smaller than 0.5, it can be presumed that there is a correlation of some significance between the two variables.
- The null hypothesis can also be rejected as the t-value is greater than zero at 3.6686.
- The correlation coefficient is 0.5994151, indicating a moderate uphill (positive) relationship between the two variables.
- This proves the confidence interval to be correct as the correlation coefficient falls between the 2 values (0.2761903 and 0.8008273).

4.4.2. Suicides and Average Monthly Rent per County

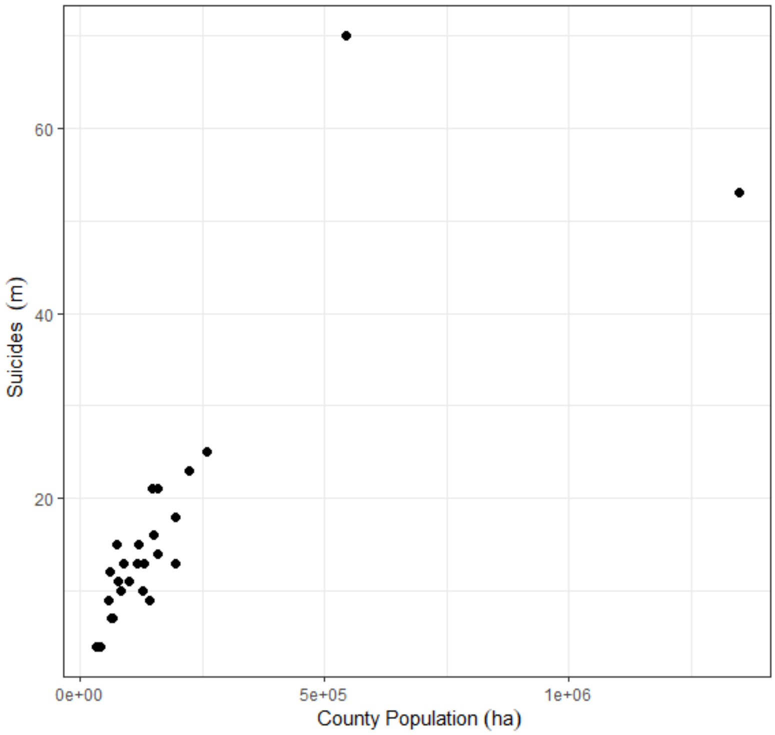


```
        Pearson's product-moment correlation

data:  all$Rent and all$Suicide
t = 2.4949, df = 24, p-value = 0.01988
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08063026 0.71540426
sample estimates:
      cor
0.4538087
```

When observing the plotted data and summarised data for the correlation between suicides per county and average monthly rent rates per county, there were a number of significant findings:

- The p-value of the correlation is 0.01988. As this is less than 0.05, it can be determined that there is some correlation between the two variables being tested.
- The t-value is greater than zero meaning that the null hypothesis can be rejected.
- The correlation coefficient is 0.4538087, indicating that there is a weak uphill (positive) correlation between suicides and rent rates.
- As the correlation coefficient falls between the interval of 0.08063026 and 0.71540426, we can determine that the 95% confidence interval prediction is correct.

4.4.3. Suicides and Population per County

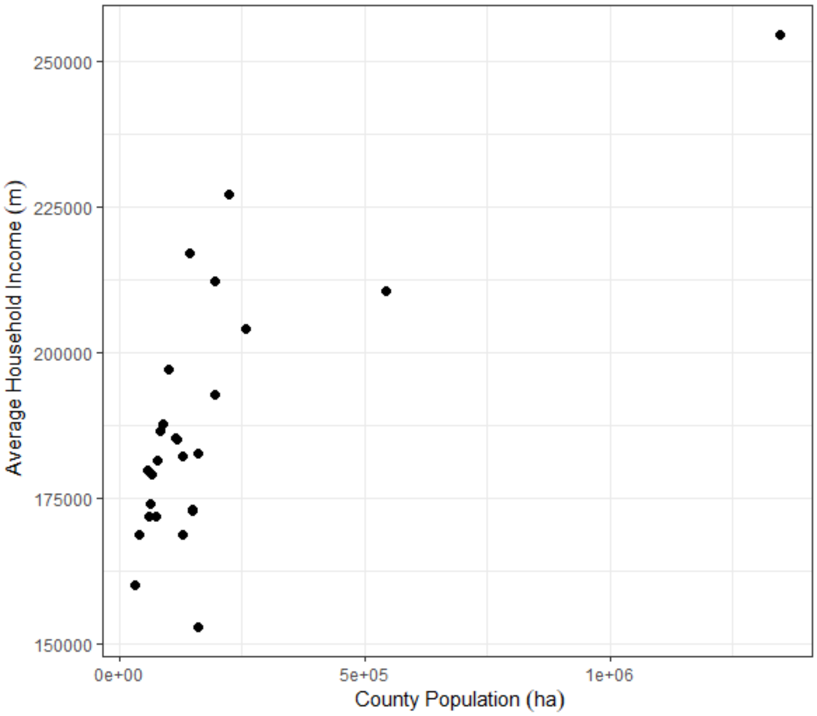


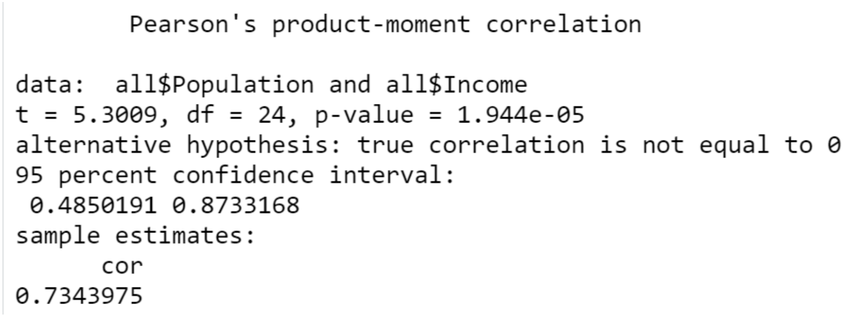
```
        Pearson's product-moment correlation

data:  all$Population and all$Suicide
t = 6.2841, df = 24, p-value = 1.701e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5778324 0.9008222
sample estimates:
      cor
0.7886624
```

The observable output values for the correlation between suicides per county and population per county:

- This correlation test outputs the highest correlation coefficient yet, at 0.7886624, meaning that there is strong (positive) uphill correlation between Suicides and county populations.
- The p-value however is greater than the sum of 0.05 at 1.701, meaning that there is strong evidence that the null hypothesis is true and that we reject the alternative hypothesis.
- The 95% confidence interval is proven to be correct as the correlation coefficient falls between the two predicted values (0.5778324 and 0.9008222).

4.4.4. Average Household Income and Population per County





```
        Pearson's product-moment correlation

data:  all$Population and all$Income
t = 5.3009, df = 24, p-value = 1.944e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4850191 0.8733168
sample estimates:
      cor
0.7343975
```

Similarly, to the previous correlation test, the correlation between average annual household income per county and population per county are quite strongly correlated as the below figures show:

- Whilst the t-value (5.3009) is greater than zero, the p-value (1.944) is greater than 0.05, showing that there is significant evidence that the null hypothesis is true and that we reject the alternative hypothesis that the correlation is not equal to zero.
- The correlation coefficient is 0.7343975, meaning that that there is strong (positive) uphill correlation between average household income and population per county in Ireland.
- The 95% confidence interval is proven to be correct as the correlation coefficient falls between the two predicted values (0.4850191 and 0.8733168).

(Pyle, 2003)

(Carter, 2019)

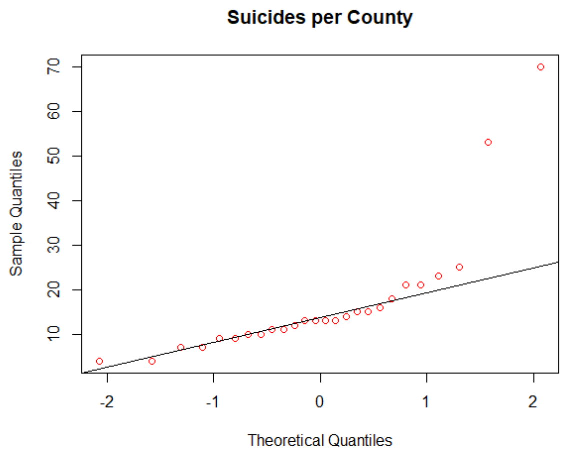
(Grabowski, 2016)

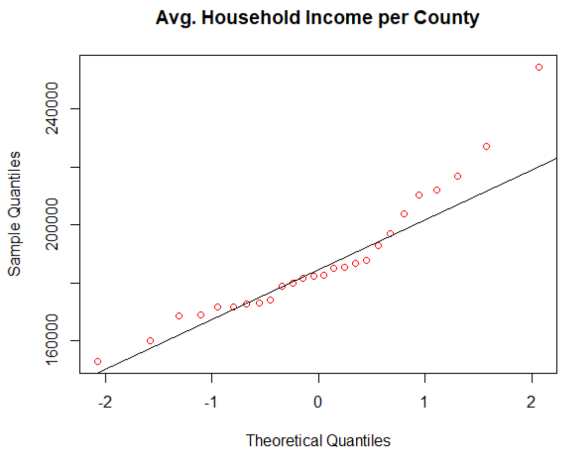
(Rego, 2015)

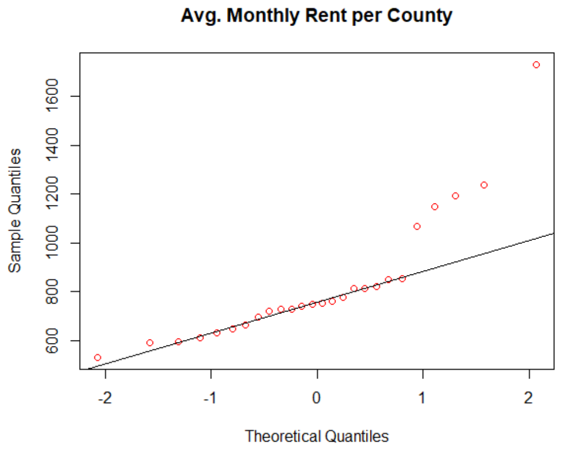
5.0 Evaluation & Results

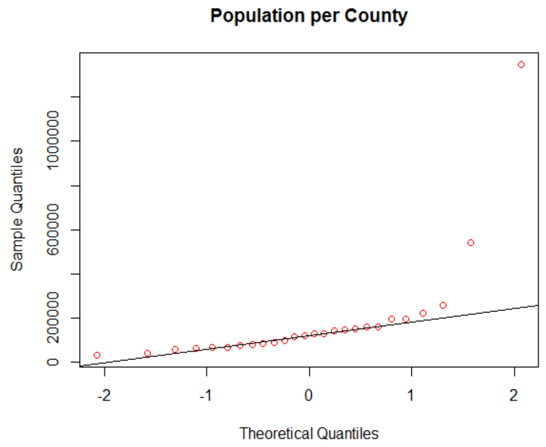
5.1. Linear Regression Models/Assumptions and Linear Distribution

Below are Q-Q plots representing the distribution of the data for each of the project variables:









From these plots, I can be interpreted that the data, for the most part, is linearly distributed, as most points fall in and around the line. Suicides and rent however, deviates from the linear distribution significantly however as a result of outliers in their data. Population also has one very significant outlier in Dublin as it as a far greater population than any other county.

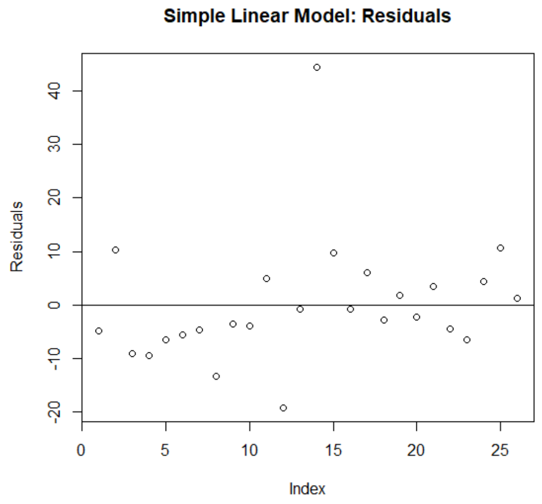
(Chambers, et al., 2018)

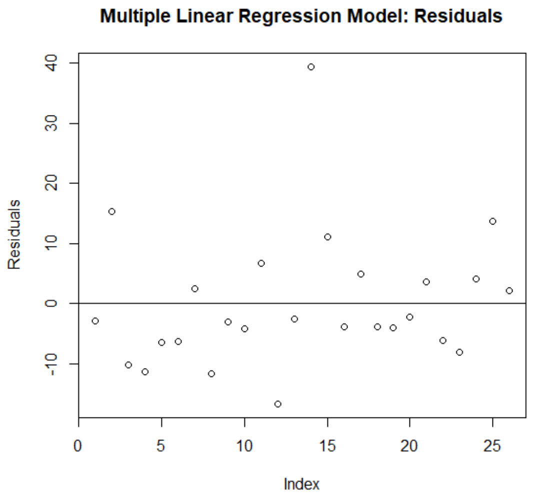
Assumptions

- We can determine that X an Y are linearly related as they both increase at the same time.
- We know that there are more observations than parameters with our given data, meaning that this assumption is proven to be true.

5.2. Linear Regression Models

Below are the scatter plots for the residuals for both the simple linear model and the multiple regression model. They are used to determine if the residuals are consistent with random error. Observations can be made upon viewing these scatter plots:





- Both plots are virtually identical in that they are randomly scattered across the zero line.
- This means that the residual standard error is random and does not follow a pattern indicative of another potential error.
- This indicates that both regression models' predictions are correct on average as opposed to being systematically too high or too low throughout.

(Frost, 2019)

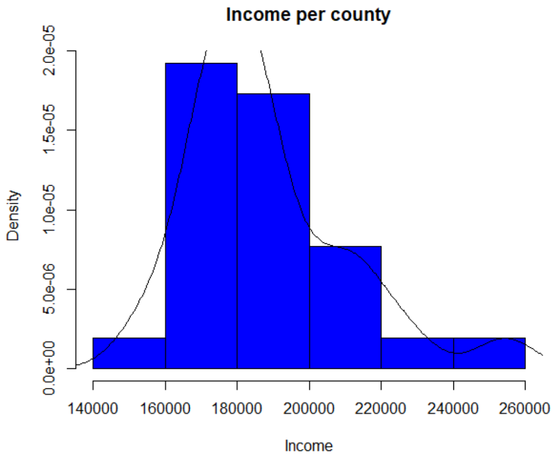
5.3. Skewness & Kurtosis

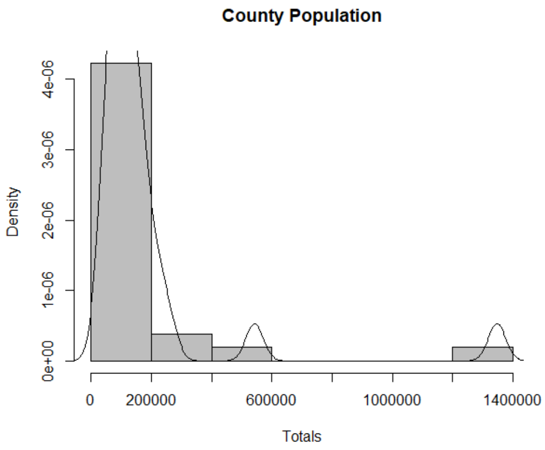
Variable	Mean	Median	Standard Deviation	Skewness	Kurtosis
Suicides per County	16.81	13	14.3945	2.455525	5.793593
Avg. Household Income	187651	182364	22294.12	1.118698	1.103928
Average Monthly Rent	825	749.5	260.0474	1.803463	3.313499
Population per County	183149	123851	257806.7	3.604246	13.16105

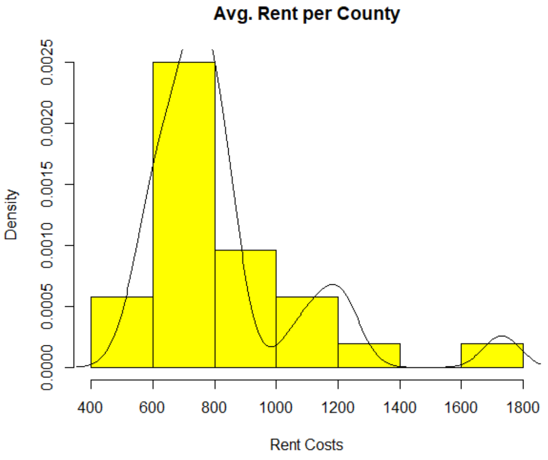
5.3.1. Skewness

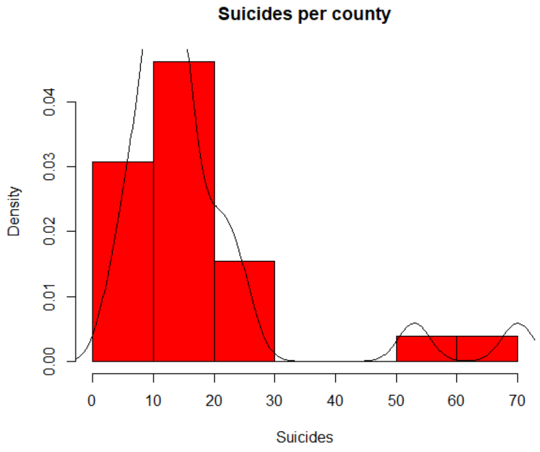
The four histograms below along with the table provided in the analysis indicate several the following:

- The skewness of all 4 variables is positively skewed as the means sit right of the medians. As all four skewness figures are also positive and not zero, it can be assumed that the distribution is asymmetric, it can be determined that the distribution is not symmetrically smooth.









- The asymmetry of the distribution for each variable can be observed in the above histograms. Significant outliers in the data cause a drastic change in the way in which the distribution is observed by the algorithm.

(Yau, 2020)

5.3.2.Kurtosis

All kurtosis values for the variables are positive, meaning that they are leptokurtic. As the figure for a normal kurtosis is 3, the kurtosis for average household income is a decreased kurtosis meaning that it has a wider peak and heavier tails as a result. We can derive that the income variable is more broadly distributed as a result.

The population variable has the highest kurtosis figure at 13.16105, meaning that its got very light tails in terms of its distribution and its peak is a narrow one. The same can be said for the kurtoses of Suicides and average rent prices as both kurtosis values are greater than 3.

(Yau, 2020)

5.4. Correlation Coefficients

The focus of this project is finding correlations between certain socioeconomic factors such as income, rent, and population density, and the number of suicides per county in Ireland (2016). Upon analysing the data via the Pearson Correlation Coefficient function in R, I observed several significant findings:

- All confidence intervals accurately predicted that the correlation coefficient figures for each respective test fell between the specified intervals.
- There is a moderate and uphill correlation between suicide and average household income at just under 60%. This is a significant finding as it shows that **as income increases, so too does the rate of suicide**. This goes against my initial belief going into this project that a lower income is an influencing factor suicide rates nationwide.
- While there is an uphill correlation between average monthly rent and suicide, it's a lot weaker at 45%, making it a less influential factor than average household income.
- The greatest correlation is between suicides per county and population per county. This comes as no surprise counties with greater populations are more likely to have higher rates of suicide. The correlation between these two variables was a strong uphill one, with a correlation of almost 79%. However, because the p-value is greater than the sum of 0.05 at 1.701, meaning that there is strong evidence that the null hypothesis is true and that we reject the alternative hypothesis.

Assumptions:

- In the case of the tests carried out, it can be confirmed that the null hypothesis ai assumed to be true until proven otherwise.
- The assumption that the population contains a normally distributed bivariate can be confirmed as true as the normal distribution for each variable has being confirmed through the use of the Q-Q plots above.
- Finally, the assumption that our data must contain independent observations is correct as the dependant variable differs from the three independent variables.

6.0 Conclusions

Upon completion of the evaluation stage of my project, I was able to identify certain strengths and weakness which my project had as well as the advantages and disadvantages of applying statistical tests and methodologies to data of this nature.

6.1. Advantage and Strengths

There are several advantages and strengths which can be attributed to carrying out a project of this nature:

- The CSO provides an abundance of publicly accessible records, making finding information in these areas quite likely.
- Having carried out tests such as these in other modules like Advanced Business Data Analysis and Data Mining, I had the necessary coding resources gathered in order to make them applicable to my chosen data.
- The R programming language is quite direct in terms of its application. Knowledge gained in one area can potentially lend itself to another.

6.2. Disadvantages and Limitations

Despite having a number of strengths and advantages, this project also struggled with disadvantages and limitations too:

- Carrying out time series methods was not a realistic endeavour as I was unable to access day-by-day records of variables such as income and rent, which ended up being plotted as annual figures.
- Going into this project, I had planned to output my findings through a web output format which ultimately was not possible due to time constraints.
- Some of the tests conducted were designed to be carried out on data from a business standpoint, so interpreting the output in a way which could be translated to my data proved to be challenging.

7.0 Further Development or Research

With additional time and resources, there are several potential directions which this project could potentially take:

- If I was able to acquire access to daily records of the independent variables used (rent, population, income), they could be converted into time series data. In doing this there I the potential to create a forecasting model allowing the project to determine future trends based on forecasted variables. Models such as the ARIMA (auto-regressive integrated moving average) or prophet would be ideal candidates as they are both used for forecasting time-series data.
- Implementing my concept into a web application format through using R Shiny. This would allow for a more interactive output and could be combined with the prophet model forecast plots to make acquiring information a smoother process.

(Akinkunmi A, 2019)

References

Aiguo, Z. & Jiang Lanling, S. P., 2011. *Application of Data Mining in Supermarkets,* ShenYang: IEEE Xplore.

Akinkunmi A, M., 2019. *Business Statistics with Solutions in R.* 1st ed. Yola: Walter de Gruyter GmbH.

Azevedo, A. & Santos, M. F., 2008. *KDD, semma and CRISP-DM: A parallel overview,* Porto: Azevedo, Ana.

Carter, D. J., 2019. *3.5 Interpretation of Confidence Intervals.* [Online]
Available at: https://bookdown.org/danieljcarter/r4steph/
[Accessed 14 May 2021].

Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A., 2018. *Graphical Methods for Data Analysis.* 2nd ed. Boca Raton: CRC Press - Taylor & Francis Group.

CSO, 2016. *Census 2016 Small Area Population Statistics.* [Online]
Available at:
https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/
[Accessed 15 March 2021].

Frost, J., 2019. *Check Your Residual Plots to Ensure Trustworthy Regression Results!.* [Online]
Available at: https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/
[Accessed 16 May 2021].

Grabowski, B., 2016. “P < 0.05” Might Not Mean What You Think: American Statistical Association Clarifies P Values. *JNCI: Journal of the National Cancer Institute,* VIII(8).

Hayes, A., 2021. *Confidence Interval.* [Online]
Available at: investopedia.com/terms/c/confidenceinterval.asp
[Accessed 12 May 2021].

Hayes, A., 2021. *Linear Relationship Definition.* [Online]
Available at: https://www.investopedia.com/terms/l/linearrelationship.asp
[Accessed 10 May 2021].

McCue, C., 2015. *Data Mining and Predictive Analysis.* 2nd ed. s.l.:Butterworth-Heinemann.

Pyle, D., 2003. Error and Confidence . In: L. Homet, H. Severson, C. Derman & M. Peterson, eds. *Business Modeling and Data Mining.* San Francisco : Morgan Kaufmann Publishers, pp. 68-70.

Rego, F., 2015. *QUICK GUIDE: INTERPRETING SIMPLE LINEAR MODEL OUTPUT IN R.* [Online]
Available at: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
[Accessed 10 May 2021].

Roiger, R. J., 2017. *Data Mining: A Tutorial-Based Primer.* 2nd ed. Boca Raton: Taylor & Francis Group.

Salem Press, 2014. *Business Reference Guide: Forecasting & Statistical Analysis.* eBook ed. Ipswich, Massachusetts: Salem Press.

Skrubbletrang, G. & Vachon, F., 2012. *Knowledge Discovery in Databases for Competitive Advantage,* St.Catherines: ResearchGate.

Weston, C., 2017. *Revealed: The average rent in each county in Ireland.* [Online]
Available at: https://www.independent.ie/business/personal-finance/property-mortgages/revealed-the-average-rent-in-each-county-in-ireland-36317785.html
[Accessed 6 May 2021].

Yau, C., 2020. *R Tutorial With Bayesian Statistics Using Stan.* ebook ed. Cupertino, California: Chi Yau.

[This page is left intentionally blank]

Appendix

National College of Ireland

Project Proposal

07/11/2020

BSHTM4

Data Analytics

Academic Year: 2020/2021

Owen Thompson

X17516666

X17516666@student.ncirl.ie

Contents

8.0 Objectives

The objectives of this project will be to identify the correlation between different counties' suicide rates and their wealth.

It will be to find a correlation between levels of wealth, social status and the suicide rates in their respective counties.

I will then draw conclusions based on that data and record it them in the final documentation of my project.

9.0 Background

I will be gathering data from a number of sources in order carry out the necessary manipulations and queries later on in my project.

CSO: The Irish central statistics office provides suicide statistics as well as Irish regional income statistics.

This data will be paramount when conducting my analysis.

Due to my lack of experience in data analytics and because I've only started studying it recently, it is likely that I will make changes here and there throughout the year and will be sure to keep my supervisor up to speed with whatever changes I make.

It is possible that the scope of my project may change in the future depending on my findings, and the development of my knowledge of data programming improves.

I will also be aiming to implement as may aspect of my existing and upcoming modules in semester 2, into this project.

Much of the practical aspects of my project will take place during semester 2 once I have completed my programming for big data module in semester 1 and have a solid foundation to build my project on.

The next couple of months will be spent, researching, learning and drawing up an educated concept of how my project will take shape in semester 2.

This is shown in the project plan and Gannt chart under the heading below.

10.0 Technical Approach

Since my project is centred around taking a data analytics approach, and I've only started doing data analytics in 4th year, my technical approach is likely to vary as the year progresses.

Because of my lack of prior experience in this field, I will be taking on as much guidance as possible from my supervisor as he is highly experienced in this area.

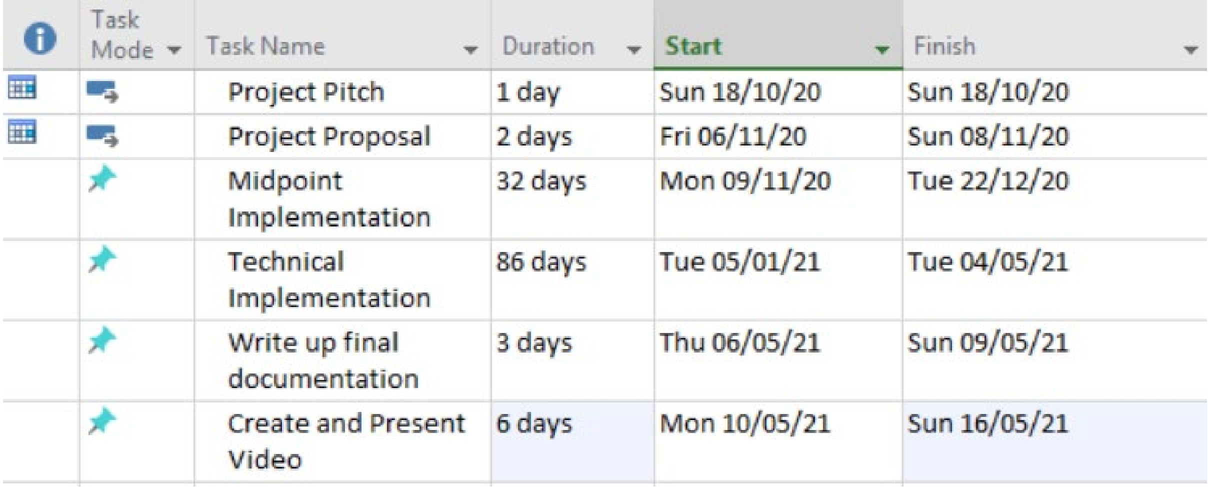
I will be doing exploratory data analytics in order to ensure that the datasets I'm using for my project are clean before implementing them in my project.

I will be taking data and manipulating it in R and then presenting my findings and analysis on a dashboard software, likely to be Tableau as it has an interactive maps function. Another possible dashboard software to consider is R shiny, which is being discussed in my data programming module.

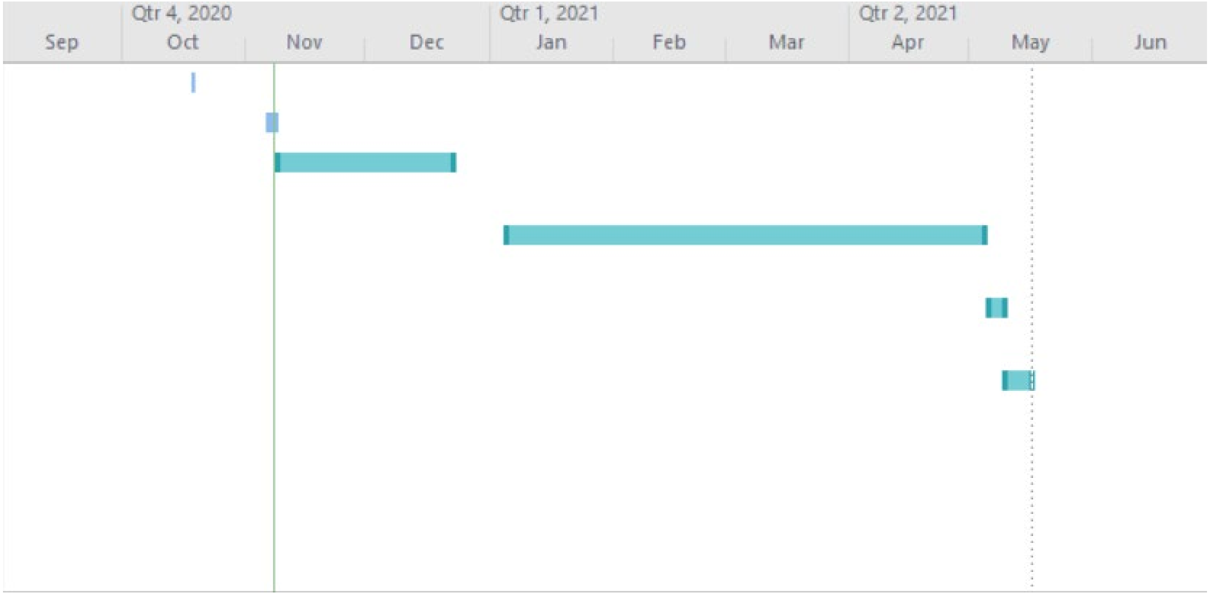
11.0 Special Resources Required

N/A

12.0 Project Plan



ℹ	Task Mode	Task Name	Duration	Start	Finish
▦	➔	Project Pitch	1 day	Sun 18/10/20	Sun 18/10/20
▦	➔	Project Proposal	2 days	Fri 06/11/20	Sun 08/11/20
	📌	Midpoint Implementation	32 days	Mon 09/11/20	Tue 22/12/20
	📌	Technical Implementation	86 days	Tue 05/01/21	Tue 04/05/21
	📌	Write up final documentation	3 days	Thu 06/05/21	Sun 09/05/21
	📌	Create and Present Video	6 days	Mon 10/05/21	Sun 16/05/21



13.0 Technical Details

‘R’ will be the language which I will use to manipulate datasets.

I will be using a number of different libraries in R.

Programming will take place on R Studio IDE.

Findings will be displayed on a dashboard, possibly an interactive map, using R Shiny or Tableau

14.0 Evaluation

The system will be presented to the end user via a dashboard displaying the various metrics either on R Shiny or Tableau dashboards.

With regards to ethics, due to the fact that I will not be using names and the information is public domain. There will be no issues relating to the ethics of this project.