National College of Ireland

Bachelor in Computing

Data Analytics

2021/2022

Douglas Masotti

18151493

x18151493@student.ncirl.ie

Analysis of the Irish Language In Ireland

Contents

Executive Summary

The key idea of this project is to analyse the Irish (Gaelic) as an endangered language, or as it is also said, a dying language. Some argue that it has died due to the globalization and English has become an official language in several nations which had the standard to be called as Lingua Francia. Thus, in order to analyse the current situation of the Irish language, to determine the frequency of people speaking Irish language by age, gender, city, province, and to predict the future of the Irish language in the upcoming years, this analysis was conducted with basis on a dataset of Irish speakers publicly available. To further comprehend the dataset, various data mining approaches were performed. For some parameters, the null/alternative hypothesis was tested. Following the study, it should be simpler to develop new laws and regulations to keep the Irish language alive.

1.0 Introduction

1.1. Background

The Irish language is a Celtic dialect just like English is a German language, and it's one of the branches of the Celtic linguistic group.

> *"The term "Gaelic" in English is derived from Gaeilge, the Irish term for the language itself. Nevertheless, while using English, the Irish linguistic is commonly called as "Irish," not "Gaelic"."* [1].

Irish is presumed to be a dead language. Irish is a native language of Ireland, and until the late 1800s, it was the primary language of the population. Primarily through the national census performed to Irish and non-Irish speakers, based on the Ireland population, has presented that only 4.2 percent of Irish speakers use it as their native tongue, while 10.5 percent use it on a regular or weekly basis.

> *"There's a big discussion going on about whether it is dead or not. Some people argue that Irish Gaelic might be a dead language. Outside of Ireland's Gaeltacht areas, hardly anyone uses it. Furthermore, people who have learned it in a Gaelscoil do not use it outside of school. Unfortunately, those who study it are eager, while those who know it and can carry on a conversation in Irish are less likely to do so. But on the other hand, some argue that it is not dead in the sense that you understand it. It is truly dead in the sense that it is no longer used by the community, except for Gaeltacht places. English signs were used instead of Irish signs in shops, restaurants, cafes, and other institutions"* [2].

This report discusses the condition of Irish Gaelic. The main objectives of our analysis are to find out the frequency of Irish speakers by age, gender, city, province, and to predict the number of Irish and non-Irish speaker for the coming year.

According to the Atlas of World´s Languages in Danger [3], the Irish linguistics is clearly vulnerable since there are just a few families in Ireland where the tongue is passed down from parents to offspring. There are, obviously, some few outliers, such as authentic Gaeltacht communities and pure Irish-language houses where members of the family communicate in

Irish language. The assistance of these kind of homes, where the tongue is more than just a sign, is critical to its resuscitation. The overall growth in Irish tongue speaking homes, as well as the desire to form these homes, may inhibit a reduction in Irish speaking people and, to some degree, overturn the linguistic transition [4].

According to Mufwene [5], globalism has forced stronger languages like English to act as an official language. As a result, indigenous regional dialects like Irish Gaelic have suffered [5]

Social attitudes have shifted from unfavourable to favourable. In regards of the people's first ever domestic and legal linguistic revival, public perception varies greatly. Even though many concerns are raised about the survival of the Irish language [6] [7] [8], restoration initiatives can be effective in the long run and be beneficial for building cultural and national identification. Nevertheless, this can only occur if an increasing number of individuals practice Irish on a regular schedule and transfer it on to their kids.

According to Gibbs [9], a recurring element when a tongue is vulnerable is indeed the society's common concerns about the language's relevance. Likewise, people may believe that their native tongue is weaker than the dominant tongue. When this happens, people seem to discontinue speaking the minority tongue in all circumstances. This mindset is transferred on to future generations. Irish Gaelic is the Republic of Ireland's 1st official language; however, this designation was only granted in order to rebuild the dwindling dialect [10].

Irish Gaelic is an excellent instance of a nation trying to change the flow of linguistic transfer by making it legal. Regardless of the fact that the tongue has never been used in national assembly, the authorities sought to move from English to Irish via the National School system [11]. As per Grenoble and Whaley [12], education has an influence on macro-variables. Education is applied at a macro-level in several situations, like as Irish Gaelic, although it may yet crash due to some other issues [12].

1.2. Aims

This report discusses the condition of Irish Gaelic. The main objectives of our analysis are to find out the frequency of Irish speakers by age, gender, city, province, and to predict the number of Irish and non-Irish speaker for the coming years.

This paper will base on a prediction and analysis for determining that whether Irish language is dying or not. So, for this purpose we carried out our analysis. At first, the datasets regarding the statistics of Irish people by age, gender, city, and province were to be downloaded from an authentic source, which collected the data primarily from the national census.

1.3. Technology

To perform this study, I decided to use R language as the technology to handle the dataset data, based on the ability that this languages can provide us with to wrangle data, and easily retrieve cross validation analysis and data visualisation plots. It will be rendered using the official RStudio application.

After acquiring the datasets, data mining techniques were used to allow extract the relevant and necessary data from the uncleaned dataset, and create a more organized and workable data frames which can be modelled and implemented to analyse the dataset in order to get a better insight of the analysis to know what the data actually holds. This is very helpful for data analysis and prediction analysis.

After the data mining algorithms, I performed regression modelling as it is a widely used statistical tool that establishes a relationship model among two variables. Among which one is the dependent variable and the other is the independent variable.

To perform all actions on R, I've used the support of R's library as:

- tidyverse, assists in the creation of tidy data, cleaning and tidying actions have the effort reduced. Essentially, tidy data refers to datasets in which each cell represents a single value, each row represents an observation, and each column represents a variable.
- ggplot2, is a popular data visualisation toolkit that is based on the "Grammar of Graphics.". Making it able to create graphs with one variable, two variables, and three variables, as well as category and numerical data. Grouping can also be done using symbols, sizes, colours, and other factors.
- dplyr, is a software that is used to manipulate data by restricting the dataset possibilities and increasing the usability of the dataset.
- scales, is the package that includes ggplot2's core scaling architecture, as well as tools to override default breaks, labels, transformations, and palettes.
- rpart, is a package with a strong machine learning tool for creating classification and regression trees. It also includes a recursive partitioning and psych functionalities.
- psych, is a toolkit for personality, psychometric theory, and experimental psychology that can be applied in a variety of situations. Even with some functions returning basic descriptive statistics, they are mostly used for multivariate analysis and scale design utilising factor analysis, principal component analysis (PCA), cluster analysis, and reliability analysis.

1.4. Structure

The dataset that will be used on this analysis is stored in Kaggle website in a 5-separate-file CSV format and required to be merged and cleaned, as detailed in chapter 2 of this study.

2.0 Data

The datasets used in this research are widely available in Kaggle website and contain license of authenticity and free public license by Creative Commons.

Irish Speakers dataset: *"This data set contains statistics on the number of Irish speakers and Non-Irish speakers gathered by the central statistics office(CSO), a government agency responsible for gathering statistics on the Irish population, primarily through the national census.*

Their database can be found here: https://www.cso.ie/en/databases/

Most of the data is only available as of 2011 since the national census only began a detailed survey on Irish speaking ability as of 2011.

The only exception is number of Irish speakers broken down by Province, which contains data dating back to 1861." [13]

Licensed by: CC0: Public Domain

The dataset [13] about statistics of Irish speakers is divided by age, gender, city, and province. The dataset consisted of the statistics regarding the number of Irish and non-Irish speakers collected by the Central Statistics Office (CSO) which is an agency of government having a responsibility of collecting data regarding the population of Irish people, mainly via national consensus.

3.0 Methodology

The analysis in the datasets will be performed during the second semester of this college degree. The idea is to start reducing the data by the range 2016 to 2018 and the same areas (by longitude and latitude) to ensure that both datasets contain useful data and guarantee that is possible to compare one against the other.

After this point, I will clean the data and remove all outliers and empty or misleading data. Then, I start to analyse the correlation between both datasets and start researching for common patterns in the data.

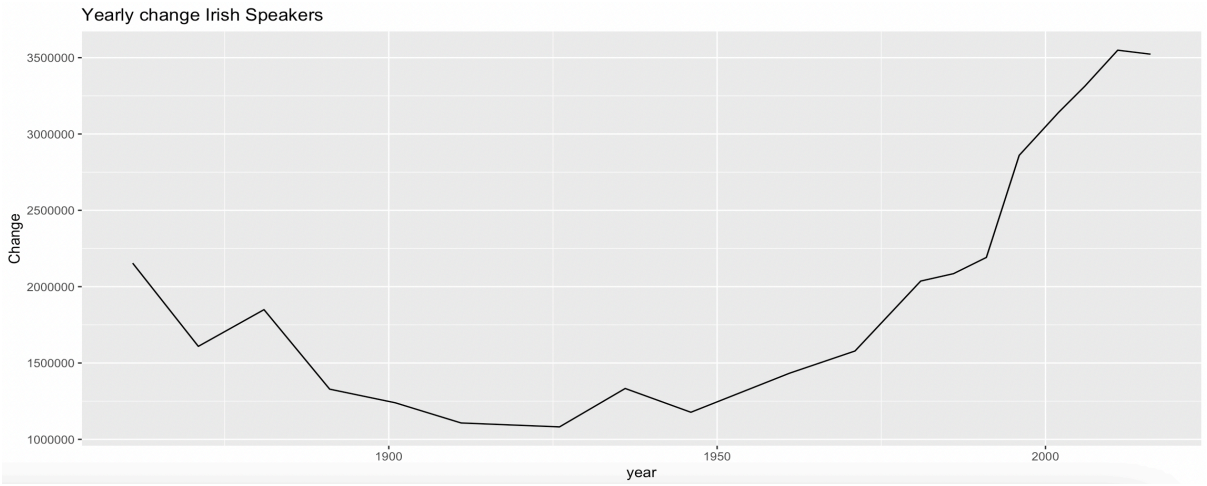
4.0 Analysis

The CSV's that structure the dataset (source: [13]) about the statistics of Irish speakers consist of the number of Irish and non-Irish speakers collected by the Central Statistics Office (CSO) which is an agency of government that has the responsibility of collecting data regarding the population of Irish people, mainly via national consensus.

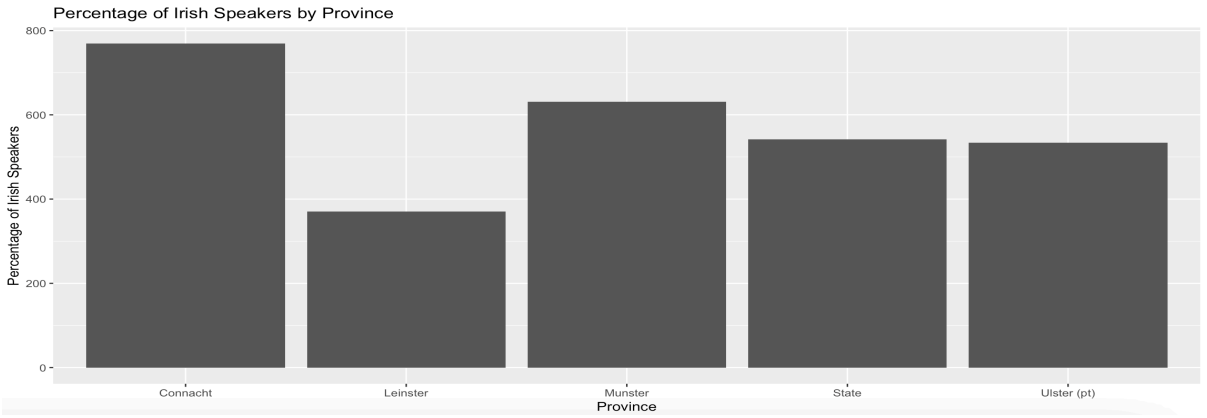
After downloading the CSV's that structure this dataset, the next step was to import the required libraries to perform this analysis and then start cleaning the data to ensure that empty and misleading information would affect the results.

Data Analysis.

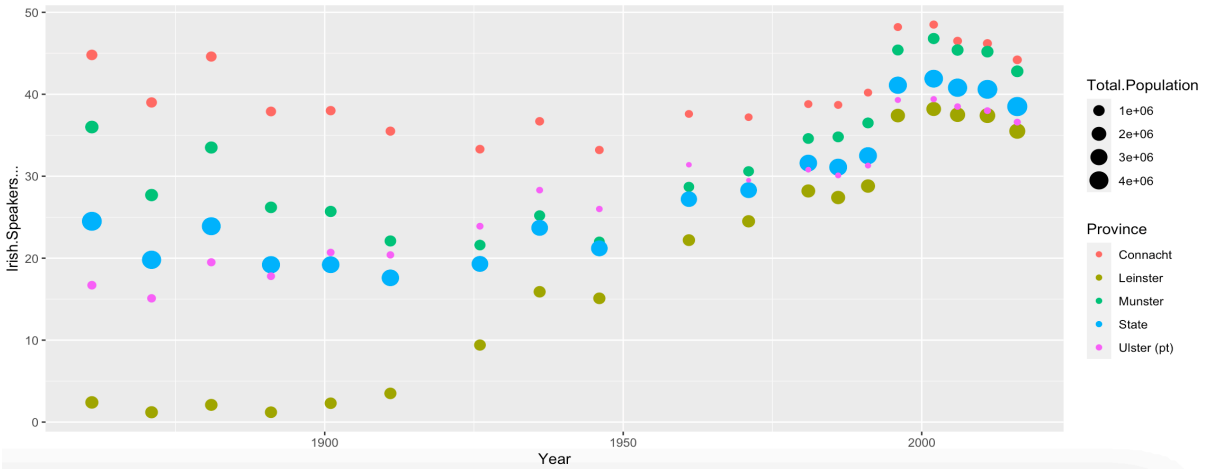
In this study, various relationships and pattern were found. Foremost, I used a line graph to shows the linear relationship from year 1860 to 2016 which clearly demonstrate to us that the number of Irish speaker are decreasing and doesn't have linear trend from 1860 to 1900, however, it rapidly and substantially increases in the following years.



In the next step, I used bar graph to display the relationship between the different provinces and the number of speakers residing in each province of Ireland. The highest number of Irish speaker was noted on Connacht, while the lowest number was in Leinster.

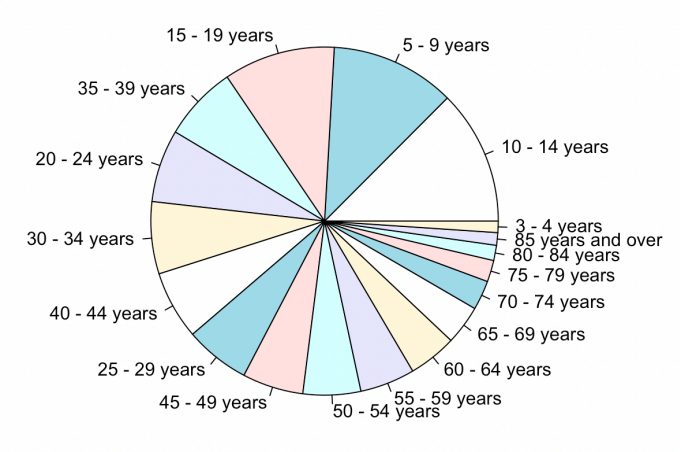


In the next step, I decided to cross validate three variables which are the province, the total percentage of Irish speakers and years, plotting its result across the years on base of axis x, and resizing each dot from scatterplot by its total population.



We can visualise, that the total population for each province is not significantly changed, however the amount of people learning in each province has increased, as we can significantly outline Leinster that moves up from barely 0 to over 35%. Although, all provinces are still leaded by Connacht that has not changed much since the first entries of this dataset on 1860, maintaining its percentage around 45% of the population.

In the next step, I decided to plot on a pie chart to display how each age group is divided for Irish speakers, and how each group impacts when compared against the other age groups and the overall of all Irish speakers.




```
# A tibble: 18 × 3
   Age.Group            n    sum
   <chr>            <int>  <int>
 1 10 - 14 years       19 474418
 2 5 - 9 years         20 439782
 3 15 - 19 years       19 393753
 4 35 - 39 years       18 264076
 5 20 - 24 years       18 256085
 6 30 - 34 years       18 254233
 7 40 - 44 years       17 241222
 8 25 - 29 years       19 229638
 9 45 - 49 years       19 215222
10 50 - 54 years       19 203622
```

It response clearly shows us that teenagers (10-14 years old) have the higher percentage to be an Irish speaker, followed by 5-9 years old and 15-19 years old. Which can lead us that the teaching of Irish language has a decent rate of delivery in schools, although it's not present in the following years, contributing to lose the knowledge by not having practice.

Data Mining and Prediction.

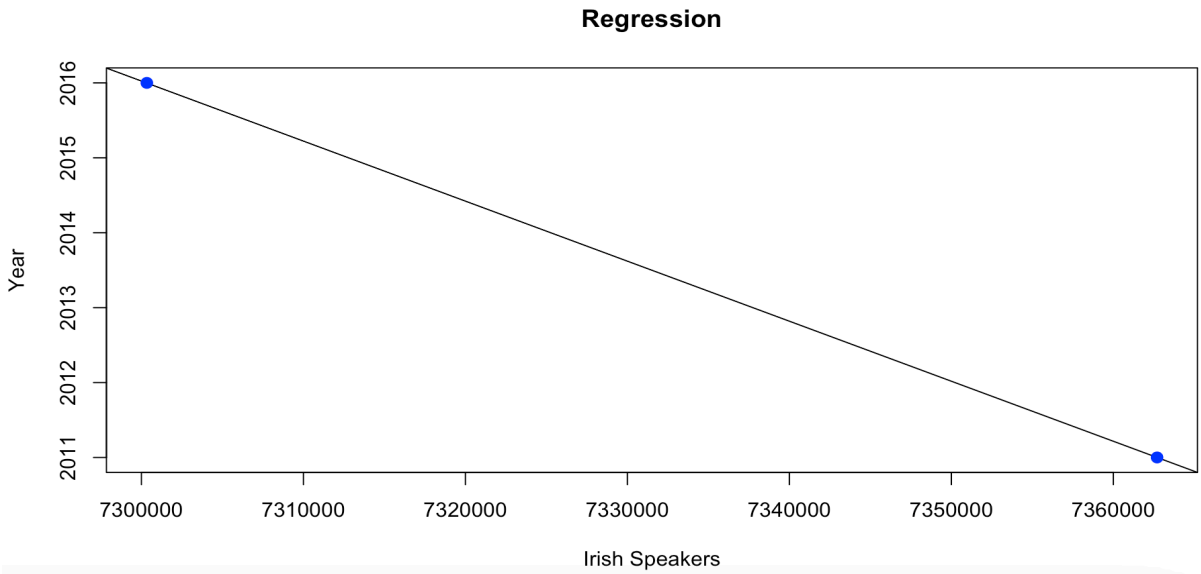
The next step was to perform the data mining techniques on the downloaded and cleaned data. These actions, and the data visualisation plots were also done in R.

Initially, is important to highlight that regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

For this purpose, I decided to use regression modelling to try to predict the future of Irish language based on data available in the CSV's structure of the dataset.

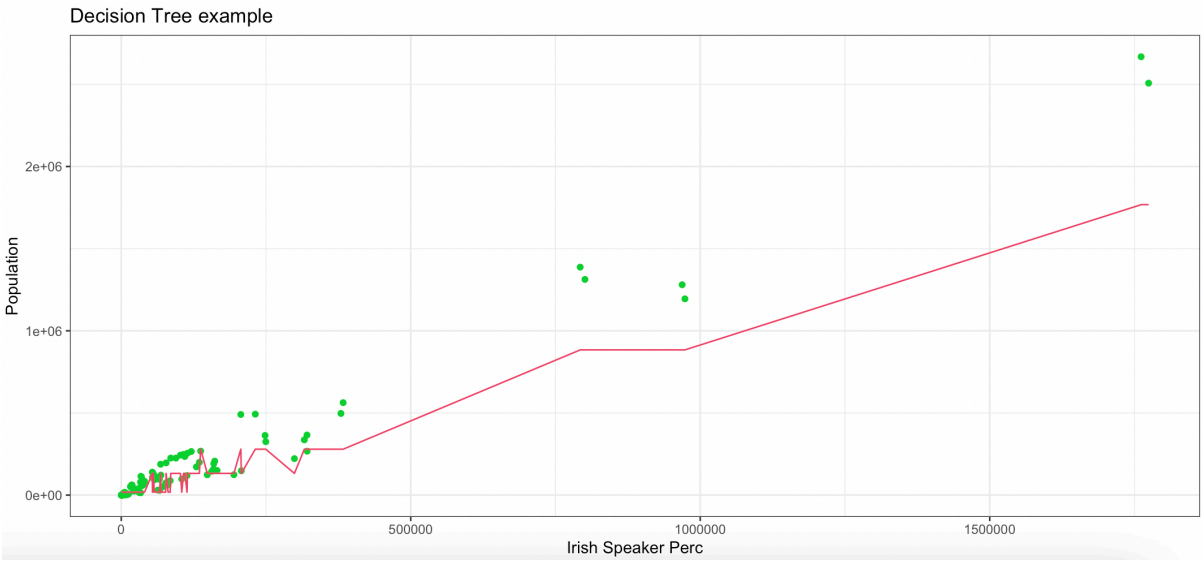
A linear regression model was established in for the year 2021, Irish speakers was predicted and plotted, and the response shows us a modest negative relationship between

variables meaning by the increase of one, the decrease of the other, so with the increase of the years the number of Irish speakers decreases significantly.



Then, I move to the prediction part of this study, starting by applying the techniques of Decision Tree, where only the two columns were kept and stored in a new data frame *df*.

The predictive regression algorithm was applied to the data frame with the Irish speakers and Non-Irish Speakers parameters and the min split parameter of 3. This has indicated to us that the number of nodes that need to be in a node before it can be split. The algorithm then tris to find the relationships between the two variables and creates a model for the data. So, a new and separate data frame is generated with only a height column that will be used for testing. At the end a line graph was plotted, and the predicted values included within the trained data set's range.



This graph is returning us the relationship between the Irish speakers and the predicted values which is received from decision tree while the green points are showing relationship between Irish speakers and non-Irish speakers. Which lead us that the predicted

values are falling within the trained data set's range meaning the line is initially intersected with the points.

After this analysis, the third step was to perform correlation analysis, which has returned the covariance between the variables given as:

	Irish_Speakers	Irish.Speakers...	NonIrish_Speakers	Not.Stated	Population
Irish_Speakers	65032451611	-1108924.8815	93538973513	4097982096.2	162669407220
Irish.Speakers...	-1108925	298.5238	-2010272	-112143.5	-3231340
NonIrish_Speakers	93538973513	-2010271.9938	138316175283	6366185498.7	238221334295
Not.Stated	4097982096	-112143.5269	6366185499	396933158.7	10861100754
Population	162669407220	-3231340.4022	238221334295	10861100753.6	411751842269

While the correlation is given as:

	Irish_Speakers	Irish.Speakers...	NonIrish_Speakers	Not.Stated	Population
Irish_Speakers	1.0000000	-0.2516794	0.9862592	0.8065779	0.9940839
Irish.Speakers...	-0.2516794	1.0000000	-0.3128448	-0.3257812	-0.2914577
NonIrish_Speakers	0.9862592	-0.3128448	1.0000000	0.8591799	0.9982203
Not.Stated	0.8065779	-0.3257812	0.8591799	1.0000000	0.8495665
Population	0.9940839	-0.2914577	0.9982203	0.8495665	1.0000000

The negative values shows us an inverse correlation, which if one variable increases the other will decrease. Moreover, the positive correlation shows us the contrary, if one variable is increasing, the other is also increasing. While zero shows no correlation between the two or more variables.

That defined, I decide to use the Spearman method to calculate the correlation. The Spearman correlation does not require continuous-level data (interval or ratio), because it uses ranks instead of assumptions about the distributions of the two variables. This allows us to analyse the association between variables of ordinal measurement levels. Moreover, the Spearman correlation does not assume that the variables are normally distributed, and are given as:

	Irish_Speakers	Irish.Speakers...	NonIrish_Speakers	Not.Stated	Population
Irish_Speakers	1.0000000	-0.5769447	0.9657061	0.9628981	0.9852252
Irish.Speakers...	-0.5769447	1.0000000	-0.7431043	-0.7107220	-0.6859682
NonIrish_Speakers	0.9657061	-0.7431043	1.0000000	0.9838736	0.9933878
Not.Stated	0.9628981	-0.7107220	0.9838736	1.0000000	0.9848894
Population	0.9852252	-0.6859682	0.9933878	0.9848894	1.0000000

To finalise this study, the chi square test was employed to assess one null/alternative hypothesis, for the columns population and Irish speakers.

H0: The two variables are independent.

H1: The two variables relate to each other

```
        Pearson's Chi-squared test

data:  chiSquared
X-squared = 20448, df = 20306, p-value = 0.24
```

We've accepted the null hypothesis and inferred that the two parameters are indeed independent because the p-value resulted as 0.24, which is considerably larger than the initial significance level set as 0.05 for the p-value, and we are confident to reject the null hypothesis and confirm that both columns are dependent.

Evaluation.

In this stage, I evaluated and compared the mining trends, regulations, and dependability to the goal defined. In this section, we look at the pre-treatment procedures and how they affect the Data Mining method outcomes, this section concentrates on the influenced model's readability and usability.

To address the concerns raised initially, we employed the decision tree for prediction, and the regression model for correlation analysis. The most common trends were found on which age group, and province the Irish Speakers were found. We utilized strategies to figure out and illustrate the solution.

This study has presented us that the largest percentages were seen in teenagers between 9 and 19 years old, being its majority in the range of 10-14 year age group, for both genders of all socioeconomic classes, and living in Connacht province.

The next concern was to forecast the number of Irish and non-Irish speakers for the future period. We employed many strategies to discover the relationships/patterns that led to the outcomes, and it was also discovered that State (as a province defined by the dataset) ranked first in terms of the number of Irish and non-Irish speakers. The Tree decision helped us to confirm if the predicted values and the given values were closer or not, meaning if the prediction was done correctly, and the hypothesis helped us to know if the two variables were independent or not.

This study has presented us that the language Irish is gradually reducing and the amount of Irish speakers are decreasing, which can lead us that the Irish (Gaelic) language is indeed a endangered language and is dying gradually, and it's being replaced by English.

5.0 Results

Finally, by performing data mining and prediction analysis on the datasets, relationships and patterns were discovered. The major goal was to identify the key parameters or relationships that would help us in better understanding of condition of the Irish language, such as which age group, city, and province has the largest number of Irish and non-Irish speakers. Likewise, future statistics of Irish and non-Irish speaking people can assist us in launching innovative measures to protect the Irish language from dying.

This study has presented us that the largest percentages were seen in teenagers between 9 and 19 years old, being its majority in the range of 10-14 year age group, for both genders of all socioeconomic classes, and living in Connacht province.

Also, this study has presented us that the language Irish is gradually reducing and the amount of Irish speakers are decreasing, which can lead us that the Irish (Gaelic) language is indeed a endangered language and is dying gradually, and it's being replaced by English.

6.0 Conclusions

Having this study performed, the answers generated can help us to understand how the Irish language is being teach in Ireland across all ages and provinces. Also it can help us understand that even with some kids/teenagers receiving it in scholar ages, as long as the people is evolving and becoming older, there is no more use of the language and it falls into oblivion by not using it.

However, this project has a strong limitation that depends on the data that is added to the dataset by yearly census and by the self-evaluation of each person, to categorise himself/herself on a speaker and non-speaker group, and this understanding of level of Irish language vary from person to person. Also, the data is considerably as true, although there is no further validation on the real level of understanding of each person. Including the kids under 19 are the ones with the highest level of Irish Speaker, however, they do not have access to fill these forms, as the people that must fill them up, are usually parents.

7.0 Further Development or Research

When looking back to this project, with additional time and resources, I would perform a quick interview with real people to understand and set standards of Irish knowledge level and have a more accurate data to analyse.

Including a further research on where the Irish language is currently spoken as main language and understand why it has not being changed to English, so this could enhance this project, to have a better understanding on the environmental and external reasons that are currently affecting this old language to be forgotten.

8.0 References

[1] University of Notre Dame, "What is Irish?," 05 June 2018. [Online]. Available: https://irishlanguage.nd.edu/about/what-is-irish/. [Accessed 08 April 2022].

[2] P. Ó. Doirnín, "Is Irish Gaelic A Dead Language?," 07 November 2020. [Online]. Available: https://fluentirish.com/is-irish-gaelic-a-dead-language/. [Accessed 08 April 2022].

[3] C. Moseley, Atlas of the World's Languages Danger, 2018.

[4] J. A. Fishman, Theoretical and Empirical Foundations of Assistance to Threatened Languages, 1991.

[5] S. Mufwene, "How languages die," Fighting for the world's languages, pp. 377-388, 2007.

[6] D. Boxer, "Applying Socioliguistics. Domains and face-to-face interaction," 2002.

[7] L. Gibbons, "Transformations of Irish Culture," 1996.

[8] P. Mallory, "The Origins of the Irish," 2013.

[9] W. Gibbs, "Saving dying languages," Scientific American, vol. 287, no. 2, pp. 78-85, 2002.

[10] J. Brittain and M. MacKenzie, "Languages endangerment and revitalisation strategies," The Routledge handbook of linguistic anthropology, pp. 433-446, 2015.

[11] M. Bradley, "Is it possible to revitalize a dying language? An examination of attempts to halt the decline of Irish.," Open Journal of Modern Linguistics, vol. 4, no. 4, pp. 537-543, 2014.

[12] L. Grenoble and L. Whaley, "Toward a typology of language endangerment.," Endangered languages: Language loss and community response, pp. 22-54, 1998.

[13] Kaggle, "Statisitics of Irish Speakers," 2021. [Online]. Available: https://www.kaggle.com/datasets/kennethgranahan/statisitics-of-irish-speakers. [Accessed 08 April 2022].

9.0 Appendices

9.1. Project Proposal

9.1.1. Objectives.

The idea of the project is to analyse different datasets with eCommerce details for specific regions and dates and the weather forecast to understand the relation between them in regards to predict if the weather (temperature, rain levels and cloudiness) has any effect in the people minds to purchase things online for specific sectors. If this correlation is achieved, this stats would be a great advance for marketing sectors to promote sales for specific company sectors based on the weather forecast for the following days.

9.1.2. Background

I've choose to work in this project based on my willing to learn Python and found interesting work with different datasets and possibly be able to predict human behaviour based on the weather. To achieve the points described on section 1.0, I will research datasets that contains longitude and latitude details for purchase history and align the dates with weather records, and when having them aligned, will start working to understand typical actions, mode or trends. These analyses will required to take into account the seasons, festive days (e.g. Christmas) between other aspects.

9.1.3. State of the Art

A few similar researches were already done and are spread in the internet, although they are usually focused on spending done in eCommerce rather than retails for rainy days, or how the effect of covid has impacted the eCommerce sales. My project will stand out because the focuses is to predict human behaviour and the results will guide the marketing sector to take actions based on the future weather.

9.1.4. Data

The data required is an invoice historic for eCommerce sales that includes dates and longitude/latitude coordinates of the delivery (or the city, if small) and weather historic for the same period and with similar localisation details. Also taking into account that they are freely shared online although have certification that is a valid and real data.

The files will be processed in a Jupyter Notebook using the Python language and the seaborn, pandas, numpy and matplot libraries to sanitise and remove duplicated or incomplete data and make all comparisons and analyses.

9.1.5. Methodology & Analysis

The methodology used will be the exploratory analysis as the aim is find the relation between both datasets and explore for connections and hypothesis to generate a solution for the real world. And the project will be spread in tasks to be taken weekly, divided by:

- Research more datasets that can provide me with sufficient data;
- Cleaning and wrangling data
- Analyse and find correlations;
- Generate hypothesis and start predicting;
- Complete report with all finds.

9.1.6. Technical Details

As I don't have any knowledge or previous courses in Python, Jupyter Notebook and Data Analyses, these will be the technical developments assigned to this project which will have the major impact for my personal development.

9.1.7. Project Plan

The initial idea for this project is share it 2 major parts linked to both semesters.

The 1st semester, as I have more classes assigned and start to learn the aspects of statistics and data analysis, I will focus my time to researching datasets that can provide me with required data and external aspects that can affect my investigation (e.g. seasons and holidays for the regions where I'm performing this research). The expectative is to have by mid point presentation in December, all the data to perform these studies and start briefly investigating the datasets and their correlation.

The 2nd semester, as I have only 2 classes and the project, I will have more time to focus on the actual code and start generating the hypothesis and trying to predict the influence of weather in the eCommerce. Which I'm expecting to take up to 75 days, starting from mid-January, which will give me April to write the final report and still few weeks in May for emergency actions that may be required.

9.2. Ethics Approval Application

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name:	Douglas Masotti		
Student ID:	x18151493		
Programme	Bsc. (Honours) Computing	**Year:**	4th Year
Module	Software Project (BSHCEDA4)		
Project Title	Analysis of the Weather Influence on eCommerce		

Please circle (or highlight) as appropriate

This project involves human participants	No

INTRODUCTION

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	X	X	X	X	X	X	X	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	No
The project makes use of public secondary dataset(s)	Yes

The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	No

SOURCES OF DATA:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) ***does not need ethics permission****, but* ***needs a letter/email from the copyright holder*** *regarding potential use.*

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner *(e.g. institution, company, etc.)* ***of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration****. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.*

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission:

The dataset chosen has the license: https://www.kaggle.com/datasets/kennethgranahan/statisitics-of-irish-speakers >>>> https://creativecommons.org/publicdomain/zero/1.0/

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form ***OR*** Citation and link to the web site where permission is granted – provided in this form	N/A Yes

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

- Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study
- Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation
- No confidential data will be used for personal advantage or that of a third party

Informed consent

- Consent to participate in the study has been given freely by the participants
- participants have the capacity to understand the project goals.
- Participants have been given information sheets that are understandable
- Likely benefits of the project itself have been explained to potential participants
- Risks and benefits of the project have been explained to potential participants
- Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress
- The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty
- Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)
- Participants have been informed of potential conflict of interest issues
- The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: Douglas Masotti

Date: 29th of March, 2022

Appendix I

1) Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

2) Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

- the source is indicated as Eurostat
- when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Appendix II

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

9.3. Reflective Journals

9.3.1. October

Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: October

What?

Reflect on what has happened in your project this month?

I received the guidelines to work in a project to be delivered by the end of my 4th year which can demonstrate my learnings throughout these years enrolled in this bachelor. The subject from the academic project requested is free for my own choice, although it must be approved by a lecture. With that in mind in a rainy day, I started to brainstorming with myself and looking for questions where I could use datasets with actual real data and try to predict future human behaviours.

So What?

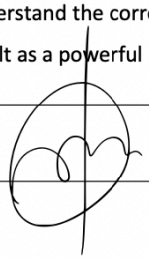
Consider what that meant for your project progress. What were your successes? What challenges still remain?

After a few hours of researching and finding interesting analysis of human compartmental. I decided to work with datasets that could help me to predict if the weather has any effect to the human actions, so I started using myself as an example and looking inwards to understand how I'm affected to a rainy and cold day. As a result, I found out that I'm much more keen to buy unnecessary things from the internet when the sky is grey.

Now What?

What can you do to address outstanding challenges?

Now, to have a real viewing if this is replicable to all humans, I decided to search for datasets that have e-commerce (with sectors clearly specified) and weather data (for determined regions matching the same date) and will start using a Jupyter notebook and Python codes to understand the correlation between them, and try to predict the human behaviour, and possibly having this result as a powerful marketing strategy.

Student Signature	

9.3.2. November

Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: November

What? Reflect on what has happened in your project this month? After the submission for the last reflection journal in October I had some updates that could not be added anymore however, by the end of October I met my Supervisor Vladimir Milosavlijevic for this project. We had a meeting through MS Teams where he give the green signal to continue with the idea I was submitted earlier and added a very useful and notable points to be considered and how should I proceed with the research before starting to actual code the dataset I found, in a way that I could find a unique way to add strength to my project.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? I had advances on my research to find some other studies performed to the same subject and I'm now gathering more examples to be able to learn which aspects were considered and what can I make it differently to have a more unique path to my analysis.
Now What? What can you do to address outstanding challenges? To address these concerns and have more material to study and compare, I will need more time to work and investigate in the internet for other examples and how they could be improved or how they were created to answer the same questions raised in my project and which path can increase value to my analysis.

Student Signature	Douglas Masotti

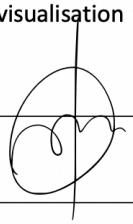
9.3.3. December

Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: December

What? Reflect on what has happened in your project this month? I started the preparation for the mid-presentation as described in the rubric provided. Created the slides and video based on the continuously research about my project idea, although along this project, we have had several submissions in a small gap of time, which does not contribute with many advances on a final result. I will focus mainly in the advances on the next semester where we have programmed less classes and more time to focus on this delivery.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? As several deliveries where happening in close time windows, the advance for this project was not substantial as expected for 2020. However, it will be the focus for the initial months of 2021.
Now What? What can you do to address outstanding challenges? I will focus on January and February, every 3 nights per week and entire Saturday to focus on this project, and I'm planning to have a more accurate position with 90% of the project ready by end of March. Where I will spend my April focusing in wrapping up my conclusions, enhancing the data visualisation techniques and creating a high level delivery.

Student Signature	Douglas Masotti 

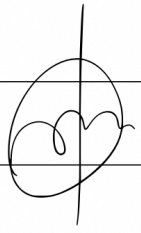
9.3.4 January

Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: January

What?
Reflect on what has happened in your project this month? I started refreshing the idea of my project and continued searching for similar projects in the internet.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? The start of the year is usually carried out with several deliveries in my professional life, although February is the month that my project will have the strongest impact in advances, my plans are to have a quick skeleton of the final delivery, use March to work in enhancements, and April to final details and deliver.
Now What? What can you do to address outstanding challenges? I will focus on February, every 3 nights per week and entire Saturday to focus on this project, and I'm planning to have a more accurate position with 90% of the project ready by half of March. Where I will spend my March and April focusing in wrapping up my conclusions, enhancing the data visualisation techniques and creating a high level delivery.

Student Signature	Douglas Masotti 

9.3.5 February

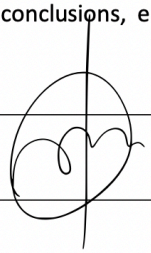
Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: February

What? Reflect on what has happened in your project this month? I started refreshing the idea of my project and continued searching for similar projects in the internet.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? The start of the year is usually carried out with several deliveries in my professional life, although March is the month that my project will have the strongest impact in advances, my plans are to have a quick skeleton of the final delivery, use April to work in enhancements, and May to final details and deliver.
Now What? What can you do to address outstanding challenges? I will focus on March, every 3 nights per week and entire Saturday (when there is no classes) to focus on this project, and I'm planning to have a more accurate position with 90% of the project ready by half of March. Where I will spend my March and April focusing in wrapping up my conclusions, enhancing the data visualisation techniques and creating a high level delivery.

Student Signature	Douglas Masotti



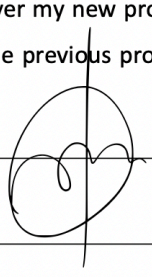
9.3.6 March

Supervision & Reflection Template

Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: March

What? Reflect on what has happened in your project this month? This month I had to stop everything gathered and advanced until this point and go back to the brainstorm step. The idea set to this project was not aligned to my capacities of delivery, so I had to withdraw my idea and resubmit all documentation about my new project.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? That defined in late March, I had separate all code lines that I could use to my new project and start analysing the online available literature on the topic Irish Speakers to understand and help me to approach this topic and have a better analysis on the subject.
Now What? What can you do to address outstanding challenges? The main challenge now is with the time I have left to organise and deliver my new project. However, the selected project does not include a link between multiple datasets as the previous project, so it will save several hours.

Student Signature	Douglas Masotti 

9.3.7 April

Supervision & Reflection Template	
Student Name	Douglas Masotti
Student Number	x18151493
Course	BSc (Honours) in Computing

Month: April

What?

Reflect on what has happened in your project this month?

The April was decisive to my project delivery. I had several challenges with the old dataset chosen and it was only visible as long as I was working with it. So, having the project idea changed so close to the deadline and assuming the responsibility to deliver a better and improve report in less time was overwhelming, although, I was able to communicate this in my current job and take some extra days off to work fully and exclusively on this outcome. The R is a language that I had no knowledge and start learning more with the development of our 2nd semester, which helped me to work along my classes and create the final analysis report in Irish Speakers.

So What?

Consider what that meant for your project progress. What were your successes? What challenges still remain?

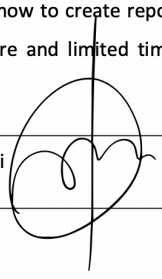
From what I have defined initially last year, unfortunately, I was unable to create and deliver with the datasets and knowledge I had until close to the end of this course. The challenges were definitely the time, although I'm happy that I decided to change the project by mid-March, so I had enough time and experience from my previous experience to start and deliver this analysis report.

Now What?

What can you do to address outstanding challenges?

The outstanding challenge of this entire project was the difficulties I faced on R code, although it helped me to create a great knowledge to work with and a good experience on how to create report analysis. Also, I must highlight the increase in my experience to work under pressure and limited time, taking decisive directions and acting precisely.

Student Signature	Douglas Masotti



9.4. Invention Disclosure Form

Not applicable

9.5. Other materials used

Not applicable for mid-presentation.