UCDPA Project Report

**Review of 1901 and 1911 Census Data for the parish of Tulsk, Co. Roscommon**

**Course Title:** Certificate in Introductory Data Analytics

**Date:** 19/07/2022

**Submitted by:** Marese Feeney

# GitHub URL

https://github.com/MareseF/UCDPA\_Marese-Feeney

# Abstract

This project has been completed as a part of a project submission for UCD Certificate in Introductory Data Analytics. The objective of the project is to review and compare the census data from 1901 and 1911 for the Tulsk Parish area in Co. Roscommon and to gain insights into the people living in the area at that time including the number of people living in the area, their ages, their occupations, and their literacy. The project uses census dataset from The National Census archives which were downloaded into a CSV and imported into the Jupyter Notebook for review. After analysis of the data, a number of insights were gained, and these are outlined in the Visualisation, Insights and Results section. The implementation uses functions, lists, merging of datasets, conditional statements, looping, Numpy and graphic visualizations to carry out the interrogation of the data.

# Introduction

In May 2022, I joined the History Society in my home village of Tulsk, Co. Roscommon. Upon joining, I was tasked with reviewing the Census data for 1901 and 1911 for the local area.

The starting of my Certificate in Introductory Data Analytics coincided nicely with this task and I saw that it was the perfect dataset to complete my project on. In reviewing the dataset, I wanted to get an insight into the people living in the Tulsk Parish area in 1901 and 1911. The area is made up of 5 Districts (DEDs) and the data was obtained from the National archives of Ireland.

# Dataset

# The dataset was taken from The National Archives of Ireland website. Due to the limitation on that site i.e. only allowed to access a max of 100 records at a time – the data was downloaded from the site for each Census year and for the 5 DED areas. The data was saved into two CSV files – one for 1901 and one for 1911.

# Below is a screenshot of the [National Archives: Census of Ireland](http://www.census.nationalarchives.ie/search/results.jsp?searchMoreVisible=&census_year=1911&surname=&firstname=&county19011911=Roscommon&county1821=&county1831=&county1841=&county1851=&parish=&ward=&barony=&townland=&houseNumber=&ded=baslick&age=&sex=&search=Search&ageInMonths=&relationToHead=&religion=&education=&occupation=&marriageStatus=&yearsMarried=&birthplace=&nativeCountry=&language=&deafdumb=&causeOfDeath=&yearOfDeath=&familiesNumber=&malesNumber=&femalesNumber=&maleServNumber=&femaleServNumber=&estChurchNumber=&romanCatNumber=&presbNumber=&protNumber=&marriageYears=&childrenBorn=&childrenLiving=#searchmore) search page.

# Graphical user interface Description automatically generated

This source was chosen as there is no other data source available for the census data. Also, no census records exist for any other years for Co. Roscommon other than 1901 and 1911. (The National Archives of Ireland, 2022)

# Implementation Process

There were a number of steps in the implementation process and these are outlined below:

1. **Data**

This project used a real-world dataset from the Census Data 1901 to 1911 for 5 DEDs Townlands in the Tulsk area of County Roscommon. The source of the data is [National Archives: Census of Ireland](http://www.census.nationalarchives.ie/search/). The data was extracted from the website and combined into one CSV file for 1901 and one CSV file for 1911.

1. **Importing**

Prior to importing the datasets, the relevant libraries were imported:

Graphical user interface, text, application

Description automatically generated

The data was imported from the two CSV Files:

* Census Data Tulsk Parish\_1901.csv
* Census Data Tulsk Parish\_1911.csv

In the first instance, Census Data Tulsk Parish\_1901.csv was imported and reviewed before the Census Data Tulsk Parish\_1911.csv file was imported and subsequently merged.





Two dataframes were created for the data sets:

* Census Data Tulsk Parish\_1901.csv = “census\_1901”
* Census Data Tulsk Parish\_1911.csv = “census\_1911”

Once the datasets were merged, a new data frame name was assigned – “df”

1. **Preparation**

As detailed aboved, pandas dataframes were created for each of the datasets with the merged dataframe referred to as “df”. A review was carried out on each dataset and subsequently the merged dataset for example, a check for Null values was carried out and the replacement of these:

|  |  |
| --- | --- |
| 1901 | 1911 |
|  |  |

In carrying out the initial review, it was evident that the 1911 census data was missing column 18 – “Census Year” and this was then added to the Census 1911 dataset before merging the two files:

Graphical user interface, table

Description automatically generated

It was observed that the inclusion of the “Census Year” column in the 1911 dataset prior to merging with the 1901 dataset had not included the row value “1911” – hence NaN appeared. NaN was replaced in this column using the “.fillna” function with the correct reference “1911”.

Where null values were observed, these were replaced in the datasets – example shown below:

Table

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generatedThe merged censuses dataframe, “df”, was reviewed under a number of criteria to understand the shape of the dataset, the headers and to confirm no null values remained. Additionally, the Index column was renamed as “Ref\_Number”

The combined data resulted in 19 columns and 6048 rows of data.

1. **Data Cleansing**

Once the datasets were merged, further data cleansing was carried out. Outliers were identified e.g. When the surnames were placed in alphabetical order, a “?” appeared in the data set. This was replaced with “Not Detailed”.

Text

Description automatically generated with medium confidence

In reviewing the dataset, it was noted that in error, an additional column “Census” had been added which captured the “Born” details and needed to be removed.

Graphical user interface, text, application, table

Description automatically generated

A check was carried out to ensure no duplicates existed:

Graphical user interface, text

Description automatically generated with medium confidence

At the time that the data was taken from the National Archives website, the age of the individuals was not captured. In order to review the ages of the individuals, an additional column needed to be added “calculated\_age”. In addition, prior to the adding of this column, a review on the data type needed to be completed. As noted below, the two columns being used in the calculation were “Census Year” (Dtype: object) minus “Born” (Dtype, int64).

Table

Description automatically generated with low confidence

Once the update was made to the Census Year, it was possible to include the calculation for the “calculated\_age”.

Table

Description automatically generated

1. **Analysis**

Analysis was carried out on the data using conditional statements, looping, groupby, Numpy function and addition of a new list value. These are outlined below:

* **Conditional statement**

A conditional statement was first added to identify the number of infants which was 491.

Table

Description automatically generated with medium confidence

However, it was realized that additional statements would need to be applied to make the information more meaningful and give a more expansive age demographic. The **conditional statement** was made using **Numpy** and by adding a new value **list:**

Table

Description automatically generated

The combination of the functions meant that there were no null values for the “Age Demographic” column now added.

* **Groupby** was used to get an understanding of the data:
  + Grouped by DED which informed that there were more people in the DED areas of Cloonyquin, Tulsk and Killukin than there were in Ogulla and Baslick DED.

Table

Description automatically generated The use of a **looping** function also indicated this as illustrated overleaf.

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generated

1. **Visualisations, Insights & Results**

Seaborn was used to generate several charts to gain further insights into the dataset. The insights are detailed below with the associated charts included for reference. The code for the creation of the charts is captured in the Jupiter notebook under the heading “Visualisation” but has also been included below for reference.

1. **Baslick and Ogulla DED**

From the above Groupby and Looping functions, it was assumed that the number of respondents in the Balick and Ogulla DED areas was low. However, using a Seaborn count plot chart it was identified that there was no census data for the Baslick area in 1911 and the approx. 613 responses were solely from 1901. It can be assumed that the 1911 census data for Baslick was not available on the National Archives website and is something that needs to be followed up on. It may be that the records were destroyed which is the case for census data prior to 1901. Additionally, the total number of respondents in the Ogulla DED area for both years was 647 but through graphic visualization, it is clear that there were more people in the Ogulla DED area in 1901 with a slight drop off in 1911 of about 20 people. Of all five DEDs in the Tulsk area, Ogulla had the smallest population.

A picture containing text

Description automatically generated

Chart, bar chart

Description automatically generated

1. **Calculated age**

There was a greater increase in the number of people in the age group 65-80 in 1911 than would have been expected based on the data available. It can be assumed that some changes in reported year “born” from the 1901 to the 1911 can be attributed to the introduction of the Old Age Pensions Act 1908 which introduced a non-contributory pension for 'eligible' people aged 70 and over. From the 1s January 1909, people over seventy, whom had an income less than 10s a year and having passed a 'character test' and being found to be of 'good character' could receive a weekly pension (Morgan, 1999). For this reason, it is thought that people increased their age in the 1911 census in order to be eligible sooner for the newly introduced pension.

Chart, histogram

Description automatically generated

1. **Age Demographic**

Most respondents were found to be teenagers with a total of 1722 recorded over the two censuses. In addition, the number of individuals who were in the 90+ age group was 6 in 1911 but 0 in 1901. The low numbers in population for the age demographic of 30 to 90+ compared to the number of teenagers may will be attributed to the Irish Famine which ripped through Ireland in 1845 to 1849.

A picture containing diagram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

1. **Sex of Individuals and their literacy**

There were more Males than Females respondents living in the Tulsk parish area over 1901 and 1911 except for Baslick where we have noted almost an equal split and as we know from earlier observations, this DED area only has data for 1901.

A picture containing calendar

Description automatically generated

Chart, bar chart

Description automatically generated

Additionally, a review of the literacy of respondents indicates that more Males than Females could “Read and write”:



Graphical user interface, application

Description automatically generated

A fewer number of people were recorded as being able to “Read and write” in the 1911 census compared to the 1911 census however, this may be due to the lack of data for Baslick area at that time.



Graphical user interface, application

Description automatically generated

1. **Religion**

Whilst there were entries for “Church of Ireland” and “Protestant Irish Church” – the most common/ popular religion in the area in 1901 and 1911 was Roman Catholic which accounted for almost 100% of respondents in both censuses. The shortfall in the 1901 vs 1911 data could again be attributed to there being no data for the Balick DED area in 1911.

A picture containing diagram

Description automatically generated

Graphical user interface, text, application, email, Teams

Description automatically generated

1. **Birthplace**

Almost 100% of individuals were born in Co. Roscommon or elsewhere in Ireland. There were some individuals who were identified as having their birthplace in Australia, United States, Scotland and England.

**Graphical user interface, application

Description automatically generated**

1. **Occupations**

Looking at the occupation of respondents over the 10-year period, Farming featured very strongly. The number of farmers recorded in 1901 was 588 compared to a drop of 50 in 1911. However, there were many terms used to describe someone involved in farming e.g. Herdsman, Farmer’s Son, Farm Labourer, Shepard, Agricultural laborer. Given the higher volume of men than women it’s not surprising that farming features so strongly.

In the 1911 census, there was a big increase in the number of “No Occupation detailed” numbers rising from 41 to 121 – a rise of 70%. In addition, the number of “House Keeper’s” recorded in 1901 was 187 vs 0 in 1911. Both of these anomalies could be attributed to the guidance for completing the Occupation section of the 1911 census form i.e.” *No entry should be made in the case of wives, daughters, or other female relatives solely engaged in domestic duties at home.”*

A pivot table was created to identify the above insights – extract below:

Graphical user interface, text, application

Description automatically generated

1. **Relation to Head of Household**

The relation to head of household strongly featured more sons than daughters. Looking at the data set, there was almost 20% more sons than daughters recorded. This may be due to the fact that women usually married sooner than the men and went to live with their husbands (Guinnane, 1992). For example, in 1901 census there were 921 sons recorded compared to 716 daughters.

**Chart

Description automatically generated**

1. **Machine Learning**

## In Machine Learning, one of the simplest prediction models is Linear Regression. It allows for predictions on the data for this reason, it is suggested to use a simple linear regression model based on the census data available could be used to predict future populations for the Tulsk parish areas.

# References

* Guinnane, T. W. (1992, Sept). Age at Leaving Home in Rural Ireland, 1901-1911. *The Journal of Economic History*, 651-674. Retrieved from https://www.jstor.org/stable/2122889
* Morgan, E. (1999, September 21). *Irish TImes*. Retrieved June 22, 2022, from https://www.irishtimes.com/culture/1900-1909-1.229664#:~:text=January%201st%2C%201909%3A%20The%20first%20old-age%20pension%20of,to%20be%20rejected%20by%20Lords%20in%20200%20years.
* The National Archives of Ireland. (2022, June 1). *Census National Archives*. Retrieved from National Archives: http://www.census.nationalarchives.ie/search/