|  |  |
| --- | --- |
|  | |
| Data Visualisation Project  Assignment 2 – Report  Ireland Census 2011 - 2016: Changing Socio-Economic Factors | |
| Module code : SPEC9995: 2022-23 | |
| Ciaran Finnegan  Student No : d21124026  TU060 Yr. 2  Data Science MSc  8/12/2022 |  |
|  |  |

Table of Contents

1 Assignment Overview 3

1.1 Introduction 3

1.2 Problem Statement 3

1.3 Datasets and Intended Audience 3

2 Part 1: Data Pre-processing 4

2.1 Data Exploration 4

2.2 Data Wrangling 5

3 Part 2: Visualisations 6

3.1 Visualisation One – Employment Rates per County 6

3.2 Visualisation 2 – Socio-economic Groups in Ireland 7

3.3 Visualisation 3 – Impact of Socio-economic Group on Health 8

3.4 Previous Iterations with Assignment Visualisations 9

4 Appendices 10

4.1 Appendix 1 – R Source Code – Visualisation One 10

4.2 Appendix 2 – R Source Code – Visualisation Two 10

4.3 Appendix 3 – R Source Code – Visualisation Three 10

# Assignment Overview

## Introduction

This document covers the design, implementation, and observations on all parts of the December 2022 CA(2) for the Data Visualisation module (TU060 – DS – Year 2).

This assignment is based on Irish Census data from 2011 and 2016.

The three sets of visualisations provide a comparison on a number of social and economic data points gathered by both census.

## Problem Statement

The 2011 Census took place in the midst of significant turmoil, when Ireland required external support to avoid economic collapse. This period of austerity had a significant impact on employment and prosperity for the Irish population.

Five years later, in 2016, Ireland was considered, by some analysts, to be starting to emerge from this period of uncertainty.

The visualisations in this assignment attempt to show graphically, using a certain subset of information, that 2016 data points show an improving situation for the Irish population since the difficult days of 2011.

## Datasets and Intended Audience

The datasets for the visualisations in the assignment come from two sources;

* Graphical information on Irish county boundaries from the Ordnance Survey of Ireland (OSI) website.
* Census ‘theme’ data for 2011 and 2011, compiled at Irish county level by the Central Statistics Office (CSO), in the following domains;
* Employment rates per county.
* The size of various socio-economic groups in Ireland.
* The perceived health/wellbeing rating of the population in each county.

The intended audience is anyone looking to understand if economic actions taken in Ireland in 2011 resulted in a measureable improvement in employment and prosperity five years later.

# Part 1: Data Pre-processing

## Data Exploration

### Geographical Data

The first visualisation uses a map of Ireland with county boundaries. Code libraries are available within the ***R*** language to read detailed geographical information and then augment with additional attributes.

The OSI provide a 2019 data file with Irish country boundaries, in a ***.shp*** format, at his location;

<https://data-osi.opendata.arcgis.com/maps/osi::counties-osi-national-statutory-boundaries-2019-generalised-20m>

The *readShapePoly* function in **R** reads this format and allows manipulation and addition of data points for use in the *ggplot* graphing function.

The map data is ‘clean’ but it is combined to additional data elements, as described in Section 2.2 below, to provide the richness of detail in Visualisation One.

### ‘Themes’ from Census Data

The 2011 and 2016 Census used a largely similar grouping of numerical data points under 15 categories or ‘Themes’. Within many of these were sub-tables of data, breaking the Themes down into lower levels of granularity.

A glossary of the census themes is located here; <https://www.cso.ie/en/media/csoie/census/census2016/census2016boundaryfiles/SAPS_2016_Glossary.xlsx>

This assignment looks at the Census numbers collated at the Irish county level for 2011 and 2016 for the following Themes;

* **Theme 8:** *Principle Status*. The focus in on the population of unemployed in the workforce for each county.
* **Theme 9**: *Social Class and Socio-Economic Group*. This captures the number of people in each county who identify as members of a given employment category. As an example, Class A and B refer to Employers and Higher Professionals respectively while Class G is for ‘Unskilled’.
* **Theme 12**: *Disability, Carers and General Health*. For this assignment, we only consider sub-table 3 that captures numbers of people by how they perceive their general health/wellbeing.

The 2011 Census theme data is available from the CSO here: <https://www.cso.ie/en/media/csoie/census/documents/saps2011files/AllThemesTablesCTY.csv>

2016 theme data was less straightforward to find on the CSO website, but a copy is stored on Kaggle at;

<https://www.kaggle.com/datasets/stephenofarrell/irish-census-2016>

## Data Wrangling

As expected, the map and census data has no missing values. However, in order to improve the visualisations in this assignment the following ‘data wrangling’ steps were required;

* The OSI map data file contains all the information to render the 26 counties of the Republic of Ireland, but contains no data points on attributes of the actual human population in each county. Thus the ‘theme’ files were loaded into a separate dataset for both 2011 and 2016 and ***joined*** separately to the county map dataframe, based on County name.
* There are 26 counties in the Republic of Ireland and therefore 26 rows in the OSI map ***.shp*** file read in the assignment code. The 2011 and 2016 Census County files breakdown certain metropolitan areas into further subdivisions. For example, Dublin County census data is split into the following sub areas; *Fingal*, *South* *Dublin*, *Dublin City*, and *Dún Laoghaire-Rathdown*. Thus it is necessary to ***re-group*** the metropolitan areas back into a single County row first before merging with the map dataset. This reduces both ‘theme’ dataframes from 30+ rows back to 26.
* To add complexity, the 2011 and 2016 Census differ slightly on the metropolitan divisions. For example. Limerick City and Limerick County are different areas (with separate rows in the dataset) in the 2011 Census, but are treated as a single area in 2016. Visualisation One had to cater for the slight structural changes in data capture between these census years.
* The ‘theme’ files only capture absolute numbers of peoples or households. It was necessary to add calculated fields to determine the ***rate*** of certain attributes per County. For example, the columns named ***T8\_1\_LFFJT*** and ***T8\_1\_ULGUPJT*** capture the population in each County looking for a first job or unemployed respectively. These values have to be divided by the total possible workforce (***T8\_1\_TT***) to find the rate of unemployment in each county.
* A similar calculation is applied to find the ratio of Socio-economic group A and B in each county (as a proportion of total population).
* The percentage of those in ‘*Very Good Health’* is derived from the number recorded in that column for each county divided by the overall county population.
* 26 separate unemployment rates generated an unwieldy legend and graph for Visualisation One. It was necessary to add another derived field that broke down this continuous set of values into a manageable, and more meaningful, range of ***categorical*** percentage descriptions.
* The 2016 theme ***.csv*** file stores the *Dún Laoghaire-Rathdown* region for Dublin with a different text format than the 2011 file. The *fada* over the ‘u’ caused a ***data corruption*** error with the rendering of 2016 ggplot graphs. Thus, it is necessary to directly update this dataframe cell in **R** code before rendering the graphs in Visualisation Three (*Dún Laoghaire-Rathdown* is specifically labelled in the Scatter Plot).

# Part 2: Visualisations

## Visualisation One – Employment Rates per County

How did rates of employment change across Ireland between 2011 and 2016?

The OSI map data provides an outline of county boundaries in the Republic of Ireland and the rates of employment in each county for 2011 and 2016 are represented below.

<map>

A graded colour scale is used to show the areas of highest unemployment (darkest colours).

The scale to measure the range of unemployment is kept as close as possible between 2011 and 2016.

It is evident that employment statistics improved in the five years after 2011. That said, areas like Donegal continue to rank lowest in employment metrics.

## Visualisation 2 – Socio-economic Groups in Ireland

Did the population of Ireland, as a whole, perceive that their social and economic status had improved in 2016, compared with 2011?

The Irish Census questionnaire each respondent elects the socio-economic group into which they believe they belong. This reflects a person’s sense of their status, and possible affluence.

The bar chart below is a comparison of the population numbers in each socio-economic group in both 2011 and 2016.

<map>

Of note in this graph is that these are absolute population numbers, aggregated to the national level from the individual county data.

Overall population in Ireland would have risen slightly in the period from 2011 to 2016.Thus, if other factors remained static, one would expect to see a general increase for all groups in 2016.

In fact, the graph shows that lower skilled employment groups (F, G, H) show zero or negative growth from 2011. The A and B categories appear to be slightly larger than one might expect.

Thus there is an apparent increase in the general perceptions of prosperity across the Irish population in 2016.

## Visualisation 3 – Impact of Socio-economic Group on Health

Does social standing impact on an individual’s health? Has this relationship changed between 2011 and 2016?

The Irish Census does not directly record income levels but it does measure the population by ‘social class’, which is an indicator of general affluence.

Do those who are ‘better off’ financially generally have better health (possibly either through lifestyle or access to health care)?

<map>

Both scatter plots show a relationship between those who identify in Socio-economic groups A and B and how many consider themselves in ‘Very Good Health’, according t the question on the Irish Census forms.

The data is represented for each county in Ireland and we use this 26 rows of data points to determine if a correlation exists.

So in Ireland, are the county areas with the highest percentage of population in Class A and B, the places with the healthiest people?

That positive correlation can be seen in both graphs, but it is not very strong. It has also not changed in any significant magnitude between 2011 and 2016. Such changes may take much longer?

For illustration, the county areas at the top and bottom of the scales are highlighted. Dún Laoghaire-Rathdown is certainly the area with the wealthiest and healthiest people, but is somewhat of an outlier.

## Previous Iterations with Assignment Visualisations

### Visualisation One

The first map graph just used the ***absolute numbers*** of unemployed in each county to generate the colour gradients. This distorted the true picture because populations with low populations will have lower numbers of unemployed workers, but may still have high ***rates*** of unemployment.

### Visualisation Two

Originally this was going to be another two graph comparison of data, similar to Visualisation One and Two.

However, comparing this type of bar chart data is much more effective with one graph that grouped twin bars together by year/socio-economic group.

### Visualisation Three

The original scatter pot contained just points for all Irish counties. This was sufficient for the general message of the visualisation but lacked ‘sparkle’.

Add labels to represent the county areas at the lower and upper areas of the graph added some useful context in terms of the spread of affluence across Ireland.

Initially two R files were used to generate the 2011 and 2016 graph data for Visualisation 3. However, as there was no complication with joining the OSI boundary data in this visualisation, the code was combined into a single file with a ***for…loop*** to generate multiple graphs.

# Appendices

## Appendix 1 – R Source Code – Visualisation One

Although there are

## Appendix 2 – R Source Code – Visualisation Two

Although there are

## Appendix 3 – R Source Code – Visualisation Three

Although there are