# Immigration’s role in the 2016 UE Referendum

*Jose Raúl Ruiz García, 2017*

# Summary

The immigration policy played a major role in the past UE Referendum in the UK. Working through two different datasets we will try to frame and delimit what was the real impact of these figures regarding the outcome of the referendum. One of the datasets presents the results of the referendum along with some other pieces of information. The second one features population related estimates along with immigration figures for the UK across the years 2001 – 2015.

By categorising the data using a *k*-means technique, I identified an association between the net immigration figures for each district and the outcome of the referendum. I also analysed and draw some conclusions on how population and immigration figures are distributed across the UK.

# Aims and objectives

The report I will produce will be aimed at a fictional NGO that supports immigrants living all over the UK and assist them in legal, work and accommodation related issues. After the referendum, they want to have a closer look at how the new situation will affect immigrants living in United Kingdom. To do so, they need to analyse how immigrant population spreads over the UK, what did their districts voted and what percentage of people turned out to vote.

My first objective is to describe how immigration figures and population change affects each district and country within the UK. Using the data provided by the UKMYE dataset I will analyse these figures both by district and by country and will try to see the impact of this figures in each of them.

The second question posed by the analysis seeks to find out if there is a correlation between the immigration received by a region in the studied period and the outcome of the referendum.

# Background to the investigation

During the 23rd of June 2016, place in UK had place one of the most important European political event of the XXI century so far. During that day, every British, Irish, Commonwealth citizen living in the UK and some British expat were eligible to vote in the EU Referendum to decide if the UK should remain a part of the European Union or otherwise separate from it (Express, 2016). As we know, the referendum ended up giving the victory to the Leave option with a 52% of votes against the Remain option, with just the 48%.

Before, during and after the referendum, many newspapers, TV, and media outlet put their analysts to work on the Brexit figures. Specifically, many analysts saw immigration and border control as major aspects of the problem. In this report, we will focus on analysing how two figures affected or related to the triumph of each options in each district, region or country.

# Scope and sources of data

Full details of data sources and licences are given in Appendix 2.

## The referendum results data

The Referendum results data was obtained from the Electoral Commission Website (The Electoral Commission, 2016 b), saved to a CSV file and then imported to a IPython Notebook as a Pandas Data Frame object for its later analysis and manipulation. Complete details regarding its preparation, cleaning and import are given in TMA 02 (Appendix 3).

## Population Estimates for UK, England and Wales, Scotland, and Northern Ireland data

The Population Estimates for UK, England, wales, Scotland, and Northern Ireland data was obtained from the Office for National Statistics website(Office for National Statistics, 2016). It consists in 5 Microsoft Excel Worksheets and 3 CSV that depicts data regarding population figures, international and national immigration, sex, and density. Most of the data is aggregated at national, regional and district levels.

The dataset that I will be using the most during my investigation is located within the file *MYEB3\_summary\_components\_of\_change\_series\_UK\_(0215).csv*. Within it, each row represents one local authority district. It also features a large number of columns that offer information regarding the district’s *lad2014\_name*, *lad2014 code*, *country*, and then figures for each year’s *population*, *births*, *deaths*, *net change*, *internal incoming*, *outcoming*, and *net migration*, *international incoming*, *outcoming*, and *net migration*, and *other change*. The period of time covered by the population figures goes from 2001 up to 2015.

The dataset covers the clear majority of UK authority districts for every year between 2001 and 2015. However, there are some important omissions that we should consider and investigate further:

* The only data featured from 2001 is the population.
* The Northern Ireland 2015 figures are missing. They are represented in the CSV file with a semicolon, which may cause problem on import.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lad2014\_code | lad2014\_name | country | population\_2001 | births\_2002 | deaths\_2002 |
| E06000001 | Hartlepool | E | 90152 | 1017 | 1042 |

Table 1 Section of the csv file

The information within the dataset do not present any special considerations regarding privacy or anonymity since it is publicly available. It does not help identifying individuals based on the data it contains or combined with other datasets. Some of the CSV files contained within the dataset that could lead to more sensitive observations are anonymised by aggregation and the figures are large enough as to not be able to point at one specific person.

# Analysis pipeline

## Referendum results data import and preparation

The first dataset used is the Referendum results one. Due to the nature of the data within the file it did not require of any particular cleaning or reshaping on a first instance. It was imported to an in-memory structure through an IPython Notebook. I thought the best way to work with it would be using a Pandas Data Frame since the csv fits perfectly in the standard Data Frame structure and some of its columns (specially *Area\_Code* and *Area*) would turn out very useful when we merge both datasets for a further analysis.

The column names and the format in which they are imported to a Pandas Data Frame are below.

|  |  |
| --- | --- |
| id | int64 |
| Region\_Code | object |
| Region | object |
| Area\_Code | object |
| Area | object |
| Electorate | int64 |
| ExpectedBallots | int64 |
| VerifiedBallotPapers | int64 |
| Pct\_Turnout | float64 |
| Votes\_Cast | int64 |
| Valid\_Votes | int64 |
| Remain | int64 |
| Leave | int64 |
| Rejected\_Ballots | int64 |
| No\_official\_mark | int64 |
| Voting\_for\_both\_answers | int64 |
| Writing\_or\_mark | int64 |
| Unmarked\_or\_void | int64 |
| Pct\_Remain | float64 |
| Pct\_Leave | float64 |
| Pct\_Rejected | float64 |

Table 2 Referendum dataset column names and data types

## Population Estimates data import and preparation

The import of the Population estimates datasets was as straightforward as the Referendum results one. It was also stored in a Pandas Data Frame in an IPython Notebook since it fitted the structure perfectly and was not large enough as to cause any space of performance issues.

However, when looking at the different column data types I spotted some columns which had been imported as object despite being numeric types. A further investigation into this helped me to find out a couple of reasons why these two groups of columns had been converted to the wrong data types:

* The thousands where parsed as objects because they had commas separating numbers.
* As I mentioned before, all the data for Northern Ireland 2015 estimates was missing, what caused the column to be parsed as string objects.

|  |  |
| --- | --- |
| lad2014\_code | Object |
| lad2014\_name | Object |
| country | object |
| population\_2001 | int64 |
| births\_2002 | int64 |
| ... | ... |
| births\_2015 | object |
| death\_2015 | object |
| natchange\_2015 | object |

Table 3 Population estimates column names and data types

The approach I used to clean the thousand commas issue was straightforward: add the *thousands = ','* parameter when importing the CSV file. This way pandas would interpret commas between numbers as thousand commas instead of regular ones.

The missing data from Northern Ireland during the 2015 could not be fixed at all, since the data was simply missing. Therefore, I opted for leaving out all the 2015 columns from any further analysis.

## Combining datasets

When it came to merge both Data Frames, a further cleaning and harmonising was required. In order to be sure that the merge will be carried out successfully I had to check that the number of codes and the codes itself from both Data Frames were the same. I did this by creating two different sets with the codes from both Data Frames and finding out discrepancies with the python command:

set\_codes\_missing = set\_immigration\_codes ^ set\_referendum\_codes

The command above will select the codes that appear only in one of the Data Frame sand assign them to the set\_codes\_missing variable. Then, selecting the rows of those specific codes in their respective Data Frames I verified that the following rows did not appear in both:

|  |  |
| --- | --- |
| Area | Country |
| Antrim and Newtownabbey | N |
| Ards and North Down | N |
| Armagh City, Banbridge and Craigave | N |
| Belfast | N |
| Causeway Coast and Glens | N |
| Derry City and Strabane | N |
| Fermanagh and Omagh | N |
| Mid and East Antrim | N |
| Lisburn and Castlereagh | N |
| Mid Ulster | N |
| Newry, Mourne and Down | N |

Table 4 Rows that only appear in the referendum results data frame

|  |  |
| --- | --- |
| Area | Country |
| Northern Ireland | N |
| Gibraltar | GI |

Table 5 Rows that only appear in the population data frame

To work with the most accurate and complete data we need to reshape these rows somehow so we end up with the best possible data frame. As it happened before with the missing data, if the row for Gibraltar is shown only in one Data Frame there is no option but to drop it during the merge. However, we can still manipulate the Northern Ireland data so it is present in the merged Data Frame.

The approach I took consisted in aggregating the data (adding the figures) from every Northern Ireland Region from the Referendum results Data Frame and add them into a newly created Northern Ireland Row. So, at the end I came up with only one row for the whole country but I did not lose all the data in the merger. This process is performed in the notebook *Immigration\_Import\_cleaning* under the header ‘Aggregating data from Northern Ireland’.

## Analysing data

In order to answer the question “how do immigration and population figures affect the outcome of the referendum?” I decided to aggregate several columns into a single one resulting of the sum of those numbers over the years. I created in total three new columns per row that contained the information shown below.

|  |  |  |
| --- | --- | --- |
| Data obtained for aggregation | Columns used | Column name |
| Aggregated figures for net immigration | international\_net\_***year*** | international\_net\_total |
| Aggregated figures for incoming immigration | International\_in\_***year*** | international\_net\_total |
| Aggregated figures for net population change | population\_2014 and population\_2001 | population\_change |

Table 6 Overview of aggregation operations performed on data

The aggregation process can be seen in the notebook *Immigration\_Import\_cleaning* under ‘Aggregating data over the years’.

During the analysis phase of the pipeline I produced several visualisations that will be explained in detail under the [findings](#_Findings_(750)) section. The most significant of these visualisations was the one corresponding to the *k*-means data mining technique used to find out possible categorisations within the data. I applied this technique using 2 clusters to categorise the data and I obtained with the following scatter plot, which represents both clusters and will be analysed in the findings section.

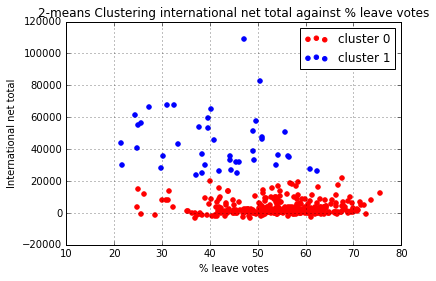


Figure 1 2-means clustering international net total against % leave vote

The plot shows a clustering done in a scatter plot of the leave vote (%) against the aggregated net international migration received in each district.

# Findings

## Findings on how UK population, population change and immigration figures are distributed across districts and countries

In order to generate aggregated values per country I grouped the data in the UKMYE dataset and added the values, obtaining the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country | Population\_2014 | International in total | International net total | Population change |
| E | 54316618 | 6691100 | 2722988 | 4866872 |
| N | 1840498 | 168493 | 33365 | 151660 |
| S | 5347600 | 489200 | 155300 | 283400 |
| W | 3092036 | 180314 | 50679 | 181804 |

*E = England, N = Northern Ireland, S = Scotland, W = Wales*

Table 7 Demographic data per country

From the table below, I produced relative values for both incoming and net immigration figures obtaining the plot below.

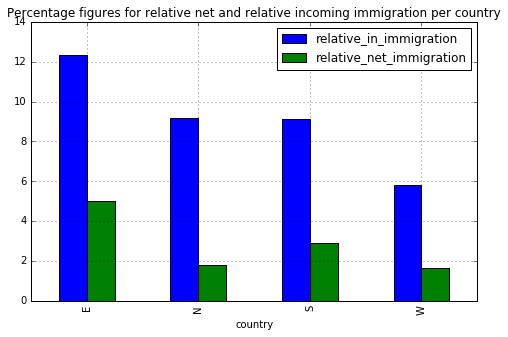


Figure 2 % immigration figures

From the table and the plot it can be seen how England, being the country with the highest population figures for 2014 also has the highest relative incoming and net immigration figures. It is also interesting to note how, although the relative incoming immigration figures for Northern Ireland are slightly higher than the Scottish ones, Scotland’s relative net immigration figures are one point higher than the Northern Ireland ones. Meaning this that more immigrants stay longer than a year in Scotland than in Northern Ireland.

Then I also calculated the relative population change values per country and plotted them against the relative net immigration so I could visualise what is the real impact of immigration figures over the relative population change (Details can be found in the notebook *Immigration\_Investigation*). I obtained the plot below.

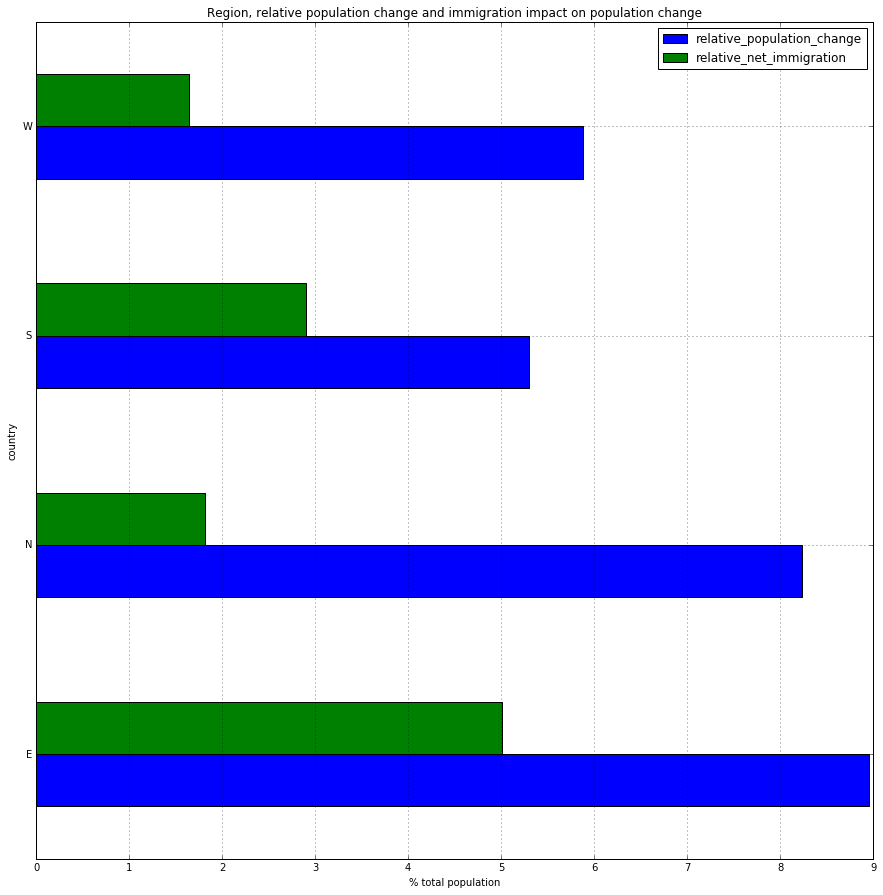


Figure 3 relative population change and net immigration

Looking at the plot we see that England and Northern Ireland are the countries whose population is increasing faster, although every country is growing over the 5% of its total population. However, when looking at which of these countries immigration figures have a higher impact, Scotland surpasses Northern Ireland for roughly one point. England is still first in both values, being the country with the highest level of relative population change and the net immigration figures.

When looking at the districts individually, I obtained a couple of conclusions that even though might seem intuitive, they need to be statistically proven in order to be accepted. The districts with higher figures for population change are also the districts with higher figures for population in 2014 and net immigration values. Both premises can be supported by the scatterplots below and the two Pearson’s R2 tests performed for both conclusions, resulting in values of 0.8 and 0.77 respectively and extremely small *p* values (Notebook *Immigration\_Investigation*).

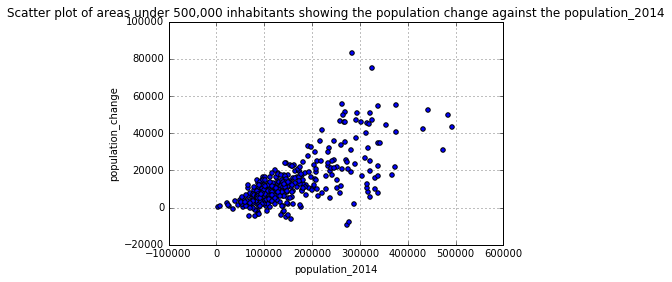


Figure 4 Population change against total population

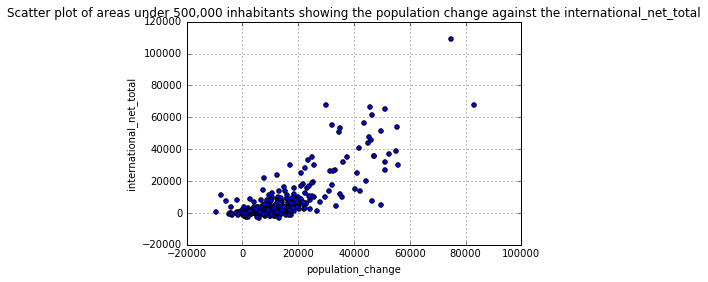


Figure 5 Population change against immigration net figures

Both scatterplots were produced only for districts under 500,000 inhabitants to better perceive a trend in the data.

## Immigration figures influenced the results of the referendum

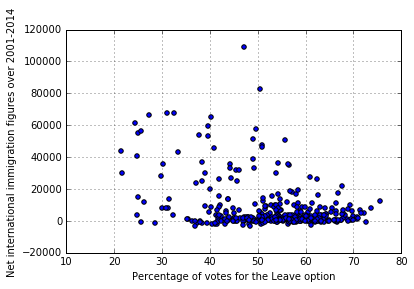
One of the main findings of my investigation was the negative correlation detected between the aggregate net international immigration figures and the vote for the leave option. I started sensing this possible correlation by looking at the scatterplot below:

Figure 5 Net immigration figures against % leave votes

To confirm my initial suspicions, I decided to perform a Pearson’s R2 test for correlation between this two variables. The process, which can be seen in detail in the Notebook I*mmigration\_referendum\_Pearson's\_R2* returned a p value of 6.98 x 10-14 and a correlation coefficient of -0.37 (2 s.f.), which indicates a moderate to low negative correlation. We therefore can affirm that even being a related factor, net immigration figures are not very tightly related to vote outcome.

In order to discover some interesting subgroup, I performed the *2*-means data mining technique on the two columns stated above, obtaining the results shown in figure 8 below.

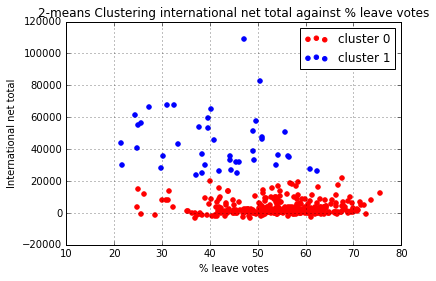


Figure 6 2-MEANS CLUSTERING INTERNATIONAL NET TOTAL AGAINST % LEAVE VOTE

The objective of the technique is to categorise data into groups of entries that are more similar between them than they are with the rest of the data.

As it can be seen in the figure, the *k*-means technique categorised the data using two clusters, one allocated to districts with roughly over 20,000 net immigration figures and the other one to districts below that amount. This classification somehow points at the finding that immigration figures do affect the distribution of the results.

## Districts with more than 20,000 net immigrants voted remain but had a lower turnout rates.

Once I had these two clusters, I annotated the Data Frame to look for further differences between both. Looking at the cluster 1 (district with over 20,000 net immigration figures), the first thing to note is how they voted in majority for the remain option, as it can be seen in the plot below:

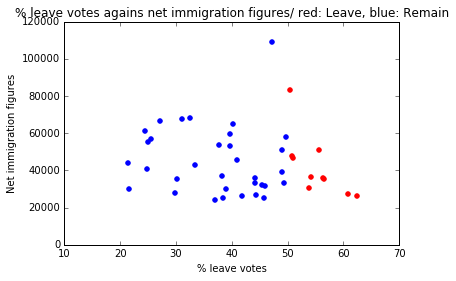


Figure 7 % leave votes against net immigration

When we count the number of districts voting leave in this plot we find that only 10 out of 43 districts did it, against 33 of them that voted remain (notebook *Immigration\_referendum\_anotated\_dataframe)*.

I also noted that not only the percentage of people voting for the leave option varies significantly but also the turnout rates. While the mean turnout rate for the districts within the cluster 1 is 74.5, for the ones in cluster 0 is 67.5. A total of 8 points below. Difference can be perceived by region in the Figure 8.

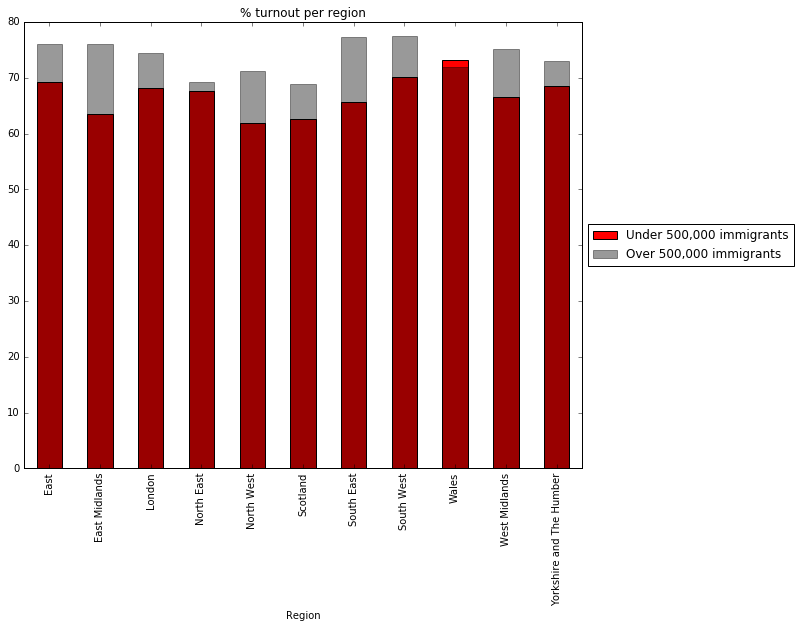


Figure 8 % Turnout per region

# Conclusions and recommendations

In order to provide with a better support to migrant population in the UK, our NGO should take note of the following:

* Every country in the UK is currently receiving high figures of international immigration.
* There is a slight difference between of incoming and net immigration.
* The fact that districts with more emigration voted remain over leave might indicate that immigration played a role in the decision
* Turnout rates were lower in those districts with high immigrants. Why did this happen.

To face this issues and plan for future actions, your NGO should continue analysing other datasets in order to find out how is the political involvement of migrant population, how are they characterised demographically, what plans are working in districts that receive more immigration, etc.

# Reflections

The referendum results data was processes directly from the csv file into interactive Python notebooks and did not require any special consideration regarding missing or dirty data, as described in the TMA 02.

The second dataset was a little bit more complicated to work with. Even though it was also small enough as to be imported directly into a Python Notebook, there was a good deal of missing data and format issues that required some cleaning. The dataset was also large enough as to result quite complex. The approach I took was aggregating the data from diverse years into a single column so that my work with it would ease up considerably.

When facing missing data, I considered using data mining techniques to obtain an approximation of the missing values. However, I ended up deciding that it will be easier and more accurate to just exclude the missing data columns entirely from the analysis.

The reason behind storing data in notebooks instead of using a database system was mainly practical, since it was the technique most used during the module and the one I was more conformable with. I stored intermediate results in csv files that were used by other notebooks.

Near the end of my analysis I started wondering if it could have been easier to work with a different dataset. From my investigation, I ended up thinking that the conclusions and findings are not excessively original nor striking. It somehow turned out quite difficult to elaborate a good report without this element.

Word count – 3008 words (excluding headers, references and appendices)

# References

* The Electoral Commission (2016, [a]) ‘EU referendum results’ [Online]. <http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information> (Accessed 15 April 2017).
* The Electoral Commission (2016, [b]) ‘The 2016 EU Referendum’ [Online]. Available at <http://www.electoralcommission.org.uk/__data/assets/pdf_file/0008/215279/2016-EU-referendum-report.pdf> (Accessed 15 April 2017).
* Express (2016) ‘EU referendum 2016: Who was allowed to vote on Brexit?’ [Online]. Available at <http://www.express.co.uk/news/politics/649517/EU-referendum-2016-Voting-Voters-Allowed-British-Irish-Commonwealth-Citizens-European> (Accessed 15 April, 2017).
* Office for National Statistics (2016) ‘Population Estimates for UK, England and Wales, Scotland and Northern Ireland’ [Online]. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland> (Accessed 15 April, 2017).
* BBC Website (2016) ‘EU Referendum’ [Online] <http://www.bbc.co.uk/news/politics/eu_referendum/results> (Accessed 15 April 2017).
* The Guardian (2017) ‘EU Referendum and Brexit’ [Online]. Available at https://www.theguardian.com/politics/eu-referendum (Accessed 15 April 2017).

# Appendix 1 Notebooks

|  |  |
| --- | --- |
| *Notebook* | *Contents* |
| [Immigration\_Import\_cleaning.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_Immigration_Import_cleaning.ipynb) | Import and cleaning of the population estimates data alone |
| [Immigration\_Investigation.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_Immigration_Investigation.ipynb) | Main investigation of the population estimates data alone |
| [Immigration\_referendum\_data\_mining.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_immigration_referendum_data_mining.ipynb) | *k*-means data mining technique of the merged data |
| [Immigration\_referendum\_Pearson's R2.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_immigration_referendum_Pearson's%20R2.ipynb) | Pearson’s R2 test of the merged data |
| [Immigration\_referendum\_turnout\_rates.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_immigration_referendum_turnout_rates.ipynb) | Investigation on the turnout rates |
| [Immigration\_referendum\_ anotated\_dataframe.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_immigration_referendum_work_with_anotated_dataframe.ipynb) | Work with the annotated Data Frame |
| [Merging\_referendum\_population.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_EMA_Merging_referendum_population.ipynb) | Merging the two datasets |
| [c8666430\_project\_diary.ipynb](http://127.0.0.1:35180/notebooks/c8666430-tm351-tma02-16j/c8666430_project_diary.ipynb) | Project diary |
| c8666430\_TMA02\_Question2b.ipynb’ | TMA 02 Question 2 notebook |

# Appendix 2 Data catalogue

The ‘Population Estimates for UK, England and Wales, Scotland and Northern Ireland’ dataset combines several csv files containing different demographic columns such as sex, ages, immigration figures and population estimates However, this analysis focuses on the MYEB3\_summary\_components\_of\_change\_series\_UK\_(0215).csv file. The files were obtained from The Office for National Statistics website at <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland> [accessed 15/04/17]. The data held in the Office for National Statistics websites in licensed under a Open Government License v.3.0 (http://webarchive.nationalarchives.gov.uk/20160105160709/https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/) which allows users to

* Copy, publish, distribute and transmit the information
* Adapt the information
* exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.

but forces the user ‘to acknowledge the source of the Information in your product or application by including or linking to any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence’

The data dictionary for the dataset is included within the ukmye2015 folder in a file named MYEB\_information\_note.pdf.

The referendum results data contained only one file with information on the outcome of the UE referendum for each district in the UK. The files were obtained from http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information [Accessed 15/04/17]. The data is held under an Open Government License v.2.0 (http://webarchive.nationalarchives.gov.uk/20160105160709/https://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/), which is similar to the v.3.0 stated above regarding what it allows users to do. It only differs slightly on the wording of the responsibility of the users when using the information provided: ‘acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence’.

# Appendix 3 TMA 02 Question 2

* Aims and objectives

This study investigates the results of the June the 23rd of 2016 referendum on the UK’s membership of the European Union. Using the dataset, we will ask questions and pose hypothesis regarding the results of the referendum, vote patterns across different areas and regions and other factors variables such as turnout rates.

* Background

Last year British politic scenario was strongly influenced by the upcoming referendum on the UK’s membership of the European Union. Different media (The Guardian, 2017) and governmental organisations covered the pre, during and post Brexit referendum down to the second.

* Sources of data

The dataset used for the investigation was obtained thought the UK Electoral Commission website (The Electoral Commission, 2016 b) and covers the relative and absolute number of leave and remain votes within every region and area in the UK (except Gibraltar) along with other variables such as turnout rates, electorate, and rejected ballots (and the reasons of the rejection).

* Analysis pipeline

The analysis phase of the data pipeline had place mainly in the form of data mining. The objective was to find a story with the data presented by the dataset. An initial approach to the dataset using the method describe() gave us some insight into which columns (such as *Leave*, *Remain*, *Area* or *Pct\_Turnout*) would turn out as useful to our investigation.

Since the dataset does not include any demographic information about who cast the votes (data is anonymised by aggregation) I had to limit myself to analyse vote patterns across regions. So I proceeded grouping data by regions and then projecting it along with several different columns, such as result percentages, electorate inscribed or turnout rates. Figure 9 shows one of the Data Frames obtained using such method.

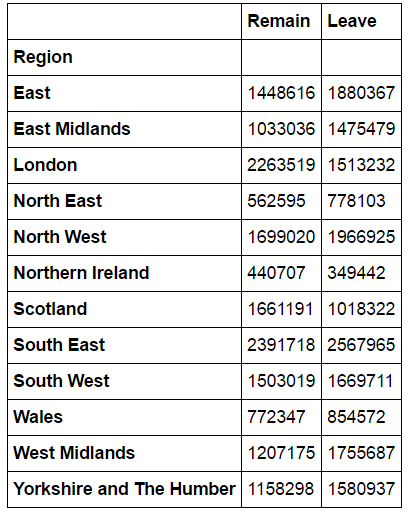


figure 9

The second strategy I used across the notebook was performing merges of the DataFrames obtained by grouping the Data. I used this method to compare number of votes for leave or remain based on turnout rates or electorate numbers for each region. Figure 10 shows the result table from one of these merges.



figure 10

* Findings

If we have a look at the number of votes for each region, it seems very clear that the remain option won in only 3 out of 12 regions (Figure 11)

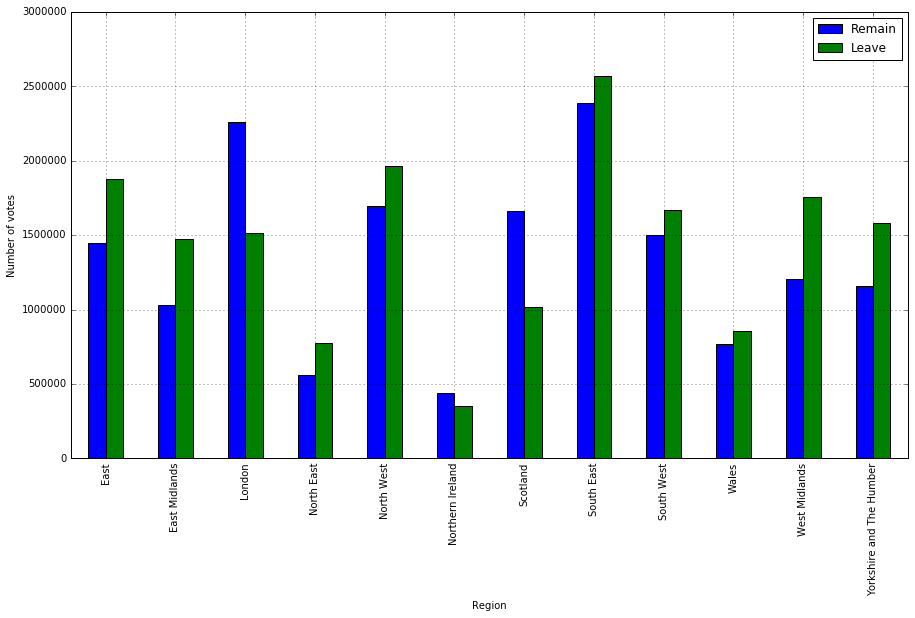


figure 11

Investigating this distribution further I found something that could use some deeper analysis: The regions where remain won are amongst the ones with the lowest turnout values. This can be observed in figure 12. The fact that we do not have many other indicators makes it hard to research this question in depth.

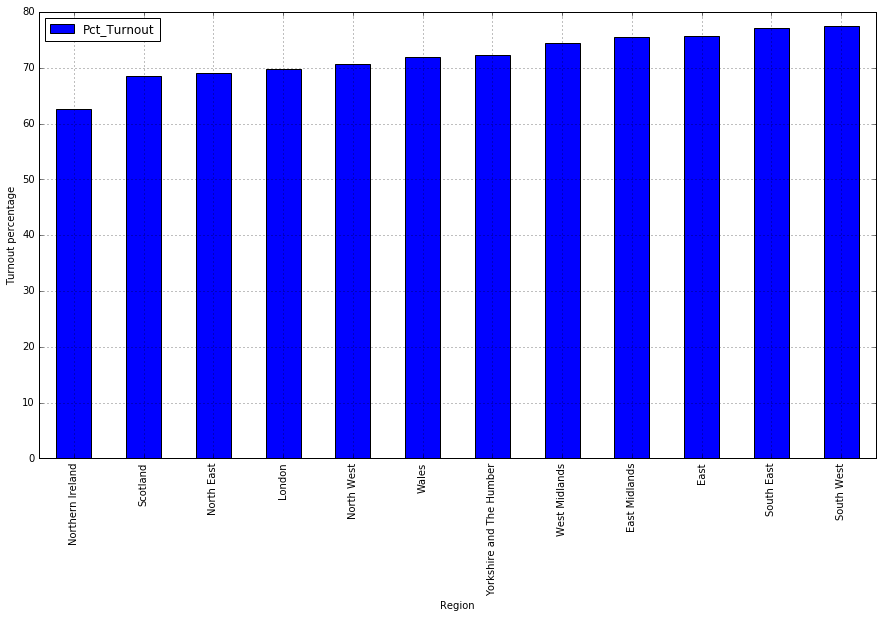


figure 12

I wanted to check if the size of the electorate of each region had any effect on which option they voted for. On a first attempt I plotted the % of leave and remain votes on an axis and the size of the electorate on a separate one to see if any pattern was discernible. However, the image was not very significant (Figure 13)

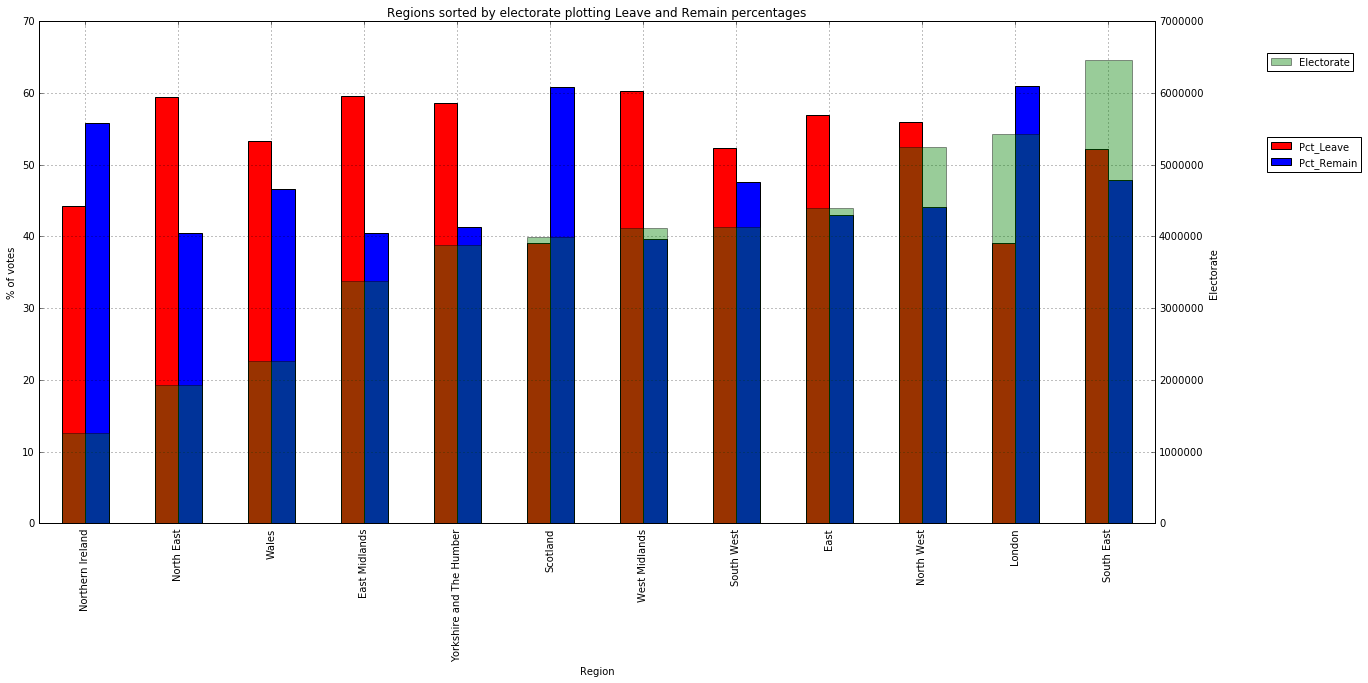


Figure 13

Still, I wanted to test this hypothesis a bit deeper, so I performed a Pearsons’ *R2* test using the electorate and the remaining columns. I obtained a p value of 0. 003.. and an *R2* of 0. 148... These values tell us that the test is statistically significant and suggest a small correlation between the size of the electorate in each entry and the likely to vote remain.

* Conclusions

The data seems to indicate some correlation between the size of the electorate and the vote for the remain option. It also shows how regions where remain won display some of the lowest turnout rates.

In order to analyse the dataset further, we will need to supplement it with additional datasets on demographic and socio-economic data. Also, the investigated dataset is missing relevant information, such as the referendum results for Gibraltar.

* References

References can be found at the end of the document