Brexit Visualization

Students: Dario Benvenuti, Alba Puy Tapia

Sapienza, Università di Roma

benvenuti.1562938@studenti.uniroma1.it, alba.puy.tapia@gmail.com

06/07/2018

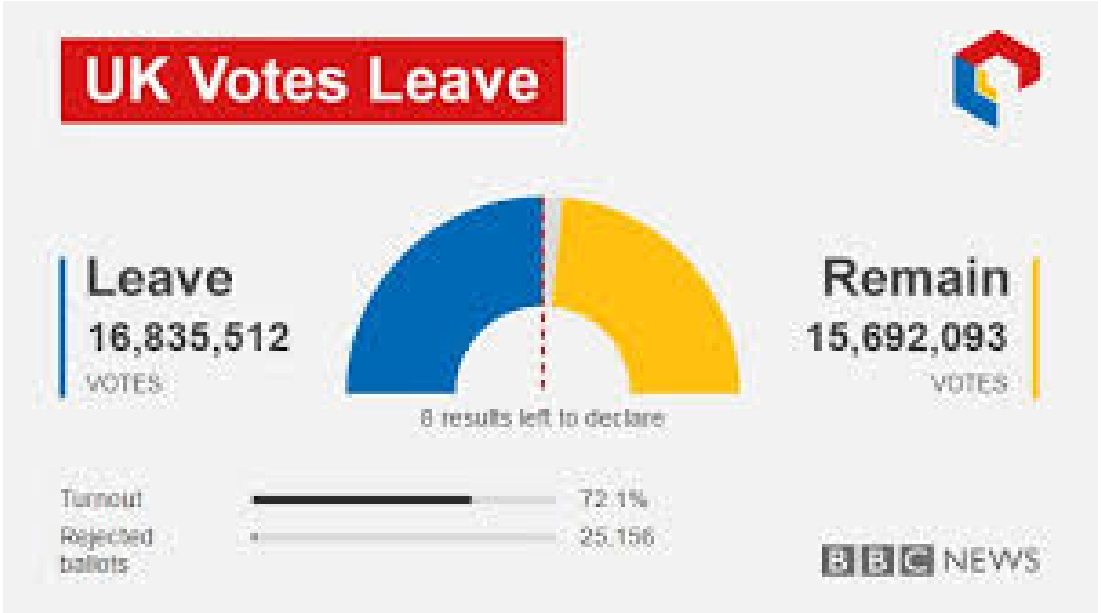
Overview

Historical contest

- David Cameron [1]
- 23rd June 2016 [2]

EU REFERENDUM
BBC

Result



$\Rightarrow$ *LEAVE*

Two datasets:

- Census
- Referendum

$$\Rightarrow Brexit\_data$$

Dimension of our dataset $= (375, 38)$, getting a value $375 \times 38 = 14250$

Some plots

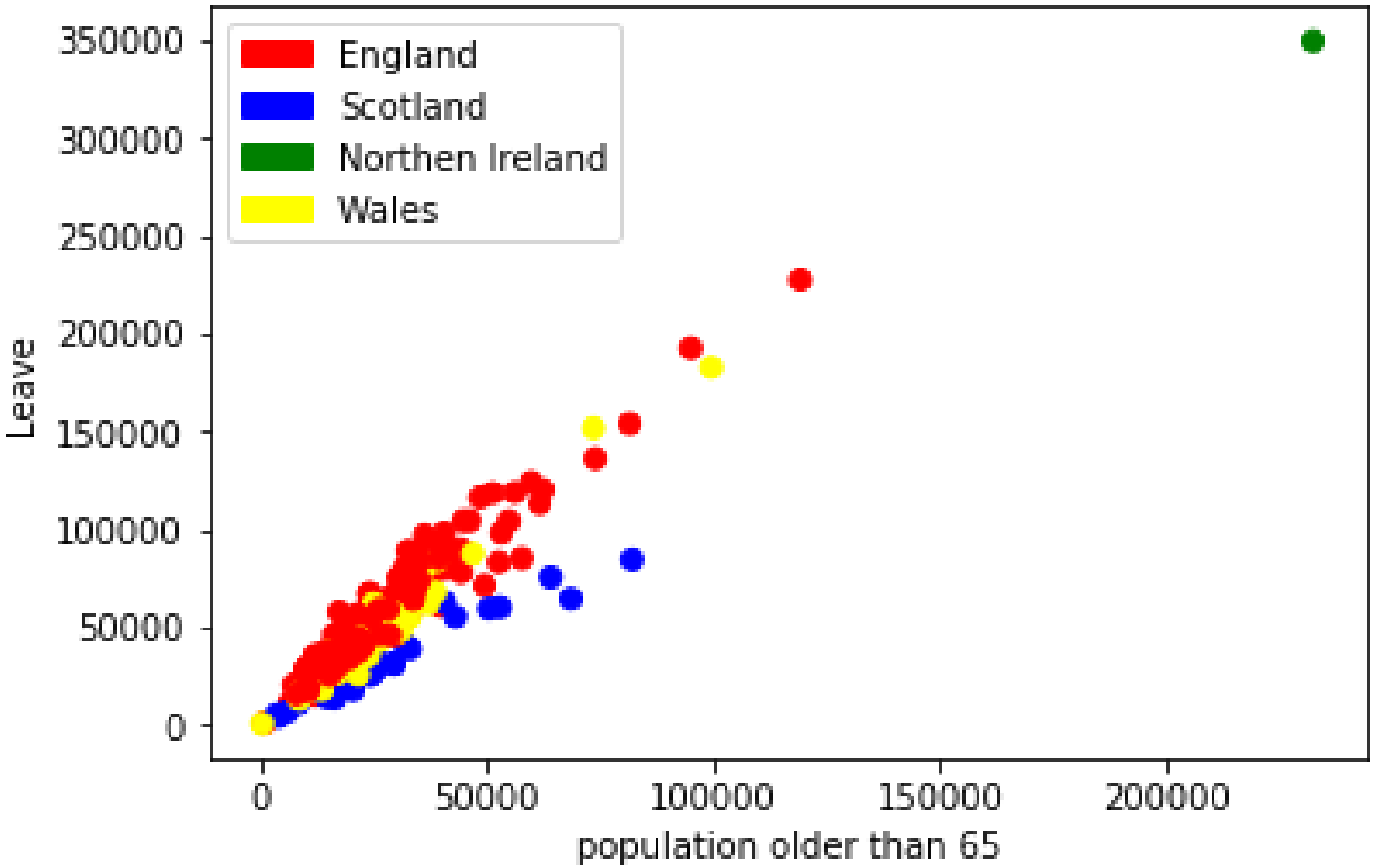


Figura: Old part of the population

Some plots

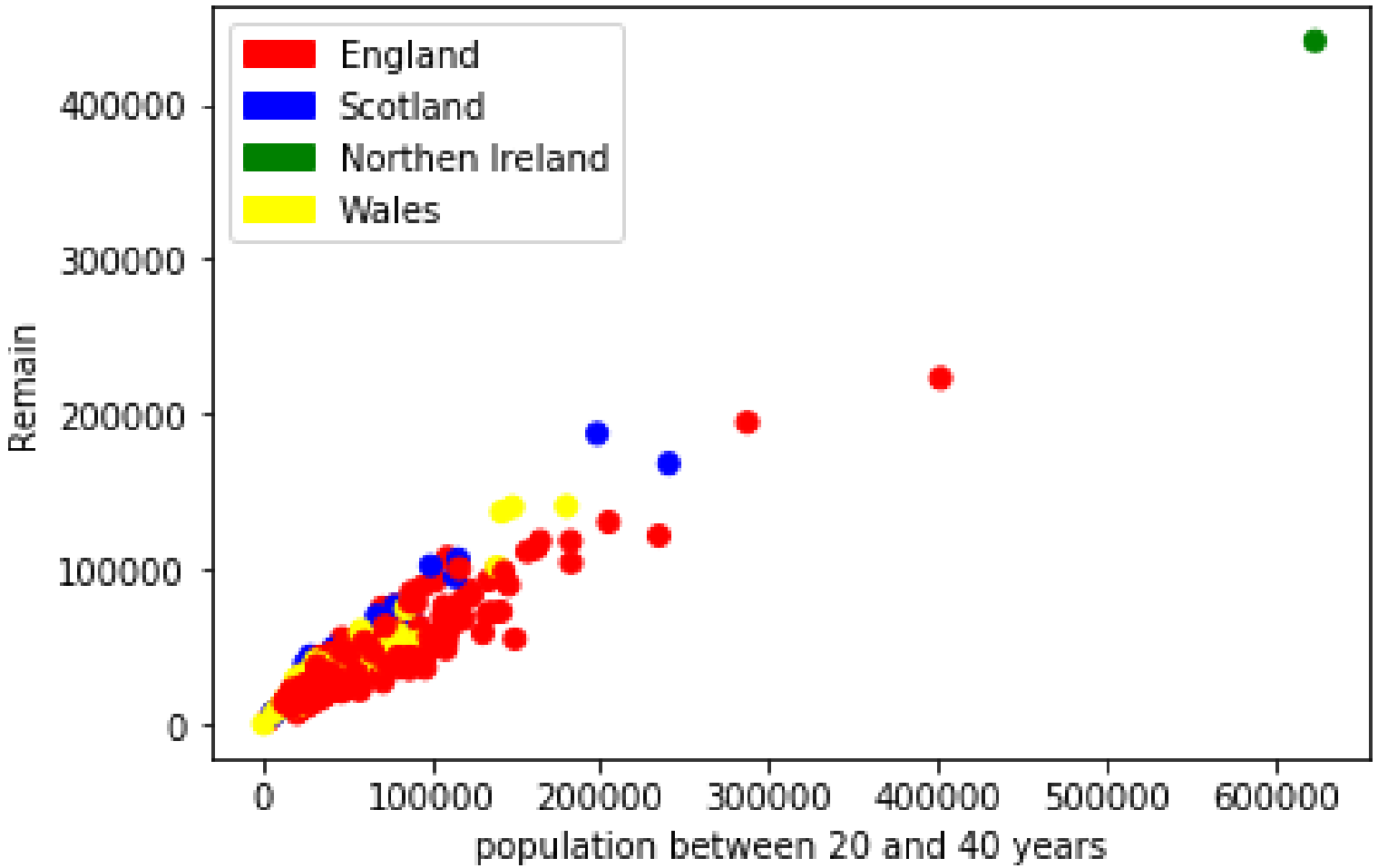
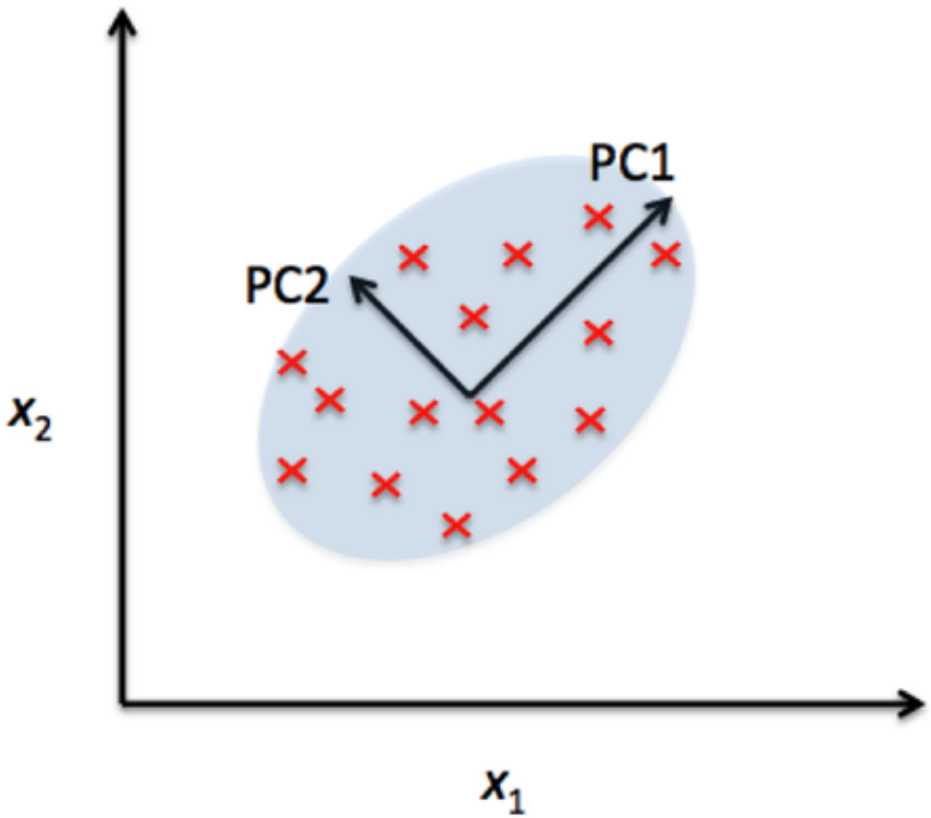


Figura: Young part of the population

Principal Component Analysis

PCA projects the features onto the principal components. The motivation is to reduce the features dimensionality while only losing a small amount of information.

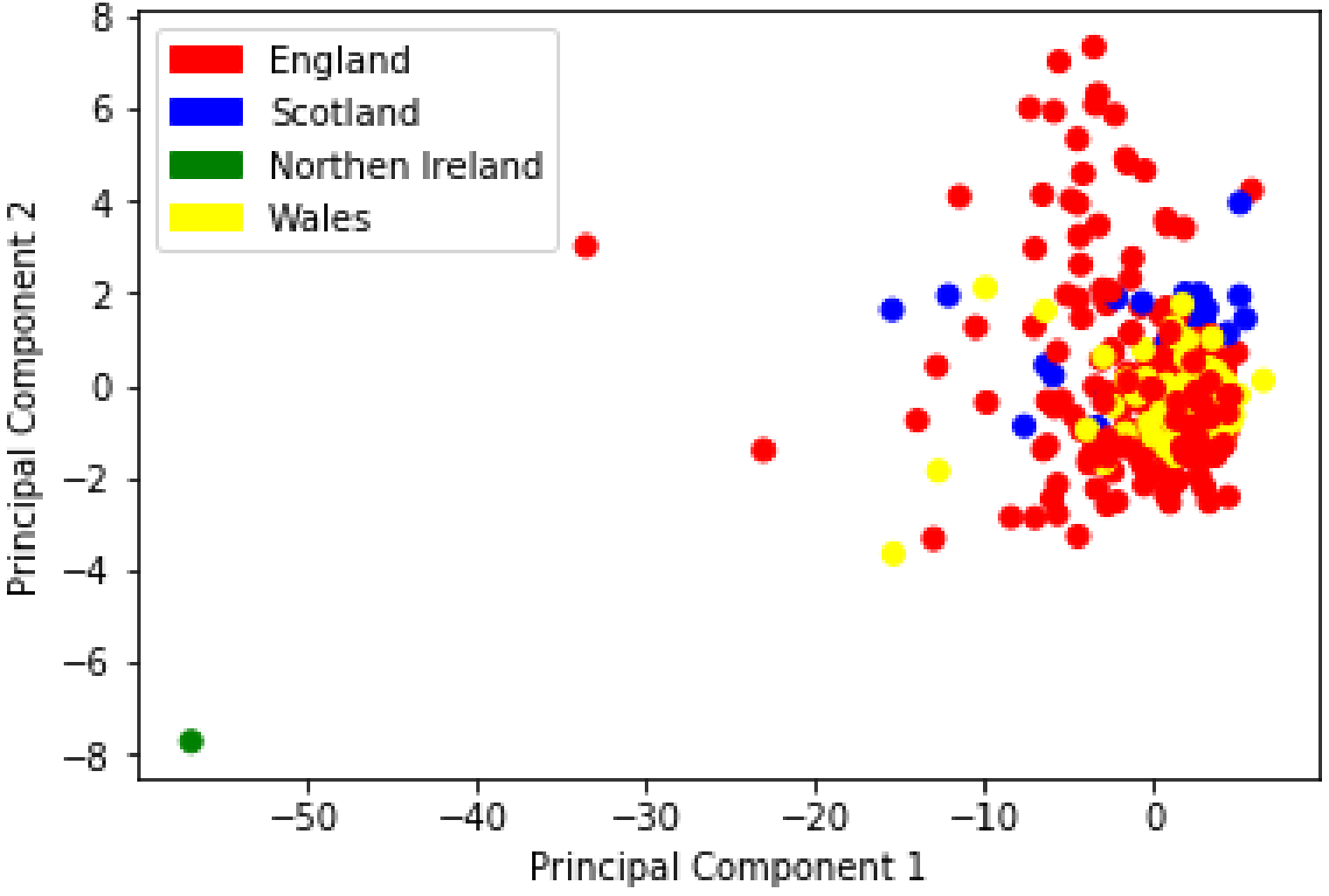


PCA in python

$\Rightarrow$ Only numerical data

- Using mlabPCA
  - data as an array
  - no need to standardize
  - $.Y[:, 0]$ gives us directly the first coordinate of all our data points.
- Not using mlabPCA:
  - We have to convert our data $\Rightarrow$ Standardize
  - Compute the eigenvalues and eigenvectors of the covariance matrix of our numerical data.
  - Eigenvector with largest eigenvalue is the first principal component.

Applying PCA (using mlabPCA)



Deleting some variables (not using mlabPCA)

PCA components till	cumulated variance
1	0.77
2	0.86
3	0.90
4	0.92
5	0.95
6	0.97

Deleting some variables (not using mlabPCA)

PCA components till	cumulated variance
1	**0.77**
2	**0.86**
3	0.90
4	0.92
5	0.95
6	0.97

Deleting some variables (not using mlabPCA)

$\Rightarrow$ LOADINGS

Let's define λ_j is the jst largest eigenvalue with eigenvector v_j. The importance of the variable i in the jst PCA component is:

$$importance\_PCA_j[i] = abs\{\sqrt{\lambda_j} * v_j[i]\} \in [0, 1]$$

Deleting some variables (not using mlabPCA)

$\Rightarrow$ Variables that does not have an importance bigger than 0.75 in none of our two first PCA components.

delete={'No Official Mark', 'Percent Turnout', 'Writing or Mark'}

Dimension of our dataset $= (375, 35)$, getting a value
$375 \times 35 = 13125$

Goals

The visualization that we come up with tries to achieve 2 goals:

- Let the user visualize easily the correlation, if there is any, between the result and the age of UK's population.
- Let the user visualize easily the correlation, if there is any, between the result and the geography (physical or cultural) of the UK.

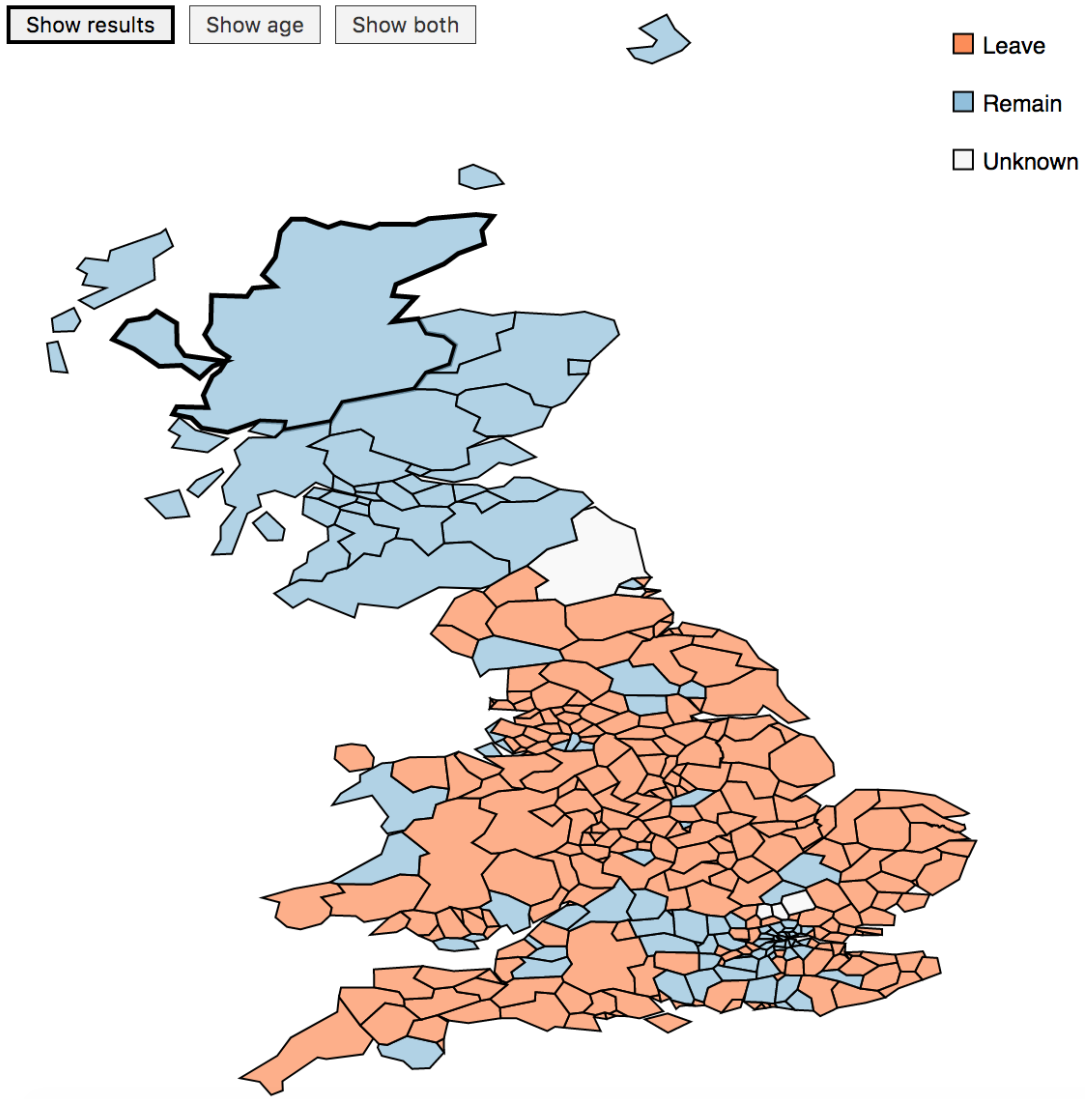
Visualizing general information about UK

In order to achieve these goals, first we show the map of UK to the user, with some data codified on it through color schemes taken from colorbrewer [3]. Next, below the map, we show to the user a simple bar chart showing the how the whole UK's population is distributed wrt various age intervals.
Hovering the mouse over one bar, will project it on the Y axis, to make the user be able to easily understand the amount of people belonging to that age interval.
The user can decide to visualize on the map:

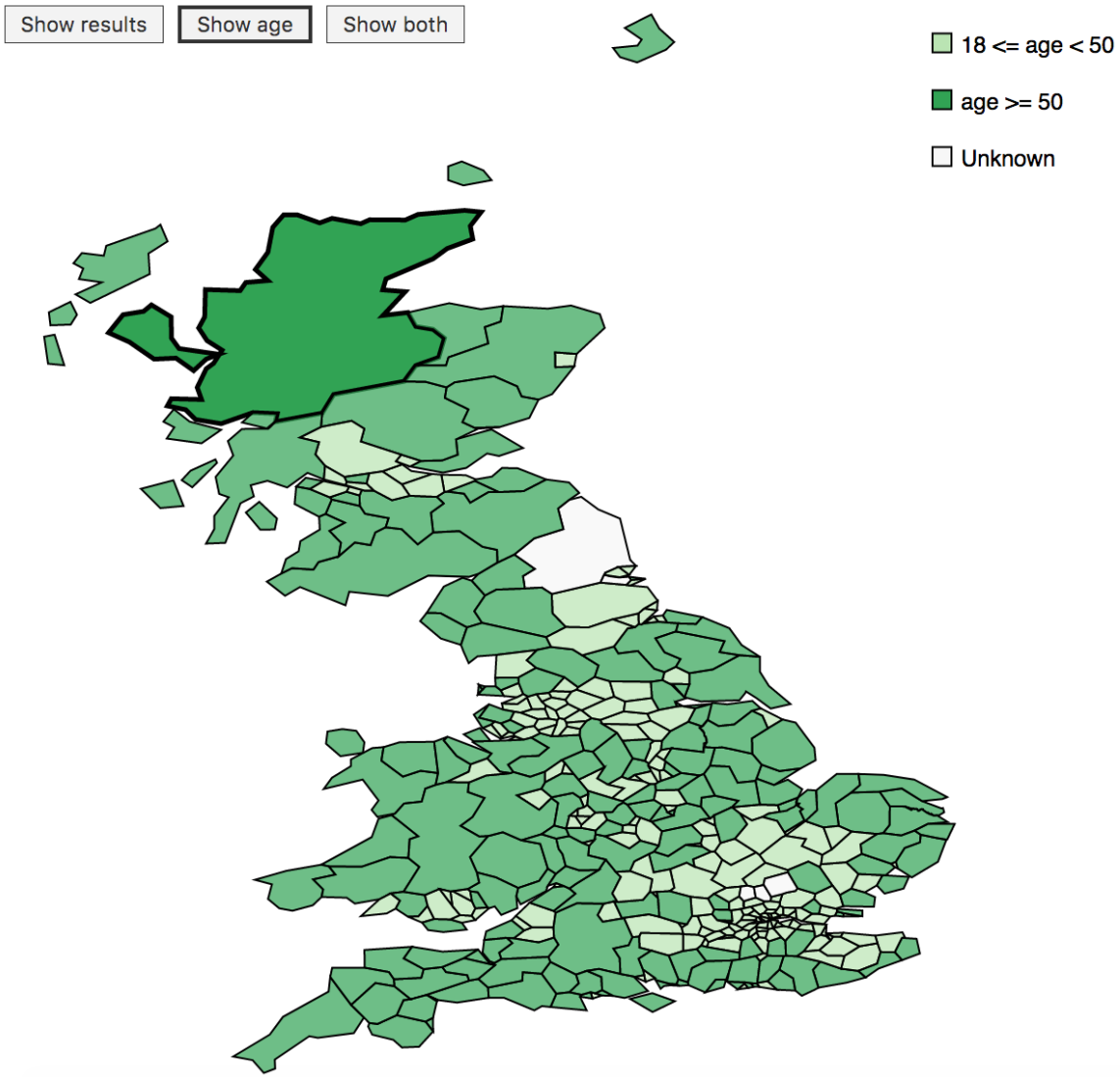
- On each region, if the result was to leave or to remain.
- On each region, if the majority of people is young or "old"[4].
- On each region, the combination of previous data.

The user can switch between these codifications through simple buttons.

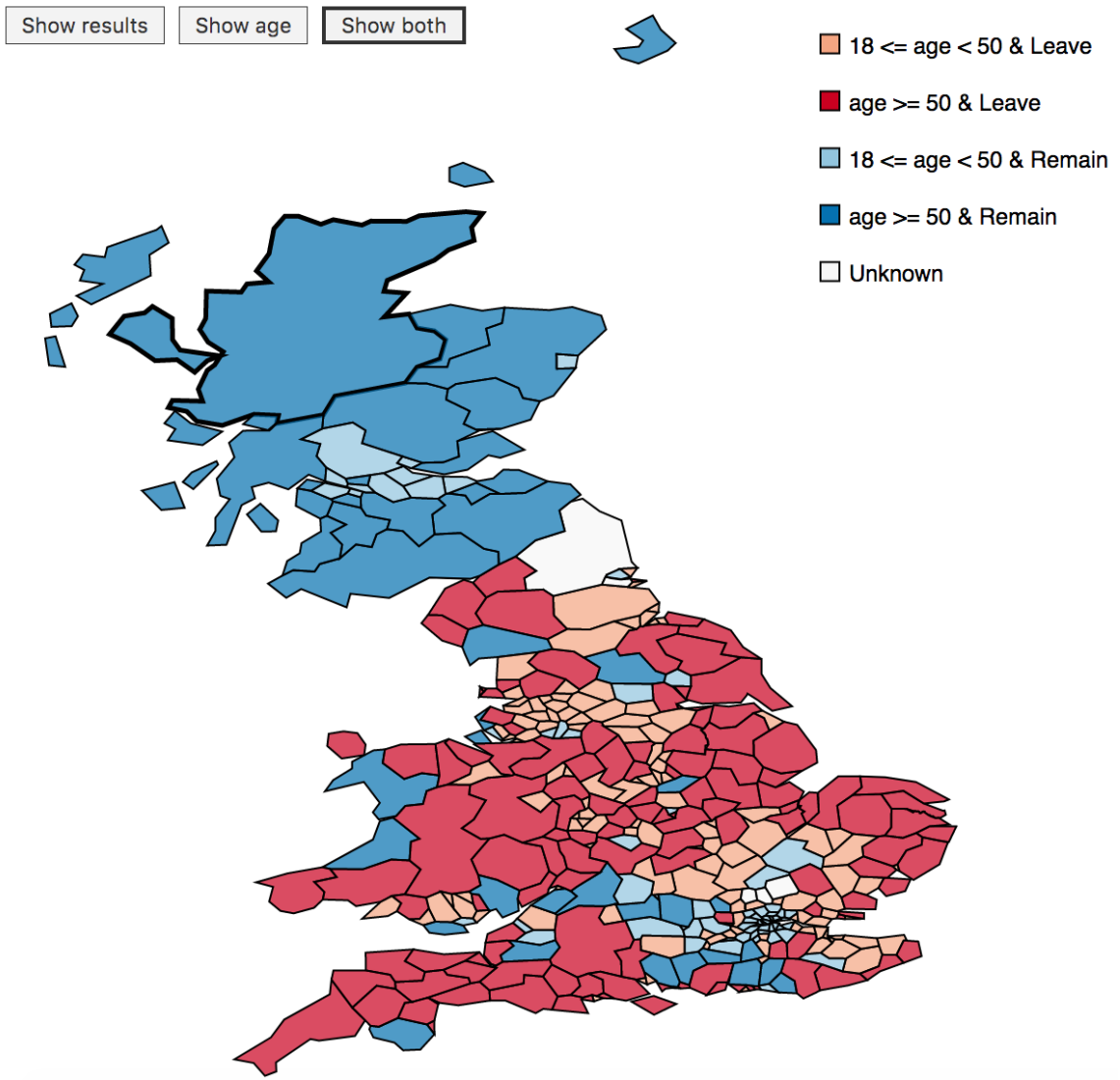
Map showing results



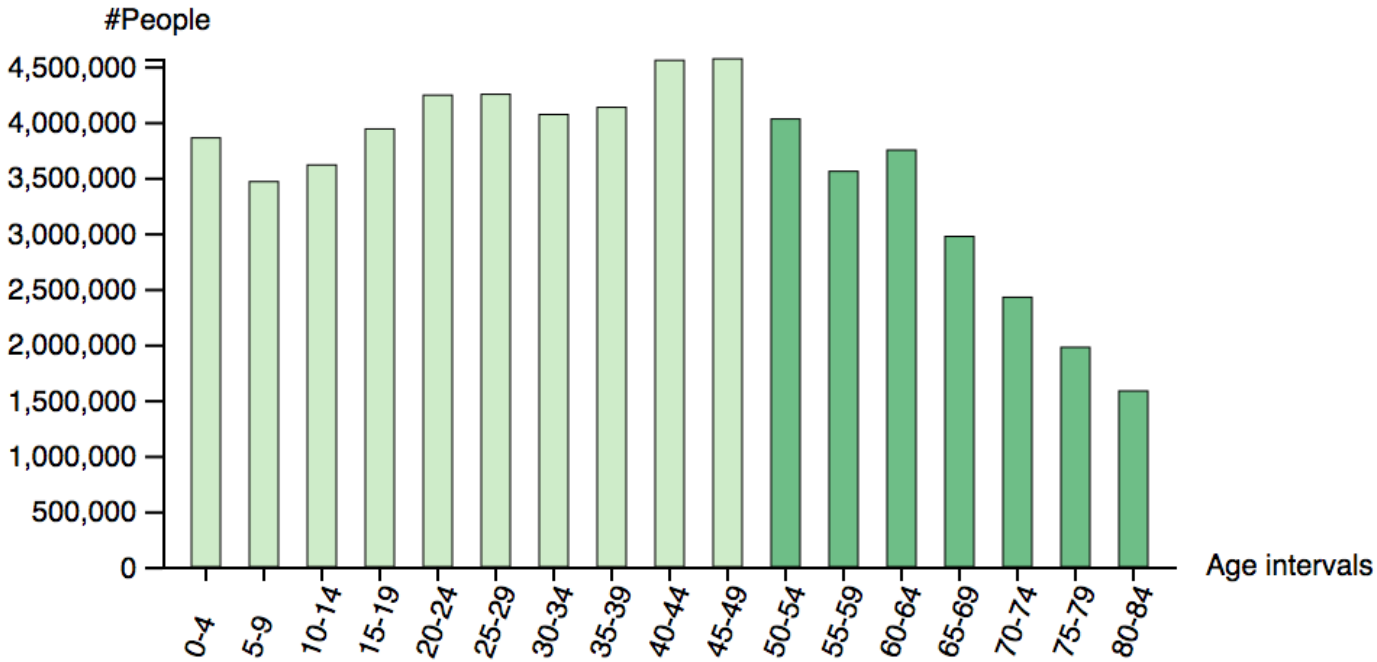
Map showing age



Map showing results and age



Bar chart showing population's distribution wrt age



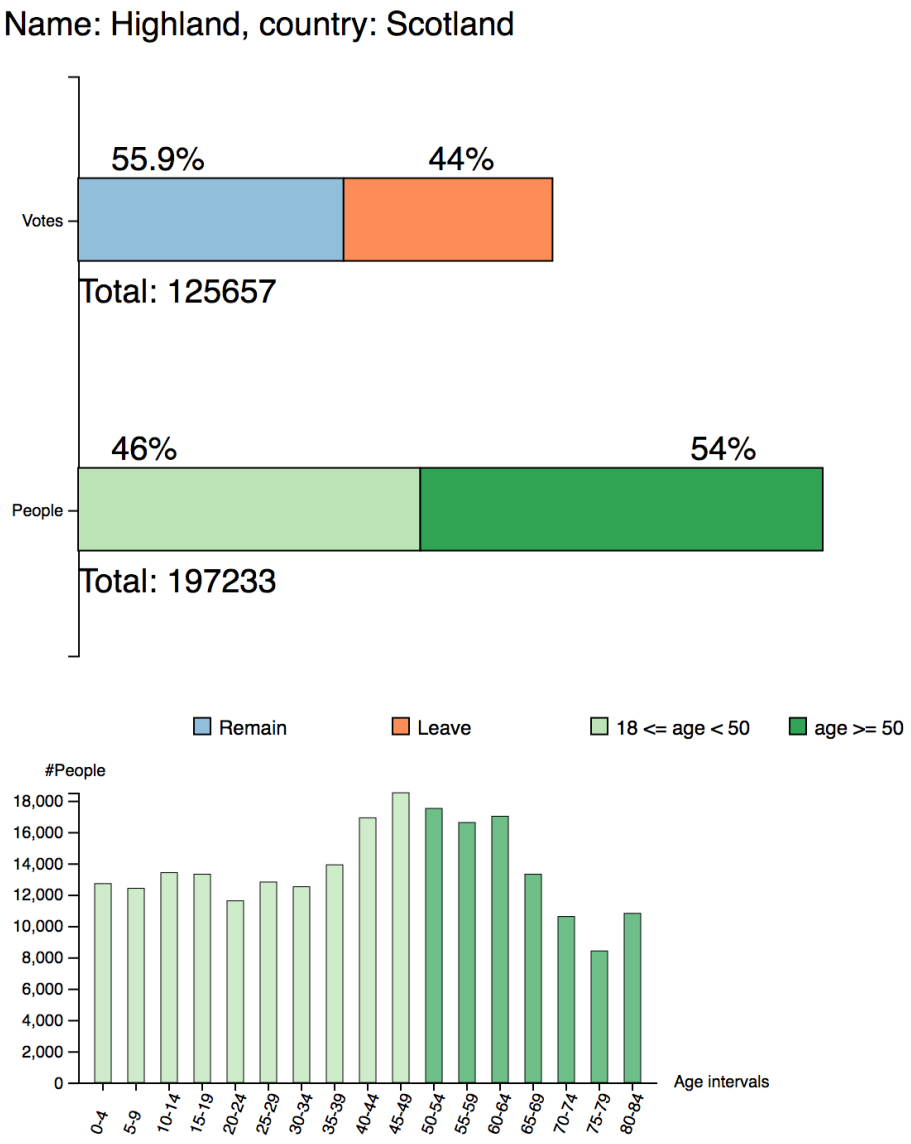
Visualizing regions' details - 1

Clicking on one region, the user can see, in another visualization, on the right of the map, more details about that specific region:

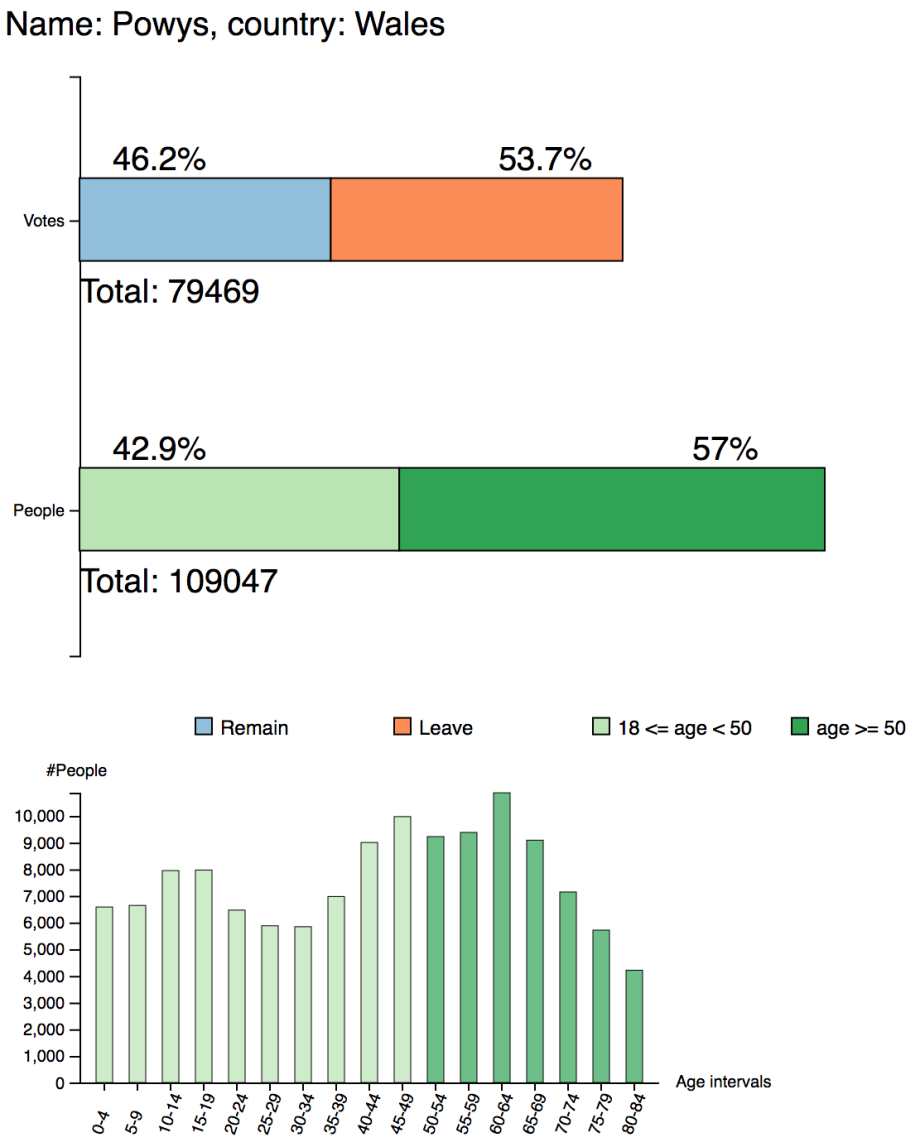
- Total number of people.
- Total number of votes.
- Difference, through percentages, between the amount of young people and old people, and between the amount of votes to remain and to leave.
- How the population, in that region, is distributed wrt various age intervals.

Through two bar charts, one horizontal and one vertical, the user will be able to easily evaluate all these information. Hovering the mouse over one bar in the vertical bar chart, will project it on the Y axis, to make the user be able to easily understand the amount of people belonging to that age interval.
By default, Highland is selected automatically when the user loads the page.

Visualizing regions' details - 2



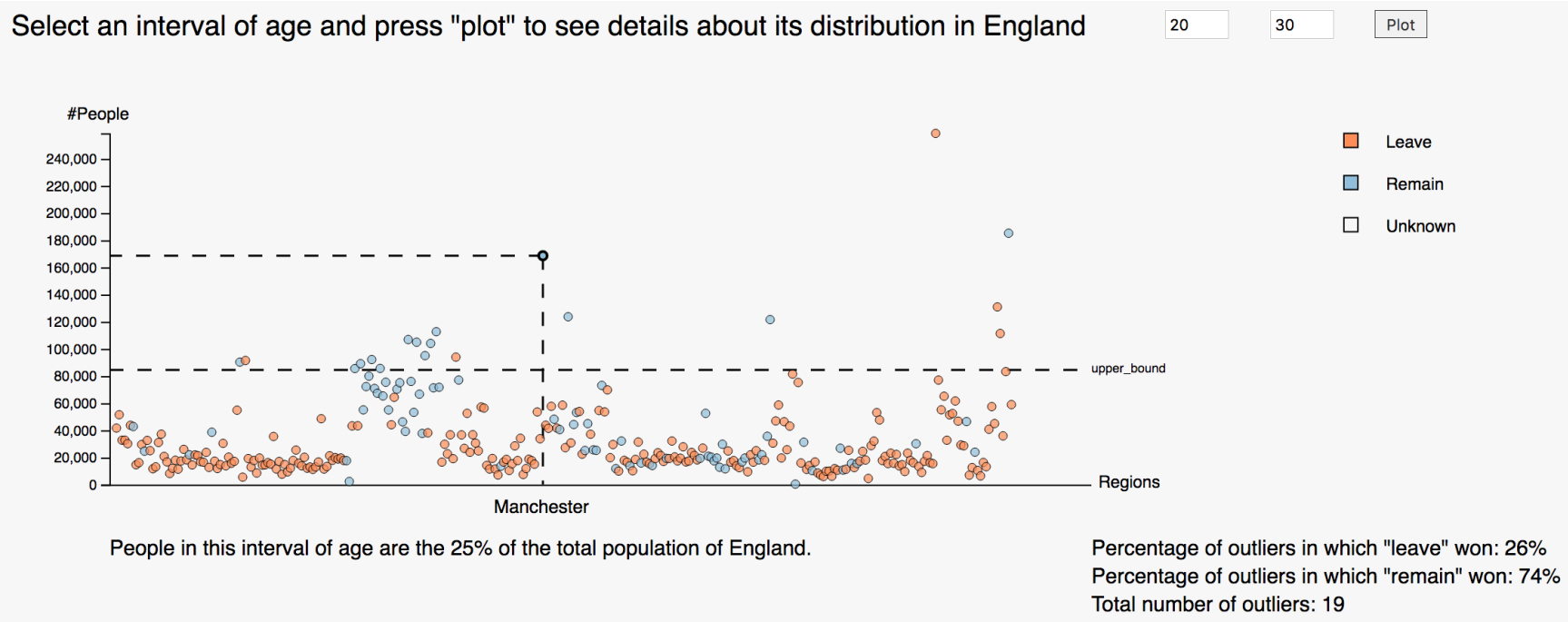
Visualizing regions' details - 3



Getting general information

- Looking at the UK map showing both the results and the age of the population, it's clear that in England the situation it's quite more complex than in Wales and Scotland.
- In order to understand better if there is really a correlation between the age and the result, we will let the user choose an arbitrary interval of age.
- We will, through a scatter plot, let him see how the population belonging to that interval of age is distributed among England's regions, and which is the result in those regions.

Analytics view



- On the Y axis we have the amount of people in the interval.
- On the X axis we have the regions, and the colors represent the result.
- Hovering the mouse on one circle will project it onto the axis, showing the amount of people from the interval in that region and the region's name.

What can be understood

From the analytics view the user can see:

- The weight that people in this interval had on the referendum.
- Upper outliers: regions with a really high amount of people in that interval, significant wrt the population; if the majority of the outliers' region has one particular result, it could mean that indeed votes depend from age.

Quick facts about the code

This software has been developed following XP [5] and some common knowledge from software engineering:

- We let the user write his own starting and ending value for the age interval. To avoid allowing him to lead the software in error states, we let him write only 2 characters. Every category of error in the input has been analyzed and for each of them, and for each possible combination of them, a precise message will be shown to the user.
- The source code has a density of comments [6] value of 0,23.
- The software has been successfully tested by 5 users with no expertise in the visual analytics field.

The visualization in action

We can have a look at the visualization in action, to see better how it works and how the software flows.

Discussion - Lessons learned from the study

Interesting features discovered exploring the data:

- In Scotland the vast majority voted to remain, independently from age.
- In Wales the population it's quite old.
- In England young people voted to remain, but their weight on the result was too little.
- The whole UK's population, wrt to age, got 2 peaks: one between 0 and 4 years, and the other around 50 years.

Discussion - Extending the visualization

In the future, this visualization can be improved to:

- Let the user filter on region/country/area.
- In the analytics part, let the user choose one value to maximize, and automatically find the interval of age that does it. For example the user could choose to maximize the number of outliers in which "remain"won.
- Give to the user the possibility to zoom on the map.
- Add Northern Ireland to the map.
- Add missing data.

References

- David Cameron, on Wikipedia
- Brexit, on Wikipedia
- Colorbrewer2
- Psychology today article
- Extreme programming, definition from Agile Alliance glossary
- Density of comments, definition

Thanks for your attention