Brexit: A Visual Analytics Approach

Author: Tareq Galala

Abstract: In this paper, we look at the 2016 referendum results data. First, will be visually exploring the spatial distributions for different variables from the 2011 U.K. census that we assume to be discriminating, aggregated to local authority level and we will identify the key drivers in England behind the leave vote and compare it with London. Second, will be looking into patterns and relationships using a visual analytics approach. And third, we will use regression analysis to capture how did different ethnics vote in different regions of England using national identity and what characteristics they hold. We will showcase a combination of visual techniques and computational methods across this paper.

Index Terms: Brexit, Referendum, National Identity, Spatial Distribution, Regression Analysis, Visual Analytics

1 Problem Statement

Brexit was the defining matter of 2019 elections, with the conservatives 'Get Brexit Done' slogan and labour promising a second referendum. To understand this election, we should go back to the 2016 referendum which resulted to leave the EU. The unexpected results of the Brexit referendum were a shock to many and better understanding the cause of this phenomenon is important, and identifying the reasons behind it. Our research questions are:

1. Is the characteristics of the voting population key drivers of the leave vote in 2016 referendum in England?
2. How does that compare to London?
3. How did ethnic minority groups vote in different regions of England?
4. What are the characteristics of different ethnic groups communities and their levels of support for the leave or remain vote?

These questions are the focus of our analysis. The approach used here is to understand the characteristics variables from the 2011 Census and national identity. This might help us understand our study on a Local Authority (LA) level in England and in London.

2 State of the Art

A number of existing papers have examined the referendum results such as by Beecham, Slingsby, and Brunsdon [1]. They demonstrated if Leave vote explanation holds equally to different parts of Britain using geographical-weighted statistics and spatial distribution of residuals with Lasso selection. In addition, they used the variable EU-born as a proxy for diversity and compared it with Leave vote. Another paper discussing the 2016 referendum is by Goodwin & Heath [2]. Similar to Beecham, Slingsby, and Brunsdon, they used multivariate models and used the leave vote proportional to the total vote on a LA level. These papers are very valuable in understanding computational methods and in generating valuable variables that could be discriminating which could lead to hidden patterns in the data. Goodwin saw that the highest effect in explaining the vote Leave was the education variable between local authorities. This is an extremely important finding in the understanding of the referendum results and it is of our interest to find other methods that can confirm this result.

Focusing on England, this paper explores the characteristics of the voting population key drivers of the vote leave in 2016 referendum, and looks into differences between London and the rest of England. A number of questions regarding the Leave vote remain to be addressed, specifically about how other ethnic groups voted in the referendum. The majority of prior research has focused on the white ethnicity to explain the Leave vote, but we are interested to see how did other ethnic groups vote and did this vary geographically. A general behavior will be explained on a local authority level but it's important to understand that this will not be on an individual-level analysis. We will take the share of vote in favour to Leave as our single outcome. We will use visual and computational analytics techniques to try to understand our questions and look into details and focus on the outcome variable vote leave.

3 Properties of the Data

The data used here is sourced from the 2011 UK census to which it contains important key sociology-economic variables for the whole UK, although we will be focusing on the data related to England only. This census is taken every ten years and provides a detailed statistic of the population and its characteristics. The census will be merged with the national identity census and then merged with the referendum results which are available only at local authority level (LA) and describes voting results in the UK referendum 2016. Also, we will import shapefiles that contain boundaries files for England and London to show choropleths and make comparisons. The census data contains 72 variables and the referendum result data set contains 7 variables. The two files will be merged through a lookup file to join them on a LA level. The outcome variable is the share of vote Leave on a Local Authority (LA) level. Table 1 shows several variables that are grouped, created and normalized with respect to the population in order to compare the differences in magnitude, as the raw numbers would not be helpful. In figure 1 we observe the distribution through their quartiles of some of the most discriminating variables.

The office of the national statistics issued the following statement on the assurance of the census 2011 data quality:

'The 2011 Census population estimates have been through a rigorous quality assurance (QA) process to ensure that users of census data have confidence in the quality and accuracy of the information.'

Variable	Description
adults	Residents from 45 to 84
young_adults	Residents from 20 to 44
white	All white ethnicity race
asian	All Asian ethnicity race
black	All black ethnicity race
mixed	Mixed ethnicity race
arab	All Arab ethnicity race
christian	Christian religion
no_religion	No religion
english_speaking	Can speak English
no_english	Can't speak English
homeowners	Households owning the house they occupy
private_rented	Households renting privately
social_rent	Households renting from council
no_rent	Residents not paying rent
not_good_health	Not good health
degree_educated	Qualification at undergraduate level or above
no_car	Households with no cars available
Professional	Professional jobs & high earners
British only	Residents who are considered British only
English only	Residents who are considered English only

Table 1. Demographics Variables normalized with the total population

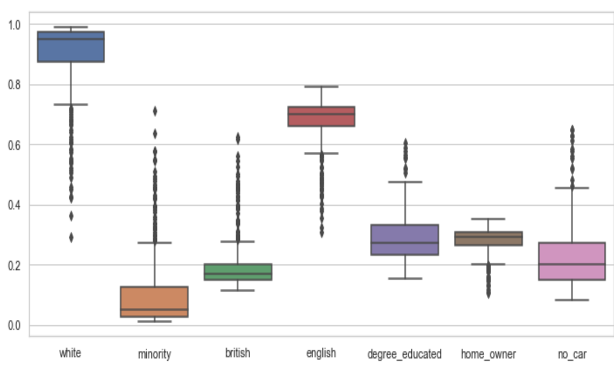


Figure 1. Boxplot of the most discriminating variables normalized

4 ANALYSIS

4.1 Analysis Approach

"Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets". [3] A great way to solve the problem and answer our research questions is the known visual analysis diagram proposed by Keim in figure 2. [5] The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. [4]. This loop presents a great workflow for our approach and refines the results for our process "Analyse first, show the important, zoom/filter, analyse further, details on demand". [5]

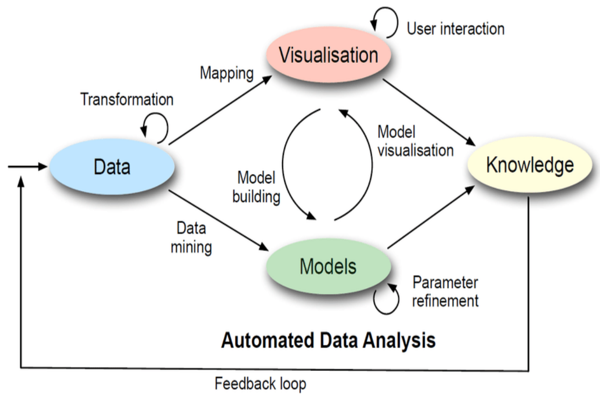


Figure 2. Visual analysis process by Keim

We will try to make sure that computational methods do what they do best and human do what they do best which is the reasoning in the analysis process. We understand that applying some algorithm will not be sufficient and that human reasoning needs to be supported by visual displays.

To achieve answers to the questions above, we will be using the following tasks:

1. Explore spatial distribution visually.

   We will explore spatial variation with GeoPandas and Altair for Leave vote and other explanatory variables by visualizing Choropleth maps of voting preference by local authority in England & then London, where we will do a comparison. We will inspect the visual output if we need to discretise the variables further. The goal is to stop when a level of understanding is achieved.

2. Explore relationships and patterns visually between explanatory variables and the target leave vote.

   Scatter plots can be great in showing relationships between our target variable and the explanatory variables. We will derive new explanatory variables and plot them. this will be refined to achieve the desired results. That will be our stopping condition.

3. Explore demographic variables with leave vote with regression analysis.

   Regression analysis can be used to check each demographic factor and explain the differences in leave vote by building models to explain this. The visualization techniques used here will be correlation matrix and scatter plots with regression model that fits across a facet grid. This technique is very effective in finding relationships across conditional subsets of a dataset. Computational methods include data grouping and fitting the models. Lastly will try to apply multiple linear regression of our target variable leave onto demographic variables in different regions.

4.2 Analysis Process

This section will apply the approaches and steps explained in the previous section and will iterate through the visual analysis diagram show above.

1. Exploring spatial distribution

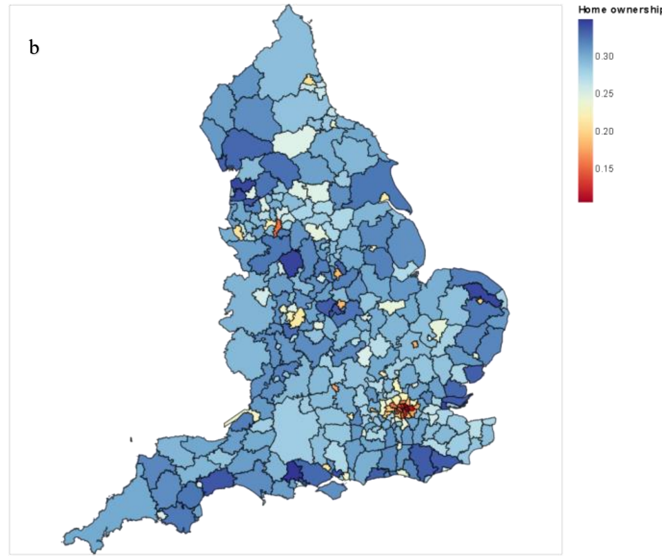
Here will be visually exploring the spatial distributions for some variables from table 1. The Choropleth maps will be coloured to the LA level with red, yellow, blue diverging colours which is a good option as the light yellow indicates the map average. It's worth noting that Altair discretise the colour efficiently but if needed we will discretise the variables further.

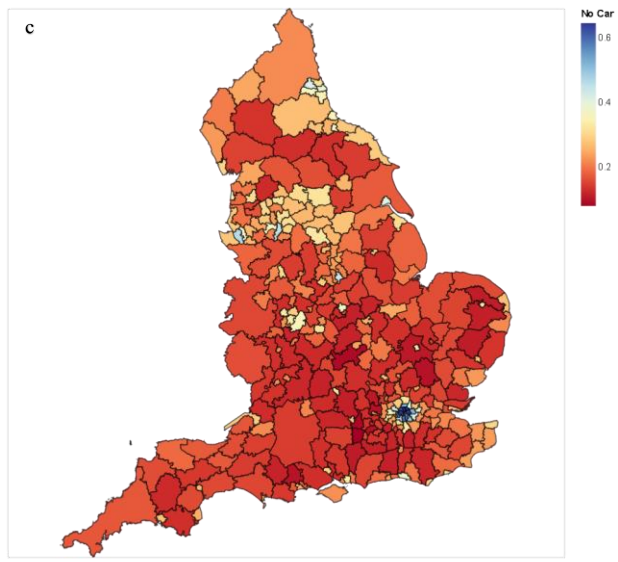
Examining the choropleth in figure 3.a, we observe that the highest concentration of population that voted leave in England is in the East Midlands and large northern cities including Sheffield Birmingham, and the south east of England. While the lowest is in London. This confirms several studies and statistics about the voting distribution in England.

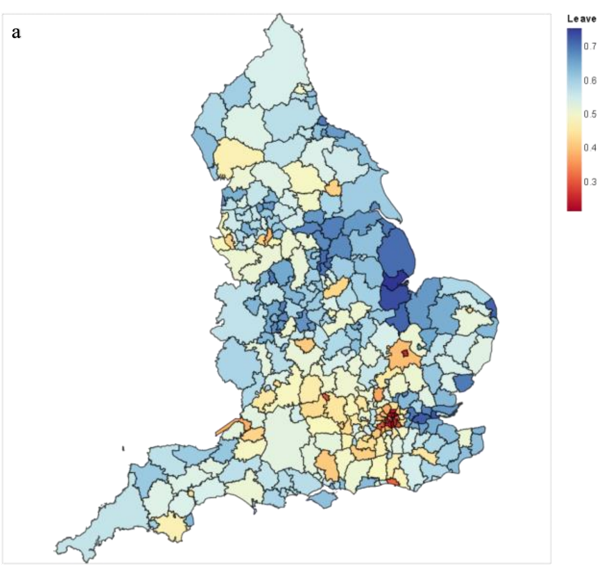
When we compare Leave vote from figure 3.a with Education from figure 3.d bottom, some interesting patterns emerge. The choropleth confirms the findings about the correlation of the education demographic variable with our target output Leave. This comparison suggests a higher correlation that previously assumed. The maps look opposite to each other.

So, is Brexit caused by lower education? It seems if the government spent more money on education and universities that Britain will still be a member of EU.

We can observe from figure 3.c that no car variable is negatively correlated with Leave vote in London and East midlands but the relationship fades in other areas. Home ownership in figure 3.b looks interesting and is the opposite of No car. This variable is strongly correalted with the Leave vote in London and around but also the correlation fades in other areas in England.







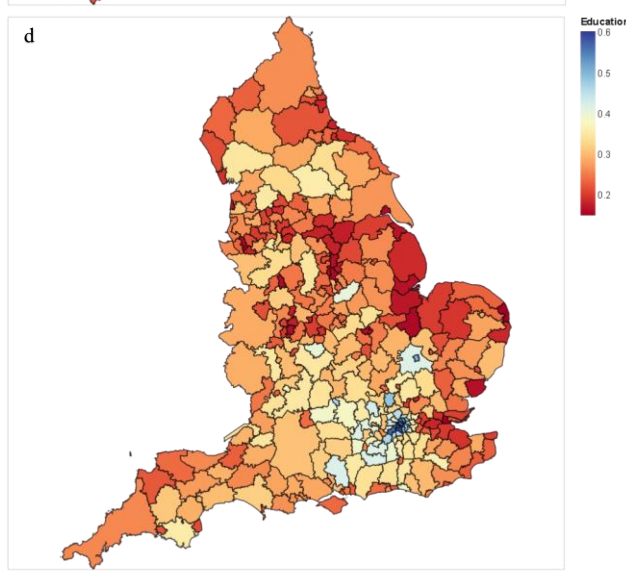


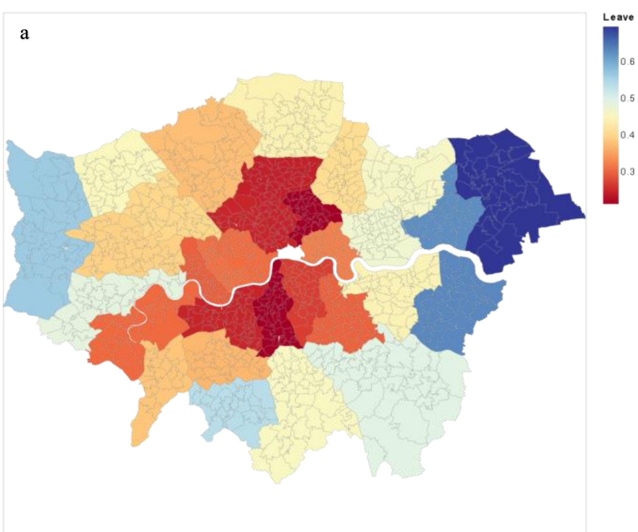
Figure 3. Choropleth maps of London measuring the variable being displayed: **3.a.** Leave vote. **3.b.** Home ownership. **3.c.** No car. **3.d.** Education

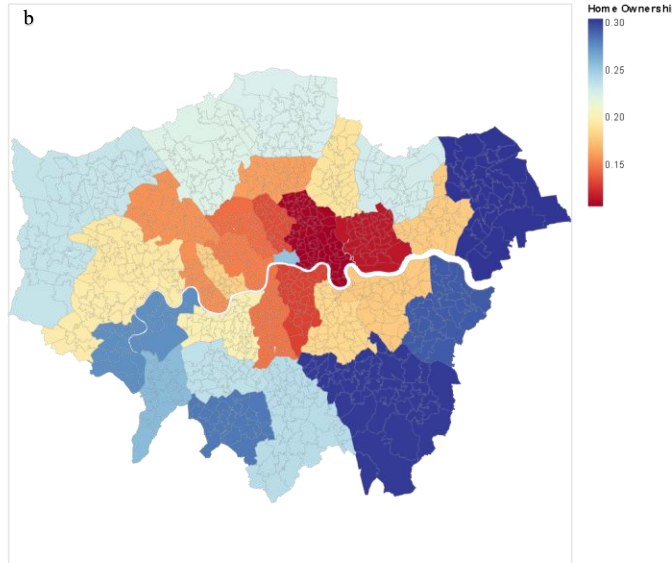
Let's take a closer look at London. London as a whole voted against leaving the EU, according to the electoral commission it was a 60/40 split in favour of remain as following:

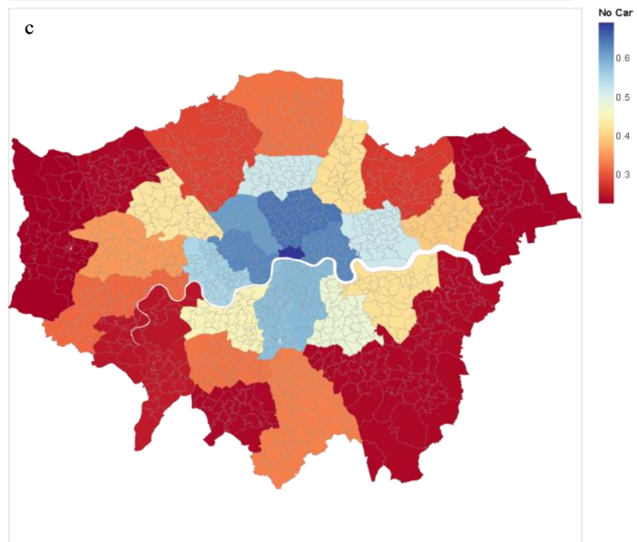
Remain	**Leave**
2,263,519	1,513,232

The vote strength varied widely across the capital. From figure 4.a, we can see the inner part of London voted remain and the outer east of London voted leave. Reports confirms this as Lambeth borough emerged as with the lowest leave voters in the entire country with a 21% leave vote only. On the other hand, Havering had the highest leave vote of 70% in London.

Figure 4.d suggests that the highest concentration of education variable is in the centre of London where they voted remain and the lowest is in the outer east of London which voted leave. No car in figure 4.c is negatively correlated in the centre and east of London. Similar to England, homeownership looks opposite from No car variable. The general pattern between the 3 variables explained here and the vote leave are clear now in regard of geographical differences, so we can stop now.







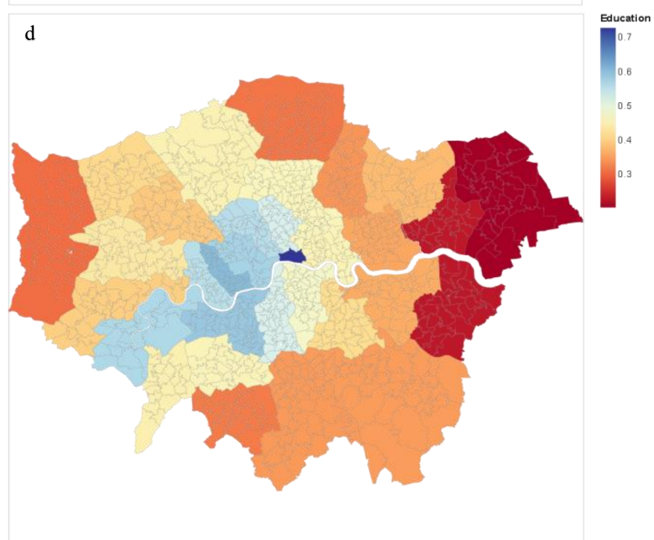


Figure 4. Choropleth maps of London measuring the variable being displayed: **4.a.** Leave vote. **4.b.** Home ownership. **4.c.** No car **4.d.** Education

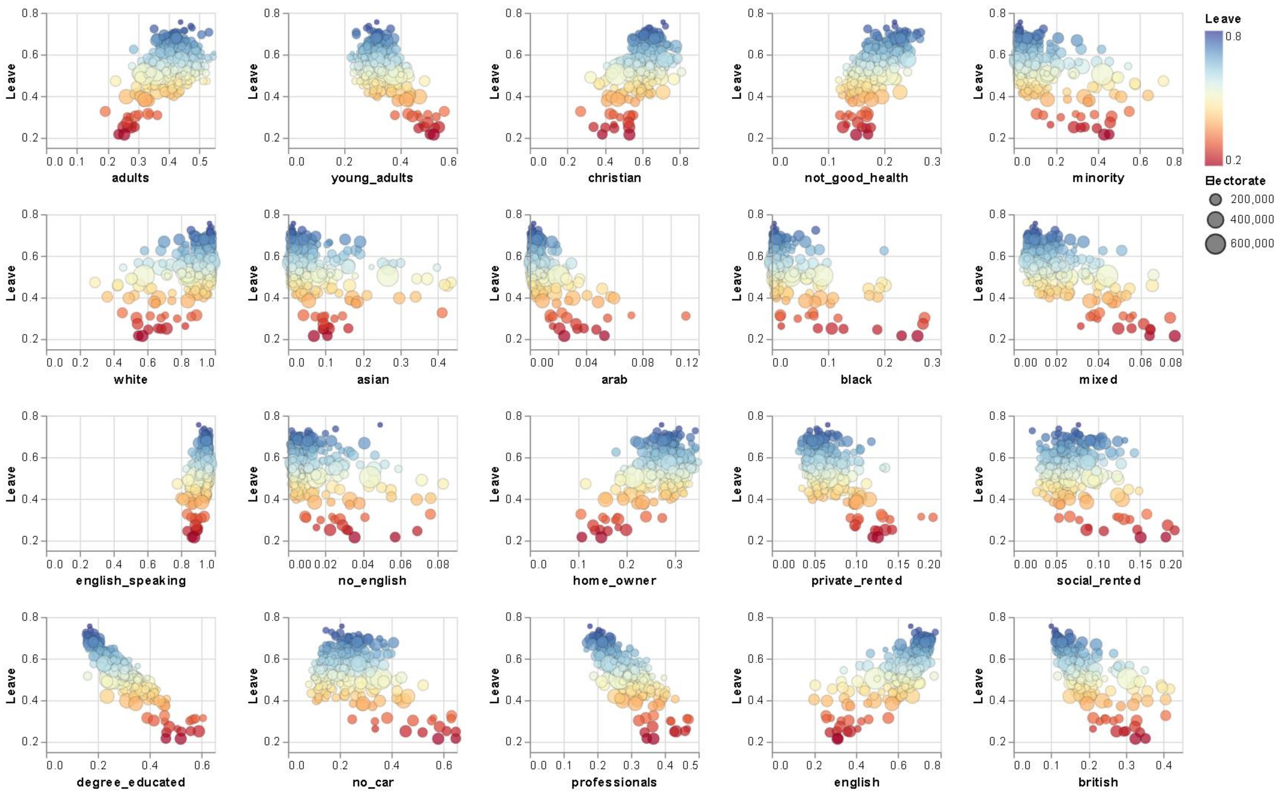


Figure 5. Scatterplots showing our most discriminating variables and Leave vote in England

2. Exploring relationships and patterns

Scatter plots can be great in exploring relationships and patterns between explanatory variables and the leave vote.

The following table summarizes the strong correlations figure 5 between our target variable leave and the explanatory variables:

Positive Correlation	**Negative Correlation**
adults, christian, white not_good_health, home_owner, english	Young_adults, no_religion, arab, black, mixed, no English, private_rented, social_rented, degree_educated, no_car, professional, british

The findings are very pleasing. There is a clear relationship between the leave vote in England and the age group. The older residents were in favour of leaving the EU, however the younger group were against leaving.

Other relationships are the race group. It is clear that whites tended to vote leave and all other ethnic minorities voted generally remain except for Asians which the relation is negative in general but not clear. I think we should investigate more about the relationships of between different races and leave vote. It would be interesting to see how this group voted in different regions of England.

The strongest correlation with the leave vote from figure 5 is degree educated and this confirms our findings previously from choropleth maps and with by Goodwin & Heath's [2] findings. Professional variable also is very strong negative correlation and goes side by side with the degree educated. That makes sense as its clear that the higher your education the better employment level. Another interesting finding is residents with poor health, where they tend to be in in favour of Brexit.

National identity came as a surprise in figure 5 shows a strong positive correlation between leave and English, but a strong negative with British.

It's important to note here that we are trying to capture general trends and understanding of our variables with the Leave output. The data we have does not represent individuals rather its aggregated on a LA level.

We can summarize our findings in this section that a general pattern emerged which looks like English, poverty, low health and low skills were key drivers in vote leave, while British, degree educated with professional jobs voted against Brexit.

3. Exploring with Regression Analysis

We explored different variables with our target leave vote in the last task and now we have found some variables that correlates with the target variable leave vote and may further deepen our understanding of our research questions. In order to understand how ethnic minorities voted in the referendum, we can try to build models to consider the effects of these variables on the vote.

Based on these assumptions, we added the national identity to the census, where they might play a role to understand more about the ethnic minorities and the way they voted. National identity tends to summarise something about the groups such as ideas, traditions and relationships. [6]

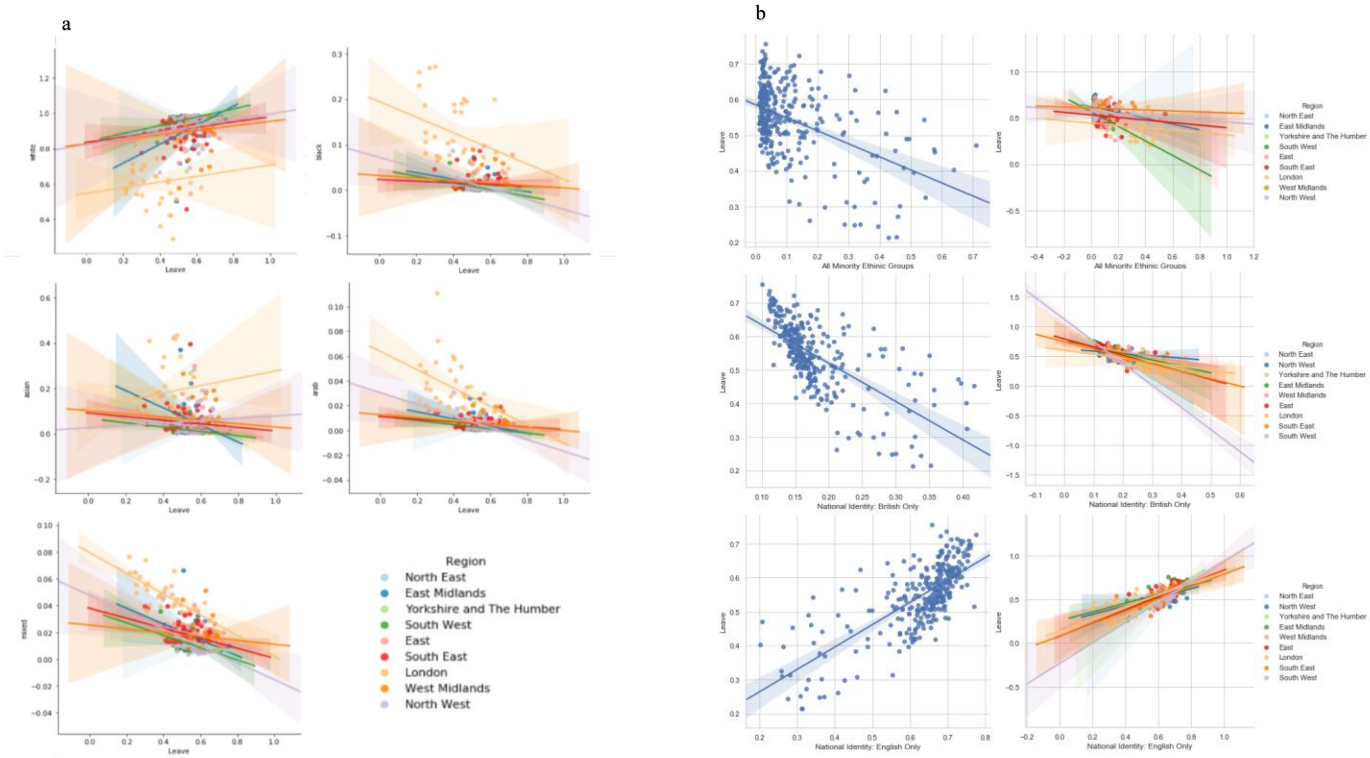


Figure 6. Regression Analysis. **6.a.** Single Ethnic groups with Leave vote in England regions. **6.b.** National identity (English & British) with leave vote and in England regions.

Figure 6.a, suggests that different ethnic groups against the leave vote. Since we don't have individual level data then we cannot be certain of our results. The graph gives insights into how that region in England voted and the corresponding size of a particular ethnic group in the region. According to the British Election Study, 2015, the strongest ethnicity group that were more likely to vote to stay in the EU were the black ethnicity, but our findings were that the mixed ethnicity has the strongest negative correlation in across all ethnicities in all regions followed by the black ethnicity. According to the policy exchange [6], national and ethnic identities are probably related where white groups are more likely to be identified as 'English only' and ethnic minorities are more likely to be identified as 'British only'. See Table 2.

	British only	British only AND British & other UK	English only identity	Any form of UK identity
White	14%	25%	64%	95%
Indian	58%	61%	12%	75%
Pakistani	63%	67%	15%	84%
Bangladeshi	71%	74%	8%	84%
Black African	43%	45%	10%	59%
Black Caribbean	55%	60%	26%	88%

Table 2. Source: A portrait of modern Britain, Policy Exchange, 2014

The top row In Figure 6.b represents the total ethnic minorities from the census data which shows a negative general correlation with Leave. The strongest slope was the south west and lightest was in the west midlands. The second row was aggregated from the national identity identified as British only which we assumed that it represents the ethnic minorities, the R-squared (R2) is 0.44 with the strongest negative correlation in south west region which matches the minority variable. Finally, the last row is English only. The regression line has a strong R-squared (R2) of 0.55 and shows a general trend of the white ethnicity towards leave. The English only suggest that Yorkshire region has the strongest white population that voted Leave. An interesting development from the 2019 election could throw light on the interaction between "Englishness" and the leave vote. Labour lost nine seats to the Conservatives across Yorkshire for the first time in a generation. It would be interesting to compare the vote distribution to our areas, because these are areas that are less ethnically diverse and tend to have a higher perception of English rather than British identity. This helps explain why the superficial correlation between ethnic diversity should not be read as causation.

At the end, a simple explanation on how ethnic minorities voted in the 2016 referendum is not an easy task and cannot be only explained by socio-economic background. Afterall, there is a list of other factors and issues that. happened around the referendum campaigns like fear of immigration and sovereignty.

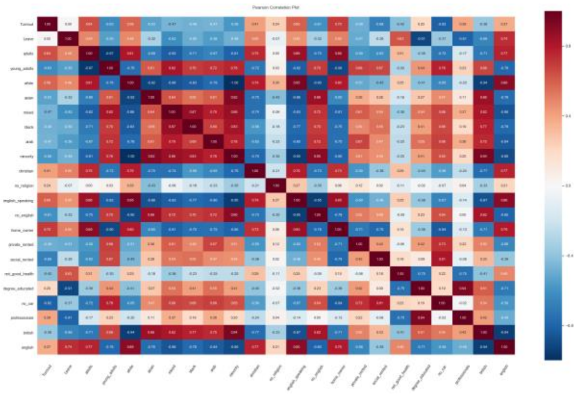


Figure 7. Pearson correlation plot

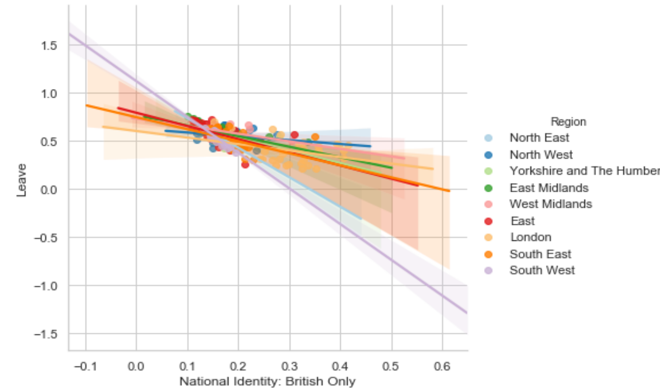


Figure 8. National identity: British only

Collinearity help us to reduce the number of variables that can explain the leave vote, it can initially be inspected through studying pairwise correlation between each explanatory variable using a correlation matrix, see figure 7. The goal is to try to cluster our variables into two groups that will summarize the ethnic groups communities, based on the policy exchange study that white groups are likely to be identified as English only and ethnic minorities are likely to be identified as British only and setting the threshold to r=0.6 we found the following:

Group 1: English Only
Heavy positive correlated with Adults (0.77) white (0.89), Christian (0.77), English speaking (0.9), home owner (0.79), and correlated with our target leave (0.74).

Group 2: British Only
Heavy positive correlated with Young Adults (0.68), Asian (0.86), black (0.77), mixed (0.82), Arab (0.75), no English (0.82), degree educated (0.61). And negatively correlated (-0.66) with leave variable.

4.3 Analysis Results

It is worth discussing our research questions and evaluate if our analysis answered the subject of this paper.

A general pattern emerged from our analysis which looks like English, poverty, low health and low skills were key drivers in vote leave, while British, degree educated and professionals voted against Brexit. Education has a stronger relationship with leave vote that previously assumed and is the most important driver of the leave vote. It is probable that if population held more university level degrees, then they will still be in the EU.

The characteristics were very similar in England and London.

National identity added a considerable complexity and played an important role to understand more about the ethnic minorities, see figure 8 which captured a general behaviour trend that explained how they voted in different regions of England. It seems plausible to conclude how did ethnic minority groups voted in different regions of England.

Under certain assumptions, this can be construed as the characteristics of different ethnic groups communities as shown in the analysis above with group 1: English only having a positive correlated support of Leave up to r=0.74 and group 2: British only with a negative correlation of r=-0.66.

.

5 Critical Reflection

In this section we will reflect on our chosen approach and analysis process. We will also discuss the limitation and generalisability of our findings and potential improvements.

The visual analytics diagram proposed by Keim was very effective where we found our approach in a loop which created a great workflow for refining our results. The analysis was simple yet effective.

It's important to mention that we tried to capture general trends and understanding of our variables with the Leave output. This type of analysis shows relationships at an LA level, so we don't know how certain groups or ethnic groups voted within LA level. The analysis does not represent individuals rather its aggregated on a LA level.

Initially we started our analysis with the referendum results on a ward level sourced from the BBC, but because of the lack of data provided where only half of the areas provided, we decided to not go further and instead we used the referendum results on a Local Authority-level (LA), the voting can be related to the demographic characteristics of those authorities. The most discriminating variable was the degree educated level in England. This result ties well with previous papers where several methods were used. [2]

The choice of choropleths maps for geographic relationships visuals worked well, allowing clear visual interpretation of results, we compared maps and reasoned about it and found visual understanding that helped to answer our questions. The process of the computational methods based on an interpreted visual output helped us with the model outputs and improved our understanding.

We introduced new variables based on the national identity that was not in the census 2011 data. This variable helped explain the variation in voting preference among ethnic groups. This was an important step as our initial analysis based on the census 2011 ethnic different groups didn't explain the group relationship with Leave but rather the LA these groups are living in. These minorities had small numbers compared with the total population at that level. The British only variable sourced from the national identity was probably the best way to capture these minorities behaviour, where they are more likely to be identified as 'British only' and white groups are more likely to be identified as 'English only'. This assumption was based on the UK's leading think tank policy exchange which is an important study that explains the portrait of modern Britain in 2016.

Our Analysis was simple, but domain knowledge is very important to gain better results. It could be beneficial to use GWR (Geographical weighted regression) techniques to map the variables and target across England regions. Also clustering techniques could be very helpful in grouping demographics linked to the Leave vote.

The methods and techniques here are very useful for the political domain where local demographics are key in elections. Targeted campaigns can be used to win key targets in different local authorities. Also, this analysis is highly generalisable and could be applied in different sectors like health, education and crime data analysis.

REFERENCES

[1] Roger Beecham, Aidan Slingsby, and Chris Brunsdon. Locally-varying explanations behind the United Kingdom’s vote to leave the European Union. Journal of Spatial Information, 2017.

[2] Goodwin, M. and Heath, O. The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result. The Political Quarterly 87, 2016.

[3] D.A. Keim, J. Kohlhammer, G. Ellis, F. Mannsmann (Eds.). Mastering the Information Age. Solving Problems with Visual Analytics, Eurographics Association, Goslar, 2010.

[4] Kohlhammer, Jörn & Keim, Daniel & Pohl, Margit & Santucci, Giuseppe & Andrienko, Gennady. Solving Problems with Visual Analytics. Procedia Computer Science. 7. 117–120, 2011.

[5] D.A. Keim, F. Mansmann, J. Schneidewind, H. Ziegler. Challenges in visual data analysis, in: Information Visualization (IV 2006), Invited Paper, July 5-7, London, United Kingdom, IEEE Press, 2006.

[6] A portrait of modern Britain, Policy Exchange, 2014.

[7] Andrienko GL, Andrienko NV. Interactive maps for visual data exploration. International Journal of Geographical Information Science, 1999.