Section 5 Project

United Kingdom
2019 General
Election

FLATIRON
SCHOOL

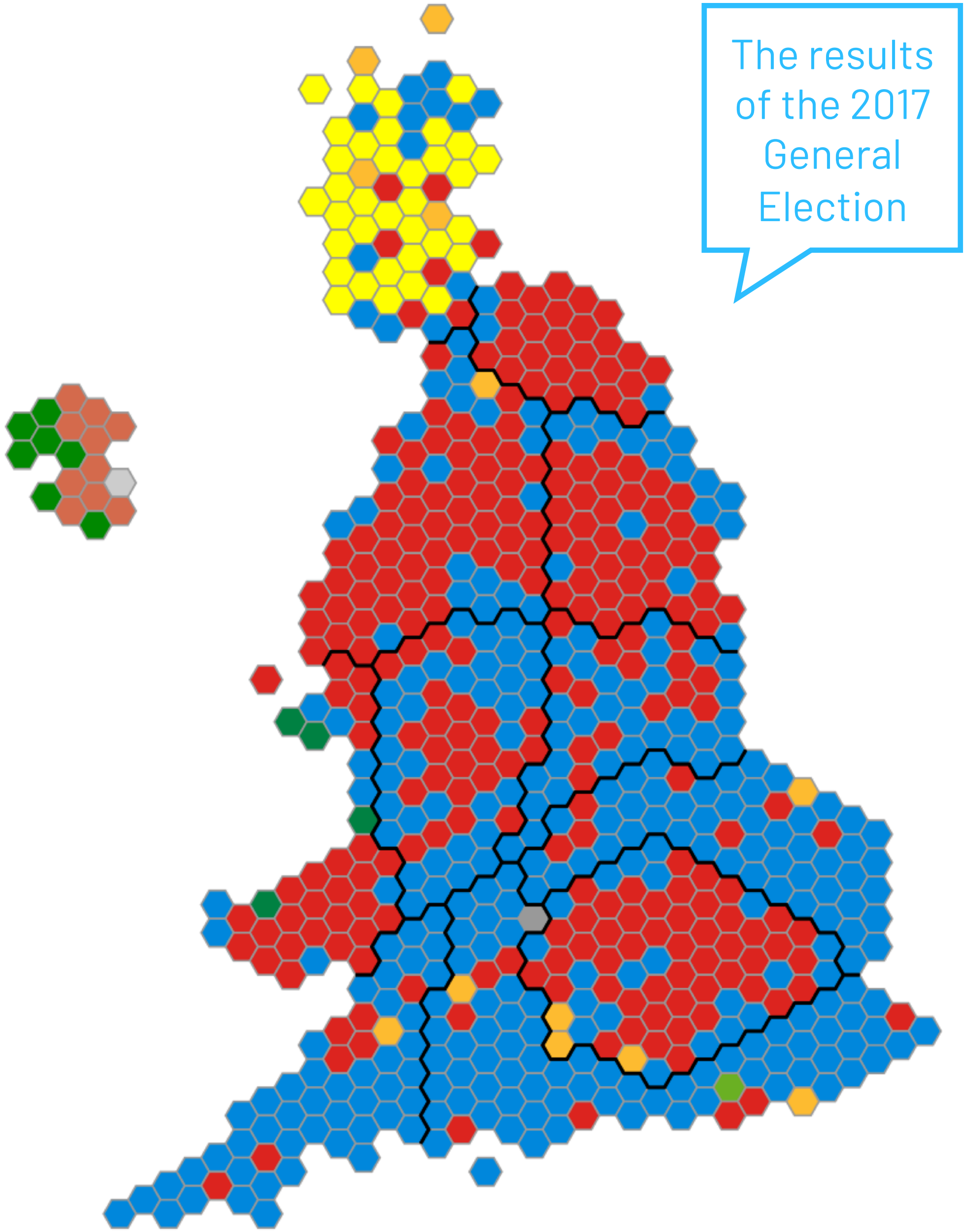
Executive summary

- A brief introduction to UK Politics
- The case for data science
- Clustering analysis
- Predictive modelling
- Possible next steps

Appendix contains additional analysis

UK Politics – General Elections

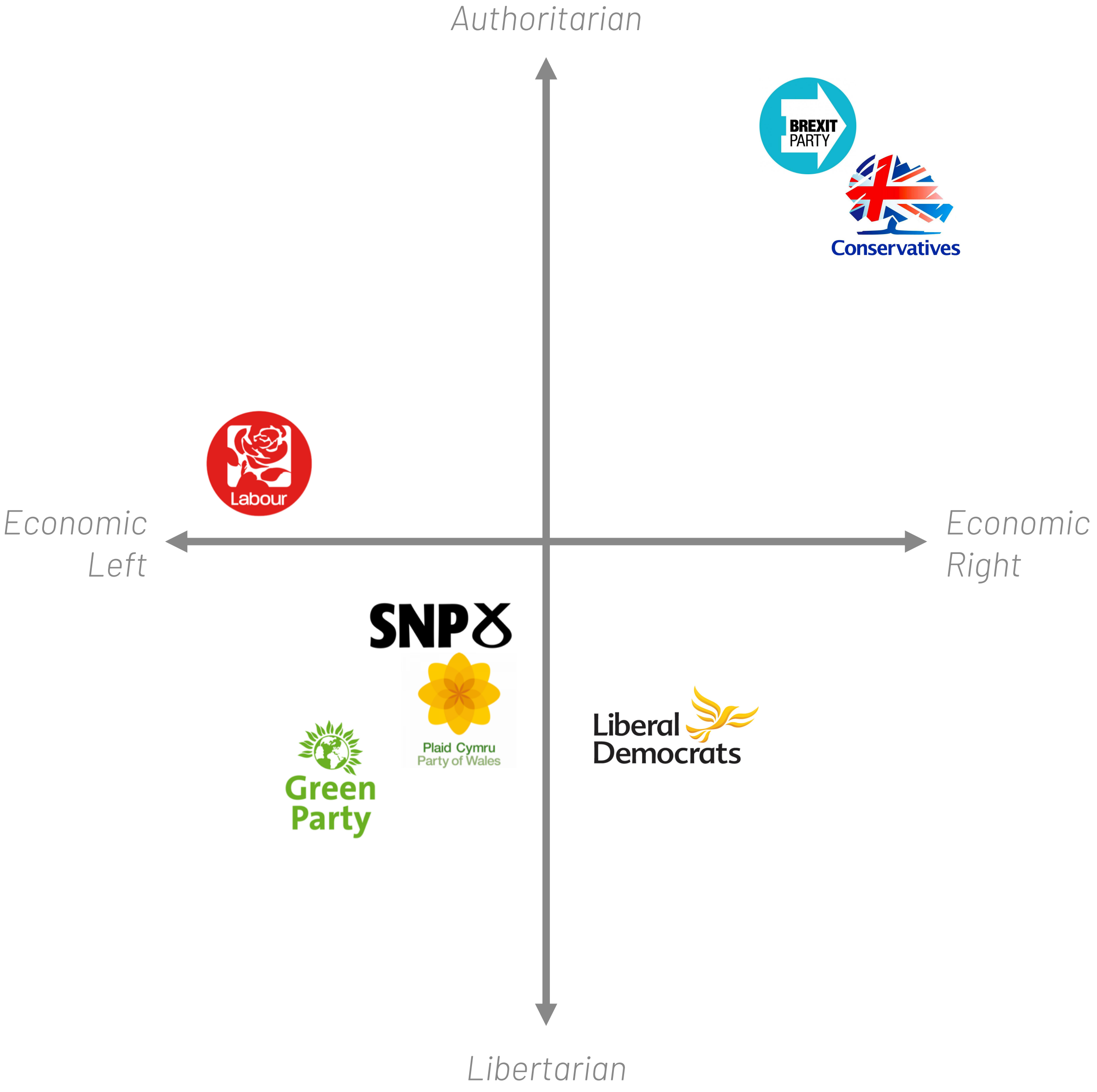
- Political power in the UK lies in the House of Commons
- 650 MPs attend parliament – one from each of the UK's 650 'constituencies'
- In a General Election, each constituency votes for a single MP from a selection of candidates unique to that constituency
- The one candidate that wins the most votes is elected
- The political party that has the most elected MPs wins the election, and typically forms the government



The results of the 2017 General Election

UK Politics –Parties

- The two main parties are the Conservatives and Labour. They have won every election since 1922*
- Scotland and Wales have nationalist parties (SNP and Plaid Cymru respectively) that only stand in those nations
- Northern Ireland has a different set of parties entirely. For this reason, NI is mostly excluded from our analysis



* This is a function of the UK's 'First Past The Post' voting system, which often leads to significant disparities between overall popular vote share, and the share of seats won in the House of Commons. In 2015 The Conservative Party managed to win more than half the seats with just 37% of the popular vote. On the other end of the spectrum, UKIP won just one seat (0.015% of the total seats) with 13% of the national vote.

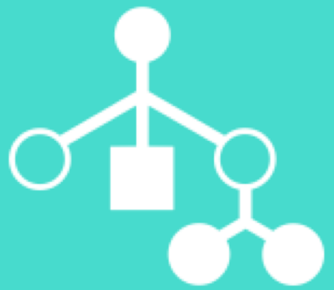
The case for data science

- Though there is emphasis on the national campaign, UK General Elections are, in essence, 650 mini-elections
- Voters in different parts of the country can have very different concerns. Local campaigns may need tailoring at a constituency level
- Resources are tight – there are strict campaign spending limits (and a finite number of party activists to deploy)

Which data science techniques could be useful here?

Clustering

By clustering constituencies based on their characteristics, we can see if there are any seats that parties are underperforming in, allowing them to better target resources



Classification

By creating a classification algorithm that predicts how a constituency votes (based on a range of local KPIs) we can better understand what factors drive voters of the different parties

The data used in this analysis

Data is taken from official Government sources. It includes the following KPIs for each constituency:

- Type of constituency (big city, small town, etc.)
- Population / population density
- Age demographics
- Incorporated businesses (3 KPIs)
- Unemployment rates (absolute and growth)
- Median weekly wages
- Share of jobs paying below local living wage*
- House prices (absolute and growth)
- House ownership status (4 categories)
- Share of families with children owning own house*
- Share of people working in given sector (12 categories)
- Share of people in managerial or professional roles*
- Share of people by qualification level (5 categories)
- Share of people by ethnicity (4 categories)
- Share of people by religion (7 categories)
- Share of people by country of birth (11 categories)
- Broadband quality in constituency
- Change in Local Authority spending power*
- Share of early-years childcare providers rated at least 'Good' by government inspectorate*
- Social mobility KPIs for children in poverty (8 KPIs)*
- Per-pupil school funding (absolute and growth)*
- School quality metrics (4 KPIs)*
- Share of people with chronic diseases (5 categories)*
- Average number of chronic diseases per capita*
- Share of people voting to leave the European Union

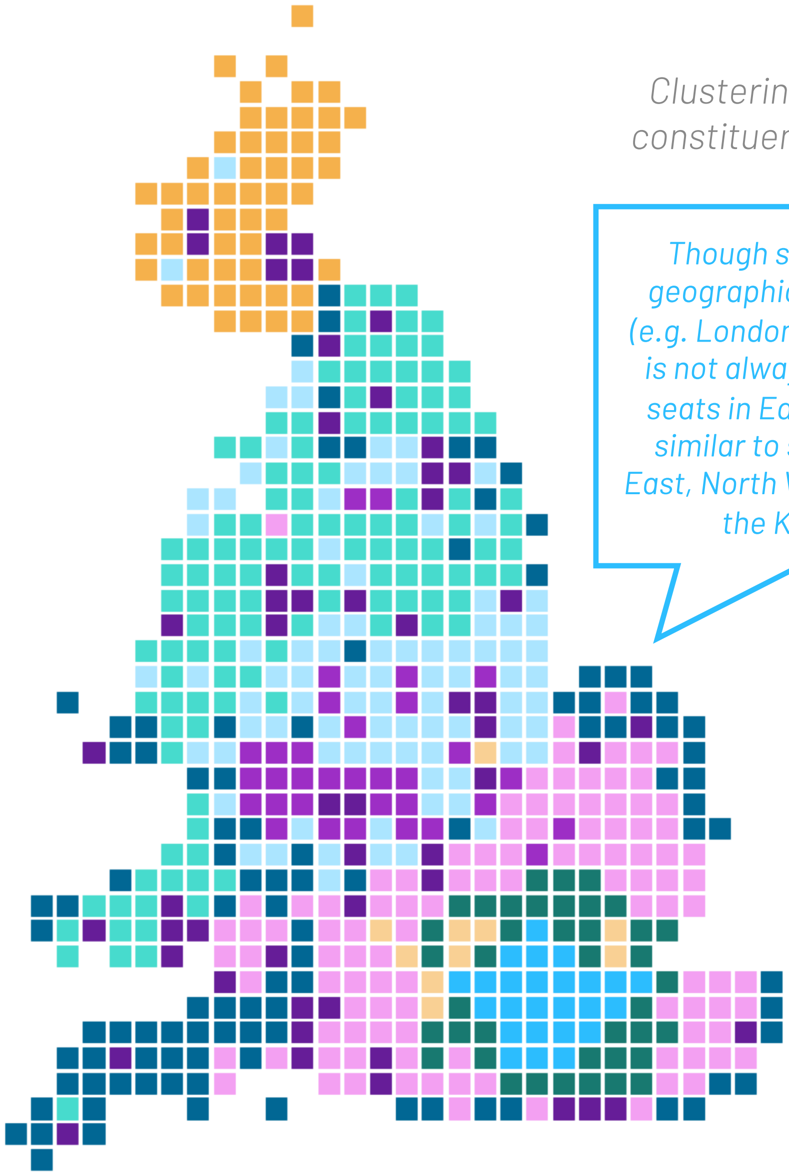
*Data for English constituencies only

Much of this data can be found online, at https://commonslibrary.parliament.uk/local-data-browse-by-topic/

Clustering constituencies

- Conventional thinking might assume seats are similar based on shallow assumptions, e.g. how they voted historically
- We can use a clustering algorithm to see which constituencies are demographically similar (and should hence vote similarly)
- Any seats that vote anomalously compared to others in their clusters could therefore be targeted in election campaigns

*Clustering analysis of UK constituencies, using HAC**



Though some clusters are geographically concentrated (e.g. London and Scotland), this is not always the case – many seats in East Anglia look very similar to seats in the South East, North Wales, Cumbria, and the Kentish Coast

* Hierarchical Agglomerated Clustering. See appendix for further details on this technique

A use case for clustering – Sefton Central

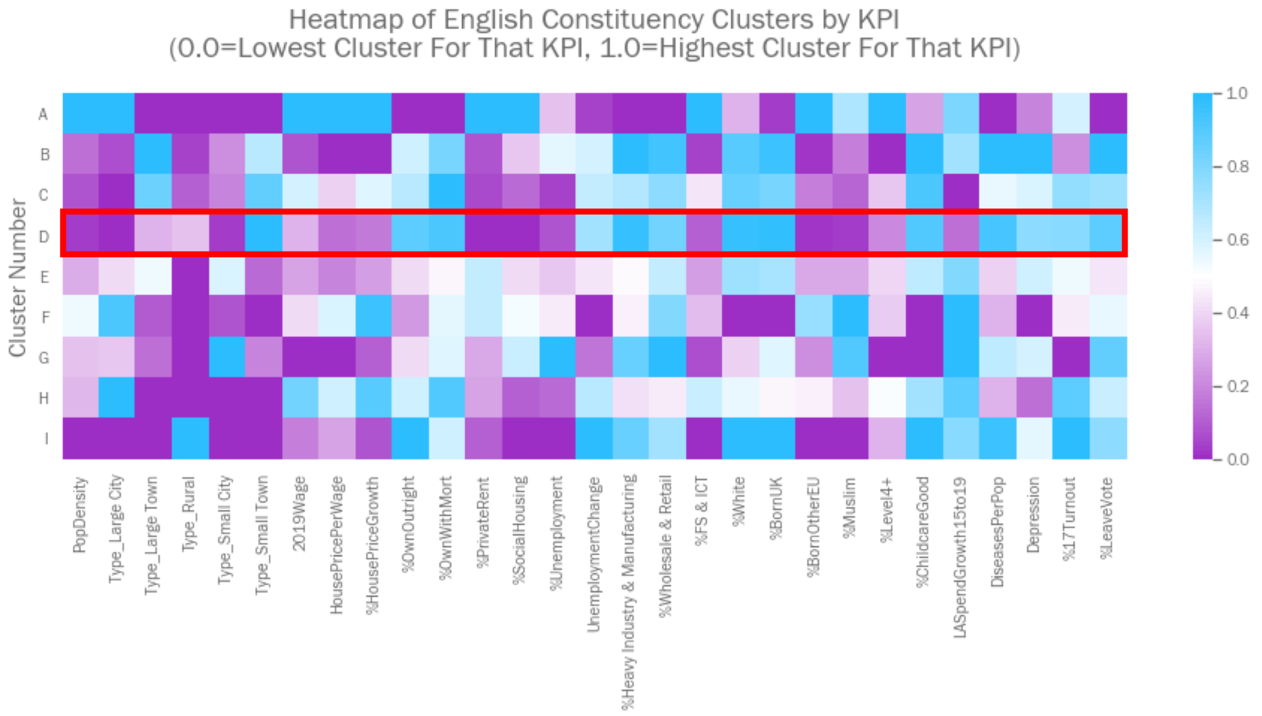
- English constituencies in Cluster D are overwhelmingly Conservative safe seats. They would therefore want to target the one Labour safe seat in that cluster (Sefton Central)

- They could look at the average demographics of Cluster 3 to help further tailor messaging

Seats In Cluster By Type	A	B	C	D	E	F	G	H	I
Conservative Safe Seats		47	67	96	3		2	11	30
Conservative / Labour Marginals	3	55	11	23	11	4	14	9	1
Labour Safe Seats	20	35		1	26	20	15	1	
Conservative / Lib-Dem Marginals	2	2	5	5	2			3	2
Lib-Dem Safe Seats					2			2	
Labour / Lib-Dem Marginals					1				
Green Safe Seats					1				

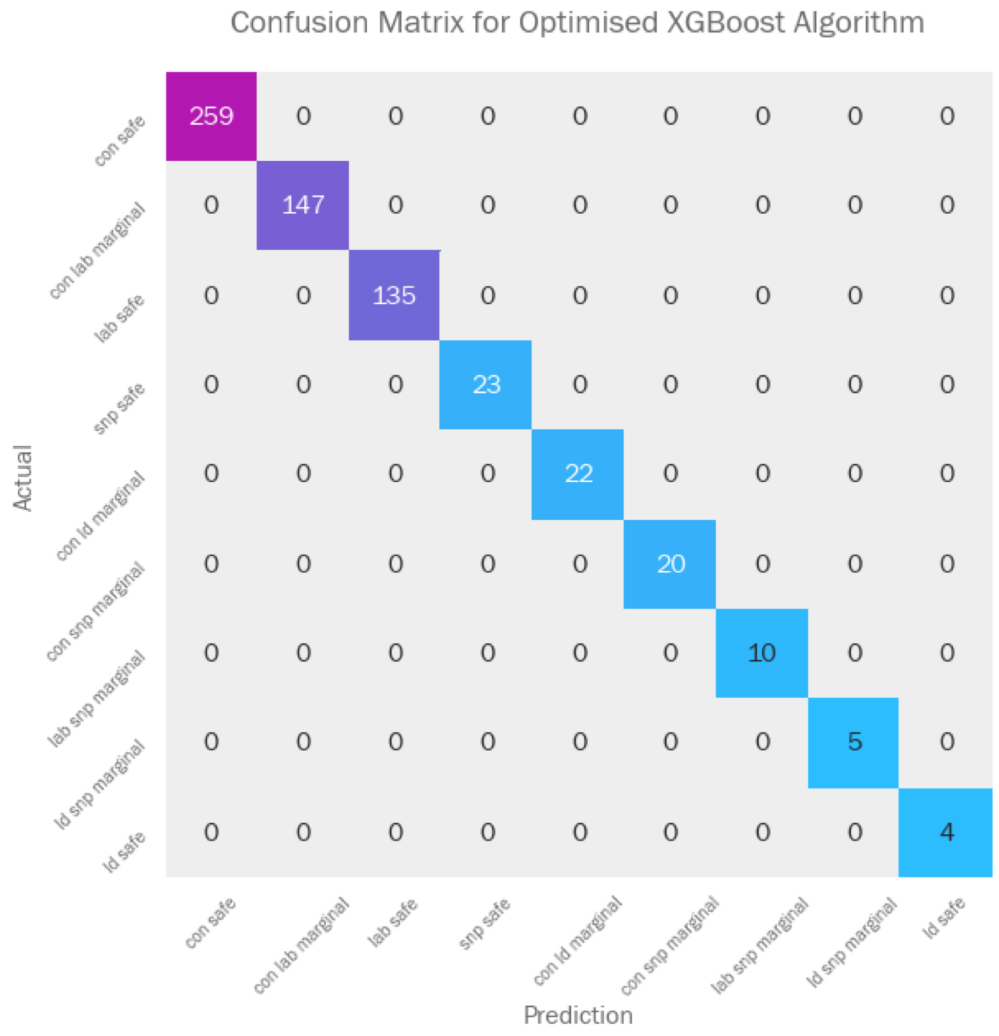
Cluster D Characteristics

- ✓ *Constituents have high rates of home ownership, with low levels of social housing*
- ✓ *House price growth has been low*
- ✓ *A high share of people work in Heavy Industry, and wages are below average*
- ✓ *Most people are white and UK-born*
- ✓ *Early-years childcare is decent, but education levels are generally low*
- ✓ *Local authority funding has been cut heavily*
- ✓ *There is a high rate of chronic illness*
- ✓ *Many people voted to leave the EU*



Categorising constituencies

- CB91 has created a model that perfectly predicts a constituency's voting preference
- The model makes its predictions using the constituency demographic data (as seen in the clustering exercise)
- We can use this model to further understand what drives the voting in each constituency

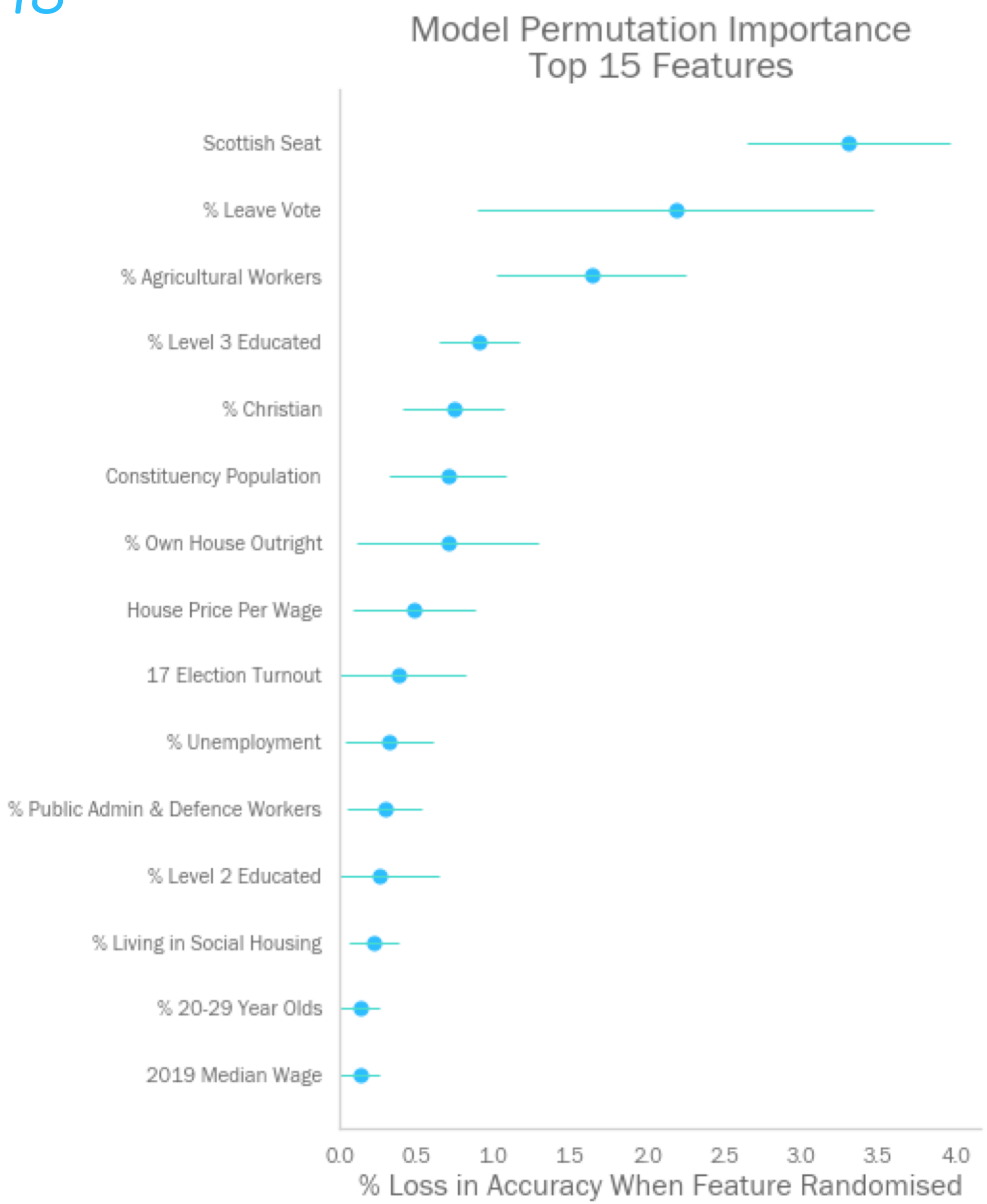


- Voting preference inferred from YouGov's national MRP poll as at November 27th 2019 (n=100,319).
- Details can be found here: https://yougov.co.uk/uk-general-election-2019
- A safe seat is given as one where the preference margin for the most popular party is over 15%

Explaining the model's decisions

CB91 used 'Permutation Importance' to see which factors had the biggest impact on the model:

- For each feature in turn, shuffle the values for each constituency
- Re-run the data through the model and see how much the model's accuracy decreases
- This procedure breaks the relationship between each feature and the seat type, thus the drop in accuracy is indicative of how much the model depends on that feature



Analysing important features

- The seats that Labour was competitive in cover a very wide range of values
- Thus, a single policy position would have failed to appease all potential Labour voters simultaneously
- The Labour / Conservative marginal seats look much more similar to safe Conservative seats than to safe Labour seats
- Thus, the Conservatives were able to have more focussed policy, without the risk of alienating either its core base, or its potential voters in Labour marginals

Three of the most important features, as identified by the Permutation Importance exercise on the previous page

Possible next steps

There are further steps we could take to improve the analysis. We could:

- Go into more detail across a full range of marginal constituencies
- Investigate feature importance with other libraries, e.g. LIME
- Re-run analysis against final results from 2019 general election
- Improve the accuracy of the feature dataset following the 2021 census

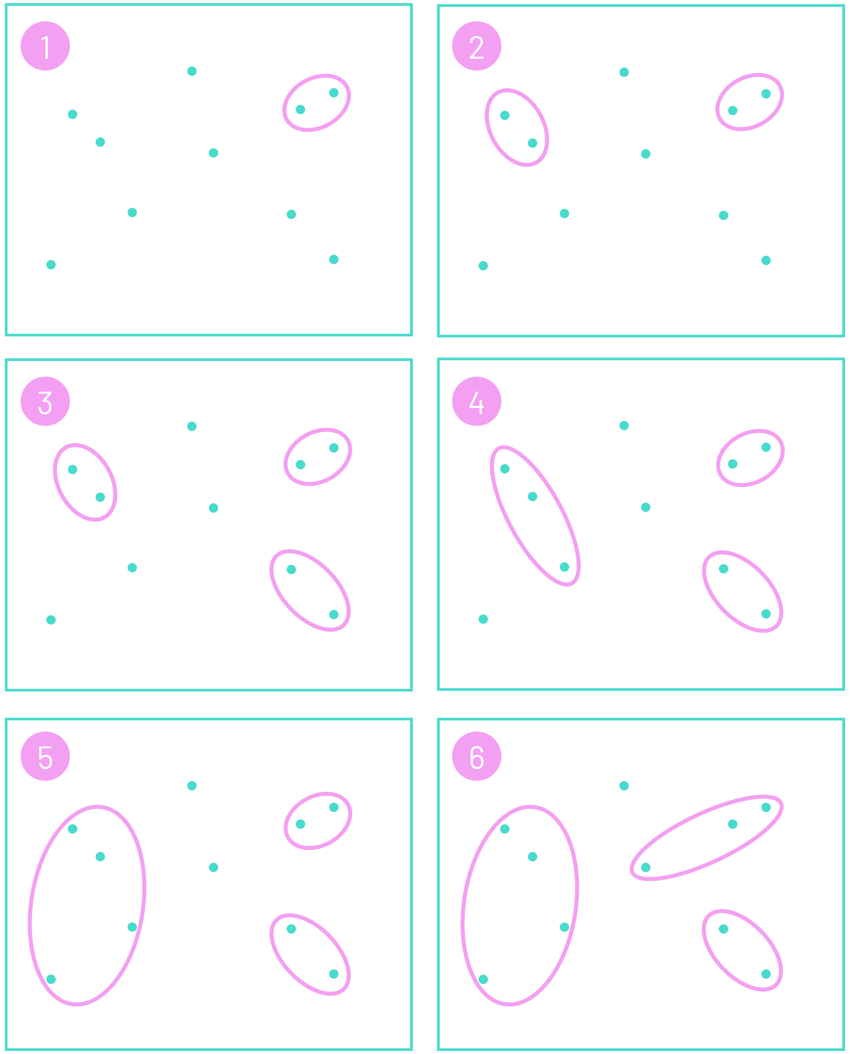
Thank you for your time

Further detail on the model, and additional analysis of the dataset is included in further pages

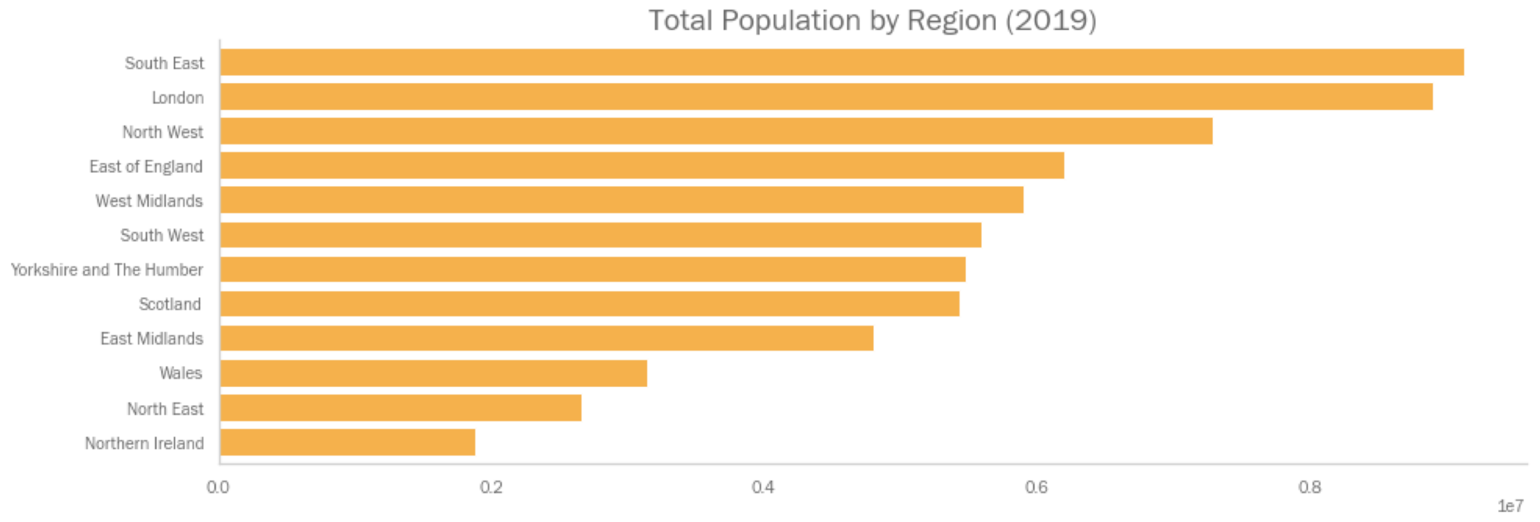
Hierarchical agglomerated clustering – in more detail

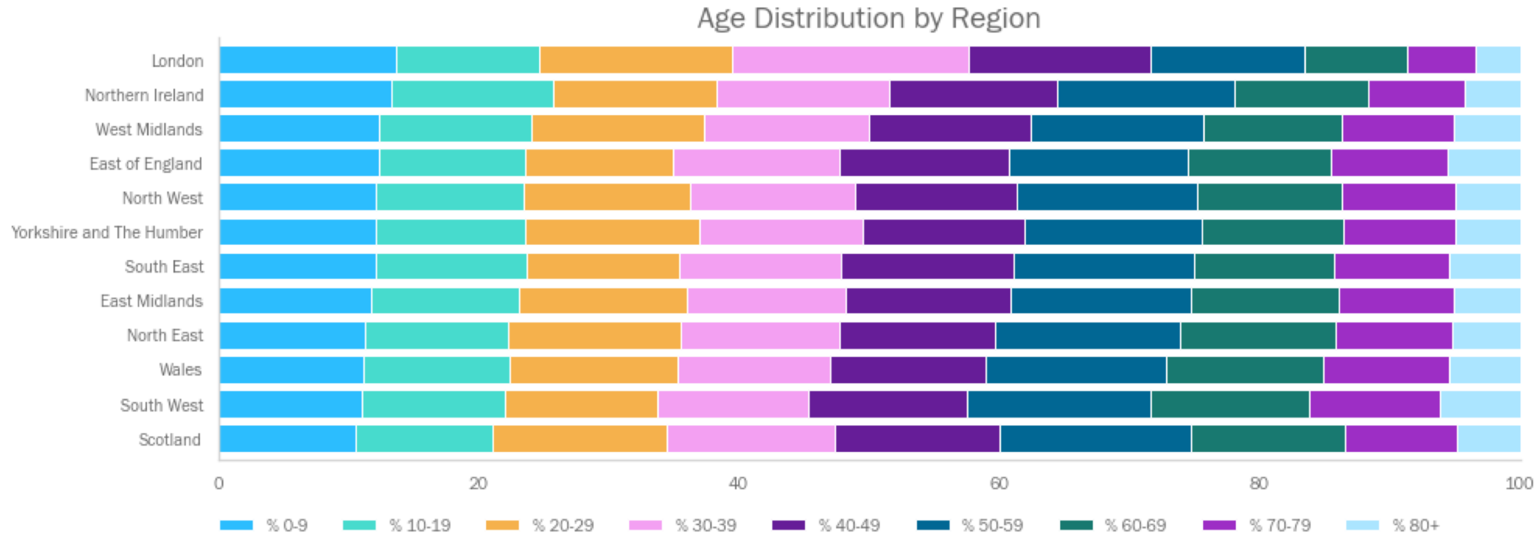
- HAC creates clusters by gradually joining datapoints that are 'close' to each other. Closeness is defined by Euclidean distance, so all features are scaled (so that the natural magnitude of a feature doesn't influence the clustering).
- HAC starts by treating each datapoint a cluster. It then joins the two closest clusters (even if this cluster is simply a single datapoint).
- It continues to join the next closest 'clusters' iteratively (measuring distances between the clusters' centres) until you have only 'n' clusters remaining. This level of 'n' is set before the algorithm is run.

First 6 stages of an HAC algorithm

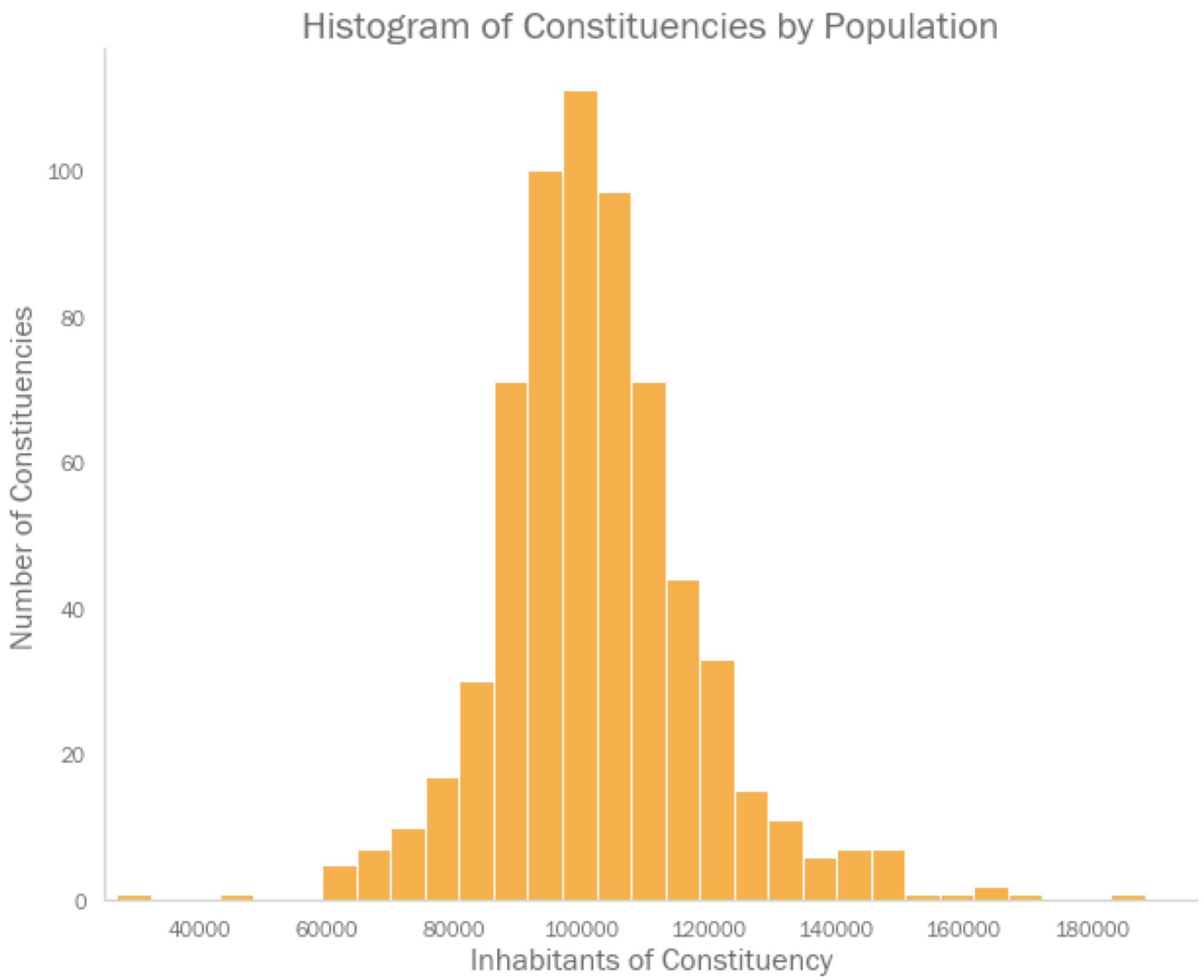


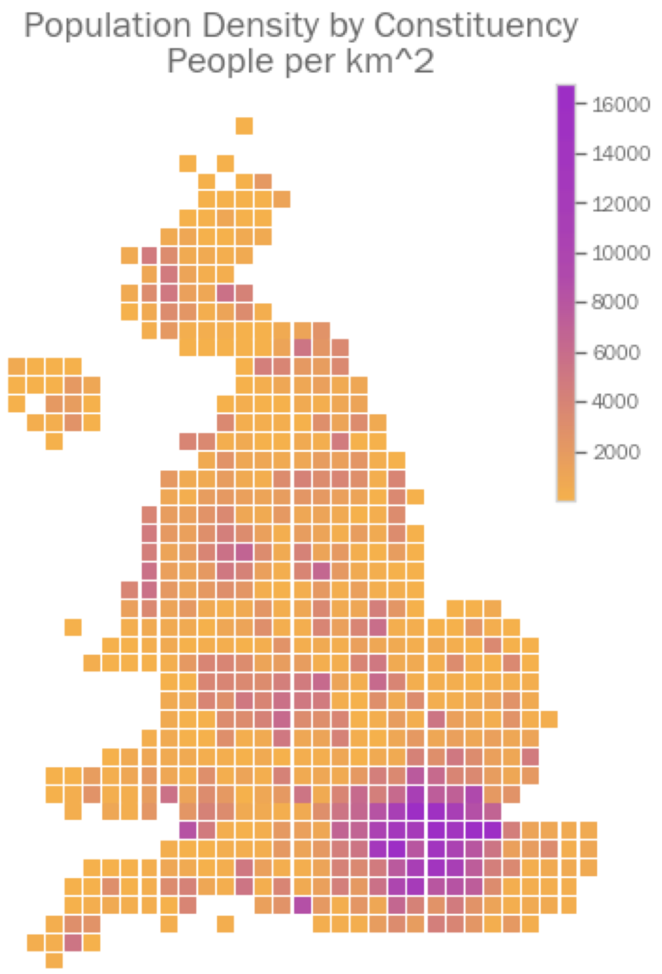
Population and age distribution by region



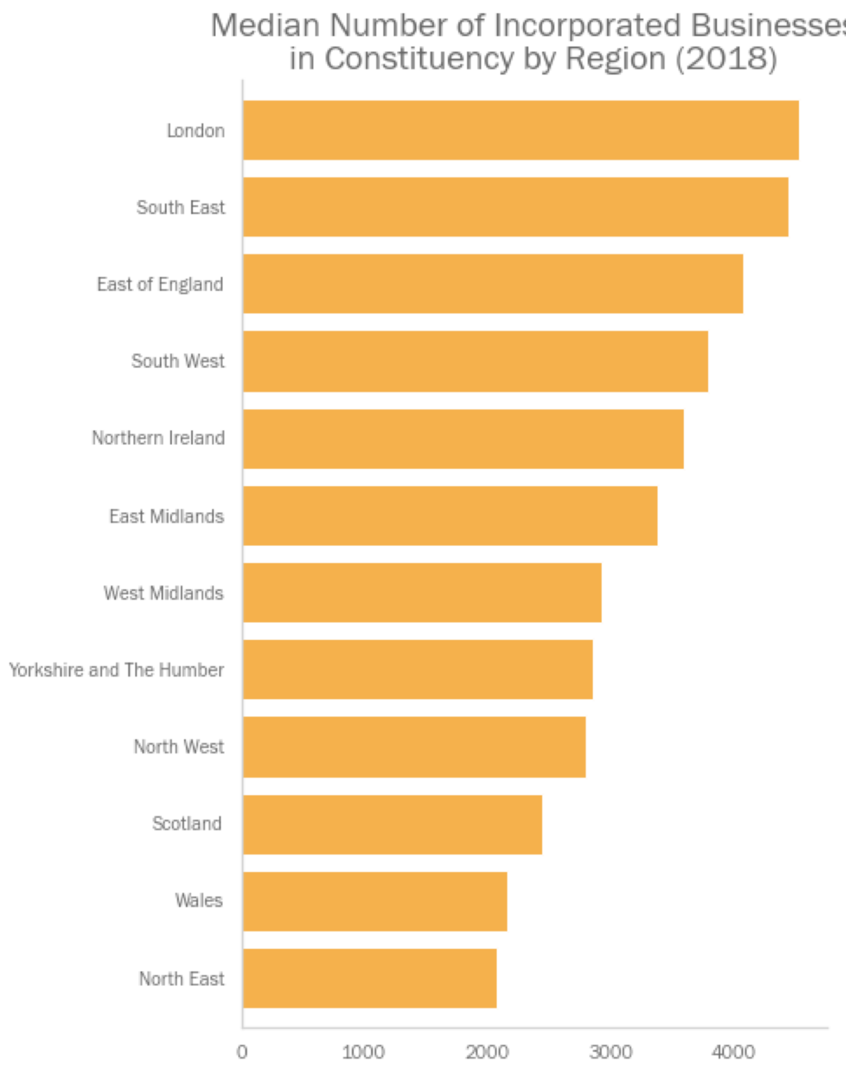


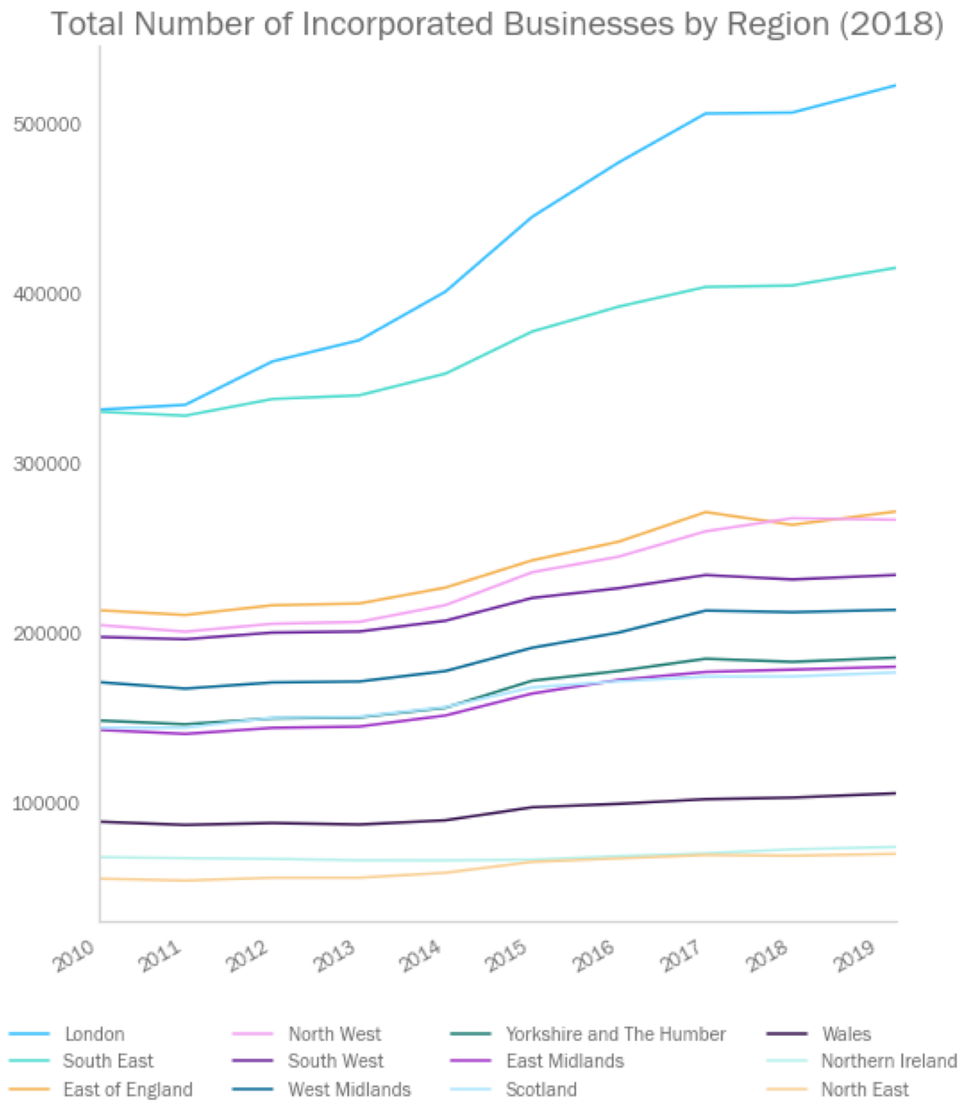
Population and population density by constituency



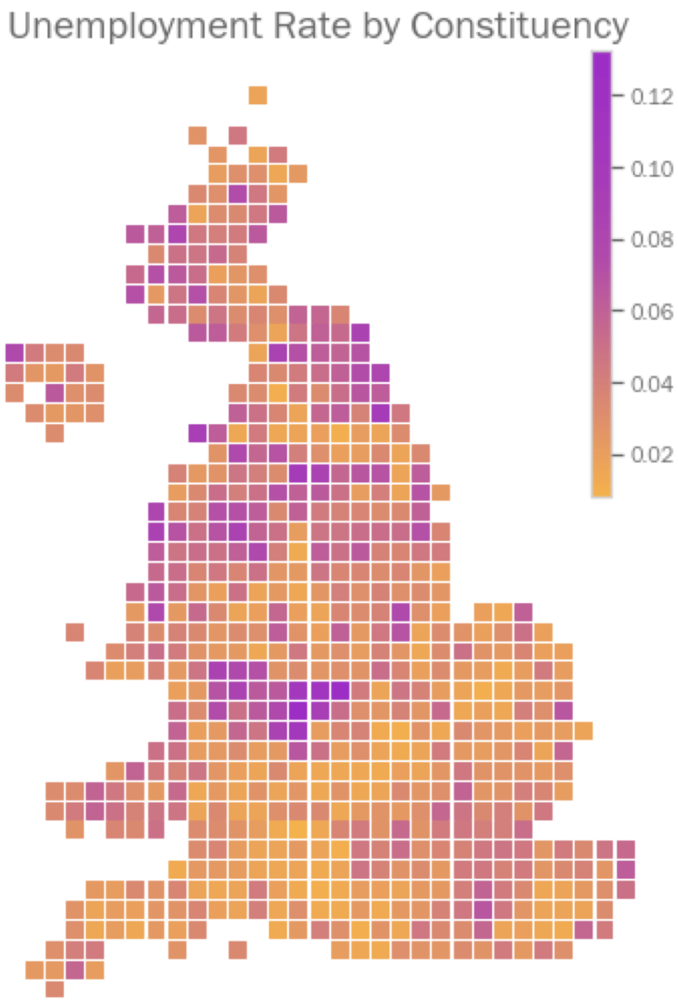


Incorporated businesses by region

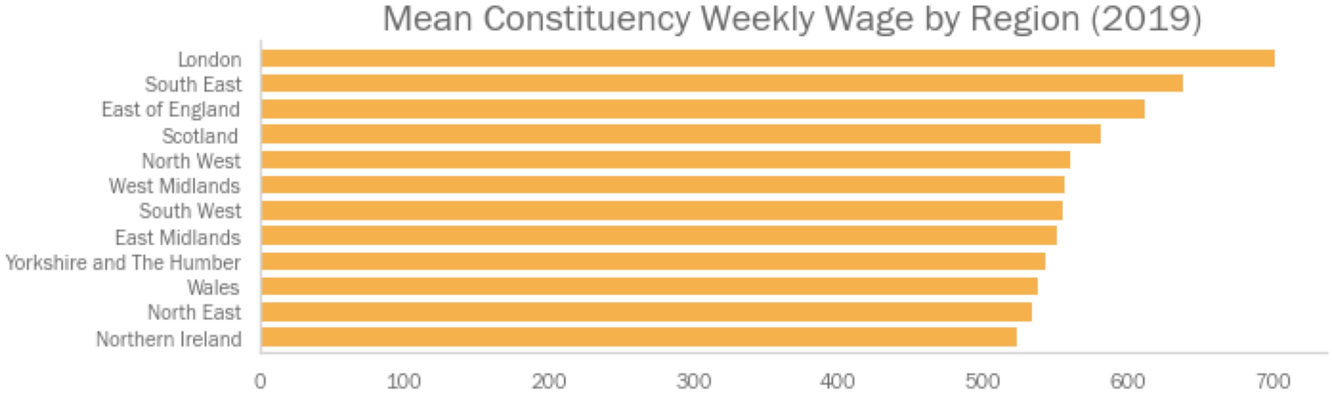


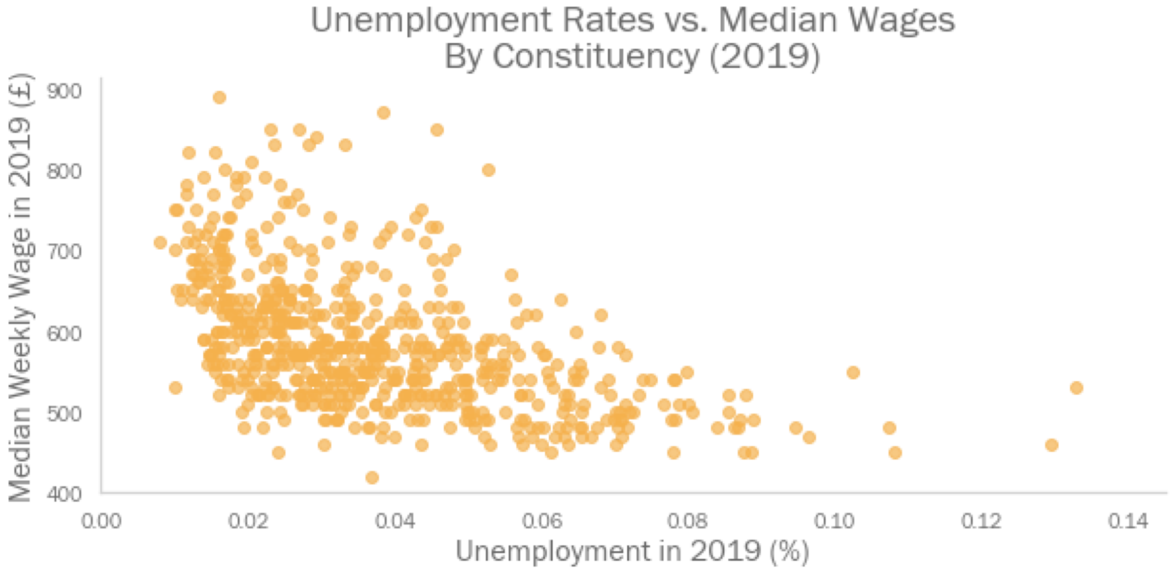


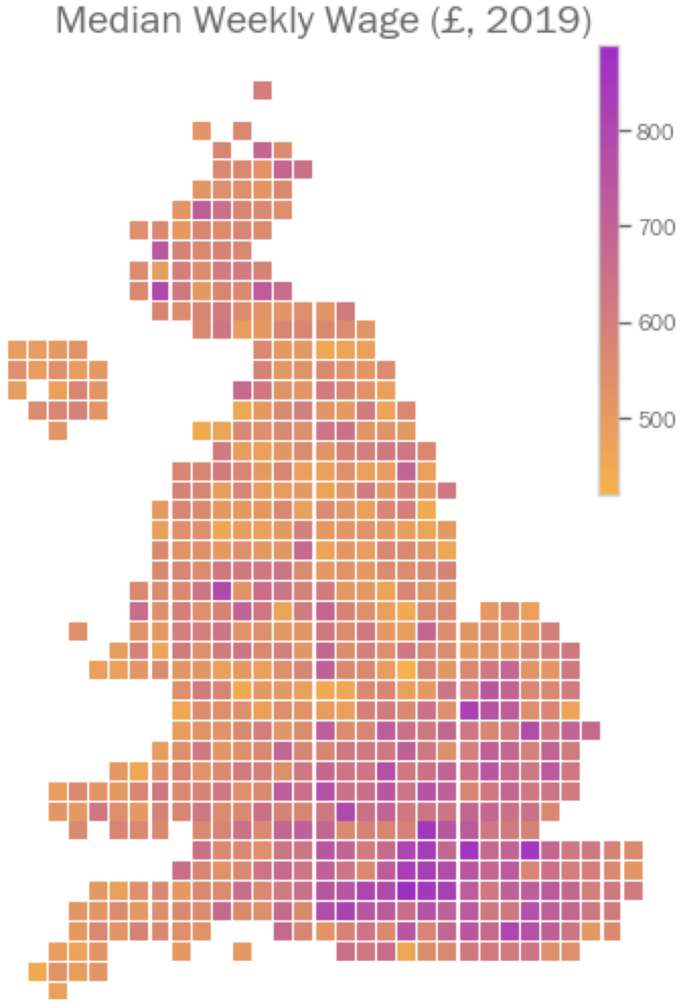
Unemployment



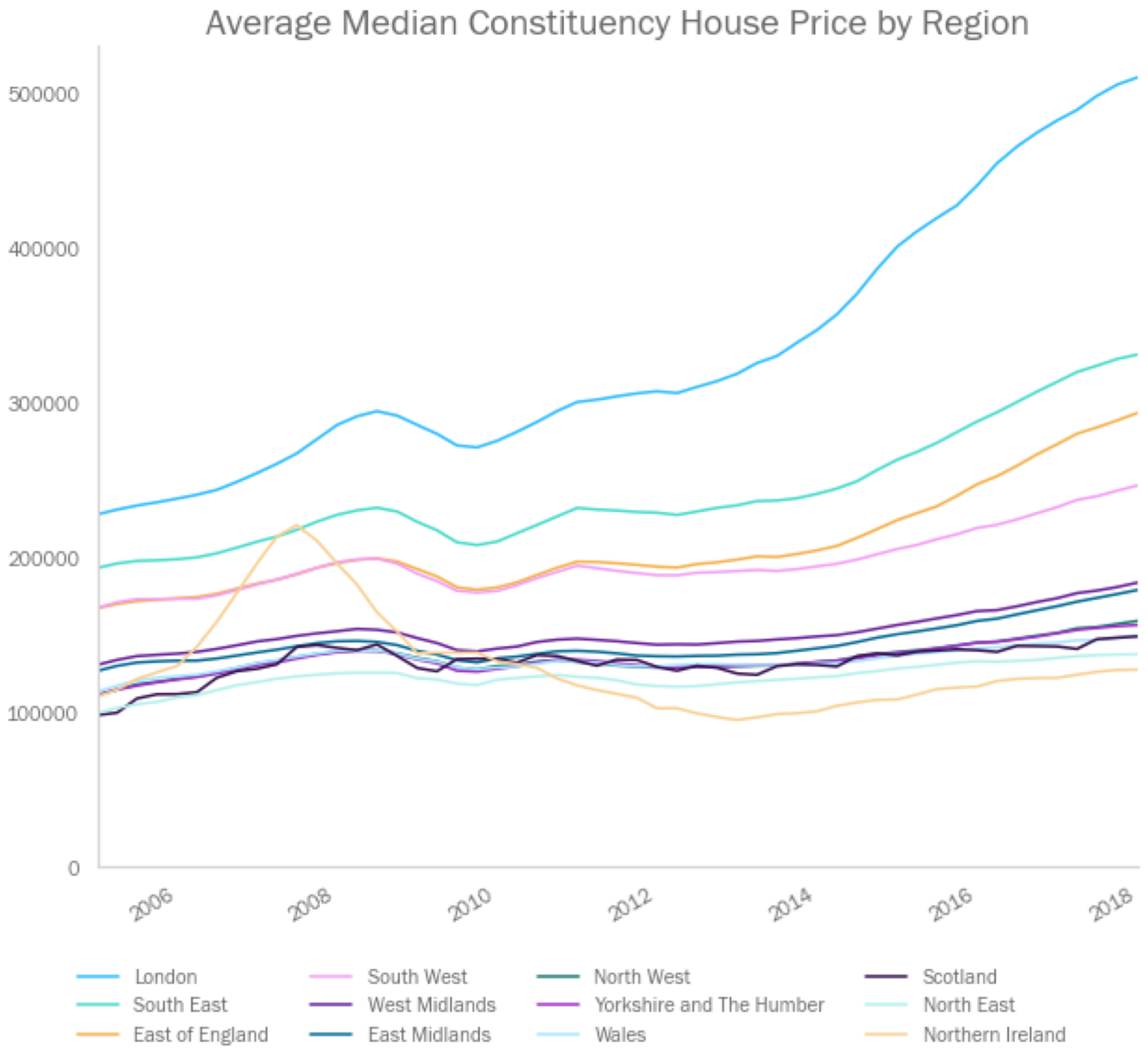
Wages

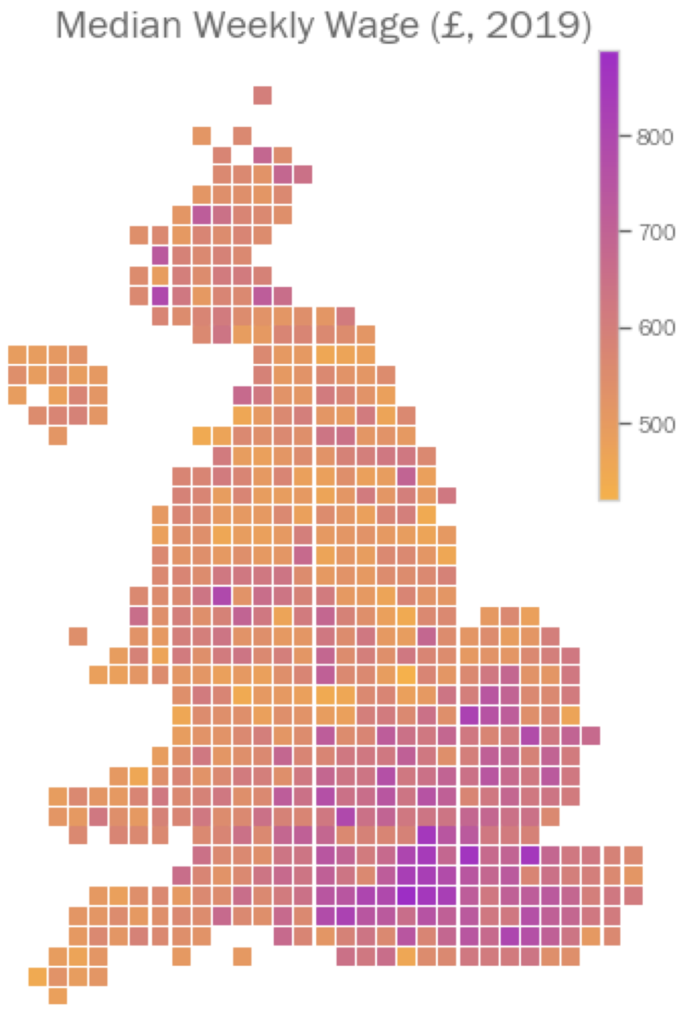

Mean Constituency Weekly Wage by Region (2019)
London
South East
East of England
Scotland
North West
West Midlands
South West
East Midlands
Yorkshire and The Humber
Wales
North East
Northern Ireland
0
100
200
300
400
500
600
700


Unemployment Rates vs. Median Wages
By Constituency (2019)
Median Weekly Wage in 2019 (£)
900
800
700
600
500
400
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Unemployment in 2019 (%)

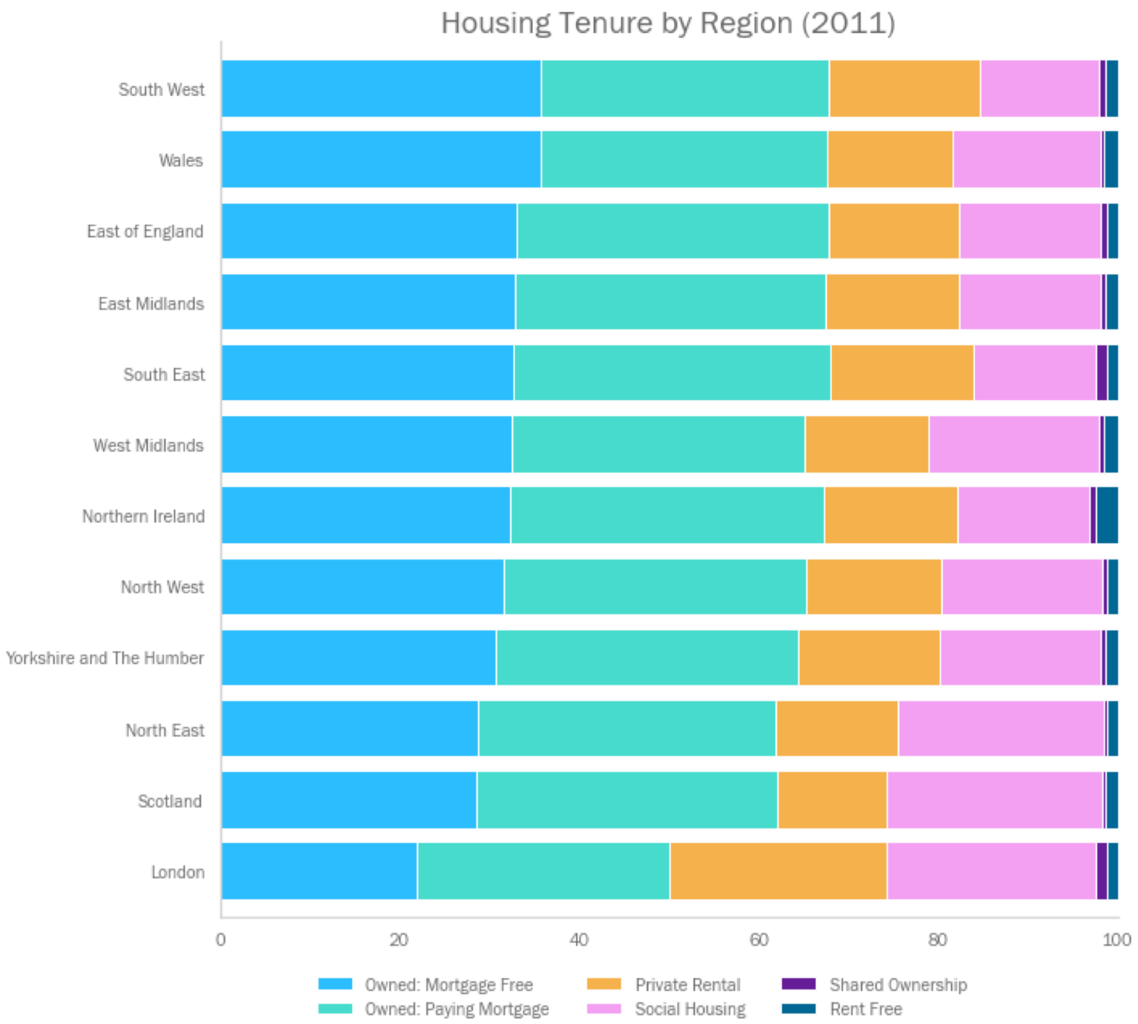

Median Weekly Wage (£, 2019)
800
700
600
500

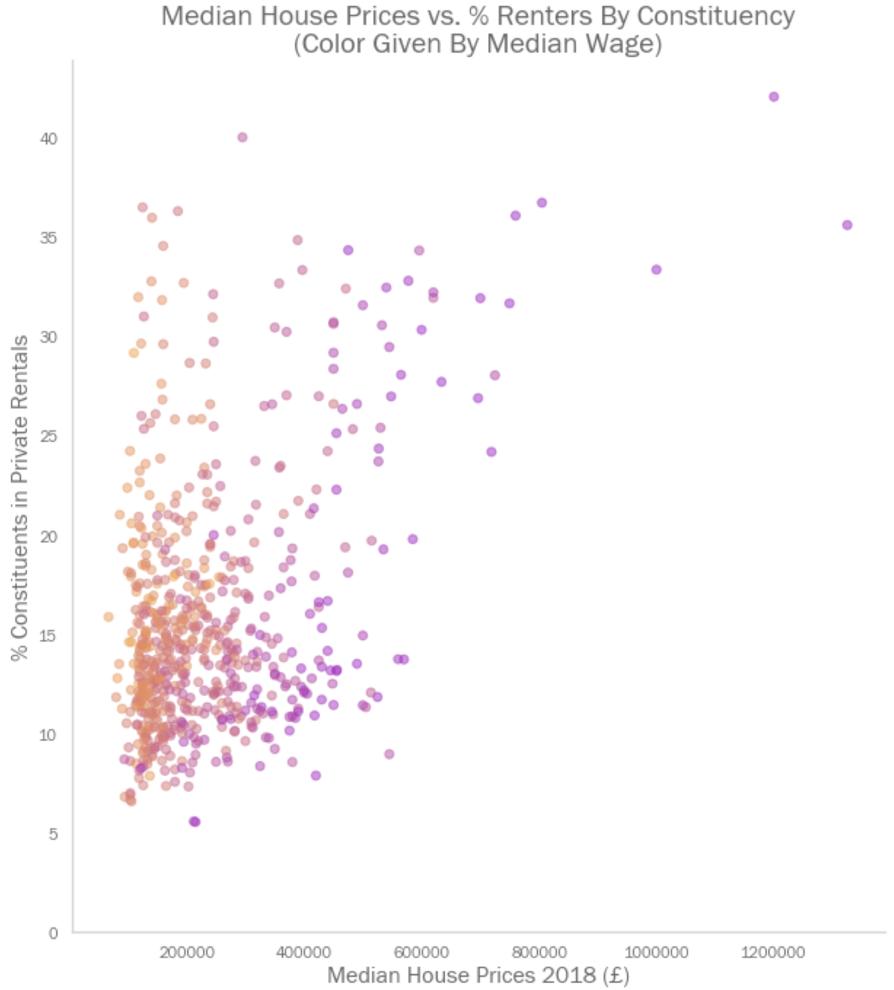
House prices





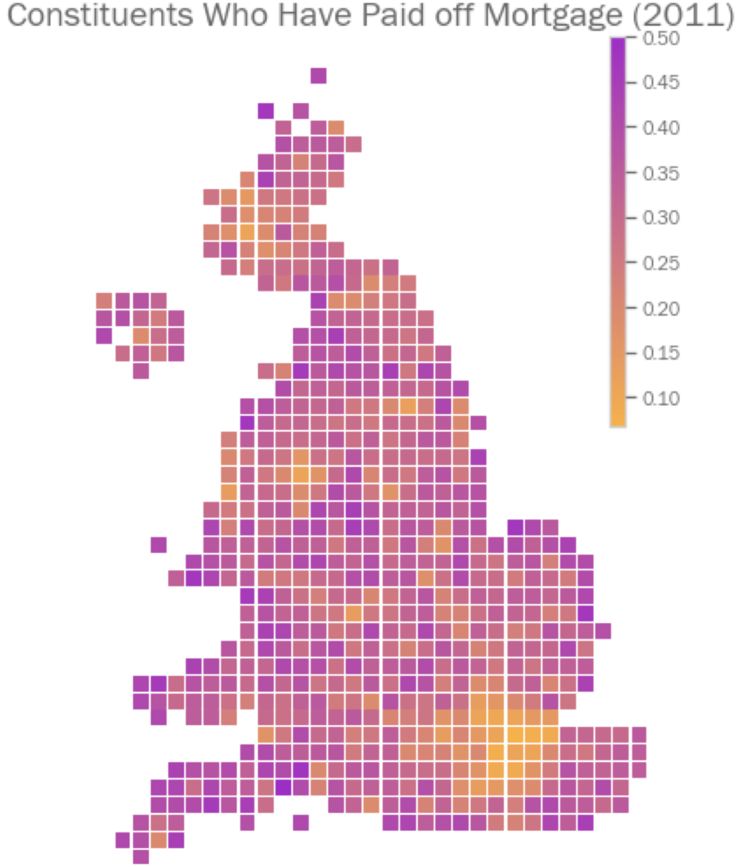
Housing tenure

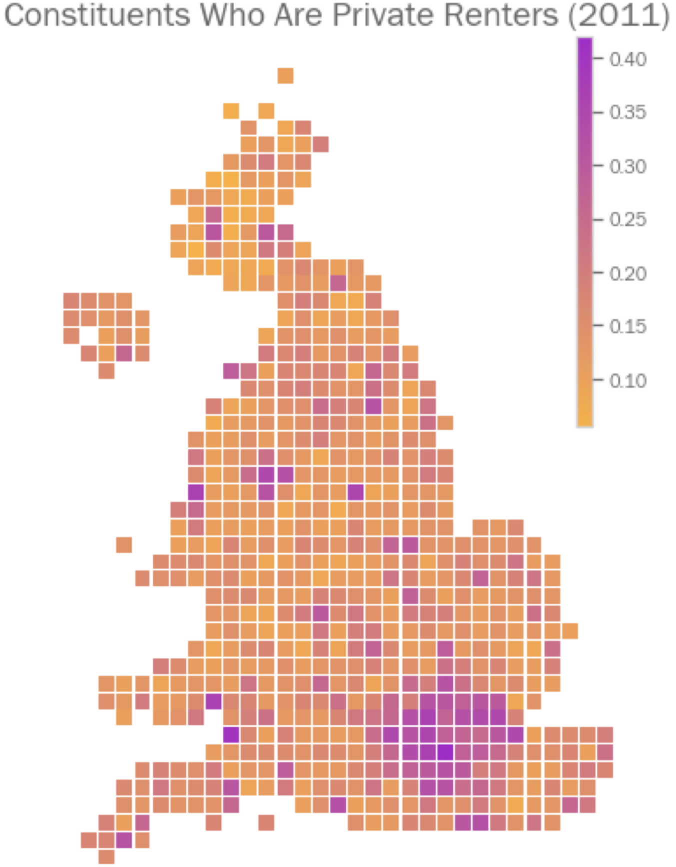


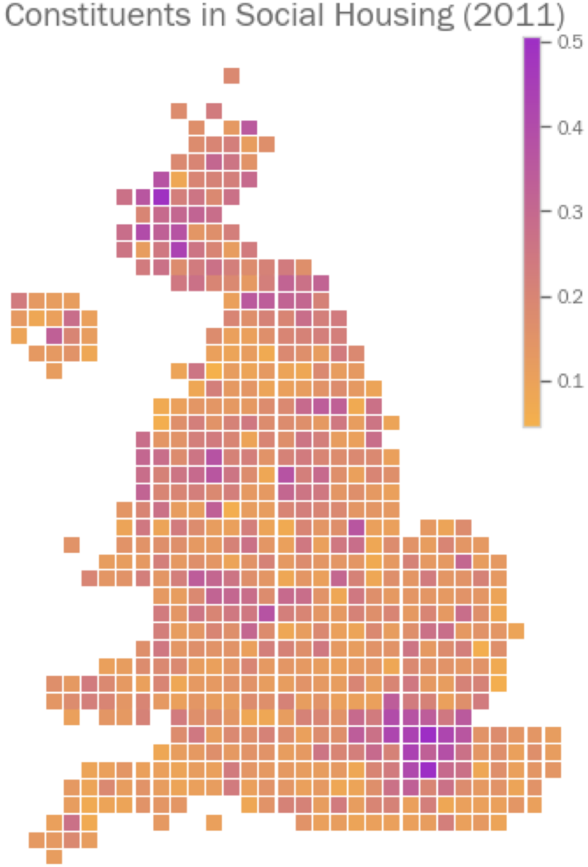


CB91

House ownership

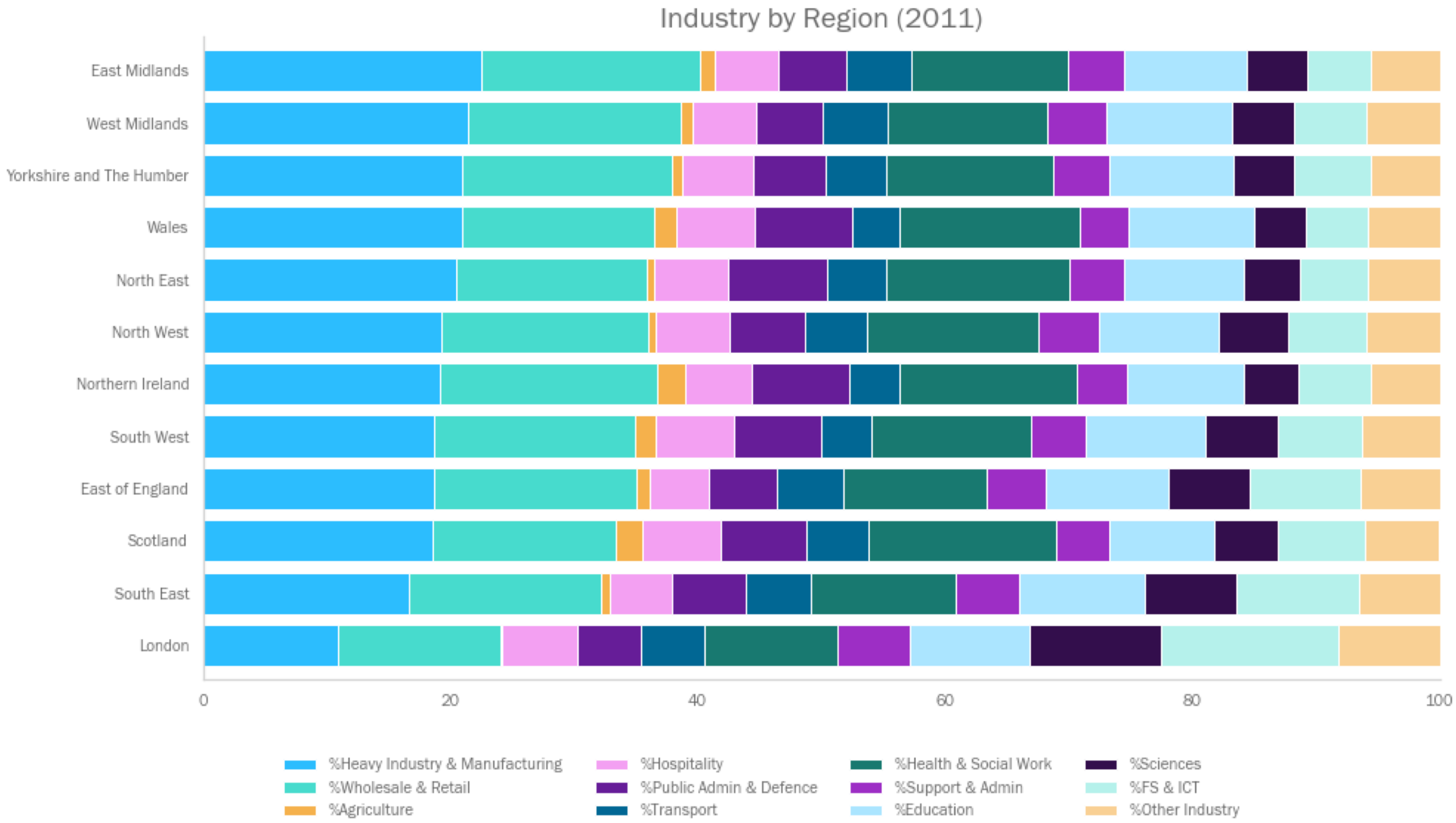






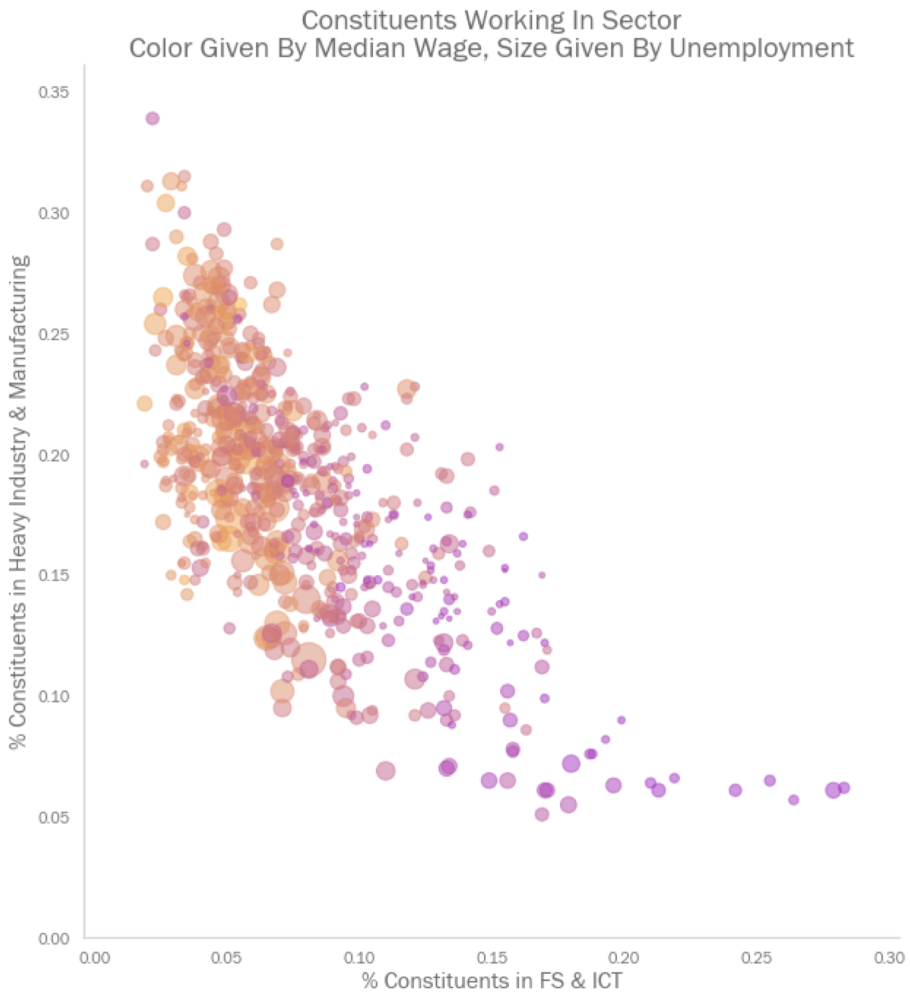
Additional Analysis

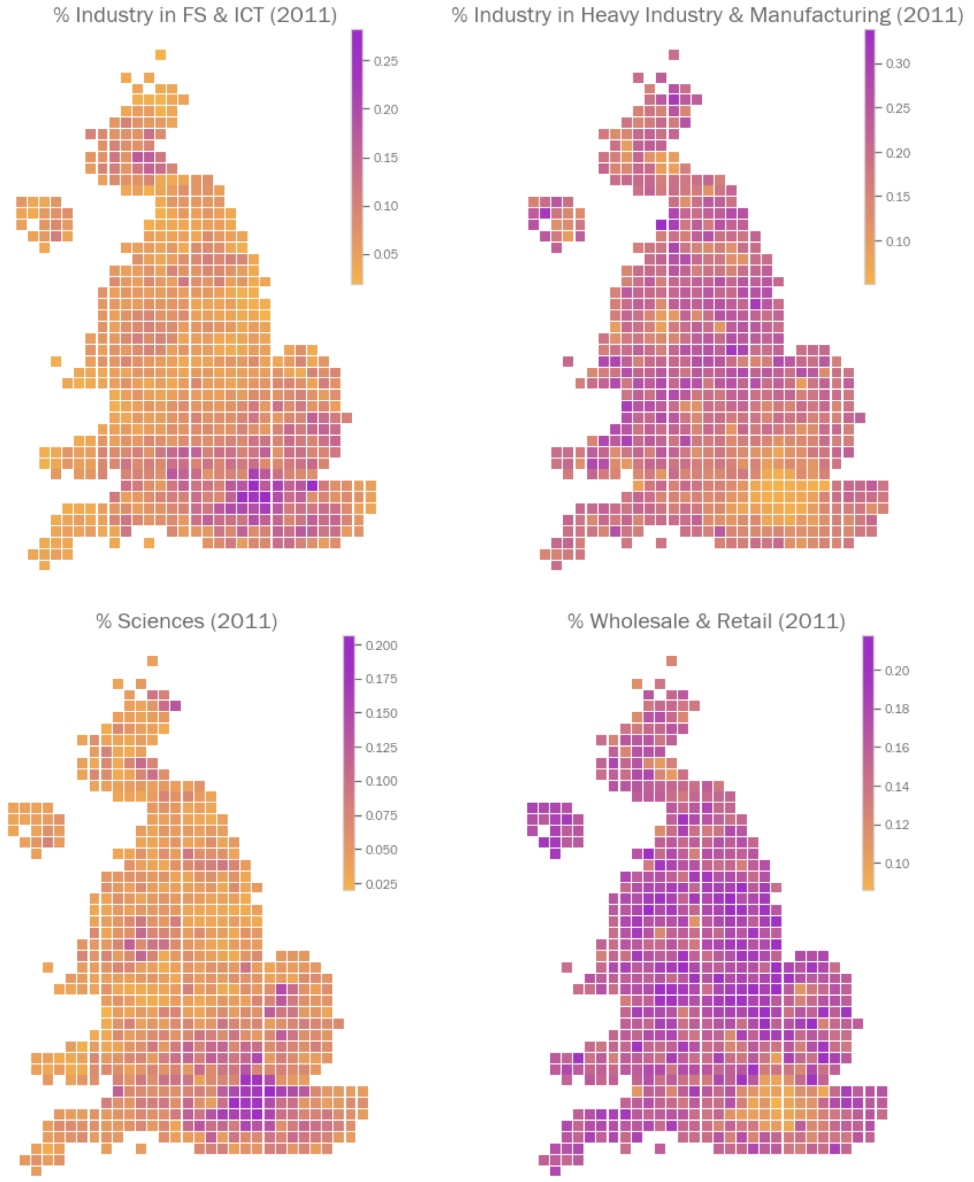
Industry



Additional Analysis

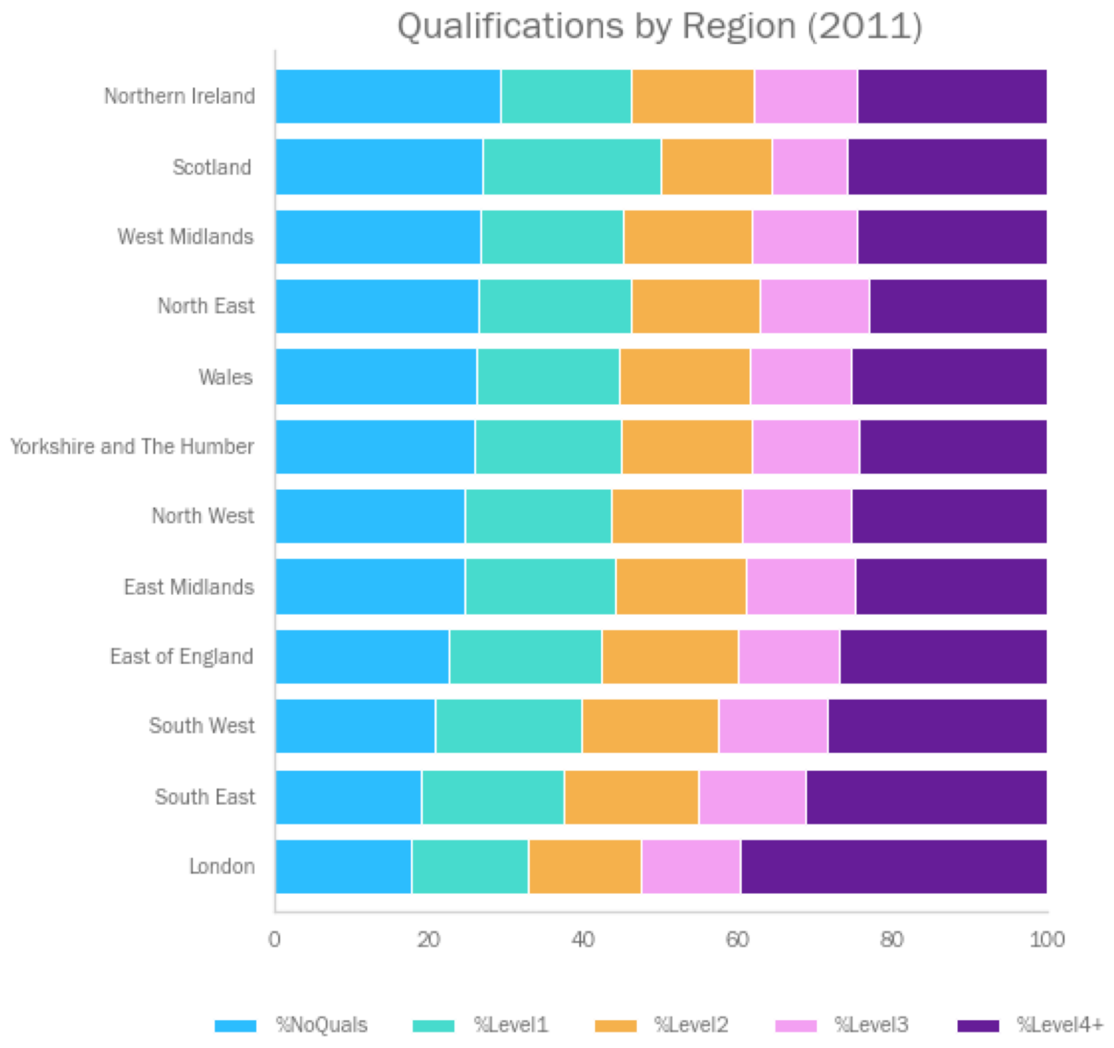
Employment by industry

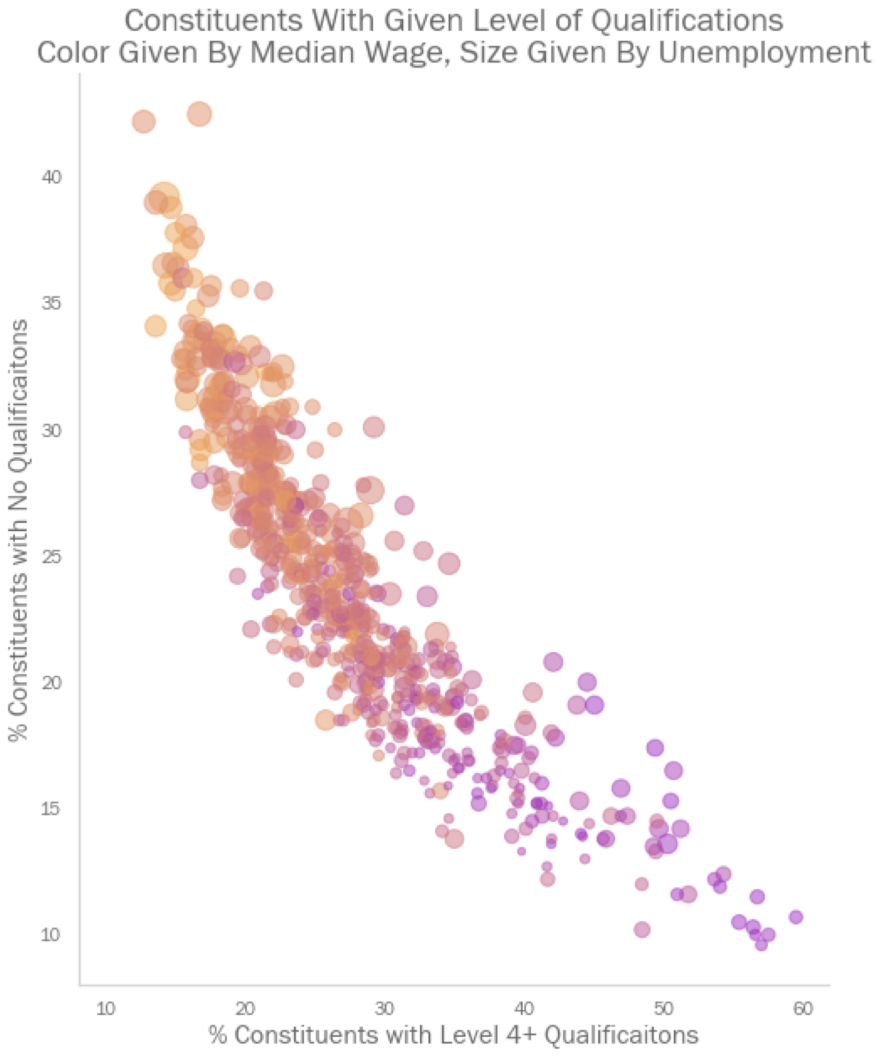




CB91

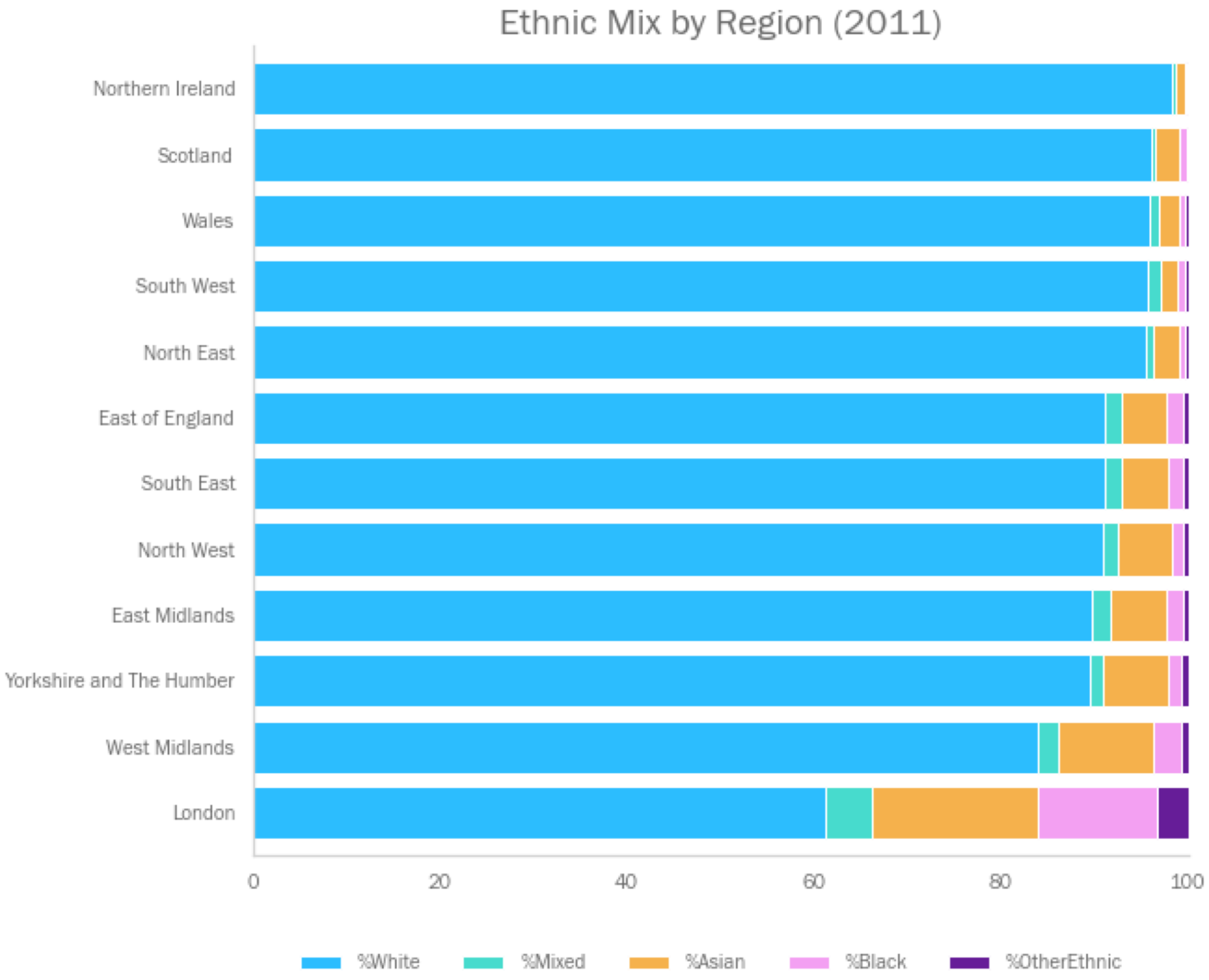
Qualification levels

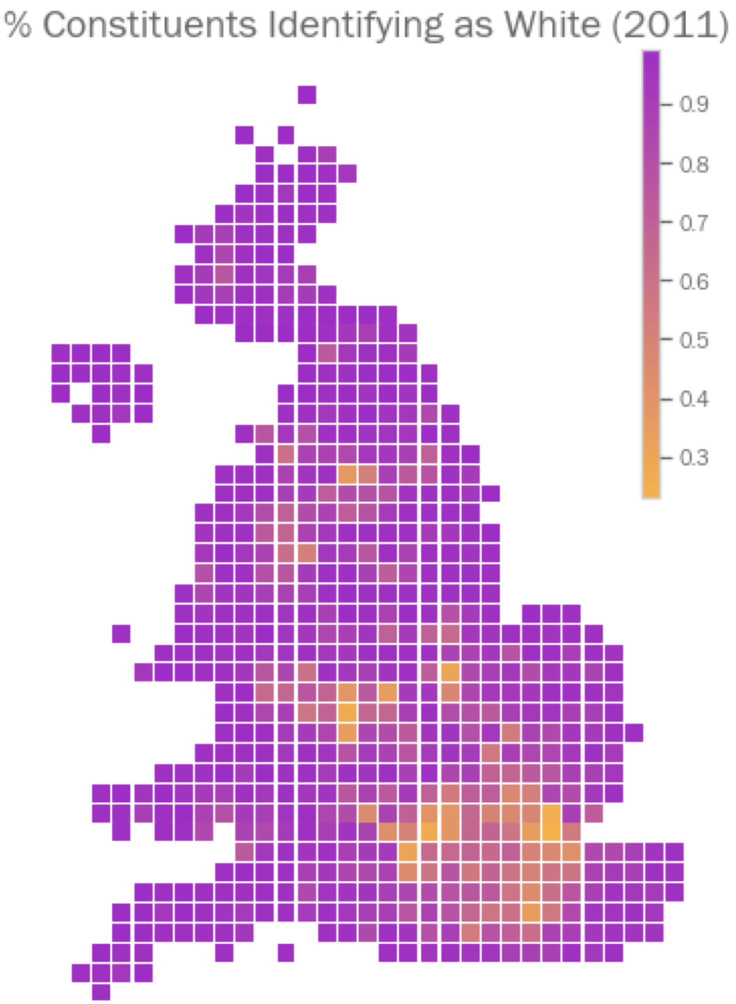




CB91

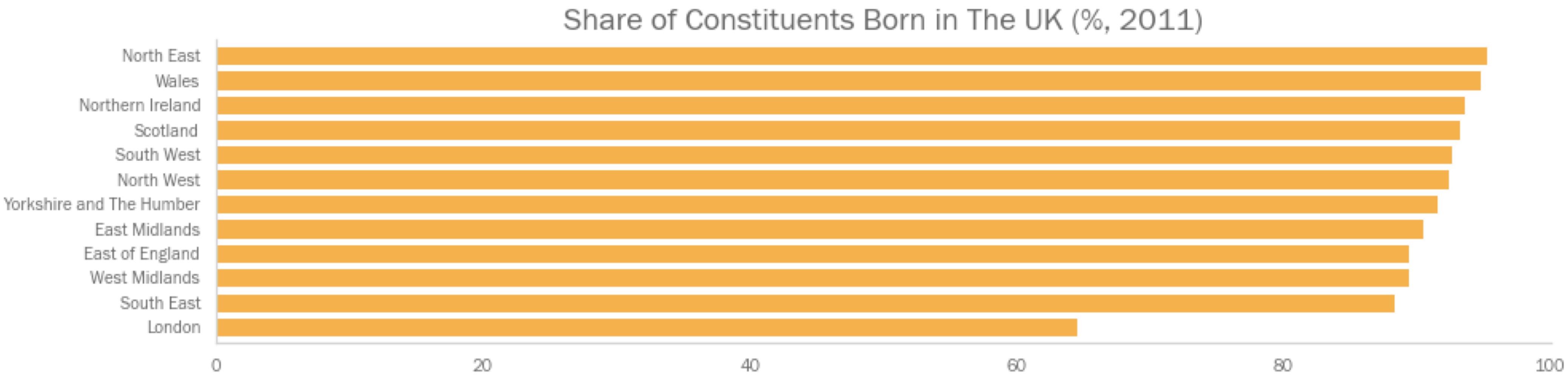
Ethnicity

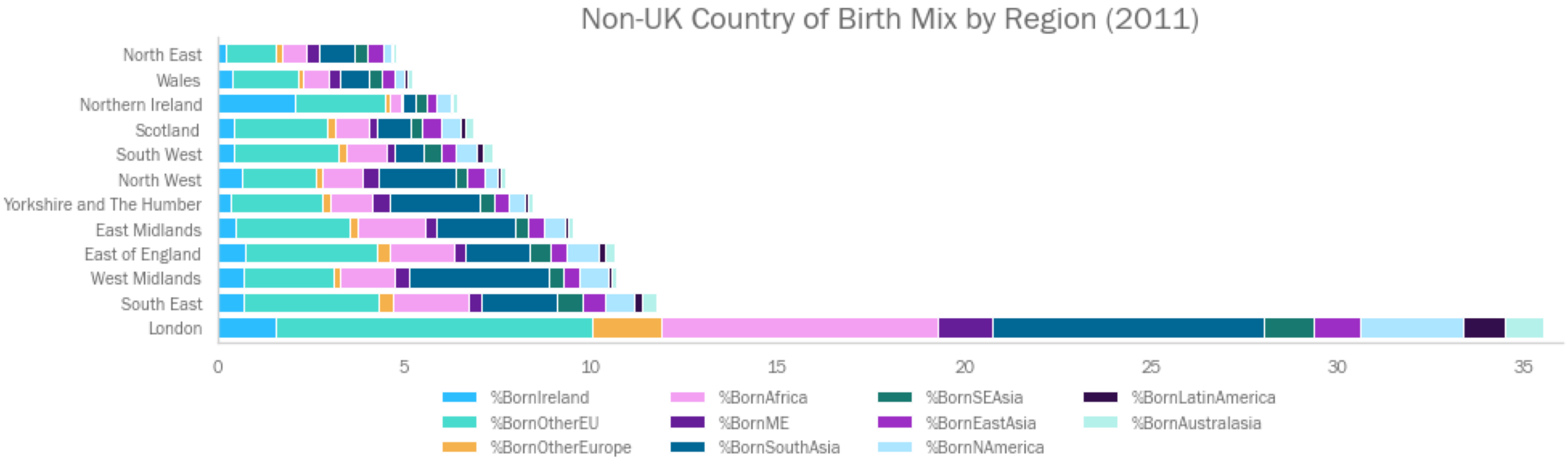




CB91

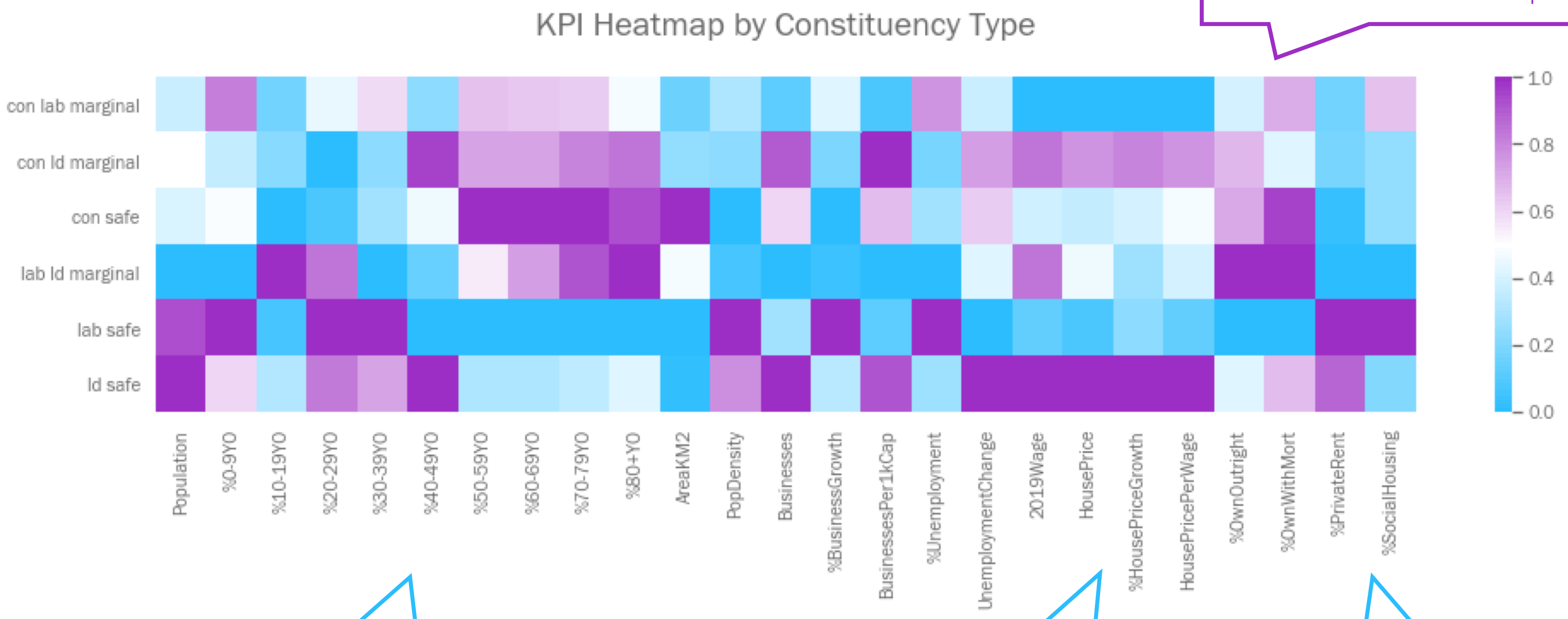
Country of birth by region





English seat type comparison (1/4)

Analysis considers the mean value for constituencies in each category, then normalises these means on a 0-1 scale to see how the different voter tribes compare.



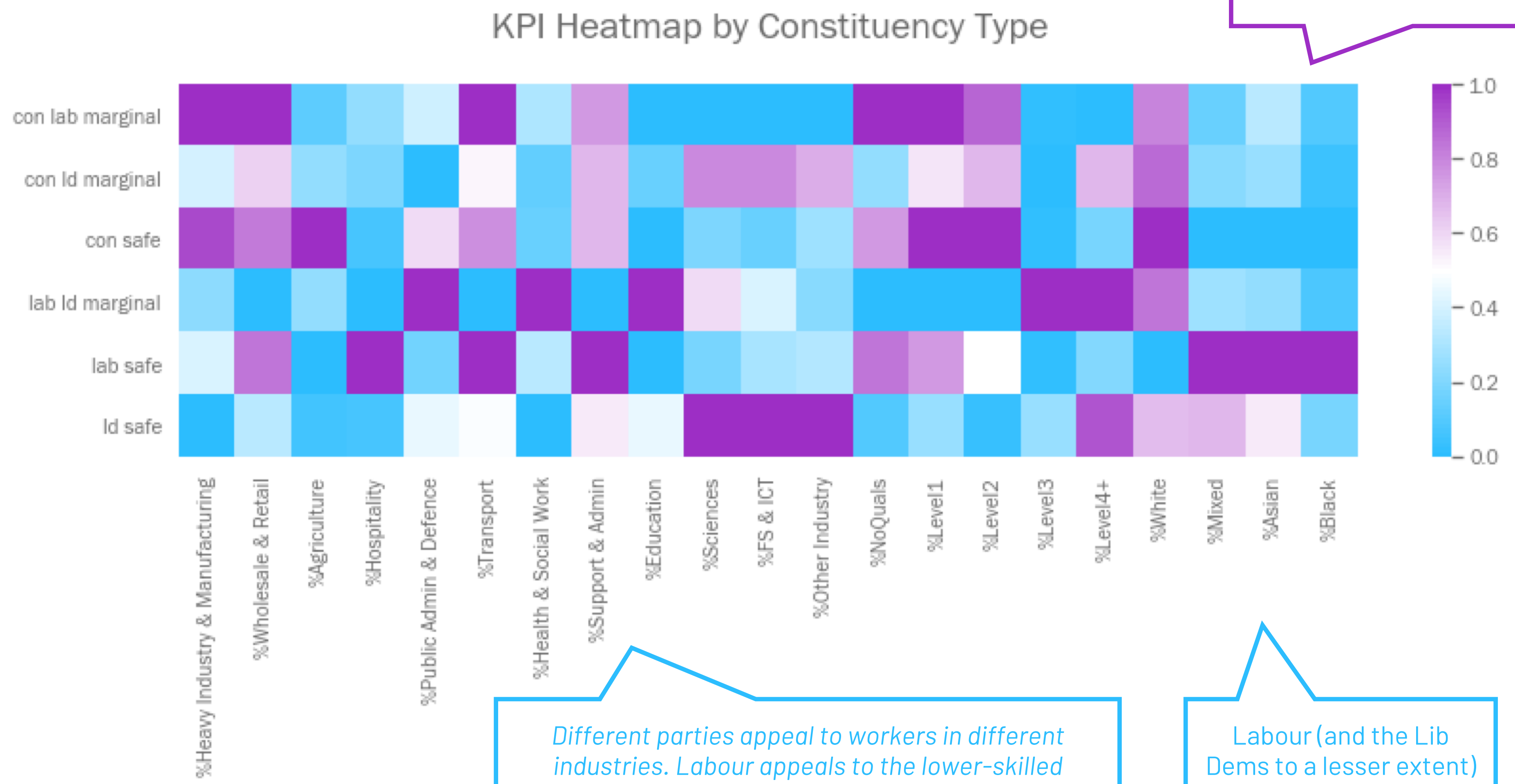
Younger constituencies tend to vote Labour and Liberal Democrat, whereas older voters tend Conservative.

Lib Dems were consistently strongest in areas with high wages, and high house prices.

Labour were especially strong in areas with high levels of renting and social housing.

English seat type comparison (2/4)

Analysis considers the mean value for constituencies in each category, then normalises these means on a 0-1 scale to see how the different voter tribes compare.

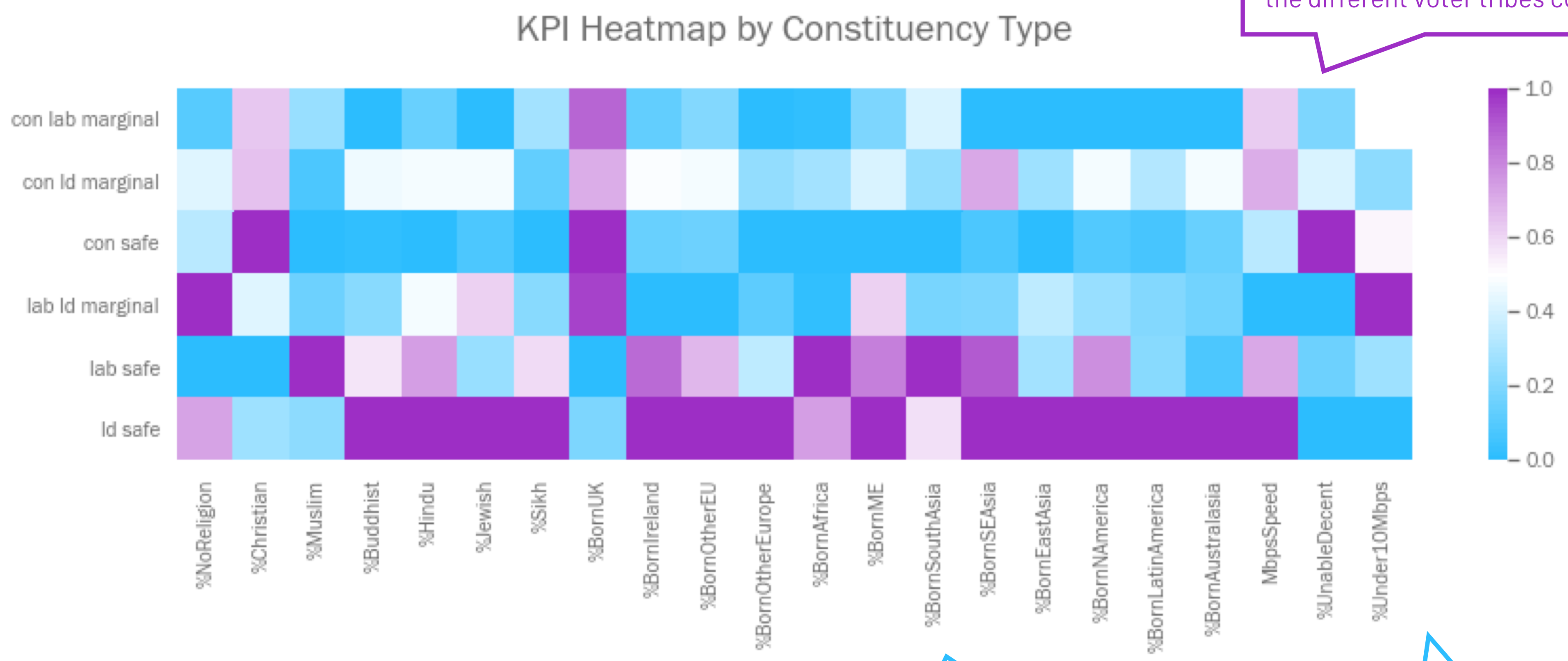


Different parties appeal to workers in different industries. Labour appeals to the lower-skilled service industries, Conservatives to heavy industry, and Lib Dems to high-powered white-collar work.

Labour (and the Lib Dems to a lesser extent) did better in areas with high ethnic diversity.

English seat type comparison (3/4)

Analysis considers the mean value for constituencies in each category, then normalises these means on a 0-1 scale to see how the different voter tribes compare.



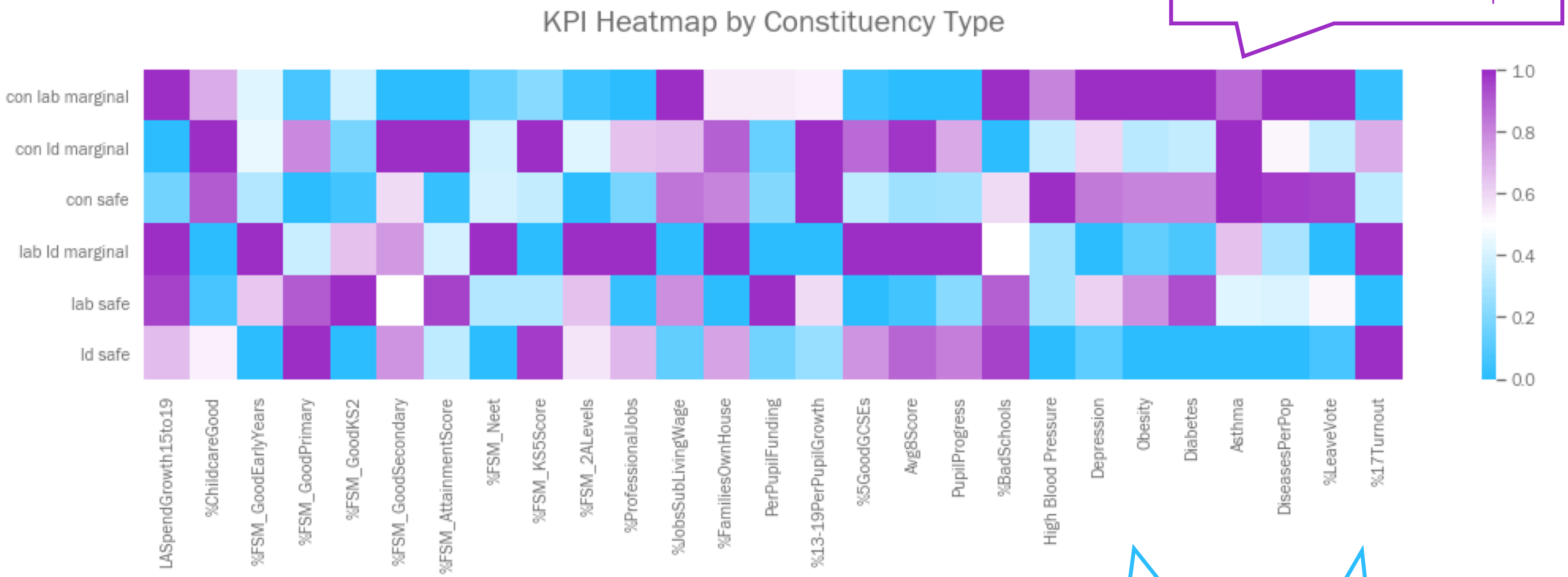
Conservatives strong in areas with high levels of self-declared Christians (typically older, whiter voters)

Lib Dems and Labour strong in areas with high levels of religious and ethnic diversity (Conservatives the opposite)

Conservatives very strong in places unable to get good broadband connections…

English seat type comparison (4/4)

Analysis considers the mean value for constituencies in each category, then normalises these means on a 0-1 scale to see how the different voter tribes compare.



Conservatives stronger in places with higher rates of chronic diseases. Again, this is likely to be due to older voters.

Competitive Lib Dem seats tend to have high levels of political engagement.