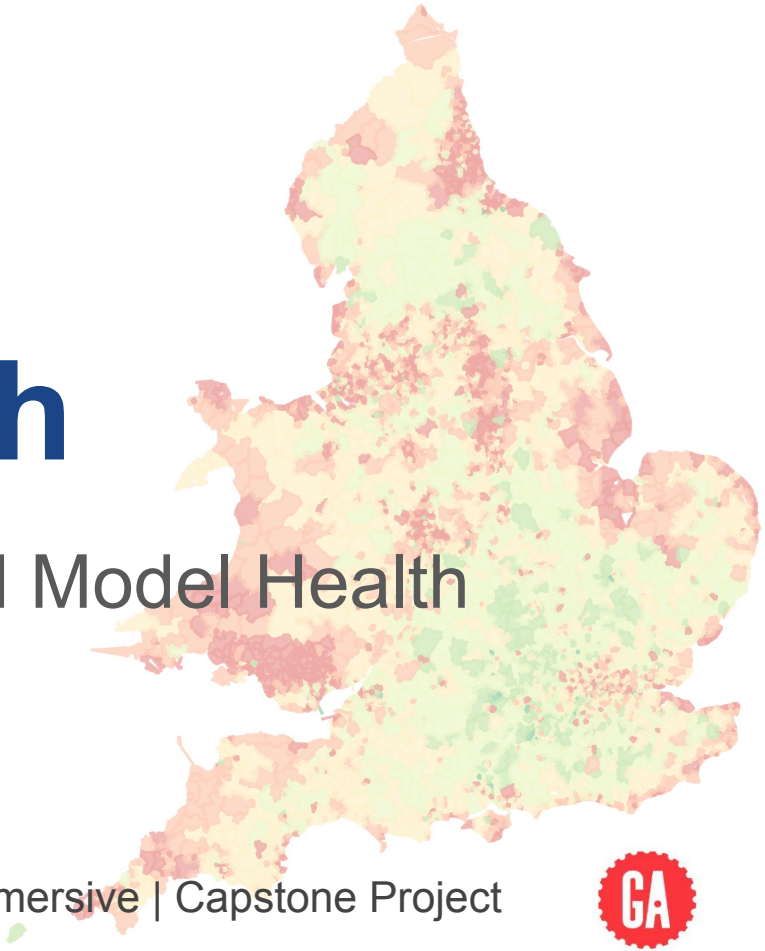
Picture of Health

Using Census Data to Map and Model Health in England and Wales



Catriona Reader | General Assembly Data Science Immersive | Capstone Project


We send the EU £350 million a week
let's fund our NHS instead
Vote Leave
Let's take back control

Motivation and Goals

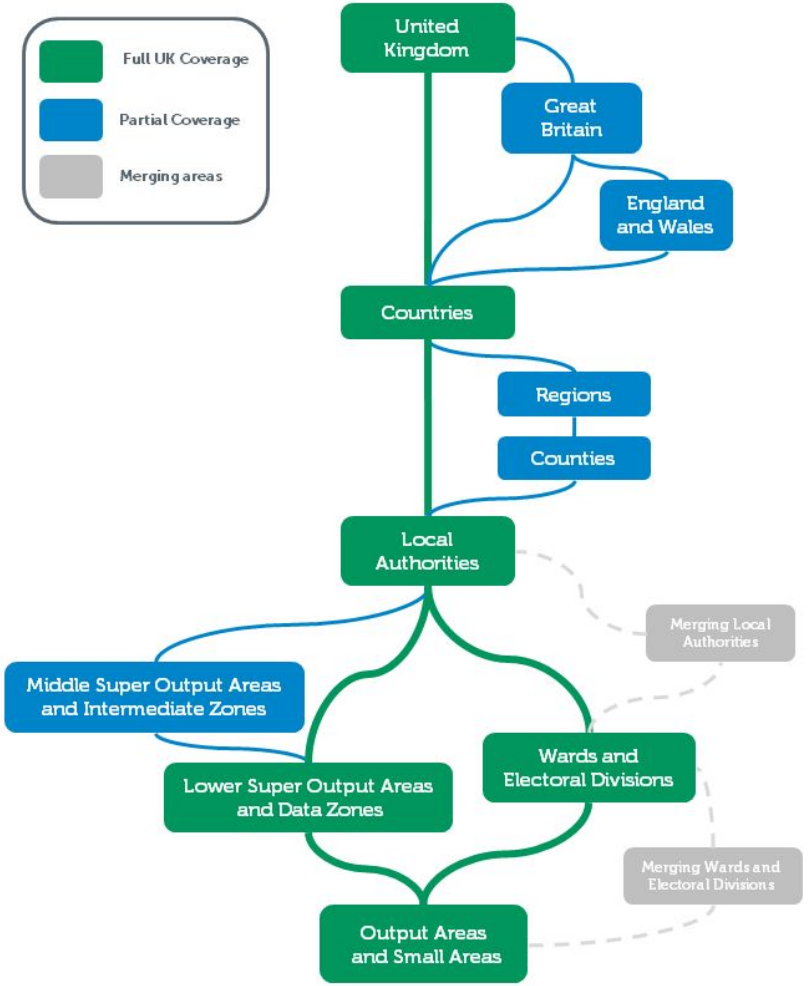
- NHS is under huge financial pressure.
- Want to model areas of the country that have a greater healthcare burden.
- The census is carried out once every 10 years, and a wealth of information from it is made available for public study.
- **This project seeks to use data from the most recent 2011 Census for England and Wales to map and model areas of poor health.**
- A model will be successful if it can be used to accurately predict levels of health based off information made available in the census, *as well as return useful information about what influences the model's results.*

Data

- 2011 Census
- Office for National Statistics
- Aggregated Data
- **Middle Super Output Area (MSOA)**

MSOAs are a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales.

They are built from groups of contiguous Lower Layer Super Output Areas. The minimum population is 5000 and the mean is 7200.



Overall Approach

- **Cleaning and joining data** - important to really understand the data you are working with
- **Exploratory Data Analysis** and mapping - let the data speak
- **Defining the Target**
- **Modelling**
- **Interpreting results**

Cleaning and joining data

Census downloads are grouped by Key Statistic. This project used:

- Usual Resident Population
- Living Arrangements
- Health and provision of unpaid care
- Education and Qualifications

Joined by unique geographic identification code for the MSOA.

Used lookup tables from the ONS Geography Portal to link the MSOAs to Local Authorities.

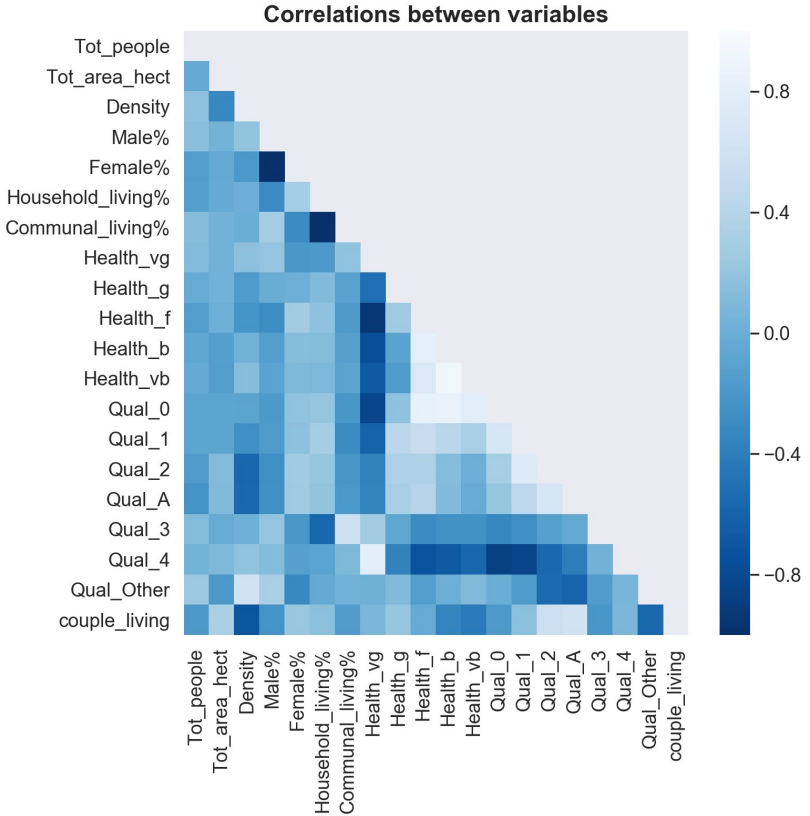
Boundary data available from the ONS geography portal in the format of shapefiles, read with the PyShp library.

```
print(data.shape)
data.head(2)
```

```
(7201, 26)
```

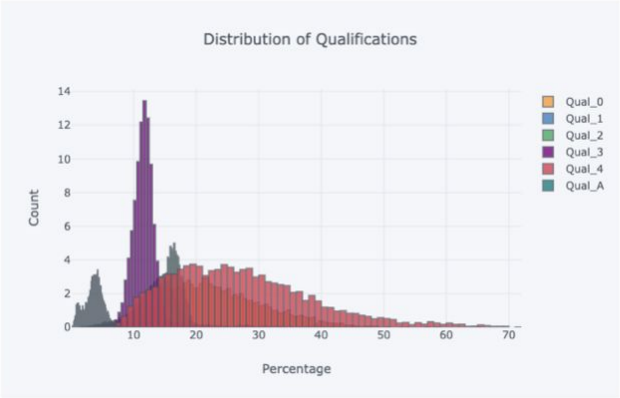
	GeographyCode	coords	MSOA	LACode	LA	Tot_people	Tot_area_hect	Density	Male%	Female%	...	Health_f	Health_
0	E02003113	[(368710.883, 173101.639), (368667.312, 173141...	South Gloucestershire 024	E06000025	South Gloucestershire	10167.0	17027.36	0.597098	49.522966	50.477034	...	10.907839	2.86220
1	E02001245	[(393356.551, 398022.687), (393367.892, 398040...	Tameside 017	E08000008	Tameside	7848.0	262.79	29.864150	49.490316	50.509684	...	14.513252	5.72120

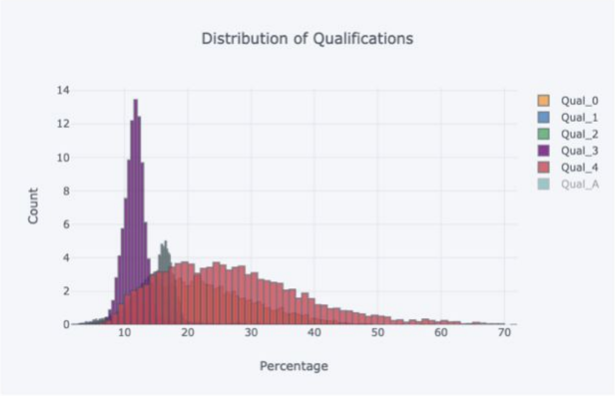
Exploratory Data Analysis

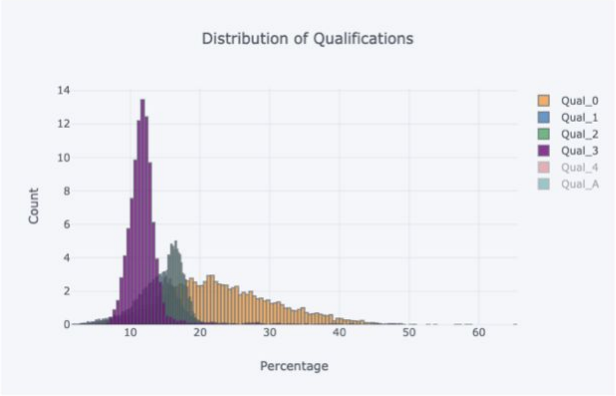


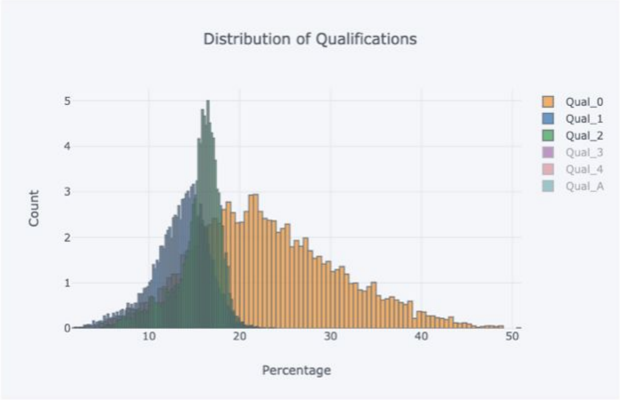
- Some entries are too highly correlated to be included in predictor features.
- This multicollinearity causes problems with modelling and interpreting the results.
- Drop these prior to modelling

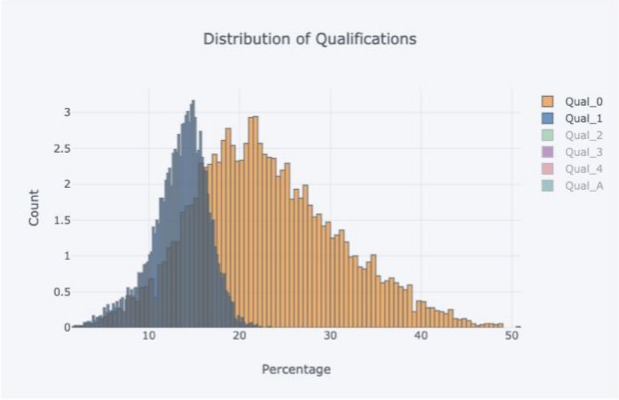
EDA

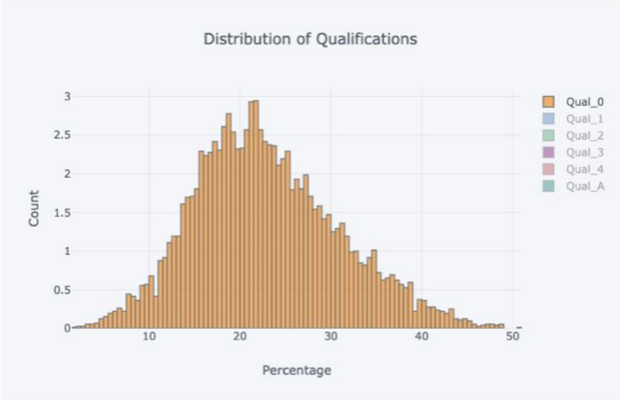

Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A


Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A

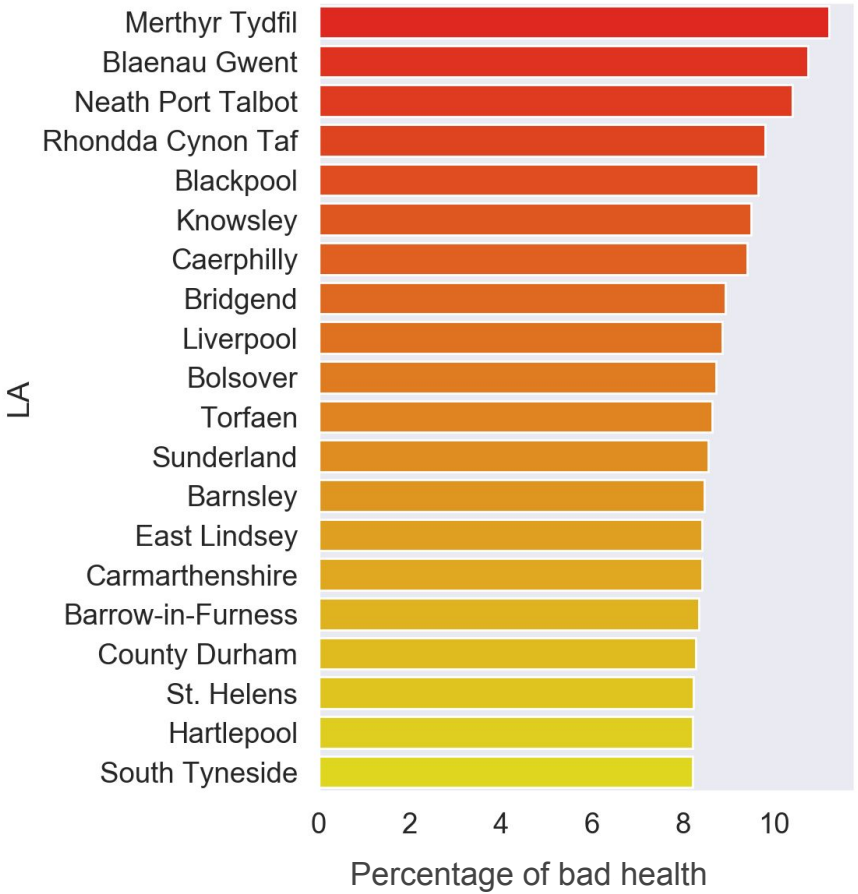

Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A

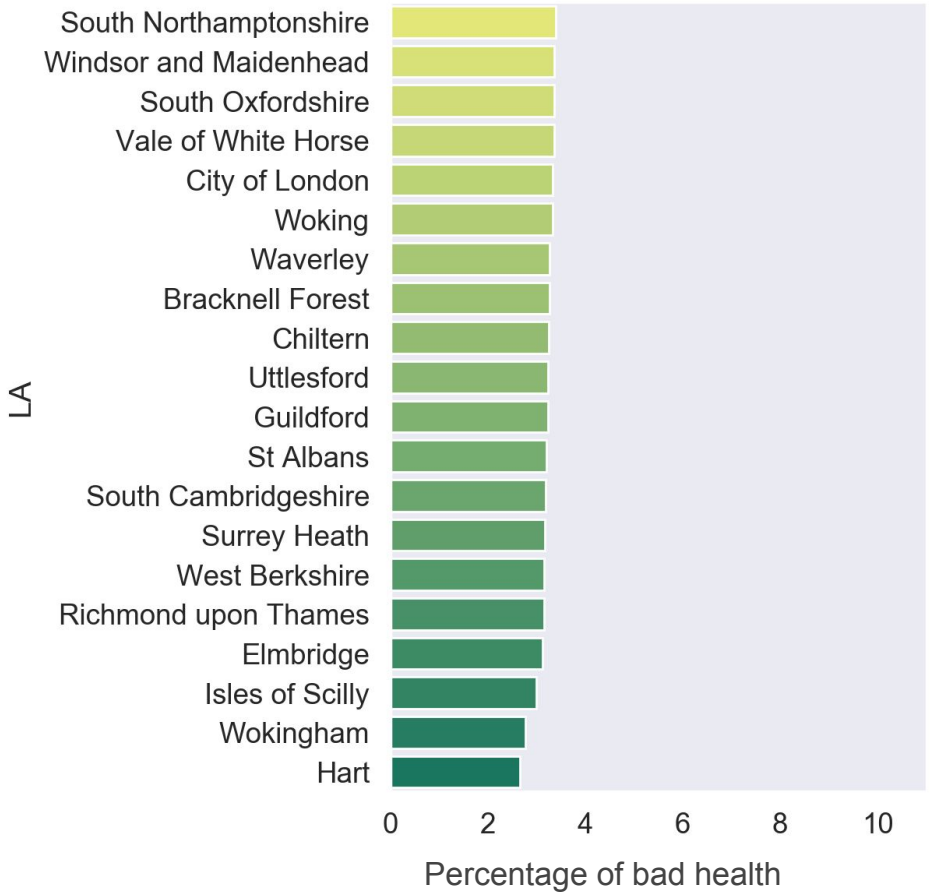

Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A


Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A

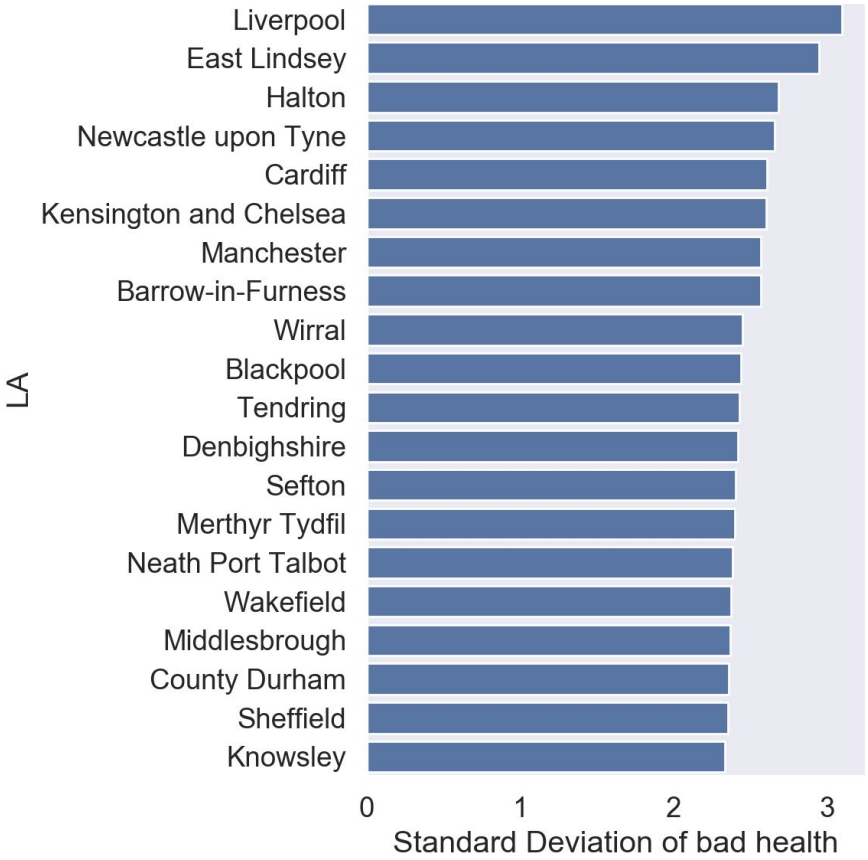

Distribution of Qualifications
Count
Percentage
Qual_0
Qual_1
Qual_2
Qual_3
Qual_4
Qual_A

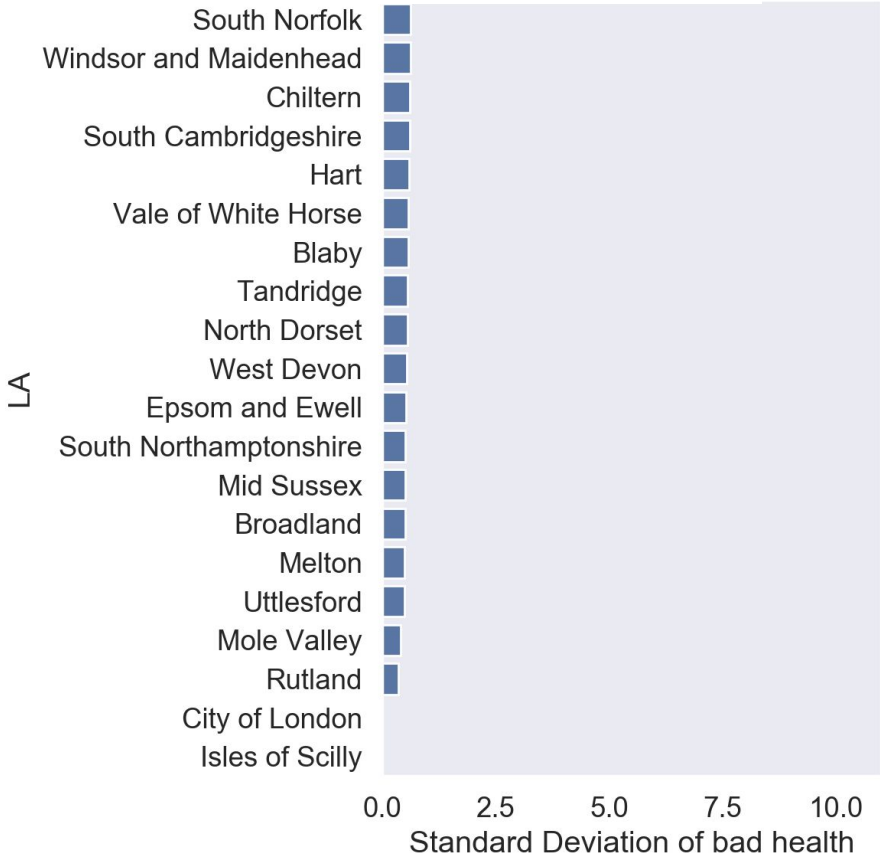
Local Authorities with highest (left) and lowest (right) density of bad health



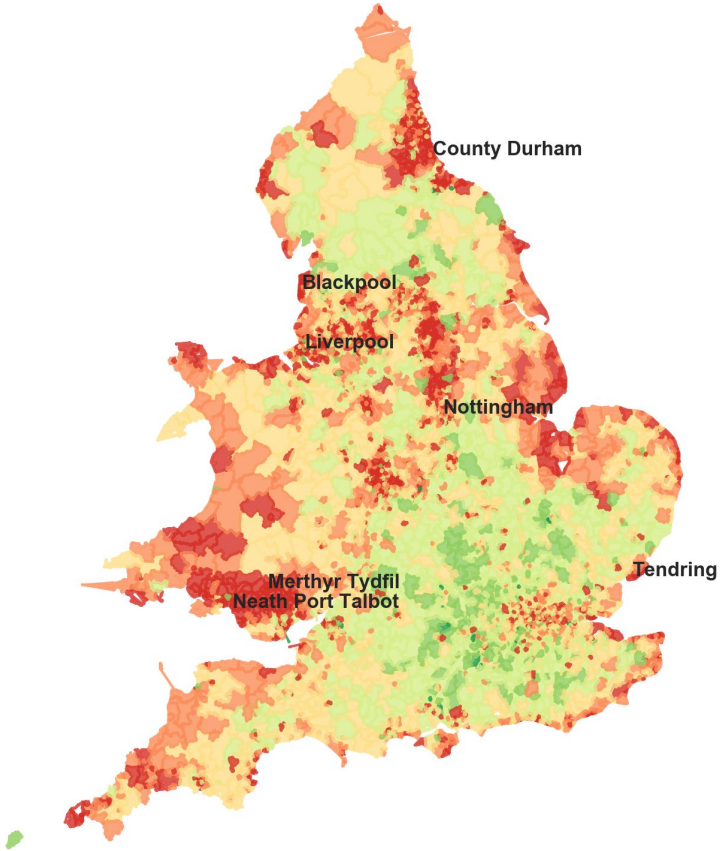


Local Authorities with highest (left) and lowest (right) variation of bad health





Heatmap showing areas of poor health in England and Wales



Percentage of bad health per MSOA

- Less than 2%
- Less than 3%
- Less than 4.11%(lower quartile)
- Less than 5.24% (second quartile)
- Less than 6.82% (upper quartile)
- More than 6.82%

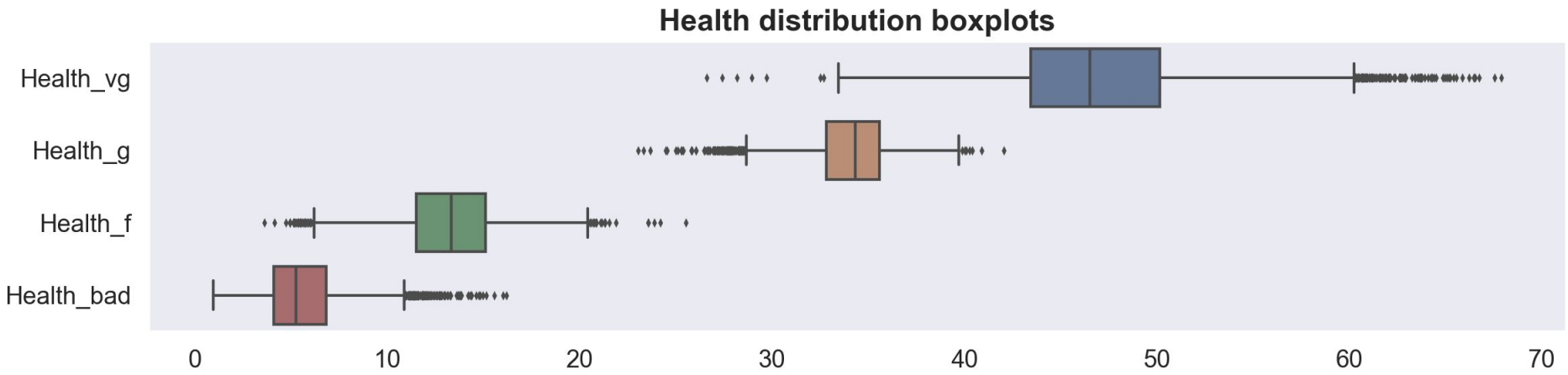
Names of Local Authorities are shown if an MSOA in that area has more than 14% reported bad health.

There is a strong concentration of green around Greater London and its surrounds, as well as the Yorkshire Dales.

Areas of poor health are more concentrated at the fringes or edges. Wales looks particularly bad, as does the area around Newcastle, and the Lincolnshire coast.

Defining the target

- Regression model with the percentage of bad health ('health_bad') as the target
- Boxplots show that the 'health_bad' distribution has many outliers - choose not to remove them
- In terms of policy implementation to improve areas of poor health, it is most important to be able to identify the worst affected areas, and the features or conditions that influence them, even if this means that my model performs less well overall.



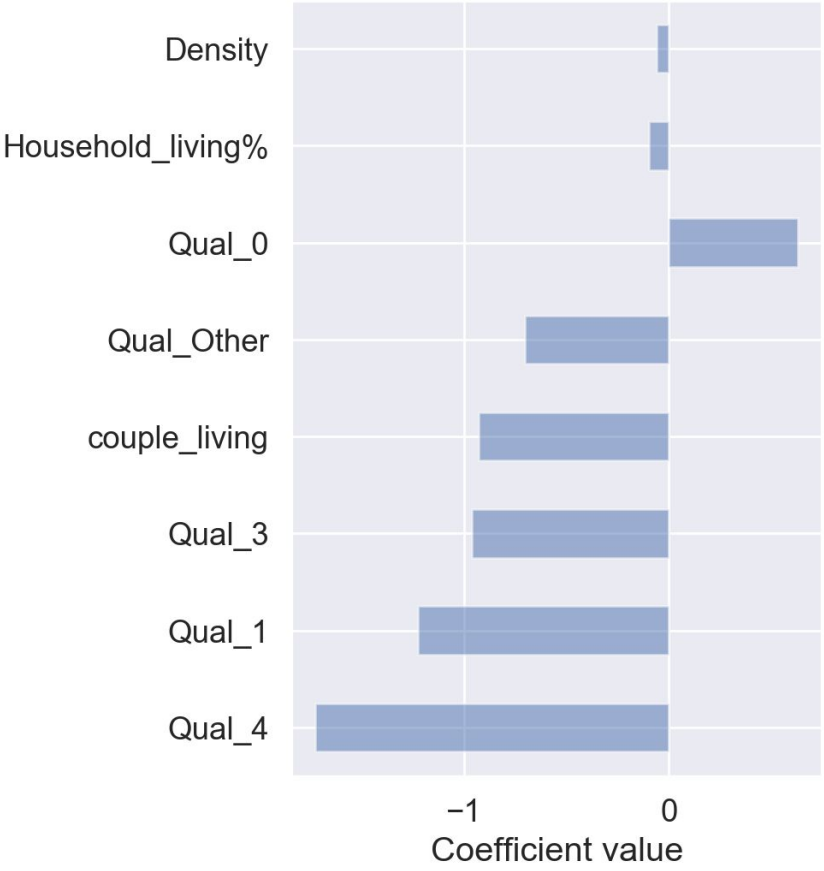
Modelling with sklearn

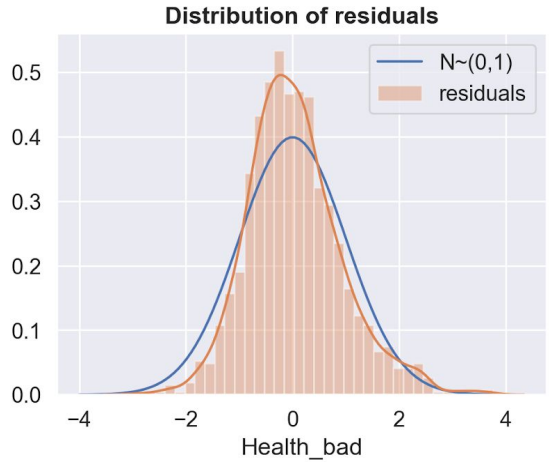
List of models I applied, along with their highest cross validated r-squared scores on the training data:

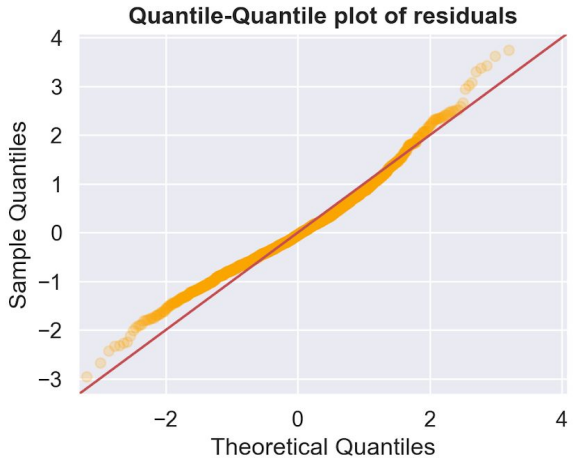
- **Linear regression model with lasso penalty: 0.82084**
- Linear regression model with ridge penalty: 0.81938
- Linear regression model with elastic-net penalty: 0.82040
- KNeighborsRegressor: 0.759533
- DecicisionTreeRegressor: 0.723733
- RandomForestRegressor: 0.7971391
- BaggingRegressor (decision tree base estimator): 0.82644

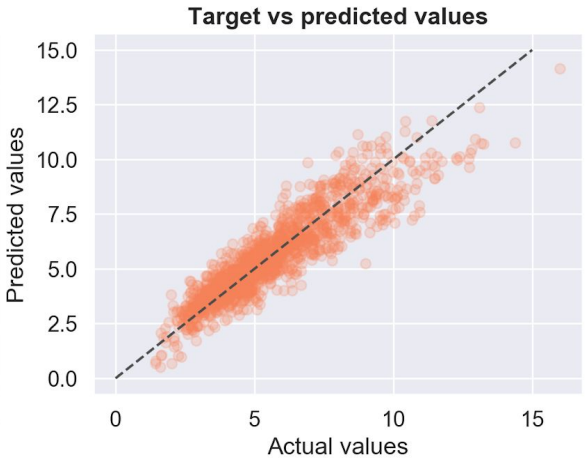
Although the BaggingRegressor had the highest cross validated score, it is a black-box model, making it less useful for interpretation - will disregard.

Lasso Feature Importances




Distribution of residuals
N~(0,1)
residuals
0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
Health_bad


Quantile-Quantile plot of residuals
Sample Quantiles
4
3
2
1
0
−1
−2
−3
−2
0
2
4
Theoretical Quantiles

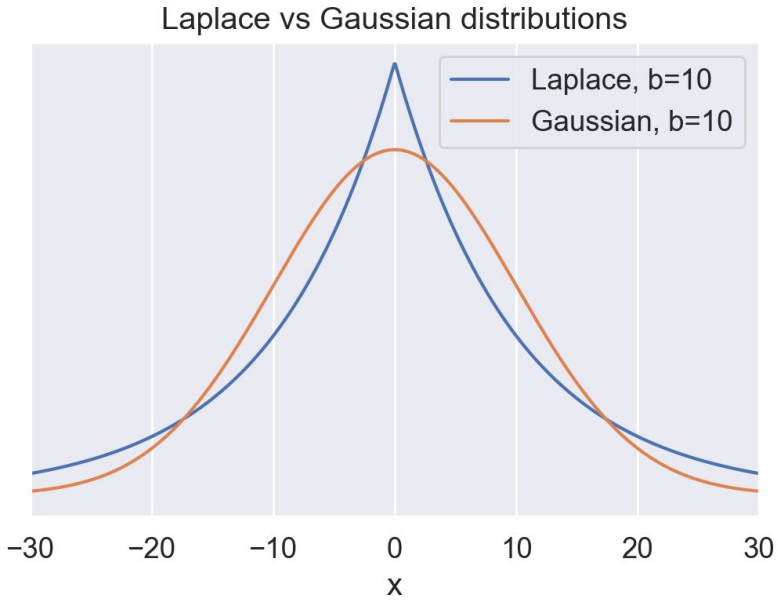

Target vs predicted values
Predicted values
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
5
10
15
Actual values

Modelling with PyMC3

I decided to implement a Bayesian linear regression, so that I could add credible intervals to the beta coefficients.

These credible intervals will quantify uncertainty in the model.

I used priors with a Laplace distribution for the betas to simulate the lasso penalty of my best linear regression model.



```
with pm.Model() as reg_laplace:

    #defining the priors
    y_std = pm.Uniform('error_std', lower=0.01, upper=10.)
    intercept = pm.Normal('intercept', mu=df_bayes.y.mean(), sd=df_bayes.y.std())
    Tot_people_beta = pm.Laplace('Tot_people_beta', mu=0, b=10)
    Tot_area_hect_beta = pm.Laplace('Tot_area_hect_beta', mu=0, b=10)
    Density_beta = pm.Laplace('Density_beta', mu=0, b=10)
    Female_beta = pm.Laplace('Female_beta', mu=0, b=10)
    Household_living_beta = pm.Laplace('Household_living_beta', mu=0, b=10)
    Qual_0_beta = pm.Laplace('Qual_0_beta', mu=0, b=10)
    Qual_1_beta = pm.Laplace('Qual_1_beta', mu=0, b=10)
    Qual_2_beta = pm.Laplace('Qual_2_beta', mu=0, b=10)
    Qual_3_beta = pm.Laplace('Qual_3_beta', mu=0, b=10)
    Qual_4_beta = pm.Laplace('Qual_4_beta', mu=0, b=10)
    Qual_Other_beta = pm.Laplace('Qual_Other_beta', mu=0, b=10)
    couple_living_beta = pm.Laplace('couple_living_beta', mu=0, b=10)

    #defining the likelihood
    E_y = pm.Normal('y_mean',
                    mu=intercept + (Tot_people_beta * df_bayes.Tot_people) +
                    (Tot_area_hect_beta * df_bayes.Tot_area_hect) +
                    (Density_beta * df_bayes.Density) +
                    (Female_beta * df_bayes.Female) +
                    (Household_living_beta * df_bayes.Household_living) +
                    (Qual_0_beta * df_bayes.Qual_0) +
                    (Qual_1_beta * df_bayes.Qual_1) +
                    (Qual_2_beta * df_bayes.Qual_2) +
                    (Qual_3_beta * df_bayes.Qual_3) +
                    (Qual_4_beta * df_bayes.Qual_4) +
                    (Qual_Other_beta * df_bayes.Qual_Other) +
                    (couple_living_beta * df_bayes.couple_living),
                    sd=y_std,
                    observed=df_bayes.y)
```

```
with reg_laplace:
    trace_laplace = pm.sample(6000, tune=2500, n_jobs=3,
                      nuts_kwargs={'target_accept':0.95})
```

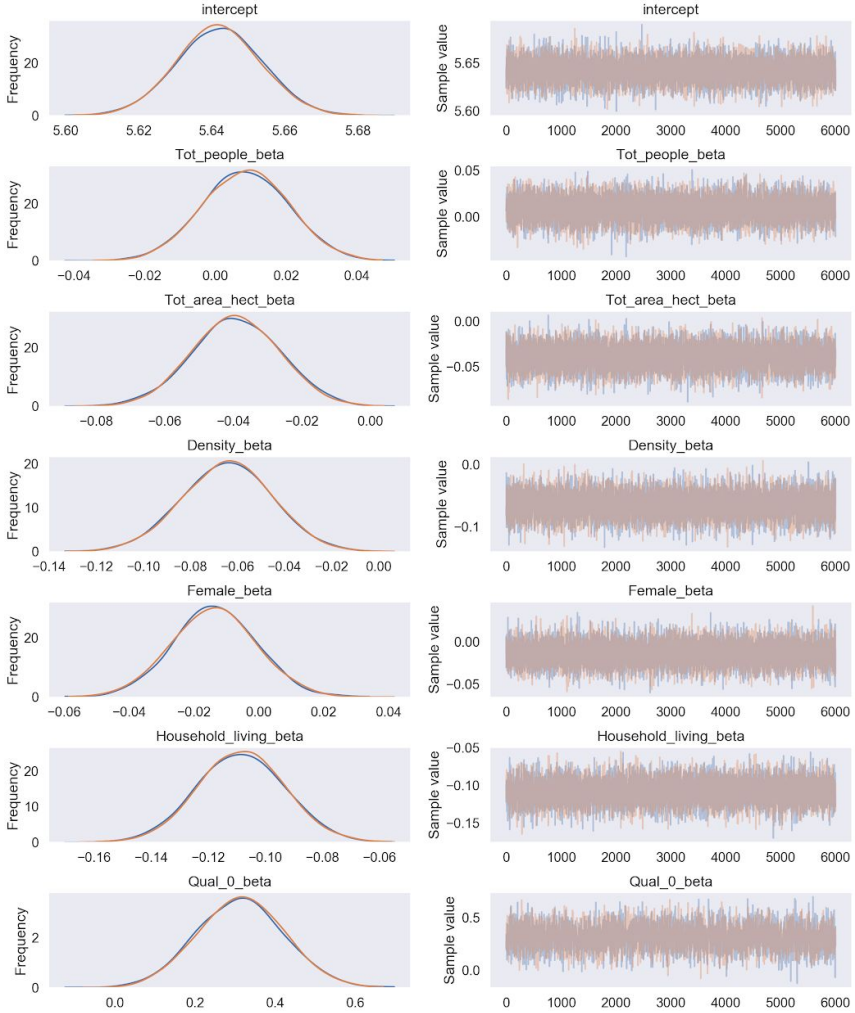
1. **Set up prior distributions**
2. **Define the likelihood function**
3. **Start the sampling procedure**

Posteriors: results will be distributions for the range of most likely model parameters

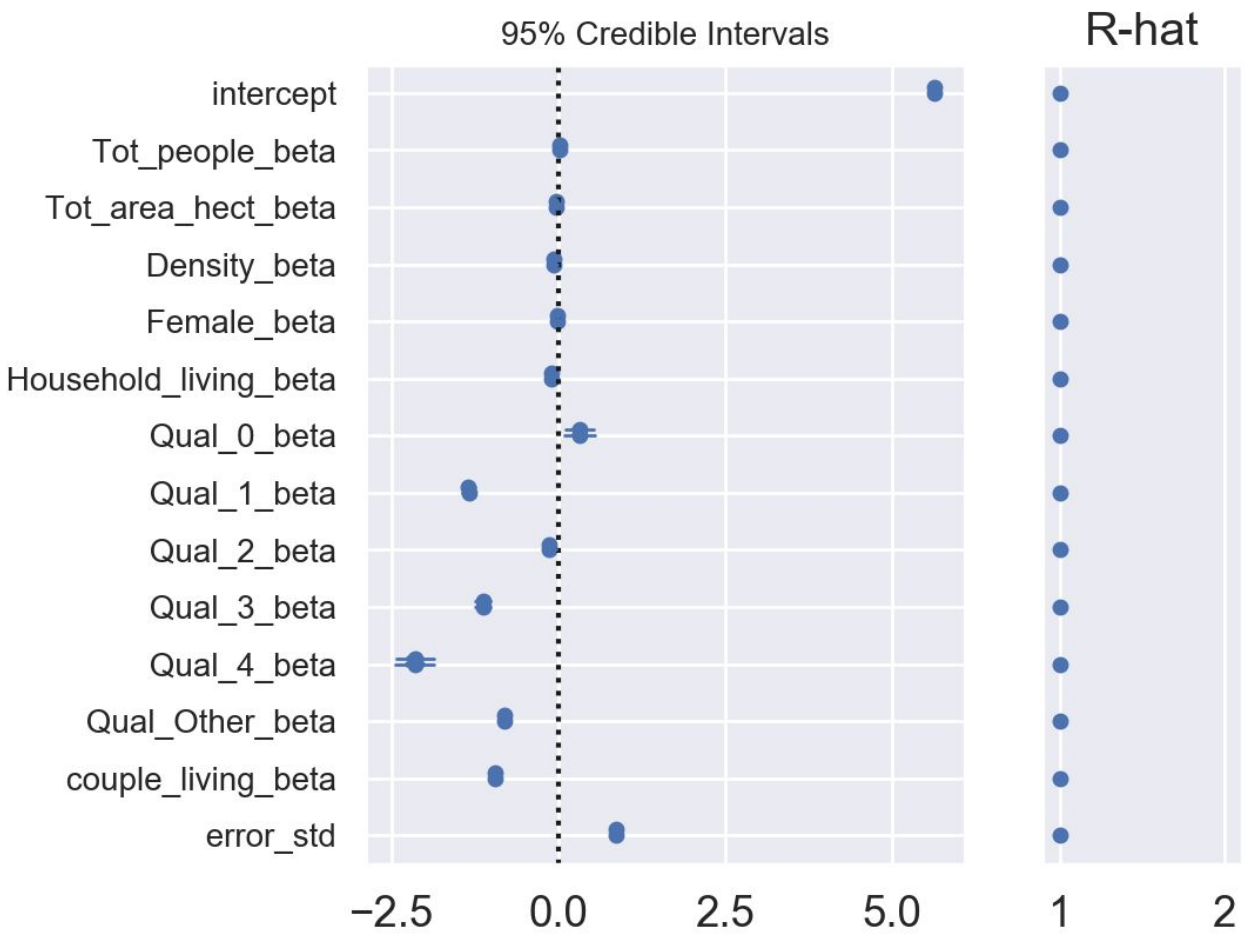
These distributions express levels of uncertainty in the model

As the amount of observed data increases, the influence of the priors decreases.

traceplots

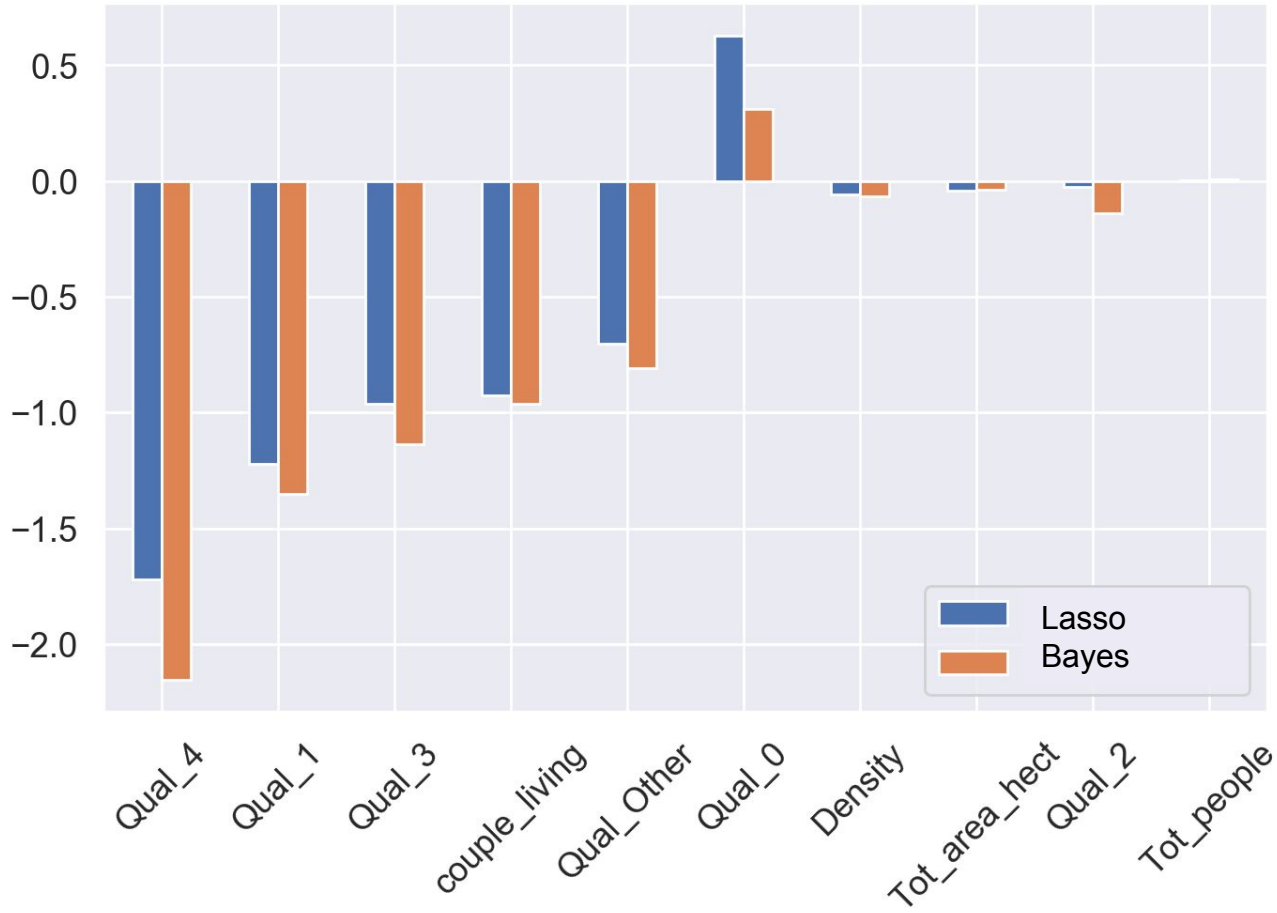

intercept
Frequency
Sample value
Tot_people_beta
Tot_area_hect_beta
Density_beta
Female_beta
Household_living_beta
Qual_0_beta
Qual_1_beta
Qual_2_beta
Qual_3_beta
Qual_4_beta
Qual_Other_beta
couple_living_beta
error_std

forest plot



R-squared score: 0.8302

comparison: lasso vs bayes



Findings

- Education levels are strong indicators of a population's health.
- An area with a high proportion of its population without any qualifications is likelier to have higher levels of bad health.
- Conversely, any level of education adds to predicting lower levels of bad health, and the term that contributes most negatively to bad health within an area (i.e. contributes most positively to predicting better health) is the percentage of the area's population with tertiary qualifications.
- The percentage of an area's population that is living in a couple contributes to a lower level of bad health.

Limitations

- Limited predictor features
- Skewed distribution of poor health

Next steps

- Find ways to further validate this model. Test on Scotland and Northern Ireland, or see if the model works as well on different levels of aggregate data.
- Find and isolate an area of England or Wales that has focussed on increasing education spend - does health improve? Do healthcare costs go down?
- Putting the code into a model that could be used by a Local Authority or policy holder to predict changes in overall health based on budgeting decisions. For example: if x amount more is spent on education, what would the overall increase in health look like?

Questions or comments?

Thank you