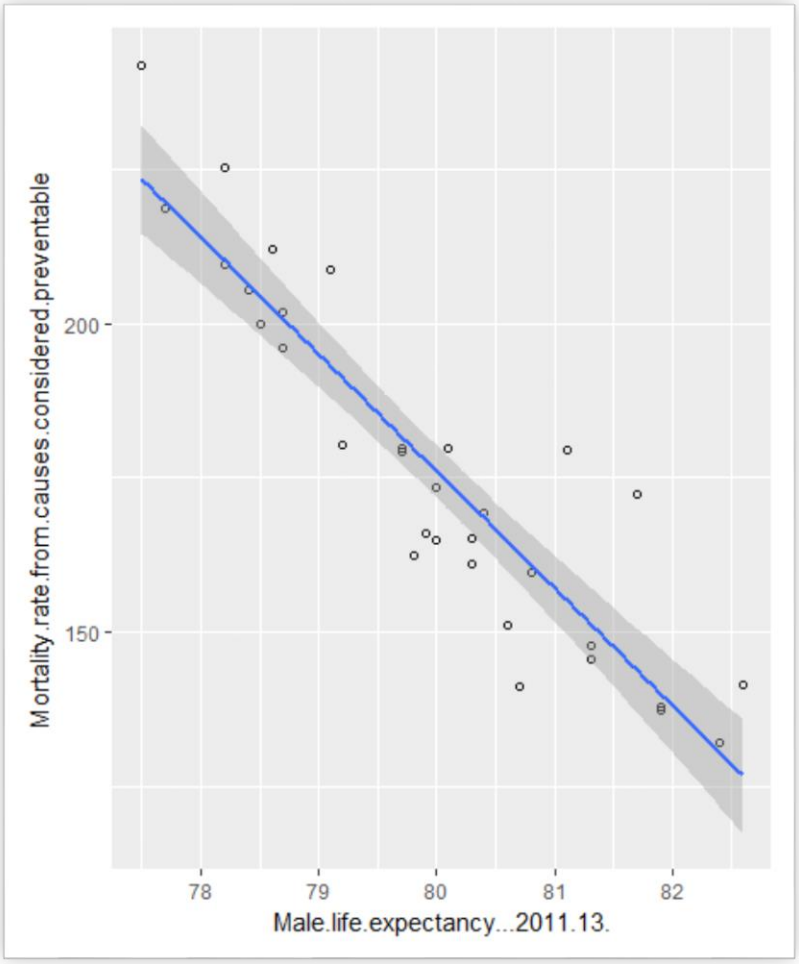


Understanding London using Open Data

A DATA DRIVEN ESSAY

RQJG1 | GEOGG125 Principles of Spatial Analysis | January 9, 2017

Contents

Word count: 1985 words

1 – Introduction

The purpose of the following report is to analyse and critically assess the use of open data sources published on the London.gov.uk website to understand London as a city. The data that will be evaluated include the *London Borough Profiles* dataset and the *Statistical GIS Boundary Files* to visualise the data. The aims and objectives are detailed in table 1.

Aim	Analyse, visualise and critically assess open data sources using a combination of mathematical and visual data representation methods to understand the dynamics of Greater London.
Objective 1	Statistical evaluation of the data will be conducted and the relative merits discussed and critically reviewed in the programming language R.
Objective 2	Data and analysis of the data will be visualised regarding its spatial element.
Objective 3	A comparison between the purely mathematical statistical analysis and the visualisation of the data will be made and the merits and limitations critically assessed.

Table 1 - Table detailing the aims and objectives of the following report.

Thorough data analysis will be performed using the statistical methods: Pearson, Spearman and Kendall correlations, regression analysis and visual analysis based on patterns and correlations which are made apparent when data is graphically plotted.

2 – Data sources

The data which will be analysed in the following report has been published by the Greater London Authority and offers *"...demographic, economic, social and environmental datasets for*

each borough" (London Datastore, 2015). The two files that will be downloaded are the *London Borough Profiles* dataset in the form of a Comma Separated Value (.csv) file type; and the *Statistical GIS Boundary* Files ESRI shapefile (.shp). As an open data source, the data can be freely analysed and re-used subject to the terms of the Open Government Licence. The origin of the data can be traced back through Census data, Ordnance survey data and Office of National Statistics data with temporal coverage of between 06/03/2011 to 31/03/2016.

3 – Methodology

The purpose of this report is to identify what characteristics the data can tell the reader about London. Therefore, two '*for loops*' were employed to run correlation analysis of each dataset with each other dataset available and then return only those of extremely high correlation (>0.95) or low correlation (<-0.95) in the programming language R. The code implemented can be viewed in figure 1. The values of >0.95 and <-0.95 were selected in order to produce a reasonable number of results which can be reviewed in this report. Between 5 and 10 unique pairs of datasets were chosen to ensure accuracy of correlation and sufficient variety of results for analysis.

```
#two for loops are executed to perform pearson correlation analysis on each dataset and only return those that are highly correlated (>0.95)
for(i in 1:ncol(borough_profiles_edit_loop)) {
  for(j in 1:ncol(borough_profiles_edit_loop)) {
    pearson <- cor(as.numeric(borough_profiles_edit_loop[, i]), as.numeric(borough_profiles_edit_loop[, j]), method = "pearson", use = "complete.obs")
    if (pearson > 0.95) {
      output <- c(i, j, pearson)
      print(output)
    }
  }
}

#two for loops are executed to perform pearson correlation analysis on each dataset and only return those that are highly correlated (<-0.95)
for(i in 1:ncol(borough_profiles_edit_loop)) {
  for(j in 1:ncol(borough_profiles_edit_loop)) {
    pearson <- cor(as.numeric(borough_profiles_edit_loop[, i]), as.numeric(borough_profiles_edit_loop[, j]), method = "pearson", use = "complete.obs")
    if (pearson < -0.90) {
      output <- c(i, j, pearson)
      print(output)
    }
  }
}
```

Figure 1 - For loops used to calculate datasets that were most highly positively or negatively correlated

After the results were produced, the identified columns were used to perform correlation analysis once again in order to confirm accuracy of the findings. Following this step, the same '*for loops*' were modified to calculate the "Spearman" and "Kendall" correlation values and return the most extreme examples to the researcher. The results of these can be viewed in Table 1, section 4.

Once the highly correlated datasets had been identified, the data was visualised graphically and spatially in order to fully investigate to what extent the two variables are linked and if the link is merely coincidental. R packages rgdal, dplyr, tmap, ggplot2, ggmap were installed to allow the data to be visualised both mathematically (using x, y graphs) and spatially (using a map of London).

Mathematically, the data identified as most strongly correlated was illustrated using the *ggplot* command which plots all the points of the two datasets on an XY graph with the regression analysis superimposed above for comparison. Then, the attributes will be modified to vary the point size by population, for each borough in question, in order to quantify the relevance of the results by the number of people that are affected.

Spatially, the data will be displayed using the tmap and ggplot2 packages and mapped within the borough boundaries using the *Statistical GIS Boundaries London* shapefiles published by the Greater London Authority (London Datastore, 2015).

The limitations of such a methodology include the reality that the report merely identifies correlations between the data. Therefore, further study would be required to confirm if the correlation was the result of a connection between the two variables, or occurs as merely coincidence. The highest reasonable correlation coefficient was selected in order to minimise the probability of a coincidental correlation. One method to minimise this likelihood further is to use more datasets, more accurate data collection methods, datasets with greater temporal resolution or comparisons with other data collection methods.

4 – Results

Type of Correlation	Highly Positively Correlated Data Column Numbers (>0.95)	Highly Negatively Correlated Data Column Numbers (<-0.95)
Pearson	1, 2	10, 63
Pearson	4, 5	70, 79
Pearson	8, 11	
Pearson	23, 68	
Pearson	39, 40	
Pearson	46, 50	
Spearman	8, 11	63, 65
Spearman	23, 68	70, 79
Spearman	39, 40	
Kendall	None identified	None identified

Table 2 - Results of the initial correlation analysis, identifying datasets with extremely high positive or negative correlation (please note that a correlation of 1 was returned for each dataset when compared to itself. These values have been removed from the results).

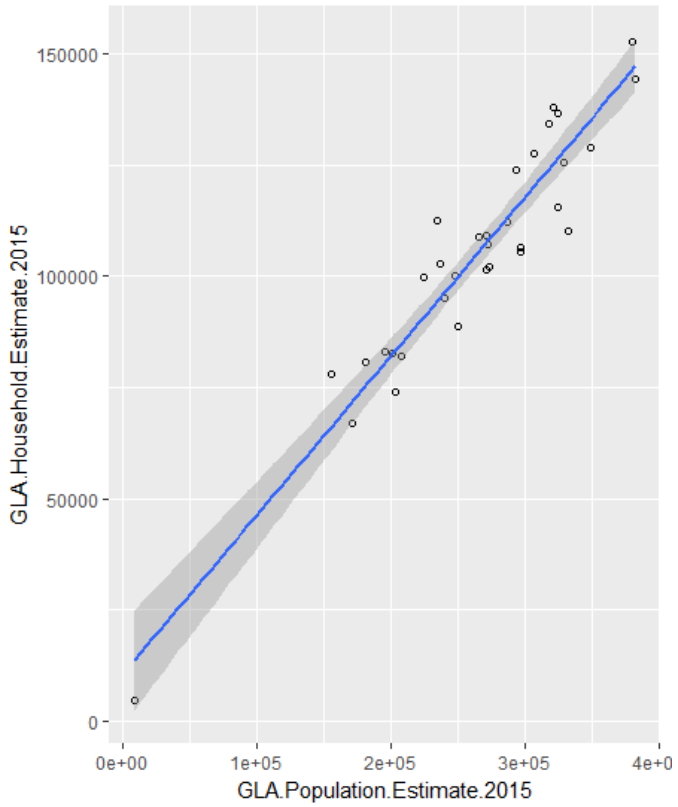
Table 2 displays the results of the correlation analysis run on each dataset, calculated in comparison to each other dataset, and only pairs of datasets scoring an extremely high or low correlation coefficient value are returned. The number of unique pairs of datasets identified was 8, which fell between the boundaries of 5 and 10 unique dataset pairings set in the methodology (section 3).

Puth et al. (2015) critically review the differences between Pearson, Spearman and Kendall correlation methods, and conclude that the most accurate is the Kendall procedure. This conclusion is supported in the results in Table 1, as the Kendall correlation analysis identified no highly correlated pairs of results. Thus, it is reasonable to deduce that none of the datasets are directly comparable, but have general links and further contributing factors.

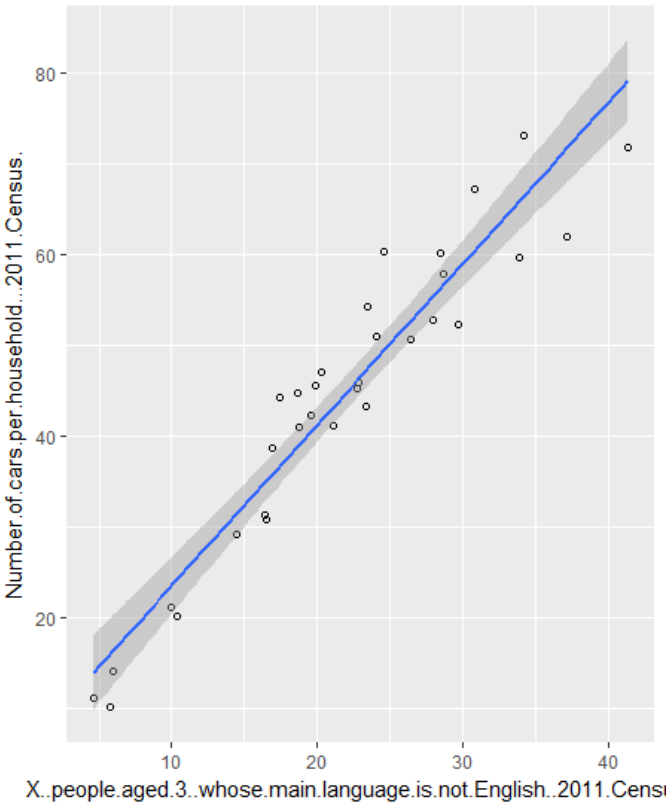
The outputs from the data analysis were 3 different representation methods (2 mathematical and 1 spatial) of 8 unique pairs of highly positively or negatively correlated datasets. Identification of the datasets, analysis and visualisation of the datasets were all completed in RStudio in the programming language R, and exported into a portable network graphic file type (png) for inclusion in this report. The two mathematical representation methods were:

1) linear regression model superimposed over a scatter graph, and
2) scatter graph with point colour representing the population of the data.

The spatial representation of the data utilised the qtm command from the tmap package to vary the colour of each borough of London based on the data analysed. All outputs from the study are included in the appendix (fig.1 – fig.33), however figures which were deemed worthy of note by the researcher are analysed in section 5.

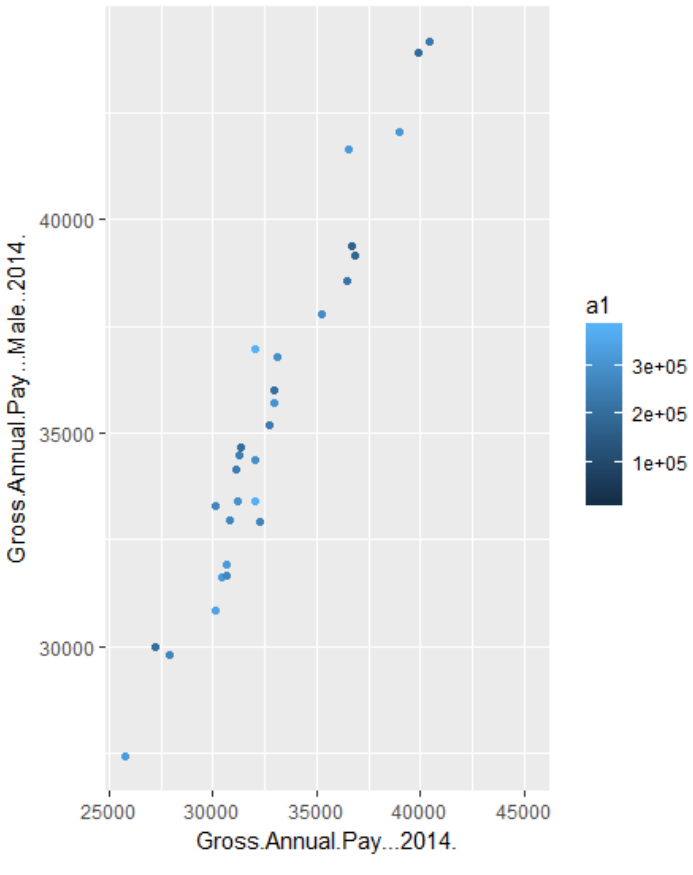


(a)

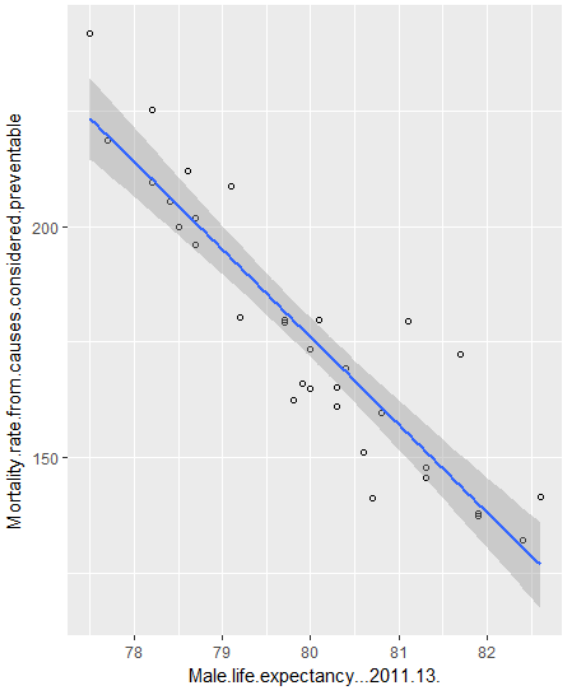


(b)

Figure 2 - Mathematical visualisation of open source 'Population data' and 'Household data' (a) and 'number of cars per household' and 'number of people aged 3 whose main language is not English' (b) published by the Greater London Authority in linear regression model in R.



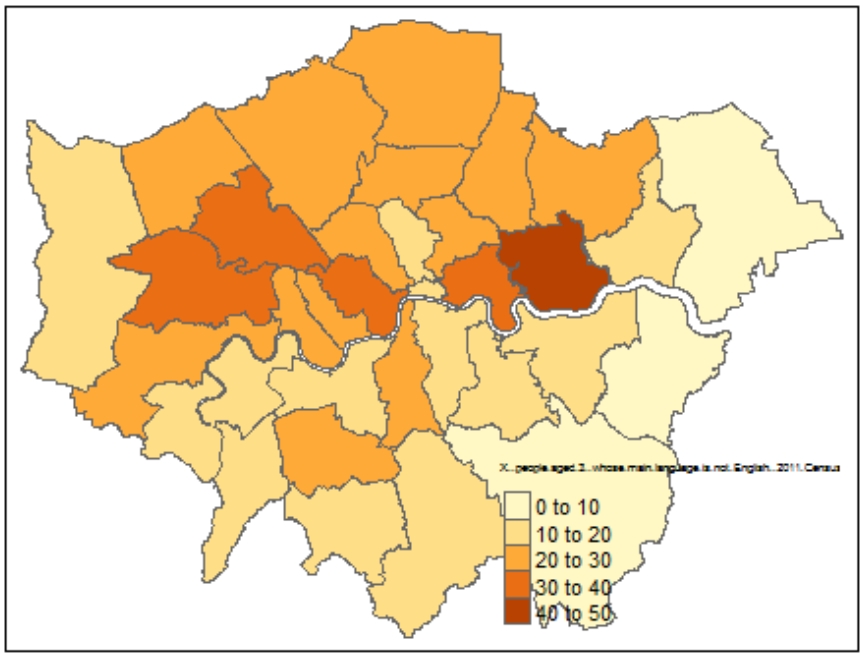
(a)



(b)

Figure 3 - Mathematical visualisation of 'Gross Annual pay' and 'Male gross annual pay' (a); and 'Male life expectancy and Mortality rate from causes considered preventable' (b). A linear regression model with the colour of the point representing the population of the borough.

(a)



(b)

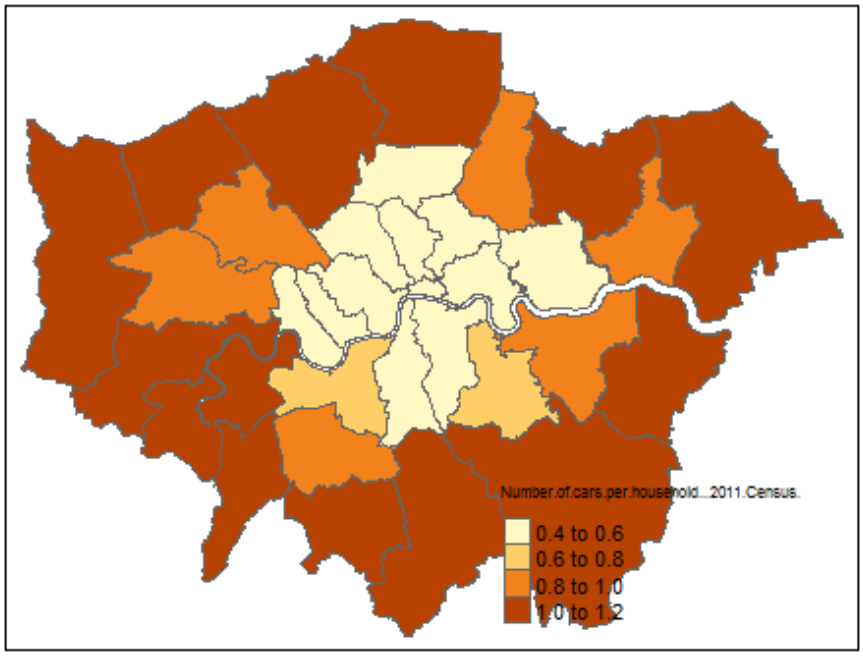


Figure 4 - Spatial visualisation of 'number of people aged 3 those first language is not English' (a) and 'number of cars per household' (b).

5 – Discussion

The purpose of this report is to understand London using open data sources. Figure 2a is a graphical representation of open source data detailing population and number of households in Greater London by borough. It is the first pair of datasets identified by both the Pearson and Spearman correlation methods as 'highly correlated' and therefore implies that generally, in

Greater London, the more households there are in a borough, the greater the population of the borough.

Figure 2b is also a graphical representation of a dataset identified as strongly, positively correlated. The two datasets in question are the *number of people aged 3 whose first language is not English* and the *number of cars per household*. Therefore, the data implies that for each borough of London, the more likely a child's first language is not English, the more likely the household is to have a greater number of cars. A possible reason for this is that when an immigrant couple whose first language is not English choose to start a family, they also decide to move to the outskirts of the city for more space and therefore require a car. It is also worth noting that a correlation could be between the number of cars per household and number of children per household, regardless of the child's language.

Figure 3a is slightly different, it shares in common with figures 2a and 2b the fact that it visualises two datasets identified as highly correlated; however, each point of the scatter graph is a different colour, based on the population of the borough represented. The purpose of this is to indicate whether the gross annual pay statistic varies over borough population figures. This is relevant in figure 3a and not figures 2a and 2b as the linear regression model was necessary to detail the y-intercept.

The most interesting aspect of figure 3a is not the data which was identified, but the data which was not identified. The two datasets which were calculated as 'highly correlated' were *Gross Annual Pay* and *Male Gross Annual Pay*; however, the following dataset in the series was *Female Gross Annual Pay* and this was not highlighted as highly correlated. Therefore, the open data published by the Greater London Authority implies that in a borough where the average gross annual pay is higher, it is likely that the average male gross annual pay is higher, however, average female gross annual pay is not likely to be higher. Furthermore, male pay is likely to be proportional to the wealth of the area they live, whereas female pay is not necessarily likely to be.

Figure 3b is the first figure detailing data which was identified as 'highly negatively correlated' or inversely proportional. Thus, the more likely a male resident of an area is to live longer, the less likely residents of the same area are to die of a preventable disease. Although, this correlation could be considered common sense, similarly to figure 3a, the most interesting aspect is that the following dataset *Female life expectancy* is not highly inversely correlated. It could be considered that the reason for this is that female life expectancy is higher than their male counterparts, but it could be higher and also proportional (but it is not). Therefore, it could be because women are equally likely to die from a preventable disease, but the majority who don't, generally live longer than men and therefore the data is less correlated. Further study is required to understand if there is a connection between these three variables or if the correlation is coincidental.

Figure 4a is the first spatially represented dataset graphic. The data used is the same as Figure 2b; the reason it was selected for analysis in the report is that despite being selected as a 'highly correlated' pair of datasets, when spatially represented, there appears to be no spatial correlation at all. One cause of this could be the method of categorising the data into colours distorts the data in a way which is not a fair representation of the data.

6 – Conclusion and future work?

In conclusion, open source data is an effective and informative way to offer information about the complex dynamics of a city. The methodology employed used mathematical correlation statistics in order to identify connections and similarities between the data in order to give an indication of some less tangible aspects of the city such as the differences between male life expectancy and number of preventable deaths in an area.

The limitations of the results include the resolution of the data. As there are only 33 different boroughs in London, for each variable, there are only 33 values for comparison. Few anomalous values are more likely to decrease the accuracy of the results than a data source of hundreds of values. It is also important to consider if parts of the data is under-representative or over-representative of the study area and if the data collected was fit for purpose or incidental. This data is collected by the national census or by the office of National Statistics and therefore fit for purpose, up to date (2015) and sufficiently reliable for this study.

7 – References

© Crown copyright 2012

You may re-use this information (not including logos or Northern Ireland data) free of charge in any format or medium, under the terms of the Open Government Licence.

However, the following attribution statement must be acknowledged or displayed on any product using ONS data:

Contains Ordnance Survey data © Crown copyright and database right 2012

Contains National Statistics data © Crown copyright and database right 2012

Puth, M.-T., Neuhäuser, M., and Ruxton, G. D. (2015) 'Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits'. *Animal Behaviour* 102, 77–84

8 – Appendix

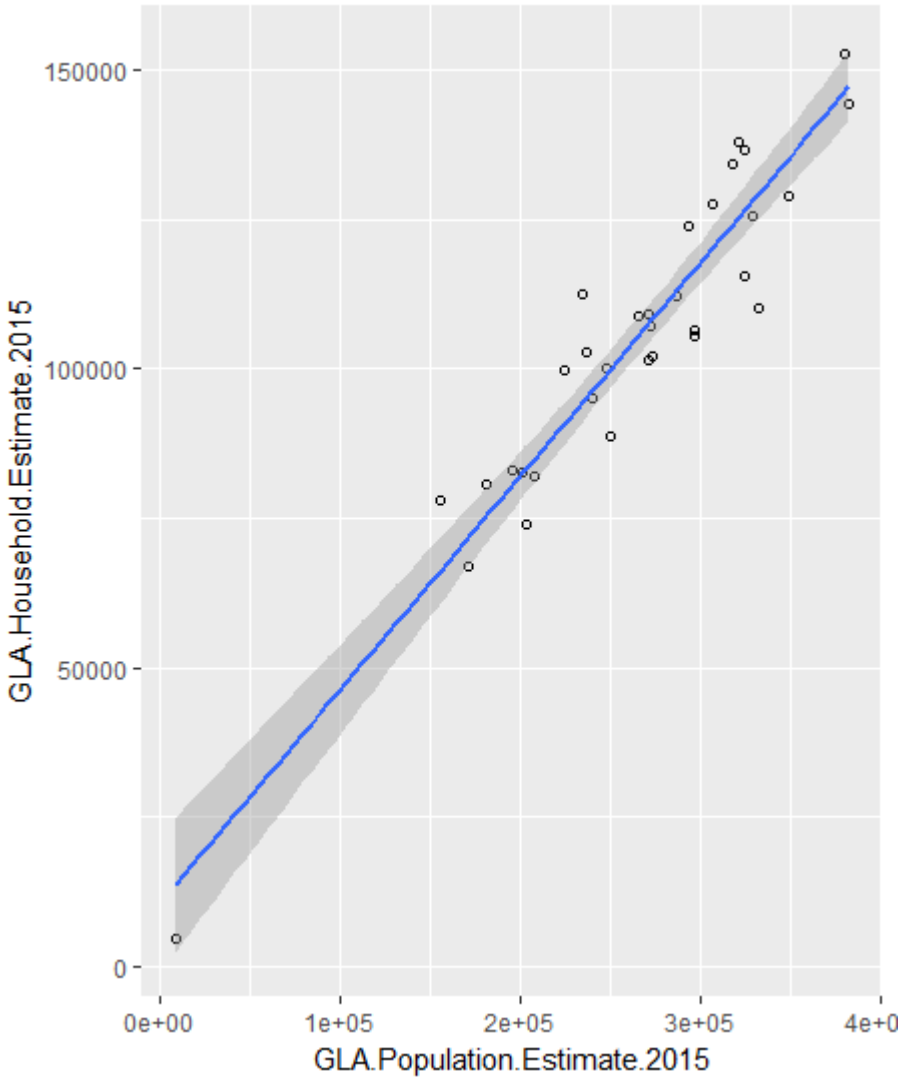


Fig. 1 - Regression Analysis of population estimate and household estimate datasets

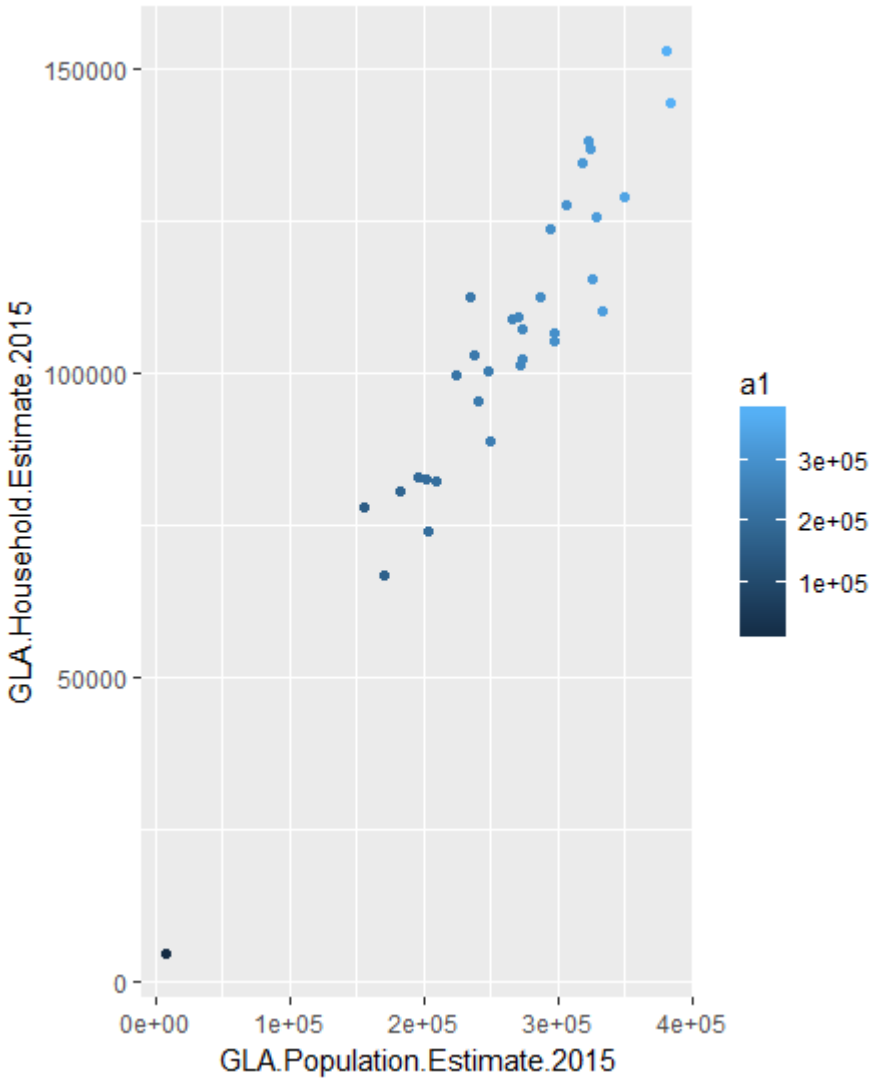


Fig. 2 - Regression Analysis of population estimate and household estimate datasets with colour proportional to population of the borough

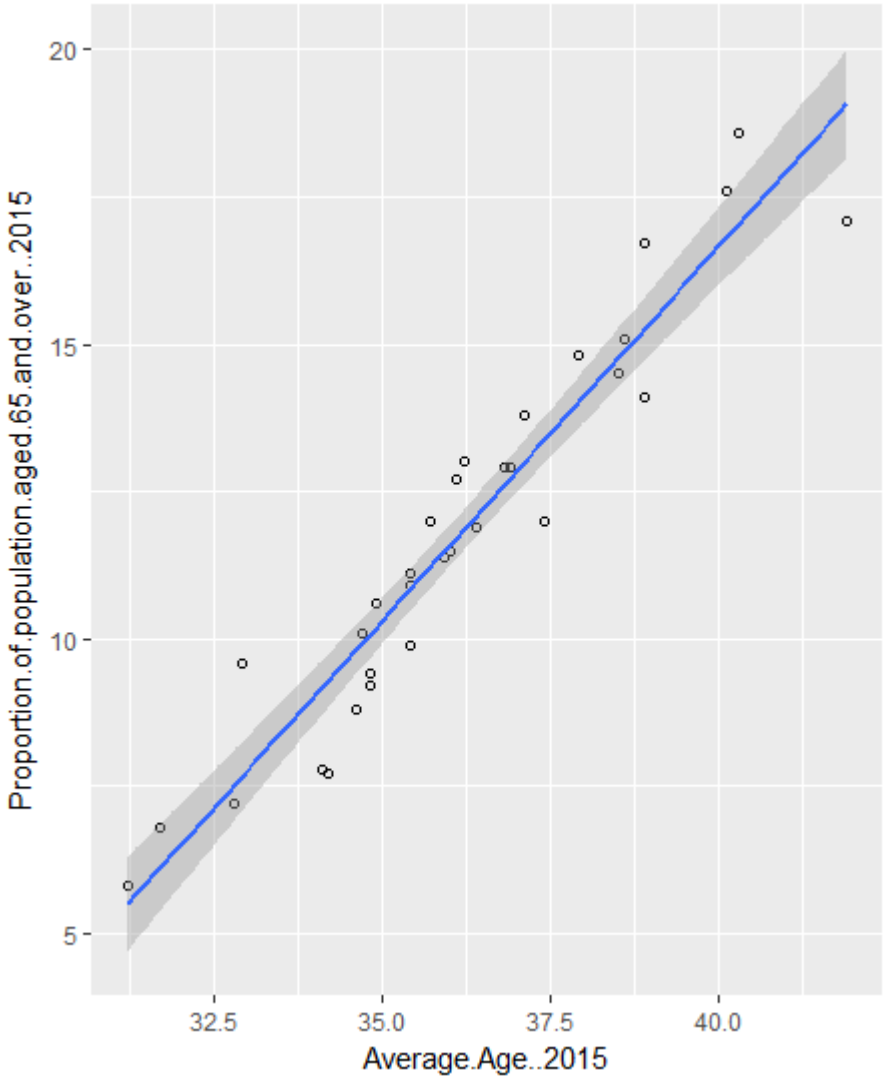


Fig. 3 - Regression Analysis of borough average age and proportion of population aged 65 and over datasets

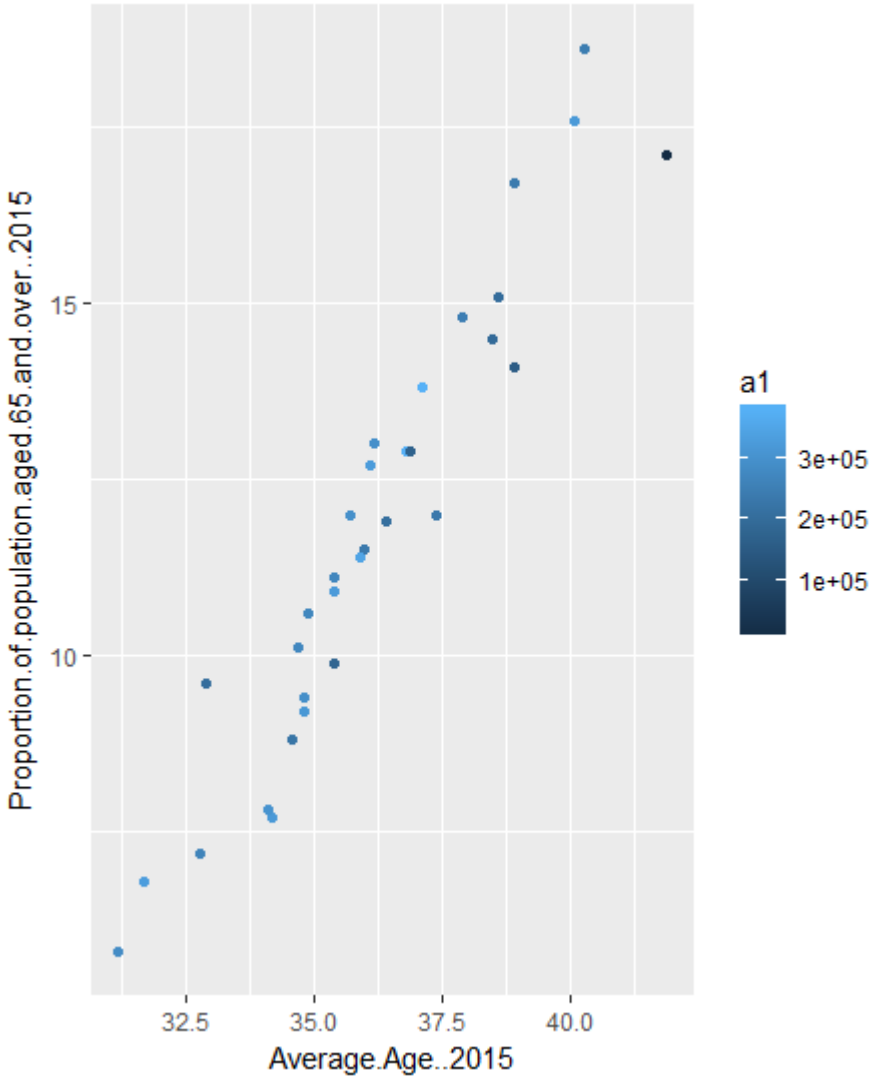


Fig. 4 - Regression Analysis of borough average age and proportion of population aged 65 and over datasets with colour proportional to population of the borough

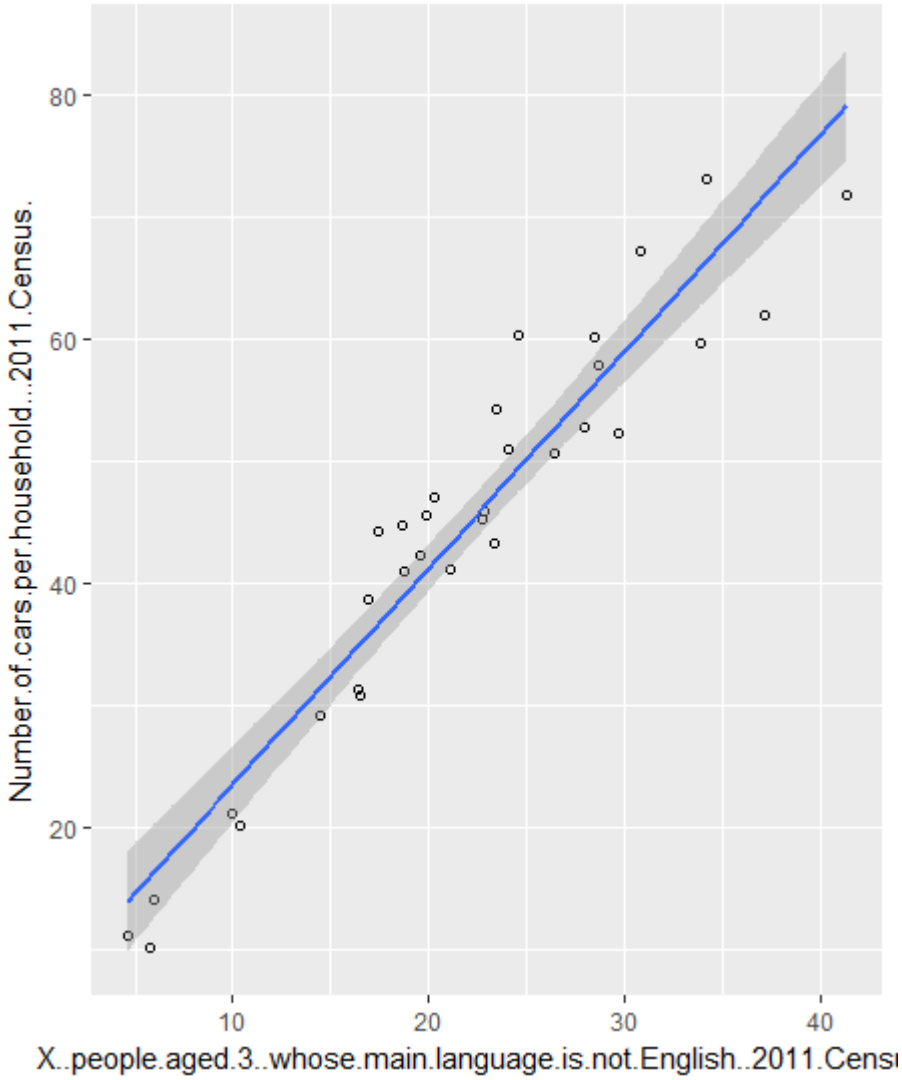


Fig. 5 - Regression Analysis of number of people whose main language is not English and number of cars per household datasets

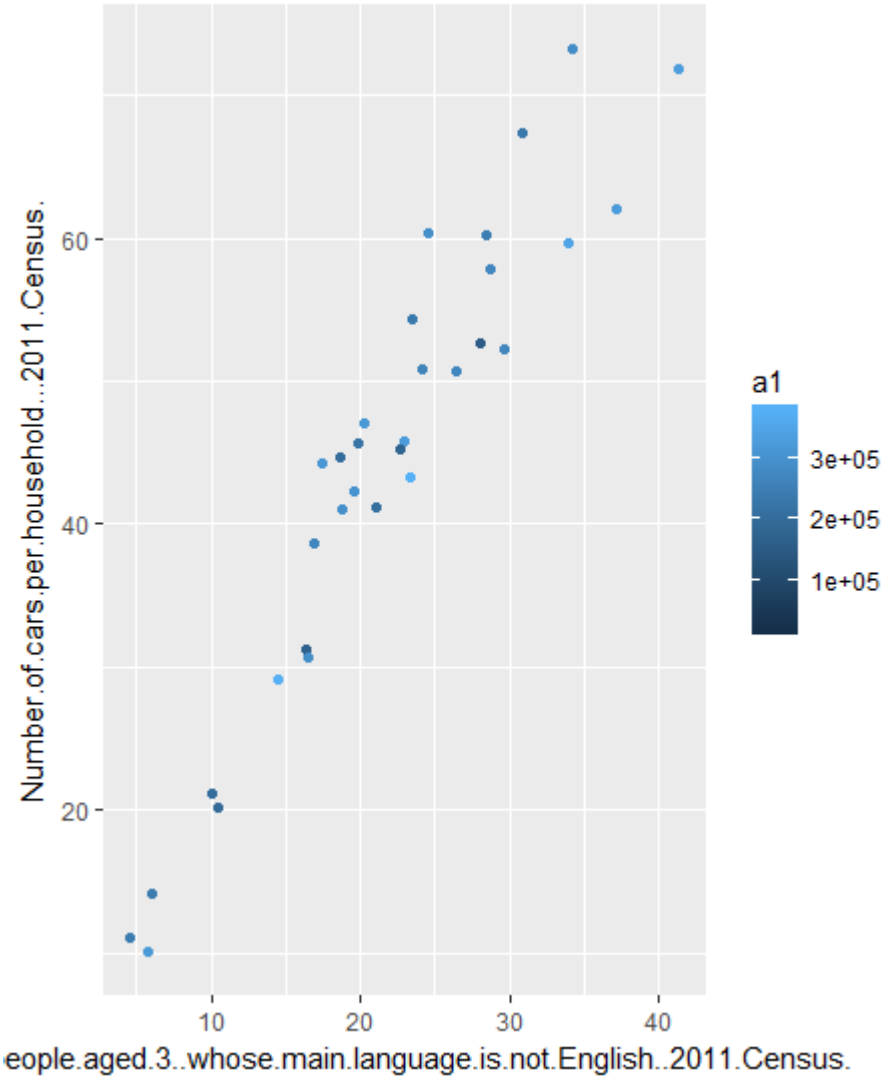


Fig. 6 - Regression Analysis of number of people whose main language is not English and number of cars per household datasets with colour proportional to population of the borough

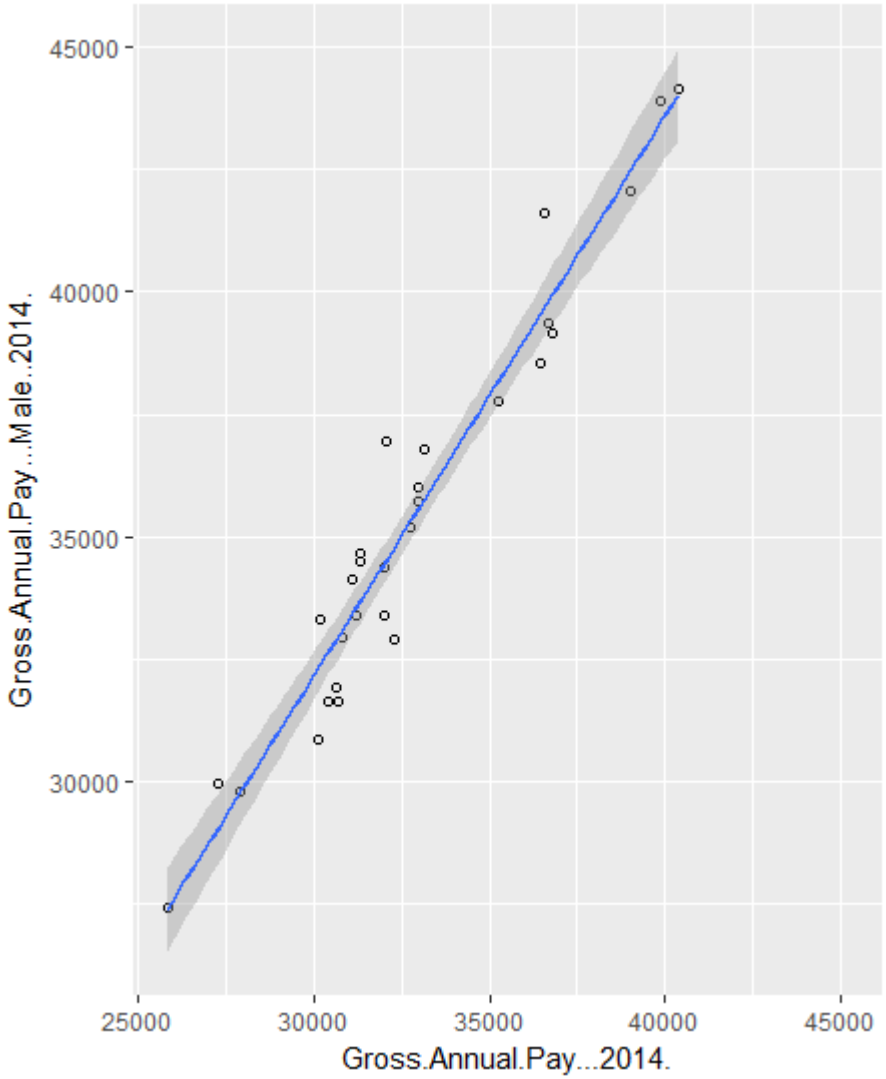


Fig. 7 - Regression Analysis of gross annual pay and male gross annual pay datasets with

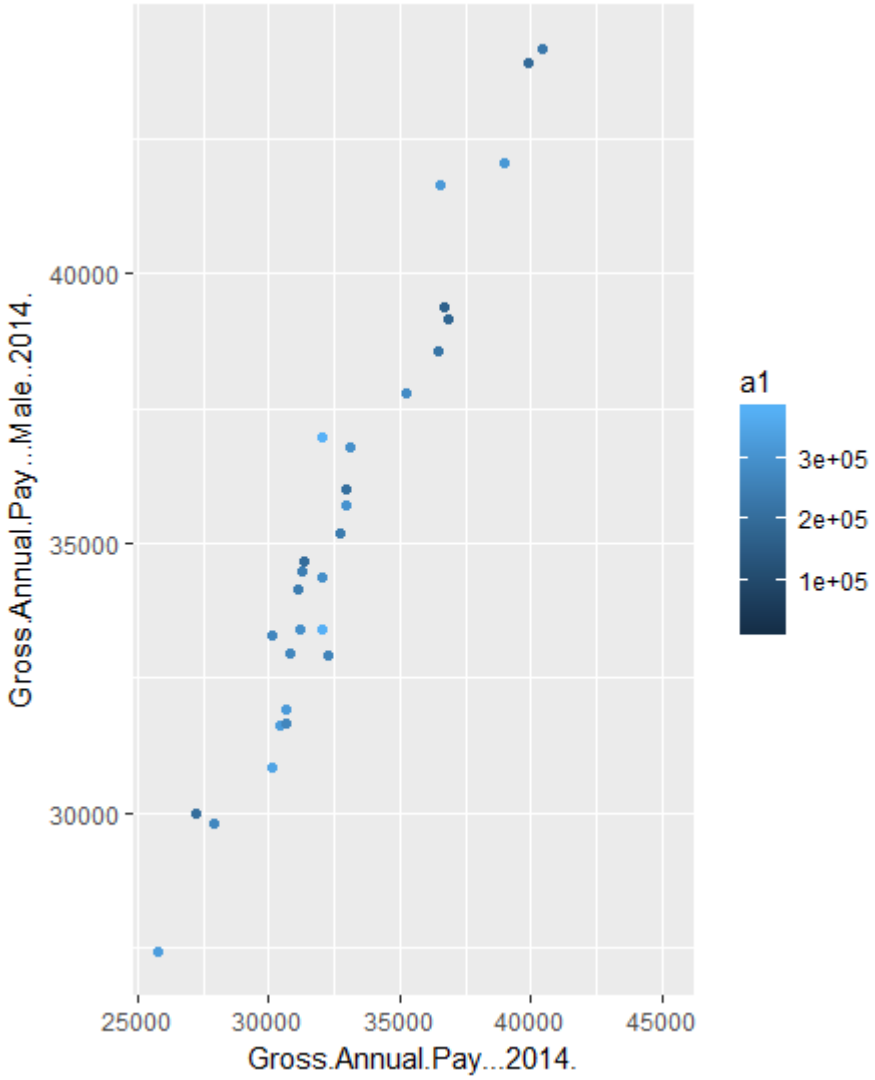


Fig. 8 - Regression Analysis of gross annual pay and male gross annual pay datasets with colour proportional to population of the borough

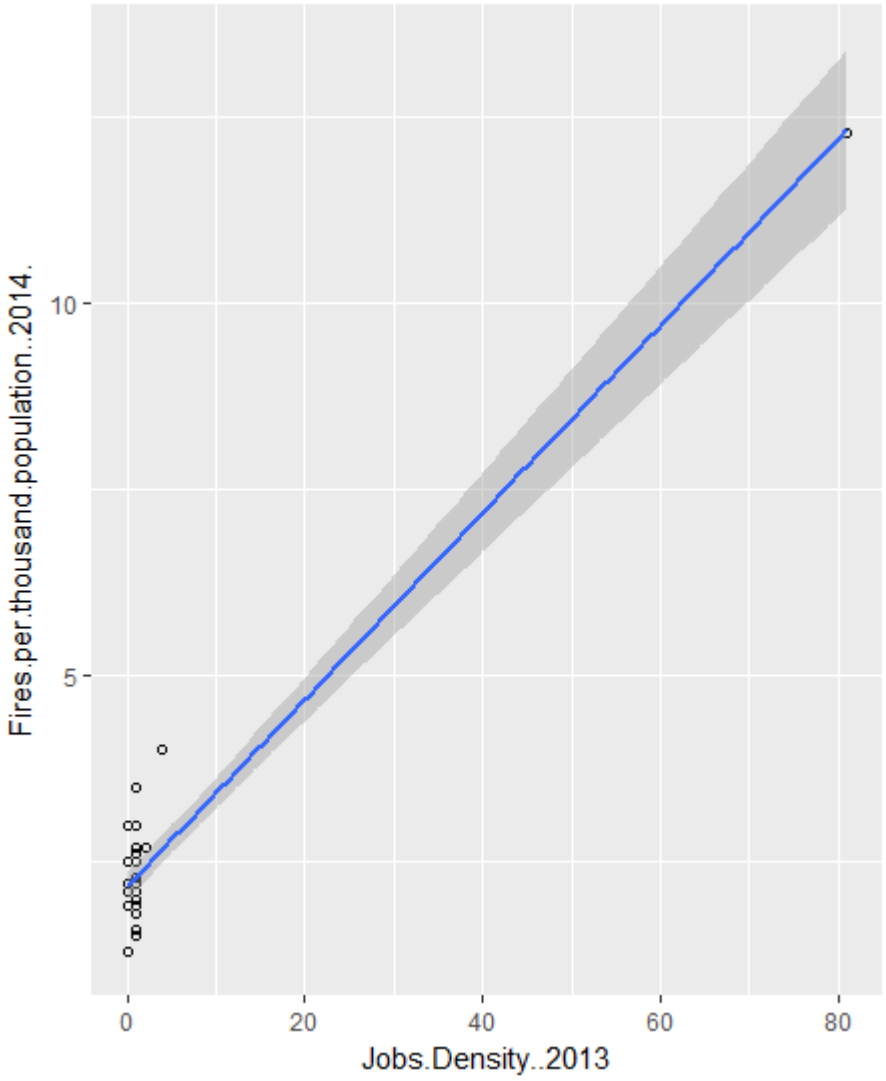


Fig. 9 - Regression Analysis of jobs density and fires per thousand population datasets

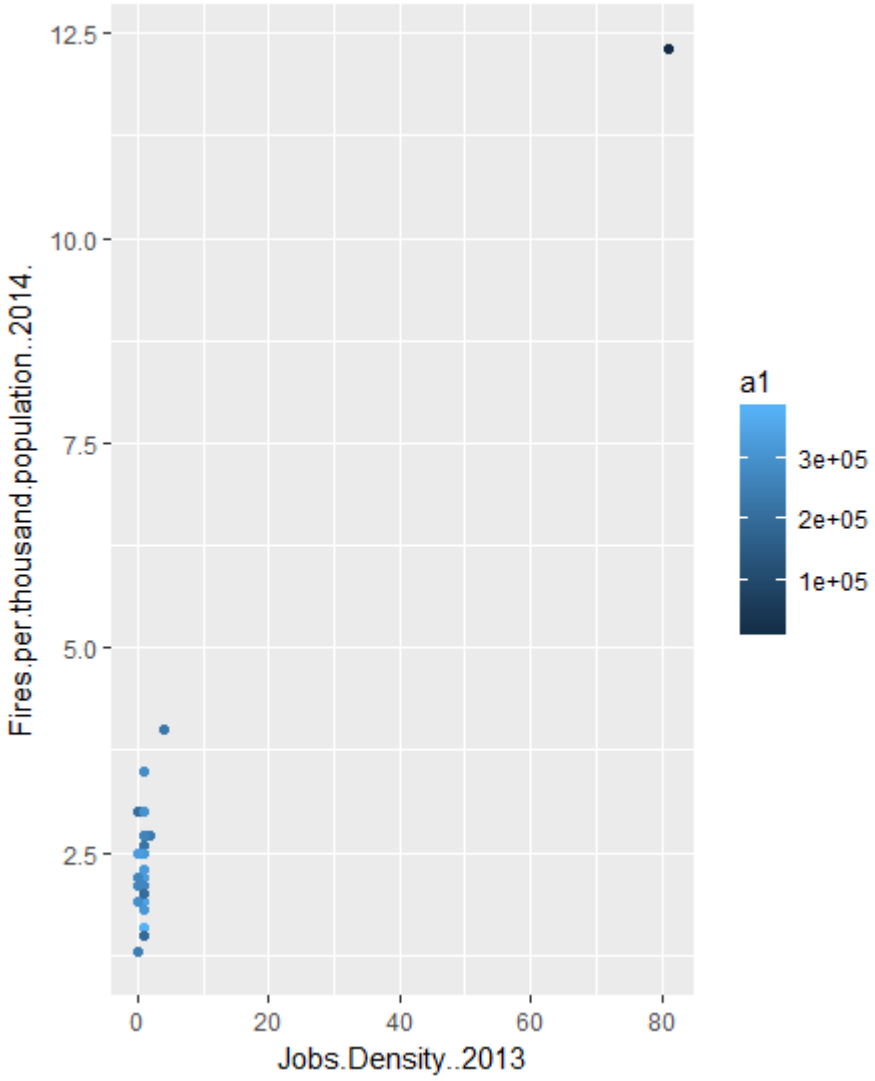


Fig. 10 - Regression Analysis of jobs density and fires per thousand population datasets with colour proportional to population of the borough

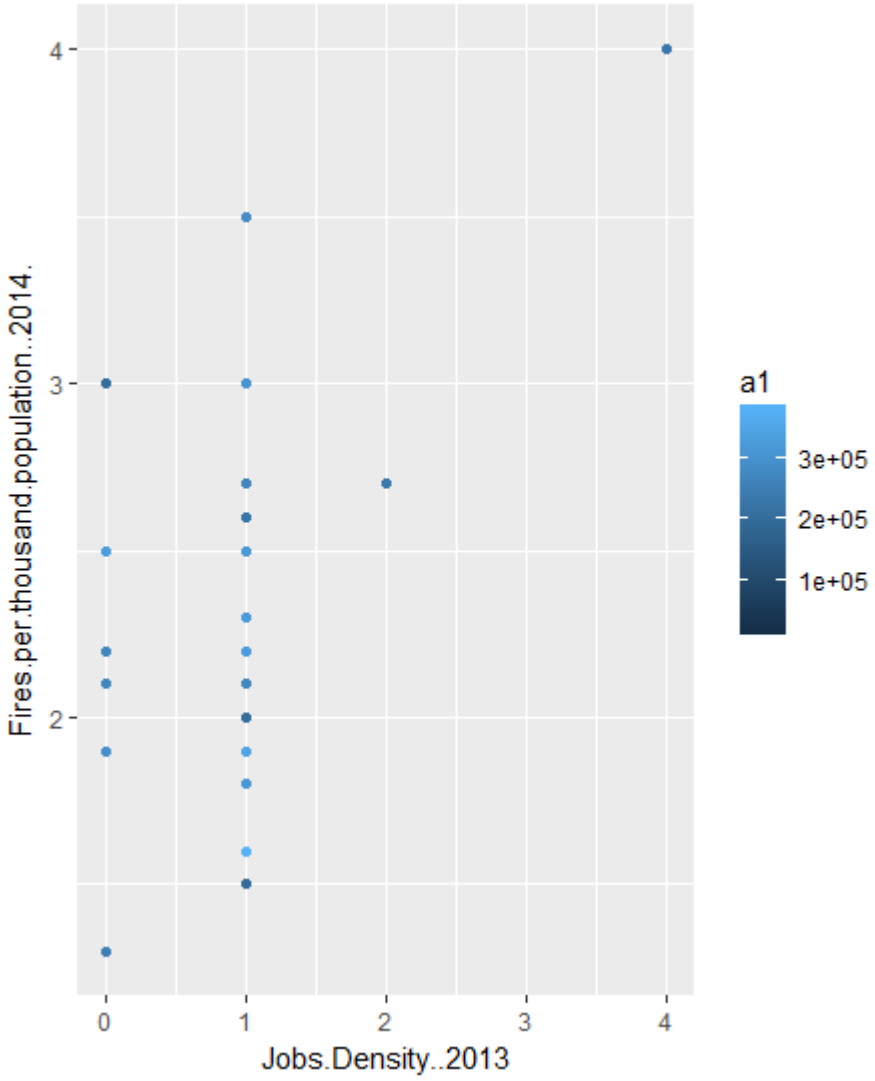


Fig. 11 - Regression Analysis of jobs density and fires per thousand population datasets with colour proportional to population of the borough (with anomalous value excluded)

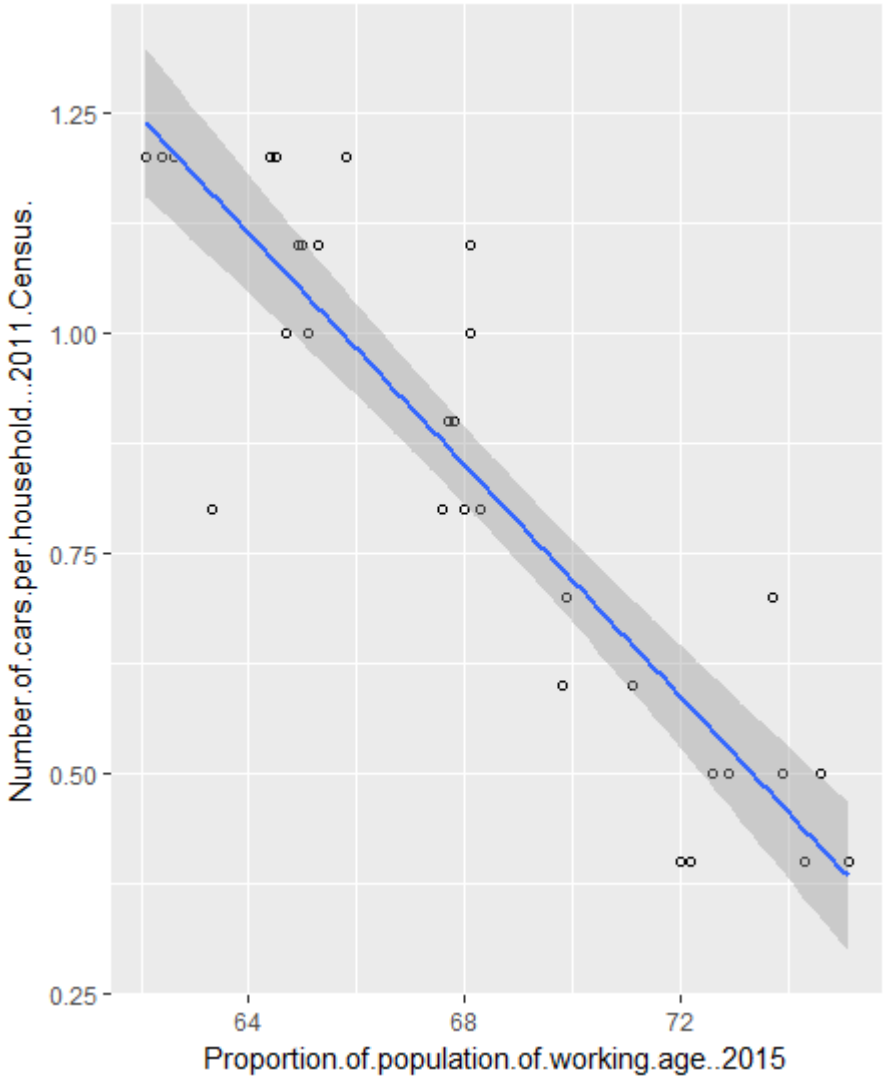


Fig. 12 - Regression Analysis of proportion of population of working age and number of cars per household datasets

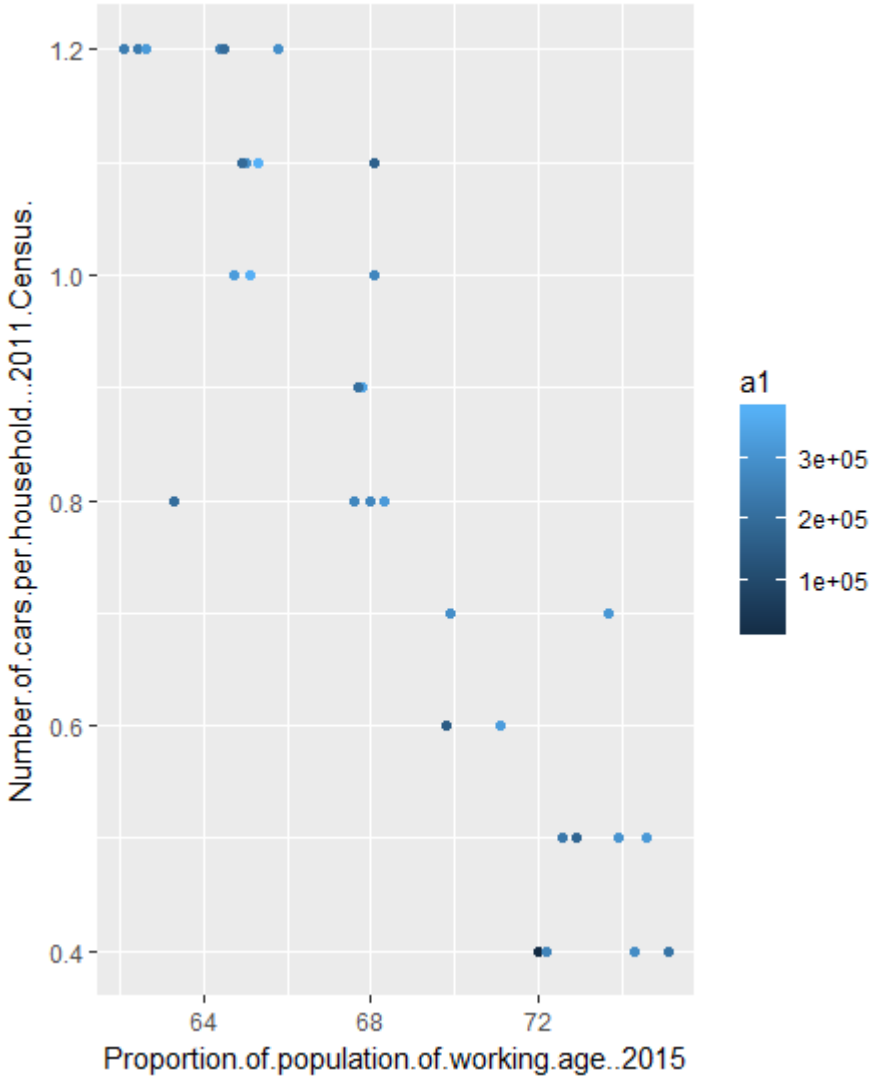


Fig. 13 - Regression Analysis of proportion of population of working age and number of cars per household datasets with colour proportional to population of the borough

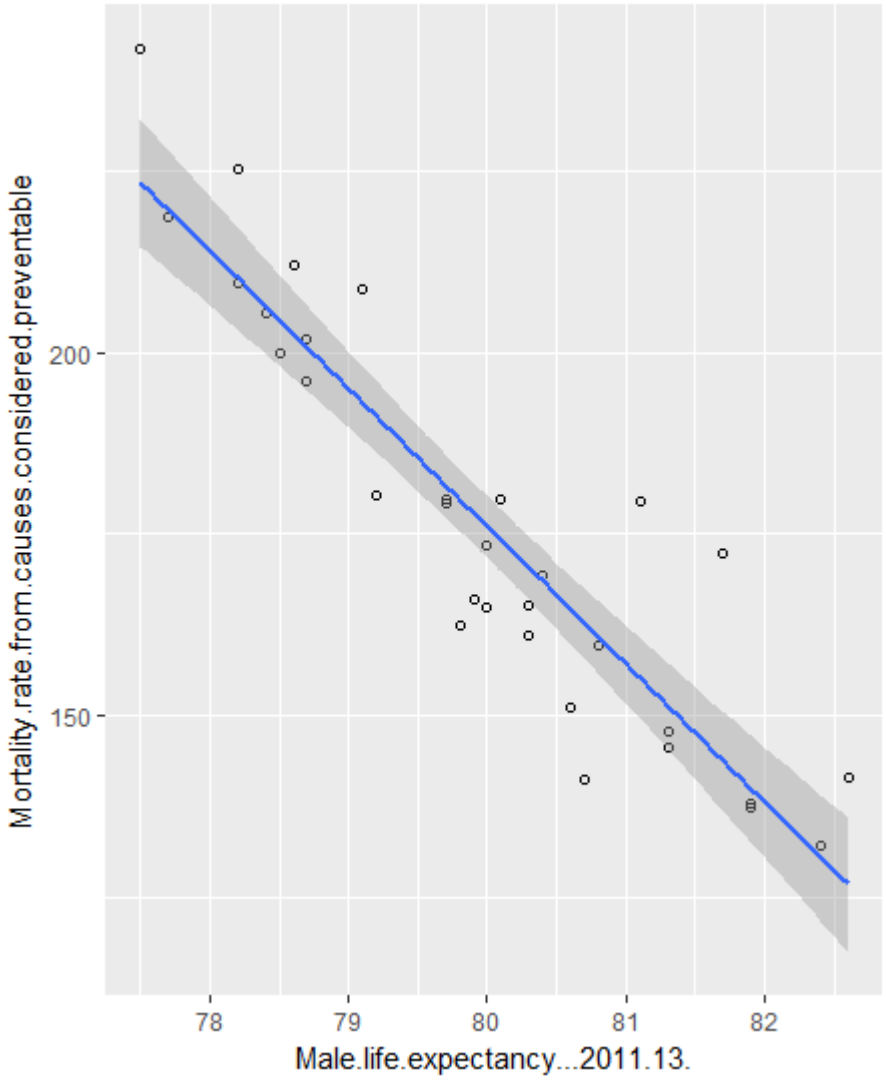


Fig. 14 - Regression Analysis of male life expectancy and mortality rate from causes considered preventable datasets

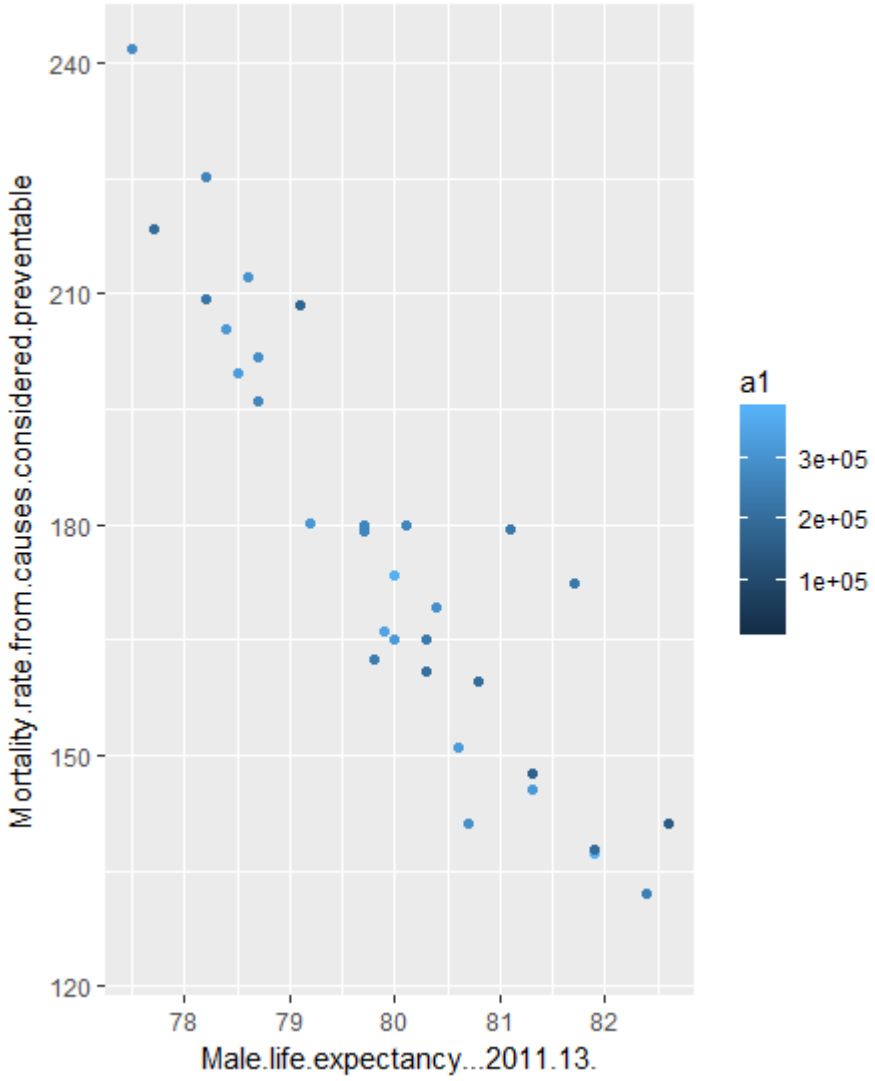


Fig. 15 - Regression Analysis of male life expectancy and mortality rate from causes considered preventable datasets with colour proportional to population of the borough

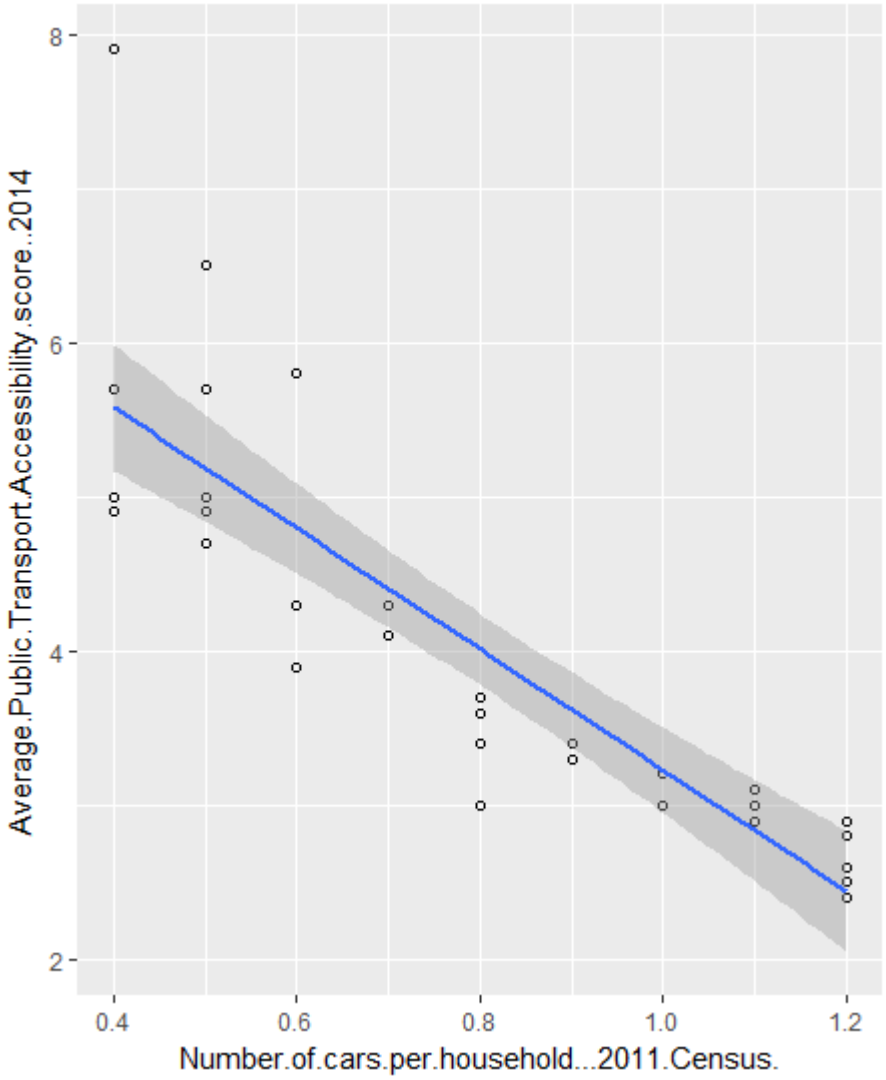


Fig. 16 - Regression Analysis of number of cars per household and average public transport accessibility score datasets

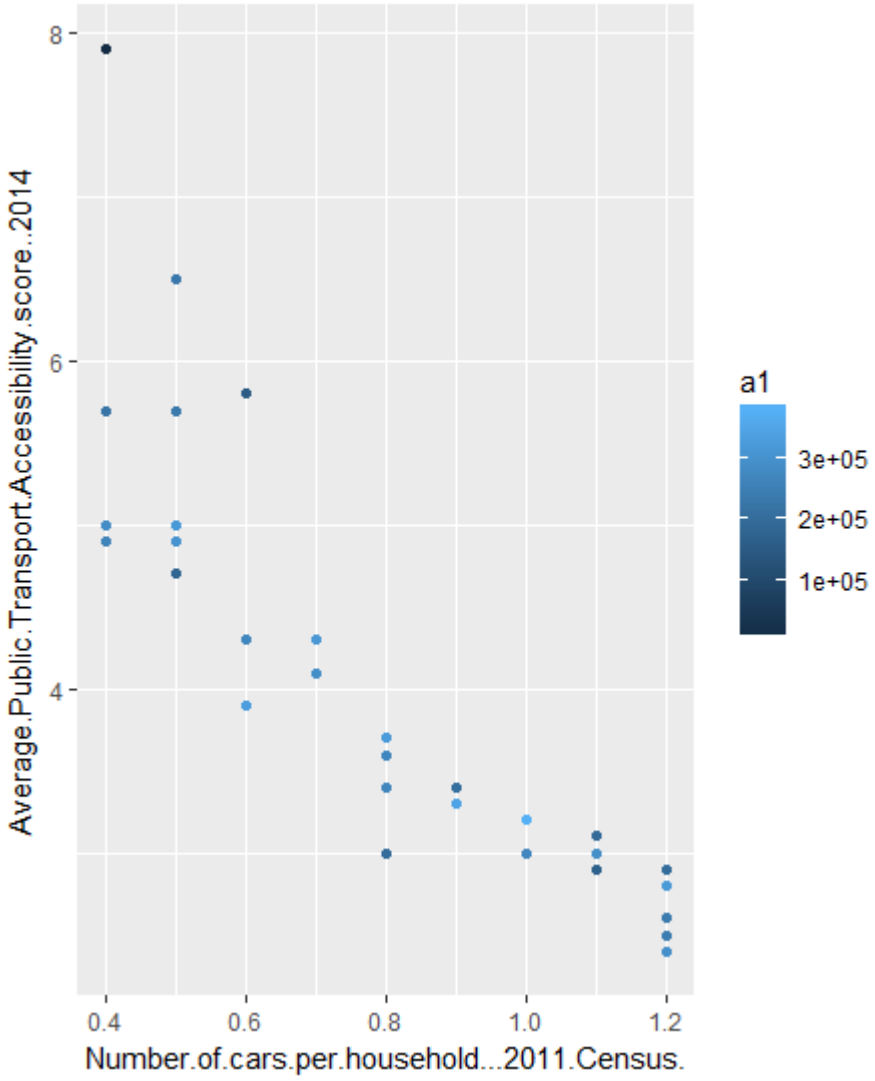


Fig. 17 - Regression Analysis of number of cars per household and average public transport accessibility score datasets with colour proportional to population of the borough

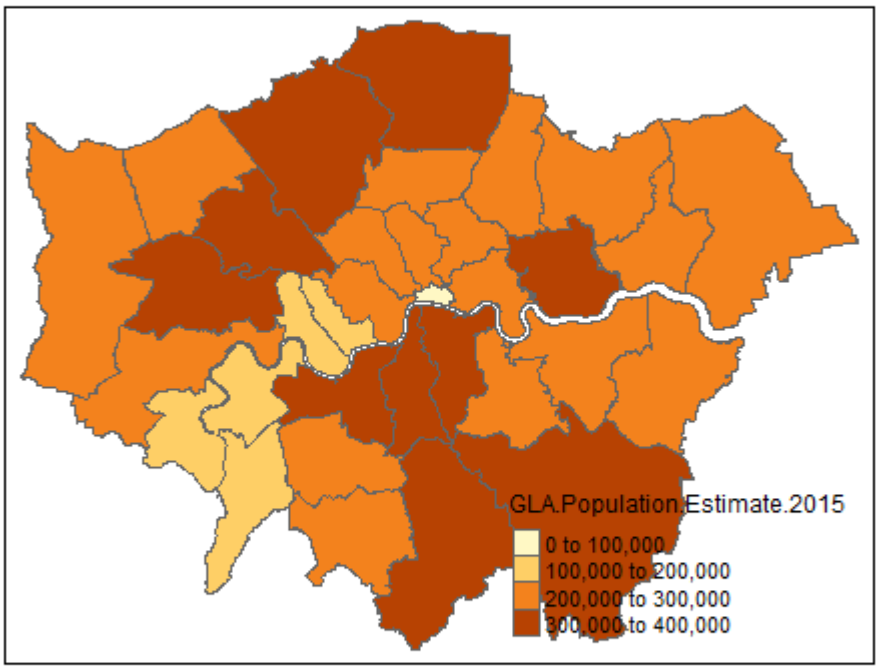


Fig. 18 - Spatial visualisation of Population Estimate data

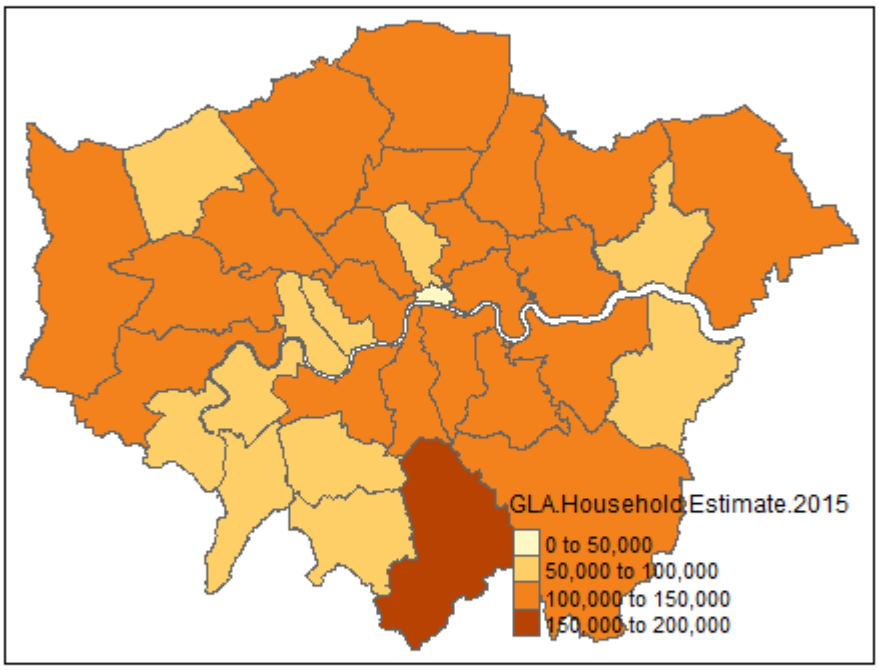


Fig. 19 - Spatial visualisation of Household Estimate data

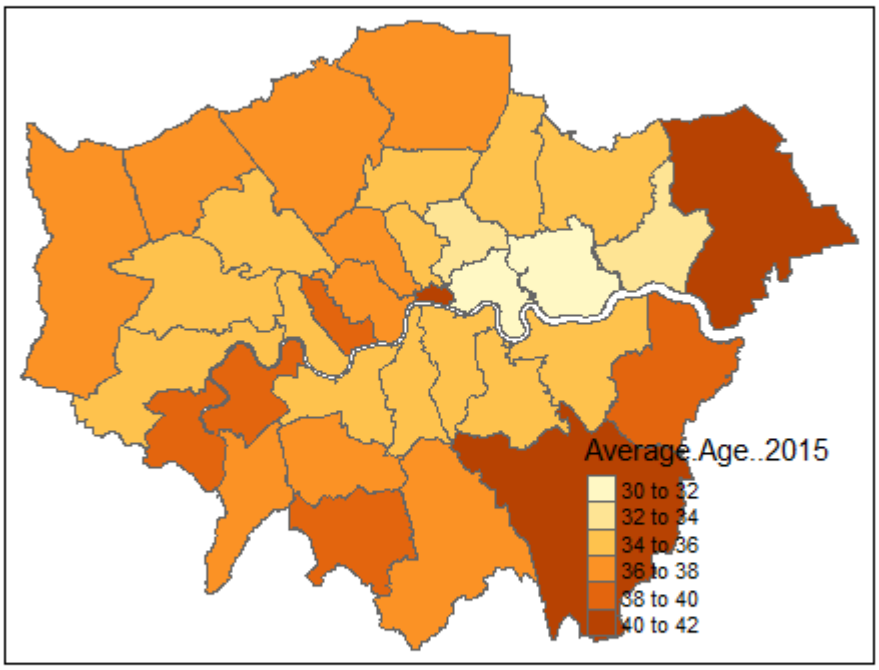


Fig. 20 - Spatial visualisation of Average Age data

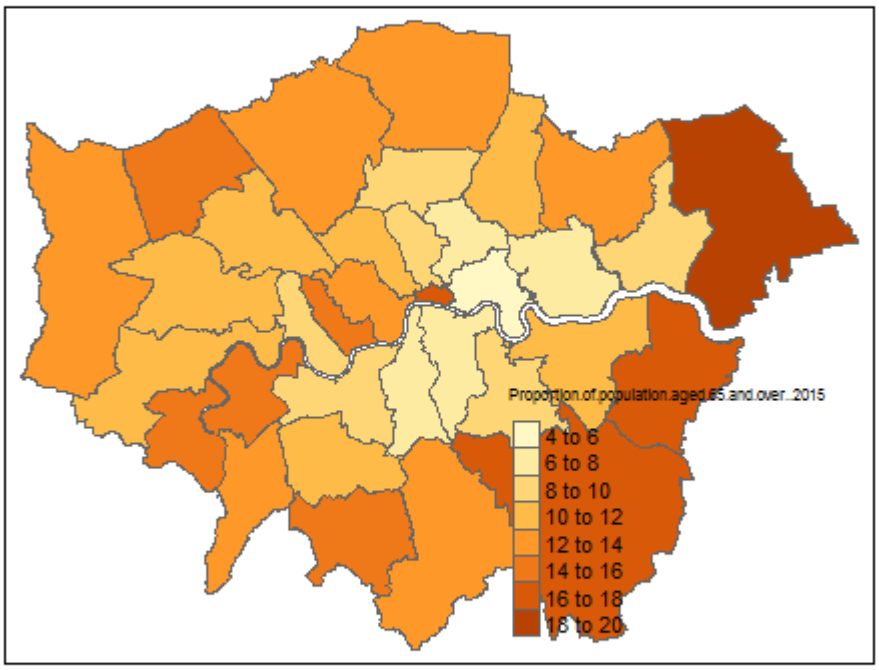


Fig. 21 - Spatial visualisation of Proportion of population aged 65 and over data

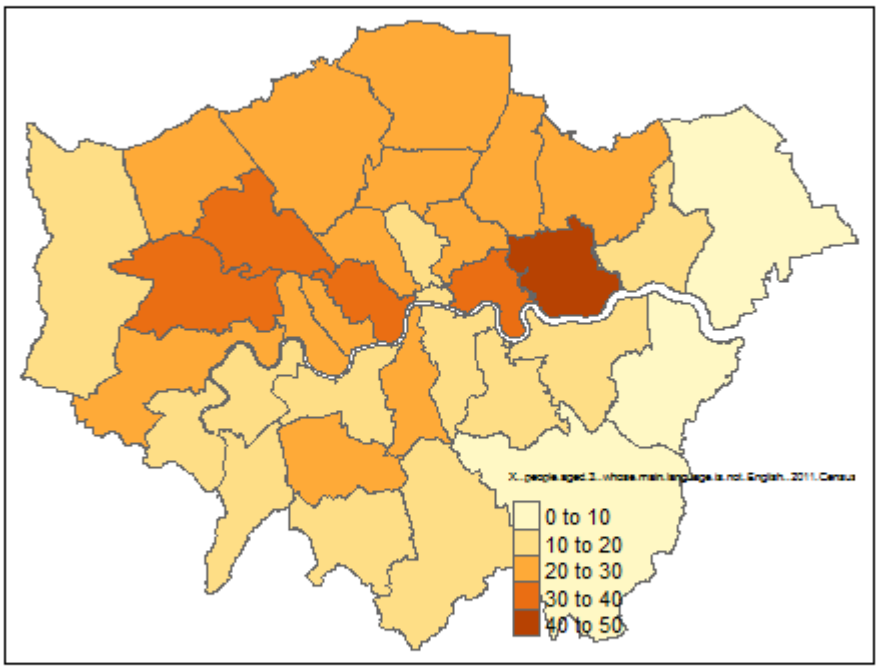


Fig. 22 - Spatial visualisation of Number of people aged three whose main language is not English data

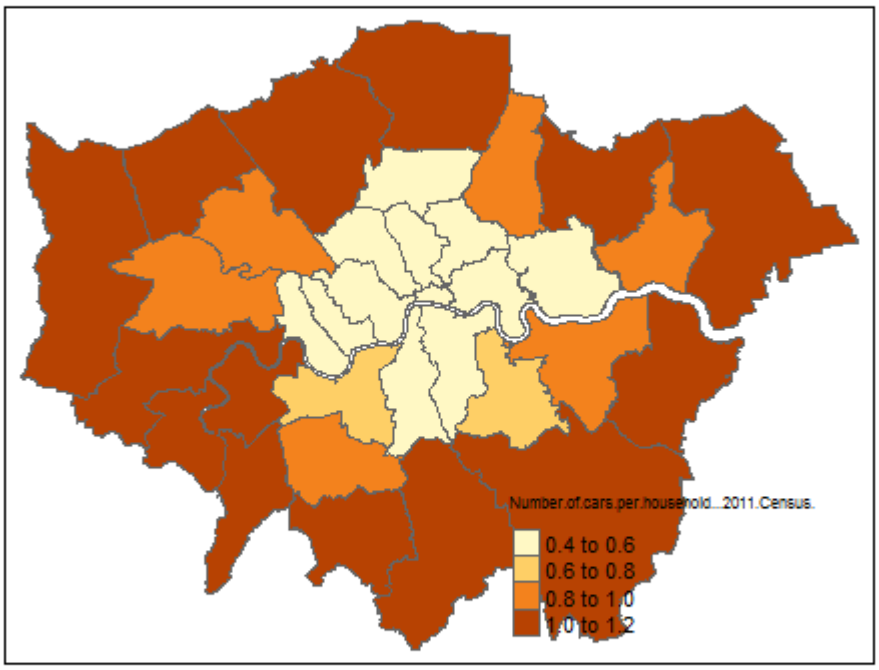


Fig. 23 - Spatial visualisation of Number of cars per household data

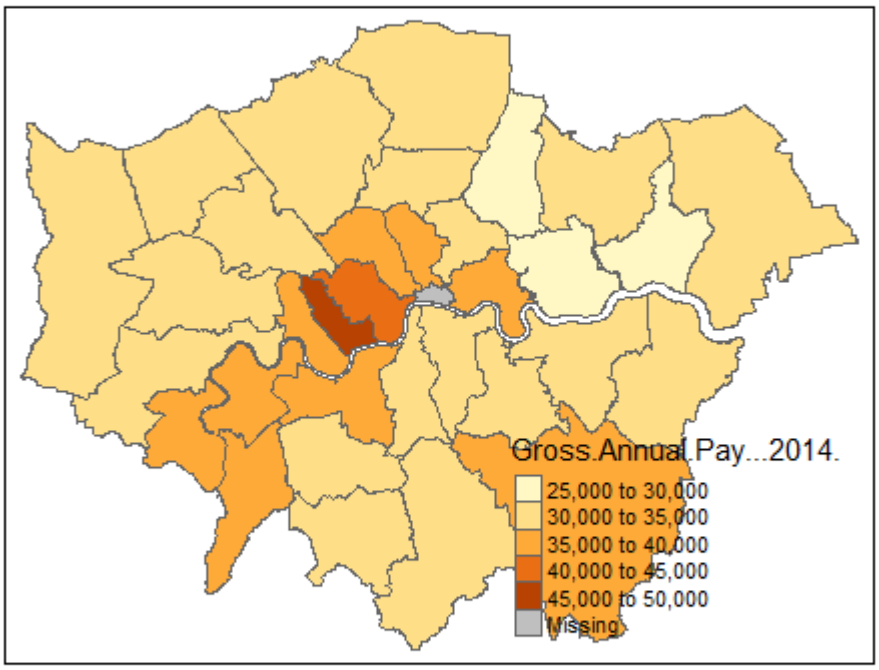


Fig. 24 - Spatial visualisation of Gross Annual Pay data

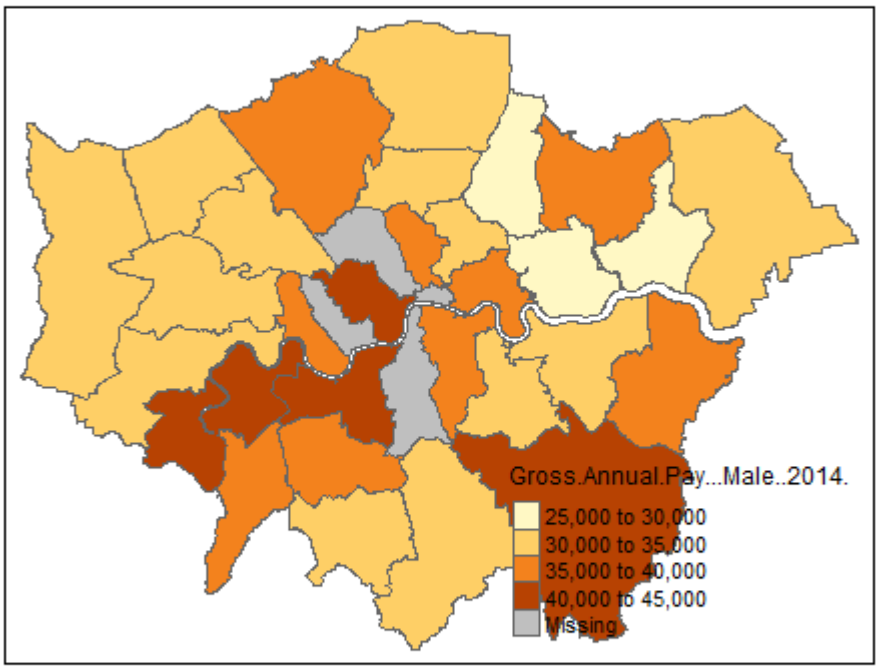


Fig. 25 - Spatial visualisation of Male Gross Annual Pay data

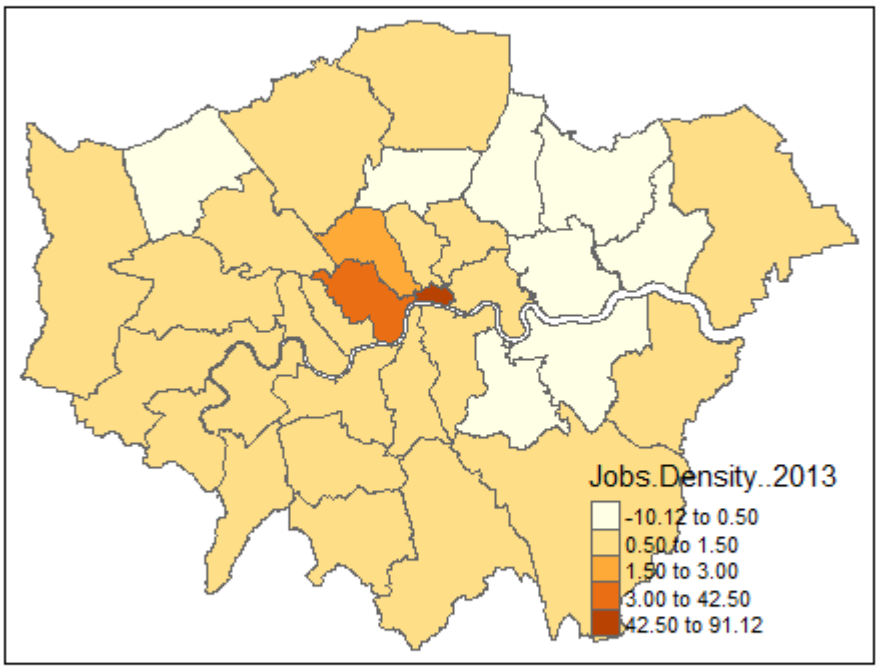


Fig. 26 - Spatial visualisation of Jobs Density data

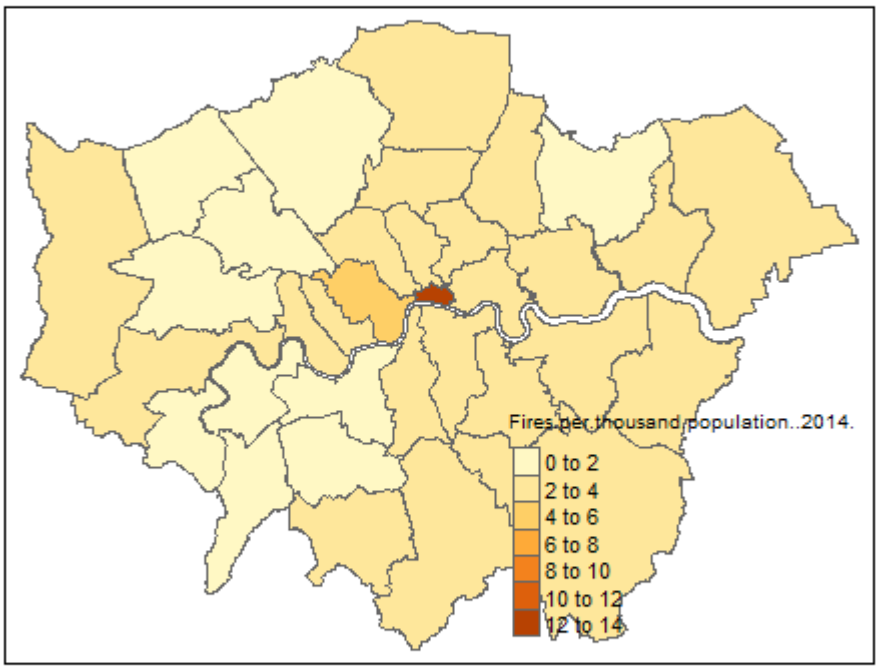


Fig. 27 - Spatial visualisation of Fires per thousand population data

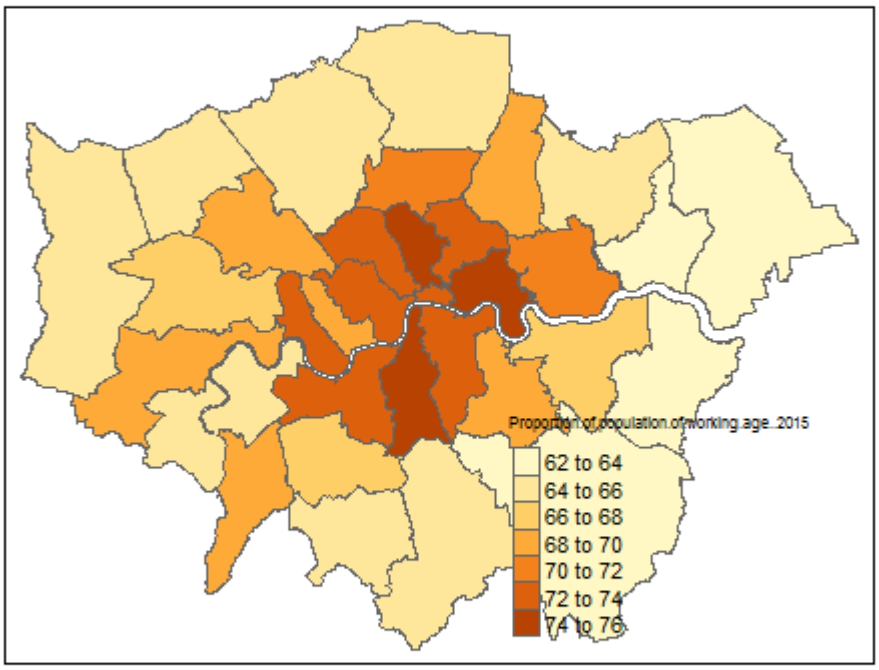


Fig. 28 - Spatial visualisation of Proportion of population of working age data

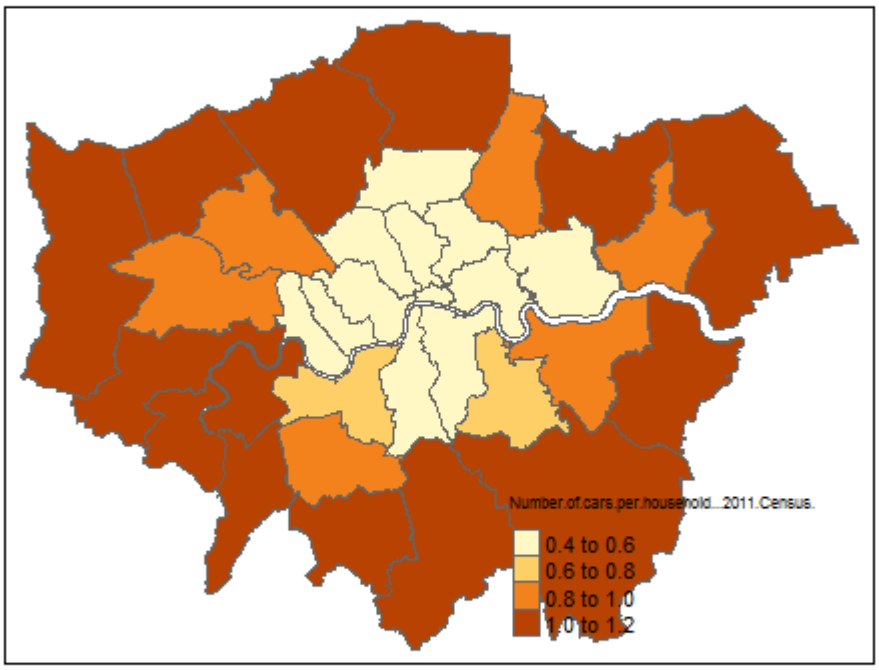


Fig. 29 - Spatial visualisation of Number of cars per household data

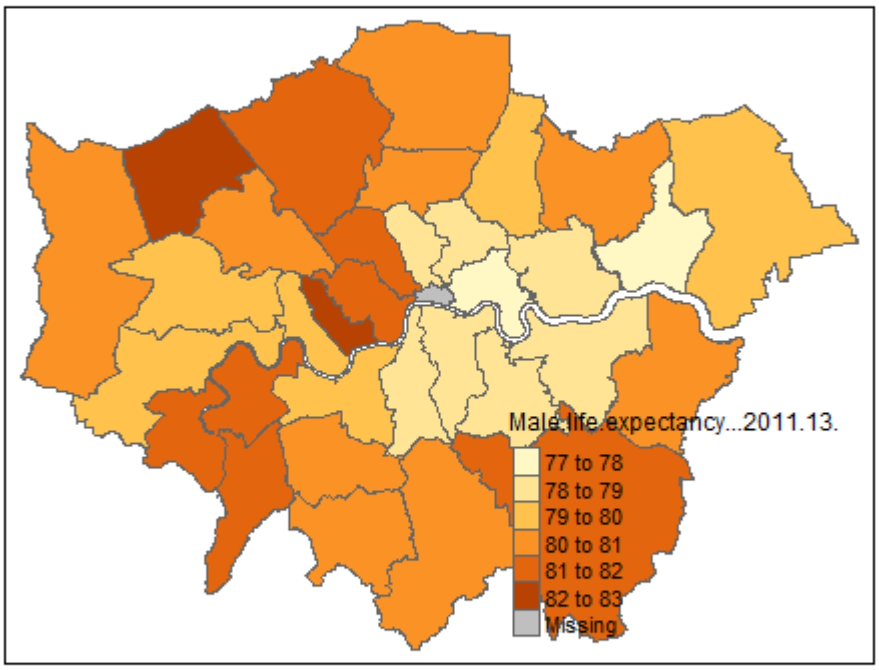


Fig. 30 - Spatial visualisation of Male life expectancy data

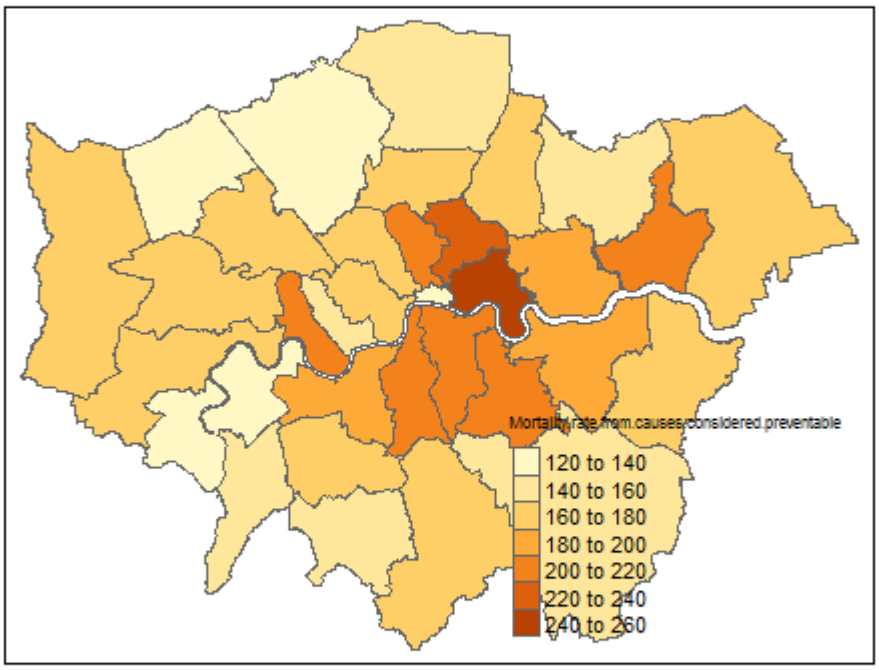


Fig. 31 - Spatial visualisation of Mortality rate from causes considered preventable data

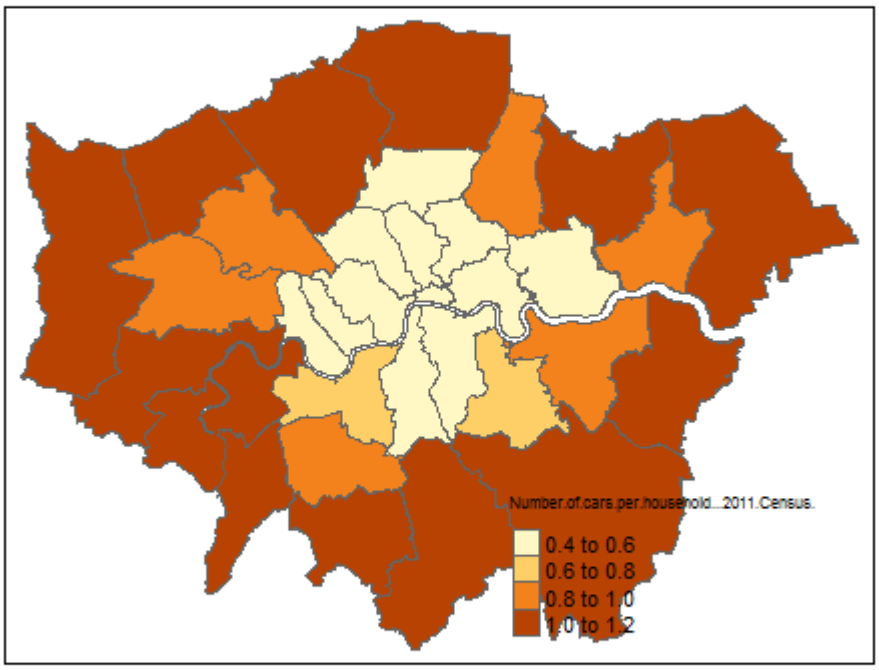


Fig. 32 - Spatial visualisation of Number of cars per household data

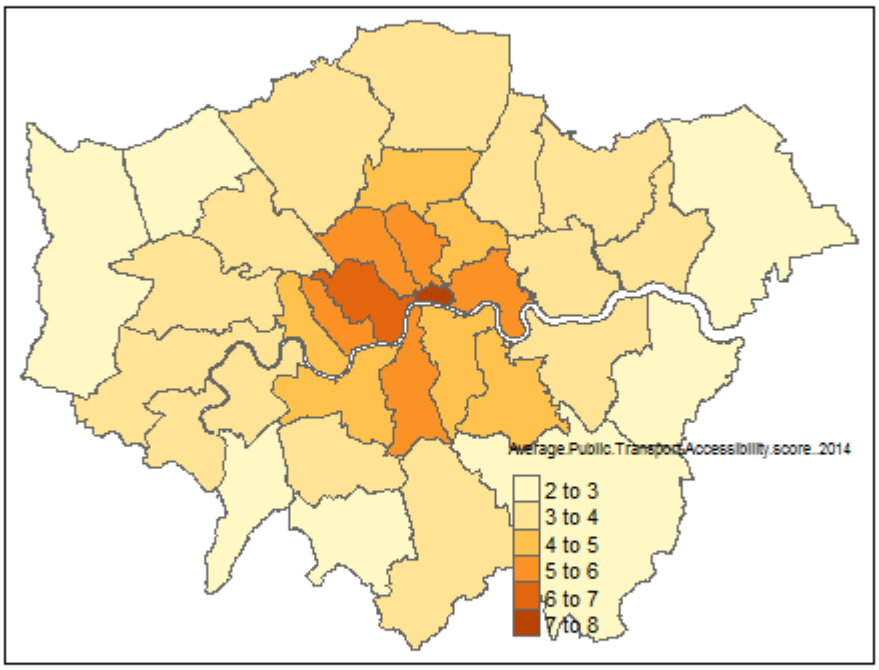


Fig. 33 - Spatial visualisation of Average Public transport accessibility score data