Demographic Analysis of London: Focusing on ethnic disparity among citizens and boroughs.

Coursework Report for Module INM433 "Visual Analytics"

Mussa Yousef

MSc Data Science

Abstract— This project explores the inequalities between ethnicities in London. We investigate disproportions amongst boroughs, focusing on the disadvantages, if any, within the black ethnic minority. Using visual analytic tools, we attempt to demystify quality of life by scrutinising Social, Education, Health, and Employment discrepancies among ethnicities in London. We analysed each segment separately, expressed visually using geospatial analysis and linking density and age dispersity to find links between boroughs, age, and ethnicity. We also used Machine Learning, building regression models, and utilised clustering to discover trends and understand the underlying reasons for these inconsistencies.

We found disproportions among employment and social status, where a high number of citizens among the white ethnic group dominating these positions, regardless of qualifications, where we also find Black and Asian ethnics surpassing their peers. These findings allowed us to focus on the key age group between 25-29, scrutinising the differences within this group brought insight into the depth of how embedded the inequality within London is. Furthermore, upon reflection, we have only scratched the surface in understanding the depth of inequality in London. This project sheds light on the problem London and United Kingdom faces, however, this isn't the solution.

1 Problem Statement

Black Lives Matter (BLM). The recent rise of the BLM movement was a focal point in 2020. After the killing of George Floyd in the United States of America, in what felt like the final straw, we saw a huge uproar detesting policing of black citizens and highlighting social inequality in the UK. Although the BLM charity was established 6 years ago to combat police brutality, they have also highlighted the social disparity and lack of opportunities for black citizens. This project will be analysing the societal construct of London, using the census 2011 data collected from citizens. We will be using data visualisation tools to identify inequalities amongst boroughs and how this affects an individual's quality of life. Also, scrutinising if black citizens are at a disadvantage in education and career prospect based on where they live. The dataset we will be using is high in dimensionality, structure is built by 610 Columns describing most aspects of society, from age and economic activity to health and qualification. The dataset is rich in information and has 649 rows from all districts which are fundamental in the structure of boroughs and ultimately the capital city of the United Kingdom, London. Thus, the census dataset will enable a full scope of exploration, allowing connections to be made between boroughs.

2 State of the Art

The fight against inequality has been a reality since the beginning of time, with gender equality still a striving discussion we have yet to solve. We find societies particularly in western countries to have vast differences in race, religion, cultures, and wealth. Measuring inequality is often considered in terms of income or consumption [4], however, this is as important to other societal dimensions such as education or health which both have a huge impact on the quality of life. Researchers have investigated inequality from several perspectives more notably in policy work, where the European Commission, has uncovered inequality links to societal effects i.e., higher crime rate/worse health outcomes. We will be intensifying these findings with the motivation of 'recognising inequality to be structural and deeply rooted' [1], hence focusing our research on the aspect of the race, more specifically the black race.

There are three papers we have analysed, which all share a different perspective and varying methodologies using visual analytics to depict inequality. "Ethnic Inequalities in London" [1] is the most tailored research paper we have examined, this paper studies the inequalities in London using both the 2001 and 2011 census and comparing the changes between boroughs. This paper has divided their findings and research into Education, Employment, Health, and Housing in London. Both 2D stacked bar charts and geospatial graphs were used to illustrate exploration, some interesting regression analyses also used, to examine the correlation of racial inequality between various dimensions.

Although this brought great findings, we felt only the surface was scraped and this could also be translated using geospatial visualisation. We will be using the same approach in breaking down sectors to investigate, however the technique to scrutinise each sub-dimension has been inspired by "The changing Anatomy of Economic Inequality in London 2007 – 2013" [3]. This Journal focuses on the disparity concerning economic inequalities between different ethnic groups. This report took a more statistical approach, utilising the calculation of indexes between boroughs and representing these using geospatial and dual-sided histograms. We will be exploiting this technique, however incorporating the visual analysis seen in the report [1].

The third journal we will be referencing; 'Inequality Briefing-defining and Measuring Inequality' [4] takes inequality from a global perspective, this journal explores methods of measuring inequality, the dimension in which inequality could be measured in and therefore considering factors behind inequality. This report will inspire theory, by understanding the effects of inequality, and introduce how we will be able to quantify inequality.

3 Properties of the Data

The Office for National Statistics reports directly to the UK Parliament, conducting statistical research on the general population. These reports are released forming the census dataset released every 10 years. The most recent census release was in 2011, therefore we must consider that the social structures may have slightly changed in the last 9 years. Nevertheless, findings in this report will fuel and motivate us to reconstruct our research and compare when the 2021 census is released.

The census dataset was extremely fruitful with information, collected from citizen households in London. The dataset has minimal missing information with 17 districts missing values which accumulates to 2.62% of the entire dataset. Hence, we will be deleting these rows, all of which are districts on the outskirts of London. The dataset has 610 columns describing the demography of 649 districts and 33 boroughs, ranging from religion, qualification, social status, economic activity, and crime. The data has general information on each district calculating total population, gender, and breaking down each age range. General information has also been incorporated with data, creating the percentage of statistical columns that add insight and will be useful with analysis.

The dataset is high in dimensions; hence we must consider the 'curse of dimensionality' [5] where studies have been made to find the structures which have been embedded in a large dataset. Therefore, we will first utilise multidimensional scaling (MDS) to represent the dataset in a low dimensional space. MDS allows us to preserve the variance between data points, mapped on a 2D Cartesian plane hence easily deriving hidden structures.

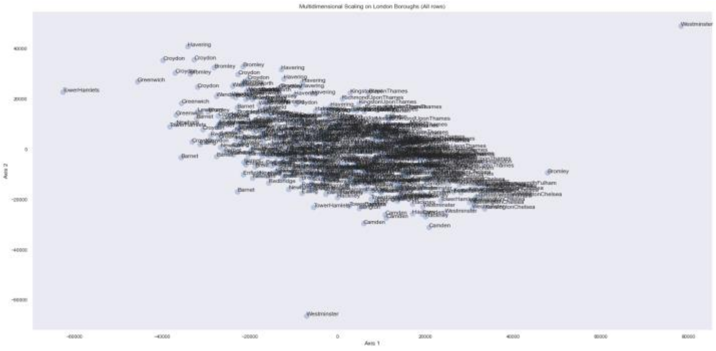


Figure 1: Multidimensional Scaling of all columns in the census dataset

Figure 1 expresses all the information in the census data using Multidimensional scaling. We first created a pairwise matrix which shaped a 624 by 624 matrix, we then normalised the data centralising all data at zero, hence calculating the Euclidean distance between each district and mapping these (Figure 1). Both Axis 1 and Axis 2 doesn't necessarily signify anything, we rather focus on the distance between points to uncover dissimilarities between boroughs. Initial visual analysis of the dataset shows a cluster of boroughs sharing characteristics. However, a few outliers identified particularly, Westminster, Tower Hamlets, Croydon, Havering, and Greenwich, tells us that there potentially some differences in characteristics within these boroughs. We will interpret this with caution, considering high dimension data points become almost equidistant from one another as dimensionality increases, therefore we will examine a subset dataset taking only columns describing the ethnicities of each borough.

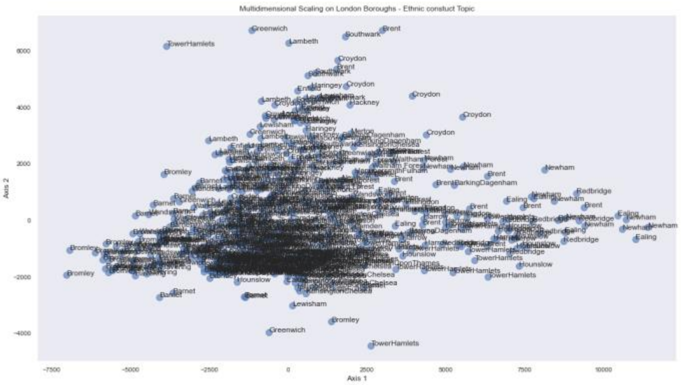


Figure 2: Multidimensional Scaling of subset data only considering Ethnic characteristics a in the census dataset

The subset (Figure 2) allows an analysis of the construct of the ethnic composition of each borough in London. We also find a cluster suggesting similarities in diversification between boroughs. However, we find the same outliers in Tower Hamlets, Greenwich, and Croydon. With the addition of Bromley, Brent, and Lambeth. The visual outliers found early in the structure indicates boroughs are not as uniform in subcategories, ethnic differences between boroughs fundamentally play a key role in all other indicators.

4 Analysis

4.1 Approach

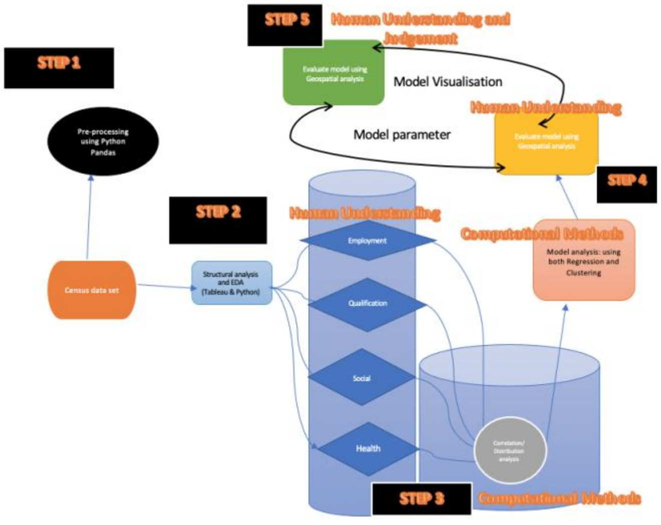

STEP 1
Pre-processing using Python Pandas
Census data set
STEP 2
Structural analysis and EDA (Tableau & Python)
Human Understanding
Employment
Qualification
Social
Health
STEP 3
Computational Methods
Correlation/ Distribution analysis
Computational Methods
Model analysis: using both Regression and Clustering
STEP 4
Human Understanding
Evaluate model using Geospatial analysis
Model parameter
Model Visualisation
STEP 5
Human Understanding and Judgement
Evaluate model using Geospatial analysis

Diagram 1: Tailored Project workflow diagram showing the Visual analytic process carried out in this project. Broken down into 5 steps.

Our approach will be broken down into 5 steps (Diagram 1). Step 1, we begin by pre-processing the Census dataset, excluding null entries and configuring headers for each subcategory. Step 2, we deconstructed the embedded structures looking at how boroughs are dispersed by solely looking at the variance between columns. Both Figures 1 and 2 are the results of this analysis.

The core of the project begins at step 3 where we navigated the direction of our analysis. From [3] we find that dimensions usually analysed when studying inequality are Education, Health, Nutrition, Security, Power, Social inclusion, Income, and Assets. Likewise, in [1] 4 key areas 'Education', 'Employment', 'Health' and 'Housing' were analysed to gain insight on inequality. We, therefore, wanted to choose the dimensions in which would be efficient in computational understanding yet maximise human understanding. For instance, the computational analysis of education may uncover gender disproportion through human understanding. Hence studying both education and gender demography wouldn't be resourceful. We, therefore, utilised Principal Component analysis (PCA). We reduced the dataset from 610 to 2 dimensions which allowed us to keep 63% of the variance. We then examined how each ethnic group is dispersed and found varying results. Checking the weighting of the 10 highest loadings in magnitude we found, Principal Component 1 and Principal Component 2 had the majority of features which described Social Status, Qualification, Economic Activity, and Health. Therefore, these are the four areas we chose to focus our analysis on.

Step 4, we now constructed regression models analysing our four focus areas related to White, Black, Asian and Mixed ethnics, we also looked at how residuals are distributed which helped us navigate and choose models that fit best without overfitting. We also applied these relationships across all London boroughs examining territories that have accordance with varying order of polynomials. Ideally, we would want to scrutinise each borough building multiple models then cluster each borough to an inequality index. However, this would take a lot of computational resources. We rather compared age and area density to better show the relationship between our dimension through our models. This then allowed us to refine our model, by clustering using each ethnicity and examining the distribution of each ethnicity proportional to each area of analysis.

Step 5, we looked at refining our model and comparing a single model to several partial models. In this part, we pay particular attention to how we measure the accuracy and complexity of our final model. Hence, we investigated different parameters and its ability to capture the complex dependencies. Finally, understandability, where we utilised our ability as a human analyst to understand the model output when we use different inputs, balancing the trade-off between model complexity and understandability. Once all five steps are completed, we would have a complete understanding of how Social Status, Qualification, Economic Activity, and Health are proportionally varied between White, Black, Asian and Mixed ethnicities. Therefore, we would be able to build a complete critical reflection.

4.2 Analysis Process

Our analytical process began with the use of Principal Component Analysis (PCA). We first established our pathway to breakdown each dimension analysed in the census data. We initially tried to use MDS to find the differential variance between subcategories. Although MDS allowed initial analysis of structure, this didn't bring as much insight to the ethnic discrepancies in the dataset. PCA allowed us to retain variance and investigate which dimensions had influence on the distribution of ethnicities. For instance, PC1 showed heavy weighting in Health and Qualification dimensions whereas PC2 had heavy weighting in Social and Economic Activity. As seen from Figure 2 we find both PC1 and PC2 to have a high influence on both White and Black distribution, we also applied this to Asian and Mixed Ethnic groups and found varying results. Hence moving into the core of this project with the help of PCA we have chosen our subcategories to focus on which will uncover why such high variance are embedded in the demography.

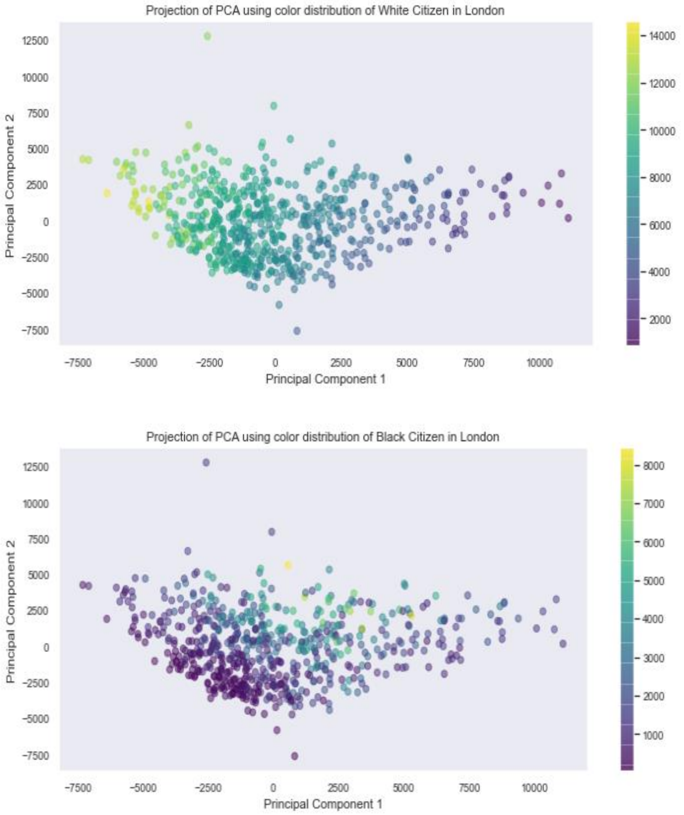


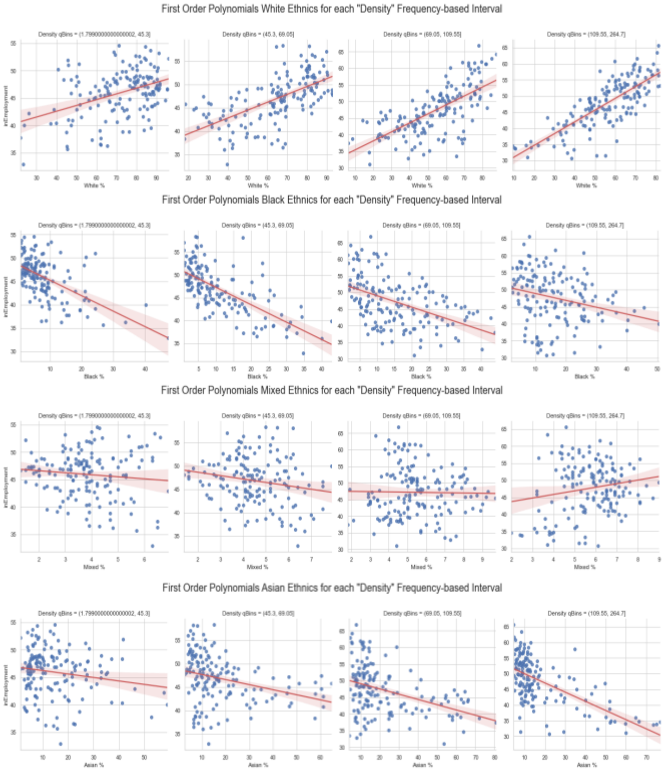
Figure 3: Principal Component analysis of dataset reduced to 2 dimension and colored using white and black ethnic weighted distribution

With the help of PCA we segment the data into subcategories, we split the data respectively into Ethnic, Social, Economic activity, Qualification, and Health including general information of each district i.e., Borough, Density and Mean Age. We then merged each category being analysed with the ethnic subcategory to prepare for analysis.

Focusing on the economic activity of citizens in London. Statistics show that ethnic minorities in England and Wales have a history of higher rates of unemployment [1]. Applying Pearson's correlation analysis, we uncover aligning results in our research. Pearson's correlation coefficient calculation shows White ethnic groups to have the only positive correlation with Employment, Self-employment, and Retired features. Whereas Black, Asian, and Mixed ethnic groups have negative coefficients, with the black ethnic group sharing a high correlation coefficient with unemployment.

Employment encapsulates most of the other dimensions and showed a strong linear relationship between all ethnic groups. Therefore, we fitted 1st – 4th order polynomials to assess these visually. The distribution of residuals showed the most random results for 1st and 2nd order polynomials. Considering 'Law of parsimony' [6] which states, "when presented with competing hypothetical answers to a problem, one should select the one that makes the fewest assumptions". We will use 1st order polynomial to assess the relationship between employment and ethnicity as it provides the least complexity. Furthermore, the initial Pearson's analysis fuelled this choice, as it measures the linear relationship hence this seems like a good fit.

Figure 4: Regression analysis of subcategory, split between White, Black, Asian and Mixed ethnic groups in Employment. Binned by Area Density of London Boroughs



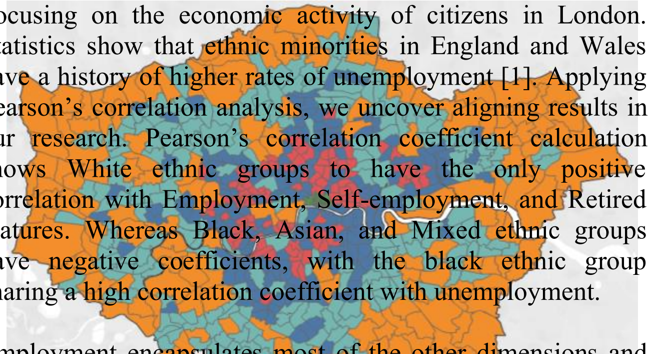


Figure 5: Geospatial image of clustering using density binning in the regression model built in figure 4. (Orange cluster=Low density) (Light blue cluster=MidLow Density) (Dark Blue cluster=Midhigh Density) (Red Cluster=High Density)

We find areas which are highly dense (Figure 5), white ethnic is far more likely to be employed or self-employed. We also discover residuals diminishing as the density increases which increases our confidence in the model. These findings are linked to the construct of London council estates being far more concentrated in the inner London boroughs, figure 4 shows the densely populated areas to be boroughs such as Islington, Newham, and Southwark, parallel to the [1] report which also show these boroughs are among homes which are also overcrowded.

Furthermore, interestingly, areas that are less dense Black and Asian ethnics have a higher rate of employment, considering the general growth of an individual's career 'Maslow's hierarchy of needs' suggests seeking security is inevitable. Therefore, movement into areas where the majority are working-class and white is the general trend, which makes up these statistics. Now, this brings me to my next area of analysis where we look at the social status and type of roles done by each ethnic group.

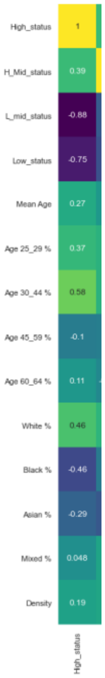
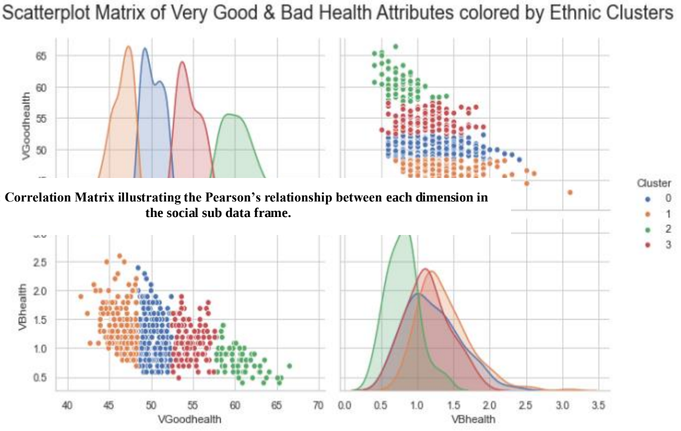


Figure 6: Correlation Matrix illustrating the Pearson's relationship between each dimension in the social sub data frame.

We
socia
status



follows similar trends to the employment analysis. We find a polar opposite correlation for White and Black ethnics who are in top-end roles such as Management/professional roles. The majority of those who are working who are of black ethnicity are in Semi-skilled and unskilled manual occupation. We investigated the interrelation differences between territories for black ethnics, by building regression models for each borough. Interestingly, we find those territories that have a low density in population, such as Kensington and Chelsea, Bromley, and Westminster show high levels of high-status job professions. This confirms the notion previously found that black individuals that are working and have increased income tend to migrate out of social housing. Although we find some insight from aligning the investigation with the density, this did not allow us to justify inequality. Therefore, looking at the correlation (figure 6) we find an interesting correlation with citizens aged 25-29. This age group will be the running force of London's foreseeable economy, however, unemployment amongst this age group is greater than the rest of the United Kingdom [2]. Pearson's coefficient within this age group was calculated to -0.29, 0.2, and 0.14 for White, Black and Asian Ethnics, hence this shows that the majority of the population within this age group of black and Asian ethnic backgrounds.

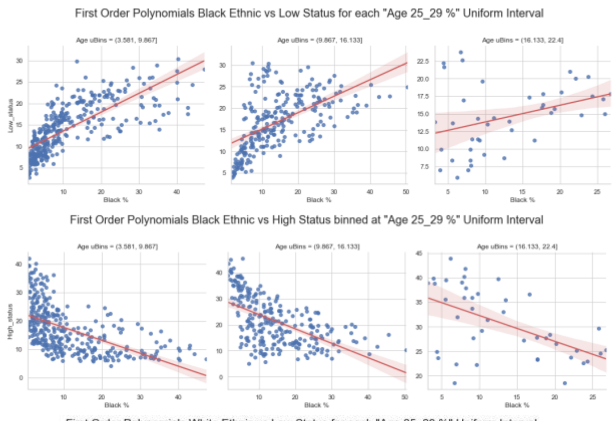
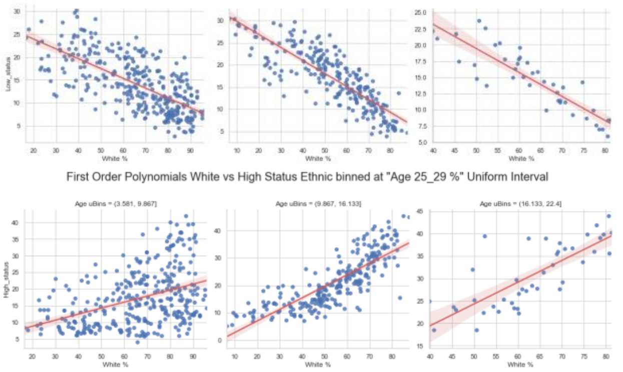


Figure 7: Regression models of White and Black Ethnic groups binned at 'AGE 25-29%'



We built further regression models binning into 25-29 Age group. We found that areas that are highly populated, are much more dispersed in social status among white ethnic citizens compared to black ethnic's citizens. The regression also provided far more confidence for white ethnic's, in areas that are highly populated with the age group. This suggests that these areas are best indicated by boroughs such as Camden, Ealing, and Harringay. Have the highest rate of inequality among ethnic groups, with equal levels of ethnic groups within these boroughs we find white ethnics between 25–29-year-olds are far more likely to have a better paying career. Considering at the age of 25, the most common credential an individual would have which would propel their career is a qualification, hence this will be our next segment of analysis.

The dataset collected levels of education from a community level, however it is often studied in terms of age groups [3]. Therefore, specifically looking at the age group between 25-29, there are approximately 800,000 citizens in London who are between this age range with 1.2% of those not having a single qualification. We find all boroughs have some citizens with no qualification, however, boroughs that are the least dense and have the highest recording of non-black citizens seem to share the highest rate of no qualification. Interesting since these boroughs also share the highest rates of individuals who are in managerial and professional roles.

Furthermore, we build cluster models using ethnicity to be able to visually differentiate between varied characteristics between health levels and ethnic groups, using the two extremes of health (Figure 8). We find Asian ethnic groups have the best health recorded, however, the worst of health were of black ethnic citizens. This could be due to an external factor related to habits or lifestyle perhaps however tailoring these findings to our research, levels of bad health may impair an individual to be able to withhold a career or attain a qualification, hence affecting social status.

We explored other external dimensions in the data such as Religion and Crime, which affect the demographic statistics. We found religion among citizens diminishing among younger citizens with the majority of the age group between 25-29 claiming they abide by no religion. Furthermore, we find the majority of crimes in London are also being committed within this age group. To be able to visually get an indication of the relationship between crime and ethnicity, we used regression analysis. We found low levels of residuals in first-order models analysing crime and ethnic backgrounds. We can deduce that areas with a dense population of the white ethnic group tend to have lower rates of crime compared to black and Asian densely populated boroughs such as Newham and Lewisham. Moreover, as density decreases for both ethnicities, we find Asian and Black ethnic experience less crime within their territory whereas white citizen uncovers a rise. 'Inequality is often a significant factor behind crime and social unrest' [4], Therefore from our previous discoveries crime rate tends to be a result of low social status and unemployment where we find the majority of these areas are found to be in districts which have high number social housing, low employment rate.

We begin our final step by merging the partitioned data which was used to analyse inequality. Once merged we split the data into train and test samples. We utilised the Wrapper method [7] to feature select using all four ethnicities mutually exclusive as our target variable. The Wrapper method is prone to overfitting, therefore as our main focus here isn't the actual model built, rather use the indicators extracted from the process. Hence overfitting the dataset is ideal for our next step which is to build regression models by clustering using ethnicity. Therefore, using the wrapper method for feature ranking respectively for each ethnicity targeted, we find 'Age 25-29 %', 'inEmployment','Qualification_Level2'and 'VGoodHealth'. We will therefore use these dimensions for our final model.

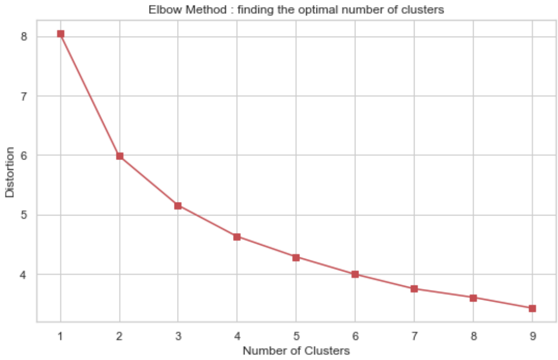


Figure 9: Elbow method: Iteration of clusters for regression models to minimize distortion.

We used the Elbow method in order to determine the number of clusters which minimises the level of distortion (figure 8). The optimal level of clusters to use in our model is 4, as we will be building multiple models to assess inequality.

4.3 Results

Our first major finding from our analysis is that inequality amongst boroughs varies, however densely populated boroughs where one ethnicity dominates, or the majority are from the Asian or Black community have polar opposite social status. We built final models to assess how the relationship between dimensions has when binned with each respective ethnicity. However, before we examine our model, we assessed which polynomial order provides the lowest root squared mean error (RSME), on average using each ethnicity as the target variable we found the second-order polynomial to provide an average of 1.9% lower RSME.

Interpretations from final models:

- Black ethnics are the most likely to be in a low-status career compared to any other ethnic group.
- As employment increases among each ethnic group, we find health generally does get better however a more complex relationship is seen among black ethnics suggest a possible external influence.
- University degrees have the least influence on white or mixed ethnic job prospects; further, we find the least number of black individuals in high social status careers.
- Black and Asians tend to have better health when in retirement compared to white ethnics.

5 Critical reflection

This project gave us great insight into the discrepancies between ethnic and societal constructs in London. However, the topic of inequality is deeply complex and cannot be justified with only the census data set. Here we were able to 'begin' an inquest as to where discrepancies exist. Our initial research questions immediately brought limitations to our analysis. Firstly, to be able to identify inequalities among boroughs greater information was needed, as each local authority essentially has its own respective downfalls which aren't expressed in the census dataset.

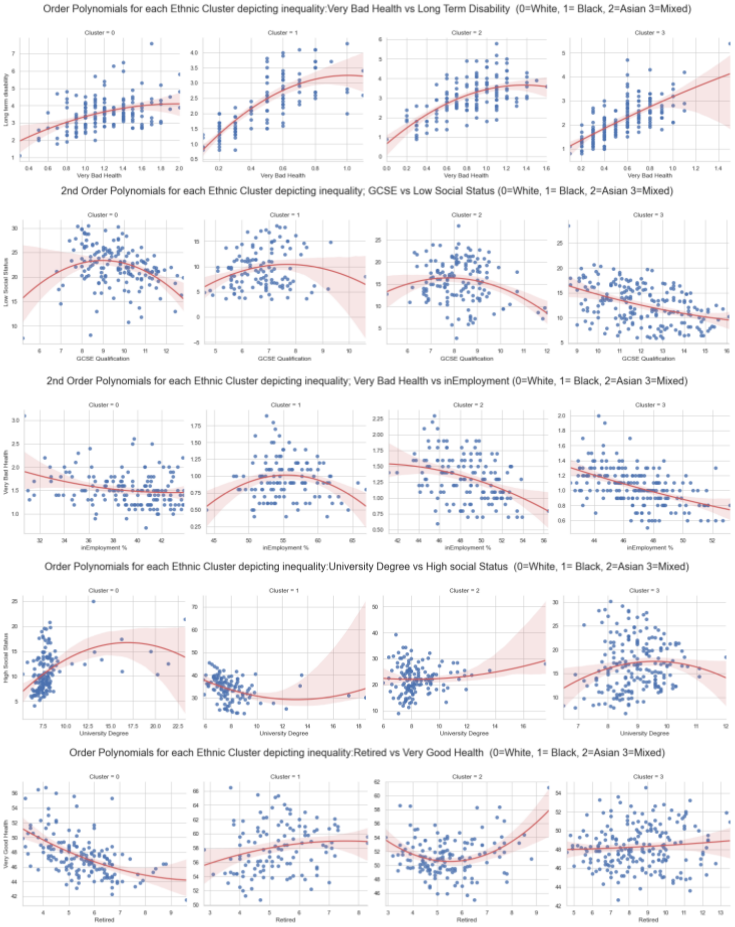


Figure 9: Regressions models using Clusters built using Ethnic splits from previous analysing interconnecting each segment chosen from the feature selection.

We were able to identify inequalities between boroughs, however, to be able to justify the scale to which this would impact an individual's quality of life, we have only scratched the surface on this. Quality of life is almost subjective differs among cultures, religions, and each individual's aspiration. Therefore, being able to deduce and justify the quality of life a baseline needs to be set to be able to compare to, potentially if we used the 2001 Census, we would have been able to compare our findings to previous years as [1] have done. Although this would have been able to shed light on any deteriorations or improvements in dimensions analysed, this still wouldn't have justified effects on an individual's quality of life.

Our approach allowed a meticulous analysis to be made of each dimension influencing the variance of ethnicities in London. Equally, this project may have been more insightful if one dimension was scrutinised i.e., Education, Social status, or health, among different ethnic groups. This would have allowed us to use the census dataset and also bring in other datasets such as housing information or medical data. In turn, this would bring more flexibility and derestrict our research, the same approach would be done for each partitioned dataset to then indicators calculated to express inequality among ethnic groups.

My advice to future analysts,

breakdown each dimension and use the census data as an applicator. For instance, as we have researched economic activity here, we would suggest prerequisite research to be exploited beforehand. Data from the DWP perhaps, specifying amount received from tax and by who and which borough. Furthermore, business information on different ethnicities who own businesses, and those who are work voluntarily. Once the entire picture is painted, we can apply this to the London, demographic data, this is what I mean by the applicator. Throughout this project, we have felt the census commands a 'facilitator' role for more descriptive datasets. We were able to gain some insight but not to the standards to which we can say the research question has been answered. We have merely touched light on the inequality within London. We will be revisiting this dataset once Census 2021 is released, this time much more vigorously.

Table of word counts

Problem statement	217/250
State of the art	413/500
Properties of the data	487/500
Analysis: Approach	496/500
Analysis: Process	1498/1500
Analysis: Results	188/200
Critical reflection	459/500

References

[1] Capital For All- "Ethnic Inequalities in London"

https://trustforlondon.fra1.digitaloceanspaces.com/media/documents/London-Inequality-report-v3.pdf

[2] T.F. Cox, M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.

http://usir.salford.ac.uk/id/eprint/57930/1/The%20UK%20is%20not%20innocent.pdf

[3] Social Policy in Cold Climate- 'The Changing Anatomy of Economic Inequality in London (2007-2013)

https://trustforlondon.fra1.digitaloceanspaces.com/media/documents/RR06.pdf

[4] Inequality Briefing – 'Defining and Measuring Inequality'-Univerisity of Nottingham

https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/3804.pdf

[5] Raul Rojas - The Curse of Dimensionality

https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf

[6] Occam's Razor –'Law of UX'

https://lawsofux.com/occams-razor

[7] A comprehensive guide to Feature Selection using Wrapper Methods in Python
https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/