**Investigating whether ethnicity affected Covid-19 susceptibility in London in 2021**

**Group:** Data 1, Group 4

**Group members:** Sin Nee Teoh, Yvonne Cowell, Sabreen El-Zeftawy, Lola Pwasanga, Yuliana Turchyn, Lucy Stephenson

1. **INTRODUCTION**

**1.1 Aims**

To investigate whether ethnicity affected Covid-19 susceptibility in London in 2021, and the factors contributing to this. A variety of statistical methods and visualisation techniques will be applied to this question, and the investigation will utilise data from multiple sources.

**1.2 Objectives**

* Investigate the population structure in terms of ethnicity for each region of London (North, East, South, West & Central).
* Establish the difference in Covid-19 susceptibility between London regions related to ethnic group demographics?
* Investigate which ethnic group is the most susceptible.
* Investigate whether patterns of susceptibility between London boroughs with above and below average proportions of black and ethnic minority (BAME) are different throughout 2021.
* To investigate whether vaccine uptake is different between London boroughs of above and below average proportions of BAME population.
* To investigate whether patterns between race and Covid-19 susceptibility be explained by other factors such as socioeconomic background/underlying medical conditions.
* To attempt to predict deaths from vaccination uptake in London in 2021

**1.3 Roadmap of the report**

The report introduces our project in context to the wider issue of race and Covid-19 in London in the Background section. The group approach is outlined during the Steps specifications section and methodology during the Implementation and Execution section. Our key findings are presented along with our interpretation of our analysis in the Key Findings section, followed by our Conclusions.

1. **BACKGROUND**

**2.1 Introduction**

The Covid-19 pandemic involved devastating numbers of deaths and cases across London in 2021. London is an ethnically diverse city, with different ethnic group population structures associated with different regions. Numerous UK-wide and US-wide studies show that Black and Hispanic people were more likely to die from Covid-1915,16. This study differs from previous studies in its scope focusing on the ethnically diverse city of London and on the year 2021, when Covid-19 vaccine rollout began in London.

**2.2 Target audience**

The analysis presented in this report highlights potential disparity in Covid-19 affecting different ethnic groups, and therefore forms part of wider efforts to improve racial inequality in London. By understanding the driving factors behind potential differences in susceptibility between ethnic groups, efforts can be targeted to attempt to reduce differences. These results may be useful to local authorities and non-governmental organisations running such targeted programmes. For example, in Birmingham a charity ran a programme for elderly Pakistani communities with inherent medical mistrust, to try and improve vaccine uptake within their community. It can aid decision making by local authorities, including funding vaccine rollout programmes and allocation of resources within London healthcare centres.

**2.3 Questions and analysis**

We asked the question is there a difference in Covid-19 susceptibility between London regions related to ethnic group demographics. This question was answered using correlation analysis and linear modelling.

The question of how the pattern of Covid-19 cases changed over the course of 2021 and whether this differed between high and low proportions of ethnic minority groups was tackled by using a time series analysis. Since vaccination rollout and changing lockdown enforcements in London happened during this time period there is likely to be changing patterns over this time period that may have affected different ethnic groups differently.

In order to answer questions about the underlying factors behind disparity in deaths and cases, data exploration using univariate and bivariate analysis was undertaken followed by multiple regression (ordinary least squared) to determine the most significant variables. To select the optimal model all variables were standardised to a sample variance of 1 and sample mean of 0 due to very different standard deviations. Six OLS regression models were applied.

Since vaccination uptake is so strongly linked to ethnicity, being able to predict deaths from vaccination uptake is key to understanding how ethnicity might affect death rates. In order to see if vaccination uptake data could be used to forecast Covid-19 deaths, a machine learning approach was taken, using a forecasting model.

1. **STEPS SPECIFICATIONS**

**3.1 Approaching the key steps of data analysis**

The project was split into multiple steps and each step assigned different members of the team to tackle it. We split the project up into the following steps:

* + Framing the problem
  + Data gathering
  + Data cleaning
  + Data analysis
  + Machine Learning
  + Data visualisation
  + Report writing
  + Presentation slides
  + Project Submission

Our team decided to work in an agile format in sprints where each sprint - separate topic from the list above. And for each task either few people were assigned to (for example data gathering or machine learning) or whole team participated (for example data analysis)

**3.2 Data Sources**

Multiple data sources were utilised, found through google searches and using git resources containing lists of public APIs.

Gov.uk data was a key source used, accessed by a usable front end portal specific to Covid-19 data1. This contained data on deaths, cases and vaccination uptake across the time period in question and could be subsetted to the region of London as a whole, as well as to individual Boroughs. The public API associated with this data was used to access deaths, cases and vaccination for the region of London for part of the machine learning portion of our project2.

The London Data Store was another data source, from which we accessed a range of data relating to explanatory variables associated with Covid-19 cases such as ethnicity, underlying health conditions and deprivation3. Finally data was taken from NHS England’s website12, 13.

1. **IMPLEMENTATION AND EXECUTION**

**4.1 Development approach and team member roles**

Our Notion project management tool and git were set up by Sabreen. Data gathering was undertaken by Yvonne and Lucy. Data cleaning was undertaken by Sin Nee, Yuliana and Lola. Meeting minutes were kept by Lola. Data analysis was undertaken by the whole group, each member tackling different research objectives. Lucy and Yuliana took Machine Learning. Report writing and Presentation was also a whole group effort.

The elements of agile development that our team used included weekly scrum meetings. We also split the project into sprints relating to the different stages in the data cycle. We attempted to implement iterative development in the way we structured task allocation but time limitation meant this wasn’t always achieved.

**4.2 Tools and libraries**

For data cleaning, exploration and analysis Python libraries Pandas, Numpy, Matplotlib and seaborn were used. The Gov.uk Covid-19 API was accessed using ‘get’ from the requests library and ‘dumps’ from the json library. For machine learning Python libraries keras and sklearn were used. Data was taken from an SQL database of boroughs and sorted into regions using the libraries ‘sqlacademy’ and ‘pymysql’.

**4.3 Implementation process**

Achievements were made early in outlining our questions and finding appropriate data sources. Different analyses techniques were successfully applied to the data.

We decided to change our approach to some of our research objectives. Data was not available for specific minority ethnic groups on deaths, vaccination uptake and cases in London. So instead we reframed the questions to work with proportional data taken from boroughs, rather than raw population numbers for each ethnic group.

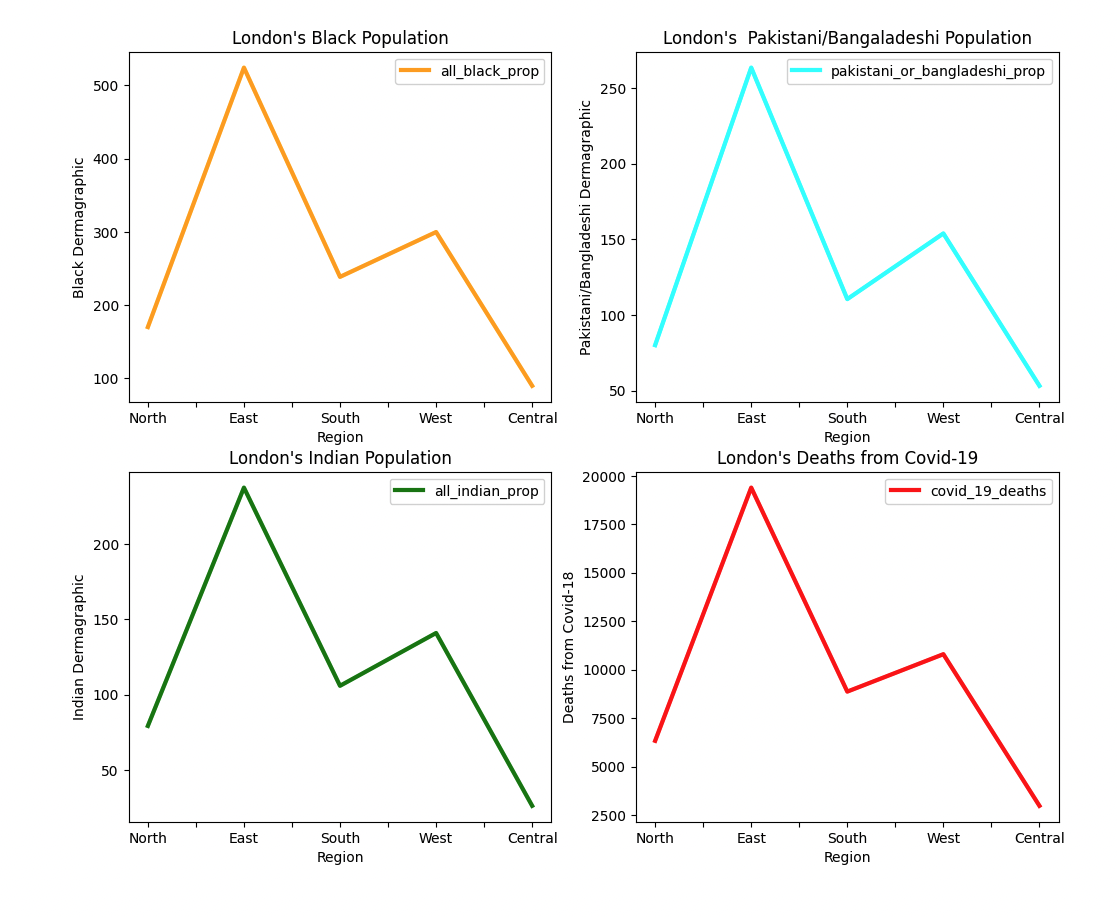
**4.4 Implementation challenges**

Joining data from different sources was a particular technical challenge we faced. We tackled this by having debugging meetings and trialling different solutions from stack overflow. The biggest challenge was meeting all of our objectives in the given time frame, with relatively short time periods available for sprints.

1. **RESULT REPORTING**

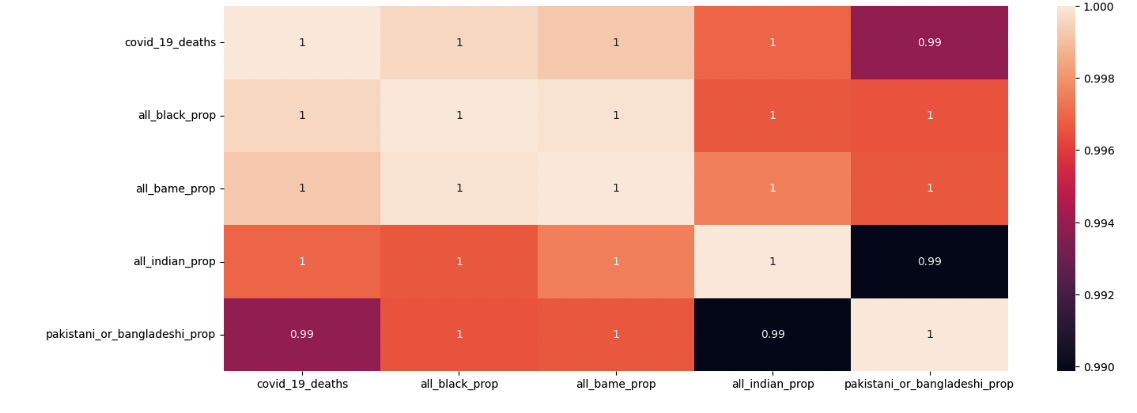
**5.1 What is the ethnic group population structure for each region of London?**

The line graph below shows the ethnicity specific demographic for each region versus Covid-19 deaths. Each plot has a near identical shape supporting the theory that the trend of Covid-19 death is closely linked with population ethnicity. The factor of ethnicity is seen to have a correlation with death rate. This can be seen more clearly on the heat map below.



**Figure 1: Showing the variation in ethnic minority group demographics in the wider London regions.**

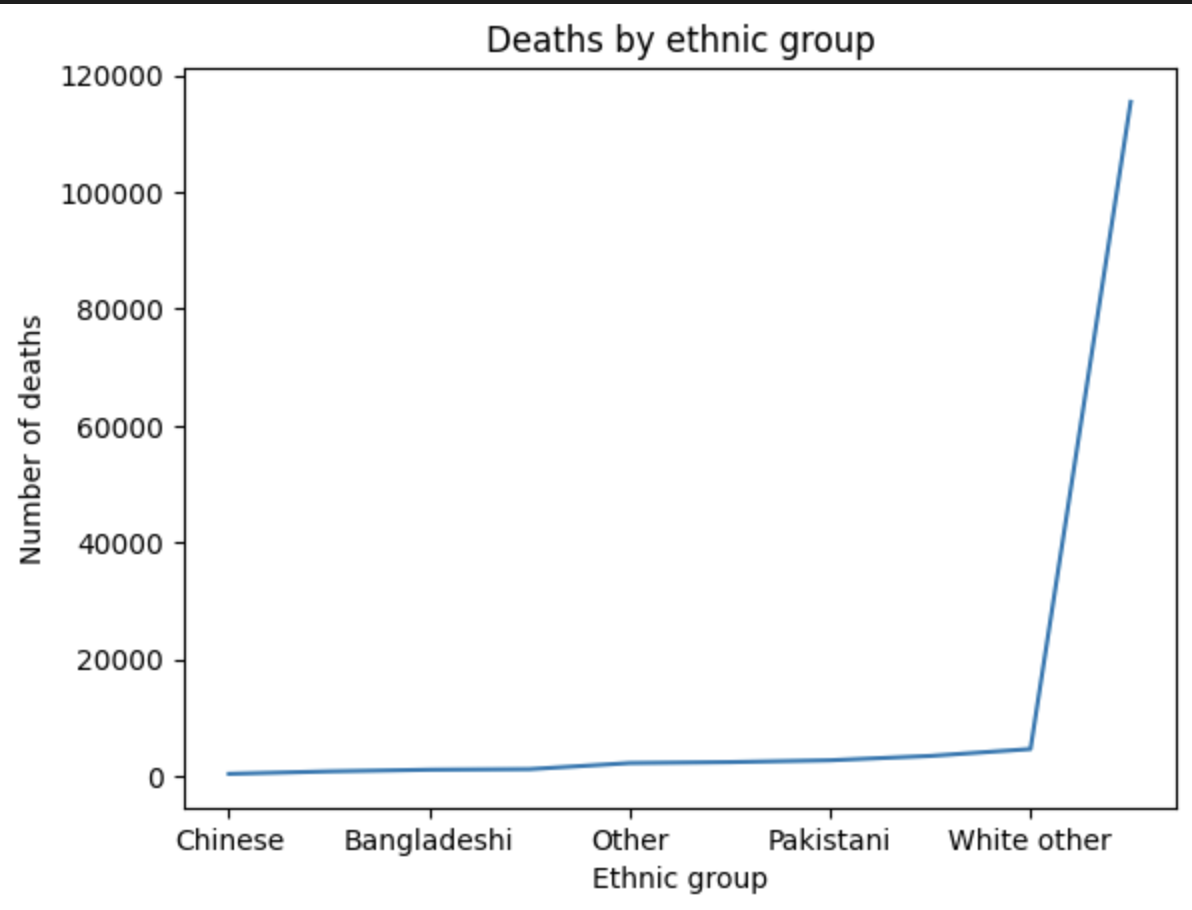
**5.2 Is the COVID-19 death rate higher in London Regions with a high ethnic demographic?**

****

**Figure 2: Covid-19 deaths and proportion of populations in a correlation matrix.**

Figure 2 shows that the proportion of black people in a borough has a strong positive correlation with the number of Covid-19 deaths. As does the proportion of overall BAME people. We can infer a linear relationship between the two. Proportion of population in boroughs who are in the Indian ethnic minority group has a strong negative correlation with those in the Pakistani or Bangladeshi ethnic minority group, showing the communities are largely living in different boroughs in London.

**5.3 Which ethnic group is the most susceptible to Covid-19?**



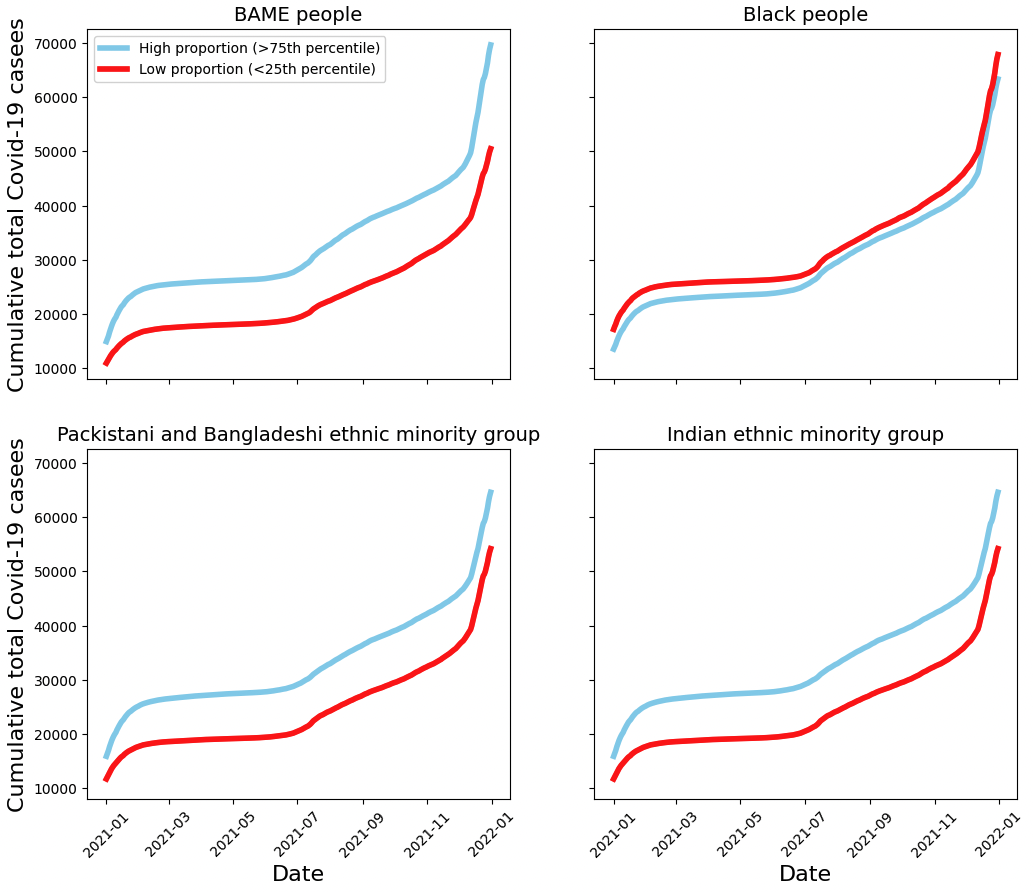
**Figure 3: Showing the variety of deaths between different ethnic groups in London.**

Figure 3 shows that White British and White other ethnic groups are showing higher number of deaths in comparison with others, which does not reflect the susceptibility of white people to Covid-19 but rather the higher number of white people in London. This graph may look different were proportion plotted as opposed to raw numbers. The least number of deaths are among Chinese ethnicity followed by Bangladeshi.

**5.4 Does the pattern of Covid-19 cases change over 2021 in boroughs of high proportion of ethnic minorities compared to those with low proportion of ethnic minorities?**

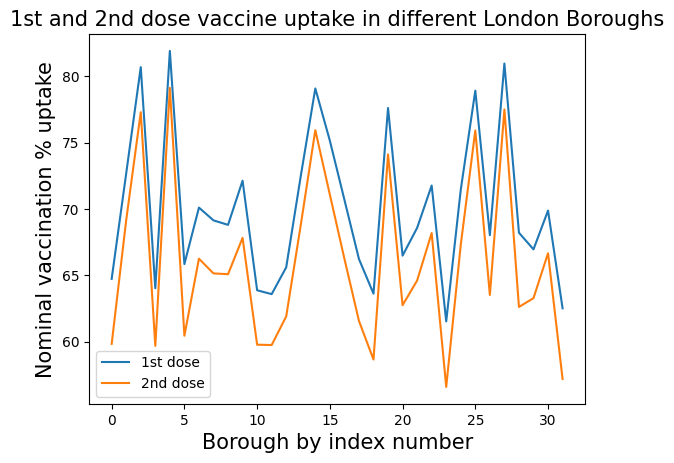
To answer this question high and low proportions were defined as having more than the 75th percentile proportion of an ethnic group and less than the 25th percentile respectively.

By looking at figure 4 it is clear that the pattern in Covid-19 cases during 2021 is very similar for all ethnic minority groups. Cases are increasing steeply at the start of 2021, and then plateau between January and May after the first dose vaccination rollout and introduction of another lockdown in London. They then begin to increase more sharply as lockdown is lifted and immunity wanes in July 2021.

The pattern is not very different between ethnic minority cases. Overall boroughs with a high proportion of BAME population have a higher Covid-19 cases over the course of 2019 than those with lower proportions. This is also true for boroughs with high proportions of Packistani and Bangladeshi and Indian communities. A gap of around 5,000 Covid-19 cases is maintained between high and low proportions throughout 2021. However, the gap is narrower between boroughs with high and low proportions of black people and the trend is reversed - boroughs with high proportions of black people have lower numbers of cases throughout 2021. ****

**Figure 4: Trend in Covid-19 cases over 2021, comparing high and low proportions of different ethnic groups.**

**5.5 Is vaccine uptake different between different regions in London?**

****

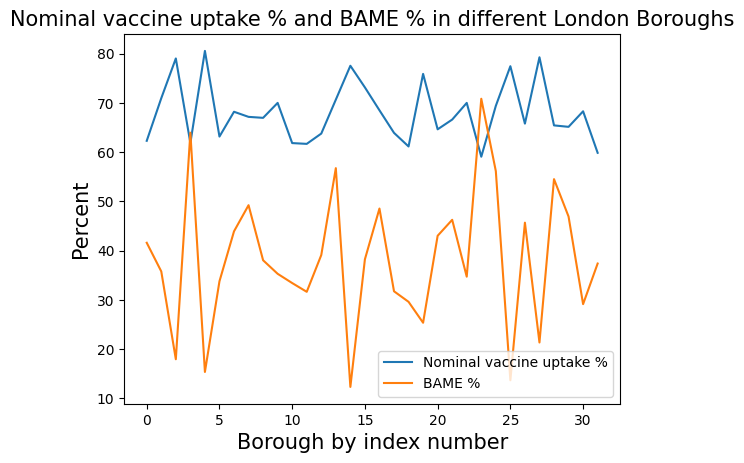
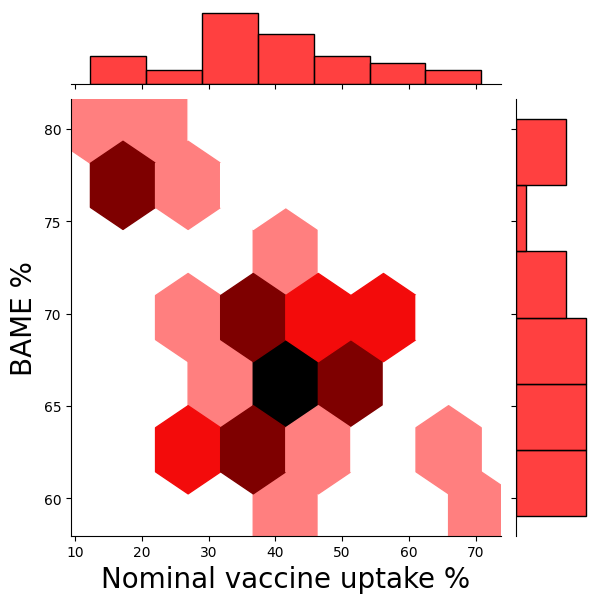
**Figure 5: Overlaid graph of first and second vaccine dose uptake % in different London Boroughs**

Vaccine uptake in all boroughs has a similar trend between the first and second vaccine doses. No significant difference is observed with mean difference of both doses at 5.9% (std <6%). Second dose is consistently slightly lower than the first dose.

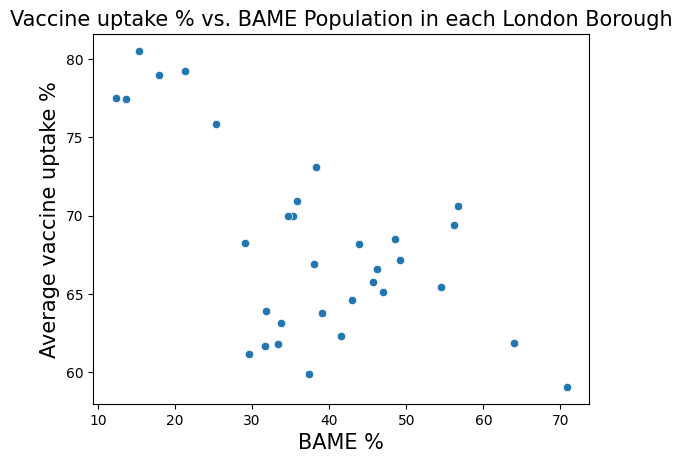
In order to summarise the vaccine uptake level in each borough, dose uptake % was categorised into different levels: Very low: 51-60%; Low: 61-70%; Medium: 71-80%; High: 81-90%; Very high: 91-100%

The result shows that 23 out of 32 boroughs (72%) have medium vaccine uptake level, 2 boroughs (6%) have low vaccine uptake level and 7 boroughs (22%) have high vaccine uptake level.

**5.6 Does vaccine uptake trend correspond to the proportions of ethnic groups in London?**

**a)b)**

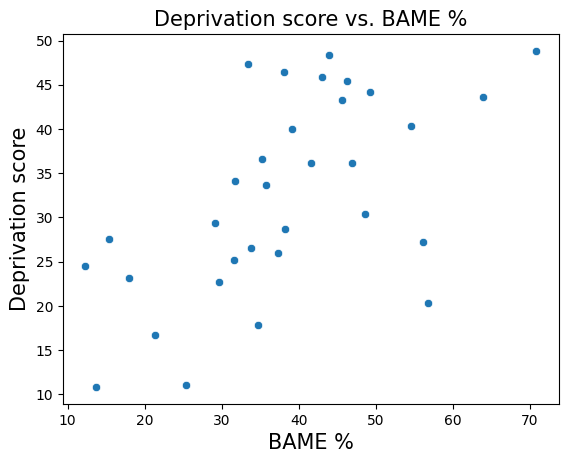
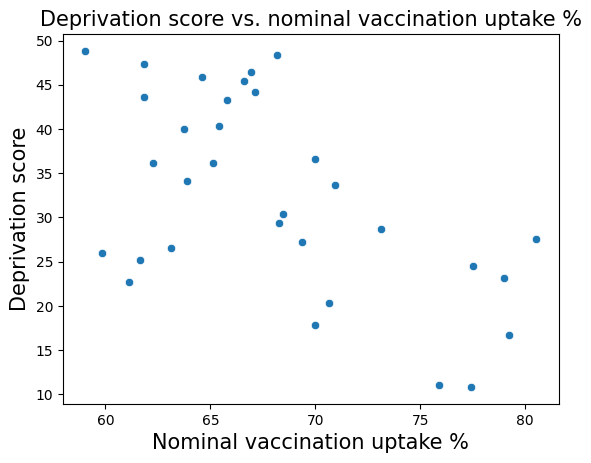
**Figure 6: a) Trend plot of both nominal vaccine uptake % and BAME % by index of Borough, b) Jointplot of BAME% vs nominal vaccine uptake %**

****

**Figure 7: Scatter plot of vaccine uptake vs BAME% in different London Boroughs**

The trend plot and joint plot show that vaccine uptake % decreases as the BAME population increases in some Boroughs. The scatter plot shows that there is a strong negative correlation at R squared value -0.62 between the average vaccine uptake % and the BAME %.

**5.7 Does vaccine uptake correspond to poverty in London?**

**a) b)**

**Figure 8: a) Scatter plot of deprivation score against BAME % and b) Scatter plot of deprivation score against nominal vaccination uptake %**

Deprivation score vs. BAME % scatter plot shows a moderately positive correlation at 0.59 between BAME % and deprivation score which means that Borough with a higher percentage of BAME population has a higher deprivation rate.

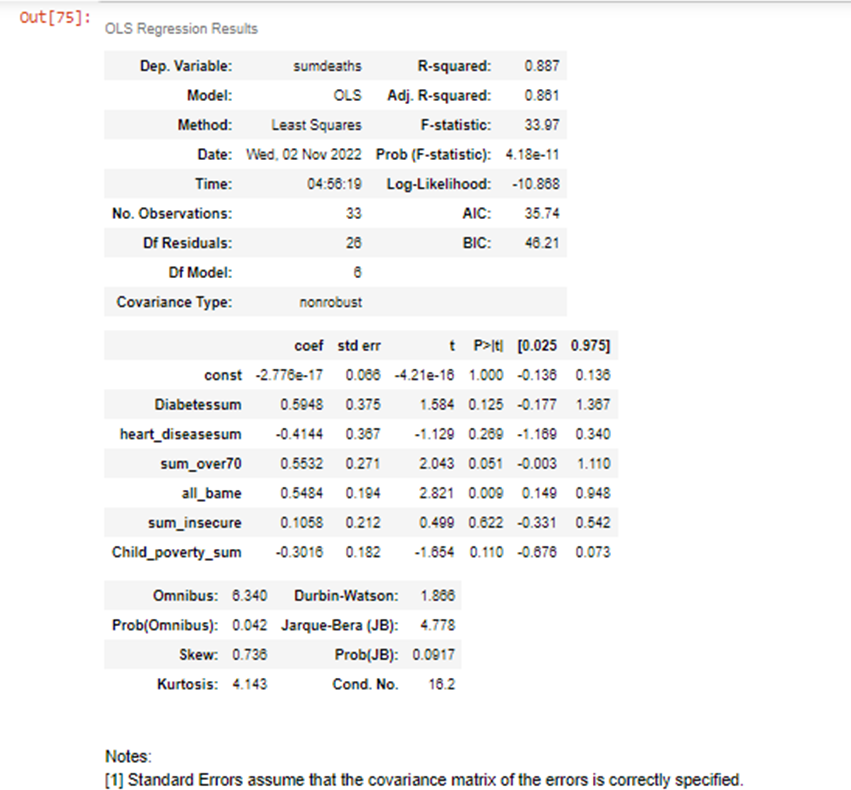
Deprivation score vs. nominal vaccination uptake % shows a moderately negative correlation at -0.57 between vaccine uptake and deprivation score which shows that Borough with high deprivation score has low vaccine uptake.

**5.8 Can patterns between race and COVID-19 susceptibility be explained by other factors such as socioeconomic background/underlying medical conditions?**

Analysis results are presented of some significant predictors of Covid19 deaths with a focus on ethnicity. Linear regression analysis was combined with multiple regression to identify statistically significant factors. Sixteen variables were analysed within the categories of ethnicity, child poverty, deprivation, medical conditions, occupation and population. This report finds that there is a positive correlation between Covid deaths and the BAME (black and minority ethnic) independent variable. The coefficients for bame , over 70s, Child poverty, insecure jobs, heart diseases and diabetes form the equation used to predict covid deaths. The results enable better understanding of the correlates between covid deaths and specific socio economic variables associated with the population in general.

**Regression results model4B**

Model4b has a high adjusted R-squared (86%) reflecting the fit and predictive power of the model, providing a better fit to the data than regression Model 5 with an Adjusted R squared of (58% )and an R-squared of 60% (see fig 1)..



**Table 1: Model4b OLS regression output**

*The Multi linear Regression equation for predicting covid deaths is:*

*Covid deaths = (constant)-2.776e17+0.5484\*allbame+(0.55220\*sum\_over70+ (-0.4144)\*heart\_diseasesum+(0.5948)\*Diabetessum+(0.1058)\*sum\_insecure+( -0.3016)\*Child\_poverty\_sum*

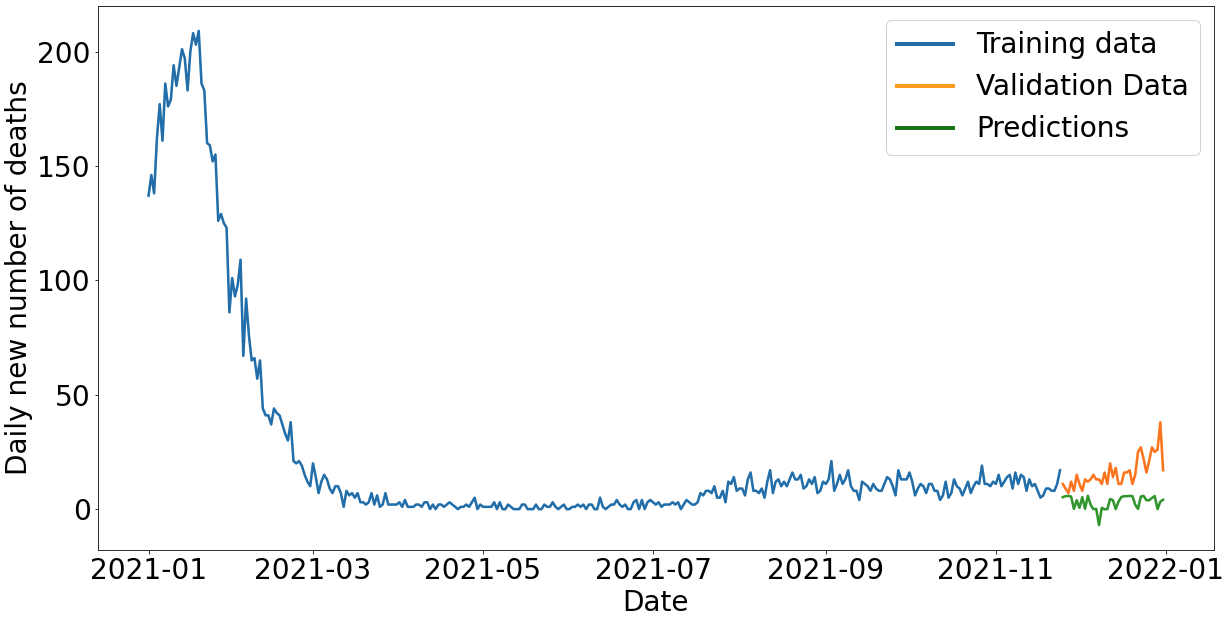
* Coefficients:For a given proportion of Diabetes, over70s, insecure jobs, child poverty and heart diseases, an increase in proportion by 1 unit of bame is associated with an increase in 5 covid deaths.
* The variables over 70s , Diabetessum, heart\_diseasesum, sum\_insecure and Child\_poverty\_sum all have large p-values indicating that there is no association with covid deaths therefore we fail to reject null hypothesis two.
* The bame variable is positively associated with covid deaths so null hypothesis one is rejected. (see jupyter notebook for details of all hypothesis)

**5. 9 Is there a relationship between BAME and Covid deaths? How strong is that relationship? Can this relationship be used to predict Covid-19 deaths?**

After fitting model 5 for BAME, the adjusted R-sq. value is .58 indicating a moderate fit. The r squared is 0.60, this indicates that BAME accounts for .60 proportion of the variability of the total covid deaths. The coefficient of determination shows that for every 1 unit increase in BAMEs within local authorities in London there is an association of an additional 5 covid deaths. A "unit" increase in proportion of population who are BAME is associated with a 8.63 "unit" increase in covid deaths. Here the coefficient represents an association between these factors. In this case, the p-value for bame is far less than 0.05 therefore we can reject the null hypothesis. This leads to the conclusion that there is a relationship between bame and covid deaths

Given a certain proportion OF BAME, can covid deaths be predicted? Here predictions are extracted using the statsmodels predict function: lm1.predict(X\_new). If there were an increase in proportion of BAMEs in a borough population by a proportion of 0.3, what would we predict for covid deaths? The prediction for covid deaths will be 51 using this model.

**5.10 Can Covid-19 deaths in London be predicted from vaccination uptake data using a machine learning approach?**

****

**Figure 9: Visualisation of training data, validation data and predictions using an LSTM sequential model to predict Covid-19 deaths from vaccination data.**

Our model underpredicts the number of deaths at the end of 2021. The model itself is not performing well using the given data. It could be because vaccination uptake is not the only factor affecting deaths, so a model that includes further parameters/terms would predict better - for example hospital capacities, social mixing etc. Another reason for underperformance may be the amount of data used to train the model or bias induced by the extreme downwards trend in our training data.

Comparing our predictions in new deaths to those made by Office of National Statistics for weekly deaths, similar predictions were made for the end of 202110. Other predictions made by UCL were also similar to ours11. However, these models did not use vaccination uptake for predictions. In contrast, one NHS funded study used a modelling approach which better reflected the real number of deaths recorded at the end of 202112. They included far more parameters than our model did, better accounting for the factors behind Covid-19 deaths. Our machine learning approach was rudimentary and limited by time. Given further opportunity, we would have tweaked the model, or tried different models and then performed model selection.

1. **CONCLUSION**

We successfully answered all the questions we set out to. Analyses showed that the variation in ethnicity (of overall BAME, Black, Pakistani and Bangladeshi and Indian ethnic groups) resulted in correlations between deaths and proportions of these populations. Vaccination uptake of both first and second doses varied between boroughs and was lower in boroughs with higher proportions of BAME populations. Regions with higher ethnicity had higher deprivation scores, which helped explain lower vaccination uptakes Vaccination uptake could be used partially to predict Covid-19 deaths. In trying to explain the relationship between ethnicity and Covid-19 death further, multivariate analysis revealed that incorporation of health conditions, child poverty and age demographics didn’t better explain the Covid-19 deaths than just ethnicity.

Our investigation was limited by the lack of data available for real numbers of deaths, cases and vaccinations in each ethnic group. Whilst our approach of using proportions within a region was helpful, this data would have enabled a more robust analysis. Models for predictions could have been improved using better implementation of machine learning practices.

Overall, we found that there was disparity in Covid-19 deaths and cases in boroughs with higher ethnicity proportions and that this could be explained by vaccination uptake patterns.

1. **REFERENCES**

[1][Coronavirus Download data (Gov.uk) 2022](https://coronavirus.data.gov.uk/details/download)

[2] Coronavirus API Documentation (Gov.uk) 2022 https://coronavirus.data.gov.uk/details/developers-guide/main-api

[3] Mayor of London - [London Datastore](https://data.london.gov.uk/dataset/coronavirus--covid-19--cases)

[4] London Borough - [Covid-19 Deaths](https://data.london.gov.uk/dataset/coronavirus--covid-19--cases)

[5] London Borough - [Vaccines](https://data.london.gov.uk/dataset/coronavirus--covid-19--cases)

[6] [Office of National Statistics](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland)

[7] [Hidden-London](https://hidden-london.com/miscellany/london-boroughs-map/) - London Maps

[8] [Wikipedia](https://en.wikipedia.org/wiki/List_of_London_boroughs)  - List of London Boroughs

[9] [Wikipedia](https://en.wikipedia.org/wiki/List_of_sub-regions_used_in_the_London_Plan)  - List of London Regions

[10] <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/deaths>

[11] <https://www.fil.ion.ucl.ac.uk/spm/covid-19/forecasting/>

[12] <https://covid19.healthdata.org/united-kingdom?view=cumulative-deaths&tab=trend>

[13] [NHS England](https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/)

[14] [NHS England](https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/)

[15] Golestaneh et al., The association of race and COVID-19 mortality. (2020) https://www.sciencedirect.com/science/article/pii/S2589537020301991

[16] Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States (2020)

<https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2769369>

[17] <https://data.london.gov.uk/dataset/sub-regions-london-plan-consultation-2009>