Disproportionate Effect of COVID-19 on BAME Population in the UK

Marisa Dinko mjd4xs@virginia.edu
Isha Gangal ig4ga@virginia.edu
Pranitha Maddi pm3mg@virginia.edu
Zetao Wang zw3hk@virginia.edu

ABSTRACT

In this report, we will provide background information on the effects of the COVID-19 pandemic globally, before focusing on how machine learning skills can be used to contribute to our understanding of the virus and its effects. Specifically, this paper outlines the disproportionate impact of COVID-19 on BAME populations in the United Kingdom, while using multiple regressions to examine the relationship between this disproportionate impact and various confounding societal factors.

1. MOTIVATION

COVID-19 has affected the lives of people globally, but has disproportionately impacted BIPOC communities. In America, we have seen the intersection between issues of race, healthcare, and access to resources influencing the ways that COVID-19 affects populations. The prevalence of these issues has become even more public as the Black Lives Matter movement surged in the summer of 2020 following the deaths of Breonna Taylor and George Floyd. While the United States has its own specific history related to race and inequality, this is a phenomena not unique to our nation. As we begin to explore issues of COVID-19, we want to analyze the role that race and ethnicity play in experiences of health and safety globally as well -- COVID-19 is a universal issue that has different effects in each nation and this transnational effect is what we hope to explore with our project.

2. BACKGROUND

The UK has healthcare that is provided by the state to everyone. It is meant to protect and improve the nation’s health and wellbeing as well as reduce health inequalities. While this is a different system than what is seen in the United States, it is still a potential contributing factor to the impact and deaths resulting from COVID-19.

The data used in this paper is from the Respiratory Data Mart. The Second Generation Surveillance System was used for information about all the samples tested, and the results of these tests were from the public health, NHS, and private laboratories. Also, newly admitted hospital patients with COVID-19 were reported to the COVID-19 Hospitalization in England Surveillance system by acute NHS trusted secure web portals. Moreover, the Mortality rates were found using 3 sources: NHS England, Health Protection teams, and Demographic Batch Service.

The data collected focuses on raw numbers of COVID-19 related deaths, population distributions, ethnicities, key/furloughed workers, etc. While the data collected is valuable on its own, there is room for analysis into what this data means more broadly by using machine learning to examine the relationships between each factor. A further explanation of the gaps our work fills in the data collected will be included in future sections.

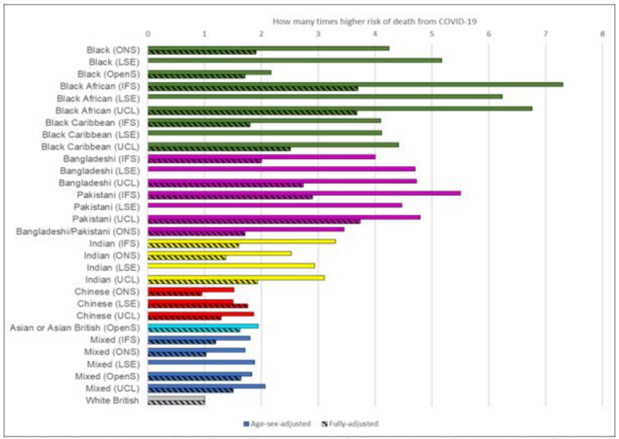
3. RELATED WORK

COVID-19 has had a disproportionate effect on ethnic minorities in multiple countries, including Brazil, France, the United States, and the United Kingdom (Bachelet). Numerous organizations have investigated this effect, trying to determine the factors at play. Examples of such organizations include the Pew Research Center, which investigated this effect in the United States; The Lancet, which has looked at the United States, United Kingdom, and Brazil; the National Health Service in the UK; and Public Health England in the UK (Baqui; “Beyond the data”; Douglas; Lopez). Each of these analyses identified multiple socio-economic factors whose effects may be compounded in minority communities, such as employment (as essential workers), overcrowded housing, chronic health issues, and access to healthcare. Even in the UK, where there is universal access to health care, such disparities remain.

4. CLAIM AND TARGET TASK

As seen in figures provided with our data, there is a trend that exists between ethnicity and excess deaths from COVID-19. However, there is an absence in the understanding of how other factors combined with ethnicity/race can be used to predict COVID-19 death rates. Moreover, just like we have explored in America, many societal structures impact this trend -- from disproportionate poverty, to redlining, to inadequate healthcare in the United States. To build upon this data, we will search and learn from previous research, and train and test our data with more optimic ways. We aim to map the relationship between ethnicity, excess deaths, population density and deprivation and how they impact the transmitting tendency. Ultimately, our target task is to predict COVID-19 death rates based on regression models that consider the aforementioned factors. We believe that this task will have social impact exemplifying and visualizing the relationship between social experiences and the current impacts of the COVID pandemic. A more detailed explanation of the implementation and impact of this target task will be in the following sections of this paper.

In order to emphasize the importance of this task, consider the following figure. The figure below comes from a report by the University of Bristol and the National Institute for Health Research on the impact of COVID-19 on minority communities. It depicts how many times higher the risk of different minority groups dying from COVID-19 is compared to white British patients. Our project aims to investigate some of the key risk factors that may be causing this effect to get a better understanding of compounding risk factors in the UK.



Notes: The figure is based on 5 primary studies[1,2,3,7,10]*, referred to as ONS*[1]*, LSE*[3]*, OpenS*[2]*, IFS*[10]*, and UCL*[7]*. ONS data was reported separately for males/females, we have produced a weighted average based on proportion of males/females in each ethnic group to simplify the figure. LSE data was reported as excess deaths above expected as % of expected, we have converted this to observed/expected to be in line with other results. LSE and UCL data have been normalised so that results can be compared to a value of '1' for white British to simplify the figure.*

5. PROPOSED SOLUTION

Since our data is tabular and supervised, we are using multiple regression models to examine the relationships between race and ethnicity, excess deaths, and other confounding variables. As mentioned above, this allows us to see what factors influence COVID-19 deaths the most in the United Kingdom. In addition, our ML model will be aided by supplementary data analysis, such as visualizations of the different factors we are investigating. Based on the timeline and the ethnicity distribution in the supervised data, we will map out the COVID-19 transmission tendency among the region and population density.

Moreover, we will create visualizations that reflect the correlation between rates of deprivation in each locality, and the amount of excess deaths per ethnicity. Through this regression model, we aim to represent the connections between living conditions prior to COVID-19, race and ethnicity, and the amount of excess deaths.

We want to see the effects on minority groups especially. Even though health care is provided by the state we want to see whether that truly levels the playing field for all United Kingdom citizens. These citizens and residents of the UK all range in their financial, and socioeconomic backgrounds making it even harder to understand and analyze which groups need the most help.

5.1 Contributions

A critical part of using machine learning to examine COVID-19 data is ensuring that we are contributing to progress in society through our work. Ultimately, through our visualizations, analysis, and generated models, we believe that our project makes two key contributions to the field: quantifies the impact of confounding variables (1) and provides the basis for a global conversation regarding race and inequality during a global health crisis (2).

The first contribution that our project will make is quantifying the impact of confounding variables on the resulting COVID-19 death rates in the United Kingdom. While the raw data shows the number of furloughed workers, the proportions of each race and ethnicity, and rates of deprivation, there is no model to show how these factors correlate with each region's COVID-19 death rates.

The second contribution that we believe our project makes is an expanded opportunity for a more globally minded dialogue surrounding inequality during a global health crisis. As aforementioned, the Black Lives Matter movement gained international coverage in the summer of 2020, and reflects a long history of marginalization and inequity in the United States. Directly connected to institutions of inequality is the disproportionate number of COVID-19 cases affecting Black and minority communities. Through our various models, we are providing an expanded lens for this conversation surrounding equity in global health. By understanding the disproportionate effects of COVID-19 on minority groups in the UK, there becomes space for a cross analysis of the similarities cross-nationally.

6. IMPLEMENTATION

In order to conduct our analysis, we created a number of visualizations representing our data as well as three regressions (linear, Lasso, and Ridge) that were run on two datasets with different feature dimensions. The rationale behind each component of our implementation is presented below, as are further details about them.

6.1 Technical Challenges

Our biggest challenge in implementing our model came from our data. Since the data were collected from a variety of sources, not all of it could be joined together into a single representative dataset. For example, while all of our data included some sort of location, the level of that location varied from city to region to country across our different datasheets.

Another challenge stemming from our data was that our data set was rather small, especially in terms of what a machine learning algorithm generally requires for effective training. Since our data focused on the United Kingdom and the early impacts of the pandemic, we ended up with a little over 300 samples in the two final datasets we used for model training. In our granular data set, each sample had 123 features representing the proportions of different ethnic groups, age brackets, and furloughed workers in each location. Given the high feature to sample ratio, this made it very difficult to create a test set that was completely left out of training.

6.2 Technical Solutions

To address our first challenge, we implemented two strategies. First, we created a wide range of visualizations to better understand the data and reveal important correlations between individual factors. For example, the data measuring the death rate across different minority groups solely existed at the aggregate group level and was therefore not suitable to be input into a model. Yet, it is exactly that kind of information that is necessary to understanding the disproportionate effect of COVID-19 on minority groups. Visualizations were created comparing those features, as well as many other pairings of features; for example, percent of population belonging to a minority group vs. percent of furloughed workers, and death rates within minority groups separated by age bracket and gender. This relates to our first contribution because collectively, the many visualizations we created help display the confounding factors affecting the impact of COVID-19 that would otherwise remain hidden. Furthermore, it also connects to our second contribution because visualizations are incredibly portable and usually very quick and intuitive to understand. While the general population may not understand the complexities of a state-of-the-art machine learning model, they can grasp the meaning of a well-made visualization at a glance. Each figure offers another starting point for conversation about the disproportionate effects of COVID-19, which can then be translated into actions and policy to help address them.

Our second strategy to address the challenge of disparate datasets was to simply create two different broad datasets upon which to run our regression. While both feature locations as the key, the first dataset was a granular look with 123 features representing the proportion of the population belonging to a number of minority groups, the number of people at each age from 0 to 90+ as well as normalized age brackets, and the percent of furloughed workers. The second dataset was a summary dataset that collapsed the different ethnic groups into a single BAME percentage, percent furloughed, and percent of the population that are key workers. For both datasets, the target variable was the COVID-19 death rate per 100,000 people.

Finally, to address the second challenge of having a limited dataset not suited for a full train-validate-test split, we relied on K-Fold Cross Validation training. We believe this was a rational choice given that we're not attempting to predict current COVID-19 death rates so much as understand which regions are being the most affected and how that relates to the underlying population demographics of the area. We relied on 5-Fold Cross Validation, presenting the average validation accuracy for each model. Moreover, this location-based approach once again connects to our second contribution in trying to understand how across locations (whether cities or regions), globally minorities groups have been disproportionately affected by COVID-19.

6.3 Technical Novelty

The novelty of our implementation stems largely from the broad approach we took in creating visualizations and our model. Given the challenges presented above, we created a wide range of different visualization types designed to present the data in many different ways. These visualizations included bar charts, pie charts, line graphs, and scatter plots with lines of best fit. We also used state-of-the-art machine learning packages (sklearn) to run multiple regressions to allow for a comparison of model success based on different types of regularization/feature extraction. Because we have contributed visualizations that reflect correlations and relationships that were not previously visible in our data, we have also provided a means for a more informed global conversation surrounding the impacts of COVID-19.

6.4 Method Variations

In conducting our analysis, we relied on three different types of regressions. Using the sklearn libraries, we created a multiple linear regression, a Lasso regression, and a Ridge regression. The linear regression had no hyperparameters to tune, but for the Lasso and Ridge regressions, we tested a range of alpha values (0.1, 0.25, 0.5, 1) and then returned the cross validation accuracy of the best model found during training. Each of the three models was run on both datasets.

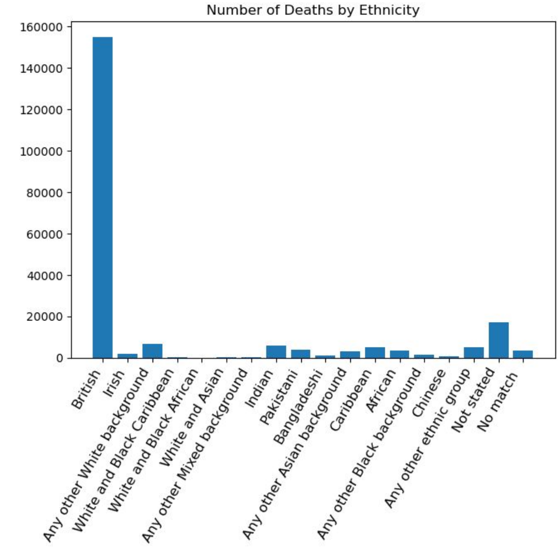
We also performed additional feature extraction by conducting a Principal Components Analysis (PCA). The PCA was conducted solely on the granular dataset, which has 123 features, and not the summary dataset which already has only 3 features. Here, again we tested several values for the number of principal components (2, 5, 10, 25, 50, 75, 100, 123). The transformed data was input into each of the three regression models, creating a graph showing how accuracy varied with the number and returning the cross validation accuracy of the best model. All results are presented below in Section 8: Experimental Data.

7. DATA SUMMARY

There are multiple sheets of data used in our project that all contributed to the models and visualizations we created. In this section, we will outline what data we had access to in our work.

The first datasheet we had access to contained information on weekly hospital deaths in the UK from April 2020 to June 2020 for various ethnicities. The data provided raw numbers mapped to an ethnicity and the corresponding ethnic group (ex. White, Irish, 161 deaths).

Figure 1: Number of Deaths by Ethnicity



Then, we accessed the UK population data which is from the UK 2011 population investigation. Compared with our death cases from April 2020 to June 2020, We determined the death ratio in different ethnicity groups. Also, we did a parallel comparison between different ethnicity groups' death rates.

Figure 2: Population Distribution in the UK

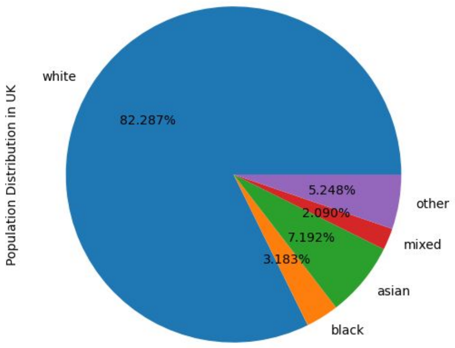
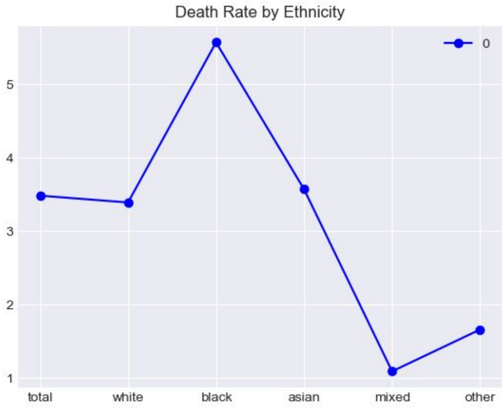
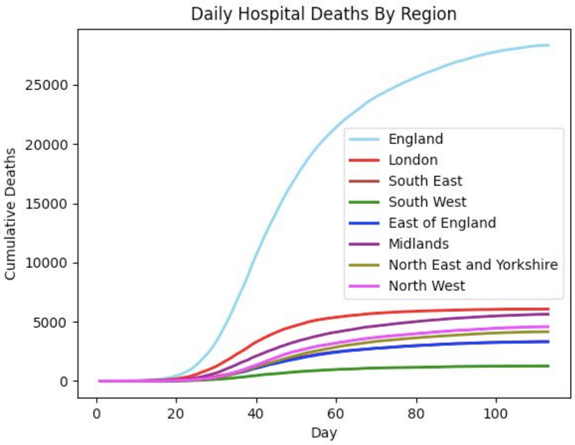


Figure 3: Death Rate by Ethnicity (Per 1,000 People)



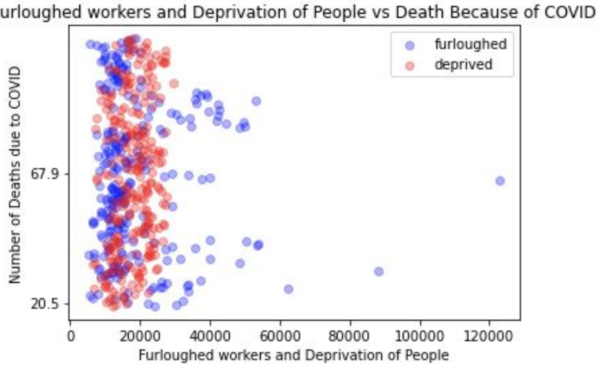
Another data sheet we used provided information regarding the daily hospital deaths by region in the United Kingdom from March 1, 2020, to June 21, 2020. This data grouped multiple towns into larger regions, so provided a broader image of what areas of the UK were facing the most hospital deaths. The following figure is one that we generated to visualize and better understand this data.

Figure 4: Daily Hospital Deaths by Region



Another integral datasheet that was used in our project outlined the proportion of COVID-19 and non-COVID-19 related deaths to ethnicity, sex, and a corresponding age group. This data allowed for more specific features to be used in creating our model since we could categorize COVID-19 related deaths as they related to more than just race or locality. (See Appendix A for graphs)

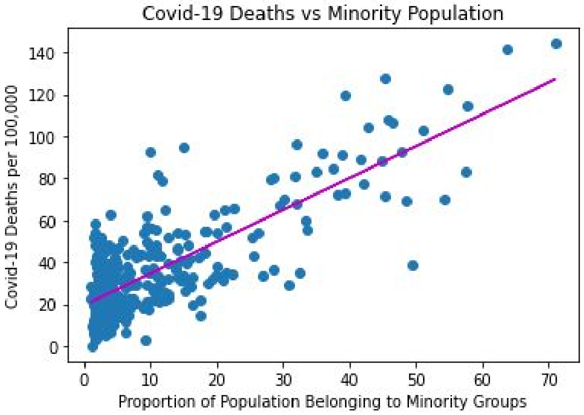
Figure 5: Furloughed Workers and Deprivation vs COVID-19 Rates



Beyond the aforementioned datasets, we also had access to datasheets that gave raw numbers of furloughed workers and key workers in towns/regions throughout the UK. Similarly, we had access to data that gave rates of deprivation in the regions as well. As shown in the graph above there is no real relationship between the furloughed people and COVID death rate. This is also the same with deprivation and its relationship with deaths due to COVID.

In addition, we had data representing the ethnic breakdown of locations across the UK that was also combined with the death rate to create a high level view of the effect we are trying to investigate. The line of best fit in the graph below had an R^2 value of 0.625, meaning over 60% of the variation in COVID-19 death rate could be explained just by looking at the minority population proportion in each location.

Figure 6: COVID-19 Deaths vs Minority Population



Through the multiple datasets we had access to, we created our model that considered how each factor impacted and correlated to the rates of COVID-19 related deaths in a region. See the Appendices for more visualizations of the data. The results of working with this data is included in the following sections of this paper.

Overall, we had real world data sets that covered information related to factors that similarly affect COVID-19 data in the United States. As previously mentioned in our contributions section, we believe that our project sparks a conversation on the global impact of COVID and perpetuating inequalities. In having this conversation, it is important to consider factors like furloughed workers or regional differences that affect both the US and the UK. A proper international analysis would not be possible without considering mutually influential factors. Moreover, because our data provided raw information on a variety of factors, we are able to contribute new understandings of the different corollary relationships between each feature.

8. EXPERIMENTAL RESULTS

In addition to the visualizations described above, we also conducted multiple analyses using different regression models and feature representations. The 5-fold cross validation accuracy of each model is presented in the table below. The loss function for all regressions was ordinary least squares with the Lasso regression adding L1 regularization and the Ridge regression adding L2 regularization. While the visualizations largely aimed to address the second contribution, the regressions aimed to address the first by attempting to use demographic data to predict death rates in each location.

Table 1: Model Accuracy Given Feature Representation

	Feature Representation		
Regression	**Granular**	**PCA**	**Summary**
Linear	0.563559219	0.707357881	0.599653702
Lasso	0.716098470	0.705027707	0.602445155
Ridge	0.734978015	0.749603955	0.596756986

The "baseline" model was the granular linear regression which had an accuracy of 0.56. Adding regularization using the Lasso and Ridge regressions resulted in accuracies of 0.72 and 0.73, respectively. The Principal Component Analysis regressions had the following accuracies: linear = 0.71, Lasso = 0.71, Ridge = 0.75. Finally, the summary dataset resulted in a linear regression accuracy of 0.60, a Lasso regression accuracy of 0.60, and a Ridge regression accuracy of 0.60.

For the Lasso and Ridge regressions, training the model also involved selecting the best value for one hyperparameter, alpha. Those values are presented in the following table.

Table 2: Best Alpha Values for Each Model

	Granular	**Summary**
Lasso	0.1	0.1
Ridge	0.5	0.25

The following three figures demonstrate how the cross validation accuracy varied for each of the three regression models based on the number of principal components (PC). The linear model peaked at PC=25, and then had progressively worse accuracy. However, the Lasso and Ridge regressions had a local max at PC=25, decreased as the principal components increased, and then began to increase again for another local max at PC=123 (i.e. the full number of features).

Figure 7: Linear Regression CV Accuracy vs PC Number

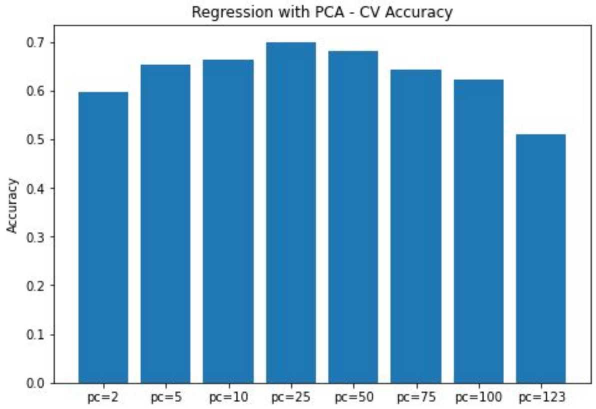


Figure 8: Lasso Regression CV Accuracy vs PC Number

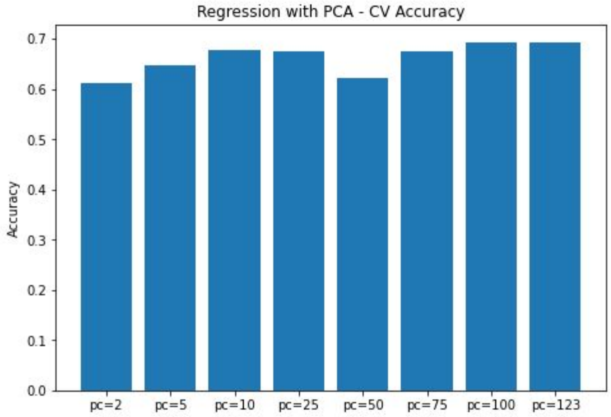
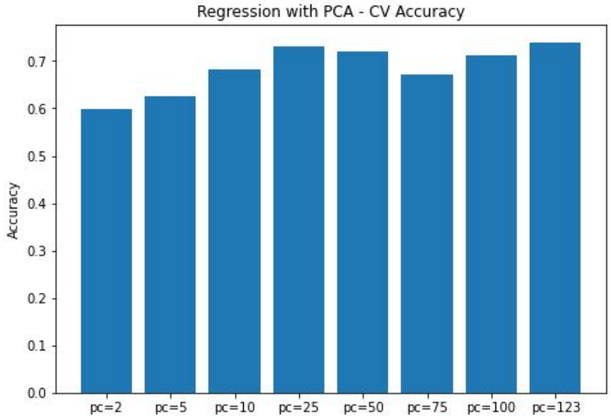


Figure 9: Ridge Regression CV Accuracy vs PC Number



9. EXPERIMENTAL ANALYSIS

By doing three regressions on multiple feature representations, we were able to make a multiway comparison across models and feature selection methods to develop our best model.

As previously stated, the baseline granular dataset linear regression model performed the worst of all the models, with a cross validation accuracy of 0.56. Introducing regularization by penalizing the L1 and L2 norms through the Lasso and Ridge regressions improved the model's performance significantly. This makes a lot of sense because the granular dataset had a lot of features relative to the sample size. Moreover, the granular dataset included a lot of highly correlated and collinear data. For example, it included both raw numbers representing the number of people at each age as well as percentages representing the proportion of the population belonging to a specific age bracket. Such regularization was likely necessary to prevent overfitting and improve the accuracy from that of the baseline linear model.

Similar to the regularization, the PCA also added some necessary feature selection, greatly improving the performance of the linear model. However, it didn't have a similar effect on the Lasso or Ridge regressions; the Lasso regression performed slightly worse than on the plain granular data and the Ridge regression performed slightly better, but neither saw the large change the linear model did. All models, with just two principal components saw about a 0.60 accuracy. This makes sense 0.625 R squared value from the simple linear regression of minority population vs death rate presented above. Looking at Figure 6, the linear model peaked at PC=25 and then began to decrease again, likely due to the increasing collinearity of the data. Interestingly, the Lasso and Ridge regressions had a local peak at PC=25, but also began to increase as the number of PCs approached the full number of features=123. This is likely due to a tradeoff between the feature selection of the regularization and the feature selection of the PCA.

Finally, the summary data performed slightly better than the baseline linear model, with accuracies around 0.60 for all three regression models, however this was much worse than any of the results using the granular data with regularization. The Lasso and Ridge regressions provided virtually no benefit on the summary data because it already only had 3 features represented. Furthermore, it's difficult to determine what exactly caused the slight increase in performance compared to the granular linear regression. It is likely that the introduction of the new variable, percent of the population that serves as key workers, provided new information that boosted performance. However, it's also possible that the locations that the data was grouped on varied enough from those in the granular data to cause the increased performance. Without far greater knowledge of the geography of the UK and the ability to combine these disparate datasets together, we cannot truly know what the effect the key workers feature had.

Regardless, the high performance of the granular models with regularization/feature selection showed that looking at many demographic features in conjunction is very important. Simply having the ethnicity breakdown of an area allowed us to predict the COVID-19 death rate in an area with relatively high accuracy. Much of this data is either relatively constant or already collected and reported by local authorities on an ongoing basis. By conducting such an analysis, we have a straightforward way to determine which areas are at a relatively high risk of getting/suffering high death rates from COVID-19. This type of analysis is very important because as a COVID-19 vaccine becomes a real possibility, there will be a lot of questions about how it should be distributed. Of course, essential workers and the elderly (groups of people who are generally known to be at high risk) should be the first to get the vaccine. However, beyond that, it's more of an open question of what the best distribution method is. Knowing which areas are more susceptible to the disease may be greatly beneficial to making that decision.

10. CONCLUSION

Our project aimed to investigate the disproportionate effect of COVID-19 on the BAME population in the United Kingdom. Through our analysis, we hoped to 1) determine the impact of confounding variables that may be present, and 2) provide the basis for a conversation about race and inequality on a global scale. Our hope is that the many visualizations we created as part of this project in conjunction with our regression analysis serves to further emphasize just how unequal a global health crisis can be, even in a country with universal health care. The benefits of understanding the interplay of ethnicity and risk are great, and the consequences of not have already shown to have a drastic impact.

Future improvements to our model would include taking a deeper look at the model results itself, adding greater complexity, and expanding our dataset. First, at the moment each of our models is like a black box. While we have the final cross validation accuracy scores, we don't have the coefficients or p-values of any of our features. One of the major benefits of a regression is they are reasonably easy to understand, a higher coefficient would mean a certain feature has a larger impact on the death rate (since all of our data were standardized, this is an acceptable assumption to make). While our visualizations aim to fill this gap, looking at these values would allow us to better investigate how different minority groups are being affected by COVID-19. In addition, we would expand upon the model itself. The regressions we implemented were all single layer models. Trying more complex models, such as an MLP, in the future may allow us to further improve our accuracy when predicting death rates by locality.

Finally, as we have stated throughout this report, our dataset was rather small. In the future, we would like to expand our data set to not only include more features but also more time series data. This would enable us to get a better understanding of what confounding features are causing such a devastating effect on minority groups as well as whether/how the effect of COVID-19 on our groups of interest is changing over time. In addition, we would like to expand our dataset to include more countries to better study this effect on a global scale.

WORKS CITED

Bachelet, Michelle Jeria. "Disproportionate impact of COVID-18 on racial and ethnic minorities needs to be urgently addressed." *Office of the United Nations High Commissioner.* 2 June 2020. <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25916>.

Baqui, Pedro, Ioana Bica, Valeria Marra, Ari Ercole, and Mihaela van der Schaar. "Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study." *The Lancet.* 2 July 2020. <https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30285-0/fulltext>.

"Beyond the data: Understanding the impact of COVID-19 on BAME groups." *Public Health England.* June 2020. <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/892376/COVID_stakeholder_engagement_synthesis_beyond_the_data.pdf>.

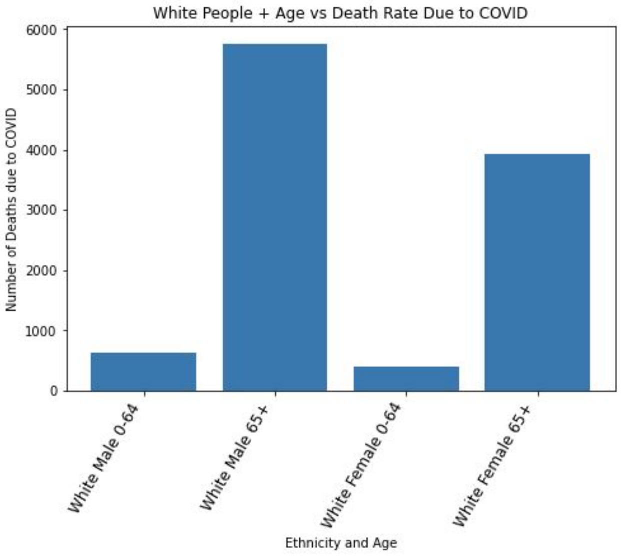
Douglas, Jason. "Research Lays Bare Covid-19's Outsize Impact on Ethnic Minorities." *The Wall Street Journal.* 30 June 2020. <https://www.wsj.com/articles/as-in-u-s-covid-19-risk-in-britain-is-higher-for-minority-groups-11593431777>.

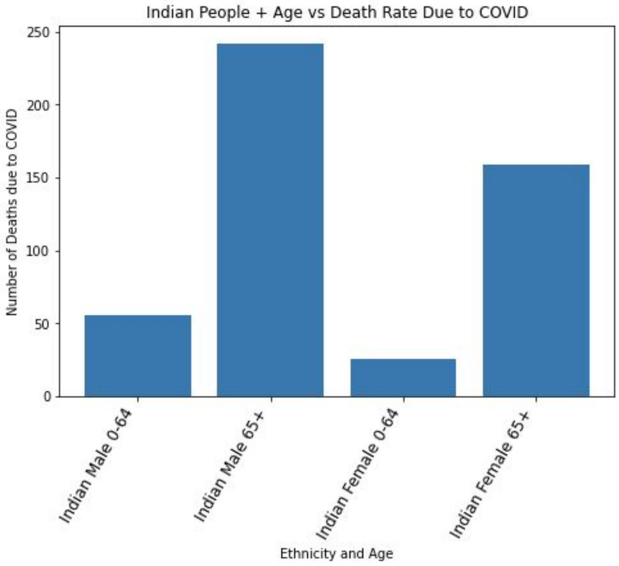
Lopez, Mark Hugo, Lee Rainie, and Abby Budiman. "Financial and health impacts of COVID-19 vary widely by race and ethnicity." *Pew Research Center.* 5 May 2020. <https://www.pewresearch.org/fact-tank/2020/05/05/financial-and-health-impacts-of-covid-19-vary-widely-by-race-and-ethnicity/>.

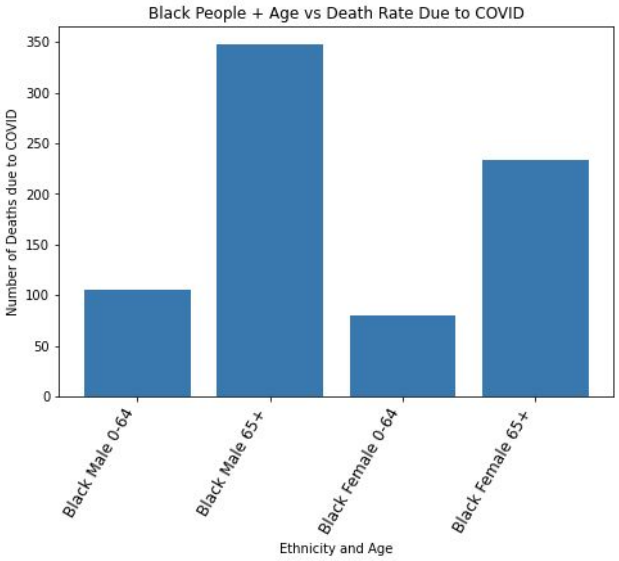
Mamluk, Loubaba and Tim Jones. "The impact of COVID-19 on Black, Asian, and minority ethnic communities." *National Institute for Health Research* . 20 May 2020. <https://arc-w.nihr.ac.uk/Wordpress/wp-content/uploads/2020/05/COVID-19-Partner-report-BAME-communities-BCC001.pdf>.

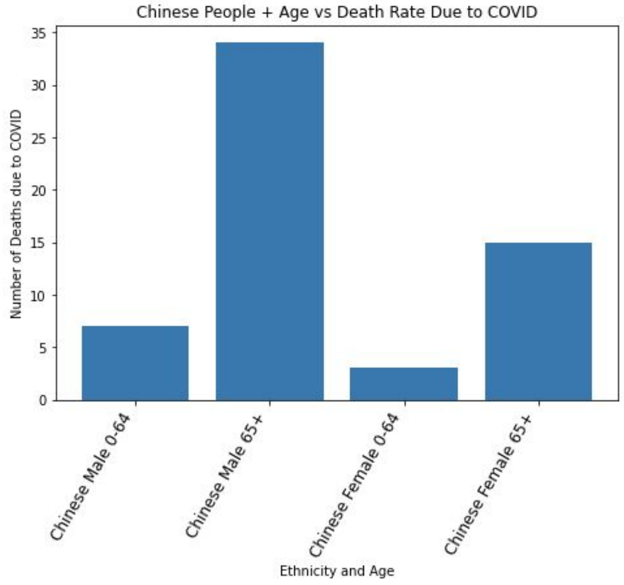
APPENDICES

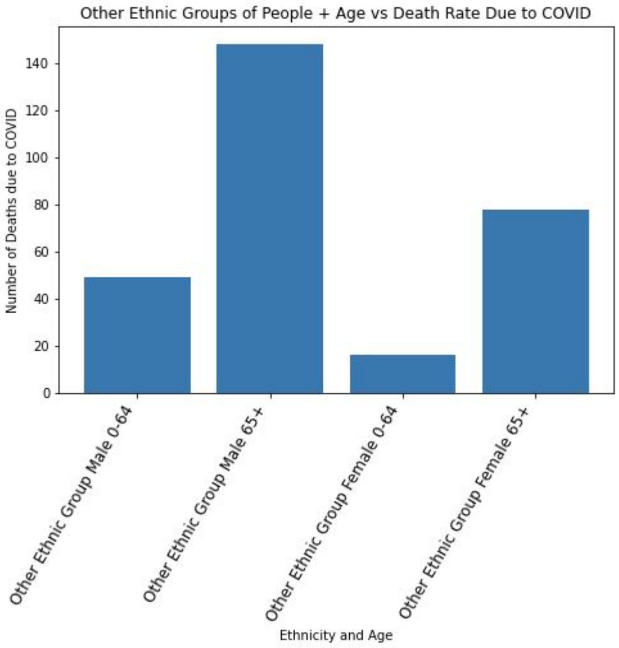
APPENDIX A: *Death Rate by Ethnicity & Age*

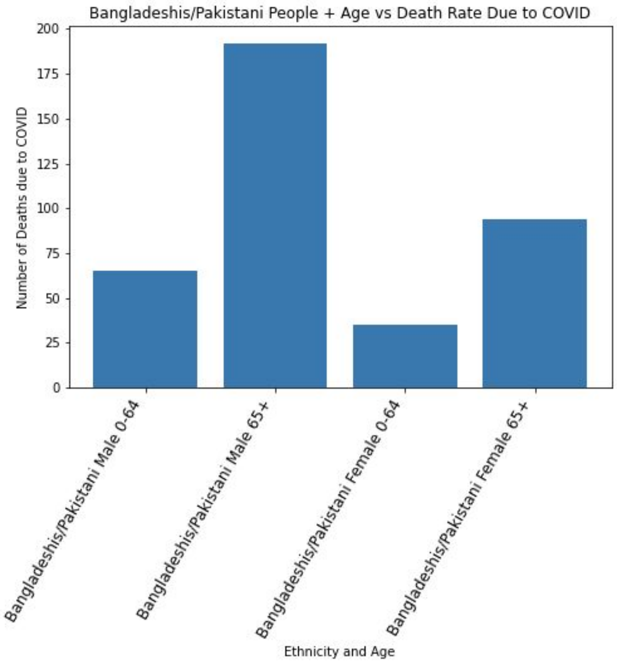


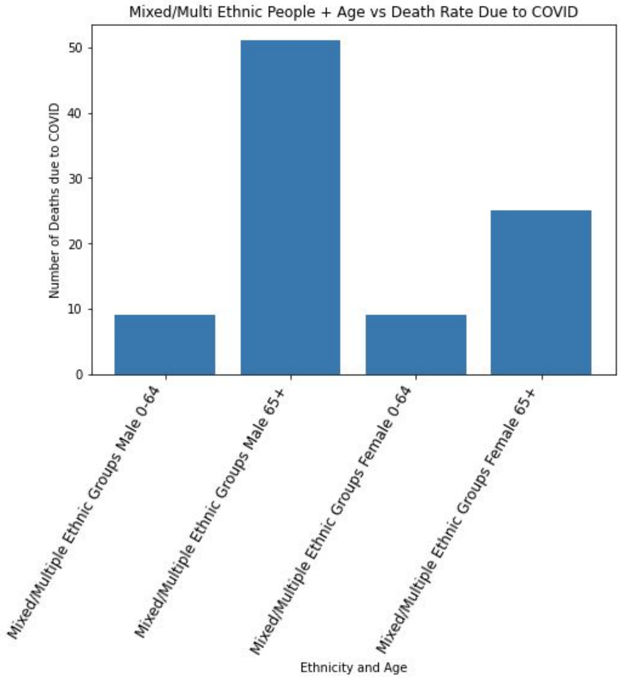






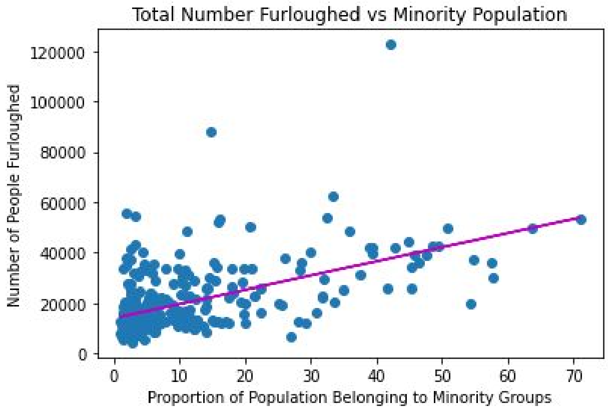

Other Ethnic Groups of People + Age vs Death Rate Due to COVID


Bangladeshis/Pakistani People + Age vs Death Rate Due to COVID


Mixed/Multi Ethnic People + Age vs Death Rate Due to COVID

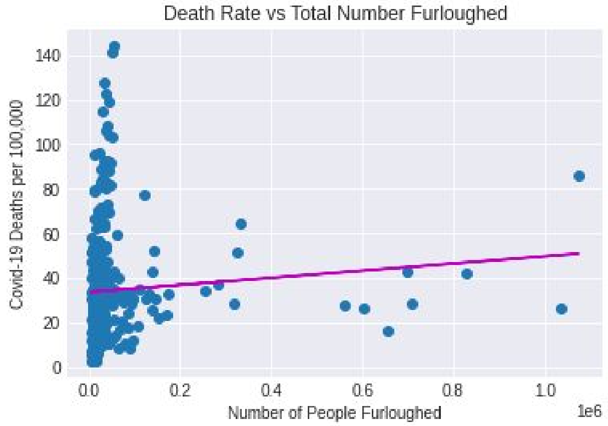
APPENDIX B

Figure B.1: Total Number Furloughed vs Minority Population



Regression R^2 =0.3053684320521

Figure B.2: Total Number Furloughed vs Death Rate



Regression R^2 =0.006325630888414313