**Title: Predicting COVID-19 Deaths by Race and Income**

Name: Bon Trinh

The COVID-19 pandemic is currently the world's biggest concern since it is not only the cause of millions of deaths but also obstructs global activities and severely affects the economy. People with a not strong financial condition are suffering because they cannot afford their living without working for too long. Furthermore, there are recent racial incidents such as racism towards the Asian community and violence towards the Black community, which accidentally creates more gathering event and increase the spread of the virus. Therefore, my project aims to create a predictor of the number of COVID-19 deaths based on the racial distribution and income of areas. This can help the authorities foresee the potential disease outbreaks in areas so that the right people and place are provided with in time medical aids or other methods of prevention.

I have combined three datasets on the [US CDC website](https://www.cdc.gov/) and [US Census Bureau website](https://www.census.gov/en.html). The first one is [COVID-19 Death Counts by County and Race](https://data.cdc.gov/api/views/k8wy-p9cg/rows.csv?accessType=DOWNLOAD) which provides information about deaths by each race in each county in the U.S. (in both number counts and percentage of the county population). This data has 21 columns, but I only use the columns for the state, county names, and the columns for the race (Hispanic, White, etc.) which indicate the deaths by COVID-19 in that race. The second one is the [County Median Income](https://www.census.gov/data/datasets/2019/demo/saipe/2019-state-and-county.html) dataset containing data on the median income for each county. Finally, I explored the [County Population by Characteristics](https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html) dataset which gives details about the race distribution in different age groups. It has more than 60 columns and 700,000 rows; hence, it cannot be read into memory at once. I read and processed it in small chunks one by one, filter the data for the most recent year only (2019), and aggregate data for all age groups in each county to get the figures for the whole county. After preprocessing all three datasets to get the necessary data, I joined them by state and county name.

In Figure 1, I visualize the data on deaths by COVID-19 with median income, colored by racial groups. From the graph, most deaths are White people (more than 80% in almost all income groups) as in most states, the majority are White. However, the percentage of Black people who died by COVID-19 in the lower-income group is relatively higher than other groups except for the White, but the percentage tends to be lower in the higher income group (above 80,000). The reason might be that there are fewer Black people in the high-income group or their probability of deaths by COVID-19 lowers as they earn more. In contrast, the percentage of deaths by Latino tends to increase as we move to the higher income group. This suggests that there is a correlation between deaths, race, and income.

In Figure 2, as a step of preprocessing, I perform PCA on the combined dataset which has 8 columns: 6 columns for each race (White, Black, Native Indian, Asian, Hispanic, and Native Hawaiian), a column for median income, and one for deaths by COVID-19. From Figure 2, it is reasonable to reduce the dimension with PCA as the reduced data with 3 components can still capture 99% of the original data. Figure 2 is plotted on data scaled with the StandardScaler and since there is no difference between the explained variance when performing PCA on scaled and non-scaled data, I only use the reduced data in this plot.

Figure 3 illustrates the coefficients for each feature of the Linear Regression model in my Pipeline. The sklearn Pipeline has three parts: (1) a StandardScaler, (2) a PCA 3 components, (3) a PolynomialFeatures with degree 2, and (4) a LinearRegression model. The explained variance scores on this model on the training and test data are 0.92 and 0.86 respectively. With PolynomialFeatures degree of 1, the scores drop substantially to 0.86 and 0.86 while increasing it to 3 only increase the figures by 0.02 for each set, and making the degree 4 will overfit and make the score on test data 0.46. Therefore, degree 2 is the optimal number. In Figure 3, regarding the coefficients, component 1 is weighted largely positive and in this component, the explained variance ratio of the percentage of White people is the largest while that of income is very low (only 0.13). On the other hand, component 2 and 3 are weighted negatively to the number of predicted deaths and the highest explained variance ratio in those two are that of median income (0.74 and 0.62 respectively). The formula of the linear regression is:

In conclusion, there seems to be a correlation between races, income, and deaths by COVID-19, especially the most important PCA component suggests that the area with a high proportion of White people and low income are more vulnerable. Meanwhile, the higher-income areas are less likely to have many deaths.

Scatter chart

Description automatically generated

**Figure 1**

Shape

Description automatically generated

**Figure 2**

Chart, waterfall chart

Description automatically generated

**Figure 3**