DS-1001 Capstone Project

Group Name: Principal Components

Group members: Ameya Shere (as12366), Shuvadeep Saha (ss15592)

Investigating factors relating to COVID-19 cases and deaths

Introduction

Ever since it began in December 2019, the COVID-19 pandemic has caused widespread illness and death, and has wrought significant damage to many industries worldwide, affecting the global economy. However, much remains unknown about the various factors facilitating such rapid spread of the SARS-CoV-2 virus, and thus the pandemic is a subject of ongoing research. As a contribution to that research, we aim in this report to use hypothesis testing, multiple regression, and machine learning methods to gain insight into the characteristics of U.S. states that affect their total counts of COVID-19 cases and deaths.

The dataset we use is sourced from the paper, "Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking," by Satyaki Roy and Preetam Ghosh. In this paper, the authors examine numerous possible contributing factors to COVID-19 cases and deaths in the United States. Combining data from a number of public sources, Roy and Ghosh create a dataset summarizing the data for these factors across all fifty states [1]. We access this data from the file "Main.xlsx" in the GitHub repository included along with the paper, and we convert this file to ".csv" format for easier use with the numpy, pandas, scipy, and scikit-learn Python libraries we use in our analysis [2]. The 50 rows of the dataset each represent one U.S. state, and the 127 columns include possible factors affecting COVID-19 cases and deaths, as well as the counts of those cases and deaths themselves.

Our interest in this dataset is motivated by its feature-rich nature. Since it includes many different possible contributing factors to COVID-19 across different areas, like demography, socioeconomics, environment, and health, it offers many options for us to explore as we attempt to answer our research questions. In addition, this dataset is different from many other COVID-19 related datasets that we have encountered throughout our research process in that it contains what we assume to be independent samples. In other COVID-19 datasets, the focus is on trends with respect to the passage of time, so much of the data is time series, and thus the rows are not necessarily independent of each other. Since our interest is in analyzing independent samples, this dataset is a good fit for our project in that sense as well.

Question 1

We begin by exploring the effect one of the factors from the dataset has on COVID-19 cases and deaths. Our first question is the following: do densely populated states have more positive cases or deaths than sparsely populated states?

We conduct one hypothesis test each for the cases and for the deaths, assuming a significance level of $\alpha = 0.05$ for both. We want to compare whether the distribution of test statistics for the densely populated states is stochastically greater than the distribution for the sparsely populated states. We also want to be cautious about outliers in the data, so we would prefer to compare the medians of samples instead of the means. Finally,

we are not sure about the distribution of the samples, so we seek a test that does not assume normality. For all these reasons, we choose a Mann-Whitney U test.

Test 1: COVID-19 cases

Let us start by stating the hypotheses:

$$H_0 := \text{Densely populated areas do not have more positive cases than less-dense areas}$$
$$H_1 := \text{Densely populated areas have more positive cases than less-dense areas}$$

Performing the U-test, we find a p-value of $p = 0.0327 < \alpha$. Thus, we reject the null hypothesis in favor of the alternative hypothesis, stating that there is a statistically significant difference in the positive COVID-19 cases depending on the population density, and that more densely populated areas have a higher number of cases.

Test 2: COVID-19 deaths

We again state the hypotheses:

$$H_0 := \text{Densely populated areas do not have more deaths than less-dense areas}$$
$$H_1 := \text{Densely populated areas have more deaths than less-dense areas}$$

After performing the test, we obtain a p-value of $p = 0.013 < \alpha$. Thus, we reject the null hypothesis in favor of the alternative hypothesis, stating there is a statistically significant difference in deaths depending on the population density, and that, as before, the more densely populated states have a higher number of deaths.

The results of the hypothesis tests indicate that for both COVID-19 cases and deaths, there is significant statistical evidence that population density has an effect. Intuitively, this conclusion makes sense given that the more tightly packed people are into one area, the more easily the virus is able to spread among the population.

Question 2

After investigating one factor affecting COVID-19 cases and deaths, we broaden our research to include multiple factors. Our second research question is as follows: from a set of selected covariates, which factors have the most effect on COVID-19 cases and deaths?

To answer this question, we begin by selecting certain potential predictors of cases and deaths from the dataset: GDP, population density, health score, busy-airport score, and days under lockdown. The reason for choosing these specific factors is that, looking at the correlation matrix for the dataset, these potential predictors show the most promise for correlation with cases and deaths. To further investigate these factors' effects, we normalize the data and then fit two multiple regression models, one for cases and one for deaths. For the models, we use L_2-regularized linear regression, also known as ridge regression, and we loop over different regularization parameter values to determine the parameter that provides the best tradeoff between bias and variance in the model. To fit these models, we perform an 80:20 train-test split and evaluate model performance using mean-squared error. We obtain the following results. For both models, the optimal regularization parameter is 1. Both the cases and the deaths models fit the data well, having mean-squared errors of 0.44 and 0.35, respectively. The results for the coefficients of each model can be found in Table 1.

Coefficient values for ridge regression models

	Airport activity score	Population density	GDP	Days under lockdown	Health score
Cases model	0.498	0.436	0.151	-0.145	0.001
Deaths model	0.541	0.572	-0.095	-0.137	0.003

Table 1: Coefficient values for each selected covariate for the cases and deaths models, rounded to three decimal places. Note that a higher coefficient value in absolute value indicates a stronger contribution to the output.

The results from the first model indicate that among the selected covariates, the most significant predictors of COVID-19 cases are both the busy-airport score and the population density, with nearly equal share. This makes sense because if a region is very densely populated, then the transmission of disease will be higher, and the same will be true for a high airport population flux. Indeed, airport influx/outflux has had such an effect throughout the pandemic on virus transmission that it has motivated many border closures in different countries around the world. In addition to the above, another related factor is the GDP, which is also positively correlated with COVID-19 cases. One explanation for this result could be that a higher GDP indicates a large number of people employed in a region, and thus a higher activity level in that region, leading to a higher risk of disease transmission. We also find that the days-under-lockdown factor is negatively correlated with the number of cases, as more people staying at home will negatively affect the disease transmission. Lastly, the health index of a region, which is a measure of how well-prepared it is to deal with health-related events, in terms of availability of hospitals, doctors, nurses, ICUs, beds, etc., has very little or no impact on the number of positive cases.

In the second model, the leading predictors are again population density and busy-airport scores, which are positively correlated with the number of deaths from COVID-19. As before, a possible explanation for this result is that the more densely populated an area is, the more a disease is able to spread and, additionally, the less available healthcare facilities are to patients with severe cases. The days-under-lockdown again has a negative impact on the number of deaths, most likely for a similar reason as in the first model. Unlike the first model, the GDP has a negative, albeit weak, impact on the number of deaths. One explanation for this effect could be that poorer countries lack proper health-care facilities, intensive care units, or ventilators, leading to huge numbers of deaths. Lastly, the healthcare index has a positive correlation with the number of deaths. This is counterintuitive, because one would assume that countries better equipped to handle healthcare emergencies would have less deaths. It is likely that this result is just due to some noise in the model, especially since the value of the coefficient in absolute value is low, indicating a weak contribution to the output variable.

From these analyses, we conclude that the factors that have the most effect on both the cases and deaths are the population density and the busy-airport score.

Question 3

After observing these results from the regression analysis, we are interested in exploring if certain states can be clustered based on characteristics they have in common. We consider the same factors as in Question 2, in addition to the total number of COVID-19 cases and deaths for each state. Our third and final research

question is as follows: can the states be grouped into distinct clusters based on population, health and epidemiological characteristics?

To cluster the states, we begin by normalizing the data and then performing dimensionality reduction on these seven features using principal component analysis (PCA). According to the Kaiser criterion, only two out of the seven principal components represent a large proportion of the variance of the dataset, so we perform clustering on these two components. The scree plot for the PCA can be seen in Figure 2.

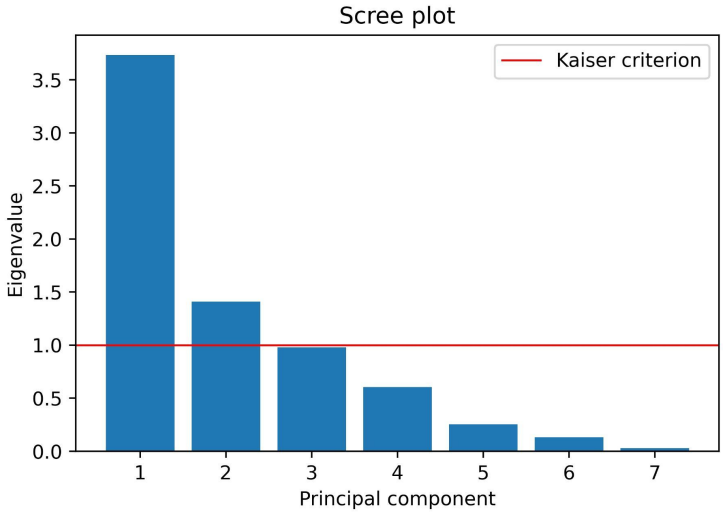

Figure 2: Scree plot for principal component analysis using Kaiser criterion

Inspecting the loadings of these principal components, we interpret them as follows. The first component points away from GDP, COVID-19 cases and deaths, and airport activity, so we interpret this component as representing low cases/deaths, low activity, and low GDP. The second component points away from health score and population density, so we interpret it as representing unhealthiness and sparse population. The complete loadings for each component can be seen in Table 3.

Approximate loadings for top two principal components

	GDP	Health Score	COVID-19 cases	COVID-19 deaths	Lockdown length	Population Density	Airport activity
Component 1	**-0.426**	0.118	**-0.499**	**-0.463**	-0.151	-0.336	**-0.454**
Component 2	0.313	**-0.678**	-0.119	-0.274	0.143	**-0.510**	0.270

Table 3: Loadings for the top two principal components, rounded to three decimal places

We attempt to cluster the data with respect to these two components using the k-means algorithm. To determine the number of clusters, we use the silhouette method, which results in an optimal cluster number of 2.

Using k-means, we group the data into two clusters, which have cluster centers that are well-separated. The plot of the clusters can be seen in Figure 4.

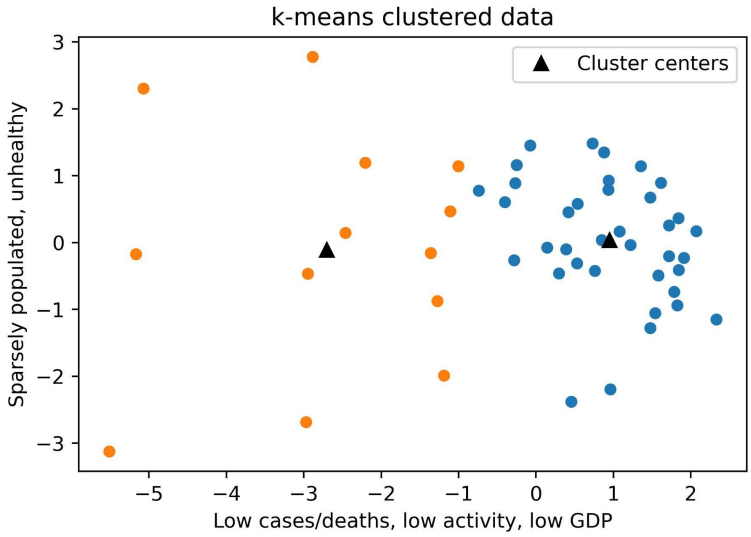

Figure 4: Data for the top two principal components grouped into two clusters

From this result, we conclude that with respect to these seven features, the states can be separated into two groups. Since the separation of cluster centers is along the first component's axis, the difference between these two clusters is that one has either lower COVID-19 cases and deaths, lower airport activity, lower GDP, or some combination of these, whereas the same properties for the other cluster are higher.

Conclusion

After performing these three analyses, we can conclude that overall, there are many factors that influence COVID-19 cases and deaths for each state in the U.S., including, but not limited to, population density, airport activity, and GDP. Indeed, the results from Questions 1 and 2 align with general intuition about how these factors would facilitate or limit the spread of disease or prevalence of death. We can also conclude that, based on the characteristics examined in Question 3, the states in the dataset can be divided into two groups with respect to the principal components that account for the most variance in the dataset.

Although we are able to obtain interesting and intuitive results for these questions, our research does have some limitations. The most obvious drawback of this project is that the dataset used only has 50 samples, but it has more than twice as many features. This limits the amount of analysis that can be performed, because using all the features would require experiments in the high-dimensional statistical regime, where results may not be representative of the actual effects occurring in the given states. To avoid this problem, we choose only a small subset of these features for use in our research, but even so, 50 is a very small sample size, and so the statistical power of the tests we used are quite low, and our results may be wrong.

Another limitation is that for some of the "score" features, like the health score, the calculation of the scores is done by a third party, and thus is somewhat opaque and difficult to soundly interpret. For other "score" features, like the infected and death scores, the data has been rounded to discrete values, and is thus more

ordinal than numerical. For this reason, we avoid using those features, which has the effect of decreasing the richness of our analysis.

In fact, after data exploration, it becomes clear that a lot of the features are not useful at all, and that the dataset is of a much lower dimension than we had originally thought. There are many categorical variables that are not useful to us, and a lot of the columns are essentially repeats of others, like the normalized columns. If we could repeat this experiment with better data, we would choose data with richer covariates and more samples, so that we could perform regression and clustering with more features, and perform hypothesis testing with high statistical power. If this were the case, our regression model probably would have had a much better fit, and we might have also found more principal components and more defined clusters. We would encourage future research in this area to seek better data sources. It would be interesting to see if the results we found would still hold with improved data.

Acknowledgements

We would like to thank Professors Pascal Wallisch and Milan Bradonjic for their time and effort in teaching us this semester. We credit their course materials, the DS-GA 1001 Lecture Notes and Lab Code files, with enabling us to complete this project.

References

[1] Roy, Satyaki, and Preetam Ghosh. “Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking.” *PloS one* vol. 15,10 e0241165. 23 Oct. 2020, doi:10.1371/journal.pone.0241165

[2] satunr. “Main.xlsx.” *COVID-19/US-COVID-Dataset*, Master branch. https://github.com/satunr/COVID-19.