COVID 19 in Pakistan

Allison Fischer
allisonfischer@lewisu.edu
DATA-51000001, Summer 2021
Data Mining and Analytics
Lewis University

I. Introduction

In the early months of the COVID 19 epidemic the concern has been less about community spread and more about international travel. I sought to uncover similarities within the data which would give a better understanding of the spread of the virus, the deaths associated with the virus, and possible extrinsic factors contributing to outcomes as of May 2020. This analysis involves COVID 19 data in Pakistan, especially looking at cities, regions, deaths, cases, recoveries, and travel [1]. With Orange 3 data visualization and analysis software, I have used principal component analysis (PCA), k-means clustering, and hierarchical clustering to investigate similarities [2].

In the future sections, I will describe the data, methodology, results, and conclusion. In section II, I provide the data used and descriptions. In section III, I present the analysis methodology. In Section IV, I describe the results and analysis is discussed. Finally, in section V the conclusions are discussed.

II. Data Description

This data includes the dates, number of new cases, number of new deaths, and number of new recovered cases reported in Pakistan. This can be seen in Table I, which also includes the province of the case, the city of the case, and whether the particular case engaged in international travel.

TABLE I. DATA ATTRIBUTES

Attribute	Type	Example Value	Description
DATE	Nominal (string)	02/26/2020	Date of reporting
CASES	Numeric (integer)	5	Number of new cases
DEATHS	Numeric (integer)	2	Number of new deaths
RECOVERED	Numeric (integer)	5	Number of new recovered cases
TRAVEL HISTORY	Nominal (string)	Dubai	International travel
PROVINCE	Nominal (string)	Federal Administration Tribal Area	Province reporting
CITY	Nominal (string)	Islamabad	City reporting

The numeric attributes used for statistical analysis are cases, deaths, and recovered. Descriptive statistics are provided in Table II.

TABLE II. ATTRIBUTE DESCRIPTIVE STATISTICS

Attribute	Mean	Standard Deviation	Range
CASES	22.64759	3.977344	180
DEATHS	0.455572	1.804996	37
RECOVERED	2.46762	1.861503	66

From February 26, 2020 through May 10, 2020 there was a total of 30,076 COVID 19 cases, 605 deaths, and 3,277 recovered cases reported. The Sindh and Punjab provinces reported the most cases, deaths, and recovered cases. This can be seen in Figure I.

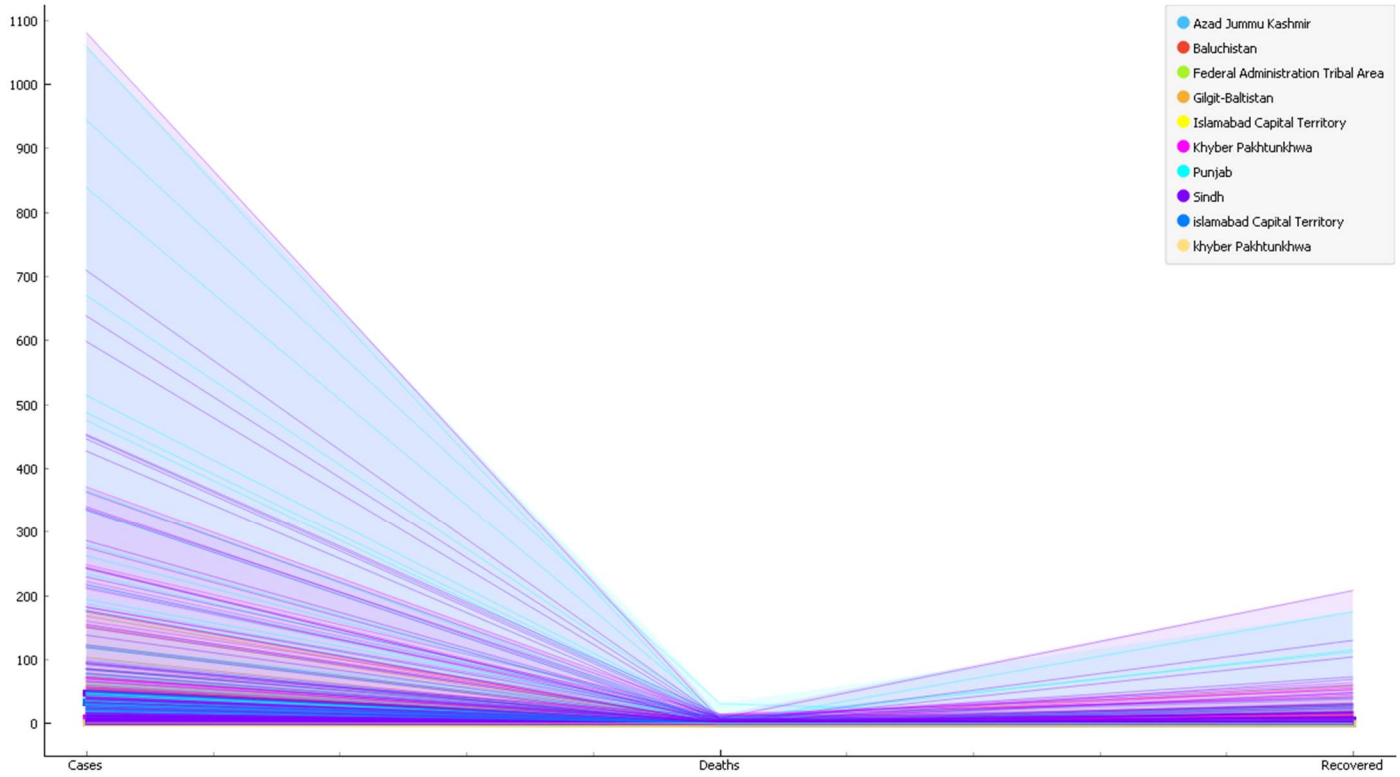

Fig I. Cases, Deaths, and Recovered by Province

III. Methodology

First, the data was checked for suitability. This was done by inspecting for missing data. Since there were no missing data, further cleaning was not necessary. Next, a line plot was created to visualize data (Fig I) and look for any differences by Province. Then PCA and k-means testing were performed [3]. K-means testing produced a silhouette score of 0.725 for two clusters [4]. Scatterplots were created to visualize clusters by principal components and by attributes. Then hierarchical clustering was performed to investigate linkage[5]. Finally, conclusions were made from data. Figure II outlines the methodology.

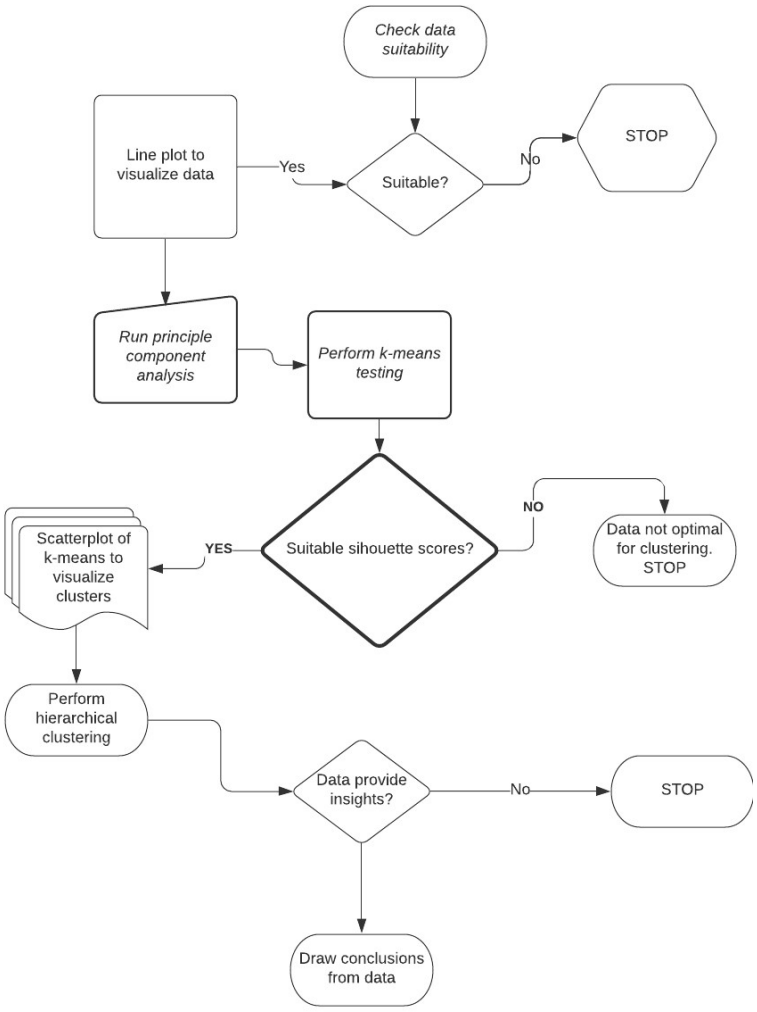

Fig II. Methodology Flowchart.

IV. Results and Discussion

Using the scatterplot of Cases vs. Deaths (Fig III) it can be seen that cases and deaths have a slight positive association. It can also be seen that cluster 2 tends to fall in the area of higher deaths and higher cases. Cluster 2 is also less dense than cluster 1.

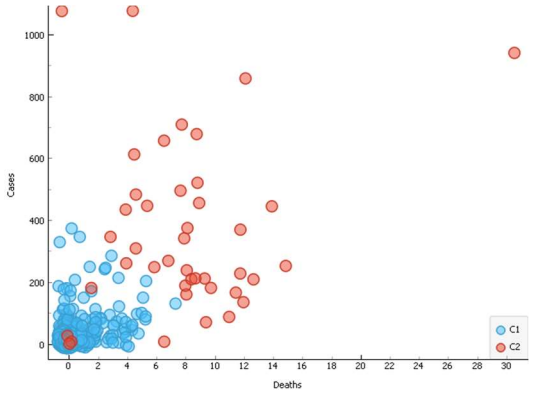

Fig III. Cases vs. Deaths Scatterplot.

From just visualizing the data three or even four clusters may seem appropriate (Fig. IV). However, the k-means test produced a silhouette score of 0.725 for two clusters, 0.707 for three clusters, and 0.131 for four clusters. Therefore, the two cluster scatterplot of PC1 vs. PC2 (Fig V) was used for further analysis.

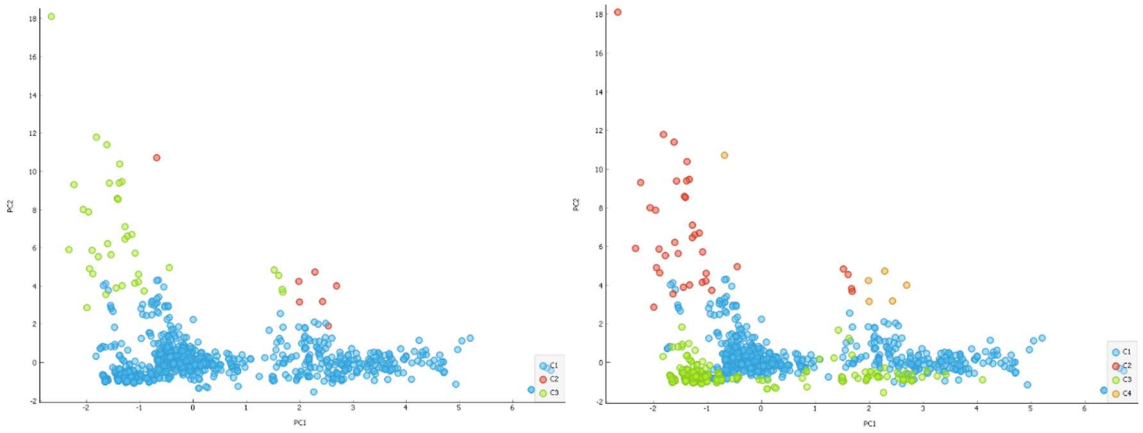

Fig IV. PC1 vs. PC2 Clusters.

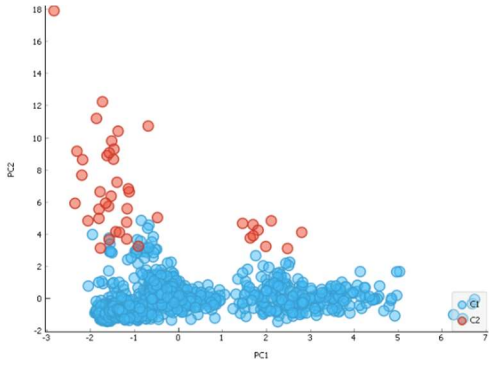

Fig V. PC1 vs. PC2 2 Clusters.

Hierarchical clustering using Euclidean distance was also performed, and two noticeable clusters seem to be divided by city (Fig. VI).

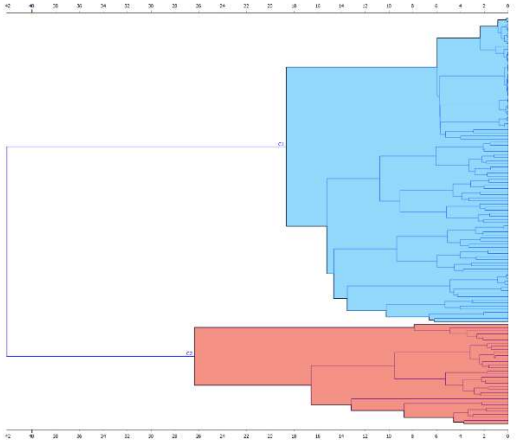

Fig VI. Hierarchical Clustering.

After visualizing the clusters, I did further research into the city differences and found that the data in cluster 1 (Fig. V) tended to be from less populous cities [6]. The hierarchical clustering was mostly divided by presence or absence of international travel and international travel was more often associated with cases in larger cities.

V. Conclusions

These research and analysis were performed to see if there was evidence of local community spread in addition to the expected spread amongst international travelers and the higher expected spread in more densely populated areas. After the data was reviewed for suitability, PCA, k-means testing, and hierarchical clustering were performed. The results were visualized with plots. Further investigation into the division of clusters revealed that there is indeed local community spread.

[1] Z.-ul-hassan Usmani, "Pakistan Corona Virus Dataset," Kaggle, 05-Jun-2020. [Online]. Available: https://www.kaggle.com/zusmani/pakistan-corona-virus-citywise-data?select=PK%2BCOVID-19-10may.csv. [Accessed: 04-Jun-2021].

[2] Pearson, K., "On lines and planes of closest fit to systems of points in space," The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, vol. 2, 1901.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *projecteuclid.org*, 01-Jan-1967. [Online]. Available: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings%20of%20the%20Fifth%20Berkeley%20Symposium%20on%20Mathematical%20Statistics%20and%20Probability,%20Volume%201:%20Statistics/chapter/Some%20methods%20for%20classification%20and%20analysis%20of%20multivariate%20observations/bsmsp/1200512992. [Accessed: 04-Jun-2021].

[4] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, 01-Apr-2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257?via%3Dihub. [Accessed: 04-Jun-2021].

[5] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[6] T. Brinkhoff, "Asia," *Population Statistics, Maps, Charts, Weather and Web Information*. [Online]. Available: https://www.citypopulation.de/en/pakistan/cities/. [Accessed: 04-Jun-2021].