Delft University of Technology

Introduction to Urban Data Science
EPA1316

Final Project: COVID-19 in India

Authors:
Heqi Wang (5216397)
Jiayu Zhang (5623499)
Kelton Huang (5266688)
Yuqi Meng (5256178)
Zelin Xu (5207142)

November 4, 2021

Abstract

From the beginning of 2020, together with the global epidemic of COVID-19, more than 34 millions infections and 459 thousands coronavirus related deaths have been reported in India. India has administered at least 1,069,718,646 doses of COVID vaccines so far(REUTERS, 2021). In order to know more about the relationship between COVID-19 infections and other factors, many research focus on age structure, regional economic situation, medical treatment level, etc have been started in the past year. However, there are still few research about the relationship between vaccination rate, hygiene habits and the COVID-19 epidemic.

In this report, we focus on the vaccination rate condition and percentage of house-holds with soap available for hand-washing in different states in India, and try to know more about how the factors impact the COVID-19 epidemic in the country. Because many studies have proved that the vaccine can significantly reduce the risk of severe and death, and has a protective effect on mild disease and infection. In the meanwhile, washing hands with soap can effectively prevent the contact and transmission of bacteria and viruses and prevent infectious diseases. We also use K-means cluster method to find if there is other interesting relationship between the factors.

1 Introduction

1.1 Research question

- What's the relationship between vaccination rate and COVID-19 epidemic?
- What's the relationship between percentage of house-holds with soap available for hand-washing and COVID-19 epidemic?
- Is there any significant link between soap hand washing rate and vaccination rate?
- What is the relationship between rate of elderly people aged 65 years or older and total mortality rate?

1.2 Important findings

- According to the cluster results, we found that states with higher vaccination rates have higher rates of soap hand washing. This result is Consistent with our first assumption before. We think that this situation might mostly cause by local economic situation. In general, places which have higher economy level are more likely to have the higher class families. These families pay more attention to personal hygiene status and epidemic prevention. In addition to this, the high rates of soap hand washing also reflect the high level of better piped water system and sanitation. That's why these states have lower rates of vaccination.
- On the other hand, there is clustering relationship between soap hand washing rates and vaccination rates. States with higher hand-washing rates had higher vaccination rates. This phenomenon is even more significant than the previous hypothesis. This also might have connection to the rules and notification messages of local governments.
- One phenomenon which interests us is that we found states with highest infection rates have the lowest mortality rates, while states with highest mortality rates have the lowest infection rates. This is very different from our prediction. We think maybe states with highest infection rates also have better medical conditions and more domestic attention. That's why the treatment is more efficient in these area.
- In addition to this, we found states with higher rates of people aged 65 years or older have higher mortality rates. This is the same as our expectation. Because in general, older people have more underlying diseases and lower resistivity. It is much more difficult for old people to get over the diseases and complications compared to younger people, which leads to higher mortality rates.

1.3 Why the findings are important

- It is essential to know the relationship between vaccination rate and COVID-19 epidemic because understanding the ability of the vaccine to block the transmission of the virus is extremely critical. For the perspective of public health, this involves whether the epidemic can be controlled by the vaccine.

- In the meanwhile, the relationship between percentage of house-holds with soap available for hand-washing and COVID-19 epidemic also helps us to decide if it is necessary to take other public health measures and give more effective suggestion to public to avoid faster transmission of the virus.

- In addition, relationship between rates of infection, rates of people aged over 65 years and mortality provides us reference in emergency medical resource allocation for the governments.

2 Related Work

- A series of action can be taken to effectively prevent the spread of COVID-19, and one of the most important actions is cleaning hands frequently with alcohol-based hand rub or soap and water (WHO, 2021). In India, especially in the rural areas in India, investigation does not show there is a good hygiene situation in handwashing. Half of the deaths in children under five years were a result of pneumonia and diarrhoeal diseases (WaterAid India, 2017), and the main cause of this situation is unsafe water, sanitation and poor hand hygiene. (Institute for Health Metrics and Evaluation, 2017) This undoubtedly speed up the spread of COVID-19 in India. Therefore, the hand wash condition in India is a key factor influence the spread of COVID. In this report, handwashing condition will be taken as an important factor to investigate the influence factors of COVID spread situation in India.

- Corona vaccination is one of the most effective way to prevent since it obviously reduce the possibility of infection and the risk of getting severe illness from COVID-19. (Asian Institute of Medical Science, 2021) From January 16, 2021, India started its COVID-19 vaccine drive(Om Prakash Choudhary, 2021), and since June 21, all adults in India can get frees vaccination. (India Today, 2021) By the end of October, about 30% Indian people are fully vaccinated and more than 707 million people have got their first shot. (BBC, 2021) And investigation show that vaccinations reduce chance of Covid death in India.(Bloomberg, 2021) That makes the vaccination situation in India also an important factor affect the spread of COVID-19 in India.

- India is a country with large population, and in some main Indian cities, the population density is extremely high. (IIPS India, 2021) COVID-19 can be spread within the distance of 1 meter(WHO, 2020), but in India, especially in some main cities, keeping a social distance is very difficult and people are always standing neck-by-neck.(USnews, 2020) That gives a good condition for the spread of COVID-19 which contributes the massive surge of infection in India.(Nature, 2021) All of these information makes population density an unignorable factor in the study of COVID-19 spread in India.

- There is also a massive aging population in India. In 2019, over 139 million Indian people's age are over 60, which is over 10% of the total population in India.(HelpAge, 2019) These aging people in India are the most dangerous group to infect and get severe illness from COVID-19, since physiological changes come with ageing and potential underlying health conditions make them in significant risk.(WHO Europe, 2020) People elder than 60 is also the first priority of vaccination in Indian COVID-19 vaccine drive.(COVID-19 AND OLDER ADULTS IN LOW AND MIDDLE INCOME COUNTRIES, 2021) Because of this, the amount/proportion of elder people in India will also be considered in this study.

- The statistics of the vaccination proportion of different age groups in India is incomplete, since this is an important factor that influence the spread of COVID-19, so there are probably uncertainties in the study about the influence of older people number and vaccination condition to the spread of COVID-19. In addition, the vaccination situation in India is developing since the vaccine drive is still running and more people are getting vaccination in each day. Therefore, a up-to-date check on statistic data is useful to the study in this report.

- The natural environment in different areas is quite different. In some remote mountain regions of desert regions in India, the population density is much lower than the plain areas. In this remote regions, because of poor traffic conditions or development condition, the proportion of vaccination is probably lower than the central regions, but the cases infected or death in those areas are also lower than other areas because of lower population density. This special condition is necessary to consider in the study.

- To prevent the spread of COVID-19, different states in India also released different policies to deal with the pandemic of COVID-19, the effects of these policies are different and some policies are probably not be well executed. Information about this is limited and this also add the uncertainty of study. But the different policies in different states do need to be considered in the study.

3 Explanatory Data Analysis

3.1 Overview of the study

- The data we used in this study are mainly obtained from the open source data provided by DDL (Development Data Lab of India), including the data set "DDL COVID India" which comes from SHRUG (The Socioeconomic High-resolution Rural-Urban Geographic Platform for India) (Asher, Lunt, Matsuura, & Novosad, 2020). Shape files used for mapping and visualising the data are downloaded from the Database of Global Administrative Areas (GADM, 2021). The data have different levels of resolution, some are at state level, while others are at sub-district level. Hence, due to this issue, the data were processed and analysed at state level. Then data sets were merged and tidied, also scaled in order to further implement K-means cluster method via sklearn. Apart from that an explanatory data analysis was conducted for observing the data characteristics.

3.2 Data description

- The data sets downloaded from the web page of DDL COVID India (DDL COVID India, 2021) were:
  - agmark
  - covid
  - demography
  - hospitals
  - mortality
  - nfhs

  While after exploring the metadata and data we excluded "agmark" and "mortality", as the former focused on farm product data which seems irrelevant to our research question, and the latter one actually described total death counts of all causes instead of deaths caused by COVID in our expectation. Besides, in the data set "hospitals" the data were lack of completeness, thus this data set was also dropped.

- The data set downloaded from the web page of the SHRUG (the SHRUG, 2021) was:
  - shrug-v1.5.samosa Population and Economic Censuses

  This data set is mainly used for examining the population statistics of each state.

- The research scope is therefore the COVID, demography and health survey data of India.

3.3 Data interrogating and analysing

- All data sets were stored in .csv format, aggregated to state level as well as yearly level. The reasons are as follows:
  1. .csv files can be directly and visually observed, ensuring an intuitive knowledge to the data.
  2. By exploring the .csv data we found the data at district/sub-district level are occasionally missing, and the district names changed a lot between the year census was conducted and the year of pandemic. Hence the geographical unit of the study was set to state, combining the consideration of policy implementation.
  3. Notably the COVID cases and deaths data was downloaded on 27th Oct. 2021, and the data were aggregated until this date since 1st Mar. 2020. As not all data were available to daily or even monthly level, for completeness other variable data were aggregated to yearly level.

- After preliminary EDA, in district-level population density data there was an outlier found to have extremely large value, and it was removed.

- There were a few data points found to be missing after aggregating, and these were filled with the mean value of the variable among all the other states. This is because the total number of states is quite small, rows containing NaN values cannot be arbitrarily removed. Hypothesizing mean values can represent the "real" average status, the NaNs were filled.

- As the cases, deaths and vaccination data were up to date, while the other data were based on the census of 2011, a big difference existed: the establishment of states Telangana and Ladakh. These 2 states were originally parts of other states, i.e. Andhra Pradesh and Jammu Kashmir. Different from previously mentioned NaNs, the whole entries of these states are blank. We believed it was not plausible to "create" two new entries which would bring relatively large impact to the study, therefore they were removed from further analyses.

- Variables used in this study could be find in Appendix A, along with the corresponding description and source. The selection of variables took both our research questions and the characteristics of data into consideration.

- As clustering was the method selected, one essential thing is to ensure all variables have the same scale. Otherwise variables with larger scale (e.g. population density and un-scaled hand wash statistics) would become dominant in the clustering and lead to results skewing towards them. In this study Min-Max scaling was applied, which is a common scaling method in data pre-processing.

- .csv files were read into the format of dataframe using Pandas; shape files were read by Geopandas. Then these were joined together after cleaning, filling and scaling. The scaling method applied here is Min-Max scaling ensuring all data values lay between 0 and 1, and this was achieved by lambda function. EDA was conducted, which can be found in the last bullet point in this section. K-means clustering was the specific clustering algorithm used in this study due to its universality and generality, which was achieved by Sklearn. However as the number of clusters should be artificially determined, and it is too arbitrary to decide a value intuitively, Elbow method was implemented for assisting to find the most suitable number of clusters. At the same time different K values were also tested to see whether the practical artificial choice stays in line with the elbow method.

3.4 Limitations

- A big limitation of the data is the time dimension. Some data are available daily, while others are only yearly. This prevents us to find a more accurate pattern of the data, especially in such a pandemic where things may totally change in a week or so. This similarly holds to the geographical dimension, with data missing at district level, it is not feasible to analyse the potential huge difference between various districts, especially between urban and rural areas.

- Another big limitation is, the COVID data are up to date, while the demography data and health survey data came from 10 years ago. India is a rapidly changing developing country, a period of 10 years could tell a completely different story, which has no benefit to our study. After the new census in 2021 is conducted, more valuable and consistent results can be found through the same study.

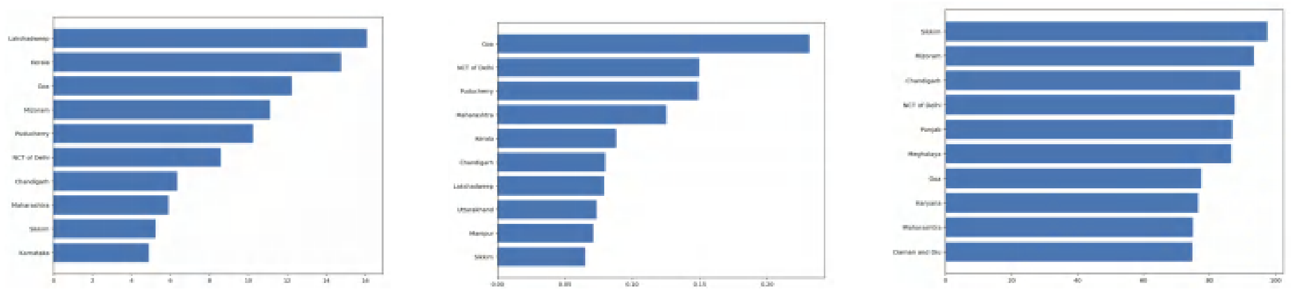
- Only taking hand washing with soap as the representative of hygiene habits is also too simple, as the viruses spread in the air and through the droplets of coughing and sneezing. If data about respiratory protection measures are available, our research question can be explored in a more targeted manner.

- The data of COVID cases and deaths actually depends on how many people were tested. More people get tested, more people will be (found to be) infected. And of course only the deaths of recorded cases can be regarded as COVID deaths, despite many were already passed away, not even found to be infected. If case and death numbers can be adjusted using tested numbers, the data would be more helpful.

- K-means clustering is a nice algorithm, however it also has disadvantage in our case. As none of the project team members are Indian, lack of cultural background prevents us from identifying the problems and characteristics of the data, e.g. the potential outliers in the data set. Unfortunately K-means clustering is especially sensitive to outliers, as the centroids can easily dragged away by them thus the results will be highly affected.

3.5 Figures and EDA

- Bar plots were made to see which states have higher value in terms of each variable, which can be found in Appendix B Fig. 9. As examples, some barplots of focused variables are displayed in Fig. 1.



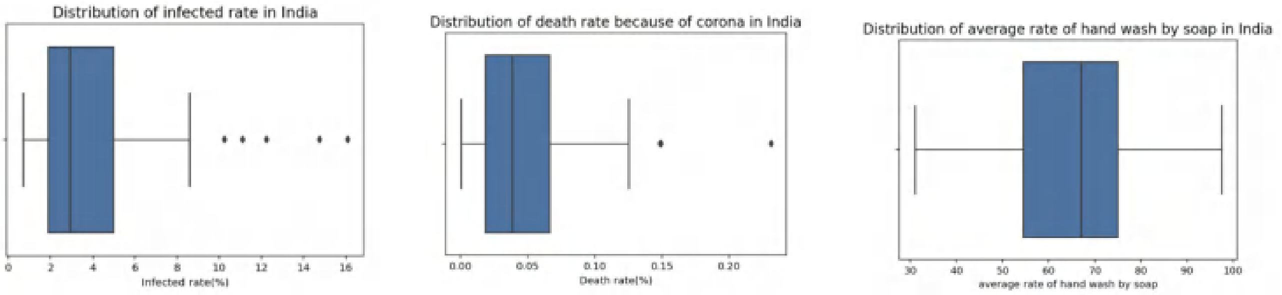
(a) Infected rate top 10 states

(b) Death rate top 10 states

(c) Soap washing hands top 10 states

Figure 1: Focused variables bar plots

- Distribution of the data was shown in box plots and histograms in Appendix B Fig. 10 and 11. As examples, some box plots of focused variables are displayed in Fig. 2.



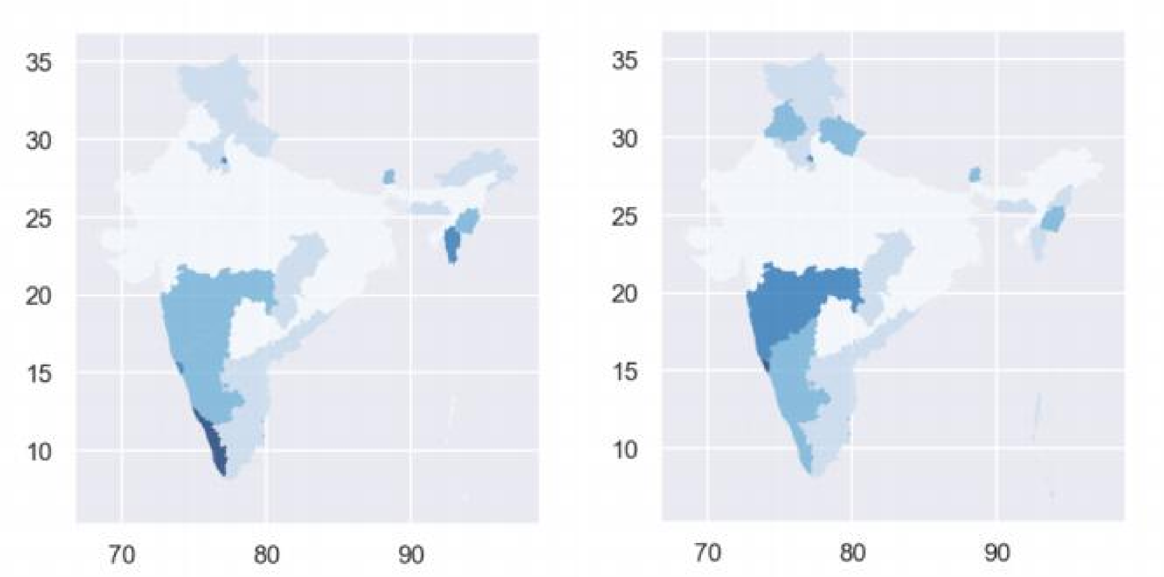
(a) Infected rate distribution

(b) Death rate distribution

(c) Soap washing hands distribution

Figure 2: Focused variables box plots

- Choropleths for each variable were shown in Appendix B Fig. 8, revealing the spatial distribution of the variables. The choropleths of infected rate and death rate are shown in Figure .



(a) Infected rate

(b) Death rate

Figure 3: The choropleths of infected rate and death rate

- Five outliers were found in terms of infected rate (i.e., Lakshadweep, Kerala, Goa, Mizoram, and Puducherry), and death rate also contained two outliers (Goa and NCT of Delhi). Although the percentage of household with soap available for washing hands showed no outliers, this large variation between 30% (Odisha) and 97.5% (Sikkim) is noticeable.

4 Analysis

- The hypothesis of the topic is the vaccination rate and hygiene habits (represented by percentage of households with soap available for hand-washing) would impact the infection and mortality rates of COVID-19 epidemic.
- Through aforementioned explanatory data analysis of the variables selected, it could be observed that there is a big difference of some variables between states. For instance the population density in NCT of Delhi is much higher than any other states, which could be considered as a outlier of the data. Thus the hypothesis validating progress couldn't not only focusing on the absolute value of one variable, multi variables should be considered at the same time instead, finding out the similarity between states through a normalised clustering algorithm.
- The clustering algorithm selected in this research is the k-means clustering algorithm. For the k-means clustering algorithm uses the Euclidean distance as the criteria of similarity judgement, the normalise progress of data is essential. The normalising method for the data is min-max normalization.

  The key of the k-means clustering algorithm is define a proper value of the k, i.e. the number of the clusters. To find the most appropriated k value, the elbow method was used. The elbow method is based on SSE (sum of the squared errors), evaluating the distance of points in a cluster to the cluster centroid. The plot of SEE values for different k value is shown in Figure 4.

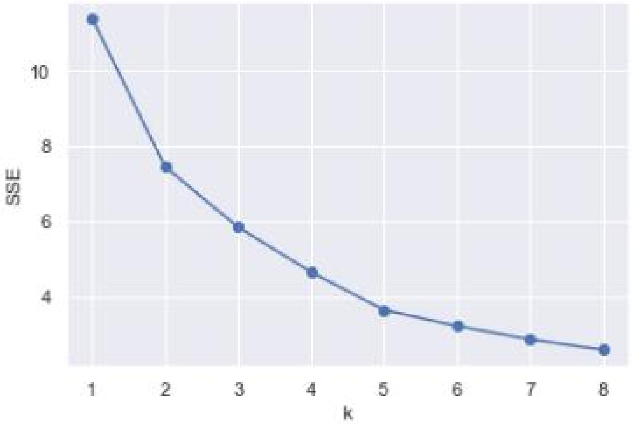
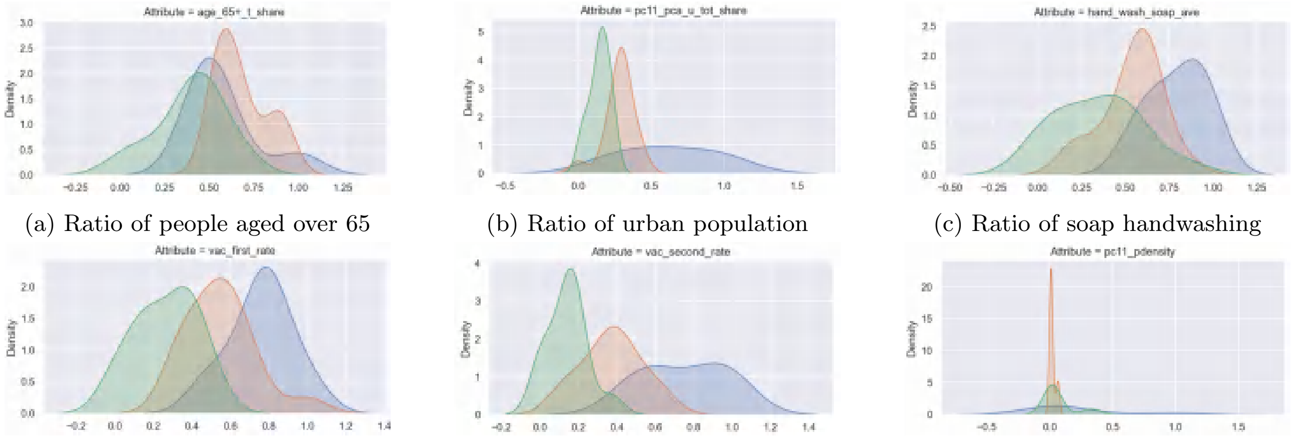


Figure 4: Elbow method of finding appropriate K value for clustering

Based on the plot of the elbow method, the "elbow point" of the figure is considered to be k=3 or k=5. Complemented by the results of the clustered distribution of each variable, there would be a better distinction between clusters when k=3, at the same time having a certain number of states in each cluster (Appendix C). Thus, the following analysis would based on the k=3 clustering results, the clustering map is visualised in Figure 5a.



(a) Ratio of people aged over 65

(b) Ratio of urban population

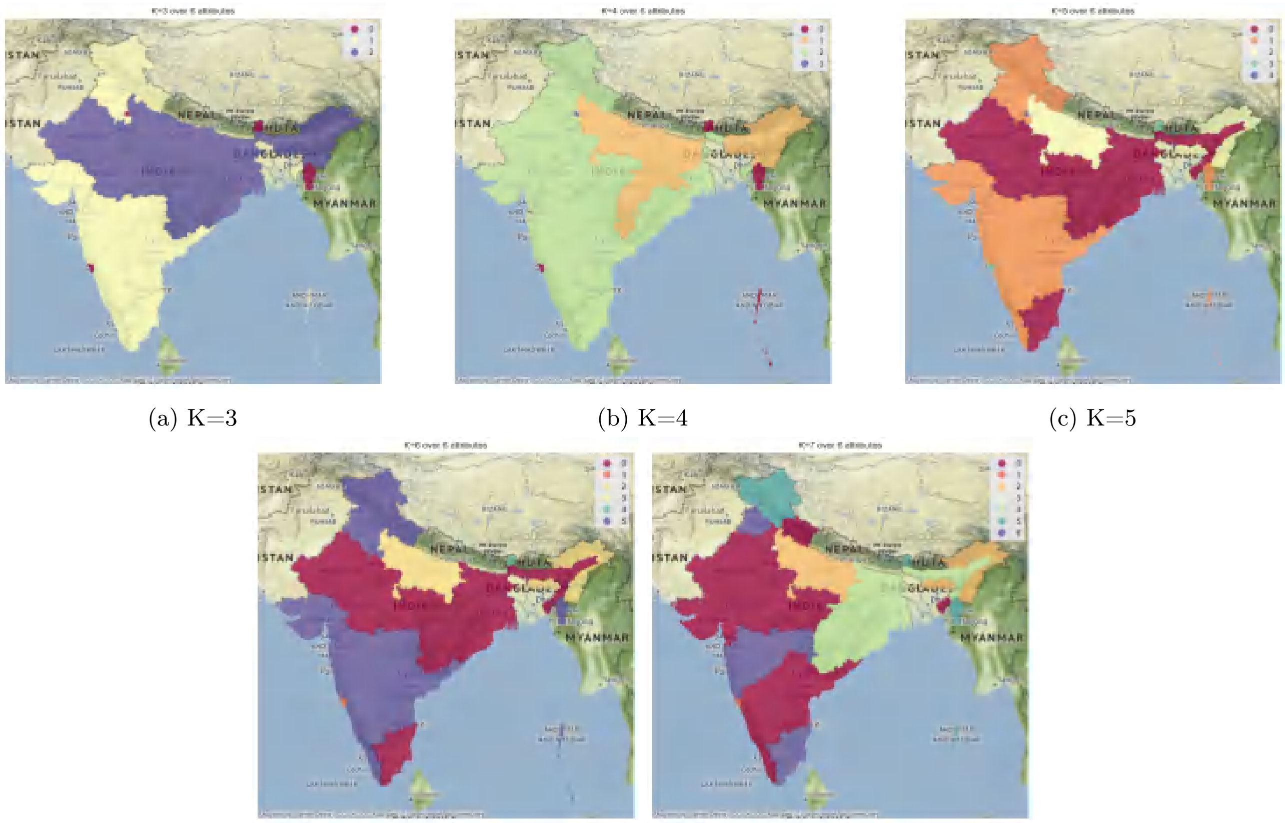
(c) Ratio of soap handwashing

(d) Vaccination rate of 1st dose

(e) Vaccination rate of 2nd dose

(f) Population density

Figure 6: KDE plots of variables in clusters when K=3 (Blue represents cluster 0, orange represents cluster 1, Green represents cluster 2)



(a) K=3

(b) K=4

(c) K=5

(d) K=6

(e) K=7

Figure 5: K-means clustering map results over 6 attributes with different K values

4.1 Clustering result Analysis

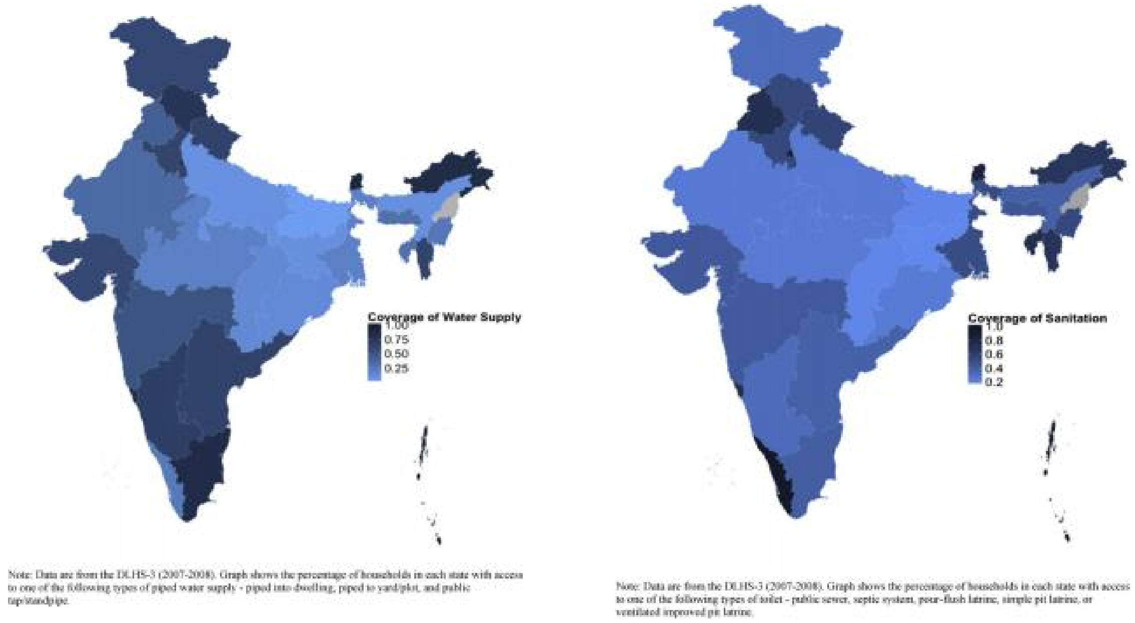
- Comparing the clustering map with different k values, we found some states are always keep in the same cluster, for example, the Rajasthan and the Madhya Pradesh, the Telangana and the Karnataka state. This means that these states share some same features in their vaccination and the hygiene habits. These features may related to the social-economy environment or policies in these states.

- In comparison between clustering with different k value, the KDE plots of vaccination rate of first dose and the second dose always have the most pronounced disparity between clusters. This means the vaccination rate variables are the strong drivers of the clustering.

- Horizontally, the variables' KDE plots also show some patterns among the states in India. We found that the states in higher first-dose vaccination rate (Figure 6d)also have higher ratio of washing hand with soap(Figure 6c), higher second-dose vaccination rate (Figure 6e)and higher ratio of urban population(Figure 6b). This could correspond to the assumption we made before that vaccination status and regional economic level are related, and people in higher economic areas tend to have better hygiene practices.

4.2 State Analysis

- Since the research assumption is focusing on the vaccination and the hygiene practices, which could also represent the sanitary conditions, the following analysis would based on the sanitary conditions, COVID vaccine policies and the infection rate, fatality rate in the states of India.
- The hand washing with soap represent not only a hygiene practice in daily life, but also reflecting the sanitary conditions and water supply in the areas. According to the research (Nandi, Megiddo, Ashok, Verma, & Laxminarayan, 2017), the maps of piped water supply coverage at baseline and improved sanitation coverage at baseline are shown in Figure 7a and Figure 7b. The geographic distributions in the figures are well corresponding to the clustering plot (K=3). States have higher piped water supply and higher coverage of improved sanitation are concentrated locating in cluster 1. On the contrary, the overlap between lower piped water supply and improved sanitation and the states in cluster 2 is obvious. In the clustering results, the states in cluster 1 have higher rate of soap hand washing and two-dose vaccination. The similarity of maps could shows the relationship between higher rate of soap hand washing and better piped water supply and improved sanitation, the low rate of soap hand washing could be limited by the infrastructure and the sanitation.



Note: Data are from the DLHS-3 (2007-2008). Graph shows the percentage of households in each state with access to one of the following types of piped water supply - piped into dwelling, piped to yard/plot, and public tap/standpipe.

(a) Geographic distribution of piped water supply coverage at baseline

Note: Data are from the DLHS-3 (2007-2008). Graph shows the percentage of households in each state with access to one of the following types of toilet - public sewer, septic system, pour-flush latrine, simple pit latrine, or ventilated improved pit latrine.

(b) Geographic distribution of improved sanitation coverage at baseline

Figure 7: Geographic distribution maps of India in related researches (Nandi et al., 2017)

- India central government released the new vaccine policy in June 2021, and the new policy would be effective since 21st June. According to the new policy, all India citizens aged 18 and over could get free COVID-19 vaccines at central and state government centres. In the earlier policy, free vaccination is only available for people aged from 18 to 44 at centres run by state government. Besides the central and state government centres, all citizens would have to pay for vaccination at private centres. In the report from (The Indian Express, 2021), three quarters of vaccines would be distributed among states for free vaccination, and the remaining 25% would be sold in the private centres. Under this policy, the states have smaller population or adequate medical resources per capita (Geography and you, 2018) would get rapid vaccination, such as NCT Delhi, Goa, Sikkim and Mizoram. These states are located in cluster 0. With similar medical resources per capita, the states with higher GDP (Knoema, 2021) could get vaccinated at private health centres, faster than states with lower GDP. The policy and the economy condition of states are coincides with the clustering results of vaccination.

5 Discussion/Conclusion

5.1 Summary

- Through a series of studies in the literature, hand washing and vaccination were found to be two effective ways to prevent the spread of new crowns, leading to the hypothesis of "The vaccination rate and hygiene habits (represented by percentage of households with soap available for hand-washing) would impact the infection and mortality rates of COVID-19 epidemic. "

- After cleaning the data, the normalization method: min-max normalization was used because of the order-of-magnitude differences between the raw data for each variable. The different metrics were solved by converting the original data linearized to the range [0,1] and changing the weights of the variables in the analysis.

- EDA analysis was performed on several variables that may be associated with the presence of the covid: percentage of people aging 65 and older; percentage of urban population; total infection rate, total death rate, percentage of households with soap available for hand washing; percentage of people get first/second vaccination and population density. The data distribution of each variable and data points with special significance, such as the plural, the median and the magnitude of its maximum value, are shown through box plots and bar charts.

- Cluster analysis was done with the the level of Indian states, cluster was performed at the state level in India and variables were clustered into two categories, one for total death rate and total infection rate, the other for factors that may have influenced the pandemic. The results show that there is a more pronounced clustering characteristic between the different states when K=3, especially for variables: percentage of households with soap available for hand washing and percentage of people get first/second vaccination.

- Hypothesis of this paper is vaccination and hand washing with soap are effective means of preventing covid in the Indian region. The content of the hypotheses was not fully confirmed, however, two related hypotheses were reflected in the results.

  1. The rate of elderly people aged 65 years or older is associated with total mortality, and states with more elderly people are found to also typically have higher mortality rates, while states with fewer elderly people also have lower mortality rates.
  2. There was clustering relationship between soap hand washing rates and vaccine first/second dose vaccination rates. States with higher hand-washing rates also had higher vaccination rates for the first/second vaccine, and states with lower hand-washing rates also had lower vaccination rates, a more significant phenomenon than the previous hypothesis.

5.2 Conclusions

As mentioned, the results of the clustering support two new hypotheses; states with higher rates of people aged 65 years or older have higher overall mortality rates; and states with higher rates of soap hand washing have higher rates of first/second doses of vaccination.

- For the first phenomenon, 65 older people are relatively less healthy with cardiovascular disease, respiratory disease, immune deficiencies, etc. Most of them have underlying diseases. Covid is is a serious respiratory disease caused by corona virus, so it is often difficult for older people to resist compared to younger people, which leads to a higher mortality rate.

- For the relationship between soap hand-washing rates and vaccination rates, part of this is due to government policies. The government may call on the public through the news media, newspapers, or by placing vaccination notification messages in public areas. Since all three variables are currently important tools for preventing the spread of covid in each country, it is reasonable that they have similar clustering relationships. Besides, even though vaccination remains voluntary in India because of its fundamental privacy and human rights implications, some states have issued policies, but some states, such as Uttar Pradesh and Orissa, make it mandatory for traders and suppliers to be vaccinated before resuming business activities. Similarly, Assam requires all government employees to be vaccinated. (POOJA SHREE A, 2021)

- A phenomenon of interest is the relationship between infection and mortality rates in the clustering results: states with the highest infection rates have the lowest mortality rates, while states with the highest mortality rates have the lowest infection rates. This is contrary to our expectation and may be explained by the fact that areas with higher diagnosis rates have better medical conditions and can be treated effectively in the early stages of infection without progressing to severe disease. During the second wave of the covid outbreak, overcrowding in hospitals and shortage of medical facilities (e.g., ventilators) due to the weak public health care system and infrastructure in some areas resulted in a lack of access to timely and systematic treatment for many patients. Therefore, we can propose a reasonable guess that for areas with abundant medical resources, the cure rate of cure rates will be relatively much higher.(Tiyara, Inc., 2021)

5.3 Limitations

- First, from the data point of view, since this study uses data related to the 2011 census, especially the variable "percentage of elderly people over 65 years old", the elderly population should be more volatile during the ten-year interval, and the absence of the most recent data may lead to inaccurate estimation results.

- The present study is limited in its analysis of factors that may be associated with the prevalence of COVID-19 infection, considering first the nature of COVID-19 as a viral respiratory infection disease that can be transmitted by aerosols and droplets. (N.van Doremalen, 2020) The high population density in India and the more or less restrictive travel policies received since the spread of the epidemic have led to an increase in the time spent indoors. Additional transmission risk of COVID-19 has been demonstrated in indoor environments, such as schools, hospitals, etc. (G.M.Abbas,I.G.Dino, 2021) At the same time, the openness of the environment and the ventilation facilities also determine whether people are more at risk of being exposed to the virus. This implies that there may be a clustering relationship between some indices of people's housing density, such as average home occupancy/vacancy, and infection rates, with states with higher average home occupancy likely to have higher infection rates.

- In addition, the policies of individual states are also related to the state of the spread of COVID-19. Because the data are calculated on an annual basis, within a year, policies in different states are adjusted accordingly to the pandemic spread, resulting in fluctuations in the two variables of “infected rate” and “mortality rate” based on policies introduced at different times, and this may be a factor contributing to inaccurate results. For example, in the southern Indian states of Kerala, Karnataka, and Maharashtra, which were initially the states with the most severe pandemic spread, the government introduced a policy of restricting public events, i.e., mass gatherings from 10th March and total closure of public places such as theaters, shopping malls, etc on 16th March. The spread of COVID-19 was controlled with the mandatory closure of public places. (S.Sharma,M.Zhang,J.Gao,H.Zhang,S.H.Kota, 2020)

- For the rate of getting the first/second vaccination, some deviating values (vaccination rate greater than 1) appear in the processing of the data cleaning, which indicates the possibility that residents of neighboring states come to this state to get vaccinated. The reason could be that some states have a large population with a shortage of vaccine supply, or that people choose to travel to the nearest state for vaccination because the appointment slots at available vaccination sites are full. While this is only representative of a minority group, it does lead to bias in the analysis of cluster effects on infection rates, mortality rates, and vaccination rates based on geographic location of India.

5.4 Future research possibilities

- This paper could be studied in more depth in several directions in the future. First, the analysis needs to be based on the most recent demographic data, for pandemics, the situation is unpredictable and rapidly changing, and data from ten years ago are not sufficient to be the basis of the analysis.

- There are some similar variables that could be included in the analysis more than the variable of hand washing rate with soap, such as the relationship between cleaning with sanitizer and changes in infection rates, or a more extended analysis of the effect of having sanitizer at the entrance of buildings with public gatherings on the prevention of new crowns within the building (indoor environment).

- As for the clustering effect of vaccination rates, infection rates and mortality rates based on geographic location, it may be necessary to combine traffic data, vaccine data and scheduled data of vaccination sites together for analysis to classify the vaccination rate of residents from other states or residents live in this particular state, of course, access to such data requires attention to privacy and human rights issues.

- In addition to this, pandemic infection and mortality rates are associated with waves of the epidemic, i.e., mutations of the strain, and a study of waves of the epidemic in eastern Uttar Pradesh showed that young people were more affected when the second wave of the epidemic came. (Mahendra M. Reddy, Kamran Zaman, Shailendra Kumar Mishra, Priyanka Yadav, Rajni Kant, 2021) Therefore, one direction that could improve this study in the future would be to consider a joint analysis of epidemic waves in combination with the age structure of the population to determine the infection and mortality rates of various younger groups at different stages of the outbreak.

References

Asher, S., Lunt, T., Matsuura, R., & Novosad, P. (2020). *The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG).* (World Bank Economic Review (Revise and Resubmit))

Asian Institute of Medical Science. (2021). *Benefits of getting a covid-19 vaccine.* Retrieved July 19, 2021, from https://www.aimsindia.com/blog/benefits-of-getting-a-covid-19-vaccine/

BBC. (2021). *Covid vaccine: India administers more than one billion covid jabs.* Retrieved June 21, 2021, from https://www.bbc.com/news/world-asia-india-56345591

Bloomberg. (2021). *Vaccinations reduce chance of covid death in india to 0.4%.* Retrieved July 16, 2021, from https://www.bloomberg.com/news/articles/2021-07-16/about-0-4-died-of-covid-19-after-vaccination-in-delta-hit-india

COVID-19 AND OLDER ADULTS IN LOW AND MIDDLE INCOME COUNTRIES. (2021). *India sidelines older people from covid-19 vaccination.* Retrieved May 13, 2021, from https://corona-older.com/2021/05/13/india-sidelines-older-people-from-covid-19-vaccination/

DDL COVID India. (2021). *Open-source covid-19 data.* Retrieved from https://www.devdatalab.org/covid

GADM. (2021). *Gadm maps and data of india.* Retrieved from https://gadm.org/download_country_v3.html

Geography and you. (2018). *Access to healthcare in india.* Retrieved June 7, 2018, from https://geographyandyou.com/access-to-healthcare-in-india/

G.M.Abbas,I.G.Dino. (2021, 4). The impact of natural ventilation on airborne biocontaminants: a study on covid-19 dispersion in an open office. *Engineering Construction and Architectural Management*. Retrieved from https://doi.org/10.1108/ECAM-12-2020-1047

HelpAge. (2019). *Aging population in india.* Retrieved from https://ageingasia.org/ageing-population-india/

IIPS India. (2021). *Top 10 populated cities in india.* Retrieved May, 2021, from https://iipsindia.org/top-10-populated-cities-in-india/

India Today. (2021). *Free covid-19 vaccination for adults in india from today: All you need to know.* Retrieved June 21, 2021, from https://www.indiatoday.in/india/story/free-covid-vaccination-adults-india-june-details-update-modi-decision-policy-1817390-2021-06-21

Institute for Health Metrics and Evaluation. (2017). *Global burden of disease-india.* Retrieved June 20, 2017, from http://www.healthdata.org/india

Knoema. (2021). *India gdp per capita by state, 2020.* Retrieved August 2, 2021, from https://knoema.com//atlas/India/ranks/GDP-per-capita

Mahendra M. Reddy, Kamran Zaman, Shailendra Kumar Mishra, Priyanka Yadav, Rajni Kant. (2021). Differences in age distribution in first and second waves of covid-19 in eastern uttar pradesh. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, *15*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1871402121003477

Nandi, A., Megiddo, I., Ashok, A., Verma, A., & Laxminarayan, R. (2017). Reduced burden of childhood diarrheal diseases through increased access to water and sanitation in india: A modeling analysis. *Social Science Medicine*, *180*, 181-192. Retrieved from https://www.sciencedirect.com/science/article/pii/S0277953616304853 doi: https://doi.org/10.1016/j.socscimed.2016.08.049

Nature. (2021). *India's massive covid surge puzzles scientists.* Retrieved April 21, 2021, from https://www.nature.com/articles/d41586-021-01059-y

N.van Doremalen. (2020). *Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7121658/pdf/NEJMc2004973.pdf

Om Prakash Choudhary, I. S., Priyanka Choudhary. (2021, 9). India's covid-19 vaccination drive: key challenges and resolutions. *The Lancet*, *21*, 1483-1484. doi: https://doi.org/10.1016/S1473-3099(21)00567-3

POOJA SHREE A. (2021, 7). *Covid-19 and mandatory vaccination policies: A fundamental rights perspective.* Retrieved from https://www.theleaflet.in/covid-19-and-mandatory-vaccination-policies-a-fundamental-rights-perspective/

REUTERS. (2021). *Covid-19 global tracker india.* Retrieved from https://graphics.reuters.com/world-coronavirus-tracker-and-maps/countries-and-territories/india/

S.Sharma,M.Zhang,J.Gao,H.Zhang,S.H.Kota. (2020). Effect of restricted emissions during covid-19 on air quality in india sci total environ. Retrieved from https://www.sciencedirect.com/science/article/pii/S0048969720323950

The Indian Express. (2021). *India's new covid-19 vaccination policy: A quixplained.* Retrieved June 25, 2021, from https://indianexpress.com/article/explained/india-covid-vaccination-policy-quixplained-7355731/

the SHRUG. (2021). *The socioeconomic high-resolution rural-urban geographic platform for india.* Retrieved from https://www.devdatalab.org/shrug

Tiyara, Inc. (2021). *Nurse shortage impacting covid-19 situation in india.* Retrieved from https://www.tiyara.org/blog/nurse-shortage-impacting-covid-19-situation-in-india?/

USnews. (2020). *Across india, a struggle to mind the gap.* Retrieved May 18, 2021, from https://www.usnews.com/news/best-countries/articles/2020-05-18/india-struggles-to-maintain-social-distancing-amid-coronavirus-pandemic

WaterAid India. (2017). *Spotlight on handwashing in rural india.* Retrieved from https://www.outlineindia.com/othermedia/1531392037.Hand-hygiene-study.pdf

WHO. (2020). *Coronavirus disease (covid-19): How is it transmitted?* Retrieved December 13, 2020, from https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted

WHO. (2021). *Advice for the public: Coronavirus disease (covid-19).* Retrieved October 1, 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

WHO Europe. (2020). *Supporting older people during the covid-19 pandemic is everyone's business.* Retrieved April 3, 2020, from https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/4/supporting-older-people-during-the-covid-19-pandemic-is-everyones-business#:~:text=Although%20all%20age%20groups%20are,potential%20underlying%20health%20conditions.

Appendices

A Variables, Description and Source

Table 1: Variables, description, and source

	Variable	Description	Source
1	age_65+_pop	Population aged over 65 in the state	http://www.devdatalab.org/covid Demography data
2	hand_wash_soap_ave	Percentage of households with soap available for hand washing in the state	http://www.devdatalab.org/covid Nfhs data
3	pc11_pca_u_tot	Urban population in the state	http://www.devdatalab.org/covid Demography data
4	pc11_pdenity	Population density of the state (population per km^2)	http://www.devdatalab.org/covid Demography data
5	vac_first_pop	Number of people registered for first vaccination	http://www.devdatalab.org/covid Covid data
6	vac_second_pop	Number of people registered for second vaccination	http://www.devdatalab.org/covid Covid data
7	infected_cases	Number of cases infected COVID	http://www.devdatalab.org/covid Covid data
8	deaths	Number of deaths of COVID	http://www.devdatalab.org/covid Covid data
9	population	Population of the state	http://www.devdatalab.org/covid Demography data & https://www.devdatalab.org/shrug Population Census data
10	age_65+_t_share	Share of population aged over 65 in the state	age_65+_pop/population
11	pc11_pca_u_tot_share	Share of urban population in the state	pc11_pca_u_tot/population
12	vac_first_rate	Rate of population registered for first vaccination	vac_first_pop/population
13	vac_second_rate	Rate of population registered for second vaccination	vac_second_pop/population
14	infected_rate	Rate of population infected COVID	infected_cases/population
15	death_rate	Rate of population dead of COVID	deaths/population

B EDA

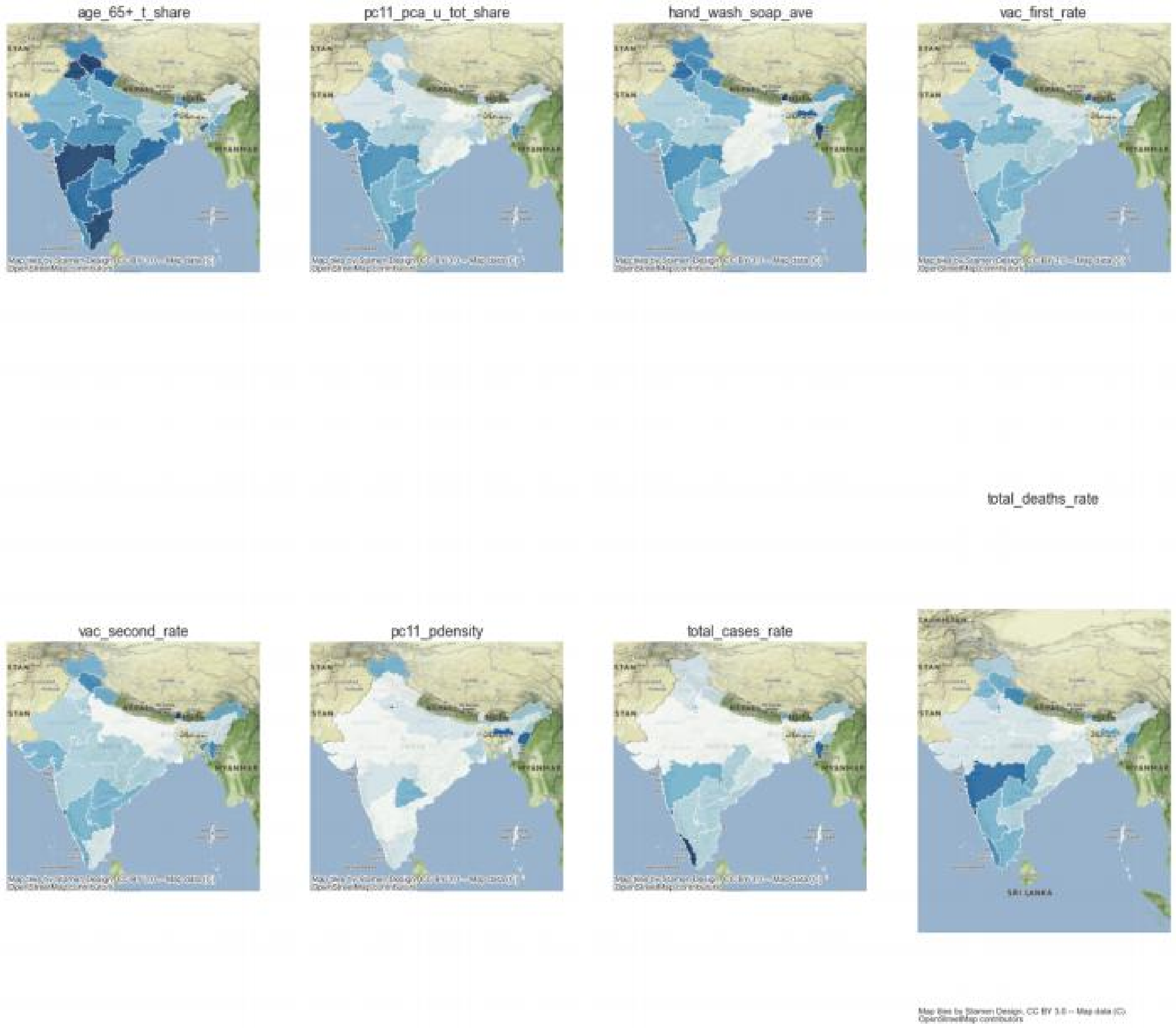


Figure 8: Choropleths of Different Variables

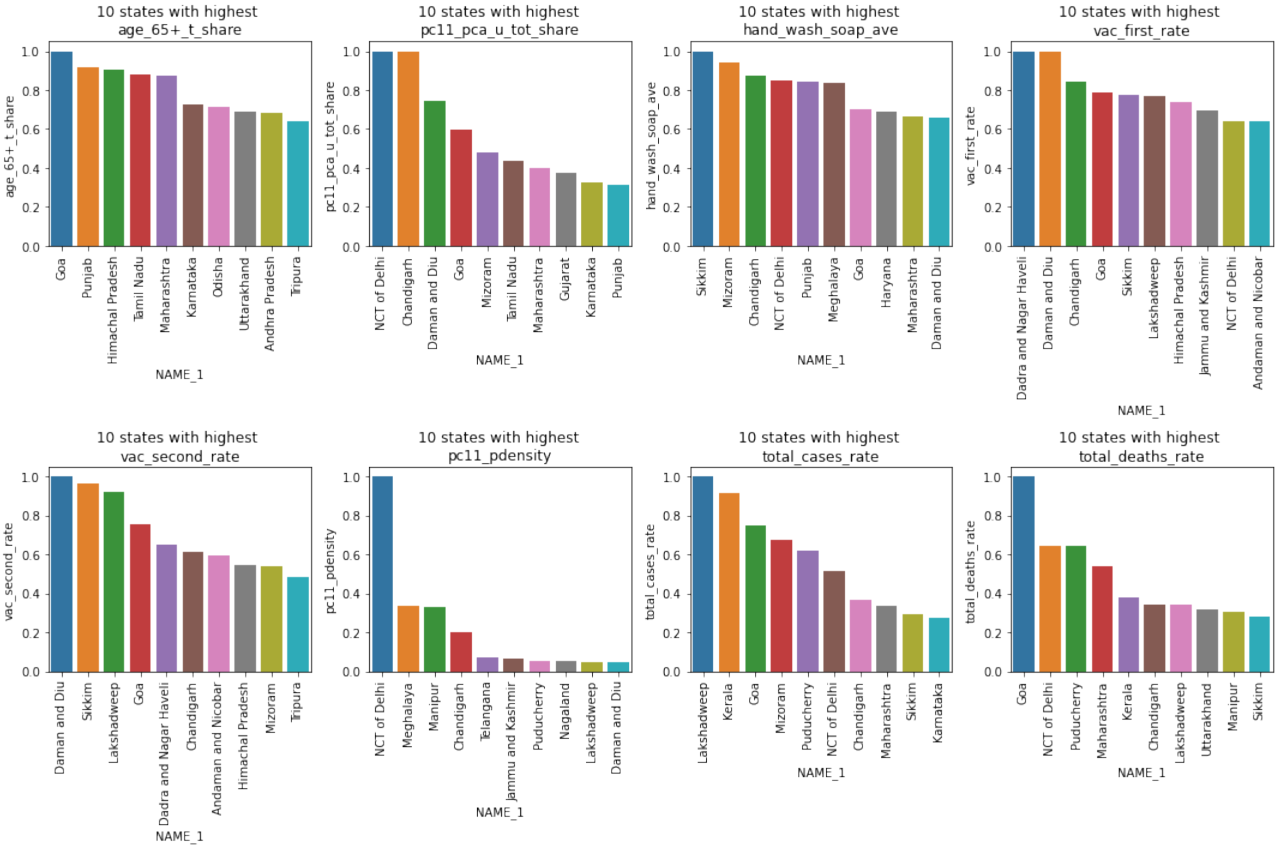


Figure 9: Bar plots showing the 10 states with the highest value of all 8 variables

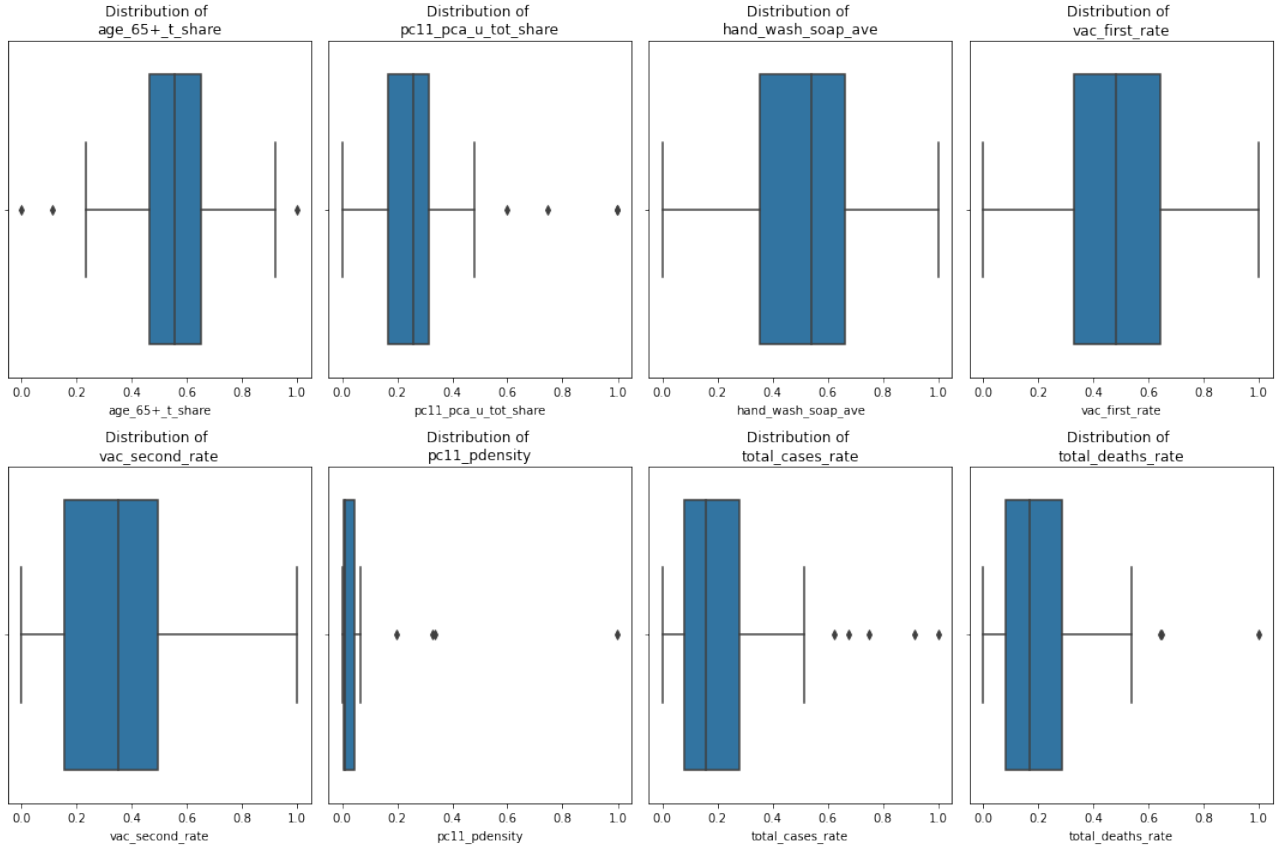


Figure 10: Box plots showing the distribution of all 8 variables

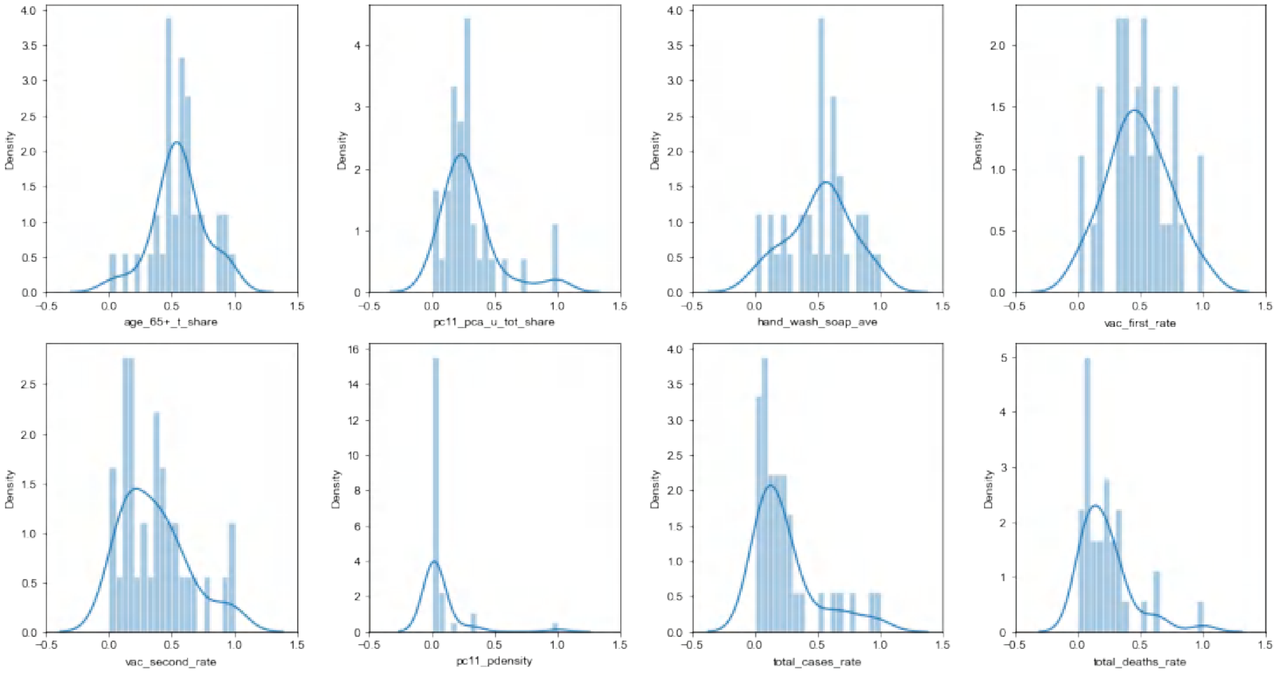
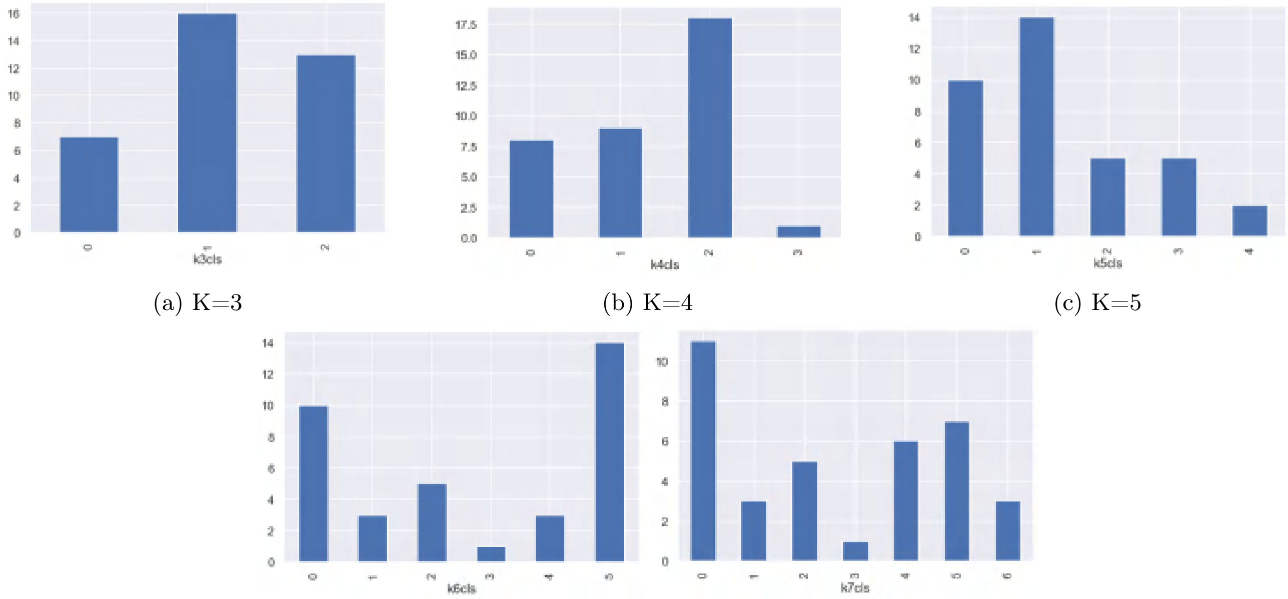


Figure 11: Histograms showing the distribution of all 8 variables

C Clustering results



(a) K=3

(b) K=4

(c) K=5

(d) K=6

(e) K=7

Figure 12: Histogram of states number in clusters with different K values