UNIVERSITY OF KARACHI

Department Of Computer Science

Academic Year: 2021

Mcs Final (Morning)

Data Mining And Warehousing

Project Report

Entitled: Covid-19 Data Analysis and Visualization

SUBMITTED BY:

Name	*Seat No*
Amir Raza	*P19101006*
Saba Islam	*P19101059*
Numrah Alauddin	*P19101052*
Sheikh Muhammad Shayan Iqbal	*P19101066*

Table of Content

Abstract

Data mining is one of the promising and constantly evolving fields in the field of data analytics. Data mining has led to solutions of various unfathomable jobs, events, diseases and evaluations. The racist grouping of techniques that falls under the data mining field makes it a formidable force for data scientists. The existing paper reviews the various papers published on COVID-19 using data mining techniques to address the pandemic in terms of its explanation, assessment and solution. The current paper reviews the work done by various authors using data mining techniques. The paper contributes uniquely to the literature by filling up the gap of review on COVID-19 related work. The following paper shows a brief detail about the all happening events took place after epidemic begins. A short review about SARS-CoV-2 tried to be summaries in the given paper. The idea to work on SARS-CoV-2 is to represent the effects of the pandemic all across the world, how it changes human behaviour, way of living, their thoughts about nature.

Keywords: COVID-19, data mining, review, pandemic, disease.

Covid-19 Data Analysis and Visualization

Since December 2019 the world is experiencing a deadly disease caused by a novel coronavirus termed as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease associated with this virus is known as COVID-19.

COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. WHO declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. Naturally, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level. The deadly impact of COVID-19 is driving a massive amount of research that aims at understanding the various characteristics of the pandemic. While there is no vaccine, considerable effort has been devoted to understanding the spread of the disease in different places in the world. The speed with which the disease has spread throughout the world demands agile solutions to understand and estimate the disease progression.

According to World Health Organization (2020), The 2019 novel coronavirus termed as SARS-CoV-2 caused pneumonia outbreak in Wuhan, China resulting in the 2019- 2020 coronavirus pandemic declared by World Health Organization (WHO). It belongs to the Orthocoronavirinae subfamily. It is distinct from Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome corona virus (SARS-CoV). Research by (Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Yu T) suggested that the infected patients showed clinical manifestations of dry cough, fever, confusion, sore throat, rhinorrhea, chest pain, dyspnea, bilateral lung infiltrates on imaging, nausea, vomiting and diarrhea. The disease caused by SARS-CoV-2 known as COVID-19 can be deadly. This happens when the severity of the disease onset results in massive alveolar damage with progressive respiratory failure with a 2% case fatality rate.

On 30 January 2020, World Health Organization (WHO) declared that COVID-19 outbreak as the sixth public health emergency of international concern, following H1N1 (2009), polio (2014), Ebola in West Africa (2014), Zika (2016), and Ebola in the Democratic Republic of Congo (2019) (Yoo JH, 2020). Ever since the emergence of COVID -19 and its consequent spread across continents engulfing both advanced and developing nations, there has been a lot of research papers publications on various aspects of COVID-19. Apart from researches being undertaken in the domain of vaccination, drug therapy and other clinical aspects, considerable research work is also being carried out with patients as the fulcrum- patients who have recovered; patients with co-morbidities and the incidence of virus etc. Thorough analysis is being performed on the people who recovered so as to shed some light on how to deal with the active cases. Data scientists all over the world are busy in making sense out of the available data and predict the near future. Finding trend pattern, feature selection, forecasting techniques are being applied in and out to come to a conclusion (Rajan Gupta, Saibal Kumar, 2020).

Literature Review

Coronaviruses belong to the Coronaviridae family in the Nidovirales order. Corona represents crown-like spikes on the outer surface of the virus; thus, it was named as a coronavirus. Coronaviruses are minute in size (65–125 nm in diameter) and contain a single-stranded RNA as a nucleic material, size ranging from 26 to 32kbs in length, as shown inz The subgroups of coronaviruses family are alpha (α), beta (β), gamma (γ) and delta (δ) coronavirus. The severe acute respiratory syndrome coronavirus (SARS-CoV), H5N1 influenza A, H1N1 2009 and Middle East respiratory syndrome coronavirus (MERS-CoV) cause acute lung injury (ALI) and acute respiratory distress syndrome (ARDS) which leads to pulmonary failure and result in fatality. These viruses were thought to infect only animals until the world witnessed a severe acute respiratory syndrome (SARS) outbreak caused by SARS-CoV, 2002 in Guangdong, China (N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K. Chan, *et al*, *2003)*.

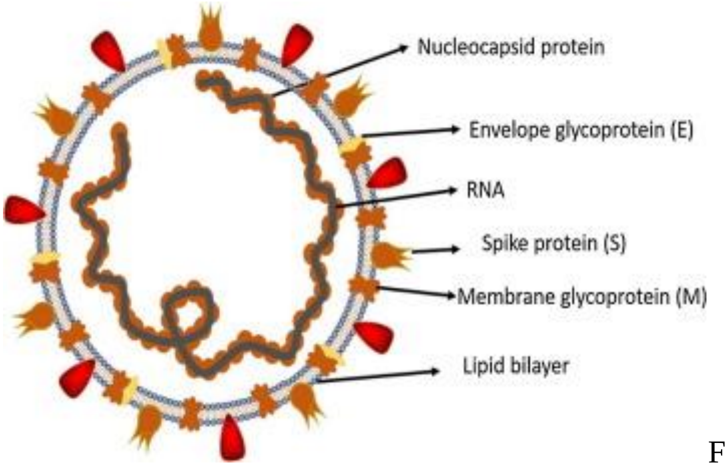


Fig. 1

In the history, SRAS-CoV (2003) infected 8098 individuals with mortality rate of 9%, across 26 countries in the world, on the other hand, novel corona virus (2019) infected 120,000 individuals with mortality rate of 2.9%, across 109 countries, till date of this writing. It shows that the transmission rate of SARS-CoV-2 is higher than SRAS-CoV and the reason could be genetic recombination event at S protein in the RBD region of SARS-CoV-2 may have enhanced its transmission ability. In this review article, we discuss the transmission of human coronaviruses briefly. We further discuss the associated infectiousness and biological features of SARS and MERS with a special focus on COVID-19.

The Analysis of Spread of Corona Virus

Research analysis suggested that the source of origination and transmission are important to be determined in order to develop preventive strategies to contain the infection. In the case of SARS-CoV, the researchers initially focused on raccoon dogs and palm civets as a key reservoir of infection. (B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, *et al*, 2005*)*.

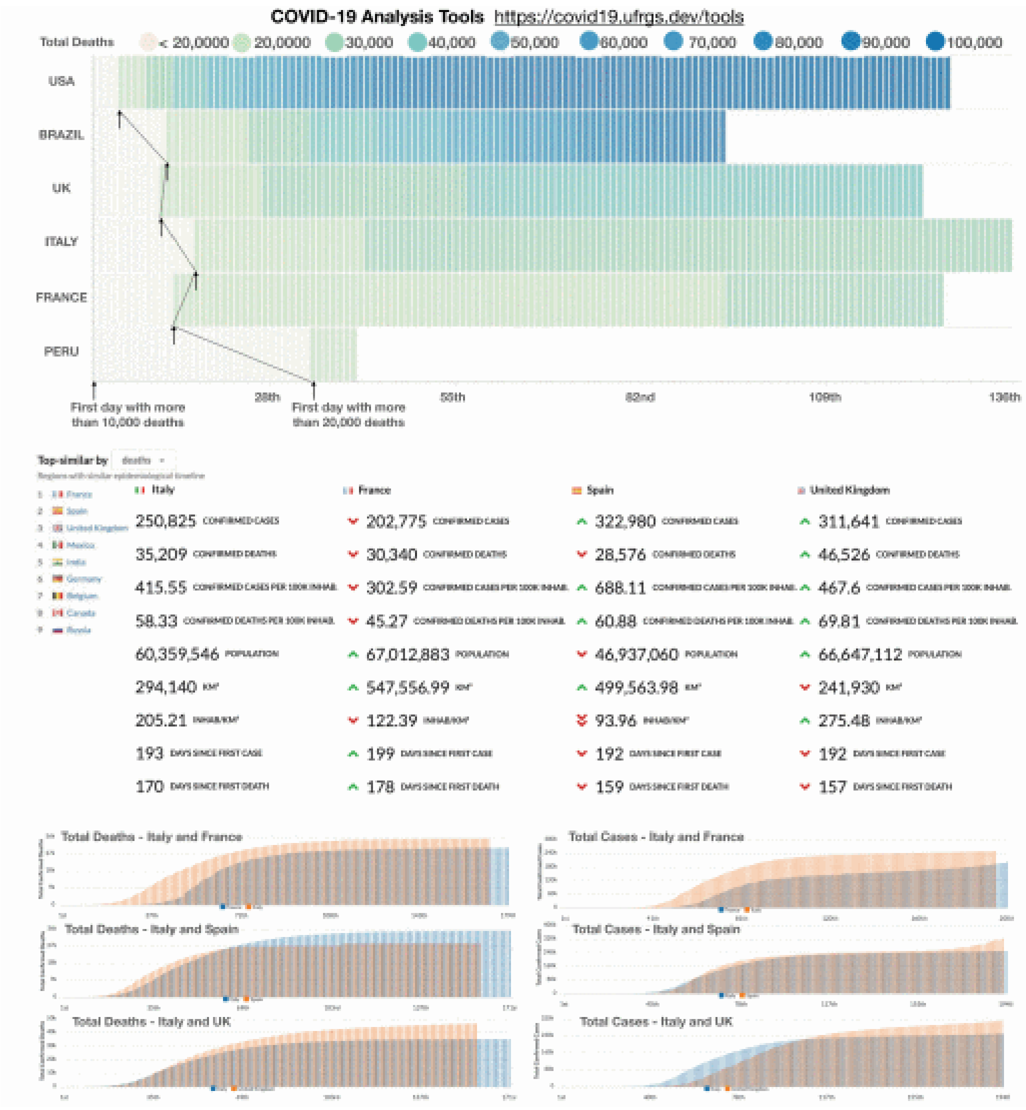
However, a study by B.J. Zheng, Y. Guan, K.H. Wong, J. Zhou, K.L. Wong, B.W Young, et al in 2004 reported that, only the samples isolated from the civets at the food market showed positive results for viral RNA detection, suggesting that the civet palm might be secondary hosts. In 2001 the samples were isolated from the healthy persons of Hongkong and the molecular assessment showed 2.5% frequency rate of anti-bodies against SARS-coronavirus. These indications suggested that SARS-coronavirus may be circulating in humans before causing the outbreak in 2003.

Moreover, in 2018, Eltahir, et al, Later on Rhinolophus bats were also found to have anti-SARS-CoV antibodies suggesting the bats as a source of viral replication. The Middle East respiratory syndrome (MERS) coronavirus first emerged in 2012 in Saudi Arabia, MERS-coronavirus also pertains to beta-coronavirus and having camels as a zoonotic source or primary host.

In a recent study by R. Lu in 2020 found that MERS-coronavirus was also detected in Pipistrellus and Perimyotis bats, Initially, a group of researchers suggested snakes be the possible host, however, after genomic similarity findings of novel coronavirus with SARS-like bat viruses supported the statement that not snakes but only bats could be the key reservoirs as shown in comparative data table. 1.

Features	**SARS-CoV**	**SARS-CoV-2**	**Year**
Emergence date	November 2002	December 2019	2003, 2004, 2020
Area of emergence	Guangdong, China	Wuhan, China	
Date of fully controlled	July 2003	Not controlled yet	

Features	SARS-CoV	SARS-CoV-2	Year
Key hosts	Bat, palm civets and Raccon dogs	Bat	2011, 2020
Number of countries infected	26	109	2020
Entry receptor in humans	ACE2 receptor	ACE2 receptor	2020
Sign and symptoms	fever, malaise, myalgia, headache, diarrhoea, shivering, cough and shortness of breath	Cough, fever and shortness of breath	2003,2020
Disease caused	SARS, ARDS	SARS, COVID-19	2003, 2020
Total infected patients	8098	123882	2020
Total recovered patients	7322	67051	
Total died patients	776 (9.6% mortality rate)	4473 (3.61% mortality rate)	


COVID-19 Analysis Tools https://covid19.ufrgs.dev/tools
Total Deaths
< 20,0000
20,0000
30,000
40,000
50,000
60,000
70,000
80,000
90,000
100,000
USA
BRAZIL
UK
ITALY
FRANCE
PERU
First day with more than 10,000 deaths
28th
First day with more than 20,000 deaths
55th
82nd
109th
136th
Total Deaths - Italy and France
Total Cases - Italy and France
Total Deaths - Italy and Spain
Total Cases - Italy and Spain
Total Deaths - Italy and UK
Total Cases - Italy and UK

Top-similar by deaths

Regions with similar epidemiological timeline

1. France
2. Spain
3. United Kingdom
4. Mexico
5. India
6. Germany
7. Belgium
8. Canada
9. Russia

Italy	France	Spain	United Kingdom
250,825 CONFIRMED CASES	▼ 202,775 CONFIRMED CASES	▲ 322,980 CONFIRMED CASES	▲ 311,641 CONFIRMED CASES
35,209 CONFIRMED DEATHS	▼ 30,340 CONFIRMED DEATHS	▼ 28,576 CONFIRMED DEATHS	▲ 46,526 CONFIRMED DEATHS
415.55 CONFIRMED CASES PER 100K INHAB.	▼ 302.59 CONFIRMED CASES PER 100K INHAB.	▲ 688.11 CONFIRMED CASES PER 100K INHAB.	▲ 467.6 CONFIRMED CASES PER 100K INHAB.
58.33 CONFIRMED DEATHS PER 100K INHAB.	▼ 45.27 CONFIRMED DEATHS PER 100K INHAB.	▲ 60.88 CONFIRMED DEATHS PER 100K INHAB.	▲ 69.81 CONFIRMED DEATHS PER 100K INHAB.
60,359,546 POPULATION	▲ 67,012,883 POPULATION	▼ 46,937,060 POPULATION	▲ 66,647,112 POPULATION
294,140 KM²	▲ 547,556.99 KM²	▲ 499,563.98 KM²	▼ 241,930 KM²
205.21 INHAB/KM²	▼ 122.39 INHAB/KM²	▼ 93.96 INHAB/KM²	▲ 275.48 INHAB/KM²
193 DAYS SINCE FIRST CASE	▲ 199 DAYS SINCE FIRST CASE	▼ 192 DAYS SINCE FIRST CASE	▼ 192 DAYS SINCE FIRST CASE
170 DAYS SINCE FIRST DEATH	▲ 178 DAYS SINCE FIRST DEATH	▼ 159 DAYS SINCE FIRST DEATH	▼ 157 DAYS SINCE FIRST DEATH

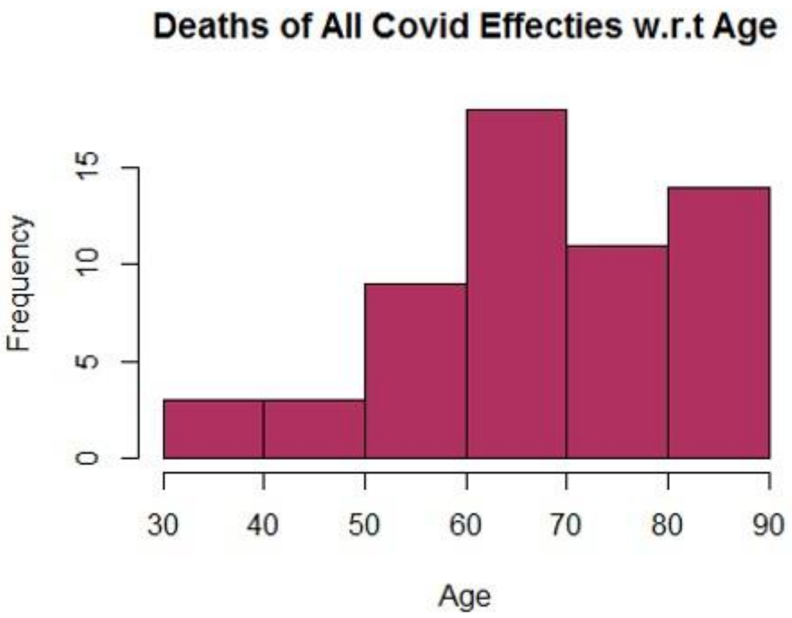
Figure 2.

Description: (Top) Heatmap matrices are useful for comparing time series such as the total deaths for different countries. Columns can be aligned by the first date after reaching a certain threshold, which allows us to compare when countries passed through specific checkpoints. (Bottom) Searching for places with similar timelines of deaths to Italy.

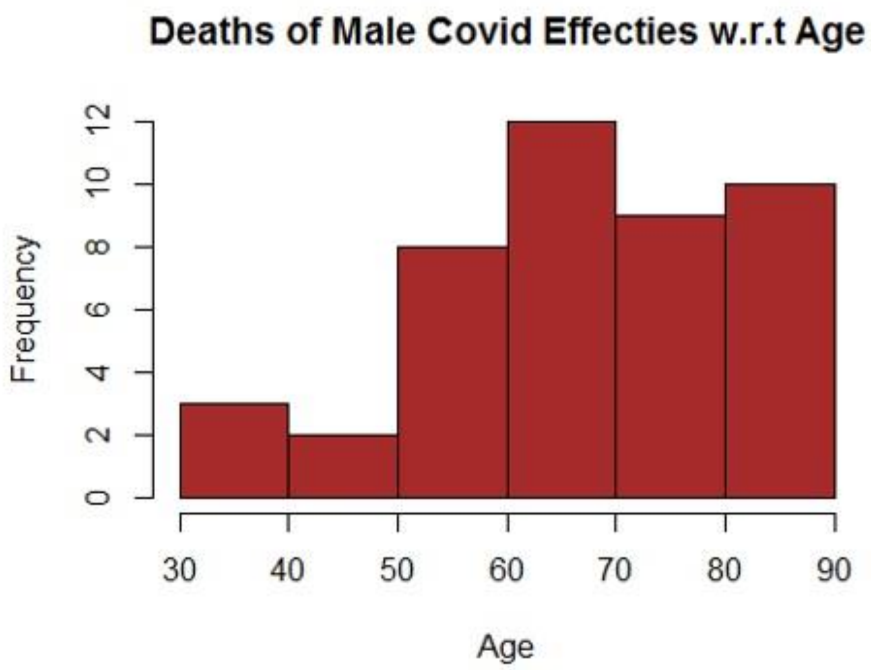
A vast collection of community-developed dashboards and interactive tools about COVID-19 are available. Good starting places to look are the data hub hosted by Tableau and the top 100 R-resources organized by Soetewey. Moreover, In-depth analysis is available at sites, such as *Our World in Data*, *Bing*, and the *COVID Tracking Project*, among others. After developing the Brazilian dashboard, we devoted our efforts to create a set of tools to compare the spread of COVID-19 data in different regions of the world. We collected data from a website, and that interface has different charts that support visualizing multiple locations in a single chart. Since the pandemic is at different stages in the world, article allow the user to align the time-series of data by a certain data the series passes a given threshold (e.g., after 100 cases). This representation is useful to observe when different locations passed through specific checkpoints taken from (covid19.ufrgs.dev).

Visualization of Covid 19 Data

Visualizations had always been easy to understand the raw data. Here we are going to compare the gender related to Covid effectives with age group as histogram shows that counts of overall death increases with the increase of age. Thus, age factor imposes a great impact on death due to COVID virus. Therefore, we can construct a result that greater the age, greater will the chance of death due to virus, Table 2.



Furthermore, the histogram in Table 3, shows that male counts of death increase with the increase of age. Thus, age factor imposes a great impact on death due to COVID virus. Therefore, we can construct a result that greater the age, greater will the chance of death due to virus.



Another Histogram in Table 4. indicates that female counts of death increases with the increase of age. Thus, age factor imposes a great impact on death due to COVID virus. Therefore, we can construct a result that greater the age, greater will the chance of death due to virus.

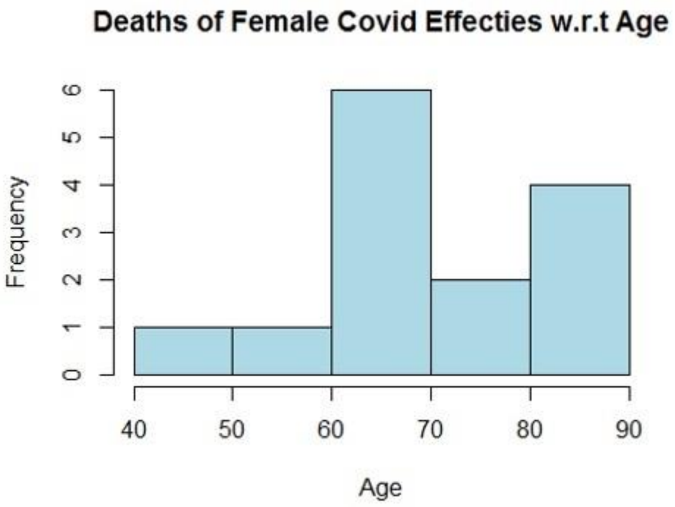
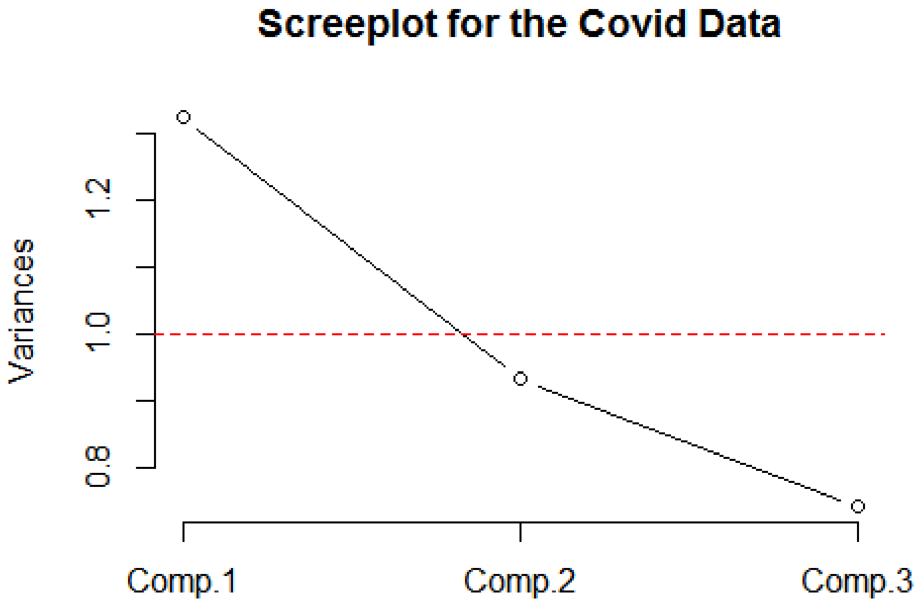
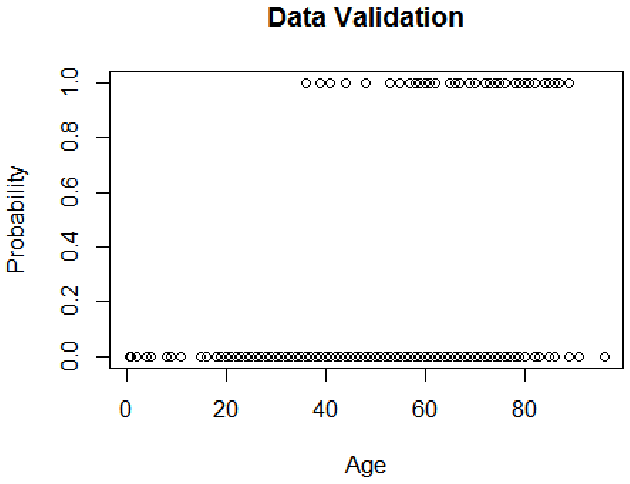


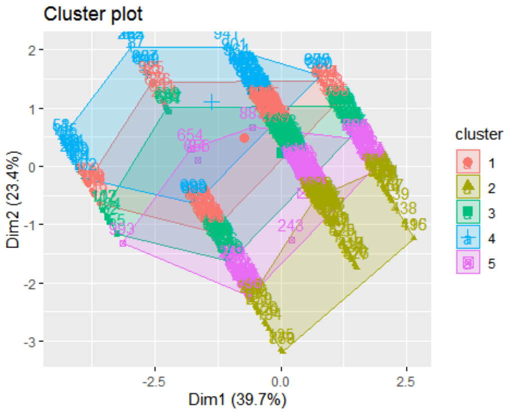
Table 5, Scree plot here indicates the out of 3 principal components i.e. (i) age, (ii) visiting Wuhan and (iii) from Wuhan, data majorly shows deviation from the Age components. It shows that whole data majorly variates with the age variable.



Apart from Regression, figure 3. shows that the data does not have any error. Patient/effected person is either death due to virus contraction or survived after contracted the virus. Also, it shows that majority of the patients were recovered from virus while the death ones are concentrated in older ages above 35-years.



K-mean clustering in Fig 4. shows the deviation of the prescribed 04 principal components from their centroid values i.e., Age, Visiting Wuhan, From Wuhan and Deaths. 5 clusters are made and displayed on right side of the cluster plot. To clean the data that included unoccupied column, we used to omit command which ultimately deletes the rows with one or more vacant columns.



Conclusion

COVID-19 is swiftly spreading worldwide. Within a few months, the mortality rate and morbidity rate has reached unexpected levels. The clinicians are working to invent treatments and vaccine to prevent this infection. The extreme situation is yet to occur. However, if we take one step toward self-isolation, it could save the entire community and the risk will decline immediately. This is a situation where each individual has to take steps toward minimizing the risk by staying in the house and immobilizing themselves. The airborne, contact transmission can only be disinfected if proper handwashing protocols are followed and each individual carry out precautionary measures to safe other individuals from this debilitating virus. World has a tremendous potential in public health, and different sectors can work together to address the challenges by the engagement of society and community along with policy initiatives.

Impact of COVID-19 by Age

This study aims to synthesize the existing research on the impact of the COVID-19 pandemic, and associated isolation and protective measures, on older adults. The secondary objective is to investigate the impact of the COVID-19 pandemic, and associated isolation and protective measures, on older adults with Alzheimer disease and related dementias.

While COVID-19, the disease caused by the new coronavirus, can lead to hospitalization and even death for young and middle-aged adults, it has caused the most severe health issues for adults over the age of 60 — with particularly fatal results for those 80 years and older. This is due in no small part to the number of underlying health conditions present in older populations. Conditions like diabetes, heart disease, and other chronic illnesses can lead to more intense symptoms and complications in the disease. Additionally, as people age, their immune system gradually loses its resiliency, meaning that they are more susceptible to infection of any kind, especially a new one like COVID-19.

Older people are clearly among the groups most at risk of serious illness and death from COVID-19. Data from China following the initial few months of the outbreak in Wuhan showed the risks of serious illness and death from COVID-19 for people in older age, with a fatality rate of close to 15 per cent in people aged 80 and over, compared to an overall rate of 2.3 per cent across the population.

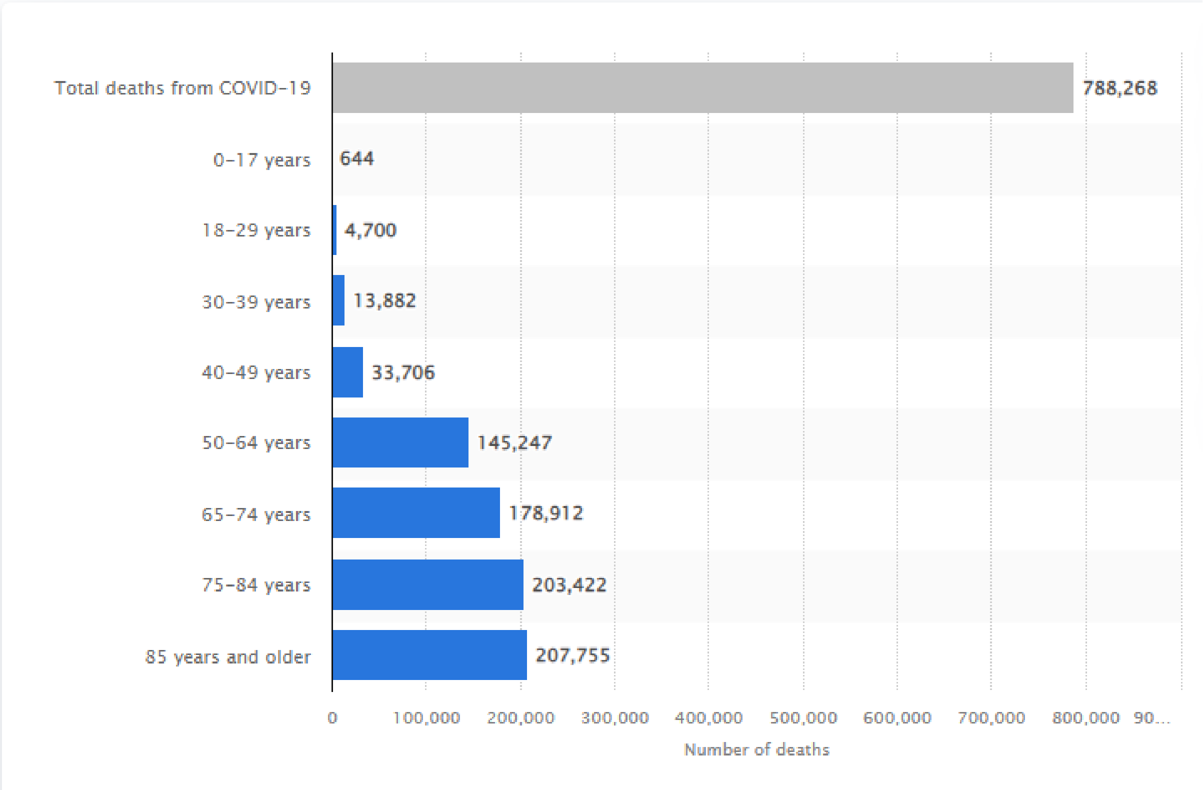
Age is *not* just a number

Older adults are the main victims of the novel COVID-19 coronavirus outbreak and elderly in Long Term Care Facilities (LTCFs) are severely hit in terms of mortality. This REPORT presents a quantitative study of the impact of COVID-19 outbreak in Wuhan during first stages of the epidemic, focusing on the effects on mortality increase among older adults over 80 and its correlation with LTCFs. The study of growth patterns shows a power-law scaling regime for the first stage of the pandemic with an uneven behaviour among different regions as well as for the overall mortality increase according to the different impact of COVID-19. However, COVID-19 incidence rate does not fully explain the differences of mortality impact in older adults among different regions. We define a quantitative correlation between mortality in older adults and the number of people in LTCFs confirming the tremendous impact of COVID-19 on LTCFs. In addition a correlation between LTCFs and undiagnosed cases as well as effects of health system dysfunction is also observed. Our results confirm that LTCFs did not play a protective role on older adults during the pandemic, but the higher the number of elderly people living in LTCFs the greater the increase of both general and COVID-19 related mortality. We also observed that the handling of the crises in LTCFs hampered an efficient tracing of COVID-19 spread and promoted the increase of deaths not directly attributed to SARS-CoV-2.

Although we should be very cautious in dichotomizing age when it comes to deciding who needs intensive care by giving the resources to young people and not to the elderly age is clearly the most important factor in predicting the odds of surviving the COVID-19 disease. In fact, this is among the most robust consensus among scientists. In an ideal world, that alone should justify the expectation that older people will be more dutiful in terms of following public health recommendations. However, the issue is not that straightforward.

COVID-19 is not the only disease for which the consequences are known to be the worst among older people. This is the case for invasive pneumococcal disease or heat stroke, among others. In both cases, there are existing preventive measures.

That said, the mortality rate of COVID-19 among elderly people is, objectively speaking, much greater than other diseases such as pneumococcal disease or heat stroke. Moreover, the governments' and citizens’ reactions and measures to minimize the consequences of that disease are of a totally different scale. Hence, it is very reasonable to expect that older people will be more dutiful when it comes to following the recommendation of public health agencies and governments. Following graph shows number of deaths according to age group.



References

A. Soetewey, "Top 100 R resources on novel COVID-19 coronavirus", [online] Available: https://www.statsandr.com/blog/top-r-resources-on-covid-19-coronavirus.

B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, *et al.* **Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms.** J Virol, 79 (18) (2005), pp. 11892-11900

B.J. Zheng, Y. Guan, K.H. Wong, J. Zhou, K.L. Wong, B.W.Y. Young, *et al.* **SARS-related virus predating SARS outbreak, Hong Kong** Emerg Infect Dis, 10 (2) (2004), p. 176

C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, *et al.* **Clinical features of patients infected with 2019 novel coronavirus in Wuhan China.** The Lancet (2020)

C. Paden, M. Yusof, Z. Al Hammadi, K. Queen, Y. Tao, Y. Eltahir, *et al.* **Zoonotic origin and transmission of Middle East respiratory syndrome coronavirus in the UAE.** Zoonoses Public Health, 65 (3) (2018), pp. 322-333

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Yu T. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. The Lancet. 2020; 395(10223):507-513.

Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analysing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis (EDA) Approach. Journal of Medical Virology; 2020.

F. Pontes, F. Pinto, W. Leuschner and J. L. D. Comba, "COVID-19 analysis tools at Instituto de Informática-ufrgs", [online] Available: https://covid19.ufrgs.dev

Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. International Journal of Infectious Diseases. 2020; 91:264–6.

J.F.-W. Chan, S. Yuan, K.-H. Kok, K.K.-W. To, H. Chu, J. Yang, *et al.* **A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.** Lancet (2020).

N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K. Chan, *et al.* **Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003.** The Lancet, 362 (9393) (2003), pp. 1353-1358

R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, *et al.* **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.** The Lancet (2020)

Rajan Gupta, Saibal Kumar. Trend Analysis and Forecasting of COVID-19 outbreak in India. Rajan Gupta, Saibal Kumar PalmedRxiv, 2020.

Sohrabi C, Alsafi Z, O’Neill N, Khan M, Kerwan A, Al-Jabir A, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery; 2020.

Tian X, Li C, Huang A, Xia S, Lu S, Shi Z, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. bioRxiv; 2020.

Vara V. Coronavirus outbreak: The countries affected. 11 MARCH 2020; Available from: https://www.pharmaceutical-technology.com/features/coronavirus-outbreak-the-countries-affected/.

Yoo JH. The fight against the 2019-nCoV outbreak: an arduous march has just begun. *J Korean Med Sci*. 2020;35: e56 10.

Appendix

Code:

```
rm(list=ls())
library(Hmisc)
library(factoextra)
library(dplyr)


data <- read.csv("C:/Users/ARSALAN
IQBAL/Downloads/COVID19_line_list_data.csv")
data <- select(data,-c(X.1:X.6))
data <- select(data,-c(X))
describe(data)
summary(data)
data_copied <- data
data_copied$death_dummy <- as.integer(data$death !=0)

data_copied <- select(data_copied,-c(2,9,10,11,12,13,18,19))
data_copied <- na.omit(data_copied)

death_all = subset(data_copied,death_dummy == 1)
alive_all = subset(data_copied,death_dummy == 0)

hist(death_all$age)

men <- subset(data_copied, gender == "male")
women <- subset(data_copied, gender == "female")
mendeath <- subset(men,death_dummy == 1)
womendeath <- subset(women,death_dummy == 1)

menalive <- subset(men,death_dummy == 0)
womenalive <- subset(women,death_dummy == 0)

mean_mendeath <- mean(men$death_dummy)
mean_womendeath <- mean(women$death_dummy)

mean_age_death <- mean(death_all$age)
mean_age_alive <- mean(alive_all$age)

hist(mendeath$age, col='brown')
hist(womendeath$age, col='light blue')


t.test(alive_all$age, death_all$age, alternative="two.sided",
conf.level = 0.95)

t.test(men$death_dummy, women$death_dummy, alternative="two.sided",
conf.level = 0.99)

#Principle Component Analysis

head(data_copied)
data_copiedpc <- data_copied[,c(7,8,9)]
pc.data <- princomp(data_copiedpc, cor = TRUE)
```

```
names(pc.data)
summary(pc.data)
eigenvectors <- pc.data$loadings
eigenvalues <- pc.data$sdev *pc.data$sdev
screeplot(pc.data,type="l", main="screeplot for the covid data")
abline(1,0,col= 'red',lty=2)

#REGRESSION

plot(data_copied$age,data_copied$death_dummy)

#CLUSTERING

data_ultra <- select(data_copied,-c(1,2,3,4,5,6,10,11,12))
data_ultra <- na.omit(data_ultra)

km <- kmeans(data_ultra, centers = 5, nstart = 100)
fviz_cluster(km, data = data_ultra)
```