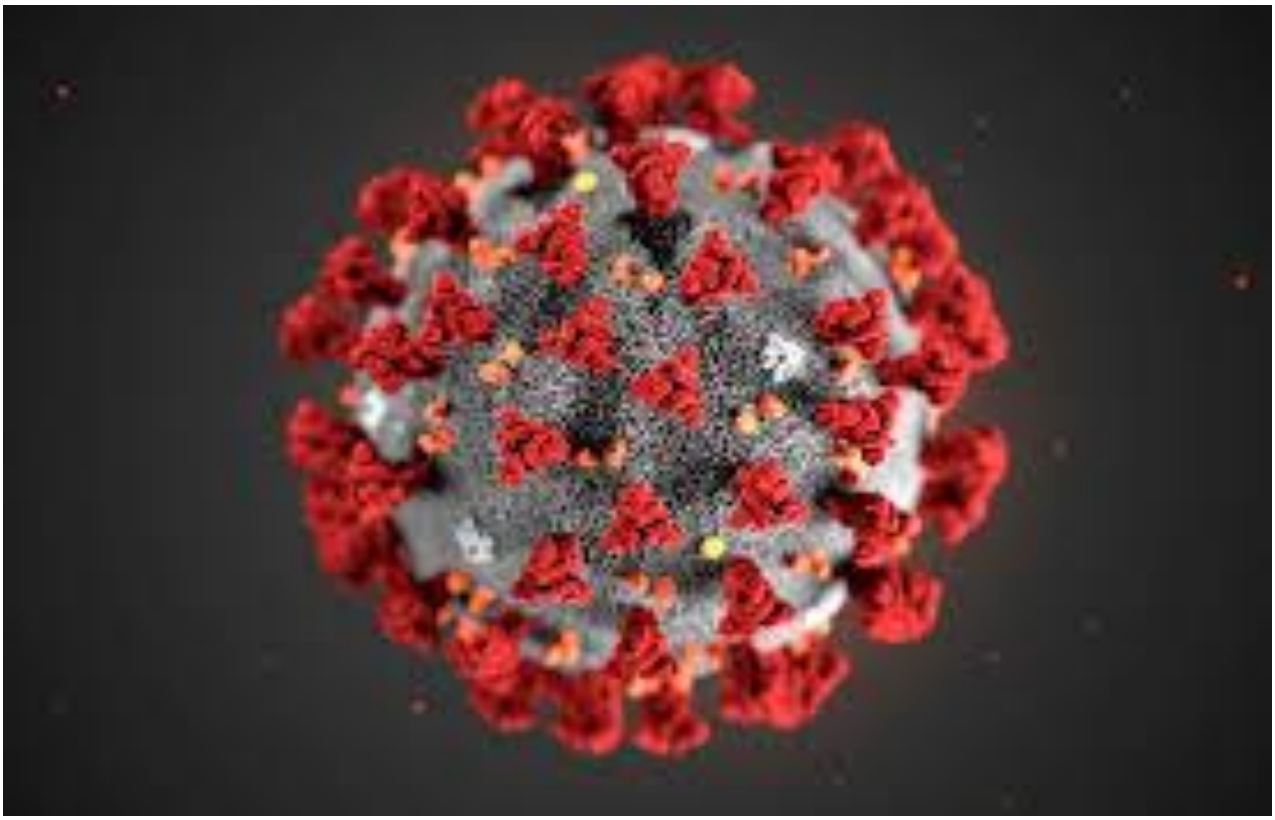


CORONA VIRUS PANDEMIC: STATISTICS AND RESEARCH

IS665 Data Mining Warehousing & Visualization Project: Phase One

Jatin Bhowad, Meenal Sawant, Komal Ghugare and Pradeepti Upadhyayula

PACE UNIVERSITY

Table of Contents

PART ONE

INTRODUCTION

According to the research, coronavirus is a group of related RNA viruses that cause diseases in mammals and birds. In humans, these viruses cause respiratory track diseases that can range from mild to lethal. There are yet no vaccines or antireal drugs to prevent or treat human coronavirus infections. Insight into the spread of the disease can help leaders respond more effectively to the epidemic. A recent article in *Forbes* mentioned that analytics can help authorities identify the most vulnerable communities. We are facing a global crisis, however, as data science has evolved, the decisions leaders will make over the coming weeks will shape the world for years to come. From a public health perspective, to combat an epidemic, officials must take several actions, such as: build awareness, set guidelines for health professionals, target infection clusters, limit population movements, and allocate scarce resources. These decisions will influence how many people will survive and how many will die over the coming days, weeks and months. Leaders must act quickly and decisively in order to save lives.

For our project we will conduct exploratory analysis on Covid 19 dataset for January to April (2020) globally. It's so hard to see into the future of Covid 19. The most difficult thing for an epidemiological model to predict human behavior. Dr. Brian Monahan, the attending physician of the U.S. Congress, predicts 70 million-150 million U.S. coronavirus cases are expected. In order, to make every best possible way to reduce this virus to spread data needs to be analyzed.

How it helps people and decision leaders?

Awareness among the society is important since it does not a cure yet and would help in controlling the spread of virus. Decisions leaders will make over the coming weeks will shape the world for years to come. The public health leaders who make the hard choices are lacking high quality, high resolution data on key questions, such as: Where is the disease likely to spread? Are there priority areas that we need to contain to limit further propagation? Where are the most vulnerable communities? Leaders must act quickly and decisively in order to save lives.

Why data analysis is needed?

Data analysis is needed for this since data is a powerful tool to be used alongside traditional scouting methods to find the best solutions to avoid the spread of disease and prevent getting infected. Data analysis on Covid 19 statistics, pulls out insights and information about the areas highly infected and the insights that might be responsible for the pandemic that you would not spot easily without visualization. Having a lot of covid19 data stats does not help anyone unless you analyze the data and use it to make decisions.

How Covid19 data analysis will improve decision making?

Government will have to undergo several changes in the rules and regulation, set guidelines in public places, travel advisory .From a public health perspective, to combat an epidemic, officials would be able to such as build awareness among people, set guidelines for health professionals, target infection clusters, limit population movements, and allocate scarce resources. These decisions will influence how many people will survive and how many will die over the coming days, weeks and months.

DATA

Data Source: We perceived the dataset of coronavirus from GitHub. So, this dataset contains the count of recovered, deaths, confirmed cases from January 2020 to April 2020 classified on categorized variables which helped us in getting the insights of the source for the disease, when did the disease spread, impact on people and areas globally. We selected this source because it has stats aggregated to the countries globally with attributes that would be helpful for our analysis.

Significant Data Variables:

Significant variables in our analysis include: Confirmed, Deaths, Recovered, Date, HealthCare Index Rank, Total Population, Median Age.

Country	168 countries in the world
Date	Range from 01/21/2020:04/07/2020
Confirmed	Infected people by Covid19
Deaths	Count of people caused death due to Covid19
Recovered	Count of people recovered from Covid19
Total Population	Population of the country consists of all persons falling within the scope of the census
Population Density	The number of people per square mile of land area
Coastal/Non-Coastal	Countries classified based on coastal areas
Median Age	Average of every country
Health Care Index	Ranking of countries based on their health facilities
Classification	Based on Developed, Under-Developed, Developing

Source: Datasets sources: https://github.com/datasets/covid-19/blob/master/data/countries-aggregated.csv

Country Population: https://www.worldometers.info/world-population/population-by-country/

Healthcare system rank: https://photius.com/rankings/healthranks_alpha.html

Classification:https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf

Age median of countries: https://en.wikipedia.org/wiki/List_of_countries_by_median_age

Coastal/Non-coastal countries : https://en.wikipedia.org/wiki/List_of_countries_by_length_of_coastline

DESCRIPTIVE ANALYSIS:

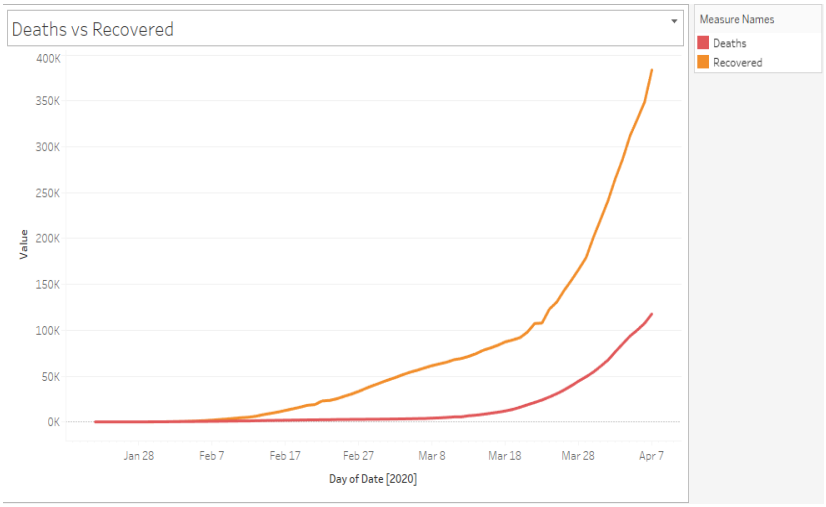
The mean death rate of people from season January to April is 88.91

The mean confirmed rate of people from season January to April is 1735

The mean recovered rate of people from season January to April is 403.8

EXPLORATORY ANALYSIS:

1) Trend of deaths vs recovered cases of COVID-19

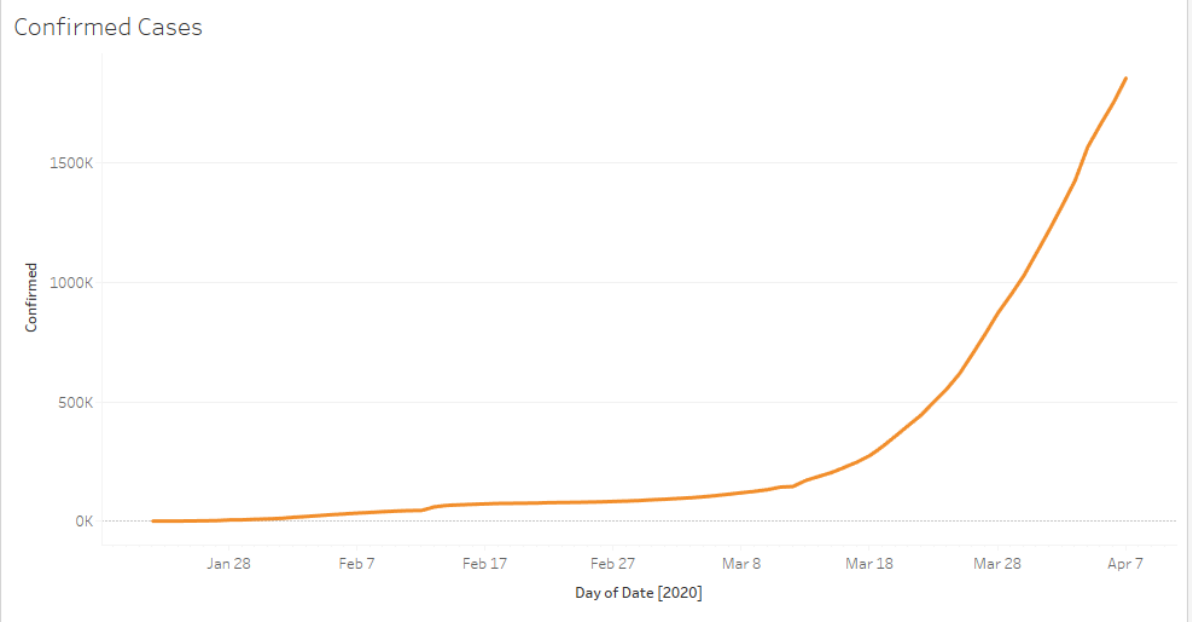


We can see that the number of deaths and recovered cases has had an exponential rise post from Feb 2020.

Probable Causes:

- Pre-existing medical conditions of the patients.
- The ones with good immunity and proper medical care have had recoveries.
- No vaccine discovered yet.
- Patient promptly getting tested as and when symptoms showed up.

2) Trend of confirmed cases

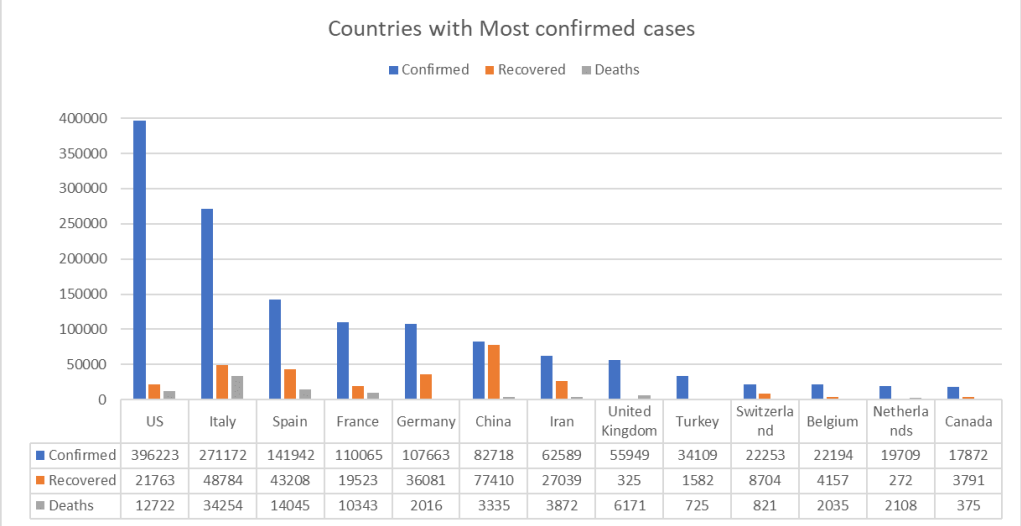


We can see that the number of confirmed cases has been on an exponential rise post from Mar 2020.

Probable causes:

- Prompt testing
- Less awareness of social distancing
- Delayed shutdown in some countries

3) Which countries are most affected by COVID – 19 ?

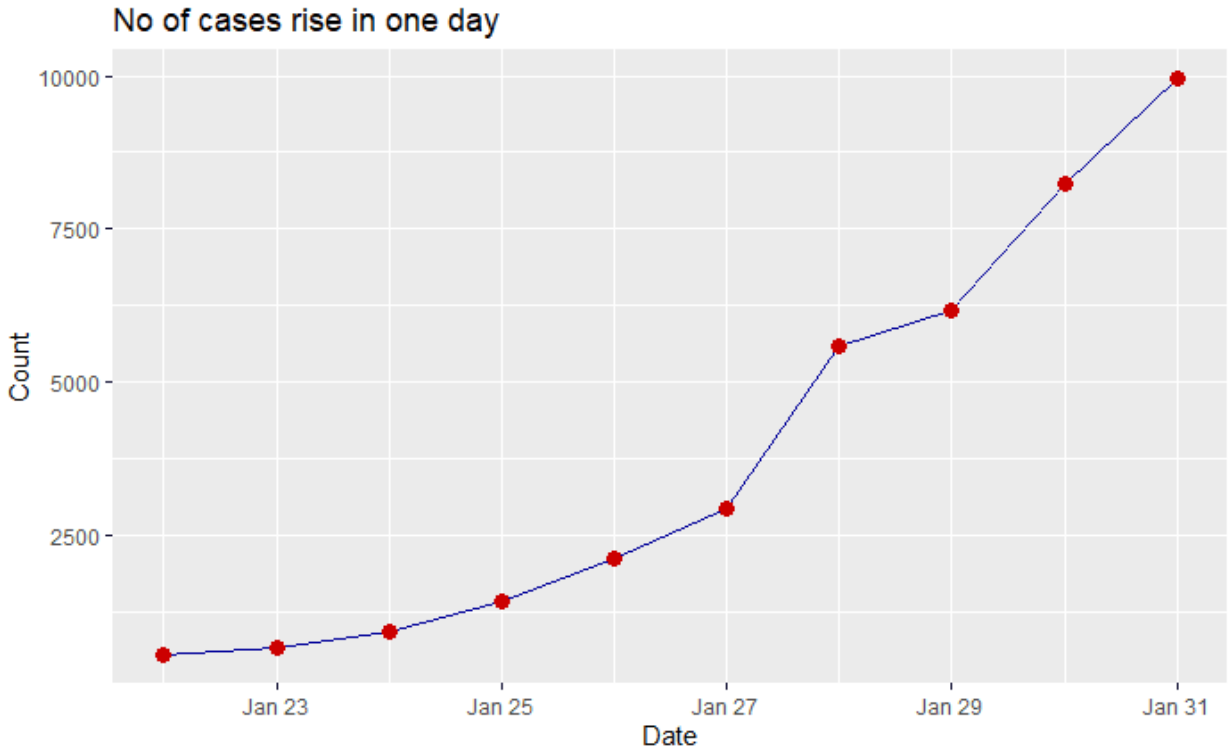


	US	Italy	Spain	France	Germany	China	Iran	United Kingdom	Turkey	Switzerland	Belgium	Netherlands	Canada
Confirmed	396223	271172	141942	110065	107663	82718	62589	55949	34109	22253	22194	19709	17872
Recovered	21763	48784	43208	19523	36081	77410	27039	325	1582	8704	4157	272	3791
Deaths	12722	34254	14045	10343	2016	3335	3872	6171	725	821	2035	2108	375

Summary

- US has the most confirmed cases, followed by Italy and Spain.
- China has the most recovered cases.
- Italy has the most deaths.

4) COVID – 19 Number of cases increasing per day globally

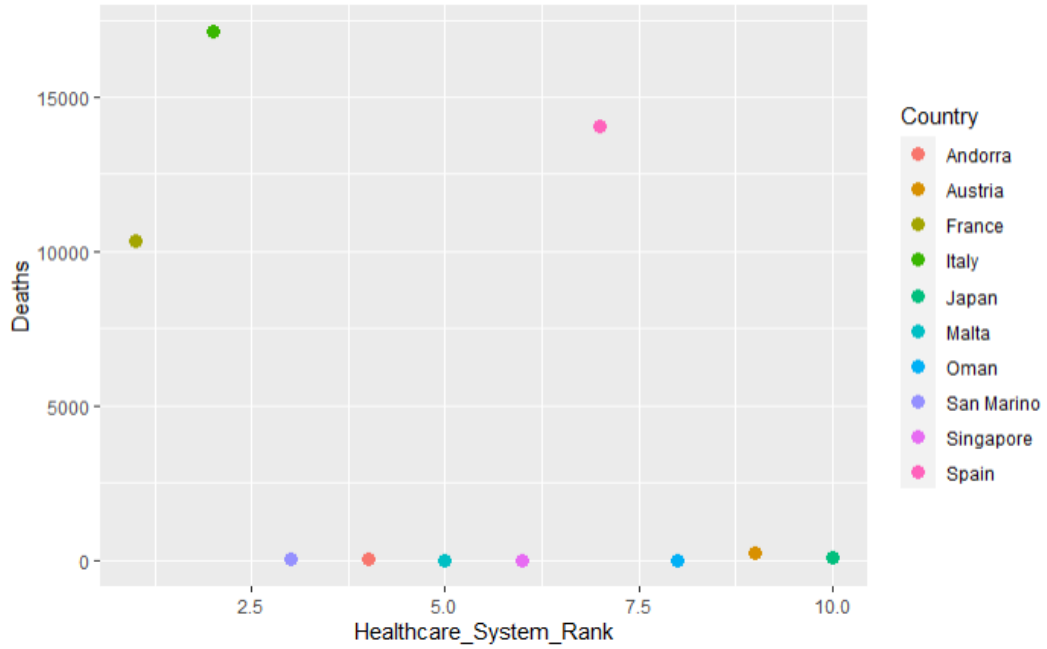


There has been a steep rise in the cases from Jan 22nd globally

Probable causes:

- More International travel
- Incubation period of 14 days
- Lack of early preventive measures by the governments from Jan 2020

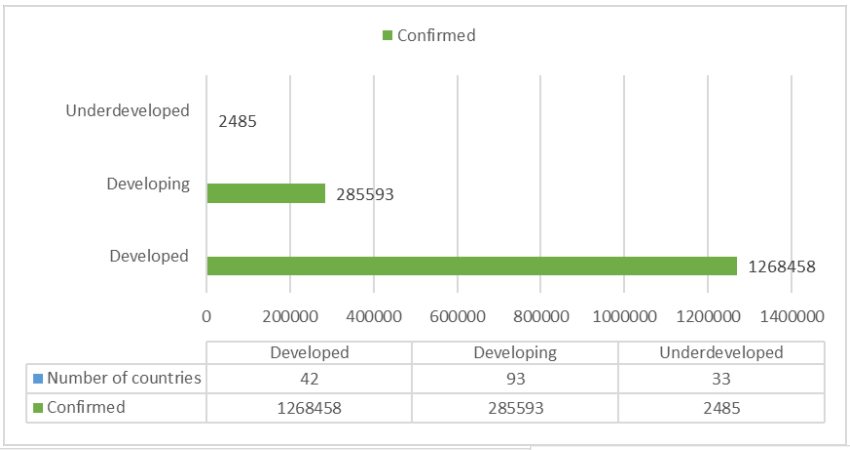
5) Amongst the top 10 countries ranking in healthcare index which country has highest death rate?



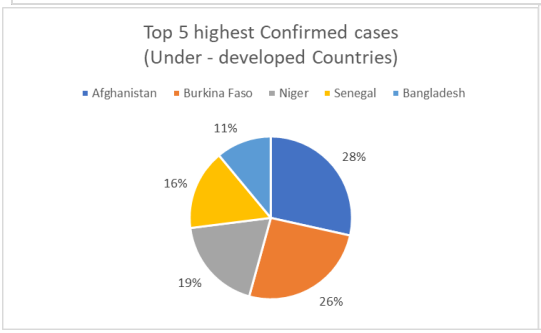
Answer: France

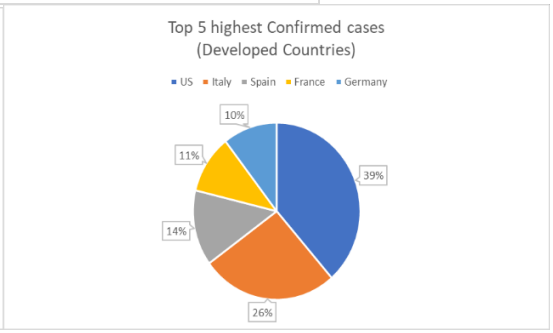
France and Italy after China are observed with the highest death rate since researchers have found they have the oldest population and being a developed country with a good healthcare index draws attraction of tourists and opportunities resulting into populated area which can be lead to a lot of human interaction and more human interaction.

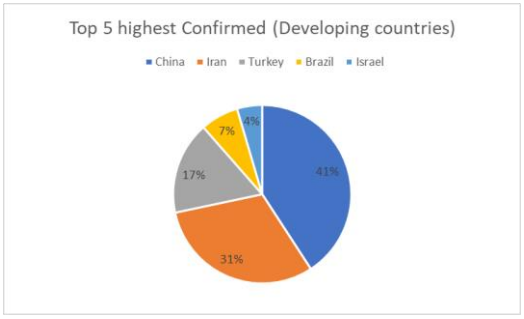
6) COVID – 19 confirmed cases based on Development Classification



	Developed	Developing	Underdeveloped
Number of countries	42	93	33
Confirmed	1268458	285593	2485



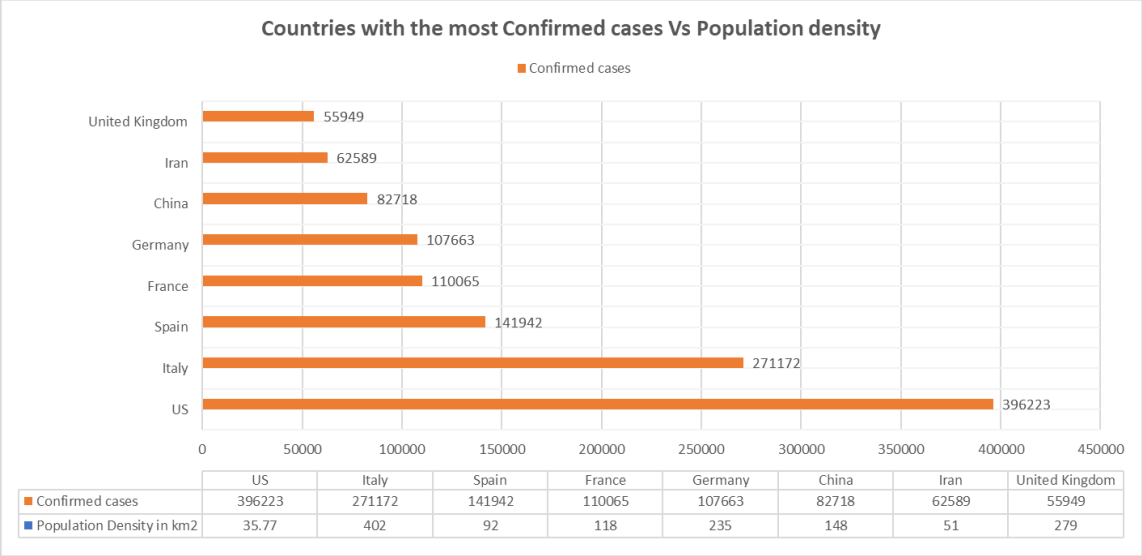




US, China, Afghanistan have the most confirmed cases in the three categories respectively.

Reasons: International travel for tourism or business purpose was observed more when the virus was spreading globally which led the virus to travel and infect more people.

7) How is the population density of a country influencing cases?



	US	Italy	Spain	France	Germany	China	Iran	United Kingdom
Confirmed cases	396223	271172	141942	110065	107663	82718	62589	55949
Population Density in km2	35.77	402	92	118	235	148	51	279

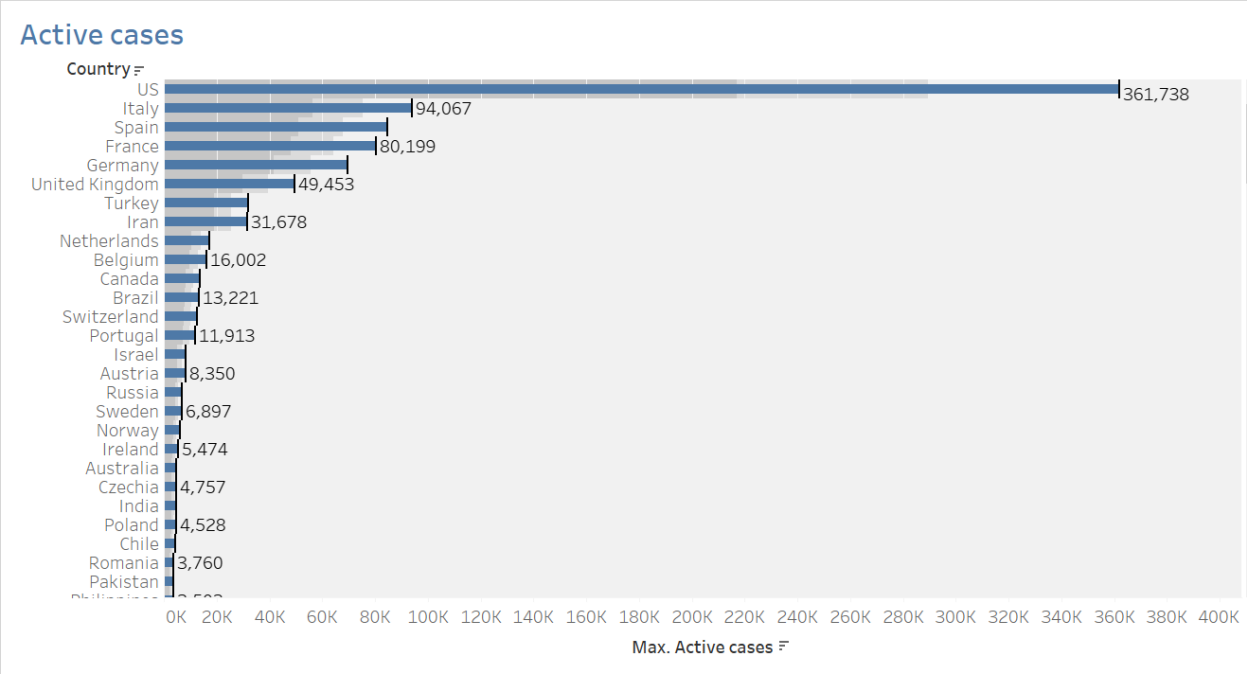
Summary

- For a fast spreading communicable disease, we found no direct relationship between the population density and number of cases in a country.

Probable causes could include:

- Difficulty in practicing social distancing in densely populated cities
- Lack of proper standards/measures in testing

8) Which country has the highest number of active cases?



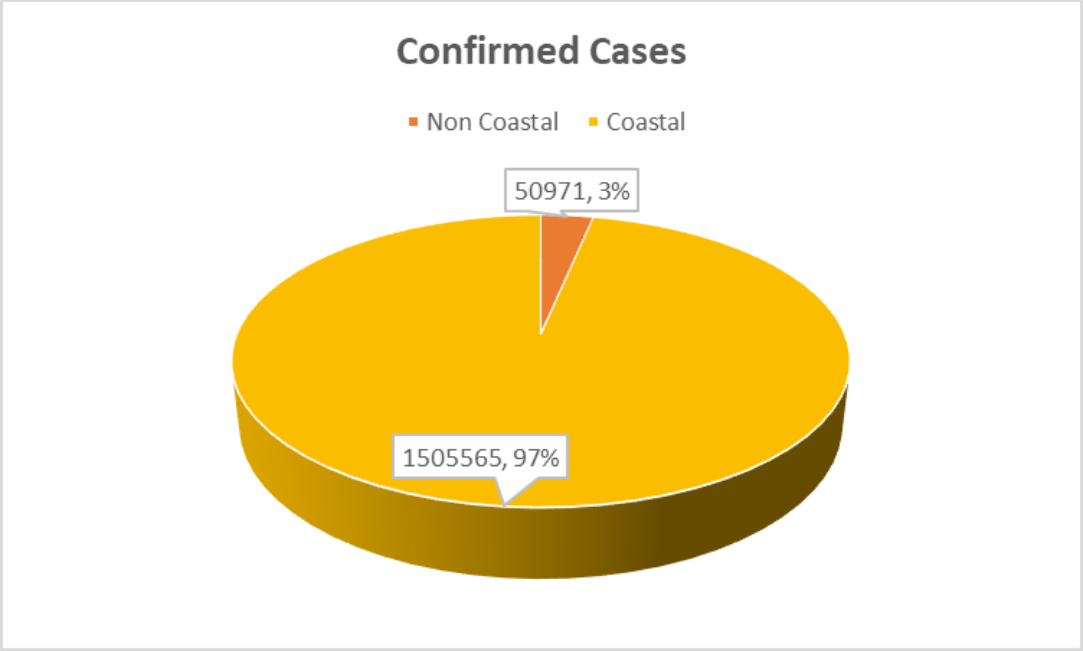
Summary

- Active cases = Confirmed – (death+ recovered)
- US has the most confirmed cases, followed by Italy and Spain.

Causes

- Late shutdown
- Testing at airports was not done on travelers from all the countries initially
- Lack of checks on travel history
- Lack of necessary medical equipment

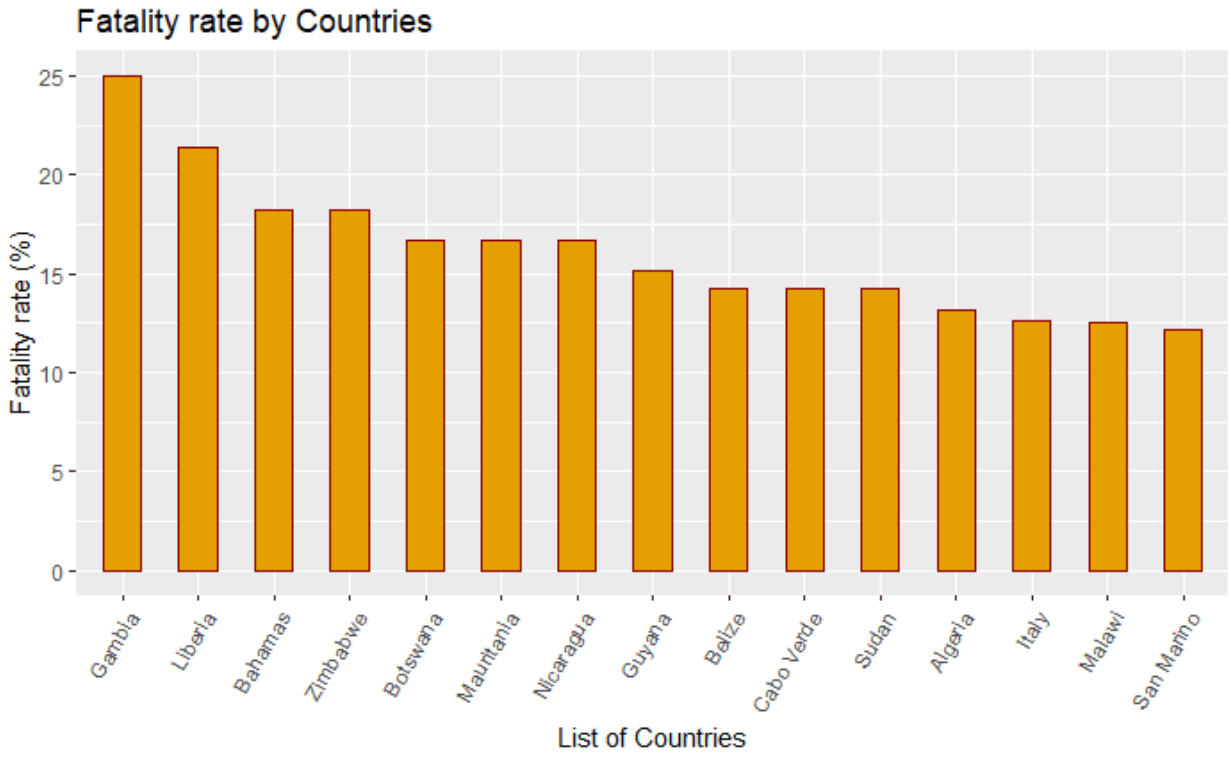
9) Confirmed Cases by Coastal/Non-coastal classification



Observations and Causes:

- Most developed countries are costal
- Humidity also may contribute to the virus spread
- Cruise ships landing in costal countries

10) Which country has the highest fatality rate?

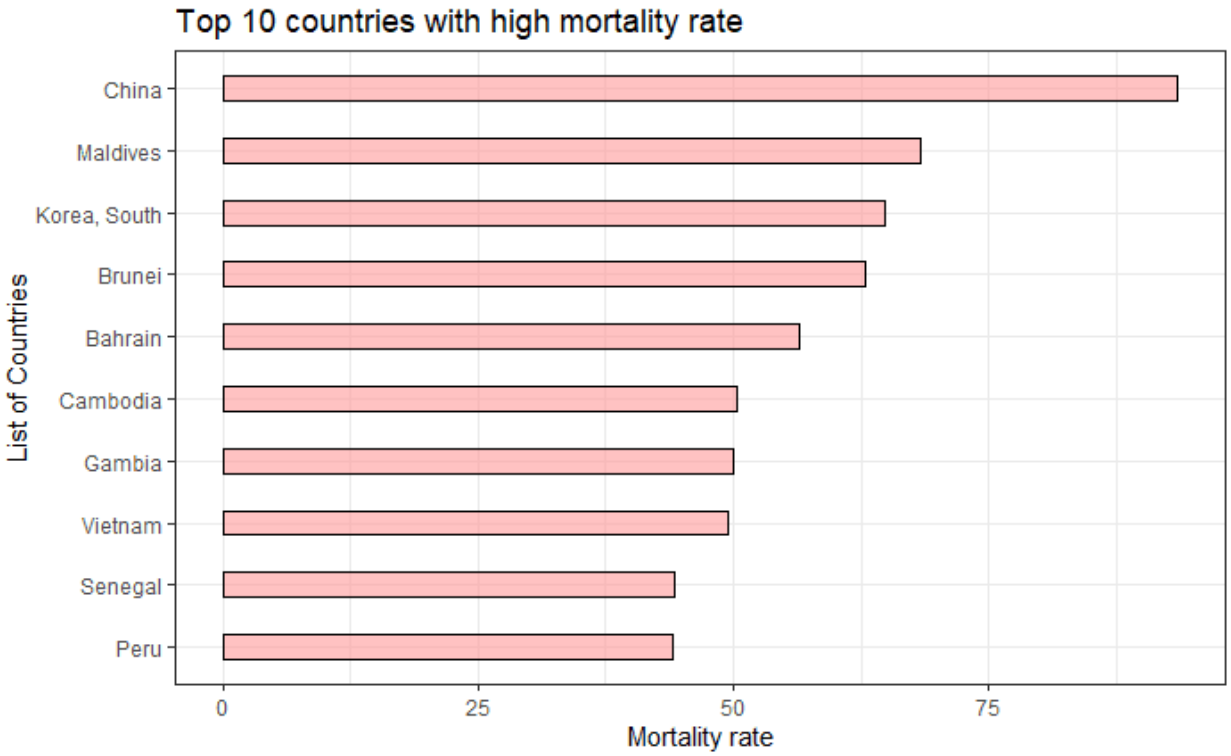


Fatality rate = (Death/confirmed) *100

Gambia has the highest no of death count vs confirmed count.

Lack of necessary medical facilities.

11) Which country the highest mortality rate?



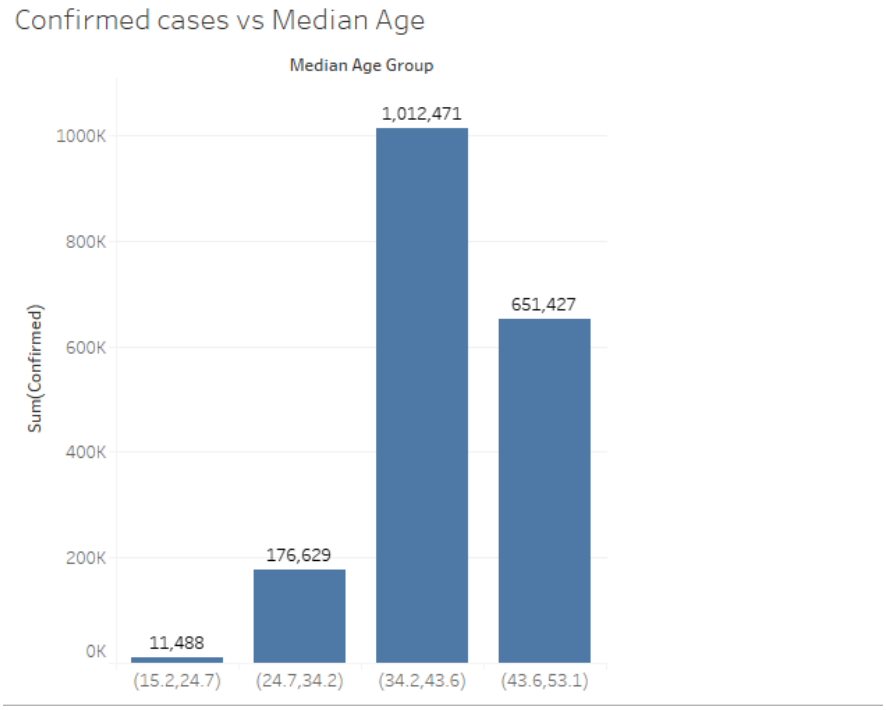
Answer: China

Mortality rate = (Recovered/confirmed) *100

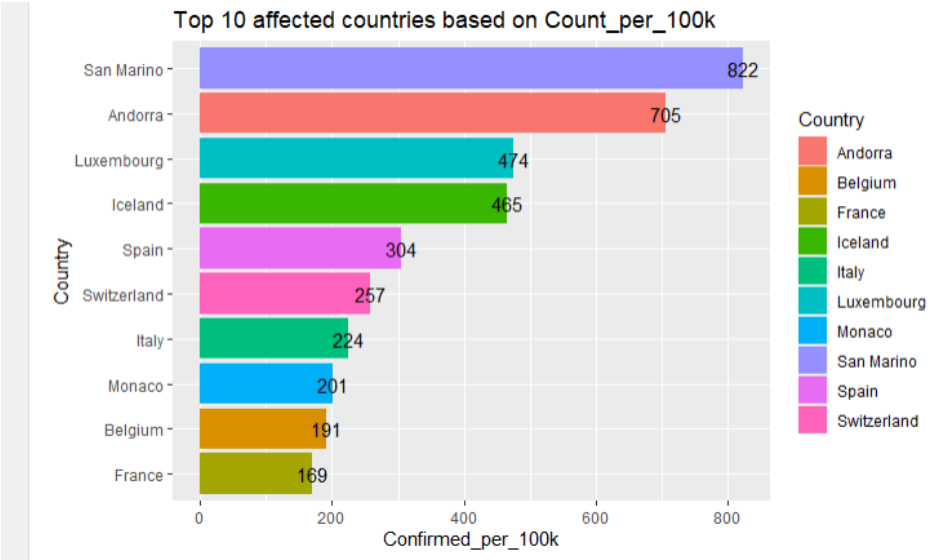
Probable causes:

- China has the highest recovery rate may be because of the intense lockdown and social distancing measures

12) How is the age affecting the number of confirmed cases?



13. Countries with most affected cases based upon count per 100k people?



Data calculated according to per 100k we observe that San Marino is highly affected due to Covid 19.

CONCLUSION:

- Developed countries have experienced the highest number of cases.
- Countries with a good healthcare index are worst hit, but are trying to flatten the curve with strict lockdowns and social distancing measures.
- Age is a great contributing factor. People of 34 years and older are at a higher risk. Pre-existing medical conditions can cause more complications.
- Population density surprisingly has not been a contributor for the widespread cases. Countries like US, Italy, Spain with very less Population density are the worst hit.
- Coastal countries have recorded more cases than non-coastal ones.
- Governments should not re-open the states/countries soon as there could be a chance of a second wave of infections.
- We can learn from some countries on how to implement strict lockdown guidelines, so that we achieve good recovery rate and reduce the burden on healthcare system.