IS 665 - Exploratory Analysis on COVID-19

By Group Fast and Serious

Jatin **Meenal** **Komal** **Pradeepti**

Contents

- COVID – 19 Global Timeline
- About the Dataset
- Descriptive Statistics
- Exploratory Analysis

COVID – 19 Global Timeline

Dec - Jan

- **31 Dec** China reported a cluster of cases of pneumonia in Wuhan
- **01 Jan** Wuhan Seafood Market closes
- **11 Jan** China records its first death
- **13 Jan** First case outside of China(Thailand)
- **20 Jan** First US case reported in Washington
- **30 Jan** WHO declares a global public-health emergency

Feb - Mar

- **09 Feb** Death toll in China surpasses that of the 2002-2003 SARS epidemic, with 811 deaths recorded
- **19 Feb** Iran outbreak begins
- **21 Feb** Italy outbreak begins
- **11 Mar** WHO characterized COVID-19 as a pandemic
- **23 Mar** NYC confirms 21,000 cases, making it the biggest epicenter of the outbreak in the US

Apr - May

- **02 Apr** The world passes 1 million COVID-19 infections
- **10 Apr** The global death toll surpasses 100,000
- **15 Apr** Number of global COVID-19 cases surpasses 2 million
- **22 Apr** US reports the highest single-day death toll for any country: more than 2,600
- **01 May** Roughly 3.3 million confirmed cases of COVID-19, with about 1.2 million being active and ongoing cases, roughly 1 million recoveries, and 234,000 deaths

About the Dataset

Variables in the dataset

- Country
- Date
- Confirmed cases
- Recovered cases
- Deaths
- Classification of the country (Developed, Underdeveloped, Developing)
- Total Population
- Population density
- Healthcare system index ranked by WHO
- Median age of people in the country
- Costal / Non costal

- 168 countries included
- Date range: Jan 2020 to Apr 2020

Descriptive Statistics of Global cases as on 4/7/2020

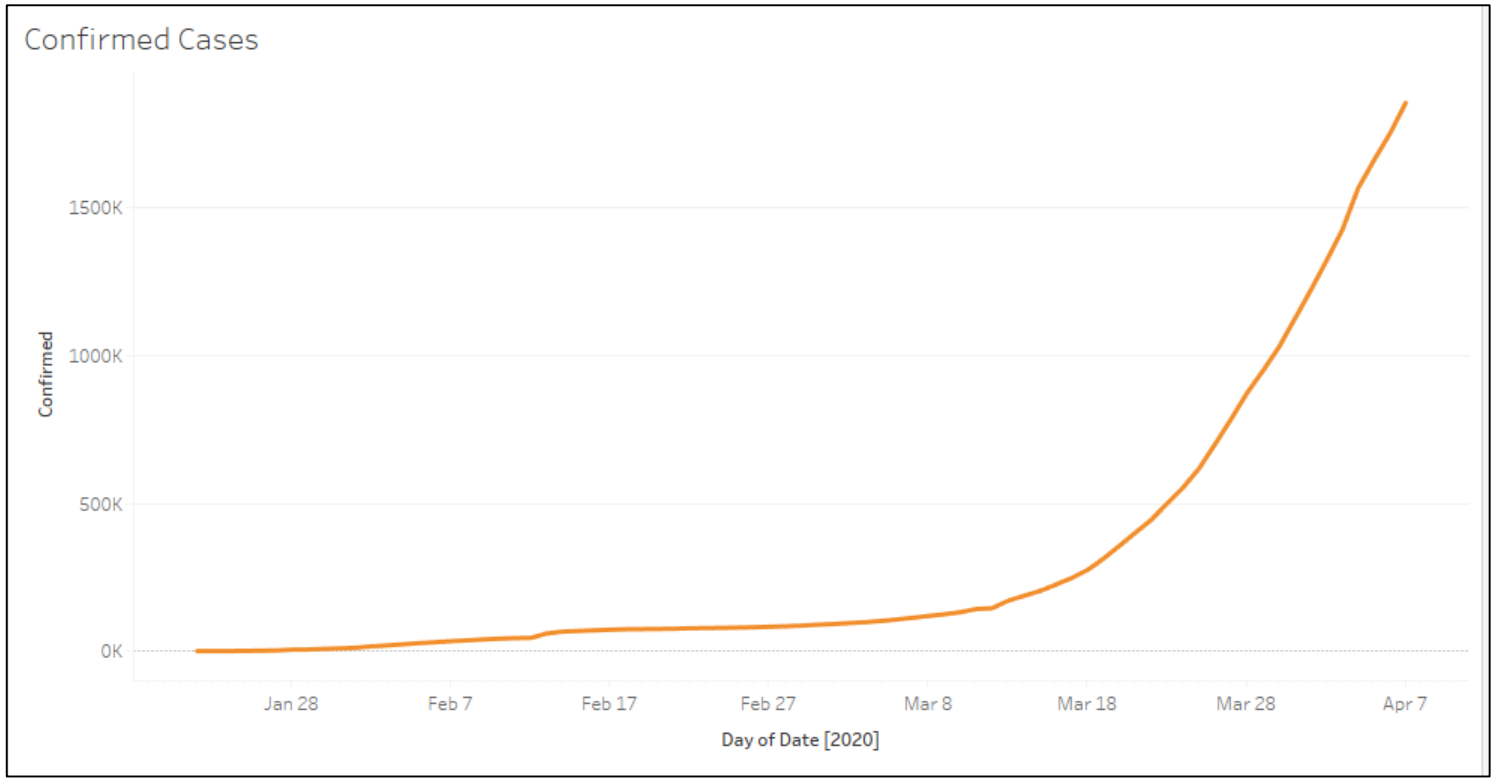
Confirmed Cases	
Mean	8458.035714
Standard Error	2833.287927
Median	383.5
Mode	8
Standard Deviation	36723.60876
Sample Variance	1348623440
Kurtosis	76.92656928
Skewness	8.001781428
Range	396221
Minimum	2
Maximum	396223
Sum	1420950
Count	168
Confidence Level(95.0%)	5593.678068

Recovered Cases	
Mean	1781
Standard Error	625.9778311
Median	37.5
Mode	0
Standard Deviation	8113.600012
Sample Variance	65830505.16
Kurtosis	51.33835175
Skewness	6.637580047
Range	77410
Minimum	0
Maximum	77410
Sum	299208
Count	168
Confidence Level(95.0%)	1235.849852

Deaths Cases	
Mean	486.4821429
Standard Error	168.7920882
Median	6
Mode	0
Standard Deviation	2187.795511
Sample Variance	4786449.197
Kurtosis	35.84836414
Skewness	5.863448001
Range	17127
Minimum	0
Maximum	17127
Sum	81729
Count	168
Confidence Level(95.0%)	333.2413175

Exploratory Analysis

Trend of confirmed cases of COVID-19



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

Probable causes:

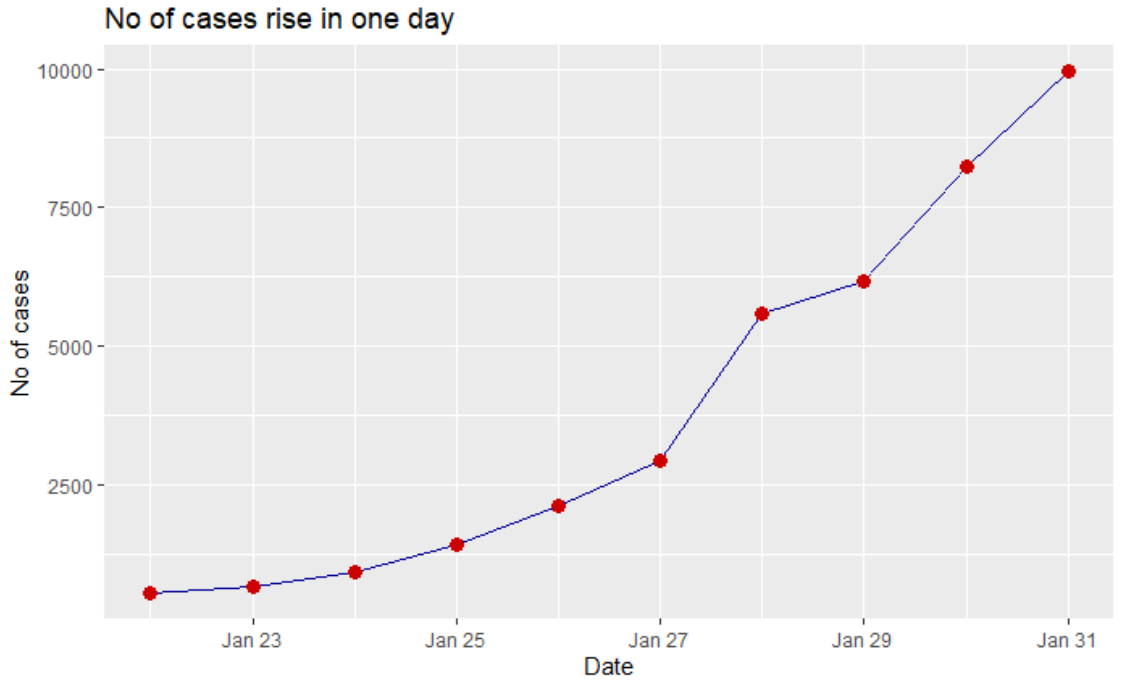
- More testing
- Less awareness of social distancing
- Delayed shutdown in some countries

COVID – 19 Number of cases increasing per day globally

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

Probable causes:

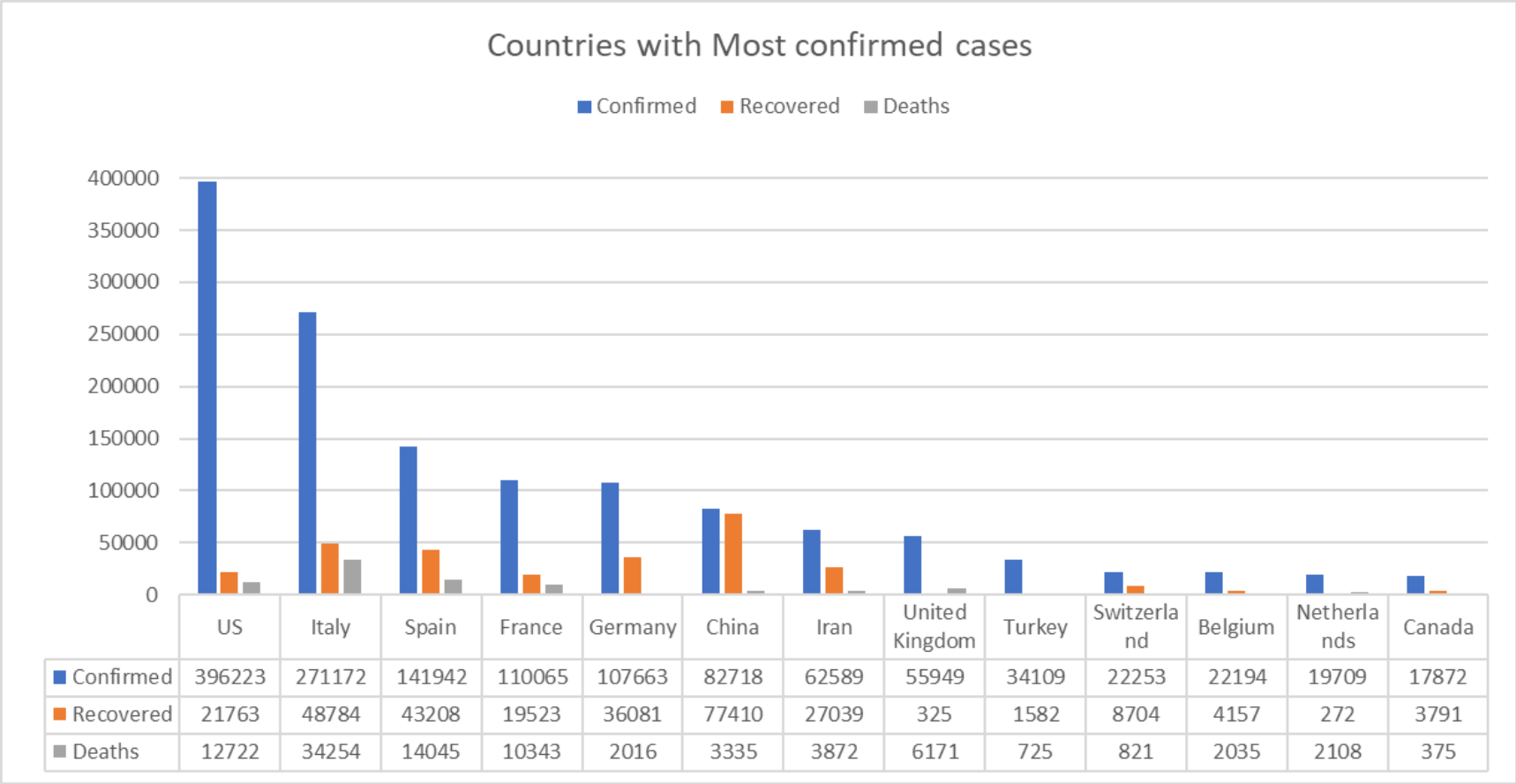
- More International travel
- Incubation period of 14 days
- Lack of early preventive measures by the governments from Jan 2020



Date <date>	No_of_cases_increased_per_day <dbl>
2020-01-22	556
2020-01-23	655
2020-01-24	942
2020-01-25	1436
2020-01-26	2122
2020-01-27	2931
2020-01-28	5587
2020-01-29	6176
2020-01-30	8247
2020-01-31	9950

1-10 of 77 rows

Which countries are the most affected by COVID – 19

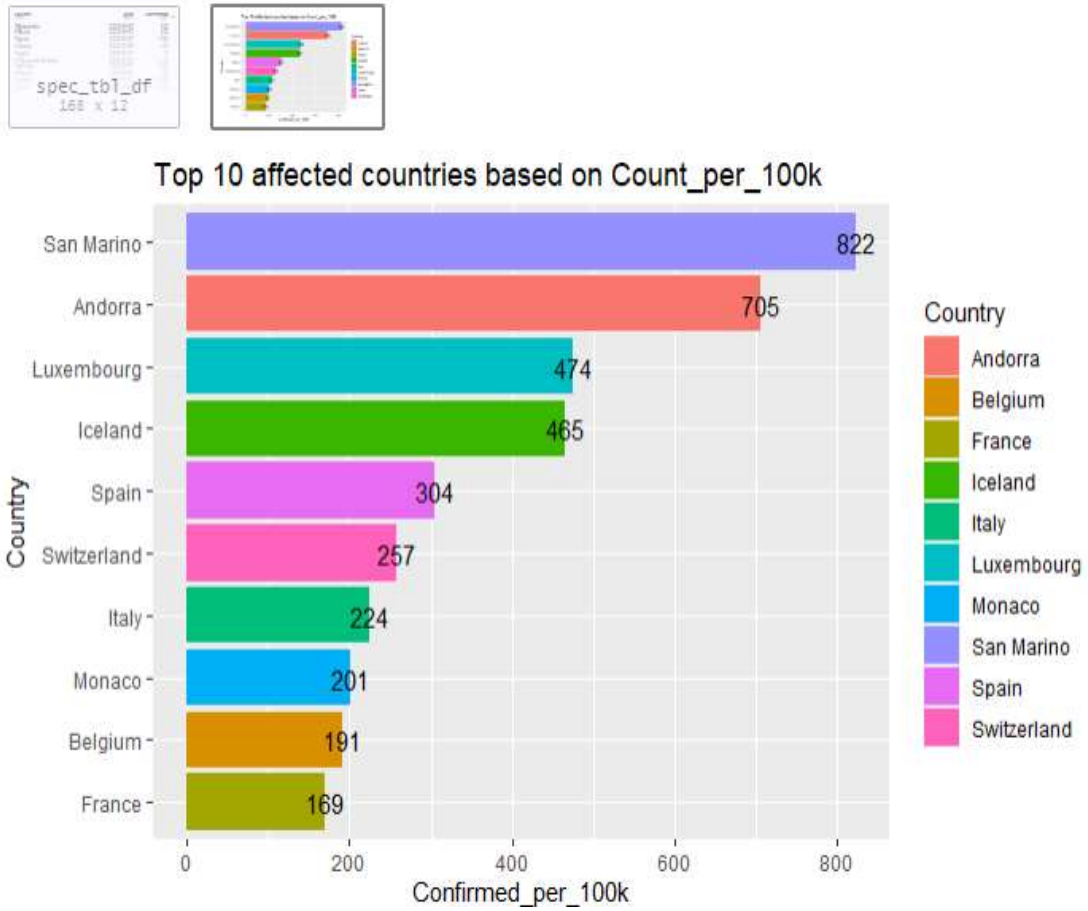


	US	Italy	Spain	France	Germany	China	Iran	United Kingdom	Turkey	Switzerland	Belgium	Netherlands	Canada
Confirmed	396223	271172	141942	110065	107663	82718	62589	55949	34109	22253	22194	19709	17872
Recovered	21763	48784	43208	19523	36081	77410	27039	325	1582	8704	4157	272	3791
Deaths	12722	34254	14045	10343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

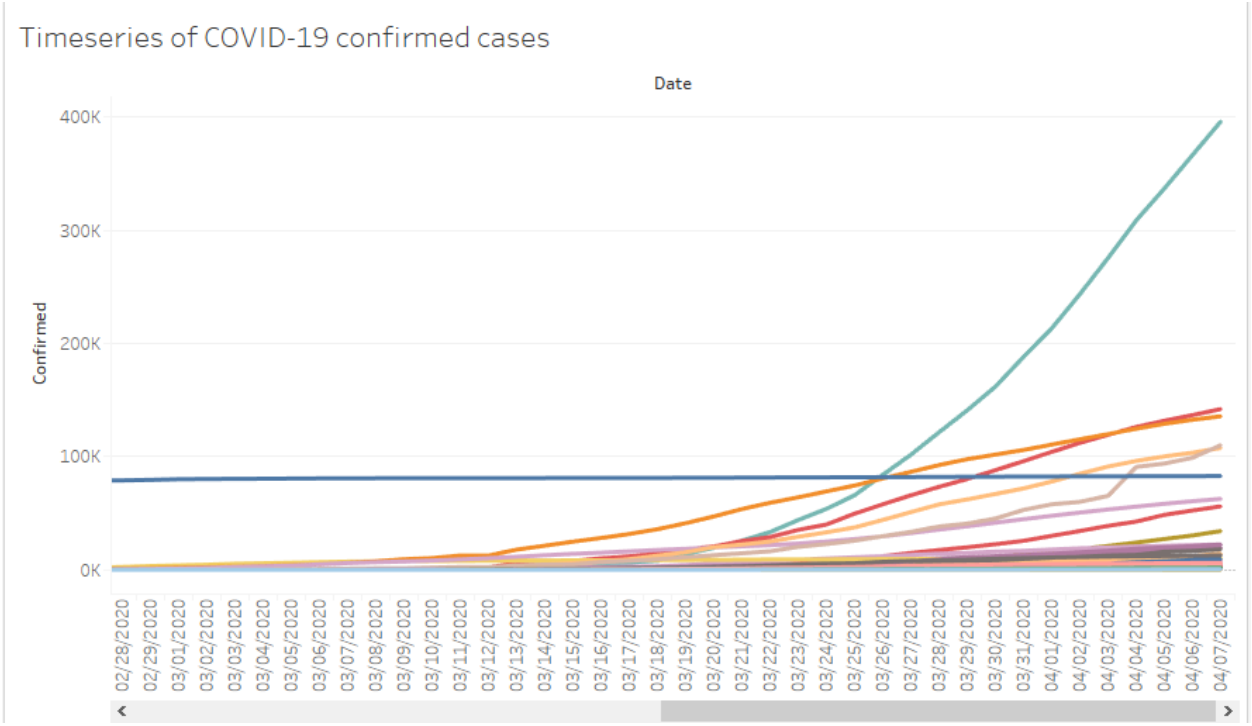
Summary

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

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



When was COVID-19 declared a pandemic and why?

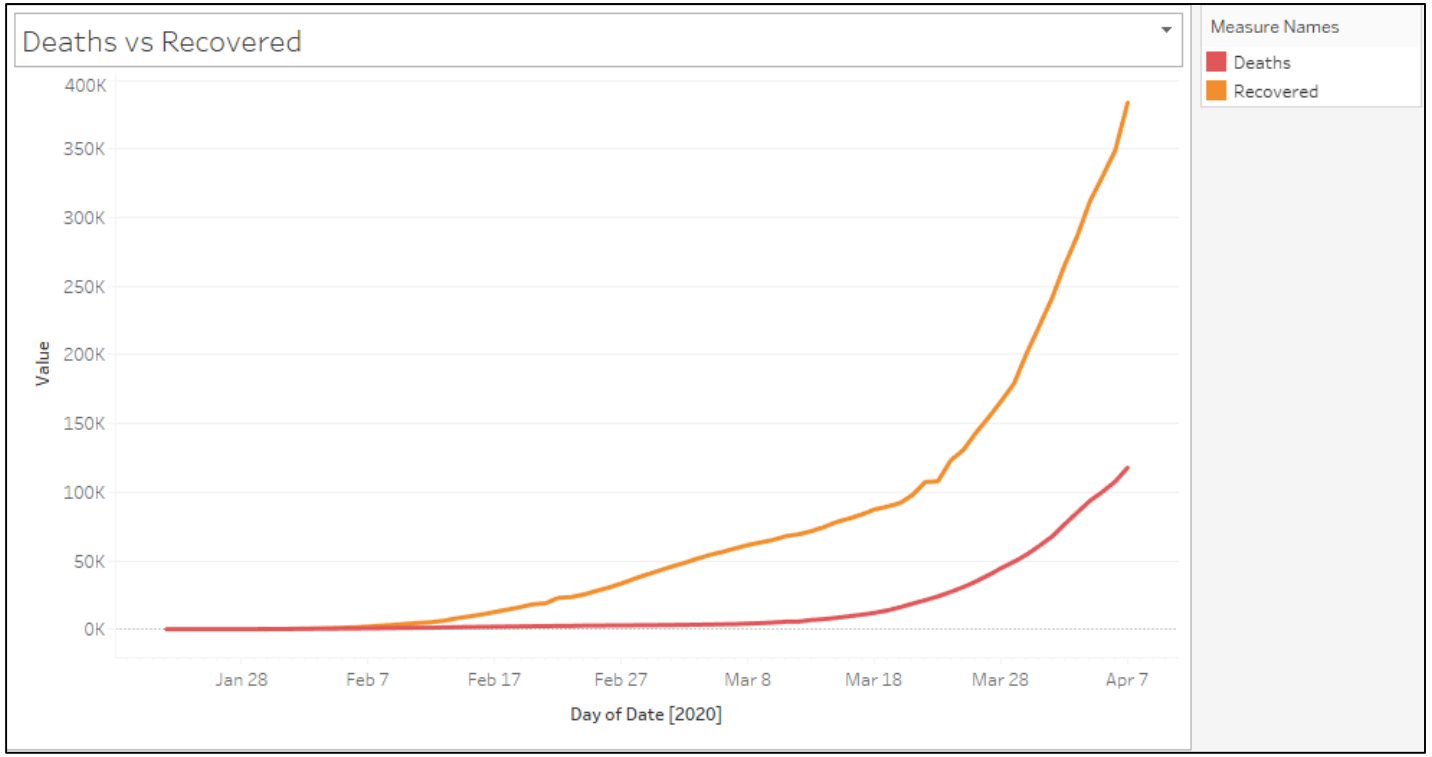


The number of cases observed world wide suddenly increased from March 2020

Probable causes:

- Virus was not contained in China
- Fast spreading nature of the virus(respiratory and human contact)
- Lack of awareness globally

Trend of deaths vs recovered cases of COVID-19

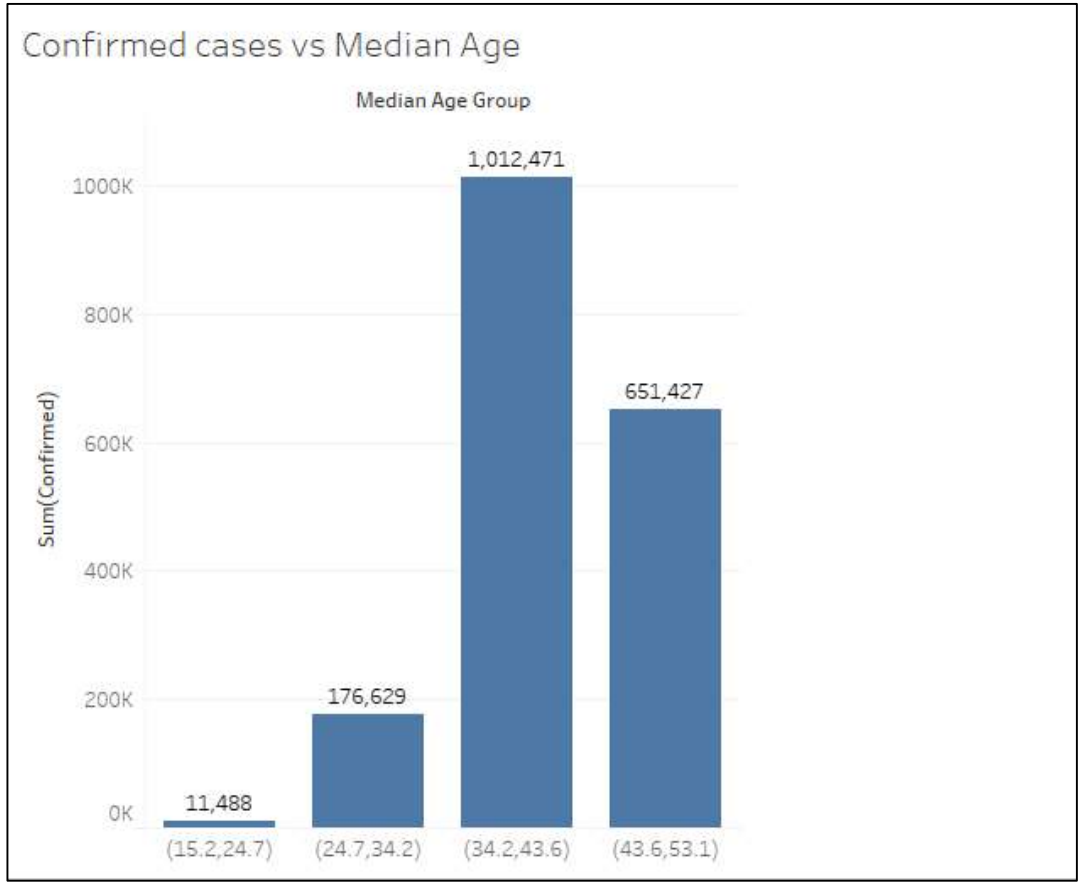


We can see that the number of deaths and recovered cases has had an exponential rise post from Feb 2020.

Probable causes:

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

How is the age affecting the number of confirmed cases?

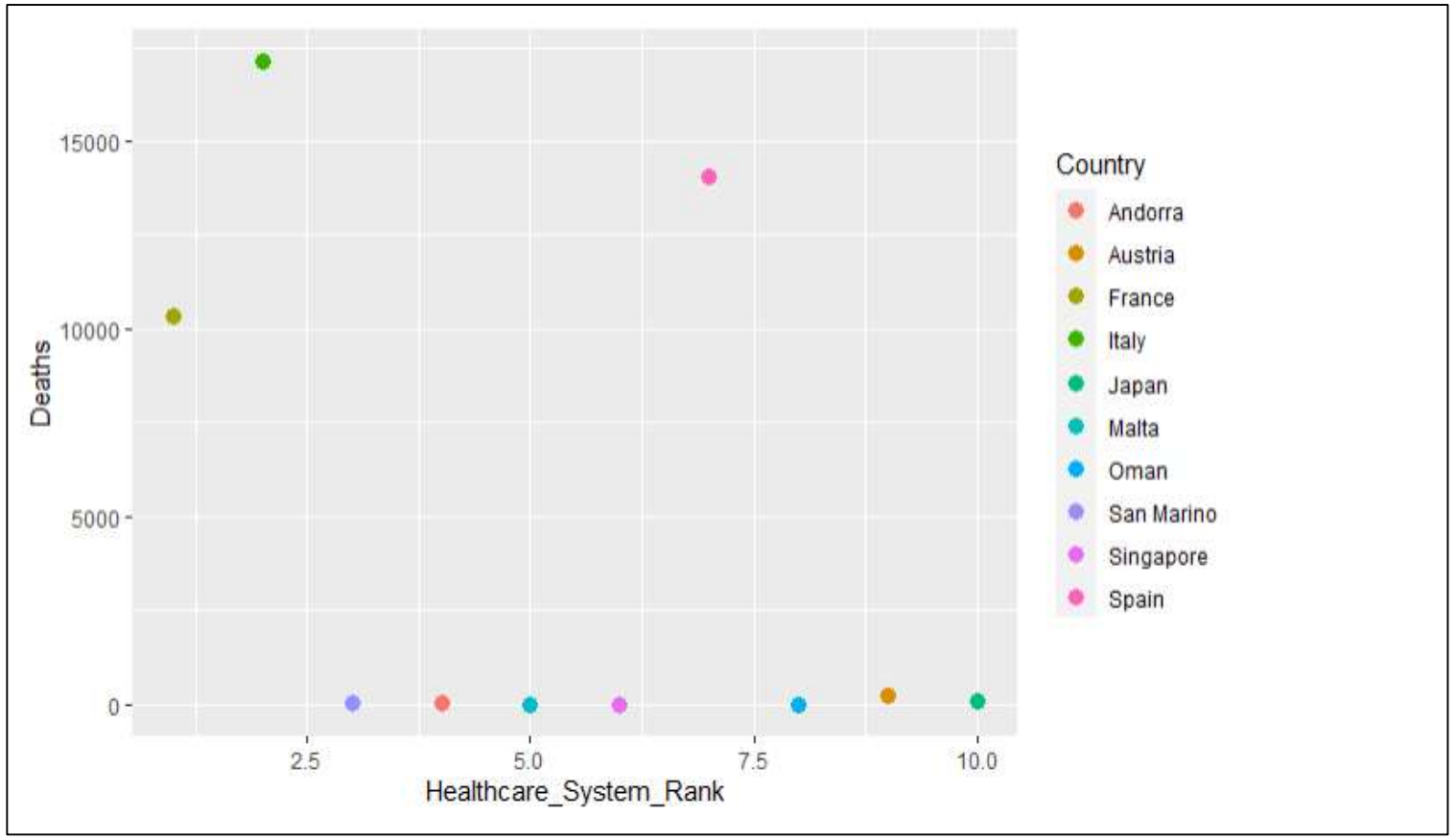


- We can see that people of ages 34 and older are majorly affected by COVID-19.
- As older people are prone to conditions such as hypertension, diabetes, cardiovascular, etc., they are at a greater risk.

Median_Age_Group <fctr>	sum(Confirmed) <dbl>	num_of_countries <int>
(15.2,24.7]	11488	53
(24.7,34.2]	176629	57
(34.2,43.6]	1012471	52
(43.6,53.1]	651427	13

4 rows

Which country has been most affected even though they have good healthcare facilities?



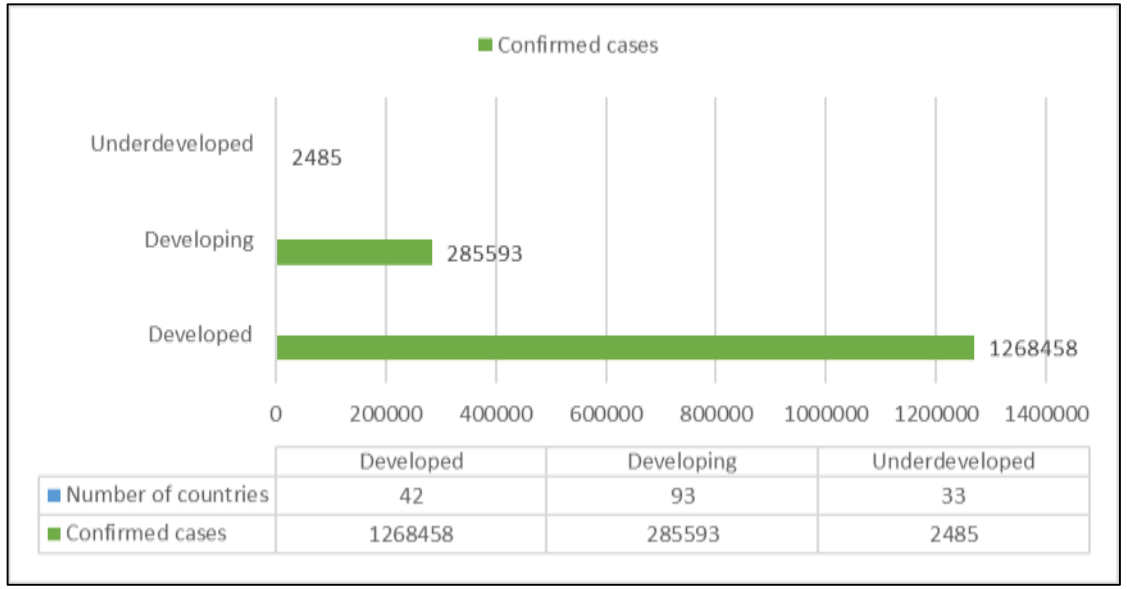
Summary

France, Italy and Spain are observed with the highest death rate.

Probable causes:

- Older population
- More tourism
- Densely populated

COVID – 19 confirmed cases based on development classification

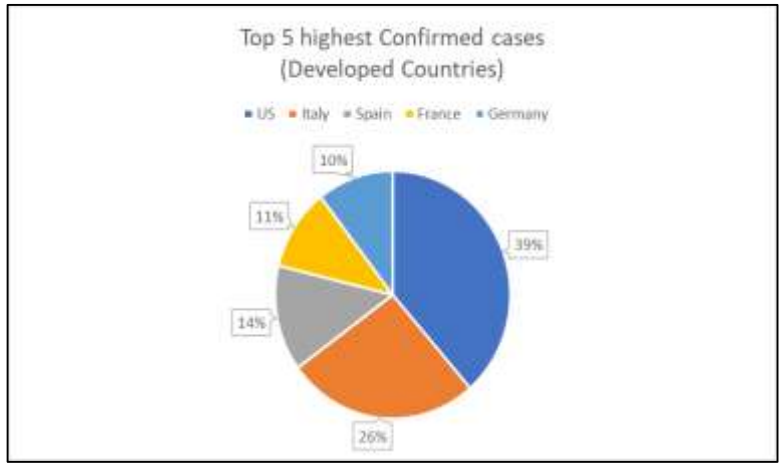


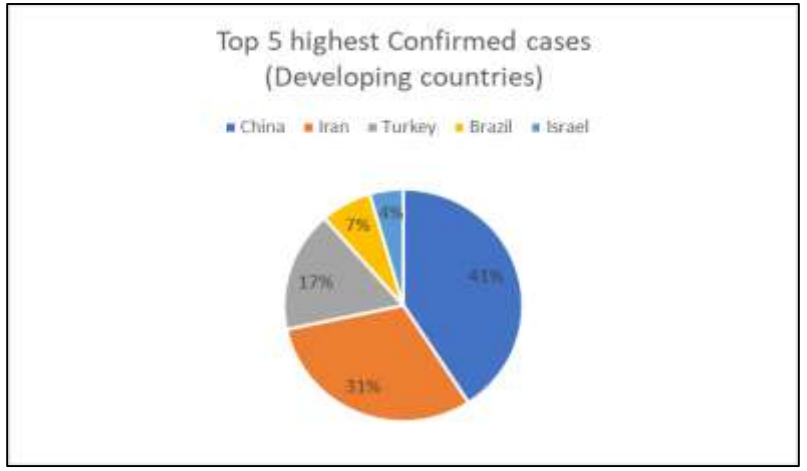
	Developed	Developing	Underdeveloped
Number of countries	42	93	33
Confirmed cases	1268458	285593	2485

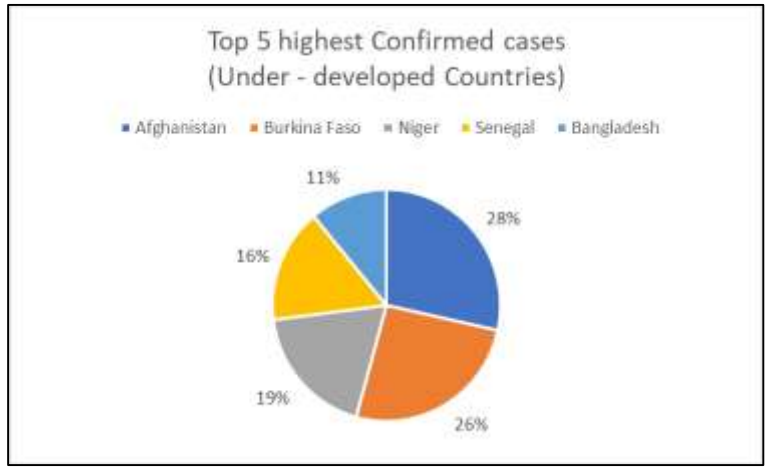
US, China, Afghanistan have the most confirmed cases in the three categories respectively.

Probable causes

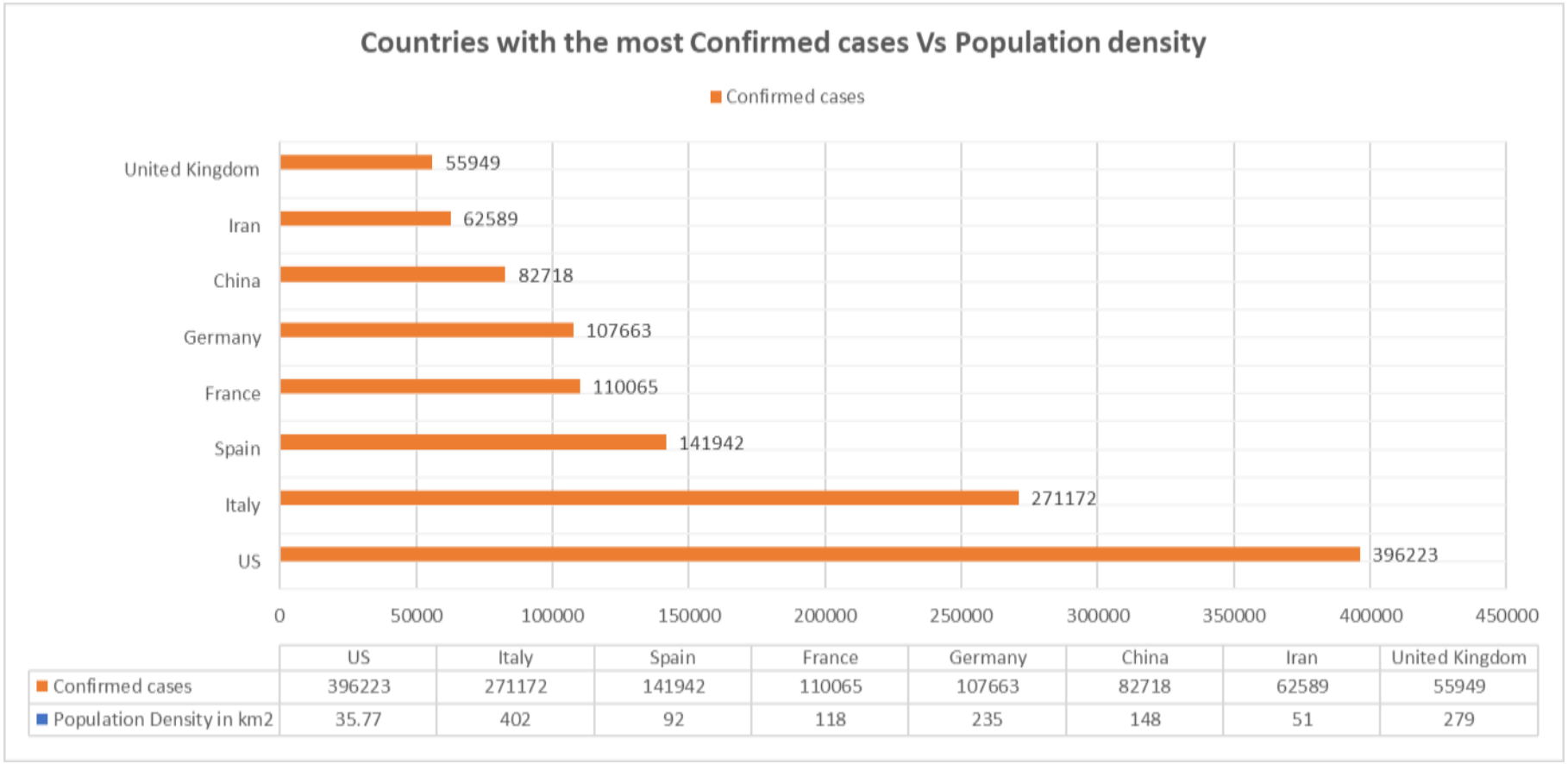
- More international travel, tourism in countries.







How is the Population Density of a country influencing the 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

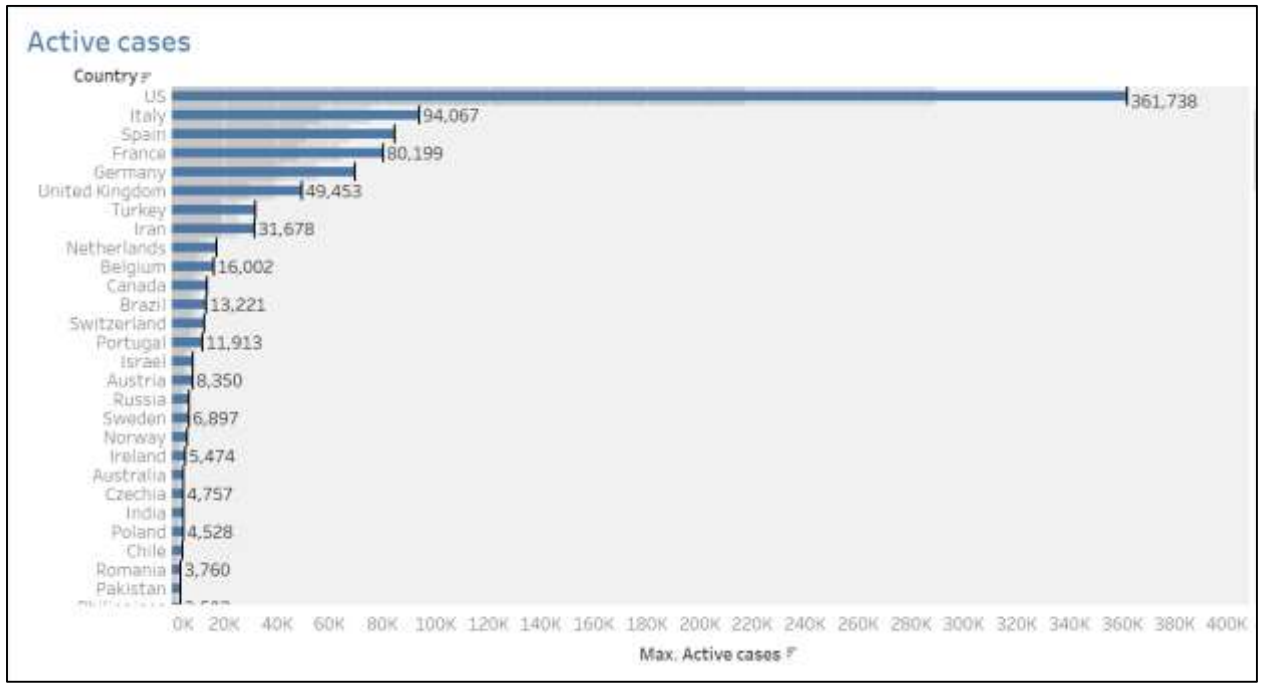
Summary

- 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

Which country has the highest number of active cases?

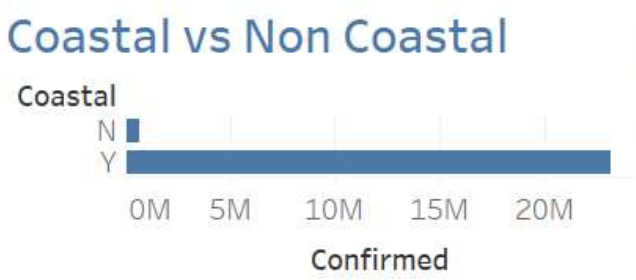


- Active cases = Confirmed – (death+ recovered)
- US has the most confirmed cases, followed by Italy and Spain.

Causes

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

Confirmed Cases by Coastal/Noncoastal classification



Coa..	
N	22,253
Y	396,223

Max. Confirmed

22,253 396,223

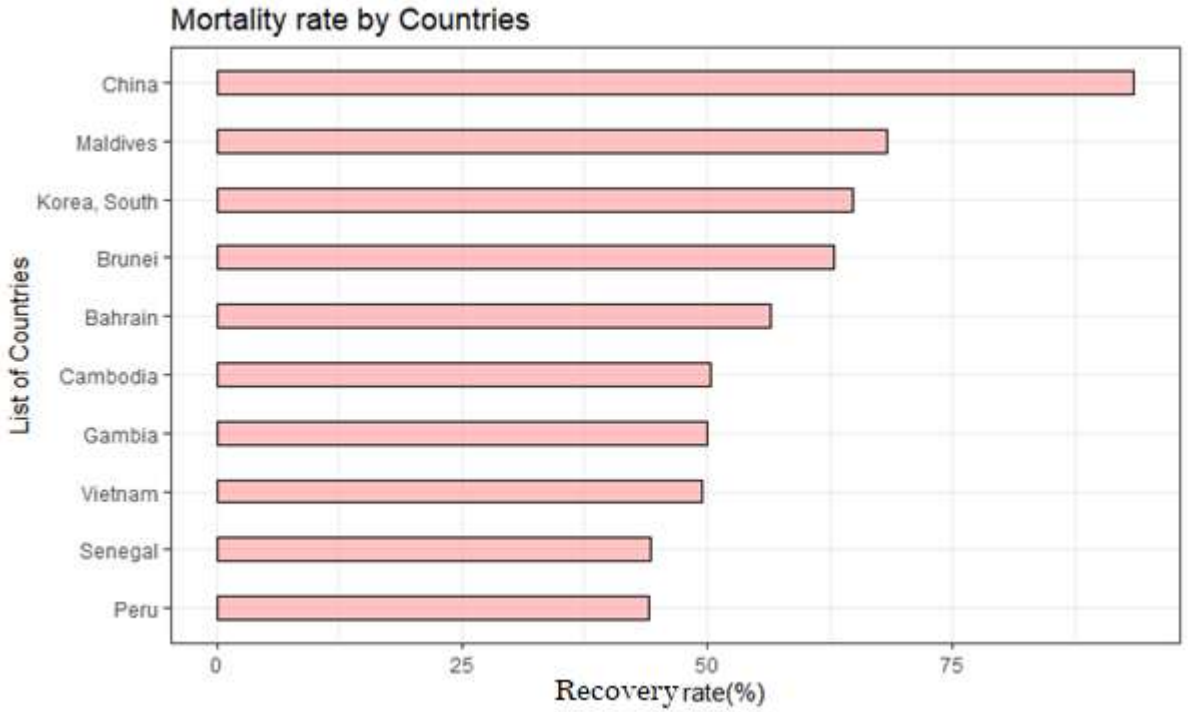
Summary

Most cases were found in Coastal countries.

Causes:

- Apart from airport entry there are many points of entries

Which country has the highest recovery rate?



Recovery rate = (Recovered/confirmed)*100

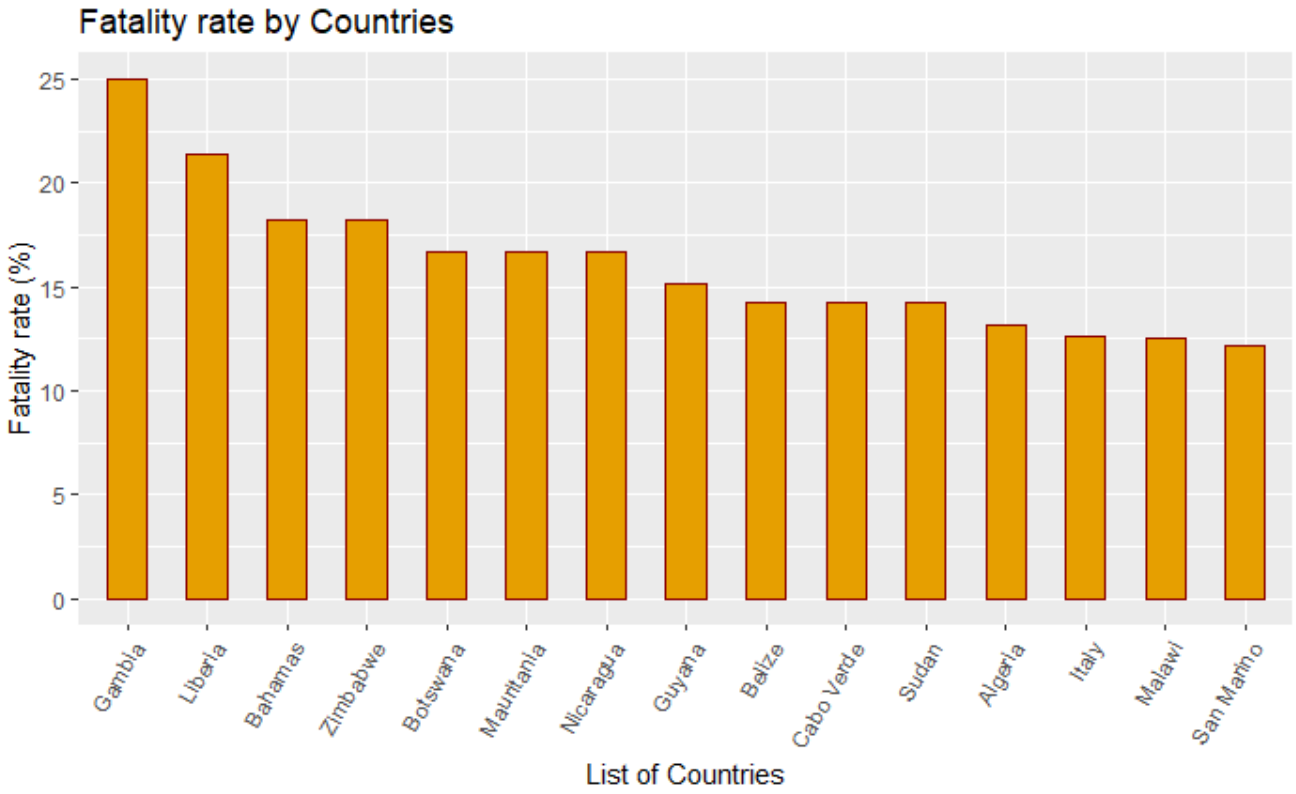
Probable causes:

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

Country <chr>	Max_Recoveryrate <dbl>
China	93.58
Maldives	68.42
Korea, South	64.80
Brunei	62.96
Bahrain	56.47
Cambodia	50.43
Gambia	50.00
Vietnam	49.40
Senegal	44.30
Peru	44.04

1-10 of 168 rows Previous 1 2 3 4 5 6 ... 17 Next

Which country has the highest fatality rate?



- Fatality rate = (Death/confirmed)*100
- Gambia has the highest no of death count vs confirmed count.

Probable Causes

- Lack of necessary medical facilities
- Lack of appropriate quarantine, isolation measures and lack of contact tracing

Country <chr>	Max_Fatality_rate <dbl>
Gambia	25.00
Liberia	21.43
Bahamas	18.18
Zimbabwe	18.18
Botswana	16.67
Mauritania	16.67
Nicaragua	16.67
Guyana	15.15
Belize	14.29
Cabo Verde	14.29

1-10 of 168 rows Previous 1 2 3 4 5 6 ... 17 Next

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.

References

- https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19
- https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3
- https://abc7ny.com/coronavirus-travel-to-new-york-nyc-china/6084186/
- https://www.aljazeera.com/news/2020/01/countries-confirmed-cases-coronavirus-200125070959786.html
- Datasets source: 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
- https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_Gambia