1. *Introduction*

2019 Novel Coronavirus (2019-nCoV) is a coronavirus identified as the cause of an outbreak of respiratory illness first officially been detected in Wuhan, China. There are various of symptoms including fever, dry cough, sore throat and shortness of breath. The virus can spread from person to person and have an incubation period of on average 5 to 6 days from being infected. However it can take up to 14 days for symptoms to show which makes coronavirus extreme hard to identify once infected. By the end of April 2020, coronavirus has already spread quickly and widely across the world.

China as the first country encounter with the outbreak, have recently reopen its lockdown city Wuhan. (2020 Hebei Lockdowns, 2020) With China seems have controlled the spreading of virus within the country, China is able to provide the world with data and experience in dealing with Covid-19. Thus, China's data is valuable to analysis in order to help the world to overcome the Covid-19 crisis.

The objective of this report is to investigate in following questions:

1. Which provinces have the highest confirmed cases and how fast is Covid-19 spreading?
2. Which age group are at a higher risk from Covid-19 in China?
3. How is China's response to COVID 19 compare to other countries?

2. *Data Wrangling*

The analysis conduct in this report uses the following data sources:

1. Novel Corona Virus 2019 data set, the data set contains daily level information on the number of affected cases, deaths and recovery from 2019 novel corona-virus given province and country. The dataset has 12k rows x 8 columns, contains tabular data and time series. The records start from 22th Jan, 2020 and is updating daily. The dataset used in this report is updated to 20th April, 2020.
   Link: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

2. Corona Virus data set with patients' details, this data set contains more information of infected patients, including age, gender, symptoms etc. The dataset is 13k rows x 44 columns, contains spatial data, dates, urls, tabular data. The time series recorded from 22th Jan, 2020 to the end of February.
   Link: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

3. The location of countries and cities from World cities database 2019, 15k rows x 11 columns spatial data.
   Link: https://simplemaps.com/data/world-cities

<u>Wrangling process:</u>

1. Structure: Input all three dataset into R studio and check the data types of each column. Make sure all columns are in the most appropriate data type for analysis. The date need to convert from factor to date format, command “as.Date” was used to reformat date variables and keep all date variables consistent. The “age” data is inputted as factor, in order to change it to numerical type while making sure data is correct, the data is first changed to character than to numeric.

2. Rename the “provinces” columns in the new Covid-19 China dataset and spatial dataset to make sure they are the same and merge the two datasets. Also, rename “Mainland China” as “China”, as tableau only recognize China when mapping.

3. Cleaning: Filter the both Novel Corona Virus 2019 data set and patient detail dataset into subsets containing covid-19 information only related to China with “filter” command in the dplyr package. For comparison between countries, filter out irrelevant information. (E.g. Diamond Princess confirmed cases)
4. Enriching: Merge the dataset with inner join, the new dataset contains both covid-19 infection information in china and location of provinces in china.

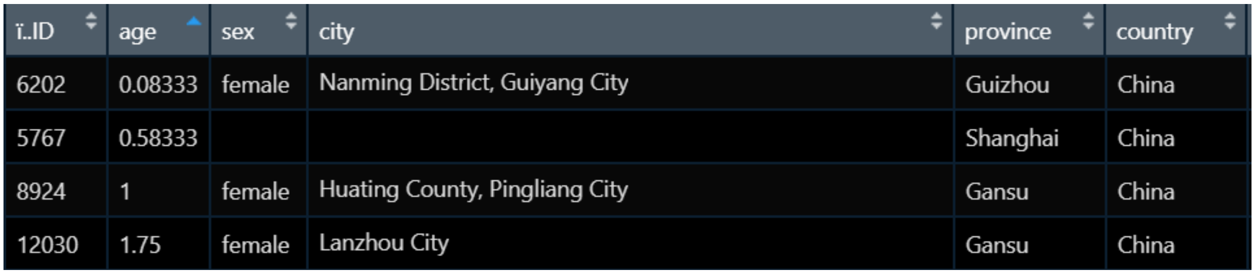
5. Validating: Scanning through datasets, inconsistency is discovered, some age information is represented as “18-70”, these data is substituted with the average of that given range to keep the data consistent. (E.g. (18+70)/2)

6. Before output the data as csv, go through error checking process.

3. Data Checking

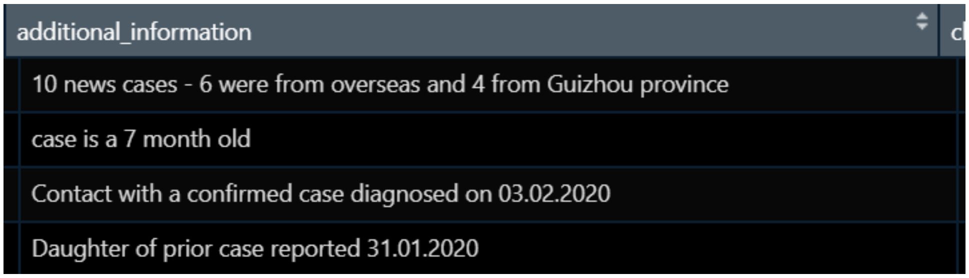
1. **<u>Entry error</u>**

Columns of Datasets are sorted to check for ambiguous values in the datasets. As shown in Figure 1, in the “age” column, age with 0.08333 and 0.58333 is detected. Through Investigation of the ‘additional information’ column (Figure 2.) , age value of 0.08333 is to be consider as an error. The row is deleted to make sure the accuracy of further analysis.



ï..ID	age	sex	city	province	country
6202	0.08333	female	Nanming District, Guiyang City	Guizhou	China
5767	0.58333			Shanghai	China
8924	1	female	Huating County, Pingliang City	Gansu	China
12030	1.75	female	Lanzhou City	Gansu	China

Figure 1. Entry Error

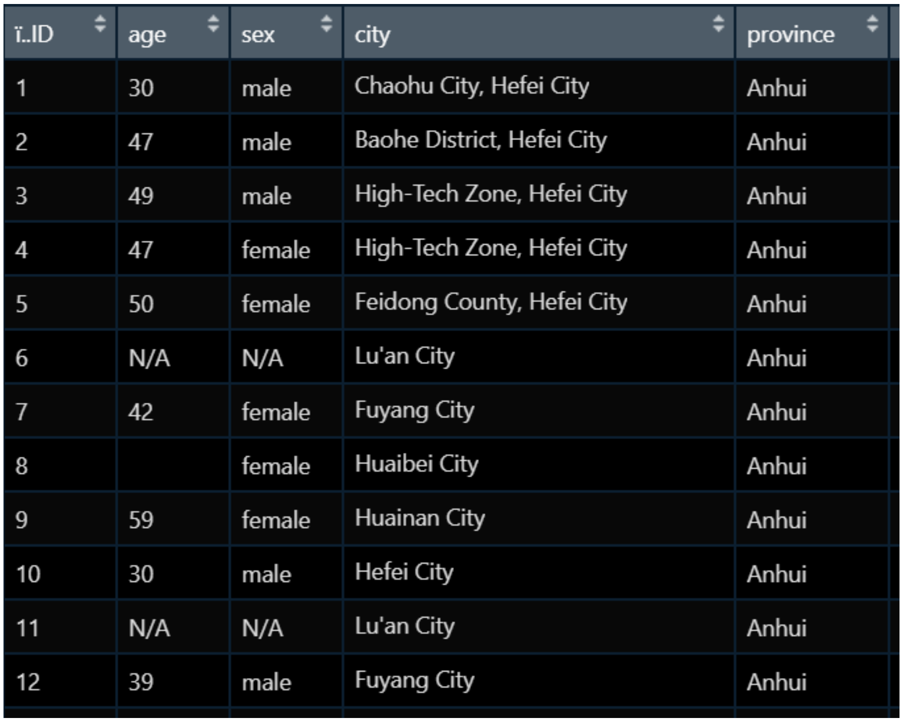


additional_information	cl
10 news cases - 6 were from overseas and 4 from Guizhou province	
case is a 7 month old	
Contact with a confirmed case diagnosed on 03.02.2020	
Daughter of prior case reported 31.01.2020	

Figure 2. Investigation on error

2. Unknown values

Through Excel and command in R studio, NAs and empty values are discovered in the dataset that need to be cleaned or substituted. (Figure 3) Instead of the replacing method, as the dataset contains more than 10k rows, removing some rows on large dataset is considered to have a relatively smaller effect on analysis. Thus, it is decided that all unknown values are replaced with "NA" and the dataset is filtered with "!is.na" to remove all unknown values from the dataset for analysis.



ï..ID	age	sex	city	province
1	30	male	Chaohu City, Hefei City	Anhui
2	47	male	Baohe District, Hefei City	Anhui
3	49	male	High-Tech Zone, Hefei City	Anhui
4	47	female	High-Tech Zone, Hefei City	Anhui
5	50	female	Feidong County, Hefei City	Anhui
6	N/A	N/A	Lu'an City	Anhui
7	42	female	Fuyang City	Anhui
8		female	Huaibei City	Anhui
9	59	female	Huainan City	Anhui
10	30	male	Hefei City	Anhui
11	N/A	N/A	Lu'an City	Anhui
12	39	male	Fuyang City	Anhui

Figure 3. NAs and empty values

4. Data Exploration

1. Which provinces have the highest confirmed cases and how fast is Covid-19 spreading?

There are 26 provinces in China, with Covid-19 outbreak initially start in Wuhan, Hubei province is expected to be the province with the most infected cases. Figure 4 is a virus distribution map created through tableau by inputting provinces and spatial data. It supports the hypothesis that Hubei is the province with most confirmed cases in China by 20th of April. Apart from Hubei, the provinces around Hubei also shows a relatively high number of confirmed cases compare to other provinces that are far away from the central of outbreak. Where close provinces such as Hunan, Anhui, Jiangxi, Zhejiang, Guangdong and Henan has around 1000 cases. Provinces that are far away from Hubei have lower cases. Therefore, the spreading of covid-19 can be considered highly related to the geo-graphic location in this case.

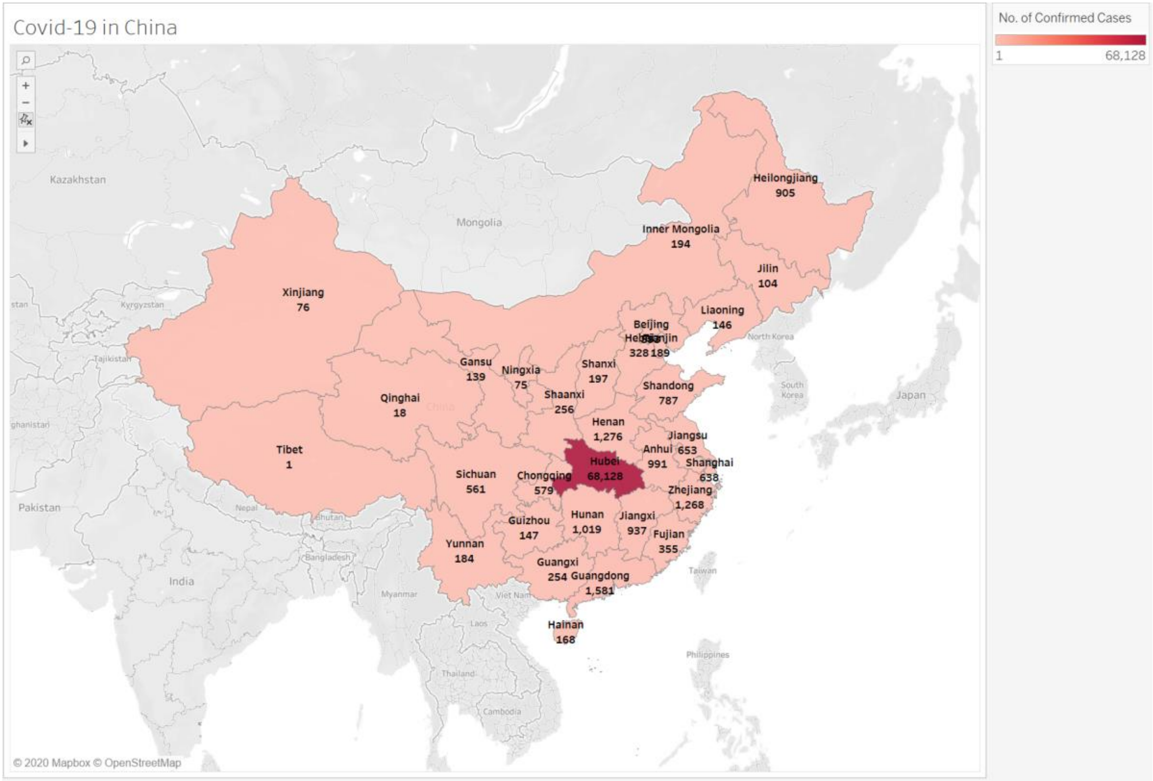


Figure 4. Covid-19 distribution map

It is known that on 23rd of January 2020, the central government of China imposed a lockdown in Wuhan and also other cities in Hubei. (Wikipedia, 2020) The action is aim to quarantine the centre of covid-19 outbreak and minimize any potential transmission of the virus. However, even with the "Wuhan lockdown" and other lockdown on cities in Hubei, coronavirus continue to transmit in China at a fast pace. Figure 5 contains two graph generated through tableau, the top graph shows the compound growth rate of confirmed cases per day and the lower graph shows the confirmed cases on match date. The growth rate was once reached to its peak 41.43% at the end of January. This means by the end of January, number of people infected by coronavirus are increasing by 41.43% per day. The speed of Covid-19 spreading in January was unbelievable, the numbers of infection can be doubled in nearly two days. Overall, the compound growth rate graph demonstrates an upward trend with high gradient in January and turns to downtrend in February. The downtrend follows an exponential shape that is expected to continues to flatten out in future. By the 20th of April, the compound growth rate has lowered to 5.89%, indicates number of people infected by coronavirus in China have been growing by 5.89% per day start from 22th of January. This is supported by the second graph in figure 5, the spreading of Covid-19 is clearly under control in March, the curve is flatten out and the number of confirmed cases increase only by less than 100 per day.

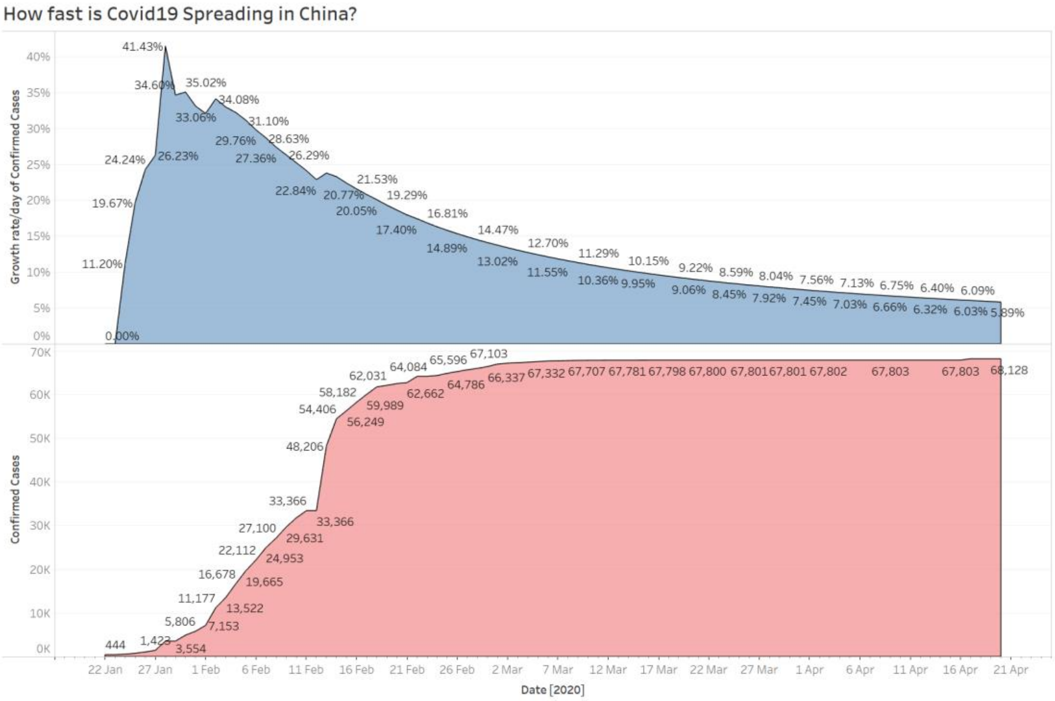
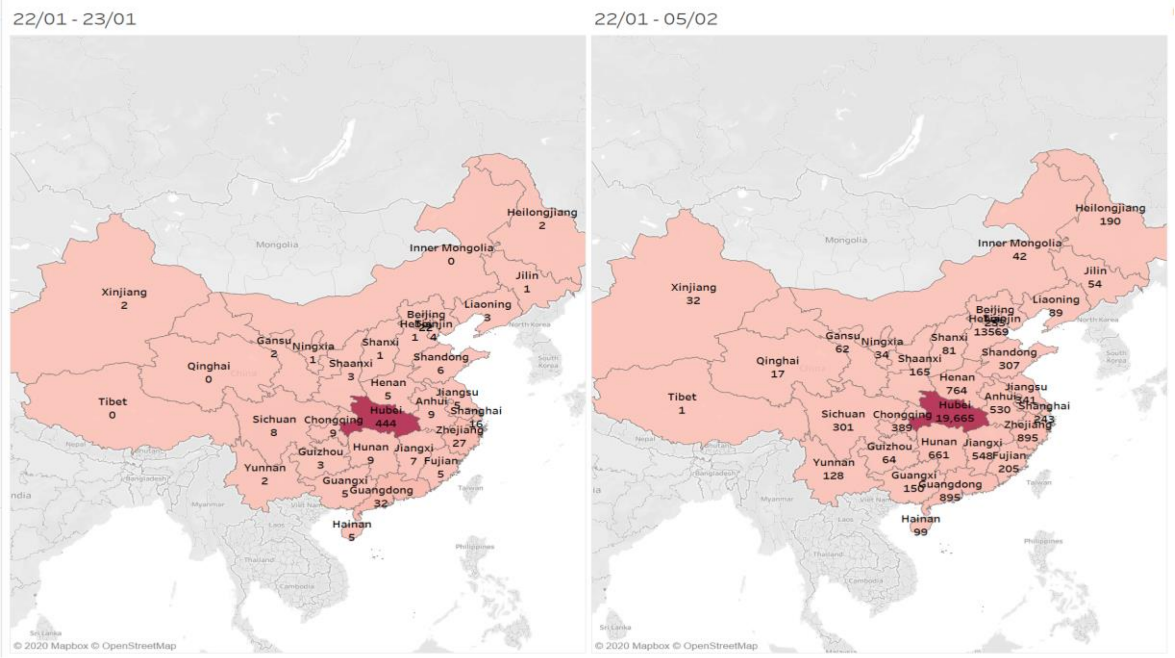


Figure 5. Growth rate of covid_19 in China


22/01 - 23/01

22/01 - 05/02

Figure 6. Before & After "Wuhan lockdown"

Figure 6 visualized the number of confirmed cases in china before and after the “Wuhan lockdown” on map. WHO estimates the incubation period of Covid-19 ranges from 1 to 14days, the graph shows number of cases in China on the lockdown date and 14 days after lockdown. According to the mayor of Wuhan, about 5 million people escape from Wuhan before the official lockdown of Wuhan city.(Collman, 2020) Although the government has made a fast decision on lock downing Wuhan, there are potential asymptomatic carriers escaping from Wuhan and thus accelerate the spreading and increase the radiance of spreading of Covid-19. In particular Hubei province, confirmed cases within the province nearly increase by 50 times in 14days. This again emphasis how fast coronavirus is able to spread among people and across provinces.

2. Which age group are at a higher risk from Covid-19 in China?

On 2 April 2020, WHO announced a statement that older people are at the highest risk from Covid-19 and all must act to prevent community spread. To investigate this issue, a bar chart of proportion of age distribution on confirmed cases in china is created through tableau with age group indicate by different colours. The chart shows that the age of infected people distributes in a normal distribution shape, indicates age distributes across all age group with a median at the 40s. Overall, only 35% of total confirmed cases are people under 40, where 65% of people being infected are over 40. Age group of 40 to 50 being the highest proportion among people infected by Covid-19. Thus, they are the most vulnerable age group in the crisis, and kids under 10 years old being the most unlikely age group to be infected. The proportion of people infected increase by 6 times from 10s to 20s, and again almost doubled from 20s to 30s, these information on the graph indicates that there are a huge difference in numbers between people older than 20years old and younger. This may leads to the reason being older people need to get in contact with others due to work or social events, therefore they are more likely to enter public area and being exposed to the coronavirus.

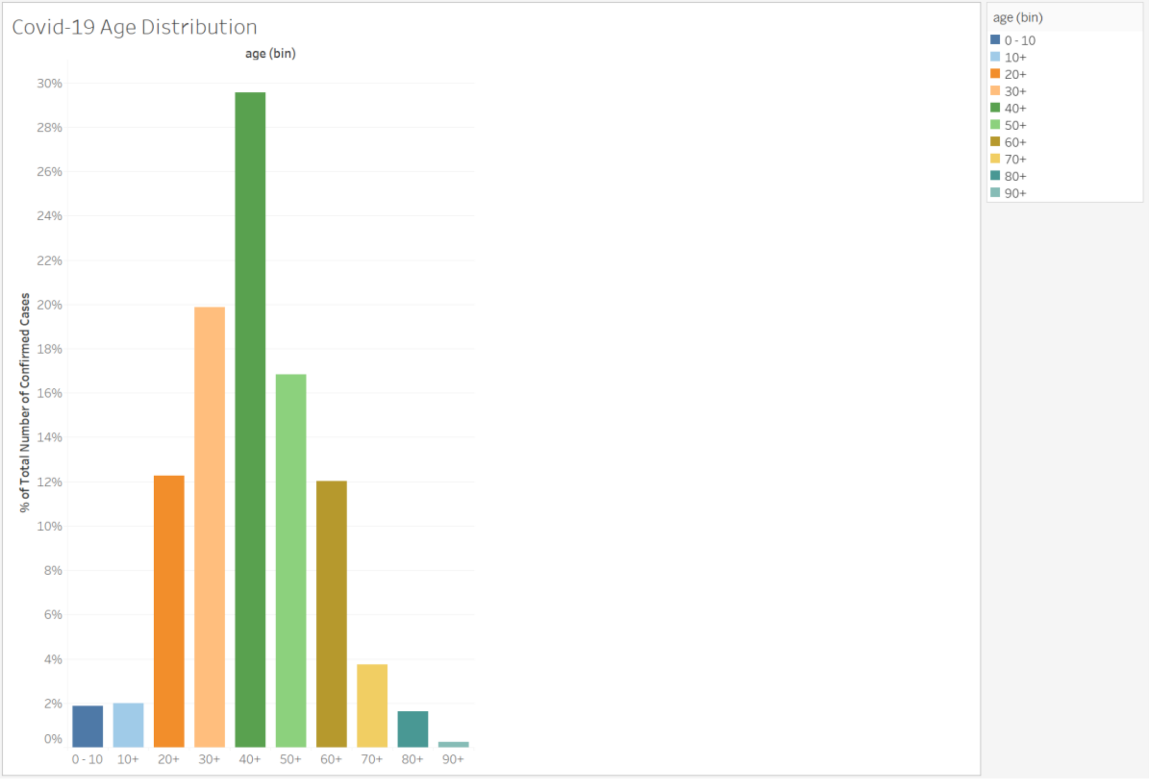


Figure 7. Age Distribution of Confirmed Cases

3. How is China’s response to COVID 19 compare to other countries?

Although Covid-19 cases was first officially identified in Wuhan, China, it didn’t take coronavirus too long to spread all over the world. From figure 8, its clearly that coronavirus has invade nearly all other countries at the beginning of March. It is observed that, by the end of March, people in some of the developed countries have already been seriously harmed by Covid-19, and have more confirmed cases in the country compare to China. On the other hand, looking at China’s curve and time line, it only took china one month to flatten the curve and take control of the disease within the country. On 20th April, China is ranked at the 8th where US, Spain, Italy, France and Germany have passed china with more than 100k cases, United States in particular have over 750k cases as the highest among the world. Also, apart from China, most of the developed countries on the graph are showing a uptrend of confirmed cases, indicates the coronavirus may continue to spread all over the world epically in US and the Europe region.

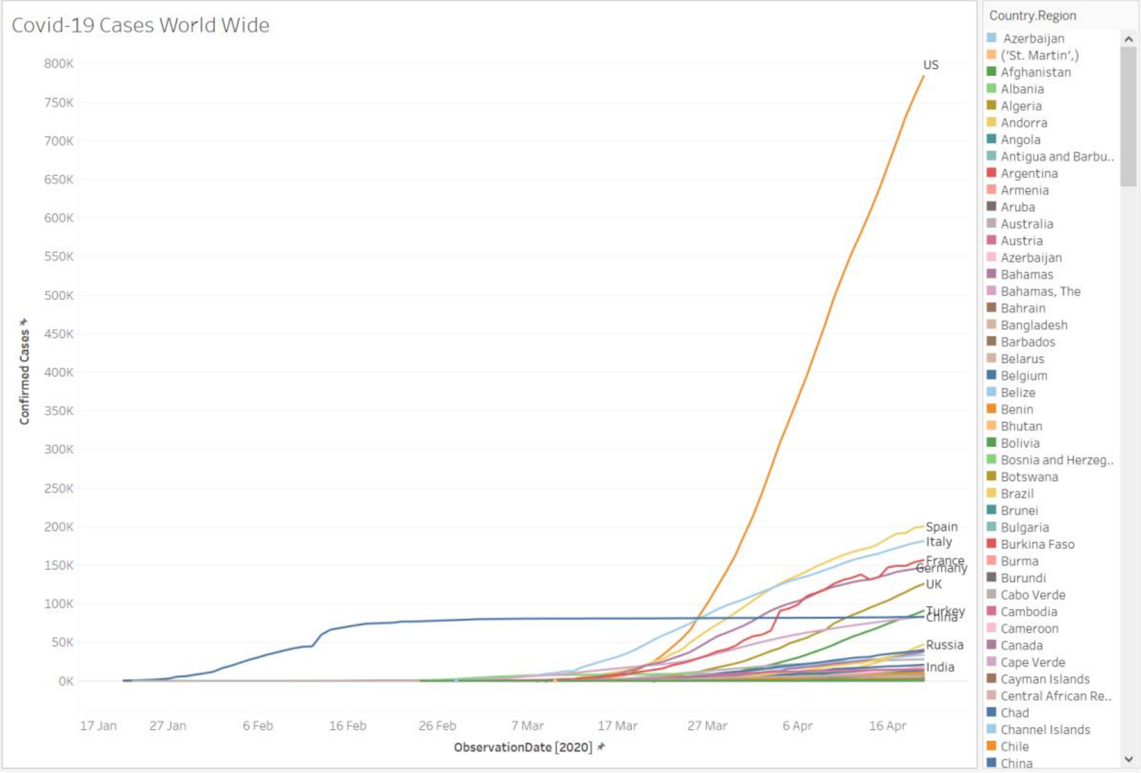


Figure 8. Confirmed Cases in countries

US, Italy and Spain as top three most infected counties are selected to analysis the performance of China react to Covid-19 outbreak. Figure 9 shows numbers and rate of recovery and death from the day of outbreak until 20th of April. The recovery rate is computed by (recovered cases /confirmed cases)*100 and death rate is computed by (death/confirmed cases)*100. China has a leading recovery rate among the four counties at 93.17%, where most patients are able to recover from the disease. In comparison to China, US have the lowest recovery rate, only 9% of patients can be cured for now. The recovery rates in Italy and Spain are increasing with death rate increasing at a lower degree over time.

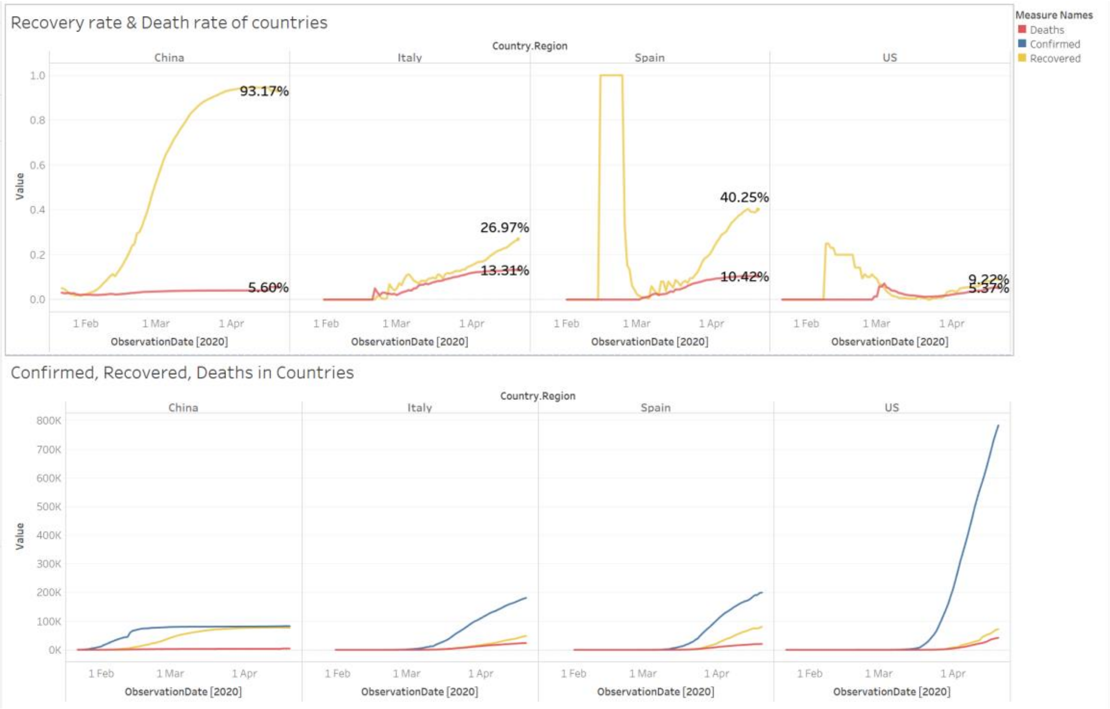


Figure 9. Death & Recovery rate in countries

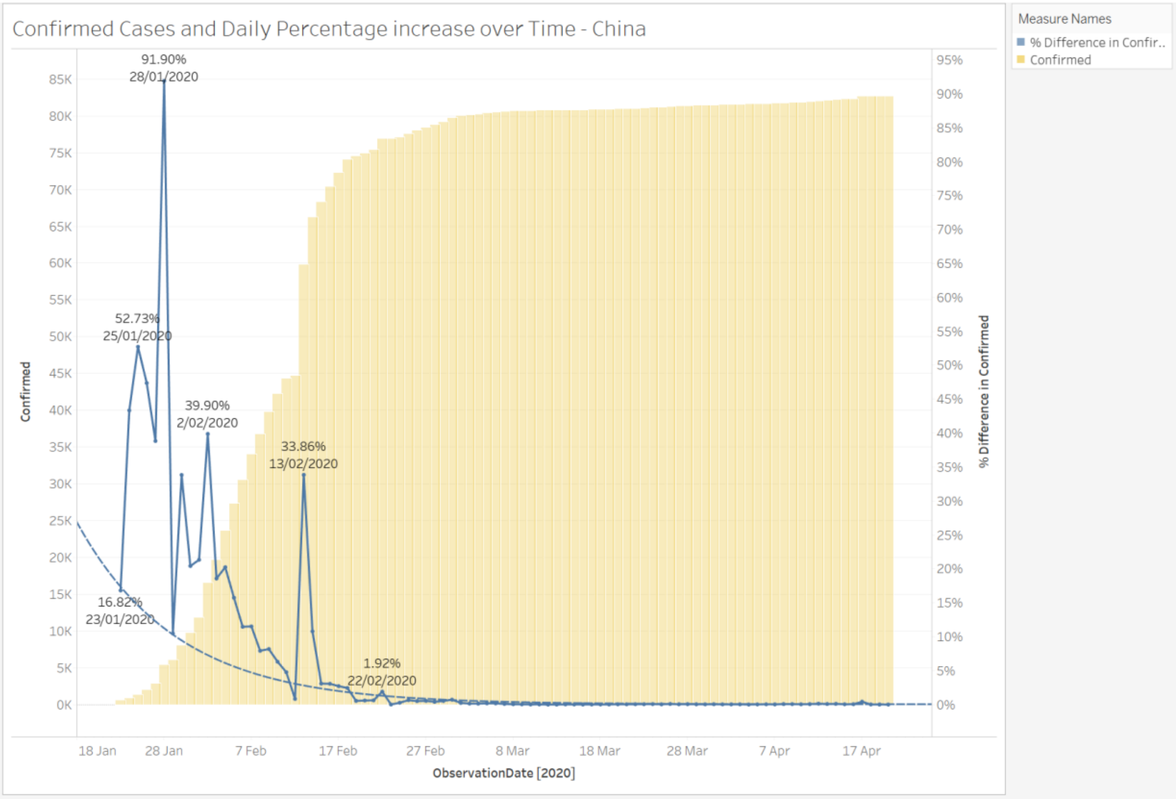


Figure 10. Percentage Increase Over time

Figure 10 is created to help investigate further on how China reacted against the outbreak and the effectiveness of the strict measures China took. "Wuhan lockdown" happened at 23rd Jan., the spreading of covid-19 reached the peak of 91.9% increase in cases one week after the lockdown. On 22rd Jan, the Wuhan government orders all residents to wear face masks in public to slower the outbreak of coronavirus.(You, 2020) After two weeks of incubation period, daily percentage increase once again reached a second peak of 40% on 2rd Feb but much lower compare to the first one. After the second peak a downtrend is observed in February from the graph. China took more steps, new hospitals were built in rush on 2 Feb. 10 more hospitals were built in February providing thousands of beds for patients. (Hilary Brueck, 2020) Also, China started disinfect the entire Wuhan city on 9th Feb., disinfect twice a day. The third outbreak, happened on 13th Feb, confirmed cases increased by 34% which lower to previous peaks. From the three peaks, we can observe that the rate of increase in China is declining In two weeks' time, and in two weeks' time, daily growth rate in China lowered to 2%. Overall, the daily increase percentage follows a convex shape and gradually approach zero in the end of February. Thus, from the observation on statistics and graphs, it is fair to say that compare to other counties, China is able to react to Covid-19 efficiently and effectively minimize the spreading and infection inside the country.

5. Conclusion

Although it is perceived that China has overact to the outbreak of coronavirus compare to other countries, the coronavirus still spread widely and quickly across the country from January to February. Luckily, China's overaction and early lockdown on cities have managed to controlled most of confirmed cases inside Hubei province. From investigation, it is found that older age group are the most at risks. Nearly all of the patients are over 20 years old and especially people in 40s. To prevent the spreading of Covid-19, Chinese government demonstrates a highly rated performance in the crisis compare to the rest of developed counties in the world. Through doing massive testing, building hospitals, postpone non-urgent medical care, city disinfection etc. China successfully takes control of Covid-19 in nearly one month. Over 90% of the patients are recovered with a death rate of 5%. While other counties still suffered from the crisis with a higher death rate higher than 10%.

6. Reflection

In the report, R studio are used in data wrangling and cleaning, providing convince and efficiency for the analysis and visualizing through tableau. The analysis and visualizations are preferred with tableau as tableau creates more beautiful and clear graphs for pattern discovery and visualizing the data. Especially on the mapping created through tableau, with support of colours and labels, audiences are able to understand the data quickly and identify patterns within the data and time line accurately. However, if the analysis and graph were created by R studio, extra testing and fitting could be added, provide further evidences and evaluation of the dataset. Statistical testing such as T-test could be applied easily through R to test for the correlation between variables to discover hidden relations.

7.Bibliography

1. *What you need to know about coronavirus (COVID-19)*. Australian Government Department of Health. (2020). Retrieved 27 April 2020, from https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/what-you-need-to-know-about-coronavirus-covid-19.

2. *2020 Hubei lockdowns*. En.wikipedia.org. (2020). Retrieved 27 April 2020, from https://en.wikipedia.org/wiki/2020_Hubei_lockdowns.

3. Collman, A. (2020). 5 million people left Wuhan before China quarantined the city to contain the coronavirus outbreak. Business Insider Australia. Retrieved 27 April 2020, from https://www.businessinsider.com.au/5-million-left-wuhan-before-coronavirus-quarantine-2020-1?r=US&IR=T.

4. Hilary Brueck, S. (2020). *China took at least 12 strict measures to control the coronavirus. They could work for the US, but would likely be impossible to implement.*. Business Insider Australia. Retrieved 27 April 2020, from https://www.businessinsider.com.au/chinas-coronavirus-quarantines-other-countries-arent-ready-2020-3?r=US&IR=T.

5. You, T. (2020). *China coronavirus: ALL Wuhan residents must wear face masks*. Mail Online. Retrieved 27 April 2020, from https://www.dailymail.co.uk/news/article-7916613/Wuhan-government-orders-residents-wear-face-masks-public.html.

6. *China streets coated in disinfectant spray to halt outbreak*. NewsComAu. (2020). Retrieved 27 April 2020, from https://www.news.com.au/world/asia/chinese-authorities-cover-streets-in-clouds-of-disinfectant-to-halt-spread-of-deadly-coronavirus/news-story/890ccb3a115ca6f622e9f07d7067c97e.