COVID-19 Virus - Understanding the spread and its extent

Project Description

We first learned about the Corona virus late December 2019. The first cases were seen in Wuhan city in China. Since then, the virus has spread to most countries. Over 110K people to this day have been infected with over 3800 death reported. Whilst, the figures can be terrifying, thorough analysis of the data can help manage the containment and allocating appropriate resources.

The main hurdle is the quality of data which appear to be collected from various sources which may not be fully trusted. In some countries, accurate details are not reported, inaccurately reported or they are underestimated. This causes the results to be skewed and needs to be taken into consideration. For example, figures in Iran where figures are under-estimated or in some African countries where they are not reported may need to be analysed individually and not compared.

What are we trying to uncover?

- Is the situation improving or worsening in a particular region?
- How do the figures compare with other locations within the same country?
- Is there a spread from one country to another?
- How is the quality of treatment, i.e. mortality rate?
- What is the relation between age and prognosis?

Motivations & Values

- Enables effective and clear communication of the issue
- Helps predicting the extent of the problem for the next 2 weeks and mobilize resources accordingly
- Helps health sector to come up with plans to accommodate affected persons
- Helps Governments devise strategies to prevent new infections and contain existing infections
- Helps government devise strategies to deal with the implications by identifying the areas affected most

Scope & Plan

The scope of this project includes:

- Data sets available on Kaggle and the links it refers to. We are not aiming to fill missing values from other sources or augment the data with additional data that may be available from elsewhere. We will however try to find additional sources of data.
- Data cleansing & wrangling is done at 2 stages. In the first stages, only useful columns are included and rows with multiple columns missing are removed. In the second stage, column level transformation is applied. This will be done using R. The initial data examination identifies the type of data transformation required.
- A combination of Tableau, R and D3 is used to understand the data for filtering, wrangling and analytics.

Appropriate Visualization Methods:

At this stage considering the following methods.

- Tabular data and bar charts to show number of confirmed, death and recovered
- Tree Maps to show the scale of infections over geographical areas
- Line chart to show the trend of the confirmed, death and recovered over 31 days.
- Bubble charts showing the trends across locations using number of deaths, infections and recoveries

Planning

When	What	Milestones
Week 3	Initial project description, column level transformations and some preliminary findings	Project proposal
Week 4	Comprehensive data cleansing and wrangling.	Cleansed data sets
Week 5	Develop data visualization units	Visual artefacts
Week 6	Document visualization and analysis results.	Documented findings

What is done so far

So far, I have:

- Documented the project proposal, examined data sets and appropriate transformations:
    - Examined each dataset in terms of its columns, format, and significance.
    - Identified which data sets to use and which columns to include
    - Identified what transformation and cleansing is required.
    - Explored what we can use the data sets for.

Intended Audience

They include:

- General public
- Health Practitioners
- Policy Makers

Data Sources

The source of data is the "Novel Corona Virus 2019 Dataset" published by Kaggle. The data is said to be updated every day but in fact covers the period of late January to late February. More recent version of data set(s) is obtained from the referring link but they also contain data no later than mid-march. Kaggle itself sets the following tasks:

- Predict how the epidemic will end
- Predict spreading of the virus

Our objective is rather understanding the current trend not so much prediction. Analysis of the data through visualization by itself however may help answer those analytical questions. Some data appear to be duplicated across data sets. Below is a detailed description of the selected data sets and recommended transformations.

Data Description & Transformations

1-COVID19_line_list_data.csv

This dataset provides high level view of number of cases, age group, location and symptoms. We could use it to

- Analyse the direction of the spread of the virus into and out of Wuhan.
- Analyse age groups and examine symptoms for varying age groups
- Analyse the proportion of death
- Analyse the spread over geographical locations

Column	Encoding	Null?
id	Categorical:Nominal	N
case_in_country	Categorical:Nominal	Y
reporting.date	Date	Y
X		N
summary	Categorical	N
location	Categorical	N
country	Categorical	N
gender	Categorical	N
age	Continuous:Sequnetial	N
symptom_onset	Date	Y
If_onset_approximated	Continuous	Y
hosp_visit_date	Date	Y
exposure_start	Date	Y

Column	Encoding	Null
exposure_end	Date	Y
visiting.Wuhan	Categorical	N
from.Wuhan	Categorical	N
death	Categorical	N
recovered	Categorical	Y
symptom	Categorical	Y
source	Categorical	N
link	Categorical	N
X.1		Y
X.2		Y
X.3		Y
X.4		Y
X.5		Y
X.6		Y

Highlighted in Pink indicates the column is filtered out during the first stage of the cleansing process.

Data Wrangling, Cleansing & Transformation

1. id: Not useful
2. case_in_country: Not useful
3. Summary: Drop column Summary as it contains information covered in columns.
4. Location: Remove Chinese characters
5. X: Remove column as it contains null only.
6. Summary: Drive a new value called ‘Precedence’ with values ‘New’, ‘First’, ‘Death’
7. Age: Break into groups of 1-7: Child, 8-12: Very Young, 13-19: ‘Teen’, 20-28: Young, 29-50: Adult, 51-70: Senior, 71-: Aged
8. Case_in_country: Replace null with zero
9. Symptom_onset: Convert year 20 to year 2020
10. If_onset_approximated: Replace NA with 0
11. Hosp_visit_date, exposure_start, exposure_end: Convert year 20 to year 2020
12. Visiting.Wuhan, from.Wuhan and death are Boolean and hence categorical. 0 and 1 are to be replaced with N and Y respectively.
13. Recovered: This is a Date. Zero values to be replaced with null.
14. Source: Not required.
15. Link: Remove column
16. X.1 – X.6: Remove as they contain null only

2-covid_19_data.csv

This data set appears to be an aggregate of data sets numbered 4, 5 and 6 below and therefore we will not be using it.

Column	Encoding	Null?
SNo		N
ObservationDate	Categorical:Ordinal	N
Province.State	Categorical: Nominal	Y
Country.Region	Categorical: Nominal	N
Last.Update	Categorical:Ordinal	N
Confirmed	Continuous::Sequnetial	N
Deaths	Continuous::Sequnetial	N
Recovered	Continuous::Sequnetial	N

File Name: 3-2019_nCoV_data.csv

This dataset appears to be a duplicate of the above. We will not be using this dataset.

4-time_series_covid_19_confirmed.csv

This data set provides detailed view of the confirmed cases from 22nd of Jan to 20th of February 2020.

The significance of this data set is that we can analyse the trend over 31 days and across different locations.

Column	Encoding	Null?
Province.State	Categorical: Nominal	Y
Country.Region	Categorical: Nominal	N
Lat	Categorical: Ordinal	N
Long	Categorical : Ordinal	N
X1.22.20	Continuous	N
X1.23.20	Continuous	N
X1.24.20	Continuous	N
X1.25.20	Continuous	N
X1.26.20	Continuous	N
X1.27.20	Continuous	N
X1.28.20	Continuous	N
X1.29.20	Continuous	N
X1.30.20	Continuous	N
X1.31.20	Continuous	N
X2.1.20	Continuous	N
X2.2.20	Continuous	N
X2.3.20	Continuous	N

Column	Encoding	Null?
X2.4.20	Continuous	N
X2.5.20	Continuous	N
X2.6.20	Continuous	N
X2.7.20	Continuous	N
X2.8.20	Continuous	N
X2.9.20	Continuous	N
X2.10.20	Continuous	N
X2.11.20	Continuous	N
X2.12.20	Continuous	N
X2.13.20	Continuous	N
X2.14.20	Continuous	N
X2.15.20	Continuous	N
X2.16.20	Continuous	N
X2.17.20	Continuous	N
X2.18.20	Continuous	N
X2.19.20	Continuous	N
X2.20.20	Continuous	N

Data Wrangling, Cleansing & Transformation

1. Pivot transforming to a new column called ‘Day’ and ‘Count. Each row is multiplied by 31 days.

5-time_series_covid_19_deaths

- Has a similar structure to file ‘time_series_covid_19_confirmed’? Except that it contains death figures. 6-time_series_covid_19_recovered.csv
- Has a similar structure to file ‘time_series_covid_19_confirmed’? Except that it contains number of recovered.

Transformation over data sets 4, 5 and 6

Join data sets 4, 5 and 6 as they have same key columns to get the number of death, recovered and confirmed together. They will join using state, region, latitude, longitude and day.

Initial Data Exploration

Data Quality

Initial examination of the data set shows some inconsistency in the data. For example, the number of infected captured for China is far less those captured for US. This needs further investigation. For the purpose of this exercise we base findings on what the data provides:

Figure 1: Tabular visualization of data shows

- China, Hong Kong, Taiwan and Vietnam were hit first during this period of January 21st to February 25th.
- Other than China itself where the number of males infected were much greater, gender does not appear to be a factor
- Grouping of Age shows the number of infections and particularly death is much greater among more senior persons.
- Filtering to China, we can see that the initial death and infections starting from Wuhan.

Figure 2:

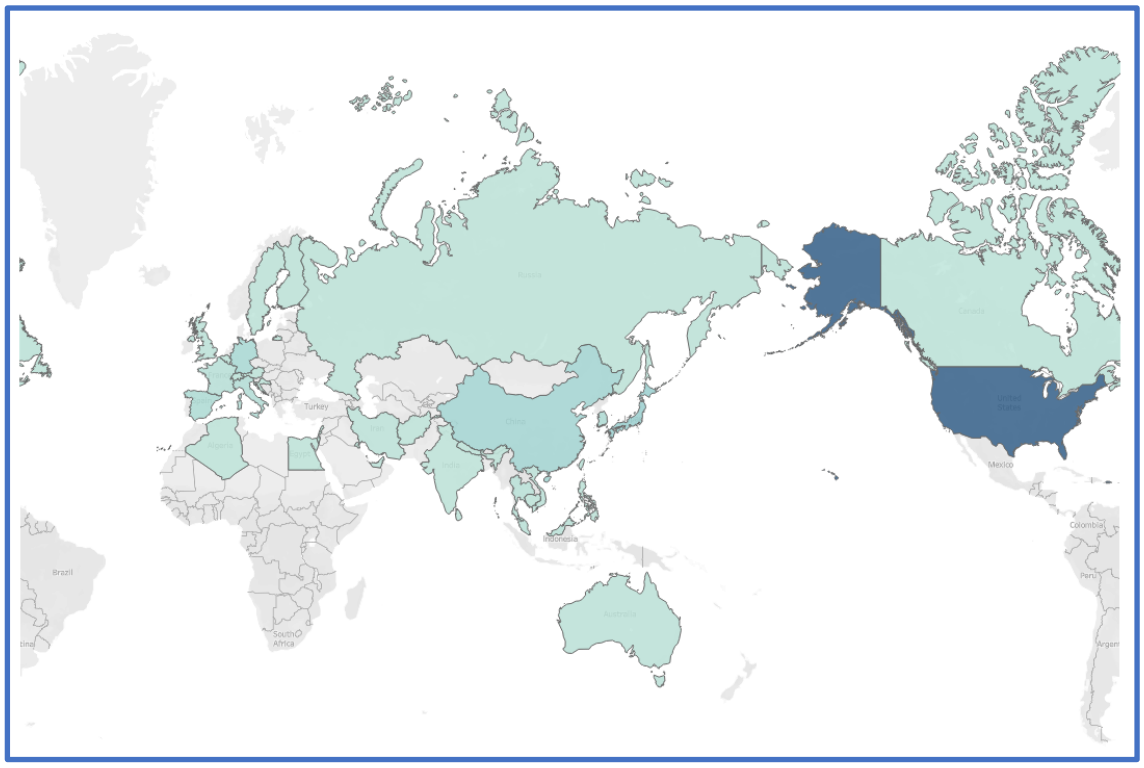

Figure 2: Showing the trend in the number of infections.

Figure 3: Geographical representation of the scale of infections across the world.

Figure 4: Highlight Tables and Tree maps appear to be a better visualization of the ranking of the spread when compared with the geographic map.

Figure 5: Showing the date range for when death occurred was between 21st of Jan to 23rd of Jan 202. Again, this needs further investigation as it seems doubtful.

Sheet 4

country	gender	IF [AgedNumer..	Number..	death
USA	Null	AgedNumericd	2	
		Child	1	
		Mature	12	
		Senior	1	
		Teen	1	
		Unknown	1,169	
		Young	6	
	20	Unknown	1	
	female	AgedNumericd	48	0
		Mature	52	0
		Senior	34	0
		Teen	4	0
		Unknown	46	0
		Young	32	0
	male	AgedNumericd	36	0
		Mature	76	0
		Senior	53	0
		Teen	4	0
		Unknown	81	0
		Young	33	0
	NA	Teen	1	0
		Unknown	75	0
Japan	female	AgedNumericd	20	0
		Mature	36	0
		Senior	20	0
		Teen	1	0
		Young	22	0
	male	AgedNumericd	34	0
		Child	4	0
		Mature	58	0
		Senior	37	0
		Teen	1	0
		Young	18	0
	NA	Child	1	0
		Unknown	5	0
China	female	AgedNumericd	6	5
		Child	3	0
		Mature	21	1
		Senior	16	6
		Teen	3	0
		Unknown	1	0
		Young	23	0
	male	AgedNumericd	12	11
		Mature	48	4

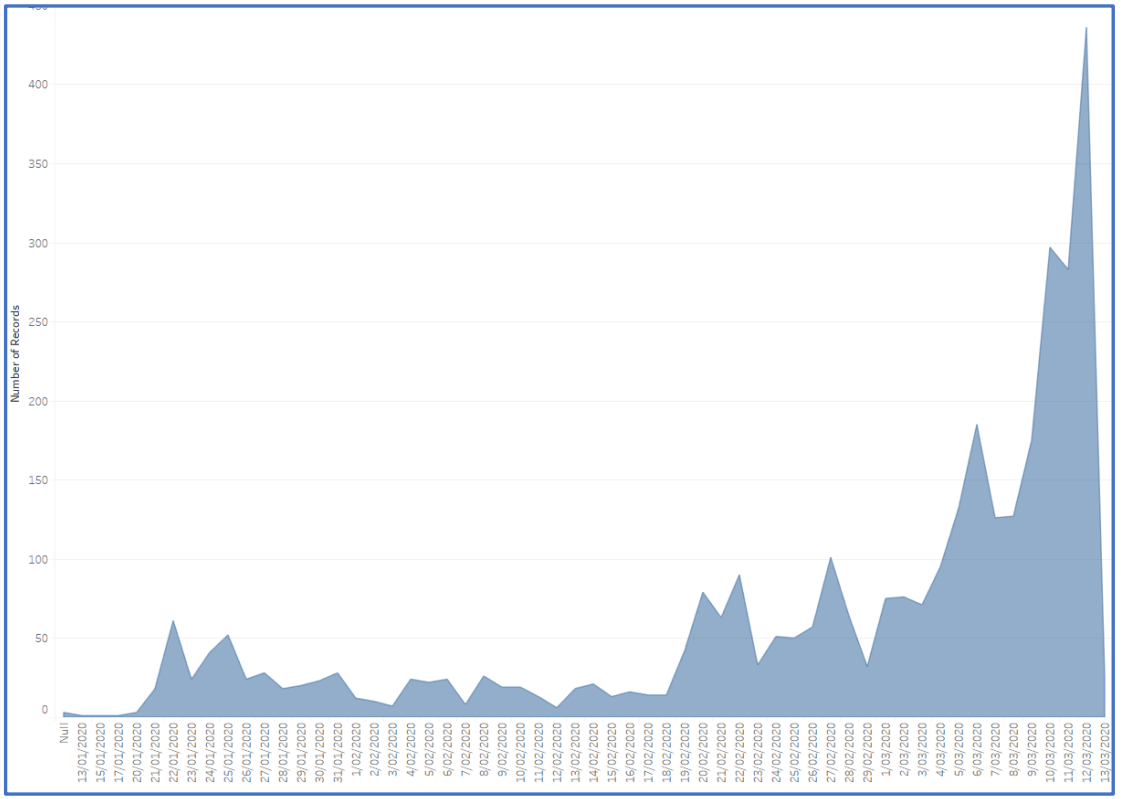

Figure 2 – Showing the trend

country	
Null	165
Afghanistan	1
Algeria	1
Australia	15
Austria	2
Bahrain	17
Belgium	1
Cambodia	1
Canada	12
China	197
Croatia	1
Egypt	1
Finland	1
France	56
Germany	168
Hong Kong	102
India	3
Iran	18
Israel	1
Italy	86
Japan	257
Kuwait	9
Lebanon	1
Malaysia	23
Nepal	1
Phillipines	3
Russia	2
Singapore	112
South Korea	114
Spain	116
Sri Lanka	1
Sweden	1
Switzerland	9
Taiwan	34
Thailand	41
UAE	21
UK	20
USA	1,768
Vietnam	16