Group 13:

Yashna Sachdev
Ujjval Sharma
Zhengyan Shi
Xingyin Lou

Preparing Ourselves Better for The Next Pandemic

Learnings from COVID-19 Datasets

Professor:
Dr. Kylie Ariel Bemis

4/15/2021

Introduction

Problem:

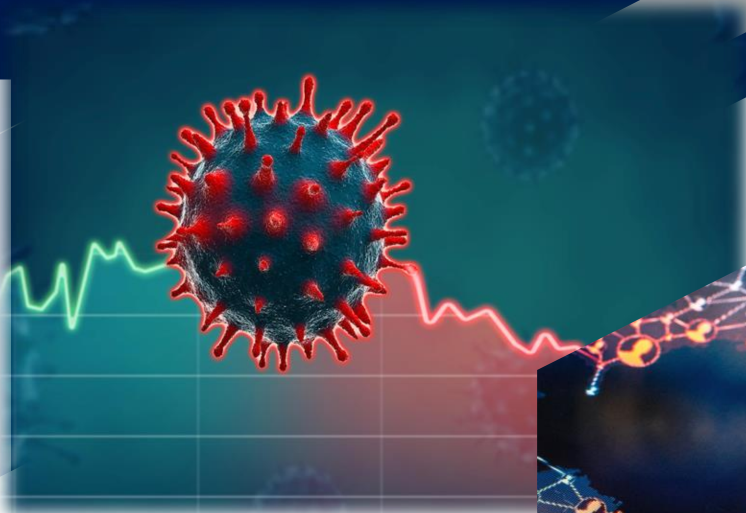
- COVID-19 and its effect on the whole world
  - Unprecedented challenge to our lives
- Impact on different regions is not the same
  - Response time in taking immediate action varies
  - survival rate, cases per region

Main Datasets used:

- Geo Distribution- Ex. cases by region, deaths
- Government Response- Ex. confirmed cases, schools closing, contact tracing, testing

Questions:

1. Who can use this analysis and datasets?
2. What patterns or trends can we visualize?
3. How do we plan to plan to manipulate and analyze multiple large datasets?



Goals - Objectives & Methods Used

Unsupervised Machine Learning

Mission 1

Compare trends between countries with cases increasing/decreasing at a similar time

Clustering Analysis (K- means)

More Data Visualization

Mission 2

Calculate the survival rate for continents and the survival rate over time

Time Series

Mission 3

Compare countries' government response rate in implementing measures due to the pandemic

Time Series

Unsupervised Machine Learning

Mission 4

Explore a general order of government's response to the situation

Association Rules

Methods (continued)

- R packages
- Unsupervised Machine Learning
- Data transformation/tidying
- Relationship exploration
- Visualizations

Details will be discussed further...

Mission 1

Grouping countries with similar number of cases around the same time

- Best category attribution based on the similarity between the point distance.
- Similar samples grouped together; clusters formed
- K = 5
- Each day’s number of cases in each country; divide those countries into clusters.

➢ **Clustering:**

A collection of data points aggregated together because of certain similarities using only input vectors without labelled outcomes.

K-means Algorithm:

Identifies *k*, then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Cluster means:

	20200303	20200314	20200316	20200317	20200318	20200319	20200320
1	4.160428	7.358289	10.12299	9.475936	11.69519	13.74332	18.10695
2	1.000000	14.500000	41.00000	33.000000	34.50000	82.50000	109.50000
3	14.000000	511.000000	823.00000	887.000000	1766.00000	2988.00000	4835.00000
4	33.058824	136.588235	165.23529	123.647059	171.64706	177.35294	232.52941
5	71.714286	889.428571	983.14286	1344.428571	1450.57143	1650.14286	2687.28571

	20200321	20200322	20200323	20200324	20200325	20200326
1	19.11765	22.82353	26.22995	26.79679	31.92513	34.00535
2	161.50000	156.50000	268.50000	199.00000	190.00000	159.50000
3	5374.00000	7123.00000	8459.00000	11236.00000	8789.00000	13963.00000
4	260.94118	268.52941	231.52941	329.11765	368.70588	496.88235
5	2611.00000	2566.42857	2347.57143	3188.57143	3019.57143	3533.85714

Results

mx	my	mz	na	nc	ne	ng	ni
4	1	1	1	1	1	1	1
nl	no	np	nz	om	pa	pe	pf
4	1	1	1	1	1	4	1
pg	ph	pk	pl	pr	ps	pt	py
1	1	1	5	1	1	1	1
qa	ro	rs	ru	rw	sa	sb	sc
1	4	1	5	1	1	1	1
sd	se	sg	si	sk	sl	sm	sn
1	1	1	1	1	1	1	1
so	sr	ss	st	sv	sx	sy	sz
1	1	1	1	1	1	1	1
tc	td	tg	th	tj	tl	tn	tr
1	1	1	1	1	1	1	4
tt	tw	tz	ua	ug	uk	us	uy
1	1	1	4	1	5	3	1
uz	va	vc	ve	vg	vi	vn	vu
1	1	1	1	1	1	1	1

Clusters of similar number of cases over time

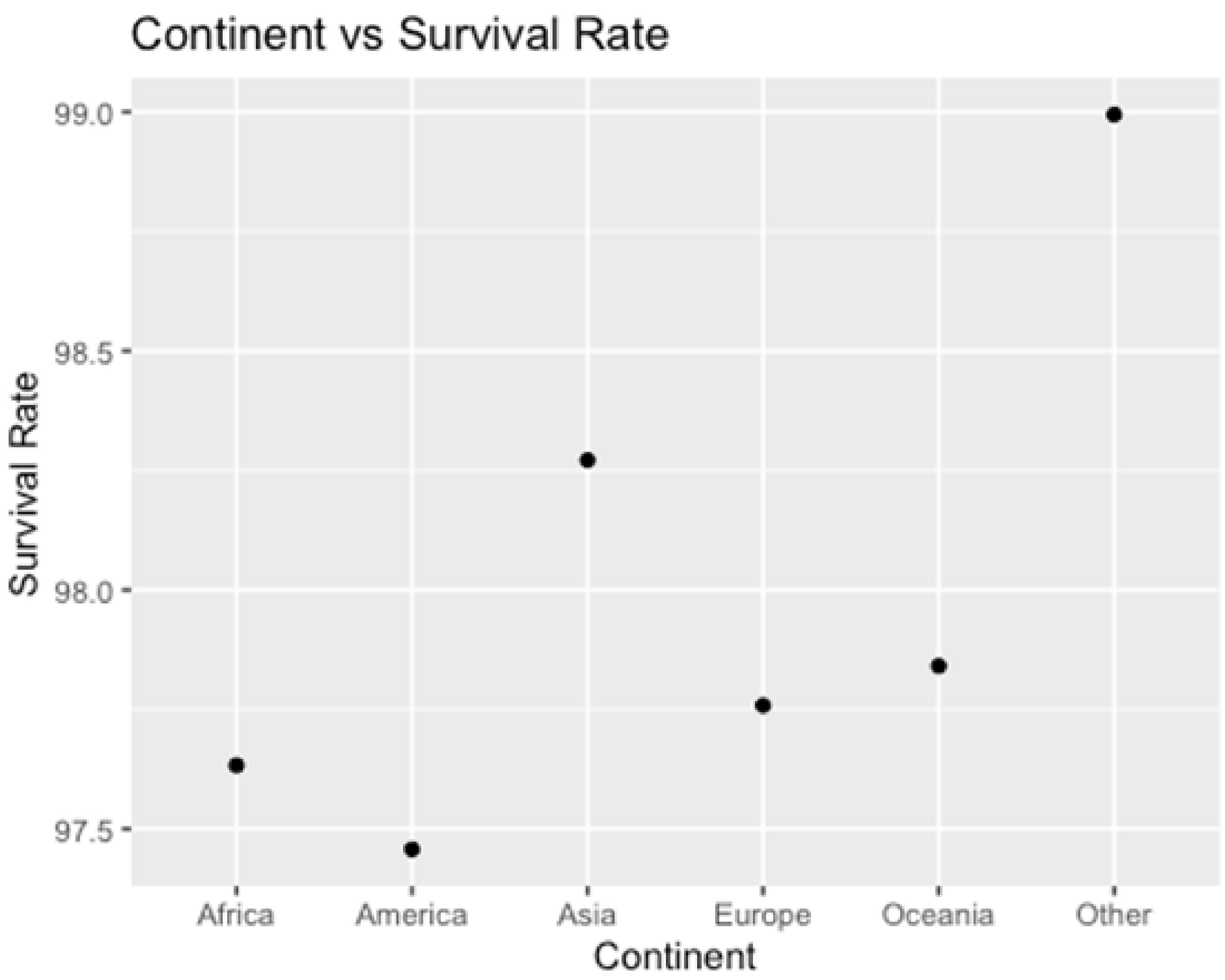
1	2	3	4	5
New Caledonia, Norway, Panama, Saudi Arabia, Uruguay	Brazil, India	America	Belgium, Canada, Peru	Spain, Russia, United Kingdom

- 5 clusters demonstrating countries with similar tendency of increasing and decreasing cases around the same time.
- Index of clusters each country belongs to

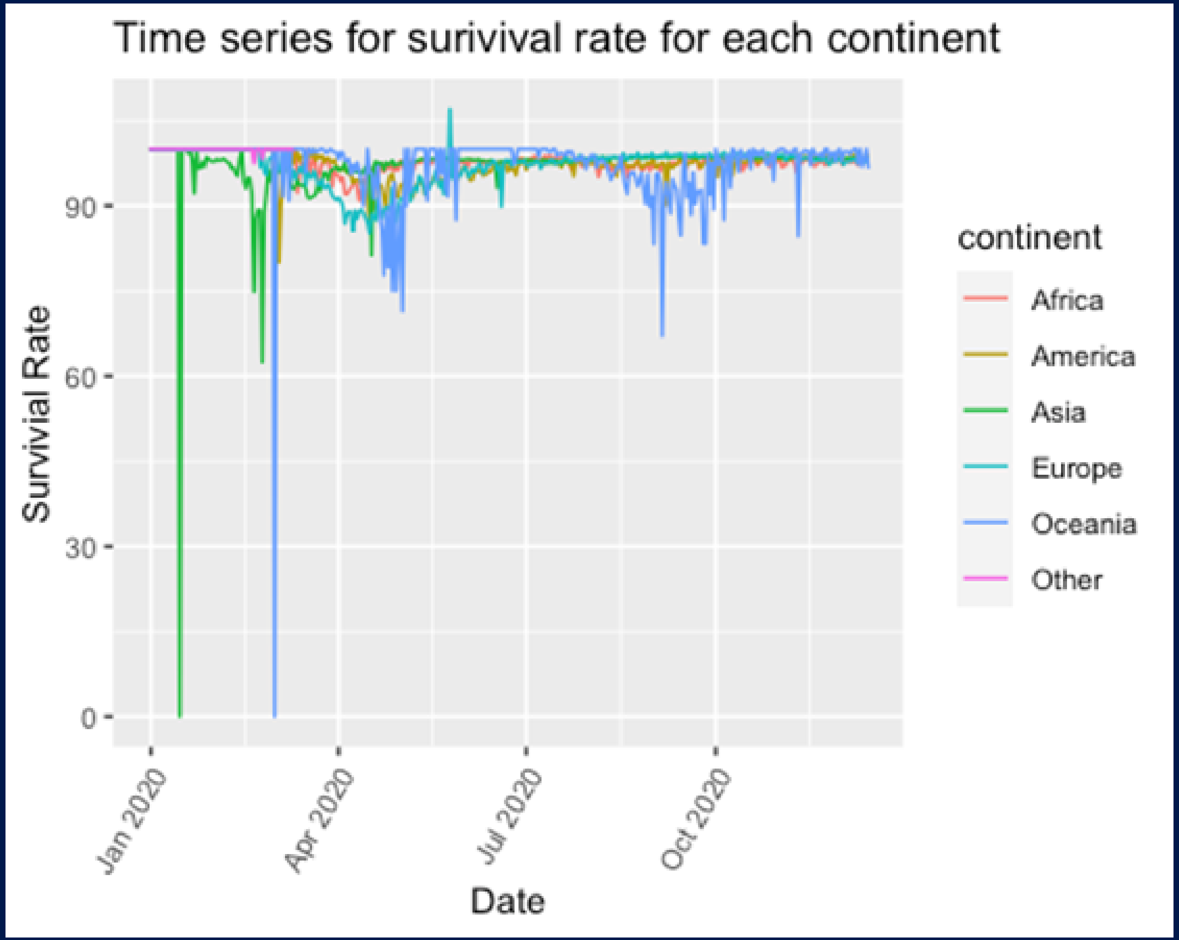
Mission 2

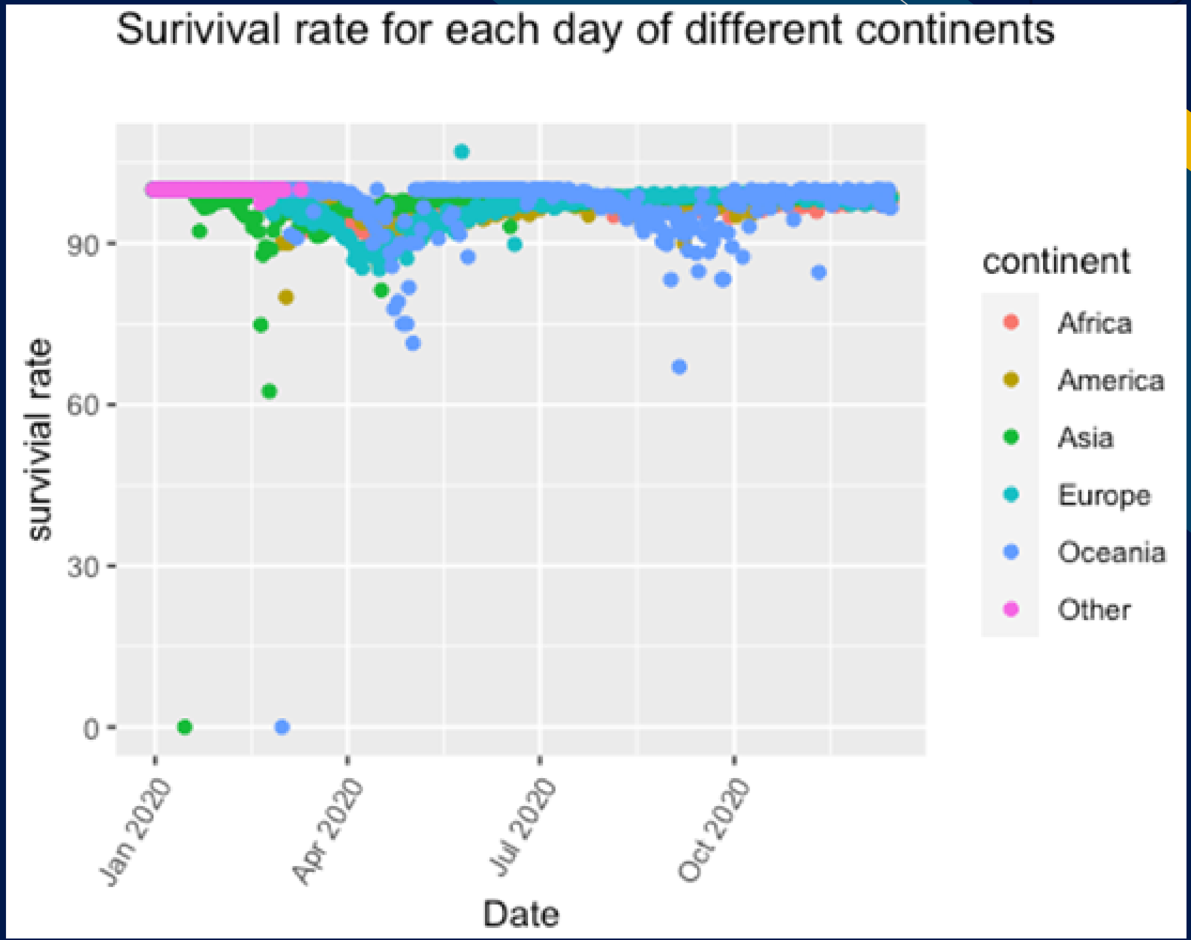
Survival Rate

Here we can see that the overall survival rate for each continent during Covid was more than 95%. However, America wasn't able to recover well in comparison to other continents.



Survival Rate over The Time





We can see that the survival rate from Jan 2020 to Dec 2020 was mostly 100%, but there few days in between where the survival rate was 0% as there were no cases reported. We can see an interesting point that number of people healthy each day were more than the number people died. Therefore, the survival rate never dropped less than 0%. Also, only Asia and Oceania reported 0 cases between Jan and Apr. Moreover, Africa' survival rate is most 100%.

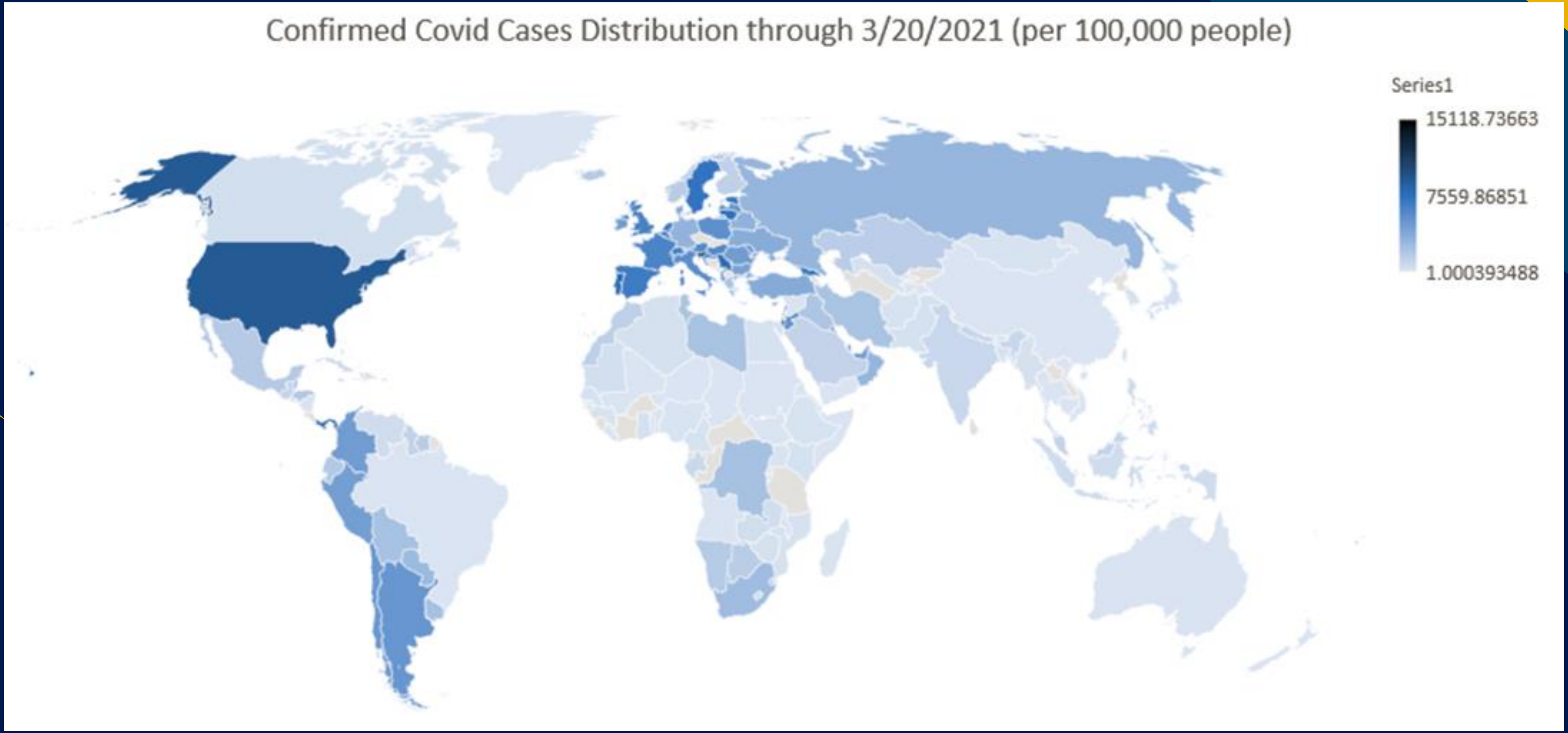
Mission 3

Government Response

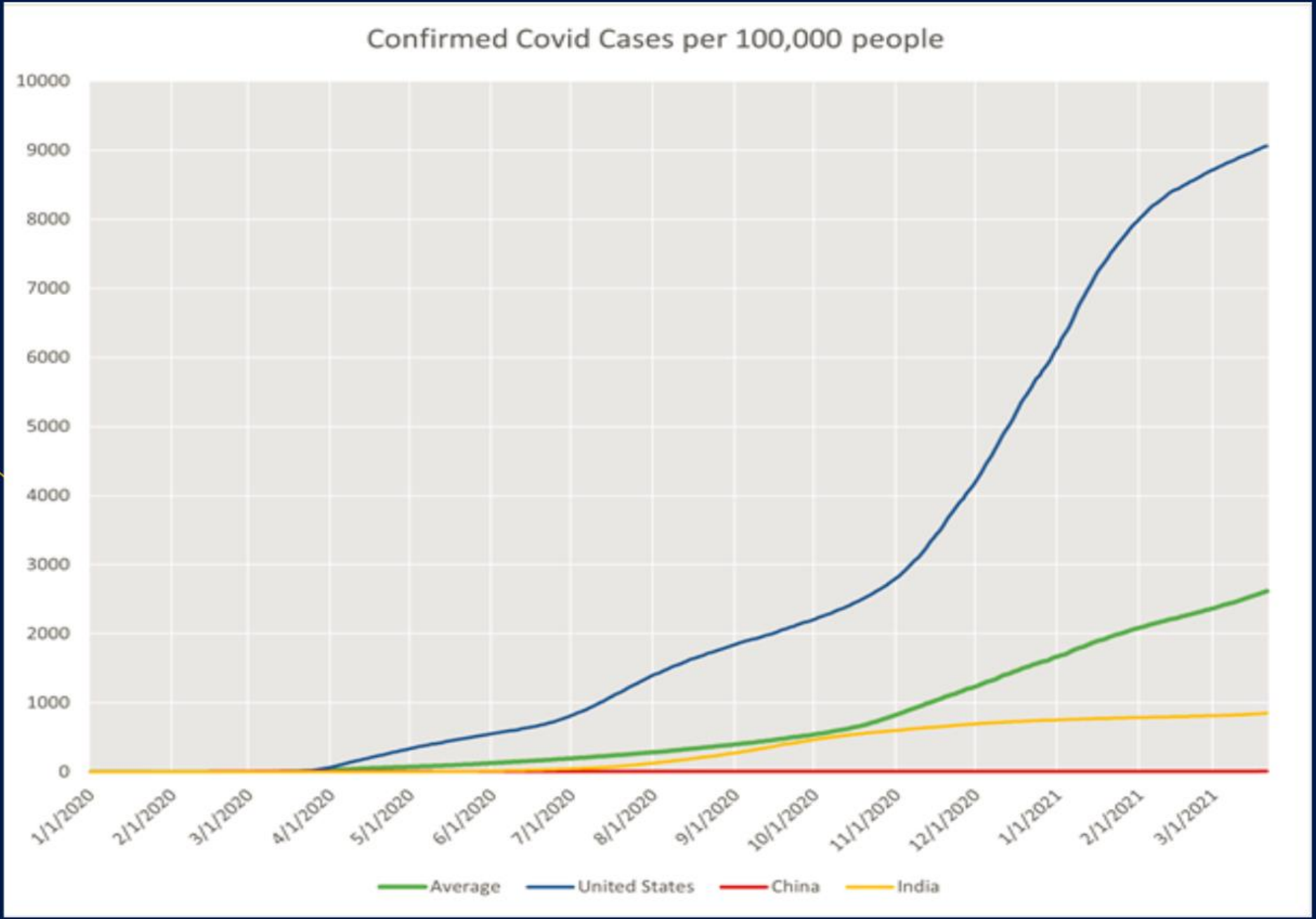
What are governments' responses to the situation and how fast does each country respond

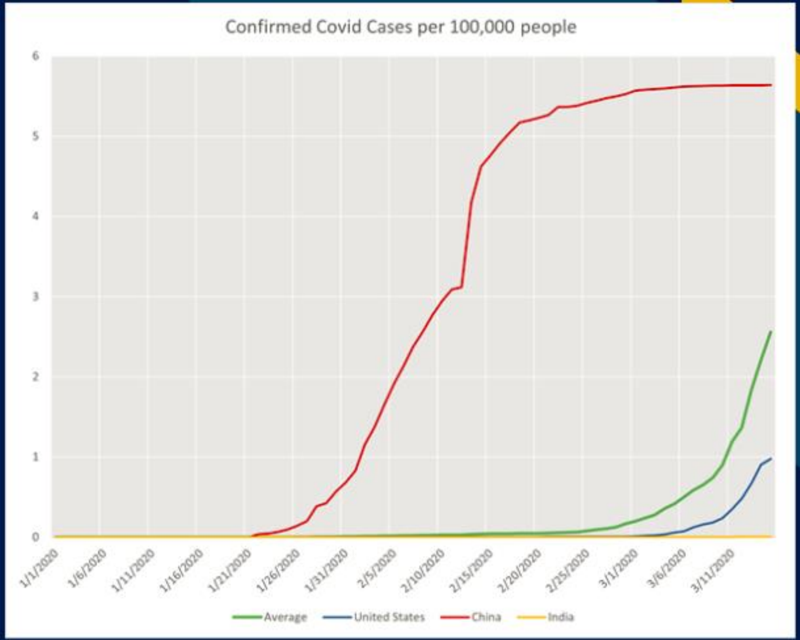
- Measures
  - OxCGRT Index
- Responding speed
  - Rank (time series)
  - Situations (# confirmed cases)

Covid-19 Confirmed Diagnosis Rate Distribution

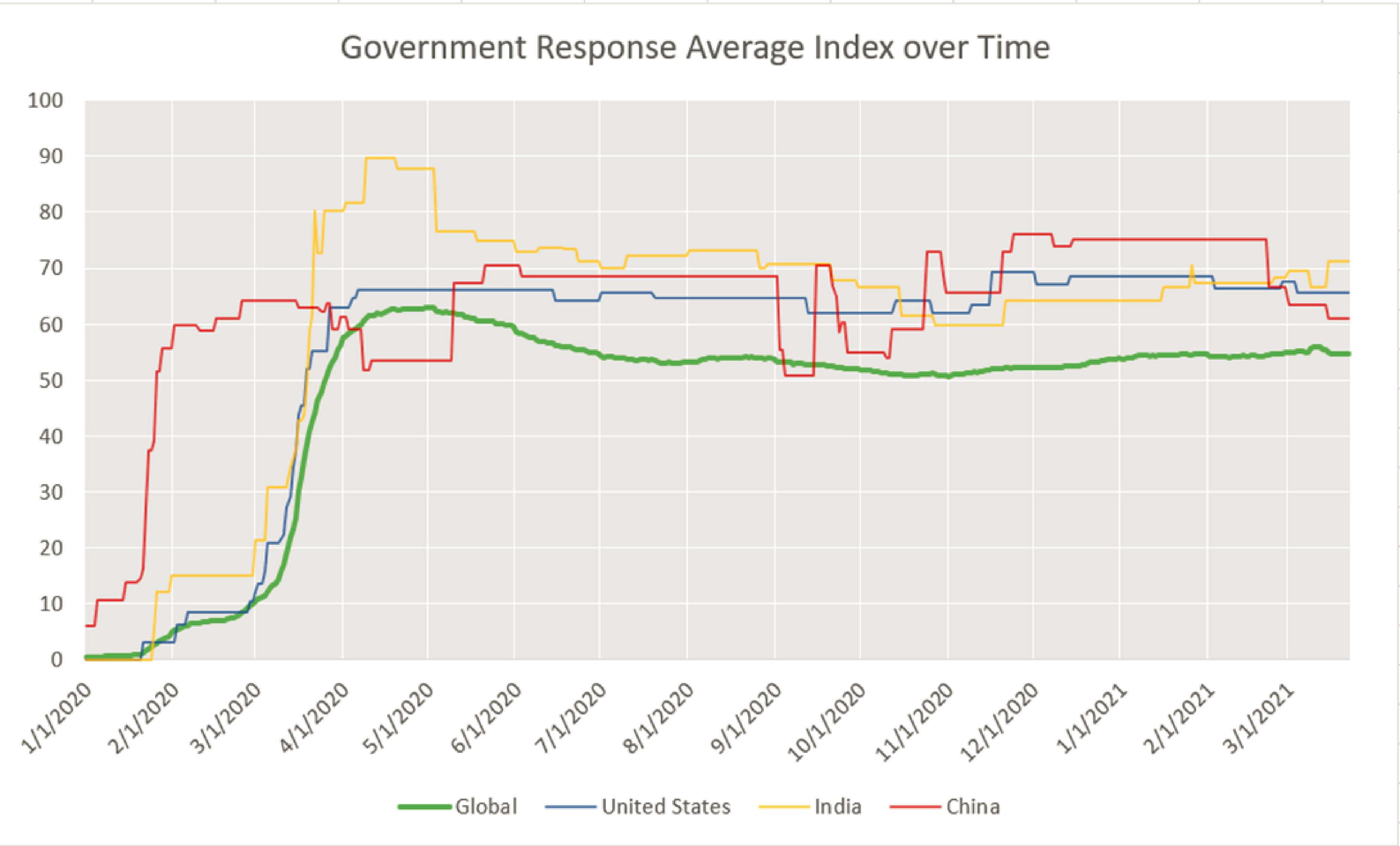


Dynamic Trend of Covid-19





The Oxford COVID-19 Government Response Tracker (OxCGRT Index)



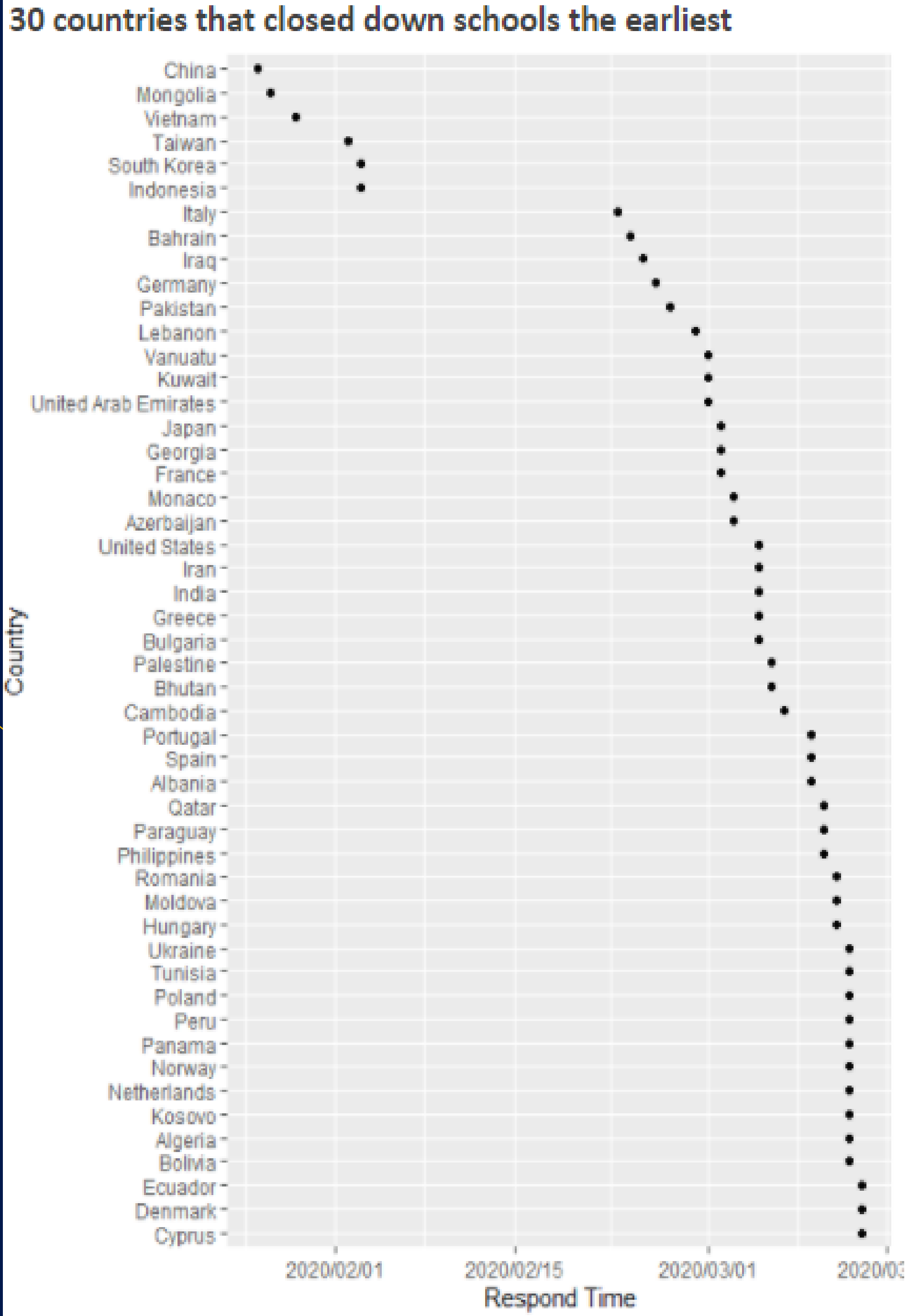
ggplot2 for visualization
LOCF for imputation

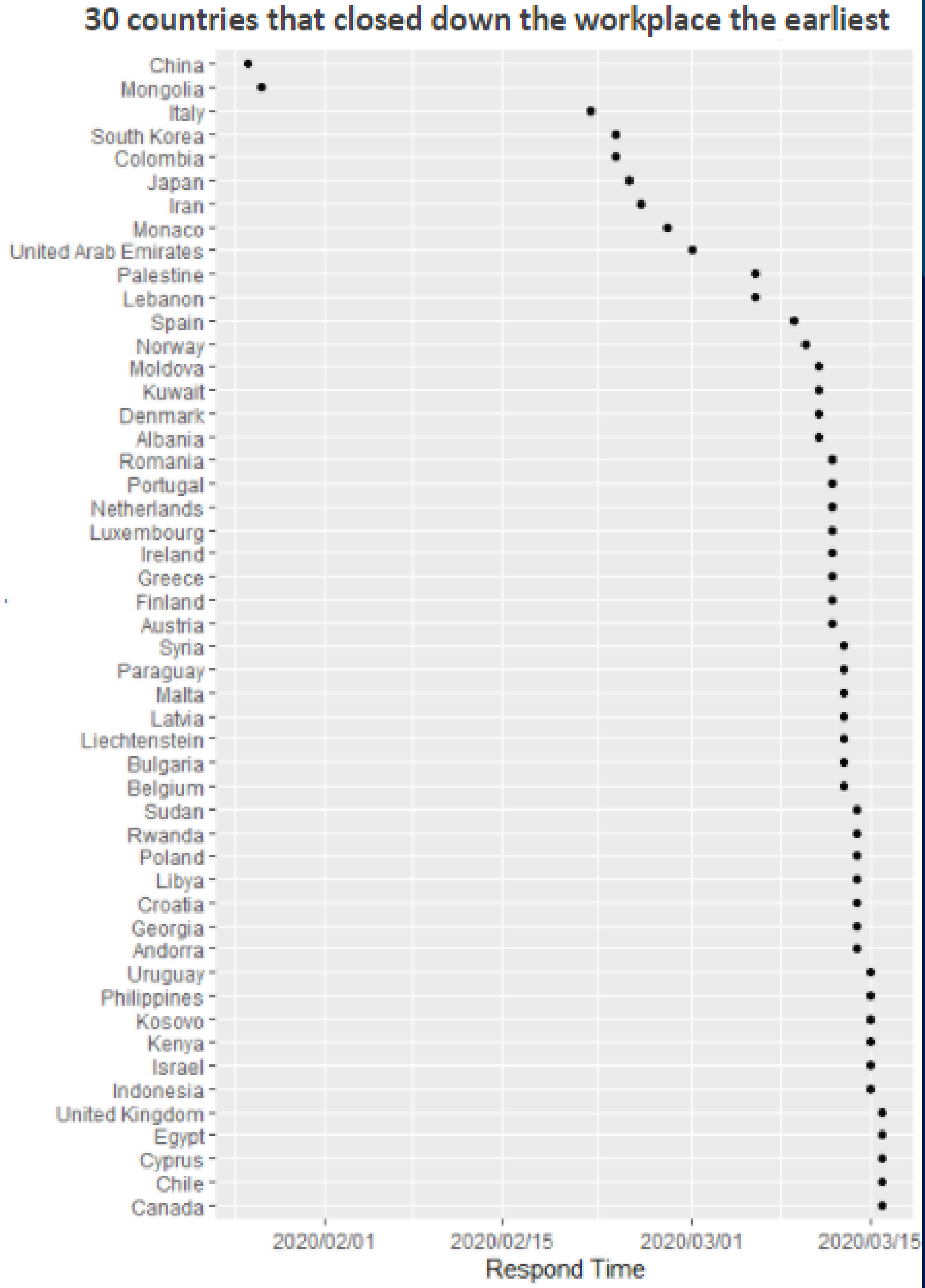
Speed matters more than scale!

8 policy indicators
- **travel controls**
- **school closure**
- **facial covering**
- **stay at home**
- **…**

4 economic indicators
- **fiscal support**
- **unemployment benefits expansion**
- **…**

5 healthcare indicators
- **contact tracing**
- **vaccination**
- **…**





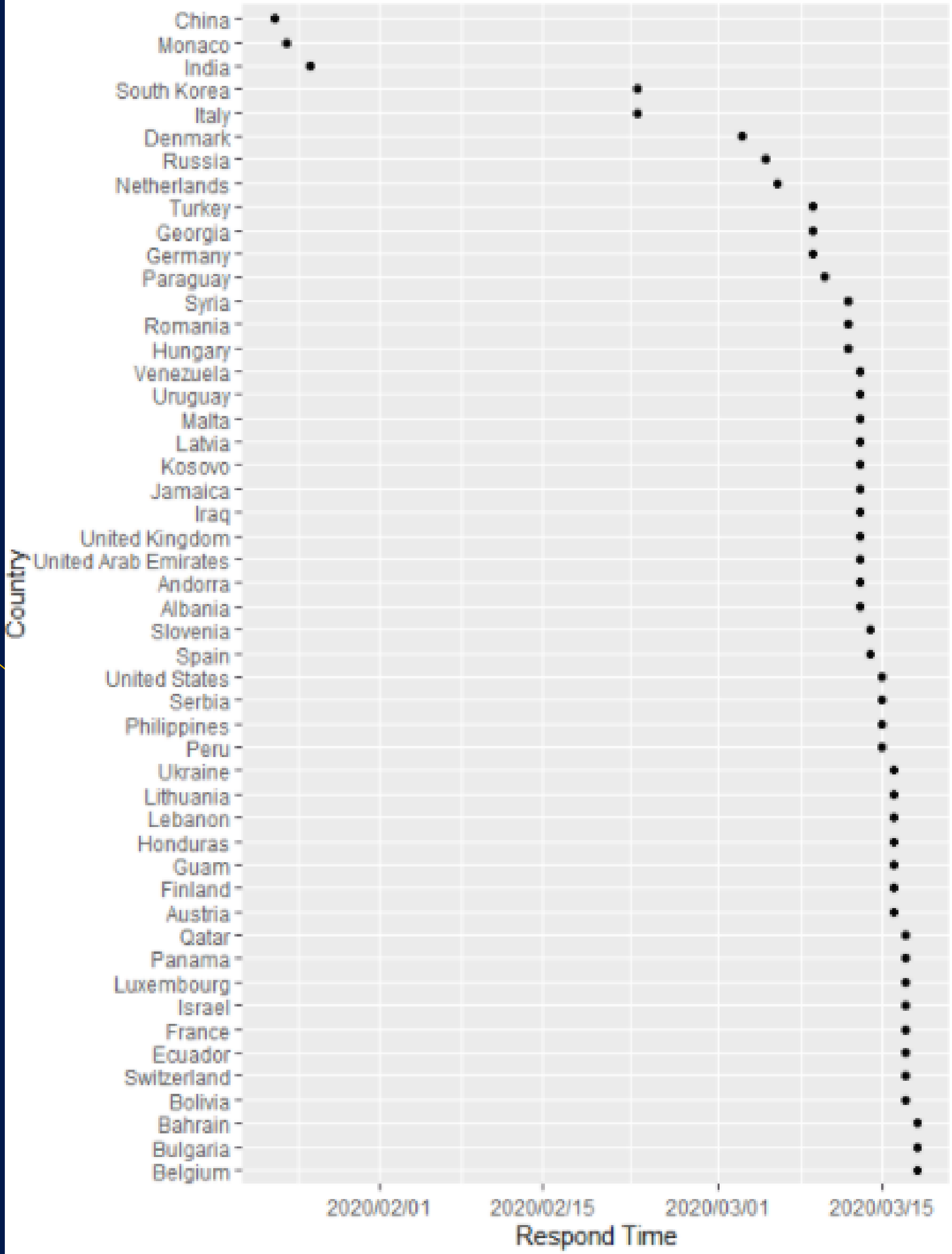
as.date transformation missing data removed

The gap between 1st and 10th is more than 75 days

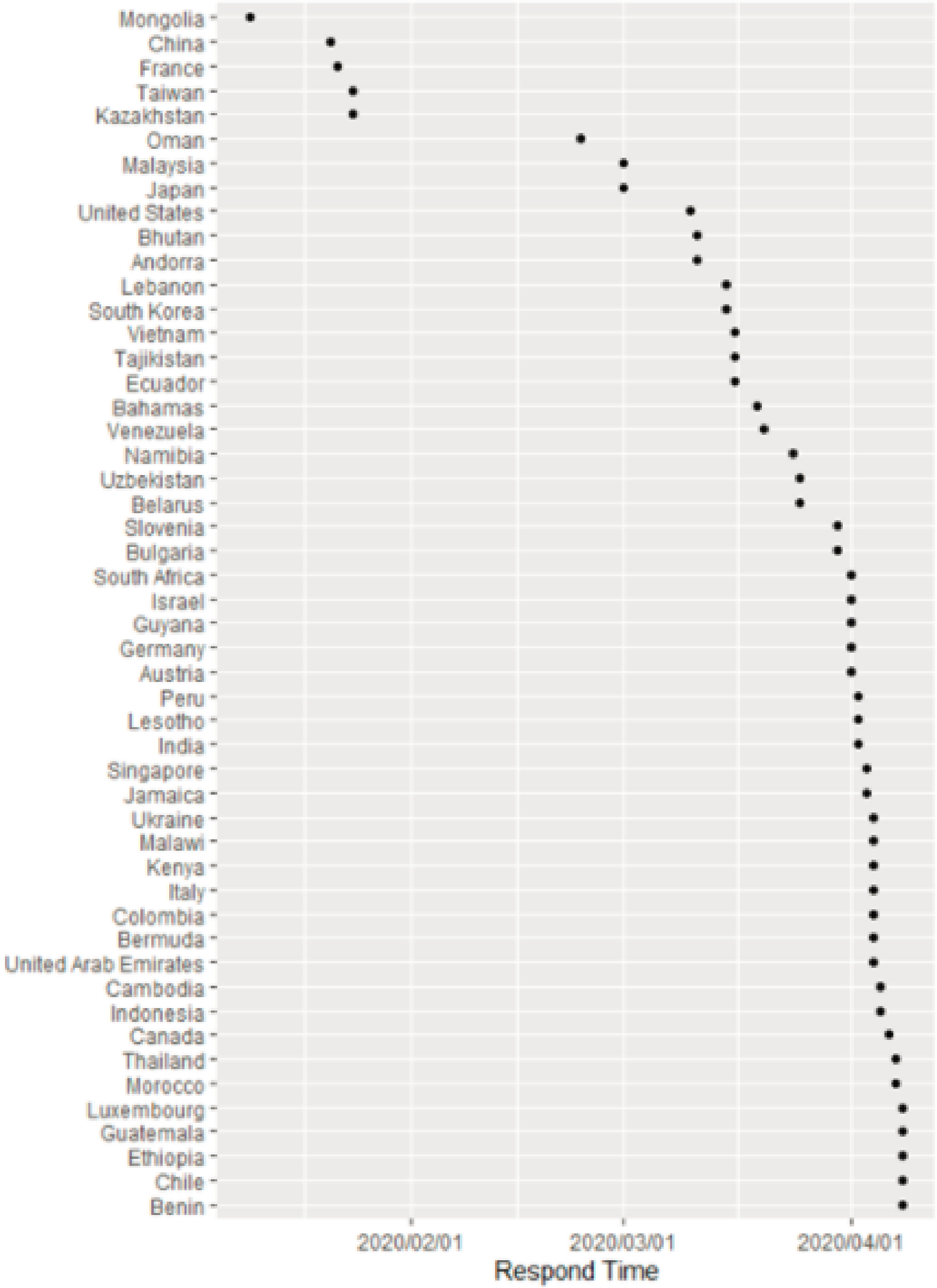
The five countries that close down the schools are all Asian countries

Italy was one the three countries that close down the workplace earliest, way ahead of most of European countries.

30 countries that requires the residents to stay at home



30 countries that requires facial covering the earliest



as.date transformation
missing data removed

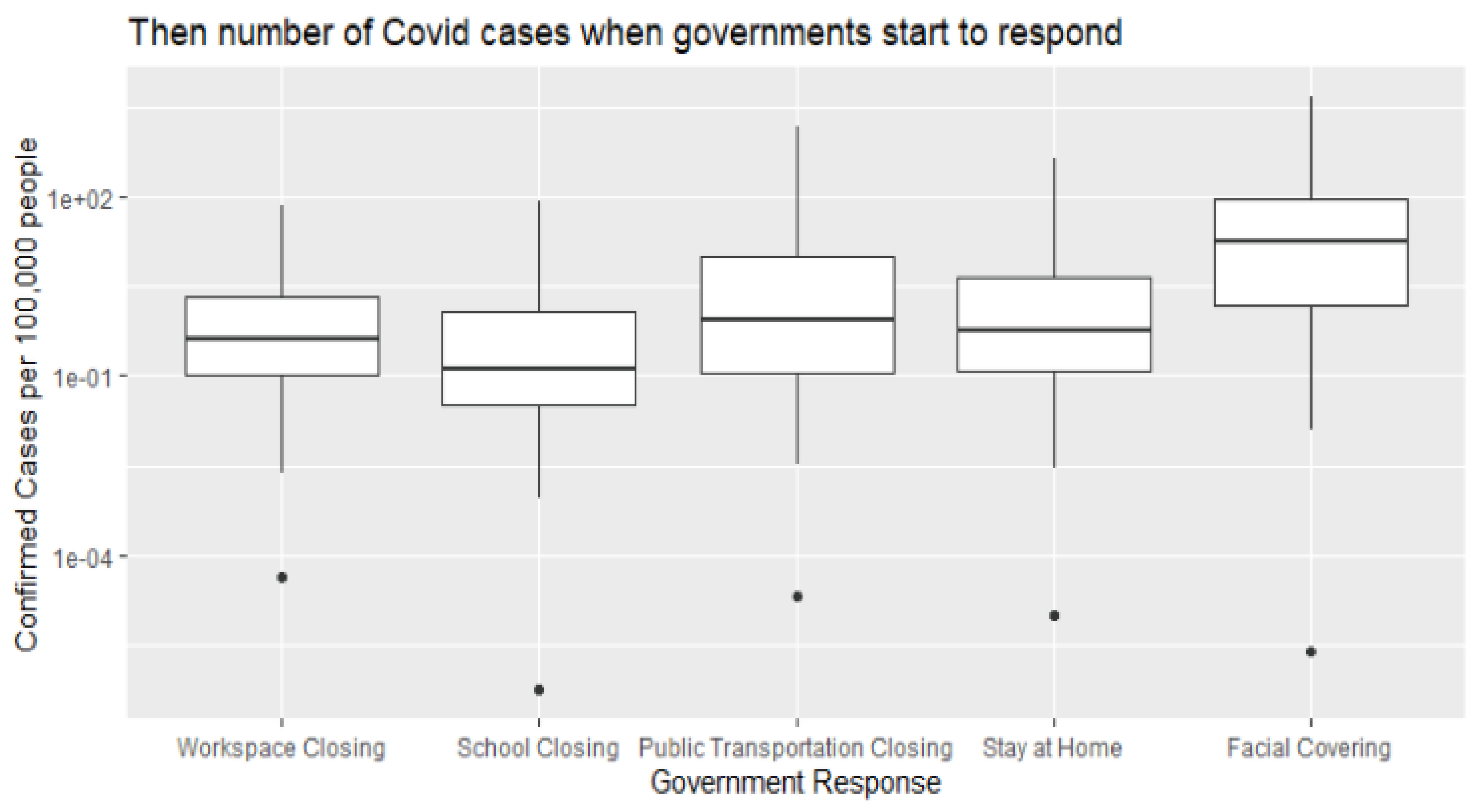
The gap between 1st and 10th is more than 75 days

Compared with other measures, most of the countries hesitate longer to take “stay-at-home” and facial covering policies
•the late response of facial covering might due to the shortage of masks at that time

France was one of the three earliest countries that requires facial covering.

School -> workspace -> stay at home -> public transportation -> facial covering

LOCF for imputation
Log transformation
ggplot2 for data visualization



Indicators/ References

	Res <fctr>	median <dbl>
1	Workspace Closing	0.29945448
2	School Closing	0.09784929
3	Public Transportation Closing	0.67216495
4	Stay at Home	0.44882533
5	Facial Covering	10.67479468

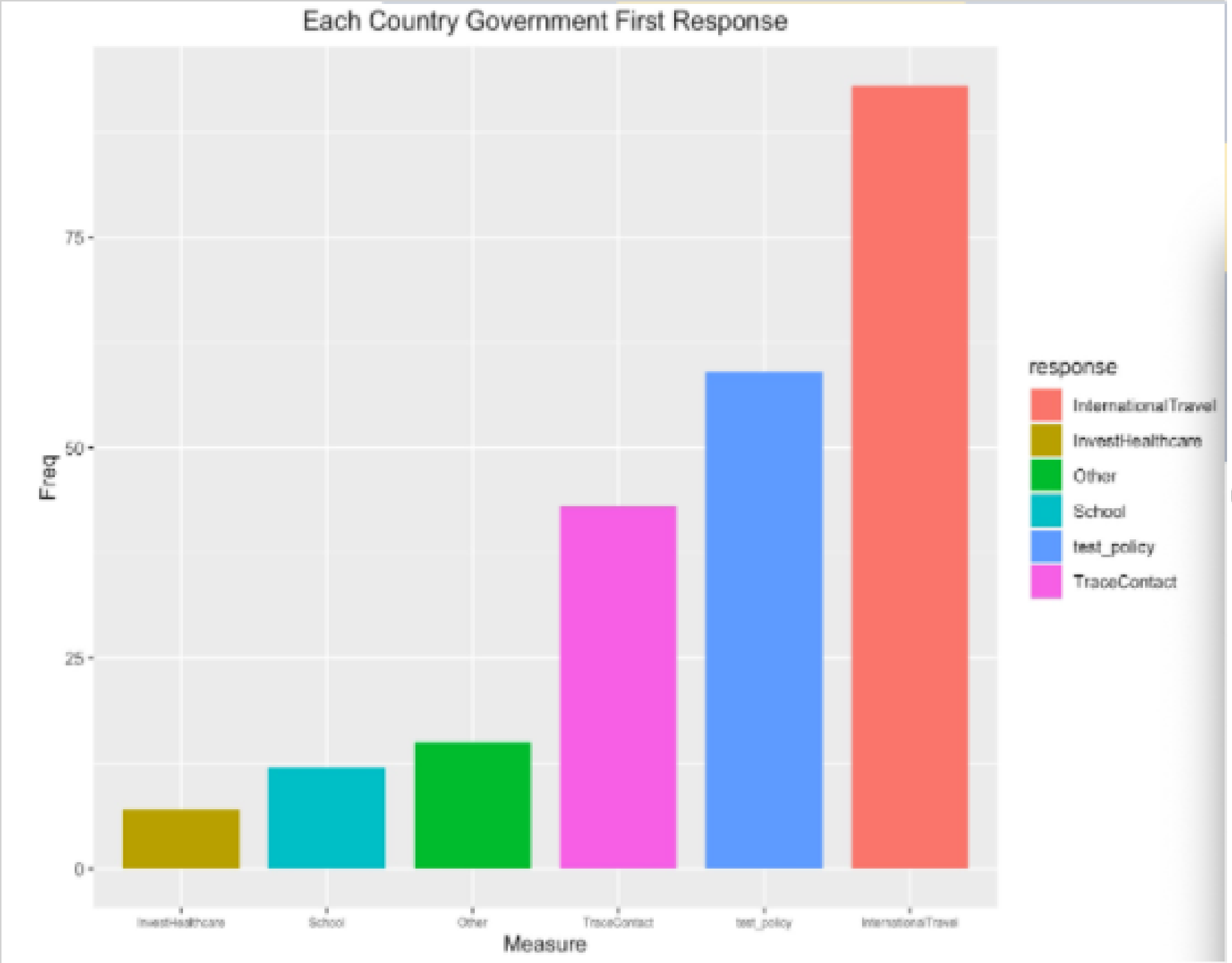
	Res <fctr>	mean <dbl>
1	Workspace Closing	3.561801
2	School Closing	2.504467
3	Public Transportation Closing	49.411082
4	Stay at Home	7.085524
5	Facial Covering	118.466966

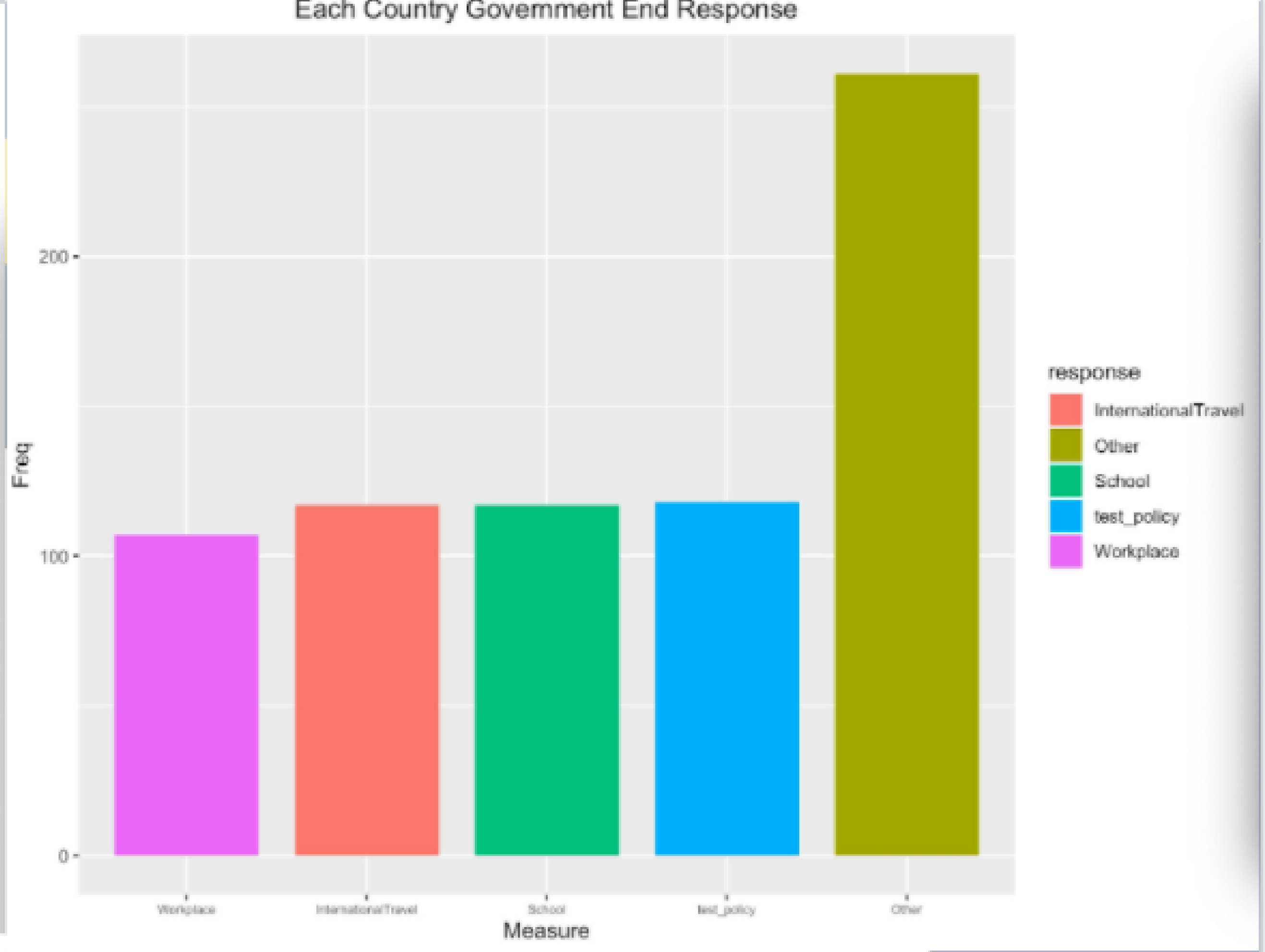
When each government start to take mobility restriction measures, what is the confirmed diagnosis rate?

Mission 4
Association Rules

Explore a general order of government's response to the situation

- Rank governments responses
- During different time periods
- Grouping Affinity
- Introduce ‘support’, ‘confidence’, ‘lift’
- Association Rules



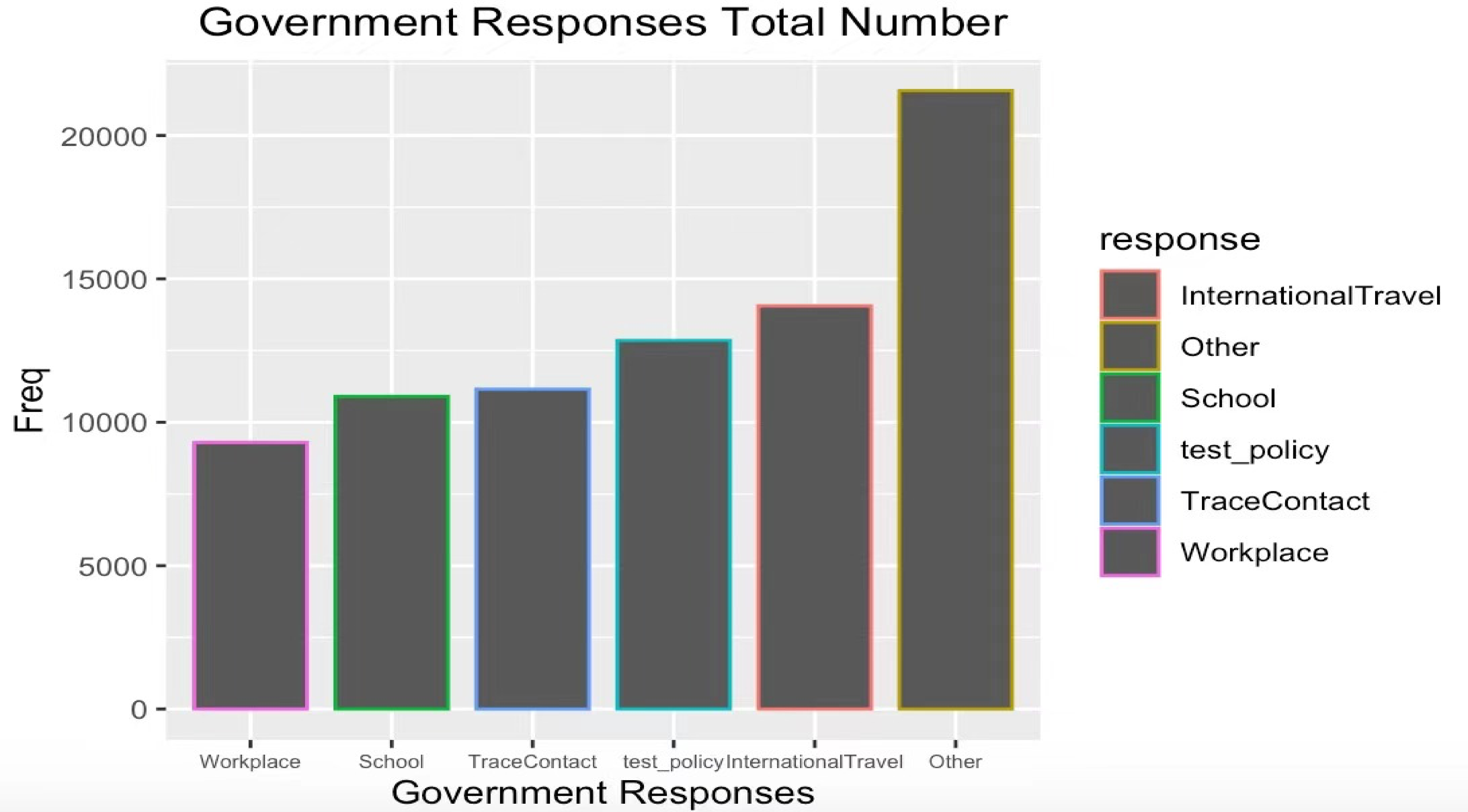


There are 93 countries governments' first thing is to control international travel (Rank No.1)

There are 59 countries governments' first thing is to implement testing policies (Rank No.2)

118 countries governments' last response is to implement testing policies

Rank No.2 of the last thing is to close the school and restrict the international travel


Government Responses Total Number
Freq
20000
15000
10000
5000
0
Workplace
School
TraceContact
test_policy
InternationalTravel
Other
Government Responses
response
InternationalTravel
Other
School
test_policy
TraceContact
Workplace

Association Rules

Association Rules (Association Rules) reflect the interdependence and association between one thing and other things. If there is a certain association relationship between two or more things, then one thing can be predicted by other things.

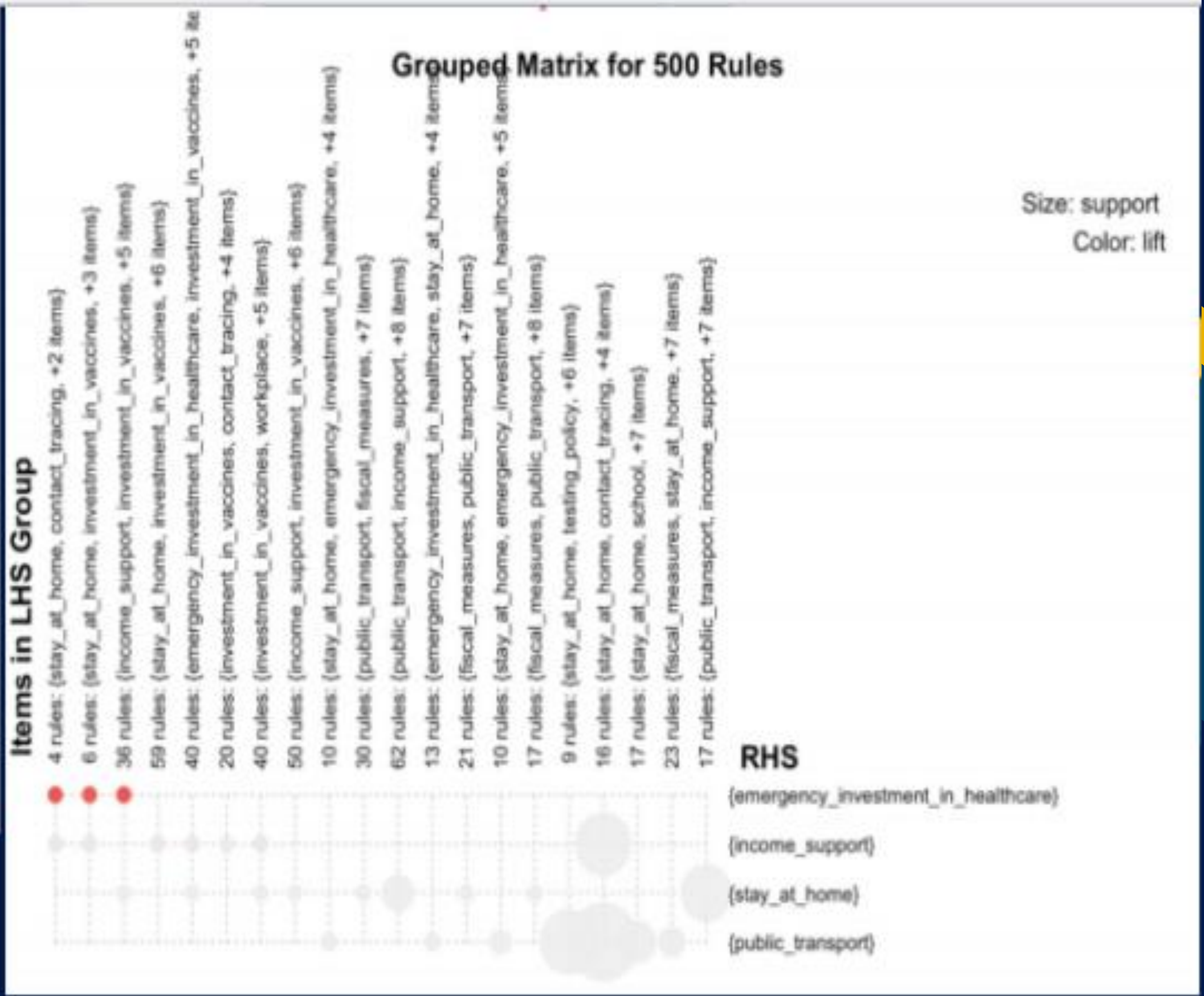
For all the emergency measures being implemented in a certain country at a certain moment, we call it a transaction.

‘Support’=P{A, B} is the ratio of the number of transactions that include both A and B in the transaction set to the number of all transactions.

'Confidence'=P{B|A}: In the case of including A, the probability of including B. That is to say: It is the ratio of the number of transactions containing A and B to the number of transactions containing A.

'Lift'=P{B|A} / P{B}: The ratio of (In the case of including A, the probability of including B) and (the probability of B). Actually, this measures the increasing effect of A on the probability of B.

```
> inspect(rules0.sorted_lift[1:20])
    lhs                              rhs                                support     confidence coverage    lift     count
[1] {contact_tracing,
     income_support,
     investment_in_vaccines,
     stay_at_home}                => {emergency_investment_in_healthcare} 0.001251095 0.5263158 0.002377080 24.10798 20
[2] {contact_tracing,
     income_support,
     investment_in_vaccines,
     stay_at_home,
     testing_policy}              => {emergency_investment_in_healthcare} 0.001251095 0.5263158 0.002377080 24.10798 20
[3] {contact_tracing,
     income_support,
     international_travel_control,
     investment_in_vaccines,
     stay_at_home}                => {emergency_investment_in_healthcare} 0.001251095 0.5263158 0.002377080 24.10798 20
[4] {contact_tracing,
     income_support,
     international_travel_control,
     investment_in_vaccines,
     stay_at_home,
     testing_policy}              => {emergency_investment_in_healthcare} 0.001251095 0.5263158 0.002377080 24.10798 20
[5] {contact_tracing,
     income_support,
     investment_in_vaccines,
     school}                      => {emergency_investment_in_healthcare} 0.001188540 0.5135135 0.002314525 23.52157 19
[6] {contact_tracing,
     income_support,
     investment_in_vaccines,
     school,
     stay_at_home}                => {emergency_investment_in_healthcare} 0.001188540 0.5135135 0.002314525 23.52157 19
```

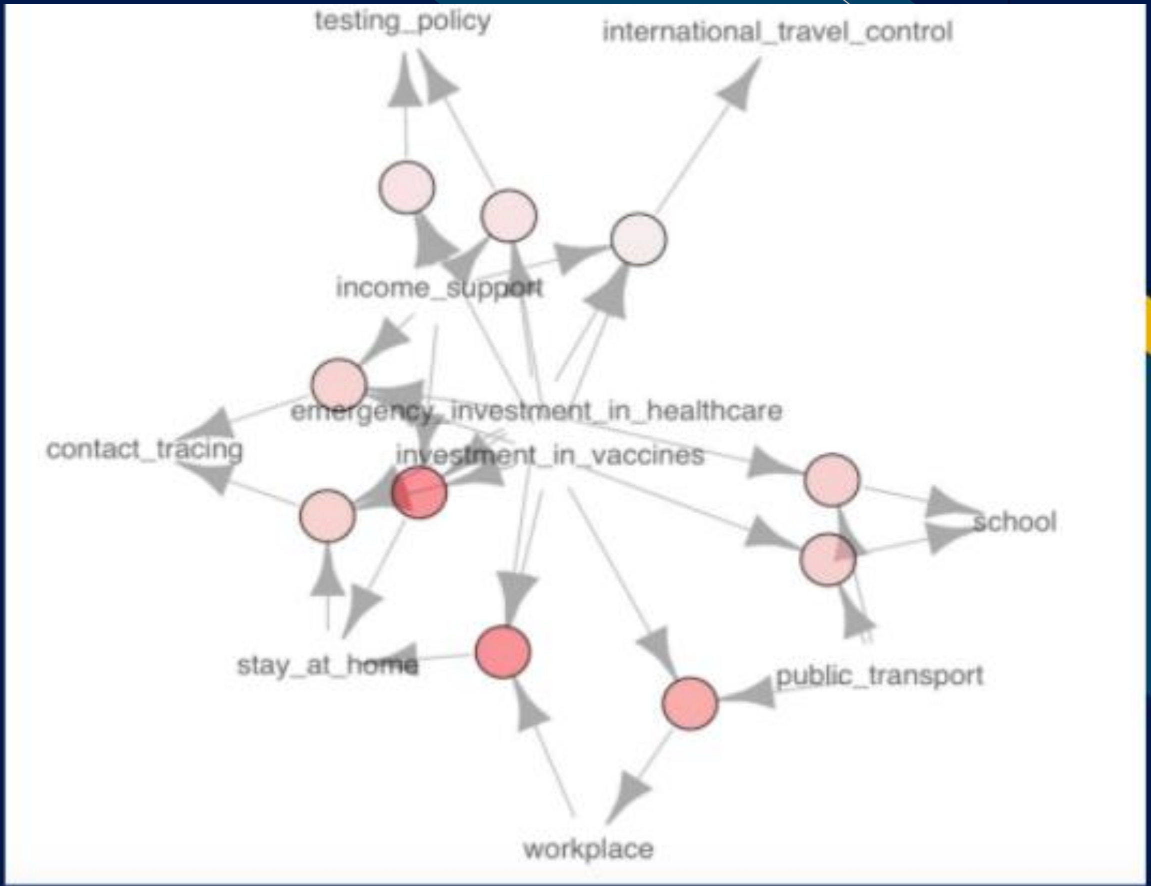


Here we set a pre-decided threshold to filter some rules
(We set support0.001, confidence0.5)
There are 4295 rules in all.

We sorted these rules by 'lift', and here are first several rules.

Grouped Graph of first 500 rules

```
> inspect(rules0.sorted_confidence[1:20])
     lhs                                  rhs                             support confidence    coverage     lift count
[1]  {income_support,
      investment_in_vaccines}          => {testing_policy}            0.002627299          1 0.002627299 1.243079    42
[2]  {investment_in_vaccines,
      public_transport}                => {workplace}                 0.001063431          1 0.001063431 1.719110    17
[3]  {investment_in_vaccines,
      public_transport}                => {school}                    0.001063431          1 0.001063431 1.466202    17
[4]  {emergency_investment_in_healthcare,
      public_transport}                => {school}                    0.012198173          1 0.012198173 1.466202   195
[5]  {emergency_investment_in_healthcare,
      income_support,
      investment_in_vaccines}          => {stay_at_home}              0.001251095          1 0.001251095 1.887367    20
[6]  {emergency_investment_in_healthcare,
      income_support,
      investment_in_vaccines}          => {contact_tracing}           0.001251095          1 0.001251095 1.432951    20
[7]  {emergency_investment_in_healthcare,
      income_support,
      investment_in_vaccines}          => {testing_policy}            0.001251095          1 0.001251095 1.243079    20
[8]  {emergency_investment_in_healthcare,
      income_support,
      investment_in_vaccines}          => {international_travel_control} 0.001251095     1 0.001251095 1.136823    20
[9]  {emergency_investment_in_healthcare,
      investment_in_vaccines,
      workplace}                       => {stay_at_home}              0.001125985          1 0.001125985 1.887367    18
[10] {emergency_investment_in_healthcare,
      investment_in_vaccines,
      stay_at_home}                    => {contact_tracing}           0.001376204          1 0.001376204 1.432951    22
```



This time: Sort by 'confidence'

Graph of first 10 rules

Each node circle means a rule. The word being pointed to is the rhs(end side), the start side's word is the lhs.

Summary

Transformation	Log transformation	tidyverse
Tidying	LOCF imputation	dplyer
Relationship Exploration	Data visualization Boxplot, time series	Map, ggplot2
Unsupervised learning	Clustering Analysis (kmeans)	Association Rules (aprior from arules, arulesviz)

Purpose ☑
Overall goals ☑
Methods ☑

Group 13:

Yashna Sachdev
Ujjval Sharma
Zhengyan Shi
Xingyin Lou

Professor:
Dr. Kylie Ariel Bemis