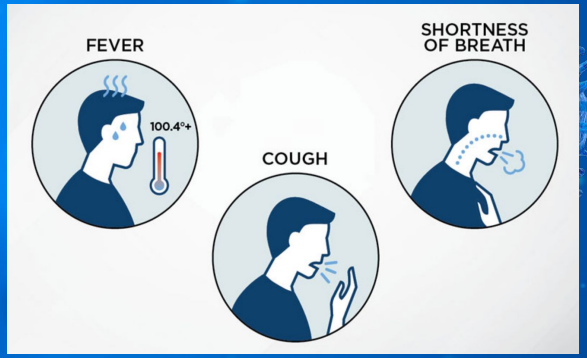
COVID-19 : Analysis & Prediction of Fatality

By : Arushi Sharma, Spatika Krishnan, Vidhey Oza

About COVID-19

- Also known by *novel coronavirus* or *nCoV-19*
- Essentially a new mutation of the already existing coronavirus
- Symptoms close to flu, but much more infectious and deadly

Symptoms



Goal

- To perform statistical analysis and use exploratory and confirmatory data analysis to predict fatality outcome of patients.
- To analyse and predict the progress of CoVid-19 in USA.

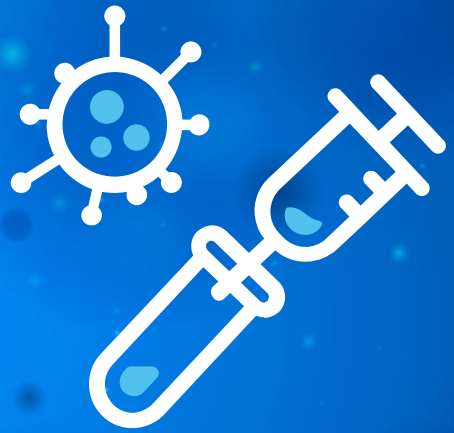
About the Data

Kaggle open dataset; 2 logically separate datasets:

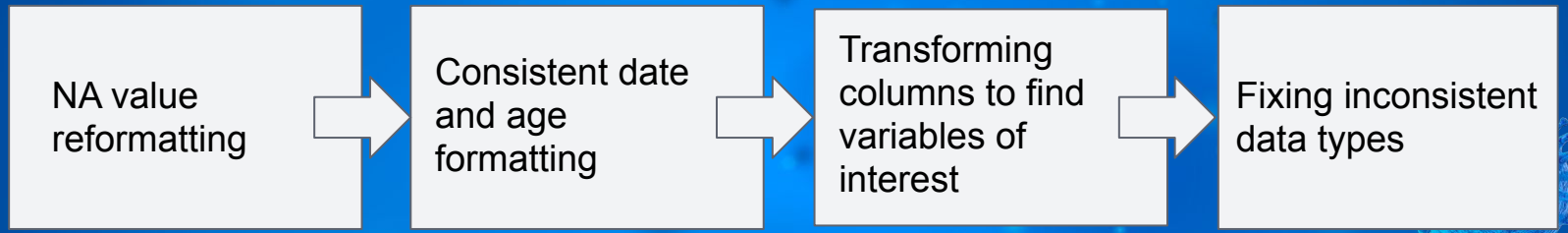
1. Day-wise data points on the number of affected cases, deaths and recovery from CoVID-19
   a. Data available from 22nd January to 8th April

2. Patient-wise data points on who has been confirmed, and their current status as infected, recovered or dead.
   a. Data of about 260,000 patients with 44 features.

Data Cleaning & Processing

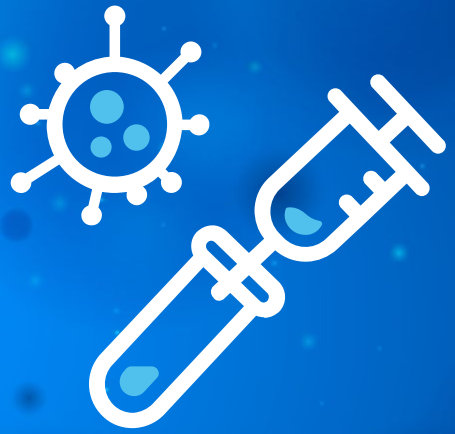


Data Cleaning & Pre-Processing

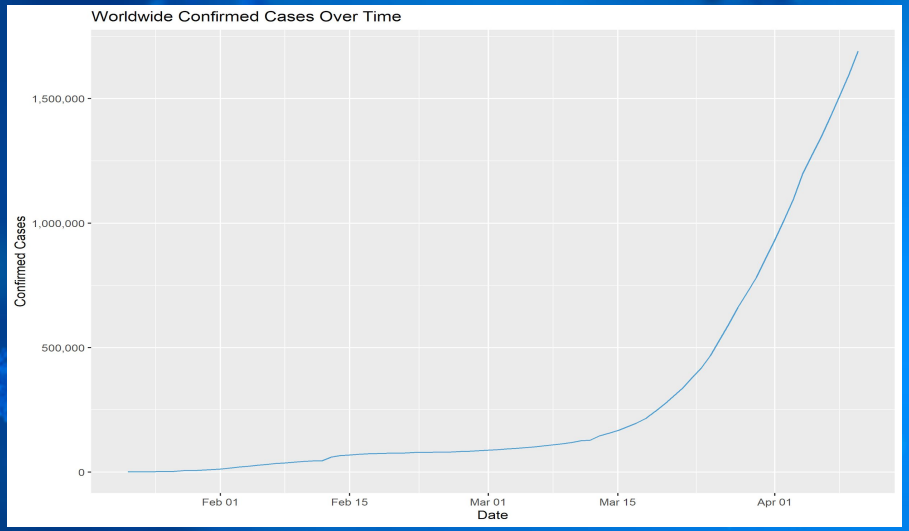


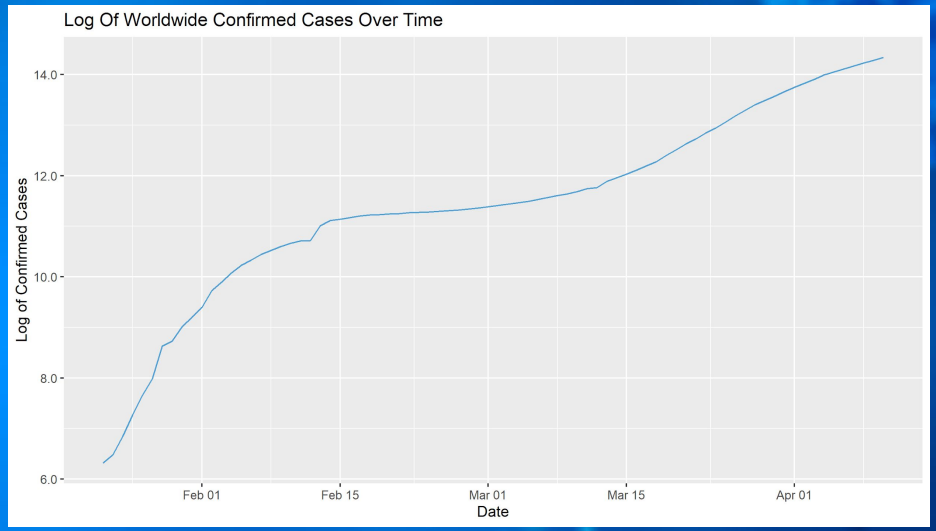
- Data for some countries was available province/state-wise, but not for all
  - Summarized country wise and state-wise for nation-wise analysis
- Generated new columns like number of days to first case in country
  - If newer case, lesser medical resources may be available

Exploratory Data Analysis



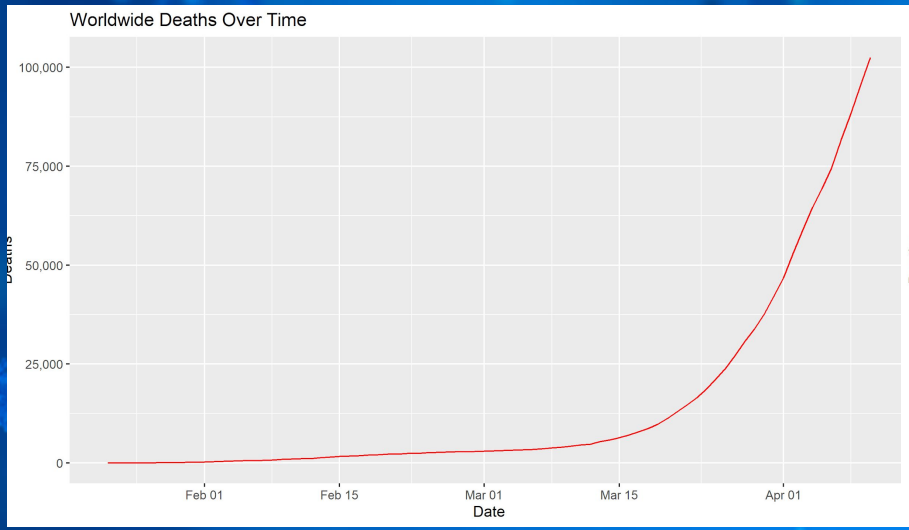
Worldwide confirmed cases (linear & log scale)

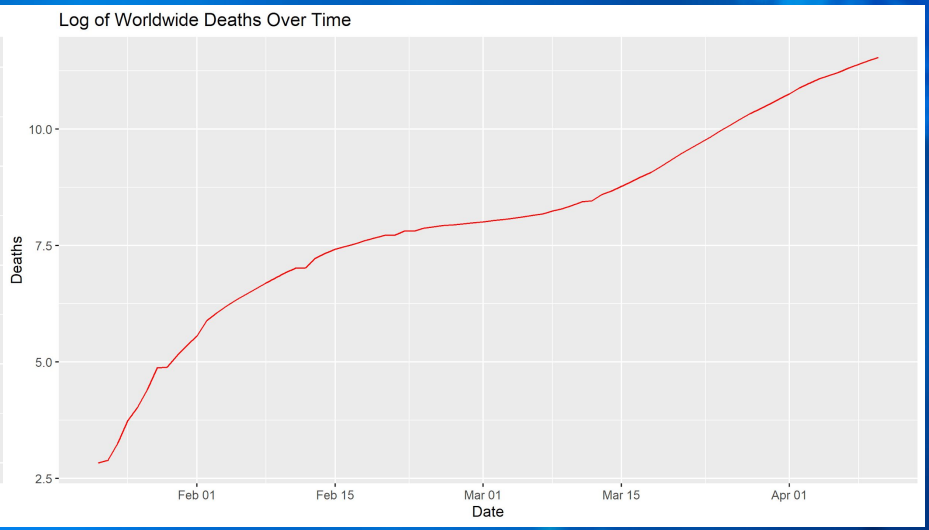




- Pandemic spread is an exponential graph
- Difficult to understand in linear scale – increases too fast
- Solution – logarithmic scale – linear slope means exponential growth

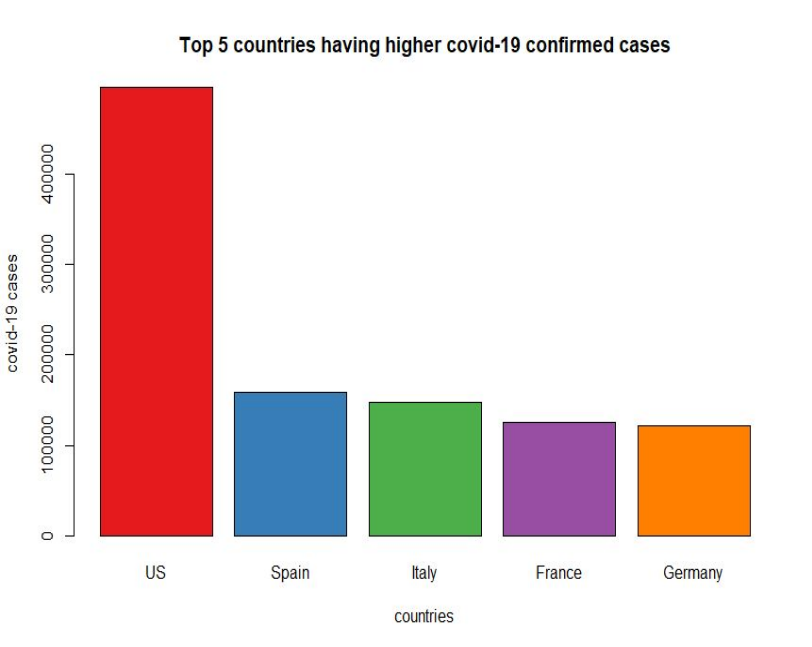
Worldwide deaths (linear & log scale)





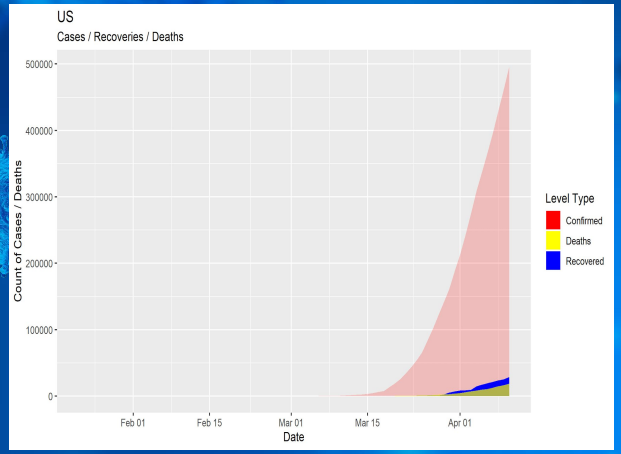
- Surprisingly similar growth to confirmed
- Hints to a nearly constant death to confirmed rate

Most affected countries

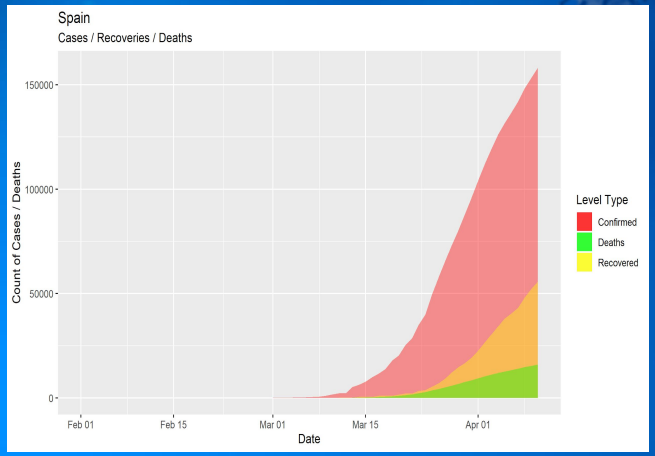


- Top 5 countries affected based on confirmed cases count – US, Spain, Italy, France, Germany
- China not on list - 2 opinions
  - Execution of testing & isolation performed well
  - Data tampered heavily (unfalsifiable)

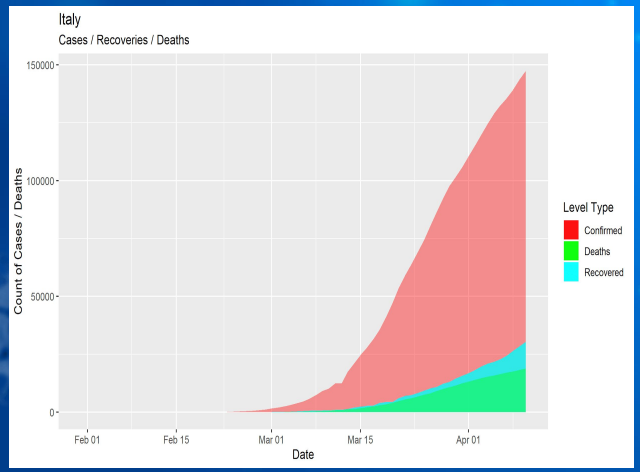
Country-wise Analysis of Cases, Recoveries and Deaths



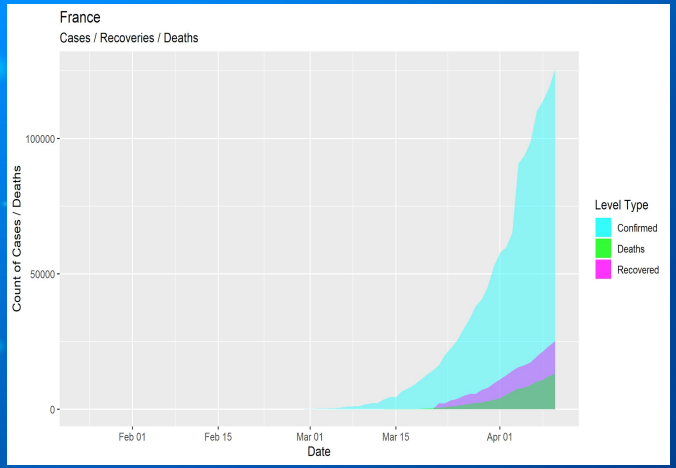
USA



Spain

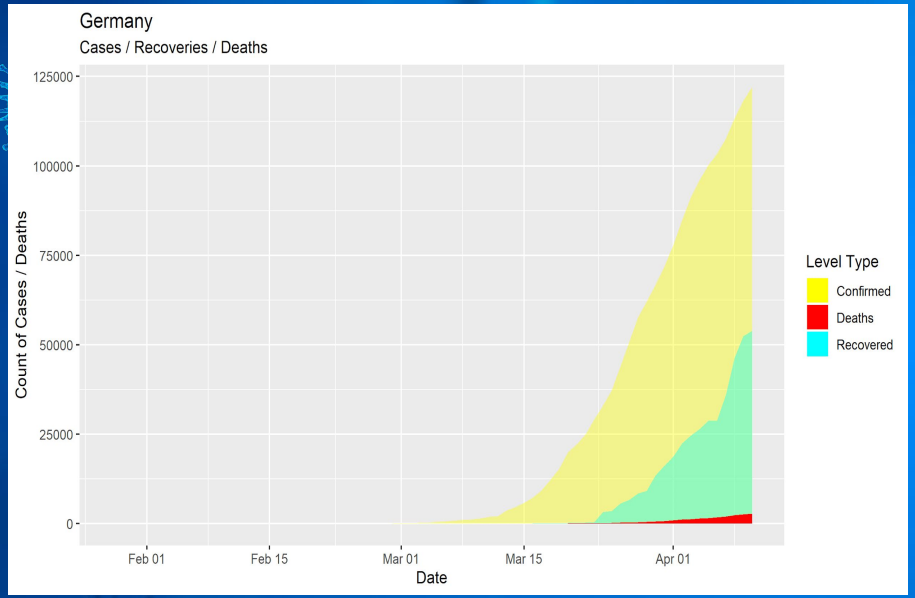


Italy



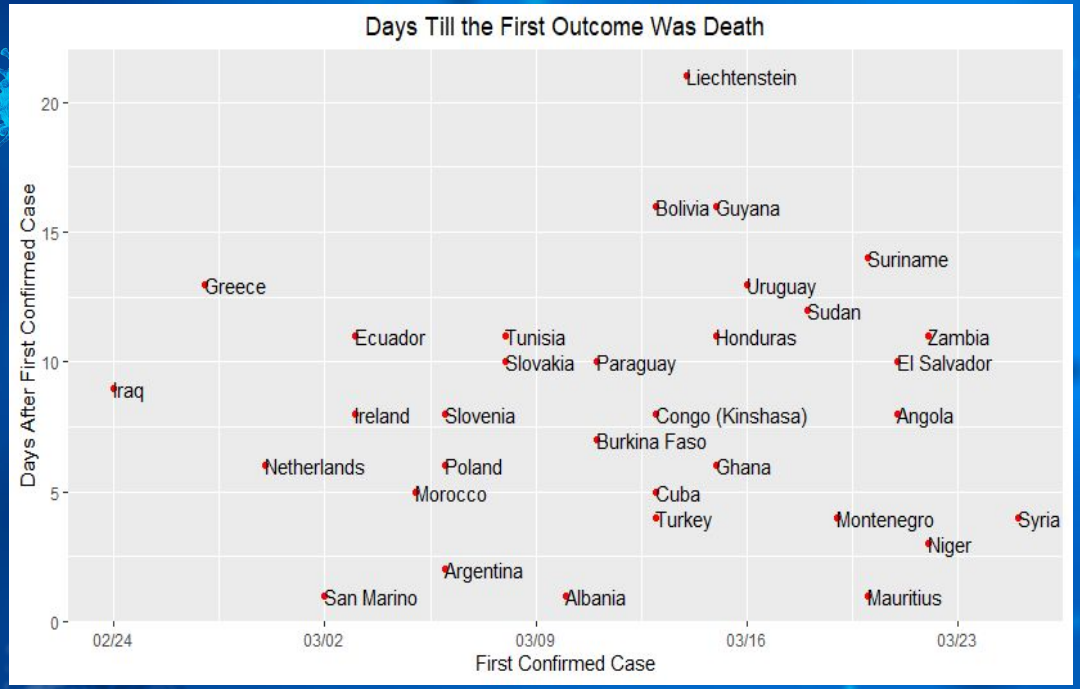
France

Country-wise Analysis of Cases, Recoveries and Deaths



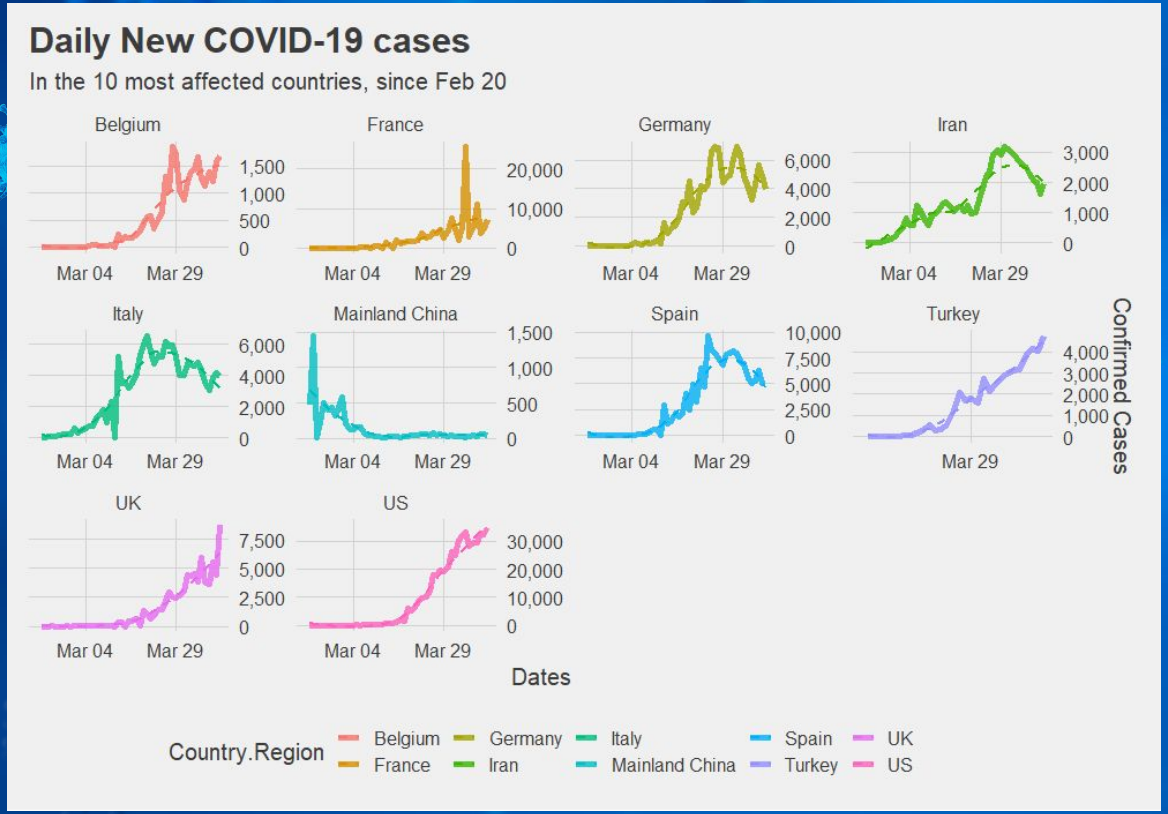
- US and Italy didn't enforce travel restrictions fast enough; open EU borders spread it to other European countries quickly
- Germany response fast – high recovered to confirmed ratio
- US response still slow – low recovered to confirmed ratio
  - Might reach full capacity of medical services if daily cases don't drop

First death across countries



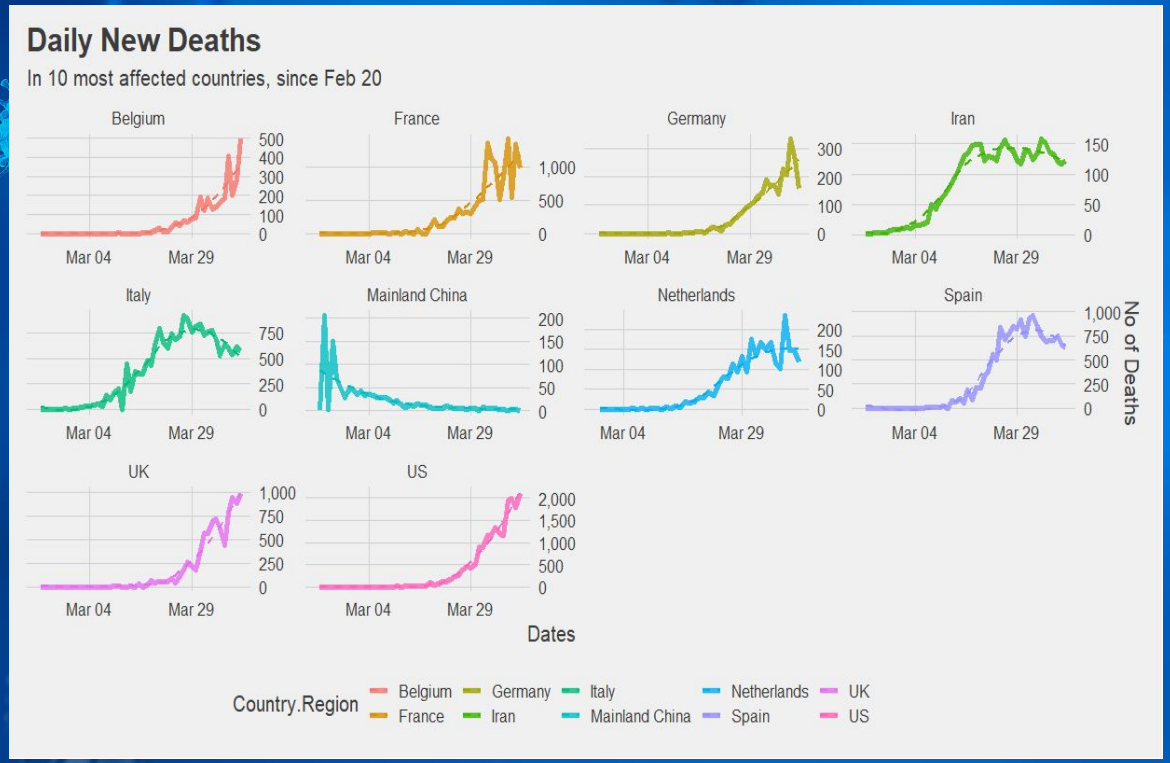
- Graph shows the most extreme situations – days between first confirmed case and first death
- Many third world countries – inadequacy of testing or isolation
- Surprising participants – European countries like Ireland, Netherlands

Comparing containment of countries (CASES)



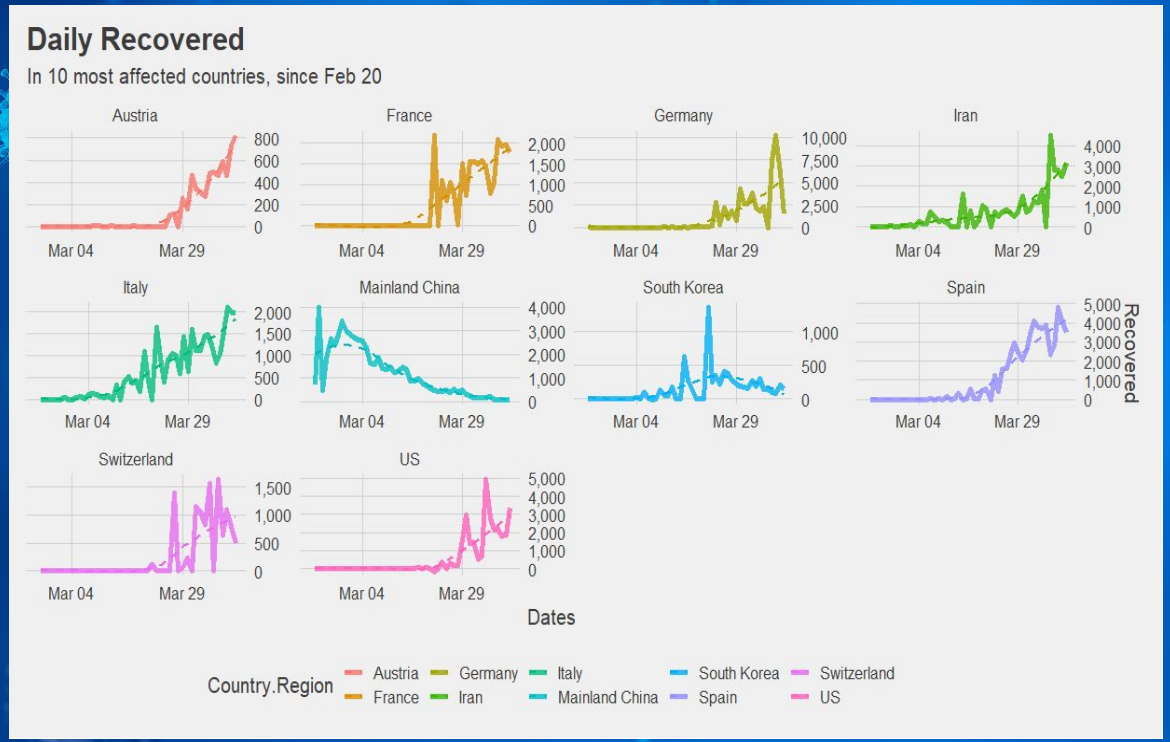
- US, UK, Turkey still on the exponential path
- Italy, Spain, Germany, Iran contained effectively
- Similarity of graphs without the time axis
- Clear indication of similar outcome for similar measures

Comparing containment of countries (DEATHS)



- Resemblance of deaths graph with confirmed graph even inside countries
- Inclusion of Netherlands
- France still experiencing deaths
- Need of cure clear

Comparing containment of countries (RECOVERED)

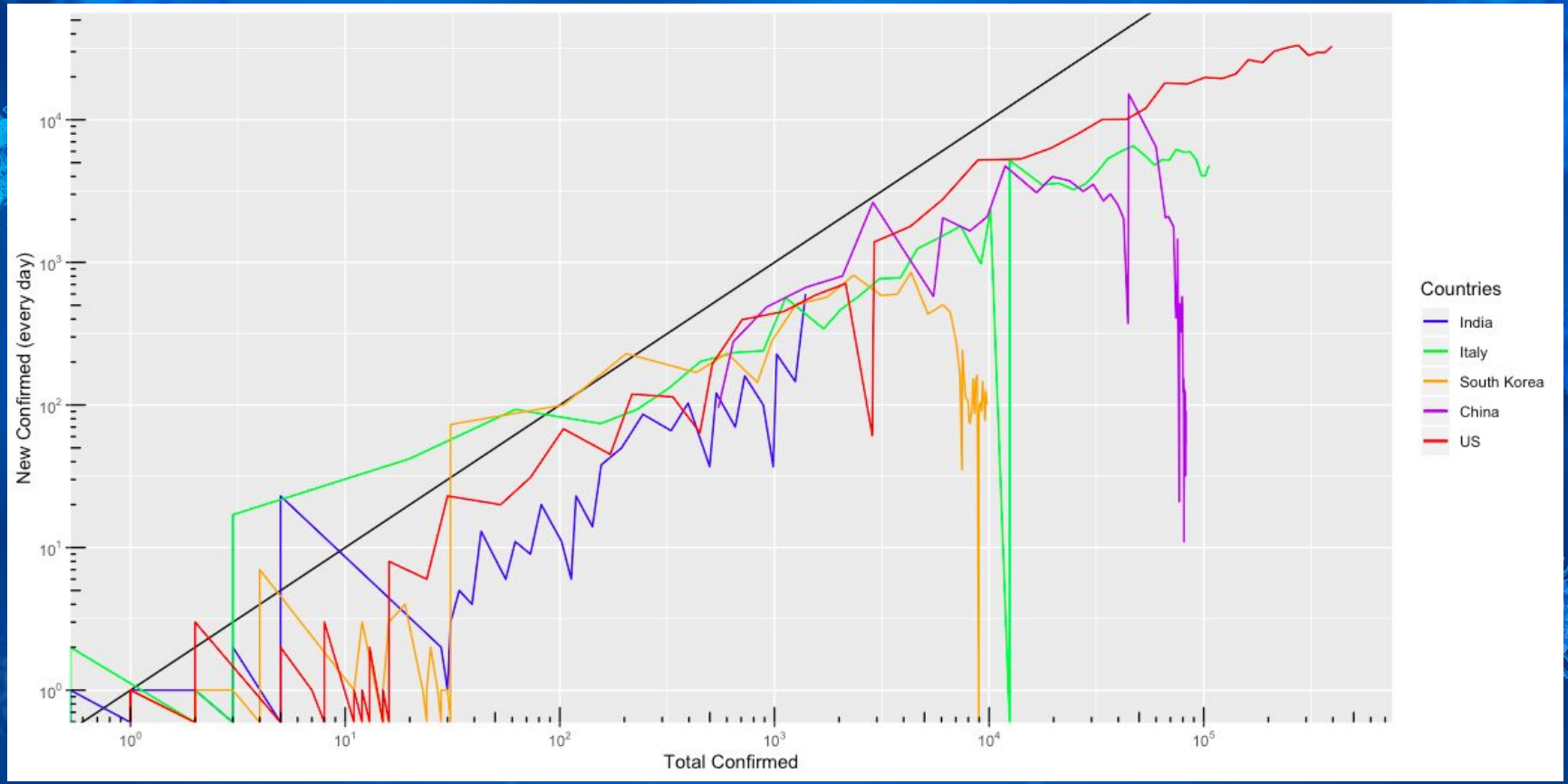


- New inclusions – South Korea, Switzerland, Austria
- South Korea success story; very quick response by testing and containment

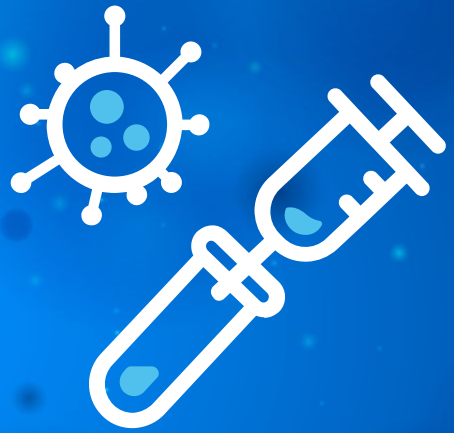
Removing time from the axes

- Intuition: pandemic spread not time-dependent, but spread dependent
- Change-based graph as opposed to total numbers to clearly see whether we still have more cases per day every day
- Logarithmic scale to see this change clearly
- Surprisingly clear analyses on where a community stands on its control of the pandemic spread

Total confirmed cases vs Daily new cases



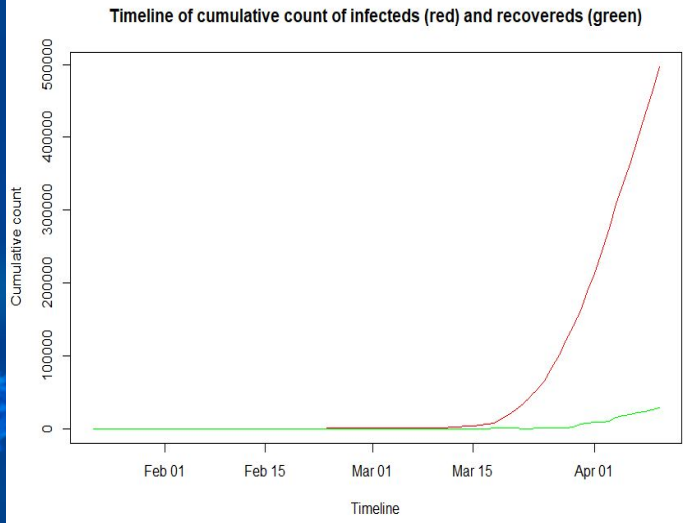
Modeling

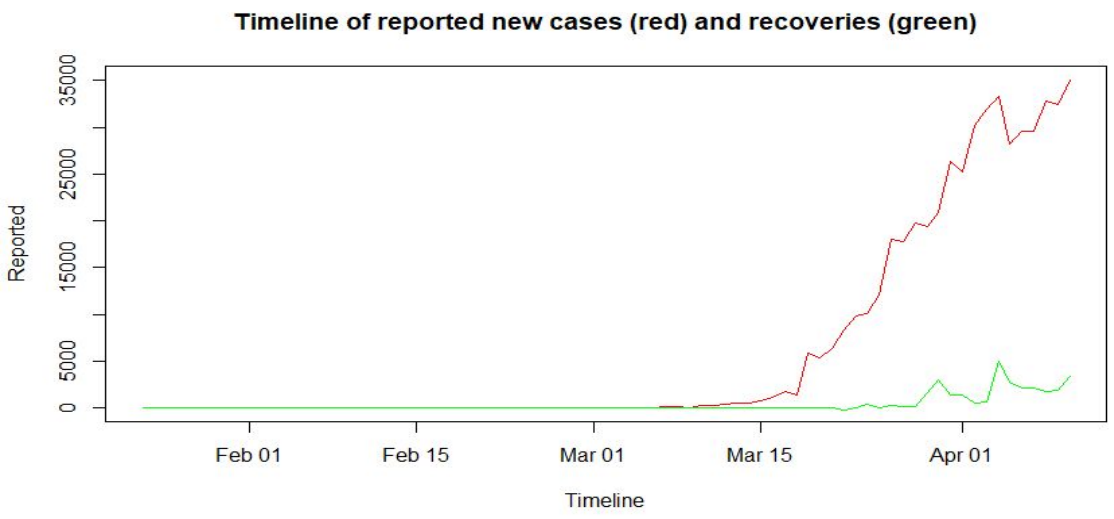


First model: Epidemiology modeling with SIR

- US currently has most affected cases, but highly localised spread
- Closer look required – state-wise epidemiology analysis
- SIR modeling
  - S: susceptible
  - I: infected
  - R: recovered (dead or recovered, can’t affect spread anymore)
- Training data till March 31st, prediction till August.

Data Analysis for USA



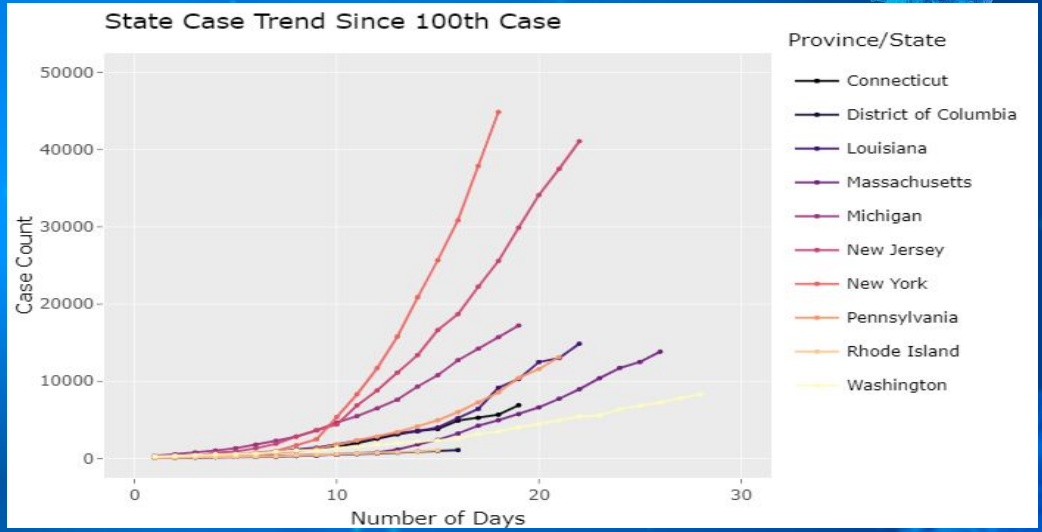


- Total cases crossed 500,000 couple of days ago
  - Almost impossible to see any pattern in the exponential graph
- Daily increase in cases as high as 35,000 (and counting!)

State-wise Analysis

Pre-processing the dataset to get state-wise data

	States	Cases	Cases Per Cap	Deaths	Deaths per Cap	Mortality
1	New York	139875	706.59	5489	27.73	3.924%
2	New Jersey	44416	495.82	1232	13.75	2.774%
3	Louisiana	16284	348.64	582	12.46	3.574%
4	Massachusetts	15202	223.74	356	5.24	2.342%
5	Connecticut	7781	216.69	277	7.71	3.560%
6	Michigan	18970	191.18	845	8.52	4.454%
7	District of Columbia	1211	180.15	22	3.27	1.817%
8	Washington	8692	121.22	400	5.58	4.602%
9	Rhode Island	1229	116.35	30	2.84	2.441%
10	Pennsylvania	14853	116.02	247	1.93	1.663%



Total population : 2016 US census data to calculate the Cases per capita and deaths per capita in each region.

SIR Modeling

Fit an SIR model on

```
Call:
lm(formula = y ~ x + 0)

Residuals:
   Min     1Q Median     3Q    Max
-28758   1320   4850   8283  10409

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x -0.190534   0.005307   -35.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7567 on 50 degrees of freedom
Multiple R-squared:  0.9627,    Adjusted R-squared:  0.9619
F-statistic:  1289 on 1 and 50 DF,  p-value: < 2.2e-16
```

1st model on the infected growth data :
dS/dt=-b SI/N ; N- total population
Low p-value and high R squared value suggests relation between growth and time.

```
Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-4971.2   145.5   826.1  1705.1  3988.8

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x 0.018606   0.000621   29.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1693 on 50 degrees of freedom
Multiple R-squared:  0.9472,    Adjusted R-squared:  0.9462
F-statistic: 897.6 on 1 and 50 DF,  p-value: < 2.2e-16
```

2nd model on the Recovery data :
dR/dt= cI;
Low p-value and high R squared value suggests relation between recovery and time.

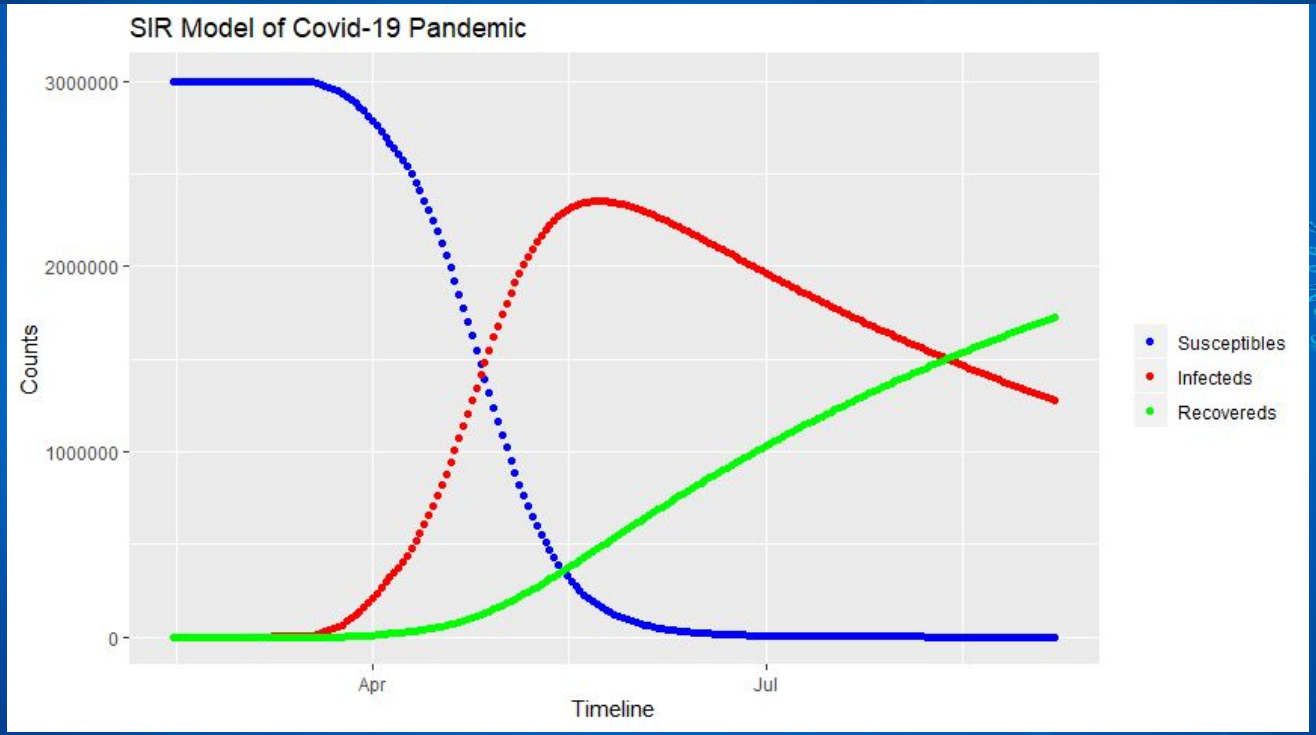
Predicted vs Actual

Predicting infected and recovered count in the future and comparing with actual data

	Dates	Predicted.Infecteds	Predicted.Recovereds	Actual.Infecteds	Actual.Recovereds
1	2020-04-07	1115373	296397.2	1126042	300054
2	2020-04-08	1157125	317149.9	1182443	328661
3	2020-04-09	1193549	338679.4	1241375	353975

SIR PLOT

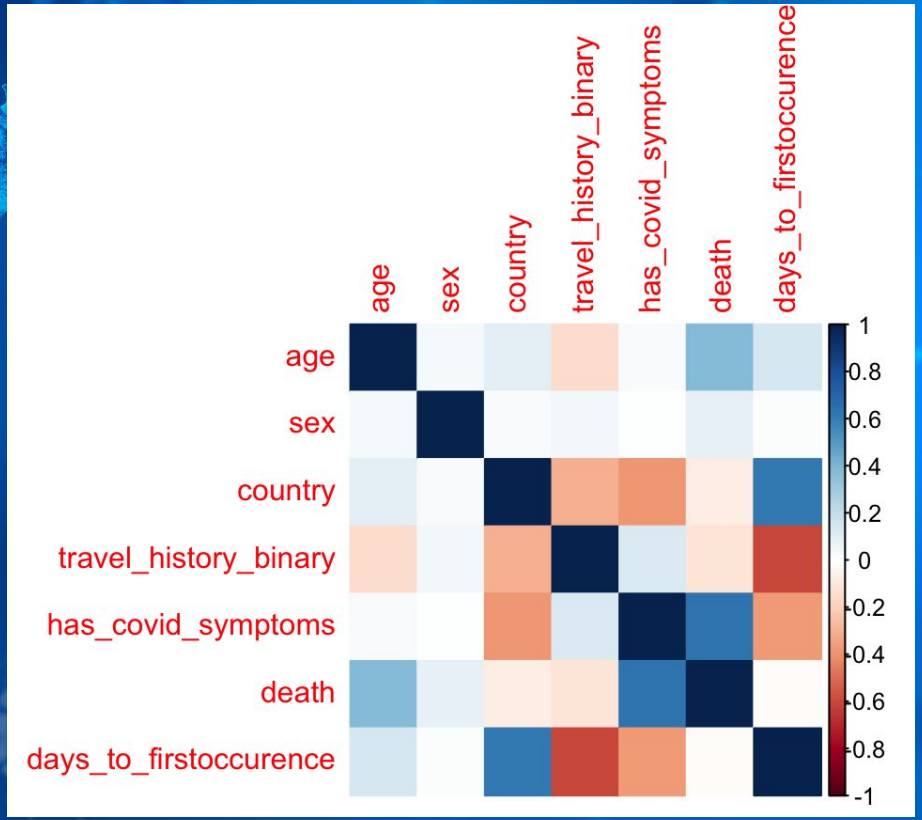
Plotting future numbers till August



Second model: Death prediction of Patient

- Prediction of a given patient's outcome based on features such as age, gender, travel history etc.
- Intuition – not for direct fatality prediction, but for assessing priority in giving medical help.
- Models tested
  - Logistic Regression
  - SVM
  - Random Forest
  - SVM One-Class-Classification

Let's find out the relations! - Correlation Matrix

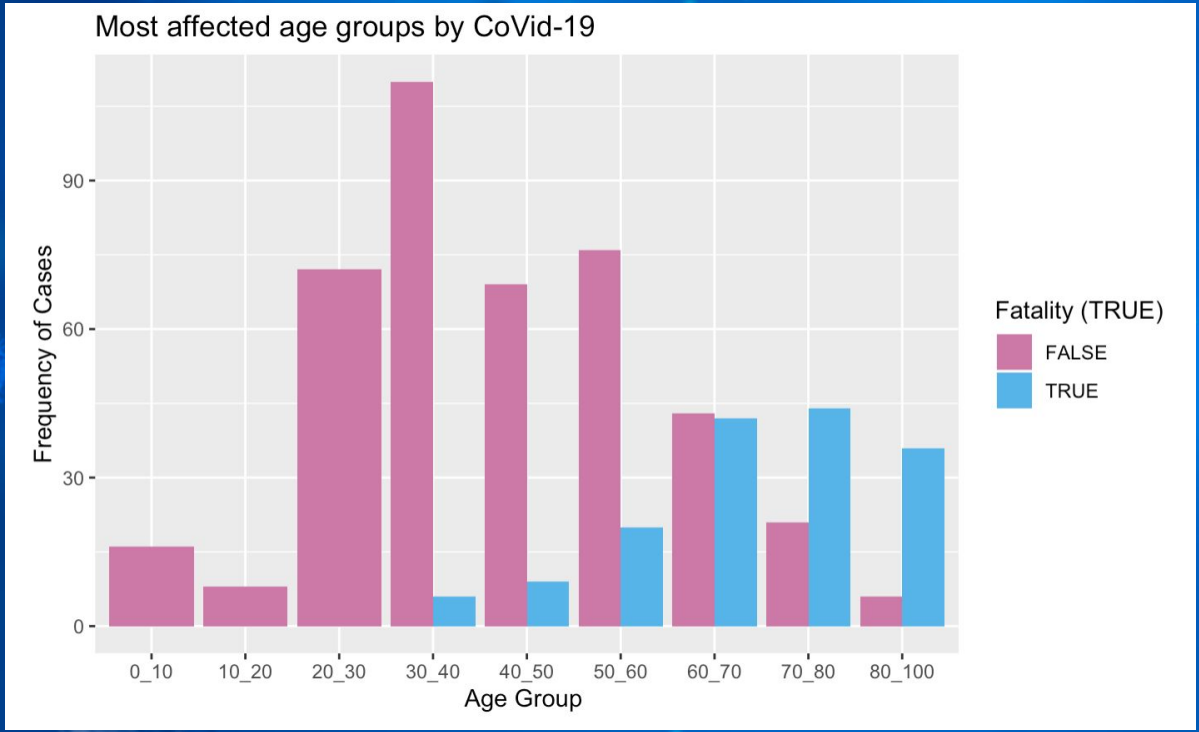


Target Variable : death

- Two most important features are:
  - has_covid_symptoms
  - age

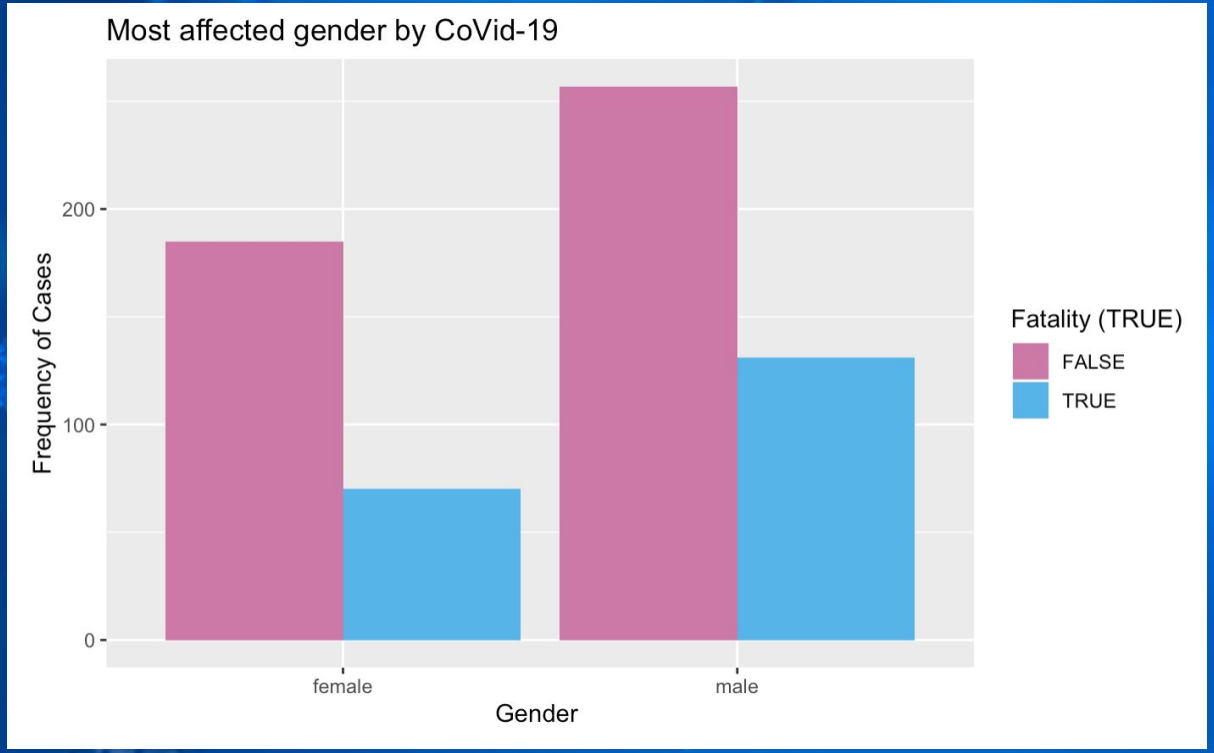
Package used : hetcor

How is age related to fatality?



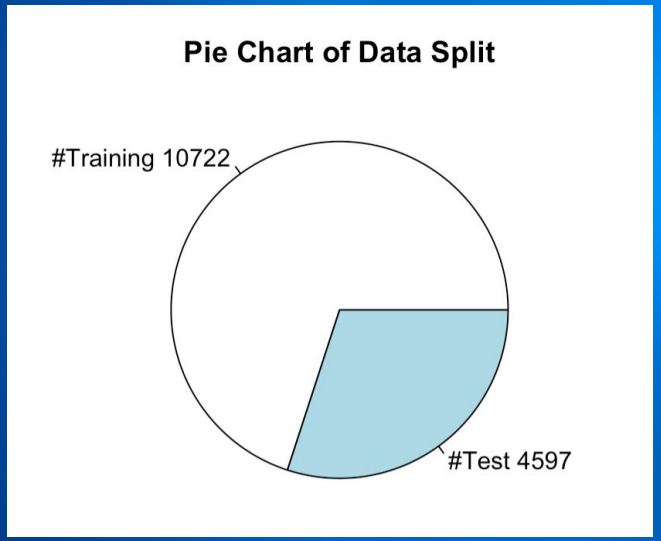
- Clear relation of fatality with age
- Side takeaway – take special care of elders
  - But young can still spread, so social distancing crucial

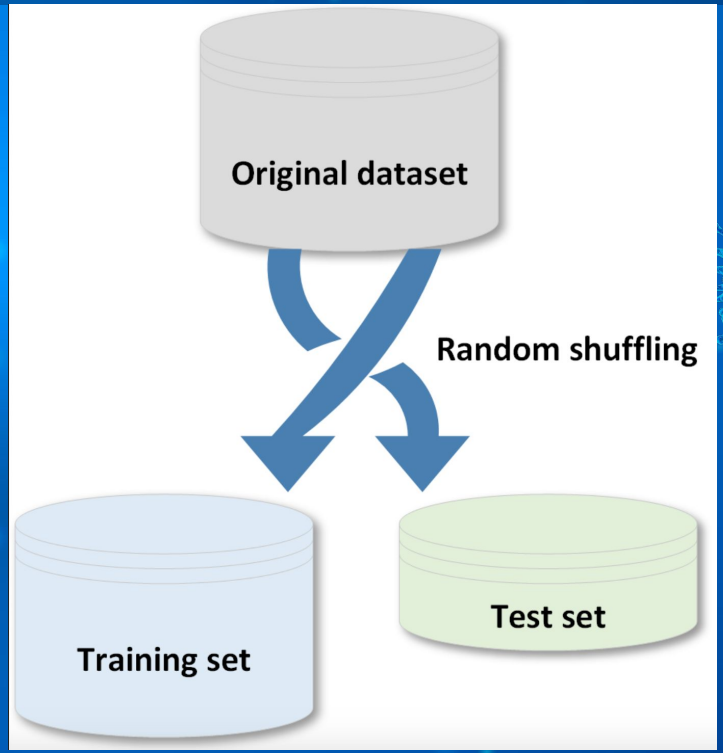
Is there a relation between Fatality and Gender?



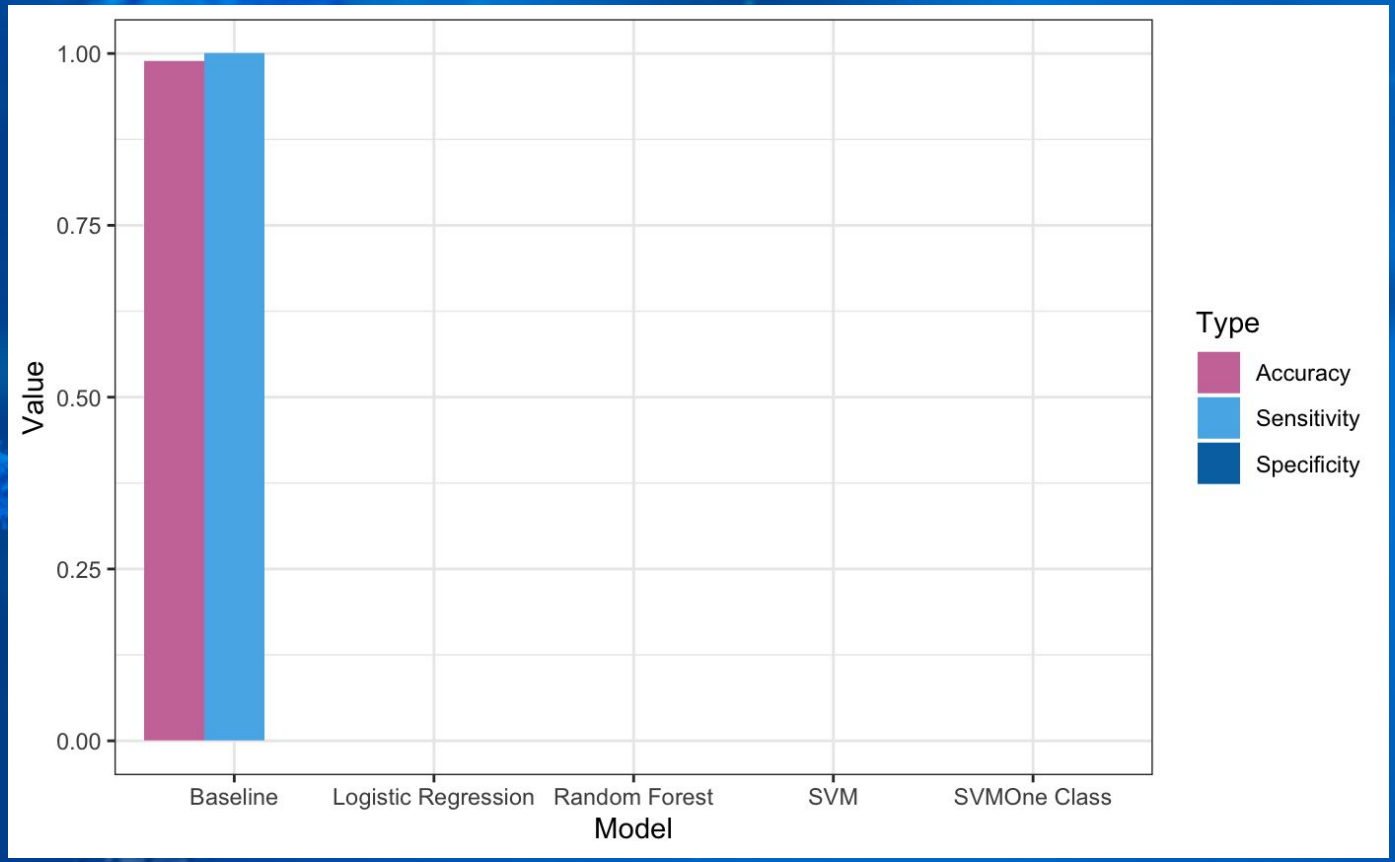
- Weak relation with gender
- Partly due to unbalanced dataset
- Relation too weak for EDA, but may be useful for modeling

Data Split into Training & Test



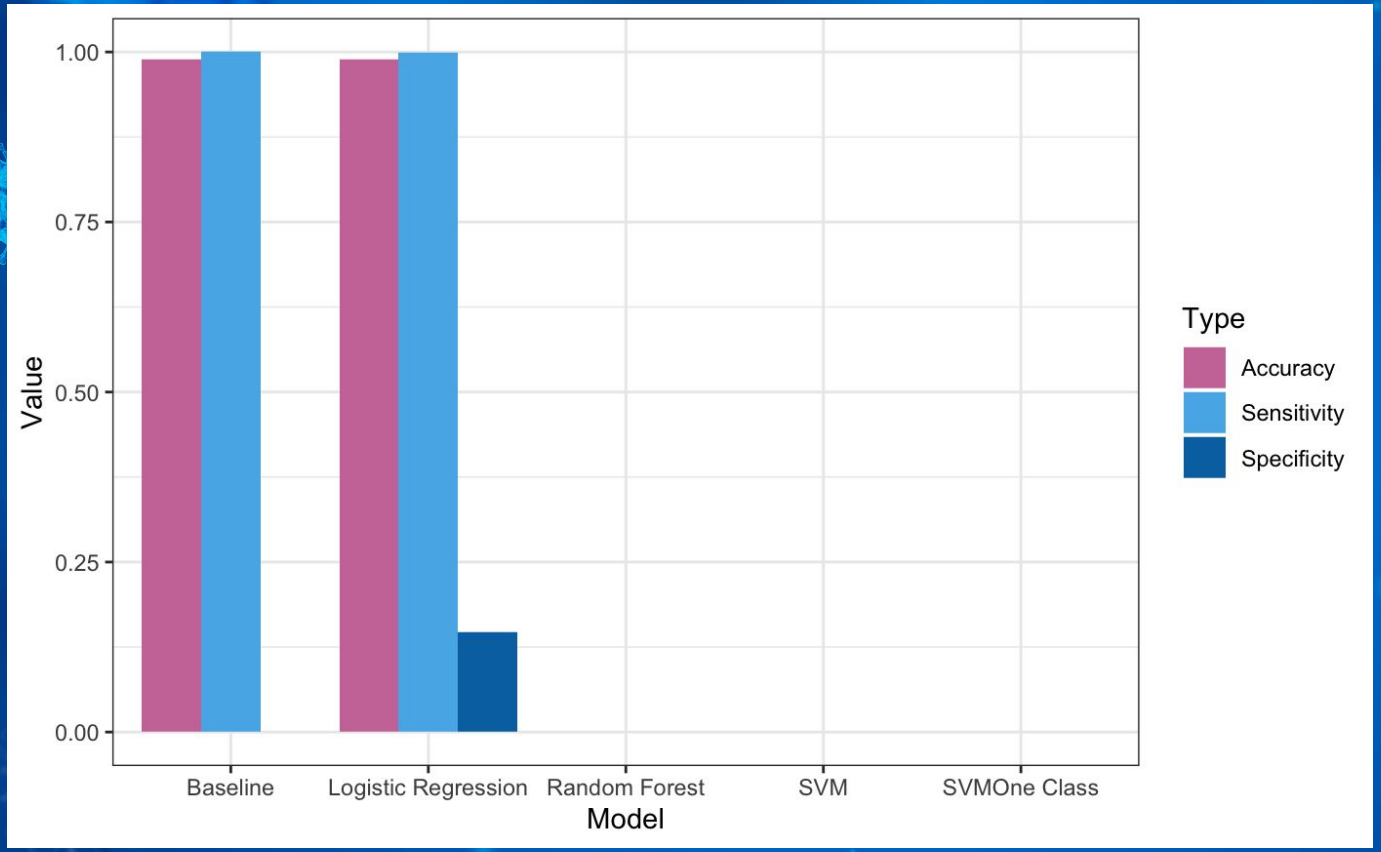


Models : Baseline (Majority Class Prediction)



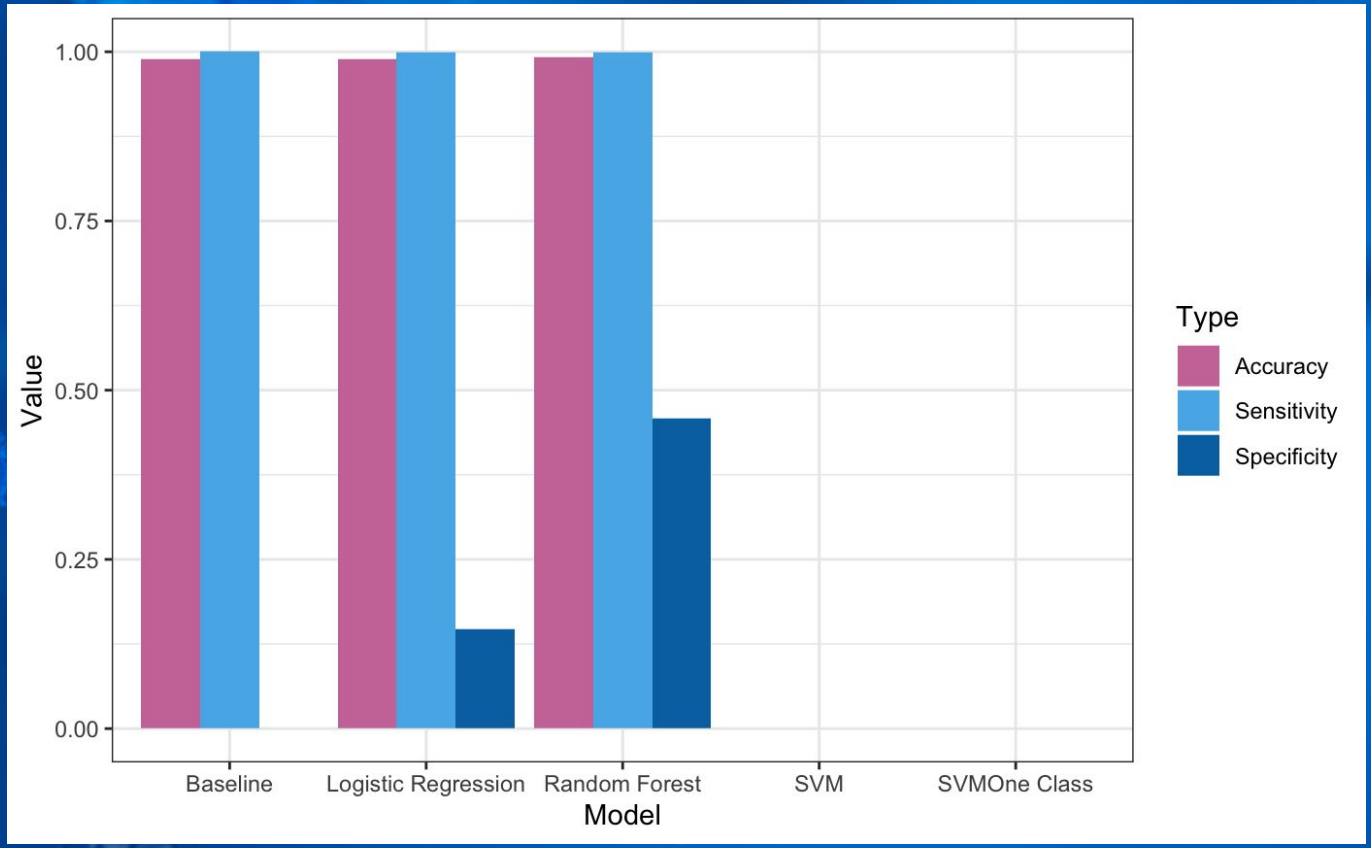
- Predicts false for the whole dataset.

Models : Logistic Regression



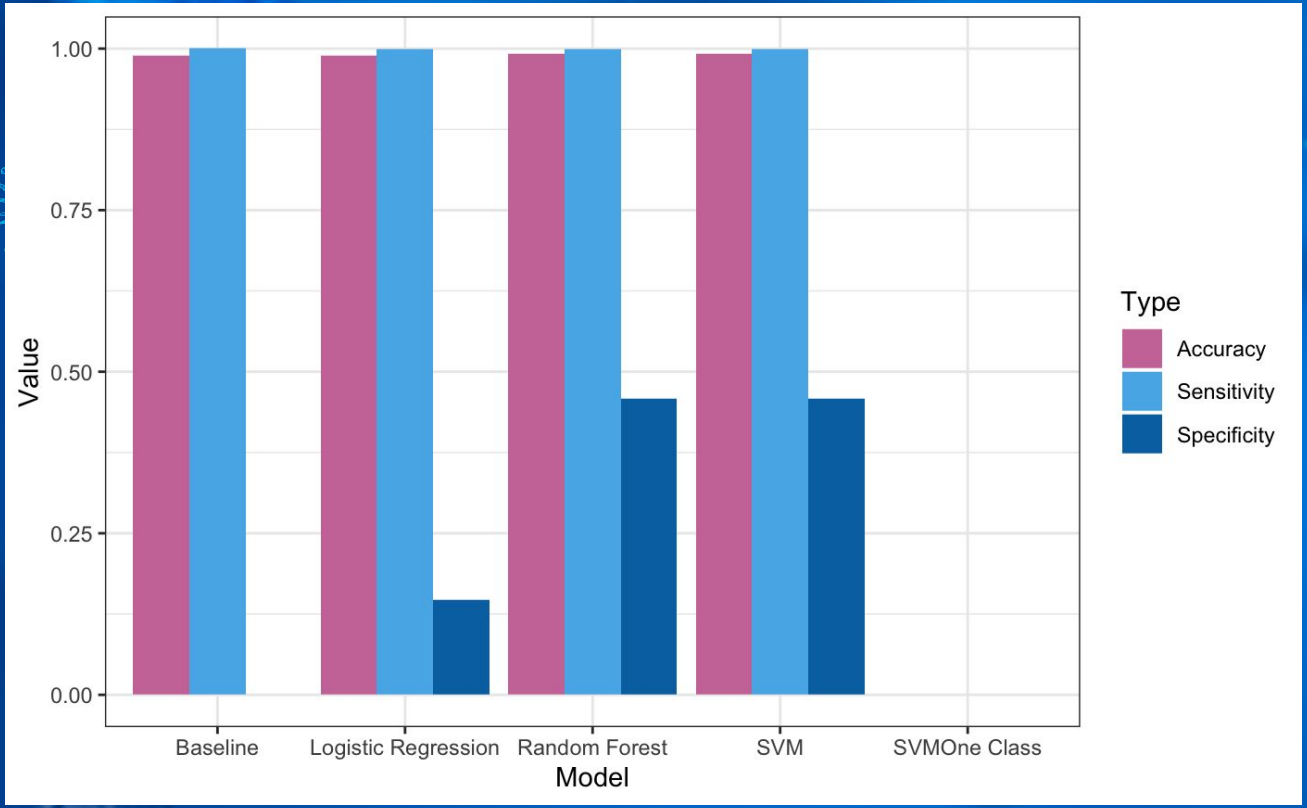
- Accuracy and Sensitivity remains same but specificity increases.

Models : Random Forest



- Observes an increase in specificity in Random Forest.

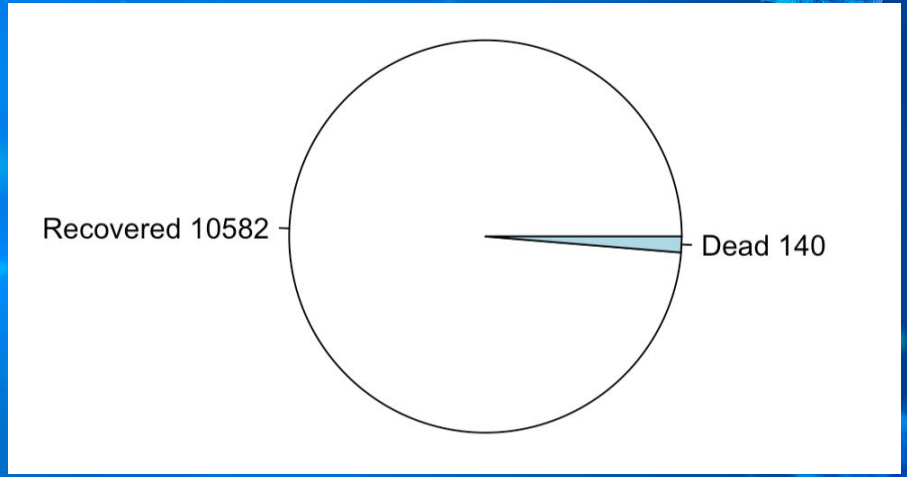
Models : SVM



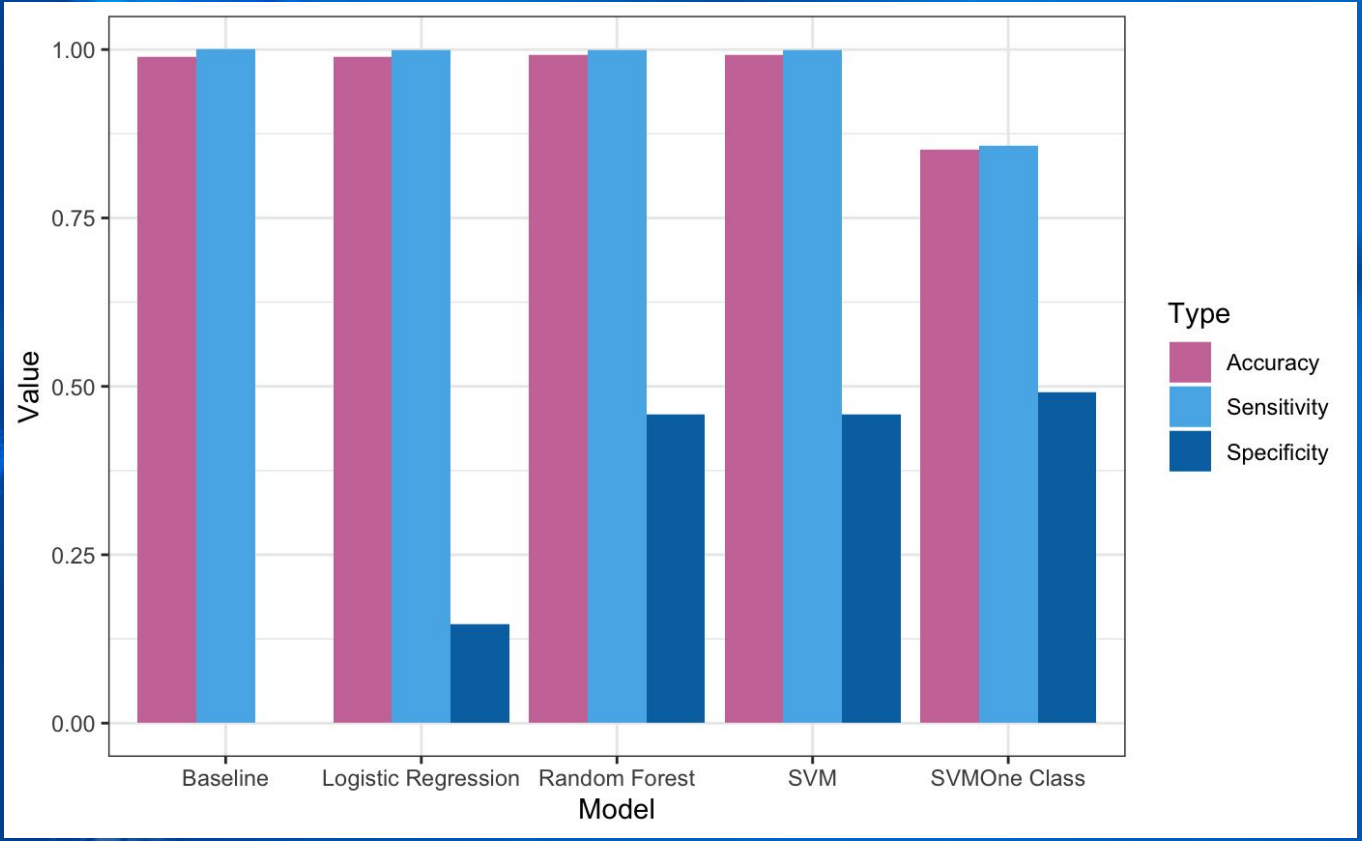
- Performance is almost same as RF.
- And we find SVM to be minutely better than Random Forest.

Data imbalance problem

- Highly imbalanced data to predict Death of COVID-19 patients
- SVM One-Class used for novelty detection (learning rare events)
- Goal: Learn rare death events and use that information while predicting



Models : SVM- One Class Classification



- While specificity increases, accuracy and sensitivity decreases hugely.

Conclusion

- On predicting a particular patient's outcome, SVM & Random Forest performed the best overall.
- We can use this model to help in assessing priority of patients in giving medical help.
- Countries are on different stages on containing the virus – but the dread of exponentiality has hit every country
- From epidemiology model, we found that USA is yet to see the peak, likely in the first week of May and then there is an expected decline in the infected cases by the end of May.

References

1. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
2. https://www.worldometers.info/coronavirus/
3. https://github.com/CSSEGISandData/COVID-19

Thank You!

Stay Safe!