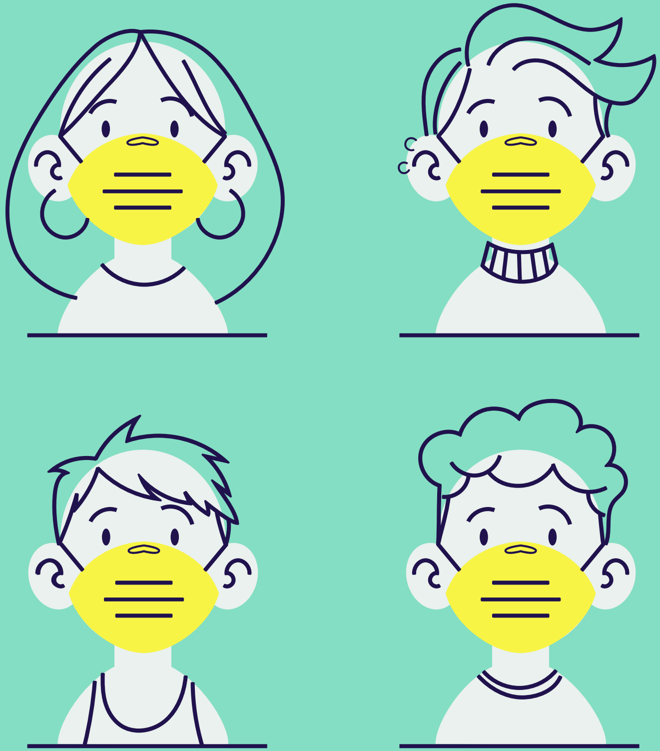
COVID-19 CHINA'S PROJECTIONS



By: Peng Zhang, Brenda Mas & Mohu Sah

AGENDA

01

Problem

02

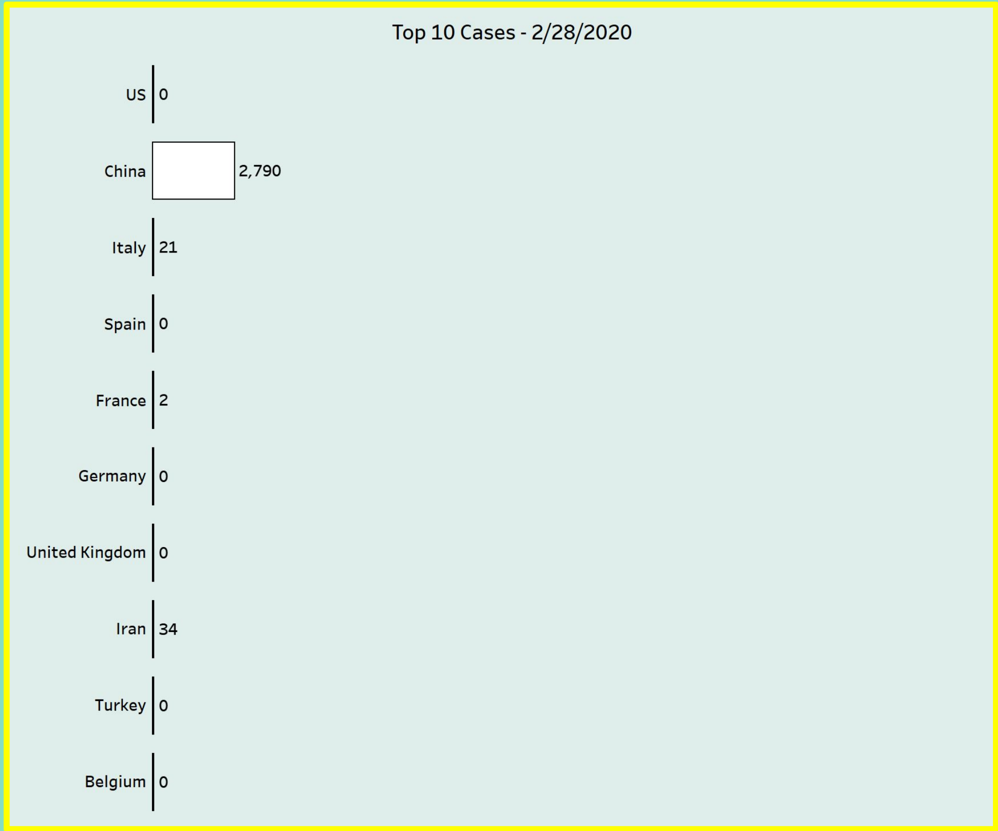
Data PreProcessing

03

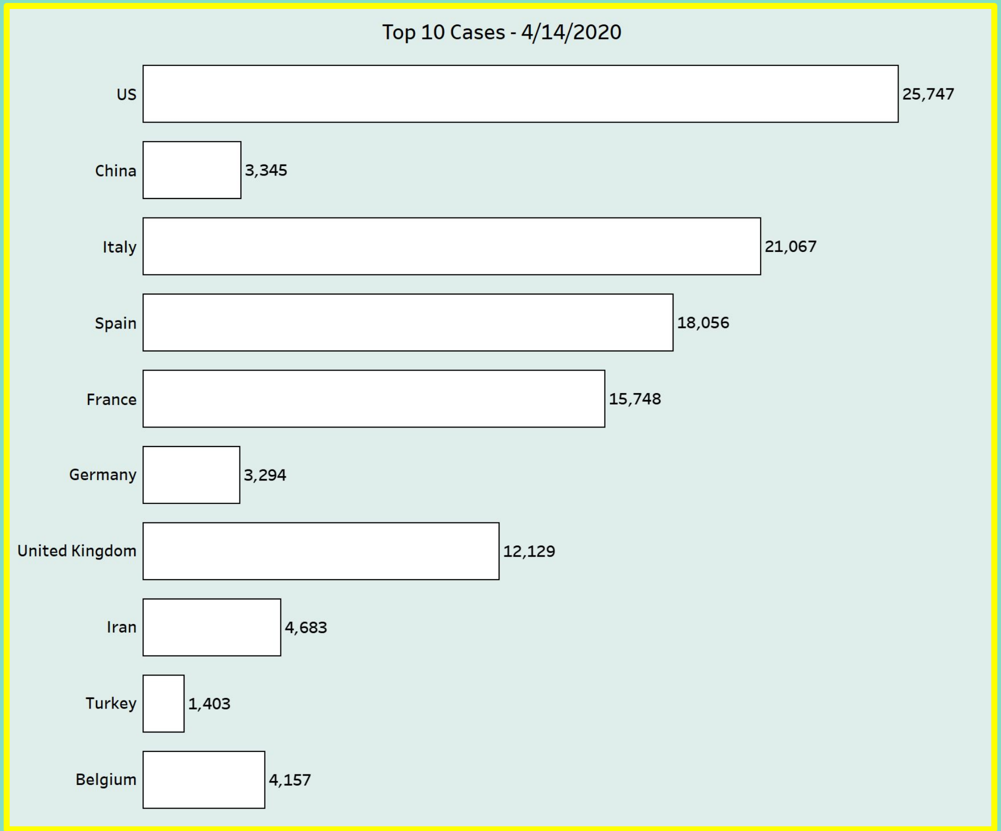
Model Choices

04

Conclusion



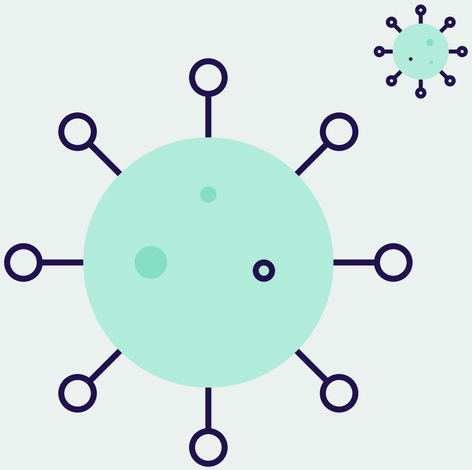
WORLD CASES ON FEB 28



WORLD CASES ON MAR 28

QUESTION

Is China Under Reporting
Their Cases?

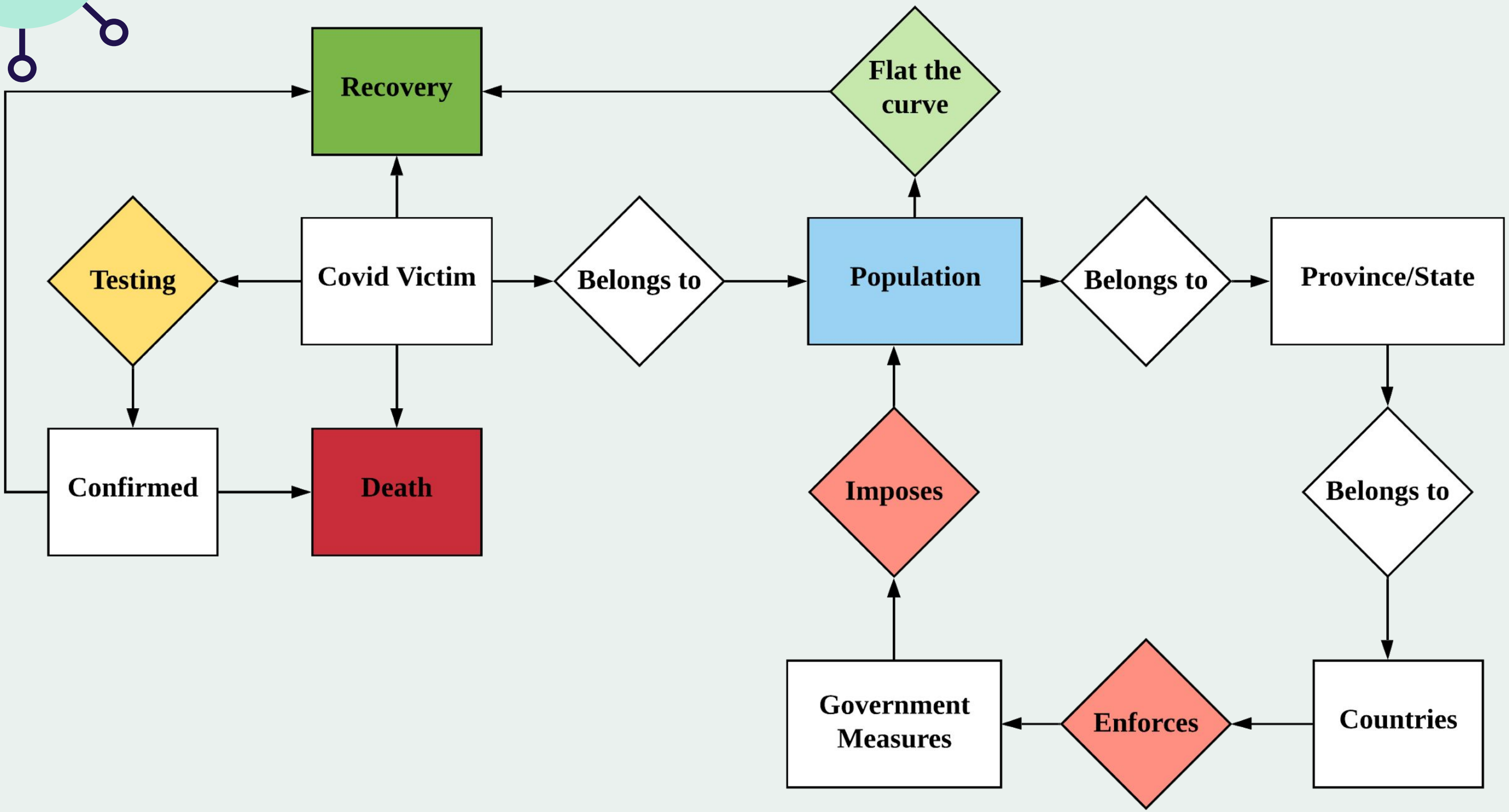


LITERATURE REVIEW

Are China's COVID Statistics Reliable?

- Autocratic & Dictatorship Government in the past have inflated their statistics - General line of reasoning
- Governance system rewards positive news
- Efforts to downplay the impact of novel Coronavirus
- China's COVID numbers are likely much higher as previously stated
- A drop in mobile phone & landline usage witnessed in China during the time of quarantine. One would rather expect an increase in mobile usage.
- Mortality Rate in Italy - 9% , suggests numbers to be misrepresented

ENTITY RELATIONSHIP DATABASE



FEATURE SELECTION

Entity

Stringent Index

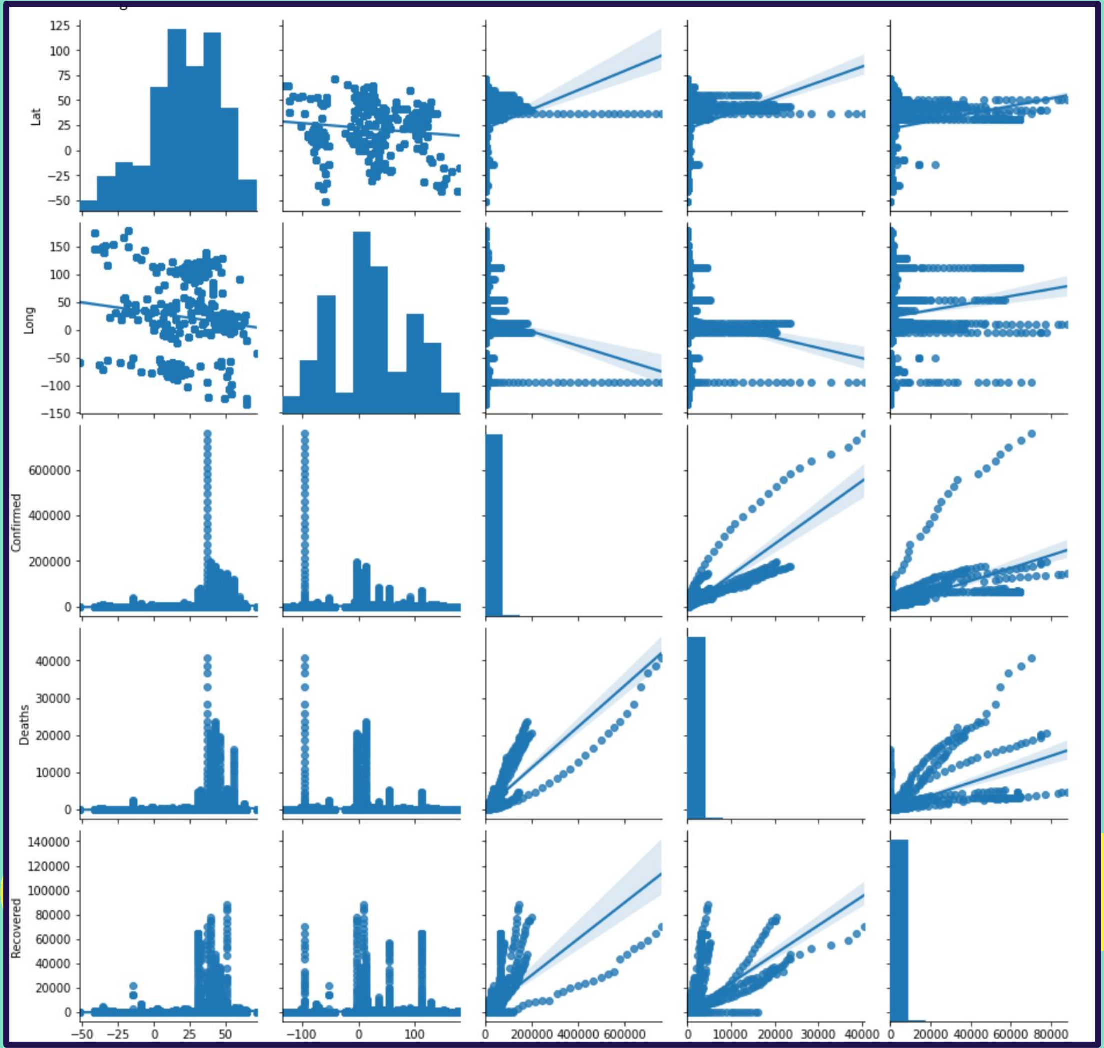
Density

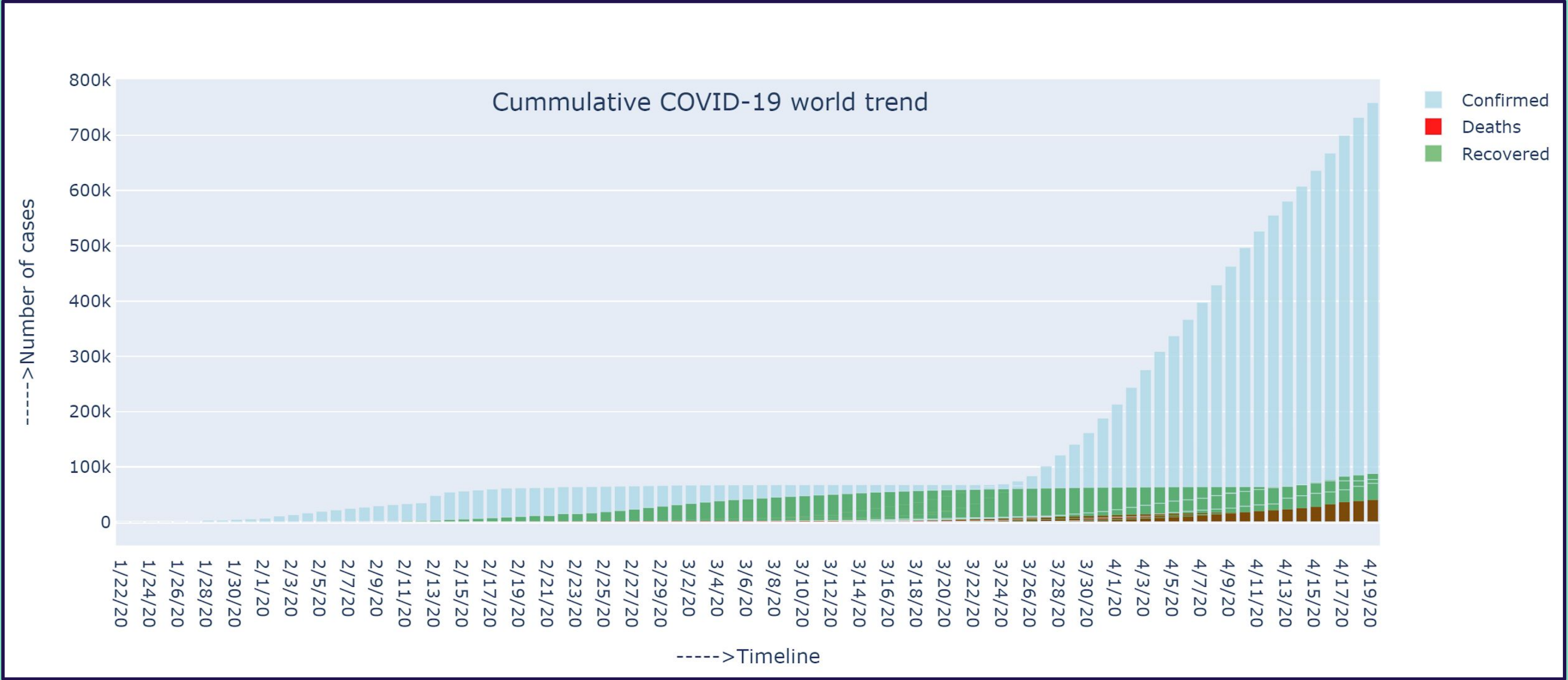
Age Group

02.

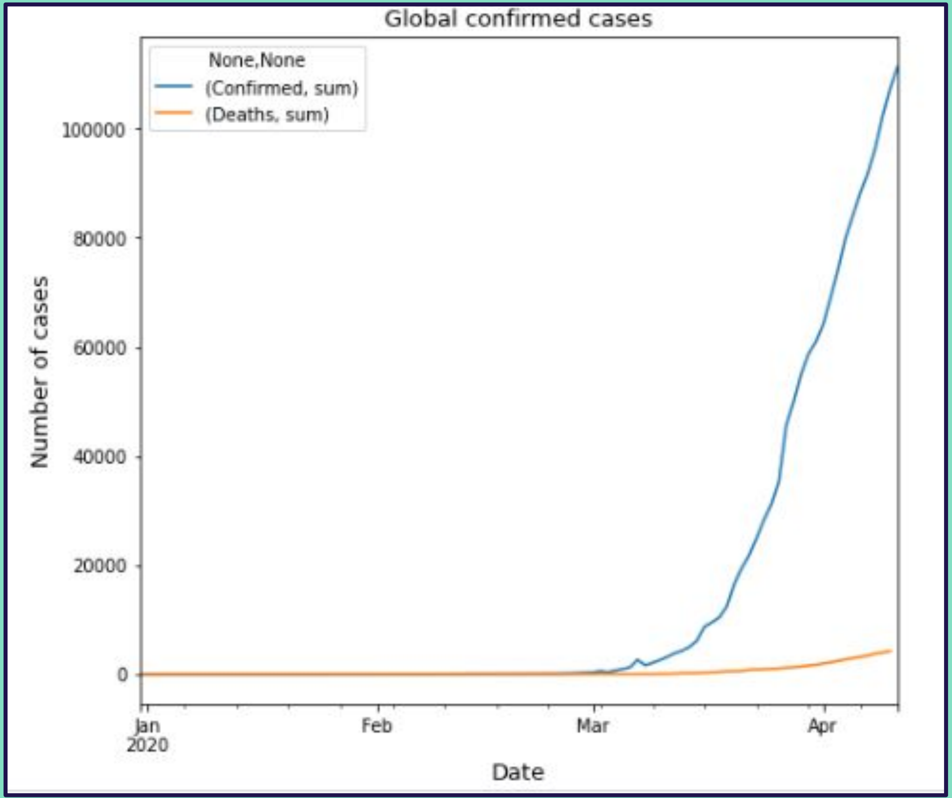
DATA PRE-PROCESSING

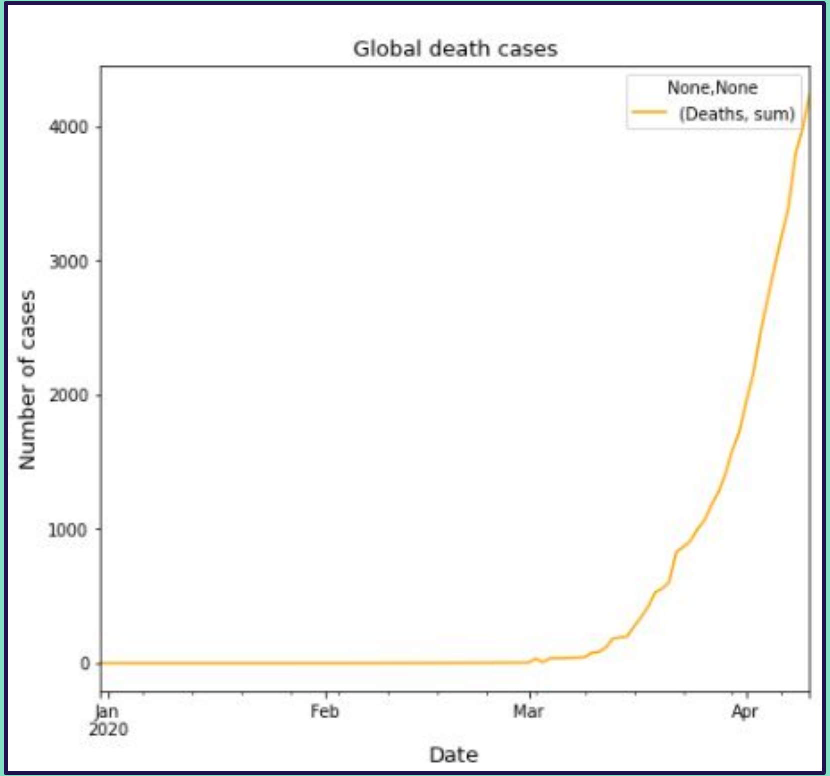
PAIRPLOTS





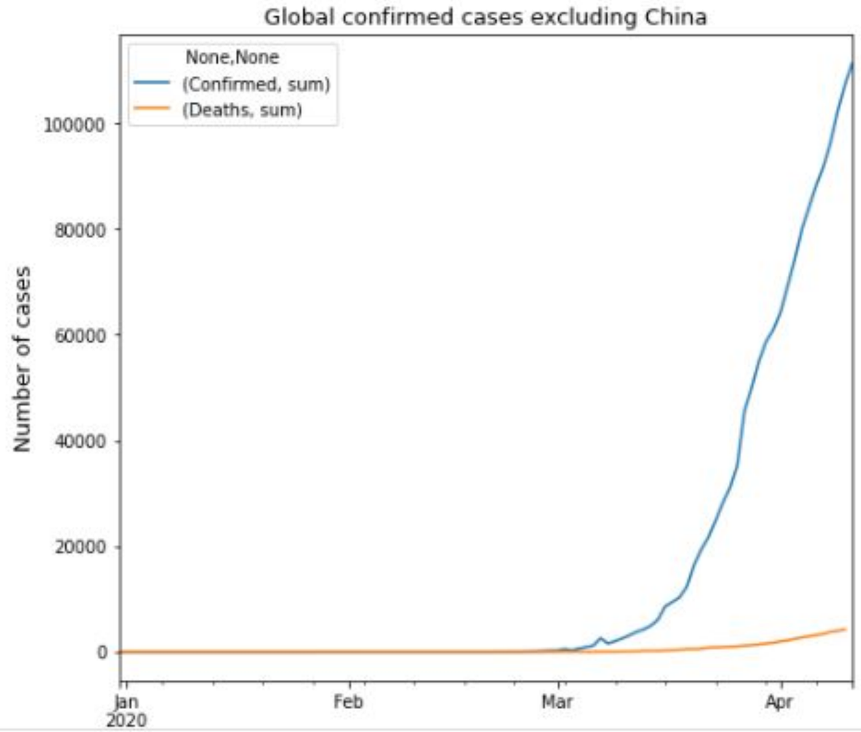
INCREMENTAL COVID-19 TREND

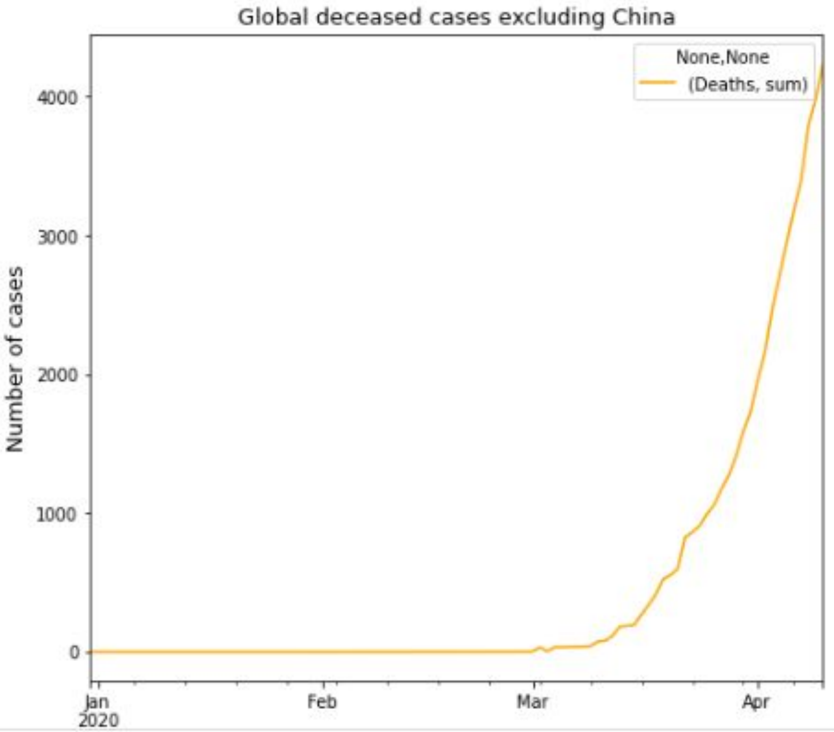




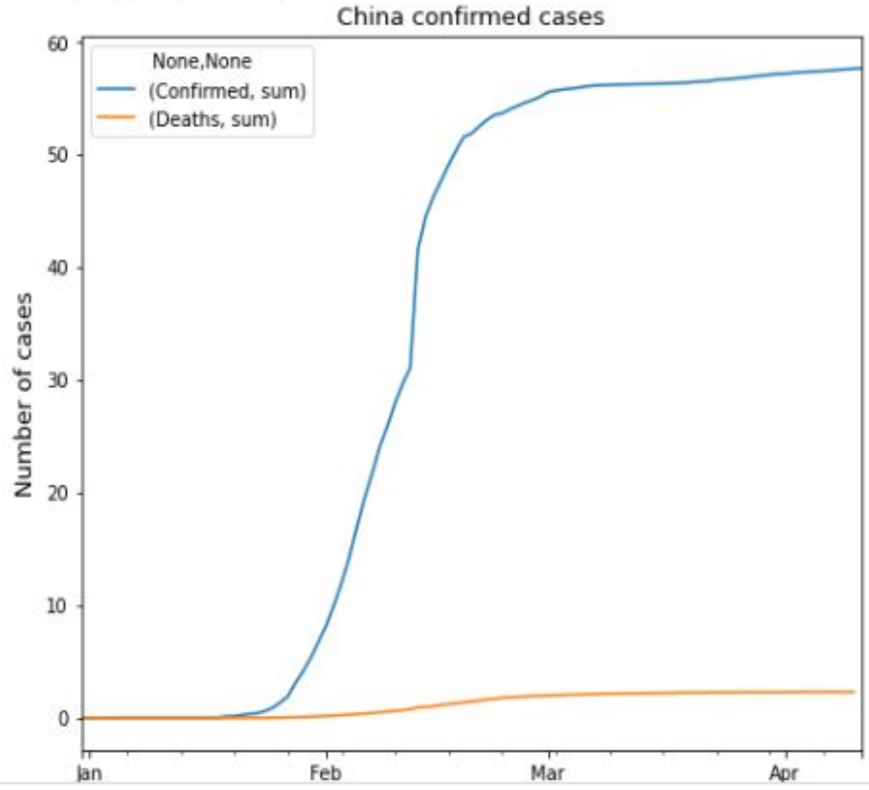
GLOBAL TREND

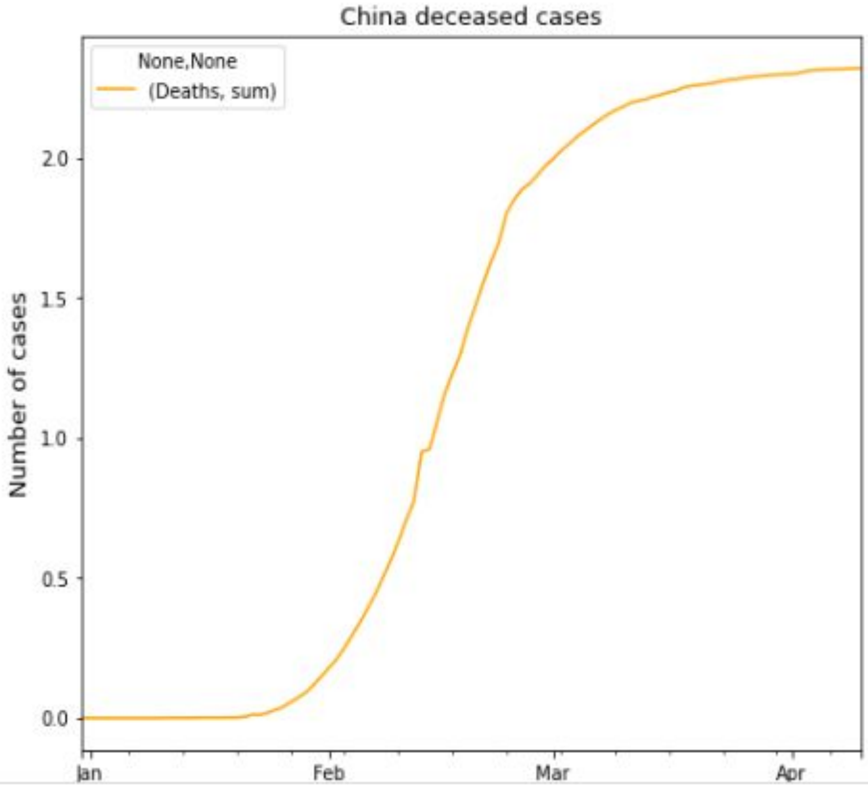
TRENDS OF THE WORLD WITHOUT CHINA





COVID 19 SITUATION IN CHINA





METHODOLOGY

- Clean & Merge datasets
- Population: Median age, population density & urban population
- Stringency (Government imposed measures):

  Lockdown, investment in healthcare, tracing, International support

POPULATION & STRINGENCY DATA

Country (or dependency)	med_age	urban_pop	density	land_area	world_share
China	38	61 %	153	9388211	18.47 %
India	28	35 %	464	2973190	17.70 %
United States	38	83 %	36	9147420	4.25 %
Indonesia	30	56 %	151	1811570	3.51 %
Pakistan	23	35 %	287	770880	2.83 %

H3_Contact tracing	E4_International support	investment in healthcare	H5_Investment in vaccines	H2_Testing policy	H3_Contact tracing.1	E1_Income support	StringencyIndex
1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0

STANDARDIZE DATA OVER WEEKS

- Standardized (Population & Time): Daily new cases per week, per million
- To compare apples-to-apples
- Every country's Day 1: When infection started

Entity	Week NUM	Confirmed	Daily	Deaths	med_age	density	H3_Contact tracing
Afghanistan	**1**	0.025688	0.026	0.000000	18.0	60.0	1.0
	2	0.565141	0.539	0.000000	18.0	60.0	1.0
	3	1.926617	1.361	0.025688	18.0	60.0	1.0
	4	4.932139	3.006	0.102753	18.0	60.0	1.0
	5	10.866119	5.935	0.359635	18.0	60.0	1.0

AGGREGATE BY COUNTRY

- Aggregate data per country
- One row per country

	Confirmed	Daily	Deaths	med_age	density	H3_Contact tracing	E4_International support
Entity							
Afghanistan	13.383565	2.230667	0.385323	18.0	60.0	1.0	0.0
Africa	9.661359	1.073667	0.469939	0.0	0.0	NaN	NaN
Albania	144.554868	28.910400	7.992216	36.0	105.0	1.0	0.0
Algeria	40.158680	6.693500	5.359052	29.0	18.0	0.0	0.0
Andorra	7778.424901	1555.684400	323.561768	0.0	164.0	1.0	0.0

DATA CLEANING

- Remove null values
- Remove arbitrary country names like Africa, World, Oceania
- Date ready for feature modelling

Entity	Confirmed	Daily	Deaths	med_age	density	H3_Contact tracing	E4_International support
Afghanistan	13.383565	2.230667	0.385323	18.0	60.0	1.0	0.0
Albania	144.554868	28.910400	7.992216	36.0	105.0	1.0	0.0
Algeria	40.158680	6.693500	5.359052	29.0	18.0	0.0	0.0
Andorra	7778.424901	1555.684400	323.561768	0.0	164.0	1.0	0.0
Angola	0.578100	0.192333	0.060853	17.0	26.0	0.0	0.0

03.

MODEL CHOICES

CHOICES

01. CLUSTERING

02. SMOTE

Deal with Imbalanced Data

03. REGRESSION

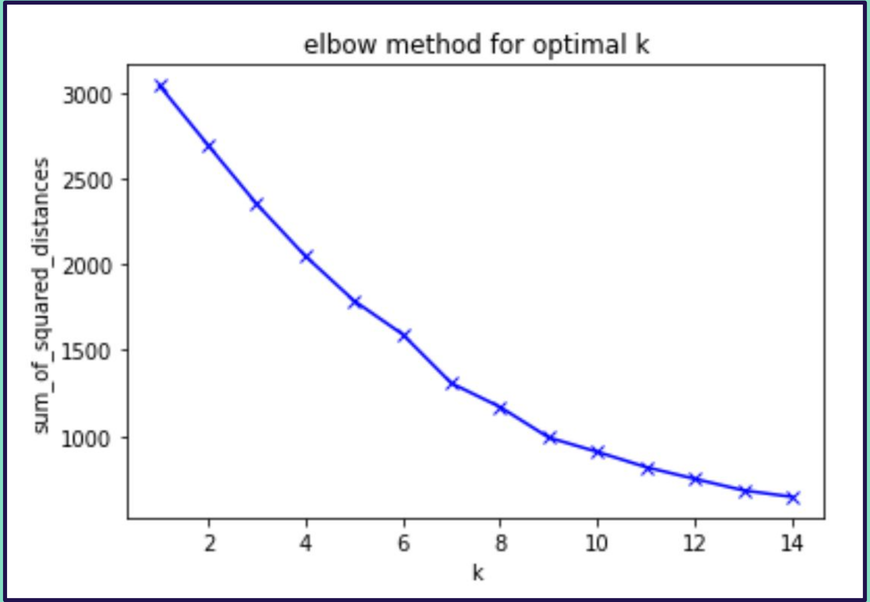
For Feature Selection

03. RANDOM FOREST

Determine Trend

04. LSTM

Determine Variation



K-MEANS CLUSTERING

Unsupervised learning

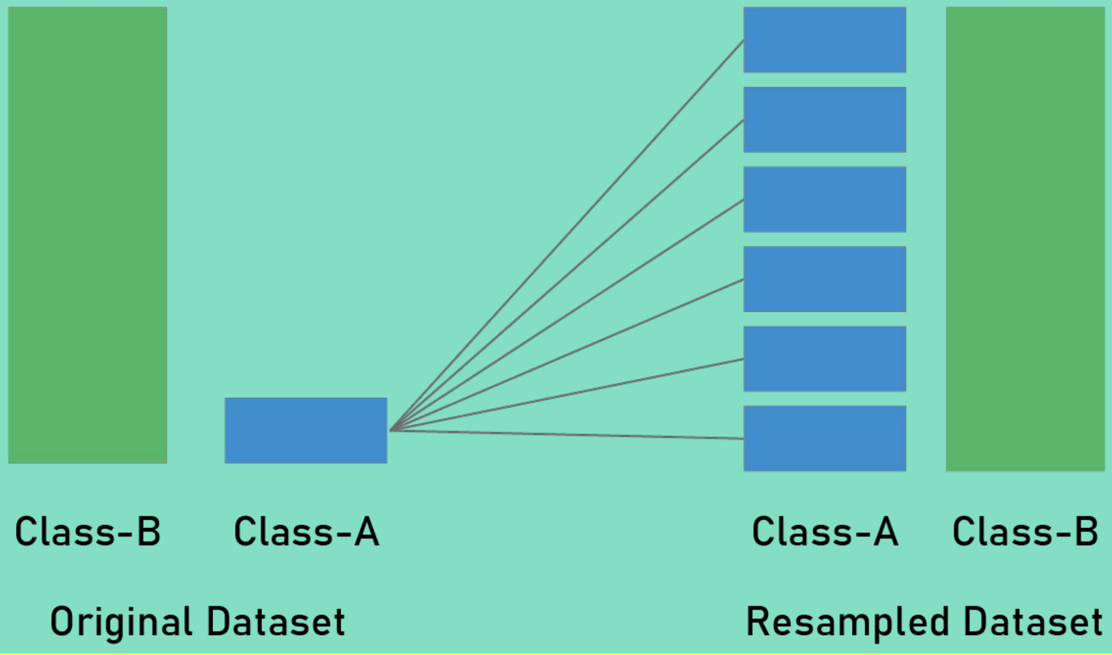
Only numerical input

Drop the country column

OPTIMAL # OF CLUSTERS: 7

Silhouette Coefficient: 0.2825; Calinski Harabasz Score: 42.37

Over Sampling

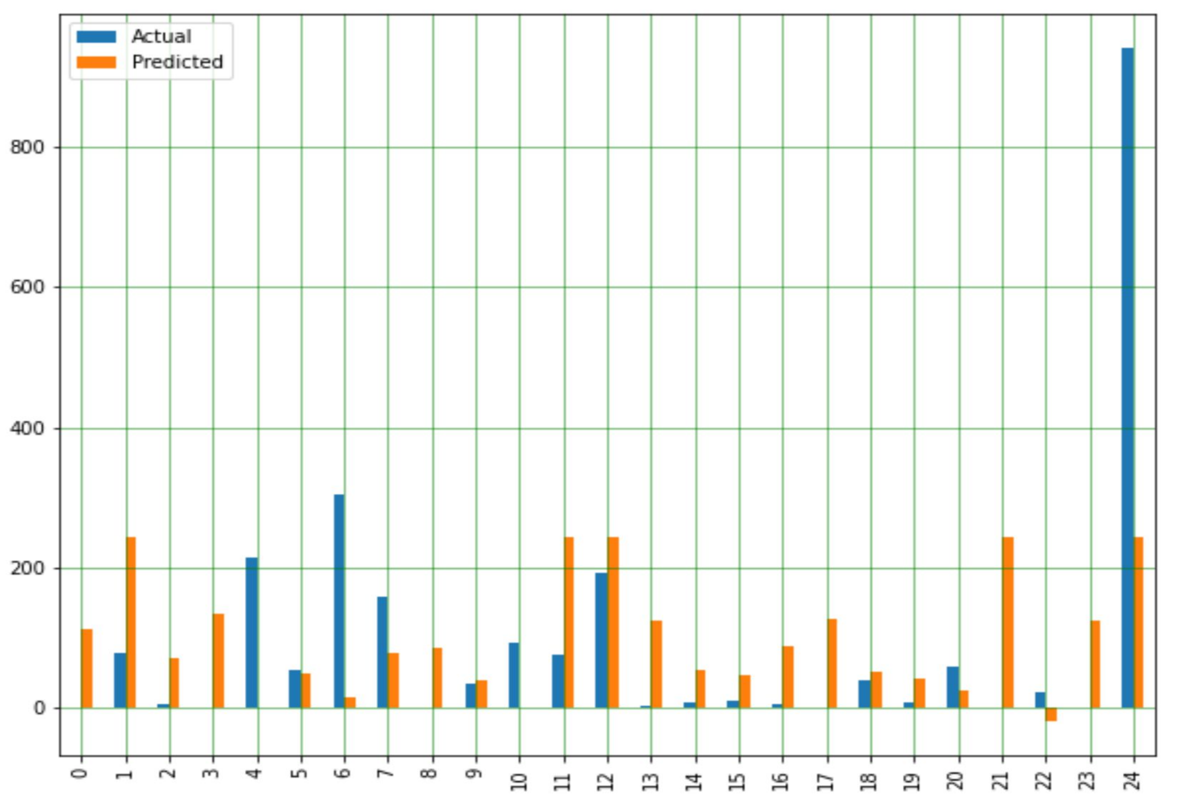


For Binary Only

SMOTE

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                  Daily   R-squared (uncentered):                   0.076
Model:                            OLS   Adj. R-squared (uncentered):              0.028
Method:                 Least Squares   F-statistic:                              1.577
Date:                Fri, 08 May 2020   Prob (F-statistic):                       0.147
Time:                        00:41:38   Log-Likelihood:                         -1015.6
No. Observations:                 142   AIC:                                      2045.
Df Residuals:                     135   BIC:                                      2066.
Df Model:                           7
Covariance Type:            nonrobust
=========================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
med_age                                 0.0786      1.446      0.054      0.957      -2.781       2.938
density                                 0.0201      0.011      1.768      0.079      -0.002       0.043
H3_Contact tracing                    -13.6201     42.493     -0.321      0.749     -97.658      70.418
E4_International support            -6.605e-06   2.56e-05     -0.258      0.796   -5.72e-05    4.39e-05
H4_Emergency investment in healthcare 7.846e-09   1.62e-07      0.048      0.961   -3.13e-07    3.28e-07
H5_Investment in vaccines           -5.608e-06   5.59e-05     -0.100      0.920      -0.000       0.000
H2_Testing policy                      56.0657     34.407      1.629      0.106     -11.981     124.113
E1_Income support                   -1.062e-12   4.09e-12     -0.260      0.795   -9.15e-12    7.03e-12
==============================================================================
Omnibus:                      185.312   Durbin-Watson:                   2.017
```

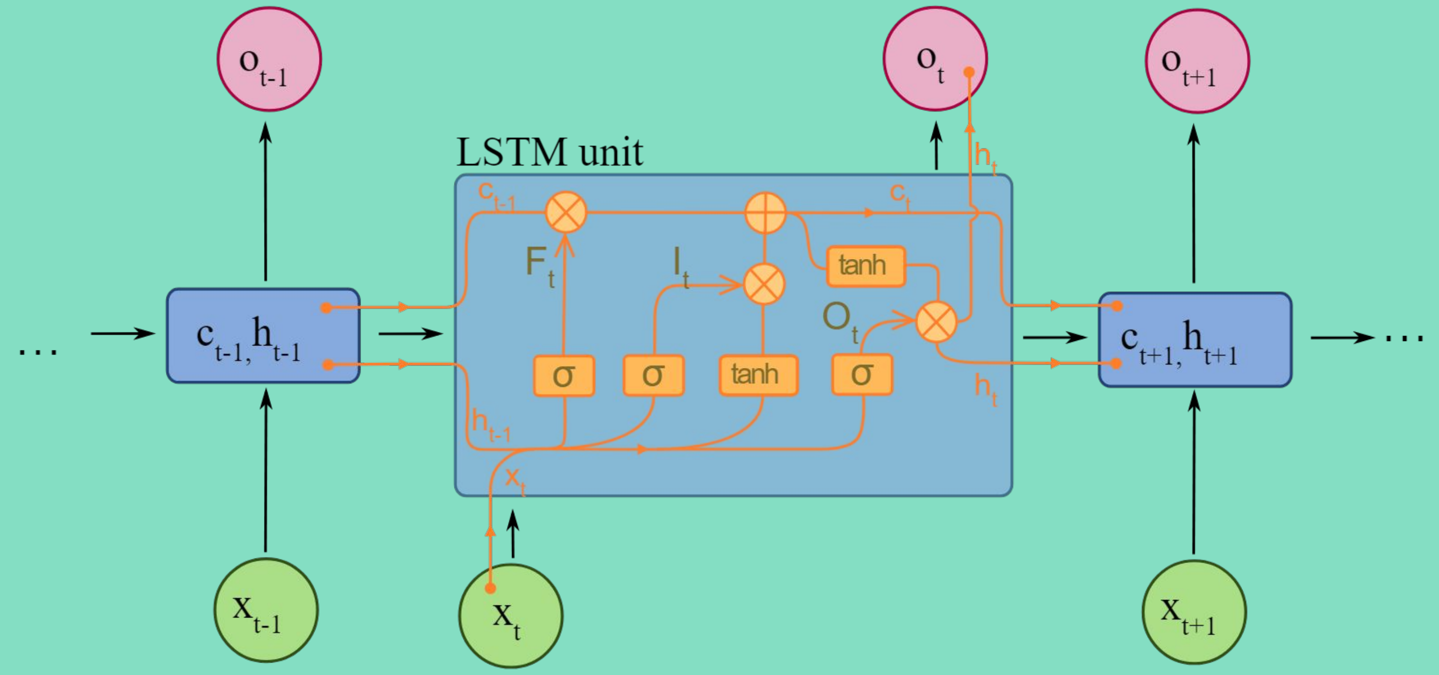
REGRESSION


Actual
Predicted
800
600
400
200
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

ACTUAL

```
Mean Absolute Error (MAE): 91.5225513570864
Mean Squared Error (MSE): 32004.957400494695
Root Mean Squared Error (RMSE): 178.89929401899465
Mean Absolute Percentage Error (MAPE): 24.22
Accuracy: 75.78
```

RANDOM FOREST REGRESSOR



LSTM NEURAL NETWORK

WHY LSTM?

- LSTM = Long Short Term Memory
- RNN (Recurrent Neural Network) that overcome technical problems
- RNNs fail to learn in the presence of time lags
- LSTM are better for time window-based feedforward networks
- Recall patterns that are very far into the past (or future)
- Resistant to noise (i.e. fluctuations in inputs that are random/irrelevant to predicting correct output)
- Parameters are trainable (in reasonable time)
- LSTM used for: handwriting recognition & generation, language modeling & translation, acoustic modeling of speech, analysis of audio, and video data

Italy

- ★ Use Italy as Comparable
- ★ Predict Italy from May 1 to May 9
- ★ Compare Predicted to Actual
- ★ Good predictor?

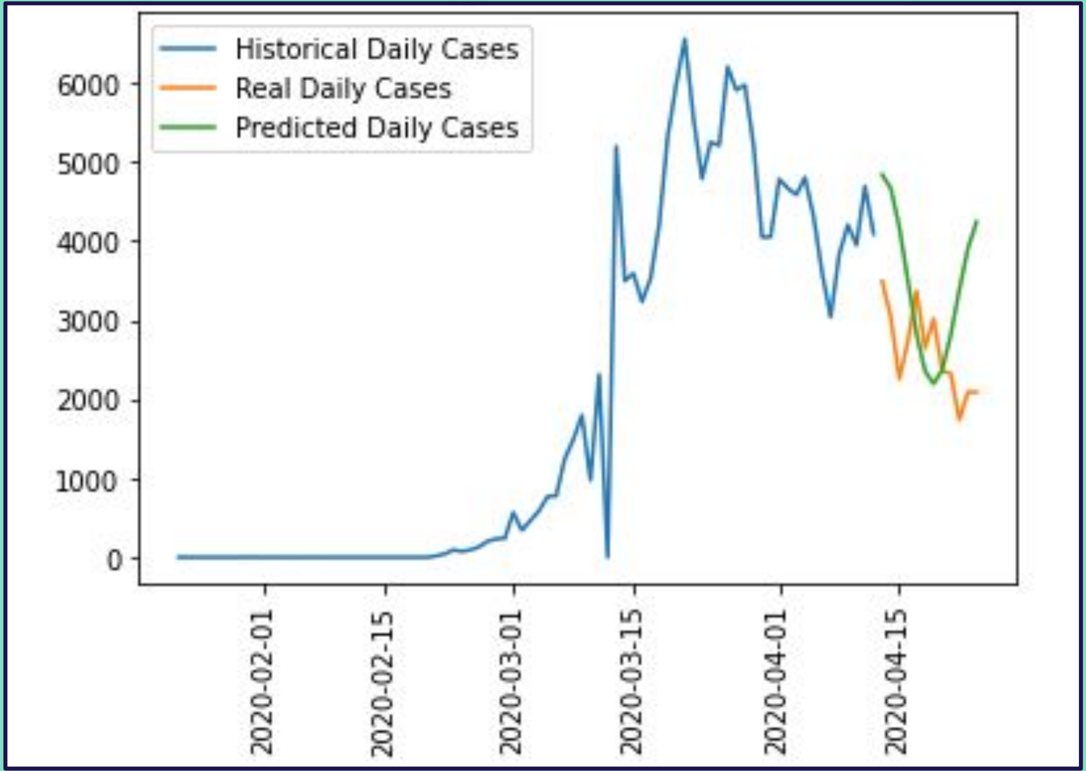
  Yes = Use to Predict China

  No = Seek Other Method

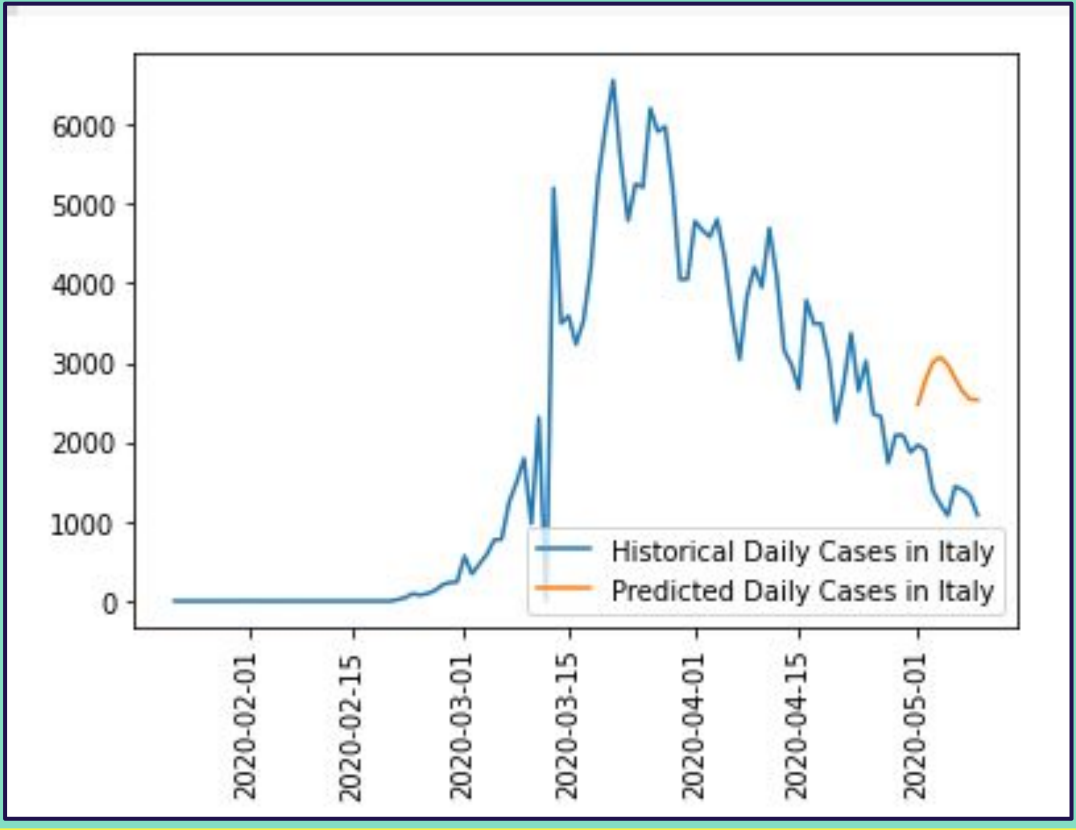
METHODOLOGY

Training

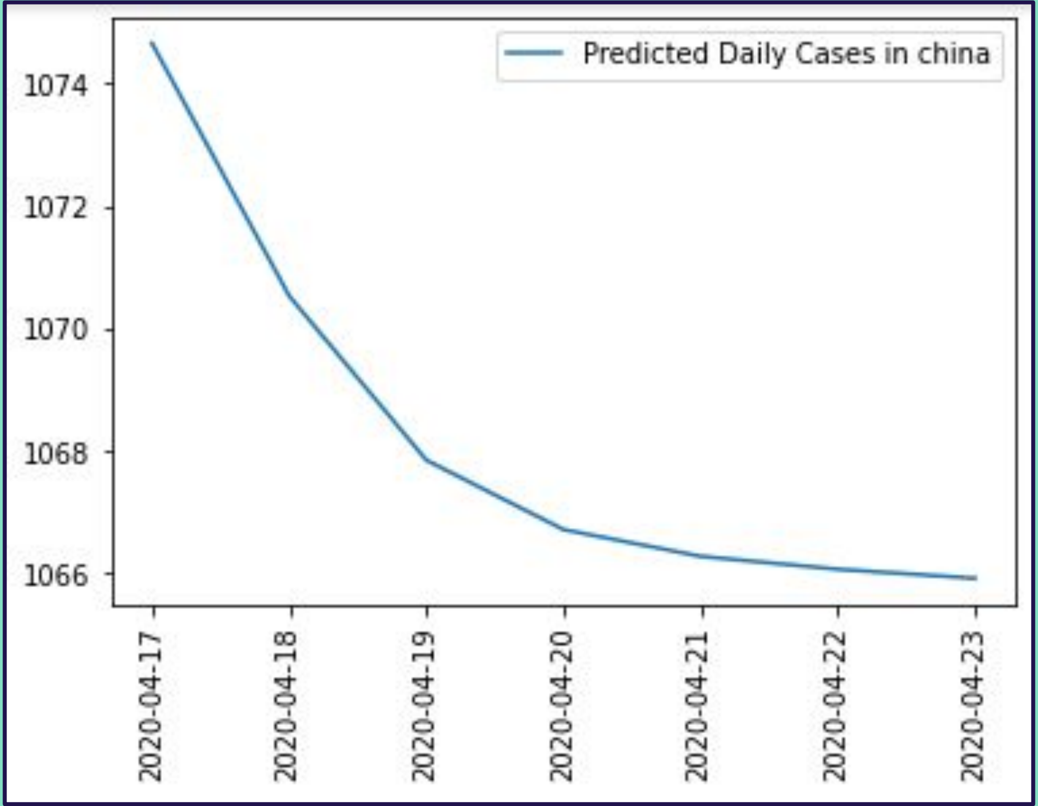
GOOD PREDICTOR



ITALY PREDICTION



ITALY PREDICTION VS ACTUAL



Apply LSTM

Using China's Data

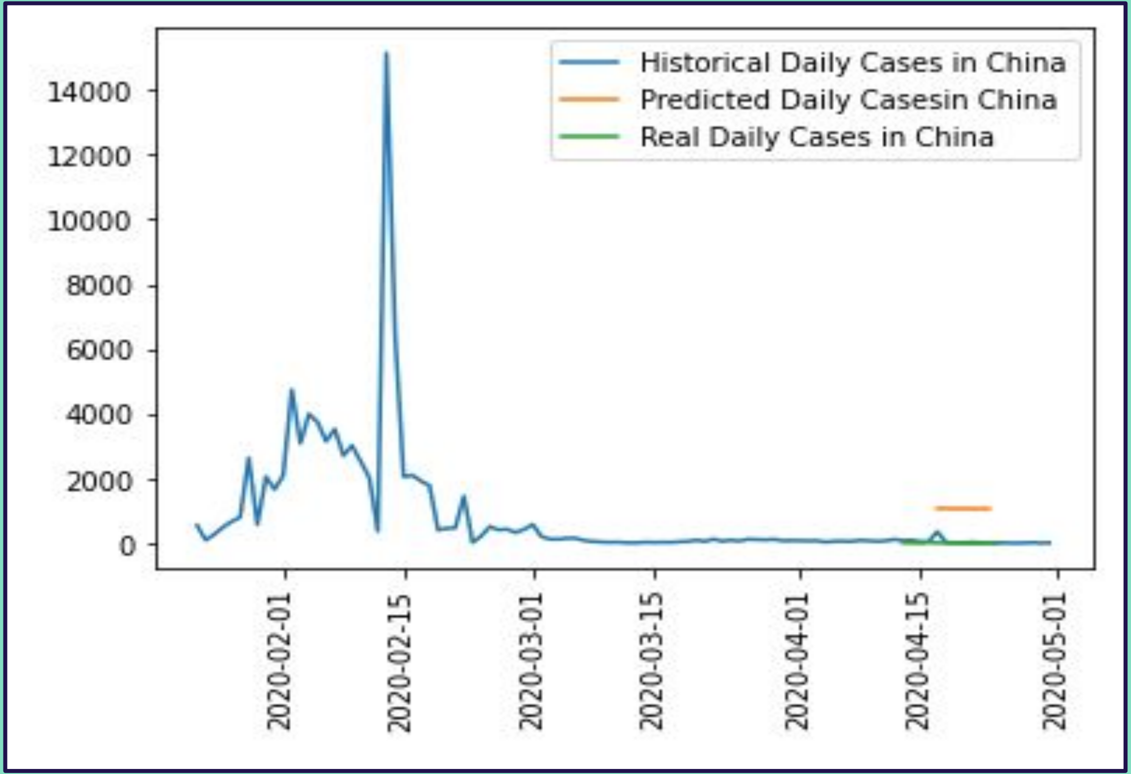
CHINA PREDICTION DAILY

CHINA PREDICTION

RESULTS: Inconclusive

Leveling off

Consistent w/Italy



04.

CONCLUSION

SUMMARY

01

CHINA

Figures Suspiciously Low

02

FEATURES

Government Stringency
Standardized: Population & Time

03

CLUSTERING

7 Clusters Optimal

04

SMOTE

Oversampling for Imbalanced
Data when Variable is Binary

05

RANDOM FOREST

Features Don't Correlate
Accuracy: 75.78

06

LSTM (Good Predictor)

Italy as Sample
China Results: Inconclusive

REFERENCES

- https://thediplomat.com/2020/03/can-chinas-covid-19-statistics-be-trusted/
- https://www.theguardian.com/world/2020/apr/09/the-cluster-effect-how-social-gatherings-were-rocket-fuel-for-coronavirus
- https://arxiv.org/pdf/2002.12298.pdf
- https://towardsdatascience.com/machine-learning-methods-to-aid-in-coronavirus-response-70df8bfc7861
- https://www.washingtonpost.com/politics/2020/03/23/china-is-reporting-big-successes-coronavirus-fight-dont-trust-numbers/
- https://www.youtube.com/watch?v=3kz54wBKi2c
- https://www.preventionweb.net/news/view/70092
- https://sph.hku.hk/en/news/press-releases/2020/nowcasting-and-forecasting-the-wuhan-2019-ncov-outbreak
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7014672/
- https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf
- https://www.bbc.com/news/world-52103747
- https://www.google.com/covid19/mobility/https://ourworldindata.org/coronavirus-data
- https://data.worldbank.org/indicator/sp.pop.totl
- https://www.kaggle.com/imdevskp/corona-virus-report
- https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
- https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker Visualizations: https://www.bbc.com/news/world-52103747
- http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm
- https://www.curiousily.com/posts/time-series-forecasting-with-lstm-for-daily-coronavirus-cases/
- https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203
- https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f

THANKS!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Please keep this slide for attribution