UCLA ANDERSON DATATHON 2020

COVID19 Analysis in relation to jobs in the United States
May 22, 2020

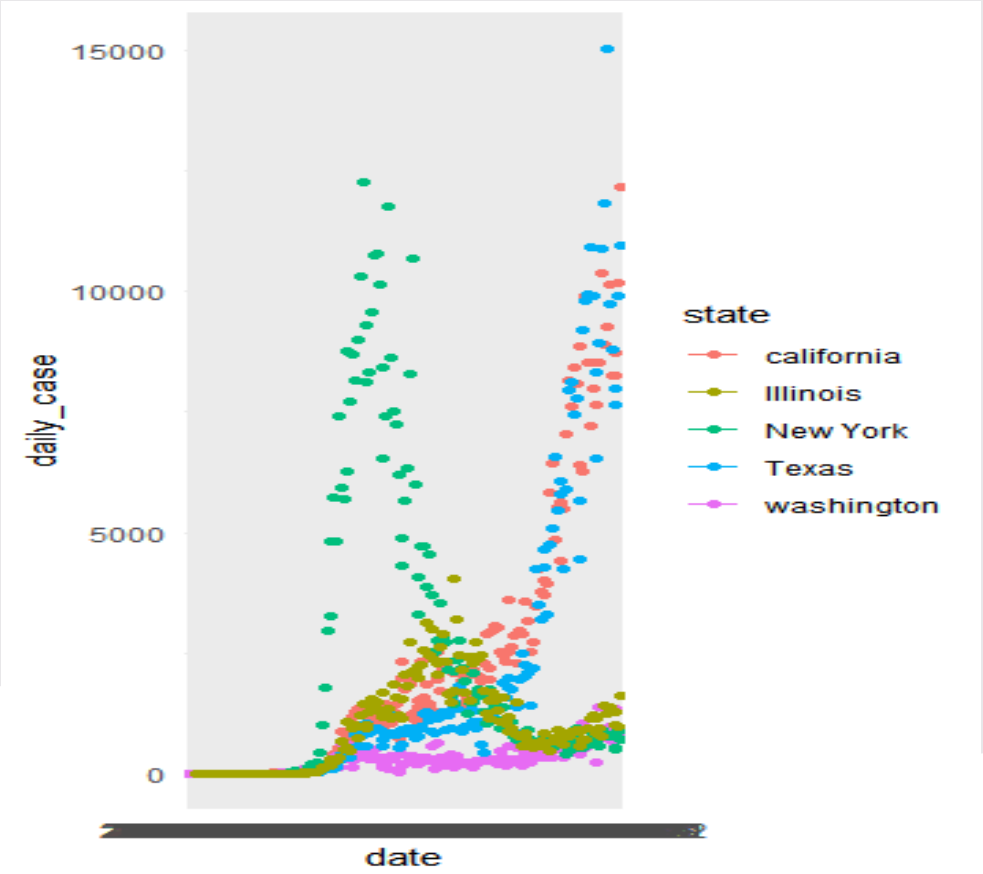
George, Hua, Taufeeq, Qiumeng

CONTENTS

PART 1 – COVID19 DATA ANALYSIS

States	Cumulative Cases	Cumulative Deaths
Washington	51,442	1,534
California	422,528	8,038
New York	413,595	32,228
Illionois	167,142	7,544
Texas	366,561	4,522

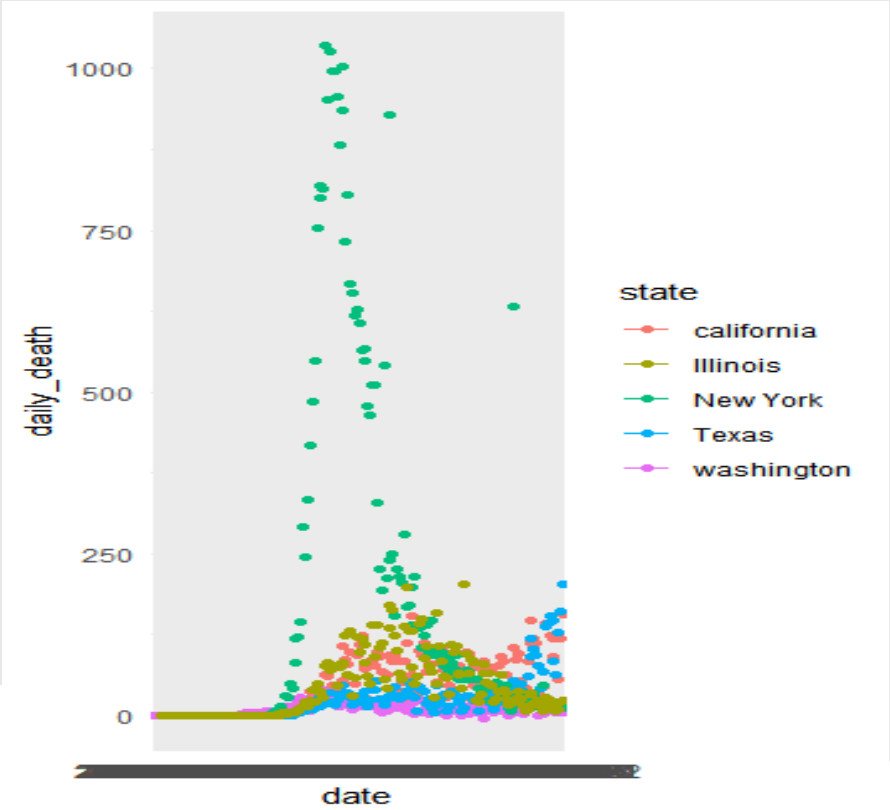
Exploratory Data Analysis



Based on the Visualization

- New York experienced a spike earlier during April in cases and deaths
- California and Texas are rapidly increasing in cases in the later months
- Illinois has had a smaller spike in between the New York peak and California increasing period

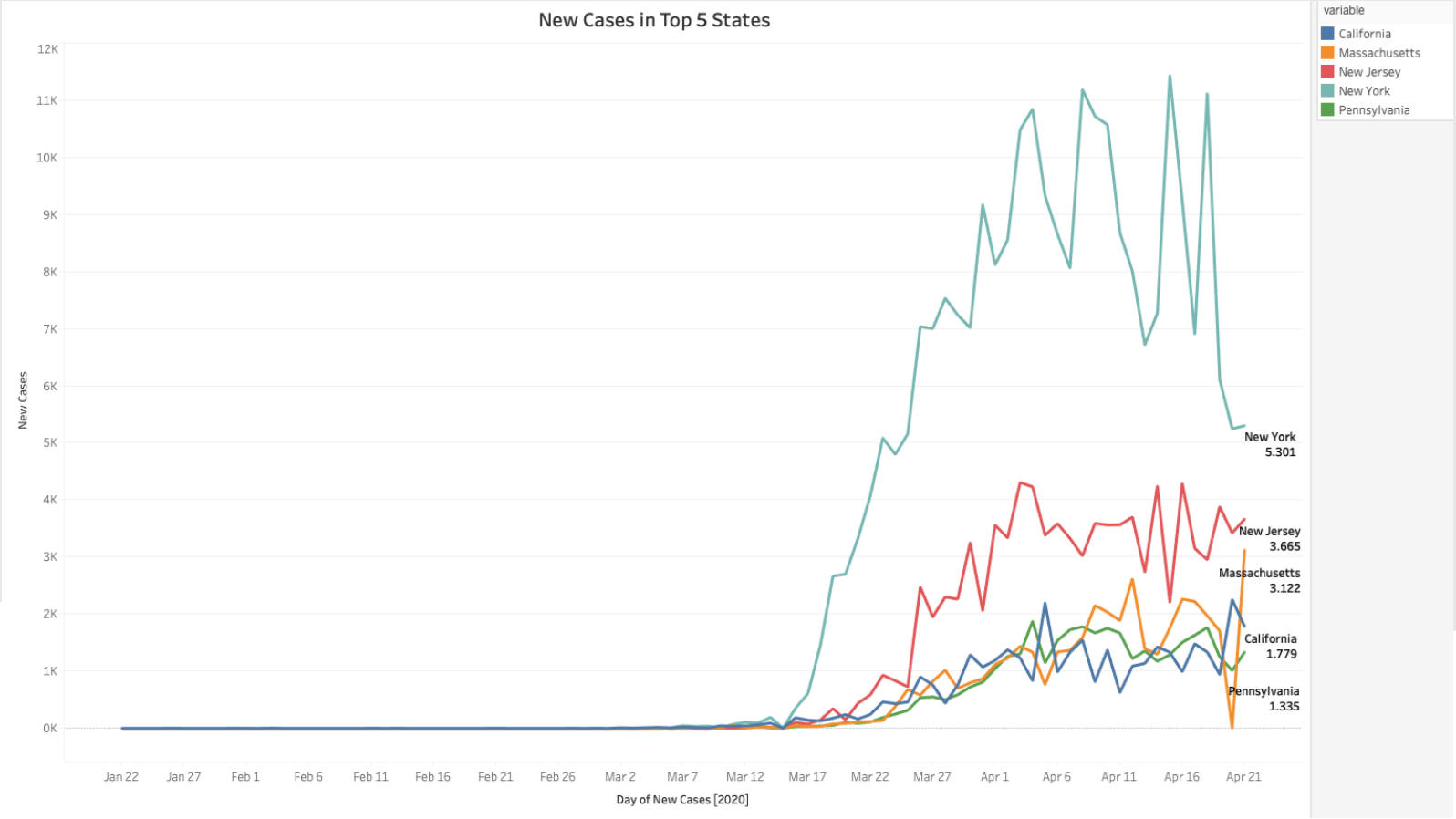
Exploratory Data Analysis



Based on the Visualization

- New York experienced the highest number of deaths due to its high spike in April
- California and Texas is increasing in terms of deaths
- Commonality between these three states is that they are highly densed compared to Illinois and Washington
- Less distance and increase chance of physical contact with infected people increasing cases/deaths in 3 states

Daily New Confirmed Cases by State

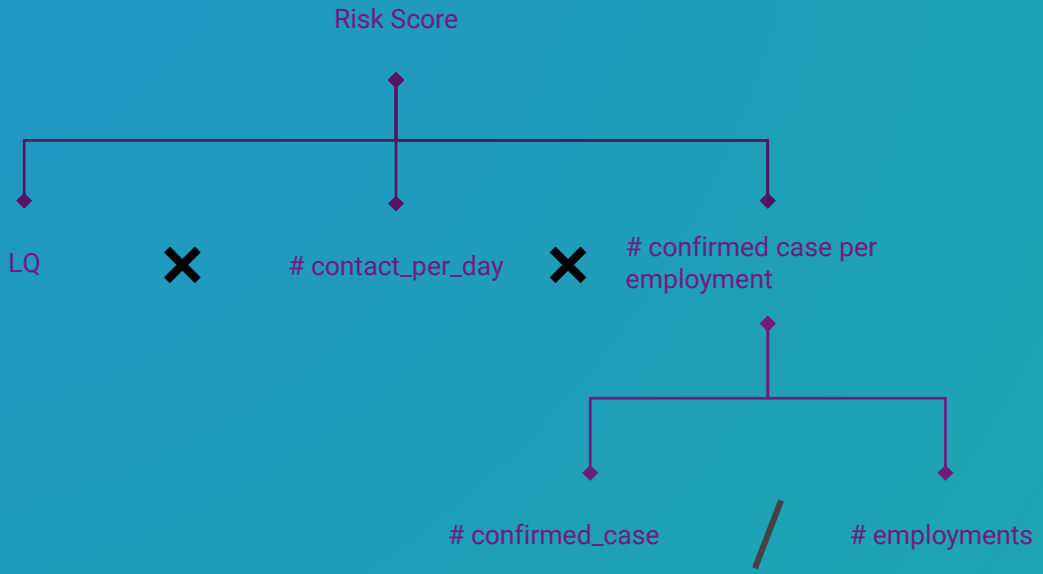


UCLA Anderson
SCHOOL of MANAGEMENT

PART 2 – JOB LOCATION QUOTIENT

Risk Score Modeling, High-Risk Occupations

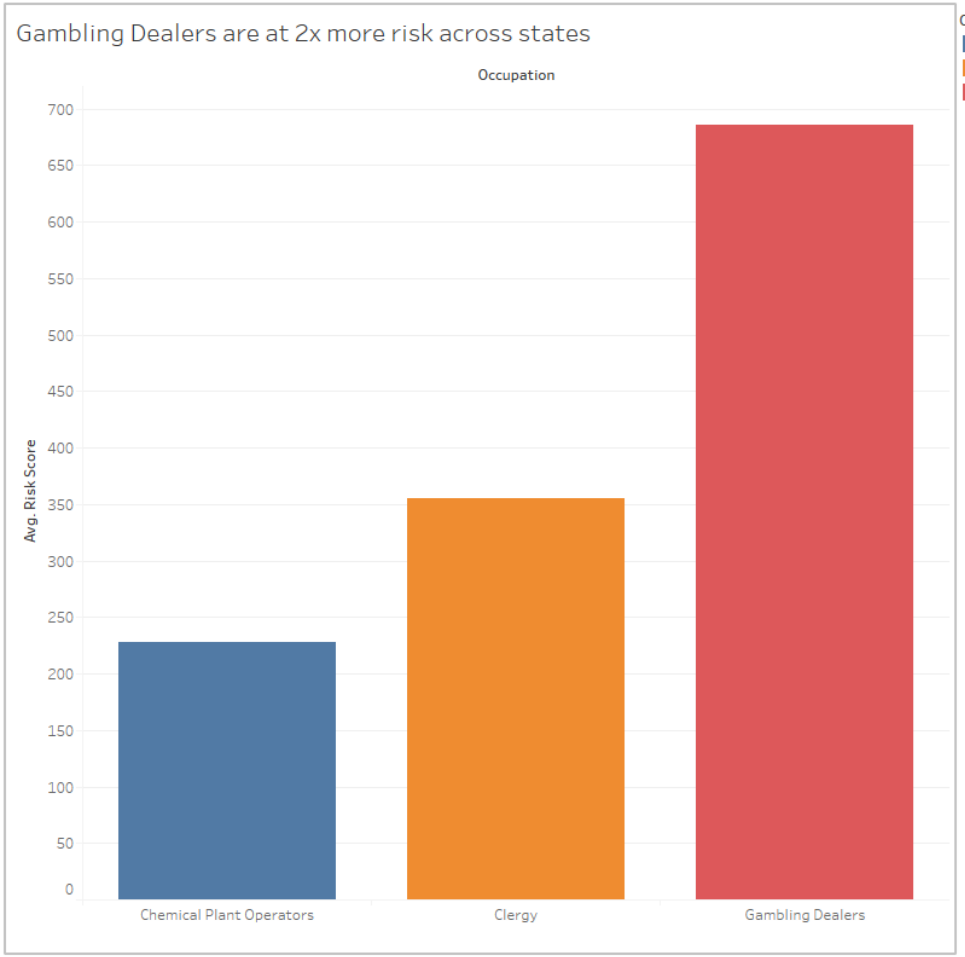
Investigating occupations at risks in these states



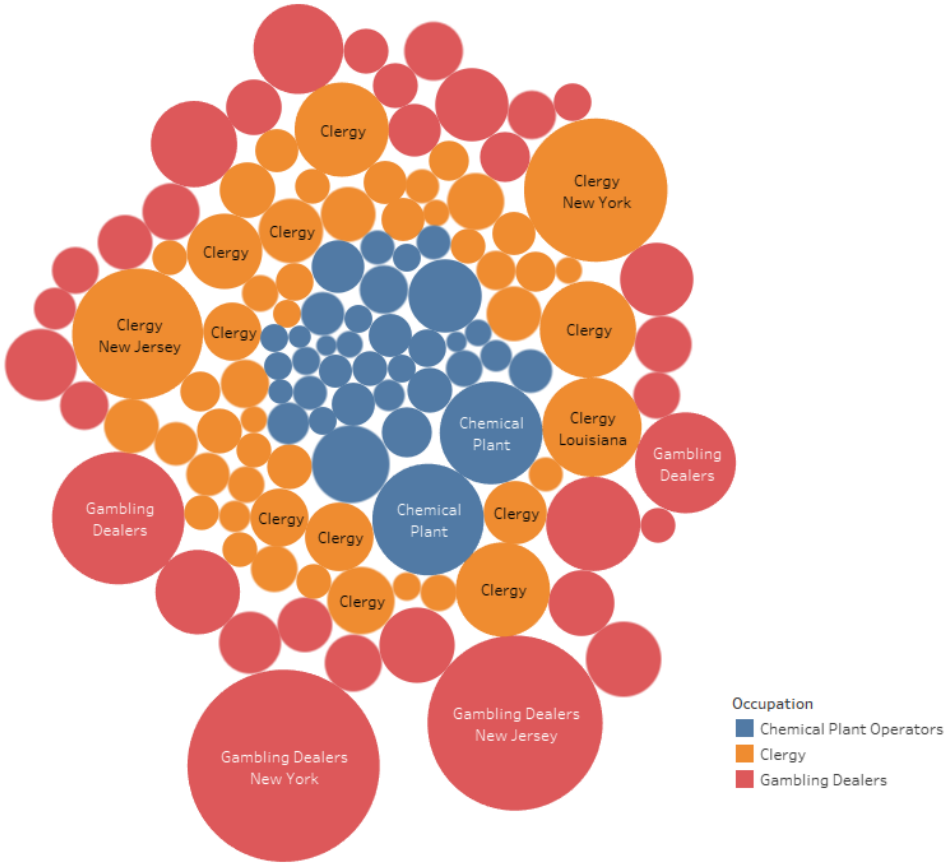
Clergy

Gambling Dealer

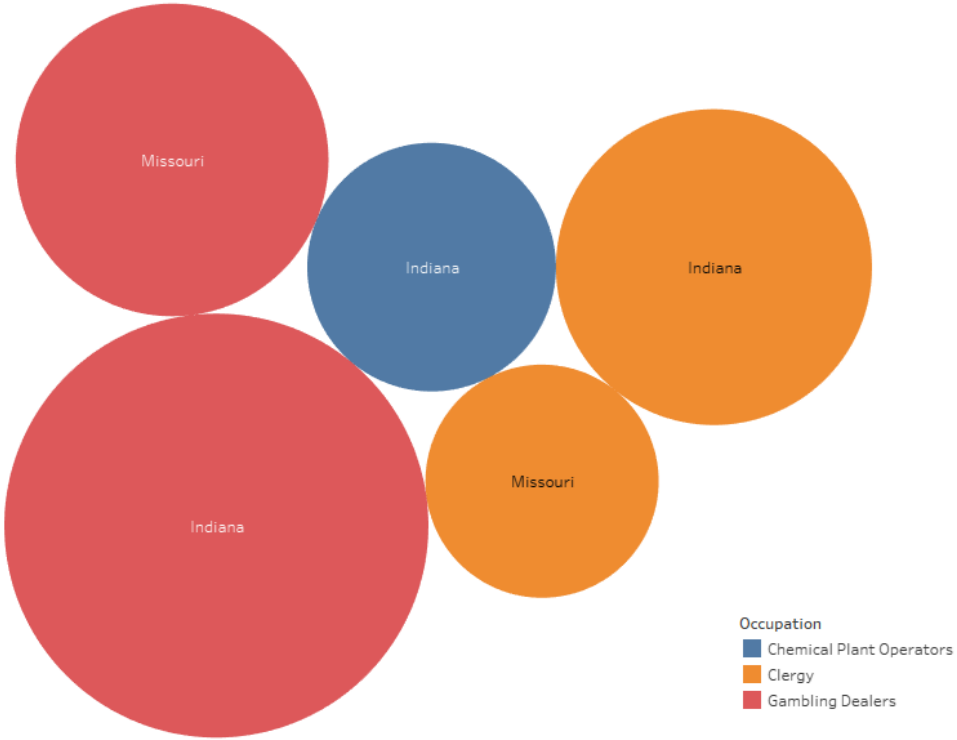
Chemical Plant Worker


Gambling Dealers are at 2x more risk across states
Occupation
Avg. Risk Score
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
Chemical Plant Operators
Clergy
Gambling Dealers
Occupation
Chemical Plant Operators
Clergy
Gambling Dealers

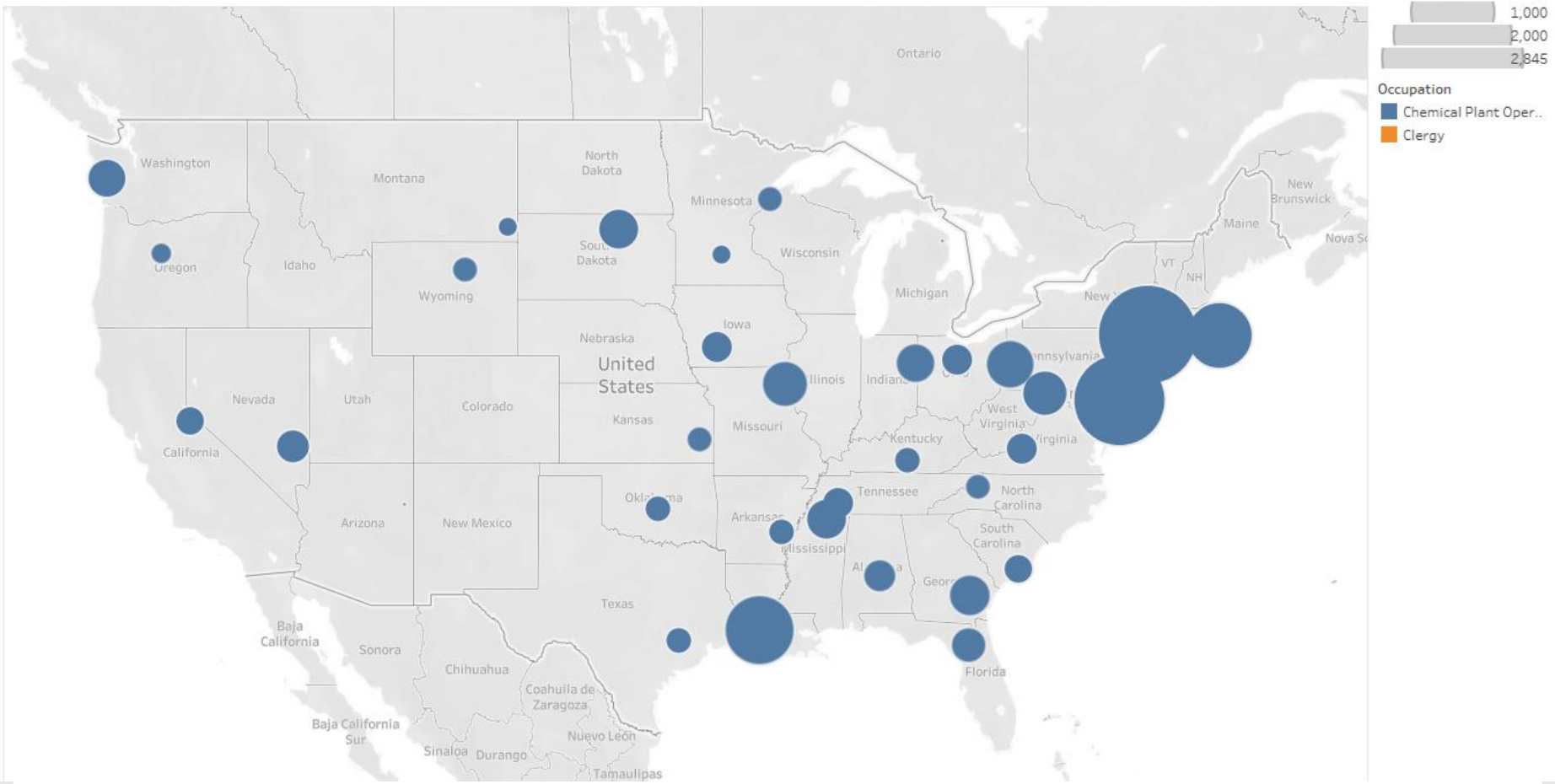
Across occupations, Gambling Dealers in New York and New Jersey are at most risk



Your response varies drastically depending on the state your business is in



States with highest comfirmed cases are at a higher risk irrespective of the occupation



PART 3 – DEEPER DIVE INTO COVID19

Possible Method for Forecast - 1

RNN - LSTM Network

Recurrent Neural Network - Long, Short Term Network is used often to predict the future with backpropagation (i.e.: backstep) sequence

Good for addressing difficult sequence, large data problems in machine learning & used often in predicting customer purchasing patterns in eCommerce

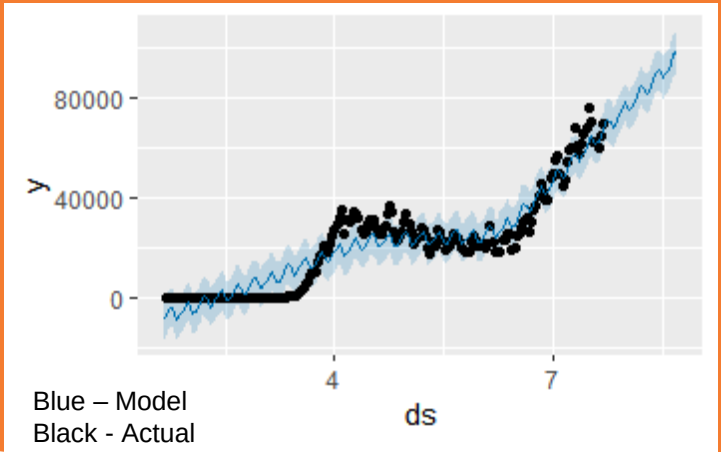
Drawbacks:

1) Normalizing the data doesn't seem appropriate – may misrepresent the increase
2) Accuracy of the model was sub 20% after training

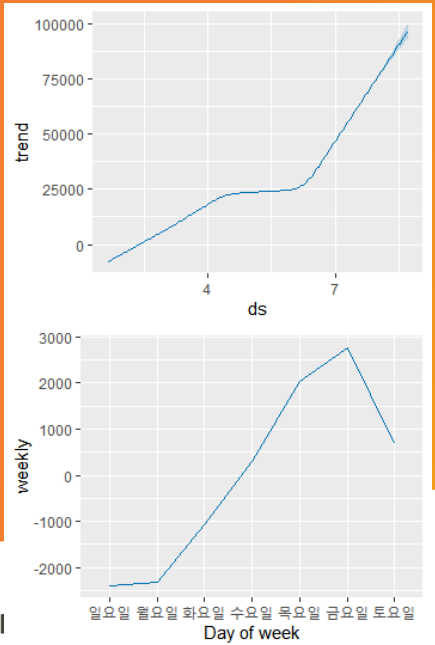
Possible Method for Forecast - 2

Prophet from Facebook

Use non-linear with seasonality modeling to predict time-series based data



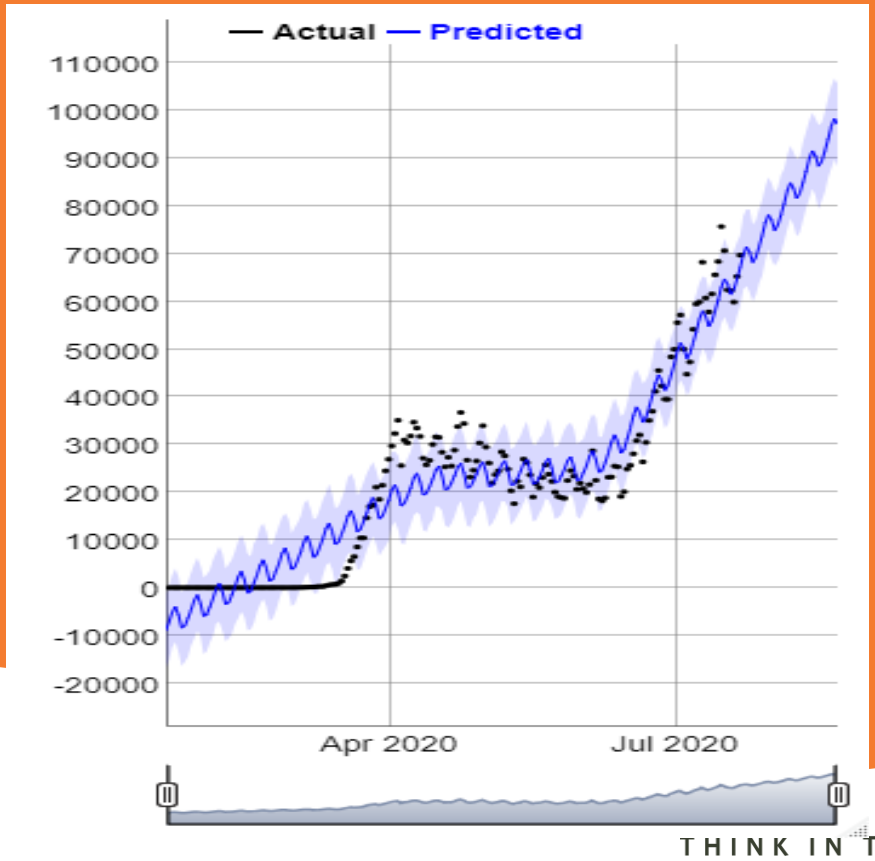
Blue – Model
Black - Actual



Interesting Insights

- U.S. Daily cases stay steady between April and July because testing capacity was limited
- Cases skyrocket since late June, signaling that testing more reveals more people that were infected
- Positive test results show highest infection numbers in later half of the week since most it takes 2-3 days to get results back

Possible Method for Forecast - 3



Model prediction results

- Compared to less complex models (e.g.: CART), Prophet shows higher forecast accuracy
- Accuracy of model compared using August data as holdout set
  (Refer to Git repository for codes, results)

Metrics	Prophet	CART	Diff
MSE	1.78E+08	3.57E+08	-1.8E+08
RMSE	13356.01	18903.56	-5547.56
MAE	8269.061	14387.44	-6118.37


THINK IN THE NEXT