COVID-19

Case Count Analysis

By Sourav Mohan, Harrison Fisher, Zhuoquan Chen, Shumo Zheng

Not So Fun Facts about Covid-19

- Type A blood is more susceptible
- Reinfection is possible, having it in the past does not mean you’re immune
- Can live on surfaces for days
- Anosmia - loss of smell
- Is more dangerous for small children and the elderly
- Texas and California both surpass 1 million confirmed cases
- Carriers are often Asymptomatic

Project Goals / Objectives

- The primary aim of this analysis is to highlight the key factors that contributed to confirmation of Covid-19 cases in the United States of America.
- Multiple Models will be assessed in order to determine the best and most effective predictor of confirmed Covid-19 cases.
- The goal is to create a model that can be used to predict future Covid-19 cases and cases of a future disease of similar magnitude.

BigQuery Datasets

List and Describe them:

- Mobility - change from baseline activity
- Open - various information such as temperature, hospitalized, recovered, etc.
- Policy - what policies were in place in certain counties
- Mask - how often it is advised to wear a mask
- Symptoms - the symptoms in different counties

Cleaned Covid Data Dictionary

Change from Baseline Mobility	Aggregation Level	Policies	Economic Measures	Stringency Index
Mobility trends in various places	Level of cluster forming	Numeric scale measuring how strict government policy was	Various economic relief measured in USD	Scale of 1-100 of how well the government responded to Covid 19
Retail and recreation, Parks, Transit, Work place, Residential, etc.		Testing policy, Stay at home order, Gathering restrictions, etc.,	Debt, Relief, stimulus, Vaccine investment, etc.	

Analysis of Covid Confirmed Cases by Policy, Climate, & Mobility

Policy Dataset Features

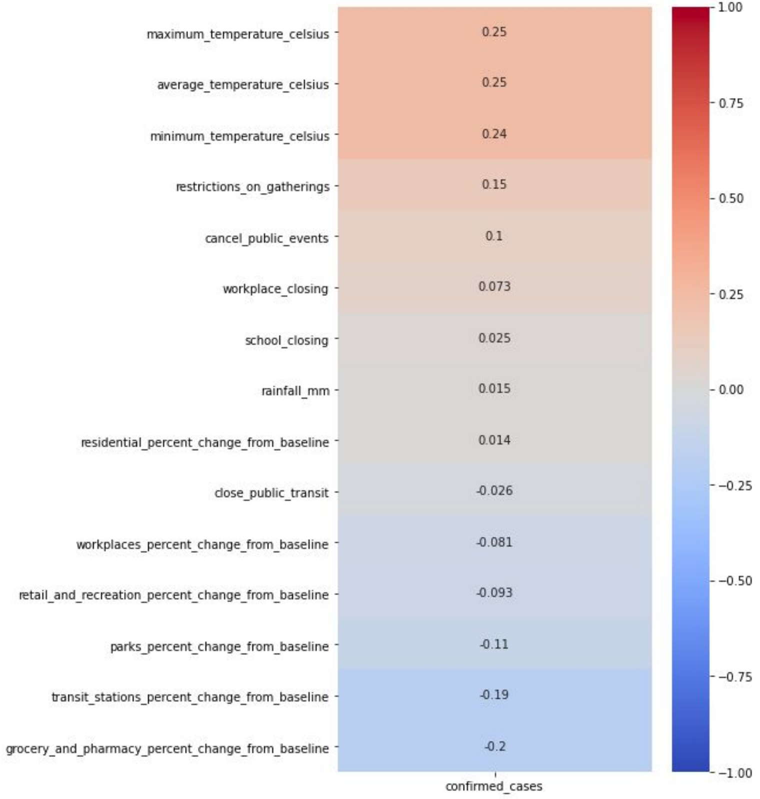
- Cancel Public Events
- Restriction on Gatherings
- International Travel Controls
- School closing policies
- Public Information Campaigns
- Testing Policy
- Workplace Closing policies

Mobility Dataset Features

- Grocery Store and Pharmacy Traffic Change
- Parks Percent Traffic Change
- Transit Station Traffic Change

Open Dataset Features

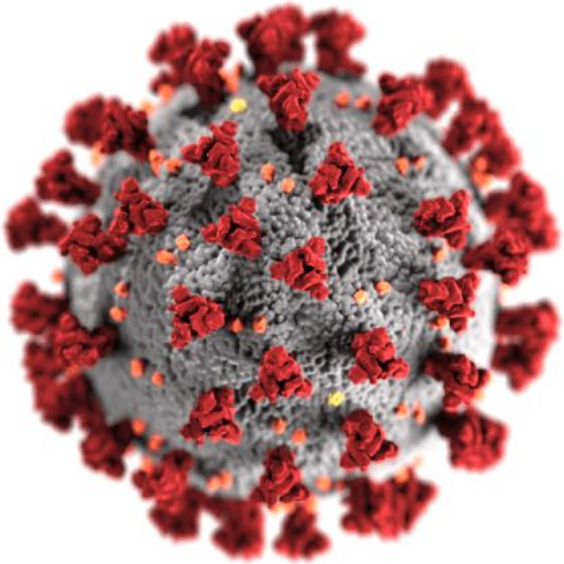
- State of California, New York, Florida and Texas
- Rainfall(mm), Average, Maximum, and Minimum Temperature (Celsius)
- Date (DD, MM, YYYY)



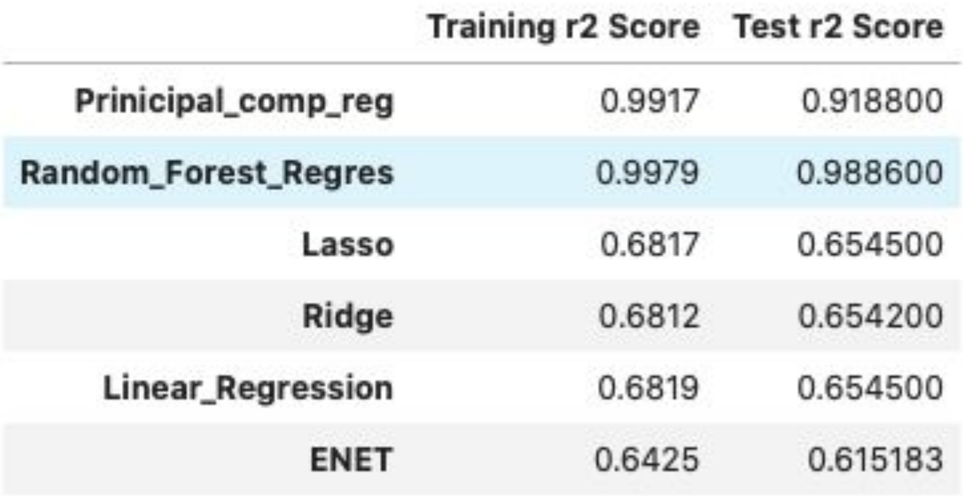
Models Assessed on Mobility, Policy, location and Climate Dataset

Models Fitted:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net
5. Random Forest Regression
6. Principal Component Regression



Covid Confirmed Cases: Model Scores

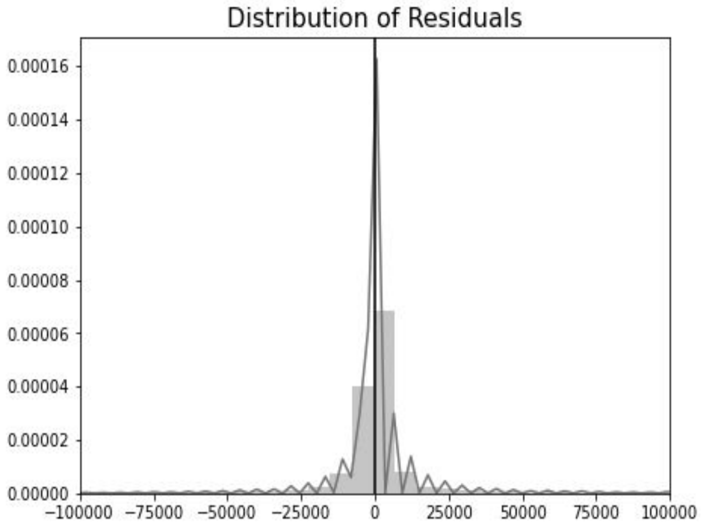


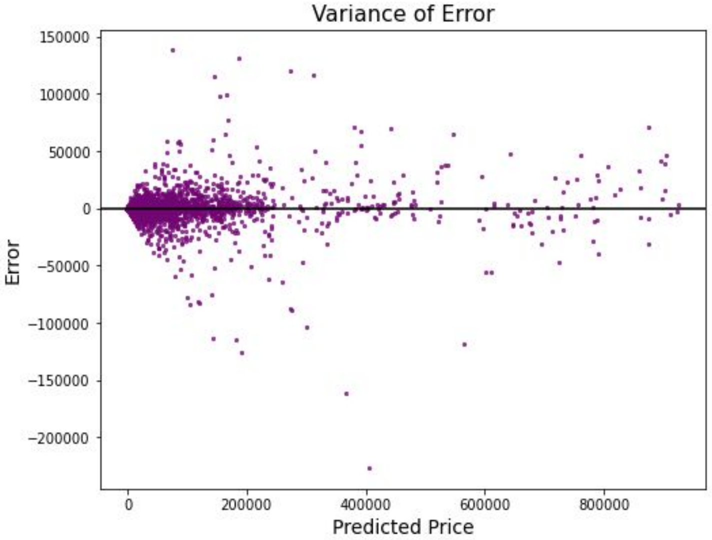
	Training r2 Score	Test r2 Score
Prinicipal_comp_reg	0.9917	0.918800
Random_Forest_Regres	0.9979	0.988600
Lasso	0.6817	0.654500
Ridge	0.6812	0.654200
Linear_Regression	0.6819	0.654500
ENET	0.6425	0.615183

Top Two Models:

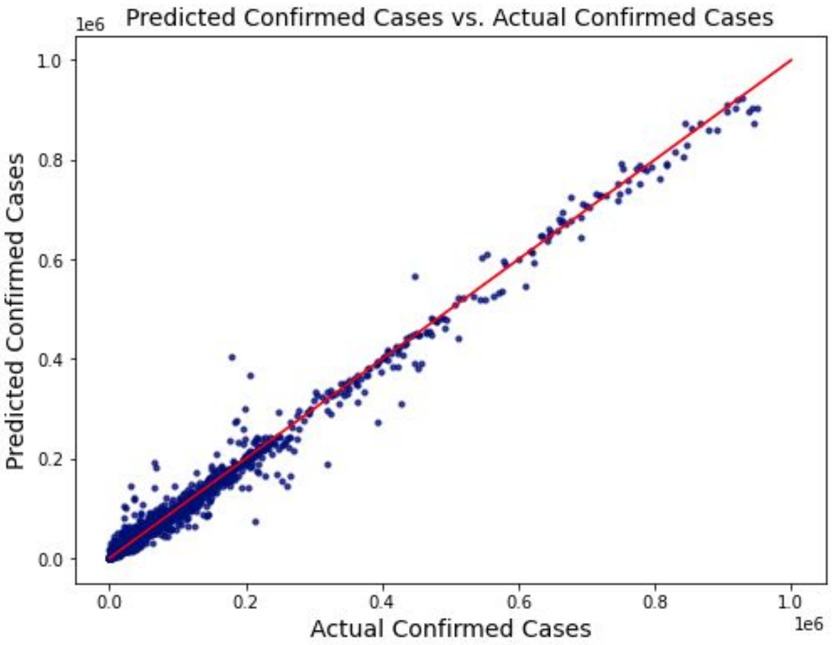
1. Random Forest Regression
2. Principal Component Regression (using Random Forest Regression)

Random Forest Regression





The Model’s prediction



- Random Forest Regression Model has an Training r2 score score of 99.8%
- The Model’s r2 score on testing data is 98.9%.
- The RMSE for the Model was around 5,649.

Google Search Symptoms for 2020 & 2019

To public health workers and to researchers in the scientific community, thank you

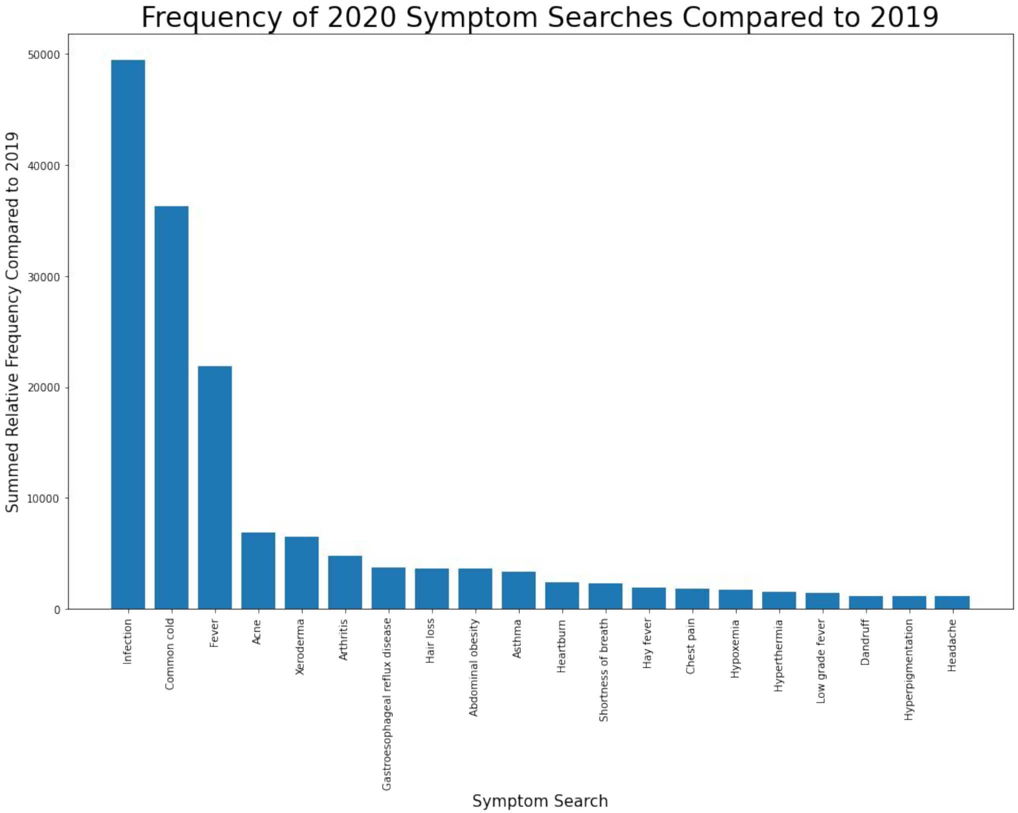
Photo by: Google

Data:

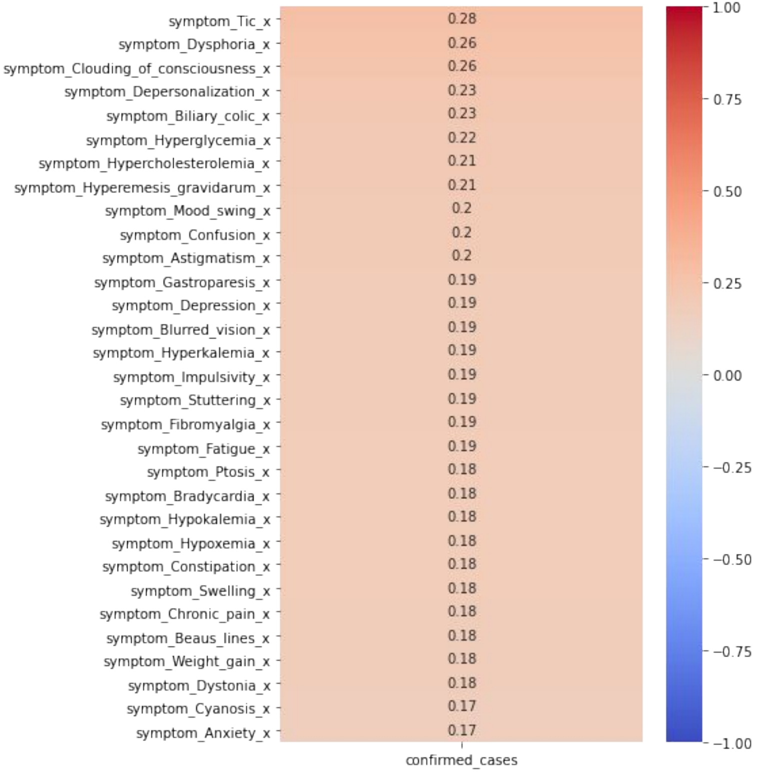
- Contains aggregated trends in Google Searches for health symptoms
- The data set is further organized by date and location
- 422 symptom searches were studied

Change in Google Search Symptoms: 2020 to 2019

1. Infection
2. Common cold
3. Fever
4. Acne
5. Xeroderma
6. Arthritis
7. Gastroesophageal Reflux Disease
8. Hair Loss
9. Abdominal Obesity
10. Asthma
11. Heartburn
12. Shortness of Breath
13. Hay Fever
14. Chest Pain
15. Hypoxemia
16. Hyperthermia
17. Low Grade Fever
18. Dandruff
19. Hyperpigmentation
20. Headache

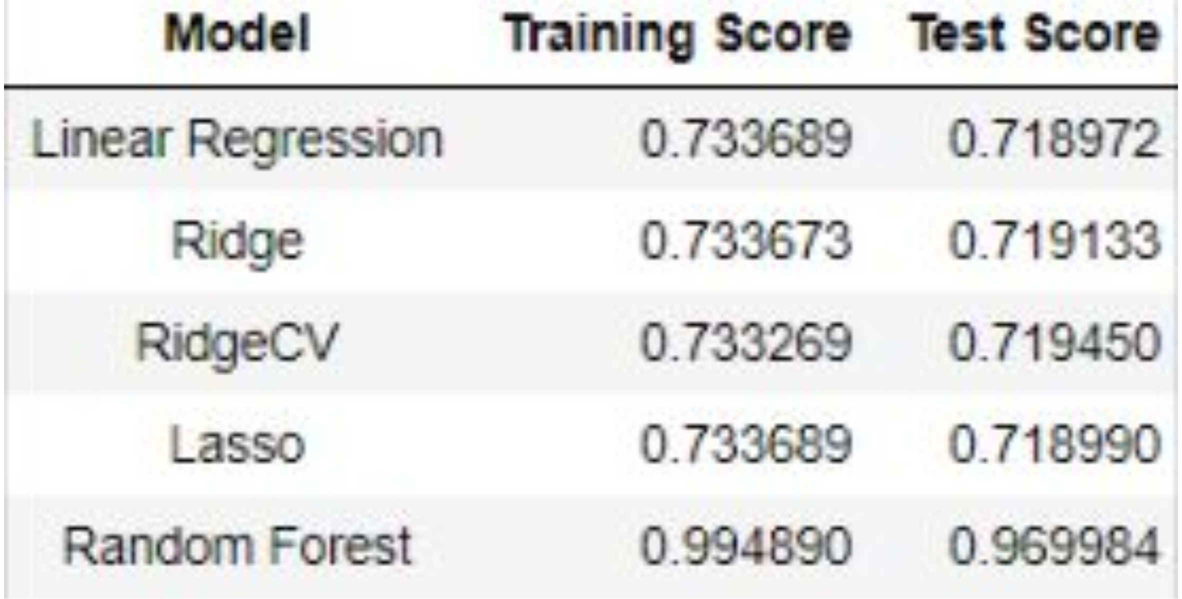


Top Correlation to Covid Case Count



- Behavioral Symptoms:
  - Dysphoria - a state of unease
  - Clouding of Consciousness
  - Mood Swings
  - Confusion
  - Depression
  - Impulsivity
  - Anxiety
- Others:
  - Fatigue
  - Weight Gain

Models Using Search Data



Model	Training Score	Test Score
Linear Regression	0.733689	0.718972
Ridge	0.733673	0.719133
RidgeCV	0.733269	0.719450
Lasso	0.733689	0.718990
Random Forest	0.994890	0.969984

Mask Data Set

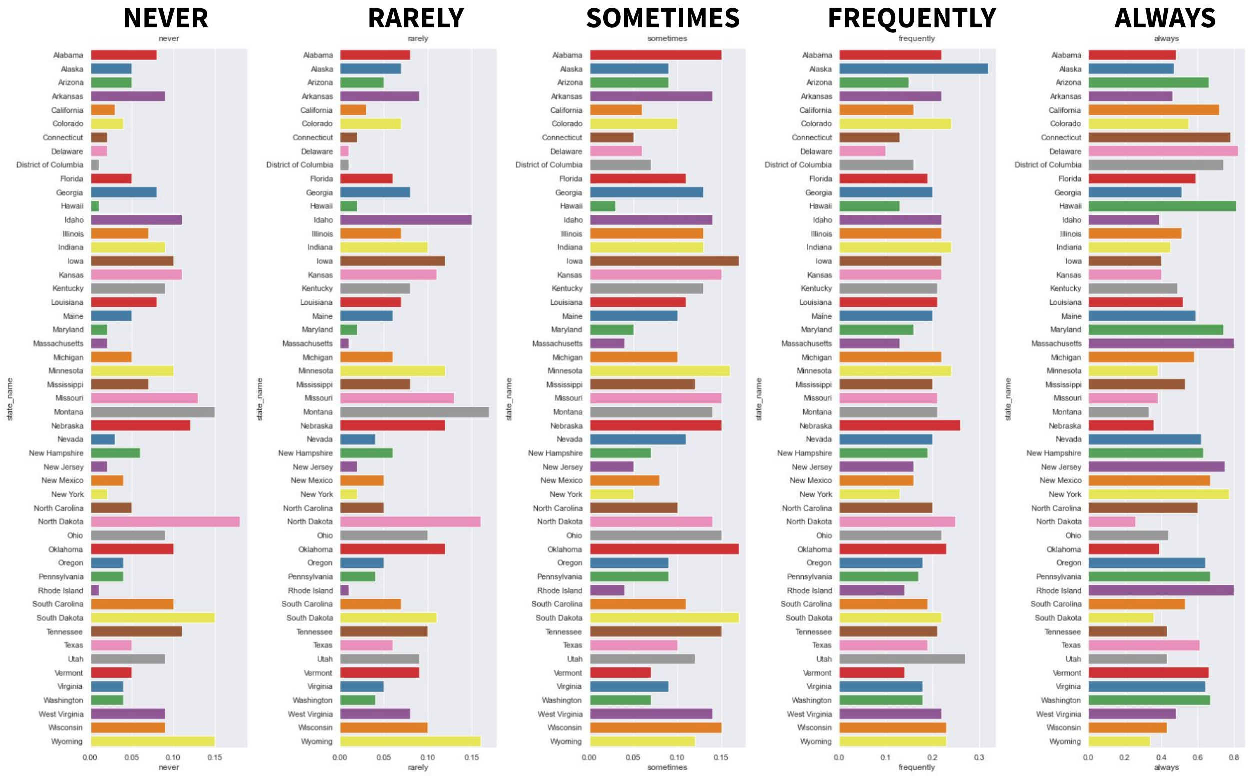
Survey Response conducted by Dynata for The New York Times

- 250,000 survey responses between July 2 and July 14

How often do you wear a mask in public when you expect to be within six feet of another person?”

- NEVER
- RARELY
- SOMETIMES
- FREQUENTLY
- ALWAYS

Mask Use in the U.S.



SAY NEVER

North Dakota **18%**

SAY SOMETIMES

Iowa **17.36%**

Oklahoma **17.41%**

South Dakota **17%**

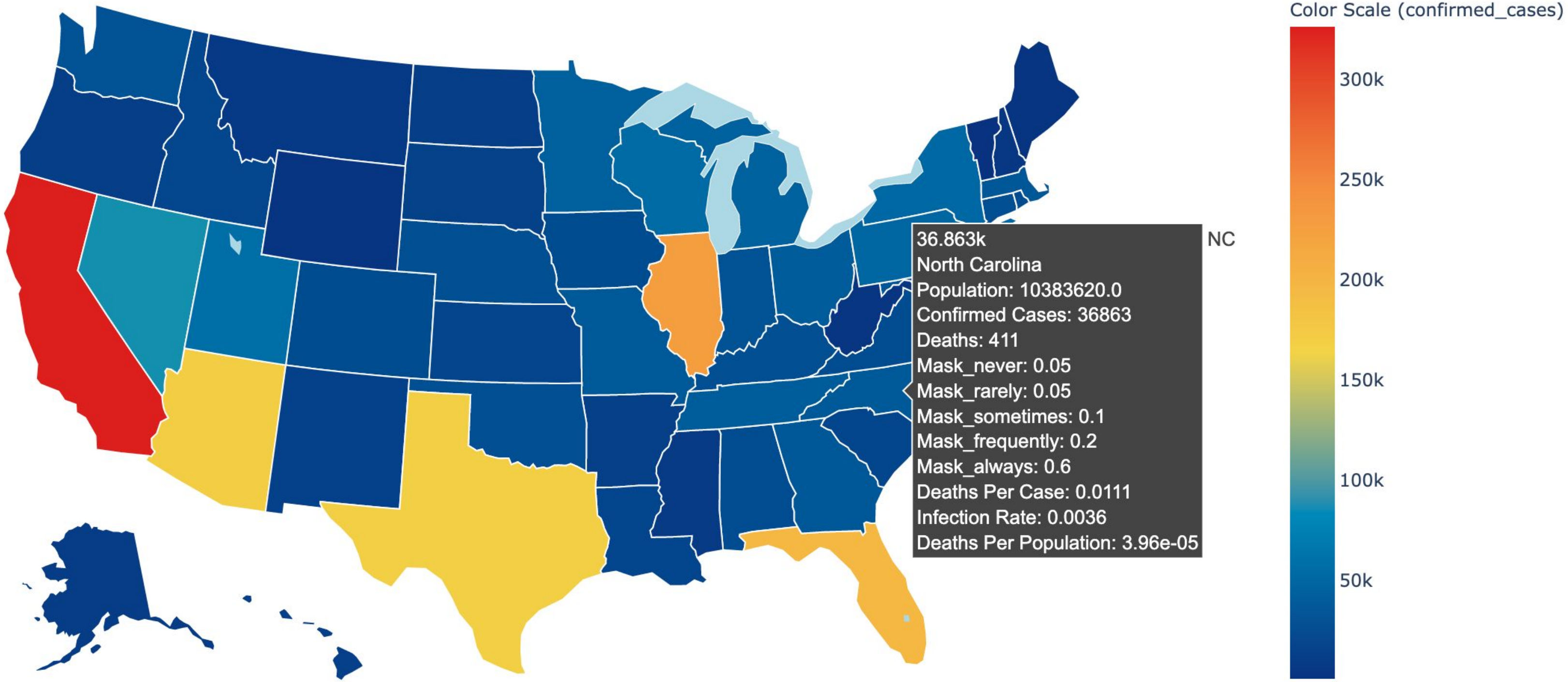
SAY ALWAYS

Delaware **82%**

NOT INTUITIVE ?

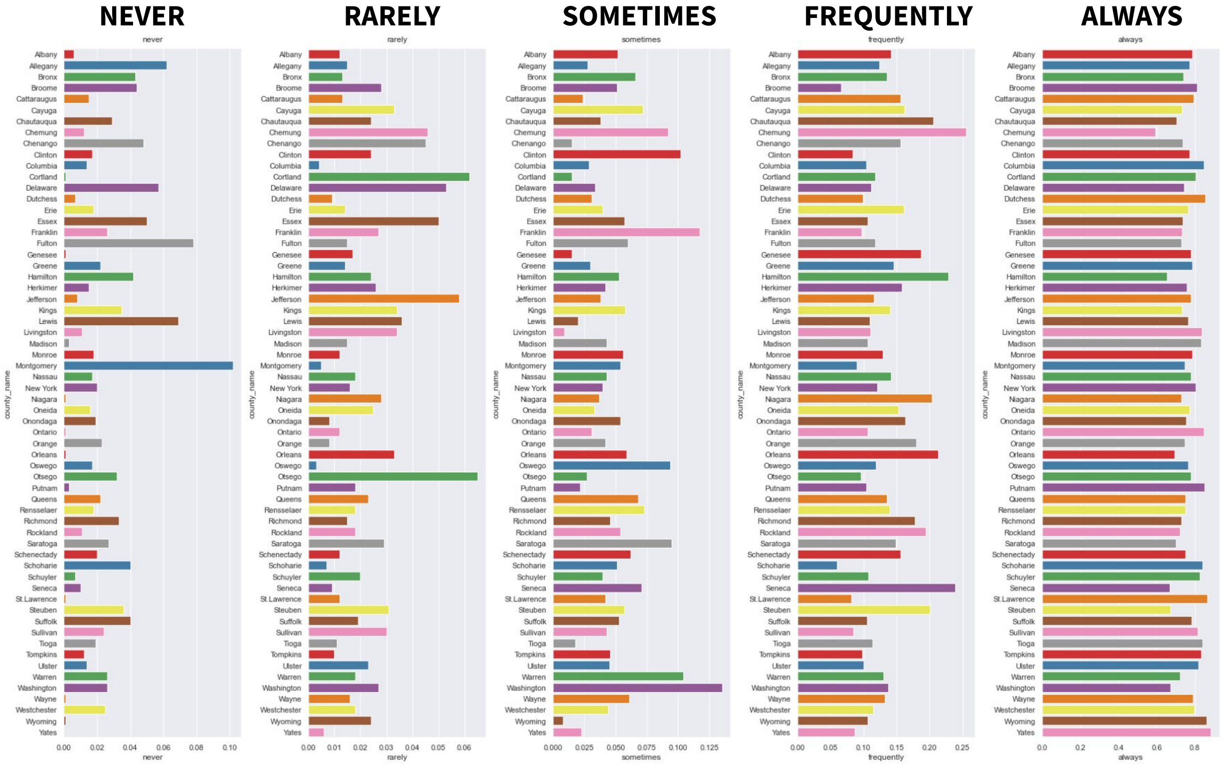
Let’s combine with a choropleth map.

2020 US COVID-19 Status by State



36.863k
North Carolina
Population: 10383620.0
Confirmed Cases: 36863
Deaths: 411
Mask_never: 0.05
Mask_rarely: 0.05
Mask_sometimes: 0.1
Mask_frequently: 0.2
Mask_always: 0.6
Deaths Per Case: 0.0111
Infection Rate: 0.0036
Deaths Per Population: 3.96e-05

Mask Use in the NY



NEVER

RARELY

SOMETIMES

FREQUENTLY

ALWAYS

The most **NEVER** use mask people are living in Montgomery county.

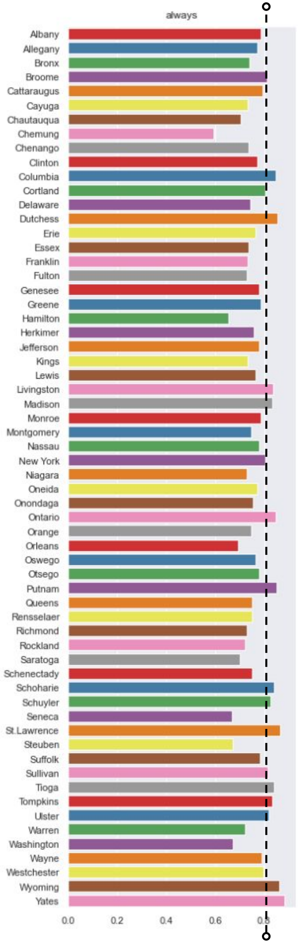
Chemung county gets the biggest number in **FREQUENTLY** for mask but not in **ALWAYS**.

In all of counties, Wyoming has the least people use mask in **SOMETIMES** & **FREQUENTLY.**

Yates gets the most in **ALWAYS**

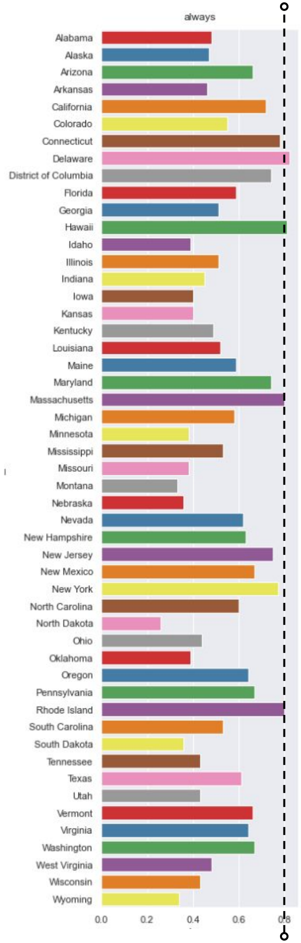
NY

SAY ALWAYS



VS

Based line 80%

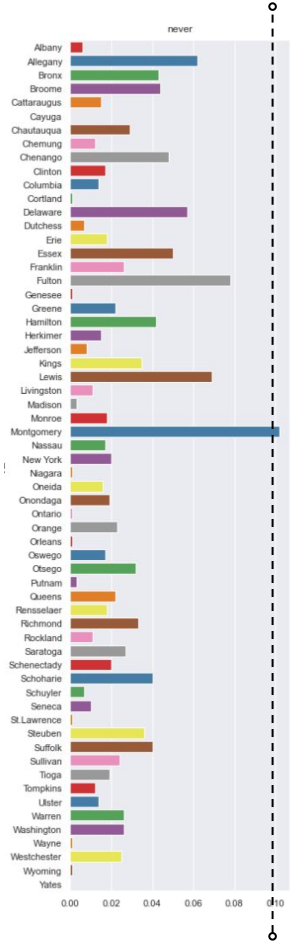


U.S.

SAY ALWAYS

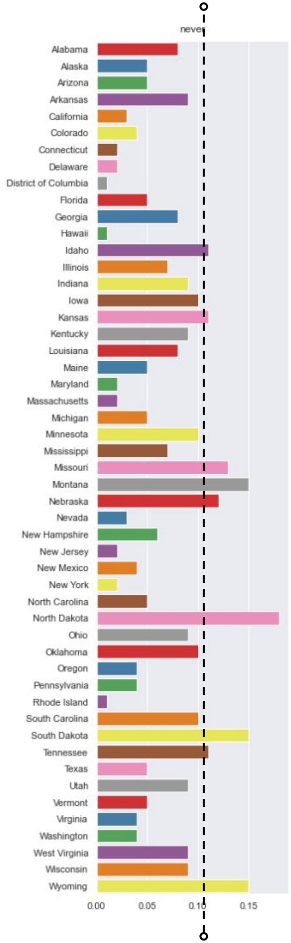
NY

SAY NEVER



VS

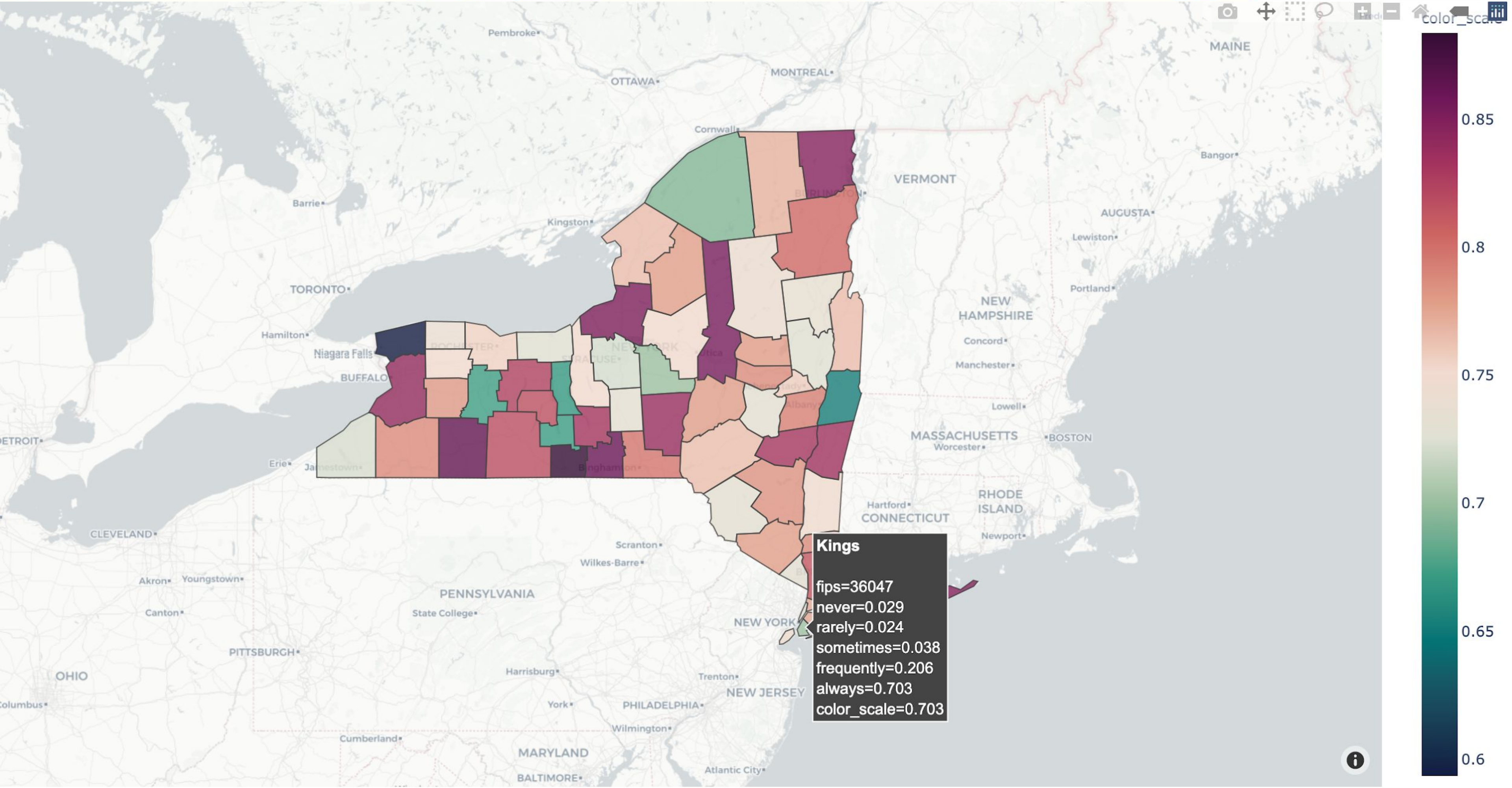
Based line 10%



U.S.

SAY NEVER

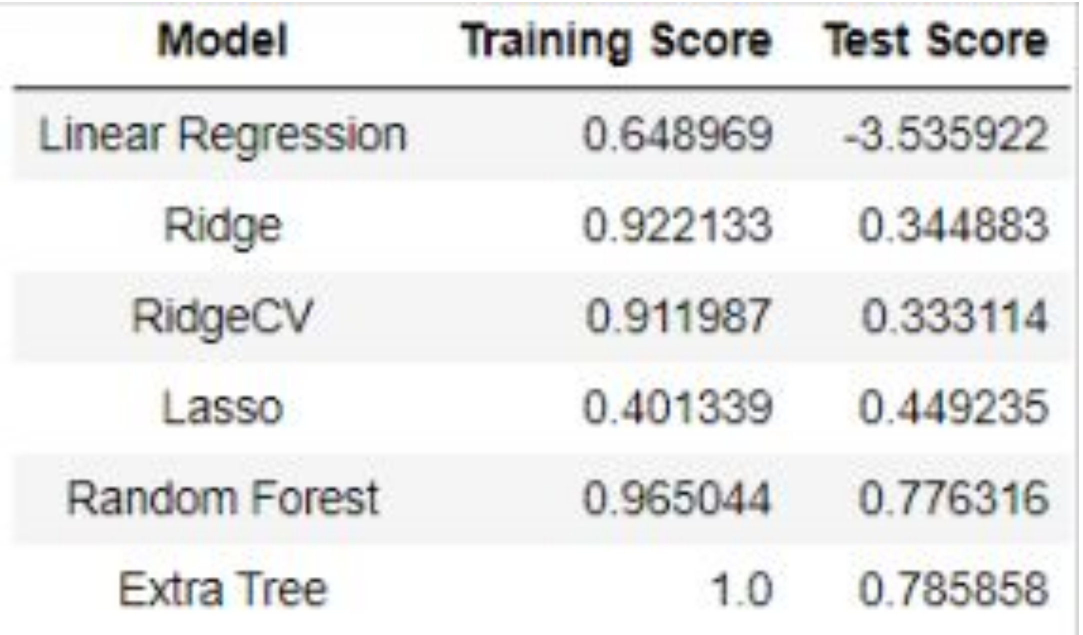
WELCOME TO NY



Kings

fips=36047
never=0.029
rarely=0.024
sometimes=0.038
frequently=0.206
always=0.703
color_scale=0.703

Modeling with Mask Data



Model	Training Score	Test Score
Linear Regression	0.648969	-3.535922
Ridge	0.922133	0.344883
RidgeCV	0.911987	0.333114
Lasso	0.401339	0.449235
Random Forest	0.965044	0.776316
Extra Tree	1.0	0.785858

Conclusions & Future Directions

- Random Forest Regressor -
    - R2 score on Testing Data: 98.9%
    - RMSE - 5,649 cases
- Symptoms Search-
    - Random Forest Model provided the best score (.96 CV score)
    - The most correlated symptoms seem to be behavioral symptoms
- Mask Model -
    - Our models displayed high variance
        - May be due to balance between high covid cases leading to high mask use and lower mask use leading to high covid case count (requires further study)
    - The data itself was a snapshot of mask usage based on a survey from July
- Future Direction:
    - Combine all the datasets into one model
    - Model using Time Series