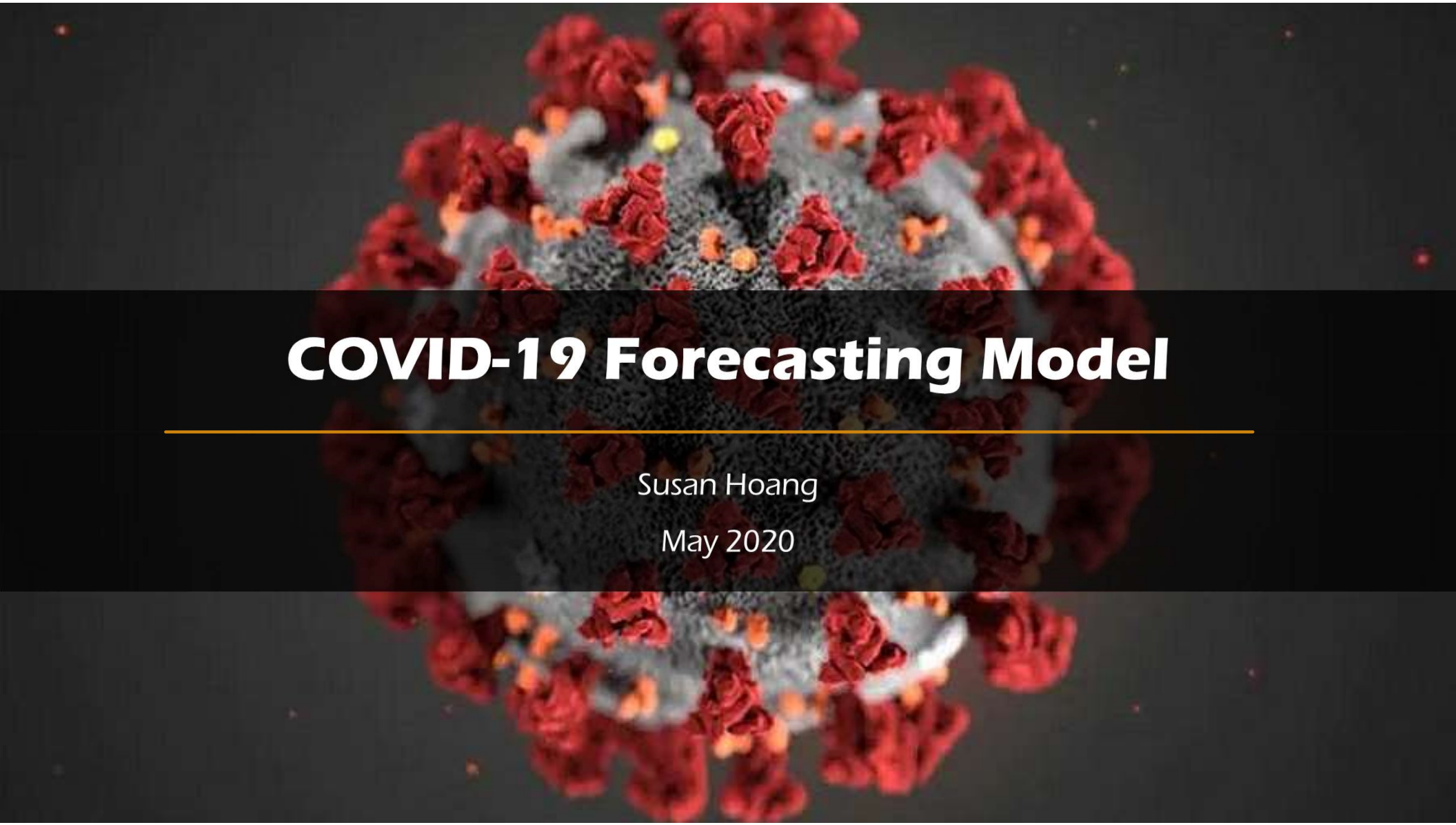


# COVID-19 Forecasting Model

Susan Hoang

May 2020

# Overview

- Business goal
- Data Science Framework
- Business Questions
- Analysis Plan
- Data Collection: Data Sources, Data Management Approach
- Known Data Issues and Proposed Solutions
- Key Preliminary Insights

# Business Goal

To aid recovery planning by understanding state-level COVID-19 variations and forecast Texas COVID-19 cases

# Business Question

How might we better understand Texas COVID-19 patterns to improve recovery planning?

- Business Question
- Analysis Plan
- Data Collection
- Insights
- Recommendation

## Current U.S. and Texas COVID-19 Pandemic Challenges:

- The U.S. has been challenged to pursue a successful containment strategy, and the impact of COVID-19 has been uneven throughout the country.
- While the temporary shutdown has been credited with slowing down the spread, it has cause significant negative impact to the economy.
- Texas, like other parts of the nation, is trying to balance economic recovery whilst preventing an outbreak resurgence.
- Widespread testing strategy is considered a key pillar to safely re-opening.

## Key Business Considerations:

- Access to timely and quality COVID data will be important to building a model.
- The COVID pandemic is unprecedented and fluid. It will be challenging to build a forecasting model based on dynamic data.

# Analysis Plan

To develop a forecasting model that predicts Texas' confirmed cases, deaths and test cases.

- Business Question
- **Analysis Plan**
- Data Collection
- Insights
- Recommendation

**Hypothesis:**

- From a recovery planning perspective, a 7-day forecast model that predicts the number of confirmed cases, deaths and tested cases will be more valuable than relying on conjecture and instinct.

**Data Required:** Historical state COVID-19 testing data

**Methodology:**

- Create a time-series forecasting model based on historical daily data and compare its accuracy to ground truth to assess its usability.

**Project Steps:**

- Evaluate historical data from April 1 through May 8, 2020
  - Compare and contrast patterns for California, Texas and New York state
- Build and test a 7-day forecasting model
- Test model with ground truth from May 9-15, 2020 to determine prediction accuracy.

# Analysis Plan

A time-series forecasting model to predict 7 days into the future may be useful for planning purposes

Business Question

Analysis Plan

Data Collection

Insights

Recommendation

**Forecast Model:** Prophet from Facebook (https://facebook.github.io/prophet/)

- Open source procedure for forecasting times series data
- It is robust to missing data and shifts in trends , and typically handles outliers well

**Data Source:**

- Historical daily data of Texas level COVID-19
- Date Range: April 1– May 8, 2020
- Records: 38
- Key Information Available:
  - State: Texas
  - Date: Daily
  - Confirmed Cases (Positive COVID-19 Diagnosis)
  - Death Cases
  - Tested (Accumulated tested cases)

**Key Consideration:** since COVID outbreak is a recent phenomenon, seasonality was not factored.

# Data Collection

Data Source: Historical state data will be leveraged to build the model

**Data Source:** The COVID Tracking Project

- Historical daily data of state level COVID-19; Downloaded 5/9/2020 (https://covidtracking.com/api)
- Date Range: April 1– May 8, 2020
- Records: 2750
- Key Information Available:
  - State
  - Date
  - Confirmed Cases (Positive COVID-19 Diagnosis)
  - Death Cases
  - Tested (Accumulated tested cases)

**Key Consideration:** since COVID outbreak occurred recently (mid-March), data is limited and quality is in flux. The start date of April was selected because data was incomplete prior to that time. With time, the model can be refined.

# Insights

Variation exists at the state level. Three large states with varying COVID impacts were evaluated.

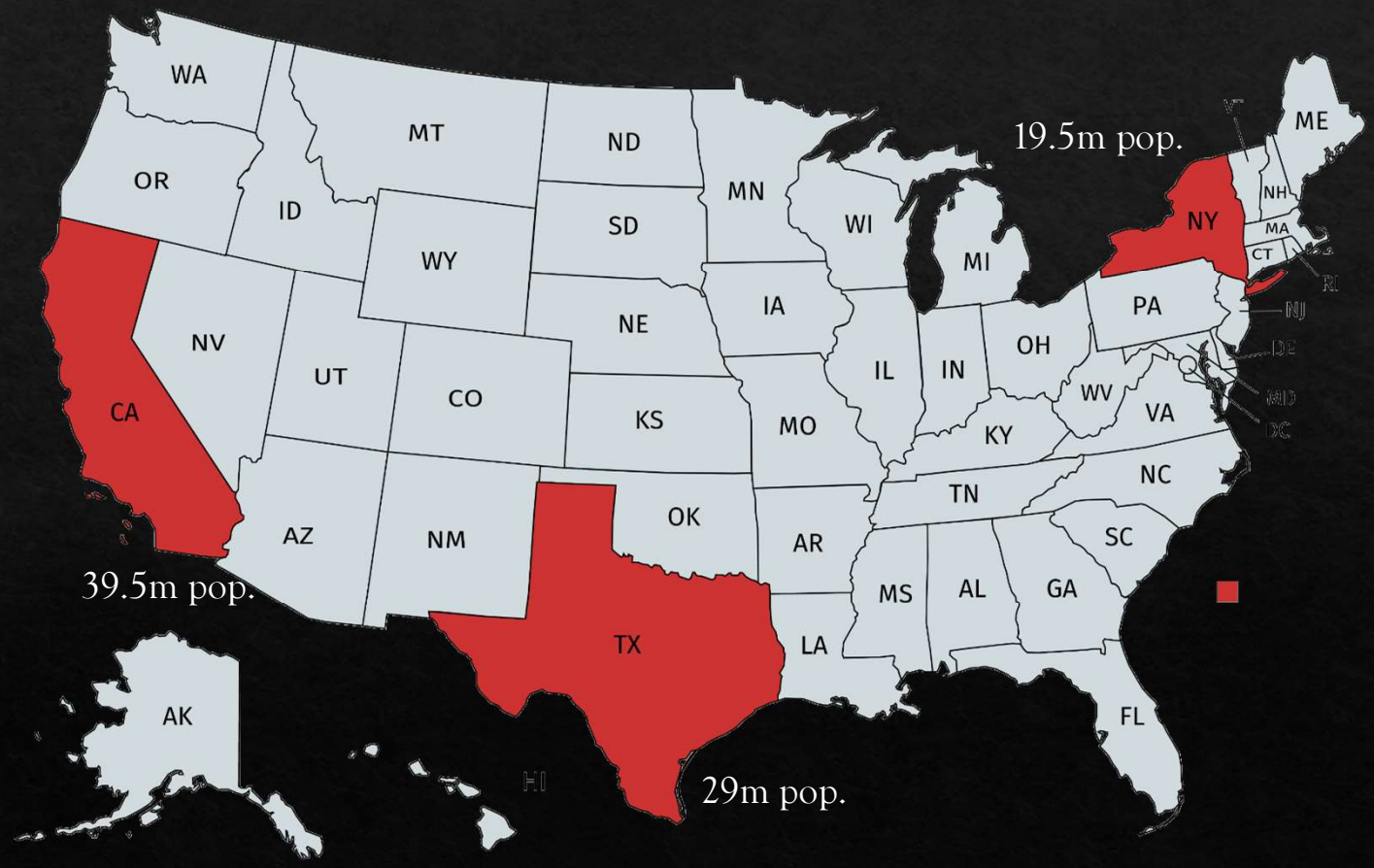
Business Question

Analysis Plan

Data Collection

Insights

Recommendation



# Insights

Variation exists at the state level. Three large states with varying COVID impacts were evaluated.

Business Question

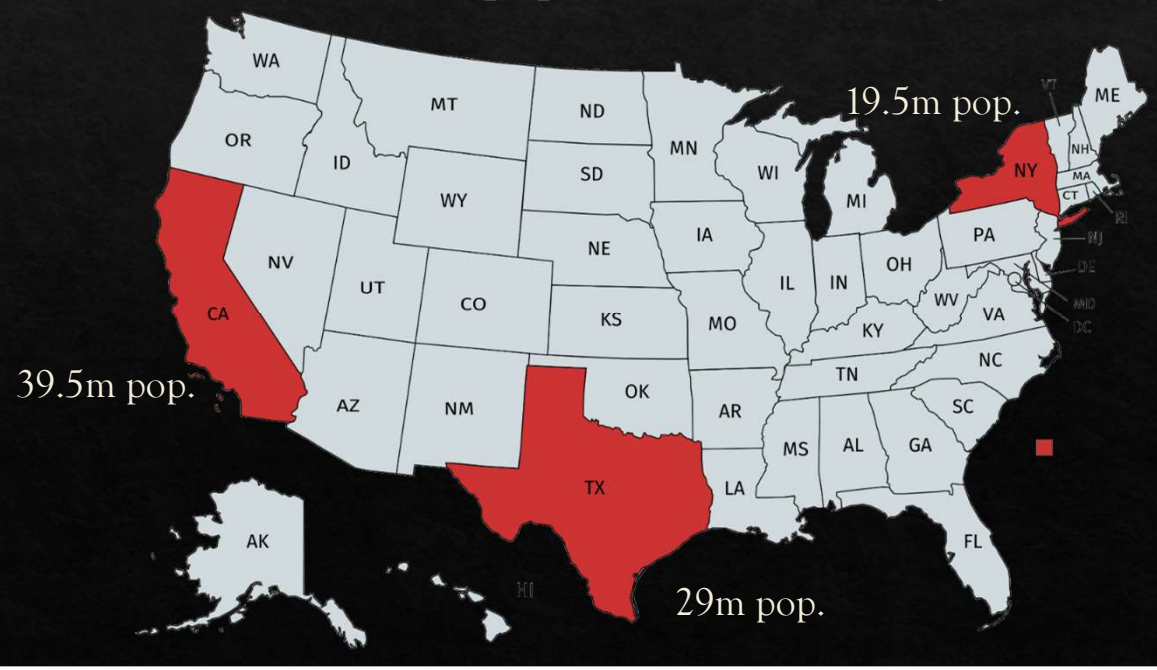
Analysis Plan

Data Collection

**Insights**

Recommendation

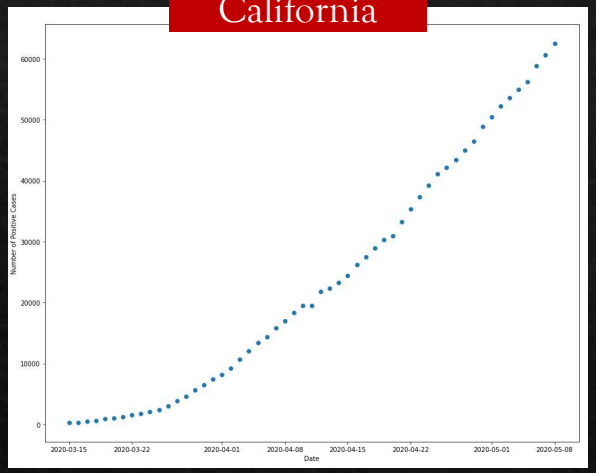
- New York had the highest positive cases, deaths and tested
- California had the highest negative cases relative to tested
- Texas had moderately high negative cases relative to tested
- California and Texas test rates vs population is relatively low



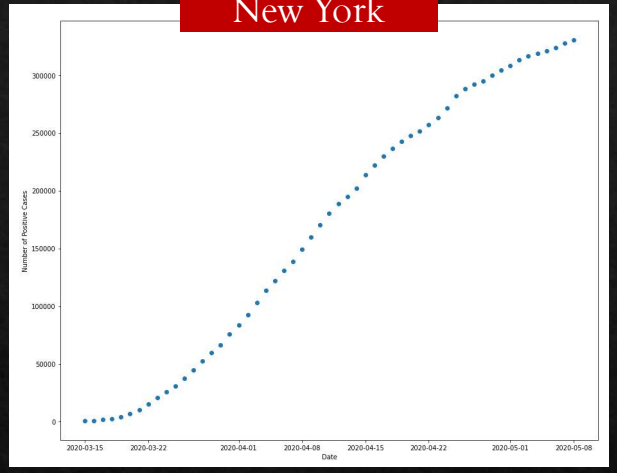
# Insights

## As of May 8th, only New York has seen flattening of the number of confirmed cases.

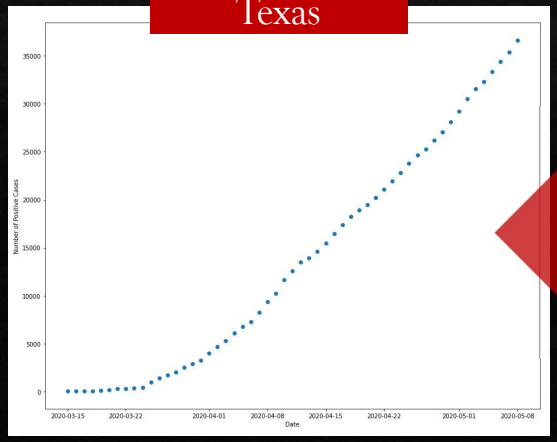
California



New York




Texas

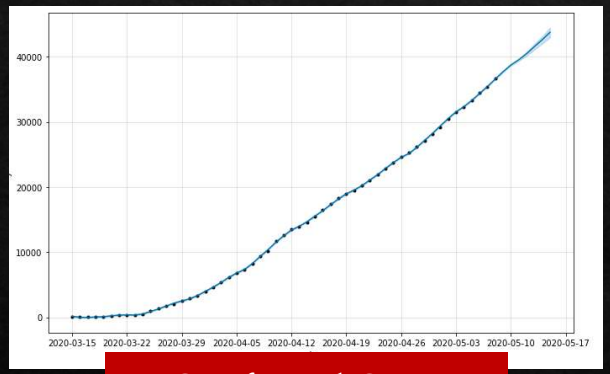



Texas confirmed cases continue to rise, so are deaths and tested cases as of May 8th.
The state began to re-open May 1st.

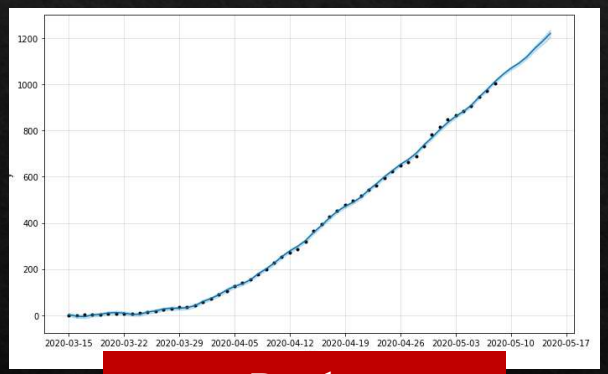
Date Range: 3/15-5/8/20
Number of Confirmed Cases per Day

# Insights

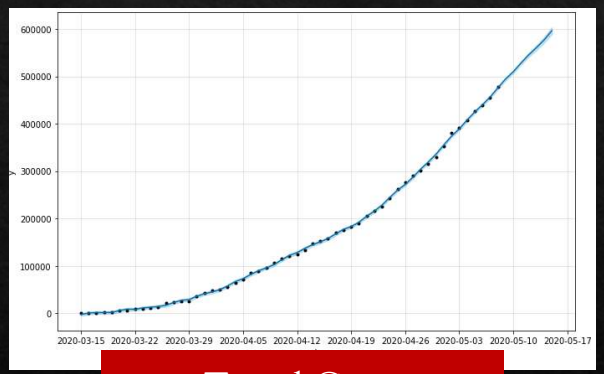
Forecast model project that Texas has not peaked in the number of cases




Confirmed Cases




Deaths




Tested Cases

# Insights

A time-series forecasting model to predict 7 days into the future may be useful for planning purposes

TEXAS COVID-19 CONFIRMED CASES

| DATE | ACTUAL | PREDICTED | LOWER LIMIT | UPPER LIMIT |
|---|---|---|---|---|
| 20200509 | 37860 | 37693 | 37589 | 37790 |
| 20200510 | 38869 | 38727 | 38600 | 38858 |
| 20200511 | 39869 | 39527 | 39330 | 39707 |
| 20200512 | 41048 | 40462 | 40144 | 40747 |
| 20200513 | 42403 | 41550 | 41070 | 41968 |
| 20200514 | 43851 | 42605 | 41989 | 43157 |
| 20200515 | 45198 | 43756 | 42942 | 44473 |

TEXAS COVID-19 RELATED DEATHS

| DATE | ACTUAL | PREDICTED | LOWER LIMIT | UPPER LIMIT |
|---|---|---|---|---|
| 20200509 | 1049 | 1043 | 1036 | 1050 |
| 20200510 | 1088 | 1069 | 1061 | 1077 |
| 20200511 | 1100 | 1091 | 1083 | 1099 |
| 20200512 | 1133 | 1118 | 1109 | 1127 |
| 20200513 | 1158 | 1154 | 1143 | 1168 |
| 20200514 | 1216 | 1186 | 1172 | 1120 |
| 20200515 | 1272 | 1221 | 1203 | 1238 |

TEXAS COVID-19 TESTED CASES

| DATE | ACTUAL | PREDICTED | LOWER LIMIT | UPPER LIMIT |
|---|---|---|---|---|
| 20200509 | 489294 | 49437 | 490958 | 497504 |
| 20200510 | 501776 | 509632 | 506205 | 512950 |
| 20200511 | 525697 | 528085 | 524338 | 531647 |
| 20200512 | 538172 | 545155 | 54943 | 549665 |
| 20200513 | 587431 | 560463 | 555060 | 565487 |
| 20200514 | 623284 | 576842 | 570688 | 583016 |
| 20200515 | 645992 | 596206 | 588879 | 60119 |

- The forecasting accuracy was best for deaths (table left).
- The lower level of accuracy for the forecasting model is expected given the dynamic nature of the COVID-19 pandemic and short timeframe of data.
  - This model is within the same range of prediction as expert models
- There were several mitigating factors that likely impacted the accuracy:
  - Texas data quality graded "B" per The COVID Tracking Project.
  - The testing and identification of COVID is influx.
  - Concerns that not enough tests have been completed in TX.

# Recommendation

Using forecasting models whilst being cautious about their limitations is better than using conjecture

## Conditions that impact forecasting models:

- **Public policies and people behaviors impact projections.**
  - EX: Texas began to re-open May 1, 2020. This likely will increase the number of confirmed cases and deaths in the future.
- **Building reliable COVID-19 models is challenging due to imperfect data.**
- **The variables will change in response to each other.**
  - EX: As testing continues to ramp up, expect the number of confirmed cases to go up.

# Summary

Forecasting models can help us better understand Texas COVID-19 patterns to inform recovery planning

- Variation exists at the state level. Texas confirmed, death and tested cases continue to rise.
- Public policies and people behaviors impact projections, Texas' re-opening will increase cases.
- This 7-day forecasting model may be useful for planning purposes, and is in range w/ expert models.
- Using forecasting models whilst being cautious about their limitations is better than using conjecture.

"COVID-19 models aren't made to be unquestioned oracles.
They're not trying to tell us one precise future, but rather the range of possibilities given the facts on the ground."
– FiveThirtyEight

https://projects.fivethirtyeight.com/covid-forecasts/; accessed 5/17/2019