# **Covid-19 Prediction with Comparison of Different Models**

Yue Wang

Department of Data Science, Brown University

DATA 1030: Hands-on Data Science

Professor Andras Zsom

Oct 12, 2021

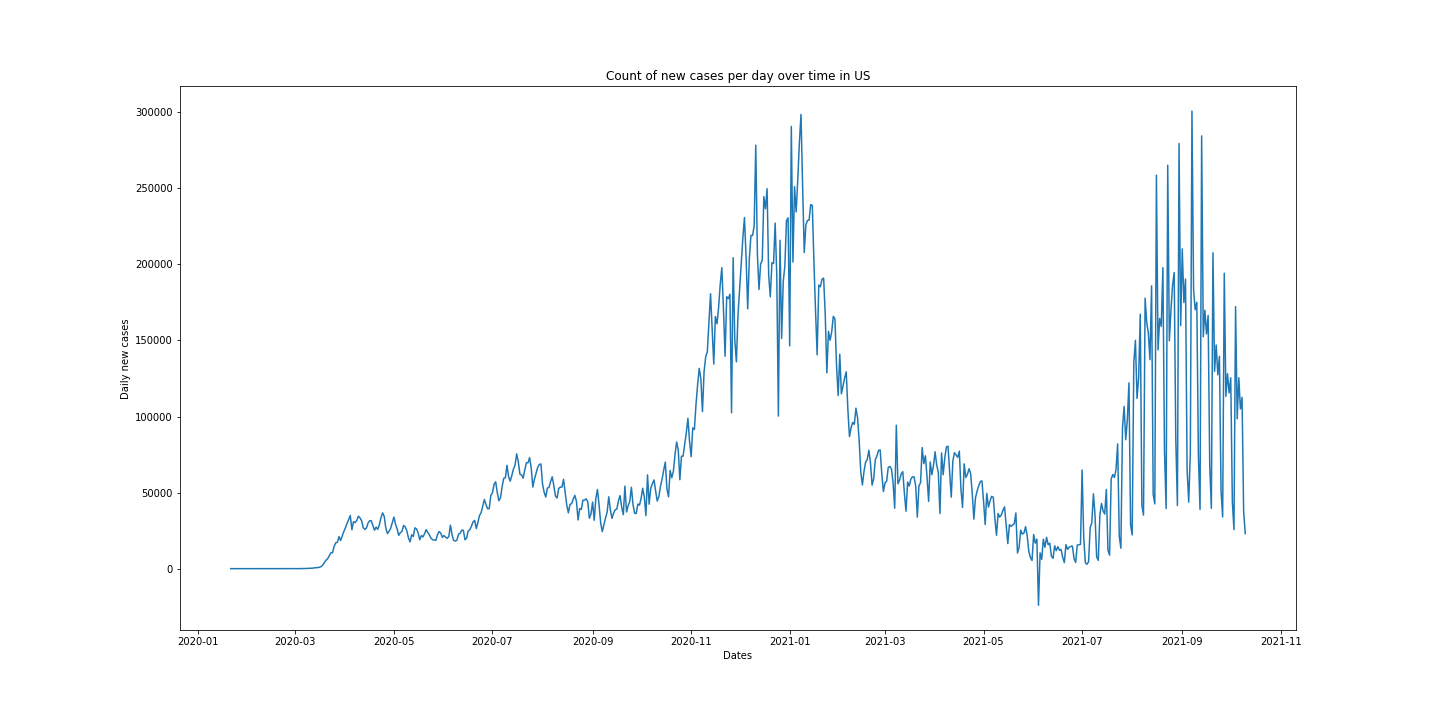
## Introduction

Covid-19 was first diagnosed and reported in December 2019, and surprisingly, continued to affect the world for almost two years with no signs of extinction so far. The prediction of Covid-19 cases seems to be an important way to measure the severity of it and to determine which protocols should be practiced and to what extent to contain it worldwide.

This project aims to predict the number of future daily Covid-19 cases by the previous daily counts of cases in the US with regression models. Data were maintained and renewed by Tableau, with two data sources. The US data was retrieved from New York Times along with other 1,929 countries, and the rest of the world’s data was from JHU CSSE Global Timeseries. The dataset has 13 columns which represent the cumulative number of people with positive Covid cases, county name, province/ state name, report date, continent name, data source name, the number of daily new deaths, county FIPs number, country alpha-3 code, country short name, country alpha-2 code, daily new positive cases, and cumulative deaths, ordered as the dataset. Each row provides the above info per day per county. There are 2,236,500 entries from 219 countries in total and this project mainly concentrates on the 2,008,440 rows representing the US only, which is still a large proportion of the data due to the lack of information on county or province in most of the other countries.

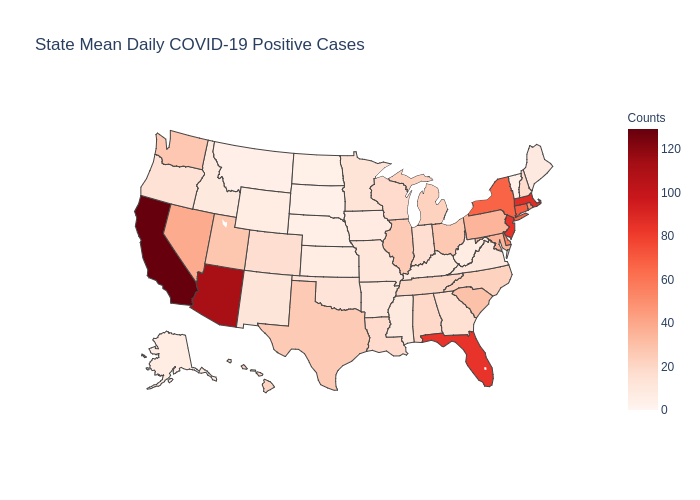
## Exploratory Data Analysis

Figure 1 shows the overall trend of the daily cases varying over time. It appears that there could be a seasonal correlation where cases increase slowly during winter and more rapidly during summer, but more data over at least 3 years would be needed to draw the conclusion. The first and only negative count appeared in June 2021.

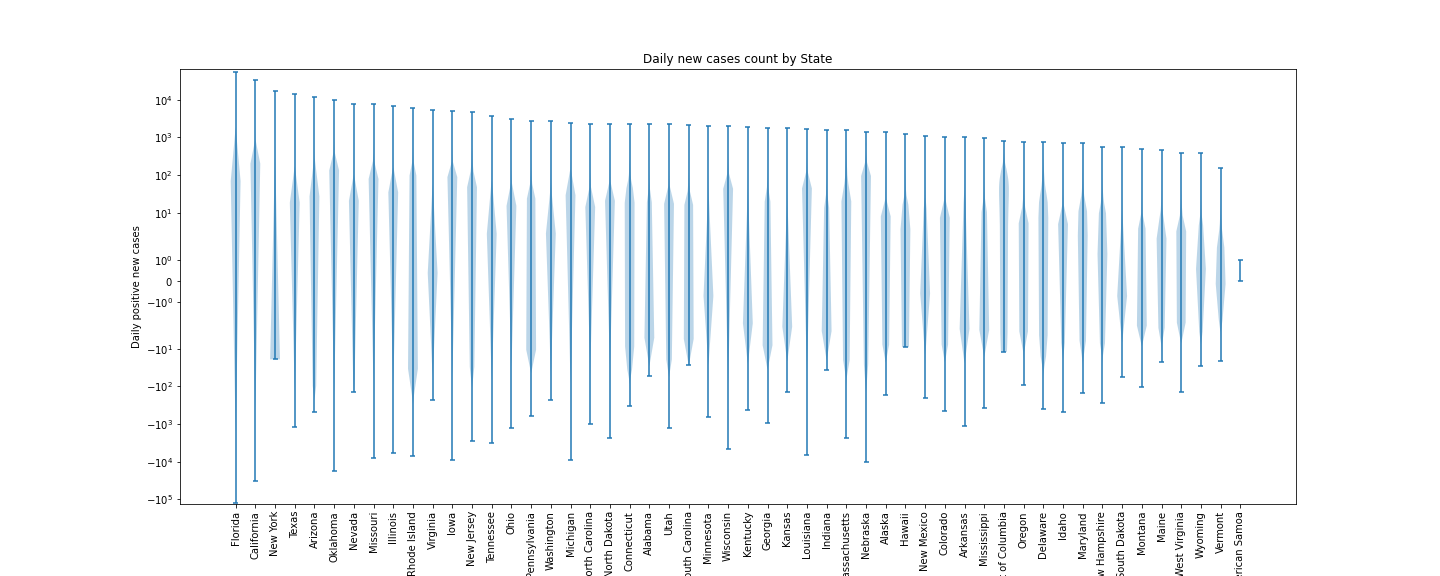


**Fig. 1** US Daily Count of Positive Cases vs Time

Figure 2 displays the average daily reports among different states in the US. California with the most population also reports the most daily cases on average. Whereas， Arizona falls behind it while being the 14th most populated state and the 33rd state ranking by the density of population (both from Census 2020). Fig. 3 presents the violin plot of daily cases grouped by different states. Symmetrical log scale was applied on both positive and negative sections of the y-axis to improve the readability of the graph. Higher variances are commonly exhibited in states with the most daily reported cases, alongside some outliers, such as New York.

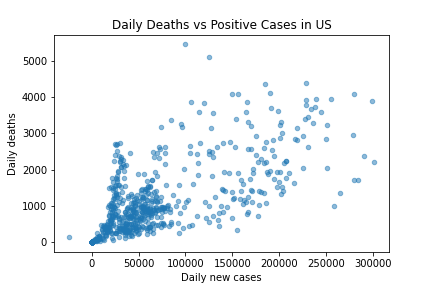


**Fig. 2** US Daily Count of Positive Cases by State



**Fig. 3** US Daily Count of Positive Cases by State Violin Plot

Figure 4 exhibits a positive correlation between the number of daily deaths and the number of daily cases via a scatter plot.



**Fig. 4** US Daily Deaths vs Daily Count of Positive Cases

## Methods

**Data Preprocessing**

The data being used has a time-series structure, so it is not i.i.d. However, 30 days were designed to be the look-back period, so 30 columns were created with the recorded number of new positive cases on the previous day, one day before, two days before… to be used as features in order to predict the target variable. The first 30 rows were dropped as a consequence of the inability to retrieve the Covid counts 30 days earlier than them. After feature engineering, this new model-ready data frame, with 30 feature variables and 1 target variable, is now i.i.d. since each row does not rely on other values besides its own fields. Additionally, the target variable is the count of daily cases of the US instead of states or counties, so this new data frame has no group structure either. Therefore, it’s safe to perform a train-test-split. 60% data was used as the training set, 20% as the validation set, and the last 20% was designed to be the testing set.

All columns of the model-ready data frame are continuous. As the last step of preprocessing, a standard scaler was fit on the features of the training set and transformed on the feature variables of all three sets.

## References

Tableau. (2020, October12). *Covid-19 (coronavirus) Data Resource Hub.* Tableau. <https://www.tableau.com/covid-19-coronavirus-data-resources>

Liebeskind, M. (2020, April 22). *5 Machine Learning Techniques for Sales Forecasting*. Towards Data Science. <https://towardsdatascience.com/5-machine-learning-techniques-for-sales-forecasting-598e4984b109>