COVID-19: Report

Modeling COVID-19 growth in California using Logistic and Exponential models and Clustering California counties using K-Means and Agglomerative methods

Authors: Abhishek Roy, Gayatri Babel, Suyash Jaju

Table Of Contents

---

1. Abstract

As the world toils to battle the COVID-19 pandemic, the US has taken a massive brunt of the impact due to the staggering number of cases and deaths. Though California has handled the pandemic well, there's news that the state is planning to reopen soon. But is that the right move, especially with cases touching north of 65,000 and still rising? Are some counties still at risk? Are some regions in the state truly doing better than other parts?

These ideas prompted us to conduct research on various demographic features along with data on changing incidence and mortality rates of each county in California and analyze their performance. Specifically, we wanted to achieve two overarching goals:

- Cluster the counties into Red, Orange, and Green zones to see which counties are still at risk at which ones face relatively less danger
- Using Regression models, find a way to predict when the incidence and mortality rates will start to plateau in high-risk counties and assess whether it is safe to ease lockdown restrictions.

2. Experimental Question and Hypothesis

2.1 Experimental Questions

Based on the background research conducted regarding the latest data obtained from JHU's Github resource, we derived a few questions we hope to address through our data analysis (CSSEGISandData). The questions can be broken down between two geographic tiers:

- State-level
  - What features can be utilized to measure the risk posed by the pandemic in different counties?
  - How can this methodology help with grouping counties into three main risk levels?

- County-level
  - What kind of model will result in the best accuracy to predict new COVID-19 cases and deaths for a given county?
  - How can we build an adaptable model that can be applied to other counties to determine the level of risk associated with restarting the local economy?

2.2 Our Rationale

Several countries, such as India, have divided cities and districts into Red, Orange, and Green zones (Red being high-risk, Orange being medium-risk, Green being low-risk). This aids in monitoring specific areas and taking surgical measures in each area accordingly. However, zoning goes beyond only accounting for cases, deaths, and testing. Instead, we approached this problem using features to measure the susceptibility and resistance of the county's population to the virus.

At a county-level, we plan to model the confirmed cases and death coronavirus trends on the time series dataset and observe the effects of the shelter-in-place order to flatten the curve. This methodology will assist in providing recommendations regarding the reopening of counties.

2.3 Our Hypothesis

By analyzing data on California counties, we hope to conduct risk analysis and cluster California counties into three different risk zones: Red (high risk), Orange (medium risk), and Green (low risk). In parallel, we will develop and test a variety of models in order to determine a predictive model that can be applied to counties to project the number of confirmed cases and deaths due to COVID-19. Our hypothesis is that California should extend the lockdown for a longer period since there are still some counties that have a high-risk level and

cases in the counties will still take time to plateau. Therefore opening up might increase cases drastically. We hypothesize that a logistic model would fit best because of its sigmoid shape and since California has been faring well, we expect less Red zones and more Orange and Green zones.

3. Exploratory Data Analysis and Visualizations

Given the large amount of COVID-19 available, a strong strategy was required to effectively explore all relevant features. Coronavirus is known to target the lower respiratory system of individuals, hence, we mapped the respiratory mortality rate for each county in California. *Figure 3a* exhibits the resulting hexbin plot of California counties' varying respiratory mortality rates.
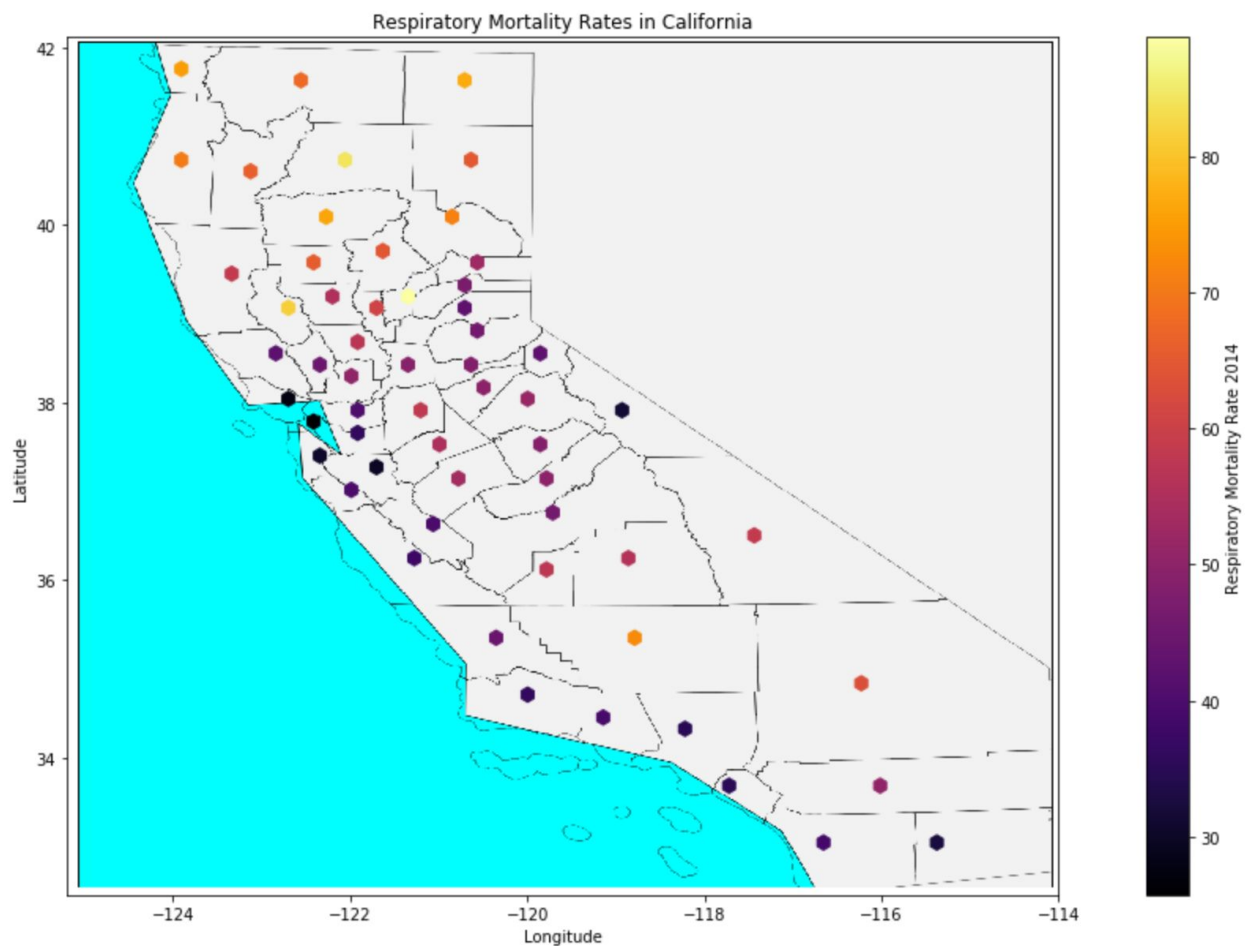

Figure 3a: Hexbin plot of respiratory mortality rates of every county in California.

Figure 3a indicates that for counties in California, health factors such as Respiratory Mortality occur in geographical clusters. This is particularly intriguing, as respiratory mortality is a strong potential to help differentiate low-risk, medium-risk, and high-risk zones for COVID-19 as a feature.

Furthermore, we created correlation heatmaps to better understand the correlation between other mortality features and the number of COVID-19 confirmed cases and deaths, as seen in *Figure 3b* and *Figure 3c*.
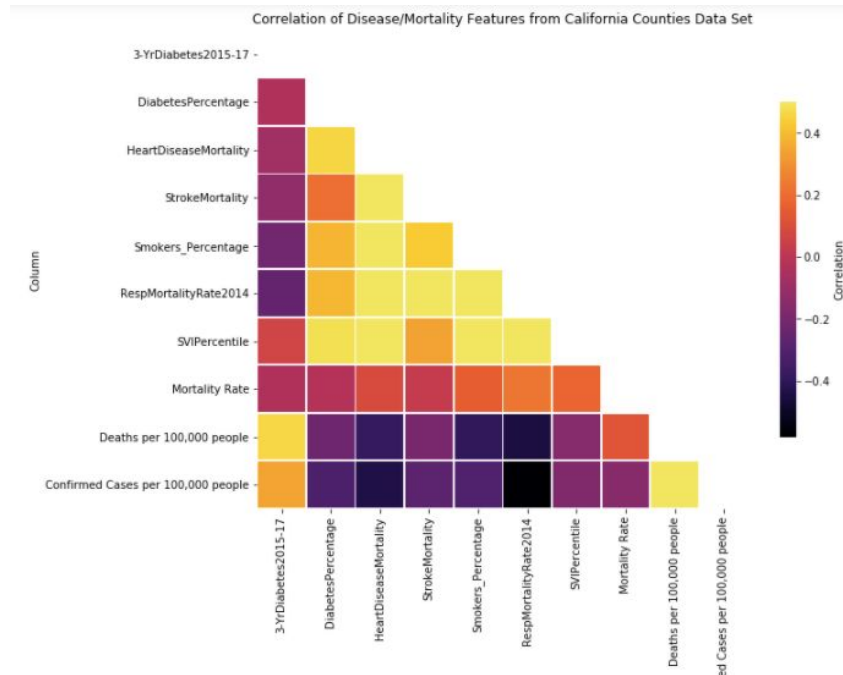

Figure 3b: Correlation Heatmap of Mortality Features and SVIP.

The plot above illustrates a strong correlation between the number of deaths, the number of confirmed cases, and 3-Yr Diabetes data (estimated percentage of people with diabetes). Additionally, it was observed that SVIPercentile is strongly correlated with diabetes percentage, heart disease mortality, stroke mortality, and the percentage of smokers. In other words, there is a greater social vulnerability in counties with higher mortality rates, greater percentage of people with diabetes, and a higher % of smokers. Thus, in order to avoid overfitting, only SVIPercentile served as a feature as an indicator of counties with higher mortality rates due to stroke, heart disease, and smoking, ensuring the models are applicable across counties.

4. Zone Clustering

4.1 Clustering Features

The EDA process, described in the previous section, provided insight into which features would serve as the strongest indicators of a county's ability to control the spread of the virus. It was found that hospital-related features, such as hospitals per 100,000 people, were expected to provide a strong measure of a county's ability to mitigate the impact of coronavirus. However, the EDA process found that this was false. For the metric regarding hospitals per 100,000 people, all county hospitals are not designated as COVID-19 treatment centers. On a higher-level, several hospital features served as similar metrics and thus, provided limited variance.

Specifically, this was the case for features People per MD and People per ICU Bed. Thus, only one hospital-related feature was included for clustering.

The final clustering features maintained a general trend: the higher the feature's value, the more risk the virus poses to the county. These particular features were found to serve as the best markers for a county's COVID-19 risk level:

Mortality Rate: Calculated as Number of people died/Population of County in last two weeks
Incidence Rate: Calculated as Number of new cases/Population of County in last two weeks
 # People per ICU Bed: Calculated as number of people per ICU Bed
RespMortality: Used given data - deaths per 100,000
Mortality25-44%: Used Mortality rate data for 3 age brackets
SVIP: Used given Social Vulnerability Index Percentile data

4.2 Clustering Methods

Clustering can be executed and visualized in two dimensions. Thus, after conducting dimensionality reduction through singular value decomposition and selecting the first two principal components, a 6-dimensional feature matrix with 2 dimensions was approximated. The scree plot below illustrates the variance captured by the first two principal components:
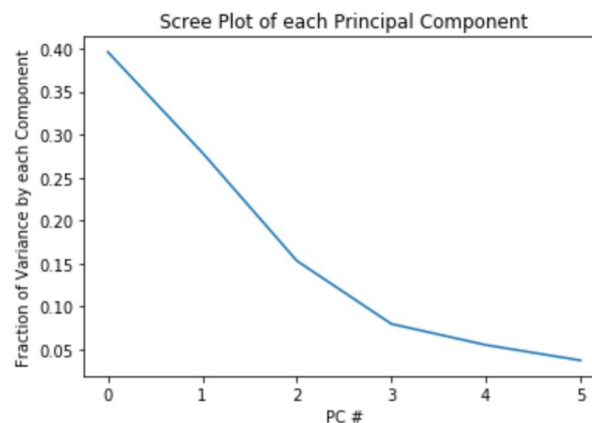

Figure 4.2a: Scree Plot of Variance captured by individual Principal Components

4.2.1 K-Means

The first attempt at visualizing the three zones involved using K-Means clustering. Due to the nature of the general plot, it was believed that grouping the points in roughly spherical groups would yield suitable zones.
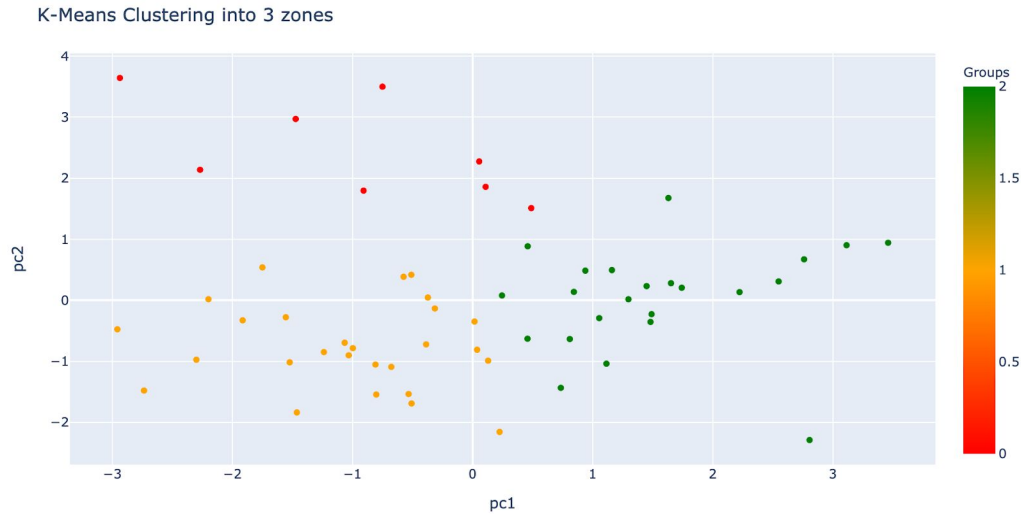
K-Means Clustering into 3 zones

Figure 4.2.1a: K-Means Clustered counties in Red, Orange, Green zones

This methodology yielded three clear zones (shown in the figure above):
- ***Red Zone*** contains large, high-risk counties like Los Angeles, Riverside, Santa Barbara, etc.
- ***Orange Zone*** consists of lower risk counties like Orange, Marin, Sonoma, Napa, San Luis Obispo, San Francisco, San Diego, Santa Cruz, Monterey, etc.
- ***Green Zone*** consists of even smaller, very low-risk counties like Yuba, Lake, Sutter, Shasta, Humboldt, Kern, Siskiyou, etc.

It was deduced that as the graph progresses from left to right, the risk of the county decreases. For example, high-risk zones, like LA, have a low PC1, high PC2 while low-risk zones like Yuba and Sutter tend to have lower PC2 and higher PC1. Furthermore, through this method, some medium-sized counties are classified as high-risk zones, such as Imperial, Tulare, and Kings.

However, there are some inconsistencies found in the clustering. For example, though Fresno and San Bernardino are near each other, they are classified as two different extremes: Fresno is Green and San Bernando is Red. This indicates that this clustering procedure is not optimal and implementing another could provide more continuity in the progression of counties from Red to Orange to Green.

4.2.2 Agglomerative

The second clustering attempt involved Gaussian Mixture and Spectral clustering. Spectral clusters were not ideal as it produced too many green-zoned counties while Gaussian Mixture yielded very similar groups as Agglomerative. Hence, the hierarchical clustering method, Agglomerative, was chosen. The resulting clusters are displayed below:
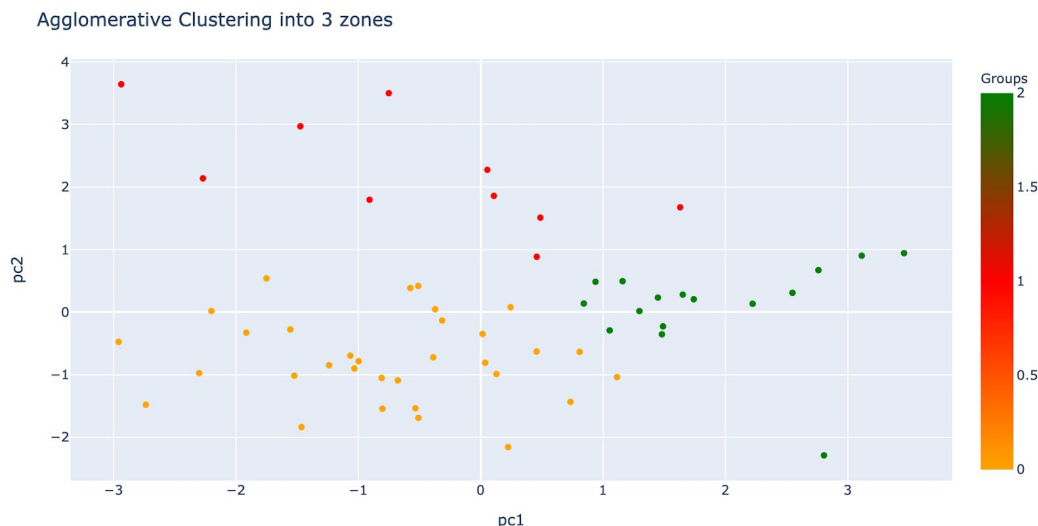
Agglomerative Clustering into 3 zones

Figure 4.2.2a: Agglomerative Clustered counties in Red, Orange, Green zones

The new clusters are similar, as Fresno and Kern, previously Green zones, are now classified as Red zones. However, it is evident that this is not the superior model as it does not align with the intuition that the higher the PC1, the safer the county as though Fresno is located in between, Kern is situated too far right on the graph to be classified as a red zone.

4.2.3 Clustering Model Evaluation

The **K-Means model** was superior to the Agglomerative model as it is better for a green zone to be classified as a red than vice versa. Furthermore, a K-Means clustering model provides a more accurate representation of zones that are high and low risk, prioritizing zones that are at risk rather than the size of the county. This rationale is supported by several news sources, cited in the Jupyter Notebook.

The **K-Means model** provides a good representation of clustering California counties most impacted by COVID-19, particularly considering factors, such as the ICU beds per person, respiratory mortality rates, general 3-year mortality rates, and social vulnerability percentiles.

4.3 Clustering Assumptions and Limitations

Assumption: Selected features, including age-related mortality rates, respiratory disease mortality, and information regarding ICU bed reflect the counties currently.

Limitation: The mortality and incidence rates are calculated from the past two weeks thus it is assumed that these rates will be similar, if not the same. Features do not account for future predictions, testing, and recovery rates for the counties.

5. Modeling

In order to find a model to predict the projected number of confirmed cases and deaths due to Covid-19, it is necessary to first evaluate the current data. *Figure 5a and 5c* display the graphs of the confirmed number of Covid-19 cases and Covid-19 deaths in California counties were similar in that both the number of infected individuals and deaths due to Covid-19 were increasing; however, it was difficult to determine if counties' confirmed cases and deaths are growing larger exponentially or linearly. In order to determine this, the logarithmic graphs of confirmed Covid-19 cases and Covid-19 deaths in each county were visualized, as seen in *Figure 5b* and *Figure 5d*. The logarithmic graphs revealed that, for most counties, the Covid-19 is spreading at an exponential rate; however, it is approaching a plateau. These findings indicate that either an exponential model or logistic model is best suited to predict the spread of Covid-19 and the corresponding deaths.

In order to test the accuracy of each model, the data on the Covid-19 confirmed cases and the Covid-19 deaths in SF were split into train data (all data until April 30th) and test data (all data beyond May 1st). Once each model produced predictions for confirmed cases and deaths, the root-mean-squared-error (RMSE) of the predictions and test data to determine which model is most accurate.
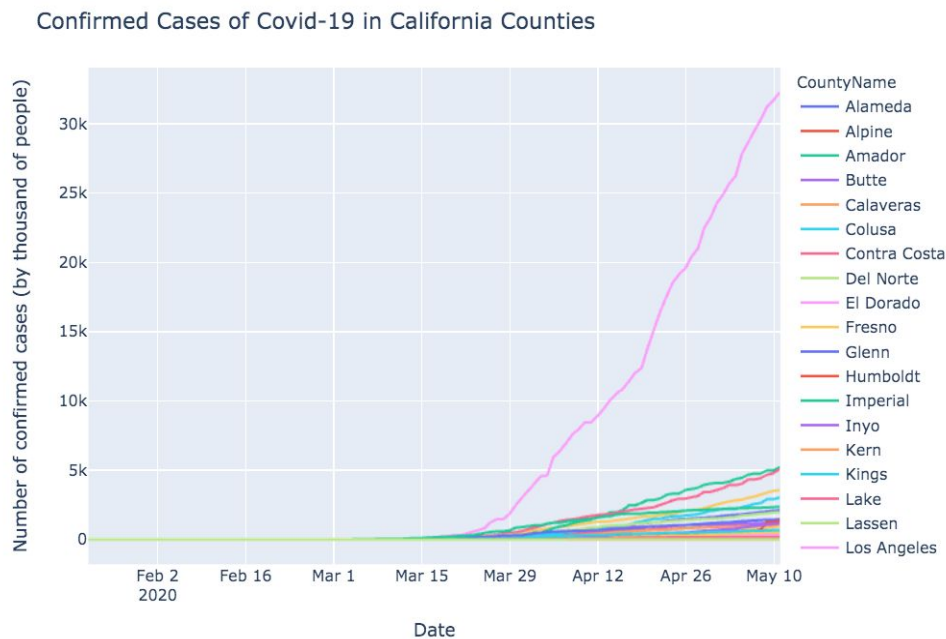

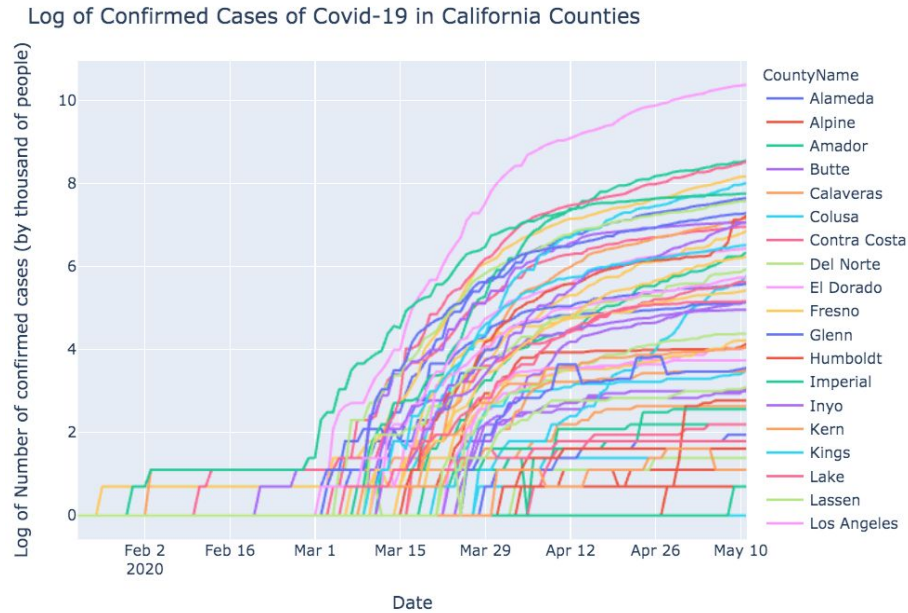*Figure 5a: Graph of Confirmed Covid-19 Cases in California Counties*

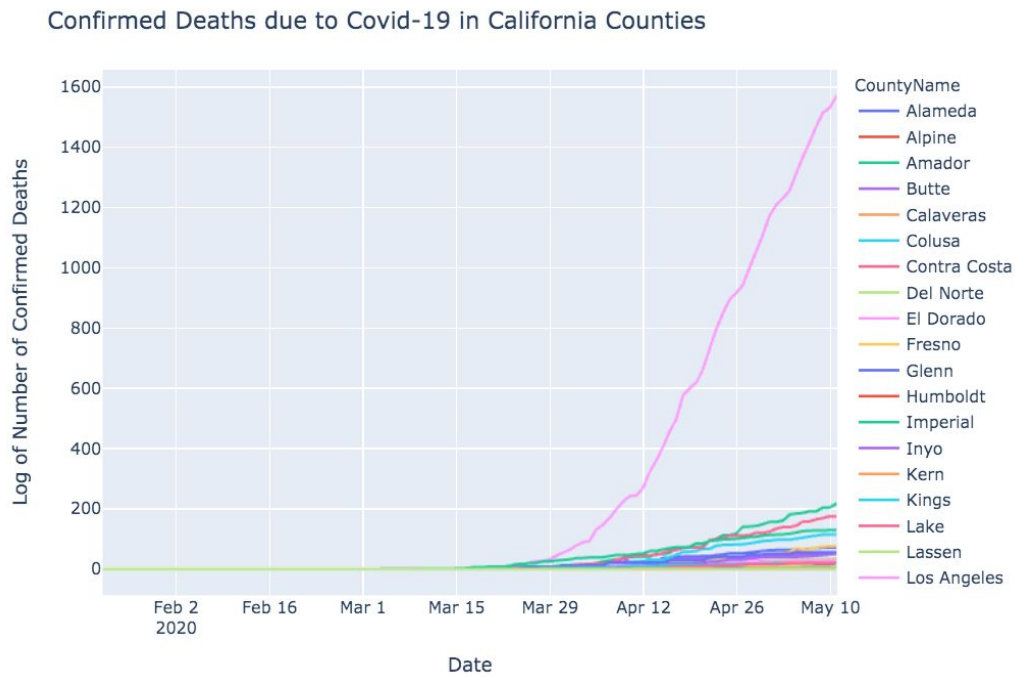*Figure 5b: Graph of Log of Confirmed Covid-19 Cases in California Counties*

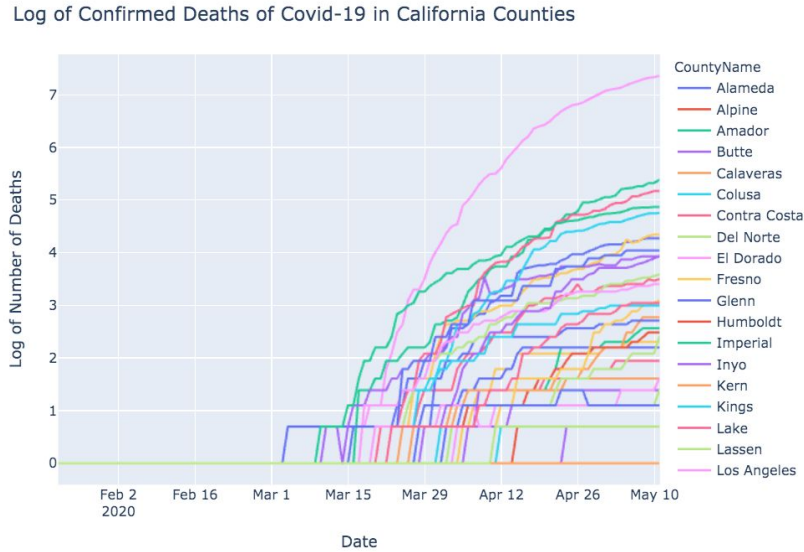*Figure 5c: Graph of Confirmed Covid-19 Deaths in California Counties*

Figure 5d: Graph of Log of Confirmed Covid-19 Deaths in California Counties

5.1 Exponential Model

The graphs of Covid-19 cases and deaths illustrated that during the initial infection period, the virus spreads at an exponential rate, thus an exponential model was explored as a potential model for prediction. While researching, it was found that there are two methods to approach the exponential model development: Scikit-Learn and a function.
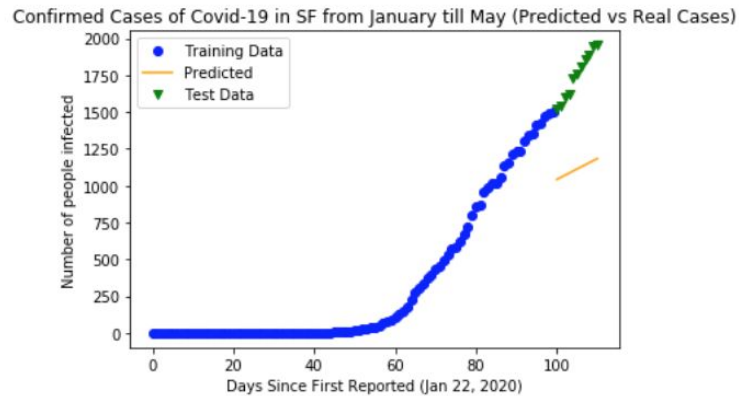

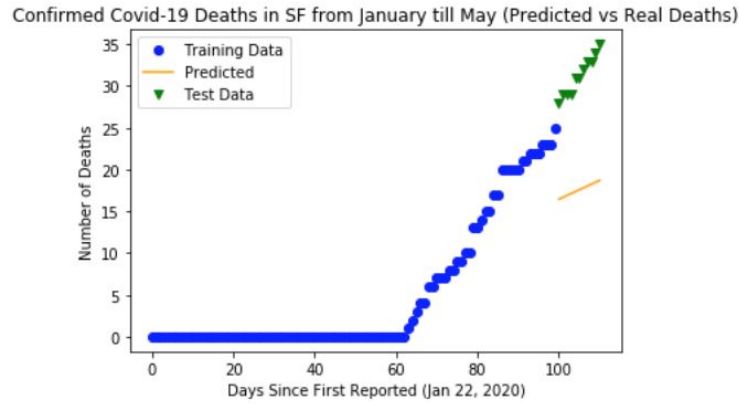*Figure 5.1a: Scikit-Learn Exponential Model with a RMSE of 642.74 cases*

Figure 5.1b: Scikit-Learn Exponential Model with a RMSE of 13.75 deaths

The first model was developed utilizing the linear regression function from Scikit-Learn. However, the corresponding RMSE error for the predicted number of confirmed cases was approximately 642 cases using this model. Thus, the linear regression model was determined as not suitable to predict the spread of Covid-19 in California counties. The graphs produced by this model for both predicted cases and deaths can be seen above in *Figure 5.1a* and *Figure 5.1b*.
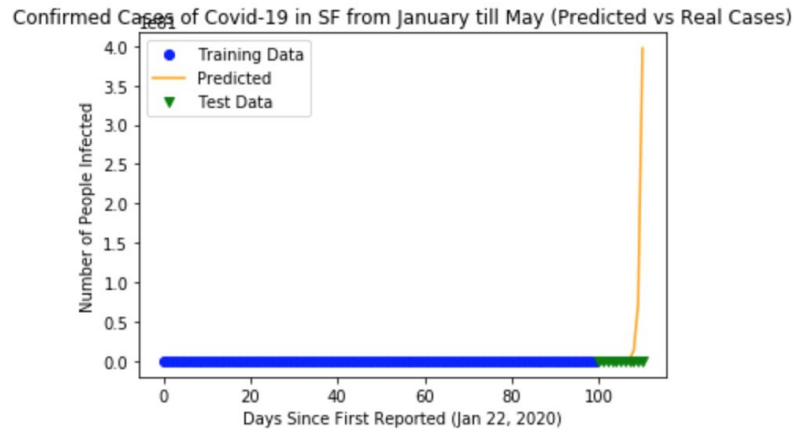

Figure 5.1c: Exponential Function Model with an RMSE of $1.22 * 10^{22}$ *cases*
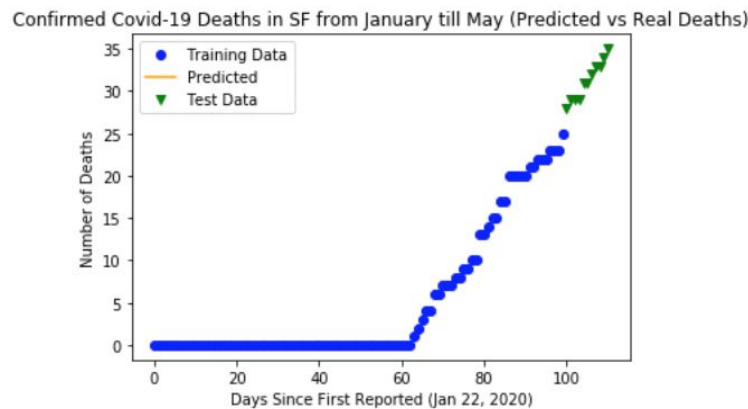

Figure 5.1d: Exponential Function Model with an RMSE of infinity

To obtain an exponential model with a better fit, an exponential function was utilized to fit the training data in order to predict the number of confirmed cases and deaths for the month of May, but the error for this model was much larger. For example, the error for the predicted number of deaths was infinity. Hence, the exponential function model is not an accurate prediction of confirmed cases and deaths.

The testing of exponential models revealed that an exponential model was not the best model to accurately predict confirmed coronavirus cases and deaths. Hence, the feasibility of utilizing a logistic model was explored.

5.2 Logistic Model

Though the graphs of confirmed cases and deaths in California exhibited an exponential growth rate, an exponential model was not apt in predicting the number of infected people and deaths due to coronavirus. This is primarily due to the fact that though infections spread at an exponential rate, it does not continue to spread at this rate forever. Thus, after exploring exponential models and consulting research papers, a logistic model was developed. Similar to how the exponential model was tested in two different manners, through Scikit-Learn and a function, the logistic model was tested.
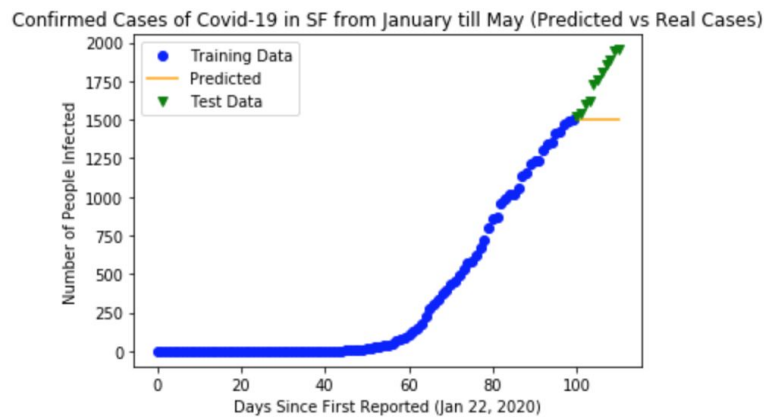

Figure 5.2a: Scikit-Learn Logistic Model with an RMSE of 289.7 cases
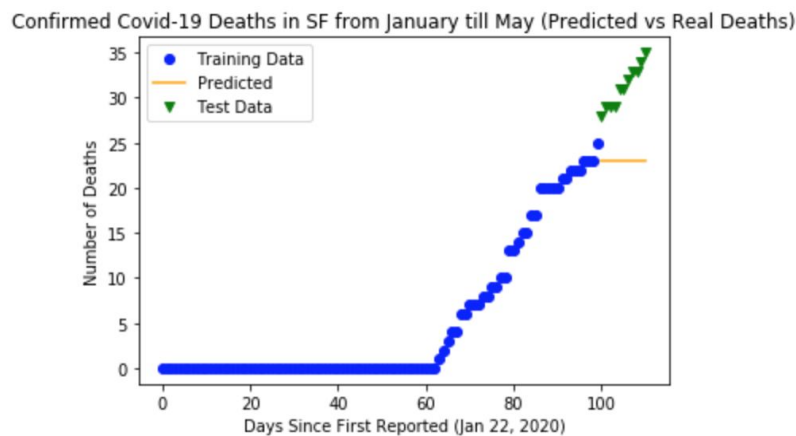

Figure 5.2b: Scikit-Learn Logistic Model with an RMSE of 8.6 deaths

The first logistic model was developed using the Scikit-Learn package, using the Logistic Regression function. This model was found to have performed much better than previous exponential models with an RMSE of nearly 290 for the confirmed cases and an RMSE of approximately 8 deaths. However, a second model fitted from a function was also tested in order to attain a smaller error.
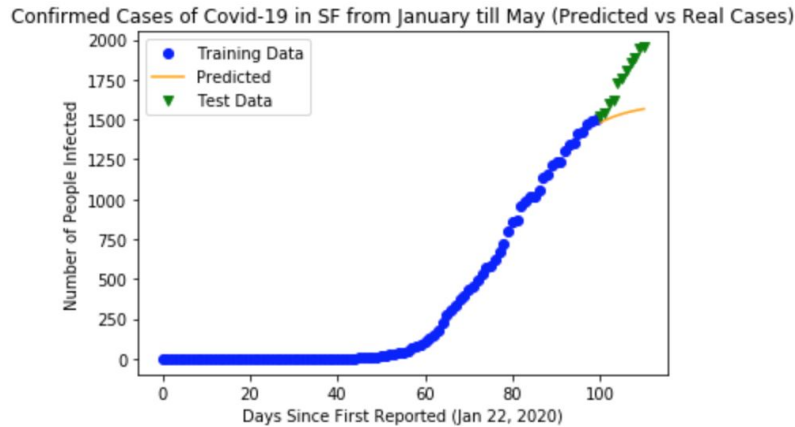

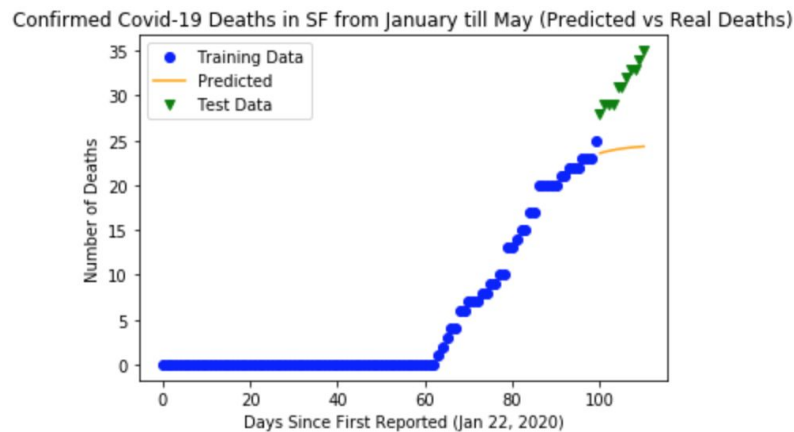*Figure 5.2c: Logistic Model Function with an RMSE of 249 cases*

Figure 5.2d: Logistic Model Function with an RMSE of 7.5 deaths

The second logistic model was developed using an equation function in order to better fit the data. This is evident by the fact that this model had lower RMSEs in comparison to the first logistic model, as well as the other two exponential functions. It is for this reason that the logistic model by function was chosen as best suited to predict confirmed cases and deaths due to coronavirus.

5.3 Modeling on Three Risk Zones

Once the best model to predict the number of infected individuals and deaths due to Covid-19 was chosen, the model was utilized to predict the number of coronavirus cases and coronavirus deaths within the next month to better understand the applicability of this model to counties within each of the three risk zones. Insights from these predictions would be used to recommend shelter-in-place orders for the next month.

5.3.1 High-Risk Zone: Los Angeles

From the high-risk zone cluster, Los Angeles was chosen to fit the logistic model in order to predict the number of infected individuals and deaths that will occur by mid-June. According to *Figure 5.3.1a* and *Figure 5.3.1c*, Los Angeles is leading in the number of confirmed coronavirus cases and deaths in the entire state of California. Overall, it was found that the logistic model was a good fit for counties in high-risk zones. For Los Angeles specifically, under the current shelter-in-place order, it was found that by mid-June, the number of coronavirus cases and deaths would just be approaching a plateau, seen below in the figures. Thus, it is recommended that the shelter-in-place order continue to be strictly enforced in high-risk counties like Los Angeles to limit the spread of coronavirus within the county.

Figure 5.3.1a: Predicted Covid-19 cases in Los Angeles until June 13th
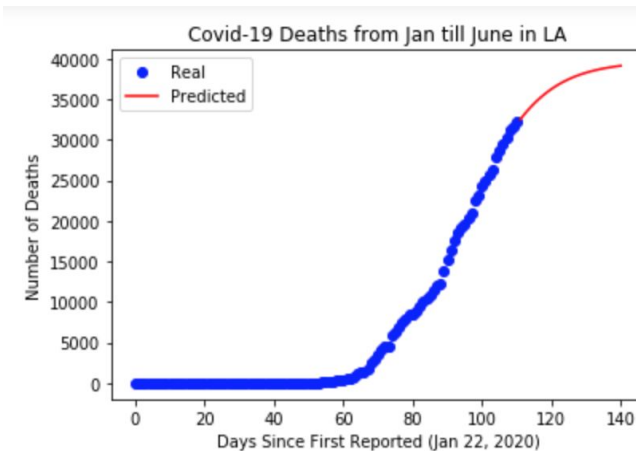
Figure 5.3.1b: Predicted Covid-19 deaths in Los Angeles until June 13th

5.3.2 Medium Risk Zone: San Diego

San Diego was part of the medium-risk cluster. The results were similar to those found in the previous section, in that the logistic model could be applied to counties in medium risk zones and the progression of the virus within these counties also followed a logistic model. While the

data points for Covid-19 deaths within San Diego were somewhat disjointed, evident in *Figure 5.3.2b*, the path still follows a logistic growth. Furthermore, similar to Los Angeles, San Diego was predicted to only be near the beginning of a plateau by mid-June (see *Figure 5.3.2a*), and thus, it is recommended that the shelter-in-place order continue to be enforced for the next month to limit the spread of coronavirus within medium-risk zones.
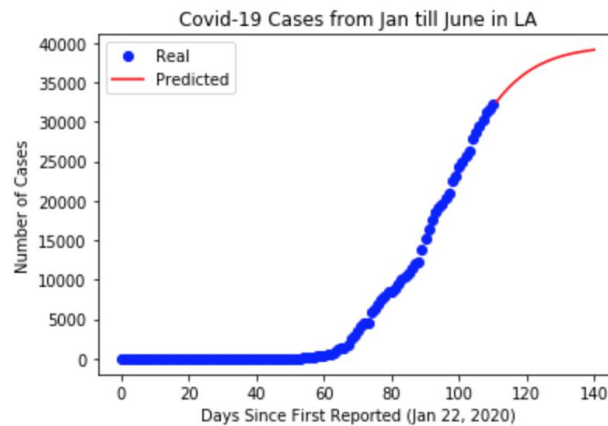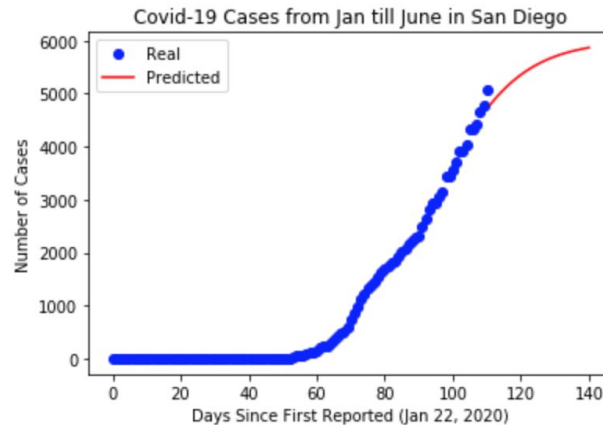

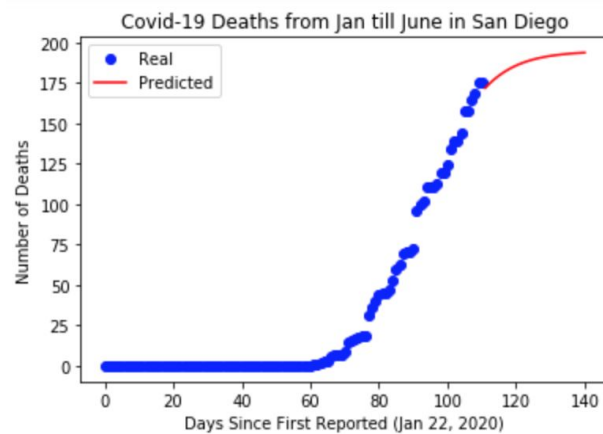*Figure 5.3.2a: Predicted Covid-19 cases in San Diego until June 13th*

Figure 5.3.2b: Predicted Covid-19 deaths in San Diego until June 13th

5.3.3 Low-Risk Zone: Shasta

Finally, the logistic model was applied to Shasta, a county from the low-risk cluster. Both graphs for confirmed cases and deaths due to Covid-19 exhibited disconnected data points, similar to that of a piecewise function, as illustrated in *Figure 5.3.3a* and *5.3.3b*. This is likely due to smaller population size, lower number of infected individuals or death, or minimal data. As a result, it was difficult to accurately model the predictions, evident by the fact that in *Figure 5.3.3a*, the projected infected individuals was 29 by mid-June while the actual number of cases was 32 by mid-May. Even so, it is important to note that even so, the data points follow a logistic growth trend. Overall, the logistic model is not best suited to be applied to predict Covid-19 cases and deaths for counties classified as low-risk zones. Furthermore, the shelter-in-place order may not need to be enforced as it is likely the number of cases and deaths will remain constant

for the next month, it is recommended that a less-restricted shelter-in-place order continue given the chance that individuals may travel from high or medium-risk zones to low-risk zones.
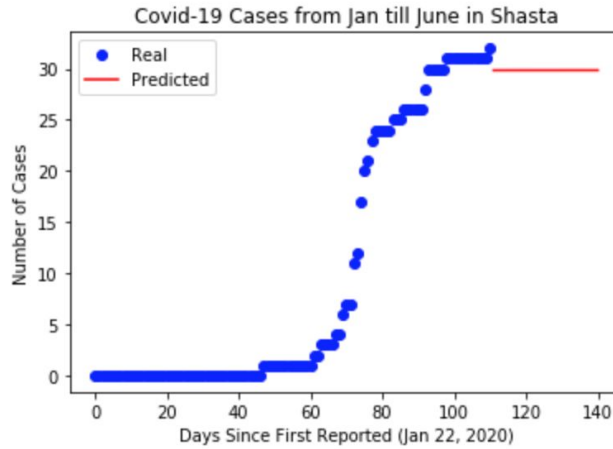

Figure 5.3.3a: Predicted Covid-19 cases in Shasta until June 13th
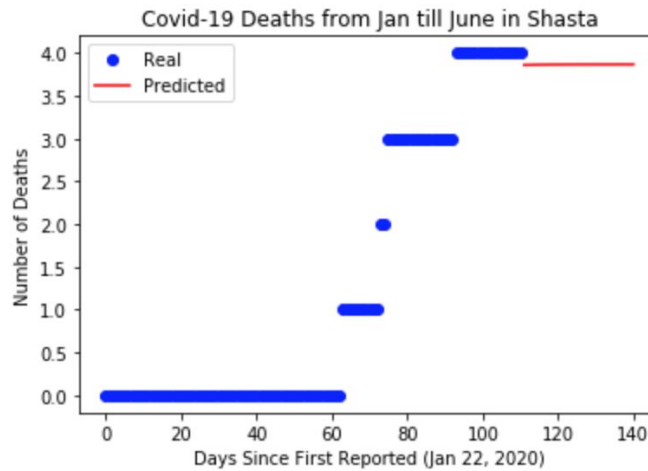
Figure 5.3.3b: Predicted Covid-19 deaths in Shasta until June 13th

5.3.4 Evaluation of Applying Logistic Model

After applying the logistic model to counties in each of the three risk zones in order to predict the number of infected individuals and deaths due to coronavirus in each selected county, it was found that the model is best suited for counties classified as high-risk and medium-risk zones. While all the counties selected indicated a logistic growth pattern, counties classified as high-risk or medium-risk zones had a higher number of cases and deaths as well as smoother data points and thus, the logistic model was able to better predict the number of cases and deaths in mid-June.

Overall, it was found that the number of Covid-19 cases and Covid-19 deaths are expected to continue to increase over the next month in high-risk and medium-risk counties, thus it is imperative that strict shelter-in-place orders continue in counties in these clusters. For

low-risk zones, while it is likely the number of cases and deaths will not rapidly rise, a relaxed shelter-in-place order is recommended in order to minimize the spread of the virus and discourage residents in medium and high-risk zones from traveling to low-risk zones.

6. Logistic Regression Analysis and Observations

6.1 Challenges With the Data

Initially, while developing our model, we hoped to train one model on the New York dataset, given that New York counties are beginning to reach a plateau, and utilize this same model to predict the occurrence of a plateau in California counties. However, since our initial model only utilized dates as a feature, this model was unable to adapt to new counties. This led us with two options: vary the features we utilized or our methodology for building and adapting the model to new counties. We decided on the latter, utilizing counties' features to instead aid in clustering and classifying counties into risk zones and testing the feasibility of a broader model on each risk zone.

6.2 Model Assumptions

While developing the model, it was assumed:

- The data in the given datasets was not misreported; the data is an accurate reflection of the given counties/states.
- There will not be any second spikes in cases or deaths due to coronavirus (or any other extreme changes to virus growth trends).
- The clustering accurately classified counties into three risk zones.
- There will be no drastic population or demographic changes within the state or counties.
- If the model could be applied to one county from each risk zone successfully, it was well-suited to be implemented to predict cases and deaths in other counties too.
- The spread of coronavirus is similar to other infections in that overall, the growth of coronavirus will ultimately result in a logistic pattern.

6.3 Model Limitations

While the logistic model could be successfully applied to predict the number of infected individuals and deaths in high and medium risk zones, the developed model has limitations:

- The model can only be successfully applied to counties with a greater number of infected people or deaths due to Covid-19. This was evident in that the model was not suited for counties classified as low-risk zones.
- The model assumed that the spread of the virus will eventually plateau and thus, cannot account for any drastic changes in virus growth or death patterns.
- The model was developed and chosen based on current data; however, it is unclear whether it can continue to be fitted to future data points.

7. Ethical Evaluation and Next Steps

7.1 Ethical Dilemmas

Some ethical dilemmas residents may face when zoning is applied to California counties:

- It is difficult in practice to apply stricter rules for one county while the neighboring county's lockdown rules are relaxed – this would mean that cottage industries, mom and pop stores, and home-grown businesses shall be struggling in one part of California, while the other part gets an opportunity to restart. Some might deem this as unfair and discriminatory, especially given the fact that a positive correlation was observed between a county's social vulnerability level and the number of COVID-19 related deaths. This would essentially mean that counties with greater social vulnerability, who are more at risk of not only COVID-19 but also losing their livelihoods with the economy being shut down, will suffer the most from stricter implementations. The best way to address this is by conducting a cost-benefit analysis of reopening the economy in high-risk zones and making the data transparent for people to see.

- The second ethical dilemma we face by introducing zones across California is that richer people might just travel or buy a temporary accommodation in the counties that are declared low-risk. The affluent Californians will prefer enjoying the "freedom" of being able to leave their houses and pursue normal activities. The influx of people migrating will once again increase the level of risk in these counties, defeating the entire purpose of introducing zones. Hence, one way to address this would be to consider prohibiting relocation so that the zones can efficiently serve their purpose.

7.2 Next Steps

If this project were continued further, the next steps taken would include:

- Collecting additional data on recovery rates and testing rates to strengthen our analysis further. Moreover, there are studies on the differential effects of the pandemic on ethnicities, so maybe more data on that for each county might be an interesting metric to use.

- Computing a Comorbidity Index using the data on Diabetes, Heart disease, Smoking, and Stroke Mortality – this might yield a more elaborate representation of a county's demographic to increase the accuracy of our clustering in assigning zones.

- Further developing the logistic model's accuracy in order to apply this form of predictive modeling to other states and countries.

8. Conclusion

Based on the following primary patterns in our data, we can conclude that it would be best if California extends its Lockdown measures at least for another 30 days:

- The presence of a large number of orange zones in the K-Means clustered counties that could potentially become red if the shelter-in-place is eased.

- The projected increase in cases and deaths in high-risk counties even now until they start to plateau. Only after they have started to plateau should counties consider lifting Lockdown.

Our conclusions were recently supported by the extension of Lockdown in several parts of California (Gorman).

9. Works Cited

Github Resources:

CSSEGISandData. "CSSEGISandData/COVID-19." *GitHub*, 11 May 2020,
    github.com/CSSEGISandData/COVID-19.

News Resources, Government pages, and Papers:

"(PDF) Estimation of the Final Size of Coronavirus Epidemic by the Logistic Model."
    ResearchGate,
    www.researchgate.net/publication/339240777_Estimation_of_the_final_size_of_coronavir
    us_epidemic_by_the_logistic_model.
"California Coronavirus Deaths Hit Stubborn Plateau; Experts Fearful about Future." *Google*,
    Google,
    news.google.com/articles/CAIiEJPrmXVOQZTYmbxAUKe703MqFwgEKg4IACoGCAo
    wjKxcMOvhCzCukLcB?hl=en-IN&gl=IN&ceid=IN:en.
"Coronavirus Update: Outbreak at Fresno Veterans Home; Tulare County Hits 50 Deaths."
    Google, Google,
    news.google.com/articles/CAIiEJCoSeT3zknawiSaYd3CsnwqGQgEKhAIACoHCAowqt
    mECzDJyIEDMIn4gwY?hl=en-IN&gl=IN&ceid=IN:en.
Gorman, Steve. "Los Angeles Mayor Hedges on Plan to Extend Coronavirus Lockdown for
    Three Months." *Reuters*, Thomson Reuters, 12 May 2020,
    www.reuters.com/article/us-health-coronavirus-usa/los-angeles-mayor-hedges-on-plan-to-
    extend-coronavirus-lockdown-for-three-months-idUSKBN22O1Y7.

Kfsn. "All Fresno County Parks, Campgrounds Are Back Open." *ABC30 Fresno*, 11 May 2020,
abc30.com/hobbies/all-fresno-county-parks-campgrounds-are-back-open/6173552/.

Kgo. "COVID-19 in California: Coronavirus Deaths in CA Projected to Rise Higher than Earlier
Prediction." *ABC7 Los Angeles*, 11 May 2020,
abc7.com/coronavirus-california-deaths-covid-19-bay-area-study/6172614/.

"Principles of Epidemiology." *Centers for Disease Control and Prevention*, Centers for Disease
Control and Prevention, 18 May 2012,
www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html.

Rowan, Harriet. "Watch: The Stunning Growth of Coronavirus Cases in California's Hardest-Hit
Counties." *Daily News*, Daily News, 4 Apr. 2020,
www.dailynews.com/2020/04/03/watch-the-stunning-growth-of-coronavirus-cases-in-calif
ornias-hardest-hit-counties/.