|  |  |
| --- | --- |
|  | Analysis of COVID-19 Across US States    **Michael Zaghi**  **Jason Cao**  **Course:** CMPT 353 Computational Data Science  **Term:** Summer 2020  **Instructor:** Greg Baker  **Date:** August 13, 2020 |

# Problem Statement

COVID-19 is an infectious disease causing fever, cough, shortness of breath and other respiratory symptoms. While the majority of cases are mild to moderate in severity, those individuals who are elderly or have underlying health conditions are more susceptible to the virus. As of August 7th, 2020, there are 5.06M confirmed cases and 162K deaths in the United States alone [1]. The goal of our analysis is twofold. First, we collect and analyze the underlying socioeconomic and racial factors in order to stratify and understand the state of the pandemic in the US. Next, we examine suspected attributes as possible driving forces contributing to COVID death and utilize the previous analysis results as a framework to construct a time series model that could predict short term future deaths in the US.

# The Data

###### Sources and Collection Method

We used three different sources of data for our analysis. The *COVID Tracking Project* dataset is used for daily death and race data by state. For socioeconomic data, we used a dataset from *Kaggle* which aggregates *John Hopkins* *US Death* data with *ACS* (American Community Survey) *US Census* data. Lastly, we collected US weather data from the *NOAA* (National Centers for Environmental Information) database. All data were gathered in csv format. Links to the original data can be found under the *Data Sources* section.

###### Cleaning

Data cleaning scripts transformed the collected data into a format that is more suitable for future analysis and modeling. In general, this means selecting, reformatting, merging, and grouping the datasets into an aggregated ‘base data’ that could be used for additional manipulations during statistical analysis and model construction. These base datasets can be found in the *output* folder as *daily\_cases\_with\_race.csv, total\_cases\_with\_race.csv,* and *total\_cases with\_socioecon.csv*.

Daily weather data from 25 states were requested from NOAA. Data cleaning was partially manual because some states mistakenly had weather data from other states. A selective set of weather attributes (wind speed, temperature and precipitation) were chosen as possible features that might affect the number of COVID deaths. For each state, missing data was imputed via Bayesian Ridge Regression [2]. The aim was to produce the most likely value for missing temperature, wind speed and precipitation on a weather station using known values from other weather stations in the same state. This method is appropriate, assuming these attribute values are not very different within each state. All states daily weather data were concatenated and then joined with *daily\_cases\_with\_race.csv* to form the base data, *dailyCases.csv*, for subsequent feature selection and time series analysis. At the end of data cleaning, 12 of the 25 states had the necessary data to continue.

# Descriptive Analysis

###### Choropleth Maps

Choropleth maps are a good way to provide a broad understanding of the geographical impact of the virus. *Figure 1* illustrates the total deaths for all 50 US states. It is clear that New York and its neighboring States have taken the brunt of the impact. On the west coast, California has the greatest number of deaths. Meanwhile, Middle America (the more rural US interior) has been significantly less affected by the virus.

*Deaths by Population Over a*nd *Deaths by Population Density* (*Figures 2* and *3* respectively), use relative measures which may allow us to make more insightful comparisons across States. The IFR (infection fatality rate) is significantly greater for individuals over the age of 65 (5.6%) compared to the next category down which is 50-64 (0.14%) [3]. *Figure 2* gives a better measure of how states are performing compared to absolute deaths. One state that differentiates itself from the rest is Louisiana. It has a relatively small population over 60 compared to their total deaths. We will arrive back at the results of *Figure 3* once we can show that population density is a useful measure.

###### Linear Regression of Socioeconomic Factors

The goal behind this analysis is to identify aggregate level drivers of the US death rate which could be used for prediction. We collected several interesting independent variables such as *Household Income Mean Earnings*, *Total Households with More than One Occupant per Room*, *Age Over 60 in the Civilian Labor Force*, and the *CDC SVI index* (a measure of socioeconomic, minority, language, and disability status). The only meaningful predictor of death we were able to uncover that had a linear relationship was *Population Density* (see *Figure 4*), which had an of 0.575 and a *p-value* less 0.05.

Now that *Population Density* has been established as statistically significant, we can see how a low ratio of death to population density is desirable. Looking at *Figure 3,* we can see that New York performs relatively better under this measure while Arizona performs the worst even though there was no indication of this using the two previous statistics.

# Race Analysis

###### Correlation Matrix

By observing the correlation coefficients between different racial groups, we may be able to identify some general patterns or trends that explain how each group is impacted by the virus. Note that, currently, the only states reporting cases and deaths for every group are CA, CO, and WA. Therefore, any information drawn from this data may not be representative of the entire US population. As expected, *Figure 7* shows that there is a positive correlation between cases and deaths for all groups. Probably the most notable pattern is the strong positive correlation between Black and White in all categories (cases on cases, death on death, and cases on death). Another interesting pattern is the relatively weak or negative correlation between, for example, AIAN and every other group.

These general trends could give some indication that, for example, the White and Black communities are tightly coupled in a way that leads to a similar infection rate. Likewise, we could theorize that the AIAN is the most segregated, which may explain why their cases and deaths do not correlate with other groups. Of course, deeper analysis are still need to validate these claims, but it is an interesting starting point for further analysis.

###### Daily Death Inference

For this part of the analysis, we were interested in the question “Is the mean daily death rate for each race different when weighted by their respective population?”. We used an ANOVA test to answer this question (our alternative hypothesis). Again, we had the problem of insufficient data and had to broaden our scope from the correlation analysis to AK, CA, CO, GA, IL, LA, MN, NC, and WA by dropping AIAN, NHPI and Multiracial. Of course, this is not representative of the entire US population. Some challenges in working with the data included negative adjustments to daily deaths, distinguishing between days with no reporting and zero deaths, and transforming the data so that it is normal enough to use the ANOVA test.

These issues were handled in the follow ways: the few negative adjustments were removed since some adjustments were large and led to less conservative values, the distribution was heavily right-skewed, so multiplying all death rates by a scalar and then taking the logarithm of the data helped this issue, days with zero deaths were removed because it became evident by looking at the transformed distribution that few days had actual zero deaths. The end result of running the ANOVA test was that the p-value was less than 0.05, allowing us to conclude that there was a difference in the means between groups. We were able to further reduce this information by utilizing Tukey’s HSD test to compare pairs. The results are shown in *Figure 8*. You can see that the greater death rate of Black’s over every other racial category is statistically significant at α=0.05.

# Feature Selection

###### Analysis of Attributes

Feature selection is part of the machine learning pipeline responsible for choosing the most useful attributes for model prediction. This also holds true for time series analysis. In feature selection, correlation analyses were done on bivariate relationships between combinations of the following variables: daily COVID positive cases, daily hospitalized number, daily temperature, daily wind speed, daily precipitation and daily deaths. These relationships are visualized via a scatter plot matrix. Our objective was to remove irrelevant features, indicated by low or no correlation, that could potentially negatively impact our prediction model.

The result was less than what we hoped for since some states have many correlated attributes, shown in *Figure 9*, while other states do not. What stood out from our analysis was the positive correlation among the following three variables: daily COVID positive cases, daily hospitalized and daily deaths. A large number of states have positive correlation in one or many of these relationships: death vs hospitalization, death vs positive COVID test and hospitalization vs positive COVID cases. However, none of these paired relationships are universally common to all states. In addition, only 2 of the 12 states have a negative correlation between increase in temperature and decrease in hospitalization and deaths. Consequently, given that these correlation results are not definitive enough to yield any meaningful feature attributes, our time series analysis had to be limited to using only past COVID deaths as the sole predictor to forecast future COVID deaths.

# Univariate Time Series

###### Short-term Daily Death Forecast

To forecast future outcomes based on known historical data, we used Facebook’s open sourced Prophet library [4] to predict the number of deaths in the next 8 days. Of the 12 states, each has 100 data points, spanning from March 26, 2020 to July 15, 2020. The last 8 days of the 100 days were withheld to validate the trained model.

Prophet API, similar to scikit-learn, fits the training data and then lets us specify the time periods we would like to forecast into the future. Its built-in plot function allows us to visualize the historical data points (black dots), the basic forecast plot (blue line) and the confidence interval of that forecast (see *Figure 10*). Its component function further breaks down the time series plot into underlying components (trend, weekly and hourly seasonality) that accounts for the entirety of our model (see *Figure 11*).

###### Forecast Validation and Transformation

In the simplest terms, forecast validation compares the data we initially withheld with our model’s prediction. The quality of the evaluation is decided by how frequent our model is wrong (MAPE in %) and by how much on average our model is wrong (MAE in number of deaths). Before any adjustment, the MAPE and MAE values are unusually high (see *Figure 14*). Also, from the graph, many of the actual data points lie outside the confidence interval. We postulated that the aforementioned issues may be due to unstable variance *i.e*. as the response variable (death) changes, so does the variance. To stabilize the variance, we applied a log transformation (Box-Cox transformation) [5] on the dataset then repeated the forecast, once again, using the transformed data.

###### Post-Transformation Forecast and Analysis

We created a new Prophet model and repeated the fit-predict cycle on the transformed dataset. According to the new time series graph, more data points lie within the confidence interval (see *Figure 12*). Errors (MAPE and MAE) from the post-transformed data are much lower since most MAPEs are down to the 30% range and most MAEs are within 10 deaths (see *Figure 14*).

Transformation made a great difference in some cases. Nevertheless, the high errors are indicative of other factors are at play. Clearly, using only past COVID deaths to predict future COVID deaths is not enough.

# Findings and Conclusion

Our initial statistical analysis reveals two major findings: 1) Arizona has surpassed all other states as the hardest hit state by COVID-19 based on population density and 2) the higher death rate of Blacks over every other racial category is statistically significant.

With respect to short term forecasting, we built a univariate time series model using Facebook’s time series tool, Prophet. Our prediction model has high errors despite log normalization, indicating other confounding factors are influencing COVID death. Consequently, more curated data is required to build better features to be used in multivariate time series models such as LSTM.

# Project Experience Summary

###### Jason Cao

* Requested (via email), collected and cleaned daily weather data (over 1.7 million rows) from *NOAA* for 25 states (descending from the most populous state).
* Imputed missing values for temperature, wind speed and precipitation via Baysian Regression
* Attempted to use correlation matrix to select attributes that might have strong correlation with COVID deaths.
* Built univariate time series model for 12 states to predict COVID deaths.
* Used Box-Cox log transformation to normalize unstable variance in COVID deaths.
* Performed a state-by-state error analysis comparing untuned prediction vs log transformed prediction; revealed log transformation greatly reduced the amount of errors in the prediction model.

###### Michael Zaghi

* Worked on a group-based data science project related to COVID-19 that resulted in useful findings related to socioeconomic factors and race.
* Collected, cleaned, and formatted data from multiple sources so that it could be further analyzed in the data science pipeline.
* Performed linear regression and polynomial regression on socioeconomic factors in order to quantify their usefulness for predicting COVID death.
* Analyzed death by race by using ANOVA and Tukey’s HSD tests resulting in statistically significant results.
* Generated correlation matrices by race between cases and deaths in order to find interesting trends and potential areas for further analysis.

**Glossary of Terms**

**AK** – Alaska

**AZ** – Arizona

**CA** – California

**CO** – Colorado

**FL** – Florida

**GA** – Georgia

**IL** – Illinois

**LA** – Louisiana

**MA** – Massachusetts

**MD** – Maryland

**MN** – Minnesota

**NC** – North Carolina

**OH** – Ohio

**SC** – South Carolina

**TN** – Tennessee

**NY**– New York

**WA** – Washington

**VA** – Virginia

**AIAN** – Native American and Alaskan Native

**NHPI** – Native Hawaiian and Pacific Islander

**MAE** – Mean Absolute Error

**MAPE** – Mean Absolute Percentage Error

**Repository**

git@csil-git1.cs.surrey.sfu.ca:mzaghi/cmpt-353-project-covid19.git

### References

[1] Centers for Dieses Control and Prevention (CDC). Accessed: 2020-08-12. Source: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[2] Scikit-Learn: Imputing Missing Values with Variants of IterativeImputer. Accessed: 2020-08-12. Source: https://scikit-learn.org/stable/auto\_examples/impute/plot\_iterative\_imputer\_variants\_comparison.html

[3] American Council on Science and Health. Accessed 2020-08-12. Source: https://www.acsh.org/news/2020/06/23/coronavirus-covid-deaths-us-age-race-14863

[4] Prophet: Quick Start Python API. Accessed 2020-08-12. Source: https://facebook.github.io/prophet/docs/quick\_start.html

[5] Forecasting in Python with Prophet. Accessed 2020-08-12. Source: https://mode.com/example-gallery/forecasting\_prophet\_python\_cookbook/

### Data Sources

###### COVID-19 Death and Related Statistics

https://covidtracking.com/

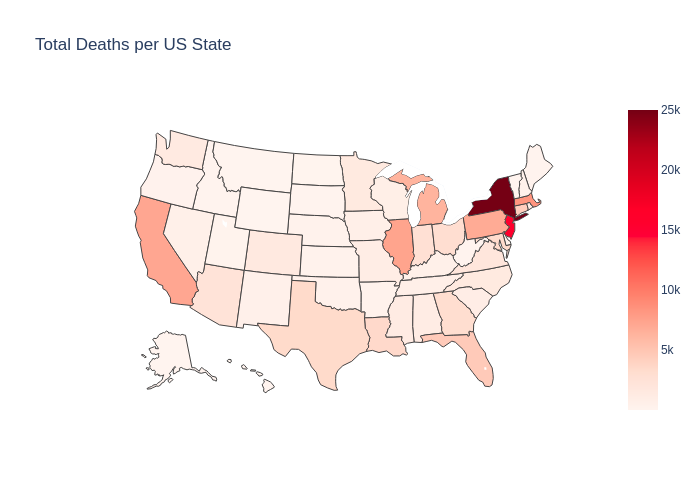
###### US Socioeconomic Data

https://www.kaggle.com/jtourkis/us-county-level-acs-features-for-covid-analysis

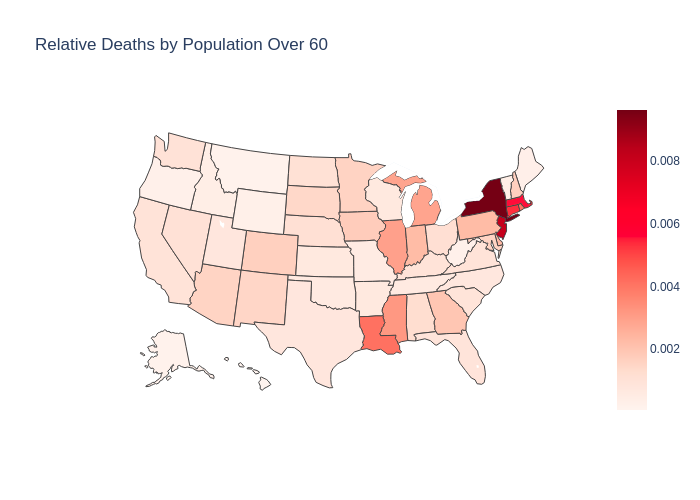
###### Weather Data

https://www.ncdc.noaa.gov/cdo-web/

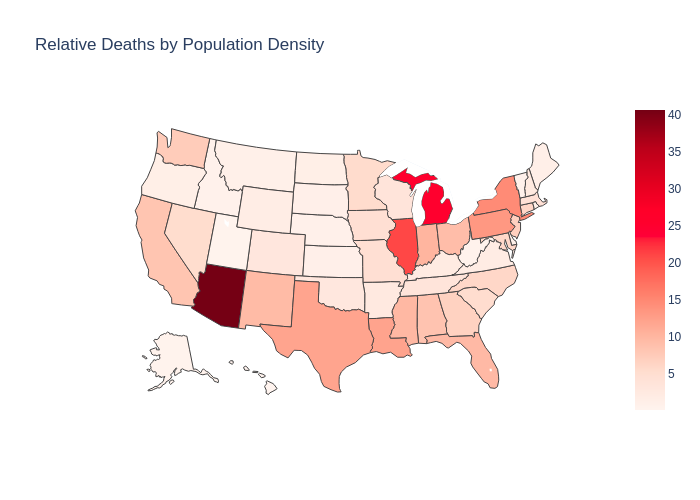
**Appendix**



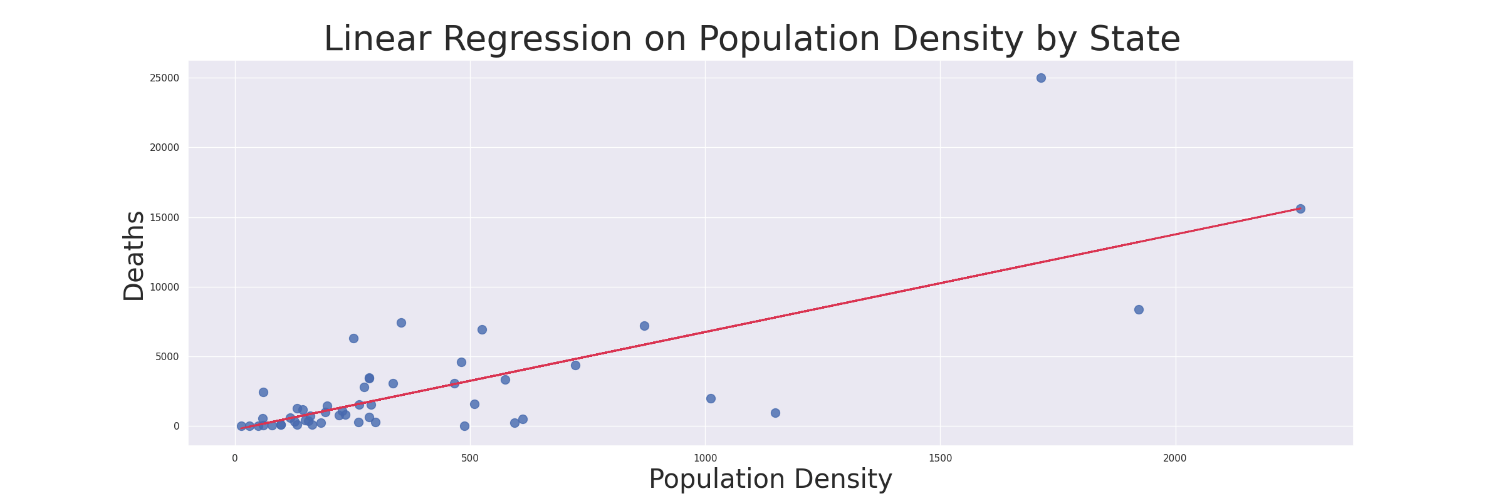
**Figure** 1**:** A choropleth map of the US showing the total deaths in the US by state.



**Figure** 2**:** A choropleth map of the US showing the ratio of deaths to population over 60 in the US by state. Computed by dividing total death by the over 60 population.



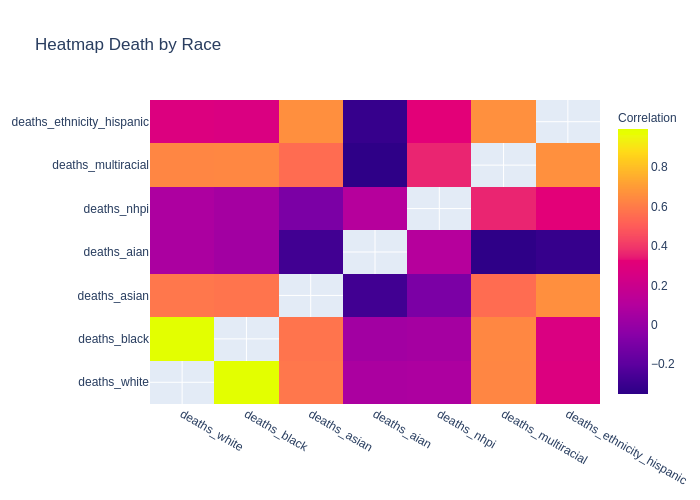
**Figure** 3**:** A choropleth map of the US showing the ratio of deaths to population density. Calculated as deaths divided by population density, where population density is computed as population per square mile.



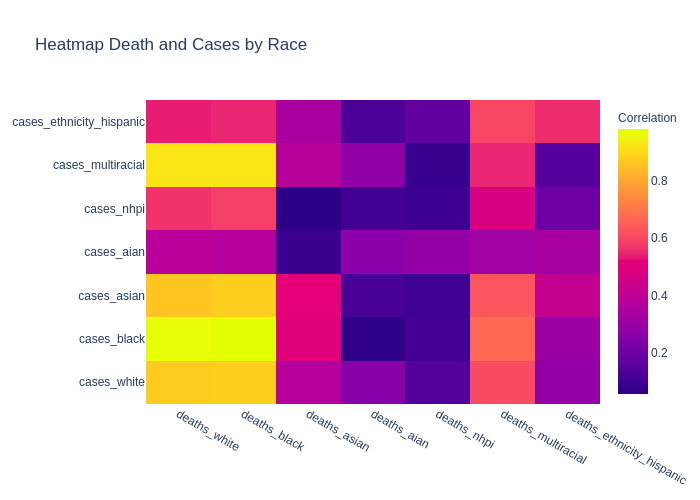
**Figure** 4**:** Linear regression of deaths on population density for all US states with an and a p-value of less than 0.05 (there is a statistically significant linear relationship).



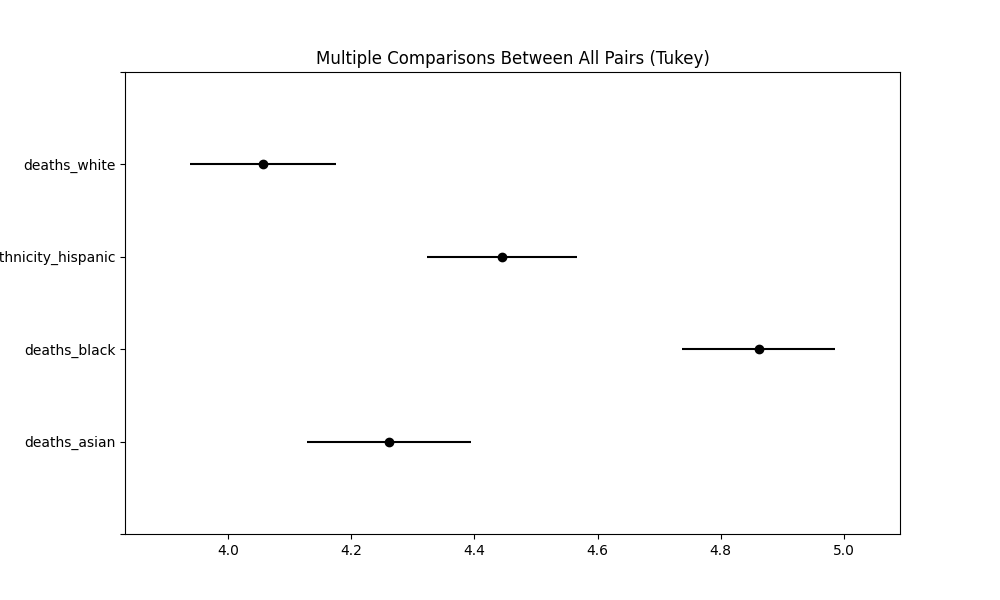
**Figure** 5**:** A heatmap showing the correlation between cases across racial groups for CA, CO, and WA. The diagonal entries are removed (the correlation is one) so that neighboring values are easier to distinguish. Data is for 2020-05-03 to 2020-07-15.



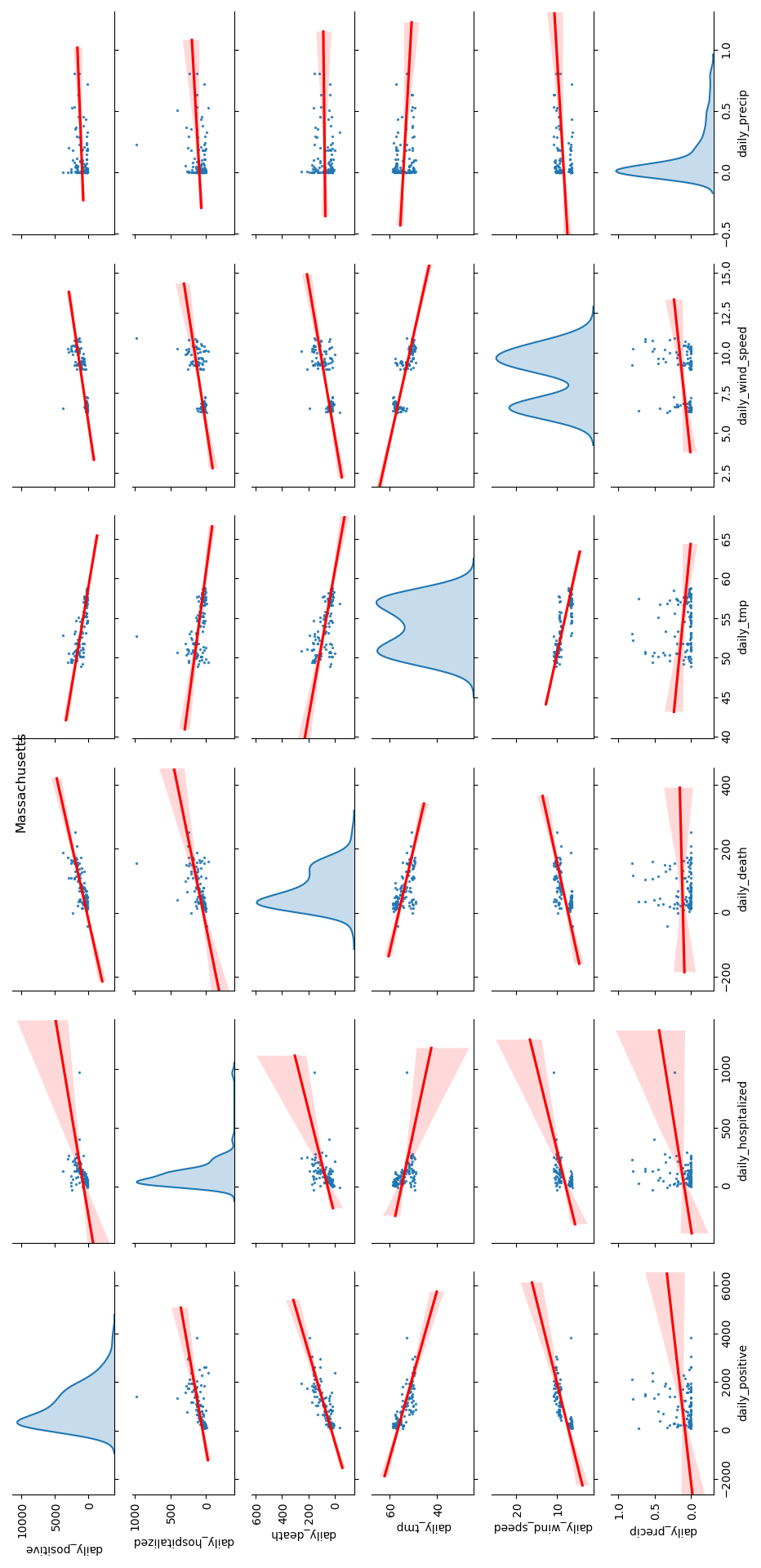
**Figure** 6**:** A heatmap showing the correlation between deaths across racial groups for CA, CO, and WA. The diagonal entries are removed (the correlation is one) so that neighboring values are easier to distinguish. Data is for 2020-05-03 to 2020-07-15.



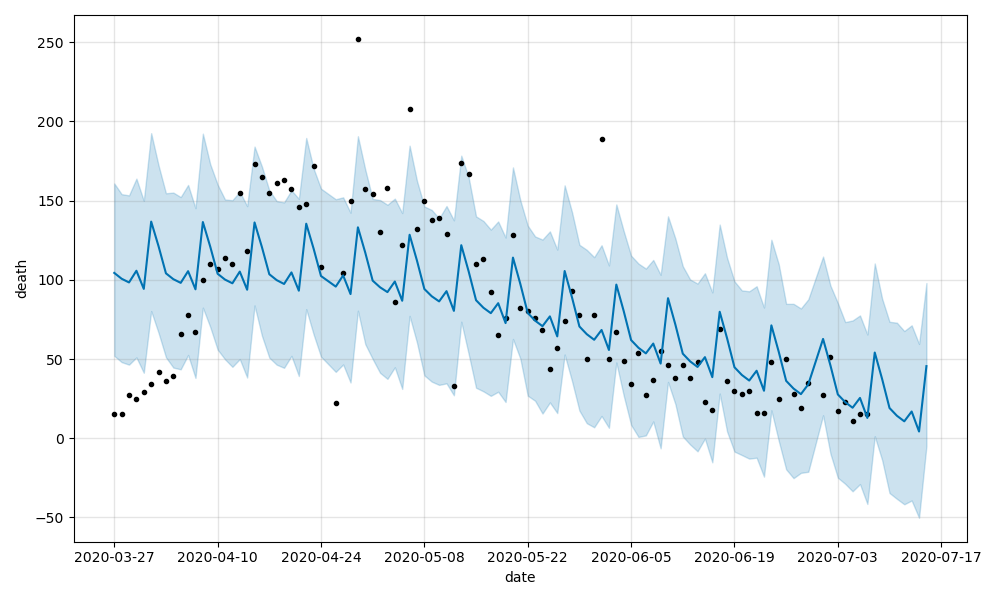
**Figure** 7**:** A heatmap showing the correlation between deaths and cases across racial groups for CA, CO, and WA. Data is for 2020-05-03 to 2020-07-15.



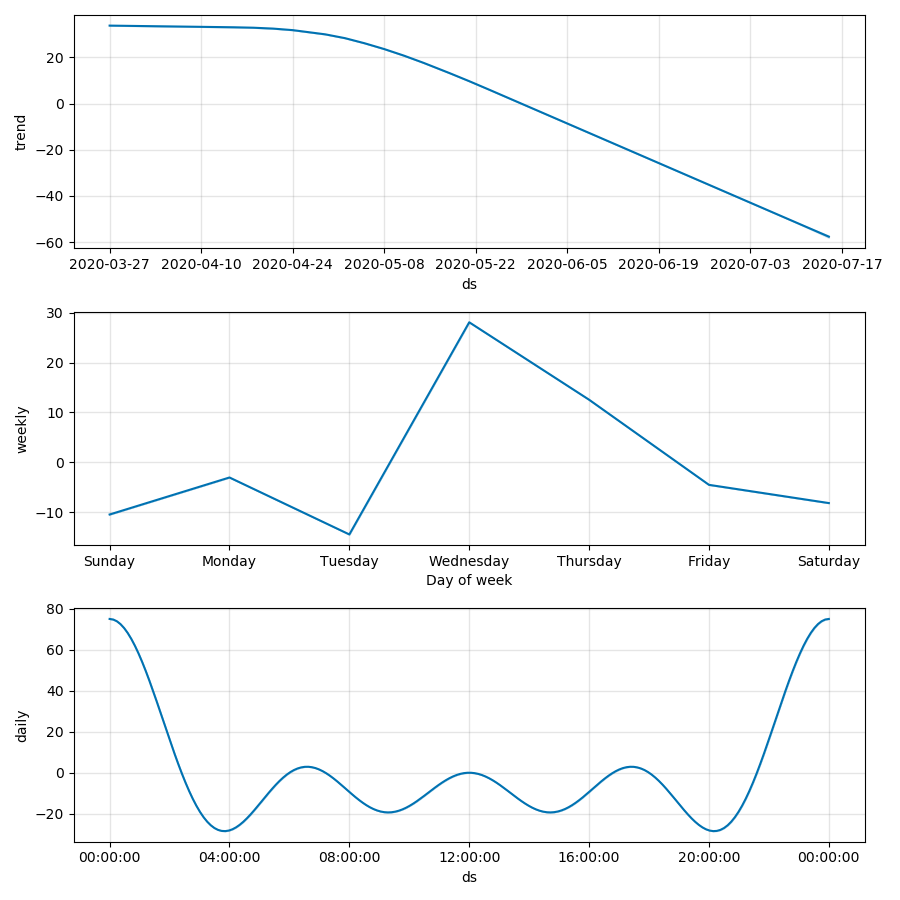
**Figure** 8**:** The results of the Tukey’s HSD test on the average daily death rate for AK, CA, CO, GA, IL, LA, MN, NC, and WA. Shows that black deaths are significantly greater than all other races and that Hispanic deaths are significantly greater than White deaths. The difference in the death rate of White and Asian is not statistically significant, as is the difference in death rate between Asian and Hispanic. X-axis is scaled and logged. The actual death rates for each race are: White - 0.0009%, Black - 0.0019%, Hispanic - 0.0012% and Asian - 0.0010%. Data is for 2020-04-26 to 2020-07-15.



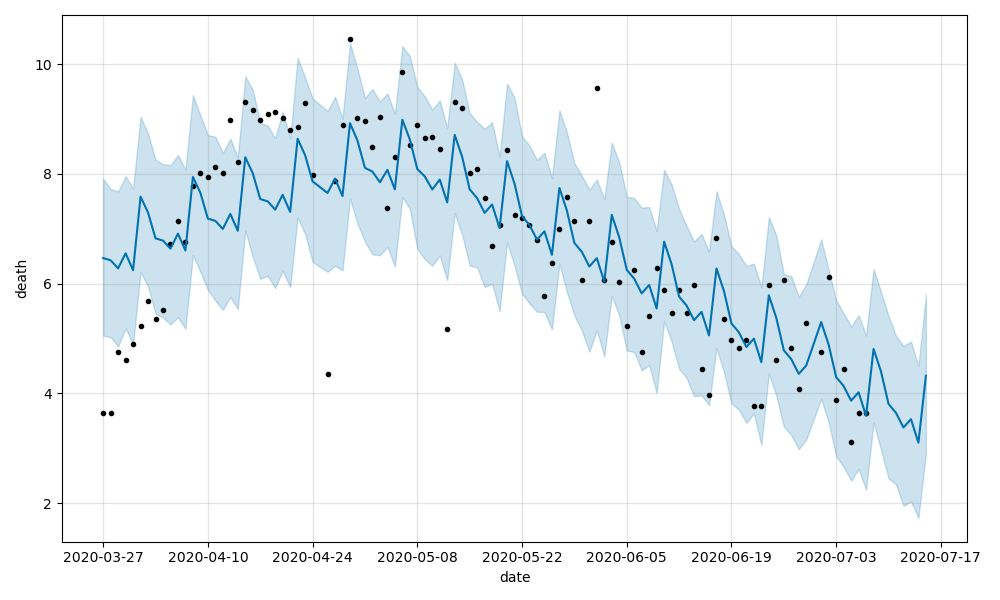
**Figure** 9**:** Scatter matrix plot was used to visualize bivariate relationships between combinations of paired attributes in order to determine how one affects the other. This method was an attempt to construct a set of features for building multivariate time series model. However, given the variability of these relationships across the 12 states, no consistent feature set could be built.



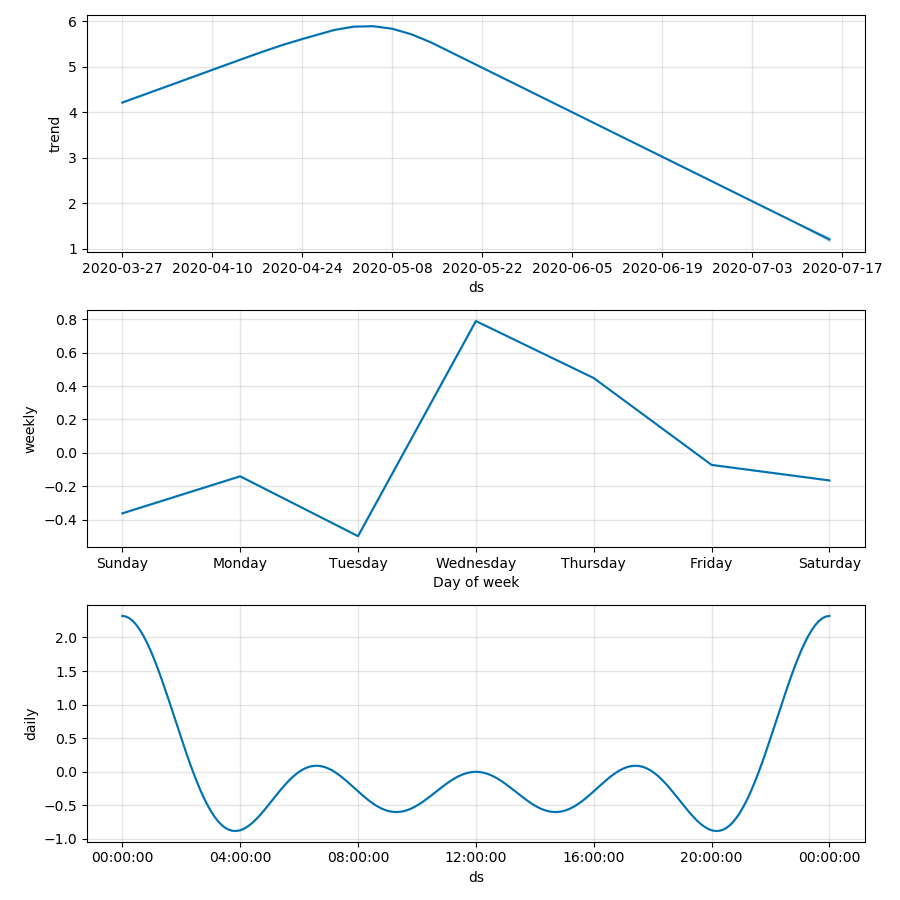
**Figure** 10**:** Univariate time series forecast of the state of Massachusetts before log transformation. Time series plot is represented by the blue line. Most data points (black dots) are distributed near the plot and within the confidence interval (light blue region).



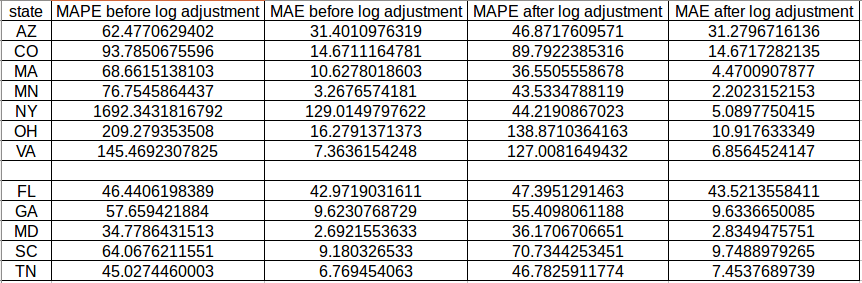
**Figure** 11**:** The break down of the entire time series plot into 3 components: trend, weekly and daily seasonality. Trend describe the direction of the time series plot is moving towards. Weekly and daily seasonality describes how the trajectory of the time series plot is affected by the daily and weekly trends.



**Figure** 12**:***Univariate time series forecast of the state of Massachusetts after log transformation. The new time series plot bends closer to the data points than before the log transformation. Its confidence interval covers more data points.*



**Figure** 13**:** Log transformation is a corrective method used to by the prediction model to fit the data better and minimize errors. The components of time series plot are not altered by this parameter tuning technique.



**Figure** 14**:** Evaluation of time series model was done by comparing the training data against the validation data. MAPE is the % error the prediction is from the reality. MAE is the magnitude of that error. More than half of the states (7) responded to log transformation and greatly reduced their prediction errors. Log transformation did not reduce the prediction errors on the other 5 states.