Intro to Problem

COVID-19 is a coronavirus outbreak that began in late 2019. The virus has since spread across the entire world and with no effective antiviral nor vaccination, people have relied on social distancing to slow the spread of the virus and hopes of eventually reaching herd immunity. The consequences of social distancing (job and/or income loss, trouble affording rent or other essentials, economic crash) have made this an untenable option for many. Moreover, the hopes for herd immunity fade as new mutations and recurrent outbreaks continue. There is also an increased prevalence of the virus in certain populations (those with preexisting conditions and those with nutrient deficiencies). It follows that the incidence of COVID-19 could both be predicted by the rate of certain diseases, nutritional, and social factors and perhaps mitigated by acting to improve the condition of those with existing disease/nutrition factors.

In order to determine if this assumption is correct, the use of disease and nutritional factors may be used to predict the day at which 100 cases of the disease are reached. The day that 100 cases are reached is an informative value due high correlation to future rates of infection and ultimately to the total number of cases.

Data Wrangling

Data on disease, nutritional, and social factors was sourced from the following sources APIs or excel/csv downloads:

- Annual number of deaths
- Nutritional factors dataset
- Mobility Reports
- COVID cases confirmed in each country

This data was then cleaned and formatted into tables of similar type. Those tables were ultimately merged on the country name.

Once the data had been merged, examination of feature distribution showed what type of data was missing and where it was missing (Figure 1). With the visualizations in Figure 1, it is clear which countries have more or less variables associated with them and which types are feature types are present, or not, most often.

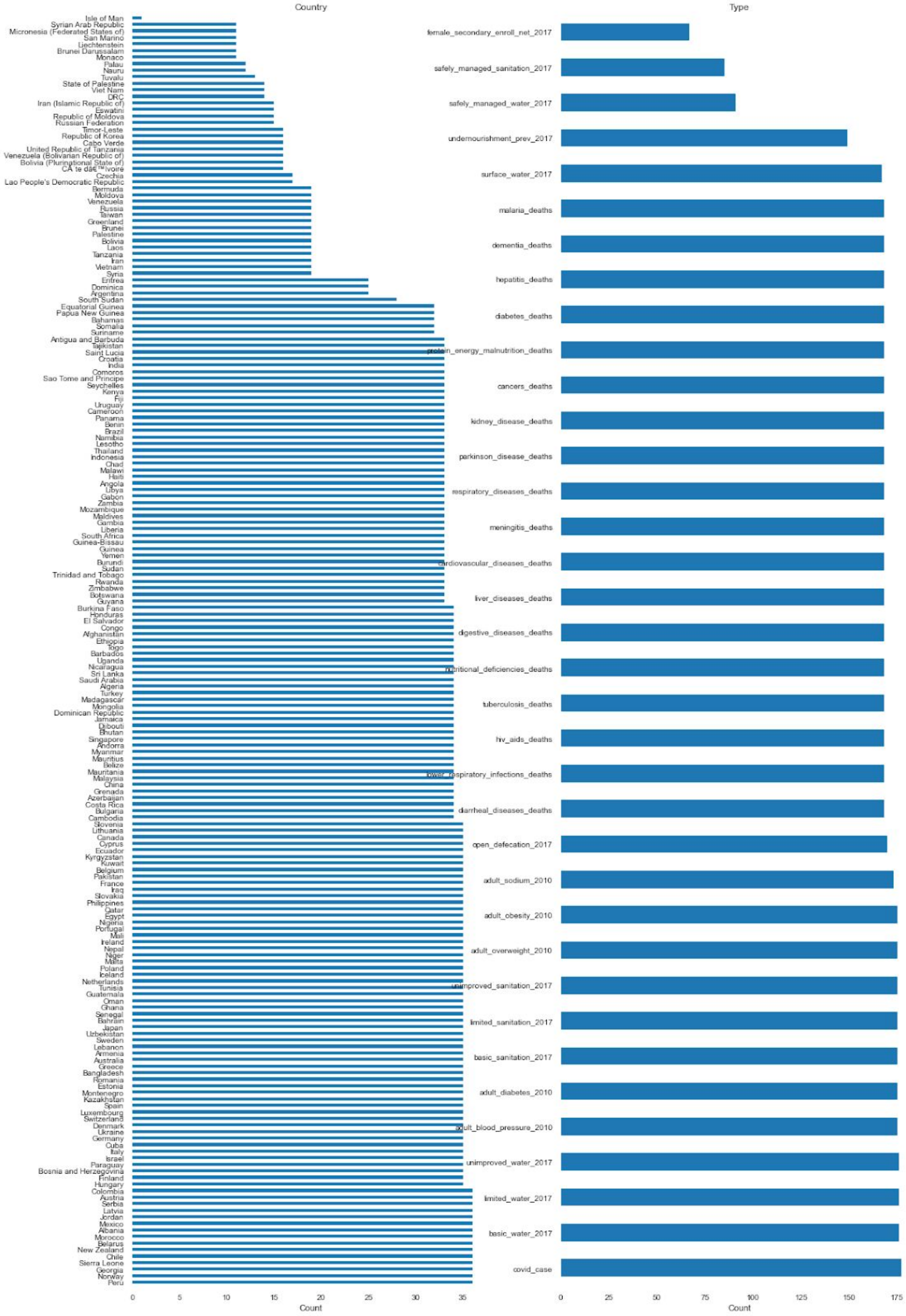


Figure 1:Feature distribution amongst countries (right) and by feature type (left)

Taking a look at the numerical distributions in Figure 2, there are some outliers, however, the occurrences make sense. For example, it would be expected that features like limited water or undernourishment would have a range in which most countries fall, but may also occur at higher percentages in some countries based on available resources. Upon initial inspection of the numeric data, the features do have a broad range of values and this is considered prior to modeling.

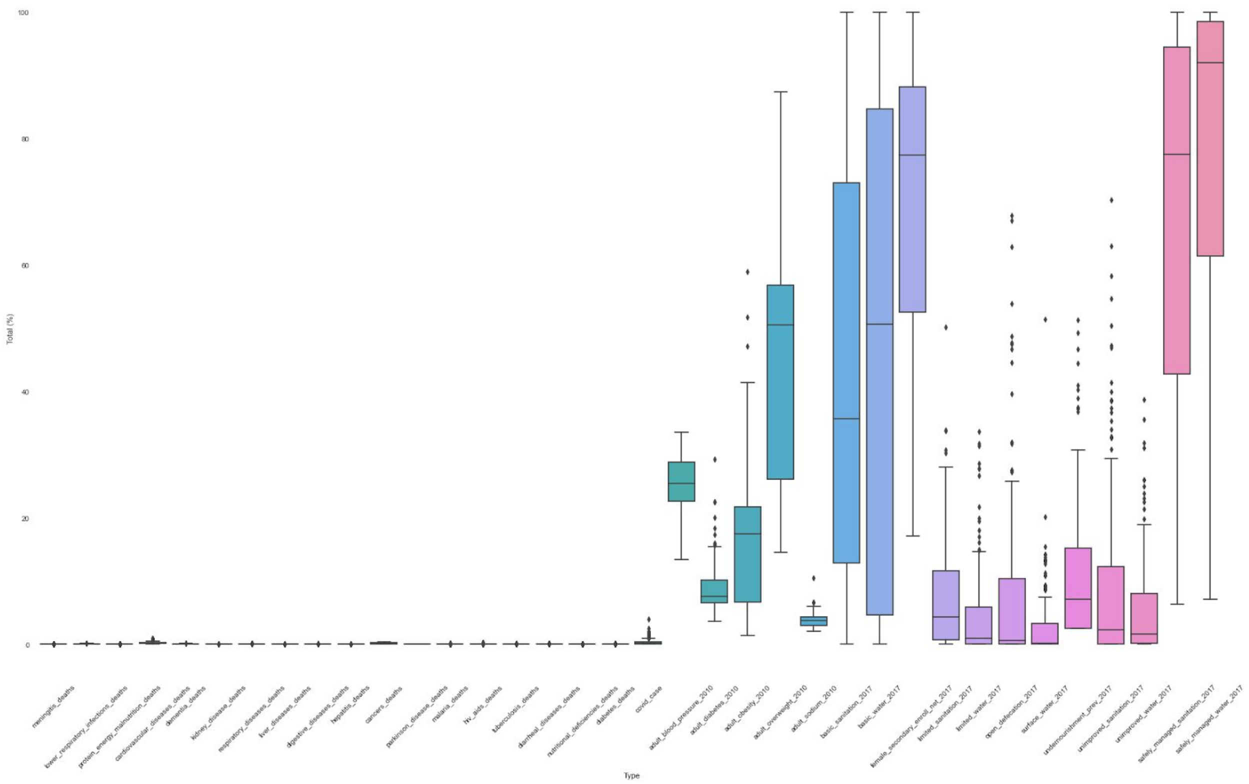


Figure 2: Numerical Feature Distributions

For a look at each step taken during data wrangling, see the R-markdown and Jupyter notebook files.

Exploratory Data Analysis

During exploratory data analysis, the data was examined in more depth, variables correlation was examined, data was imputed and scaled, and higher dimensional analysis was conducted.

During examination of the data, the date was converted to a day number. This day number represents the number of days from the beginning of the pandemic to the day that the country reached 100 or more cases. Recall, this is the value that will be predicted during modeling.

A correlation matrix was created once the day_number column was created. This matrix indeed confirmed a relationship between the COVID cases per capita and the day number meaning the day number has some representation of the greater COVID picture. This is later confirmed again during principal component analysis. The correlation matrix also showed which features showed minimal correlation, despite intuition that they would correlate. For example, the mobility features (data relating to social distancing) was minimally correlated. This phenomena can be seen through the scatter plots in Figure 3. While counter-intuitive, this could be explained through a few people not social distancing and infecting many others or by the fact that social distancing was not implemented early on in the pandemic.

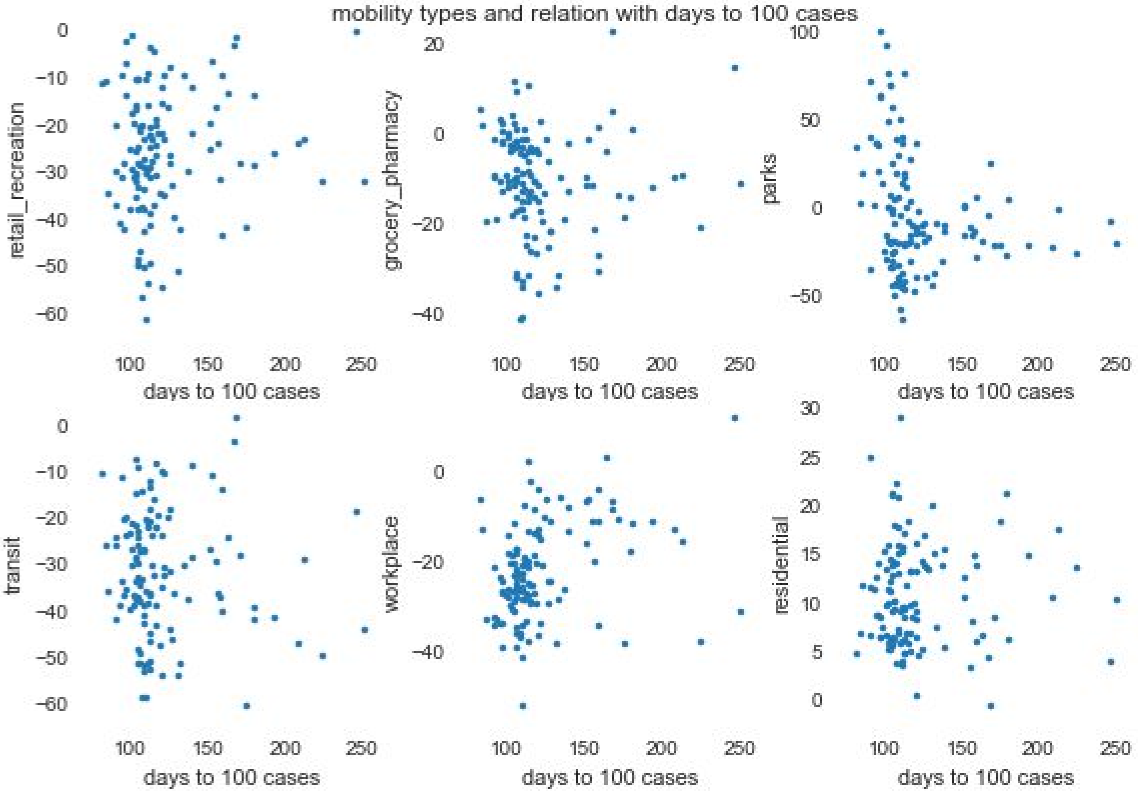


Figure 3: Against intuition, little correlation exists between people's mobility and the days to 100 cases.

Some other relationships were highlighted by the correlation matrix and are discussed further below.

The macroeconomic rating showed high correlation. Upon closer examination (Figure 4), the 'banding' characteristic of a discrete variable is evident. This variable could be used as a numeric or categorical, however, in this case, it makes sense to use the numeric value since these values are true numbers with the interval between values having meaning.

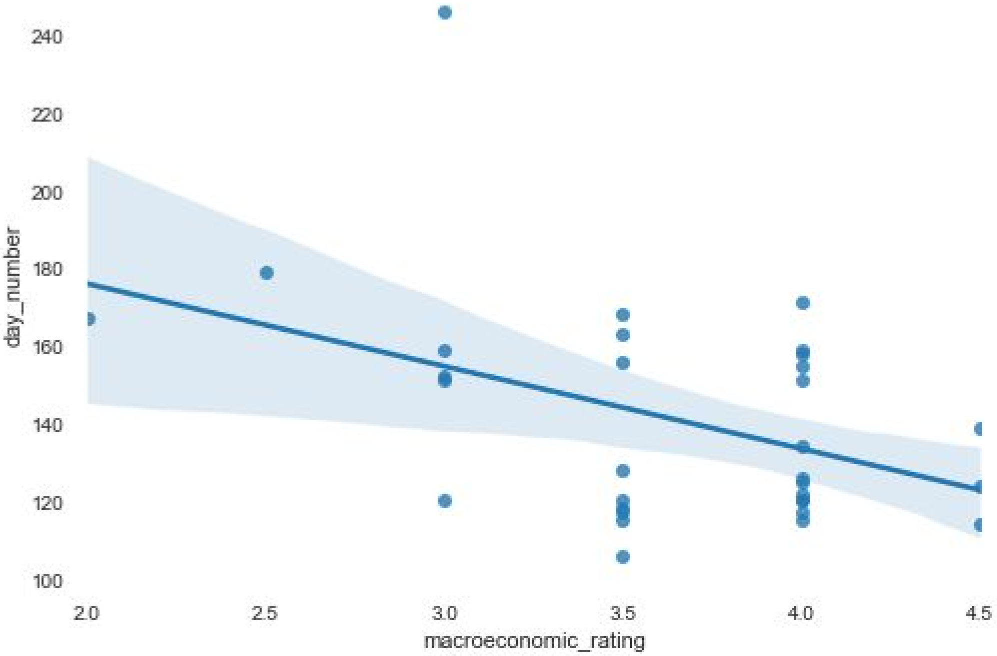


Figure 4: Scatter plot with best fit line for a closer look at the relation between the day number and macroeconomic rating feature.

One of the strongest corollary relationships was with the physicians feature (Figure 5). Unfortunately, this feature was also one of the least populated and ultimately required imputation.

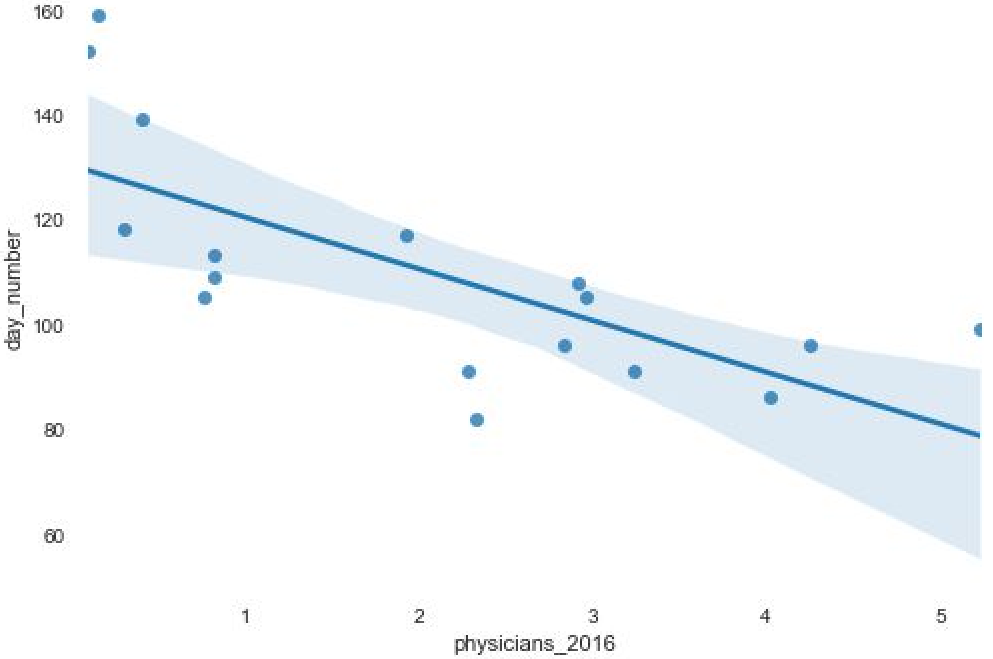


Figure 5: Scatter plot with best fit line for closer look at the relation between the day number and physician feature.

Once the data was examined, it was clear that data needed to be imputed in several places. Based on the assumption that all values were correlated (at least somewhat) to the day number, the dataframe was sorted by dataframe and all null values were linearly imputed based on this order.

Despite having determined some potentially relevant features for the prediction of the day when 100 cases are reached, there are still some concerns about the predictive variables. Thus, principal component analysis (PCA) was used to find linear combinations of the features that are uncorrelated with one another. First the features are scaled so the mean is approximately 0 and the standard deviation is 1. Then, features derived through PCA are visualized (Figure 6) in lower dimension to understand how much variance the representation explains. The original features contributions to the derived features provide insight for modeling.

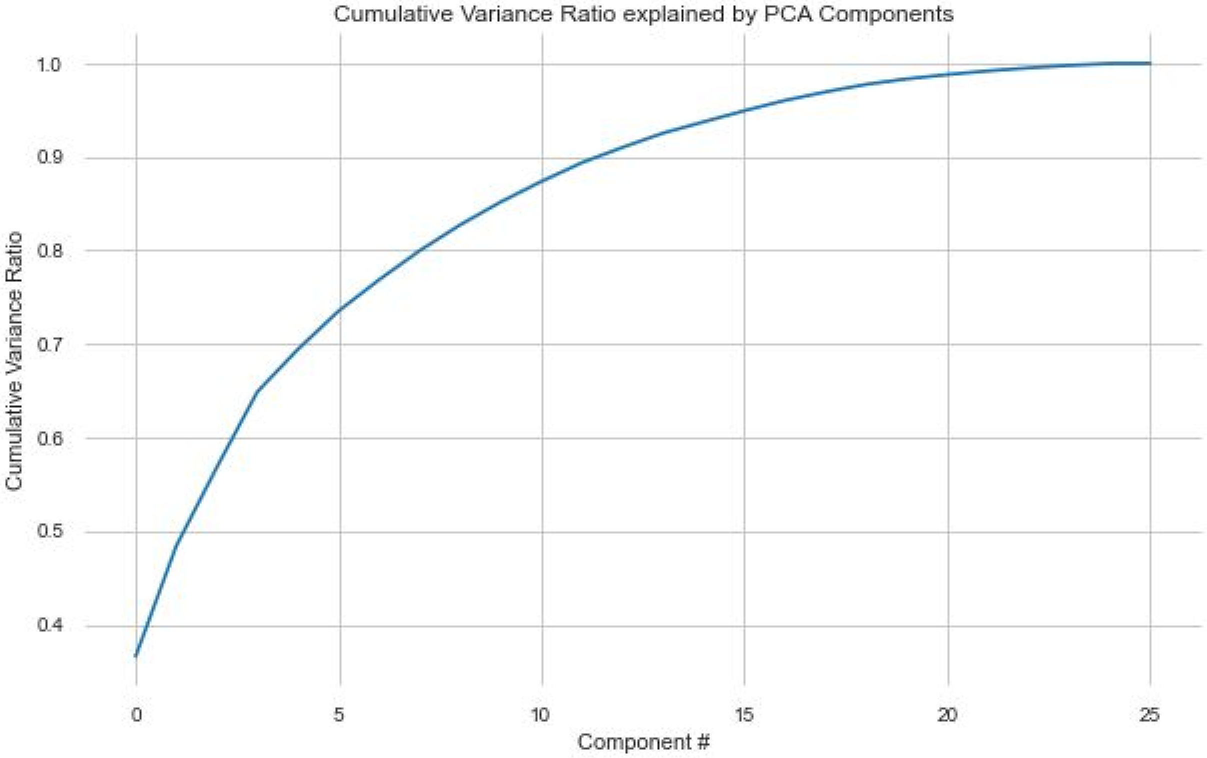


Figure 6: Variance explained through PCA features.

To determine the effectiveness of the principal components, scatter plots with the first two components (explaining 48% of the variance) and color coding by the day number quantile range was performed (Figures 7 and 8).

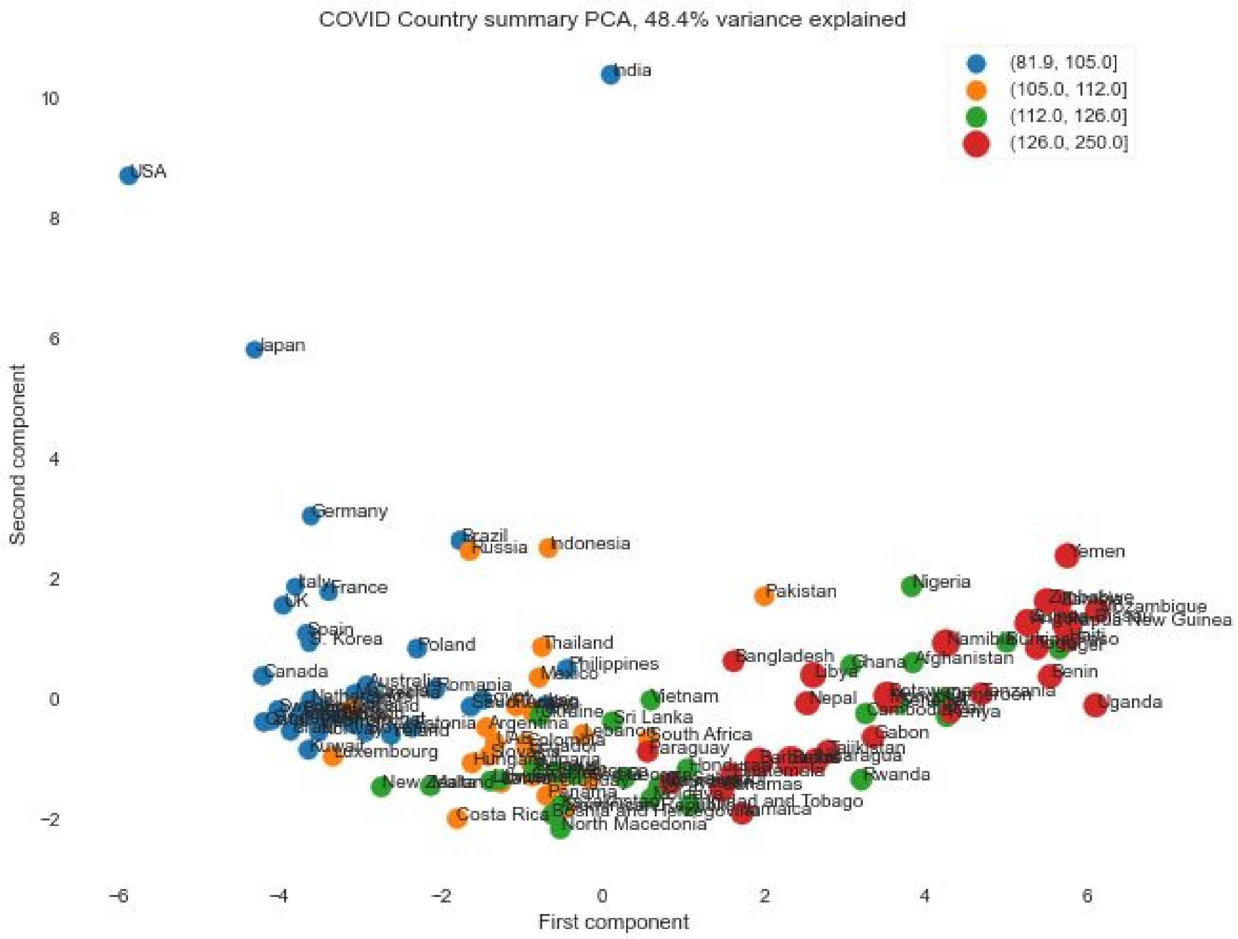


Figure 7: Scatter plot between the first two principal components with color coding by four quantile ranges of the day number.

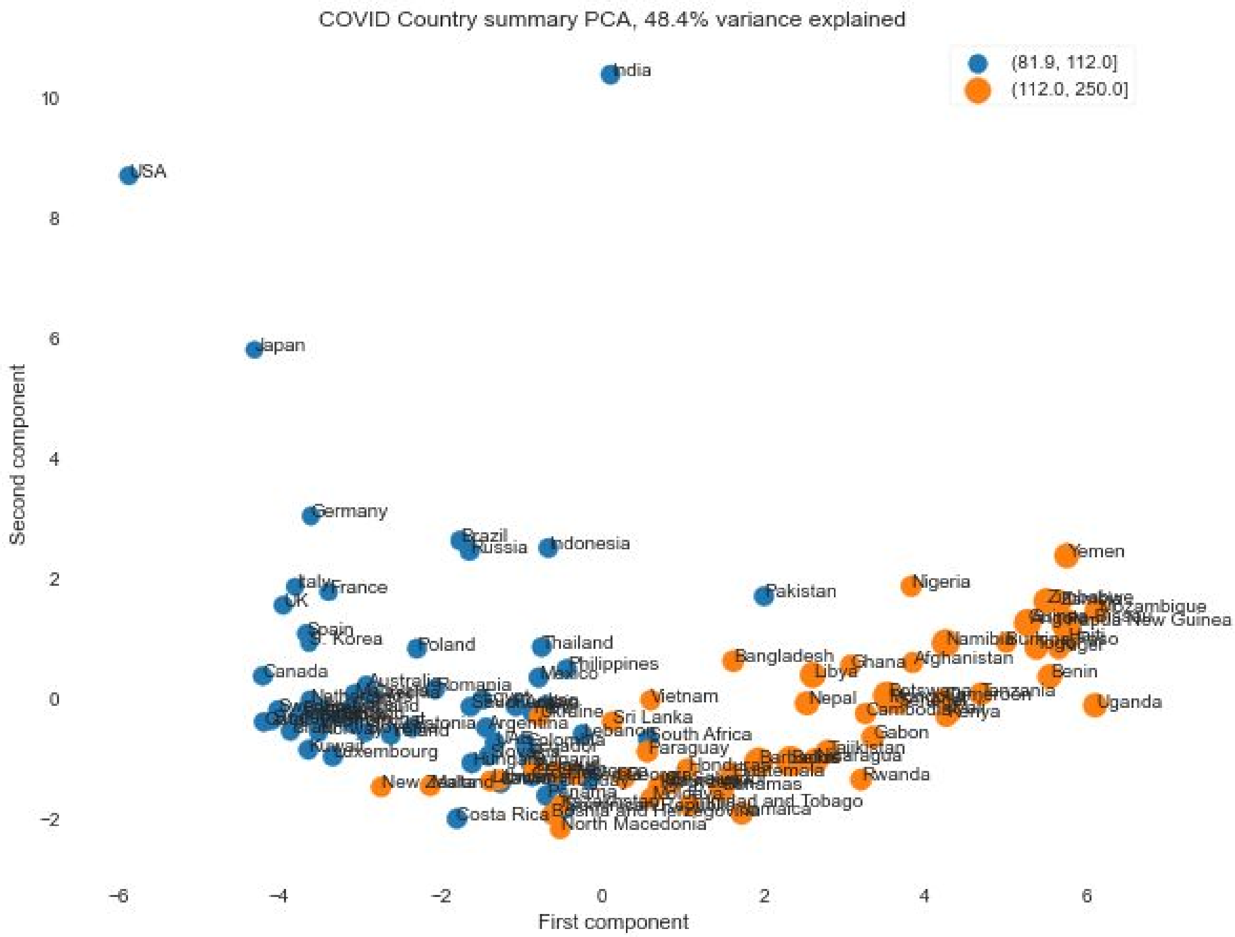


Figure 8: Scatter plot between the first two principal components with color coding by two quantile ranges of the day number.

Figures 7 and 8 show the distribution of countries based on the first two derived principle components. Additionally, grouping of each range can be observed in each plot where the day number data is divided into four ranges in Figure 7 and two ranges in Figure 8. There is some overlap of groupings in each plot indicating additional variance that is not explained by the first two principal components.

Figure 9 shows the highest impact coefficients for predicting day number.

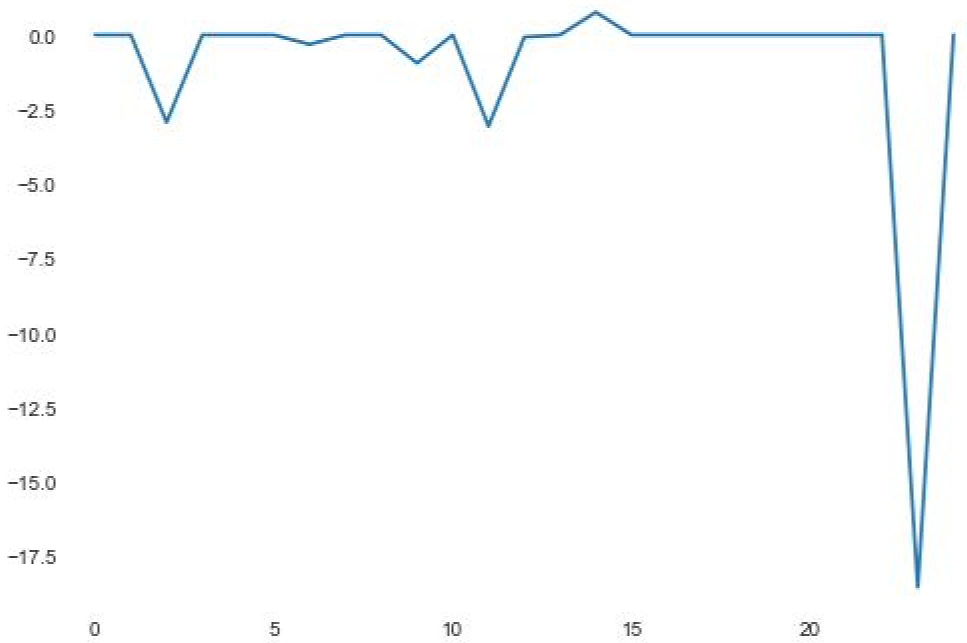


Figure 9: Lasso Coefficient Visualization of Highest Impact Coefficients.

By looking at the greatest value coefficients, the subset of predictive features is clear. Moving forward, these features will be utilized.

For the complete exploratory data analysis process, see the Jupyter notebook.

Modeling

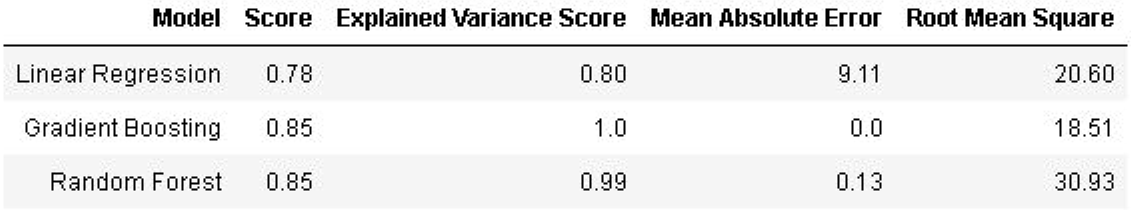
Data preprocessing for modeling includes the scaling of data for standard magnitude and creation of dummy features. In this dataset, all features are treated as numeric so no dummy features are required. Data was scaled using the MinMaxScaler function with a range of -1 to 1. Later investigation showed that scaling the data resulted in a similar modeling result as the unscaled data, therefore, scaling was neglected.

A categorical column of week categories was added to the dataset to ensure modeling options. This category contained 9 categories representing sets of weeks. Ultimately it was found that logistic regression depended upon the day number to predict the week category and without the day number, prediction was extremely poor. Thus, this method was quickly eliminated.

Data was split into a predictor variable dataframe (X) and a target variable dataframe (y). These dataframes were then split into train and test data for use in modeling.

The complete preprocessing and test/train data is available here.

Three model types were created to predict the day number feature: linear regression, gradient boosting, and random forest. Several variations of the gradient boosting and random forest models were tried in order to optimize parameters. Once optimal parameters were determined, these were utilized in final modeling. The summary of the results from these models are given in Table 1.



Model	Score	Explained Variance Score	Mean Absolute Error	Root Mean Square
Linear Regression	0.78	0.80	9.11	20.60
Gradient Boosting	0.85	1.0	0.0	18.51
Random Forest	0.85	0.99	0.13	30.93

Table 1: Summary of results from each model.

Based on the score and explained variance score, it would seem that linear regression should be ruled out since gradient boosting and random forest perform much better. However, given the mean absolute error of the gradient boosting model, it seems this model may be overfit. The random forest, while having a good score, has a much worse root mean square error, meaning the points are not well fit to the regression line. Since gradient boosting had the same score performance and better root mean square error, random forest can be eliminated.

To review the performance of the models a set of plots were created (Figure 10) to be compared to plots of the actual data (Figure 9). For each model, the frequency of predicted day numbers are shown. These prove to be relatively similar for all the models, however, there is definitely less similarity in the linear regression model. When looking at the True day vs the Predicated Day plots, it is very obvious that linear regression is the worst performer. Meanwhile, gradient boosting and random forest both perform similarly to Figure 9--the ideal case.

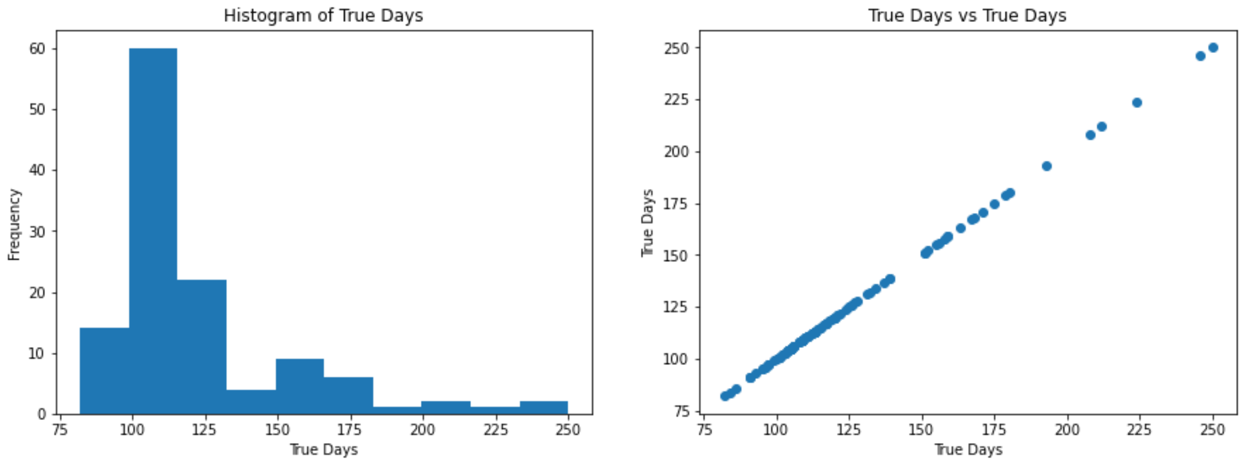


Figure 9: Real data plots to visualize best case scenario.

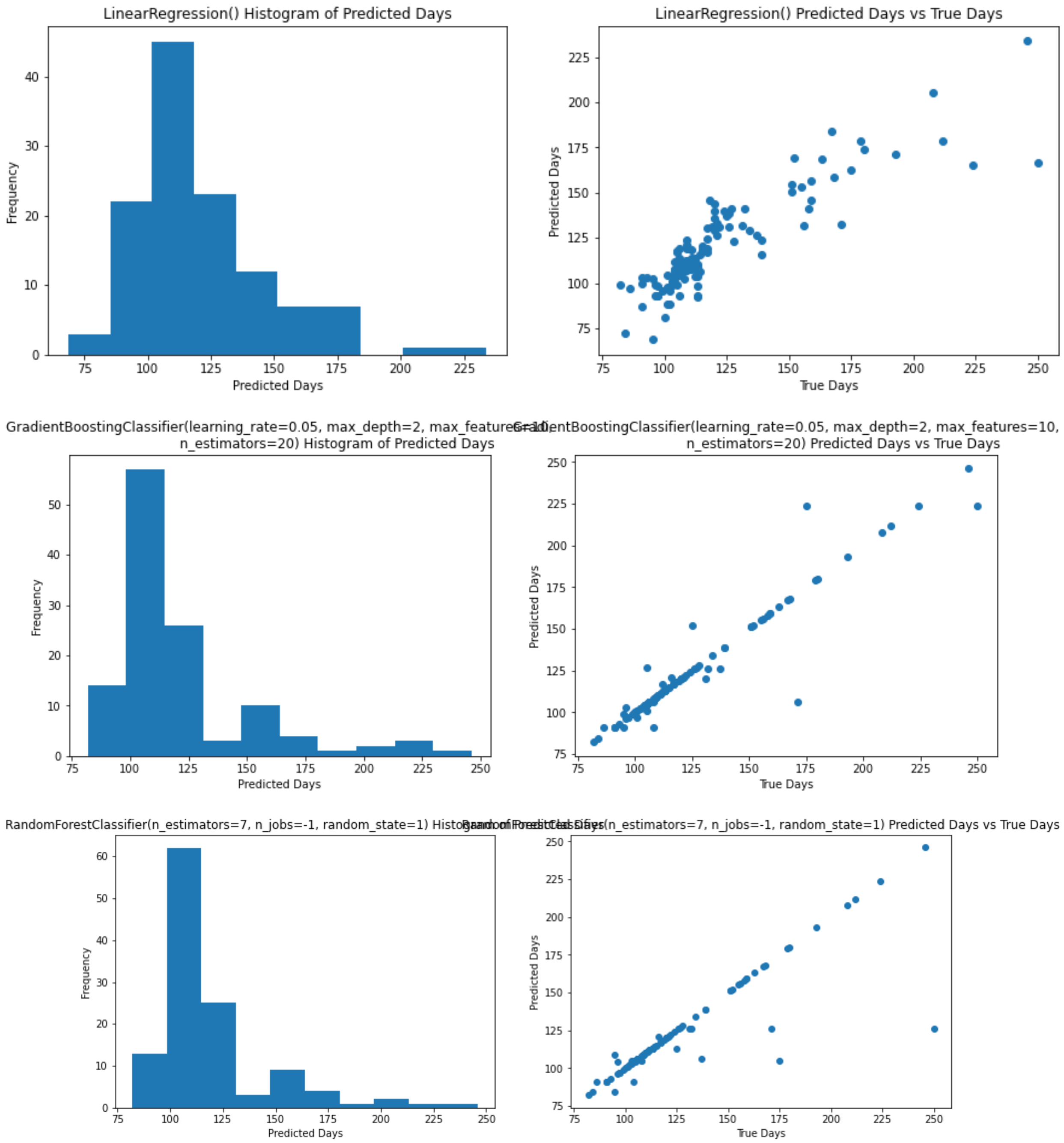


Figure 10: Plots of the each model type, showing linear regression is the worst predictor and gradient boosting is the best predictor.

The best model proves to be the gradient boosting model since the misclassification errors it does have are much less extreme than the misclassification errors in the random forest model. Ultimately, the choice between gradient boosting and random forest is a difficult one, since they perform similarly, but the root mean square error and plots show that the gradient boosting makes less extreme errors.