Group Name: #ItsNotBadWork #ItsBadData

Team Members: Jess Bunnag (tb2817), Clara Kim (sk9615), Tiffany Lin (tl3493), Anna Zuo (az2575)

Motivation:

Our group is interested in exploring which factors contribute to the rise of COVID-19 cases in each U.S. state. We are particularly interested in 1) whether there is a difference in COVID-19 cases for states by political affiliation, 2) whether we can classify a state's political affiliation using state metadata and COVID-19 cases, and 3) whether there is a difference in the number of new COVID-19 cases based on travel frequencies.

To explore these questions, we will use datasets on COVID-19 cases, trips traveled by and metadata of each U.S. state to conduct hypothesis testing and classification methods. And to ensure consistency across our datasets, we will use a common time frame from 2020/01/01 to 2020/12/31.

Datasets:

Dataset 1: United States COVID-19 Cases and Deaths by State over Time[1]

- **Provider**: Center for Disease Control and Prevention (CDC)
- **Description**: The dataset contains daily records of COVID case and death counts for each U.S. state and territory.
- **Rows 0-41759**: Information by date (1/22/20 - Present)
- **Column 1**: Submission date (date case counts were logged)
- **Column 2**: State code
- **Column 3-7**: COVID case counts (total, confirmed, probable, new, new probable, etc.)
- **Column 8-12**: Death counts (total, confirmed, probable, etc.)
- **Column 13**: Created At (when the data was recorded)
- **Column 14-15**: Consent Information (states' consent to including confirmed and probable case counts)

Dataset 2: COVID-19 State Data[2]

- **Provider**: Kaggle
- **Description**: The dataset contains aggregated COVID case data per state, along with state demographic and health information
- **Rows 0-50**: Information for each state (50 U.S. State, District of Columbia)
- **Column 1**: State (state name)
- **Column 2-4**: COVID information (aggregated counts over time for tested, infected, and deaths)
- **Column 5-7, 9-13, 19-26**: State Demographic Information (population, population density, Gini Index, Income, GDP, Unemployment, Sex Ratio, Smoking Rate, Pollution, Med-Large Airport Count, Temperature, Precent Urban, Percent Age Groups (0-25, 26-54, 55+), and School Closure Date)
- **Column 8, 14-18**: Health Information (ICU Beds, Flu Beds, Respiratory Deaths, Physicians, Hospitals, Health Spending)

Dataset 3: Blue and Red States[3]

- **Provider**: GKGIGS
- **Description**: political standing of the 50 states during the 2020 presidential election
- **Rows 0-49**: Information on each State
- **Column 1**: State name
- **Column 2**: State code
- **Column 3**: Political standing of the states either blue/democratic or red/republican

Dataset 4: State Trips[4]

- **Provider**: Bureau of Transportation Statistics (BTS)
- **Description**: Mobility statistics captured by a mobile device data panel on how often people stay at home versus how often they travel
- **Rows 0-53805**: Information by date (2019/01/01 - 2021/11/20)
- **Column 1**: Indexing
- **Column 2**: Level of travel either state or national
- **Column 3**: Date
- **Column 4**: State code
- **Column 5-6**: Population staying at home and population not staying at home
- **Column 7**: total number of trips
- **Column 8-17**: Counts for different trip frequencies (< 1, 1-3, 3-5, 5-10, 10-25, 25-50, 50-100, 100-250, 250-500, >= 500)
- **Column 18**: Unique Row Identifier
- **Column 19**: Week number
- **Column 20**: Month number

Question 1

To begin, we tested whether there is a difference in the number of COVID-19 cases between blue and red states. As per the media, we suspected that there would be a difference in the number of COVID-19 cases based on the state's political party. However, like every jurisdiction, we assumed innocence initially, thus our null hypothesis was that there is no difference in the mean number of COVID-19 cases between blue and red states.

Before testing our hypothesis, we performed data cleansing. We filtered the CDC data on COVID testing for each state by their political party in order to form the two populations: democratic and republican states. Then we aggregated all the daily new cases in 2020 for each state within the two groups respectively. Due to an imbalance of population across states, we normalized the aggregated cases by the population of each state. We then proceeded to graph and compare the calculated ratios for each state's COVID case to observe whether there is a difference in the mean of COVID cases between blue and red states.

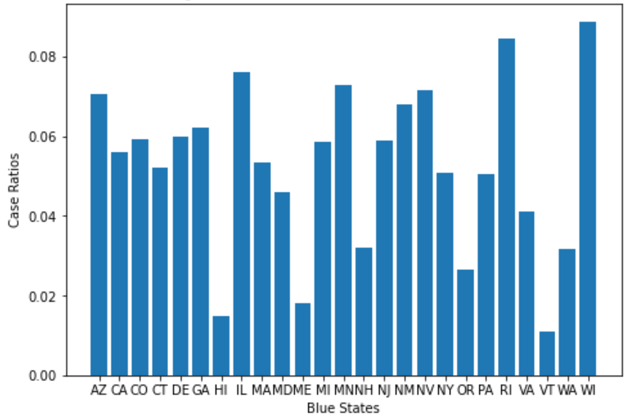


Figure 1: Democratic/Blue State Case Ratios

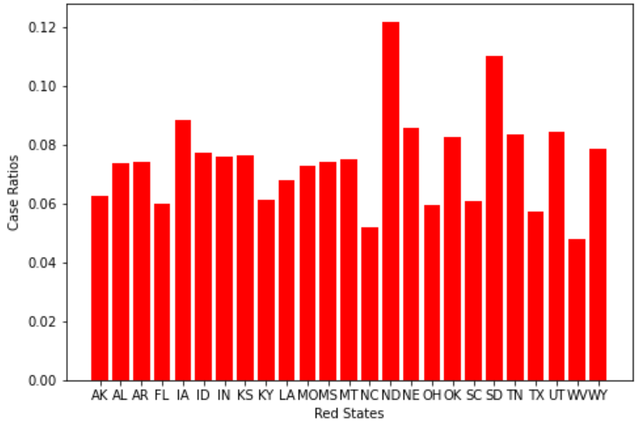


Figure 2: Republican/Red State Case Ratios

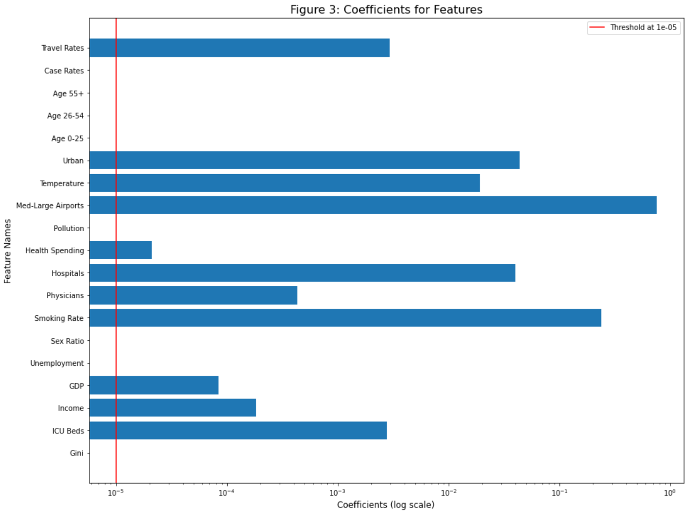
Figure 1 and **Figure 2** show the COVID case ratios for each state and you will notice in the blue population there are more divots in the bar graph compared to the red population. On average, the highs in the republican states were larger than the democratic states and the lows in the democratic states were lower than the republican states.

Most importantly we ran an independent t-test on the two groups to confirm our observations. The p-value of the test result was 0.0001655 which is smaller than our set alpha=0.05. Therefore, the observations are inconsistent with chance and we can reject the null hypothesis that there is no difference in the mean number of COVID-19 cases between blue and red states.

Question 2

Next, we wanted to further explore the differences between blue and red states through classification. From our results in Question 1, we presume COVID-19 case rates could be a strong indicator of a state's political affiliation. We were also interested in exploring how travel rates and various demographic and public health factors could be used to predict whether a state is red or blue.

So, we decided to build classifier models to predict a state's political affiliation based on the relevant features that are available in Datasets 1, 2, and 3. Before building the models, we performed data cleaning and feature engineering. We aggregated the daily total case count per state as well as the daily trips taken per state for all of 2020, and combined these with the state political, demographic, and public health data. Then, similar to in Question 1, we normalized the aggregated case counts and aggregated trip counts by the state's population to obtain the case rate and travel rate per state.



Using this cleaned data, we built three different classification models to select for the best performing model given our data. In each model, we took whether a state is blue or red as the label (dependent variable), and used our remaining relevant columns as the features (independent variables). We initially eliminated irrelevant columns such as 'Deaths', 'Flu Deaths', etc., since we are focusing on the number of COVID-cases and not deaths. We further eliminated 'Population' since it was already highly correlated with other features such as 'ICU Beds', 'Physicians', 'Med-Large Airports', and was also used to normalize 'Case Rates' and 'Travel Rates'.

Model 1, Logistic Regression: we began by building a logistic regression for binary classification. Before building the model, we selected the features with the highest coefficients in our regression. Using Lasso regression, we removed 8 redundant features for which the coefficient values dropped to zero, and kept only the 11 features with coefficient values above our threshold (**Figure 3**). The selected features for our model were: ICU bed count, income, GDP, smoking rate, physician count, hospital count, health spending, count of medium to large airports, temperature, urban percentage, and travel rates. After selecting our features, we used stratified K-folds cross-validation with five folds to train and test our model.

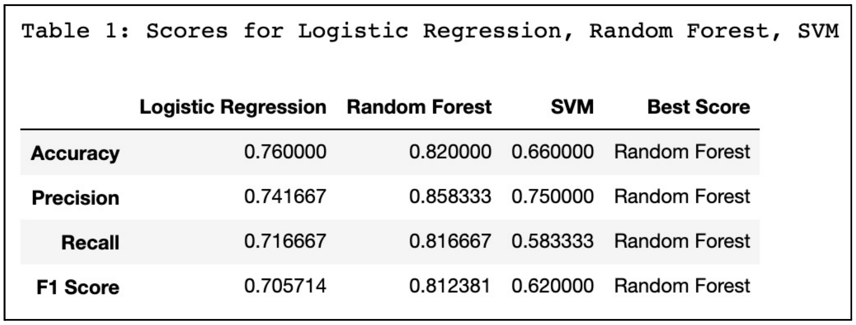


Table 1: Scores for Logistic Regression, Random Forest, SVM

	Logistic Regression	Random Forest	SVM	Best Score
Accuracy	0.760000	0.820000	0.660000	Random Forest
Precision	0.741667	0.858333	0.750000	Random Forest
Recall	0.716667	0.816667	0.583333	Random Forest
F1 Score	0.705714	0.812381	0.620000	Random Forest

Model 2, Random Forest Model: Next, we built a random forest model to see if it could be a stronger classifier. As with the logistics regression model, we used K-folds cross-validation with 5 folds to build a random forest model with 100 estimators.

Model 3, SVM Model: Finally, we built a SVM model to see if this could be a better separator of our classes. As with the other models, we used K-folds cross-validation with 5 folds to build a random forest model with 100 estimators.

To evaluate and compare the performance of our three models, we began by assessing the accuracy, precision, recall, and F1-score averaged across all K folds for each model (**Table 1**). As shown in our comparison table, the random forest model scored the highest for all metrics, indicating that the random forest had the strongest performance amongst the three models.

Figure 4: Cumulative Confusion Matrix for Logistic Regression

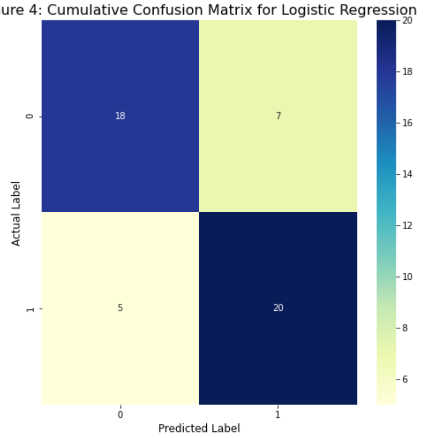


Figure 5: Cumulative Confusion Matrix for Random Forest Classifier

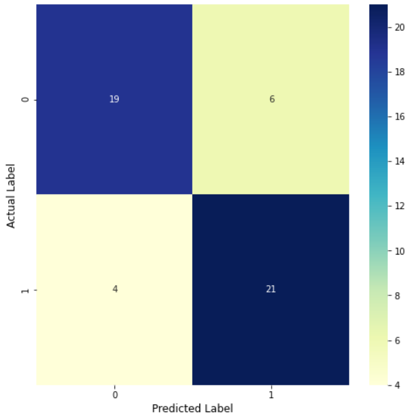
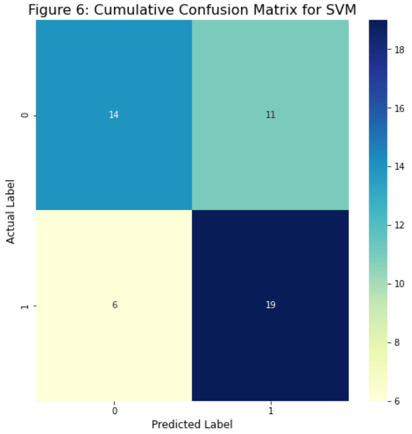


Figure 6: Cumulative Confusion Matrix for SVM



To evaluate the performance of each model further, we looked at the confusion matrices. We created each confusion matrix by summing the number of true positives, false positives, true negatives, and false negatives for each of the K folds (**Figure 4**, **Figure 5**, and **Figure 6**). We observed that both the logistic regression and random forest model had high true positives (blue states correctly classified) and true negatives (red states correctly classified) with lower numbers of false positives (red states classified as blue states) and false negatives (blue states classified as red states). The random forest model performed slightly better, with one fewer false negatives than the logistic regression. The SVM model performed more poorly than the other two models, with significantly higher counts of false positives, and fewer correct predictions.

Finally, we decided to plot the ROC curves for each model, which is a plot of the true positive rate versus the false positive rate for each possible classification decision threshold. We plotted the ROC for each of the 5 folds for the three models, and referenced the average AUC as an additional evaluation metric.

Figure 7: ROC Curve (Logistic Regression)

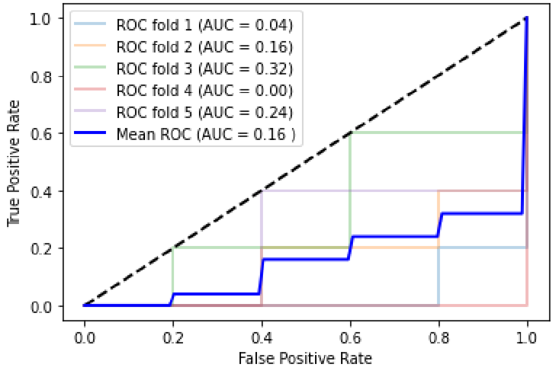
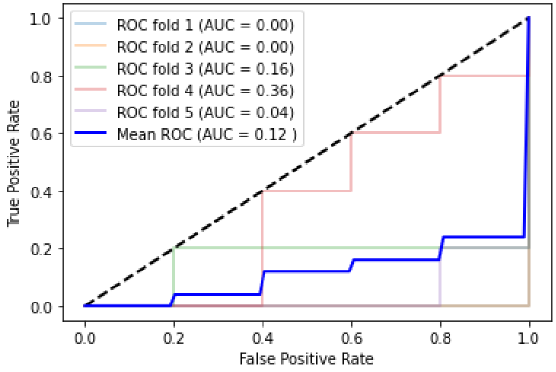


Figure 8: ROC Curve (Random Forest Classifier)



AUC is a measure of how well the classifier separates class, and the mean AUC metrics were relatively low for all three models, at 0.16 for logistic regression, 0.12 for random forest, and 0.59 for SVM (**Figure 7**, **Figure 8**, and **Figure 9**). However, since our dataset is balanced between our two classes (red vs. blue state), traditional metrics like accuracy, precision, recall, and F-1 score should be sufficient for evaluating the performance of our models. Hence, as per our traditional metrics, our best performing classifier of whether a state is red or blue was the random forest model.

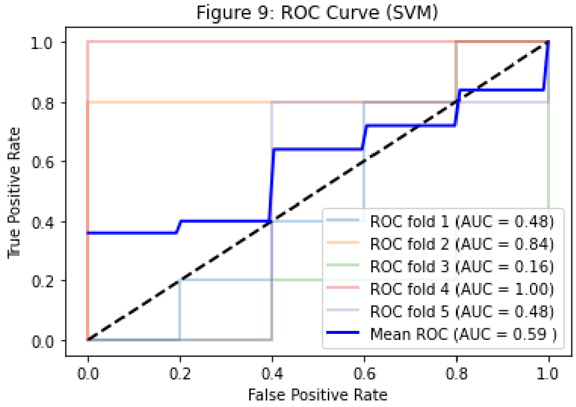


Figure 9: ROC Curve (SVM)

Question 3

Lastly, we wanted to test whether there is a difference in the number of new cases between more traveled states and less traveled states in 2020. Specifically, we will test whether the more traveled states have more new cases (per population). We define more traveled states as states having greater number of trips per unit of population. As in Question 1, our null hypothesis is that there is no difference in the number of new cases between the two state groups, and the alternative hypothesis is that the more traveled states have more new cases than the less traveled states.

The data preprocessing for this test is similar to that in Question 1. Since there are no missing values in the data, no imputation method was needed. However, since we are only interested in the data in 2020, we filtered data only from that year. Because the number of new cases and the number of trips hypothetically depend on the population, we also normalized both the number of cases and the number of trips by population.

In order to determine the more traveled and less traveled states, we performed a median split on the number of trips (per population) by state. **Figure 10** and **Figure 11** illustrate the number of cases by state in the two groups. We can see that there are more states in the less traveled states group with lower case ratio than in the more traveled states group.

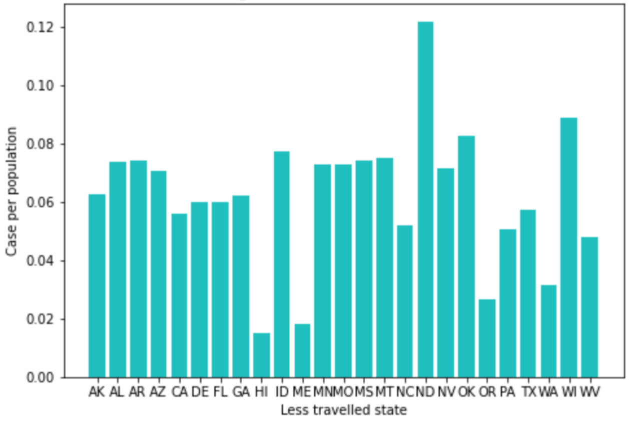


Figure 10: Less travelled state

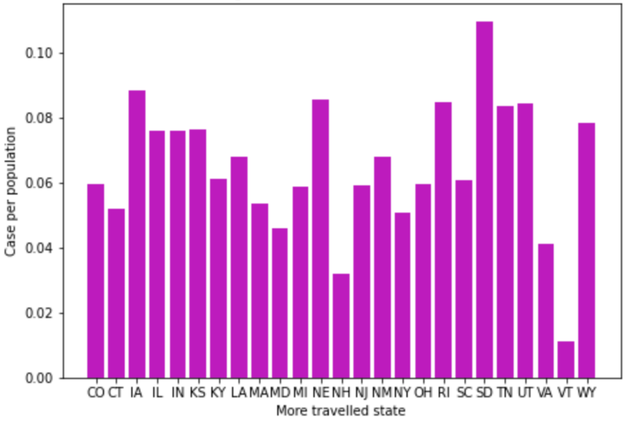


Figure 11: More travelled state

To test our hypothesis, we ran a one-tailed independent t-test. The p-value of the test result was 0.0242 which is smaller than alpha value=0.05. Therefore, it suffices to reject the null hypothesis and conclude that the more traveled states have higher numbers of new COVID-19 cases per population than the less traveled states.

Conclusion

In summary, we found that there is a difference in the number of COVID-19 cases between blue and red states, random forest performed best in classifying political affiliation using COVID-19 rates and other factors, and states with high travel frequencies have more COVID-19 cases than those with low travel frequencies. Overall, our analysis showed that the number of COVID-19 cases was a fine factor in distinguishing states.

From this project, we learned that certain datasets will not always provide sufficient information to address the question at hand. Since we took several different datasets, and combined the information for our analyses, we also brought in potentially irrelevant data, which led to an increase in the amount of data manipulation needed. Also, some of our datasets had potentially inaccurate information - for example, Dataset 1 contains COVID-19 case count data that could be inaccurate due to potential delays in testing and reporting, differences in how states report cases, infected individuals not getting tested or seeking medical care, etc. Additionally, since our datasets come from different sources, the information between sources may be inconsistent. For example, the date on which the number of trips was recorded in Dataset 4 might not align with the date on which the number of cases was reported in Dataset 1. Also, due to the delay in testing and reporting of COVID cases, the reported number of cases on a given date could have resulted from travel patterns and frequencies from a previous date. Lastly, there were inherent limitations to our problem structure - for example, in Question 2, we structured the problem as classification on the state level based on U.S. state-specific information. This limited the number of data points available to the number of states (50), which constrained us to a small number of data points for training and testing our classification model.

Overall, our biggest takeaway from this project was the importance of the datasets. In the future, we will select a sufficiently large and consistent dataset so that our analysis is representative of the true population.

References

1. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36
2. https://www.kaggle.com/nightranger77/covid19-state-data
3. https://www.gkgigs.com/list-of-blue-states-and-red-states/
4. https://www.kaggle.com/ramjasmaurya/trips-by-distancefrom-2019-to-nov-2021?select=State_trips.csv