COVID-19 Exploration - Impact of Government Decisions and Adherence on Death Rates (Progress Report I)

Linwood Hall halll05@students.ecu.edu
Ping Wang wangp19@students.ecu.edu
Brian Crawford crawfordbr20@students.ecu.edu

1 Introduction

Our final project will be analyzing various data sets around COVID-19 and the global pandemic. Specifically we will attempt to analyze the impact government decisions influenced the mortality rates for North Carolina. The overall goal of this project will be to show through big data analysis what impact various government decisions had on overall infection and mortality rates. The results of this project can help influence government decisions in the future and will enable leadership to make more informed and better decision on what restrictions or other actions should be imposed to limit the spread and mortality of a deadly disease.

2 Dataset Acquisition

The dataset we will be using is from two sources:

1. The COVID Tracking Project
2. COVID-19 State and County Policy Orders

The COVID Tracking Project. This organization is a volunteer based effort designed to collect and publish COVID-19 data. Sponsored by *The Atlantic*, this organization is dedicated to providing accurate data through daily validation and cross-checking.

COVID-19 State and County Policy Orders. This dataset, from Data.gov, contains 4220 records with 10 variables. The information contained within this dataset includes the policy type, as well as start/end dates. We took information regarding **state level** policies for North Carolina (Shelter In Phase, Phase 1, and Phase 2), and created a discretized variable (**policy**) within our processed dataset.

The North Carolina dataset contains daily information on the COVID-19 pandemic. As of 3/6/21, there were 366 non-duplicated records with 42 variables. Some of the variables included are:

1. death: total deaths during this time period
2. deathConfirmed: death that is confirmed to be COVID-19 related
3. positive: total number of probable cases reported
4. totalTestResults: a metric describing viral testing results

There are many more variables to consider, but the previous list gives a general summary of what the dataset contains.

This dataset is adequate for reporting deaths and testing results. However, there was some missing data in variables such as:

1. hospitalized
2. recovered
3. negativeTestsAntibody
4. positiveTestsAntigen

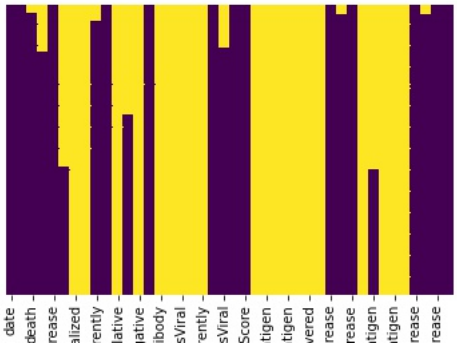

Figure 1: NaN Values Per Variable

The reason for this missing data is largely due to reporting standards being inconsistent across each state, as well as inconclusive science on antibodies. For the purposes of this project, these fields have been dropped. **Figure 1** references a selection of fields with 'NaN' values, colored in yellow.

Examining the remaining variables shows that deaths are negatively correlated to the presence of a government policy (fig. 2), such as a 'Shelter-in-place' order, or a 'Phased' reopening approach.

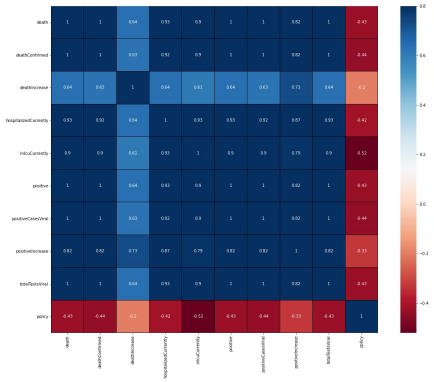

Figure 2: Correlation between COVID-19 Variables

3 Initial Results

3.1 Distribution Check

In order to build models to predict deathConfirmed and deathIncrease, we first check the distribution of dependent variables.

As shown in **Figure 3** and **Figure 4**, both deathConfirmed and deathIncrease are right skewed and not following normal distribution. For this reason, we tried to build a model with decision tree regression.

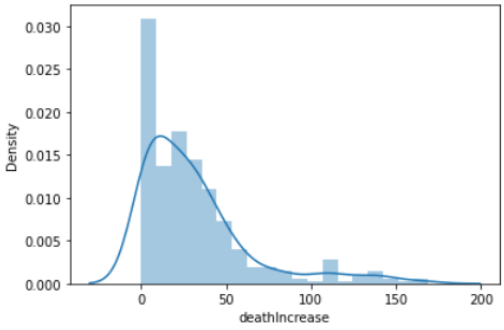

Figure 3: Distribution of Death Increase

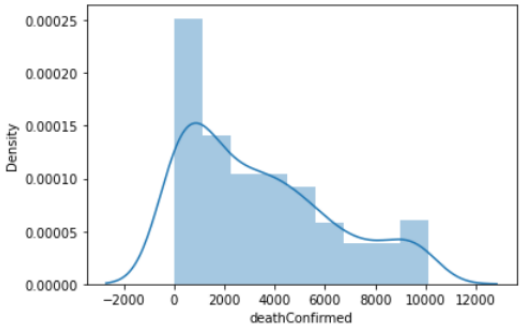

Figure 4: Distribution of Death Confirmed

Features included as independent variables: policy,totalTestsViral, positiveCasesViral, inIcuCurrently

Predicting deathConfirmed

The R-square value of this model is -3.44, indicating that this model is not appropriate for our dataset.

3.2 Predicting deathIncrease

The R-square value of this model is -1.69 indicating that this model is not appropriate for our dataset.

For the next step, we will transform the response variables with variable transform methods(1/x transformation) for normalization purpose. Also, more feature selection methods(Boruta) will be applied to better decide features included in the model.

4 Issues

So far, most of the difficulty with this project has been the data preprocessing. Finding a complete dataset with the inconsistent COVID-19 reporting procedures has been difficult. Even with the cross-referenced data provided by The COVID Tracking Project, there were several variables with no data. This also creates challenges in visualizing the data as it's required to be formatted in a way that requires transformation to visualize correctly.

Additional issues include figuring out how to visualize the data in a meaningful way, and figuring out how to tell the story we want. As a result of this we have created a number of mock-ups to play with how we can visualize and create an interactive view for the data and insights garnered from the machine learning algorithms.

So far, our group has been in agreement on examples provided, and the work has been divided evenly.

5 Milestones & Schedule

All milestones will be prepared together as a team. This ensures all parties are aware of the most up to date models and progress we are making as individual sub-tasks will be distributed to individual members.

The major items we wish to complete are listed below.

1. ~~Search and Collect initial data (9 days)~~
2. ~~Analyze, pre-process, clean and build database (5 days)~~
3. ~~Normalize the data, encode as needed, and prepare the data for modeling (3 days)~~
4. ~~Build initial models (4 days)~~
5. ~~Build initial Dashboard Mock-ups (4 days)~~
6. ~~Prepare and Submit Progress Report I (2 days)~~

   $\surd Milestone$: $ProgressReportI(March11, 2021)$
7. Continue to refine and build additional models (14 days concurrent)
8. Build and experiment with visualizations (14 days concurrent)
9. Experiment and build user interface (14 days concurrent)
10. Prepare and Submit Progress Report II (2 days)

    $\surd Milestone$: $ProgressReportII(March25, 2021)$
11. Finalize models, complete error checking (15 days)
12. Prepare database and visuals for final report (15 days)
13. Prepare Final Report (6 days)
14. Run through final report, refine and adjust (4 days)

$\surd Milestone : FinalReport(April20, 2021)$

6 Data Visualization Mock-Ups

The ideal solution for a data visualization or set of data visualizations would be ones that are clear and concise and convey the results of the machine learning algorithm applied as well as enabling the user to explore additional policies and the outcome on infections and deaths.

This is proposed to be achieved utilizing Tableau Public and will be accessible to anyone with internet access.

You can view the examples and interact with the current versions by going to the hyperlink linked in the caption under the mock-up.

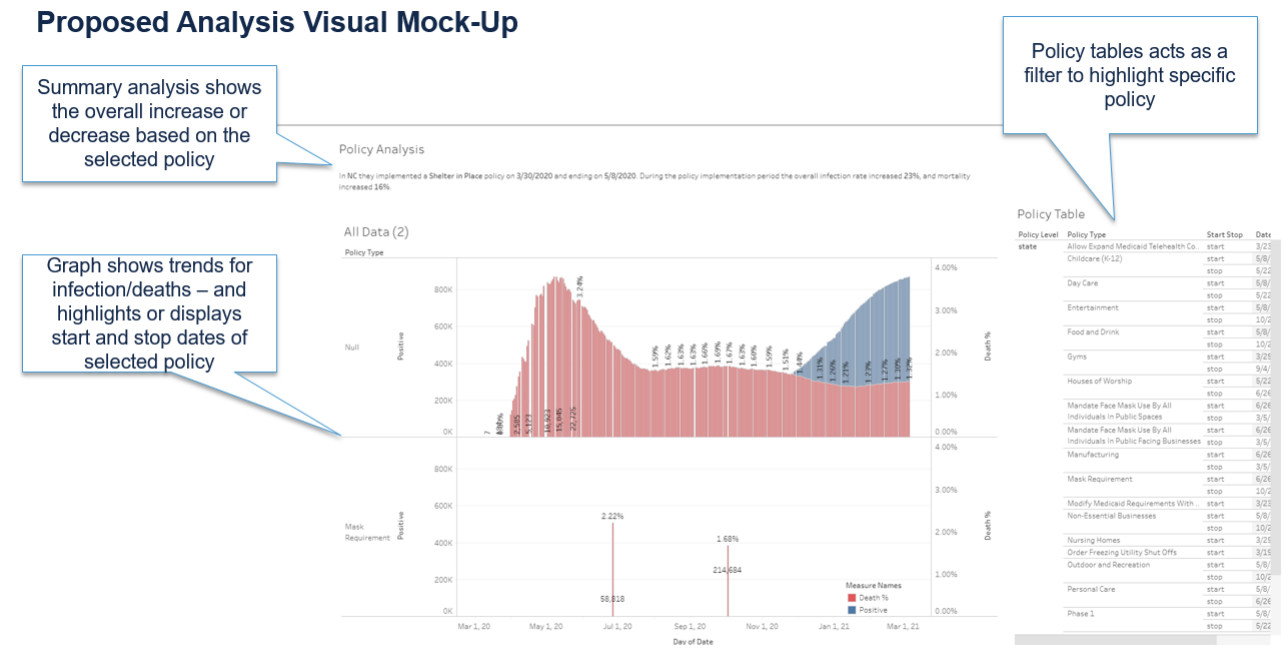

Figure 5: Link: Dashboard Mock-Up 1

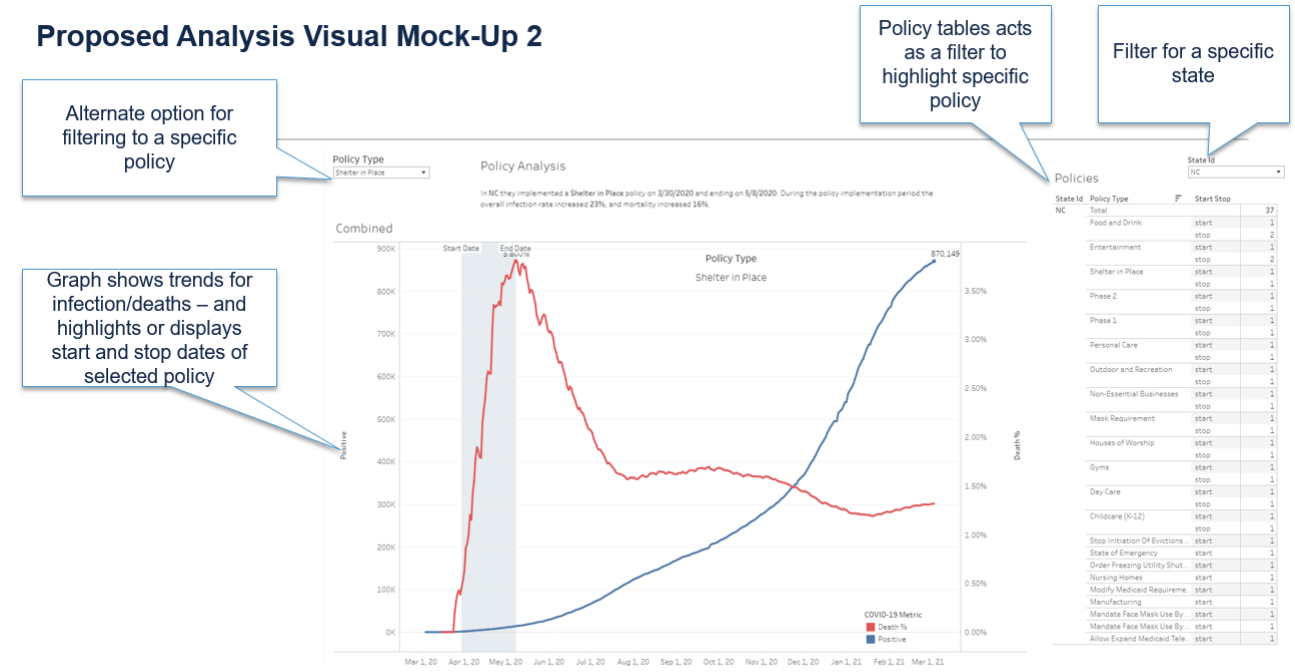

Figure 6: Link: Dashboard Mock-Up 2

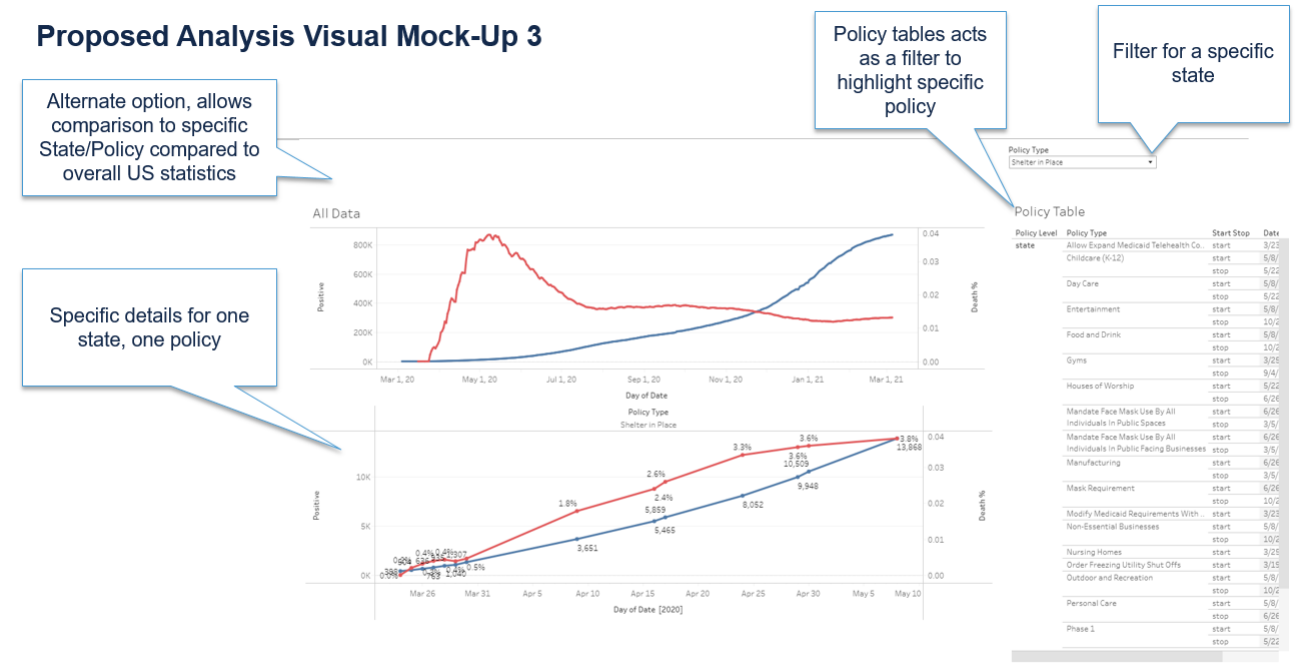

Figure 7: Link: Dashboard Mock-Up 3