**Part 2: Washington Mobility Changes during COVID-19 Final Report**

Kara Shibley + Griffin Reichmuth

**Research questions and Results**

1. What factors are most important in determining whether or not a county has made significant changes in retail/entertainment nonessential travel following the announcement of the Washington stay at home mandate?

We found that which candidate a county voted for in the 2016 presidential election was one of the most important factors in determining whether or not a county has made significant changes in nonessential travel during the COVID-19 pandemic. The other most relevant factor was who the county voted for in the 2016 gubernatorial election.

2. How have the different groups adjusted their retail/entertainment nonessential travel over time during the COVID-19 Pandemic?

We found that counties that voted for Trump tended to have a smaller reduction in non essential travel. Additionally, the counties that voted for Inslee tended to have a larger reduction in non essential travel.

3. How do mobility changes regarding the six location types in the data compare at the WA state level to the national average?

We found that Washington tended to have a similar reduction in the different categories of travel compared to the United States as a whole. The main two categories with differences are seen in Workplace travel, where Washington saw a smaller reduction in travel, and Park travel, where Washington saw much larger increases in the travel to these locations.

**Motivation and background**

COVID-19 is currently at the center of society, shaping people's current and future lives. Google map search history can indicate where a person is traveling. As this data compares search history percent changes from the baseline levels (based on median travel for that day of the week from Jan 3rd to Feb 6), we think this data can indicate the changes people are making in their travel in response to the current pandemic. As people travel from place to place, the virus is then given the chance to spread more easily which then increases the probability of the virus infecting a new host and increasing the magnitude of the pandemic. We think that this data can indicate which counties are following the stay at home order the best and will be the most diligent in stopping the spread of the virus. Additionally, we want to determine which factors most heavily influence how diligent a county is in stopping the spread of COVID-19, as well as determine if there are unchangeable factors (e.g. the dominant industry) that are skewing the travel data.

**Data Sets**

1. **Google maps mobility data:** Shows how visits and length of stay at different places change compared to a baseline. Calculated changes using the same kind of aggregated and anonymized data used to show [popular times](https://support.google.com/business/answer/6263531?hl=en) for places in Google Maps

Found at:

<https://www.google.com/covid19/mobility/>

1. **CDC Data on County Population Classification:** Classification defined by the US OMB and the USDA. There are 6 different classifications listed on the CDC site: large central metro (1), large fringe metro (2), medium metro (3), small metro (4), micropolitan (5), non-core (6)

Related data and classification scheme found at:

<https://www.cdc.gov/nchs/data_access/urban_rural.htm#Data_Files_and_Documentation>

Data under: <https://www.cdc.gov/nchs/data/data_acces_files/NCHSURCodes2013.xlsx>

1. **2016 Federal and WA state election Data:** Provided information on county voting for the 2016 Gubernatorial and 2016 Presidential race.

Found at:

<https://results.vote.wa.gov/results/20161108/Export.html>

Data:

<https://results.vote.wa.gov/results/20161108/export/20161108_AllCounties.csv>

1. **2017 USDA Ag Census Data:** Looks at a variety of farm related statistics, including how much farms are making and how large the farms are - separated on a county level.

Found at: <https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_2_County_Level/Washington/st53_2_0001_0001.pdf>

**Methodology (algorithm or analysis)**

ML section (coding intensive part of project):

* Get various feature data (e.g. industry, population, etc)
* Clean data, remove NaN rows when applicable, fix titles and delete unnecessary columns
* Combine data into one working data set comprised of the relevant features and movement related data
* Perform one hot encoding of categorical data. One hot encoding is a process by which categorical variables are converted into a form (1 for an affirmative variable, 0 for a negative variable) that could be provided to ML algorithms to do a better job in prediction.
* Run a decision tree classifier for feature selection (will produce the importance of all features). We can then interpret the features that the model deems most important as predictors for how a county will respond to stay at home orders. Meaning that the most important factor is the strongest predictor in how the county is reacting to COVID-19 legislation through their movement.
* Run the decision tree classifier to classify all counties based on the various features selected. This allows for counties with similar mobility prediction features to be analyzed together. Test the efficacy of this classifying by using a 70/30% train/test split of the data.
* Use a random forest regression model of 100 regression trees and the county level data grouped by selected features to predict movement within each county group. Use data from 2020-03-23 to 2020-04-13 as the training set and 2020-04-27 to 2020-05-18 as the testing set. Find the mean error of the model for each date

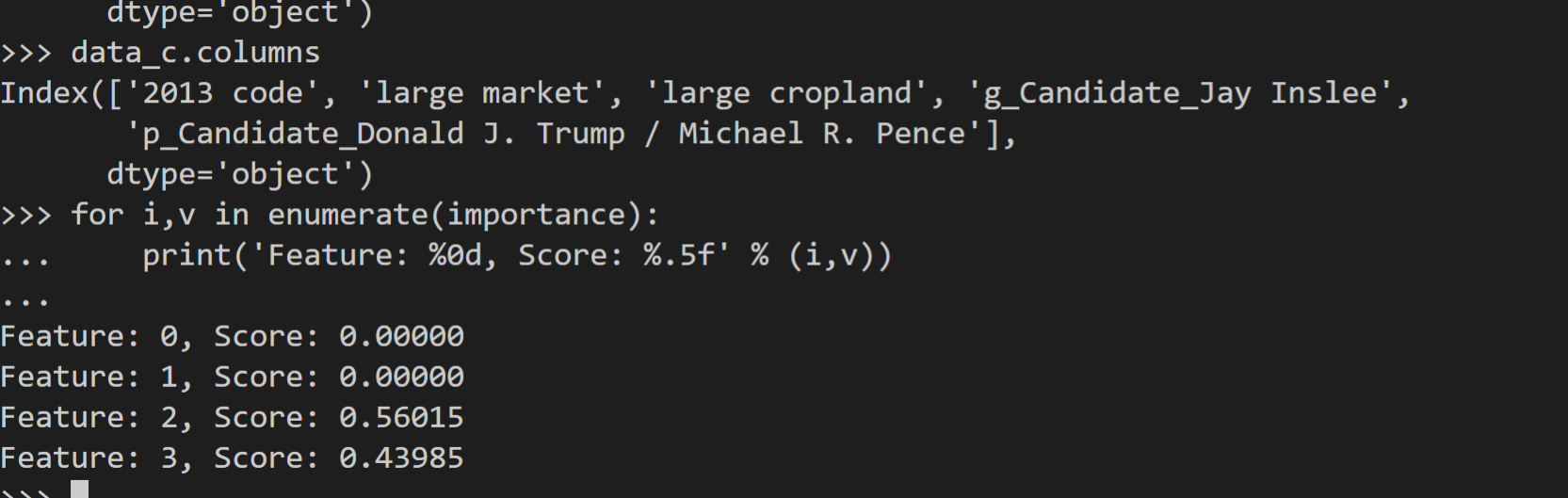
For Non ML section (additional analysis/visuals):

* Create a scatter plot using Plotly to visually show the percent changes in retail/entertainment nonessential travel over the course of the COVID-19 pandemic
* Create 6 subplots using matplotlib, one for each category of travel, and on each plot show the US percent change in travel vs the WA percent change in travel

**Results**

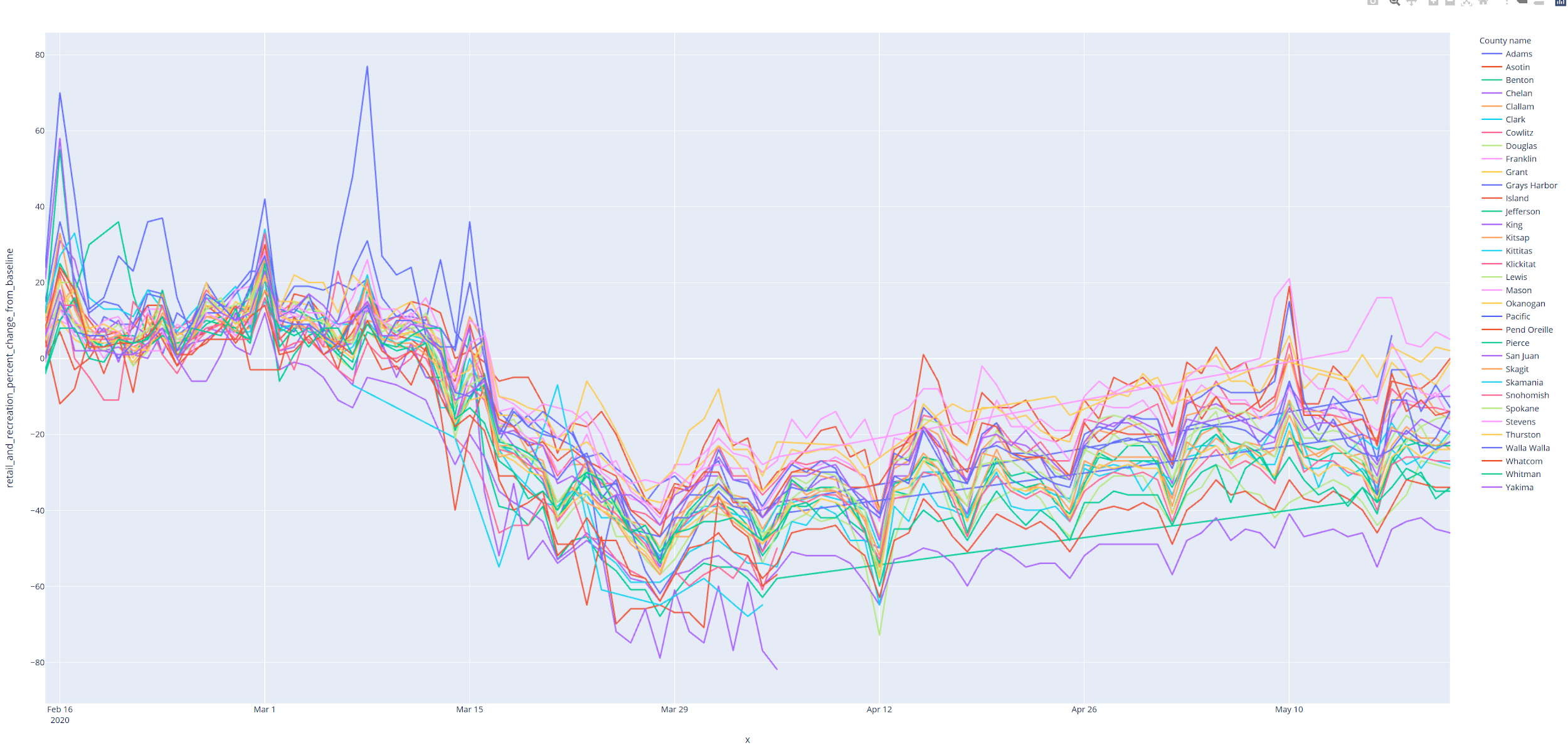
Our first question attempted to find a factor(s) or characteristic(s) that could be used to predict people's response to the Stay at Home mandate, but what we found was that responses to this crisis are nearly impossible to predict from the narrow data we could gather. As we were analyzing our models, we built several different decision trees and realized that due to the nature of the data, but there was only so much we could do with our data that would build a meaningful regression without overfitting the data. We then decided that no model was a great fit for this data, thus we would take creative liberties and use our models in different ways to try to gather any information we could. The US government puts counties into 6 different classifications, based on population density, so we decided to create a decision tree to attempt to classify what category a county is, based on the other features that we gathered. Our thought process behind this was that we could then determine which factors most heavily “make a county a county” - ie what decisions county members make, either political or in industry types pursued, our model deems most important in predicting the “type” of county.

We thought that counties would hold similar people, and that typically divisive categories like political views, dominant industries, and population thresholds could create a predictive model but in a time like this, individual human responses are not nearly as predictable as we thought. Our initial decision tree classifier models, created to determine the most important factors for a model aimed at classifying counties, found that the way a county leans politically is one of the most predictive factors. We expected dominant agriculture counties, having large numbers of essential agricultural workers, to be important in classification. But, our model found almost no predictive correlation from agriculture. Instead we found that the counties winning candidates for the 2016 presidential election and for the 2016 gubernatorial election were the two most important factors. We re-ran our analysis several times and consistently found despite the presidential and gubernatorial election results alternating which was more important, these factors were consistently the two most important factors.

**

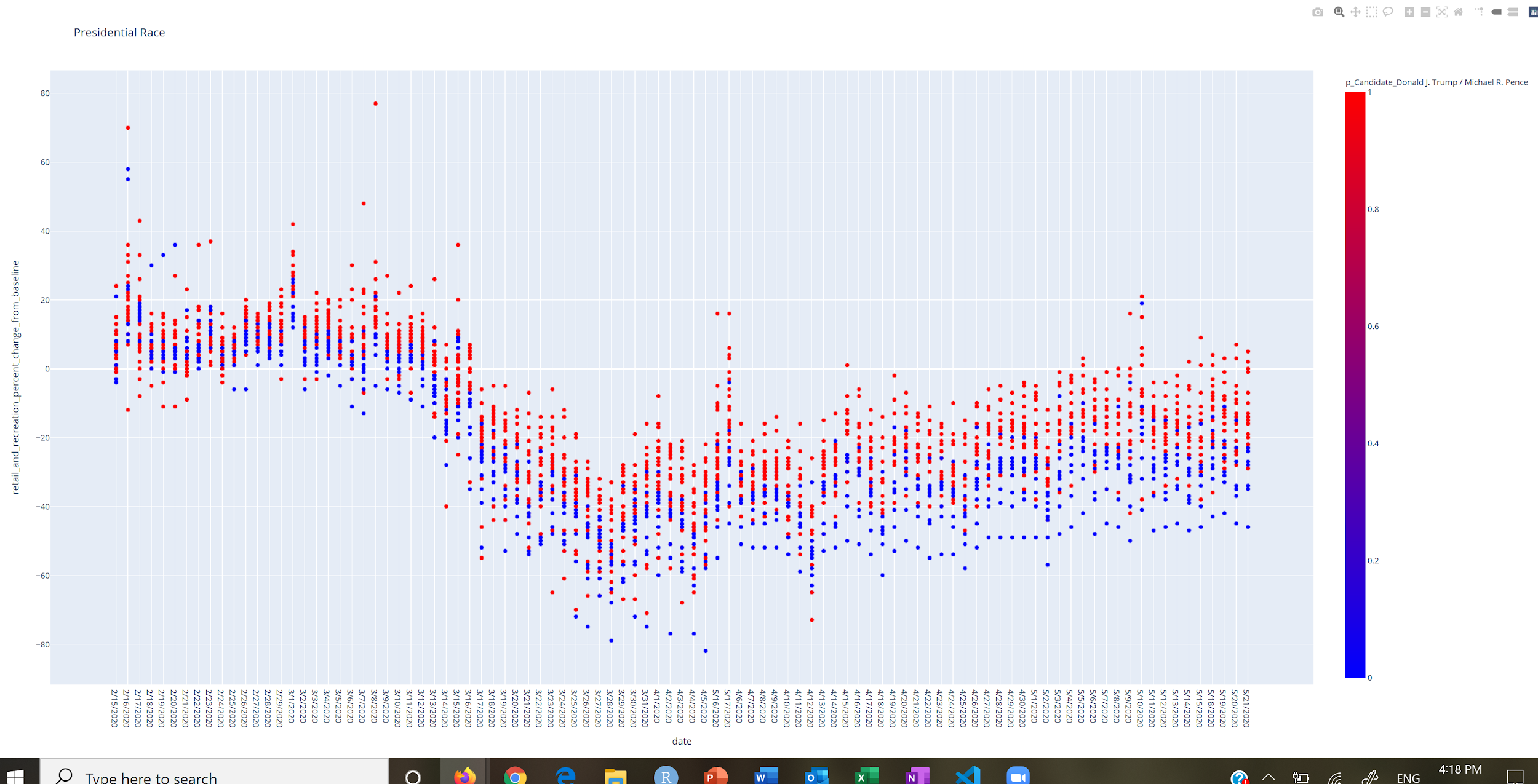
*Figure 1: The importance scores of the four main features, as they were used to predict “2013 Code,” showing that political county classifiers are most import*

In conclusion, we found that politics are one of the best predictors in determining a county’s response to the Stay at Home mandate, but these classifiers still produced a rather poor model at predicting movement. We saw an average error of about 18%, meaning that our model was 18% off on predicting daily movement in Retail/Recreation when trained on data from Mon March 23rd - Mon April 13th, and tested on data from Mon April 27th - Mon May 18th. Because of this huge error, we decided to graph the movement data for every county in Washington over the course of the pandemic, in hopes of finding a trend that explained this error.



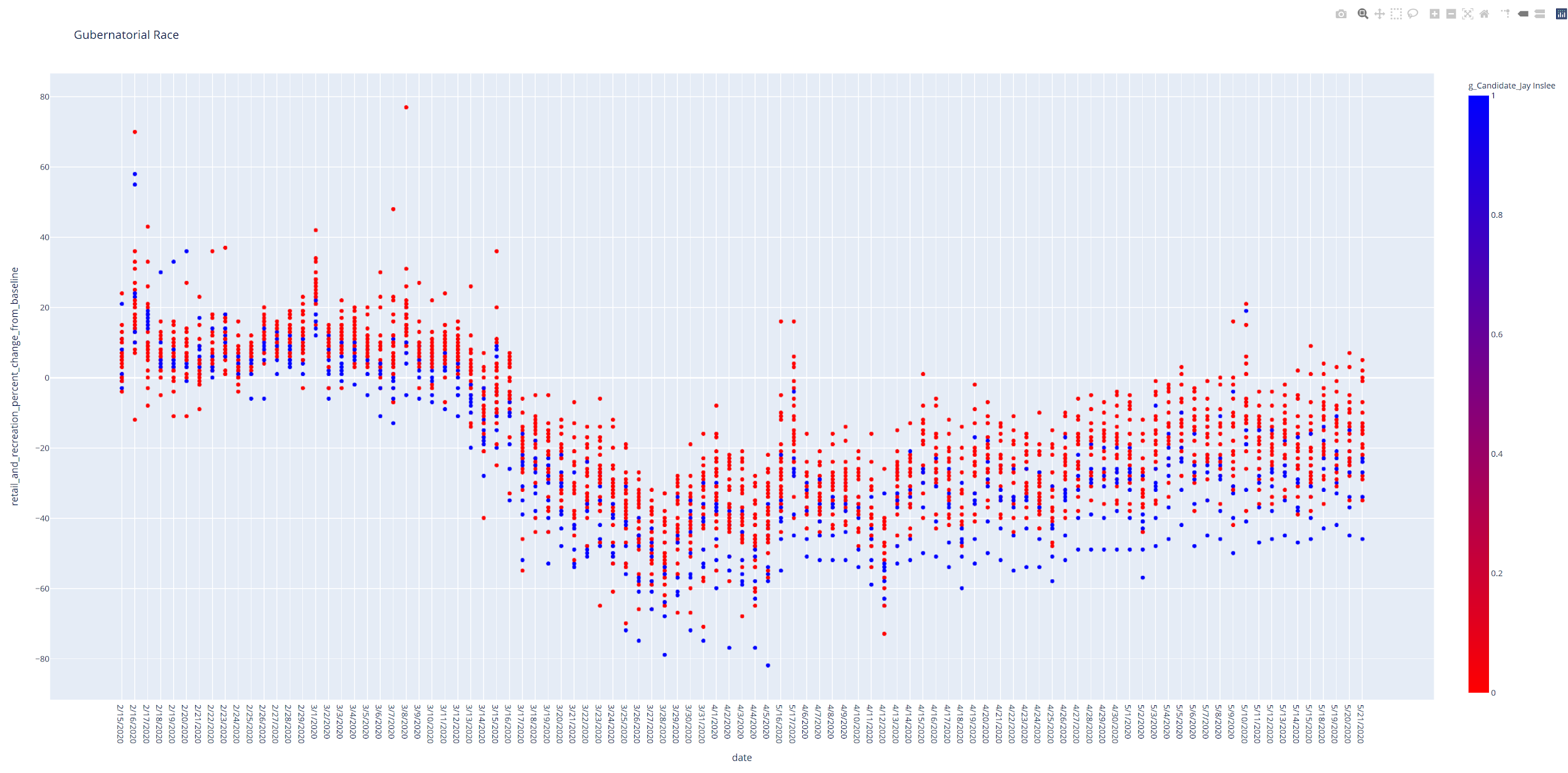
*Figure 2: Percent change in retail/entertainment travel over the course of the COVID-19 pandemic for every county in Washington.*

Based on the statewide movement trends, we attribute this error to the fact that regardless of dominant political party, every county started to see an increase in non essential travel a couple weeks into the stay at home order. We believe this increase in movement was due to the initial fear wearing off, and individuals getting fed up with being cooped up - not that this is any excuse to break the order, but rather humans are hard to control and value freedom and autonomy. This trend made our first plans for ML rather meaningless and inaccurate, so we pivoted slightly and focused more on the analysis of groups rather than predicting movement. As we initially planned, our second question built upon our previous analysis, using the most important factors we had previously found. As the 2016 presidential candidate was one of the most predictive factors, we explored the differences in the percent change in nonessential retail/entertainment travel between the counties that had versus the counties that had not voted for Trump. We found that counties that had not voted for Trump in 2016 had a larger decrease in retail/entertainment than those who did, but there was still significant overlap between the two groups.



*Figure 3: Percent change in retail/entertainment travel over the course of the COVID-19 pandemic, shown with counties that voted for Trump in 2016 in red and those who did not in blue.*

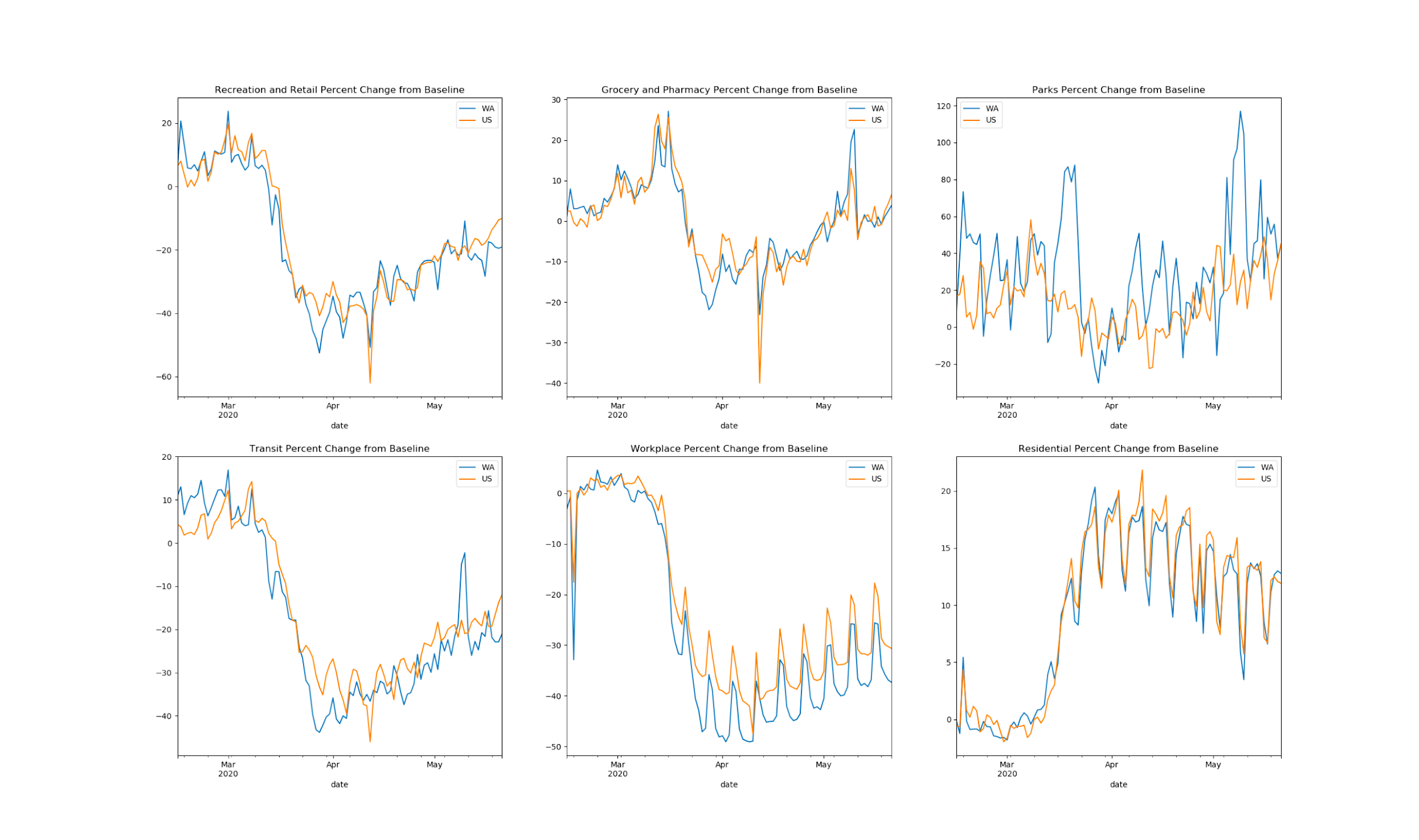
Additionally, we found that the gubernatorial candidate from 2016 was also an important factor in predicting county movement, so we also looked at the difference in percent change in nonessential travel for the counties that voted for Inslee in 2016 and those who had not. We found that the party trends demonstrated in the presidential election matched those in the gubernatorial election, with those who voted for Inslee showing a bigger decrease in travel than those who had not. This is unsurprising, as the majority of counties that voted for Trump in 2016 did not vote for Inslee, and vice versa.



*Figure 4: Percent change in retail/entertainment travel over the course of the COVID-19 pandemic, shown with counties that voted for Inslee in 2016 in blue and those who did not in red.*

Overall, we saw Washington counties could be further divided into political parties and that changes in travel in retail/entertainment over the course of the COVID-19 pandemic were distinct between the two groups. Republican Counties tended towards a smaller decrease in non-essential travel and Democratic Counties had a larger decrease in non-essential travel.

Our third question explored the differences and similarities between the way Washington has responded to the COVID-19 pandemic vs the way the country as a whole has responded, based on travel data in 6 different categories. We saw that Washington followed the same patterns as the United States as a whole in 4 of the 6 categories - Recreation and Retail, Grocery and Pharmacy, Transit, and Residential. One of the most significant differences was seen in Workplace Travel, where Washington had a smaller percent change from baseline levels which indicates that fewer Washingtonians have seen a difference in where they are working over the pandemic. This could be due to a variety of factors, but one hypothesis is that there is a significant number of essential workers in agriculture that some states would not have. Another area of significant difference from Washington compared to the US as a whole was that there was a much larger increase in travel to parks in the state of Washington. The country as a whole saw an increase in travel to parks starting primarily in April - likely due to the increasingly good weather and people looking for a safe and healthy activity to do during quarantine. Washington saw a much larger increase than the rest of the country, which may be due to the more “outdoorsy” reputation of many people in the state. Additionally, Washington is known for having a lot of outdoor recreational areas, which makes it easier for the majority of Washingtonians to access parks, where some states may have a more limited amount of outdoor recreation areas.



*Figure 5: Percent change in 6 different categories of travel over the course of the COVID-19 pandemic, demonstrating the changes seen in Washington (blue) compared to the rest of the country (orange).*

Overall, we found that with the limited county level data that we could find, it was relatively difficult to create an accurate and meaningful predictive model. Humans are difficult to predict, and our graphs show that despite stay at home orders not changing and restrictions not being lifted people gradually started to travel more and more. However, we did find that politics play a large role in how a county reacts to the stay at home mandate, and how much people are reducing nonessential travel. Beyond that we saw that Washington follows the United States trends in travel fairly closely, with meaningful deviations in just a third of the travel categories.

**Challenge Goals**

We initially planned to use **New Libraries** and **Machine Learning** as our challenge goals - which we believe we met - however we think that we also accomplished the goals of **Multiple Datasets** and **Messy Data** as well. We spent a significant amount of time gathering, cleaning, and compiling our data - as it is difficult to find county level data that is consistent across each county - and this quickly became a large portion of our project. We think the amount of work we put into prepping and combining data, and by using 4 different datasets, shows that we were able to accomplish the goals of Messy Data and Multiple Datasets as well. We used Machine Learning to create several decision tree classifiers and regression models so we believe we met our challenge goal there. We also created an interactive plot that has a hoverable plot that indicates what county is attached to each data point in Plotly (a new library) so we believe we accomplished that goal as well.

**Work Plan Evaluation**

Our Plan:

1. Gather necessary data and county classifications (3 hours)

*This was a really poor estimate of time - this specific step took over 10 hours to gather, clean, and combine all of the data. It was very difficult to find meaningful county level data that had standard formatting and classifications across all counties.*

1. Perform initial Random forest classifier (0.5 - 1 hour)

*This took closer to 1-1.5 hours as it took more time than we expected to create the desired one hot encodings and a useful model.*

1. Perform decision tree classifier to classify all counties into groupings (0.5 hr)

*This was a fairly accurate step.*

1. Predict movement for groups using the a decision tree regressor (0.5 hr)

*This step ended up taking closer to 2 hours as we spent a significant amount of time attempting to create a meaningful and accurate model. Unfortunately as we examined the limitations to our data and graphs of trends over time we realized that our data and the nature of humans in the pandemic did not lend itself well to the models we had planned.*

1. Create a line graph to show the percent changes in retail/entertainment nonessential travel over the course of the COVID-19 pandemic (0.5 hrs)

*This took closer to an hour as we had initial difficulty understanding the new packages and getting the desired graphs.*

1. Create a graph showing how mobility changes regarding the six location types in the data compare at the WA state level to the national average (0.5 hours)

*This step was fairly accurate.*

1. Interpret results and compile into final report (4 hours)

*It ended up taking us longer to compile into the final report because there was a significant amount of interpreting the meaning of our data, this step took closer to 8 hours.*

We ended up primarily doing all of the coding over zoom as we found it easiest to talk through steps and make decisions together. There was still some individual work but primarily we worked together.

**Testing**

As our project was based on machine learning and plotting, we used accuracy testing to test the validity of our code. We ran accuracy testing of our ML model and found that we had an average of 18.129 difference from what our model predicted for percent change from baseline vs what was actually seen on that day. This is a huge error and the main reason that we do not trust our model, which led us to pivot into gathering any information we could from our model and analyzing data in groups based on what we found.

**Collaboration**

We primarily used ED as a reference to get a layout for our code, and then used the official documentation for Plotly, MatPlotLib, Pandas, and SkLearn to refine and tailor our code to our intended purpose. Additionally we used the website “MachineLearningMastery.com” to find more information about ML in python as well as to determine feature importance as well as “DataCamp.com” to learn more about random forests.