Quinn Shim and David Ding
CS 396
MS2

1. Updates on your original proposal. In many cases, what you proposed may not be accomplishable anymore after EDA. This is common. In this case, you might want to switch or change your project goal slightly based on the result of EDA. If this is the case, please describe what change is made and why they are needed. If no change is needed, please describe how your EDA results reconfirm your project goal and help you proceed to data modeling.

After our EDA phase, we decided not to study whether or not rates of COVID cases and deaths changed after election results, as mentioned in our original proposal. We found that our data was better suited to focus on the question of whether COVID was correlated with election outcomes. The EDA we conducted showed a possible correlation between mask usage in different counties and election outcomes, so we decided to explore that further in our data modeling, along with cases and deaths data.

2. Data Modeling. As we mentioned in class, data modeling is just a way to understand the data. We have covered numerous statistical and machine-learning modelings: some are basic while others are more complicated. Discuss what you have tried, why you try and what's the result. Note that failed attempts (which may actually happen more) should also be discussed. For ML models, describe what evaluation metric is used. Is there any attempt to reduce overfitting? If text processing is needed in your project, you can discuss them too.

For this project, we tried Pearson and Spearman correlation coefficient, Chi Squared testing, K nearest neighbors and Linear Regression. For the Pearson coefficients, we did not find significant relationships between the percent of voters which voted for a candidate and the number of cases or deaths for that county. These values were close to zero, showing a small correlation between these two variables. However, Pearson did show a medium-strength, positive relationship between voting for Biden and wearing a mask frequently. This also showed about the same strengthened negative correlation between Trump voters and not wearing a mask frequently. The Pearson correlation coefficient showed that there was a slightly higher monotonic relationship between the variables than was found using Pearson. This increase, however, was insignificant and did not provide any new perspectives on the correlations given from Pearson.

We also conducted a Chi-Squared test of independence between Trump-majority counties, Biden-majority counties, and whether or not their per-capita COVID case rate was above or below the average of all counties. Our calculated p-value for this test was 0.16, which is above our significance level of 0.05. Therefore, we cannot reject the null hypothesis that states the two categories are independent, and conclude that there is not a statistically significant relationship between the two.

We also tried KNN modeling on our data to try and find a relationship, but we found that it was not a very effective predictor of election outcomes given COVID cases and deaths data. We tried several values for k, including 5 neighbors, 7 neighbors, and 10 neighbors. Anything less than 5 neighbors would lead to overfitting, given our dataset size. We used euclidean distance as the evaluation metric. Plotting the data, we saw that Trump counties and Biden counties

were all plotted fairly close to each other, without distinctive clumps. There were also far more counties where Trump received the majority of votes. In fact, around 83% of US counties we studied had a majority of Trump votes over Biden. Our KNN accuracy was found to be around this range as well, between 83-86% accuracy. Our model was not much more effective at predicting the outcome than simply picking “Trump” for every county.

Our linear regression model did not provide a strong model for our data. In the models for voting preference and covid cases and deaths, there was a very small R2 value for our combinations. This shows that the explanatory variables for our models do not provide a great model for the relationship between voting and covid statistics. The R2 value for voting preferences and mask usage was slightly higher, though still relatively small at around .15. This again shows that mask usage does not provide a great explanation for the voting preferences of counties.

With these results combined, the models show that our original hypothesis cannot be supported from the data we used. Although there was a correlation found between voting preference and mask usage, the models show that there is no significant evidence that covid statistics can be used to explain the voting preferences in counties.

3. Difficulties that you have encountered. Discuss those that you have solved and those you have not solved. I will provide with you feedbacks to help you succeed!

Our first difficulty we encountered was when we tried using just the number of covid cases in our data models. This turned out to not provide fair and accurate results because we did not consider using cases per capita, especially since there is a political split between highly populated counties and sparsely populated ones. We fixed this by adding a new column to account for the cases and deaths per capita.

Our second difficulty came with the linear regression. For more accurate measurements, we would need to add more explanatory variables to our regression. We could then use ridge regression to hone in on the best variables. However, we did not know which explanatory variables to use as well as had trouble combining data into 2D arrays to be used by the sklearn library.

4. Plan on visualization. You do not need to do this yet, but discuss a short plan on what results can be visualized.

For visualization, we will visualize KNN and the Linear Regressions. We plan on visualizing these plots for multiple combinations of variables. We will also make a histogram for the correlation coefficients we found from Pearson and Spearman.