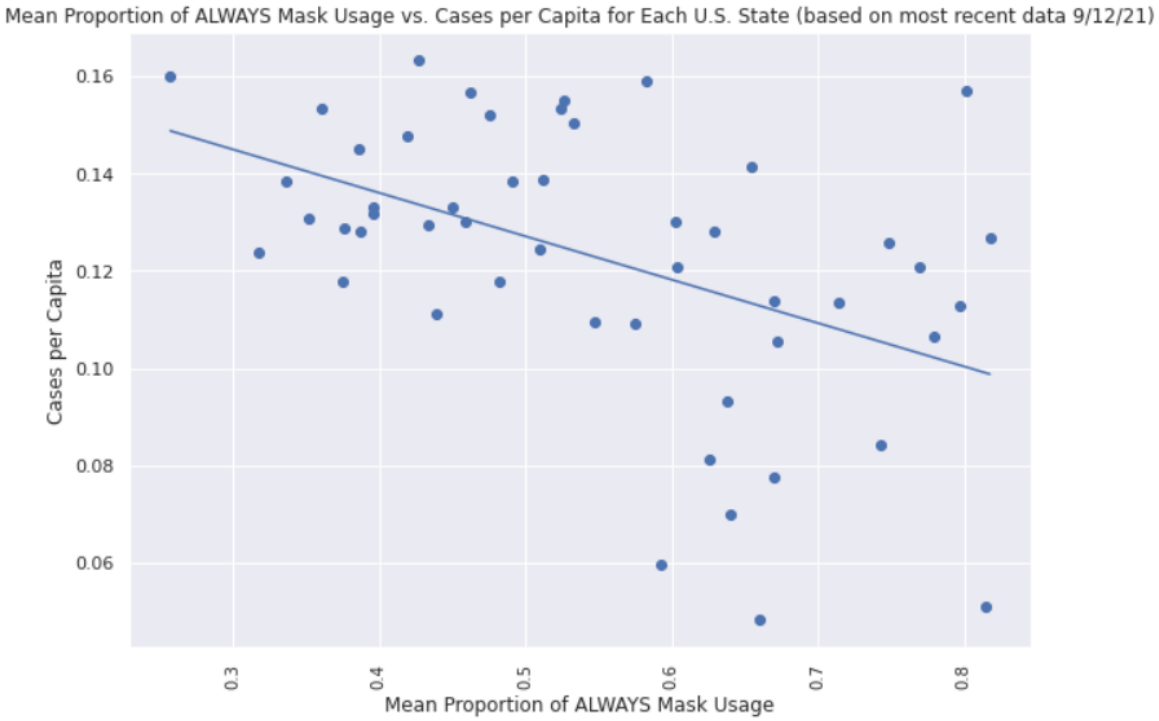
OPEN ENDED EDA (CLOBBER)

Improved Hypothesis: Partaking in more COVID-19 precautions (higher mask usage, higher vaccinations per capita, and more social distancing) is correlated to less COVID-19 cases per capita for each county in the U.S.

EDA 1:



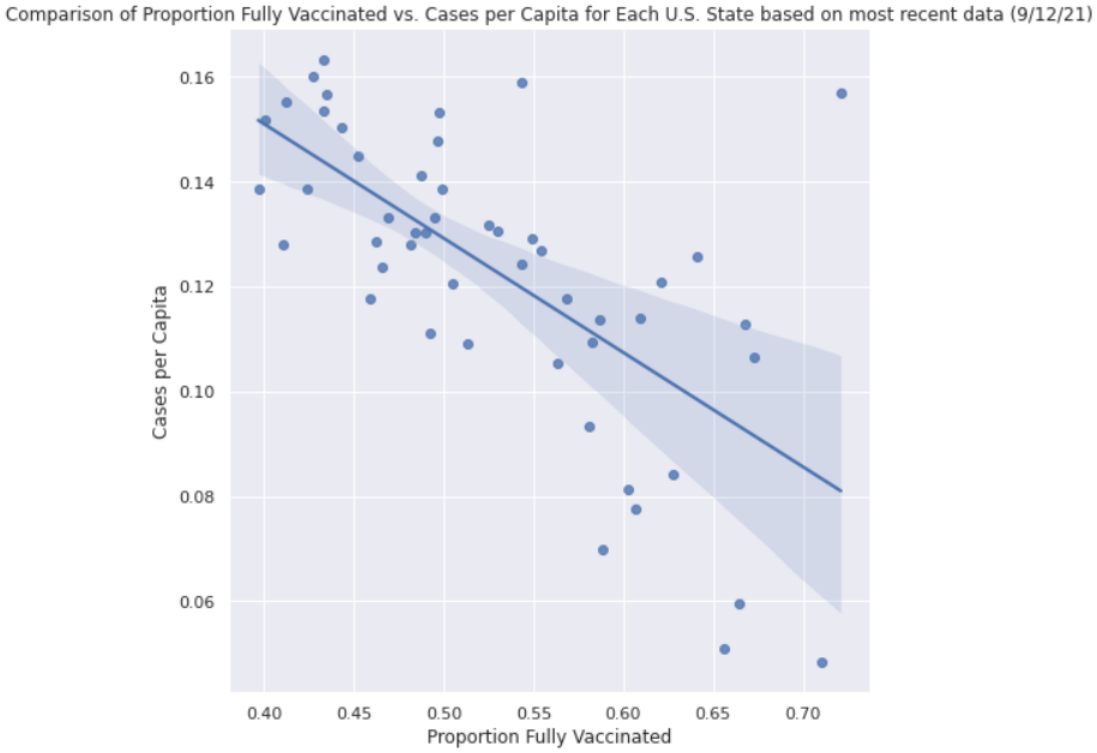
The data used to create the above visualization is a joined dataset from the cases, counties, and mask_use dataframes. Each row in the dataset represents a particular state in the U.S. There are two features corresponding to each state: cases per capita and the mean proportion of the population whose mask usage is maximum ('ALWAYS') within each state. Each state's population estimate in 2020 ('POPESTIMATE2020') was used to derive these features. This data is from 9/12/20, the most recent count of total cases for each U.S. state.

This scatter plot analyzes the mean proportion of each state who wears masks at the maximum "Always" frequency (that was aggregated by taking the mean of the proportions for all counties in a particular state) and the Covid-19 cases per capita based on September 12, 2021 data. This visualization indicates that states with a higher mean proportion of "Always" mask usage tend to be correlated with having fewer cases per capita as we can notice a downward trend in the data. This conclusion supports our hypothesis, which initially stated that states participating in more Covid-19 precautions (i.e. mask usage) are correlated with fewer Covid-19 cases compared to states who don't in the United States.

In this visualization, a higher mean proportion of the Always mask usage seems to correlate to less COVID cases per capita for each state. Therefore, this would provide a justification for our hypothesis, and we are intrigued to determine whether including social

distancing as a feature would also have an effect on improving/reducing the correlation between COVID precautions and the number of cases for each county. Specifically, how does social mobility in various areas (e.g. workplaces, residential, transit stations) correlate to the number of COVID-19 cases per capita for each county in the U.S.? In other words, does reducing the frequency of close contact between people contribute to minimizing the cases per capita?

EDA 2:



The data set used to create this visualization is a joined dataset from the cases, counties, and mask_use dataframes in order to create new columns labeled ‘cases per capita’ and prop_fully_vaccinated.’ Each row in the dataset represents a state in the U.S. There are two features corresponding to each state: cases per capita and the proportion of the population within each state that is fully vaccinated. The population used to derive these features was based on ‘POPESTIMATE2020’ (each state’s population estimate in 2020). In addition, this data is only using data from 9/12/20 which represents the most recent count of total cases and vaccinations for each U.S. state.

This scatterplot shows a moderate negative correlation between the proportion fully vaccinated and cases per capita for each state. This means that as the proportion fully vaccinated increases in states, the cases per capita overall decreases. As the proportion fully vaccinated decreases, the cases per capita overall increases. In relation to our hypothesis, this visualization proves that participating in COVID-19 precautions like getting vaccinated is correlated to less cases per capita for each state. Therefore, there is a likelihood that social distancing is also correlated to less COVID cases per capita and possibly lower COVID cases overall.

For further EDA, we are curious to see how social mobility is correlated to the number of

COVID-19 cases per capita for each state. Does lower social mobility in a particular area (such as workplace over retail/recreational areas) have a higher correlation with cases per capita? In other words, which locations are the most important for people to practice social mobility measures?

Experimental Design

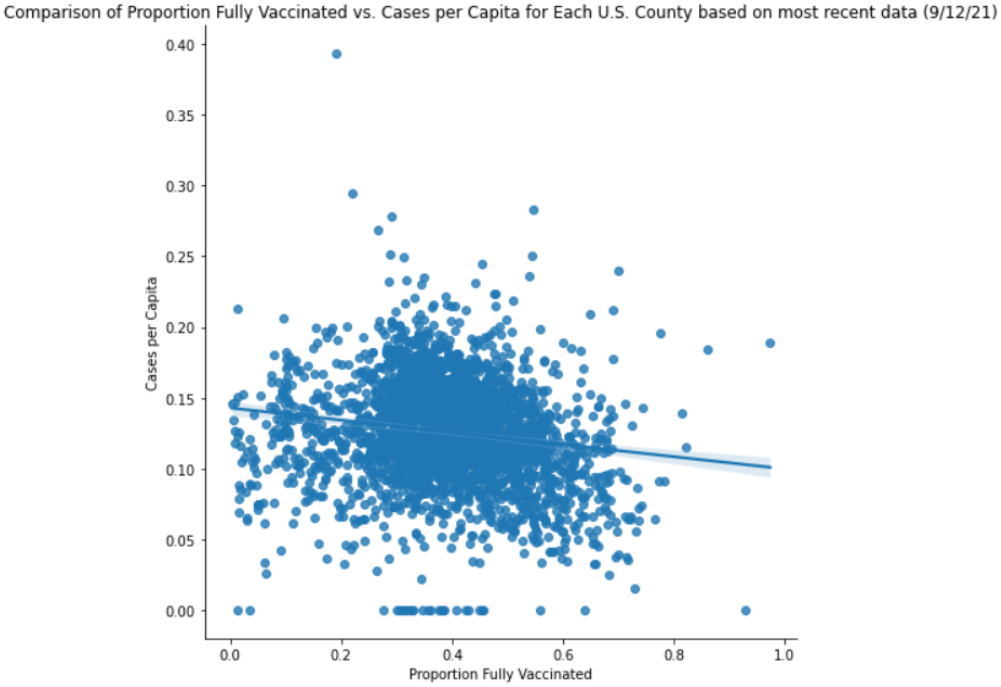
- We will be doing multivariate modeling and using a Random Forest Regression model to predict cases per capita for each state based on COVID precautions. To do this, we will use the data already provided in the data frames "vaccinations," "cases," "mask_use," and "counties." We will also look for external data about social distancing in each county and vaccination rate for each county.
- We will introduce a new dataset through Google's COVID-19 Community Mobile Reports. For each state, there is information about 'retail & recreation', 'grocery & pharmacy', 'parks', 'transit stations', 'workplaces', and 'residential'. The data is organized temporally in which they measure the daily increase or decrease in each of the categories outlined in comparison to a baseline of 0% for each state. We will gather the median for each of these categories amongst all days recorded starting from January 23, 2020 to September 12, 2021. In addition we will get data about the value of fully vaccinated populations in each county from CDC.
- We will predict total number of cases per capita for each county based on the most recent data (up until 9/12/21) by using a Random Forest Regression model based on 12 features: proportion of mask wearing (ALWAYS, NEVER, SOMETIMES, RARELY, FREQUENTLY) population, proportion of fully vaccinated population, percent of social mobility for each category for each state. We will use a train/test split of 80/20%. We will improve this model through cross validation and regularization which will in particular, help identify the best features of social mobility to use.
- We will visualize the accuracy of our model by calculating the RMSE and visualizing predictions vs. the actual number of total cases in a scatterplot. A scatterplot with a high correlation will mean that our hypothesis is proven to be true, while a low correlation will conclude that our hypothesis is not justifiable as the hypothesized COVID-19 precautions would not be highly efficient at predicting COVID cases per capita for each county. We can also visualize the accuracy of our hypothesis by creating a scatterplot between each feature and cases per capita to see if that particular COVID precaution has a positive or negative correlation with cases per capita. If lower cases per capita are correlated with more people doing a COVID precaution, then our hypothesis can be proven to be true.

Open Ended Modeling Report

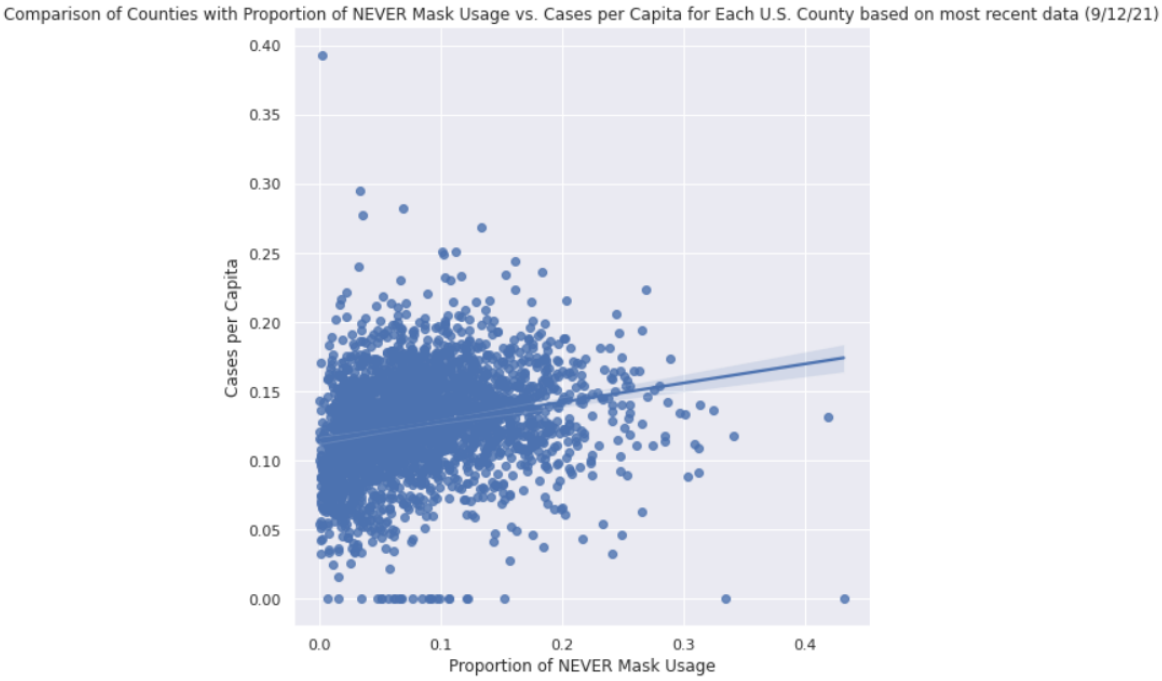
Hypothesis: Partaking in more COVID-19 precautions (higher mask usage, higher vaccinations per capita, and more social distancing) is correlated to less COVID-19 cases per capita for each county in the U.S.

Objective: We will predict total number of cases per capita for each county based on the most recent data (up until 9/12/21) by using a Random Forest Regression model based on 12 features: proportion of mask wearing (ALWAYS, NEVER, SOMETIMES, RARELY, FREQUENTLY) population, proportion of fully vaccinated population, percent of social mobility for each category for each state. We will use a train/test split of 80/20%. We will improve this model through cross validation and regularization which will in particular, help identify the best features of social mobility to use.

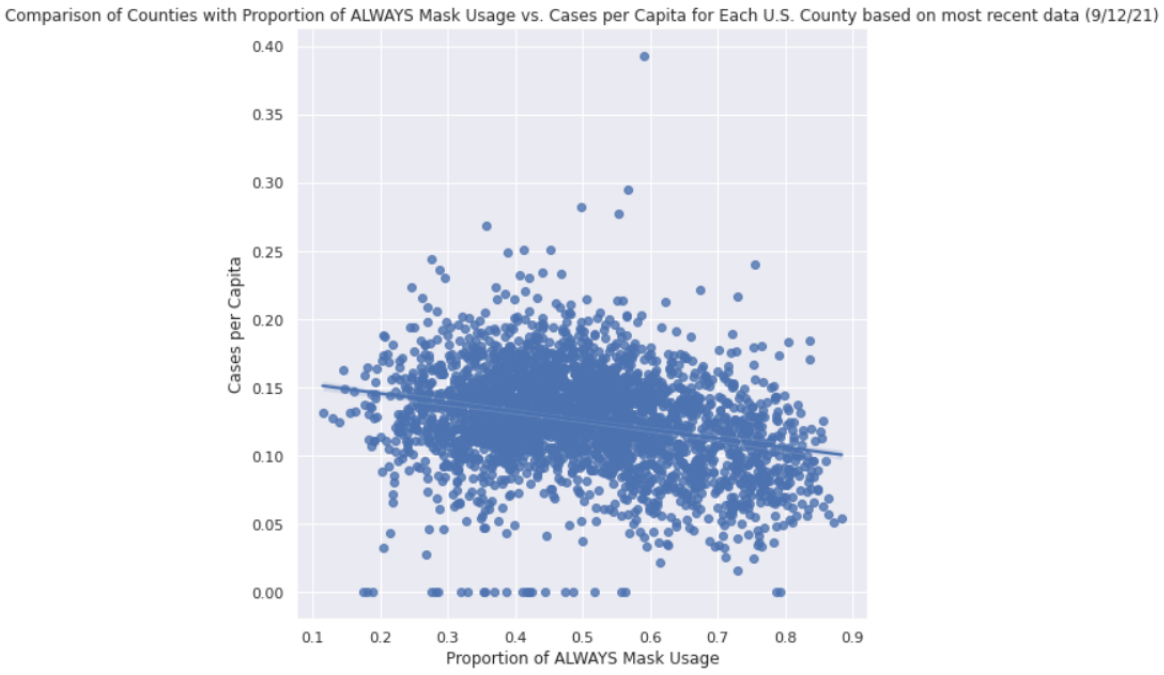
- *Problem:* We can visualize the accuracy of our hypothesis by creating a scatterplot between each feature listed above and cases per capita to see if that particular COVID precaution has a positive or negative correlation with cases per capita. If lower cases per capita are correlated with more people doing a COVID precaution, then our hypothesis can be proven to be true. We can confirm this hypothesis with existing datasets, including an external dataset based on the Google Community Mobility Reports.
- *Answer:* Plots labeled 1 - 4 below prove this hypothesis to be true.



PLOT 1 - There is a moderately weak negative linear correlation between the proportion fully vaccinated and the cases per capita in each county for data updated on 9/12/21. This shows that a higher proportion of a fully vaccinated population is correlated to less cases per capita for each county.



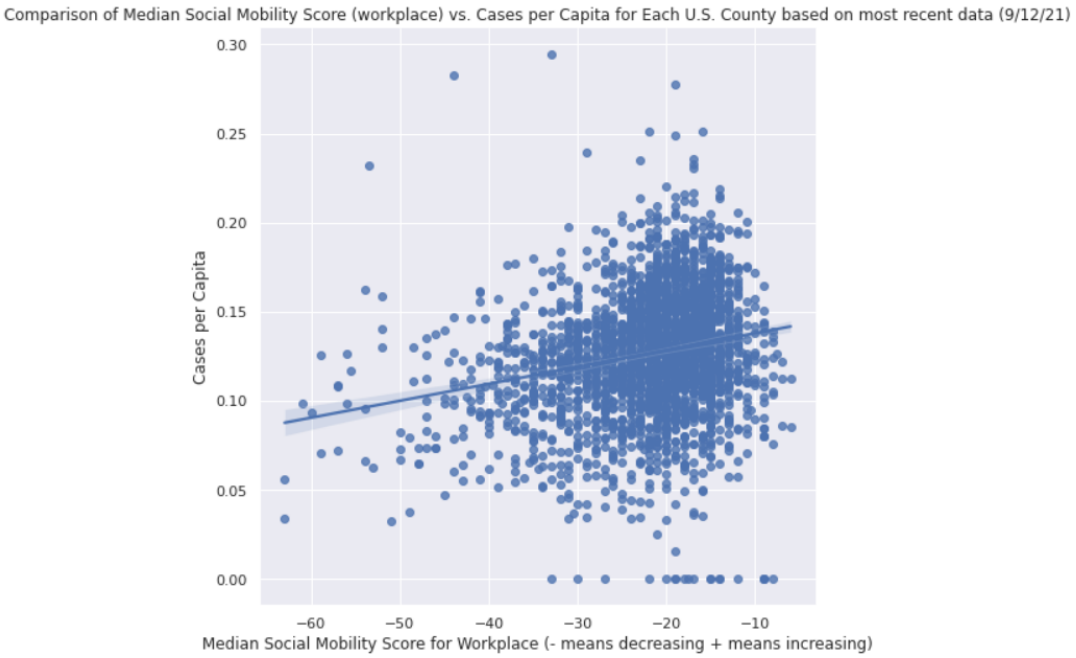
PLOT 2 - There is a moderately weak positive linear correlation between the proportion of NEVER mask usage and the cases per capita in each county for data updated on 9/12/21. This shows that a higher proportion of never wearing masks is correlated to higher cases per capita for each county.



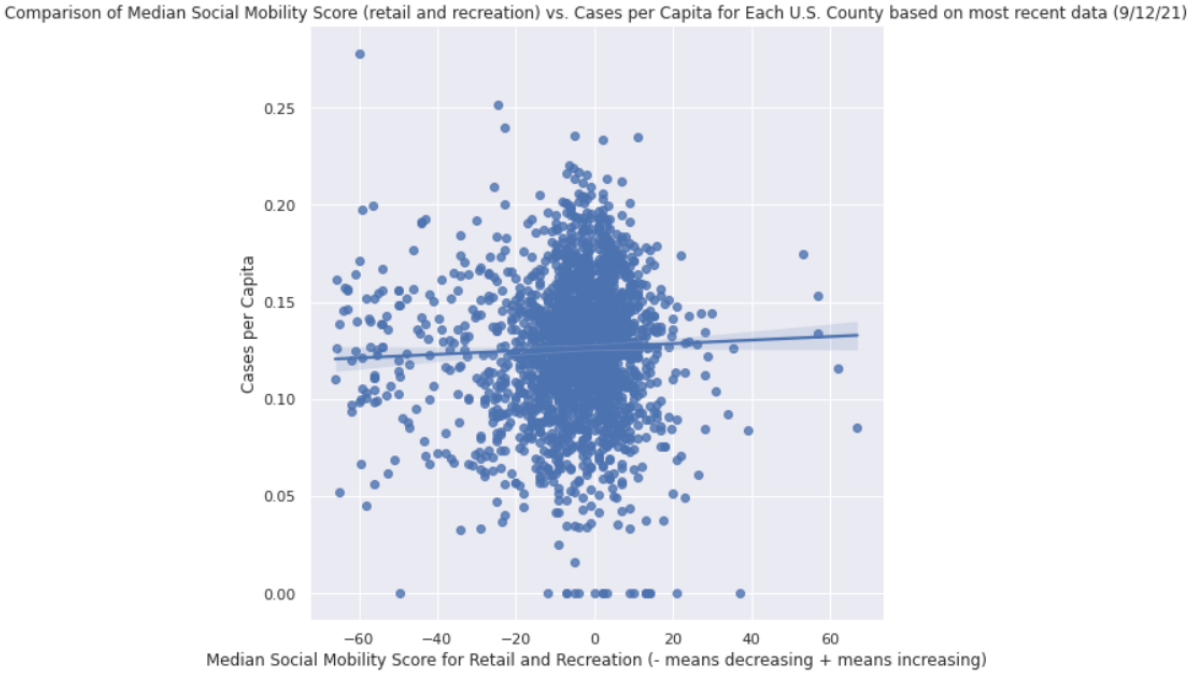
PLOT 3 - There is a moderate negative linear correlation between the proportion of ALWAYS mask usage and the cases per capita in each county for data updated on 9/12/21. This shows that a higher proportion of always wearing masks is correlated to lower cases per capita for each county.

The three plots above showing linear relationships between the proportion of fully vaccinated, NEVER mask usage, and ALWAYS mask usage with cases per capita are example of ways to prove our improved hypothesis: Partaking in more COVID-19 precautions (higher mask usage, higher vaccinations per capita, more social distancing) is correlated to less COVID-19 cases per capita for each county in the U.S.

IMPORTING A NEW DATASET FOR SOCIAL MOBILITY SCORES (A COVID-19 precaution)



PLOT 4 - There is a moderately weak positive linear correlation between the median social mobility score for workplaces and the cases per capita in each county for data updated on 9/12/21. This shows that a higher social mobility score of workplaces (more people are going to work) is correlated to higher cases per capita for each county. This plot above is another example of how to prove our improved hypothesis: Partaking in more COVID-19 precautions (higher mask usage, higher vaccinations per capita, more social distancing) is correlated to less COVID-19 cases per capita for each county in the U.S. This plot shows that more social distancing (in the workplace) is correlated to less cases per capita.



PLOT 5 - There is little to no linear correlation between the median social mobility score for retail and recreation and the cases per capita in each county for data updated on 9/12/21. This feature's lack of a linear relationship with the output (cases per capita) will be explored later. This introduces the idea that not all of our features that we plan to use in our model will have a linear relationship with cases per capita although most of them do.

BASELINE MODELS

In order to determine the best baseline model, we trained our data on a Multiple Linear Regression model and a Random Forest Regression model.

The inputs for both of the baseline models were:

- ‘prop_vac’ - proportion of fully vaccinated population in each county updated on 09/12/21 based on POPESTIMATE2020 and external dataset by CDC
- ‘retail_and_recreation_percent_change_from_baseline’ - median value for social mobility in the geographic category: retail and recreation over the span of time from 01/23/21 to 09/12/21 (based on Google Community Mobility Reports)
- ‘grocery_and_pharmacy_percent_change_from_baseline' - median value for social mobility in the geographic category: grocery and pharmacy over the span of time from 01/23/21 to 09/12/21 (based on Google Community Mobility Reports)
- 'parks_percent_change_from_baseline' - median value for social mobility in the geographic category: parks over the span of time from 01/23/21 to 09/12/21 (based on Google Community Mobility Reports)
- 'transit_stations_percent_change_from_baseline' - median value for social mobility in the geographic category: transit stations over the span of time from 01/23/21 to 09/12/21 (based on Google Community Mobility Reports)
- 'workplaces_percent_change_from_baseline' - median value for social mobility in the geographic category: workplaces over the span of time from 01/23/21 to 09/12/21 (based on Google Community Mobility Reports)
- "NEVER" - NEVER mask usage proportion from mask_use data for each county
- "RARELY" - RARELY mask usage proportion from mask_use data for each county
- "SOMETIMES" - SOMETIMES mask usage proportion from mask_use data for each county
- "FREQUENTLY" - FREQUENTLY mask usage proportion from mask_use data for each county
- "ALWAYS" - ALWAYS mask usage proportion from mask_use data for each county

Additional Notes on Inputs:

- We used the median for our six categories of social mobility scores in order to get an overall representation of each county’s social mobility score over the time frame of when cases began to be reported (1/23/2021 to 9/12/21). Median was preferred over mean, because of having both positive and negative numbers present in each category.
- We excluded 'residential_percent_change_from_baseline' in the baseline model, because there were too many null values in the original dataset, and this meant the median was not calculable for most counties.
- Any null values in the other social mobility categories were replaced with 0 to signify no movement from the base of 0% change in social mobility.

Outputs:

- ‘cases_per_capita’ - determined by taking the total number of cases recorded by 09/12/21 and dividing it by POPESTIMATE2020 for each county.

Sources Explanation:

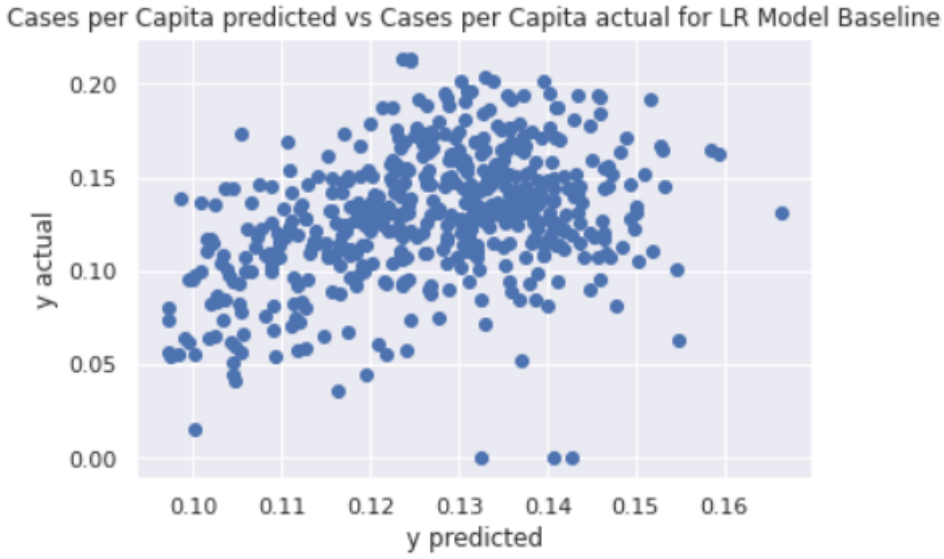
- Deriving number of fully vaccinated people for each county: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh
    - For baseline model - We took data from 09/12/2021 which was the total number of fully vaccinated people ('vaccines_counties.csv')
    - For improvement 1 of the model - We first gathered data only from 04/12/2021 to 09/12/2021 in a Google Colab notebook and then imported a new csv file to the project (allows for faster processing of the code in this notebook) ('vaccines_counties_time.csv')
- Deriving social mobility scores (all 6 categories): https://www.kaggle.com/bigquery/covid19-google-mobility
    - For baseline model - Data was collected from '2020-01-23 to 2021-09-12' in order to match with cases_per_capita data in provided cases and county dataset. We also calculated median of each sub category of social mobility taken from these dates as a representation of the county's social mobility overall in a Google Colab notebook and then transferred it to a new csv file ('social_mobility (1)')
    - For time-based model - We took data from 07/02/2021 to 07/14/2021 in Google Colab and then imported the new csv file to this project (allows for faster processing since this a very large dataset) ('social_mobility_july')

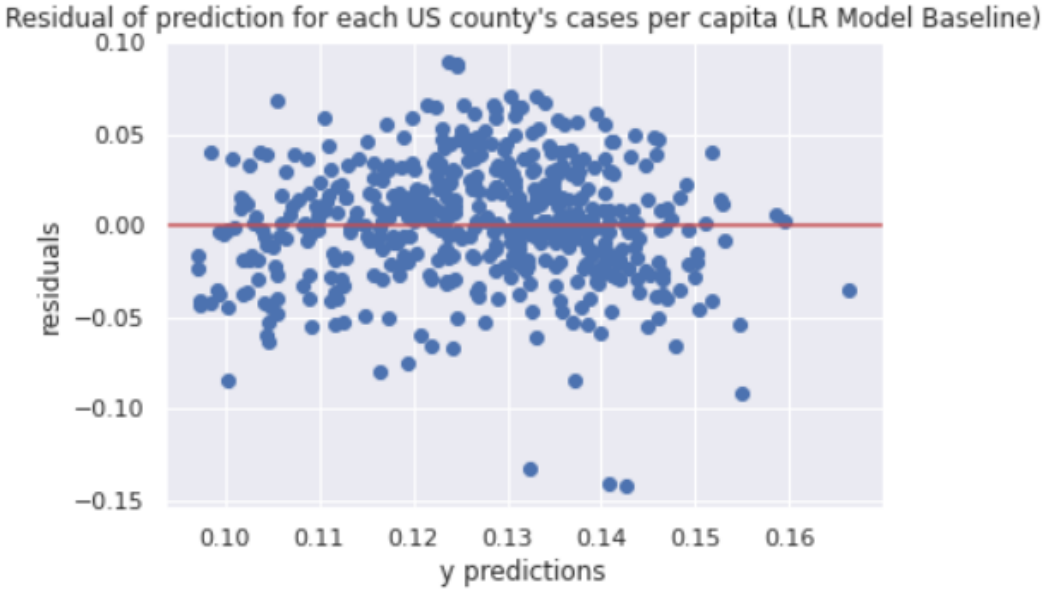
Link to Google Colab notebook:
https://colab.research.google.com/drive/1mU2fpfqfxZiIKYHYAFRqfXw8IOxewjrN?usp=sharing

We are training a Multiple Linear Regression model first, because we notice a linear relationship between mask usage, some social mobility scores and cases per capita for each county. This linear relationship can be seen in Plots 1-4.

RESULTS:

(train_rmse, test_rmse) = (0.03401421801316921, 0.034901767438255 63)



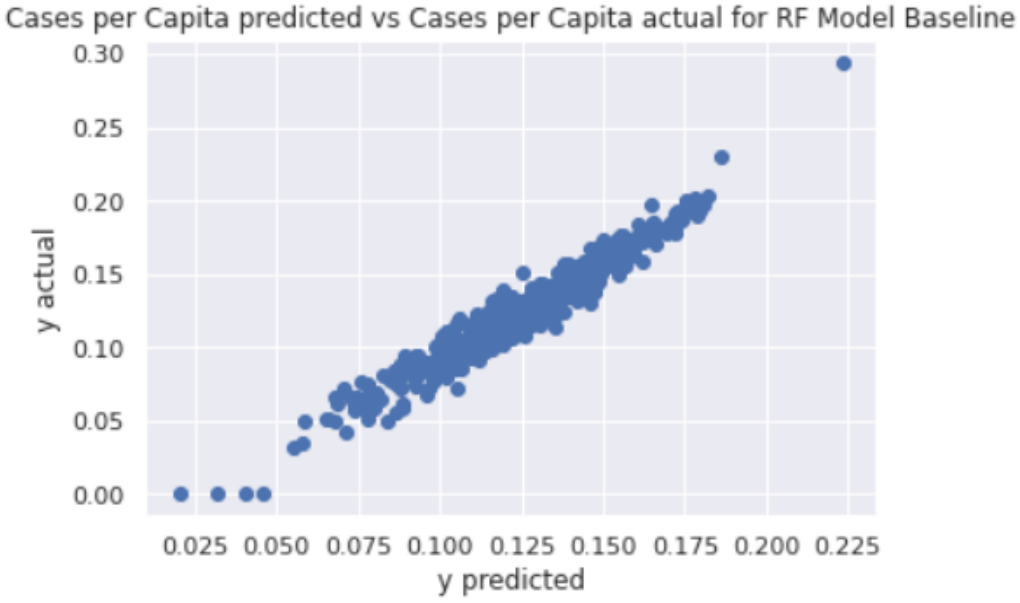


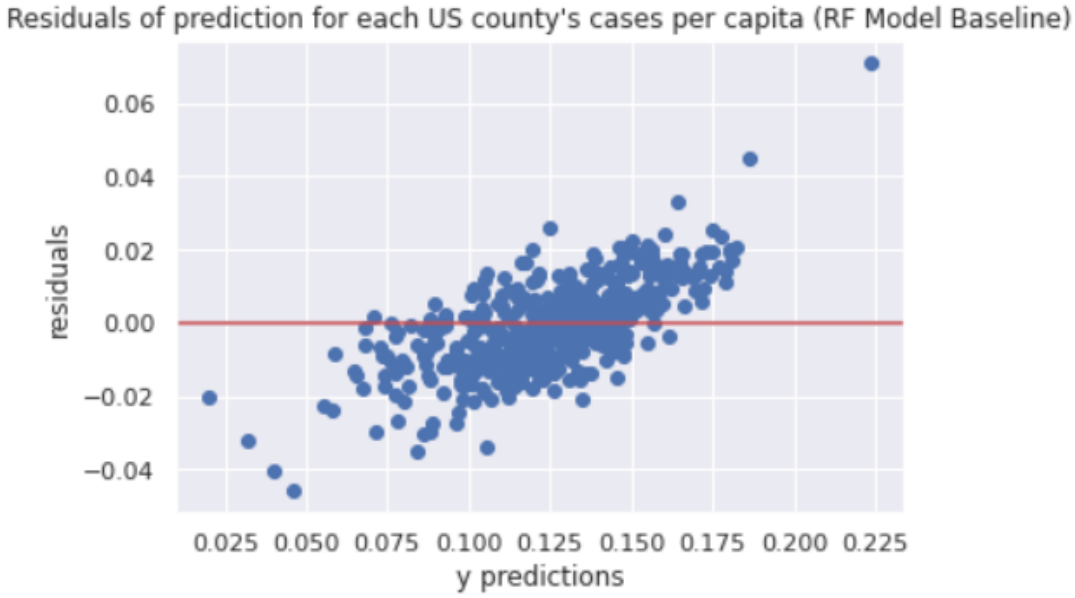
Train Score is : 0.14140134629689272
Test Score is : 0.13432259576407968

Despite having many features that have a linear relationship with the output (cases per capita), there are some features that do not have as strong or any linear relationship at all (as seen in PLOT 5) with the target variable like 'retail_and_recreation_percent_change_from_baseline'. Since not all features have a linear dependency, Random Forest Regression (RFR) might be better at representing the complex dependencies between the features. RFR can also represent large datasets well, handle missing values like the few in the social mobility categories, and it is more efficient at predicting amongst complicated relationships and outliers.

RESULTS:

(train_rmse, test_rmse) = (0.012273640158940823, 0.010944068945704943)





Train Score is : 0.891453339556849
Test Score is : 0.8860004944194824

The Random Forest Regression model (baseline) performed best, so we decided to improve upon the Random Forest Regression baseline model. The scatterplot 'Cases per Capita predicted vs Cases per Capita actual for RF Model Baseline' in comparison to the scatterplot 'Cases per Capita predicted vs Cases per Capita actual for LR Model Baseline' clearly shows that the RF model is a much better fit as there is a stronger linear relationship between the y_predicted and the actual y in comparison to the Linear Regression model.

Additionally, the train score R^2 value was 0.891 and test score R^2 was 0.886 for the RF model, while the LR model had a much lower train R^2 score of 0.141 and lower test R^2 score of 0.134. The train RMSE and test RMSE were both also much lower (0.011795235855206662, 0.011255643168113673) in comparison to (0.0324670825406989 and 0.03388256235809448). The higher R^2 scores and lower RMSE values shows the RF model fits better to our data.

FINDING WAYS TO IMPROVE THE MODEL - FIRST TRY - CROSS VALIDATION

```
Trying first 1 features
        RMSE: 0.04015896428694099
Trying first 2 features
        RMSE: 0.03700889461812737
Trying first 3 features
        RMSE: 0.03596695094000228
Trying first 4 features
        RMSE: 0.035360005795992945
Trying first 5 features
        RMSE: 0.03532486639457322
Trying first 6 features
        RMSE: 0.034406407795098896
Trying first 7 features
        RMSE: 0.03303803529356388
Trying first 8 features
        RMSE: 0.03242175115741901
Trying first 9 features
        RMSE: 0.0316912233983354
Trying first 10 features
        RMSE: 0.031556072595192275
Trying first 11 features
        RMSE: 0.0313372137401589
Best choice, use the first 11 features
```

- *K-Fold Cross-validation errors when k = 10 show that all of the features should be used to optimize the model.*

IMPROVEMENT #1

Changes to inputs:

- "rate_of_vaccination" (replaces ‘prop_vac’) - represents the percent increase of fully vaccinated individuals starting from 07/02/2021 to 07/14/2021. (percentage is the number of fully vaccinated individuals each day divided by the total number of vaccinated (by 9/12/21)).
- New feature: 'rate_cases_per_capita' - the daily increase of cases (from 07/02/2021 to 07/14/2020) divided by the population estimate for each county (POPESTIMATE2020).
- New feature: ‘residential_percent_change_from_baseline' - can take the raw social mobility scores for each day in the specified time frame (07/02/2021 to 07/14/2020), so there are more data points that are not null. Therefore, we can use this feature, and fill any NAN values with 0.
- All of the mask use features outlined in the baseline model are still used.
- All of the social mobility features outlined in the baseline model, but instead of taking the median, we take the raw social mobility score because we are deriving values for each day in the time frame.

Additional Notes:

- We used the time frame from 07/02/2021 to 07/14/2021, because the data regarding mask usage is also taken from this time frame.

Outputs:

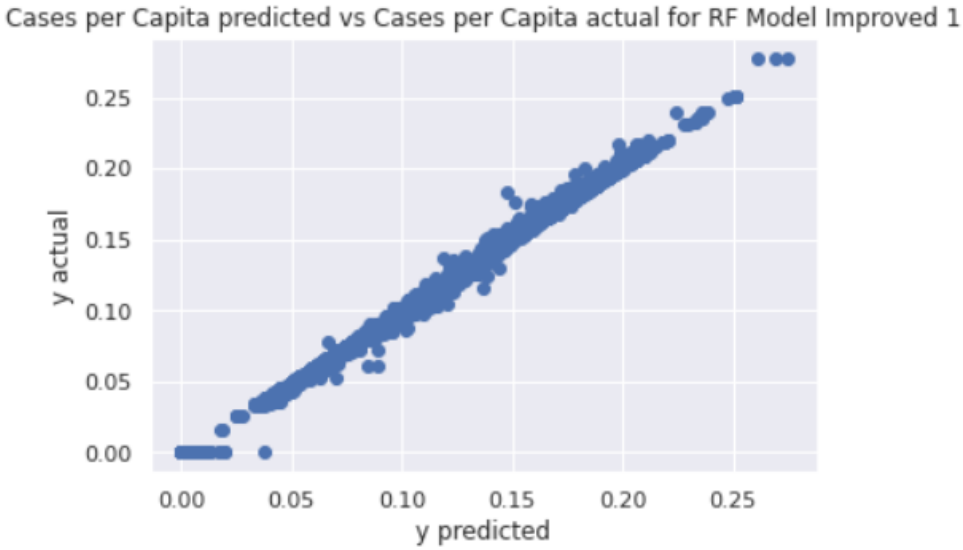
- *same as baseline model: (‘cases_per_capita’ determined by taking the total number of cases recorded by 09/12/21 and dividing it by POPESTIMATE2020 for each county.)*

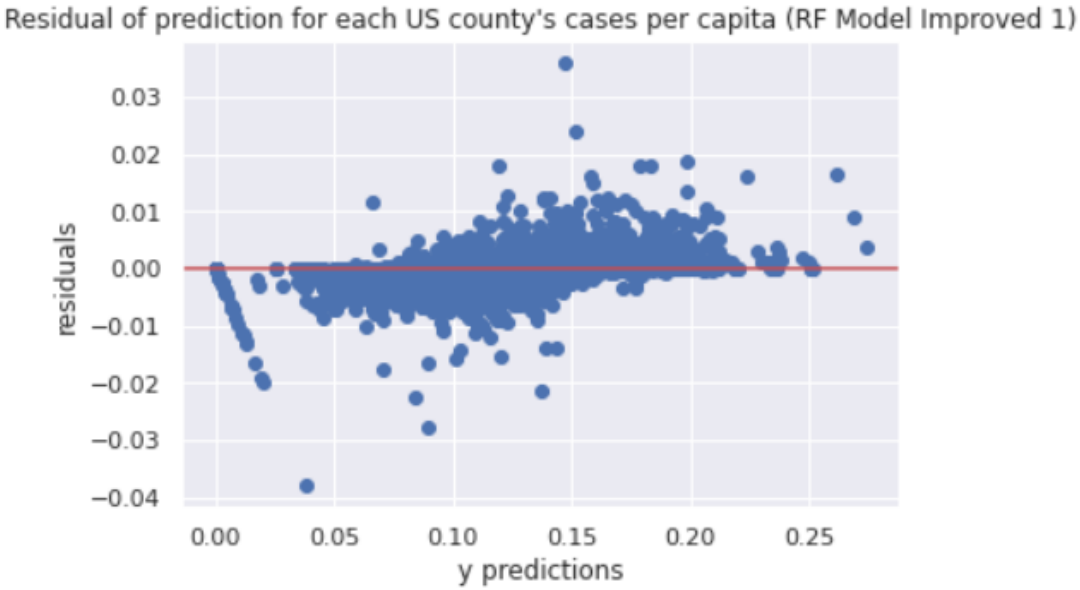
***Problem**:* Our baseline model performed quite well, but one improvement would be introducing more data. You can see in the scatterplot "Residuals of prediction for each US county's cases per capita (RF Model Baseline)" that the residual values are increasing for points that are at the lower and higher ends of the output value (cases per capita). This means that the model needs improvement in predicting for points that are outliers or for data points represented less in the data.

Solution: We are trying to predict total cases per capita for each county (updated until 9/12/21), so it would be beneficial to add a time component to our features. The accuracy of our model will increase as more data is introduced for each county, because Random Forest Regressor will be able to interpret and include more complex relationships between our features and the output variable. We collected data on our features from the time frame from 07/02/21 to 07/14/21. The changes we made in comparison to the baseline model was that we added two features, and updated what values we would use for social mobility scores. Instead of using the proportion fully vaccinated (‘prop_vac’), we replaced it with a ‘rate_of_vaccination’ value that represents the percent of fully vaccinated people in each county daily. We also created a feature called ‘rate_cases_per_capita” which represents the daily increase of cases per capita for each county.

Instead of using the median of social mobility scores overtime, we selected the raw values from each day instead.

After First Improvement





Train Score is : 0.9962266001619805
Test Score is : 0.9963166978050538

Result: Our intuition was correct, and the model improved.

Takeaways from the plots and statistics shown above:

- In the plot 'Cases per Capita predicted vs Cases per Capita actual for RF Model Improved 1', you can see that the linear relationship became stronger. The overall coefficient of determination increased (from ~.89 in the baseline model to ~.99 in the improved model).
- The residuals for points at the higher ends of the output values (y predictions) also decreased by ~.02, which is quite a large quantity considering our output values are scaled to be from 0.00 to 0.25.
- The training and test RMSE also decreased by more than half of the baseline model's value (from ~.01 to ~.002 for both training and test RMSE).

SECOND IMPROVEMENT - FEATURE ENGINEERING

'transit_stations_percent_change_from_baseline' AND 'parks_percent_change_from_baseline'

Inputs:

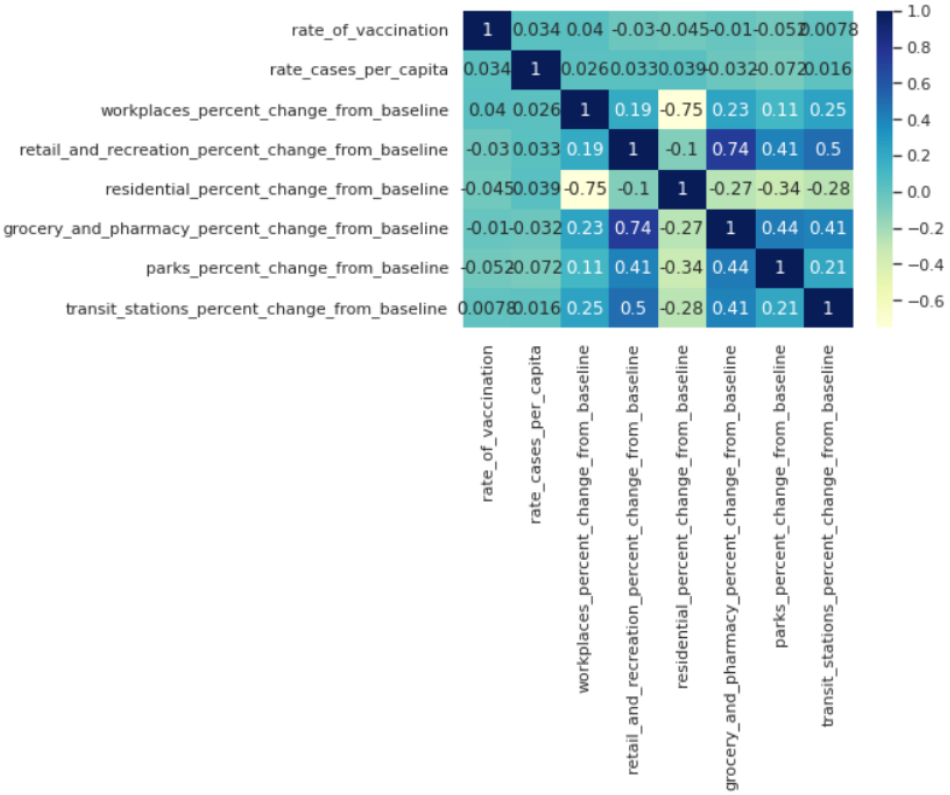
- Same as the improved model above, but we applied feature engineering to the features: 'transit_station_percent_change_from_baseline' and 'parks_percent_change_from_baseline' in order to fill the null values existing in these features.

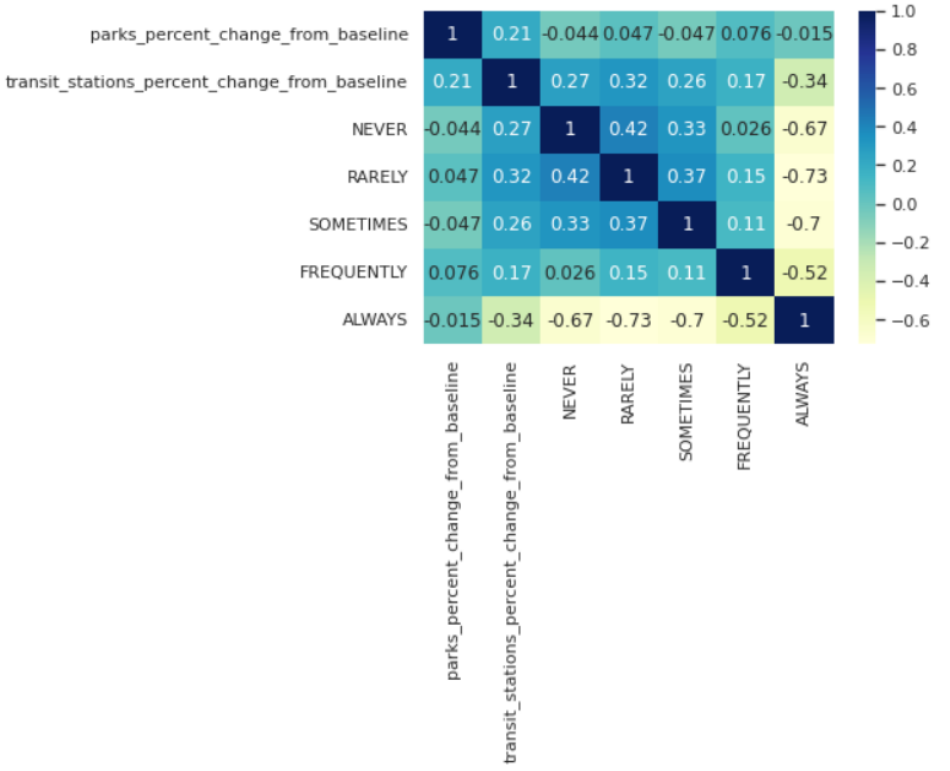
Outputs:

- Same as baseline model: ('cases_per_capita' determined by taking the total number of cases recorded by 09/12/21 and dividing it by POPESTIMATE2020 for each county.)

Problem: More than half of the values in the dataset had null values for the features: 'parks_percent_change_from_baseline' (21364/27053) and 'transit_stations_percent_change_from_baseline' (17342/27053). These features for social mobility are not as efficient in the model, because they all will be replaced with 0. We should make these features more meaningful, and the Random Forest Regression model will be able to add more layers of complexity in terms of relationships between features and the output which might increase the accuracy rate of predictions.

Solution: We filled these null values in the two features listed above with predictions determined from a Random Forest Regression model in order to make full use of these features, add more layers of complexity to our model, and potentially lower variance.


rate_of_vaccination 1 0.034 0.04 -0.03 -0.045 -0.01 -0.052 0.0078
rate_cases_per_capita 0.034 1 0.026 0.033 0.039 -0.032 -0.072 0.016
workplaces_percent_change_from_baseline 0.04 0.026 1 0.19 -0.75 0.23 0.11 0.25
retail_and_recreation_percent_change_from_baseline -0.03 0.033 0.19 1 -0.1 0.74 0.41 0.5
residential_percent_change_from_baseline -0.045 0.039 -0.75 -0.1 1 -0.27 -0.34 -0.28
grocery_and_pharmacy_percent_change_from_baseline -0.01 -0.032 0.23 0.74 -0.27 1 0.44 0.41
parks_percent_change_from_baseline -0.052 0.072 0.11 0.41 -0.34 0.44 1 0.21
transit_stations_percent_change_from_baseline 0.0078 0.016 0.25 0.5 -0.28 0.41 0.21 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
rate_of_vaccination
rate_cases_per_capita
workplaces_percent_change_from_baseline
retail_and_recreation_percent_change_from_baseline
residential_percent_change_from_baseline
grocery_and_pharmacy_percent_change_from_baseline
parks_percent_change_from_baseline
transit_stations_percent_change_from_baseline


parks_percent_change_from_baseline 1 0.21 -0.044 0.047 -0.047 0.076 -0.015
transit_stations_percent_change_from_baseline 0.21 1 0.27 0.32 0.26 0.17 -0.34
NEVER -0.044 0.27 1 0.42 0.33 0.026 -0.67
RARELY 0.047 0.32 0.42 1 0.37 0.15 -0.73
SOMETIMES -0.047 0.26 0.33 0.37 1 0.11 -0.7
FREQUENTLY 0.076 0.17 0.026 0.15 0.11 1 -0.52
ALWAYS -0.015 -0.34 -0.67 -0.73 -0.7 -0.52 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
parks_percent_change_from_baseline
transit_stations_percent_change_from_baseline
NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS

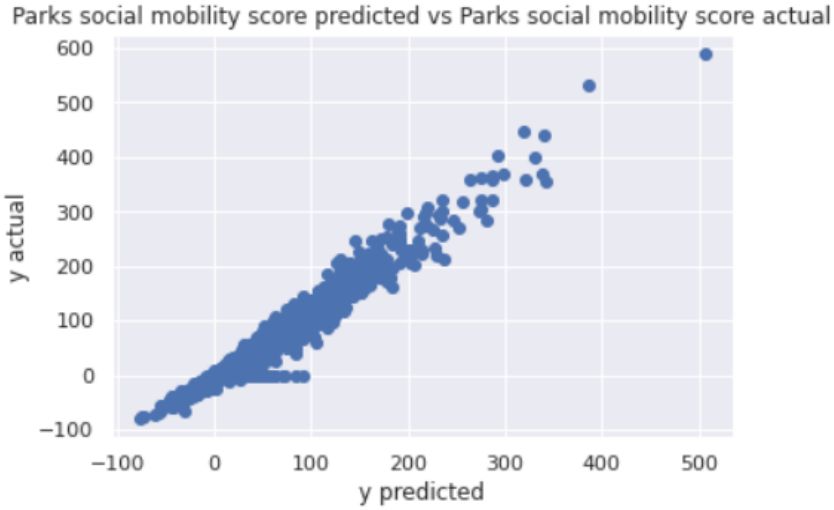
Feature Engineering Parks

Inputs:

- 'rate_of_vaccination'
- 'rate_cases_per_capita'
- 'grocery_and_pharmacy_percent_change_from_baseline'
- 'retail_and_recreation_percent_change_from_baseline'
- 'transit_stations_percent_change_from_baseline'
- 'workplaces_percent_change_from_baseline'
- 'residential_percent_change_from_baseline'
- "NEVER"
- "RARELY"
- "SOMETIMES"
- "FREQUENTLY"
- "ALWAYS"

All of the inputs are ones taken from the improved model above.

The inputs were chosen based on the above correlation matrices. These matrices show that there is a varying range of linear correlations with the inputs we chose and our output which is 'park_percent_change_from_baseline'. It ranges from a correlation of -.044 to .44. Since some features do not have a very strong linear correlation while some features have a moderate linear correlation, we will use a Random Forest Regression model to make a prediction about the values in our dataframe that are null for the 'park_percent_change_from_baseline' value.




```
]: baseline.score(X_test, y_test)

]: 0.9411360562121968
```

- Model's overall R^2 score = 0.9411360562121968
- *This model is determined to be good enough to predict the social mobility scores for 'parks_percent_change_from_baseline' as it has a R^2 value of around .941.*

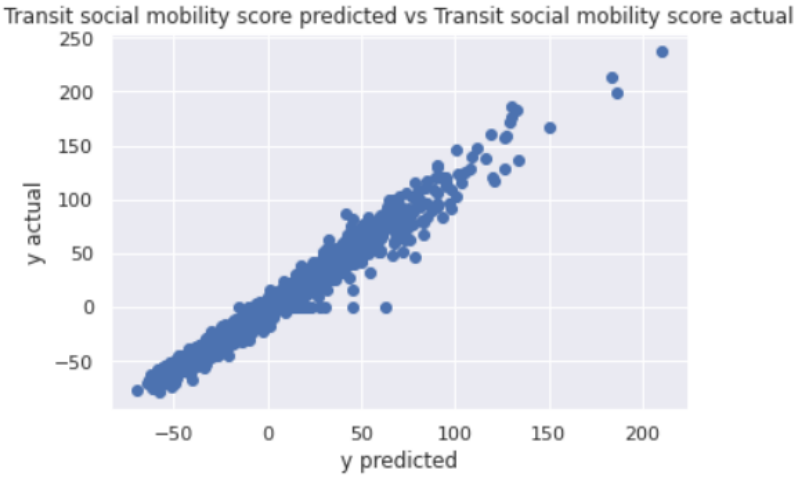
Feature Engineering Transit

Inputs:

- 'rate_of_vaccination'
- 'rate_cases_per_capita'
- 'grocery_and_pharmacy_percent_change_from_baseline'
- 'retail_and_recreation_percent_change_from_baseline'
- 'parks_percent_change_from_baseline'
- 'workplaces_percent_change_from_baseline'
- 'residential_percent_change_from_baseline'
- "NEVER"
- "RARELY"
- "SOMETIMES"
- "FREQUENTLY"
- "ALWAYS"

All of the inputs are ones taken from the improved model above.

The inputs were chosen based on the above correlation matrices. These matrices show that there is a varying range of linear correlations with the inputs we chose and our output which is 'transit_stations_percent_change_from_baseline'. It ranges from a correlation of -.0078 to .41. Since some features do not have a very strong linear correlation while some features have a moderate linear correlation, we will use a Random Forest Regression model to make a prediction about the values in our dataframe that are null for the 'transit_stations_percent_change_from_baseline' value.




```
8]: baseline.score(X_test, y_test)

8]: 0.9441877933315731
```

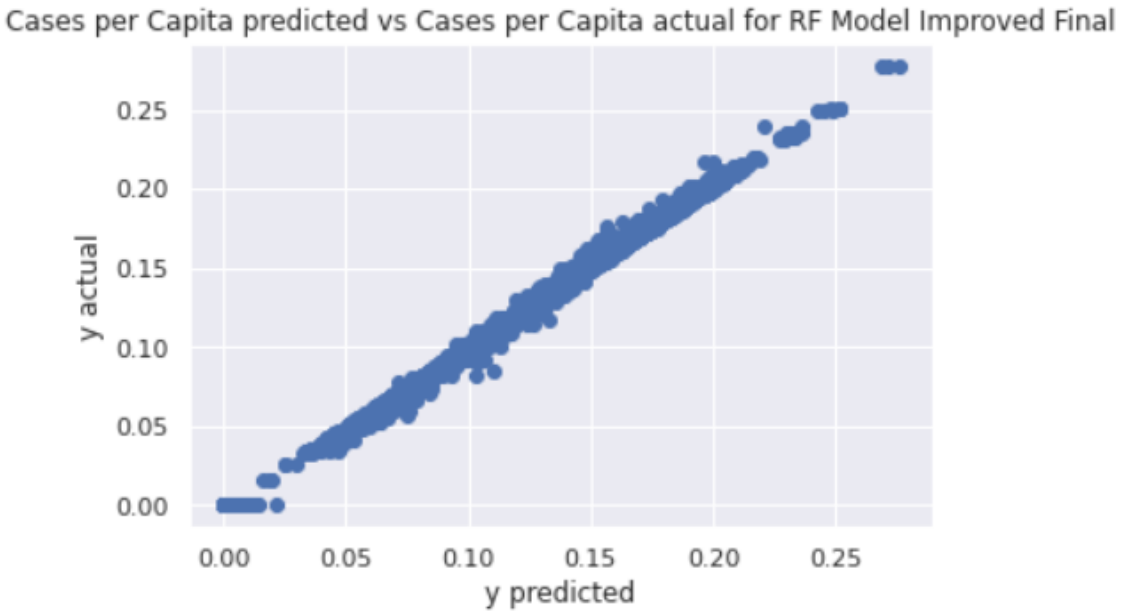
- Model's overall R^2 score = 0.9441877933315731

- *This model is determined to be good enough to predict the social mobility scores for 'transit_stations_percent_change_from_baseline' as it has a R^2 value of around .944.*

FINAL MODEL AND EVALUATION

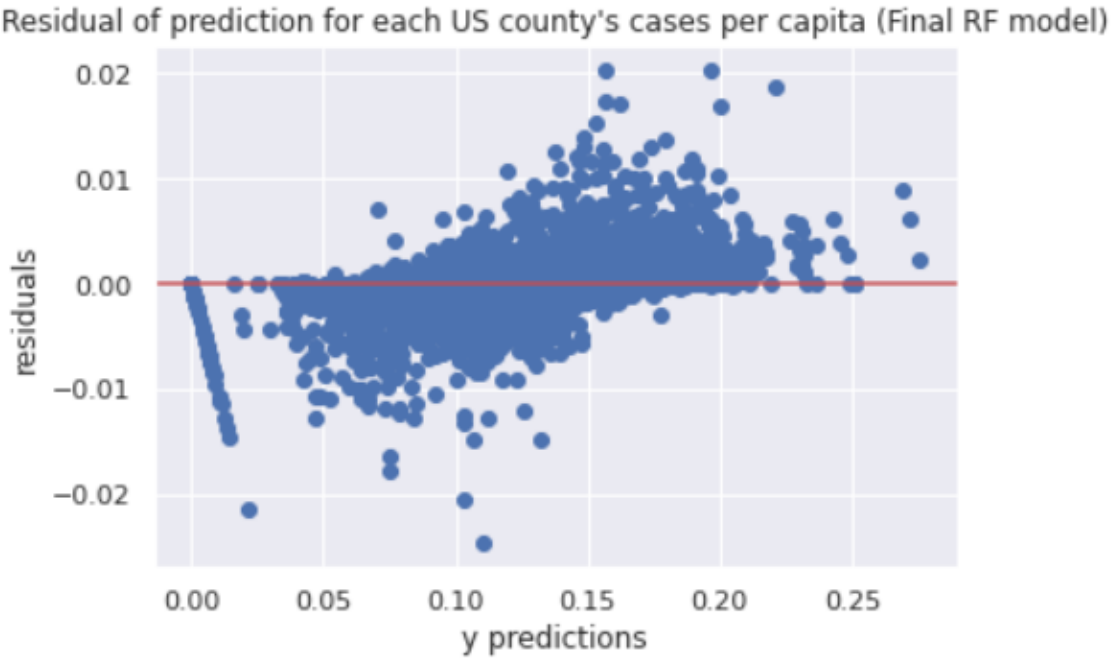
train_rmse, test_rmse = (0.0024452706222808152, 0.002176856467826995)

```
Trying first 1 features
        RMSE: 0.04023079827295893
Trying first 2 features
        RMSE: 0.038194488390326985
Trying first 3 features
        RMSE: 0.03600658809284159
Trying first 4 features
        RMSE: 0.034954274101133345
Trying first 5 features
        RMSE: 0.03392486483964276
Trying first 6 features
        RMSE: 0.033587447209131524
Trying first 7 features
        RMSE: 0.03272816763172469
Trying first 8 features
        RMSE: 0.031506289716131806
Trying first 9 features
        RMSE: 0.026179473156886623
Trying first 10 features
        RMSE: 0.019615156503424377
Trying first 11 features
        RMSE: 0.01397822694189905
Trying first 12 features
        RMSE: 0.011150313403513813
Trying first 13 features
        RMSE: 0.009599774536970277
Best choice, use the first 13 features
```



Takeaway: There is a strong positive correlation between cases per capita predicted and the actual cases per capita for each county after training our final RF model with both of our improvements. This means our R^2 value is most likely very high meaning our model has fit to our data very well. The predictions are very accurate to the actual cases per capita value for most points.

Connection to project: In comparison to the scatterplot 'Cases per Capita predicted vs Cases per Capita actual for RF Model Improved 1', this scatterplot shows a slight decrease in distance of the y predicted value to the y actual value. This essentially means that the residuals have gotten smaller for certain points in this final model in comparison to the previous model.



Takeaway: The residuals are low for our y predictions in this final model meaning our predictions are close to the actual values of each county's cases per capita.

Connection to project: The residuals have decreased for the data in the lower range of our output (cases per capita) meaning our model has gotten more accurate at predicting cases per capita for counties that have a low cases per capita value (0.00 to 0.05).

- Train Score is : 0.995055285798306
- Test Score is : 0.9960049564451132

Result:

- In the plot 'Cases per Capita predicted vs Cases per Capita actual for RF Model Improved Final', you can see that the linear relationship stayed mostly the same. Both the train R^2 and test R^2 values decreased slightly (by ~.001). The train R^2 went from 0.9962266001619805 to 0.995055285798306 and the test R^2 score went from 0.9963166978050538 to 0.9960049564451132. This change is too small for it to be a drastic reduction in the model's performance, but it can still be considered a slight decrease in our model's performance.
- The residuals for points at the higher and lower ends of the output values (y predictions) also decreased more in comparison to the model after the first improvement (by ~.02 for the lower end of the output values and by ~.01 for the upper ends of the output values). The range of values for our residuals decreased, which can be considered an improvement to our model.
- The training RMSE increased slightly (by ~.0003) and the test RMSE increased even more slightly (by ~.00001). (from (0.002120425806614615, 0.002121575606160645) to (0.0024452706222808152, 0.002176856467826995)). This can be seen as a very slight decrease in our model's performance.

A possibility as to why this improvement did not overall help in improving the model by much but instead slightly hindered its performance could be due to the fact that when model complexity increases, model variance also increases. The dataset might have gotten noisier which can in turn cause the model to capture this noisy data overall resulting in overfitting. Our improvement added more complexity to our model by making two of our features more representative of each county instead of just being 0 for most data points, but this in turn can make too complex of a model especially when using decision trees and many features in our Random Forest Regression model. Overfitting will overall decrease the accuracy of our model hence why our R^2 value was slightly lower and our RMSE values slightly higher.

FEATURE IMPORTANCE TABLE: FINAL MODEL

feature	importance
ALWAYS	0.280
RARELY	0.146
NEVER	0.141
SOMETIMES	0.135
FREQUENTLY	0.125
transit_stations_percent_change_from_baseline	0.038
parks_percent_change_from_baseline	0.037
retail_and_recreation_percent_change_from_baseline	0.036
grocery_and_pharmacy_percent_change_from_baseline	0.021
rate_cases_per_capita	0.013
workplaces_percent_change_from_baseline	0.012
rate_of_vaccination	0.010
residential_percent_change_from_baseline	0.005

FEATURE IMPORTANCE TABLE:AFTER IMPROVEMENT 1 MODEL

feature	importance
ALWAYS	0.283
RARELY	0.147
NEVER	0.145
SOMETIMES	0.142
FREQUENTLY	0.130
retail_and_recreation_percent_change_from_baseline	0.036
transit_stations_percent_change_from_baseline	0.033
parks_percent_change_from_baseline	0.027
grocery_and_pharmacy_percent_change_from_baseline	0.020
rate_cases_per_capita	0.013
workplaces_percent_change_from_baseline	0.011
rate_of_vaccination	0.008
residential_percent_change_from_baseline	0.005

FEATURE IMPORTANCE TABLE: BASELINE MODEL

feature	importance
prop_vac	0.170
ALWAYS	0.166
SOMETIMES	0.100
NEVER	0.086
RARELY	0.081
FREQUENTLY	0.080
retail_and_recreation_percent_change_from_baseline	0.077
workplaces_percent_change_from_baseline	0.075
grocery_and_pharmacy_percent_change_from_baseline	0.067
transit_stations_percent_change_from_baseline	0.055
parks_percent_change_from_baseline	0.042

The ranking of importance of our final model's features are interesting. One thing to note is that after doing our second improvement to the model, you can see that the importance of the features: 'transit_stations_percent_change_from_baseline' and 'parks_percent_change_from_baseline' slightly increased. This can be due to the fact that the predictions we made were quite accurate and this helped the model to determine cases per capita more accurately. The importance of the features for mask usage became higher after improving the model in comparison to the baseline model. In addition, since we removed 'prop_vac' as a feature, 'ALWAYS' became a more important feature overall. The social mobility scores overall became weighted less important in comparison to the baseline model which could be beneficial to the overall model, because it seems like it contains noisier values.

Future Work

Future work could look more into safety precautions and how to slow down COVID-19 cases so that in the case of something like this happening again, we will have a better idea on how to handle the situation and have statistics to back it up. For example, we are currently looking at how these precautions affect the number of cases per capita for each county. We could dig deeper by looking at the rates of change for precautions like mask usage. This way, we would be able to see even more how correlated these precautions are with slowing down the spread of COVID-19. If some of these precautions correlate with a greater rate of change for a decrease in cases, we would know that this precaution is more effective. We would also visualize how having a combination of precautions could result in the greatest rate of change rather than just following one precaution.

Furthermore, there are other precautions we could add to our model such as booster shots

and surveillance testing. It would be interesting to see how these also affect cases per capita as well as the rates of change for cases for each county. Surveillance testing would be especially interesting, as surveillance testing would hopefully result in people being more cautious and being able to not spread the virus if they do have it by being able to know that they do have it. This could be correlated to lower social mobility scores as well. On the other hand, more testing may also just result in more cases as these cases are being detected. This is similar to other precautions as while precautions may intuitively result in fewer cases, more cases may also result in more precautions. Because of this, it would be interesting to model these variables and see how they interact upon one another so that we will be able to better understand how to prevent cases from increasing or spreading.