Assessing the Impact of Cyclical Noise on Machine Learning Outcomes: Investigating Daily vs. Weekly Covid-19 Caseloads as Predictive Factors for Future Caseloads

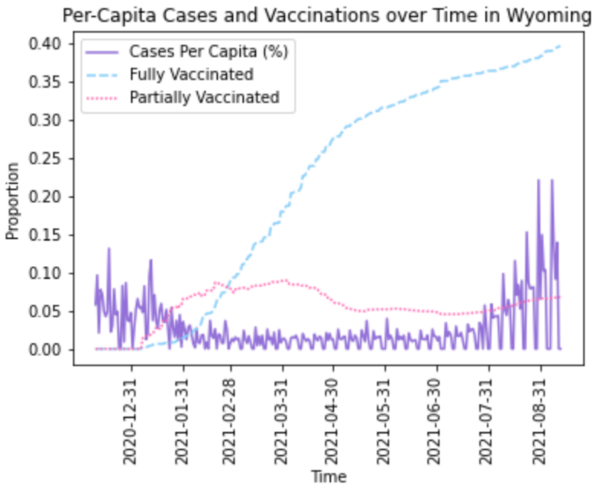
Project Problem, Open-Ended EDA, and Hypothesis

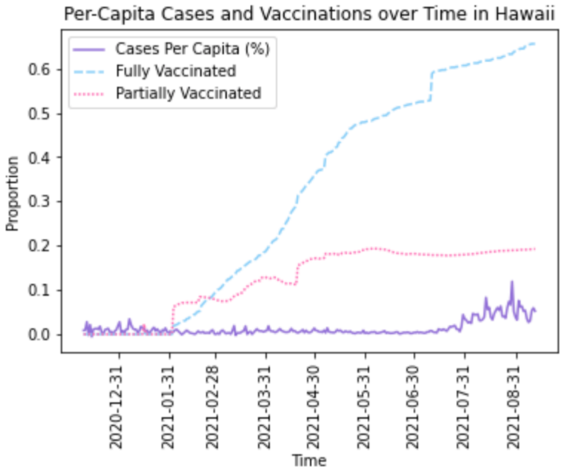
Introduction: Project Problem

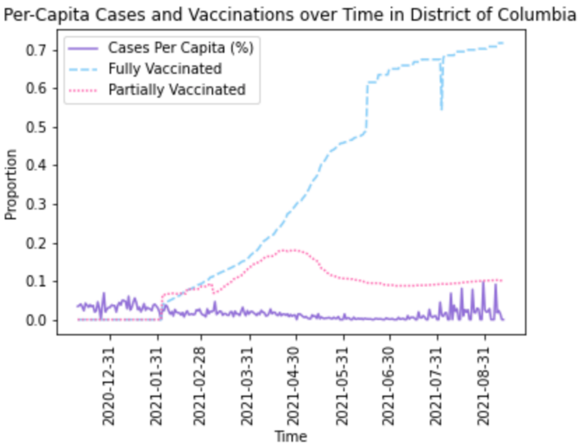
In this project, Aryana Far, Olivia McCauley, and Sydney Loura seek to understand the relative importance of daily versus weekly county-level COVID-19 caseload values in predicting future caseloads. More specifically, our team attempted to assess whether daily or weekly changes in county-level COVID-19 cases functioned as a better predictor of future state-level caseloads.

With the rate of new vaccinations slowing since early-mid summer, new COVID-19 variants such as Delta and Omicron accounting for growing proportions of the total caseload in the United States, and widely-available vaccines possessing decreased efficacy against new variants, the national COVID-19 caseload has spiked significantly since July. It is important to be able to understand the trajectory of the pandemic in terms of caseloads, as being able to predict which counties, and by extension, states, will experience higher caseloads can help with resource and staffing allocation. Being able to properly allocate resources, such as PPE and ventilators, and ensure that hospitals in counties projected to experience higher caseloads have sufficient staffing, can significantly improve COVID-19 patient outcomes, as well as reduce trauma to front-line health responders. Since federal funding for such resources are often allocated to state governments, we determined that it would be an interesting and useful investigation to aggregate county-level predictions across states by averaging these predictions.

Open-Ended EDA







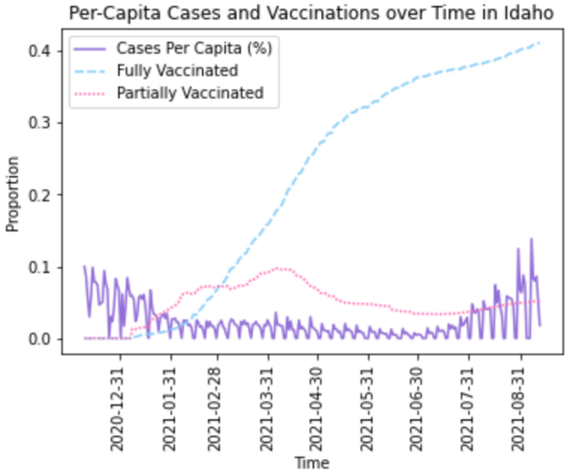


Figure 1 Analysis: In the visualizations above, we see that there are definite linear trends in the cases per-capita for selected states. However, we can also see that there are sharp spikes and drops within those linear trends. This made us interested in how a predictive model that took those variations into account would perform in comparison to a model that used more general trends in per-capita cases. In addition, the linear trends in per-capita cases led us to conclude that multiple linear regression models would work well in this experiment.

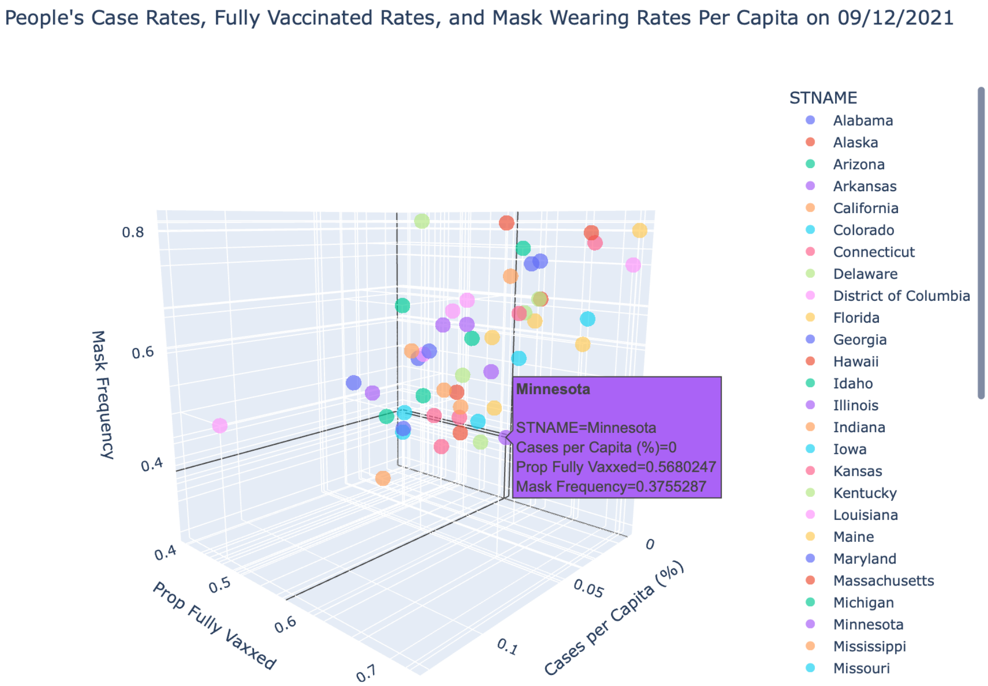

People's Case Rates, Fully Vaccinated Rates, and Mask Wearing Rates Per Capita on 09/12/2021

Figure 2 Analysis: In this visualization, we plotted the proportion of fully vaccinated individuals in a state against state-wide cases per-capita, and state-wide mask use frequency. By rotating this 3-D visualization, created in the open-ended EDA section, we can see that there is a clear positive, linear correlation between these three variables. This further reinforced our decision to use multiple linear regression as our model. (see Open-Ended EDA in part 1 notebook for the interactive version)

Open-Ended Questions: What are some of the context-dependent reasons that might cause COVID-19 to spike more and more day to day?

What are some of the States that we would like to look into more? Are there any outliers in the scatterplot that catch our attention?

What might be the reason for the small drop in Fully Vaxxed folks in the line plots for DC around July of 2021? Might there be missing data there?

Hypothesis

We hypothesize that daily, county-level per-capita changes in caseloads are a better predictor of future per-capita county-level caseload changes than weekly, county-level per-capita changes in county-level caseloads.

Plausible negation of the hypothesis:

It is possible that weekly changes in county-level caseloads are a better predictor of future caseloads, because it could be more representative of the trends in caseloads, which could reduce the impact of day-to-day noise from daily caseload data. For example, maybe significantly more people get tested for COVID-19 on certain days of the week than others, and introduce more caseload variation/oscillation between days that reduces the model's ability to predict the general trajectory of caseloads. See Figure 1 for visualization of varying caseloads over time.

Criteria for Confirming vs. Rejecting the Hypothesis:

Two of the models created for this investigation will be trained and tested on random splits of the data (67/33, training/testing), one using daily, county-level per-capita caseload increases to predict state-wide increases in new cases per-capita. All other included features will be the same. If the model using daily data is significantly more accurate than the model using the weekly data, we would accept the hypothesis. If the reverse was found, then we would reject the hypothesis.

Confirmability of the Hypothesis:

Given that we are testing the relative effectiveness of predicting state-wide increases in new cases per-capita using county-level with daily vs. weekly new cases per-capita, and we have access to all that data, the hypothesis is able to be confirmed or rejected in principle with the datasets provided for the project. However, it is important when comparing these models to understand that changes in past or recent caseloads are not the only factors influencing future caseloads. Other factors, like masking compliance, population density, etc. must be considered in the model in order to put caseloads as a predictive factor in context.

Hypothesis Creativity and Feature Creation:

Neither hypothesis X-variables, per-capita daily changes in case load nor per-capita weekly changes in caseload were original features of the dataset. In order to compute them, we first calculated per-capita total daily cases per date and divided by another feature, POPESTIMATE2020 (Census population estimates for the year 2020), to get the total cases per capita to date for each day present in the dataset. Next, we took the difference of each previous column from the next column to compute the per-capita change by day.

The Y-variable for our model was the per-capita change in caseload for counties, which was calculated in the same way as the X-variables.

We believe that these features represent significant effort to exploit the available features in the dataset, since we wanted to ensure that all predictions would be occurring on the same scale (per-capita) rather than raw case values. In addition, to improve our models in order to better evaluate the hypothesis, we imported two additional datasets. One had national election data for 2020, and the other had geographic data on county sizes.

Modeling

Model 1: Baseline

Notebook Location: Part 2, Problem 6a

Model Type: sk-learn Linear Regression

Explanation of Model Choice:

As observed in the Open-Ended EDA section above, there is a linear association between caseload and other features, and the trajectory of caseloads over time is roughly linear. Subsequently, we determined that Linear Regression was the best choice for modeling.

Model Inputs:

For each state in the contiguous US and Alaska, a new autoregressive model is trained on each county's data. The data for each county was a dataframe, where the rows were all dates between 2020-12-13 and 2021-09-12, and the columns were that date's new cases per-capita and the previous day's new cases per-capita. The models for each county were fitted to a random sample of dates, which constituted 67% of the dates. The X-training variable was the previous date's new cases, and the Y-training variable was that date's new cases. Once the model was trained, the model was tested on 33% of that county's data with the same X and Y variables.

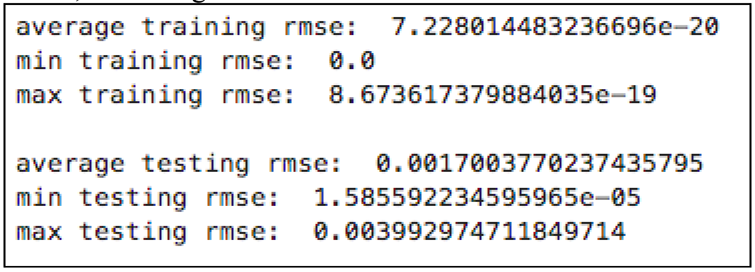
Model Output:

For each county, the model generated a set of predictions for the corresponding test set. It used the previous day's new cases per-capita to generate predictions for the next day's new cases per-capita. The same protocol was applied to generate predictions for the training set. The predictions for the test set and training set, as well as the actual testing and training values, were averaged for each county and stored in an array. The average training and testing RMSEs were then calculated across all counties in each state, creating one average training RMSE and one average testing RMSE value for each state.

Model 1 Evaluation:

Evaluation metric: average testing and training RMSE values across all counties in a state.

Model Performance: Poor, due to high variance. See values below:



```
average training rmse:  7.228014483236696e-20
min training rmse:  0.0
max training rmse:  8.673617379884035e-19

average testing rmse:  0.0017003770237435795
min testing rmse:  1.585592234595965e-05
max testing rmse:  0.003992974711849714
```

As shown above, the average training RMSE is significantly lower than the average testing RMSE. Moreover, the minimum testing RMSE is significantly larger than the maximum training RMSE, indicating that the model performs far better on the training data than it does on the testing data, due to overfitting. These results are visualized below:

Baseline Model for Training Data with RMSEs by State

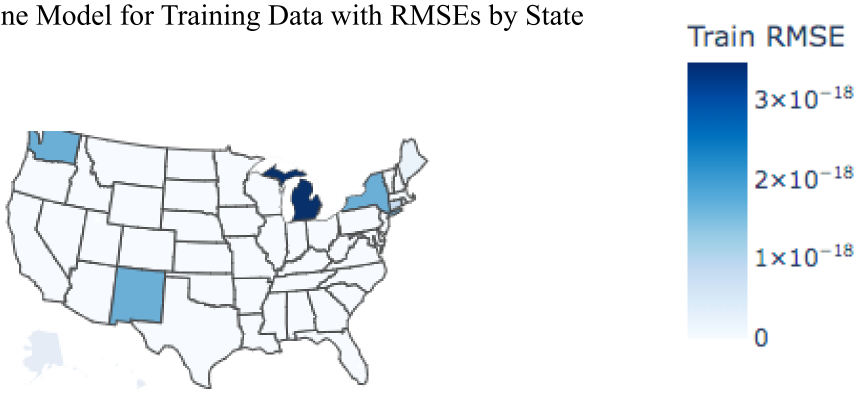


Figure 3 Analysis: The majority of states, as shown in the visualization, have RMSEs of almost zero. This indicates that the model performs extremely well on the training data, as the highest RMSE value is still extremely small. See 6a in the part 2 notebook for interactive visualization.

Baseline Model for Testing Data with RMSEs by State

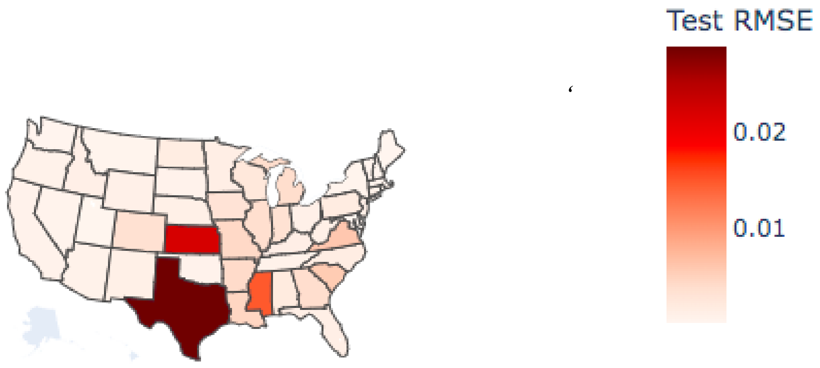


Figure 4 Analysis: The majority of states, as shown in the visualization, still have relatively low RMSEs. However, the lowest RMSE on the test set is still higher than the highest RMSE in the training set. This shows that the model performs significantly worse on the testing dataset. See 6a in the part 2 notebook for interactive visualization.

Overall Takeaways: The states that had the models with the worst testing performance were Texas, Kansas, and Missouri, all of which had RMSEs of 0 on the training data. In contrast, the states with the worst RMSE performance for the training dataset, Michigan, Washington, New Mexico, and New York, had relatively low testing RMSEs. Taken together, the results as visualized in the two plots above indicate that the models were significantly overfitted to the training data, which led to relatively high variance. This indicates that to improve future models, we should attempt to reduce overfitting to the training data. Moreover, the absolute values of the test RMSEs were high compared to the per-capita caseload values, as both RMSE and per-capita caseload values are primarily decimal numbers with non-zero numbers starting in the 10s-place. This indicates that the test RMSEs are on the same scale of magnitude as the per-capita caseload values themselves, which shows extremely poor model performance.

Model Improvements

Improvement 1: PCA as feature engineering

Notebook Location: Part 2, Problem 6b

Problem: The baseline model predicted new cases per-capita for each day based on the previous day's new cases per-capita. However, this led to overfitting, and one possible reason identified was the lack of other features in the model to prevent the model from relying to heavily on the previous day's data.

Solution: To resolve this issue, we performed PCA to determine which other features from the data might be useful. Since relevant features might vary from week to week, it was decided that predictions for the improved models would be generated on a week-to-week basis. To determine relevant features on a weekly basis, we created two functions:

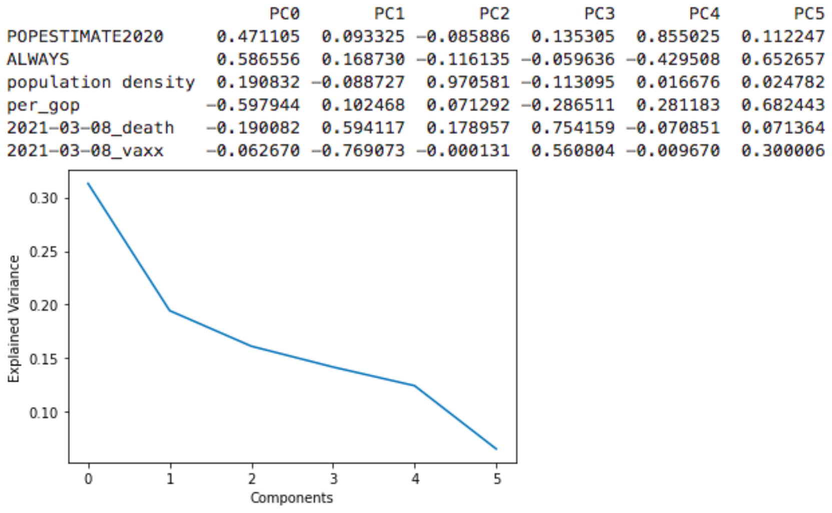
make_features_df(week):

> *Takes in a datetime object, and creates a table with the following features: 'STATE', 'STNAME','POPESTIMATE2020', 'COUNTYFP', 'ALWAYS', 'population density', total per-capita vaccinations to date, total per-capita deaths to date, and the proportion of GOP (Republican) votes in each county.*

pca(week):

> *Takes in a datetime object, and uses the make_features_df function above to generate a df for that week. Performs PCA on that df, returns principal component values and creates a scree plot.*

Result: PCA was performed on the week of 2021-03-08, which was the week we used when running our linear regression models (except the baseline model). Then, the principle components visualized in a scree plot:

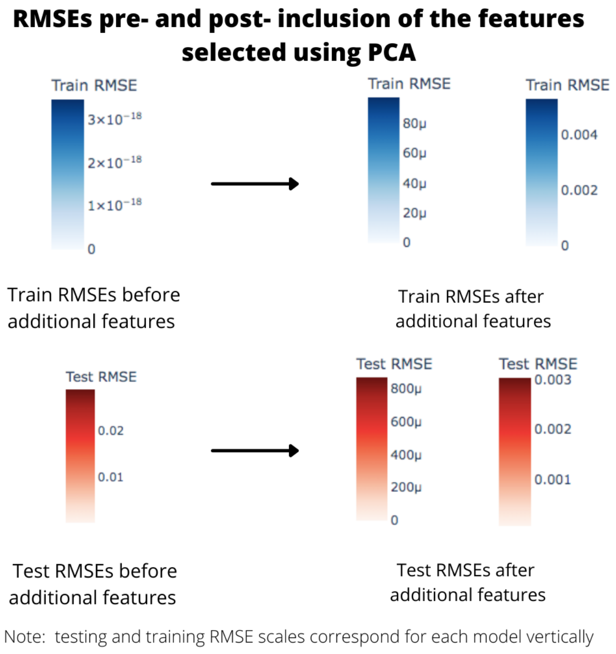


```
                         PC0       PC1       PC2       PC3       PC4       PC5
POPESTIMATE2020     0.471105  0.093325 -0.085886  0.135305  0.855025  0.112247
ALWAYS              0.586556  0.168730 -0.116135 -0.059636 -0.429508  0.652657
population density  0.190832 -0.088727  0.970581 -0.113095  0.016676  0.024782
per_gop            -0.597944  0.102468  0.071292 -0.286511  0.281183  0.682443
2021-03-08_death   -0.190082  0.594117  0.178957  0.754159 -0.070851  0.071364
2021-03-08_vaxx    -0.062670 -0.769073 -0.000131  0.560804 -0.009670  0.300006
```

Figure 5 Analysis: There is no clear "elbow" in the scree plot of principal components, indicating that all features included in the PCA were significant. As a result, we decided to include all of the features that we performed PCA on.

When these features were included in the subsequent models, the absolute values of the test RMSEs decreased, as well as the difference in scale between training and test RMSEs, decreased significantly. The decreased scale of RMSE values indicates that the absolute scale of error decreased significantly, while the convergence of test and training RMSE values in terms of scale indicates that overfitting was also reduced. This pattern is illustrated below:

Figure 6:



Note: testing and training RMSE scales correspond for each model vertically

Taken together, the convergence of test and training RMSE values and decreasing scale of the test RMSEs indicate a significant improvement in model performance.

Improvement 2: Using Weekly Data instead of Daily Data to reduce overfitting

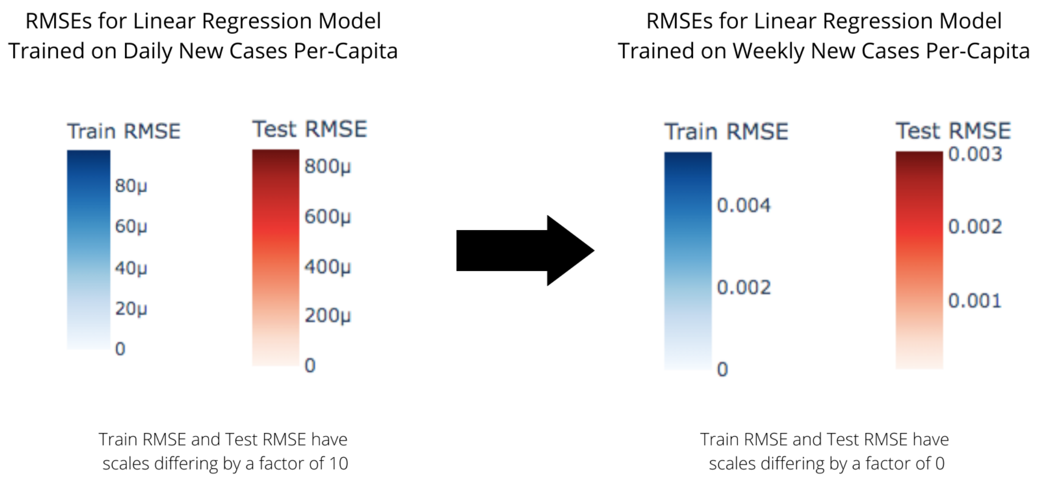
Notebook Location: Part 2, Problems 6c and 6d

Problem: In order to study our hypothesis, we created two models (See our Testing the Hypothesis using Modeling section for more details). While both models used the features selected using PCA, one model used all weekly, county-level changes in per-capita caseload changes to date to predict per-capita county-level caseloads for the selected date. The other model used all daily, county-level changes in per-capita caseloads to predict per-capita county-level caseload changes for the selected date. Based on the difference in scale between the test and training RMSEs, we suspected that the model using daily new cases per-capita was overfitting.

Solution: We believed that since new weekly cases per-capita represents change over a longer time-period, a model trained on this data might have lower variance than the model trained on daily case change data. In order to assess this, we compared the difference in scale between the testing and training RMSEs for both models.

Results: Our analysis shows that training the model with weekly, rather than daily, per-capita caseload change data did decrease overfitting. The results are illustrated below:

Change in RMSE scales between Daily and Weekly Models



As illustrated in the visualization above, training and testing the model on weekly data significantly reduced the difference in scale between the training and testing RMSEs. However, there is a tradeoff with this reduction in overfitting, namely, that the absolute scale of the RMSE values increased by a factor of 1000 when using weekly data. Depending on how this model was being put into use, users might prioritize consistency of error between training and testing datasets, or prefer smaller absolute RMSE values in general (See the following section for more details on how our team evaluated this tradeoff).

Testing the Hypothesis using Modeling

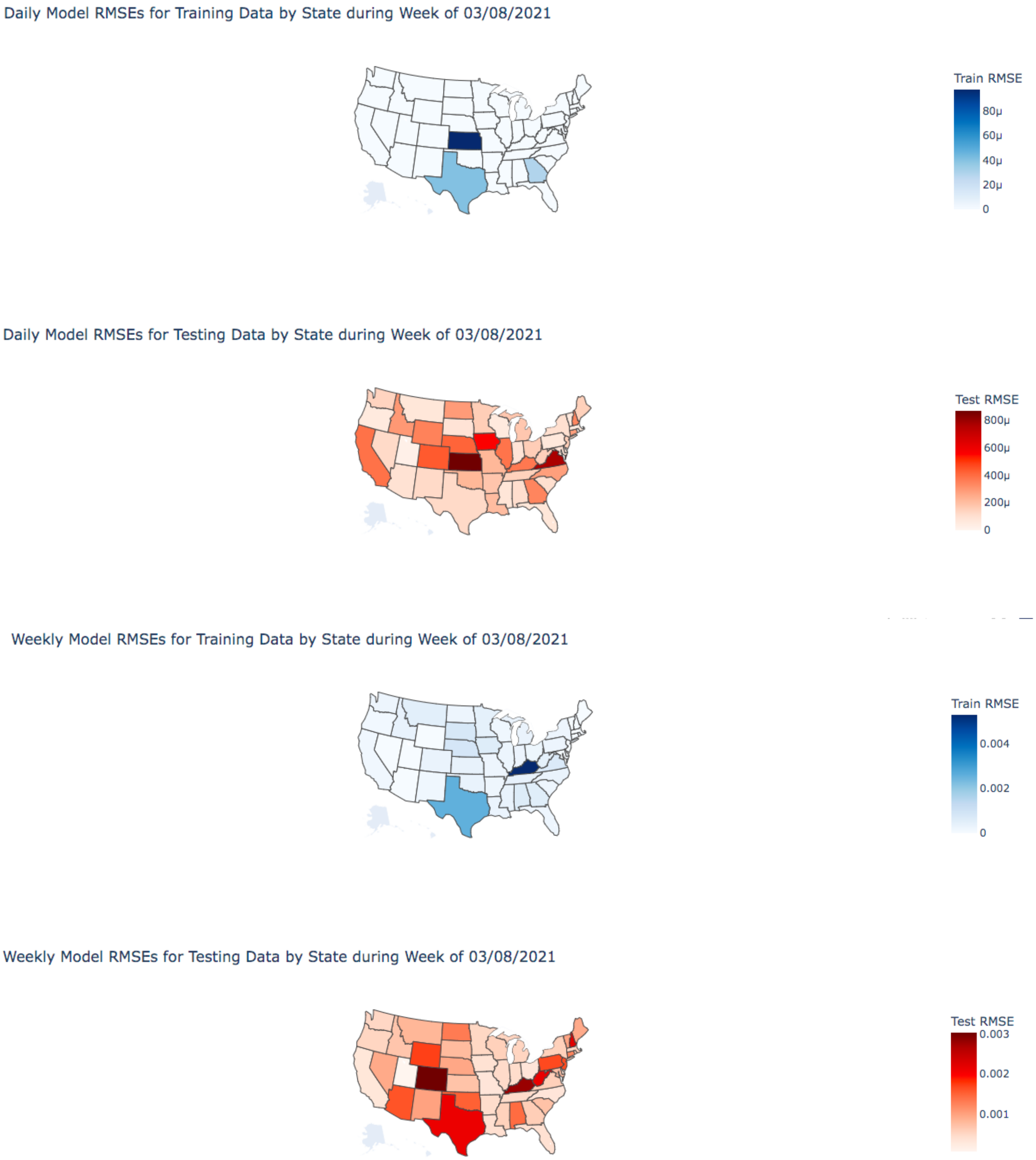
Experimental Structure

In order to test our hypothesis, we created two models. One model was trained on data with a weekly temporal split, while another model was trained with a daily temporal split. Both weekly and daily data represented the per-capita caseload change for all counties in the contiguous US, as well as Alaska. As determined above in *Criteria for Confirming vs. Rejecting the Hypothesis,* both models were trained and tested on the same random sample from the data (which included the features added after PCA selection, see Improvement 1 for more details), and then compared against each other to determine the conclusion of the experiment.

Results & Visualizations

The experiment did not show that one model was strictly better than the other. Rather, the model results showed that there is a tradeoff between absolute model accuracy, and consistency in prediction accuracy between training and testing sets.

A single date was used to visualize each model's results to keep the interactive map interpretable:



Daily Model RMSEs for Training Data by State during Week of 03/08/2021

Daily Model RMSEs for Testing Data by State during Week of 03/08/2021

Weekly Model RMSEs for Training Data by State during Week of 03/08/2021

Weekly Model RMSEs for Testing Data by State during Week of 03/08/2021

Figure 7 Analysis: The plots above show that the model performed significantly better with the daily data. See Figure 6 for more in-depth analysis on the RMSE values, see 6c/6d in the part 2 notebook for interactive version of the maps.

The models were trained and tested further on every county in the contiguous United States and Alaska for every week between 2020-12-13 and 2021-09-12. The results for each week were aggregated by averaging the RMSE for each county. The results are plotted below:

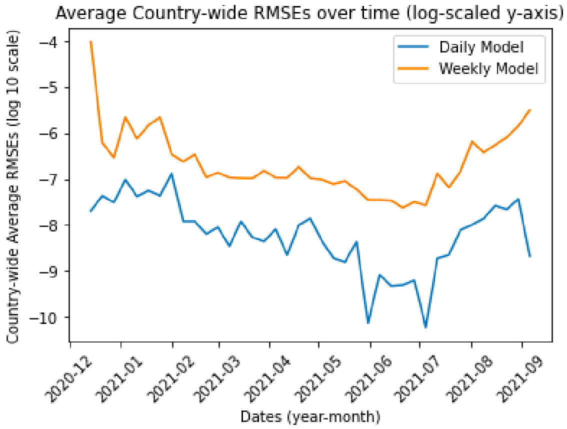


Figure 8 Analysis: The y-axis was scaled by log-10 for interpretability. The graph largely confirms the expectation that the weekly model has higher absolute errors, but lower variance. The daily model has lower absolute error, but higher variance.

Conclusions Regarding the Hypothesis

While either model could be considered "better" depending on how it was utilized for external purposes, we concluded that the model trained with a daily temporal split performed better than the model trained with a weekly temporal split. We came to this conclusion primarily because of the absolute size of the RMSE values for the model trained using a daily temporal split. Even though the model trained and tested on the daily temporal split (referred to as the "daily model" from here on) had significantly higher variance between the testing and training set prediction accuracies, the absolute size of those errors was still extremely small, especially in comparison to the actual per-capita caseload change values.

Specifically, the RMSE values for the daily model's training set predictions ranged between 0 and $\sim\ 100\mu$, while the RMSE values for the testing set prediction ranged from~ $100\mu$to $\sim\ 800\mu$. The scale between them is different by a factor of 10 between the training and testing RMSEs, but the absolute size of the RMSE is extremely small. The RMSE values for the model trained and tested on the weekly temporal split vary less between testing and training set predictions, with the training RMSEs ranging between ~0 to ~0.005, and the testing RMSEs ranging between ~0 to ~0.003. The RMSEs of training and testing splits for the weekly model are on the same scale, however, the absolute size of the weekly model's RMSEs is significantly bigger than the daily model's — a difference of scale by a factor of 1000. We concluded that given the scale factor differences, and the absolute size of the RMSEs for the daily model, that the model trained on the daily temporal split performed significantly better than the model trained on the weekly temporal split.

Thus, we conclude by asserting that our data supports the hypothesis: that daily, county-level per-capita changes in caseloads are a better predictor of future per-capita county-level caseload changes than weekly, county-level per-capita changes in county-level caseloads.

Future Research

One way that this model could potentially be improved would be to aggregate the county-level predictions into state-level predictions. While we averaged RMSE values across counties as a metric for how accurate the model was for predicting caseloads in each state, our predicted values were per-capita. It could be more useful and accurate to weight these per-capita county-level predictions of caseload change over time during aggregation. For example, weighting by ratio of county residents to hospital beds could emphasize model accuracy in areas where hospitals are more likely to become overwhelmed. This could improve the usefulness of the model, and help better inform resource allocation on the federal level.