Spring 2023 Data C102 Final Project

Research Topic: Transportation & Human Mobility in Covid Period

Jie Zheng
Jiaping Huang
Qingyi Fang
Jiacheng Yao

Contents

Data Overview

We explored three datasets: daily mobility, covid, and monthly transportation data.

Data Source and Bias Discussion
The daily community mobility data was generated by Google and aggregated from usage data in products such as Google Maps. It is a sample that represents people's movement trends over time. However, since the data was generated using only Google Maps users' information, we are concerned about selection bias and convenience sampling. People who own a smartphone and use Google service might behave and react differently to Covid than people who can not afford a smartphone.

The covid data was generated by Opportunity Insights to track the economic impacts of Covid-19. It is generated by combining a lot of publicly available information during the pandemic period such as infection and vaccine data published by CDC and WHO. The potential bias we should be considering here would be that some people might be positive but did not get a test. We should also think about the performance of different covid tests, and might need to deal with problems such as false positive and false negative results.

The monthly transportation data was published by the Bureau of Transportation Statistics. The Bureau of Transportation Statistics made this data by keeping track of various factors of transportation such as total airline and total railway transportation. We will focus on the total airline in our study.

Privacy Discussion
- For the mobility dataset, Google Maps users agree to participate in this data by accepting the terms of service and they have the option to opt out at any time, and the data was an aggregated result to preserve anonymity.
- For the covid dataset, the public were aware of the release of this kind of data. Since the data were aggregated to reflect the situation of a region or the country, they could not be linked back to an individual, and the aggregation process also applied differential privacy to preserve privacy.
- For the transportation dataset, the data was gathered from the performance of different transportation industries in the United States. Each individual participates in the data by taking public transportation or placing online shopping orders. However, in this case, the data eventually reflect how people in the whole country are doing rather than how each individual is doing. Therefore, we do not need to worry too much about personal privacy being violated here.

Granularity
- Each row in the mobility dataset represents the mobility information on one date for a specific state in the U.S.
- Each row in the covid dataset contains information on Covid indicators such as vaccine rate and hospitalized rate on one day for a specific state in the U.S.
- Each row in the transportation dataset contains information on the total airline for one month for a specific state in the U.S.

Data Cleaning and Preprocessing

We looked up documentation of each dataset and replaced the numeric column state fips that represent states with the abbreviations of each state. We then merged the covid data with the daily mobility data on date so that each row contains covid and mobility information on one day for a specific state in the U.S. There are some missing values for hospitalized_rate and we decided to drop those rows. We observed that there are some very extreme outliers for multiple columns, so we filtered rows that are between quantile(0.1) to quantile(0.9) for each column. We also created a season column in order to observe if there is any seasonal effect on variables.

In addition, we noticed that each row in the covid dataset represents a day while each row in the monthly transportation dataset represents a day. We would like each row in our data frame this part to represent a month. In order to do this, we wrote a function to figure out how many month_after_2020 for a given row and applied it to the covid dataset. We filtered out rows that are between Jan 2020 and Dec 2022 since our study would only focus on the Covid-19 period. We then aggregated the hospitalized_rate by the maximum of a month because we would like to see how bad it is for the worst day in a month. We then created another data frame by merging them so that each row now contains information for one month for a specific state in the U.S. We only kept the following columns that are relevant to our interests: state, month_after_2020, hospitalized_rate, and airline.

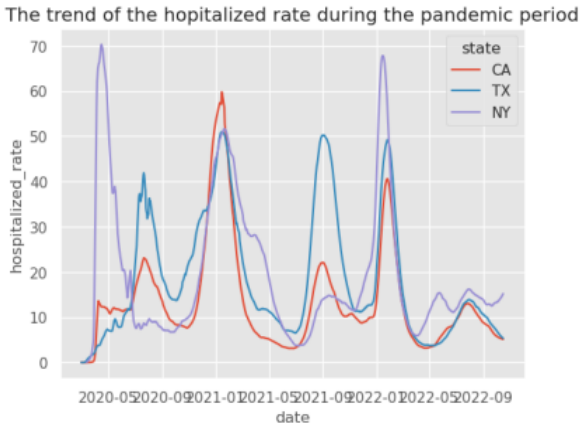
	state	month_after_2020	hospitalized_rate	airline
0	AL	4	10.10	3010000.0
1	AL	5	12.90	8050000.0
2	AL	6	17.70	16530000.0
3	AL	7	43.00	24060000.0
4	AL	8	43.00	25830000.0
...	...	...	...	...
1423	WY	27	9.11	72480000.0

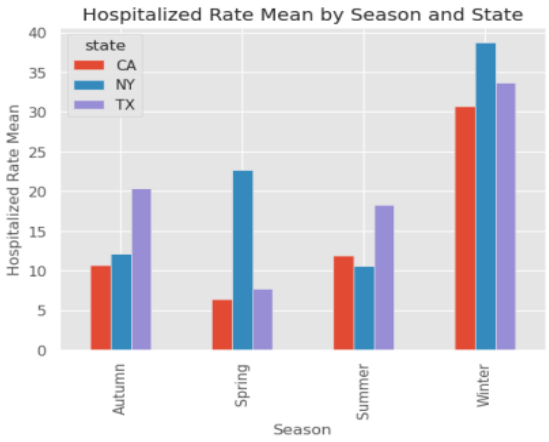
Research Question

1. How will the hospitalization rate change in response to human mobility?
2. Among different states in the United States, is there a significant association between the total airline traffic and the hospitalized rate in the pandemic period?

EDA

Question 1





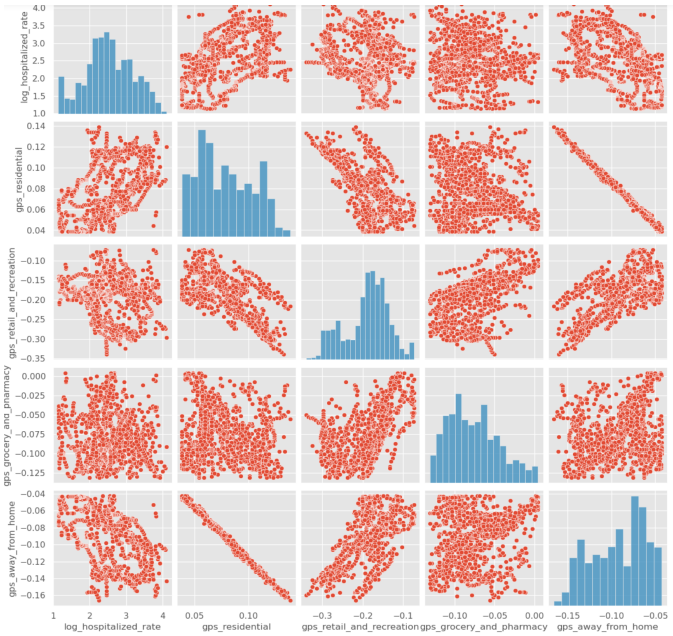
Trends and Relationships:
From the above plots, we saw that hospitalized rates between states are significantly different, NY had the highest hospitalized rate in the beginning of the pandemic period. We also observed that the hospitalized rate has some seasonal effects, where Winter has the highest hospitalized rate.

Data Cleaning Steps:
After observing the different hospitalized rates between states, we decided to one-hot encode the state and use it as one of the features in our prediction model. We also add season as one of the features in our prediction model.

Why are these relevant:
For our first research questions, we would like to predict hospitalized rates from many different variables. The visualizations suggest that we can use state and season as features in our prediction model.

Trends and Relationships:

- From the top 5 plots of the pair plot on the right, we can see that the linearity between the response variable and explanatory variables is not clear.
- Also, we can see some significant collinearity between gps_residential and gps_away_from_home.
- From the beginning of the EDA process, we also observed that some variables have very extreme outliers and we now get rid of those so that they don't have much outliers now.
- From the beginning of the EDA process, we observed that the distribution hospitalized rate follows some kind of exponential distribution.



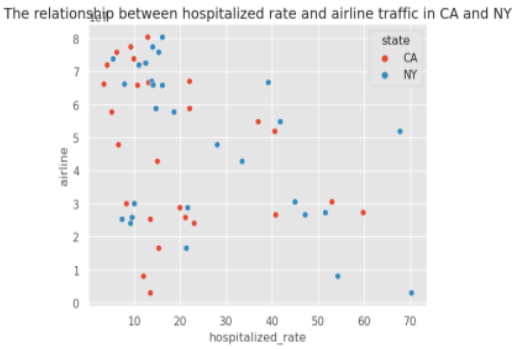
Data Cleaning Steps:

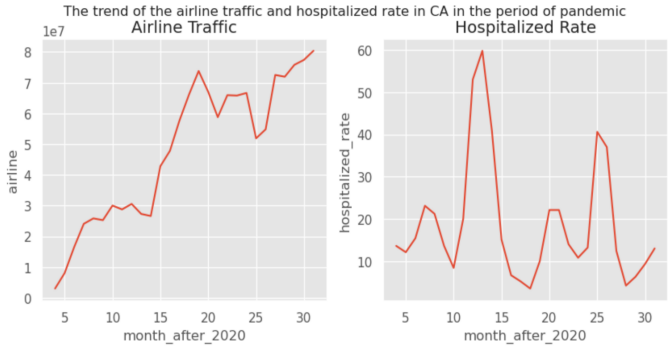
- After observing that two variables are negatively correlated, we decided to only include one of them in our prediction model to avoid multicollinearity.
- We removed outliers for our feature variables and only kept quantile(0.1) to quantile(0.9).
- We performed log transformation for hospitalized rate so that the distribution of our target variable now looks like normal distribution.

Why are these relevant:

- By making this pairplot, we can visualize any correlation between variables and the distributions of each variable. It can also help us validate some of our assumptions for GLMs. We would like to see if any variables have a strong linear relationship with hospitalized rate which is the prediction target of our model.
- The visualization suggests that we may need to apply feature engineering to some of the variables such as gps_grocery_and_pharmacy and hospitalized_rate.
- We also realized that there are some outliers that can become noisy data and affect the performance of the model. Removing outliers should ideally make our model perform better.

Question 2





Trends and Relationships:

- In the first plot above, we choose two states (CA and NY) to observe if there is any clear association between hospitalized_rate and total airline. We can see that there is a very weak negative correlation between the two variables.
- We also specifically choose CA to see if there exist seasonality effects in both variables to better study any association between them. We can see that total airline traffic was low during the beginning of the pandemic since many places have lockdown policies and people started to work from home which would decrease the consumption of air traffic. As it entered the middle stage of the pandemic, the airline traffic started to gradually increase, and this was about the time vaccines and boosters became available. People started to plan their travel as more people were fully vaccinated and the airline policy in response to Covid became less strict.
- We also observed that the hospitalized rate was highest near Dec 2020, and decreased rapidly since Feb 2021. We think that this is also due to the release of vaccines and boosters which make people get less sick. After being fully vaccinated, people were likely not ended up to be hospitalized after they caught covid.

Why are these relevant:

- The observation that hospitalized rate and airline have weak association motivates the decision of our second research question. It inspired us to conduct a multiple-hypothesis test for each state to study if this kind of correlation is significant.
- Our group thought that it makes sense for airline traffic to be low when the pandemic got worse (indicated by the hospitalized rate). We would like to use the opportunity of this project to validate our hypothesis.

Q1: Prediction with GLMs and nonparametric methods

In this project, we aim to predict hospitalization rates using human mobility data, states, and seasons. We are using data that has mobility information from GPS, and our analysis suggests that these factors could influence hospitalization. To make accurate predictions, we are testing three methods: Frequentist GLM, Bayesian GLM, and a Non-parametric approach. Since hospitalization rates are a continuous variable. To evaluate our models, we'll use metrics like MSE, RMSE, MAE, and R-squared. By calculating these for both training and testing sets, we can assess potential overfitting and overall performance.

GLMs

Model Choice:

- Frequentist GLM with normal likelihood
- Bayesian GLM with normal likelihood and chosen prior distribution

Feature Choice: In Frequentist GLM method, we include gps_retail_and_recreation, gps_grocery_and_pharmacy, gps_parks, gps_transit_stations, gps_workplaces, gps_residential, gps_away_from_home, NY, TX, Autumn, Summer and Winter as the features we use. The NY, TX, Autumn, Summer, Winter are the dummy variables we created. Besides, we take the log transformation for hospitalized_rate to be the response variable

Assumptions: For this model, after taking the log transformation for the hospitalized rate, it is reasonable to make the assumption of our response variable that the log_hospitalized_rate is normally distributed. Besides, we make the assumption of the linearity between the features and the response variable.

Frequentist GLM

Results:

Generalized Linear Model Regression Results

Dep. Variable:	log_hospitalized_rate	No. Observations:	1112
Model:	GLM	Df Residuals:	1099
Model Family:	Gaussian	Df Model:	12
Link Function:	identity	Scale:	0.17732
Method:	IRLS	Log-Likelihood:	-609.55
Date:	Sun, 07 May 2023	Deviance:	194.87
Time:	00:29:56	Pearson chi2:	195.
No. Iterations:	3	Pseudo R-squ. (CS):	0.7570
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.3102	0.169	7.751	0.000	0.979	1.642
gps_retail_and_recreation	1.9679	0.659	2.988	0.003	0.677	3.259
gps_grocery_and_pharmacy	-2.4208	0.785	-3.083	0.002	-3.960	-0.882
gps_parks	-0.3034	0.148	-2.053	0.040	-0.593	-0.014
gps_transit_stations	-0.5547	0.487	-1.139	0.255	-1.510	0.400
gps_workplaces	-1.2821	0.606	-2.116	0.034	-2.470	-0.095
gps_residential	-93.6973	9.791	-9.569	0.000	-112.888	-74.507
gps_away_from_home	-83.2411	7.750	-10.740	0.000	-98.431	-68.051
NY	0.6450	0.045	14.377	0.000	0.557	0.733
TX	0.5666	0.088	6.465	0.000	0.395	0.738
Autumn	0.2945	0.035	8.347	0.000	0.225	0.364
Summer	0.3737	0.043	8.636	0.000	0.289	0.459
Winter	0.7439	0.048	15.340	0.000	0.649	0.839

The picture above is from the summary of the Frequentist GLM. We can see that the estimated parameters for gps_grocery_and_pharmacy is -2.4208, meaning that with one unit increased in this feature, the

log_hospitalized_rate would decrease 2.4208 and for the original prediction of hospitalized rate, it would increase 8.3%. Also, we can observe the Log-likelihood is -609.55 and the Pearson chi2 is 195. Since there is some randomness in the data sample, in this model, we would introduce the confidence interval to measure the uncertainty for the estimations of coefficients.

Uncertainty Quantification

As we can see above, the uncertainty in this model for coefficients of some features are bigger than we expect. Let's take a closer look at the confidence interval of the feature gps_park. The true parameter of this feature is 95% confident lying on the interval [-0.593, -0.014].

	coef	CI_lower	CI_upper
const	1.310234	0.978903	1.641564
gps_retail_and_recreation	1.967904	0.676890	3.258919
gps_grocery_and_pharmacy	-2.420846	-3.959824	-0.881869
gps_parks	-0.303406	-0.593012	-0.013799
gps_transit_stations	-0.554715	-1.509518	0.400088
gps_workplaces	-1.282132	-2.469518	-0.094745
gps_residential	-93.697349	-112.888198	-74.506501
gps_away_from_home	-83.241056	-98.431341	-68.050770
NY	0.645027	0.557096	0.732958
TX	0.566630	0.394857	0.738404
Autumn	0.294488	0.225335	0.363640
Summer	0.373722	0.288901	0.458542
Winter	0.743887	0.648844	0.838930

Reflection: To better measure the performance of this model, we introduce RMSE for training data and for test data here. The RMSE for training data is 6.6688 and the RMSE for test data is 6.3267. Clearly, the RMSE is large but reasonably convincing. Meanwhile, for a better look at if the Frequentist GLM model in this set up is a good fit for the data and the new data (in this case, for the test data), we would visualize the fitting by plot fitted value versus the observed value in test set(in log scale). From the result of RMSE and visualization, we would claim this Frequentist GLM model does not make a good enough fit for the data.

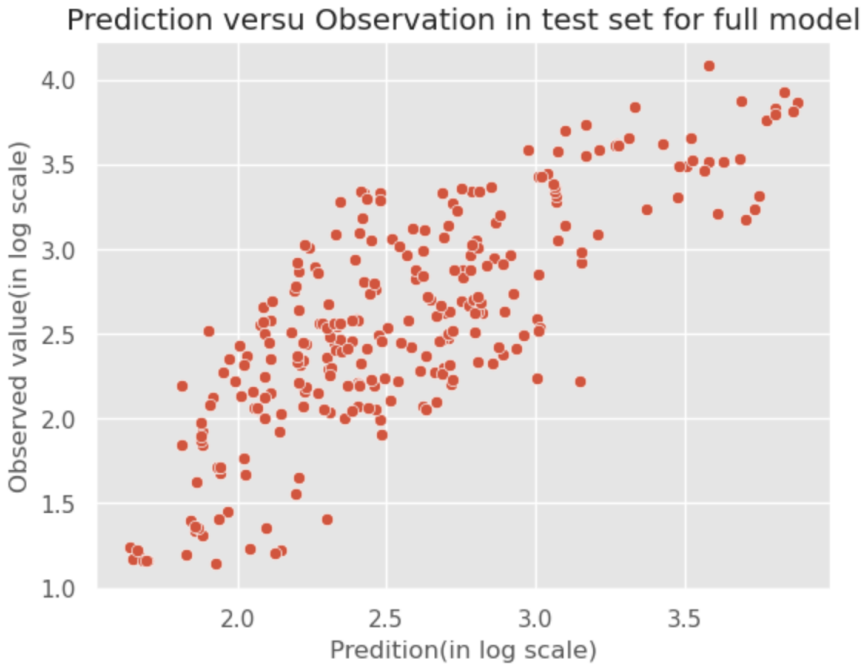
Bayesian GLM

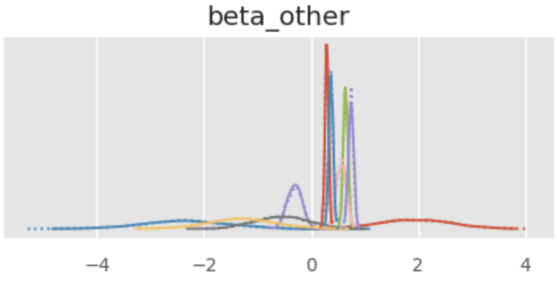
Prior Choice: we choose the prior for the coefficients of features instead of defaultly choosing "flat" prior. The chosen prior distributions are based on the result of the Frequentist GLM. From the result of Frequentist GLM, we would say the coefficients of gps_residential and gps_away_from_home are normally distributed with mean as -90 and standard deviation as 1000. The coefficient of the intercept is normally distributed with mean as 1 and standard deviation as 10. The coefficient of the sigma of the likelihood of y is half-normally distributed with standard deviation as 1000.
Assumptions: Since the Bayesian model would produce a predictive posterior distribution for each prediction and each coefficient we estimated and as we mentioned before, we assume the response

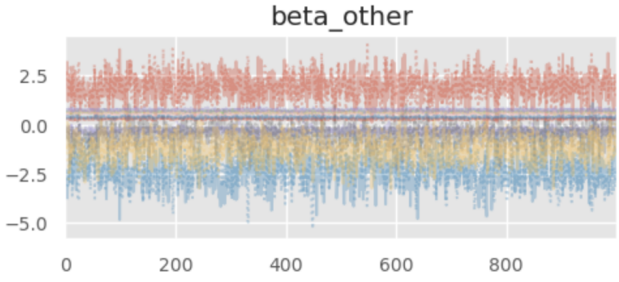
variable and the coefficients are normally distributed, it is reasonable to take the mean of the predictive posterior distribution as representation to measure the performance of the Bayesian model.

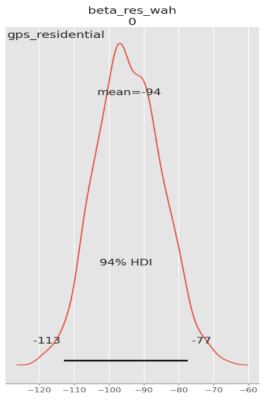
Results

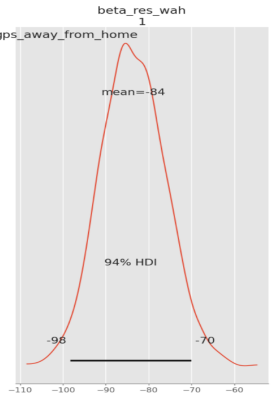
In this method, we use credible intervals for prediction and the coefficients to measure the uncertainty. As we learned, for the estimated coefficients, the randomness is from the drawing sample for the coefficient in the posterior distribution and for the prediction, the randomness is from the coefficient which captures the variation on average y and from the variation of the actual observations around its average y. The pictures below are the result of the Bayesian GLM model, providing the predictive posterior distribution for each parameter and its credible interval.

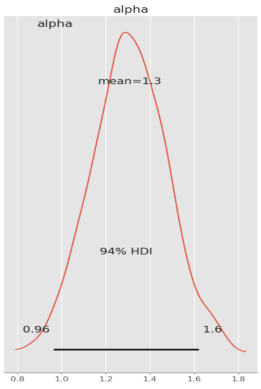


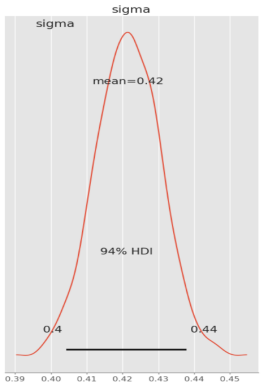


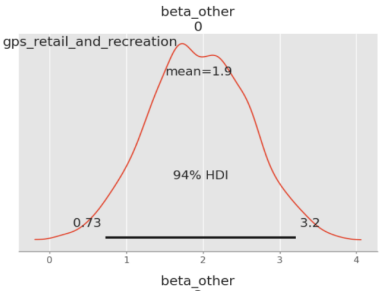


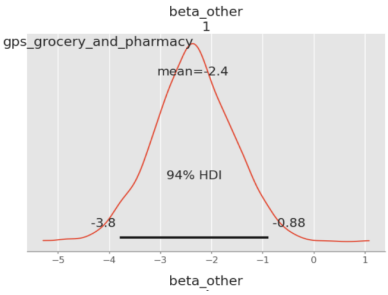

beta_res_wah
0
gps_residential
mean=-94
94% HDI
-113
-77

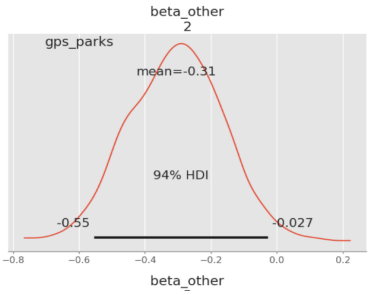

beta_res_wah
1
gps_away_from_home
mean=-84
94% HDI
-98
-70

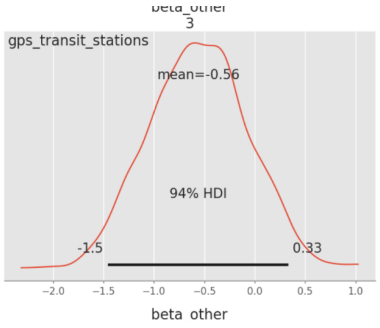

alpha
alpha
mean=1.3
94% HDI
0.96
1.6

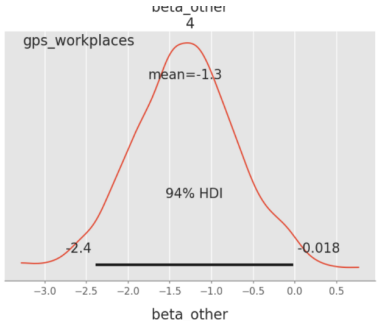

sigma
sigma
mean=0.42
94% HDI
0.4
0.44

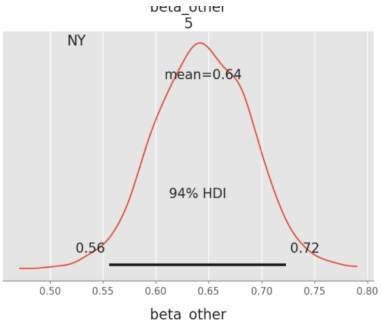

beta_other
0
gps_retail_and_recreation
mean=1.9
94% HDI
0.73
3.2
beta_other

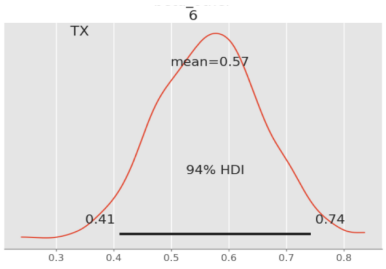

beta_other
1
gps_grocery_and_pharmacy
mean=-2.4
94% HDI
-3.8
-0.88
beta_other

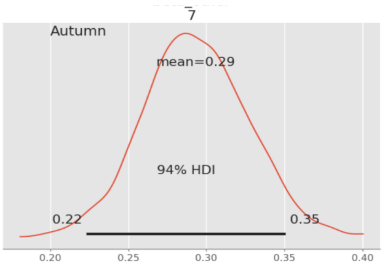

beta_other
2
gps_parks
mean=-0.31
94% HDI
-0.55
-0.027
beta_other

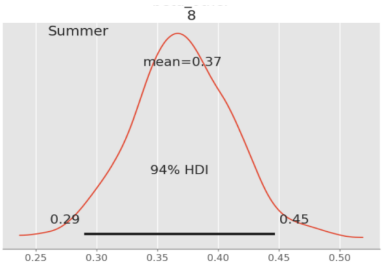

3
gps_transit_stations
mean=-0.56
94% HDI
-1.5
0.33
beta_other

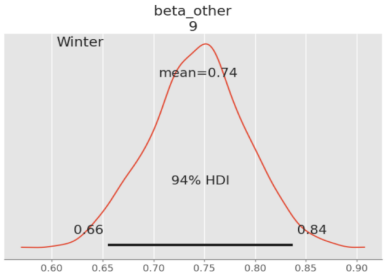

4
gps_workplaces
mean=-1.3
94% HDI
-2.4
-0.018
beta_other


5
NY
mean=0.64
94% HDI
0.56
0.72
beta_other


6
TX
mean=0.57
94% HDI
0.41
0.74


7
Autumn
mean=0.29
94% HDI
0.22
0.35

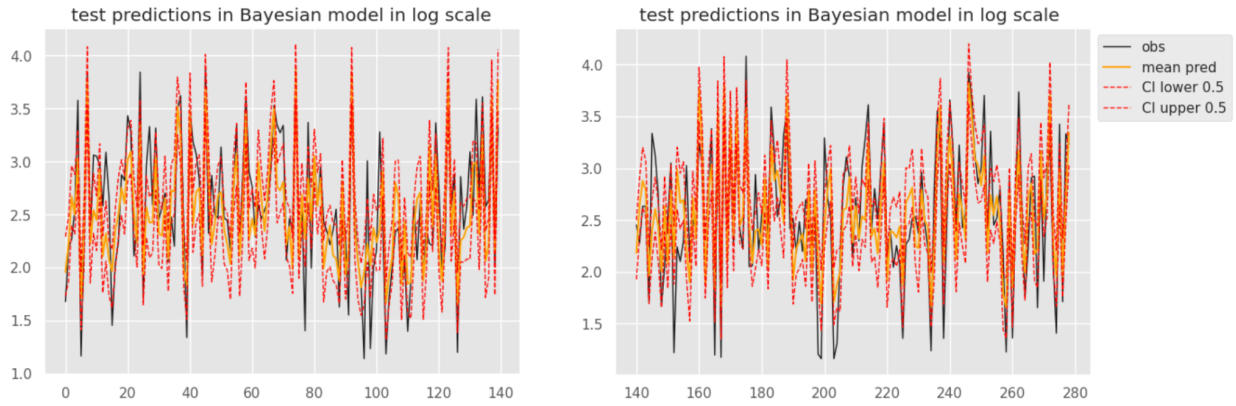

8
Summer
mean=0.37
94% HDI
0.29
0.45


beta_other
9
Winter
mean=0.74
94% HDI
0.66
0.84

On the other hand, for the prediction in test data, we also can tell the performance of this model by visualization and the credible intervals for predictions.

Uncertainty Quantification

Here, learning from other experiments, we would introduce other kinds of credible intervals. The result is shown below.(Just the the top 5 results)



	log_hospitalized_rate	pred_mean	CI_lower	CI_upper
0	1.678964	1.956857	1.730655	2.290486
1	2.187174	2.233860	1.874260	2.475247
2	2.271094	2.667031	2.396011	2.959299
3	2.533697	2.497308	2.316746	2.877062
4	3.575151	3.027670	2.779280	3.296764

Reflection: Same as the part in the Frequentist model, we plot the fitted value and the observed values in test data to see if the model is good to predict some new data. We also compute the RMSE for training data and test data. The RMSE for training data is 6.6796 and the RMSE for test data is 6.3138.

Non-parametric Method

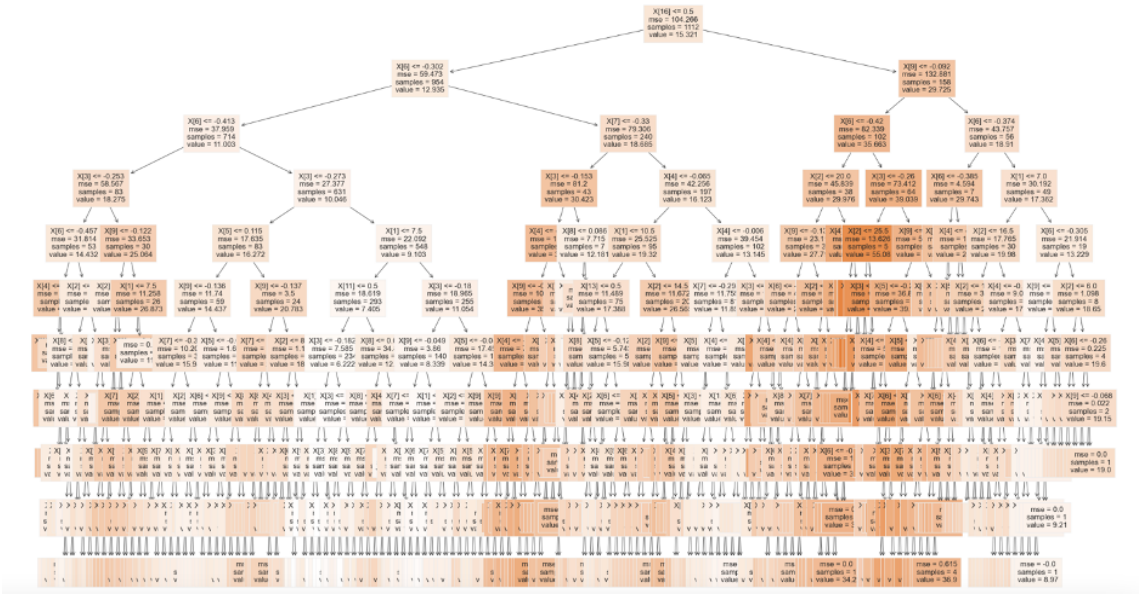
Model Choice: Nonparametric methods we are using in this project are Decision Tree Regressor and Random Forest Regressor. Both models can model complex relationships and interactions between variables. Decision Tree Regressor can be interpreted by visualization, and it is a simple but powerful model that could reduce the features by itself. Meanwhile, Random Forest Regressor often produces better results than Decision Tree due to its reduction in the variance in the predictions by averaging the results of multiple decision trees, which leads to less chance of overfitting.
Assumptions: The assumption being made by both Decision Tree and Random Forest is that the relationships between the input features and the outcome variable are non-linear and complex.

Decision Tree

Assumptions: The data can be partitioned into parts with similar response values.

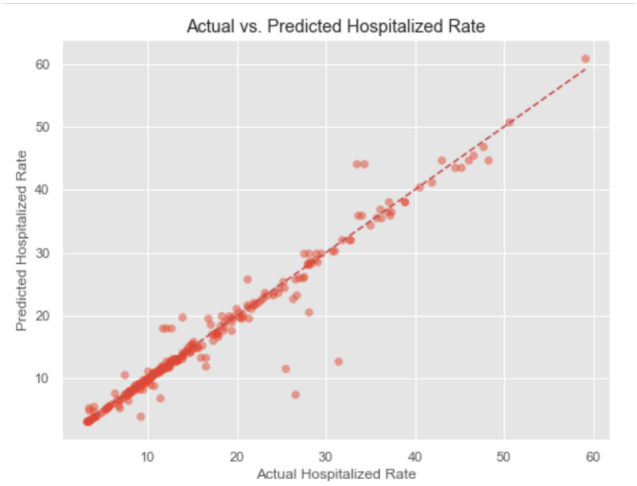
Results

Training set RMSE is 0.742, and test set metrics are MSE is 5.626, RMSE is 2.372, MAE is 0.912, and R2 is 0.949. As we can see from the metrics, the test set RMSE is higher than training set RMSE, but still is relatively low, which indicates that the Decision Tree Regressor is fitting the seen data well, and generally well to unseen data. The R2 value of the testing set is 0.949, which means that the model explains about 94.9% of the variance in hospitalized rate. The MAE indicates that the model's predictions deviate from the actual values by approximately 0.912 units on average, which is generally low.



As we can see from the above graph, the first few layers include the most important features which are X[16], X[9], X[6], X[7], X[3], X[4], X[2], X[1], which corresponding to features Winter, gps_away_from_home, gps_transit_stations, gps_workspaces, gps_retail_and_recreation, gps_grocery_and_pharmacy, day, and month, which make perfectly sense and support the initial

hypothesis since both human mobility features and geographical features are important in predicting hospitalized rate.



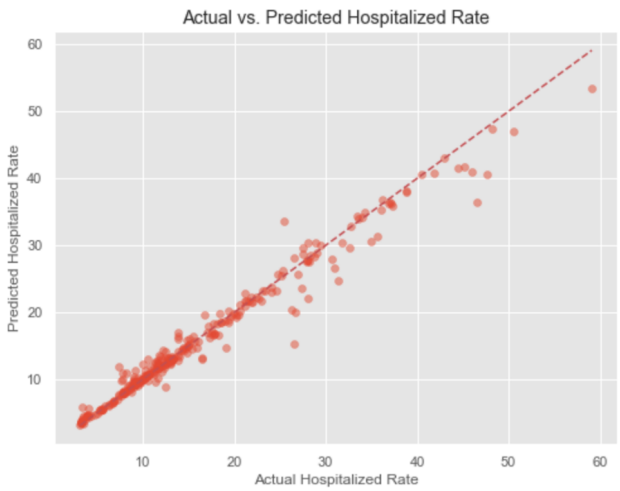
Lastly, as we can see from the above graph, the result for the decision tree is performed well since it joins the y=x line pretty well, so we conclude that the decision tree model is a decent model for our research question.

Random Forest

Assumptions: The individual trees' predictions can be combined to produce a more robust prediction.

Results

Training set RMSE is 1.007, and the test set metrics are MSE is 3.422, RMSE is 1.850, MAE is 0.983, and R2 is 0.969. As we can see from the metrics, and compared to Decision Tree Regressor Result, the training set RMSE for random forest is worse than the training set RMSE for decision tree, since random forest is reducing some overfitting prone from decision tree. The testing set RMSE for random forest is better than testing RMSE for decision tree, so that random forest performs better to unseen data. The R2 value of the testing set is 0.969, which means that the model explains about 96.9% of the variance in hospitalized rate. The MAE indicates that the model's predictions deviate from the actual values by approximately 0.983 units on average, which is generally low and slightly higher than decision tree's MAE.



Lastly, as we can see from the above graph, the result for random forest performs well since it joins the y=x line pretty well, so we conclude that random forest model is a decent model for our research question.

Discussion

Result Summary: Below is the prediction summary table for 4 different models we used in this section. 5 results is sufficient for comparing the performance, and this result represents in log scale for the hospitalized rate.

	pred_frequentist	pred_bayesian	pred_random_forest	pred_decision_tree	observed
0	1.943022	1.956857	1.704268	1.837370	1.678964
1	2.217877	2.233860	2.382519	2.214846	2.187174
2	2.660890	2.667031	2.296297	2.290926	2.271094
3	2.501682	2.497308	2.519908	2.653242	2.533697
4	3.070967	3.027670	3.446118	3.572346	3.575151

Clearly, the prediction from two non-parametric methods achieves quite excellent results as it is generally close to the observed one. Also, we apply RMSE for the test data to measure the performance for each model, and here RMSE is computed by the original hospitalized rate instead of the ones in the log scale.

	Model	rmse
0	Frequentist	6.326717
1	Bayesian	6.313851
2	Random Forest	1.925693
3	Decision Tree	2.265319

RMSE is good to measure how the model would perform when meeting the new data upcoming. From the table above, we can see the error would be about 2% using the non-parametric methods and the error would be about 6% using the Frequentist GLM and Bayesian GLM.

Above all, we can see non-parametric methods have a high confidence level in predicting the hospitalized rate. Though the Frequentist model and Bayesian model have a significantly higher error, 6% is still a reasonable error for predicting the result.

Difference between Frequentist GLM and Bayesian GLM: There are some differences between these two GLM implementations, including the assumptions of the model, the representation of the estimations of the parameters, and uncertainty quantification.

The assumptions of the frequentist GLM are:

- The linearity between the features and the log-transformed hospitalized rate.
- The log-transformed hospitalized rate is normally distributed.

The estimation of the Frequentist model is quite straightforward to catch up, since each estimation for each parameter is just a single estimation.
Here, the uncertainty of the estimation is from the randomness of data. To measure the uncertainty of the parameter estimation, the model summary provides the 95% confidence interval.
For the Bayesian GLM, the assumptions are:

- The linearity between the features and the log-transformed hospitalized rate.
- The likelihood distribution for the the log-transformed hospitalized rate
- The prior distributions for each parameter of interest.

The estimation of the Bayesian model is not a single estimation. Instead, from Pymc3, it provides the predictive posterior distribution for each parameter and each prediction. Therefore, we choose the mean of distribution to be our estimations/predictions.
The uncertainty for Bayesian estimation is from the sampling from posterior distribution and from the predictive posterior distribution. To measure the uncertainty, we use the credible interval.
Interpretation: For the Frequentist GLM model, let's take the estimated parameter for gps_retail_and recreation as an example. The estimation is about 1.98, meaning that with one unit increased in gps_retail_and recreation, the log hospitalized rate would increase by 7.24% (7.24 is the exp(1.98)). For the Bayesian GLM model, also take the mean of posterior distribution for gps_retail_and recreation as an example. The mean is 1.9, meaning that with one unit increased in gps_retail_and recreation, the log hospitalized rate would increase by 6.68% (6.68 is the exp(1.9)). For the Decision Tree method, the interpretation is the graphical tree, and we can not interpret the Random forest result.
Limitation:

- For all the models we constructed, we believe that the features are not sufficient to capture, therefore the estimations are still quite biased.
- Also, the assumptions of our GLM models are not really sufficiently held. The linearity is not really held since the correlation between the features and response variable is not significant.
- In the EDA part, the aggregate process may erase some influential data points.
- We are not really confident about the data we explored, since there might be significant biases in the recording of these data

Potentially Useful Data: The data we have in this section just include the CA, NY, TX states. Since this choice may cause some bias in estimations, we think including the data in all states would be helpful to improve the performance.
Uncertainty Realization: The uncertainty of estimation for the GLM model is a bit high. The uncertainty is coming from the choice of the prior distribution for Bayesian GLM and from the sample data for the frequentist GLM. Also, we believe that the uncertainty in these two models is partly caused by the measurement error.

Q2: Multiple hypothesis testing / decision making

Method

In this part of the project, we focus on exploring our second research question: is there a significant association between COVID-19 cases and airline traffic? To find out, we have compiled a data frame with monthly information on hospitalization rates and airline traffic for all of the 51 states of the USA. Below are the first few rows of our data:

	state	month_after_2020	hospitalized_rate	airline
0	AL	4	10.1	3010000.0
1	AL	5	12.9	8050000.0
2	AL	6	17.7	16530000.0
3	AL	7	43.0	24060000.0
4	AL	8	43.0	25830000.0

We now set up our hypotheses:

Null hypothesis: There is no significant association between hospitalization rates and airline traffic.
Alternative hypothesis: There is a significant association between hospitalization rates and airline traffic.

We chose to perform a hypothesis test using the above hypotheses on each of the 51 states in our data frame, all with a 5% significance level. Testing multiple hypotheses can provide a more comprehensive understanding of the data, increase the robustness of our conclusions, control for confounding variables, and potentially uncover new insights.

Our method of testing will be the Pearson correlation coefficient and p-value for testing non-correlation. The Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables, ranging from -1 to 1, with 0 being no correlation at all. We chose to use this method because, for each correlation coefficient, we can also obtain its associated p-value that can tell us the probability of observing such a correlation coefficient, or a more extreme coefficient, under the assumption that the two variables are actually not correlated. This method of choice directly aligns with our null and alternative hypotheses, and we can also quantify the strength of the relationship between hospitalization rate and airline traffic and evaluate whether such a relationship is due to chance.

At first, we took a shot at this by throwing all of our data from the hospitalization column and airline column into our Pearson correlation test and achieved a p-value of nearly zero. Surprising or not, while this is more than sufficient to reject the null hypothesis and be in favor of the alternative, we thought that it wasn't very meaningful to perform such a test consisting of all of the values in our data frame and that there could be plenty of confounding factors yet to be taken into account.

We then approached the testing process as proposed, performing a hypothesis test on each of the 51 states. Since we have data starting from the fourth month after 2020 up until the 31st month after 2020, there will be 28 pairs of data in the Pearson correlation testing process, which we thought would be a sufficient amount of data per test.

Result

Our results demonstrate that 11 out of the 51 states can achieve a p-value below our 5% threshold of statistical significance, in which case we would reject the null hypothesis. These states are Arizona, Connecticut, the District of Columbia, Illinois, Louisiana, Massachusetts, Maryland, Maine, New Jersey, New York, and Rhode Island. Specifically, we can see that Connecticut, Massachusetts, and Rhode Island have returned p-values below the 1% threshold in terms of the correlation between hospitalization rates and airline traffic. Additionally, it was quite interesting to note that all of these 11 states are on or close to the east coast of the United States besides the state of Arizona. Below are the detailed p-values we obtained from these 11 states:

```
11

state
AZ    0.034052
CT    0.002183
DC    0.026175
IL    0.048737
LA    0.027373
MA    0.000759
MD    0.019537
ME    0.045679
NJ    0.032705
NY    0.011844
RI    0.000818
```

Of course, on the other hand, we were not able to reject the null hypothesis for the rest of the 40 states. Interestingly, most of the p-values for these 40 states were far above our 5% threshold. In fact, 8 of these states had p-values close to 1, suggesting that there could not be a correlation between our two variables whatsoever. Considering that we only obtained 11 states with statistical significance, this figure seems rather surprising. These discoveries really raise the question of whether there is a correlation between hospitalization rates and airline traffic overall, and plant a red flag on our initial p-value of nearly zero when we performed a single test on all of our data.

Controlling for Family-Wise Error Rate (FWER) and False Discovery Rate (FDR)

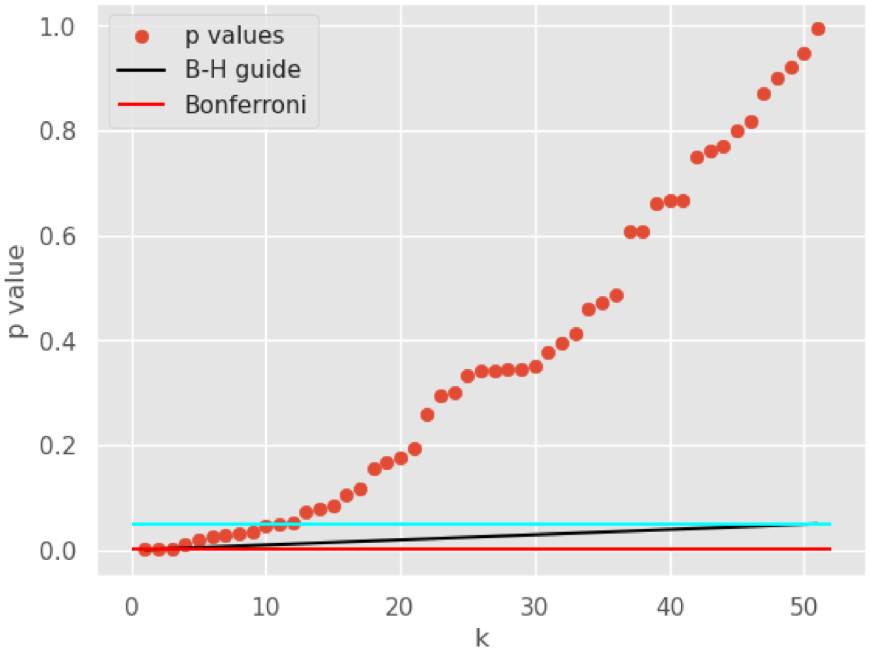
We will now be applying two correction methods on top of our discoveries made at the 5% significance level, known as the Bonferroni correction and the Benjamini-Hochberg procedure. Both methods are used to control the false positive rate in multiple hypothesis testing.

The Bonferroni correction method controls for the Family-Wise Error Rate, which is the probability of obtaining at least one false positive test. It assigns a new significance level to each of our individual tests with a threshold of our original threshold divided by the number of tests performed. In our case, it would be the original 5% threshold divided by a total number of 51 tests, so we would end up with a new threshold of less than 2% of the original one. This method is very conservative because it strictly controls the probability of having a single false positive test to be within the overall significance level.

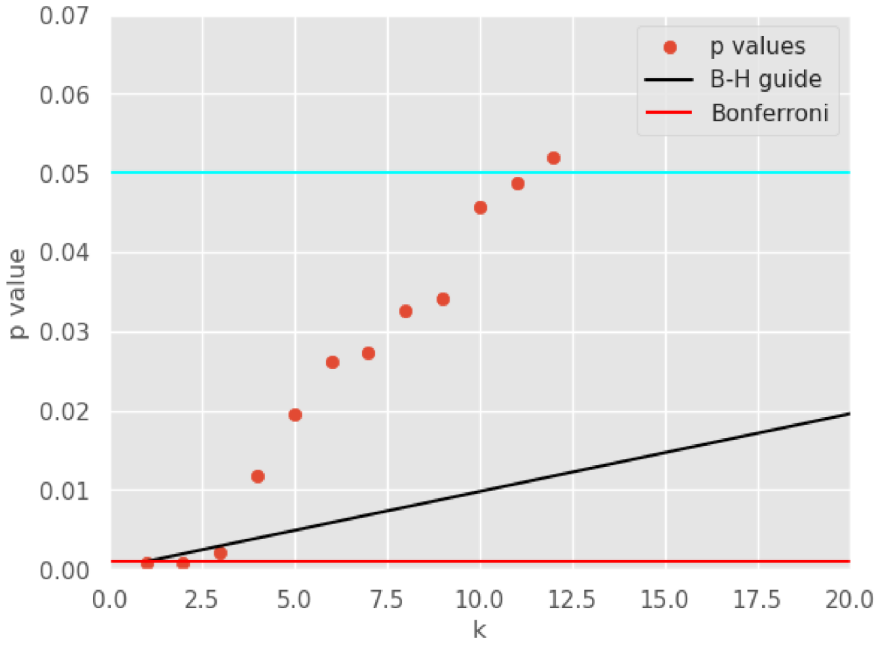
The Benjamini-Hochberg procedure, however, is a less conservative method that controls for the False Discovery Rate, which is the overall expected proportion of false positive tests. The less conservative part is that as we perform more and more tests on our data, our tolerance of a false positive test will increase, as opposed to strictly limiting the number of false positives in the Bonferroni. This procedure follows a threshold that increases each time we have made a discovery, allowing it to make more discoveries that the Bonferroni method would never make.

There were 11 null hypothesis rejections found in the p-values in our previous testing and after applying the Bonferroni and Benjamini-Hochberg correction procedures, controlling for FWER made only 2 discoveries and controlling for FDR made 3 discoveries. The Bonferroni correction tends to be more conservative, resulting in fewer discoveries, whereas FDR control methods like the Benjamini-Hochberg procedure are generally more powerful, often leading to a higher number of discoveries. In our case, both methods found much fewer discoveries among the p-values we initially found.

Below is a visualization of how the aforementioned procedures work in our case of multiple hypothesis testing. The p-values here are ranked in order from smallest to largest, and we have a cyan-colored horizontal line indicating our original threshold of 5%. The black and red lines represent the Benjamini-Hochberg threshold and the Bonferroni threshold, respectively, and we can see that both are below the original threshold. Also notice that as we “see” more p-values coming in, the BH guide is more tolerant with false positives, and by the very end, we would be as tolerant as our original threshold is in terms of filtering out false positive discoveries.



Now, let's take a closer look at which p-values are really still below the thresholds after the correction methods have been applied:



As we can see from above, the Bonferroni threshold lies very low, only making two discoveries: Massachusetts and Rhode Island. The BH guide makes a third discovery, Connecticut, which is our third

lowest p-value from our original list above. These three states after the correction are all on the east coast of the United States and in the New England region.

Discussion

Overall, we have come to the conclusion that from testing with our original threshold of 5%, the 11 rejections of the null hypothesis are statistically significant and that within those states, hospitalization rate and airline traffic are indeed correlated. However, when looking at our results overall, we would conclude that the correlation between these two variables is not as strong as we would hope for, as we had 40 states without significant evidence to reject our null hypothesis. However, we did discover that our initial single test of the whole data frame (returning a p-value of basically zero) should not be meaningful and had extreme bias based on our further testing on every individual state and realized that in fact, only about 20% of the states showed significance on the figures.

We do believe that there are certain limitations in the dataset we have chosen. When compiling the data together, we selected the highest hospitalization rate in each month in each state because we wanted to focus on its relationship with airline traffic at the worst times of COVID-19 each month. Doing so may have led to some inaccuracies when the Pearson correlation test establishes the correlation coefficient, as the results could be different if we had a different approach to filtering the dataset. Furthermore, there could still be confounding factors and other biases in this dataset, and we could include more data from different perspectives and attempt to reduce these factors in future testing. At the end of the day, based on the dataset we put together, we were hoping to discover a correlation between our variables because we expect there to be one, but the results did not seem like we were able to do so.

Future testing can include datasets on other countries besides the United States, as some may not utilize planes as the primary source of travel and we are trying to model how people's travel preferences and plans have changed in times such as COVID-19. Our current dataset could also include more detailed categories and classifications so that we can differentiate between which factors play more important roles in people's travel plans during the worst times of the pandemic. Predicting, testing, and modeling these figures can help all of us better understand how people's life would change according to such unpredictable situations and would potentially make us better prepared for another similar catastrophe in the future.

Conclusion

Key Findings

For research question 1, in the topic to predict the hospitalized rate by mobility indexes in the pandemic period, the prediction from two non-parametric methods achieves quite excellent results as it is generally close to the observed one. Though the Frequentist model and Bayesian model have a significantly higher error, 6% is still a reasonable error for predicting the result.
For research question 2, in the topic to explore the association between hospitalized rate and the airline traffic in the pandemic period, we have come to the conclusion that from testing with our original threshold of 5%, the 11 rejections of the null hypothesis are statistically significant and that within those states, hospitalization rate and airline traffic are indeed correlated. However, when looking at our results overall, we would conclude that the correlation between these two variables is not as strong as we would hope for, as we had 40 states without significant evidence to reject our null hypothesis.

Result Realization

For research question 1, we restricted our data in state CA, NY, TX. It is large enough for the prediction but if we want to predict the hospitalized rate around the whole country, we would need to generalize our data to all states.
For research question 2, since we construct 51 tests for 51 states, the data involve all the states in the U.S, we would say the testing is general enough.

Limitation

The main limitation is that we are not really confident about the data we explored, since there might be significant biases in the recording of these data.

Future Exploration

We believe that mobility is only a part of the equation when it comes to influencing changes in hospitalization rates. We can introduce more factors, such as employment data, unemployment rate and retail sales profit, to help us get a better prediction. Also, for exploring association, we may introduce other measurements for airline traffic.