Data 102 Final project

Chengming Liu(lcm_19berkeley.edu), Cheryl Liu, David Hsieh, Newton Szeto

May 8, 2022

Contents

1 Data Overview

- How were your data generated? Is it a sample or census?
  - For the data that we used in research question 1, the Highway Vehicle Miles Traveled is a monthly statistic that is sampled and then estimated by the Federal Highway Administration.
- If you chose to use your own data, describe the data source and download process.
  - N/A
- If you chose to add additional data sources, explain why.
  - For our first research question, we included an additional data source regarding the number of COVID-19 cases in the United States. We felt that this was necessary as our first research question was measuring the impacts that COVID-19 had on travel, and the data set would provide us with the sufficient data required to create our model based on the number of new cases per month. For our second research question, we didn't include additional data sources.
- If your data represents a sample:
- If your data represents a census:
- To what extent were participants aware of the collection/use of this data?
- What is the granularity of your data? What does each row represent? How will that impact the interpretation of your findings?
- Are any of the following concerns relevant in the context of your data?
  - Selection bias
  - Measurement error
  - Convenience sampling
- Are there important features/columns that you wish you had, but are unavailable? What are they and what questions would they help you answer?

2 Research questions

2.1 Question 1

- What is the research question? What real-world decision(s) could be made by answering it?
  - Our first research question is: Has the COVID pandemic caused an increase in the monthly number of highway vehicle miles traveled within the U.S.? Based on our results, we believe that this could determine the impacts that COVID-19 had on travel, which would assist those in the tourism industry in calculating how much of their business was lost or gained because of the pandemic compared to other potential factors.
- Explain why the method you will use is a good fit for the question
  - For this question, we will be using causal inference as the main method of data analysis. Specifically within casual inference, we will be focusing on matching, where we match our two confounders of winter months and school months, differentiating it by whether or not a treatment was applied. Through matching, we will be able to get the average difference between the treatment and control variables. This will reduce the effects of the confounding variables. By using matching to eliminate the confounders, we are able to minimize the external factors and focus purely on the impact of the treatment on the outcome.

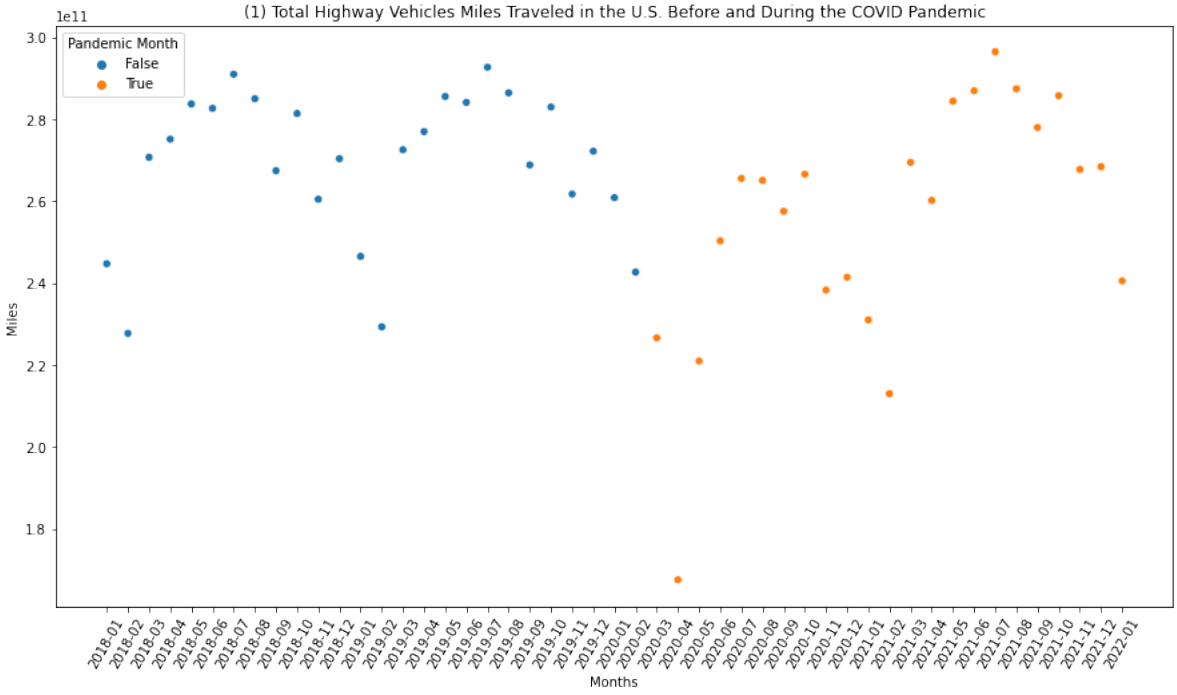
2.2 Question 2

- What is the research question? What real-world decision(s) could be made by answering it?
  - We are predicting air carrier fatalities from transit ridership, personal and government construction spending, employment in the industry. This can shed light into what are the contributing factors in prediction of fatalities and what future measures can be suggested to cope with increasing in fatalities.
- Explain why the method you will use is a good fit for the question
  -

3 EDA

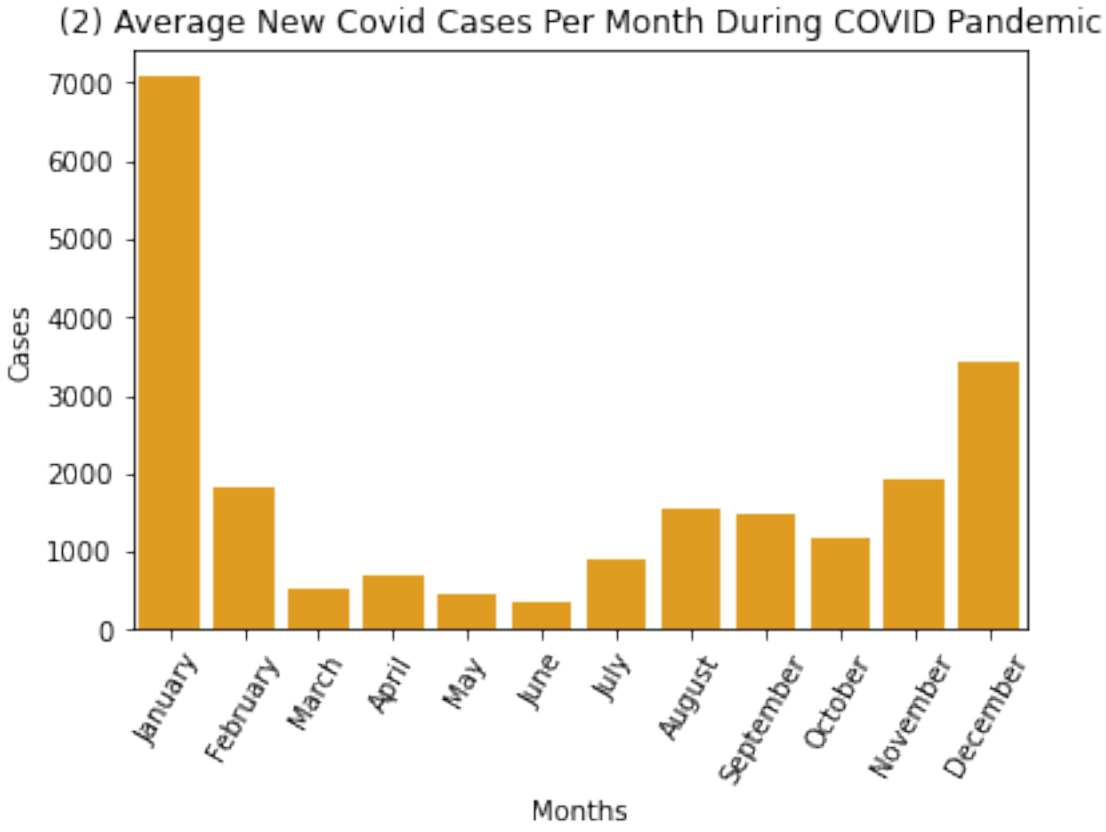
3.1 Question 1

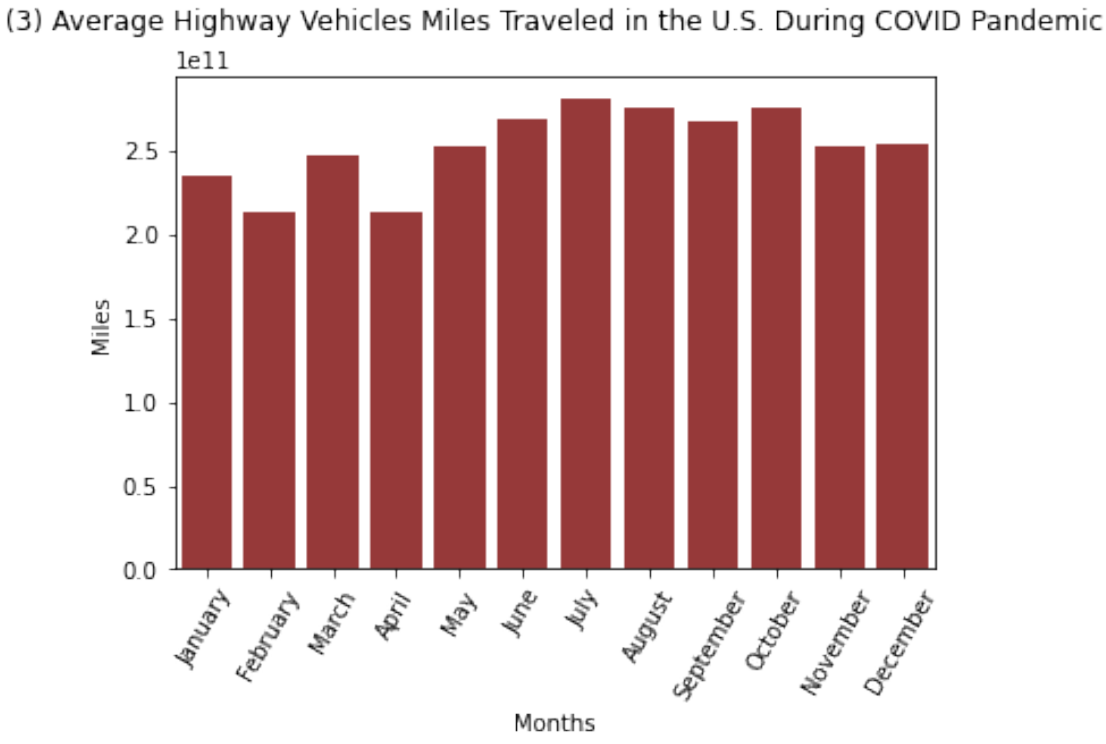
- Visualize at least two quantitative variables and two categorical variables. Your visualizations must be relevant to your research questions!
- Describe any trends you observe, and any relationships you may want to follow up on.



-

Plot (1): The plot shows a clear pattern in total highway miles driven over time for data recorded before the pandemic was officially announced. In particular, the total highway miles driven consistently peaked during the summer months and reached the lowest point during the winter months. However, after March 2020 when the pandemic was officially announced, the total highway miles driven decreased significantly enough in April 2020 to make it an outlier. The peak during the summer also was not as high as in previous years, until 2021. The nonconforming trend in the plot in 2020 is an indication that the COVID pandemic might have caused a change in normal driving behavior on highways, which is what we will be investigating in our causal inference.





Plots (2), (3): So far, we recognize that there appears to be an inverse trend between the average number of COVID cases per month and the number of vehicle miles traveled per month. This is a promising start for us as we explore the impact COVID had on vehicle miles traveled, however, we have to be wary of confounders such as less travel during winter months. Looking at the scatterplot, we can see that while the parabolas are generally shaped the same and at the same scale, the pandemic months show that the parabola has shifted downwards, which could indicate the pandemic's effect on vehicle miles traveled. This can also help account for our confounding variables, like the lack of travel during winter months due to the weather, because proportionally the upside down parabolas have the same scale, but are just all shifted equidistantly lower, from what we can infer is the impact of COVID-19.

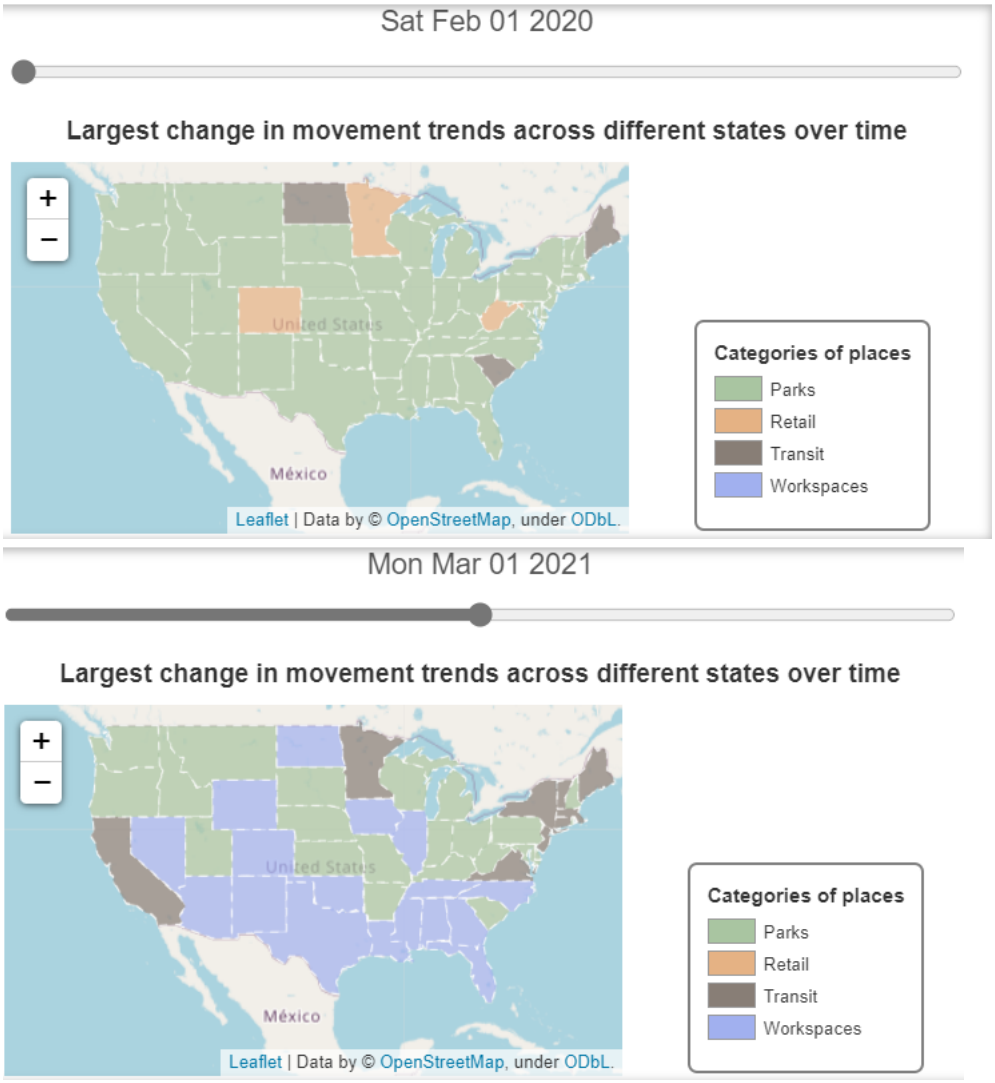
- Describe any data cleaning steps you took. How will these decisions impact your models and inferences?
  - For plot (1), we performed some data cleaning and manipulation steps. First, focusing on the transportation dataset, we felt that it was best to filter out the dates before January 2018 and after January 2022, as there is no data outside of these dates for the column we are looking for: "Highway Vehicle Miles Traveled - All Systems." We also had to filter the columns to only show the "Highway Vehicle Miles Traveled - All Systems" column. We then removed the day of the dates and added a column named "Pandemic Month" to indicate whether the month occurred during the pandemic. The data was already collected on a monthly basis, which seems optimal for identifying previous trends in highway miles driven and how they changed during the pandemic. Thus, it is best if our units for our causal inference is time in months.
  - For plot (2), we also had to clean the dataset for our imported COVID dataset. First, we started by selecting the columns we wanted, which were the date and the new cases. After converting the dates into datetime objects, we filtered out data before the beginning of the pandemic, March 2020. We then removed the years and days from the dates and grouped the data by month, aggregating using the mean, giving us the mean number of cases per month of the pandemic. Grouping based on month revealed that winter months were associated with higher case counts on average during the pandemic than the other months. This indicated that winter and school year months could be confounding variables.
  - we further manipulated the transportation dataset by filtering out months before the pandemic, grouping by month, and aggregating using the mean, giving us the average total highway vehicle miles driven per month of the pandemic. From this we confirmed the confounding variable's effect on the outcome.
- Explain how your visualizations should be relevant to your research questions: either by motivating the question, or suggesting a potential answer. You must explain why they are relevant.
  - Plot (1) served as the main motivator for our analysis of the pandemic possibly causing a change in total highway vehicles miles driven in the U.S. Since the scatter plot shows a very clear change in the previous repeating pattern of total highway miles driven over time after the pandemic began, this encourages us to perform causal inference to see whether the pandemic actually caused this change to happen.
  - Plots (2), (3): For the graph "Average New Covid Cases Per Month During COVID Pandemic," we visualized the average number of COVID cases per month for the months of the pandemic using our COVID data. We wanted to check this to show evidence of confounders affecting the outcome. Ultimately, we saw that during the winter months, the average number of new COVID cases per month was higher relative to the summer months. This indicates to us that a month occurring during winter or during the regular school year may be causing the change in new covid cases, which would confirm the confounders' affect on the treatment.

    With our graph "Average Highway Vehicles Miles Traveled in the U.S. During COVID Pandemic," we are seeking to measure how highway vehicle miles changed throughout the year during the COVID pandemic on a monthly basis. We see that the number of highway miles driven on average peaks in July and then slowly decreases over time. This implies that the confounder may be causing the outcome as well. However, there could be other confounders to this such as better weather in most states and summer break for those in school.

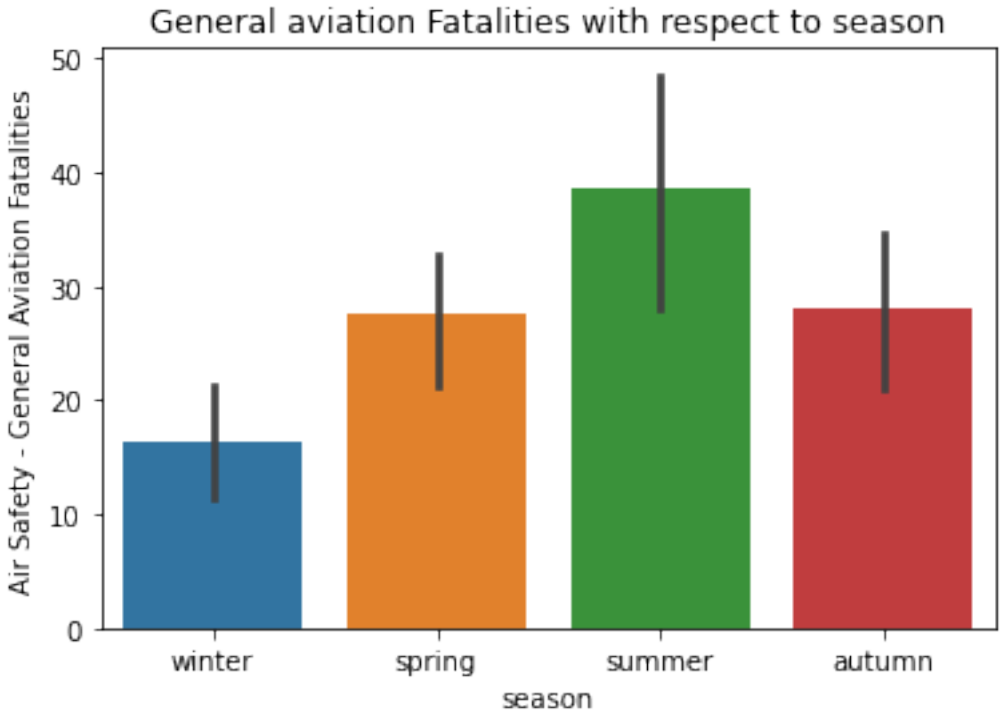
    Since the status of a month occurring during winter or during the regular school year is likely a confounding variable based on these plots, it must be accounted for in our causal inference analysis.

3.2 Question 2

- Visualize at least two quantitative variables and two categorical variables. Your visualizations must be relevant to your research questions!



The "% change with respect to state and time" plot shows the change categorical variable "Categories of places" with time from Feb 2021 to April 2022. This visualization would help us understand which category of places experiencing the most drastic % change(in absolute value) for each state in US by month. "Categories of places" could be used as a feature in the prediction model of air carrier fatalities.



–

By visualizing the value we want to predict with respect to season, a feature that contains more information than the raw dataset is provided, we are hoping to discover whether season is a meaningful feature engineering step

- Describe any trends you observe, and any relationships you may want to follow up on.
  - Geoplot(largest change in movement trends across different states over time): We notice that the residential and grocery_and_pharmacy percent change from baseline are not experiencing drastic changes. It's reasonable because people always need residential areas, grocery and pharmacy to live, no matter where they are. For January 2021, the % change of transit in most of the state is the greatest in absolute value.
  - Barplot(season vs fatility): From the graph, we can tell that fatalities in aviation have a higher value in summer comparing to all other three seasons, while it is the lowest during the winter. Although the value being lowest should be related to the time COVID-19 broke out, which was prior to 2020 spring.
- Describe any data cleaning steps you took. How will these decisions impact your models and inferences?
  - For the google region reports data, we first identify the nan values, which is mainly the data for North Dakota July 2020, July 2021, and August 2021. We remove the nan values by filling the nan values of August 2021 by the data of August 2020. Also, we use the average of June and August data to fill the July data.
  - For the DOT transportation data, we limit the dataset to the same span of time the google region reports data provides, and fill the nan value with an average value of the data 3 months prior to the nan value entry. The filling procedure may heavily impact data from 2022, since there are a lot of columns being nan, and the intrinsic value of these columns are often decided by policies, domestic or international events, which means they often cannot be accurately portrayed by an average.
- Explain how your visualizations should be relevant to your research questions: either by motivating the question, or suggesting a potential answer. You must explain why they are relevant.
  - Since the prediction model requires the transit ridership as a feature, we would like to investigate in the movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. The different movement trends may affect the fatality in the air travel.
  - From the dot dataset, we are only provided with so much information. Extracting seasonality out of date is to suggest the possibility of externalities with respect to these given data: Is there a trend of maintenance employment or working pressure that cannot be reflected directly by our given dataset, but might be partially captured by our extracted features?

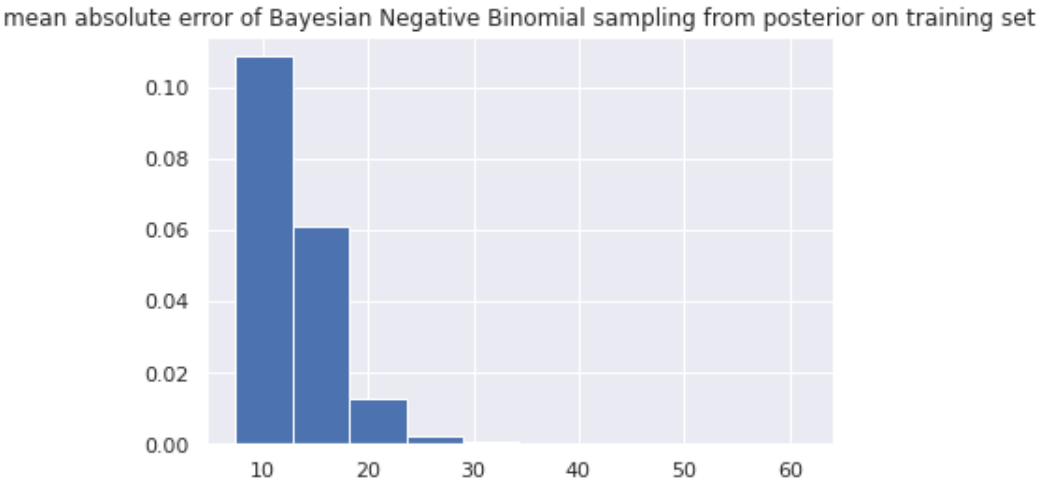
4 Inference and Decisions

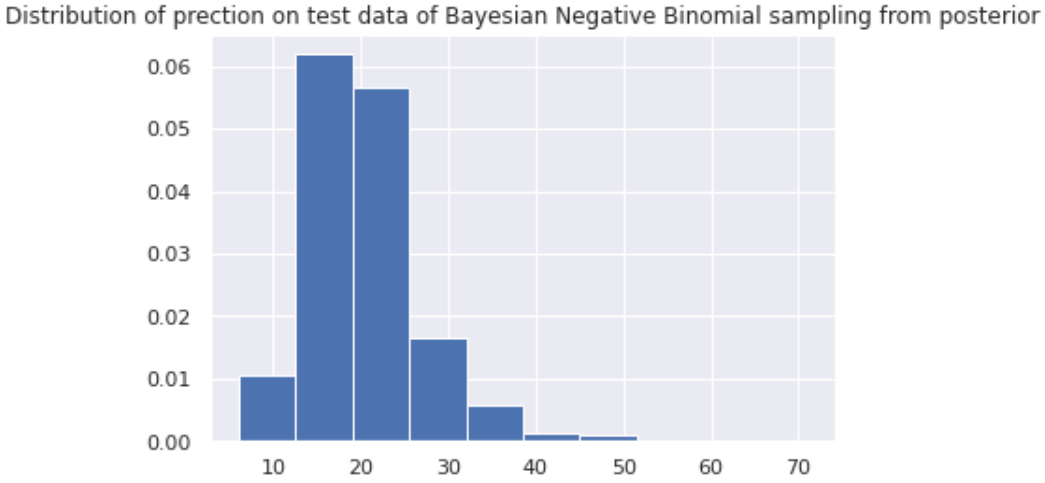
4.1 Question 1

- Methods
  - Describe which variables correspond to treatment and outcome
    - The treatment variable is binary, with 1 representing that a month has above the average number of monthly new COVID cases during the pandemic, and 0 otherwise. The pandemic is defined as beginning in March 2020 and lasting up to the present. The outcome is the total number of highway vehicle miles traveled in a month.
  - Describe which variables (if any) are confounders. If the unconfoundedness assumption holds, make a convincing argument for why.
    - We identified two major confounding variables. The first confounder would be whether or not the month we are examining is a winter month. Winter months were described as months from November to February. We felt that this was a confounder because during the winter months, people are less inclined to drive long distances due to the turbulent weather. For the second confounding variable, we identified it as school months, which was defined as the months from September to May. We recognized this as a confounding variable because during the school year, people have less time to travel long distances given the time constraints of assignments, projects, exams, and other involvement.
  - What methods will you use to adjust for confounders?
  - Are there any colliders in the dataset? If so, what are they?
- Results
  - Summarize and interpret your results, providing a clear statement about causality (or a lack thereof) including any assumptions necessary.
    - For our analysis, we filtered out data for months before March 2020, the start of the pandemic, to avoid possible confounders such as the official announcement of the COVID pandemic. Our treated units were months that had an above average number of new COVID cases, while untreated units were months that had below the average number of new COVID cases. In order to test for causality, we used matching causal inference, which was based on two binary confounders: whether a month occurred during winter and whether it occurred during the school year. We were able to calculate the average treatment effect, or ATE, from our model, which resulted in an output of -104,604,914.9 miles. This showed us that for months during the COVID-19 pandemic that had a higher number of cases relative to the average, people in the United States traveled less on highway systems. We mainly made two assumptions in calculating this number. The first one was that people travel less in the winter months given the weather. The second assumption was that people travel less during school months as they do not have as much free time.
  - Where possible, discuss the uncertainty in your estimate and/or the evidence against the hypotheses you are investigating.
    - One potential uncertainty in our causal inference analysis is that there might not be enough data to make a definitive statement about the average treatment effect of the pandemic on the total number of highway vehicle miles traveled per month. This is mainly due to the fact that we filtered out data for months before March 2020, the start of the COVID pandemic. Another possible uncertainty is that it is possible that there may be additional confounders that may be affecting the result, which would not completely satisfy the unconfoundedness assumption.
- Discussion
  - Elaborate on the limitations of your methods.
    - One limitation of our matching causal inference analysis is that there might not be enough data points to accurately predict the ATE. After performing data cleaning, we have 23 units (months), seven of which are in the treatment group and the rest in the untreated group. Additionally, the treatment variable, whether or not the number of new COVID cases of a month was above the monthly average number of new COVID cases, was chosen somewhat arbitrarily. Although we would likely reach the same conclusion, a different calculation of the treatment variable would potentially lead to a different ATE value. Lastly, choosing a different granularity for our units, such as weeks, could also reveal a different pattern in highway vehicle miles driven for the treated and untreated groups.

- What additional data would be useful for answering this causal question, and why?
  - Additional data on employee paid time off (PTO) in the U.S. during COVID pandemic months might be useful in identifying this as a potential confounder. Months during the pandemic with more employees on PTO or with greater total days spent on PTO can potentially cause an increase or decrease in new COVID cases per month as well as total highway vehicle miles driven.
- How confident are you that there's a causal relationship between your chosen treatment and outcome? Why?
  - Based on the magnitude of total highway vehicle miles driven per month in our cleaned data compared to that of our calculated ATE, we are not very confident that there is a causal relationship between months having above the average number of new COVID cases and total highway vehicle miles driven. The magnitude of total highway vehicle miles driven for all the months in our cleaned dataset is on the order of 100 billion miles, while the magnitude of our ATE from this dataset is only on the order of 100 million miles.

4.2 Question 2

- Methods
  - Describe what you're trying to predict, and what features you're using. Justify your choices.
    - We are trying to predict general aviation fatalities. We are using U.S. Airline Traffic, Government Construction Spending on aviation related field, unemployment status for air transportation, unemployment and labor participation rate in the U.S., U.S. Air Carrier Cargo amount, international and domestic, and U.S. marketing air carriers on-time performances.
  - Describe the GLM you'll be using, justifying your choice. Describe any assumptions being made by your modeling choice.
    - We are using negative binomial regression, since the value we are trying to predict is count data: fatalities are in integer and cannot be negative. We are assuming linearity in model parameters, independence of individual observations, and the multiplicative effects of independent variables.
  - Describe the nonparametric method(s) you'll be using, justifying your choice. Describe any assumptions being made by your modeling choice.
    - We are using XGBoost's implementation of decision trees. This is one of the most popular implementation of decision trees. Since this is a non-parametric approach, there are no underlying assumptions about the distribution of the errors or the data.
  - How will you evaluate each model's performance?
    - We decided to use mean absolute error to evaluate each model's performance.
- Results
  - Summarize and interpret the results from your models
    - 



For GLM in Bayesian implementation, by plotting the distribution of the mean absolution error calculated between predicting over the training set by sampling from posterior and the ground truth, we can observe that majority of the errors we obtain is ranging from 7.5 to roughly 20. By plotting the distribution of the prediction on the test set by sampling from posterior, we can obtain a 95% probability credible interval that our prediction will fall between 11 and 36.

```
mean_absolute_error(train.iloc[:, 0], np.ceil(negbin_results.
↪predict(train[negbin_model.exog_names].values, linear=False)))
```

```
7.916666666666667
```

```
np.ceil(negbin_results.predict(test[negbin_model.exog_names].values[0],␣
↪linear=False))
```

```
array([19.])
```

* For GLM in Frequentist implementation, we can see that by the standard of log-likelood, our model is not a good fit. By Chi-square test, our model is quite okay in fitting the data. Our negative binomial model has a mean absolute error of 7.916666666666667 over the training set, and predict that on the test set, the value of fatalities will be 19.

```
mean_absolute_error(train_y, np.ceil(xgb_best.predict(train_X)))
```

```
2.3333333333333335
```

```
mean_absolute_error(val_y, np.ceil(xgb_best.predict(val_X)))
```

```
6.5
```

```
np.ceil(xgb_best.predict(test_X))
```

```
array([26.], dtype=float32)
```

* For non-parametric method, we choose XGBoost's implementation of decision tree. We split the first 24 months of data into training set, the following 2 months of data into validation set, and the final month as test set. On training set, we achieved a mean absolute error of 2.3333333333333335, and on validation set, we achieved a mean absolute error of 6.5. On test, we predict the value of fatalities will be 26.

– Estimate any uncertainty in your GLM predictions, providing clear quantitative statements of the uncertainty in plain English.

* For GLM in Bayesian implementation, we can see from the posterior predictive checks that for any of the features that we used, none of them are very confident in terms of being a variable over the training set.

Generalized Linear Model Regression Results

Dep. Variable:	Air Safety - General Aviation Fatalities	No. Observations:	26
Model:	GLM	Df Residuals:	13
Model Family:	NegativeBinomial	Df Model:	12
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-110.66
Date:	Sun, 08 May 2022	Deviance:	3.7264
Time:	15:58:16	Pearson chi2:	3.46
No. Iterations:	9	Pseudo R-squ. (CS):	0.08950
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
U.S. Airline Traffic - Total - Seasonally Adjusted	1.584e-08	3.34e-08	0.473	0.636	-4.97e-08	8.14e-08
State and Local Government Construction Spending - Runway	1.39e-09	3.68e-08	0.038	0.970	-7.08e-08	7.35e-08
State and Local Government Construction Spending - Air Passenger Terminal	2.749e-09	3.66e-08	0.075	0.940	-6.89e-08	7.44e-08
State and Local Government Construction Spending - Air	-3.169e-12	3.61e-08	-8.78e-05	1.000	-7.07e-08	7.07e-08
State and Local Government Construction Spending - Other Public Safety	7.891e-09	4.71e-08	0.167	0.867	-8.45e-08	1e-07
State and Local Government Construction Spending - Public Safety	-7.094e-09	2.69e-08	-0.264	0.792	-5.98e-08	4.57e-08
State and Local Government Construction Spending - Infrastructure	-1.521e-09	2.11e-08	-0.072	0.943	-4.29e-08	3.99e-08
Transportation Employment - Air Transportation	-6.27e-06	2.27e-05	-0.277	0.782	-5.07e-05	3.81e-05
Unemployment Rate - Seasonally Adjusted	2.4368	21.978	0.111	0.912	-40.639	45.513
Labor Force Participation Rate - Seasonally Adjusted	9.1882	25.158	0.365	0.715	-40.120	58.496
U.S. Air Carrier Cargo (millions of revenue ton-miles) - International	-2.615e-10	2.03e-09	-0.129	0.897	-4.24e-09	3.72e-09
U.S. Air Carrier Cargo (millions of revenue ton-miles) - Domestic	-1.023e-11	3.22e-09	-0.003	0.997	-6.33e-09	6.31e-09
U.S. marketing air carriers on-time performance (percent)	-0.3973	5.411	-0.073	0.941	-11.002	10.207

* =

```
Bootstrap std error for U.S. Airline Traffic - Total - Seasonally Adjusted: 0.000
Bootstrap std error for State and Local Government Construction Spending - Runway: 0.000
Bootstrap std error for State and Local Government Construction Spending - Air Passenger Terminal: 0.000
Bootstrap std error for State and Local Government Construction Spending - Air: 0.000
Bootstrap std error for State and Local Government Construction Spending - Other Public Safety: 0.000
Bootstrap std error for State and Local Government Construction Spending - Public Safety: 0.000
Bootstrap std error for State and Local Government Construction Spending - Infrastructure: 0.000
Bootstrap std error for Transportation Employment - Air Transportation: 0.000
Bootstrap std error for Unemployment Rate - Seasonally Adjusted: 117.226
Bootstrap std error for Labor Force Participation Rate - Seasonally Adjusted: 68.961
Bootstrap std error for U.S. Air Carrier Cargo (millions of revenue ton-miles) - International: 0.000
Bootstrap std error for U.S. Air Carrier Cargo (millions of revenue ton-miles) - Domestic: 0.000
Bootstrap std error for U.S. marketing air carriers on-time performance (percent): 16.605
```

For GLM in Frequentist implementation, we confirm the uncertainty of our fit by comparing those with the result of bootstraping, which in nature should be less confident than our fitting

results.

- Discussion
  - Which model performed better, and why? How confident are you in applying this to future datasets?
    - From the perspective of mean absolute error, our non-parametric method - XGBoost's implementation of decision tree - achieved the best performance. However, we are not extremely confident in applying this model to future datasets.
  - Discuss how well each model fits the data.
    - For both GLM implementations, the model didn't fit on the training set particularly well, with Bayesian implementation's error ranging from 7.5 to close to 30, and Frequentist implementation's error being at 7.91. The non-parametric approach fits the data the best, achieving mean absolute error of 2.3 on training set and 6.5 on validation set.
  - Explain any differences you observed between the Bayesian and frequentist implementations of your GLM.
    -
  - Interpret the results from each model. You may choose to not provide interpretations, but you must justify this choice.
    -
  - Elaborate on the limitations of each model.
    - GLM Bayesian implementation:
    - GLM Frequentist implementation:
    - Non-parametric model:
  - What additional data would be useful for improving your models?
    - It would be extremely helpful to obtain dataset that contains information about aircraft's maintenance status, maintenance staff's attendance form (anonymized), and runways' and other aircraft related infrastructure's maintenance status.

5 Conclusion

5.1 Question 1

- Summarize your key findings
  - Through our EDA analysis, we observed that during the COVID pandemic, there was an initial decrease in monthly total highway vehicle miles driven. However, after performing matching causal inference with our defined treatment variable, we calculated a relatively small magnitude negative ATE value.
- How generalizable are your results? How broad or narrow are your findings?
  - Our matching causal inference analysis is narrow in that it just tries to discover a potential causal relationship between new monthly COVID cases and total monthly highway vehicle miles driven. On the other hand, our research question is more broad in that it seeks to find a causal relationship between any aspect of the COVID pandemic and total monthly highway vehicles miles driven. For example, the potential causal effect of news outlets reporting on the pandemic on the outcome variable was not analyzed. However, our analysis provides evidence that the number of new monthly COVID cases in the U.S. did not have a causal effect on the outcome.
- Based on your results, suggest a call to action. What interventions, policies, real-world decisions, or action should be taken in light of your findings?
  - Our causal inference analysis does not identify a problem to solve or a potential solution. It found a lack of a causal relationship between new monthly COVID cases and total monthly highway vehicles miles driven.
- Did you merge different data sources? What were the benefits and/or consequences of combining different sources?

  - Yes, we merged the "Bureau of Transportation Statistics: Monthly Transportation Statistics" dataset with an external "United States COVID-19 Cases and Deaths by State over Time" dataset from the CDC. This was helpful because it allowed us to base our binary treatment variable on new monthly COVID cases. On the other hand, it also required us to create a somewhat arbitrary way of separating the treated and untreated groups.

- What limitations are there in the data that you could not account for in your analysis?
  - Both of the datasets we used mostly provided sufficient data for performing our matching causal inference analysis. However, the CDC dataset likely does not have completely accurate COVID case numbers due to potentially low COVID test rates.

- What future studies could build on your work?
  - Future studies could evaluate a potential causal relationship between new monthly COVID deaths or news outlet pandemic reporting on total monthly highway vehicle miles driven.

5.2 Question 2

- Summarize your key findings
  - In conclusion, the non-parametric method works the best among the GLM Bayesian implementation, the negative binomial GLM Frequentist implementation, and the XGBoost's non-parametric method, because we achieved the lowest mean absolute error of 2.33 on the training set, and of 6.5 on the validation set. On test, we predict the value of fatalities will be 26.

- How generalizable are your results? How broad or narrow are your findings?
  - Our model works okay on the dataset we got, but we believe it is not generalizable to the reality. The findings are based on the data collected, and the implications of our model to the real data could generate different results.

- Based on your results, suggest a call to action. What interventions, policies, real-world decisions, or action should be taken in light of your findings?
  - Based on our findings, since the predicting process is black-boxed, we could not suggest any specific policies or actions regarding to reduce the travel fatalities.

- Did you merge different data sources? What were the benefits and/or consequences of combining different sources?
  - We merged different data sources: "Google: Daily Community Mobility Data" and "Bureau of Transportation Statistics: Monthly Transportation Statistics". The google dataset provides more features for our model.

- What limitations are there in the data that you could not account for in your analysis?
  - We only have access to features that not directly related to the fatality rate. If there exists flight data, accident rates of the plane, and etc., we could improve our model.

- What future studies could build on your work?
  - Future studies related to how pandemic is related to travel safety could build on our work.