Data 102 Final Project, Fall 2022

James Shi, Peng Zheng, Joe Zhou, Chloe Liu

Data Overview

Our data is generated by the census. The data is published by the Bureau of Transportation Statistics. The dataset includes information on airline traffic, transit ridership, transportation employment, construction spending, and transborder. I chose to add an additional data source since the fatalities we have in the original dataset are only the death number reports. However, the plain number cannot represent the fatality rate. There are no groups that were systematically excluded from your data. The participants were totally aware of the data because the data are widely used by researchers. In this way, it is highly possible that they will read articles or watch videos referencing this data. This data has a high level of granularity since it has a large number of individual pieces of information, each row represents the corresponding data of that month. The granularity of data can affect how it is used and analyzed and can impact the accuracy and usefulness of the results. I don't think any selection bias, measurement error, and convenience bias concerns are relevant in the context of our data. We wish that we had features like the specific distribution of air accident reasons, the specific government funding distribution, and the specific construction timeline of government spending. These questions can help our group better understand the effective time of government spending and it can help our model to be more accurate. Furthermore, the lack of consistent non-empty entries within a certain period makes it difficult to model and predict, since each feature is likely to have different availability and thus a different number of usable entries, even if we fixed the timeframe to be more recent.

EDA

For research question 1, we initially looked at the entire dataset and discovered that many columns related to our topic of interest consist of a large proportion of null values, which would be hard to make use of during prediction modeling. In particular, there are multiple columns with more than 90% of null values, and solving this was our first goal in data cleaning. After exploring and deciding to select a subset of the dataset where data is more available between

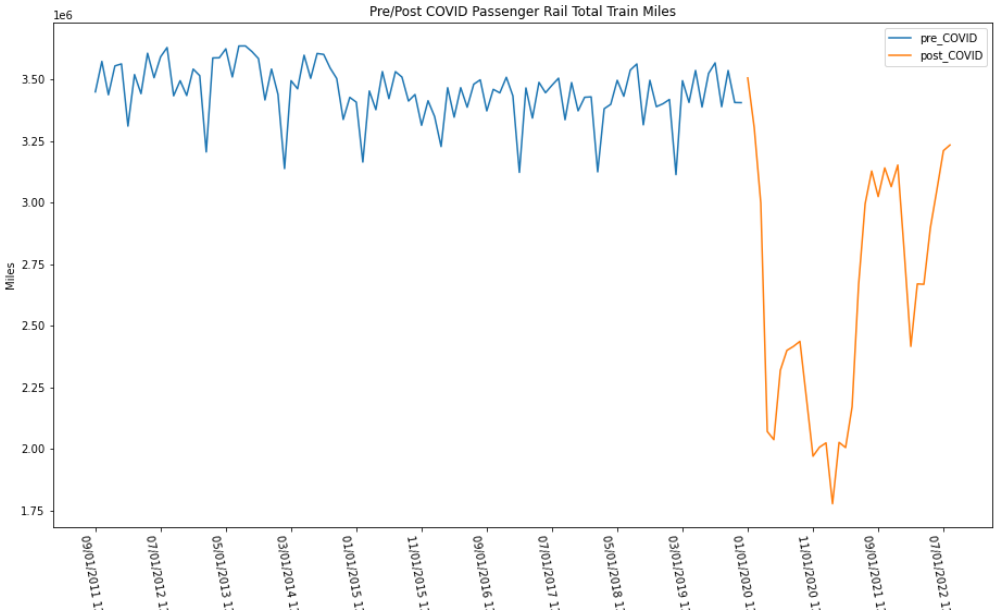
1975 and 2020, we wrote a function that would calculate the precise portion of null values in each column. An example of the function output would be as follows:

	variables	na%
0	Index	0.000000
1	Air Safety - General Aviation Fatalities	0.381579
2	Highway Fatalities Per 100 Million Vehicle Mil…	0.802632
3	National Highway Construction Cost Index (NHCCI)	0.763158
4	Highway Fuel Price - On-highway Diesel	0.618421
5	Highway Fuel Price - Regular Gasoline	0.565789
6	Unemployment Rate - Seasonally Adjusted	0.763158
7	Labor Force Participation Rate - Seasonally Ad…	0.763158
8	Passenger Rail Total Reports	0.368421
9	Amtrak On-time Performance	0.907895
10	Rail Fatalities	0.368421
11	Rail Fatalities at Highway-Rail Crossings	0.368421

We initially suspected that a large number of null values were due to the nature of the data being recorded in discrete monthly timeframes, for example, yearly data might only have 1 record every 12 rows, but even after using regular expressions to create a “year” column and grouping with years, some columns still heavily lack sufficient data required for modeling, so we decided to set the arbitrary cutoff and only consider columns with less than 50% null values. As a result, we discovered that a lot of railroad-related data actually have a decent amount of data. After some data exploration, we decided to switch away from the original research question related to general ridership to rail ridership. In particular, we will be trying to predict passenger rail total miles using other variables in this dataset.

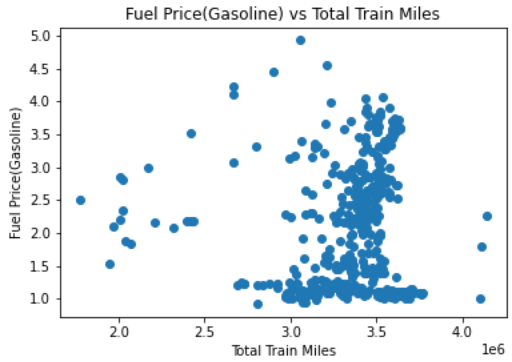
For our EDA, we decided to visualize several variables. First, thinking that the COVID-19 pandemic might be related to changes in our y: passenger rail total train miles, we created a

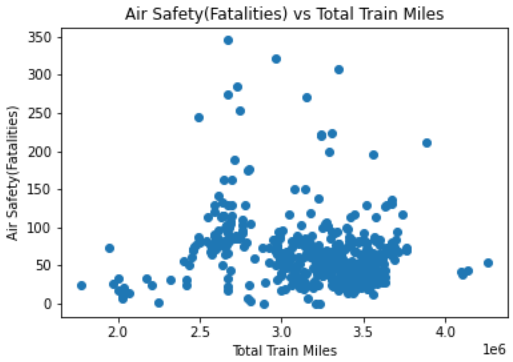
categorical variable that indicates whether or not the year is within the pandemic(2020-2022):

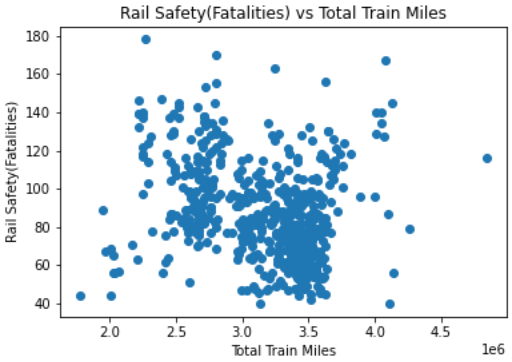


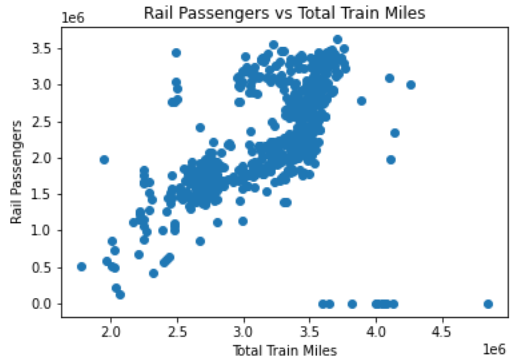
When plotting the trend of the train miles and highlighting the COVID period in orange, we can clearly see that passenger train traffic has been impacted by the pandemic. So naturally, we would want to include this information as part of our model features. To generalize, we have decided to create a categorical variable indicating if the data comes from the pre-COVID times or post-COVID period.

Furthermore, we also scatter-plotted our y-variable vs several other variables to research the correlation between them:







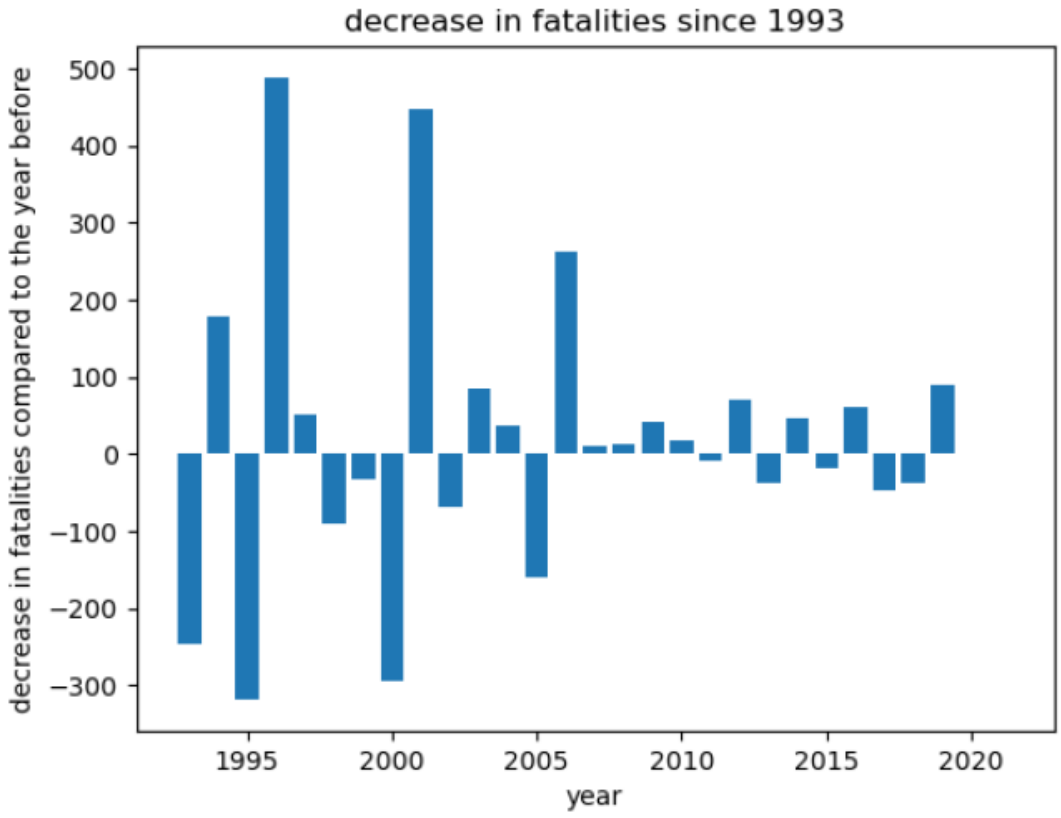


We see that the number of passengers is actually highly positively correlated with the total train miles. Therefore, we might also consider using this in our prediction modeling.

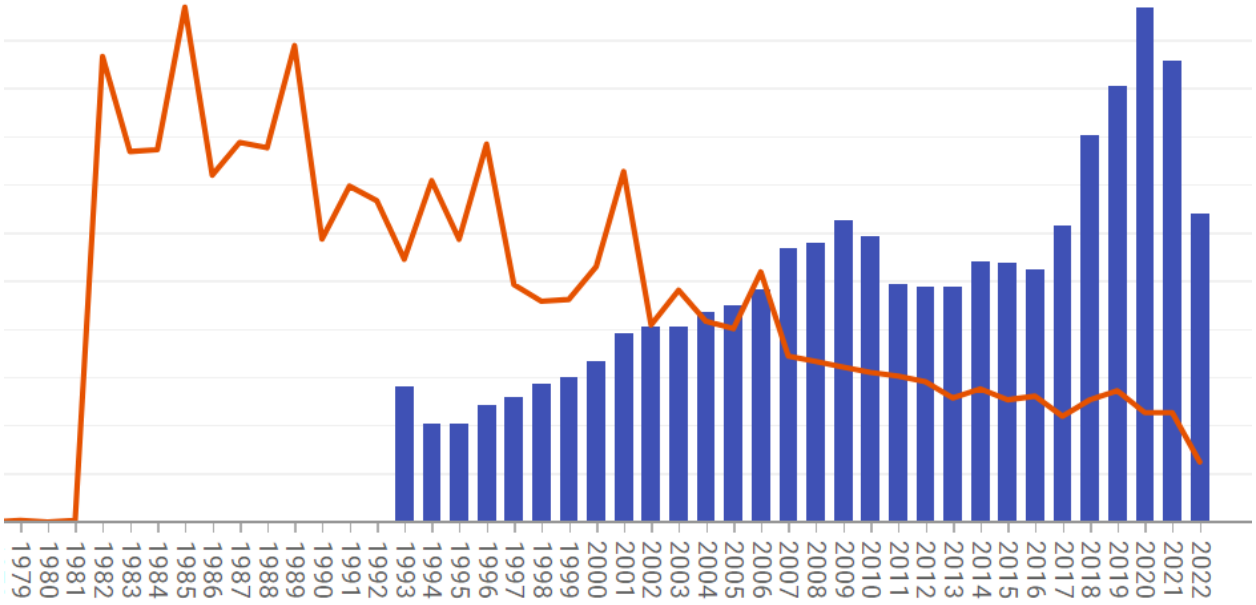
For question 2, we found that when State and Local Government Construction Spending on Air accumulates at a specific level, the general aviation fatalities would drop rapidly. The possible reason for that phenomenon is that air infrastructures like airports, air traffic controls, and aircraft were finished building and implemented. For example, a new airplane model which is designed to be safer under any weather conditions was funded five years ago and got updated in most airlines last year. Then there will be a great chance that the general aviation fatalities of this year drop greatly.

I first created a new dataframe called “air” which only includes the columns related to air. Then, I fill all NA as 0 and this should be fine because the NA in spending and fatalities means 0. I changed the data type of spending as an integer for further calculation. Here I make a new dataframe by grouping the following data by year. In this way, I can get the total spending on air infrastructure per year and general aviation fatalities per year. Then, I dropped rows with no

general aviation fatalities and named them “after”. I want to calculate the decrease in fatalities each year, so I subtract two columns and add a negative sign to show the decrease. Then, I use a barplot as the visualization for the decrease in fatalities. On the other hand, I use a barplot to show the total expense of Construction Spending on Air and on Air terminals at the same time. Then I combine two graphs and we can easily find that as different funding levels are reached, the fatalities will be decreased a lot.



By combining the chart of fatalities decrease and the chart of total Construction Spending on Air, the graph may suggest that as the funds reached a specific level, the fatalities would decrease tremendously. Upon my research on funding levels, the government usually reauthorizes the five-year funding plan into law. And we can see that as the plan and total spending reached that level, the decrease in fatalities is obvious.



For research question 2, we first create a new dataset containing only the variables related to air. Our main variables of interest are "General Aviation Fatalities" and "State and Local Government Construction Spending - Air". The original National Transportation Statistics dataset contains data for each month. We want to make the monthly data annually because upon the granularity of air fatality data, monthly fatalities are too separated. We have to consider the data under the time scope of the year so that the annual fatalities can be treated as proper data. The monthly fatalities can be influenced by many noises like seasonal weather or an increase in ridership because of specific festivals. We averaged the data for each month so that each row represents data for each year from 1993 to 2020. After this, we change the variable fatality rate into the yearly aviation fatality rate which is the total fatalities per 100,000 flight hours.

Research Question

Research Question 1: How can we predict rail passenger total miles with the variables in the data set?

We hope that the prediction will help relevant transportation regulatory organizations to better forecast the distance of trains given other features.

Research Question 2: Is there a causal relationship between state and local government construction spending in air and general airline fatality rate from 1993 - 2020 in the U.S.?

We will conduct causal inference on the observational data of State and Local Government Construction Spending on Air from 1993 - 2020 in the US. We excluded the time period after COVID-19 because of the unusual shock of the pandemic on the airline industry. We also excluded the year before 1993 due to a lack of reliable sources of data on government expenditure. Our research question can help the government to better visualize the effect of different air funding. We use causal inference for this research question because the causal assumptions are necessarily context-specific, and the reasons for fatality rates are more complex and multidimensional than a simple fit criterion based on squared loss.

Inference and Decision

Part I: GLM and Non-Parametric Method for Prediction

Non-parametric method:

For the non-parametric method we decide to use the decision tree and random forest regressor to predict the 'Passenger Rail Total Train Miles'. The reason why tree regressor is our top choice is due to three main reasons: first, we have a large dataset that could provide a huge amount of training data for the model, which is exceptionally good for tree regressor than other non-parametric methods; second, we want to make our model more explainable to anyone who is interested in how does the prediction actually take place; third, due to huge amount of features in our dataset, tree regressor are able to automatically choose the best features that could decrease the impurity of the whole tree. In addition, since the tree regressor model has a hyperparameter, we could use the GridSearchCV function from the model_selection package of Sklearn to tune those parameters to get better performance.

Data Selection:

Since the tree regressor will self-select the feature to split to minimize the impurity, the most ideal circumstance is to fit the model with all available variables in our dataset. However, this dataset has the following drawbacks that don't allow us to do so:

1. First, the dataset has a huge amount of missing data before 1975, which makes it difficult for the model to fit the data.
2. Second, variables all have an imbalanced portion of missing values, and we can't fit the model with that poor input.

```
X_names = list(nn[nn['na%']<0.6]['variables'])
X_names
```

```
['Index',
 'Date',
 'Air Safety - General Aviation Fatalities',
 'Freight Rail Intermodal Units',
 'Freight Rail Carloads',
 'Highway Fuel Price - Regular Gasoline',
 'Passenger Rail Passengers',
 'Passenger Rail Passenger Miles',
 'Passenger Rail Total Train Miles',
 'Passenger Rail Employee Hours Worked',
 'Passenger Rail Yard Switching Miles',
 'Passenger Rail Total Reports',
 'Rail Fatalities',
 'Rail Fatalities at Highway-Rail Crossings',
 'Trespasser Fatalities Not at Highwaya-Rail Crossings',
 'Heavy truck sales',
 'Light truck sales',
 'Auto sales',
 'Heavy truck sales SAAR (millions)',
 'Light truck sales SAAR (millions)',
 'Auto sales SAAR (millions)',
 'Year']
```

To solve these problems and get appropriate training and test data set for our model, we first filter the data only between 1975 and 2020. Also, we wrote a function that could display the portion of missing values for each column, and we only select the columns that have less than 40% of their data missing. The final list of available variables is listed here. (Excluding the Index and the date).

Training and Testing the data:

We use both the decision tree and random forest method to fit the training data with the variables listed above. We expect the random forest to perform better than the decision tree since it is the ensemble method. As we are using the R^2 to measure the performance of our tree regressor, it turns out that the training accuracy for the decision tree and the random forest is 0.82 and 0.9, respectively. Moving into the testing part, we found that the test set accuracy for the decision tree and the random forest is 0.726 and 0.834. It showed that indeed that random forest performs better than decision trees on both training and test data. To further improve our model, we decide to use a GridSearchCV on the max_feature parameter of the random forest model to carry out a

10-fold cross-validation on max_feature ranging from 1 to 19. It turns out that the final test accuracy increases to 0.86 which is not a significant improvement and the optimal max_feature number is 9.

Interpret the result:

Since we are using a tree regressor, the R^2 is the best measurement of performance. The R^2 is 1 – SSE/Baseline Variance, which means that decision tree model variance only composes 27.4% of the baseline model (average prediction) and random forest model variance only composes 16.6% of the baseline model.

We could say that the result is a fair prediction but not a very good one, and it certainly has many limitations which will be further analyzed in the discussion part where the non-parametric method result will be evaluated against the GLM result.

Prediction with GLM:

Model Choice and Assumption(Bayesian vs Frequentist)

For predicting the 'Passenger Rail Total Train Miles' using GLM we need to make some assumptions on the choice of the likelihood of our prediction target. First, since we are predicting a discrete count whole number, our choice of likelihood is limited to Poisson and negative binomial distribution. However, the Poisson GLM assumes that the variance of our target equals its average. We did a quick check and find out that the variance is way bigger than the average, so we rule out the Poisson distribution and the only choice left is negative binomial.

It is actually a good choice since negative binomial distributions are usually used to approximate random variables in which the variance is bigger than the mean. And for prior to our target, we assume that the distribution should be normal since total train miles are like another real-life scenario that most data falls on the average and very high and very low Rail miles are rare to see. The link function we used in Negative Binomial GLM is the exponential function, which makes sure that all the data inputs are positive for the GLM to work.

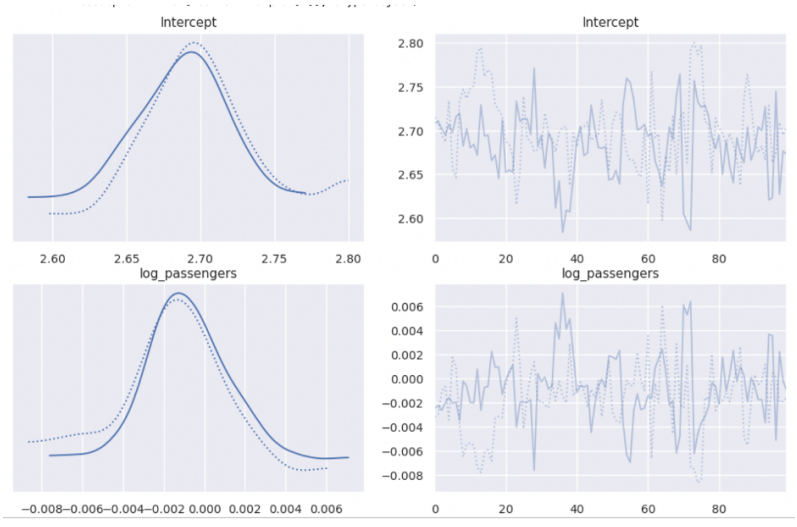
For the frequentist part of GLM, it assumes that the parameter beta is fixed and it would generate the estimation of beta using the maximization of likelihood.

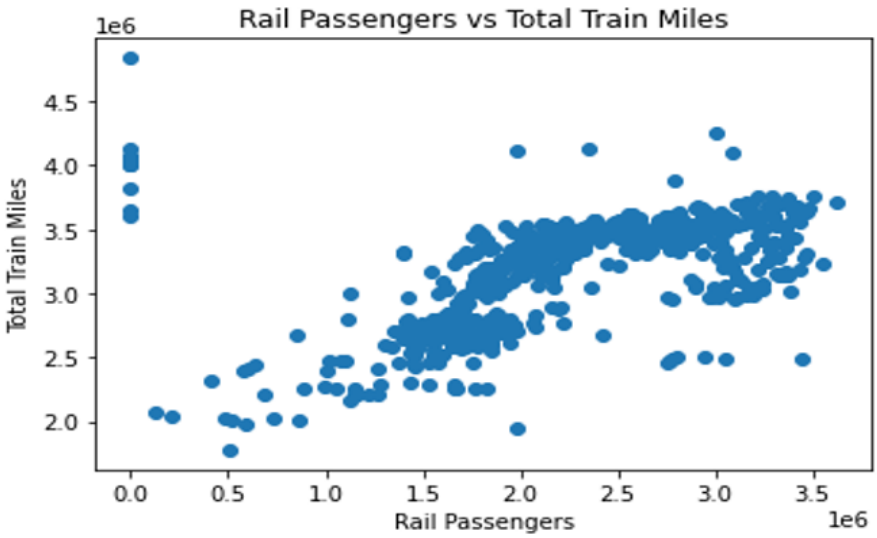
Feature Selections:

As in the non-parametric method, we first narrow down our choice to features that has less than 40% of missing values, and in that list of variable we use domain knowledge of railroad common sense to draw multiple scatter plot of variable in the EDA section. After several explorations, we assume that there might exist a positive linear relationship between the number of passengers to the total rail miles and the number of passengers may serve as a good predictor for predicting total train miles.

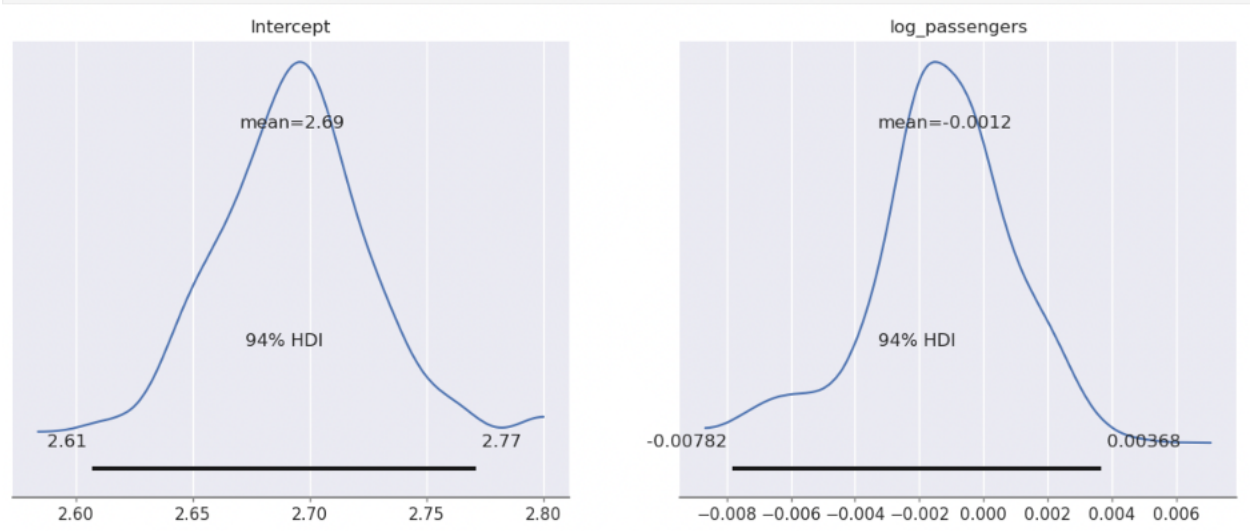
Results & Uncertainty Quantification:

For Bayesian regression, as shown in the graph below, the posterior distribution of intercept and log_passengers shows an average of approximately 2.69 and -0.0012, respectively. To interpret the result: we could see that the coefficient for log_passenger is very small, and it is negative, which is contradicted our EDA that shows a positive correlation between the two variables. However, since we take the log of both total miles and the number of passengers in order for the bayesian model to run successfully, a unit increase in the number of passengers actually means an e^-0.0017 change in the total miles. This means that the effect of the number of passengers still has a positive impact on the total miles, but it just shows a decreasing growth rate. This actually aligns with our EDA graph in which we could interpret the shape of the graph as a function that is increasing (first derivative > 0) but with decreasing growth rate (second derivative < 0).





To quantify the uncertainty in our bayesian regression, we could use the credible interval for both intercept and coefficient for log_passengers. As shown below:



We could say that based on the credible intervals, the highest density interval, in which 94% of falls within it, are between 2.61 and 2.77 for the intercept term. For the coefficient of log_passenger, there is also 94% of chance that the coefficient will fall within -0.00782 and 0.00368. The implication of these indicators will be further discussed in the discussion section.

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	Poisson	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-755.42
Date:	Wed, 17 Feb 2021	Deviance:	1366.3
Time:	12:51:51	Pearson chi2:	1.20e+03
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.9697	0.023	219.386	0.000	4.925	5.014
year	0.1829	0.001	132.547	0.000	0.180	0.186

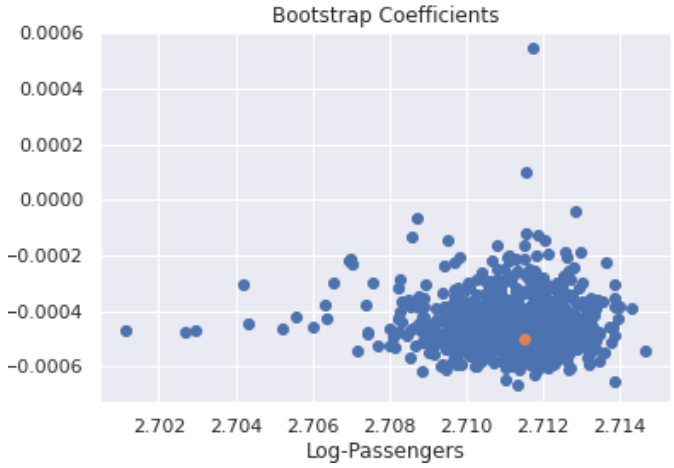
Frequentist GLM Summary:

From the output of the stats model, we see that the 95% confidence interval for the constant is [2.384, 3.039], whereas the 95% confidence interval for our coefficient of log passengers is [-0.023, 0.022]. Comparing this back to our bayesian GLM, we see that indeed the results are consistent with the constant being centered around 2.7 and our coefficient being slightly centered left to 0.

To further investigate the uncertainty, we coded up a bootstrap function that uses random samples from our original dataframe to calculate the coefficients, and we get the standard deviations for both coefficients to be much lower than the standard error provided by the frequentist sm model:

constant sd: 0.0014
log-passengers sd:0.00012

We also plotted our bootstrap results and we can see that indeed the interval is very tight:



Discussion, Evaluation, and Comparison between GLM and Non-Parametric Method:

1. To compare the performance between GLM and Non-Parametric method we use the mean of the posterior distribution of intercept and coefficient to get a single variable regression model to predict the total rail miles but ends up getting a negative R^2 of -0.244, which means that the GLM performs way worse than the non-parametric method of decision tree and random forest if we are only looking at the R^2. The negative R^2 of GLM means that the GLM did worse than merely guessing the average of rail total miles. However, this comparison might not be so adequate since the decision tree and random forest uses up to 19 features to come up with the prediction while our GLM only has one feature.
2. For the non-parametric method, decision trees, and random forests have relatively high training accuracy. We expect that tree regressor to have an over 90% training accuracy while fitting, however, maybe due to the lack of feature engineering such as standardization and normalizing the scales of different features, the non-parametric method only has a fair fit to the data. For GLM I would say that even though the coefficient of the number of passengers aligns with the trend of decreasing growth trend of total miles against the number of passengers, it did fit the data poorly by showing that all the prediction tends to be too small compared to the real test set. I would say that the negative binomial likelihood for approximating total miles of rail isn't a good choice. From the EDA part, we could see that the total mile of rail show more of a left-skewed distribution which is unmatched by the negative binomial distribution.

3. I would say that both the non-parametric method and GLM are ready to use for prediction for total rails in the future since they didn't yield a satisfactory R^2 in the training and test set. For the NP method, more feature engineering is needed and hopefully, the ticket type of each passenger could be new data that could greatly improve the NP method training process. For GLM, the most important part is to determine the likelihood of total miles, and due to a large number of missing values and imbalanced features, it is quite difficult for us to come up with an appropriate likelihood that fits GLM well, so I would say a more comprehensive dataset would help the GLM the most.

Part II: Causal Inference

Methods

The treatment variable here is the yearly state and local government construction spending on air in the U.S.. The data is provided by the U.S. Census Bureau's monthly estimates of the total dollar value of construction work done in the United States. And the outcome variable is the yearly aviation fatality rates (total fatalities per 100,000 flight hours). We originally planned to use the "Air Safety - General Aviation Fatalities" column in the Monthly Transportation Statistics. Yet, we soon realized that it would be more inappropriate to use the fatality rate instead of absolute fatality numbers. So we found the extra dataset of fatality rate calculated per 100,000 flight hours. The unit here is the U.S. from 1993 - 2020. To account for the potential delaying effect of construction spending, we will investigate the impact of government spending on the fatalities rate 2 years.

Before adjusting for confounding factors, we first take a look at the correlation between the two variables of our interest. Given the scale of the two variables are different - expenditures come in billions of dollars while fatality rates are between 0 - 3. We first take the log of expenditure variables before drawing a scatter plot to visualize the relationship between the two. Another feature engineering here is that we take into account the time-delaying effect of expenditure. The construction can take a few years to complete. Therefore, we calculate the "lagged fatality rates" for 2 years. We will repeat the causal inference analysis for each situation. From the scatterplots

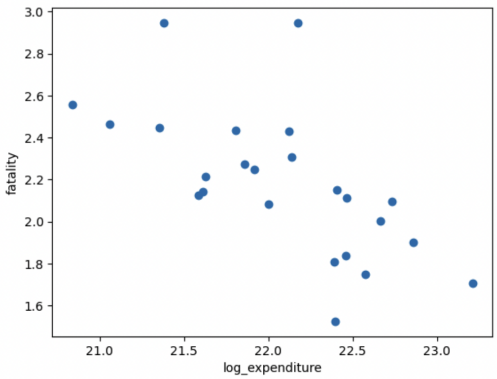
below, we can see there seems to be a negative correlation between the two variables. And the coefficient of an OLS regression confirmed that as government expenditure increases, fatality drops.

OLS Regression Results

Dep. Variable:	lagged_fatality_rate2	R-squared:	0.416
Model:	OLS	Adj. R-squared:	0.390
Method:	Least Squares	F-statistic:	15.69
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	0.000663
Time:	22:24:39	Log-Likelihood:	-2.0507
No. Observations:	24	AIC:	8.101
Df Residuals:	22	BIC:	10.46
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.7341	2.158	4.973	0.000	6.258	15.211
log_expenditure	-0.3873	0.098	-3.961	0.001	-0.590	-0.185

Omnibus:	9.438	Durbin-Watson:	1.292
Prob(Omnibus):	0.009	Jarque-Bera (JB):	7.834
Skew:	0.994	Prob(JB):	0.0199
Kurtosis:	4.971	Cond. No.	850.



The next step would be to deal with confounding variables. Three confounding variables we adjusted for our annual national GDP and averaged gas price index in the U.S.. We consider GDP to be a confounding variable because when the economy is going well, the government spending on air construction increases as well. It also affects fatality rates through airline operations. As demand for airline services increases, the airline industry becomes more profitable, investing more in safety infrastructures and eventually decreasing fatalities. We found extra data from FRED on yearly GDP. The other confounding variable is gas price. Gas price increases usually cause the government to tighten the expenditure on construction. It will also increase the cost of airline operations, decreasing demand. As airline operations are in difficulties, safety measures might be compromised, leading to more accidents. We found the dataset on the gas price index from FRED (US Regular All Formations Gas Price). The last but not least confounding variable here is time. Over the years, fatality drops gradually, and government expenditure increases. We acknowledge that the passing of time does not directly affect our treatment and outcome variables, but we use “year” here to adjust for other confounding variables we can’t easily quantify, such as the advancement of technology, and the development of society.

We first used inverse probability weight to adjust for the confounding variables. We learn about using inverse propensity weight on binary treatment variables in the class. However, we are dealing with a continuous variable here - expenditure. From searching on the internet, we found a way of calculating propensity scores for continuous variables using densities - how likely does our observed variable appear given a predicted distribution. We first ran a regression on "log expenditure"(A), our treatment variable against all the confounders (X), and obtained a function $g(x)$. Then define a conditional distribution $Normal \sim (g(x_i),\ var(a_i - g(x_i)))$. The conditional mean of each sample is its prediction and the variance is the variance of the predicted residuals. We defined the density as the value of a_i from D_i. And the weight is the inverse of the density. Once we obtain the weight, we regress the outcome against the treatment, weighted by the IP-weights we obtained. Then we can interpret the coefficients as causal effects. One assumption we are making here is the linear relationship between our treatment and outcome variables.

$$A = g(X) + \epsilon = \alpha_0 + \alpha_1 X$$
$$\tilde{A} = A - g(X)$$
$$A_i \mid X_i \sim Normal\left(g\left(X_i\right), Var\left(\tilde{A}\right)\right)$$

WLS Regression Results

Dep. Variable:	lagged_fatality_rate2	R-squared:	0.556
Model:	WLS	Adj. R-squared:	0.536
Method:	Least Squares	F-statistic:	27.58
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	2.87e-05
Time:	22:35:41	Log-Likelihood:	-7.1997
No. Observations:	24	AIC:	18.40
Df Residuals:	22	BIC:	20.76
Df Model:	1		
Covariance Type:	nonrobust		

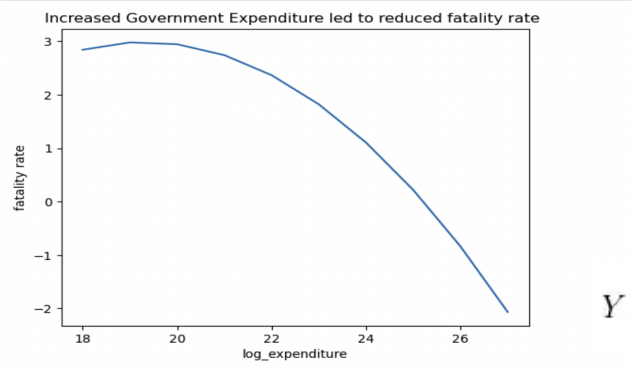
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.3241	1.908	6.458	0.000	8.366	16.282
log_expenditure	-0.4544	0.087	-5.252	0.000	-0.634	-0.275

Omnibus:	11.758	Durbin-Watson:	1.433
Prob(Omnibus):	0.003	Jarque-Bera (JB):	10.607
Skew:	1.215	Prob(JB):	0.00497
Kurtosis:	5.167	Cond. No.	731.

Notes:

We can also have a non-linear transformation of the main treatment variables, by adding the squared term of "log expenditure". *(Marginal Structural Model)* But then we won't be able to interpret the treatment effect as the coefficient of treatment covariate. Instead, we make counterfactual outcome predictions by setting some treatment value and running it though our

model to get predicted outcomes. We can see how the counterfactual outcome prediction changes as a function of different dosages. From the graph, we can see that as log_expenditure increases, the fatality rate drops, confirming a negative causal effect of government expenditure in air construction on air fatality rates.



$$Y = \beta_0 + \beta_1 A + \beta_2 A^2, \text{ weighted by } w$$

The second technique we use to adjust for confounding variables is outcome regression. Regressing the outcome variable on both treatment variables and confounders, we will use the coefficient to estimate the treatment effect. Two feature engineering we did here is taking the log of GDP variables given its large scale and constructing a variable called “year effect” which is the “year” column minus the start year. The aim is to scale the variables to the same level. We are making three assumptions here. First, we assume that GDP and gas price are the only two confounding variables here. Second, the linear model correctly describes the interactions between the variables. Third, each confounder affects our treatment and outcome variables equally. We will try and see how it performs before discussing the limits of our three assumptions.

We run OLS regress on lagged-fatality rate for 2 years against “log_GDP”, “gas”, “year_effect” and our treatment variable “log_expenditure”. To our surprise, the coefficient of “log_expenditure” becomes positive, contradicting our original hypothesis that increases in government spending reduce. Furthermore, the coefficient of “log_GDP” is also negative. The coefficient of the “year effect” is positive, aligning with the trend that fatality decreases with time. We also find that both “log_GDP” and “log_expenditure” coefficients have relatively high p-value, meaning that they are not statistically significant. Relating back to the assumption we made above, one possible explanation for the result is that we have not yet considered all the

significant confounders. Their relationship might not be linear as well given airline accidents are random events. We also have a limited size of samples due to the lack of data.

OLS Regression Results

Dep. Variable:	lagged_fatatility_rate2	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.582
Method:	Least Squares	F-statistic:	11.69
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	0.000121
Time:	22:24:39	Log-Likelihood:	3.6433
No. Observations:	24	AIC:	0.7134
Df Residuals:	20	BIC:	5.426
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.3244	7.369	-0.587	0.564	-19.696	11.048
log_expenditure	0.2967	0.251	1.184	0.250	-0.226	0.819
log_GDP	0.0779	0.958	0.081	0.936	-1.921	2.076
year_effect	-0.0582	0.032	-1.799	0.087	-0.126	0.009

Omnibus:	10.504	Durbin-Watson:	1.596
Prob(Omnibus):	0.005	Jarque-Bera (JB):	9.238
Skew:	1.071	Prob(JB):	0.00986
Kurtosis:	5.156	Cond. No.	4.42e+03

Notes:

Conclusion:

For research question 2. Through inverse propensity weight, we found that log government expenditure has a negative causal effect of -0.54 on the fatality rate. However, with outcome regression, we had a contradicting result. Even though the result of outcome regression is not statistically significant. It still weakens the previous hypothesis. Overall, we conclude that we are unable to confirm there exists a negative causal relationship between government expenditure on fatalities without further evidence support. But our analysis shed light on many other aspects of the problems, which is very valuable for possible future analysis. For example, we found that government expenditure can be well predicted by GDP, gas price, and year with a R_sqaured value of 0.912. Given that government expenditure is closely related to economic development, it is crucial to exclude effects of economic growth on fatalities. For further analysis, one possible approach is to use instrument variables, which might require experts' knowledge and delve deep into the factors that affect government expenditure. Given that our results are constrained to 1993 - 2020 in the U.S... due to a lack of data sources, it would also be desirable to gather more data on government expenditure and fatality rate. In this analysis, we merged data from three sources

and combined them using years as the index. Combining different data sources gives us the freedom to include the confounders of our interest.

From our findings, In the 2010s, with the development and progress of technology and improvement of systems under continuous government funding, the number of aviation accidents caused by human errors and mechanical failures decreased significantly. Therefore, it can be seen that the probability of aviation accidents actually has room for downward pressure.

We have a lot of limitations in the data that we could not account for in our analysis. Upon research, the main cause composition of aviation accidents accounted for half of the human causes such as pilot operation errors, followed by objective causes such as mechanical failures and bad weather. We can't find any useful data that can explain the effects of bad weather conditions, pilot operation errors, and mechanical failures on air fatalities. Also, the construction time of funding aviation infrastructures should be considered. According to the government document, the government usually employs a five-year FAA reauthorization signed into law. The effect of government funding needs to be considered in the real world. The original dataset has its own limitation, which is that the relevant data are limited to apply.

The reason that our study only considers limited features like GDP, fuel price, and government spending, future research can merge more possible features from different datasets, such as specific distribution of air accident reasons, the specific government funding distribution, and the specific construction timeline of government spending.