DATA SCIENCE 102

DATA, INFERENCES, AND DECISIONS

FINAL PROJECT - SPRING 2023

Contents

Data Overview	3
Research Questions	5
Exploratory Data and Analysis (EDA)	7
Question 1	14
Question 2	17
Conclusion	20

Data Overview

We obtained the *Bureau of Transportation Statistics: Monthly Transportation Statistics* dataset to do our analysis for this project, which was generated by the U.S. Department of Transportation. They collect information through state, local, and federal agencies about transportation utilization and spending. After collecting data, they then assemble the information to prepare it to be published and used to determine how to allocate funds. This published data is a sample of current transportation information from 1947 to 2023.

In this data, there is information about personal spending on transportation, motor vehicle sales, fuel prices, state and local construction, and more. Looking at the data there is information from 1947 to 2023, however there is quite a bit of missing data from 1947-1966. That timing makes sense because the world is evolving and we had a few variables we did not think to account for until time passed. According to the data, after each half decade from 2000 to 2020 there is an increase in government spending on transportation. That is to be expected since the world continues to grow and develop for the increasing population. Another similarity seen in the data is that the unemployment rate should be rising and falling due to economic cycles and the unemployment rate indeed goes through a pattern of decreasing to increasing.

Furthermore, in the data there is a key difference to what we would expect to happen within the population. There is a decrease in personal spending on gas in 2010 however there is a high increase again in 2015. We would assume that there is a constant increase due to the constant growth of population. In addition, since there are more roads, more people would want to travel and explore.

This affects the generalizability of our results because it allows us to learn a bit about our past data to predict future data. We can potentially assume that the unemployment will continuously increase and decrease on a loop. As for government spending on transportation, it can potentially continue to increase because of the continuous efforts to grow our society. In addition, based on what we see for personal spending we can assume that it would change based on significant changes in our societies economy.

In terms of the participants, they were potentially not aware of the collection of the data if they don't keep track of or stay up to date on the U.S. Department of Transportation data collection. The U.S Department of Transportation does publish the information so there is transparency. However, there is no advertisement of the data being delivered to people's houses.

The granularity of the data is in days. Each row has a designated date for all the up-to-date information per day. This will impact our findings because having each day of the year and a huge range of years will force us to group the data and create different means based on the groupings. We are not grouping by changes in the data, like an increased spike, but we are grouping based on half decades, so that we are not grouping with too wide a range. Due to our choice of half decades, we are able to analyze the data in an efficient manner. We don't have to worry about separating outliers yet and we don't need to concern ourselves with more data than we can handle.

A concern that we do have is how we are choosing to divide up the government spending: by half decades. It could blind us to a clear connection from 2000 and 2005. In

addition, another concern we have is whether some features are more helpful than others when it comes to the effect government spending has on unemployment. Furthermore, we are worried about the potential idea that personal spending on gas is increasing for another reason, like scarcity. There are many outside factors we are trying to control, but we can only do so much. Overall, we are concerned about selection bias, in terms of how we group government spending (by decades, high/med/low levels of government spending and personal spending). The dataset is not modified for differential privacy. All features in the dataset are more than helpful. They contain values and rates of a specific group of information.

There are quite a bit of missing values in some of the dates listed in the dataset. This means that they were not collecting that information at that time. They slowly started accounting for other information as the time passed. In our analysis we dealt with this by focusing our analysis on the time we are able to fully analyze to the best of our ability. That is why our analysis is about the 2000's to 2023 rather than the 1940's to 2023.

First we made sure to narrow the data to our independent variables and dependent variables so that we could look at all the data in an efficient way. Afterwards we renamed the variables, since they were quite long in the original data. By renaming the variables, we are also able to understand our data better with more efficient names. Also, we dropped all the none values, so that we can condense to usable data. In addition, we used regex to change the date to show the year only. That way we could analyze by the year and not just the day, which would be too much information. These decisions impacted our models because instead of focusing on the per column values we created our own groupings. Even though the different groupings were not causing a great change, depending on what values we looked at, there was still a slight change that was enough to change the mean of values and the variance. It changed the effect we would see from 2000-2005 to 2000-2010.

Research Questions

Question 1: How does government spending on transportation affect personal spending on gas?

We want to explore the impact of government spending on transportation, in terms of personal spending on gasoline. This can help answer many real-world questions, influencing important decisions regarding government expenditures on public works projects so that they are most efficient. Often times, the government spends extensive amounts of money on projects which have little impact; this is a waste of money. We want to make decisions which maximize the amount of benefit per dollar for the public.

For example, the government may want to encourage alternative transportation methods to driving in order to reduce effects of climate change. We can see if devoting government spending towards growing public transit reduces the frequency at which people use personal cars (as measured by how much they spend on gas). We can also see where else the government can devote funds in order to reduce driving traffic and pollution.

We will be using causal inference methods to answer this question. This approach is a great fit for our problem as we are trying to predict the effect of a treatment on the population. Government spending is our treatment. We can observe the impact of high vs low government spending, prescribing spending to “high” vs. “low” after inspecting our data and choosing appropriate threshold (the mean, in this case). Personal gasoline spending is our outcome. We make note of other possible confounders and observe their effect on personal spending through creating a regression model.

Since this is an observational study, we cannot rely on randomness to take care of confounders. There may be many other confounders in this case, and we must rely on techniques that deal with confoundedness to make our model more accurate. We use inverse propensity weighting to do this. Also, another limitation of our method is that we have to categorize government spending into “high” or “low”, and we are choosing this threshold ourselves. Some instances may be very close to each other (near the mean), yet one may be categorized as “low” and one may be categorized as “high,” although they are not that different. This may not lead to the most accurate result.

Question 2: Does better mass transit transportation funding imply lower levels of unemployment?

We want to explore how public transportation levels impact unemployment. This is quite important for government officials to observe before making decisions on where and how much funding to devote on public projects, especially with unemployment in mind. During times of economic crisis, the government can also choose a spending plan which minimizes unemployment and boosts economic activity. We want to explore whether funding mass transit transportation decreases unemployment because it helps workers travel to jobs, interviews, etc. when they do not have/cannot afford other transportation.

For this question we used prediction methods such as Generalized Linear Models (GLM) and non-parametric methods like random forests and decision trees. We thought this would be

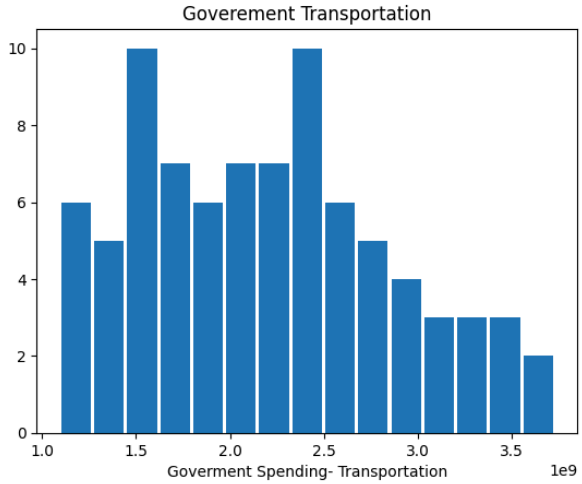
the best fit as we can predict how much government spending is needed to affect unemployment by the desired amount. We wanted to see if mass transit had an effect and how much this effect was on unemployment. We chose the Normal GLM with the identity link/model because we believed that the unemployment rate follows the normal distribution the closest. Decision trees and Random Forests were also important for our prediction because we aren't very certain about our assumptions for our model.

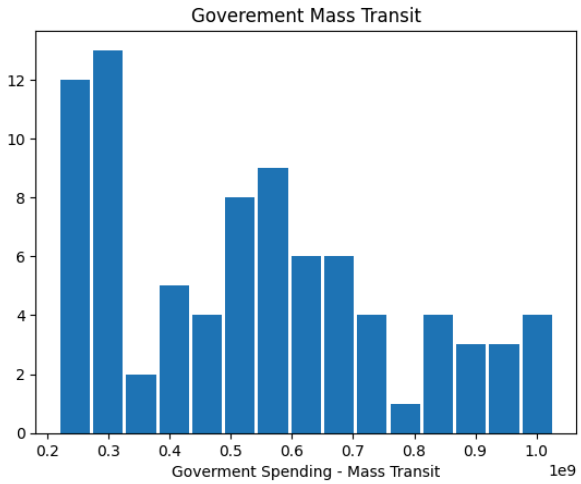
Some limitations of GLM are that we are choosing our prior distribution to be Normal. We are not sure if this is accurate enough. We may be getting an incorrect estimate if our distribution is not close to Normal. Thus we need non-parametric methods to compare to make the best decision on our predictions.

Exploratory Data and Analysis

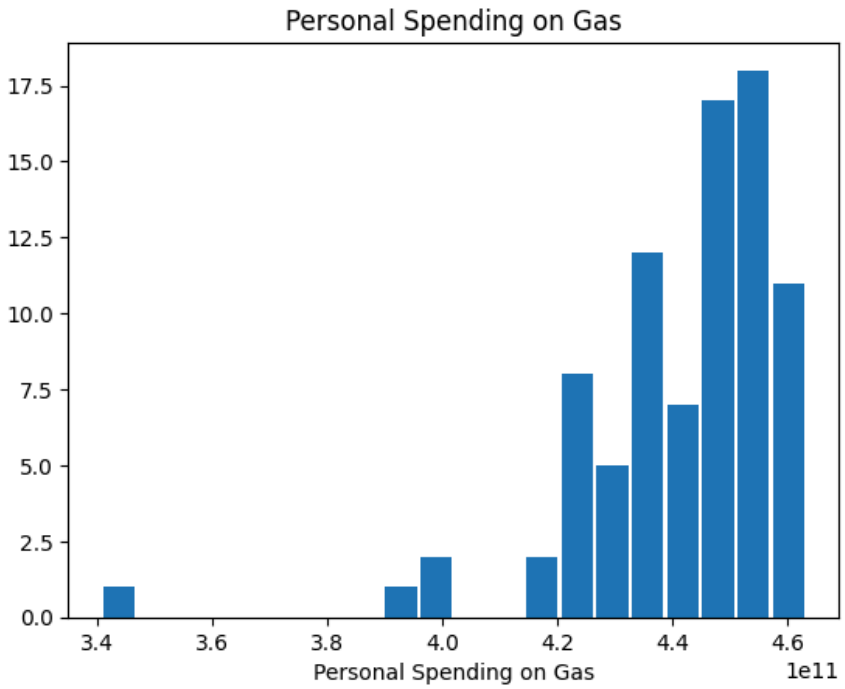
Histogram and KDE of Categorical Variables

By creating a histogram of our independent variables, it will allow us to analyze the government transportation and government mass transit per year. Through a histogram we will be able to analyze the values and their densities. For the histogram of government transportation, it allowed us to create the high and lows of government spending. In addition, we created a histogram for personal spending to determine the high and lows of personal spending. Overall these values will help determine whether we want to create the line between high and low spending using these values or by using certain percentiles.

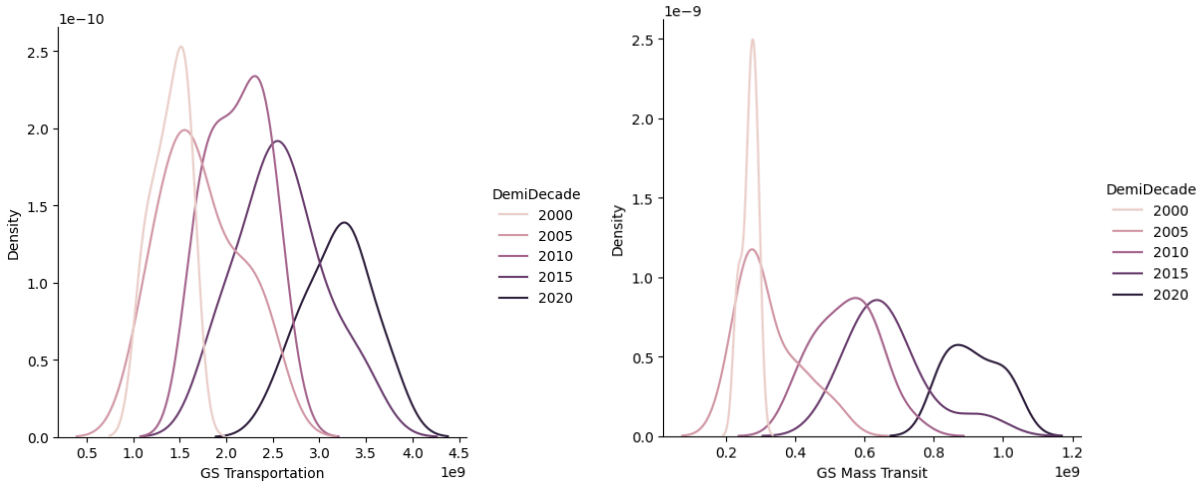




These graphs are the histograms of the government spending on mass transit and transportation. By looking at the histograms we can try to get an idea of how we might classify low and high spending. In this case, we picked with the mean as there is no distinct bimodal look to the distribution. Without a bimodal distribution, it would be hard to justify any other type of classification of high and low spending. We can also see that transportation is more normally distributed in terms of spending than mass transit is. Both graphs also have a slight right skew.

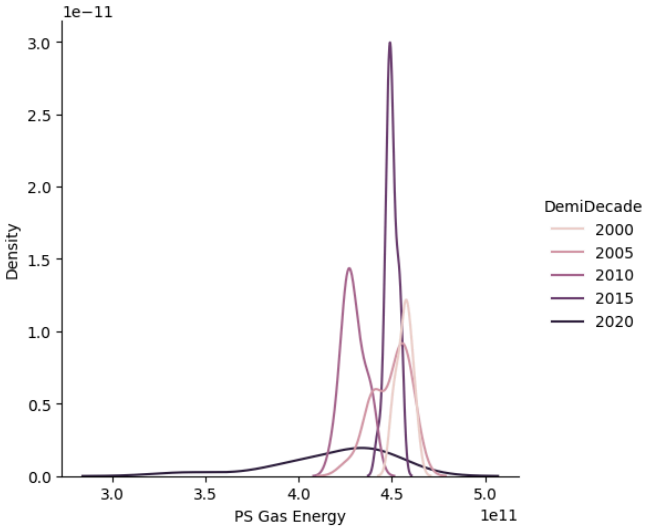


This histogram of personal spending on gas is important to note as it allows us to see the dispersion of spending. Our histogram has a tight distribution with a left skewed distribution and possibly outliers. With little variance in spending it might make it difficult to assess whether the treatment has a large effect on spending since it changes so little from year to year. Also if our outcome variable isn't normally distributed it might make our linear regression results questionable as we fail to meet one of the necessary assumptions.



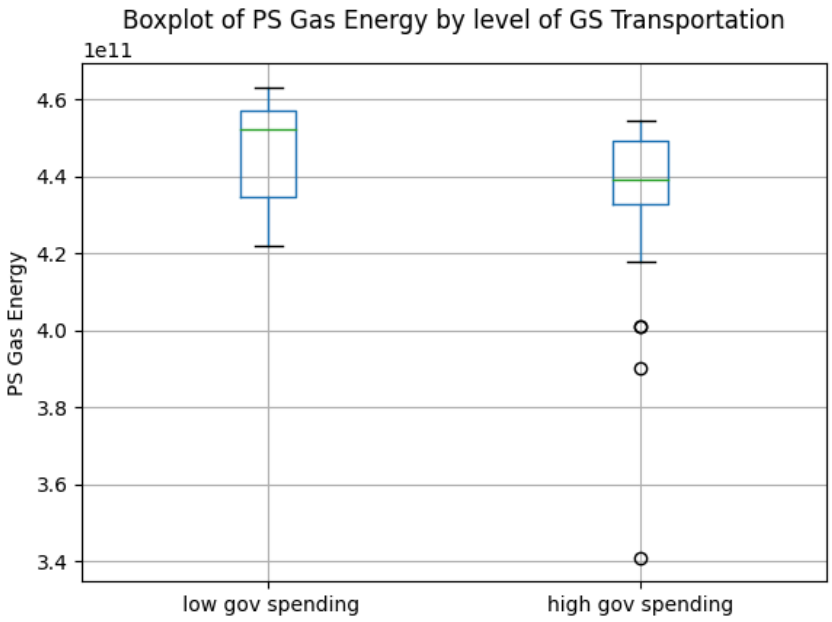
In order to analyze the independence for question 1, we first transformed the Date column to have only the year so that it could be used to determine the densities of each

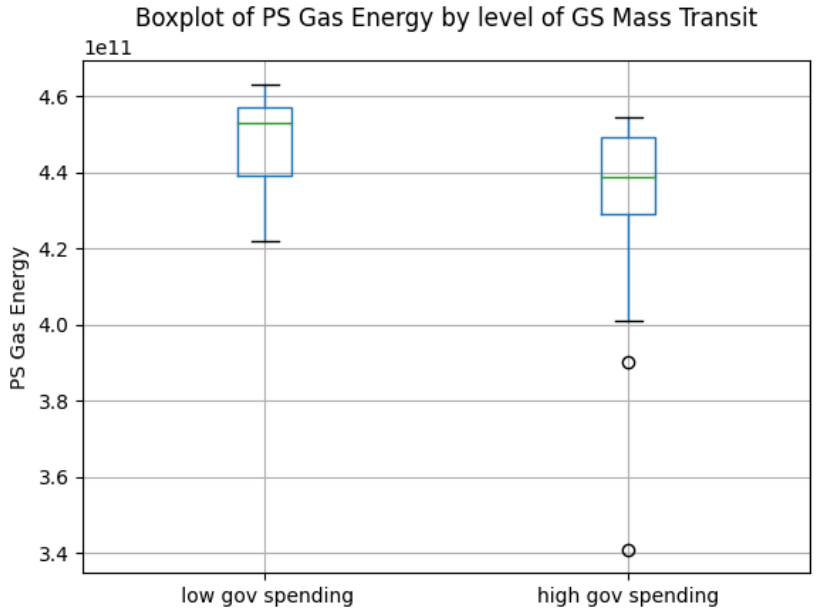
independent variable based on every half decade for the KDE graphs. This KDE plot assists us in seeing the difference in government spending based on the time of year and the government spending in mass transit and seeing if the year had a significant effect on the treatment selection. As we can see there are high densities in government spending on transportation and the range is much larger than the government spending on mass transit. We can also see that mass transit depends more on year and thus we can't assume randomized treatment as comfortably for mass transit as for transportation.



This KDE plot will assist us in noticing the changes of personal spending over the years. It will give us the ability to notice any patterns that we need to keep in account when determining whether the government spending has an effect on personal spending. By looking at this graph we are able to see that the personal spending is around $4.6 * 10^{11}$ in 2005 and goes down and then in 2020 we reach about $4.5 * 10^{11}$. It decreases in 2010 and goes up a bit in 2020. There seems to be little correlation between year and spending.

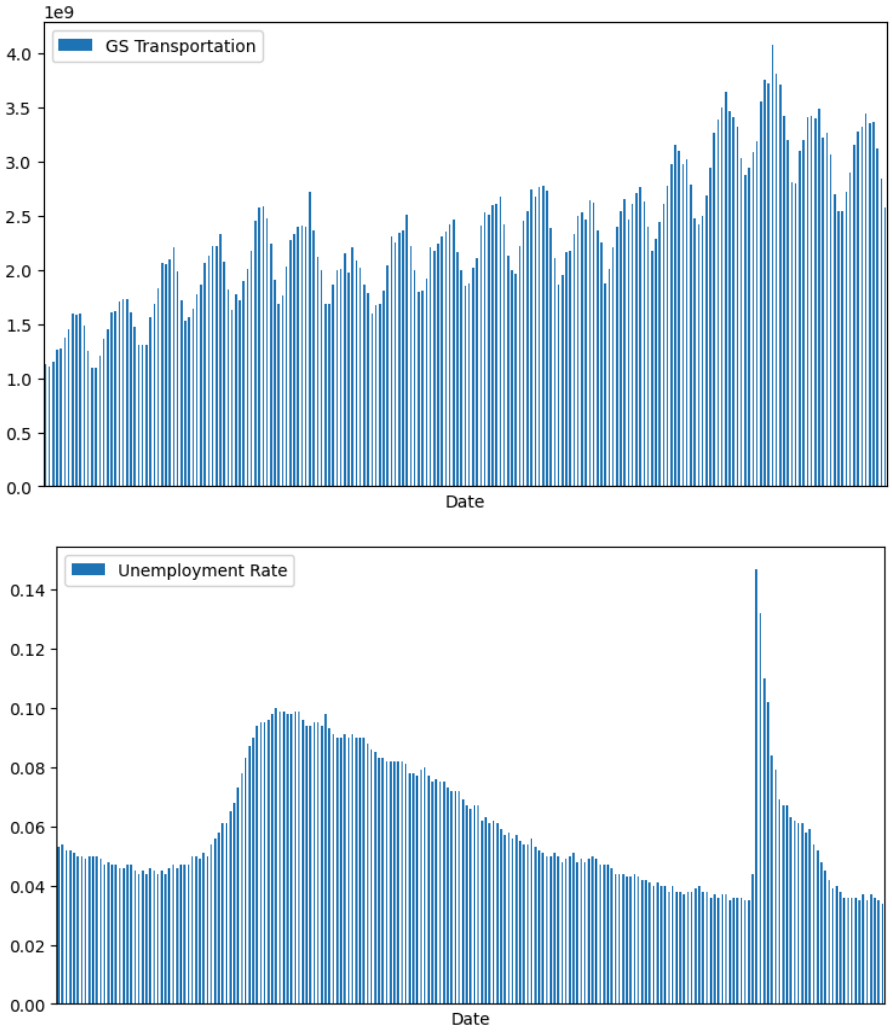
Boxplot of Categorical Variables



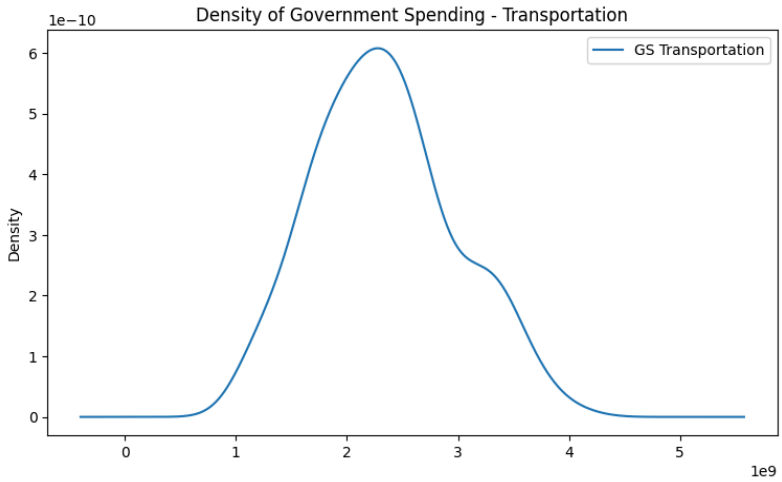


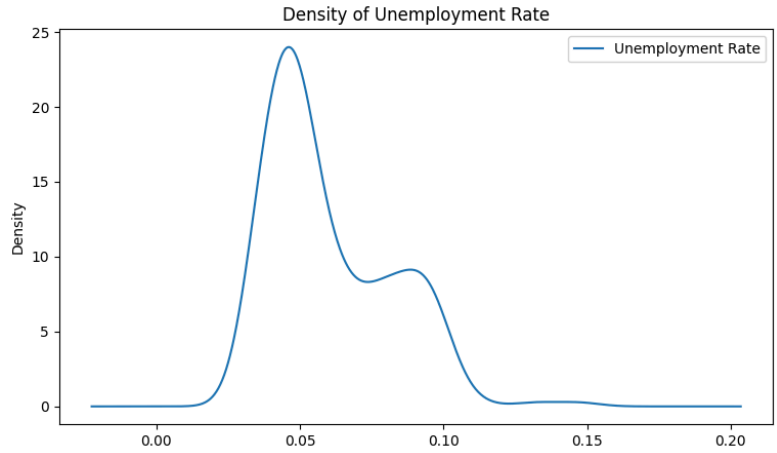
Based on the boxplots above we can see that there is a negative correlation between government spending and personal spending on transportation. We can see this is true for both types of government spending. This helps to understand the effect of high and low government spending periods on personal spending. This is important for our causal inference model as we are seeing if government spending has an effect on personal spending on transportation. It also raises other questions that might need to be answered, such as does more government spending result in lowered personal spending in later periods.

Histograms and KDEs of Quantitative Variables



For question 2, we decided to plot the histogram to look at the distribution of government spending on transportation and unemployment rate to look at the quantitative variables and how they have changed over the years. They are graphed over the same periods of times after dropping all NA's from both columns. We will go on to look at other confounding variables that might have led to the changes in values for the two.

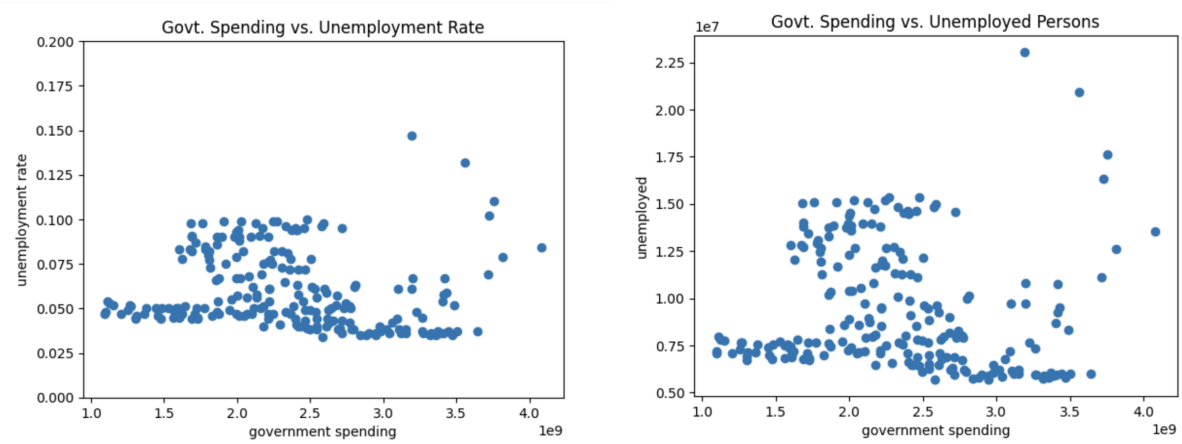




We also plotted the KDE plots for both columns to see the density. Even though we didn't see much similarity in the histogram plots, the density plots show a similar pattern for the two values. This could indicate slight correlations between the two values and it is something we can go on to look at more in depth in this project. We might also find similar trends in the density of the other confounding variables.

Scatterplot of Quantitative Variables

Our two quantitative variables are government spending (on transportation) and unemployment. There are two variables that represent unemployment: “unemployment rate,” and “unemployed” (which is number of unemployed persons). In question 2, we are trying to see the impact that government spending on transportation has on unemployment. Looking at this graph can help us observe any important trends in our analysis.



We can see that both “unemployment rate” and “unemployed” follow the same trends (as expected). Thus, it would be safe to use either for our model. We can also see a slight linear relationship between government spending and unemployment rate, but also a lot of noise. This noise could skew our results as there are many factors that affect unemployment rate that are not always quantifiable.

Question 1 (Causal Inference)

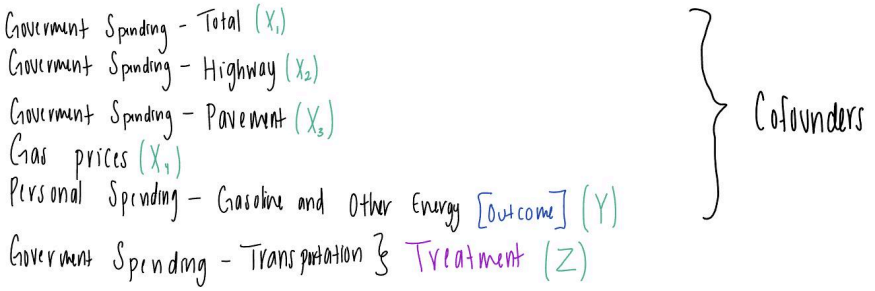
How does government spending on transportation affect personal spending on gas?

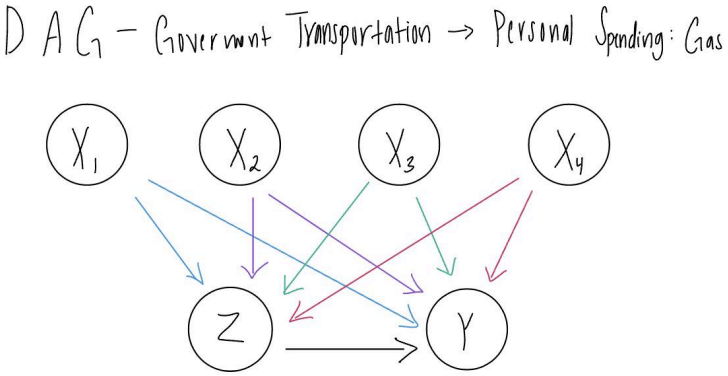
Methods

For our causal inference model, our outcome variable was Personal Spending on Transportation on Gas and Other Energy, and our treatment variable was Government Spending on Construction in Transportation. The confounders we chose were Total Government spending on Construction, Government Spending on Construction in Pavement, Price of Gasoline, and Government Spending on Construction on Highways and Streets. These confounders lead to a higher regression model accuracy when regressing the treatment on the outcome. They also lead to a shift in the causal effect of government spending on transportation on personal spending on gasoline and other energy. The unconfoundedness assumption doesn't hold due to our confounding variables.

Through conditional independence of the confounding variables, this can be mitigated. This was implemented by outcome regression using a linear model. Since our model had a very high R-Squared of 0.979 we believe this to be the best predictor of the causal effect of Government Spending on Transportation on Personal Spending on Gas and Other Energy.

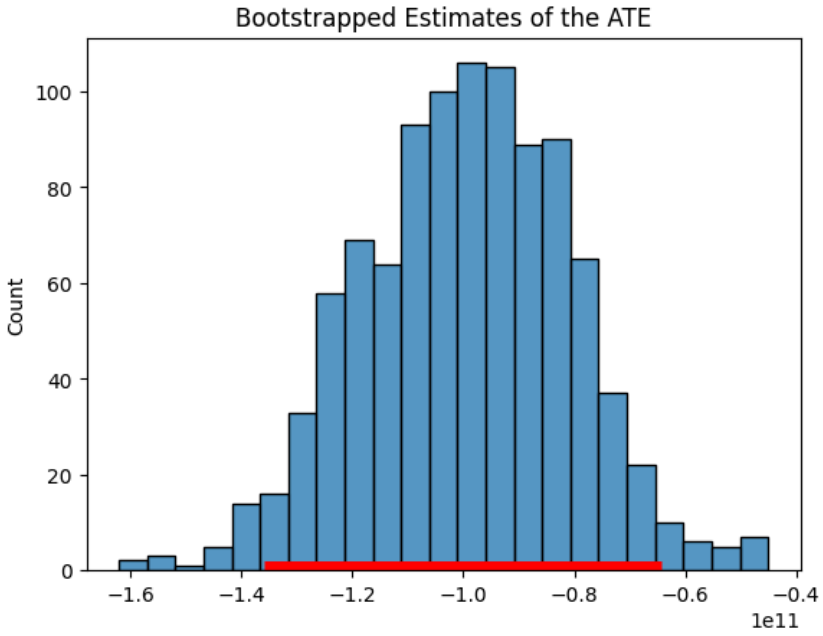
There are no colliders as none of the variables are affected by the treatment and outcome variables. While one could argue that gas price is a collider, we believe that gas price is by its supply rather than the demand for gas through personal spending. Gas conglomerates like OPEC control their supply to reach a desired demand and thus its price.



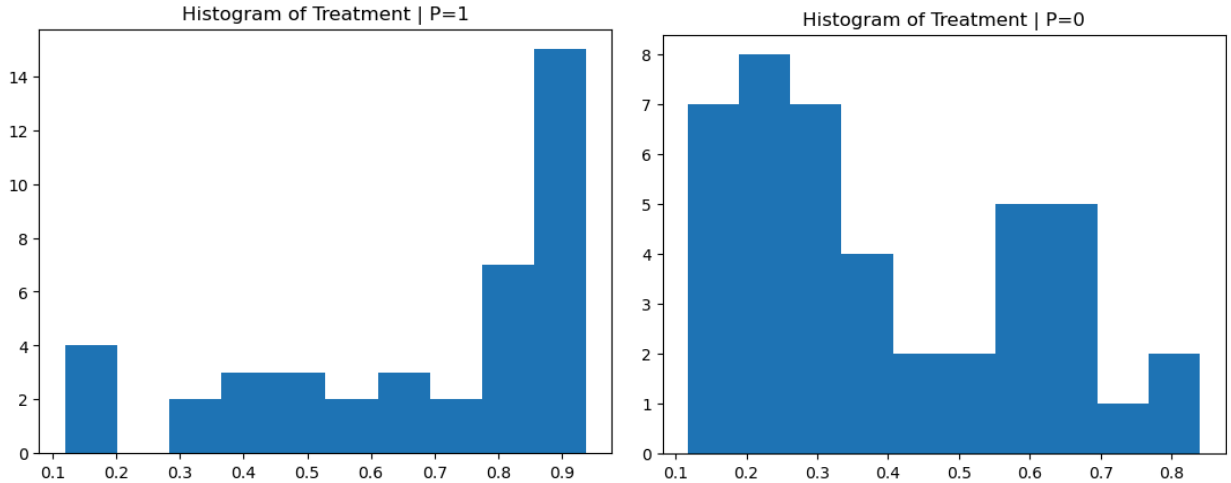


Results

The causal effect of High versus Low Government Spending on Construction on Transportation without confounders was $ -115 billion. However, when accounting for our confounders our causal effect was $-97.8 Billion. This causal effect was statistically significant as seen in the histogram including our 95 percent confidence interval for our Average Treatment effect.



We believe using a linear regression model works best as our propensity score weighting was not effective. This is because the distribution of propensity scores for treatment and control were not the same. This method gave an average treatment effect of $-48.7 Billion. The histogram of propensity scores on the left is for treatment and the right is for the control group.



Discussion

Some limitations of this analysis is that this data is of the whole US and the results might not be easy to generalize to local or state government budgets. Data for each state might give a better idea of the effect of government spending on personal spending. This is because generalizations about the whole United States are not useful as they can contain many different confounding variables from other government spending budgets that we can't model.

Additionally, consumer habits can change from year to year. Today we have more electric vehicles, which allow for people to spend less on energy goods. Having some data on the shift of consumer spending could help to give a better understanding of the effect of government spending on personal spending on gas and energy.

From the dataset, "government spending on transportation" is very vague. There is very little explanation as to how variables are defined and how the data is collected. It would be interesting to get an inside look into the analysis from the Bureau of Transportation Statistics.

While it makes sense that more government spending leads to lower personal spending on gas and other energy, there are a lot of factors that affect government and consumer spending. It is quite possible that our results are coincidental as there aren't many reliable ways to model the causal effect without extensive data collection. There are many confounding variables that we are most likely not accounting for despite our high model accuracy.

Question 2 (GLMs and nonparametric methods)

Does better public transportation imply lower levels of unemployment?

Methods

We are trying to predict the unemployment rate. The feature we are using is the log of government spending on mass transit. We chose government spending specifically on mass transit because we believe there would be a correlation between spending on mass transit and unemployment. This is because if the government makes mass transit more accessible through increased spending, it will help unemployed people take on jobs they previously couldn't reach due to lack of transportation.

We used a Normal distribution GLM as we believed a normal prior and linear regression model fit best for our GLMs. We believed the distribution of unemployment rate followed a roughly normal distribution based on our EDA. Also due to this normality assumption we used a linear regression model for a frequentist GLM model.

For our non-parametric methods, we are using both a decision tree and a random forest. The idea behind using both is to see if one is better than the other, or if the two models tell us different parts of the story. Nonparametric methods make no assumptions about the distributions of the parameters.

For the Bayesian GLM, we will run a posterior predictive check. For the Frequentist GLM, we will evaluate the model based on its R-Squared value. To evaluate the decision tree and random forest, we find the training set error and test set error by calculating the RMSE (root mean square error). We will also look at the uncertainty of the models by the confidence intervals and credible intervals of the slope and intercept of the model.

Results

From our Frequentist GLM, our regression found no correlation between government spending on mass transit and unemployment rate. This is because our confidence intervals for our coefficients include zero. For our Bayesian GLM model, we have a similar conclusion as our highest density intervals also include zero for our coefficients. The posterior-predictive check and R-squared showed that the models were not very good predictors.

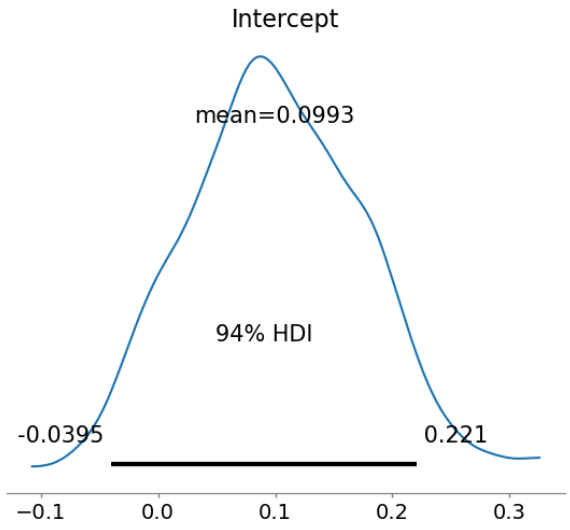
Additionally, our non-parametric models had very high RMSE for our data. Despite this, the non-parametric models were better as they made no assumptions on the data. Our random forest model performed better than our decision tree in our test set marginally, but our Decision tree was twice as good in the training set.

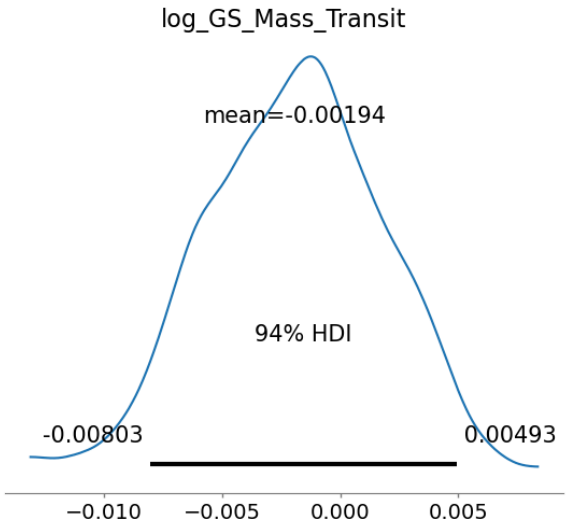
The non-parametric models were a better fit for our prediction as our data was not very normally distributed as under our assumptions. While, it may seem that we had incorrect assumptions, few posterior distributions would have matched the unemployment rate distribution. Thus if we eliminated any assumptions, we would likely have a better prediction.

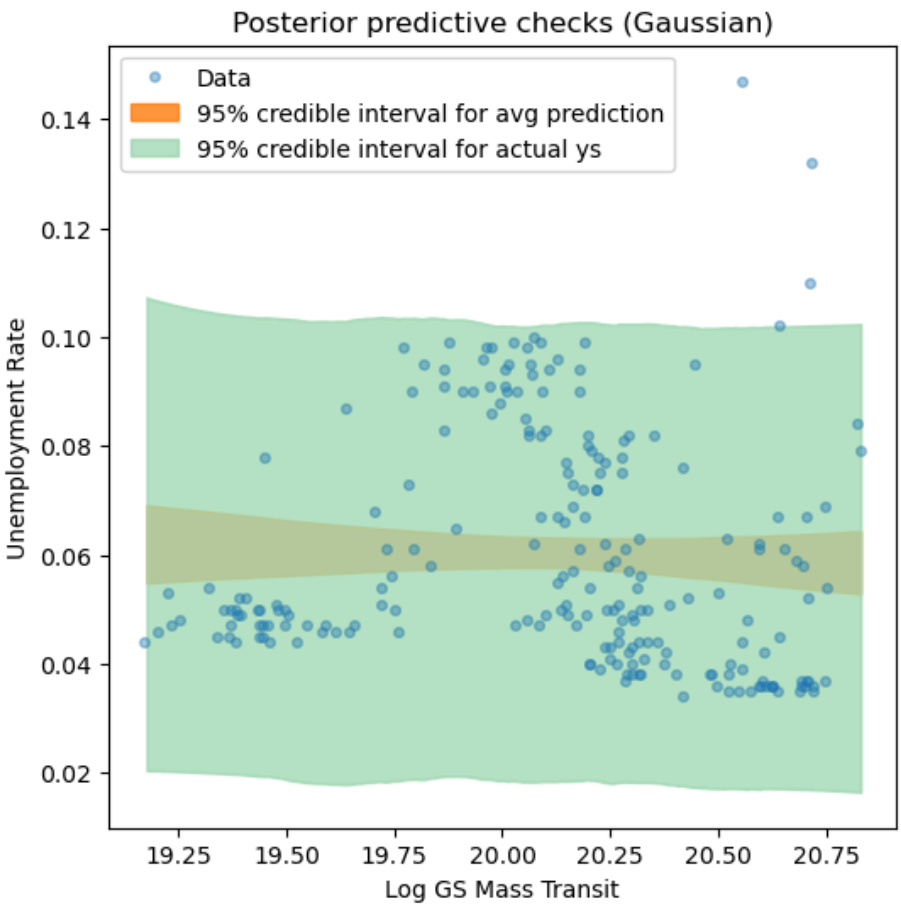
Based on the 94 percent Highest Density Intervals of our coefficients below we can't reject the null hypothesis that mass transit has an effect on unemployment rate. This is because our intervals include zero, thus they are not statistically significant. This also means that the models are very uncertain in their prediction.

Additionally, based on our posterior predictive check, we can see that our model shows very little correlation between unemployment and government spending on transportation. We can also see that our Bayesian GLM is not close to the true distribution of the unemployment rate. This may be because the distribution of the unemployment rate is not very normal, leading to a discrepancy between our model and the true distribution.

In our frequentist GLM, our intercept and slope confidence intervals were very large leading to a lot of uncertainty in the regression. This was also true for our non-parametric methods as our RMSE was high as well. Despite this, the non-parametric method of random forest had the least uncertainty since the error was lowest. Both our GLMs came to the same results for regression coefficients.







Discussion

The non-parametric method performed the best as it has the least amount of uncertainty. While the Frequentist and Bayesian methods came to the same conclusion, they had very high uncertainty in their predictions based on the confidence and credible intervals of coefficients.

However, none of the models fit the data very well because they all have high uncertainty. This is because mass transit spending has little correlation with unemployment to begin with. The implementation of Bayesian and Frequentist GLM were fairly similar, which is corroborated by their similarity in prediction of model coefficients. They both had similar levels of uncertainty as well.

If we were to create a new model, finding a dataset of people that use public transportation and their levels of unemployment would improve our uncertainty. By having a more specific population, it would be much easier to see the effect of government spending on mass transit on the unemployment rate.

Conclusion

For question one, as we looked at the models, we concluded that the best model was the linear regression model. The regression model accounted for the confounders, (government spending-total, government spending-highway, government spending- pavement, gas prices, and personal spending- gasoline and other energy). In addition, after modeling the propensity scores, we saw that they were not uniform. Overall, we had a good choice in confounders, therefore we had a good R^2.

For question two, since it was hard to find a good prior distribution, the nonparametric model turned out to be better. Despite this, our GLM model helped us see that there was no correlation between mass transit and unemployment. However, this could be due to other factors that we can't quantify well as seen in our data analysis which showed a slight correlation for certain data points in our scatterplot.

For the generalizability of our results, our dataset is a compilation of national statistics. Therefore, our findings only apply to the U.S. It is not very generalizable since the whole world's outcome for personal spending and unemployment could not be predicted by this. In addition, these results describe the U.S. as a whole. It cannot necessarily be applied to smaller regions, such as cities, accurately. (Large cities and rural areas may not have the same reality.)

Given the results of our first research question, a call to action for government agencies would be to consider spending more on transportation. Considering the fact that over the last several years we have increased our greenhouse gas emissions, it would be smart to increase spending on transportation to mitigate our effect on climate change. This way we can slow down our effects on the earth in order for us to make time for better solutions. Additionally, with less personal spending on gas, we have less cars on the road, and thus less overall traffic nationwide. This could lead to improved driver safety and decreased time spent on the road.

However, for our second research question, our results have shown us that there is no need to spend more on mass transit in hopes of decreasing unemployment. There is too much uncertainty in our results and too many confounders that can't be accounted for in this model. There may be better ways in which we can decrease unemployment. Thus, it might be better to invest our time and money in other factors that have a more direct correlation to unemployment.

Even with these results, we must consider the limitations of our data which is where and how the Bureau of Transportation collects data. It is aggregated from several smaller agencies and we cannot verify that it is all collected the same way. Another limitation is that the categories are quite vague. For example, we know there is a category for "Government Spending - Mass Transit" but we do not know specifically what is included in "Mass Transit" spending.

To delve deeper into our results from question one, new studies could look at the relationship between personal spending on gas or government transportation spending and pollution. We might also want to see how people change their transportation habits given pollution levels to see if pollution is a confounder. Both of these questions could be important to understanding how government spending on transportation affects personal spending on gas and greenhouse gas emissions.

Since our data for our second question is collected for the whole country, it is too broad and makes it hard to make any generalizations for more local government spending. We believe that our results will vary significantly from place to place due to people's varying living

conditions, habits, and many other factors. Future studies that compare government spending in mass transit versus unemployment rates in more specific locations and places where mass transit is more commonly used could help to give more conclusive results.

While working on this project, we learned that the results we concluded might be completely different from what we initially expected from our initially held assumptions. Although we went into the project thinking we had enough information, the data was still too general. We didn't have enough data on many factors that we wish we could have. We need datasets that are more inclusive and have more granularity to make stronger connections between variables. We also learned that choosing which model to use requires much more experimentation before finding the best fitting one. All in all, this project was a great learning experience in the importance of good data collection and proper model usage.