Written Report

Data Overview

The dataset used is called Monthly Transportation Statistics provided by the U.S. Department of Transportation, the Bureau of Transportation Statistics. It is a compilation of national statistics on transportation and contains over fifty time series from nearly two dozen data sources. The Bureau of Transportation Statistics brings together the latest data from across the Federal government and transportation industry. According to the official website of the Bureau of Transportation Statistics, the data are collected following the principles of the 'Guide to Good Statistical Practice in the Transportation Field[1], where the collection "instruments" are forms, questionnaires, automated collection screens, and file layouts used to collect the data from data suppliers. Based on the description above, the data is intended to be a complete enumeration, hence it is produced as a census. Participants are not being explicitly asked to report their transportation data so they are not aware of the collection of this data. There are not any groups that were systematically excluded from the data except that some data from the underdeveloped area might be excluded. In terms of granularity, the data set is broken down into months, where each row in the data represents a particular month in a given year. Since each row is monthly data, the overall dataset is not large so it might affect the findings. Furthermore, according to the Bureau of Transportation Statistics, measurement errors can still be present in the data but it is trying to minimize the possible data calculation and conversion mistakes. A feature that breaks the ridership according to the region might be helpful. For example, coastal areas compared to inland. The geographical location might affect ridership as well. This way, we can further break down the research questions into specific regions and then compare.

Research Questions

For this project, there are two research questions we wanted to answer using the dataset we have:

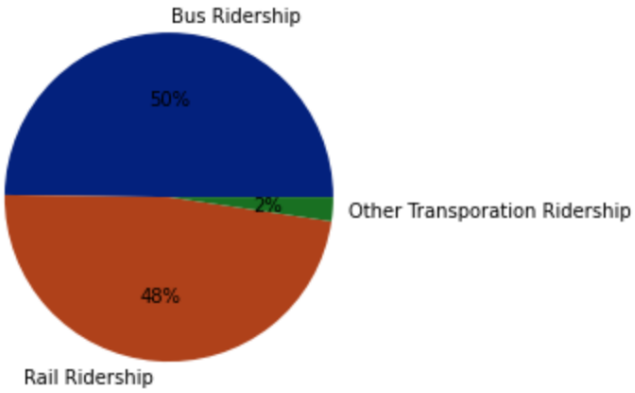
1. Did U.S. traffic volume decrease in response to the Covid-19 pandemic?
    - By answering this question, we can find out how human mobility, especially in transportation, has been impacted by the Covid-19 pandemic.
    - By using multiple hypothesis testing with the A/B test, we are able to decide whether the two numerical samples come from the same underlying distribution and whether it is statistically significant that one is lower than the other.
2. Is there a causal relationship between mass transit construction spending and mass transit utilization?
    - By answering this question, the state and government can decide whether to spend and how much they should spend on mass transit construction if there's a relationship between mass transit investment and utilization
    - By using causal inference, we are able to determine whether changes in mass transit construction spending cause changes in mass transit utilization.

EDA

Data cleaning:

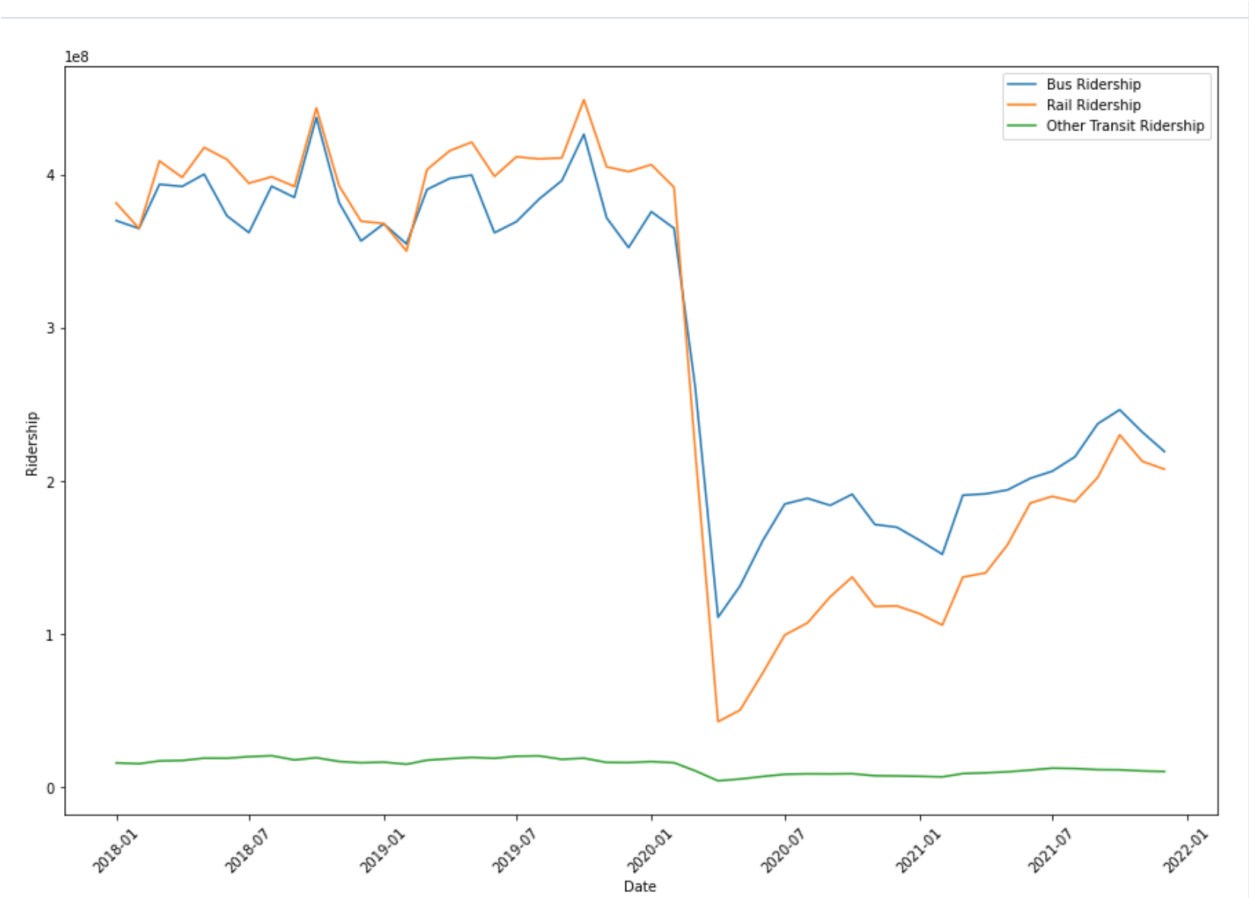
For question 1 we are comparing the transportation trend before and during the covid pandemic. We used International Airline Traffic, Domestic Airline Traffic, Highway Vehicle Miles Traveled, Bus Ridership, Rail Ridership, and others as the features. We select the years 2018, 2019, 2020, and 2021, and classify 2018, 2019, and 2020 from January to March as before the pandemic and the rest of the 2020 and 2021 as during the pandemic. We drop the Na of the data

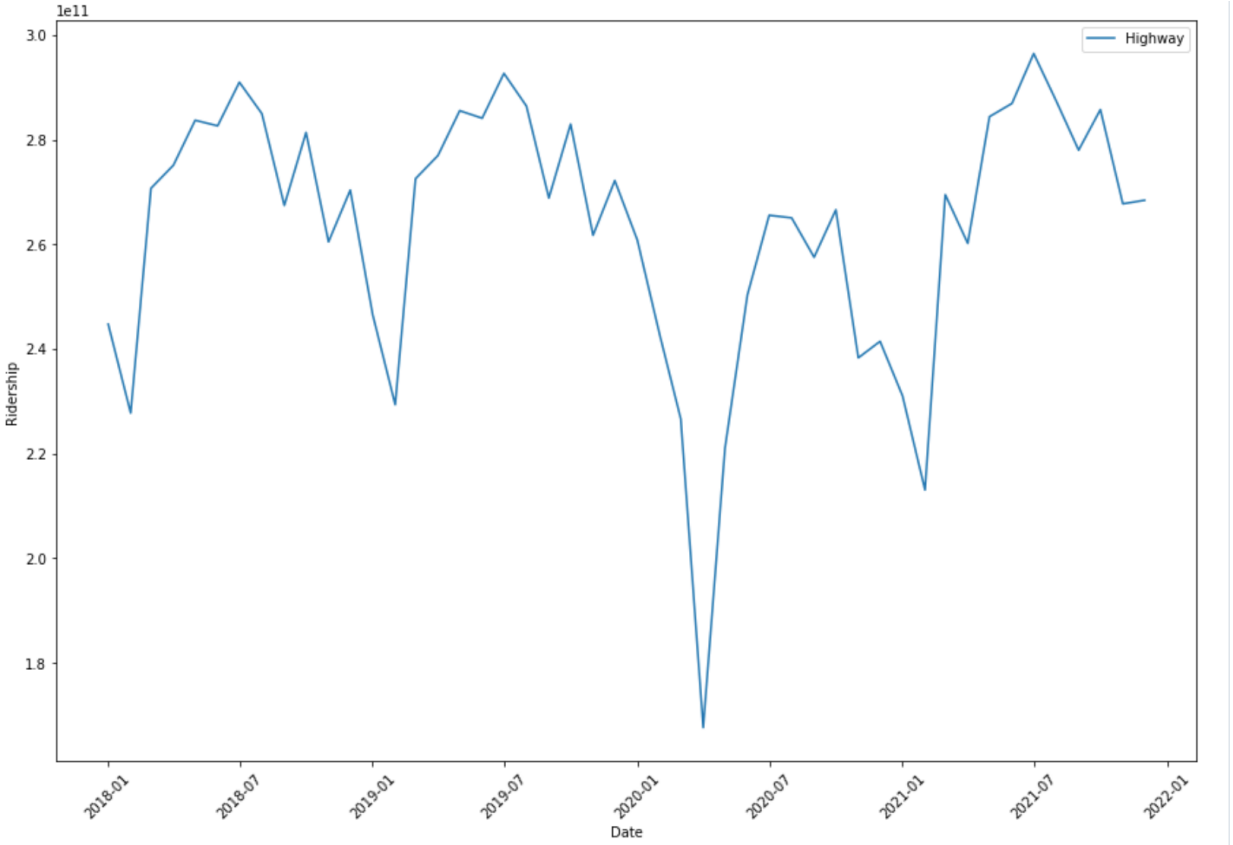
and it happens to have no effect on the data we need. For q2 in order to find out if there is a relationship between total ridership and spending on mass transit. In the model, State and Local Government Construction Spending - Mass Transit is the treatment; total ridership (which is the added value of three different ridership) is the outcome; State and Local Government Construction Spending - Highway and Street is the instrumental variable; Real Gross Domestic Product - Seasonally Adjusted is the confounding variable. These features are selected and then Na is dropped, which leaves with 80 rows. This could potentially affect the model since the data are significantly less. The EDA for question one includes one pie chart and three line plots.

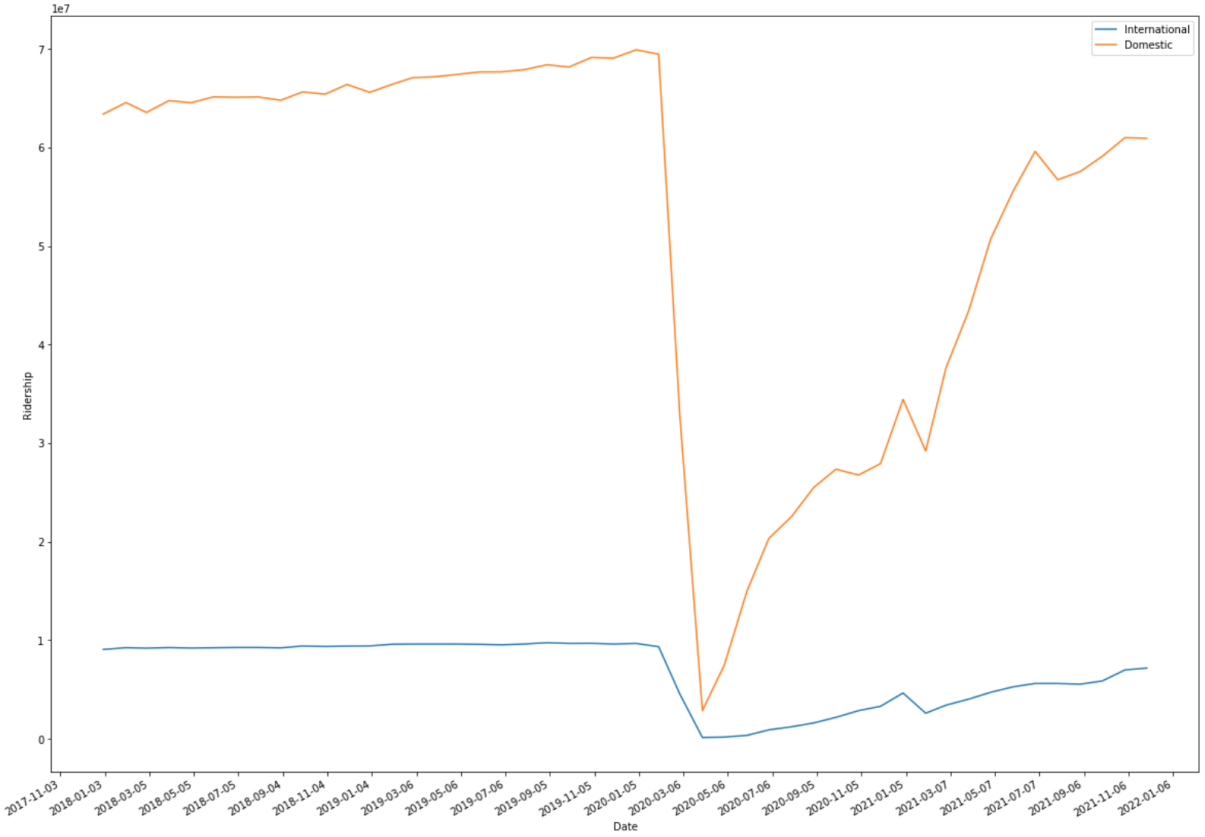


The pie chart shows the proportion of the different ridership. Here it shows that there is a rather even split between Bus Ridership and Rail Ridership in terms of percentage, and only a small portion (2%) of ridership is from other than bus and rail. Relationships that can be assumed is that perhaps Bus ridership is affected the most since it has the largest proportion.

The following line graphs all show a significant decrease in ridership when the pandemic happened.

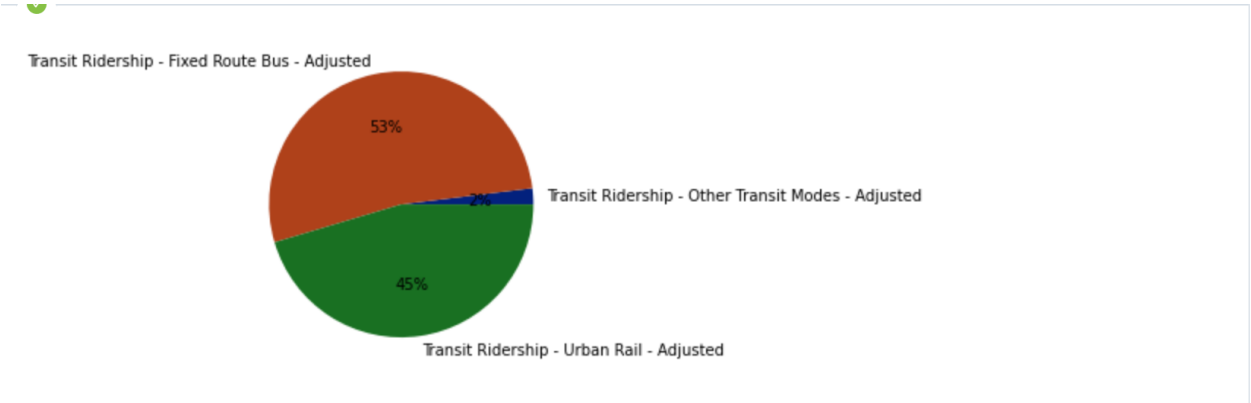



1e11
3.0
2.8
2.6
2.4
2.2
2.0
1.8
Ridership
Highway
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
Date

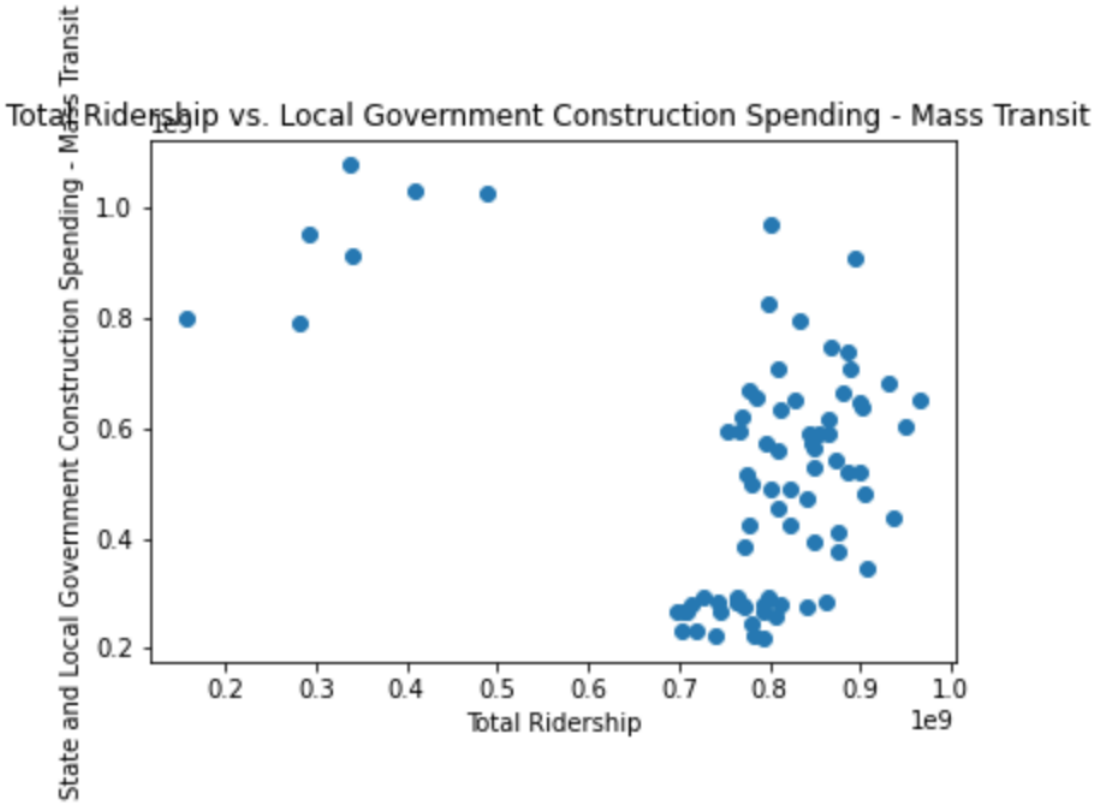

1e7
International
Domestic
Ridership
Date
0
1
2
3
4
5
6
7
2017-11-03
2018-01-03
2018-03-05
2018-05-05
2018-07-05
2018-09-04
2018-11-04
2019-01-04
2019-03-06
2019-05-06
2019-07-06
2019-09-05
2019-11-05
2020-01-05
2020-03-06
2020-05-06
2020-07-06
2020-09-05
2020-11-05
2021-01-05
2021-03-07
2021-05-07
2021-07-07
2021-09-06
2021-11-06
2022-01-06

Using the airline graph as an example, both international and domestic airline traffic had a sharp "dip" to 0 in April 2020. Domestic airline traffic thereafter demonstrated a generally increasing trend, eventually getting back to a similar level as before the pandemic started in January 2022. International airline traffic shows a similar trend, with less fluctuation in the process of getting back to the level before the pandemic started.

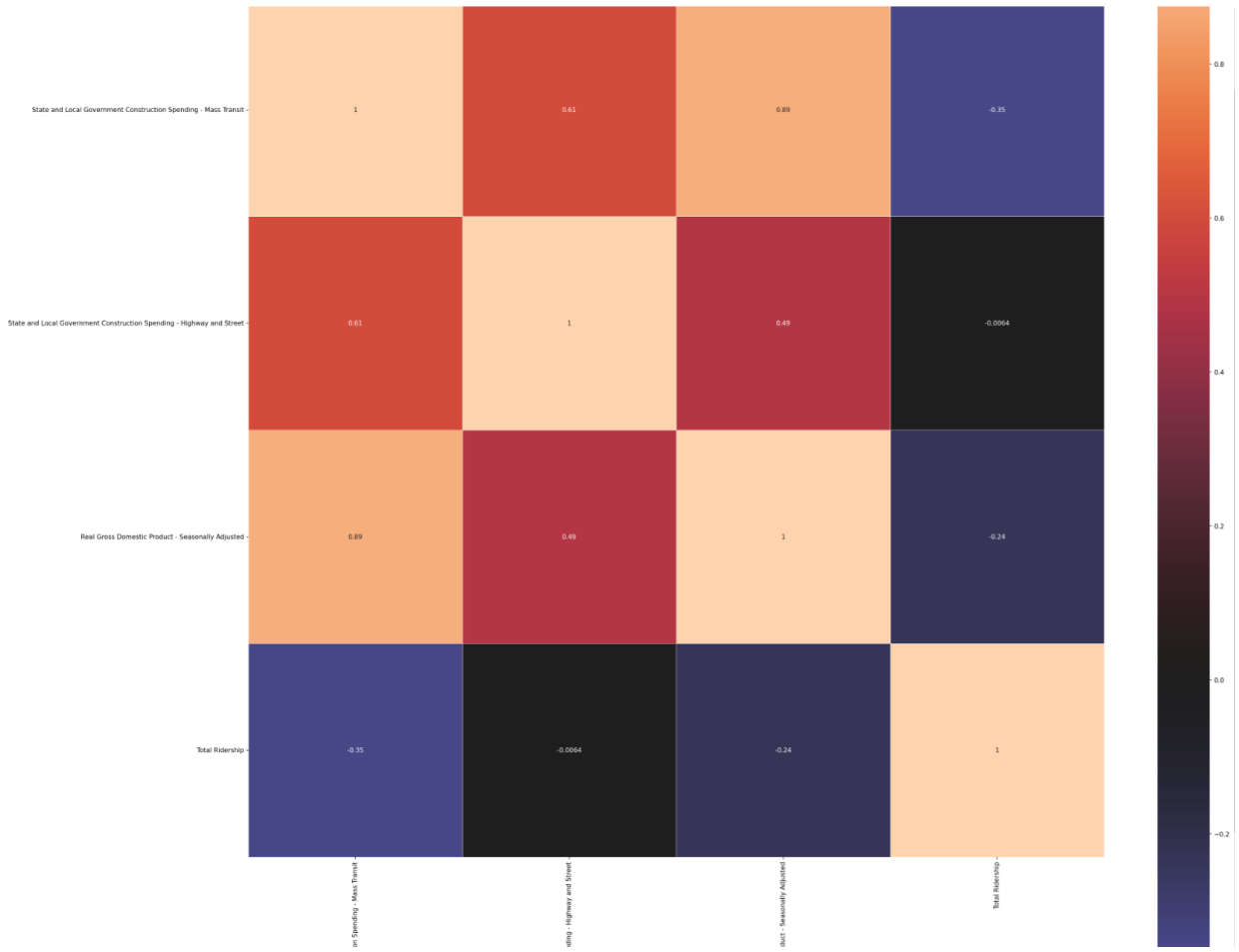
The EDA for question 2 includes a pie chart, scatter plot, and a correlation graph.



The research question is about finding out whether there is a relationship between total transit ridership and the "State and Local Government Construction Spending on Mass Transit ", and here the pie chart shows the proportion of each type of transit ridership. Ridership for buses is 53%; ridership for urban rail is 45% and ridership for other transit modes is 2%. Here transit ridership - fixed-route bus could be the most impactful since it has the largest proportion.



The scatter plot shows a weak relationship between total ridership (added value of each type of ridership) and government spending on mass transit, suggesting that total ridership might not have a strong direct relationship with the spending on mass transit, which is also shown in the later result.



This graph shows the correlation between each feature. We can see that Real GDP is highly correlated to government spending on mass transit and has a small negative correlation with ridership, which makes GDP a potential confounder. Different types of transit ridership are correlated with each other. Spending on highways and streets has 0.61% correlation with spending on mass transit, which could be a potential instrumental variable.

Multiple Hypothesis Testing/Decision Making

To answer the first research question, "Did U.S. traffic volume decrease in response to the COVID-19 pandemic?", we performed six hypothesis tests, each on different variables that measure different aspects of traffic volume, listed below:

1. International Airline Traffic
2. Domestic Airline Traffic
3. Highway Vehicle Miles Traveled
4. Rail Ridership
5. Bus Ridership
6. Other Transit Ridership (Demand response-taxi, vanpool, and ferryboat)

Methods:

Traffic volume is a broad term and can be measured in different aspects, so we broke it down into six relevant traffic variables to better understand the change in traffic volume. From the trends graphed for each of the variables above, we noticed all of the variables have a sharp drop after the pandemic starts. Therefore, we wanted to test if it is statistically significant that the traffic volume after the pandemic is lower than that before the pandemic. The null and alternative hypotheses are the same for the six variables above.

- Null hypothesis: There is no difference in the average value of the traffic volume variable before and after the Covid-19 pandemic.

- Alternative hypothesis: The traffic variable has a lower volume after the pandemic, on average, than that before the pandemic.

For each of the six variables, the test statistics is the difference between the means of the two groups (pre-pandemic as 0 and post-pandemic as 1). Small values (that is, large negative values) of this statistic will favor the alternative hypothesis. An A/B Test was performed by shuffling the labels and simulating the test statistic under the null hypothesis 1000 times. This way, we are able to see how the test statistic should vary under the null hypothesis. Then, the empirical p-value is computed based on the observed statistics and the predicted behavior of the statistic under the null hypothesis.

Results:

- The p-value for International Airline Traffic is 0.
- The p-value for Domestic Airline Traffic is 0.
- The p-value for Highway Vehicle Miles Traveled is ≈ 0.028.
- The p-value for Other Transit Ridership is 0.
- The p-value for Rail Ridership is 0.
- The p-value for Bus Ridership is 0.

Under the naive p-value threshold of 0.05, we reject all of the null hypotheses since all of the p-values are smaller than the threshold. This means that it is statistically significant to conclude that the volume of international airline traffic, domestic airline traffic, highway vehicles traffic, rail ridership, bus ridership, and other ridership is lower than those after the pandemic.

Bonferroni Correction:

Using Bonferroni Correction, we want to control the Family Wise Error Rate at 0.05, so all the p-values are thresholded at $0.05/6 \approx 0.008$. In this case, we fail to reject only one hypothesis. That is, there is no significant evidence to conclude that the "highway vehicle miles traveled" decreased after the pandemic. This was implied by the trends graphed in the EDA. Highway vehicle miles traveled decreased sharply in March 2020 and bounced back to their original state shortly.

Benjamini-Hochberg Correction:

Under the Benjamini-Hochberg procedure, all the null hypotheses are rejected since all the sorted p-values are less than or equal to $\frac{0.05 \times k}{6}$. It suggests that the volume of international airline traffic, domestic airline traffic, highway vehicle, rail ridership, bus ridership, and other transit ridership all decreased after the pandemic. This method controls the False Discovery Rate at 0.05.

Discussion:

When looking at the individual tests, since the values for 5 of the 6 tests are 0.0, and the only nonzero 0.028, it can be safely predicted that at least five of the tests have small p-values, which means the number of people using the transit significantly decrease due to the pandemic and the real world decisions that can be made is that the transportation company must act to reduce their loss due to the pandemic. For example, they can minimize their loss by reducing the number of employees or ask for funds from the government. When looking at the results from the aggregate form, when a similar pandemic happens, the transportation company and the government can react faster and come up with better strategies to deal with the dilemma.

One limitation is that since we only use for years in the data, which only have 48 months, we have a small dataset. A dataset that has a small number of data might make the result less accurate. To avoid p-value hacking, Benjamini-Hochberg Correction and Bonferroni Correction were used other than the naive threshold.

If more data is provided, tests on miles traveled by the international cargo ships can be conducted, thus the effect of the pandemic on cargo ships can be observed with some real world meaning.

Causal Inference

Causal question:

Is there a relationship between mass transit investment and mass transit utilization?

Variables:

Z = treatment = State and Local Government Construction Spending - Mass Transit *(dollar)*

W = instrumental = State and Local Government Construction Spending - Highway and Street *(dollar)*

Y = outcome = Transit Ridership - Other Transit Modes - Adjusted + Transit Ridership - Fixed Route Bus - Adjusted + Transit Ridership - Urban Rail - Adjusted *(number of transportation vehicles dispatched)*

X = confounder = Real Gross Domestic Product - Seasonally Adjusted *(dollar)*

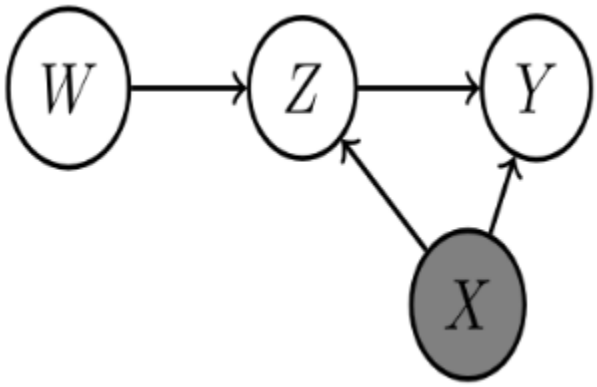
Methods:

To answer this question, two Ordinary Least Square Estimator models were performed using the variables identified above. The model assumes that the outcome is generated as a linear function of the confounder X (the Real GDP) and the treatment Z (Mass Transit Spending), with additive noise epsilon:

$$Y = \beta_1 Z + \beta_2 X + \epsilon.$$

The goal is to estimate the first coefficient, beta_1, the true causal effect of government mass transit spending on mass transit ridership. The result would then provide insight into the relationship between the treatment variable and the outcome variable.

The first model is a Two-Stage Ordinary Least Square Regression. To account for the potential bias due to confounder X highly correlated with Z, the model uses the instrumental variable W (Highway/Street Spending) to first estimate Z using OLS (denoted Z_hat). Then using Z_hat to regress on the outcome Y, instead of using Z directly.



In the second model, a One-Stage Ordinary Least Square Regression is performed to estimate the coefficient beta_1 for outcome Y (mass transit ridership) using Z (treatment) and X (confounder) directly.

In both models, the unconfoundedness assumption does not hold because not all confounding variables are observed and being accounted for. There could be various factors in the real world affecting treatment Z and outcome Y other than Real GDP. For instance, extreme weather due to climate change could affect how much the government spends on mass transit and how often people go out, but it is difficult to quantify that via data. Additionally, there are no colliders in the dataset.

Results:

Model 1: Using mass transit spending to predict total ridership (2SLS)

In stage 1, we fitted the OLS parameters to predict treatment Z (mass transit spending) using the instrumental variable W (highway spending). The coefficient for highway spending is 0.0571, which denotes a weak but still present casual relationship between highway spending and mass transit spending. The limitation of highway spending as an instrumental variable will be discussed in more detail in the discussion section.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Transit Spending   R-squared:                       0.368
Model:                           OLS   Adj. R-squared:                  0.360
Method:                Least Squares   F-statistic:                     45.41
Date:               Sun, 08 May 2022   Prob (F-statistic):           2.46e-09
Time:                       05:00:25   Log-Likelihood:                -1632.9
No. Observations:                 80   AIC:                             3270.
Df Residuals:                     78   BIC:                             3275.
Df Model:                          1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.374e+08   5.97e+07      2.302      0.024    1.86e+07    2.56e+08
Highway Spending     0.0571      0.008      6.739      0.000       0.040       0.074
==============================================================================
Omnibus:                        5.554   Durbin-Watson:                   0.494
Prob(Omnibus):                  0.062   Jarque-Bera (JB):                4.489
Skew:                           0.472   Prob(JB):                        0.106
Kurtosis:                       2.325   Cond. No.                     2.09e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.09e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(Model 1: First stage OLS result)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Total Ridership   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.013
Method:                 Least Squares   F-statistic:                  0.003151
Date:                Sun, 08 May 2022   Prob (F-statistic):              0.955
Time:                        05:00:25   Log-Likelihood:                -1621.1
No. Observations:                  80   AIC:                             3246.
Df Residuals:                      78   BIC:                             3251.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                      7.782e+08   6.83e+07     11.395      0.000    6.42e+08    9.14e+08
PredictedTransitSpending    -0.0072      0.128     -0.056      0.955      -0.262       0.248
==============================================================================
Omnibus:                       52.760   Durbin-Watson:                   0.466
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              160.072
Skew:                          -2.287   Prob(JB):                     1.74e-35
Kurtosis:                       8.205   Cond. No.                     2.11e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.11e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(Model 1: Second stage OLS result)

In the second stage of the 2SLS, the coefficient on Z_hat (predicted transit spending) that was produced in the first stage is -0.0072. The expectation was that higher mass transit spending will result in more total ridership. However, the 2SLS result shows the opposite. Since the coefficient is negative, when the transit spending increases by 1 unit of total dollar value, there will on average be a 0.0072 decrease in the number of vehicles dispatched.

Model 2: Using real GDP and mass transit spending to regress on total ridership (1SLS)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Total Ridership   R-squared:                       0.146
Model:                            OLS   Adj. R-squared:                  0.124
Method:                 Least Squares   F-statistic:                     6.600
Date:                Sat, 07 May 2022   Prob (F-statistic):            0.00226
Time:                        06:21:24   Log-Likelihood:                -1614.8
No. Observations:                  80   AIC:                             3236.
Df Residuals:                      77   BIC:                             3243.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const             5.16e+08   2.68e+08      1.923      0.058   -1.82e+07    1.05e+09
Real GDP         2.955e-05   2.05e-05      1.443      0.153   -1.12e-05    7.03e-05
Transit Spending    -0.4390      0.155     -2.829      0.006      -0.748      -0.130
==============================================================================
Omnibus:                       23.818   Durbin-Watson:                   0.744
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               37.102
Skew:                          -1.209   Prob(JB):                     8.78e-09
Kurtosis:                       5.300   Cond. No.                     2.75e+14
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.75e+14. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(Model 2: 1SLS result)

In the 1SLS, the coefficient for real GDP is positive but extremely small. This indicates that holding the transit spending constant, an increase in real GDP will result in a slight increase in the number of transportation vehicles dispatched. On the other hand, the coefficient of transit spending is -0.4390. This indicates that holding the effect of real GDP constant, an increase in one unit of total dollar value in transit spending will result in a 0.4390 decrease in the number of vehicles dispatched on the highway.

In both the 2SLS model and 1SLS model, the coefficient of transit spending is negative when used to predict total ridership. From these results, we conclude that as transit spending increases, it will result in a slight decrease in total ridership. Therefore, there is a relationship between mass transit investment and mass transit utilization despite the unexpected negative relationship from the findings.

Discussion:

In the 2SLS, highway spending was used as an instrumental variable with the assumption that it has a causal effect on transit spending. However, based on the result of the first stage OLS, the coefficient of highway spending when being regressed by transit spending is only 0.0571. This coefficient is relatively small, hence the suitability of highway spending as an instrumental variable is also limited. This limitation does not necessarily indicate that the result of the OLS regression is highly unreliable. Instead, it signifies that further exploration of the causal relationship between highway and transit spending should be done to provide stronger evidence for the causal relationship between transit spending and total ridership.

Another limitation of both models is that the OLS regressions are run based on the assumption that there are no variables other than transit spending. More specifically in the 2SLS we assume that even if there exist confounding variables, their effects are controlled by highway spending as the instrumental variable. This might not be the most logical assumption because there are confounders such as environmental and geographical conditions that can affect total ridership not only through their impacts on highway spending if any. For example, the frequency of natural disasters could directly impact the number of vehicles dispatched hence the total ridership due to

the danger of traveling during the time of the disaster. Effects of such confounders are not controlled by the instrumental variable highway spending but are also not taken into account when predicting the total ridership in both of the models.

Additional data that may be useful for answering this causal question could be that instead of measuring mass transit ridership in the number of transportation vehicles dispatched, it is measured in the number of people riding mass transit each month. Such a change in the unit could significantly improve the interpretability of the model. If the data provide more economic indicators, that could be useful as well. For example, the monthly average of the S&P 500 index, could provide a more in-depth insight into the U.S. economy, and how that affects government spending and citizen transportation.

Given that both of the models employed suggest a small negative correlation between the treatment variable and outcome variable (-0.0072 & -0.4390), we may conclude that there's a causal relationship between the chosen treatment and outcome. However, due to the relatively small coefficients, we are not fully confident that a causal relationship holds in the real world.

Conclusions:

For the first research question, we can conclude that it is statistically significant that transportation traffic after the pandemic is lower than transportation before the pandemic. The result is only generalizable to the U.S., so it might not apply to the other regions of the world. The finding is broad since it includes basically all the transportation types people use: airlines, vehicles, buses, rail, demand response-taxi, vanpool, and ferryboat. Based on the results, we see

that the pandemic does have a huge impact on human transportation. It is reasonable to assume that related public transit or transportation companies might have suffered a loss after the pandemic started. Future research and studies could further investigate this matter as well as how human mobility changed in different regions in the United States. Furthermore, studies can compare how human mobility differs after the pandemic by categorizing variables such as education level, income level, age, etc.

In the second research question, the finding suggests that there is a negative causal relationship between the amount of government spending on mass transit and the utilization of mass transit. The causal relationship is not very strong because of the assumption that all other confounding variables have been controlled for. In addition, such a finding could be relatively narrow because we combined the different categories of mass transit ridership into one column by taking the total sum. Also, we know that mass transit spending is particularly referring to the dollar value of construction work done on mass transit. We are essentially narrowing it down to the spending and utilization of mass transit systems in the United States, hence the result and findings do not apply to non-U.S. territories.

Based on the results from this causal inference analysis, one prominent real-world decision might be to allocate government resources to different sectors/areas of transportation if the nation wants to increase mass transit ridership. Increasing the investment in mass transit system construction would not lead to more citizens riding mass transit. Such a shift in spending policy could have a significant impact on climate change, because if the U.S. government focuses its resources on other aspects of transportation, more people might start riding public transportation, thus greatly

reducing the national carbon footprint. A future study that could be built upon this research is the study of how the government can find the most efficient way to allocate its spending that minimizes the cost of construction while maximizing the mass transit ridership.

Reference

1. https://www.bts.gov/archive/publications/guide_to_good_statistical_practice_in_the_transportation_field/chapter_03#:~:text=Principles,used%20to%20collect%20the%20data.