Group: Delilah Catron, Dea Bardhoshi, Derek Punaro, Quoc Huynh

Data 102 Final Project Report

1. Data Overview

We are using the provided US Bureau of Transportation Statistics: **Monthly Transportation Statistics** dataset and **Google Daily Community Mobility dataset.**

The US Bureau of Transport dataset contains data relevant to transportation spending, use, and safety here in the United States, with some variables containing data going back to 1947. This dataset contains over 50 time series as generated from census data collected from about two dozen data sources, those primarily being governmental agencies. Because some of the data is more historic than others, and relevant transportation variables have dramatically evolved since the last century, not every time series is equally complete.

It could be argued that data on some modes of transportation and any respective governmental spending towards them in the US such as rideshare vehicles, bicycles, motor scooters or more are being excluded. However, since the dataset simply reports the data/statistics that different agencies deem relevant to collect, and can accurately collect, it doesn't appear any groups were systematically excluded from the dataset. Further, the US Bureau of Transport invites anyone to give them feedback about what other statistics they would like to see included in the dataset, there is always potential for further groups to be included if ever deemed relevant.

Most of the data "participants", in this case primarily meaning users of US transportation services and governmental agencies, would not have necessarily been aware of the use and collection of this data. However, since each of the statistics is an aggregate per month, i.e. monthly amount of governmental spending in a specific category, or number of miles traveled on highways, their data could be considered to be anonymized for any one individual. So, the granularity of the data is monthly, with each row representing different months of a year. This level of detail beneficially impacts the interpretation of our findings in that we can more accurately understand how time would affect variables.

Neither selection bias, measurement error, nor convenience sampling seem to be concerns in this dataset because of the census reporting format of the data generated. However, because the dataset is so large and has so many sources and features, accuracy of the data at all times and over time could be a potential concern. For the analysis we did with the dataset, the only features we wish we would've had were more features we could've used as confounding variables for our causal inference.

The Google Daily Community Mobility dataset represents samples of location data from the population across the world generated by aggregated, anonymized sets of data from Google account users who have turned on their Location History setting. The data is at an international scale, and is conveniently broken up by country, and then where applicable and possible, regions and subregions of the country. Specifically, the data reports how visits and length of stay at different categories of places change compared to baseline values right before the pandemic from January-February 2020. The specific categories include "Retail & recreation", "Grocery & Pharmacy", "Parks", "Transit stations", "Workplaces", and "Residential".

The group that would be considered systematically excluded from the worldwide dataset would include people who don't possess phones/Google accounts/cellular data, because they wouldn't have a choice to include their data in the dataset. Participants are considered to be aware of the collection of the data because their Location History setting is automatically turned off by default, so in turning it on, the user would have been made aware that Google services would be collecting the data. However, it is unclear whether or not the user would have been made explicitly aware of the use of this data for this particular data analysis, because oftentimes privacy and data usage notices aren't very transparent or are difficult to understand. Google claims to use differential privacy precautions on the dataset, and says users are allowed to delete their movement data anytime[1].

The data as presented by Google was the daily percent change from the baseline in each particular category in a region, which first began at the beginning of the COVID pandemic in February 2020, and has since stopped as of October 15, 2022. So, each row represents how mobility trends on that particular day changed from the baseline value for each of the place categories. This beneficially impacts our interpretation of the data because we are able to look at how each variable changed over time, in a high degree of detail.

The primary concern for the data would be selection bias, because this data is only created by a sample of the population of users with phones/Google accounts/cellular data, who have their Location History setting on. So, people who have access to these things, and are willing to accept the Location History setting, may have different movements as compared to those who do not. As such, selection bias in the data may result in inaccurate data generation. A secondary concern is the accuracy generation of the baseline comparison data for the region. Google includes specific questions about whether local events and seasonal changes might have biased the baseline. Because if the baseline was biased to begin with, the rest of the percent difference data wouldn't be accurate and wouldn't make sense.

There are no other features or columns we wish we had or were unavailable.

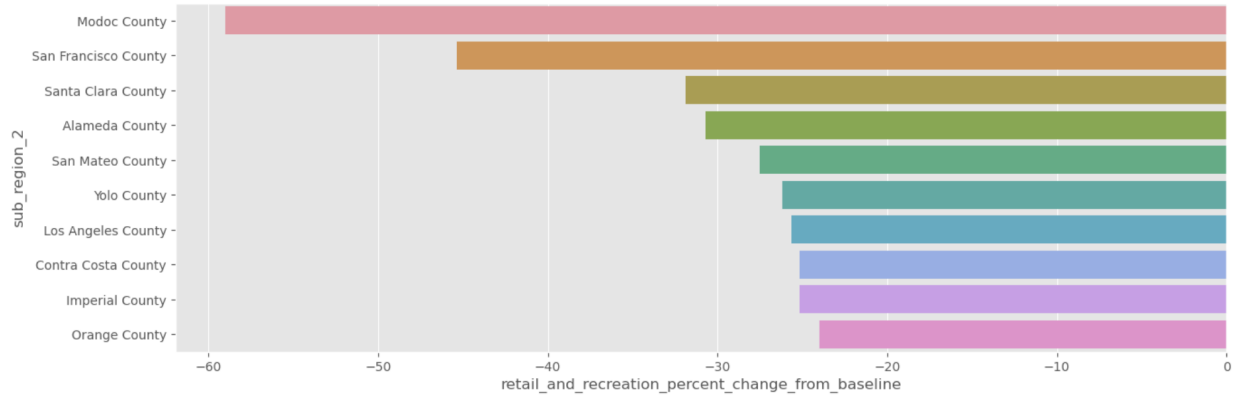
2. EDA

Question 1- Multiple Hypothesis Testing Method

How did COVID affect transportation in urban versus rural counties in California differently?

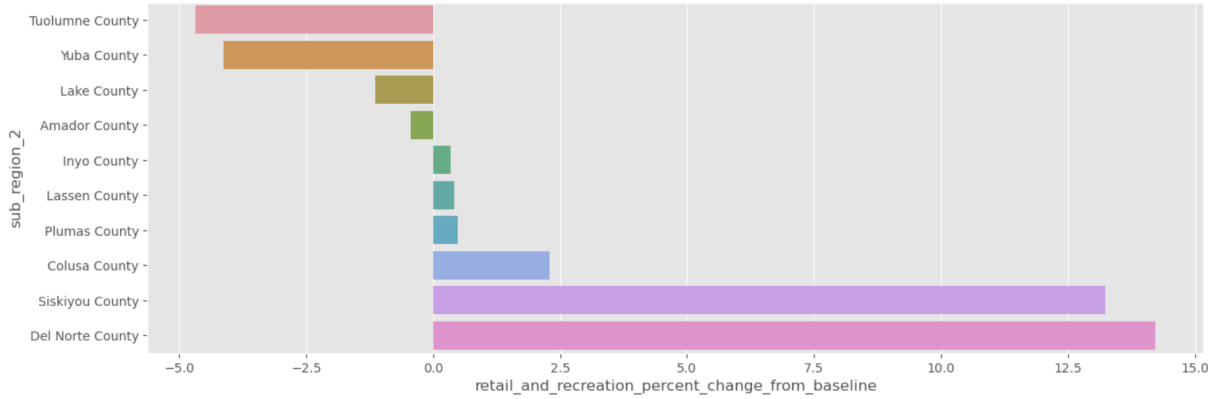
[1] Google.com, 2020, www.google.com/covid19/mobility/data_documentation.html?hl=en.

Change in "Retail and Recreation" Mobility Type: top 10 counties (most negative change)



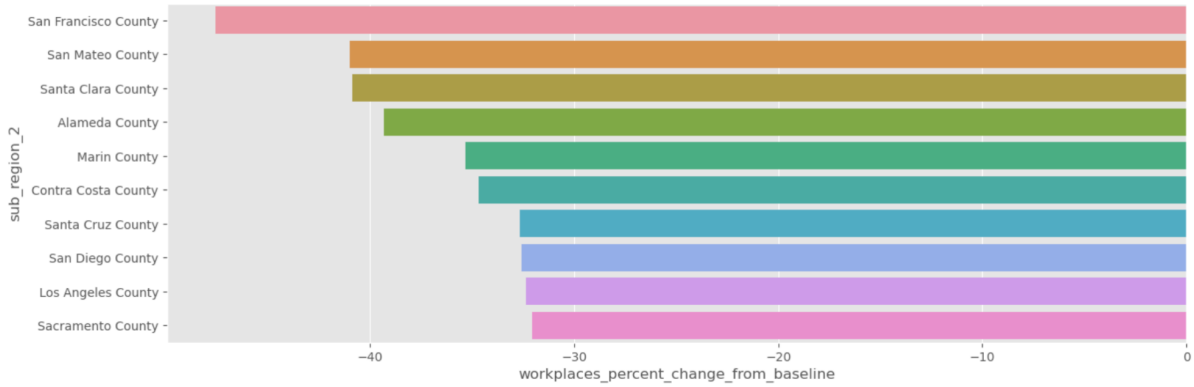
In this barplot we notice that the county with the highest mean change from baseline is Modoc county followed by San Francisco, Alameda and Santa Clara. The latter three are geographically adjacent to each-other which can explain the similarity in the values, whereas Modoc county has a very small population (3rd smallest in California). The majority of the counties in the plot have a very large population however, which is interesting in our analysis of question 1 as well (since we are trying to determine whether urban and rural counties' mobility was affected differently from the pandemic.) So, this visualization helps to motivate the question.

Change in "Retail and Recreation" Mobility Type: top 10 counties (most positive change)



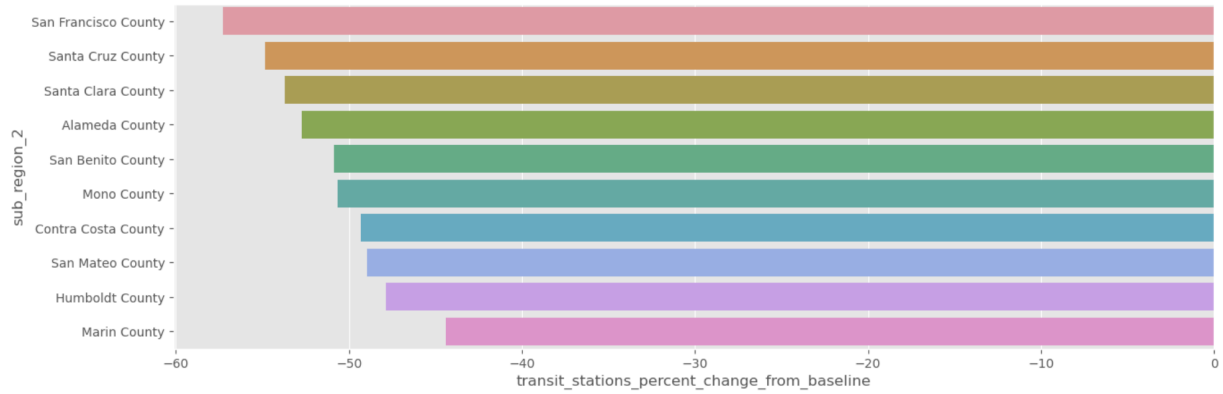
In this plot, we notice that a few of the counties' retail and recreation mobility category average actually increased, and Del Norte, Siskiyou and Colusa have small populations, but are actually closer to natural parks (Redwood State Park, Shasta-Trinity State Forest and Colusa National Refuge respectively) which can explain why perhaps more tourists visited them during the pandemic. This is relevant to our research question because we are trying to determine whether urban and rural counties' mobility was affected differently from the pandemic. So, this visualization helps to motivate the question.

Change in "Workplaces" Mobility Type: top 10 counties (most negative change)



In this visualization, we notice that the counties with the most decrease in the mean workplace mobility were San Francisco, San Mateo, Santa Clara, Marin and Contra Costa (all in the Bay Area) followed by the San Diego and Los Angeles areas. This is informative for our analysis as it shows that highly urbanized areas saw a great reduction in workplace mobility. So, this visualization helps to motivate the question.

Change in "Transit Stations" Mobility Type: top 10 counties (most negative change)



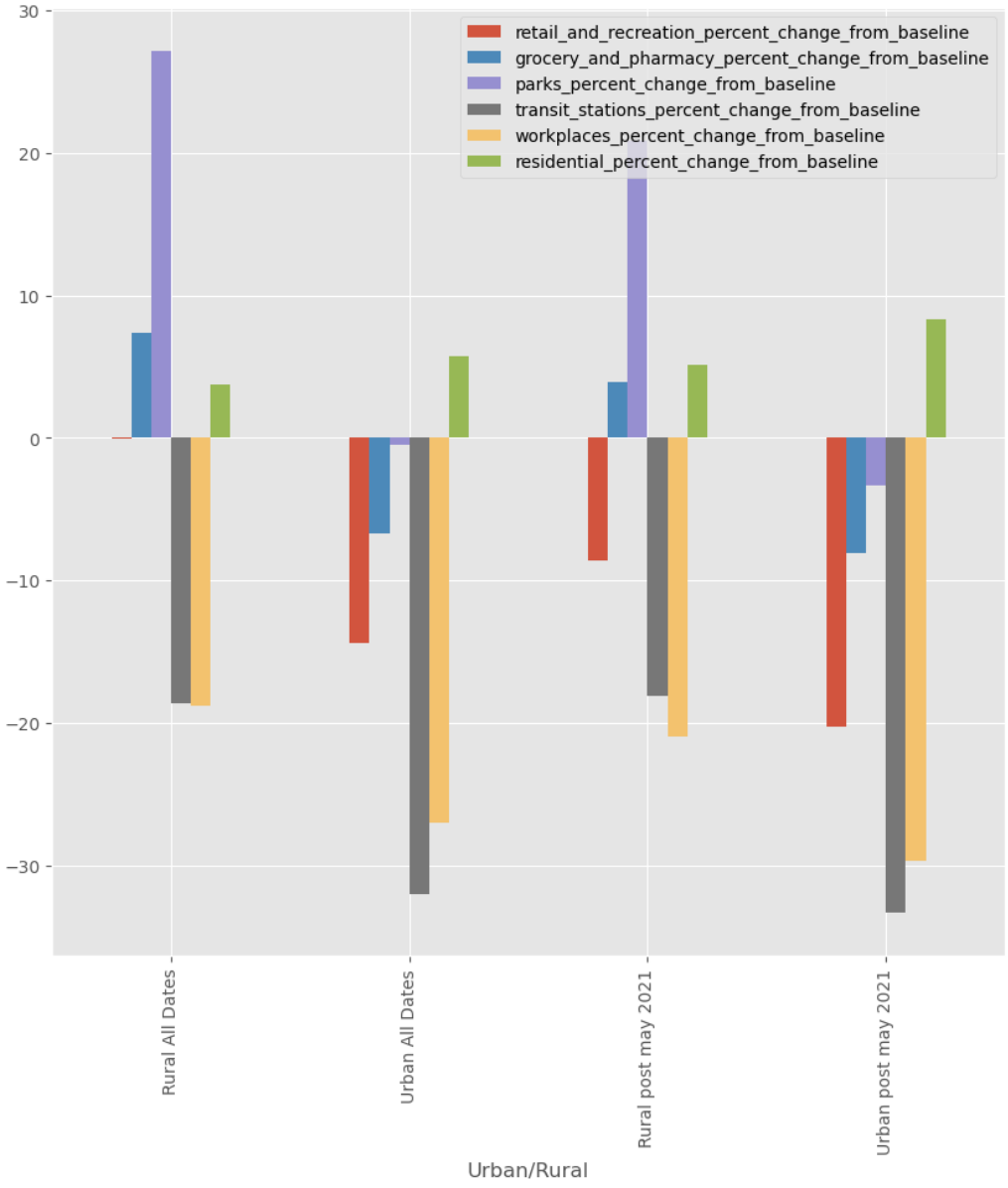
We notice that Bay Area counties seem to have the largest negative changes in transit station mobility, which can presumably be linked to less work-related mobility in general. Again, this is consistent with our question's aim of testing whether places like the Bay Area were impacted by COVID-19 differently compared to other less populated and rural areas in California. So, this visualization helps to motivate the question.

Rural vs Urban Counties Mobility Comparisons

Using the website (https://totalescape.com/destinations/list-of-rural-counties-in-california/) we are categorizing California Counties as urban versus rural. Rural counties are defined to be the following:

- Alpine Mariposa Sierra Trinity Amador Calaveras Inyo Lassen Modoc Mono Plumas Siskiyou Butte Colusa Del Norte El Dorado Glenn Humboldt Kern Lake Mendocino Nevada Placer San Luis Obispo Shasta Tehama Tulare Tuolumne Yolo Yuba.

Rural v Urban Counties Mobility Comparison: Post-May 2021 versus all dates



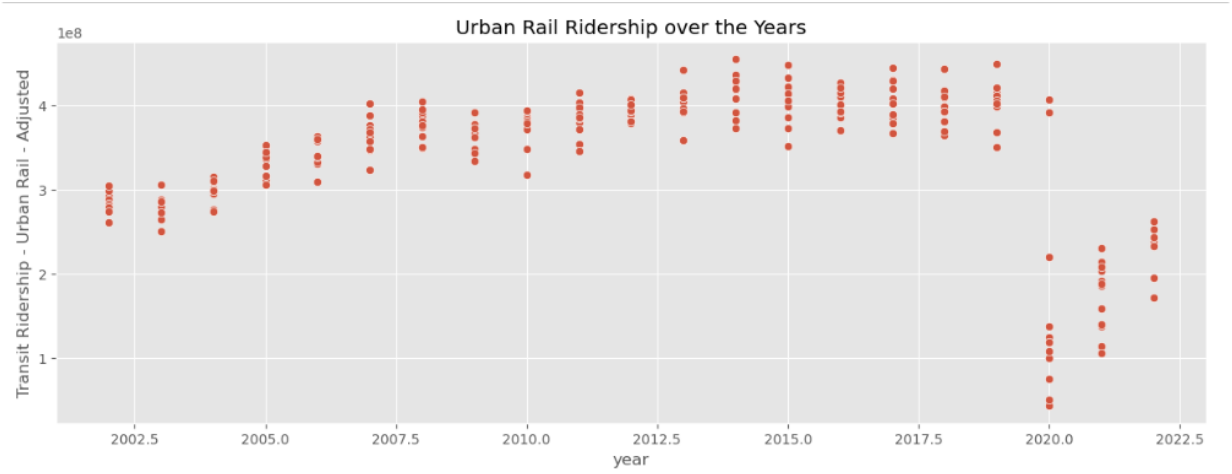
We were also interested in looking at how choosing differing time cutoffs versus looking at all the data would change our observations of mobility. Above we can see mobility metrics for rural/urban counties divided by “post-May 2021” and “all dates”. There are some differences in urban counties: retail movement mean decrease is less pronounced post May 2021, and residential mean change seems to have increased in the post May 2021 period. In rural counties, there is less of an increase in “parks” mobility

compared post May 2021 compared to overall dates. So, this visualization helps to suggest a potential answer.

Question 2- Causal Inference Testing Method

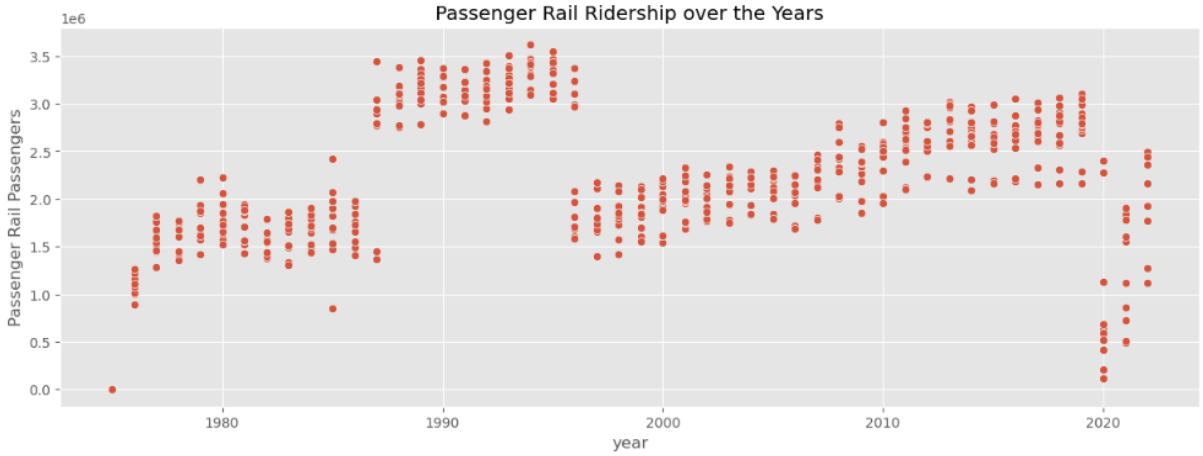
What is the effect of government spending, specifically, infrastructure spending on road safety?

Urban Rail Ridership Confounding Variable



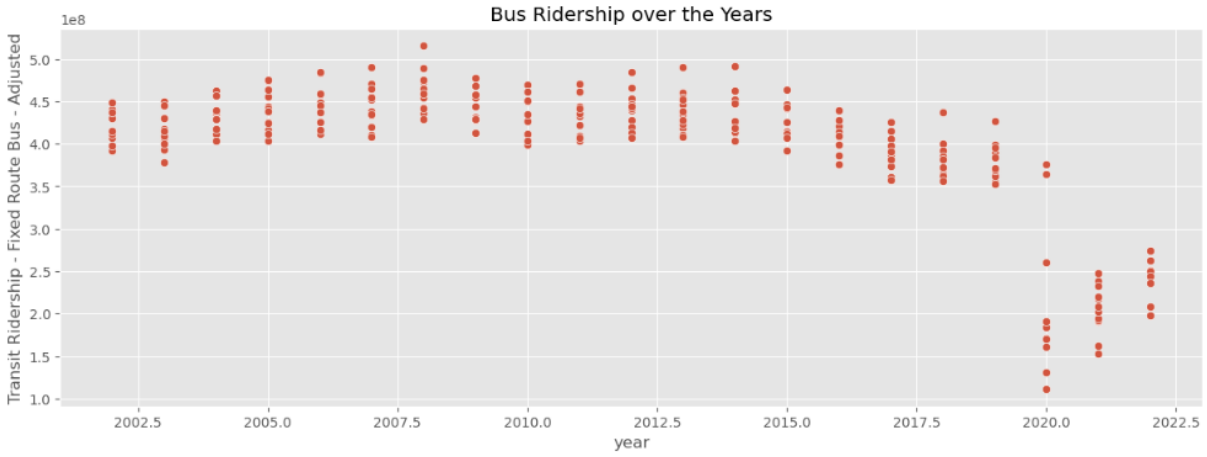
In order to explore confounding variables for our causal inference model, we looked at some potential confounding variables and their interaction over time. First looking at the popularization of other forms of public transit, we created a scatterplot of Urban Rail Ridership per Year. This graph seems to be steadily increasing since the early 2000s, but has a sharp dropoff during the beginning of the COVID pandemic in 2020. Since then, urban rail ridership appears to be increasing again. This graph could be used to suggest that the popularization of other forms of transit besides driving personal automobiles could lead to less people driving on roads which would lead to less accidents, and hence government spending may not be the only cause of improved road safety. So, popularization of other forms of transit is a good choice for a confounding variable.

Passenger Rail Ridership Confounding Variable



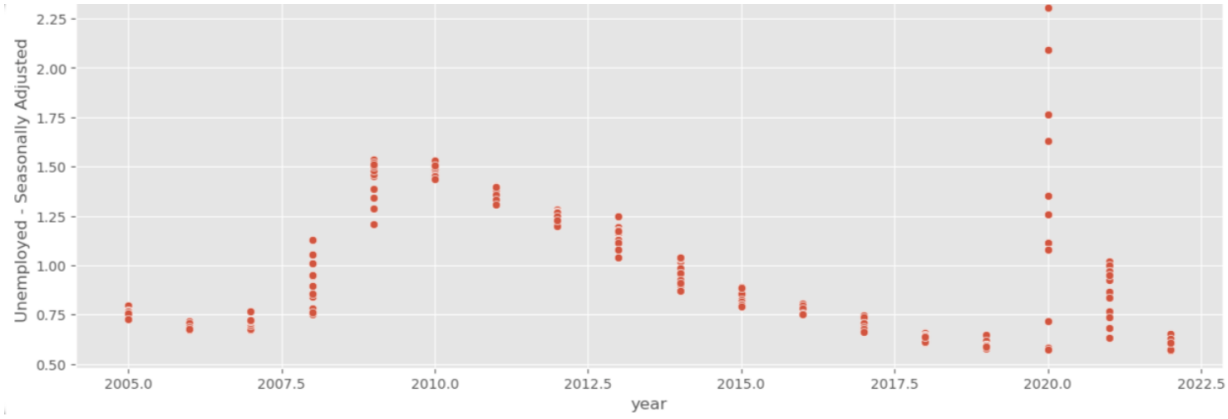
Similarly, we created a scatterplot of passenger rail ridership over time and found that there is a steady increase in passenger rail ridership since the early 2000s, but a noticeable dropoff in 2020 during the beginning of the pandemic. This graph interestingly includes a really high spike in the 90s that would require domain knowledge to investigate. This graph could be used to suggest that the popularization of other forms of transit besides driving personal automobiles could lead to less people driving on roads which would lead to less accidents, and hence government spending may not actually be the only cause of improved road safety. So, popularization of other forms of transit is a good choice for a confounding variable.

Bus Ridership Confounding Variable



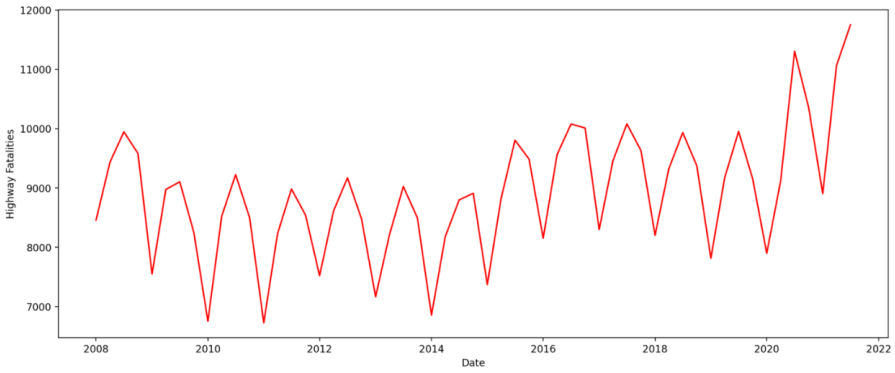
However, when we created a scatterplot of bus ridership over time, we found instead that bus ridership seems to be slowly decreasing since the early 2000s, with a noticeable dropoff during the beginning of the 2020 COVID pandemic. Since then, there has been a slight increase in ridership, but nowhere near pre-pandemic levels. Even though this graph doesn't suggest an increase in popularity in this form of public transit, the fact other modes have been increasing is still enough to suggest it's a useful confounding variable to use.

Unemployment Rate Confounding Variable



To examine another confounding variable, unemployment, we created a scatterplot of the national unemployment rate over time. This graph demonstrates a steady increase in unemployment since the 2008 recession, with a very large spike in unemployment at the beginning of the 2020 COVID pandemic. Thereafter, unemployment rates decrease almost to pre-pandemic levels. This graph could be used to suggest that high unemployment rates might lead to less commuters and hence less people on the roads and hence less accidents, therefore government spending may not be the only cause of improved road safety. So, unemployment is a good choice for a confounding variable for our model.

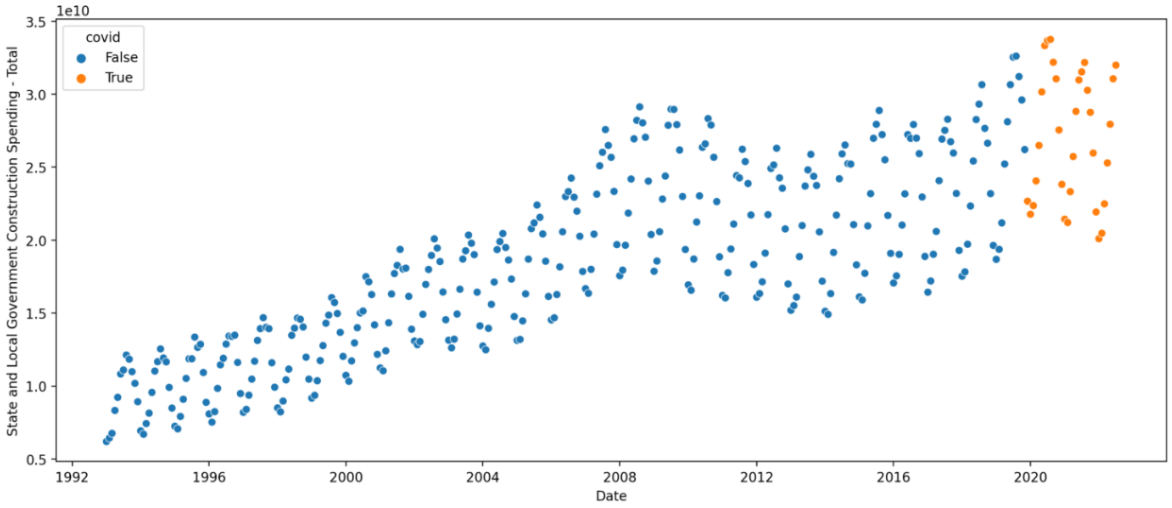
Highway Fatalities over Time



	Date	Highway Fatalities
830	2016-03-01T00:00:00	<NA>
831	2016-04-01T00:00:00	9,563.0000
832	2016-05-01T00:00:00	<NA>
833	2016-06-01T00:00:00	<NA>
834	2016-07-01T00:00:00	10,078.0000
835	2016-08-01T00:00:00	<NA>
836	2016-09-01T00:00:00	<NA>
837	2016-10-01T00:00:00	10,011.0000
838	2016-11-01T00:00:00	<NA>
839	2016-12-01T00:00:00	<NA>

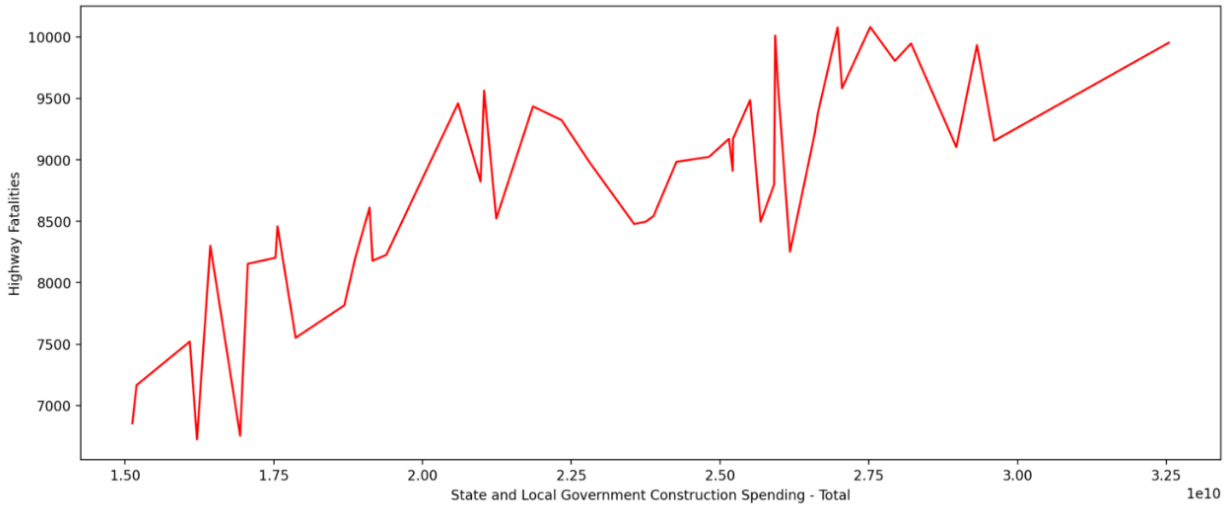
This graph shown above is the monthly highway fatalities from 2008 through 2022. We notice that highway fatalities are only calculated every other two months leading to our strange pattern. Nonetheless, our graph shows a slight increase in highway fatalities over the years with the increase happening around 2015. Strangely, we notice that from 2020 and onwards, the highway fatalities actually increased despite the pandemic that rendered less drivers on the road.

Total Government Spending over Time

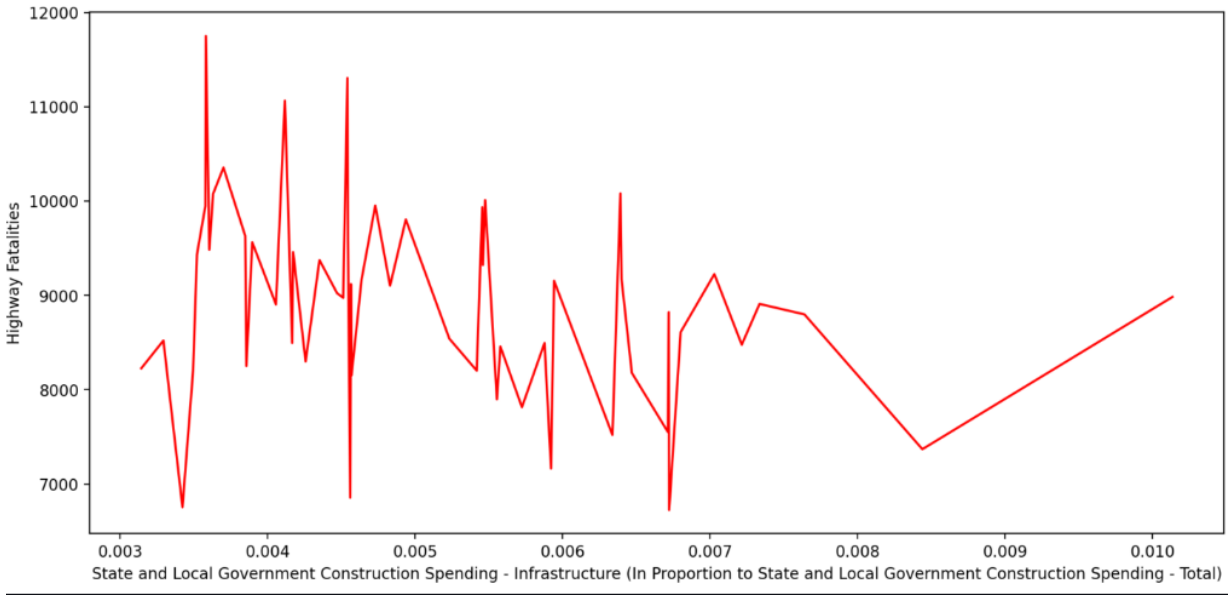


The following scatterplot graph above displays the yearly total government spending with Covid Times (2019 and onwards) being the hue and each dot representing months. We can see a strong increase in government spending over the years with a major hump in 2008. During the COVID Pandemic, we still see a higher increase in government spending but not to an extreme extent in proportion to the previous years.

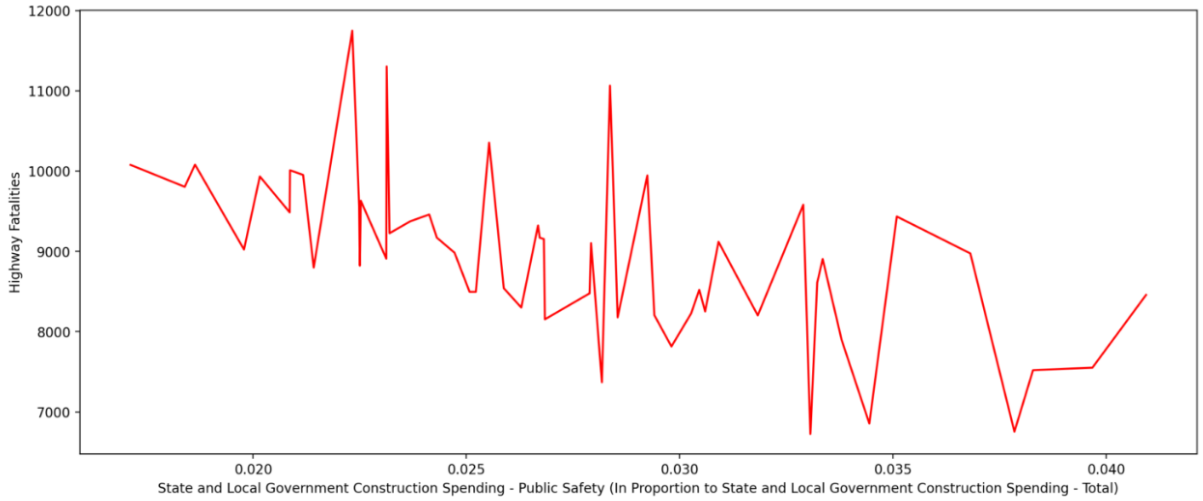
Highway Fatalities vs Government Spending



Here, the graph displays the relationship between Total Government Spending and Highway Fatalities before the COVID Pandemic. We notice that despite some volatility and dips, there is still an underlying increase in highway fatalities for every increase in Government Spending. With our causal question being the effect of Government Spending on Highway Fatalities still in mind, this graph is showing the story that increases in Government Spending leads to higher Highway Fatalities. However, it's still important to note that total Government Spending includes spending in other areas unrelated to vehicles/roads such as education, waste water, etc. Regardless, it's still something that is interesting to keep in mind and to dig deeper.



The graph above displays our main variable Government Infrastructure in proportion to total spending versus highway fatalities in our entire dataset. Looking at our plot, we see that there is no clear correlation between the two variables. It is extremely volatile and currently appears to show that government spending on infrastructure does not have any effect on highway fatalities. This is extremely important to keep in mind as we will test for causal inference utilizing regression later. For now, it appears there is no causal relationship or effect.



The graph above displays the Government Spending on Public Safety in proportion to Total Government Spending versus Highway Fatalities. We can see that there is a general decrease in fatalities as the government spends more on public safety. This makes sense because with higher spending on public safety, we expect there to be less fatalities as more drivers will become more cautious and safer. It was important that we looked at proportions in respect to Total Government Spending because total spending

is always different for each month. Due to this, we must take into account these different proportions because though it might appear that spending of public safety for a particular month might be higher than others, it does NOT indicate that it was prioritized in the government spending budget. In other words, we need to account for varying spending and seeing if truly increasing the spending will have an influence.

3. Research Questions

a. Multiple Hypothesis Testing

Our research question is, broadly, "Was mobility in rural versus urban California counties impacted differently from the COVID-19 pandemic?". We were interested in learning more about how different types of mobility (as defined by the Google Mobility dataset) had changed in urban and rural counties, and whether there were differences between these two types of counties. We believe that this question can be informative for agencies that look at transportation and urban planning, and it could have some applications in how we understand ease of access to basic necessities like food, pharmaceuticals, but also work commuting patterns and activities like recreation. Using multiple hypothesis testing is a good fit for this problem because it can answer this question with the data we have, provided in the Google Mobility dataset and given the various mobility metrics this dataset provides, we can test multiple hypotheses at once. Furthermore, applying two different correction algorithms (Benjamini-Hochberg and Bonferroni) will allow us to correct for the False Discovery Rate and the Family Wise Error Rate respectively.

Our hypotheses were:

- **Hypothesis 1**: Rural/Urban counties had the same mean change in parks movement during the period post-May 2021 (null hypothesis).
    - Alternative hypothesis: Rural/Urban counties had a different mean change in parks movement during the period post-May 2021 (null hypothesis).
- **Hypothesis 2:** Rural/Urban counties had the same mean change in retail/recreation movement during the period post-May 2021 (null hypothesis)
    - Alternative hypothesis: Rural/Urban counties had a different mean change in retail/recreation movement during the period post-May 2021 (null hypothesis).
- **Hypothesis 3:** Rural/Urban counties had the same mean change in grocery/pharmacy movement during the period post-May 2021 (null hypothesis)
    - Alternative hypothesis: Rural/Urban counties had a different mean change in grocery/pharmacy movement during the period post-May 2021 (null hypothesis).
- **Hypothesis 4:** Rural/Urban counties had the same mean change in workplaces movement during the period post-May 2021 (null hypothesis)
    - Alternative hypothesis: Rural/Urban counties had a different mean change in workplaces movement during the period post-May 2021 (null hypothesis).
- **Hypothesis 5:** Rural/Urban counties had the same mean change in transit station movement during the period post-May 2021 (null hypothesis)
    - Alternative hypothesis: Rural/Urban counties had a different mean change in transit station movement during the period post-May 2021 (null hypothesis).
- **Hypothesis 6:** Rural/Urban counties had the same mean change in residential movement during the period post-May 2021 (null hypothesis)

- Alternative hypothesis: Rural/Urban counties had a different mean change in residential movement during the period post-May 2021 (null hypothesis).

b. Causal Inference

Causal Inference Question: What is the effect of government spending, specifically, infrastructure spending on road safety? The confounding variables include: **popularization of other forms of transportation:** less people driving leads to less accidents, **(Un)employment:** might lead to less traffic since people are commuting less. Our **units** are our months (since that is the granularity of the data). Our **outcome** is highway fatalities, and our **treatment** is government spending, and the **technique** we'll be using is outcome regression. Outcome regression is a good choice because we do not have instrumental variables, and we can use the dataset to identify potential confounding variables which we use as features for the model. Some real-world decisions to be made include understanding how and where to allocate spendings, and whether there is reason to believe that a large amount of spendings leads to significant reductions in fatalities or not.

4. Inference and Decision

a. Multiple Hypothesis Testing

i. Methods

Given the Google Mobility data contained daily changes from baseline values for each of California's counties across all 6 mobility variables, we first computed a column of "Urban/Rural " for each entry of the data. We also needed to account for possible discrepancies in mobility patterns in different months of the pandemic. For instance, temporary address changes rose dramatically early in the pandemic as we saw in our EDA, whereas later in 2021 these changes were more stabilized. Thus, we restricted the dates of our analysis to May 1st 2021, up to the present. Furthermore, to obtain one observation per county, we aggregated the daily baseline changes into a mean value for every county across all mobility metrics (thus now one county had 6 mobility values corresponding to it, as well as a tag of urban/rural).

We chose to test 6 different hypotheses in order to answer our research question, and then apply the Bonferroni correction and Benjamini-Hochberg algorithm to our obtained p-values. We will choose a threshold of alpha = 0.05, and the Bonferroni Correction will allow us to account for the Family Wise Error rate, whereas the Benjamini-Hochberg will allow us to correct for the False Discovery Rate. In section 3a, we outline all 6 of our individual hypotheses.

To perform the testing, we tentatively decided to run a two-sample t-test for each of the six hypotheses, but given that the assumptions of the t-test were not satisfied, we chose a non-parametric test, the Mann-Whitney U-test instead. Specifically, the two-sample t-test requires that the samples from both populations follow a normal distribution, and we tested this assumption using QQ-plots. None of the categories for the urban or rural samples passed the QQ-plot test, as is visible in our code submission (under Parametric: two-sample t-test). The Mann-Whitney U test makes no assumptions on the underlying

distribution of the samples, and it tests under a null hypothesis that says the distribution of the two samples is the same, versus the alternative hypothesis: the distribution of the two samples is not the same. Thus, we decided to run the Mann-Whitney U test for all 6 of our hypothesis questions.

ii. Results

Here is a table summarizing each p-value per hypothesis we conducted:

	categories	p_values	Reject/Fail_to_reject_Bonferroni	Reject/Fail_to_reject Benjamini Hochberg
0	grocery_and_pharmacy_percent_change_from_baseline	5.200643e-08	True	True
1	parks_percent_change_from_baseline	1.007068e-03	True	True
2	transit_stations_percent_change_from_baseline	4.198928e-02	False	True
3	retail_and_recreation_percent_change_from_base...	1.296275e-07	True	True
4	workplaces_percent_change_from_baseline	1.136166e-05	True	True
5	residential_percent_change_from_baseline	1.164532e-02	False	True

From the table, **using a threshold of 0.05**, we would reject the null hypothesis for all of our hypothesis tests.

Using the new Bonferroni cutoff of 0.008333333333333333, we fail to reject the null hypothesis for "Rural/Urban counties had the same mean change in transit station movement post May 2021" and "Rural/Urban counties had the same mean change in residential movement post May 2021". We reject the null hypothesis for the remaining 4 hypotheses.

Using Benjamini-Hochberg, the cutoff was: 0.04198927663307768. Thus, we reject the null hypothesis for all 6 of our hypotheses.

iii. Discussion

After we applied our correction methods, four of the discoveries remained significant (using Bonferroni) and all of them were significant with the Benjamini-Hochberg and as individual tests. There were some limitations in our analysis due to the nature of the Google Mobility dataset, which is created using the samples from the people who have opted-in to the location tracking settings of Google Maps, that the company states "may or may not represent the exact behavior of the entire population"[2]. If we had more data, especially more detailed data at the level of individual businesses or sites that were being visited during the pandemic, we could conduct analysis using mobility categories that were more specific. For instance, we could look at variables such as types of restaurants and how people's eating preferences changed during COVID-19 given safety measures or other restaurant metrics. In aggregate, we can say that our

[2] Google.com, 2020, www.google.com/covid19/mobility/data_documentation.html?hl=en.

results showed differences in mobility patterns for rural and urban counties, and individually for each category we observed that these differences were significant, which could lead to decisions about site selecting different amenities, especially basic needs such as food in urban and rural areas.

b. Causal Inference

i. Methods

Our question consists of determining the effect **government spending on infrastructure** has on **highway fatality count**. Because our dataset contains multiple variables such as total government spending, unemployment rate, etc, there was a potential for different effects on highway fatalities. To test for government spending infracture's effect, we utilized the **outcome regression approach**.

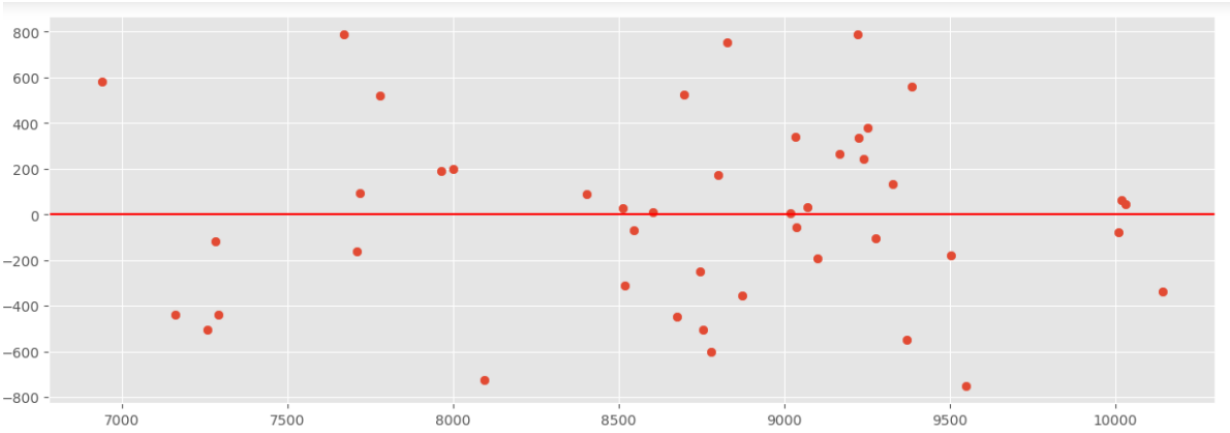
Assumptions of Regression:

Linear regression makes several assumptions:

- **There is a linear relationship between the features and the predicted variable.**
    - To test the existence of a linear relationship, we calculate the correlation coefficient between each pair of (feature, predicted variable), and we observe a negative correlation between proportion of health, transportation and infrastructure spendings in total government spendings, and positive correlations in transit ridership and total government spendings:

```
{'proportion_health': -0.7051895280284202,
 'proportion_transport': -0.33834311050369315,
 'Unemployment Rate - Seasonally Adjusted': -0.24860641332862055,
 'Unemployed - Seasonally Adjusted': -0.24692127639626243,
 'proportion_inf': -0.224533757932157,
 'Transit Ridership - Fixed Route Bus - Adjusted': -0.09567031475924015,
 'Labor Force Participation Rate - Seasonally Adjusted': -0.007550887482103298,
 'Transit Ridership - Urban Rail - Adjusted': 0.37471067134286223,
 'Transit Ridership - Other Transit Modes - Adjusted': 0.719230399749151,
 'State and Local Government Construction Spending - Total': 0.7816455705035091}
```

- **The residuals from the regression are identically and independently distributed following a normal distribution with mean 0 and constant variance.** Here is the plot of our residuals (y-axis) and the predicted values, in which we can observe no heteroskedasticity and residuals distributed around 0.



We then picked variables to reduce **multicollinearity** by dropping highly correlated variable pairs. The following table displays variables that have considerably high collinearity. We decided to drop Unemployment Rate due to its high collinearity, Government Spending Total because it is directly used in creating the proportional variables, and Labor Force Participation Rate as it is the converse of unemployment rate.

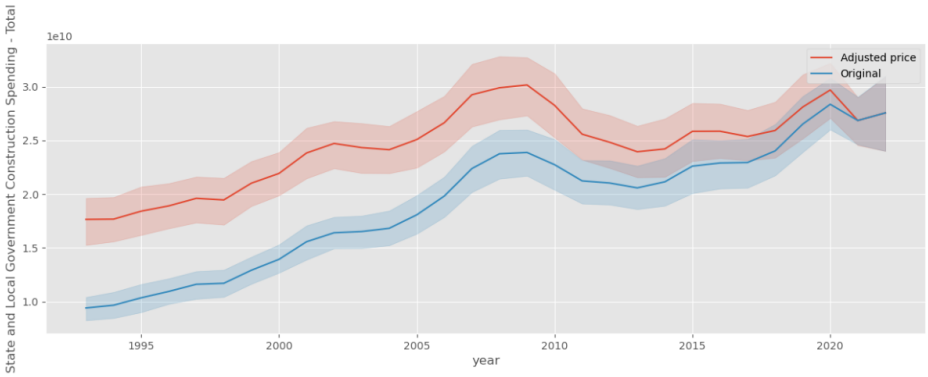
	Transit Ridership - Other Transit Modes - Adjusted Transit Ridership - Fixed Route Bus - Adjusted
Transit Ridership - Other Transit Modes - Adjusted Transit Ridership - Urban Rail - Adjusted	0.73704
Transit Ridership - Urban Rail - Adjusted Transit Ridership - Other Transit Modes - Adjusted	0.73704
Unemployment Rate - Seasonally Adjusted Unemployed - Seasonally Adjusted	0.99967
Unemployed - Seasonally Adjusted Unemployment Rate - Seasonally Adjusted	0.99967

Feature Engineering:

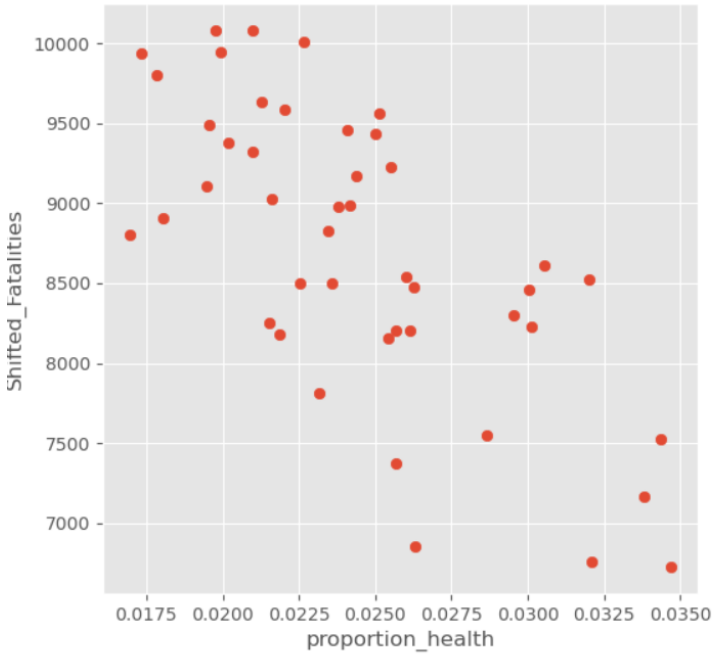
To normalize the dataset, we need to adjust and account for inflation. Since our dataset goes back to the 1950s and above, we need to convert all price values in the dataset to current day currency. Utilizing the Consumer Price Index (CPI), we adjusted all prices to 2021 currency as there is no CPI yet for 2022. We also needed to create new proportion variables (i.e government spending on infrastructure proportional to total government spending) to account for changes in total government spending changing each month and year.

To account for Time Delayed Effects, we also decided to shift all of our data forward by one year to ensure that the policy or change has indeed had time to cause effect. As there is no standard amount of time needed for policy effect to take place, one year will be our assumption.

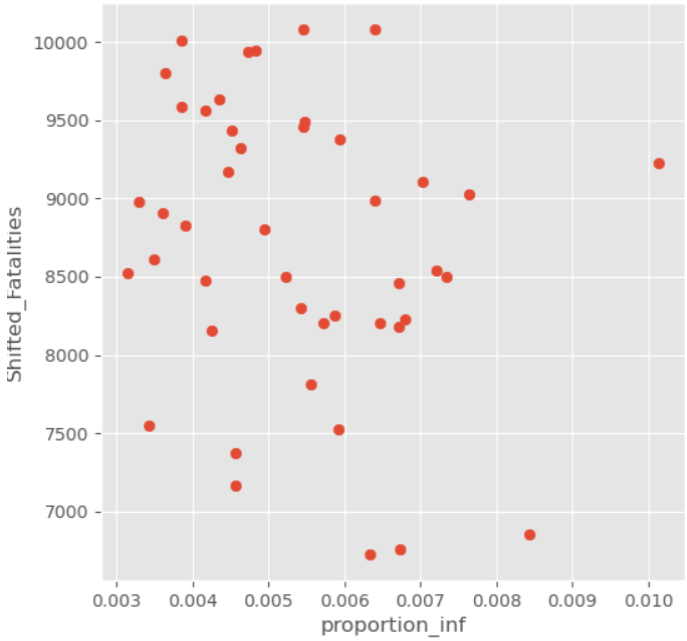
The following graph below displays the conversion of the dataset prices to current day prices for Government Spending Total. As we expect, we see prices to be slightly higher than what they were originally and does not drastically change the economic intuition behind the total spending.

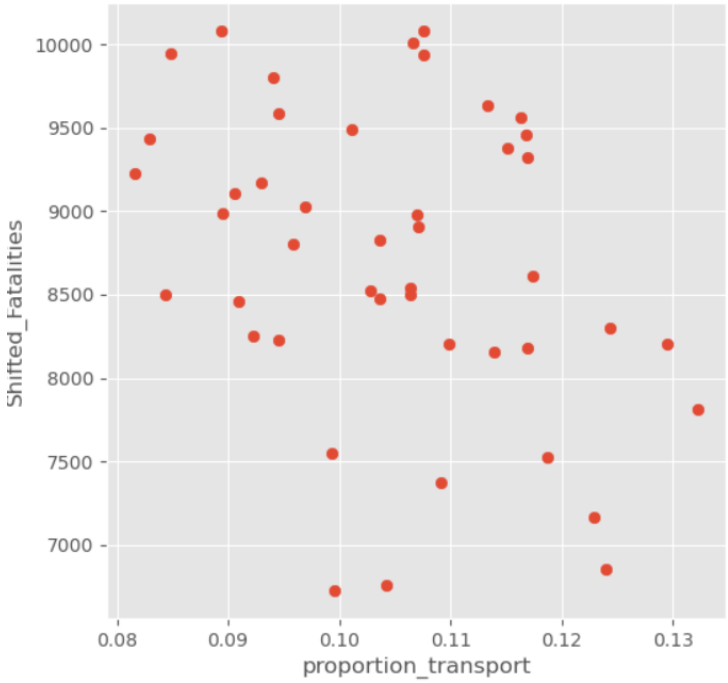


The following graphs below will showcase our variables and confounders that we have picked out and their correlation against our outcome variable of fatalities (where Time Delay has already been implemented). We see that most will have a linear correlation holding our current assumptions true.

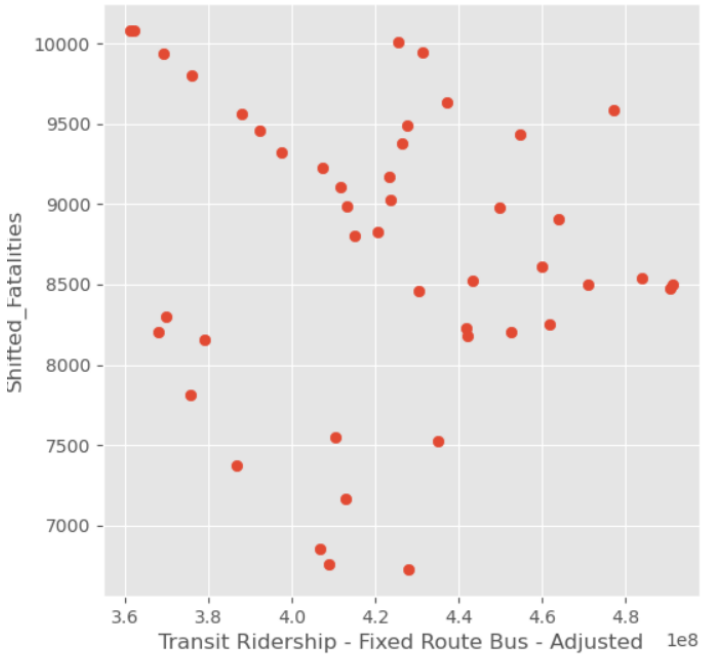


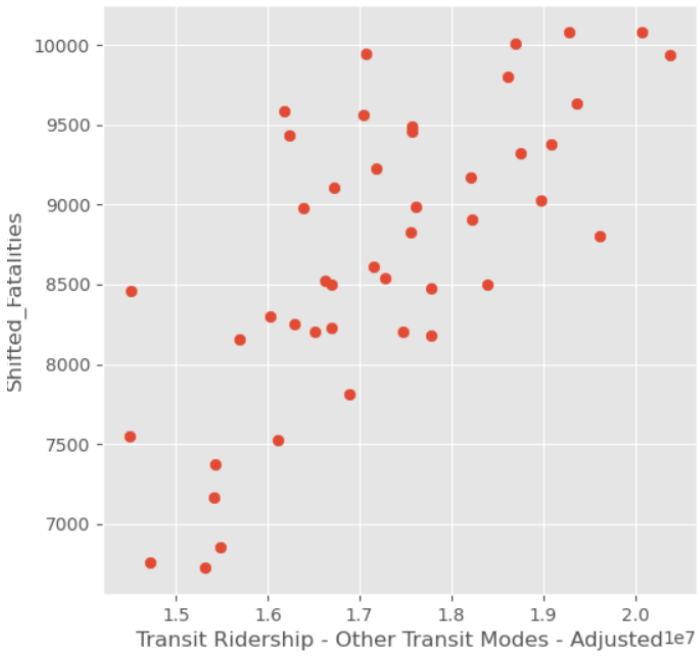
Group: Delilah Catron, Dea Bardhoshi, Derek Punaro, Quoc Huynh

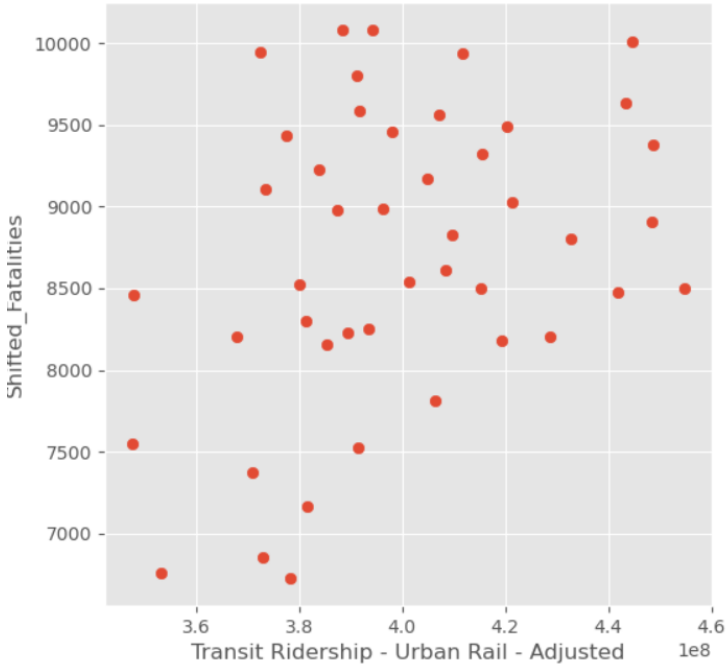


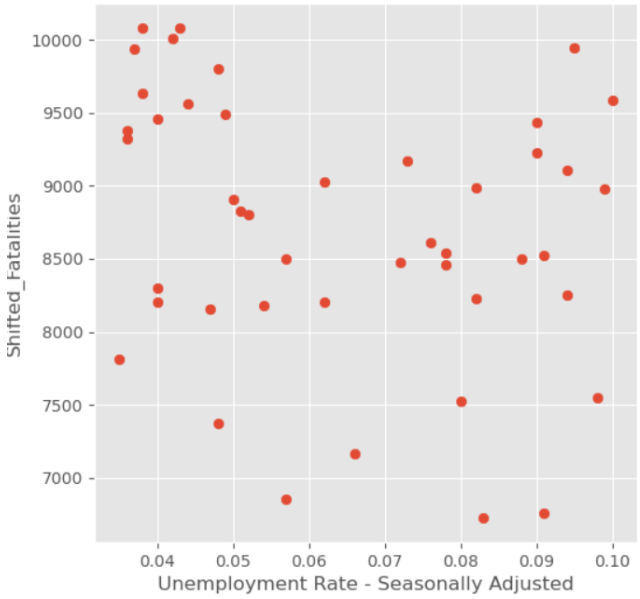


Group: Delilah Catron, Dea Bardhoshi, Derek Punaro, Quoc Huynh









The final model for our regression was:

Highway_fatalities = unemployment_rate + proportion_infrastructure_spendings + proportion_healthcare_spendings + proportion_transportation_spendings + transit_ridership_rail + transit_ridership_bus + transit_ridership_other + first_quarter_of_year

ii. Results

OLS Regression Results

Dep. Variable:	Shifted_Fatalities	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	18.75
Date:	Sun, 11 Dec 2022	Prob (F-statistic):	1.05e-10
Time:	13:45:55	Log-Likelihood:	-332.96
No. Observations:	45	AIC:	683.9
Df Residuals:	36	BIC:	700.2
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.55e+04	2735.549	5.666	0.000	9950.495	2.1e+04
Transit Ridership - Other Transit Modes - Adjusted	0.0001	0.000	1.026	0.312	-0.000	0.000
Transit Ridership - Fixed Route Bus - Adjusted	-1.533e-06	4.93e-06	-0.311	0.758	-1.15e-05	8.46e-06
Transit Ridership - Urban Rail - Adjusted	-1.132e-05	7.32e-06	-1.546	0.131	-2.62e-05	3.53e-06
proportion_transport	-1.364e+04	8839.049	-1.543	0.132	-3.16e+04	4285.890
proportion_health	-2.762e+04	2.31e+04	-1.197	0.239	-7.44e+04	1.92e+04
proportion_inf	-1.064e+05	4.61e+04	-2.311	0.027	-2e+05	-1.3e+04
Unemployment Rate - Seasonally Adjusted	-1.321e+04	8553.327	-1.544	0.131	-3.06e+04	4137.224
first_month	-1320.8776	334.282	-3.951	0.000	-1998.833	-642.922

A form of metric we used to verify the quality of our regression model was utilizing the Adjusted R-Squared (AR2) in which we chose the model with the highest AR2. Based on our results, we notice that our AR2 is 0.763 which indicates that our regression is modeling our data well. Looking at our focused variable *proportion_inf,* we notice that its coefficient is a pretty large value indicating that there is a large effect. Looking at its p-value we notice that it is below our threshold of .05 indicating that this variable is indeed statistically significant.

Something we need to take note of is the standard error, which has a large value. This is implying that there is a high level of uncertainty for this specific value. Below is a table from above displaying each variable's confidence interval values.

Variable Name	**0.025 (Lower Bound 95% Percent Confidence interval)**	**0.975 (Upper Bound 95% Percent Confidence interval)**
Constant	9950.495	2.1e+04
Transit Ridership - Other Transit Modes - Adjusted	-0.000	6.62e-05
Transit Ridership - Fixed Route Bus - Adjusted	-1.15e-05	8.46e-06
Transit Ridership - Urban Rail - Adjusted	-2.62e-05	3.53e-06
proportion_transport	-3.16e+04	4285.890

proportion_health	-7.44e+04	1.92e+04
proportion_inf	-2e+05	-1.3e+04
Unemployment Rate - Seasonally Adjusted	-3.06e+04	4137.224
first_month	-1998.833	-642.922

From our regression summary, there is indeed an effect of government spending on infrastructure on highway fatalities. Specifically, for every unit increase of spending on infrastructure, we expect to see a decrease of 1.064e+05 fatalities holding all else constant. As our p-value is also statistically significant, it further supports and reinforces the alternative hypothesis that there is indeed an effect.

iii. Discussion

Some further things we can do to potentially create even more accurate results is to look even deeper into the time delay aspect of the dataset and try to utilize other sources to determine a more specific time shift that we can incorporate instead of generalizing to one year. Though our results support the intuition that more infrastructure spendings leads to less fatalities, our uncertainty in the estimate is relatively high: we believe this is a result of needing more data points, and perhaps data separated by state. In addition, because the data is indexed by time, there is some time dependency between the observations, which might require other time series approaches to address. Additionally, COVID creates different shifts in fatalities and in how spendings were decided by governments, which would need closer inspection.

5. Conclusion

For question one, our findings showed that urban and rural counties' mobility was impacted differently in the period post May 2021. For question two, we used causal inference to estimate the effect of infrastructure spendings on highway fatalities and for every unit increase of spending on infrastructure, we expect to see a decrease of 1.064e+05 fatalities holding all else constant. Our findings for question one were relatively narrow as we only looked at California counties, but similar research could be done for other states as well. For question two, we reported our causal effect with a large margin of uncertainty, which could be reduced by additional analysis using other confounding variables. We did not merge datasets for question

one because the Google Mobility dataset contained the county names and the mobility metrics/ For question two, our current dataset was so broad that it would be difficult to incorporate other datasets.

Based on our question one results, we would recommend studying the different types of mobility from an urban planning perspective as well: for instance, how are grocery shops and pharmacies spread out or clustered in rural areas, or whether there are any patterns in transport use between these two types of counties? Furthermore, through question two we determine that It is feasible that increased government spending on highways could lead to safer roads and fewer fatalities. For example, if the government spends money on improving road infrastructure, such as by adding guardrails, widening lanes, or improving signage, this could make roads safer and reduce the number of accidents and fatalities. We believe there are many types of studies that can build on this work, including economical analysis of the cost and benefits of infrastructure spendings, impacts the mobility has on social lives and in our health. As a call to action, we recommend combining our transportation-related findings to perhaps building more road infrastructure and thinking about the areas (rural versus urban) where this new infrastructure would be most effective in terms of safety but also allowing ease of mobility for residents.