Berkeley
Division of Computing, Data Science, and Society

Data Science 102
Data, Inferences, and Decisions

Final Project - Fall 2022

Exploring the Relationship Between Mobility and US GDP and Investigating the Causes of Traffic Fatalities

Kaitlyn Du
Daniel Huang
Matthew Koen
Curtis Wong

Preface

This report was written in December 2022 at UC Berkeley for the Data C102 Final Project. During this project, we investigated United States Transportation Data collected from the COVID pandemic era and its' relationship to economic indicators such as GDP. The overarching goal was to explore inefficiencies, if any, in the Department of Transportation's spending, discover to what extent government investments in local infrastructure improved the lives of residents, and how the government can decrease fatality rate on roads.

Abstract

We chose to focus on these aspects of economic efficiency in relation to transportation: (i) During COVID year 2020, is there a relationship between the change in mobility and the change in GDP?; (ii) What are the causes of traffic fatalities and their corresponding prevention measures?

Contents

1 Introduction

We intend to take a deeper dive into exploring the United States' economic efficiency by viewing the effectiveness of government spending from a data perspective. Given how the sudden halt in mobility during the COVID pandemic caused an economic downturn, we can assume that there's a strong relationship between mobility and the economy; this would not be an uncharacteristic assumption as we've seen many examples over the past few years from how essential workers have been the backbone of the economy. With this, we intend to investigate the relationship between investing government funds in different facets of transportation, and how efficiently these investments support the economy. An increase in mobility and transportation generally implies more traffic fatalities, our secondary research question is to explore the causes of these fatalities and any preventative measures to take to reduce these numbers.

2 Data Overview

2.1 Bureau of Transportation Statistics: Monthly Transportation Statistics

One of the provided census data sets sourced from the Department of Transportation's website - each row is a monthly statistic for the United States as a whole ranging from "Highway Vehicle Miles Traveled" to "State and Local Government Construction Spending - Mass Transit". This is fairly granular data for the past few years and possibly the most optimal granularity for this type of data set - it's nearly impossible to collect economic data daily, and it captures the seasonal/monthly fluctuations of transportation while eliminating daily noise.

According to the descriptions provided, these are monthly estimates of the United States as a whole. Data points, such as fatalities, were collected via Fatality Analysis Reporting System - another census collected from police reports, and airline traffic was collected from commercial air carriers. As most of these reported data points are estimates, it's unlikely that participants were aware that their mobility data was used in this survey; additionally, given that a few participants were fatalities, it is certainly unlikely they were unaware of their participation in this census.

Something to take into account would be measurement error, as most of the data points are estimates. Additionally, we should be cautious of systematic exclusion - given that data points originated from other local government and private entities, geographic areas where these resources are not as available might not be included in our overall data. Given that some other data sets were more granular in terms of geographic regions, it would be interesting to see how transportation spending affected smaller localities, such as cities.

2.2 Google: Daily Community Mobility Data

The following provided data sets were sourced from Google's COVID data collection initiative; we used data from the years 2020 - 2022. Each row represented an individual region's data, in the case of the US, this was state data, which we then grouped into average monthly data for the country so that it was of the same granularity as GDP data.

This data was sourced from Descartes [4], or anonymized/de-identified mobile device data locations, to measure mobility. This was a census of the population that uses mobile devices, as such, eliminates any groups that do not have mobile devices for any reason from personal preferences to economic status. While it seems like most of the population does have some form of a mobile device, this potentially excludes the older and youngest generations. It is unsure that participants were aware of the collection or use of their data; it's likely that there were terms of service but unlikely that many read the terms of service. Despite being anonymized data, for obvious reasons, this yields clear ethical concerns about the data collection. Additionally, Google discloses that all devices have locations turned on automatically - users have to go the extra step to turn off location sharing in order to have their data withheld. [2]

The data provided is average baseline changes in percent, it would be interesting for us to explore baseline percentages in terms of spending as well in each region. Did spending on recreation, groceries, or other household necessities increase? How is mobility impacted by socioeconomic status?

2.3 U.S. GDP 1960 - 2022

This data set was one that we sourced and imported from a federal website. [1] This data was very granular in that it included monthly US GDP from 1960 up to 2022. The data was provided as a CSV on the FRED website, and we chose to include this data set to have more insight into more granular GDP data to compare with our mobility and transportation data. As this is the national GDP, this is a statistic, calculated by national census data about private consumption, private investment, government investment, and export/import data points. Less so than our previous data sets, participants were probably more aware that their data was being collected by the government in batches for research purposes - it is not a surprise for a semi-regular census or GDP report to be released, and is a major talking point in political campaigns. Some interesting features to explore would be GDP calculated per locality, such as a city or town, instead of as a nation. It would allow us to explore how the pandemic has disproportionately impacted each city.

3 Research Questions and Implications

3.1 Question 1: During COVID year 2020, is there a relationship between the change in mobility and change in GDP?

Mobility supports economic mobility, in terms of being able to move efficiently between jobs, schools, shopping centers, and other locations. We want to know if changes in GDP have affected mobility in the United States. Due to the size of the data set, our group decided to narrow down our focus to just the state of California. We believe this question is worth investigating because mobility is usually a sign of healthy and active communities and an indicator of high quality of life; we want to know if the change in GDP during COVID has a relationship with the various mobility categories: retail/recreation, grocery/pharmacy, parks, transit stations, workplaces, and residential areas.

For this research question, we decided to use multiple-hypothesis testing. The reason for this is that as enumerated above, there are many categories for mobility, meaning we can conduct a hypothesis test for each one. More specifically, since we are exploring the relationship between a categorical variable and a quantitative variable, we will be employing A/B testing.

3.2 Question 2: What are the causes of traffic fatalities and their corresponding prevention measures?

Traffic fatalities are devastating losses; we want to discover the main causes of these fatalities and try to implement data-driven decisions to implement prevention measures to reduce fatalities. While there exist pure accidents, inclement weather, and other factors play a role in fatalities. If there are systematic fatalities in transportation safety legislation, from public transportation to aviation safety that could be improved and have been perhaps overlooked, we want to explore better government policies. For example, could adding more traffic lights improve road safety? Could requiring more vehicle maintenance checks save more lives? Should the strategy to invest more into medical facilities, to reduce fatalities instead? Furthermore, given health data on the state of drivers - how often they'll drive sober or sleepy can help policymakers know where to invest more time. If we obtain data on an increase in drunk driving among teenagers, they'll be able to invest into anti-drinking programs, while if accidents are caused by sleepy drivers, they'll be able to increase the number of rest stops along long stretches of highways for sleepy drivers.

For this research question, we decided to approach this using GLM's and non parametric methods, using Poisson regression as the GLM of choice, and random forests as the non-parametric model of choice. This was the optimal choice, as infrastructure spending and fatalities share too many confounding variables.

4 Exploratory Data Analysis

4.1 EDA: Research Question 1

Before looking into mobility, we wanted to check on a more granular level if infrastructure spending is correlated with GDP over the years. For the Department of Transportation data set, we cleaned the data to include a year, month, and date column. During this process, we also discovered that the data set contained quite a large number of null values, so we decided to calculate the percentage of null values in each column. With these results, we deliberated whether or not a column was significant and hand-selected which features we believed were relevant to our research questions; dropping the rest. We believed that features with more than 50% null values were not worth imputing, as this could bias our models or create irregularities due to the sparsity of the feature matrix. With this data set properly cleaned, we plotted the United States' cumulative GDP over the years and the total amount of money spent on infrastructure. To no surprise, the numerical values at such different scales, so we normalized the data by calculating its z-scores.

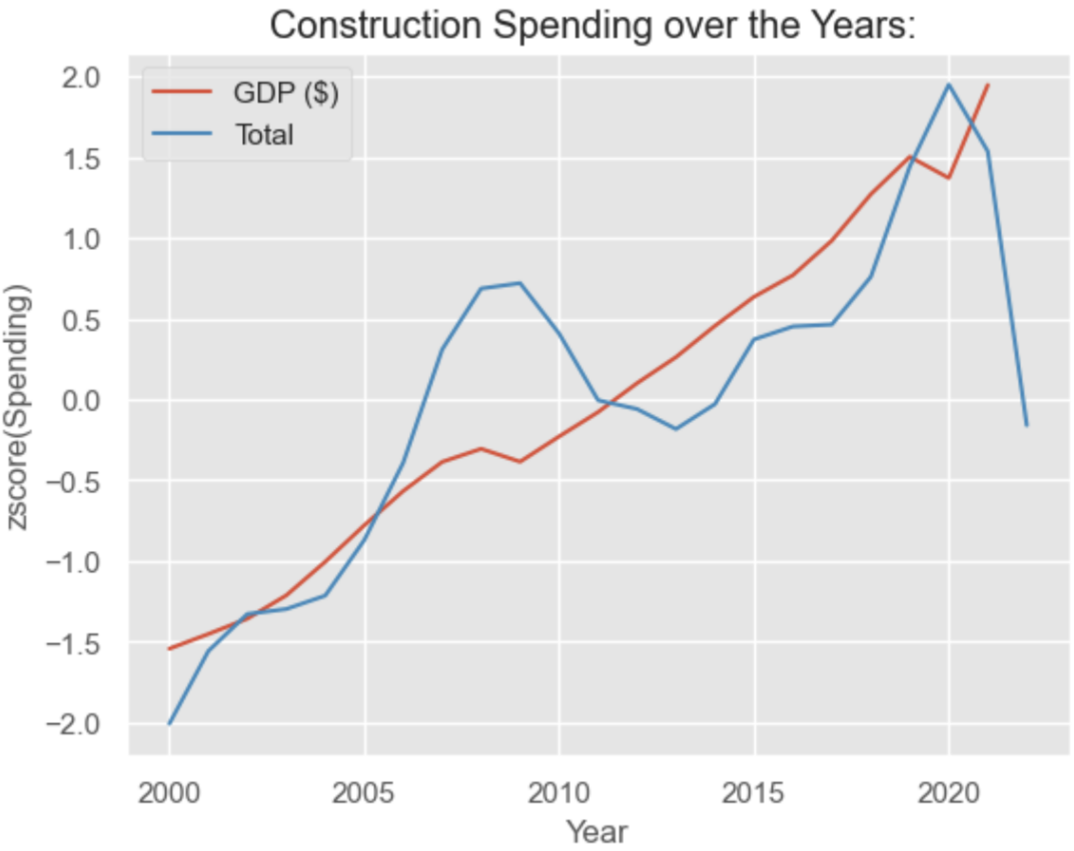


Figure 1: Line plot of US GDP (red) and US total infrastructure spending (blue) normalized.

As seen in the figure above, it appears that cumulative GDP does have a positive correlation with total infrastructure spending. However, while GDP seems to follow a mostly linear increase, infrastructure spending has two distinct spikes, one around 2008 and another and another in 2020. The former might have something to do with the 2008 financial crisis and this could be another potential topic to explore, but we won't be delving deeper into this as our research question is focused on COVID. The latter however is exactly when COVID happened, and although there was a spike, a very sharp drop followed, tanking infrastructure spending levels back down to that of 2013. Although we cannot make any causal statements now, we hypothesize that this drop could be possibly explained COVID. During that time the lockdown and quarantine greatly limited travel and outside work, which may have had an influence on infrastructure spending, or in this, the lack thereof.

The Google Daily Community Data contains mobility data for many countries and their regions (ie. states) spanning over a wide range of years. Due to the size of this dataset, we narrowed our focus on just the state of California and the year 2020, since that was when COVID was at its peak. Taking a closer look at the percent change in mobility over the months gave us the following line plot:

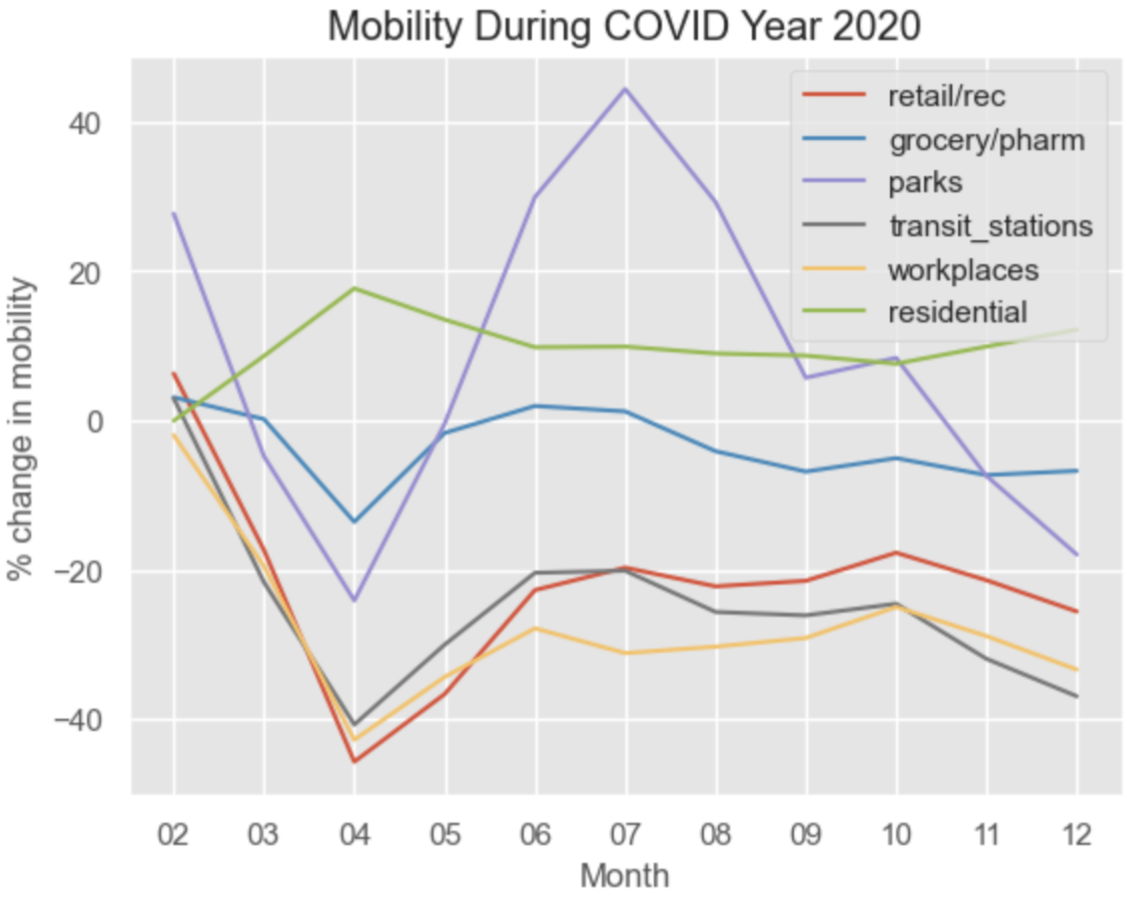
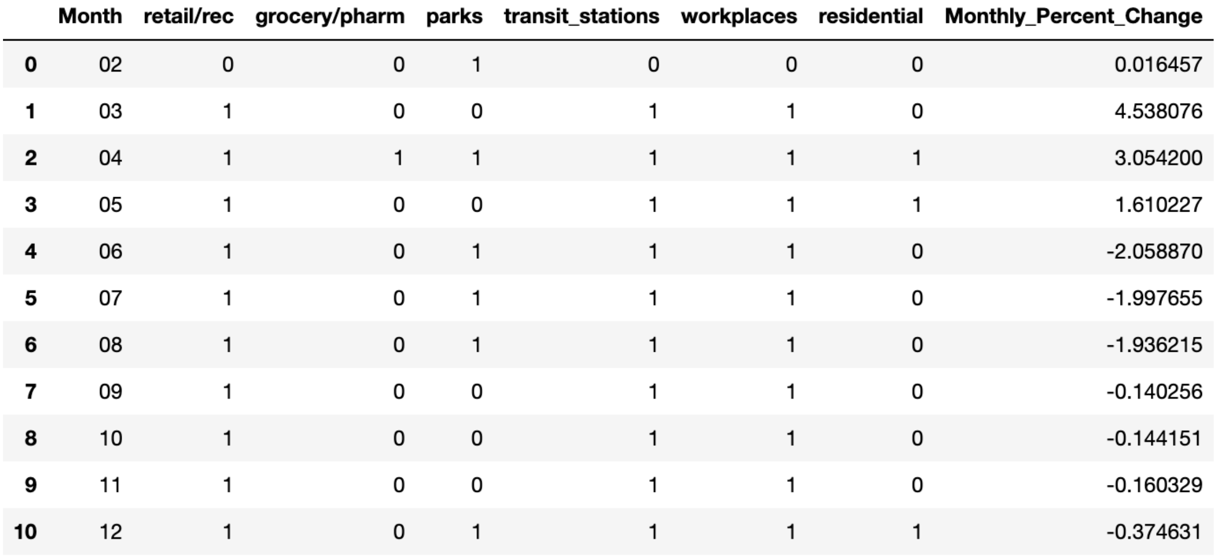


Figure 2: Line plot of the percent change in each mobility category.

From January to April, there is a clear nosedive in mobility for almost all categories. In sharp contrast, residential mobility was the only category that increased during that time period. Changes in retail/recreation, workplaces, and transit stations largely remained constantly decreasing at 20%. Once again, this could be explained by quarantine and lockdown measures where all of these categories were limited; retail shops and recreation locations closed down, workplaces mostly shifted to work from home, and most forms of public transportation closed down. Change in mobility for the grocery and pharmacy category also largely remained constant but close to 0% change. We think this is because commodities like food and medicine are inelastic goods, meaning that there are necessities with no acceptable substitutes. Residential mobility, as mentioned previously, experienced a spike. We think this may be influenced by the decrease in workplace mobility as people shifted to work from home. And finally, the last abnormality to address is the huge spike in the change of mobility for parks. Although a lockdown was in place, we recall that "essential" activities were still permitted and "exercise" happened to fall under that definition. Thus, with all the extra time people had, many may have gone to parks, either for exercise or, what we believe to be more likely, escapism.

Currently, each row contains numerical data for each category. However, in order to answer our first research question, we needed to convert our columns into categorical variables. From looking at the line plot, we decided to define a "significant change" to be any absolute percent change greater than 10%. With this definition, we were able to binarize our categories with 1 indicating a significant change and 0 otherwise. After that, we then merged with our supplementary data US GDP 1960-2022. In doing so, each row became a month containing whether or not there was a significant change in mobility for all the aforementioned categories and the percent change in GDP from the previous month. This is the final table that will be used to answer our first research question: During COVID, is there a relationship between the change in mobility and the change in GDP?



	Month	retail/rec	grocery/pharm	parks	transit_stations	workplaces	residential	Monthly_Percent_Change
0	02	0	0	1	0	0	0	0.016457
1	03	1	0	0	1	1	0	4.538076
2	04	1	1	1	1	1	1	3.054200
3	05	1	0	0	1	1	1	1.610227
4	06	1	0	1	1	1	0	-2.058870
5	07	1	0	1	1	1	0	-1.997655
6	08	1	0	1	1	1	0	-1.936215
7	09	1	0	0	1	1	0	-0.140256
8	10	1	0	0	1	1	0	-0.144151
9	11	1	0	0	1	1	0	-0.160329
10	12	1	0	1	1	1	1	-0.374631

Figure 3: The final preprocessed table used for multiple hypothesis testing.

4.2 EDA: Research Question 2

Upon first glance at the graphs, there appears to be seasonal spikes for fatalities, transportation spending, as well as highway miles travelled. Noticeably, in 2020, there is also a significant decrease in miles travelled due to the COVID lockdown, while generally, miles travelled increases over the summer holidays. This is likely due to families travelling during summer vacation causing these spikes. Government spending on general transportation increases steadily over the year in a normal curve, to spike in approximately July each year. However, there appears to be spikes in fatalities around the beginning of each season, especially so in the winter months. This is likely due to haphazard road conditions in colder states with icy roads, blizzards, and other dangerous driving conditions.

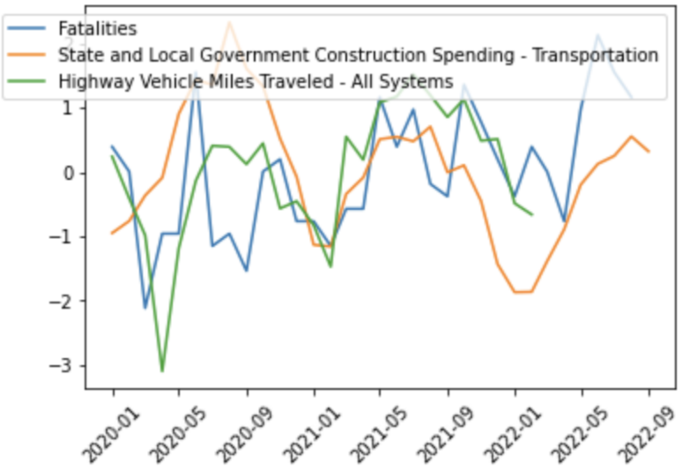


Figure 4: Fatalities and Spending

Oddly, the first months of the lockdown seemed to have the largest spike in state and local transportation spending, despite most staying in residential properties on lockdown. We are unsure of the reasoning of the spike, but we believe that this may have been influenced by commercial and private transportation - there became increased demand for online shopping for all goods. However, this should have been offset by most people working from home, eliminating the daily workforce transportation numbers.

Similarly to our first data set, each row contains numerical data for each category. However, two columns that we wanted to further investigate were changed into categorical variables - Highway Fatalities per 100 million, and Pavement spending, based on significant spending. Significant spending was defined as spending, or deaths above average. With this definition, the data was binarized, with 1 indicating a significant change and 0 otherwise.

Before binarizing the data, our fatalities data set was sourced from 3 years worth of COVID mobility data sets from 2020 - 2022 on Google. The data was extremely granular by zip code and by date, so they were then condensed into a national monthly average. Our fourth data set was the fatalities data set from the Department of Transportation, we selected columns that we wanted to explore further, and left merged them with the three mobility data sets. [3]

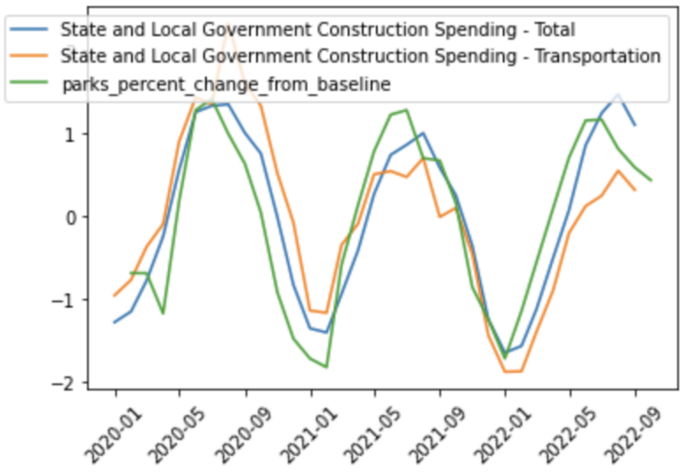
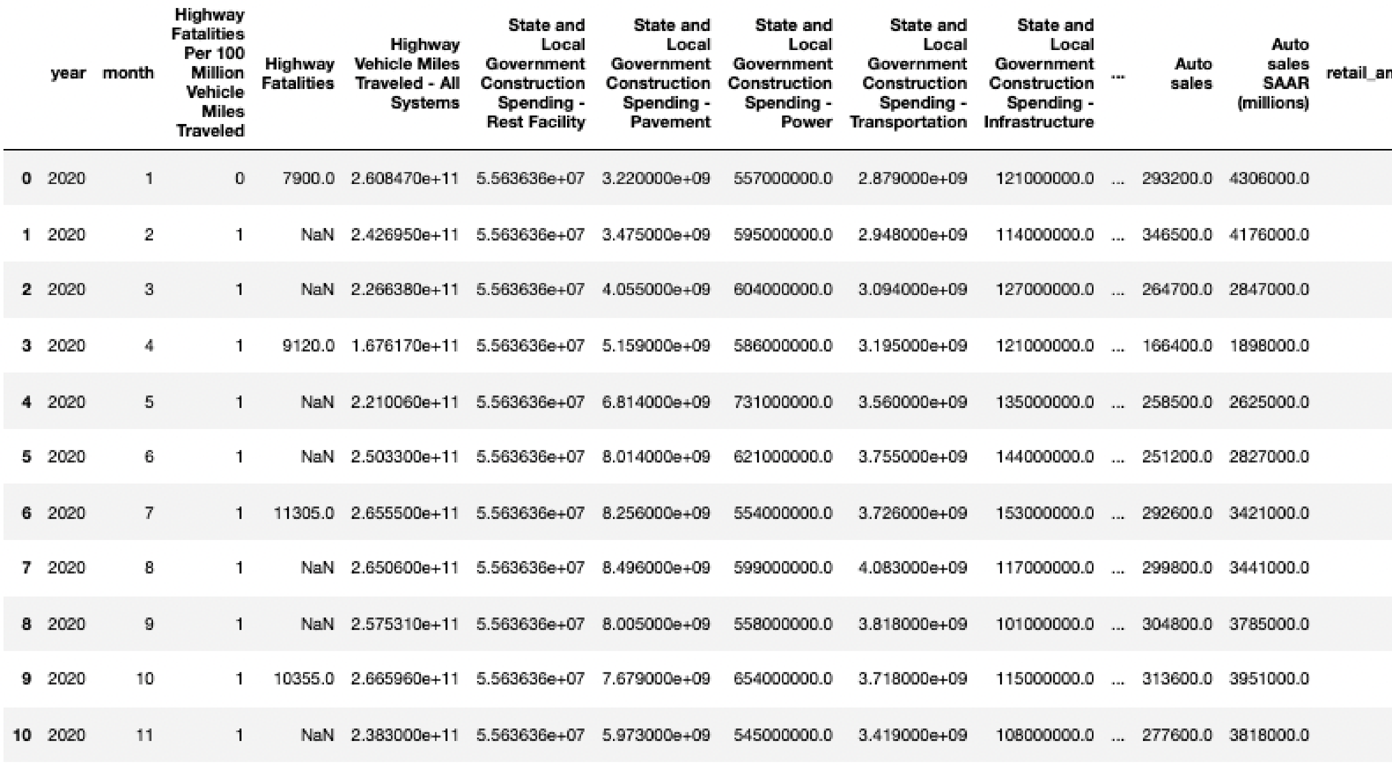


Figure 5: Fatalities and Spending

From this, we were able to preliminary visualize how current transportation patterns matched government spending across different features. For example, transportation spending and recreational mobility change were very tightly correlated - this is likely due to COVID, in which parks/outdoor recreation were some of the only things that were permitted under lockdown, and their usage has increased in popularity post-lockdown as well. With more investment into public spaces, this likely seemed to encourage higher quality parks, and more people to use them.



	year	month	Highway Fatalities Per 100 Million Vehicle Miles Traveled	Highway Fatalities	Highway Vehicle Miles Traveled - All Systems	State and Local Government Construction Spending - Rest Facility	State and Local Government Construction Spending - Pavement	State and Local Government Construction Spending - Power	State and Local Government Construction Spending - Transportation	State and Local Government Construction Spending - Infrastructure	...	Auto sales	Auto sales SAAR (millions)	retail_an
0	2020	1	0	7900.0	2.608470e+11	5.563636e+07	3.220000e+09	557000000.0	2.879000e+09	121000000.0	...	293200.0	4306000.0	
1	2020	2	1	NaN	2.426950e+11	5.563636e+07	3.475000e+09	595000000.0	2.948000e+09	114000000.0	...	346500.0	4176000.0	
2	2020	3	1	NaN	2.266380e+11	5.563636e+07	4.055000e+09	604000000.0	3.094000e+09	127000000.0	...	264700.0	2847000.0	
3	2020	4	1	9120.0	1.676170e+11	5.563636e+07	5.159000e+09	586000000.0	3.195000e+09	121000000.0	...	166400.0	1898000.0	
4	2020	5	1	NaN	2.210060e+11	5.563636e+07	6.814000e+09	731000000.0	3.560000e+09	135000000.0	...	258500.0	2625000.0	
5	2020	6	1	NaN	2.503300e+11	5.563636e+07	8.014000e+09	621000000.0	3.755000e+09	144000000.0	...	251200.0	2827000.0	
6	2020	7	1	11305.0	2.655500e+11	5.563636e+07	8.256000e+09	554000000.0	3.726000e+09	153000000.0	...	292600.0	3421000.0	
7	2020	8	1	NaN	2.650600e+11	5.563636e+07	8.496000e+09	599000000.0	4.083000e+09	117000000.0	...	299800.0	3441000.0	
8	2020	9	1	NaN	2.575310e+11	5.563636e+07	8.005000e+09	558000000.0	3.818000e+09	101000000.0	...	304800.0	3785000.0	
9	2020	10	1	10355.0	2.665960e+11	5.563636e+07	7.679000e+09	654000000.0	3.718000e+09	115000000.0	...	313600.0	3951000.0	
10	2020	11	1	NaN	2.383000e+11	5.563636e+07	5.973000e+09	545000000.0	3.419000e+09	108000000.0	...	277600.0	3818000.0	

Figure 6: Final combined fatalities and baseline percentage changes

5 Multiple Hypothesis Testing

5.1 Methods

In this section, the following hypothesis test was used:

- H_0 : A significant percent change in mobility category in either direction (increase or decrease) **has no relation** with the percent change in US GDP during that time. Any patterns that may arise is due to random chance.
- H_1 : A significant percent change in mobility category in either direction (increase or decrease) **is correlated** with the percent change in US GDP during that time.

As evident from our data overview and EDA, it makes sense to have multiple tests as the six categories are all unique, and our EDA further proved that there was a range of weak to strong associations with change in US GDP. Since there are multiple categories, A/B testing is perfectly suited for this purpose. To correct our multiple hypothesis tests, we will first use a naive $\alpha = 0.2$. Bonferonni correction will be applied to set a family-wise error rate (FWER) of 0.2. Lastly, the Benjamini-Hochberg (B-H) procedure will be used to limit a false discovery rate (FDR) of 0.2.

5.2 Results

Performing A/B testing with a sample size of $25,000$ on the six categories, retail/recreation, grocery/pharmacy, parks, transit stations, workplaces, and residential, yielded the following results:

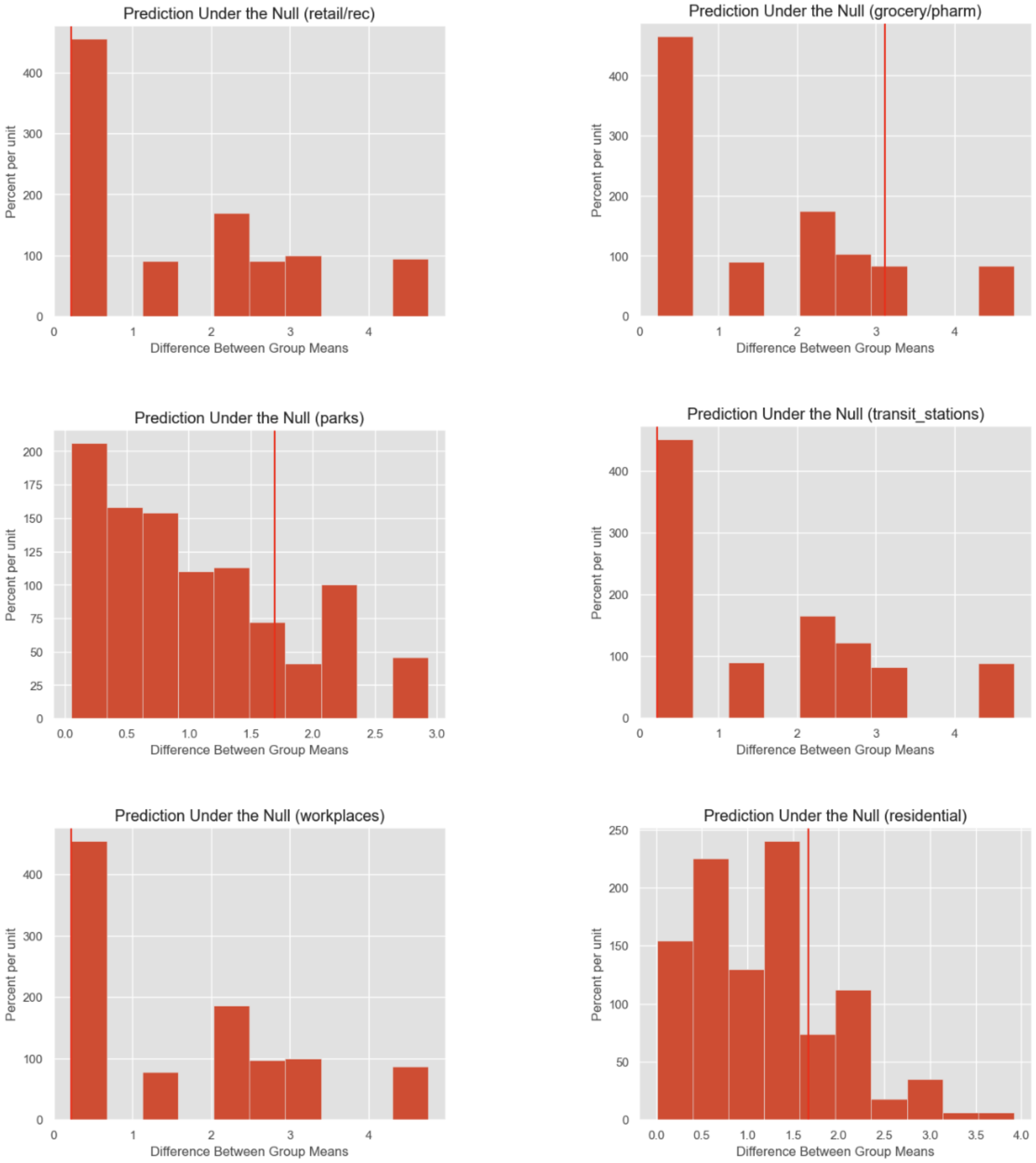


Figure 7: Results from A/B tests on all six mobility categories.

For retail and recreation, the p-value computed was 1.0 For grocery and pharmacy, the p-value computed was 0.167. For parks, the p-value computed was 0.194. For transit stations, the p-value computed was 1.0. For workplaces, the p-value computed was 1.0. Finally, for residential, the p-value computed was 0.217. With a naive $\alpha = 0.2$, we reject a total of 2 tests, concluding that significant changes in grocery/pharmacy and parks is related to changes in US GDP.

Moving on to correcting our multiple hypothesis tests, Bonferonni correction was applied with FWER $= 0.2$. This means that the probability of at least one of our discoveries being incorrect is 0.2. With six tests, this meant that our new $\alpha = \frac{0.2}{6} = 0.033$. This is a very low threshold and resulted in zero discoveries.

Next, we applied the less restrictive B-H procedure, setting a false discovery rate to 0.2. This means that the expected number of false discoveries is equal to $0.2 \times 6 = 1.2$. The algorithm computed a p-value threshold of 0.167, which meant that there was only 1 discovery; significant changes in grocery and pharmacy mobility is related to changes in US GDP.

5.3 Discussion

With no correction procedures and using an $\alpha = 0.2$, two discoveries were significant, grocery/pharmacy and parks. After applying Bonferonni correction, no discoveries were found. We believe this occurred because the error correction procedure is too restrictive, causing the α threshold to be 0.033. After applying the more lenient Benjamini-Hochberg procedure, one discovery was made, the test for significant changes in grocery/pharmacy was rejected.

In terms of limitations, we were greatly disadvantaged by the amount of data points as there are only twelve months in a year. If there was data on **daily** percent change in mobility and US GDP, then our tests would most definitely be better as our samples would be more representative. We do not believe p-hacking is relevant in our procedures as 25000 samples per bootstrap is generally a normal amount of samples.

Additional tests could be conducted if we had widened our research question. For example, instead of just analyzing data during COVID year 2020, maybe the data from the start of the 21st century could be investigated, as this would include months from 20000 to 2022. Although this exploration could be expanded to include the 20th century as well, our first dataset, Bureau of Transportation Statistics: Monthly Transportation Statistics, contains many null values during this time period, and our second dataset, Google: Daily Community Mobility Data, does not contain mobility data for that time.

6 Gaussian Mixture Model (Bayesian Hierarchical Model) and Prediction with GLMs and Nonparametric Methods

6.1 Methods and Discussion

In this section, we used feature selection through Decision Trees to test whether a feature is a good predictor for traffic fatalities based on feature importance. The model calculates importance through the Gini impurity weighted by the probability to determine the split of reaching the next node to make a tree. Choosing a maximum depth of 14 would prove to show the overall best results of the model. The classifier yielded these results:

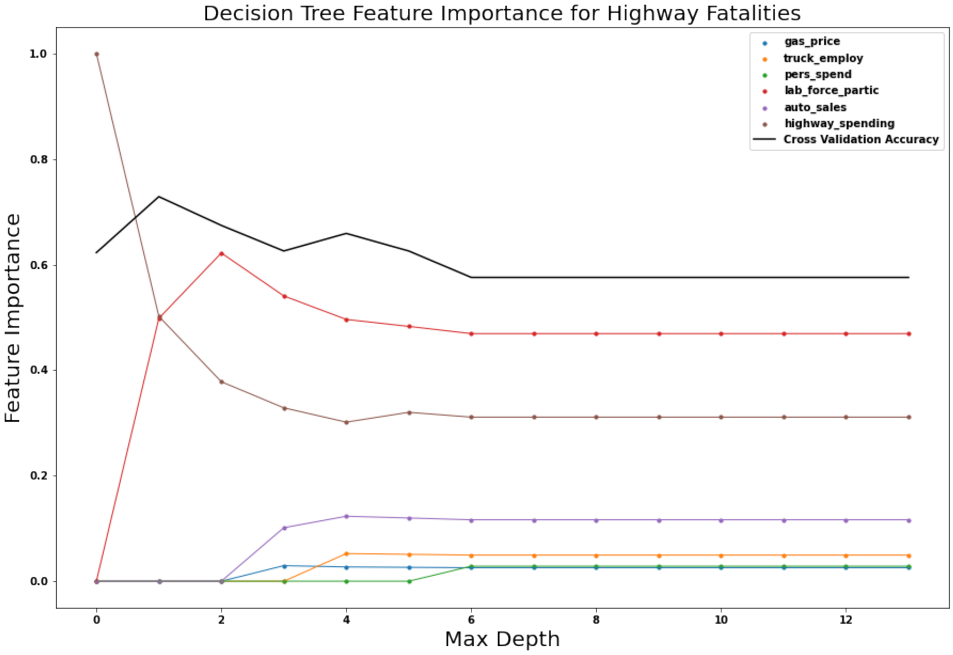


Figure 8: Decision Tree Feature Importance on Highway Fatalities

Upon initial inspection, we see that labor force participation has the highest importance followed by highway spending, then auto sales. Likewise, we used Linear Regression modeling to test the accuracy of each feature and how good they can predict the amount of fatalities.

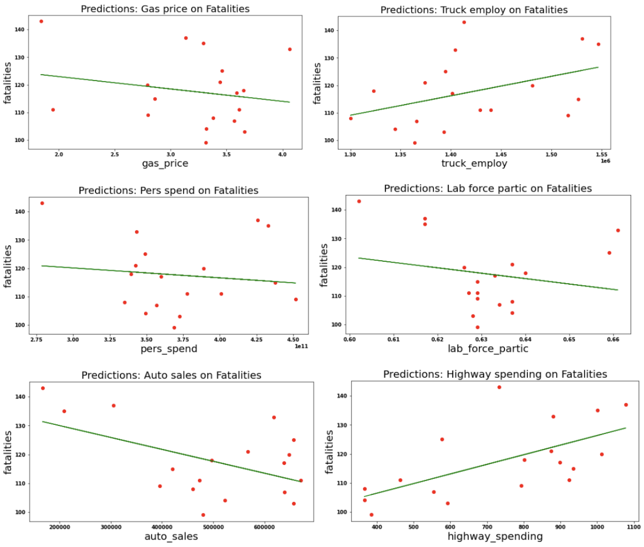


Figure 9: Linear Regression Modeling (Test Set Plot)

However, we cannot make any assumptions or causality that these features have on fatalities. The research question here is to figure out how one can lower traffic fatalities, and what contributes to traffic fatalities. The original method was to find an instrumental variable to determine the causal effect of highway infrastructure spending on traffic fatalities. However, after some EDA and seaborn pairplot examination, it appeared near impossible. The reasoning is that traffic fatalities and infrastructure spending share many confounders. For an instrumental variable, we are attempting to find something that only affects highway infrastructure spending, independent of traffic fatalities, independent of all confounders. Unfortunately, from external research and EDA, economic states and government types are the primary drivers for infrastructure spending. These, through the seaborn pairplot above, are highly linked to traffic fatalities through confounders themselves. So, instead, we took the approach to model traffic fatalities using GLM's and nonparametric methods, using Poisson regression as the GLM of choice, and random forests as the non-parametric model of choice. To evaluate each model's performance, we use root mean squared error and compare against each model.

```
Mean of Fatalities Distribution: 117.12068965517241
Variance of Fatalities Distribution: 115.96763460375072
```

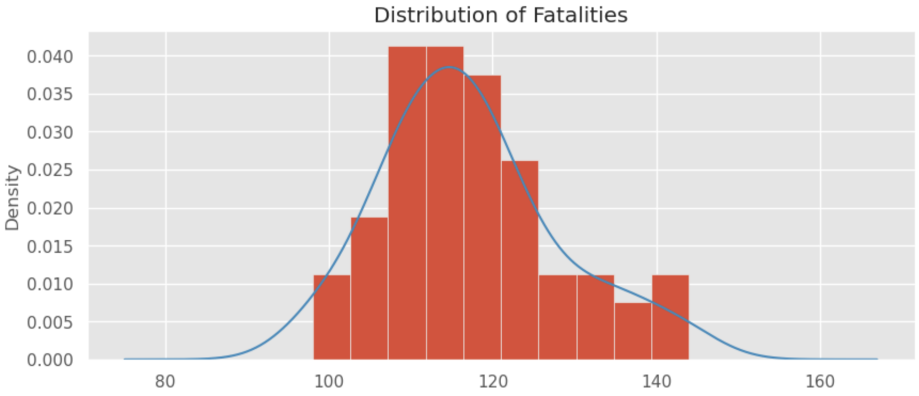


Figure 10: Distribution of Fatalities.

Poisson seemed like the most logical choice as fatalities is 1. A Poisson process (a count of events occurring) and 2. The mean and variance of the fatality distribution are near identical, lending itself to a Poisson distribution. we also tested linear regression and negative binomial regression against this hypothesis.

To test this hypothesis, we modeled fatalities under a weak Gamma prior with no covariates:

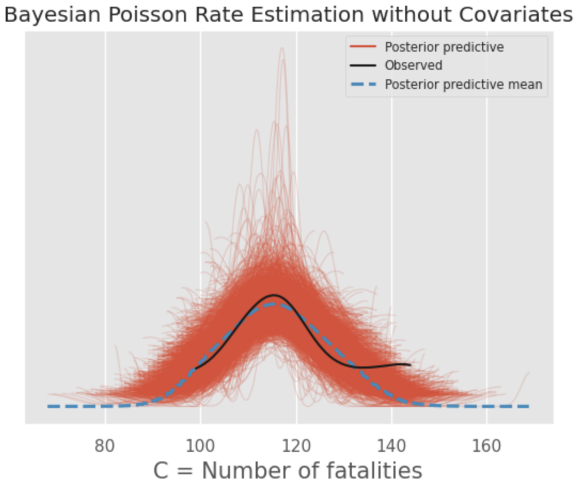


Figure 11: Poisson model of Fatalities under a weak Gamma prior with no covariates

As you can see, we achieved a relatively close fit. This suggests highway fatalities, while still a process dependent on confounders, is a relatively random Poisson process.

The first exploration was to consider how highway infrastructure spending is associated with different fatality levels. Because we are assuming highway fatalities follows a mostly Poisson process, most of the features selected have to do with the rate of car arrivals (which in turn causes rates of car crashes). Highway Infrastructure spending is the most unique of all of our features. This is because all other features, besides infrastructure spending, affect the rate of arrival of cars. However, infrastructure spending (while affecting rate of arrival of cars) also affects safety of roads, which isn't captured in the other features. To explore this further, we chose to create a Gaussian Mixture model. This is because highway infrastructure spending appears to be a mix of two Gaussian distributions:

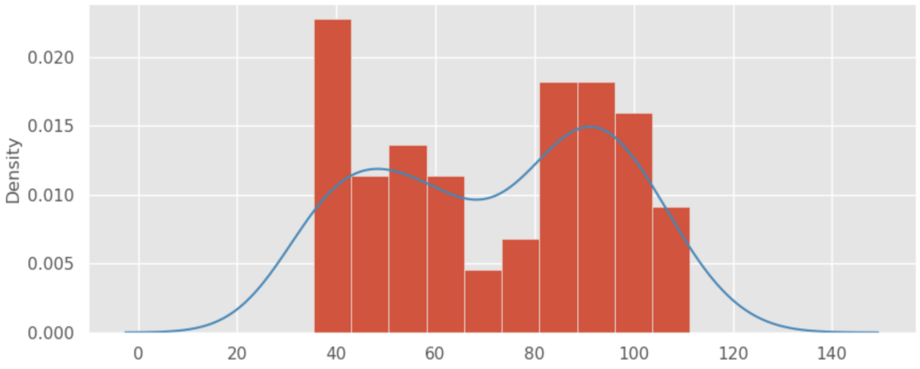


Figure 12: Distribution of Highway Infrastructure Spending

To model using a Gaussian Mixture model, we defined the followed variables and related distributions:

Let $z_i \sim Bernoulli(\pi)$ be each level of infrastructure spending. $q_i \sim Exponential(\lambda)$ be our rate of fatalities and $x_i|z_i, q_0, q_1 \sim Poisson(q_i)$, our number of fatalities.

- Let $z_i \sim Bernoulli$ distribution for z_i because our level of infrastructure spending comes from two different normal distributions: lower and higher infrastructure spending as seen above.
- Let rate of fatalities $\sim$ *Exponential* distribution as it is a rate value, and will be plugged into a Poisson distribution.
- Let number of fatalities $\sim$ *Poisson* distribution because of the above exploration of fatality data, and because it is a Poisson process.

6.2 Results

The results of the Gaussian Mixture Model were surprising, and not consistent with our chosen priors. We chose priors for z_i to be Bernoulli(.45) as the data appeared to skew slightly to the left. We chose parameters for the prior of q_0 to be Exponential(5) and q_1 to be Exponential(1), as we expected less highway infrastructure spending to be associated with more highway fatalities.

The results from our posterior distribution of q_i are as follows:

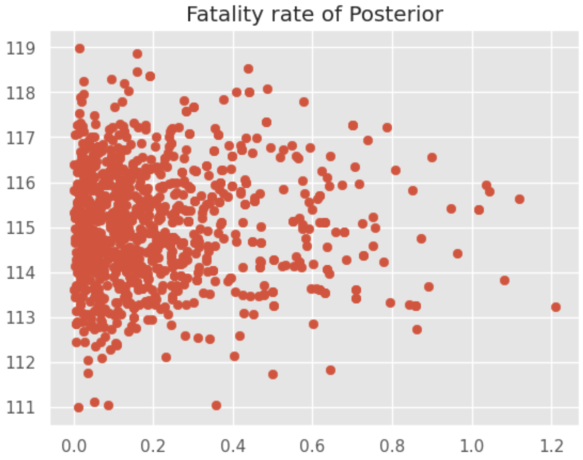
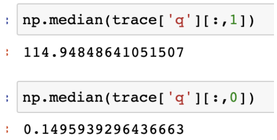


Figure 13: Posterior Results for Gaussian Mixture model of q_i, the rate of arrival of fatalities for the Poisson Process



```
: np.median(trace['q'][:,1])
: 114.94848641051507
```

```
: np.median(trace['q'][:,0])
: 0.1495939296436663
```

Figure 14: Median of Results for Gaussian Mixture model Posterior of q_i, the rate of arrival of fatalities for the Poisson Process

As you can observe, highway fatalities arrived at a far higher rate for values in the higher distribution of infrastructure spending, and far lower for the lower distribution of infrastructure spending.

Next, we chose to compare the results of different forms of GLM's and non-parametric methods, specifically Poisson regression, Negative Binomial Regression, Linear Regression, and a Random Forest. The models performed as follows:

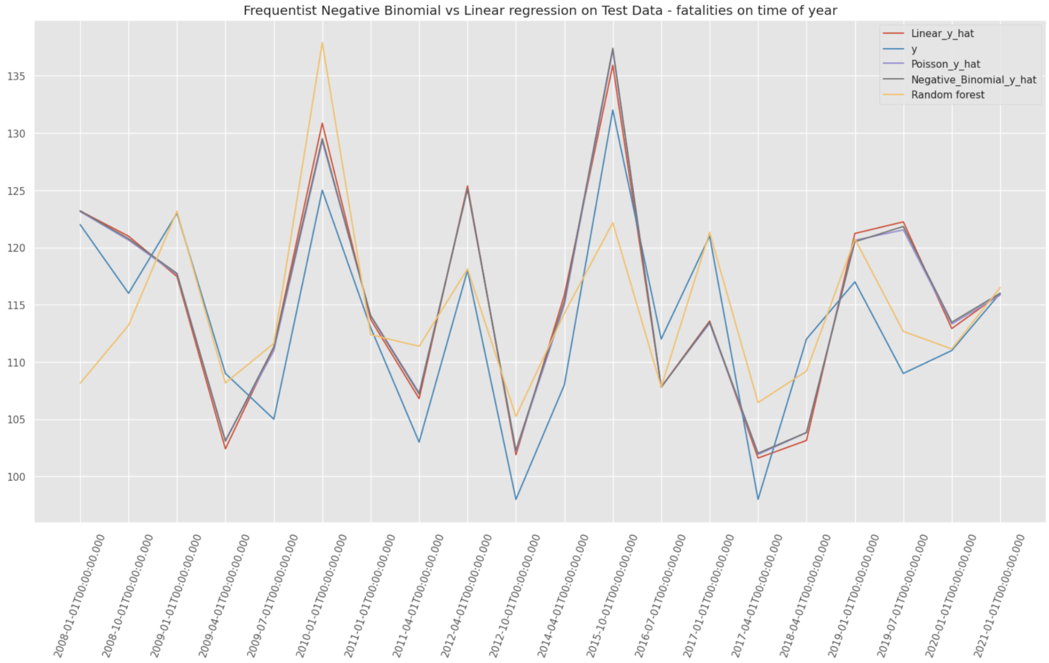


Figure 15: Model performance, graph

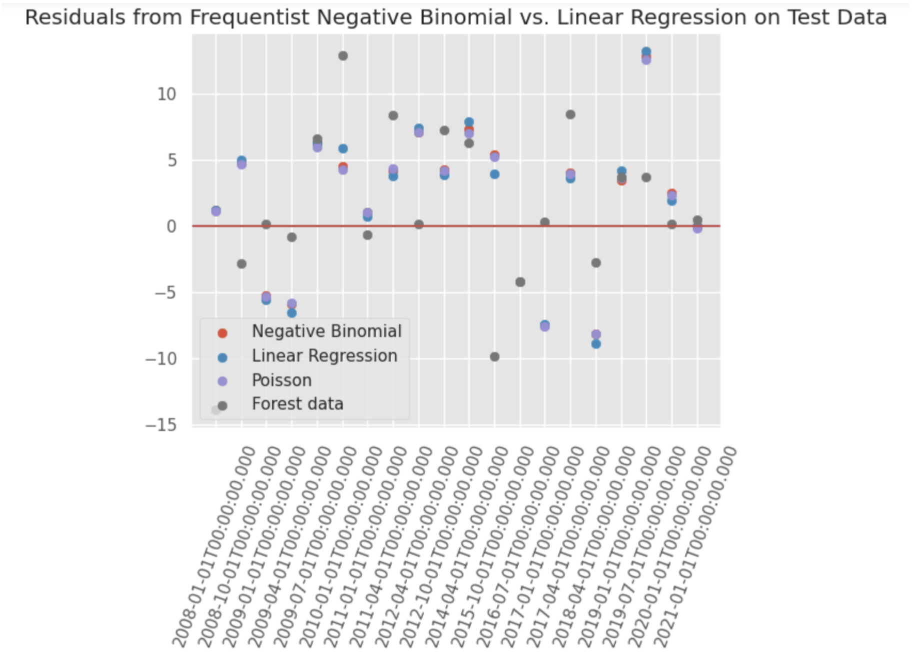
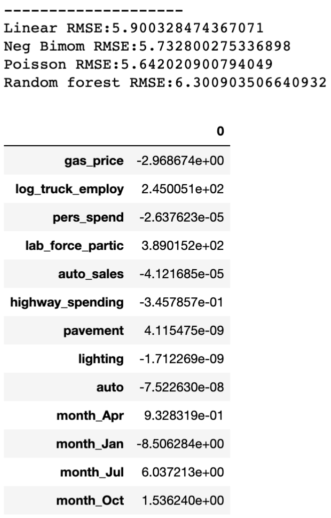


Figure 16: Model performance, residuals



```
--------------------
Linear RMSE:5.900328474367071
Neg Bimom RMSE:5.732800275336898
Poisson RMSE:5.642020900794049
Random forest RMSE:6.300903506640932
```

	0
gas_price	-2.968674e+00
log_truck_employ	2.450051e+02
pers_spend	-2.637623e-05
lab_force_partic	3.890152e+02
auto_sales	-4.121685e-05
highway_spending	-3.457857e-01
pavement	4.115475e-09
lighting	-1.712269e-09
auto	-7.522630e-08
month_Apr	9.328319e-01
month_Jan	-8.506284e+00
month_Jul	6.037213e+00
month_Oct	1.536240e+00

Figure 17: Model performance, RMSE and Poisson (best performing model) Coefficients

7 Conclusion

For Multiple Hypothesis Testing, after performing A/B testing on the six mobility categories, retail/recreation, grocery/pharmacy, parks, transit stations, workplaces, and residential, in the Google Daily Community Mobility dataset, two discoveries, grocery/pharmacy and parks, were made with an $\alpha = 0.2$ Applying Bonferonni correction with FWER=0.2, no discoveries were made, which we believe is due to the restrictive nature of the procedure. Performing the Benjamini-

Hochberg procedure resulted in one discovery, grocery/pharmacy.

As mentioned in our EDA, the final table used for testing was created by merging the Google Mobility dataset with supplementary data set containing monthly US GDP. Our findings for this research question were quite limited in scope as it was focused on just the state of California. Additionally, our study only considered the year 2020, when COVID was at its peak, and given that only monthly data was available for US GDP, this greatly limited the number of data points. However, we believe future studies can be conducted that builds on our work; incorporating more years, like the entirety of the 21st century instead of just 2020. For mobility, we did combine data sets, and in doing so, we did lose some granularity - one data set had daily data, while one only had monthly data. In order to merge these data sets, we averaged the daily data, which allowed us to see broader comparisons, and lost some daily noise.

Based on our results, COVID ironically did not have a significant impact on many categories of changes in residential mobility. Given the type of correction (or lack of correction) applied, only one to two discoveries were found. Nevertheless, these discoveries still imply that changes in mobility may be correlated with changes in GDP, which, if moved in the negative direction, can have impacts on the livelihoods of citizens.

For the Gaussian Mixture Model, we must be careful in interpreting these results too. From our model, we observed that the lower distribution of infrastructure spending was associated with an arrival rate of fatalities of ~ 0.15, whereas the higher distribution of spending resulted in an arrival rate of ~ 115. This suggests that more infrastructure spending results in far higher fatalities. While this may be true, this doesn't imply causation. Instead, from EDA, we explored how higher infrastructure spending was associated with higher economic stimulation, such as a higher labor participation rate, higher trucking participation rate, and higher level of personal spending. Each of these appear to individually contribute to higher highway fatalities, as seen in our regression analysis. So, it is hard to draw any specific conclusions and perhaps requires a more complex hierarchical model to control for these confounders. One suggestion we could possibly give is to focus on highway infrastructure spending in recessionary periods, in order to control for highway fatalities caused by economic booms. Or, perhaps these assumptions are wrong, and that highway infrastructure spending distracts drivers, or causes more construction on the roads, and therefore causes more highway fatalities.

After exploring highway fatalities, given the data, we are confident in assuming highway fatalities follows a Poisson Process. When interpreting these results, we must be careful as we have not used causal inference to determine any causal link. However, our team believes we use enough confounders to speak in general relationship terms. We see that a $1 increase in gas price corresponds to a fall of 1.47 traffic fatalities per 100 million vehicles. This in general makes sense, as the logic follows that higher gas prices means less people drive, drive not as far, or drive slower. Next, labor force participation has a large effect on traffic fatalities, with every percentage point increase in labor force participation corresponding to 280 traffic fatalities per 100 million vehicles. This follows as well, as a significant portion of traffic frequency corresponds to how many people are having to drive to work. The other values, such as month coefficients, are most likely linked to weather differences/activity differences. The primary take away from each feature is that they are linked to the Poisson arrival process, specifically our parameter for the Poisson distribution. Each feature has its own effect on the rate of arrival of cars on the road, except for infrastructure spending, and that our model is primarily based on the assumptions of a Poisson Process.

Our Poisson Process followed the following structure:

$$\lambda_w e = exponential(\beta \chi)$$

$$\chi_w e\tilde{}Poisson(\lambda_i)$$

Unfortunate bugs or incompetence with running PyMC3 prevented us from modeling using all these features in a Bayesian approach, so we took a Bayesian Poisson approach to modeling based on time of month and gas prices:

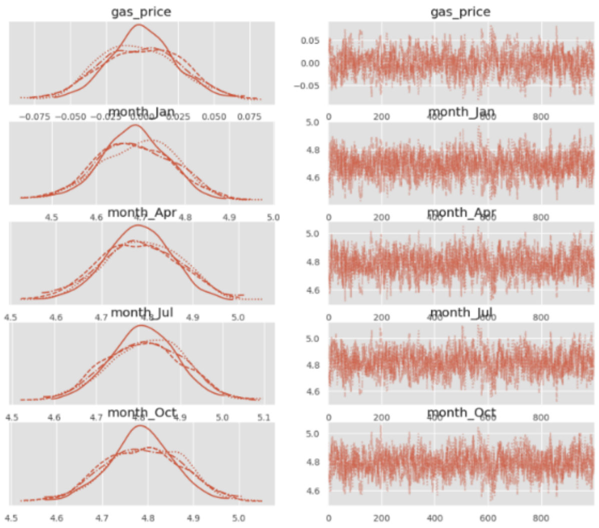


Figure 18: Bayesian Poisson Regression

with the posterior predictive check:

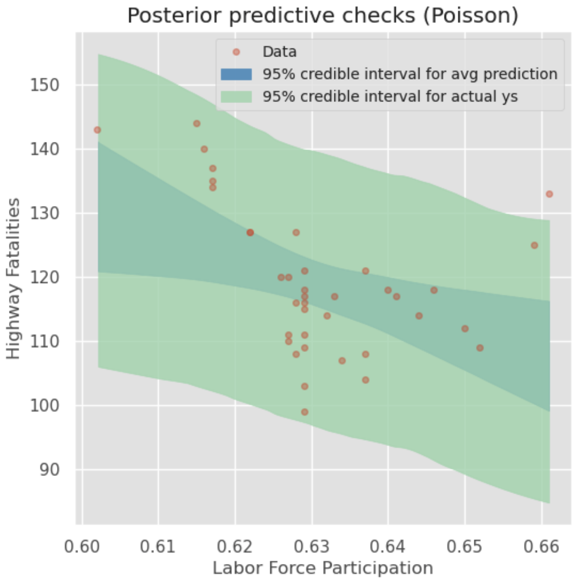


Figure 19: Bayesian Poisson Regression Posterior Predictive Check

References

[1] St. Louis FED. *Leading Indicators OECD: Reference series: Gross Domestic Product (GDP): Normalised for the United States.* URL: `https://fred.stlouisfed.org/series/USALORSGPNOSTSAM`. (accessed: 12.06.2022).

[2] Google Research. *Google: Daily Community Mobility Data.* URL: `https://www.google.com/covid19/mobility/`. (accessed: 12.06.2022).

[3] Ramond Robinson. *Bureau of Transportation Statistics: Monthly Transportation Statistics.* URL: `https://data.bts.gov/Research-and-Statistics/Monthly-Transportation-Statistics/crem-w557`. (accessed: 12.06.2022).

[4] Michael Warren. *Mobility Changes in Response to COVID-19.* URL: `https://arxiv.org/pdf/2003.14228.pdf`. (accessed: 12.06.2022).