ANALYSIS OF THE TRANSPORTATION SPENDING ON UTILIZATION RATES IN THE UNITED STATES

Prepared for

Data 102 Course Staff
University of California, Berkeley

Prepared by

Alexander Lee, Russell Algeo, Santhosh Mathialagan, and Liya Zhang

Analysis of Transportation Spending on Utilization Rates in the United States.

Data Overview

The United States Bureau of Transportation Statistics keeps a thorough record of transportation spending and investment across sectors. In addition, Google also has several publicly available data sets on Daily Community Mobility. Finally, we used outside data sources from the CDC to gather information about COVID-19 daily cases in the United States.

Research Questions

In this project, we sought to analyze the effect of transportation investment on human mobility. Our first research question was how changes in human mobility impact trends in transit stations utilization. To do this, we first created a generalized linear model to predict percent changes in transit stations based on various other percent changes in mobility. We also used nonparametric models such as neural networks to further analyze these relationships. Our second research question was: Is there a causal relationship between transportation investment and transportation utilization? We used causal inference techniques such as outcome regression to analyze the effect of transportation investment on flight data.

EDA

To address our first research question of how changes in mobility patterns affected the utilization of transit stations, we compiled Google's dataset of mobility trends in the United States from 2020 - 2022. We truncated our data to the last 3 years in order to ensure our dataset contained only the most relevant trends and to limit unnecessary computational expenses. For data cleaning, we removed any entries with missing values for one or more mobility trends. This was done to avoid skewing any of our regression coefficients, and our interpretations of them, by mistakenly regressing on incomplete entries. Since the data was provided neatly as whole real number percent changes in mobility trends, no further data manipulation was required. For one of our EDA components, we analyzed the distribution of our target/response variable, transit station usage, across states in the USA. Using the GeoPandas library, we created a map to visualize whether transit station usage increased or decreased, on average, over the three year period (Figure 1). Figure 2 is the same plot, but with the average change in transit usage by state as a continuous variable. Figure 3 is a high level bar graph to visualize the distribution and specific magnitude of average transit changes by state, across the board. We noticed that more inland and traditionally rural states tended to actually see increases in transit usage. Furthermore, all coastal states besides South Carolina saw decreases in transit usage. The differences and patterns in transit usage from state to state motivated us to answer this research question using GLMs, and to try to find out exactly how other mobility trends were correlated with transit and how they could be used to predict future trends in transit usage.

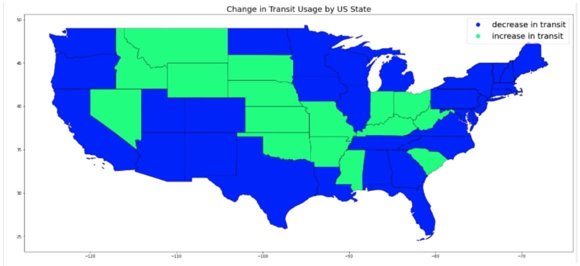


Figure 1: Direction of Change in Transit by State

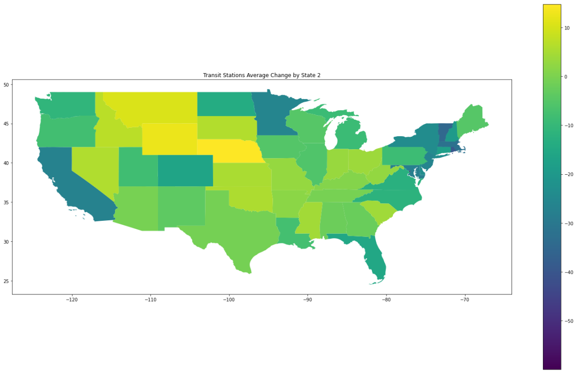


Figure 2: Average Change in Transit by State (%)

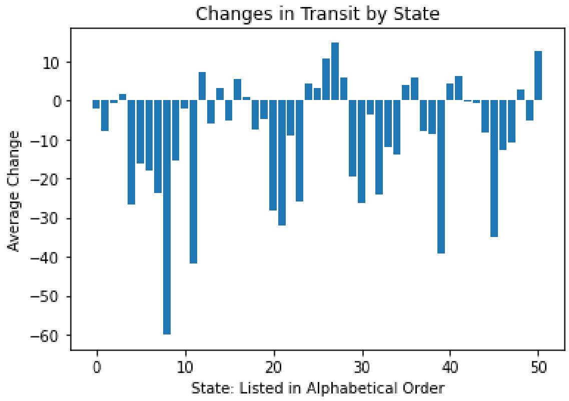


Figure 3: Distribution of Average Changes in Transit by State (%)

In Figure 4, we created a heatmap summarizing various correlation coefficients between columns in the Google Dataset. Here, the darker colors represent a higher (Red for positive, blue for negative) correlation coefficient between variables.

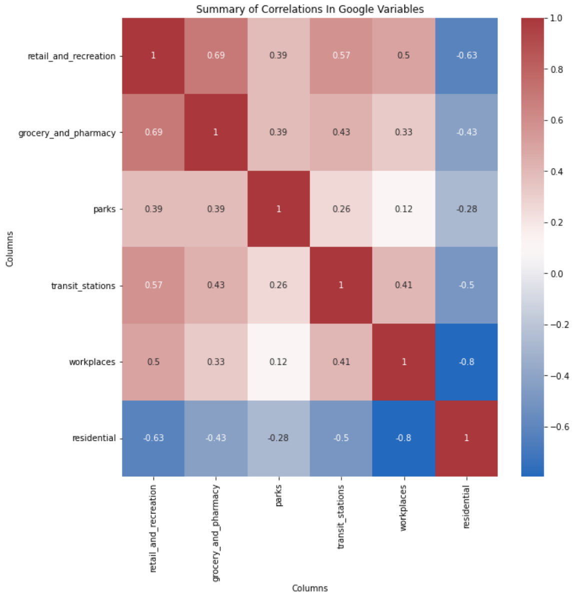


Figure 4: Summary of Correlations In Google Variables

We notice that travel to the workplace and residential areas have a very strong negative correlation. Residential and retail and recreation have a moderate negative correlation. Transit stations as well as workplaces have a moderate positive correlation with retail and recreation. Grocery and pharmacy have a strong positive correlation with retail and recreation. Given that we are looking to predict transit stations' percent change from other variables, the correlations (both positive and negative) with this variable are promising.

Next, we visualized both the Google Transportation data set and the Government Public Mobility Data Set together, which is the main data set for question 2.

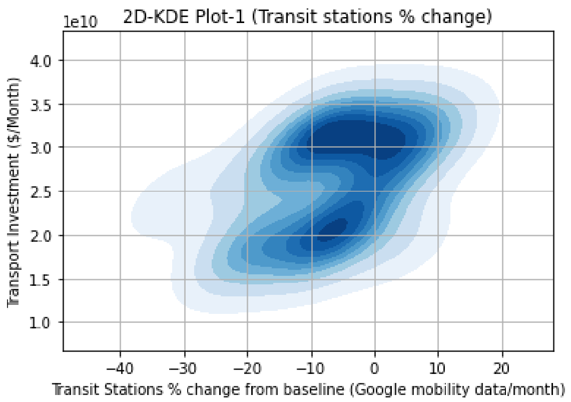


Figure 5: 2D-KDE Plot-1

There are two high density areas at $x =- 10$ to $- 5, y = 1.75$ to 2.25 and $x =- 5$ to $0; y = 3$ to 3.5.

The shape of the distribution (positive slope) does suggest that as the Investment in transport is increased, there is an increase in the value of transit_stations_percent_change_from_baseline (from google mobility data). For instance, for a Transport investment of $1.5e10$ the value of transit_stations_percent_change_from_baseline ranges from $- 30$ to 5 whereas for an investment of $3.5e10$ it ranges from $- 20$ to 15. So both the minimum and maximum values of the range increase.

Thus a positive correlation is seen between the two metrics and we can say that we expect to see a positive coefficient for predicting the corresponding regression coefficient (beta)

To obtain the Transport Investment (dollar/month) for all regions combined we directly added the data of following columns from the government dataset (State and Local Government Construction Spending -) - Bridge - Lighting - Pavement - Highway and Street - Power - Dock / Marina - Water - Mass Transit - Land Passenger Terminal - Land - Runway - Air Passenger Terminal - Air - Transportation. The mean of all values of "Transit stations % change from baseline" in a month across all regions is taken to obtain the value of "Transit stations % change from baseline" for a particular month for all regions combined.

In case data for a particular region is missing in the Google dataset or if it has not been considered in the government dataset (in dollar/month calculation), the inferences might not be true for that region.

The positive correlation between the two variables, supports a causal relationship between transportation investment and transportation utilization (thereby validating our research question). This is because an increase in the value of "Transit Stations % change from baseline", implies an increase in usage of public transport hubs such as bus/subway/train stations and thus increased utilization of transport when the government investment in transportation is increased.

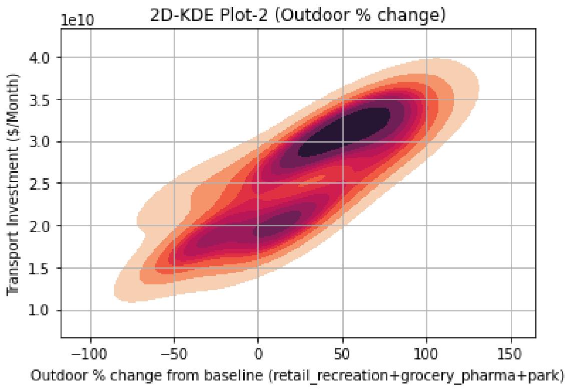


Figure 6: 2D-KDE Plot-2

There are two high density areas at $x =- 25$ to 25, $y = 1.75$ to 2.25 and $x = 50$ to 75, $y = 3$ to 3.25.

The shape of the distribution (with a positive slope) implies that as the Investment in transport is increased, there is an increase in the value of total_outdoor_percent_change_from_baseline (calculated from google mobility data). For instance, for a Transport investment of $1.5e10$ the value of total_outdoor_percent_change_from_baseline ranges from $- 75$ to 50 whereas for an investment of $3.5e10$, it ranges from 0 to 125. So both the minimum and maximum values of the range increase with increased government investment in Transportation.

Thus a strong positive correlation is seen between the two metrics (one obtain from google data and the other obtained from government data) and we can say that we expect to see a positive coefficient for predicting the corresponding regression coefficient (beta)

To obtain the Transport Investment (dollar/month) for all regions combined we used the same procedure as described in Analysis for Figure 5.

To obtain the Outdoor percent change from baseline for all regions combined we added the percent change from baseline values in Retail & Recreation, Grocery & Pharma and Parks. Thereafter, the mean of all values of "Outdoor % change from baseline " in a month across all regions is taken to obtain the value of "Outdoor % change from baseline" for a particular month for all regions combined. This variable captures the increase/decrease in mobility trends for all other outdoor places except Workplaces and Residential.

How is this plot relevant to our research question? The idea is that if the government increases investment in Transportation, people would get better transportation facilities and thus would like to use them more often. Where would they go more frequently using better transportation facilities? To one of these outdoor places. Thus, if they are visiting these outdoor places more often on increasing investment in Transportation, it implies increased Utilization of transport validating our research question. Note that there is no compulsion for people to visit these outdoor places (like a workplace where people have to go every day irrespective of whether they like it or not since they are getting paid for the job). That's why we have not included the workplace in the calculation of "Outdoor % change from baseline".

The positive correlation between the two variables (from the KDE plot) supports a causal relationship between Transportation investment and Transportation utilization (thereby again validating our research question).

However, we need to be mindful of the fact that causality cannot be concretely established from Observational studies as there might be confounding (hidden) variables. So to eliminate that possibility, we need to consider other possible variables like Covid cases, time of the year, shutdown, etc which might also be the cause for the increase in "Outdoor % change from baseline" and not the Increasing Government investment in Transportation (which might be a coincidence).

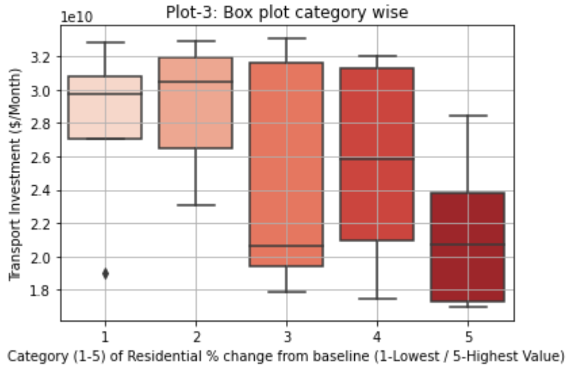


Figure 7: Box plot category wise

There is a clear downtrend in the box plots. So on Decreasing Transport Investment, there is an increase in the Category of "Residential % change from baseline". It means people are staying at Residential places more frequently when investment in Transport is decreased (Dollar/Month).

So the plot suggests Higher the government investment in Transportation, people will stay in residential places less frequently as they probably use the better transportation facilities to stay more frequently at non-residential (outdoor) places. This again supports our research question that there is a positive causal relationship between transportation investment and transportation utilization.

To obtain the Category(1-5) of "Residential % change from baseline", we calculate the ranks of all the values, and then the lowest values are assigned Category 1 and the next set of lowest values are assigned Category 2 and so on such that all categories get an almost equal number of data points. Finally the highest values of "Residential % change from baseline" are assigned to Category 5. This way we transform the "Residential % change from baseline" variable from continuous to Categorical with 5 different categories (1-Lowest to 5-Highest value).

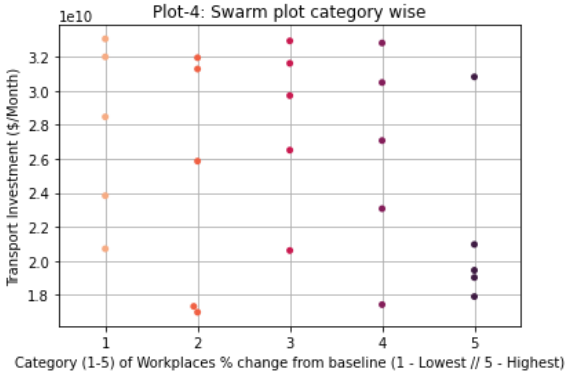


Figure 8: Swarm plot category wise

From the swarm plot, we observe the data points are all over the place and there is no clear trend.

This is in line with our expectations. Increasing investment in transportation should not affect mobility trends for the workplace. This is because people visiting these workplaces are most probably employees who are getting paid to visit the workplace every day. So regardless of whether the transportation facilities are good or bad, they are forced to use them and visit the workplace as it is a part of their job.

Thus this plot suggests that increased transportation investment might not imply

Increased transport utilization (there might not be a causal relationship) if the transportation is used to commute to workplaces only.

However, if the transportation facilities are being used to commute to other outdoor places (excluding workplaces), we already saw that there might be a strong positive causal relationship between Transport investment and Transport Utilization.

Note that the Category (1-5) of "Workplaces % change from baseline" has been calculated in the same way as the Category (1-5) of "Residential % change from baseline" has been calculated in part-3 (Before analysis of Figure 7).

For our second research question, our first step was to convert all the columns to the correct formats needed, such as making the date column into date time form. One issue with the US Bureau of Transportation Statistics data set is incomplete rows during earlier years. For example, the dataset contains around 100 years worth of data, but complete rows do not begin to appear until 2018. Thus, we decided to only focus on data from January 2020 onwards where we had complete information in the data set on both investment and spending. For this question, we decided to focus on just one sector and chose the airline industry. Thus we only used a subsection of our data set on investment and utilization within the airline industry. In addition, we added government data on COVID-19 cases. Because our transportation data was at the level of months, we needed to aggregate the COVID data to the levels of months.

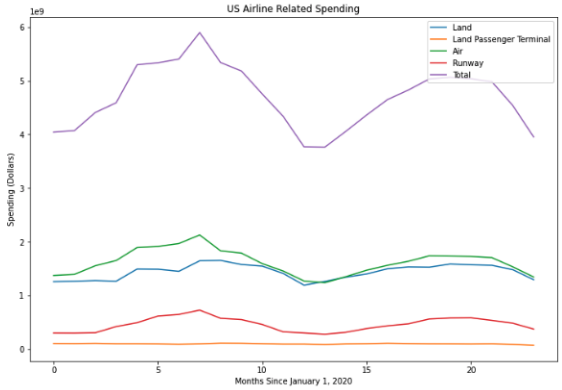


Figure 9: Investment in the Airline Industry 2020-present.

As shown in Figure 9, We notice that spending falls greatly in all sectors–besides land passenger terminals–starting at about 5 months after January 1st, 2020. It is likely that the reason land passenger terminals do not fluctuate is that the data used is only for about 2 years, so it is unlikely that land passenger terminals experienced major renovations or investments in this short of time (i.e. this spending is not a yearly or frequent expense). For the other categories of spending, there is a steady decline until about 11 months after Jan 1, 2020. After this, spending begins to rise until about 19 months after, when it peaks again and begins to fall. Given that we are looking to deduce a causal effect of transit investment on usage, this visualization is relevant because now that we know what trends exist in investment, we can compare these to trends in usage to understand their relationship.

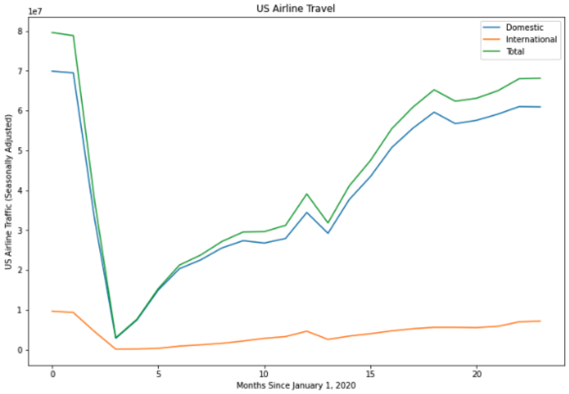


Figure 10: US Airline Travel since January 2020.

As shown in Figure 10, We notice that US domestic airline travel falls greatly from 0 months after Jan 1, 2020 until about 2 months after. This trend is mirrored in US international airline travel as well. After both reach a minimum, they slowly increase over the next 18 months. The green line depicts the summation of both domestic and international travel.

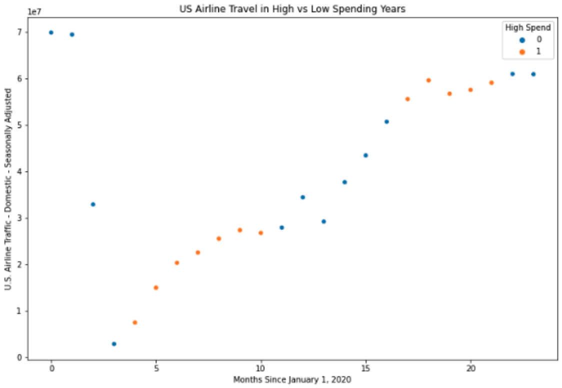


Figure 11: US Airline Travel in High and Low Spending Years

As shown in Figure 11, we notice that initially after January 1, 2020, spending is less than average. Spending then rises to above average from 3 months to 8 months after January 2020. It then once again falls below average from months 9 to 14, and rises to above average from months 15 to 20. Lastly, it falls beneath average from months 21 to 22. When spending is below average, it appears that the rate of growth in domestic airline travel is shrinking. On the other hand, in the months of high spending, the rate of growth is large and growing. This visualization is relevant because it can help us understand how investment affects airline travel if we look at above average vs below average spending months.

Prediction with GLMs and Nonparametric Methods

Frequentist GLM

To predict how mobility trends affected transit usage, we chose to use Frequentist and Bayesian Generalized Linear Models and a Neural Network. GLMs were a good fit for answering this research question because there are multiple predictor variables (mobility trends) and results of GLMs are easy to interpret. GLMs give us a clear understanding of how each of the individual mobility trend variables influence transit station usage in an unbiased manner. Our predictor variables/features are the percent changes in the traffic volume of retail and recreation, grocery and pharmacy, parks, workplace, and residential sectors. Our outcome variable is the percent change in transit station usage. The assumption that we are making here is that there is a linear relationship between our predictor variables (mobility trends) and our outcome variable.

Statistic	**Value**
Log-Likelihood	-1.7852e+06

k-Fold CV MSE	494.58

Table 1: Frequentist GLM Model Checking

Feature	0.025	coeff	0.975
Retail & Recreation	0.566	0.572	0.575
Grocery & Pharmacy	0.139	0.147	0.15
Parks	0.022	0.023	0.024
Workplace	0.143	0.152	0.162
Residential	-0.819	-0.787	-0.773

Table 2: Bootstrapped 95% Confidence Intervals

For our Frequentist GLM, our log likelihood was astronomically low, meaning that our GLM did not fit the data very well. Our k-Fold cross validation mean squared error was around 495, which meant our model had very high bias in a prediction setting as well. We think this is due to confounding factors outside of our dataset playing a role in how various mobility trends may affect transit usage from region to region. One such example is the regional discrepancies in the quality of public transit infrastructure that would cause a predictor like workplace mobility trends to be positively correlated in cities where public transit can easily and effectively be utilized for workplace commuting, and negatively correlated in rural places where commuting can only be done via personal vehicle, and increased workplace attendance simply detracts from opportunities to use public transportation.

Although a prediction task may be impractical with this model, there is tremendous consistency with the coefficients for our predictor variables on average over 100 bootstrapped GLM coefficients. In our 95% confidence intervals for these coefficients, none of our lower bounds for positive coefficients are negative and none of our upper bounds for negative coefficients are positive. Furthermore, all but one of our CI intervals are within a 0.02 percent span between bounds. This means that we can be confident that variables with positive coefficients (Retail & Rec, Grocery & Pharmacy, Parks, and Workplace) are positively correlated with transit usage and variables with negative coefficients (Residential) are negatively correlated with transit usage.

Bayesian GLM

We implemented our Bayesian GLM in Bambi using a Normal/Gaussian likelihood function. Because we do not have any domain knowledge regarding mobility trends and their impact on public transit, it would have been inappropriate to impose new prior distributions. Our results and 95% credible intervals (Table 3) were essentially the same as our Frequentist confidence intervals, so we know our Monte Carlo Markov Chain sampled our posterior distributions accurately (Figure 13).

Feature	2.5%	mean	97.5%
Retail & Recreation	0.568	0.573	0.580

Grocery & Pharmacy	0.138	0.145	0.152
Parks	0.022	0.023	0.024
Workplace	0.144	0.152	0.158
Residential	-0.808	-0.787	-0.768

Table 3: 95% Credible Intervals

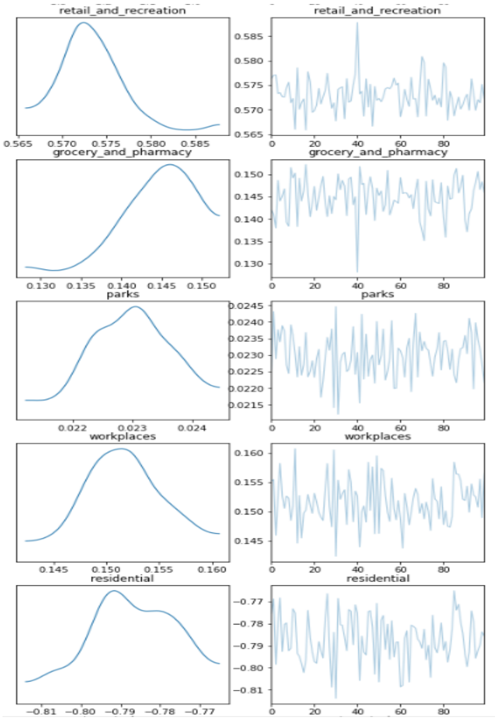
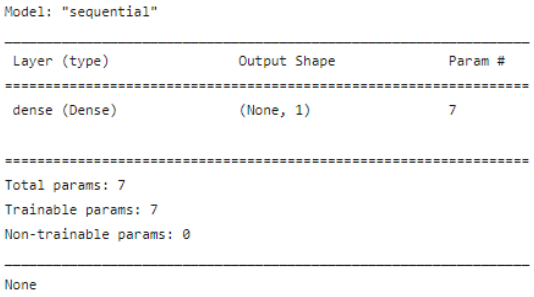


Figure 13: Bayesian GLM Posteriors

Nonparametric Model: Neural Network

k-Fold CV MSE	477.92

Our nonparametric model, a Neural Network built using the Keras package, a high level implementation of TensorFlow NN tools, featured a single hidden layer with a ReLU activation function. We selected a ReLU activation function because our model was designed as a regression task with a continuous, real outcome variable. This is also why we chose mean squared error to be our loss function, upon which we used the ADAM gradient descent optimizer to calculate our weights. We only used a single layer due to computational constraints during the backpropagation algorithm, with such a large dataset of our size. Because this is a nonparametric model, and specifically a neural network, it is incredibly difficult to create interpretations of the predictor variables' effects on the outcome. That being said, our k-Fold MSE was 477.92, so this model has better promise for prediction tasks than for leveraging insights.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 1)                 7

=================================================================
Total params: 7
Trainable params: 7
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 14: Keras Model Results

Causal Inference

Statistic	**Value**
Log-Likelihood	- 437.67
Total Spend Coeff	- 0.0148
New Case Coeff	- 1.5858

Table 4: Summary of Results from Outcome Regression Model 1

For our first model (Summarized in Table 3), we used outcome regression to analyze the causal inference of investment on travel.

We used an OLS outcome regression model because of our assumptions that our data was continuous and linear. We sought to analyze whether investing more into airports and related infrastructure for air travel have a causal effect on the number of passengers and airline utilization in general. Given that our data is from recent years, one big confounder on the data is COVID-19. In an attempt to deal with this confounder, we added in outside data for the number of COVID-19 cases at the time (from CDC). Our model had a log likelihood of -437.67 (average log likelihood of -18.23 per data points) and the confidence interval for our total spending coefficient contains zero, suggesting that this model is not a great model of causal inference.

Some of the assumptions that went into this model include that we assume the relationship between causal variables and outcome variables is linear, which may not be true. Additionally, our linear model assumes that each regressor affects total travel independently of each other, which may not be true. For example, the number of new covid cases per month may not act independently of the total investment amount (as suggested by the strong multicollinearity).

Statistic	**Value**
Log-Likelihood	- 429.27
Land Passenger Terminal Coeff	.8247
Land Coeff	- .0008
Runway Coeff	- .2493
Air Passenger Terminal Coeff	- .5299
Air Coeff	.3228
New Case Coeff	- .3.2929

Table 5: Summary of Results from Outcome Regression Model 2

In our second model, instead of using total spending, we trained the model with multiple regressors: one for each of the different spending metrics on airlines. Again, we noticed a poor log likelihood value (-429; -17.28 on average per data point). In addition, all of the individual regressors contain zero in the confidence interval, further complicating our analysis of causal inference.

Conclusions

To address our first question of how various trends in mobility patterns affect public transit usage, one of the models we developed was a Linear GLM. We assumed a linear relationship between our features and targets, and chose to use a Linear GLM because we are not working with counts or binary data. One of the limitations of this assumption is that the model cannot respond to other non-linear interactions or relationships.

In our Frequentist GLM, retail and recreation had the biggest positive correlation with our target variable, with a 0.572 expected percent increase in transit usage per 1 percent increase in retail and recreation. This makes sense because retail and recreation areas tend to be centered in highly populated areas and typically have lots of infrastructure surrounding them, such as public transit stations. On the other hand, our most negatively correlated predictor was residential trends, with an expected decrease in transit by 0.787 percent per 1 unit increase in residential trends. This was an unsurprising result, seeing as how increases in residential trends mean less travel overall. Our Bayesian GLM model echoed these findings exactly. For our neural network, we had the best predictive accuracy, but we faced computational constraints with such a large data set, limiting our hidden layer density, epoch size, and gradient descent batch size. If fully optimized, our findings indicate that a neural network would be the ideal prediction model due to its ability to capture non-linear relationships and higher order features with added layers.

For our causal inference model, it is possible that there are other confounders that are not controlled for in this situation which may be leading to the poor outcome regression model. For example, we attempted to control for COVID through cases but it is possible that other colliders may exist such as lockdown regulations. In addition, international travel may be affected differently because our investment data is only for the United States and not other parts of the world.

Part of the challenges with our model is we were inferring that the linear model described and explained the relationship between these variables; however, it is possible that this is a non linear relationship. In addition, we assumed that each regressor affected total travel independently of each other, which may not be true. Particularly, because we have a different causal variable for each category of airline investment, we are assuming they each have independent effects -- which is likely not true as each is affected by the total amount of investment in airlines allocated for that month. Furthermore, both models are based at the level of months due to the datasets that we were working with. In reality, it is likely that monthly spending on infrastructure does not explain the transportation utilization during the same month. For example, big investment projects may take multiple months for the construction or other improvements to be completed -- meaning the initial investment may come months before its effects are seen. It is possible that the completed projects may have a stronger effect on utilization, meaning that the effect of investment may be delayed by a few months.

Works Cited

U.S. Department of Transportation, B. of T. S. (2022, April 12). *Monthly Transportation Statistics: Tyler Data & Insights*. Bureau of Transportation Statistics. Retrieved May 9, 2022, from https://data.bts.gov/Research-and-Statistics/Monthly-Transportation-Statistics/crem-w557

Covid-19 Community mobility reports. Covid-19 Community Mobility Reports. (n.d.). Retrieved May 10, 2022, from https://www.google.com/covid19/mobility/

Centers for Disease Control and Prevention. (n.d.). *United States covid-19 cases and deaths by State over time*. Centers for Disease Control and Prevention. Retrieved May 9, 2022, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

102 Course Staff. (n.d.). *Lab 8: Estimating Causal Effects Using Unconfoundedness*. Lab08 - Jupiter Notebook. Retrieved May 9, 2022, from https://data102.datahub.berkeley.edu/user/russellalgeo/notebooks/sp22/labs/lab08/lab08.ipynb