# Modeling Report: Predicting Uber Rides in a post-lockdown world

**Problem:**

- Overall idea:
    - How can Uber maximize their ridership? Answer: By highlighting the features of Uber that BART fails to offer and by offering competitive rates during the most popular BART times. The busiest BART times are often the least safe, least comfortable, and slowest times to use this form of public transportation. If we can create a model to predict the number of people who will be taking BART on any given day, Uber can use this data to take advantage of their safety, comfort, and accessibility features and promote reduced fares to increase their ridership.
- Hypothesis:
    - We can create a model to somewhat accurately predict the number of riders on BART on any given day. This information would allow Uber to promote ridership by offering discounts on these busy days as a safer and more comfortable alternative. Uber should lower fares on Friday since that is the day with the highest consistent average number of BART riders. To maximize risk-averseness Uber could offer a discount only between 3pm and 9pm in order to minimize ride cancellations and to match the popular BART times.
- How will the hypothesis be confirmed or rejected?
    - The hypothesis will be accepted if our model can accurately predict, within a reasonable measure, the number of riders per day and reflects that Friday will consistently have the highest average number of BART riders. The hypothesis will be accepted if we can create a model that can predict 30 days of BART users with a 15% confidence interval and that the average number of riders on Friday are at least 1 person higher than the next closest day average.
- Can you confirm or reject this hypothesis in principle, assuming you have unlimited access to all data?
    - Yes, we can confirm or reject this hypothesis in principle assuming we have unlimited access to all data. The basis on whether we confirm or reject this hypothesis is whether the model accurately predicts the number of riders on any given day with a 15% confidence interval and if the average number of riders on Fridays are 1 person higher than the next closest day's average. So if we had unlimited access to all data we could confirm or reject this hypothesis.
- Can you confirm or reject this hypothesis with existing datasets? Or is it reasonable to expect the dataset exists?
    - Yes, we can confirm or reject this hypothesis with existing datasets. We believe that with existing datasets we can predict the number of riders on any given day with a 15% confidence interval and that the average of riders on Fridays will be higher than the next highest by at least 1 person.

- Consider a ”creative” data source or feature in the hypothesis.
    - Our entire EDA is based around a ‘creative’ data source, a BART dataset in which we provide essential information to Uber so they can influence people who would normally take BART to instead use their service. With this ‘creative’ BART dataset we begin by finding the average number of riders who ride from Downtown Berkeley to the Civic Center in SF per day. We will train our data to predict the average number of riders traveling from Berkeley to SF per day and compare different times of the day to see the most popular hours of the most popular days so that Uber can offer their promotion specifically during this time.

**Answer**:

- After we completed our EDA model through analyzing the BART data from Berkeley to SF, we were successfully able to predict the number of riders on a given day with a confidence interval of at least 15%. Our confidence interval, before making any initial improvements, was 18%  and our estimates of average number of Friday riders were consistently above 1 person (and still within a reasonable range) which means that both criteria were above our threshold for accepting our hypothesis, thus we can indeed confirm our hypothesis. To calculate this confidence interval we used a linear regression model and predicted the ridership for 30 days of BART from Berkeley to SF and subtracted this prediction from the actual values and repeated this to calculate the confidence interval of 18%. For our average number of riders for Friday we found the average number of riders per day of the week for the 30 days that we predicted and found that Friday was at least 1.2 people greater than the next closest day (Saturday). Thus after these calculations and analyses we were able to confirm our hypothesis.

**Modeling**:

- Mentions model to train
    - For our model, what we used was linear models, we used multiple linear models to visualize data and compare it, to notice the linear trends, basically utilizing time and intervals as quantifiers in the linear models in comparison to location and amounts of people.
- Describes inputs to the model
    - The inputs that we put into the model were ridership data from Bart, where we had data that gave us the amount of people that traveled from one bart station to another, and time. So the inputs to the model would be amounts of people, location and time, where time is broken into different scales, such as hour, day, month, and year, where locations are based on different bart stations.
- Mentions output
    - The outputs that we created with our data were different comparisons of the inputs, where we basically wanted to see certain information about ridership, like

we looked at average bart riders by hour per day of the week, average bart riders by hour on every day of the week, which we did individually charting out the average bart riders by hour for every day of the week separately, actual ridership across all Bart for the whole year of 2021 and then we used the temporal data and location data to compare the forecasted ridership versus the actual ridership during a given time period(11/3/21-12/2/21) going from Berkeley to SF.

- Includes an explanation of the model choice, potentially related to EDA or insight.
    - Our model choice was to use linear models because with linear models it is very easy to see and compare different outputs over time and most of our data deals with a temporal element where we are creating our outputs by comparing certain inputs over time, such as amounts of people, also taking into account location. When using linear models we can set the models to the same scale, which is useful for our specific EDA because what we want to do is notice ridership trends. So with the linear models in our EDA one of the things we did was make individual linear models for every day of the week, plotting out the 24 hours of the day versus the average riders, which allows for us to do two main things, one being that we get to see what time of the day people are riding bart the most on any given day of the week and then also being able to see on which day of the week there is most ridership. The linear models allow us to use our data and compare it on an even scale in an easily understandable manner. We went further with this and plotted the average bart riders per day of week so that we can compare this to the individual models and confirm the time periods with the most ridership and which days have the most ridership. Linear models were furthermore also useful in terms of scale for our EDA in terms of plotting and visual ease, where with a linear model you can easily plot multiple plots on the same chart, such as we did on the Forecasted vs Actual Ridership Plot and the Average Bart Riders by hour per day of week, because with linear models it is easier to compare these trends, as they can easily be represented on one chart in comparison to other model methods.

**Model Evaluation and Analysis**:

- Includes evaluation results for the model.

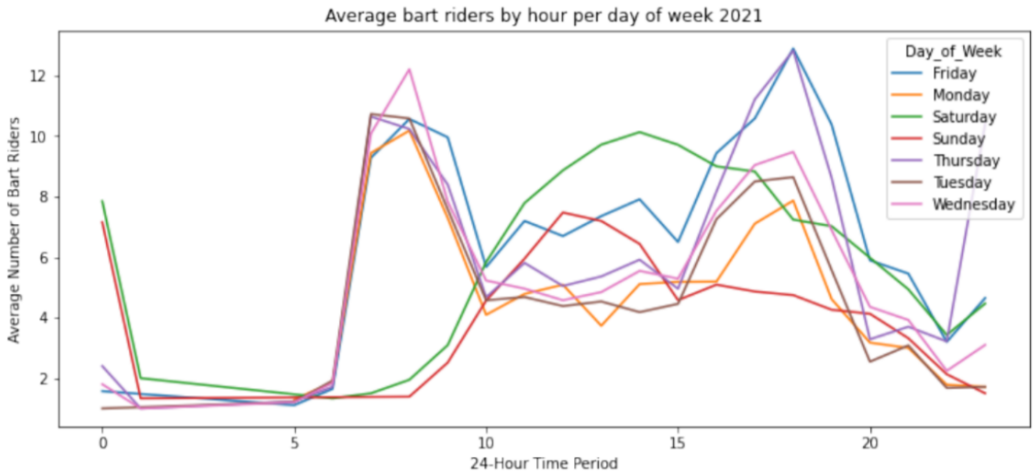
    In our task to create a model that can help uber somehow, we looked at Bart data, with one important output that we analyzed being the number of people who travel from Berkeley to SF and conversely from SF to Berkeley. We created multiple linear models and charts around this interest and found some interesting results. When looking at the average bart riders by hours during weekdays, we found that interestingly enough that on Friday's people travel most from Berkeley to SF on average. Our model not only showed the day in which there was the most travel, but also the time. The model showed us that throughout the week,

travel from Berkeley to SF was most common from around 15 a clock to 17 a clock, in military time which equals around 3-5 pm in layman terms. This means that with our models we were able to pinpoint not only the day, but also the time when there is most Bart ridership. We compared the forecasted versus actual ridership and found that this trend was actually very accurate, with the trend simply not reaching the maxima or minima, but the max days and times remained the same because our model correctly predicted the trend.

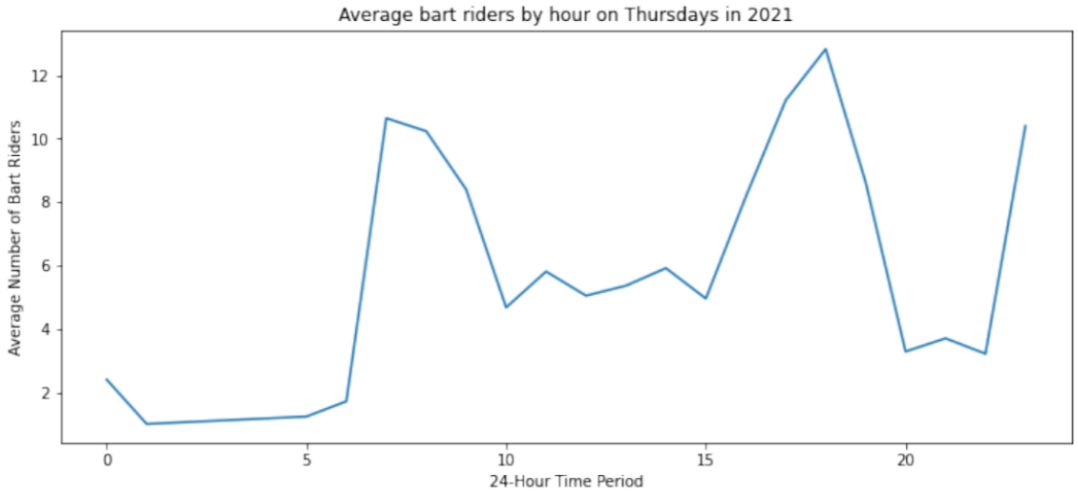
- Evaluation method is appropriate for the task and model.
    - For the evaluation of our results, the most helpful part of our modeling we found was that with the linear models we could have plots plotted on top of each other, which allowed for us to see how the predictions vary from the data. It allows for us to look at something important stated previously, which was the confidence, and what we found was that generally the confidence level was pretty low with a confidence of around 20 % at first, but as we adjusted our model a bit, it seems to change every time, and increase a bit but didn't dip lower than the 20% so that would be the base confidence level or the lowest confidence level, even though it wavers higher. We found that the trend lines look similar, and that in general this is a good model because it really captures the trend pattern. So if we have a base solid confidence level, which seems to be able to improve using the model we are using, then the method seems appropriate because it is an adaptable method that could be improved to better help with our EDA. It shows the trend and patterns easily, so it would be exemplary of the real data, as long as we reach the overall maxima and minima by improving the confidence level, then we could definitely have a very good model easily. Our model benefits us and our hypothesis because it allows us to create comparisons to uber, using the information that we have gathered on the trend of how on average throughout the week people use public transportation, specifically Bart, being able to pinpoint when down to the day and time. This can be beneficial because it allows for Uber to create strategic plans in order to lower their rates at this time during the max Bart ridership, offering competitive pricing in order to encourage more Uber ridership, as well as having the location data, this allows for Uber to increase marketing in these areas to lure in more customers. This would be beneficial because it reduces the clustering of people on Bart, which is beneficial, especially right now during times of covid.
- Includes an explanation of whether the model result is "good" or "bad".
    - In our EDA we found that this model is a good model but can definitely be improved , because the model seems to capture the overall trend fairly accurately, mimicking it very well, but did not exactly reach the overall maxima, nor the overall minima of the actual values. This means that the in some sense the model is not the best because since it doesn't reach the overall maxima or minima, then the confidence interval must be generally low, but since it captured trend and

pattern then definitely the model works pretty well because it is able to mimic the real data, this means that the model is not totally bad because our next task to take this model from good to great would be to work on increasing the confidence interval. Since we noticed the model was generally good, but could use some improvement, that is what we did in the following section, in an attempt to make the model better. Not only that but in the Future Work Section there can also improvement of the model.

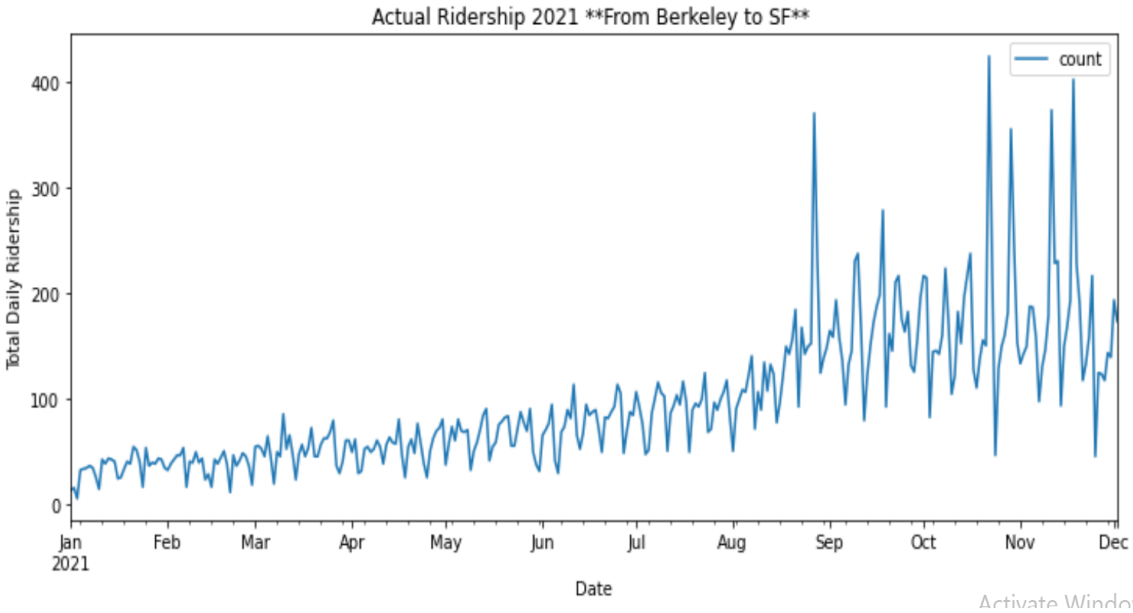
- Plots :
    - Included below are 4 examples of the plots we created using linear models in order to better understand our data and to better compare it. Each plot served a different purpose and helped us better understand the data and piece together the narrative the data is telling us.
        - The first plot depicts average bart riders by hour per day of week 2021, and what this plot shows us is a comparison of the average number of riders on a 24 hour scale, with each day of the week plotted as a different line. This allows for us to compare every day of the week to each other to see which day has the most ridership, as well as being able to analyze the data further and see specifically how many people are riding and at what times, allowing for us to see the day and time with most ridership




- The second plot is an example of one of the average bart riders by hour on any given day, which is basically where we plotted the average bart riders by hour on every day, but individually separated by day, so that we could better see the trends on individual days, allowing for us to better see the amounts of people and the time frame.

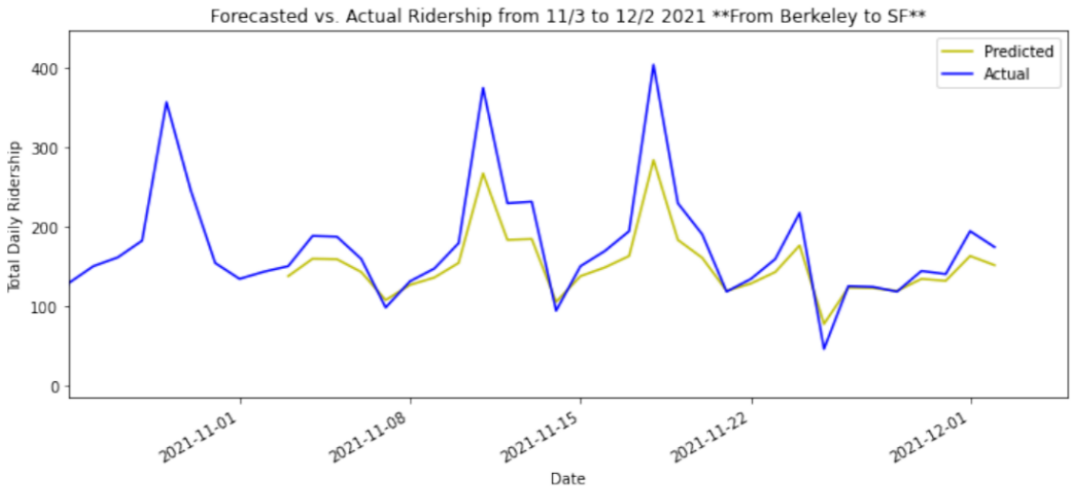



- The third plot is a plot of the actual ridership during 2021 from Berkeley to Sf. This allows for us to look at the ridership and analyze how throughout the year the ridership trends change, specifically being able to see which months there is the most and least ridership and take into account covid and the impact it had on the Bart ridership based on month, while also being able to see how many people ride Bart and what is an average amount of riders  and what are the extrema which allow for us to plan accordingly.




- The fourth plot takes the previous plot a bit further and serves to analyze our model, as it compares the forecasted versus the actual ridership from a given time period (11/3/2021 to 12/2/2021) from Berkeley to SF. This model allows for us to visualize the confidence level and see how accurately our model follows the trend in comparison to the actual data in

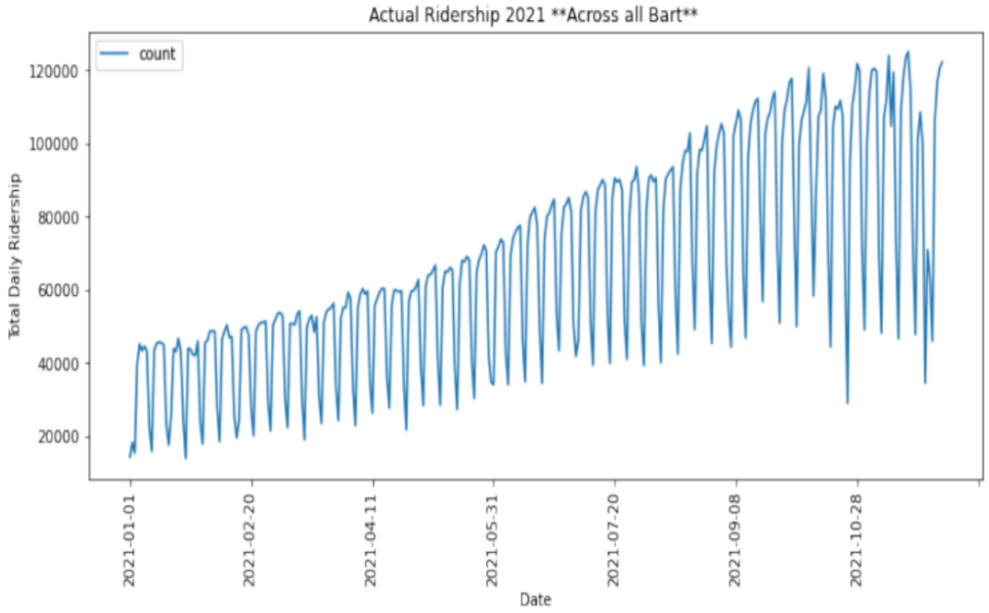
order to see if our predictions are correct, which in this case, yes they are generally correct.



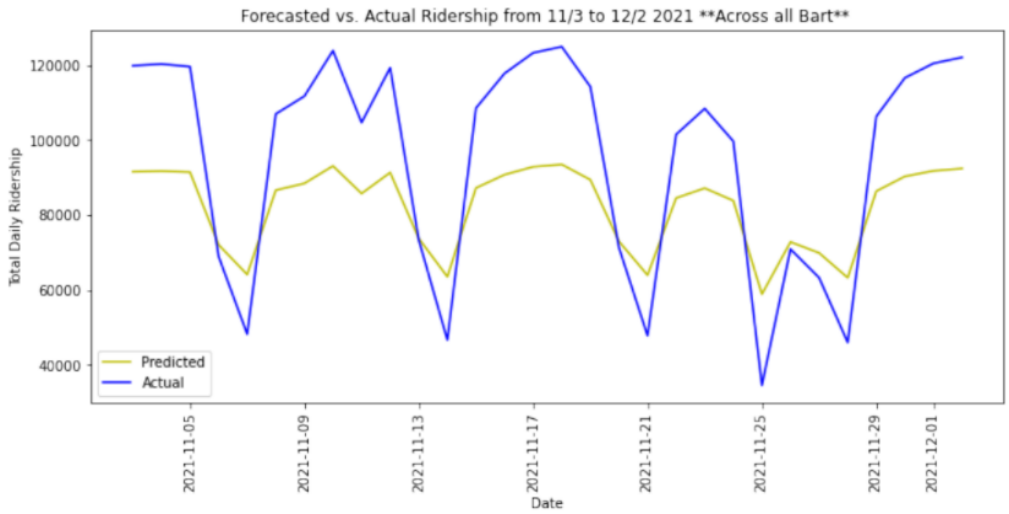

**Model Improvement**:

- Problem:
  - While we were able to create a model that could somewhat accurately predict the average number of BART riders from Berkeley to SF on a given day, an 18% confidence interval (i.e. accuracy) is pretty low (despite passing our threshold for confirming our hypothesis). We are only looking at a very small subset of the data, ridership from one BART stop in Berkeley (DBRK) to one BART stop in San Francisco (CIVC), which does not give us a great scope of the overall trend of BART traffic and Uber would be missing out on information on more BART traffic to increase their own ridership.
- Solution:
  - We believed that our accuracy was low because we were looking at averages rather than totals and because we were looking at a very specific trip (Berkeley to SF) rather than the entirety of all BART trips. Thus we made two improvements: First, we expanded the scope of BART usage to all BART rides, not just Berkeley to SF rides anymore; and secondly, we counted the total number of riders, rather than the average number of riders. These changes would allow us to take a step back and look at the patterns of BART traffic as a whole and hopefully improve our prediction accuracy. Observing the actual ridership across all BART would allow for us to more accurately predict ridership and allow Uber to reach a broader audience to which they can gain more users. Pictured below is the Actual ridership per day through January to November of 2021. As you can see as the year progresses, there is a linear increase in the number of BART riders; while if we compare it with the actual ridership during the same time period with just trips

from Berkeley to SF we would predict that the amount of riders before August would be much lower than they actually are.




- Result
    - After observing the entire dataset of trips, rather than just those from Berkeley to SF, we were able to predict 30 days of BART ridership with a 36% confidence interval (doubling our previous accuracy). We used the same method that we did in the initial EDA, before improvements, by splitting the data into training and test, then using linear regression to predict 30 days of BART ridership and comparing these predictions to the actual values. Similarly to the previous forecasted vs actual ridership graph discussed in the Model Evaluation and Analysis section, our predicted values fail to reach the relative maximas and minimas that the actual ridership reaches but indeed underlines the main trend/pattern of the actual data line plot. Pictured below is the Forecasted vs Actual Ridership from November 3rd to December 2nd (a 30 day period). As you can see the predicted does not exactly match the actual figures but it still outlines the general trend. One thing to note is that now we are comparing total number of riders while before we were comparing averages, which is why the forecasted line plot does not reach the same maximas and minimas as the actual line plot .




**Future Work**:

- Describe a direction for future work:
    - As for the direction for our future work, we would ask ourselves to delve deeper into the comparison between Uber and Bart and create more models comparing different values. Eventually we want to ask the question, is BART worth it? We want to really scrutinize Bart and Uber and compare the two meticulously, looking at the average price for one, two, or three people to take an uber and compare the travel time and price to Bart. Not only that but compare the different rider options, as Bart has discounts for the elderly, disabled and children, while Uber gives out occasional discounts and has staggering prices based on vehicle size and the luxury of the car. We could make our model better by taking into consideration holidays and how this affects Uber and Bart, as well as the accessibility of both transportations to the disabled or those with disabilities. Another thing wee could do, is factor in events that happen in SF that could cause a lot of congestion in terms of traffic or create moments where there is a high demand for access to transportation, such as Giants and Warriors games, concerts, Carnaval and Pride, as these days greatly impact the transportation systems. As we include this additional data in future work, it would also improve our model and make it more appealing as well in terms of information is can providing, because it would become more all encompassing in terms of import trends that allow for us to offer to Uber, which in turn they could use to create more opportunities to have special promotions, tailor promotions, and overall allow for Uber to increase ridership. In future work we would consider not only transportation between Berkeley and SF, but delve deeper into all possible routes for generalized traffic in the Bay Area. We would be able to generate a more all encompassing look at the ridership in 2021 in both Uber and Bart and make sure to enclude what is possible of 2020 in order to make our model more prepared to

deal with more possible lockdowns because this improvement would allow for us to look past just the one route, and think bigger, letting it be all data, looking at total riders per day across all routes, so Uber can better offer discounts and market on busy days to promote their services. Our end goal with our model would be getting it to the point where it could accurately predict ridership patterns in a complex way using the inputs in order to make an extremely accurate ridership model, which Uber could then use this ridership model to predict when to market and how to market their services. Our model would be extremely accurate and be able to estimate ridership precovid and post covid, even though the ridership dropped as covid happened for bart because this would also be something taken into account. Our model is already pretty good as is, being decently accurate and allowing for us to really complete this EDA, but in future would we can easily improve our model to yield even better EDA results.