MGT 6203 Team 16

Reducing rideshare rides to improve CO2 emissions in Austin Texas

Jeffrey Finucane, Justin Heinzekehr, Jingya Ye, Ting Sit, Warren Ehrenfried

Overview of Problem Statement

The study is to analyze if there is an existing relationship between the number of ride shares vs public transportation services. Do rideshare services serve as a complement to public transportation or cannibalize public transit utilization? What motivates a passenger to switch mode of transportation?

The analysis leverages the dataset of rideshares in the Austin, Texas area and the ride volume from CapMetro, the main bus service agency in the region. The outcome of the analysis will be used to provide recommendations to CapMetro to improve bus services, which could help attract more passengers to take public transit instead of using rideshares.

The main benefits of the study are:

- Reduce greenhouse gas emission by shifting mode of transportation to public transit, which has a 69% lower emission (Anair et al., 2020) than rideshares services
- Mitigate traffic congestion as a result of reducing number of cars on roads from rideshare
- Achieve 50/50 mode of transportation goal for Austin

Literature Survey

Existing studies on the relationship among rideshare and public transportation is inconclusive. While the fear of rideshare service may replace public transit usage sounds logical, the substitution effect may not be completely valid (Hoffmann Pham et al., 2019). In fact, a complementary effect has been suggested, which provides opportunities of integration among the public transit and ridesharing systems (Zhang & Zhang, 2018). Cats et al. (2022) argues that both substitution and complementary effects could co-exist depending on the circumstance, including the personal monetary evaluation of travel time, service accessibility, and costs, that transition to a different mode of transportation may not be voluntary. This argument agrees with observations from Hoffmann Pham et al. (2019) study that a rise in Uber and Lyft services leads to a decline in Yellow and Green Taxi ride volume, but no obvious decline in subway usage.

Nonetheless, some aspects of rideshare service show strong substitution effect, although the effect is uneven across various demographic factors (Pan & Qiu, 2022). Quantifying those effect factors could become the blueprint for public transit agencies to improve service quality to make public transit options more desirable for passengers. Increasing the service quality of public transit services often leads to an increase in ridership (Erhardt et al., 2022), and we assume natural migration of trips from rideshare as public services improve. Our study focuses on identifying such factors for Austin residents to provide CapMetro ideas on what investment area they can prioritize. Examples of factors as suggested by research study includes spatial disparities, availability (Cats et al., 2022), poverty and unemployment rates, minority (Pan & Qiu, 2022), cost, travel time, and trip purpose (Hoffmann Pham et al., 2019).

One area that studies agree on is the environmental benefits of public transit, which serve as the common motivation for conducting the research. Anair et al., (2020) paper that looks at CO2 emissions confirmed that public transit does indeed provide a more environmentally friendly approach to transit compared to ridesharing.

Data Sources

- Austin Ride Volume: This data set summarizes the total volume of rideshare on each day from June 16th, 2016 to August 31st, 2016. Number of rideshare will be the response to the model to be trained for. (https://data.world/andytryba/rideaustin)
- Rideshare Austin Data: This data set contains key information which includes geographic start and end locations of the rideshare, detailed vehicle list and weather condition for each registered rideshare. (https://data.world/andytryba/rideaustin)
- Vehicle Fuel Economy: This will help to get fuel consumption in both city and highway with CO2 emission on each specific vehicle. (https://www.fueleconomy.gov/feg/download.shtml)
- Capmetro Shapefiles: Dataset contains the geographic locations of current public transit stops in the Austin CapMetro public transit city. (https://www.capmetro.org/destinations)(https://data.texas.gov/Transportation/CapMetro-Shapefiles-AUGUST-2017/5d4c-snum)

Approach

1. Performing exploratory analysis on ride share data to identify potential independent variables which may explain the passenger behaviors of using rideshare. Variables may include hour of day, day of week, weather condition, distance traveled, duration, location, and costs. For each variable, perform a statistical test to determine if the change in distribution is significant on each variable. Also, perform tests to confirm the independence among the selected variables.
2. Calculating the distance of the n-nearest bus stops to each rideshare pick-up and drop-off point to test if we can account for some variance in our regression model through this factor even though our rideshare coordinate data is coarse and may not be accurate enough to act as a key indicator on choice of mode of transportation.
3. Performing similar exploratory analysis on bus ride volume and coming up with similar distributions as rideshare on variables identified before. Also run the same set of statistical test on variable independence and significance
4. Identify if the distribution among rideshare rides and bus rides are similar or contrast each other, which will help inform whether potential relationships exist. Perform statistical tests to assess if there are significant correlations among the two ride volumes information.
5. If all the statistical tests are passed, a linear regression model can be applied, and the equation could be like this:
   Rideshares ~ Bus Ride + independent variables identified in steps 1, 2, & 3
6. Once the model is established, a recommendation will be made on bus services that will factor in the independent variable and lead to an anticipated increase in bus rides. Rideshares volume will be reduced should the correlation be negative.
7. The reduced rideshare volume will be converted to a CO2 calculation model to estimate the greenhouse emission reduction once compared to original CO2 consumption.

Data Cleansing

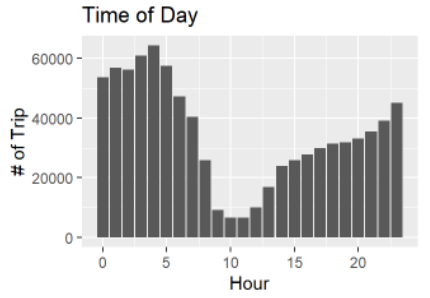
- Standardize all distance measure in km
- Remove rideshare trip > 20 km from the analysis by assuming such distance would be feasible to be traveled by bus

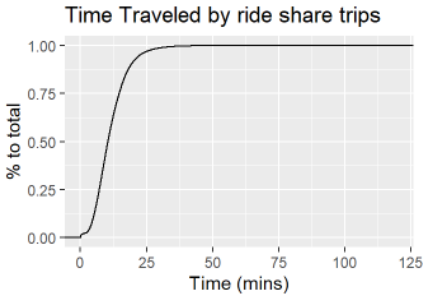
- Remove data which have start or end lat/lon information missing
- Remove rides that do not have a corresponding emissions data in Gov't file
- Select 1 emission profile per ride (sometimes multiple depending on if car is hybrid or two releases in the same year)
- Calculating the number of trips for each rideshare start and end location
- Using Euclidean measurement to find the average distance of the n-nearest bus stops for each pick up and drop off location

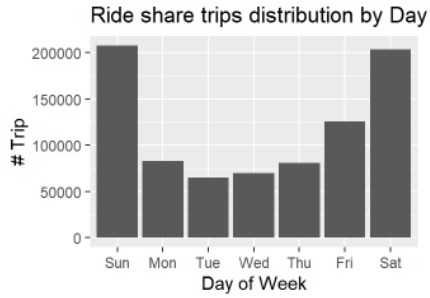
Initial Hypothesis

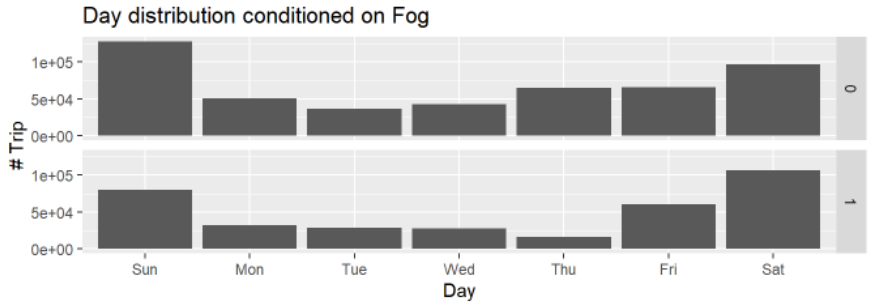
- Control on day of week, hour of day, and weather condition, **higher** bus ride volume lead to **lower** rideshare volume (negatively correlated)
- For the same to and from location, the **longer** the difference in traveled time among rideshare and public transposition, the **higher** the rideshare volume
- The shorter the distance between a pickup/dropoff location from a bus stop, the **lower** the share ride volume
- The **fewer** rides on the road through rideshare the **less** fuel consumed and **less** CO2 emissions (positive correlation)

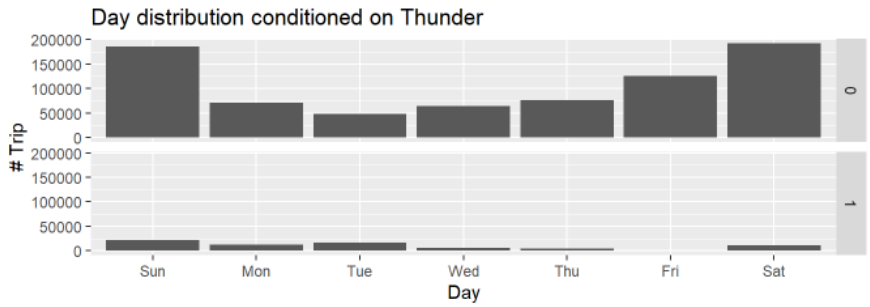
Exploratory Analysis











Observations on Rideshare dataset:

- Rideshares tend to be more populated around early morning time and late evening
    - Hypothesis to Test: public transportation service hours impact passenger choice
- Almost 90% of rideshare complete in < 25 minutes
    - Hypothesis to Test: travel time in public transportation impact passenger choice
- Rideshare happen more frequently over weekend

    - Hypothesis to Test: public transportation frequency during weekend impact passenger choice
    - Hypothesis to Test: rideshare over weekend is more effective in terms of time/cost relative to public transportation
- Do not observe significant distribution with/without weather conditions
    - Hypothesis to Test: Passenger preference to use rideshare is not impacted by weather conditions

Statistical Testing

Based on the exploratory analysis, the variables below are identified to have an impact on the volume on rideshare usage. Relevant statistical tests will be carried to confirm their significance.

Day of week, hour of day – apply further grouping on these two variables to reduce the number of category variables and further increase the prediction power of the time metric.

- Sunday & Saturday are labeled as weekend, while the rest of the days are labeled weekday, under new column *wkd*
- 10pm to 5am is grouped as "night time", 6am to 9am as "rush hour", 10am to 4pm as "work hour", 5pm to 10pm as "evening" under new column *hr_cat*
- Aggregated rideshare volume by hours and by days (obtain dataset rideshare_group), perform a linear regression to check if coefficient meet 5% significant level

```
Call:
lm(formula = vol ~ hr_cat + wkd, data = rideshare_group)

Residuals:
   Min     1Q Median     3Q    Max
 -5395  -1804   -165   1596   8426

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5774.3      380.3  15.183  < 2e-16 ***
hr_catevening    -3122.2      576.0  -5.420 2.11e-07 ***
hr_catrush_hour  -3075.0      618.8  -4.970 1.68e-06 ***
hr_catwork_hour  -5105.4      522.9  -9.763  < 2e-16 ***
wkdweekend        4722.1      456.6  10.343  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2673 on 163 degrees of freedom
Multiple R-squared:  0.5572,    Adjusted R-squared:  0.5464
F-statistic: 51.28 on 4 and 163 DF,  p-value: < 2.2e-16
```

The day of week and hour of day indicate **significant relation** with rideshare volume.

Weather Conditions (Fog, Thunder) - knowing the day of week has a relationship with rideshare volume, a Kolmogorov–Smirnov (K-S) test is used to check the ride share volume distribution by day week conditioned on whether Fog or Thunder exist. The null hypothesis is that the existence of weather conditions do not impact rideshare volume distribution by day. Due to the imbalance nature of the occurrence of fog or thunder, a downsample technique is applied to ensure the two distributions have the same number of data points.

K-S Test for Fog

```
        Exact two-sample Kolmogorov-Smirnov test

data:  no_fog$n and fog$n
D = 0.28571, p-value = 0.9627
alternative hypothesis: two-sided
```

K-S Test for Thunder

```
        Exact two-sample Kolmogorov-Smirnov test

data:  no_thunder$n and thunder$n
D = 0.2381, p-value = 0.9627
alternative hypothesis: two-sided
```

Weather conditions such as fog and thunder **do not appear to influence** rideshare volumes.

In order to know how these possible factors affect ridesharing volume, we need to first be able to compare the distribution of riders to a baseline - in our case, the usage pattern of public transportation during the same period of time.

Exploratory analysis revealed to us that the location data from RideAustin was reported in 0.01 increments of latitude and longitude, the equivalent of a square about six city blocks wide. However, in the densest area of Austin, bus stops are located closer than six blocks apart, and their locations are reported in 0.0001 increments. In order to meaningfully compare the two, we decided to group all rideshare trips and public transportation trips by latitude and longitude locations rounded to the nearest 0.01, rather than by specific bus station.

We merged the grouped ridesharing and public transportation datasets together on the latitude and longitude pair to see whether certain locations, such as the busier downtown area, were more in demand for both modes of transportation, or if the data was not actually correlated by the area of the city. Since we only have data for one non-profit ridesharing service instead of the whole ridesharing market, it was more meaningful to compare the *percent* of ridesharing trips and public transportation trips corresponding to particular latitude/longitude pairs.

The percent of ridesharing and public transportation trips by location were positively correlated and statistically significant (see regression result below). We can conclude that areas of the city in which more people use public transportation are also more likely to have a higher volume of ridesharing. Based on the adjusted R-squared value, about 35% of the variance in ridesharing volume across the city of Austin can probably be explained simply by the variance in overall demand for transportation to certain locations in the city (i.e., higher demand for more highly trafficked areas, and lower demand as you move away from downtown).

```
Call:
lm(formula = pctrideshare ~ pctpubtrans, data = merged_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.030596 -0.000995 -0.000537 -0.000130  0.077901

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0005940  0.0004053   1.466    0.144
pctpubtrans 0.7570990  0.0541688  13.977   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007205 on 363 degrees of freedom
  (545 observations deleted due to missingness)
Multiple R-squared:  0.3499,    Adjusted R-squared:  0.3481
F-statistic: 195.3 on 1 and 363 DF,  p-value: < 2.2e-16
```

If our goal is to figure out what factors affect people’s decision to use public transportation or ridesharing services, our goal is to find variables to add to this basic regression model that either reduce the correlation between the two distributions, or improve our ability to explain the variance between them.

Our first attempt, for example, was to add variables for the day of the week and hour of the day that a trip was made (measured at the start of the trip). This time, we grouped both datasets by latitude and longitude pair as well as by day of the week (1=Sunday, 2=Monday, etc.) and by hour of the day (12=noon, 13=1:00 pm, etc.). For the sake of simplicity, we added binary variables for weekday vs. weekend and for work hours (6 am to 6 pm) vs. off-hours. In future we could refine the regression model to account for specific days or hours if there appear to be significant variation aside from the weekday/weekend and business hours/off-hours split. Re-running the regression analysis with the two binary variables shows the following:

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0003066 -0.0000259 -0.0000111  0.0000020  0.0059577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.739e-05  1.508e-06   38.07   <2e-16 ***
pctpubtrans  2.577e-01  8.817e-03   29.23   <2e-16 ***
weekday1    -3.113e-05  1.501e-06  -20.74   <2e-16 ***
workhours1  -2.590e-05  1.397e-06  -18.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000117 on 29715 degrees of freedom
  (32438 observations deleted due to missingness)
Multiple R-squared:  0.04364,   Adjusted R-squared:  0.04354
F-statistic:   452 on 3 and 29715 DF,  p-value: < 2.2e-16
```

In this model, all variables were statistically significant, and the correlation between public transportation and ridesharing trips was much reduced. Even though the demand for public transportation in a particular area of the city was still the biggest predictor of ridesharing demand, the day of the week and hour of the day made a significant difference as well. People were less likely to use ridesharing services on weekdays and during business/commuting hours.

One possible interpretation of this result might be that people are more likely to use public transportation for regular trips, like commuting to work, when they know in advance exactly where they will need to be traveling and will need to make the same trip on a regular basis. Or, it may be that people want to use ridesharing services at times when public transportation is less

reliable, or at times of day when it feels safer to use ridesharing services than public transportation.

In future iterations, we plan to add other variables into our regression model to suggest more specific strategies for the city. One of the most important considerations will be whether the distance from a bus station affects how willing people are to use public transportation, which could tell us whether building new bus stations is worth the city's investment, and if so, by how much.

Another variable we would like to study is the impact of passenger's choices on commuting. Another binary variable to add into regression can be whether the route of public transportation is under construction or not. According to the mobility project (2016-2017) of the city of Austin(data.austintexas.gov, 2016), there are 17 projects which means 17 different streets which are under construction. It will be interesting to see the pattern in those areas whether it will have more rideshare trips or less especially compared between weekdays and weekends. If the result comes back that there is a significant decrease in public transportation volume then it will make more sense to add stops or bus trips nearby and also by how much when making recommendations to achieve the 50/50 goal.

Formation of Regression Model

Based on the above analysis, the linear regression model for this study is as follow

Rideshare Volume ~ Bus Ride Volume + Day of Week + Hour of Day + Distance to Bus Stop

Unexpected Problems

CO2 Calculation

- Some rideshare rides did not have corresponding Vehicle emissions data
    - 43% of data was NA values
    - Solution- due to large data exclude data points with NAs
- Multiple year, make, model vehicles had multiple rows of data in the government source
    - Solution- randomly select one type of car to be represented
- Multiple fuel sources from hybrids

- Solution- if a multi-fuel vehicle take average otherwise take value of designated fuel type

Rideshare Coordinate Accuracy

- The degrees of our coordinates for the rideshare data set are very coarse
  - (x, y) coordinates are only precise to two decimal places
  - Our bus stop data set coordinates are accurate to the 5th decimal place
  - Solution- calculate the distance of the coarse rideshare pick-up/drop-off coordinates to each bus stop location (over 2000 locations) and take the median or mean of the n-nearest bus stop locations
  - In the future may look to use a spatial binning method of bus stop locations to get a better representation of nearest bus stops weighted by number of stops in each bin to estimate the most accurate average distance

Hypothesis Testing

- Many hypotheses that set out to be tested may not be able to evaluate with the existing datasets. Additional efforts are need to gather more information but may not be feasible within the project timeline

Reference

Anair, D., Martin, J., Pinto de Moura, M. C., & Goldman, J. (2020, February 25). *Ride-Hailing Climate Risks | Union of Concerned Scientists*. https://www.ucsusa.org/resources/ride-hailing-climate-risks

Cats, O., Kucharski, R., Danda, S. R., & Yap, M. (2022). Beyond the dichotomy: How ride-hailing competes with and complements public transport. *PLOS ONE*, *17*(1), e0262496. https://doi.org/10.1371/journal.pone.0262496

Erhardt, G. D., Hoque, J. M., Goyal, V., Berrebi, S., Brakewood, C., & Watkins, K. E. (2022). Why has public transit ridership declined in the United States? *Transportation Research Part A: Policy and Practice*, *161*, 68–87. https://doi.org/10.1016/j.tra.2022.04.006

Hoffmann Pham, K., Ipeirotis, P. G., & Sundararajan, A. (2019). Ridesharing and the Use of Public Transportation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4099122

Pan, Y., & Qiu, L. (2022). How Ride-Sharing Is Shaping Public Transit System: A Counterfactual Estimator Approach. *Production And Operations Management*, *31*(2), 906–927. https://doi.org/10.1111/poms.13582

Zhang, Y., & Zhang, Y. (2018). Exploring the Relationship between Ridesharing and Public Transit Use in the United States. *International Journal of Environmental Research and Public Health*, *15*(8), 1763. https://doi.org/10.3390/ijerph15081763