MGT 6203 Group Project Proposal Template

Please edit the following template to record your responses and provide details on your project plan.

TEAM INFORMATION (1 point)

Team #: 16

Team Members:

1. Jeffrey Finucane; jfinucane (EdX)
   Graduated from Purdue University in 2021 with Bachelors in Industrial Engineering. Took multiple classes involving machine learning. Completed IE332 which required us to build an autonomous drone model that gathered air quality data in hypothetical cities. Also completed a regression analysis of aircoil propulsion and its effect on jet engines. I am currently a Logistics Analyst for the KraftHeinz company.

2. Justin Heinzekehr; HeinzekehrJ (EdX)
   Graduated from Goshen College with a BA in Religion and from Claremont School of Theology with an MA and PhD in religion. Now works as the Director of Institutional Research at Goshen College, which includes strategic analysis of enrollment, human resources, finances, and student learning assessment data.

3. Jingya Ye; yeeze215y(EdX)
   Graduated from University of Windsor with a master degree in Material Engineering. I have been working in the automotive industry as a manufacturing engineer since graduation. I am involved in building up the product's defect categories for cost saving and six sigma purposes for different plants. Production projection for different processes within the product life cycles and visualization with VBA for data presentation. Other projects include building up the network infrastructure to digitize information in legacy machines and linking in traceability to products that align with OEM requirements.

4. Ting Sit; sittingman(Edx)
   Graduated from UCLA with a degree in Economics and a minor in Accounting. Started out in Finance, I transitioned into Supply Chain field, then Demand Planning which I am currently in. Have industry experience in semiconductor, consumer packaged goods, retail, e-commerce, and material science manufacturing. I learned about the field of Data Science about 6 years ago and had been trying to obtain tangible knowledge and experience since. I have completed bootcamp curriculums with General Assembly and SpringBoard and have done Data Science projects on classification and regression (github)

5. Warren Ehrenfried; WarrenEhrenfried(EdX)
   Graduated from UTK with a BS in Geology. Within Geology I became proficient in GIS software and analyzing geospatial data although I have no real analytics background. Currently I work as a GIS Specialist at a natural gas company and review and update our infrastructure maps as well as review raw data captured in the field to QC our data capturing methods. I am currently working on an R Shiny application to help my team visualize gaps in our data capture methods in managing the large number of data points we capture every year.

OBJECTIVE/PROBLEM (5 points)

Project Title: Driving adoption of public transport as the preferred mode of commute in Austin, Texas

Background Information on chosen project topic:

A 2020 report by the Union of Concerned Scientists estimates that the average U.S. "ridesharing" trip results in 69% more pollution than the transportation choices it displaces . These results challenge claims by companies like Uber and Lyft that their services have a positive environmental impact by reducing the number of vehicles on the road.

Public transportation, on the other hand, is the most environmentally friendly mode of transportation short of walking or biking, producing on average six times fewer carbon emissions than a standard ridesharing trip (103 g CO2 compared to 683 g CO2, respectively). The impact of ridesharing can be limited by pooling trips and by using hybrid or electric vehicles, but even the best case ridesharing scenario - a pooled trip in an electric vehicle - would produce an estimated 146 g CO2, still 42% more than public transportation.

The city of Austin, Texas provides a unique opportunity to analyze the relationship between ridesharing and public transportation. In 2016 and 2017, a nonprofit ridesharing company (RideAustin) made ride data publicly available. The public transportation system in Austin, CapMetro, also made ridership and station data publicly available around the same time as part of a restructuring and rebranding effort. By using both datasets, we hope to explore factors which cause people to choose ridesharing over public transportation and make recommendations about how Austin could have strategically adjusted bus routes to encourage residents to use public transportation instead of ridesharing apps.

Under the company's goal of Community, CapMetro is dedicated to developing regional transit plans in partnership with the community to achieve a transit system that is accessible to everyone in Austin. The outcome of this project will highlight opportunities for CapMetro to invest capital on fleets of buses investment to improve connectivity of local residents across districts. This initiative will receive funding not only from CapMetro, but also from The City of Austin's Project Connect Office.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

Our group will provide a holistic bus route implementation strategy to the city of Austin by utilizing regression, clustering and optimization to inform public transit capital investment to meet the city's 50/50 mode share target by 2039, with the following key objectives:

- Reducing harmful emissions
- Providing more affordable and accessible transit services to citizens.
- Mitigating traffic congestion as city's population growth

State your Primary Research Question (RQ):

How do we improve the servicing route coverage, operation hours, plus customer preference to improve the utilization of the CapMetro public transportation services?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. How is ridesharing related to the availability of public transportation in a given area? Is ridesharing more common to and from locations that are further away from public transportation?
2. If so, how should CapMetro adjust its routes or schedule to capture more of the preferred mode of commute in the city?
3. What would be the estimated environmental impact of such changes, measured in CO2 emissions?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Under Austin Strategy Mobility Plan (ASMP) in 2019, Austin aims to achieve a 50/50 mode share by 2039, which refers to half of the commute in the city will use modes of transportation other than drive alone. As of 2021, only ~34% mode share has been achieved, with ridesharing (8%) ranking first in commute modes, while public transit (3%) ranks second.

This analysis will help shifting commute mode from vanpool/ridesharing to public transit. With a more convenient public transit network, it is likely to further facilitate the usage shift from drive alone, and drive toward 50% mode share goal and achieve Austin sustainability city development.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

https://data.world/andytryba/rideaustin

https://www.fueleconomy.gov/feg/download.shtml

https://data.texas.gov/Transportation/CapMetro-Shapefiles-AUGUST-2017/5d4c-snum

https://www.capmetro.org/destinations

Data Description (describe each of your data sources, include screenshots of a few rows of data):

1. Ride Volume. The 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.

Ride Volume

Date	Rides		Weekend	Total for Week	Running Total				
Thursday, June 16	190				190			Day 1	190
Friday, June 17	345				535			Day 2	535
Saturday, June 18	411				946			Day 3	946
Sunday, June 19	202		958	1148	1148			Day 4	1148
Monday, June 20	100				1248			Day 5	1248
Tuesday, June 21	124				1372			Day 6	1372
Wednesday, June 22	220				1592			Day 7	1592
Thursday, June 23	144	-24%			1736			Day 8	1736
Friday, June 24	397	15%			2133			Day 9	2133
Saturday, June 25	503	22%	21%	52%	2636			Day 10	2636
Sunday, June 26	256	27%	1156	1744	2892			Day 11	2892
Monday, June 27	97	-3%			2989			Day 12	2989
Tuesday, June 28	108	-13%			3097			Day 13	3097
Wednesday, June 29	122	-45%			3219			Day 14	3219
Thursday, June 30	203	41%			3422			Day 15	3422
Friday, July 1	514	29%			3936			Day 16	3936
Saturday, July 2	767	52%	65%	40%	4703			Day 17	4703
Sunday, July 3	622	143%	1903	2433	5325	297		Day 18	5325
Monday, July 4	402	314%			5727			Day 19	5727
Tuesday, July 5	199	84%			5926			Day 20	5926
Wednesday, July 6	195	60%			6121			Day 21	6121
Thursday, July 7	248	22%			6369			Day 22	6369
Friday, July 8	450	-12%			6819			Day 23	6819
Saturday, July 9	654	-15%	-22%	4%	7473			Day 24	7473
Sunday, July 10	378	-39%	1482	2526	7851			Day 25	7851

2. RideAustin_Weather. 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. By knowing the vehicle information, it helps to compute the consumption of fuel with another data set. It is crucial to understand the location points to identify demanding neighborhoods and use clustering to group similar neighborhoods together.

completed_on	distance_tr	end_locati	end_locati	started_on	driver_ratir	rider_ratin	start_zip_c	end_zip_cc	charity_id	requested	free_credit	surge_fact	start_locati	start_locati	color	make	model	year
2016/6/4 4:35	285	30.27	-97.75	2016/6/4 4:34	5	5				REGULAR		0	-97.75	30.27	Black	Cadillac	XTS	2013
2016/6/4 4:51	1029	30.27	-97.74	2016/6/4 4:45	5	5				REGULAR		0	-97.75	30.27	Black	Cadillac	XTS	2013
2016/6/4 5:27	8459	38.68	-121.04	2016/6/4 5:18	5	5				REGULAR		0	-121.07	38.65	Gray	Bentley	Continenta	2013
2016/6/4 6:51	443	38.68	-121.04	2016/6/4 6:50	5	5				REGULAR		0	-121.04	38.68	Gray	Bentley	Continenta	2013
2016/6/4 8:17	568	38.68	-121.04	2016/6/4 8:16	3	5				REGULAR		0	-121.04	38.68	Gray	Bentley	Continenta	2013
2016/6/4 15:13	4051	30.27	-97.74	2016/6/4 15:05	5	5				REGULAR		0	-97.76	30.25	Black	Cadillac	XTS	2013
2016/6/4 15:26	790	30.27	-97.75	2016/6/4 15:24	5	5				REGULAR		0	-97.75	30.27	Black	Cadillac	XTS	2013
2016/6/5 3:50	2171	30.27	-97.75	2016/6/5 3:40	0.5	5				REGULAR		0	-97.75	30.26	Black	Cadillac	XTS	2013
2016/6/5 4:33	10260	30.27	-97.75	2016/6/5 4:17		5				REGULAR		0	-97.77	30.2	Black	Infiniti	QX60	2015
2016/6/5 7:12	5294	30.24	-97.78	2016/6/5 6:57	3	5				REGULAR		0	-97.75	30.27	Black	Cadillac	SRX	2012
2016/6/5 7:36	9768	30.2	-97.77	2016/6/5 7:26	4.5	5				REGULAR		0	-97.74	30.27	Silver	Toyota	Highlander	2008
2016/6/5 20:59	12169	30.2	-97.67	2016/6/5 20:47	5	5				REGULAR		0	-97.75	30.24	White	Nissan	Murano	2015
2016/6/5 21:12	9859	30.27	-97.74	2016/6/5 20:56	5	5				REGULAR		0	-97.8	30.21	Black	Cadillac	XTS	2013
2016/6/5 22:19	4289	30.31	-97.75	2016/6/5 22:11	5	5				REGULAR		0	-97.74	30.29	Black	Cadillac	XTS	2013
2016/6/5 23:35	3305	30.29	-97.74	2016/6/5 23:29	3	5				REGULAR		0	-97.75	30.31	Black	Cadillac	XTS	2013
2016/6/5 21:55	2290	30.27	-97.75	2016/6/5 21:47		5				REGULAR		0	-97.74	30.26	Black	Cadillac	SRX	2012
2016/6/5 22:23	2411	30.27	-97.75	2016/6/5 22:15	3	5				REGULAR		0	-97.75	30.25	Black	Cadillac	SRX	2012
2016/6/5 23:57	2107	30.25	-97.75	2016/6/5 23:51		5				REGULAR		0	-97.75	30.27	Black	Cadillac	SRX	2012
2016/6/5 23:33	2120	30.27	-97.74	2016/6/5 23:24	5	5				REGULAR		0	-97.75	30.27	Gray	Toyota	Highlander	2012
2016/6/6 0:07	9771	30.26	-97.75	2016/6/5 23:47		5				REGULAR		0	-97.77	30.31	White	Nissan	Murano	2015
2016/6/6 1:28	3468	30.31	-97.73	2016/6/6 1:20		5				REGULAR		0	-97.75	30.29	White	Chevrolet	Tahoe	2013

3. Fuel economy -Vehicle. This data set is associated with the RideAustin_Weather data set. This will help to get fuel consumption in both city and highway with CO2 emission on each specific vehicle.

barrels08	barrelsA08	charge120	charge240	city08	city08U	cityA08	cityA08U	cityCD	cityE	cityUF	co2	co2A	co2Tailpipi	co2Tailpipi	comb08	con
14.16714	0	0	0	19	0	0	0	0	0	0	-1	-1	0	423.1905	21	
27.04636	0	0	0	9	0	0	0	0	0	0	-1	-1	0	807.9091	11	
11.01889	0	0	0	23	0	0	0	0	0	0	-1	-1	0	329.1481	27	
27.04636	0	0	0	10	0	0	0	0	0	0	-1	-1	0	807.9091	11	
15.65842	0	0	0	17	0	0	0	0	0	0	-1	-1	0	467.7368	19	
13.52318	0	0	0	21	0	0	0	0	0	0	-1	-1	0	403.9545	22	
11.9004	0	0	0	22	0	0	0	0	0	0	-1	-1	0	355.48	25	
12.39625	0	0	0	23	0	0	0	0	0	0	-1	-1	0	370.2917	24	
11.44269	0	0	0	23	0	0	0	0	0	0	-1	-1	0	341.8077	26	
11.9004	0	0	0	23	0	0	0	0	0	0	-1	-1	0	355.48	25	
11.44269	0	0	0	23	0	0	0	0	0	0	-1	-1	0	341.8077	26	
14.16714	0	0	0	18	0	0	0	0	0	0	-1	-1	0	423.1905	21	
12.39625	0	0	0	21	0	0	0	0	0	0	-1	-1	0	370.2917	24	
14.16714	0	0	0	18	0	0	0	0	0	0	-1	-1	0	423.1905	21	
22.88538	0	0	0	12	0	0	0	0	0	0	-1	-1	0	683.6154	13	
12.93522	0	0	0	20	0	0	0	0	0	0	-1	-1	0	386.3913	23	
14.8755	0	0	0	18	0	0	0	0	0	0	-1	-1	0	444.35	20	
14.16714	0	0	0	19	0	0	0	0	0	0	-1	-1	0	423.1905	21	
15.65842	0	0	0	17	0	0	0	0	0	0	-1	-1	0	467.7368	19	
15.65842	0	0	0	17	0	0	0	0	0	0	-1	-1	0	467.7368	19	
18.59438	0	0	0	14	0	0	0	0	0	0	-1	-1	0	555.4375	16	
18.59438	0	0	0	14	0	0	0	0	0	0	-1	-1	0	555.4375	16	
22.88538	0	0	0	11	0	0	0	0	0	0	-1	-1	0	683.6154	13	
12.93522	0	0	0	21	0	0	0	0	0	0	-1	-1	0	386.3913	23	
15.65842	0	0	0	17	0	0	0	0	0	0	-1	-1	0	467.7368	19	
22.88538	0	0	0	11	0	0	0	0	0	0	-1	-1	0	683.6154	13	
14.16714	0	0	0	18	0	0	0	0	0	0	-1	-1	0	423.1905	21	

4. CapMetro Shapefile. Dataset contains the geographic locations of current public transit stops in the Austin CapMetro public transit city

STOP_ID	STOP_NAME	STOP_ABBR	STREET_NMB	ON_STREET	AT_STREET	CITY	ZIP	BAY	STOP_TYPE	PLACEMENT	CORNER	STATUS	LATITUDE	LONGITUDE
66	4925 Craigwood/FM	CRFMS	4925	CRAIGWOOD	FM 969	AUSTIN	78725		Bus Stop	Nearside	Southeast	Active	30.2841709	-97.65985415
252	200 Trinity/2nd	2TRS	200	TRINITY	2ND	AUSTIN	78701		Bus Stop	Mid-Block	Northeast	Active	30.2638421	-97.74042677
462	851 Rutland/Park Vil	S1	851	RUTLAND	PARK VILLAGE	AUSTIN	78758		Bus Stop	Mid-Block	Southeast	Active	30.36547	-97.69752
466	8740 Lamar/Payton	S1801	8740	LAMAR	PAYTON GIN	AUSTIN	78758		Bus Stop	Mid-Block	Southwest	Active	30.35680916	-97.70106551
467	8630 Lamar/Fairfield	S63	8630	LAMAR	FAIRFIELD	AUSTIN	78758		Bus Stop	Farside	Southwest	Active	30.35529611	-97.7031279
468	Lamar/Thurmond	S15	8400	LAMAR	THURMOND	AUSTIN	78758		Bus Stop	Farside	Southwest	Active	30.35313912	-97.70608203
469	8320 Lamar/Meadov	S62	8320	LAMAR	MEADOWLARK	AUSTIN	78758		Bus Stop	Nearside	Northwest	Active	30.3522556	-97.70729488
471	7720 Lamar/Stobaug	S1911	7720	LAMAR	STOBAUGH	AUSTIN	78757		Bus Stop	Mid-Block	Southwest	Active	30.34604	-97.71394
472	7520 Lamar/Morrow	S11	7520	LAMAR	MORROW	AUSTIN	78757		Bus Stop	Mid-Block	Southwest	Active	30.34313363	-97.71565611
474	6814 Lamar/Justin	S56	6814	LAMAR	JUSTIN	AUSTIN	78757		Bus Stop	Farside	Southwest	Active	30.33620568	-97.72000833
475	6600 Lamar/Brentwo	S54	6600	LAMAR	BRENTWOOD	AUSTIN	78757		Bus Stop	Nearside	Northwest	Active	30.33411003	-97.72132254
476	6200 Lamar/Denson	S7	6200	LAMAR	DENSON	AUSTIN	78757		Bus Stop	Nearside	Northwest	Active	30.33042504	-97.72363629
480	5528 Lamar/Koenig	S48	5528	LAMAR	KOENIG	AUSTIN	78756		Bus Stop	Mid-Block	Southwest	Active	30.32424886	-97.72752964
482	5300 Lamar/North Lo	S6	5300	LAMAR	NORTH LOOP	AUSTIN	78756		Bus Stop	Nearside	Northwest	Active	30.32130258	-97.72939072
483	5106 Lamar/51st	S44	5106	LAMAR	51ST	AUSTIN	78756		Bus Stop	Nearside	Northwest	Active	30.31904976	-97.73080613
484	Triangle Station (SB	S1805	4600	GUADALUPE	LAMAR	AUSTIN	78751		Rapid Station	Mid-Block	Southwest	Active	30.3146309	-97.73252431
485	4500 Guadalupe/45t	200SB	4500	GUADALUPE	45TH	AUSTIN	78751		Bus Stop	Nearside	Northwest	Active	30.311163	-97.73295788
486	Guadalupe/43rd Stre	S39	4300	GUADALUPE	43RD	AUSTIN	78751		Bus Stop	Nearside	Northwest	Active	30.30883127	-97.73439524
487	Guadalupe/41st Stre	S37	4100	GUADALUPE	41ST	AUSTIN	78751		Bus Stop	Nearside	Northwest	Active	30.30670338	-97.73575572
489	Guadalupe/Maiden	S1806	3500	GUADALUPE	MAIDEN	AUSTIN	78705		Bus Stop	Nearside	Northwest	Active	30.30167414	-97.73896028
490	Guadalupe/34th Stre	S31	3402	GUADALUPE	34TH	AUSTIN	78705		Bus Stop	Nearside	Northwest	Active	30.30045562	-97.73970878
492	Guadalupe/30th Stre	S27	3000	GUADALUPE	30TH	AUSTIN	78705		Bus Stop	Nearside	Northwest	Active	30.29680363	-97.74202809
494	Guadalupe/27th Stre	S14	2700	GUADALUPE	27TH	AUSTIN	78705		Bus Stop	Nearside	Northwest	Active	30.29219693	-97.74130995
495	Guadalupe/26th Stre	S2	2600	GUADALUPE	26TH	AUSTIN	78705		Bus Stop	Nearside	Northwest	Active	30.29067306	-97.74140804
497	UT West Mall Statior	S19	2246	GUADALUPE	23RD	AUSTIN	78705		Rapid Station	Mid-Block	Southwest	Active	30.28606366	-97.74181517

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Dependent variable: Rideshare volume number, bus stops in Austin, CO2 emissions in Austin Area, probability of a rideshare trip to be replaced by the newly defined bus route (classification)

Independent variables: start and end locations; weather, bus on schedule, number of bus stops in between, number bus routes in between locations, fuel consumption for investigated model cars in between locations;

New variables: MPG*distance*barrel price to account for the independent variable for the purpose of cost comparison between public transportation to rideshare. co2*MPG*distance to create the independent variable for emission comparison based on the rideshare volume. Average daily public transportation volume by station (vol_pubtrans), average daily volume of rideshares associated with each station (vol_rideshares), average daily distance from station to start and end points of associated rideshares (dist_to_pubtrans),also the difference in commute time (diff_in_commute time) among bus vs rideshare and difference in cost of commute.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Our approach to this question follows four stages, an exploratory stage and three stages that correspond to our supporting research questions.

1. After exploring the data, is the scope of our research question reasonable?

   The first step of our analysis will be to determine the scope of our research question. This partially depends on the granularity of the RideAustin geolocation data. If the data is coarser than the density of public transportation stops in Austin, then it would be difficult to draw any conclusions about the relationship between station location and ridership patterns. Our preliminary mapping of the RideAustin data shows that latitude and longitude coordinates are accurate within 6 blocks. Our next step is to map the CapMetro stations. If, in the densest areas of the CapMetro network, stations are more than six blocks apart, no modifications to our research question are necessary. However, if CapMetro stations are less than six blocks apart, we may need to redefine our project so that we

consider only areas of the city where stations are more spread out. The new research question could be, for example, how CapMetro could increase the use of public transportation specifically for people commuting into the city from a certain distance away from downtown.

Other exploratory questions might include:

a. Physical - are there areas where ride trip start or end do not have bus services coverage?
b. Timing - how many of the rideshare trip fall outside bus service operation?
c. Service - what are the trips that are covered by both physical & timing perspective, but the duration is too long as comparing to taking buses?

2. Does distance affect the percent of people who use ridesharing instead of public transportation?

   To compare ridesharing patterns with public transportation patterns, we need to assign four new variables to each ride in the RideAustin dataset:

   1) a binary variable indicating whether a ride could have been taken with public transportation. This would be true if the start and end location of the ride were within **¾ miles of a station**, and false otherwise. (Our distance factor is based on research suggesting that ¾ mile is an inflection point where pedestrians become less willing to walk to public transportation.)

   2 and 3) If variable 1 is true, a categorical variable indicating the closest CapMetro station to the beginning and end points of the trip. (For now, we plan to use straight-line distance. We may need to adjust at least for the Colorado River, which runs through the city and can be crossed only where there is a bridge.)

   4) If 1 is true, the total distance of the closest start and end stations to the start and end of the trip, as a continuous variable.

   Some locations will have heavier traffic just because of their location. For example, downtown locations will likely have more people using public transportation and more people using ridesharing, because there are more total people using transportation there. To know whether ridesharing is affected by distance to public transportation, we need to compare the geographical distribution of ridesharing with the geographical distribution of public transportation use. If the two distributions are approximately the same, we cannot assume that distance has a significant effect on people's choice of transportation. If the two distributions are different, then distance may be a significant factor.

   We can use a regression model to explore the relationship of these two distributions. Using the CapMetro ridership data, we will calculate average daily public transportation volume by station (vol_pubtrans). Based on the closest station of rideshares, we will find average daily volume of rideshares associated with each station (vol_rideshares). We can also calculate average daily distance from station to start and end points of associated rideshares (dist_to_pubtrans). The following regression model will show us whether rideshare distribution differs significantly from public transportation patterns, and if so, whether distance is a significant factor:

   - vol_rideshares = a + b(vol_pubtrans) + c(dist_to_pubtrans) + d(diff_in_commute time among bus vs rideshare) + e(diff in cost of commute).

If c is significant and negative, then distance from a station is probably a factor in people's choice to use public transportation or ridesharing. If not, we will need to look for other factors: day of the week, weather, station efficiency, etc.

If b is significant but negative, this could be an even stronger indication that ridesharing is more common where public transportation tends to be less utilized, whether because of availability or some other factor.

If b is significant and positive and c is not significant, then we cannot assume that mere availability of public transportation will change people's behavior. It's more likely that people choose ridesharing for convenience or preference, regardless of how close they are to public transportation. In this case, we might advise the city to invest more in their own version of ridesharing (i.e., VanPool), or we could expand the regression model to include other variables that might do a better job explaining ridesharing patterns compared to the overall distribution of travel.

3. How can the city adjust routes, schedule, or other factors to encourage more people to use public transportation?

   Our approach here will depend on stage 2. If distance appears to be the main factor, then we could focus on rideshares that were not served at all by public transportation (i.e., start or end points were more than a certain distance away from a station) and suggest new stations to bring these locations into the public transportation network. K-means clustering of start and end points would organize those rideshares into natural clusters. The center of each cluster would represent a potential new station.

   If distance is not a factor, then we would need to shift focus in this stage, possibly looking at improving station efficiency (reducing wait times, number of late arrivals, increasing hours of operation etc.) or developing Austin's public ridesharing program using similar clustering and optimization methods as described below.

4. What would be the estimated environmental impact of this plan?

   The final stage is to figure out which potential improvements would lead to an optimal result in terms of ridership and environmental impact. We would need to estimate whether, for example, a new station is cost effective based on distance from other routes, and whether the cost of building the bus stop would justify the amount of rides we could expect to gain and the environmental impact of this change. An estimate from Durham, NC suggests that a new bus stop might cost $35,000 to $50,000 to build, depending on the amenities included. Any additional length added to a bus route would add fuel cost and, in some cases, require the city to add vehicles so that wait time was not affected too severely. We could estimate the cost of each station as: construction_cost + length*fuel_cost + length_factor*new_vehicle_cost.

   Our regression equation from step 2 would suggest the number of riders we might expect to gain from a new station. Based on the type of vehicles used in the clusters of 2016 rideshares, we could use EPA data on mpg by vehicle make and model to convert our proposal to gallons of fuel saved and the amount of reduction in CO2 emissions.

   If we assigned some monetary value to CO2 emissions (The current estimate of the total social cost of CO2 emissions, for instance, would suggest $51 per ton of CO2, but may soon rise to $190 based on EPA recommendations), then we could translate this into an optimization problem: maximize value of

CO2 reduction based on adding a certain set of stations with an associated cost of construction, fuel, and vehicles.

Data transformations required:

- Clean up make and model fields to match RideShare and mpg datasets. RideShare is often less specific about make and model than MPG file (2D/4D, RWD/4WD/AWD, etc.) We will need to decide how to aggregate the MPG data by general make and model.
- Transforming latitude and longitude data into distance. (Probably straight-line distance? May be difficult to get street distance.)

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

After completing the initial data exploration, we expect to find that there is a correlation between riders choosing to take a ride share and the distance to the bus stop. However, if there is not, a hypothesis test may be required to assess the rider's true preference to switch mode of transportation. It is expected that it will not be the sole factor, however, it should be significant enough to warrant a complete analysis on the public transportation in Austin.

Per our exploration, the next conclusion that is expected to be drawn is that bus stops are not available in all areas where the population takes rides. Through looking at the geographic locations and utilizing a predetermined radius, the group expects that there is an opportunity to enhance the current transit system in Austin. Due to some logistical and political constraints, the current setup of public transit stops will not be the optimal cluster model that maximizes coverage for Austin's population and there is an opportunity to right size and adjust CapMetro stops in the city.

Lastly, through marginal changes in bus routes, it is expected that the overall CO2 emissions from reducing rideshares and increasing public transit will have a net positive effect in the Austin area.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Through our analysis we hope to provide the city of Austin city infrastructure decisions that can impact their use of public transit.

Some benefits of this proposal will help reduce harmful emissions, provide more affordable and accessible transit services to citizens, and reduce traffic congestion while supporting the population growth in the city.

In addition to our recommendations to make smarter city infrastructure decisions, the government/city can partner with companies to promote local business. The rideshare opportunities however, may show a descending trend and the people who lost their jobs can be repositioned to the new infrastructure projects.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Decide on algorithms to pursue (flexible as more info unveiled) - June 21st

Complete Data ingestion and cleansing - June 28th

Project Proposal Video- July 2nd

Models for each component of analysis completed - July 5th

Progress Report 1 - July 9th

Integrate all aspects into comprehensive script - July 12th

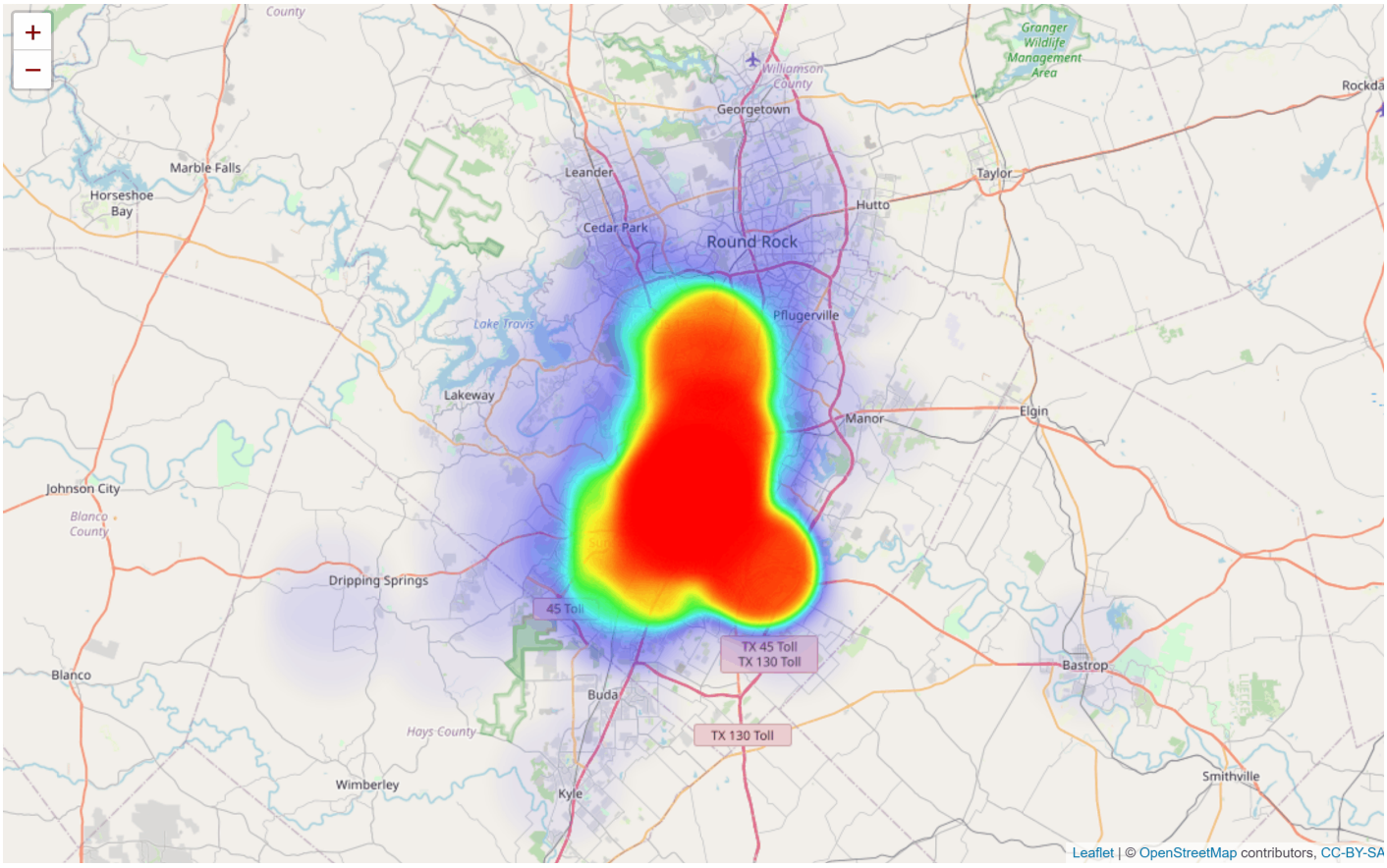
Code finalized - July 17th

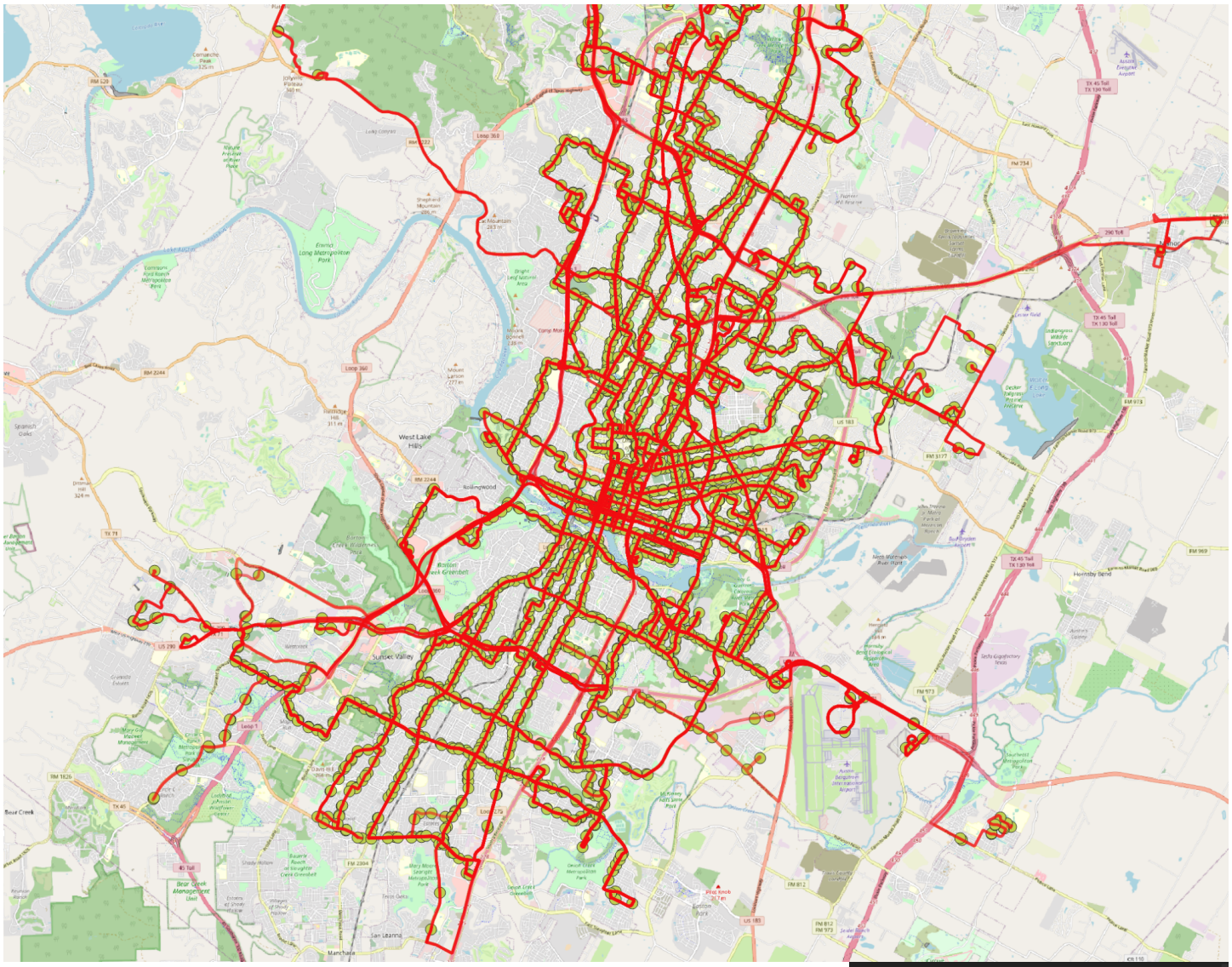
Final Report 1 - July 20th

Final Video Presentation - July 23rd

Appendix (any preliminary figures or charts that you would like to include):

<u>Heatmap of 5000 samples points to display highest traffic areas</u>





Bus routes and bus stops displayed in GIS