Trip Prediction of Shared Micromobility in Austin

1. Abstract

Shared micromobility has become more popular in the last few years, offering benefits such as reduced traffic congestion and lower carbon emissions. However, increased shared micromobility vehicles usage also presents challenges related to rising demand and profitability. Implementing a model to predict trip duration can help operators anticipate demand and optimize profitability. In this project, we implemented a multiple linear regression model using trip duration as the response variable and trip distance, month, hour, day of week, the district of trip origin and the vehicle type as the predictors. The regression analysis revealed that trip duration is influenced by all these factors, with varying magnitudes. For example, longer distance was associated with longer trips, while scooter trips were shorter than bicycle trips. Additionally, the results provided insights into the temporal, spatial and monthly effects on trip duration, which is information that can be used to improve the profitability of shared micromobility.

2. Introduction

The use of shared micromobility has risen in the last few years. More people are opting to use shared e-scooters and bicycles over personal vehicles for short distance travel. In comparison to 2022, the number of micromobility trips “increased by 16% in 2023” in the United States, reaching a total of 133 million trips [1]. Using shared micromobility over private vehicles offers several benefits, such as reduced traffic congestion, lower carbon emissions and less need of parking space.

While the rise of shared micromobility usage comes with several benefits, this has also presented several challenges. For instance, as demand for shared micromobility increases, the shared vehicle operators need to ensure that supply keeps pace, meaning that they need to be able to predict the demand for these shared vehicles. Another challenge operators have faced is profitability. In 2023, the shared micromobility company Bird filed for bankruptcy, while Micromobility.com was removed from the Nasdaq stock exchange after its stock price failed to meet the minimum share price requirements. In this same year, Superpedestrian shut down operations in the U.S due to financial problems [2]. One of the factors contributing to this profitability challenge is cost management, particularly the cost of the vehicle manufacturing [3].

To address these challenges, a multiple linear regression model that can predict trip duration can be implemented. Predicting the trip duration can be used to better anticipate demand and optimize the shared micromobility distribution and operations. Additionally, the multiple linear regression model can help operators identify key factors influencing travel duration, providing insights that can be used to optimize revenue.

3. Problem Goal

In this project, we would like to address two challenges that shared micromobility operators face: the rising demand and profitability issues. Can we implement a model that can predict trip duration using variables such as trip distance, month, hour, day of week, the district of trip origin and the vehicle type? How do these factors influence trip duration, and which have the strongest effects?

Using a dataset of shared micromobility vehicle trips in Austin, Texas, the goal of this project is to fit a multiple linear regression model that can predict trip duration given the trip distance, month, hour, day of week, district origin and the vehicle type. In addition, we would like to understand the relationship between these predictors and the trip duration, as well as identify predictors with greater impact on trip duration.

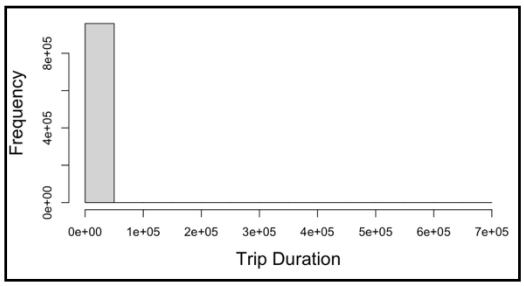
4. Data Description

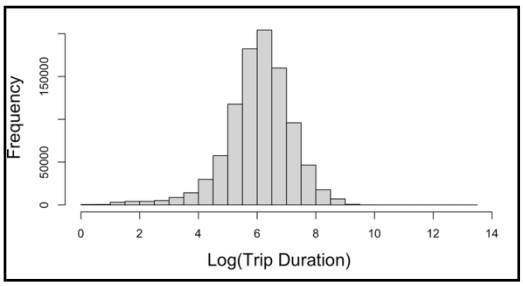
The dataset consists of shared micromobility vehicle trips from 2018 - 2022 in Austin [4]. It contains over 15 million rows and 18 columns. Due to the large size of this dataset, I limited the data downloaded to only the rows corresponding to January, May and September of 2021, resulting in 959,790 records.

The table below contains the columns used in this project along with their definitions:

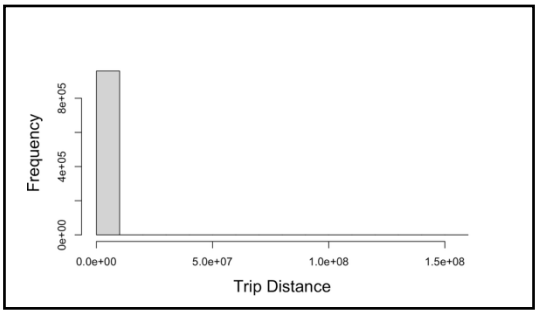
Variable	**Definition**
Trip.Duration	This is the duration of the trip in seconds.
Trip.Distance	This is the distance of the trip in meters.
Day.of.Week	This is the day of the week on which the trip started, represented as 0 - 6, with 0 being Sunday.
Vehicle.Type	This represents the vehicle used during the trip. This can be “moped”, “scooter”, or “bicycle”.
Hour	This is the hour in which the trip started, represented as 0 – 23.
Council.District..Start.	This represents the council district number in which the trip started from.
Month	This represents the month in which the trip started from, represented as 1-12.

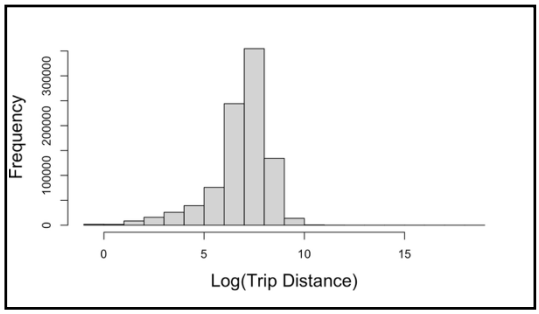
5. Exploratory Data Analysis



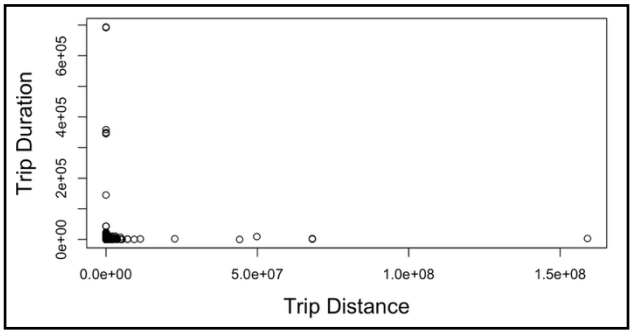


The left image above shows that while most trip durations are short, there are some trip durations that are very long, meaning the data is right-skewed. The right image shows that applying a log transformation to trip duration can reduce the effects of the long trip durations, allowing us to observe a more normal distribution of the trip duration values.

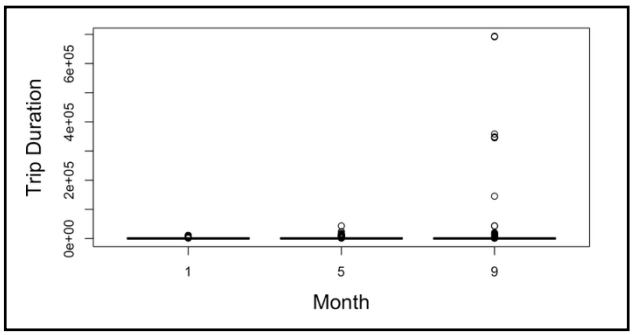




The figures above display a similar behavior to trip duration. Applying a log transformation to trip distance reduced the effect of the extreme values to the right of the histogram.



The above is a trip duration versus trip distance plot. It reveals that most data points have both a short distance and duration, and there doesn't appear to be a linear relationship.



The boxplot above for the month variable revealed there are high values for the trip duration in September. Since school typically starts towards the end of August or early September, it is possible these higher values could be due to the start of school. For instance, in 2021, the University of Texas at Austin started the Fall semester on August 25th [5]. As another possibility, there could have been events that led people to have longer trips. For example, University of Texas had several games in September, which could have led to people using the shared micromobility vehicles more and with a longer trip duration [6]. However, it is also possible that these extreme values are a result of errored reporting data in September.

6. Methods

6a Baseline model

$$\text{Trip. Duration} \sim \text{Trip. Distance} + \text{Day. of. Week} + \text{Vehicle. Type} + \text{Hour} + \text{Council. District. Start} + \text{Month}$$

We fitted the above multiple linear regression model using the lm() function in R, which estimates the model parameters using ordinary least squares (OLS). In this model, the response variable is the trip duration. The predictor variables include the numerical predictor trip distance and the following categorical variables: the day of the week, vehicle type, hour, council district origin and month. The lm() function automatically converted the categorical predictors into dummy variables.

6b Transformation

From the data exploration, the trip distance and the duration were both heavily right-skewed. In addition, the trip duration versus distance plot revealed a non-linear relationship between these two variables. A previous study on trip prediction showed that applying log transformation can not only address skewness but also improve the performance of the model [7]. Therefore, to address these issues, we performed log transformation to the two variables.

6c Multicollinearity

Multicollinearity exists when the predictor variables are related to each other. In such cases, this can produce large coefficient estimates as well as incorrect t-test results, making it difficult to identify the significant predictors. In this project, we check for multicollinearity using the variance inflation factor (VIF). Large VIF values indicate the presence of multicollinearity.

6d Model Selection

To select the "best" model, we used the step() function to perform backward selection based on the Akaike Information Criterion (AIC). This starts from the full model and iteratively removes predictors to minimize the AIC.

6e Outliers and Influential Points

In this project, we used the Bonferroni correction to identify potential outliers. After identification of these outliers, we analyzed the outliers to determine if these were reasonable outliers or if they are more likely to be errored data. Using Cook's Distance, we can identify potential influential points and determine if they are influential enough that can cause a large effect in our model if we remove them.

6f Assumptions

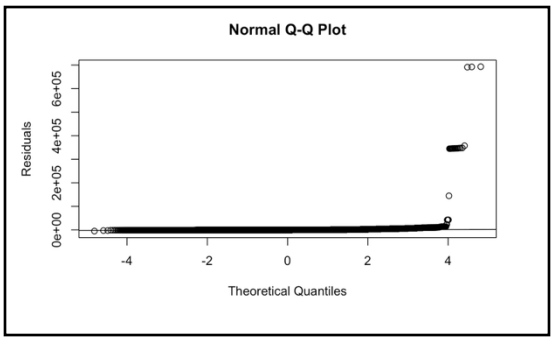
To ensure the validity of our linear regression model, we checked the regression error assumptions. By plotting the residuals against the fitted values, we can verify the constant variance assumption. A constant mean function equal to 0 would indicate linearity, while homoscedasticity would indicate constant variance. To verify the normality assumption of the residuals, we generated the QQ-plot and verified if most data points are aligned with the normal line.

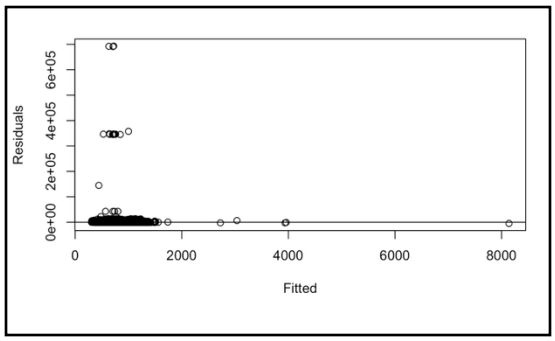
6g Metrics

To fit and evaluate the models, the data was separated into a training and a testing set in a ratio of 7:3. The metrics used during for this evaluation were R^2 and the RMSE.

7. Results

7a Baseline Model





After fitting the initial model, this resulted in a model with a low R^2 of 0.002 and RMSE of 2115.83. The above QQ-plot showed that the residuals are not normally distributed, while the residuals vs fitted plot showed that most residuals were grouped together on the lower left of the graph with several outliers.

7b Transformation

After applying the log transformation, our new model is:

$$\log(Trip.Duration) \sim \log(Trip.Distance) + \text{Day.of.Week} + \text{Vehicle.Type} + \text{Hour} + \text{Council.District.Start} + \text{Month}$$

The summary of the transformed model demonstrated that compared to the initial model, the R^2 increased to 0.4963.

7c Multicollinearity

Variable	VIF
Hour3	3.256
Council.District..Start.6	2.71
Month5	2.547
Month9	2.522
Council.District..Start.10	2.109
Hour23	1.85
Hour1	1.807
Hour2	1.799
Hour21	1.757
Hour6	1.751

The table above shows the top 10 highest VIF values. The highest VIF value is 3.256 for the predictor Hour3, indicating there is some multicollinearity but not high, meaning that it is not necessary to remove any of the predictor variables.

7d Model Selection

After performing backward selection based on the AIC, the results showed that no predictors were dropped, meaning that no variable removal improved the AIC. The final model's AIC was -425,066.8.

7e Outliers and Influential Points

Using Bonferroni's correction, we identified 320 potential outliers. After analyzing these outliers, we noticed two categories of outliers: outliers with extreme high speeds and outliers with extreme low speeds. The observations with extremely high speeds (calculated as trip distance divided by trip duration) appeared implausible, meaning that these observations could have been a result of data collection errors.

The table below shows a few examples of such observations. The column "speed" was added to find the speed of each data point:

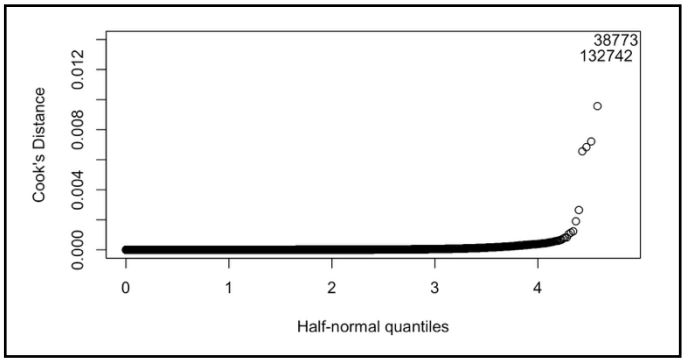
Trip.Duration	Trip.Distance	Vehicle.Type	speed_mps
19	44,141,260	scooter	2,323,224.21
25	4,876,529	scooter	195,061.16
13	2,412,790	scooter	185,599.23
1080	68,153,166	scooter	63,104.78
32	1,780,141	scooter	55,629.41

From the table above, we can see that in the first observation, the speed was 2,323,224.21 m/s, which does not appear to be plausible. Moreover, shared scooters systems typically have a maximum speed of 20 - 25 km (5.56 m/s – 6.94 m/s) for safety reasons [8]. In addition, these guidelines show that the maximum speed for e-scooters should be 15 mph (6.7056 m/s) [9]. Hence, we applied a threshold on the possible highest speed such that we removed the outliers in which the speed was greater than 7 m/s. The second category of outliers were the extremely low speed observations, in which the trip distance was very small compared to the trip duration. The table below shows a few of these idle observations:

Trip.Duration	Trip.Distance	Vehicle.Type	speed_mps
347,268	54	scooter	0.0001555
345,885	62	scooter	0.00017925
3,224	1	scooter	0.00031017
2,677	1	scooter	0.00037355
5,287	2	scooter	0.00037829

Per the table above, the first row contains a data point in which the distance traveled was 54 meters, but the trip duration was 347,268 s (96.463 h), which is more than one day. Since these data points do not appear to be plausible and may greatly affect our model fit, I removed the outliers in which the speed was less than 0.05 m/s and the trip duration was longer than 300 s.

After analyzing the outliers, we identified the influential points using Cook's distance. In the graph below, we can see the observations with a Cook's distance larger than 0.004 appear to be more influential than the rest.

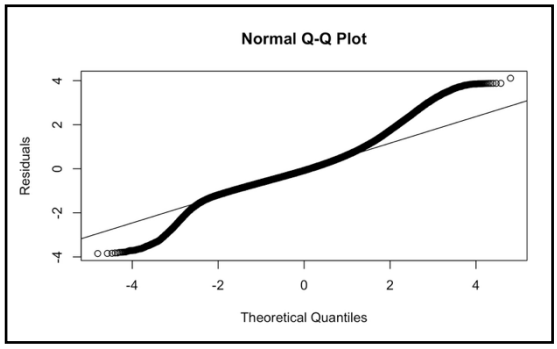


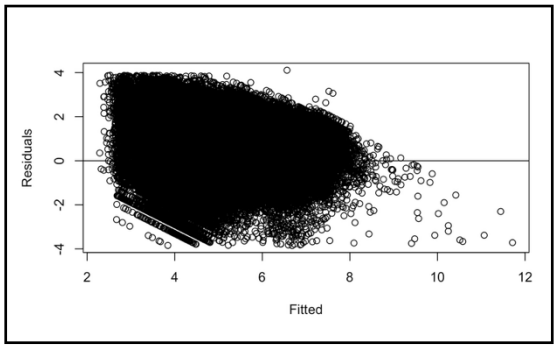
The table below contains the observations that were flagged as more influential than other points:

Trip.Duration	Trip.Distance	Vehicle.Type
5477	1149	scooter
1040	17	scooter
696	17	scooter
476	1724	scooter
924	3210	scooter
637	1728	scooter

After removing these influential points and re-fitting the model, however, the summary regression coefficients and the p-values indicated there were no significant changes, so these points were not removed from the data.

7f Regression Assumptions





After identifying and removing the outliers as described previously, we re-assessed the regression assumptions. The QQ-Plot above shows that the middle points aligned well with the line while the tails did not, suggesting non-normal residuals. The residuals against fitted values plot shows that most residuals are centered around 0, meaning there is a constant mean of 0 for the residuals. However, since there is a greater spread towards the higher fitted values, this indicates there is heteroscedasticity, meaning that the plot does not appear to show constant variance.

7g Final Model Summary

The final model fitted based on ordinary least squares (OLS) was the following with an R^2 of 0.5036:

$$\log(Trip.Duration) \sim \log(Trip.Distance) + \text{Day. of. Week} + \text{Vehicle. Type} + \text{Hour} + \text{Council. District. Start} + \text{Month}$$

Predictor Group	Predictor	Effect size
Distance	*Log(Trip.Distance)*	0.49% trip duration increase per 1% increase
Vehicle Type	*Vehicle.Typescooter* (Scooter)	-7.33%.
Month	*Month5* (May)	-8.99%
	Month9 (September)	-11.97%
Council District Origin	*Council.District.Start.6* (District 6)	233.28%
	Council.District.Start.3 (District 3)	-18.63%
Hour	*Hour7-Hour9* (7:00 A.M. – 9:00 A.M.)	4.4% to 20.1%
	Hour10-Hour17 (10 A.M. – 5:00 P.M)	-26.4% to -4.5%
	Hour19-Hour21 (7:00 P.M. – 9:00 P.M.)	1.3% to 1.7%
Day of Week	*Day.of.Week.1-Day.of.Week4* (Monday – Thursday)	-12.3% to - 4.3%
	Day.of.Week.5-Day.of.Week6 (Friday – Saturday)	3.7% to 5.6%

The table above is a summary of the regression results. The full output of summary() can be found on Table 1 in the Appendix. Based on the summary output, most predictors are statistically significant, meaning that their p-value was less than 0.05. The first significant predictor is trip distance, which has a positive linear relationship with trip duration- a 1% increase in trip distance results in a 0.49% increase in trip duration. This relationship is expected since as trip distance increases, we would expect trip duration to increase as well.

"Vehicle.Type" and the rest of the predictors are categorical and are represented by dummy variables, meaning their regression coefficients are relative to the corresponding reference category. For example, "Vehicle.Typescooter" has a regression coefficient of -0.0762, which indicates that scooter trips are shorter than trips by bicycle by 7.33%. Note that since we applied log transformation, the effect is calculated as $(e^{\beta} - 1) \times 100$.

Both months ("Month5" and "Month9") in the output are significant and have negative regression coefficients, meaning that trips in May and September are typically shorter than in the reference month January by 8.99% and 11.97%, respectively. The difference in the coefficient magnitudes suggest there is a monthly effect, and the trip duration differs across months.

Most of the “Council.District..Start.” predictors were significant except for “Council.District..Start.2” and “Council.District..Start.7”. Overall, this indicates that the trip origin has a meaningful effect on trip duration, with some districts associated with longer trips and others with shorter trips relative to the reference district (first district). “Council.District..Start.6” had the largest positive regression coefficient at 1.20 with trips starting from this district longer than those from the first district by 233.28%. On the other hand, “Council.District..Start.3” had the lowest negative regression coefficient at -0.206, so this district is associated with the shortest duration compared to the first reference district.

Most of the “Hour” predictors were statistically significant, meaning that trip duration can vary during a day. Since “Hour10”-“Hour17” had negative regression coefficients, trips from 10:00 A.M. to 5:00 P.M (midday to afternoon) are about 4.5-26.4% shorter than the reference hour (12:00 AM). On the other hand, 7:00 AM–9:00 AM and 7:00 PM–9:00 PM were associated with longer trips by around 4.4-20.1% and 1.3-1.7% longer, respectively, relative to midnight. These longer durations in the morning could be linked to the rush-hour travel commute to school or work, while the evening longer durations may be related to the commute to home or leisure travel. Overall, the results for “Hour” indicate there is a temporal pattern- depending on the time of the day, the trip may be shorter or longer.

All the “Day.of.Week” predictors were statistically significant, confirming that trip duration differs throughout the week. The Monday-Thursday predictors had negative regression coefficients, indicating that these trips are typically shorter than on the reference day (Sunday). Friday and Saturday had positive coefficients, meaning that trips are longer on Fridays and Saturdays by about 3.7-5.6% than on Sunday. This pattern may suggest that the shorter trips are associated with commuting to work or school, while the longer trips on Fridays and Saturdays may be associated with leisure travel.

After analyzing the summary of the final model, this final model was then evaluated using the test data, resulting in an RMSE of 170.98. To compare the RMSE, the original model (without any transformations), was first re-fit using the cleaned training data with the outliers removed and then re-evaluated using the test data. This resulted in an RMSE of 563.54, meaning the final model reduced the RMSE by 69.66% and outperformed the original model.

8. Conclusion

In this project, I successfully implemented a multiple linear regression model such that given the trip distance, month, hour, day of week, the district trip origin and the vehicle type, we can predict the trip duration. To address the skewness of the trip distance and duration data, we applied a log transformation

to both variables. This model achieved an R^2 value of 0.5036 and an RMSE of 170.98, outperforming the original model RMSE by 69.66%.

The regression analysis of the final model revealed that trip duration is influenced by distance, vehicle type, month, district origin, time of day and day of week. A longer trip distance is associated with longer trips, and trips by scooter are shorter than when using a bicycle. There is a monthly effect on trip duration since both May and September had shorter trips than in January. The model revealed that the trip duration can vary depending on the trip origin district, with District 6 having significantly longer trips compared to District 1, and District 3 having the shortest trips compared to District 1. Moreover, the time of day influenced the trip duration- 7:00 AM–9:00 AM and 7:00 PM–9:00 PM were associated with longer trips compared to the reference, suggesting these were related to commuting to work or school, and commuting back home. Finally, the day of the week also influenced trip duration with Fridays and Saturdays having longer trip durations than Monday to Thursday, which may suggest that leisure travel is associated with longer trip durations.

This multiple linear regression model provides a tool for micromobility system operators to predict trip duration to better anticipate the demand and optimize the shared micromobility distribution and operations. The regression analysis insights provide information that can be used to create strategies to optimize revenue. However, it is important to note that this regression analysis had some limitations. Despite adding the log transformation, there was heteroscedasticity in the residuals versus fitted values graph. In addition, the QQ-plot demonstrated that the tails deviated from the normal line. Furthermore, while the RMSE (170.98) of the final model improved when comparing it to the RMSE of the original model, this was still relatively large since the mean and median of the test data were 175.9 and 183.0, respectively. Overall, the results suggest that while the model captures some structure of the data, its prediction capabilities could be improved. Future work should consider adding interaction terms, additional transformations or robust methods to address the heteroscedasticity and non-normality issues.

9. References

[1] National Association of City Transportation Officials, *Shared Micromobility Report 2023*. NACTO, 2023. [Online]. Available: https://nacto.org/publication/shared-micromobility-report-2023/

[2] L. Lazo, “E-scooter cities settle after turbulence in micromobility,” *Smart Cities Dive*, Aug. 5, 2024. [Online]. Available: https://www.smartcitiesdive.com/news/escooter-cities-settle-after-turbulence-micromobility/723264

[3] McKinsey & Company, *Micromobility's Emerging Road to Profitability*. McKinsey Center for Future Mobility, 2022. [Online]. Available: https://www.mckinsey.com/features/mckinsey-center-for-future-mobility/mckinsey-on-urban-mobility/micromobilitys-emerging-road-to-profitability

[4] City of Austin, *Shared Micromobility Vehicle Trips (2018–2022)*. Austin Open Data Portal, 2022. [Online]. Available: https://data.austintexas.gov/Transportation-and-Mobility/Shared-Micromobility-Vehicle-Trips-2018-2022-/7d8e-dm7r/about_data

[5] University of Texas at Austin, *Academic Calendar 2021–2022*. UT Austin Registrar, 2021. [Online]. Available: https://registrar.utexas.edu/calendars/21-22

[6] University of Texas Athletics, *2021 Football Schedule*. Texas Longhorns, 2021. [Online]. Available: https://texaslonghorns.com/sports/football/schedule/2021

[7] Y. Zhang and X. Li, "Micromobility and sustainable urban transport," in *Smart Transportation Systems 2022*, Singapore: Springer, 2022, pp. 15–34. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-19-2894-9_2

[8] International Transport Forum, *Safer Micromobility: Technical Report*. OECD/ITF, 2021. [Online]. Available: https://www.itf-oecd.org/sites/default/files/safer-micrombility-technical-report.pdf

[9] National Association of City Transportation Officials, *Guidelines for Shared Micromobility*. NACTO, 2019. [Online]. Available: https://nacto.org/wp-content/uploads/NACTO_Shared_Micromobility_Guidelines_Web.pdf