CSE 6242

Data & Visual Analytics

OptiRide: An Interactive Approach To Enhance Uber's Atlanta Experience with Real-Time Recommendations

Team 11 “The Data Drivers”

James Ashworth
Sean C Kapsal
Irina Low
Allison N Martin
Ronald Mosness

December 3, 2023

Contents

Introduction: Team 11 is creating a proof-of-concept design for a real-time travel time information and recommendation system with an interactive user interface for Uber's Atlanta territory to optimize ride hailing services for Uber customers (Georgia Tech constituents), Uber Drivers, and City Planners modeled on a year's data (2019Q2 - 2020Q1) of the Uber movement data project [20]. This dataset contains over 23 million mean travel time observations aggregated by month of year, hour of day, and day of week. There are roughly 239K unique trips (directed source/destination pairs) between 831 locations around the greater Atlanta area represented. We supplemented this data set with latitude and longitude information from the United States Census Bureau. We also used the Google geocoding API to collect name information for each geography in our data

Problem Definition: Today, Uber customers (Georgia Tech students) rely on the travel time estimates provided by their travel app, Uber drivers use their own experience, travel pay rates, and largely guess on the best place to position themselves to maximize earnings, and city planners commission expensive traffic studies [11,12] and try to use complicated geographical information system solutions to gain insights into travel patterns. Travel apps can provide misleading travel times to Tech students if they use average travel time only [13]. Guessing is not an effective strategy for Uber drivers. Traffic studies, which city planners rely on, can be expensive and take a long time to complete [11,12]. They do not provide large volumes of data over the periods and areas that can be accomplished using data collection techniques similar the Uber movement data project [20].

Travelers need travel time data that is reliable and accurately conveys how long their trip could take. There are real consequences for being late. "A trip that usually takes a half-hour, with little or no warning, takes an hour. Now the motorist is late for work, has missed a doctor's appointment, or is facing hefty childcare penalties for picking up the kids late. Maybe a trucker gets held up in unexpected traffic, making shipments late to the manufacturer, disrupting just-intime delivery, and losing the competitive edge on other shippers" [13]. Being late is more important to travelers than being early. We need to communicate to travelers how long a trip could take on a bad traffic day. Calculating travel time by simple average is not enough when the consequences of being late can be so important.

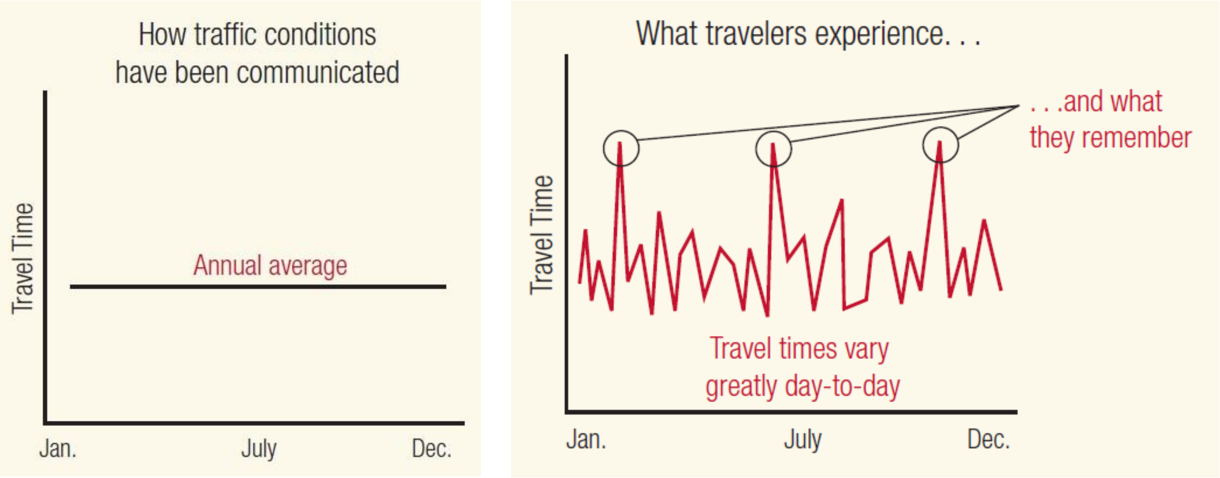


Figure 1a and 1b [13]: Communicating travel time by average (1a left) vs communicating peak travel time day information (1b right).

Communicating travel time by average (1a left) does not provide the level of detail that travelers need. The peak travel time day information (1b right) gives the travelers more detail and provides the worst-case scenario information.

Drivers make their own schedule, so they need information to tell them the best time to work. By leveraging information such as when is the busiest hour, day of the week, and month, they can best prepare their schedule to be optimal for profitability.

City Planners can monitor congestion levels, identify traffic bottlenecks, and analyze traffic flow patterns. This information allows them to adjust traffic signal timings, implement temporary traffic diversions, or deploy resources effectively to alleviate congestion during peak hours. For instance, if the system detects a sudden traffic build-up on a specific route, city planners can promptly redirect traffic to alternative routes or adjust signal phases to optimize the flow of vehicles. [4]

Literature Survey:

The travel time handbook [10] establishes guidelines and provides advice on all facets of collecting travel time data while [13] does point out flaws in collected travel time data. References [2, 3, 11, 14, 15] help understand how a major city works with transit agencies and currently uses this data to maintain a balance between old and new transportation services. There is context on study costs and prices [1, 12, 16] to help the monetization process of the study. The targeting of Georgia Tech students in our implementation was aided by the fact that “younger, better-educated individuals are more likely to adopt on-demand ride services,” [17] which is backed up by a survey conducted in San Francisco where “84% of ride-sourcing customers had a bachelor’s degree or higher” [18].

Proposed method and Innovations:

The uniqueness of our project is in the integration of several crucial components from the field of transportation data analysis, making it a special and useful tool for various stakeholders. First and foremost, we are using a comprehensive dataset that spans from the second quarter of 2019 to the first quarter of 2020. This ensures the accuracy and relevancy of the data. We will be able to identify and examine trends and patterns over time thanks to this historical dataset, something that has not been done thoroughly in the context of Atlanta's Uber data.

If our proof-of-concept system is accepted and developed into a customer product, users will have an intuitive experience thanks to real-time, zone-to-zone traffic flows, and congestion heat maps.

We are providing a deep understanding of transportation patterns and trends, enabling well-informed judgments, thanks to our utilization of extensive historical data. We are promoting collaboration and concurrently solving several pain areas by providing for the requirements of students, Uber drivers, and municipal planners. This multifaceted strategy guarantees a large user base. The use of visual components improves the understanding and enables users to swiftly draw important conclusions from this large dataset (23M observations [20]).

Data, calculations, and analysis:

We used data from the Uber Movement project [20]. This dataset consisted of 23M mean travel time observations from 239,290 trips (directed pickup/dropoff pairs) across 831 Locations from the Atlanta area. Uber provided a shapes file which we used as a crosswalk between the Uber location identifiers and the census tract geo locations. We also used this shapes file to draw the census tract shapes in the visualization. We supplemented the Uber movement data with central tract latitude and longitude from the U.S. Census TIGER/line data [21]. For the census data, we had to convert the provided Dbase file to csv using excel. We also use the Google geocoding API [22] to perform reverse geocoding. Once we had latitude and longitude for each trip, from the census data and a Google API key with the correct permissions, we were able to call the Google geocoding API to do reverse geocoding to get collections of nearest address to that latitude and longitude. We used this address information to provide names for each of the, otherwise unnamed, census tracts.

From this data, we calculated the degree out and degree in for each of the 831 locations in our trip data. These metrics can be used by our city planners to identify busier and less busy travel locations. This could be used by our drivers to identify where to station themselves to ensure business.

We calculated minimum, mean, and maximum travel time observed for each of the 239,290 trips. We also calculated the metrics recommended in the Travel Time Reliability Study [13], 95th percentile travel time, buffer index, planning time index, buffer time, and planning time for each trip. All of these measures help give the traveler a better understanding of how long the trip could take. They allow our system to provide more accurate and reliable estimates of travel time for each trip.

The 95th percentile travel time is a metric for what the worst travel times or worst traffic days look like for each trip. The buffer index is a calculated ratio of how much extra "buffer time" a traveler should expect to add to their normal travel time. "The buffer index is computed as the difference between the 95th percentile travel time and average travel time, divided by the average travel time." [13]. A traveler would multiply the buffer index by the mean travel time to get the "buffer time". The buffer time is then added to the mean travel time to get the travel time estimate. Buffer time is the extra time required for the trip. "Buffer time is calculated as the difference between the 95th percentile travel time and the average travel time." [13]

The Planning Time Index is the estimated total travel time multiplier. To ensure on-time arrival, we would multiply the planning time index by the traffic-free (free-flow) travel time. "The planning time index is computed as the 95th percentile travel time divided by the free-flow travel time." [13]. We used the minimum travel time for the trip as the traffic-free travel time. Planning time is "the total travel time, which includes buffer time (i.e., calculated as the 95th percentile travel time or mean travel time plus buffer time)." [13]

An additional method we used to analyze and predict travel times in our system was to create a linear regression model for each of the 239,290 trips. We trained 239,290 Linear Regression models, one model per trip (ordered pickup drop off pairs). We used these models to predict travel time during morning rush hour, evening rush hour, weekday non-rush-hour, and weekends using feature engineering. We found that these models could explain roughly 40 percent of the variance in travel time (R squared = 0.4037). We predicted the travel time for each of the travel time types (AM Rush, PM Rush, etc) and calculated the 95 percent (alpha = 0.05) prediction intervals (upper and lower) for each travel time type.

We used the Mapie (Model Agnostic Prediction Interval Estimator) Python package to perform 5-fold cross validated regressions and calculate the 95% prediction intervals for our trip travel time estimations. Mapie will only calculate a guaranteed prediction interval if you have more than (1 / alpha) observations available. For our 95% (0.05 alpha) confidence level, we needed more than 20 observations. We had 29,928 trips with 20 or fewer observations. For these trips we calculated the prediction intervals after prediction using the formula: prediction plus or minus the one minus alpha quantile of the absolute error.

We split the training into two sets to be run on two different machines. One machine to train the models for trips with more than 20 observations and one to train the trips with 20 or fewer observations. Training and predictions for the 20 or fewer observation trips (n = 29,928) ran for just over 12 hours on an intel-based PC with an i7 processor and 64 GB of RAM. We used a gaming laptop with an intel based i9 processor and 64GB of RAM to train the models for the trips with more than 20 observations. That training was occurring at a rate of approximately 500 models and predictions per minute. We trained 209,362 models in roughly 7 hours on this machine.

The difference that the computing power had on model training and prediction speed was substantial. This is a consideration to keep in mind when training large models or scoring a large number of observations using cost intensive cloud service provider compute resources. Careful consideration should be given to cost per hour between different sized/spec'd machines. It may be worth it to provision a larger machine to complete training and scoring significantly faster.

The model training time will be a challenge that needs to be solved before we can turn this proof of concept into a real time solution. We would have to work to speed up model training and scoring. Another option would be to separate training and scoring. Possibly train the models weekly or on some other cadence but be able to use the trained model to score/predict trip travel time in real time.

To reduce risk from a computer crash or model error, we created a loop to build regression models and after each trip iteration, the predictions and prediction intervals were appended to a csv file on disk. This allowed us to train on two separate machines and combine the results into one data file later for visualization. This would also allow us to split the data further and separate the training and scoring between more computers, reducing training time further.

These various calculations, predictions, and prediction intervals, enable us to provide robust estimations of how long a given trip would take for various travel times. A selected subset of this information is presented to our users in the interactive visualization.

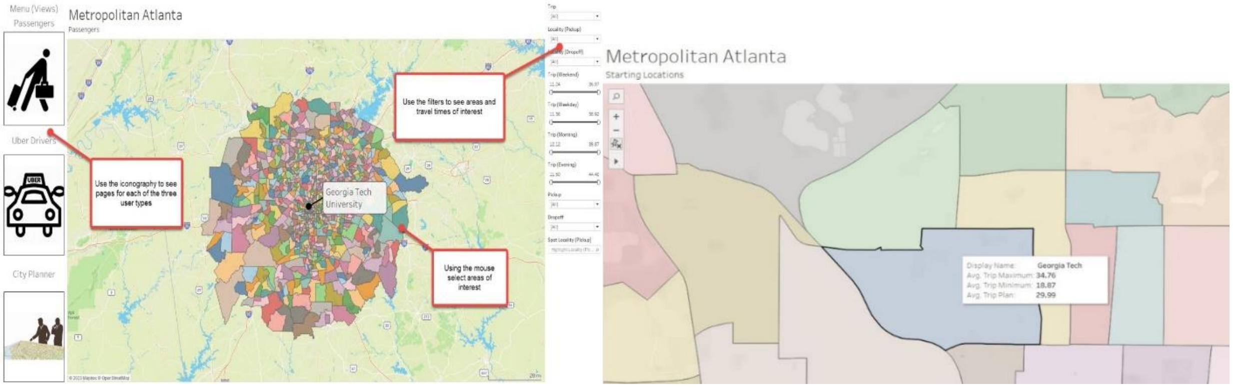
For GA Tech Students and other institute travelers, our system goes beyond the typically provided average/estimated travel duration between locations. It calculates the metrics proposed in reference [13], described above, and provides travelers with time recommendations that help them account for worst case scenario travel days and increase their likelihood of arriving on time. “Travel time reliability measures are relatively new, but a few have proven effective… The most effective methods of measuring travel time reliability are 90th or 95th percentile travel times, buffer index, and planning time index” [13]

We calculated these leading travel time reliability metrics, as well as performed linear regression to predict travel time and a 95% confidence interval for how long their trip will take. We calculated and provided a subset of these metrics specific to their selected trip, travel month, day, and time of day. We may also be able to recommend better travel parameters (Month, Day, Hour).

For Drivers we provide similar metrics to the student’s perspective by utilizing linear regression results with a 95% prediction interval using the same data, but the scenario differs in that it will display the best case for profitability. By visualizing the busiest times, drivers can determine the peak hours in their given market. Uber’s surge pricing entices the drivers to pick up riders during times of high demand, which can increase the driver’s compensation on a given night.

To support City Planners, our tool incorporates a clustering mechanism where each route within the Atlanta territory is systematically assigned to one of three distinct clusters based on two key factors: mean travel time and the number of trips. City Planners can discover clusters of routes that experience similar traffic patterns or identify regions within the city or service area with distinct travel time characteristics.

User Interface Description: In Tableau, we created a user-friendly interface for analyzing Uber trip data in and around the Atlanta area, with a visually informative spider chart (origin-destination or flow maps) extending outward from the Georgia Tech campus as the central point of reference. The interface provides users with a dynamic map that displays pick-up and drop-off locations, color-coded to represent different trip characteristics like trip duration. Additionally, users can filter the data based on various parameters such as date, time of day, and specific Uber service types. The spider chart is integrated into the interface, enabling users to visualize trip patterns and distances from Georgia Tech, with radiating lines representing trip destinations and their respective frequencies. This intuitive interface empowers GT students, Uber drivers, and city planners and gives them the ability to explore Uber trip data effectively, gaining valuable insights into travel patterns and optimizing transportation options in the Atlanta area.



Use the iconography to see pages for each of the three user types

Use the filters to see areas and travel times of interest

Using the mouse select areas of interest

Display Name: Georgia Tech
Avg. Trip Maximum: 34.76
Avg. Trip Minimum: 18.87
Avg. Trip Plan: 29.99

Figure 2a and 2b: The menu allows selection of one of the three customer categories: Passengers, Uber Drivers, or City Planners. The filters allow for selection of routes, localities, or desired times in minutes. The map allows selection by census tract from pickup or drop-off locations. It then displays the travel times to the census tract the mouse is hovering on. Times included (minutes) AM Rush Hour, PM Rush Hour, Weekday, and Weekend.

Experiments & Evaluation:

Our technology stack includes csv, JSON, and SQLite for data storage, Python and R for data exploration, analysis, and modeling. Our user type specific Interactive Visualization tool is Tableau. After data cleaning, we are leveraging the data to have a comprehensive dataset which we are using to run methods on to provide three various perspectives for viewing.

For Students and other travelers, we calculated trip specific travel times by travel time type: AM Rush Hour, PM Rush Hour, Weekday non rush hour, and Weekend. We also used linear regression to predict travel time and a 95% prediction interval travel time range for each trip. We used R-squared values and MSE to evaluate the performance of our modeling. We were able to explain 40% of the variance in travel times with our models (R^2 = 0.4037). We calculated a buffer index and planning time index. This analysis is non-trivial due to the size of the dataset, the number of calculations, regressions, and predictions performed, and the requirement to provide the results in a way that enables a responsive interactive user interface. There are 23M observations in our dataset comprised of 239,290 unique trips.

For Drivers we conducted a similar procedure using the same comprehensive dataset and performed linear regression to predict 95% prediction interval travel time ranges, but for the opposite scenario. We will also used R-squared for evaluation of our models, and we display the best times specific to the driver, rather than from the student's perspective.

City Planners to pinpoint peak hours, we rely on historical hourly average travel times. Our analysis reveals that travel times significantly extend during hours 6-8 and 15-18. We employed K-means clustering, employing the Elbow Method to determine the optimal cluster count. Through this method, we segmented the Atlanta territory into three discernible clusters, each representing unique travel behaviors: interconnected highway communities, the MARTA transit corridor, and the suburbs. These clusters offer insight into the varied traffic dynamics across the city. For example, the 'interconnected highway communities' exhibit high-speed traffic prone to sporadic congestion at certain times. In contrast, the 'MARTA transit corridor' indicates more consistent but potentially slower-paced traffic influenced by public transit. Meanwhile, the 'suburbs' cluster portrays a pattern of steady, lower-density traffic. Visualized through our user interface's 'Cluster (3 Set)' filter, these findings empower city planners to tailor traffic management strategies and urban planning initiatives according to the distinct characteristics of each area.

Conclusions, discussion, future work:

This 23 million observation dataset, supplemented by census data and google API calls, calculations, combined with our non-trivial data analysis, calculations, regression models, and highly interactive visualization has allowed us to complete a proof-of-concept design for providing more reliable travel time estimates for Uber passengers, drivers, and city planners. The main goal of our project has been completed.

Along the way, through our comprehensive literature survey, our analysis and visualization work, we learned a great deal about analysis and visualization of big data. We have learned a great deal about travel time estimation and geographic area visualization in Tableau.

One challenge has been that our dataset provider website shut down as of October 1, 2023. We are fortunate that we downloaded the required data files and shapes information before that happened.

For future work, we may present our proof-of-concept platform to Uber, our project beneficiary, for their consideration. If they like our work, our team would be happy to partner with Uber and develop the real-time version of our solution using their real-time trip data. Work is needed in speeding up our regression algorithm and scoring process to allow it to be used in a real-time system. The compute resources and cost would have to be improved as well. We would also like to incorporate text-based prompts to provide recommendations for better times to travel. For city planners, we would like to add supplemental data about the areas in our visualization. This could include population, demographic, food availability, economic, and healthcare access data. These efforts were all beyond the time available for this proof-of-concept design.

Team Activity Statement: All team members are actively involved, attending team meetings, and contributing a similar effort.

References

[1] UC Davis (n.d.). Sustainable Transportation. New Research on how Ride-Hailing Impact Travel Behavior By Regina R. Clewlow, Ph.D. - STEPS (Ucdavis.edu). https://steps.ucdavis.edu/new-research-ride-hailing-impacts-travel-behavior/

[2] Bloomberg (2021, September 21). Ride Hailing Hidden Costs. Adding Up Ride-Hailing's Hidden Environmental Costs - Bloomberg. Retrieved September 23, 2023, from https://www.bloomberg.com/news/articles/2021-09-30/adding-up-ride-hailing-s-hidden-environmental-costs?utm_content=citylab&utm_campaign=socialflow-organic&utm_medium=social&utm_source=twitter

[3] Mi Diao, Hui Kong & Jinhua Zhao (2021, February 1). Impacts of transportation network companies on urban mobility. Impacts of Transportation Network Companies on Urban Mobility. Retrieved September 23, 2023, from https://www.nature.com/articles/s41893-020-00678-z

[4] Xiao Han., Yun Yu., Gao, Z. Y., & Zhang, H. M. (2021). The value of pre-trip information on departure time and route choice in the morning commute under stochastic traffic conditions. Transportation Research Part B: Methodological, 152, 205-226. https://doi.org/10.1016/j.trb.2021.08.006

[5] Arnott, R., de Palma, A., & Lindsey, R. (1991). Does providing information to drivers reduce traffic congestion? Transportation Research Part A: General, 25(5), 309-318. https://doi.org/10.1016/0191-2607(91)90146-H

[6] Liu K. (2022). Analysis of the Conflict between Car Commuter's Route Choice Habitual Behavior and Traffic Information Search Behavior. Sensors (Basel, Switzerland), 22(12), 4382. https://doi.org/10.3390/s22124382

[7] Barceló, J. (2010). Fundamentals of Traffic Simulation (1st ed., pp. XVIII, 442). Springer New York, NY. https://doi.org/10.1007/978-1-4419-6142-6

[8] D. L. Gerlough (1965). Simulation as a Tool in Traffic Control System Evaluation. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-1722-7_5

[9] Boxill, S. A., & Yu, L. (2000). An Evaluation of Traffic Simulation Models for Supporting ITS Development (Report No. PB2001102338). Retrieved from https://ntlrepository.blob.core.windows.net/lib/17000/17500/17586/PB2001102338.pdf

[10] Turner, Eisele, Benz & Holdener (1998). Travel time data collection handbook (Report No. FHWA-PL-98-035) https://doi.org/10.21949/1404545 Retrieved from https://www.fhwa.dot.gov/ohim/tvtw/natmec/00020.pdf

[11] Uber Movement Team (2018, Jan 30). Visualizing Access to Healthy Food Options in Cincinnati. Retrieved from https://www.theoceancleanup.com/updates/whales-likely-impacted-by-great-pacific-garbage-patch

[12] B.W. Yarger (2023, Oct 5) What are Traffic Studies. Retrieved from https://www.yargerengineering.com/articles/study.html

[13] Taylor (2005, Dec 1) Travel Time Reliability (Publication No. HOP-06-070) Retrieved from https://transportationops.org/publications/travel-time-reliability-making-it-there-time-all-time

[14] Pyrialakou, V. D., Hajibabaee, P., Williams, A., & Farrokhvar, L. (2023). Integrating ride-hailing services with transit: An exploratory planning framework. *Journal of Public Transportation*, *25*, 100056.

[15] Oviedo, D., Scorcia, Y., & Scholl, L. (2021). Ride-hailing and (dis) Advantage: Perspectives from Users and Non-users.

[16] Bashir, M., Yousaf, A., & Verma, R. (2016). Disruptive business model innovation: How a tech firm is changing the traditional taxi service industry. Indian Journal of Marketing, 46(4), 49-59.

[17] Cho, Y., & Rodgers, G. (2018). What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. Transportation Research Board. https://doi.org/10.1016/j.tbs.2018.06.002

[18] Cervero, R., Chan, N., Dai, D., Rayle, L., & Shaheen, S. (2016). Just a better taxi? A survey-based comparison of taxis, transit, and ride sourcing services in San Francisco. Transport Policy, 45, 168–178. https://doi.org/10.1016/j.tranpol.2015.09.010

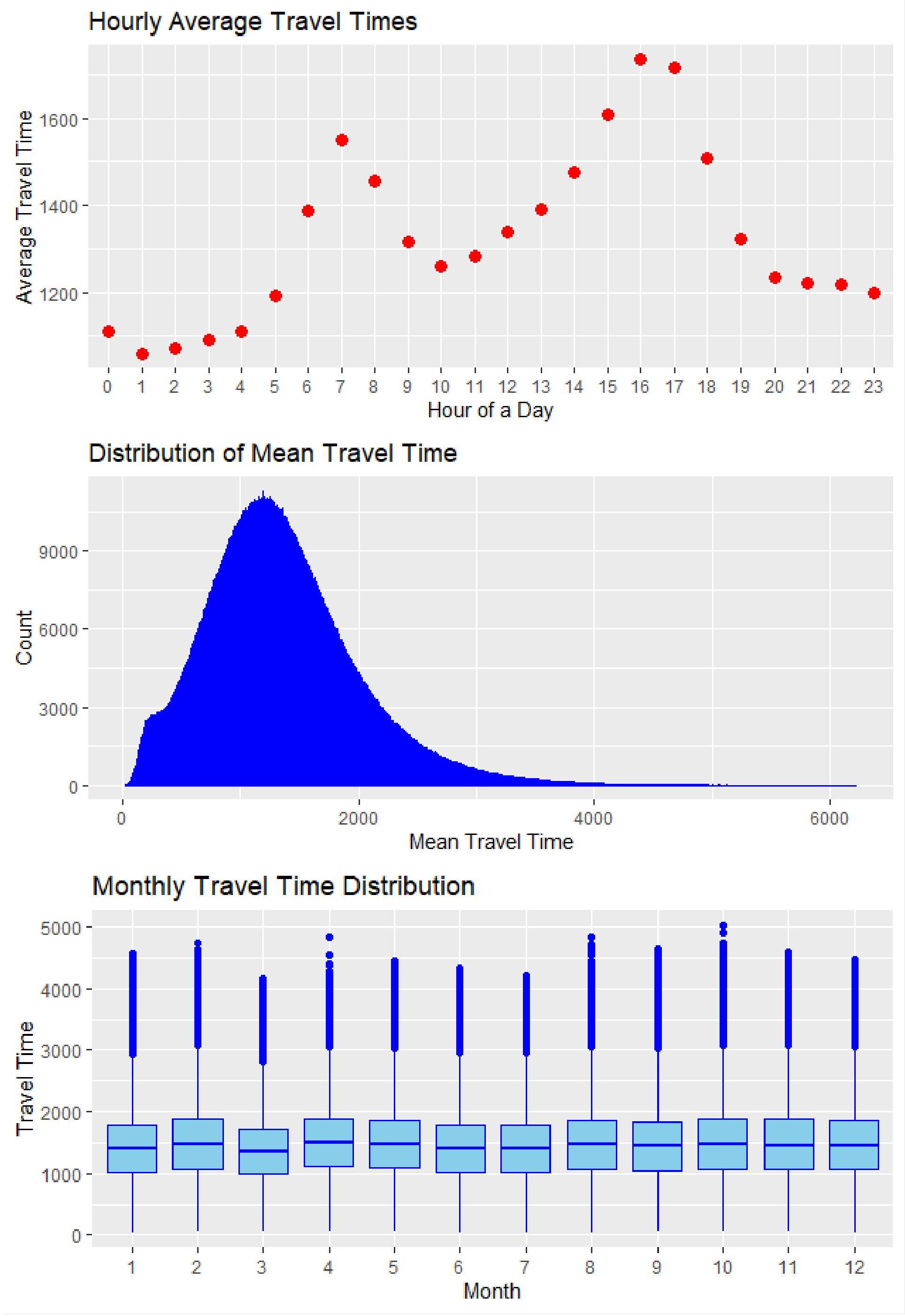
[19] Wang, M., & Mu, L. (2018). Spatial disparities of Uber Accessibility: An exploratory analysis in Atlanta, USA. Computers, Environment and Urban Systems, 67, 169–175. https://doi.org/10.1016/j.compenvurbsys.2017.09.003

[20] Uber. Uber Movement data project. Retrieved Sept 2023 from https://movement.uber.com

[21] United States Census Beureau. TIGER/Line. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2019.html, https://www2.census.gov/geo/tiger/TIGER2019/TRACT/tl_2019_13_tract.zip

[22] Google. Geocoding API. Retrieved from https://developers.google.com/maps/documentation/geocoding/overview, https://developers.google.com/maps/documentation/geocoding/requests-reverse-geocoding

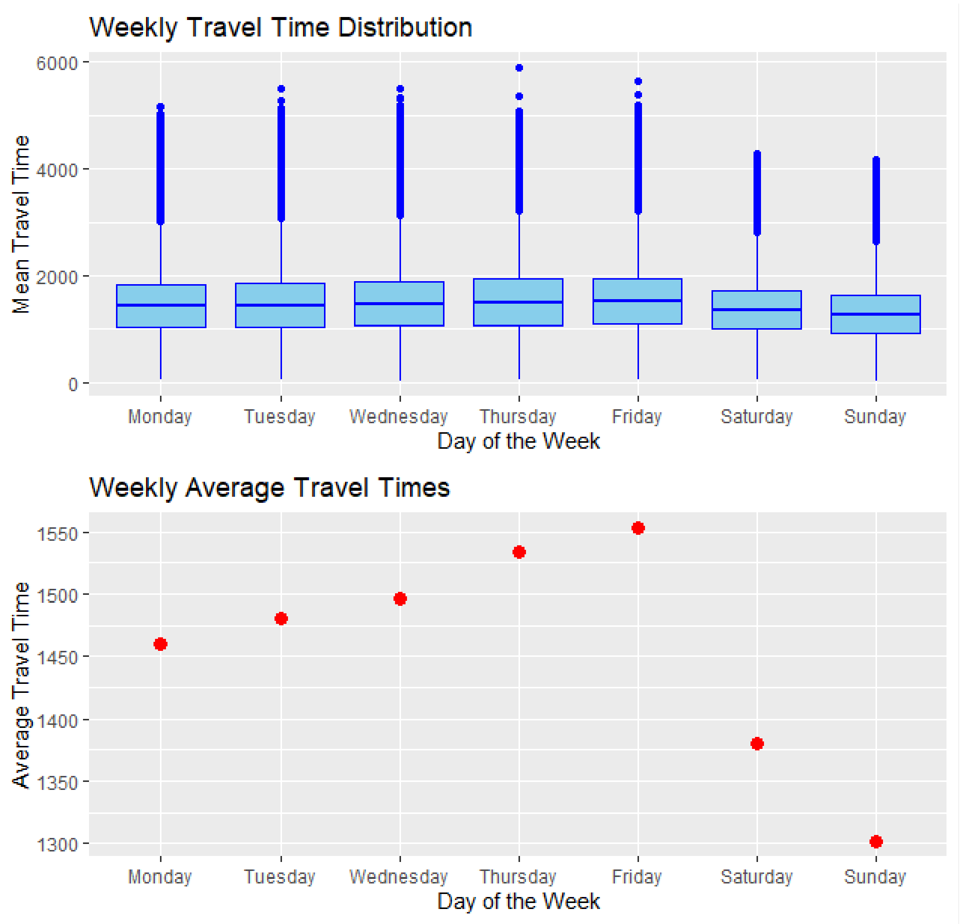
Appendix



Hourly Average Travel Times

Distribution of Mean Travel Time

Monthly Travel Time Distribution



Appendix Acknowledgements

This project was made possible by the Uber Movement data project [20]. Without the access to this open dataset, this project would not be possible. We also must acknowledge the provider of our shape information.

- Data © OpenStreetMap contributors, http://www.openstreetmap.org/copyright
- Geoid from JSON file can be used to get details about the tract https://datacommons.org/browser/geoId/13121008301
- The census geocoder API can give more info from lat and long https://geocoding.geo.census.gov/geocoder/geographies/coordinates?x=-84.444288&y=33.757668&benchmark=2020&vintage=2020
- (TIGER/Line® files) U.S. Census Bureau, http://www.census.gov/