The UChicago Lyft Ride Smart Program: Effects on Ridesharing

By Abe Burton, Ridhi Purohit, Harsh Vardhan Pachisia, and Rohit Kandala
11/29/2023

Research Question

How much did UChicago's Lyft Ride Program impact ridesharing in Hyde Park and is it worth continuing?

- We used the "Transportation Network Providers" dataset from the Chicago Data Portal.
- This program launched September 2021, but expanded after a student had been killed on November 9, 2021. It was scaled back in July 2023 due to "environmental concerns".
    - This constriction has been met with criticism from The Maroon as official threaten further cuts.

Pre-Program

Time period with no Lyft Ride Smart.

Program Begins

Students can take 10 rides per month up to $15/ride. On Thu, Fri, Sat from 9PM to 4AM.

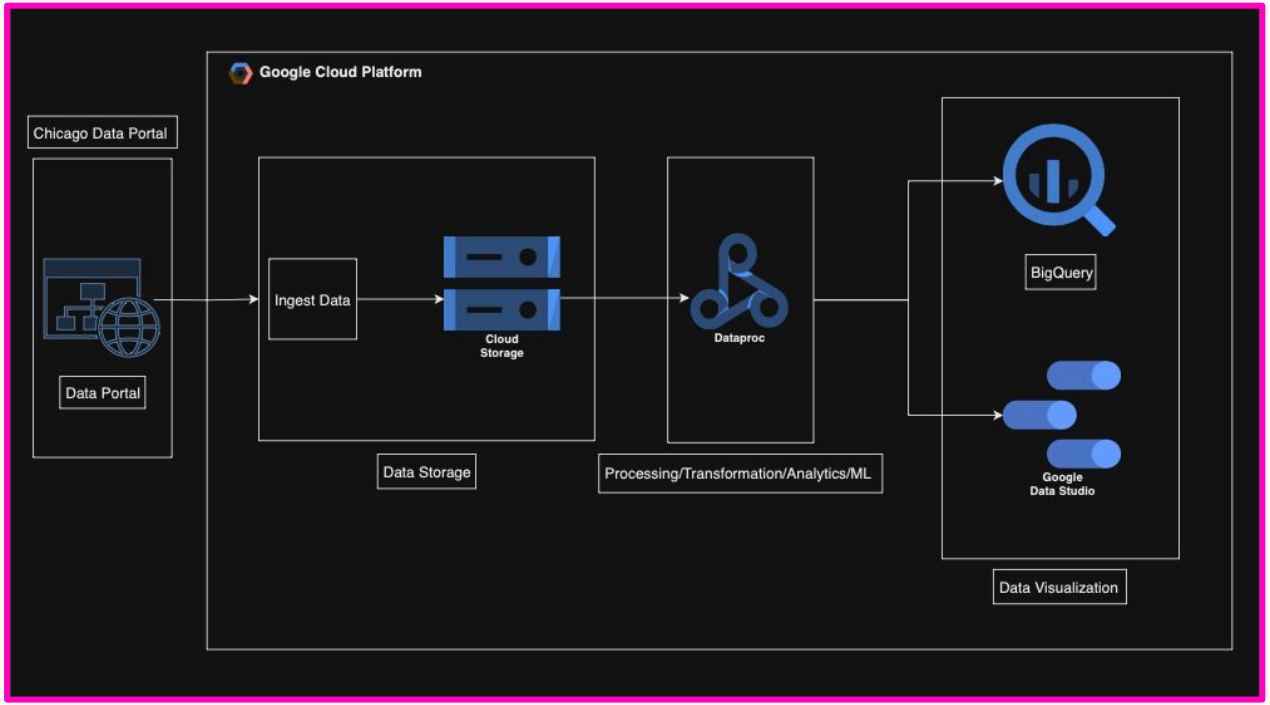
Expansion

Students can use the pass every day from 5PM to 4AM. Still 10 rides/month up to $15/ride.

Constriction

Program constricted to 7 rides/month up to $10/ride with the same time(s).

Project Architecture



Google Cloud Platform (GCP) had everything we needed for storage, processing, analysis, machine learning, and visualization.

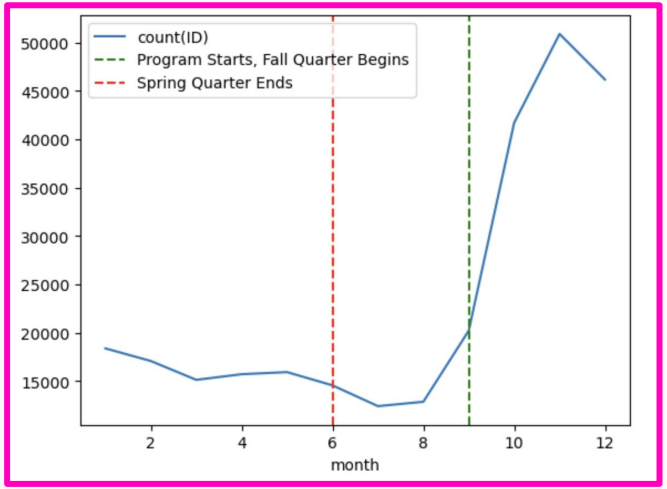
Our total data was **84.4 GB**.

Exploratory Data Analysis (EDA): Structure of Data

- The Chicago Data Portal is easy to use, and it was great to have each ride be a row in the dataset with 21 associated columns. However, we ran into a couple of challenges with our EDA:
  - Not grouped by rideshare provider.
    - Further complicated our research question.
  - We excluded 2020 data as there were too many confounding variables.
  - 2023 data was incomplete (up to September) and contained more errors on average.
  - Could not fully take advantage of geospatial analysis with our user-managed environments.
  - Great data size disparities by year.
- Initially, we had planned to layer in Chicago's crime(es) dataset in addition to the rideshare and weather dataset, but had to exclude crime to reduce scope.

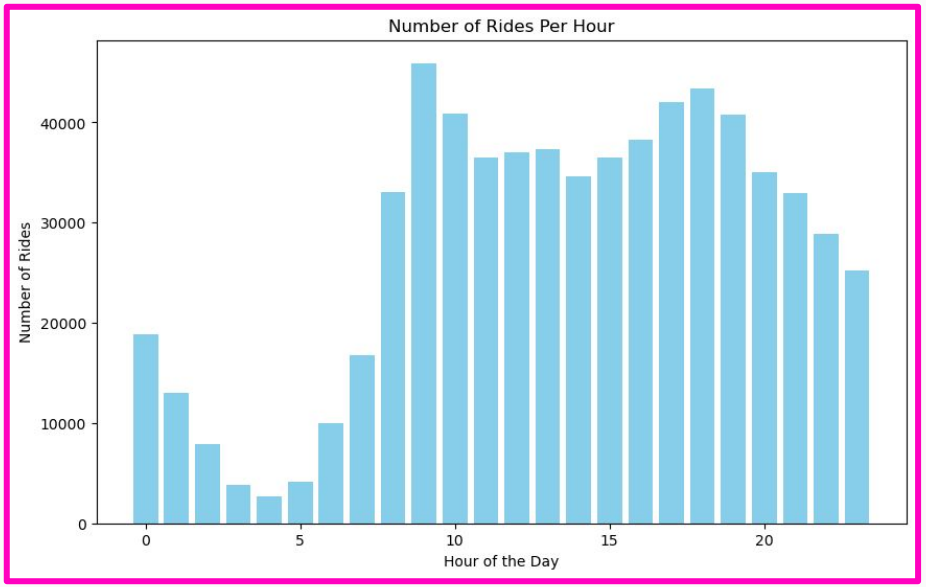
Exploratory Data Analysis (EDA): Analysis Results

- To ensure standardization, we ran the same PySpark code on each dataset (split by year) and found the following about rideshares in Hyde Park:
  - 2019 had the highest number of rides; 2020 had the lowest.
  - Before the program, more rides were taken during the evening.
    - **BUT,** after the program was implemented, not only did the count of rides increase, but also there would be a sharp spike at 5 PM.
  - Regardless of year (except 2020), rideshares drop off around June (summer starts), and pick back up when the Fall Quarter starts, with an exaggerated effect in 2021 (when program started)

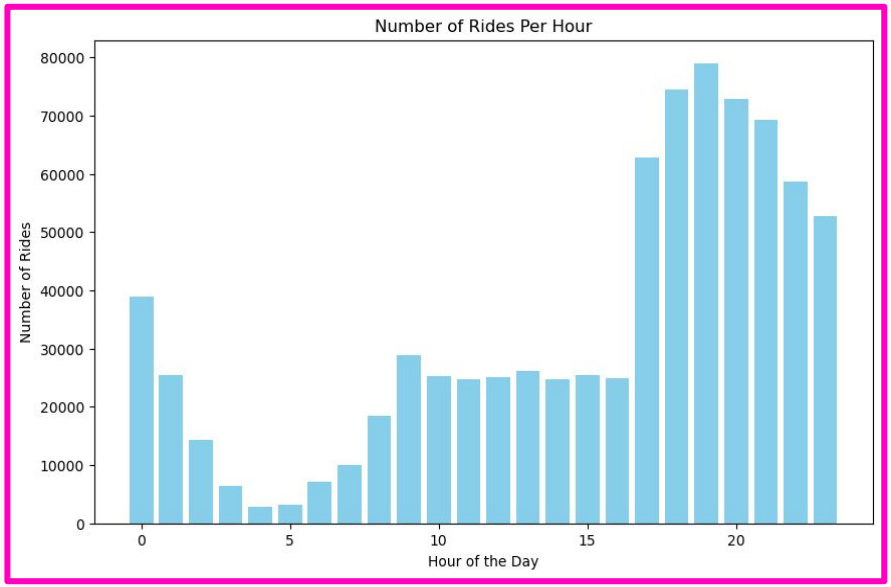


Rideshares by Month (2021)

Exploratory Data Analysis (EDA): Analysis Results



Rides per Hour (pre-program)



Rides per Hour (post-program)

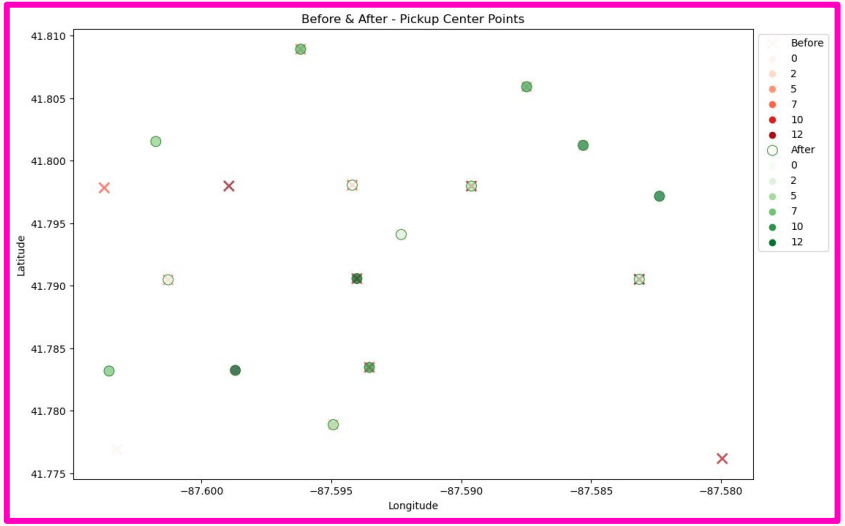
Unsupervised Machine Learning: Problem/Model

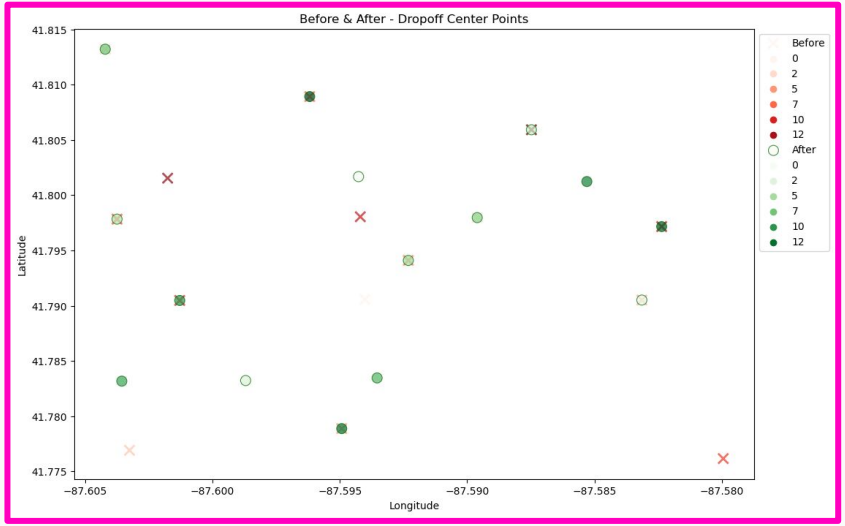
- The question we want to address with unsupervised learning is if there are any patterns in the locations where riders book a cab from and how that has changed with the introduction of the Lyft Ride Program.
    - Could the locations where rides are booked from and locations where dropoff happens be grouped together?
    - How do these clusters relate to well-known spots in Hyde Park where students take rides from?
    - **Model: K-Nearest-Neighbors (KNN)**
        - Features: Latitude & Longitude.
        - Evaluation metric: Silhouette Score.
        - Iterating to find “best-k” (15).
        - Create clusters based on pickup and dropoff locations for data from before the program (2018, 2019) and after the program (2021, 2022)
- **Hypothesis #1:** The program has an effect on pickup locations.
- **Hypothesis #2:** Time of day and day of the week affect ridership.
- **Hypothesis #3:** Given key UChicago locations, the program affects the locations where rides are taken from.

Unsupervised Machine Learning: Results

- We ran five different types of analysis (clustering, centermost, before/after, temporal, and comparison to existing points of interest).
  - The **program did not affect where people were calling rideshares from** in Hyde Park.
  - There was slight **variation in the centerpoint locations** of the pickup clusters before and after the program.
  - Limiting the analysis to clusters of pickup locations before and after the program that share the same centerpoint locations led to the following observations:
    - On average, **trip duration increased** post-program
    - On average, **number of trips increased** post-program
    - On average, **fare increased** post-program
    - **More late-night trips** were taken post-program
  - Focusing on three locations near popular university hotspots (51st, 54th and 57th streets), we found that:
    - Trip duration decreased, possibly indicating that students were substituting routes they otherwise would have walked with free Lyfts.
  - Cluster mapping to centerpoint location ‘‘5746 S Ellis Ave, Chicago, IL 60637, USA’ (a central UChicago location) shows a **marked increase in the proportion of rides taken at the end of month.** Other common clusters show an increasing trend too.
    - This could possibly be an indication of students making use of their pending free Lyft rides before the Lyft Pass resets at the end of the month.

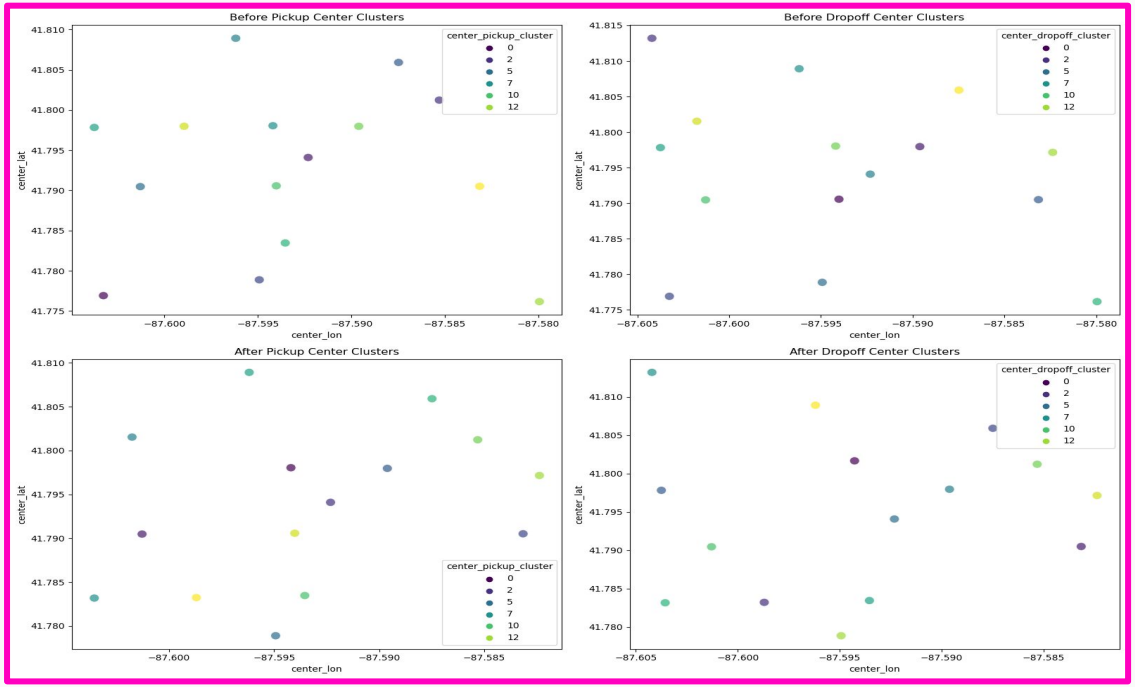
Unsupervised Machine Learning: Results





These two graphs (before and after the program respectively) indicate that centerpoints of pickup and dropoff locations had significant overlap despite the introduction of the program.

Unsupervised Machine Learning: Visualization



The four figures on the left show the before/after clustering and pickup/dropoff locations. The changes are not dramatic.

Table below shows the centerpoint locations which were common amongst the pickup clusters pre vs. post program

center_address
5534 S Dorchester Ave, Chicago, IL 60637, USA
5142 S Hyde Park Blvd, Chicago, IL 60615, USA
6358 S Kimbark Ave, Chicago, IL 60637, USA
4922 S Cornell Ave, Chicago, IL 60615, USA
5746 S Ellis Ave, Chicago, IL 60637, USA
1208 E 47th St, Chicago, IL 60653, USA
1322 E 54th St, Chicago, IL 60615, USA
6122 S Kenwood Ave, Chicago, IL 60637, USA
5719 S Kimbark Ave, Chicago, IL 60637, USA
1455 E 54th St, Chicago, IL 60615, USA
5700 S DuSable Lk Shr Dr, Chicago, IL 60637, USA

Supervised Machine Learning: Problem

- Aim: Showcase the **impact of the Lyft program** on daily ridership counts within the program area (Hyde Park, Kenwood, Woodlawn) by **predicting how behavior would have been** without the program (predict counter-factual)

- Processing dataset:
  1. Calculate daily ridership count by the 77 community areas of Chicago
  2. Merge in daily weather data: temperature, snow, precipitation, time of sunset, etc
  3. Y: daily count of rides in-program area, Features: Daily Rides in all other community areas + weather variables
  4. We split our data into three subsets: before the program (up to October 2021), during the program (Oct 2021 to June 2023), and after the 2023 contraction (July 2023 to Present)

Supervised Machine Learning: Approach

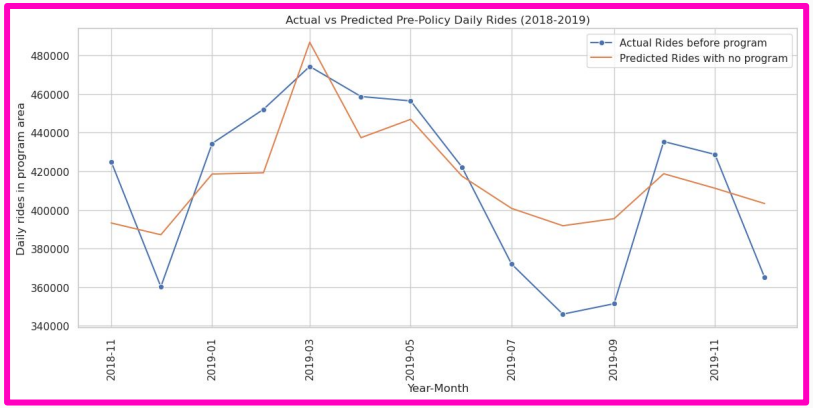
Hypothesis: Actual ridership counts (with program) would be far greater than the model's predictions (w/o program).

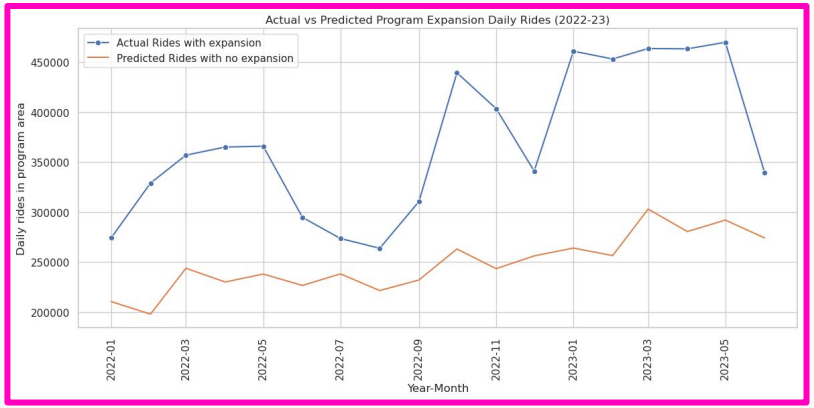
Two-model approach:

1. First model: Train model on pre-program dataset (2018-2019) to predict ridership counts for:
    a. **Program start:** Predict daily counts for October, November, and December 2021 (on weekends only)
    b. **Program Expansion:** Predict daily counts for January 2022 - June 2023 (all days)
2. Second model: Train model on January 2022 - June 2023 dataset:
    a. **Program Contraction:** Predict daily counts for July 2023-September 2023 when the program contracts (7 rides, <$10/ride, all days)

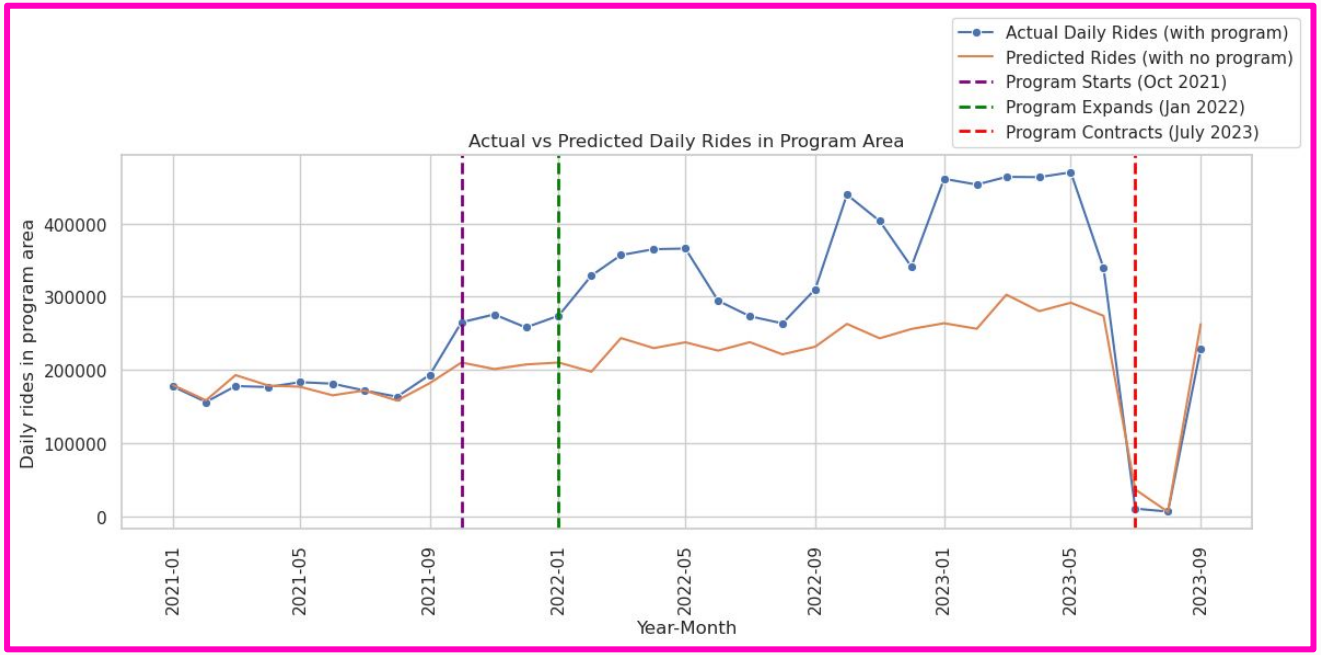
Supervised Machine Learning: Model

- **Model: Linear Regression (Elastic Net)**
  - Features (x): year, month, 74 daily community area ride counts, temperature, precipitation, snow (binary), snow depth, and time of sunset (int).
  - Label (y): ridership counts of in-program areas
  - Powerful model with R^2 of 0.95, but the model had a relatively high RMSE & MAE.





Supervised Machine Learning: Visualization



There was a **clear increase** in ridership counts after the program was implemented.

Usage increase breaks down to about 4 rides per student per month

Conclusion: Lessons Learnt

How much did UChicago's Lyft Ride Program impact ridesharing in Hyde Park and is it worth continuing?

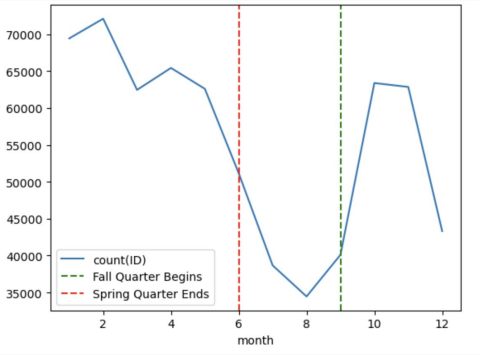
- Through our supervised model, we can say that the ridership count amongst all providers greatly increased in Hyde Park. This increase may be partially due to other factors (COVID-19) but evidence for the program's effect is clear.
    - Rides attributed to the program are a **major factor** because our model follows the actual counts closely when there's no program. **Usage Estimate: 4 rides per student per month.**
- More rides are taken later in the evening supporting a **safety-motivated hypothesis** for Lyft usage.
- Lyft Ride Smart amplify student habits - similar destinations to campus, shopping, and apartments with higher frequency and more trips taken at night.
    - Rides are not for superfluous spending - they facilitate necessary student trips in a wider range of times with increased safety. This is clarified with our clustering analysis.

Appendix

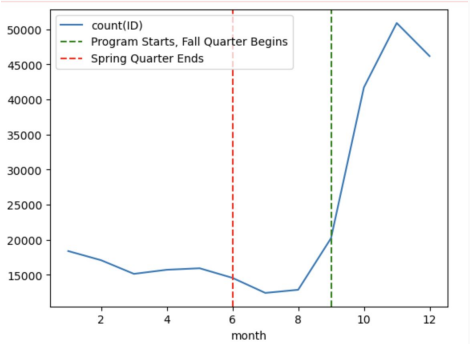
Appendix: Lessons for Next Time

- We would change/expand upon on this project in the following ways:
  - Conduct more analysis into 2020. We know that ridership dropped, and slowly picked up, but do not know much beyond that.
  - Layer in more datasets. We did this with the weather dataset, but comparing this with the crime dataset would be valuable as UChicago expanded the program due to safety concerns.
    - Moreover, this policy is a way to address student safety concerns without changes to police presence.
  - Conduct analysis into average fare and how that affects tips.
    - Does this also increase the number of pooled trips due to higher fares?
    - Do additional charges impact the rideshare counts?

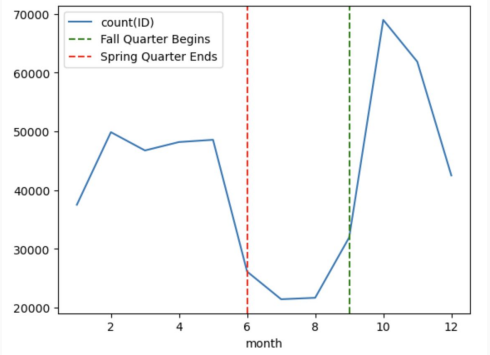
Appendix: Ride Counts



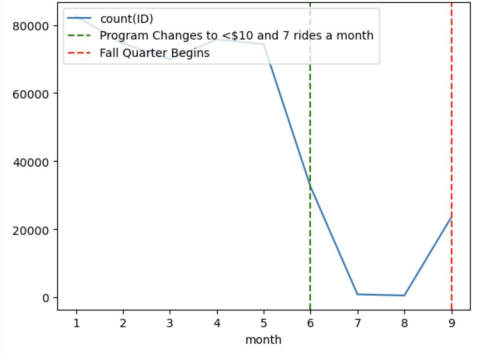
2019 Rides



2021 Rides

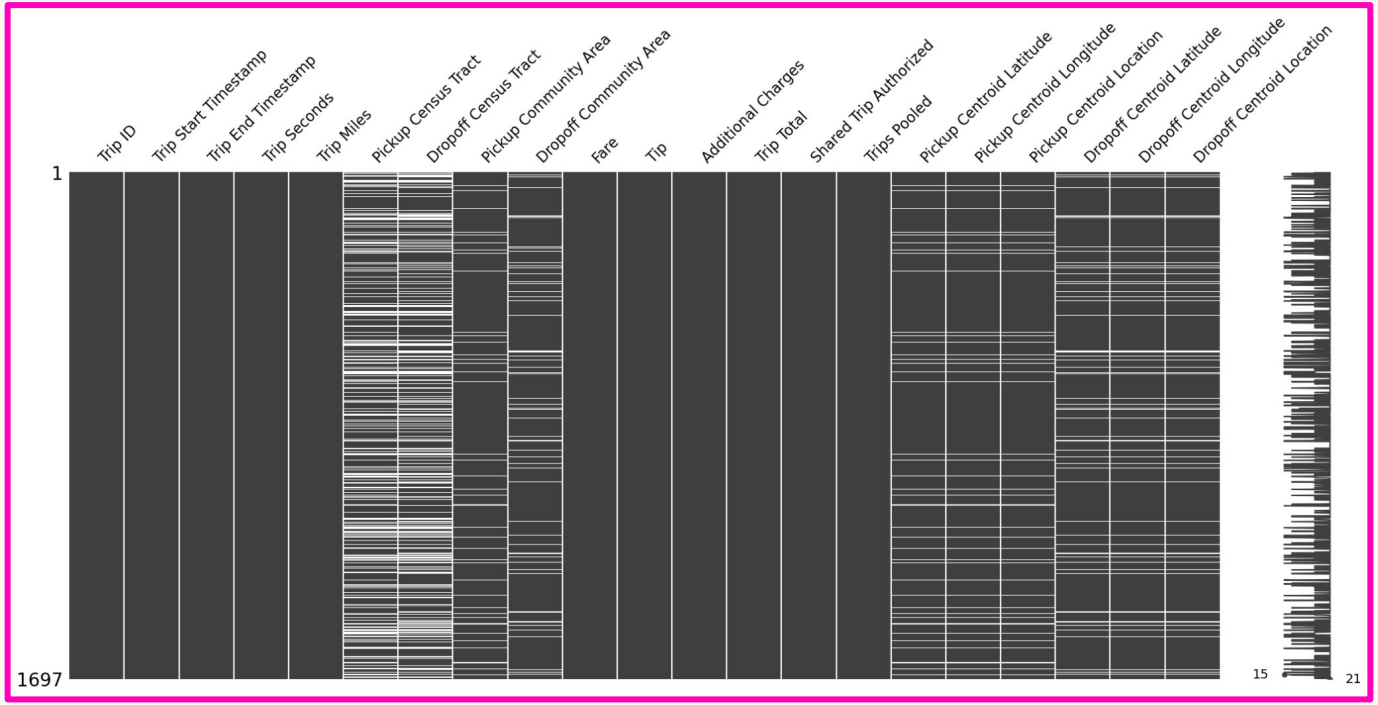


2022 Rides

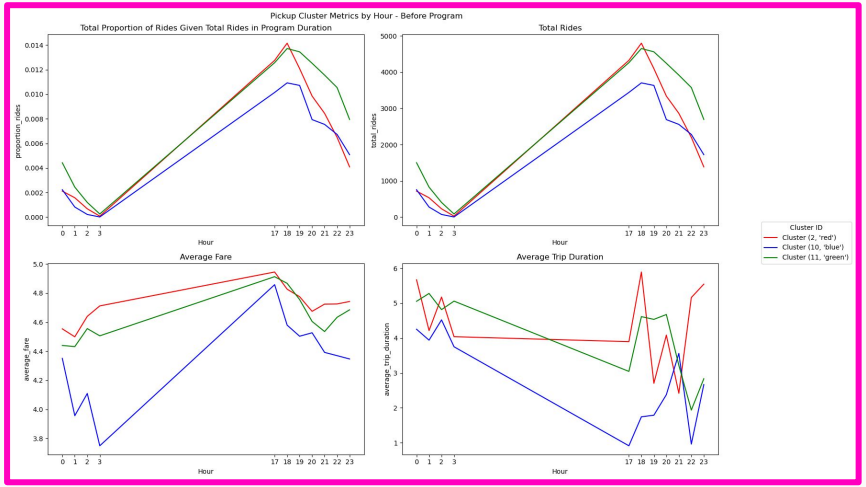


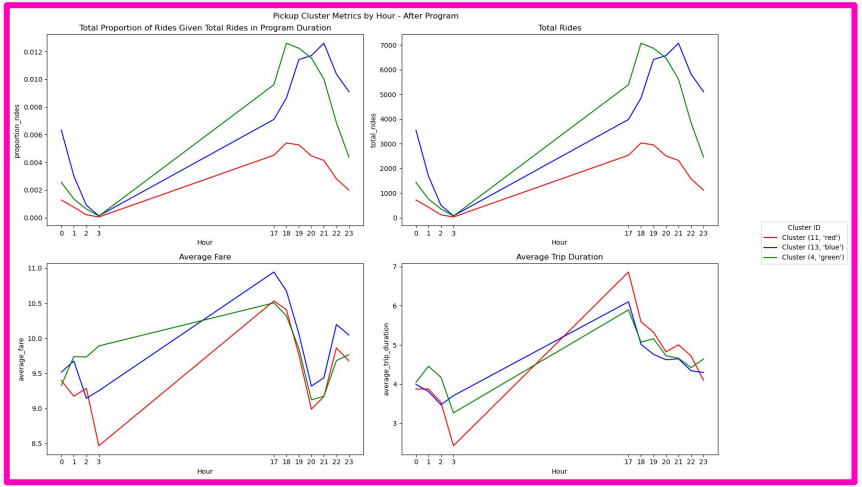
2023 Rides

Appendix: Missing Data for each column



Appendix: Temporal Clustering





These two graph sets show pickup clusters by hour, and are measuring total proportion of rides, ride count, average fare, and average trip duration. They are limited to addresses around the university.

Appendix: Geo-visualization

Chicago Rideshares Geo-Visualization

Heatmap of Rides in the Program Area