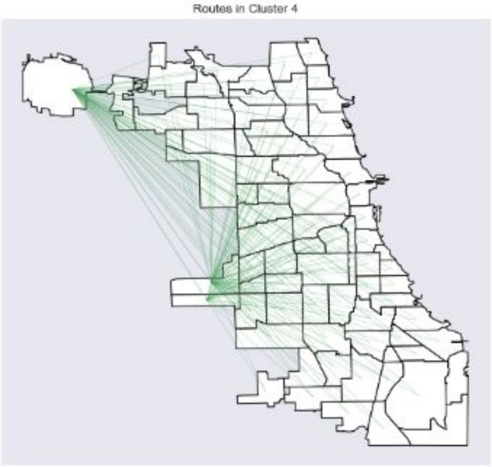
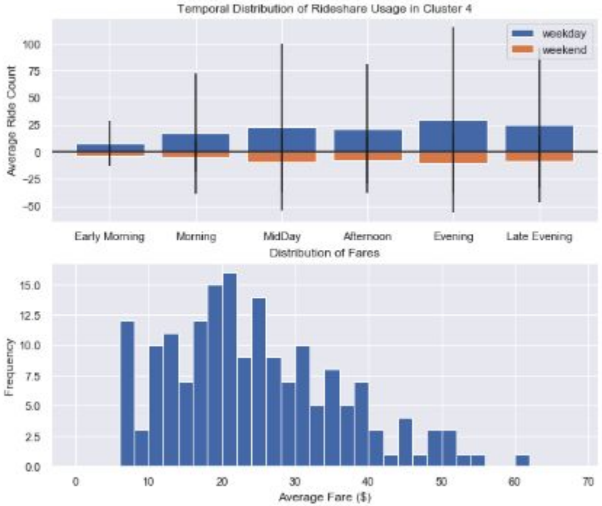
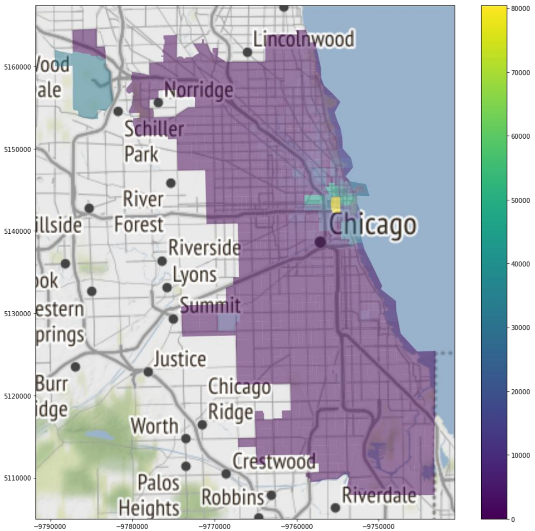






Machine Learning with the Chicago Ridesharing Trips Data Set

Lisa Taylor
Completed as Capstone Assignment 1
Springboard Data Science Bootcamp

Chicago Rideshare Dataset

Since November 2018, Chicago has required “Transportation Network Providers” operating within the City (Lyft, Uber, etc.) to report basic rideshare information.

Trips Dataset:

- > 70 million individual trips.
- Anonymized: trip starting and ending locations are generalized to the nearest census tract, and drivers cannot be linked to particular rides they provided.
- Fields: start/end census tract, start/end time, fare, distance, duration, tip, whether pooling authorized, number of trips in pool.
- November 2018 through March 2019

Source: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

Business Case

Why does Chicago collect this data?

- Gain insight on travel patterns within the City. Apply understanding to:
    - Target improvements in public transit offerings
    - Improve infrastructure planning
    - Ensure that access to rideshare is provided equitably

Other Data Users:

- Drivers: Understanding usage patterns helps them effectively position themselves within the City and avoid circulating without a rider.
- Environmental agencies/citizen groups: Better capability to quantify air emissions and traffic impacts associated with ridesharing, demonstrate need for better transit alternatives.

Big Questions

- Can ride demand be predicted by census tract for a given day of week and time of day?
- Are there clusters of ride patterns that could be potentially better served by a targeted transit option, such as a shuttle bus?

Process

- Cleaning, Wrangling
- Data Enhancement
- EDA
- Statistical Tests
- Machine Learning
  - Demand Prediction
  - Clustering

Cleaning and Wrangling, Enhancement

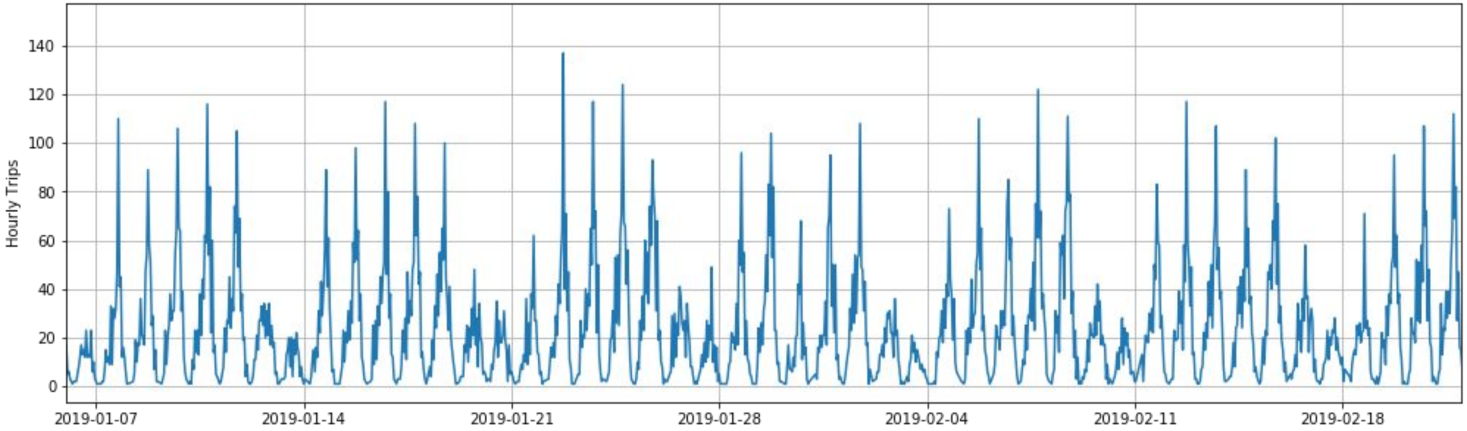
- Download and import
  - 12 GB file
  - Use 5% random subsample for evaluation
- Remove unusable records (zero fare, no census tract)
- Aggregation by hour
  - Datetime manipulation
  - Calculate ride counts, average trip distance, duration, fare by census tract, hour
- Integrate weather data
  - Hourly precipitation, temperature, wind speed from Midway AIrport (NOAA)
  - Merge on ride date time after synchronizing to nearest hour.

More Enhancement

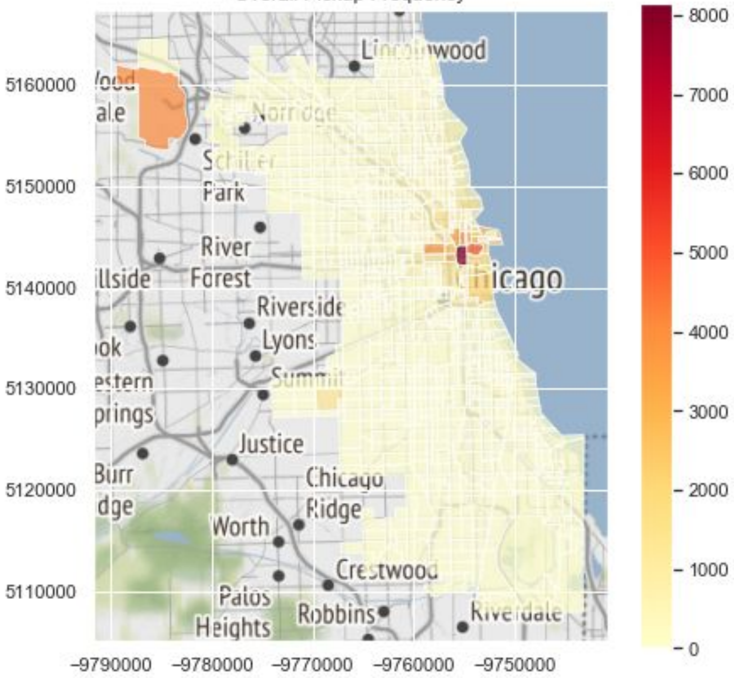
- Census geodata
  - Import census tract geojson to geopandas geodataframe object
  - Enables mapping of data by census tract and geospatial operations on the data (distance, area, etc.)
  - Join rideshare data (multiindex on census tract, time) to census geodataframe with pandas merge
- Census population data
  - Use Census API to acquire population and income data, merge on ride pickup census tract
- Derived fields (per tract):
  - Distance from downtown
  - Population density
  - Median Income

EDA Findings: Temporal Pattern

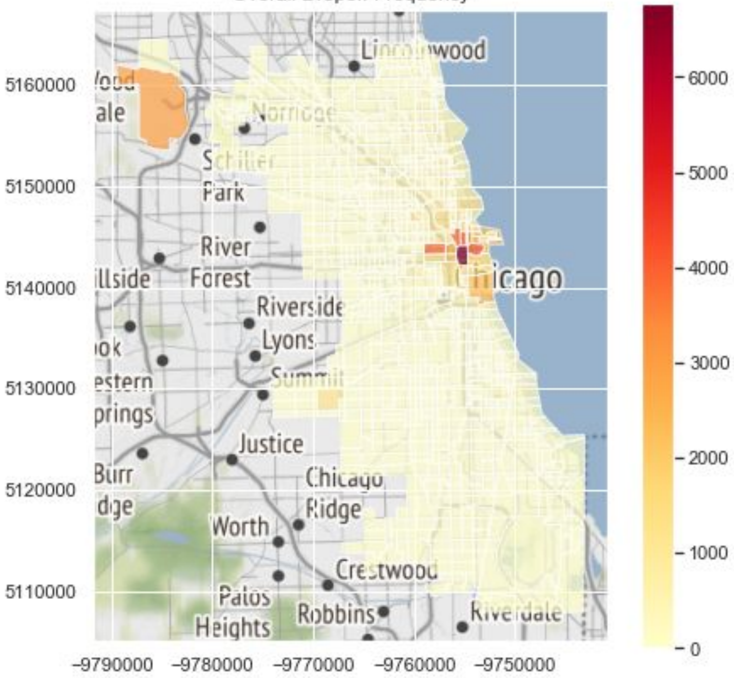
- Total system usage is cyclical
- Hourly pattern superimposed on weekly pattern
- Holiday effects



EDA Findings: Spatial Pattern

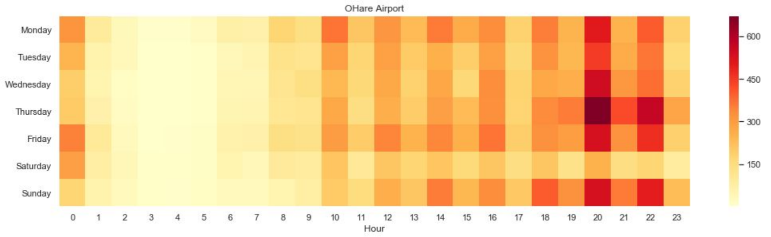


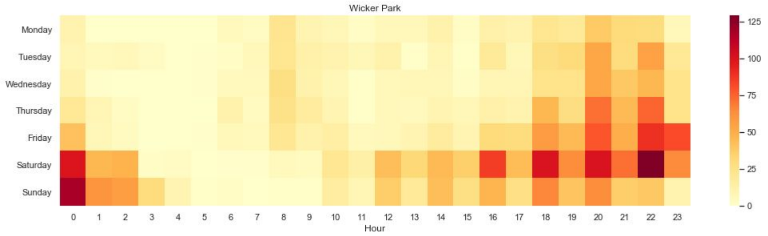
Overall Pickup Frequency

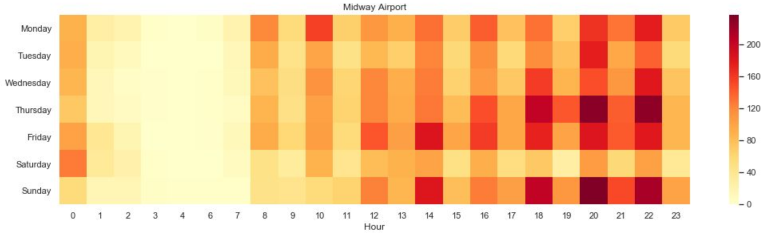


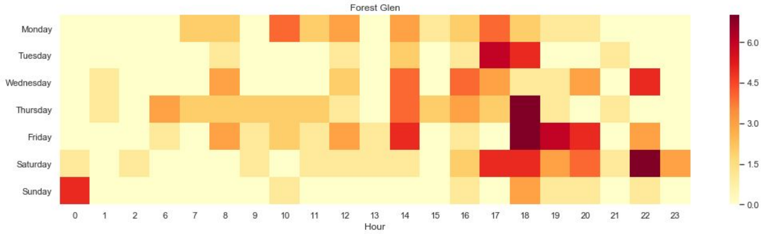
Overall Dropoff Frequency

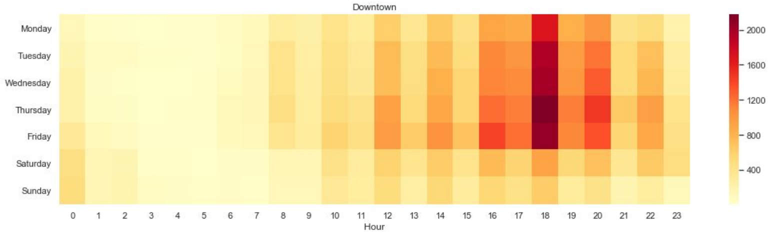
EDA Findings: Spatiotemporal Patterns

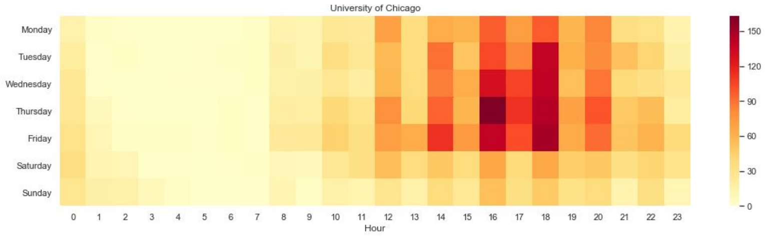






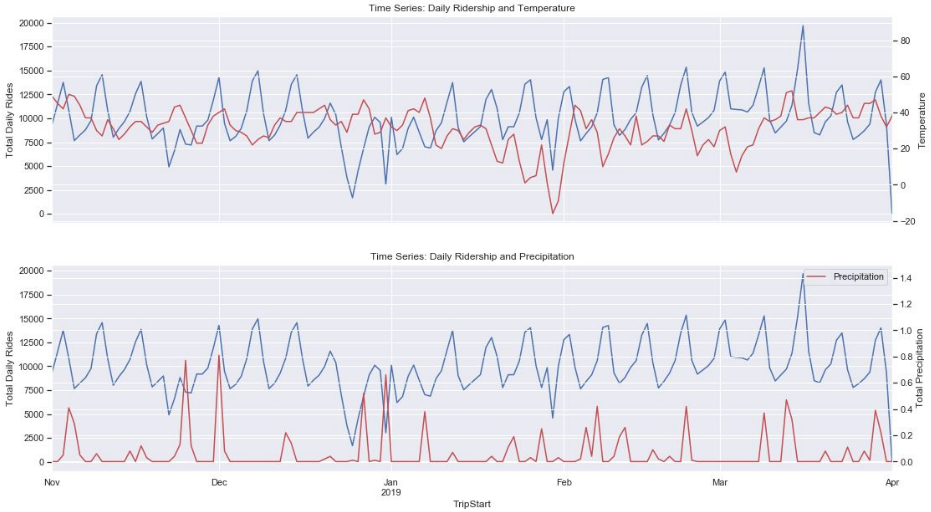


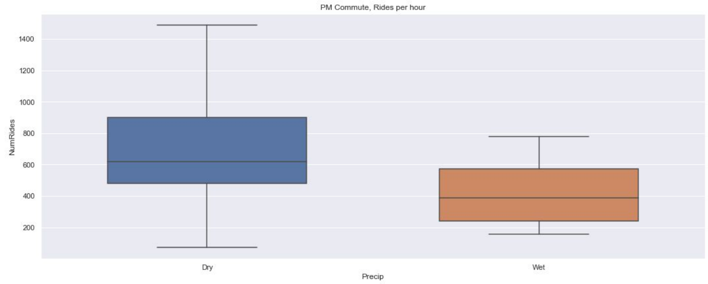


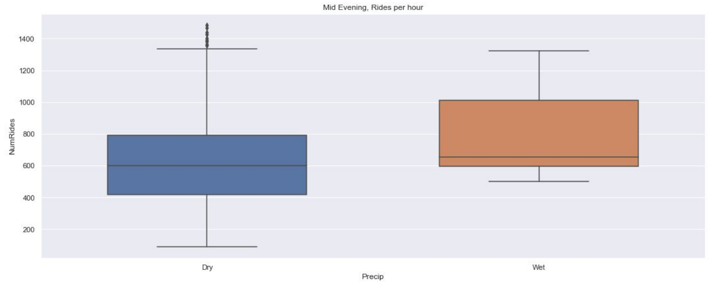


EDA Findings: Weather Effects?

- Not apparent at daily level
- Focused impact within specific time windows

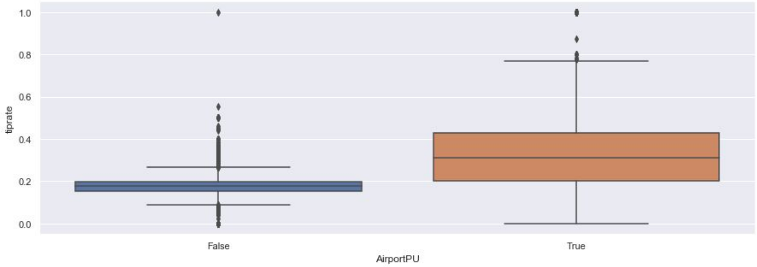


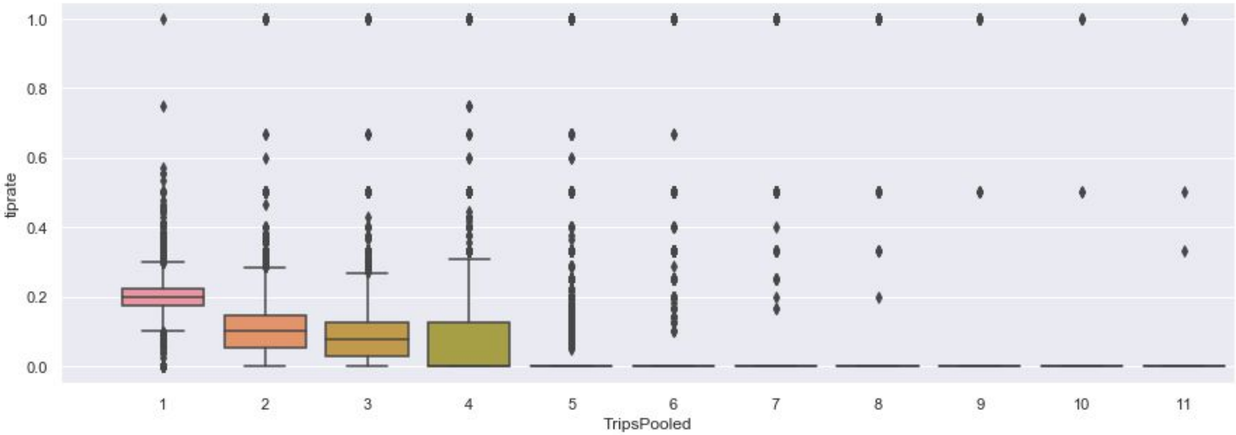




EDA Findings: Tipping

- Overall 20% tipping rate
- Statistically significant increase in probability of tipping for airport pickups.
- Decrease in tipping probability for pooled rides


tiprate
False
True
AirportPU


tiprate
TripsPooled