Analyzing Dockless eVehicles Trip Data for Predicting Ridership Demand in Louisville, KY

Author: Abhiram Muktineni

Commonwealth Of Kentucky

Content

1. Exploratory Data Analysis
2. Geo Spatial Data Analysis
3. Machine Learning Analysis
4. Model Comparisons
5. Recommendations

Exploratory Data Analysis (EDA)

- Info about dockless scooters imported from Louisville Open Data
- Dataset Size: 641224 rows x 13 columns (each row represents a trip)
- Timeframe: August 2018 to December 2020
- Python Libraries used in this project:
  - import pandas
  - import matplotlib
  - import seaborn
  - import scipy

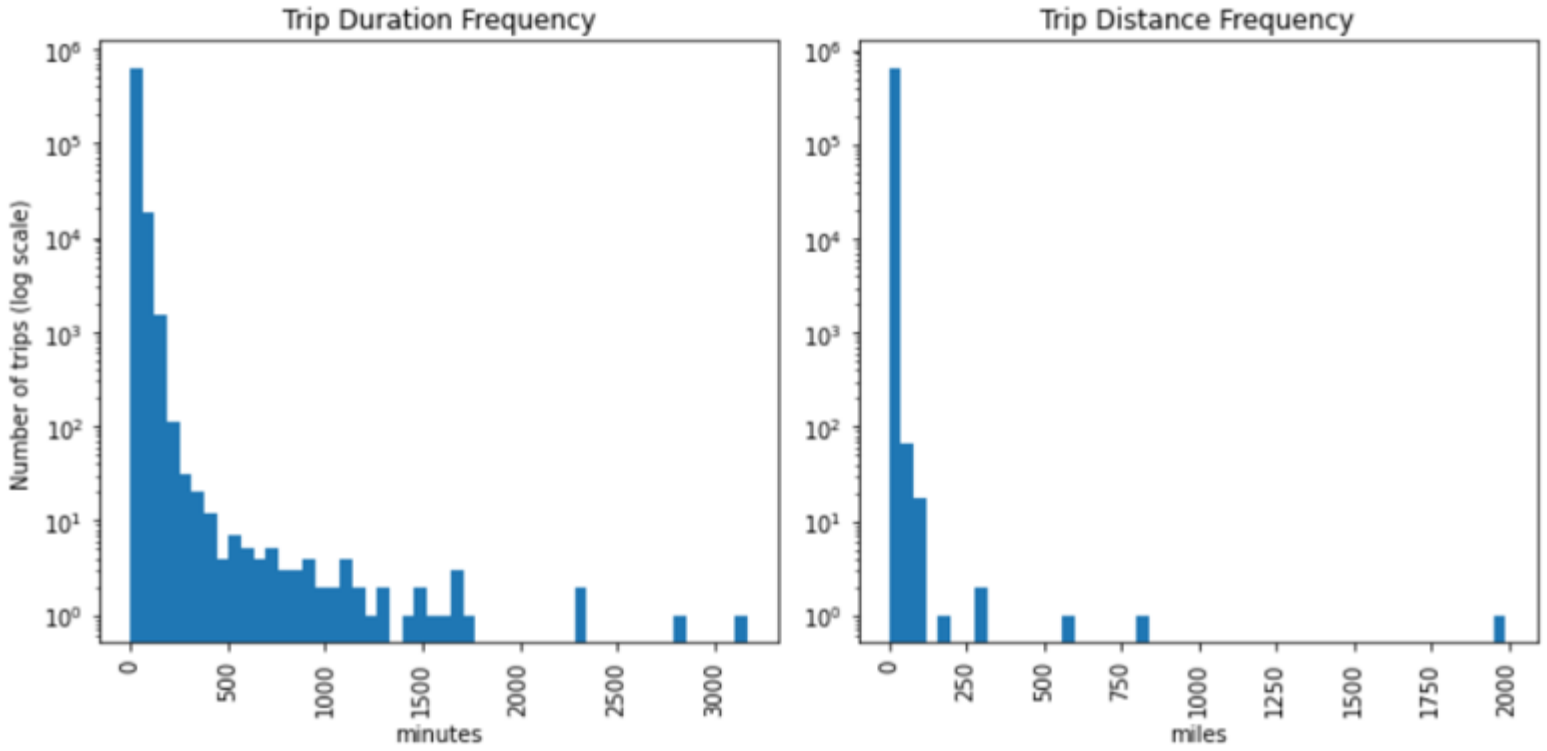
Data source: https://data.louisvilleky.gov/dataset/dockless-vehicles

Feature Variables

- **TripID** - a unique ID created by Louisville Metro
- **StartDate** - in YYYY-MM-DD format
- **StartTime** - rounded to the nearest 15 minutes in HH:MM format
- **EndDate** - in YYYY-MM-DD format
- **EndTime** - rounded to the nearest 15 minutes in HH:MM format
- **TripDuration** - duration of the trip minutes
- **TripDistance** - distance of trip in miles based on company route data
- **StartLatitude** - rounded to nearest 3 decimal places
- **StartLongitude** - rounded to nearest 3 decimal places
- **EndLatitude** - rounded to nearest 3 decimal places
- **EndLongitude** - rounded to nearest 3 decimal places
- **DayOfWeek** - 1-7 based on date, 1 = Sunday through 7 = Saturday, useful for analysis
- **HourNum** - the hour part of the time from 0-24 of the StartTime, useful for analysis

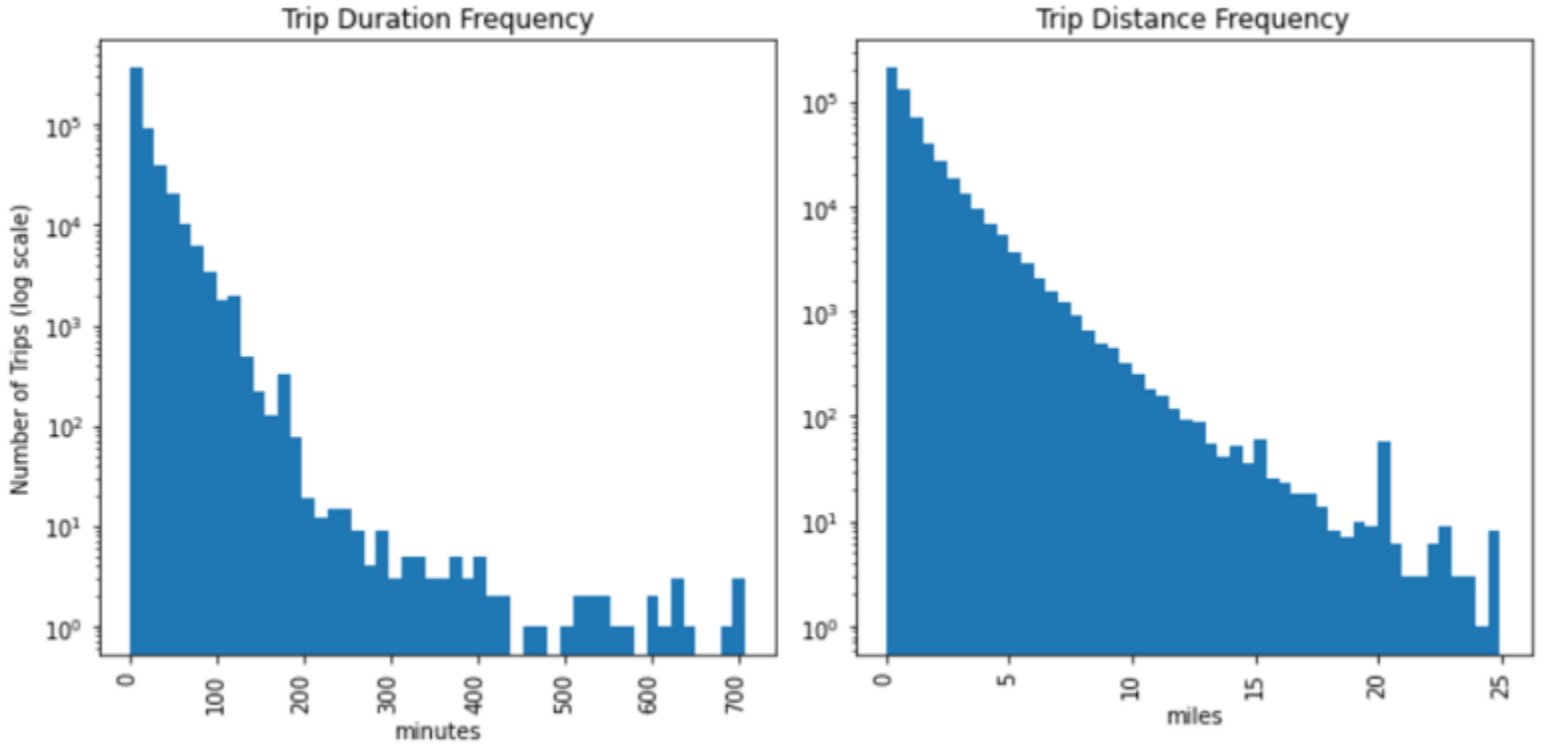
Trip Duration & Distance

(before removing Outliers)

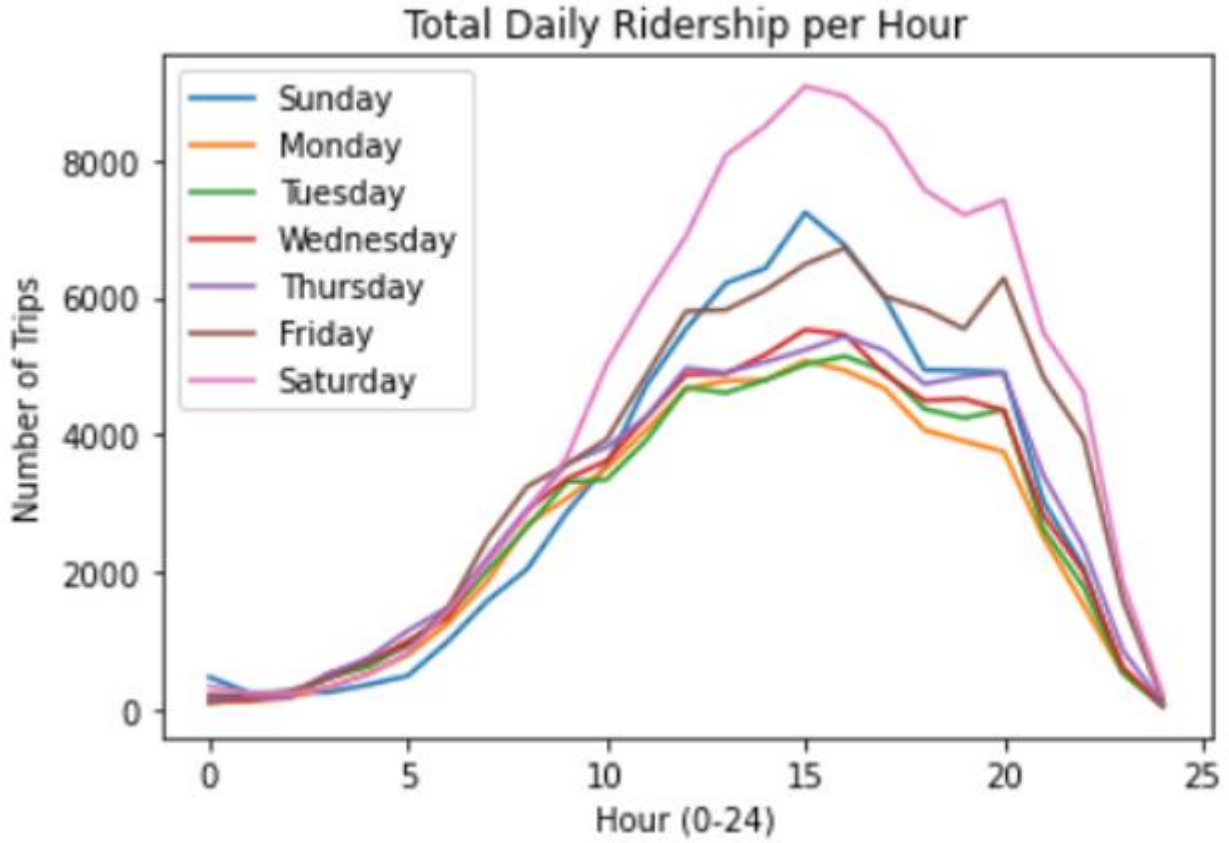


Trip Duration & Distance

(after removing Outliers)

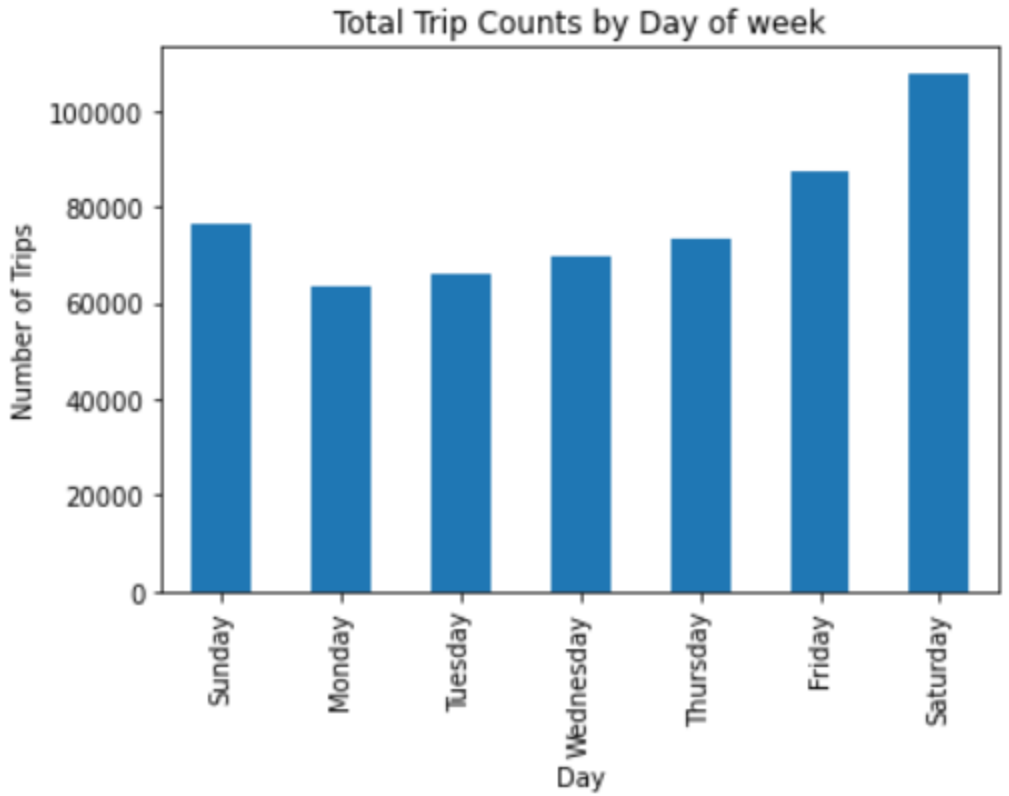


Usage by Time



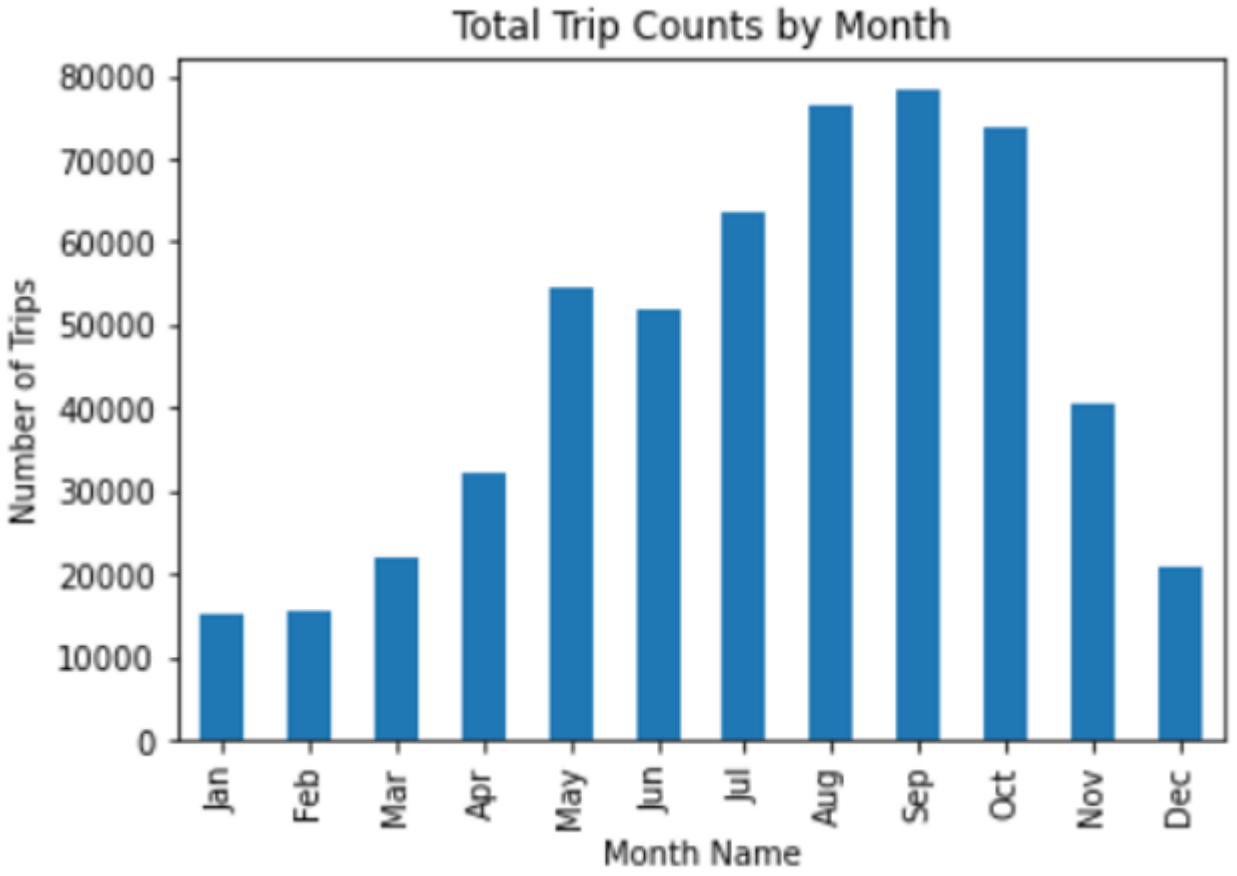
The number of rides in a certain time period will vary depending on time of day and day of the week. In the above plot, each day of the week has a different sized curve, but each day's usage peaks in the mid afternoon.

Usage by Day of Week



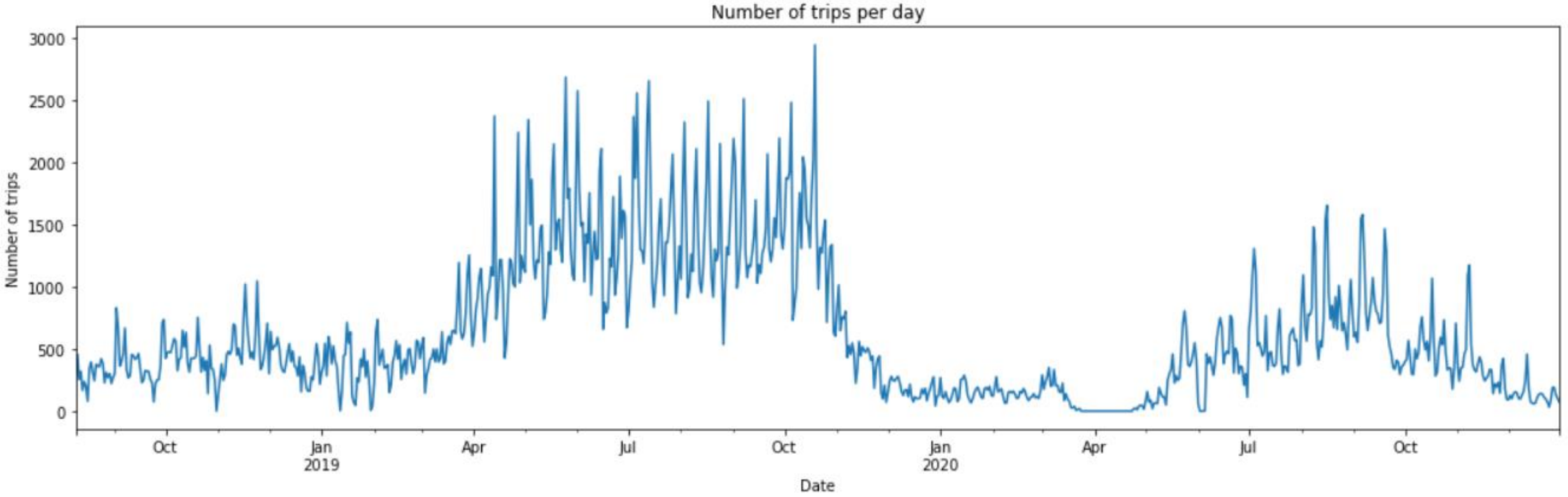
General ridership bottomed out during the week, and increases into the weekend.

Usage by Month



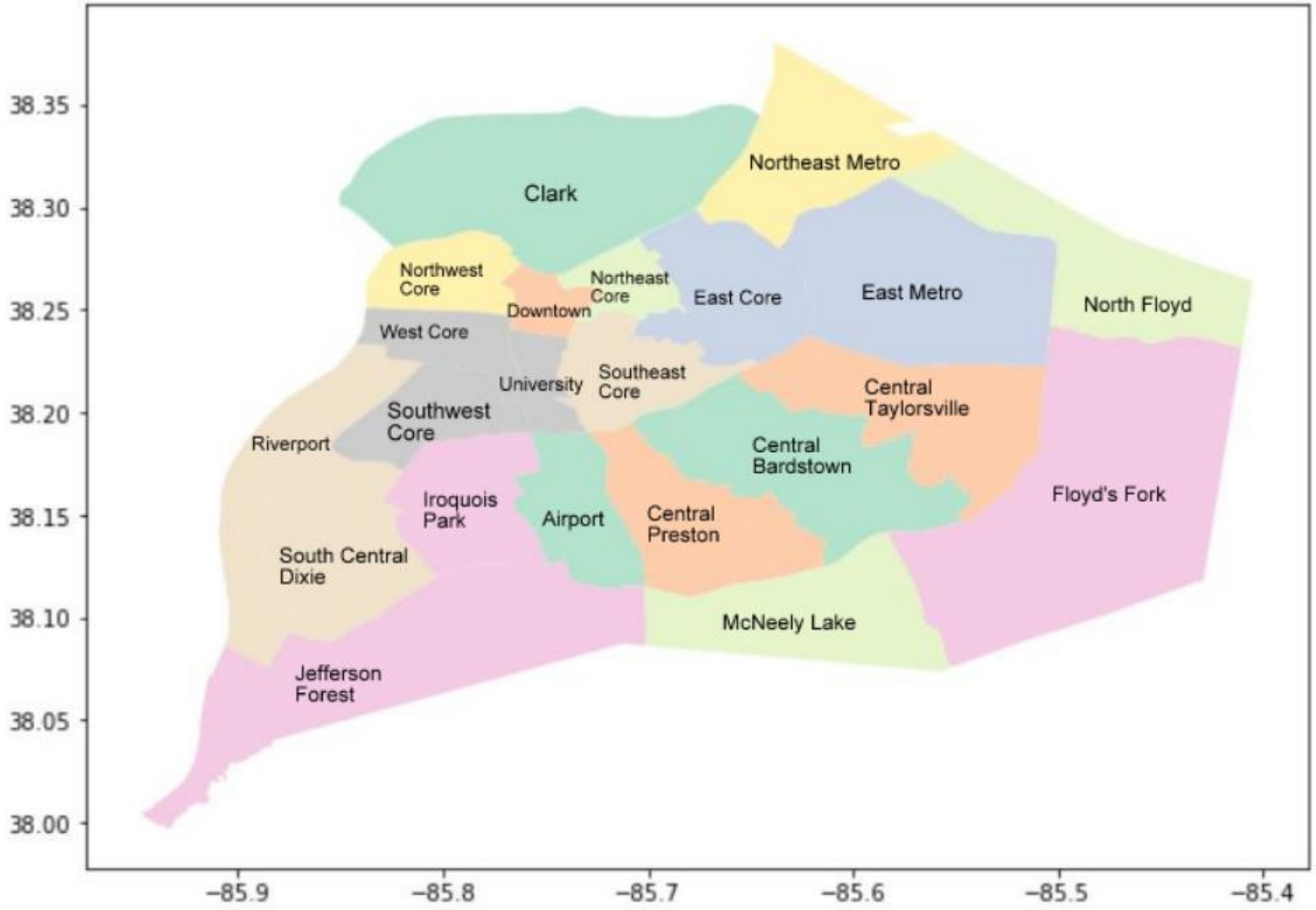
In the above plot, there appears to be more ridership in the warm summer months through mid fall, with peak usage in July through October.

Overtime Usage



In mid-October of 2019, there is a peak that coincides with Louisville's HalloScream, Jack-O-Lantern and Haunted Park after Dark geocatching Halloween events. The regularly spaced spikes in the data appear to fall on weekends. The graph also shows a drastic decrease in March/April 2020 as scooters lost popularity across the US due to increasing Covid cases during the early pandemic days.

Geo-Spatial Data Analysis

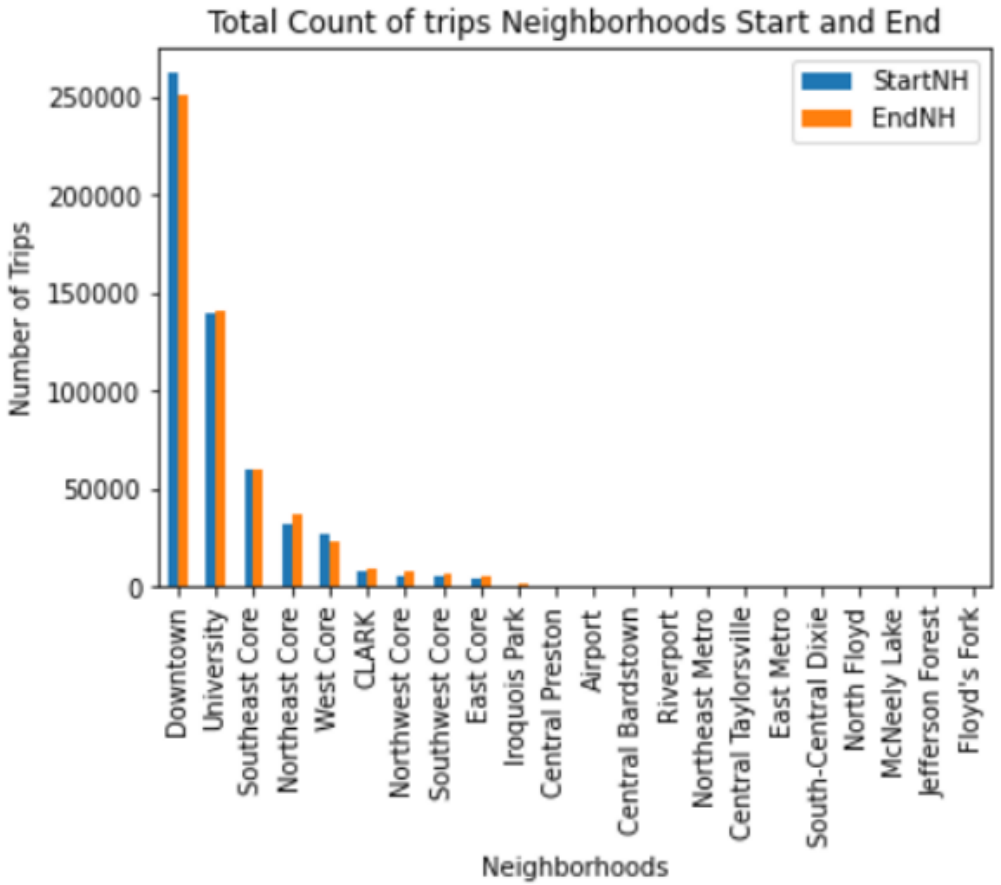


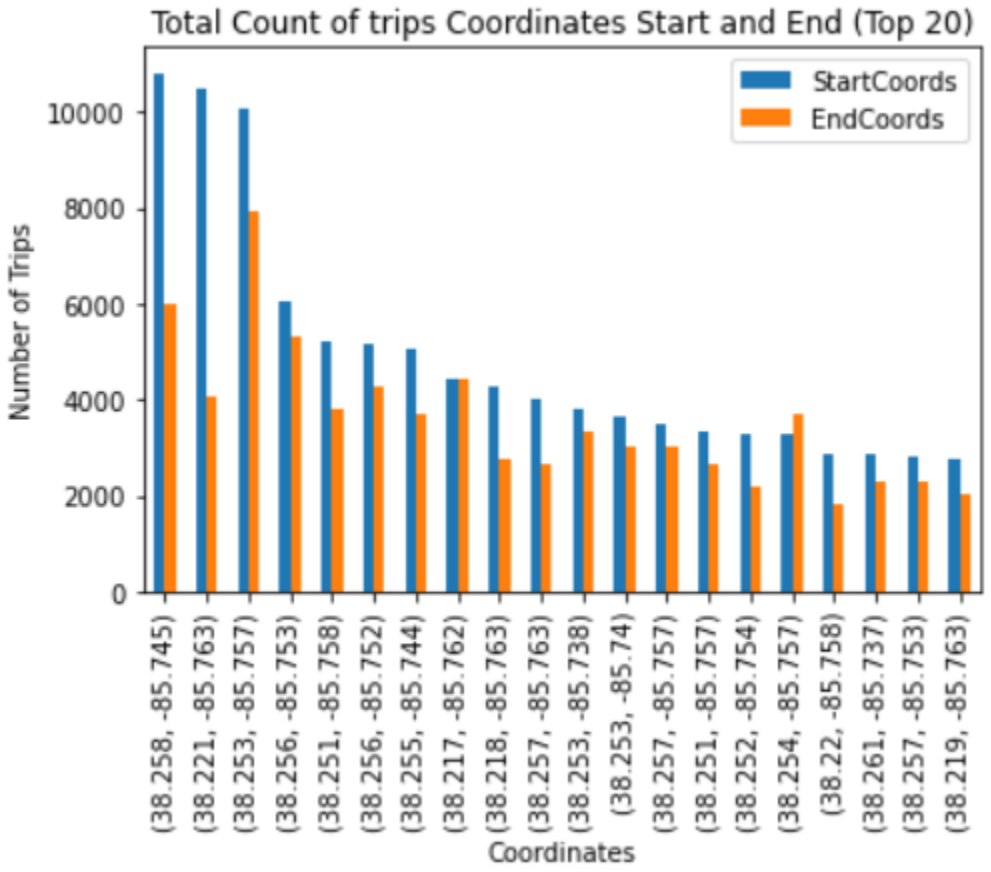
Geo-Spatial Data Analysis

- What is Geospatial data?
  - Data with location information such as Latitudes and Longitudes.

- Importing geospatial data with GeoPandas
  - import geopandas

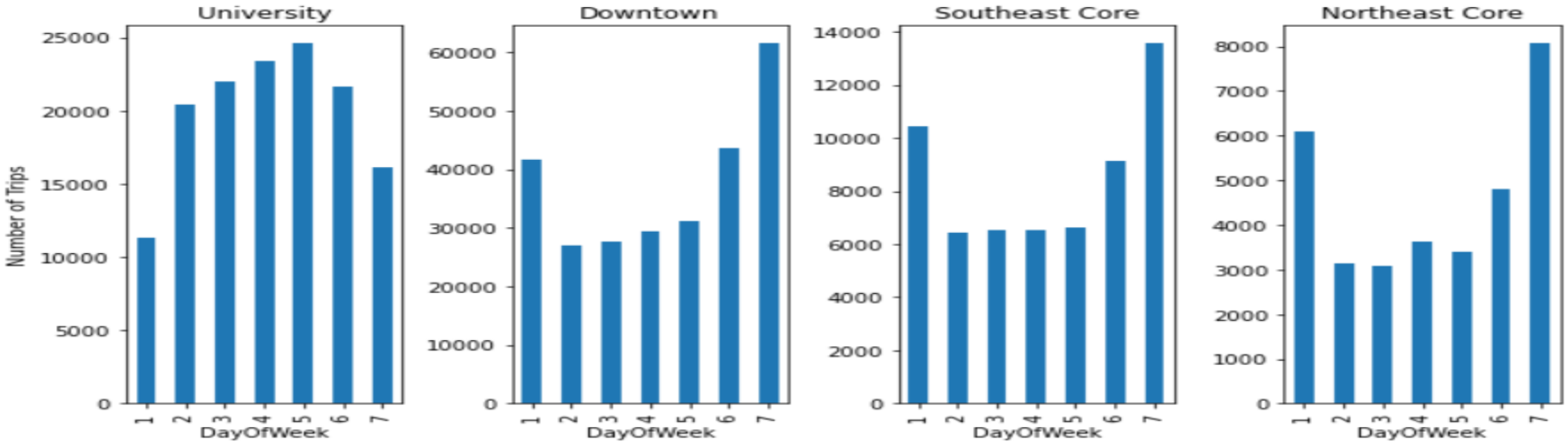
Usage by Location



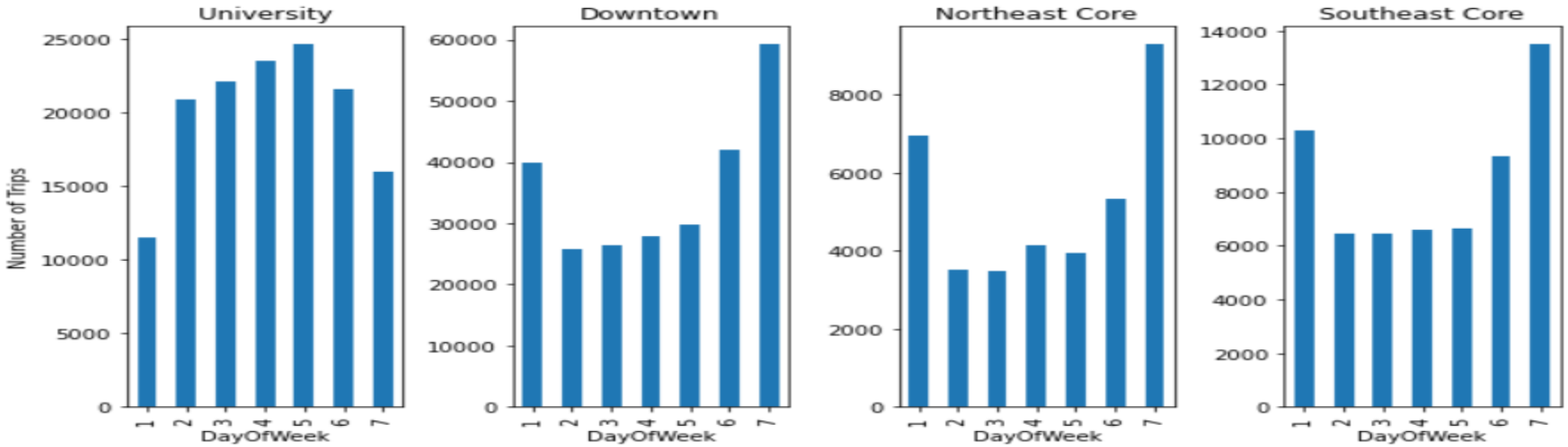


Top 4 Neighborhoods

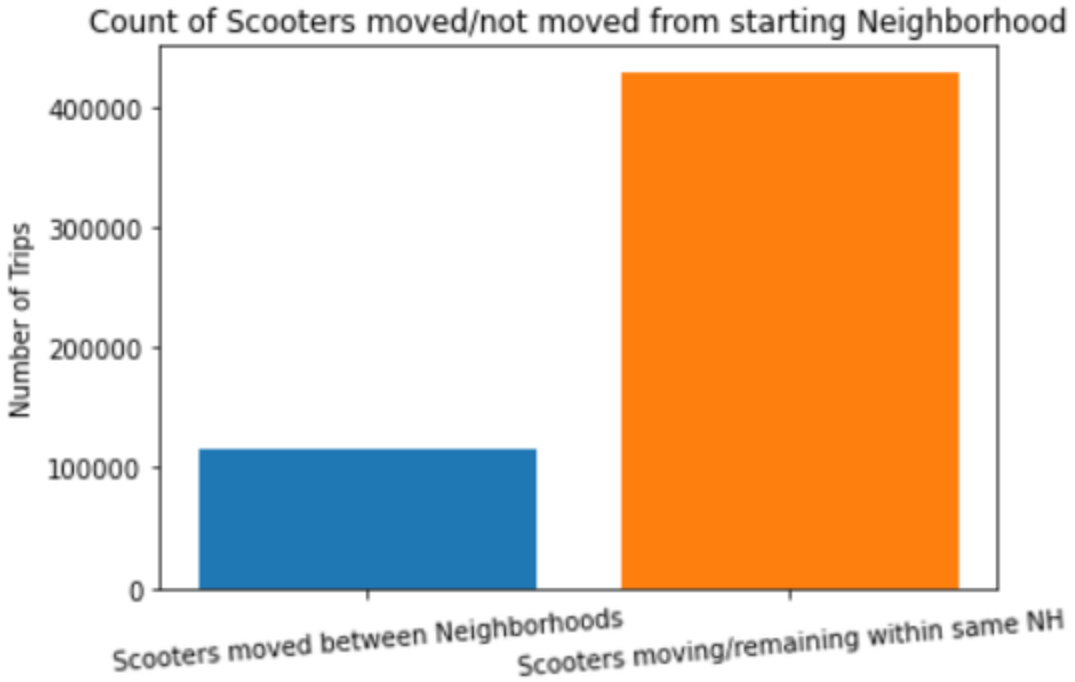
Top 4 starting location counts by day:



Top 4 ending location counts by day:

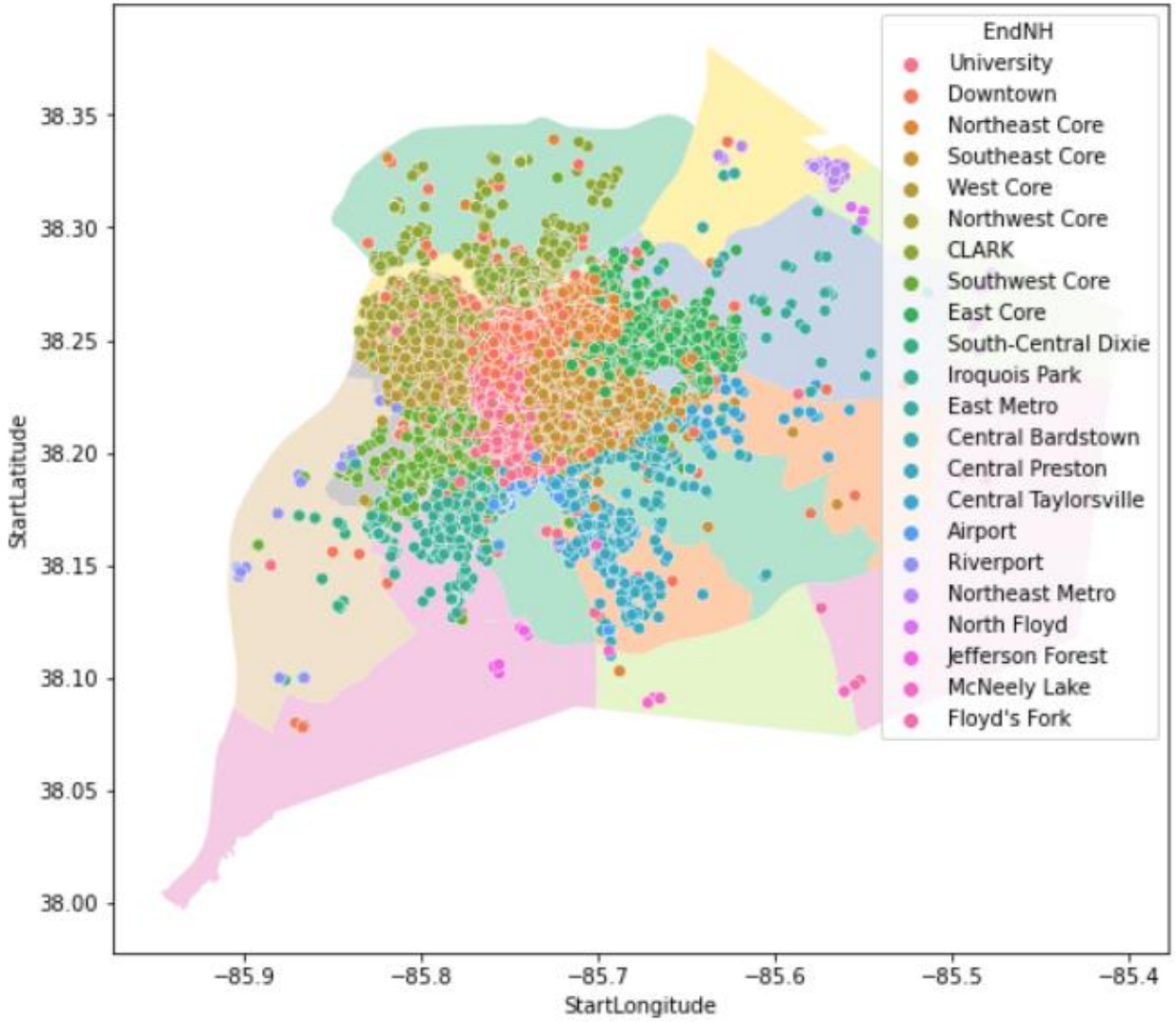


Louisville Urban Neighborhoods



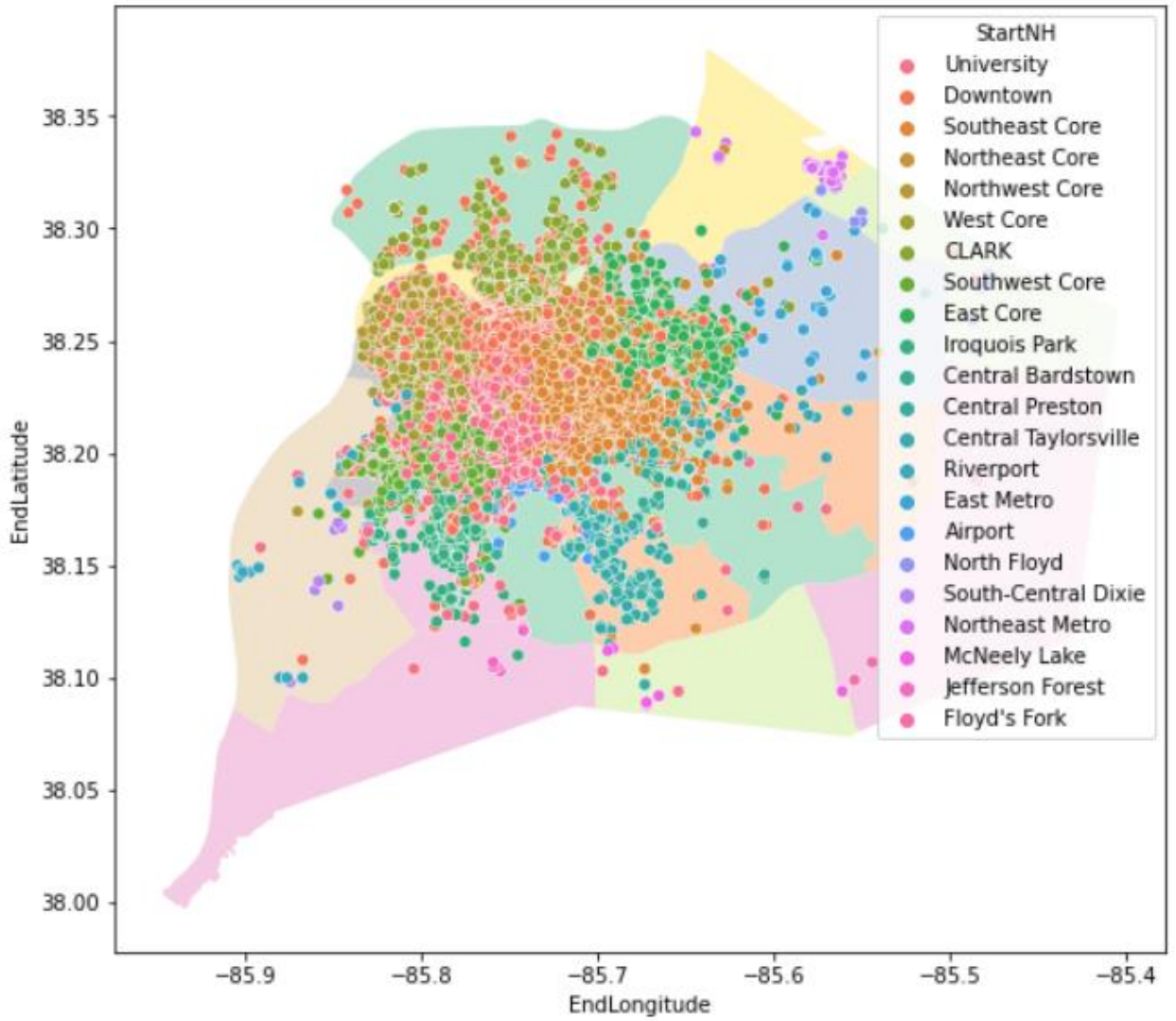
Louisville Urban Neighborhoods

(Plot containing Starting Locations of trips)



Louisville Urban Neighborhoods

(Plot containing Ending locations of trips)



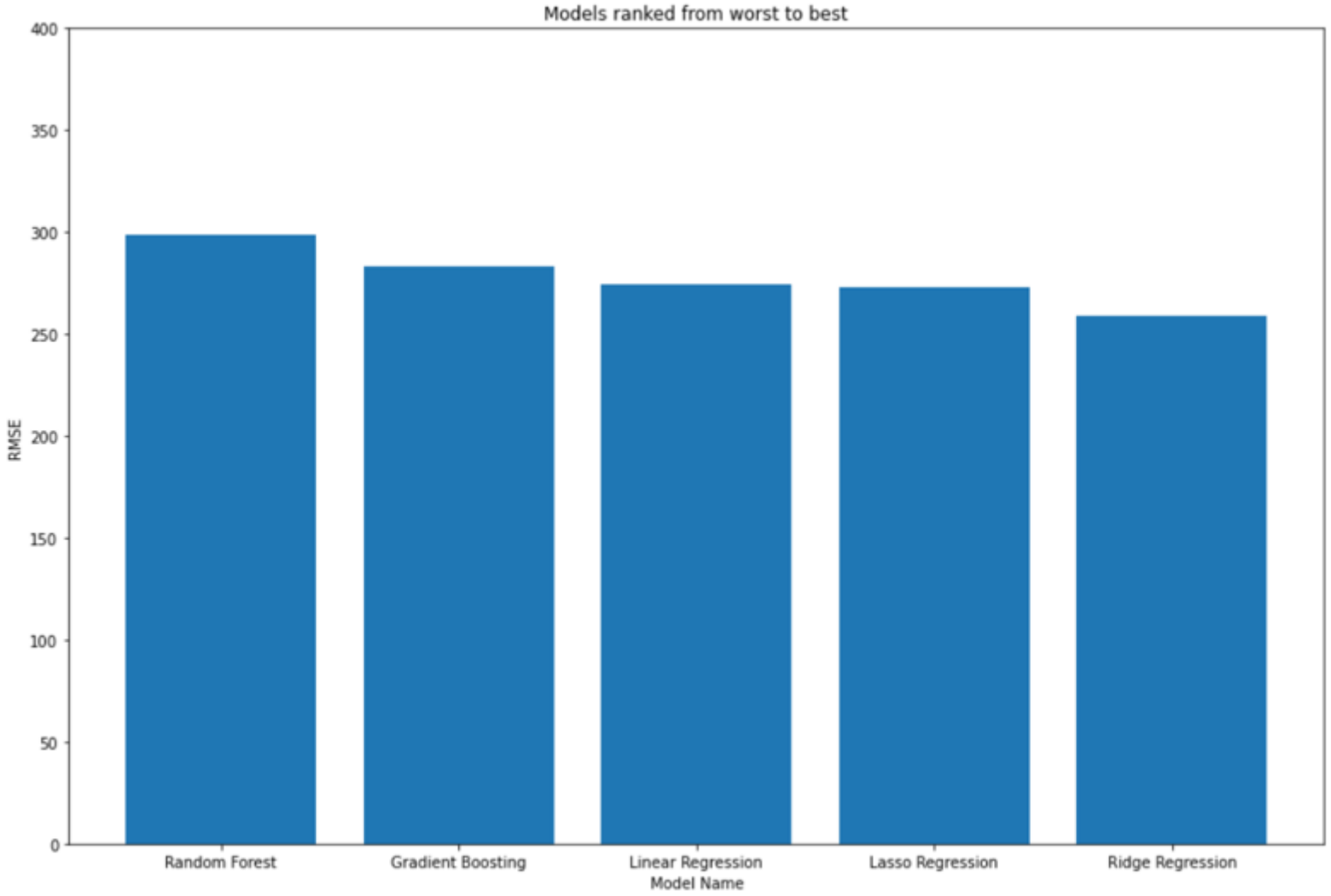
Machine Learning Analysis

- Following Machine Learning methods have been applied:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
  - Random Forest Regression (Ensemble Method)
  - Gradiant Boosting Regression (Ensemble Method)

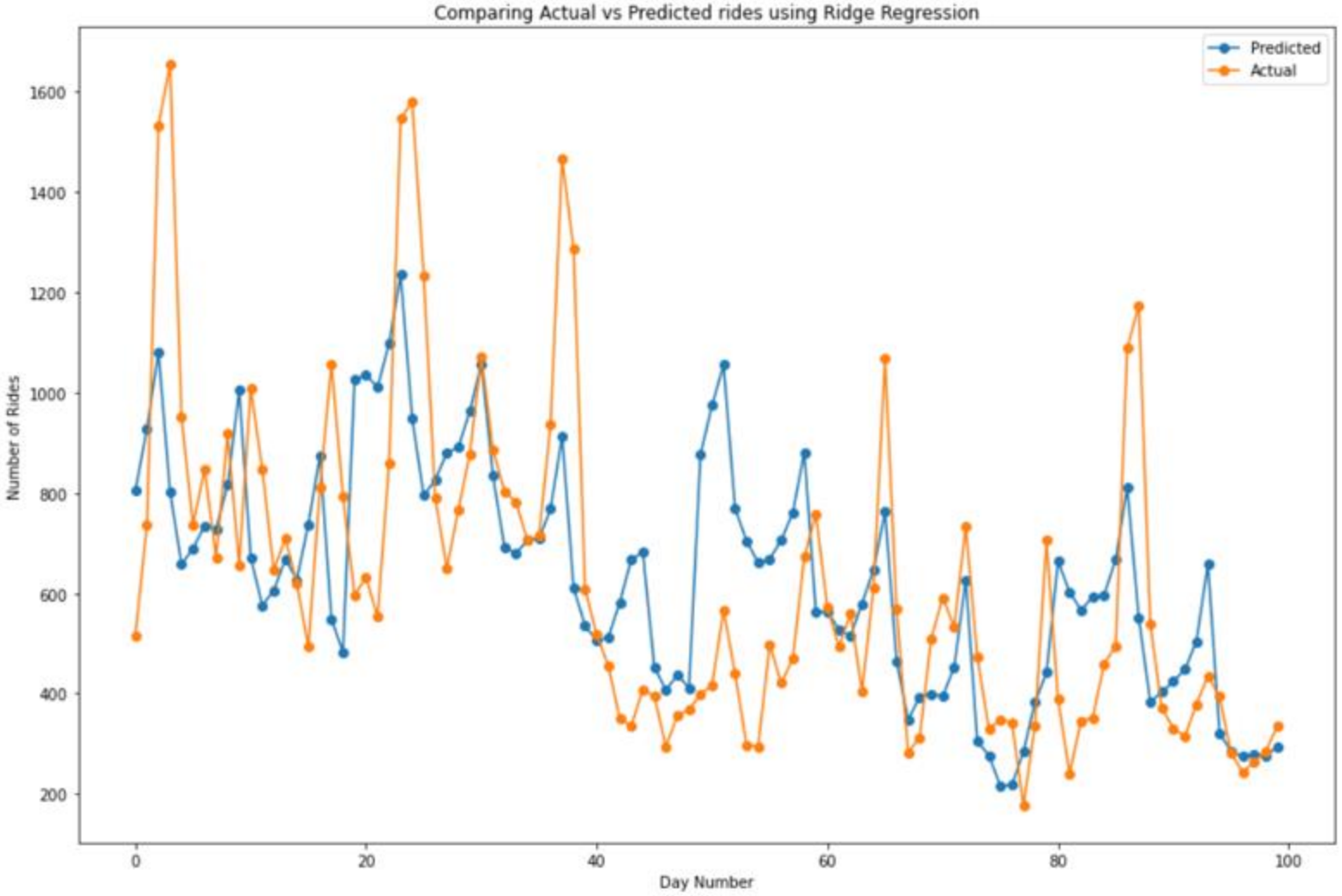
Model Scores

Model	MAE	MSE	RMSE
Linear Regression	216.74	75410.23	274.60
Ridge Regression	198.76	66867.16	**258.58**
Lasso Regression	213.94	74443.99	272.84
Random Forest	212.08	89163.24	298.60
Gradient Boost	201.15	78962.08	281

Model Comparison



Best Model - Ridge Regression



The ridge regression model gave a **root mean square error of 258.58** when ran against the test set making it the **best performing model**

Recommendations

It turns out that **Ridge Regression** model has been the **best performing model** over other models. Hence, we recommend Louisville Department of Public Works to deploy the Ridge Regression model in order to forecast the total daily ridership demand of electric find and ride vehicles in Louisville.

Thank You – Keep Riding