# Analysis of E-Scooters in Washington DC

Team: Enchen Fu, Guanzong Zi, Haoyang Li, Taren Daniels, Ying Hua, Yining Fan
Advisor: Professor Y. Samuel Wang
Cornell Bowers School of Computing and Information Science
Client: Dr. Pramita Bagchi, George Mason University

## Introduction

### *Project Overview*

E-scooter popularity has risen as a source of local transportation in Washington, DC – how we can use the data to understand transportation patterns?

Our data set Includes: Location (longitude and latitude), Battery Level and other information about e-scooters.

Goals:

Understand basic spatial patterns of e-scooters;

Determine and compare dynamics (i.e. hotspots, low use regions, and high drop-off regions) daily and weekly;

Use exploratory data analysis techniques to understand and interpret e-scooter data.

### *Data Description*

After cleaning the data, we have a total of 48 csv files & 5 variables:

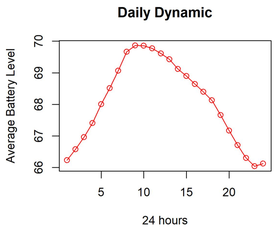
-lat(latitude) & lon(longitude): location of each E-Scooter

-battery_level: ranges from 0 –100 (in percentage)

-is_disabled: Weather a scooter's battery is lower than 20%

-time: the moment that the scooter sent data to the server

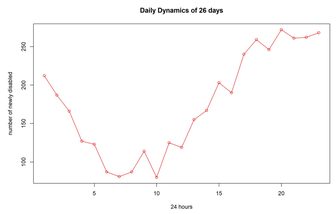
## Data Analysis




Change in average battery level can represent usage of e-scooters. On the left is daily dynamics of average battery level of all e-scooters.

### *Newly-added Disabled E-scooters*

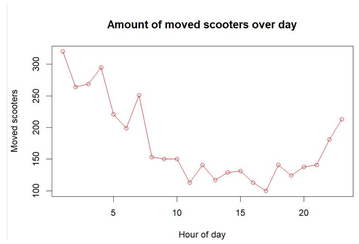
Increase in disabled e-scooters amount indicates high usage periods of a day. Below is the daily dynamics of newly added disabled e-scooters.




### *Removed Disabled E-scooters*

We assume that when an e-scooter is disabled, the company would pick them up for recharge, and then they are removed from the data set. Analyzing their amount over the day helps us explain the battery level dynamics.

Here is the amount of removed disabled e-scooters over the 26 days that we have full data.




As this result suggests, most e-scooters were taken away during later hours of the nights and very early mornings. It might be because the company didn't want to bother the users and chose those hours to do the job.

### *Newly Added Fully Charged E-scooters*

We assume that the company is aware of the e-scooters' usage situation and they would put fully charged e-scooters back to high usage regions. Here is a map that records the location of newly added e-scooters.

Since the company put e-scooters at same locations, each point on this map is a stack of many points. We can use this graph later to verify our clustering results.

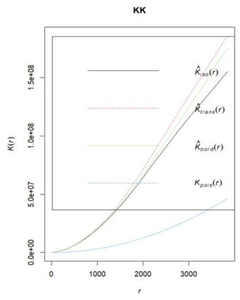
## Methods

### *Ripley's K Function*

The function can determine whether features' distribution exhibits statistically significant clustering or dispersion over a range of distances. Formula:

$$K(r) = \frac{1}{n\lambda}\sum_{i=1}^{n} N_{p_i}(r)$$

Theoretically, if the observed value of K-Function is greater than some expected K for a specific r, the distribution should be considered more clustered.




The figure on the left is the plot of applying 'Kest' function in 'spatstat' package on the data at 9am, August 14th.

All the 3 estimate-curves are above the theoretical curve. Thus the distribution pattern of the e-scooters is clustered, making it feasible to apply some clustering method!

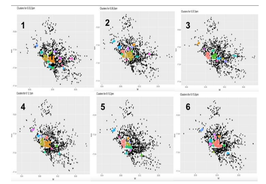
### *DBSCAN*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm used in machine learning to identify clusters of data points in a given dataset. DBSCAN works by grouping data points that are close together based on their distance and density. It defines clusters as areas with high density and identifies outliers as points that do not belong to any cluster. Unlike some other clustering algorithms, DBSCAN does not require the user to specify the number of clusters beforehand. Instead, it dynamically determines the number of clusters based on the data.

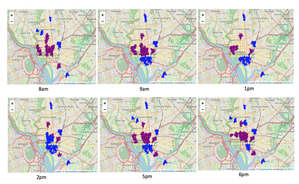
Its parameters are ε and *MinPts*. A larger value of *MinPts*.will lead to a larger number of clusters. A larger value of ε will yield clusters with larger size, as clusters will merge together, Vice versa.

## Results

### *Spatial Point Pattern Analysis Results:*




*(cluster plots from 6 days for different hours)*




*(distributions for August 3rd)*

### *General Spatial Patterns*

Approximately 10 clusters that are mainly distributed in the middle of city.

**High drop-off regions:** More likely in the downtown area, including the President Park (especially east side), Union Station, Foggy Bottom

**Low-use regions:** More likely at the edge of the city

Typical Clustering Areas Include the triangle area located at the north of the Washington Channel and south of Independence Ave

**Uncertain (sometimes high drop-off, sometimes low-use)**: Logan Circle, China town, The Senate Park

### *Daily Dynamics*

In the very beginning of the day, the size and number of low-use regions are larger than the high drop-off regions for most days.

In the evenings around 6pm, the number of high drop-off regions is greater than low use regions for almost all days.

### *Weekly Dynamics*

During the morning peak hours, there are more high drop-off regions compared to low use regions for weekdays, while patterns are not seen during the weekends.

During Weekdays:

-Greater number of high drop-off regions in the mornings

-Decreasing number of high drop-off regions in the afternoons

-Increasing number of high drop-off regions in the evenings

During Weekends:

Increasing number of high drop-off regions from morning to evening

### *Further Analysis and Regression*

**Goal:** Analyze other extraneous variables that may have a significant impact on e-scooter battery levels. Metrics for weather include: hourly temperature, indicator variable for rain, windspeed, and total precipitation levels.

Exploratory Data Analysis Results shows that there are substantial instances of rain in the dataset to parse out the effects, and the variation in average hourly temperature appears to follow a defined a pattern.

```
Subset selection object
Call: regsubsets.formula(battery_level ~ precip + temp + windspeed +
    rain, data = subset_all, nvmax = 3)
4 Variables  (and intercept)
          Forced in Forced out
precip        FALSE      FALSE
temp          FALSE      FALSE
windspeed     FALSE      FALSE
rain1         FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         precip temp windspeed rain1
1  ( 1 ) " "    "*"  " "       " "
2  ( 1 ) " "    "*"  " "       "*"
3  ( 1 ) " "    "*"  "*"       "*"
```

Lowest Test MSE

$$\widehat{BatttryLevel} = 66.07 + .03(Temperature) + .42(Rain) + e_i$$

We trained a model to find relationship between weather and battery level. We can conclude that temperature and raining circumstances are closely related to battery use.

## Conclusion

DBSCAN method identifies hotspots, spatial features like high drop-off regions and low-usage regions, as well as daily and weekly dynamics. However, it is not recommended for use with unbalanced data as this may have negative impact on the clustering and complicate analyses across different time periods.

Also, from the model we built, we can conclude that temperature and the presence of rain has a significant effect on battery levels.