Analysis of impact of socio-economic indicators on Micro Transit Occupancy

Samir Gupta
Vanderbilt University

Abstract**—** To analyze e-scooter micro transit data along with socio-economic indicator big data using big data technology such as Spark, Cloud technology such as Amazon Athena and EMR clusters. The insights drawn from this analysis can be used to drive further research into optimization and e-scooter placement across the city. Further this paper is also used to show case how Spark can be used for processing of big data on machines with low memory (RAM).

Keywords— Big Data, Spark, E-scooter usage, socio-economic indicators.

I. INTRODUCTION

Due to their affordability and convenience for short distances, e-scooters have grown in popularity as a form of mobility in cities [15]. In addition, during the COVID-19 pandemic, the use of shared e-scooters as an alternative to public transportation for short-distance travel is increasing [16]. The usage of e-scooters is rapidly increasing, and people are using them not only as an alternative to but coherently with public transit systems such as buses and metros for their transportation.

This gives us the motivation for optimization and help guide research in this field. One of the biggest problems that is faced with e-scooters is how many scooters should be present in an area and where to place them i.e., the placement of e-scooters [15]. Another problem that requires attention is the charging of these scooters. This paper can help guide further research in both areas. This paper’s main objective is to provide visualizations that can help understand the correlation between socio-economic factors and scooter usage in the city of Nashville, more specifically Davidson County.

Another major problem that this paper talks about is handling of Big Data on machines that do not have enough memory to store data in its memory. This paper makes use of Spark created at UC Berkeley in 2009, which is a very powerful tool that makes use of distributed computing framework for processing of Big Data [1]. In this paper we can successfully process big data of size 3.5 GB on a single machine with 4 cores in minutes.

Section 2 of this paper discusses related work in this field. The next section describes the problem statement. The further sections describe the methodology, results obtained and Future work for this paper.

II. RELATED WORK

The popularity of e-scooter use in recent years has led academics to investigate a few variables impacting e-scooter occupancy. In this part, we go over the pertinent research on e-scooter occupancy analysis.

In Indianapolis during September to November 2018, Liu et al.'s [18] analysis of shared e-scooter trip data examined travel patterns. 60% and 65% of the total journeys were under 10 minutes and less than one mile, respectively. Like this, Mooney et al. (2019) looked at the spatial equity of dockless bikes in Seattle, Washington and discovered that a higher density of dockless bikes related to a higher education level among the locals.

Karim et al. (2020) [17] investigate the potential usage of machine learning methods for prediction of e-scooter occupancy. The authors of these papers make use of e-scooter data and weather patterns to predict the occupancy of e-scooters. They were able to predict the occupancy with an accuracy of 80%.

Ricardo et al [19] have shown a data driven method that makes use of clustering algorithms for establishing shared electric scooter (SES) parking locations and assessing their anticipated utilization. Their findings indicate that there was a 13% decrease in overall trip collection and 300% hazardous narrow sidewalk trips were recorded.

III. PROBLEM STATEMENT

The goal of this paper is to analyze big data primarily Micro transit e-scooter data and gain insights from it. The hypothesis for this paper is that socio-economic factors have a direct proportionality to the e-scooter usage in Davidson County. The motivation behind this paper is that there are socio economic factors that govern the usage of scooters area wise. Socio economic factors can include median household income, and other indicators of income as well. The data for socio economic factors is readily available online using census website for United states of America. Further, the goal of this paper is also to demonstrate the potential ability of big data technologies such as Spark.

In this paper we process 3.5 Giga bytes of trip data using Spark on a machine with 16 GB RAM and Intel i7 – 11th Gen (4 Core processor). In a real-world situation not all the 16 GB RAM is available for processing and loading data into RAM (which was the case in this paper).

We had initially attempted to load all the data into the RAM using Pandas DataFrame, however that was unsuccessful as it led to out of memory errors and system crashes. That is where Apache Spark came to the rescue, we were able to successfully partition the data to process on my 4-core machine and process and perform various transformations on the data. We made use of various powerful tools in Spark such as User Defined Functions (UDFs) and Structured Query Language (SQL) tools available inside of Spark.

In this paper, we also replicated my code using cloud technologies on Amazon AWS cloud service, namely, Amazon Athena, EMR cluster and S3 buckets. We discuss in the later sections the issues faced by me when using these technologies and use cases of these technologies in the later sections.

From this paper we were successfully able to analyze big data and create various visual plots that can help guide further research and optimization of the placement and utilization of e-scooters in the city of Nashville. Effectively the knowledge and insights gained from the visualizations of this paper can further utilized in various other cities and it can be identified whether a similar trend is happening. This paper can also help guide optimizing the placement of charging stations by placing charging stations in areas for riders to end their trips in those areas.

IV. Technology Used

To complete this project, we made use of Python as my coding language, the python version used was 3.9.13. We made use of python as it provides an easy coding interface to work with over Spark, Pandas and Geopandas, we will explain the advantages of these technologies later in this section. Further, python has useful libraries for creating visual plots using libraries such as plotly, matplotlib and seaborn. Below is an introduction of all the technologies used:

i) Spark

Apache Spark developed at UC Berkeley AMPLab in 2009 [1] is a powerful distributed computing framework [1]. This open-source technology makes it possible to quickly analyze and handle enormous volumes of data [2]. Spark is widely used in the industry today, by various companies for managing and handling big data, it is heavily used in sectors that deal with a lot of time series and big data such as Finance, Healthcare, and retail [3]. Through a language-integrated API akin to DryadLINQ in Scala, a statically typed functional programming language for the Java VM, Spark offers the RDD abstraction [1]. Spark has the ability of lazy computation, i.e., using two main concepts transformations and actions, it only runs the transformations on the data whenever an action is called on it. Until then spark goes on to build the RDD (Resilient Distributed Dataset) using DAGs (Directed Acyclic Graphs) [1]. There are various applications and benefits of using Spark such as Real-time data processing – Spark can be heavily used for processing of data from IoT device sensors [3]; Large Scale Data Analysis - Spark can handle data sets of several terabytes or petabytes, making it an excellent choice for big data applications[3]; Machine Learning – Spark has a machine learning library called MLlib that can be used to run various machine learning algorithms for classification, regression and clustering; Stream Processing – this is a very powerful tool that can be used for real time stream processing of data from various input sources such as PubSubClient, Kafka Messaging System, Databases and even IoT sensors [4].

Benefits of using Spark include [1]; High Performance - Spark is built to operate quickly and effectively, it is a great option for processing and analyzing massive amounts of data since it can carry out computations up to 100 times quicker than Hadoop; Fault tolerance: Spark is built to gracefully manage errors. Data processing and analysis can continue uninterrupted since it can recover from node failures or network partitions; Cost-effectiveness: Spark is a free to use open-source framework. Commodity hardware, which is considerably less expensive than proprietary gear, can be used to install it and Finally; Simple API: Spark's straightforward API makes it simple for programmers to create sophisticated data processing and analysis applications. Additionally, it supports a variety of programming languages, such as Python, Java, and Scala. For this paper Spark played a crucial role wherein it was used to read and process the 3.5 GB e-scooter dataset.

ii) Pandas

Pandas is a robust toolkit for Python programming that allows for data manipulation and analysis [6]. It offers data structures and operations for successfully managing and adjusting huge and intricate data collections [7]. Pandas can read data of any type that includes files of type JSON, parquet, pickle, text, etc. The main features of Pandas include; Data structures – includes two main types series and dataframe (looks similar to tables in SQL) [7]; Data Manipulation – pandas includes various functions for sorting, filtering, merging and reshaping, it also has the ability to make use of SQL query over the dataframes; Time series Analysis and integration with other libraries – pandas can be easily integrated with other libraries such as numpy and pyspark (can make udf's that can be used with big data).

iii) Geopandas

GeoPandas adds support for geographical data to the well-known data science library pandas [8]. It provides pandas with the ability to read geometry data that is present in geospatial data. This data can be viewed inside a python notebook easily using this library.

iv) Plotly

An interactive, open-source plotting toolkit for Python, plotly provides over 40 different chart types for a variety of statistical, financial, geographic, scientific, and three-dimensional use-cases [9]. It is built on top of the JavaScript library (plotly.js) and can be used to create various plots and save them as a HTML useful in web applications, PNG, or various other formats. Most of the plots for this paper were created using plotly, the integration of plotly with mapbox (explained later) was key for creating visualizations in this paper.

v) Mapbox

The robust data visualization tool Plotly Mapbox enables users to build dynamic maps with numerous layers of data [10]. It is based on the Mapbox mapping platform and offers several tools for designing beautiful and useful maps [11].

vi) Cloud Technologies Used

In this paper, we made use of Amazon Web Services cloud service available using AWS Learner lab that provides 100$ free credit for students in the Topics in Big Data class for the Spring 2023 semester. Amazon AWS is a very powerful cloud service offering various technologies for dealing with Big Data namely EC2 instances, wherein one can create a virtual machine with say 100 GB RAM and can be used to read and process Big Data. However, for this paper we made use of three very useful technologies offered by AWS:

A) Athena: Using conventional SQL, it is simple to evaluate data directly in Amazon Simple Storage tool using Amazon Athena, an interactive query tool [12]. Athena can be used to use your data stored in Amazon S3 and start using conventional SQL to conduct ad-hoc queries and obtain answers in seconds with a few clicks in the AWS Management Console [12]. This is a very powerful tool that provides query results on Big Data of the size of petabytes in time ranging between seconds and minutes. This tool was used in this project to check the potential of this tool on big data of size 3GB.

B) EMR Cluster:

The cluster is the main part of Amazon EMR. Amazon Elastic Compute Cloud (Amazon EC2) instances are grouped together to form clusters [13]. A node is any instance inside a cluster [13]. The node type describes the function that each node plays inside the cluster. Each type of node is given a role in a distributed application like Apache Hadoop by Amazon EMR, which also installs various software components on each node type [13].

C) S3 Bucket:

An object storage service called Amazon Simple Storage Service (Amazon S3) provides performance, security, and scalability that are unmatched in the market [14]. For a variety of use cases, including data lakes, websites, mobile applications, backup and restore, archives, enterprise applications, IoT devices, and big data analytics, customers of all sizes and sectors can use Amazon S3 to store and protect any amount of data [14]. In this project all the data was stored in S3 bucket and data was easily accessed by EMR clusters and Amazon Athena with the help of S3 Buckets.

V. Methodology And Visualization

This section highlights the methodology used for the project and the visualizations created for the analysis of E-scooter and socio-economic indicator data.

We made use of three main datasets for this project:

a) E-scooter availability data.

b) E-scooter trips data.

c) Socio Economic data (for USA).

The data was first preprocessed/cleaned, this included fixing the format of various columns in the data; fixing the data type of various columns, for example, certain columns had data type as a list and had only a single element inside the list such data types were fixed and replaced with a string or integer depending on the data type of the element in the list; further the data was of JSON format which stored everything as a key value pair certain key-value pairs were broken and only the value part was retained (this was not done for geometry column); Geometry column was fixed and converted to shapely geospatial data format – this was done to ensure that the data was readable in GeoPanda dataframes. Finally, only those columns were selected from the dataset that were required for the final visualization and processing in this paper. Processing of socio-economic data only included selecting columns that were required i.e., GEOID, Locations, median income, and geometry columns. Finally, all the duplicate records were removed from all the datasets.

Once the preprocessing was complete, visualizations were created on the GeoPanda dataframes using plotly and matplotlib. The starting points of each trip was also determined for analyzing the occupancy per area for each trip, the starting point was then joined with the socio-economic dataset was determining the area under which the trips started, this was achieved with the help of a spatial join (sjoin) available in geopanda's library.

Below we explain each visualization created as part of this paper.

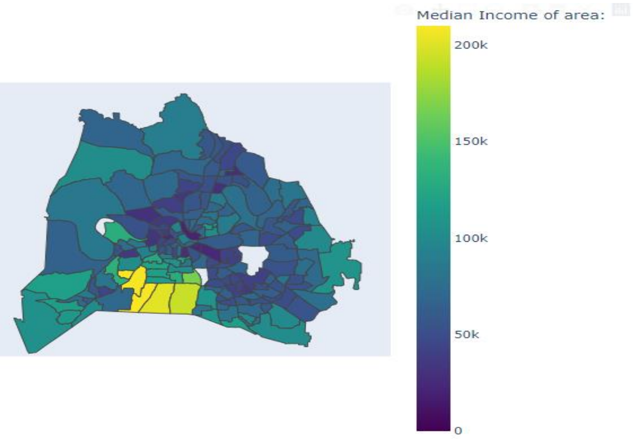

Fig 1: Socio-eco indicator

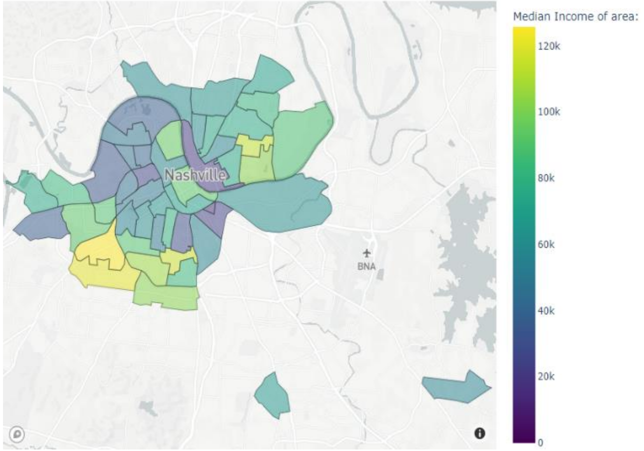

Fig 2: Socio-eco indicator only trips

The figures (1) and (2) are visualizations for socio-economic indicators for Nashville Davidson County. In both these graphs the legend indicates the median of each area ranging from 0 dollars to 120,000 dollars. Figure (2) is a subset of fig (1) it only contains those areas for which trip data was available. This graph was merged with trip start data points for further visualization.

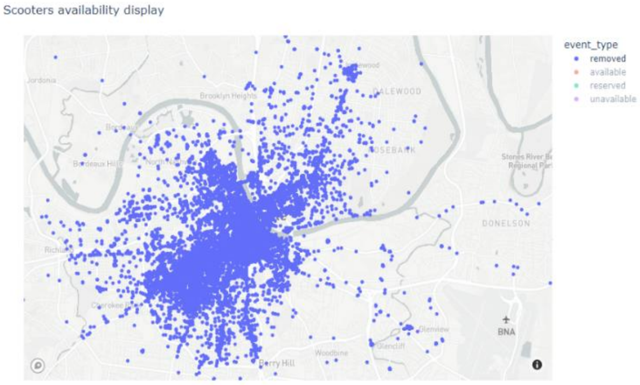

Fig 3: Scooter's removed visualization

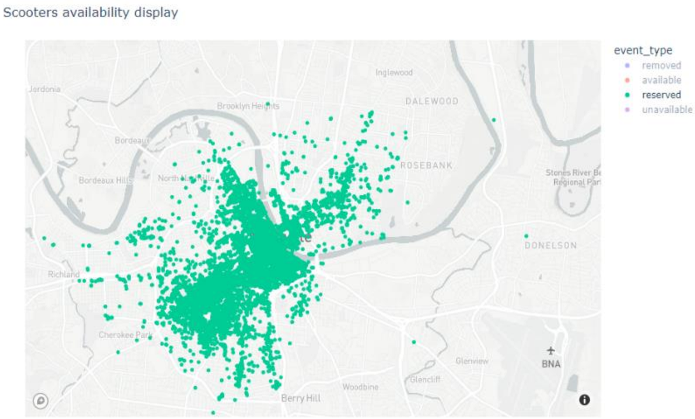

Fig 4: Scooter's reserved visualization

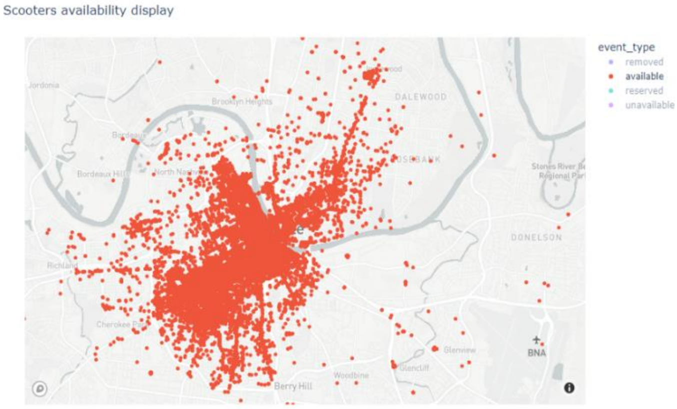

Fig 5: Scooter's available visualization

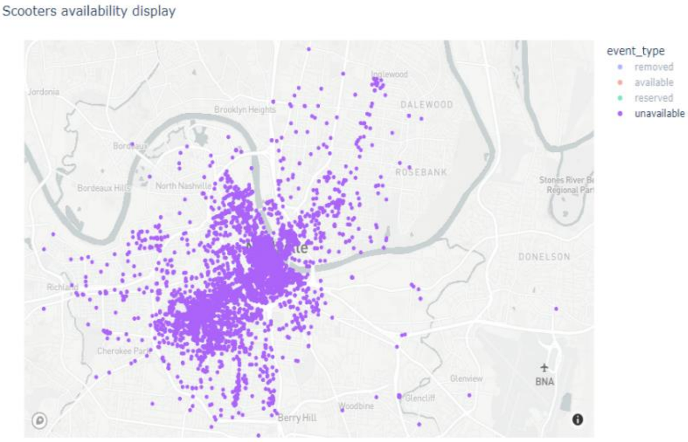

Fig 6: Scooter's unavailable visualization

Figures (3), (4), (5) and (6) are used to indicate the scooter availability in Davidson County. This includes scooters that were removed (Fig 3) these can include scooters that were removed due to some physical issue with the scooter such as a non-working battery, physical problem with the scooter such as a broken handle or problem with the braking system or even some technical problem such as software glitches. Fig 4 depicts scooters that are reserved at any time of the day. Fig 5 shows all the scooters that are available throughout Davidson County across the entire lifetime of the scooters. Fig 6 indicates scooters that are not available at any given time. This can happen due to issues such as no battery in scooter or a software update required for the scooter, etc.

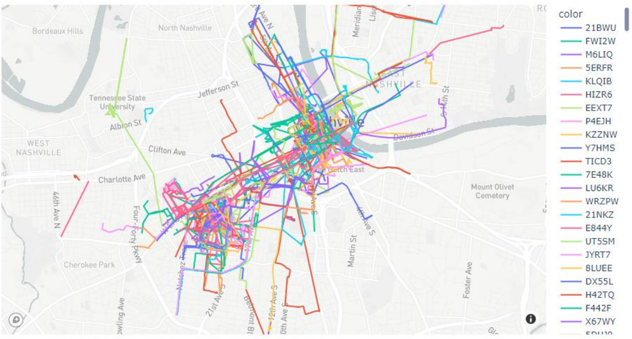

Fig 7: 500 E-scooter Trips in Davidson County

Figure (7) shows 500 random trips taken by e-scooter users in Davidson County. For visibility not all the 97000 trips were plotted on the graph. This graph was plotted using plotly. The legend in the graph corresponds to the ID associated with each e-scooter. There can be like with the same color as a single scooter can be used by multiple users at different times. The starting position of these can change as well as the scooters are manually picked up and charged by the scooter provider.

VI. Results

The results obtained in this paper were coherent with the hypothesis for the paper i.e., higher median income indicated more e-scooter usage. This statement is supported with the help of Figures (8) and (9) below.

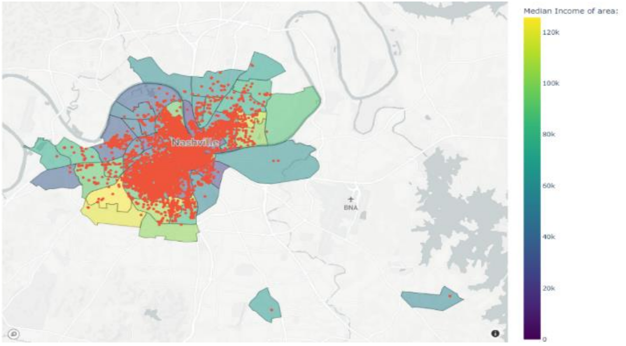

Fig 8: Socio-eco and trips merged

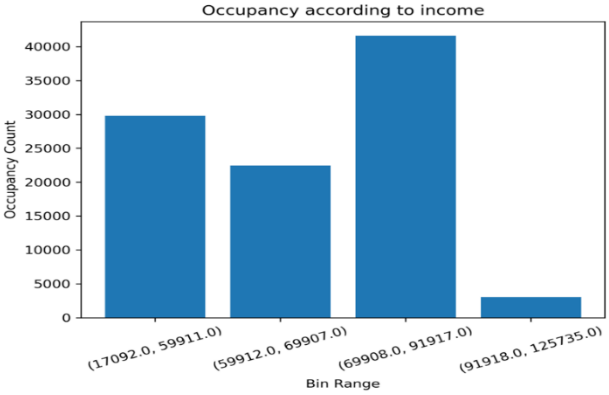

Fig (9): Occupancy of e-scooter

In figure (9) the socio-economic median income was divided into 4 bins:

1) (17092.0, 59911.0) – Low
2) (59912.0, 69907.0) – Medium
3) (69908.0, 91917.0) – High
4) (91918.0, 125735.0) – Very High

The trend for occupancy can be easily understood by referring to figures (2) and (8). The occupancy for bin range 3 is the highest indicating high occupancy where socio economic factors are high and from bins 0 and 1 low socio-economic factors co-relates to low e-scooter usage.

The important question is "why does bin 4 has low e-scooter usage?" as it completely breaks the hypothesis made for this paper. The reason for that can be understood from figure 2, as we can see that bin 4 mainly consists of area that are away from the city center, areas where people tend to use cars are their primary means of travel as they primarily travel long distances and fig (5) supports this hypothesis wherein there is low availability of e-scooters in areas away from the city center.

Spark played an important role in this paper as it was used for the entire processing of the big data used in this paper. It can in handy in a situation where pandas failed to read data and gave out of memory errors. Spark was able to process big data and convert it into a compressed and partitioned parquet file of 200 MB in a total of 3-5 minutes.

VII. Conclusion And Future Work

The work done in this paper can help drive further research in the optimization of e-scooter placement and charging. Future work also includes how weather, office hours and holidays play a role in e-scooter usage. The above-mentioned data can be visualized further to determine further insights into usage of e-scooters.

This work can also be further expanded to understand usage in other cities around the USA as well. Another potential research area is cost optimization of e-scooters, this problem can make use of weather data and an online generative model can be created with the help of Reinforcement learning to optimize the cost charged to users.

This paper can be further expanded by analyzing the end locations near the city center for creating charging stations for these e-scooters, where users can end trips at charging stations close by.

Acknowledgment

I would like to acknowledge My Professor DR. Abhishek Dubey for guiding us in this Project and for providing me with the data required for this Project.

References

[1] Steven Zaharia, M., et al. (2012). Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI 12).

[2] Apache Spark https://spark.apache.org/.

[3] Kumar, R. and Kant, R. (2017). Big Data Analytics with Spark. Cham: Springer.

[4] Farooq, U. and Sultana, S. (2018). Stream Processing with Apache Spark. Birmingham: Packt Publishing.

[5] McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

[6] Reback, J., et al. (2020). Pandas Development Team. pandas-dev/pandas: Pandas, version 1.1.4.

[7] VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.

[8] https://geopandas.org/en/stable/getting_started/introduction.html

[9] https://plotly.com/python/getting-started/

[10] https://plotly.com/python/maps-mapbox/

[11] https://www.mapbox.com/

[12] https://docs.aws.amazon.com/athena/latest/ug/what-is.html

[13] https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html

[14] https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html

[15] Kim S, Choo S, Lee G, Kim S. Predicting Demand for Shared E-Scooter Using Community Structure and Deep Learning Method. *Sustainability*. 2022; 14(5):2564. https://doi.org/10.3390/su14052564

[16] Campisi, T.; Akgün-Tanbay, N.; Nahiduzzaman, M.; Dissanayake, D. Uptake of e-Scooters in Palermo, Italy: Do the Road Users Tend to Rent, Buy or Share? In *International Conference on Computational Science and Its Applications*; Springer: Cham, Switerland, 2021; pp. 669–682

[17] Karim, S., et al. (2020). "E-scooter occupancy prediction using machine learning techniques." Transportation Research Part C: Emerging Technologies.

[18] Liu, M.; Seeder, S.; Li, H. Analysis of E-scooter Trips and Their Temporal Usage Patterns. *ITE J*. **2019**, *89*, 44–49

[19] Ricardo Sandoval , Caleb Van Geffen a, Michael Wilbur, Brandon Hall, Abhishek Dubey, William Barbour, Daniel B. Work Data driven methods for effective micromobility parking. https://www.sciencedirect.com/science/article/pii/S2590198221000750