Analyzing the Los Angeles Shared Bike Data

Weilin Ouyang[1], Mengjun Wang[2], Tianhao Wu[3], and Haoyu Li [4]

Abstract— **This analysis presents an in-depth exploration of a bike-sharing system, aiming to understand its efficacy and impact on urban transportation. The study methodically dissects various aspects of the system, including rush hour patterns, the relationship between trip durations and passholder types, and a comparative analysis of one-way and round-trip usage. Using sophisticated data processing techniques, such as regular expressions and K-Means clustering, alongside visualization tools like Pyecharts, the research reveals significant insights. The findings illustrate key trends in user behavior, peak usage times, and the distribution of different passholder types. This comprehensive examination not only sheds light on current usage patterns but also provides valuable information for policy development and potential enhancements to the system. The study underscores the critical role of bike-sharing in alleviating traffic congestion, reducing air pollution, and promoting sustainable urban mobility, thereby contributing to the discourse on the future of urban transportation solutions.**

Keywords— Urban Transportation, Bike-Sharing Analysis, User Behavior Trends, Sustainable Mobility

I. INTRODUCTION

The purpose of this analysis is to provide a comprehensive examination of the bike-sharing system, delving into its various aspects to gain insights into its impact, utilization, and potential areas for improvement. As bike sharing addresses issues such as traffic congestion, air pollution, and the need for sustainable transportation alternatives, it is increasingly important to understand the nuances and effectiveness of bike sharing programs. Our goal is to provide a balanced assessment that informs policy decisions, enhances user experience, and analyzes the sustainability of shared bikes for urban transportation.

II. OBJECTIVES

This paper aims to provide an in-depth analysis of the Los Angeles Metro Bike Share program and provide a comprehensive understanding of its operational efficiency and user behavior. Through data visualization technology, analyze and display the distribution of popular starting stations in the use of shared bicycles. This analysis will help us understand our users' preferred geographies and likely usage patterns. We also study peak hours of shared bike usage to identify daily travel patterns and temporal correlations. This will provide a basis for optimizing vehicle distribution and adjusting service plans. In order to understand the diversity of user needs and improve the membership program, we also analyzed the correlation between the length of bicycle use and the types of membership cards held, and explored the differences between different types of membership card holders (such as daily commuters, occasional users, etc.) Differences in length of bike use. Understand the nature of user journeys and destination choices by comparing data for one-way and round-trip journeys, providing insights for service improvements. Through these analyses, this paper hopes to provide practical suggestions for the management and development of the Los Angeles Metro bicycle sharing program, and also provide reference for bicycle sharing programs in other cities.

III. DATA

A. Data Collection

This study used the dataset provided by the Metro Bike Share which is a bicycle sharing system in the Los Angeles, California metropolitan area. The system uses a fleet of about 1,400 bikes and includes 93 stations in Downtown Los Angeles, Venice, and the Port of Los Angeles. Recent six years data are being collected from 2017 July to 2023 September. There are four quarters a year while each .csv file contains data for one quarter of the year. The provided data contains trip ID, trip duration, start and end time, start and end station, the geo-location for the stations, bike id, passholder type, trip route category and some other information.

B. Preprocessing

The dataset underwent initial filtering by the system, where staff service and test trips, as well as trips under one minute, were removed to ensure data relevance. Further processing was applied in this study: 1. Quarterly datasets were combined into a single dataset. 2. All time units were standardized to minutes. 3. Date formats were made consistent. 4. Trips exceeding 8 hours were excluded, under the assumption that such durations are humanly improbable and these riders fall outside the target customer group. 5. Trips with plan_duration values of 999, 150, and blank were excluded, as these are presumed to be used for staff testing. 6. To resolve inconsistencies in pass types, 'Flex Pass' was reclassified as 'Annual Pass' based on similar plan_duration characteristics. 7. Inconsistencies in 'Walk-up' passholder_type durations were corrected; instances marked as 1 were changed to 'One Day Pass' to reflect accurate duration. 8. Trips involving station "3000" were excluded, as

*This work was not supported by any organization

[1]W. Ouyang is student in Computer science. University of Tennessee, wouyang2 at vols.utk.edu

[2]M. Wang is student in Civil engineering. University of Tennessee, mwang43 at vols.utk.edu

[3]T. Wu is student in Computer Science, University of Tennessee, twu21 at vols.utk.edu

[4]H. Li is student in Computer Science, University of Tennessee, hli102 at vols.utk.edu

this station is a virtual one used by staff for bike transportation or operation. 9. Bikes modified from standard to smart were re-categorized under a new type: "standard/smart", to account for changes in bike type across different periods under the same bike_id. These steps were taken to enhance the accuracy and relevance of the data for analysis.

C. Feature Section

To facilitate the analysis of optimal bike station placement, key geographical features of stations were selected, including station ID, latitude, and longitude. This information is crucial for identifying 'hot' stations and aiding in future bike station planning. Concurrently, for a detailed examination of peak usage times and types of trips, various trip-related data points were extracted. These include trip duration, start and end times, trip route category, and plan duration. This comprehensive data collection will enable a thorough analysis of rush hour patterns and trip characteristics, providing valuable insights for improving bike-sharing services.

IV. Methodology

Four analyses are investigated in this study. The distribution of hot stations, rush hour analysis and the round trip/one way comparison analysis.

A. The Distribution of Hot Stations Analysis

The study utilizes data from the Los Angeles Metro Bike Share system. This dataset includes detailed information on bike sharing start stations across the LA metro area, including geographic coordinates (latitude and longitude) and usage statistics.

We performed an initial processing of the data to eliminate any inconsistencies or missing values, particularly in geographic coordinates and site usage counts. The cleaned dataset is then structured to focus on key variables: 'Starting Station Latitude', 'Starting Station Longitude', and 'Counts' (representing the number of times a station has been used as a starting point). We employ a scatter plot to visually represent the data. This method is chosen for its effectiveness in displaying geographical distributions and allowing for the easy identification of patterns and outliers.

We used the plot function from matplotlib.pyplot to create the scatter plot. Because the number of the count is too large, the size of the scatter points (s) is scaled, reduced by a factor of 10, with full opacity set by alpha=1. The plot is focused on the specific longitude (-118.28 to -118.22) and latitude (34.02 to 34.07) ranges, highlighting the key geographic area.

This methodology, utilizing matplotlib.pyplot for data visualization, provides a detailed and visually accessible exploration of popular bicycle sharing start stations in Los Angeles. The approach is designed to reveal insights into station distribution and usage patterns, offering valuable information for future development and optimization of the bike-sharing system.

B. Rush Hour Analysis

In this rush hour analysis, the methodology begins with data cleansing, involving the removal of missing values from the bike_info DataFrame. Regular expressions are then utilized to extract time information from the dataset. Specifically, two patterns are employed: pattern_time to capture the complete time format (hh:mm:ss), and pattern_hour to isolate the hour component.

The analysis iterates through the Start Time column, applying pattern_time to extract the full time stamp. In cases where time data is missing, a default handling mechanism is incorporated. The extracted times are further processed using pattern_hour to isolate the hours.

This hour data is then used to construct a line chart with Pyecharts, a tool for creating interactive web-based charts. The x-axis (attr) and y-axis (v) data are derived from the analysis. The chart includes customizations such as labels, mark points for maximum and minimum values, and a title, culminating in a comprehensive visualization of rush hour trends. The chart is finally rendered to an HTML file, 'rush-hour-line.html', for detailed examination.

C. Round Trip/One way Comparison Analysis

In this comparative analysis of 'Round Trip' and 'One Way' bike trips, the methodology involves segregating the data into two distinct groups based on the 'Trip Route Category' column in the 'bike_info' DataFrame. Two subsets, 'one_way_trip' and 'round_trip', are created to facilitate separate analyses of each category. For each group, the 'Start Hour' values are counted and sorted. This data is then used to construct line charts using Pyecharts, visually comparing the departure times for 'One Way' and 'Round Trip' categories. These charts highlight the peak hours for each trip type. Next, the proportion of different passholder types within each trip category is analyzed. The value counts for 'Passholder Type' are normalized and visualized using pie charts, again employing Pyecharts. This offers insights into the distribution of passholder types between 'One Way' and 'Round Trip' users. Finally, a descriptive statistical analysis of trip durations for each category is performed. This involves summarizing the 'Duration' data for 'One Way' and 'Round Trip' trips separately, providing an overview of the typical trip lengths for each category. This comprehensive methodology enables a nuanced understanding of the differences and similarities between 'One Way' and 'Round Trip' bike trips in terms of time, passholder types, and trip duration.

D. Relevancy Analysis Between Duration & Passholder Type

In this relevancy analysis between duration and passholder type, the methodology involves selecting specific columns from the 'bike_info' DataFrame to form 'bike_trip_info'. K-Means clustering, a popular unsupervised machine learning technique, is employed to classify trip durations into four categories: short, medium, long, and very long.

The KMeans model is trained on the 'Duration' data, and the resulting cluster centers are used to define the

duration classes. These classes are then mapped back to the ‘bike$_{t}rip_{i}nfo$‘$DataFrametocategorizeeachtrip$.

To understand the distribution of passholder types across these duration categories, value counts are computed for each class and normalized. This analysis leads to the creation of a radar chart using Pyecharts. The chart visualizes the proportion of different passholder types within each duration category, offering insights into user behavior patterns. The final output is rendered into an HTML file, ’duration-rate-radar.html’, for a detailed and interactive representation of the findings.

V. Results & Conclusion

A. The Distribution of Hot Stations Analysis

Fig. 1 illustrates the map view of Los Angeles, while Fig. 2 shows an example view of the bike route. After analyzing the hot station geolocation, Fig. 3 shows the distribution of the hot stations.

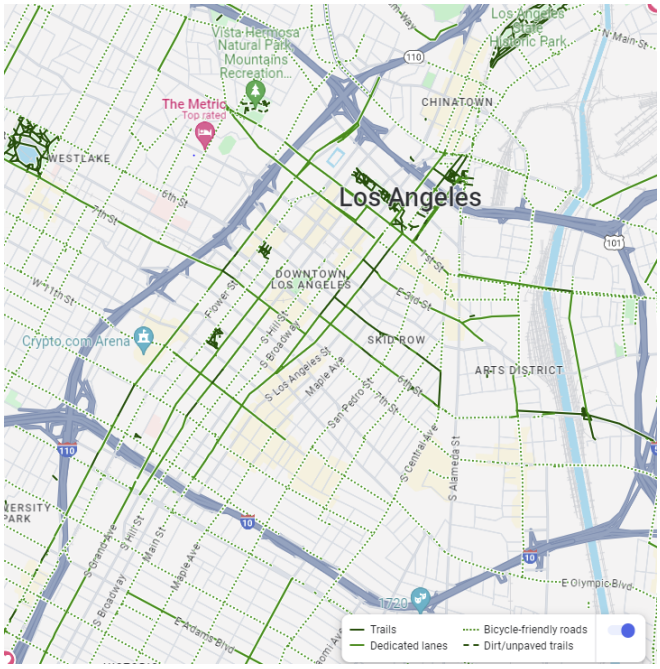

Fig. 1. Los Angeles city map.

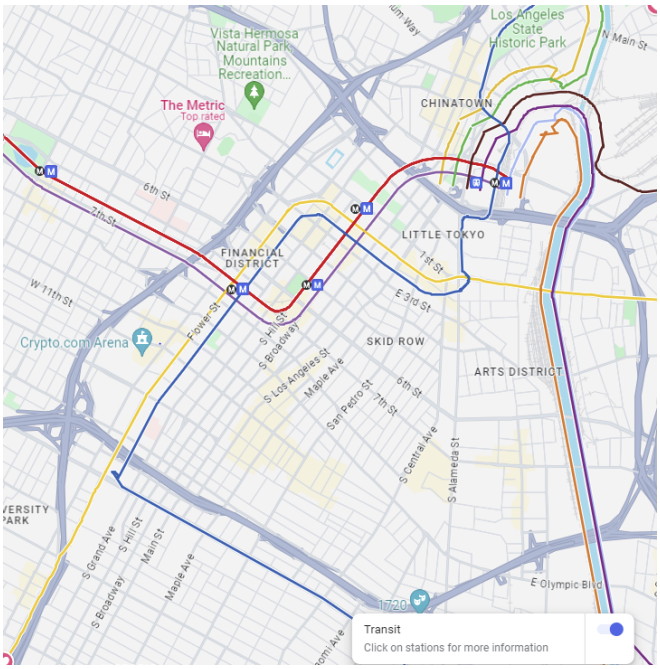

Fig. 2. The distributed bike route overview.

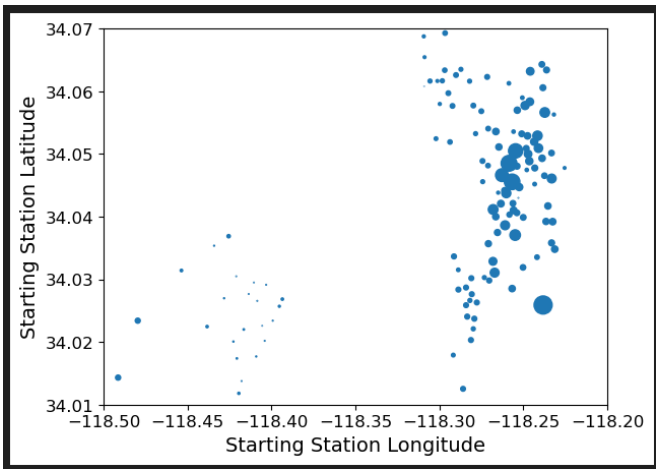

Fig. 3. The distribution of hot stations.

B. Rush Hour Analysis

Fig. 4 shows the user pattern for different time slots. The relative analysis is also described below.

- Shared-bike riders usually choose to start their trip at 7 a.m.–11 p.m.
- 7 a.m.–10 a.m. morning peak period, 10 a.m.–13 p.m. afternoon peak period and 16 p.m.–18 p.m. evening peak period show a significant increase in shared-bike use
- The peak of use is 18 p.m., which is presumed to be the evening peak
- The lowest point of use is 4 a.m.; a few shared-bikes are used in mid-night

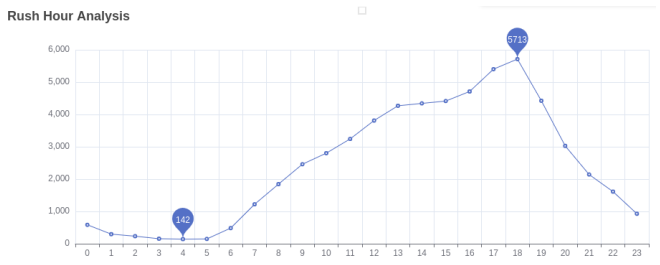

Fig. 4. Number of uses of bikes in different time slots.

- From the peak to the lowest point, there is a complete downward trend, with the vehicle gradually decreasing.

C. Round Trip/One Way Comparison Analysis

Figs. 5 and 6 show the percentages for different memberships for single trips and round trips, respectively.

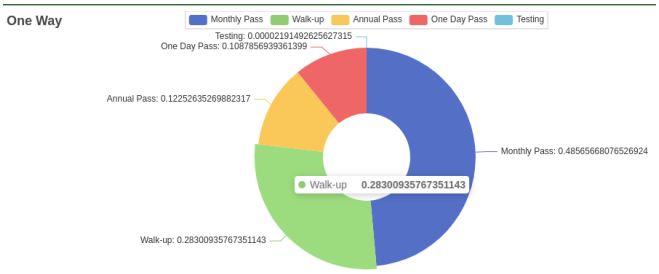

Fig. 5. Precentage for different membership (Single Trip)

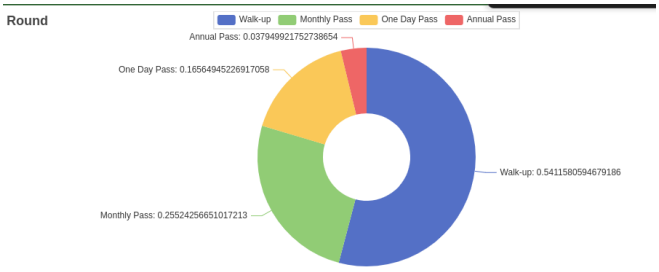

Fig. 6. Precentage for different membership (Round Trip)

Fig. 7 also compares the departure time differences among different trip types.

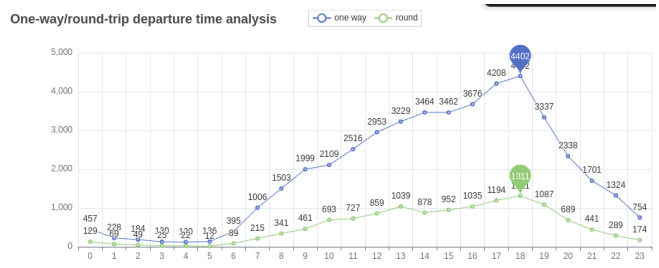

Fig. 7. Distribution of Departure Time.

- The proportion of users with membership cards for One Way Trip's shared bicycles is very high, accounting for over 70
- The majority of Round Trip's shared bicycle users are casual users, with only thirty percent having membership cards, and most of these members are also monthly card holders.

D. Relevancy Analysis Between Duration and Passholder Type

Fig. 8 shows the radar plot of the duration rate analysis.

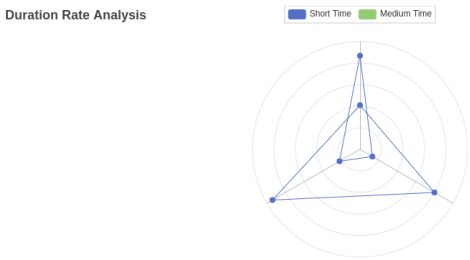

Fig. 8. Duration rate analysis.

- The majority of users who use shared bikes for a very long time are those who use them immediately.
- Over 60
- The percentage of those who possess Flex-Pass among short-time trip bike users is also the highest among all types.

E. Analysis of the relationship between the usage duration of one-way, Round, and Shared Bicycles

Fig. 9 shows a detailed comparison of the single-trip and round-trip. Based on the mean, median, and upper and lower quartiles in the table above, it is evident that the duration of use for Round Trip bicycles is generally longer than that of One Way Trip bicycles, which is consistent with expectations. VI. Limitation

	One Way	Round
count	45631.000000	12780.000000
mean	39.475839	68.067136
std	136.662651	129.701820
min	1.000000	1.000000
25%	8.000000	18.000000
50%	15.000000	35.000000
75%	27.000000	75.000000
max	1440.000000	1440.000000

Fig. 9. The time parameter analysis for both types of trips.

Managing and filtering the vast dataset to eliminate irrelevant or misleading information proved more challenging than anticipated. As traffic analysis can be influenced by environmental factors, this study focused solely on analyzing the existing data and its characteristics. Consequently, it may lack robustness in adapting to changes.

VII. Future Work

This study concentrates on analyzing data to identify key characteristics that could inform recommendations for bike traffic distribution in Los Angeles. However, the current approach has limitations in adapting to environmental changes and does not account for unforeseen factors. In future research, the authors aim to develop a more resilient analytical method. This may involve incorporating machine learning prediction models to yield more reliable and insightful results.

VIII. Org Chart

A. Distribution

Mengjun Wang: Mengjun Wang is responsible for gathering and preparing the necessary data for our project. This includes collecting relevant datasets, cleaning and organizing the data, and ensuring its readiness for analysis. Haoyu Li: Haoyu Li's primary role is to conduct data analysis and create visualizations that help interpret the data effectively. He will employ various analytical techniques and tools to extract insights from the collected data. Weilin Ouyang: Weilin Ouyang is tasked with performing statistical analyses on the data. His responsibilities include running statistical tests, modeling, and drawing meaningful conclusions from the data to support our research objectives. Tianhao Wu: Tianhao Wu is responsible for documenting our project's progress and findings. He will prepare written reports, documentation of methodologies, and any necessary presentations to communicate our results clearly.

B. milestone

09/28–10/10 data cleaning and preprocessing; 10/11–11/15 data and result analysis; 11/16–12/05 report writing

Fig. 10 below shows the timeline overview of this project.

Fig. 10. Project Timeline Overview.

Fig. 11 below shows the organization chart for this project.

Fig. 11. Organization Chart.