a bike-sharing business analysis project
(first edition)

by Qi Zhou / qiqizhou1996@gmail.com
date: March 18, 2025

Introduction:

This project [1] demonstrates a fundamental data analysis process: asking questions, preparing datasets, cleaning data, analyzing and visualizing data, and sharing a compelling data story. The NY Citi bike-sharing program [2] aims to enhance its business strategy, thereby improving user experience.

Asking questions:

1. What are the most popular pick-up locations across the city for NY Citi Bike rental?
2. How does the average trip duration vary across different age groups?
3. Which age group rents the most bikes?
4. How does bike rental vary across the two user groups (one-time users vs. long-term subscribers) on different days of the week?
5. Does user age impact the average bike trip duration?

Preparing datasets:

The dataset [3] is chosen to answer the above questions, thus revealing how the bike-sharing service serves its users and to drive informed decision-making accordingly.

In this project, Google Sheets is the tool used to store, clean, and analyze the data. The raw data, which totals 20,400 records, spans three months, starting on January 1, 2017, and ending on March 31 of the same year, as shown in Figure 1.

(Figure 1: Raw Data)

Cleaning data:

After obtaining the raw data, it is first backed up. Then, the backup data is cleaned, and the raw data remains intact to ensure traceability and reproducibility of subsequent cleanup operations. Regarding this data, the cleanup includes three steps - identifying and removing duplicates, trimming whitespaces, and handling missing data points as well as outliers.

Compared with the raw data, the cleaned data totals 16,843 records, as shown

in Figure 2. Besides, the attributes are highlighted with a light green background and bold font so as to help understand the table quickly and easily.
(Figure 2: Cleanup on Backup Data)

Analyzing & Visualizing data:

To answer the first question, a pivot table is created to summarize the use frequency of each starting station. After sorting in descending order, Grove St PATH station is the most popular starting station among the 50 stations, with 2,115 out of 16,843 data records (note that the average is 337). Moreover, as the bar chart in Figure 3 suggests, Exchange Place, Sip Ave, and Hamilton Park stations are quite highly used. Thus, a possible business strategy is supplying these stations with sufficient bikes promptly to meet the large demand.
(Figure 3: Q1 - Top 15 Pickup Locations)

For the second question, a pivot table summarizes the average trip duration for seven age groups (18-24, 25-34, 35-44, 45-54, 55-64, 65-74 and 75+). As the column chart in Figure 4 suggests, most users aged 18 to 74 spend an average of 9 minutes on a short ride, while users over 75 spend an average of 50 minutes on long trips. Thus, the business strategy would focus on the ten-minute ride.
(Figure 4: Q2 - Trip Duration vs. Age Groups)

For the third question, a pivot table summarizes the number of bike rentals for the seven age groups. As the pie chart in Figure 5 suggests, the 35-44 age group is the largest user group, with 7,697 of 16,843 rentals, accounting for nearly half of the total. In addition, the 25-34 and 45-54 age groups account for another 41.5% of rentals, so they are also vitally important user groups. Thus, the business strategy would mainly target users aged 25 to 54.
(Figure 5: Q3 - Bike Rental vs. Age Groups)

For the fourth question, a pivot table summarizes the number of bike rentals per weekday by one-time users and subscribers. As the stacked stepped area chart in Figure 6 suggests, most users are subscribers who primarily use the service during the workday, with 13,528 of 16,843 rentals. On the other hand, one-time users tend to rent bikes for the weekend. Thus, the business strategy would greatly retain subscribers and try to attract one-time users on the weekends.
(Figure 6: Q4 - User Type vs. Weekday)

Regarding the last question, descriptive statistics, in which the average trip

duration is 9 minutes (shown in Figure 7), are used to generate a general intuitive understanding of the whole 16,843 data records and help effectively eliminate outliers. As the scatter plot in Figure 7 suggests, the 16,843 pairs of data indicate no significant relationship between trip duration and user age. Thus, it seems that user age does not impact the average bike trip duration. For further questions, more data is needed to conduct a detailed analysis.
(Figure 7: Q5 - User Age vs. Trip Duration)

Data storytelling:

This project leverages descriptive analysis to understand how the NY Citi bike-sharing program operates, plus its customer behaviour, thus providing specific solutions for upgrading the business strategy.

The raw dataset contains 20,400 records, spanning three months, from January 1 to March 31, 2017. After data cleaning, the dataset retains 16,843 records. According to the cleaned data, there are 50 starting stations, and the average number of services provided by the stations over these three months is 337. However, four stations - Grove St PATH, Exchange Place, Sip Ave, and Hamilton Park - each provide over 1,000 times of service. Therefore, these stations should be supplied with sufficient bikes promptly to meet this high demand. Additionally, more than 80% of customers are subscribers, primarily aged between 25 and 54, who are accustomed to renting bikes for ten-minute short rides during workdays. Thus, the business strategy should focus on retaining these customers through subscriber benefits and improved riding experiences.
(Report)
(Presentation)