Project Report

DIVVY BIKES DATA ANALYSIS PROJECT

Introduction

Bicycle-sharing systems, which can provide shared bike usage services for the public, have been launched in many big cities. In bicycle-sharing systems, people can borrow and return bikes at any stations in the service region very conveniently. Therefore, bicycle-sharing systems are normally used as a short distance trip supplement for private vehicles as well as regular public transportation

Divvy Bikes is Chicagoland's bike share system across Chicago and Evanston. Divvy provides residents and visitors with a convenient, fun and affordable transportation option for getting around and exploring Chicago.

Divvy, like other bike share systems, consists of a fleet of specially-designed, sturdy and durable bikes that are locked into a network of docking stations throughout the region. The bikes can be unlocked from one station and returned to any other station in the system. One can become an Annual Member or buy a Pass from a Divvy station kiosk or the Divvy App.

People use bike share to explore Chicago, commute to work or school, run errands, get to appointments or social engagements, and more. Divvy is available for use 24 hours/day, 7 days/week, 365 days/year, and riders have access to all bikes and stations across the system.(https://divvybikes.com/about)

Analysis Objectives

This dataset covers the period for the first quarter of 2019. The purpose of this project is to gain some insight into city-wide biking trends by analyzing the Divvy trip data. My objectives here are to look into the number of Divvy trips and bikes used, popular start and end stations, demographic information of Divvy Bikes users and the days and times Divvy bikes are used.

I'll go though my analysis by answering the below questions.

1. How many trips occurred in the first quarter of 2019?

2. How many bikes were used in the first quarter of 2019?

3. Which bikes were used most often for trips?

4. What is the average trip duration based on user type?

5. What are the most popular start stations?

6. What are the most popular end stations?

7. What are the most popular routes?

8. How is Divvy Bikes usage divided between customers and subscribers?

9. Are Divvy Bikes used more by female or male users?

10. What age are Divvy Bikes users?

11. At which times are Divvy Bikes most used?

12. How is Divvy Bikes usage by day of the week?

Dataset description

This is a Historical trip data of Divvy Bikes, it's available for public use and was downloaded from (https://divvybikes.com/system-data). The dataset is provided according to the Divvy Data License Agreement, and released on a monthly schedule.

This data was downloaded as a CSV file "Divvy_Trips_2019_Q1" , it contains 365069 rows and 12 columns, each row represents a trip record in the dataset with a unique ID. We can know the trip start and end time as well as the corresponding origin and destination bike stations. The trip record also indicates whether the user is an annual membership holder or just a one-day pass holder, who are called the "subscriber" and "customer" respectively. The trip record data also includes users' gender and birth year information.

Column data types and description:

Column Name	Data Type	Column Description
trip_id	int64	Unique number representing each trip
start_time	object	Trip start date and time
end_time	object	Trip end date and time
bikeid	int64	Unique number representing each bike
tripduration	object	The trip length in seconds
from_station_id	int64	Unique number representing each station
from_station_name	object	Name of the station where the trip started
to_station_id	int64	Unique number representing each station
to_station_name	object	Name of the station where the trip ended
usertype	object	User type (subscribers for annual members, customer for one trip and one_day pass)
gender	object	User gender as disclosed by users
birthyear	float64	User birth year as disclosed by users

Data Preparation

In this project I will be using some libraries for analysis like pandas to clean, explore, and manipulate data, numpy to perform mathematical operations, matplotlib.pyplot to create animated visualizations, datetime to manipulate date and time object data and calendar to display days of the week.

First of all, for the sake of smoothing the later analysis, the dataset is under a cleansing process in order to confirm its integrity and comprehensibility. Most of the errors, such as rows with wrong information or wrong data type will be amended and corrected. Additionally, I will make some changes to the data by adding useful new columns based on calculations of already existing columns in order to facilitate data analysis and arrive at more insightful conclusions.

When exploring the data types of each column, I observed that the dataset has no duplicate rows but it contains null values only in two columns. The ‘gender’ column has 19711 and the 'birthyear' column has 18023 null values. As these are too many null values I will avoid removing these records and instead use a different solution in order not to lose valuable data from the columns.

start_time and end_time columns are of object data type, which makes it harder to do useful analysis on them, so they were casted to date format. Converting these columns to date format will allow extracting useful date information like day and time from these columns. In this analysis I focused on start_time column and used the extracted information to create two new columns. See created columns below.

Column Name	Data Type	Column Description
start_hour	int64	Trip start hour of the day
day_of_week	object	Trip start day of the week

The tripduration column had few data issues to fix before starting analysis. The tripduration column is of object data type, and needed to be casted to Integer data type to do calculations on it. When converted to Integer, it prompted an error, since some big values contained “,” as separator, so I removed the “,” from these values first and then converted the data type to Integer.

When exploring the tripduration column, I noticed that some trips have extremely long durations up to 2952 hours. Trips with such long durations are likely due to an error, so I needed to verify this irregularity. When I looked at the Divvy Bikes official website (https://divvybikes.com/pricing/day-pass) I found that Divvy Bikes offer a Day Pass subscription.The Day Pass includes unlimited classic bike rides in a 24-hour period, up to 3 hours each. If users want to take a ride longer than 3 hours, they can either check their bike in and check another bike out to avoid extra time fees, or pay just an extra $0.17 per minute to keep the same bike for more than 3 hours.

Taking a closer look at these long trips, I can see 763 entries where trip duration varied from 3 to 2952 hours. It appears that these long trips are likely to be an error due to users forgetting to return a bike or a bike not being docked properly. Having these incorrect values will affect the overall analysis. Since these entries are not too many, I decided to calculate the mean duration for all trips less than three hours and replaced the long trip duration with the mean duration.

I also used the "tripduration" column to create the “tripduration_min” column for the trip duration in minutes.

Column Name	Data Type	Column Description
tripduration_min	Int64	The trip length in minutes

The “gender” column has 19711 null values, which will affect the data analysis. I decided not to delete these entries as they are so many that it would mean losing all valuable data in other columns, so I filled all null values with “Missing" and I disregarded “Missing” values when doing analysis on “gender” column.

The “birthyear” column has 18023 null values which I left untouched as I only used the "birthyear" column to create a new “user_age” column by subtracting birth year from current date.

Column Name	Data Type	Column Description
user_age	Int64	User age based on year of birth

Taking a quick look at the new “user_age” column, I found that the minimum user age is 20 years while the maximum user age is 123 years old. As the maximum age doesn’t seem realistic for a bike sharing business, I realized that it could be an error or human mistake while entering user data. When I looked for ages above 80 years, I found 280 entries. I decided to calculate the average age for users below 80 years which is 41 years. I replaced the extreme ages above 80 with the average age.

The “user_age” column was created based on the year of birth so it contained 18023 null values. I decided against deleting these entries so as not to lose all valuable data in other columns. Hence, I filled these null values with the average user age 41 years old.

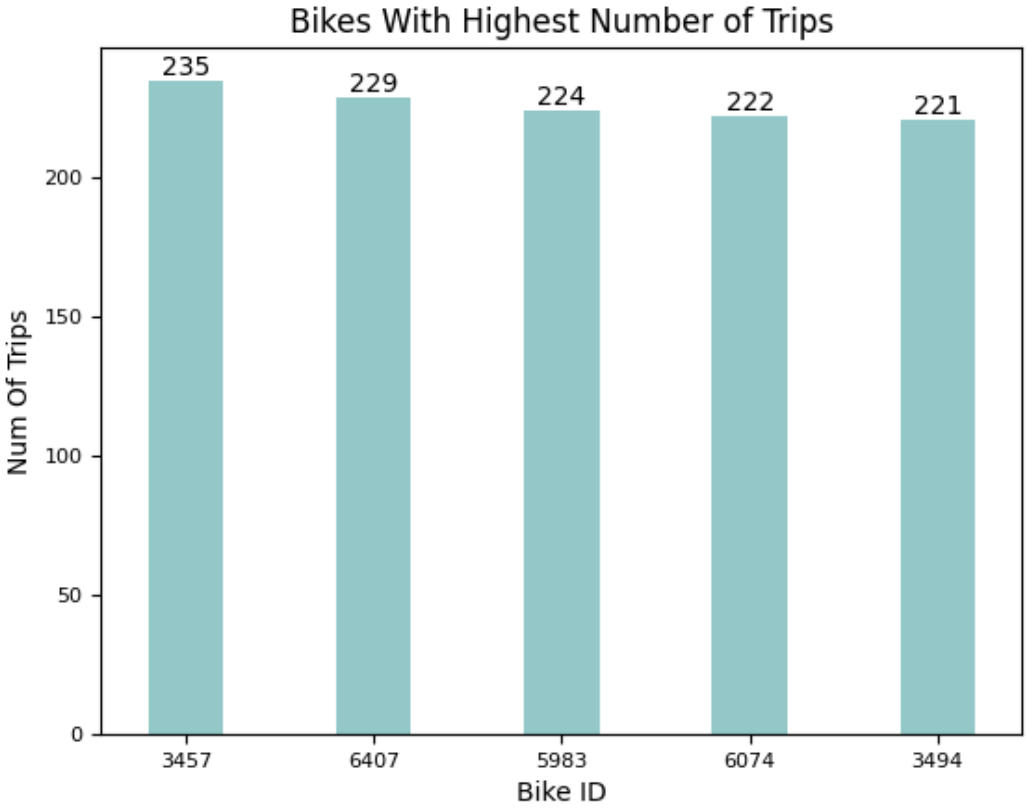
Data Insights

After cleaning and preparing the dataset for analysis, it is time to use the data to draw useful insights and answer the predefined questions.

1. Number of Divvy Trips, Most Used Bikes and Average Trip Duration

Analyzing the dataset shows that during the first quarter of 2019, there were 365069 trips completed and 4769 bikes were used for these trips.

The bikes used for the highest number of rides were (3457, 6407, 5983, 6074, 3494)



The mean trip duration for all users is 12 minutes. The mean trip duration of subscription riders is 11 minutes, which is lower than the mean trip duration of all trips, while it is exactly the opposite for casual riders, whose mean trip duration is 31 minutes, which is higher than the the mean trip duration of all trips. This tells us that customer riders usually take the bikes out for a longer duration compared to subscribers.

2. Popular Start and End Stations and Popular Trip Routes

Identifying the popular stations can provide valuable strategic insights for the business. It allows to focus maintenance efforts on highly frequented stations. Moreover, knowledge of the unique factors of popularity of these stations can be leveraged to increase traffic and number of customers at other locations, or even suggest ideal spots for opening new stations.

The “from_station_name” column was analyzed to get the most popular start stations for most trips. Clinton St & Washington Blvd station is the most popular start station with 7699 trips.

Trip Start Station	Number of Trips
Clinton St & Washington Blvd	7699
Clinton St & Madison St	6565
Canal St & Adams St	6342
Columbus Dr & Randolph St	4655
Canal St & Madison St	4571

The “to_station_name” column was used to get the most popular end stations for most trips. Clinton St & Washington Blvd station is also the most popular end station with 7699 trips.

Trip End Station	Number of Trips
Clinton St & Washington Blvd	7699
Clinton St & Madison St	6859
Canal St & Adams St	6744
Canal St & Madison St	4875
Michigan Ave & Washington St	4412

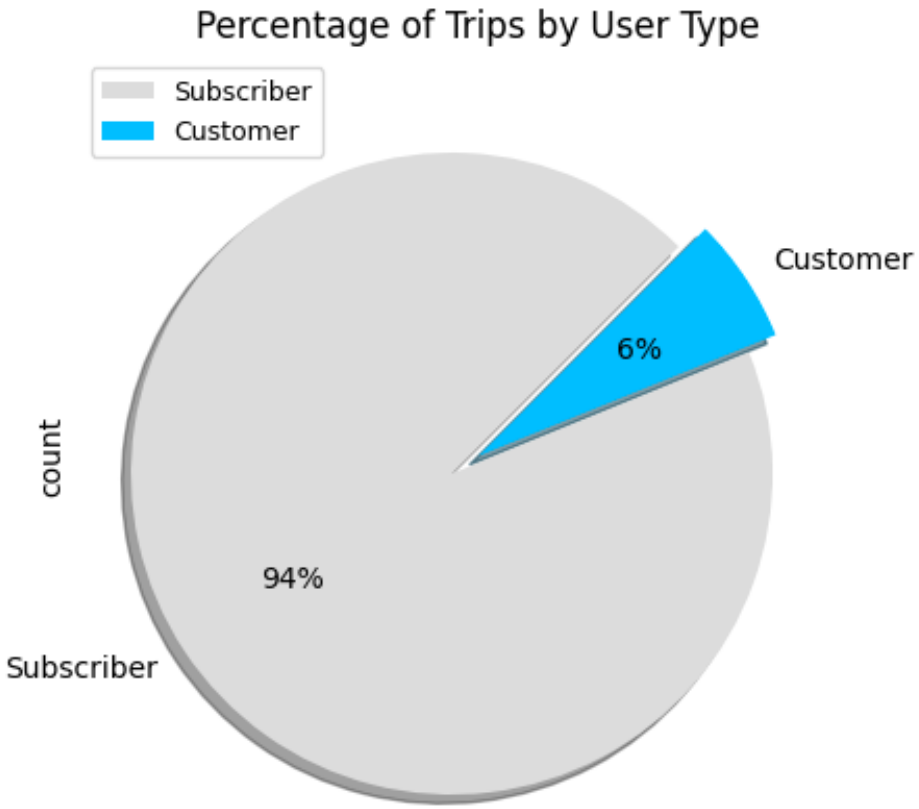
The “from_station_name” and “to_station_name” columns were grouped together to get the most popular trip routes taken. The trip from Michigan Ave & Washington St station to Clinton St & Washington Blvd station is the most popular route with 513 trips.

Trip Start Station	Trip End Station	Number of Trips
Michigan Ave & Washington St	**Clinton St & Washington Blvd**	513
Columbus Dr & Randolph St	**Clinton St & Washington Blvd**	491
Michigan Ave & Washington St	**Canal St & Adams St**	482
Canal St & Madison St	**Michigan Ave & Washington St**	479
Canal St & Adams St	**Michigan Ave & Washington St**	447

3. Divvy Bikes User Types and Demographic Information

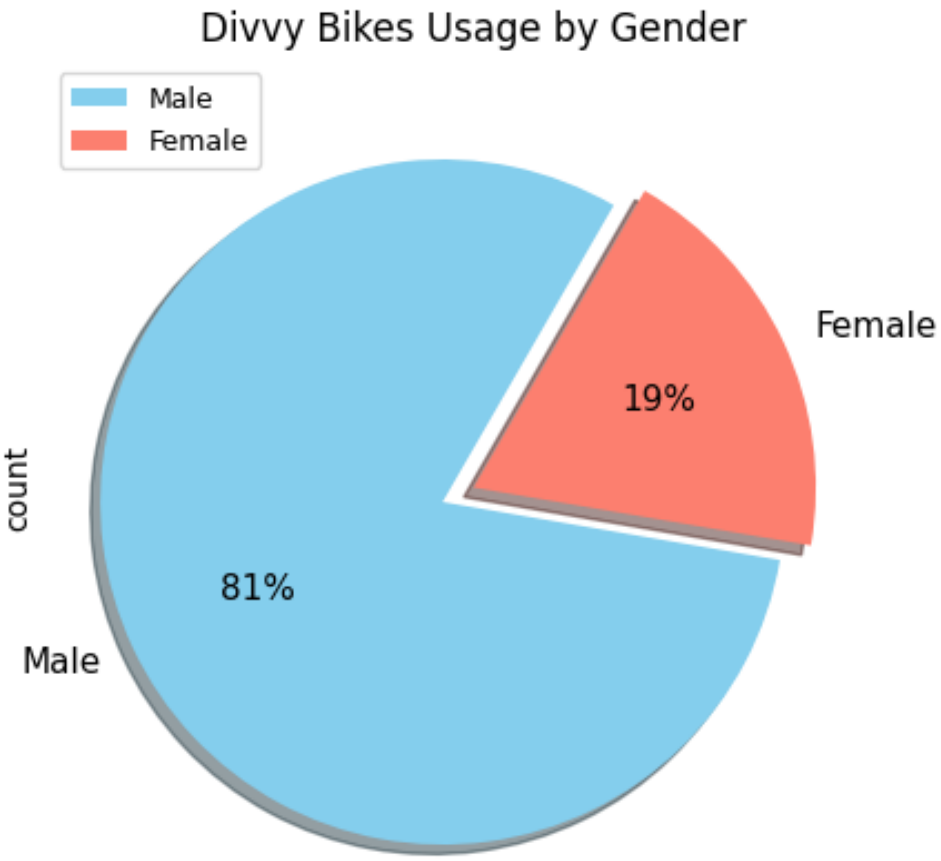
- **Divvy Bikes Usage by Users' Ages**

I analyzed the "usertype" column to calculate the percentage between trips done by subscribers and customers. There are 341906 trips by subscribers and 23163 trips by customers. Subscriber trips represents 94% of total trips while customer trips represent 6%. Subscriber trips make up a majority of rides.



- **Divvy Bikes Usage by Users Gender**

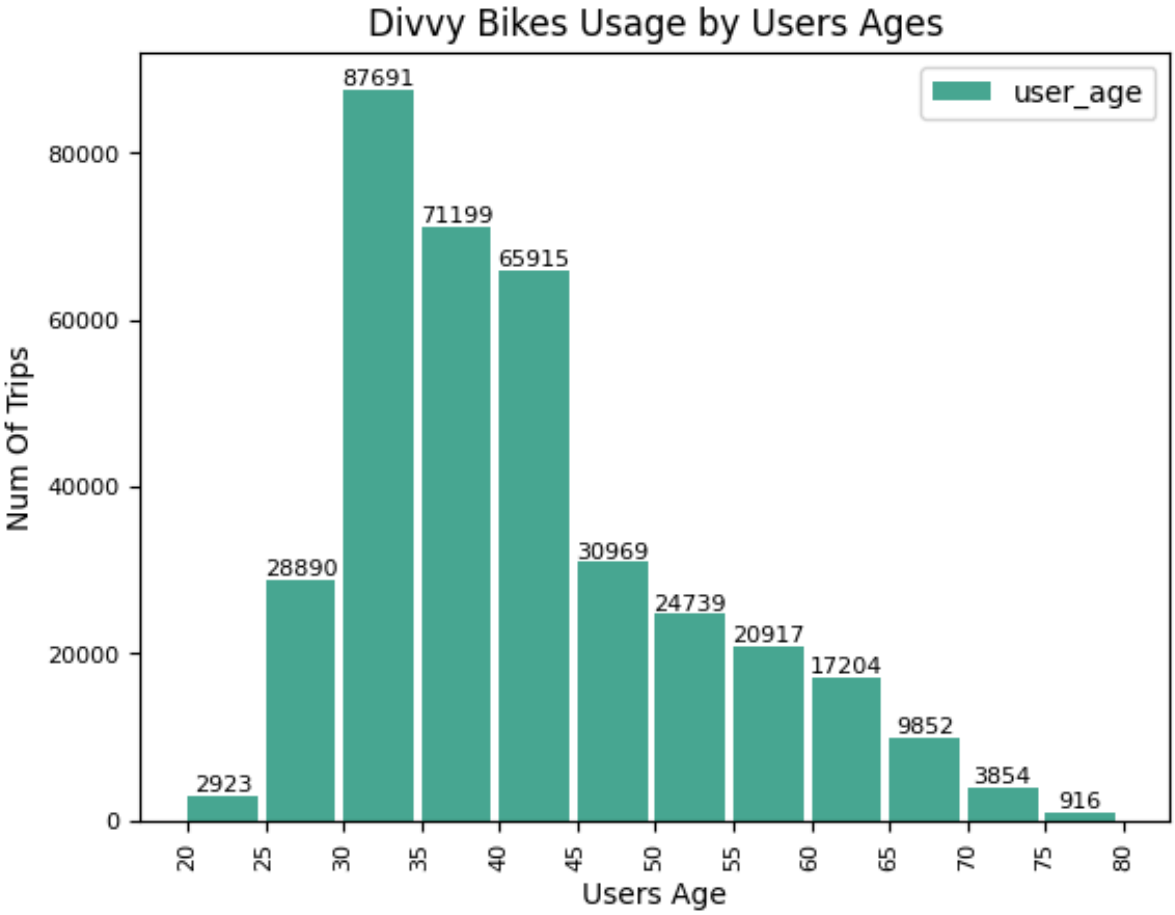
Divvy bikes are used by male and female users, and to get the percentages of who used Divvy bikes more, I analyzed the data in the “gender” column. First I disregarded 19711 missing values and then I discovered 278440 trips by male and 66918 trips by female users, which translates to 81% of trips made by male users and 19% by female users.



- **Divvy Bikes Usage by Users' Ages**

After calculating user age based on birthyear, I analyzed the "age" column to see what age group uses Divvy Bikes most. The minimum user age is 20 years and the maximum user age is 80 years. By dividing user age into 12 age groups, with each group spanning 5 years, it appears that users between 30 and 35 years have the highest number of trips at 87691, while users between 35 and 40 years come in second with 71199 trips.

The two age groups with the lowest number of trips are users between 75 and 80 years, who did 916 trips, and users between 20 and 25 years with 2923 trips.

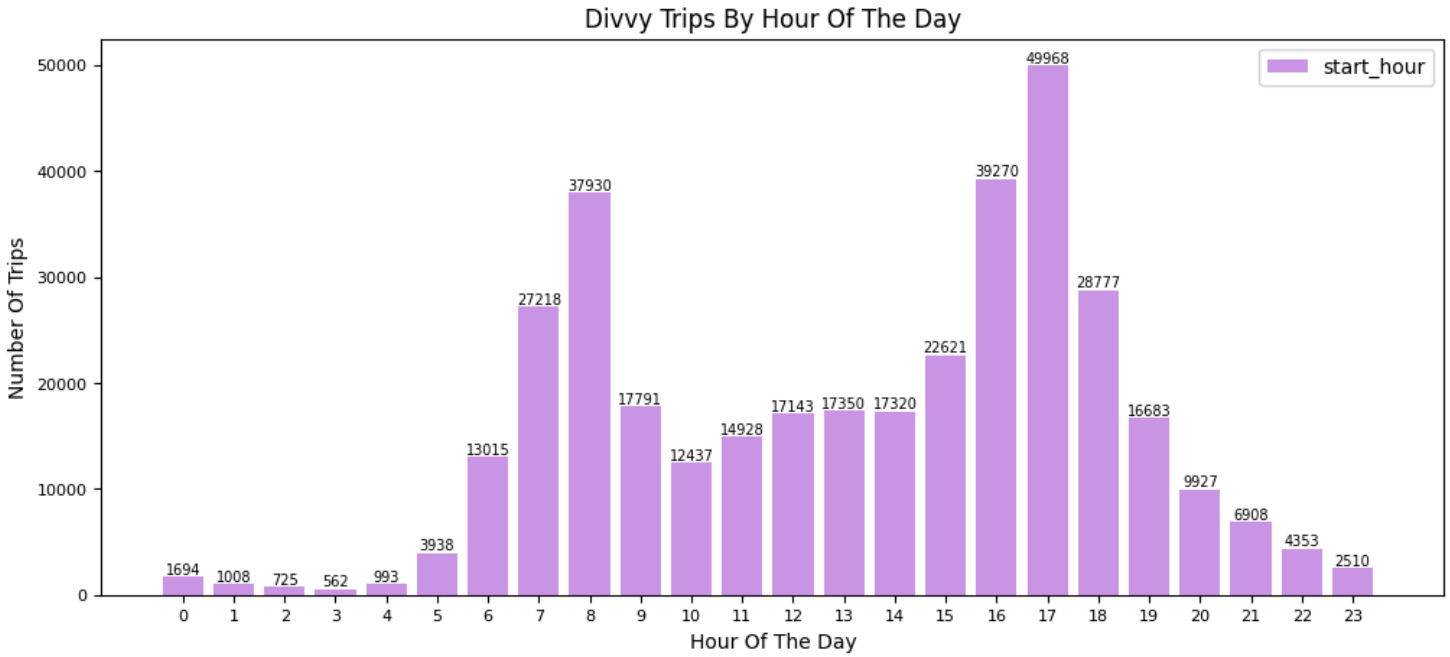


4. Cyclical Pattern in Each Hour and Each Day

Identifying which days and what times are the busiest will help adjusting resources accordingly.

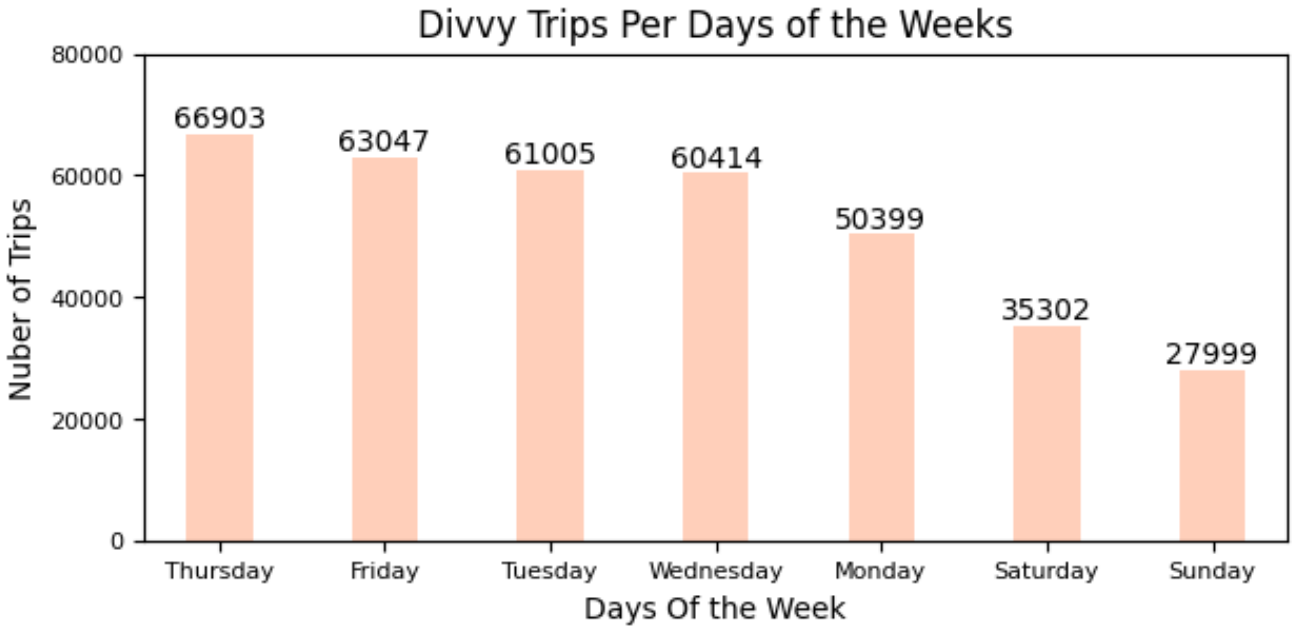
- **Cyclical Pattern By Hour**

The “start_hour” column can be displayed in a histogram chart to show when Divvy bikes were used most during the day. I noticed that user ride behavior is influenced by the workday. Peak times can be observed in the afternoon at 5 pm and 4 pm with 49968 and 39270 trips respectively, and in the morning at 8 am and 7 am with 37930 and 27218 trips each which indicate that users used Divvy Bikes to travel to and from work. The lowest usage of Divvy Bikes occurs between midnight and 5 am.



- **Cyclical Pattern by Day**

Analyzing the “day_of_week” column shows us that the number of trips during weekdays is relatively similar and does not differ too much. Thursday has the highest number of trips at 66903. Saturdays and Sundays show 35302 and 27999 trips respectively, reducing the number of trips during the weekend by 50% compared to weekdays.



References

Divvy Bikes Official Website

Dataset downloading link