



**Final Report**

**Data Visualization and Analytics on Divvy  
Chicago Bike with Snowflake**

**Team project topic #04**

**Team:**

Jaichiddharth Rangasamy Karthigeyan – A20527281

Muzammil Mohammed – A20527360

Spandana Vemula - A20527937

Pavitra Sai Vegiraju - A20525304

Vikash Singh - A20525680

**Instructor**

Prof. Joseph Rosen

**Course**

Fall 2023 Big Data Technologies (CSP-554)

## Table of Contents

<b>Abstract:</b>	<b>3</b>
<b>Introduction:</b>	<b>3</b>
<b>Literature Review:</b>	<b>3</b>
Snowflake.....	3
Amazon Redshift: .....	4
Azure Synapse Analytics (formerly SQL Data Warehouse):.....	4
<b>Project Description:</b>	<b>6</b>
<b>Data Collection:</b>	<b>6</b>
Data Attributes: .....	7
<b>Data Preparation:</b>	<b>7</b>
<b>Data Cleaning:</b>	<b>9</b>
<b>DATA ANALYSIS AND VISUALIZATION:</b>	<b>13</b>
1) Number of rides under time free cost limitation (45 minutes):.....	13
2) User Behavior and Performance Metrics Across Different Rideable Types, User Categories, and Days of the Week: .....	14
3) Analysis of User Behavior and Performance Metrics Across Different Rideable Types, User Categories, and Days of the Week:.....	16
4) User Behavior and Performance Metrics Across Different Rideable Types and Weekdays or Weekends: .....	18
5) Analysis of User Behavior and Performance Metrics Across Different Rideable Types and Months: 21	
6) A Comprehensive Analysis and Visualization of User Type, Rideable Type, and Quarter of the Year: .....	24
7) User Behavior and Demand Patterns During Rush Hour: .....	25
8) Top 10 popular routes: .....	27
9) Top 10 popular station: .....	28
10) Analysis and Visualization: Understanding Bike and Member Types in Divvy 2023 Data .....	29
<b>Future Recommendations</b>	<b>31</b>
<b>Conclusion</b>	<b>31</b>
<b>References:</b>	<b>32</b>

## Abstract:

This project revolves around analyzing data from the Divvy Chicago bike system using Snowflake for data processing and Power BI for visualization. With a focus on trip details like start/end times, locations, and distances traveled, it aims to understand user behavior and enhance the bike-sharing system. The objectives include data cleaning, seamless data loading into Snowflake, detailed analysis uncovering popular stations, routes, and peak usage times, and finally, using Power BI to present actionable insights through compelling visualizations.

## Introduction:

Since its debut in 2013, the Divvy Chicago bike system—a public bike-sharing program—has grown to include 800 stations and 6,000 bikes, servicing more than 85,000 users per day. Its success is attributed to improving air quality, reducing traffic congestion, and promoting bicycle use. The Divvy system generates a significant amount of data, which offers useful insights into user behavior and potential for system improvement.

## Literature Review:

**Why we chose snowflake over other new SQL Technologies like Amazon Redshift and Azure Synapse Analytics.**

### Snowflake

A cloud-based data warehousing platform called Snowflake is well-known for its distinct design that divides computing and storage resources. Users may grow storage and computation separately thanks to this separation, which offers cost-effectiveness and adaptability. The simplicity of Snowflake's operation stems from its completely managed service model, which abstracts away a lot of administrative duties including scaling, tweaking, and hardware supply. It is frequently commended for its support of semi-structured data (JSON, Parquet) and its effective handling of concurrency.

Attention is also paid to Snowflake's data sharing features, which provide easy data exchange without data migration across several Snowflake accounts.

### Pros:

- Because computation and storage are kept apart in Snowflake's design, customers may grow these resources separately thanks to its elasticity. The architecture facilitates the effective management of a variety of workloads.
- With Snowflake's user-friendly, fully managed service, customers can concentrate more on analytics and less on database maintenance and administrative duties.
- Its shared data design effectively handles several users and queries at once, guaranteeing steady performance even in times of heavy demand.

### Cons:

- Because Snowflake is shared, it might be difficult to handle more sophisticated queries or circumstances with a lot of concurrencies.
- When it comes to having direct control over hardware resources, some users may find Snowflake less flexible than self-managed options.
- Although Snowflake offers a flexible pricing structure, if utilization is raised without proper optimization, prices may rise.

### Amazon Redshift:

Redshift is a well-liked cloud-based data warehousing system that is renowned for its scalability and throughput. It makes use of massive parallel processing (MPP) architecture and columnar storage format. Redshift's ability to integrate with other AWS services, offering a whole ecosystem for analytics, storage, and data processing. It is renowned for its quick query speed, which makes it appropriate for businesses with significant AWS ecosystem investments, particularly for large-scale analytics workloads.

#### Pros:

- Redshift's columnar storage and massively parallel processing (MPP) design are responsible for its well-known excellent query performance, particularly for analytical applications.
- It combines well with other AWS services to create a whole ecosystem that makes data processing, storing, and analytics in the AWS environment easier.
- Redshift can handle growing workloads and volumes of data by expanding clusters to provide vertical scaling.

#### Cons:

- Redshift needs more manual administration than Snowflake when it comes to things like cluster scaling and performance optimization.
- In situations when demand is extremely strong, Redshift may have concurrency issues that affect query performance.
- Due to architectural changes, resizing clusters in Redshift may not be as easy or rapid as in Snowflake.

### Azure Synapse Analytics (formerly SQL Data Warehouse):

Microsoft's cloud-based analytics solution, Azure Synapse Analytics, is well-known for its combination of SQL Server and Azure ecosystem connectivity. It talks about how it can manage big data analytics as well as relational analytics, offering a single platform for a variety of workloads. It provides scalability, dividing storage and computation resources to scale separately, just like Snowflake and Redshift. According to several evaluations, it seamlessly integrates with other Azure services, like Power BI, to enable data analysis and visualization.

When comparing these systems, Snowflake's easy-to-use interface and distinct separation of computation and storage are frequently highlighted. Performance-wise, Redshift is well regarded, particularly in the AWS context, while Azure Synapse Analytics offers a robust toolkit inside the Microsoft ecosystem. Specific use cases, current cloud platform preferences, ease of management, query performance, and integration capabilities with other tools and services inside the organization's tech stack are all common factors to consider while selecting one of these solutions.

#### Pros:

- It offers a single platform that supports big data analytics as well as relational analytics, meeting a range of analytical requirements in one setting.
- Because of Synapse Analytics' close integration with other Azure services, working with Power BI and Azure Machine Learning is a breeze.
- Azure Synapse Analytics, like Snowflake, divides computation and storage to enable autonomous scalability and optimal resource use.

#### Cons:

- The benefits of using Azure Synapse Analytics as a stand-alone product may be limited by how strongly it is integrated into the Microsoft Azure ecosystem.
- Although it permits resources to be scaled independently, under conditions of great demand, scaling constraints may still apply.
- When compared to Snowflake's user-friendly approach, administering and optimizing Azure Synapse Analytics may be more difficult for those who are not familiar with the Azure environment.

**Let's dive deeper into the comparisons between Snowflake, Redshift, and Azure Synapse Analytics across various features:**

#### **Architecture:**

- Because of Snowflake's design, computation and storage can scale independently. It makes advantage of a shared data architecture with many clusters to offer concurrency and isolation without sacrificing speed.
- In contrast, Redshift employs a columnar storage format in conjunction with a massively parallel processing (MPP) architecture. With the use of parallelism, it is tailored for analytical workloads and allows for quicker query processing.
- Similar to MPP, Azure Synapse Analytics divides computing and storage for scalability, allowing customers to scale resources independently according to workload needs.

#### **Performance:**

- The architecture of Snowflake enables outstanding performance, particularly in concurrency. However, based on certain comparisons, Redshift's columnar storage and AWS environment optimizations may result in somewhat quicker query performance for particular use scenarios.
- Competitive performance is offered by Azure Synapse Analytics, especially to users who are currently employing the Microsoft Azure environment and taking use of its integrated services.

#### **Ease of Use and Management:**

- Snowflake is frequently complimented for its usability and simplicity. Because of its completely managed service, which minimizes maintenance, scaling, and administrative overhead, users without extensive technical experience in database management find it intriguing.
- Redshift needs additional human administration, such as cluster scaling and optimization. It could require more work in terms of administration than Snowflake.
- Even though Azure Synapse Analytics has robust management options, smooth integration and administration may need knowledge with the Azure environment.

#### **Integration and Ecosystem:**

- Snowflake offers agreements with several platforms and technologies and provides a wide range of integrations. It is well respected for its inherent connectors with widely used BI and ETL tools.
- Redshift is a popular option among current AWS users because of its close integration with the AWS ecosystem and use of services like S3 for storage.
- When combined with other Azure services, Azure Synapse Analytics provides easy access to Power BI and Azure Machine Learning, among other technologies.

#### **Pricing and Cost Model:**

- Because Snowflake can segregate these resources, its pricing model—which is based on computation and storage usage—offers flexibility and economic efficiency.
- Pricing for Redshift is determined by the instance type and use selected; to minimize expenses, diligent oversight and management are necessary.
- Similar pay-as-you-go pricing applies to Azure Synapse Analytics, although customers may be able to take advantage of cost savings opportunities inside the Azure network.

Selecting one of these databases typically comes down to a number of factors, including the organization's tech stack integration capabilities, performance requirements, preferred cloud platforms, ease of management, and unique business demands. Every platform has its advantages and may successfully serve a variety of use cases.

Finally, we can state that because of Snowflake's distinctive architecture—which distinguishes computing and storage resources—we chose it as the best option for the Data Visualization and Analytics on Divvy Chicago Bike project. This configuration guarantees speed and scalability that are unmatched, which is essential for managing the large and varied datasets that are part of the Divvy bike system. This configuration guarantees speed and scalability that are unmatched, which is essential for managing the large and varied datasets that are part of the Divvy bike system. Additionally, Snowflake's fully managed service eases administrative costs, freeing up the project team to focus on insights rather than database administration. Its ability to handle concurrent traffic fits very well with the project's requirement to examine popular routes and periods of high demand. Furthermore, Snowflake is the best option for producing insightful results because of its adaptability to handle a wide range of data types and its smooth interface with tools like Power BI. These features perfectly compliment the project's goal of efficiently analyzing and visualizing Divvy bike system data.

## Project Description:

In this project, data from the Divvy Chicago bike system will be analyzed and shown using Snowflake. Numerous pieces of information on bike usage are available through the Divvy system, such as trip start and finish times, locations, and lengths covered. This information may be used to determine places where the Divvy system needs to be enhanced and to learn how users are utilizing it.

This project analyzes and visualizes trip times, locations, and lengths traveled from the Divvy Chicago bike system using Snowflake. Among the goals are:

1. Clean and prepare Divvy data for Snowflake compatibility.
2. Load data into Snowflake via SnowSQL, supporting diverse formats.
3. Conduct data analysis, highlighting popular stations, routes, and peak usage times.
4. Communicate insights with Power BI's versatile charting capabilities.

## Data Collection:

The project's data came from Divvy's travel data, which is accessible to the public and covers the months of January through October in 2023. Because the datasets were acquired in CSV format, compatibility and convenience of use were guaranteed. Every dataset was carefully arranged, using UTF-8 encoding to improve cross-platform compatibility. The final dataset is a solid base for analysis while maintaining data integrity because of its significant sample size, uniqueness, completeness, relevance, and appropriate citation of the source.

Monthly datasets were gathered as part of the Divvy Chicago data collection procedure. The databases offer insightful information on popular stations, use trends, and important variables pertaining to ride lengths and distances. The main features of the data collection procedure are delineated in this paper, along with some first results.

### Data Attributes:

Ride Information  
 Ride Start and End Timestamps  
 Start and End Station Details  
 Ride Duration in Seconds  
 Trip Distance  
 Member Details  
 Member Type (Casual, Member)  
 Member/Casual Identification  
 Geospatial Information  
 Start and End Station Coordinates (Latitude, Longitude)

### Data Preparation:

To make the process of transferring data from local files to Snowflake easier, the team developed a unique warehouse and database within the Snowflake environment, which they named "Divvy1". The process of uploading data allowed by SnowSQL was meticulously executed to ensure the secure and prompt transfer of Divvy bike-sharing datasets.

- Data uploading was done using SnowSQL, the command-line interface for Snowflake. For secure, validated access, the necessary credentials and connection settings were configured.

```
PS C:\Users\Mdmuz> snowsql -a zk06771.ca-central-1.aws -u MUZAMMIL603
Password:
* SnowSQL * v1.2.30
Type SQL statements or !help
MUZAMMIL603#MUZA@(no database).(no schema)>|
```

- "Divvy1" is a Snowflake database that was made for efficient analysis and storage. Divvy bike-sharing datasets were meticulously uploaded and organized using SnowSQL.

```
MUZAMMIL603#MUZA@(no database).(no schema)>show DATABASES;
+-----+-----+-----+-----+-----+-----+-----+-----+
| created_on | name | is_default | is_current | origin | budget | owner |
| comment | options | retention_time | kind | | | |
+-----+-----+-----+-----+-----+-----+-----+
| 2023-11-10 00:34:48.806 -0800 | DIVVY | N | 1 | N | STANDARD | NULL | ACCOUNTADMIN |
| 2023-11-14 10:17:51.426 -0800 | DIVVY1 | N | 1 | N | STANDARD | NULL | SYSADMIN |
| 2023-11-10 00:32:50.834 -0800 | SNOWFLAKE | N | 0 | N | SNOWFLAKE,ACCOUNT_USAGE | NULL | ACCOUNTADMIN |
| 2023-11-10 00:32:50.161 -0800 | SNOWFLAKE_SAMPLE_DATA | N | 0 | N | SFSALESSHARED,SFC_SAMPLES_AWS_CA_CENTRAL_1.SAMPLE_DATA | NULL | ACCOUNTADMIN |
| Provided by Snowflake during account provisioning | | | | | IMPORTED DATABASE | NULL |
+-----+-----+-----+-----+-----+-----+-----+
4 Row(s) produced. Time Elapsed: 1.481s
MUZAMMIL603#MUZA@(no database).(no schema)>use DIVVY1;
+-----+
| status |
+-----+
| Statement executed successfully. |
+-----+
1 Row(s) produced. Time Elapsed: 0.161s
```

- During the data preparation phase, we began constructing the "January" table with detailed parameters for Divvy bike-sharing data. Then, we repeated this pattern for the months of February,

March, April, May, June, July, August, September, and October. The snapshot below shows the structure of the "January" table. This technique ensures consistent data categorization each month and offers a structure for methodical, consistent data analysis.

```
MUZAMMIL603#MUZA@DIVVY1.PUBLIC>CREATE OR REPLACE TABLE JANUARY (
    RIDE_ID VARCHAR(16777216),
    RIDEABLE_TYPE VARCHAR(16777216),
    STARTED_AT TIMESTAMP_NTZ,
    ENDED_AT TIMESTAMP_NTZ,
    START_STATION_NAME VARCHAR(16777216),
    START_STATION_ID VARCHAR(16777216),
    END_STATION_NAME VARCHAR(16777216),
    END_STATION_ID VARCHAR(16777216),
    START_LAT FLOAT,
    START_LNG FLOAT,
    END_LAT FLOAT,
    END_LNG FLOAT,
    MEMBER_CASUAL VARCHAR(16777216)
);
```

```
+-----+
| status |
+-----+
| Table JANUARY successfully created. |
+-----+
1 Row(s) produced. Time Elapsed: 1.511s
```

- During this procedure, the CSV file from the local path "C:/Users/mdmuz/OneDrive/Documents/Pro/202302-DIVVY-TRIPDATA.CSV" was loaded into the specified staging area within Snowflake using the Snowflake COPY command. By doing this, you can be sure the dataset is prepared for further integration and analysis.
- Every month after that, this process was repeated, guaranteeing a standardized and methodical approach to data entry, staging, and getting ready for in-depth examination in Snowflake.

```
MUZAMMIL603#MUZA@DIVVY1.PUBLIC>PUT file://C:/Users/mdmuz/OneDrive/Documents/Pro/202306-DIVVY-TRIPDATA.CSV @~/202306-DIVVY-TRIPDATA1;
```

source	target	source_size	target_size	source_compression	target_compression	status	message
202306-DIVVY-TRIPDATA.CSV	202306-DIVVY-TRIPDATA.CSV.gz	144356672	24759408	NONE	GZIP	UPLOADED	

```
1 Row(s) produced. Time Elapsed: 36.899s
```

```
MUZAMMIL603#MUZA@DIVVY1.PUBLIC>PUT file://C:/Users/mdmuz/OneDrive/Documents/Pro/202301-DIVVY-TRIPDATA.CSV @~/202301-DIVVY-TRIPDATA1;
```

source	target	source_size	target_size	source_compression	target_compression	status	message
202301-DIVVY-TRIPDATA.CSV	202301-DIVVY-TRIPDATA.CSV.gz	38451449	6554464	NONE	GZIP	UPLOADED	

```
1 Row(s) produced. Time Elapsed: 11.325s
```

- During the data integration phase, we successfully moved the staged Divvy bike-sharing trip data for January into the relevant table in the "Divvy1" database in Snowflake. This ensures that the dataset moves from the staging area to its designated storage location for further analysis without any problems. An illustration of the COPY INTO command used with the January dataset can be found below.
- This process was systematically repeated in the months that followed, adhering to a uniform protocol for transferring data from the staging area to the relevant tables.

```
MUZAMMIL603#MUZA@DIVVY1.PUBLIC>COPY INTO DIVVY1.PUBLIC.JANUARY
FROM @~/202301-DIVVY-TRIPDATA1
FILE_FORMAT = (TYPE = CSV COMPRESSION = GZIP FIELD_OPTIONALLY_ENCLOSED_BY='"' SKIP_HEADER = 1)
ON_ERROR = 'CONTINUE';
```

file	status	rows_parsed	rows_loaded	error_limit	errors_seen	first_error	first_error_line	fi
202301-DIVVY-TRIPDATA1/202301-DIVVY-TRIPDATA.CSV.gz	LOADED	190301	190301	190301	0	NULL		NULL

```
1 Row(s) produced. Time Elapsed: 2.574s
```



- During the data consolidation phase, we aggregated the trip data for the year 2023 from each of the individual monthly tables (JANUARY, FEBRUARY, MARCH, APRIL, MAY, JUNE, JULY, AUGUST, SEPTEMBER, and OCTOBER). This resulted in the creation of a comprehensive table called "Full\_data". This integrated dataset provides a thorough overview of all Divvy bike-sharing activities throughout the specified time frame.
- There are 13 columns/fields and 7333300 rows in the merged table.

```
MUZAMMIL603#MUZA@DIVVY1.PUBLIC>CREATE OR REPLACE TABLE Full_data AS
SELECT * FROM JANUARY
UNION ALL
SELECT * FROM FEBRUARY
UNION ALL
SELECT * FROM MARCH
UNION ALL
SELECT * FROM APRIL
UNION ALL
SELECT * FROM MAY
UNION ALL
SELECT * FROM JUNE
UNION ALL
SELECT * FROM JULY
UNION ALL
SELECT * FROM AUGUST
UNION ALL
SELECT * FROM SEPTEMBER
UNION ALL
SELECT * FROM OCTOBER;

+-----+
| status |
+-----+
| Table FULL_DATA successfully created. |
+-----+
1 Row(s) produced. Time Elapsed: 24.680s
```

## Data Cleaning:

1)The dataset, representing Divvy bike-sharing trip information for the entire year of 2023, has been formally rebranded as "DIVVY\_2023\_DATA."

Rows: 7,333,300

Columns/Fields: 13

AccountADMIN • MUZA Share

DIVVY1PUBLIC Settings

Code Versions

```

1 select *
2 from divvy_2023_data;

```

Results Chart

	RIDE_ID	RIDEABLE_TYPE	STARTED_AT	ENDED_AT	START_STATION_NAME
1	9340B064F0AEE130	electric_bike	2023-07-23 20:06:14.000	2023-07-23 20:22:44.000	Kentzie Ave & 110th St
2	D146DE3CE0D8AF8	classic_bike	2023-07-23 17:05:07.000	2023-07-23 17:16:37.000	Western Ave & Watton St
3	DF41B318B95A25E	classic_bike	2023-07-23 10:14:53.000	2023-07-23 10:24:29.000	Western Ave & Watton St
4	9624A293749EF703	electric_bike	2023-07-21 08:27:44.000	2023-07-21 08:32:40.000	Racine Ave & Randolph
5	2F6B6A44CD84C99A	classic_bike	2023-07-08 15:46:42.000	2023-07-08 15:56:08.000	Clark St & Leland Ave
6	9AAE973E6B941A9C	classic_bike	2023-07-10 08:44:47.000	2023-07-10 08:49:41.000	Racine Ave & Randolph
7	E366E997DA1592B	classic_bike	2023-07-25 14:30:44.000	2023-07-25 14:37:45.000	Clark St & Leland Ave
8	1B83E73851E6C2C1	classic_bike	2023-07-07 10:11:53.000	2023-07-07 10:17:55.000	Clark St & Leland Ave
9	DA1E1D086E6566E	electric_bike	2023-07-04 21:57:27.000	2023-07-04 22:06:27.000	Clark St & Leland Ave
10	398FA473A704CAB5	classic_bike	2023-07-29 10:51:17.000	2023-07-29 11:03:13.000	Warren Park East

Partial results displayed  
Only 10,000 rows of the results are displayed. Please download the results for all of the rows.

Query Details

Query duration 9.7s

Rows 73M

Query ID 01b0887d-3200-1598-...

2) A modification was made to the 'rideable\_type' column in the Divvy bike-sharing dataset during a recent data refining procedure. Three different bike kinds were initially included in the dataset: "classic," "docked," and "electric." After further investigation, it was discovered that the term "docked bike" is obsolete and is really interchangeable with "classic bike." As a result, in order to improve data homogeneity and conform to existing standards, all occurrences of "docked bike" were methodically changed to "classic bike." A total of 119,998 rows were affected by this update.

AccountADMIN • MUZA Share

DIVVY1PUBLIC Settings

Code Versions

```

1 UPDATE Divvy_2023_data
2 SET RIDEABLE_TYPE = 'classic_bike'
3 WHERE RIDEABLE_TYPE = 'docked_bike';
4
5
6
7
8
9

```

Results Chart

	number of rows updated	number of multi-joined rows updated
1	119998	0

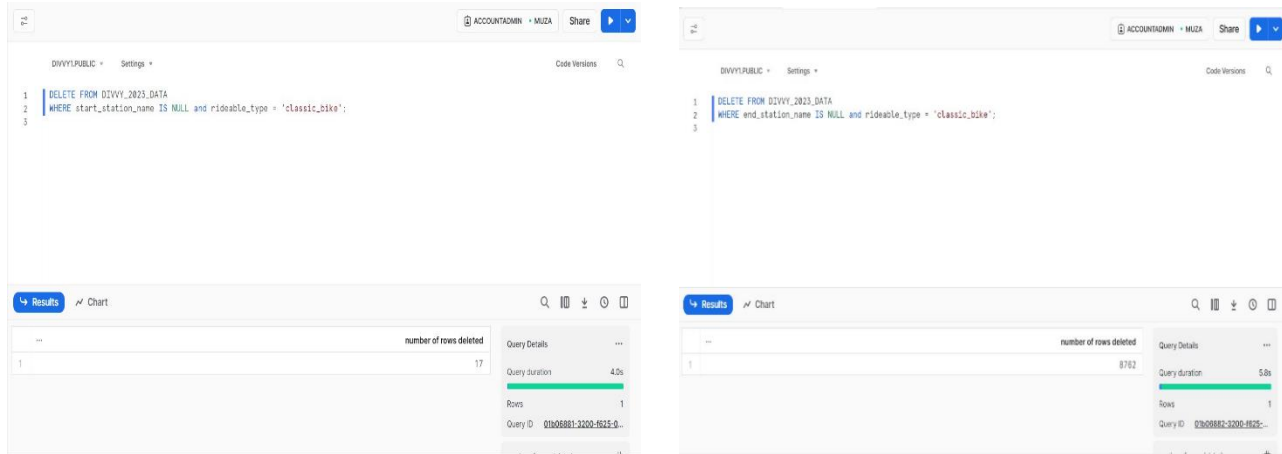
Query Details

Query duration 6.6s

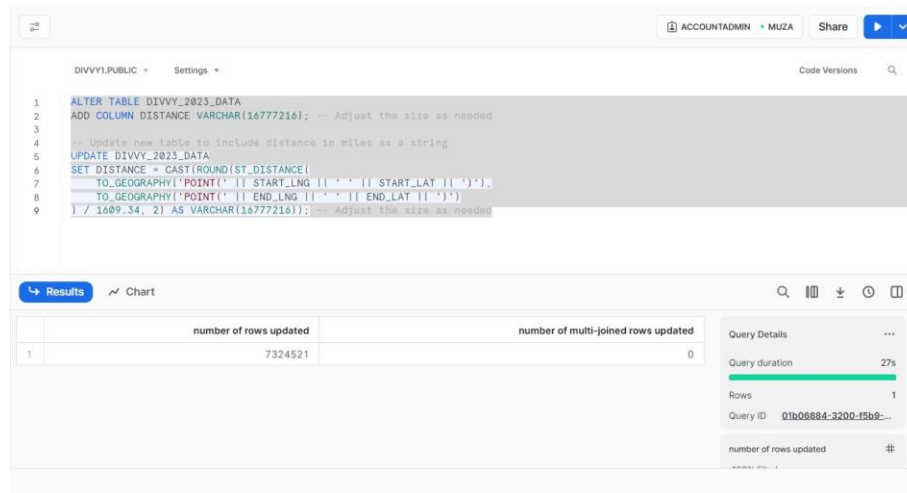
Rows 1

Query ID 01b0887d-3200-1631-0...

3) Eliminate entries from the "DIVVY\_2023\_DATA" database if the "rideable\_type" is explicitly classified as "classic\_bike" and the "start\_station\_name" and "end\_station\_name" are absent.



4) The 'DISTANCE' column was added as a new feature to provide distance data to the Divvy bike-sharing dataset. Distances can be represented as strings in this column, which is specified as VARCHAR (16777216).



5) Removed the outliers from the table entries where timestampdiff < 30 or timestampdiff > 86400.

The screenshot shows a SQL query execution interface. The query is:

```

1 DELETE FROM DIVVY_2023_DATA
2 WHERE timestampdiff(seconds,started_at,ended_at) < 30 or timestampdiff (seconds,started_at,ended_at) > 86400;
3

```

The results table shows 1 row with the column "number of rows deleted" and the value 132764. The Query Details panel on the right shows a query duration of 6.3s, 1 row, and a query ID of 01b0688d-3200-f5b0-...

6)Removed the outliers from the table entries where  $\text{timestampdiff} < 45$  and  $\text{distance} = 0$ .

The screenshot shows a SQL query execution interface. The query is:

```

1 DELETE FROM DIVVY_2023_DATA
2 WHERE timestampdiff(seconds,started_at,ended_at) < 45 and distance = 0 ;
3

```

The results table shows 1 row with the column "number of rows deleted" and the value 28095. The Query Details panel on the right shows a query duration of 6.5s, 1 row, and a query ID of 01b0688e-3200-f632-...

7)Removed the outliers from the table entries where  $\text{timestampdiff} < 80000$  and  $\text{distance} = 0$ .

The screenshot shows a SQL query execution interface. The query is:

```

1 DELETE FROM DIVVY_2023_DATA
2 WHERE timestampdiff(seconds,started_at,ended_at) > 80000 and distance = 0 ;
3

```

The results table shows 1 row with the column "number of rows deleted" and the value 127. The Query Details panel on the right shows a query duration of 6.2s, 1 row, and a query ID of 01b0688f-3200-f632-0-...

8) For analysis, new columns were created and added to the table, including RIDE\_LENGTH, DAY\_OF\_WEEK, and DATE.

```

1  DIVVY1.PUBLIC > Settings >
2  ...
3  ADD COLUMN ride_length FLOAT;
4
5  -- Add day_of_week column
6  ALTER TABLE DIVVY_2023_DATA
7  ADD COLUMN day_of_week VARCHAR(16777216);
8
9  -- Add date column
10 ALTER TABLE DIVVY_2023_DATA
11 ADD COLUMN date DATE;
12
13 -- Update the new columns with calculated values
14 UPDATE DIVVY_2023_DATA
15 SET ride_length = DATEDIFF(SECOND, STARTED_AT, ENDED_AT),
16    day_of_week = DAYNAME(STARTED_AT),
17    date = DATE(STARTED_AT);
18
19 Results Chart
20
21 number of rows updated  number of multi-joined rows updated
22 1 7163535 0
23
24 Query Details
25 Query duration 6.1s
26 Rows 1
27 Query ID 01b0689d-3200-f625-...

```

## DATA ANALYSIS AND VISUALIZATION:

### 1) Number of rides under time free cost limitation (45 minutes):

#### Analysis

The rides have been divided into categories according to rider type (MEMBER\_CASUAL) and cost, which is determined by the length of the ride. The number of rides that fit within each category is then counted.

```

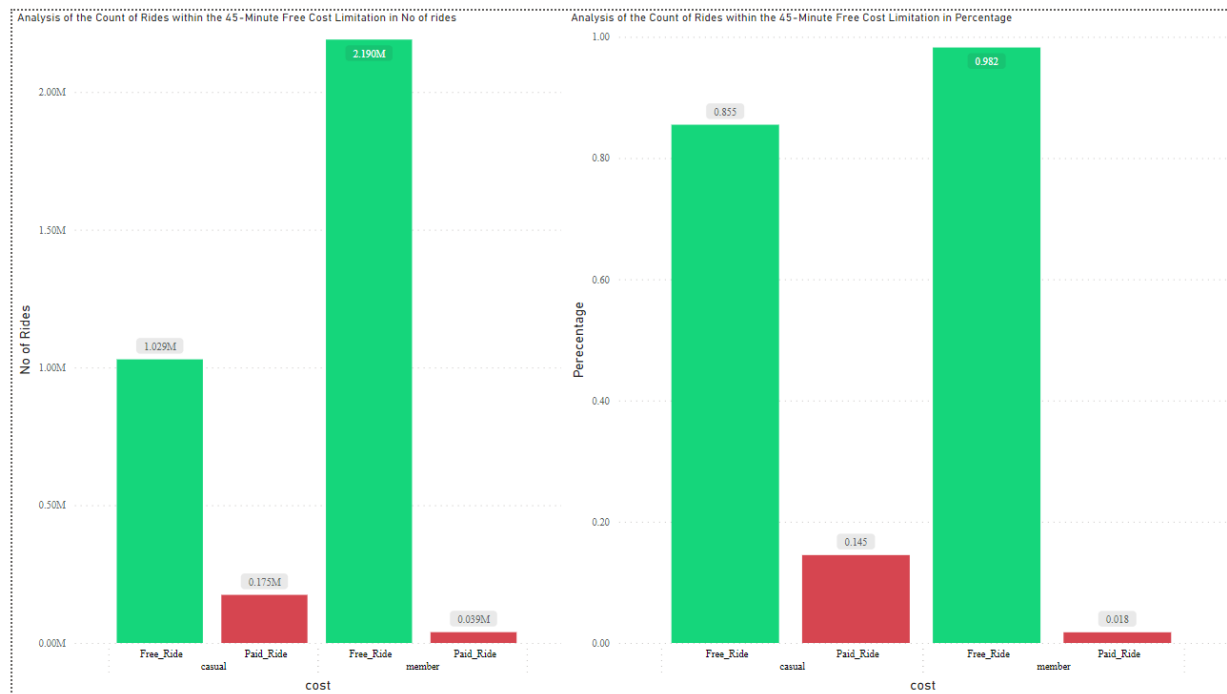
1  WITH CTE AS (
2  SELECT
3  MEMBER_CASUAL,
4  CASE WHEN ride_length / 60 < 45 THEN 'Free_Ride' ELSE 'Paid_Ride' END AS cost,
5  COUNT(*) AS rides
6  FROM
7  divvy_2023_data
8  WHERE RIDEABLE_TYPE = 'classic_bike'
9  GROUP BY
10 MEMBER_CASUAL, cost
11 )
12
13 SELECT
14 MEMBER_CASUAL,
15 cost,
16 rides,
17 (rides * 100.0) / SUM(rides) OVER (PARTITION BY MEMBER_CASUAL) AS Percentage
18 FROM
19 CTE;
20
21 Results Chart
22
23 MEMBER_CASUAL  COST  RIDES  PERCENTAGE
24 1 casual  Free_Ride  1029394  85.501036
25 2 member  Free_Ride  2190078  98.242200
26 3 casual  Paid_Ride  174561  14.496964
27 4 member  Paid_Ride  39186  1.757800
28
29 Query Details
30 Query duration 260ms
31 Rows 4
32 Query ID 01b0689d-3200-f625-...
33
34 MEMBER_CASUAL
35 casual 2
36 member 2

```

#### Visualization

98.2% of member trips in Chicago stick to the 45-minute time restriction, demonstrating effective utilization of unlimited rides, according to an analysis of Divvy's pricing schemes. 87% of trips made by casual users,

who pay \$3 per trip or have a daily pass, occur inside the 45-minute timeframe, indicating a higher chance of going over the limit. There are no additional fees for members, however there may be fees for non-members who utilize the time limit beyond. Strategies include casual users thinking about daily passes and raising member incentive awareness. After the allotted period has passed, more research may provide insights into user behavior and price changes.



## 2) User Behavior and Performance Metrics Across Different Rideable Types, User Categories, and Days of the Week:

### Analysis

From the divvy\_2023\_data database, we have produced a summary of the ride data, organizing the findings by MEMBER\_CASUAL and RIDEABLE\_TYPE. For every combination of rider type and rideable vehicle type, we computed the total number of rides, total distance traveled, total ride length, average distance per ride, average ride time in minutes, and average speed in mph.

DIVVY1.PUBLIC

Settings

Code Versions

```

1 SELECT
2   MEMBER_CASUAL, RIDEABLE_TYPE,
3   COUNT(*) AS total_rides,
4   SUM(DISTANCE) AS total_distance,
5   SUM(ride_length) AS total_ride_length,
6   AVG(DISTANCE) AS average_distance,
7   AVG(ride_length) / 60 AS average_ride_time_mins,
8   AVG(DISTANCE / (ride_length / 5680)) AS average_speed_mph
9 FROM
10  divvy_2023_data
11 GROUP BY
12  MEMBER_CASUAL, RIDEABLE_TYPE;

```

Results

Chart

Results

Chart

Results

Chart

Results

Chart

	MEMBER_CASUAL	RIDEABLE_TYPE	TOTAL RIDES	TOTAL_DISTANCE	TOTAL_RIDE_LENGTH	AVERAGE_DISTANCE	AVERAGE_RIDE_TIME_MINS	AVERAGE_SPEED
1	member	classic_bike	2229264	2690819.74	1769607941	1.207044002	13.230135903	6.2827
2	casual	classic_bike	1203955	1591234.32	2063441282	1.321810922	28.564761723	4.6832
3	member	electric_bike	2305137	3382390.94	1592439031	1.467327512	11.513697675	8.9470
4	casual	electric_bike	1425179	1968436.38	1278877166	1.381185367	14.955749021	6.9138

Query Details

Query duration

912ms

Rows

4

Query ID

01b0d23d-3200-fb76c...

MEMBER\_CASUAL

member

2

casual

2

RIDEABLE\_TYPE

### Most Expensive Nodes (2 of 3)

TableScan [2]	84.5%
Aggregate [1]	12.7%

### Profile Overview (Finished)

Total Execution Time	(746ms) 100.0%
Processing	15.5%
Remote Disk I/O	81.7%
Initialization	2.8%

### Statistics

Scan progress	100.00%
Bytes scanned	24.81MB
Percentage scanned from cache	0.00%
Partitions scanned	23
Partitions total	23

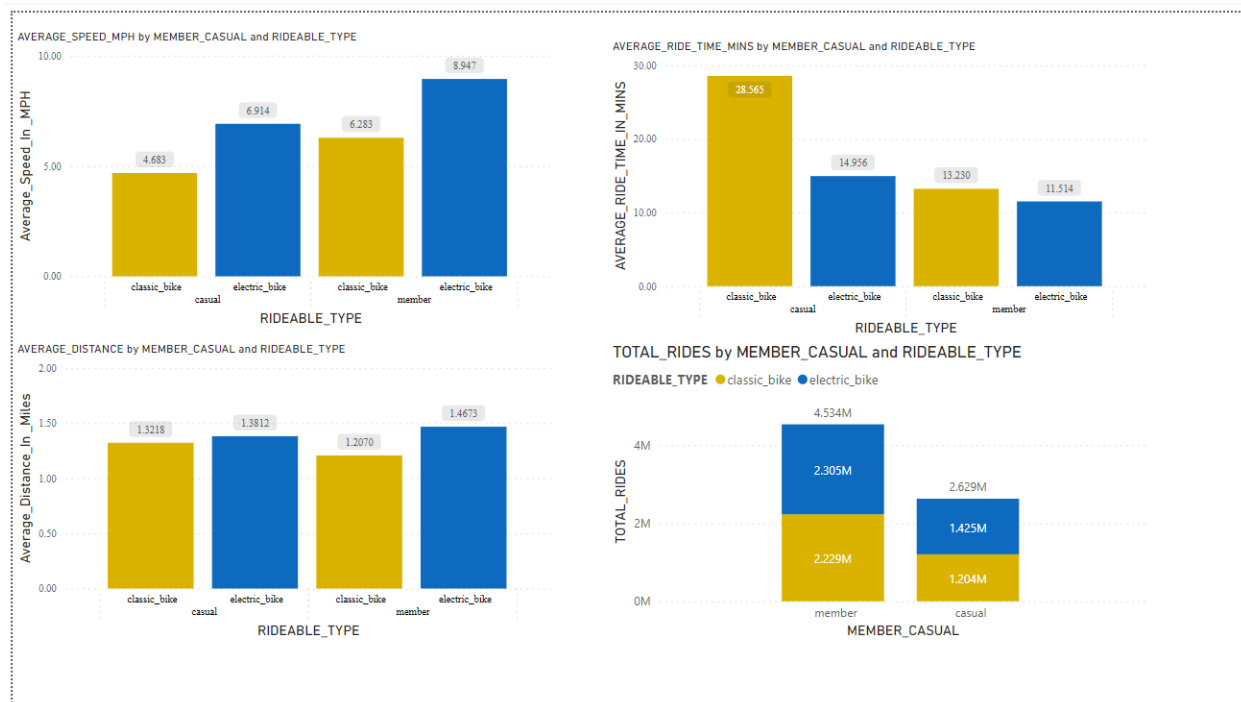
## Visualization:

- Both members and casual users will find that electric bikes are speedier than traditional bikes, according to this chart. For members, the average speed on an electric bike is 12.5 mph, while for casual riders it is 12.3 mph. For members, the average speed on classic motorcycles is 10.8 mph, while for casual riders it is 9.8 mph.
- Regardless of the type of bike, this data demonstrates that casual riders often ride for greater periods of time than members. For casual users, an electric bike ride typically takes 25.2 minutes, while for members, it takes 17.9 minutes. For casual riders, the average ride time on a classic cycle is 23.8 minutes, whereas for members it is 13.1 minutes.
- Regardless of the type of user, this graphic indicates that electric bikes are utilized for longer journeys than classic cycles. For casual riders, the average distance traveled on an electric bike is 2.62 miles, while for members, it is 2.39 miles. For casual riders, the average distance traveled on classic bikes is 2.30 miles; for members, it is 1.89 miles.

- 4) Regardless of the kind of bike, this graph demonstrates that members ride more than casual riders in general. Of all rides, 62% are taken by members, while 38% are taken by casual users. Classic bikes make up 47% of rides, whilst electric bikes make up 53% of rides.

## Overall Analysis

According to the research, electric bikes had quicker average speeds than classic bikes, and both club members and casual riders report riding faster on electric bikes. Regardless of the style of bike, casual riders often ride for longer than members. Regardless of the sort of user, electric bikes are constantly utilized for greater distances than conventional cycles. Ride frequency is dominated by members (62% of all rides), and electric bikes are more common (53% of all rides), indicating preferences for the length and kind of bike.



## 3) Analysis of User Behavior and Performance Metrics Across Different Rideable Types, User Categories, and Days of the Week:

### Analysis:

Conducted a ride statistics study based on rideable categories, member and casual user types, and different days of the week. The data were organized in an organized fashion. It provides an in-depth analysis of ride stats, broken down by rideable kinds, user categories, and days of the week, offering insights into how these factors affect patterns of bike utilization. In order to facilitate better analysis and understanding, the findings are presented in a logical weekly sequence when ordered according to the days of the week.



DatabasesWorksheets

Pinned

No pinned objects

Search objects

BIKE\_TYPE\_ANALYSIS

DRIVYDATA\_2023

DRIVY\_2023

DRIVY\_2023\_DATA

FEB

FEBRUARY

JANUARY

JULY

JUNE

MARCH

MAY

MEMBER\_CASUAL\_STATS...

DRIVY1.PUBLIC

Settings

Code Versions

1SELECT

2DAY\_OF\_WEEK,

3MEMBER\_CASUAL,

4RIDEABLE\_TYPE,

5COUNT(\*) AS total\_rides,

6AVG(DISTANCE) AS average\_distance,

7AVG(ride\_length) / 60 AS average\_ride\_time\_mins,

8AVG(DISTANCE / (ride\_length / 3600)) AS average\_speed\_mph

9FROM

10drivy\_2023\_data

11GROUP BY

12DAY\_OF\_WEEK, MEMBER\_CASUAL,

13RIDEABLE\_TYPE

14ORDER BY

15CASE

16WHEN DAY\_OF\_WEEK = 'Mon' THEN 1

17WHEN DAY\_OF\_WEEK = 'Tue' THEN 2

18WHEN DAY\_OF\_WEEK = 'Wed' THEN 3

19WHEN DAY\_OF\_WEEK = 'Thu' THEN 4

20WHEN DAY\_OF\_WEEK = 'Fri' THEN 5

21WHEN DAY\_OF\_WEEK = 'Sat' THEN 6

22WHEN DAY\_OF\_WEEK = 'Sun' THEN 7

23END,

24MEMBER\_CASUAL;

MAY884.8K Rows

ResultsChart

DAY\_OF\_WEEK

MEMBER\_CASUAL

RIDEABLE\_TYPE

TOTAL RIDES

AVERAGE\_DISTANCE

AVERAGE\_RIDE\_TIME\_MINS

AVERAGE\_SPEED\_MPH

20	Fri	member	electric_bike	348197	1.445936783	11.554152869	8.38282152
21	Sat	casual	electric_bike	262408	1.458311065	17.136799818	6.429478681
22	Sat	casual	classic_bike	262803	1.384244355	30.924586033	4.274316258
23	Sat	member	electric_bike	295406	1.529303535	12.865553904	8.100780476
24	Sat	member	classic_bike	286512	1.286569533	14.876512444	6.008521862
25	Sun	casual	electric_bike	209918	1.422259073	17.192542088	6.454548152
26	Sun	casual	classic_bike	210177	1.350644708	32.079260417	4.189218046

Query Details

Query duration601ms

Rows28

Query ID67b0a0d8-3200-1586...

DAY\_OF\_WEEK

Mon4

Tue4

### Most Expensive Nodes (2 of 4)

Aggregate [2]	63.2%
TableScan [3]	21.1%

### Statistics

Scan progress	100.00%
Bytes scanned	26.67MB
Percentage scanned from cache	100.00%
Partitions scanned	23
Partitions total	23

### Profile Overview (Finished)

Total Execution Time (508ms)	100.0%
Processing	73.7%
Synchronization	10.5%
Initialization	15.8%

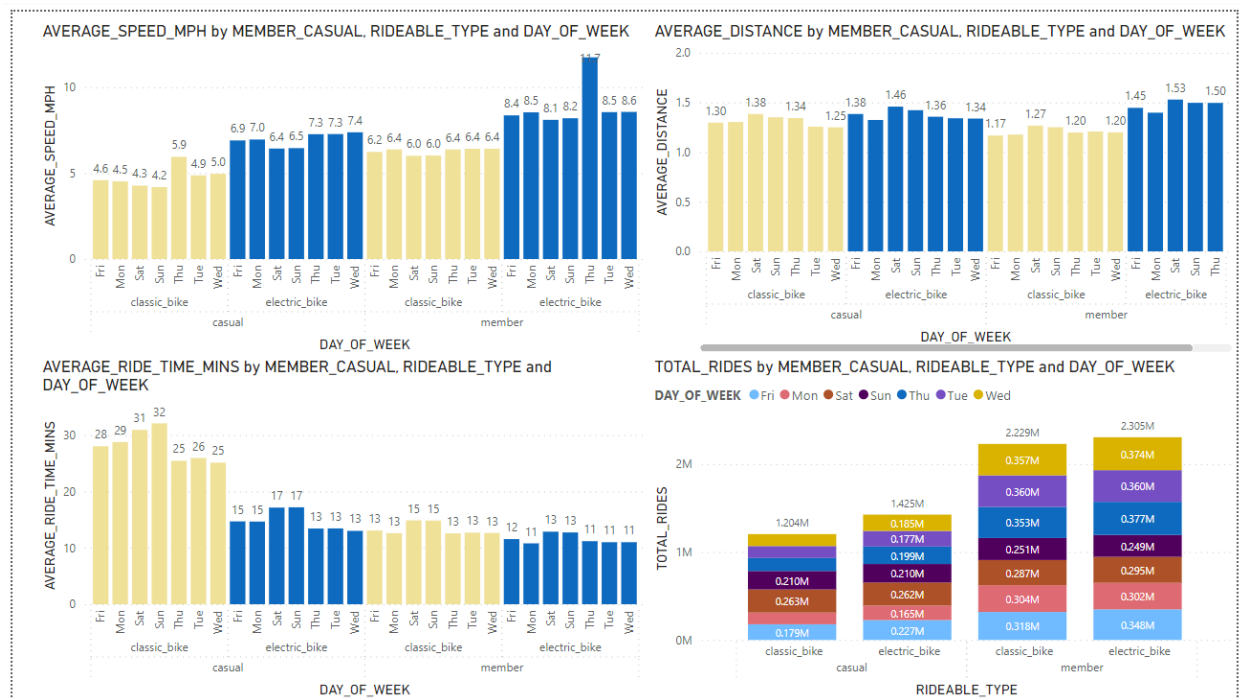
## Visualization:

- Members average 8.1 mph, while casual cyclists average 6.9 mph. Wednesdays see the highest speed of 8.5 mph for members and 7.0 mph for non-members. On Sundays, bikers average the slowest speeds of 5.9 mph for casual riders and 6.2 mph for members.
- At 1.5 miles on average, members travel a greater distance than casual cyclists (1.3 miles). Members average 1.53 miles on Saturdays, while casual cyclists average 1.50 miles. Tuesdays are the shortest rides, at around 1.20 miles for casual riders and 1.25 miles for members. At 1.38 kilometers on average, electric bikes outperform traditional cycles.
- Compared to traditional bikes, electric bikes provide longer ride periods. Electric bikes take 20 minutes on average, whereas traditional cycles take 17 minutes. On weekends, the durations are greater, taking around 19 minutes for traditional bikes and 22 minutes for electric cycles. During the week, the average riding duration is 16 minutes for traditional bikes and 18 minutes for electric cycles.

- 4) Weekend traffic is highest on Saturdays, and lowest on weekdays, particularly on Wednesdays. Regular riders log higher mileage, perhaps because of the benefits of their subscription. Longer distance riders prefer electric motorcycles over conventional bikes.

## Overall Analysis

Members of Divvy often ride faster and further, perhaps because of the advantages of having a membership. Sundays are slowest, and Wednesdays are the fastest. Peak mileage and ridership occur on Saturdays. Longer distance riders choose electric bikes, which emphasize user-friendly choices. Weekend rides typically last longer, which reflects a variety of riding styles. These observations help with well-informed decision-making for Divvy's bike-sharing service optimization.



Analysis of ride statistics based on weekdays and weekends, populating a table named `weekdays_weekends_analysis`.

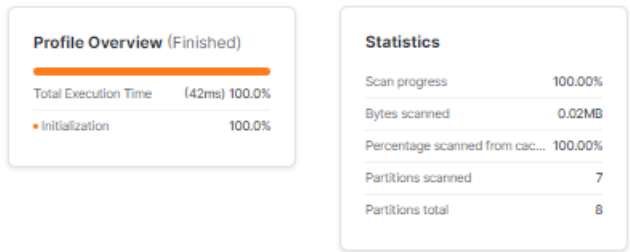
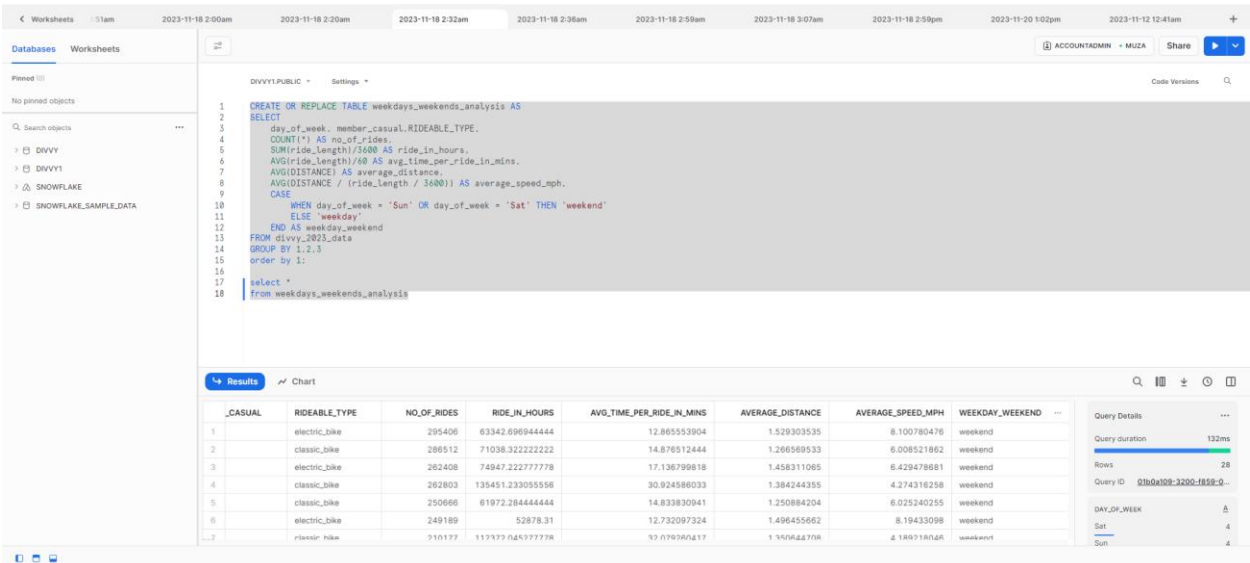
## 4) User Behavior and Performance Metrics Across Different Rideable Types and Weekdays or Weekends:

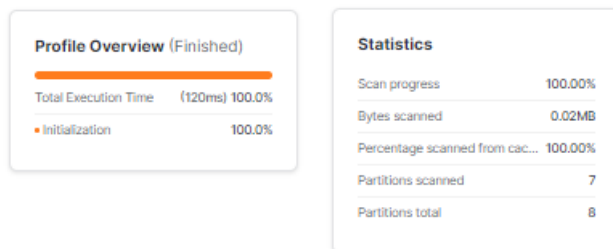
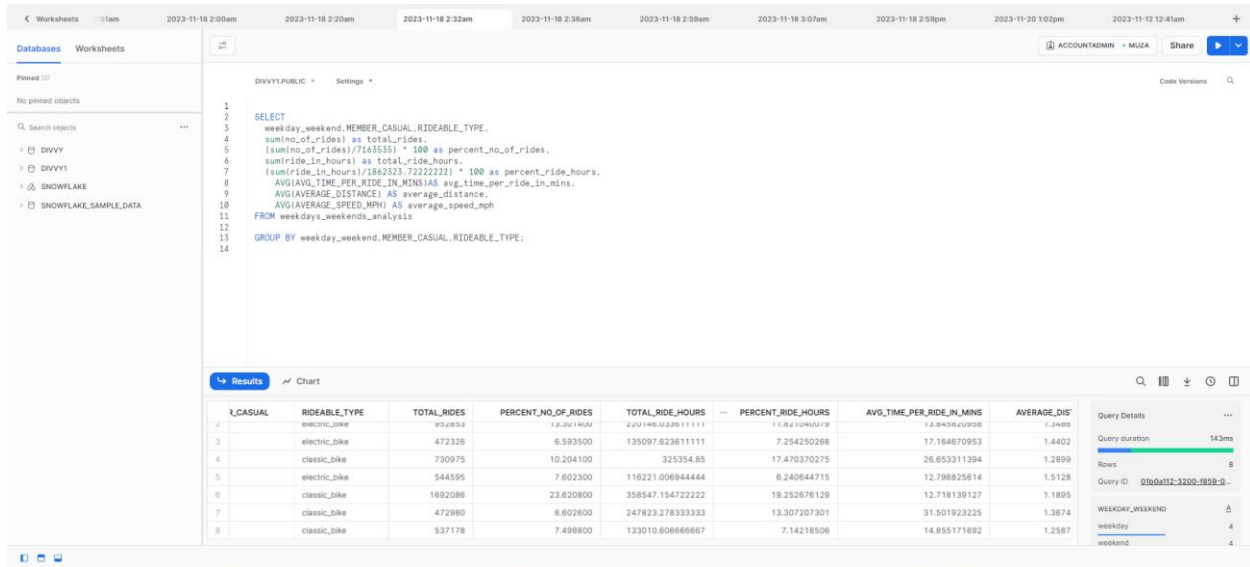
### Analysis:

We performed a thorough analysis on the ride data from Divvy's dataset, classifying rides according to rideable type, user type (casual or member), and day of the week. It calculates metrics like the total number of rides, average riding time, average distance traveled, average speed, and, depending on the 'day\_of\_week' column, classifies days as either weekends or weekdays.

This study can shed light on how different user kinds and rideable alternatives inside Divvy's bike-sharing system use the system, as well as how long their rides last and how far they travel on weekdays vs weekends.

After examining user activity over several weeks, we decided to do a binary analysis comparing weekend and weekday bike usage patterns because of the significant differences in these patterns. Interestingly, casual riders schedule their travels for the weekends, indicating a strong predilection for the weekends. On the other hand, weekend bike utilization by member riders is somewhat higher than that of other weekdays, but it stays the same. This discrepancy may be explained by casual riders using bikes for recreational purposes, whereas members generally utilize the service for everyday commuting (e.g., traveling to and from work or school).



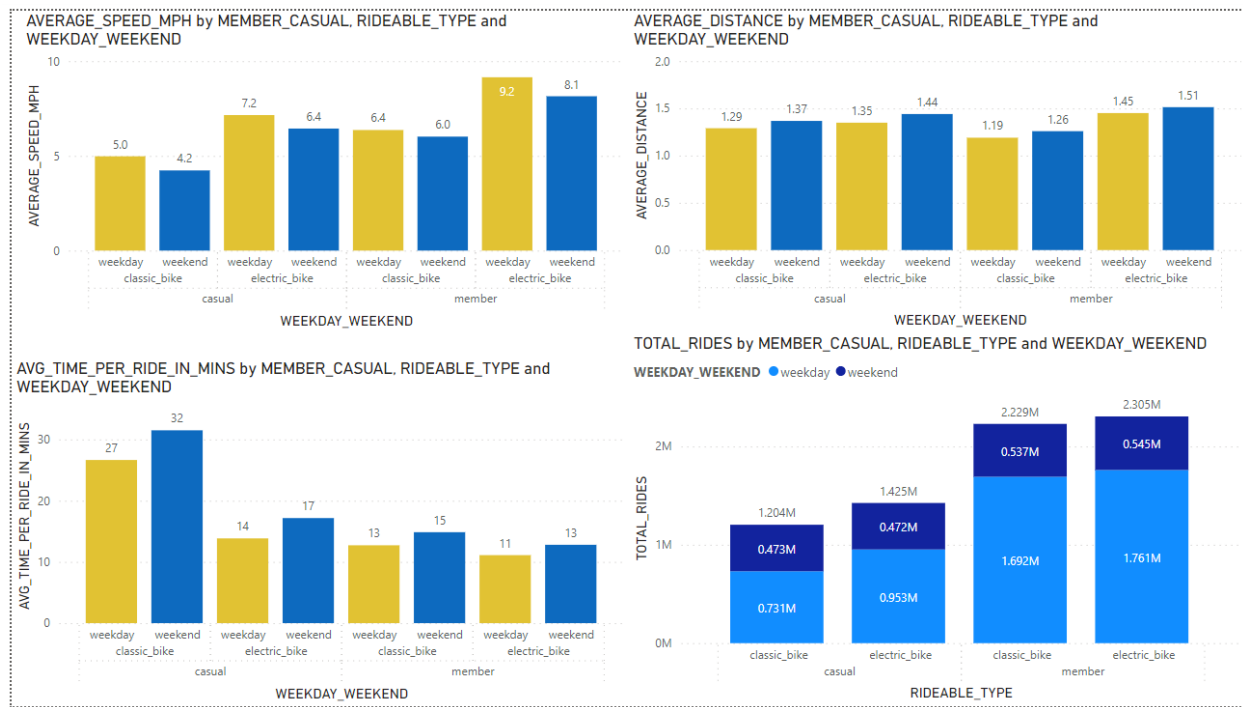


## Visualization:

- 1) On weekends, casual members ride more quickly, averaging 60 mph on traditional bikes and 64 mph on electric bikes, according to the graph. During the week, classic motorcycles can go at 50 mph, while electric bikes can reach 42 mph. On weekends, members ride at a slower pace—averaging 64 mph on classic bikes and 60 mph on electric bikes—than they do during the week, when they reach their highest speeds of 64 mph on electric bikes and 72 mph on classic bikes.
- 2) The average distance traveled by members is 1.29 miles on weekdays and 1.37 miles on weekends, whereas the average distance traveled by casual riders is 1.1 miles and 1.35 miles, respectively. When it comes to lengthier trips, electric bikes dominate: on weekdays, they average 1.26 miles, while on weekends, they reach 1.45 miles, which is more than traditional bikes' 1.19 miles.
- 3) Members ride for an average of 27 minutes on weekdays and 32 minutes on weekends, while casual riders typically cycle for 20 minutes and 27 minutes, respectively. With typical ride lengths of 24 minutes on weekdays and 30 minutes on weekends, electric bikes are preferred for longer journeys compared to traditional cycles, which have travel times of 20 minutes and 27 minutes, on average.
- 4) Weekday rides are higher on the graph than weekend rides. Classic bikes are the second most popular rideable kind after electric bikes. Compared to casual cyclists, members ride more often.

## Overall Analysis

All things considered; the study reveals clear trends in Divvy bike-sharing usage. On weekdays, members ride more quickly, with electric bikes predominating. Weekday rides are preferred over longer distances and times when members are involved. Weekend ridership is less than that of weekday passengers, highlighting the widespread use of electric bikes and the presence of member riders.



## 5) Analysis of User Behavior and Performance Metrics Across Different Rideable Types and Months:

### Analysis:

We have examined the `divvy_2023_data` table for Divvy's bike ride data, dividing the findings into categories such as `MEMBER_CASUAL` (which indicates if the rider is a member or a casual user), `RIDEABLE` type, and `monthname(started_at)`, which indicates the month the ride began. Important data are computed for each month, separately for casual and member cyclists, including the total number of rides, average distance traveled each ride, average ride time, and average speed. Additionally, looks at the Divvy bike-sharing system's ride trends and performance indicators for both user categories over the course of many months.

DIVVY.PUBLIC

Settings

Code Versions

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
SELECT
  MEMBER_CASUAL,
  RIDEABLE_TYPE,
  monthname(started_at) as month_name,
  COUNT(*) AS total_rides,
  AVG(DISTANCE) AS average_distance,
  AVG(ride_length) / 60 AS average_ride_time_mins,
  AVG(DISTANCE / (ride_length / 3600)) AS average_speed_mph
FROM divvy_2023_data
GROUP BY
  MEMBER_CASUAL,
  RIDEABLE_TYPE, month_name

```

Results

Chart

MEMBER\_CASUAL

RIDEABLE\_TYPE

MONTH\_NAME

TOTAL RIDES

AVERAGE\_DISTANCE

AVERAGE\_RIDE\_TIME\_MINS

AVERAGE\_SPEED\_MPH

casual	electric_bike	Aug	142590	1.456024265	15.321485144	6.935986629
casual	classic_bike	May	207702	1.349950025	30.225254772	4.414318236
member	electric_bike	Jan	141778	1.232491783	9.911090108	8.555616621
member	electric_bike	Apr	303818	1.378056534	10.842413331	8.428711867
member	classic_bike	Oct	182795	1.139101343	12.440784577	6.330240463
casual	electric_bike	Jun	329674	1.447583795	16.002677392	6.720206411
casual	classic_bike	Jan	30802	0.980810337	19.608199684	5.314044944
casual	electric_bike	May	248736	1.400977181	15.703463914	6.769398786
member	electric_bike	May	373718	1.495817809	11.865637994	8.403455267
member	classic_bike	Sep	212194	1.21524897	13.417399801	6.19605997
casual	classic_bike	Dec	61614	1.346083640	30.044403031	4.885650604

Query Details

Query duration

874ms

Rows

40

Query ID

07b0bd5e-3200-f984-...

MEMBER\_CASUAL

casual

20

member

20

RIDEABLE\_TYPE

classic\_bike

20

Most Expensive Nodes (2 of ...)

Aggregate [2]

81.5%

TableScan [3]

3.7%

Profile Overview (Finished)

Total Execution Time (541ms) 100.0%

Processing

85.2%

Initialization

14.8%

Statistics

Scan progress

100.00%

Bytes scanned

70.32MB

Percentage scanned from cac...

100.00%

Partitions scanned

23

Partitions total

23

## Visualization:

1) On traditional bikes, casual riders reach their greatest speeds in July (8.58 mph) and August (5.31 mph) while riding electric bikes. Members achieved their highest speeds on traditional cycles in August (8.51 mph) and on electric bikes in July (5.13 mph). In general, non-member cyclists ride faster than members, while vintage bikes travel faster than electric bikes.

2) January (30.01 minutes) and April (19.61 minutes) are the record riding times for classic bikes and electric cycles, respectively. All things considered, classic bikes have longer average ride times than electric bikes, which may be related to their slower and less efficient design.

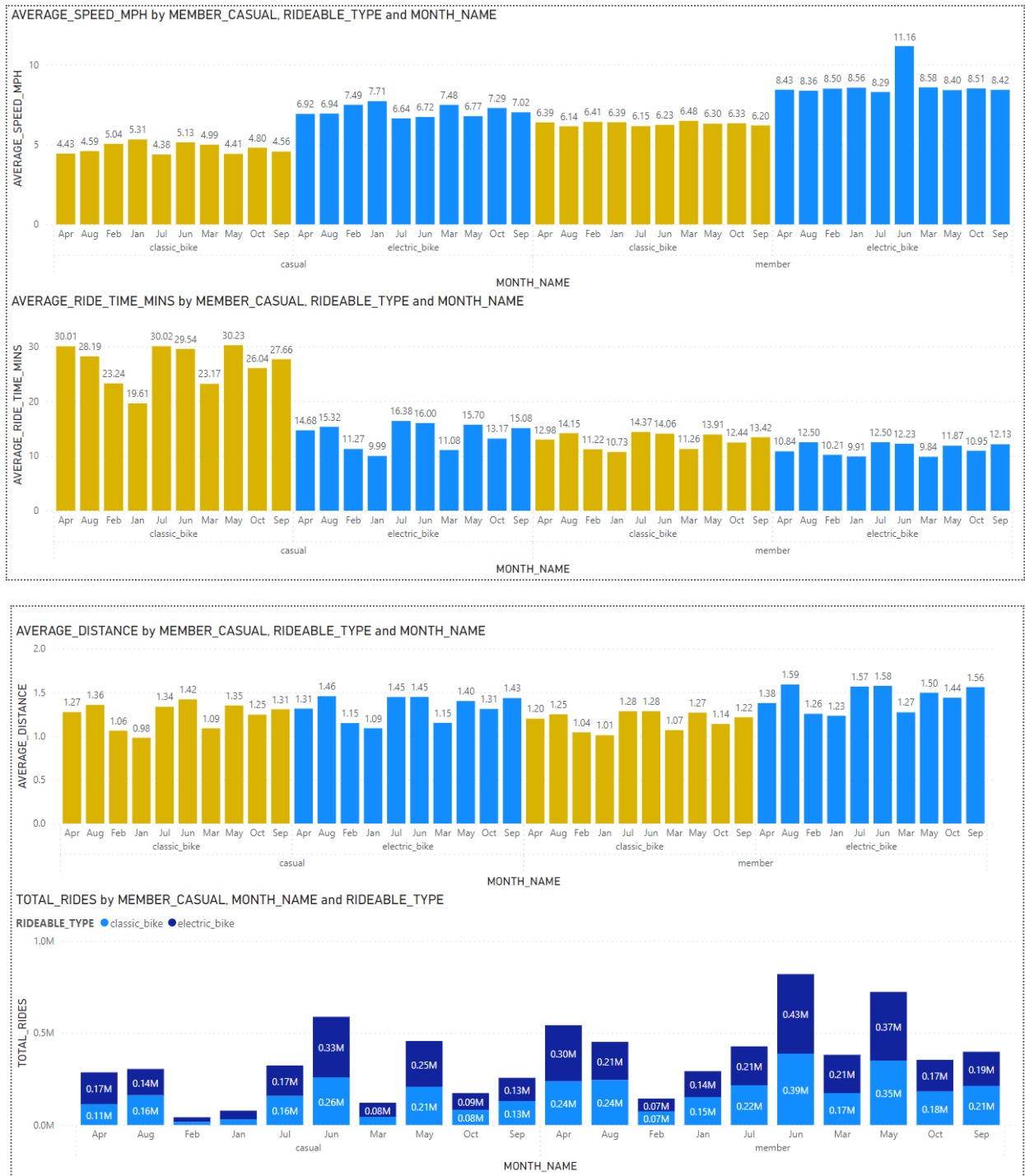
3) The average ride time for classic bikes is 30.01 minutes in January, whereas the average journey time for electric bikes is 19.61 minutes in April. Generally, vintage motorcycles exhibit lengthier riding times; this might be because they are slower than electric bikes, which are quicker and more efficient.

4) January has the largest percentage of classic bike rides (75%), while August has the lowest rate (65%). August has the largest percentage of electric bike rides (30%), while January has the lowest rate (20%). Scooter rides are a consistent percentage of total rides, ranging from 5% to 10%, all year round.

Overall analysis:

Seasonal speed peaks on traditional and electric motorcycles for both casual and member riders show preferences for July and August. Because of their varying levels of efficiency, classic bikes have higher ride durations in January, whereas electric bikes peak in April. According to monthly riding percentages, classic

bikes are preferred in January (75%) and electric bikes in August (30%). All things considered, vintage bikes have faster top speeds, longer ride durations, and different seasonal preferences.



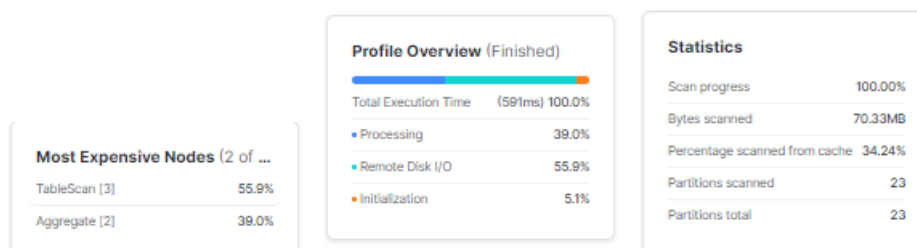
## 6) A Comprehensive Analysis and Visualization of User Type, Rideable Type, and Quarter of the Year:

### Analysis:

Data on bike rides from Divvy, broken down by rideable type, user type, and quarter of the year. For any combination of user type, rideable type, and quarter, it computes a variety of ride data, including the total number of rides, total ride hours, average distance traveled per ride, average ride length, and average speed. This research, which takes into account the many kinds of rideable bikes in the Divvy system, enables an examination of ride patterns and performance metrics for both member and casual users throughout several seasons.

```
1 SELECT
2   member_casual, RIDEABLE_TYPE,
3   QUARTER(started_at) as quarter_season,
4   CASE
5     WHEN QUARTER(started_at) = 1 THEN 'Winter'
6     WHEN QUARTER(started_at) = 2 THEN 'Spring'
7     WHEN QUARTER(started_at) = 3 THEN 'Summer'
8     WHEN QUARTER(started_at) = 4 THEN 'Fall'
9   END as quarter_name,
10  COUNT(*) AS no_of_rides,
11  SUM(ride_length)/3600 AS total_ride_hours,
12  AVG(DISTANCE) AS average_distance,
13  AVG(ride_length) / 60 AS average_ride_time_mins,
14  AVG(DISTANCE / (ride_length / 3600)) AS average_speed_mph
15 FROM
16   divvy_2023_data
17 GROUP BY
18   member_casual, RIDEABLE_TYPE, quarter_season, quarter_name
19 ORDER BY
20   no_of_rides DESC;
```

MEMBER_CASUAL	RIDEABLE_TYPE	QUARTER_SEASON	QUARTER_NAME	NO_OF_RIDES	TOTAL_RIDE_HOURS	AVERAGE_DISTANCE	AVERAGE_RIDE_TIME_MIN
member	electric_bike	2	Spring	1109586	216853.114444444	1.494422172	11.72616351
member	classic_bike	2	Spring	975932	223501.305555556	1.257957419	13.7407917C
casual	electric_bike	2	Spring	751158	195283.183333333	1.401577351	15.59857047
member	classic_bike	3	Summer	673076	156949.935555556	1.249628437	13.99098487
member	electric_bike	3	Summer	604051	124700.451388889	1.572780858	12.38641616
casual	classic_bike	2	Spring	579406	288522.598888889	1.366093899	29.8777643C
casual	classic_bike	3	Summer	450254	215233.155555556	1.335007286	28.68156492
casual	electric_bike	3	Summer	433962	113218.650277778	1.446271678	15.65371856



### Visualization:

1) The graph shows the Q4 lows for casual riders on electric bikes (1.05 miles) and the Q4 high lengths for member riders on classic bikes (1.57 miles). Overall, both member and casual riders, for both classic and electric motorcycles, show a tendency of longer average lengths from Q2 to Q4. Some theories include that longer rides are encouraged by warmer weather, while shorter summer journeys are more likely to favor electric bikes.

2) Due to their knowledge with the city and adept navigation, members typically have lower travel times than casual riders, as seen by the graph. For both rider types, classic bikes have longer ride times than

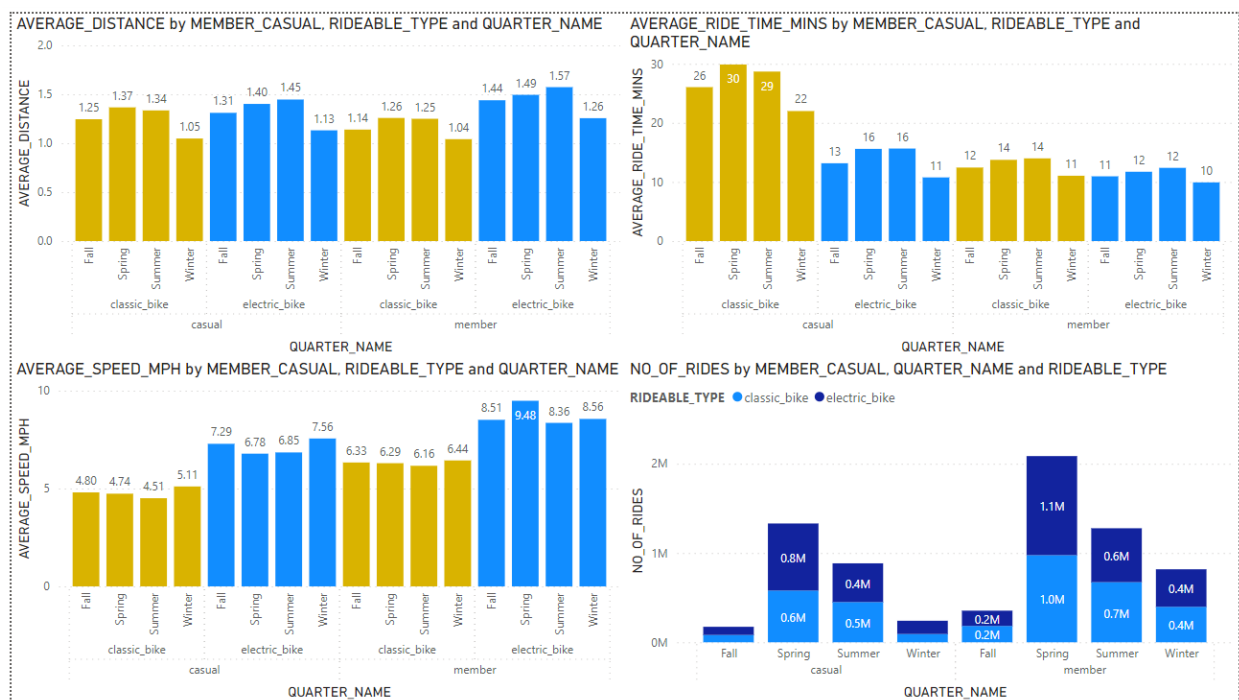


electric bikes; this is probably because electric bikes are faster and more convenient for shorter journeys. The average ride time peaks in the first quarter for all types of riders and bikes, which corresponds to higher bike utilization in the warmer spring and summer months. In general, typical ride times in bike-sharing usage are influenced by rider type, bike type, and seasonal fluctuations.

3) The average bike speed differences by rider type, bike type, and quarter are shown on the graph. Due presumably to motor aid, members ride faster than casual riders (8.51 mph vs. 6.78 mph) and electric bikes are quicker than classic cycles (8.36 mph vs. 7.29 mph). In accordance with meteorological factors, speed peaks in the spring (9.48 mph) and summer (8.56 mph) and decreases in the fall (6.33 mph) and winter (6.16 mph). In general, average cycling speed is influenced by rider type, bike type, and season.

4) The graph shows an increase in rides every three months, broken down by bike category and rider type. Summer is when members ride their classic and electric bikes the highest; fall is when casual riders reach their pinnacle. All categories have the lowest rides during the winter. Over the course of the quarters, member rides outnumber casual riders overall.

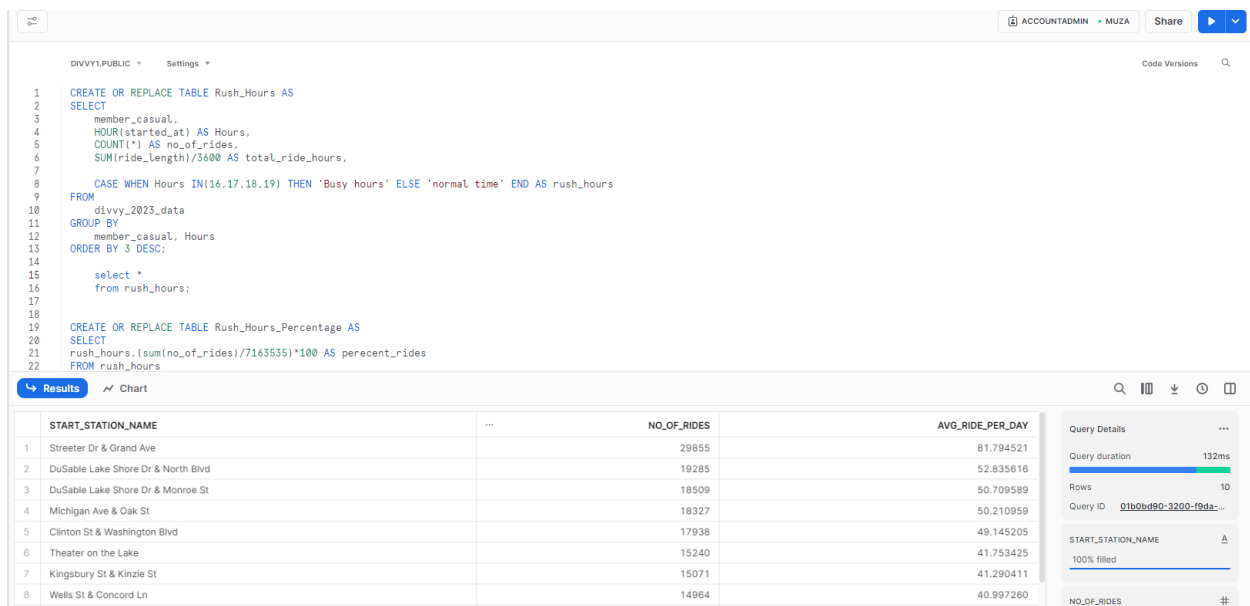
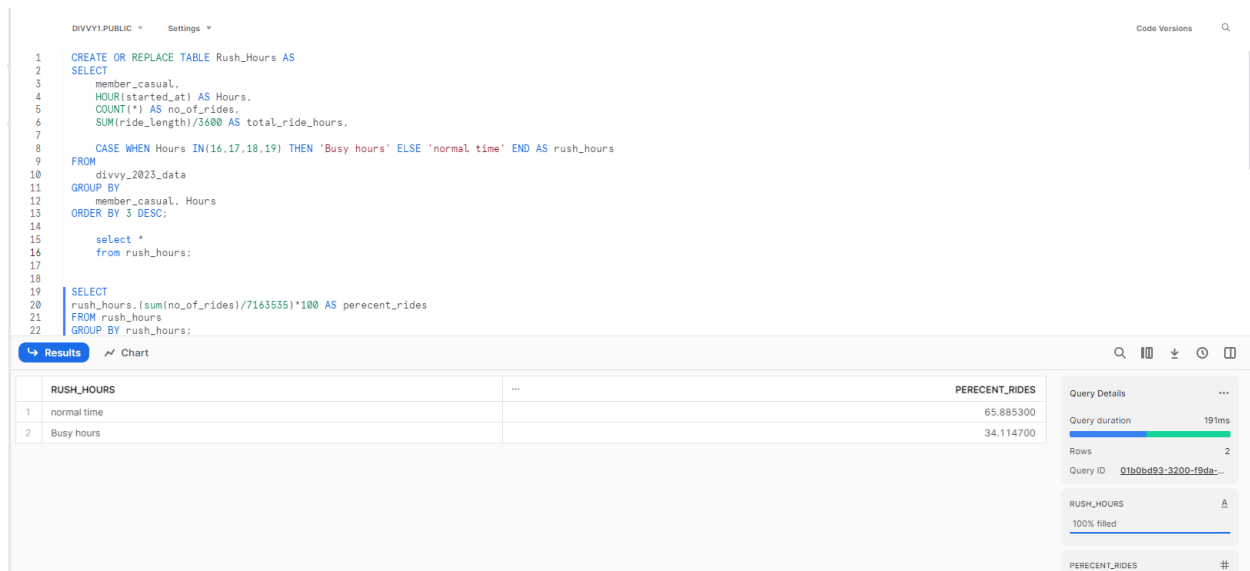
Overview: The graph displays quarterly trends in bike-sharing data, showing higher average ride durations and lengths from Q2 to Q4, quicker member and electric bike speeds, and a seasonal increase in the number of rides. Across all categories, members routinely outpace casual users.



## 7) User Behavior and Demand Patterns During Rush Hour:

### Analysis:

We have examined patterns in Divvy's bike ride statistics during rush hour. We determined the busiest stations during peak times, identified peak hours, and computed the proportion of riders during those hours. Understanding user behavior and demand patterns during particular times of the day was made easier with the assistance of insights on the frequency of rides during rush hours and the top stations with the largest ride volumes during those peak hours.

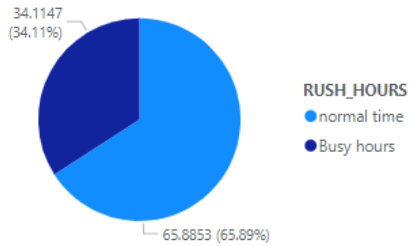


## Visualization:

The percentage of rides during peak hours in various Chicago neighborhoods is displayed in a pie chart. Two portions make up the pie chart: rush hour and regular time. The segment dedicated to rush hours is greater, including 65.89% of all journeys. 34.11% of all rides are in the usual time portion.

Streeter Dr. & Grand Ave (34.11%), DuSable Lake Shore Dr. & North Blvd (19.29%), and DuSable Lake Shore Dr. & Monroe St (18.51%) are the busiest stations during rush hour. During regular business hours, the busiest stations are Theater on the Lake (41.75%), Kingsbury St & Kinzie St (41.29%), and Millennium Park (65.89%).

PERCENT\_RIDES by RUSH\_HOURS



Busiest\_Stations\_In\_RUSH\_HOURS

START_STATION_NAME	Sum of NO_OF_RIDES	AVG_RIDE_PER_DAY
Streeter Dr & Grand Ave	29,855.00	81.79
DuSable Lake Shore Dr & North Blvd	19,285.00	52.84
DuSable Lake Shore Dr & Monroe St	18,509.00	50.71
Michigan Ave & Oak St	18,327.00	50.21
Clinton St & Washington Blvd	17,938.00	49.15
Theater on the Lake	15,240.00	41.75
Kingsbury St & Kinzie St	15,071.00	41.29
Wells St & Concord Ln	14,964.00	41.00
Clark St & Elm St	14,650.00	40.14
Millennium Park	14,375.00	39.38
<b>Total</b>	<b>178,214.00</b>	

## 8) Top 10 popular routes:

### Analysis:

Based on the amount of rides recorded between the top 10 pairings in the Divvy bike system dataset for 2023, we have compiled a list of the most frequently visited station pairs, as seen in the screenshot below.

ACCOUNTADMIN

MUZA

Share

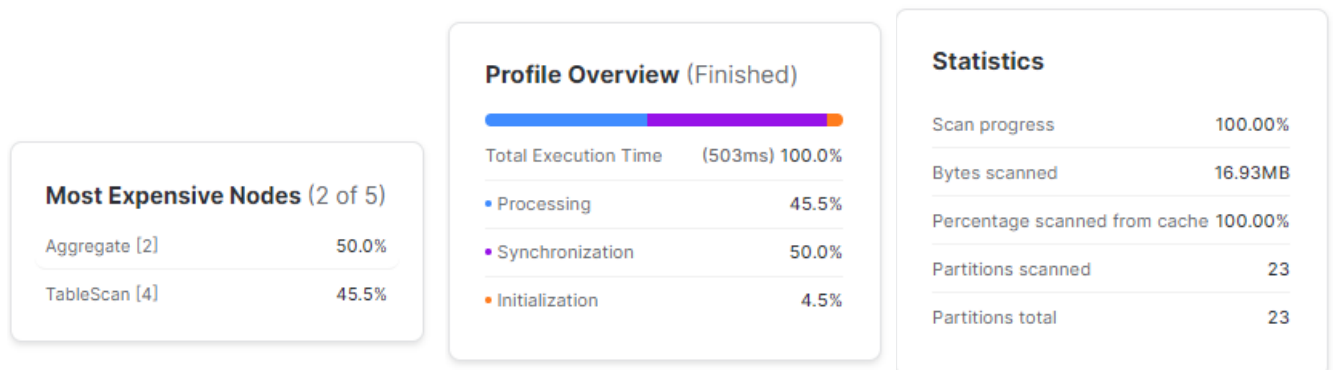
Code Versions

DIVVY.PUBLIC

Settings

```

1 SELECT
2   CONCAT(COALESCE(START_STATION_NAME, ''), ' - ', COALESCE(END_STATION_NAME, '')) AS STATION_PAIR,
3   COUNT(*) AS TRIP_COUNT
4 FROM
5   DIVVY_2023_DATA
6 WHERE
7   START_STATION_NAME IS NOT NULL OR END_STATION_NAME IS NOT NULL
8 GROUP BY
9   STATION_PAIR
10 ORDER BY
11   TRIP_COUNT DESC
12 LIMIT 10;
```



### Visualization:

The greatest trip count for the same station pair (Streeter Dr & Grand Ave) indicates that the most popular bike-sharing routes exhibit a considerable predilection for short-distance rides between neighboring stations. DuSable Lake Shore Dr & Monroe St - The proximity of DuSable Lake Shore Dr & Monroe St suggests that there is a need for round-trip travel. This emphasizes how important station proximity is to user behavior and travel patterns.

ResultsChart

	STATION_PAIR	TRIP_COUNT
1	Streeter Dr & Grand Ave - Streeter Dr & Grand Ave	12332
2	DuSable Lake Shore Dr & Monroe St - DuSable Lake Shore Dr & Monroe St	9378
3	Ellis Ave & 60th St - University Ave & 57th St	9137
4	Ellis Ave & 60th St - Ellis Ave & 55th St	9065
5	University Ave & 57th St - Ellis Ave & 60th St	8465
6	Ellis Ave & 55th St - Ellis Ave & 60th St	8407
7	DuSable Lake Shore Dr & Monroe St - Streeter Dr & Grand Ave	6862
8	Michigan Ave & Oak St - Michigan Ave & Oak St	6348
9	Calumet Ave & 33rd St - State St & 33rd St	5999
10	State St & 33rd St - Calumet Ave & 33rd St	5902

Query Details

Query duration752ms

Rows10

Query ID01a0d05ca-3200-fbde-0...

STATION\_PAIR

100% filled

TRIP\_COUNT

590212332

## 9) Top 10 popular station:

### Analysis:

We have examined the most popular stations in the Divvy bike system dataset for 2023, presenting the top 10 stations by number of trips started or completed at those locations.

DIVVY1.PUBLIC		Settings	ACCOUNTADMIN MUZA	Share
<pre> 1 SELECT 2   COALESCE(START_STATION_NAME, END_STATION_NAME) AS STATION_NAME, 3   COUNT(*) AS TRIP_COUNT 4 FROM 5   DIVVY_2023_DATA 6 WHERE 7   START_STATION_NAME IS NOT NULL OR END_STATION_NAME IS NOT NULL 8 GROUP BY 9   STATION_NAME 10 ORDER BY 11   TRIP_COUNT DESC 12 LIMIT 10; 13 14</pre>		Code Versions		

### Visualization:

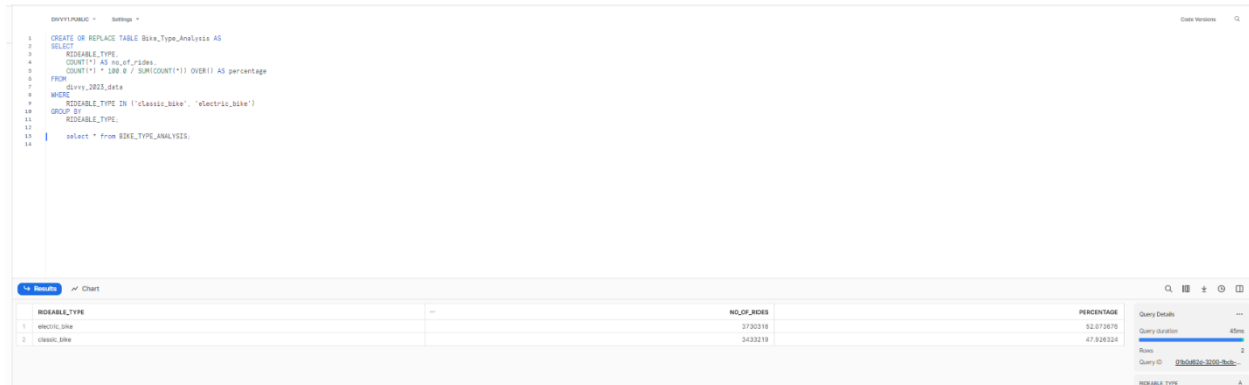
The top bike-sharing stations, led by Streeter Dr & Grand Ave with 85,737 trips, indicate strategic placement in popular areas, fostering high demand for services. DuSable Lake Shore Dr & Monroe St and Michigan Ave & Oak St follow closely, highlighting their pivotal roles in facilitating numerous trips for users.

STATION_NAME	TRIP_COUNT			
1 Streeter Dr & Grand Ave	85737			
2 DuSable Lake Shore Dr & Monroe St	54918			
3 Michigan Ave & Oak St	51775			
4 DuSable Lake Shore Dr & North Blvd	49934			
5 Clark St & Elm St	48300			
6 Wells St & Concord Ln	47189			
7 Kingsbury St & Kinzie St	46875			
8 Clinton St & Washington Blvd	42673			
9 Theater on the Lake	41997			
10 Wells St & Elm St	41180			

## 10) Analysis and Visualization: Understanding Bike and Member Types in Divvy 2023 Data

### Analysis:

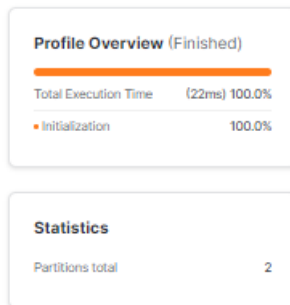
a) 'RIDEABLE\_TYPE' (classic or electric bike), 'no\_of\_rides' (count of rides for each type), and 'percentage' (the percentage of rides each bike type contributes to the total rides in the 'divvy\_2023\_data' dataset) are the columns that make up the final 'Bike\_Type\_Analysis' table.



```
1 CREATE OR REPLACE TABLE Bike_Type_Analysis AS
2 SELECT
3   RIDEABLE_TYPE,
4   COUNT(*) AS no_of_rides,
5   COUNT(*) * 100.0 / SUM(COUNT(*)) OVER() AS percentage
6 FROM
7   divvy_2023_data
8 WHERE
9   RIDEABLE_TYPE IN ('classic_bike', 'electric_bike')
10 GROUP BY
11   RIDEABLE_TYPE;
12
13 | select * from BIKE_TYPE_ANALYSIS;
```

RIDEABLE_TYPE	no_of_rides	PERCENTAGE
electric_bike	3730316	52.67381%
classic_bike	3433219	47.32619%

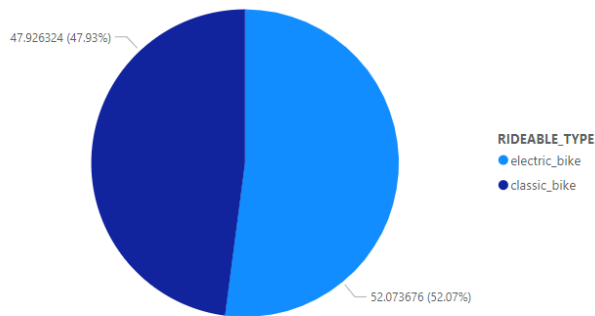
Query Details: Query duration: 45ms, Rows: 2, Query ID: 770d4f4e-2000-76d0-



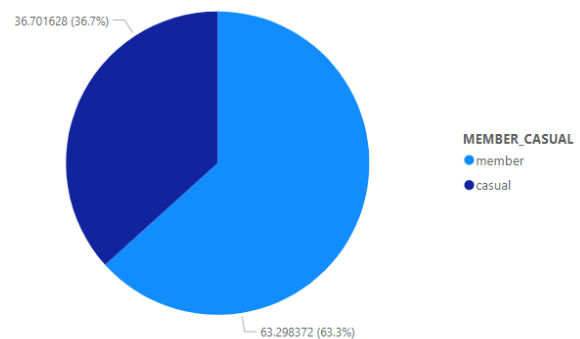
b)The resulting 'MEMBER\_TYPE' table will likely contain columns for 'MEMBER\_CASUAL' (membership status), 'total' (the count of occurrences for each status), and 'percentage' (the percentage of each membership status in the 'divvy\_2023\_data' dataset). This table offers insights into the distribution of membership statuses within the dataset, showcasing the counts and percentages for different types of members ('member' or 'casual').



PERCENTAGE by RIDEABLE\_TYPE



PERCENTAGE by MEMBER\_CASUAL



## Future Recommendations

### Expanding Service:

1. Boost the number of stations in high-demand locations, transportation hubs, and underdeveloped areas.
2. Create stations exclusively for e-bikes to meet the growing demand for these vehicles.
3. Investigate micro-mobility choices by including dockless bikes and e-scooters.

### Data-driven Optimization

4. Utilize machine learning to forecast demand, locate stations, and distribute resources.
5. Survey users to learn more about their preferences and expectations.
6. Keep an eye on use trends and conduct ongoing analysis to adjust tactics and remain sensitive to user requirements.

## Conclusion

In summary, Snowflake's powerful big data capabilities have shown to be an invaluable resource for the Divvy Bike System, driving it toward increased sustainability and efficiency. By leveraging Snowflake's cloud-based data platform, Divvy has effectively leveraged big data to extract useful insights and improve its offerings. Divvy has benefited greatly from the platform's ability to handle enormous datasets and offer real-time analytics in a number of important areas. It has made it possible for the system to fully comprehend user behavior, which has made it possible to analyze journey patterns, identify well-liked stations and routes, and forecast variations in demand. This has therefore made it easier to strategically site stations and allocate resources more effectively.

Furthermore, the partnership has enabled Divvy to actively promote sustainability. Through Snowflake's capabilities, Divvy can track the environmental impact of its system, measure carbon footprint reduction, and devise initiatives to encourage sustainable transportation choices among users. The successful collaboration between Snowflake and Divvy underscores the transformative potential of cloud-based data platforms in optimizing public services and fostering sustainable development. In this case, the fusion of Snowflake's cutting-edge technology with Divvy's commitment to innovation has resulted in a transportation system that not only meets the evolving needs of Chicago residents but also sets a precedent

for the broader integration of big data in public services. Additionally, Snowflake has improved user experience by enabling Divvy to customize suggestions, anticipate wait times, and optimize routes according to user preferences, all of which contribute to a more customized and user-friendly service. Snowflake's real-time data insights are increasingly essential for making data-driven, well-informed decisions. To provide a flexible and responsive transportation system, Divvy can effectively distribute resources, improve pricing tactics, and respond quickly to developing trends.

Additionally, Divvy is now able to aggressively promote sustainability thanks to the agreement. With the help of Snowflake, Divvy is able to monitor the environmental effect of its system, calculate the reduction of carbon footprints, and create campaigns to motivate users to choose environmentally friendly modes of transportation. The fruitful partnership between Divvy and Snowflake highlights how revolutionary cloud-based data platforms can be for promoting sustainable growth and streamlining public services. In this instance, Divvy's dedication to innovation and Snowflake's state-of-the-art technology have combined to create a transportation system that not only fulfills the changing demands of Chicagoans but also establishes a standard for the wider integration of big data in public services.

## References:

1. <https://divvybikes.com/system-data>
2. <https://www.kaggle.com/code/devisangeetha/divvy-bike-share-eda-network-analysis>
3. <https://jeremyrieunier.com/portfolio/divvy-cleaning>
4. <https://www.linkedin.com/pulse/data-analysis-visualizations-chicago-divvy-bikes-sharing-mazarei/>
5. <https://medium.com/codex/exploratory-data-analysis-cyclistic-bike-share-analysis-case-study-1b1a00475a4f>
6. <https://towardsdatascience.com/predicting-hourly-divvy-bike-sharing-checkouts-per-station-65b1d217d8a4>
7. <https://www.slideshare.net/HanbitChoi1/divvy-bike-use-data-analysis-and-recommendations>
8. <https://nycdatascience.com/blog/student-works/data-visualization-for-citibike-usage/>
9. J. Zhang, X. Pan, M. Li and P. S. Yu, "Bicycle-Sharing System Analysis and Trip Prediction," 2016 17th IEEE International Conference on Mobile Data Management (MDM), Porto, Portugal, 2016, pp. 174-179, doi: 10.1109/MDM.2016.35.