Uber Dataset Analysis In SQL

By Purva Pharat

Introduction

My name is Purva Pharat. I worked on this project using the Uber dataset to enhance my skills in data analysis and SQL. This project allowed me to explore real-world data, practice writing efficient SQL queries, and gain deeper insights into analysing and interpreting data. It was a valuable experience that helped me improve my technical expertise and problem-solving abilities.

Data Description

The UberDataset_cleaned.csv file contains ride details collected from Uber trips. It captures various attributes of the trips, including time, location, purpose, and distance. The dataset comprises 1154 entries and 11 columns, which are described below:

Column Descriptions:

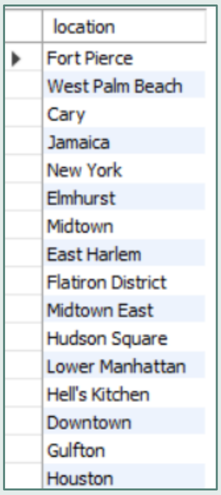
1. START_DATE: The timestamp marking the start of the trip.
2. END_DATE: The timestamp indicating the end of the trip.
3. CATEGORY: Categorizes the trip as either "Business" or "Personal."
4. START: The starting location of the trip.
5. STOP: The destination location of the trip.
6. MILES: The total distance of the trip measured in miles.
7. PURPOSE: Specifies the reason for the trip (e.g., "Meeting," "Airport Drop-off").
8. TIME_OF_DAY: Groups the trip into periods such as morning, afternoon, or evening.
9. MONTH_OF_THE_RIDE: Identifies the month in which the trip occurred.
10. DAY_OF_THE_RIDE: Specifies the day of the week for the trip.
11. DURATION_OF_THE_RIDE: Duration of the trip in hours and minutes format.

Database Link: Click Here | Clean Data Link: Click Here | Python File: Click Here

Basic Level

- List all unique pickup locations to identify distinct areas of service.

```
select distinct START
from uberdataset;
```

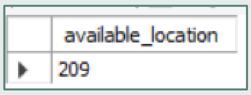


location
Fort Pierce
West Palm Beach
Cary
Jamaica
New York
Elmhurst
Midtown
East Harlem
Flatiron District
Midtown East
Hudson Square
Lower Manhattan
Hell's Kitchen
Downtown
Gulfton
Houston

There are 177 locations available for pickup

- List all unique available locations to identify distinct areas of service.

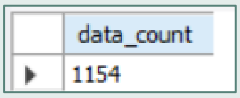
```
select count(*) as available_location
from (
        select distinct(START) as area
        from uberdataset
        union
        select distinct(STOP) as area
        from uberdataset) as data ;
```



available_location
209

- Determine the total number of rides recorded in the dataset.

```
select count(*) as data_count
from uberdataset;
```

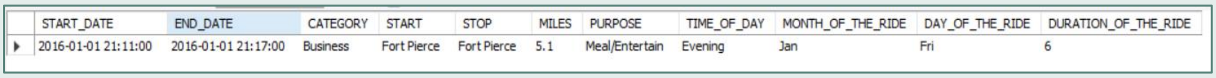


data_count
1154

- Display the earliest and latest pickup_datetime data.

    a. Earlier Pickup Time

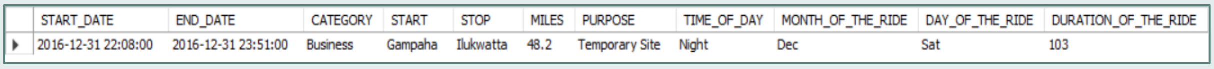
```
select *
from uberdataset
where START_DATE = (
        select min(START_DATE)
        from uberdataset);
```



START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	TIME_OF_DAY	MONTH_OF_THE_RIDE	DAY_OF_THE_RIDE	DURATION_OF_THE_RIDE
2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	Evening	Jan	Fri	6

b. Latest Pickup Time

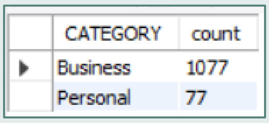
```
select *
from uberdataset
where START_DATE = (
        select max(START_DATE)
        from uberdataset);
```



START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	TIME_OF_DAY	MONTH_OF_THE_RIDE	DAY_OF_THE_RIDE	DURATION_OF_THE_RIDE
2016-12-31 22:08:00	2016-12-31 23:51:00	Business	Gampaha	Ilukwatta	48.2	Temporary Site	Night	Dec	Sat	103

- List all unique values in the CATEGORY column (e.g., business, personal).

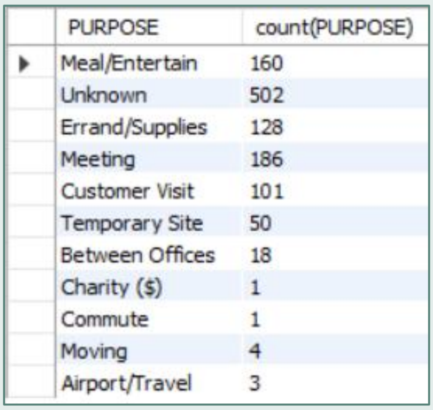
```
select CATEGORY , count(CATEGORY) as count
from uberdataset
group by CATEGORY;
```



CATEGORY	count
Business	1077
Personal	77

- Count the total number of rides for each PURPOSE category.

```
select PURPOSE,count(PURPOSE)
from uberdataset
group by PURPOSE;
```

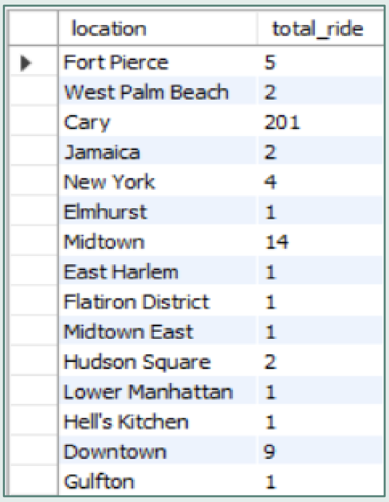


PURPOSE	count(PURPOSE)
Meal/Entertain	160
Unknown	502
Errand/Supplies	128
Meeting	186
Customer Visit	101
Temporary Site	50
Between Offices	18
Charity ($)	1
Commute	1
Moving	4
Airport/Travel	3

Intermediate Level

- Calculate the total number of rides for each pickup_location.

```
select START as location , count(START) as total_ride
from uberdataset
group by START;
```

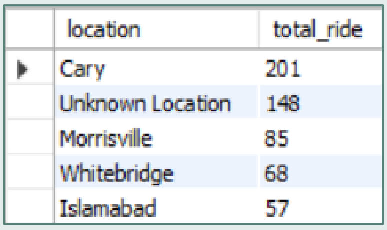


location	total_ride
Fort Pierce	5
West Palm Beach	2
Cary	201
Jamaica	2
New York	4
Elmhurst	1
Midtown	14
East Harlem	1
Flatiron District	1
Midtown East	1
Hudson Square	2
Lower Manhattan	1
Hell's Kitchen	1
Downtown	9
Gulfton	1

There are 177 locations available for pickup

- Find the top 5 busiest pickup_locations.

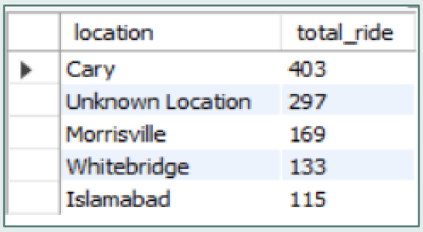
```
select START as location , count(START) as total_ride
from uberdataset
group by START
order by count(START) desc
limit 5;
```



location	total_ride
Cary	201
Unknown Location	148
Morrisville	85
Whitebridge	68
Islamabad	57

- Find the top 5 busiest locations.

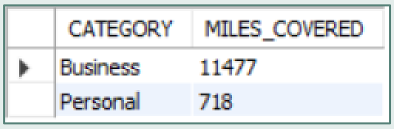
```
SELECT START as location , SUM(total_ride) as total_ride
from (   select START,count(START) as total_ride
         from uberdataset
         group by START
         union all
         select STOP,count(STOP) as total_ride
         from uberdataset
         group by STOP) as data
group by START
order by SUM(total_ride) desc
limit 5;
```



location	total_ride
Cary	403
Unknown Location	297
Morrisville	169
Whitebridge	133
Islamabad	115

- Calculate the total distance traveled (MILES) for each ride category (CATEGORY).

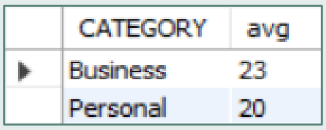
```
select CATEGORY , round(SUM(MILES)) as MILES_COVERED
from uberdataset
group by CATEGORY;
```



CATEGORY	MILES_COVERED
Business	11477
Personal	718

- Calculate the average DURATION_OF_THE_RIDE(min) for each ride category.

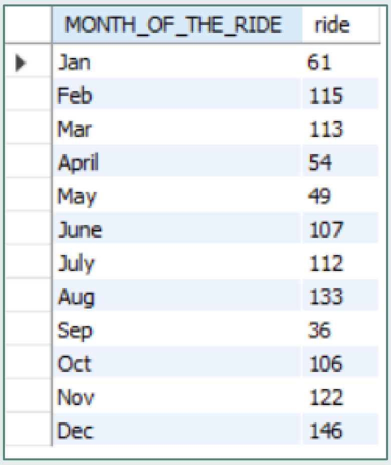
```
select CATEGORY,round(avg(DURATION_OF_THE_RIDE)) as avg
from uberdataset
group by CATEGORY;
```



CATEGORY	avg
Business	23
Personal	20

- Analyze the total number of rides by MONTH_OF_THE_RIDE.

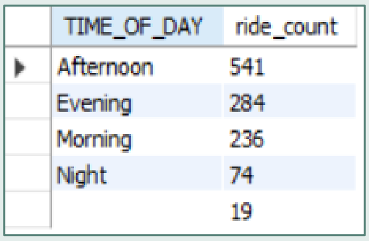
```
select MONTH_OF_THE_RIDE,count(MONTH_OF_THE_RIDE) as ride
from uberdataset
group by MONTH_OF_THE_RIDE
order by count(MONTH_OF_THE_RIDE) desc
limit 1 ;
```



MONTH_OF_THE_RIDE	ride
Jan	61
Feb	115
Mar	113
April	54
May	49
June	107
July	112
Aug	133
Sep	36
Oct	106
Nov	122
Dec	146

- Identify the time of day (TIME_OF_DAY) when most rides occur.

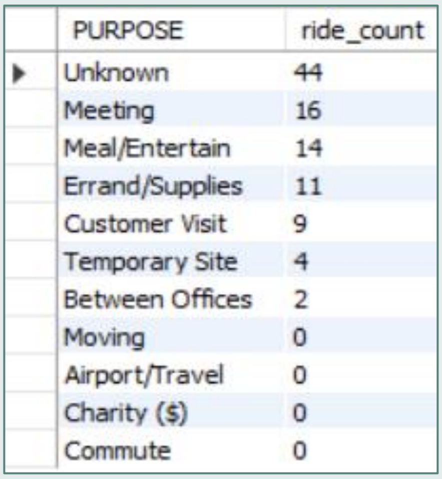
```
select TIME_OF_DAY, count(TIME_OF_DAY) as ride_count
from uberdataset
group by TIME_OF_DAY
order by count(TIME_OF_DAY) desc;
```



TIME_OF_DAY	ride_count
Afternoon	541
Evening	284
Morning	236
Night	74
	19

- Compute the percentage of rides for each PURPOSE to understand their relative importance.

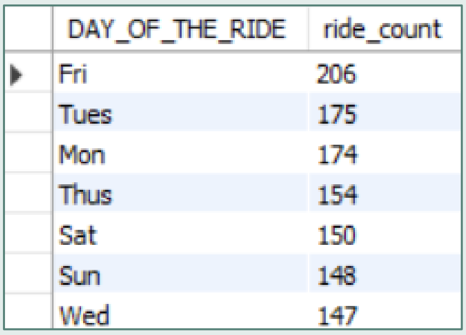
```
select PURPOSE,
round(count(PURPOSE)/(select (count(*)) as ride_count from uberdataset) * 100)  as
ride_count
from uberdataset
group by PURPOSE
order by count(PURPOSE) desc;
```



	PURPOSE	ride_count
▶	Unknown	44
	Meeting	16
	Meal/Entertain	14
	Errand/Supplies	11
	Customer Visit	9
	Temporary Site	4
	Between Offices	2
	Moving	0
	Airport/Travel	0
	Charity ($)	0
	Commute	0

- Find the day of the week (DAY_OF_THE_RIDE) with the highest number of rides.

```
select DAY_OF_THE_RIDE,count(DAY_OF_THE_RIDE) as ride_count
from uberdataset
group by DAY_OF_THE_RIDE
order by count(DAY_OF_THE_RIDE) desc;
```

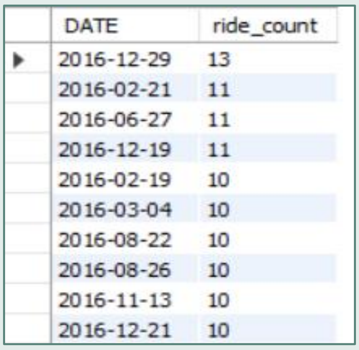


	DAY_OF_THE_RIDE	ride_count
▶	Fri	206
	Tues	175
	Mon	174
	Thus	154
	Sat	150
	Sun	148
	Wed	147

Advanced Level

- Identify the day with the highest number of rides.

```
select date(START_DATE) AS DATE , count(START) as ride_count
from uberdataset
group by date(START_DATE)
order by count(START) desc;
```

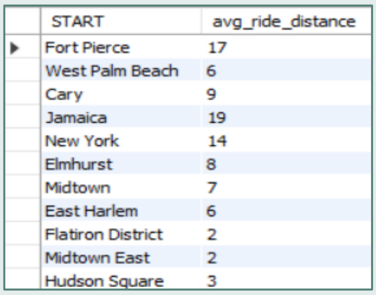


	DATE	ride_count
▶	2016-12-29	13
	2016-02-21	11
	2016-06-27	11
	2016-12-19	11
	2016-02-19	10
	2016-03-04	10
	2016-08-22	10
	2016-08-26	10
	2016-11-13	10
	2016-12-21	10

294 rows are return

- Find the average ride distance for each pickup_location.

```
select START , round(avg(MILES)) as avg_ride_distance
from uberdataset
group by START;
```

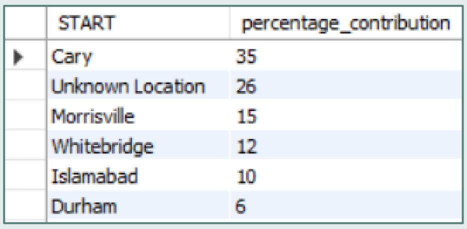


	START	avg_ride_distance
▶	Fort Pierce	17
	West Palm Beach	6
	Cary	9
	Jamaica	19
	New York	14
	Elmhurst	8
	Midtown	7
	East Harlem	6
	Flatiron District	2
	Midtown East	2
	Hudson Square	3

177 rows are return

- Find the percentage contribution of each pickup_location to the total rides.

```
select START , round(sum(total_ride)/(select count(*) from uberdataset)*100) as
percentage_contribution
from (   select START,count(START) as total_ride
         from uberdataset
         group by START
         union all
         select STOP,count(STOP) as total_ride
         from uberdataset
         group by STOP) as data
group by START
order by sum(total_ride) desc;
```



	START	percentage_contribution
▶	Cary	35
	Unknown Location	26
	Morrisville	15
	Whitebridge	12
	Islamabad	10
	Durham	6

209 rows are return

Insights

There are 177 pickup locations available.