My Divvy Trips Exercise

Handayani Utomo

Collect Data

The data available for analysis consists of 4 quarter data. The data is both interesting and challenging, the amount is large but the structure is not so good.

The data appears in various formats, csv and non-csv. Data in the form of non-csv is opened using notepad and then converted into csv form.

It can be seen that the data for each quarter has a different structure with different field names even though it is intended to accommodate the same information.

```
Upload Divvy datasets (csv files) here
q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
q1_2020 <- read_csv("Divvy_Trips_2020_Q1.csv")
```

Wrangle Data and Combine into a Single File

The first step is to tidy up the data structure, the fields for quarters other than q1_2020 are adjusted so that they match the q1_2020 data fields.

q1_2020	q2_2019
ride_id	01 - Rental Details Rental ID
rideable_type	01 - Rental Details Local Start Time
started_at	01 - Rental Details Local End Time
ended_at	01 - Rental Details Bike ID
start_station_name	01 - Rental Details Duration In Seconds Uncapped
start_station_id	03 - Rental Start Station ID
end_station_name	03 - Rental Start Station Name
end_station_id	02 - Rental End Station ID
start_lat	02 - Rental End Station Name
start_lng	User Type
end_lat	Member Gender
end_lng	05 - Member Details Member Birthday Year
member_casual	

q3_2019	q4_2019
trip_id	trip_id
start_time	start_time
end_time	end_time
bikeid	bikeid
tripduration	tripduration
from_station_id	from_station_id
from_station_name	from_station_name
to_station_id	to_station_id
to_station_name	to_station_name
usertype	usertype
gender	gender
birthyear	birthyear

Transform the q3_2019 and q4_2019 fields to be compatible with the q1_2020 fields

ride_id	=	trip_id
rideable_type	=	bikeid
started_at	=	start_time
ended_at	=	end_time
start_station_name	=	from_station_name
start_station_id	=	from_station_id
end_station_name	=	to_station_name
end_station_id	=	to_station_id
member_casual	=	usertype

Transform the q2_2019 and q4_2019 fields to be compatible with the q1_2020 fields

ride_id	=	"01 - Rental Details Rental ID"
rideable_type	=	"01 - Rental Details Bike ID"
started_at	=	"01 - Rental Details Local Start Time"
ended_at	=	"01 - Rental Details Local End Time"
start_station_name	=	"03 - Rental Start Station Name"
start_station_id	=	"03 - Rental Start Station ID"
end_station_name	=	"02 - Rental End Station Name"
end_station_id	=	"02 - Rental End Station ID"
member_casual	=	"User Type"

To determine how long a bicycle is used, I created a new variable that I named ride_length_minutes which will store the time difference between ended_at and started_at in minutes.

I did this after I failed to process data using R. Various R functions (as.Date(), as.POSIXct(), as.POSIXlt() and strptime()) I have used to change the data class from character to date, but the result always returns to NA.

The same thing I did for the weekday field because the weekday field reads data from the started_at field.

After all fields are transformed, all quarter data is merged into one large data with the name all_trips.

Clean Up and Add Data to Prepare for Analysis

The original data frame consist of more than 4 million records and after I ran it using R, there are several data that must be deleted, especially data with a negative and zero value in the ride_length_minutes field.

In q4_2019 data frame there are 13 negative value data in the ride_length_minutes field, in q1_2020 data frame there are 25 negative value data in the ride_length_minutes field and there are several thousand data with zero value in ride_length_minutes field. All data with negative and zero value must be removed before data analysis can be carried out. After data cleaning process is done, there are total number of 3,276,657 records remain.

Other fields that are not used in data analysis are also deleted such as fields start_lat, start_lng, end_lat, end_lng, gender and birthday. I named the clean data all_trips_v2. The customer type is changed from Customer and Subscriber to Casual and Member.

Conduct Descriptive Analysis

The descriptive analysis of all_trips_v2 can be summarized as follows:

all_trips_v2$ride_length_minutes

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1.00	7.00	11.00	14.87	20.00	59.00

aggregate summary in minutes

	Mean	Median	Max	Min
casual	24.07	22.00	59.00	1.00
member	12.18	10.00	59.00	1.00

aggregate summary days_of_week in minutes

	Casual	Member
Sunday	24.52406	12.43192
Monday	24.34846	12.30749
Tuesday	23.27740	11.85203
Wednesday	24.29647	12.03826
Thursday	23.91126	12.20291
Friday	24.22400	12.35381
Saturday	22.93388	11.62878

aggregate summary average_duration

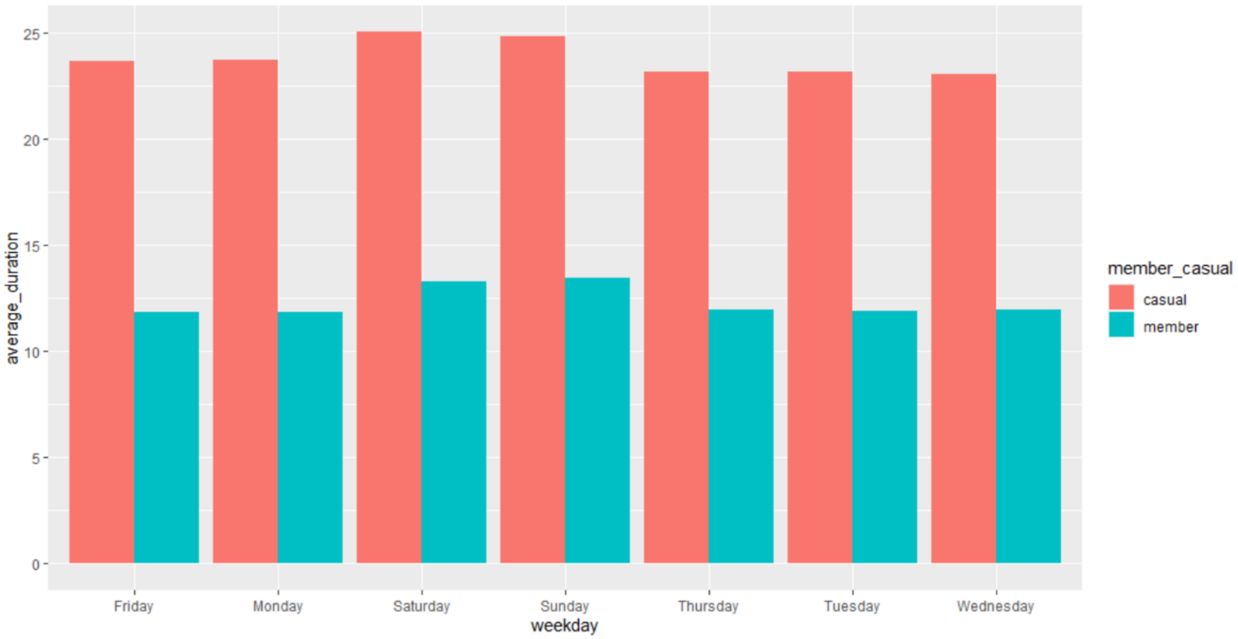
	Casual	Member
Sunday	24.8	13.5
Monday	23.7	11.9
Tuesday	23.2	11.9
Wednesday	23.1	12.0
Thursday	23.2	11.9
Friday	23.7	11.8
Saturday	25.0	13.3

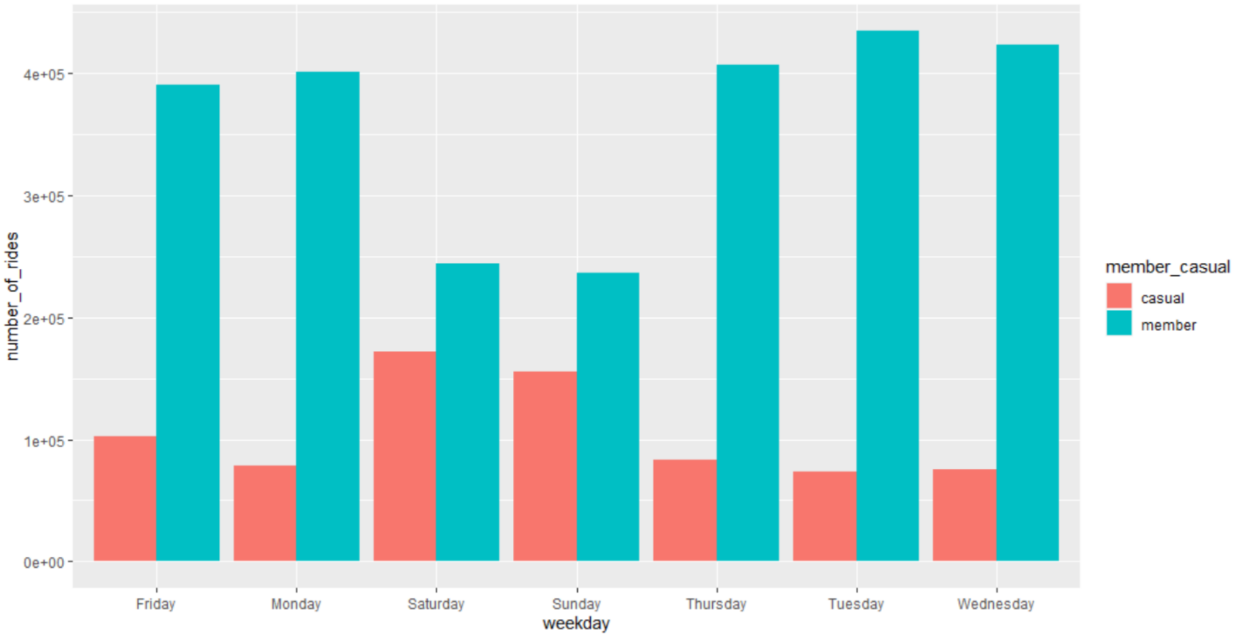
aggregate summary number_of_rides

	Casual	Member
Sunday	155,859	236,234
Monday	78,800	401,155
Tuesday	73,727	434,349
Wednesday	75,362	422,710
Thursday	83,292	406,703
Friday	102,738	390,088
Saturday	171,367	244,273

From the table above, it can be seen that Members use bicycles more often even though the duration of their use is on average shorter than Casuals. On the other hand, Casuals on average spends more time on a bicycle but uses it less frequently.

Visualize the Result


average_duration
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
weekday
member_casual
casual
member


number_of_rides
4e+05
3e+05
2e+05
1e+05
0e+00
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
weekday
member_casual
casual
member