Ask

Stakeholders

The list of stakeholders is as follows:

- Cyclistic company: a bike-share company located in Chicago, United States.
- Lily Moreno: the director of Marketing responsible for the development of campaigns.
- Marketing Analytics team: a team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
- Junior data analyst: part of the Marketing Analytics team. In charge of making this report.
- Cyclistic executive team: they make decisions over the marketing strategies.
- Casual riders: customers who purchase single-ride or full-day passes.
- Cyclistic members: customers who purchase annual memberships.
- Cyclistic's finance analyst: a team of data analysts that have concluded that annual members are much more profitable than casual riders.

Business Task

The question that this report will try to address is: How do annual members and casual riders use Cyclistic bikes differently? It is believed that there is an opportunity to convert casual riders into members, that's why it is important to understand how differently they behave and use the bike-share service to be able to address future marketing campaigns.

Prepare

The data used for the present project has been made available by Motivate International Inc. under this license: https://divvybikes.com/data-license-agreement

The data is located in the following URL: https://divvy-tripdata.s3.amazonaws.com/index.html

The files that are going to be used are the most recent quarterly files:

- Divvy_Trips_2019_Q1.zip
- Divvy_Trips_2019_Q2.zip
- Divvy_Trips_2019_Q3.zip
- Divvy_Trips_2019_Q4.zip
- Divvy_Trips_2020_Q5.zip

Each zip file contained a CSV file with the same name, for example: Divvy_Trips_2019_Q1.csv.

The following chart presents a detailed explanation of the columns that each CSV file has, their data type, and a description of what it is:

Column Name	Datatype	Description
trip_id	Number	Identification number of the trip.
start_time	Date and time	Start date and time of the trip.
end_time	Date and time	End date and time of the trip.
bikeid	Number	Identification number of the bike.
tripduration	Number	Duration of the trip in seconds.

from_station_id	Number	Identification of the initial bike station.
from_station_name	Text	Location of the initial bike station.
to_station_id	Number	Identification of the final bike station.
to_station_name	Text	Location of the final bike station.
usertype	Text (list)	Type of user: Subscriber or Customer.
gender	Text (list)	Gender of the user: Male or Female.
birthyear	Number	Year of birth of the user.

However, one of the CSV files contains a different structure:

Text in yellow means a new column. ~~Stroked text~~ means a column that no longer exists.

Column Name	Datatype	Description
ride_id	Number	Identification number of the trip.
rideable_type	Text	There is only one value there: docked_bike
started_at	Date and time	Start date and time of the trip.
ended_at	Date and time	End date and time of the trip.
~~bikeid~~		
~~tripduration~~		
start_station	Number	Identification of the initial bike station.
start_station_name	Text	Location of the initial bike station.
end_station	Number	Identification of the final bike station.
end_station_name	Text	Location of the final bike station.
member_casual	Text (list)	Type of user: Member or Casual.
start_lat	Decimal	Latitude value of the initial bike station
start_lng	Decimal	Longitude value of the initial bike station
end_lat	Decimal	Latitude value of the final bike station
end_lng	Decimal	Longitude value of the final bike station
~~gender~~		
~~birthyear~~		

As part of this process, the Divvy_Trips_2019_Q1.csv file was loaded in RStudio for an initial assessment.

The dataset contains 365,069 rows and 12 columns. The column datatypes correspond to the table presented above.

```
── Data Summary ────────────────────────
                           Values
Name                       Q12019
Number of rows             365069
Number of columns          12
_________________________
Column type frequency:
  character                4
  numeric                  6
  POSIXct                  2
_________________________
Group variables            None
```

Some general comments about the initial analysis of the data:

- The variables gender and birthyear contain missing values.
- The variable gender contains two possible values (Male or Female), the same as usertype (Subscriber or Customer).
- The minimum value of birthyear is 1900 which seems unlikely.
- The maximum trip duration corresponds to 10628400 seconds ~ 123 days which also seems unlikely.

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>
1	from_station_name	0	1.0000000	10	43	0	594
2	to_station_name	0	1.0000000	10	43	0	600
3	usertype	0	1.0000000	8	10	0	2
4	gender	19711	0.9460075	4	6	0	2

	skim_variable	n_missing	p0	p25	p50	p75	p100
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	trip_id	0	21742443	21848765	21961829	22071823	22178528
2	bikeid	0	1	1777	3489	5157	6471
3	tripduration	0	61	326	524	866	10628400
4	from_station_id	0	2	76	170	287	665
5	to_station_id	0	2	76	168	287	665
6	birthyear	18023	1900	1975	1985	1990	2003

Process

In this part of the process, any cleaning or manipulation of data will be documented.

For this part of the process, two roads were taken, one using Microsoft Excel and another using R.

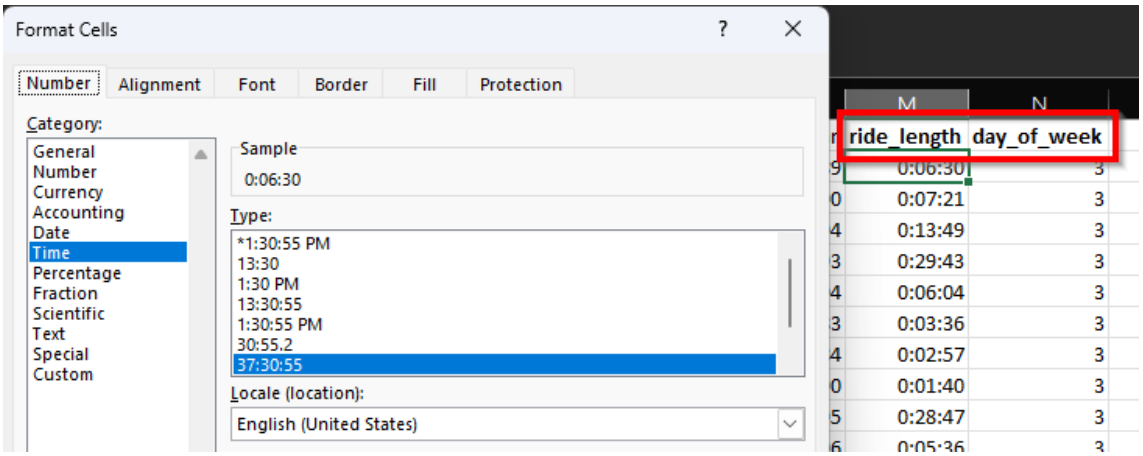
Microsoft Excel

The CSV files were converted to XLSX files. The files obtained are the following:

- Divvy_Trips_2019_Q1.xlsx
- Divvy_Trips_2019_Q4.xlsx
- Divvy_Trips_2020_Q1.xlsx

Then, for each of the files, the next steps were applied:

- Creation of a new column called ride_length.
- Application of a formula that subtracts start_time minus end_time.
- Application of this particular format: HH:MM:SS to the results.
- Creation of a new column called day_of_week.
- Application of a formula that brings the day of the week (=WEEKDAY()) of the start_time.



R

The following logic was applied in the R markdown file:

- The five CSV files were loaded and merged into a single dataframe.
- The variable tripduration already existed in four of the CSV files. In the one CSV file where it didn't exist, the column was obtained by applying the formula that subtracts start_time minus end_time:

```
#Calculate the value of the trip duration
Q12020 <- Q12020 %>%
  mutate(tripduration = as.numeric(difftime(end_time, start_time, units = "secs")))
```

- Creation of the day_of_week variable:

```
Cyclistic <- Cyclistic %>%
  mutate(day_of_week = weekdays(start_time))
```

```
head(Cyclistic)
```

A tibble: 6 × 12

to_station_id <dbl>	to_station_name <chr>	usertype <chr>	gender <chr>	birthyear <dbl>	day_of_week <chr>
84	Milwaukee Ave & Grand Ave	Subscriber	Male	1989	Tuesday
624	Dearborn St & Van Buren St (*)	Subscriber	Female	1990	Tuesday
644	Western Ave & Fillmore St (*)	Subscriber	Female	1994	Tuesday
176	Clark St & Elm St	Subscriber	Male	1993	Tuesday
35	Streeter Dr & Grand Ave	Subscriber	Male	1994	Tuesday
49	Dearborn St & Monroe St	Subscriber	Female	1983	Tuesday

6 rows | 7-12 of 12 columns

- Creation of the start_hour variable:

```r
Cyclistic <- Cyclistic %>%
  mutate(start_hour = hour(start_time))
```

```r
Cyclistic %>%
  select (start_time, start_hour)
```

A tibble: 4,244,774 x 2

start_time <S3: POSIXct>	start_hour <int>
2019-01-01 17:12:00	17
2019-01-01 17:15:00	17
2019-01-01 17:15:00	17
2019-01-01 17:15:00	17
2019-01-01 17:16:00	17
2019-01-01 17:17:00	17
2019-01-01 17:18:00	17

-

Analyze

Microsoft Excel

For the first part of the Analyze step, some pivot tables were created in the following Excel files:

- Divvy_Trips_2019_Q1.xlsx (2019 Q1)
- Divvy_Trips_2020_Q1.xlsx (2020 Q1)

Some images for each file are being considered in the next table:

<table>
<tr><th colspan="3">2019 Q1</th><th colspan="3">2020 Q1</th></tr>
<tr><th colspan="6">Number of rides per day</th></tr>
<tr><td colspan="3"><table>
<tr><th>Row Labels</th><th>Number of rides</th><th>Percentage</th></tr>
<tr><td>Sunday</td><td>27,999.00</td><td>7.67%</td></tr>
<tr><td>Monday</td><td>50,399.00</td><td>13.81%</td></tr>
<tr><td>Tuesday</td><td>61,005.00</td><td>16.71%</td></tr>
<tr><td>Wednesday</td><td>60,414.00</td><td>16.55%</td></tr>
<tr><td>Thursday</td><td>66,903.00</td><td>18.33%</td></tr>
<tr><td>Friday</td><td>63,047.00</td><td>17.27%</td></tr>
<tr><td>Saturday</td><td>35,302.00</td><td>9.67%</td></tr>
<tr><td>Grand Total</td><td>365,069.00</td><td>100.00%</td></tr>
</table></td><td colspan="3"><table>
<tr><th>Row Labels</th><th>Number of rides</th><th>Percentage</th></tr>
<tr><td>Sunday</td><td>50,850.00</td><td>11.91%</td></tr>
<tr><td>Monday</td><td>66,778.00</td><td>15.64%</td></tr>
<tr><td>Tuesday</td><td>74,961.00</td><td>17.56%</td></tr>
<tr><td>Wednesday</td><td>69,911.00</td><td>16.38%</td></tr>
<tr><td>Thursday</td><td>66,140.00</td><td>15.49%</td></tr>
<tr><td>Friday</td><td>60,663.00</td><td>14.21%</td></tr>
<tr><td>Saturday</td><td>37,584.00</td><td>8.80%</td></tr>
<tr><td>Grand Total</td><td>426,887.00</td><td>100.00%</td></tr>
</table></td></tr>
<tr><th colspan="6">Average ride time per day (member or subscribers)</th></tr>
<tr><td colspan="3"><table>
<tr><td>usertype</td><td>Subscriber</td></tr>
<tr><th>Row Labels</th><th>Average of ride_length</th></tr>
<tr><td>Sunday</td><td>0:16:48</td></tr>
<tr><td>Monday</td><td>0:14:38</td></tr>
<tr><td>Tuesday</td><td>0:14:22</td></tr>
<tr><td>Wednesday</td><td>0:12:06</td></tr>
<tr><td>Thursday</td><td>0:12:01</td></tr>
<tr><td>Friday</td><td>0:13:53</td></tr>
<tr><td>Saturday</td><td>0:16:59</td></tr>
<tr><td>Grand Total</td><td>0:13:54</td></tr>
</table></td><td colspan="3"><table>
<tr><td>member_casu</td><td>member</td></tr>
<tr><th>Row Labels</th><th>Average of ride_length</th></tr>
<tr><td>Sunday</td><td>0:15:49</td></tr>
<tr><td>Monday</td><td>0:12:59</td></tr>
<tr><td>Tuesday</td><td>0:11:32</td></tr>
<tr><td>Wednesday</td><td>0:11:40</td></tr>
<tr><td>Thursday</td><td>0:11:33</td></tr>
<tr><td>Friday</td><td>0:12:37</td></tr>
<tr><td>Saturday</td><td>0:15:30</td></tr>
<tr><td>Grand Total</td><td>0:12:41</td></tr>
</table></td></tr>
<tr><th colspan="6">Average ride time per day (customer or casual)</th></tr>
</table>

usertype	Customer
Row Labels	**Average of ride_length**
Sunday	0:41:35
Monday	0:44:27
Tuesday	0:40:27
Wednesday	0:51:57
Thursday	2:13:48
Friday	0:59:54
Saturday	1:00:20
Grand Total	**1:01:57**

member_casual	casual
Row Labels	**Average of ride_length**
Sunday	1:35:11
Monday	1:13:54
Tuesday	1:24:38
Wednesday	1:15:00
Thursday	2:05:52
Friday	1:58:18
Saturday	1:40:17
Grand Total	**1:35:47**

2019 Q1 (Number of rides per day per)

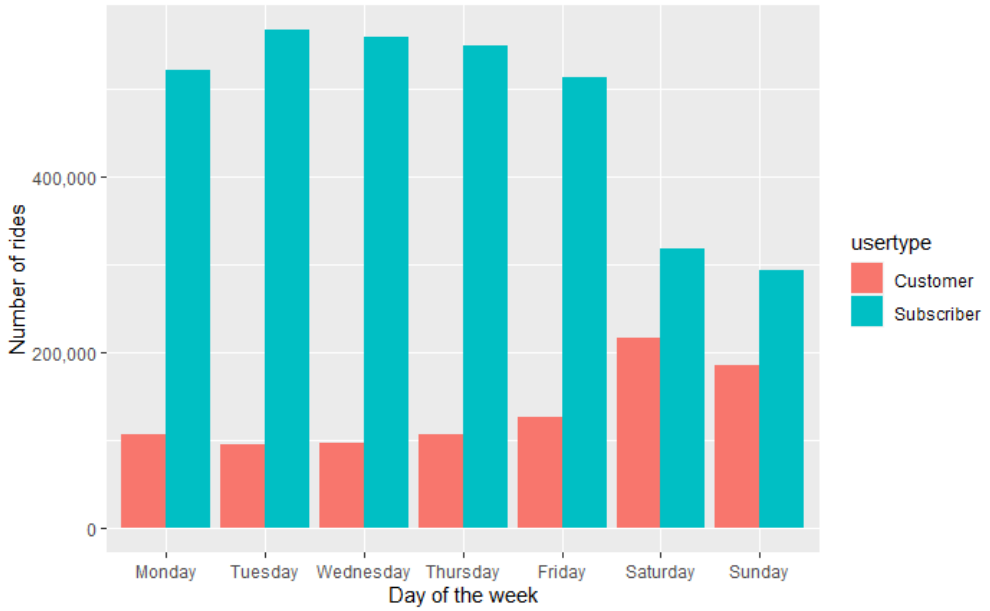
Count of trip_i	Column Labels							
Row Labels	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Grand Tota**
Customer	3,766.00	1,892.00	2,728.00	2,489.00	2,920.00	3,375.00	5,993.00	23,163.0
Subscriber	24,233.00	48,507.00	58,277.00	57,925.00	63,983.00	59,672.00	29,309.00	341,906.0
Grand Total	**27,999.00**	**50,399.00**	**61,005.00**	**60,414.00**	**66,903.00**	**63,047.00**	**35,302.00**	**365,069.0**

2020 Q1 (Number of rides per day per)

Count of ride_id	Column Labels							
Row Labels	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Grand Tota**
casual	14,886.00	4,855.00	5,264.00	5,933.00	4,895.00	5,167.00	7,480.00	48,480.0
member	35,964.00	61,923.00	69,697.00	63,978.00	61,245.00	55,496.00	30,104.00	378,407.0
Grand Total	**50,850.00**	**66,778.00**	**74,961.00**	**69,911.00**	**66,140.00**	**60,663.00**	**37,584.00**	**426,887.0**

R

A bar chart was generated to identify the number of rides per type of user throughout the week:



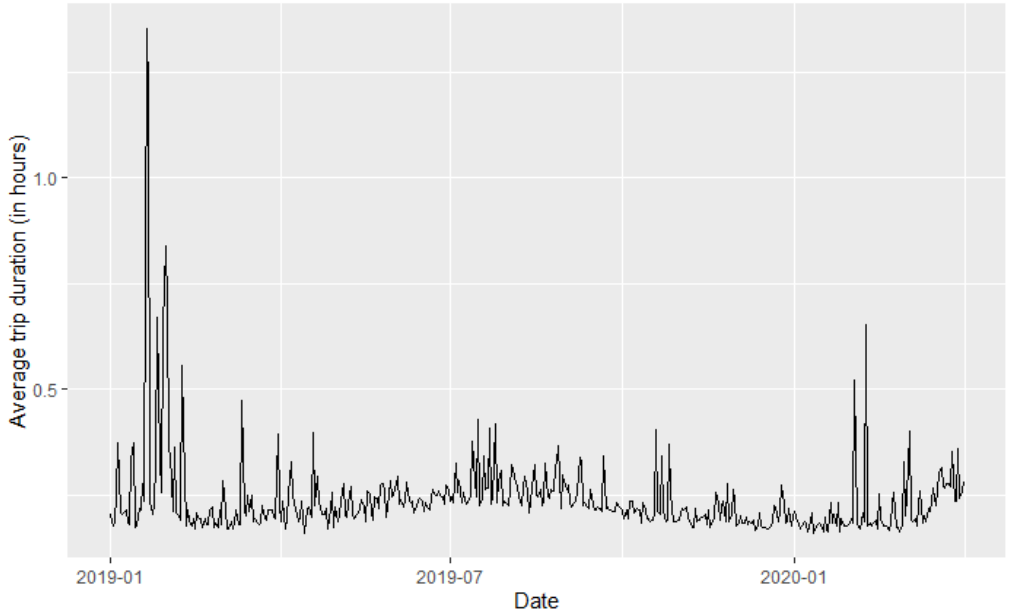
Some observations can be taken from this:

- In the number of rides, there is a clear difference between the types of users. Subscribers present a much larger number of rides.
- Both subscribers and customers present a steady distribution during weekdays and weekends. Subscribers do more rides during weekdays, whereas customers do more

rides during the weekend; however, even on weekends, the number of rides for subscribers is greater than for customers.

A line chart was generated to identify the average trip duration (in hours) during the complete dates presented in the data analyzed.

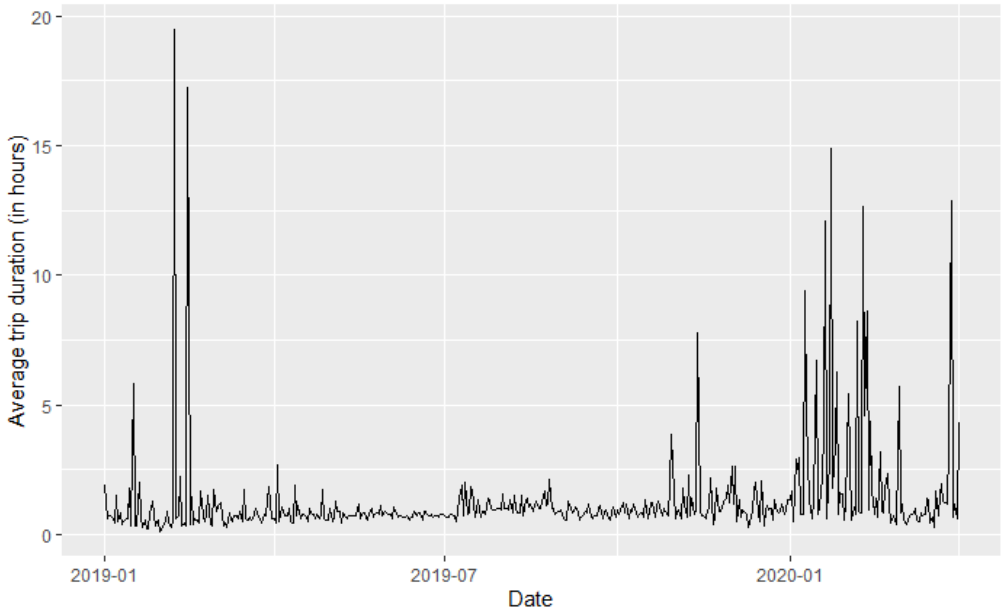
Subscribers:



day_of_week <fctr>	mean_tripduration <dbl>
Monday	0.2348917
Tuesday	0.2304404
Wednesday	0.2259868
Thursday	0.2254482
Friday	0.2292561
Saturday	0.2703796
Sunday	0.2574934

- During the timeframe of the data (5 quarters), it can be seen that the average hours per day is around the same except for a couple of peaks around the first quarter.
- On weekdays, the average hours of the subscribers is 0.225 hours ~ 13.5 minutes.
- On weekends, the average hours of the subscribers is 0.264 hours ~ 15.8 minutes.

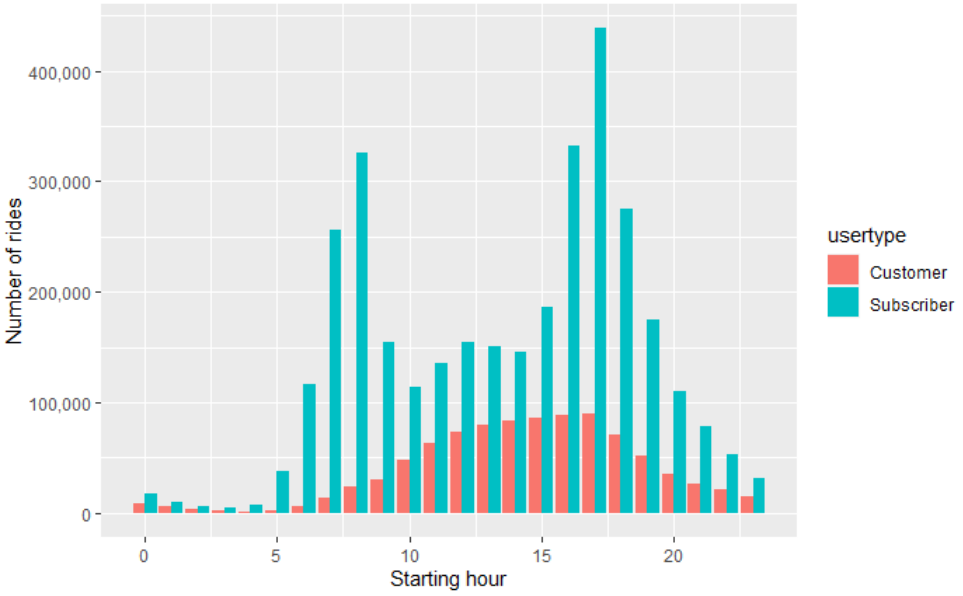
Customers:



day_of_week <fctr>	mean_tripduration <dbl>
Monday	0.9231958
Tuesday	0.9823283
Wednesday	1.0211159
Thursday	1.0500175
Friday	1.0426786
Saturday	0.9277288
Sunday	0.9886233

- During the timeframe of the data (5 quarters), it can be seen that the average hours per day is around the same except for a first group of peaks around the first quarter, and a second one around the fifth quarter.
- On weekdays, the average hours of the subscribers is 1.0036 hours ~ 60.2 minutes.
- On weekends, the average hours of the subscribers is 0.958 hours ~ 57.4 minutes.

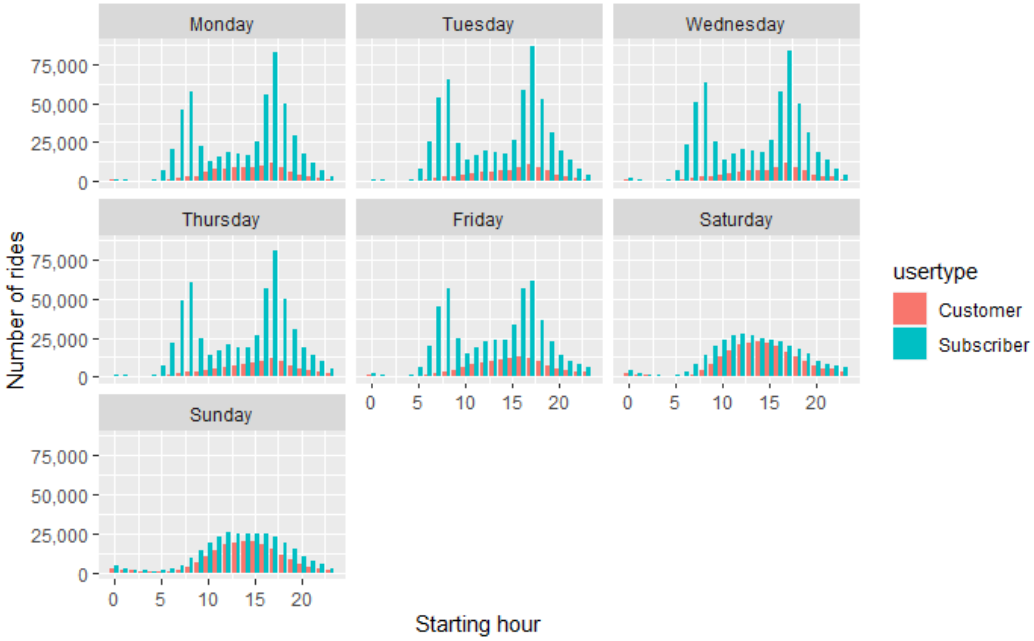
A bar chart showing the number of rides per hour and type of user was made.



Some observations can be taken from this:

- Subscribers seem to present a bimodal distribution whereas customers seem to present a normal distribution.
- The maximum number of rides for subscribers has two peaks: the first one around 7 am and 8 am, and the second one around 4 pm, 5 pm, and 6 pm.
- There is minimal activity between 12 am and 5 am.

The same report was done per day:



Some observations can be taken from this:

- Subscribers seem to present a bimodal distribution but only during weekdays. Customers seem to present a normal distribution.

- The maximum number of rides for subscribers has two peaks: the first one around 7 am and 8 am, and the second one around 4 pm, 5 pm, and 6 pm, but this is only during weekdays.
- During the weekend, the distribution of rides is the same between subscribers and customers.

Act

According to the analysis performed and the reports generated, there are some relevant conclusions:

- Subscribers and customers definitely present a different behavior when it comes to the usage of the service.
- Subscribers have on average more rides per day than customers. Subscribers have much more rides during weekdays than on weekends. The opposite applies to customers, more rides during weekends.
- Subscribers have an average time of use of 14 minutes per ride whereas customers have an average time of use of 58.5 minutes per ride.
- Subscribers present a bimodal distribution in the number of rides during weekdays. The times at which there are peaks of use are consistent with the use of the service for commuting to and from work (around 7 am and 8 am, and around 4 pm, 5 pm, and 6 pm).