Bike_share

Fazlay Rabby

2022-04-16

Contents

Future Analysis Scopes: **14**

Scenario

The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, it's crucial to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the marketing team will design a new marketing strategy to convert casual riders into annual members.

Data Source:

Cyclistic Bike Share Data-set (Apr,2021 - Mar,2022)

https://divvy-tripdata.s3.amazonaws.com/index.html

Project Objective:

Analyzing last 12 months of trip data from Cyclistic Bike Share Data-set to get insight of how Casual riders & Annual members differ, why casual riders would buy membership & how digital media could effect their marketing tactics.

Library setup for necessary packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(readr)
library(ggplot2)
```

Combine data of last 12 months into one file for further exploration

```
DF_raw <- read_csv(list.files(pattern="*tripdata.csv"))
```

```
## Rows: 5657545 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

First look at the data

```
skim_without_charts(DF_raw)
```

Table 1: Data summary

Name	DF_raw
Number of rows	5657545
Number of columns	13
Column type frequency:	
character	7
numeric	4
POSIXct	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	16	16	0	5657545	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	730842	0.87	3	53	0	861	0
start_station_id	730839	0.87	3	44	0	852	0
end_station_name	781250	0.86	10	53	0	861	0
end_station_id	781250	0.86	3	44	0	853	0
member_casual	0	1.00	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.05	41.64	41.88	41.90	41.93	45.64
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-73.80
end_lat	4853	1	41.90	0.05	41.39	41.88	41.90	41.93	42.17
end_lng	4853	1	-87.65	0.03	-88.97	-87.66	-87.64	-87.63	-87.49

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2021-02-01 00:55:44	2022-03-31 23:59:47	2021-08-14 18:03:18	4730641
ended_at	0	1	2021-02-01 01:22:48	2022-04-01 22:10:12	2021-08-14 18:27:11	4724534

Data Limitations:

In this data-set we have Trip Id but we don't have the Rider Id which may impact on the analysis. Here we can analyze the number of trips taken by Casual or Annual Members but we can't figure out how many casual members or annual members we are talking about.

Processing data

```
DF_trimmed<- DF_raw |>
  rename(subscription_type = member_casual,
         start_time=started_at,
         end_time =ended_at,
         trip_id = ride_id,
         bike_type =rideable_type,
         start_station = start_station_name,
         end_station= end_station_name) |>
  mutate(trip_length = as.numeric(end_time-start_time),
         month= month(start_time, label = TRUE),
         day = wday(start_time,label = TRUE),
         date = day(start_time),hour = hour(start_time))
```

Removing empty rows to avoid bias in analysis.

```
DF_trimmed <- drop_na(DF_trimmed)
```

Review Data

```
skim_without_charts(DF_trimmed)
```

Table 5: Data summary

Name	DF_trimmed
Number of rows	4595213
Number of columns	18
Column type frequency:	
character	7
factor	2
numeric	7
POSIXct	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
trip_id	0	1	16	16	0	4595213	0
bike_type	0	1	11	13	0	3	0
start_station	0	1	3	53	0	860	0
start_station_id	0	1	3	44	0	851	0
end_station	0	1	10	53	0	860	0
end_station_id	0	1	3	44	0	852	0
subscription_type	0	1	6	6	0	2	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
month	0	1	TRUE	12	Jul: 692321, Aug: 674409, Sep: 621150, Jun: 608778
day	0	1	TRUE	7	Sat: 807901, Sun: 710642, Fri: 650811, Wed: 627229

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.04	41.65	41.88	41.90	41.93	45.64
start_lng	0	1	-87.64	0.02	-87.83	-87.66	-87.64	-87.63	-73.80
end_lat	0	1	41.90	0.04	41.65	41.88	41.90	41.93	42.17
end_lng	0	1	-87.64	0.02	-87.83	-87.66	-87.64	-87.63	-87.52
trip_length	0	1	1291.01	10996.25	-3354.00	411.00	722.00	1309.00	3356649.00
date	0	1	15.36	8.78	1.00	8.00	15.00	23.00	31.00
hour	0	1	14.24	5.01	0.00	11.00	15.00	18.00	23.00

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_time	0	1	2021-02-01 01:07:04	2022-03-31 23:59:47	2021-08-10 12:26:57	3938305
end_time	0	1	2021-02-01 01:47:45	2022-04-01 22:10:12	2021-08-10 12:48:14	3931592

A Quick overview of how riding patter changes over the time period of a year.

```
monthly_trips <- DF_trimmed |>
  select(month, trip_id) |>
  group_by(month) |>
  summarise(number_of_trips_thoushand = n_distinct(trip_id)/1000)
```

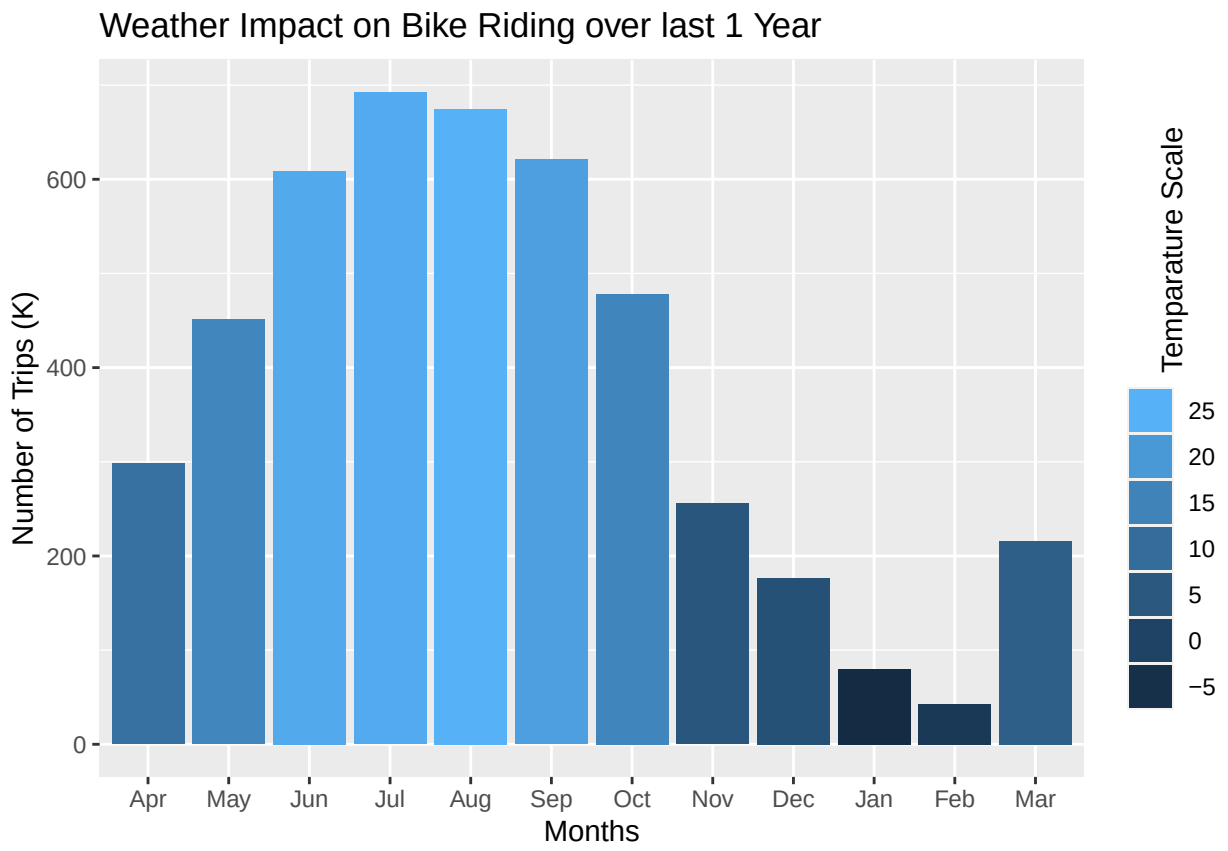
##Observation: A gradual increase is observed from Mar-2021 to Jul-2021 & from then there is also a gradual decrease which may aligned to the weather patter of Chicago City.

Adding weather record of last 12 years for comparison with number of rides.

```
temperature <- read.csv(list.files(pattern= "*temparature*")) |>
  select(month,temperature) |>
  mutate(temperature, temperature = (temperature-32)*(5/9))
#join temperature from Chicago Temperature record
monthly_trips <- left_join(temperature,monthly_trips, by = "month")
```

How the weather/temperature impact on the riding pattern?

```
monthly_trips <- monthly_trips
monthly_trips$month <- factor(monthly_trips$month,
                              levels = c("Apr", "May", "Jun", "Jul", "Aug","Sep", "Oct", "Nov", "Dec",
ggplot(data= monthly_trips,
       mapping = aes(x=month, y= number_of_trips_thoushand,fill = temperature)) +
  geom_col() +
  labs(x= "Months", y= "Number of Trips (K)",
       title = "Weather Impact on Bike Riding over last 1 Year") +
  theme(legend.position = "right", legend.title=element_text(angle = 90)) +
  guides(fill = guide_legend(title = "Temparature Scale", reverse = TRUE))
```



Observation

It's clear that riding tendency increase with the temperature & decrease again as it gets colder in that locality.

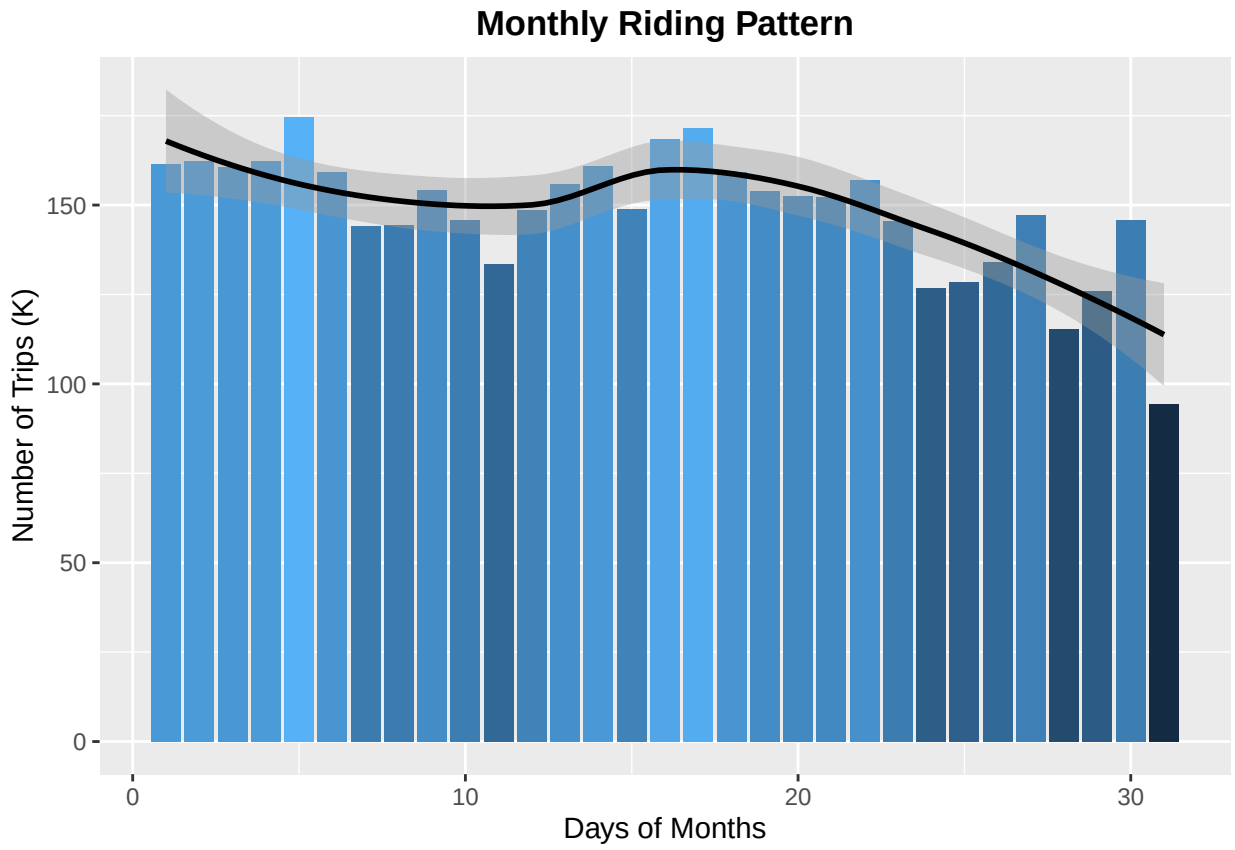
What about month, week & daily pattern..?

```
#Average monthly riding pattern
days_of_month <- DF_trimmed |>
  select(date,trip_id) |>
  group_by(date) |>
  summarise(number_of_trips_thoushand = n_distinct(trip_id)/1000)

ggplot(days_of_month, mapping = aes(x= date,y=number_of_trips_thoushand, fill = number_of_trips_thoushar
  geom_col() +
  geom_smooth(color = "black") +
  labs(x= "Days of Months",
       y= "Number of Trips (K)",
       title = "Monthly Riding Pattern")+
  theme(legend.position = "right",
        legend.title=element_text(angle = 90)) +
  guides(fill = guide_legend(title = "Avg Trips/Day", reverse = TRUE))+
```

```
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", hjust = 0.5))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Observation:

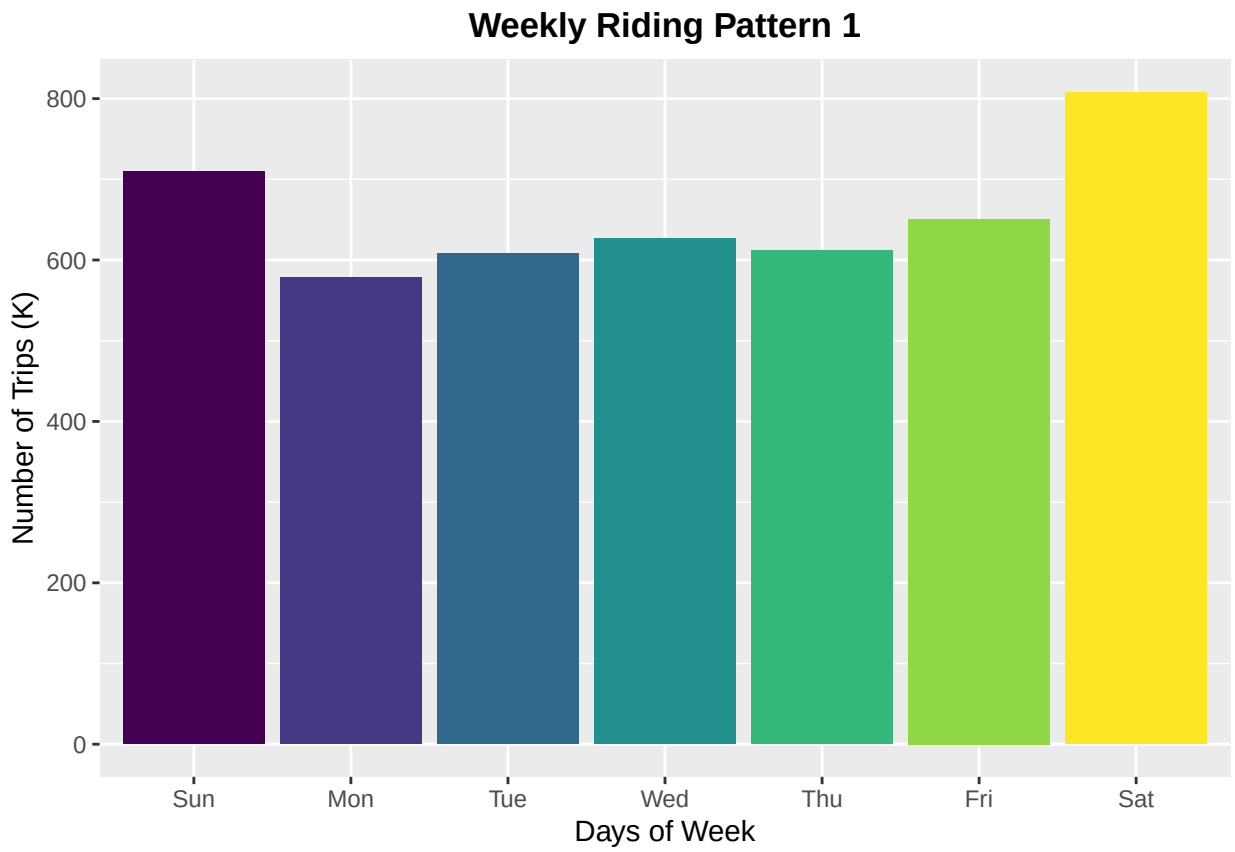
The result show a downtrend on last 10 days of a given month according to the trip record compared to the first 20 days.

Average weekly riding pattern analysis

```
days_of_week <- DF_trimmed |>
  select(day,trip_id, subscription_type) |>
  group_by(day, subscription_type) |>
  summarise(number_of_trips_thoushand = n_distinct(trip_id)/1000)
```

```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
## argument.
```

```
ggplot(days_of_week,
       mapping = aes(x= day,y=number_of_trips_thoushand, fill = day))+
  geom_col()+
  labs(x= "Days of Week",
       y= "Number of Trips (K)",
       title = "Weekly Riding Pattern 1")+
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", hjust = 0.5))
```



Observation:

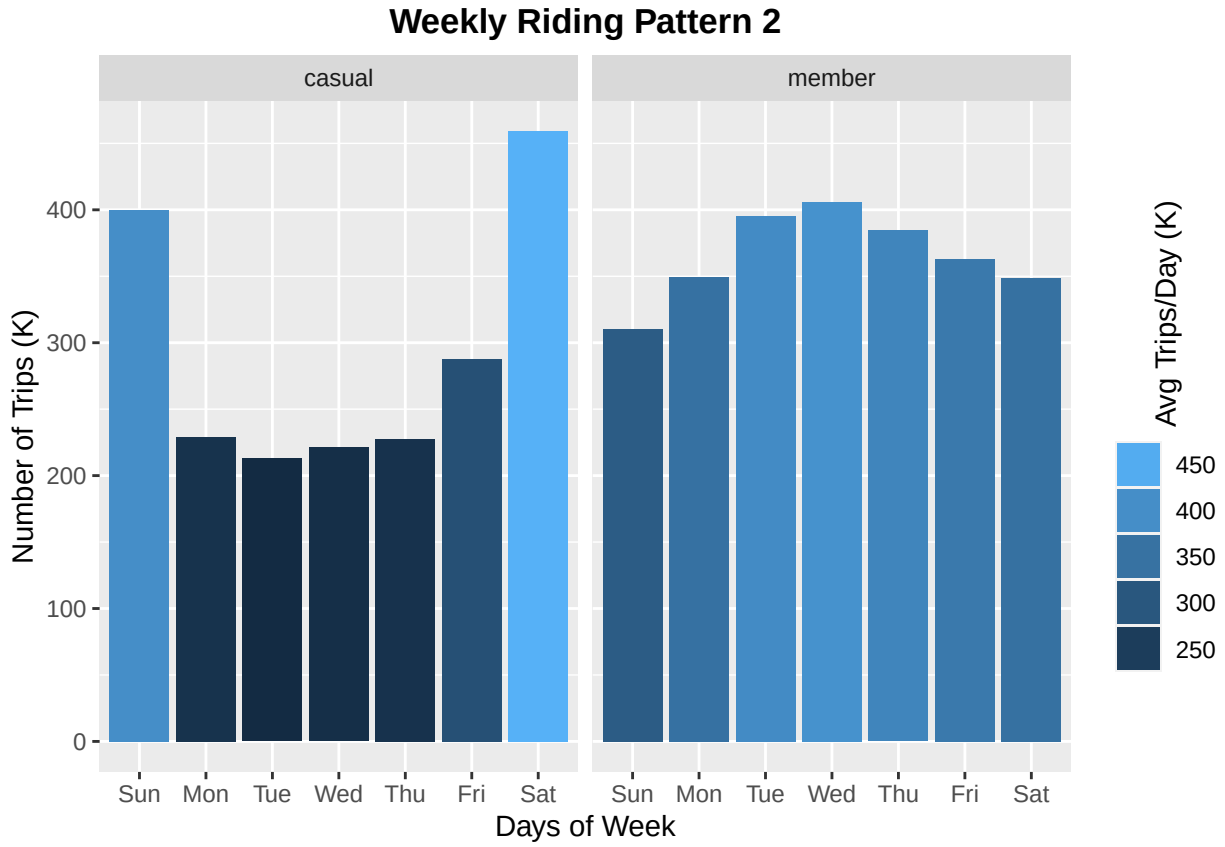
The highest trips are made on weekends. Is it the same for both types of riders...?

Further drill down to differentiate casual vs annual riders riding patter on weekly basis

```
#Average weekly riding pattern by riders type
days_of_week_2 <- DF_trimmed |>
  select(day,trip_id,subscription_type) |>
  group_by(day, subscription_type) |>
  summarise(number_of_trips_thoushand = n_distinct(trip_id)/1000)
```

```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
## argument.
```

```
ggplot(days_of_week_2, mapping = aes(x= day,y=number_of_trips_thoushand, fill = number_of_trips_thoushan
  geom_col()+
  facet_wrap(~subscription_type)+
  labs(x= "Days of Week",
       y= "Number of Trips (K)",
       title = "Weekly Riding Pattern 2")+
  theme(legend.position = "right",
        legend.title=element_text(angle = 90),
        plot.title = element_text(face = "bold", hjust = 0.5)) +
  guides(fill = guide_legend(title = "Avg Trips/Day (K)", reverse = TRUE))
```




```
# Data is showing that casual riders ride bikes on weekend in higher number
```

Observation:

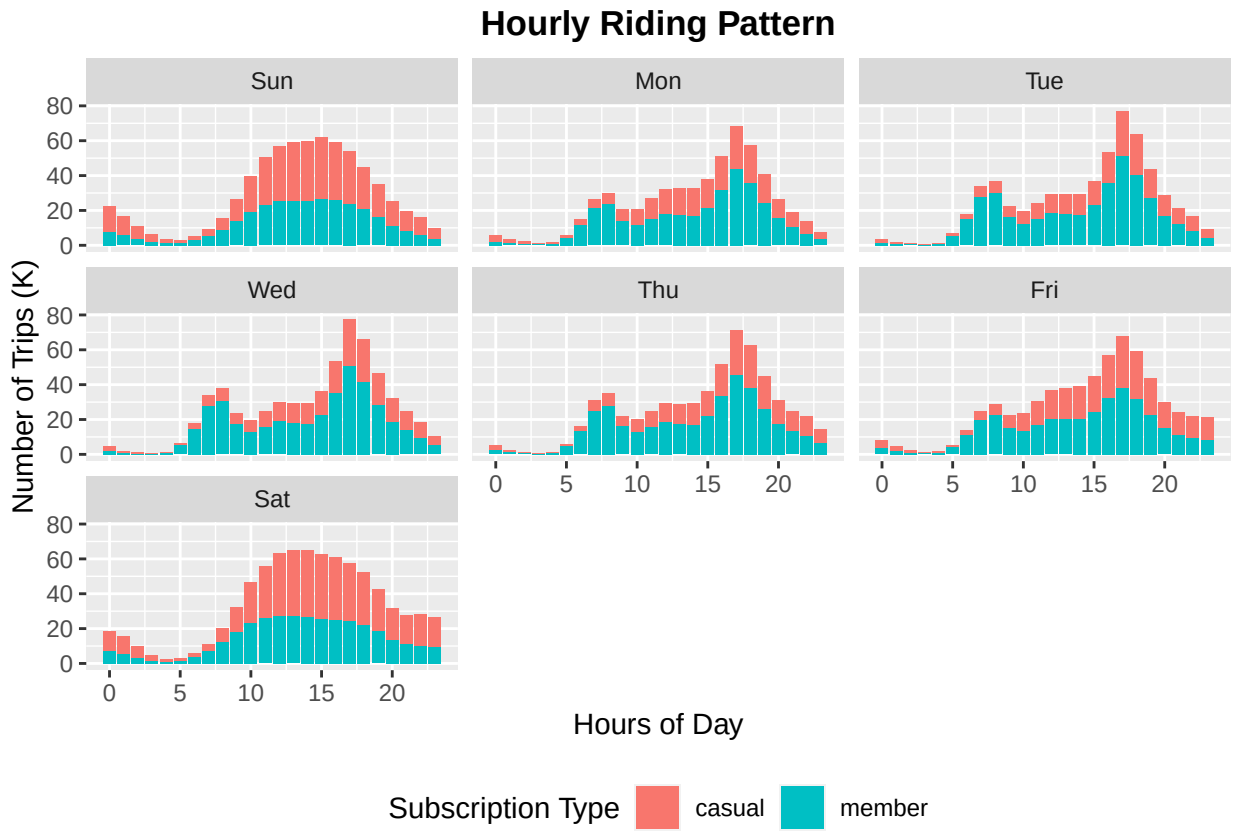
Here we can see an interesting difference in riding patter. Annual members more likely to use bikes on week days apparently as commute to the work but Casual riders use bike mostly on weekends (Sat & Sun) which seems like recreational.

In this stage of analysis we shall look at Hourly patter all through the week

```
hourly <- DF_trimmed |>
  select(hour,trip_id,subscription_type, day) |>
  group_by(hour, subscription_type, day) |>
  summarise(number_of_trips_thoushand = n_distinct(trip_id)/1000)
```

```
## `summarise()` has grouped output by 'hour', 'subscription_type'. You can
## override using the `.groups` argument.
```

```
ggplot(hourly, mapping = aes(x= hour,y=number_of_trips_thoushand, fill= subscription_type))+
  geom_col()+
  facet_wrap(~day)+
  labs(x= "Hours of Day",
       y= "Number of Trips (K)",
       title = "Hourly Riding Pattern")+
  theme(legend.position = "bottom",
        plot.title = element_text(face = "bold", hjust = 0.5)) +
  guides(fill = guide_legend(title = "Subscription Type"))
```



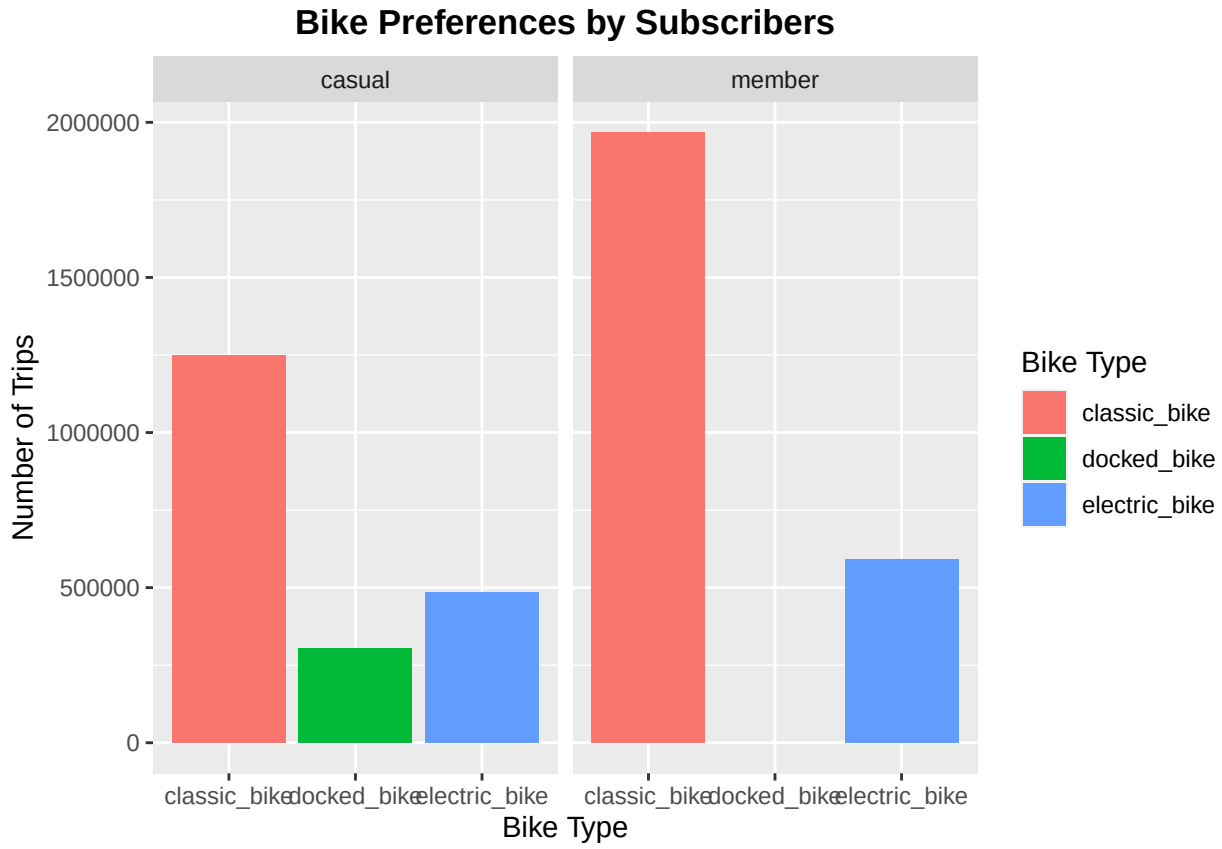
Observation:

As earlier we observed the riding patter of Casual members, here we see nice and smooth increasing & decreasing pattern on weekends which also signifies the number of trips made by casual members are different

than the other week days of the week.

Bike Preference

```
bike_type <- DF_trimmed |>
  count(bike_type,subscription_type)
ggplot(bike_type, aes(x= bike_type,y=n, fill = bike_type)) +
  geom_col()+
  facet_wrap(~subscription_type)+
  labs(x= "Bike Type", y= "Number of Trips",
       title = "Bike Preferences by Subscribers")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5)) +
  guides(fill = guide_legend(title = "Bike Type"))
```



Observation

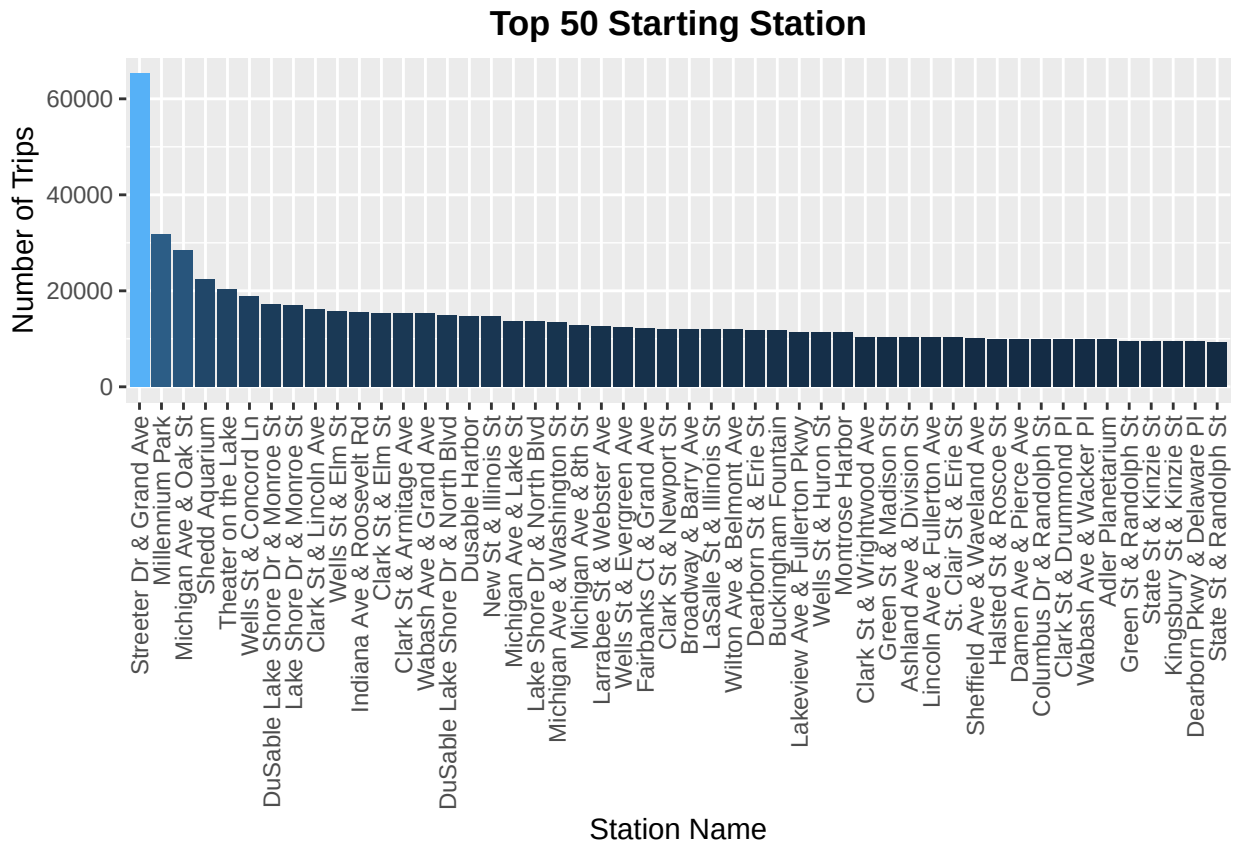
Seems like docked bike is popular only to the Casual Riders & a significant difference is usage of Classic Bike

Let's have a glimpse of hot-spots/popular starting points for casual bikers

```
popular_start_station <- DF_trimmed |>
  filter(subscription_type == "casual") |>
  select(start_station, subscription_type, trip_id, start_lat,start_lng) |>
  group_by(start_station) |>
  summarise(start_count = as.numeric(n_distinct(trip_id))) |>
  arrange(desc(start_count)) |>
  top_n(50)
```

```
## Selecting by start_count
```

```
popular_start_station |>
  ggplot(aes(x= reorder(start_station,-start_count),
             y= start_count, fill= start_count))+
  geom_col()+guides(x = guide_axis(angle = 90))+
  labs(x= "Station Name",
       y= "Number of Trips",
       title = "Top 50 Starting Station")+
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", hjust = 0.5))
```



Observation:

Here we have made a list of Top 50 Start Station based for casual riders based on previous record which can be used as top priority to launch the marketing campaign

Important Insights from the Analysis

- The main focus of the digital campaign shall be to spread the awareness of benefits of riding bikes on everyday (instead of weekend only)
- From April we shall start to provide early bird offers
- Saturday & Sunday is the best time to campaign if we want to best use our resources
- We have a list of top 50 popular Start Stations from where Casual Members start ride which can be on our priority list to launch the campaign

Future Analysis Scopes:

- Once the campaign is launched we can collect riders Demographic information to further analyse by Age group, gender etc & only then we can compare exactly how many casual/annual riders are riding our bikes & better understand their riding pattern.
- Cyclistic Management may take initiative to introduce a Customized Mobile Application for the Riders with can be helpful for both riders & the company.