Cyclistic Data Analysis - Yearly Overview

Gayatri Paul

2023-01-06

1. Introduction

Cyclistic is a bike share company launched in 2016. It has more than 5000 bicycles locked into more than 600 stations across Chicago. Bikes can be unlocked from one station and returned to any station in the system. Cyclistic has flexible pricing plans: single ride passes, full-day passes and annual memberships. Customers with annual memberships are called Cyclistic members and others with passes are called casual riders.

Cyclistic's finance analysts have concluded that annual members are more profitable and maximizing the number of annual memberships will be key to growth. Mareno- the director of marketing believes that rather than creating a campaign to target new customers, we can try to convert existing casual riders into members as they are already aware of the program and have chosen Cyclistic for their mobility needs.

The goal here is to design marketing strategies aimed at converting casual riders into annual members. For that, we need to understand how members and casual riders differ, why would they opt for an annual membership and how digital media be used to influence their decision?

Pricing:

Classic Bike:

Single Ride: $3.30 - Ride for 30 mins - Delay charges: $0.15/minute Day Pass: $15 - Ride for 180 mins - Delay charges: $0.16/minute Annual Membership: $108 per year, 45 mins each ride - Delay charges: $0.15/minute

Electric Bike:

Casual riders: $1 to unlock + $0.39 per minute Members: $0 to unlock + $0.16 per minute

In this case study, we will analyze the key differences between casual riders and annual members and provide top 3 recommendations based on our analysis. The recommendations will be then shared with the stakeholders- Lily Moreno, the director of marketing and Cyclistic executive team.

Data Source:

The datasets have been collected from following website:

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

The data has been made available by Motivate International Inc. under this license:

https://ride.divvybikes.com/data-license-agreement (https://ride.divvybikes.com/data-license-agreement)

We are using the data from October 2021 to October 2022.

2. Prepare

● How is the data organized?

The data has been arranged in Wide format. With following columns:

Column Name Description

ride_id:	Ride ID (Alphanumeric of varying lengths)
rideable_type:	Bike Type (Classic / Dock/ Electric)
started_at:	Date and time @ the start of the ride (format: mm/dd/yyyy hh:mm)
ended_at:	Date and time @ the end of the ride (format: mm/dd/yyyy hh:mm)
start_station_name:	Name of the station where ride started
start_station_id:	ID of the station where ride started
end_station_name:	Name of the station where ride ended
end_station_id:	ID of the station where ride ended
start_lat:	Latitude where the ride started
start_lng:	Longitude where the ride started
end_lat:	Latitude where the ride ended
end_lng:	Longitude where the ride ended
member_casual:	Rider type. 1-Casual, 2-Member

● Are there issues with bias or credibility in this data?

As the data has been directly downloaded from the website and the records are directly taken from tracked rides, there are no bias or credibility issues in the data. The data is ROCCC: Reliable, Original, Comprehensive, Current and Cited.

● How did you verify the data's integrity?

The data source and license to use the data for non-commercial purposes has been provided and linked in the document. The data does not include any personal information such as the rider's name, address or their credit card details. Hence it abides by privacy and security regulations.

● How does it help you answer the question?

The data contains information regarding type of rider and number of trips. We know the prices of each type of pass; hence we can utilize this information for our analysis and gain valuable insights.

● Are there any problems with the data?

1. Some columns of the data sets contain missing information.
2. The data does not contain any information which can help us understand how many times the same person used the bike service. This would have been beneficial for the analysis.
3. There are a few rides for which recorded bike return time is before the start time, we will have to remove those rows.
4. There is two options available for casual riders. Full day pass and single ride pass. There is no data which can tell us which pass was used for the ride. This data is crucial as it will tell us how much money the rider spent on the ride.

3. Process

The data from October 2021 to October 2022 has been downloaded. Each file is in csv format. Let's import it to R.

```
oct_21 <- read.csv("C:/Users/paulg/Downloads/csv/1_oct_21.csv", header=T, na.strings=c
("","NA"))

nov_21 <- read.csv("C:/Users/paulg/Downloads/csv/2_nov_21.csv", header=T, na.strings=c
("","NA"))

dec_21 <- read.csv("C:/Users/paulg/Downloads/csv/3_dec_21.csv", header=T, na.strings=c
("","NA"))

jan_22 <- read.csv("C:/Users/paulg/Downloads/csv/4_jan_22.csv", header=T, na.strings=c
("","NA"))

feb_22 = read.csv("C:/Users/paulg/Downloads/csv/5_feb_22.csv", header=T, na.strings=c("","
NA"))

mar_22 = read.csv("C:/Users/paulg/Downloads/csv/6_mar_22.csv", header=T, na.strings=c("","
NA"))

apr_22 = read.csv("C:/Users/paulg/Downloads/csv/7_apr_22.csv", header=T, na.strings=c("","
NA"))

may_22 = read.csv("C:/Users/paulg/Downloads/csv/8_may_22.csv", header=T, na.strings=c("","
NA"))

jun_22 = read.csv("C:/Users/paulg/Downloads/csv/9_jun_22.csv", header=T, na.strings=c("","
NA"))

jul_22 = read.csv("C:/Users/paulg/Downloads/csv/10_july_22.csv", header=T, na.strings=c
("","NA"))

aug_22 = read.csv("C:/Users/paulg/Downloads/csv/11_aug_22.csv", header=T, na.strings=c
("","NA"))

sept_22 = read.csv("C:/Users/paulg/Downloads/csv/12_sept_22.csv", header=T, na.strings=c
("","NA"))

oct_22 = read.csv("C:/Users/paulg/Downloads/csv/13_oct_22.csv", header=T, na.strings=c
("","NA"))
```

All the data frames contain 13 variables, lets combine them for the purpose of analysis.

```
df= rbind(oct_21, nov_21, dec_21, jan_22, feb_22, mar_22, apr_22, may_22, jun_22, jul_22,
aug_22, sept_22, oct_22)
head(df)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 620BC6107255BF4C electric_bike 2021-10-22 12:46:42 2021-10-22 12:49:50
## 2 4471C70731AB2E45 electric_bike 2021-10-21 09:12:37 2021-10-21 09:14:14
## 3 26CA69D43D15EE14 electric_bike 2021-10-16 16:28:39 2021-10-16 16:36:26
## 4 362947F0437E1514 electric_bike 2021-10-16 16:17:48 2021-10-16 16:19:03
## 5 BB731DE2F2EC51C5 electric_bike 2021-10-20 23:17:54 2021-10-20 23:26:10
## 6 7176307BBC097313 electric_bike 2021-10-21 16:57:37 2021-10-21 17:11:58
##         start_station_name start_station_id end_station_name end_station_id
## 1 Kingsbury St & Kinzie St     KA1503000043             <NA>           <NA>
## 2                     <NA>             <NA>             <NA>           <NA>
## 3                     <NA>             <NA>             <NA>           <NA>
## 4                     <NA>             <NA>             <NA>           <NA>
## 5                     <NA>             <NA>             <NA>           <NA>
## 6                     <NA>             <NA>             <NA>           <NA>
##   start_lat start_lng end_lat end_lng member_casual
## 1  41.88919  -87.6385   41.89  -87.63        member
## 2  41.93000  -87.7000   41.93  -87.71        member
## 3  41.92000  -87.7000   41.94  -87.72        member
## 4  41.92000  -87.6900   41.92  -87.69        member
## 5  41.89000  -87.7100   41.89  -87.69        member
## 6  41.89000  -87.7100   41.93  -87.70        member
```

```
print(nrow(df))
```

```
## [1] 6386920
```

```
colnames(df)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
summary(df)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:6386920     Length:6386920     Length:6386920     Length:6386920
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:6386920     Length:6386920     Length:6386920     Length:6386920
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.64   Min.   :-87.84   Min.   :41.39   Min.   :-88.97
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.65   Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :45.64   Max.   :-73.80   Max.   :42.37   Max.   :-87.30
##                                   NA's   :6319    NA's   :6319
##  member_casual
##  Length:6386920
##  Class :character
##  Mode  :character
##
##
##
##
```

Rename the column named “member_casual” to “rider_type” for clarity.

```
library(dplyr)
df=rename(df,"rider_type"="member_casual")
head(df)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 620BC6107255BF4C electric_bike 2021-10-22 12:46:42 2021-10-22 12:49:50
## 2 4471C70731AB2E45 electric_bike 2021-10-21 09:12:37 2021-10-21 09:14:14
## 3 26CA69D43D15EE14 electric_bike 2021-10-16 16:28:39 2021-10-16 16:36:26
## 4 362947F0437E1514 electric_bike 2021-10-16 16:17:48 2021-10-16 16:19:03
## 5 BB731DE2F2EC51C5 electric_bike 2021-10-20 23:17:54 2021-10-20 23:26:10
## 6 7176307BBC097313 electric_bike 2021-10-21 16:57:37 2021-10-21 17:11:58
##         start_station_name start_station_id end_station_name end_station_id
## 1 Kingsbury St & Kinzie St     KA1503000043             <NA>           <NA>
## 2                     <NA>             <NA>             <NA>           <NA>
## 3                     <NA>             <NA>             <NA>           <NA>
## 4                     <NA>             <NA>             <NA>           <NA>
## 5                     <NA>             <NA>             <NA>           <NA>
## 6                     <NA>             <NA>             <NA>           <NA>
##   start_lat start_lng end_lat end_lng rider_type
## 1  41.88919  -87.6385   41.89  -87.63     member
## 2  41.93000  -87.7000   41.93  -87.71     member
## 3  41.92000  -87.7000   41.94  -87.72     member
## 4  41.92000  -87.6900   41.92  -87.69     member
## 5  41.89000  -87.7100   41.89  -87.69     member
## 6  41.89000  -87.7100   41.93  -87.70     member
```

Remove duplicate Rows

```
df=df[!duplicated(df), ]
print(nrow(df))
```

```
## [1] 6386920
```

Number of rows remained the same, which indicates that the data frame contained no duplicate data.

Find columns with missing information.

```
colSums(is.na(df))
```

```
##            ride_id      rideable_type         started_at           ended_at
##                  0                  0                  0                  0
## start_station_name   start_station_id   end_station_name     end_station_id
##             986387             986387            1054844            1054844
##          start_lat          start_lng            end_lat            end_lng
##                  0                  0               6319               6319
##         rider_type
##                  0
```

Columns namely: start_station_name, start_station_id, end_station_name, end_station_id, end_lat, end_lng contain missing values. As we are not using these columns for analysis, we will keep them as they are. If we decide to use these columns, we will first remove rows with NA and then proceed further.

Convert date columns to proper date formats & split date and time into 2 columns.

```
library(lubridate)

df$started_at <- ymd_hms(df$started_at)
df$ended_at <- ymd_hms(df$ended_at)

df$start_date <- as.Date(df$started_at)
df$start_time <- format(as.POSIXct(df$started_at), format = "%H:%M:%S")

df$end_date <- as.Date(df$ended_at)
df$end_time <- format(as.POSIXct(df$ended_at), format = "%H:%M:%S")
```

Calculate ride length in mins

```
df$ride_length <- round(difftime(df$ended_at,df$started_at, units = "mins"), 2)

df$ride_length <- as.numeric(df$ride_length)
```

keep rows with only positive ride lengths

```
df <- filter(df, ride_length > 0)
head(df)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 620BC6107255BF4C electric_bike 2021-10-22 12:46:42 2021-10-22 12:49:50
## 2 4471C70731AB2E45 electric_bike 2021-10-21 09:12:37 2021-10-21 09:14:14
## 3 26CA69D43D15EE14 electric_bike 2021-10-16 16:28:39 2021-10-16 16:36:26
## 4 362947F0437E1514 electric_bike 2021-10-16 16:17:48 2021-10-16 16:19:03
## 5 BB731DE2F2EC51C5 electric_bike 2021-10-20 23:17:54 2021-10-20 23:26:10
## 6 7176307BBC097313 electric_bike 2021-10-21 16:57:37 2021-10-21 17:11:58
##         start_station_name start_station_id end_station_name end_station_id
## 1 Kingsbury St & Kinzie St     KA1503000043             <NA>           <NA>
## 2                     <NA>             <NA>             <NA>           <NA>
## 3                     <NA>             <NA>             <NA>           <NA>
## 4                     <NA>             <NA>             <NA>           <NA>
## 5                     <NA>             <NA>             <NA>           <NA>
## 6                     <NA>             <NA>             <NA>           <NA>
##   start_lat start_lng end_lat end_lng rider_type start_date start_time
## 1  41.88919  -87.6385   41.89  -87.63     member 2021-10-22   12:46:42
## 2  41.93000  -87.7000   41.93  -87.71     member 2021-10-21   09:12:37
## 3  41.92000  -87.7000   41.94  -87.72     member 2021-10-16   16:28:39
## 4  41.92000  -87.6900   41.92  -87.69     member 2021-10-16   16:17:48
## 5  41.89000  -87.7100   41.89  -87.69     member 2021-10-20   23:17:54
## 6  41.89000  -87.7100   41.93  -87.70     member 2021-10-21   16:57:37
##     end_date end_time ride_length
## 1 2021-10-22 12:49:50        3.13
## 2 2021-10-21 09:14:14        1.62
## 3 2021-10-16 16:36:26        7.78
## 4 2021-10-16 16:19:03        1.25
## 5 2021-10-20 23:26:10        8.27
## 6 2021-10-21 17:11:58       14.35
```

There are rides which lasted less than a minute which seems odd but, but as we don't have relevant data or source of information which can be used to confirm that these entries are wrong, we will assume that these

are correct and continue with the analysis.

Calculate number of days the ride lasted

```
df$no_of_days <- as.numeric(difftime(df$end_date, df$start_date, units = "days" )+1)

count(filter(df, no_of_days > 1))
```

```
##       n
## 1 38417
```

38417 rides lasted more than a day. If we assume this data is correct, these users had to be charged extra, which means their ride must costed a lot, this case can be used to advocate annual memberships.

Calculate day of week for the ride

```
df$day_of_week_num <- wday(df$start_date)

df <- df %>%
  mutate(day_of_week=
           ifelse(day_of_week_num==1,"sunday",
                  ifelse(day_of_week_num==2, "monday",
                         ifelse(day_of_week_num==3, "tuesday",
                                ifelse(day_of_week_num==4, "wednesday",
                                       ifelse(day_of_week_num==5, "thursday",
                                              ifelse(day_of_week_num==6, "friday",
                                                     "saturday" )))))))
```

Add month and Season columns.

```
library(tidyverse)

df$month_col <- month(df$started_at)

df <- df %>%
  mutate(season=
           ifelse(month_col==12 |month_col==1 |month_col==2,
                  "winter",
                  ifelse(month_col==3 |month_col==4 |month_col==5,
                         "spring",
                         ifelse(month_col==6 |month_col==7 |month_col==8,
                                "summer",
                                "fall"))))

df$year <- year(df$started_at)

df$month_year= paste(df$month_col, df$year)

head(df)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 620BC6107255BF4C electric_bike 2021-10-22 12:46:42 2021-10-22 12:49:50
## 2 4471C70731AB2E45 electric_bike 2021-10-21 09:12:37 2021-10-21 09:14:14
## 3 26CA69D43D15EE14 electric_bike 2021-10-16 16:28:39 2021-10-16 16:36:26
## 4 362947F0437E1514 electric_bike 2021-10-16 16:17:48 2021-10-16 16:19:03
## 5 BB731DE2F2EC51C5 electric_bike 2021-10-20 23:17:54 2021-10-20 23:26:10
## 6 7176307BBC097313 electric_bike 2021-10-21 16:57:37 2021-10-21 17:11:58
##         start_station_name start_station_id end_station_name end_station_id
## 1 Kingsbury St & Kinzie St     KA1503000043             <NA>           <NA>
## 2                     <NA>             <NA>             <NA>           <NA>
## 3                     <NA>             <NA>             <NA>           <NA>
## 4                     <NA>             <NA>             <NA>           <NA>
## 5                     <NA>             <NA>             <NA>           <NA>
## 6                     <NA>             <NA>             <NA>           <NA>
##   start_lat start_lng end_lat end_lng rider_type start_date start_time
## 1  41.88919  -87.6385   41.89  -87.63     member 2021-10-22   12:46:42
## 2  41.93000  -87.7000   41.93  -87.71     member 2021-10-21   09:12:37
## 3  41.92000  -87.7000   41.94  -87.72     member 2021-10-16   16:28:39
## 4  41.92000  -87.6900   41.92  -87.69     member 2021-10-16   16:17:48
## 5  41.89000  -87.7100   41.89  -87.69     member 2021-10-20   23:17:54
## 6  41.89000  -87.7100   41.93  -87.70     member 2021-10-21   16:57:37
##     end_date end_time ride_length no_of_days day_of_week_num day_of_week
## 1 2021-10-22 12:49:50        3.13          1               6      friday
## 2 2021-10-21 09:14:14        1.62          1               5    thursday
## 3 2021-10-16 16:36:26        7.78          1               7    saturday
## 4 2021-10-16 16:19:03        1.25          1               7    saturday
## 5 2021-10-20 23:26:10        8.27          1               4   wednesday
## 6 2021-10-21 17:11:58       14.35          1               5    thursday
##   month_col season year month_year
## 1        10   fall 2021    10 2021
## 2        10   fall 2021    10 2021
## 3        10   fall 2021    10 2021
## 4        10   fall 2021    10 2021
## 5        10   fall 2021    10 2021
## 6        10   fall 2021    10 2021
```

Add holiday column.

```
df <- df %>%
  mutate(holiday=
           ifelse(start_date=="2021-11-11" |start_date=="2021-11-25"|start_date=="2021-12-
24"
                  |start_date=="2021-12-31"|start_date=="2022-01-01"|start_date=="2022-03-
17"
                  |start_date=="2022-05-30"|start_date=="2022-07-04"|start_date=="2022-09-
05"
                  |day_of_week=="sunday"|day_of_week=="saturday",
                  "holiday",
                  "workday"))
```

4. Analyze

Statistics

```
average_ride_length = round(mean(df$ride_length), 2)
print(paste0("Average ride_length = ", average_ride_length))
```

```
## [1] "Average ride_length = 19.41"
```

```
max_ride_length = round(max(df$ride_length), 2)
print(paste0("Max. ride_length = ", max_ride_length))
```

```
## [1] "Max. ride_length = 41387.25"
```

```
tb_no_of_days <- df %>%
  group_by(rider_type) %>%
  summarise(min_no_of_days=min(no_of_days),max_no_of_days=max(no_of_days),avg_no_of_days=r
ound(mean(no_of_days),digits=2))
tb_no_of_days
```

```
## # A tibble: 2 × 4
##   rider_type min_no_of_days max_no_of_days avg_no_of_days
##   <chr>               <dbl>          <dbl>          <dbl>
## 1 casual                  1             30           1.01
## 2 member                  1              3           1
```

```
rides_by_day_of_week <- df %>%
  count(day_of_week)

mode_of_day_of_week <- rides_by_day_of_week %>%
  filter(n == max(n)) %>%
  select(day_of_week)

print(paste0("Mode of the day of week = ", mode_of_day_of_week))
```

```
## [1] "Mode of the day of week = saturday"
```

```
df_casual <- df %>%
  filter(rider_type == 'casual')

rides_by_day_of_week_casual <- df_casual%>%
  count(day_of_week)

mode_of_day_of_week_casual <- rides_by_day_of_week_casual %>%
  filter(n == max(n)) %>%
  select(day_of_week)

print(paste0("Mode of the day of week for casual rider = ", mode_of_day_of_week_casual))
```

```
## [1] "Mode of the day of week for casual rider = saturday"
```

```
df_member <- df %>%
  filter(rider_type == 'member')

rides_by_day_of_week_member <- df_member%>%
  count(day_of_week)

mode_of_day_of_week_member <- rides_by_day_of_week_member %>%
  filter(n == max(n)) %>%
  select(day_of_week)

print(paste0("Mode of the day of week for member = ", mode_of_day_of_week_member))
```

```
## [1] "Mode of the day of week for member = wednesday"
```

Money spent by casual riders per ride

```
keep <- c("ride_id", "rideable_type", "rider_type", "ride_length")
df_casual <- df_casual[keep]

df_casual_classic <- df_casual %>%
  filter(rideable_type=="classic_bike" & ride_length>180)
nrow(df_casual_classic)
```

```
## [1] 8595
```

There are significant number of rides which exceeded the time limit.

We do not know whether the casual riders had day pass or single ride pass We don't know if each ride in day pass has been given new ride ID , lets assume considering minimum extra charges: classic bike if ride length <= 30 - single ride pass + no extra charge if ride length <= 60 single ride pass + extra charge if ride length 60< x < 180 - day pass + no extra charge if ride length > 180 - day pass + extra charges electric bike No time limit as bike is charged per minute docked bike pricing same as classic but without extra charge

As we are making assumptions, the value won't be exact, but we can get a general idea.

```
casual_rider_charges <- df_casual %>%
  mutate(ride_length = ceiling(ride_length)) %>%
  mutate(charges=
           ifelse(rideable_type=="electric_bike", 1+0.39*(ride_length),
                  ifelse(rideable_type=="docked_bike", 3.30,
                         ifelse(ride_length <= 30, 3.30,
                                ifelse(ride_length <= 60, 3.30+0.15*(ride_length-30),
                                       ifelse(ride_length <= 180, 15,
                                              15+(ride_length-180)*0.16))))))
```

```
avg_charge_per_ride_casual_rider = round(sum(casual_rider_charges$charges)/nrow(casual_rid
er_charges), digits=2)
print(paste0("avg_charge_per_ride_casual_rider = ", avg_charge_per_ride_casual_rider, "
$"))
```

```
## [1] "avg_charge_per_ride_casual_rider = 6.38 $"
```

```
max_charge_per_ride_casual_rider = max(casual_rider_charges$charges)
print(paste0("max_charge_per_ride_casual_rider = ", max_charge_per_ride_casual_rider, "
$"))
```

```
## [1] "max_charge_per_ride_casual_rider = 235.8 $"
```

```
# Casual riders who spent more than annual membership charges

Target <- casual_rider_charges %>%
  filter(charges > 108) %>%
  nrow()

Target
```

```
## [1] 4105
```

At least 4105 Casual riders paid more than Annual membership for a single ride. These people will be easy to convert into members. More data will be needed for further analysis.

```
# Casual riders who spent more than annual membership charges

Target1 <- casual_rider_charges %>%
  filter(charges > 100) %>%
  nrow()

Target1
```

```
## [1] 4296
```

Members vs Casual Riders at a glance

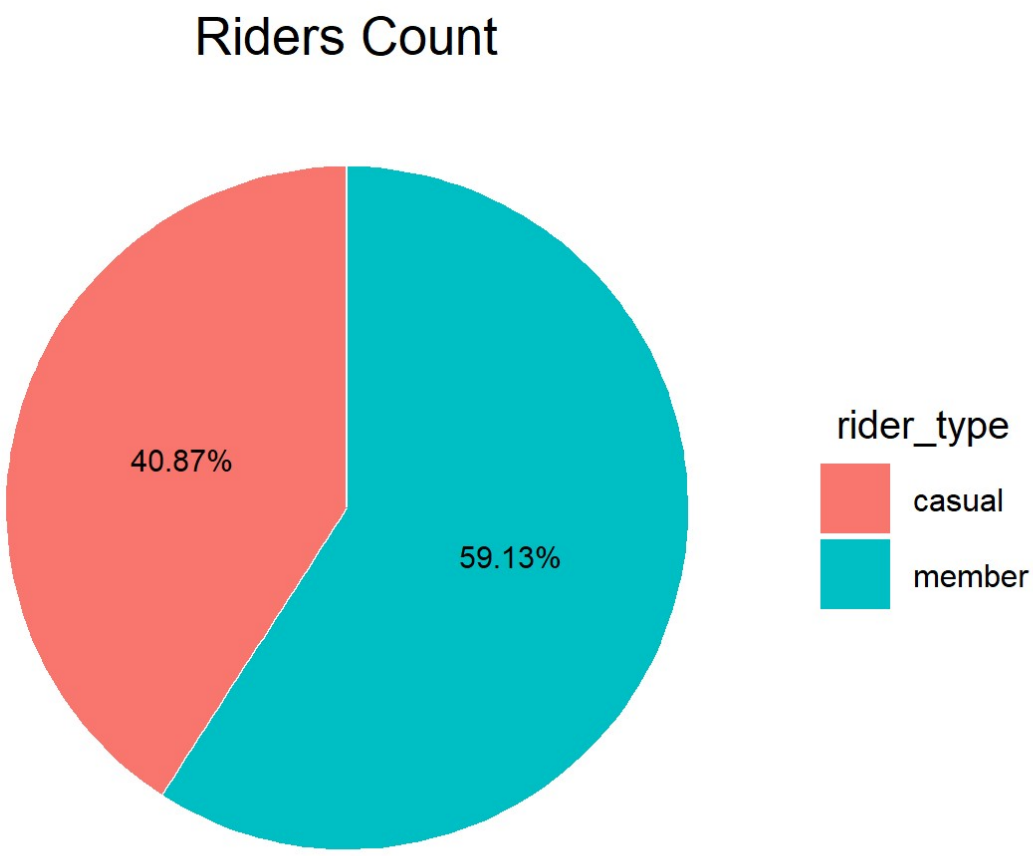
Number of rides

```
riders_count <- df %>%
  group_by(rider_type) %>%
  summarize(count = n())

total_rides <- nrow(df)

riders_count$percentage <- round((riders_count$count/ total_rides)*100, digits = 2)
```

```
riders_count_freq_tbl <- riders_count%>%
  bind_rows(summarise(., across(where(is.numeric), sum),
                      across(where(is.character), ~'Total')))
```

```
percentage <- round(riders_count$percentage, digits=2)
rider_type <- riders_count$rider_type

ypos = cumsum(percentage) - 0.5 * percentage
ypos = 100 - ypos

pie1 = ggplot() + theme_bw() +
  geom_bar(aes(x = "", y = percentage, fill = rider_type),
           stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  ggtitle("Riders Count") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid=element_blank(),
        panel.border = element_blank()) +
  theme(legend.text=element_text(size=12),
        legend.title = element_text(hjust = 0.5, size=15),
        legend.key.size = unit(1,"cm")) +
  geom_text(aes(x = "", y = ypos, label = paste0(percentage, "%")), size = 4)
```

```
pie1
```




```
riders_count_freq_tbl
```

```
## # A tibble: 3 × 3
##   rider_type   count percentage
##   <chr>        <int>      <dbl>
## 1 casual     2609952       40.9
## 2 member     3776332       59.1
## 3 Total      6386284      100
```

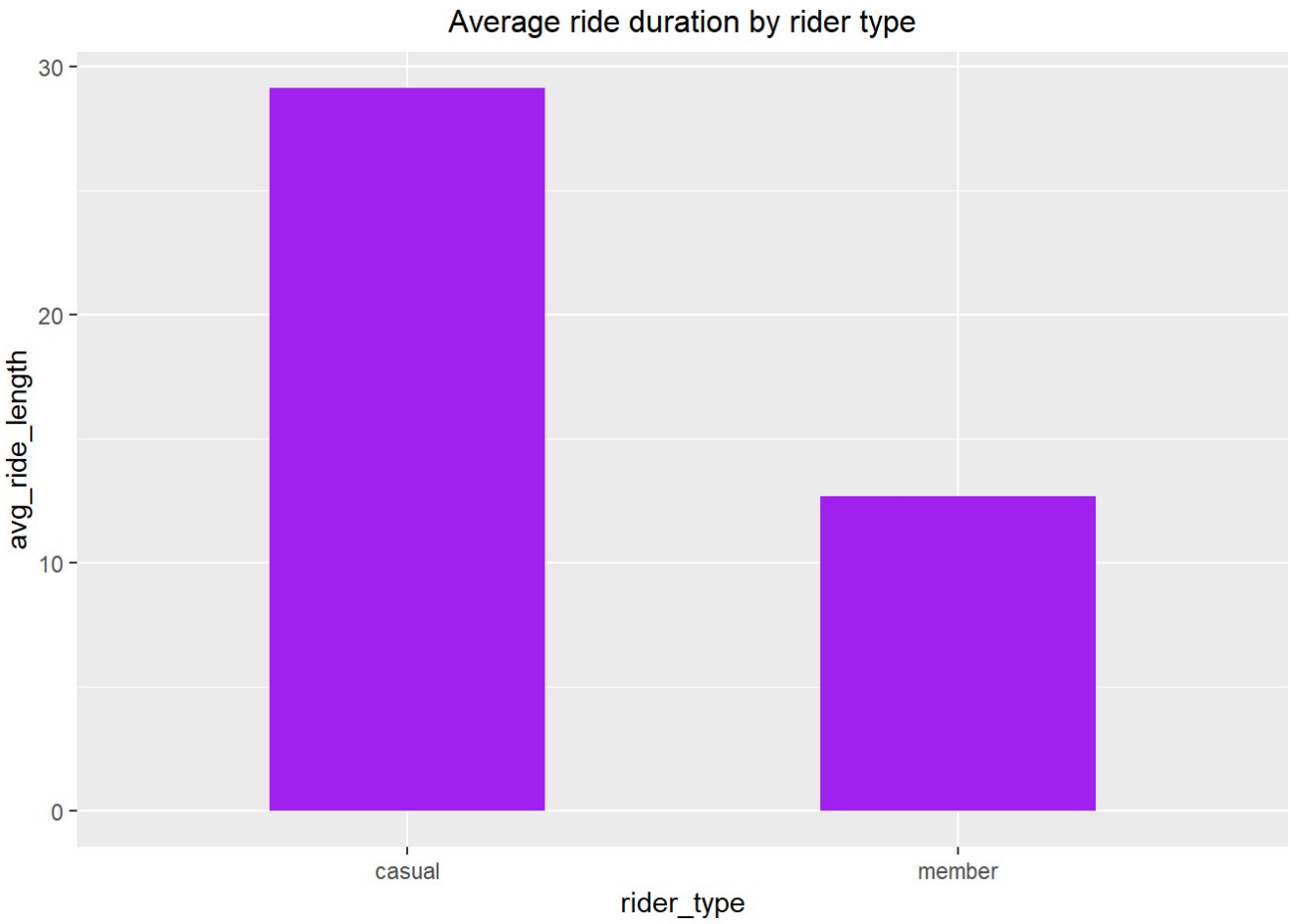
- Bike usage was observed to be more in Annual members compared to casual riders.

Average ride duration

```
avg_ride_length <- df %>%
  group_by(rider_type) %>%
  summarise(avg_ride_length = round(mean(ride_length), digits=2))
```

```
bar1 <- ggplot(data=avg_ride_length, aes(x=rider_type, y=avg_ride_length)) +
  geom_bar(stat="identity", width = 0.5, fill="purple")+
  ggtitle("Average ride duration by rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar1
```




```
avg_ride_length
```

```
## # A tibble: 2 × 2
##   rider_type avg_ride_length
##   <chr>                <dbl>
## 1 casual                29.1
## 2 member                12.7
```

- Average ride duration for members was 12.69 mins.
- Average ride duration were higher in Casual riders (29.13 mins).

  This may indicate that :

- Members used bikes for their daily commute to work or University &
- Casual riders used the bikes for longer distances or leisure activities.

Trends based on day of the week

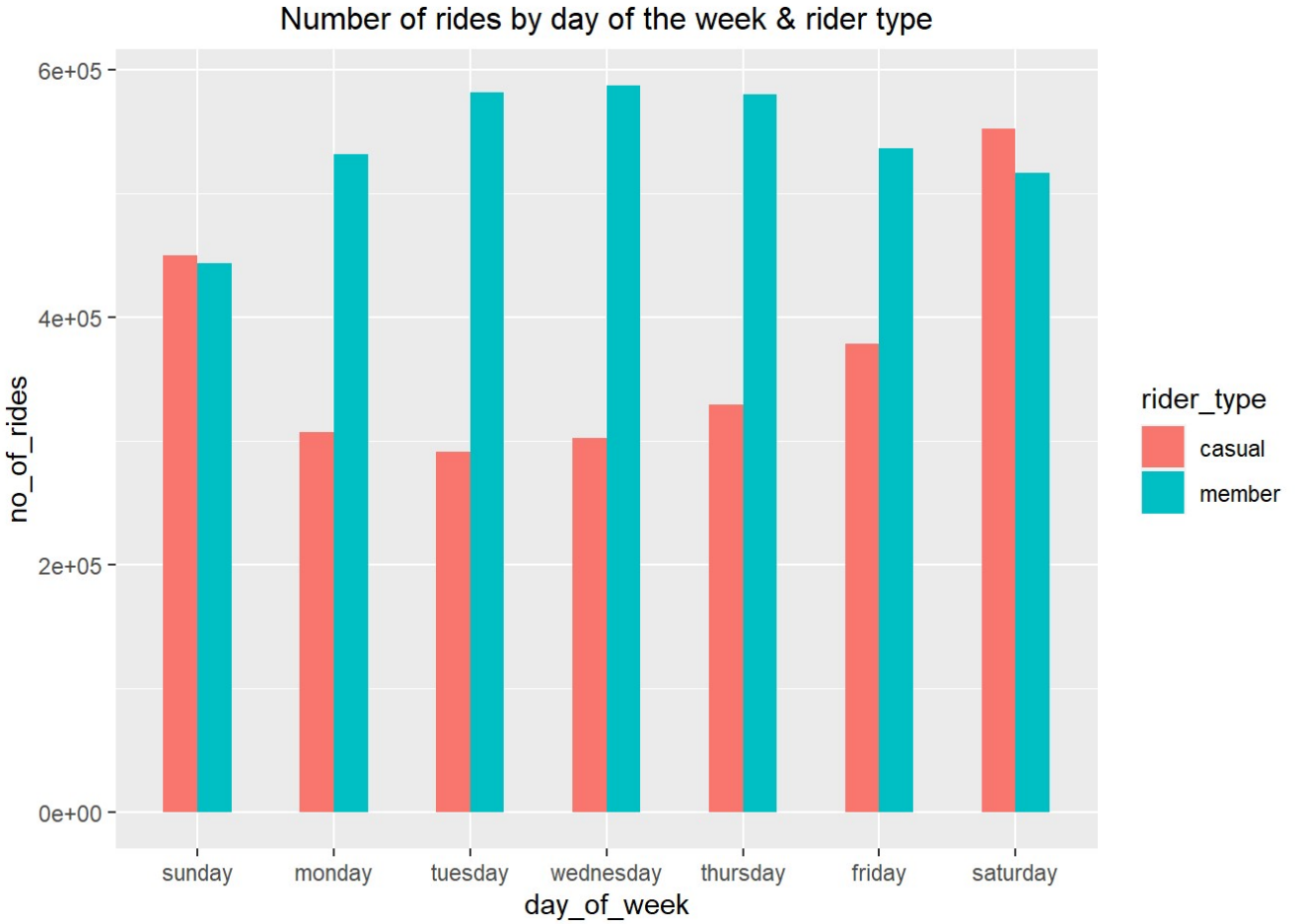
Number of rides

```
no_of_rides_by_day <- df %>%
  group_by(day_of_week, rider_type) %>%
  summarise(no_of_rides = n())
```

```
no_of_rides_by_day_tbl <- spread(no_of_rides_by_day, key= "rider_type", value= "no_of_ride
s")
```

```
positions <- c("sunday","monday","tuesday","wednesday","thursday","friday","saturday")

bar2 <- ggplot(data=no_of_rides_by_day, aes(x=day_of_week, y=no_of_rides, fill=rider_typ
e)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  scale_x_discrete(limits = positions)+
  ggtitle("Number of rides by day of the week & rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar2
```




```
no_of_rides_by_day_tbl
```

```
## # A tibble: 7 × 3
## # Groups:   day_of_week [7]
##   day_of_week casual member
##   <chr>        <int>  <int>
## 1 friday      378440 536453
## 2 monday      306995 531311
## 3 saturday    551952 516322
## 4 sunday      449653 443931
## 5 thursday    329245 579743
## 6 tuesday     291437 581226
## 7 wednesday   302230 587346
```

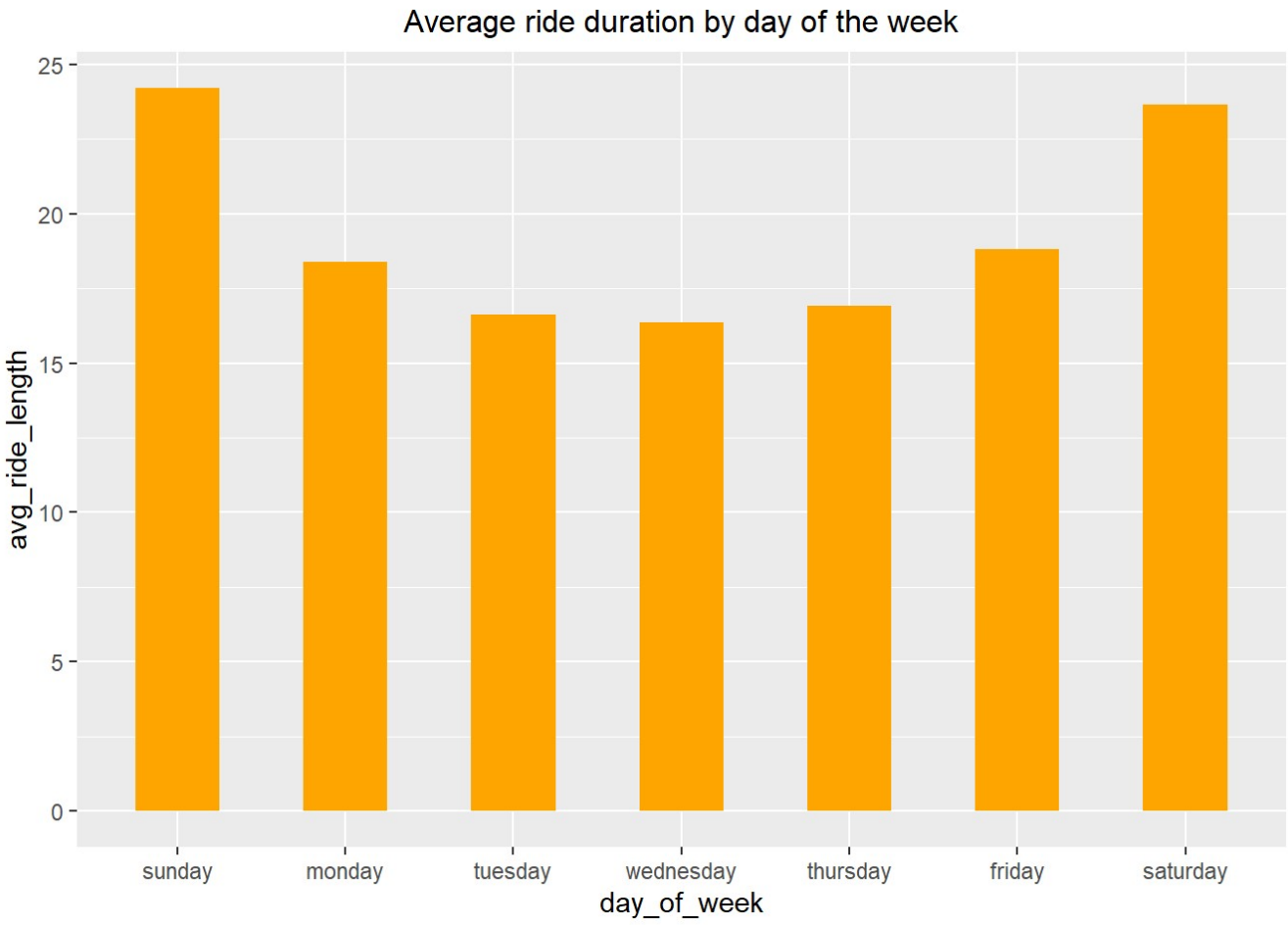
- Number of rides for annual members were seen to be higher on weekdays. This helps us reaffirm our theory that members use the bikes for short distance daily commutes to and from work or University.
- Number of rides were highest on Saturdays and Sundays for casual riders which again confirms our previous assumption that the casual riders used bikes for leisure activities like exploring the city, going to movies, cafes or restaurants.
- Number of rides by casual riders were marginally higher than members on Saturdays and Sundays. On all other days, number of rides by members significantly higher than casual riders. This is also an indication that the significant proportion of casual riders may be tourists.

Average ride duration

```
avg_ride_length1 <- df %>%
  group_by(day_of_week) %>%
  summarise(avg_ride_length = round(mean(ride_length), digits=2))
```

```
positions <- c("sunday","monday","tuesday","wednesday","thursday","friday","saturday")
bar3 <- ggplot(data=avg_ride_length1, aes(x=day_of_week, y=avg_ride_length)) +
  geom_bar(stat="identity", width = 0.5, fill="orange")+
  scale_x_discrete(limits = positions)+
  ggtitle("Average ride duration by day of the week")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar3
```




```
avg_ride_length1
```

```
## # A tibble: 7 × 2
##   day_of_week avg_ride_length
##   <chr>                 <dbl>
## 1 friday                 18.8
## 2 monday                 18.4
## 3 saturday               23.6
## 4 sunday                 24.2
## 5 thursday               16.9
## 6 tuesday                16.6
## 7 wednesday              16.4
```

- Average ride duration was highest on Sundays followed by Saturdays.
- The ride lengths remained in the similar range from Tuesdays to Thursdays but were slightly higher on Mondays and Fridays. This may be due to the long Weekends.

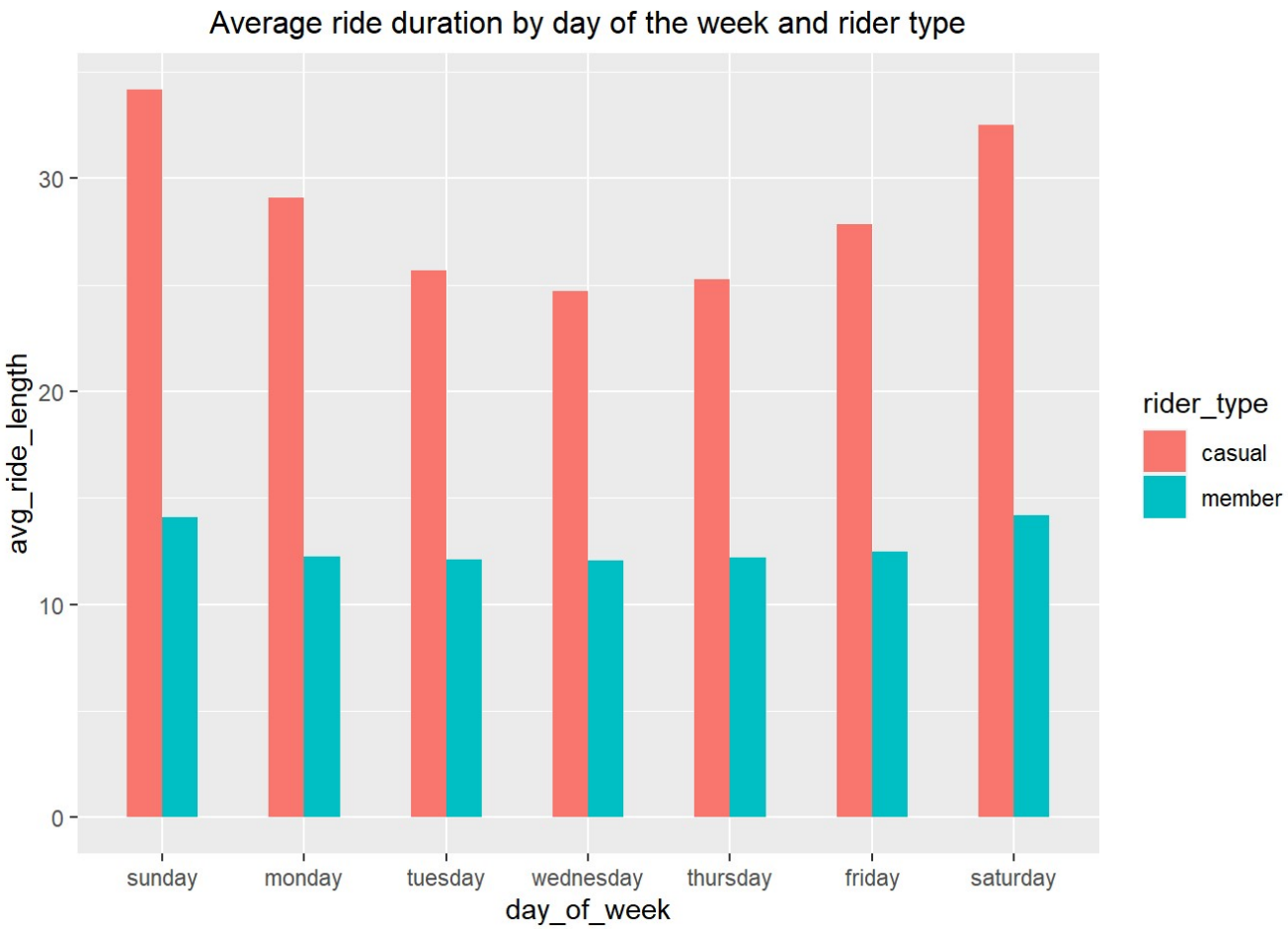
Average ride duration by day of the week & rider type

```
avg_ride_length_by_day <- df %>%
  group_by(day_of_week, rider_type) %>%
  summarise(avg_ride_length = round(mean(ride_length), digits=2))
```

```
library(tidyr)

avg_ride_length_by_day_tbl <- spread(avg_ride_length_by_day, key= "rider_type", value= "av
g_ride_length")

avg_ride_length_by_day_tbl <- avg_ride_length_by_day_tbl %>%
  rename( "avg_ride_length_casual" = "casual",
          "avg_ride_length_member" = "member")
```

```
bar4 <- ggplot(data=avg_ride_length_by_day, aes(x=day_of_week, y=avg_ride_length, fill=rid
er_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  scale_x_discrete(limits = positions)+
  ggtitle("Average ride duration by day of the week and rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar4
```




```
avg_ride_length_by_day_tbl
```

```
## # A tibble: 7 × 3
## # Groups:   day_of_week [7]
##   day_of_week avg_ride_length_casual avg_ride_length_member
##   <chr>                        <dbl>                  <dbl>
## 1 friday                        27.8                   12.5
## 2 monday                        29.1                   12.2
## 3 saturday                      32.5                   14.2
## 4 sunday                        34.2                   14.1
## 5 thursday                      25.3                   12.2
## 6 tuesday                       25.7                   12.1
## 7 wednesday                     24.7                   12.0
```

- Average ride duration for casual riders (25 mins or more) was consistently higher than annual members (less than 15 mins) regardless of the day of the week.
- It was seen to be higher on weekend compared to weekdays in Casual riders as well as members.

Bike-type wise trend

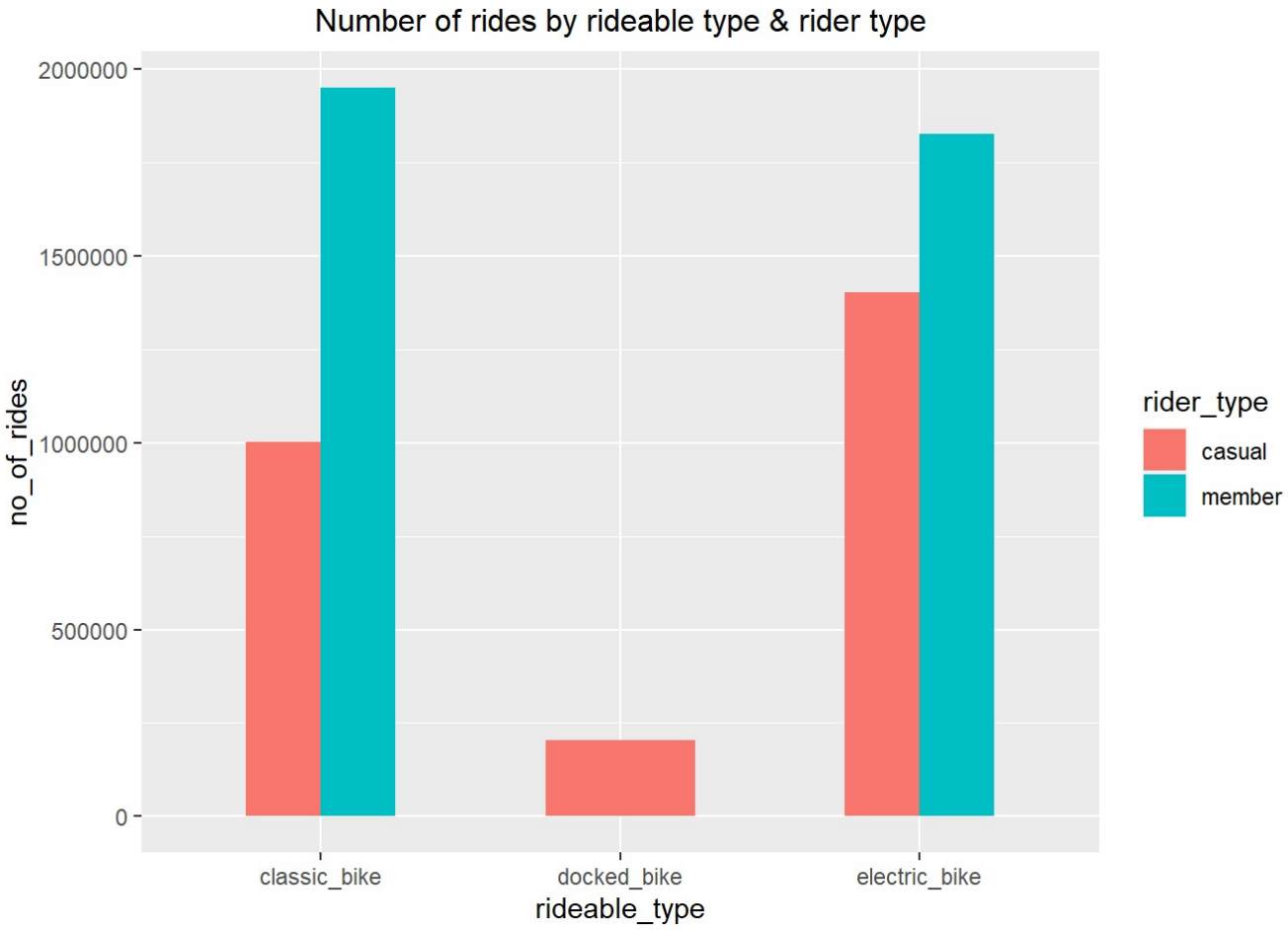
Number of rides by rider type

```
bike_type <- df %>%
  group_by(rideable_type, rider_type) %>%
  summarise(no_of_rides = n())
```

```
bike_type_tbl <- spread(bike_type, key= "rider_type", value= "no_of_rides")
```

```
bar5 <- ggplot(data=bike_type, aes(x=rideable_type, y=no_of_rides, fill=rider_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  ggtitle("Number of rides by rideable type & rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar5
```




```
bike_type_tbl
```

```
## # A tibble: 3 × 3
## # Groups:   rideable_type [3]
##   rideable_type  casual  member
##   <chr>           <int>   <int>
## 1 classic_bike  1002953 1950970
## 2 docked_bike    205078      NA
## 3 electric_bike 1401921 1825362
```

- Classic bike was the bike of preference for members, closely followed by electric bike. Members did not use docked bike even once throughout the year.
- Casual riders preferred electric bikes over classic bikes. Casual riders did opt for docked bikes occasionally.

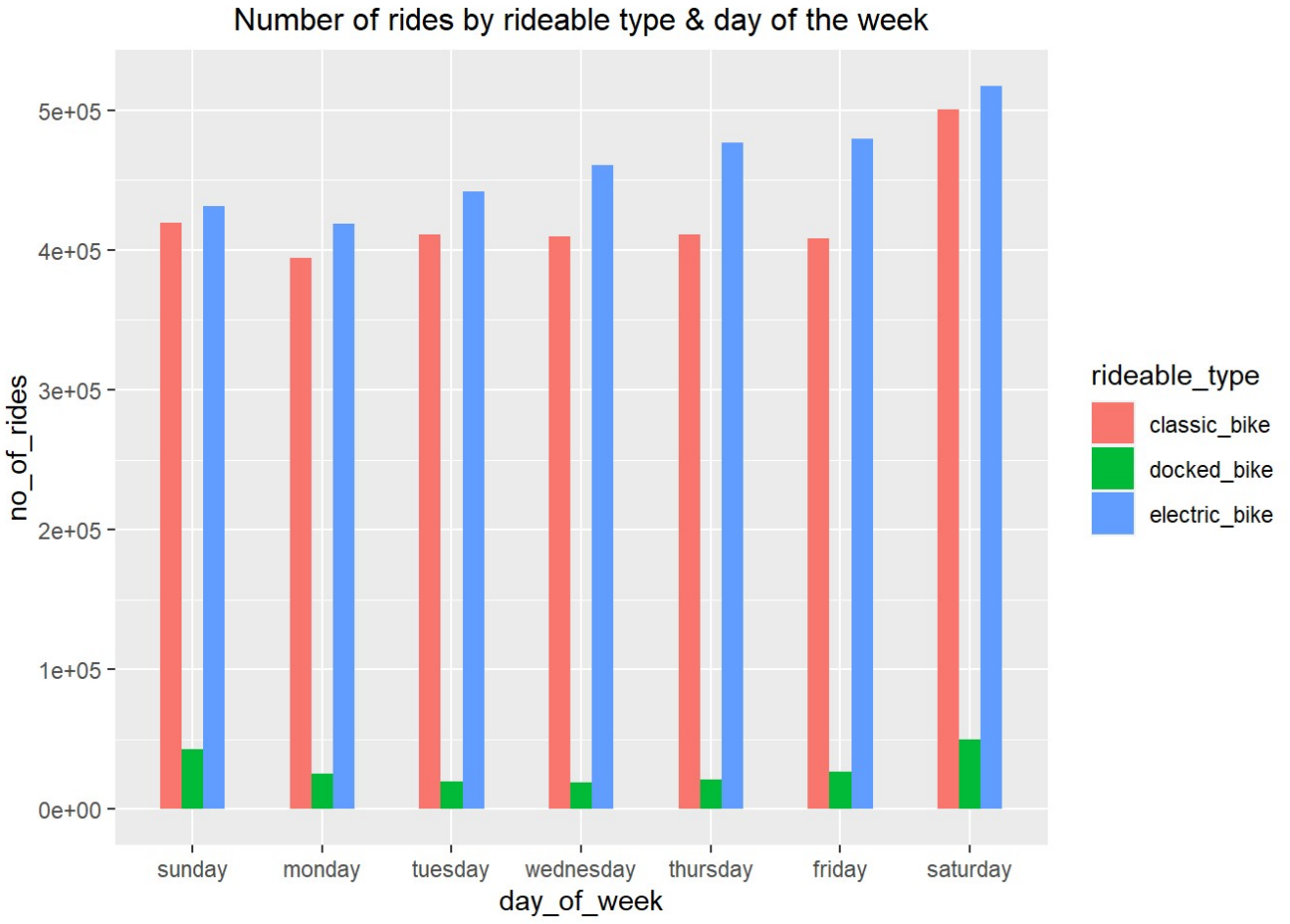
Number of rides by day of the week

```
bike_type2 <- df %>%
  group_by(day_of_week,rideable_type) %>%
  summarise(no_of_rides = n())
```

```
bike_type2_tbl <- spread(bike_type2, key= "rideable_type", value= "no_of_rides")
```

```
bar6 <- ggplot(data=bike_type2, aes(x=day_of_week, y=no_of_rides, fill=rideable_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  scale_x_discrete(limits = positions)+
  ggtitle("Number of rides by rideable type & day of the week")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar6
```




```
bike_type2_tbl
```

```
## # A tibble: 7 × 4
## # Groups:   day_of_week [7]
##   day_of_week classic_bike docked_bike electric_bike
##   <chr>              <int>       <int>         <int>
## 1 friday            408071       27120        479702
## 2 monday            394140       25235        418931
## 3 saturday          500859       49902        517513
## 4 sunday            419319       42886        431379
## 5 thursday          410990       20938        477060
## 6 tuesday           410890       19720        442053
## 7 wednesday         409654       19277        460645
```

- Number of bike rides on all three bikes was highest on Saturday followed by Sunday.
- Classic and docked bikes usage was less than electric bikes throughout the week.
- This may be due to the fact that, electric bikes are faster and take less effort.

This must have resulted in

- more people choosing electric bikes.
- Faster rides means faster bike returns and in turn more availability.

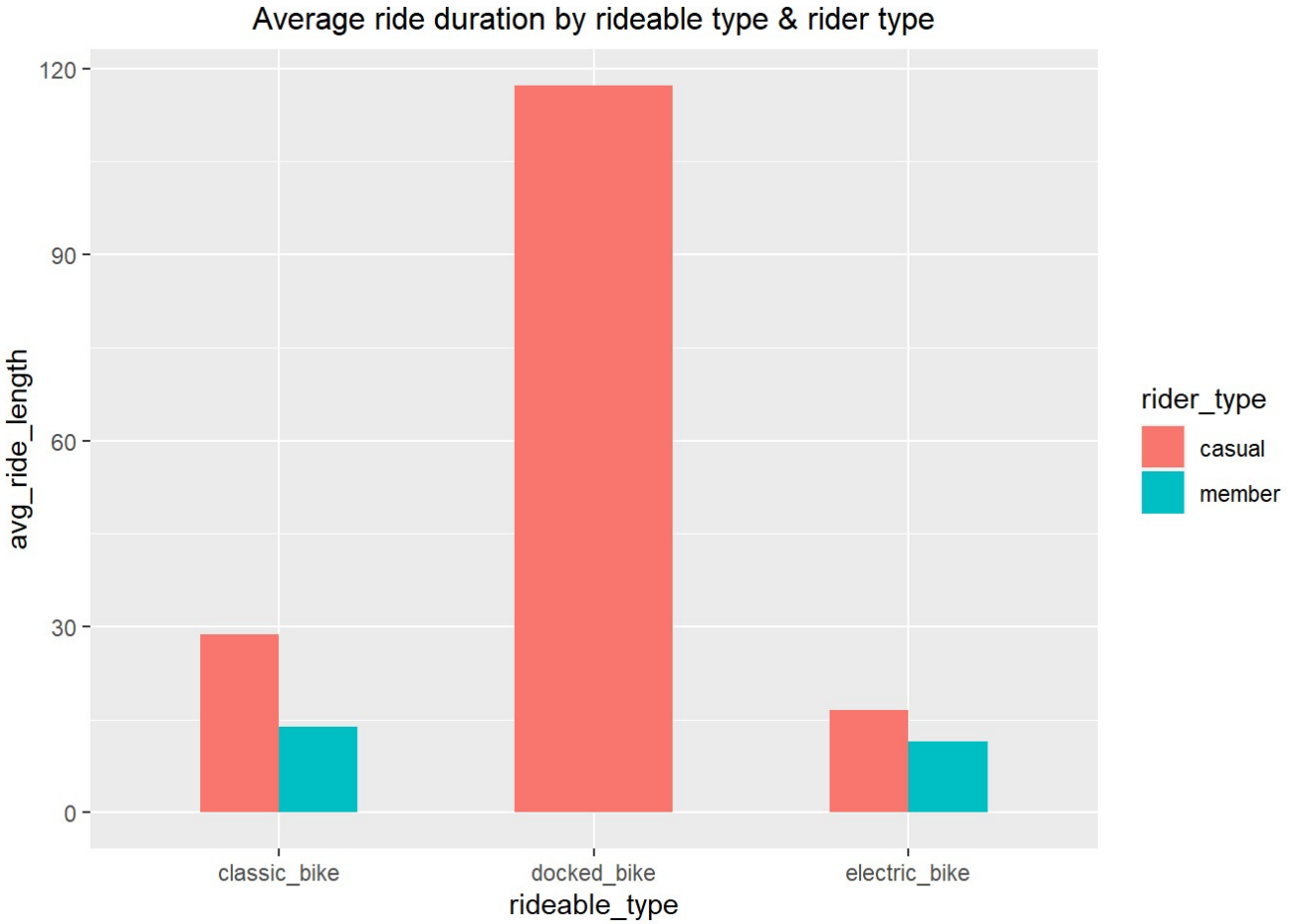
Average ride duration by rideable type & rider type

```
avg_ride_length_by_ride_type <- df %>%
  group_by(rideable_type, rider_type) %>%
  summarise(avg_ride_length = round(mean(ride_length), digits=2))
```

```
avg_ride_length_by_ride_type_tbl <- spread(avg_ride_length_by_ride_type, key= "rider_typ
e", value= "avg_ride_length")

avg_ride_length_by_ride_type_tbl <- avg_ride_length_by_ride_type_tbl %>%
  rename( "avg_ride_length_casual" = "casual",
          "avg_ride_length_member" = "member")
```

```
bar7 <- ggplot(data=avg_ride_length_by_ride_type, aes(x=rideable_type, y=avg_ride_length,
fill=rider_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  ggtitle("Average ride duration by rideable type & rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar7
```




```
avg_ride_length_by_ride_type_tbl
```

```
## # A tibble: 3 × 3
## # Groups:   rideable_type [3]
##   rideable_type avg_ride_length_casual avg_ride_length_member
##   <chr>                          <dbl>                  <dbl>
## 1 classic_bike                    28.8                   13.8
## 2 docked_bike                    117.                    NA
## 3 electric_bike                   16.5                   11.5
```

- This graph shows that rides were faster on electric bikes compared to classic / docked bikes.

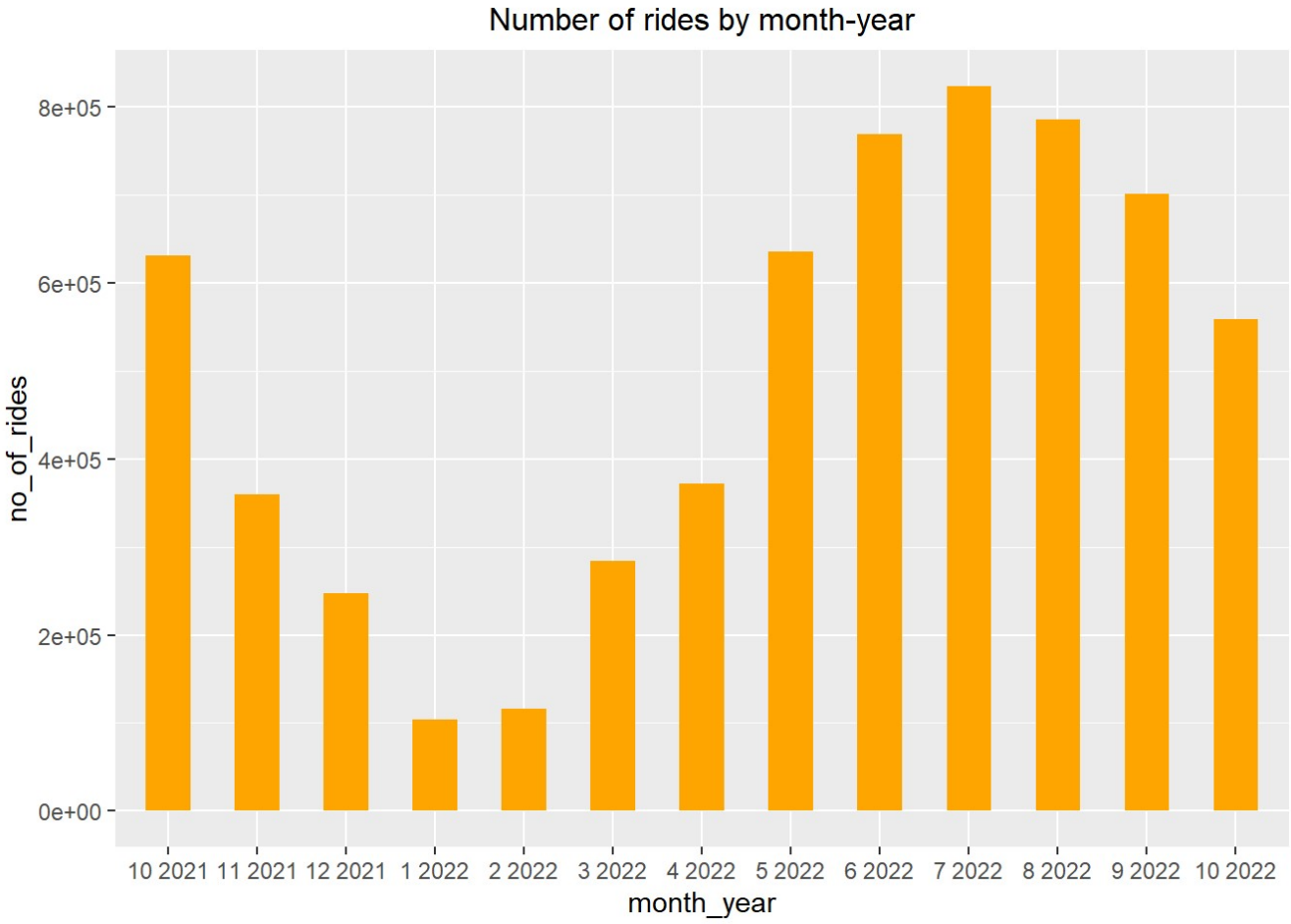
Month-wise Trends

Number of rides by month-year

```
monthwise_trend <- df %>%
  group_by(month_year) %>%
  summarise(no_of_rides = n())
```

```
positions3 <- c("10 2021","11 2021","12 2021","1 2022","2 2022","3 2022", "4 2022", "5 202
2", "6 2022", "7 2022", "8 2022", "9 2022", "10 2022")

bar8 <- ggplot(data=monthwise_trend, aes(x=month_year, y=no_of_rides)) +
  geom_bar(stat="identity", width = 0.5, fill="orange")+
  scale_x_discrete(limits = positions3)+
  ggtitle("Number of rides by month-year")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar8
```




```
monthwise_trend
```

```
## # A tibble: 13 × 2
##    month_year no_of_rides
##    <chr>            <int>
##  1 1 2022          103765
##  2 10 2021         631156
##  3 10 2022         558620
##  4 11 2021         359892
##  5 12 2021         247519
##  6 2 2022          115604
##  7 3 2022          284024
##  8 4 2022          371218
##  9 5 2022          634810
## 10 6 2022          769138
## 11 7 2022          823416
## 12 8 2022          785855
## 13 9 2022          701267
```

- This graph shows how the number of rides gradually increased around summer and decreased during winters.

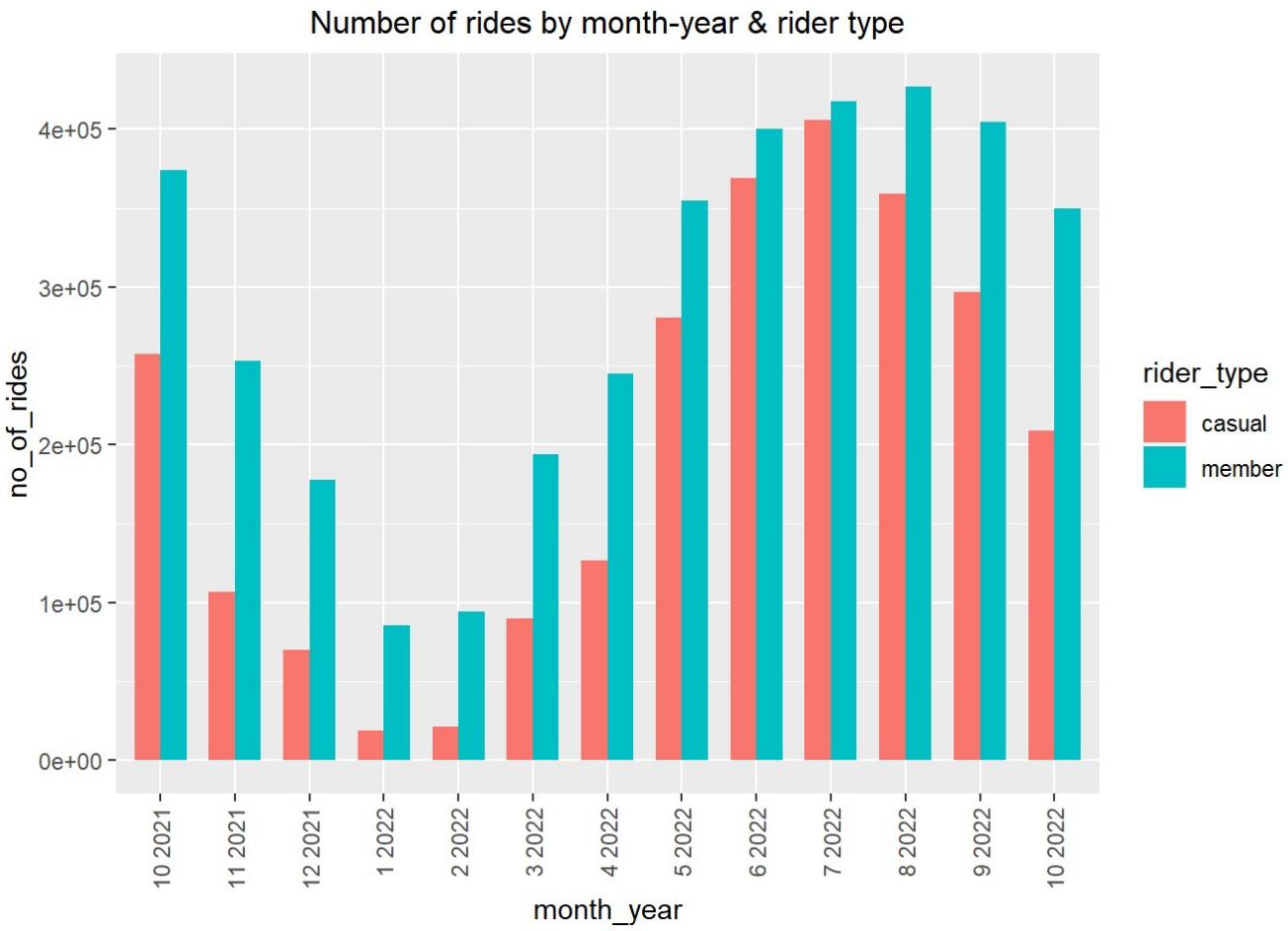
Number of rides by month-year & rider type

```
monthwise_trend1 <- df %>%
  group_by(month_year,rider_type) %>%
  summarise(no_of_rides = n())
```

```
monthwise_trend_tbl1 <- spread(monthwise_trend1, key= "rider_type", value= "no_of_rides")
```

```
positions3 <- c("10 2021","11 2021","12 2021","1 2022","2 2022","3 2022", "4 2022", "5 202
2", "6 2022", "7 2022", "8 2022", "9 2022", "10 2022")

bar9 <- ggplot(data=monthwise_trend1, aes(x=month_year, y=no_of_rides, fill=rider_type)) +
  geom_bar(stat="identity", width = 0.7, position = 'dodge')+
  scale_x_discrete(limits = positions3)+
  ggtitle("Number of rides by month-year & rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
bar9
```




```
monthwise_trend_tbl1
```

```
## # A tibble: 13 × 3
## # Groups:   month_year [13]
##    month_year casual member
##    <chr>       <int>  <int>
##  1 1 2022      18517  85248
##  2 10 2021    257203 373953
##  3 10 2022    208961 349659
##  4 11 2021    106884 253008
##  5 12 2021     69729 177790
##  6 2 2022      21414  94190
##  7 3 2022      89874 194150
##  8 4 2022     126398 244820
##  9 5 2022     280387 354423
## 10 6 2022     369022 400116
## 11 7 2022     406013 417403
## 12 8 2022     358886 426969
## 13 9 2022     296664 404603
```

- Number of rides by members were always higher than casual riders.
- There was marginal difference between number of rides by casual riders and members June and July.
- The difference kept increasing as the winter approached, and during winter even rides by members decreased.

Season-wise Trends

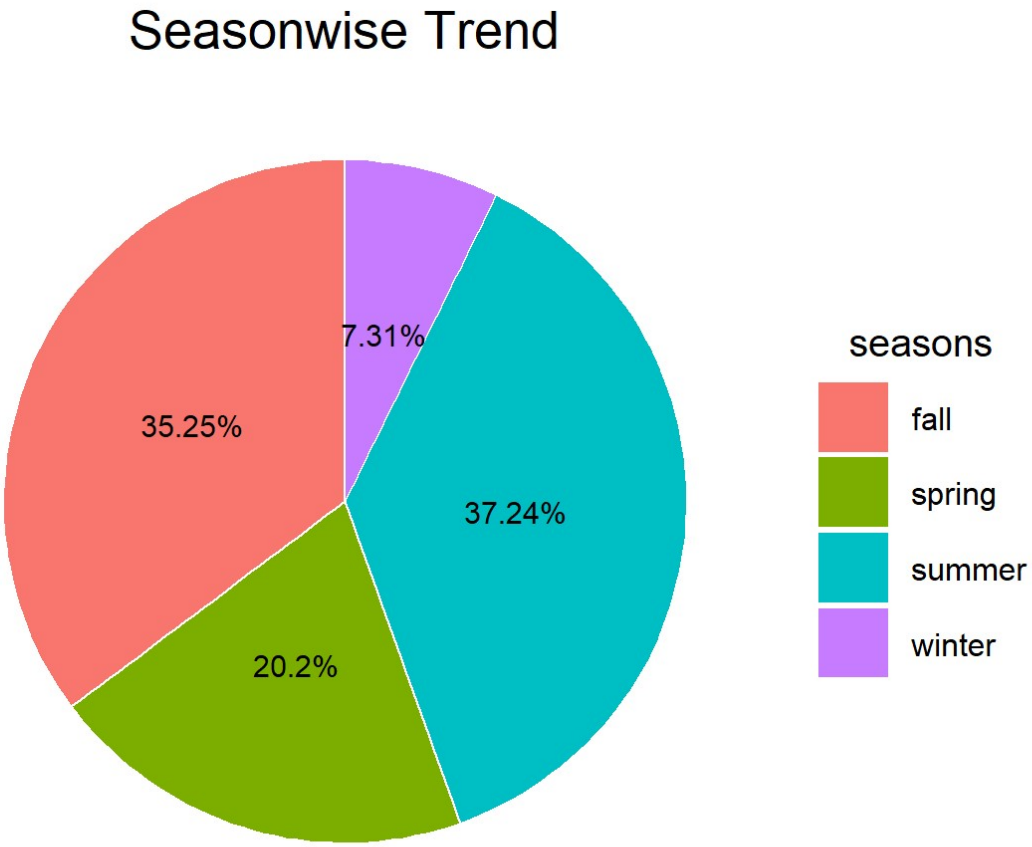
Number of rides

```
season <- df %>%
  group_by(season) %>%
  summarise(no_of_rides = n())

season$percentage <- round((season$no_of_rides/ total_rides)*100, digits=2)
```

```
seasonal_freq_tbl <- season%>%
  bind_rows(summarise(., across(where(is.numeric), sum),
                      across(where(is.character), ~'Total')))
```

```
percentage2 <- round(season$percentage, digits=2)
seasons <- season$season

ypos1 = cumsum(percentage2) - 0.5 * percentage2
ypos1 = 100 - ypos1

pie2 <- ggplot() + theme_bw() +
  geom_bar(aes(x = "", y = percentage2, fill = seasons),
           stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  ggtitle("Seasonwise Trend") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid=element_blank(),
        panel.border = element_blank()) +
  theme(legend.text=element_text(size=12),
        legend.title = element_text(hjust = 0.5, size=15),
        legend.key.size = unit(1,"cm")) +
  geom_text(aes(x = "", y = ypos1, label = paste0(percentage2, "%")), size = 4)
```

```
pie2
```




```
seasonal_freq_tbl
```

```
## # A tibble: 5 × 3
##   season no_of_rides percentage
##   <chr>        <int>      <dbl>
## 1 fall       2250935      35.2
## 2 spring     1290052      20.2
## 3 summer     2378409      37.2
## 4 winter      466888       7.31
## 5 Total      6386284     100
```

- Highest numbers of rides were observed in Summer followed by fall.
- Lowest number of rides were seen in Winters.
- As the data is for Chicago, which has tremendous amounts of snow and winds during winters, such low numbers of rides during winters are justified.

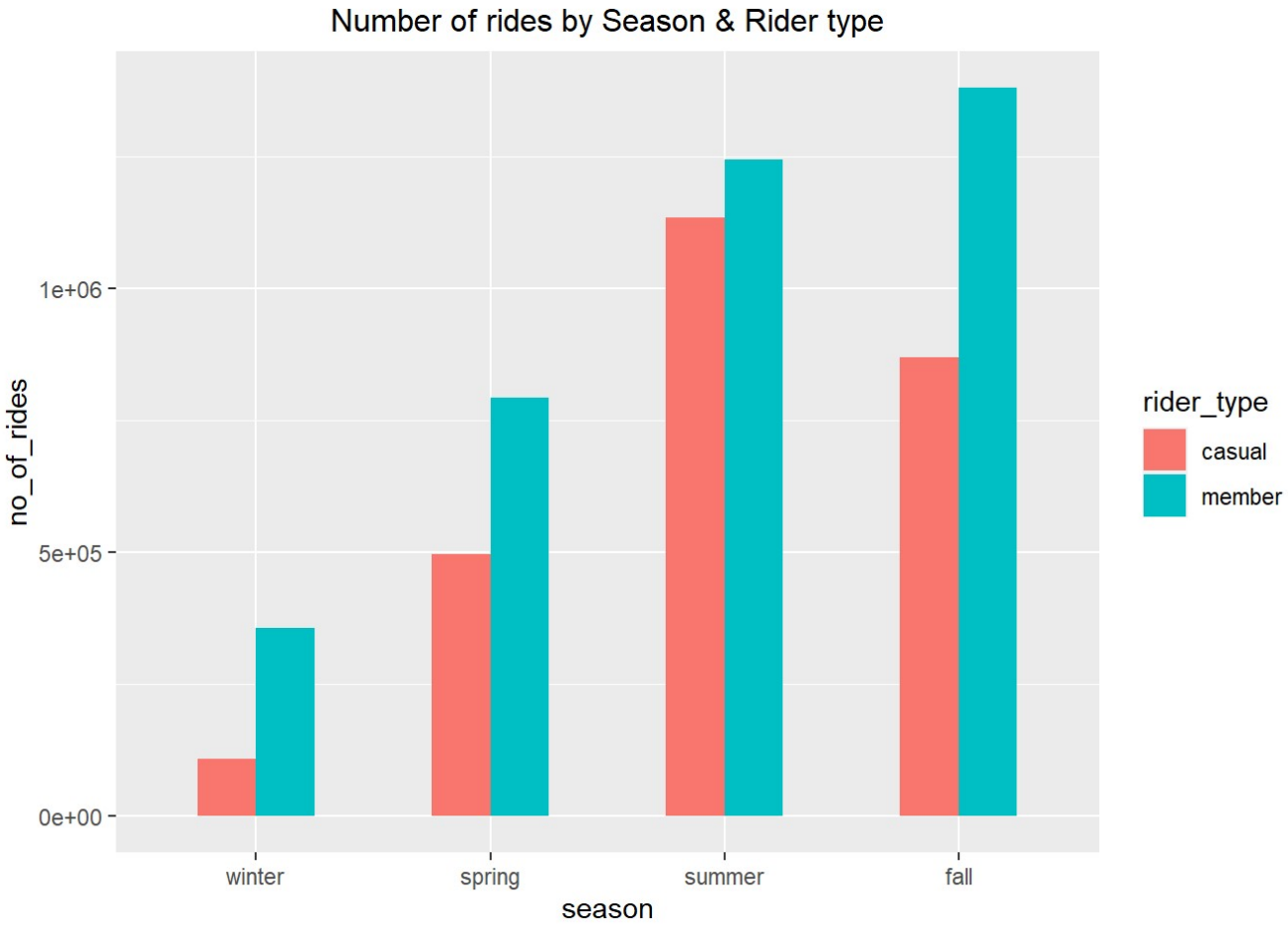
Number of rides by Season & Rider type

```
seasonwise_rides <- df %>%
  group_by(season,rider_type) %>%
  summarise(no_of_rides = n())
```

```
seasonwise_rides_tbl <- spread(seasonwise_rides, key= "rider_type", value= "no_of_rides")
```

```
positions2 <- c("winter","spring","summer","fall")

bar10 <- ggplot(data=seasonwise_rides, aes(x=season, y=no_of_rides, fill=rider_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  scale_x_discrete(limits = positions2)+
  ggtitle("Number of rides by Season & Rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar10
```




```
seasonwise_rides_tbl
```

```
## # A tibble: 4 × 3
## # Groups:   season [4]
##   season  casual  member
##   <chr>    <int>   <int>
## 1 fall    869712 1381223
## 2 spring  496659  793393
## 3 summer 1133921 1244488
## 4 winter  109660  357228
```

- Number of rides by members were higher than Casual riders each season.
- Fall was season in which members used maximum number of bike rides.
- Maximum number of rides by casual riders were seen in Summer.

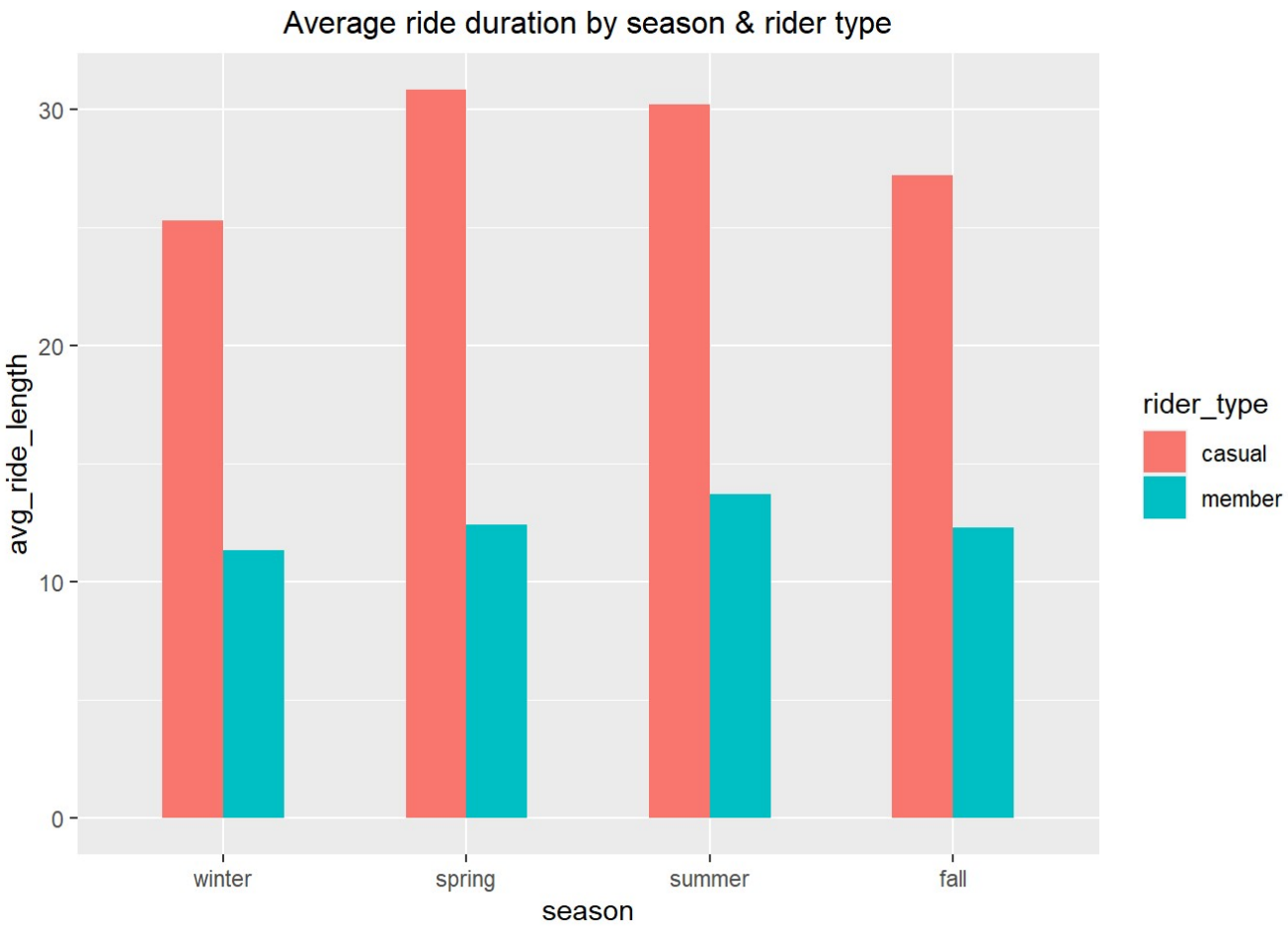
Average ride duration by Season & Rider type

```
avg_ride_length_by_season <- df %>%
  group_by(season, rider_type) %>%
  summarise(avg_ride_length = round(mean(ride_length), digits=2))
```

```
avg_ride_length_by_season_tbl <- spread(avg_ride_length_by_season, key= "rider_type", valu
e= "avg_ride_length")

avg_ride_length_by_season_tbl <- avg_ride_length_by_season_tbl %>%
  rename( "avg_ride_length_casual" = "casual",
          "avg_ride_length_member" = "member")
```

```
positions2 <- c("winter","spring","summer","fall")

bar11 <- ggplot(data=avg_ride_length_by_season, aes(x=season, y=avg_ride_length, fill=ride
r_type)) +
  geom_bar(stat="identity", width = 0.5, position = 'dodge')+
  scale_x_discrete(limits = positions2)+
  ggtitle("Average ride duration by season & rider type")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar11
```




```
avg_ride_length_by_season_tbl
```

```
## # A tibble: 4 × 3
## # Groups:   season [4]
##   season avg_ride_length_casual avg_ride_length_member
##   <chr>                   <dbl>                  <dbl>
## 1 fall                     27.2                   12.3
## 2 spring                   30.8                   12.4
## 3 summer                   30.2                   13.7
## 4 winter                   25.3                   11.3
```

- Average ride length remained within the range of 25 to 31 mins.
- Average ride duration remained within the range of 11 to 14 mins.

Holiday-wise Trends

```
#find unique dates in data and create a new data frame

unique_dates <- unique(df$start_date)
unique_dates_df <- as.data.frame(unique_dates)
```

```
#add day column to new data frame

unique_dates_df$day_num <- wday(unique_dates_df$unique_dates)

unique_dates_df <- unique_dates_df %>%
  mutate(day=
           ifelse(day_num==1,"sunday",
                  ifelse(day_num==2, "monday",
                         ifelse(day_num==3, "tuesday",
                                ifelse(day_num==4, "wednesday",
                                       ifelse(day_num==5, "thursday",
                                              ifelse(day_num==6, "friday",
                                                     "saturday" )))))))
```

```
#add holiday column to new data frame

unique_dates_df <- unique_dates_df %>%
  mutate(holidays=
           ifelse(unique_dates=="2021-11-11" |unique_dates=="2021-11-25"|unique_dates=="20
21-12-24"
                  |unique_dates=="2021-12-31"|unique_dates=="2022-01-01"|unique_dates=="20
22-03-17"
                  |unique_dates=="2022-05-30"|unique_dates=="2022-07-04"|unique_dates=="20
22-09-05"
                  |day=="sunday"|day=="saturday",
                  "holiday",
                  "workday"))
```

```
holiday_tbl <- unique_dates_df %>%
  group_by(holidays) %>%
  summarise("no_of_days"=n())

holiday_tbl
```

```
## # A tibble: 2 × 2
##   holidays no_of_days
##   <chr>         <int>
## 1 holiday         122
## 2 workday         274
```

```
no_of_rides_by_holiday <- df %>%
  group_by(holiday,rider_type) %>%
  summarise("no_of_rides"= n())

no_of_rides_by_holiday <- no_of_rides_by_holiday %>%
  spread(key = rider_type, value=no_of_rides)

no_of_rides_by_holiday$total = no_of_rides_by_holiday$casual + no_of_rides_by_holiday$memb
er

no_of_rides_by_holiday$num_of_days <- holiday_tbl$no_of_days

no_of_rides_by_holiday <- as.data.frame(no_of_rides_by_holiday)

no_of_rides_by_holiday
```

```
##   holiday  casual  member   total num_of_days
## 1 holiday 1060024 1019811 2079835         122
## 2 workday 1549928 2756521 4306449         274
```

```
rides_by_holiday <- no_of_rides_by_holiday %>%
  select(holiday,casual,member,num_of_days)

rides_by_holiday <-rides_by_holiday %>%
  gather(key=rider_type, no_of_rides, casual:member)

rides_by_holiday <- rides_by_holiday %>%
  mutate(avg_no_of_rides = round(no_of_rides/num_of_days,digits=2))
```

```
total_rides_by_holiday <- no_of_rides_by_holiday %>%
  select(holiday,total,num_of_days) %>%
  rename("no_of_rides" = "total") %>%
  mutate(avg_rides = no_of_rides/num_of_days)
```

```
all_riders_tbl <- no_of_rides_by_holiday %>%
  select(holiday, total, num_of_days) %>%
  mutate(avg_rides= total/num_of_days)
```

Holidays vs Workdays

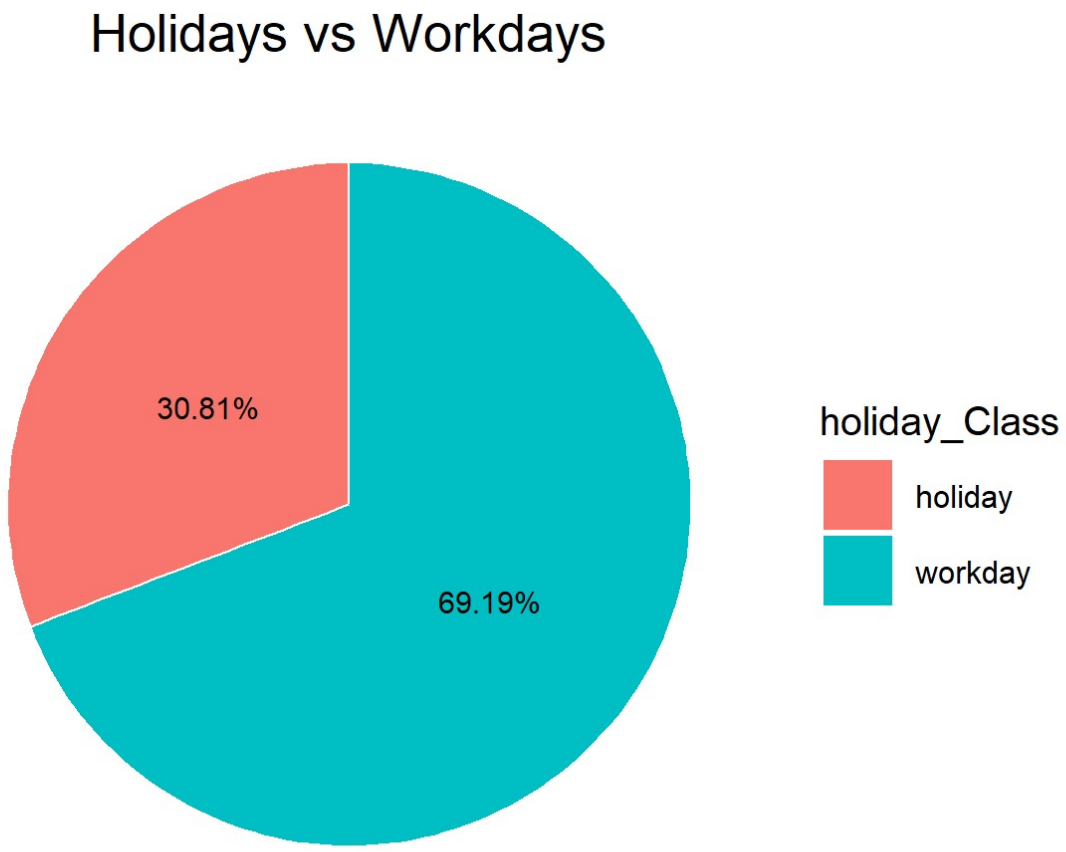
```
total_num_of_days = sum(total_rides_by_holiday$num_of_days)

holidays <- total_rides_by_holiday %>%
  mutate(percentage = round(num_of_days/total_num_of_days*100, digits=2)) %>%
  select(holiday, num_of_days, percentage)
```

```
percentage3 <- round(holidays$percentage, digits=2)

holiday_Class <- holidays$holiday
holiday_Class
```

```
## [1] "holiday" "workday"
```

```
ypos = cumsum(percentage3) - 0.5 * percentage3
ypos = 100 - ypos


pie3 = ggplot() + theme_bw() +
  geom_bar(aes(x = "", y = percentage3, fill = holiday_Class),
           stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  ggtitle("Holidays vs Workdays") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid=element_blank(),
        panel.border = element_blank()) +
  theme(legend.text=element_text(size=12),
        legend.title = element_text(hjust = 0.5, size=15),
        legend.key.size = unit(1,"cm")) +
  geom_text(aes(x = "", y = ypos, label = paste0(percentage3, "%")), size = 4)
```

```
pie3
```



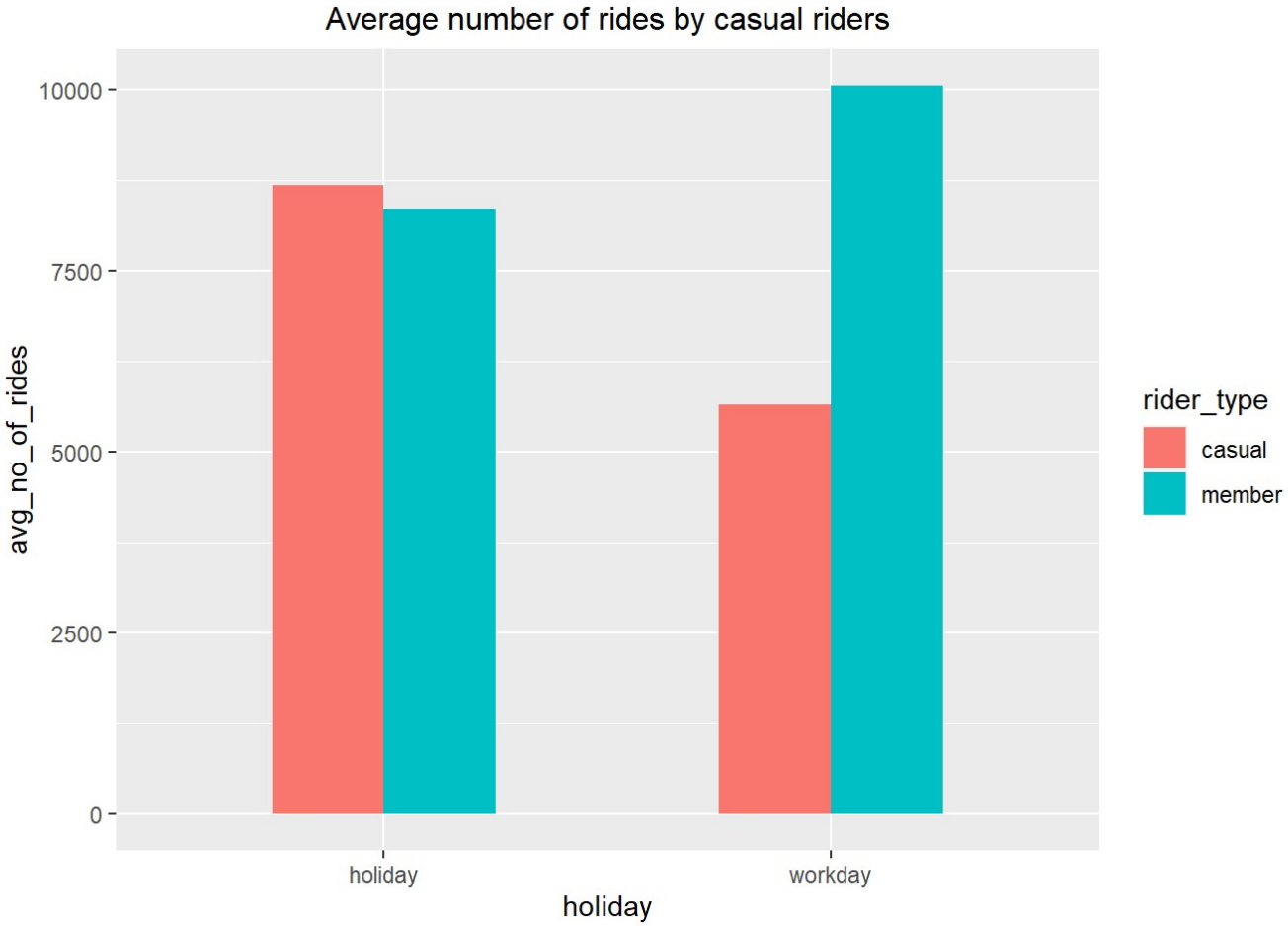

```
rides_by_holiday
```

```
##   holiday num_of_days rider_type no_of_rides avg_no_of_rides
## 1 holiday         122     casual     1060024         8688.72
## 2 workday         274     casual     1549928         5656.67
## 3 holiday         122     member     1019811         8359.11
## 4 workday         274     member     2756521        10060.30
```

Average number of rides

```
positions5 <- c("holiday", "workday")

bar12 <- ggplot(data=rides_by_holiday, aes(x=holiday, y=avg_no_of_rides, fill=rider_type))
+
  geom_bar(stat="identity", width = 0.5, position = "dodge")+
  ggtitle("Average number of rides by casual riders")+
  scale_x_discrete(limits = positions5)+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar12
```



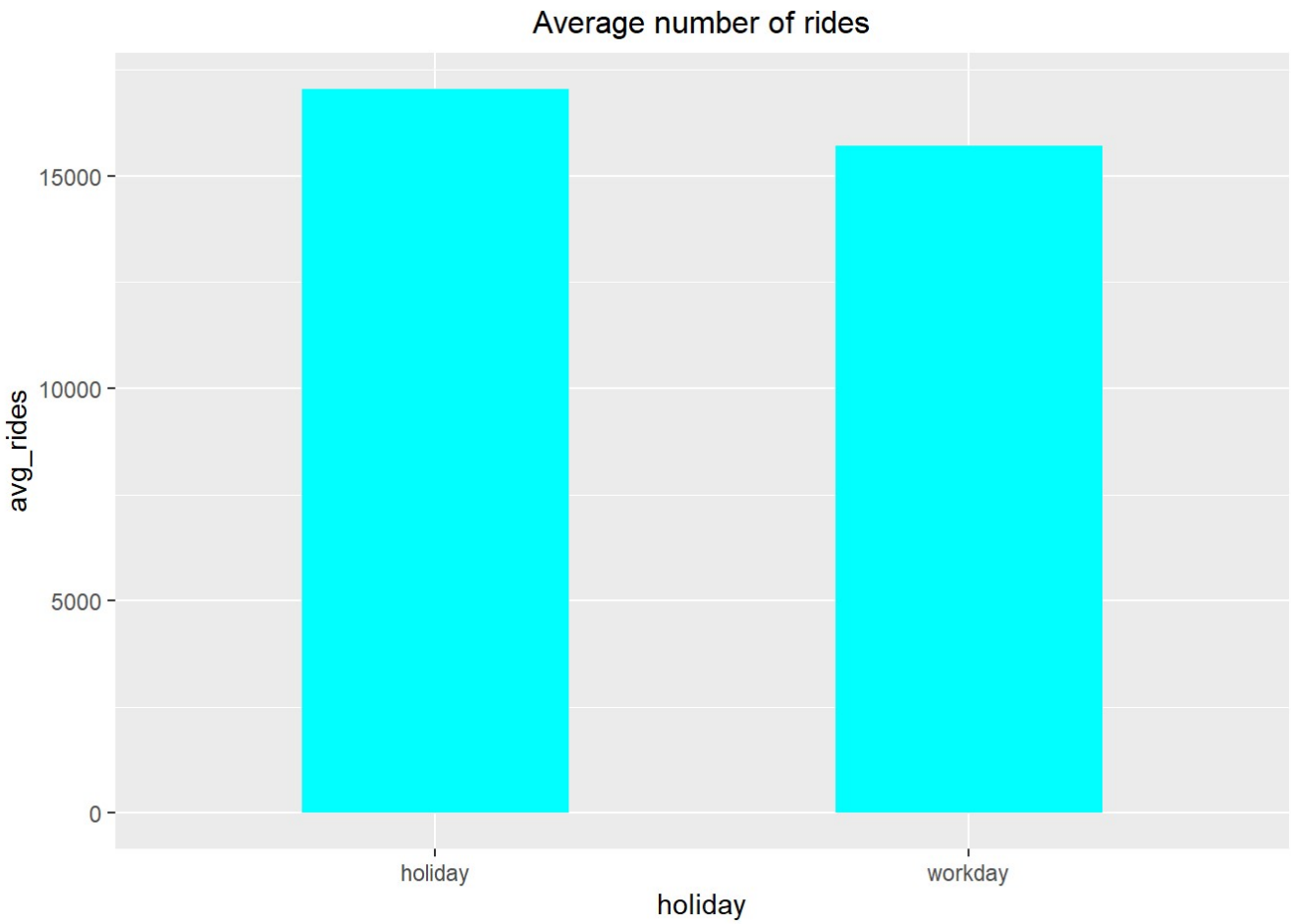

```
rides_by_holiday
```

```
##   holiday num_of_days rider_type no_of_rides avg_no_of_rides
## 1 holiday         122     casual     1060024         8688.72
## 2 workday         274     casual     1549928         5656.67
## 3 holiday         122     member     1019811         8359.11
## 4 workday         274     member     2756521        10060.30
```

Casual riders used more numbers of bikes during holidays, members preferred to use more bikes on Workdays. During holidays, number of casual riders was more than members. During non-holiday days, number of members was more than casual riders.

```
bar13 <- ggplot(data=all_riders_tbl, aes(x=holiday, y=avg_rides)) +
  geom_bar(stat="identity", width = 0.5, fill="cyan")+
  ggtitle("Average number of rides")+
  scale_x_discrete(limits = positions5)+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar13
```




```
all_riders_tbl
```

```
##   holiday   total num_of_days avg_rides
## 1 holiday 2079835         122  17047.83
## 2 workday 4306449         274  15716.97
```

In general average number of bikes used on holidays was more than workdays.

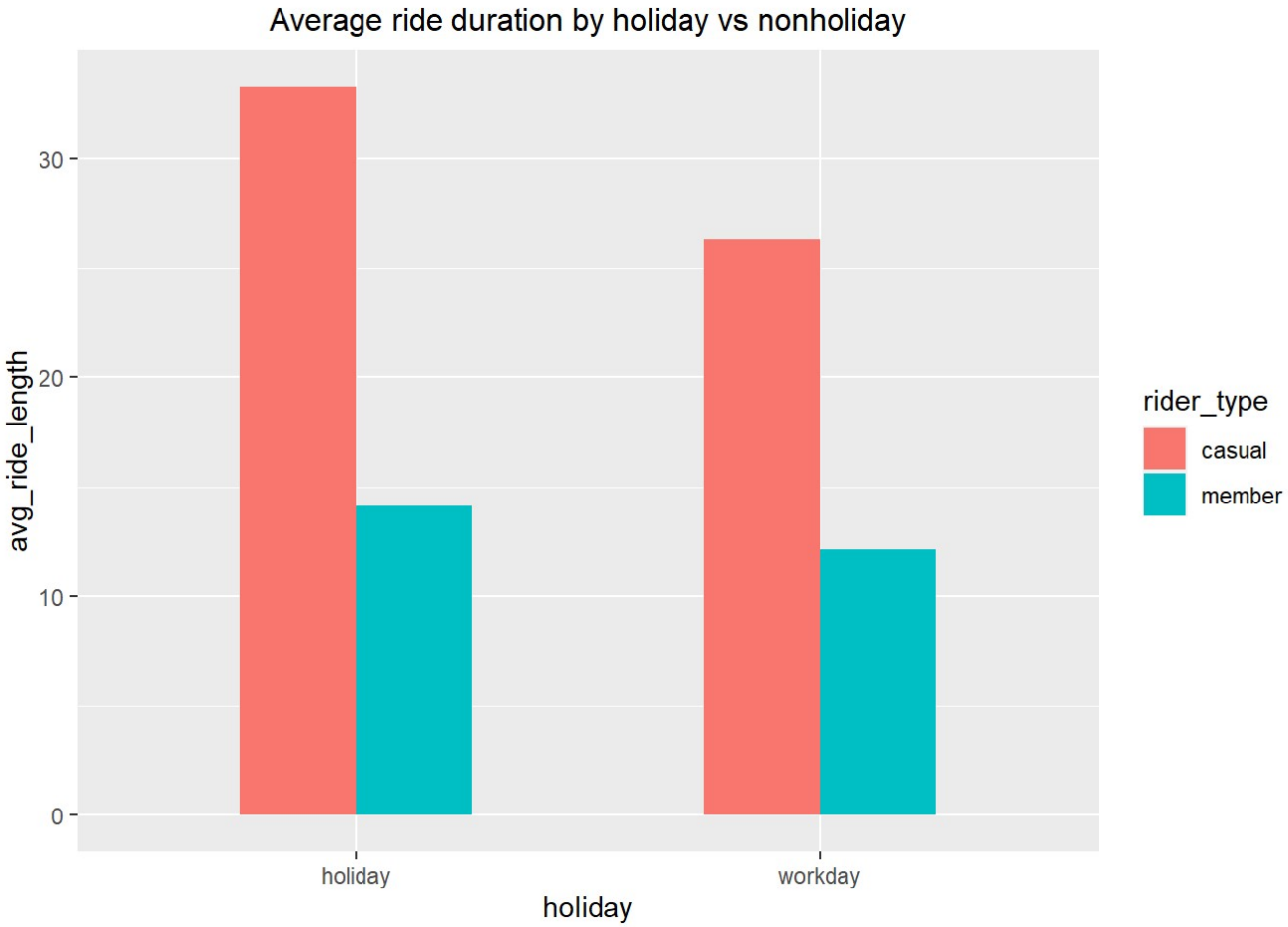
Average ride duration

```
avg_ride_length_by_holiday_tbl <- df %>%
  group_by(holiday, rider_type) %>%
  summarise(avg_ride_length = round(mean(ride_length),digits=2))

avg_ride_length_by_holiday_tbl
```

```
## # A tibble: 4 × 3
## # Groups:   holiday [2]
##   holiday rider_type avg_ride_length
##   <chr>   <chr>                <dbl>
## 1 holiday casual                33.3
## 2 holiday member                14.1
## 3 workday casual                26.3
## 4 workday member                12.2
```

```
bar14 <- ggplot(data=avg_ride_length_by_holiday_tbl, aes(x=holiday, y=avg_ride_length, fil
l=rider_type)) +
  geom_bar(stat="identity", width = 0.5, position = "dodge")+
  ggtitle("Average ride duration by holiday vs nonholiday")+
  scale_x_discrete(limits = positions5)+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```

```
bar14
```




```
avg_ride_length_by_holiday_tbl
```

```
## # A tibble: 4 × 3
## # Groups:   holiday [2]
##   holiday rider_type avg_ride_length
##   <chr>   <chr>                <dbl>
## 1 holiday casual                33.3
## 2 holiday member                14.1
## 3 workday casual                26.3
## 4 workday member                12.2
```

Rides duration were seen to be more on holidays compared to work days.

5. Share

Understanding how members and casual riders differ from one another is a key to understanding how we

can target and convert the casual riders into members.

Findings:

- Number of rides by members almost 20% higher than casual riders, which indicates that members used bikes more frequently.
- On average, one ride lasted less than 15 minutes for a member, and around 30 minutes for casual rider. This indicates that most members used the bike to travel shorter distances while casual riders used it to travel comparatively longer distance. This may also mean that, casual riders may have stopped in between the ride- may be for site-seeing or coffee breaks or taking pictures.
- Bike usage for members was more on weekdays and least on Sundays. Contrarily, number of casual riders was seen to highest on weekends. The number of casual riders was more on Fridays and Monday compared to other weekdays. It can be due to long weekends, but more analysis will be needed to confirm our theory.
- Slighty longer rides duration were observed in members on weekends compared to weekdays. It implies that members used the bikes to just to travel to work or university from their place of residence on weekdays and for leisure activities or fixed chores on weekends. May be some members use the bike to exercise daily on their regular routes and opt for longer routes on weekends.
- In casual riders, the rides were shortest on Wednesdays and they gradually kept on increasing till Sundays again to start reducing as the week starts. This suggests that, most casual riders riding the bikes on Weekends are visitors or tourists and those using the bikes from Tuesdays to Thursdays also comprise of local population, trying to get to work or somewhere important. These local people can be targeted in next marketing campaign as they are already familiar with Cyclistic and have plausibly use the service frequently.
- Casual riders prefer electric bike while members prefer classic bike. Given the fact he electric bikes are faster than classic bikes, the ride lengths must shorter for casual riders. But it’s the opposite, this reconfirms the notion that casual riders are using the bike for longer distances and stopping in between while members travel daily on a fixed short route. None of the members have used docked bikes, while a few casual rides have chosen docked bikes sometimes.
- For all the 13 months we analyzed, number of rides by members were always higher than number of rides by casual users. The difference was highest during November and December (Winter Season) and least during June and July (Summer).
- Maximum rides were taken in Summer followed by Fall. Winter saw least number of rides. which is understandable as Cyclistic is located in Chicago which is known for extremely harsh winters.
- Members took maximum rides during fall season while Casual riders took maximum rides during summer.
- Average ride duration of members was consistent throughout the year, while for casual riders it varied with season.
- Holiday wise analysis showed that casual members used the service more on holidays while members used the service more on workdays. Rides spans were more on holidays and and less on workdays.

Recommendations:

Following groups can be targeted for marketing campaign:

- Casual riders comprise mostly of visitors, targeting frequent visitor who have used Cyclistic more than

4-5 times can be easily persuaded for annual memberships.
- The trends show that, casual riders also contain certain proportion of local residents, who have rented the bike but have not enrolled, with proper marketing and special offers even they can be seen as potential members.
- There are at least 4100 casual riders which paid more than 108$ for a single ride which is the price of annual membership. They can be easily convinced for memberships too.

More data will be needed to exactly identify these groups of people.

- As summer has been the most popular season among casual riders, the campaign can be run around the mid-winter to throughout the summer, to engage with them.
- Cyclistic has all-day pass, but for some visitors- monthly memberships can be more beneficial. Cyclistic can release more plans and discounts for such customers.
- Referral discounts during busiest seasons are a great way to attract more customers.

The marketing campaign can focus on following points while marketing:

- Bike is the cheapest option compared to other public transports.
- Bikes are readily available, instead of waiting for a bus or a lyft, people can ride the bike on their own schedule.
- Biking is a great exercise.
- No parking fees or gas money needed.
- Biking is great for environment.