Cyclistic

JQH

10/4/2021

Cyclistic Data Analysis Project (Track 1 of Google Data Analytics Certificat

ASK Step

The client has provided a number of files split into quarters and contain the ride sharing data for 2021. this data will be loaded , explored and cleaned as needed to prepare for analysis.

Business Task The main business Task required from this part of the project is :

> Understanding the main difference between the Casual and Member riders of the Cyclistic company. and based on the insights and Data , a new marketing strategy will be formed.

Stakeholders The main team involved for this project are :

- Lily Moreno -> director of marketing and will be responsible to form the intended new strategy based on the data insights
- Cyclistic marketing analytics team -> the collected insights from this analysis will be combined with other work from this team for the final analysis
- Cyclistic Exec team -> final approval on going ahead with the program will come from this team.

Prepare Step

Data Structure and location The data sets were downloaded from client and stored temporarily for analysis. all the data sets for 2021 were given in a csv file format and will be inspected for any issues by using R analysis tools and later after cleaning.

Data Ethics and Privacy Since the data is coming directly from Motivate international and is assumed to be fit for the business case study purpose. The data is under a licence agreement which prohibits storing or distributing this data in any way. Inferring or atempting to relate this data to any individuals or persons who have used the cyclistic service.

Process Step

The process step will be to combine all the datasets into one main set and start the cleaning and analysis process. since we are using an R markdown file all cleaning steps will be documented below in R code blocks.

Load and stage the data for cleaning first we load our required libraries for data manipulation and processing

```r
# examine each file to ensure consistent col names and investigate the data

str(df1)
```

Examine the data to determine problems

```
## 'data.frame':    96834 obs. of  13 variables:
##  $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-0
##  $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-0
##  $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St" "California Av
##  $ start_station_id  : chr  "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```r
str(df2)
```

```
## 'data.frame':    49622 obs. of  13 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E7
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 19:10:18" "2021-0
##  $ ended_at          : chr  "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 19:19:10" "2021-0
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake S
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Rand
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num  42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```r
str(df3)
```

```
## 'data.frame':    228496 obs. of  13 variables:
##  $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-03-16 08:32:30" "2021-03-28 01:26:28" "2021-03-11 21:17:29" "2021-0
##  $ ended_at          : chr  "2021-03-16 08:36:34" "2021-03-28 01:36:55" "2021-03-11 21:33:53" "2021-0
##  $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields A
##  $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted
##  $ end_station_id    : chr  "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num  41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(df4)
```

```
## 'data.frame':    337230 obs. of  13 variables:
##  $ ride_id           : chr  "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887262AD101C6
##  $ rideable_type     : chr  "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-04-12 18:25:36" "2021-04-27 17:27:11" "2021-04-03 12:42:45" "2021-
##  $ ended_at          : chr  "2021-04-12 18:56:55" "2021-04-27 18:31:29" "2021-04-07 11:40:24" "2021-
##  $ start_station_name: chr  "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blvd & 84th S
##  $ start_station_id  : chr  "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr  "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loomis Blvd &
##  $ end_station_id    : chr  "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "member" ...
```

```
str(df5)
```

```
## 'data.frame':    531633 obs. of  13 variables:
##  $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D39110C
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-05-30 11:58:15" "2021-05-30 11:29:14" "2021-05-30 14:24:01" "2021-
##  $ ended_at          : chr  "2021-05-30 12:10:39" "2021-05-30 12:14:09" "2021-05-30 14:25:13" "2021-
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(df6)
```

```
## 'data.frame':    822410 obs. of  13 variables:
##  $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8A
##  $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 11:30:55" "2021-
##  $ ended_at          : chr  "2021-07-02 15:19:58" "2021-07-07 17:16:09" "2021-07-25 11:48:45" "2021-
##  $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Ave
##  $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubbard
##  $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

```
str(df7)
```

```
## 'data.frame':    822410 obs. of  13 variables:
##  $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8A
##  $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 11:30:55" "2021-0
##  $ ended_at          : chr  "2021-07-02 15:19:58" "2021-07-07 17:16:09" "2021-07-25 11:48:45" "2021-0
##  $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Ave &
##  $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubbard
##  $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

```
str(df8)
```

```
## 'data.frame':    804352 obs. of  13 variables:
##  $ ride_id           : chr  "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFAF1
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-08-10 17:15:49" "2021-08-10 17:23:14" "2021-08-21 02:34:23" "2021-0
##  $ ended_at          : chr  "2021-08-10 17:22:44" "2021-08-10 17:39:24" "2021-08-21 02:50:36" "2021-0
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

```
##  [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"
```

Looking at the Structure of each dataset we can see that they can be easily combined into one master file for easier cleanup and exploring

```
# Create one Variable Data frame with all the df's combined with rbind function (this requires the data

bike_rides_2021 <- rbind(df1,df2,df3,df4,df5,df6,df7,df8)

# View a Summary of the data to see the data structure and investigate any issues with the data

skim_without_charts(bike_rides_2021)
```

Table 1: Data summary

Name	bike_rides_2021
Number of rows	3692987
Number of columns	13
Column type frequency:	
character	9
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	2870577	0
rideable_type	0	1	11	13	0	3	0
started_at	0	1	19	19	0	2403002	0
ended_at	0	1	19	19	0	2399964	0
start_station_name	0	1	0	53	370303	747	0
start_station_id	0	1	0	36	370301	734	0
end_station_name	0	1	0	53	399161	746	0
end_station_id	0	1	0	36	399161	734	0
member_casual	0	1	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	0.04	41.64	41.88	41.90	41.93	42.07
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-87.52
end_lat	3371	1	41.90	0.04	41.54	41.88	41.90	41.93	42.15
end_lng	3371	1	-87.65	0.03	-88.07	-87.66	-87.64	-87.63	-87.49

As we can see from the data skim function , there is some cleaning needed :

- Some missing values in the end_lat and end_long data , start_station_name , end_station_name)
- data types need to be fixed (dates : started_at , ended_at)

```
# drop na and null values from the Dataset

bike_rides_2021 <- drop_na(bike_rides_2021)

# Change data types of the array data

bike_rides_2021$started_at <- lubridate::as_datetime(bike_rides_2021$started_at)
bike_rides_2021$ended_at <- lubridate::as_datetime(bike_rides_2021$ended_at)

# Calculate ride length and add new column

bike_rides_2021$ride_length <- difftime(bike_rides_2021$ended_at,bike_rides_2021$started_at, units = c(

# add the day of the week to the data
bike_rides_2021$day_of_week <- lubridate::wday(bike_rides_2021$started_at , label = TRUE )

#clean the ride_length to remove negative values
bike_rides_2021 <- bike_rides_2021 %>%  filter(bike_rides_2021$ride_length > 0 )
```

```
# View a Summary of the data after cleaning to check

skim_without_charts(bike_rides_2021)
```

Cleaning the Data

Table 4: Data summary

Name	bike_rides_2021
Number of rows	3689232
Number of columns	15
Column type frequency:	
character	7
difftime	1
factor	1
numeric	4
POSIXct	2

Table 4: Data summary

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	2867635	0
rideable_type	0	1	11	13	0	3	0
start_station_name	0	1	0	53	370294	747	0
start_station_id	0	1	0	36	370292	734	0
end_station_name	0	1	0	53	395530	746	0
end_station_id	0	1	0	36	395530	734	0
member_casual	0	1	6	6	0	2	0

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ride_length	0	1	1 secs	3235296 secs	779 secs	20037

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
day_of_week	0	1	TRUE	7	Sat: 697777, Sun: 568473, Fri: 557653, Thu: 485592

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.04	41.64	41.88	41.90	41.93	42.07
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-87.52
end_lat	0	1	41.90	0.04	41.54	41.88	41.90	41.93	42.15
end_lng	0	1	-87.65	0.03	-88.07	-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	2021-01-01 00:02:05	2021-08-31 23:59:35	2021-07-12 21:25:28	2400807
ended_at	0	1	2021-01-01 00:08:39	2021-09-01 17:21:36	2021-07-12 21:48:21	2397716

Analyze Step

In this step we will work to perform some calculations on grouped data as well as some basic trend and relationship insights

```
# create ride length variable by converting original from factor to numeric

bike_rides_2021$ride_length_num <- as.numeric(bike_rides_2021$ride_length)

#focus the analysis on the Member vs Casual data so will aggreagte based on the two conditons of memebe

mean_rides <- aggregate(bike_rides_2021$ride_length_num ~ bike_rides_2021$member_casual , FUN = mean)

max_rides <- aggregate(bike_rides_2021$ride_length_num ~ bike_rides_2021$member_casual , FUN = max)

mean_rides_day_of_week <- aggregate(bike_rides_2021$ride_length_num ~ bike_rides_2021$member_casual + bi

#aggregate the data based on the station location

loc_rides_start <- aggregate(bike_rides_2021$ride_length ~ bike_rides_2021$start_station_name+bike_ride

#export the data for external analysis and plotting using Tableu software

write_csv(loc_rides_start , file= "~/Bike_data_location_based.csv" )
```

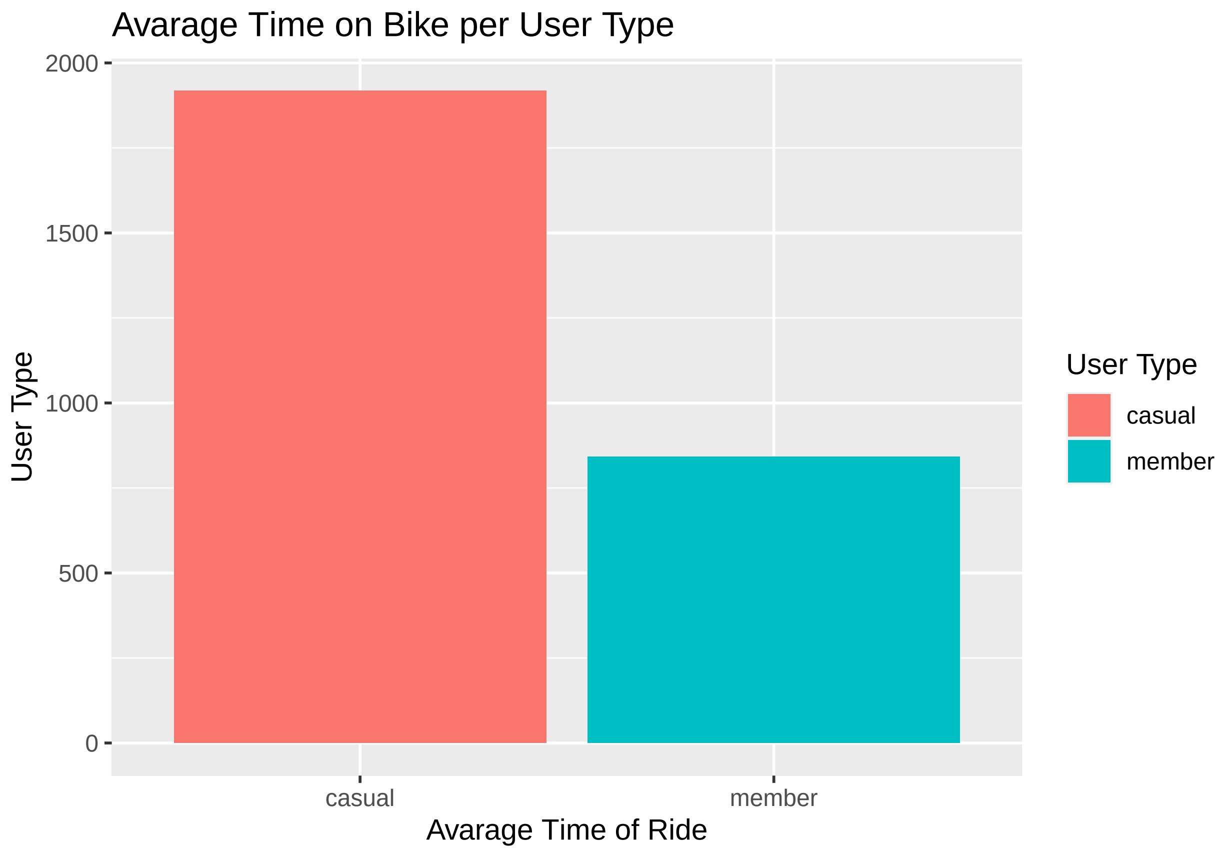
Visualize and gain more insight

```
# Start to visualize the data

#loading the ggplot for creating charts
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

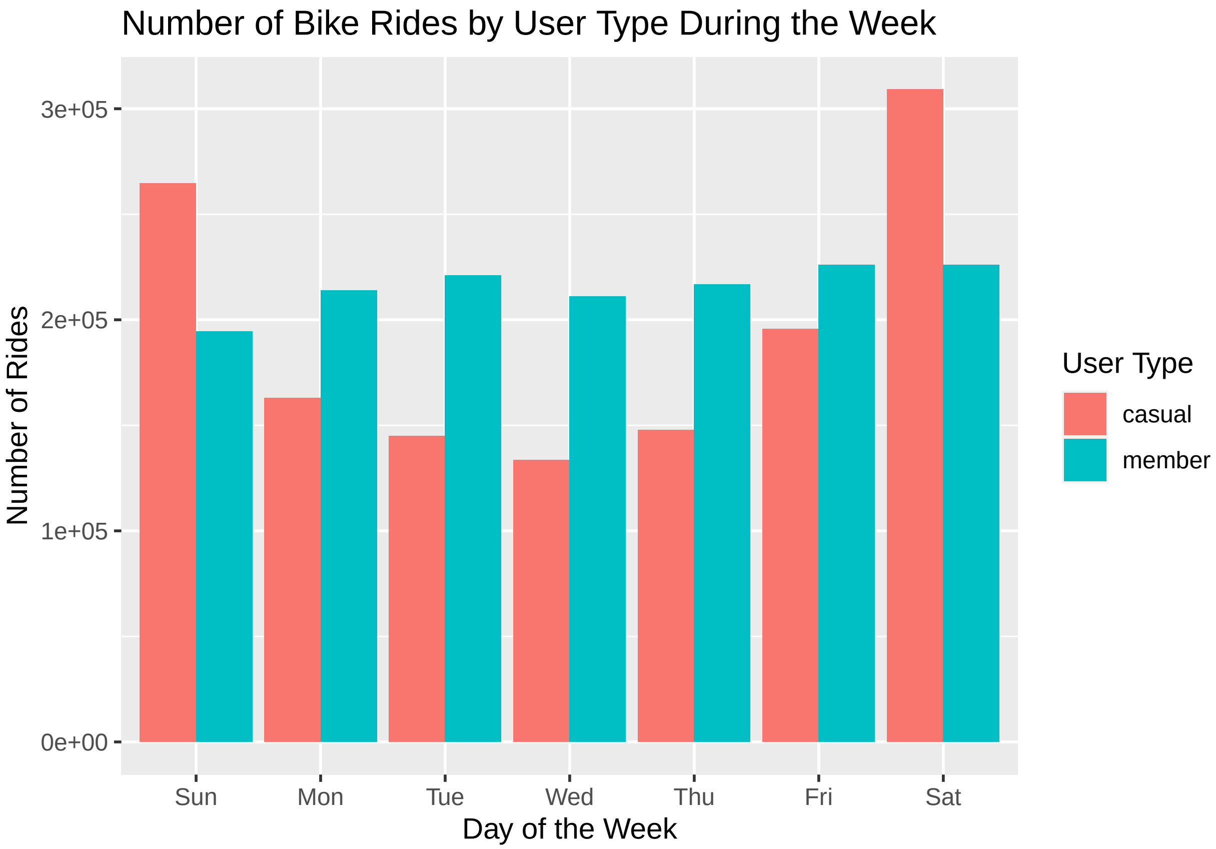
```
# create a col plot to show the avg time per user type using the service
bike_rides_2021 %>% group_by(member_casual) %>% summarise(avg = mean(ride_length)) %>%
  ggplot( aes(x =member_casual , y = avg  , fill = member_casual))+
  geom_col() +
  labs(title = "Avarage Time on Bike per User Type"  , x = "Avarage Time of Ride" , y = "User Type" , f
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



```
# view the same type of data but by week day to understand the behavior through out the week
bike_rides_2021 %>% group_by(day_of_week , member_casual) %>% summarise(number_rides = n_distinct(ride_
  ggplot( aes ( x = day_of_week , y = number_rides , fill = member_casual ))+
  geom_col(position = "dodge")+
  labs(title = "Number of Bike Rides by User Type During the Week"  , x = "Day of the Week" , y = "Numbe
```

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



```
# creating a density plot to understand high density ride length value
density <- bike_rides_2021 %>% filter(ride_length<3600)  %>%
  ggplot( aes(x = ride_length , fill = member_casual)) +
  geom_density()+
  labs(title = "Density of  Ride Times for the User Groups"  , x = "Ride Time (length)" , y = "Density"

histogram <- bike_rides_2021 %>% filter(ride_length<3600)  %>%
  ggplot( aes(x = ride_length , fill = member_casual)) +
  geom_histogram()+
  labs(title = "HistoGram of  Ride Times for the User Groups"  , x = "Ride Time (length)" , y = "Densit

grid.arrange(density , histogram , ncol =2 )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.

## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```