Bike_sharing_markdown

NiYa Wang

2021/10/23

Prepareation

Install and load packages

```
install.packages("tidyverse")
install.packages("dplyr")
library(tidyverse)
library(skimr)
library(dplyr)
```

```
## -- Attaching packages --------------------------------------- tidyve
rse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

Import datasets

```
q1_2019 <- read_csv("Divvy_Trips_2019_Q1.csv")
q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
q1_2020 <- read_csv("Divvy_Trips_2020_Q1.csv")
```

Data from dirty to clean

Change column usertype value 'Subscriber' and 'Customer' to 'member' and 'casual'

```
q1_2019$usertype <- gsub("Subscriber", "member", q1_2019$usertype)
q2_2019$usertype <- gsub("Subscriber", "member", q2_2019$usertype)
q3_2019$usertype <- gsub("Subscriber", "member", q3_2019$usertype)
q4_2019$usertype <- gsub("Subscriber", "member", q4_2019$usertype)
q1_2020$usertype <- gsub("Subscriber", "member", q1_2020$usertype)

q1_2019$usertype <- gsub("Customer", "casual", q1_2019$usertype)
```

```
q2_2019$usertype <- gsub("Customer", "casual", q2_2019$usertype)
q3_2019$usertype <- gsub("Customer", "casual", q3_2019$usertype)
q4_2019$usertype <- gsub("Customer", "casual", q4_2019$usertype)
q1_2020$usertype <- gsub("Customer", "casual", q1_2020$usertype)
```

Consistent column type of trip_id as character

```
q1_2019$trip_id <- as.character(q1_2019$trip_id)
q2_2019$trip_id <- as.character(q2_2019$trip_id)
q3_2019$trip_id <- as.character(q3_2019$trip_id)
q4_2019$trip_id <- as.character(q4_2019$trip_id)
q1_2020$trip_id <- as.character(q1_2020$trip_id)
```

Combine all datasets into a variable cyclistic_all_data

```
cyclistic_all_data <- bind_rows(q1_2019,q2_2019,q3_2019,q4_2019,q1_202
0)
```

Transfer birthyear, day_of_week and ride_length as integer format

```
cyclistic_all_data$birthyear <- as.integer(cyclistic_all_data$birthyea
r)
cyclistic_all_data$day_of_week <- as.integer(cyclistic_all_data$day_of_
week)
cyclistic_all_data$ride_length <- as.integer(cyclistic_all_data$ride_le
ngth)/60
```

Glance of data

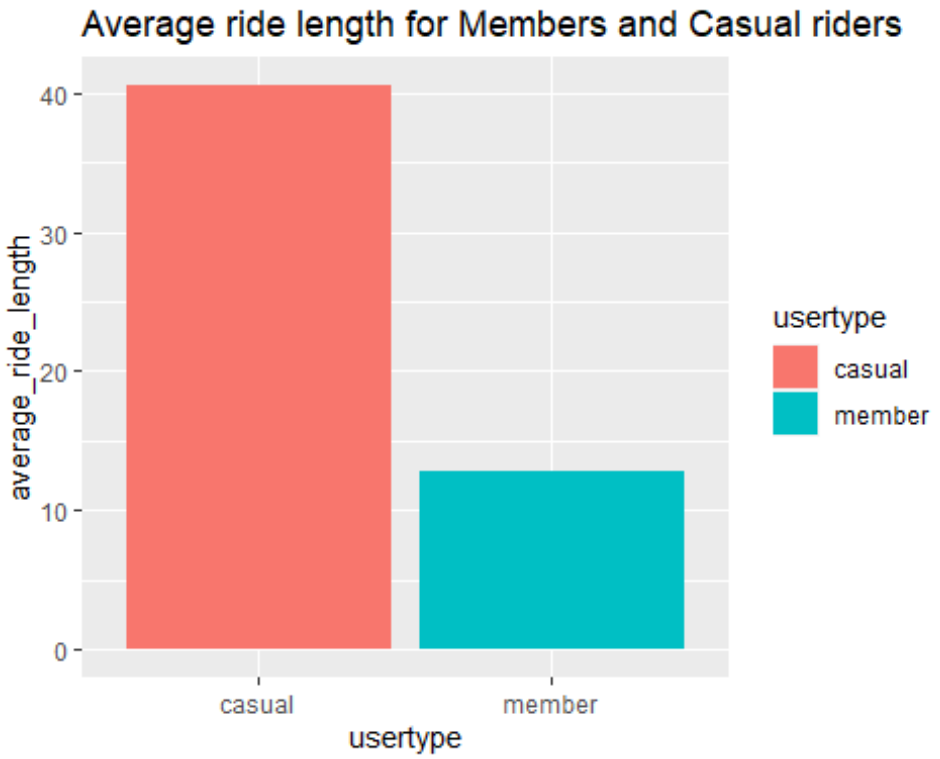
```
head(cyclistic_all_data)

## # A tibble: 6 x 7
##   trip_id  usertype day_of_week ride_length from_station_name  to_st
ation_name
##   <chr>    <chr>          <int>       <dbl> <chr>              <chr>

## 1 21743665 member             2        1.02 Blue Island Ave &~ Blue
Island Ave ~
## 2 21763499 casual             6        1.02 Michigan Ave & Id~ Michi
gan Ave & I~
## 3 21794469 member             4        1.02 Canal St & Adams ~ Canal
 St & Adams~
## 4 21809886 member             7        1.02 Sedgwick St & Sch~ Sedgw
ick St & Sc~
## 5 21847368 member             2        1.02 Field Museum       Field
 Museum
## 6 21863790 member             2        1.02 Sheffield Ave & W~ Sheff
ield Ave & ~
## # ... with 1 more variable: birthyear <int>

glimpse(cyclistic_all_data)
```

```
## Rows: 3,592,937
## Columns: 7
## $ trip_id           <chr> "21743665", "21763499", "21794469", "21809
886", "218~
## $ usertype          <chr> "member", "casual", "member", "member", "m
ember", "m~
## $ day_of_week       <int> 2, 6, 4, 7, 2, 2, 4, 3, 4, 2, 3, 1, 5, 1,
1, 4, 5, 2~
## $ ride_length       <dbl> 1.016667, 1.016667, 1.016667, 1.016667, 1.
016667, 1.~
## $ from_station_name <chr> "Blue Island Ave & 18th St", "Michigan Ave
 & Ida B W~
## $ to_station_name   <chr> "Blue Island Ave & 18th St", "Michigan Ave
 & Ida B W~
## $ birthyear         <int> 1985, NA, 1989, 1993, 1970, 1998, 1981, 19
63, 1970, ~
```

Visualizations

Average ride_length for members and casual riders

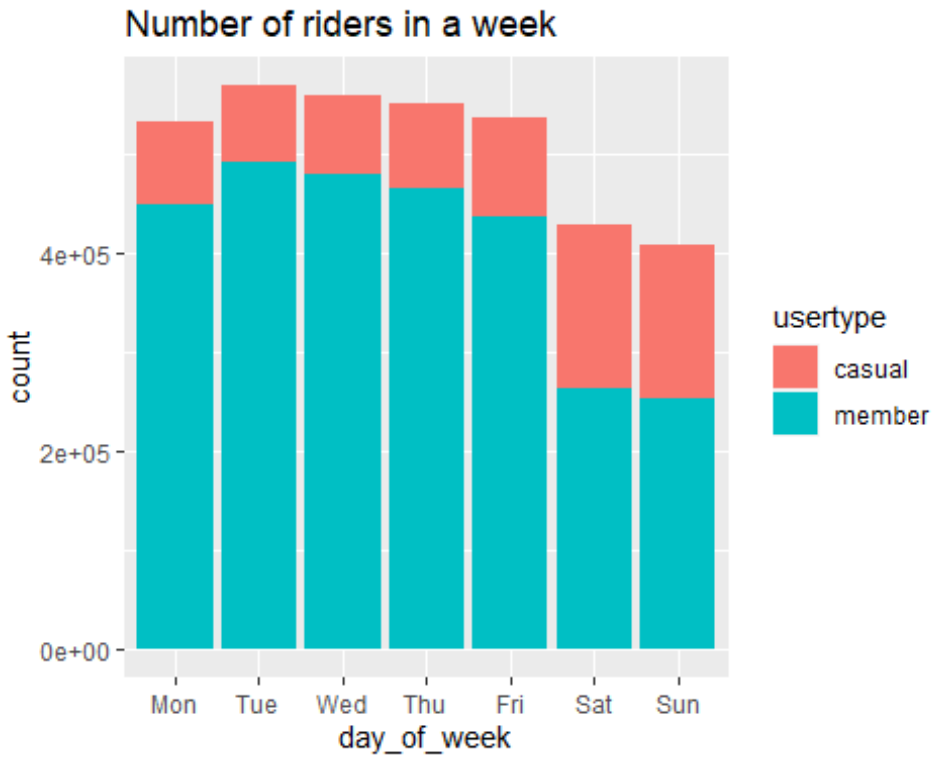
```
average_length <- cyclistic_all_data %>%
  group_by(usertype) %>%
  summarize(average_ride_length = mean(ride_length))

ggplot(data = average_length) +
  geom_bar(stat = "identity", mapping = aes(x = usertype, y= average_ri
de_length, fill = usertype)) +
  labs(title = "Average ride length for Members and Casual riders")
```



Average

Number of riders in a week

```
day_week <-c('Mon','Tue','Wed','Thu','Fri','Sat','Sun')
ggplot(data = cyclistic_all_data) +
  geom_bar(mapping = aes(x = day_of_week,fill = usertype)) +
  labs(title = "Number of riders in a week")+
  xlim(day_week)
```



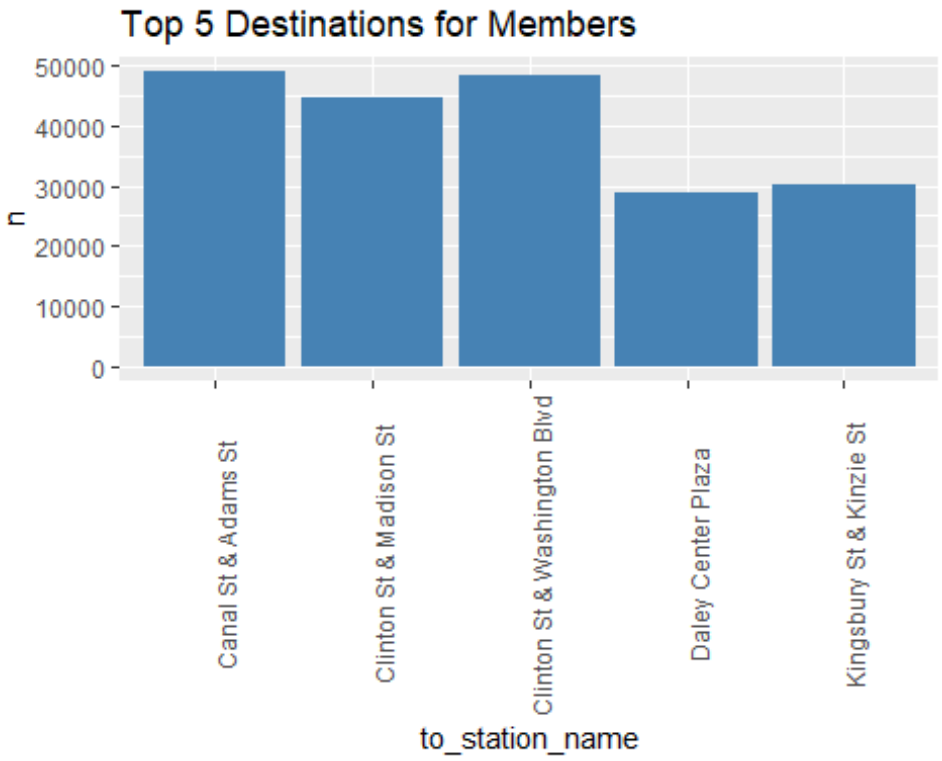
Investigate Destinations

```
num_station <-cyclistic_all_data %>%
  drop_na(to_station_name) %>%
  group_by(to_station_name) %>%
  count(usertype, sort=(decreasing=T))
num_station %>% arrange(num_station)
```

```
## # A tibble: 1,285 x 3
## # Groups:   to_station_name [644]
##    to_station_name            usertype     n
##    <chr>                      <chr>    <int>
##  1 2112 W Peterson Ave        casual      95
##  2 2112 W Peterson Ave        member     491
##  3 63rd St Beach              casual     569
##  4 63rd St Beach              member     454
##  5 900 W Harrison St          casual     654
##  6 900 W Harrison St          member    5468
##  7 Aberdeen St & Jackson Blvd casual     985
##  8 Aberdeen St & Jackson Blvd member   10707
##  9 Aberdeen St & Monroe St    casual     803
## 10 Aberdeen St & Monroe St    member    8645
## # ... with 1,275 more rows
```

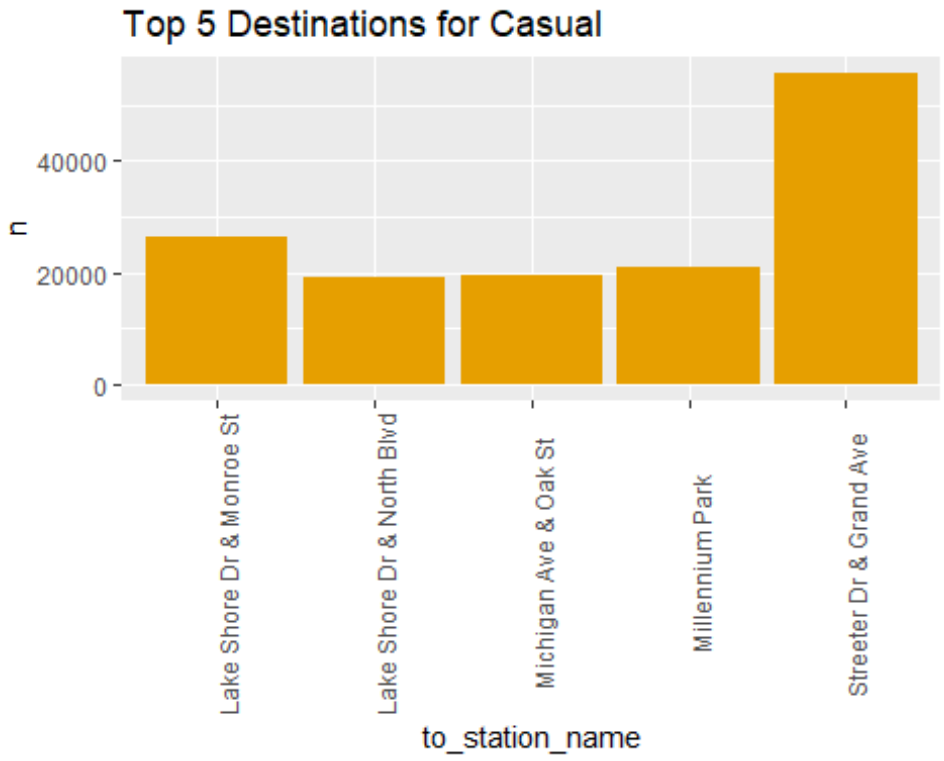
Top 5 Destinations for Members

```
top_member <- subset(num_station, usertype =="member") %>%
  head(5, )

ggplot(data = top_member) +
  geom_bar(stat = "identity", mapping = aes(x = to_station_name, y =
n), fill="steelblue") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Top 5 Destinations for Members")
```



Top 5 Destinations for Casuals

```
top_casual <- subset(num_station, usertype =="casual") %>%
  head(5, )

ggplot(data = top_casual) +
  geom_bar(stat = "identity", mapping = aes(x = to_station_name, y =
n), fill="#E69F00") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Top 5 Destinations for Casual")
```



Distribution of rider's ages

```
age <- 2021 - cyclistic_all_data$birthyear
ggplot(data = cyclistic_all_data, aes(x=age, fill=usertype)) +
  geom_histogram(bins =12, position="dodge") +
  xlim(10,80) +
  scale_y_continuous((name="number")) +
  labs(title="Distribution of user's age")
```