```
Bike Trips analysis
Musa Karimli
3/7/2022
Data importing
Let's first import data from sql:
```

```
Because data was too large for my memory to handle, I exported CSV files to a local SQL server(Postgres), then imported data from here.
```

```
library(tidyverse)
library('RPostgreSQL')
library(dplyr)
con <- dbConnect(drv =PostgreSQL(),</pre>
                 user="postgres",
                 password="sinif555a",
                 host="localhost",
                 port=5432,
                 dbname="customer_data")
dbListTables(con) #list all the tables
## [1] "customer_purchase" "workbook"
                                                 "movies"
   [4] "customer"
                            "nation"
                                                 "orders"
                            "region"
                                                 "customer_address"
## [7] "part"
                            "departments"
                                                 "automobile_data"
## [10] "employees"
```

```
## [13] "warehouse"
                              "orders"
                                                    "avocado"
## [16] "tripdata"
# query the order table
trips <- dbGetQuery(con, "SELECT * from trip.tripdata;")</pre>
trips<-as_tibble(trips)</pre>
glimpse(trips)
```

```
## Rows: 5,717,608
## Columns: 13
## $ ride_id <chr> "203C67DBB306DC92", "4121F200FFC2B153", "B75B91A856~
## $ start_station_name <chr> "Woodlawn Ave & 55th St", "Bissell St & Armitage Av~
## $ start_station_id <chr> "TA1307000164", "13059", "KA1504000171", "TA1308000~
## $ end_station_id <chr> NA, "TA1309000006", NA, NA, NA, NA, NA, NA, NA, NA, "13042"~
## $ start_lat
                 <dbl> 41.79529, 41.91844, 41.97164, 41.95445, 41.85787, 4~
## $ start_lng
## $ end_lat
## $ end_lng
                 <dbl> -87.59647, -87.65222, -87.65022, -87.64807, -87.640~
                 <dbl> 41.79000, 41.91808, 41.94000, 41.95000, 41.88000, 4~
                 <dbl> -87.60000, -87.64375, -87.67000, -87.65000, -87.620~
## $ end_lng
                 <chr> "member", "casual", "casual", "casual", "~
## $ member_casual
```

```
summary(trips)
   ride_id
                rideable_type
                              started_at
                                            ended_at
## Length:5717608
                Length: 5717608
                             Length: 5717608
                                          Length: 5717608
  Mode :character Mode :character Mode :character Mode :character
```

```
##
##
##
##
##
   start_station_name start_station_id end_station_name end_station_id
   Length: 5717608 Length: 5717608 Length: 5717608
                                              Length: 5717608
   Class :character Class :character Class :character Class :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
   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. :42.17 Max. :-87.49
                   NA's :4831 NA's :4831
##
   member_casual
   Length: 5717608
   Class :character
   Mode :character
##
##
##
Data cleaning and transformation
```

excluding trips which are lasted zero seconds or below trips <- trips %>% filter(trip_duration>0) # extracting date components

there isn't any duplicate in "ride_id" column

finding the shortest distance between two locations

finding approximate speed of the ride

trips <- trips %>% mutate(distance_ctd = distHaversine(cbind(start_lng, start_lat),

trips<-transform(trips, speed=distance_ctd/as.double(trip_duration, units='secs'))</pre>

setting abnormal values to NA so it won't affect to the calculations

summarise("avg_duration_trip"=mean(trip_duration, na.rm=T),

"n_of_rides"=format(n(), scientific=F))

<drtn>

docked_bike 4929.1841 secs

electric_bike 753.2742 secs

"avg_speed"=mean(speed, na.rm=T),

"n_of_rides"=format(n(), scientific=F))

... with 2 more variables: avg_speed <dbl>, n_of_rides <chr>

summarise("avg_duration_trip"=mean(trip_duration, na.rm=T),

"n_of_rides"=format(n(), scientific=F))

"avg_speed"=mean(speed, na.rm=T),

—these values are not considered in several below analyses.

Number of rides by membership

· Casual riders move more on weekends because of additional free time.

trips %>% group_by(start_year,start_quarter,start_month,member_casual) %>%

"avg_trip_distance"=mean(distance_ctd, na.rm=TRUE),

"avg_speed"=mean(speed, na.rm=T),

"avg_trip_distance"=mean(distance_ctd, na.rm=TRUE),

member_casual rideable_type avg_duration_trip avg_trip_distance avg_speed

trips['trip_duration']=trips\$ended_at-trips\$started_at

1. Dealing With date and time:

converting string to datetime

calculating trip duration

trips\$ended_at <- ymd_hms(trips\$ended_at)</pre> trips\$started_at <- ymd_hms(trips\$started_at)</pre>

library(lubridate)

```
trips$start_year <- year(trips$started_at)</pre>
trips$start_month <- month(trips$started_at)</pre>
trips$start_quarter <- quarter(trips$started_at)</pre>
trips$start_week <- week(trips$started_at)</pre>
trips$start_wday <- wday(trips$started_at)</pre>
trips$start_day <- day(trips$started_at)</pre>
trips$start_hour <- hour(trips$started_at)</pre>
trips$start_year_month <- floor_date(as_date(trips$started_at),"month")</pre>
  2. Dealing with null and duplicate values:
# replacing NA stations names with no info
trips <- trips %>%
    mutate(start_station_name = case_when(start_station_name==''~'No info',
```

is.na(start_station_name)~'No info',

TRUE ~start_station_name))%>%

is.na(end_station_name)~'No info',

mutate(start_station_name = case_when(end_station_name==''~'No info',

TRUE ~end_station_name))

```
trips %>% count(ride_id) %>%
    filter(n>1)
## # A tibble: 0 x 2
## # ... with 2 variables: ride_id <chr>, n <int>
 3. Cleaning String Data:
library(stringr)
trips <- trips %>%
    mutate(start_station_name = str_trim(start_station_name, side='both')) %>%
    mutate(end_station_name = str_trim(end_station_name, side='both')) %>%
    mutate(start_station_name = str_to_title(start_station_name)) %>%
    mutate(end_station_name = str_to_title(end_station_name))
 4. Adding additional columns for data analysis:
```

```
mutate(speed = case_when(speed==0 ~ NA_real_,
                            is.infinite(speed) ~ NA_real_,
                            TRUE ~ as.numeric(speed)))
 trips <- trips %>%
     mutate(distance_ctd = case_when(distance_ctd==0 ~ NA_real_,
                              is.infinite(distance_ctd) ~ NA_real_,
                              TRUE ~ as.numeric(distance_ctd)))
Data analysis
Looking at the data with summarization:
 trips %>% group_by(member_casual, rideable_type) %>%
```

cbind(end_lng, end_lat)))

<dbl>

2360.

2084.

2461.

2548. 1.64

2627. 3.10

<dbl>

2.30

2.87

3.70

1931.

trips %>% group_by(start_year,start_quarter,start_month) %>% summarise("avg_duration_trip"=mean(trip_duration, na.rm=T), "avg_trip_distance"=mean(distance_ctd, na.rm=TRUE),

A tibble: 5 x 6

<chr>

5 member

13

A tibble: 26 x 8

hours.

300 K -

300 K

200 K -

100 K -

0 K

Number of Rides 300 K - 200 K - 100 K

0 K

400 K

300 K

200 K

100 K

2 casual

Groups: member_casual [2]

<chr>

1 casual classic_bike 1738.8018 secs

3 casual electric_bike 1183.2579 secs
4 member classic_bike 845.7443 secs

... with 1 more variable: n_of_rides <chr>

library("geosphere")

trips <- trips %>%

A tibble: 13 x 7

```
## # Groups: start_year, start_quarter [5]
    start_year start_quarter start_month avg_duration_trip avg_trip_distance
##
        <dbl>
              <int> <dbl> <drtn>
                                                          <dbl>
## 1
                                                          2134.
  2
                                                          2299.
##
## 3
                                                          2332.
## 4
                                                          2381.
## 5
                                                          2431.
## 6
                                                          2447.
## 7
                                                          2437.
## 8
                                                          2420.
## 9
                                                          2260.
## 10
                                                          2062.
## 11
                                                          2039.
## 12
                                                          1877.
```

```
## # Groups: start_year, start_quarter, start_month [13]
            start_year start_quarter start_month member_casual avg_duration_trip
##
                      <dbl> <int> <dbl> <chr> <drtn>

      2021
      1
      2 casual
      2962.6862 secs

      2021
      1
      2 member
      1081.4072 secs

      2021
      1
      3 casual
      2289.6601 secs

      2021
      1
      3 member
      838.2379 secs

      2021
      2
      4 casual
      2281.5631 secs

      2021
      2
      4 member
      881.4493 secs

      2021
      2
      5 casual
      2294.1080 secs

      2021
      2
      5 member
      878.4177 secs

      2021
      2
      6 casual
      2227.5568 secs

      2021
      2
      6 member
      880.7195 secs

## 1
## 2
##
     4
## 5
## 6
## 7
## 8
## 9
## 10
## # ... with 16 more rows, and 3 more variables: avg_trip_distance <dbl>,
## # avg_speed <dbl>, n_of_rides <chr>

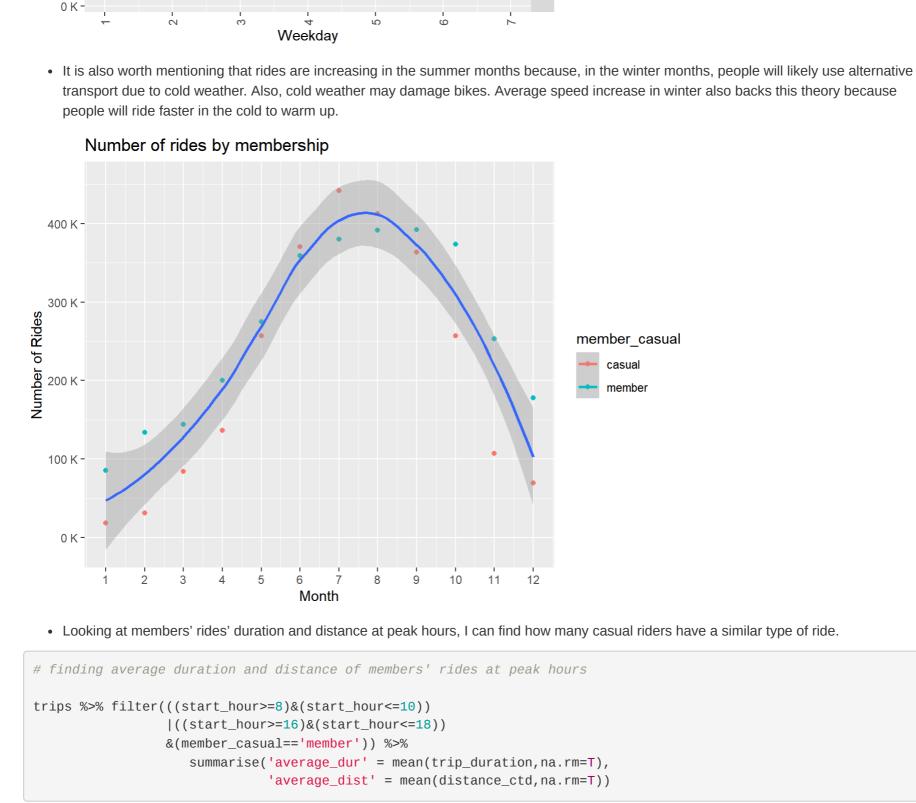
    We can see that members ride bikes faster and shorter rides than casual riders.

   • It is clear that members use a bike to commute to work daily. Also, exact start and end locations indicate that people didn't ride bikes to work
```

• The graph below supports this theory by showing that members' rides are much more than casuals' at the beginning and end of working

```
Number of Rides
                                                                                       member_casual
                                                                                           casual
                                                                                            member
     100 K
                                          10
                                               12
                                                     14
                                                           16
                                                                 18
                                            Hour
          Number of rides Trend
     400 K -
```

member_casual casual member



average_dur average_dist

finding similar type of casual rides

group_by(start_month, member_casual) %>%

summarise(mean(avg_rides)))/22

mean(avg_rides)

Let's dive deeper into rideable types:

1306.451

1

0.75 M

0.50 M -

0.25 M -

0.00 M

Wrap Up

bikes.

casual

member

filter((start_hour>=8&start_hour<=10)

|(start_hour>=16&start_hour<=18)) %>%

cat('Casual rides in peak hours:', nrow(casual_rides), sep = ' ')

summarise("avg_duration_trip"=mean(trip_duration, na.rm=T),

"avg_speed"=mean(speed, na.rm=T),

summarise('avg_rides'=mean(n_of_rides)) %>%

Number of rides by membership

"avg_trip_distance"=mean(distance_ctd, na.rm=TRUE),

"n_of_rides"=as.numeric(format(n(), scientific=F)))%>%

filter((member_casual=='casual')&(distance_ctd<=2287)</pre> &(trip_duration<duration(964, units = 'second')))

1 963.6844 secs

(casual_rides %>%

casual_rides <- trips %>%

Casual rides in peak hours: 344903 # Assuming those casual riders used bikes for commuting to work, dividing the average of these rides by working d ays, we could know how many casual riders we are missing from membership.

```
1.25 M
                                                                                 1.00 M -
                                                                                 0.75 M -
                                                                                 0.50 M -
0.25 M on the second of the se
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              rideable_type
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           classic bike
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            3
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           docked_bike
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               electric_bike
```

member

• It is evident that people prefer classic bikes. Docked bikes seem to be a new type of ride, electric bikes are less popular than classic bikes,

but it has the potential to become more popular, we can see in the trend line. Number of rides Trend 200 K 100 K Number of Rides docked_bike member_casual casual member 200 K 100 K -0 K -2021-03-01 2021-11-01 2021-12-01 2022-01-01 2021-04-01 2021-05-01 2021-06-01 2021-07-01 2021-08-01 2021-09-01 2021-10-01 2022-02-01

billboards may assist people in understanding the importance of lowering dependency on non-renewable energy.

accordingly. For example, bicycles may need better quality tires.

casual

Membership

I would propose incentivizing individuals to use cycles as daily transport to commute. Acknowledging that using bikes reduces carbon emissions and traffic jams will help people use them. There are still people who don't subscribe as members who use bicycles. Campaigns, emails, and

Further, seasonal subscriptions, for example, making summer and winter plans differently, will draw more people in the summer to subscribe. But

Investing in electric bikes may allow more people to commute because some people are also afraid of sweating, and it is less tiring than traditional

to expand membership in the winter, investing extra funds to have winter equipment to rent may benefit. Also, bikes need to be serviced