Problem Set template

Sandra Dai & Qianyu Shao

10/11/2021

```
library(tidyverse)
library(dplyr)
```

1 Front matter

This submission is my work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **Qianyu Shao & Sandra Dai**

Late coins used this pset: 0. Late coins left: 9.

2 Problems

2.1 Git merge conflicts

1. Succinctly explain, why did you have a merge conflict? because you and your partner both changed the name and Git is unable to automatically resolve differences in code between two commits.

2.2 Exploratory analysis

2.2.1 Download the data

1.

```
setwd("C:/Users/Administrator/Documents/GitHub/applied-problem-set-1-sandra_qianyu")
df<-read.csv("trips_mcma.csv")
```

2.

```
#print(df)
#head(df) quickly look at the data structure
#str(df) give the information of data type
#glimpse(df)
#View(df) open a new window to show the data completely
#summary(df) #give information about some statistics of the data
```

2.3 Let's see what's inside this data set:

1

```
summary(df)
```

```
##        X              id_trip        trip_number       mun_origin
##  Min.   :     1   Min.   :     1   Min.   : 1.000   Min.   :  1.00
##  1st Qu.:132899   1st Qu.:132899   1st Qu.: 1.000   1st Qu.:  7.00
##  Median :265798   Median :265798   Median : 2.000   Median : 16.00
##  Mean   :265798   Mean   :265798   Mean   : 1.696   Mean   : 34.76
##  3rd Qu.:398696   3rd Qu.:398696   3rd Qu.: 2.000   3rd Qu.: 57.00
##  Max.   :531594   Max.   :531594   Max.   :14.000   Max.   :318.00
##                                                     NA's   :6464
##   state_origin     dto_origin       mun_dest        state_dest
##  Min.   : 2.00   Min.   :  1.0   Min.   :  1.00   Min.   : 1.00
##  1st Qu.: 9.00   1st Qu.: 40.0   1st Qu.:  7.00   1st Qu.: 9.00
##  Median : 9.00   Median : 85.0   Median : 16.00   Median : 9.00
##  Mean   :11.99   Mean   :100.6   Mean   : 34.72   Mean   :11.99
##  3rd Qu.:15.00   3rd Qu.:156.0   3rd Qu.: 57.00   3rd Qu.:15.00
##  Max.   :31.00   Max.   :888.0   Max.   :318.00   Max.   :32.00
##  NA's   :1545    NA's   :20390   NA's   :7920     NA's   :2213
##     dto_dest          sex             age           origin
##  Min.   :  1.0   Min.   :1.000   Min.   : 6.00   Length:531594
##  1st Qu.: 40.0   1st Qu.:1.000   1st Qu.:22.00   Class :character
##  Median : 85.0   Median :2.000   Median :35.00   Mode  :character
##  Mean   :102.3   Mean   :1.519   Mean   :36.44
##  3rd Qu.:157.0   3rd Qu.:2.000   3rd Qu.:50.00
##  Max.   :888.0   Max.   :2.000   Max.   :97.00
##  NA's   :22846
##      dest              reason             dep_hour       dep_min
##  Length:531594      Length:531594      Min.   : 0.0   Min.   : 0.00
##  Class :character   Class :character   1st Qu.: 8.0   1st Qu.: 0.00
##  Mode  :character   Mode  :character   Median :12.0   Median : 0.00
##                                        Mean   :12.5   Mean   :12.62
##                                        3rd Qu.:16.0   3rd Qu.:30.00
##                                        Max.   :23.0   Max.   :59.00
##                                        NA's   :382    NA's   :382
##     arr_hour        arr_min       mode_trans            day
##  Min.   : 0.00   Min.   : 0.00   Length:531594      Length:531594
##  1st Qu.: 9.00   1st Qu.: 5.00   Class :character   Class :character
##  Median :13.00   Median :20.00   Mode  :character   Mode  :character
##  Mean   :13.07   Mean   :21.49
##  3rd Qu.:17.00   3rd Qu.:35.00
##  Max.   :23.00   Max.   :59.00
##  NA's   :382     NA's   :382
```

there are 20 variables, 531594 rows. Each row means an obervation. id_trip is a unique identifier for each trip

2 there are 528134 valid trips. 31 different states are the origin

```
names(df)
```

```
##  [1] "X"            "id_trip"      "trip_number"  "mun_origin"   "state_origin"
##  [6] "dto_origin"   "mun_dest"     "state_dest"   "dto_dest"     "sex"
## [11] "age"          "origin"       "dest"         "reason"       "dep_hour"
## [16] "dep_min"      "arr_hour"     "arr_min"      "mode_trans"   "day"
```

```
sum(is.na(df$state_origin))
```

```
## [1] 1545
```

```
sum(is.na(df$state_dest))
```

```
## [1] 2213
```

```
valid_trips <-
  df %>% filter(state_origin != 'NA' & state_dest != 'NA')
df %>%
  filter(state_origin != 'NA') %>%
  count(state_origin)
```

```
##    state_origin      n
## 1             2      1
## 2             3      8
## 3             7      2
## 4             9 265437
## 5            10      2
## 6            11     46
## 7            12     16
## 8            13   3589
## 9            14     13
## 10           15 260328
## 11           16     15
## 12           17    231
## 13           19      1
## 14           20     22
## 15           21    150
## 16           22    103
## 17           23      2
## 18           24      6
## 19           25     13
## 20           29     31
## 21           30     27
## 22           31      6
```

3. List the different modes of transportation available. Show your results on descending order by the number of trips

```
mode_trans_types <- df %>% count(mode_trans)
mode_trans_types %>% arrange(desc(n))
```

```
##       mode_trans      n
## 1 Public_Transit 203107
## 2           Walk 161313
## 3            Car 153693
## 4           Bike  13481
```

4. five least common reasons for trips except NAs are pick_up_someone, health, errands, other, church.

```
reason_rank <- df %>% filter(reason != 'NA') %>% count(reason)
tail(arrange(reason_rank, desc(n)),5)
```

```
##             reason     n
## 6  pick_up_someone 24891
## 7           health  5893
## 8          errands  3546
## 9            other  3031
## 10          church  1811
```

5. What proportion of the trips occur during the week and what proportion during the weekend?

```
days_dist <- df %>% count(day)
#week proportion
329580/(329580+202014)
```

```
## [1] 0.6199844
```

```
#weekend proportion
202014/(329580+202014)
```

```
## [1] 0.3800156
```

```
days_dist_church <- df %>% filter(reason == 'church')
days_prop_church <- days_dist_church %>% count(day)
# week proportion
495/(495+1316)
```

```
## [1] 0.2733297
```

```
# weekend proportion
1-(495/(495+1316))
```

```
## [1] 0.7266703
```

```
days_dist_walk <- df %>% filter(mode_trans == 'Walk')
days_prop_walk <- days_dist_walk %>% count(day)
# week proportion
107913/(107913+53400)
```

```
## [1] 0.6689665
```

```r
# weekend proportion
53400/(107913+53400)
```

```
## [1] 0.3310335
```

my observation finds that a greater proportion of people go to churches during weekends, and a greater proportion of people choose walk as the mode of transit during week days.

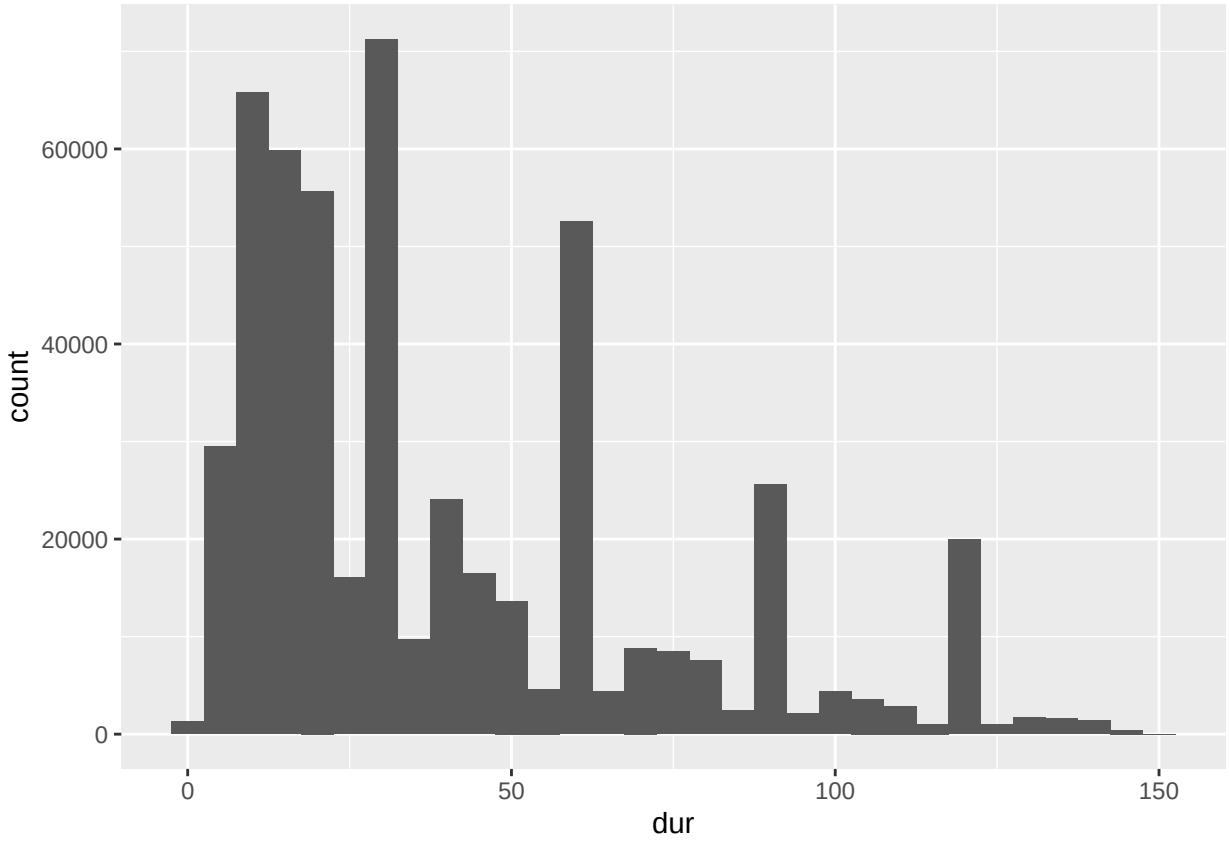
6. The average trip time is 43.68. We removed NAs in "dur" and values greater than 150 min to make the plot more informative.

```r
df_ave<-df %>%
  filter(!is.na(arr_hour)&
           !is.na(arr_min)&
           !is.na(dep_hour)&
           !is.na(dep_min)) %>%
  mutate(dur=(arr_hour-dep_hour)*60+arr_min-dep_min)


  df_ave<-df_ave %>%
    mutate(dur=ifelse(dur>0,dur,(24-dep_hour+arr_hour)*60+arr_min-dep_min))

  average_dur<-mean(df_ave$dur)

  df_ave%>%
    filter(dur<150)%>%
    ggplot()+
    geom_histogram(aes(x=dur),binwidth = 5)
```



2.4 More practical exercise

1. a.221131 b.131050 0.5926 c.83919 0.3795 0.6404

```
df %>%
  filter(state_origin==9 &
           state_dest==9 &
           !is.na(mun_origin) &
           !is.na(mun_dest) &
           !is.na(dto_origin) &
           !is.na(dto_dest)) %>%
  count()
```

```
##        n
## 1 221131
```

```
df%>%
  filter(state_origin==9&
           state_dest==9&
           !is.na(mun_origin)&
           !is.na(mun_dest)&
           !is.na(dto_origin)&
           !is.na(dto_dest)&
           (mun_origin-mun_dest)==0)%>%
  count()
```

```
##        n
## 1 131050
```

```
131050/221131
```

```
## [1] 0.5926351
```

```
df%>%
  filter(state_origin==9&
           state_dest==9&
           !is.na(mun_origin)&
           !is.na(mun_dest)&
           !is.na(dto_origin)&
           !is.na(dto_dest)&
           (dto_origin-dto_dest)==0)%>%
  count()
```

```
##       n
## 1 83919
```

```
83919/221131
```

```
## [1] 0.379499
```

```
83919/131050
```

```
## [1] 0.6403586
```

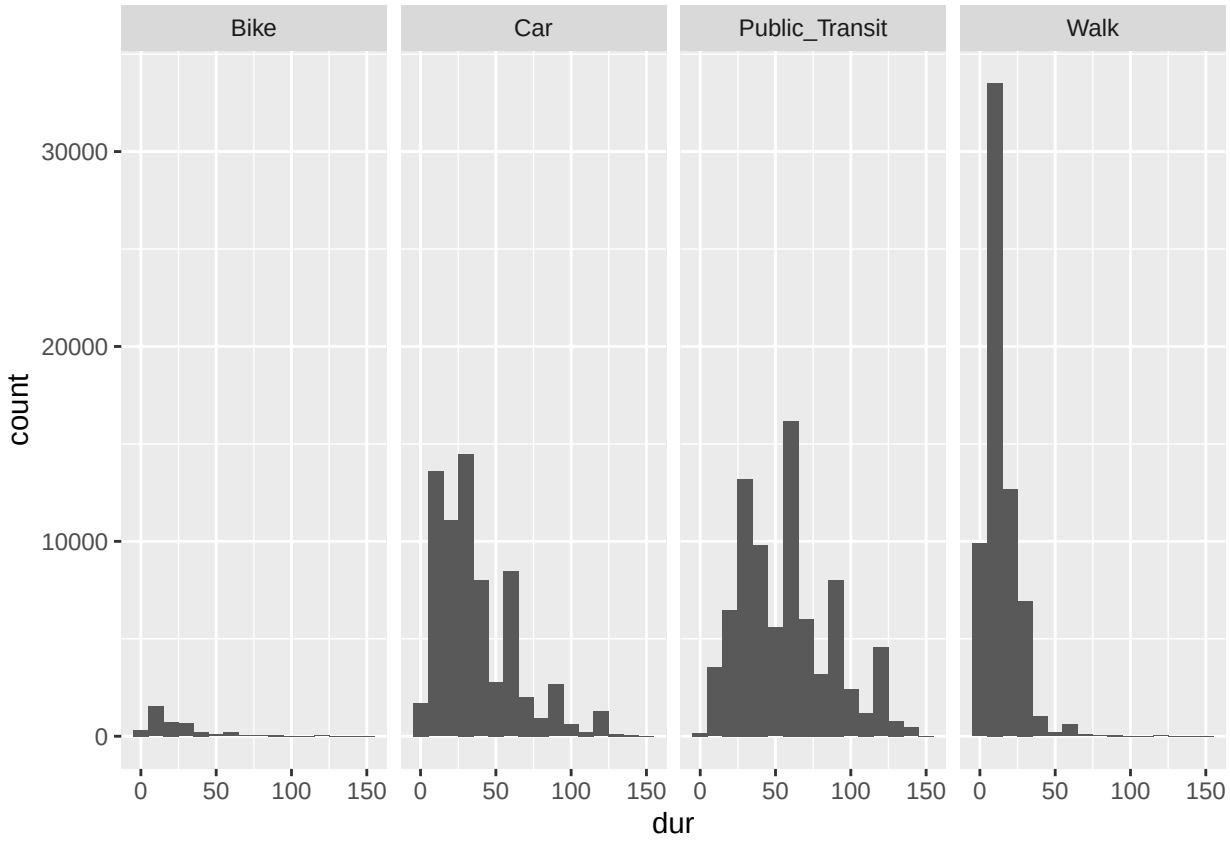
2. Bike trips are shorter than car and public transportation trips, but longer than walk trips. Since there are only a few trips last beyond 150 min and majority of them are public transportation trips, we filtered out those trips longer than that to make the plot more informative.

```
df_Mex<-df %>%
  filter(state_origin==9&
           state_dest==9&
           !is.na(mun_origin)&
           !is.na(mun_dest)&
           !is.na(dto_origin)&
           !is.na(dto_dest)) %>%
  mutate(dur=(arr_hour-dep_hour)*60+arr_min-dep_min)

df_Mex<-df_Mex%>%arrange(dur)


df_Mex<-df_Mex%>%
  mutate(dur=ifelse(dur>0,dur,(24-dep_hour+arr_hour)*60+arr_min-dep_min))

df_Mex %>%
  filter(dur<150) %>%
  ggplot()+
  geom_histogram(aes(x=dur),binwidth = 10)+
  facet_wrap(vars(mode_trans),ncol=4)
```




```
df_Mex %>% count(mode_trans)
```

```
##       mode_trans     n
## 1           Bike  3922
## 2            Car 68413
## 3 Public_Transit 83637
## 4           Walk 65159
```

3. Walk is more common for trips inside districts; Walk is more common for trips inside municipalities, too. A few people choose biking as transpotation method for trips inside districts and municipalities.

```
df%>%
  filter(state_origin==9&
           state_dest==9&
           !is.na(mun_origin)&
           !is.na(mun_dest)&
           !is.na(dto_origin)&
           !is.na(dto_dest)&
           (dto_origin-dto_dest)==0)%>%
  count(mode_trans)
```

```
##       mode_trans     n
## 1           Bike  2058
## 2            Car 16111
```

```
## 3 Public_Transit 11376
## 4           Walk 54374
```

```
df%>%
  filter(state_origin==9&
           state_dest==9&
           !is.na(mun_origin)&
           !is.na(mun_dest)&
           !is.na(dto_origin)&
           !is.na(dto_dest)&
           (mun_origin-mun_dest)==0)%>%
  count(mode_trans)
```

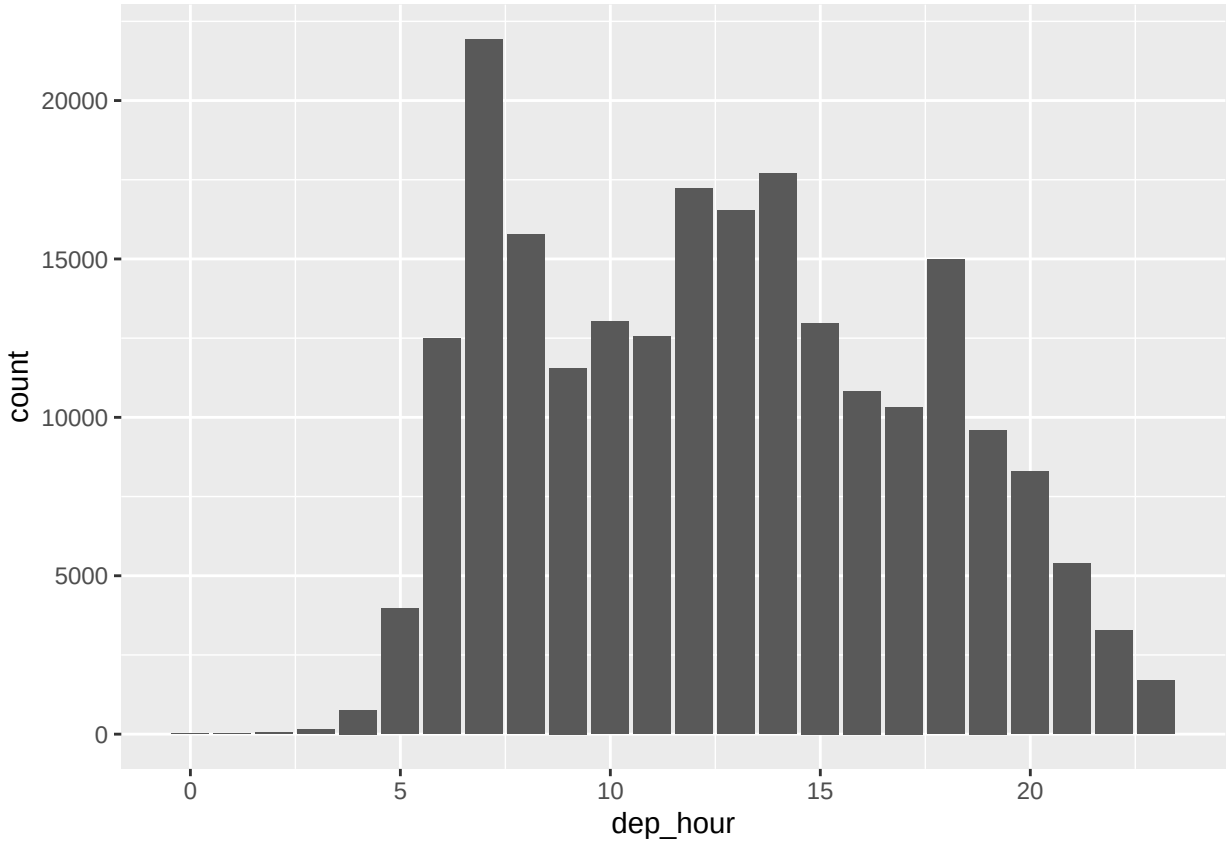
```
##       mode_trans     n
## 1           Bike  3011
## 2            Car 32461
## 3 Public_Transit 32883
## 4           Walk 62695
```

4. From the whole picture, there are several rush hours for trips, which are around 7, 13 and 18.

Compared between different transportation modes, we found that people who departure at early hours tend to use car and public transportation, those who departure at late hours tend to use bikes and walk. In the afternoon and night of a day, people like car and public transportation than biking or walking. Generally, they have similar distribution.

```
ggplot(df_Mex)+geom_bar(aes(x=dep_hour))
```

```
## Warning: Removed 64 rows containing non-finite values (stat_count).
```

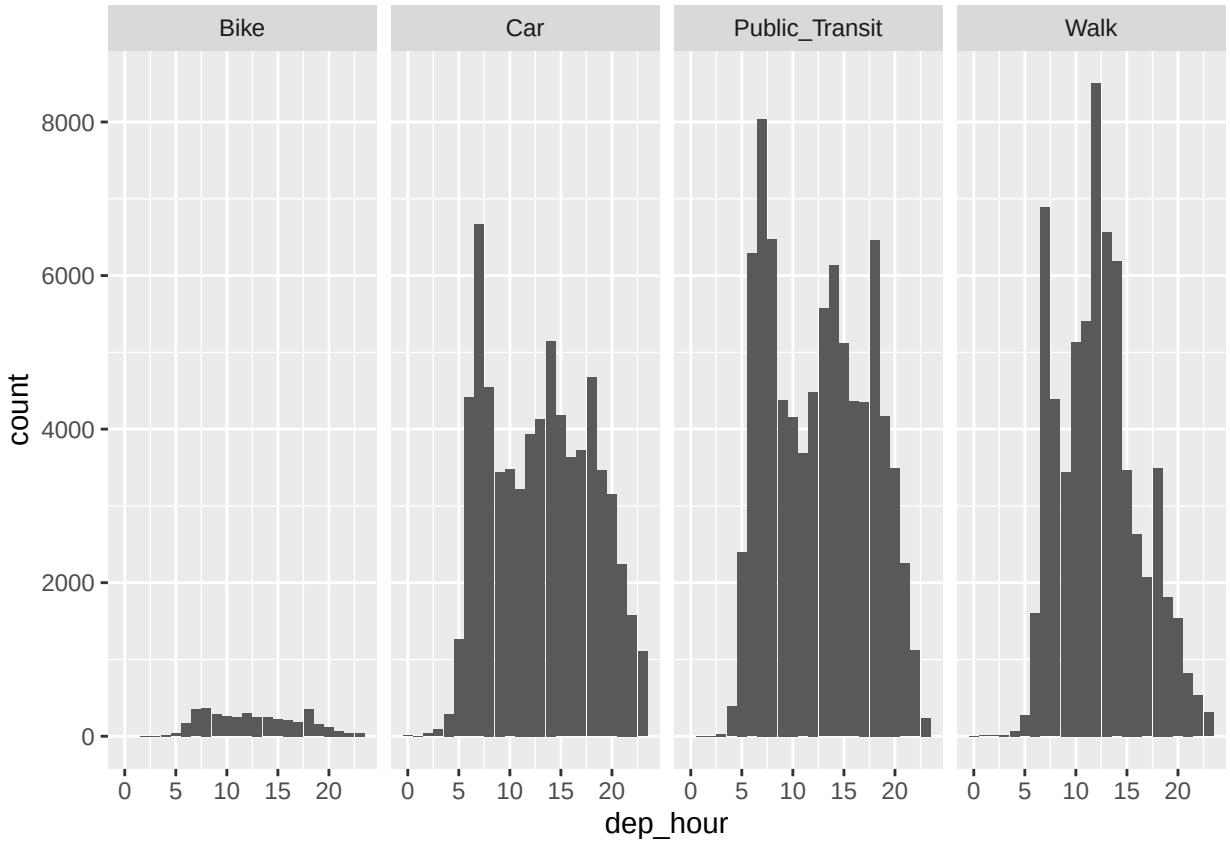



```
summary(df_Mex$dep_hour)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    9.00   13.00   12.69   16.00   23.00      64
```

```
ggplot(df_Mex)+geom_bar(aes(x=dep_hour))+
  facet_wrap(vars(mode_trans),ncol=4)
```

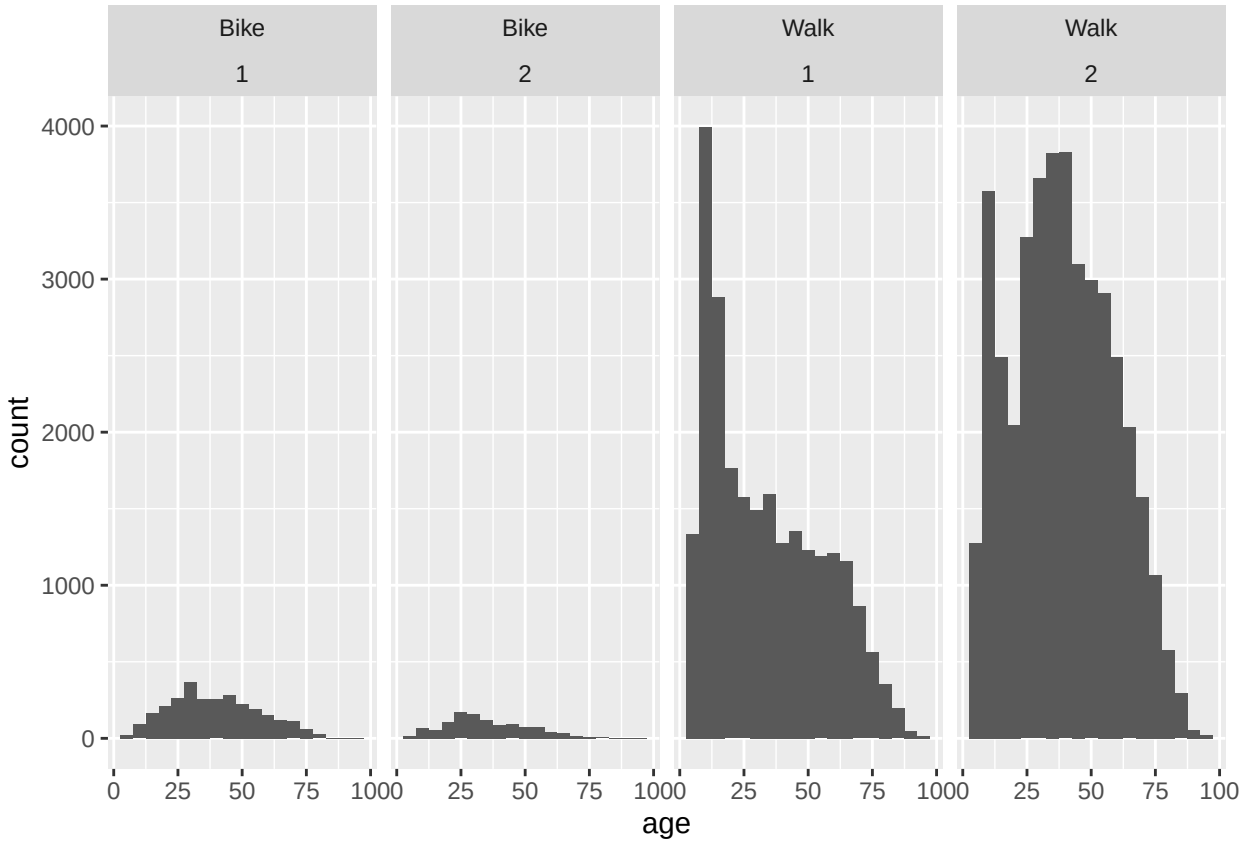
```
## Warning: Removed 64 rows containing non-finite values (stat_count).
```



5. No matter for female or male, there are much more people choose to walk than bike. As for difference in sex for biking, in general, there are more male choose to bike than female. And the age of female bikers concentrates on 25-35, while that for male spreads more widely, ranging from 25-50. As for difference in sex for Wlaking, in general, there are more female choose to walk than female. And the age of male walkers concentrates on 5-15, while that for female spreads more widely, ranging from 5-10 & 20-50.

```
df_bw<-df_Mex%>%
  filter(mode_trans=="Bike"|mode_trans=="Walk")

ggplot(df_bw)+
  geom_histogram(aes(x=age),binwidth=5)+
  facet_wrap(vars(mode_trans,sex),ncol=4)
```



6. From what have been discussed above, there are four main points about biking: First, a few people choose biking as transpotation method for trips inside districts and municipalities. Second, people who departure at early hours tend to use car and public transportation, those who departure at late hours tend to use bikes and walk. Third, no matter for female or male, there are much more people choose to walk than bike. And in general, there are more male choose to bike than female. Fourth, bike trips are shorter than car and public transportation trips, but longer than walk trips.

7. Reasons for not choosing biking:

a. lack of bike lanes: data indicates the distributions of bike lanes
b. being easily stolen: data indicates the possibilities for bikes' being stolen in a given period of time