Problem Set template

Earnest Salgado and Guillermo Trefogli

22/04/2020

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

Add your initials to indicate your agreement: **ES and GATW**

Add your collaborators: **ES and GATW**

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

1 Git merge conflicts (10 pts)

You and your partner will share a github remote respository and make changes and update the same files in your local repositories. This can be messy, but git has handy tools to deal with potential conflicts between your versions.

We are suggesting the following workflow to minimize conflicts. Divide the work into manageable chunks we refer to as "issues".

i. Branch: For each chunk, you will work on a branch. Push your changes to github often. We repeat. Add and commit code to your branch early and often!
ii. Pull request: When you have a stable section of code, make a pull request.
iii. Review: Ask your partner to review your code (this should be a high priority step!)
iv. Merge: Merge your code into master, dealing with any merge conflicts immediately.

Repeat with the next issue of your assignment.

Now, to practice, you will create your first merge conflict.

Prelude

i. Play paper, scissors, rock to determine who goes first. Call that person Partner 1.
ii. Both partners, clone the repository. On the master branch, Partner 1 rename the file pset_template to applied_ps_1.Rmd, delete the pdf file, commit and push to github. Partner 2 pull the changes. Now you both have the properly named file on your computers.

Begin merge conflict practice i. a. Partner 1, Start a branch called merge_conflict_practice_1. In applied_ps_1.Rmd replace "Fernanda Sobrino" with your name. Push your branch to github. (They call this "Publish" in github desktop). b. Partner 2, Start a branch called merge_conflict_practice_2. In applied_ps_1.Rmd replace "Fernanda Sobrino" with your name. ii. Partner 1 screen share and make a pull request. iii. Partner 2 screen share on github review the pull request. Accept your partners changes and merge the branch into master. Hooray! This is your first successful pull request! iv. Partner 2 make a pull request. v. Partner 1 scren share. On github review the pull request. There should be a merge conflict because you both changed the same line of the file. Adjust the file and then merge.

Switch roles and force a merge conflict with a different section of applied_ps_1.Rmd (e.g. the date or titles).

1. Succinctly explain, why did you have a merge conflict? We had a merge conflict caused by the competing changes made on the same line when we both replaced "Fernanda Sobrino" with our names. We were making different changes to the same line and same file on our different branches.

2 Exploratory analysis

2.1 Download the data (5 points)

1. The data you need is already in the repository. You should see a file name trips_mcma.csv. Read it into R using (make sure you are in the right directory to do this: read.csv('path/file.csv').

```
# ggplot(mpg, aes(manufacturer)) + geom_bar()
trips_mcma <- read.csv('trips_mcma.csv')
```

2. R has six description methods: print, head, str, glimpse, View, summary. Apply them to trips_mcma.csv. Dot not show them in your .pdf but do write the code necessary to get them and inspect them in your computer.

```
print(trips_mcma)
```

```
head(trips_mcma)
```

```
str(trips_mcma)
```

```
glimpse(trips_mcma)
```

```
view(trips_mcma)
```

```
summary(trips_mcma)
```

1. Are any of the methods redundant, in the sense that you don't learn anything about the data from these commands that you didn't already know from the prior methods? Make a list of the non-redundant methods (giving preference to the commands with prettier output)

Head, str, print, and glimpse are redundant after applying view to the data. Each of those commands either provide the same or a lesser amount of the information seen in view.

The non-redundant methods would be summary, and technically the view command we mentioned above.

2. Of the non-redundant methods, write a note (max two lines per command) that will quickly help someone (perhaps future you!) recall how each command is useful.

view is useful to invisibly view the argument and to see a full view of total observations and variables, variable types (integer, character, etc.) and data.

summary is useful to observe statistical and r characteristics for each variable such as min, 1st Qtr, median, mean, 3rd Qtr, max, NA's, length, class, and mode.

2.2 Let's see what's inside this data set: (35 points)

1. How many variables and how many rows does this data has? What does each row represent? Which variable is the unique identifier for each trip?

There are 20 variables and 531,594 rows in this data. Each row represents one person's individual trip. The id_trip variable is the unique identifier for each trip.

2. From how many different states are trips originating from? How many valid trips do we have? For now, define a valid trip as: a trips for which we have both origin and destination state.

There are at least 22 different states that trips are originating from. There are 1545 trips whose origin state are missed in the data.

```
trips_mcma %>%
  count(state_origin) %>%
  head(5)
```

```
##   state_origin      n
## 1            2      1
## 2            3      8
## 3            7      2
## 4            9 265437
## 5           10      2
```

There are 528,134 valid trips counted in our data after removing NA values

```
Non_NA_trips_mcma <- filter(trips_mcma, !is.na(state_origin), !is.na(state_dest))
nrow(Non_NA_trips_mcma)
```

```
## [1] 528134
```

3. Which different mode of transportation do we have? Show your result in descending order by the number of trips.

```
  trips_mcma %>%
  group_by(mode_trans) %>%
  summarise(counting_mode=n()) %>%
  arrange(desc(counting_mode))
```

```
## # A tibble: 4 x 2
##   mode_trans     counting_mode
##   <chr>                  <int>
## 1 Public_Transit        203107
## 2 Walk                  161313
## 3 Car                   153693
## 4 Bike                   13481
```

4. Which are the 5 least common reasons for trips? (ignoring NAs)

```
reason_trips_mcma <- trips_mcma %>%
  filter(!is.na(reason))

reason_trips_mcma %>%
  group_by(reason) %>%
  summarise(n=n()) %>%
  arrange(n) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   reason              n
##   <chr>           <int>
## 1 church           1811
## 2 other            3031
## 3 errands          3546
## 4 health           5893
## 5 pick_up_someone 24891
```

5. What proportion of the trips occur during the week and what proportion during the weekend? Find at least two different variables you can use to explore the differences between weekday and weekend trips. Use these variables to show how week vs weekend trips differ. Write a short paragraph about your findings.

For this question we are including all the values of 'state_origin' and 'state_dest', since it is not requested to only considering valid values. Then:

```
trips_mcma %>%
  group_by(day) %>%
  summarise(day_trips=n()) %>%
  mutate(total_day_trips = sum(day_trips),
         prop_trips = round(day_trips/total_day_trips, 2)*100)
```

```
## # A tibble: 2 x 4
##   day     day_trips total_day_trips prop_trips
##   <chr>       <int>           <int>      <dbl>
## 1 week       329580          531594         62
## 2 weekend    202014          531594         38
```

Week days have a total number of trips of 329580, which is the 62% of the total trips under analysis, whilst weekend days have 202014 (38%).

We will examine the differences of week and weekend trips in terms of the mode of transportation and reason for travel variables, as follows:

```
differences <- trips_mcma %>%
  group_by(day) %>%
  pivot_wider(names_from = day, values_from = mode_trans)

differences %>%
  group_by(week) %>%
  filter(!any(is.na(week))) %>%
  summarise(trip_modes_week=n()) %>%
  mutate(prop_trips_week =
           round(trip_modes_week/sum(trip_modes_week)*100, 2)) %>%
  arrange(desc(trip_modes_week))
```

```
## # A tibble: 4 x 3
##   week           trip_modes_week prop_trips_week
##   <chr>                    <int>           <dbl>
## 1 Public_Transit          128718           39.1
## 2 Walk                    107913           32.7
## 3 Car                      84913           25.8
## 4 Bike                      8036            2.44
```

```
differences %>%
  group_by(weekend) %>%
  filter(!any(is.na(weekend))) %>%
  summarise(trip_modes_weekend=n()) %>%
  mutate(prop_trips_weekend =
           round(trip_modes_weekend/sum(trip_modes_weekend)*100, 2)) %>%
  arrange(desc(trip_modes_weekend))
```

```
## # A tibble: 4 x 3
##   weekend        trip_modes_weekend prop_trips_weekend
##   <chr>                       <int>              <dbl>
## 1 Public_Transit              74389               36.8
## 2 Car                         68780               34.0
## 3 Walk                        53400               26.4
## 4 Bike                         5445                2.7
```

In terms of mode of transportation, we can see that in both of them public transportation is the most important one. In week days, the second most important mode of transportation is walking and the third one is traveling by car. This order changes during weekend days, in which traveling by car become the second most important and walking becomes the third one. Finally, in both type of days biking is the least choice.

```
travel_time <- trips_mcma %>%
  group_by(day) %>%
  pivot_wider(names_from = day, values_from = reason)

travel_time %>%
  group_by(week) %>%
  filter(!any(is.na(week))) %>%
  summarise(trip_reason_week=n()) %>%
  mutate(prop_trips_week =
           round(trip_reason_week/sum(trip_reason_week)*100, 2)) %>%
  arrange(desc(trip_reason_week))
```

```
## # A tibble: 10 x 3
##    week            trip_reason_week prop_trips_week
##    <chr>                      <int>           <dbl>
##  1 home                      156029           47.4
##  2 work                       72119           21.9
##  3 school                     39379           12.0
##  4 pick_up_someone            21791            6.61
##  5 shopping                   21598            6.56
##  6 leisure                     9331            2.83
##  7 health                      4224            1.28
##  8 errands                     2662            0.81
##  9 other                       1794            0.54
## 10 church                       495            0.15
```

```
travel_time %>%
  group_by(weekend) %>%
  filter(!any(is.na(weekend))) %>%
  summarise(trip_reason_weekend=n()) %>%
  mutate(prop_trips_weekend =
           round(trip_reason_weekend/sum(trip_reason_weekend)*100, 2)) %>%
  arrange(desc(trip_reason_weekend))
```

```
## # A tibble: 10 x 3
##    weekend         trip_reason_weekend prop_trips_weekend
##    <chr>                         <int>              <dbl>
##  1 home                          92909              46.0
##  2 work                          37068              18.4
##  3 leisure                       31273              15.5
##  4 shopping                      29077              14.4
##  5 school                         3345               1.66
##  6 pick_up_someone                3100               1.54
##  7 health                         1669               0.83
##  8 church                         1316               0.65
##  9 other                          1237               0.61
## 10 errands                         884               0.44
```

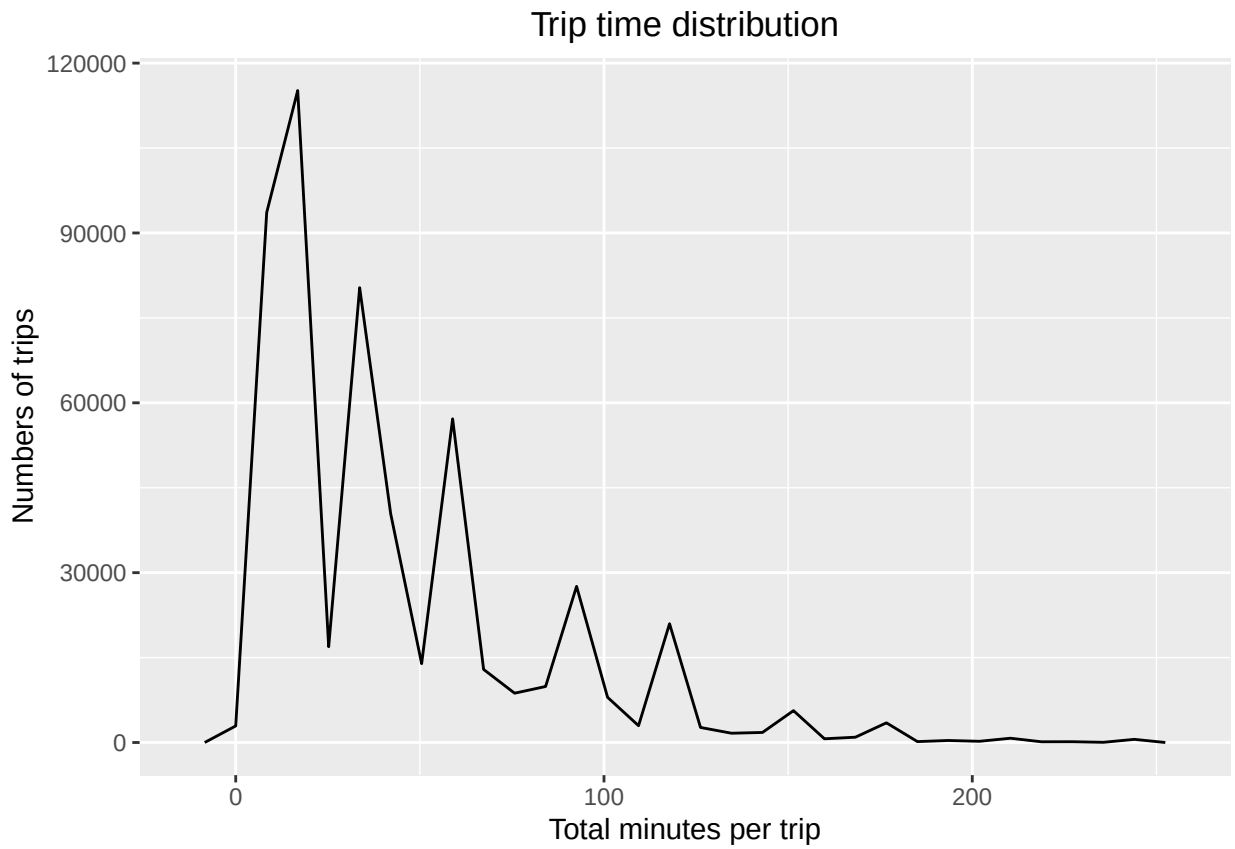
In terms of reasons for trips, we can see that for week days the three most frequent ones are going home, going to the workplace, and going to the school. In the case of weekend days, going home and going to the workplace remains the first and second most important, but leisure becomes the third one.

6. How long does the average trip takes (Save this variable for later)? Plot the distribution of travel time. Can you restrict the sample so this plot is more informative? Do you notice any weird pattern in travel times? Why do you think this is happening?

```
travel_time <- trips_mcma %>%
  mutate(trip_time =(arr_hour*60 + arr_min) - (dep_hour*60 + dep_min))

travel_time %>%
  filter(!is.na(trip_time)) %>%
  summarise(mean(trip_time))
```

```
##   mean(trip_time)
## 1         43.1956
```

```
travel_time %>%
  filter(trip_time > 0 & trip_time < 250) %>%
  ggplot(aes(x=trip_time)) +
  geom_freqpoly() +
  labs(x="Total minutes per trip", y="Numbers of trips") +
  ggtitle("Trip time distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```



The average trip is 43 minutes. There is a decreasing patterns in terms of duration of trip time and number of trips. In other words, the longer the trip, the more uncommon it is in this dataset.

2.3 More practical exercise (50 points)

The Mexico City secretary of transportation asks for your help to analyze these data. They are interested in knowing when and where people are deciding to bike. Eventually they would like to implement some policies to incentivize people to bike more but first they need to understand the current mobility patterns.

1. Restrict the sample to Mexico City (state code 9) and to trips with known municipality and district. How many trips are there inside Mexico City? How many trips are inside the same municipality? What proportion of all the Mexico City trips do they represent? How many trips are there inside the same district? What proportion of all the Mexico City trips do they represent? What proportion of the trips inside municipalities do they represent?

For this analysis, we will keep with the trips that depart and arrive inside Mexico City. On the other hand, we sill keep with valid values for municipalities and districts inside the city.

```
df_trips <- filter(trips_mcma, state_origin==9 &  state_dest==9 & !is.na(mun_origin)
            & !is.na(dto_origin) & !is.na(mun_dest) & !is.na(dto_dest))
```

There are 221,131 trips inside Mexico City:

```
df_trips %>% count()
```

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

There are 131,050 trips inside the same municipality:

```
in_muni <- mutate(df_trips, inside_mun = ifelse(mun_origin==mun_dest, 1, 0))

total_in_muni <- filter(in_muni, inside_mun == 1) %>%
  count(inside_mun)
print(total_in_muni)
```

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

The total of trips inside municipality is the 59.2 % of total trips of Mexico City:

```
total_df <- in_muni %>% nrow()

weight_in_muni <- (total_in_muni/total_df)*100
print(weight_in_muni)
```

```
##     inside_mun        n
## 1 0.0004522206 59.26351
```

There are 83,919 trips inside the same municipality:

```r
in_dto <- mutate(df_trips, inside_dto = ifelse(dto_origin==dto_dest, 1, 0))

total_in_dto <- filter(in_dto, inside_dto == 1) %>%
  count(inside_dto)
print(total_in_dto)
```

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

The total of trips inside municipality is the 37.9 % of total trips of Mexico City:

```r
weight_in_dto <- (total_in_dto/total_df)*100
print(weight_in_dto)
```

```
##     inside_dto       n
## 1 0.0004522206 37.9499
```

The 64.0 % of trips inside the same municipality is inside the same district:

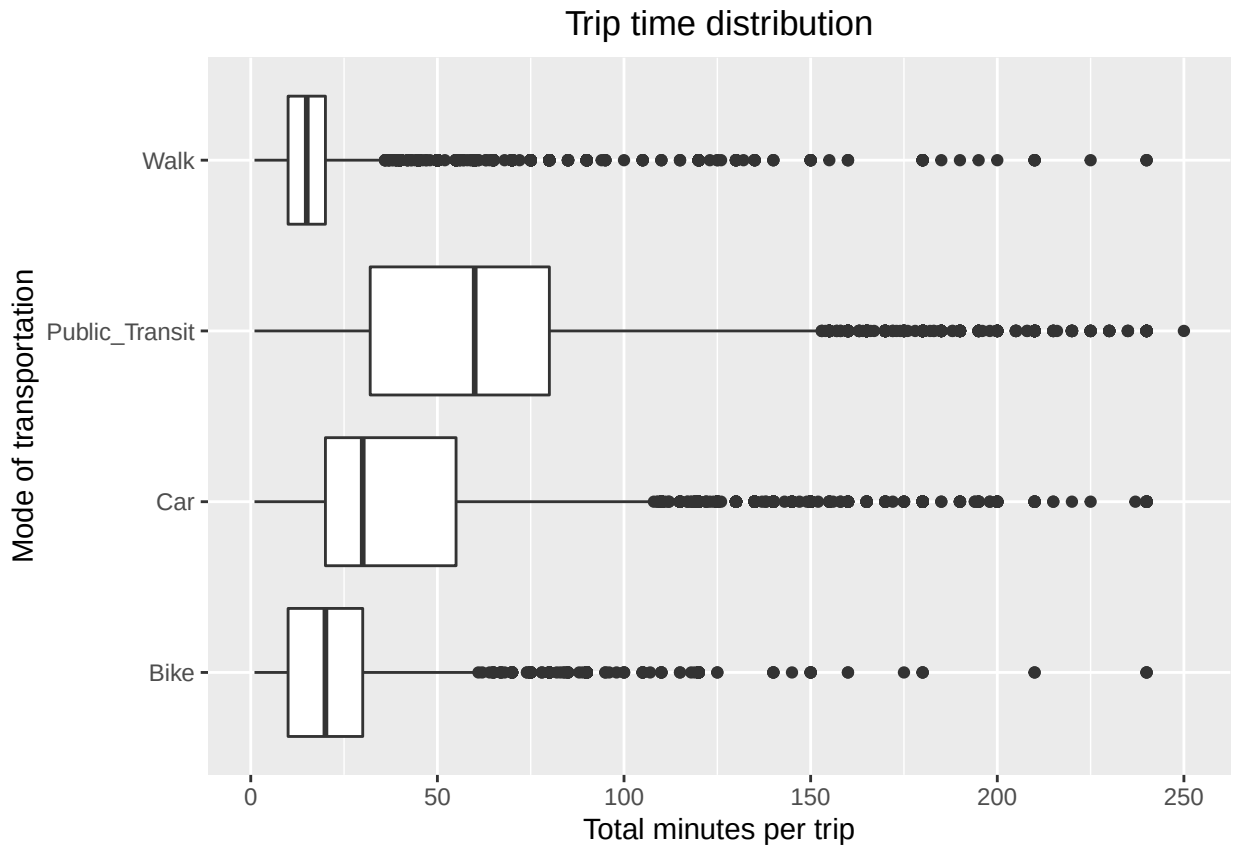
```r
weight_dto_muni <- (total_in_dto/total_in_muni)*100
print(weight_dto_muni)
```

```
##   inside_dto        n
## 1        100 64.03586
```

2. Are biking trips shorter/longer compared to other modes of transportation? Show this using a plot, remember to restrict your sample of travel times to see the distribution better, justify your threshold.

Biking trips are in average shorter than car and public transportation, but longer than by walking. To see it in a plot, we will restrict by trip time > 0 hours and removing the outliers. Our threshold is 600 since there is only one value for public transportation which is above that number:

```r
df_trips_time <- df_trips %>%
  mutate(trip_time =(arr_hour*60 + arr_min) - (dep_hour*60 + dep_min))

df_trips_time_2 <- filter(df_trips_time, trip_time > 0 & trip_time < 600)

df_trips_time_2 %>%
ggplot(aes(x = mode_trans, y = trip_time)) +
  geom_boxplot() +
  coord_flip() +
  labs(x="Mode of transportation", y="Total minutes per trip") +
  ggtitle("Trip time distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```



3. Which mode of transportation is more common for trips inside districts? How about trips inside the same municipality? What can you tell us about the biking trips?

Walking is the most common mode of transportation in both those trips inside the same municipality and inside the same district. On the other side, biking is the least mode of transportation in both groups. Respectively, biking just represents 2.3 % for trips inside the same municipality and 2.5 % for trips inside the same district.

```
df_time_type <- mutate(df_trips_time_2, inside_mun = ifelse(mun_origin==mun_dest, 1, 0))

type_in_muni <- filter(df_time_type, inside_mun==1)

type_in_muni %>%
  group_by(mode_trans) %>%
  summarise(muni_trips=n()) %>%
  mutate(prop_trips = round(muni_trips/sum(muni_trips), 3)*100) %>%
  arrange(desc(muni_trips))
```

```
## # A tibble: 4 x 3
##   mode_trans     muni_trips prop_trips
##   <chr>               <int>      <dbl>
## 1 Walk                62689       47.8
## 2 Public_Transit      32878       25.1
## 3 Car                 32441       24.8
## 4 Bike                 3011        2.3
```

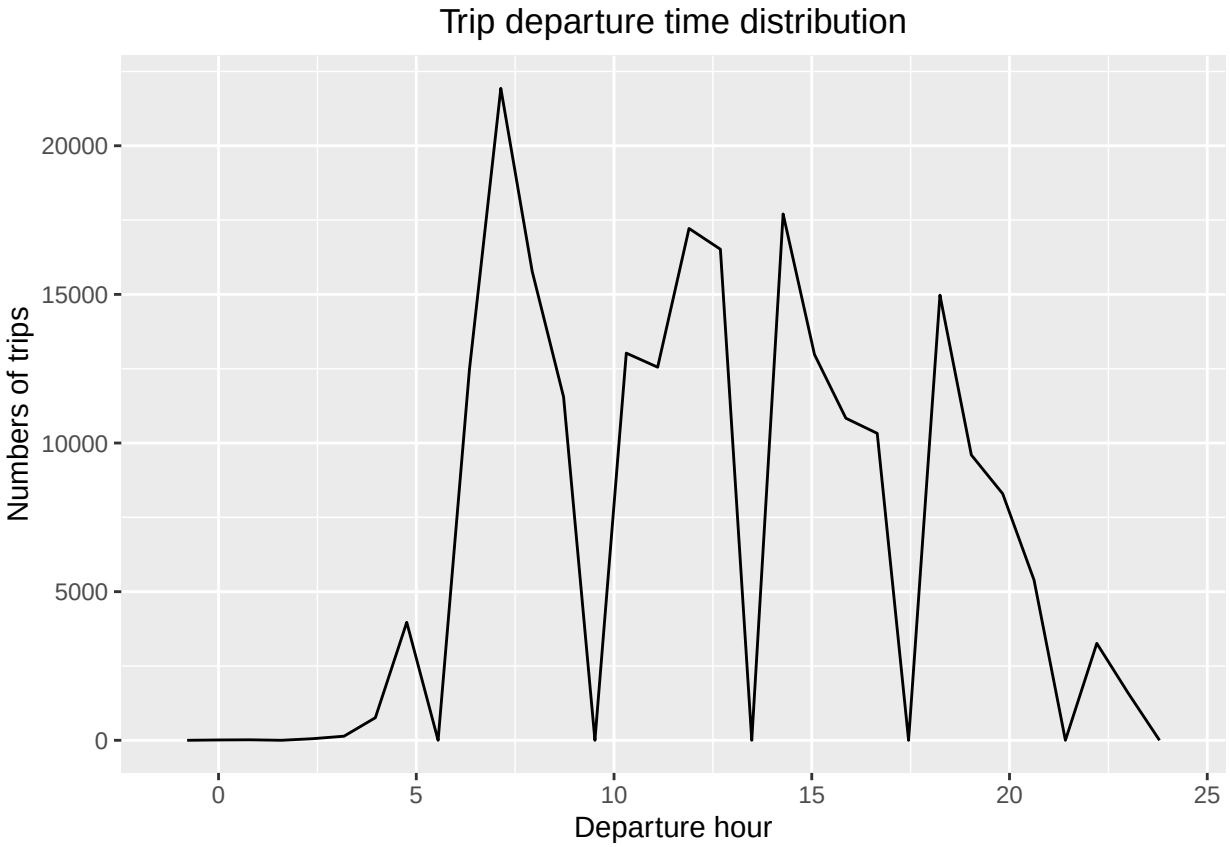
```r
df_time_type <- mutate(df_trips_time_2, inside_dto = ifelse(dto_origin==dto_dest, 1, 0))

type_in_dto <- filter(df_time_type, inside_dto==1)

type_in_dto %>%
  group_by(mode_trans) %>%
  summarise(dto_trips=n()) %>%
  mutate(prop_trips = round(dto_trips/sum(dto_trips), 2)*100) %>%
  arrange(desc(dto_trips))
```

```
## # A tibble: 4 x 3
##   mode_trans     dto_trips prop_trips
##   <chr>              <int>      <dbl>
## 1 Walk               54369        65
## 2 Car                16108        19
## 3 Public_Transit     11374        14.
## 4 Bike                2058         2
```

4. Show the distribution of when trips start, use dep_hour. Is this different by mode of transportation? Write a couple of sentences about your findings.

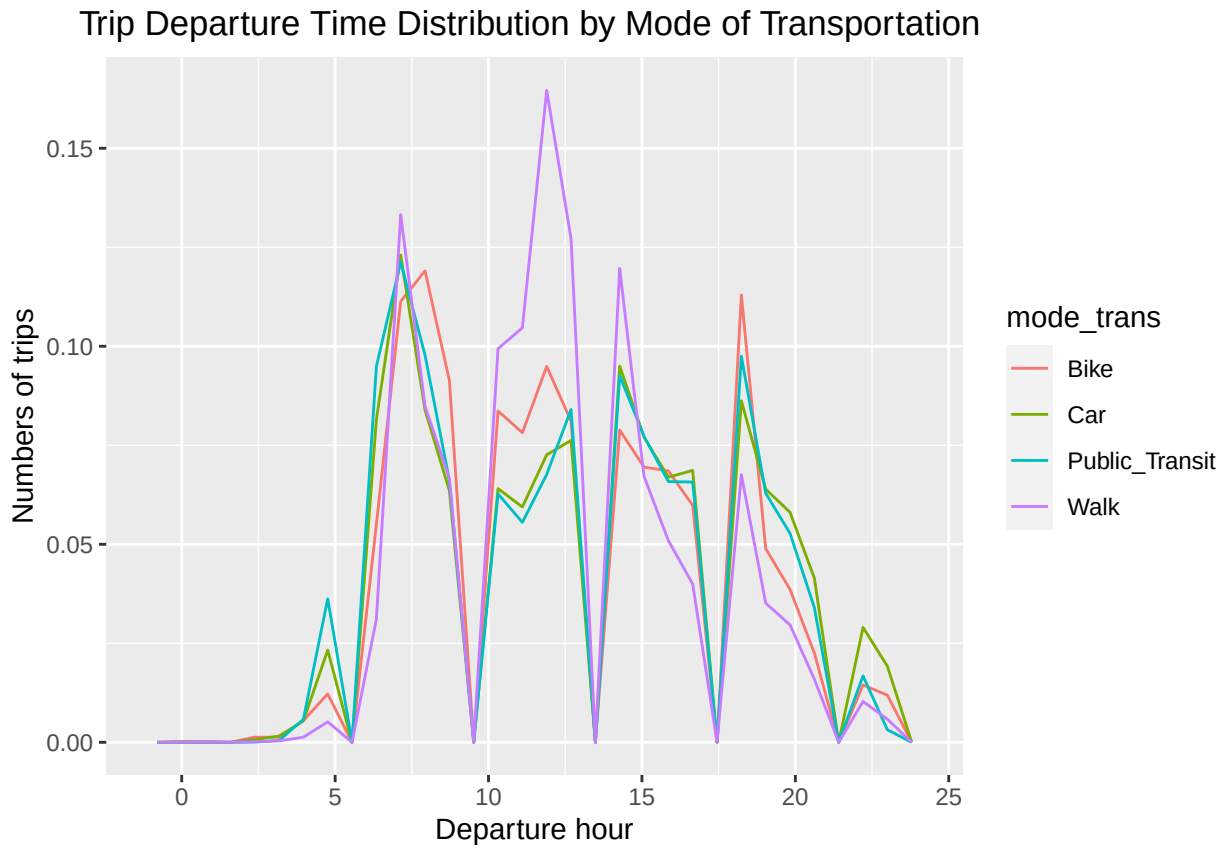
The following plot shows the frequency of trips in terms of hour departures. It can be notice that there is a peak during the day around 7 AM. Then, there are a couple more of peaks around 1 PM and roughly 3 PM. The lowest points are before 5am and after 9PM.

```r
df_time_type %>%
  ggplot(aes(x = dep_hour)) +
  geom_freqpoly() +
  labs(x="Departure hour", y="Numbers of trips") +
  ggtitle("Trip departure time distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```



In the following plot, we have standardized the frequency of trips for each mode of transportation, so that it can be seen the quantity according to departure hours in terms of the subtotal quantity of each mode of transportation:

```
df_time_type %>%
  ggplot(aes(x = dep_hour, y = ..density..)) +
  geom_freqpoly(aes(color = mode_trans)) +
  labs(x="Departure hour", y="Numbers of trips") +
  ggtitle("Trip Departure Time Distribution by Mode of Transportation ") +
  theme(plot.title = element_text(hjust = 0.5))
```

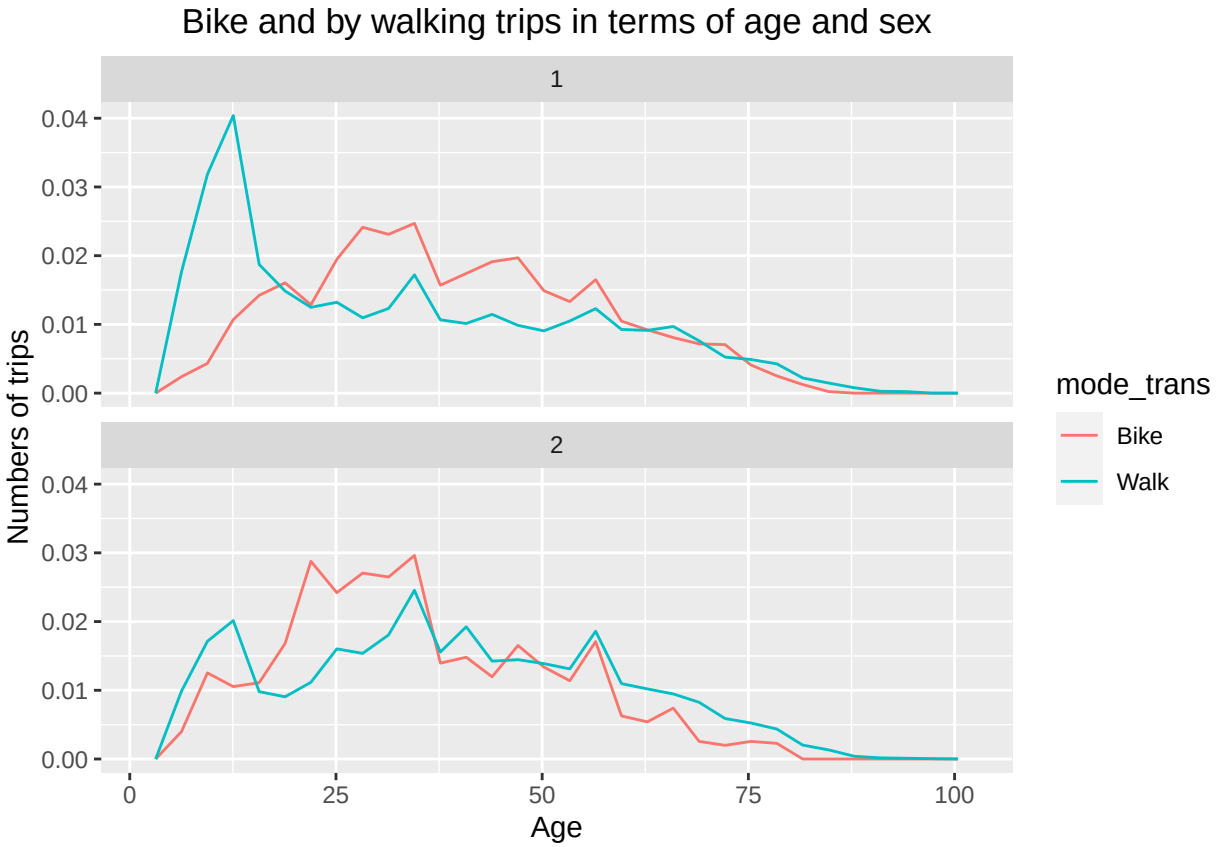


When split by mode of transportation, we can notice that the pattern of departure hour remains the same for all mode of transportation. Despite that, we can say that walking has a particular peak around noon, which makes its distribution of departure hours slightly different from the rest.

5. Explore the relationship between age, sex and the decision to walk or bike. The code for sex is 1 for male and 2 for female.

```
sex = c('male', 'female')

df_time_type %>%
  filter(mode_trans == "Walk" | mode_trans == "Bike") %>%
  ggplot(aes(x = age, y = ..density..)) +
  geom_freqpoly(aes(color = mode_trans)) +
  facet_wrap(vars(sex), nrow = 2, ) +
labs(x="Age", y="Numbers of trips") +
  ggtitle("Bike and by walking trips in terms of age and sex") +
  theme(plot.title = element_text(hjust = 0.5))
```

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



We can notice that both of them, male and female, have similar patterns in terms of biking and walking through their ages, but for different age periods. One difference is that male prefer to walk more than female in the early part of life. Then, a second difference is that male prefer biking more than walking from their twenties to their sixties, approximately. On the other hand, female prefer biking from their teenage year to their thirties. Finally, female prefer walking in their elderly years while male have equal preferences in their elderly years.

6. Write a short paragraph of what you just learn about biking trips compare to other kind of trips (from your answers to the previous questions)

The first conclusion is that people in Mexico City are not use to bike as much as all the other mode of transportation. Biking represents less than 3 % of total trips in Mexico City. There is no difference in terms of the hour of the day in which people bike in comparison with other modes of transportation. In terms of sex, male and female have similar patterns in terms of biking, with slightly differences, as they age. Furthermore, the time duration of the bike trip is by far less than traveling by public transportation and cars, but more than by walking. Finally, we also notice people's preferences to bike in Mexico city are the same during the week and weekends, however people travel much more on weekdays. The reasons to take trips changes as well during weekends, when people travel more because of leisure.

7. There are several reasons why people might decide not to bike. Can you think about 2 of them and what extra data sets you might have to have if you want to prove your theories? (you don't need to download any data or do any coding or plots for this question)

First, it is not efficient for them to bike long distances. Then, it is important to see at the data of the distances people should travel in their trips. Second, it is not convenient for them to do their trips by biking due to the lack of quality roads to bike in the city (safe and well connected, for instance). Then, we have to see at the data of roads and any quality assessment of the biking roads in the city.