Table of Contents

**1. Introduction** ............................................................................................................................... 2

**2. Plans/Requirement Analysis** .................................................................................................... 2

**3. Database Design** ........................................................................................................................ 3

**4. Application Design** .................................................................................................................... 4

**5. Demos** ......................................................................................................................................... 5

**6. Conclusions and Future Work** ................................................................................................. 5

6.1. Conclusions .......................................................................................................................... 5

6.2. Limitations ........................................................................................................................... 5

6.3. Potential Improvements or Future Work .............................................................................. 5

What is the average duration of the bike trips and does it varies according to the cities.

Load the below 4 packages:

**library**(ggplot2)

**library**(data.table)

**library**(bit64)

**library**(sqldf)

**library**(gridExtra)

**Import the 2 data files into R**

my\_trip = fread("trip.csv")

my\_station = fread("station.csv")

**Join data on stations to trips**

Join\_trips = data.table(sqldf("SELECT ta.duration, ta.start\_date as date, ta.subscription\_type, ta.zip\_code, sa.id as st\_id,sa.city as city FROM trip ta LEFT JOIN station sa ON ta.start\_station\_id=sa.id ")

**Splitting the date variable into month, day and year variables:**

Join\_trips = Join\_trips[, dates := sapply(date, function(a) strsplit(a, " ")[[1]][1]) ]

Join\_trips = Join\_trips[, times := sapply(date, function(a) strsplit(a, " ")[[1]][2]) ]

Join\_trips = Join\_trips[, hours := sapply(times, function(a) strsplit(a, ":")[[1]][1])]

Join\_trips$hours = as.numeric(Join\_trips$hours)

Join\_trips = Join\_trips[, minutes := sapply(times, function(a) strsplit(a, ":")[[1]][2])]

Join\_trips$minutes = as.numeric(Join\_trips$minutes)

**# Split date variable into month, day and year variables**

Join\_trips = Join\_trips[, month := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][1]))]

Join\_trips = Join\_trips[, day := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][2]))]

Join\_trips = Join\_trips[, year := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][3]))]

**# Add a "day of week" variable**

Join\_trips = Join\_trips[, weekday := sapply(dates , function(x) weekdays(as.Date(x,"%m/%d/%Y"))) ]

**# Clean up**

Join\_trips = Join\_trips[, c("date", "times") := NULL]

Join\_trips = Join\_trips[, weekend := ifelse(weekday %in% c("Sunday", "Saturday"), 1, 0)]

**# Create a variable for season**

Join\_trips = Join\_trips[month %in% c(12,1,2), season:="Win"]

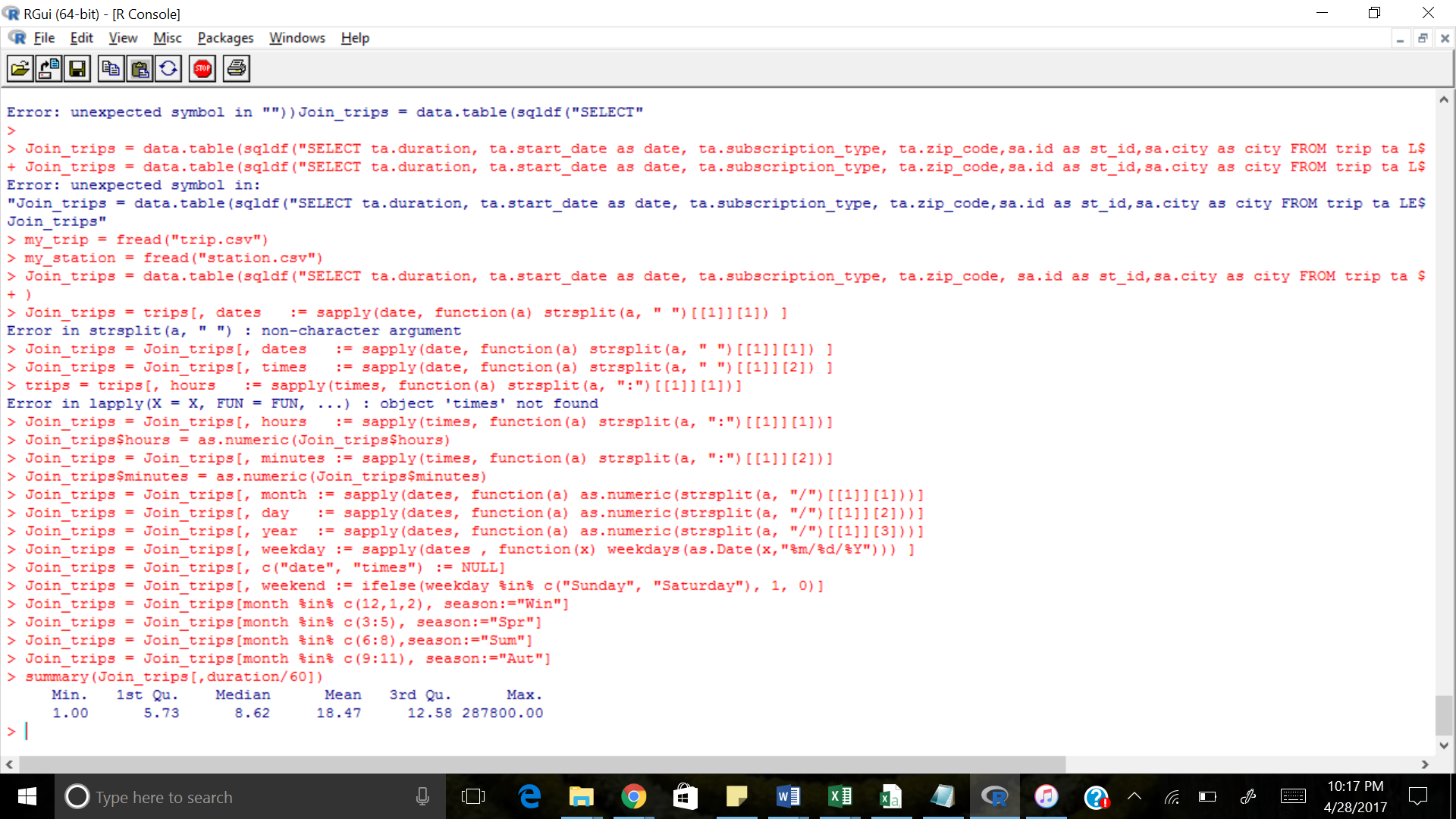
Join\_trips = Join\_trips[month %in% c(3:5), season:="Spr"]

Join\_trips = Join\_trips[month %in% c(6:8),season:="Sum"]

Join\_trips = Join\_trips[month %in% c(9:11), season:="Aut"]

**Mean trips duration in minutes by days**

summary(Join\_trips[,duration/60])



**# Trips less than 1 hour**

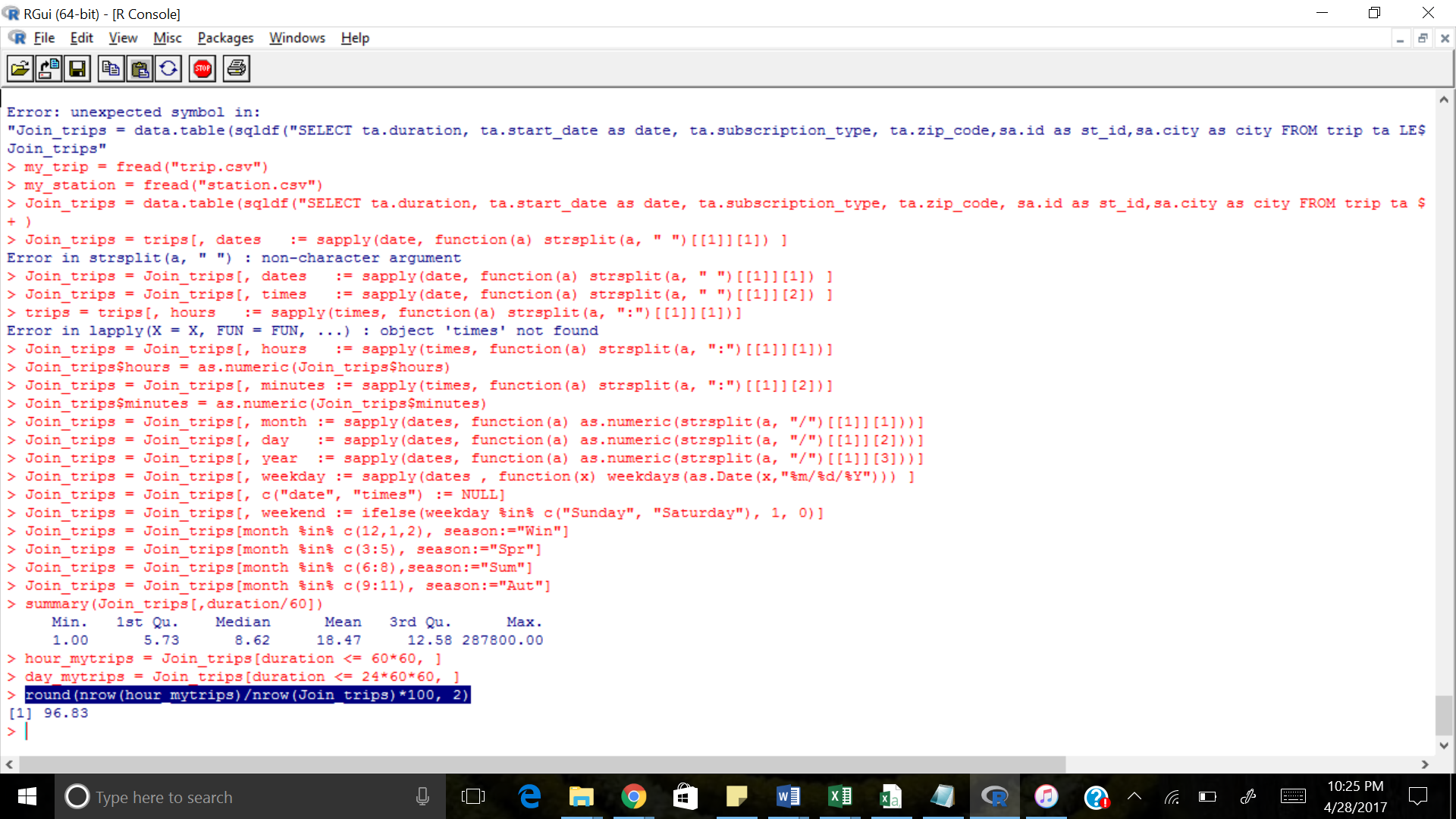
hour\_mytrips = Join\_trips[duration <= 60\*60, ]

**# Trips less than 1 day**

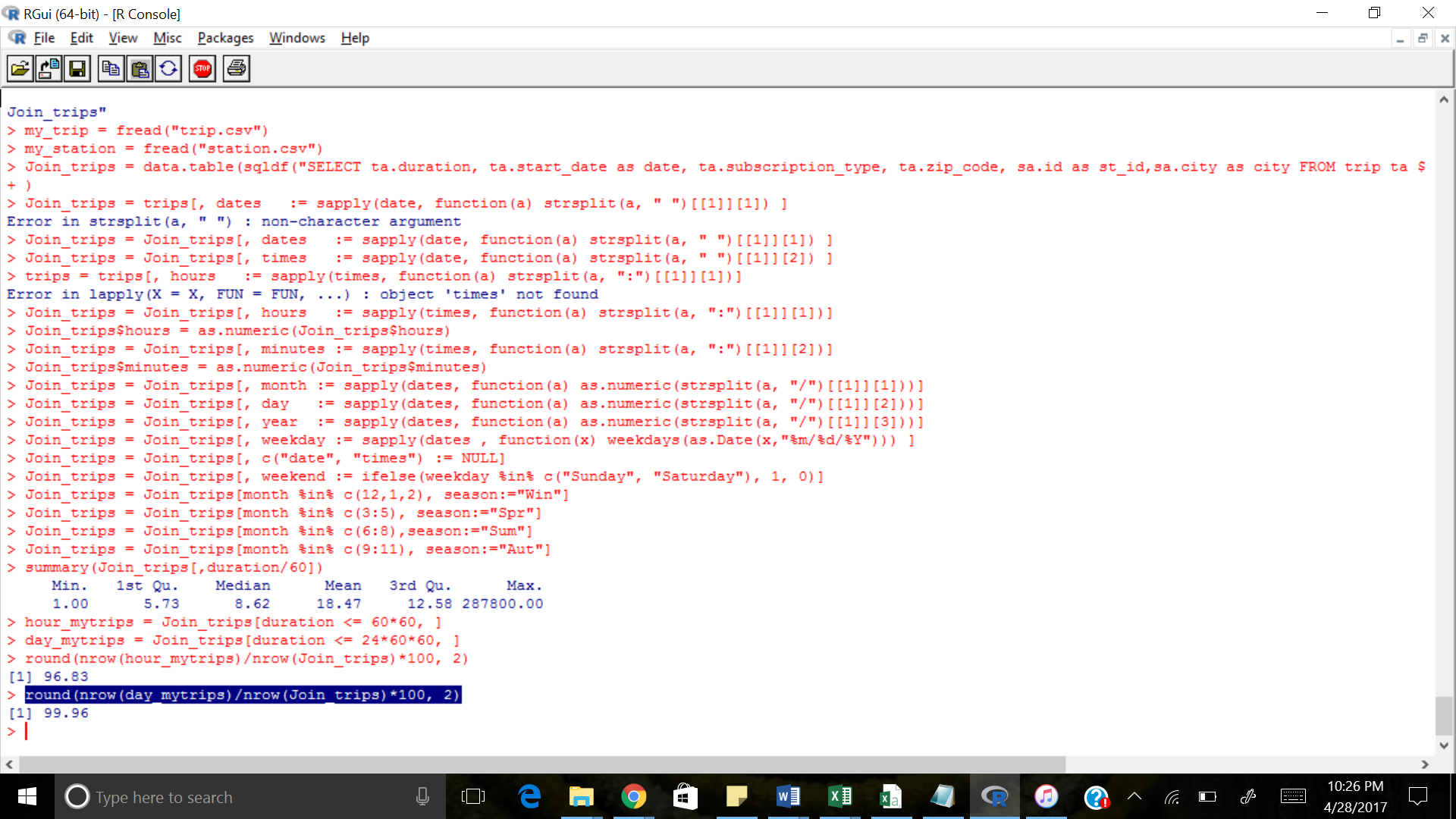
day\_mytrips = Join\_trips[duration <= 24\*60\*60, ]

**# Shares of the trips we are working with**

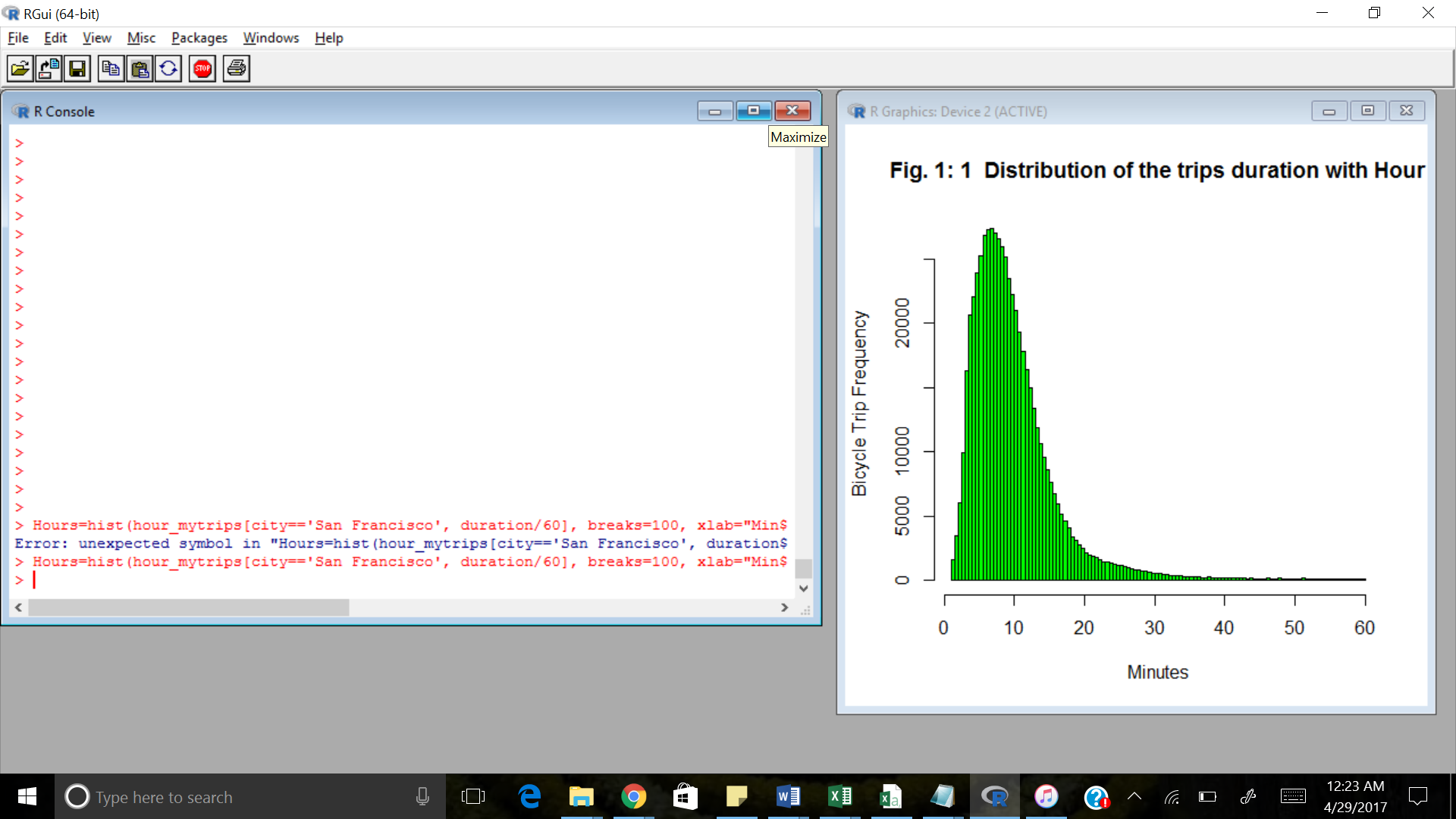
round(nrow(hour\_mytrips)/nrow(Join\_trips)\*100, 2)

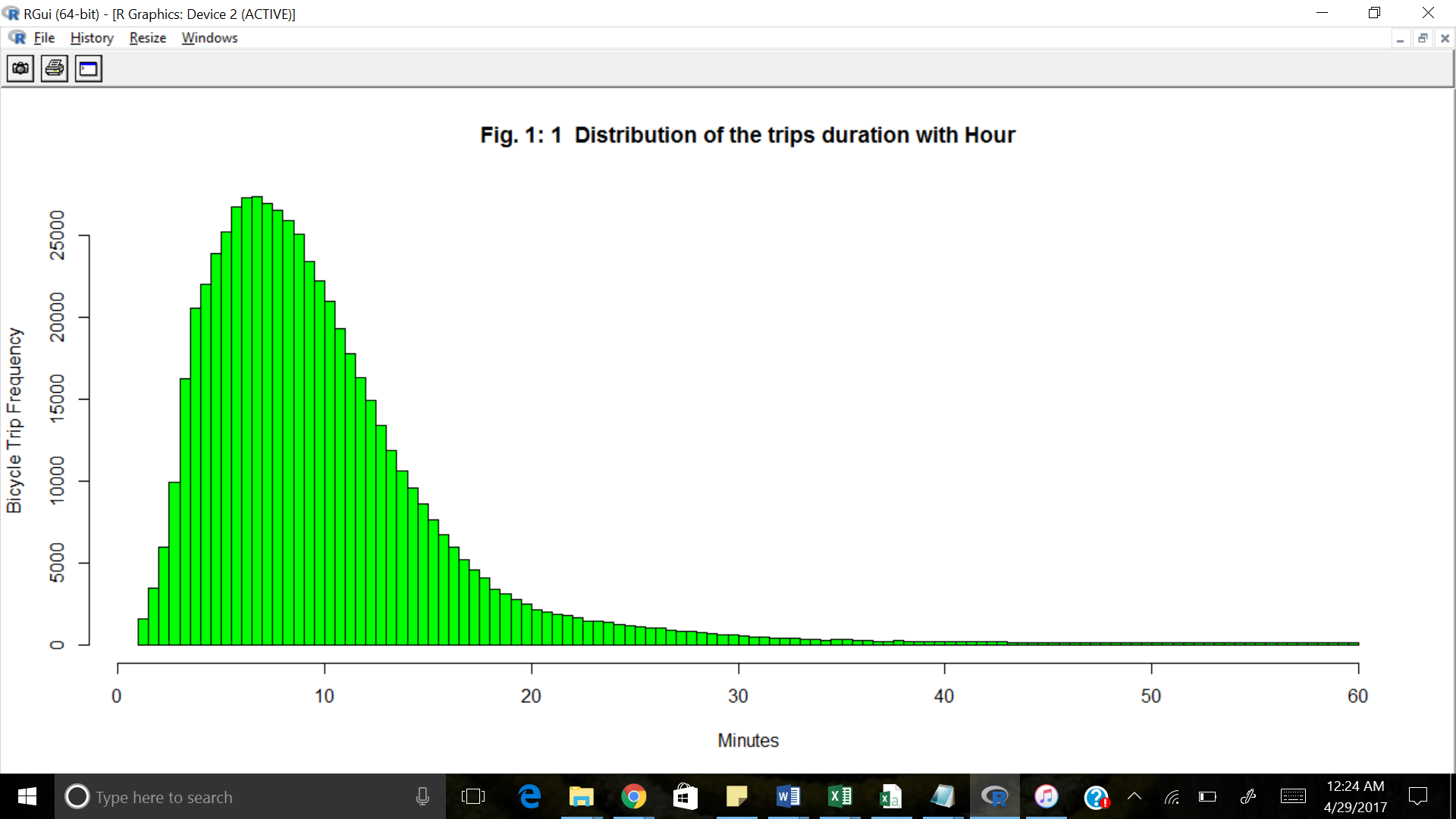


round(nrow(day\_mytrips)/nrow(Join\_trips)\*100, 2)

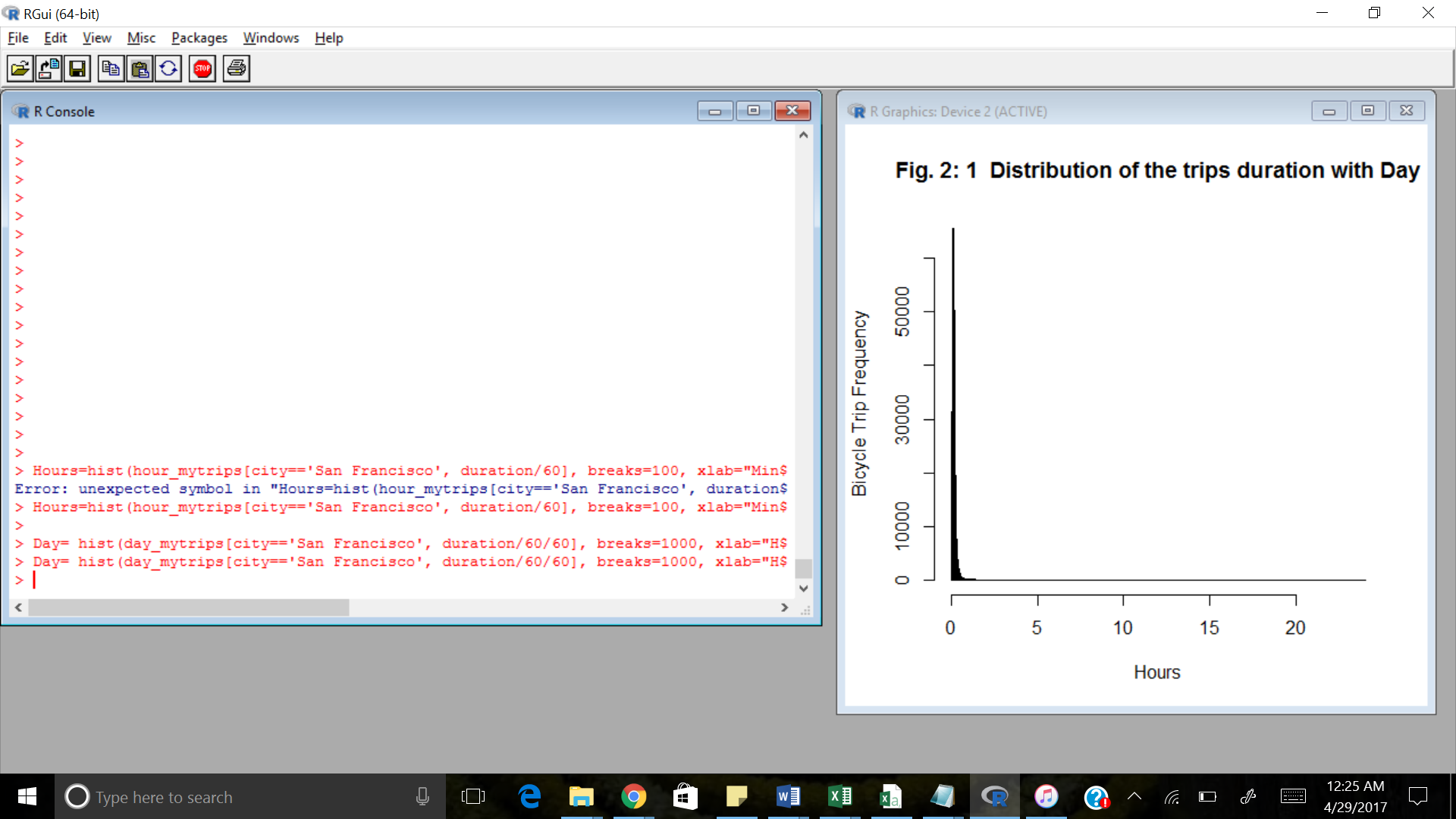


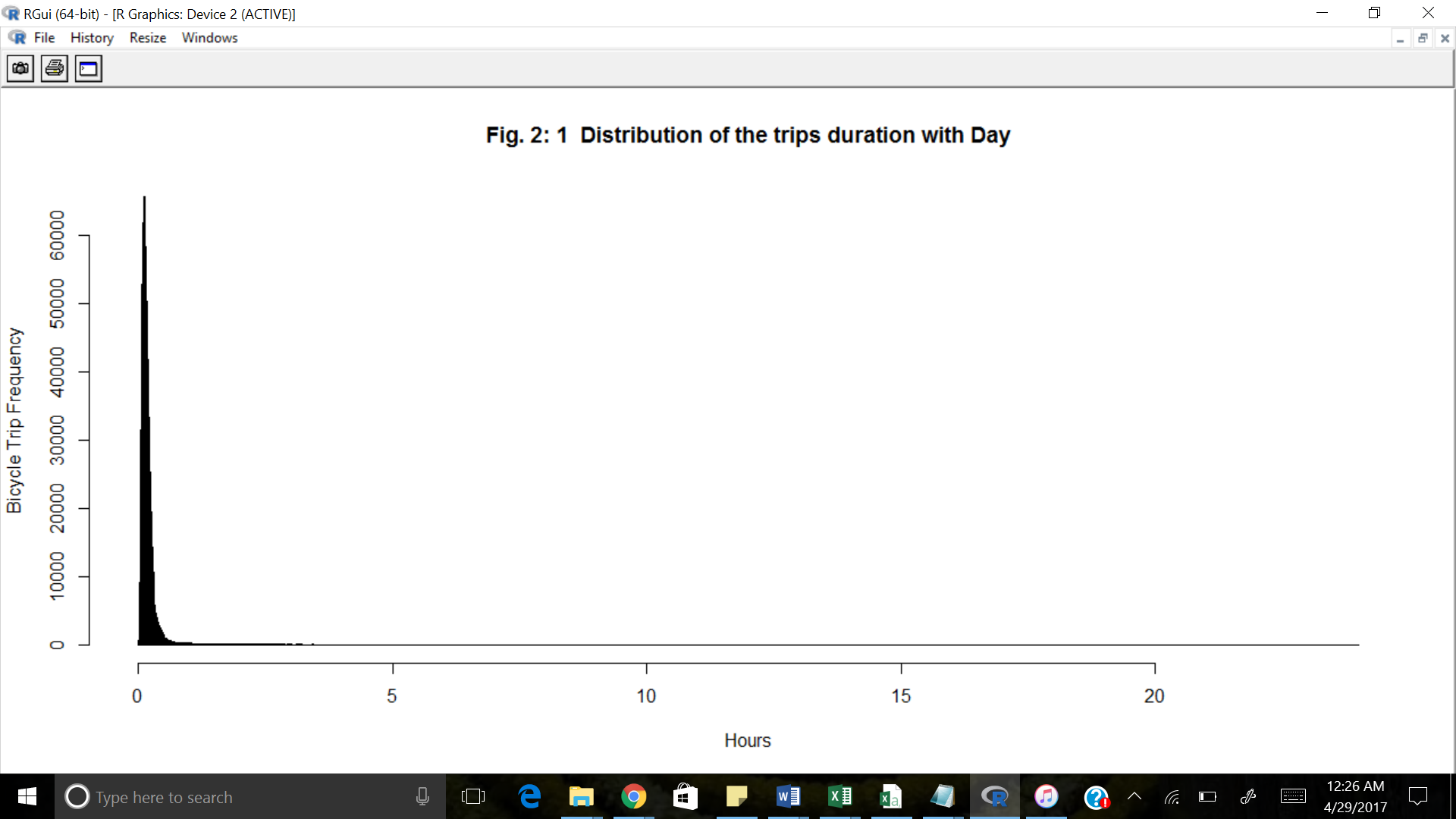
**# Check the histograms**

Hours=hist(hour\_mytrips[city=='San Francisco', duration/60], breaks=100, xlab="Minutes",ylab="Bicycle Trip Frequency", main="Fig. 1: 1 Distribution of the trips duration with Hour",col="green" )



Day= hist(day\_mytrips[city=='San Francisco', duration/60/60], breaks=1000, xlab="Hours",ylab="Bicycle Trip Frequency", main="Fig. 2: 1 Distribution of the trips duration with Day",col="red")





**Duration of trips in different cities on the workweek and weekend:**

**Average trip duration in minutes by weekdays/weekends and by city**

**day\_duration = hour\_mytrips[, mean(duration)/60, list(weekend, city)]**

**day\_duration = day\_duration[, weekend:=as.factor(weekend)]**

**day\_duration = day\_duration[, city:=as.factor(city)]**

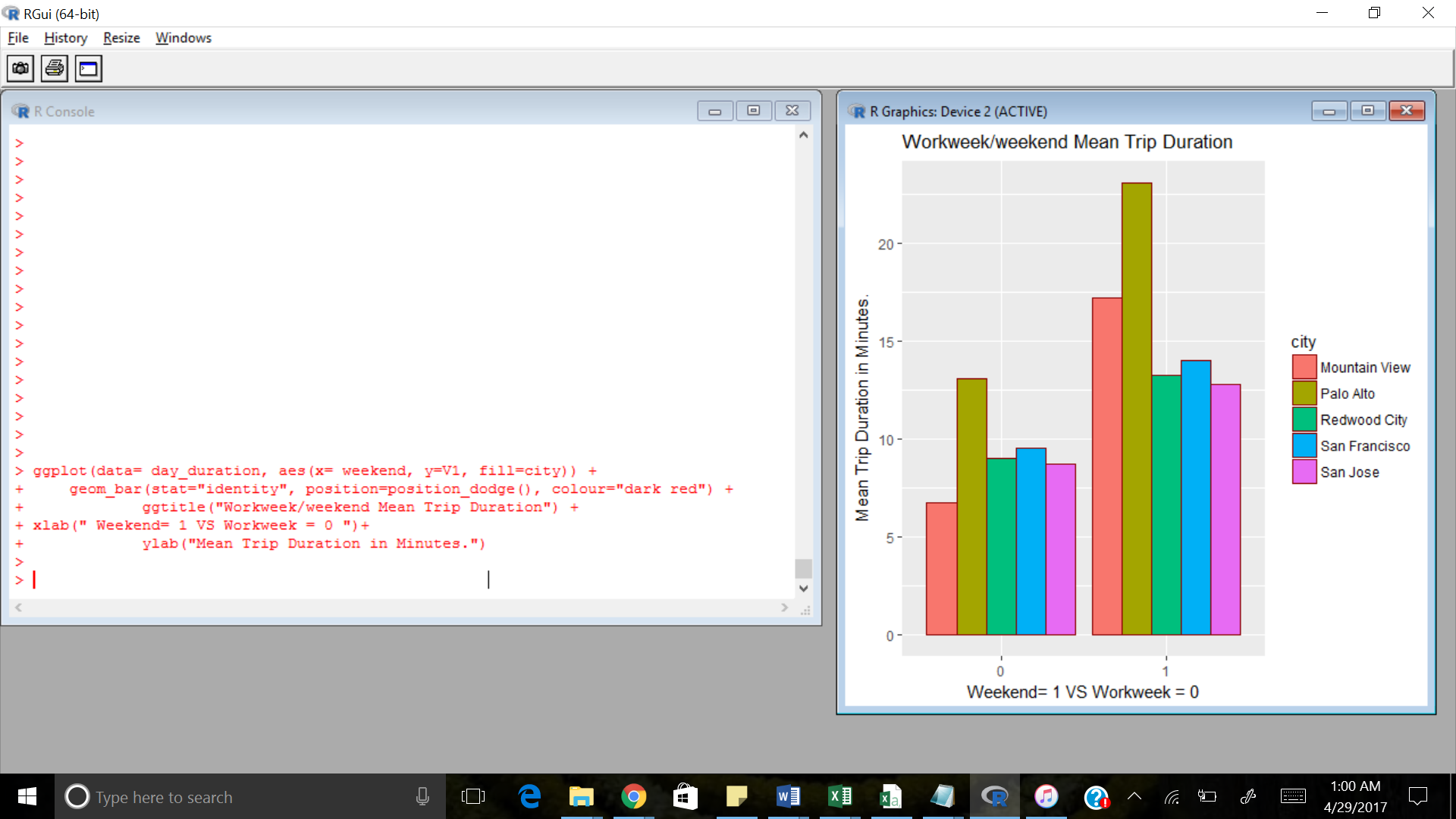
ggplot(data= day\_duration, aes(x= weekend, y=V1, fill=city)) +

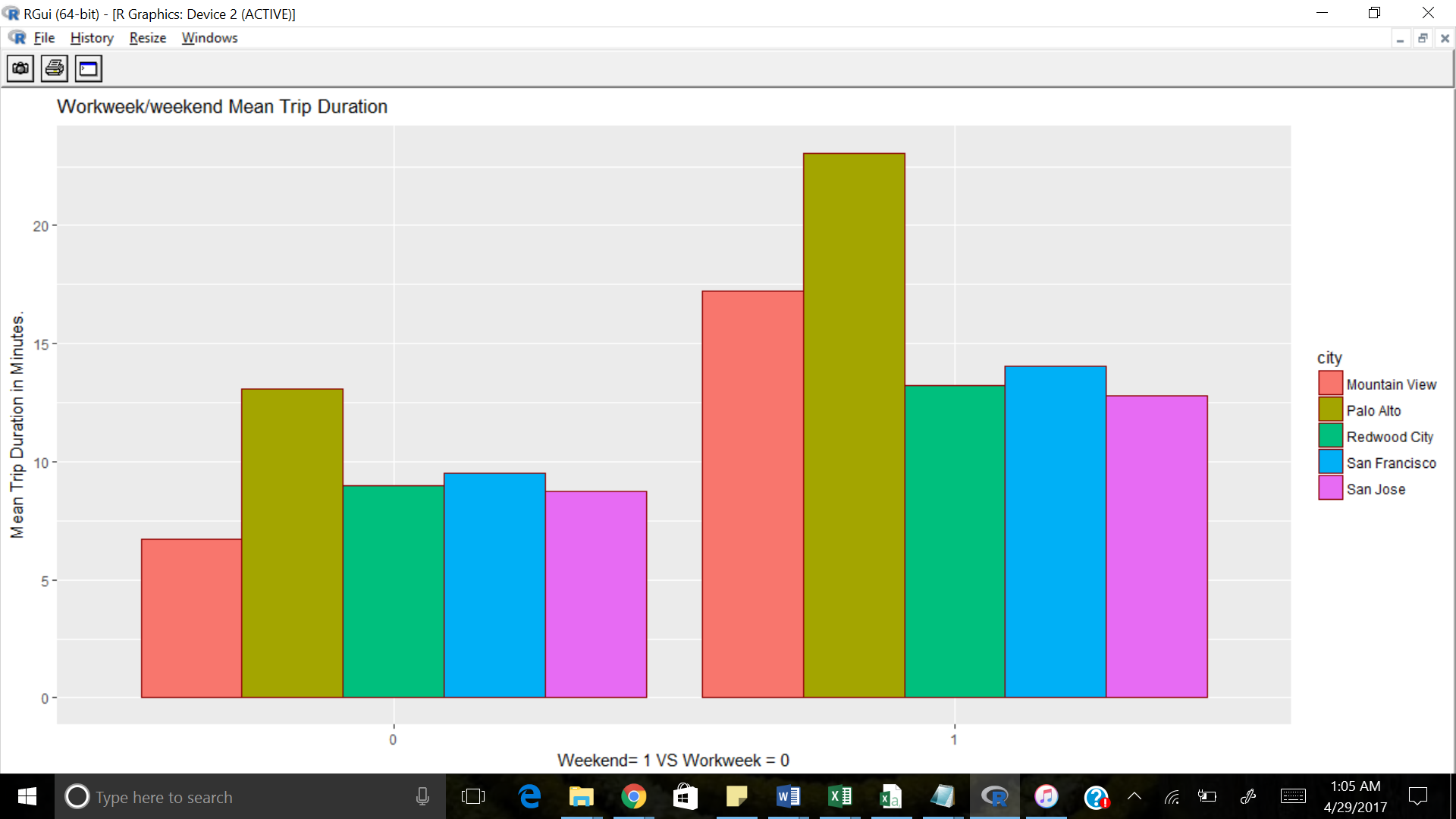
geom\_bar(stat="identity", position=position\_dodge(), colour="dark red") +

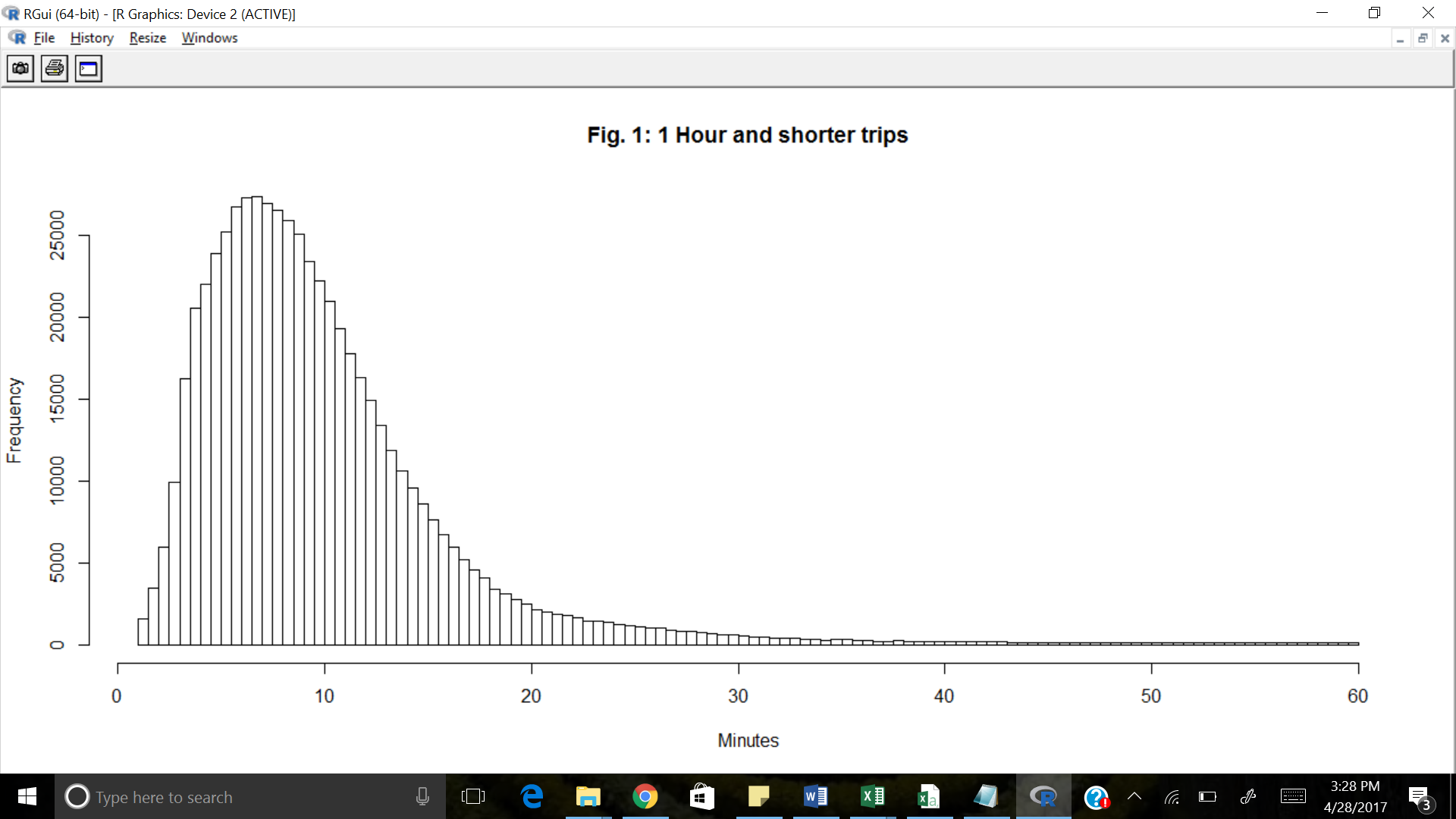
ggtitle("Workweek/weekend Mean Trip Duration") +

xlab(" Weekend= 1 VS Workweek = 0 ")+

ylab("Mean Trip Duration in Minutes.")







trips = trips[, dates := sapply(date, function(a) strsplit(a, " ")[[1]][1]) ]

> trips = trips[, times := sapply(date, function(a) strsplit(a, " ")[[1]][2]) ]

> trips = trips[, hours := sapply(times, function(a) strsplit(a, ":")[[1]][1])]

> trips$hours = as.numeric(trips$hours)

> trips = trips[, minutes := sapply(times, function(a) strsplit(a, ":")[[1]][2])]

> trips$minutes = as.numeric(trips$minutes)

> trips = trips[, month := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][1]))]

> trips = trips[, day := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][2]))]

> trips = trips[, year := sapply(dates, function(a) as.numeric(strsplit(a, "/")[[1]][3]))]

> trips = trips[, weekday := sapply(dates , function(x) weekdays(as.Date(x,"%m/%d/%Y"))) ]

> trips = trips[, c("date", "times") := NULL]

> trips = trips[, weekend := ifelse(weekday %in% c("Sunday", "Saturday"), 1, 0)]

> trips = trips[month %in% c(12,1,2), season:="Win"]

> trips = trips[month %in% c(3:5), season:="Spr"]

> trips = trips[month %in% c(6:8),season:="Sum"]

> trips = trips[month %in% c(9:11), season:="Aut"]

> summary(trips[,duration/60])

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 5.73 8.62 18.47 12.58 287800.00

> hour\_trips = trips[duration <= 60\*60, ]

> day\_trips = trips[duration <= 24\*60\*60, ]

> round(nrow(hour\_trips)/nrow(trips)\*100, 2)

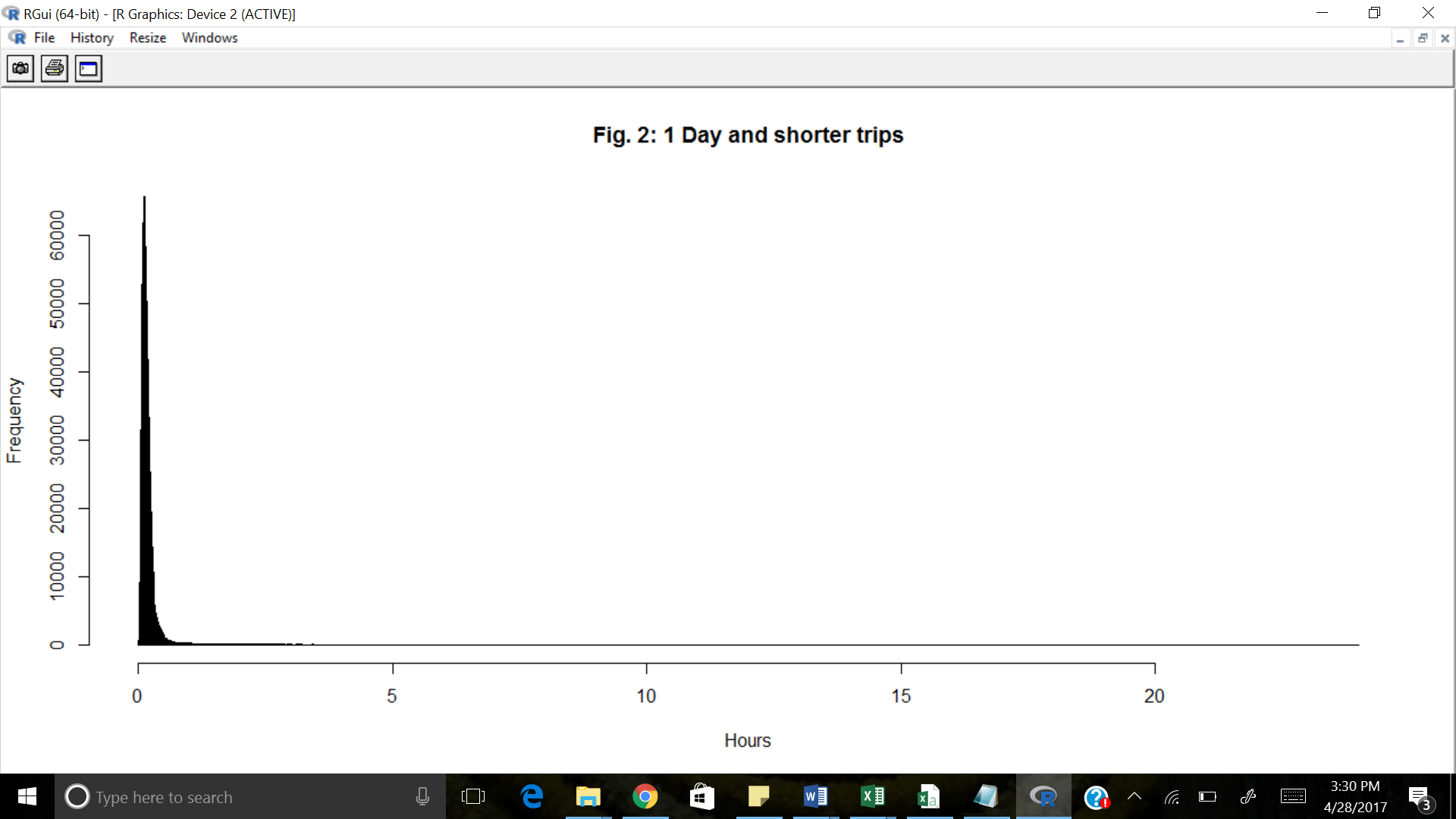
[1] 96.83

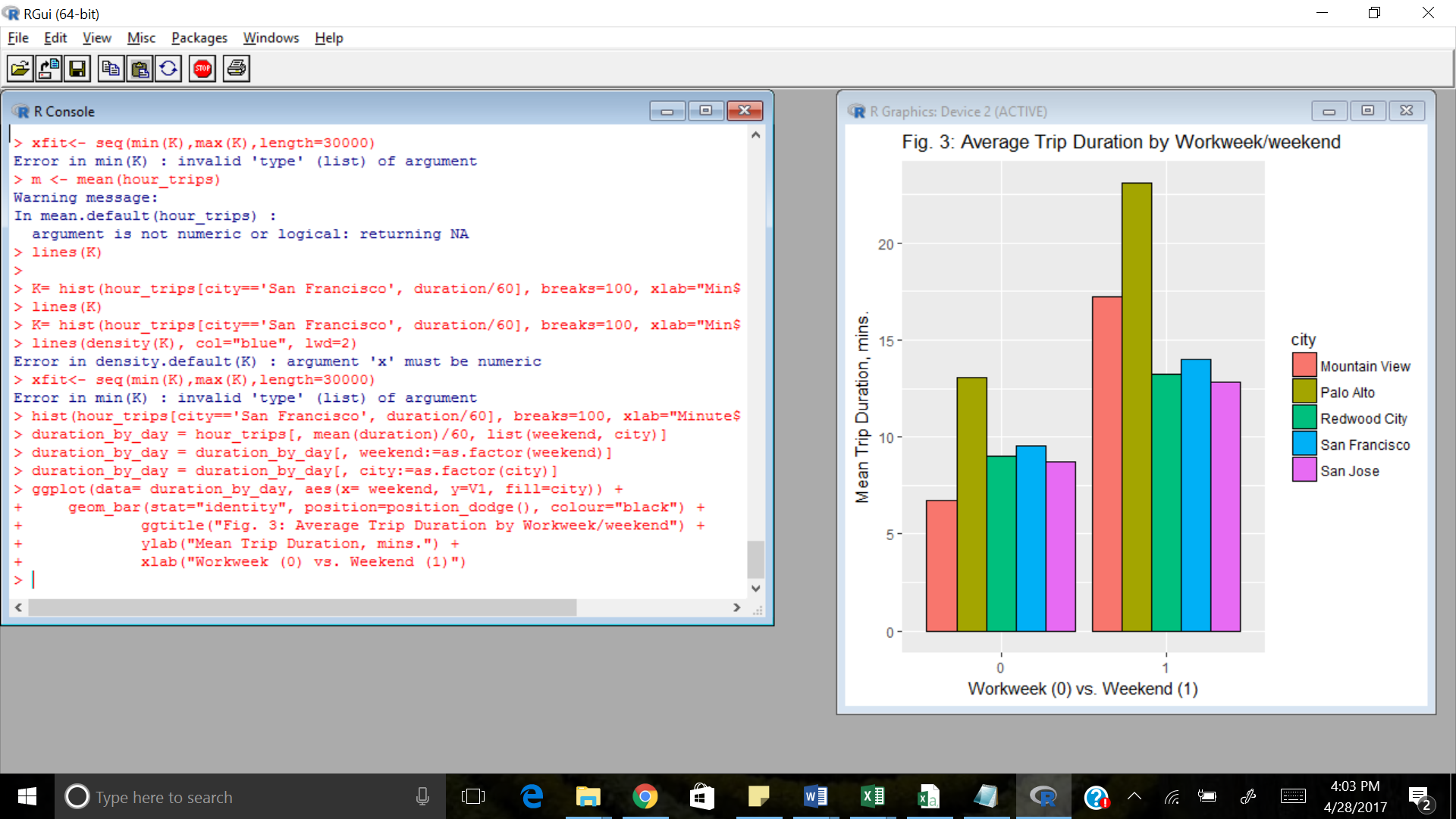
> round(nrow(day\_trips)/nrow(trips)\*100, 2)

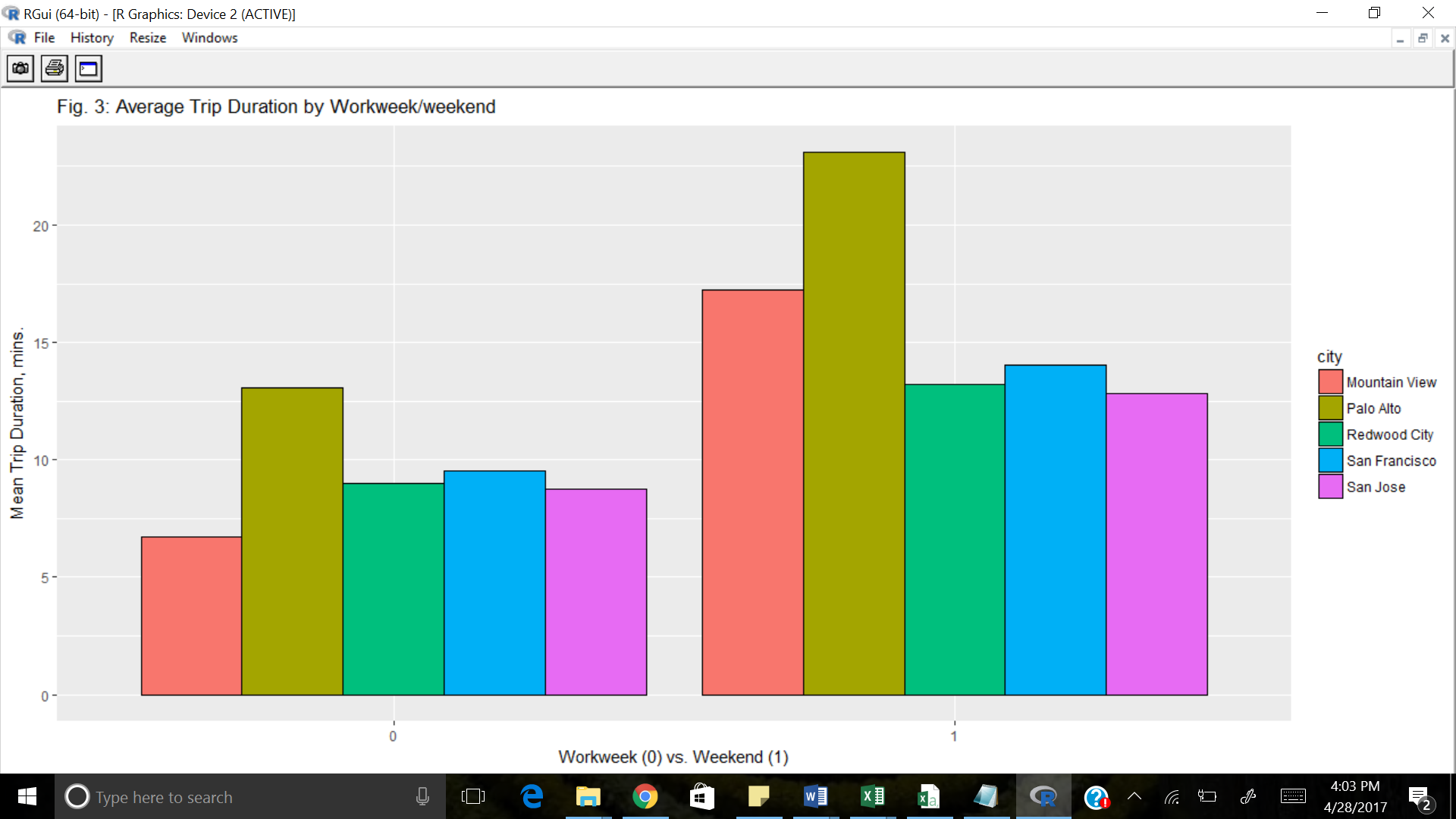
[1] 99.96

> hist(hour\_trips[city=='San Francisco', duration/60], breaks=100, xlab="Minutes", main="Fig. 1: 1 Hour and shorter trips")

> hist(day\_trips[city=='San Francisco', duration/60/60], breaks=1000, xlab="Hours", main="Fig. 2: 1 Day and shorter trips"







trip <- mutate(trip, weekend = (wday(trip$date) == 1 |wday(trip$date) == 7))