uber analysis

Krishnan Raman

10/6/2020

```
knitr::opts_chunk$set(echo = TRUE)
rm(list=ls())
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

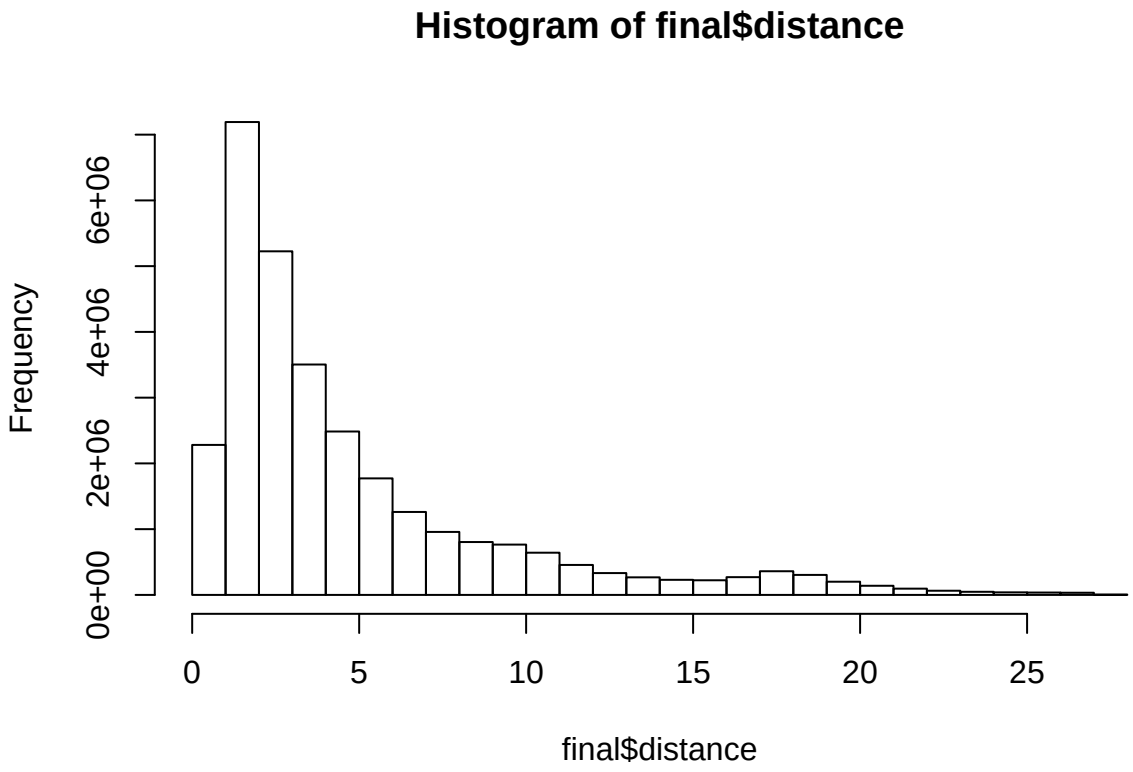
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

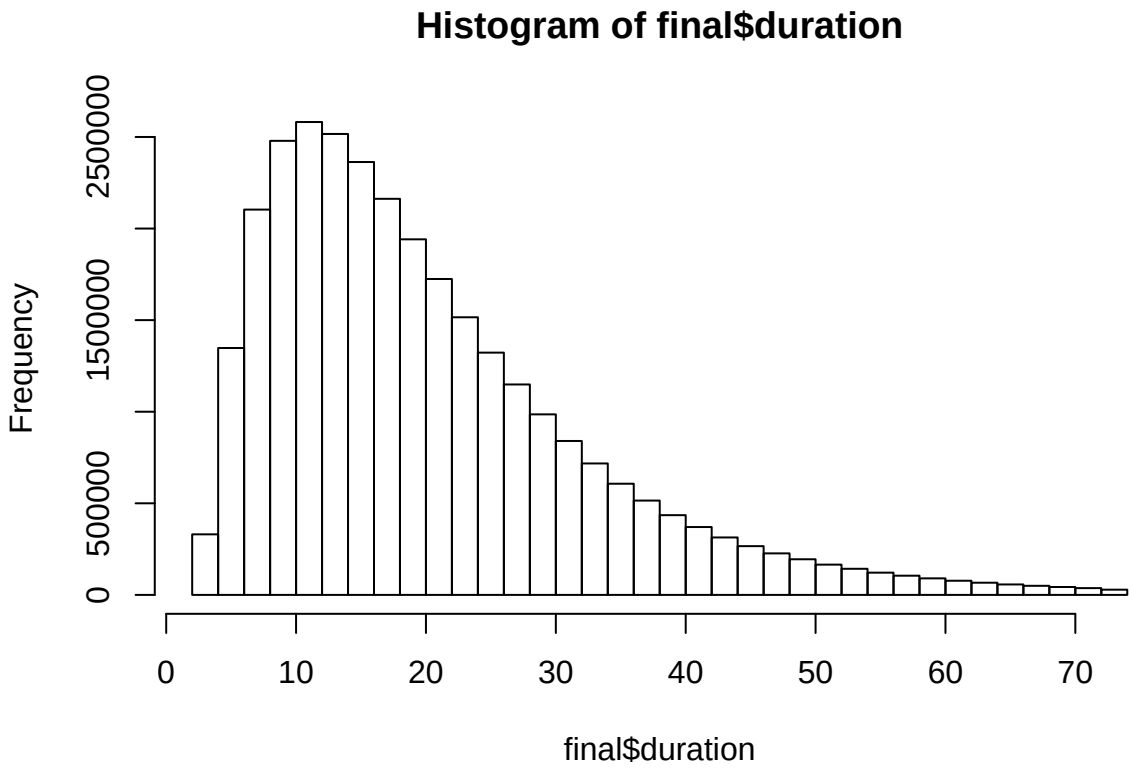
```
# CHANGE THIS TO LOCAL DRIVE WHERE uber_nyc_data.csv is located
# set nrows to 40 million
setwd("~/Desktop/695/uber-tlc-foil-response/uber-trip-data/")
df<-read.csv("uber_nyc_data.csv", nrows=1000*1000*40)

# convert factor vars to formatted numbers
df$distance = as.double(as.character(df$trip_distance))
```

```
## Warning: NAs introduced by coercion
```

```
df$duration = as.double(as.difftime(as.character(df$trip_duration), format = "%H:%M:%S", units = "mins")

# find 1% & 99% quantiles, eliminate anything beyond
# this helps with cancelled trips, overly long trips & other weird outlier cases
durq = quantile(df$duration,c(0.01, 0.99), names=F, na.rm=T)
disq = quantile(df$distance,c(0.01, 0.99), names=F, na.rm=T)

df2 = df[df$duration > durq[1] & df$duration < durq[2] & df$distance > disq[1] & df$distance < disq[2],]
df2 = select(df2,2:4, 7:8)

# remove NAs & prev dataframes
final = df2[complete.cases(df2), ]
rm(df,df2)

# now we are ready! lets look at distance distribution, duration distribution
hist(final$distance)
```




```
hist(final$duration)
```



```r
# get summary stats on distance & duration
summary(final)
```

```
##    origin_taz       destination_taz              pickup_datetime
##  2A     : 5439515   2A     : 4691257   2015-07-09 22:00:00:   12237
##  15     : 2063567   15     : 2161787   2015-06-27 22:00:00:   12113
##  4C     : 2006456   4C     : 1856022   2015-06-27 23:00:00:   11794
##  1      : 1956701   1      : 1723933   2015-06-27 21:00:00:   11453
##  6B     : 1736530   6B     : 1644425   2015-05-16 22:00:00:   11159
##  5C     : 1626098   11     : 1418915   2015-06-28 00:00:00:   11116
##  (Other):15164393   (Other):16496921   (Other)            :29923388
##     distance         duration
##  Min.   : 0.400   Min.   : 2.867
##  1st Qu.: 1.720   1st Qu.:10.967
##  Median : 3.080   Median :17.167
##  Mean   : 4.836   Mean   :20.152
##  3rd Qu.: 6.030   3rd Qu.:26.183
##  Max.   :27.170   Max.   :73.750
##
```