Lab 0 Solutions

Since a lot of you have used **dplyr** for analysis, this solution set uses an alternative approach using **base** R functions.

```
flights <- read.csv('flights14.csv')
head(flights)
```

```
  year month day dep_delay arr_delay carrier origin dest air_time distance hour
1 2014     1   1        14        13      AA    JFK  LAX      359     2475    9
2 2014     1   1        -3        13      AA    JFK  LAX      363     2475   11
3 2014     1   1         2         9      AA    JFK  LAX      351     2475   19
4 2014     1   1        -8       -26      AA    LGA  PBI      157     1035    7
5 2014     1   1         2         1      AA    JFK  LAX      350     2475   13
6 2014     1   1         4         0      AA    EWR  LAX      339     2454   18
```

Task 1

```
sort(table(flights$dest),decreasing = TRUE)[1:10]
```

```
  LAX   ATL   SFO   MCO   BOS   ORD   MIA   CLT   FLL   DCA
14434 12808 11907 11709 11609 11589  9928  9624  9471  6748
```

Task 2

```
routes <- paste(flights$origin,flights$dest,sep="-")
sort(table(routes),decreasing = TRUE)[1:3]
```

```
routes
JFK-LAX JFK-SFO LGA-ORD
  10208    7368    7052
```

Task 3

```
flights6 <- flights[flights$month==6,]
dat <- table(flights6$day,flights6$origin)
dat
```

```
   EWR JFK LGA
1  276 276 283
2  314 280 324
3  291 263 303
4  309 277 316
5  331 287 321
6  339 286 327
7  254 255 193
8  308 281 272
```

```
 9  328 284 316
10 321 278 299
11 325 267 297
12 298 262 284
13 269 258 233
14 250 258 199
15 302 285 276
16 339 291 330
17 345 285 325
18 337 289 315
19 331 294 315
20 342 293 326
21 255 269 204
22 319 293 273
23 340 292 327
24 342 293 322
25 304 275 302
26 330 296 325
27 342 293 324
28 256 272 208
29 313 295 266
30 339 295 312
```

Task 4

```
dat2 <- table(routes,flights$carrier)
dat2[dat2>0] <- 1
colSums(dat2)
```

```
 AA  AS  B6  DL  EV  F9  FL  HA  MQ  OO  UA  US  VX  WN
 24   1  56  56 103   2   2   1  32   3  48  12   7  16
```

Task 5

```
dat5 <- aggregate(arr_delay~carrier,data=flights,FUN=mean)
dat5[order(dat5$arr_delay,decreasing = FALSE),]
```

```
   carrier  arr_delay
2       AS -3.8885017
12      US  0.9997612
13      VX  3.2501563
4       DL  5.1552671
1       AA  5.4635769
11      UA  7.5645276
9       MQ  9.4957702
3       B6 10.1824681
14      WN 11.2175265
8       HA 12.4500000
5       EV 13.2214521
7       FL 13.6730616
10      OO 14.8250000
6       F9 26.6088795
```

Task 6

```
dat6 <- do.call(data.frame,
                aggregate(distance~carrier,
                          data=flights,FUN=function(x) c(total=sum(x),mean=mean(x))))
# ordering by total miles flown
dat6[order(dat6$distance.total,decreasing = TRUE),]
```

```
    carrier distance.total distance.mean
11      UA       72155634     1559.5486
4       DL       52575904     1261.3273
3       B6       47554862     1069.1531
1       AA       36407712     1384.2184
5       EV       22510199      565.3130
13      VX       11980325     2497.4620
14      WN       11715150      984.3010
9       MQ       10023882      540.1089
12      US        9178227      547.9539
2       AS        1378748     2402.0000
8       HA        1295580     4983.0000
7       FL         828067      661.9241
6       F9         759054     1604.7653
10      OO         143735      718.6750
```

```
# mean
dat6[order(dat6$distance.mean,decreasing = TRUE),]
```

```
    carrier distance.total distance.mean
8       HA        1295580     4983.0000
13      VX       11980325     2497.4620
2       AS        1378748     2402.0000
6       F9         759054     1604.7653
11      UA       72155634     1559.5486
1       AA       36407712     1384.2184
4       DL       52575904     1261.3273
3       B6       47554862     1069.1531
14      WN       11715150      984.3010
10      OO         143735      718.6750
7       FL         828067      661.9241
5       EV       22510199      565.3130
12      US        9178227      547.9539
9       MQ       10023882      540.1089
```

Task 7

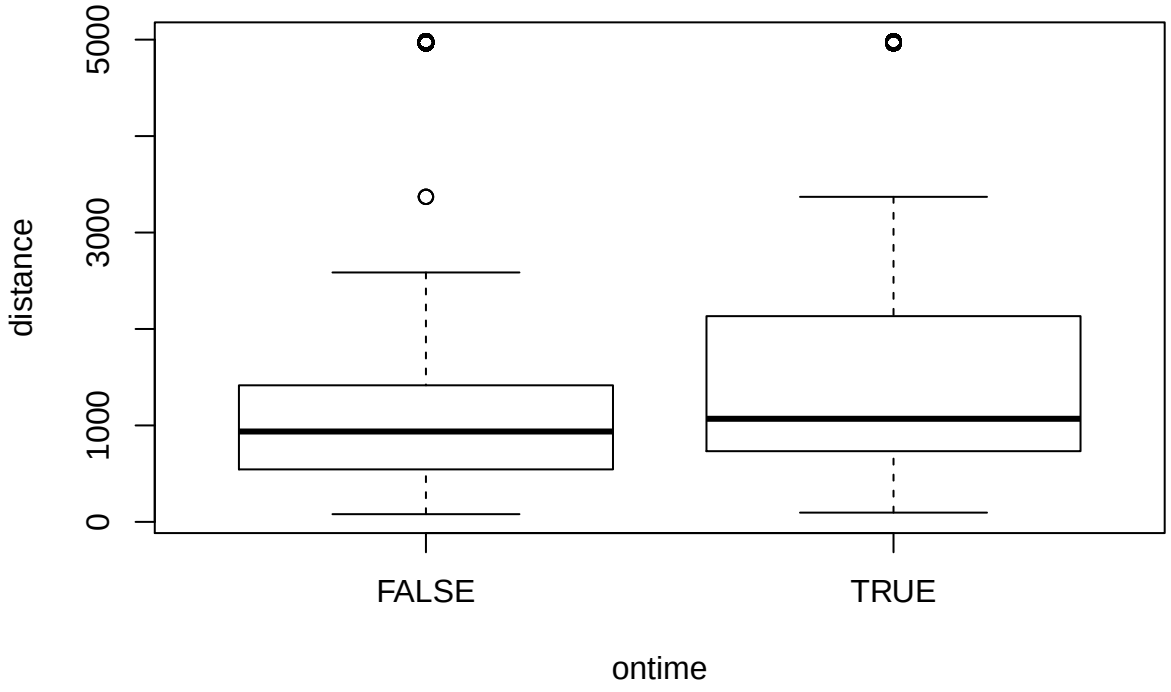
```
del_dep <- flights[flights$dep_delay>0,]
del_dep['ontime'] <- del_dep$arr_delay <= 0
print(paste("Number of flights that departed late and are on-time:",
            sum(del_dep$ontime)))
```

```
[1] "Number of flights that departed late and are on-time: 26593"
```

```
print(paste("Fraction of flights that departed late and are on-time:",
            round(mean(del_dep$ontime),3)))
```

```
[1] "Fraction of flights that departed late and are on-time: 0.267"
```

```r
with(del_dep,{
  boxplot(distance~ontime)
})
```




```r
with(del_dep,{
  boxplot(sqrt(distance)~ontime)
})
```