Project5

Elizabeth McGuckin

October 11, 2018

To begin this project I had to pull up the dataexpo2009 dataset for 2005 because I had to find the distance of all of the flights in all airlines for 2005

```
myDF <- read.csv("/depot/statclass/data/dataexpo2009/2005.csv")
head(myDF)
```

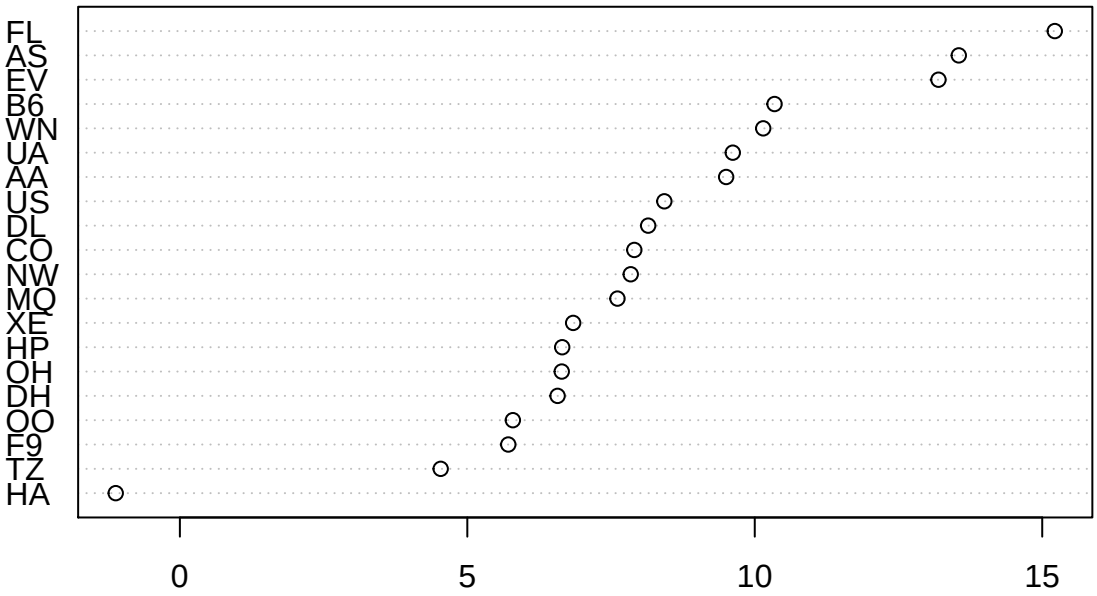
```
##   Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime CRSArrTime
## 1 2005     1         28         5    1603       1605    1741       1759
## 2 2005     1         29         6    1559       1605    1736       1759
## 3 2005     1         30         7    1603       1610    1741       1805
## 4 2005     1         31         1    1556       1605    1726       1759
## 5 2005     1          2         7    1934       1900    2235       2232
## 6 2005     1          3         1    2042       1900       9       2232
##   UniqueCarrier FlightNum TailNum ActualElapsedTime CRSElapsedTime AirTime
## 1            UA       541  N935UA               158            174     131
## 2            UA       541  N941UA               157            174     136
## 3            UA       541  N342UA               158            175     131
## 4            UA       541  N326UA               150            174     129
## 5            UA       542  N902UA               121            152     106
## 6            UA       542  N904UA               147            152      97
##   ArrDelay DepDelay Origin Dest Distance TaxiIn TaxiOut Cancelled
## 1      -18       -2    BOS  ORD      867      4      23         0
## 2      -23       -6    BOS  ORD      867      6      15         0
## 3      -24       -7    BOS  ORD      867      9      18         0
## 4      -33       -9    BOS  ORD      867     11      10         0
## 5        3       34    ORD  BOS      867      5      10         0
## 6       97      102    ORD  BOS      867      3      47         0
##   CancellationCode Diverted CarrierDelay WeatherDelay NASDelay
## 1                        0            0            0        0
## 2                        0            0            0        0
## 3                        0            0            0        0
## 4                        0            0            0        0
## 5                        0            0            0        0
## 6                        0           23            0        0
##   SecurityDelay LateAircraftDelay
## 1             0                 0
## 2             0                 0
## 3             0                 0
## 4             0                 0
## 5             0                 0
## 6             0                74
```

Including Plots

In question 1, I had to find the average distance for all the flights in the airlines in 2005. I used tapply to find the mean of the distance travelled of the unique carriers. I used na.rm=T to remove the blank data. I also had to make a dotchart, so I sorted these results from question 1.

```
question1 <- tapply(myDF$Distance, myDF$UniqueCarrier, mean, na.rm=T)
Q1 <- tapply(myDF$DepDelay, myDF$UniqueCarrier, mean, na.rm=T)
dotchart(sort(Q1))
```

```
## Warning in dotchart(sort(Q1)): 'x' is neither a vector nor a matrix: using
## as.numeric(x)
```



Before I began to answer question 2 I had to pull the data for taxis in June 2017. That is why I ran the following code

```
myCT <- read.csv("/depot/statclass/data/taxi2018/yellow_tripdata_2017-06.csv")
head(myCT)
```

```
##   VendorID tpep_pickup_datetime tpep_dropoff_datetime passenger_count
## 1        2  2017-06-08 07:52:31   2017-06-08 08:01:32               6
## 2        2  2017-06-08 08:08:18   2017-06-08 08:14:00               6
## 3        2  2017-06-08 08:16:49   2017-06-08 15:43:22               6
## 4        2  2017-06-29 15:52:35   2017-06-29 16:03:27               6
## 5        1  2017-06-01 00:00:00   2017-06-01 00:03:43               1
## 6        2  2017-06-01 00:00:00   2017-06-01 00:00:00               2
##   trip_distance RatecodeID store_and_fwd_flag PULocationID DOLocationID
## 1          1.03          1                  N          161          140
## 2          1.03          1                  N          162          233
## 3          5.63          1                  N          137           41
## 4          1.43          1                  N          142           48
## 5          0.60          1                  N          140          141
## 6         17.57          2                  N          132           74
##   payment_type fare_amount extra mta_tax tip_amount tolls_amount
## 1            1         7.5   1.0     0.5       1.86         0.00
## 2            1         6.0   1.0     0.5       2.34         0.00
## 3            2        21.5   1.0     0.5       0.00         0.00
## 4            1         8.5   1.0     0.5       0.88         0.00
## 5            1         4.5   0.5     0.5       2.00         0.00
## 6            1        52.0   0.0     0.5      11.71         5.76
##   improvement_surcharge total_amount
## 1                   0.3        11.16
```

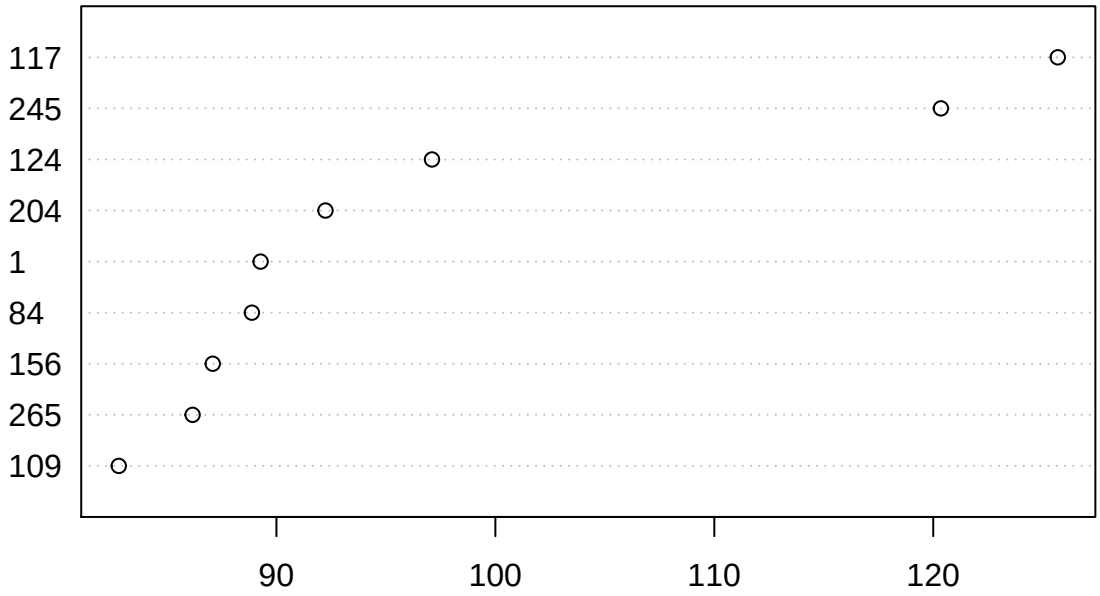
```
## 2                  0.3        10.14
## 3                  0.3        23.30
## 4                  0.3        11.18
## 5                  0.3         7.80
## 6                  0.3        70.27
```

Question 2 asked for the average total cost for taxi rides in certain locations in June of 2017.

To answer this question I used tapply to find the mean of the total amount of cost for each location. I used na.rm=T to remove the blank data I also had to create a dotchart, so I sorted my results to include only results greater than 80.

```
question2 <- tapply(myCT$total_amount, myCT$PULocationID, mean, na.rm=T)
Q2 <- tapply(myCT$total_amount, myCT$PULocationID, mean, na.rm=T)
dotchart(sort(Q2[Q2 > 80]))
```

```
## Warning in dotchart(sort(Q2[Q2 > 80])): 'x' is neither a vector nor a
## matrix: using as.numeric(x)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.