STAT 46700/CS 59000 Topics in Data Science Spring 2025

Lab 1

Vaishak Balachandra (0037831852)

Q.N. 1) Flight times (scheduled and actual) and taxi-out times of all Delta outbound flights from Atlanta in October 2004 are provided in the link below.

http://users.stat.ufl.edu/~winner/data/atltime1004a.dat

```
Variables/Columns
Flight Date  1-2
Flight Number  6-9
Tail Number  13-18
Destination city  22-24
Taxi-out time (minutes)  28-31
Scheduled flight time  35-38
Actual flight time  42-45
distance (miles)  49-52
```

Note that cancelled flights will have actual flight time=0.

a) Import the data in R without saving in your computer
b) How many flights were cancelled? Delete all the cancelled flights.
c) How many destinations does the Delta flight have from Atlanta?

```
> ########## q1
> # a
> url <- "https://users.stat.ufl.edu/~winner/data/atltime1004a.dat"
> q1 <- read.table(url)
> head(q1,6)
  V1         V2  V3 V4  V5  V6   V7
1  1  833N619DL ABQ 26 204 214 1250
2  1 1125N660DL ABQ 18 211 204 1250
3  1 1589N678DL ABQ 23 206 212 1250
4  2  833N659DL ABQ 18 204 194 1250
5  2 1125N656DL ABQ 23 211 210 1250
6  2 1589N665DN ABQ 19 206 212 1250
> tail(q1)
      V1         V2  V3 V4 V5 V6  V7
17807 30 1455N933DL VPS 13 73 59 250
17808 30 1803N980DL VPS 14 70 63 250
17809 31  370N973DL VPS 18 71 69 250
17810 31 1011N972DL VPS 19 77 69 250
17811 31 1455N911DL VPS 15 73 63 250
17812 31 1803N907DL VPS 37 70 85 250
> dim(q1)
[1] 17812     7
> # b
> names(q1)
[1] "V1" "V2" "V3" "V4" "V5" "V6" "V7"
> sum(q1$V6 == 0)
[1] 88
> cat("There are 88 flights got cancelled")
There are 88 flights got cancelled
> # or
```

```
> data_new = subset(q1, q1$V6==0)
> dim(data_new)
[1] 88  7
> cat("There are 88 flights got cancelled")
There are 88 flights got cancelled
> summary(q1)
       V1          V2                 V3                 V4               V5
 Min.   : 1   Length:17812       Length:17812       Min.   :  0.00   Min.   : 48.0
 1st Qu.: 8   Class :character   Class :character   1st Qu.: 15.00   1st Qu.: 84.0
 Median :16   Mode  :character   Mode  :character   Median : 19.00   Median :106.0
 Mean   :16                                         Mean   : 20.75   Mean   :128.4
 3rd Qu.:24                                         3rd Qu.: 25.00   3rd Qu.:138.0
 Max.   :31                                         Max.   :100.00   Max.   :580.0
       V6              V7
 Min.   :  0.0   Min.   : 132.0
 1st Qu.: 83.0   1st Qu.: 399.0
 Median :106.0   Median : 568.0
 Mean   :127.4   Mean   : 711.8
 3rd Qu.:141.0   3rd Qu.: 783.0
 Max.   :627.0   Max.   :4496.0
> # c
> length(table(q1$V3))
[1] 93
> cat("93 flights from Atlanta")
93 flights from Atlanta
```

Q.N. 2) Although all cases of AIDS in England and Wales must be reported to the Communicable Disease Surveillance Centre, there is often a considerable delay between the time of diagnosis and the time that it is reported. In estimating the prevalence of AIDS, account must be taken of the unknown number of cases which have been diagnosed but not reported. Under the Rich source of data in the Brightspace you will see the second data set "aids" .(You can access the data from the Data folder in the Brightspace too). The data set here records the reported cases of AIDS diagnosed from July 1983 and until the end of 1992.

a) Import the dataset in R.

b) Determine the dimension of the dataset.

c) List the variables included in the dataset.

d) Calculate the summary statistics of the delay.

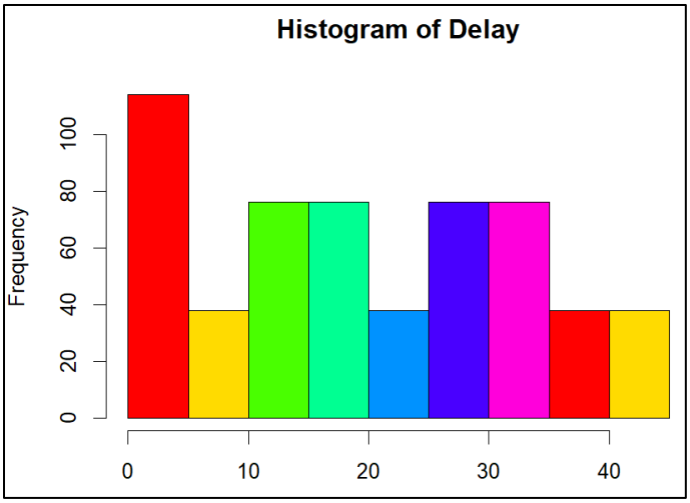
e) Draw a histogram of the delay.

f) Aggregate the data based on year to calculate the mean

g) Aggregate the data based on year and quarter to calculate the mean

```
> ########## q2
> # a
> # file.choose()
> q2 <- read.csv("C:\\Users\\PNW_checkout\\Downloads\\sem2\\0. Coursework\\Data science\\Lab\\Lab 1
\\aids.csv")
> head(q2)
  X year quarter delay dud time y
1 1 1983       3     0   0    1 2
2 2 1983       3     2   0    1 6
3 3 1983       3     5   0    1 0
4 4 1983       3     8   0    1 1
5 5 1983       3    11   0    1 1
6 6 1983       3    14   0    1 0
> attach(q2)
> # b
> dim(q2)
```

```
[1] 570   7
> cat("There are 570 rows and 7 columns")
There are 570 rows and 7 columns
> # c
> colnames(q2)
[1] "X"       "year"    "quarter" "delay"   "dud"     "time"    "y"
> # d
> summary(delay)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   20.00   20.07   32.00   41.00
> # e
> hist(delay, main = "Histogram of Delay", col = rainbow(7))
> # f
> ?aggregate
> aggregate(data.frame(q2), by = list(year), FUN = mean)
   Group.1     X year quarter    delay       dud time         y
1     1983  15.5 1983     3.5 20.06667 0.0000000  1.5  0.800000
2     1984  60.5 1984     2.5 20.06667 0.0000000  4.5  1.633333
3     1985 120.5 1985     2.5 20.06667 0.0000000  8.5  3.583333
4     1986 180.5 1986     2.5 20.06667 0.0000000 12.5  7.183333
5     1987 240.5 1987     2.5 20.06667 0.0000000 16.5 10.016667
6     1988 300.5 1988     2.5 20.06667 0.0000000 20.5 13.566667
7     1989 360.5 1989     2.5 20.06667 0.0500000 24.5 15.483333
8     1990 420.5 1990     2.5 20.06667 0.3000000 28.5 17.850000
9     1991 480.5 1991     2.5 20.06667 0.5666667 32.5 18.300000
10    1992 540.5 1992     2.5 20.06667 0.8333333 36.5 17.233333
> # g
> aggregate(data.frame(q2), by = list(year, quarter), FUN = mean)
   Group.1 Group.2   X year quarter    delay        dud time          y
1     1984       1  38 1984       1 20.06667 0.00000000    3  0.9333333
2     1985       1  98 1985       1 20.06667 0.00000000    7  3.1333333
3     1986       1 158 1986       1 20.06667 0.00000000   11  5.4666667
4     1987       1 218 1987       1 20.06667 0.00000000   15  8.9333333
5     1988       1 278 1988       1 20.06667 0.00000000   19 11.6000000
6     1989       1 338 1989       1 20.06667 0.00000000   23 14.9333333
7     1990       1 398 1990       1 20.06667 0.20000000   27 18.7333333
8     1991       1 458 1991       1 20.06667 0.46666667   31 18.0666667
9     1992       1 518 1992       1 20.06667 0.73333333   35 20.6666667
10    1984       2  53 1984       2 20.06667 0.00000000    4  1.0000000
11    1985       2 113 1985       2 20.06667 0.00000000    8  2.6666667
12    1986       2 173 1986       2 20.06667 0.00000000   12  8.0000000
13    1987       2 233 1987       2 20.06667 0.00000000   16  9.4000000
14    1988       2 293 1988       2 20.06667 0.00000000   20 14.0666667
15    1989       2 353 1989       2 20.06667 0.00000000   24 14.6000000
16    1990       2 413 1990       2 20.06667 0.26666667   28 16.3333333
17    1991       2 473 1991       2 20.06667 0.53333333   32 17.5333333
18    1992       2 533 1992       2 20.06667 0.80000000   36 21.2000000
19    1983       3   8 1983       3 20.06667 0.00000000    1  0.8000000
20    1984       3  68 1984       3 20.06667 0.00000000    5  2.0000000
21    1985       3 128 1985       3 20.06667 0.00000000    9  4.2000000
22    1986       3 188 1986       3 20.06667 0.00000000   13  7.2666667
23    1987       3 248 1987       3 20.06667 0.00000000   17 10.2000000
24    1988       3 308 1988       3 20.06667 0.00000000   21 14.9333333
25    1989       3 368 1989       3 20.06667 0.06666667   25 16.8666667
26    1990       3 428 1990       3 20.06667 0.33333333   29 17.3333333
27    1991       3 488 1991       3 20.06667 0.60000000   33 20.4000000
28    1992       3 548 1992       3 20.06667 0.86666667   37 18.2000000
29    1983       4  23 1983       4 20.06667 0.00000000    2  0.8000000
30    1984       4  83 1984       4 20.06667 0.00000000    6  2.6000000
31    1985       4 143 1985       4 20.06667 0.00000000   10  4.3333333
32    1986       4 203 1986       4 20.06667 0.00000000   14  8.0000000
33    1987       4 263 1987       4 20.06667 0.00000000   18 11.5333333
34    1988       4 323 1988       4 20.06667 0.00000000   22 13.6666667
35    1989       4 383 1989       4 20.06667 0.13333333   26 15.5333333
36    1990       4 443 1990       4 20.06667 0.40000000   30 19.0000000
37    1991       4 503 1991       4 20.06667 0.66666667   34 17.2000000
38    1992       4 563 1992       4 20.06667 0.93333333   38  8.8666667
```



Q.N. 3) Data from four different regions (*region*) related to four different products are provided in the Brightspace. Import the data in R and choose from East to display the distribution by product type.

```
> ########## q3
> # install.packages("readxl")
> library(readxl)
> # file.choose()
> q3 <- read_xlsx("C:\\Users\\PNW_checkout\\Downloads\\sem2\\0. Coursework\\Data science\\Lab\\Lab
1\\region.xlsx", sheet = "East")
> head(q3)
# A tibble: 6 × 5
  Month `Porduct - A` `Porduct - B` `Porduct - C` `Porduct - D`
  <chr>         <dbl>         <dbl>         <dbl>         <dbl>
1 Jan            1860          1202          1371          1973
2 Feb            1191          1811          1618          1289
3 Mar            1890          1724          1088          1880
4 Apr            1804          1732          1964          1541
5 May            1913          1937          1613          1357
6 Jun            1016          1895          1229          1285
> dim(q3)
[1] 12  5
> names(q3) = c("Month", "Product_A","Product_B","Product_C","Product_D")
> boxplot(q3[,-1], col = rainbow(3), main = "Boxplot Distribution of the Region \nbased on Data Typ
e")
```