<u>סטטיסטיקה / תרגיל 1#</u>

אריאל סטולרמן
קבוצה 03

(1)

- matrix - אובייקט ב-R להחזקת נתונים במערך דו-מימדי, כאשר המידע המאוחסן בכל תאיו הנו מאותו סוג (למשל character, logical וכו'). ניתן להתייחס לעמודות או לשורות כאל וקטורים.
- data.frame - אובייקט ב-R להחזקת נתונים בקבוצה המאורגנת באופן דומה למטריצה (כמו מערך דו מימדי), כאשר המידע המאוחסן בכל תאיו אינו מחוייב להיות מאותו סוג. ניתן להתייחס לעמודות כאל וקטורים.
- vector - אובייקט ב-R להחזקת נתונים במערך חד מימדי, כמו מטריצה בעלת שורה אחת, כל הנתונים מאותו סוג.
- list - אובייקט ב-R המהווה אוסף כלשהו של אובייקטים אחרים ב-R, שאינם חייבים להיות מאותו סוג. רשימות יכולות לכלול בתוכן רשימות אחרות.
- function - ניתן ליצור פונקציות המהוות אובייקטים ב-R על מנת לחסוך חזרות קוד, לקצר עבודה ולפשט. ישנן פונקציות מובנות בשפה, כגון var, plot וכיו"ב. ניתן לכתוב פונקציות ע"י שימוש בפונקציה edit.
- formula – פונקציה המאפשרת לחלץ נוסחאות הנכללות באובייקטים שונים ב-R.
- expression – אובייקט ב-R המהווה וקטור של ביטויים ב-R <u>לא</u> משוערכים, מעיין מעטפת, אותם ניתן לשערך ע"י הפקודה eval.

(2)

```
ex01q2=read.table('http://www-
stat.stanford.edu/~tibs/ElemStatLearn/datasets/ozone.data', header=T)
```

(a) בקובץ 4 עמודות, כלומר 4 משתנים (ו-111 נתונים בכל עמודה):

```
> dim(ex01q2)
[1] 111   4
```

(b) שמות המשתנים הם: ozone, radiation, temperature, wind:

```
> names(ex01q2)
[1] "ozone"       "radiation"   "temperature" "wind"
```

(c) ozone – numeric; radiation – integer; temperature – integer; wind – numeric:

```
> class(ex01q2$'ozone')
[1] "numeric"
> class(ex01q2$'radiation')
[1] "integer"
> class(ex01q2$'temperature')
[1] "integer"
```

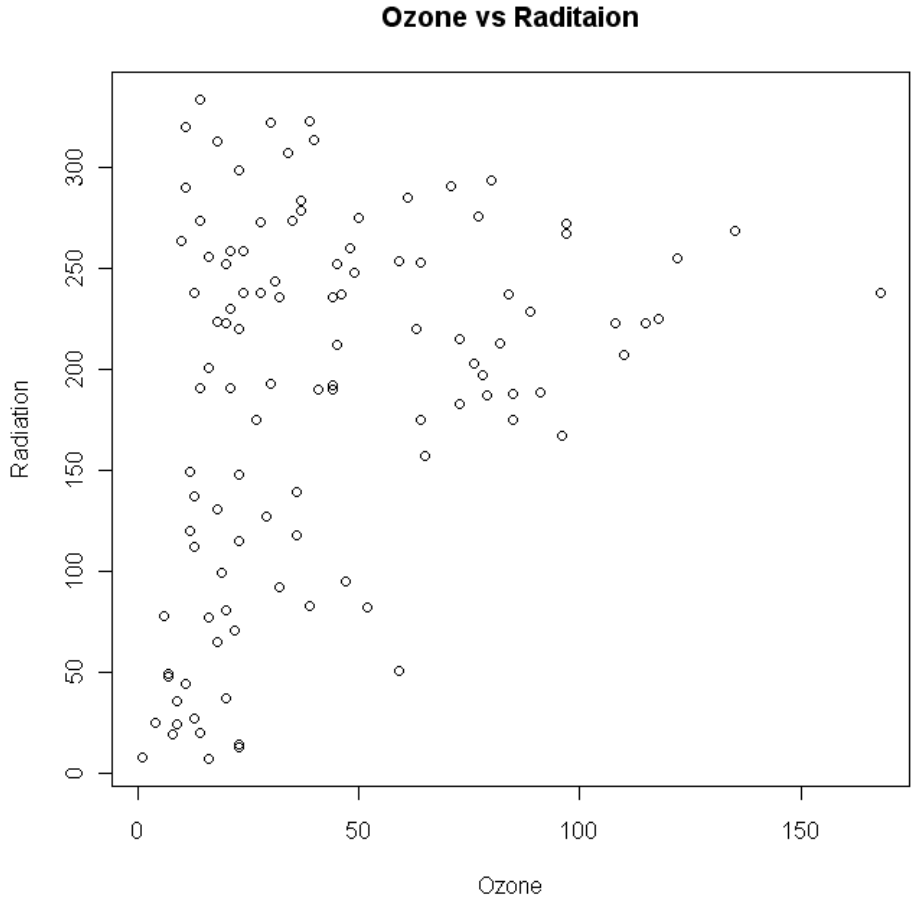
```
> class(ex01q2$'wind')
[1] "numeric"
```

(d) פלט הפקודה summary נותן תקציר על כל אחת מעמודות טבלת הנתונים. במקרה זה כל הנתונים מספריים, וניתנת האינפורמציה הבאה: מינימום, מקסימום, רבעון ראשון, חציון, רבעון שלישי וממוצע:

```
> summary(ex01q2)
     ozone         radiation      temperature        wind
 Min.   :  1.0   Min.   :  7.0   Min.   :57.0   Min.   : 2.300
 1st Qu.: 18.0   1st Qu.:113.5   1st Qu.:71.0   1st Qu.: 7.400
 Median : 31.0   Median :207.0   Median :79.0   Median : 9.700
 Mean   : 42.1   Mean   :184.8   Mean   :77.8   Mean   : 9.939
 3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:84.5   3rd Qu.:11.500
 Max.   :168.0   Max.   :334.0   Max.   :97.0   Max.   :20.700
```

(e), (f)

```
> x=ex01q2$'ozone'
> y=ex01q2$'radiation'
> plot(y~x, ylab='Radiation', xlab='Ozone', main='Ozone vs Raditaion')
```



(g)

```
> Tf=ex01q2$'temperature'
> temp.celcius=(5/9)*(Tf-32)
```

(h) טמפרטורת המינימום היא 13.88889 מעלות והמקסימום 36.11111 מעלות:

```
> min(temp.celcius)
[1] 13.88889
> max(temp.celcius)
[1] 36.11111
```

(3)

```
> a=c(1,5,4,6,8,5)
> a
[1] 1 5 4 6 8 5
> b=rep(11, times=9)
> b
[1] 11 11 11 11 11 11 11 11 11
> c=seq(from=1, to=15, by=2)
> c
[1]  1  3  5  7  9 11 13 15
```

(4)

```
> x=seq(from=23,to=99.6,by=0.3)
> y=matrix(x,ncol=16)
```

(a) תשובה: 35.9

```
> y[12,3]
[1] 35.9
```

(b) תשובה: 15680

```
> sum(x)
[1] 15680
```

(c) תשובה: 15680 (לא, זה לא מפתיע)

```
> sum(y)
[1] 15680
```

(d) תשובה: 1086227

```
> z=x*x
> sum(z)
[1] 1086227
```

(5) הערה: כיוון שהשורה הראשונה של טבלת הנתונים מכילה כותרות, הנתונים נשמרו ל-ex01q5 כאשר header=T (בניגוד לקוד הכתוב בקובץ התרגיל), והתשובות בהתאם:

```
>ex01q5=read.table('http://www-
stat.stanford.edu/~tibs/ElemStatLearn/datasets/marketing.data', header=T)
```

(a) בקובץ 14 משתנים ו-8992 תצפיות

```
> dim(ex01q5)
[1] 8992   14
```

(b) תשובה: 7

```
> ex01q5[7015,13]
[1] 7
```

(c) תשובה: בקובץ חסרים 2693 נתונים

```
> table(is.na(ex01q5))
FALSE   TRUE
123195   2693
>
```

(d) תשובה: 3.456376

```
> mean(subset(ex01q5, X2==2, select=X5))
      X5
3.456376
```

(e) תשובה: 6096

```
> sum(subset(ex01q5, X2==1, select=X3))
[1] 6096
```

(6)

(a) Tab delimited values - המאפיין קבצים אלו הוא שההפרדה בין השדות בכל שורה ושורה בטבלת הנתונים היא רווח TAB. קריאת נתונים בהם המפריד הוא TAB יכולה להיעשות בשני דרכים: האחת ע"י שימוש בפקודה read.delim או read.delim2, או שימוש פשוט ב-read.table, בה ברירת המחדל היא מפריד white space (הכולל בין היתר TAB). למרות שזו ברירת המחדל, מפורשות ניתן להגדיר זאת ע"י הוספת המתג `sep="\t"`.

(b) Comma separated values (csv) - המאפיין קבצים אלו הוא שהמספרים המיוצגים בהם מיוצגים ע"י פסיק בתור נקודה עשרונית, וההפרדה בין השדות היא ע"י נקודה-פסיק / המספרים מיוצגים עם נקודה עשרונית רגילה, ופסיק מהווה מפריד בין שדות. ניתן להשתמש בפקודות read.csv או read.csv2 בכדי לקרוא נתונים מסוג זה, או להשתמש במתגים `sep=";", dec=","` (sep למפריד ו-dec לקביעת סימן לנקודה עשרונית).

(c) xls - המאפיין קבצים אלו הוא שמקורם בקבצי xls (Exels sheet), ועל מנת לקרוא אותם יש להתקין את חבילת xlsReadWrite ולהשתמש בפקודת read.xls (שאר המתגים דומים לפקודת read.table).