**R Programming course**

**جلسه سوم**

نکته: در قسمت Environment، متغیرها نمایش داده میشوند که هر زمان که بخواهیم میتوانیم آن را پاک کنیم و دیتای جدیدی در آن وارد کنیم. پاک کردن این قسمت زمانی که نیازی به متغیرها نداریم از این جهت مهم است که ممکن است متغیر جدیدی که وارد میکنیم هم نام متغیر گذشته باشد و در این صورت اطلاعات ممکن است با هم ادغام شوند و نتیجه نادرستی را در نهایت ببینیم. راه حل دیگر برای پاک کردن متغیرهای موجود استفاده از دستور rm(list = ls()) است.

همانطور که قبلا هم گفته شده بود، برای import و export کردن یکسری اطلاعات و فایلها از سیستم خود نیاز است که پوشه ای که فایل مورد نظر در آن قرار دارد را با دستور setwd (“ “) مشخص کنیم و نام آن را در بین " " قرار دهیم. (اگر این کار را نکنیم R ، از مسیر دیفالت خودش دنبال فایلتان میگردد. با دستور getwd() میتوان به مسیر دیفالت R دست پیدا کرد و در این صورت میتوان فایل مورد نظر را در فایلی که مسیر دیفالت است قرار داد و آن را فراخوانی کرد.)

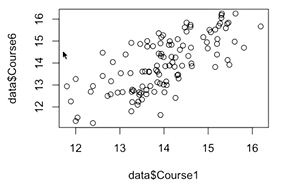
برای import کردن دیتا و دیدن اطلاعات آن از دستور data <- read.csv (“*data’s name*”, sep= “,”, row.names= NULL) استفاده میکنیم.

برای اینکه دیتافریم را مشاهده کنیم یا میتوانیم در قسمت Environment روی فایل مورد نظر double click کنیم و یا اینکه دستور view(data) را run کنیم.

**Plots:**

**نمودار هیستوگرام:**

* به طور کلی از دستور hist(data$Course2) استفاده میشود (به عنوان مثال Course2 یک متغیر و ستون در دیتافریممان است).
* با دستور hist (data$ Course2, breaks = 20) میتوان تعداد مستعطیل هایی که در نمودار به ما نشان میدهد را تغییر دهیم. (بجای 20 میتوان اعداد متفاوتی را استفاده کرد)
* آپشن: مثلا میتوان در کد بالا بعد از breaks = 20 با گذاشتن یک , و زدن گزینه tab (روی کیبورد)، آپشن هایی که میتوانیم با این کد و این نمودار داشته باشیم را مشاهده کنیم و انتخاب کنیم.
* با دستور hist(data$Course2, breaks = 20, probability=T) میتوان بجای تعداد (frequency)، تراکم یا درصد (density) را در نمودار هیستوگرام مشاهده کرد. (در اصل تعداد را درصد تبدیل میکند)
* با دستور hist(data$Course2, breaks=20, probability=T, col=”lightblue”)، علاوه بر اطلاعات بالا میتوان هیستوگرام را به نام رنگی که در col= “” نوشتیم تغییر دهیم. (نام رنگ با املا و فرمی که نوشته میشود مهم است که در R تعریف شده باشد)
* با دستور hist(data$Course2, breaks=20, probability=T, col=”lightblue”, border = “darkblue”, main = “Course2”) میتوان با کد border، رنگ اطراف نمودار را تغییر داد و با کد main، نمودار را نام گذاری کرد.
* میتوان برچسب محورهای x و y را تغییر داد: …, xlab = “Score”, ylab = “Density of scores”)
* میتوان حدود محورهای x و y را تغییر داد و آن را بیشتر یا کمتر کرد: … , xlim = c(0,20), ylim = c(0, 0.8) )
* درصورتی که بخواهیم نمودار هیستوگرام ما نام نداشته باشد میتوان اینگونه عمل کرد: …, main = “ “ و یا …, main = NULL.
* اگر بخواهیم بر روی نمودار یک خط vertical (عمودی) اضافه کنیم میتوان از دستور مقابل استفاده میکنیم: abline(v=9) ، abline( v=9, col= “red”) (در این حالت خطی که روی نمودار رسم میشود عمودی، روی داده 9 و به رنگ قرمز خواهد بود.
  + همچنین میتوان ضخامت خط کشیده شده را بیشتر یا کمتر کرد. دیفالت خط در R، 1 است و میتوان از دستور …, lwd = 2، استفاده کرد و عدد را بزرگتر از 1 (ضخیم تر) و کوچکتر از 1 (نازک تر) ترسیم کرد.
  + با استفاده از دستور lty = 3، میتوان نوع خط را تغییر داد (با تغییر اعداد در lty میتوان نوع خط را تغییر داد و مثلا آن را نقطه چین کرد).
  + برای ترسیم خط افقی بجای v از h (horizontal) استفاده میکنیم: abline (h= 0.5, col = “violet”, lwd = 2, lty = 3)



Scatter plot

در این نمودار قرار است یک متغیر را بر اساس یک متغیر دیگر plot کنیم.

* مثلا در این نمودار هر دایره مربوط به یک دانشجو است که بر

اساس عملکردش در cours1 و course6 رسم شده است.

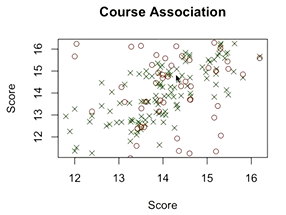
* در اصل این نمودار ارتباط دو متغیر را براساس هم نشان میدهد که

میشود همبستگی را هم از آن استنباط کرد.

* برای کشیدن آن از دستور: ( plot (x = data$Cours1, y = data$Course6
* میتوان تعیین کرد که جنس و نوع نشانگر های داخل نمودار چه چیز باشد:

Plot (x = data$Cours1, y = data$Course6, col = “darkgreen”, pch = 4) با تغییر عدد pch میتوان نوع نشانگرها را تغییر داد.

* همیچنین میتوانیم سایز نشانگرها تغییر داد: …, cex=1.5)
* برای fit کردن یک خط روی داده های نمودار از دستور مقابل استفاده میکنیم: abline (line (x = data$Course1, y = data$ Course6), main = “Course Association”, col = “orange”, lwd = 5, lty = 3)
* همچنان میتوانیم از دستور ylab= “” و xlab = “”، عنوان محورها تغییر داد.
* اگر که بخواهیم بر روی نمودار بالا ارتباط یکی از کورسهای محور x و یا y را با یک کورس دیگر (مثلا Course3) را نشان دهیم میتوان از دستور مقابل استفاده کرد: points(x = data$Course1, y = data$Course3, col=”red”, pch = 5)
  + در اصل در این نمودار محور x برای Course1



ثابت است اما محور y برای نشانگرهای سبز رنگ،

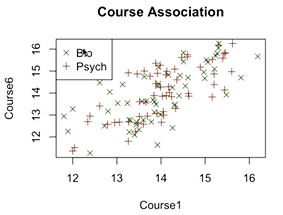
نشان دهنده Course6 است و برای نشانگر های قرمز،

نشان دهنده Course3.

* حال اگر در dataset خود یک متغیر اسمی داشته باشیم که چند variable داشته باشد و بخواهیم ارتباط دو متغیر را با تفکیک متغیرهای اسمی نشان دهیم به شیوه selection عمل میکنیم:
  + Plot(x = data$Course1[data$Field ==”Bio”], y = data$Course6[data$Field == “Bio”], col = “darkgreen”, pch = 4, main = “Course Association”, xlab = “Course1”, ylab = “Course6”)

در این حالت بر روی نمودار تنها نمره های رشته Bio را براساس نمره کورس 1 و 6 داریم.

* + حال اگر بخواهیم علاوه بر رشته Bio، رشته Psych را هم در دو کورس 1 و 6 بر روی یک نمودار ببینیم از این دستور استفاده میکنیم. Points(x= data$Course1[data$Field == “Psych”], y = data$Course6[data$Field == “Psych”], col = “darkred”, pch = 2)



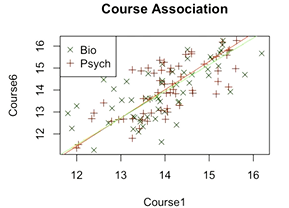
نکته: نمودار روبه رو از نظر بصری مشخص نشده است که هر علامت و رنگ مربوط به کدام متغیراسمی است که ما نسبت داده ایم (Bio و Psych). از دستور زیر برای نام گذاری علامت ها در نمودار استفاده میکنیم:

Legend(“topleft”, legend = c(“Bio”, “Psych”), pch = c(4, 2), col = c(“darkgreen”, “darkred”))

* + میتوانیم برای هر یک از متغیرهای داخل نمودار یک خط هم fit کنیم:

abline(line(x = data$Course1[data$Field == “Bio”], y = data$Course6[data$Field == “Bio”], col = “green”)

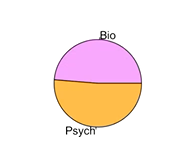
abline (line(x=data$Course1[data$Field == “Psych”], y= data$Course6[data$Field ==”Psych”], col = “red”)



**Pie chart**

این نمودار مخصوص متغیرهای اسمی است و میزان آن را در یک دایره به ما نشان میدهد.

* ابتدا با دستور data$Field میتوانیم تمام داده های مربوط به متغیر Field که یک متغیر اسمی است را ببینیم.
* دستور as.factor(data$Field) به ما نشان میدهد که متغیر Field شامل چند lable است. (در دیتاست ما دوتاست: Bio و Psych).
* و اگر از دستور tabulate(as.factor(data$Field)) استفاده کنیم به ما میگوید از هر داده متغیر Field چندتا وجود دارد. (در این دیتاست: Bio = 56 , Psych = 59 است)
* دستور نمودار دایره ای: pie(x = c(56, 59), labels = c(“Bio”, “Psych”), col = c(“violet”, “orange”))



**Bar plot**

زمانی که میخواهیم فراوانی یک متغیر اسمی (این نکته حائز اهمیت است که برای بصری کردن فراوانی متغیر های اسمی از bar plot استفاده میکنیم و برای متغیرهای عددی و غیر اسمی از histogram استفاده میکنیم)

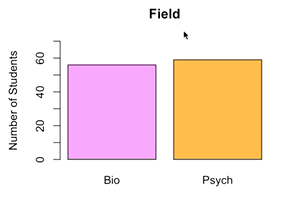
* مانند کاری که برای نمودار دایره ای انجام دادیم باید متغیر اسمی و تعداد و ارزشهایش را مشخص کنیم:

data$Field

as.factor(data$Field)

tabulate(as.factor(data$Field))

* دستور نمودار میله ای: barplot(height = c(56,59),names.arg = c(“Bio”, “Psych”), col=c(“violet”, “orange”), main = “Field”, ylim = c(0,70), ylab = “Number of Students”))



* کاربرد دیگر این نمودار برای نمایش میانگین است. مثلا میخواهیم ببینیم در یک

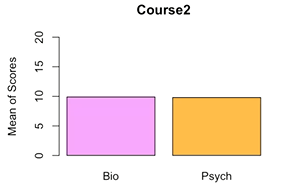
کورس خاص (مثلا Course2) دانشجویان در دو رشته Bio و Psych در چه

وضعیتی هستند و کدام یک بهتر عمل کردند. در این حالت ابتدا میانگین نمره های هر دو رشته را در کورس2 بدست میاوریم و سپس آن را plot میکنیم:

mean(data$Course2[data$Field == “Bio”], na.rm = T)

mean(data$Course2[data$Field == “Psych”], na.rm = T)

barplot(height = c(9.9, 9.8), names.arg = (“Bio”, “Psych”), col = c(“violet”, “orange”), main = “Course2”, ylim = c(0,20), ylab = “mean of scores”)

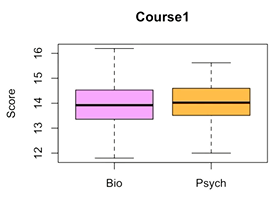


**Box Plot**

این نموار برای متغیرهای پیوسته کشیده میشود و توزیع داده ها را به صورت عمودی نشان میدهد.

* دستور نمودار جعبه ای: boxplot(Course1 ~ Field, data, col = “red”)
* میتوان بر اساس تعداد نمودار ها رنگ های متفاوتی قرار داد: boxplot(Course1 ~ Field, data, col = “Violet”, “orange”)
* نهایتا میتوان اطلاعات دیگر هم اضافه کرد:

boxplot(Course1 ~ Field, data, col = “violet, “orange”, ylim = c(10, 18), main = “Course1”, xlab = “”, ylab = “Score”)



برای اینکه بتوان plotها را در غالب یک فایل pdf داشته باشیم راه های متفاوتی وجود دارد. یکی از آنها استفاده از دستور مرتبط است:

Pdf(“Boxplot.pdf”)

boxplot(Course1 ~ Field, data, col = “violet, “orange”, ylim = c(10, 18), main = “Course1”, xlab = “”, ylab = “Score”)

boxplot(Course2 ~ Field, data, col = “violet, “orange”, ylim = c(10, 18), main = “Course1”, xlab = “”, ylab = “Score”)

…

Dev.off()

به این صورت هر تعداد پلاتی که داریم را میتوانیم در یک صفحه از pdf نشان میدهد.