به نام خدا

دانشگاه صنعتی شریف - دانشکده مهندسی کامپیوتر

آمار و احتمال مهندسی

پاییز 1401

تمرین عملی بخش دوم

طراحان: محمدجواد ماهرالنقش، محمدمهدی ابوترابی

موعد تحویل: 14 آبان

همفکری در تمامی تمرین‌های درس توصیه می‌شود. در عین حال از شما خواسته می‌شود تا تمام پیاده‌سازی را به تنهایی و بدون مشاهده کد دیگران انجام دهید.

لطفا در فایل ارسالی تمام بلوک‌های کد اجرا شده و شامل نمودارها و خروجی‌های لازم باشند.

امیررضا آذری

99101087

سوال اول

دیتافریم airquality یکی از دیتا فریم‌های biult-in است که اطلاعاتی در مورد وضعیت آب‌وهوای نیویورک در یک بازه‌ی زمانی می‌دهد.

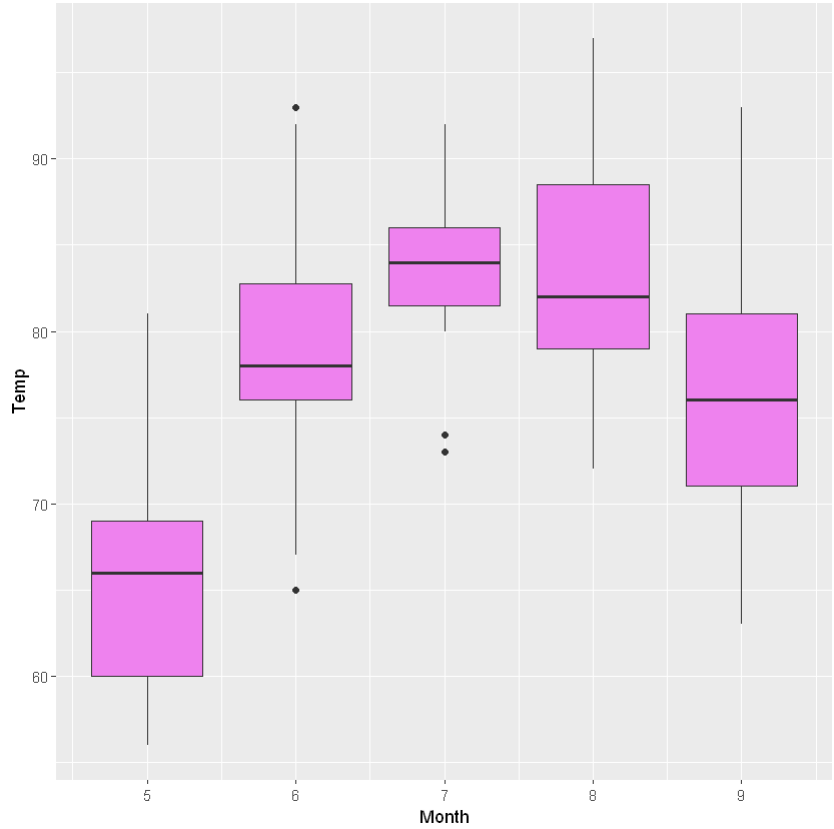
In [1]:
```
# AmirReza Azari
# 99101087
head(airquality)
```

A data.frame: 6 × 6

	Ozone	Solar.R	Wind	Temp	Month	Day
	<int>	<int>	<dbl>	<int>	<int>	<int>
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

الف) به کمک boxplotها یک نمودار مناسب ارائه دهید که وضعیت دما (temp) را بر حسب ماه‌های مختلف نشان دهد.

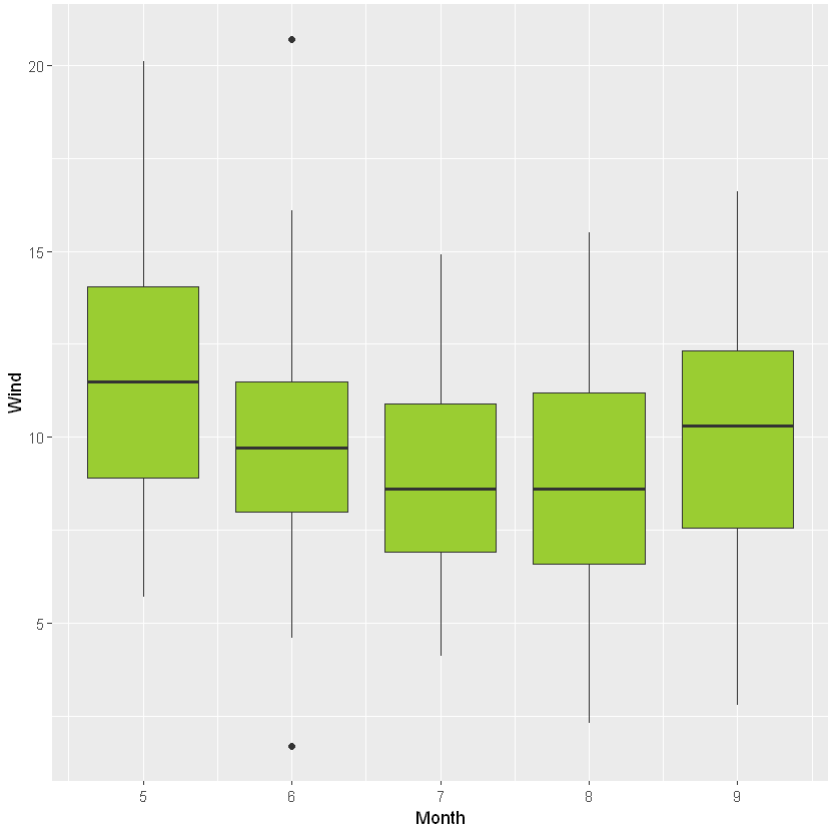
In [3]:
```
library(ggplot2)
ggplot(airquality) +
  geom_boxplot(aes(x = Month, y = Temp, group = Month), fill = "violet")
```



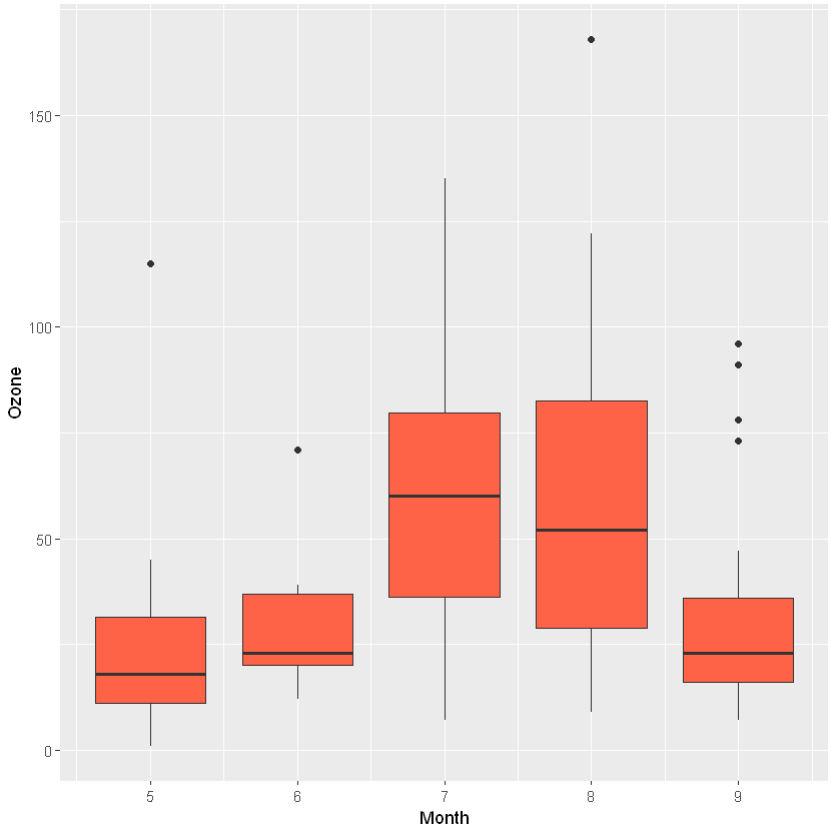
ب) بخش الف را برای میانگین باد و میانگین غلظت اوزون و میانگین تابش خورشید بر حسب ماه‌های مختلف تکرار کنید و نتایج خود را از این 4 نمودار شرح دهید.

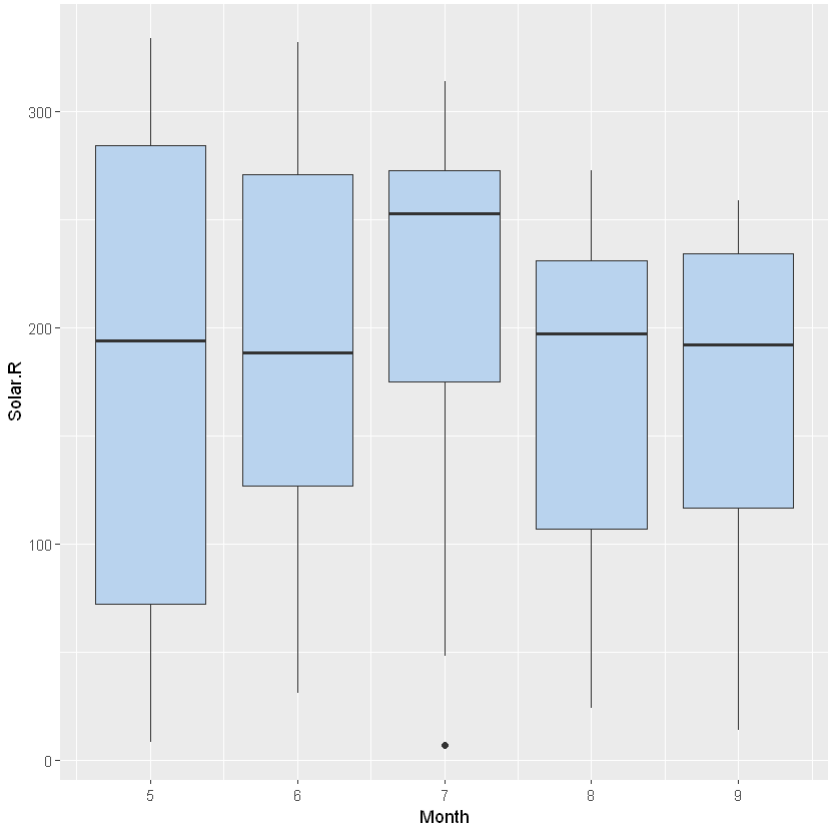
In [4]:
```
ggplot(airquality) +
  geom_boxplot(aes(x = Month, y = Wind, group = Month), fill = "yellowgreen")
ggplot(airquality) +
  geom_boxplot(aes(x = Month, y = Ozone, group = Month), fill = "tomato")
ggplot(airquality) +
  geom_boxplot(aes(x = Month, y = Solar.R, group = Month), fill = "slategray2")
# چهار نمودار ما نشان می دهند هرگاه میانگین دما در یک ماه پایین تر بوده است،
# میانگین باد در آن ماه بیشتر و میانگین غلظت اوزون نیز کمتر بوده
# است. همچنین برای میانگین تابش خورشید همانطور که مشخص می باشد،
# در اکثر ماه ها در یک رنج ثابتی بوده است اما در ماهی که بیشترین میانگین دما
# را دارا است، این میزان به حداکثر خود رسیده است.
```

```
Warning message:
"Removed 37 rows containing non-finite values (`stat_boxplot()`)."
```



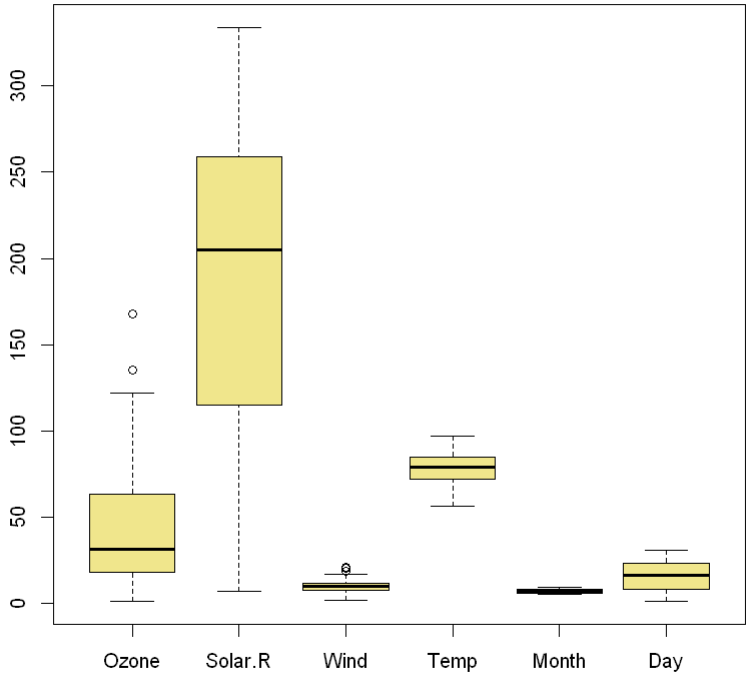

```
Warning message:
"Removed 7 rows containing non-finite values (`stat_boxplot()`)."
```


150
100
50
0
Ozone
5
6
7
8
9
Month



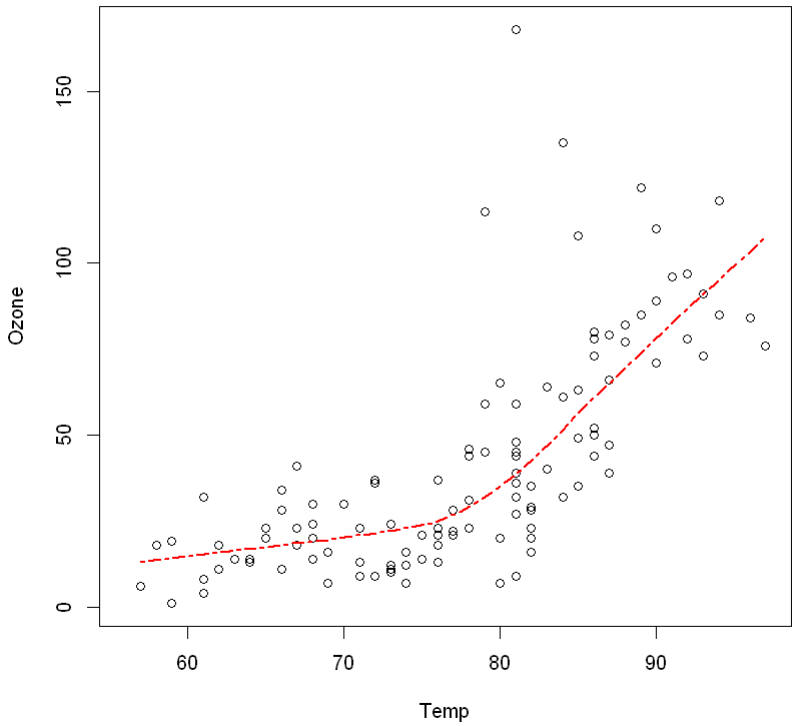
ج) به کمک boxplotها رنج اعداد مختلف در پارامترهای متفاوت را در این دیتاست را بررسی کنید.

In [2]:
```
# Temp:
# ما، همانطور که مشخص است، در ماه 5ام ، بیشتر رنج دما بین 60 تا 68 می باشد و میانگین کلی هم تقریبا برابر 66 درجه می باشد
# اما دما های دیگری خارج از این بازه هم وجود دارد که کمترین آن حدود 56 و بیشترین حدود 81 می باشد.
# ام، رنج بین 76 تا 83 می باشد و میانگین حدودا 83 می باشد. داده های خارج از محدوده که کیس های خاص ب حساب می آیند هم
# قابل مشاهده است.
# به ترتیب برای ماه های دیگر نیز، رنج دمایی و میانگین دمای آن ماه به طور مشخص قابل مشاهده می باشد.
# ********************************************************************************************
# Wind:
# همانطور که دکر کردیم، پارامتر باد و دما نسبت عکس به هم دارند.
# در ماه 5ام با میانگین دمای 66، رنج میزان باد حدود 8.5 تا 14.5 می باشد و میانگین برابر 11.5 است.
# در ماه 6ام نیز رنج به بین 6 تا 11 می باشد و میانگین چیزی حدود 9.5 است.
# ماه های دیگر نیز به همین منوال داده هایشان نشان داده شده و رابطه بین دما و باد مشخص می باشد.
# ********************************************************************************************
# Ozone:
# رابطه بین غلظت اوزون و دما نیز گویا مستقیم است و هرگاه در ماهی، میانگین دمای بیشتری داشته ایم،
# میزان غلظت اوزون هم بیشتر بوده است.
# برای مثال در ماه 5ام رنج بیشتر آن در حدود 10 تا 30 با میانگین حدودی 20 می باشد.
# اما برای مثال در ماه 7 ام طیف تفاوت داده ها بیشتر از ماه 6ام و 5ام می باشد.
# و همچنین شاهد هستیم که ماه 7ام دارای بالاترین میانگین غلظت اورون می باشد زیرا بالاترین میانگین دما نیز،
# در اختیار ماه 7ام است.
# ********************************************************************************************
# Solar.R:
# میزان تابش خورشید را می توان با مقایسه پارامتر های دیگر منطقی دانست.
# برای مثال در مقایسه ماه 7ام و 8ام، هر دو میانگین باد تقریبا یکسانی دارند،
# اما ب دلیل میانگین دمایی بالاتر در ماه 7ام، میزان تابش خورشید و همچنین میزان غلظت اوزون نیز بیشتر است.
# نکته دیگر که قابل ذکر می باشد، این است که در ماه 5ام طیف تابش خورشید بیشتر است اما در ماه 7ام،
# طیف کوچکتری را در بر گرفته است اما به دلیل بالاتر بودن این رنج، باعث افزایش دما نیز شده است.
boxplot(airquality, col = "khaki")
```



د) به کمک نمودار scatter نمودار میانگین غلظت اوزون بر حسب دما را رسم کنید و آن را تحلیل کنید.

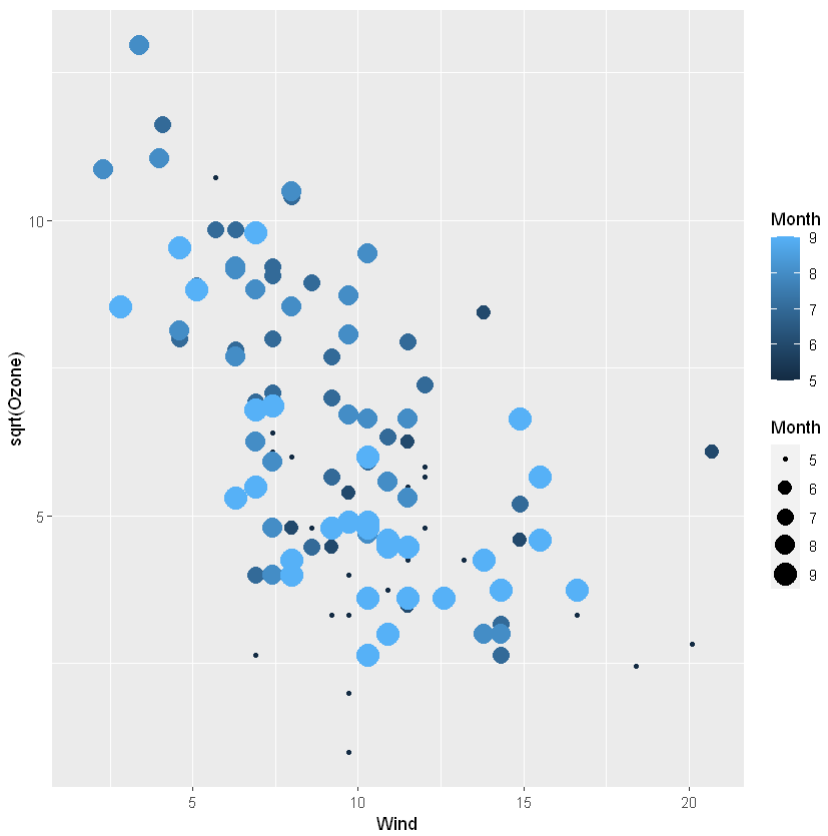
In [6]:
```
with(airquality, scatter.smooth(Temp, Ozone, lpars = list(col = "red",
                                                          lwd = 2, lty = 6)))
# همانطور که در بخش قبل نیز ذکر کردیم، رابطه تقریبا مستقیم بین دما و غلظت اوزون وجود دارد که نمودار هم این را به طور واضح
# نشان داده است.
```



ه) نمودار نقطه‌ای (scatter) را براساس ویژگی‌های جذر میانگین غلظت اوزون و ماه و باد رسم کنید. سعی کنید نمودار رسم شده اطلاعات را به خوبی نشان دهد. نتایج خود را بیان کنید. (به وابستگی میان متغیرها توجه کنید.)

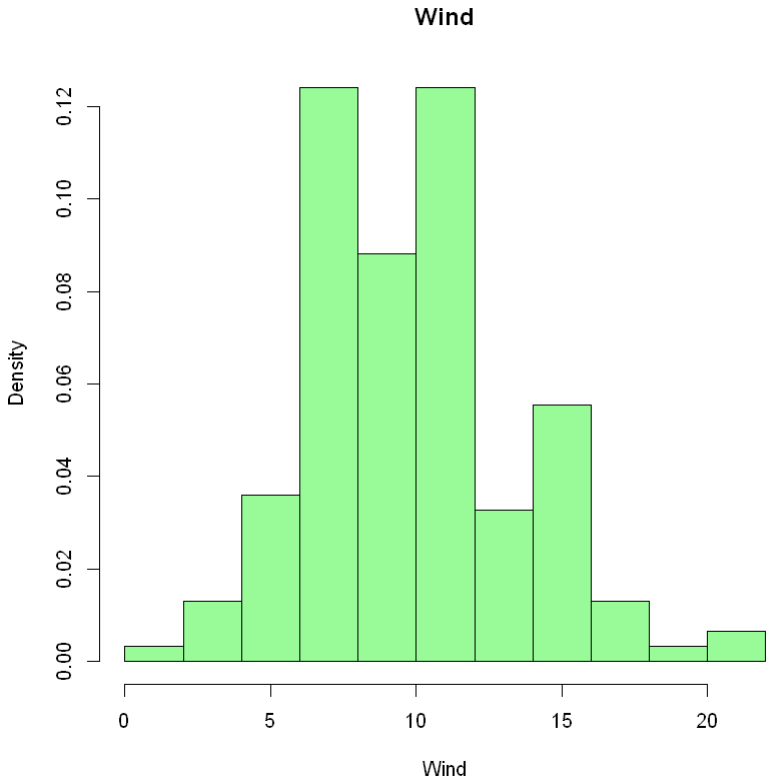
In [4]:
```
# one solution is making 3 scatter:
# with(airquality, scatter.smooth(Temp, sqrt(Ozone), lpars = list(col = "red",
#                                                                 lwd = 3.5, lty = 2)))
# with(airquality, scatter.smooth(Temp, sqrt(Wind), lpars = list(col = "blue",
#                                                                lwd = 3.5, lty = 2)))
# with(airquality, scatter.smooth(Temp, sqrt(Solar.R), lpars = list(col = "purple",
#                                                                lwd = 3.5, lty = 2)))
# another solution with 1 scatter:
ggplot(airquality, aes(x=Wind, y=sqrt(Ozone), size=Month, color = Month)) +
  geom_point()
# همانطور که مشاهده می شود، وابستگی متغیر ها به دما و نتیجه ای که از کشیدنboxplot
# ها گرفته بودیم، در این نمودار ها واضح است.
# دما و علظت اوزون رابطه ای تقریبا مستقیم دارند و هر چه دما بالاتر رفته است،
# به طور میانگین غلظت اوزون هم بیشتر شده است.
# رابطه بین دما و باد اما بدین شکل نیست و همانطور که مشخض می باشد،
# نوعی رابطه معکوس با هم دارند.
# رابطه میان دما و تابش خورشید نیز جالب به  نظر می آید
# تقریبا می توان گفت نسبت مستقیمی به هم دارن و با افزایش دما،
# تابش خورشید هم زیاد شده است و بالعکس!!
```

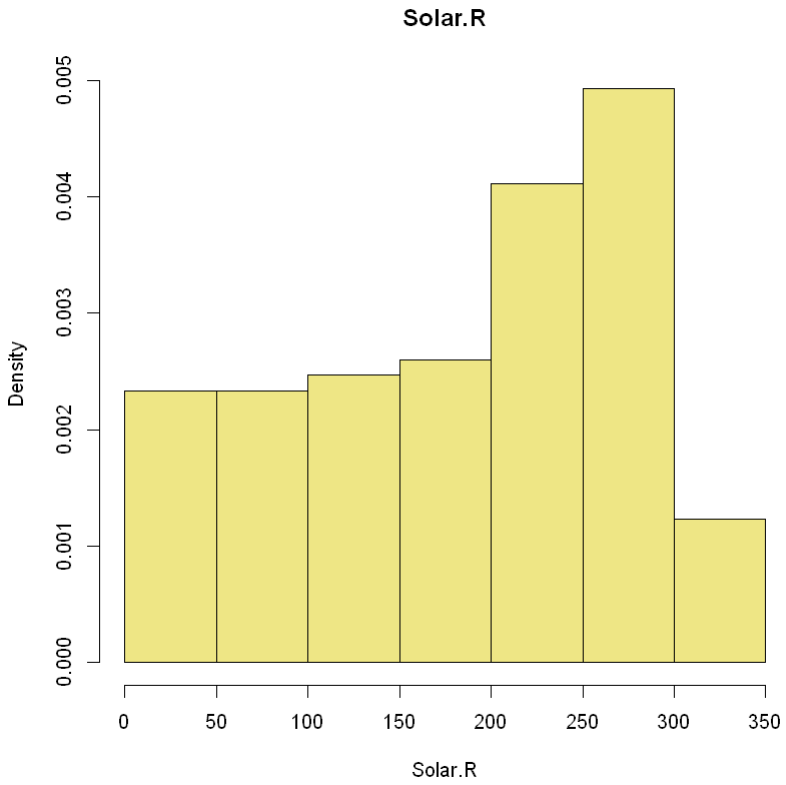
```
Warning message:
"Removed 37 rows containing missing values (`geom_point()`)."
```

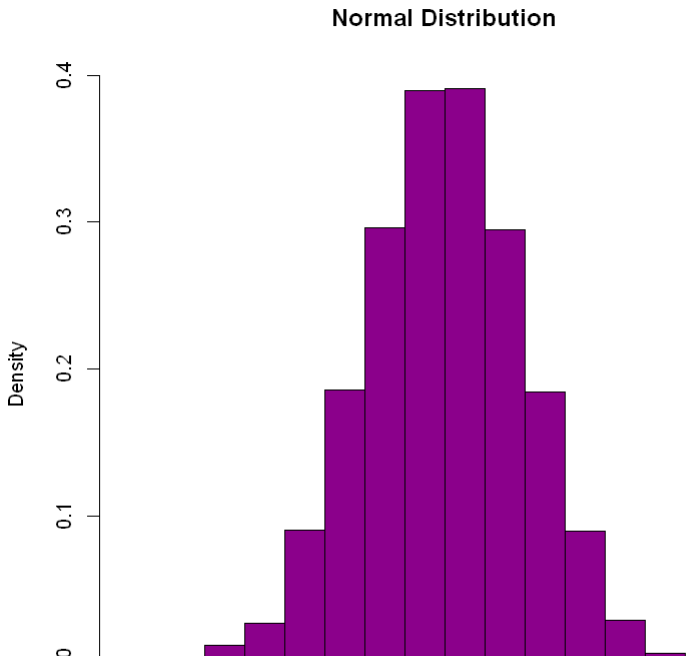


و) نمودار هیستوگرام برای باد و تابش خورشید رسم کنید، این دو نمودار را با هم و با نمودار توزیع نرمال مقایسه کنید و نتیجه‌ی خود را آن بیان کنید.

In [8]:
```
hist(airquality$Wind, xlab = "Wind", main="Wind", probability = T, col = "palegreen")
hist(airquality$Solar.R, xlab = "Solar.R", main="Solar.R", probability = T, col = "khaki2")

x <- rnorm(10000)
hist(x, main="Normal Distribution", probability = T, col = "darkmagenta")
# آنطور که نمودارها نشان می دهند، وضعیت میانگین باد به شکل قابل توجه
# و مناسبی از توزیع نرمال پیروی می کند
# در مورد تابش خورشید هم می توان گفت به گونه ای از این توزیع تبعیت می کند اما
# شاید آنچنان مشهود و دقیق نباشد.
# در مورد مقایسه دو نمودار باد و تابش خورشید هم، هر دو به گونه ای در رنج خاصی
# بیشترین فراوانی را دارند و هر چه از این رنج دورتر می شویم
# فراوانی کمتر خواهد شد.
```




Solar.R
Density
0.000
0.001
0.002
0.003
0.004
0.005
0
50
100
150
200
250
300
350
Solar.R



سوال دوم

در یک فرایند پواسون زمان بین دو اتفاق از یک توزیع نمایی پیروی می‌کند. اگر بخواهیم زمان اتفاق nام را پیش‌بینی کنیم از توزیع گاما استفاده می‌کنیم.

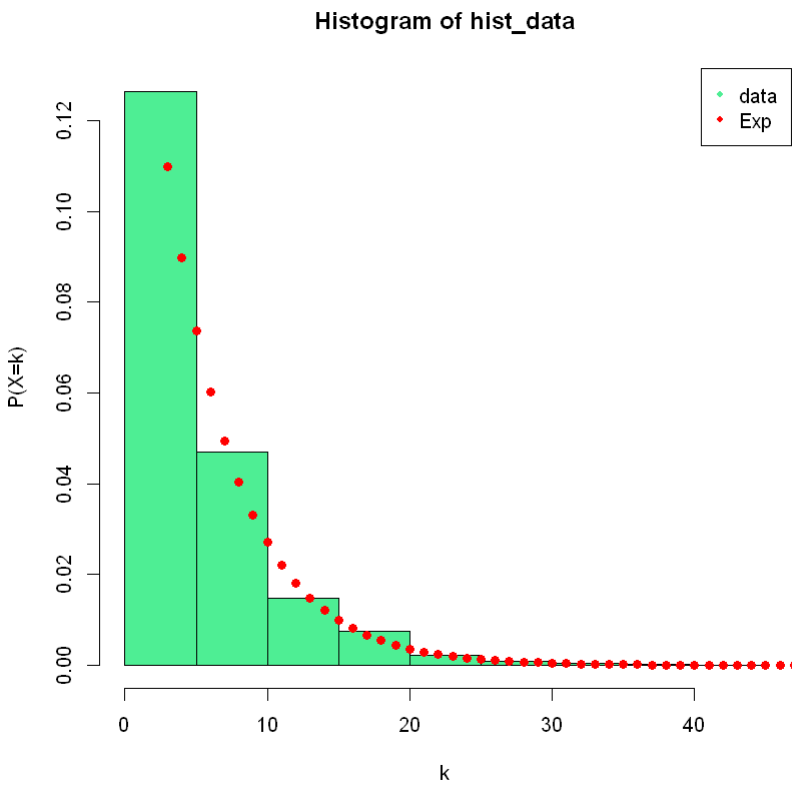
$$T_n = \sum_{i=1}^{n} X_i,\ X_i s\ are\ i.i.d\ and\ X_i \sim Exp(\lambda)$$

$$T_n \sim Gamma(n, \lambda)$$

حال فرض کنید فاصله بین ورود مشتریان به یک رستوران از توزیع نمایی پیروی می‌کند به طوری که در هر دقیقه پنج مشتری وارد می‌شوند.

الف) 2000 نمونه تصادفی برای زمان بین ورود مشتریان تولید کنید و نمودارهای هیستوگرام داده‌های تولیدشده و چگالی توزیع آن را در یک نمودار رسم کنید. نمودار هیستوگرام شما باید چگالی احتمال را برای هر نمونه را نشان دهد.

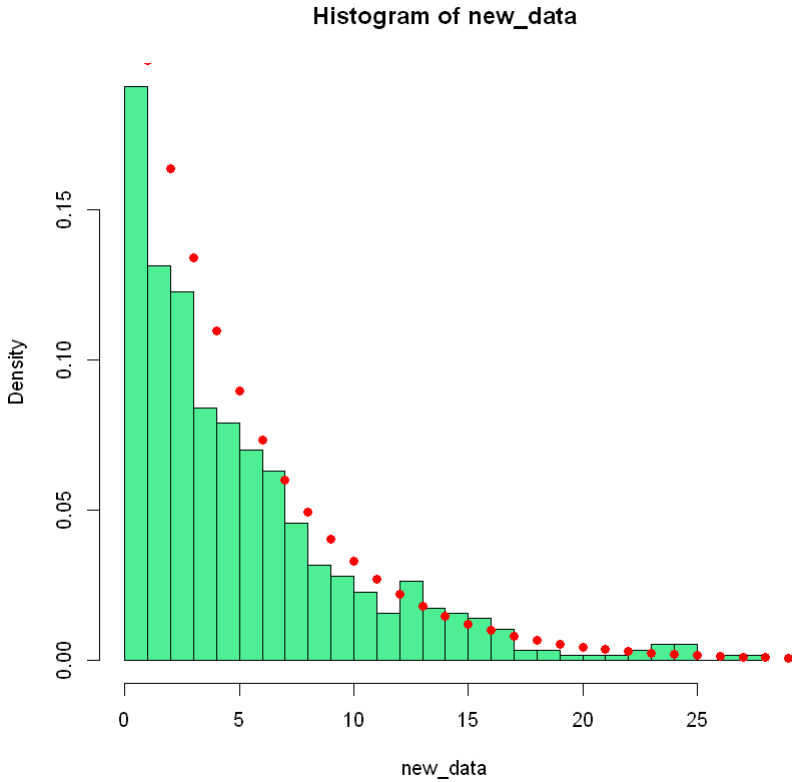
In [9]:
```
# first way:
# times <- rexp(2000, 1/5)
# plot <- ggplot(data.frame(times))+
#   geom_histogram(aes(x = times , y = after_stat(density)),colour = 3, fill = "white") +
#   geom_density(aes(x = times),lwd = 1, colour = 5,
#                alpha = 0.25)
# plot(plot)
# second way:
phe <- function(range, rate) {
hist_data <- rexp(n = 2000, rate)
hist(hist_data, probability = T, ylab = "P(X=k)", xlab = "k", col = 'seagreen2')

points_data <- dexp(x = range, rate)
points(range, points_data, col = 'red', pch = 16)
legend("topright", legend=c("data", "Exp"), pch = c(20, 20),
       col = c('seagreen2', 'red'))
}

phe(c(1:2000), 1/5)
```

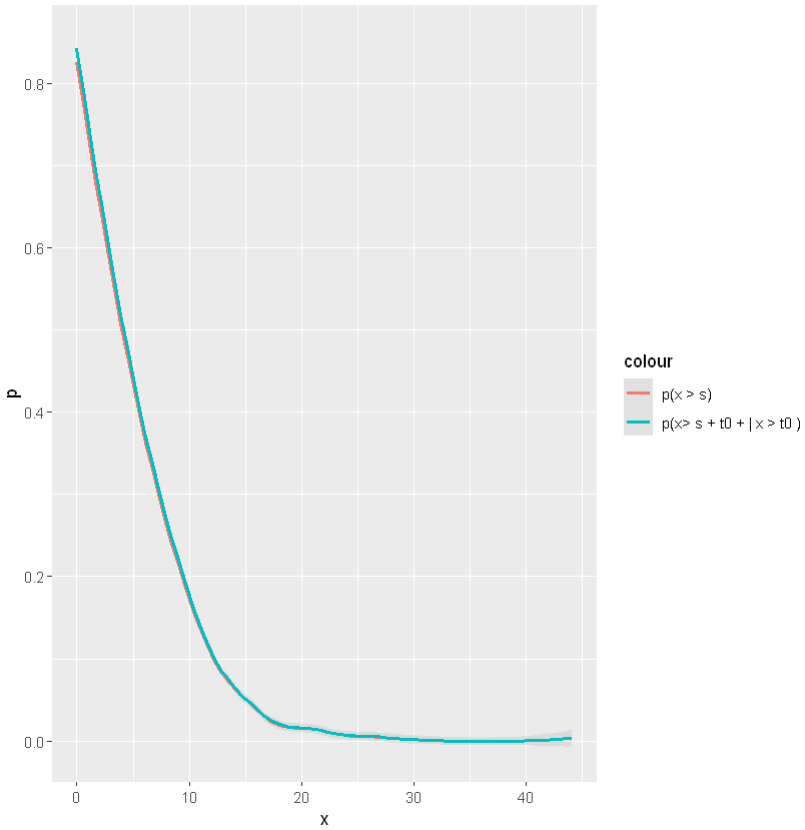


ب) بی‌حافظگی توزیع نمایی را با داده‌های تولید شده در قسمت قبل را با استفاده از داده‌های تولید شده در قسمت قبل و رسم نمودار به همان شکل نشان دهید.

In [3]:
```
data <- rexp(2000, 1/5)
new_data <- data[data > 6] - 6
hist(new_data, col='seagreen2',  probability = T,breaks = 25)
points(dexp(seq(0,30), 1/5), col='Red', pch=16)

# second solution:
times = rexp(2000, 0.2)
x <- seq(0,max(times),by = 0.1)
CDF = function(a,sample){
  return(length(sample[sample<=a])/2000)
}
t0 = 1
cdf_s = sapply(x,function(s) 1-CDF(s,times))
cdf_conditinal = sapply(x, function(s) (1-CDF(s+t0,times)) / (1-CDF(t0,times)))
d1 = data.frame(x=x , p = cdf_s)
d2 =  data.frame(x=x , p = cdf_conditinal)
p <- ggplot() +
  # blue plot = p(x > s)
  geom_smooth(data=d1, aes(x=x, y=p , colour = "p(x > s)" ) ,
              size=1,alpha = 0.1 ) +
  # red plot =  p(x> s + t0 + | x > t0 )
  geom_smooth(data=d2, aes(x=x, y=p , colour = "p(x> s + t0 + | x > t0 )" ),
              size=1 , alpha = 0.1)
plot(p)
```

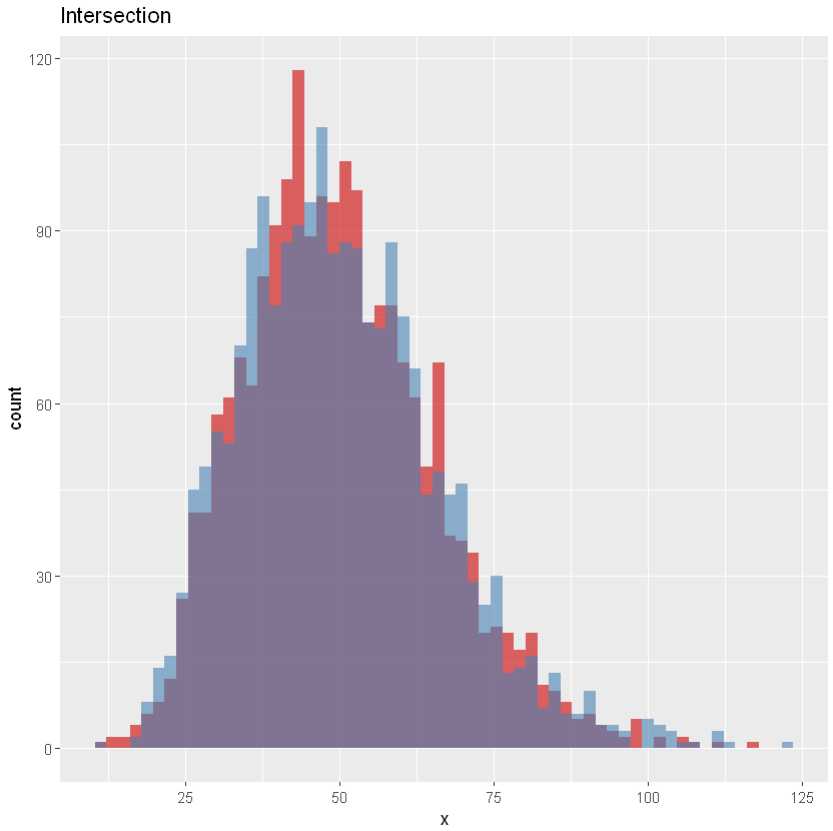
```
Warning message:
"Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead."
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```


Histogram of new_data
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
25
new_data



ج) با تولید n=10 متغیرتصادفی نمایی، ویژگی گفته شده در صورت سوال را بررسی کنید. برای این‌کار می‌توانید برای شبیه‌سازی هر متغیرتصادفی مانند قسمت الف عمل کنید.

In [18]:
```
n <- 2000
lambda <- 1/5
num <- 10

collection <- rep(0, n)
for (i in 1:num) {
  collection <- collection + rexp(n, lambda)
}

data_exp <- data.frame(x=collection)
data_gamma <- data.frame(x=rgamma(n, num, lambda))

# ggplot() +
#   ggtitle("Exponential") +
#   geom_histogram(data=data_exp, aes(x), fill="red3", bins = 60, alpha = .6)

# ggplot() +
#   ggtitle("Gamma") +
#   geom_histogram(data=data_gamma, aes(x), fill="steelblue", bins = 60, alpha = .6)

ggplot() +
  ggtitle("Intersection") +
  geom_histogram(data=data_exp, aes(x), fill="red3", bins = 60, alpha = .6) +
  geom_histogram(data=data_gamma, aes(x), fill="steelblue", bins = 60, alpha = .6)
```



سوال سوم

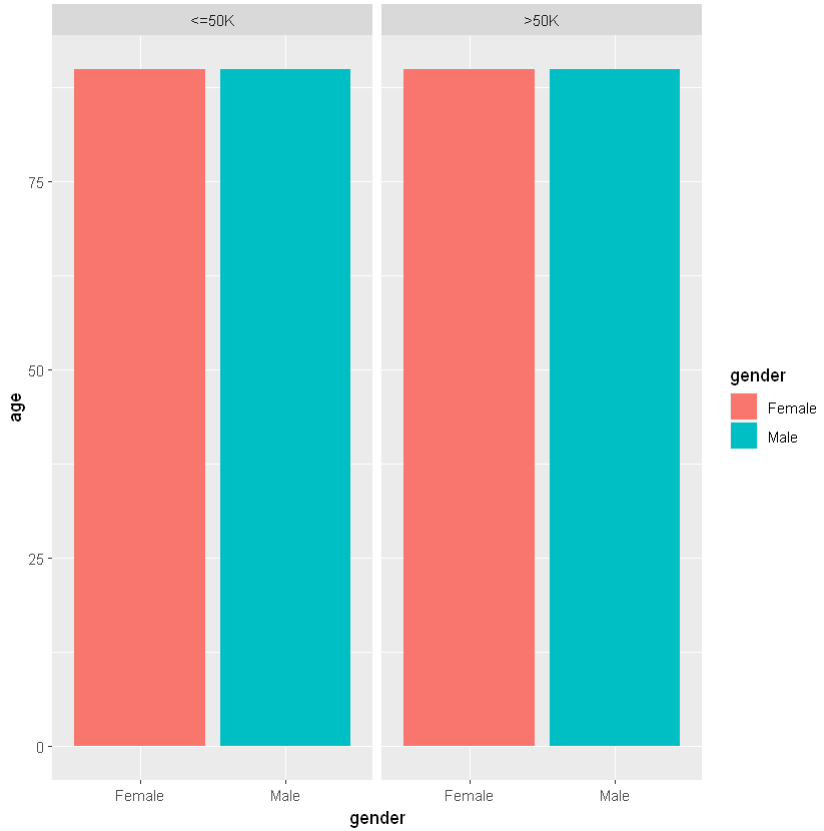
دیتاست مربوط به درآمد در پیوست موجود است.
الف) به وسیلهٔ دستور read.csv اطلاعات دیتاست را بخوانید. (دیتاست adult.csv)

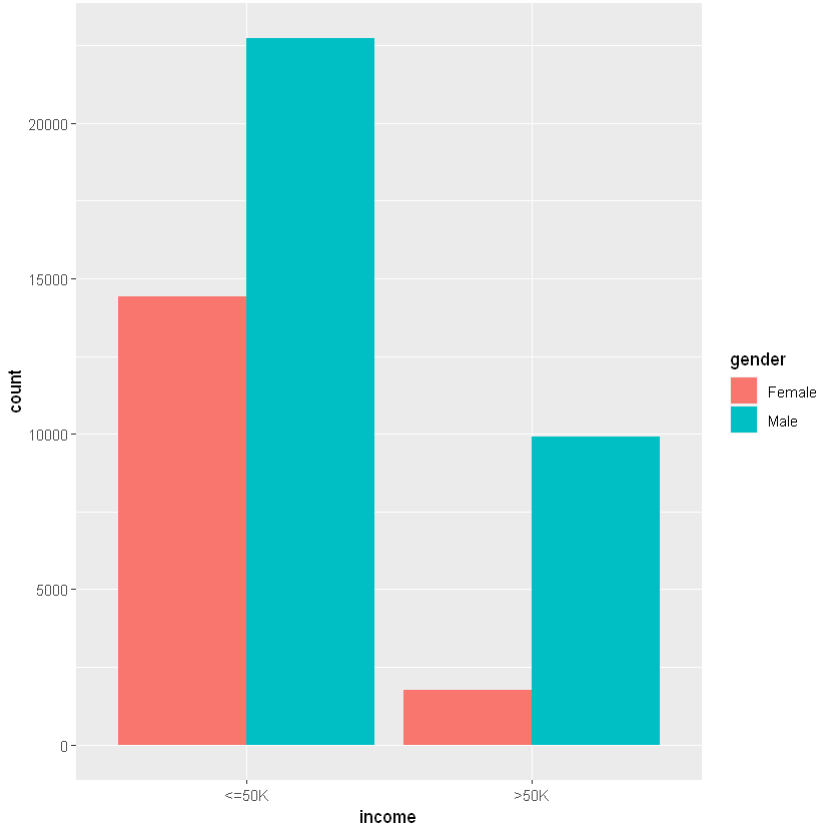
In [4]:
```
data_income <- read.csv("adult.csv")
```

نمودار مربوط به درآمد را برحسب سن/جنسیت به کمک نمودارهای زیر بکشید.

ب) نمودار Bar Plot (درآمد برحسب جنسیت و سن)

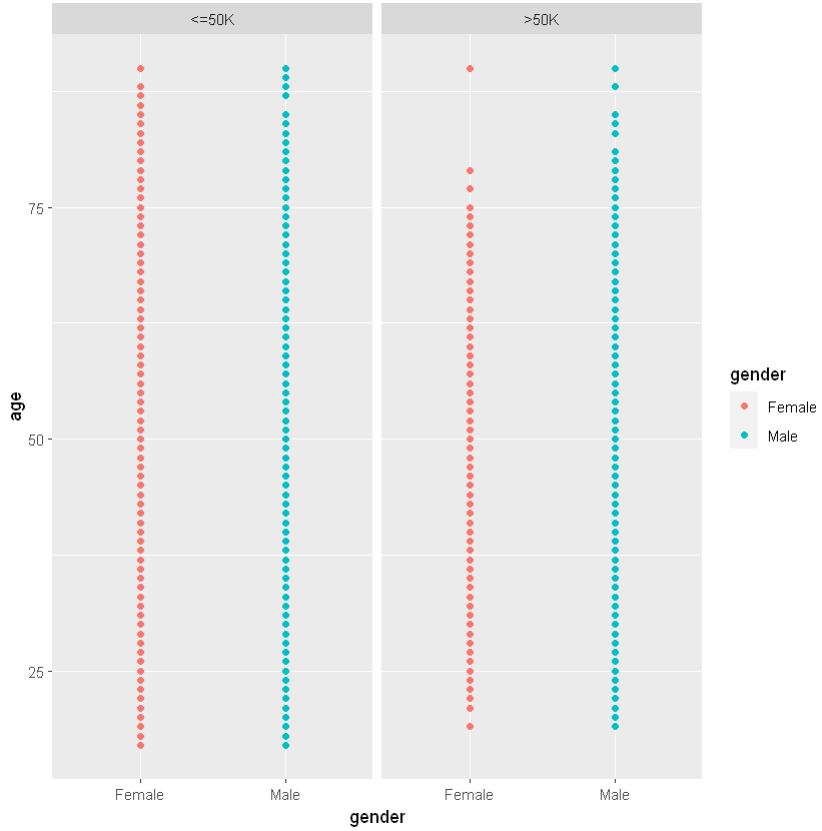
In [20]:
```
ggplot(data=data_income, aes(x=gender, y=age, fill=gender)) +
  geom_bar(stat="identity", position=position_dodge())  +
  facet_grid(~income)
# solution for counting:
ggplot(data=data_income, aes(x=income, fill=gender)) +
  geom_bar(position=position_dodge())
```





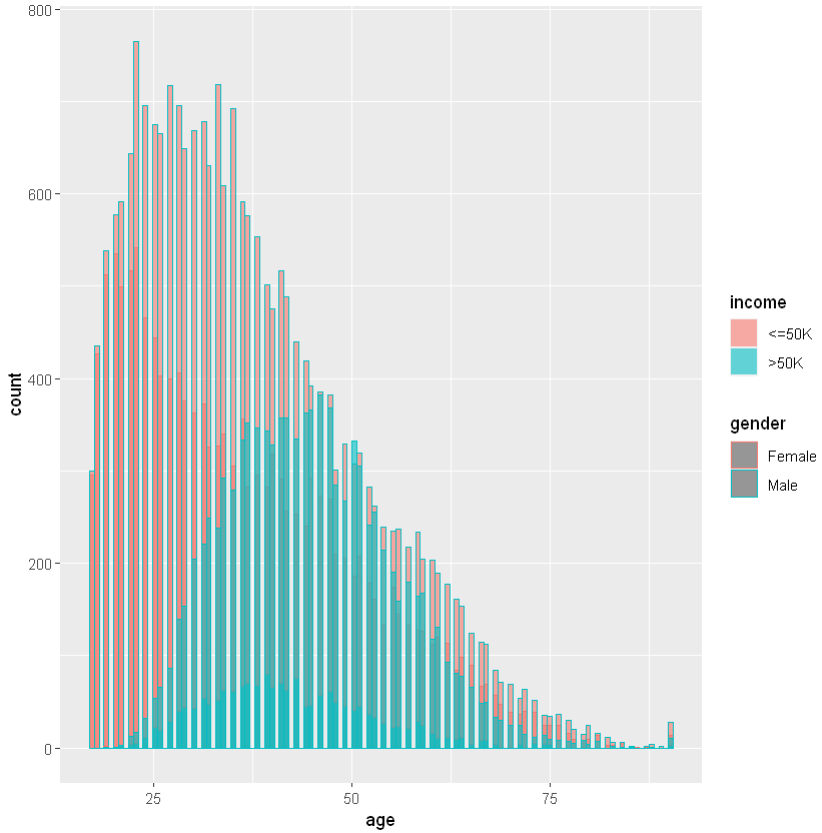
ج) نمودار Scatter Plot

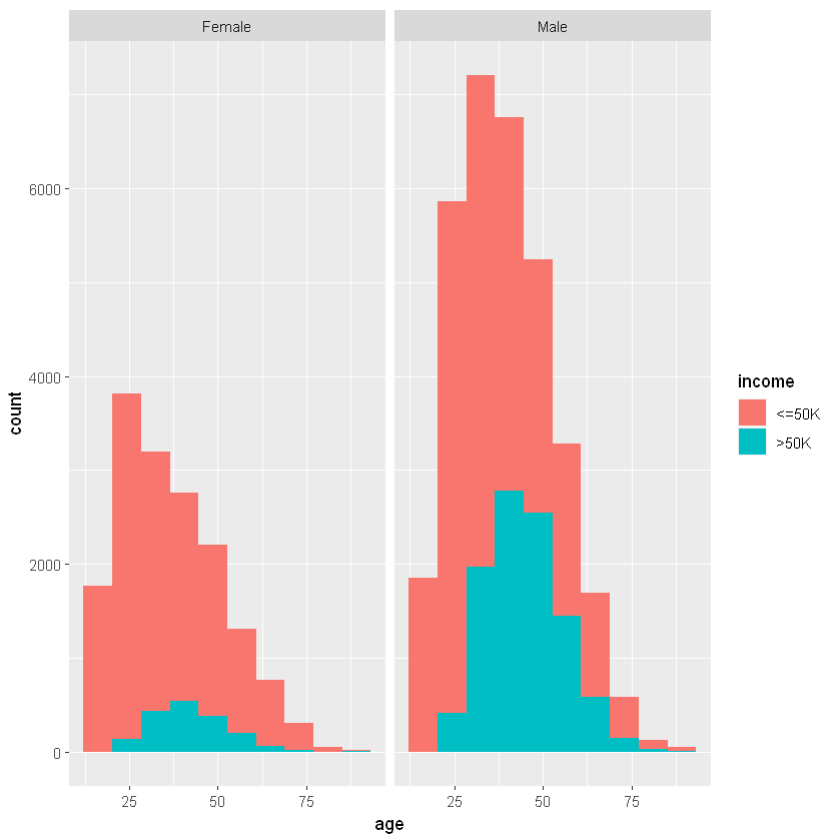
In [21]:
```
ggplot(data_income, aes(x=gender, y=age, color = gender)) +
  geom_point() +
  facet_grid(~income)
```



د) نمودار Histogram (مربوط به سن که اطلاعات زن و مرد در یک نمودار به طور تلفیقی باشد)

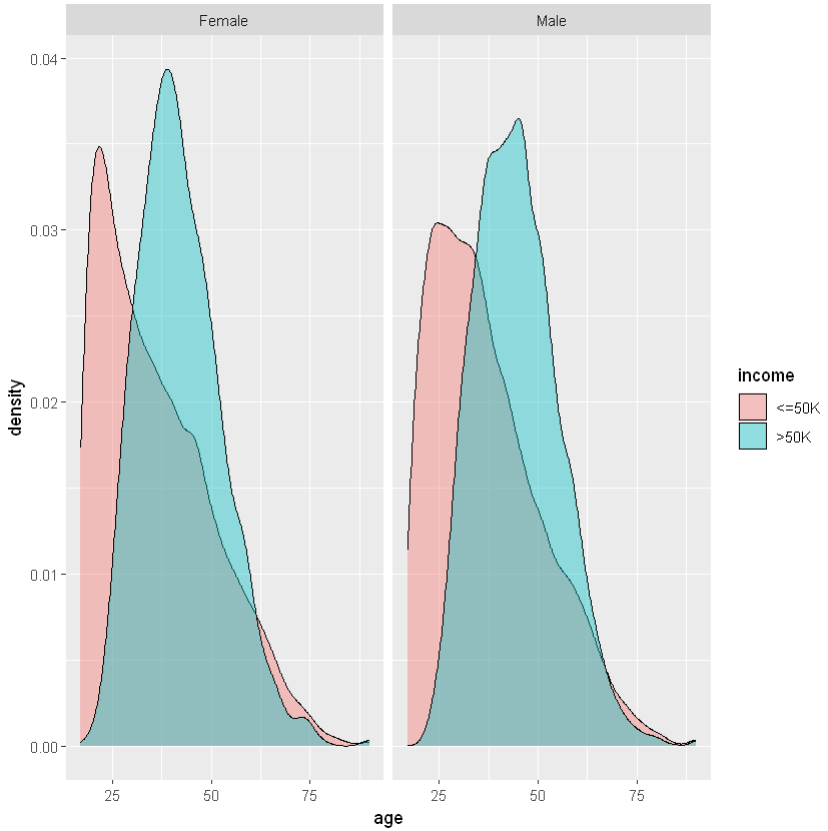
In [22]:
```
# solution 1:
ggplot(data_income, aes(x=age, fill=income, color=gender)) +
  geom_histogram(position="identity", bins=120, alpha = .6)
# solution 2:
ggplot(data_income,aes(x=age, fill=income))+
  geom_histogram(bins = 10)+
  facet_grid(~gender)+
  theme_get()
# optional solution:
# ggplot(data_income, aes(x=age, fill=gender, color=gender)) +
#   geom_histogram(position="identity", bins=120, alpha = .6)
# ggplot(data_income, aes(x=age, fill=income, color=income)) +
#   geom_histogram(position="identity", bins=120, alpha = .5)
```

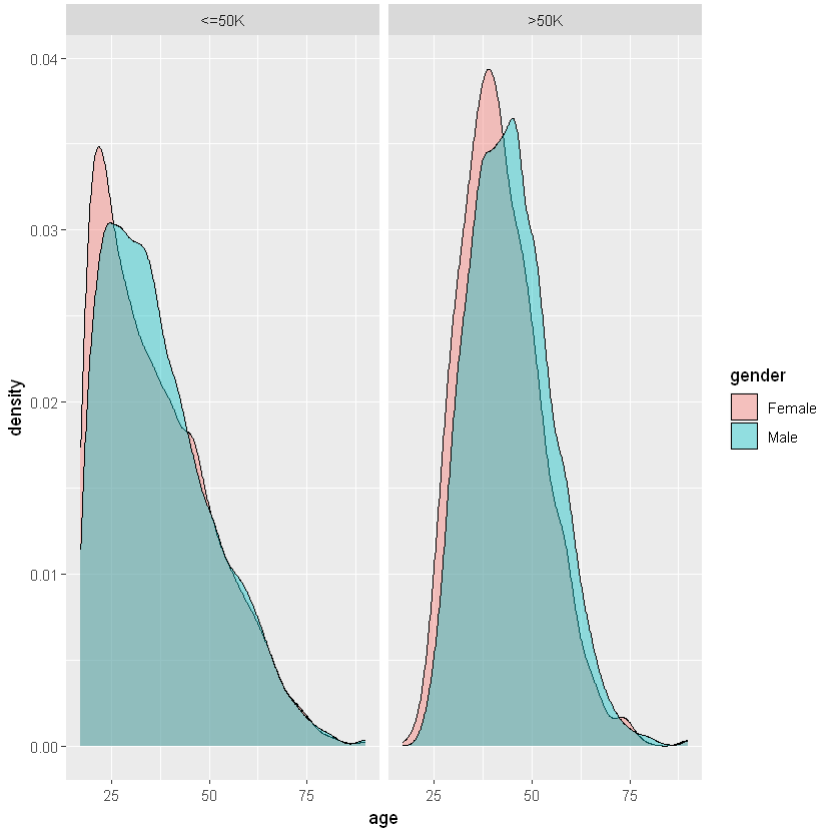




ه) نمودار Density Plot (مربوط به سن که اطلاعات زن و مرد در یک نمودار به طور تلفیقی باشد)

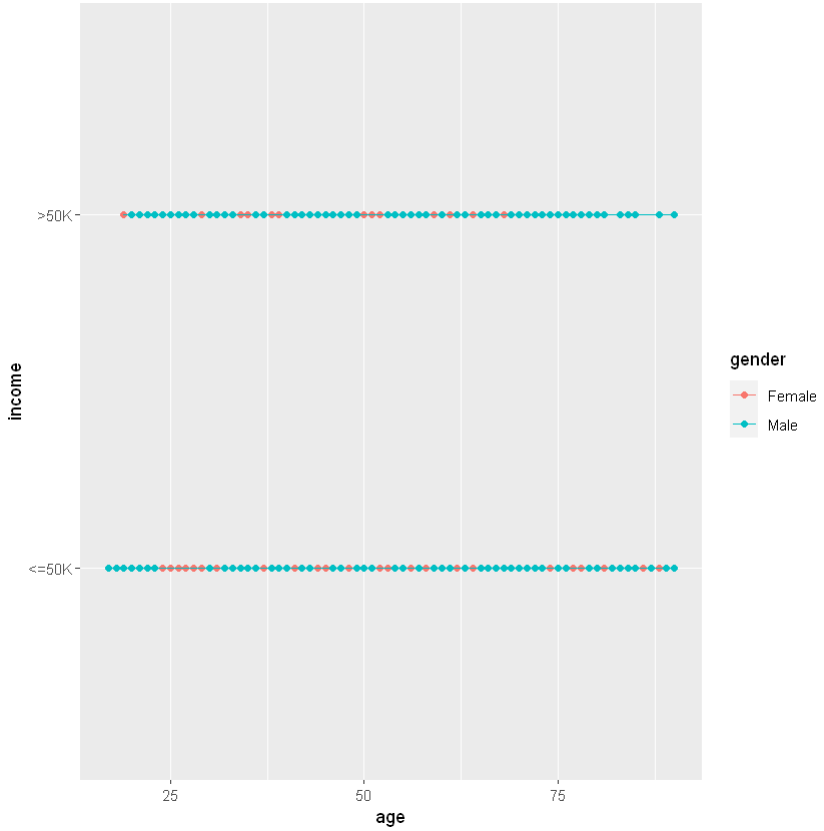
In [23]:
```
ggplot(data_income, aes(x=age, fill=income)) +
  geom_density(alpha = .4) +
  facet_grid(~gender)
# or
ggplot(data_income, aes(x=age, fill=gender)) +
  geom_density(alpha = .4) +
  facet_grid(~income)
```





Trend Plot نمودار (و

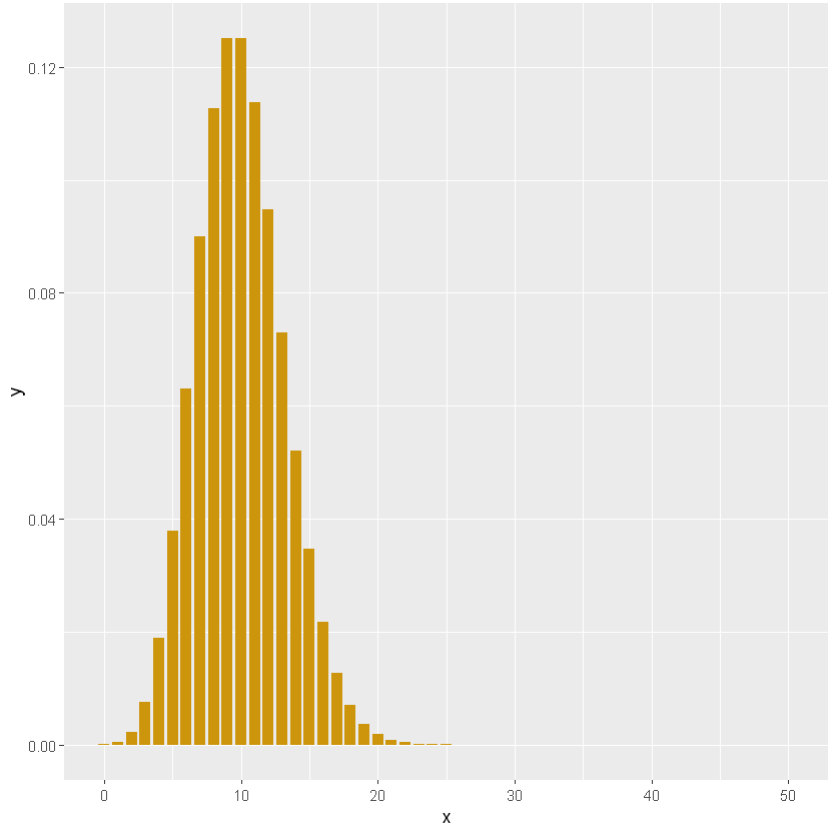
In [5]:
```
ggplot(data_income, aes(age, income, color = gender)) +
  geom_point(position="identity") +
  geom_line()
```



سوال چهارم

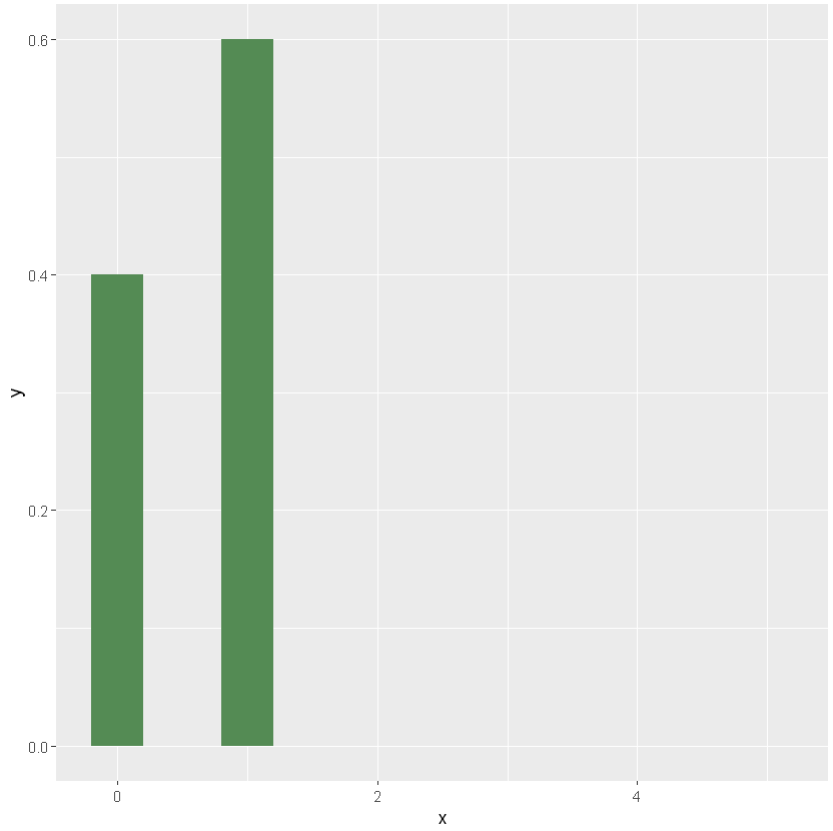
نمودار توزیع های احتمالی زیر را با توجه به به پارامتر های داده شده رسم کنید.
الف) نمودار توزیع پوآسون (در بازهٔ 0 تا 50 با پارامتر 10)

In [25]:
```
x <- seq(0,50, by=1)
df <- data.frame(x=x, y=dpois(x, 10))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = .8, fill = "darkgoldenrod3",
           position=position_dodge(width = .5)) + theme_get()
```



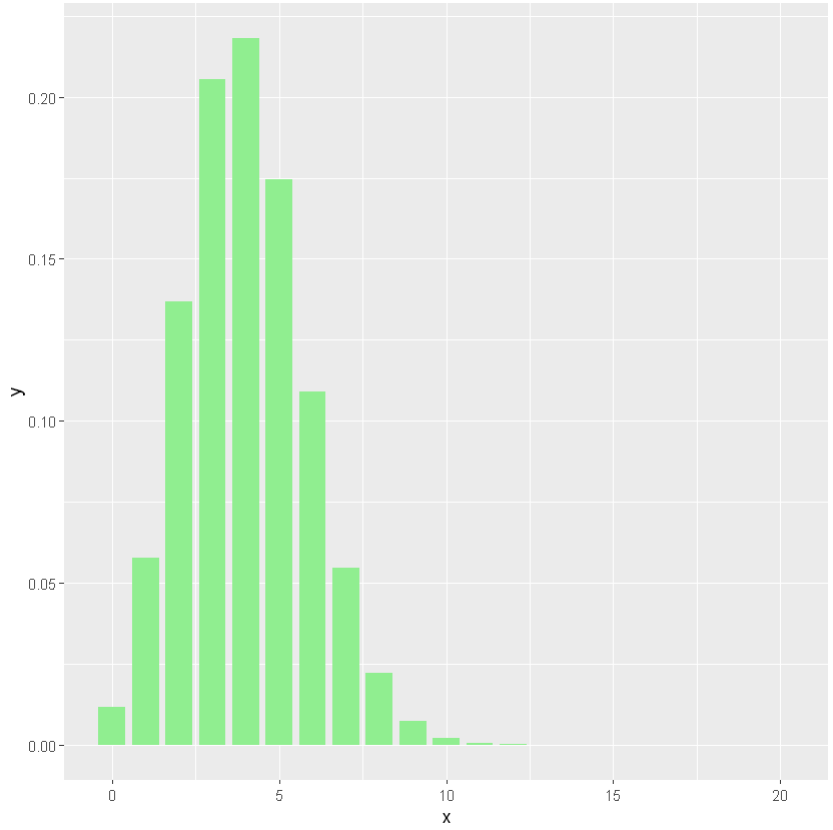
ب) نمودار توزیع برنولی (در بازهٔ 0 تا 5 با پارامتر 0.6)

In [26]:
```
df <- data.frame(x=c(0,1,2,3,4,5), y=c(.4, .6, 0, 0, 0, 0))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = .4, fill = "palegreen4",
           position=position_dodge(width = .5)) + theme_get()
```



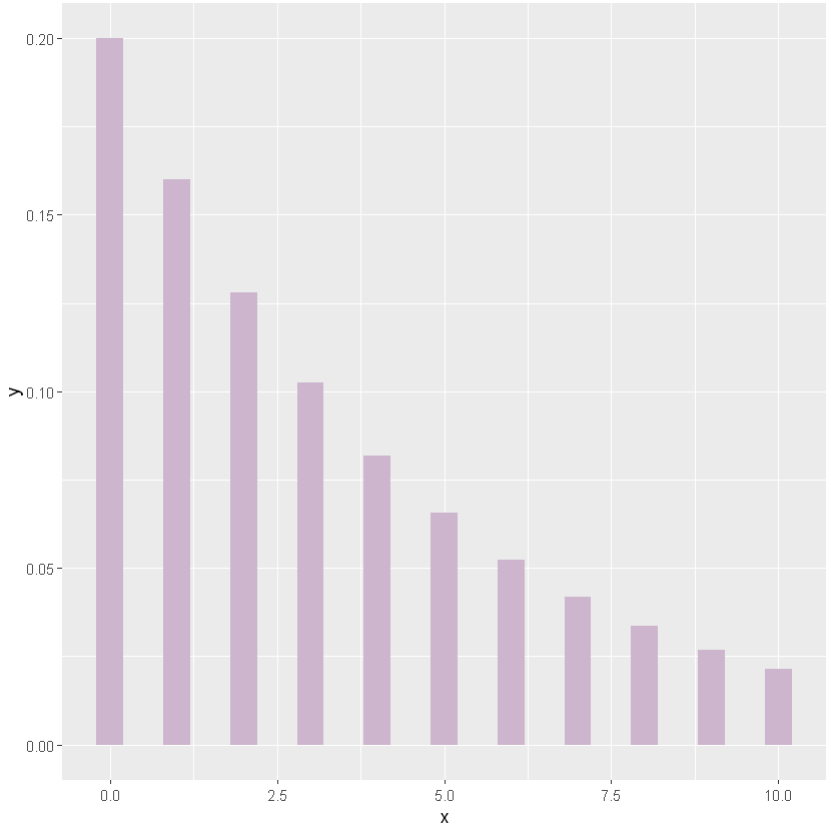
ج) نمودار توزیع دو جمله ای (در بازهٔ 0 تا 20 با پارامتر 0.2)

In [27]:
```
x <- seq(0,20, by=1)
df <- data.frame(x=x, y=dbinom(x, 20, 0.2))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = .8, fill = "palegreen2",
           position=position_dodge(width = .5)) + theme_get()
```



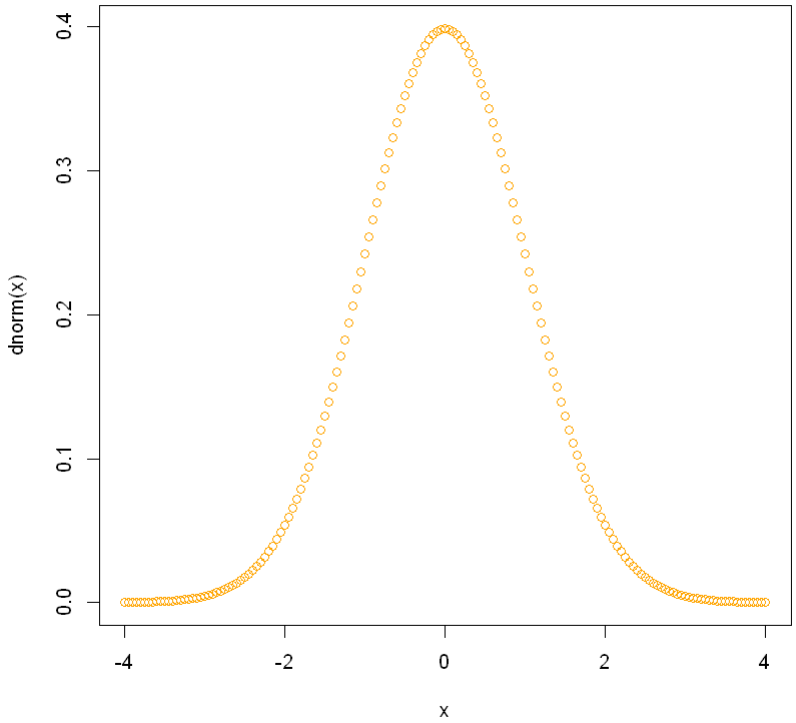
د) نمودار توزیع هندسی (در بازهٔ 0 تا 10 با پارامترهای $p = 0.2$ و $n = 3$)

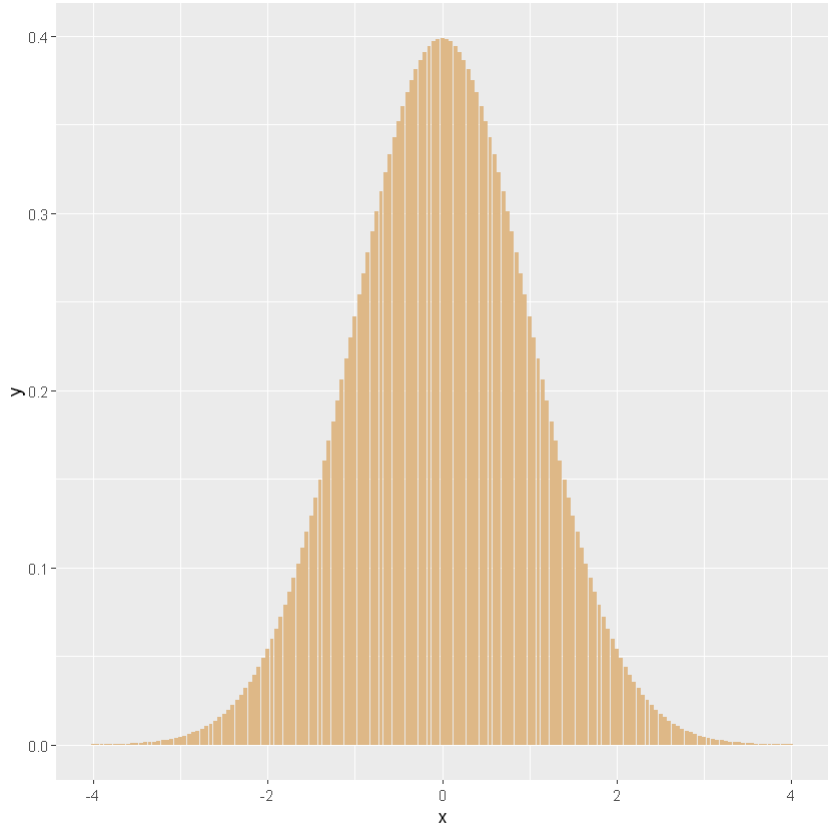
In [28]:
```
x <- seq(0,10, by=1)
df <- data.frame(x=x, y=dgeom(x, .2))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = 0.4, fill="thistle3",
           position=position_dodge(width = 0.5)) + theme_get()
```



ه) نمودار توزیع نرمال (در بازهٔ 4- تا 4)

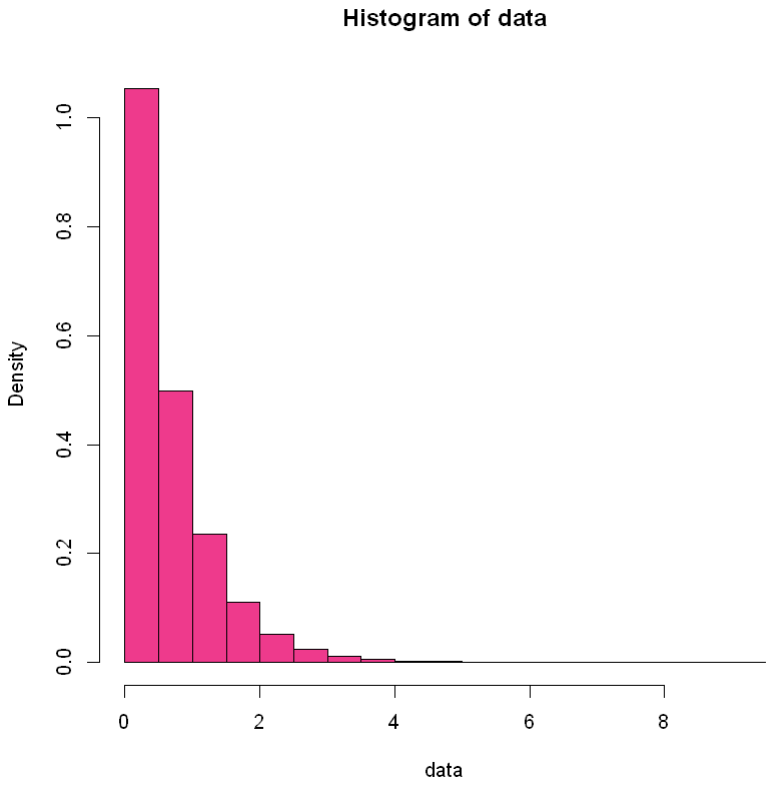
In [29]:
```
x = seq(-4, 4, by = 0.05)
plot(x, dnorm(x), col="orange")
df <- data.frame(x=x, y=dnorm(x))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity", fill = "burlywood") + theme_get()
```

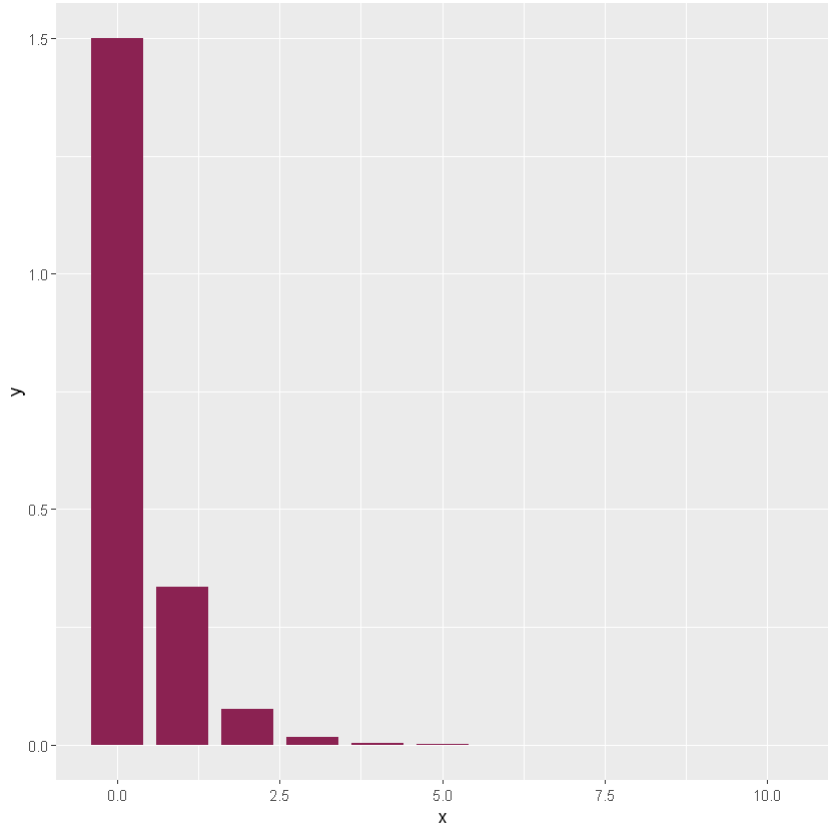




و) نمودار توزیع نمایی (در بازهٔ 0 تا 10 با پارامتر 1.5)

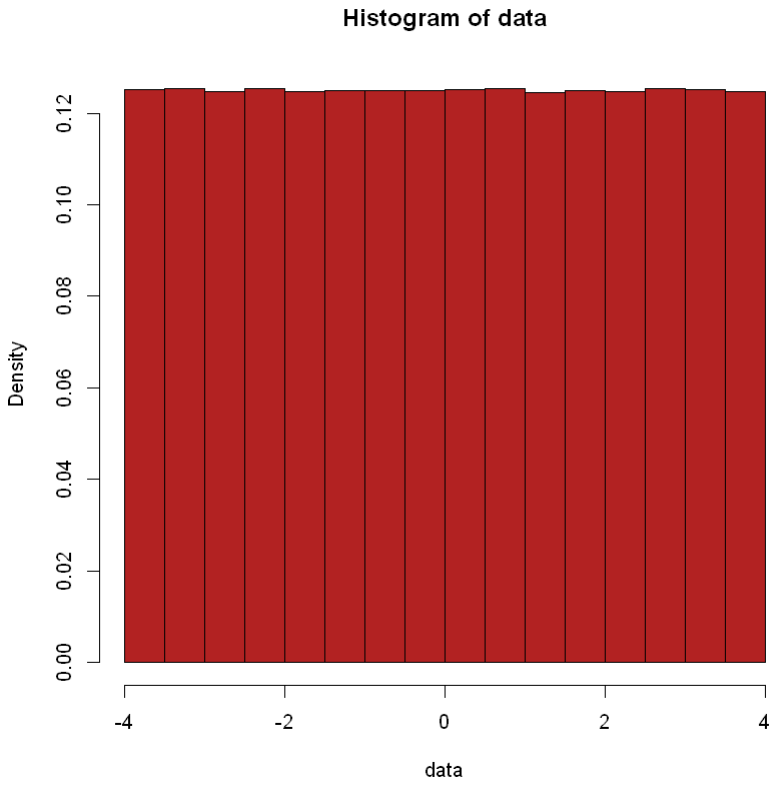
In [30]:
```
data <- rexp(1000000, rate = 1.5)
data <- data[data<=10]
hist(data, probability = T, col = "violetred2")
x <- seq(0,10, by=1)
df <- data.frame(x=x, y=dexp(x, 1.5))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = .8, fill = "violetred4",
           position=position_dodge(width = .5)) + theme_get()
```

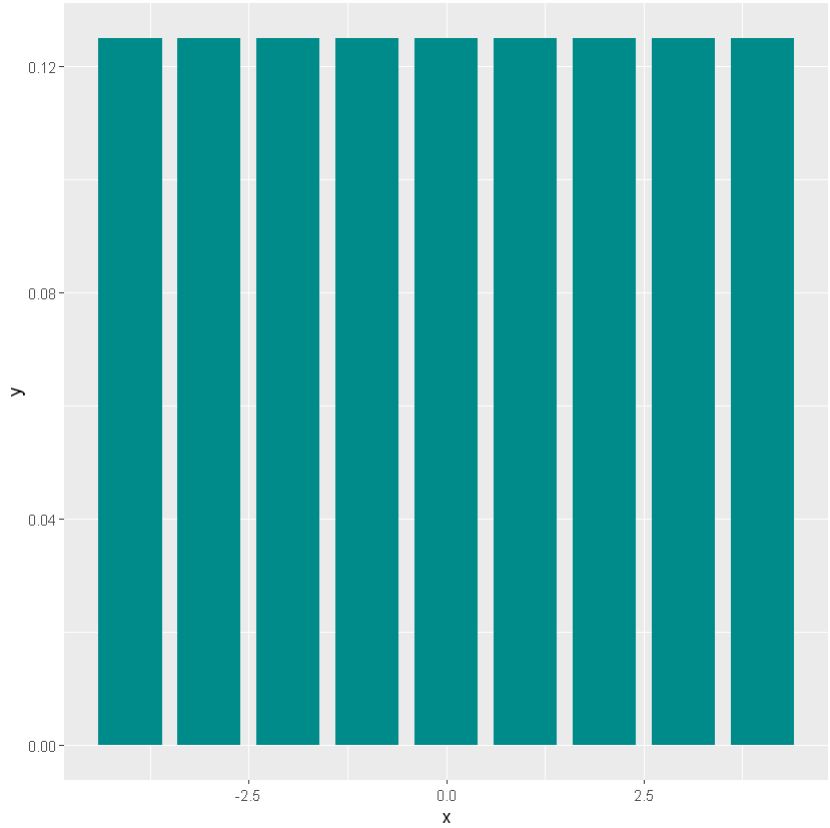




ز) نمودار توزیع یکنواخت (در بازهٔ 4- تا 4)

In [31]:
```
data <- runif(2000000, -4, 4)
hist(data, probability = T, col="firebrick")
x <- seq(-4,4, by=1)
df <- data.frame(x=x, y=dunif(x, -4, 4))
ggplot(data = df, aes(x = x, y = y)) +
  geom_bar(stat = "identity",  width = .8, fill = "cyan4",
           position=position_dodge(width = .2)) + theme_get()
```




0.12
0.08
0.04
0.00
y
-2.5
0.0
2.5
x