به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

استنباط آماری

فاز اول پروژه

نام و نام خانوادگی	مهسا تاجیک
شماره دانشجویی	810198126
تاریخ ارسال گزارش	99/2/12

QUESTION 0

a) برای شرکتهای بیمه یک کار بسیار مهم ، تعیین حق بیمه ایده آل برای هر فرد بیمه شده با توجه به چند متغیر مستقل مانند سن ، شاخص توده بدنی و جنسیت است. در مجموعه داده ی insurance

اطلاعات مربوط به حق بیمه افراد و این ویژگی ها برای 1338 نفر آمده است. مطالعه ی این مجموعه داده از این جهت می تواند جالب و مورد توجه باشد که می توانیم وجود روابط مختلفی بین متغیرهای آن بررسی کنیم . به طور مثال وجود رابطه بین سن افراد و شاخص توده بدنی و یا جنسیت هر فرد و میزان حق بیمه ی فرد و ...

b) این مجموعه داده شامل 7 متغیر(ویژگی) 1338 مورد می باشد. که این ویژگی ها عبارتند از : سن ، جنسیت ، شاخص توده بدنی ، تعداد فرزند ، سیگاری بودن ، محل زندگی و حق بیمه است.

c) خیر هیچ مقدار از دست رفته ای در این مجموعه داده وجود ندارد. این موضوع را بعد از لود کردن مجموعه داده در برنامه R با کد زیر می توانیم بررسی کنیم :

```
> insurance <- read.csv("insurance.csv")
> any(is.na(insurance))
[1] FALSE
>
```

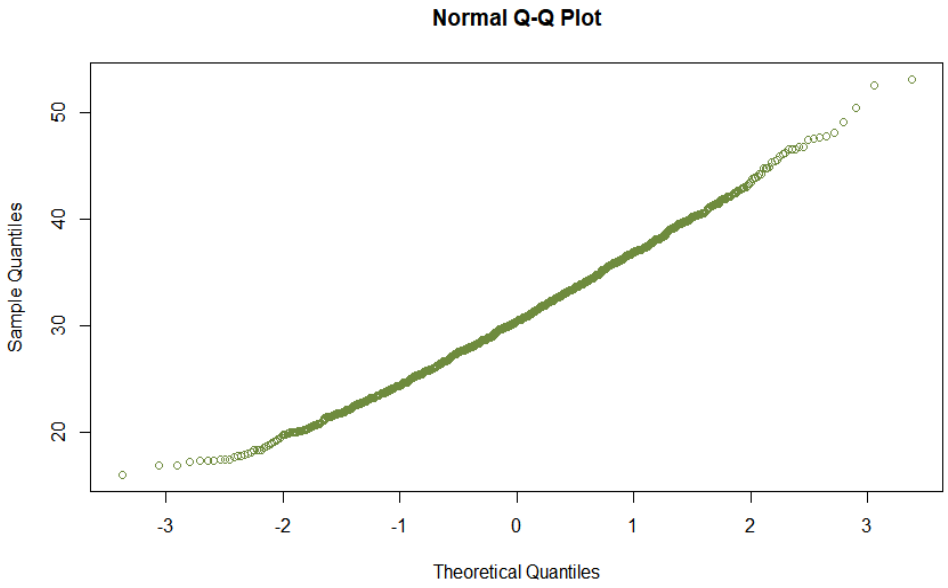
d) متغیرهای smoker و age مهم تر بوده و میتوان اطلاعات مهمی از آن ها بدست آورد زیرا مصرف سیگار روی سلامتی تاثیرگذار بوده و همچنین با افزایش سن احتمال ابتدا به بعضی بیماری ها افزایش می یابد بنابراین این دو ویژگی می توانند با میزان حق بیمه در ارتباط باشند.

QUESTION 1

a) در این قسمت نمودار Q-Q برای یکی از متغیرهای numerical خواسته شده است.

نمودار Q-Q یک ابزار گرافیکی است که به ما کمک می کند ارزیابی کنیم که آیا مجموعه ای از داده ها از برخی توزیع های نظری مانند یک توزیع نرمال آمده یا خیر. اگر دو توزیع مقایسه شده مشابه باشند، نقاط روی نمودار چندک-چندک تقریباً روی خط y = x قرار خواهند گرفت . اگر توزیع‌ها رابطه خطی داشته باشند، نقاط نمودار، تقریباً روی یک خط راست قرار می‌گیرند، ولی این خط الزاماً خط y = x نمی‌باشد.

متغیرهای age , bmi , children , charges همگی numerical هستند و ما متغیر bmi را برای رسم نمودار استفاده می کنیم.



شکل 1-1

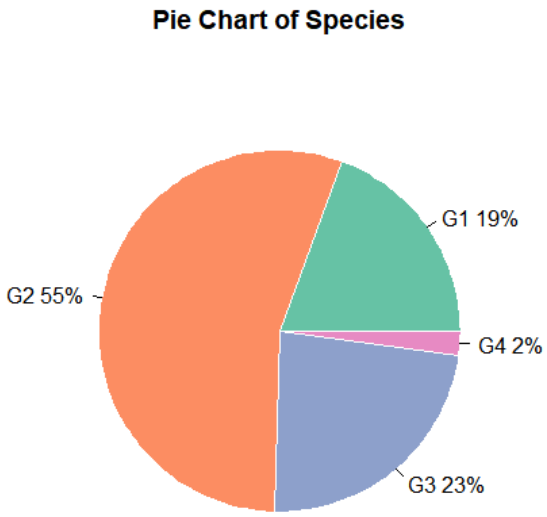
همانطور که در شکل 1-1 مشاهده می کنیم نمودار بصورت یک خط است بنابراین توزیع bmi با توزیع نرمال رابطه خطی دارد.کد مربوط به این قسمت را در زیر می بینیم:

```
qqnorm(insurance$bmi,col = "darkolivegreen4")
```

b) برای اینکه بتوانیم داده های متغیر عددی bmi را 4 گروه تقسیم کنیم ، داده های بیشینه و کمینه را پیدا کرده و فاصله ی بین آن ها را 4 دسته می کنیم که می توانیم مقادیر شروع و پایان بازه ها را در شکل 1-2 ببینیم. سپس تعدادی که در هر دسته قرار میگیرد را در یک حلقه شمارش می کنیم و نمودار پای چارت را برای آن رسم می کنیم که در شکل 1-3 قابل مشاهده است و می بینیم بیشتر افراد دارای شاخص توده بدنی در بازه ی 25 تا 34 (G2) هستند.

```
> max(insurance$bmi)
[1] 53.13
> min(insurance$bmi)
[1] 15.96
> (max(insurance$bmi)-min(insurance$bmi)) / 4
[1] 9.2925
> g1 <- 0
> max(insurance$bmi)
[1] 53.13
> min(insurance$bmi)
[1] 15.96
> (max(insurance$bmi)-min(insurance$bmi)) / 4
[1] 9.2925
> min(insurance$bmi)+(max(insurance$bmi)-min(insurance$bmi)) / 4
[1] 25.2525
> min(insurance$bmi)+2*(max(insurance$bmi)-min(insurance$bmi)) / 4
[1] 34.545
> min(insurance$bmi)+3*(max(insurance$bmi)-min(insurance$bmi)) / 4
[1] 43.8375
>
```

شکل 2-1

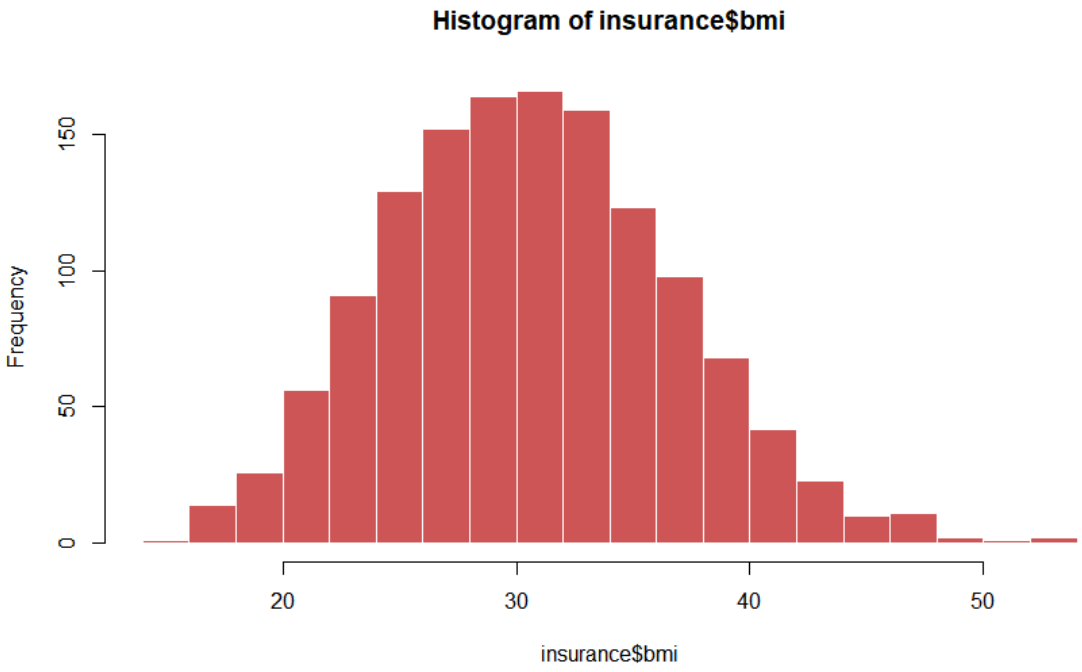


شکل 3-1

کد مربوط به این قسمت در زیر قابل مشاهده است:

```
library(RColorBrewer)
max(insurance$bmi)
min(insurance$bmi)
(max(insurance$bmi)-min(insurance$bmi)) / 4
min(insurance$bmi)+(max(insurance$bmi)-min(insurance$bmi)) / 4
min(insurance$bmi)+2*(max(insurance$bmi)-min(insurance$bmi)) / 4
min(insurance$bmi)+3*(max(insurance$bmi)-min(insurance$bmi)) / 4
g1 <- 0
g2 <- 0
g3 <- 0
g4 <- 0
for(i in 1:1338){
  if(insurance$bmi[i] <= 25.2525)
    g1 <- g1+1
  else if((insurance$bmi[i] > 25.2525) & (insurance$bmi[i] <= 34.545))
    g2 <- g2+1
  else if((insurance$bmi[i] > 34.545) & (insurance$bmi[i] <= 43.8775))
    g3 <- g3+1
  else
    g4 <- g4+1
}
myPalette <- brewer.pal(4, "Set2")
slices <- c(g1,g2,g3,g4)
lbls = c("G1","G2","G3","G4")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="")

pie(slices, labels = lbls,
    main="Pie Chart of Species\n" , border="white", col=myPalette)
```

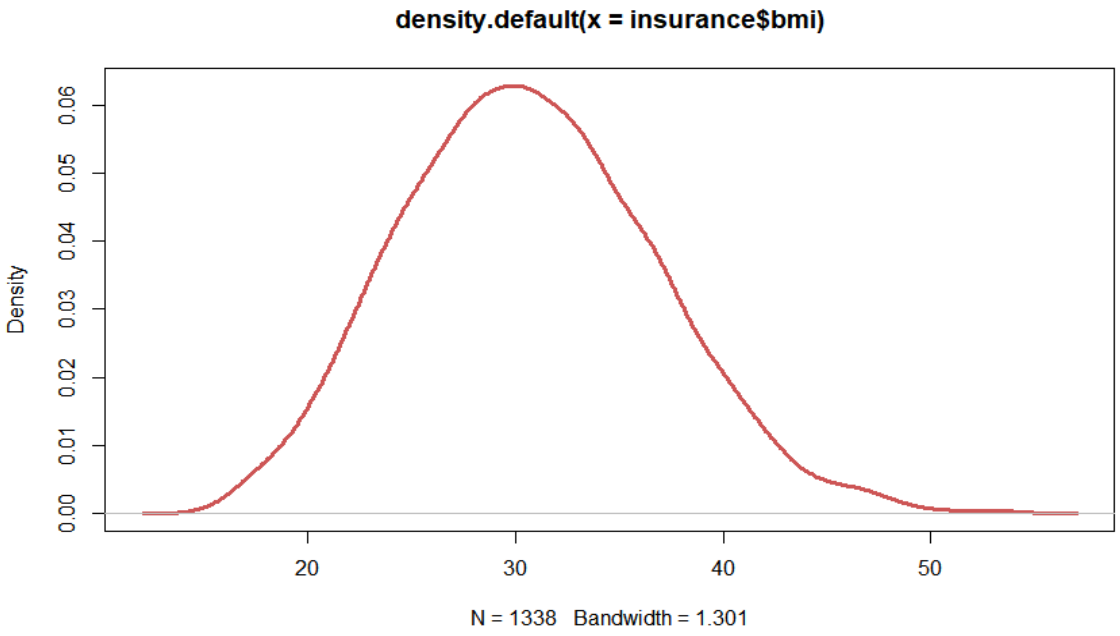
c) هیستوگرام مربوط به متغیر bmi در شکل 4-1 آورده شده و همانطور که میبینیم توزیعی شبیه نرمال دارد و در قسمت a هم دیدیم که نمودار Q-Q برای این متغیر خطی شد که صحت این موضوع را نشان می دهد.



شکل 4-1

```
hist(insurance$bmi, col = "indianred3" , border = "white" , breaks = 20)
```

d) نمودار density برای متغیر bmi در شکل 5-1 آمده است :



شکل 1-5

```
density = density(insurance$bmi)
plot(density, col = "indianred3" , lwd = 3)
```

e) مقدار skewness با کد زیر محاسبه شده و نتیجه هم قابل مشاهده است :

```
> library(e1071)
> skewness(insurance$bmi)
[1] 0.2834106
>
```

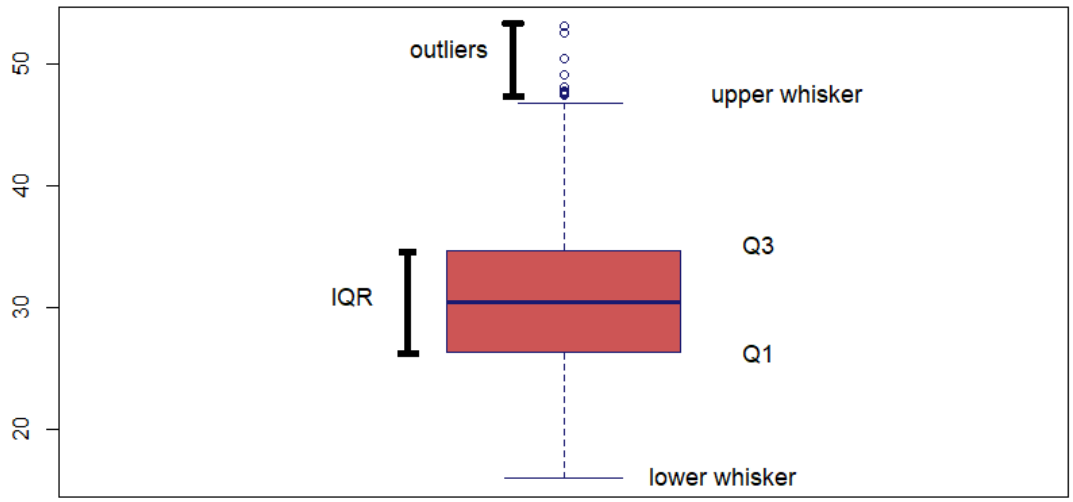
f) مقادیر میانگین ، واریانس و انحراف معیار برای شاخص توده بدنی در شکل 1-6 آورده شده است.

```
> mean(insurance$bmi)
[1] 30.6634
> var(insurance$bmi)
[1] 37.18788
> sd(insurance$bmi)
[1] 6.098187
```

شکل 1-6

g) نمودار باکس پلات برای شاخص توده بدنی در شکل 1-7 آمده است و مقادیر IQR ، Q1 ، Q3 و Whisker ها در نرم افزار R محاسبه شده و نتایج در شکل 1-8 مشاهده می شود و همچنین روی نمودار باکس پلات این مقادیر نشان داده شده است.

```
boxplot(insurance$bmi,boxwex=0.5, border=c("midnightblue"), col=c("indianred3"), xlab = "insurance$bmi")
Q1 <- max(min(insurance$bmi), quantile(insurance$bmi, c(0.25)))
lower_whisker <- Q1 - 1.5 * IQR(insurance$bmi)
Q3 <- min(max(insurance$bmi), quantile(insurance$bmi, c(0.75)))
upper_whisker <- Q3 + 1.5 * IQR(insurance$bmi)
IQR(insurance$bmi)
Q1
Q3
lower_whisker
upper_whisker
```



شکل1-7

```
> IQR(insurance$bmi)
[1] 8.3975
> Q1
[1] 26.29625
> Q3
[1] 34.69375
> lower_whisker
[1] 13.7
> upper_whisker
[1] 47.29
>
```

شکل1-8

h) برای پیدا کردن داده های پرت (outliers) ، ابتدا داده های bmi را از بین بقیه متغیرهای مجموعه داده select می کنیم سپس مقادیر بزرگتر از upper whisker که همان داده های پرت ما هستند فیلتر می کنیم. نتیجه در شکل 1-9 نشان داده شده است.

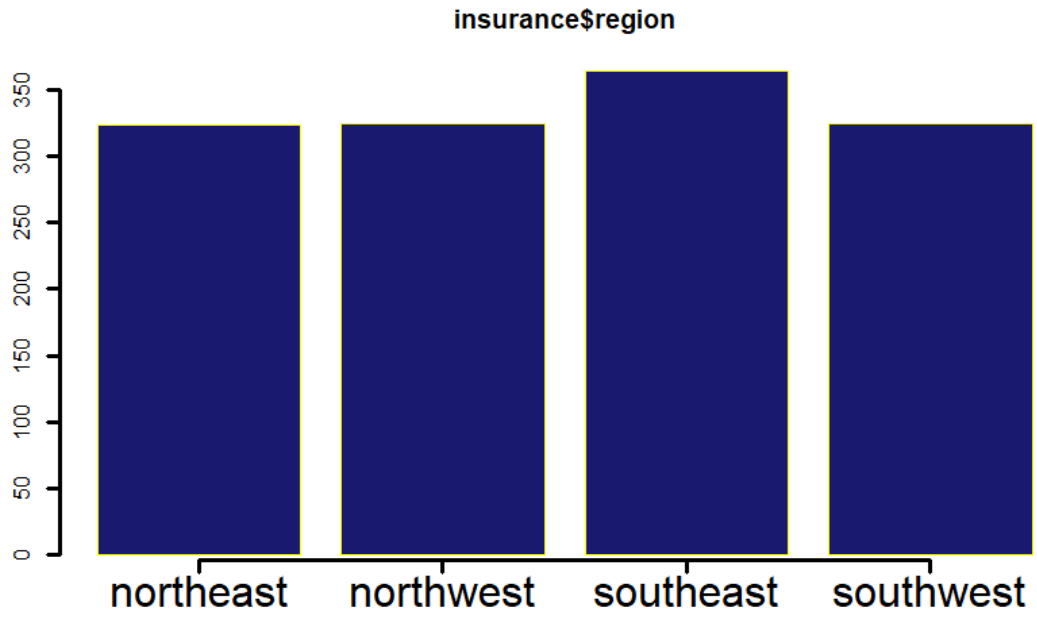
```
> filter(select(insurance, bmi), bmi > upper_whisker)
    bmi
1 49.06
2 48.07
3 47.52
4 47.41
5 50.38
6 47.60
7 52.58
8 47.74
9 53.13
>
```

شکل 9-1

QUESTION 2

در این مجموعه داده متغیرهای region و smoker ، categorical هستند. برای این سوال متغیر region را در نظر می گیریم.

a) نمودار barplot برای متغیر region را در شکل 2-1 می بینیم.



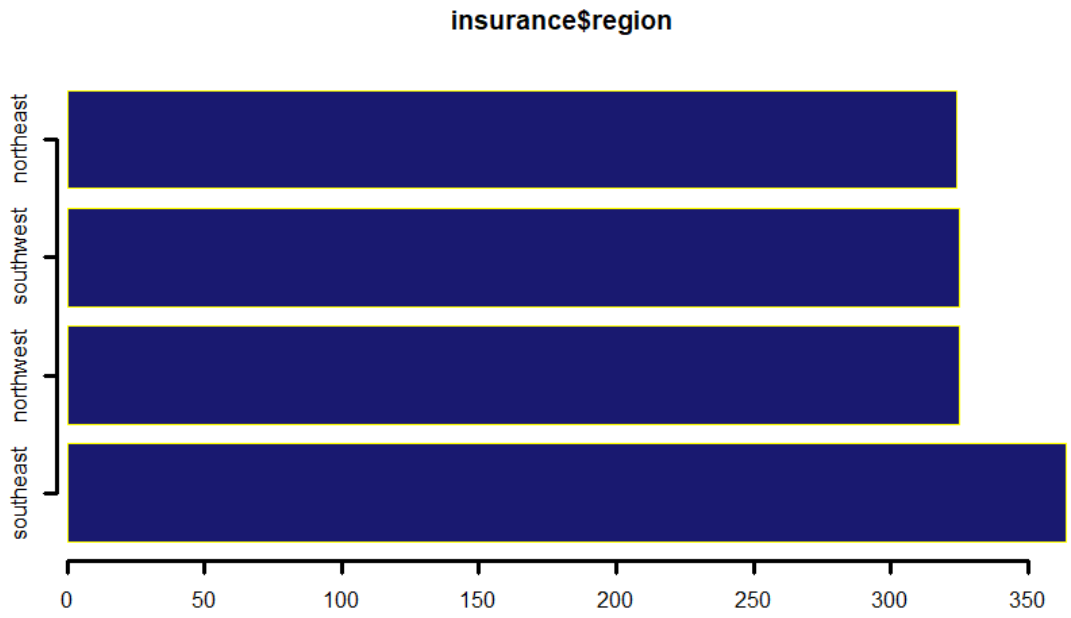
شکل 2-1

```
count <- table(insurance$region)
barplot(count, col = "midnightblue", border = "yellow",lwd = 3 , main = "insurance$region", cex.names=2, axis.lty=1 )
```

b) ابتدا مقادیر مناطق مختلف را با دستور order مرتب می کنیم و سپس با اضافه کردن ویژگی Horiz = TRUE نمودار را به فرم افقی تبدیل می کنیم کد این قسمت در شکل 2-2 آمده و نتیجه در شکل 2-3 نشان داده شده است.

```
> barplot(count, col = "midnightblue", border = "yellow",lwd = 3 ,
+         main = "insurance$region", cex.names=1, horiz = TRUE,axis.lty=1)
> count <- table(insurance$region)
> count

northeast northwest southeast southwest
      324       325       364       325
> count <- count[order(count,decreasing = TRUE)]
> count

southeast northwest southwest northeast
      364       325       325       324
> barplot(count, col = "midnightblue", border = "yellow",lwd = 3 ,
+         main = "insurance$region", cex.names=1, horiz = TRUE,axis.lty=1)
>
```

شکل 2-2



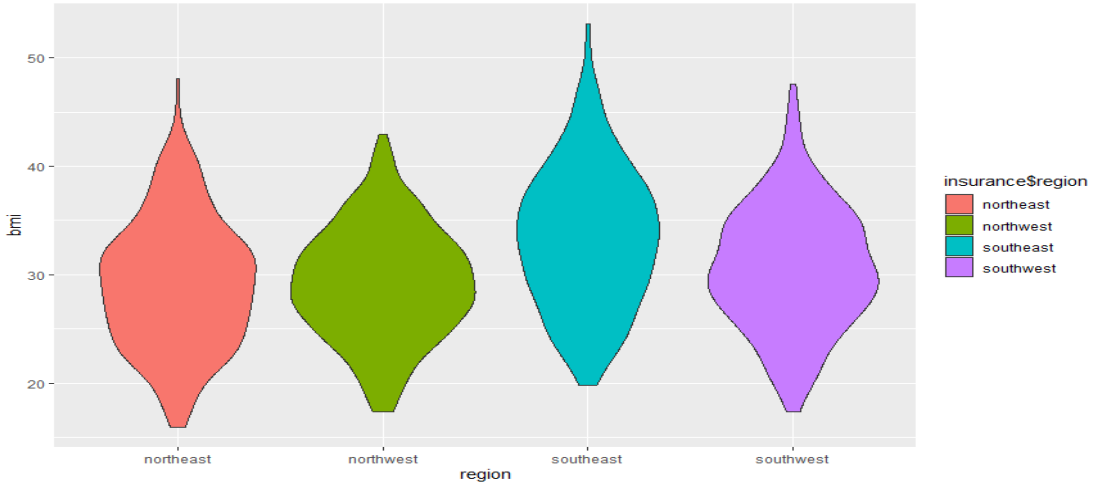
شکل 2-3

c) در این قسمت frequency table برای متغیر region خواسته شده که در شکل 2-4 آمده است.

```
> table(region)
region
northeast northwest southeast southwest
      324       325       364       325
> detach(insurance)
```

شکل 2-4

d) برای رسم violin plot علاوه بر متغیر region به یک متغیر numerical هم نیاز داریم که bmi را در نظر می گیریم و نمودار را با استفاده از کتابخانه ggplot2 رسم می کنیم . نتیجه در شکل 2-5 و کد این قسمت در شکل 2-6 نشان داده شده است.



شکل 2-5

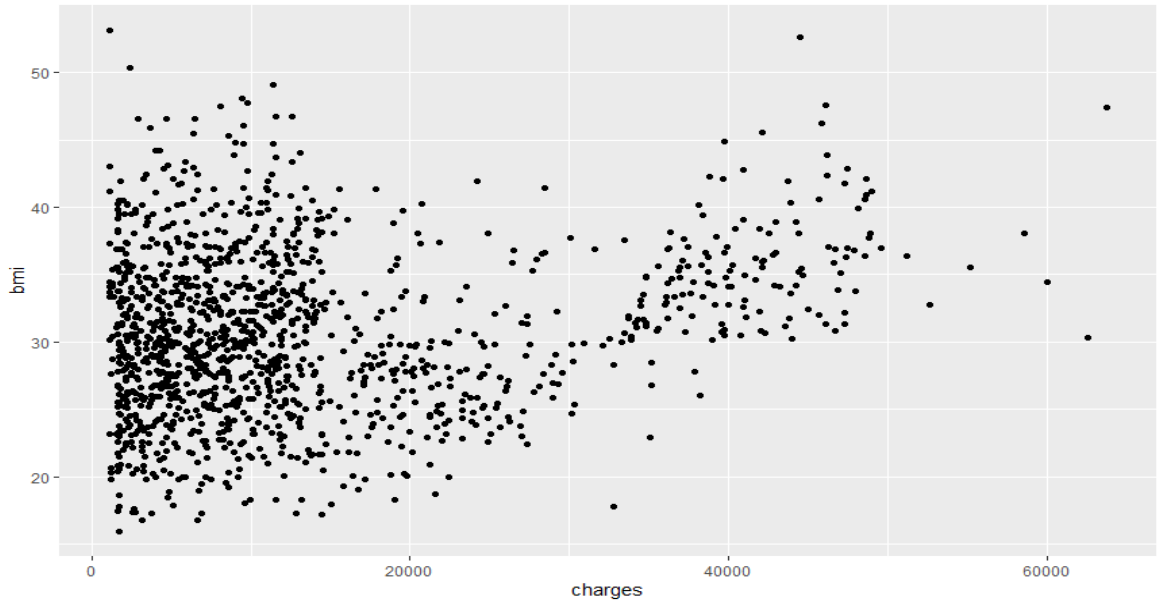
```
library(ggplot2)
ggplot(insurance, aes(x=region, y=bmi)) + geom_violin(aes(fill = insurance$region))
```

شکل 2-6

QUESTION 3

در این سوال دو متغیر عددی bmi و charges را انتخاب می کنیم و با استفاده از کتابخانه ggplot2 این بخش را انجام می دهیم.

a) بوسیله scatter plot می توان ارتباط بین دو متغیر عددی را بررسی کرد. این نمودار برای دو متغیر charges و bmi رسم شده و در شکل 1-3 و کد آن در شکل 2-3 آمده است. همانطور که مشاهده می کنیم bmi های کمتر از 30 حق بیمه کمتری دارند (بیشینه آن حدود 40000) ولی برای bmi های بیشتر از 30 مقادیر متفاوتی از بیمه را میبینیم بنابراین می توانیم بگوییم برای آن ها ارتباطی بین bmi و charges وجود ندارد.

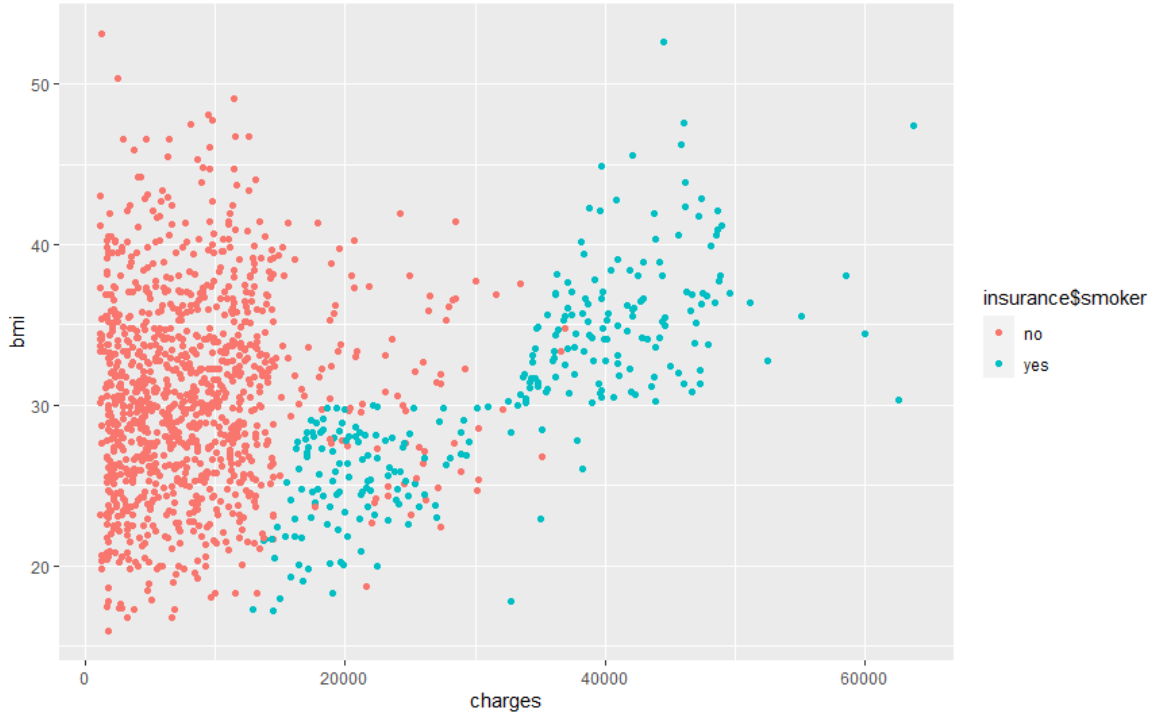


شکل 3-1

```
ggplot(insurance, aes(x=charges, y=bmi)) + geom_point()
```

شکل 3-2

b) برای این قسمت متغیر smoker را در نظر گرفتیم و scatter plot را رسم کردیم. همانطور که در شکل 3-3 میبینیم ، سیگاری ها(نقاط آبی رنگ) بیمه بیشتری دریافت می کنند و با افزایش bmi این مقدار بیشتر می شود.



شکل 3-3

```
ggplot(insurance, aes(x=charges, y=bmi, color = insurance$smoker)) + geom_point()
```

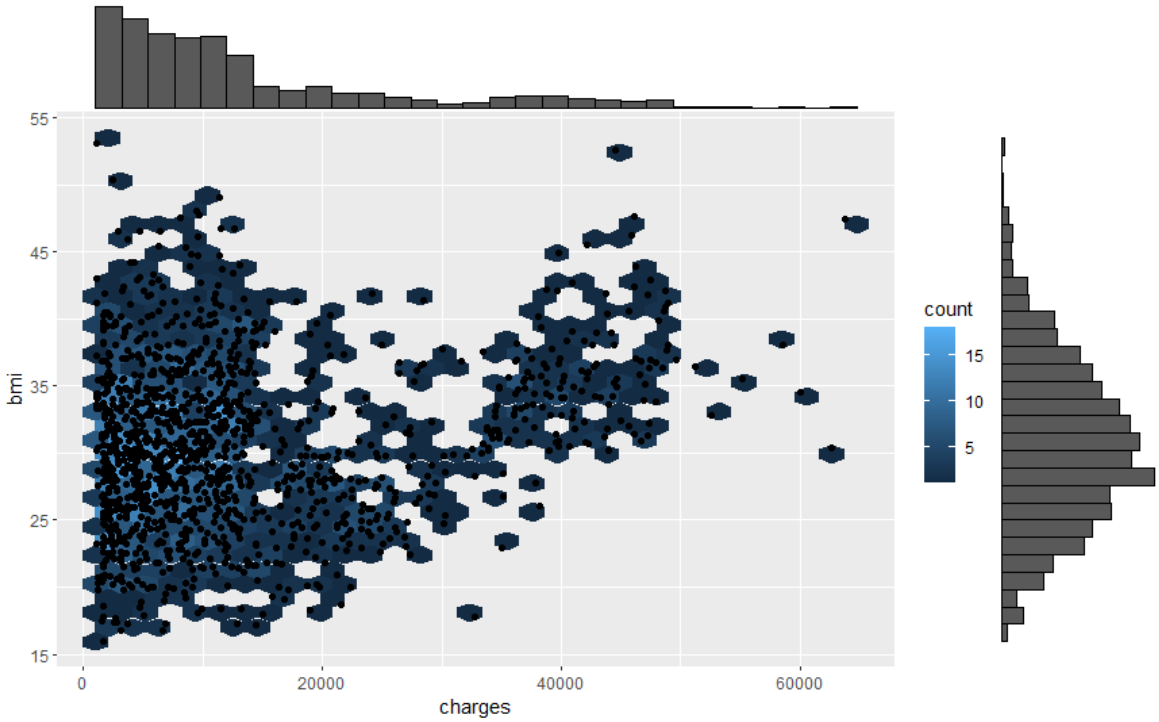
c) ضریب correlation برای متغیرهای charges و bmi ، همانطور که در شکل 3-4 دیده می شود برابر است با 0.19 . یعنی حدود 0.19 درصد دو متغیر correlated هستند که مقدار بسیار کمی است و همچنین مقدار p-value برابر است با 2.459e-13 که نزدیک به صفر است و این فرض که دو متغیر correlated هستند رد می شود.

```
> cor.test(insurance$charges, insurance$bmi, method = c("pearson"))

        Pearson's product-moment correlation

data:  insurance$charges and insurance$bmi
t = 7.3966, df = 1336, p-value = 2.459e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1463052 0.2492822
sample estimates:
      cor
0.198341
```

شکل 3-4

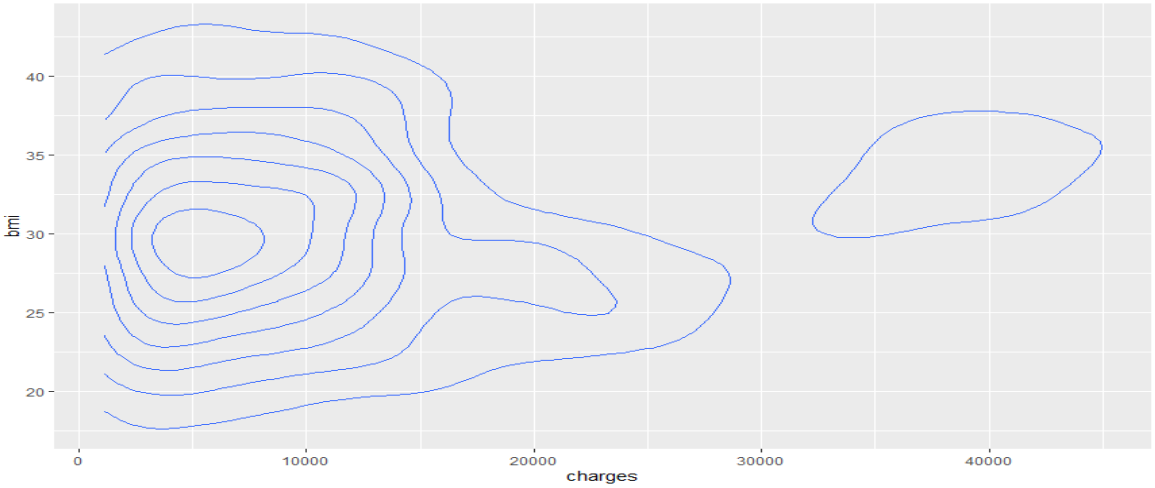
d) نمودارhexbin مانند یک هیستوگرام دوبعدی است که داده ها را داخل شش ضلعی هایی دسته بندی می کند و در واقع شش ضلعی ها مشابه bin ها در هیستوگرام یک بعدی هستند هرچقدر تعداد bin ها را بیشتر می کنیم اندازه شش ضلعی ها کوچکتر شده و تعداد داده هایی که داخل آن قرار میگیرد کمتر می شود. متغیر count در سمت راست نمودار شکل 3-5 همین مقادیر را نشان می دهد که در شکل با رنگ مشخص می شود. نمودار هیستوگرام یک بعدی هر متغیر هم روی محور متناظر با آن رسم شده است.

```
library(ggplot2)
library(ggExtra)
library(hexbin)
# bin <- hexbin(insurance$charges,insurance$bmi,xbins = 40)
# plot(bin)
p <-ggplot(insurance, aes(charges, bmi)) + stat_binhex() + geom_point()
p1 <- ggMarginal(p, type="histogram")
p1
```



شکل 5-3

e) نمودار 2d density برای دو متغیر charges و bmi در شکل 6-3 نشان داده شده است.



شکل 6-3

```
ggplot(insurance, aes(x=charges, y=bmi)) + geom_density_2d()
```

یک روش خوب برای شروع بررسی یک متغیر خاص، استفاده از هیستوگرام است. هیستوگرام متغیر را به دسته‌هایی تقسیم می‌کند، نقاط داده‌ای را در هر دسته می‌شمارد و دسته‌ها را روی محور x نمایش داده و تعداد نقاط را روی محور y نشان می‌دهد. عرض دسته(bin size) مهم‌ترین پارامتر برای هیستوگرام است و همواره باید مقادیر متفاوت عرض بررسی شوند تا بهترین مقدار برای هر مجموعه داده‌ای مشخص شود.

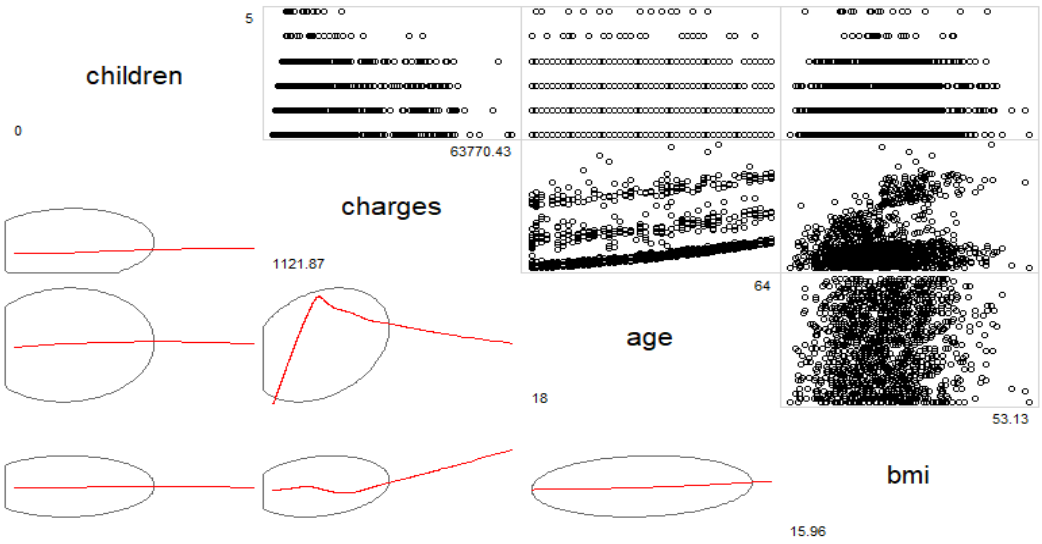
عرض‌های کم برای دسته ممکن است باعث شلوغ شدن نمودار شوند؛ اما از طرف دیگر عرض‌های بزرگ نیز ممکن است تفاوت‌های ظریف را نشان ندهند. وقتی می‌خواهیم توزیع‌های یک متغیر را در چند دسته از داده‌ها مقایسه کنیم، هیستوگرام‌ها با مانع خوانایی مواجه می‌شوند.

نمودار چگالی نسخه هموار و پیوسته‌ای از هیستوگرام است که از روی داده‌ها تخمین زده می‌شود. محور x مقدار متغیر را همانند هیستوگرام نشان می‌دهد و محور y در یک نمودار چگالی برابر با تابع چگالی احتمال است. نمودار چگالی همانند bin size در هیستوگرام پارامتری دارد که پهنای باند (bandwidth) نامیده می‌شود. این پارامتر تأثیر زیادی روی نتیجه نهایی نمودار دارد.

QUESTION 4

a) 4 متغیر عددی داریم که با استفاده از کتابخانه correlogram ، scatter plot دو به دوی آن ها را رسم می کنیم و نتیجه در شکل 1-4 آمده است.

```
library(GGally)
library(corrgram)
corrgram(insurance, order=NULL, lower.panel=panel.shade,
         upper.panel=NULL, text.panel=panel.txt)
corrgram(insurance, order=TRUE, lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt)

corrgram(insurance, order=TRUE, lower.panel=panel.ellipse,
         upper.panel=panel.pts, text.panel=panel.txt, diag.panel=panel.minmax)
ggcorr(insurance, method = c("everything", "pearson"))
```

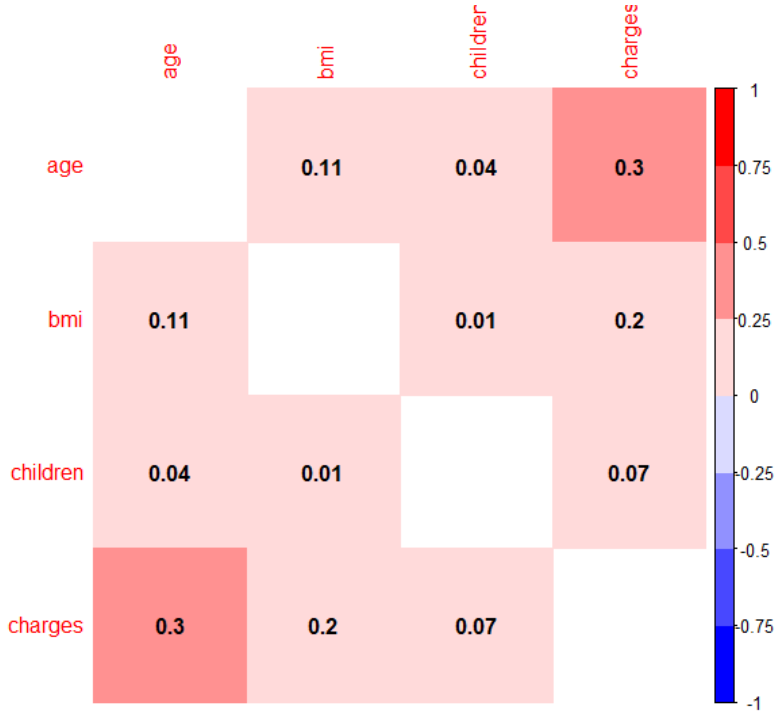


شکل 1-4

b) همانطور که در شکل 4-1 میبینیم در نمودار charges-age برای تعداد زیادی از داده ها با افزایش سن مقدار بیمه افزایش داشته است و به نظر می رسد این دو متغیر بهم مرتبطند و همچنین در نمودار charges-bmi تا حد کمتری این مسئله دیده می شود.

c)

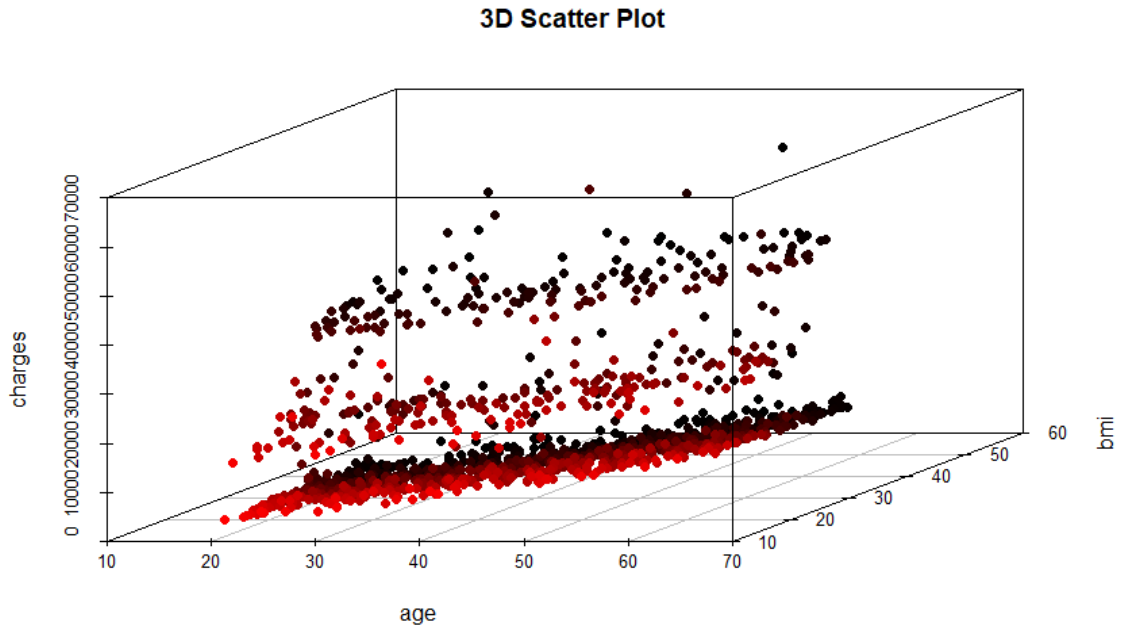
```
library(ggcorrplot)
library(dplyr)
library(corrplot)
filter_data <- filter(select(insurance, age, bmi, children, charges))
corr <- cor(filter_data[c('age','bmi', 'children', 'charges')])
color <- colorRampPalette(c("blue","white","red"))(8)
corrplot(corr, method = "color", addCoef.col = "black", diag = FALSE, col = color)
```



شکل4-2

d) همانطور که در قسمت a در scatter plot های تک بعدی مشاهده کردیم در شکل 4-3 هم می بینیم که بین داده های زیادی در مجموعه داده برای دو متغیر age , charges و همینطور charges , bmi رابطه ای وجود دارد بگونه ای که با افزایش سن و شاخص توده بدنی ، بیمه افزایش می یابد ولی این در مورد همه ی داده ها صادق نیست.

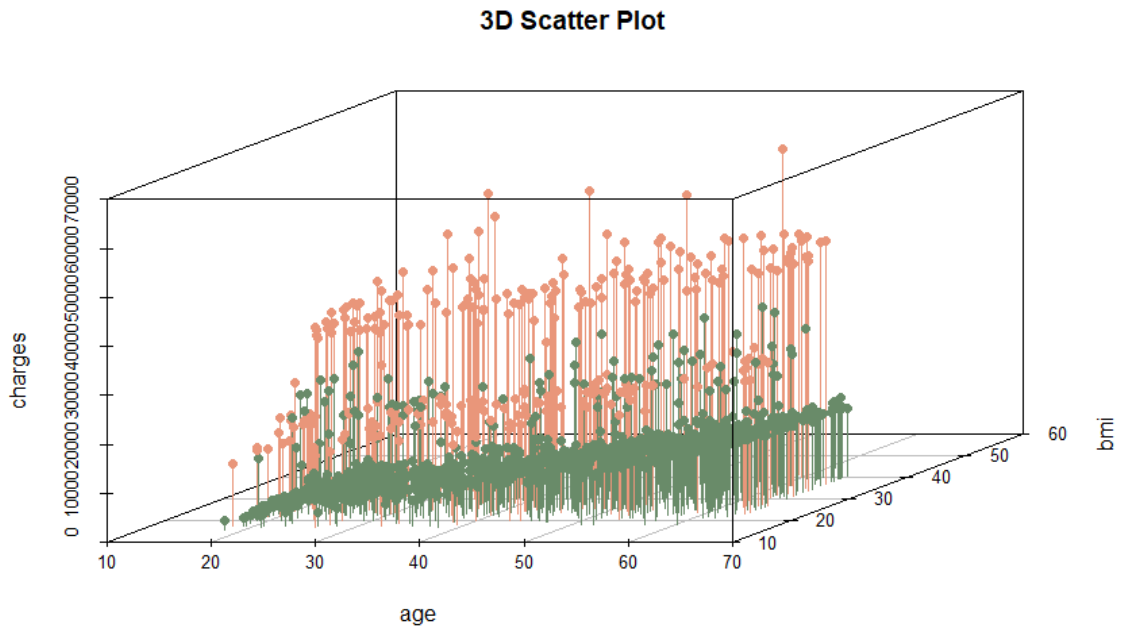
```
library("scatterplot3d")
filter_data <- filter(select(insurance, age, bmi, charges))
scatterplot3d(filter_data, main="3D Scatter Plot", pch = 16, highlight.3d = TRUE)
```



شکل 3-4

(e

```
library("scatterplot3d")
colors <- c("darkseagreen4","darksalmon")
colors <- colors[as.numeric(as.factor(insurance$smoker))]
filter_data <- filter(select(insurance, age, bmi, charges))
scatterplot3d(filter_data, main="3D Scatter Plot", pch = 16, color = colors, type = "h")
```



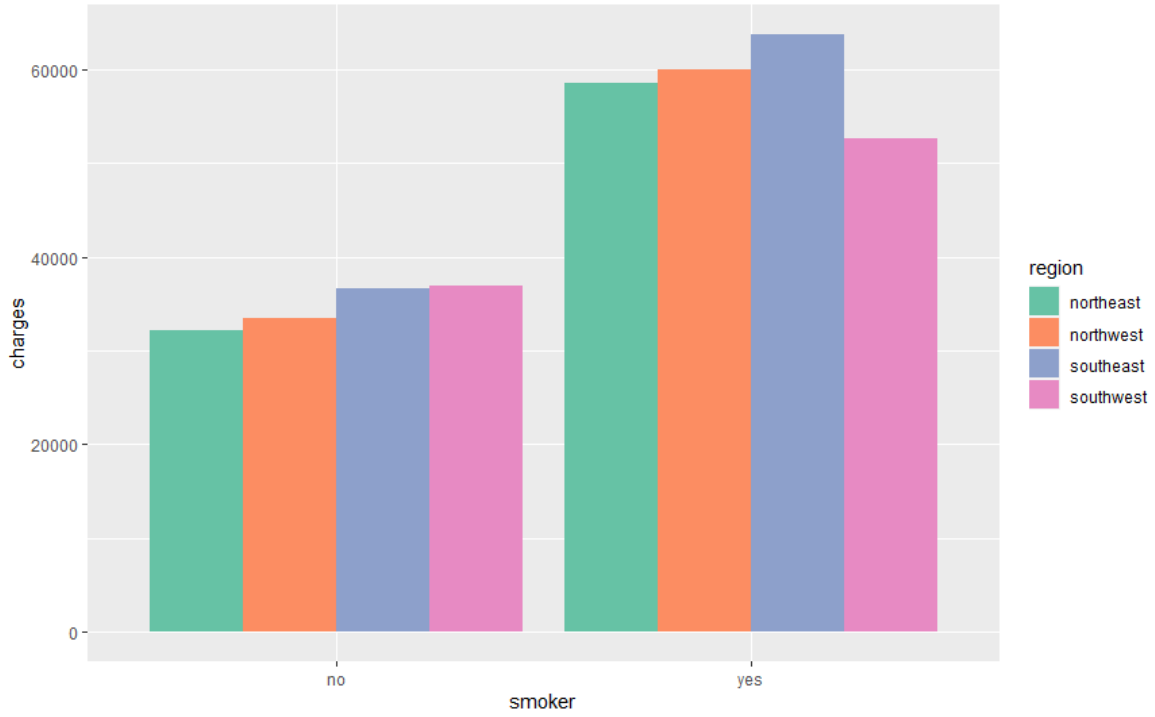
QUESTION 5

a) برای contingency table از بین متغیرهای categorical ، region,smoker را برای این سوال انتخاب می کنیم و می توانیم درصد افرادی را که در هر یک از 4 ناحیه سیگاری هستند در جدول ببینیم.

```
> ct1 <- table(insurance$smoker, insurance$region, dnn=c("insurance","region"))
> prop.table(ct1)
         region
insurance  northeast  northwest  southeast  southwest
      no  0.19207773 0.19955157 0.20403587 0.19955157
      yes 0.05007474 0.04334828 0.06801196 0.04334828
>
```

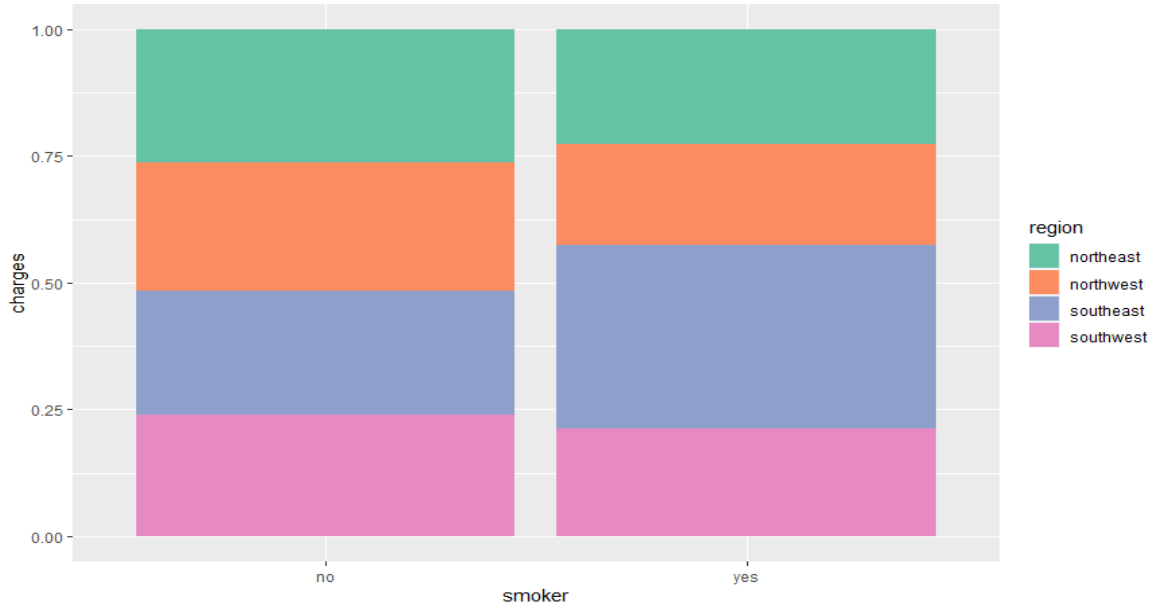
b) برای رسم grouped bar char علاوه بر دو متغیر categorical قسمت قبل ، متغیر عددی charges برای نشان دادن مقدار بیمه ای که افراد سیگاری و غیرسیگاری به تفکیک منطقه ی خود دریافت میکنند ، در نظر گرفته شده است. میبینیم که سیگاری های منطقه ی southeast بیشترین مقدار را دریافت می کنند.

```
library(RColorBrewer)
library(ggplot2)
insurance <-filter(select(insurance,smoker,region,charges))
ggplot(insurance, aes(x=smoker,y=charges,fill = region)) +
  geom_bar(position="dodge", stat="identity") + scale_fill_brewer(palette = "Set2")
```



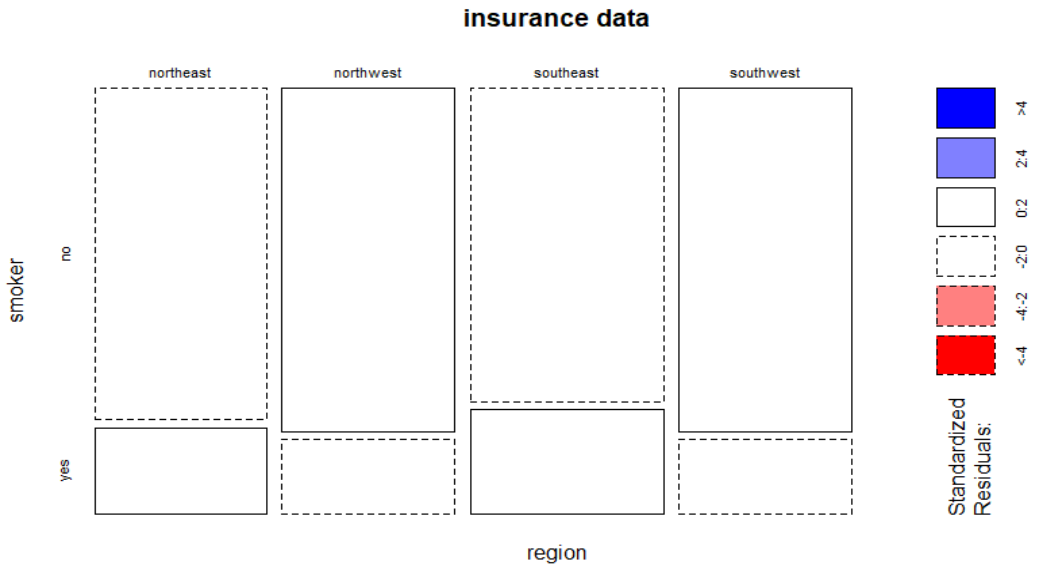
c) در segmented bar plot دو نمودار طول یکسانی دارند تا بتوان گروه ها را راحت تر نسبت بهم سنجید و در شکل هم میبینیم که مشابه نمودار قبل ، سیگاری های منطقه southeast سهم بیشتری را دریافت می کنند.

```
library(ggplot2)

ggplot(insurance, aes(x=smoker,y=charges,fill = region)) +
  geom_bar(stat="identity", position = "fill") + scale_fill_brewer(palette = "Set2")
```



(d

```
tbl <- xtabs(~region + smoker, insurance)
mosaicplot(tbl, main = "insurance data", shade = TRUE)
```



QUESTION 6

a) در شکل 1-6 نحوه ی محاسبه بازه اطمینان 98 درصد برای میانگین متغیرعددی charges نشان داده شده است

```
> library(Rmisc)
> insurance <- read.csv("insurance.csv")
> conf = CI(insurance$charges, ci = 0.98)
> low <- conf["lower"]
> up <- conf["upper"]
> conf
   upper     mean    lower
14041.52 13270.42 12499.32
>
```

شکل 1-6

b) تفسیر بازه اطمینان بدست آمده در قسمت a : 98 درصد اطمینان داریم میانگین بیمه دریافتی در مجموعه داده ی موردنظر در بازه ی [12499.32 , 14041.52] قرار دارد.

c)

d) برای طراحی آزمون فرض با سطح معناداری 0.02 ابتدا یک نمونه به سایز 35 می گیریم. میانگین (xbar) و انحراف معیار(s) متغیر charges و همچنین میانگین نمونه ی گرفته شده(m) را محاسبه می کنیم. سپس مقدار standard error و آماره ی z را از فرمول زیر محاسبه کرده و با pnorm مقدار

pvalue را بدست می آوریم. کد مربوط به این قسمت در شکل 6-2 و نتیجه در شکل 6-3 آمده و میبینمیکه مقدار pvalue از آلفا بزرگتر بوده و نمی توانیم فرض صفر را رد کنیم.

H0 : μ = m

Ha : μ ≠ m

Se = s / sqrt(n) = 12110.01 / 5.916 = 2046.996

z = (xbar - m) / se = (13270.42 – 12000.83) / 2046.996 = 0.6202323

```
library(Rmisc)
conf = CI(insurance$charges, ci = 0.98)
conf
sam <- sample(insurance[,7], 35, replace = FALSE)
xbar <- conf["mean"]
s <- sd(insurance$charges)
m <- mean(sam)
se <- s/sqrt(35)
z = (xbar-m)/se
pvalue <- 2*pnorm(-abs(z))
pvalue
```

شکل 6-2

```
> pvalue
     mean
0.5351048
```

شکل 6-3

e) بازه ی اطمینان 95 درصد برای میانگین این متغیر را در شکل 6-4 می بینیم:

```
> library(Rmisc)
> conf = CI(insurance$charges, ci = 0.95)
> conf
   upper     mean    lower
13919.89 13270.42 12620.95
```

شکل 6-4

کد مربوط به این قسمت برای محاسبه pvalue مشابه حالت قبل است فقط مقدار آلفا برای مقایسه با pvalue تغییر می کند . مقدار pvalue که در قسمت قبل بدست آوردیم برابر است با 0.5 که در مقایسه با مقدار 0.05 برای آلفا بزرگ بوده و فرض صفر باز هم رد می شود.

f) مقدار خطای نوع 2 را با استفاده از کد زیر محاسبه کردیم و مقدار آن برابر شد با 0.84 که کد و نتیجه در شکل 6-5 آمده است.

```
> q = qnorm(p=0.05,mean=m, sd=se, lower.tail=FALSE)
> beta <- pnorm(q, mean=xbar, sd=se)
> beta
[1] 0.847229
```

شکل 6-5

g) مقدار توان برابر است با : $1-\beta$. مقدار بتا در قسمت قبل محاسبه شد بنابراین مقدار توان برابر است با 0.152771 که مقدار بسیار کمی بوده و این به مقدار effect size بر میگردد.

QUESTION 7

در این سوال می خواهیم روی دو متغیر عددی آزمون فرض را انجام دهیم تا میانگین آن ها را با هم مقایسه کنیم . در سوال خواسته شده که یک نمونه 25 تایی از داده ها را برای اینکار جدا کنیم.
متغیرهای عددی bmi وcharges را برای این سوال انتخاب می کنیم.

a) زمانیکه انحراف معیار جامعه برای ما نامشخص باشد و همچنین زمانیکه سایزنمونه کمتر از 30 باشد ، از t-test استفاده می کنیم.

b) می خواهیم ببینیم آزمون فرض طراحی شده شواهدی قانع کننده از تفاوت بین میانگین دو متغیر ارائه میدهد یا خیر. فرض صفر را به این شکل تعریف می کنیم که تفاضل میانگین نمونه ها برابر صفر است یعنی دو نمونه میانگین برابر دارند و به دنبال رد این فرض هستیم.

نمونه هایی به سایز 25 از دو متغیر از مجموعه داده insurance می گیریم. میانگین و انحراف معیار دو نمونه و standard error را محاسبه کرده سپس آزمون t را انجام داده (با درجه آزادی 24 =1-25) و pvalue را محاسبه می کنیم . کد این قسمت در شکل 7-1 نشان داده شده است . برای بازه ی اطمینان 95 درصد مقدار آلفا 0.05 است و مقدار pvalue در نهایت با آلفا مقایسه می شود.

Hypothesis test :

H0 : $\mu 1-\mu 2 = 0$

Ha : $\mu 1-\mu 2 \neq 0$

```
sample_bmi <- sample(insurance[,3], 25, replace = FALSE)
sample_charges <- sample(insurance[,7], 25, replace = FALSE)
xbar1 <- mean(sample_charges)
xbar2 <- mean(sample_bmi)
s1 <- sd(sample_charges)
s2 <- sd(sample_bmi)
se = sqrt(s1^2/25 + s2^2/25)
t = (xbar1-xbar2)/se
2*pt(-abs(t),24)
```

شکل 7-1

مقدار pvalue محاسبه شده در شکل 7-2 آمده است و این مقدار بسیار کوچک و نزدیک صفر است و از مقدار آلفا یعنی 0.05 کوچکتر است پس فرض صفر رد می شود و میانگین دو متغیر تفاوت معناداری باهم دارد.

```
2*pt(-abs(t),24)
1] 1.620939e-06
```

شکل 2-7

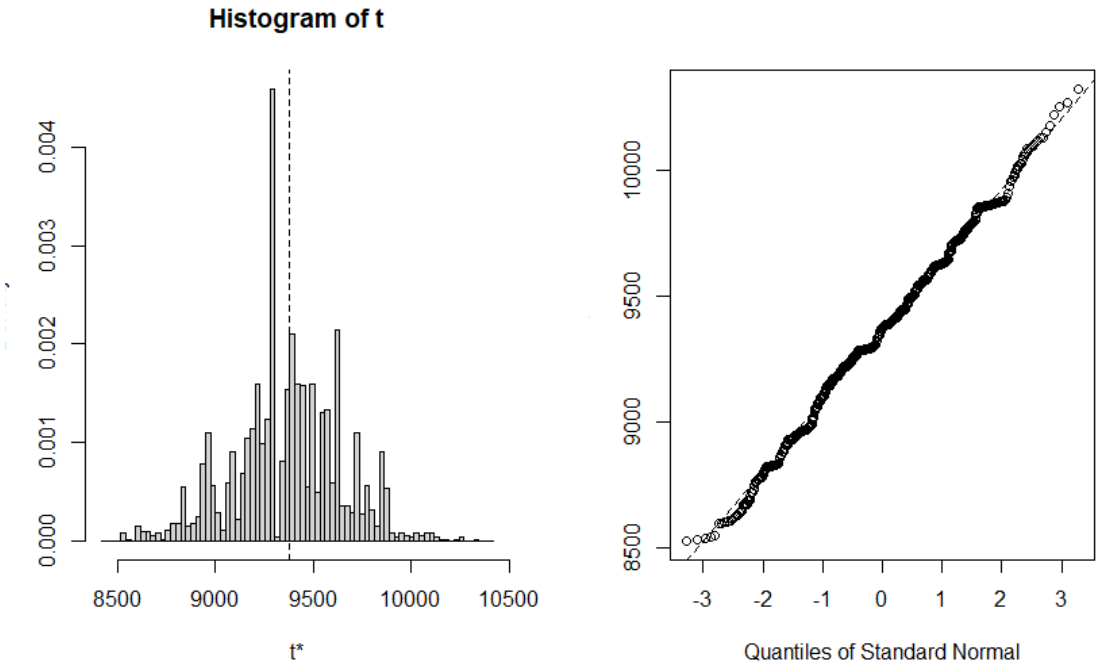
QUESTION 8

ابتدا boxplot را برای هر 4 متغیر عددی رسم می کنیم و می بینیم متغیر charges بیشترین outlier را دارد و برای این سوال این متغیر را انتخاب می کنیم . بدلیل حساس بودن میانگین نسبت به داده های پرت از میانه برای پیدا کردن بازه اطمینان استفاده می کنیم ولی بدلیل اینکه قضیه حد مرکزی برای میانه قابل استفاده نیست از دو روش percentile و standard error بازه اطمینان 95 درصد را برای میانه متغیر charges محاسبه می کنیم.

a) کد مربوط به این قسمت در شکل 1-8 آمده است. تابع calc_med را بعنوان statistic برای bootstrap تعریف کردیم. این تابع در هربار نمونه گیری در بوت استرپ میانه را بین داده ها پیدا می کند. در قسمت بعدی کد با استفاده از تابع boot در کتابخانه boot ، 1000 بار از داده ی charges در مجموعه داده insurance نمونه برداری کردیم و می توان توزیع میانه ها را با تابع plot(results) رسم کرد که نتیجه در شکل 2-8 نشان داده شده و همچنین می توان بازه اطمینان 95 درصد را با تابع boot.ci محاسبه کرد و یا می توانیم از تابع quantile با مشخص کردن حدود بازه اطمینان ، استفاده کنیم . نتیجه ی حاصل از این دو تابع برای بازه اطمینان در شکل 3-8 آمده است و برابر است با [8823.986 , 9875.168]

```
boxplot(insurance$age)
boxplot(insurance$bmi)
boxplot(insurance$charges)
boxplot(insurance$children)

library(boot)
calc_med <- function(data = insurance$charges, i) {
  d <- median(insurance$charges[i])
  return(d)
}

results <- boot(data=insurance$charges, statistic=calc_med,R=2000)
results
plot(results)
boot.ci(results)
quantile(results$t, c(0.025, 0.975))
```

شکل 1-8



شکل 8-2

```
CALL :
boot.ci(boot.out = results)

Intervals :
Level      Normal              Basic
95%   (8850, 9947 )   (8888, 9940 )

Level     Percentile            BCa
95%   (8824, 9876 )   (8824, 9876 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(results) : bootstrap variances nee
> quantile(results$t, c(0.025, 0.975))
    2.5%    97.5%
8823.986 9875.168
>
```

شکل 8-3

b) در این قسمت می خواهیم با روش standard error بازه ی اطمینان 95 درصد را برای میانه متغیر عددی charges بدست آوریم.کد این قسمت در شکل 8-4 آورده شده است. ابتدا انحراف معیار را برای خروجی bootstrap محاسبه کرده و سپس با تقسیم بر جذر تعداد نمونه های گرفته شده se را محاسبه می کنیم . و با تابع median هم میانه را پیدا کرده و مشابه روشی که در میانگین استفاده می کردیم عمل کرده و بازه ی اطمینان را محاسبه می کنیم و نتیجه را در شکل 8-5 می توان دید.

95% confidence interval for median = median ± 1.96 * se

```
sd <- sd(results$t)
se <- sd/sqrt(2000)
med <- median(results$t)
lower <- med - 1.96 * se
upper <- med + 1.96 * se
confidence_interval <- c(lower, upper)
confidence_interval
```

شکل 4-8

```
> confidence_interval
[1] 9357.354 9381.877
```

شکل 5-8

c) همانطور که در شکل های 3-8 و 5-8 مشاهده کردیم بازه ی اطمینان در دو روش با هم متفاوت بوده و در روش percentile به بازه ی اطمینان بزرگتری می رسیم در حالیکه در روش standard error این بازه بسیار کوچک و محدود بوده و بازه ی دقیق تری را به ما می دهد بنابراین استفاده از این روش بهتر است.

QUESTION 9

می خواهیم ببینیم افزایش تعداد فرزندان باعث افزایش مقدار بیمه میشود یا خیر. می توانیم با cor.test میزان همبستگی این دو متغیر را بررسی کنیم.

```
cor.test(insurance$children, insurance$charges, method = c("pearson"))
```

خروجی برابر است با 0.067 بنابراین این دو متغیر همبستگی کمی دارند.