دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس استنباط آماری

پروژه_فاز اول

اردیبهشت 1399

زهرا محقق راد

810199260

فهرست:

Question 0

(A

مجموعه‌داده مورد استفاده "HealthCare" ، شامل اطلاعات مربوط به وضعیت سلامتی 5110 فرد می‌باشد و نیز برای هر فرد 13 ویژگی آن از جمله جنسیت، سن، وضعیت تاهل، سیگاری بودن، bmi و ... ثبت شده است. باتوجه به این مجموعه‌داده می‌توان به اطلاعات جالبی دست‌یافت،به عنوان مثال ارتباط بین متوسط گلوکز و جنسیت و اینکه در کدام جنسیت این مقدار بیشتر می‌باشد، ارتباط بین جنسیت و نرخ bmi، تاثیر سیگار کشیدن بر bmi و متوسط گلوگز در بدن و نیز بر نرخ بیماری قلبی که همه این‌ها را می‌توان برحسب جنسیت به صورت جداگانه تفیک کرد و اطلاعات جالبی به‌دست‌آورد. همچنین می‌توان میزان مبلغی که هر فرد سالانه برای سلامتی خود پرداخت می‌کند را مشاهده کرد و نیز ارتباط آن را با گروه کاری که هر فرد دارد بررسی کرد، که کدام یک از تایپ‌های کاری، سالیانه مبالغ بیشتری را برای مراقبت‌های سلامتی خود صرف می‌کند. به طور کل می‌توان ارتباط بین تمام این مقادیر را برای 5110 نفر به‌دست آورد و اطلاعات جالبی کسب کرد.

(B

مجموعه‌داده مورد بررسی شامل 5110 نمونه از افراد می‌باشد که برای هر فرد 13 ویژگی آن ثبت شده‌است که این 13 ویژگی به شرح زیر می‌باشد.

- id
- Gender
- Age
- Hypertension
- Heart_disease
- Ever_married
- Work_type
- avg_glucose_level
- bmi
- smoking_status
- stroke
- health_bills
- Residence_type

ستون‌های 'age'، 'bmi'، 'avg_glucose_level' و 'health_bills'، متغیرهای numericalمی‌باشند و مابقی ستون-ها متغیر categoricalهستند.

(C

همانطور که می‌دانیم missing value، ستون‌هایی هستند که در برخی از سطرهای مربوط به آن‌ها مقدار "NA" ذخیره شده‌است. با استفاده از دستور ()is.na ستون‌هایی از مجموعه‌داده را که دارای missing value می‌باشند، پیدا کردیم. دو ستون 'bmi' و 'health_bills' به ترتیب هرکدام از 5510 سطری که دارند، 201 تا سطر از آن دارای مقدار 'NA' می‌باشد، تعداد این سطرها را با استفاده از دستور sum(is.na()) محاسبه کردیم. که تقریبا می‌توان گفت 4% از سطرهای مربوط به این دو ستون دارای missing value می‌باشند. به منظور هندل کردن این مقادیر تاجایی که امکان محاسبه‌ی مقادیر این ستون‌ها با استفاده از ستون‌های دیگر باشد، آن‌ها را حساب کرده و مقادیر آن‌ها را در سطر و ستون‌های مربوط ذخیره می‌کنیم، در غیر این اگر بخواهیم با این ستون‌ها کار کنیم و مقادیر آن‌ها قابل محاسبه از سطرهای دیگر نباشد، این سطرها را حذف می‌کنیم.

D)

باتوجه به اطلاعات و ارتباطی که می‌خواهیم آن را مشاهده کنیم، هرکدام از ستون‌ها می‌توانند مهم باشند. به عنوان مثال برای بررسی نرخ بیماری قلبی، اینکه فرد سیگاری بوده‌است یا خیر می‌تواند relevant باشد، یا جنسیت، سن یا وضعیت زندگی یک فرد که روستایی بوده‌است یا شهری، می‌تواند در نرخ بیماری یا متوسط گلوکز یا نرخ bmi آن تاثیرگذار باشد. هم‌چنین تایپ کاری افراد به میزان هزینه‌ای که به طور سالیانه برای مراقبت از سلامتی خود پرداخت می‌کند مرتبط می‌باشد و ....

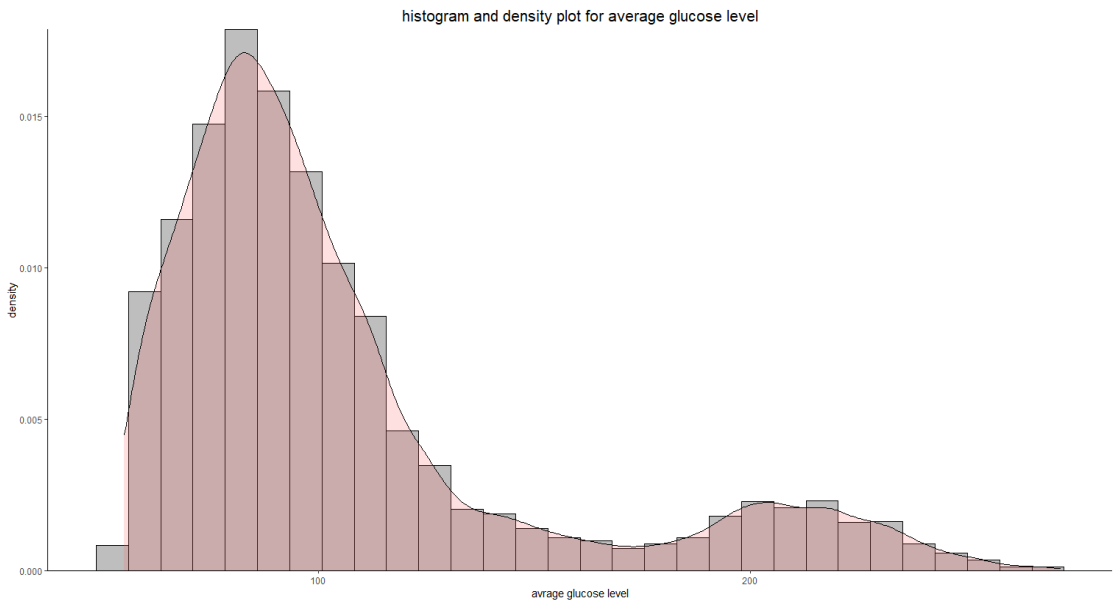
Question 1

برای این قسمت متغیر avg_glucose_level به عنوان یک متغیر numerical مورد بررسی قرار گرفته‌شده‌است.

A)

با استفاده از دستورات زیر نمودار هیستوگرام مربوط به این ستون رسم شده و نیز density بر روی آن فیت شده‌است. خروجی به شرح زیر می‌باشد: (سایز bin برابر 30 در نظر گرفته‌شده‌است)

```
ggplot(data, aes(x =avg_glucose_level))+
 geom_histogram(aes(y=..density..),bins = 30, color = "black", fill="Gray")+
 geom_density(alpha=0.2, fill ="#FF6666")+
 scale_y_continuous(expand = c(0, 0))+
 ggtitle("histogram and density plot for average glucose level")+
 labs(x= "avrage glucose level")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank,()
  panel.grid.minor = element_blank,()
  panel.background = element_blank,()
  axis.line = element_line(colour = "black"),
)
```

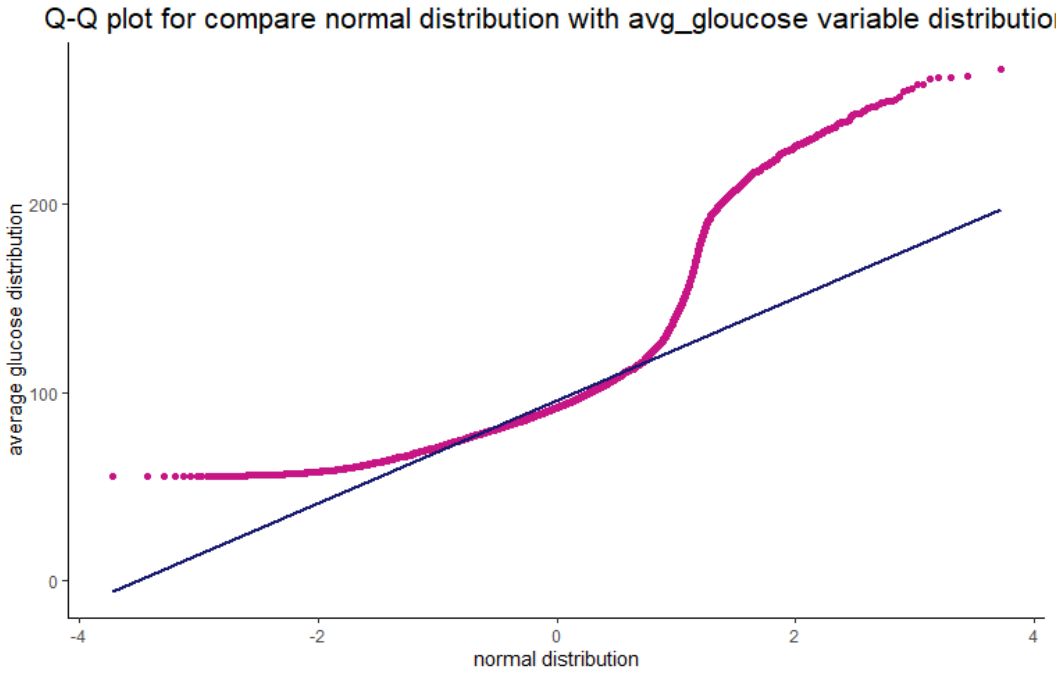


باتوجه به نموداری که در بالا مشاهده می‌فرمایید، ابتدا نمودار در یک نقطه پیک داشته و سپس مقدار آن کاهش یافته و دو باره یک پیک کوچکی داشته است که البته نمی‌توان این را به عنوان یک مد در نظر گرفت. پس توزیع این متغیر از نوع unimodal می‌باشد. (با توجه به اینکه پیک دوم خیلی کوچکتر است نمی‌توان این توزیع را bimodal در نظر گرفت). همچنین از نمودار بالا قابل مشاهده می‌باشد که توزیع right skewed می‌باشد. زیرا توزیع‌های right skewed دارای یک long tail در سمت راست می‌باشند.

(B

باتوجه به نمودار هیستوگرامی که در بالا مشاهده کردیم، این متغیر یک توزیع right skewed دارد و نیز می‌توان گفت که توزیع unimodal می‌باشد. برای مقایسه این توزیع این متغیر با توزیع نرمال از نمودار Q-Q plot استفاده کردیم که خروجی آن به شرح زیر می‌باشد:

```
ggplot(data, aes(sample=avg_glucose_level))+
  stat_qq(color = "#FFD700") + stat_qq_line(color = "#191970", lwd = 1)+
  ggtitle("Q-Q plot for compare normal distribution with avg_gloucose variable distribution")+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "black"),
    axis.title = element_blank()
  )
```



همانطور که در نمودار بالا قابل مشاهده می‌باشد، تعداد زیادی از نقاط ابتدایی از خط فاصله گرفته‌اند و این به این منظور است که از توزیع نرمال فاصله دارند. در نقاط میانی روی خط فیت شده‌اند بنابرای از توزیع نرمال پیروی می‌کنند. تعداد بسیار زیادی از نقاط در نقاط پایانی فاصله‌ی زیادی از خط گرفته‌اندکه نشان‌دهنده‌ی این است که این نقاط بسیار از توزیع نرمال فاصله دارند. به عبارتی دیگر این نوع نمودار، همان‌طور که در درس گفته‌شد، نشان دهنده‌ی یک توزیع right skewed می‌باشد.

(C

برای محاسبه‌ی skewness از تابع ()skewness موجود در کتابخانه‌ی "moment" استفاده شده‌است. خروجی و دستور استفاده شده برای این قسمت به شرح زیر می‌باشد:

```
library(moments)
skewness(data$avg_glucose_level)
```

```
> [1] 1.571822
```

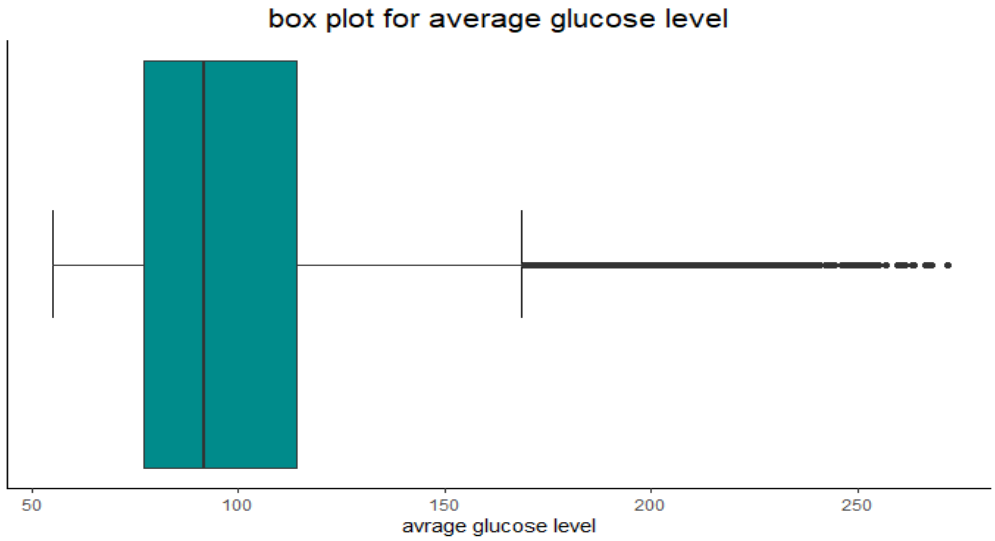
همانطور که ملاحظه می‌فرمایید مقدار به‌دست‌آمده مثبت می‌باشد که به این معنی است که right skewed داریم و نیز میانگین توزیع این متغیر از میانه آن بزرگتر بوده‌است، زیرا مقدار مثبت‌شده‌است و right skewed داریم. همچنین هرچه این مقدار از 0 بیشتر باشد، skewness بیشتری را در توزیع شاهد خواهیم بود.

(D

برخی از راه‌های نمایش outlier ها استفاده از boxplot یا استفاده از scatter plot ها می‌باشد، برای این قسمت از boxplot استفاده کردیم که به شرح زیر می‌باشد:

```
ggplot(data, aes(y = avg_glucose_level)) +
stat_boxplot(geom ='errorbar', width = 0.2) +
  geom_boxplot(fill = "#008B8B") +
  ggtitle("box plot for average glocose level")+
  labs(y= "avrage glucose level")+
  theme)
    plot.title = element_text(hjust = 0.5, size = 16),
    panel.grid.major = element_blank,()
    panel.grid.minor = element_blank,()
    panel.background = element_blank,()
    axis.line = element_line(colour = "black"),
    axis.ticks.y = element_blank,()
    axis.text.y = element_blank,()
```

)+ coord_flip()



نقاطی که در تصویر خارج از whisker بالا قرار گرفته‌اند، outlier می‌باشند که تعداد آن‌ها برابر 627 عدد است، که این مقدار را با دستور length(p$out) به دست‌آوردیم. outlier های این متغیر، نشان دهنده‌ی این می‌باشد که 627 نفر از افراد متوسط گلوکز بیشتر از مقدار 168 دارند (زیرا مقدار whisker بالا برابر 168 می‌باشد).

(E

برای محاسبه‌ی میانگین، میانه، واریانس و انحراف معیار از دستورات زیر استفاده شده‌است که خروجی آن‌ها نیز در زیر قابل نمایش می‌باشد:

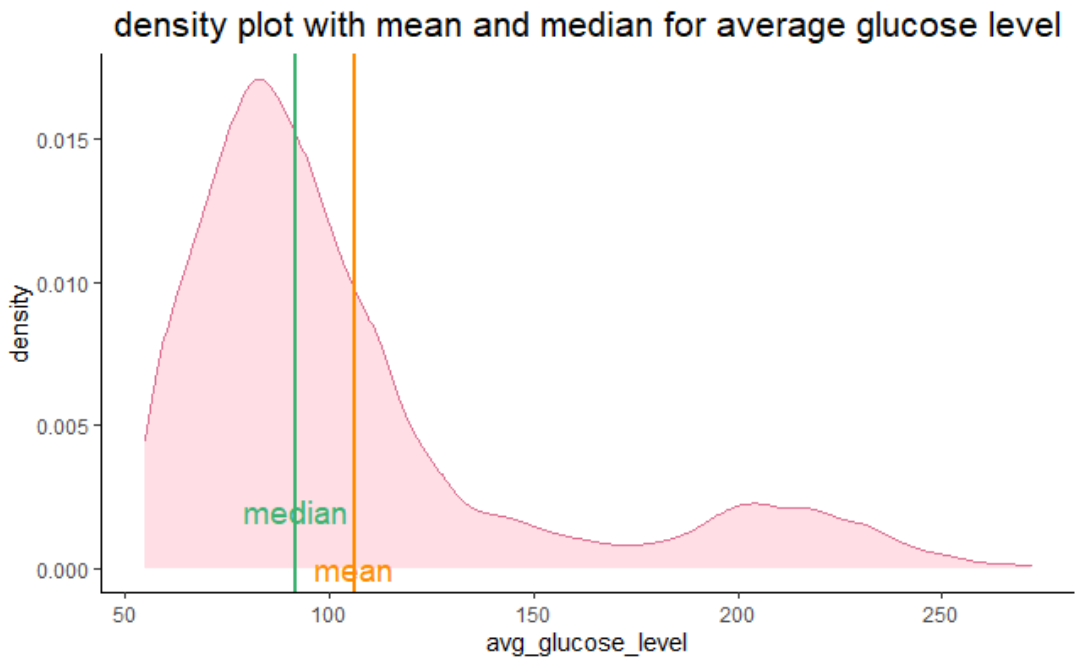
```
> mean(data$avg_glucose_level)
[1] 106.1477

> median(data$avg_glucose_level)
[1] 91.885

> var(data$avg_glucose_level)
[1] 2050.61

>sd(data$avg_glucose_level)
[1] 45.28356
```

همانطور که در خروجی‌های بالا مشاهده می‌کنید، میانگین داده‌های این ستون از مجموعه‌داده مورد بررسی برابر 106.15 می‌باشد و میانه برابر 92 می‌باشد. همانطور که در بالا هم ذکر کردیم و در نمودار توزیع مشاهده کردیم چون توزیع right skewed داشتیم، مقدار میانه باید کمتر از مقدار میانگین باشد. همچنین مقدار واریانس برابر 2051 و انحراف معیار 45.3 می‌باشد.

(F

برای رسم density plot و نیز مشخص نمودن میانگین و میانه بر روی آن از دستورات زیر استفاده شده‌است و خروجی آن به شرح زیر می‌باشد:

```
ggplot(data, aes(x = avg_glucose_level)) +
 geom_density(color = "Navy", fill="Navy", alpha=0.5)+
 geom_vline(xintercept=glu_mean, size=1, color="DarkOrange")+
 annotate("text", x=glu_mean, y=0, label= "mean", color = "DarkOrange", size = 5)+
 geom_vline(xintercept=glue_median, size=1, color="MediumSeaGreen")+
 annotate("text", x=glue_median, y=0.002, label= "median", color = "MediumSeaGreen", size = 5)+
 ggtitle("density plot with mean and median for average glucose level")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black")
 )
```



رابطه‌ی بین میانگین و میانه همانطور که در نمودار بالا قابل مشاهده‌است به صورت median < mean می‌باشد، به همین دلیل همانطور که می‌دانیم در این مواقع توزیع به صورت right skewed می‌باشد که در تصویر بالا نیز مشخص می‌باشد.

(G

با استفاده از تابع cut() ابتدا این متغیر را به یک متغیر categorical تبدیل کردیم. مقادیر که کمتر از 1/2 میانگین می‌باشند، لیبل 'A' گرفته‌اند، مقادیری که بین 1/2 تا میانگین می‌باشند، لیبل 'B' گرفته‌اند، مقادیری که بزرگتر مساوی از مقدار میانگین و کوچکتر از 3/4 میانگین می‌باشند لیبل 'C' گرفته‌اند و در نهایت مقادیر بزرگتر از 3/4 میانگین لیبل 'D' گرفته‌اند. کدهای استفاده شده برای این قسمت به شرح زیر می‌باشد:

```
category_glucose_level <- data.frame((table(cut(data$avg_glucose_level,
      breaks=c(0, glu_mean / 4  , glu_mean / 2 , 3 * glu_mean / 4, glu_mean),
      label = c("A", "B", "C", "D")))))
```

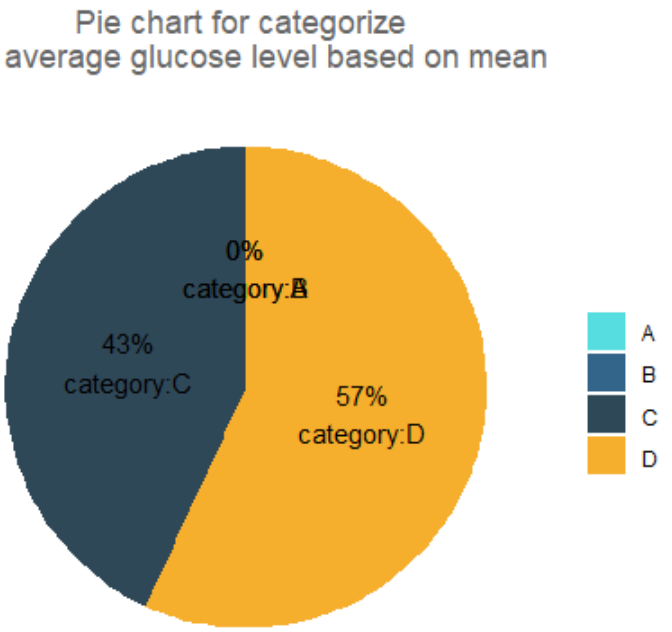
خروجی این قسمت به صورت زیر است:

	Var1	Freq
1	A	0
2	B	0
3	C	1495
4	D	1971

همانطور که قابل مشاهده می‌باشد تعداد سطرهایی که از ستون avg_glucose_level در گروه A و B می‌افتند برابر 0 می‌باشد. pie chart این متغیر به صورت زیر می‌باشد:

```
value <- category_glucose_level$Freq / sum(category_glucose_level$Freq)
```

```
ggplot(category_glucose_level, aes(x="", y=Freq, fill=Var1)) +
 geom_bar(stat="identity", width=1)+
 coord_polar("y", start=0)+
 geom_text(aes(label = paste0(round(value*100), "%")),
        position = position_stack(vjust = 0.5))+
 scale_fill_manual(values=c("#55DDE0", "#33658A", "#2F4858", "#F6AE2D"))+
 labs(x = NULL, y = NULL, fill = NULL, title = "categorize average glucose level based on mean")+
 theme_classic()+
 theme(axis.line = element_blank(),
     axis.text = element_blank(),
     axis.ticks = element_blank(),
     plot.title = element_text(hjust = 0.5, color = "#666666"))
```



همانطور که در بالا گفتیم و pie chart نیز قبل مشاهده می‌باشد، درصد سطرهایی که در گروه A و B قرار می‌گیرد برابر صفر می‌باشد. به عبارتی هیچ سطری نداریم که مقادیر آن کمتر از میانگین باشد.

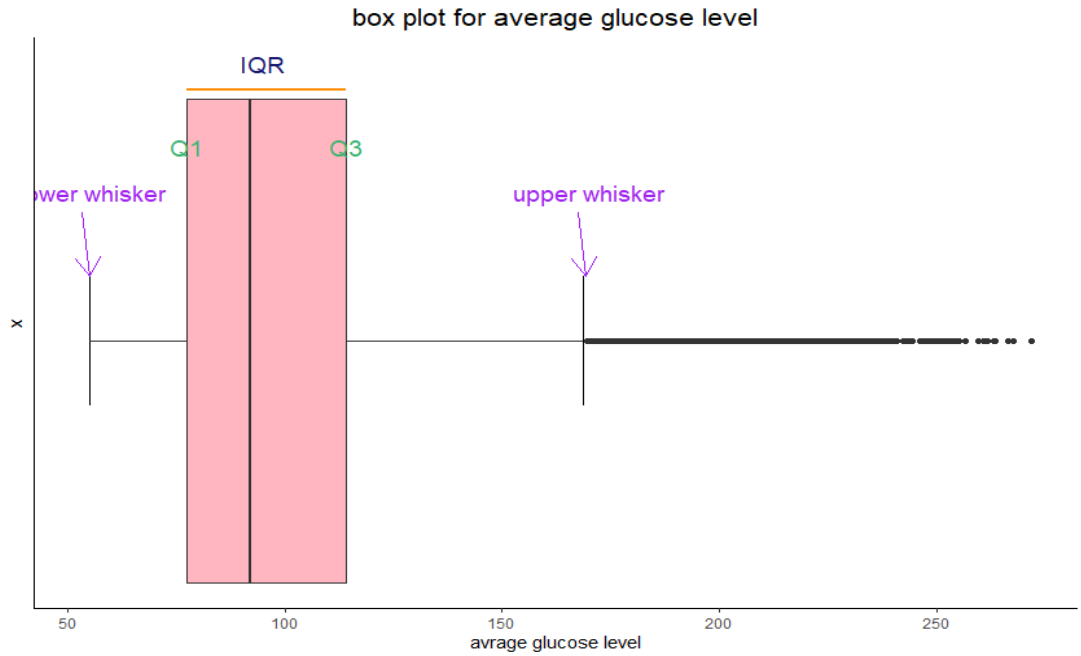
(H

برای نمایش box plot از توابع کتابخانه‌ی ggplot2 استفاده شده‌است. برای مشخص کردن IQR و whisker بر روی نمودار از توابع annotate() و geom_segment() استفاده شده‌است. کدهای زده شده برای این قسمت و خروجی آن به شرح زیر می‌باشد:

```
lower <- quantile(data$avg_glucose_level, 0.25)
upper <- quantile(data$avg_glucose_level, 0.75)
whisker_L <- min(data$avg_glucose_level)
whisker_U <- upper + 1.5*(upper - lower)
```

```
ggplot(data, aes(y = avg_glucose_level)) +
  stat_boxplot(geom ='errorbar', width = 0.2) +
  geom_boxplot(fill = "LightPink") +
  annotate("text", y=lower, x=0.3, label= "Q1", color = "#3CB371", size = 5)+
  annotate("text", y=upper, x=0.3, label= "Q3", color = "#3CB371", size = 5)+
  annotate("text", y=whisker_L, x=0.23, label= " lower whisker", color = "purple", size = 5)+
  annotate("text", y=whisker_U, x=0.23, label= " upper whisker", color = "purple", size = 5)+
  annotate("text", y=95, x=0.43, label= "IQR", color = "MidnightBlue", size = 5)+
  geom_segment(aes(x = 0.2, y = whisker_U - 2, xend = 0.1, yend = whisker_U),
               arrow = arrow(length = unit(0.5, "cm")), color = "darkorange")+
  geom_segment(aes(x = 0.2, y = whisker_L - 2, xend = 0.1, yend = whisker_L),
               arrow = arrow(length = unit(0.5, "cm")), color = "darkorange")+
  geom_segment(aes(x = 0.39, y = lower, xend = 0.39, yend = upper),
```

```
          color = "darkorange", lwd = 1)+
ggtitle("box plot for average glucose level")+
labs(y= "avrage glucose level")+
theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
  axis.ticks.y = element_blank(),
  axis.text.y = element_blank(),
)+ coord_flip()
```



همانطور که در نمودار بالا قابل مشاهده می‌باشد، Q1، چارک اول می‌باشد که از 25% دیتا بیشتر بوده و این مقدار را باستفاده از تابع quantile() محاسبه شده‌است. Q3 نیز چارک سوم می‌باشد که از 75% دیتا بیشتر بوده و از 25% دیتا کمتر می‌باشد. برای محاسبه‌ی آن نیز از quatile() استفاده شده‌است و دو مقدار Q1 و Q3 در نمودار بالا با رنگ سبز مشخص شده‌اند و همانطور که در تصویر مشخص است Q1 = 72.2 و Q3 = 114 می‌باشد. IQR نیز برابر تفاضل Q1 و Q3 می‌باشد و در نمودار بالا با رنگ سرمه‌ای نمایش داده شده‌اصست. در نهایت دو whisker بالا و پایین؛ همانطور که می‌دانیم در محاسبه‌ی whisker پایین اگر زودتر به نقطه مینیمم دیتا برسیم آن نقطه برابر با whisker پایین می‌شود در غیر این صورت مقدار آن به صورت $Q1 - 1.5 \times IQR$ محاسبه می‌شود. برای این متغیری که ما در این قسمت مورد بررسی قرار دادیم چون نقطه‌ی مینیمم جلوتر است، بنابراین این نقطه برابر با whisker پایین می‌شود که در نمودار بالا با رنگ بنفش نمایش داده‌شده است. همین‌طور برای whisker بالا، اگر زودتر به نقطه‌ی ماکسیمم دیتا برسیم، آن نقطه برابر با whisker بالا خواهد بود در غیر

این صورت به صورت $Q3 + 1.5 \times IQR$ محاسبه می‌شود.که برای متغیر مورد بررسی ابتدا به مقداری که این فرمول به ما می‌دهد می‌رسیم، پس همین مقدار به دست آمده برابر whisker بالا شده و در تصویر بالا با رنگ بنفش نمایش داده شده‌است.

Question 2

برای این قسمت متغیر Residence_type به عنوان یک متغیر categorical مورد بررسی قرار گرفته‌شده‌است.

(A

برای محاسبه‌ی تعداد در هر دسته از تابع table() بر روی این متغیر استفاده شده‌است و خروجی آن تحت عنوان یک دیتافریم ذخیره‌شده‌است. سپس با استفاده از تابع prop.table() برای محسابه درصد هر دسته استفاده‌شده است. کدهای مورد استفاده برای این قسمت و نیز خروجی آن به شرح زیر می‌باشد:

```
Residence <- data.frame(table(data$Residence_type))
Residence$per <- format(round(prop.table(data.frame(table(data$Residence_type))$Freq)
                 *100, 2), nsmall = 2)
```

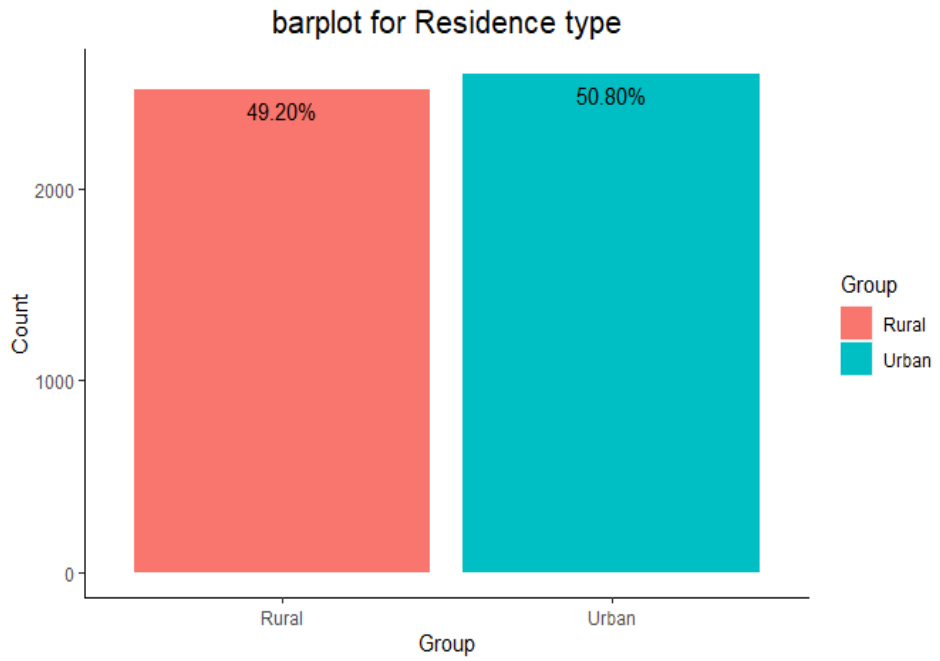
	Group	Count	percent
1	Rural	2514	49.20
2	Urban	2596	50.80

همانطور که در جدول بالا قابل مشاهده می‌باشد، داده‌ها در این متغیر به سه دسته 'Rural' و 'Urban' تقسیم شده‌اند. که بیشترین تعداد و درصد برای دسته‌ی 'Urban' می‌باشد و کم‌ترین تعداد و درصد به دسته 'Rural' تعلق دارد.

(B

به منظور رسم Bar plot برای این متغیر categorical از توابع موجود در کتابخانه‌ی ggplot2 استفاده شده‌است. کدهای زده‌شده برای این قسمت و نمودار خروجی به شرح زیر می‌باشد:

```
ggplot(Residence, aes(x=Var1, y=Freq, fill=Var1)) +
 geom_bar(stat="identity")+
 geom_text(aes(label=paste0(per,"%")), vjust=1.6, color="white", size=3.5)+
 ggtitle("box plot for average glucose level")+
 labs(x = "Group", y= "Number")+
 theme(
   plot.title = element_text(hjust = 0.5, size = 16),
   panel.grid.major = element_blank(),
   panel.grid.minor = element_blank(),
   panel.background = element_blank(),
   axis.line = element_line(colour = "black"),
 )
```



(C

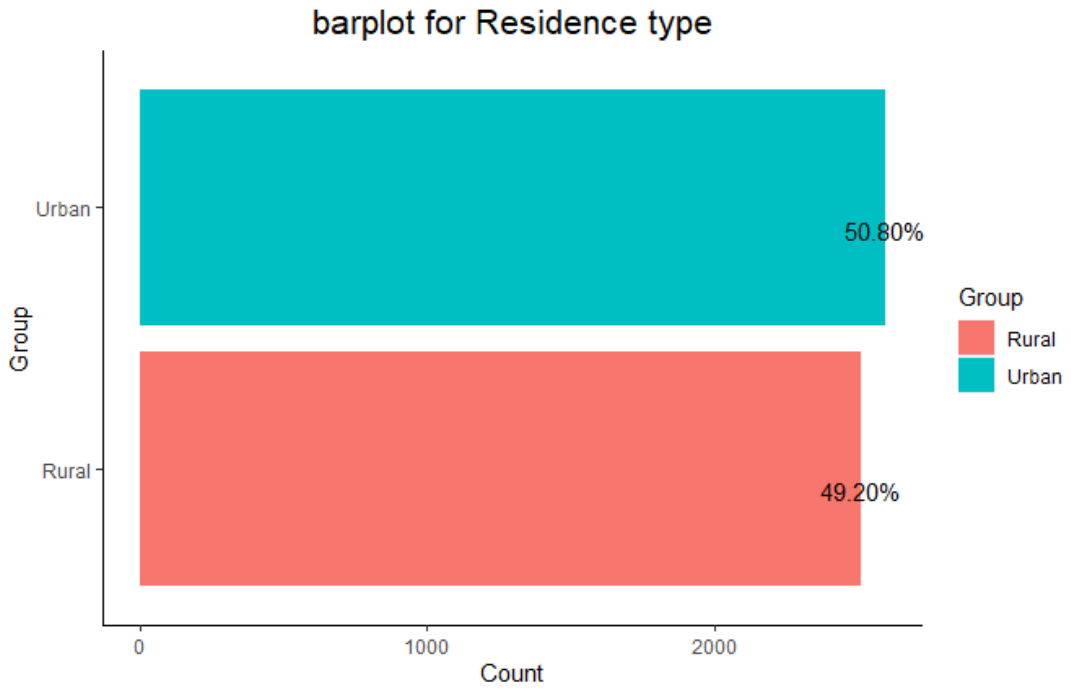
در این قسمت ابتدا با استفاده از تابع order()، دیتافریمی که در مورد 'A' درست کرده‌بودیم، مرتب نموده و سپس barplot را برای آن رسم کردیم. کدهای زده شده برای این قسمت و نیز خروجی آن به شرح زیر می‌باشد:

```
Residence <- Residence[order(Residence$Count),]
```

	Group	Count	percent
1	Rural	2514	49.20
2	Urban	2596	50.80

```
ggplot(Residence, aes(x=Group, y=Count, fill=Group)) +
 geom_bar(stat="identity")+
 geom_text(aes(label=paste0(percent,"%")), vjust=1.6, color="black", size=4)+
 ggtitle("barplot for Residence type")+
 labs(x = "Group", y= "Count")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
```

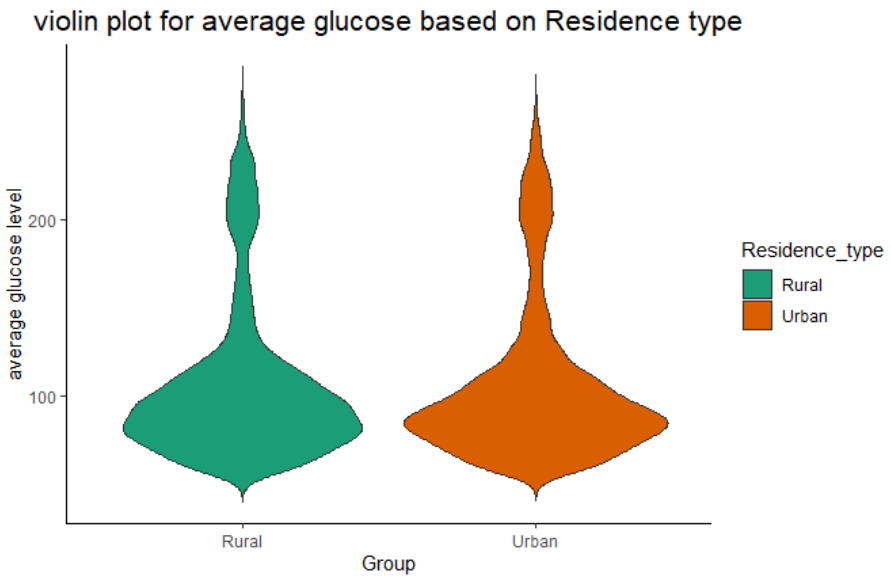
)+ coord_flip()

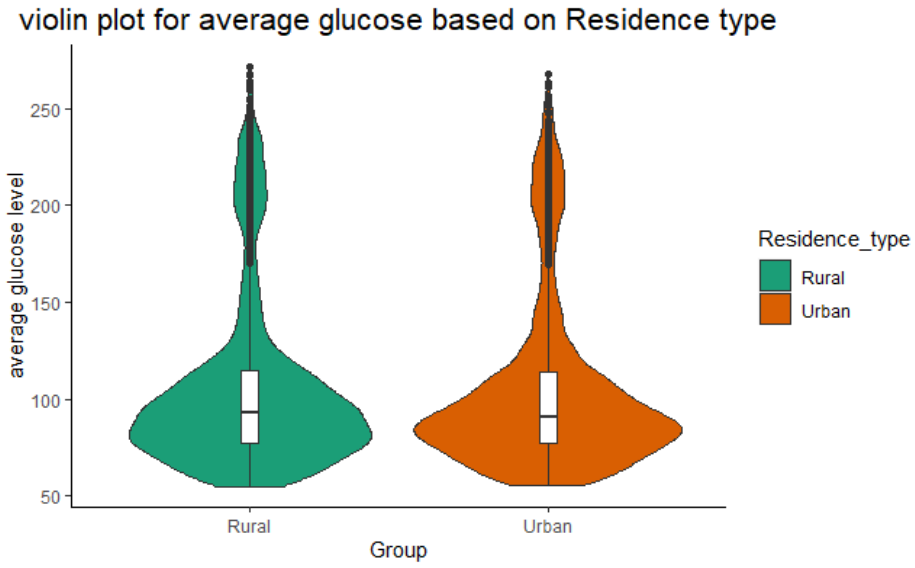


(D

با توجه به آنکه امکان رسم violin plot به تنهای برای متغیر categorical امکان پذیر نمی‌باشد، به همین دلیل به همراه متغیر numerical آن را رسم کرده‌ایم. این متغیر numerical ستون "avg_glucose_level" می‌باشد. کدهای زده‌شده برای این قسمت و نیز نمودار خروجی به شرح زیر می‌باشد:

```
ggplot(data, aes(x=Residence_type, y = avg_glucose_level, fill=Residence_type)) +
 geom_violin()+
 stat_summary(fun.data="mean_sdl", mult=1,
              geom="crossbar", width=0.2 )+
 scale_fill_brewer(palette="Dark2")+
 ggtitle("violin plot for average glucose based on Residence type")+
 labs(x = "Group", y= "average glucose level")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
 )
```





*در تصویر دوم بر روی نمودار violon plot، نمودار boxplot را نیز فیت کرده‌ایم.

Violon plot مشابه boxplot می‌باشد با این تفاوت که، این نمودار ترکیبی از box plot و density plot است. این نمودار نیز مشابه boxplot تفسیر می‌شود. نقاطی که در وسط قرار گرفته بیشترین احتمال را دارند و نقاطی که در انتهای violon plot می‌باشند، oulierها را تشکیل می‌دهندو

Question 3

برای این قسمت از دو ستون avg_glucose_level و bmi به عنوان دو متغیر numerical استفاده شده‌است. البته لازم به ذکر است که در هر دو این ستون‌ها missing value داریم و نمی‌توان از روی ستون‌های دیگر نیز این ستون‌ها را تکمیل کرد.

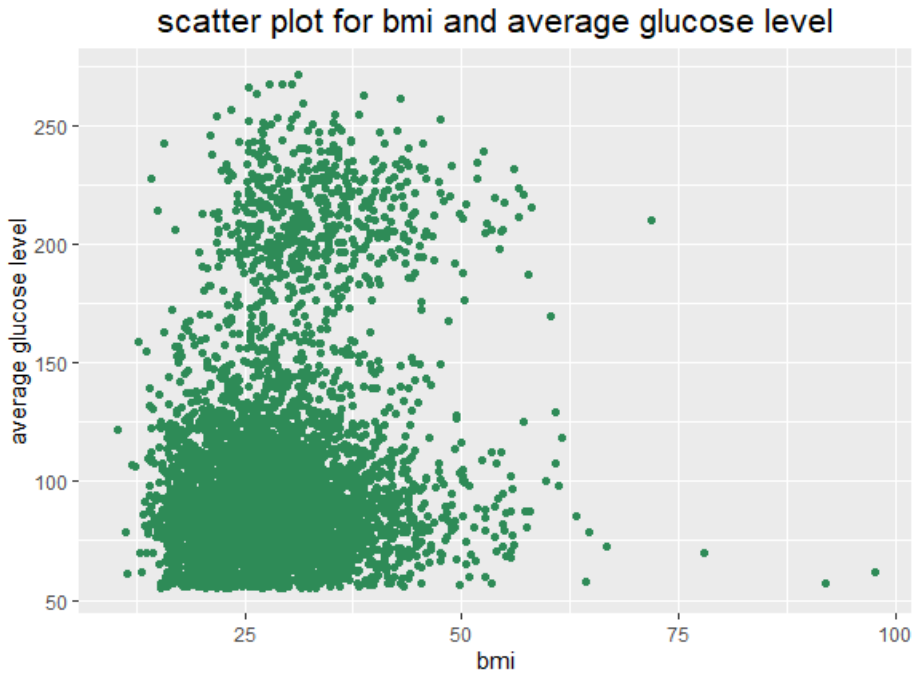
(A

طبق بررسی تقریبا 100 مقدار اول مربوط این دو ستون، در بعضی از سطرها با افزایش مقدار bmi، مقدار avg_glucose_level افزایش یافته، در بعضی نقاط نیز با افزایش bmi مقدار آن یا به مقدار کمی کاهش یافته و یا تغییر خاصی نکرده‌است یا اندکی افزایش یافته‌است. اما تقریبا تعداد نقاطی که با افزایش bmi مقدار avg_glucose_level آن‌ها افزایش یافته بود بیشتر بود، به همین دلیل شاید بتوان گفت تقریبا یک ارتباط مثبت اما از نوع ضعیف بین این دو متغیر برقرار باشد.

(B

برای رسم scatter plot از توابع geom_point موجود در کتابخانه‌ی ggplot2 استفاده‌شده است. کدهای مورد استفاده برای این قسمت و نمودار خروجی به صورت زیر می‌باشد.(لازم به ذکر است که ابتدا سطرهایی را که missing value داشتند با استفاده از دستور data[rowSums(is.na(data)) == 0,] حذف کردیم.)

```
non_miss_df = data[rowSums(is.na(data)) == 0, ]

ggplot(non_miss_df, aes(x = bmi, y= avg_glucose_level)) +
  geom_point(color = "SeaGreen")+
  ggtitle("scatter plot for bmi and average glucose level")+
  labs(x = "bmi", y= "average glucose level")+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
  )
```



(C

به منظور محاسبه‌ی correlation coefficient از تابع cor() استفاده شده‌است. که مقداری که برای این دو متغیر برمی‌گرداند به شرح زیر می‌باشد:

```
> cor(non_miss_df$bmi, non_miss_df$avg_glucose_level)
[1] 0.1755022
```

(D

Correlation coefficient راجب میزان strength و direction ارتباط بین دو متغیر اطلاعات می‌دهد. باتوجه به مقدار به دست‌آمده در 'C'، ارتباط و correlation بین دو متغیر 'bmi' و 'avg_glucose_level' از نوع مثبت بوده، اما این ارتباط ضعیف می‌باشد. زیرا می‌دانیم هرچه این مقدار به عدد '1' نزدیک‌تر باشد، ارتباط قوی‌تر است. اما عددی که برای correlation coefficient بین این دو متغیر به دست‌آوردیم، خیلی با عدد '1' فاصله دارد، بنابراین این ارتباط از نوع ضعیف اما مثبت می‌باشد. جوابی که در 'A' نیز داده شد تقریبا مشابه همین نتیجه‌ای است که با توجه به scatter plot و correlation coefficient به‌دست‌آوردیم.

(E

به منظور تست کردن significance of correlation coefficient ابتدا آزمون فرض را مشخص می‌کنیم.لازم به ذکر است که آزمون فرض را به منظور اینکه تصمیم بگیریم آیا رابطه‌ی خطی بین این دو متغیری که بررسی کردیم در جامعه وجود دارد یا خیر، انجام می‌دهیم

$H_0 : p = 0$ (The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X_1 and X_2 in the population.)

$H_1 : p \neq 0$ (The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X_1 and X_2 in the population.r = 0.176 (sample correlation coefficient)

برای انجام این تست از تابع cor.test() استفاده‌شده است که خروجی آن به شرح زیر می‌باشد:

```
cor.test(data$bmi, data$avg_glucose_level)
```

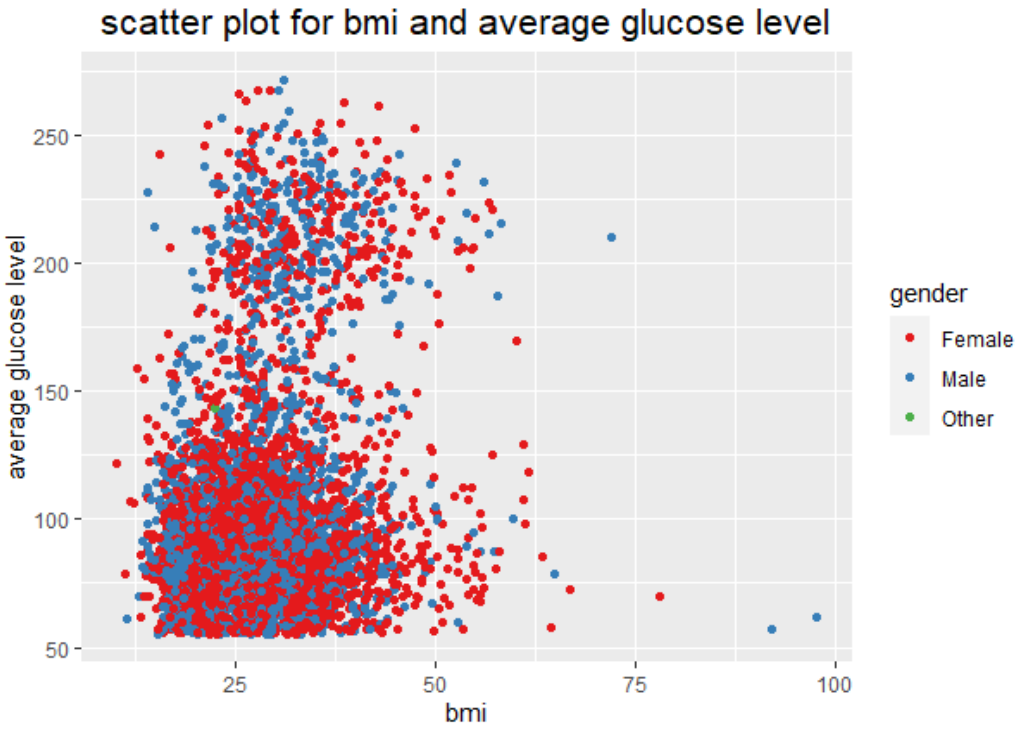
```
        Pearson's product-moment correlation

data:  data$bmi and data$avg_glucose_level
t = 12.488, df = 4907, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1482550 0.2024831
sample estimates:
      cor
0.1755022
```

همانطور که در تصویر و خروجی بالا قابل مشاهده می‌باشد، مقداری که برای p-value به دست‌آمده است بسیار کم‌تر از مقدار $\alpha = 0.05$ می‌باشد، بنابراین H_0 رد می‌شود. به عبارتی دیگر بیان می‌کند که یک رابطه خطی به‌دست آمده بین این دو متغیر به اندازه‌ی کافی strong می‌باشد و در جامعه‌آماری نیز چنین رابطه‌ی خطی وجود دارد.

(F

متغیر Categorical که برای این قسمت در نظر گرفته‌شده‌است، متغیر 'gender' می‌باشد. کدهای زده‌شده برای این سمت و خروجی آن به شرح زیر می‌باشد:

```
ggplot(non_miss_df, aes(x = bmi , y=avg_glucose_level, group = gender, color = gender)) +
 geom_point()+
 scale_color_brewer(palette='Set1')+
 ggtitle("scatter plot for bmi and average glucose level")+
 labs(x = "bmi", y= "average glucose level")+
 theme(
   plot.title = element_text(hjust = 0.5, size = 16),
 )
```



همانطور که در تصویر بالا قابل مشاهده می‌باشد، سه گروه 'female'، 'male' و 'other' (البته لازم به ذکر است که از گروه 'other' تنها یک نمونه فقط وجود دارد) براساس این دو ویژگی 'bmi' و 'avg_glucose_level' خیلی باهم درآمیخته هستند و به عبارتی دیگر قابل تفیک نمی‌باشند و نمی‌توان نتیجه‌ی خاصی گرفت که این دو ویژگی در گروه 'female'با گروه 'male' چه تفاوت‌هایی دارد. در واقع می‌توان برای هر دو گروه تغییرات این دو ستون به یک شکل می‌باشد. البته در گروه مردان همانطور که نمودار بالا قابل مشاهده می‌باشد، دوتا از نمونه‌ها خیلی outlier هستند، این دو نمونه با اینکه bmi آن‌ها افزایش یافته است، اما مقدار avg_glucose_level آن‌ها کاهش یافته‌است. البته از این outlier ها در گروه 'female' هم داریم، اما این دو outlier در گروه مردان نسبت به outlier های دیگر همین گروه و گروه 'female' خیلی برجسته‌تر می‌باشند.

که با استفاده از دستور زیر این دو outlier را پیدا کردیم و خروجی آن به شرح زیر می‌باشد:

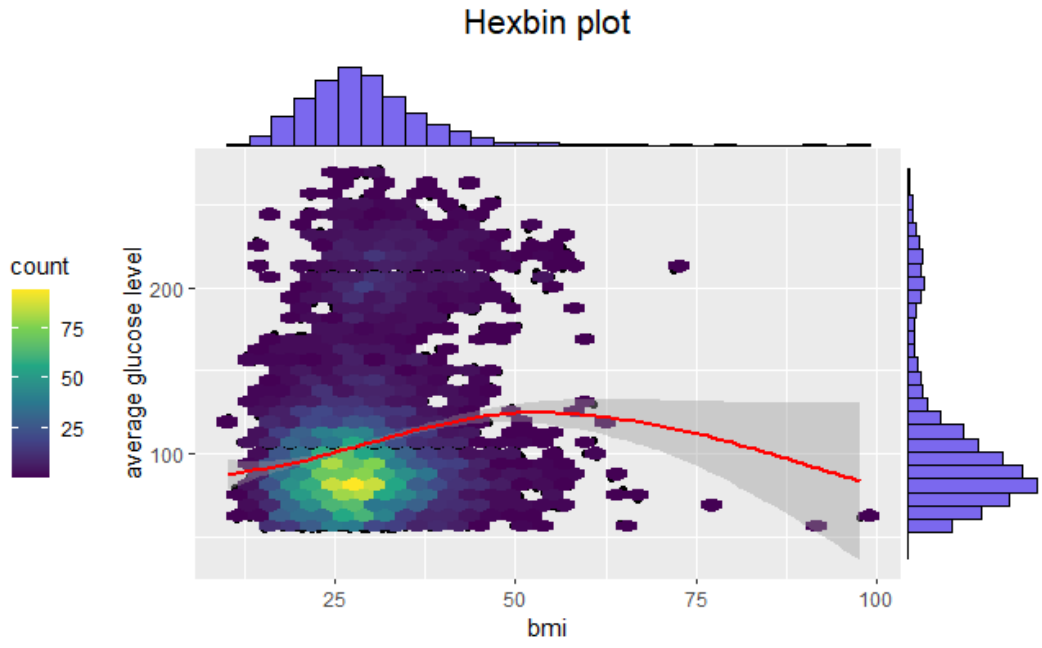
```
View(non_miss_df[which(non_miss_df$bmi > 80),])
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills
2129	56420	Male	17	1	0	No	Private	Rural	61.67	97.6	Unknown	0	5462.386
4210	51856	Male	38	1	0	Yes	Private	Rural	56.90	92.0	never smoked	0	5162.965

(G

روش اول: در قدم اول برای رسم نمودار hexbin از تابع geom_hex() استفاده شده‌است. سپس با استفاده از تابع ggMarginal() موجود در کتابخانه ggExtra ، دو هیستوگرام را بر روی آن فیت کرده‌ایم کدهای این قسمت و خروجی به شرح زیر می‌باشد:

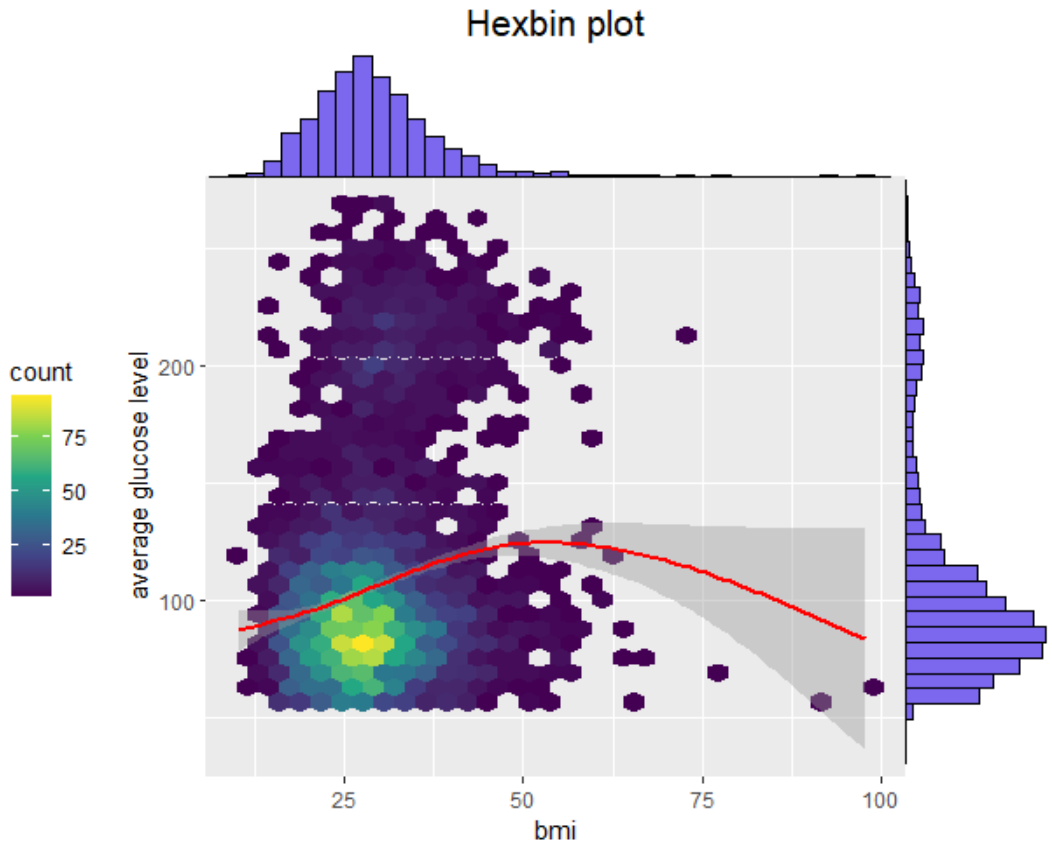
```
library(ggExtra)

hexbin <- ggplot(non_miss_df, aes(x = bmi , y=avg_glucose_level))+
 geom_point()+
 geom_hex(bins = 30)+
 geom_smooth(method = "gam", color = "red", formula = y ~ s(x, bs = "cs"))+
 labs(x = "bmi", y= "average glucose level")+
 ggtitle("Hexbin plot")+
 scale_fill_continuous(type = "viridis")+
 theme(
  legend.position = "left",
  plot.title = element_text(hjust = 0.5, size = 16),
 )

ggMarginal(hexbin,
        type = 'histogram',
        margins = 'both',
        size = 5,
        colour = 'black',
        fill = '#7B68EE')
```



روش دوم: در قدم اول برای رسم نمودار hexbin از تابع geom_hex() استفاده شده‌است. سپس نمودار histogram به ازای هر دو متغیر کشیده‌شده و در نهایت با استفاده از تابع insert_xaxis_grob() و insert_yaxis_grob() نمودارهای histogram در محور x و y نمودار hexbin فیت شده‌اند. درنهایت با استفاده از تابع ggdraw() برای نمایش نمودار نهایی استفاده شده‌است. تمام این توابع مربوط به کتابخانه cowplot می‌باشد. کدهای زده شده برای این قسمت و خروجی آن به شرح زیر می‌باشد:

```
library(hexbin)
library(cowplot)

hexbin <- ggplot(non_miss_df, aes(x = bmi , y=avg_glucose_level))+
 geom_hex(bins = 30)+
 geom_smooth(method = "gam", color = "red", formula = y ~ s(x, bs = "cs"))+
 labs(x = "bmi", y= "average glucose level")+
 ggtitle("Hexbin plot")+
 scale_fill_continuous(type = "viridis")+
 theme(
  legend.position = "left",
  plot.title = element_text(hjust = 0.5, size = 16),
 )

histTop <- axis_canvas(hexbin, axis = "x")+
 geom_histogram(data = non_miss_df, aes(bmi), bins = 40, color = "black", fill="#7B68EE")+
 theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
  axis.title = element_blank(),
  axis.ticks = element_blank(),
  axis.text = element_blank(),
  axis.line.x = element_blank(),
  axis.line.y = element_blank(),
 )

histRight <- axis_canvas(hexbin, axis = "y", coord_flip = TRUE)+
 geom_histogram(data = non_miss_df,aes(avg_glucose_level), bins = 40, color = "black",
fill="#7B68EE")+
 theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
  axis.title = element_blank(),
  axis.ticks = element_blank(),
  axis.text = element_blank(),
  axis.line.x = element_blank(),
```

```
  axis.line.y = element_blank(),
 )+coord_flip()


x <- insert_xaxis_grob(hexbin, histTop, position = "top" )
y <- insert_yaxis_grob(x, histRight, position = "right")

ggdraw(y)
```



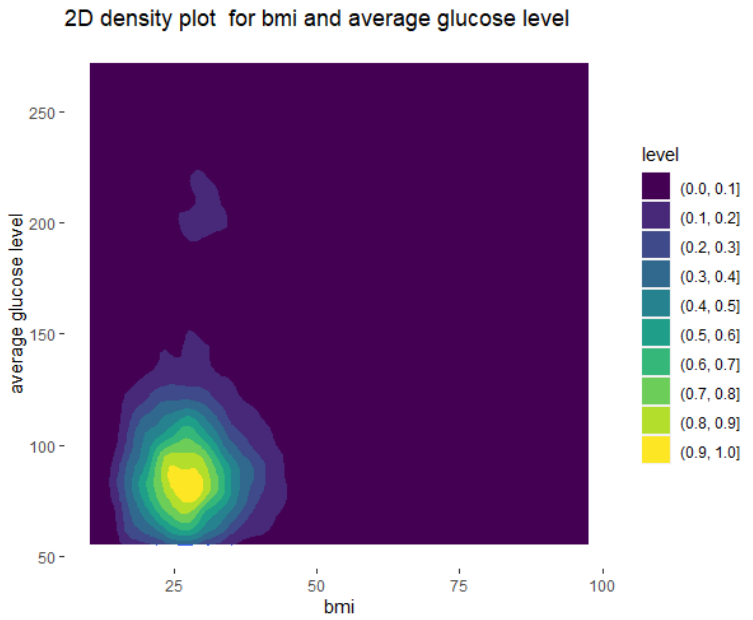
همانطور که در نمودار بالا قابل مشاهده می‌باشد در نقاطی که دو marginal histogram دارای پیک می‌باشد، نمودار hexbin نیز دارای تعداد نقاط بیشتری نسبت به بقیه قسمت‌ها می‌باشد. با توجه به legend هرچه به سمت رنگ زرد پیش می‌رویم تعداد نقاط در آن ناحیه افزایش داشته‌است. اگر بخواهیم به طور جداگانه راجب هر بُعد صحبت کنیم، ابتدا متغیر 'bmi'، همانطور که قابل مشاهده می‌باشد در حدود 25 دارای پیک می‌باشد و در در hexnbin نیز شاهده این افزایش تعداد هستیم زیرا رنگ نمودار به سمت زرد رفته‌است. در متغیر 'avg_glucose_level' نیز در حدود 90 دارای یک پیک هستیم و تعداد زیادی از نقاط نمونه، مقدار میانگین گلوکز آن‌ها حدود 90 است، که در نمودار hexbin نیز در این حدود همانطور که شاهد افزایش تعداد هستیم، زیرا رنگ نمودار به سمت زرد رفته‌است. با استفاده از پارامتر bin موجوددر تابع geom_hex() می‌توان اندازه هر hex را مشخص کرد، همانطور که در histogram با افزایش سایز bin تعداد نقاطی که در یک bin قرار می‌گیرند کاهش یافته به نوعی که اگر خیلی طول bin را کم کنیم هر دیتا پوینت در یک bin می‌افتد ب عبارتی ارتفاع همه نقاط موجود یک می‌شود و یک سری نقاط وجود نخواهد داشت. همچنین هرچه سایز آن را کم کنیم تعداد دیتا پوینت‌هایی

که در یک نقطه می‌افتد، افزایش می‌یابد به طوری که اگر طول bin را خیلی زیاد در نظر بگیریم، کل دیتاپوینت روی یک bin خواهد افتاد. در hexbin نیز به همین ترتیب است هرچه طول bin را بزرگ بگیریم اندازه hex بزرگ شده و تعداد زیادی از نقاط نمونه را در بر می‌گیرد و هرچه طول bin را کوچک بگیریم، اندازه‌ی hexها کوچک‌تر شده و تعداد کمی از نقاط نمونه را در بر می‌گیرد. به طور کلی تغییر سایز bin می‌تواند نتیجه‌ای که می‌گیریم را تغییر دهد زیرا این نمودار به سایز bin حساس می‌باشد. سایزی که برای bin چه در histogram و چه در hexbin در کد بالا در نظر گرفته شده برابر 30 می‌باشد و به عبارتی دیگر طول bin(bin width) برابر 0.3 می‌باشد.

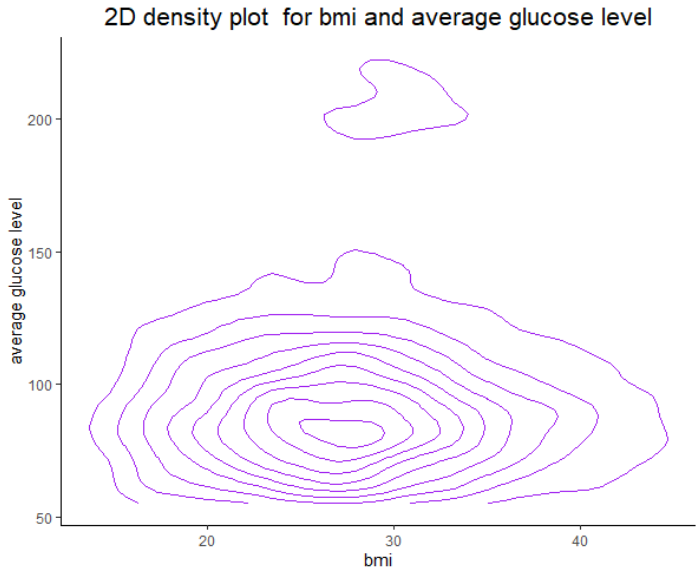
(H

به منظور رسم 2D density plot از تابع geom_density_2d() استفاده شده‌است. همچنین در نمودار (1) ، از تابع geom_density_2d_filled(contour_var = "ndensity") نیز استفاده شده‌است. کدهای زده شده برای این قسمت و خروجی به شرح زیر می‌باشد:

```
#1
ggplot(non_miss_df, aes(x = bmi , y=avg_glucose_level))+
 geom_density_2d()+
 geom_density_2d_filled(contour_var = "ndensity")+
 theme(
   panel.grid.major = element_blank(),
   panel.grid.minor = element_blank(),
   panel.background = element_blank(),
   axis.line = element_blank(),
 )


#2
ggplot(non_miss_df, aes(x = bmi , y=avg_glucose_level))+
 geom_density_2d()+
 theme(
   panel.grid.major = element_blank(),
   panel.grid.minor = element_blank(),
   panel.background = element_blank(),
   axis.line = element_blank(),
 )
```



شکل(1)



شکل(2)

تفسیر این دو نموداری که برای 2D density plot رسم شده‌است، مشابه همان scatter plot می‌باشد. همانطور که قابل مشاهده می‌باشد قسمت زیادی از دیتا در محدوده‌ای قرار دارند که اندازه 'bmi' آن حدود 25 و 'avg_glucose_level' آن حدود 100 می‌باشد زیرا همانطور که قابل مشاهده می‌باشد در این نقاط چگالی افزایش یافته‌است. مابقی دیتا نیز در همین حدود می‌باشندو فقط تعداد کمی از آن‌ها در قسمت بالایی نمودار می‌افتند همانطور که در شکل قابل مشاهده می‌باشد. که این نقاط به نوعی می‌توان گفت outlierها می‌باشند.

مزیت استفاده از hexbin و 2D density زمانی است که می‌خواهیم رتباط بین دو متغیر numerical را زمانی که تعداد زیادی نقطه(نمونه) وجود دارد و خطر overplotting در scatterplot داریم بررسی کنیم. به عبارتی زمانی که تعداد زیادی نقطه داریم و مقدار دقیق density ها در آن نقاط با استفاده از scatterplot مشخص نمی‌شود از این دو نمودار استفاده می‌کنیم. که با استفاده از ین نمودارها توزیع نقاط به طور واضح‌تری نمایش داده می‌شوند. در غیر این صورت اگر چنین خطری وجود نداشت از همان scatter plot استفاده می‌کنیم زیرا در این مواقع scatter plot ،effective بیشتری دارد.

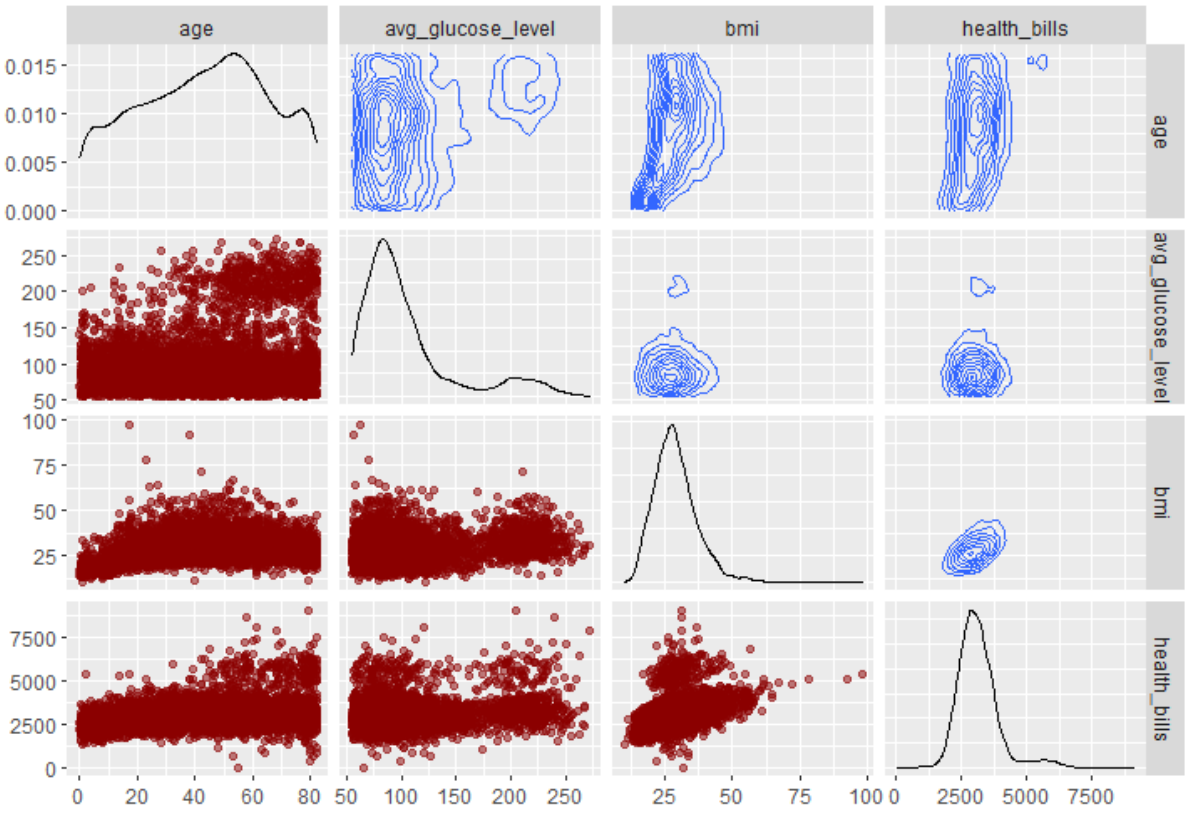
همچنین برای visualization چندین توزیع همزمان، hexbin به طور کلی بهتر از هیستوگرام ها کار می کنند.

Question 4

(A

ابتدا ستون‌های numerical موجود در دیتاست را جدا کرده و به عنوان یک دیتافریم جدید ذخیره می‌کنیم. متغیرهای numerical در مجموعه‌داده مورد بررسی عبارتند از : age، bmi، avg_glucose_level و Health_bills. همچنین به عنوان پیش‌پردازش بعدی سطرهایی از این ستون‌ها که missing value داشتند و قابل محاسبه از روی ستون‌های دیگر نبودند را حذف کردیم. سپس برای رسم correlogram از تابع ggpairs() موجود در کتابخانه GGally استفاده کرده‌ایم. کدهای این قسمت و خروجی به شرح زیر می‌باشد:

```
library(GGally)

numeric_data <- non_miss_df[, c(3, 9, 10, 13)]
ggpairs(numeric_data, progress = F,
        lower = list(continuous = wrap("points", color = "darkred", alpha = 0.5)),
        upper = list(continuous = wrap("density")))
```



علاوه بر رسم scatter plot() برای هر دو متغیر، 2D density plot نیز برای هر دو متغیر رسم شده که در بالای قطر قابل مشاهده می‌باشد. قطر اصلی نیز نماینگر نمودار density برای هر متغیر می‌باشد.

طبق نمودارهای به دست‌آمده، همانطور که مشاهده می‌فرمایید تمامی متغیرها دو به دو باهم ارتباط و correlation مثبت دارند و هیچ گونه correlation منفی بین هیچ دو متغیری را شاهد نمی‌باشیم. البته این ارتباط در بعضی از متغیرهای قوی‌تر و در بعضی‌ها ضعیف‌تر می‌باشد. به عنوان مثال بین دو متغیر avg_glucose_level و age ارتباط خیلی ضعیف بوده و correlation coefficient بین این دو متغیر بیشتر به 0 نزدیک می‌باشد، زیرا همانطور که از نمودار قابل مشاهده می‌باشد داده ها خیلی پراکنده هستند و روی یک خط با شیب مثبت متمرکز نشده‌اند. همچنین می‌توان با توجه به نمودارهای بالا گفت که قوی‌ترین correlation به احتمال زیاد مربوط به Health_bills و bmi می‌باشد. زیرا به نسبت بقیه متغیرهای بیشتر روی خط متمرکز می‌باشد.

روی قطر اصلی نیز توزیع هر متغیر نمایش داده شده‌است. که توزیع مربوط به bmi یک توزیع unimodal و right skewed می‌باشد، توزیع Health bills نیز مشابه توزیع bmi . توزیع سن نیز با اغماض یک توزیع symmetric می‌باشد.

(B

برای رسم correlogram heatmap از تابع ggcorrplot() استفاده شده‌است. ابتدا با استفاده cor() مقدار correlation coefficient بین هر دو متغیر را حساب کرده و سپس با استفاده از تابع cor_pmat() مقدار p-value را برای هر دو متغیر محاسبه کرده و سپس heatmap را براساس این مقادیر رسم کردیم. کدهای این قسمت و خروجی آن به شرح زیر می‌باشد: (لازم به ذکر است که مشابه بالا سطرهایی که missing value داشتند و قابل محاسبه از روی ستون‌های دیگر نبودند را حذف شده‌اند به این دلیل که با وجود آن سطرها نمی‌توان correlation coefficient را محاسبه کرد.)

```
cormat <- round(cor(numeric_data,  method = "pearson"), 2)
```

```
                    age avg_glucose_level  bmi health_bills
age                1.00              0.24 0.33         0.30
avg_glucose_level  0.24              1.00 0.18         0.18
bmi                0.33              0.18 1.00         0.42
health_bills       0.30              0.18 0.42         1.00
```

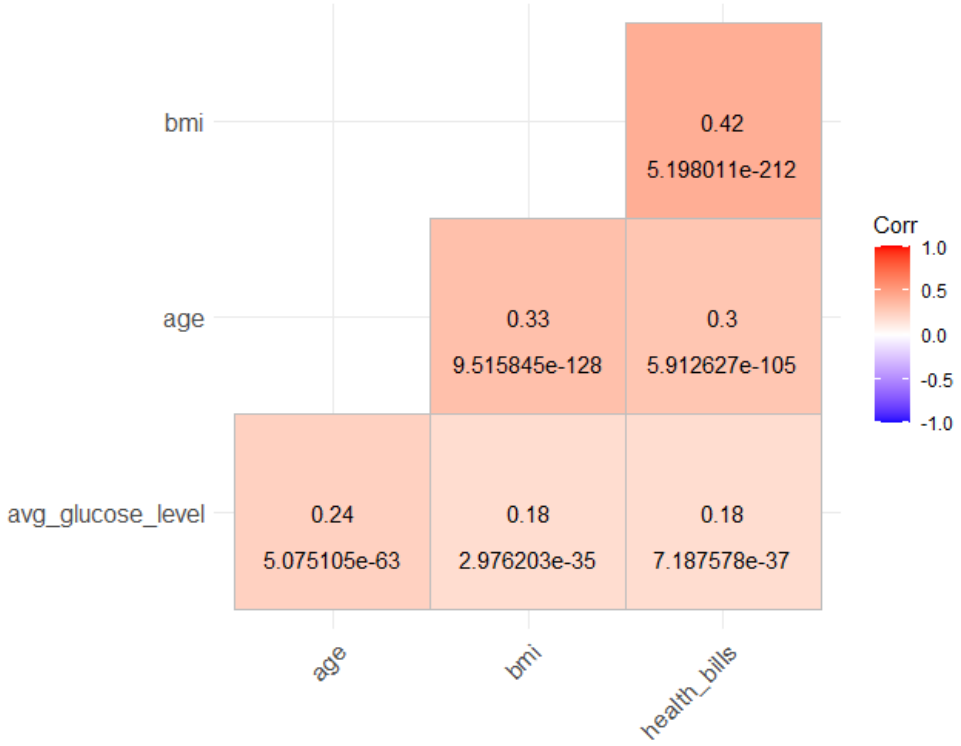
جدول بالا نمایانگر مقدار correlation coefficient بین هر دو متغیر می‌باشد.

```
p.mat <- cor_pmat(numeric_data)
```

```
                          age avg_glucose_level          bmi  health_bills
age               0.000000e+00      5.075105e-63 9.515845e-128 5.912627e-105
avg_glucose_level 5.075105e-63      0.000000e+00  2.976203e-35  7.187578e-37
bmi               9.515845e-128     2.976203e-35  0.000000e+00 5.198011e-212
health_bills      5.912627e-105     7.187578e-37 5.198011e-212  0.000000e+00
```

جدول بالا نمایانگر مقدار p-value براساس cor.test() بین هر دو متغیر می‌باشد.

```
ggcorrplot(cormat, hc.order = TRUE, type = "lower",
           lab = TRUE, p.mat =p.mat, sig.level = .05)+
 geom_text(aes(label = c(5.075105e-63, 2.976203e-35, 7.187578e-37,
                         9.515845e-128, 5.912627e-105, 5.198011e-212)),
           vjust = 3, hjust = 0.5)
```



مقداری که در بالای هر خانه از ماتریس نمایش داده شده‌است، مقدار correlation coefficient بین هر دو متغیر را نمایش می‌دهد و مقداری که در پایین هر خانه نوشته شده‌است نمایانگر مقدار p-value به دست آمده از اجرای تابع cor.test() به ازای هر دو متغیر می‌باشد. باتوجه به مقادیری که برای p-value به دست‌آمده، تمام آن‌ها از $\alpha = 0.05$ کوچکتر بوده و به همین دلیل فرض H_0 را رد می‌کنیم.

فرض H_0 به ازای هر دو متغیر به صورت زیر تعریف شده‌است:

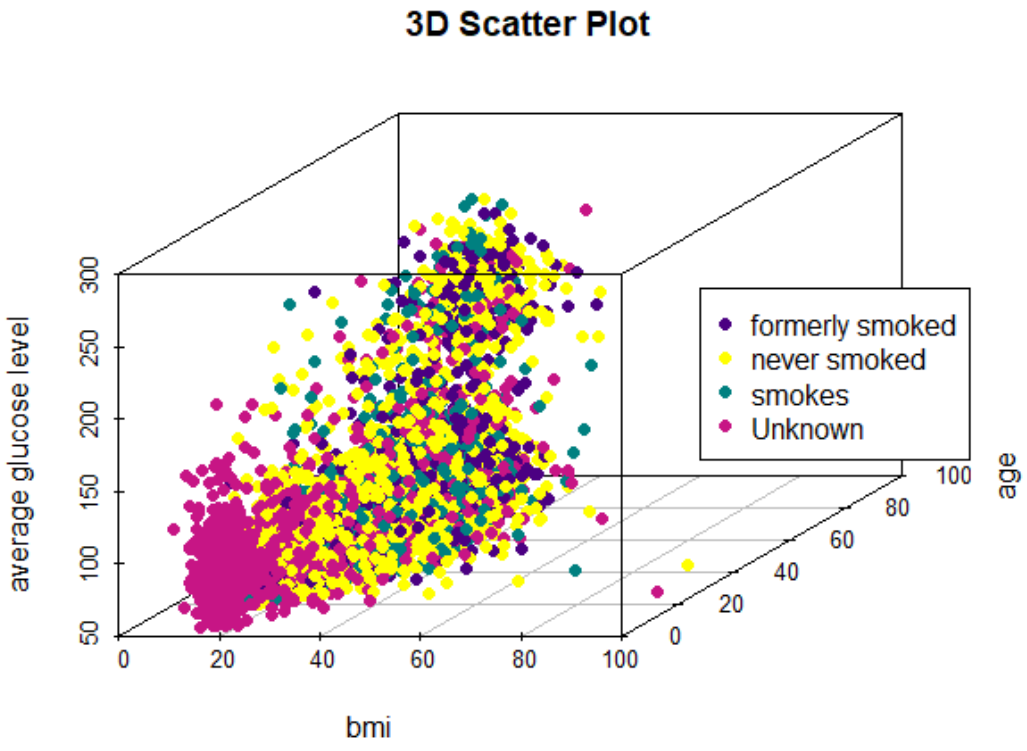
$H_0 : p = 0$ (The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X_1 and X_2 in the population.)

$H_1 : p \neq 0$ (The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X_1 and X_2 in the population.r = 0.176 (sample correlation coefficient)

(C

برای رسم 3D scatter plot از تابع scatterplot3d() استفاده شده‌است. این نمودار برای سه متغیر numerical که عبارتند از 'age'، 'bmi' و 'avg_glucose_level' رسم شده‌است که با توجه به متغیر 'smoking_status' رنگ‌آمیزی شده‌اند. کدهای این قسمت و خروجی آن به شرح زیر می‌باشد:

```
scatterplot3d(x = non_miss_df$bmi,
              y = non_miss_df$age,
              z = non_miss_df$avg_glucose_level,
              pch = 16, color = cols[group],
              main="3D Scatter Plot",
              xlab = "bmi",
              ylab = "age",
              zlab = "average glucose level")

legend("right", legend = levels(as.factor(non_miss_df$smoking_status)),
    col = cols, pch = 16)
```



همانطور که در تصویر بالا قابل مشاهده می‌باشد، گروهی از افراد که وضعیت سیگاری بودن خود را 'unknown' زده‌اند، تجمع بیشتری در نقاطی دارند که سن آن‌ها پایین بوده و نیز bmi و متوسط سطح گلوکز آن‌ها نیز پایین است. گروهی که 'never smoked' می‌باشند، دارای توزیع پراکنده‌ای هستند و تقریبا تمام رنج سنی و تمام رنج متوسط سطح گلوکز را شامل می‌شوند و رنج bmi برای این افراد تقریبا حدود 17 تا 60 می‌باشد.

افرادی که در دسته‌بندی 'formerly smoked' قرار گرفته‌اند، افراد 15 سال به بالا نمونه را شامل می‌شوند و رنج bmi آن‌ها حدود 17 تا 55 می‌باشد و متوسط سطح گلوکز آن‌ها میتوان گفت که تمام رنجی که تعریف شده‌است را در برمی‌گیرد.

در نهایت افرادی که در درسته‌بندی 'smoked' قرار گرفته‌اند، تقریبا می‌توان گفت که از رنج سنی 25 سال هستند و رنج bmi آن‌ها حدود 15 تا 45 می‌باشد و متوسط سطح گلوکز آن‌ها میتوان گفت که تمام رنجی که تعریف شده‌است را در برمی‌گیرد.

Question 5

دو متغیر cateforical که برای این قسمت استفاده شده‌است، عبارت است از: gender و smoking_type

(A

با استفاده از تابع table() ابتدا جدول contingency را ایجاد کرده و سپس با استفاده از تابع addmargins() مقدار نهایی در هر سطر و ستون را به تمام سطر و ستون‌ها اضافه کردیم. کدهای این قسمت و خروجی آن به شرح زیر می‌باشد:

```
con_table <- table(data$gender, data$smoking_status)
addmargins(con_table)
```

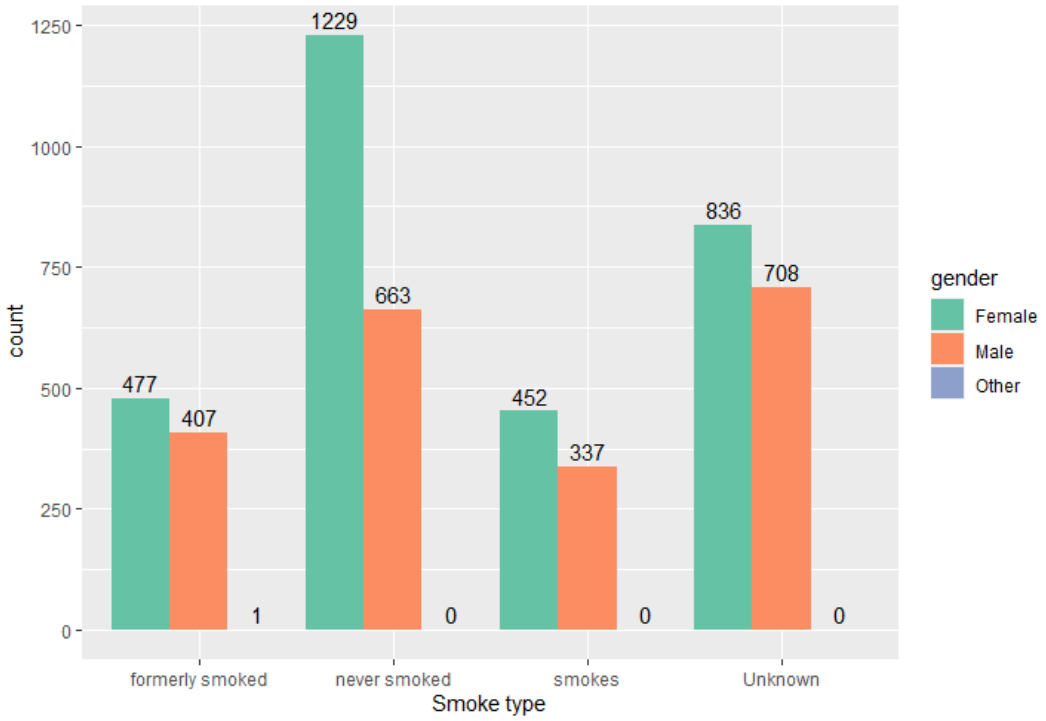
	formerly smoked	never smoked	smokes	Unknown	Sum
Female	477	1229	452	836	2994
Male	407	663	337	708	2115
Other	1	0	0	0	1
Sum	885	1892	789	1544	5110

این جدول نشان‌دهنده‌ی این است که به عنوان مثال در گروه آقایان، چند نفر سیگار می‌کشند، چند نفر گاهی اوقات سیگار می‌کشند، چند نفر هرگز سیگار نمی‌کشند و کسانی که به این سوال جواب نداده‌اند. به همین ترتیب نیز تعداد این فراد در گروه‌های دیگر قابل مشاهده می‌باشد.

(B

به منظور رسم bar chart نیز از توابع موجود در کتابخانه‌ی ggplot2 استفاده‌شده‌است. کدهای مربوط به این قسمت و خروجی آن به شرح زیر می‌باشد:

```
ggplot(df_2Catg, aes(x = smoke_type, y = count, fill = gender))+
 geom_bar(stat="identity", position=position_dodge())+
 scale_fill_brewer(palette = "Set2")+
 labs(x = "Smoke type" , y = "count")+
 geom_text(aes(label=count), vjust=-0.4, color="black", size=4,
           position = position_dodge(.9))
```

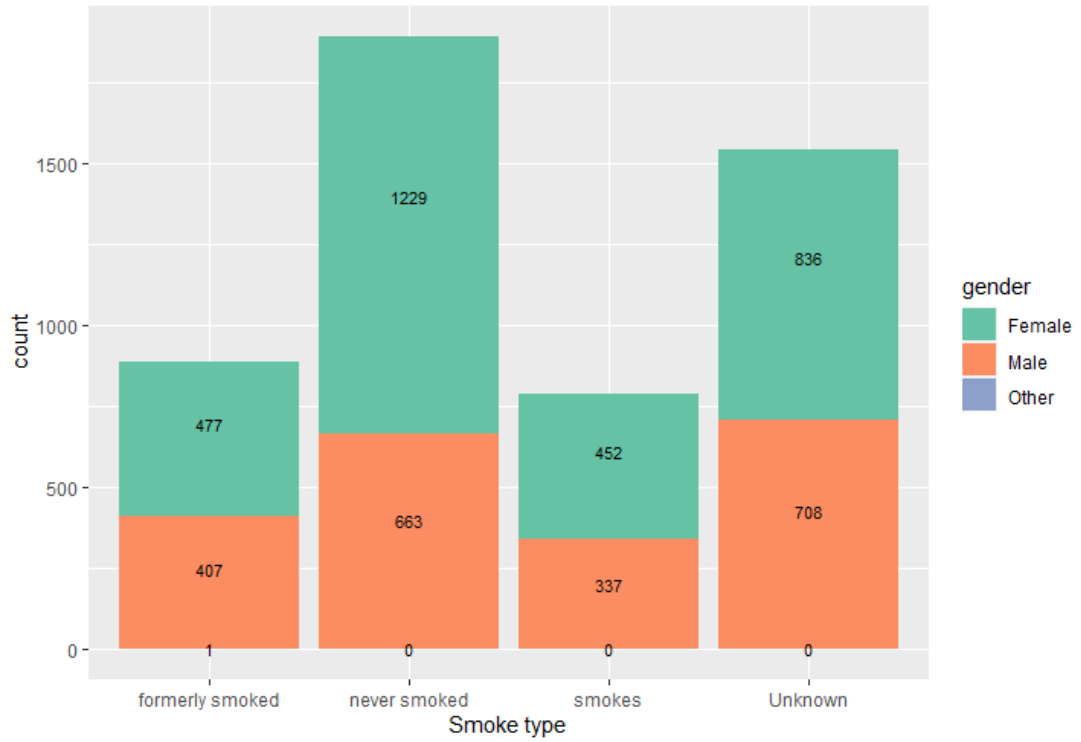


این bar chart مشخص می‌کند در بین گروه‌های 'female'، 'male' و 'other' چند نفر در دسته از smoking status قرار گرفته‌اند.

(C

به منظور رسم segmented bar plot نیز از توابع موجود در کتابخانه‌ی ggplot2 استفاده‌شده‌است. کدهای مربوط به این قسمت و خروجی آن به شرح زیر می‌باشد:

```
ggplot(df_2Catg, aes(x = smoke_type, y = count, fill = gender, label = count))+
  geom_bar(stat="identity")+
  geom_text(size = 3, position = position_stack(vjust = 0.6))+
  scale_fill_brewer(palette = "Set2")+
  labs(x = "Smoke type" , y = "count")
```



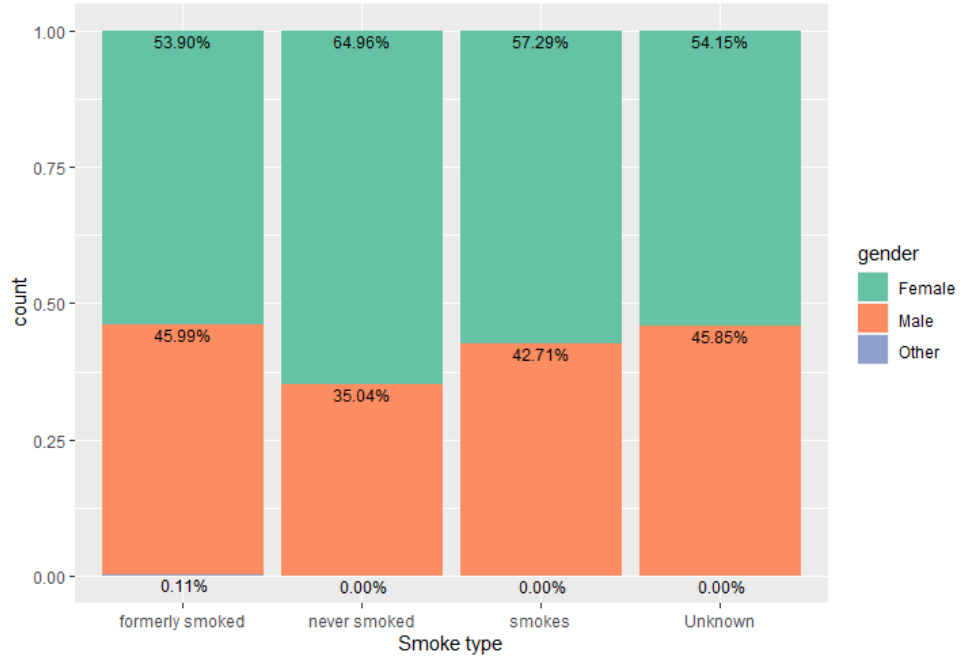
هر bar در این نمودار مشخص می‌کند در بین گروه‌های 'female'، 'male' و 'other' چند نفر در دسته از smoking status قرار گرفته‌اند.

(D

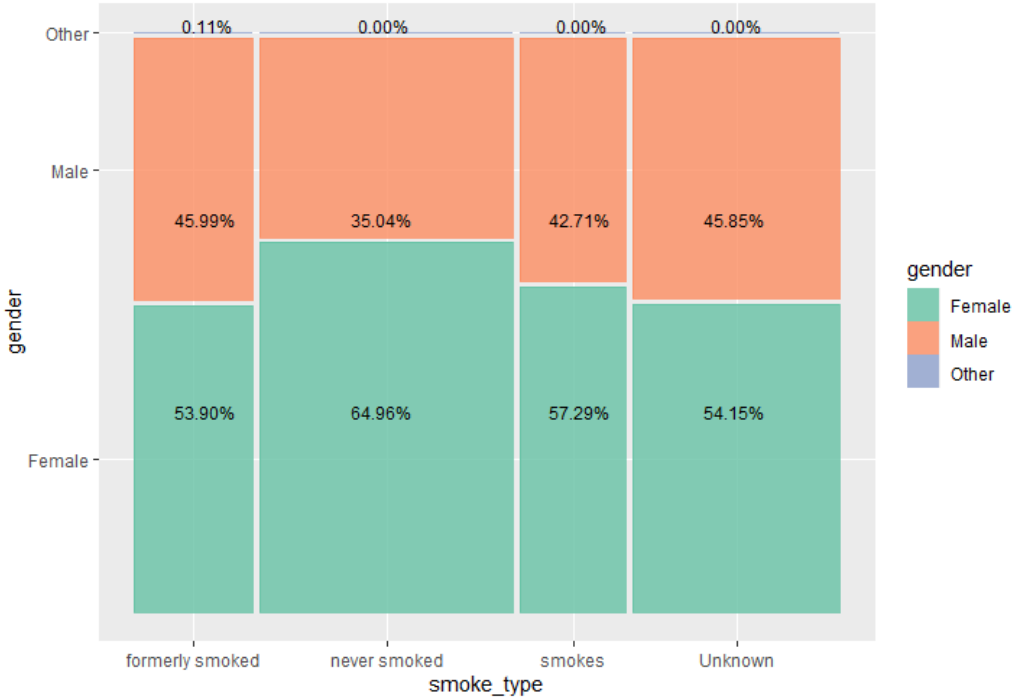
به منظور رسم mosaic plot از دو تابع استفاده شده‌است که خروجی شکل(1) مشابه آن نموداری است که در صورت سوال وجود دارد و با استفاده از توابع موجود در کتابخانه‌ی ggplot2 کشیده شده‌است. شکل(2) نمونه‌ی دیگری از mosaic plot می‌باشد که با استفاده از کتابخانه‌ی ggmosaic کشیده شده‌است.کدهای مربوط به این قسمت و خروجی آن به شرح زیر می‌باشد:

```
df_2Catg$per <- format(round(data.frame(prop.table(con_table, margin = 2))$Freq
                             *100, 2), nsmall = 2)

#sol1
ggplot(df_2Catg, aes(x =smoke_type, y = count, fill = gender))+
  geom_bar(stat="identity",position = "fill")+
  scale_fill_brewer(palette = "Set2")+
  geom_text(aes(label= paste0(per, '%')), vjust=1.2, color="black", size=3.2,
            position = "fill")+
  labs(x = "Smoke type" , y = "count")

#sol2
ggplot(df_2Catg)+
```

```
geom_mosaic(aes(x = product(smoke_type), weight = count, fill = gender))+
scale_fill_brewer(palette = "Set2")+
geom_text(aes(y =1,
          x = as.numeric(rep(c(0.1, 0.35, 0.63, 0.85), each = 3)),
          label= paste0(per, '%')), vjust=0, color="black", size=3.2,
        position = "fill")
```



شکل(1)



شکل(2)

Question 6

(A

ابتدا از مجموعه‌داده یک سمپل انتخاب کرده و بر روی متغیر 'health_bills' این نمونه، موارد خواسته شده را بررسی می‌کنیم.
با استفاده از دستور زیر نمونه را از مجموعه داده به صورتی که در بالا توضیح داده شد، انتخاب میکنیم:

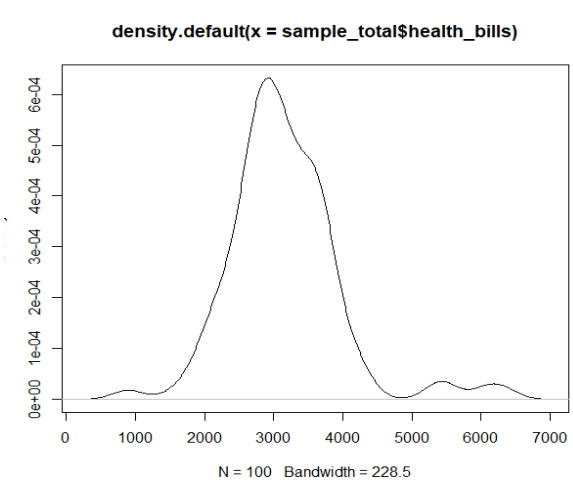
```
sample_total <- non_miss_df[sample(nrow(non_miss_df), 100),]
```

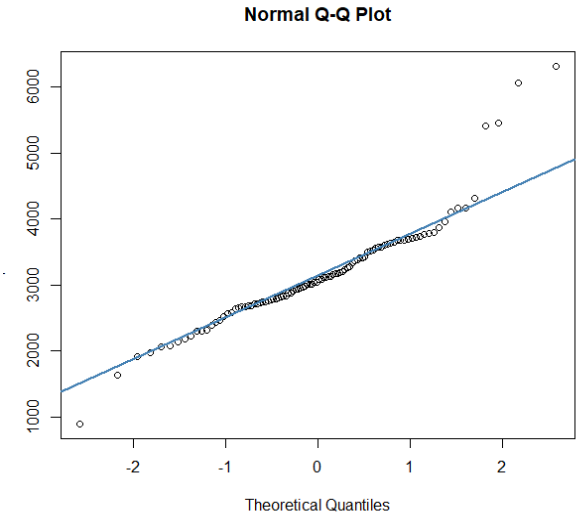
در قدم اول باید شرایط حد مرکزی را بررسی کنیم:

1- این مجموعه داده به صورت رندم از جامعه انتخاب شده‌است و نیز 100 نفر از 10% کل جامعه کمتر می‌باشد.(زیرا اعضای کل جامعه برابر 5110 نفر می‌باشند.که البته با حذف سطرهای missing value تعداد برابر 4909 می‌باشد.)

2- سایز سمپل مورد بررسی از 30 بیشتر می‌باشد و نیز توزیع سمپل تقریبا متقارن بوده و چولگی زیادی نداریم.

Density plot و Q-Q plot این توزیع در تصویر پایین قابل مشاهده می‌باشد:





حال که شرایط حد مرکزی برقرار شد به محاسبه‌ی بازه‌ی اطمینان می‌پردازیم. به منظور محاسبه‌ی بازه اطمینان 95%، از دو راه آن را به دست‌آوردیم:

1- مطابق فرمولی که در درس بود، پیش رفتیم.

$$CI = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

```
SE = sd(sample_total$health_bills)/sqrt(length(sample_total$health_bills))
conf_left = mean(sample_total$health_bills) - qnorm(0.95) * SE
conf_right = mean(sample_total$health_bills) + qnorm(0.95) * SE
```

خروجی کد بالا به شرح زیر می‌باشد:

CI = (2921, 3145.3)

2- روش دوم با استفاده از تابع CI() موجود در کتابخانه Rmisc می‌باشد. کد و خروجی به شرح زیر می‌باشد:

```
library(Rmisc)

ci <- CI(sample_total$health_bills, ci = 0.95)
```

CI = (2898, 3168.4)

(B

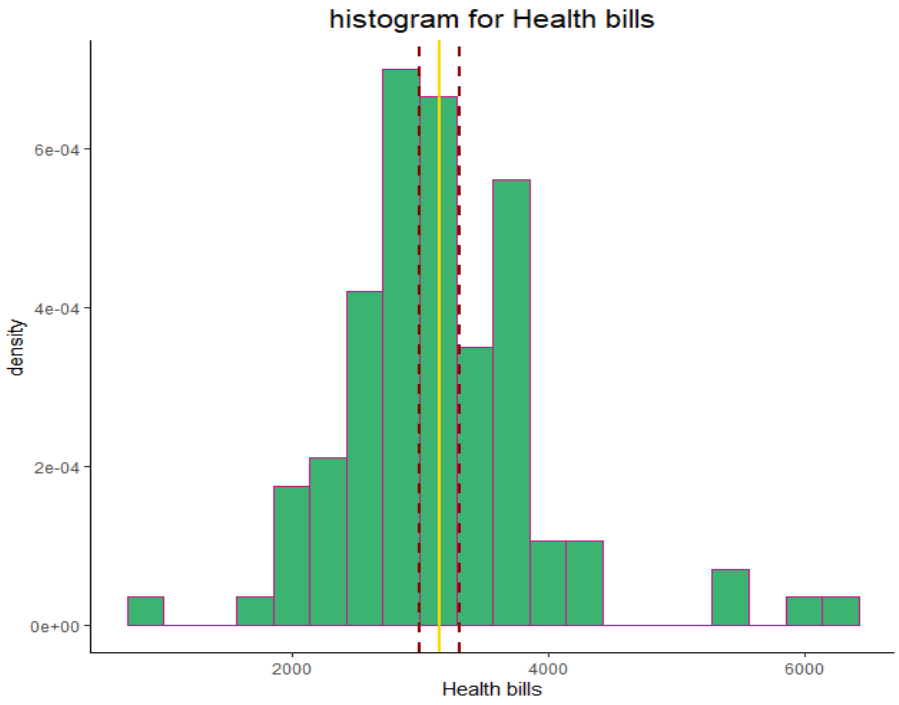
در مجموعه داده‌ای که داریم و ستونی که برای بررسی انتخاب کردیم هزینه‌ای که هر فرد سالیانه برای مراقبت‌های پزشکی خود می‌پردازند، نوشته شده‌است. حال بازه‌ی اطمینان 95%ای که در بالا به دست آمد، بیان می‌کند که ما 95% اطمینان داریم که افراد جامعه(جامعه‌ای که این افراد از آن به صورت رندم انتخاب شده‌اند) هزینه‌ای که برای مراقبت‌های پزشکی خورد به طور سالیانه می‌پردازند، به طور میانگین در بازه‌ی (2898, 3168.4) قرار می‌گیرد. یا به عبارت دیگر 95% random sample هایی که با سایز 100 از جامعه برمی‌داریم منجر به بازه‌اطمینان‌هایی می‌شود که میانگین واقعی جامعه برای هزینه-مراقبت‌های پزشکی را در برمی‌گیرد. یا به بیانی دیگر اگر نمونه‌های مختلف با سایز 100 از از جامعه برداریم و با هرکدام از این‌ها یک بازه‌ی اطمینان بسازیم، آنگاه 95% این بازه‌های اطمینان، میانگین واقعی را شامل می‌شود.

(C

برای رسم هیستوگرام و نیز فیت کردن میانگین و بازه‌ی اطمینان بر روی آن از توابع geom_histogram و geom_vline، استفاده‌شده‌است. کدهای زده‌شده برای این قسمت و خروجی آن به شرح زیر می‌باشد:

```
ggplot(sample_total, aes(x =health_bills))+
 geom_histogram(aes(y=..density..),bins = 20, color = "MediumVioletRed", fill="MediumSeaGreen")+
 geom_vline(aes(xintercept = mean(health_bills)), col = "Gold",
        size = 1)+
 geom_vline(aes(xintercept = ci["upper"]), col = "darkred",
        linetype = "dashed",size = 1)+
 geom_vline(aes(xintercept = ci["lower"]), col = "darkred",
        linetype = "dashed", size = 1)+
 ggtitle("histogram for Health bills")+
 labs(x= "Health bills")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.background = element_blank(),
  axis.line = element_line(colour = "black"),
```

)



خط زرد موجود در تصویر بالا نشان‌دهنده‌ی میانگین متغیر برای سمپل انتخاب شده و نیز خطوطی که با خطچین نمایش داده‌شده‌است نشان دهنده‌ی بازه‌ی اطمینان می‌باشد. مقدار میانگین برای سمپل مورد بررسی برابر 3033.101 می‌باشد.

(D

آزمون فرضی که برای این متغیرکه میزان هزینه‌ای که هر فرد سالیانه برای مراقبت‌های سلامتی خود می‌پردازد، را نشان می‌دهد، در نظر گرفتیم به شرح زیر می‌باشد:

ادعا می‌کنیم که متوسط هزینه‌ای برای هر فرد برابر 3100 می‌باشد و فرض مقابل را به این صورت تعریف می‌کنیم که متوسط هزینه بیشتر از 3000 می‌باشد.

$H_0 : \mu = 3100$

$H_A : \mu > 3100$

p-value ($\bar{x} > 3033.101 \mid H_0 : \mu = 3100$)

$\rightarrow \bar{x} \sim N(\mu = 3100,\ SE = \frac{s}{\sqrt{n}} = 60.14)$

Test statistic : $Z = \frac{3033.101 - 3100}{60.14} = -0.98$

p-value = $p(Z > -0.98) = 0.84$

اگر مقدار α را به طور پیش فرض برابر 0.05 در نظر بگیریم، داریم $\alpha <$ p-value. آنگاه H_0 را نمی‌توان رد کرد. به عبارتی تلویحا می‌گوییم که میانگین هزینه‌ای که هر فرد در جامعه به طور سالیانه برای مراقبت‌های پزشکی خود می‌پردازد برابر 3100 می‌باشد. همچنین چون اختلاف p-value و α زیاد می‌باشد، با اطمینان بیشتری فرض H_0 را نمی‌توانیم رد کنیم.

مراحل محاسبه p-value در R نیز به شرح زیر می‌باشد:

```
z_stat <- (mean(sample_total$health_bills) - 3000) / SE
p_val <- pnorm(z_stat, lower.tail = FALSE)
```

(E

بازه‌ی اطمینانی که در مورد 'A' به دست‌آوردیم برابر (2898, 3168.4) شد. این بازه‌ی اطمینان، فرض H_0 را شامل می‌شود به عبارتی مقداری که برای میانگین health_bills در فرض صفر در نظر گرفتیم، برابر3100 بود که این مقدار در این بازه‌اطمینان می‌افتد. به همین دلیل H_0 را می‌پذیریم. مشابه نتیجه‌ای که در مورد 'D' با محاسبه p-value به دست‌آوردیم.

(F

برای محاسبه‌ی type II error با استفاده از R به این صورت عمل شده‌است که critical value را نسبت به α به دست‌آورده و بعد با توجه به actual mean که داریم power را محاسبه‌کرده و چون می‌دانیم مقدار خطای نوع دوم (β) برابر 1 – power می‌باشد، مقداری خطای نوع دوم را به‌دست آوردیم. کدهای این قسمن و خروجی آن به شرح زیر می‌باشد:
لازم به ذکر است مقداری که برای actual mean در نظر گرفته شده‌است برابر میانگین 'health_bills' مجموعه‌داده اصلی که به عنوان جامعه اصلی درنظر گرفتیم می‌باشد.

μ_a = 3138.585

```
mu0  <- 3100
mua  <- mean(non_miss_df$health_bills)

alpha <- 0.05

# critical value for a level alpha test
crit <- qnorm(1-alpha, mu0, SE)

# power: probability for values > critical value under H1
(pow <- pnorm(crit, mua, SE, lower.tail=FALSE))


# probability for type II error: 1 - power
(beta <- 1-pow)
```

Type II error : 0.86

همانطور که می‌دانیم خطای نوع دوم به این معنی است که اگر فرض H_A صحیح باشد ولی ما نتوانیم H_0 را رد کنیم. به عبارتی تلویحا بگوییم H_0 صحیح است، آنوقت خطای نوع دوم اتفاق افتاده‌است. مقداری که برای خطای نوع دوم به دست‌آوردیم تقریبا برابر 86% می‌باشد، که به این معناست 86% احتمال دارد ما فرض H_0 (که بیان می‌کرد میانگین health bills در جامعه برابر 3100 است) را نتوانیم رد کنیم ولی در اصل این فرض غلط بوده و باید آن‌را رد می‌کردیم و تلویحا فرض H_A(که بیان می‌کرد میانگین health bills در جامعه بیشتر از 3100 می‌باشد) را می‌پذیرفتیم. علت آنکه خیلی مقدار خطای نوع دوم زیاد شد، این است که فاصله actual mean با میانگینی که در فرض صفر در نظر گرفتیم خیلی کم می‌باشد و می‌دانیم هرچه این فاصله کم‌تر شود خطای نوع دوم افزایش یافته و power کاهش می‌یابد.

(G

باتوجه به کدی که در بالا استفاده‌شده‌است، ابتدا مقدار power را محاسبه‌کردیم، که مقدار power برای این آزمون فرضی که طراحی کردیم برابر 14% می‌باشد.

Power : 0.14

همانطور که می‌دانیم خطای نوع دوم وابسته به effect size می‌باشد، که مقدار effect size برابر :

Effect size = $\bar{x} - \mu_0$

اگر میانگین واقعی جامعه، به null value (μ_0) خیلی نزدیک باشد، اینکه یک اختلافی مشاهده کنیم و بتوانیم H_0 را رد کنیم خیلی سخت‌تر می‌شود. هرچه این اختلاف و در نتیجه effect size بیشتر باشد، رد کردن H_0 و مشاهده‌ی این اختلاف ساده‌تر است و به عبارتی β (خطای نوع دوم) کاهش پیدا می‌کند. از آنجایی که power برابر β - 1 می‌باشد، چون خطای نوع دوم کاهش می‌یابد پس power افزایش پیدا می‌کند.

Question 7

(A

در قدم اول برای برداشتن نمونه‌ی 25 تایی از مجموعه داده(مجموعه داده فاقد missing value می‌باشد) از دستور زیر استفاده کرده و خروجی نیز در پایین قابل نمایش می‌باشد. لازم به ذکر است که این نمونه 25تایی برای دو متغیر 'bmi' و 'avg_glucose_level' در نظر گرفته شده‌است.

```
sample_25 <-non_miss_df[sample(nrow(non_miss_df), 25),c(9, 10)]
```

	avg_glucose_level	bmi
3620	61.54	13.2
539	155.43	27.3
1494	78.98	15.1
3591	112.69	33.5
619	158.31	32.8
528	113.85	34.0
2184	72.52	32.0
1395	90.16	28.9
2119	113.05	31.0
1123	96.84	30.2
4803	97.64	17.0
3401	73.19	33.5
3796	62.61	30.7
2186	211.35	30.7
458	74.44	45.2
3387	76.81	28.3
3151	231.43	23.0
3466	141.16	36.7
4199	71.18	23.9
3350	82.09	35.7
2792	69.38	22.1
4944	67.21	21.8
3488	83.64	29.4
4919	79.49	28.9
4316	100.80	39.3

(a

به دلیل آنکه تعداد نمونه‌ها از 30 کمتر می‌باشد و شرایط حد مرکزی رعایت نشده، به همین دلیل از t-test استفاده می‌کنیم.

(b

بررسی شرایط t-test:

استقلال درون گروهی داریم زیرا در هر گروه random sample / Assignment داریم و سایز سمپل نیز از 10% جامعه آماری نیز کم‌تر می‌باشد.

استقلا بین گروهی نداریم، زیرا هر دو گروه وابسته به هم هستند و هردو متغیر وابسته به یک نفر می‌باشد. در چنین مواقعی لزومی ندارد که از آزمون اختلاف دو میانگین استفاده کنیم. ابتدا تفاضل bmi و avg_glucose_level برای هر فرد را حساب می‌کنیم و یک متغیر جدید به نام diff ساخته می‌شود. با این کار دو متغیر را بک متغیر تبدیل کردیم. حال از همان روش‌هایی که برای استنباط روی یک متغیر داشتیم، استفاده می‌کنیم.

```
sample_25$diff <- sample_25$avg_glucose_level - sample_25$bmi
```

با استفاده از دستور بالا، ستون diff را به نمونه 25تایی که انتخاب کرده‌بودیم، اضافه می‌کنیم.

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

در واقع این فرض صفر در آزمون فرض بیان می‌کند که بین میانگین مقدار bmi و avg_glucose_level هیچ تفاوتی وجود ندارد و فرض جایگزین بیان می‌کند که بین میانگین این دو متغیر تفاوتی وجود دارد.

diff
128.13
63.88
79.19
125.51
79.85
40.52
61.26
82.05
66.64
80.64
39.69
31.91
180.65
29.24
48.51
208.43
104.46
47.28
46.39
47.28
45.41
54.24
50.59
61.50

$\bar{x}_{diff} = 74.1$

$s_{diff} = 44.7$

$T = \frac{74.1 - 0}{\frac{44.7}{\sqrt{25}}} = 8.3$

n = 25

df = 25 – 1 = 24

P-value = 2 × 8.379415e-09 = 1.675883e-08

باتوجه به مقدار به دست‌آمده برای p-value ، اگر فرض کنیم مقدار $\alpha = 0.05$ باشد، در این صورت چون p-value < α ، آنگاه H_0 را رد می‌کنیم. به عبارتی دیتایی که جمع‌آوری کردیم، شواهد قانع‌کننده‌ای برای رد کردن H_0 فراهم می‌کند. به همین دلیل تلویحا می‌توانیم بگوییم بین میانگین bmi و avg_glucose_level تفاوتی وجود دارد.

پیاده‌سازی این آزمون فرض در R نیز به شرح زیر می‌باشد:

```
t.test(sample_25$diff, data = sample_25, alternative = "two.sided")
```

```
        One Sample t-test

data:  sample_25$diff
t = 8.2867, df = 24, p-value = 1.683e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 55.61713 92.51007
sample estimates:
mean of x
  74.0636
```

(B

کدی که برای انتخاب نمونه‌های 100تایی برای هردو متغیر به صورتی که مستقل از یکدیگر باشند به این صورت تعریف شده است که از بین سطرهای اول تا نصف تعداد کل سطرها، نمونه 100تایی برای متغیر اول برداشته و برای متغیر دوم نیز یک نمونه 100 تایی از باقی مانده سطرها انتخاب کرده‌ایم. کدهای این قسمت به شرح زیر می‌باشد:

```
n1 = nrow(non_miss_df)
n2 = n1/2
```

```
smp2 <-non_miss_df[sample(1:n2+1, 100), c(1,9)]
smp1 <- non_miss_df[sample(n2+2 : n1 , 100),c(1, 10)]
```

آزمون فرضی که برای این قسمت در نظر گرفته شده‌است به شرح زیر می‌باشد:

$H_0 : \mu_{bmi} = \mu_{avg\_glucose\_level} \rightarrow \mu_{bmi} - \mu_{avg\_glucose\_level} = 0$

$H_A : \mu_{bmi} \neq \mu_{avg\_glucose\_level} \rightarrow \mu_{bmi} - \mu_{avg\_glucose\_level} \neq 0$

در این قسمت برای بررسی آزمون فرض از t-test استفاده کرده‌ایم. بنابراین بررسی می‌کنیم که شرایط t-test برقرار می‌باشد یا خیر:

- شرط Random sampling/assignment رعایت شده‌است و نیز سایز نمونه از 10% جامعه‌آماری کم‌تر می‌باشد. به همین دلیل استقلال درون‌گروهی داریم.
- استقلال بین گروهی نیز داریم، زیرا نمونه‌های هر متغیر مورد بررسی به صورت مستقل از هم انتخاب شده‌است.

برای محاسبه‌ی p-value و بازه‌اطمینان 95% هم از تابع t.test() استفاده شده‌است و هم با فرمول‌های موجود در درس پیاده‌سازی شده‌است .خروجی به شرح زیر می‌باشد:

```
#sol1
t.test(smp2$avg_glucose_level,smp1$bmi, alternative = "two.sided",
    paired = FALSE, var.equal = FALSE, conf.level = 0.95)

#sol2
df = length(smp1$bmi) - 1
SE = sqrt((var(smp1$bmi)/length(smp1$bmi)) +
        (var(smp2$avg_glucose_level)/length(smp2$avg_glucose_level)))

t_statistic <- (mean(smp2$avg_glucose_level) - mean(smp1$bmi)) / SE

left <- (mean(smp2$avg_glucose_level) - mean(smp1$bmi)) - (-qt(0.025, df= df) * SE)
right <- (mean(smp2$avg_glucose_level) - mean(smp1$bmi)) + (-qt(0.025, df= df) * SE)

p_value_orginal <- 2 * pt(t_statistic, df = df, lower.tail = FALSE)
```

خروجی تابع t.test() :

```
        Welch Two Sample t-test

data:  smp2$avg_glucose_level and smp1$bmi
t = 16.343, df = 104.05, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 75.12965 95.87935
sample estimates:
mean of x mean of y
 113.8995   28.3950
```

خروجی روش دوم:

p-value : 7.173953e-30

CI = (75.12, 95.89)

مقداری که در هردو روش به دست‌آمد یکی می‌باشد. حال با توجه به مقداری به‌دست آمده برای p-value، اگر مقدار α به طور پیش‌فرض برابر 0.05 باشد، آنگاه p-value < α بوده و آنگاه H_0 را رد می‌کنیم. به عبارتی میانگین این دو متغیری که انتخاب کردیم باهم برابر نمی‌باشد و این فرض را رد می‌کنیم.لازم به ذکر است که چون فاصله p-value تا α زیاد می‌باشد، با اطمینان بیشتری این فرض را رد می‌کنیم. همچنین باتوجه به بازه‌ی اطمینان به دست‌آمده نیز این فرض رد می‌شود، زیرا بازه‌ی اطمینان مقدار 0 را شامل نمی‌شود.

Question 8

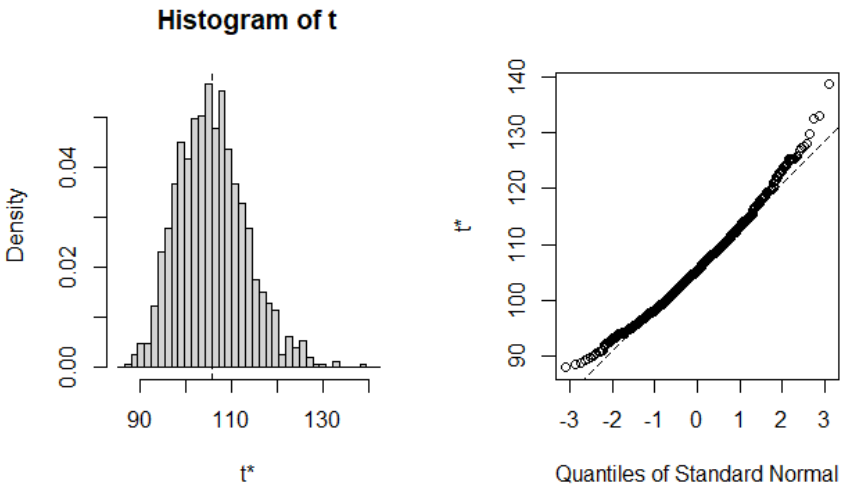
برای این سوال متغیر "avg_glucose_level" در نظ گرفته‌شده‌است. این متغیر همانطور که در قسمت‌های نمودار آن را ملاحظه کردید، دارای یک توزیع right skewedمی‌باشد و به همین دلیل اگر از میانگین استفاده کنیم زیرا این معیار نسبت به outlierها حساس می‌باشد و به همین دلیل نتیجه‌ی درستی به ما نخواهد داد. در چنین توزیع‌هایی از معیار median استفاده می‌کنیم که به outlierها حساس نیست ولی همانطور که می‌دانیم CLT برای میانه نمی‌باشد.

(A

برای حل این سوال یک نمونه 25تایی از مجموعه داده با استفاده از دستور زیر گرفته‌شده است، سپس 1000 مرتبه از این نمونه 25ایی که داریم، نمونه‌برداری کرده و در هر نمونه میانگین را محاسبه می‌کنیم تا در نهایت یک نمونه 1000تایی از میانگین‌ها داشته‌باشیم. به همین منظور از تابع boot() استفاده شده‌است. سپس با استفاده از روش percentile، بازه‌اطمینان را محاسبه می‌کنیم، که برای این قسمت از تابع quantile() استفاده‌شده‌است. کدهای زده‌شده برای این قسمت به شرح زیر می‌باشد.

```
sample_25 <-data.frame(non_miss_df[sample(nrow(non_miss_df), 25),c(9)])
colnames(sample_100) <- c("avg_glocuse_level")


Bootstrap_1 <- boot(sample_25, statistic=meanfun, R=1000)

plot(Bootstrap_1)
```

Bootstrap distribution برای این نمونه‌برداری به صورت زیرمی‌باشد:



بازه اطمینانی که با استفاده از روش percentile به‌دست آمد به شرح زیر می‌باشد:

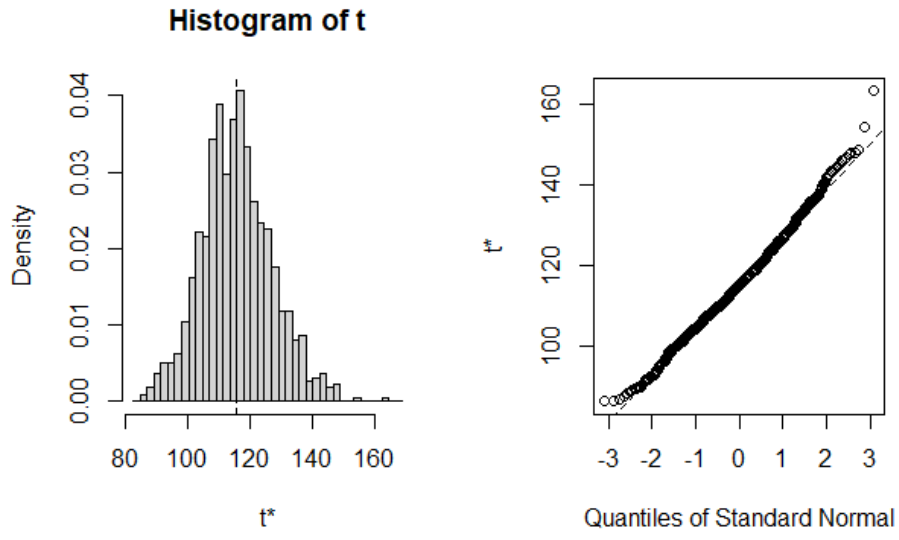
```
#compute CI with percentile method
smp_CI <- quantile(Bootstrap_1$t, c(0.025, 0.975))
```

Confidence Interval : (93.22, 123)

(B

ابتدا یک Bootstrap distribution از میانگین نمونه که سایز نمونه برابر 20 می‌باشد، می‌سازیم (لازم به ذکر است که Bootstrap sampling به تعداد 1000 بار انجام شده‌است). که خروجی به‌دست آمده به شرح زیر می‌باشد:

```
sample_20 <-data.frame(non_miss_df[sample(nrow(non_miss_df), 20),c(9)])

library(boot)

meanfun <- function(data, i){
  d <- data[i, ]
  return(mean(d))
}

Bootstrap <- boot(sample_20, statistic=meanfun, R=1000)

plot(Bootstrap)
```



حال با روش standard error برای این مقادیر به‌دست‌آمده، بازه‌ی اطمینان را محاسبه می‌کنیم. کد زده شده برای این قسمت و خروجی به شرح زیر می‌باشد:

```
df1 <- length(Bootstrap$t) - 1
ci_left <- Bootstrap$t0 - (-qt(0.025, df = df1) * sd(Bootstrap$t))
ci_right <- Bootstrap$t0 + (-qt(0.025, df = df1) * sd(Bootstrap$t))
```

Confidence Interval : (93.26, 138.11)

(C

بازه‌ی اطمینانی که با استفاده از این دو روش به‌دست آمده‌است، تفاوت جزئی باهم دارند و نزدیک به هم می‌باشند که این نشانه‌ی آن است که این روش روی این نمونه‌ای که داشتیم ، تقریبا خوب جواب داده‌است.

Question 9

برای محاسبه‌ی تست ANOVA به منظور اینکه مشاهده کنیم اختلافی بین میانگین متغیر 'health_bills' در پنج نوع گروه کاری که داریم یکسان است یا خیر از تابع aov() استفاده شده‌است. آزمون فرضی که در نظر گرفتیم طبق آنچه صورت سوال خواسته به شرح زیر می‌باشد:

$H_0 : \mu_{children} = \mu_{Govt\_job} = \mu_{Never\_worked} = \mu_{Private} = \mu_{Self-employed}$

$H_A : \mu_{children} \neq \mu_{Govt\_job} \neq \mu_{Never\_worked} \neq \mu_{Private} \neq \mu_{Self-employed}$

```
anova_one_way <- aov(health_bills ~ work_type, data = non_miss_df)
summary(anova_one_way)
```

```
              Df    Sum Sq  Mean Sq F value Pr(>F)
work_type      4 1.951e+08 48784956   76.11 <2e-16 ***
Residuals   4904 3.143e+09   641007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

باتوجه به نتیجه‌ای که در تصویر بالا قابل مشاهده می‌باشد، چون مقدار p-value به دست‌آمده در مقایسه با مقداری که به صورت پیش فرض برای α داریم که برابر 0.05 است، خیلی کم‌تر است، به همین دلیل فرض H_0 را رد می‌کنیم. به عبارتی میانگین health bills در گروه‌های کاری مختلف باهم برابر نمی‌باشد و حداقل دو گروه هستند که میانگینشان به یکدیگر متفاوت می‌باشد.

برای بررسی اینکه میانگین کدام دو گروه با یکدیگر متفاوت می‌باشد از تابع TukeyHSD() استفاده کرده‌ایم و خروجی به شرح زیر می‌باشد:

```
TukeyHSD(anova_one_way)
```

```
$work_type
                                diff        lwr      upr     p adj
Govt_job-children           551.45959  430.25790 672.6613 0.0000000
Never_worked-children       366.05178 -107.31261 839.4162 0.2157638
Private-children            576.93586  483.06641 670.8053 0.0000000
Self-employed-children      605.95221  490.74661 721.1578 0.0000000
Never_worked-Govt_job      -185.40781 -659.26093 288.4453 0.8231359
Private-Govt_job            25.47627  -70.82748 121.7800 0.9515223
Self-employed-Govt_job      54.49262  -62.70495 171.6902 0.7104665
Private-Never_worked        210.88408 -256.72516 678.4933 0.7334426
Self-employed-Never_worked  239.90043 -232.45458 712.2554 0.6367876
Self-employed-Private        29.01635  -59.62277 117.6555 0.8994894
```

ستون آخر در تصویر بالا، مقدار p-value را برای تست آزمون فرض آینکه هر دو گروه میانگین برابر دارند یا خیر دو گروه نمایش می‌دهد.که این مقادیر را باید با α^* بسنجیم.

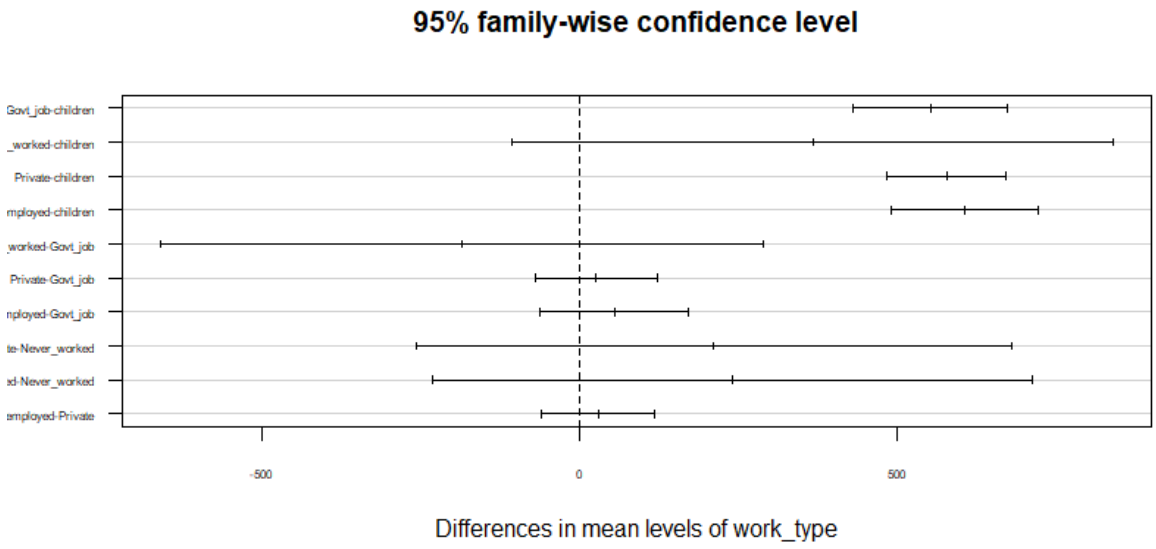
$$\alpha^* = \frac{\alpha}{k} = \frac{0.05}{10} = 0.005$$

K برابر تعداد دوبه دو گروه‌هایی است که باید باهم مقایسه شوند و از آنجایی که ما 5 تا گروه داریم، برابر $\binom{5}{2} = 10$ می‌باشد. طبق این عدد به دست‌آمده و مقادیر p-value در تصویر بالا، سه مورد از ترکیب‌ها، میانگین health_bills متفاوت از هم دارند که این سه مورد عبارتند از:

- Govt_job - children
- Private-children
- Self-employed-children

همچنین نموداری برای مقایسه میانگین دو به دو گروه‌ها رسم کرده‌ایم که صورت زیر می‌باشد:

```
plot(TukeyHSD(anova_one_way), las = 1)
```



که این تصویر نیز نشان‌دهنده‌ی این است که که کدام دو گروه میانگین health_bills شان با یکدیگر برابر نیست.