



نام و نام خانوادگی : سرمد زندی گوهرریزی

شماره دانشجویی : ۸۱۰۱۹۹۱۸۱

درس : استنباط آماری

مدرس : دکتر بهنام بهرک



ProjectPhase-1 (Report)

سوال شماره .

علم آمار تشکیل شده از ۴ جز است که جمع آوردی داده های مرتبط با مسئله ی مورد مطالعه، یکی از اجزای علم آمار تلقی می شود. در علم آمار، داده های جمع آوری شده (observations) را درون مجموعه داده ها (dataset) قرار می دهیم. در یک dataset، هر سطر نشان دهنده ی یک observation یا case است و هر ستون را متغیرهایی میدانیم که برای هر case اطلاعاتی را در اختیار ما قرار می دهند.

در این پروژه مجموعه داده ای که مورد بررسی قرار می گیرد، HealthCare نام دارد که شامل اطلاعات وضعیت سلامتی حدود ۵۰۰۰ نفر و همچنین میزان هزینه ی سالیانه سلامتی آنهاست. در واقع یک نمونه شامل ۵۱۱۰ نفر از افراد مورد مطالعه مان (جامه هدف) گرفته شده و اطلاعات وضعیت سلامتی آن ها درون این مجموعه داده قرار داده شده است.

در سوال شماره . این پروژه قصد داریم تا با این مجموعه داده بیشتر آشنا شویم.

A :

برای بدست آوردن اطلاعات بیشتر در مورد مجموعه داده Healthcare در ابتدا به کمک R آن را import می کنیم. آن چیزی که مشخص است این dataset اطلاعات فردی 5110 نفر شامل جنسیت (male - female - other)، سن و شماره شناسایی آن فرد، و همچنین اطلاعات سلامتی این 5110 نفر که شامل فشار خون (0 = فشار خون پایین و 1 = فشار خون بالا)، سابقه بیماری قلبی (0 = داشتن و 1 = نداشتن)، وضعیت تاهل (yes = متاهل بوده است و no = مجرد است)، نوع شغل، نوع محل سکونت (Urban/ Rural)، سطح متوسط گلوکز، شاخص bmi فرد، وضعیت استعمال دخانیات و سابقه سکتة مغزی (0 = داشتن و 1 = نداشتن) می باشند را به همراه میزان هزینه سالیانه سلامتی آنها در اختیار ما قرار می دهد. مطالعه این dataset از این جهت می تواند جالب و پر اهمیت باشد که میتوان با بررسی و مشاهده ویژگی های این dataset، فرضیات اولیه ای را در مورد این داده ها مطرح کرد. به عنوان مثال میتوان عوامل موثر در سکتة قلبی و سکتة مغزی را شناخت و سپس با استفاده از ابزار های آماری به اطمینان حاصل کردن از این فرضیات و نتیجه گیری های آماری پرداخت.

(import کردن dataset و مشاهده آن در R)

```
# Importing the Healthcare dataset
HealthCare <- read.csv("H:/Second Term/Statistical Inference/Projects/ProjectPhase1/HealthCare.csv")
view(HealthCare)
```

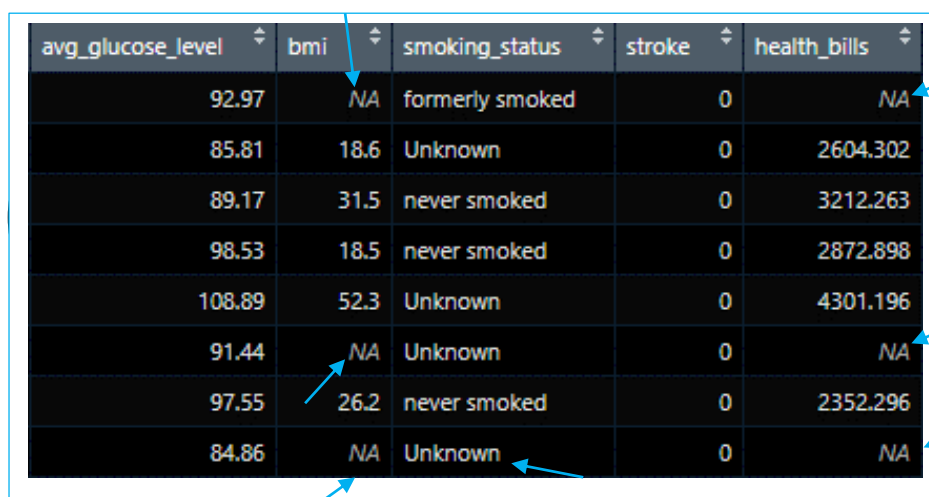
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills
1	9046	Male	67.00	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1	6011.860
2	51676	Female	61.00	0	0	Yes	Self-employed	Rural	202.21	NA	never smoked	1	NA
3	31112	Male	80.00	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1	6384.530
4	60182	Female	49.00	0	0	Yes	Private	Urban	171.23	34.4	smokes	1	5862.754
5	1665	Female	79.00	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1	5461.262

B :

در یک dataset، هر سطر نشان دهنده ی یک observation یا case است و هر ستون را متغیرهایی میدانیم که برای هر case اطلاعاتی را در اختیار ما قرار می دهند. بنابراین در این dataset تعداد سطر ها نشان دهنده ی تعداد case ها یا همان observation های ما هستند، که در این مطالعه برابر با 5110 است. همچنین هر ستون نشان دهنده ی یک متغیر است که البته ستون id در واقع identity محسوب می شود و نه متغیر. بنابراین تعداد متغیرها برابر با ۱۲ است که عبارتند از: "gender"، "age"، "hypertension"، "heart_disease"، "ever_married"، "work_type"، "Residence_type"، "avg_glucose_level"، "bmi"، "smoking_status"، "stroke" و "health_bills".

C :

در علم آمار، missing value زمانی رخ می‌دهد؛ که هیچ مقدار داده ای برای یک متغیر در یک observation ذخیره نشده باشد. در مجموعه داده Healthcare مشاهده می شود که با این پدیده رو به رو هستیم و برای برخی از observation ها برای یک متغیر مقداری وجود ندارد. دلایل مختلفی میتواند برای این پدیده وجود داشته باشد، ممکن است برخی داده ها بر اثر یک رویداد تصادفی یا حتی به دلیل بی پاسخی، از دست رفته باشند. به عنوان نمونه بخشی از dataset که شامل اطلاعات از دست رفته است در زیر آورده شده است.



avg_glucose_level	bmi	smoking_status	stroke	health_bills
92.97	NA	formerly smoked	0	NA
85.81	18.6	Unknown	0	2604.302
89.17	31.5	never smoked	0	3212.263
98.53	18.5	never smoked	0	2872.898
108.89	52.3	Unknown	0	4301.196
91.44	NA	Unknown	0	NA
97.55	26.2	never smoked	0	2352.296
84.86	NA	Unknown	0	NA

همانطور که مشاهده می شود در ستون smoking_status که وضعیت سیگار کشیدن هر فرد درون این dataset را نشان می‌دهد، می بینیم که برای برخی افراد مقدار "Unknown" وجود دارد که به معنی در دست نبودن اطلاعات برای این فرد است. اما داده گم شده حساب نمی شود. همچنین در سایر متغیرهای این مطالعه (در ستون های bmi و health_bills در dataset که از دسته متغیرهای عددی حساب می شوند) عبارت N/A مشاهده می شود که به معنی بدون پاسخ بودن این مقدار برای آن فرد (case) است.

معمولاً در تحلیل های آماری، مشاهداتی که دارای مقادیر گمشده هستند، نادیده گرفته می شوند و بدون در نظر گرفتن آن ها محاسبات صورت می گیرد. در بعضی از تحلیل های آماری به دلیل کمبود مشاهدات، گاهی داده گمشده در متغیرهای عددی را با میانگین (Mean) یا میانه (Median) جایگزین می کنند و در مورد متغیرهای categorical از مقداری که بیشترین تکرار را داشته برای جایگزینی استفاده می کنند تا تعداد نمونه، کاهش نیابد.

در این پروژه، تنها متغیرهای bmi و health_bills دارای مقادیر گم شده هستند و من از روش دوم (یعنی جایگزینی میانگین) برای رفتار کردن با داده های گمشده در این متغیر ها استفاده میکنم.

D :

از آنجایی که این dataset شامل اطلاعات وضعیت سلامت افراد درون آن است با نگاهی اجمالی میتوان دریافت که متغیرهای فشار خون ، سابقه بیماری قلبی ، سطح متوسط گلوکز ، شاخص bmi فرد ، وضعیت استعمال دخانیات و سابقه سکته مغزی از جمله متغیرهایی هستند که حاوی اطلاعات مهمی در مورد هر observation میباشدند. به دلیل اینکه مقدار هر کدام از این متغیرها برای یک observation میتواند فرضیاتی را برای ما به همراه داشته باشد. مثلاً فشار خون بالای یک فرد یا سیگاری بودن یک فرد ممکن است رابطه ای با سکته قلبی آن داشته باشند. همچنین از این متغیر ها میتوان برای پیش بینی اینکه آیا یک بیمار احتمال سکته مغزی را دارد یا خیر استفاده شود. البته در آینده ممکن است فرضیاتی پذیرفته یا رد شوند. در حال حاضر با نگاهی اجمالی میتوان صرفاً یکسری فرضیه اولیه مطرح کرد که در آینده با ابزار های آماری مورد بررسی قرار می گیرند.

متغیرهای عددی (Numerical) در یک Dataset به متغیرهای کمی معروفند که مقادیر عددی را اختیار می کنند. این متغیرها دو نوع اند: متغیرهای عددی گسسته، که تعداد قابل شمارشی مقدار می توانند اختیار کنند و متغیرهای عددی پیوسته، که در یک رنج مشخص تعداد زیادی مقدار مختلف را اختیار می کنند. برای این سوال متغیر انتخابی "age" است که در واقع سن هر فرد (Case) در این مجموعه داده را مشخص می کند.

A:

Histogram یکی از روش های مصورسازی داده های عددی است. در این نمودار داده های مجموعه داده به بازه هایی با طول مساوی تقسیم می شوند که برای هر بازه، یک bin وجود دارد که ارتفاع آن نشان دهنده ی تعداد observation های آن بازه است. از نمودار هیستوگرام برای دیدن شکل توزیع داده های عددی استفاده می شود و انتخاب کردن ساینز bin مناسب در شکل توزیع و اطلاعاتی که این نمودار به ما می دهد بسیار تاثیر گذار است. برای تحلیل یک هیستوگرام باید بدانیم که مفهوم چولگی و Modality یک توزیع به چه معناست.

چولگی یک توزیع میزان کجی آن را به سمت راست یا چپ نشان می دهد (در واقع میزان عدم تقارن توزیع) و به ما می فهماند که نیاز به نمونه ای با ساینز بزرگتر داریم. همچنین Modality تعداد نقاط Max توزیع یا همان نقاط Peak توزیع را مشخص می کند. (در واقع مشخص می کند کدام bin بیشترین ارتفاع را دارد).

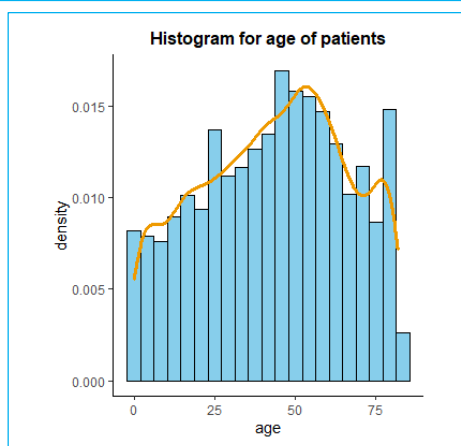
در این سوال با استفاده از کتابخانه ggplot نمودار هیستوگرام و منحنی density توزیع آن برای متغیر عددی age رسم شده است، همچنین برای مشخص کردن عرض bin مناسب چندین هیوریستیک وجود دارد که در این سوال از روش Freedman-Diaconis استفاده شده است:

$$\frac{2(IQR)}{n^{1/3}}$$

در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است:

```
# importing libraries
library(ggplot2)
library(plyr)
# HISTOGRAM and probability density function for age of observations
ggplot(Healthcare, aes(x = age)) +
# plot a histogram
  geom_histogram(aes(y=..density..),
    binwidth = function(x) 2 * IQR(x) / (length(x)^(1/3)) ,
    colour = "black", fill = "skyblue") +
# distribution of age (Density Curve)
  geom_density(color="orange2",size=1.4) +
  labs(title="Histogram for age of patients") + # the title of the plot
  theme_bw() +
# the grid and background removed from plot,
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),
    # put the title location in the center of the plot.
    plot.title = element_text(size=12,face="bold",hjust = 0.5))
```

کد:



نمودار:

همانطور که گفته شد به نقطه Peak توزیع (آن bin که بیشترین ارتفاع را در هیستوگرام دارد) Mod گفته می شود. از آنجایی که دو bin با بیشترین ارتفاع در این توزیع دیده می شود، میتوان گفت که این توزیع تقریباً bimodal است چون می توان وجود دو Peak در این توزیع را مشاهده کرد. از آنجایی که این منحنی، توزیع سن افراد حاضر در مجموعه داده را نشان می دهد، میتوان به این موضوع پی برد که تعداد زیادی از افراد (Case) حاضر در این مجموعه داده، در بازه سنی بین ۴۵ تا ۵۵ سال قرار دارند.

نکته: می دانیم که عرض bin در شکل توزیع تاثیر می گذارد، اگر binwidth را خیلی کوچک در نظر می گرفتیم شکل توزیع مان یکنواخت (Uniform) می شد و اگر آن را خیلی بزرگ در نظر می گرفتیم همه داده ها در یک bin جمع می شدند و هیچ اطلاعاتی نمیتوانستیم از هیستوگرام بدست بیاوریم. ممکن است با دستکاری سایز bin به توزیع مد نظرممان برسیم اما این کار یکی از روش های قلب آماری به شمار می رود، بنابراین از بین هیوریستیک های موجود، بهترین گزینه را انتخاب کردم تا سایز مناسبی را برای bin ها داشته باشم و می بینیم که تقریباً با توزیعی bimodal روبه رو هستیم.

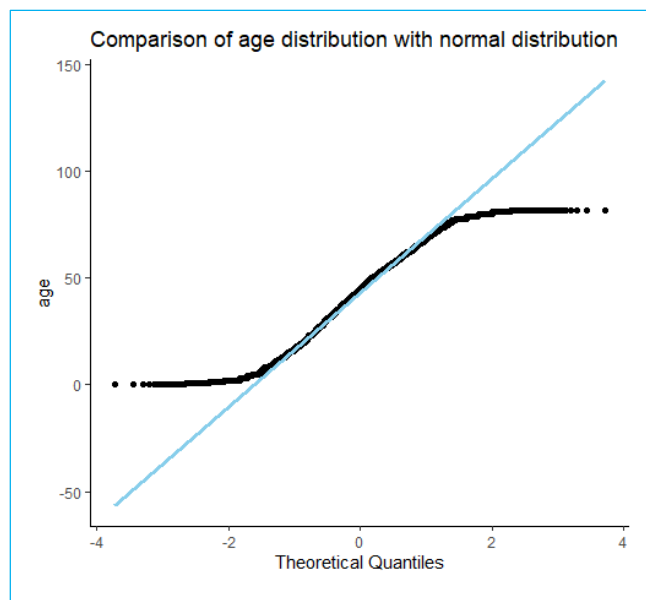
B :

هر توزیع دارای ویژگی هایی از جمله : شکل ، چولگی ، گستردگی ، مرکزیت آن و Modality آن است. همانطور که گفته شد چولگی یک توزیع میزان کجی آن را به سمت راست یا چپ نشان می دهد و به ما می فهماند که به سبب بزرگتری برای استنباط های آماری نیاز داریم، چولگی یکی از ویژگی های توزیع های عددی است و همانطور که پیداست در این سوال همانطور که از هیستوگرام رسم شده در قسمت A مشاهده می شود، با توزیعی چوله به چپ روبه رو هستیم. در قسمت A در مورد مفهوم Modality یک توزیع صحبت شد و در مورد این توزیع دیدیم که توزیع سن افراد حاضر در این مجموعه داده، به دلیل اینکه دو bin با بیشترین ارتفاع در آن دیده می شود توزیعی bimodal است. شکل یک توزیع، یکی دیگر از ویژگی های توزیع ها به شمار می رود و نشان دهنده الگوی آن است. همانطور که گفته شد شکل توزیع ما bimodal و چوله به چپ است. همچنین گستردگی و مرکزیت یک توزیع از جمله ویژگی های آن است که در قسمت های بعدی این سوال به آن ها به طور کامل خواهیم پرداخت اما در این قسمت هدف اصلی، مقایسه توزیع سن افراد حاضر در این مجموعه داده، با توزیع نرمال است تا ببینیم آیا از توزیع نرمال (گوسی) پیروی میکنند یا خیر.

توزیع نرمال توزیعی Unimodal، متقارن و به اصطلاح bell curve است. ولی ما دیدیم که توزیع سن افراد حاضر در این مجموعه داده، توزیعی bimodal و چوله به چپ است. اما ما برای مقایسه یک توزیع با توزیع نرمال معمولاً از نمودار Q-Q plot استفاده می کنیم. این نمودار درواقع یک visual check است که یکسری اطلاعات درباره توزیع ها و نحوه ارتباطشان با هم را می دهد و می گوید دو توزیع شکل یکسانی دارند یا خیر. برای رسم این نمودار از کتابخانه ggplot استفاده شده است. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است :

```
# Q-Q plot for comparison of age distribution with normal distribution
qplot(sample = age, data = HealthCare)+
# the title of the plot
labs(title="Comparison of age distribution with normal distribution",
      y = "age",x="Theoretical Quantiles")+
# draw a line with size=1.3 in Q-Q plot
stat_qq_line(size=1.3,color="skyblue") +
theme_classic()
```

کد :



نمودار :

در Q-Q plot هر چه data point ها نزدیک تر به خط راست یا حتی روی خط راست باشد گوییم توزیع مان به توزیع نرمال نزدیک تر است. با توجه به نمودار Q-Q plot رسم شده، مشاهده می شود که اکثر داده ها روی خط راست قرار دارند اما قسمت های عمده ی بالا و پایین Data point هایمان از خط راست فاصله زیادی دارند. بنابراین در این حالت اصطلاحاً میگوییم توزیع ما Long Tail است و گستردگی بیشتری نسبت به توزیع نرمال دارد.

C :

چولگی یک توزیع میزان کجی آن را به سمت راست یا چپ نشان می دهد (در واقع میزان عدم تقارن توزیع) و به ما می فهماند که به سبب بزرگتری برای استنباط های آماری نیاز داریم، همانطور که گفته شد چولگی یکی از ویژگی های توزیع های عددی است. یکی از راه های سریع برای محاسبه چولگی یک توزیع، استفاده از تعریف ساده Non-parametric آن است :

$$sk = \frac{\text{mean} - \text{median}}{\text{standard deviation}} = \frac{\mu - m}{\sigma}$$

و می دانیم که اگر از این طریق به محاسبه چولگی یک توزیع بپردازیم، داریم :

if sk > 0 : right_skewed

if sk = 0 : symmetric

if sk < 0 : left_skewed

نکته : هرچه این عدد بیشتر باشد توزیع ما چولگی شدید تری خواهد داشت.

بنابراین با توجه به توضیحات داده شده، ابتدا در R برای محاسبه چولگی توزیع سن افراد حاضر در مجموعه داده HealthCare، از پکیج moments و تابع skewness() استفاده می کنیم. سپس چولگی را با استفاده از روش Non-parametric آن که در بالا شرح داده شده است محاسبه می کنیم. در شکل زیر کد محاسبه چولگی و نتیجه حاصله از اجرای کد به هر دو روش قابل مشاهده است :

```
# importing library
library(moments)

# calculate skewness in r
skewness(HealthCare$age)

# Non-parametric
sk <- (mean(HealthCare$age)-median(HealthCare$age))/sd(HealthCare$age)
sk
```

کد :

```
> # calculate skewness in r
> skewness(HealthCare$age)
[1] -0.1370191
>
> # Non-parametric
> sk <- (mean(HealthCare$age)-median(HealthCare$age))/sd(HealthCare$age)
> sk
[1] -0.0784245
```

پاسخ :

همانطور که می بینیم، عدد در هر دو روش منفی شده است و این به معنی این است که توزیع ما چوله به چپ است. لازم به ذکر است که در هیستوگرام رسم شده در قسمت A هم شاهد کشیده شدن دم (Tail) توزیع به سمت چپ یا به عبارتی چولگی به سمت چپ هستیم. در واقع در این حالت دم سمت چپ توزیع طولانی تر از سمت راست است و این به معنی متمرکز شدن حجم داده بیشتری در سمت راست توزیع می باشد. در واقع در توزیع های چوله به چپ میانگین کمتر از میانه است و اصطلاحاً میانگین توزیع سن افراد حاضر در مجموعه داده HealthCare به سمت چپ منحنی توزیع آن تمایل دارد.

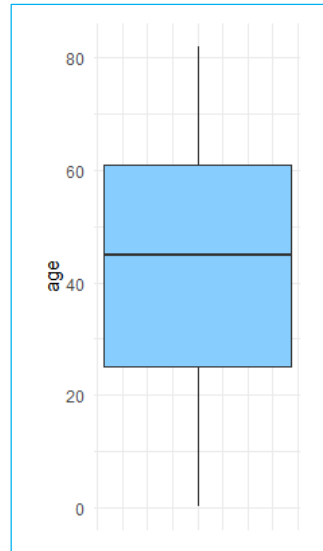
D :

outlier ها یکسری داده های پرت هستند که مطالعه آن ها میتواند منجر به اطلاعات جالبی در مورد data شود. همچنین اهمیت مطالعه این داده های پرت میتواند باعث تشخیص خطاهایی که در زمان جمع آوری داده ها اتفاق افتاده شود و چولگی را مشخص و تغییر دهد. یک راه شناسایی و مصورسازی outlier ها اسفاده از Boxplot است. در ابتدا لازم به ذکر است که متغیر عددی انتخابی من در این سوال، متغیر age است که برای هر observation سن آن را مشخص می کند. اما با رسم Boxplot برای این متغیر همانطور که در شکل زیر میبینیم، مشاهده می شود که برای این متغیر هیچ outlier یا داده پرت وجود ندارد. از این رو میتوان فهمید که در جمع آوری اطلاعات سن افراد در این مجموعه داده خطایی رخ نداده است.

```
# importing libraries
library(ggplot2)

# plot a Boxplot for "age" variable
ggplot(HealthCare) +
  aes(y = age) +
  geom_boxplot(fill = "skyblue1") +
  theme_minimal() +
  theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())
```

کد :

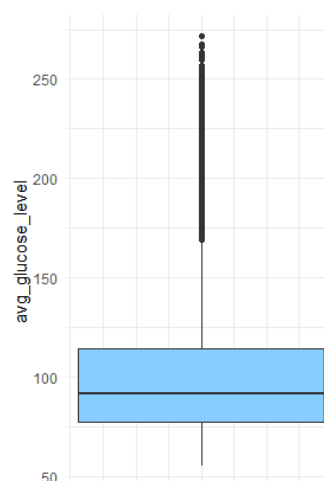


نمودار :

اما برای پاسخ دادن به این سوال، صرفاً برای این قسمت، از متغیر عددی avg_glucose_level که سطح متوسط گلوکز هر فرد (Case) درون این مجموعه داده را مشخص می‌کند استفاده می‌کنیم. برای شناسایی و مصورسازی outlier های این متغیر با استفاده از کتابخانه ggplot اقدام به رسم Boxplot طبق کد زیر می‌کنیم.

```
# plot a Boxplot for "avg_glucose_level" variable
ggplot(HealthCare) +
  aes(y=avg_glucose_level) +
  geom_boxplot(fill = "skyblue1") +
  theme_minimal() +
  theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())
```

کد :



نمودار :

همانطور که در نمودار Boxplot حاصل مشاهده می‌شود، این متغیر دارای outlier یا داده‌های پرت هست که بالاتر از upper whisker قرار گرفته‌اند.

E :

میان و میانگین از معیارهای مرکزیت یک توزیع هستند. مشخصاً میانگین معماریست که با آن متوسط حسابی observation های یک متغیر عددی مشخص می شود و میان نقطه وسط توزیع است که ۵۰ درصد داده ها از آن پایین تر و ۵۰ درصد باقی داده ها از آن بالاتر قرار می گیرند. برای محاسبه این دو معیار برای توزیع سن افراد حاضر در این مجموعه داده، میتوان به ترتیب از توابع mean() و median() در R استفاده کرد. در شکل زیر کد و نتیجه محاسبه میانگین و میان توزیع سن افراد قابل مشاهده است :

<pre>mean(HealthCare\$age) median(HealthCare\$age)</pre>	کد :
<pre>> mean(HealthCare\$age) [1] 43.22661 > median(HealthCare\$age) [1] 45</pre>	پاسخ :

مشاهده می شود که میانگین سنی افراد حاضر در این مجموعه داده برابر با ۴۳/۲۲ است و همچنین ۵۰ درصد از این افراد سنی بیشتر از ۴۵ سال دارند (یعنی از بین ۵۱۱۰ نفر حدوداً ۲۵۵۵ نفر) و همینطور ۵۰ درصد از این افراد سن شان کمتر از ۴۵ سال است.

واریانس و انحراف معیار از معیارهای سنجش گستردگی data در dataset شناخته می شوند که بسیار برای آنالیز داده ها کاربردی هستند. هر دو واریانس و انحراف معیار در آمار میزان فاصله observation را از میانگین توزیع بیان میکنند اما تفاوت آن ها در جنس آن هاست، درواقع واریانس از جنس مربع داده ها و انحراف معیار از جنس خود دیتا است. برای محاسبه این دو معیار برای گستردگی توزیع سن افراد حاضر در این مجموعه داده، میتوان به ترتیب از توابع var() و sd() در R استفاده کرد. در شکل زیر کد و نتیجه محاسبه واریانس و انحراف معیار توزیع سن افراد قابل مشاهده است :

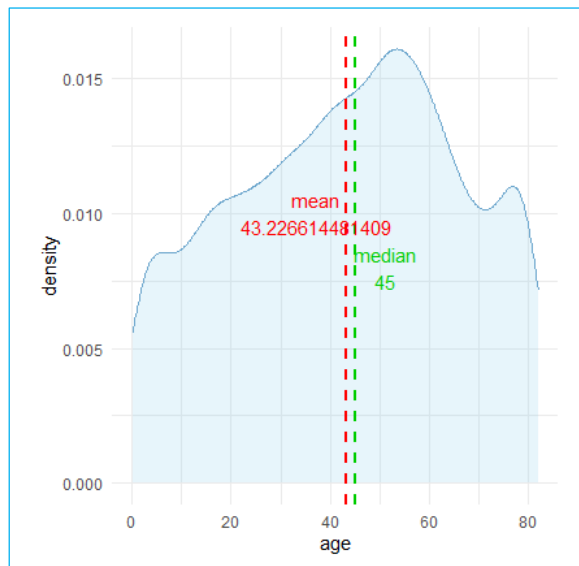
<pre>var(HealthCare\$age) sd(HealthCare\$age)</pre>	کد :
<pre>> var(HealthCare\$age) [1] 511.3318 > sd(HealthCare\$age)</pre>	پاسخ :

با توجه به اینکه متوسط سن افراد در مجموعه داده برابر با ۴۳/۲۲ است و ۵۱۱۰ فرد در مجموعه داده وجود دارد: واریانس برابر با ۵۱۱/۳۳۱۸ و انحراف معیار برابر با ۲۲/۶۱۲۶۵ است.

F :

در قسمت قبل میانگین و میان توزیع به طور کامل معرفی شدند و آن ها را حساب کردیم. همینطور در مورد مفهوم چولگی یک توزیع در قسمت های قبل بحث شد. در این قسمت میخواهیم به رابطه ای که بین این دو معیار میانگین و میان وجود دارد بپردازیم. در ابتدا با استفاده از کتابخانه ggplot یک density plot رسم کرده و با استفاده از geom_vline، یک خط برای مشخص کردن میانگین و میان توزیع رسم می کنیم. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است :

<pre># plot a density plot for age ggplot(HealthCare, aes(x=age))+ geom_density(color="skyblue3", fill="skyblue", alpha = 0.2)+ theme_minimal() + # add lines for the mean and median geom_vline(aes(xintercept=mean(age)), color="red1", linetype="dashed", size=1) + geom_vline(aes(xintercept=median(age)), color="green3", linetype="dashed", size=1) + # # add texts for the mean and median annotate(geom="text", x=mean(HealthCare\$age)-6, y=0.01, label=paste0("mean\n",mean(HealthCare\$age)), color="red1") + annotate(geom="text", x=median(HealthCare\$age)+6, y=0.008, label=paste0("median\n",median(HealthCare\$age)), color="green3")</pre>	کد :
--	------



نمودار :

در قسمت C یک تعریف Non-parametric برای محاسبه چولگی معرفی شد، اما از روی رابطه ای که بین میانگین و میانه یک توزیع density وجود دارد هم میتوان چولگی آن را تشخیص داد. اگر میانه یک توزیع از میانگین آن کوچکتر باشد، توزیع چولگی مثبت دارد و اصطلاحاً چوله به راست است همچنین z-score منفی خواهد داشت. همچنین اگر میانه یک توزیع از میانگین آن بزرگتر باشد، توزیع چولگی منفی دارد و اصطلاحاً چوله به چپ است همچنین z-score مثبت خواهد داشت. از آنجایی که با توجه به شکل، میانه توزیع سن افراد حاضر در این مجموعه داده از میانگین بزرگتر است، شاهد کشیده شدن دم (Tail) توزیع به سمت چپ یا به عبارتی چولگی به سمت چپ هستیم. در واقع میانگین توزیع سن افراد حاضر در مجموعه داده HealthCare به سمت چپ منحنی توزیع آن تمایل دارد.

G :

در این قسمت باید متغیر عددی age را بر اساس میانگین آن که ۴۳ است به ۴ بازه (0-21 , 22-43 , 44-64 , 65-82) تبدیل کنیم. در واقع می خواهیم با استفاده از یک متغیر Numerical ، یک متغیر categorical جدید به مجموعه داده HealthCare اضافه کنیم. در ابتدا برای این کار یک تابع را معرفی میکنیم که در هر بار فراخوانی یک پارامتر را از ورودی گرفته و بر اساس اینکه این پارامتر چه عددی بوده است یک label که معرف بازه آن است بر می گرداند. سپس با استفاده از تابع Lapply ، این تابع را روی تک تک مقادیر متغیر age اعمال کرده و در ستون جدید age_interval درون مجموعه داده HealthCare می ریزیم. کد این قسمت و ستون جدید ایجاد شده در مجموعه داده، در عکس زیر قابل مشاهده است :

```
# this function Categorizing the age variable into four intervals based on its mean.
func <- function(inp){
  if(inp <= 21)
    return("0-21")
  else if(inp >= 22 & inp <=43)
    return("22-43")
  else if(inp >=44 & inp <=64)
    return("44-64")
  else if(inp >=65)
    return("65-82")
}

# Add a new column to HealthCare dataset and call function on each row of age variable
# to Categorize this variable into four intervals based on its mean
HealthCare$age_interval <- lapply(HealthCare$age,func)
```

age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills	age_interval
17.00	0	0	No	Private	Urban	92.97	NA	formerly smoked	0	NA	0-21
13.00	0	0	No	children	Rural	85.81	18.6	Unknown	0	2604.302	0-21
55.00	0	0	Yes	Private	Urban	89.17	31.5	never smoked	0	3212.263	44-64
42.00	0	0	No	Private	Urban	98.53	18.5	never smoked	0	2872.898	22-43
31.00	0	0	No	Private	Urban	108.89	52.3	Unknown	0	4301.196	22-43
38.00	0	0	Yes	Private	Urban	91.44	NA	Unknown	0	NA	22-43
24.00	0	0	No	Private	Urban	97.55	26.2	never smoked	0	2352.296	22-43
80.00	0	0	Yes	Govt_job	Urban	84.86	NA	Unknown	0	NA	65-82

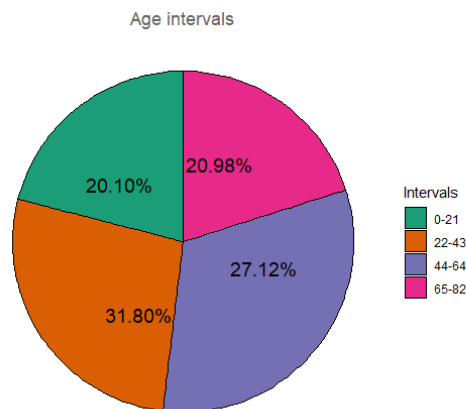
سپس یک dataframe ایجاد کرده که دو ستون شامل نام هر category و فرکانس تکرار آن را نگه داری می کند. و در آخر با استفاده از کتابخانه ggplot اقدام به رسم نمودار pie chart می کنیم که رنگ هر category در آن متفاوت است و درصد آن بر روی آن نوشته شده است. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است :

```
# create a dataframe
dataa <- data.frame(data = prop.table(table(unlist(HealthCare$age_interval)))*100)

# importing libraries
library(ggplot2)
library(scales) # for percentage scales

# plot a pie chart that visualizes the frequency of these four categories (age intervals)
ggplot(data = dataa, aes(x = "", y = data.Freq, fill = data.Var1)) +
  geom_bar(stat = "identity", color = "black", width=1) +
  labs(fill = "Intervals", title = "Age intervals") +
  coord_polar("y", start=0) +
  scale_fill_brewer(palette = "dark2")+
  theme_void() + theme(axis.line = element_blank(),
                        axis.text = element_blank(),
                        axis.ticks = element_blank(),
                        plot.title = element_text(hjust = 0.5, color = "#666666"))+
  geom_text(aes(y = data.Freq/3 + c(0, cumsum(data.Freq)[-length(data.Freq)]),
                label = percent(data.Freq/100)), size=5)
```

کد :



نمودار :

همانطور که از نمودار pie chart پیداست، حدود ۲۰ درصد از افراد حاضر در این مجموعه داده بین ۰ تا ۲۱ سال سن دارند، و حدود ۲۱ درصد از این افراد در بازه سنی بین ۶۵ تا ۸۲ سال هستند. همچنین قابل مشاهده است که اکثر افراد (حدود ۳۲ درصد) بین ۲۲ تا ۴۳ سال سن دارند و حدود ۲۷ درصد از افراد این مجموعه داده، بازه سنی ۴۴ تا ۶۴ سال را تشکیل می دهند.

H :

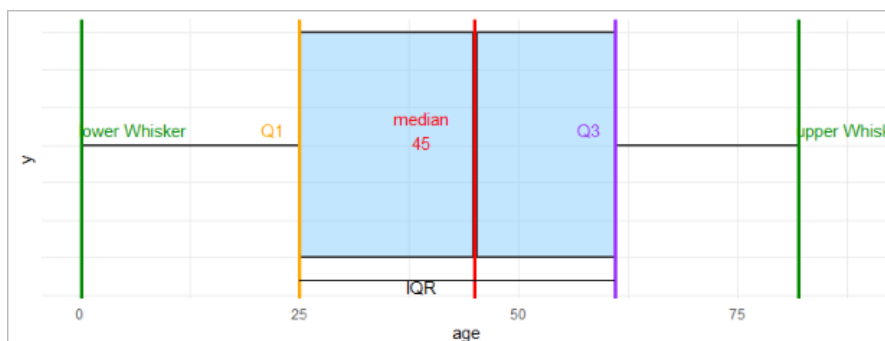
هدف از این قسمت درک آناتومی نمودار Boxplot است. می دانیم که Q1 (چارک اول) نقطه ای است که ۲۵ درصد داده ها پایین تر از آن قرار دارند. Q2 (همان میانه توزیع) نقطه ای که ۵۰ درصد داده ها پایین تر از آن قرار دارد. Q3 (چارک سوم) نقطه ای که ۷۵ درصد داده ها پایین تر از آن قرار دارد. و به فاصله بین چارک اول و چارک سوم IQR می گویند. همچنین برای مشخص کردن lower whisker و upper whisker باید بدانیم که اگر ۱/۵ برابر از IQR پایین بیاییم به سطحی می رسیم که lower whisker در آن قرار دارد و اگر به نقطه ی min زودتر برسیم، آنگاه نقطه min میشود همان lower whisker. و اگر ۱/۵ برابر از IQR بالاتر برویم به سطحی می رسیم که upper whisker در آن قرار دارد و اگر به نقطه ی max زودتر برسیم، آنگاه نقطه max میشود همان upper whisker.

برای مشخص کردن موارد ذکر شده و نمایش آنها بهترین راه استفاده از Boxplot است. در این قسمت با استفاده از کتابخانه ggplot اقدام به رسم این نمودار می کنیم و سپس موارد ذکر شده را با استفاده از geom_vline با رسم خط روی نمودار Boxplot نمایش می دهیم. همچنین برای بدست آوردن این مقادیر از تابع fivenum() استفاده شده است که به ترتیب : min، Q1، median، Q3 و Max را برای متغیر age محاسبه میکند و برمی گرداند. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است :

```
# fiveum function can return min , Q1 , Q2 , Q3 , max for age variable
five <- fiveum(HealthCare$age)
IQR <- five[4] - five[2]

# plot a boxplot to specify upper and lower quartiles, whiskers, and the IQR
ggplot(HealthCare) +
  aes(x=age) +
  geom_boxplot(alpha=0.5, fill="skyblue1", lwd=1) +
  theme_minimal() +
  theme(axis.text.y=element_blank(), axis.ticks.y=element_blank()) +
  # Q2 (median)
  geom_vline(aes(xintercept=median(age)), color="red1", size=1.3) +
  annotate(geom="text", x=median(HealthCare$age)-6, y=0.05,
    label=paste0("median\n", median(HealthCare$age)),
    color="red1") +
  # Q3 (third quartile)
  geom_vline(aes(xintercept=quantile(age,0.75)), color="purple1", size=1.3) +
  annotate(geom="text", x=quantile(HealthCare$age,0.75)-3, y=0.05,
    label=paste0("Q3"),
    color="purple1") +
  # Q1 (first quartile)
  geom_vline(aes(xintercept=quantile(age,0.25)), color="orange1", size=1.3) +
  annotate(geom="text", x=quantile(HealthCare$age,0.25)-3, y=0.05,
    label=paste0("Q1"),
    color="orange1") +
  # lower whisker
  geom_vline(aes(xintercept=max(five[2]-(1.5*IQR), five[1])), color="green4", size=1.3) +
  annotate(geom="text", x=max(five[2]-(1.5*IQR), five[1])+6, y=0.05,
    label=paste0("lower whisker"),
    color="green4") +
  # upper whisker
  geom_vline(aes(xintercept=min(five[4]+(1.5*IQR), five[5])), color="green4", size=1.3) +
  annotate(geom="text", x=min(five[4]+(1.5*IQR), five[5])+6, y=0.05,
    label=paste0("upper whisker"),
    color="green4") +
  # IQR
  geom_segment(aes(x = five[2], y = -0.45, xend = five[4], yend = -0.45)) +
  annotate(geom="text", x=median(HealthCare$age)-6, y=-0.47,
    label=paste0("IQR"),
    color="black")
```

کد :



نمودار :

سوال شماره ۲

متغیرهای Categorical در یک Dataset تعداد محدودی دسته (category) را می توانند اختیار کنند. در این سوال، متغیر انتخابی، "smoking_status" است که برای هر فرد (Case) در این Dataset وضعیت استعمال دخانیات آن را با یکی از category های : "smokes", "never smoked", "formerly smoked", یا "Unknown" مشخص می کند. لازم به ذکر است که وجود داشتن مقدار Unknown برای یک فرد (Case) به معنی در دست نبودن اطلاعات برای آن است.

A :

برای پیدا کردن فرکانس تکرار هر دسته و درصد آن در زبان R، به ترتیب از توابع table() و prop.table() استفاده میکنم تا بر روی متغیر smoking_status اعمال شود. در شکل زیر به ترتیب فرکانس تکرار هر category و درصد آن قابل مشاهده است.

```
# find the frequency of each category with table function
table(HealthCare$smoking_status)

# find the percentage of each category with prop.table function
prop.table(table(HealthCare$smoking_status))*100
```

کد :

```
> # find the frequency of each category with table function
> table(HealthCare$smoking_status)

formerly smoked    never smoked         smokes         Unknown
              885             1892             789             1544

> # find the percentage of each category with prop.table function
> prop.table(table(HealthCare$smoking_status))*100

formerly smoked    never smoked         smokes         Unknown
      17.31898      37.02544      15.44031      30.21526
```

پاسخ :

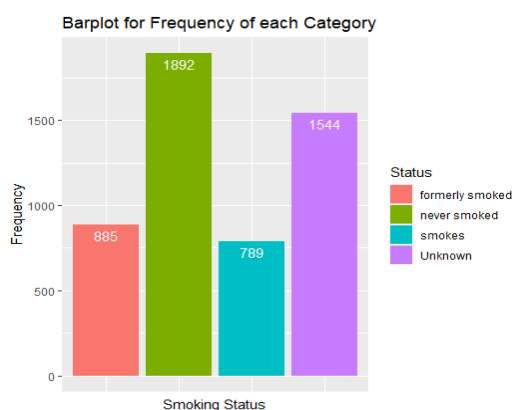
همانطور که از نتایج فرکانس تکرار هر category در متغیر smoking_status پیداست، ۸۸۵ نفر از افرادی (case) که در این Dataset قرار دارند در دسته formerly smoked قرار گرفته اند که حدود ۱۷/۳ درصد از بیماران را تشکیل میدهد. همچنین ۱۸۹۲ نفر در دسته never smoked، ۷۸۹ نفر در دسته smokes و ۱۵۴۴ نفر در دسته Unknown قرار گرفته اند که به ترتیب ۳۷ درصد، ۱۵/۴ درصد و ۳۰/۲ درصد از بیماران حاضر در این مجموعه داده را تشکیل می دهند. نکته جالبی که از نتایج استخراج می شود در دسترس نبودن اطلاعات حدود ۳۰ درصد از افراد داخل این dataset برای وضعیت استعمال دخانیات آن ها است.

می دانیم که برای مصورسازی متغیرهای categorical، نمودارهای متعددی وجود دارد که از رایج ترین آن ها می توان به BarPlot اشاره کرد. این نمودار برای مصورسازی یک متغیر categorical بکار می رود که شباهت هایی با نمودار Histogram که در سوال یک دیدیم دارد اما همانطور که می دانیم BarPlot برای مصورسازی متغیرهای categorical به کار می رود و اصلی ترین تفاوت آن با Histogram وجود داشتن فاصله بین میله ها و اهمیت داشتن ارتفاع هر میله است. در حقیقت در این نمودار ارتفاع هر میله، فرکانس تکرار آن category را نشان می دهد که در این قسمت از سوال در کنار مشخص کردن frequency هر category به کمک تابع table، از Barplot برای نمایش این مقدار برای هر category که به کمک کتابخانه ggplot رسم شده استفاده شده است. لازم به ذکر است که برای هر category از رنگ های متفاوتی استفاده شده است و تعداد افراد حاضر در هر category (فرکانس تکرار) بر روی میله مربوط به آن دسته قرار داده شده است. در شکل زیر که این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing libraries
library(dplyr)
library(ggplot2)
library(tidy)
library(scales) # for percentage scales

# find the frequency of each category with Barplot
ggplot(Healthcare, aes(x = smoking_status)) +
  geom_bar(aes(x = smoking_status, fill = smoking_status), position = "dodge") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "white") +
  labs(title="Barplot for Frequency of each Category ",
       x="Smoking Status", y="Frequency", fill="Status") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank())
```

کد :



نمودار :

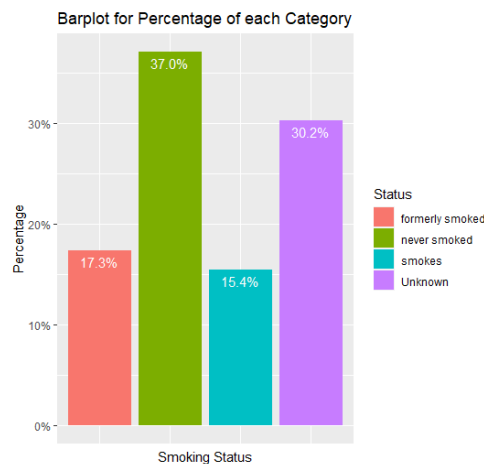
: B

همانطور که گفته شد BarPlot برای مصورسازی یک متغیر categorical بکار می رود در این قسمت از کتابخانه ggplot برای رسم نمودار barplot استفاده شده است. در این نمودار هر category با رنگ های متفاوت نمایش داده شده است. همچنین درصد هر دسته بر روی میله مربوط به آن نوشته شده است. همانطور که در قسمت A دیدیم و در این نمودار هم قابل مشاهده است، از بین افراد حاضر در این dataset، افرادی که در دسته formerly smoked قرار گرفته اند حدود ۱۷/۳ درصد از بیماران را تشکیل میدهد. همچنین ۳۷ درصد از افراد در دسته never smoked، ۱۵/۴ درصد در دسته smokes و ۳۰/۲ درصد در دسته Unknown قرار گرفته اند. در شکل زیر که این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing libraries
library(ggplot2)
library(scales) # for percentage scales

# find the percentage of each category with Barplot with different colors
# and add percentage marks to each bar
ggplot(data = Healthcare, aes(x = smoking_status,
                             y = prop.table(stat(count)),
                             fill = smoking_status,
                             label = scales::percent(prop.table(stat(count)))) +
  geom_bar(position = "dodge") +
  geom_text(stat = 'count',
            position = position_dodge(.9),
            vjust = 1.5,
            size = 4, colour = "white") +
  scale_y_continuous(labels = scales::percent) +
  labs(title="Barplot for Percentage of each Category ",
       x="Smoking Status",y="Percentage",fill="Status") +
  theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())
```

کد :



نمودار :

C :

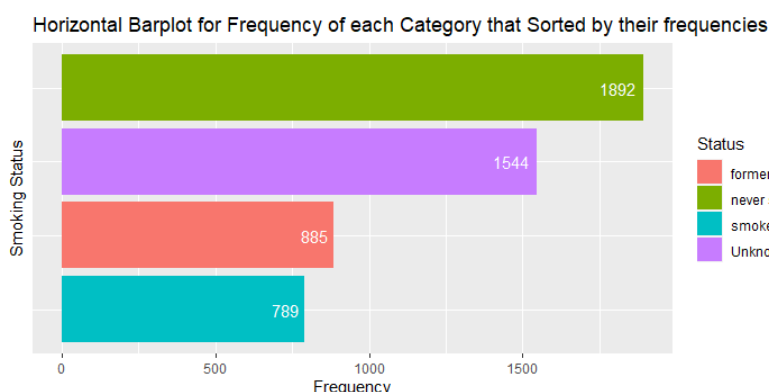
در این قسمت از سوال، برای مرتب سازی دسته های متغیر smoking_status براساس فرکانس تکرارشان، از یک تابع به نام reordering استفاده میکنم که در زمان صدا زدن آن، متغیر smoking_status را به عنوان پارامتر ورودی می گیرد و category های آن را براساس فرکانس تکرار به صورت صعودی مرتب کرده و سپس به کمک کتابخانه ggplot اقدام به رسم Horizontal BarPlot کرده. لازم به ذکر است که برای هر category از رنگ های متفاوتی استفاده شده است و تعداد افراد حاضر در هر category (فرکانس تکرار) بر روی میله مربوط به آن دسته قرار داده شده است. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing libraries
library(ggplot2)
library(scales) # for percentage scales

# A function that sorts my categorical variable categories by their frequencies.
reordering <- function(inp) {
  factor(inp, levels = names(sort(table(inp), increasing = TRUE)))
}

# plotting a Horizontal Barplot for my categorical variable that sorted by their frequencies.
ggplot(Healthcare, aes(x = reordering(smoking_status)) +
  geom_bar(aes(x = reordering(smoking_status), fill = smoking_status), position = "dodge") +
  geom_text(aes(label = ..count..), stat = "count", hjust = 1.2, size = 4, colour = "white") +
  labs(title="Horizontal Barplot for Frequency of each Category that Sorted by their frequencies",
       x="Smoking Status",y="Frequency",fill="Status") +
  theme(axis.text.y=element_blank(),axis.ticks.y=element_blank()) +
  coord_flip()
```

کد :



نمودار :

: D

نمودار ویولنی یک روش برای مصورسازی متغیرهای numerical است که بسیار شبیه به Boxplot است اما اطلاعات بیشتری را نسبت به Boxplot به ما می دهد و زمانی استفاده می شود که می خواهیم توزیع آماری تعداد زیادی داده را در Box Plot داشته باشیم، اما نمایش همه داده ها روی نمودار از خوانایی آن می کاهد. در این قسمت از سوال به دلیل اینکه باید این نمودار را برای یک متغیر categorical رسم می کردیم، و از آنجایی که این نمودار برای مصورسازی متغیرهای عددی است، من این نمودار را برای متغیر smoking_status در مقابل متغیر عددی age رسم کرده ام.

در این نمودار برای هر category علاوه بر نمودار ویولنی آن، یک boxplot هم در آن رسم شده است تا بتوانیم اطلاعاتی نظیر میانه و IQR را هم داشته باشیم. می دانیم که بخش های پهن تر این نمودار نشان دهنده این است که نمونه ها در داده مورد نظر با احتمال بیشتری این مقدار را می توانند بگیرند و هر چه برای یک مقدار این پهنای کوچکتر باشد احتمال آن کمتر است.

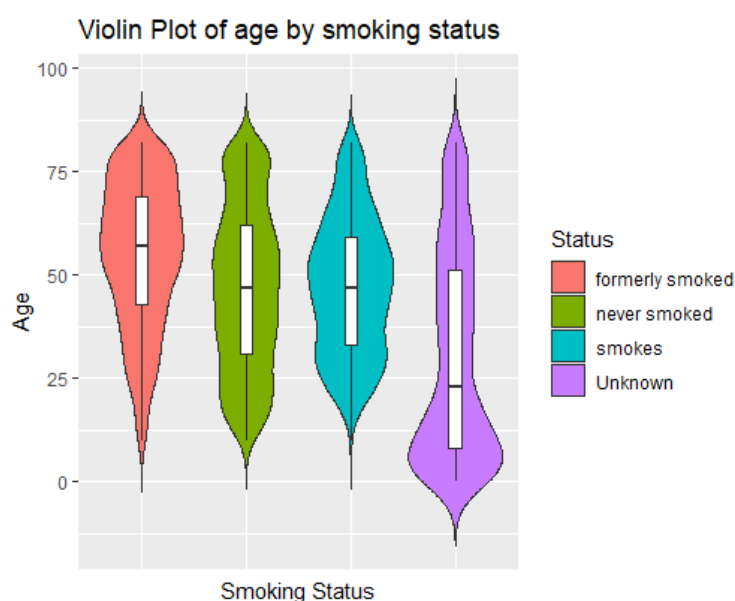
همانطور که در شکل زیر مشاهده می شود، میانه سن بیماریانی که وضعیت سیگار کشیدن آن ها formerly smoked شده است، از سه category دیگر بیشتر است، همچنین میانه سن بیماریانی که اطلاعاتی برای وضعیت سیگار کشیدن آن ها در دست نداریم و در دسته Unknown قرار می گیرند از بقیه کمتر است. و تقریباً می توان گفت میانه سن افرادی که در دسته های never smoked و smokes قرار گرفته اند با هم برابر است.

همچنین با توجه به این نمودار با احتمال بیشتری، افرادی که در مورد وضعیت سیگار کشیدن آن ها اطلاعاتی در دست نداریم و در دسته Unknown قرار می گیرند، دارای سنین کمتری نسبت به افراد حاضر در سه category دیگر هستند. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing libraries
library(ggplot2)

# Plotting a violin plot for smoking status and age variables
ggplot(HealthCare, aes(x = smoking_status, y = age )) +
  geom_violin(trim = FALSE, aes(fill = smoking_status)) +
  geom_boxplot(width = 0.12,)+
  labs(title="Violin Plot of age by smoking status",
       x="Smoking Status",y="Age",fill="Status") +
  theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())
```

کد :



نمودار :

دو متغیر انتخابی برای این سوال، عبارتند از: متغیر عددی age که سن افراد حاضر در این مجموعه داده را مشخص می‌کند، و متغیر عددی Health bills که میزان هزینه سالیانه سلامتی هر فرد (Case) درون این مجموعه داده را مشخص می‌کند.

A :

در آمار رابطه ای که بین دو متغیر یافت می‌شود ویژگی‌هایی از جمله جهت، شکل و شدت خواهد داشت، اما میدانیم که وجود هر ارتباطی بین دو متغیر دلیل بر نتیجه‌گیری علی نخواهد بود و نمی‌توان از همبستگی بین دو متغیر، نتیجه‌گیری علی کرد. اما در این قسمت هدف زدن حدس و گمان در مورد رابطه بین دو متغیر انتخابی است.

از آنجایی که متغیر age سن افراد مورد مطالعه ما را نشان می‌دهد، و از آن طرف متغیر Health bills میزان هزینه ای که آن‌ها در سال برای سلامتی خود می‌پردازند را مشخص میکند، می‌توان حدس زد که با افزایش سن افراد میزان هزینه‌های سالیانه آن‌ها هم به دلیل کهولت سن و مشکلاتی که در افرادی با سنین بالاتر رخ می‌دهد افزایش می‌یابد. اما همانطور که گفتیم این صرفاً یک حدس است و نمیتوان به دلیل وجود همبستگی بین دو متغیر، نتیجه‌گیری علی کرد.

در واقع به نظر میرسد که یک ارتباط خطی مثبت بین این دو متغیر وجود داشته باشد.

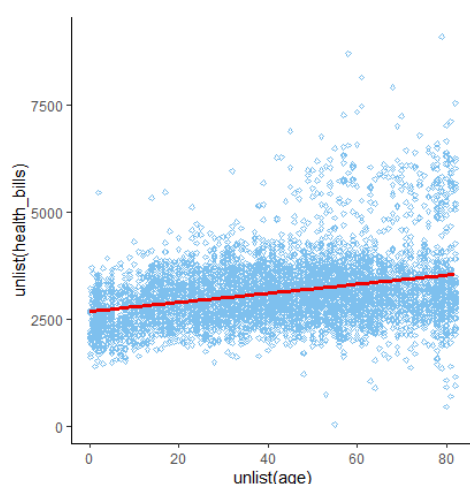
B :

Scatter plot یکی از روش‌های مصورسازی متغیرهای عددی است که اغلب از آن برای نمایش رابطه بین دو متغیر عددی استفاده میکنند. در این نمودار ارتباط بین دو متغیر را با یک خط regression ای یا یک منحنی نمایش می‌دهند که این ارتباط میتواند مثبت یا منفی باشد. همچنین در این نمودار قادر به شناسایی outlierها هم هستیم.

ابتدا با استفاده از کتابخانه ggplot در R این نمودار را رسم می‌کنیم، کد و نمودار حاصل از اجرای آن در شکل زیر قابل مشاهده است :

```
# importing library
library(ggplot2)
# plot a scatter plot
ggplot(sampp, aes(x=unlist(age), y=unlist(health_bills))) +
  geom_point(size=1, shape=23, color="skyblue2") +
  geom_smooth(method=lm, se=FALSE, color="red2", size=1.3) +
  theme_classic()
```

کد :



نمودار :

همانطور که در نمودار scatter plot مشاهده می‌شود یک ارتباط خطی و مثبت بین این دو متغیر وجود دارد و از آنجایی که اغلب data pointها به خط نزدیک هستند می‌توان گفت یک رابطه قوی بین این دو متغیر وجود دارد. البته شاهد تعدادی outlier هم هستیم اما آن چیزی که مشاهده می‌شود، با افزایش سن افراد میزان هزینه‌های سالیانه آن‌ها هم افزایش می‌یابد.

C و D : (اگر موردی نداشته باشد، این دو قسمت را با هم جواب می دهیم)

در تحلیل‌های آماری، روش‌های مختلفی برای محاسبه ارتباط یا همبستگی بین دو متغیر وجود دارد. منظور از ضریب همبستگی بین دو متغیر، قابلیت پیش‌بینی مقدار یکی برحسب دیگری است. می‌دانیم که یک ارتباط بین دو متغیر می‌تواند مقداری بین -۱ تا +۱ داشته باشد. (اگر ارتباط بین دو متغیر صفر باشد یعنی هیچ ارتباطی با هم ندارند). هر چه ارتباط بین دو متغیر به مقدار +۱ یا مقدار -۱ نزدیک‌تر باشد گوییم آن ارتباط قوی است. همچنین هر چه ارتباط بین دو متغیر به مقدار +۱ نزدیک باشد، گوییم این ارتباط مثبت است و هر چه به مقدار -۱ نزدیک‌تر باشد، نشان‌دهنده‌ی منفی بودن ارتباط این دو متغیر است.

ارتباط مثبت بین دو متغیر یعنی با افزایش یافتن یکی، مقدار دیگری هم افزایش یابد و این افزایش در واقع به سمت بالا جهت داشته باشد، و ارتباط منفی بین دو متغیر یعنی با کاهش یافتن یکی، مقدار دیگری هم کاهش یابد و این کاهش یک خط رو به پایین باشد.

با توجه به توضیحات داده شده و نمودار scatter plot رسم شده در قسمت قبل، میتوان فهمید که ارتباط بین سن افراد و میزان هزینه‌ای که سالیانه برای سلامت خود می‌پردازند، یک ارتباط خطی و مثبت است. اما این بدین معنی نیست که ما میتوانیم از این همبستگی که بین این دو متغیر وجود دارد نتیجه‌گیری علی انجام دهیم.

همچنین برای محاسبه‌ی ضریب همبستگی بین دو متغیر می‌توان از روش‌های مختلفی از جمله، پیرسون، اسپیرمن و کندال استفاده کرد. ضریب همبستگی پیرسون براساس میانگین و واریانس محاسبه می‌شود و ممکن است در زمانی که outlier داریم میزان همبستگی را به درستی نشان ندهد. در چنین مواقعی از ضریب همبستگی اسپیرمن استفاده می‌شود. اما در این سوال برای محاسبه ضریب همبستگی بین این دو متغیر عددی از تابع $\text{cor}()$ در R استفاده میکنم. کد و نتیجه حاصل از اجرای آن در شکل زیر قابل مشاهده است :

<pre>library("ggpubr") # calculate the correlation coefficient for these two variables cor(unlist(sampp\$age),unlist(sampp\$health_bills), method = c("pearson", "kendall", "spearman"))</pre>	کد :
<pre>> cor(unlist(sampp\$age),unlist(sampp\$health_bills), method = c("pearson", "kendall", "spearman")) [1] 0.2964851</pre>	پاسخ :

همانطور که در قسمت A این سوال حدس زده شد و با توجه به مقدار ضریب همبستگی بدست آمده، مشاهده می‌شود که مقدار این عدد مثبت و نشان‌دهنده‌ی مثبت بودن رابطه بین این دو متغیر است. و این یعنی با افزایش سن افراد میزان هزینه‌های سالیانه آن‌ها که برای سلامتی آن‌ها پرداخته می‌شود هم افزایش می‌یابد.

E :

آزمون correlation برای بررسی رابطه بین دو متغیر به کار می‌رود. به عنوان مثال در این آزمایش بررسی می‌شود که آیا رابطه‌ای بین سن افراد و میزان هزینه‌ای که سالیانه برای سلامتی‌شان پرداخت میکنند وجود دارد یا خیر. برای اینکار ضریب همبستگی این دو متغیر محاسبه می‌شود، و با صفر مقایسه می‌شود. در واقع این آزمون بدین صورت خواهد بود که :

$$H_0: \text{correlation coefficient} = 0$$

$$H_A: \text{correlation coefficient} \neq 0$$

فرض صفر به این معنی است که بین این دو متغیر رابطه‌ای وجود ندارد (یعنی ضریب همبستگی صفر است)
فرض جایگزین به این معنی است که بین این دو متغیر رابطه‌ای (منفی یا مثبت) وجود دارد و ضریب همبستگی صفر نیست.

در این آزمایش از یک significance level یا همان آلفا استفاده می‌شود و مقدار p-value حساب شده را با آلفا مقایسه میکنیم و تصمیم می‌گیریم که آیا شواهد کافی برای رد فرضیه صفر داریم و ضریب همبستگی برابر صفر نیست و بین این دو متغیر رابطه‌ای وجود دارد، یا برعکس شواهد کافی برای رد فرضیه صفر نداریم و ضریب همبستگی برابر صفر است و بین این دو متغیر رابطه‌ای (یا مثبت یا منفی) وجود ندارد.

بنابراین در این قسمت میخواهیم ضریب همبستگی بدست آمده در قسمت قبل را آزمایش کنیم. برای این کار از تابع $\text{cor.test}()$ در R استفاده میکنیم. کد و نتیجه اجرای این تابع برای دو متغیر انتخابی در شکل زیر آمده است :

```
# Correlation testing
cor.test(unlist(sampp$age),unlist(sampp$health_bills), method = c("pearson", "kendall", "spearman"))
```

کد :

```
data: unlist(sampp$age) and unlist(sampp$health_bills)
t = 22.187, df = 5108, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2712710 0.3212924
sample estimates:
      cor
0.2964851
```

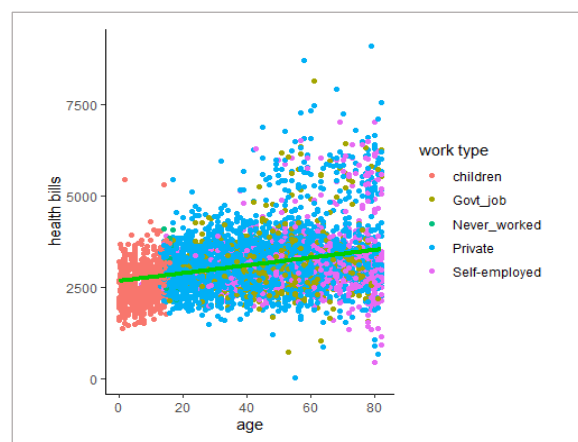
پاسخ :

مقدار p-value آزمون correlation بر اساس ضریب همبستگی محاسبه می شود و این احتمال را نشان می دهد که همبستگی بین دو متغیر به طور تصادفی رخ داده است. همانطور که از نتیجه آزمون مشخص است، شواهد کافی و قانع کننده ای برای رد فرضیه صفر داریم و ضریب همبستگی برابر صفر نیست و در واقع بین دو متغیر سن و میزان هزینه سالیانه سلامتی رابطه ای وجود دارد. و از آنجایی که ضریب همبستگی مثبت است (برابر با عددی که در قسمت قبل بدست آوردیم)، ارتباط بین این دو متغیر یک ارتباط خطی مثبت است. یعنی با افزایش یکی دیگری هم افزایش میابد. اما همانطور که گفتیم، وجود همبستگی و ارتباط بین دو متغیر دلیل بر وجود داشتن رابطه علی بین این دو نیست.

: F

در این قسمت برای نمودار scatter plot رسم شده در قسمت b این سوال ، متغیر work_type انتخاب شده است. این متغیر نوع شغل افراد مورد مطالعه ما را نشان می دهد. ابتدا با استفاده از کتابخانه ggplot در R این نمودار را رسم می کنیم، سپس رنگ و سمبل هر data point را با استفاده از متغیر work_type در این نمودار مشخص می کنیم. کد و نمودار حاصل از اجرای آن در شکل زیر قابل مشاهده است :

```
# importing library
library(ggplot2)
# plot a scatter plot
ggplot(sampp, aes(x=unlist(age), y=unlist(health_bills),
  colour = unlist(work_type))) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE,color="green3",size=1.6)+
  theme_classic()+
  labs(x="age",y="health bills",colour="work type")
```



در واقع این نمودار رابطه بین سن افراد و میزان هزینه ای که سالیانه برای سلامت خود می پردازند را براساس نوع شغل هر فرد نشان می دهد. جالب است همانطور که مشاهده می شود، افرادی که سنین پایینی دارند و به عبارتی children محسوب می شوند در پایین نمودار مشخص هستند و همچنین افراد در سنین بالاتر تمایل بیشتری به self-employed و دارند.

: G

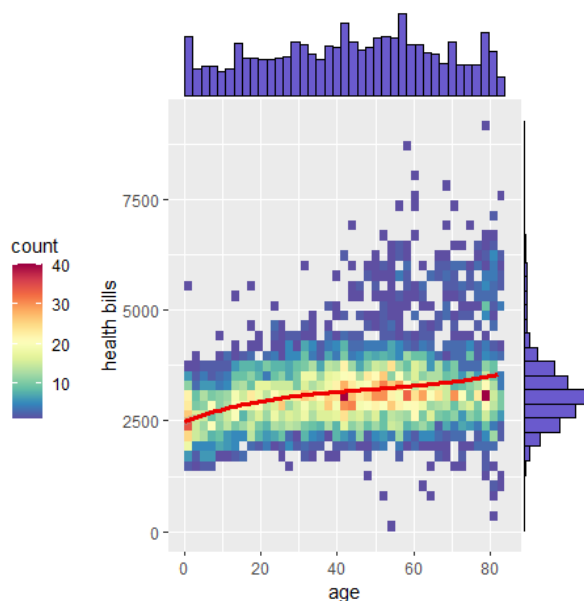
نمودار hexbin برای نشان دادن رابطه بین ۲ متغیر عددی در مواقعی که data point های زیادی وجود دارد مفید است. در این نمودار شکل نمایشی، به چندین شش ضلعی تقسیم شده است و همه داده ها در این شش ضلعی ها (hexagonal regions) بر اساس طیف رنگی (color gradient) که نشان دهنده ی تراکم data point ها در آن جا می باشد قرار می گیرند. در این نمودار نقاطی که پر رنگ تر هستند نشان دهنده ی تعداد data point های بیشتر است و این نمودار نسبت به scatter plot اطلاعات بیشتری را به ما میدهد.

در این قسمت با استفاده از کتابخانه ggplot نمودار Hexbin را با توزیع Marginal برای دو متغیر عددی age و health bill رسم می کنیم. کد و نمودار حاصل از اجرای آن در شکل زیر قابل مشاهده است :

```
# importing libraries
library(ggplot2)
library(hexbin)
library(RColorBrewer)

# plot a hexbin plot with marginal distribution and fitting curve for
# age and health bills variables
plot <- ggplot(sampp, aes(unlist(age), unlist(health_bills))) + stat_bin2d(bins=40) +
  scale_fill_gradientn(colours=r) + geom_point(size=-1) +
  theme(legend.position="left") +
  geom_smooth(method = lm, formula = y ~ splines::bs(x, 3),
             se = TRUE, color="red2", size=1.3) +
  labs(x="age", y="health bills", title="Hexbin plot")
ggMarginal(plot, type="histogram", fill = "slateblue", xparams = list( bins=40))
```

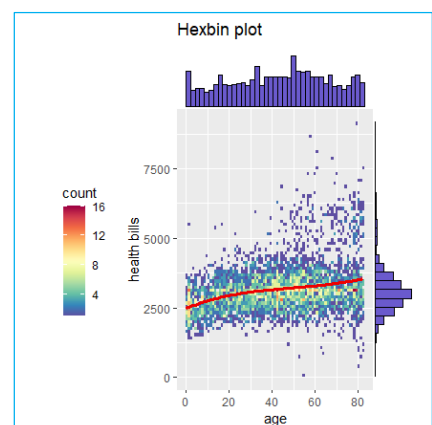
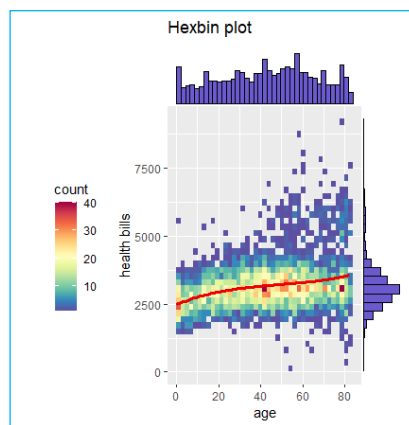
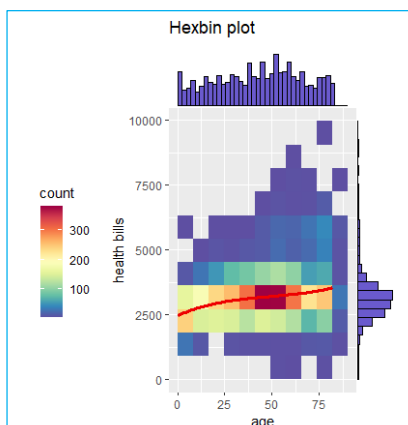
Hexbin plot



همانطور که در گراف رسم شده قابل مشاهده است، یک ارتباط مثبت بین این دو متغیر وجود دارد و از آنجایی که اغلب data point ها به خط نزدیک هستند می توان گفت یک رابطه قوی بین این دو متغیر وجود دارد. که البته همانطور که گفته شد این نمودار نسبت به scatter plot اطلاعات بیشتری را می دهد و نقاطی که پر رنگ تر هستند نشان دهنده ی تعداد data point های بیشتر است. این را می توان در این گراف مشاهده کرد که شش ضلعی های نزدیک curve پر رنگ تر و نشان دهنده ی وجود data point های بیشتری هستند. همینطور نکته جالب توجه این نمودار، هیستوگرام های آن برای دو متغیر ما است که به نوعی توزیع این دو متغیر را نشان میدهند. به عنوان مثال برای متغیر health bill نشان دهنده این است که برای اغلب افراد میزان هزینه سالیانه ای که برای سلامت خود باید پرداخت کنند، بین ۲۰۰۰ تا ۳۰۰۰ است و بیشترین تکرار را، همانگونه که هیستوگرام آن هم نشان میدهد، در اطراف نقطه ۲۵۰۰ داریم. اما طبق چیزی که در قسمت های قبل هم دیدیم، با افزایش سن افراد میزان هزینه های سالیانه آن ها هم افزایش میابد.

در این نمودار با عوض کردن سایز bin، مشخصاً تعداد شش ضلعی های آن و سایز آنها تغییر خواهد کرد. به دلیل اینکه درون این شش ضلعی ها یا همان bin ها، data point های نزدیک به هم قرار میگیرند. بنابراین هرچه سایز bin ها را افزایش دهیم، تعداد data point هایی که درون یک bin قرار میگیرند زیاد تر میشود و در نتیجه با شش ضلعی های (hexagonal regions) بزرگ تری روبه رو خواهیم شد که ناحیه بزرگتری از data point ها را پوشش داده اند. همینطور برعکس.

در شکل زیر یکبار این نمودار با سایز bin کوچک، یکبار با سایز bin مناسب، و بار دیگر با سایز bin بزرگ رسم شده اند و در کنار هم به ترتیب از راست به چپ قرار گرفته اند. به خوبی قابل مشاهده است که در نموداری که سایز bin آن بزرگ است (نمودار سمت چپ) تعداد data point هایی که درون یک bin قرار میگیرند زیاد تر میشود و در نتیجه با شش ضلعی های (hexagonal regions) بزرگ تری روبه رو خواهیم شد که ناحیه بزرگتری از data point ها را پوشش داده اند.

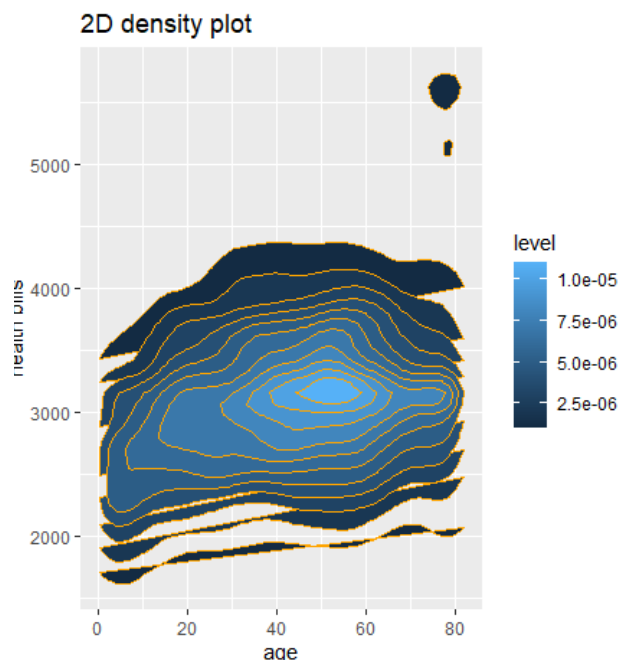


: H

نمودار density plot تابع Probability Density Function را برای یک متغیر عددی به تصویر می کشد. اما زمانی که تعداد data point ها زیاد است بهتر است که از نمودار 2D Density Plot استفاده کنیم. به دلیل اینکه این نمودار در جاهایی که بیش از حد overplotting داریم مفید است و اطلاعات بیشتری به ما می دهد. در شکل زیر با استفاده از کتابخانه ggplot نمودار 2D Density Plot را برای دو متغیر عددی age و health bill رسم می کنیم. کد و نمودار حاصل از اجرای آن در شکل زیر قابل مشاهده است :

```
# importing libraries
library(ggplot2)
# Draw the 2D density plot for age and health bills variables
ggplot(sampp, aes(x=unlist(age),
                  y=unlist(health_bills))) +
  stat_density_2d(aes(fill = ..level..),
                  geom = "polygon", colour="orange")+
  labs(x="age",y="health bills",title="2D density plot")
```

گفتیم که این نمودار برای مواقعی که داده های زیادی داریم و در scatterplot ، اصطلاحاً overplotting رخ می دهد، اطلاعات بیشتری به ما میدهد. مثلاً می توان مشاهده کرد که در نمودار رو به رو غالب افراد در بازه سنی ۴۰ تا ۶۰ سال قرار دارند، و در سال میزان هزینه ای که برای سلامتی خود پرداخت می کنند حدود ۳۰۰۰ دلار است. در این نمودار می توانیم بفهمیم که در کجا ها تراکم داده های زیادی داریم و هر چه این تراکم کمتر شود مشاهده می شود که رنگ این نمودار در آن نقاط تیره تر است.



سوال شماره ۴

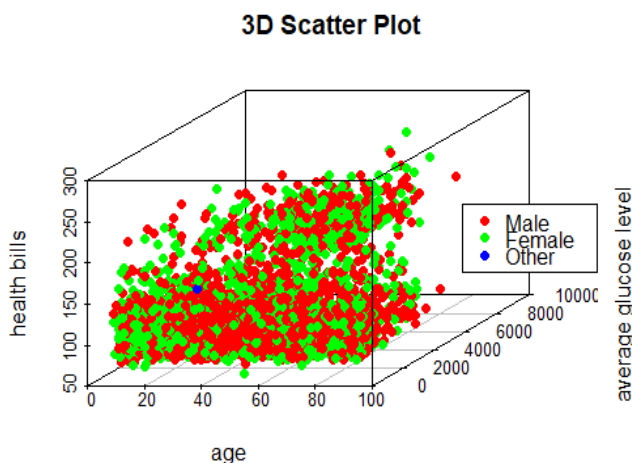
---- : A

---- : B

: C

از نمودار SD scatter plot به منظور نمایش رابطه بین دو متغیر عددی با در نظر گرفتن متغیر سوم استفاده می شود. در این سوال متغیر اول age انتخاب شده است که سن افراد مورد مطالعه در این مجموعه داده را نشان می دهد. متغیر دوم health bills است که میزان هزینه ای که در سال این افراد برای سلامتی خود می پردازند. و متغیر سوم که می خواهیم با در نظر گرفتن آن ، ارتباط بین دو متغیر دیگر را بدانیم متغیر ave glucose level است که متوسط قند خون افراد را نشان می دهد. همچنین از متغیر categorical جنسیت gender استفاده شده است تا رنگ data point ها را به سه رنگ سبز (برای زن) و قرمز (برای مرد) و همچنین آبی (برای other) دسته بندی کنیم. برای رسم این نمودار از کتابخانه ggplot استفاده شده است و کد آن به همراه نتیجه حاصل از اجرای کد در شکل زیر قابل نمایش است.

```
# importing libraries
library("scatterplot3d")
# my categorical variable is gender
catt <- c("red","green","blue")
catt <- catt[transform(unlist(sampp22$gender ),
                        id=as.numeric(factor(unlist(sampp22$gender ))))$id]
# Draw a 3D scatterplot for the numerical variables
# and use the categorical variable as points' color
scatterplot3d(unlist(sampp22$age),unlist(sampp22$health_bills),
              unlist(sampp22$avg_glucose_level),main="3D Scatter Plot",
              xlab = "age",ylab = "average glucose level",zlab = "health bills",
              color=catt, pch = 16)
# plot a legend for gender
legend("right", legend = c("Male","Female","Other"),
      col = c("red", "green","blue"), pch = 16)
```



آنچه که مشخص است غالب data point ها در پایین این نمودار هستند. اما به نظر میرسد از این نمودار نتیجه خاصی نمیتوان گرفت و نمیتوان به درستی رابطه بین دو متغیر سن و میزان هزینه های سلامتی سالیانه افراد، با در نظر گرفتن متوسط سطح قند خون آنها، مشخص کرد. پراکندگی داده ها به گونه ای است که نمی توان رابطه خاصی بین این سه متغیر پیدا کرد. اما تا حدودی مشخص است در سنین بالاتر، افراد هزینه سالیانه بالاتری برای سلامتی خود پرداخت میکنند و این میتواند تاثیر گرفته از سطح قند خون بالای آنها باشد که ممکن است مشکلاتی نظیر سکته قلبی و ... به همراه داشته باشد.

سوال شماره ۵

متغیرهای Categorical در یک Dataset تعداد محدودی دسته (category) را می توانند اختیار کنند. در این سوال، متغیرهای انتخابی عبارتند از: متغیر gender که جنسیت افراد مورد مطالعه را مشخص میکند، و متغیر work_type که نوع شغل این افراد را در پنج دسته "Self-Private"، "Never_worked"، "Govt_jov"، "children" یا "employed" قرار میدهد.

A:

می دانیم که برای مصورسازی متغیرهای categorical، نمودارهای متعددی وجود دارد که یکی از رایج ترین آن ها BarPlot بود که در سوال دوم به آن پرداخته شد. اما این نمودار تنها برای مصور سازی یک متغیر categorical بکار می رود. برای اینکه بتوانیم دو متغیر categorical را مصورسازی کنیم و رابطه بین آن ها را نمایش دهیم می توانیم به سراغ Frequency/Contingency table برویم.

ابتدا به کمک کتابخانه grid و gridExtra اقدام به رسم این جدول در R می کنیم، کد این جدول به همراه نتیجه حاصل از اجرای کد در شکل زیر قابل مشاهده است:

```
# importing libraries
library(gridExtra)
library(grid)

# draw Frequency/Contingency table for gender and work type variables
my_table = table(gender=HealthCare$gender,worktype=HealthCare$work_type)
grid.table(addmargins(my_table,FUN = sum),theme=ttheme_minimal(
  core=list(bg_params = list(fill = "azure2", col=NA),
    fg_params=list(fontface=1)),
  colhead=list(fg_params=list(col="skyblue")),
  rowhead=list(fg_params=list(col="orange"))))
```

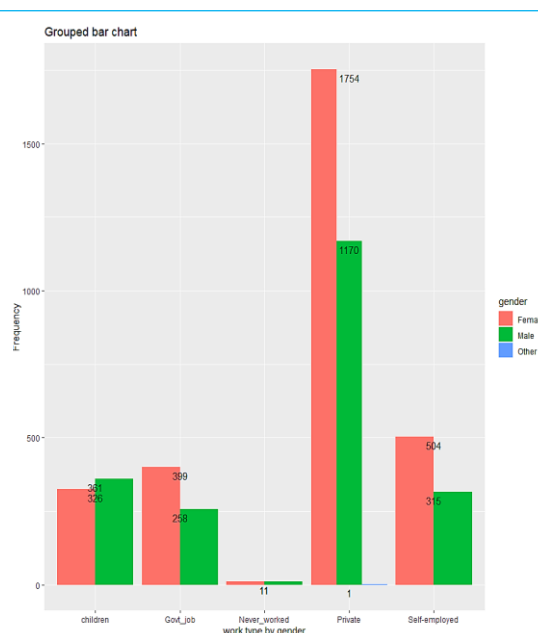
	children	Govt_job	Never_worked	Private	Self-employed	sum
<i>Female</i>	326	399	11	1754	504	2994
<i>Male</i>	361	258	11	1170	315	2115
<i>Other</i>	0	0	0	1	0	1
<i>sum</i>	687	657	22	2925	819	5110

این جدول تعداد افرادی که در هر دسته شغلی قرار میگیرند را بر اساس جنسیت آن ها نمایش می دهد. به عنوان مثال، از بین افراد مورد مطالعه ما در مجموعه داده، جمعاً ۲۹۲۵ نفر شغل خصوصی دارند که ۱۷۵۴ نفر از این افراد زن، و ۱۱۷۰ نفر مرد هستند. همچنین ۱ نفر که جنسیتی غیر از زن و مرد دارد در دسته افرادی قرار دارد که شغل خصوصی دارند.

: B

همانطور که گفته شد BarPlot برای مصور سازی یک متغیر categorical بکار می رود. در این نمودار ارتفاع هر میله، فرکانس تکرار آن category را نشان می دهد. اما در این سوال با استفاده از دو متغیر categorical اقدام به رسم این نمودار می کنیم. ابتدا به کمک کتابخانه ggplot آن را رسم کرده و فرکانس تکرار هر گروه شغلی را بر اساس جنسیت افراد گروه بندی می کنیم. در هر category از رنگ های متفاوتی استفاده شده است و تعداد افراد حاضر در هر category (فرکانس تکرار) بر روی میله مربوط به آن دسته قرار داده شده است. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing library
library(ggplot2)
# Grouped bar chart for gender and work type variables
ggplot(Healthcare, aes(x = work_type, group = gender)) +
  geom_bar(aes(x = work_type, fill = gender), position = "dodge") +
  geom_text(aes(label = ..count..), stat = "count",
    vjust = 1.5, colour = "black") +
  labs(title="Grouped bar chart",
    x="work type by gender",y="Frequency",fill="gender")
```



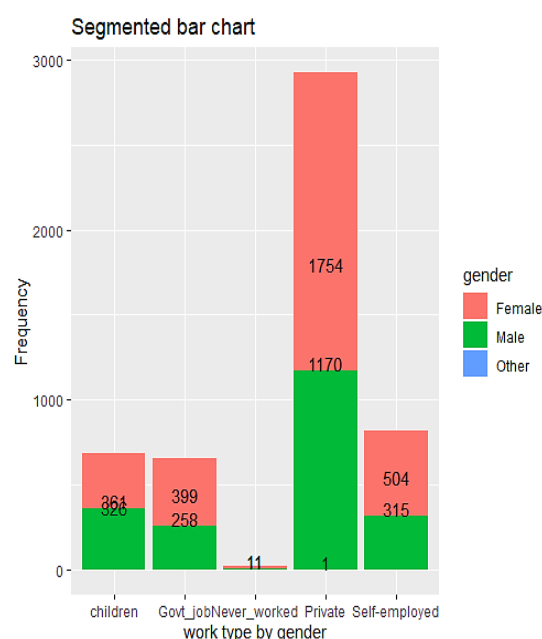
همانطور که در این نمودار قابل مشاهده است، به عنوان نمونه، از بین افراد حاضر در این مجموعه داده، جمعاً ۲۹۲۵ نفر شغل خصوصی دارند که ۱۷۵۴ نفر از این افراد زن، و ۱۱۷۰ نفر مرد هستند. همچنین ۱ نفر که جنسیتی غیر از زن و مرد دارد در دسته افرادی قرار دارد که شغل خصوصی دارند.

: C

Segmented Barplot برای مصورسازی توزیع های احتمال شرطی مفید است. در این نمودار هر Bar دارای چندین قطعه است که در هر قطعه یک category نشان داده می شود. در این سوال با استفاده از دو متغیر categorical اقدام به رسم این نمودار می کنیم. ابتدا به کمک کتابخانه ggplot آن را رسم کرده، و موقعیت هر category را به صورت stack رو هم انباشته می کنیم. در هر category از رنگ های متفاوتی استفاده شده است و تعداد افراد حاضر در هر category (فرکانس تکرار) بر روی میله مربوط به آن دسته و segment مربوطه قرار داده شده است. در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد آمده است.

```
# importing library
library(ggplot2)
# Segmented bar plot for gender and work type variables
ggplot(HealthCare, aes(x = work_type, group = gender)) +
  geom_bar(aes(x = work_type, fill = gender),
    position = position_stack(vjust = .1)) +
  geom_text(aes(label = ..count.., stat = "count",
    vjust = .01, colour = "black")) +
  labs(title="segmented bar chart",
    x="work type by gender", y="Frequency", fill="gender")
```

همانطور که در این نمودار قابل مشاهده است، به عنوان نمونه، از بین افراد حاضر در این مجموعه داده، جمعاً ۲۲ نفر تا به الان شغلی نداشته اند که ۱۱ نفر از این افراد زن، و ۱۱ نفر هم مرد هستند.



: D

یکی دیگر از روش های مرسوم برای نشان دادن رابطه بین دو متغیر categorical استفاده از نمودار Mosaic plot است. این نمودار شباهت زیادی به نمودار segmented plot دارد اما تفاوت این نمودار این است که در اینجا عرض هر Bar مهم است. در این سوال با استفاده از دو متغیر categorical gender و work_type اقدام به رسم این نمودار می کنیم. ابتدا به کمک کتابخانه ggplot آن را رسم می کنیم. در هر category درون هر Bar از رنگ های متفاوتی بر اساس جنسیت استفاده شده است و تعداد افراد حاضر در هر category (فرکانس تکرار) بر روی میله مربوط به آن دسته و segment مربوطه قرار داده شده است. همچنین درصد هر category روی آن برجسب گذاری شده است در شکل زیر کد این نمودار و نتیجه حاصله از اجرای کد قابل مشاهده است.

```
# importing library
library(ggplot2)
library(ggmosaic)

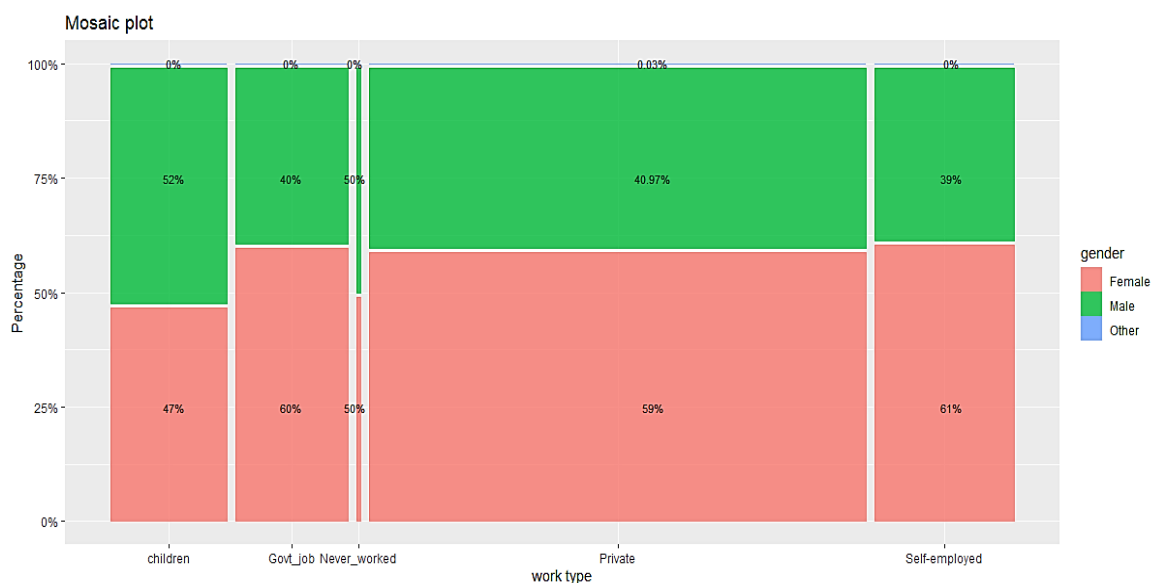
# Mosaic plot
ggplot(data = Healthcare) +
  geom_mosaic(aes(x = product(work_type),
                  group = gender,
                  weight = prop.table(stat(count)),
                  fill = gender,
                  label = scales::percent(prop.table(stat(count)))))) +
  scale_y_continuous(labels = scales::percent) +
  # for labeling
  labs(title="Mosaic plot ",
       x="work type",y="Percentage",fill="gender") +
  annotate("text",x=0.07,y=0.25,label="47%",lwd=3)+
  annotate("text",x=0.07,y=0.75,label="52%",lwd=3)+
  annotate("text",x=0.07,y=1,label="0%",lwd=3)+

  annotate("text",x=0.2,y=0.25,label="60%",lwd=3)+
  annotate("text",x=0.2,y=0.75,label="40%",lwd=3)+
  annotate("text",x=0.2,y=1,label="0%",lwd=3) +

  annotate("text",x=0.275,y=0.25,label="50%",lwd=3)+
  annotate("text",x=0.275,y=0.75,label="50%",lwd=3)+
  annotate("text",x=0.275,y=1,label="0%",lwd=3)+

  annotate("text",x=0.6,y=0.25,label="59%",lwd=3)+
  annotate("text",x=0.6,y=0.75,label="40.97%",lwd=3)+
  annotate("text",x=0.6,y=1,label="0.03%",lwd=3)+

  annotate("text",x=0.93,y=0.25,label="61%",lwd=3)+
  annotate("text",x=0.93,y=0.75,label="39%",lwd=3)+
  annotate("text",x=0.93,y=1,label="0%",lwd=3)
```



همانطور که در این نمودار مشاهده می شود، به عنوان نمونه، از بین افرادی که شغل خصوصی دارند، مردها ۴۰/۹۷ درصد، زن ها ۵۹ درصد و بقیه حدود ۰/۰۳ درصد از این افراد را تشکیل می دهند.

سوال شماره ۶

متغیرهای عددی (Numerical) در یک Dataset به متغیرهای کمی معروفند که مقادیر عددی را اختیار می کنند. این متغیرها دو نوع اند : متغیرهای عددی گسسته، که تعداد قابل شمارشی مقدار می توانند اختیار کنند و متغیرهای عددی پیوسته، که در یک رنج مشخص تعداد زیادی مقدار مختلف را اختیار می کنند. . برای این سوال متغیر انتخابی "age" است که در واقع سن هر فرد (Case) در این مجموعه داده را مشخص می کند.

A :

می دانیم که احتمال اینکه تخمین نقطه ای که برای پارامتر واقعی جامعه آماری مان در یک مطالعه با استفاده از یک sample میزنیم خیلی احتمال کمی دارد که دقیقاً برابر با پارامتری شود که به دنبال آن هستیم. در مفهوم بازه اطمینان به دنبال ارائه یک بازه هستیم که بتوانیم با اطمینان بالایی بگوییم که این بازه شامل پارامتر واقعی جامعه هدف می شود. در این سوال پارامتری که به دنبال آن هستیم "میانگین (μ)" است. می دانیم که برای ساخت بازه اطمینان در اولین قدم باید شرایط قضیه حد مرکزی (CLT) را بررسی کنیم.

در ابتدا باید بدانیم که در این سوال جامعه هدف ما، مجموعه داده HealthCare است و برای انجام این سوال اولین قدم گرفتن یک نمونه است، که من نمونه ای با سبای ۵۰ به طور تصادفی انتخاب می کنم.

کد این قسمت و نمونه گرفته شده توسط این کد در R، در شکل زیر قابل مشاهده است :

```
# Choose a random sample of 100 data points from the HealthCare dataset
My_sam <- HealthCare[sample(nrow(HealthCare), size = 50, replace = FALSE),]

# My_sam
  id gender age hypertension heart_disease ever_married work_type Residence_type avg_glucose_level bmi smoking_status stroke health_bills age_interval
942 58999 Male 60.00 0 0 Yes Govt_job Urban 100.54 30.1 never smoked 0 2582.010 44-64
163 69768 Female 1.32 0 0 No children Urban 70.37 NA Unknown 1 NA 0-21
3911 36548 Male 31.00 0 0 Yes Govt_job Urban 65.70 30.4 formerly smoked 0 2005.820 22-43
1362 33167 Female 59.00 0 0 Yes Private Urban 89.96 28.1 Unknown 0 2305.429 44-64
1000 2454 Male 4.00 0 0 No children Rural 89.11 20.1 Unknown 0 2497.102 0-21
936 25287 Male 54.00 0 0 Yes Private Urban 92.95 41.0 never smoked 0 2935.147 44-64
654 72823 Female 79.00 0 0 Yes Private Urban 70.35 23.0 formerly smoked 0 2723.354 65-82
1972 48775 Female 78.00 1 0 Yes Self-employed Rural 201.07 21.8 Unknown 0 2673.110 65-82
1663 51883 Female 52.00 0 0 Yes Govt_job Rural 69.11 35.2 never smoked 0 3357.326 44-64
```

شرط استقلال (Independence) : تمامی observation های درون این sample مستقل از هم هستند و به صورت تصادفی انتخاب شده است. همچنین $10\% < 50$ از جامعه هدف است. بنابراین شرط استقلال برقرار است. شرط sample size/skew : از آنجایی که $30 < 50$ است و با فرض نرمال بودن توزیع می توان گفت این شرط هم برقرار است و میتوانیم از قضیه حد مرکزی استفاده کنیم.

به دلیل اینکه می خواهیم بازه اطمینان برای "میانگین (μ)" بسازیم، فرم کلی این بازه اطمینان بدین صورت خواهد بود :

$$Point\ estimate \pm ME \rightarrow Point\ estimate \pm Z*SE$$

که داریم :

$$Point\ estimate = \bar{X} \text{ و } SE = \frac{\sigma}{\sqrt{n}}$$

از آنجایی که می خواهیم بازه اطمینان ۹۵٪ ایجاد کنیم برای محاسبه Z^* باید یا به سراغ جدول نرمال استاندارد رفته یا از R و دستور qnorm کمک بگیریم. در ابتدا میانگین متغیر age را بدست آورده (این می شود point estimate). سپس از آنجایی که σ را داریم باید آن را حساب کنیم تا بتوانیم standard error را محاسبه کرده و سپس با استفاده از فرم کلی بازه اطمینانی که معرفی کردیم، اقدام به ساخت این بازه بکنیم. (کد این قسمت به طور کامل در R نوشته شده است و کامنت گذاری شده است). در شکل زیر کد ساخت بازه اطمینان ۹۵٪ به همراه نتیجه نهایی از اجرای این کد قابل مشاهده است :

```
# 95% Confidence Interval for mean
# Point estimate ± Z* SE

# Point estimate (x)
point_estimate <- mean(My_sam$age)

# standard error (SE)
SE <- sd(HealthCare$age)/sqrt(length(My_sam$age))

# Z*
Z <- qnorm(0.975)

# 95% Confidence Interval for mean
CI <- c(point_estimate - (Z*SE), point_estimate + (Z*SE))
CI
```

کد :

```
> # 95% Confidence Interval for mean
> CI <- c(point_estimate - (Z*SE), point_estimate + (Z*SE))
> CI
[1] 37.99861 50.53419
```

نتیجه :

بازه اطمینان

بنابراین ما ۹۵٪ اطمینان داریم که سن افراد مورد مطالعه ما (جامعه هدف، که همان مجموعه داده HealthCare است) به طور متوسط بین ۳۷/۹ تا ۵۰/۵ سال است.

: B

همانطور که مشاهده شد در قسمت A ما یک نمونه از مجموعه داده HealthCare گرفتیم و یک بازه اطمینان ۹۵٪ برای میانگین سن ایجاد کردیم و تفسیر این بازه اطمینان بدین صورت است که: ما ۹۵٪ اطمینان داریم که سن افراد مورد مطالعه ما (جامعه هدف، که همان مجموعه داده HealthCare است) به طور متوسط بین ۳۷/۹ تا ۵۰/۵ سال است.

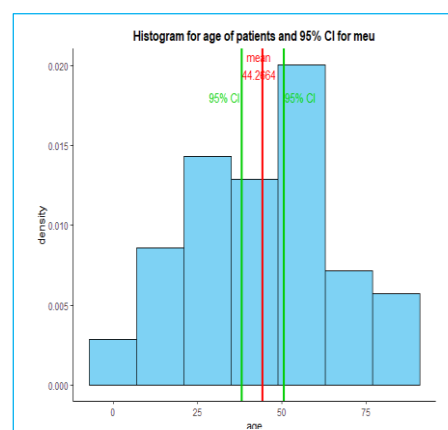
تفسیر Confidence level: زمانی که ما یک بازه اطمینان، با سطح اطمینان ۹۵٪ ایجاد می کنیم، این سطح اطمینان به این معنی است که اگر تعدادی sample با سایز ثابت (در اینجا با سایز ۵۰) انتخاب کنیم و با هر sample یک بازه اطمینان بسازیم، ۹۵٪ از این بازه اطمینان هایی که ساختیم شامل پارامتر جامعه هدف (در اینجا μ) می شوند و فقط ۵٪ از این بازه ها μ را شامل نمی شوند.

: C

در این قسمت هدف رسم کردن یک هیستوگرام برای نمایش بازه اطمینانی که در قسمت A بدست آوردیم و همچنین نمایش پارامتر میانگین برای متغیر انتخابی age است. برای این کار با کمک کتابخانه ggplot اقدام به رسم هیستوگرام برای سن افراد می کنیم و سپس با دستور geom_vline برای میانگین سن، و بازه اطمینانی که ایجاد کردیم خط عمودی رسم می کنیم. لازم به ذکر است که خط قرمز رنگ همانطور که برچسب گذاری شده است نشان دهنده میانگین و خطوط سبز رنگ نشان دهنده بازه اطمینان ۹۵٪ برای میانگین هستند. در شکل زیر کد هیستوگرام به همراه نتیجه نهایی از اجرای این کد قابل مشاهده است:

کد:

```
# importing libraries
library(ggplot2)
# HISTOGRAM and 95% Confidence Interval for mean age of observations
ggplot(My_sam, aes(x = age)) +
  # plot a histogram
  geom_histogram(aes(y=..density..),
                 binwidth = function(x) 2 * IQR(x) / (length(x)^(1/3)),
                 colour = "black", fill = "skyblue") +
  labs(title="Histogram for age of patients and 95% CI for meu") + # the title
  theme_bw() +
  # the grid and background removed from plot,
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),
        # put the title location in the center of the plot.
        plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  # mean
  geom_vline(aes(xintercept=mean(age)),color="red1",size=1.3)+
  annotate(geom="text", x=mean(My_sam$age)-1, y=0.02,
         label=paste0("mean\n",mean(My_sam$age)),
         color="red1")+
  # Confidence interval
  geom_vline(aes(xintercept=point_estimate + (Z*SE)),color="green3",size=1.3)+
  annotate(geom="text", x=point_estimate + (Z*SE)+5, y=0.018,
         label="95% CI",
         color="green3")+
  geom_vline(aes(xintercept=point_estimate - (Z*SE)),color="green3",size=1.3)+
  annotate(geom="text", x=point_estimate - (Z*SE)-5, y=0.018,
         label="95% CI",
         color="green3")
```



: D

در این مطالعه در سوال اول دیدیم که متوسط سن افراد حاضر در این مجموعه داده ۴۳ است و با توجه به بازه اطمینانی که در سوال ششم ایجاد کردیم دیدیم که با اطمینان ۹۵٪ متوسط سن افراد مورد مطالعه ما (جامعه هدف) بین ۳۷/۹ تا ۵۰/۵ سال قرار دارد. اما با نگاه اجمالی که در سوال صفر به این مجموعه داده انداختم متوجه شدم که این مجموعه داده وضعیت سلامتی حدود ۵۰۰ نفر را نشان می دهد و دیدیم که متغیرهای فشار خون، سابقه بیماری قلبی، سطح متوسط گلوکز، شاخص bmi فرد، وضعیت استعمال دخانیات و سابقه سکته مغزی از جمله متغیرهایی هستند که حاوی اطلاعات مهمی در مورد هر observation میباشند. به دلیل اینکه مقدار هر کدام از این متغیرها برای یک observation میتواند فرضیاتی را برای ما به همراه داشته باشد. در این قسمت فرضی که من در نظر گرفته ام و می خواهم برای آن آزمون فرض اجرا کنم این است که "میانگین سنی مخالف ۴۳ است" بنابراین مشخص است که با آزمون فرضی دو طرفه روبه رو هستیم.

برای انجام آزمون فرض ما یک framework داریم که به ترتیب برای این سوال آن را انجام میدهم، اما روش انجام آزمون فرض من مبتنی بر شبیه سازی نیست و من به روش تئوری براساس قضیه حد مرکزی آن را انجام می دهم :

۱- مشخص کردن فرض ها :

$$H_0 : \mu = 43$$

$$H_A : \mu \neq 43 \text{ (آزمون دوطرفه است)}$$

۲- جمع آوری دیتا ، محاسبه point estimate (برای میانگین) و انحراف معیار نمونه :

$$n = 50 \quad \bar{X} = \text{mean}(\text{age}) \quad SE = \frac{s}{\sqrt{n}}$$

۳- بررسی شرایط قضیه حد مرکزی (CLT) :

- شرط استقلال (Independence) : تمامی observation های درون این مجموعه داده مستقل از هم هستند و به صورت تصادفی انتخاب شده است. همچنین $50 > 10\%$ از جامعه هدف است. بنابراین شرط استقلال برقرار است.
- شرط sample size/skew : از انجایی که $50 > 30$ است و با فرض نرمال بودن توزیع می توان گفت این شرط هم برقرار است و میتوانیم از قضیه حد مرکزی استفاده کنیم.

۴- محاسبه آزمون Z :

$$Z = \frac{\text{point estimate} - \text{null value}}{SE}$$

۵- محاسبه p-value و تصمیم گیری با $\text{significance level} = 0.05$:

If $p\text{-value} < \alpha$, reject H_0 ; the data provide convincing evidence for H_A .

If $p\text{-value} > \alpha$, fail to reject H_0 ; the data do not provide convincing evidence for H_A .

کد این مراحل و همچنین p-value محاسبه شده برای این آزمون فرض در شکل زیر قابل مشاهده است :

```
# Hypothesis testing for mean of age variable
# 1 set the hypotheses
null_value <- 43
# 2 calculate the point estimate
n <- length(my_sam$age)
# Point estimate (X)
X <- mean(my_sam$age)
# standard error (SE)
SE <- sd(Healthcare$age)/sqrt(n)
# 3 check conditions
# 4 calculate test statistic
Z_statistic <- (X-null_value)/SE
# 5 calculate the p-value
p_value <- 2 * (1- pnorm(Z_statistic))
p_value
```

کد

```
> p_value
[1] 0.6920988
```

پاسخ :

از انجایی که $p\text{-value} > \alpha$ شده است. بنابراین نمی توانیم فرض صفر را در مقابل فرض H_A رد کنیم و از نظر آماری شواهد کافی و قانع کننده ای نداریم که نشان دهد میانگین سنی مخالف ۴۳ است.

مقدار p-value محاسبه شده نشان دهنده ی این است که ما با چه شدت و اطمینانی، نتوانستیم که فرض صفر را در مقابل فرض جایگزین رد کنیم. و بدین معنی است که اگر در حقیقت میانگین سنی افراد مورد مطالعه ما (جامعه هدف) ۴۳ سال باشد، ۶۹٪ شانس این وجود دارد که یک نمونه تصادفی از ۵۰ نفر از افراد مورد مطالعه ما (جامعه هدف) میانگین سنی بیشتر یا کمتر از ۴۳ سال را بدست بیاورند.

E :

می دانیم که یکی از راه های سریع تر انجام آزمون فرض استفاده کردن از بازه اطمینان است. اگر بازه اطمینانی که با سطح اطمینان مورد نظرم (مثلا ۹۵٪) میسازیم، شامل عددی که در فرض صفر است (null value) بشود، نمیتوانیم فرض صفر را رد کنیم. اگر بازه اطمینانی که با سطح اطمینان مورد نظرم میسازیم، شامل عددی که در فرض صفر است (null value) نشود، آنگاه فرض صفر را رد میکنیم.

با توجه به توضیحات داده شده و بازه اطمینان ۹۵٪ ساخته شده در قسمت A، مشاهده می شود که این بازه (۵/۵۰، ۳۷/۹) شامل null value که مقدارش برابر ۴۳ است می شود. بنابراین نی توانیم فرض صفر را در مقابل فرض H_A رد کنیم. همانطور که می بینیم این نتیجه با روشی که در قسمت قبل انجام شده است، همخوانی دارد.

F و G :

خطای نوع ۲ یکی از انواع خطاها در آزمون فرض است و زمانی اتفاق می افتد که تصمیم به رد نکردن فرض صفر بگیریم در صورتی که فرض جایگزین درست است و باید فرض صفر را رد می کردیم. ما احتمال رخ دادن خطای نوع ۲ را با β نمایش می دهیم. همچنین power یک آزمون، احتمال به درستی رد کردن فرض صفر را نشان می دهد و جز شاخصه های کیفیت آزمون فرض است.

(لازم به ذکر است، از آنجایی که می دانیم $Power = 1 - \beta$ است می توان با محاسبه β ، دیگری را بدست آورد. بنابراین قسمت F و G سوال ششم را با هم حل میکنم)

مقدار β بستگی به اختلاف بین actual mean و null value هم دارد، ما به این اختلاف effect size می گوئیم و می دانیم که هر چه این اختلاف بیشتر باشد انتظار داریم که β کوچکتر شود و به طبع آن، توان تست بیشتر می شود.

برای محاسبه توان و خطای نوع ۲ به یک پارامتر دیگر به نام actual man (میانگین واقعی جامعه هدف) نیاز داریم.

ما در آزمون فرض دوطرفه ای که در قسمت D انجام دادیم null value برابر با ۴۳ سال در نظر گرفته شده بود. اما با توجه به اینکه کل نمونه های مجموعه داده برابر با جامعه هدف در نظر گرفته شده اند بنابراین میانگین سنی افراد مورد مطالعه ما (actual mean) برابر با میانگین متغیر age در مجموعه داده HealthCare است. حال برای محاسبه توان تست و خطای نوع ۲ خواهیم داشت :

$$\alpha = 0.05 \quad \mu_0 = 43 \quad \mu_a = 43.22$$

احتمال رد نکردن H_0 زمانی که H_A صحیح است برابر است با : $Type II error: \beta = P(H_0 \text{ نکردن} | \mu = \mu_a)$

احتمال رد کردن H_0 زمانی که H_A صحیح است برابر است با : $Power = 1 - \beta = P(H_0 \text{ کردن} | \mu = \mu_a)$

بنابراین داریم :

$$\mu_a = 43.22$$

$$\bar{X} \sim N\left(\mu = 43.22, SE = \frac{\sigma}{\sqrt{n}} = 3.19\right)$$

$$\alpha = 0.05 \quad \left(\text{تست دوطرفه است} \frac{\alpha}{2}\right) \rightarrow P(Z > Z_{\alpha}) = 0.025 \rightarrow Z_{\alpha} = -1.96$$

ما فرض صفر را رد می کنیم اگر $Z < -1.96$ یا $Z > 1.96$ باشد (تست دو طرفه است).

$$P\left(\frac{\bar{X} - 43}{3.19} < -1.96\right) \rightarrow P(\bar{X} < 3.19 \times -1.96 + 43 = 36.74)$$

$$P\left(\frac{\bar{X} - 43}{3.19} > 1.96\right) \rightarrow P(\bar{X} > 3.19 \times 1.96 + 43 = 49.25)$$

$$P(\bar{X} < 36.74 | \mu = \mu_a = 43.22) \rightarrow P\left(Z < \frac{36.74 - 43.22}{3.19}\right) \rightarrow P(Z < -2.03)$$

$$P(\bar{X} > 49.25 | \mu = \mu_a = 43.22) \rightarrow P\left(Z > \frac{49.25 - 43.22}{3.19}\right) \rightarrow P(Z > 1.89)$$

$$Power = P(Z < -2.03) + P(Z > 1.89) = 0.0212 + 0.0294 = Poewr = 0.0506$$

$$\beta = 1 - Power = 1 - 0.0506 = 0.9494$$

بنابراین در این آزمون فرض :

احتمال رد نکردن H_0 زمانی که H_A صحیح است برابر است با 0.9494

احتمال رد کردن H_0 زمانی که H_A صحیح است برابر است با 0.0506

همانطور که گفته شد مقدار β بستگی به اختلاف بین actual mean و null value دارد، ما به این اختلاف effect size می‌گوییم. و می‌دانیم که هر چه این اختلاف بیشتر باشد انتظار داریم که β کوچکتر شود و به طبع آن، توان تست بیشتر می‌شود.

سوال شماره ۷

A :

در این قسمت به طور تصادفی یک نمونه با اندازه ۲۵ بدون جایگزاری، که شامل ۲۵ سطر از مجموعه داده HealthCare است گرفته شده. لازم به ذکر است که دو متغیر انتخابی برای این سوال، متغیر عددی age که سن افراد حاضر در این مجموعه داده را مشخص می‌کند، و متغیر عددی avg_glucose_level که سطح متوسط قند خون یا همان گلوکز هر فرد (Case) درون این مجموعه داده را مشخص می‌کند. و از آنجایی که نمونه گرفته شده شامل اطلاعات ۲۵ فرد است، برای هر فرد، مقادیر این دو متغیر انتخابی به هم وابسته هستند.

کد این قسمت و نمونه گرفته شده توسط این کد در R، در شکل زیر قابل مشاهده است :

```
# Choose a random sample of 25 data points from the dataset
My_sample <- HealthCare[sample(nrow(HealthCare), size = 25, replace = FALSE),]
```

کد :

```
> My_sample
  id gender age hypertension heart_disease ever_married work_type Residence_type avg_glucose_level bmi smoking_status stroke health_bills age_interval
640 63693 Male 37.00      0              0          No      Private      Urban      67.39 35.6      unknown      0      3395.599      22-43
631 39745 Female 60.00      0              0          Yes Self-employed Rural      58.65 30.1      never smoked      0      3241.124      44-64
4258 24854 Female 24.00      0              0          No Self-employed Urban      79.42 21.4      never smoked      0      2896.457      22-43
4892 18636 Female 26.00      0              0          Yes Govt_job Urban      72.36 35.4      never smoked      0      2770.034      22-43
4466 8623 Female 3.00      0              0          No children Urban      78.79 22.6      unknown      0      3692.651      0-21
1464 37907 Female 22.00      0              0          No Private Urban      135.64 19.5      never smoked      0      3268.001      22-43
1152 43268 Female 52.00      1              0          No Private Urban      73.00 25.2      smokes      0      3751.886      44-64
1367 35737 Male 1.08      0              0          No children Urban      86.09 19.5      unknown      0      3112.131      0-21
853 41537 Female 17.00      0              0          No Private Rural      62.49 26.9      never smoked      0      2981.514      0-21
3813 6960 Female 26.00      0              0          No Govt_job Urban      90.35 38.6      unknown      0      4286.939      22-43
996 60211 Male 1.40      0              0          No children Urban      90.51 18.9      unknown      0      1769.281      0-21
785 31956 Female 58.00      0              0          Yes Private Urban      76.99 29.0      never smoked      0      3745.807      44-64
3839 53759 Male 56.00      0              0          Yes Self-employed Urban      122.73 37.5      formerly smoked      0      3219.957      44-64
4259 22330 Female 45.00      0              0          Yes Self-employed Urban      82.94 29.3      unknown      0      2891.998      44-64
649 36814 Female 49.00      0              0          Yes Private Rural      56.11 28.7      smokes      0      2754.504      44-64
3542 18891 Male 24.00      0              0          No Govt_job Rural      99.65 50.3      never smoked      0      4270.342      22-43
4912 44813 Female 34.00      0              0          No Private Rural      69.06 29.0      smokes      0      2925.577      22-43
3743 29546 Male 71.00      0              0          Yes Govt_job Rural      99.76 33.4      formerly smoked      0      3337.343      65-82
3446 22607 Female 41.00      0              0          Yes Private Urban      103.79 28.6      never smoked      0      3490.290      22-43
3715 68657 Female 1.48      0              0          No children Urban      61.53 20.5      unknown      0      3291.467      0-21
641 34363 Female 27.00      0              0          Yes Private Rural      95.12 27.0      never smoked      0      2646.292      22-43
2684 16856 Female 69.00      0              0          Yes Private Rural      84.46 19.9      unknown      0      3279.947      65-82
2992 5780 Female 47.00      0              0          Yes Private Urban      74.63 45.3      never smoked      0      4509.476      44-64
3560 32103 Male 59.00      0              0          Yes Self-employed Urban      76.51 29.8      never smoked      0      3280.862      44-64
4085 32023 Male 4.00      0              0          No children Urban      79.16 20.2      unknown      0      3283.520      0-21
```

a :

از آنجایی که نمونه برداری به صورت تصادفی و بدون جایگزاری انجام شده است و همچنین وضعیت سلامتی افراد در این نمونه مستقل از هم دیگر است، و همینطور $10\% < 25$ جامعه هدف است بنابراین شرط استقلال برقرار است. اما در مورد شرط sample size/skew از آنجایی که اندازه نمونه گرفته شده کوچک و کمتر از ۳۰ است و همچنین به دلیل نداشتن σ ، باید به جای استفاده از z-test، از t-test و توزیع t استفاده کنیم (که به توزیع نرمال شبیه است).

دلیل استفاده از t-test آن است که نمونه گرفته شده به اندازه کافی بزرگ نیست و همچنین چون σ را نداریم و باید به جای آن از s استفاده کنیم، خود این s یک تخمین است که دارای خطاست بنابراین از توزیع t استفاده می‌کنیم که شبیه توزیع نرمال است اما نسبت به توزیع نرمال دیرتر به صفر میل می‌کند. لازم به ذکر است که همین دلیل بازه اطمینان هایی که با توزیع t می‌سازیم نسبت به توزیع نرمال بازه اطمینان های بزرگتری هستند، به دلیل اینکه این توزیع می‌خواهد در مواقعی که سائز نمونه کوچک است، پارامتر جامعه هدفمان که به دنبالش هستیم، داخل بازه قرار بگیرد.

b :

در این سوال یک sample شامل اطلاعات، و وضعیت سلامت ۲۵ فرد داریم و با توجه به دو متغیر انتخابی "سن افراد" و "متوسط سطح گلوکز بدن"، می‌خواهیم بدانیم که آیا اختلاف قابل توجهی بین میانگین این دو متغیر عددی وجود دارد یا خیر. در واقع می‌خواهیم با استفاده از آزمون فرض استنباط را برای اختلاف ۲ میانگین انجام بدهیم. می‌دانیم که آماره "اختلاف بین دو میانگین" آماره ای است که می‌توانیم برای آن آزمون فرض مبتنی بر قضیه حد مرکزی اجرا کنیم. بنابراین اولین کار بررسی شرایط قضیه حد مرکزی است که این شرایط در قسمت a بررسی شدند و متوجه شدیم که باید از توزیع t استفاده کنیم. اما نکته این

سوال اینجاست که بین دو متغیر انتخابی، استقلال بین گروهی نداریم و اصطلاحاً paired data داریم، بنابراین باید از Paired Hypothesis test برای استنباط برای اختلاف ۲ میانگین استفاده کنیم.

Framework انجام آزمون فرض برای زمانی که Paired data داریم:

۱- تبدیل ۲ متغیر عددی به یک متغیر:

در این مرحله باید تفاضل دو متغیر age و avg_glucose_level را برای هر فرد حساب کرده و در یک متغیر جدید به نام diff بریزیم. برای اینکار در ابتدا یک ستون به نام observation به sample اضافه کردم تا بتوانم با استفاده از تابع lapply روی تک تک سطرهای این sample تابعی که تفاضل این دو متغیر را حساب میکند، اجرا کنم و سپس حاصل تفاضل این دو متغیر برای هر observation را در متغیر diff میریزم.

۲- مشخص کردن فرض ها:

$$H_0 : \mu_{diff} = 0$$

$$H_A : \mu_{diff} \neq 0 \text{ (آزمون دوطرفه است)}$$

۳- محاسبه point estimate (برای میانگین diff) و انحراف معیار diff:

$$n_{diff} = 25 \quad \bar{X}_{diff} = \text{mean}(diff) \quad S_{diff} = \text{sd}(diff) \quad SE_{diff} = \frac{S_{diff}}{\sqrt{n_{diff}}}$$

۴- محاسبه آماره آزمون t:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

۵- درجه آزادی برابر است با: $df = n_{diff} - 1$

۶- محاسبه p-value و تصمیم گیری با $\text{significance level} = 0.05$:

If $p\text{-value} < \alpha$, reject H_0 ; the data provide convincing evidence for H_A .

If $p\text{-value} > \alpha$, fail to reject H_0 ; the data do not provide convincing evidence for H_A .

کد این مراحل و همچنین p-value محاسبه شده برای این آزمون فرض در شکل زیر قابل مشاهده است: (همچنین این آزمون و محاسبه p-value را با استفاده از تابع t.test() هم انجام دادم که در زیر قابل مشاهده است)

```
# Paired Hypothesis test Framework
# 1) Convert 2 numeric variables to one variable called diff

# Adding a column with consecutive numbers
My_sample$observation <- 1:25
# do calculate difference between the age and avg_glucose_level variables
funct <- function(inp){
  diff <- My_sample$avg_glucose_level[inp] - My_sample$age[inp]
  return(diff)
}

# Add a new column to sample and named it "diff"
# call function on each row of the sample with lapply.
# calculate difference between the age and avg_glucose_level variables
My_sample$diff <- lapply(My_sample$observation,funct)

# 2) set the hypothesis
# 3) calculate the point estimate
n_diff <- length(My_sample$observation)
x_diff <- mean(unlist(My_sample$diff))
s_diff <- sd(unlist(My_sample$diff))
SE_diff <- s_diff/sqrt(n_diff)

# 4) calculate test statistic
T_statistic <- (X_diff-0)/SE_diff
# 5) df
dfree <- 24
# 6) p-value
P_VALUE <- 2 * pt(T_statistic,df=dfree,lower.tail = FALSE)

# with t-test
t.test(unlist(My_sample$diff))
```

کد:

```
> P_VALUE
[1] 8.826102e-07
```

پاسخ:

```
> # with t-test
> t.test(unlist(My_sample$diff))

One Sample t-test

data:  unlist(My_sample$diff)
t = 6.5561, df = 24, p-value = 8.826e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 32.5135 62.3897
sample estimates:
mean of x
 47.4516
```

از آنجایی که $p\text{-value} < \alpha$ شده است. بنابراین فرض صفر را رد می‌کنیم و از نظر آماری شواهد کافی و قانع کننده ای داریم که نشان دهد اختلاف قابل توجهی بین میانگین سن افراد مورد مطالعه (جامعه هدف) و سطح گلوکز بدن آن ها وجود دارد.

همانطور که دیدیم، دو متغیر انتخابی برای این سوال، متغیر عددی age که سن افراد حاضر در این مجموعه داده را مشخص می کند، و متغیر عددی avg_glucose_level که سطح متوسط قند خون یا همان گلوکز هر فرد (Case) درون این مجموعه داده را مشخص می کند، انتخاب شدند. در ابتدا از جامعه هدف (از مجموعه داده) دو نمونه با ساین ۱۰۰ به صورت کاملاً مستقل گرفته شده است (یعنی افراد درون این دو نمونه مستقل از هم هستند و استقلال بین گروهی داریم). سپس از یک نمونه متغیر age و از نمونه دیگر متغیر avg_glucose_level را انتخاب می کنیم (بنابراین استقلال بین گروهی داریم) در این سوال می خواهیم بدانیم که آیا اختلاف قابل توجهی بین میانگین این دو متغیر عددی وجود دارد یا خیر. در واقع می خواهیم با استفاده از آزمون فرض استنباط را برای اختلاف ۲ میانگین انجام بدهیم.

می دانیم که آماره "اختلاف بین دو میانگین" آماره ای است که می توانیم برای آن آزمون فرض مبتنی بر قضیه حد مرکزی اجرا کنیم. بنابراین اولین کار بررسی شرایط قضیه حد مرکزی است که این شرایط در قسمت a بررسی شدند اما از آنجایی که ساین نمونه بزرگ تر از ۳۰ است می توانیم از هر دو z-test و t-test استفاده کنیم، به دلیل اینکه توزیع t در مواقعی که ساین نمونه بزرگ است بسیار به توزیع نرمال نزدیک می شود.. اما نکته این سوال اینجاست که ایندفعه بین دو متغیر انتخابی، استقلال بین گروهی داریم، بنابراین باید از Non-Paired Hypothesis test برای استنباط برای اختلاف ۲ میانگین استفاده کنیم.

Framework انجام آزمون فرض :

۱- مشخص کردن فرض ها :

$$H_0 : \mu_{glucose} - \mu_{age} = 0$$

$$H_A : \mu_{glucose} - \mu_{age} \neq 0 \quad (\text{آزمون دوطرفه است})$$

۲- محاسبه point estimate :

$$Point\ estimate = \bar{X}_{glucose} - \bar{X}_{age} \quad SE_{\bar{X}_{glucose} - \bar{X}_{age}} = \sqrt{\frac{s_{glucose}^2}{n_{glucose}} + \frac{s_{age}^2}{n_{age}}}$$

۳- محاسبه آماره آزمون t :

$$T = \frac{\bar{X}_{glucose} - \bar{X}_{age} - null\ value}{SE_{\bar{X}_{glucose} - \bar{X}_{age}}}$$

۴- درجه آزادی برابر است با : $df = \min(n_{glucose} - 1, n_{age} - 1)$

۵- محاسبه p-value و تصمیم گیری با $\alpha = 0.05$: significance level

If $p\text{-value} < \alpha$, reject H_0 ; the data provide convincing evidence for H_A .

If $p\text{-value} > \alpha$, fail to reject H_0 ; the data do not provide convincing evidence for H_A .

کد این مراحل و همچنین p-value محاسبه شده برای این آزمون فرض در شکل زیر قابل مشاهده است :

```
My_age_sample <- Healthcare[sample(nrow(Healthcare[1:2555,]),
size = 100, replace = FALSE),]
My_glucose_sample <- Healthcare[sample(nrow(Healthcare[2556:5110,]),
size = 100, replace = FALSE),]

My_age_sample_1 <- My_age_sample$age
My_glucose_sample_1 <- My_glucose_sample$avg_glucose_level

# 1) set the hypothesis
# 2) calculate the point estimate
n_glucose <- length(My_glucose_sample_1)
X_glucose <- mean(My_glucose_sample_1)
s_glucose <- sd(My_glucose_sample_1)

n_age <- length(My_age_sample_1)
X_age <- mean(My_age_sample_1)
s_age <- sd(My_age_sample_1)

p_estim <- X_glucose - X_age
SE_age_glucose <- sqrt( ( (s_glucose^2)/(n_glucose) ) + ( (s_age^2)/(n_age) ) )

# 3) calculate test statistic
T_statisticc <- (p_estim-0)/SE_age_glucose
# 5) df
dfreee <- 99
# 6) p-value
P_VALUEE <- 2 * pt(T_statisticc,df=dfreee,lower.tail = FALSE)
P_VALUEE
```

کد :

```
> P_VALUEE
[1] 3.78526e-23
```

پاسخ :

از آنجایی که $p\text{-value} < \alpha$ شده است. بنابراین فرض صفر را در مقابل فرض H_A رد می کنیم و از نظر آماری شواهد کافی و قانع کننده ای داریم که نشان دهد اختلاف قابل توجهی بین میانگین سن افراد مورد مطالعه (جامعه هدف) و سطح گلوکز بدن آن ها وجود دارد.

اگر بخواهیم برای اختلاف میانگین دو متغیری که داریم، یک بازه اطمینان ۹۵٪ بسازیم فرم کلی آن بدین صورت است :

$$\bar{X}_{glucose} - \bar{X}_{age} \pm t_{99}^* SE_{\bar{X}_{glucose} - \bar{X}_{age}}$$

کد این بازه و نتیجه اجرای آن در شکل زیر قابل مشاهده است :

```
# 95% confidence interval
t_ <- abs(qt(0.025,df=99))
CO_IN <- c(p_estim - (t_ * SE_age_glucose),p_estim + (t_ * SE_age_glucose))
CO_IN
```

کد :

```
> CO_IN <- c(p_estim - (t_ * SE_age_glucose),p_estim + (t_ * SE_age_glucose))
> CO_IN
[1] 48.94327 66.56453
```

پاسخ :

می دانیم که یکی از راه های سریع تر انجام آزمون فرض استفاده کردن از بازه اطمینان است. اگر بازه اطمینانی که با سطح اطمینان مورد نظرم (مثلا ۹۵٪) میسازیم، شامل عددی که در فرض صفر است (null value) بشود، نمیتوانیم فرض صفر را رد کنیم. اگر بازه اطمینانی که با سطح اطمینان مورد نظرم میسازیم، شامل عددی که در فرض صفر است (null value) نشود، آنگاه فرض صفر را رد میکنیم.

با توجه به توضیحات داده شده و بازه اطمینان ۹۵٪ ساخته شده در قسمت A، مشاهده می شود که این بازه (۴۸/۹۴, ۶۶/۵۶) شامل null value که مقدارش برابر ۰ است نمی شود. بنابراین فرض صفر را در مقابل فرض H_A رد کنیم.

همانطور که می بینیم این نتیجه ، همخوانی دارد.

سوال شماره ۸

در این سوال ، از متغیر عددی avg_glucose_level که سطح متوسط قند خون یا همان گلوکز هر فرد (Case) درون این مجموعه داده را مشخص می کند استفاده میکنم.

A :

در مفهوم بازه اطمینان به دنبال ارائه یک بازه هستیم که بتوانیم با اطمینان بالایی بگوییم که این بازه شامل پارامتر واقعی جامعه هدف می شود. در این سوال پارامتری که به دنبال آن هستیم "میانگین (μ)" است. می دانیم که برای ساخت بازه اطمینان در اولین قدم باید شرایط قضیه حد مرکزی (CLT) را بررسی کنیم اما از آنجایی که در این سوال میخواهیم از روش percentile برای ساخت بازه اطمینان استفاده کنیم بنابراین کاری به روش های مبتنی بر قضیه حد مرکزی نخواهیم داشت. در روش percentile method، به دلیل اینکه می خواهیم بازه اطمینان ۹۵٪ ایجاد کنیم عملاً ۲/۵ درصد بالایی و ۲/۵ درصد پایینی داده ها را دور میریزیم. یعنی در این روش به دنبال ۲.۵امین و ۹۷.۵امین percentile هستیم. برای اینکار از دستور quantile در R استفاده میکنم. در شکل زیر کد ساخت بازه اطمینان ۹۵٪ با روش percentile به همراه نتیجه حاصله قابل مشاهده است.

```
CI_95 <- c(quantile(HealthCare$avg_glucose_level,0.025),
           quantile(HealthCare$avg_glucose_level,0.975))
```

کد :

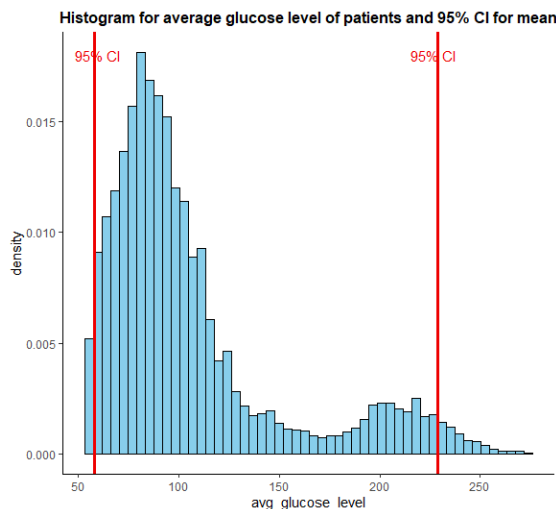
```
> CI_95
      2.5%      97.5%
57.99625 228.70000
```

پاسخ :

بنابراین طبق این بازه اطمینان، ما ۹۵٪ اطمینان داریم که سطح قند خون (گلوکز) افراد مورد مطالعه ما (جامعه هدف) به طور متوسط بین ۵۷/۹ تا ۲۲۸/۷ قرار دارد. همچنین در شکل زیر این بازه اطمینان ۹۵٪ را روی هیستوگرام متغیر avg_glucose_level هم نشان می دهیم.

```
# importing libraries
library(ggplot2)
library(plyr)
# HISTOGRAM and 95% Confidence Interval for mean avg_glucose_level of observations
ggplot(Healthcare, aes(x = avg_glucose_level)) +
# plot a histogram
  geom_histogram(aes(y=..density..),
                 binwidth = function(x) 2 * IQR(x) / (length(x)^(1/3)) ,
                 colour = "black", fill = "skyblue") +
  labs(title="Histogram for average glucose level of patients and 95% CI for mean") +
  theme_bw() +
# the grid and background removed from plot,
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),
        # put the title location in the center of the plot.
        plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
# 95% Confidence interval
  geom_vline(aes(xintercept=quantile(avg_glucose_level,0.025)),color="red2",size=1.3)+
  annotate(geom="text", x=quantile(Healthcare$avg_glucose_level,0.025)+2, y=0.018,
          label="95% CI",
          color="red2")+
  geom_vline(aes(xintercept=quantile(avg_glucose_level,0.975)),color="red2",size=1.3)+
  annotate(geom="text", x=quantile(Healthcare$avg_glucose_level,0.975)-2, y=0.018,
          label="95% CI",
          color="red2")
```

کد :



پاسخ :

: B

در این مطالعه، Original sample ما شامل اطلاعات ۵۱۱۰ نفر است. در ابتدا برای ساخت توزیع bootstrap باید از original sample تعدادی نمونه گیری (با جایگذاری) صورت بگیرد (bootstrap sample) به طوری که سائز این نمونه ها برابر با ۲۰ باشد (bootstrap population). سپس هر بار میانگین هر bootstrap sample را حساب می کنیم (bootstrap statistic) و این کار را ۱۰۰ بار تکرار میکنیم (در این سوال تعداد bootstrap sample ها را ۱۰۰ در نظر گرفتیم)، بنابراین می توانیم توزیع bootstrap را رسم کنیم. (این توزیع باید متقارن باشد) همانطور که در شکل زیر می بینیم، از نمونه original به تعداد ۱۰۰ بار نمونه گیری با جایگذاری صورت گرفته است و برای هر کدام، میانگین متغیر انتخابی من محاسبه و به عنوان یک نقطه در dotplot شکل زیر نمایش داده شده است. مشاهده می شود که توزیع bootstrap ما که به کمک R و کتابخانه ggplot رسم شده است توزیعی متقارن با میانگین و انحراف معیار این نقطه ها است.

```
#create an Empty Vector
bootstrap_dist <- c()

#for 1000 times resampling
for (i in c(1:100)) {
  #Sampling of the original sample with placement
  bootstrap_sample <- Healthcare[sample(nrow(Healthcare), size = 20, replace = TRUE),]

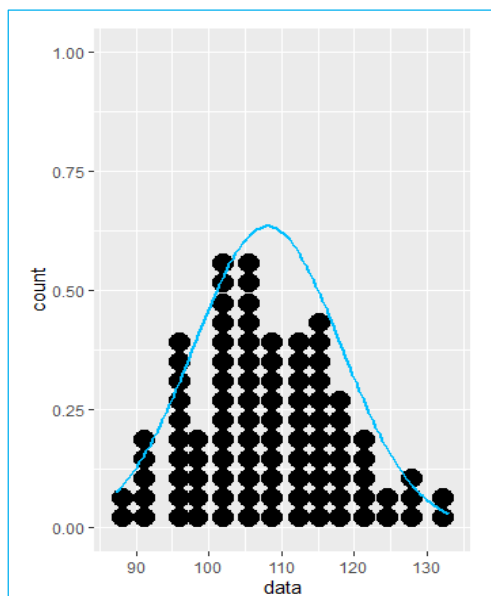
  #Using the following commands, the average of my variable is calculated.
  mean_boos <- mean(bootstrap_sample$avg_glucose_level)

  bootstrap_dist <- append(bootstrap_dist,mean_boos)
}

#To be able to visualize data with the ggplot Library, I create a data frame.
mydata_frame <- data.frame(data = bootstrap_dist)

library(ggplot2)
#I use dotplot to display the distribution shape.
ggplot(mydata_frame, aes(x = data)) + geom_dotplot(binwidth = 3) +
  stat_function(fun = function(curve) 16* (dnorm(curve,mean = mean(mydata_frame$data),
                                                sd = sd(mydata_frame$data))),
              color = "deepskyblue",size=1)
```

کد :



پاسخ :

برای ایجاد کردن بازه اطمینان با توزیع bootstrap دو روش داریم : Percentile Method و Standard error method
روش Standard error method برای bootstrap samples :

- در این روش برای ایجاد بازه اطمینان از فرمول $point\ estimate \pm t^* SE$ استفاده میکنیم.
- در اینجا از توزیع t با درجه آزادی $100 - 1 = 99$ برای بازه اطمینان ۹۵ درصد استفاده میکنیم
- Point estimate همان میانگین متغیر انتخابی من در نمونه original (در مجموعه داده) است
- از SE توزیع bootstrap استفاده می کنیم.

کد ساخت بازه اطمینان ۹۵٪ با روش Standard error method برای bootstrapping method در R طبق توضیحات
بالا به همراه نتیجه نهایی (بازه اطمینان ۹۵٪) در شکل زیر قابل مشاهده است :

```
# ***standard error method*** --> bootstrap samples confidence interval
# point estimate ± t* X SE
# point estimate --> mean of original sample
# df = 99
# SE --> bootstrap distribution

# calculate the mean
point_estimateeee <- mean(HealthCare$avg_glucose_level)

# df = n-1 = 100-1
df <- 99

# SE --> bootstrap distribution
SEE <- sd(mydata_frame$data)

# t-score with df=999
t_score <- abs(qt(0.025,df=99))

# confidence interval
# point estimate ± t* X SE
CI_boot <- c(point_estimateeee - (t_score * SEE),point_estimateeee + (t_score * SEE))
CI_boot

> CI_boot <- c(point_estimateeee - (t_score * SEE),point_estimateeee + (t_score * SEE))
> CI_boot
[1] 84.5599 127.7354
```

کد :

پاسخ :

بازه اطمینان bootstrap

: C

بله، بازه اطمینان هایی که با توزیع bootstrap ساخته می شوند محدودیت sample size و skewness را ندارند. همچنین توزیع sampling با نمونه گیری از جامعه هدف ایجاد می شود اما توزیع bootstrap از نمونه گیری از original sample بدست می آید. همانطور که در بازه اطمینان ساخته شده از توزیع bootstrap و بازه اطمینان ساخته شده از original sample

مشاهده می شود، بازه اطمینان original sample کمی محتاط تر عمل کرده و بازه اطمینان بزرگتری ایجاد کرده است. البته استفاده از روشی که برای ساخت این بازه اطمینان استفاده کردیم هم تا حدودی تاثیر گذار است. اما در کل توزیع bootstrap تحت تاثیر چولگی نخواهد بود و تا حد زیادی توزیعی متقارن است. همچنین انحراف معیار توزیع bootstrap کمتر از انحراف معیار original است و طبق نتایجی که حاصل شده است می بینیم بازه اطمینان ایجاد شده برای این دو متفاوت است.

سوال شماره ۹

در این مسئله می خواهیم بررسی کنیم که آیا اختلافی در بین میانگین هزینه های سالیانه سلامتی افرادی که در ۵ گروه شغلی "Private"، "Never-worked"، "Govt-jov"، "children" یا "Self-employed" قرار دارند وجود دارد یا خیر. از آنجایی که در این مطالعه با بیش از دو گروه رو به رو هستیم و می خواهیم میانگین این گروه ها را مقایسه کنیم باید از تست ANOVA و توزیع آماری F استفاده کنیم.

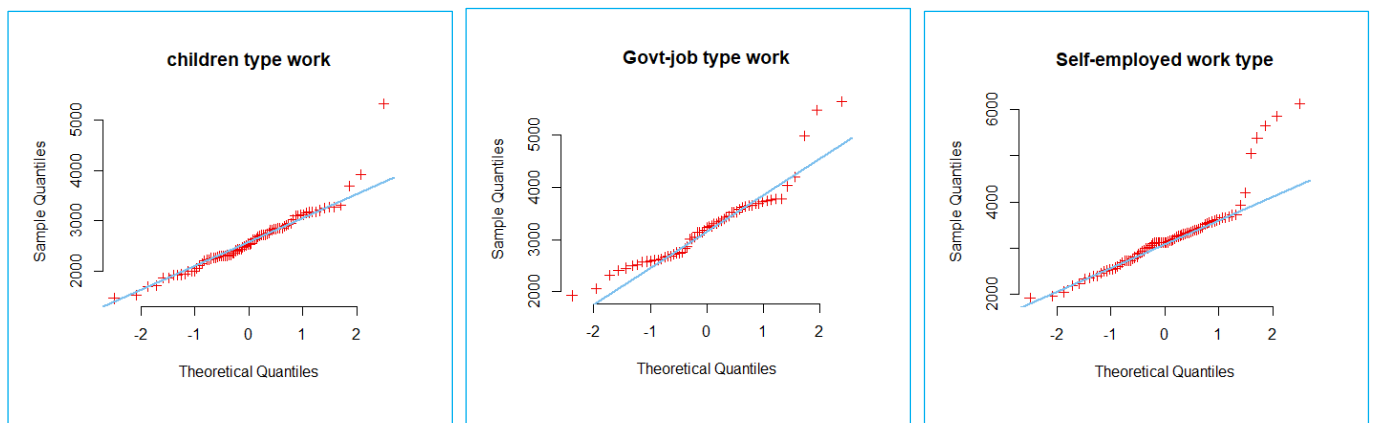
مجموعه داده ما شامل ۵۱۱۰ observation است و در این سوال آن را جامعه هدف در نظر میگیریم. بنابراین در ابتدا یک نمونه به صورت تصادفی از آن ایجاد کرده که اندازه آن ۵۰۰ است. فرکانس تکرار افراد در هر گروه (نوع شغلی) در درون این sample، در نمودار Barplot زیر که با استفاده از کتابخانه ggplot در R رسم شده است، قابل مشاهده است :

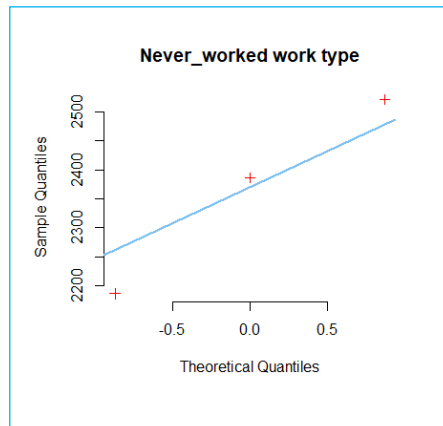
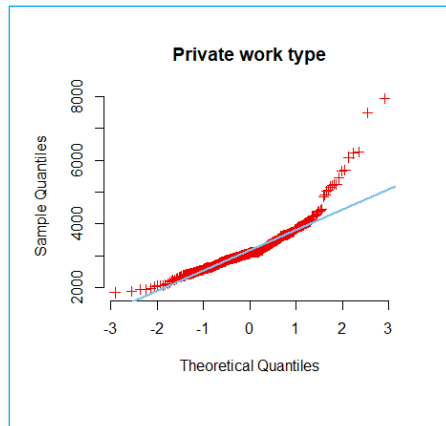


از آنجایی که نمونه گیری تصادفی بوده است و observation های درون هر گروه مستقل از هم اند بنابراین شرط استقلال درون گروهی برقرار است. همچنین از آنجایی که این پنج گروه (پنج نوع شغل) مستقل از هم هستند شرط استقلال بین گروهی هم برقرار است. بنابراین شرط independence آزمون ANOVA برقرار است.

همچنین طبق Q-Q plot های رسم شده برای هر پنج گروه مشاهده می شود که توزیع هر ۵ نوع شغلی تقریباً نرمال است. پس شرط Approximately Normal هم برقرار است. کد این قسمت در R به همراه نمودار های رسم شده در شکل زیر قابل مشاهده است :

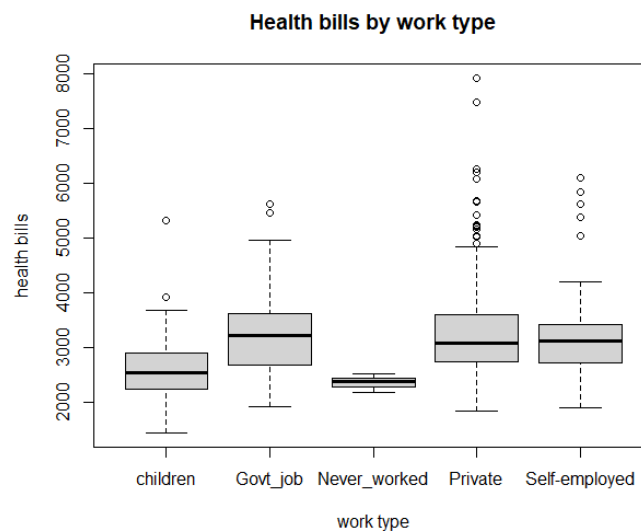
(لازم به ذکر است با یک حلقه for مقادیر health bill هر گروه را از نمونه جدا کرده و در پنج vector میریزم تا Q-Q plot هر گروه را رسم کنم. و همچنین طبق آن چیزی که در سوال صفر برای رفتار در مقابل مقادیر گمشده گفتم، برای مقادیر گمشده (N/A) در متغیر health bill از روش جایگزاری میانگین با داده های گمشده استفاده میکنم)





همچنین با توجه به side-by-side Boxplot رسم شده زیر فرض می کنیم که واریانس این پنج گروه برابر است بنابراین شرط Constant Variance (homoscedastic) هم برقرار است. کد این قسمت در R به همراه نمودارهای رسم شده در شکل زیر قابل مشاهده است :

```
# Boxplot for all work type
boxplot(unlist(mynewsample_1$health_bills)~unlist(mynewsample_1$work_type),
        main='Health bills by work type',xlab='work type',ylab='health bills')
```



بنابراین framework آزمون ANOVA بدین صورت خواهد بود :

۱- ابتدا فرض صفر و فرض جایگزین را مشخص می کنیم :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$H_A : \text{At least one pair of means are different from each other}$

$$\alpha = 0.05$$

$$F = \frac{MSG}{MSE} : \text{فرمول آماره آزمون F}$$

۳- محاسبه درجه آزادی :

$$\text{total} : df_T = n - 1 \rightarrow df_T = 500 - 1 = 499$$

$$\text{group} : df_G = k - 1 \rightarrow df_G = 5 - 1 = 4$$

$$\text{error} : df_E = df_T - df_G = n - k \rightarrow df_E = 500 - 5 = 495$$

۴- محاسبه آماره آزمون میانگین برای هر گروه :

$$n_{\text{Private}} = 277$$

$$\bar{X}_{\text{Private}} = 3247.919$$

$$n_{\text{Self-employed}} = 81$$

$$\bar{X}_{\text{Self-employed}} = 3206.088$$

$$\begin{aligned} n_{\text{Never_worked}} &= 3 & \bar{X}_{\text{Never_worked}} &= 2364.33 \\ n_{\text{Govt_job}} &= 59 & \bar{X}_{\text{Govt_job}} &= 3229.042 \\ n_{\text{children}} &= 80 & \bar{X}_{\text{children}} &= 2595.368 \end{aligned}$$

۵- محاسبه p-value و تصمیم گیری با $\alpha = 0.05$: significance level

If $p\text{-value} < \alpha$, reject H_0 ; the data provide convincing evidence for H_A .

If $p\text{-value} > \alpha$, fail to reject H_0 ; the data do not provide convincing evidence for H_A .

ما آزمون ANOVA یکطرفه را با استفاده از جدول ANOVA انجام می دهیم اما برای انجام تست ANOVA در R میتوانیم از aov که تست one way ANOVA را انجام میدهد استفاده میکنیم. (البته میتوان به صورت دستی هم محاسبات را انجام داد اما چون اینجا هدف تحلیل و استنباط است، در این قسمت از توابع R استفاده میکنیم و سپس به تحلیل، آنالیز و نتیجه گیری می پردازیم)

کد این قسمت و نتیجه تست one way ANOVA در R در شکل زیر قابل مشاهده است :

```
# ANOVA in R
anova_test<-aov(unlist(mynewsample_1$health_bills)~unlist(mynewsample_1$work_type))
summary(anova_test)
```

```
              Df    Sum Sq Mean Sq F value    Pr(>F)
unlist(mynewsample_1$work_type)  4  29524270  7381067   12.32 1.47e-09 ***
Residuals                    495 296518348   599027
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA table

		DF	Sum SQ	Mean SQ	F_value
Group	Class	4	29524270	7381067	12.32
Error	Residuals	495	296518348	599027	
	Total	499			

احتمال اینکه یک توزیع F با درجه آزادی های (۴ و ۴۹۵) مقدارش از ۱۲/۳۲ بیشتر بشود چه قدر است ؟

$$p\text{-value}: P(F_{(4,495)} > 12.32) = 1.472762e - 09$$

از آنجایی که $p\text{-value} < \alpha$ شده است. بنابراین فرض صفر را رد کنیم و از نظر آماری شواهد کافی داریم که نشان می دهد اختلاف قابل توجهی در بین میانگین هزینه های سالیانه سلامتی افرادی که در ۵ گروه شغلی "Govt-jov"، "children"، "Never-worked"، "Private" یا "Self-employed" قرار دارند وجود دارد. همچنین با توجه به side-by-side Boxplot رسم شده هم میتوان به این موضوع پی برد. (بار دیگر این نمودار که قبلا رسم شده است در زیر قابل مشاهده است)

