Statistical Inference

Project Phase 1.

—

Ali Izadi

810199102

سوال ۰:

(a

دیتاست Students از داده های دانش آموزان دو موسسه شامل نمرات آن ها در سه درس مختلف و اطلاعات دیگری از دانش آموزان از جمله سن، شغل پدر و مادر، ساعات مطالعه، غیبت و ... که بر نمرات آن ها تاثیر دارد، است.

با مطالعه این دیتاست میتوان به سوالاتی از جمله عملکرد دانش آموزان در در موسسه مختلف و همچنین بررسی معیارهای مختلف بر نمره دانش آموزان از جمله ساعات مطالعه آن ها، میزان سلامتی و غیبت آنها و همچنین شغل پدر و مادر پرداخت.

(b

۱۵ متغیر (به جز id=index)و ۳۹۵ نمونه در این دیتاست وجود دارد.

(c

خیر. missing value در دیتاست وجود ندارد.

```
any(is.na(StudentsPerformance))
```

(d

۱- بر روی نمرات سه درس افراد میزان ساعت مطالعه آن ها و غیبت های آن ها می تواند بیشترین تاثیر را داشته باشد.

۲- موسسه آموزشی می تواند متغیر بعدی بر روی نمرات افراد باشد.

۳- متغیرهایی مانند سلامتی و شغل پدر و مادر و استفاده از internet و romantic نیز در مرحله بعدی می توانند حاوی اطلاعات در میزان نمره ها باشند.

۴- متغیری مانند failure نیز از نمرات به دست می آیند که اطلاعات اضافی ای در آن وجود ندارد.

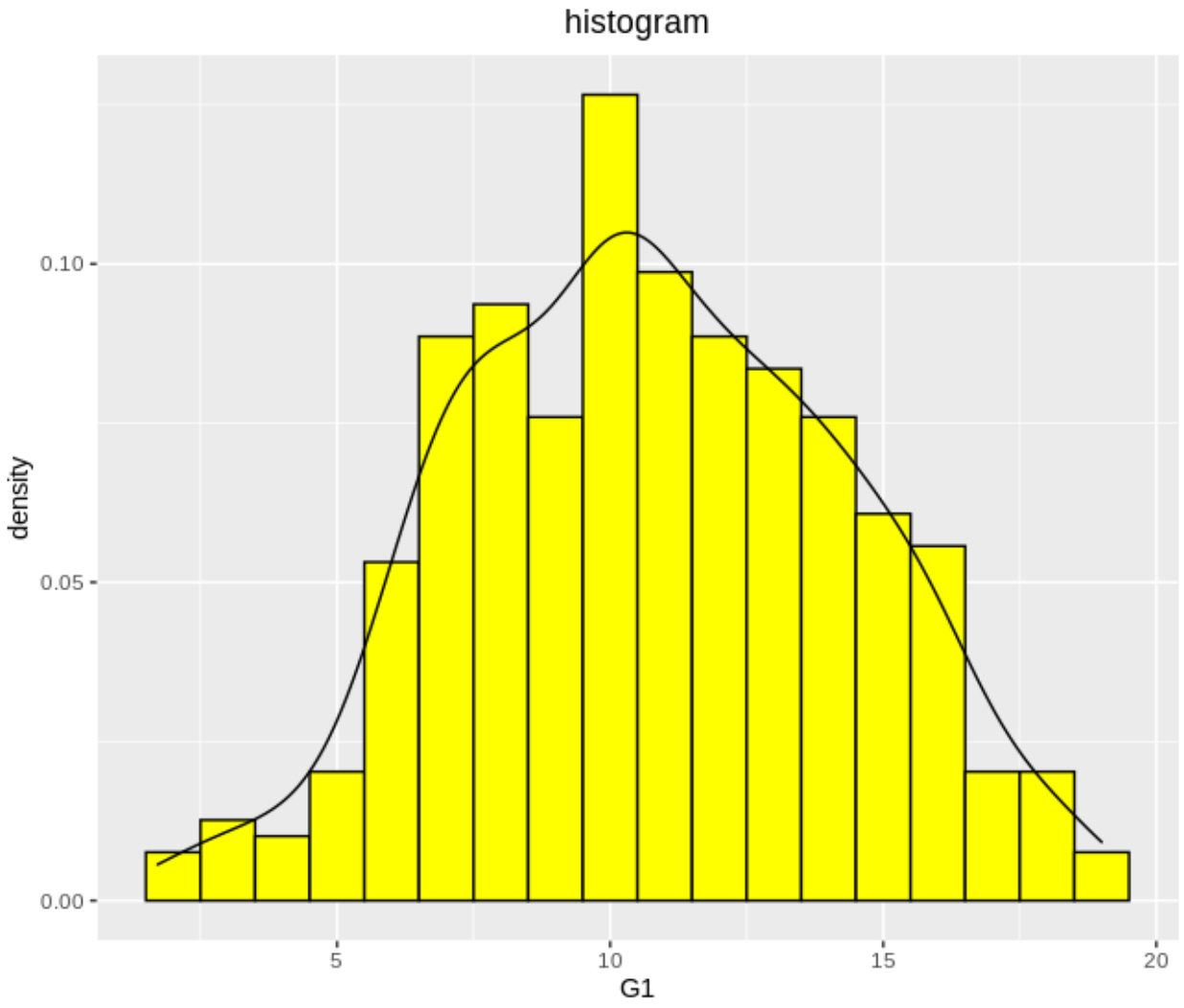
۵- متغیری مانند سن هم با توجه به این که مشخص کننده مقطع تحصیلی است ممکن است مشخص کننده نمره افراد باشد البته نه لزوما

سوال ۱)

a)

```
ggplot(StudentsPerformance, aes(x = G1)) +
  geom_histogram(aes(y =..density..), binwidth = 1, fill="yellow", color="black") +
  geom_density() +
  ggtitle("G1 histogram") +
  theme(plot.title = element_text(hjust = 0.5))
```

متغیر G1 انتخاب شده است و چون مشخص کننده نمره است پس می توان اندازه bin را یک نمره در نظر گرفت. همانطور که در شکل زیر مشاهده می شود modality خاصی در از متغیر وجود ندارد اما نمره ۸ تا ۹ بیشتر از نمره ۹ تا ۱۰ است و یک پیک در این جا به وجود آمده است که میتواند به خاطر این باشد که نمرات کمتر و نزدیک به ۱۰ برای قبول شدن دانش آموز به ۱۰ گرد شده اند:



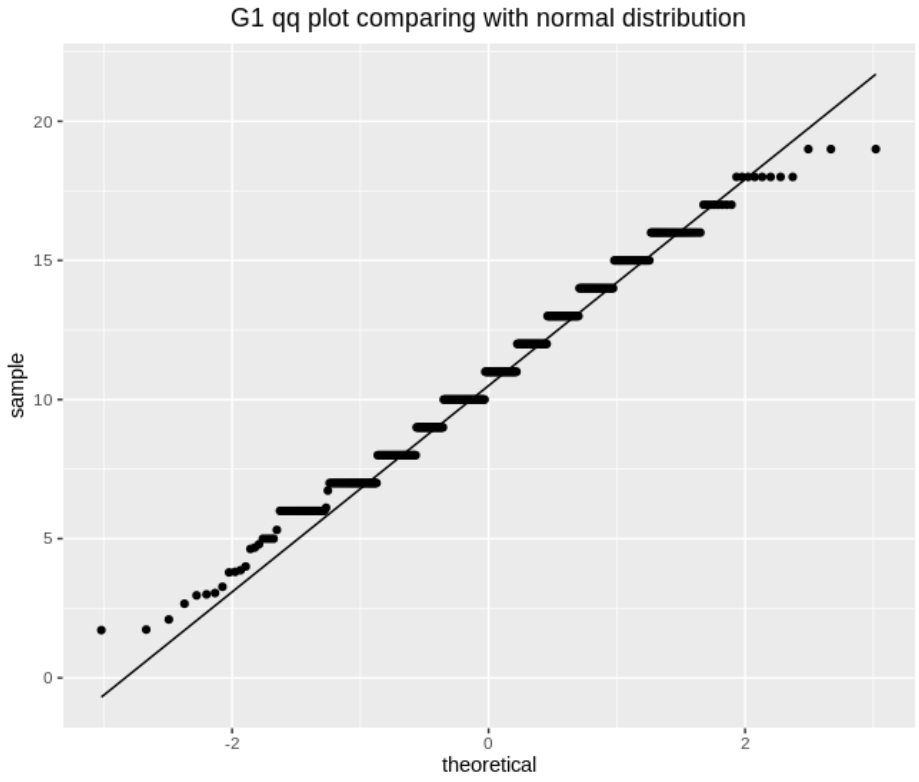
(b

با محاسبه quantile داده ها از range داده ها از ۱.۷ تا ۱۹ است.

```
> quantile(StudentsPerformance$G1)
       0%       25%       50%       75%      100% 
 1.713843  8.000000 11.000000 13.000000 19.000000 
```

با استفاده از qqplot توزیع را با توزیع نرمال مقایسه میکنیم که در دو طرف tail توزیع کمی تفاوت با توزیع نرمال وجود دارد.

```
ggplot(StudentsPerformance, aes(sample = G1)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("G1 qq plot comparing with normal distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```



(c

از رابطه زیر skewness را محاسبه میکنیم:

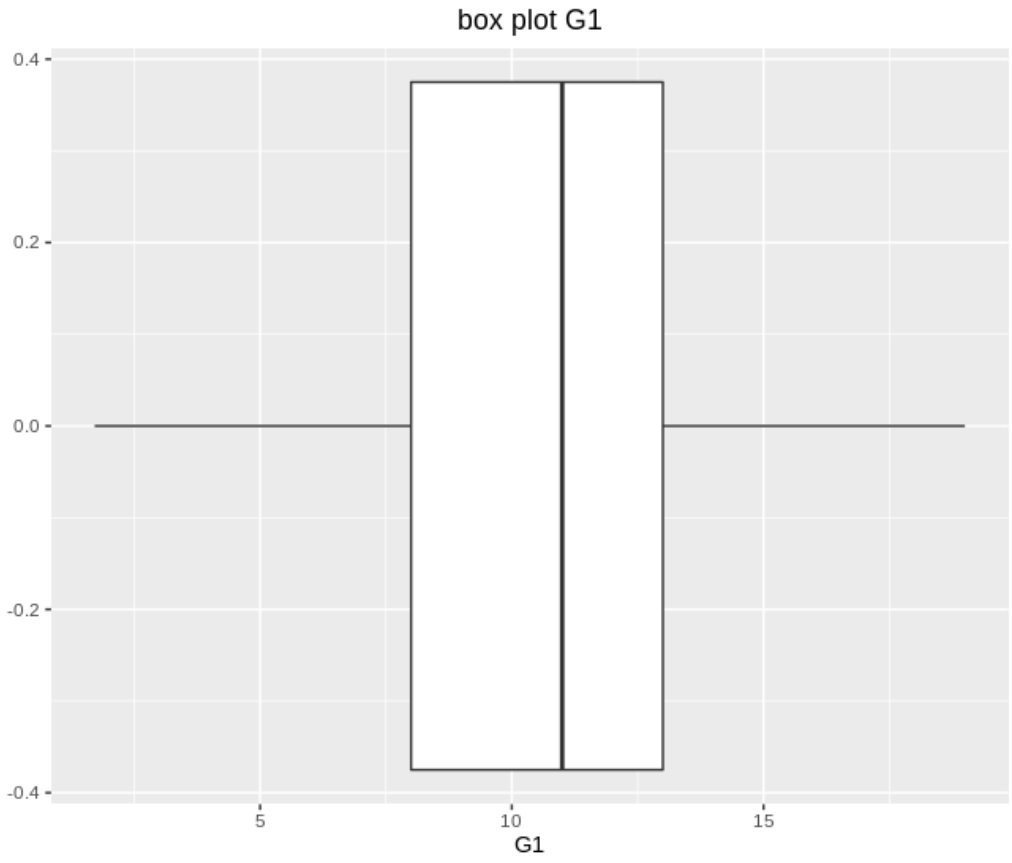
$$sk = \frac{mean\ - median}{standard\ deviation} = \frac{\mu - m}{\sigma}$$

```
> (mean(StudentsPerformance$G1) - median(StudentsPerformance$G1)) /
+     sd(StudentsPerformance$G1)
[1] -0.06167235
```

مقدار منفی نشان دهنده این است که توزیع left-skewed است.

(d

با نمایش box plot مشاهده میشود که upper whisker و lower whisker برابر با ماکسیمم و مینیمم داده ها هستند و با این روش outlierای مشاهده نمیشود.



(e

میانگین، میانه، واریانس و انحراف معیار مطابق زیر محاسبه شده است:

```
> mean(StudentsPerformance$G1)
[1] 10.78285
>
> median(StudentsPerformance$G1)
[1] 11
>
> var(StudentsPerformance$G1)
[1] 12.39784
>
> sd(StudentsPerformance$G1)
[1] 3.521057
>
```

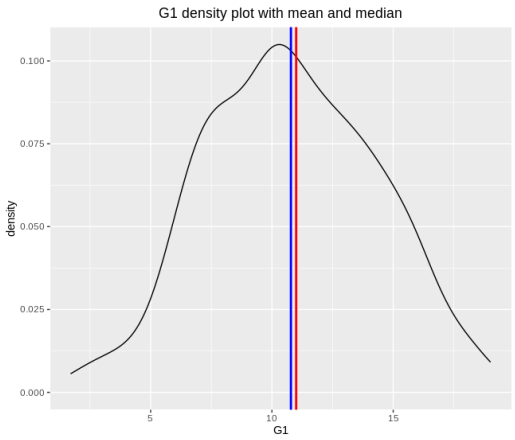
همان طور که از شکل توزیع نیز میتوان نتیجه گرفت میانگین نمرات نزدیک ۱۰ و میانه از میانگین بزرگتر و برابر ۱۱ است. انحراف معیار نیز نزدیک ۳.۵ و واریانس یعنی توان دوی انحراف معیار نیز برابر ۱۲.۳ است.

(f

میانگین با رنگ آبی از میانه با رنگ قرمز کوچکتر است.

```
ggplot(StudentsPerformance, aes(x = G1)) +
  geom_density() +
  geom_vline(aes(xintercept=mean(G1)), color="blue", size=1) +
  geom_vline(aes(xintercept=median(G1)), color="red", size=1) +
  ggtitle("G1 density plot with blue mean and red median") +
  theme(plot.title = element_text(hjust = 0.5))
```

همانطور که در شکل مشاهده می شود پیک توزیع نیز از میانه و میانگین کوچکتر است.



g) برای دسته بندی داده ها با استفاده از میانگین به صورت زیر به چهار دسته تقسیم میشوند که گروه دوم به عنوان مثال از میانگین منهای میانگین تقسیم بر ۲ بزرگتر و از میانگین کوچکتر است.

(mean - mean/2) > G1,

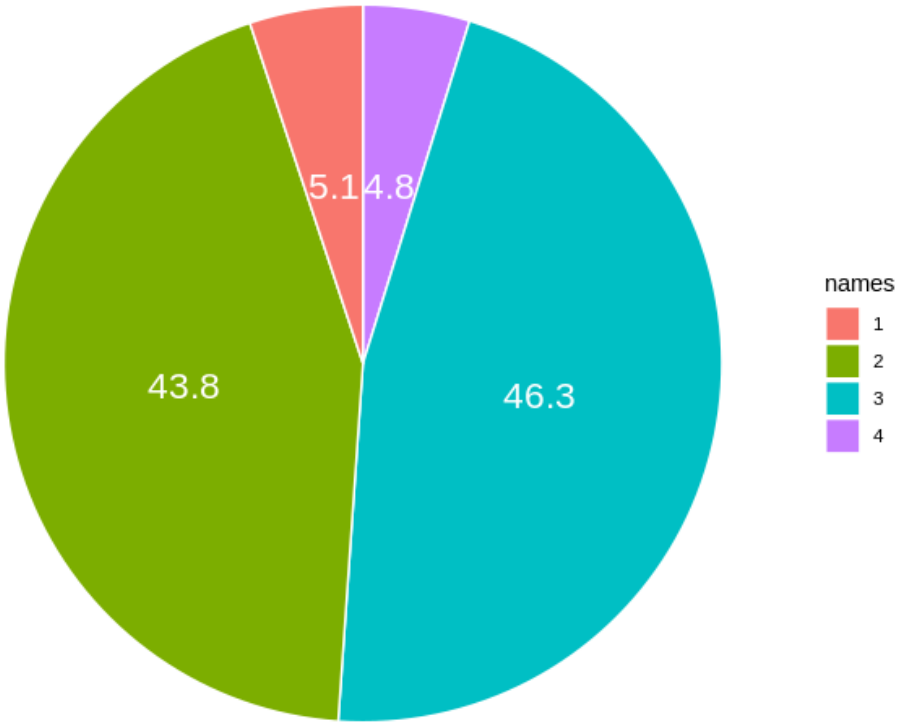
((mean - mean/2) <= G1) & (G1 < mean),

((mean + mean/2) > G1) & (G1 >= mean),

(mean + mean/2) <= G1)

محاسبات به صورت زیر انجام شده و pie chart در ادامه آورده شده است.

```
mean = mean(StudentsPerformance$G1)
G1 = StudentsPerformance$G1
frequencies = c(sum((mean - mean/2) > G1),
                sum(((mean - mean/2) <= G1) & (G1 < mean)),
                sum(((mean + mean/2) > G1) & (G1 >= mean)),
                sum((mean + mean/2) <= G1))


categories_names = c('1', '2', '3', '4')

data = data.frame(
  value=frequencies,
  names=categories_names
)

data <- data %>%
  arrange(desc(names)) %>%
  mutate(prop = round(value / sum(data$value) *100, digits=1)) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop )

ggplot(data, aes(x = "", y = prop, fill = names)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(y = ypos, label = prop), color = "white", size=6) +
  theme_void()
```

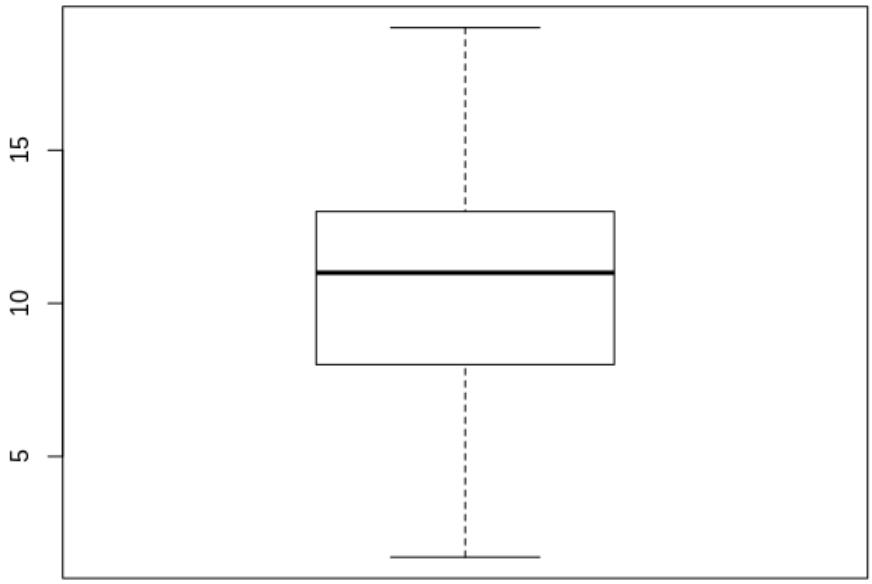


(h

با استفاده از تابع box plot مقادیر زیر به ترتیب شامل lower whisker و lower quartile و median و upper quartile و upper whisker آورده شده است. IQR نیز از تفاضل quartileها به دست می آید که برابر است با ۵

```
> bp = boxplot(G1)
> bp
$stats
          [,1]
[1,]  1.713843
[2,]  8.000000
[3,] 11.000000
[4,] 13.000000
[5,] 19.000000
```

نمودار box plot نیز در زیر آورده شده است:



سوال ۲)

(a

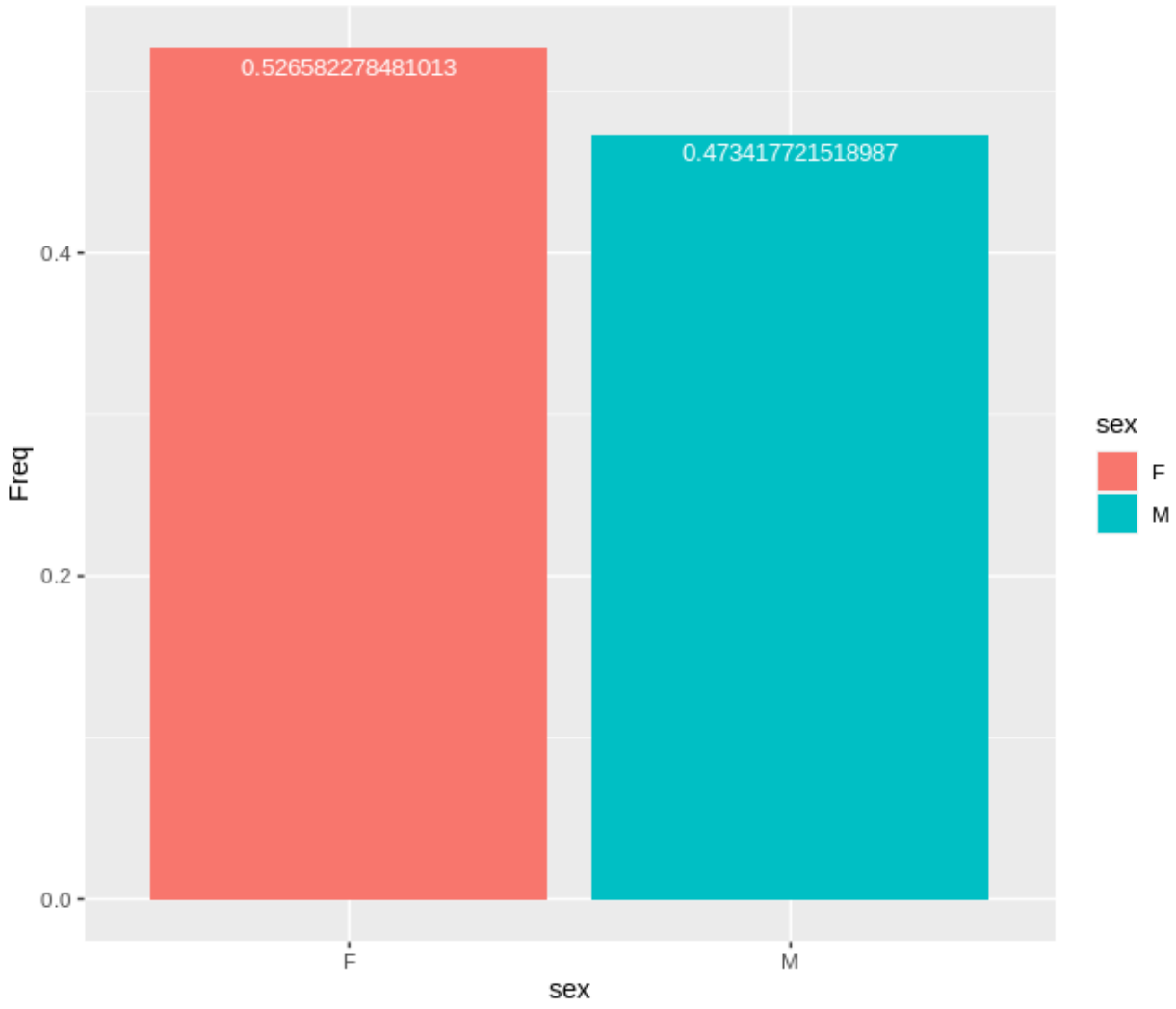
متغیر جنسیت به عنوان متغیر categorical انتخاب شده است. که frequency و percentage هر دسته در زیر محاسبه شده است:

```
> sex = StudentsPerformance$sex
>
> table(sex)
sex
  F   M
208 187
> table(sex) / sum(table(sex))
sex
        F         M
0.5265823 0.4734177
>
```

(b

با استفاده از کد زیر bar plot درصد هر گروه در زیر آورده شده است:

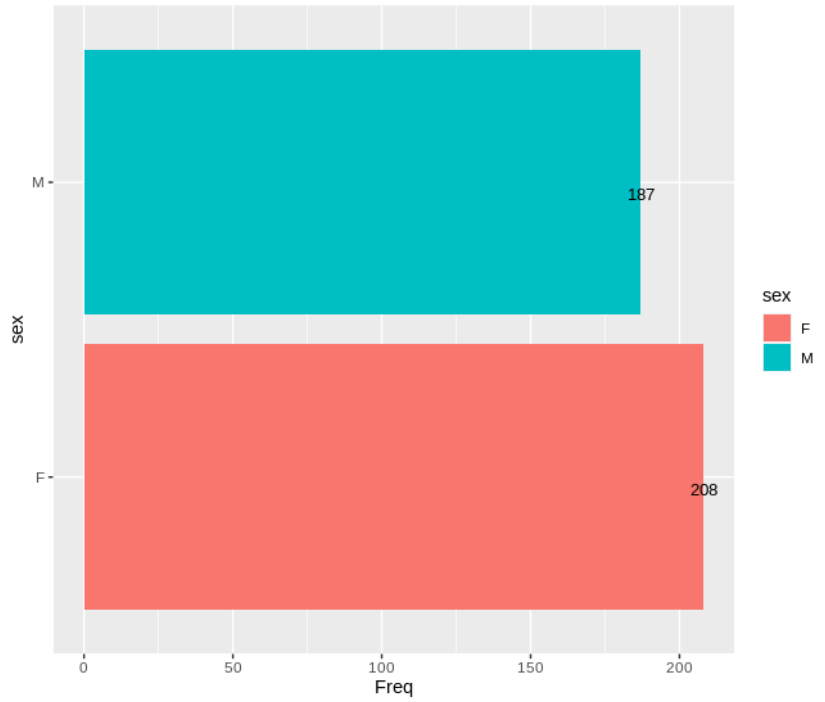
```
df  = as.data.frame(percentage)
ggplot(data=df, aes(x=sex, y=Freq, fill=sex)) +
  geom_bar(stat="identity")+
  geom_text(aes(label=Freq), vjust=1.6, color="white", size=3.5)
```



(c

مطابق با زیر ابتدا گروه ها sort شده و سپس horizontal bar plot آن ها آورده شده است:

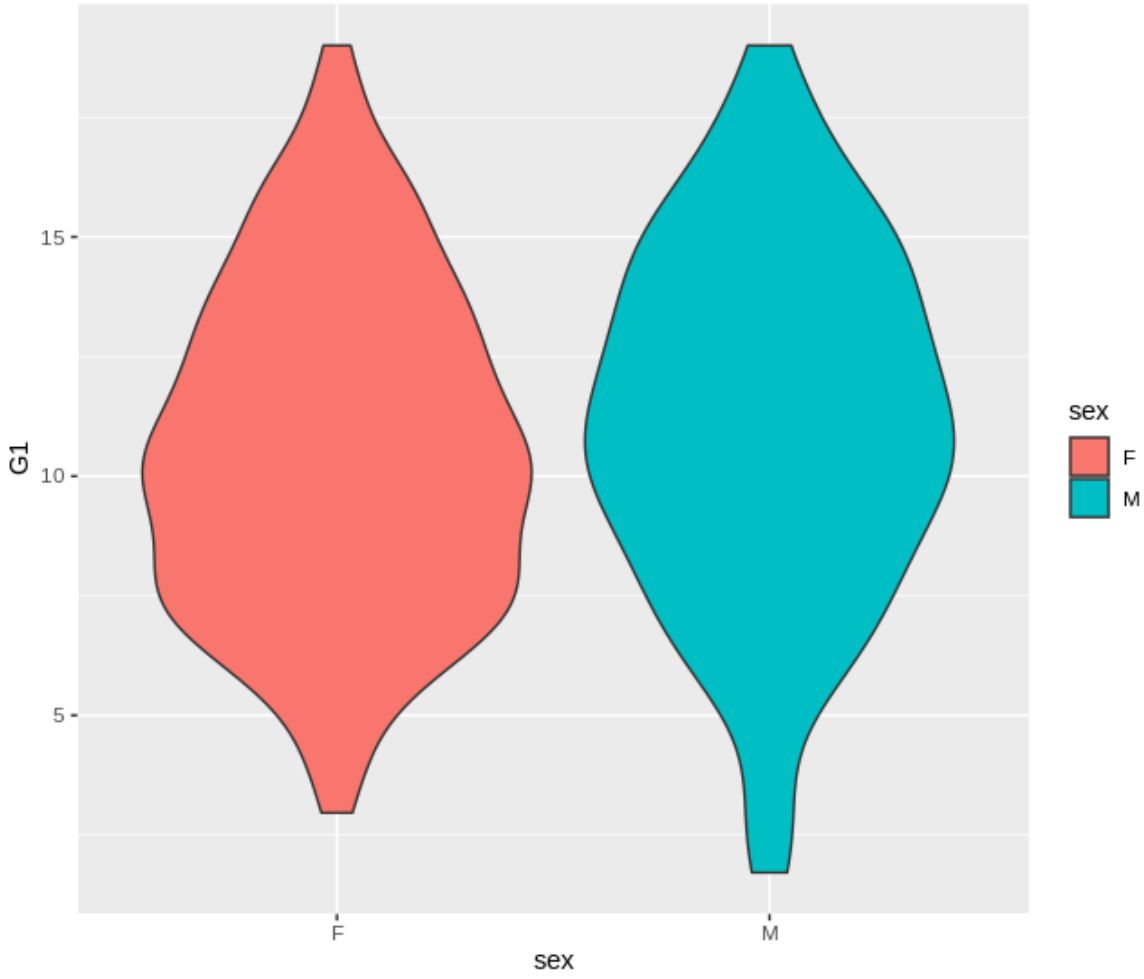
```
df = as.data.frame(freq)
df = df[order(df$Freq),]

ggplot(data=df, aes(x=sex, y=Freq, fill=sex)) +
  geom_bar(stat="identity")+
  geom_text(aes(label=Freq), vjust=1.6, color="black", size=3.5) +
  coord_flip()
```



(d

با انتخاب متغیر G1 به عنوان متغیر عددی violin plot برای متغیر sex در زیر آورده شده است:

```
ggplot(StudentsPerformance, aes(x=sex, y=G1, fill=sex)) +
  geom_violin()
```



سوال ۳)

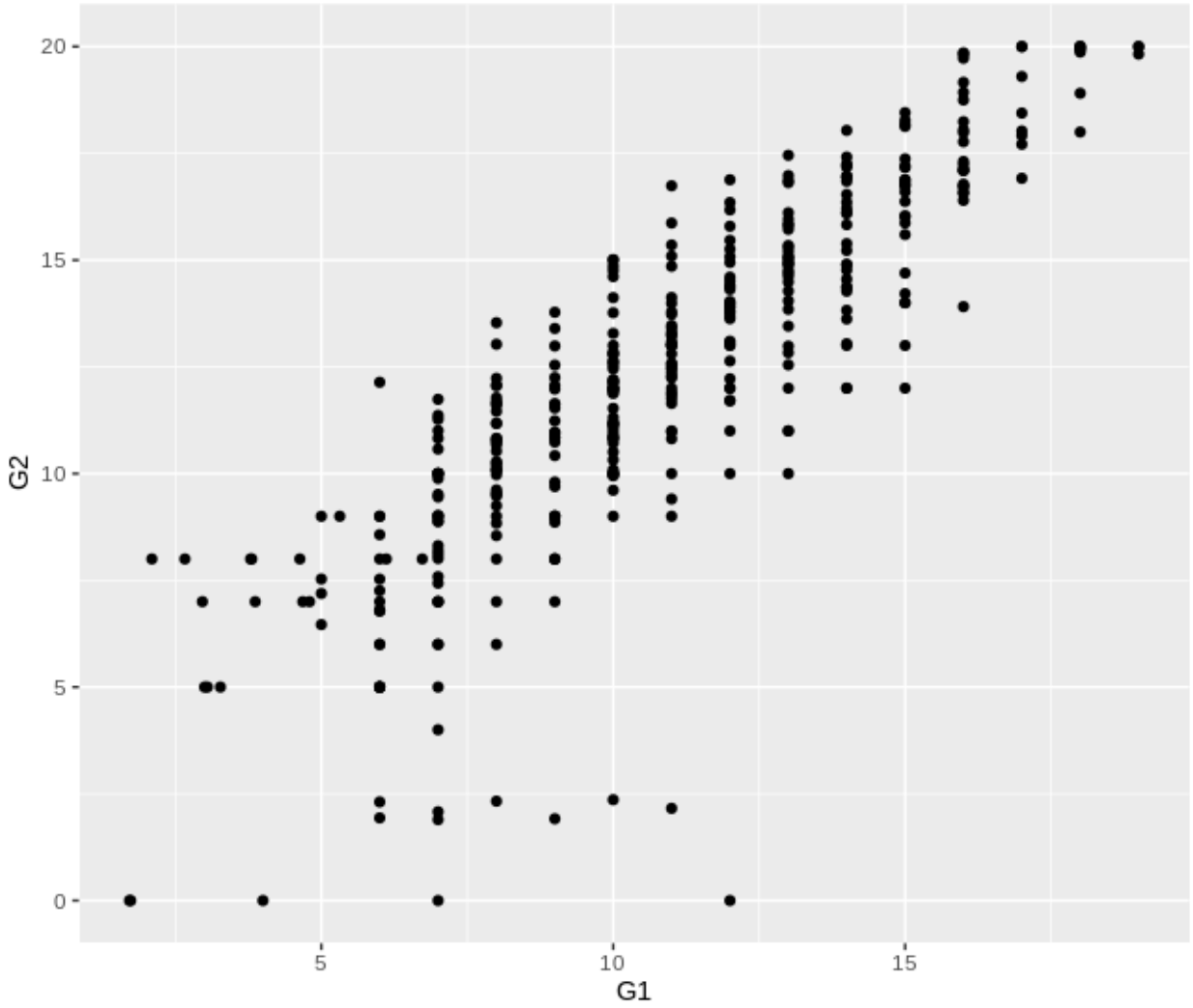
a)

دو متغیر G1 و G2 انتخاب میشوند که نمرات مربوط به دو دورس متفاوت دانش آموزان هستند. احتمالا این دو عدد با هم correlation مثبت داشته باشند چون هر دو نمرات مربوط به یک دانش آموز هستند.

(b

با استفاده از کد زیر scatter plot دو متغیر G1 و G2 رسم شده است که فرضیه قسمت اول را نیز تایید میکند.

```
ggplot(data = StudentsPerformance, aes(x = G1, y = G2)) +
  geom_point()
```



(c

مطابق با زیر correlation برابر است با 0.85

```
> cor(StudentsPerformance$G1, StudentsPerformance$G2)
[1] 0.8509365
```

(d

مقدار correlation به دست آمده مثبت و نزدیک به یک بوده بنابراین با افزایش متغیر G1 متغیر G2 نیز افزایش میابد. فرضیه اولیه در قسمت اول نیز با استفاده از این عدد به دست آمده تایید میشود و نمرات این دو درس برای دانش آموزان با هم correlation دارند.

(e

با استفاده از cor.test تست significance دو متغیر برای correlation انجام شده است. مقدار p-value تقریبا برابر با صفر است که در این حالت فرض صفر رد میشود و بنابراین correlation بین این دو متغیر صفر نبوده و ۹۵ درصد اطمینان داریم که در بازه بین 0.82 تا 0.87 قرار دارد.

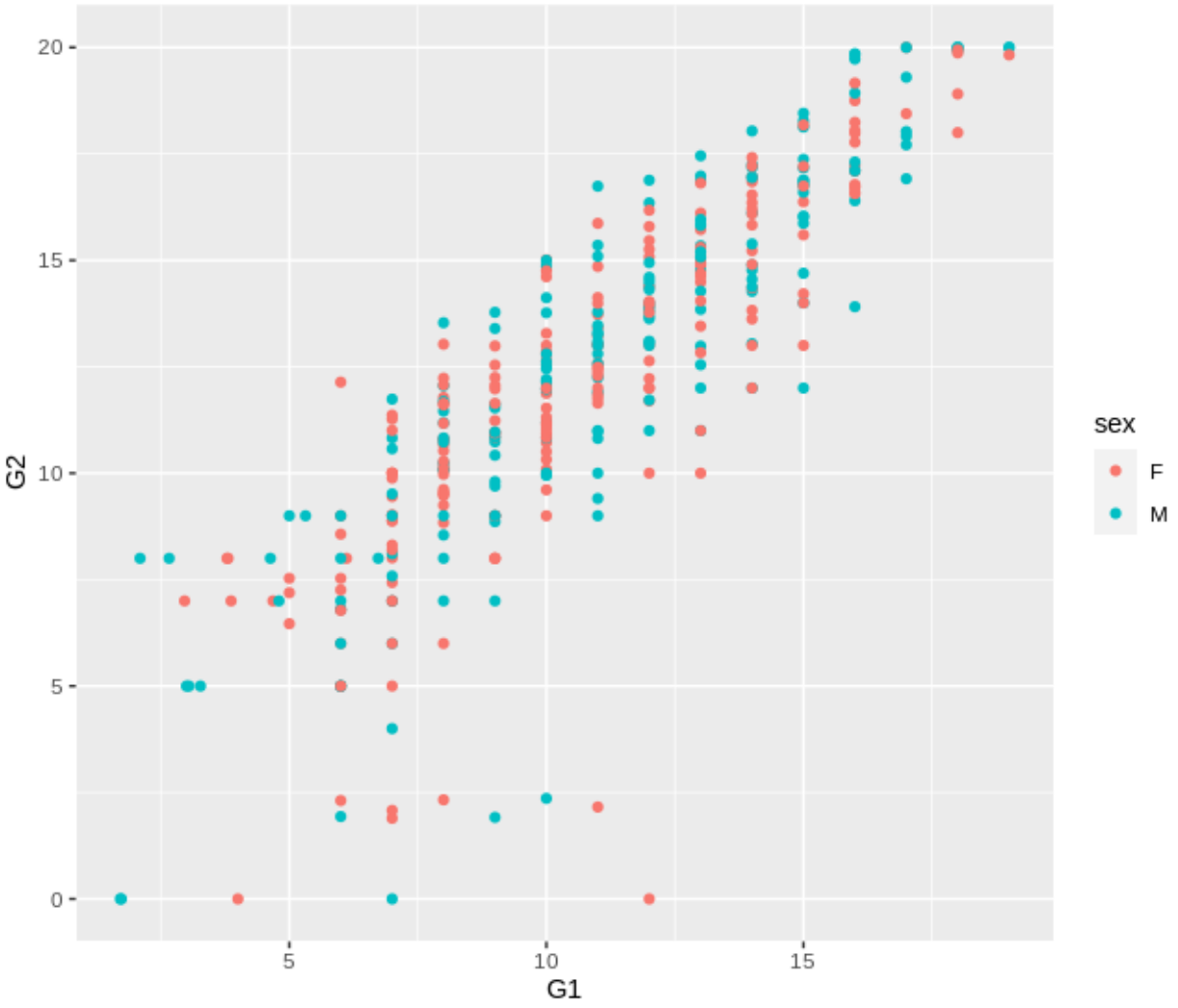
در این روش p-value با استفاده از توزیع t و با درجه آزادی n-2 محاسبه میشود.

```
> cor.test(StudentsPerformance$G1, StudentsPerformance$G2)

        Pearson's product-moment correlation

data:  StudentsPerformance$G1 and StudentsPerformance$G2
t = 32.115, df = 393, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8212172 0.8760518
sample estimates:
      cor
0.8509365
```

(f

با استفاده از کد زیر scatter plot دو متغیر G1 و G2 نسبت به کشیده شده اند و بر اساس دو گروه پسر دختر با رنگ ها جداگانه در شکل مشخص شده اند. همانطور که مشاهده می شود پسر یا دختر بودن تاثیر خاصی در نمرات این دو درس دانش آموزان وجود ندارد.

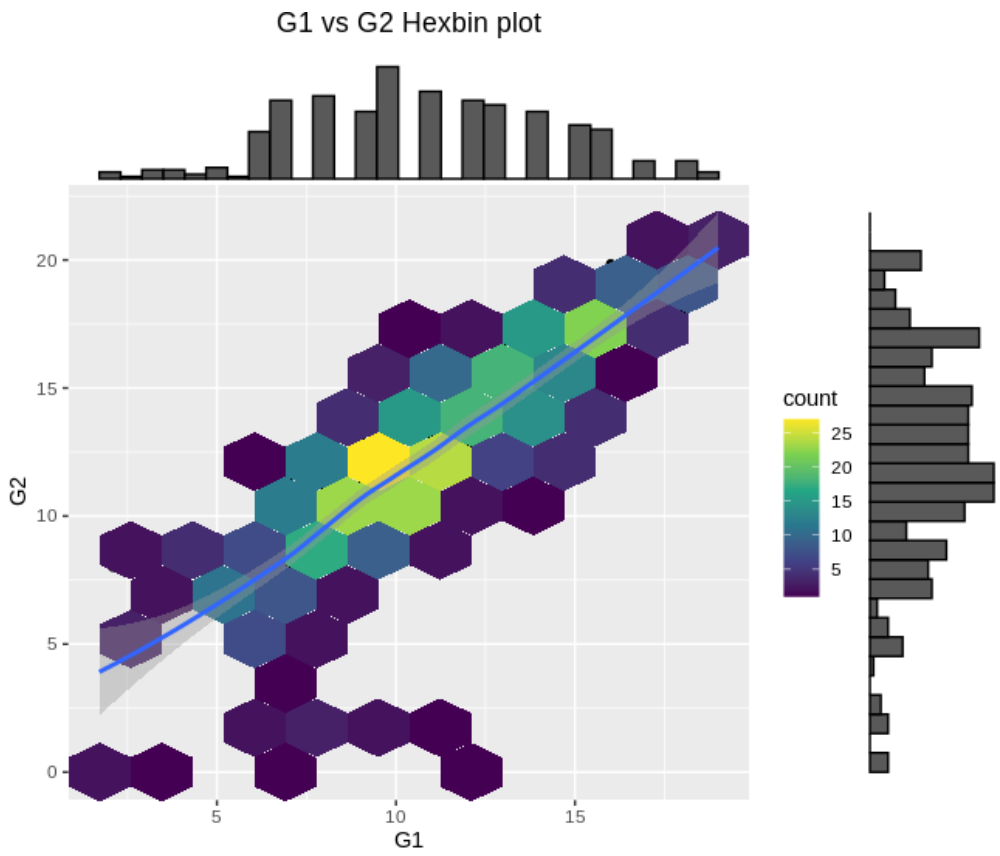
```
ggplot(data = StudentsPerformance, aes(x = G1, y = G2, color=sex)) +
  geom_point()
```



(g

با استفاده از کد زیر hexbin plot کشیده شده است. که این گراف زمانی کاربرد دارد که تعداد sample ها زیاد باشد و scatter plot نتواند به خوبی توزیع داده ها را به صورت locally مشخص کند. در این نمودار فضای مختصات به binهایی تقسیم می شود که تعداد sampleها در هر bin با رنگ مشخص میشود.

```
p = ggplot(StudentsPerformance, aes(G1, G2)) +
  geom_point() +
  geom_hex(bins = 10) +
  geom_smooth() +
  scale_fill_viridis_c() +
  ggtitle("G1 vs G2 Hexbin plot") +
  theme(plot.title = element_text(hjust = 0.5))

ggMarginal(p, type="histogram", size=5)
```

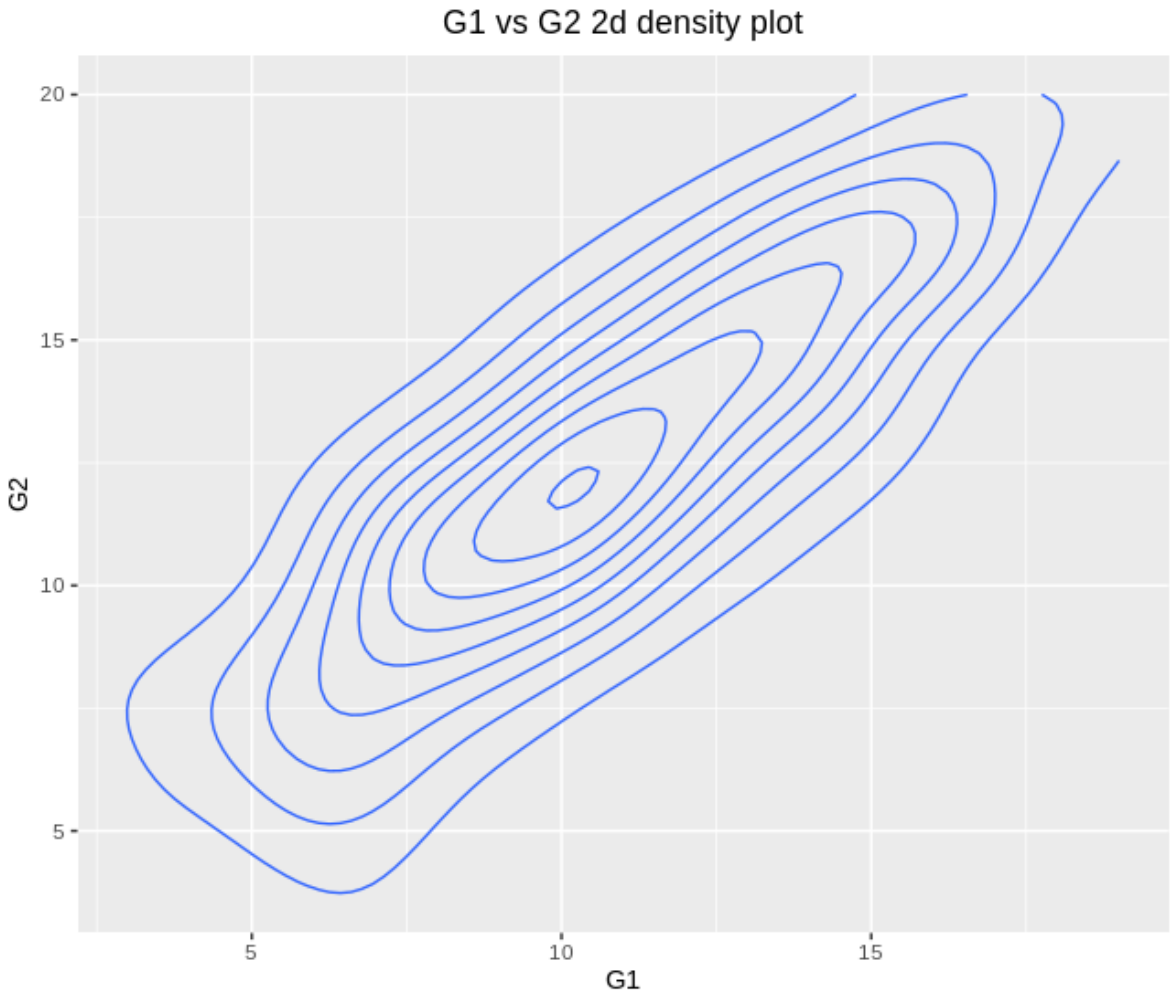


همانطور که مشاهده میشود تعداد binها ۱۰ انتخاب شده اند زیرا تعداد نمونه ها کم بوده است. هم چنین با توجه به رنگ های توزیع ها مشخص میشود که پیک نمودار نزدیک ۱۰ و هم چنین یک پیک نزدیک ۱۶ وجود دارد.

(h

```
ggplot(StudentsPerformance, aes(x=G1, y=G2) ) +
  geom_density_2d() +
  ggtitle("G1 vs G2 2d density plot") +
  theme(plot.title = element_text(hjust = 0.5))
```

با توجه به پیاده سازی بالا 2d density plot دو متغیر G1 و G2 در زیر آورده شده است.



در hexbin میتوان تعداد binها را کنترل کرد که در نتیجه شکل توزیع تغییر میکند اما در density plot خیر. در گراف بالا توزیع joint دو متغیر تنها نشان دهنده وجود یک پیک در نزدیک ۱۰ است در صورتی که در hexbin پیک دیگری نیز مشاهده شد. هم چنین توزیع locally نمونه ها در گراف بالا به خوبی مشخص نیست در صورتی که در hexbin plot با استفاده از رنگ در هر bin توزیع نمونه ها کاملا مشخص بود.

سوال ۴)

a)

```
ggpairs(StudentsPerformance[, c('age', 'goout', 'studytime', 'failures', 'health', 'absences', 'G1', 'G2', 'G3')],
        upper = list(continuous = "points", combo = "dot_no_facet"))
```

با استفاده از پیاده سازی بالا scatter plot دو به دو متغیرهای numerical در زیر آورده شده است.

همانطور که مشاهده می شود بین نمرات سه درس correlation شدید مثبت وجود دارد که منطقی است.

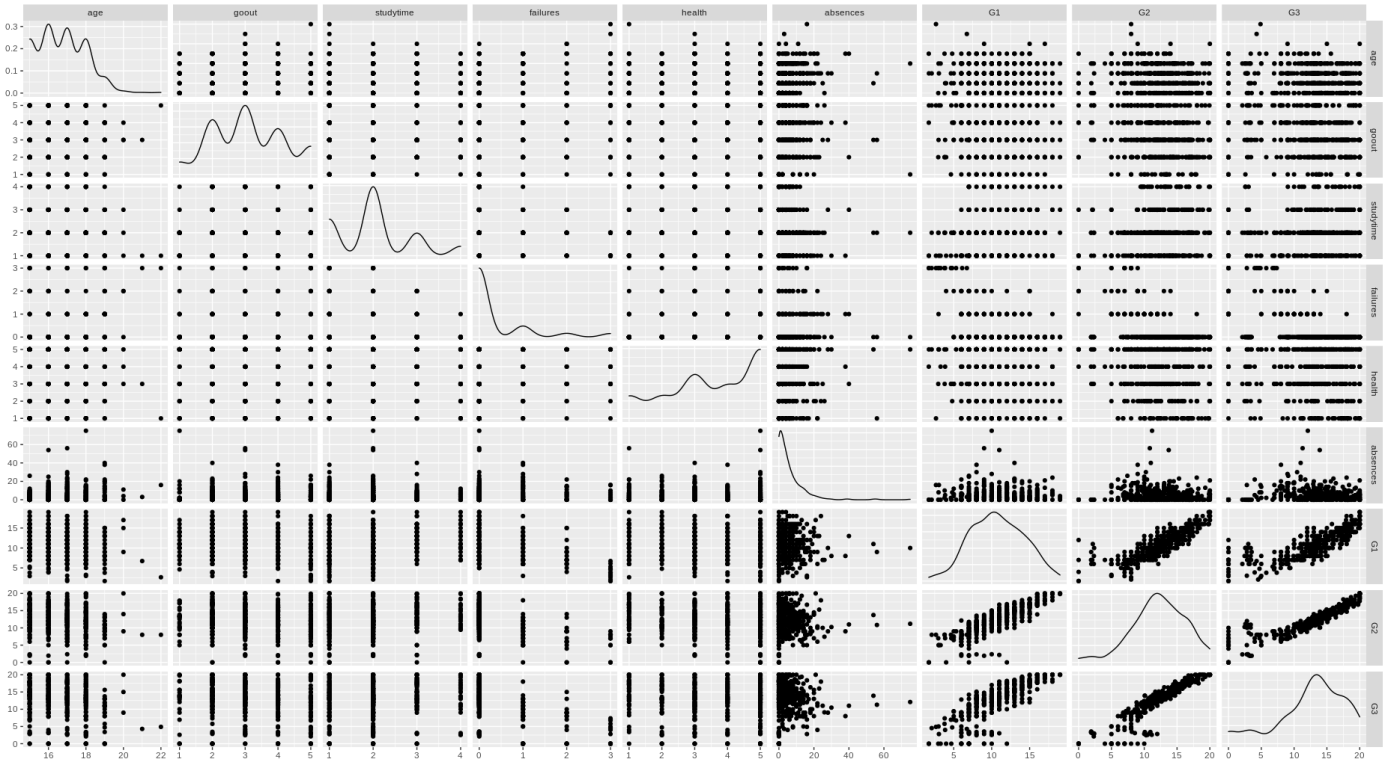
هم چنین توزیع غیبت نسبت به نمرات نیز از دو طرف دارای tail است.

طرف راست tail به خاطر این است که کسانی که نمرات بالایی گرفته اند تعداد غیبت کمتری داشته اند اما طرف چپ tail به خاطر این است که تعداد افراد کمتری نمره کم گرفته اند. اما آنهایی که غیبت زیاد داشته اند لزوما نمره آنها خیلی کم نشده است.

میزان سلامتی هم بر نمرات تاثیر گذار است و تعداد قابل توجهی از نمرات کم برای افرادی است که نمره سلامتی آن ها ۵ است یعنی سلامتی آن ها کم است. اما عکس آن لزما صادق نیست. یعنی سلامتی کم لزوما منجر به نمره کم نشده است.

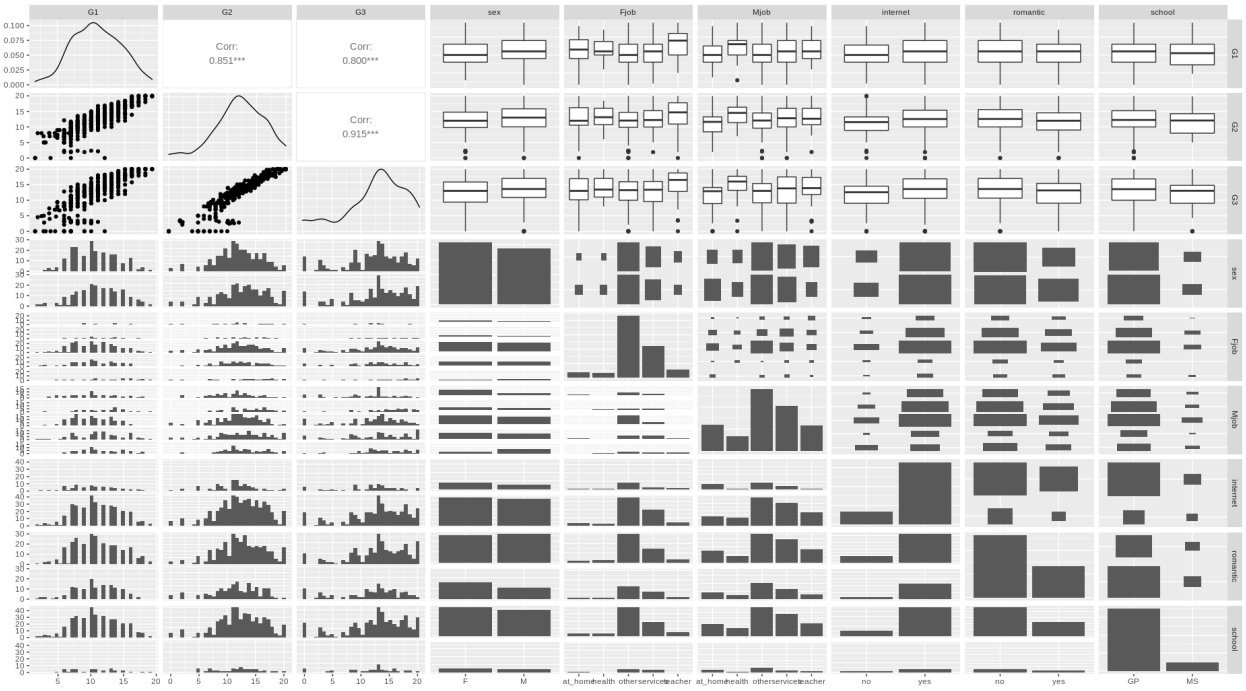
در تعداد غیبت ها هم نیز چندین outlier مشاهده میشود.

همچنین سن اکثر دانش آموزان زیر ۲۰ سال است گرچه تعدادی بالای ۲۰ سال نیز وجود دارند.


age
goout
studytime
failures
health
absences
G1
G2
G3

همچنین رابطه متغیرهای categorical با نمرات افراد در نمودار زیر مشاهده میشود که بین گروه های مختلف تفاوت قابل ملاحظه ای با مشاهده صرف وجود ندارد. گرچه ممکن است این تفاوت ها از نظر آماری significant باشند که باید تست های آماری مربوطه برای مقایسه ها انجام شود.

```
ggpairs(StudentsPerformance[, c('G1', 'G2', 'G3', 'sex', 'Fjob', 'Mjob', 'internet', 'romantic', 'school')])
```



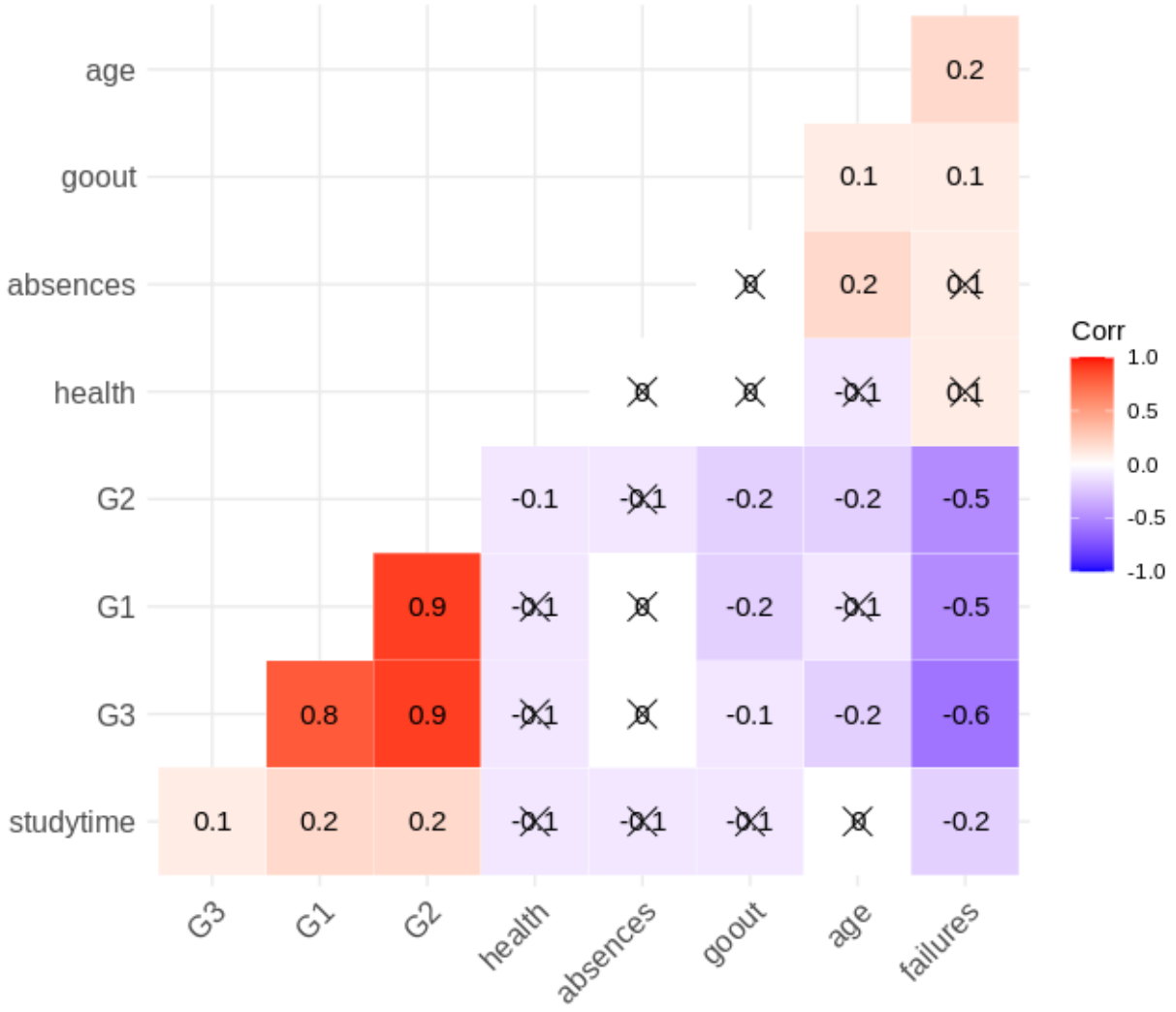
(b

با استفاده از کد زیر heatmap correlogram متغیرهای numerical کشیده است.

```
corr = round(cor(StudentsPerformance[, numerical]), 1)
p_values = cor_pmat(StudentsPerformance[, numerical])
ggcorrplot(corr, hc.order = TRUE, outline.col = "white", type = "lower", lab = TRUE, p.mat = p_values)
```

عدد داخل هر مربع correlation و ضربدرها مربوط به p-valueها هستند.

هم چنین متغیرهایی که با هم correlation دارند با استفاده از کلاسترینگ در کنار هم قرار گرفته اند.

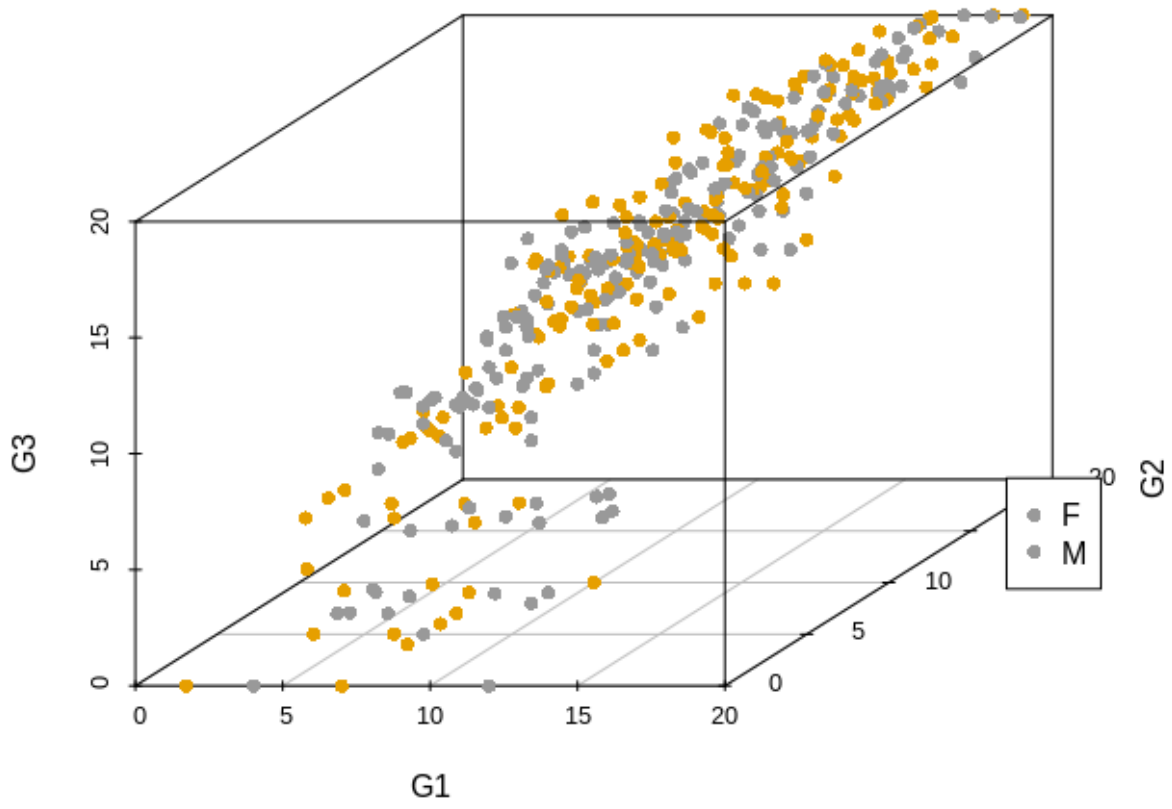


(c

با استفاده از کد زیر نمودار سه بعدی بین سه متغیر G1 و G2 و G3 نسبت به هم کشیده شده اند و بر اساس متغیر جنسیت با دو رنگ مختلف نمایش داده شده اند.

```
colors <- c("#999999", "#E69F00")
colors <- colors[as.numeric(StudentsPerformance$sex)]
sp = scatterplot3d(StudentsPerformance[,c('G1', 'G2', 'G3')], pch = 16, color=colors)
legend(sp$xyz.convert(24, 10, 4.5), legend = levels(StudentsPerformance$sex),
      col =  colors, pch = 16)
```

همانطور که مشاهده می شود این سه متغیر با افزایش همدیگر افزایش می یابند اما جنسیت افراد بر روی آنها تاثیری ندارد.



سوال ۵)

(a

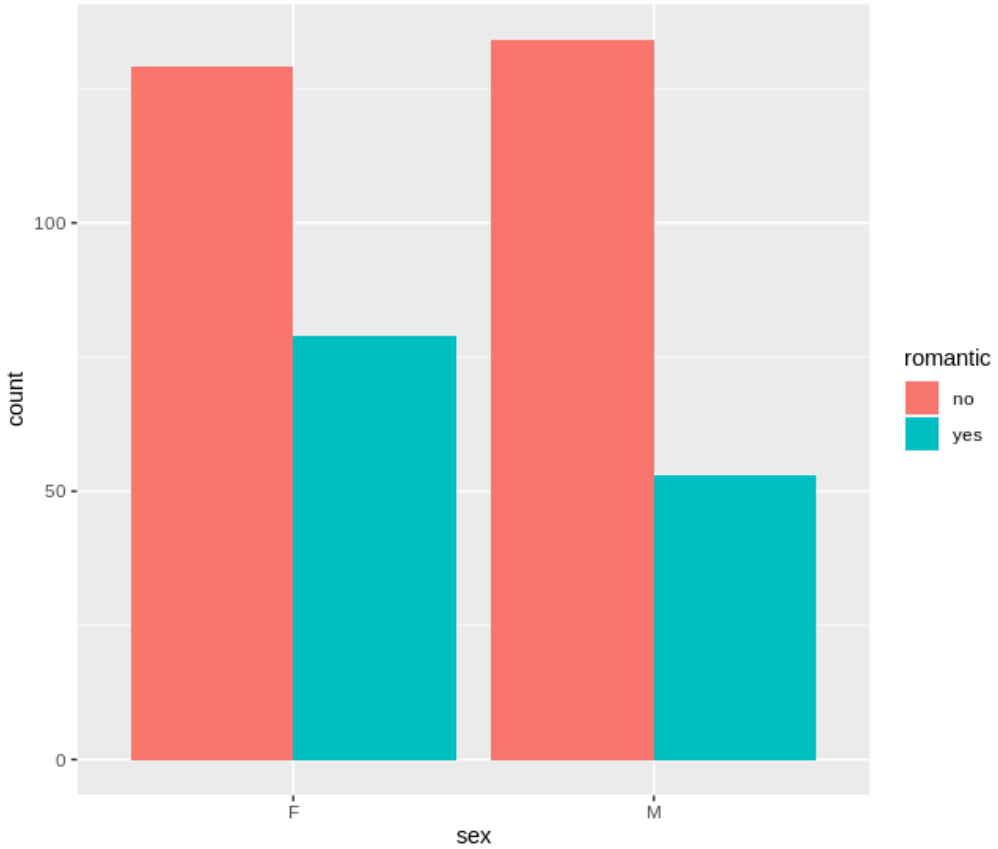
دو متغیر جنسیت و رمانتیک به عنوان دو متغیر categorical انتخاب شده اند. Table آن ها نیز در زیر آمده است.

```
> table(StudentsPerformance[,c('sex', 'romantic')])
   romantic
sex  no yes
  F 129  79
  M 134  53
>
```

(b

Grouped bar chart

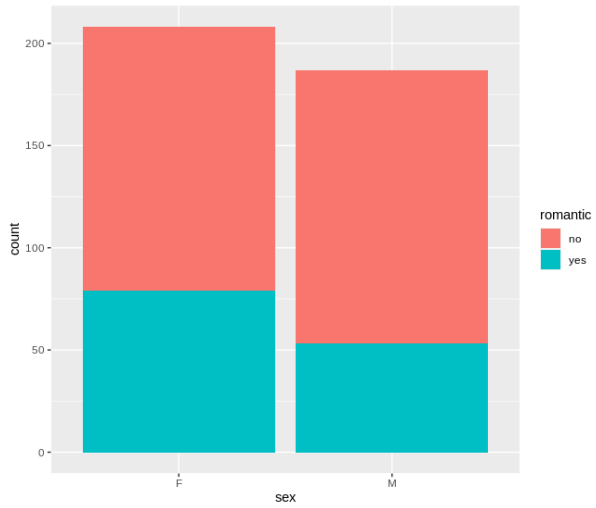
```
ggplot(StudentsPerformance, aes(x = sex, fill = romantic)) +
  geom_bar(position = "dodge")
```



(c

Segmented bar plot

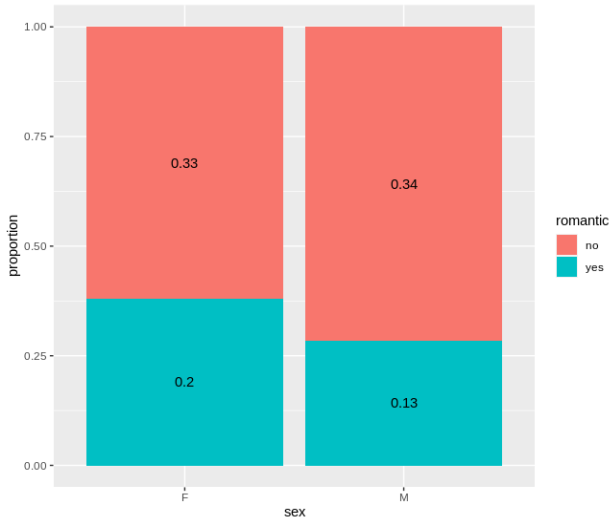
```
ggplot(StudentsPerformance, aes(x = sex, fill = romantic)) +
  geom_bar(position = "stack")
```



(d

Mosaic plot

```
ggplot(data = StudentsPerformance, aes(x = sex, fill = romantic)) +
  geom_bar(position = "fill") + ylab("proportion") +
  stat_count(geom = "text",
             aes(label = stat(round((..count..)/sum(..count..), digits = 2))),
             position=position_fill(vjust=0.5), colour="black")
```



سوال ۶)

متغیر G2 را انتخاب میکنیم. چون تعداد جامعه ۳۹۴ تاست و برای شرط independence مجبوریم کمتر از ۱۰ درصد جامعه نمونه برداری کنیم پس n=23 تا sample به صورت تصادفی از جامعه انتخاب می کنیم.

```
n_samples = 35
samples = sample_n(StudentsPerformance, n_samples)$G2
```

حال شرایط را بررسی میکنیم:

Independence: random & n = 35 < 10% of population → independence

Sample size/skew: n = 35≥ 30 & sample not skewed→ nearly normal sampling distribution

پس از توزیع نرمال استفاده می کنیم.

(a

به صورت زیر confidence interval محاسبه شده و برابر است با ۹.۷۷ تا ۱۲.۱۱

```
> upper = samples_mean + qnorm((1-0.95)/2, lower.tail = FALSE) * population_sd / sqrt(n_samples)
> lower = samples_mean - qnorm((1-0.95)/2, lower.tail = FALSE) * population_sd / sqrt(n_samples)
> upper
[1] 12.11167
> lower
[1] 9.778657
```

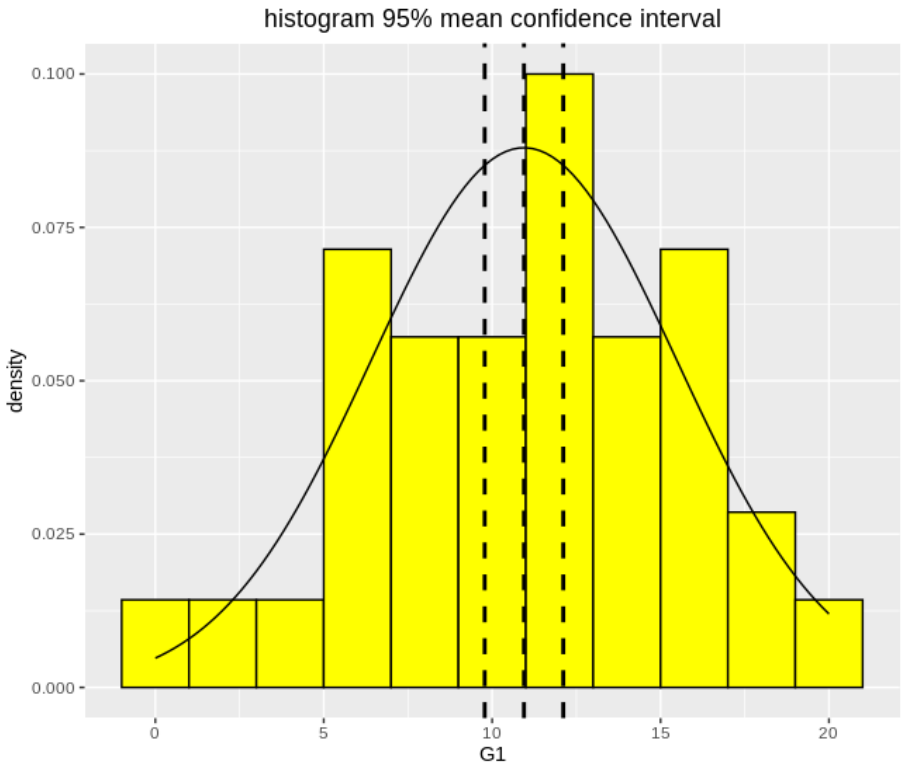
(b

۹۵ درصد مطمئنیم که میانگین نمرات G2 افراد در بازه بین ۹.۷ تا ۱۲.۱۱ قرار دارد.

(c

```
df = data.frame(
  G1 = samples
)
ggplot(df, aes(x = G1)) +
  geom_histogram(aes(y =..density..), binwidth = 2, fill="yellow", color="black") +
  stat_function(fun = dnorm, args = list(mean = samples_mean,
                                         sd = samples_sd)) +
  ggtitle("histogram 95% mean confidence interval") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_vline(aes(xintercept=lower),
             linetype="dashed", size=1) +
  geom_vline(aes(xintercept=upper),
             linetype="dashed", size=1) +
  geom_vline(aes(xintercept=samples_mean),
             linetype="dashed", size=1)
```

در شکل زیر histogram نمونه ها کشیده شده است و میانگین نمونه ها و همچنین ۹۵ درصد confidence interval برای میانگین نمونه با خط چین مشخص شده است.



(d

```
> mean(population)
[1] 10.78285
> samples_mean
[1] 10.94516
>
```

میانگین واقعی جامعه برابر ۱۰.۷۸ و میانگین نمونه ها برابر با ۱۰.۹۴ به دست آمده است. بنابراین باید تستی طراحی کنیم که آیا میانگین نمونه از نظر آماری آیا بزرگتر از میانگین واقعی است یا نه؟.

Independence: random & n = 50 < 10% of population → independence

Sample size/skew: n = 35≥ 30 & sample not skewed→ nearly normal sampling distribution

H0: mean= 10.78

H1: mean > 10.78

```
> se = population_sd / sqrt(n_samples)
> z = (samples_mean - population_mean) / se
> p_value = pnorm(z, lower.tail = FALSE)
> p_value
[1] 0.3925336
```

در نتیجه مقدار p-value > alpha=0.5 پس نمیتوان فرض صفر را رد کرد و شواهد کافی برای بیشتر بودن میانگین نمونه ها وجود ندارد. که انتظارمان نیز همین بود چون H0 میانگین واقعی جامعه بود.

(e

بله. در قسمت b به این نتیجه رسیدیم که ۹۵ درصد اطمینان داریم که نمرات بین ۹.۷ تا ۱۲.۱۱ قرار دارد. پس میانگین نمونه که 10.94 به دست آمده چون در این بازه قرار دارد پس نمیتوان با alpha=0.5 فرض صفر را رد کرد.

(g, ff

برای محاسبه power به actual mean نیاز داریم. در مرحله قبل null value را برابر میانگین جامعه گذاشتیم تا بتوانیم observation - null را محاسبه کنیم. بنابراین اینجا اگر actual mean را میانگین جامعه در نظر بگیریم type 2 error معنایی ندارد.

خطای نوع دوم احتمال عدم رد کردن h0 با فرض میانگین واقعی جدید است.

هم چنین power از یک منهای type 2 error به دست می آید. که هر چقدر effect size بیشتر باشد power تست بیشتر است و آسان تر میتوان فرض h0 را رد کرد.

سوال ۷)

(A

(a

باید از t-test استفاده کنیم چون تعداد sampleها از ۳۰ کوچکتر است.

Independence: random & n = 25 < 10% of population → independence

Sample size: 30 ≥ n = 25 & sample skewed→ t sampling distribution

(b

چون دو گروه از هم مستقل نیستند و هر نمونه مربوط به یک دانش آموز است پس تست یک paired analysis است.

بنابراین تفاضل دو گروه که نمرات G1 و G2 هستند را حساب میکنیم و میانگین آن را به عنوان observation در نظر میگیرم پس:

H0: u_diff = 0

hA: u_diff != 0

۱- ابتدا به صورت رندوم ۲۵ نمونه از جامعه انتخاب میکنیم. سپس تفاضل G1 و G2 را برای هر نمونه محاسبه کرده و در نهایت میانگین این تفاضل ها را به دست می آوریم که عدد 0.76- به دست می آید:

```
> n_samples = 25
> samples = sample_n(StudentsPerformance, n_samples)
>
> G1 = samples$G1
> G2 = samples$G2
>
> diff = G1 - G2
> u_diff = mean(diff)
> u_diff
[1] -0.7642462
>
```

در ادامه برای محاسبه t-statistic ابتدا se را با استفاده از sd جامعه محاسبه میکنیم

```
> population_sd = sd(StudentsPerformance$G1 - StudentsPerformance$G2)
> se = population_sd / sqrt(n_samples)
> se
[1] 0.4294247
```

در نهایت t-statistic را محاسبه کرده و چون تست دو طرفه است ۲ p-value برابر مقدار t کوچکتر از t-statistic به دست می آید.

```
>
> t = u_diff / se
> df = n_samples - 1
>
> p_value = 2 * pt(t, df=df)
> p_value
[1] 0.08779291
```

همانطور که مشاهده می شود مقدار p-value > alpha=0.05 است پس نمی توان فرض h0 را رد کرد و شواهد کافی برای تفاوت نمرات هر دانش آموز در این دو درس با توجه به نمونه گیری انجام شده وجود ندارد.

(B

استقلال داخل گروهی وجود دارد زیرا sampling به صورت رندوم انجام میشود. اما مشکلی که وجود دارد این است که تعداد ۱۰۰ نمونه کمتر از ۱۰ درصد جامعه نیست اما چون در صورت سوال n=100 انتخاب شده است چاره ای نداریم جز این که این شرط را قبول کنیم. استقلال بین گروهی نیز وجود دارد چون نمونه برداری از متغیر G1 و G2 مستقل از هم انجام میشود.

چون تعداد نمونه ها از ۳۰ بیشتر است پس میتوان از توزیع نرمال استفاده کرد.

و میتوان تست مقایسه دو میانگین را انجام داد.

H0: mean(G1) - mean(G2) = 0

HA: mean(G1) - mean(G2) != 0

ابتدا نمونه گیری ۱۰۰ تایی را به صورت مستقل انجام می دهیم سپس observation را که تفاضل میانگین این دو گروه است محاسبه میکنیم که عدد -2.16 به دست می آید:

```
> G1 = sample_n(StudentsPerformance, n_samples)$G1
> G2 = sample_n(StudentsPerformance, n_samples)$G2
> observation = mean(G1) - mean(G2)
> observation
[1] -2.161691
```

سپس se و df را مطابق فرمول های مقایسه دو میانگین و بر اساس sd جامعه به دست می آوریم.

```
> n_a = length(G1)
> n_b = length(G2)
> se = sqrt((sd(StudentsPerformance$G1) ^ 2) / n_a + (sd(StudentsPerformance$G1) ^ 2) / n_b)
> t = (observation - 0) / se
> df = min(n_a, n_b)
```

در نهایت در یک تست دو طرفه مقدار p-value را به دست می آوریم که نزدیک صفر است و بنابراین فرض صفر رد میشود و شواهد کافی وجود دارد مبنی بر این که میانگین نمرات این دو درس با هم متفاوت است.

```
> p_value = pt(t, df=df) * 2
>
> p_value
[1] 3.398767e-05
```

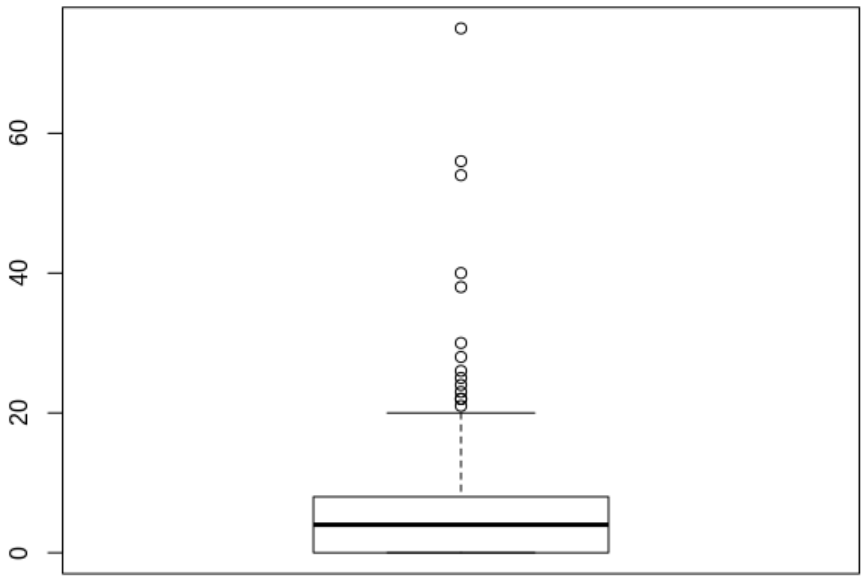
مطابق با زیر confidence interval نود و پنج درصد برای تفاوت این دو گروه بین 1.17- تا 3.14- به دست می آید که باز هم نشان دهنده رد شدن فرض صفر با alpha 0.05 درصد است یعنی ۹۵ درصد اطمینان داریم که تفاضل میانگین دو گروه در این بازه قرار دارد و صفر نیست.

```
> upper = observation + qt((1-0.95)/2, df=df, lower.tail = FALSE) * se
> lower = observation - qt((1-0.95)/2, df=df, lower.tail = FALSE) * se
>
> upper
[1] -1.173767
> lower
[1] -3.149615
>
```

سوال ۸)

a)

تعداد غیبت ها دارای outlier است پس این متغیر را انتخاب میکنیم.



ابتدا نمونه گیری رندوم با سایز ۲۰ انجام میدهیم.

```
n_samples = 20
samples = sample_n(StudentsPerformance, n_samples)$absences
```

سپس دویست bootstrap samples برای میانگین ایجاد میکنیم:

```
n_bootstrap = 200
bootstrap_samples <- c()

for(i in 1:n_bootstrap){
  bootstrap_a <- sample(1:n_samples, n_samples, replace=TRUE)
  mean <- mean(samples[bootstrap_a])
  bootstrap_samples <- c(bootstrap_samples, mean)
}
```

حال برای این توزیع bootstrap بازه اطمینان را به صورت زیر محاسبه میکنیم که برای بازه اطمینان ۹۵ درصد بازه 3.09 تا 7.5 برای میانگین به دست می آید.

```
> quantile(bootstrap_samples, probs=c(0.025, 0.975))
  2.5%  97.5%
3.0975 7.5000
```

(b

Independence: random & n = 200 < 10% of population → independence

Sample size/skew: n = 200≥ 30 & sample not skewed→ nearly normal sampling distribution

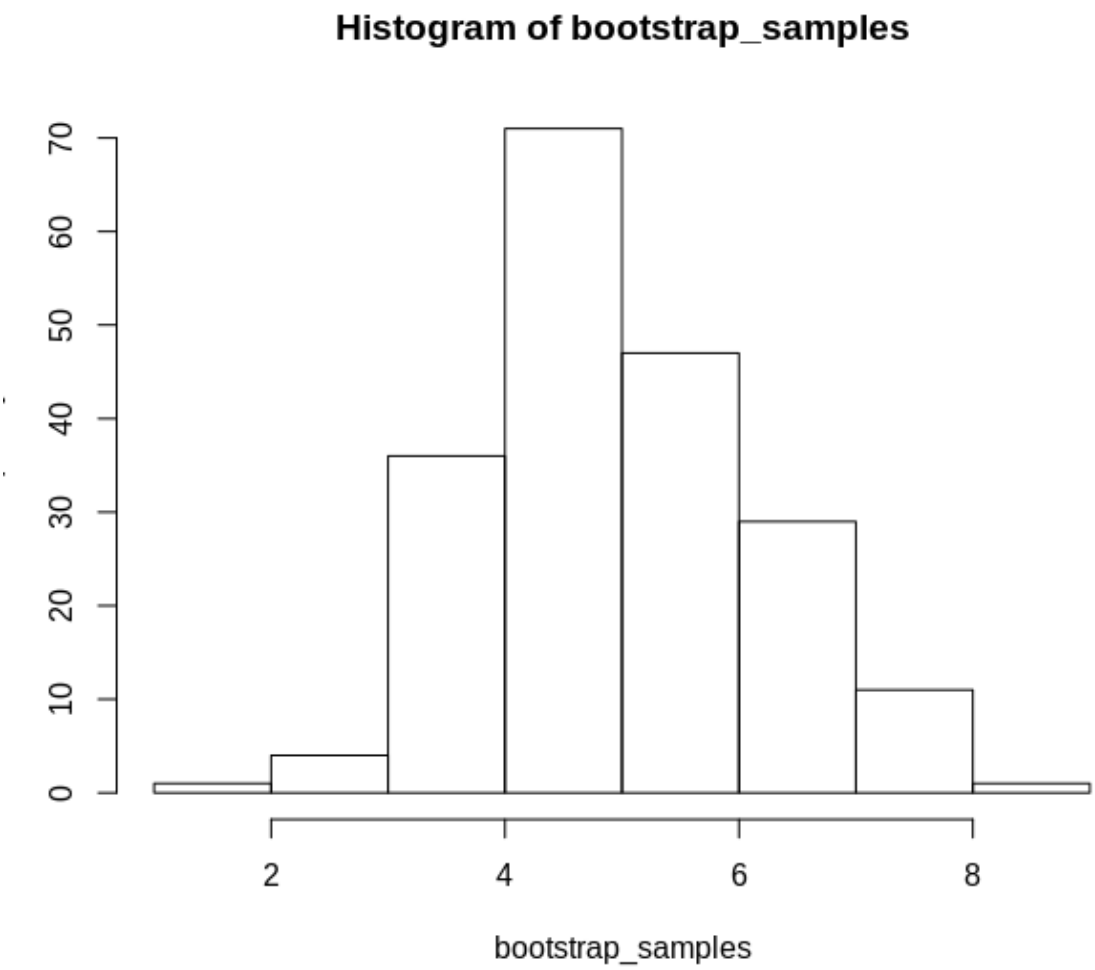
پس میتوانیم از توزیع نرمال برای محاسبه confidence interval استفاده کنیم.

مطابق با زیر بازه اطمینان ۹۵ درصد از 1.48 تا 8.7 با استفاده از روش standard error به دست می آید. Standard error نیز از تقسیم انحراف معیار توزیع bootstrap بر جذر تعداد توزیع bootstrap به دست می آید.

```
> samples_mean = mean(samples)
> bootstrap_sd = sd(boostrap_samples)
> upper = samples_mean + qnorm((1-0.95)/2, lower.tail = FALSE) * bootstrap_sd / sqrt(n_bootstrap)
> lower = samples_mean - qnorm((1-0.95)/2, lower.tail = FALSE) * bootstrap_sd / sqrt(n_bootstrap)
> upper
[1] 8.717291
> lower
[1] 1.482709
```

(c

بازه اطمینان در روش دوم بزرگتر به دست آمده است که دلیل آن میتواند به خاطر این باشد که توزیع bootstrap همچنان از توزیع نرمال پیروی نمیکند و دارای skewness است.



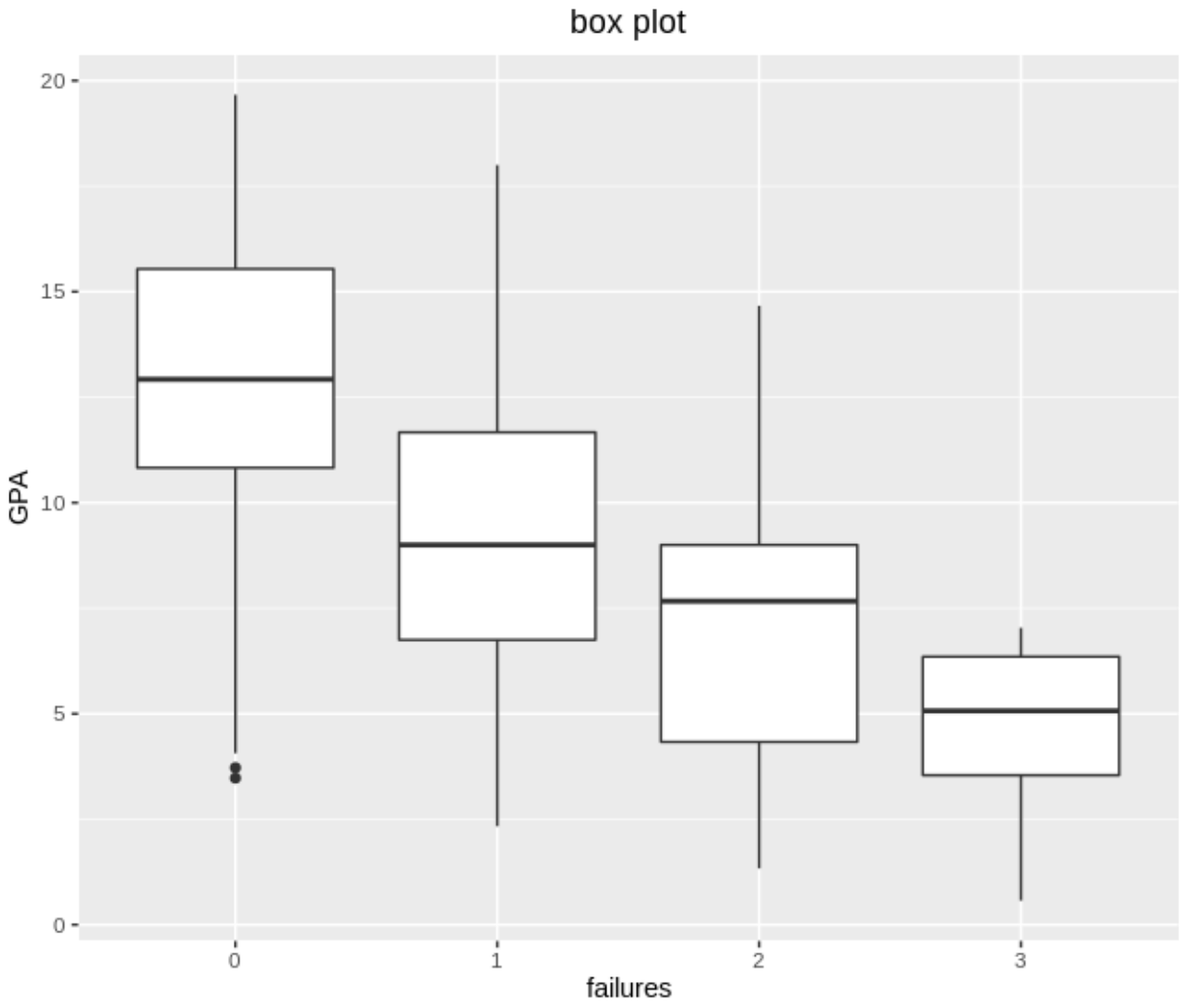
سوال ۹)

در ابتدا میانگین نمرات سه درس را محاسبه میکنیم.

```
StudentsPerformance$GPA = (StudentsPerformance$G1 + StudentsPerformance$G2 +
                           StudentsPerformance$G3) / 3
```

ابتدا با استفاده از box plot نمرات گروه های مختلف را بررسی میکنیم.

```
StudentsPerformance$failures=as.factor(StudentsPerformance$failures)

ggplot(StudentsPerformance, aes(x = failures, y=GPA)) +
  geom_boxplot() +
  ggtitle("box plot") +
  theme(plot.title = element_text(hjust = 0.5))
```



ابتدا باید شروط زیر را چک کنیم:

۱- استقلال بین گروهی: برای برقراری این شرط نیاز است random sampling است پس از هر گروه اما چون تعدادی از گروه ها کمتر از ۲۰ نمونه دارند پس فرض میکنیم این شرط نیز برقرار است.

۲- استقلال داخل گروهی که برقرار است زیرا نمونه ها نمرات دانش آموزان مختلف است.

۳- توزیع ها تقریبا نرمال هستند.

۴- توزیع ها تقریبا واریانس برابر دارند.

تست Anova:

فرض صفر: میانگین ۴ گروه failures با هم برابر هستند.

فرض جایگزین: میانگین حداقل دو گروه failures با هم تفاوت دارند.

با انجام تست Anova نتیجه زیر به دست آمده است:

```
> result = aov(GPA ~ failures, data = StudentsPerformance)
> summary(result)
             Df Sum Sq Mean Sq F value Pr(>F)
failures      3   1994   664.8   58.22 <2e-16 ***
Residuals   391   4464    11.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

بنابراین چون p-value نزدیک به صفر است بنابراین فرض صفر رد میشود و حداقل دو گروه وجود دارند که میانگین آن ها با هم تفاوت دارند.

این نتیجه از box plot کشیده شده در ابتدای جواب سوال نیز می توانست به صورت تصویری به دست آید.