

## • Question 0

### بخش A:

دیتاست StudentsPerformance داده‌های واقعی دانشآموزان دو مدرسه است که شامل نمره‌ی آن‌ها در ۳ آزمون، و اطلاعات مربوط به دانشآموزان است که شامل نام مدرسه، جنسیت، سن، شغل پدر و مادر، تعداد دفعات بیرون رفتن دانشآموزان، استفاده کردن یا نکردن از اینترنت، داشتن یا نداشتن روابط عاطفی، میزان مطالعه، تعداد درس‌های افتاده، سطح سلامتی و تعداد غیبت‌ها است.

مطالعه‌ی این دیتاست، به ما این امکان را می‌دهد که ارتباط نمره‌ی دانشآموزان با این اطلاعات را بررسی کنیم؛ و در صورتی که بتوانیم ارتباط قوی‌ای بین خروجی آن‌ها و یک یا بیشتر از این رفتارها پیدا کنیم، احتمال اینکه با کنترل برخی از این‌ها، بتوانیم بازدهی دانشآموزان را افزایش دهیم، وجود دارد. البته این مطالعه experimental نیست. وجود ارتباط مبنی بر علیت نیست.

### بخش B:

این دیتاست، ۳۹۵ نمونه دارد و برای هر نمونه ۱۵ ویژگی (شامل نمره‌ی ۳ آزمون مختلف) را داریم.

### بخش C:

این دیتاست، برای تمامی دانشآموزان، تمامی ویژگی‌ها را دربر دارد و هیچ مقدار null‌ای در آن موجود نیست. اما در صورتی که این مقادیر وجود داشت؛ یک راه برای پر کردن این مقادیر؛ استفاده از مد داده‌های است. یعنی مقداری که بیشترین تکرار در دیتاست را داشته است.

روشی مثل استفاده از میانگین هم وجود دارد؛ اما میانگین می‌تواند خیلی تحت تاثیر داده‌های پرت باشد و البته برای متغیرهای categorical هم معنا ندارد.

### بخش D:

حدس اولیه‌ی من این است که تعداد ساعت مطالعه‌ی دانشآموزان و تعداد غیبت‌ها (که احتمالاً رابطه‌ی نسبتاً نزدیکی با سطح سلامتی آن‌ها خواهد داشت) اثر بیشتری بر نمره‌ی آن‌ها داشته باشد. همچنین تعداد افتادن درس‌ها می‌تواند تخمینی از سطح تلاش دانشآموز باشد؛ در نتیجه می‌تواند در پیش‌بینی نمره‌ی آن‌ها اثرگذار باشد. تعداد دفعات بیرون رفتن از این نظر که زمانی از دانشآموز صرف می‌کند؛ در مرحله‌ی بعدی قرار دارد. اما از نظر من عواملی چون مدرسه‌ی دانشآموز، جنسیت، سن، شغل پدر و مادر، استفاده از اینترنت و یا روابط عاطفی تاثیر چندانی بر نمرات آن‌ها نخواهد داشت.

همچنین احتمال می‌دهم که نمرات G1، G2 و G3 دارای correlation مثبت و نسبتاً قوی‌ای باشند. چرا که دانشآموزان درس‌خوان احتمالاً برای هر سه، زمان خواهند گذاشت.

## • Question 1

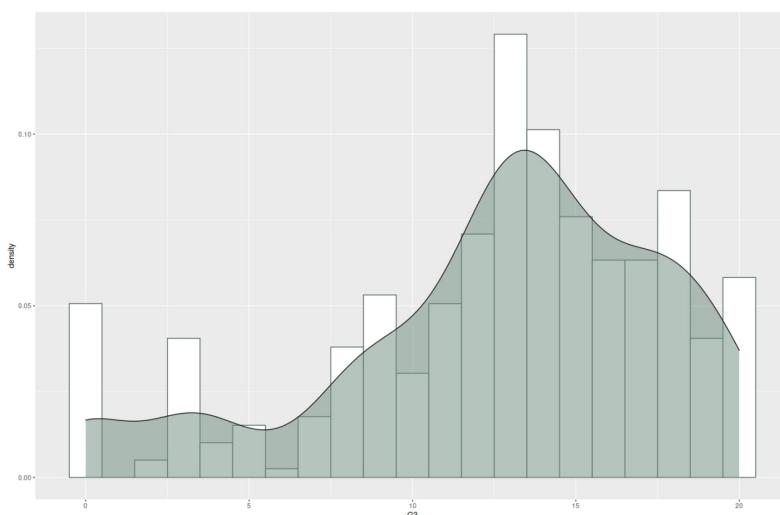
برای این بخش، از نمره‌ی G3 در این دیتاست استفاده شده است.

### بخش A:

با استفاده از کد زیر، نمودار خواسته شده را رسم کردیم.

```
ggplot(students, aes(x=G3)) +  
  geom_histogram(aes(y=..density..),  
                 bins=21, color='#475c53', fill='white'  
  ) +  
  geom_density(alpha=0.5, fill='#6c8c7e')
```

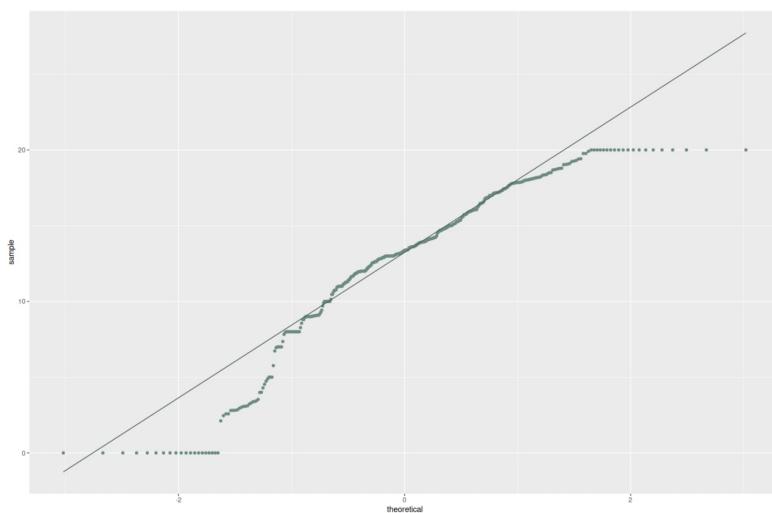
که نتیجه‌ی آن را در شکل زیر ملاحظه می‌کنید.



همانطور که مشخص است این منحنی را می‌توان 2-modal نامید. چرا که دو پیک در آن مشاهده می‌شود.

## بخش B:

از آنجایی که نمره‌ی کمتر از صفر، و نمره‌ی بیشتر از ۲۰ نمی‌توانیم داشته باشیم؛ توزیع نمی‌تواند نرمال باشد. حدسی که قبل از مقایسه‌ی این توزیع، با توزیع نرمال می‌توانیم داشته باشیم؛ این است. که احتمالاً در نمرات میانی خیلی نزدیک به. توزیع نرمال خواهد بود و در نواحی نمرات بالا و پایین از این توزیع فاصله خواهد گرفت. و همانطور که در قسمت قبل هم مشاهده کردیم، نمودار توزیع این نمرات کمی چولگی چپ دارد. حالا در تصویر زیر، می‌توانیم QQ-plot این توزیع و توزیع نرمال را مشاهده کنیم.



همانطور که حدس می‌زدیم در قسمت بالا و پایین نمودار، از توزیع نرمال دور شده‌ایم. همچنین با توجه به اینکه در این دو بخش، نقاط در زیر خط قرار دارند، یعنی توزیع ما دارای چولگی چپ است.

برای رسم QQ-plot از کد زیر استفاده شده است.

```
ggplot(students, aes(x=G3)) +  
  geom_histogram(aes(y=..density..),  
                 bins=21, color='#475c53', fill='white'  
  ) +  
  geom_density(alpha=0.5, fill='#6c8c7e')
```

## بخش C:

میزان چولگی را می‌توان با معیار  $sk = (mean - median)/sd$  سنجید. اگر این مقدار، عدی مثبت باشد؛ توزیع دارای چولگی راست و اگر منفی باشد چولگی چپ است. همچنین اگر این مقدار برابر صفر باشد، یعنی توزیع متقارن است. مقدار بدست آمده در این بخش فلان است و در نتیجه، این محاسبات هم نشان می‌دهد که توزیع نمرات دانشآموزان در این آزمون دارای چولگی چپ است. مقدار چولگی در این ویژگی هم گزارش شده است.

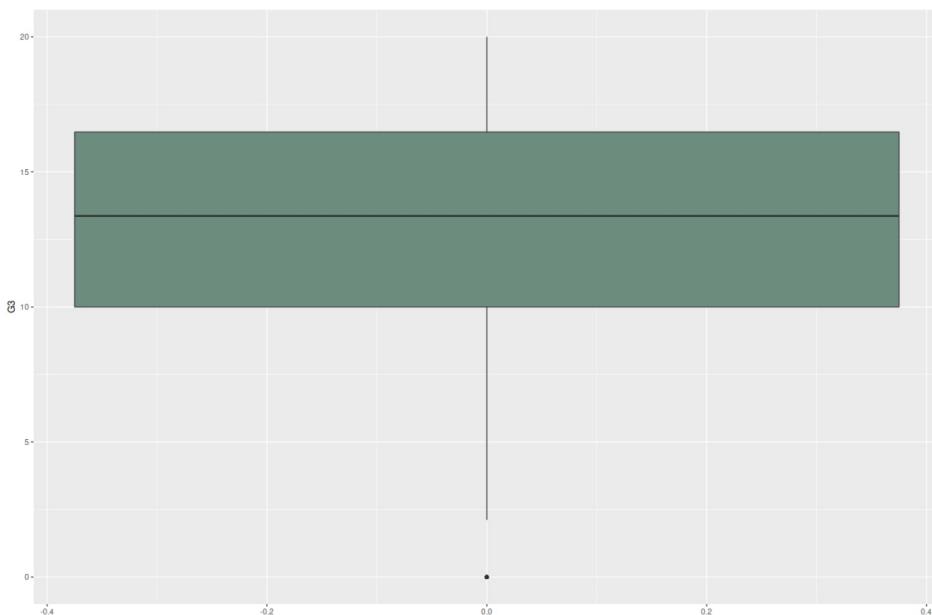
skewness	-0.141
----------	--------

محاسبه‌ی این مقدار هم تنها با یک خط کد زیر، قابل انجام است.

```
skewness = (mean(students$G3) - median(students$G3)) / (sd(students$G3))
```

## بخش D:

برای نشان‌دادن outlierها، از یک boxplot استفاده شده است.



کد این نمودار نیز به این صورت است:

```
ggplot(students, aes(y=G3)) + geom_boxplot(fill='#6c8c7e')
```

همانطور که در نمودار مشخص است؛ نقاط نقطه‌ی صفر، outlier هستند. در واقع دانشآموزانی که نمره‌ی صفر دریافت کرده‌اند، رفتاری بسیار متفاوت از دیگران داشته‌اند. مطالعه‌ی ویژگی‌های آن‌ها احتمالاً می‌تواند دید مناسبی از علت این اتفاق در اختیار ما بگذارد. عکس زیر ویژگی‌های این دانشآموزان (به جز نمرات آن‌ها) را نمایش می‌دهد.

school	sex	age	Fjob	Mjob	goout	internet	romantic	studytime	failures	health	absences
GP	M	18	other	services	3	yes	no	1	2	4	0
GP	F	15	teacher	services	2	yes	yes	3	2	5	0
GP	F	16	other	other	2	yes	yes	1	2	5	0
GP	M	17	other	other	5	yes	no	1	3	5	0
GP	F	15	services	health	2	yes	no	2	3	3	0
GP	M	18	other	other	5	yes	yes	1	3	4	0
GP	M	19	at_home	services	4	yes	yes	1	3	4	0
GP	M	17	other	at_home	2	yes	yes	1	2	5	0
GP	M	16	other	other	4	no	no	1	1	5	0
GP	M	16	other	other	5	yes	no	1	2	2	0
GP	F	16	services	at_home	5	yes	yes	2	3	3	0
GP	F	17	other	at_home	4	no	yes	3	1	5	0
GP	M	18	services	other	4	yes	no	2	1	2	0
GP	F	19	services	services	4	no	yes	2	1	3	0
GP	M	18	services	teacher	3	yes	no	2	1	2	0
GP	F	17	at_home	at_home	1	yes	yes	2	1	4	0
MS	F	17	services	other	1	yes	yes	1	1	1	0
MS	M	19	services	other	2	no	no	1	1	5	0
MS	F	19	other	services	2	yes	no	3	1	5	0
MS	F	18	other	other	1	no	no	2	1	5	0

در نگاه اول، به نظر می‌رسد که زمان مطالعه‌ی اکثر آن‌ها پایین است و هیچکس نیست که کمتر از یکبار، افتاده باشد. با مشاهده‌ی این ویژگی‌ها می‌توان بررسی‌های بیشتری برای راه حل‌های احتمالی انجام داد. چرا که این مشاهدات کاملاً چشمی، و این آزمایش هم است.

## بخش E:

جدول زیر مقادیر این آماره‌ها را گزارش کرده است:

آماره	مقدار
میانگین	12.64
میانه	13.37
واریانس	26.47
انحراف معيار	5.14

- تعریف هر کدام از آنها هم به این صورت است:
- میانگین: در واقع جمع تمامی مقادیر با یکدیگر، تقسیم بر تعداد آنهاست. این مقدار نشان‌می‌دهد که یک ویژگی به‌طور متوسط، چه مقداری دارد. همچنین از داده‌های پرت، بسیار تاثیرپذیر است و `roust` نیست.
- میانه: عددی است که نصف مقادیر قبل از آن، و نصف دیگر بعد از آن قرار دارند. اگر تعداد داده‌ها فرد باشد، این عدد، همان عدد وسط اعداد مرتب‌شده‌ی ماست و اگر زوج باشد، میانگین دو عدد وسط.
- واریانس: واریانس معیاری برای اندازه‌گیری پراکندگی است و برابر است با امید ریاضی توان دوم تفاضل هر نمونه، با میانگین. این معیار از جنس توان دوم داده است.
- انحراف معیار: برابر جذر واریانس است و از این نظر که هم‌جنس داده‌ی ماست برای بسیاری از کاربردها مناسب‌تر است.
- برای محاسبه‌ی هر یک از این‌ها تنها استفاده از توابع پیش‌فرض `R` نیاز است.

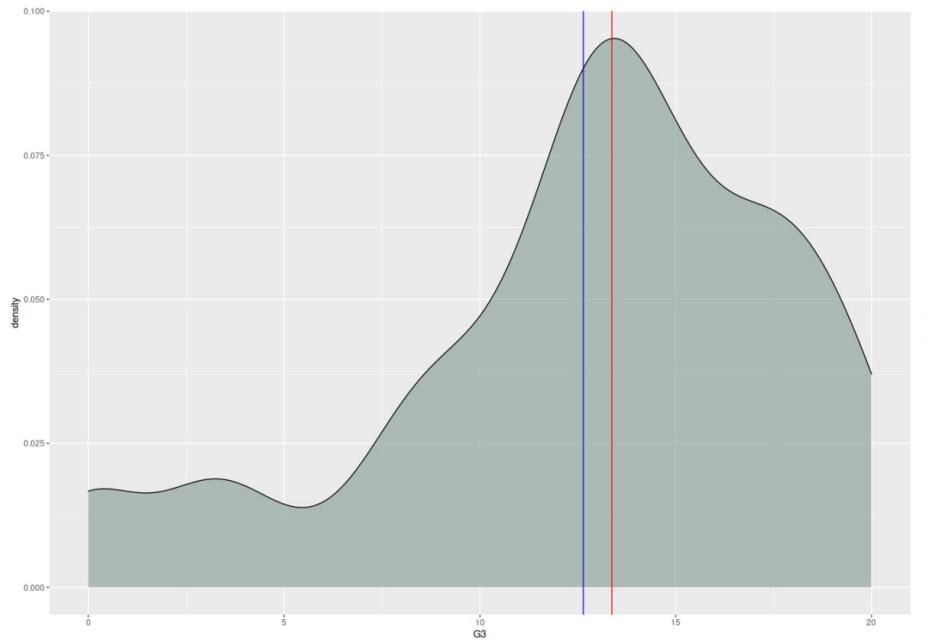
```
mean_ = mean(students$G3)
median_ = median(students$G3)
var_ = var(students$G3)
sd_ = sd(students$G3)
```

## بخش F:

با استفاده از کد زیر، نمودار خواسته‌شده رسم می‌شود.

```
ggplot(students, aes(x=G3)) +
  geom_density(alpha=0.5, fill="#6c8c7e") +
  geom_vline(aes(xintercept=mean_), linetype='mean'), color='blue') +
  geom_vline(aes(xintercept=median_), linetype='median'), color='red') +
  scale_linetype_manual(
    name = 'lines',
    values = c('mean' = 1, 'median' = 1),
    guide = guide_legend(override.aes = list(colour = c('blue', 'red'))))
```

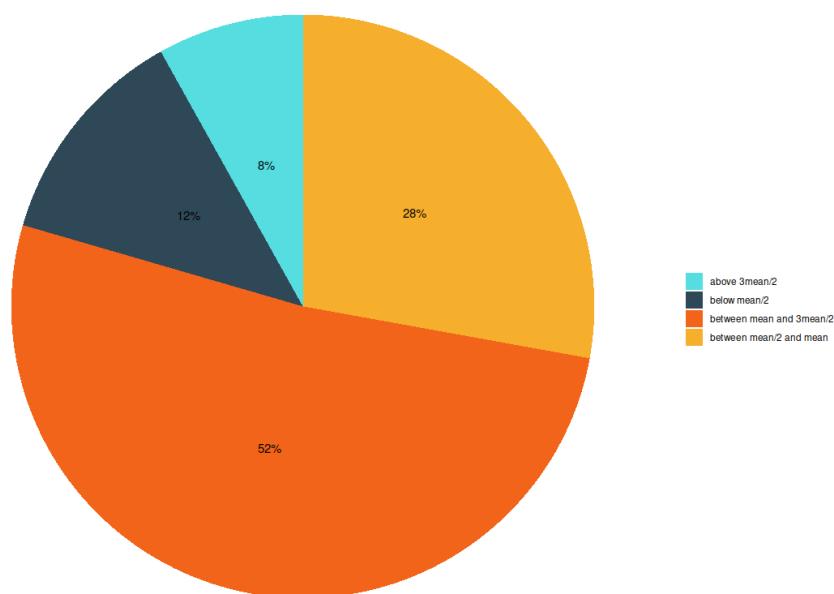
نمودار حاصل، به این صورت است.



همانطور که در نمودار مشخص است؛ میانگین در سمت چپ میانه قرار دارد و این یعنی توزیع ما چولگی به چپ دارد که در نمودار density هم مشخص است. (از آنجایی که میانگین یک آماره‌ی robust نیست، و از outlier‌ها تاثیر زیادی می‌گیرد، و همچنین اینکه در قسمت‌های قبلی مشاهده کردیم که افراد با نمره‌ی صفر داده‌های پرت ماست؛ منطقی است که میانگین ما بیشتر به سمت نمرات کوچکتر کشیده شود در حالی که تعداد آن‌ها خیلی زیاد نیست).

## بخش G:

برای تقسیم بعدی با استفاده از میانگین، داده‌ها را به چهار دسته‌ی ۱- کمتر یا مساوی نصف میانگین ۲- بزرگ‌تر از نصف میانگین و کمتر یا مساوی میانگین ۳- بزرگ‌تر از میانگین و کمتر یا مساوی ۱.۵ برابر میانگین ۴- بزرگ‌تر از ۱.۵ برابر میانگین تقسیم کردیم. pie-chart زیر سهم هر کدام از این دسته‌ها را نشان می‌دهد.



در زیر می‌توانید کدهای تقسیم‌بندی و رسم نمودار را مشاهده کنید.

```
mean_q1 = mean_/2
mean_q3 = mean_*1.5

g1_p = nrow(subset(students, G3 <= mean_q1)) / nrow(students)
g2_p = nrow(subset(students, G3 > mean_q1 & G3 <= mean_)) / nrow(students)
g3_p = nrow(subset(students, G3 > mean_ & G3 <= mean_q3)) / nrow(students)
g4_p = nrow(subset(students, G3 > mean_q3)) / nrow(students)

mean_seperated_df = data.frame(
  "group"=c('below mean/2', 'between mean/2 and mean', 'between mean and 3mean/2', 'above 3mean/2'),
  "porportion"=c(g1_p, g2_p, g3_p, g4_p)
)
ggplot(mean_seperated_df, aes(x="", y=porportion, fill=group)) + geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) + geom_text(aes(label=paste0(round(porportion*100), "%")), position=position_stack(vjust = 0.5)) +
  scale_fill_manual(values=c("#55DDE0", "#2F4858", "#F26419", "#F6AE2D")) +
  labs(x=NULL, y=NULL, fill=NULL, title="Categorized base on means") +
  theme_classic() + theme(axis.line = element_blank(), axis.text = element_blank(), axis.ticks = element_blank())
```

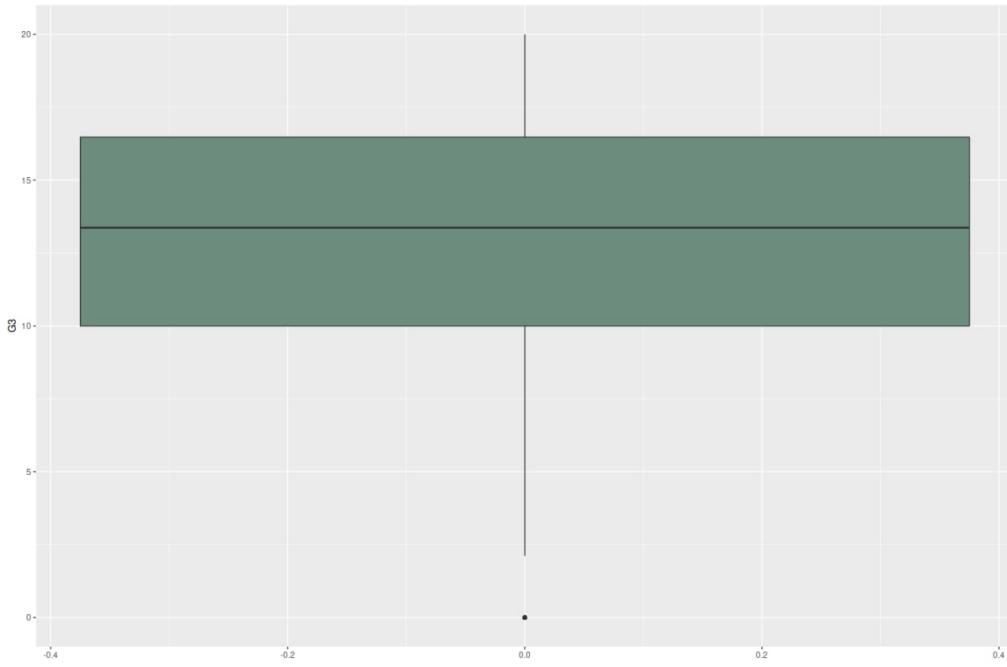
## بخش H:

مقادیر در جدول زیر گزارش شده‌اند:

آماره	مقدار
lower quartile (Q1)	10.00
median (Q2)	13.37
upper quartile (Q3)	16.47
IQR	6.47
lower whisker	0.29
upper whisker	20

که با استفاده از این قطعه کد محاسبه شده و پلات را هم بعد از آن می‌توانید مشاهده کنید.

```
ggplot(students, aes(y=G3)) + geom_boxplot(fill='#6c8c7e')
quartiles = quantile(students$G3)
IQR = quartiles[4] - quartiles[2]
upper_whisker = min(max(students$G3), quartiles[4] + 1.5*IQR)
lower_whisker = max(min(students$G3), quartiles[2] - 1.5*IQR)
```



- **Question 2**

برای این بخش، از شغل مادران (Mjob) در این دیتاست استفاده شده است.

**بخش A:**

جدول زیر انواع شغل‌ها، فرکانس و درصد آن‌ها را نشان می‌دهد.

شغل	درصد	فرکانس
at home	0.15	59
health	0.09	34
services	0.26	103
teacher	0.15	58
other	0.36	141

مقادیر آن، با کد زیر بدست آمده‌اند.

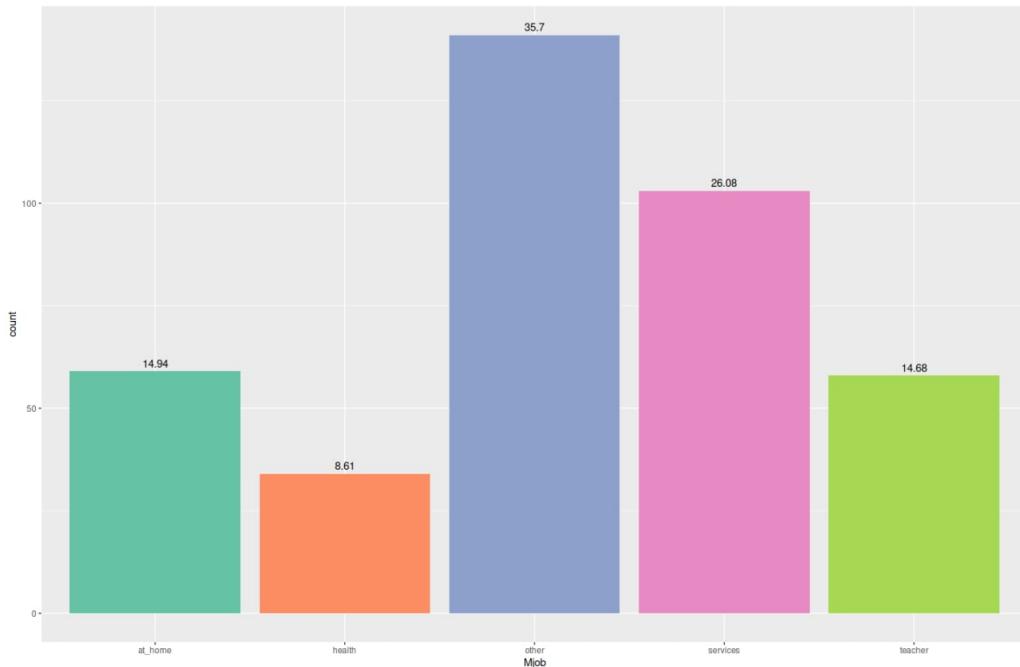
```
library(janitor)
library(dplyr)
tabyl(students$Mjob, sort = TRUE)
```

## بخش B:

نمودار با استفاده از کد زیر قابل رسم است.

```
library(RColorBrewer)
cou1 <- brewer.pal(5, "Set2")
ggplot(data=students, aes(x=Mjob)) +
  geom_bar(aes(y=..count..), fill=cou1, stat='count') +
  geom_text(aes(label=round(..count../nrow(students)*100, digits=2), y=..count..), stat="count", vjust=-0.5)
```

و نتیجه به این صورت است.



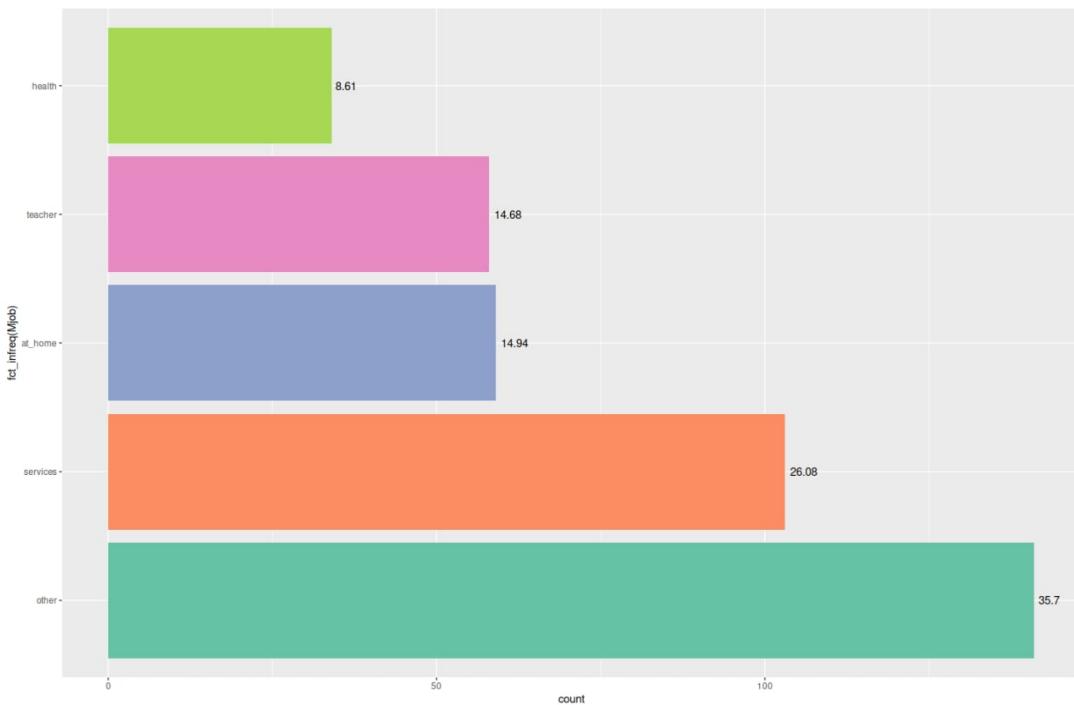
## بخش C:

برای رسم این نمودار، باید کد قبلی به این صورت تغییر پیدا کند.

```
library(forcats)
ggplot(data=students, aes(x=fct_infreq(Mjob))) +
  geom_bar(aes(y=..count..), fill=cou1, stat='count') +
  geom_text(aes(label=round(..count../nrow(students)*100, digits=2), y=..count..), stat="count", vjust=0.5) +
  coord_flip()

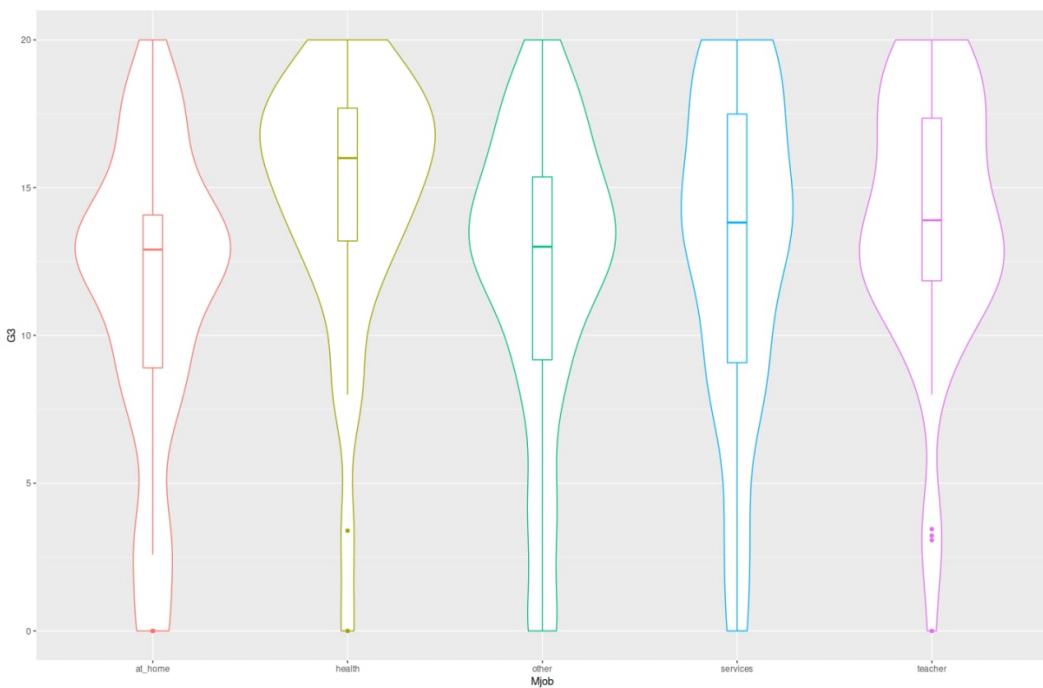
ggplot(students, aes(x=Mjob, y=G3, color=Mjob)) +
  geom_violin() + geom_boxplot(width=0.1) +
  theme(legend.position="none")
```

و به این صورت به نمایش درمی‌آید.



## بخش D:

برای رسم این نمودار، متغیر عددی G3 را انتخاب کردیم که برای هر گروه از 'Mjob' را رسم کند.



برای رسم violin-plot قطعه کد زیر استفاده می‌شود.

```
ggplot(students, aes(x=Mjob, y=G3, color=Mjob)) +  
  geom_violin() + geom_boxplot(width=0.1) +  
  theme(legend.position="none")
```

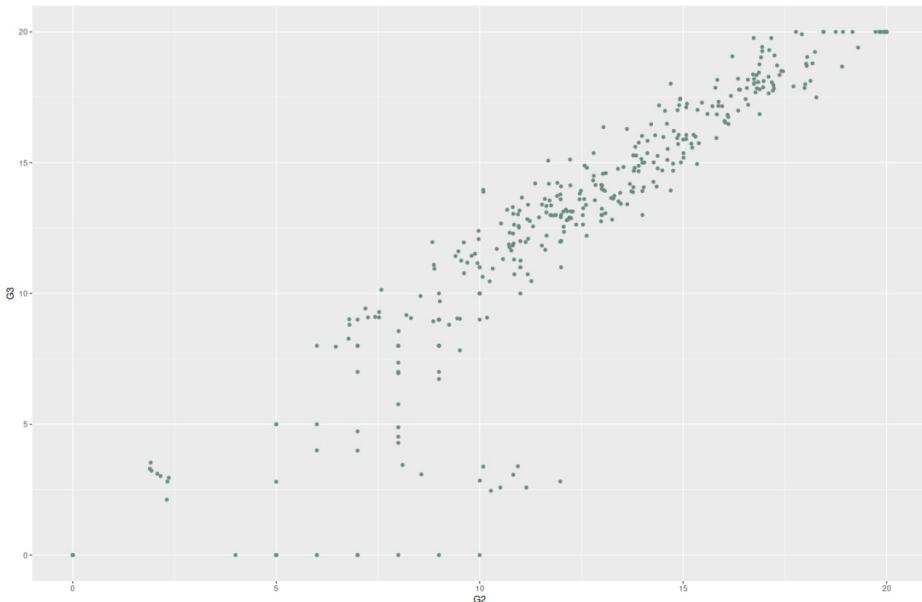
## • Question 3

برای این بخش، از نمره‌ی دانشآموزان در دو آزمون G2 و G3 استفاده شده است.  
**:بخش A:**

به نظر می‌رسد که این دو متغیر با یکدیگر correlation مثبتی داشته باشند. چرا که احتمالاً دانشآموزانی که نمره‌ی خوبی در یک آزمون دریافت می‌کنند؛ اکثربیت در سایر آزمون‌ها هم نمره‌ی بالایی خواهند داشت.

## **:بخش B:**

رسم شده، به این صورت است.



که با کد زیر رسم می‌شود.

```
ggplot(students, aes(x=G2, y=G3)) + geom_point(color='#6c8c7e')
```

در بخش نمرات بالای ۵، رابطه‌ی خطی و مثبت، کاملاً مشهود است. در نمرات پایین‌تر اما رابطه‌ی خیلی واضح نیست. همچنین علامت این وابستگی مثبت است و به نظر می‌رسد که خطی باشد.

## بخش C:

عدد حاصل از محاسبه‌ی correlation با دستور

```
corr = cor(students$G2, students$G3)
```

برابر با ۰.۹۱۵ است که نشان از رابطه‌ی بسیار قوی‌ای دارد.

## بخش D:

از آنجایی که عدد correlation چیزی بین -۱ تا ۱ است؛ قدر مطلق آن نشان‌دهنده‌ی قدرت ارتباط است؛ که در اینجا با عدد بدست‌آمده، رابطه‌ی قوی‌ای بین نمره‌ی G2 و G3 وجود دارد. همچنین علامت این عدد نشان دهنده‌ی این است که این دو رابطه‌ی مستقیم دارند و یا رابطه‌ی عکس. که رابطه‌ی آن‌ها مستقیم است. از آنجایی که این عدد توسط pearson correlation محاسبه شده و عدد قابل توجهی هم است؛ همچنین این correlation از نوع خطی است؛ در نتیجه می‌توانیم نتیجه بگیریم که رابطه‌ی دو متغیر ما هم از جنس خطی است. در نتیجه حدسیات قسمت اول، به نظر صحیح می‌رسند.

## بخش E:

برای محاسبه‌ی p-value در این correlation از این دستور استفاده شده است.

```
cor.test(students$G2, students$G3)
```

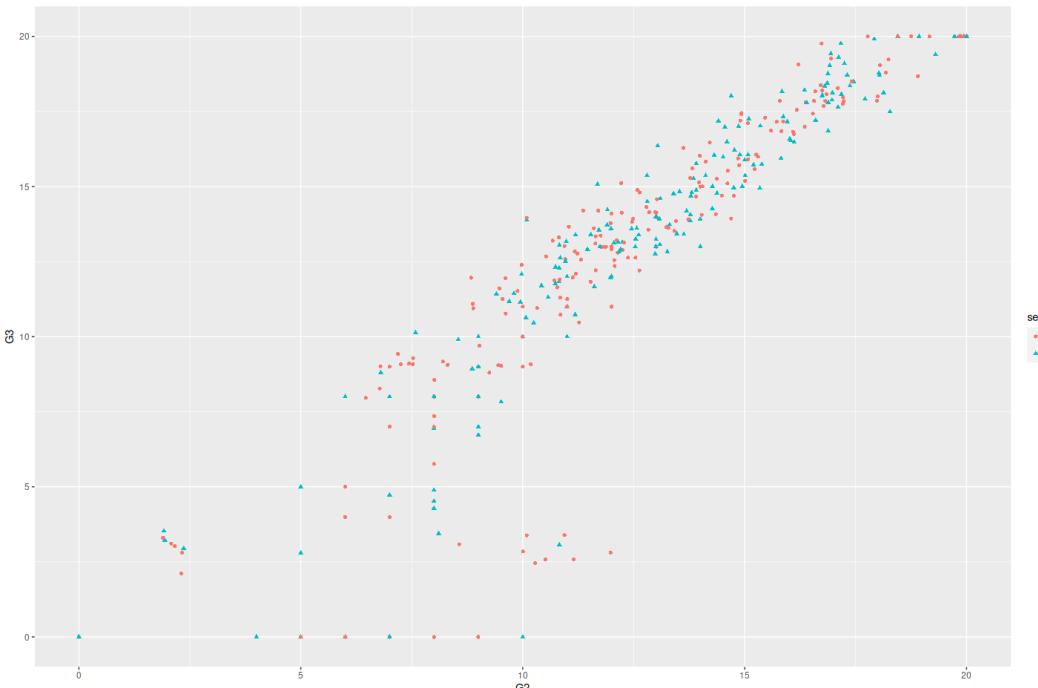
و مقدار بدست آمده برای آن برابر  $2.2e-16$  است. این عدد، بسیار عدد کوچکی است. در نتیجه ما می‌توانیم فرض صفر را -که در اینجا برابر این است که رابطه‌ای بین این دو متغیر وجود ندارد و مقدار correlation آن‌ها برابر صفر است- را رد کنیم.

در واقع p-value بیان می‌کند که اگر فرض صفر ما صحیح باشد؛ احتمال این‌که مقدار مشاهده‌شده‌ی ما رخ دهد؛ چقدر است. و اگر این عدد، عدد کوچکی باشد؛ در نتیجه احتمال

اینکه مشاهده‌ی ما تنها براساس شانس باشد، بسیار پایین می‌آید. و می‌توانیم فرض صفر را رد کنیم.

## بخش F:

تصویر بدستآمده در اینجا آورده شده است.



که بر شکل و رنگ نقاط بر اساس جنسیت افراد تعیین شده‌اند. از نمودار این طور به‌نظر می‌رسد (تنها با مشاهده) هر دو رفتارهای مشابهی دارند.  
برای رسم آن هم از کد زیر استفاده شده است.

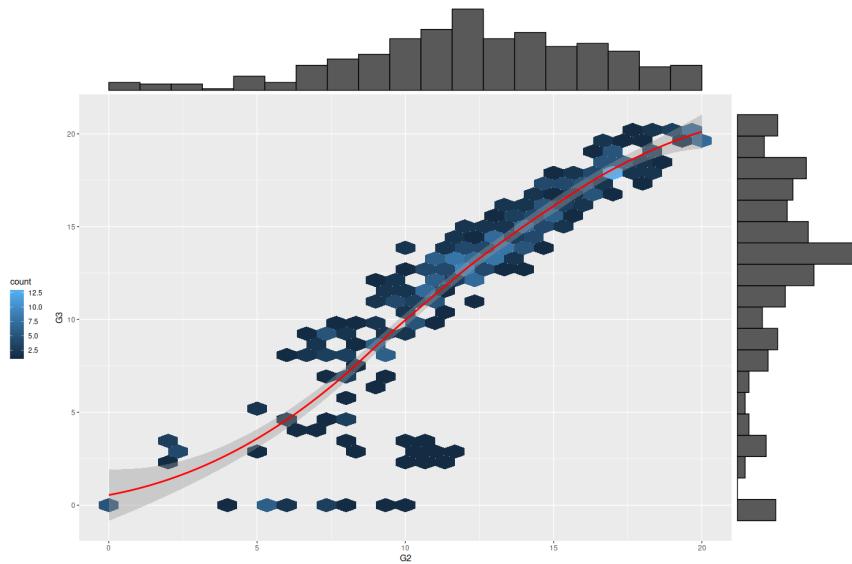
```
ggplot(students, aes(x=G2, y=G3)) +  
  geom_point(aes(shape=sex, color=sex))
```

## بخش G:

با استفاده از این کد، می‌توان یک marginal distribution hexin را به‌همراه رسم کرد.

```
p = ggplot(students, aes(G2, G3)) + geom_point(col="transparent")  
ggMarginal(  
  p+geom_hex(bins=20) + theme(legend.position='left') +  
  geom_smooth(method = "loess", color='red') , type="histogram",  
  bins=20  
)
```

نمودار بدستآمده با تعداد ۲۰ بین در شکل زیر قابل مشاهده است.

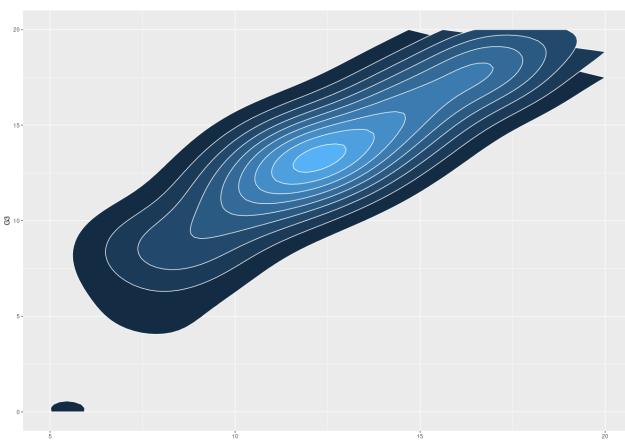


اگر به نمودار توجه کنیم؛ نقاط کمزنگ‌تر بیشتر نزدیکی خط  $y=x$  قرار دارند. از آنجایی که ارتباطی که در قسمت‌های قبل پیدا کردیم هم، ارتباطی قویا خطی بود؛ این مسئله منطقی است (افرادی که نمرات نسبتاً یکسان در دو درس دارند، تقریباً زیاد است). با تغییر bin-size اندازه‌ی ۶ ضلعی‌ها تغییر می‌کند؛ که یعنی با کاهش سایز bin هر کدام از آنجا ارتباط رنج وسیع‌تری از دو متغیر را در بر خواهد داشت. در نتیجه با بزرگ‌کردن بیش از اندازه آن، احتمالاً generalization را از دست می‌دهیم و با کوچک‌کردن آن دیگر قادر به مشاهده‌ی جزئیات نخواهیم بود. که با اجرای کد زیر حاصل می‌شود.

```
ggplot(students, aes(x=G2, y=G3) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")
```

## بخش H:

شکل زیر، 2D رسم شده برای این دو متغیر است.



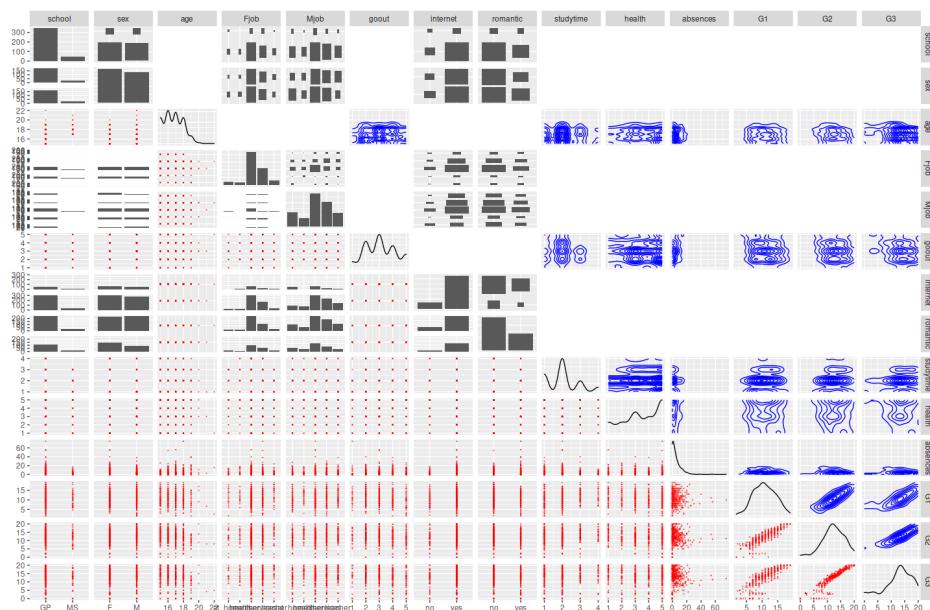
- Question 4

## بخش A:

برای رسم correlogram قطعه‌کد زیر را خواهیم داشت.

```
library(GGally)
cols = 2:16
cols = cols[-10]
ggpairs(
  students, columns=cols,
  upper=list(combo='blankDiag', continuous = wrap('density', color='blue')),
  lower=list(continuous=wrap('points', color='red', size=0.2, alpha=0.5),
             combo=wrap('points', color='red', size=0.2, alpha=0.5))
)
```

برای جفت‌های categorical و 2d-density plot، numerical نداریم و در قطر هم برای متغیرهای numerical، توزیع و برای categorical histogram داریم.



## بخش B:

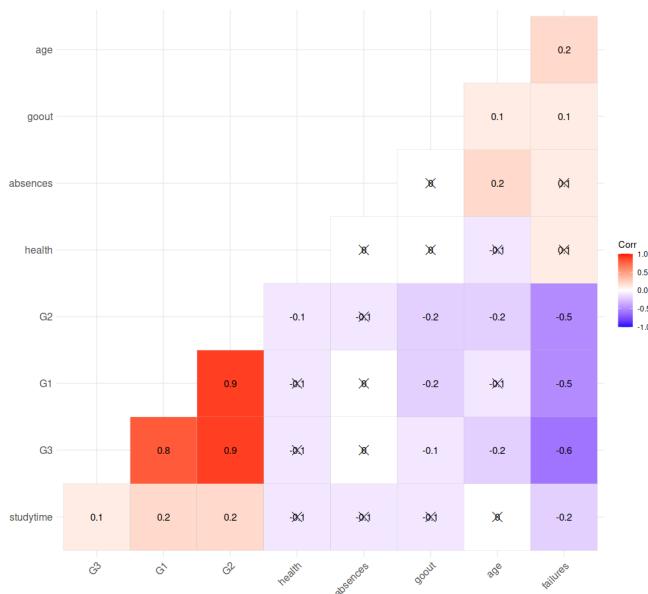
در این بخش با استفاده از قطعه‌کدی که در ادامه خواهیم دید heatmap correlogram رسم کردہ‌ایم. برای این بخش، تنها می‌توانیم متغیرهای numerical داشته باشیم. پس ابتدا متغیرهای categorical را حذف کرده و سپس correlation و p-value آن‌ها را حساب کردیم.

```

library(GGally)
cols = 2:16
cols = cols[-10]
ggpairs(
  students, columns=cols,
  upper=list(combo='blankDiag', continuous = wrap('density', color='blue')),
  lower=list(continuous=wrap('points', color='red', size=0.2, alpha=0.5),
             combo=wrap('points', color='red', size=0.2, alpha=0.5))
)

```

که نتیجه به این صورت است.



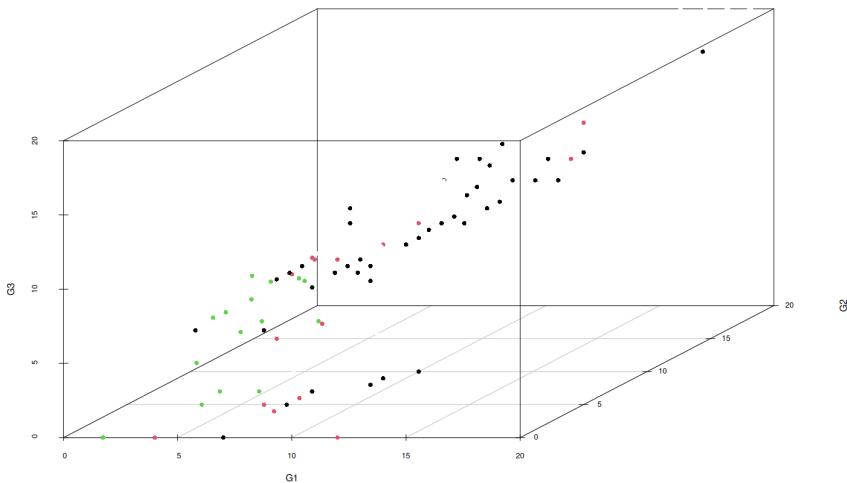
همانطور که در نمودار بالا مشخص است بیشترین ارتباط مثبت مربوط به سه نمره‌ی G1، G2 و G3 است و بیشترین ارتباط منفی هم مربوط به تعداد failure است که موید حدس‌های اولیه ماست. تعداد غیبت‌ها هم به نظر می‌رسد که هیچگونه ارتباطی نداشته باشند. همچنین زوج‌هایی که آن‌ها قادر به رد کردن فرض عدم وجود ارتباط بین آن‌ها نبوده؛ با علامت ضربدر مشخص شده‌اند.

## بخش C:

سه متغیر numerical استفاده شده، همان G1، G2 و G3 هستند و متغیر categorical تعداد failure است. کد زیر می‌تواند این پلات را رسم کند.

```
library(scatterplot3d)
scatterplot3d(students[, c('G1', 'G2', 'G3')], color=students$failures, pch=16)
```

و نتیجه به مانند زیر خواهد بود.



## • Question 5

برای این بخش دو متغیر جنسیت و رمانتیک بودن یا نبودن استفاده شده است.

**بخش A:**

برای این بخش، خروجی کد زیر

```
ft = ftable(students$sex ~ students$romantic)
```

این نتایج را در اختیار ما می‌گذارد:

	male	female
romantic=yes	53	79
romantic=no	134	129

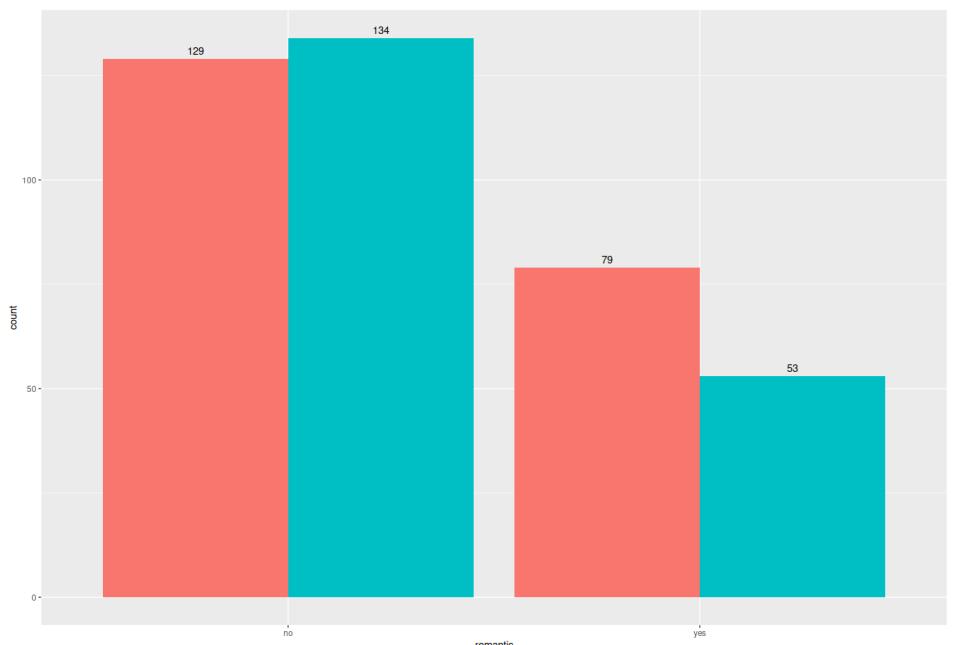
به نظر می‌رسد (تنها با مشاهده) خانم‌ها رمانتیک‌تر از آقایان هستند؛ در حالی که در مجموع اکثر افراد رمانتیک نیستند.

## بخش B:

برای رسم grouped bar chart از کد زیر استفاده شده است.

```
ggplot(students, aes(x=romantic, fill=sex)) +  
  geom_bar(position="dodge") +  
  geom_text(aes(label=..count..), stat='count', position=position_dodge(0.9), vjust=-0.5)
```

و خود نمودار در زیر آورده شده است.

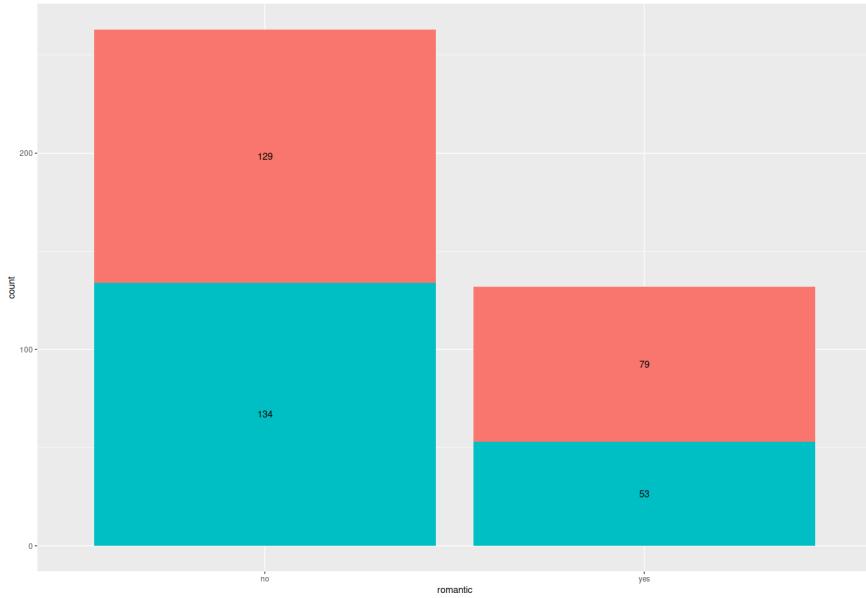


## بخش C:

با قطعه کد زیر segmented bar plot رسم شده است.

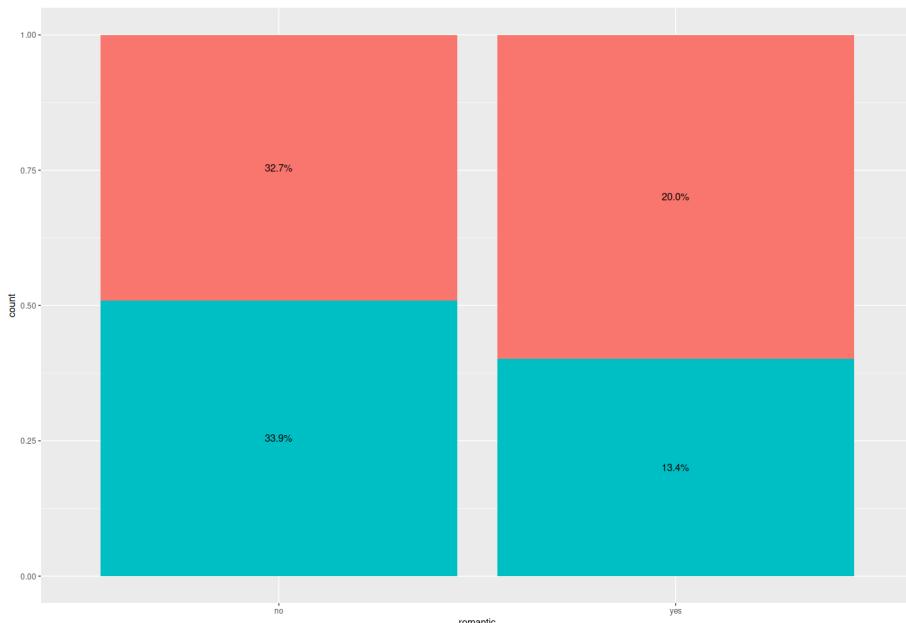
```
ggplot(students, aes(x=romantic, fill=sex)) +  
  geom_bar(position="dodge") +  
  geom_text(aes(label=..count..), stat='count', position=position_dodge(0.9), vjust=-0.5)
```

که نتیجه‌ی آن به این صورت است.



:D بخش

نمودار زیر mosaic plot



با کد زیر قابل رسم است.

```
ggplot(students,aes(x=romantic,fill=sex)) +
  geom_bar(position="fill") +
  geom_text(aes(label=scales::percent(..count../sum(..count..))),
            stat='count',position=position_fill(vjust=0.5))
```

## • Question 6

از G1 برای سوالات این بخش استفاده شده است.

### بخش A:

برای محاسبه‌ی این بخش، در ابتدا یک سمپل ۳۵ تایی از این داده انتخاب شده، که چون حداقل ۳۰ تا را داراست و کمتر از ۱۰ درصد داده است، می‌توانیم با استفاده از CLT مسائل را حل کنیم. با کد زیر مقدار میانگین،  $se$  و  $sd$  بدست‌آمد و با یافتن  $z$  بازه‌ی اطمینان ۹۵ درصدی را برای میانگین این متغیر محاسبه کردیم. (در همه‌ی بخش‌های این سوال از همین سمپل استفاده شده است).

```
sampled = students[sample(nrow(students), size=35), ]  
mean_ = mean(sampled$G1)  
sd_ = sd(sampled$G1)  
se_ = sd_/sqrt(nrow(sampled))  
z = qnorm(p=0.025, lower.tail=FALSE)  
ci = c(mean_ - z*se_, mean_ + z*se_)
```

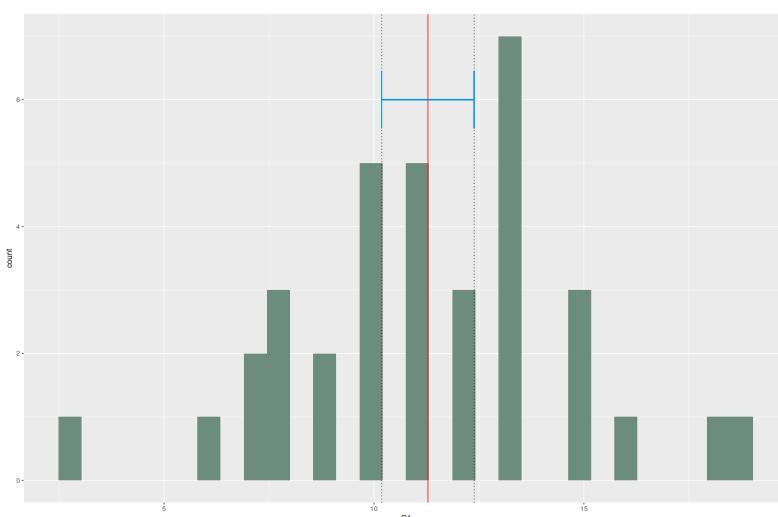
که نتیجه بازه‌ی (10.18526, 12.38617) است.

### بخش B:

مقدار بدست‌آمد در بخش قبل، بیان می‌کند که ما ۹۵ درصد اطمینان داریم که میانگین واقعی جامعه‌ی ما، در بازه‌ی بدست‌آمد قرار دارد. در واقع تعبیر بازه‌ی اطمینان این است که اگر ما تعداد خیلی زیادی سمپل ۳۵ تایی از جامعه برداریم و بازه‌ی اطمینان همه‌ی آن‌ها را محاسبه کنیم؛ ۹۵ درصد آن‌ها شامل مقدار اصلی میانگین خواهند بود.

### بخش C:

نمودار خواسته شده به این صورت است.



که با استفاده از کد زیر آن را رسم کردیم.

```
ggplot(sampled, aes(x=G1)) +  
  geom_histogram(fill="#6c8c7e") +  
  geom_vline(xintercept = mean_, color='red') +  
  geom_vline(xintercept=ci[1], linetype="dotted") +  
  geom_vline(xintercept=ci[2], linetype="dotted") +  
  geom_errorbarh(aes(y=6, x=mean_, xmin=ci[1], xmax=ci[2]), col="#0094EA", size=0.5)
```

## بخش D:

فرض صفر را این در نظر می‌گیریم که توزیع نمرات کاملاً حول مرز نقطه‌ی قبولی متقاضن هستند. در واقع میانگین این نمرات برابر ۱۰ است. سپس با استفاده از کد زیر، با محاسبه‌ی میانگین نمونه و  $se$  احتمال میانگین داده شده به شرط درستی فرض صفر را محاسبه می‌کنیم.

```
d = abs((mean_ - 10) / se_)  
p_value = 2 * pnorm(d, lower.tail=FALSE)
```

که این مقدار برابر با ۰.۰۲۲ است که از ۰.۰۵ کمتر است در نتیجه ما می‌توانیم فرض صفر را رد کنیم. در واقع این مقدار بیان می‌کند که در صورتی که فرض صفر صحیح باشد، احتمال اینکه ما این میانگین را در نمونه‌ی خود ببینیم برابر ۰.۰۲۲ است در نتیجه احتمال اینکه این داده، شناسی مشاهده شده باشد پایین است.

## بخش E:

بازه‌ی اطمینان بدستآمده برابر (10.18526, 12.38617) است که شامل نقطه‌ی ۱۰ نمی‌شود. در نتیجه از طریق بازه‌ی اطمینان ۹۵ درصدی هم ما می‌توانیم فرض صفر را رد کنیم.

## بخش F:

برای محاسبه‌ی خطای نوع دوم، کد زیر نوشته شده است.

```
a_mean = mean(students$G1)  
test_z = qnorm(p=0.05, lower.tail=FALSE)  
type_II = 1 - pnorm(mean_-test_z*se_, mean=a_mean, sd=se_) -  
  pnorm(mean_+test_z*se_, mean=a_mean, sd=se_, lower.tail=FALSE)
```

مقدار واقعی میانگین در واقع مقدار میانگین دیتاست است. در واقع احتمال اینکه ما نتوانیم فرض صفر را رد کنیم، به شرط آنکه فرض جایگزین واقعاً صحیح باشد را محاسبه می‌کند. مقدار بدستآمده برای این خطا برابر ۷۷٪ می‌باشد که خطای بسیار زیادی است.

## بخش G:

مقدار power که در واقع همان احتمال مکمل خطای نوع دو است را می‌توان از روش دیگر با استفاده از قطعه‌کد زیر محاسبه کرد.

```
a_mean = mean(students$G1)
test_z = qnorm(p=0.05, lower.tail=FALSE)
power = pnorm(mean_-test_z*se_, mean=a_mean, sd=se_) +
  pnorm(mean_+test_z*se_, mean=a_mean, sd=se_, lower.tail=FALSE)
```

که در واقع احتمال اینکه ما بتوانیم فرض صفر را رد کنیم در حالی که فرض جایگزین واقع صحیح است را محاسبه می‌کند. عدد بدستآمده همانطور که از قسمت قبل انتظار داشتیم، برابر ۰.۲۳٪ است. هرچه effect size power ما هم افزایش خواهد یافت. چرا که احتمال برداشتن سمپلی که فاصله‌ی بیشتری از میانگین واقعی داشته باشد هم افزایش خواهد یافت.

- Question 7

## بخش A:

در این بخش با برداشتن ۲۵ نمونه از جامعه، و انتخاب دو متغیر G2 و G3 هر کدام از این ۲۵ داده، کاملاً به یکدیگر وابستگی دارند. چرا که هر دو نمره، دقیقاً متعلق به یک نفر است و ما نمی‌توانیم استدلال کنیم که استقلال بین گروهی داریم. در نتیجه اختلاف آن‌ها را به عنوان یک متغیر جدید در نظر می‌گیریم و با استفاده از تست  $t$ ، فرض خود را بررسی می‌کنیم. در اینجا فرض ما این است که میانگین تفاضل این دو متغیر مخالف صفر است؛ در حالی که فرض صفر دقیقاً عکس این است. با توجه به توضیحات داده شد، مقدار p-value حاصل از قطعه‌کد زیر برابر  $1.8e-5$  است که عددی بسیار کوچک می‌باشد؛ پس ما می‌توانیم فرض صفر را رد کنیم.

```
sampled = students[sample(nrow(students), size=25), ]
sampled = sampled[, c('G2', 'G3')]
paired = sampled$G2 - sampled$G3
mean_0 = 0
sample_mean = mean(paired)
sample_sd = sd(paired)
mean_se = sample_sd / nrow(sampled)
d = abs((sample_mean - mean_0) / mean_se)
p_value = 2*pt(d, df=nrow(sampled)-1, lower.tail=FALSE)
```

## بخش B:

در اینجا چون ۱۰۰ سمپل به صورت جدا از نمرات G2 و ۱۰۰ سمپل از نمرات G3 انتخاب شده‌اند؛ دیگر نمی‌توانیم از آنالیز pair استفاده کنیم؛ چرا که دیگر هر یک از سمپل‌ها دقیقاً متناظر دیگری نخواهد بود. در نتیجه از t-test با استفاده از قوانین تعیین se برای اختلاف میانگین دو گروه استفاده کردیم و با استفاده از کد زیر، مقدار p-value برابر با ۰.۶۹ بdst می‌آید که عددی بسیار بزرگ است! در نتیجه ما نمی‌توانیم فرض صفر را رد کنیم.

```
sampled = students[sample(nrow(students), size=25), ]
sampled = sampled[, c('G2', 'G3')]
paired = sampled$G2 - sampled$G3
mean_0 = 0
sample_mean = mean(paired)
sample_sd = sd(paired)
mean_se = sample_sd / nrow(sampled)
d = abs((sample_mean - mean_0) / mean_se)
p_value = 2*pt(d, df=nrow(sampled)-1, lower.tail=FALSE)

# ----- part B
G2_sample = students[sample(nrow(students), size=100), ]$G2
G3_sample = students[sample(nrow(students), size=100), ]$G3

mean_G2 = mean(G2_sample)
mean_G3 = mean(G3_sample)

df = 99
se = sqrt(sd(G2_sample)**2/99 + sd(G3_sample)**2/99)

p_value = 2 * pt(abs(mean_G2-mean_G3), df=df, lower.tail=FALSE)
ci_t = qt(0.025, df=df, lower.tail=FALSE)
ci = c(mean_G2-mean_G3 - ci_t*se, mean_G2-mean_G3 + ci_t*se)
```

باشه اطمینان بdst آمده توسط کد بالا، باشه (0.99, 1.78) است که شامل نقطه‌ی صفر هم می‌شود و نتیجه بسیار با روش قبلی متفاوت است. در واقع اشکال اینجاست که ما در هر یک از نمونه‌ها نزدیک به ۲۵ درصد جامعه را برداشت‌هایم و دو گروه ما دیگر نمی‌توانند از یکدیگر مستقل باشند. در نتیجه نمی‌توان به نتایج تست اعتنا کرد.

## • Question 8

متغیر failures در این دیتاست برای این بخش مناسب به نظر می‌رسد.

## بخش A:

ابتدا با برداشتن ۱۰۰۰ سمپل ۲۰ تایی (با جایگزینی) از یک نمونه‌ی ۲۰ تایی انتخاب شده از جامعه‌ی آماری (بدون جایگزینی) یک bootstrap sample می‌سازیم. با حذف ۲.۵ درصد بزرگ و ۲.۵ درصد کوچک این نمونه، می‌توانیم به یک باشه اطمینان برای میانه برسیم که برابر با باشه (0, 3) بdst می‌آید. این محاسبات توسط قطعه‌کد زیر صورت می‌گیرد.

```

sampled = students[sample(nrow(students), size=20), ]
bootstrap_sample = c()
for (i in 1:1000){
  resample = sample(sampled$failures, size=20, replace=TRUE)
  bootstrap_sample = c(bootstrap_sample, resample)
}

bootstrap_median = median(bootstrap_sample)
ci = c(quantile(bootstrap_sample, probs=0.025), quantile(bootstrap_sample, probs=0.975))

```

## بخش B:

در این بخش با محاسبه‌ی بازه‌ی اطمینان، از روش se (standard error) ما را برابر با  $sd$  در bootstrap sample در نظر گرفتیم) به بازه‌ی (1.35, 1.35-1) می‌رسیم که کمی با قبلی متفاوت است.

```

bootstrap_sd = sd(bootstrap_sample)
ci_t = qt(0.025, df=nrow(sampled)-1, lower.tail=FALSE)
ci = c(bootstrap_median-ci_t*bootstrap_sd, bootstrap_median+ci_t*bootstrap_sd)

```

## بخش C:

مقدار بدستآمده در این دو بخش، با یکدیگر تفاوت زیادی دارد. یک علت آن احتمالاً این است که فرکانس تکرار مقدار ۰ در این متغیر بسیار زیاد است. در واقع اصلاً از یک توزیع متقارن پیروی نمی‌کند در نتیجه روش‌های standard error دقیق مناسبی بر روی آن نخواهند داشت. ضمن اینکه این توزیع، مقدار کمتر از صفر نمی‌تواند داشته باشد. روش اول چون تنها از روی داده‌ها تصمیم می‌گیرد؛ به این مشکل برخواهد خورد اما در روش دوم این صادق نیست. در اینجا هم از آنجایی که از توزیع میانه اطلاعی نداریم؛ استفاده از روش اول بسیار مناسب‌تر است.

## • Question 9

در ابتدا در کد زیر ستوان total که در واقع میانگین نمره‌ی دانشآموزان در ۳ آزمون است را اضافه می‌کنیم و سپس با استفاده از تست ANOVA بر روی گروه‌های با تعداد failure مختلف، تلاش می‌کنیم بررسی کنیم که آیا میانگین در این گروه‌ها متفاوت است، یا خیر.

```

students$total = rowMeans(students[, c('G1', 'G2', 'G3')])
one.way = aov(total ~ failures, data=students)

```

نتیجه‌ی این تست،  $p$ -value ای برابر  $2e-16$  است. فرض صفر در اینجا این است که تفاوتی بین میانگین این گروه‌ها وجود ندارد. در حالی که این  $p$ -value به ما اجازه می‌دهد که فرض صفر را با قدرت زیادی رد کنیم. و ادعا کنیم که با احتمال بالایی، حداقل بین دو گروه از این چند گروه، میانگین‌های متفاوتی وجود دارد.