R Notebook

Import Libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(dplyr)
```

Load our Data

```
data <- read.csv("assets/StudentsPerformance.csv")
data%>%head()
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
```

```
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

```
ndata<-mutate(data,mean=(`math.score`+`reading.score`+`writing.score`)/3) #mean of three score
table(data$parental.level.of.education)
```

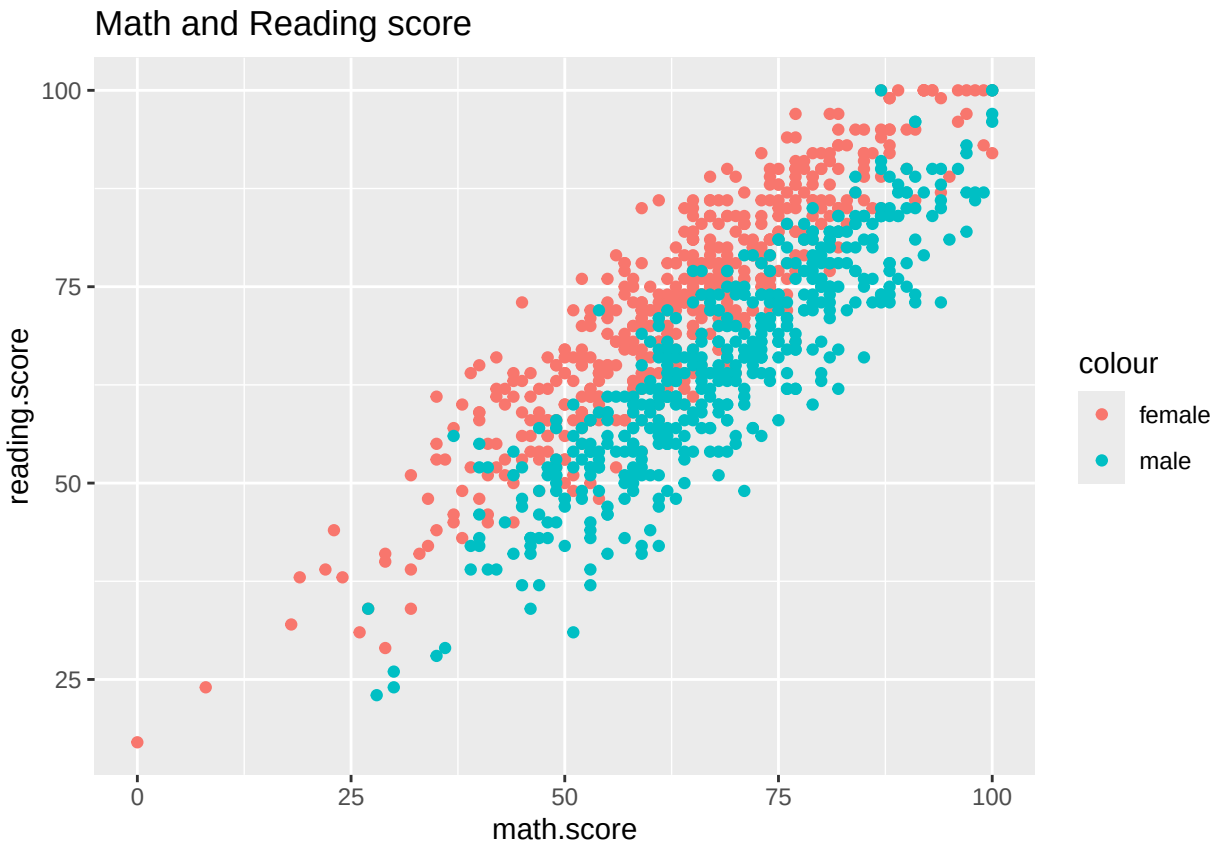
```
##
## associate's degree  bachelor's degree        high school    master's degree
##                222                118                196                 59
##       some college   some high school
##                226                179
```

```
table(data$lunch)
```

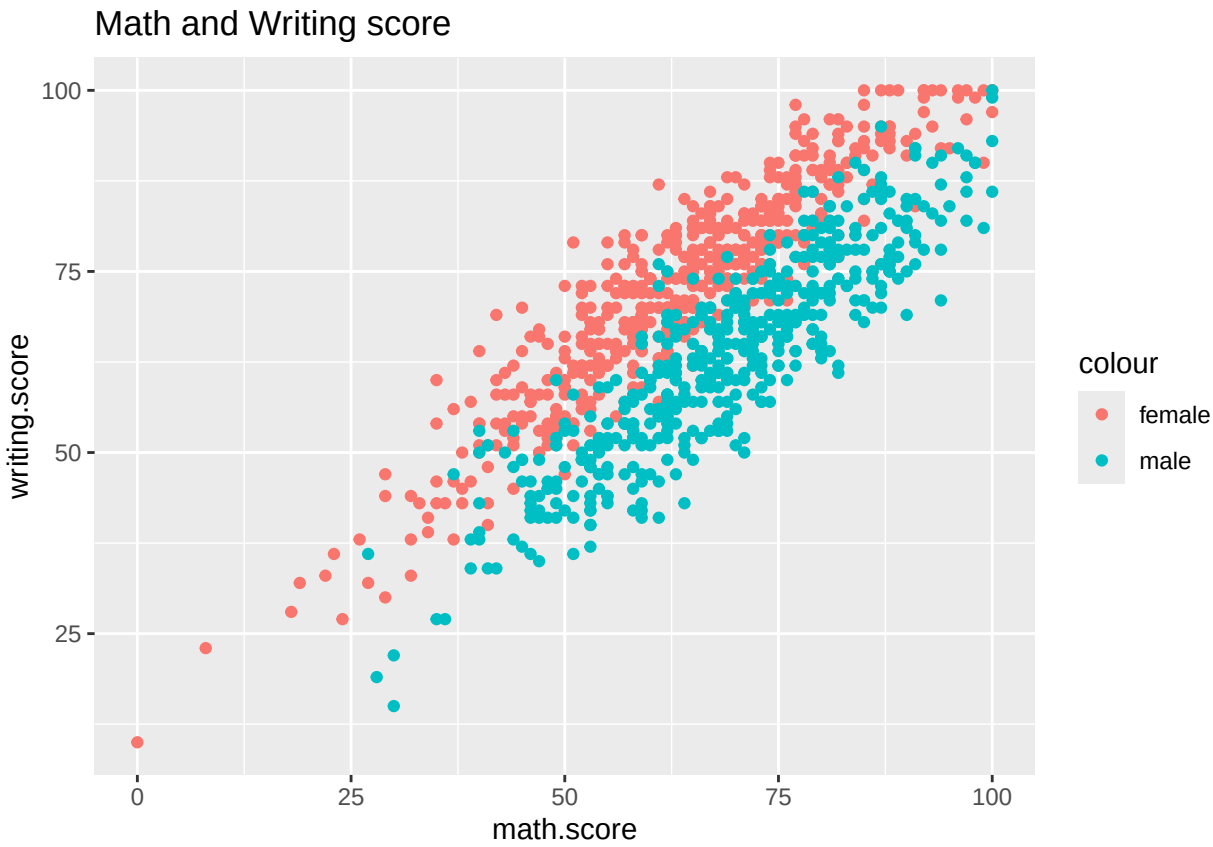
```
##
## free/reduced     standard
##          355          645
```

Comparing the Reading and Writing Scores again Math Scores

```
girl_data<-data%>%filter(gender=='female')
boy_data<-data%>%filter(gender=='male')
ggplot()+
  geom_point(girl_data,mapping=aes(`math.score`,`reading.score`,color='female'))+
  geom_point(boy_data,mapping=aes(`math.score`,`reading.score`,color='male'))+labs(title='Math and Readi
```

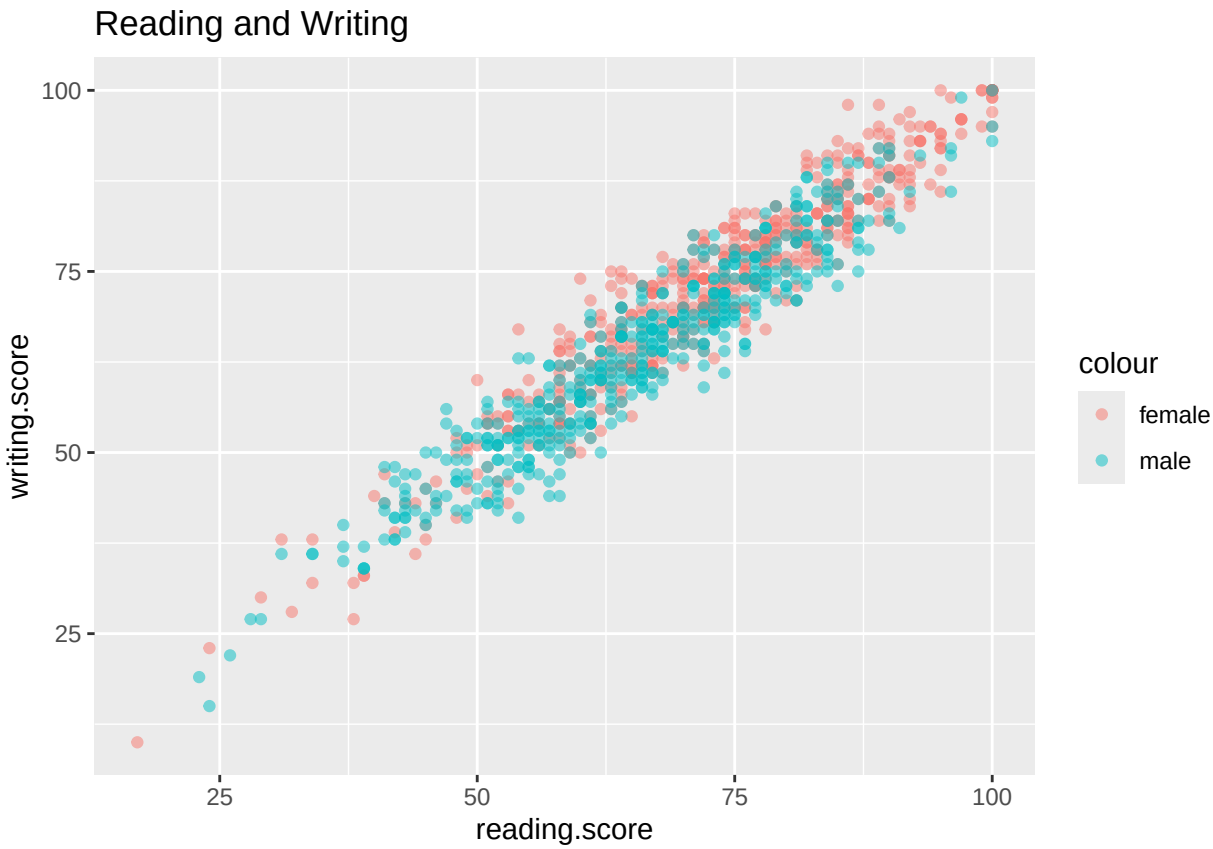



```
ggplot()+
  geom_point(girl_data,mapping=aes(`math.score`,`writing.score`,color='female'))+
  geom_point(boy_data,mapping = aes(`math.score`,`writing.score`,color='male'))+labs(title='Math and Wr
```

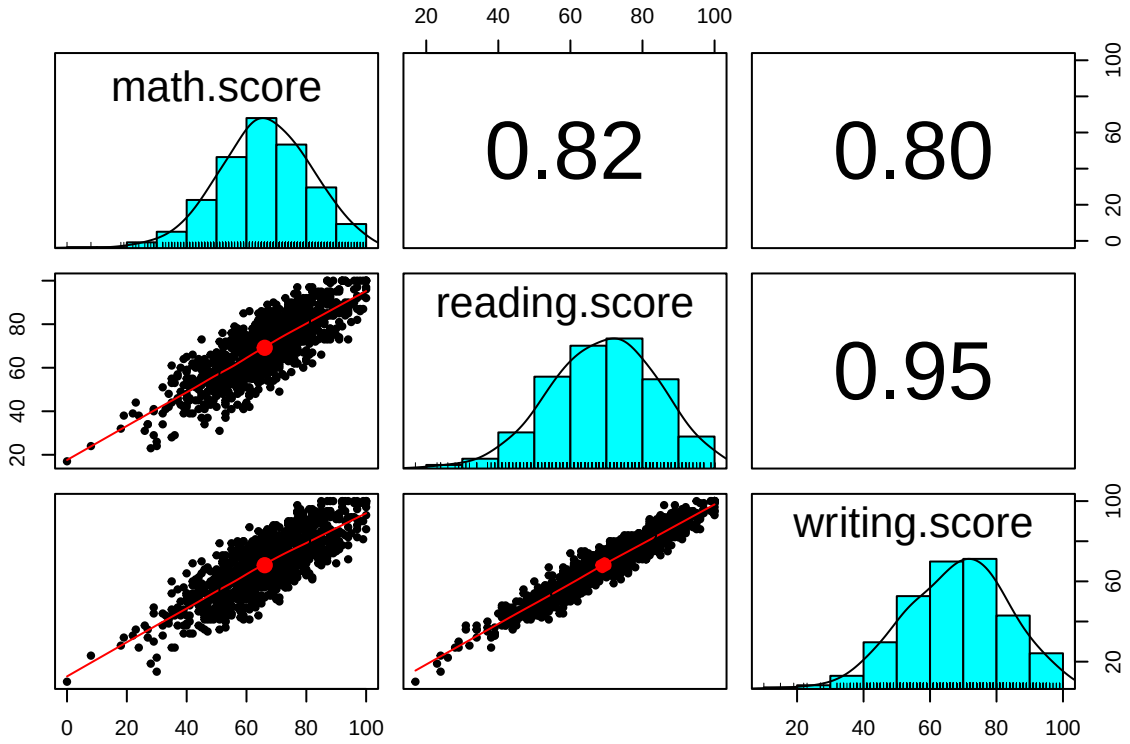


Now we comparing Reading and Writing Scores

```
ggplot()+
  geom_point(girl_data,mapping=aes(`reading.score`,`writing.score`,color='female'),alpha=1/2)+
  geom_point(boy_data,mapping = aes(`reading.score`,`writing.score`,color='male'),alpha=1/2)+labs(title=
```




```
pairs.panels(data[6:8])
```

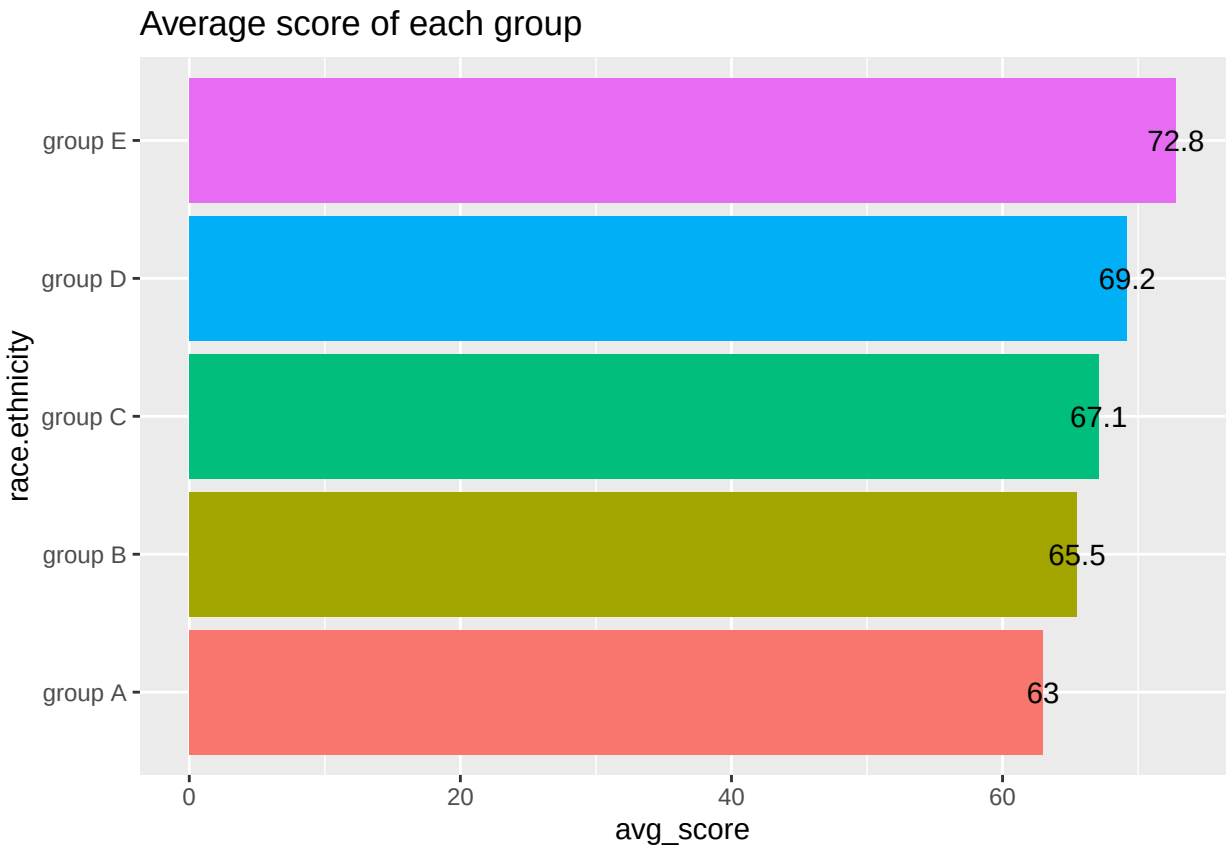


The plot above shows that females have a higher reading/wriitng score and males tend to have higher math scores

Averages Scores of Each Group

```
ndata%>%
group_by(race.ethnicity)%>%
summarize(avg_score=round(sum(mean)/n(),1))%>%
ggplot(aes(race.ethnicity,avg_score,fill=race.ethnicity))+geom_bar(stat='identity')+
geom_text(aes(label = avg_score))+coord_flip()+labs(title='Average score of each group')+guides(fill=F)
```

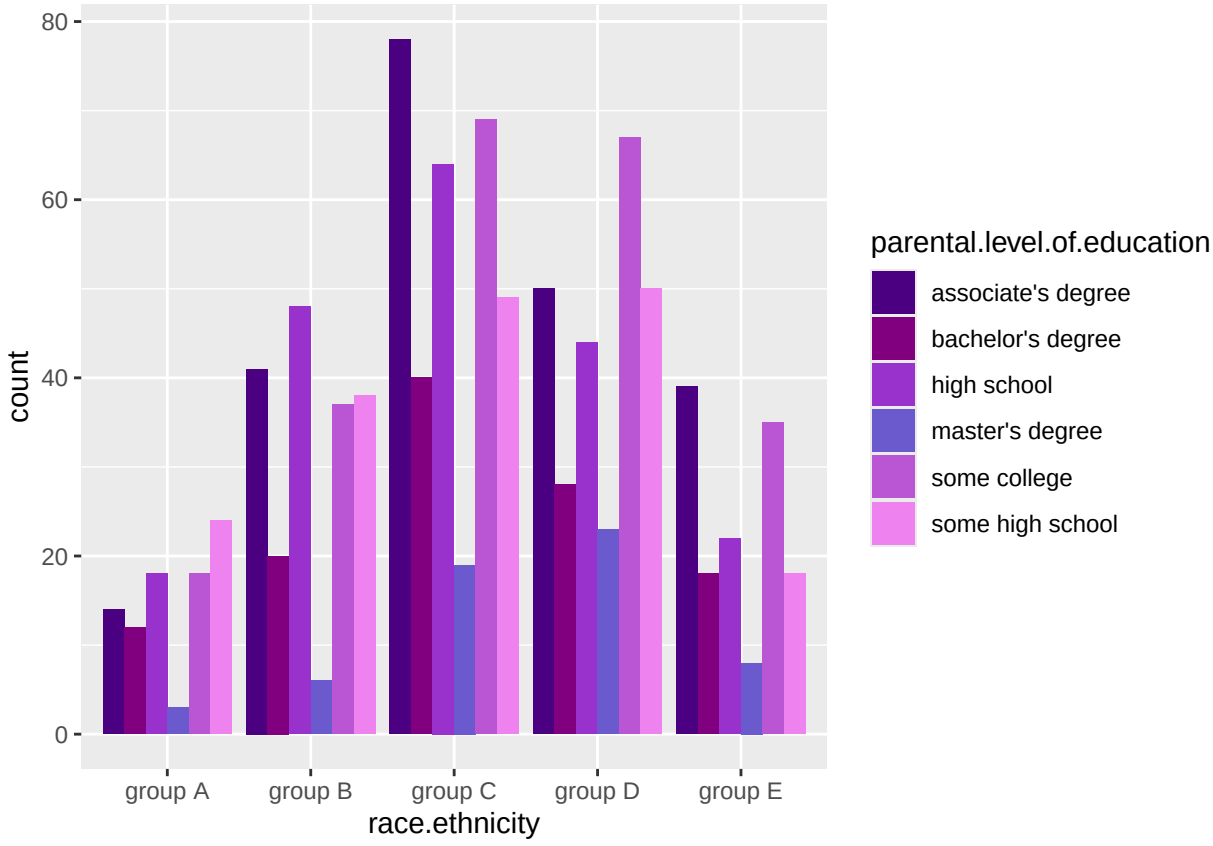
```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Level of Education of Parents

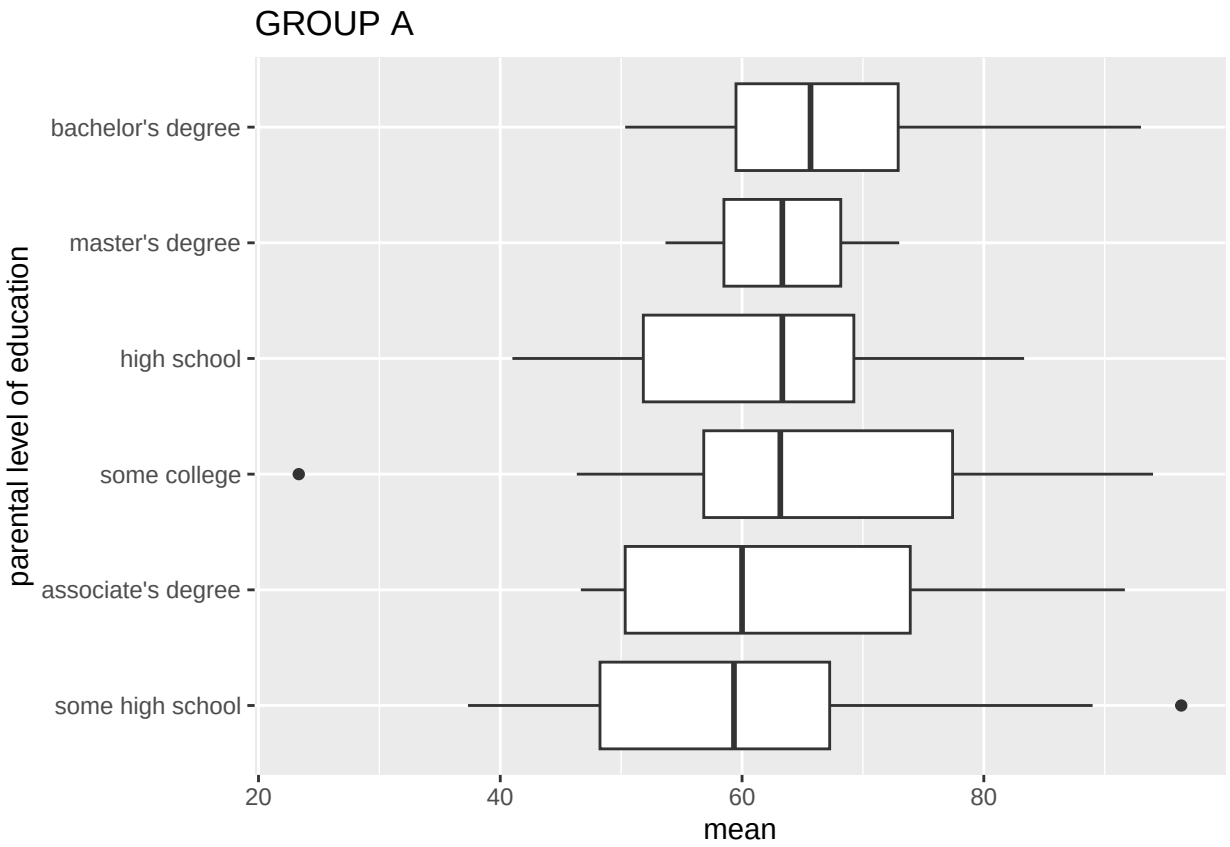
```
data%>%group_by(race.ethnicity,parental.level.of.education)%>%summarize(count=n())%>%
ggplot()+geom_col(aes(race.ethnicity,count,fill=parental.level.of.education),position='dodge')+scale_fi
```

```
## `summarise()` has grouped output by 'race.ethnicity'. You can override using
## the `.groups` argument.
```

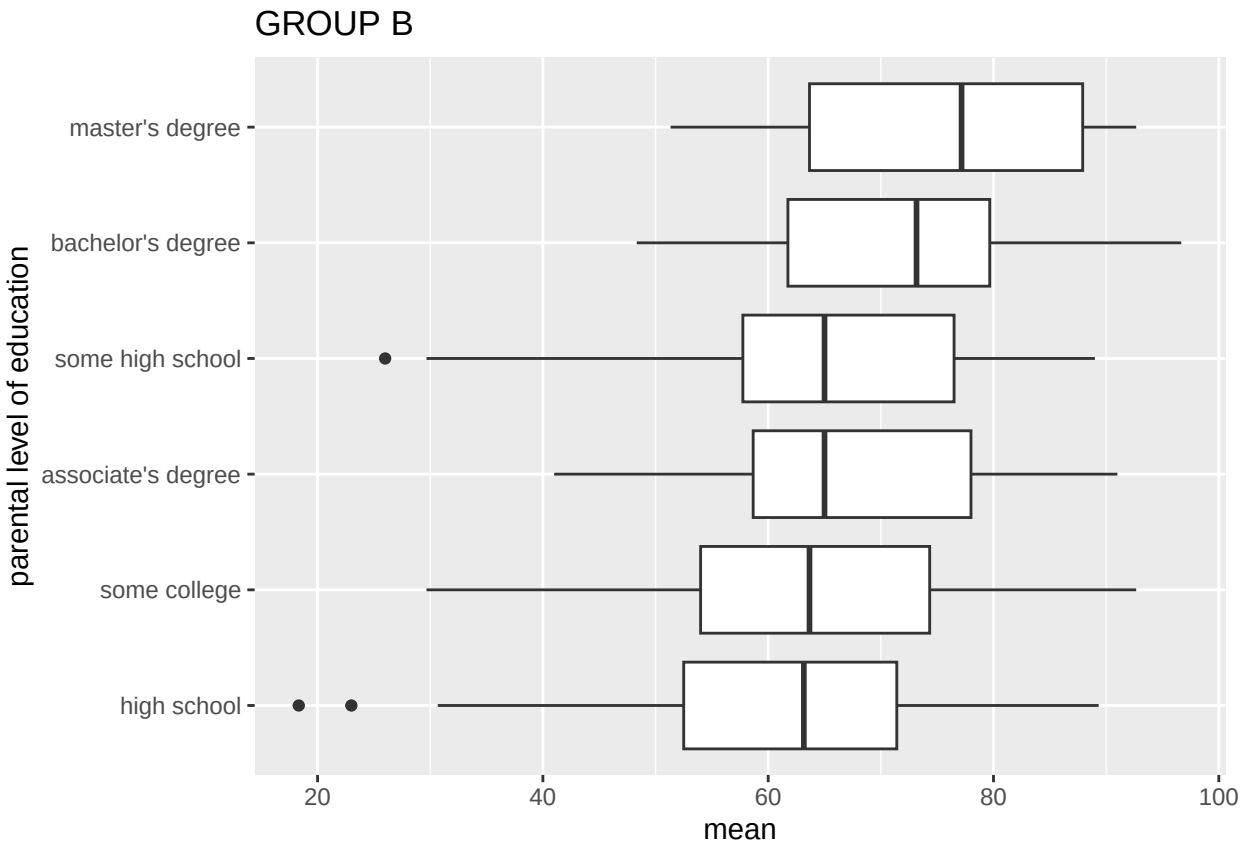


Education of Parents and Mean Score

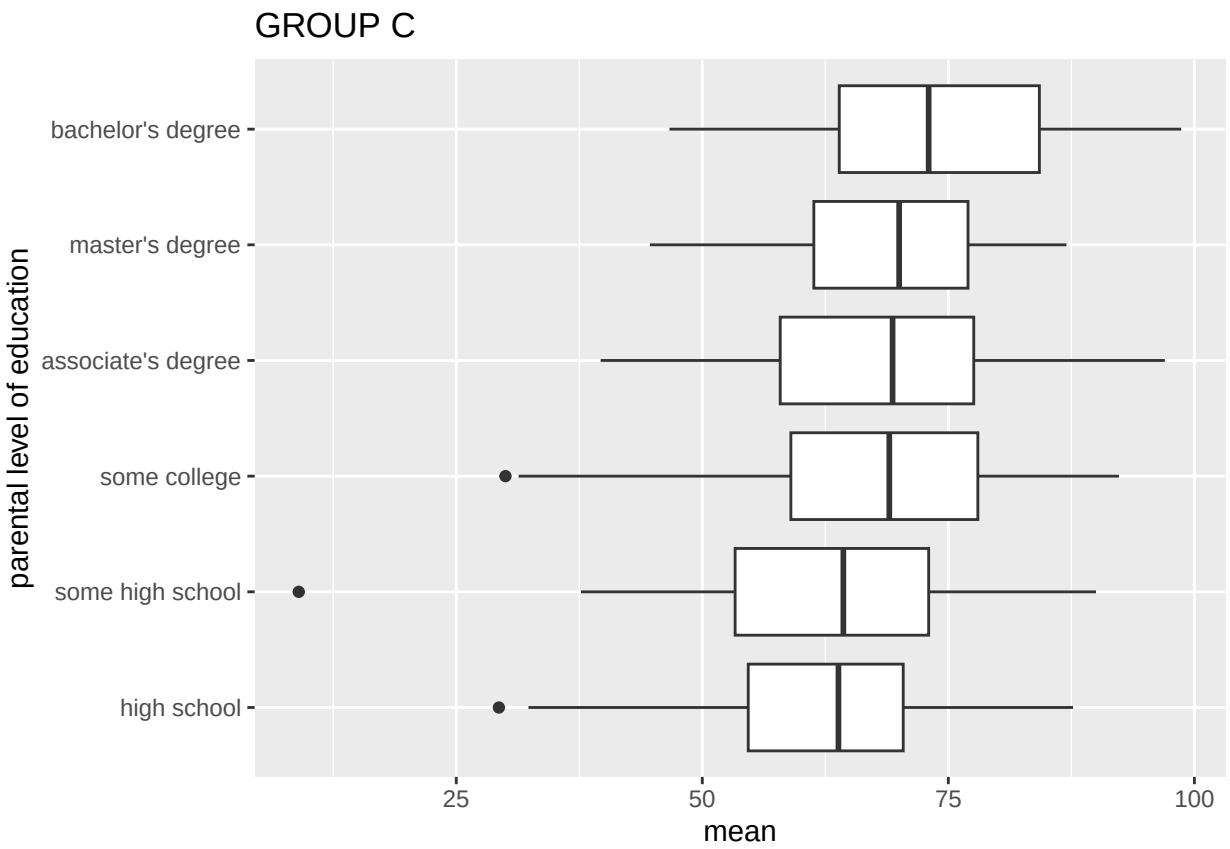
```
groupa<-ndata%>%filter(`race.ethnicity`=='group A')
groupb<-ndata%>%filter(`race.ethnicity`=='group B')
groupc<-ndata%>%filter(`race.ethnicity`=='group C')
groupd<-ndata%>%filter(`race.ethnicity`=='group D')
groupe<-ndata%>%filter(`race.ethnicity`=='group E')
ggplot(groupa)+geom_boxplot(mapping=aes(x=reorder(`parental.level.of.education`,mean,median),mean))+
  ggtitle('GROUP A')+xlab('parental level of education')+coord_flip()
```



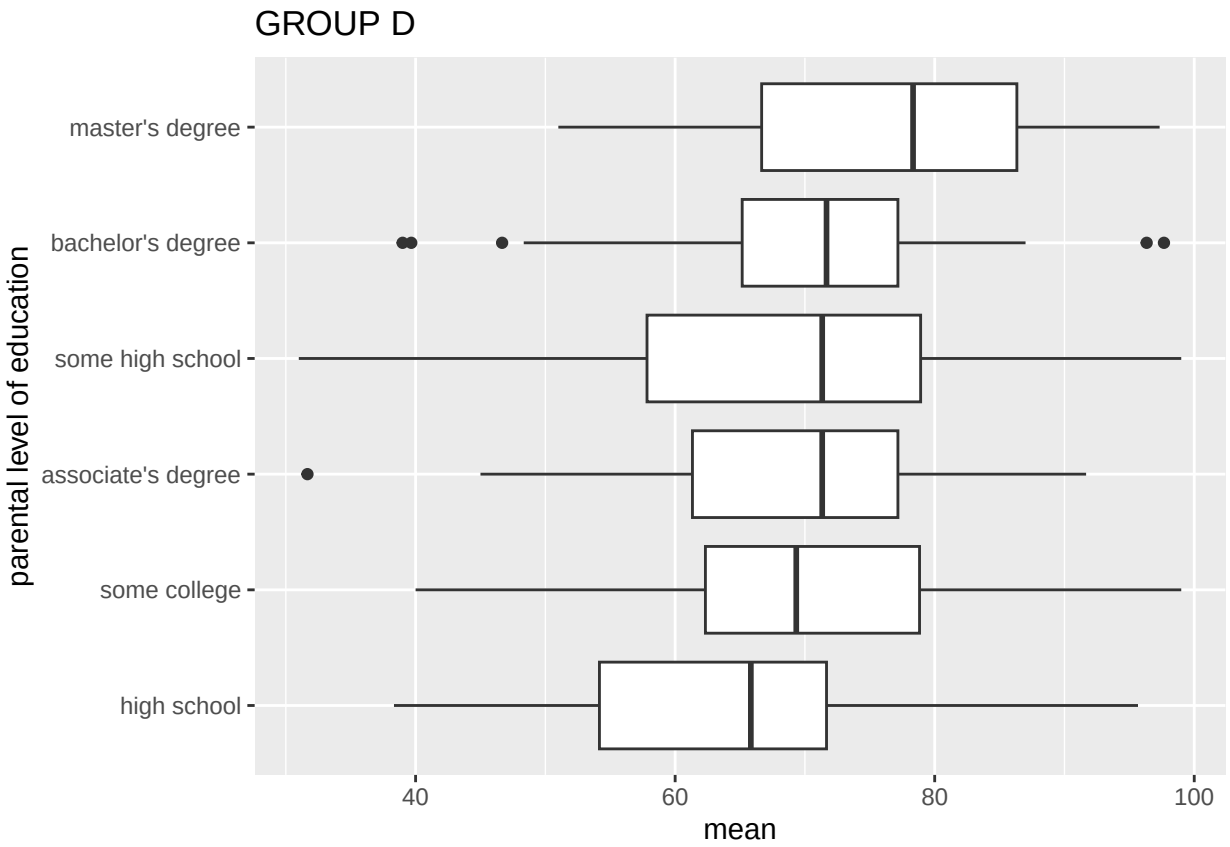

```
ggplot(groupb)+geom_boxplot(mapping=aes(x=reorder(`parental.level.of.education`,mean,median),mean))+
  ggtitle('GROUP B')+xlab('parental level of education')+coord_flip()
```



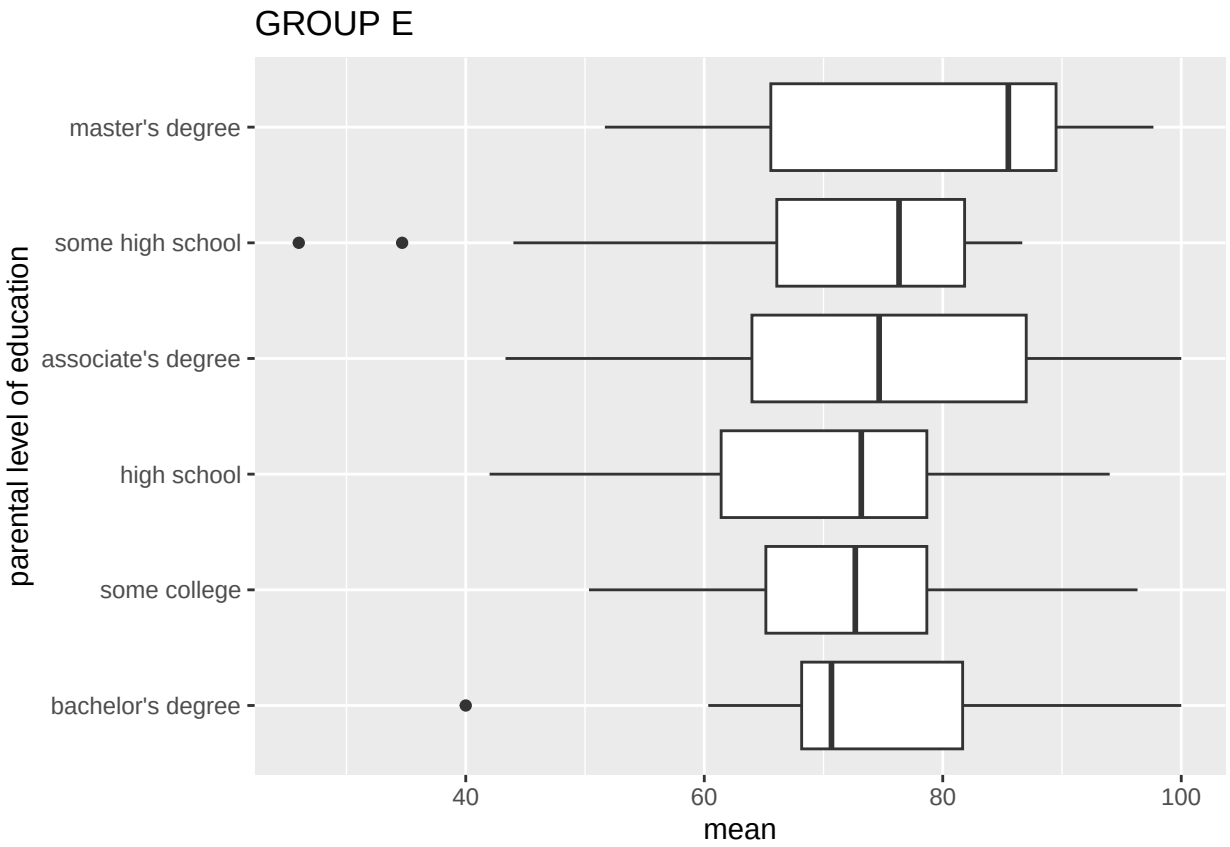

```
ggplot(groupc)+geom_boxplot(mapping=aes(x=reorder(`parental.level.of.education`,mean,median),mean))+
  ggtitle('GROUP C')+xlab('parental level of education')+coord_flip()
```




```
ggplot(groupd)+geom_boxplot(mapping=aes(x=reorder(`parental.level.of.education`,mean,median),mean))+
  ggtitle('GROUP D')+xlab('parental level of education')+coord_flip()
```



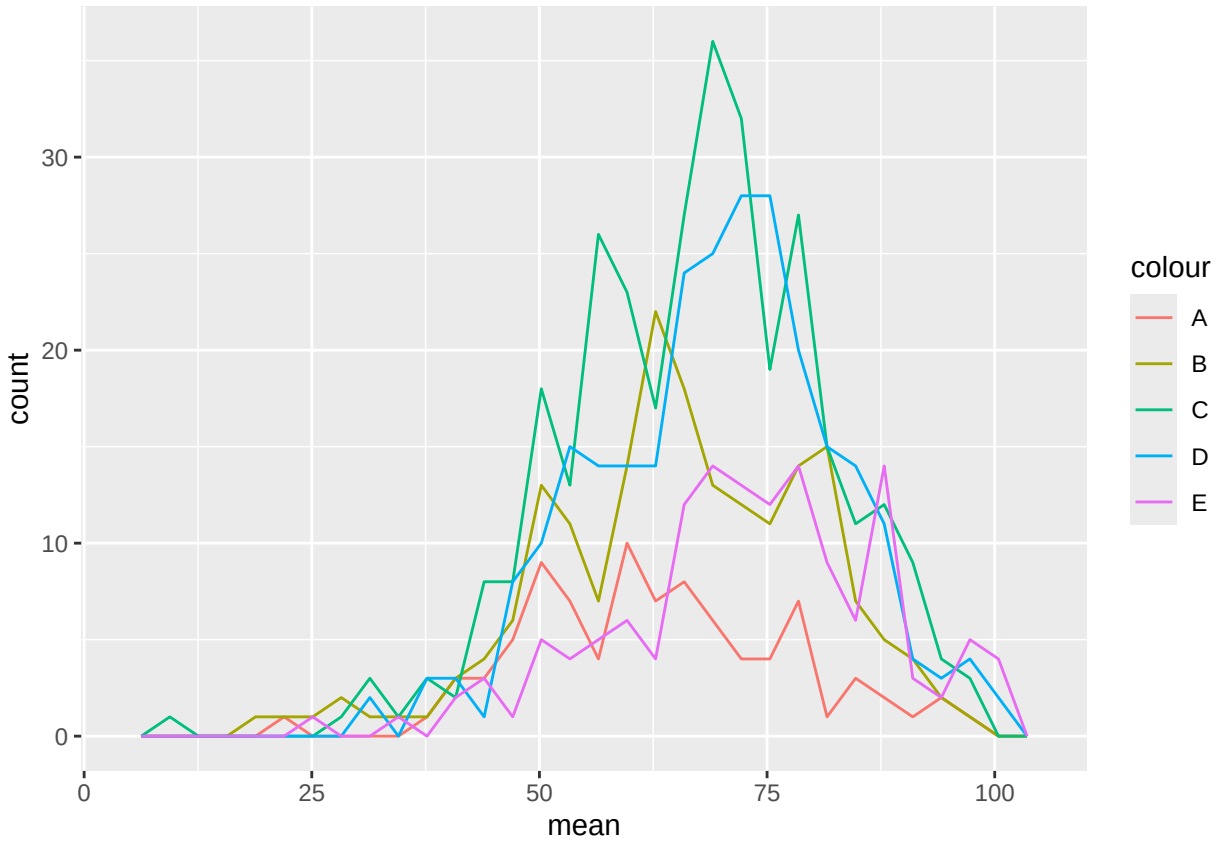

```
ggplot(groupe)+geom_boxplot(mapping=aes(x=reorder(`parental.level.of.education`,mean,median),mean))+
  ggtitle('GROUP E')+xlab('parental level of education')+coord_flip()
```



Mean Score of Each Plot

```
ggplot()+geom_freqpoly(groupa,mapping=aes(mean,color='A'))+geom_freqpoly(groupb,mapping=aes(mean,color=
  geom_freqpoly(groupc,mapping=aes(mean,color='C'))+geom_freqpoly(groupd,mapping=aes(mean,color='D'))+
  geom_freqpoly(groupe,mapping=aes(mean,color='E'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
group_A<-groupa%>%rename(mean_A=mean)%>%select(mean_A)%>%arrange(desc(`mean_A`))%>%head(89)%>%round(dig
group_B<-groupb%>%rename(mean_B=mean)%>%select(mean_B)%>%arrange(desc(`mean_B`))%>%head(89)%>%round(dig
group_C<-groupc%>%rename(mean_C=mean)%>%select(mean_C)%>%arrange(desc(`mean_C`))%>%head(89)%>%round(dig
group_D<-groupd%>%rename(mean_D=mean)%>%select(mean_D)%>%arrange(desc(`mean_D`))%>%head(89)%>%round(dig
group_E<-groupe%>%rename(mean_E=mean)%>%select(mean_E)%>%arrange(desc(`mean_E`))%>%head(89)%>%round(dig
cbi<-cbind(group_A,group_B,group_C,group_D,group_E)
cbi%>%head()
```

```
##   mean_A mean_B mean_C mean_D mean_E
## 1   96.3   96.7   98.7   99.0  100.0
## 2   94.0   92.7   97.0   99.0  100.0
## 3   93.0   92.7   97.0   98.7  100.0
## 4   91.7   91.7   94.0   97.7   99.7
## 5   89.3   91.0   93.7   97.3   97.7
## 6   89.0   90.7   93.3   96.3   97.7
```

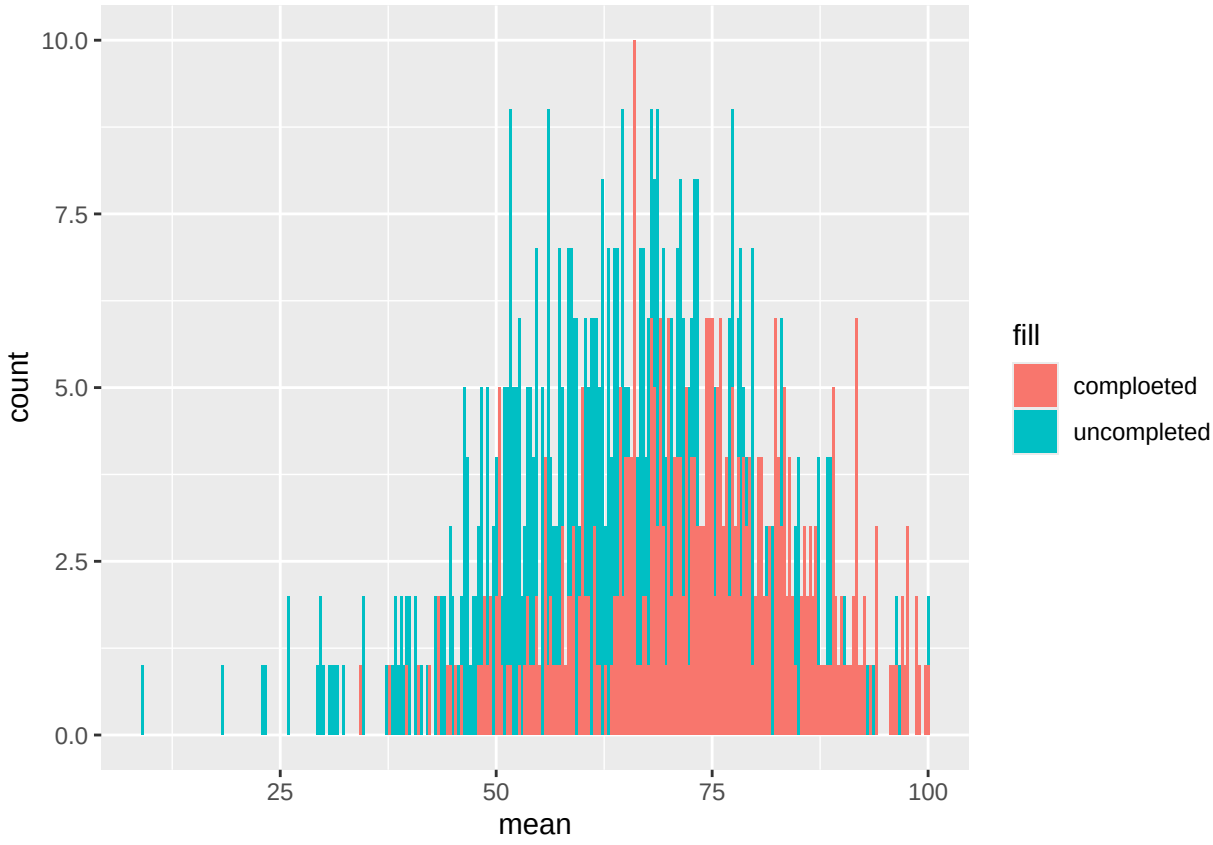
```
summary(cbi)
```

```
##      mean_A          mean_B          mean_C          mean_D
##  Min.   :23.30   Min.   :66.00   Min.   :75.70   Min.   :75.00
##  1st Qu.:52.00   1st Qu.:72.00   1st Qu.:78.00   1st Qu.:78.30
##  Median :61.30   Median :77.30   Median :82.30   Median :82.00
##  Mean   :62.99   Mean   :77.85   Mean   :83.28   Mean   :83.08
##  3rd Qu.:73.00   3rd Qu.:83.00   3rd Qu.:87.30   3rd Qu.:86.30
##  Max.   :96.30   Max.   :96.70   Max.   :98.70   Max.   :99.00
```

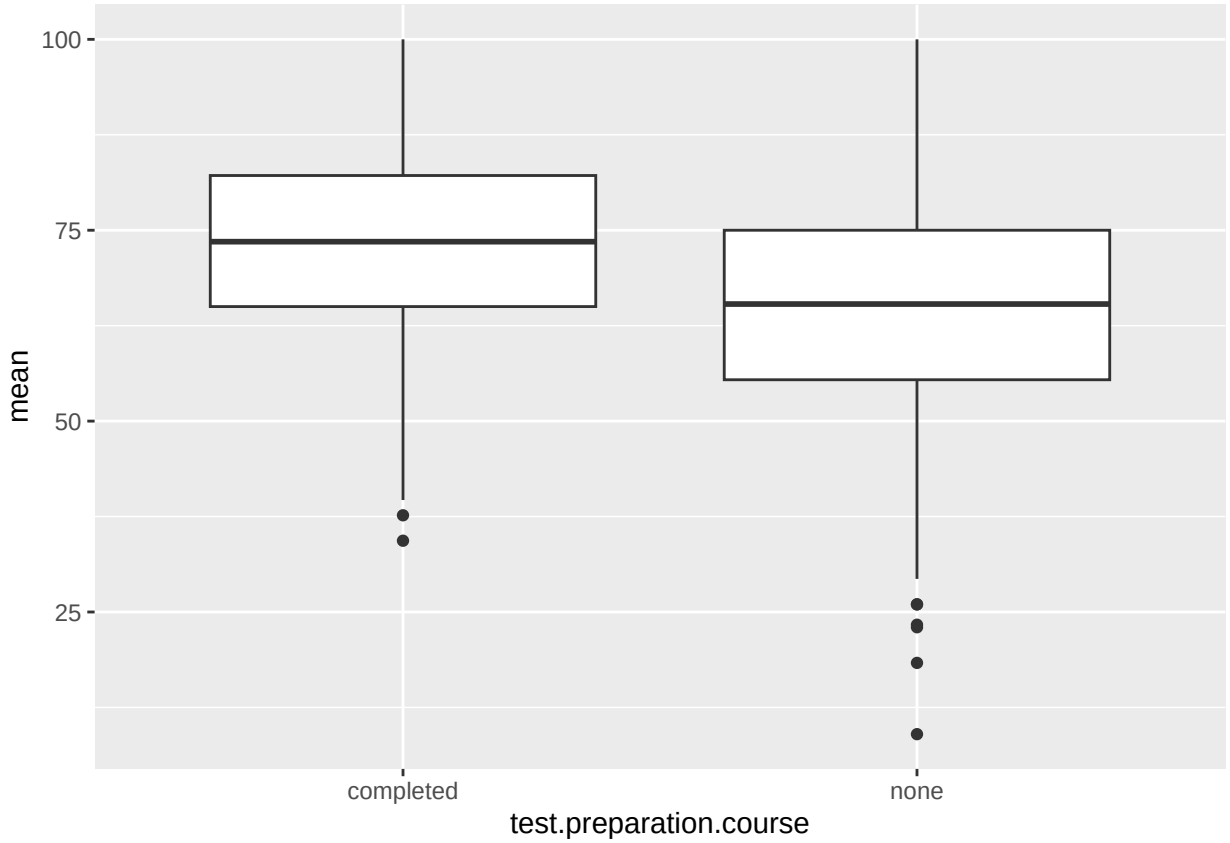
```
##      mean_E
##  Min.   : 68.70
##  1st Qu.: 74.70
##  Median : 79.70
##  Mean   : 81.37
##  3rd Qu.: 87.30
##  Max.   :100.00
```

Students who completed test prep courses have higher mean scores

```
#(3)observe test preparation course and mean
uncomplete_prep<-ndata%>%filter(`test.preparation.course`=='none')
complete_prep<-ndata%>%filter(`test.preparation.course`=='completed')
prep1<-ggplot()+geom_bar(uncomplete_prep,mapping = aes(x=mean,fill='uncompleted'))+
  geom_bar(complete_prep,mapping=aes(x=mean,fill='comploeted'))
prep2<-ggplot()+geom_boxplot(ndata,mapping=aes(`test.preparation.course`,mean))
prep1
```




```
prep2
```


100
75
50
25
mean
completed
none
test.preparation.course