R 语言数据可视化

Cooper

目录

1 实验介绍

实验目的：研究什么因素会影响学生的成绩

数据来源：网页下载http://roycekimmons.com/tools/generated_data/exams

数据说明：该数据集包括三次考试的分数以及各种对其产生影响的个人、社会因素。

2 主要内容

2.1 STEP 1 载入所有需要使用的包

```
library(pacman)
```

```
## Warning: 程辑包'pacman'是用R版本4.2.3 来建造的
```

```
p_load(dplyr, tidyverse, ggplot2, ggthemes, patchwork, corrplot, ggsci, rticles)
```

dplyr # 数据处理

tidyverse&ggplot2 # 数据可视化作图

ggthemes # 绘图的主题包 theme_stata() 和 scale_color_economist()

ggsci # 绘图的颜色包 scale_color_jco() 和 scale_color_npg()

patchwork # 用于拼接图片

corrplot # 绘制相关性图

rticles #Rmarkdown 主题包，使用 CTeX documents，输出中文 PDF

2.2 STEP 2 数据预处理

2.2.1 读入数据

```
data.st <- as.data.frame(read.csv('student-exams.csv'))
head(data.st,n=10)
```

```
   gender race.ethnicity parental.level.of.education test.preparation.course
1  female        group D            some high school                completed
2    male        group D                some college                     none
3    male        group C           bachelor's degree                     none
4  female        group C                 high school                completed
```

```
5    male        group C          associate's degree                 none
6    male        group D            some high school                 none
7  female        group E                some college                 none
8  female        group B                 high school                 none
9    male        group E                 high school                 none
10   male        group B            some high school             completed
   math.score reading.score writing.score
1          74            79            81
2          54            51            48
3          82            88            85
4          55            70            74
5          48            52            39
6          71            72            63
7          65            70            66
8          54            64            61
9          89            81            72
10         65            62            66
```

2.2.2 查看基本信息

```
glimpse(data.st)# 查看基本信息
```

```
Rows: 1,000
Columns: 7
$ gender                      <chr> "female", "male", "male", "female", "male"~
$ race.ethnicity              <chr> "group D", "group D", "group C", "group C"~
$ parental.level.of.education <chr> "some high school", "some college", "bache~
$ test.preparation.course     <chr> "completed", "none", "none", "completed", ~
$ math.score                  <int> 74, 54, 82, 55, 48, 71, 65, 54, 89, 65, 69~
$ reading.score               <int> 79, 51, 88, 70, 52, 72, 70, 64, 81, 62, 78~
$ writing.score               <int> 81, 48, 85, 74, 39, 63, 66, 61, 72, 66, 72~
```

一共 7 列，1000 行数据即 1000 个学生。

变量名解释:

变量名	含义
gender	性别
race.ethnicity	学生分组
parental.leval.of.education	父母的受教育水平
test.preparation.course	学生是否完成预科班课程
math.score	数学分数

变量名	含义
reading.score	阅读分数
writing.score	写作分数

该数据集显示了学生在数学、阅读和写作方面的表现，以及性别和其他社会因素。

2.2.3 为了方便理解和操作重命名列名

```
names_columns <- c('Gender','Race','Parent_Education','Test_Prep','M_Score','R_Score','W_Score')
colnames(data.st) <- names_columns
colnames(data.st)
```

```
[1] "Gender"           "Race"             "Parent_Education" "Test_Prep"
[5] "M_Score"          "R_Score"          "W_Score"
```

2.2.4 检查数据是否有异常值

```
summary(data.st)
##     Gender              Race           Parent_Education    Test_Prep
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     M_Score          R_Score          W_Score
##  Min.   : 18.00   Min.   : 23.00   Min.   : 20.00
##  1st Qu.: 56.00   1st Qu.: 60.00   1st Qu.: 58.00
##  Median : 68.00   Median : 70.00   Median : 69.00
##  Mean   : 66.99   Mean   : 69.78   Mean   : 68.67
##  3rd Qu.: 78.00   3rd Qu.: 81.00   3rd Qu.: 80.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
table(data.st$Gender)
##
## female   male
##    499    501
table(data.st$Race)
##
## group A group B group C group D group E
```

```
##      73     205     293     284     145
table(data.st$Parent_Education)
##
## associate's degree  bachelor's degree        high school    master's degree
##                171                144                195                 68
##       some college   some high school
##                242                180
table(data.st$Test_Prep)
##
## completed      none
##       341       659
```

三项成绩中阅读成绩平均值最高，

可以看出数据中没有 NA 或者缺失值。

2.2.5 添加新的一列平均成绩进入数据

平均成绩更加适合检验一个学生成绩表现

```
data.st <- data.st %>% mutate(Avg_Score = (M_Score+R_Score+W_Score)/3)
# 增加平均成绩这一列
head(data.st,10)
```

```
   Gender    Race   Parent_Education Test_Prep M_Score R_Score W_Score
1  female group D   some high school completed      74      79      81
2    male group D       some college      none      54      51      48
3    male group C  bachelor's degree      none      82      88      85
4  female group C        high school completed      55      70      74
5    male group C associate's degree      none      48      52      39
6    male group D   some high school      none      71      72      63
7  female group E       some college      none      65      70      66
8  female group B        high school      none      54      64      61
9    male group E        high school      none      89      81      72
10   male group B   some high school completed      65      62      66
   Avg_Score
1   78.00000
2   51.00000
3   85.00000
4   66.33333
5   46.33333
6   68.66667
7   67.00000
```

```
8   59.66667
9   80.66667
10  64.33333
```

以上展示了增加了平均成绩这一列之后的数据的前十行。

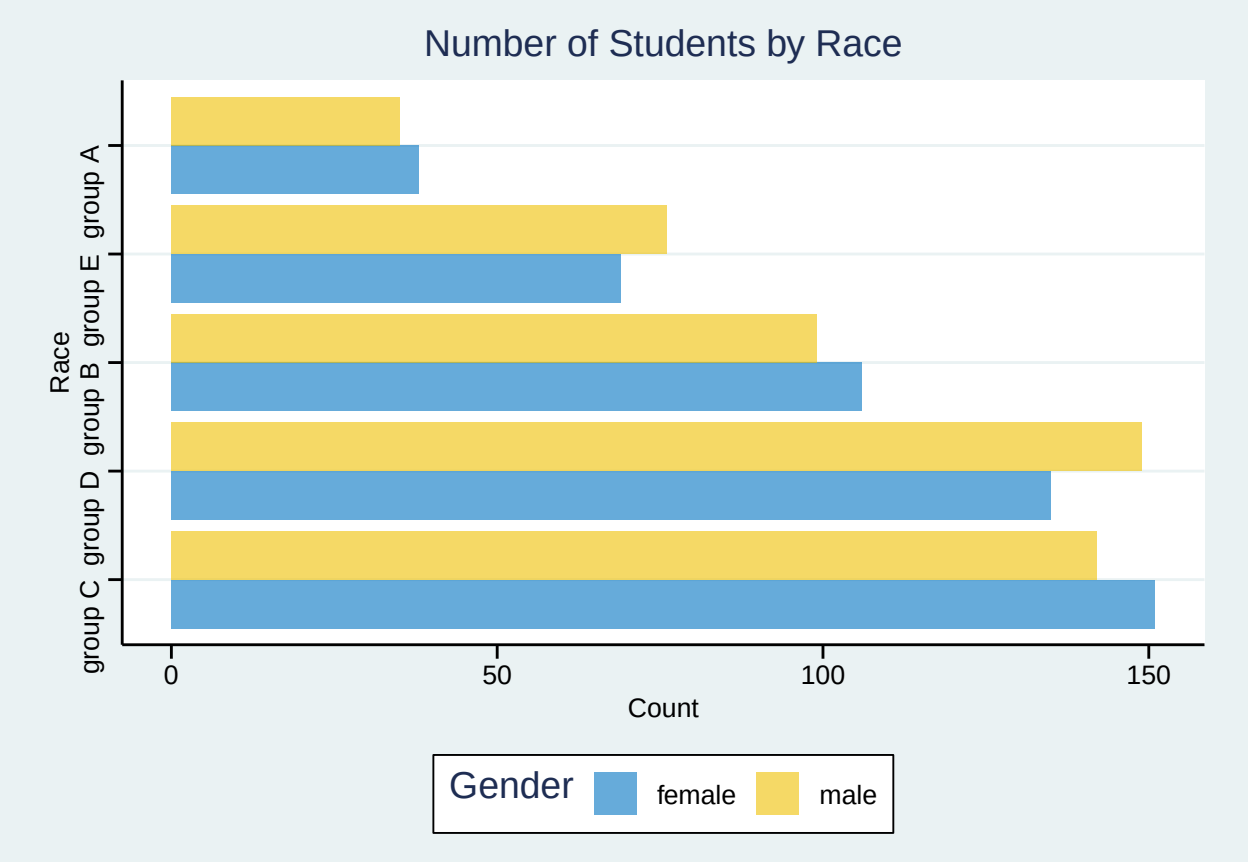
2.3 STEP 3 数据可视化

数据可视化的主要目的是为了方便进行数据分析，以下我会试用 ggplot2 这个包绘制不同测试分数与学生性别、父母教育程度、不同分组关系的图。

2.3.1 不同组中男女生的数量（直方图）

```
data.st %>% group_by(Race, Gender) %>% summarise(n = n()) %>%
  ggplot(aes(x = reorder(Race, -n),
             y = n, fill = Gender)) + #reorder() 按照男女生的数量排序从低到高
  geom_col(position = 'dodge')  + coord_flip() + #coord_flip 让整个图横过来
  labs(x = "Race",y='Count') +
  ggtitle('Number of Students by Race') +
  scale_color_jco(alpha=0.6) +
  scale_fill_jco(alpha=0.6,labels = c("female", "male")) +
  theme_stata()
```

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

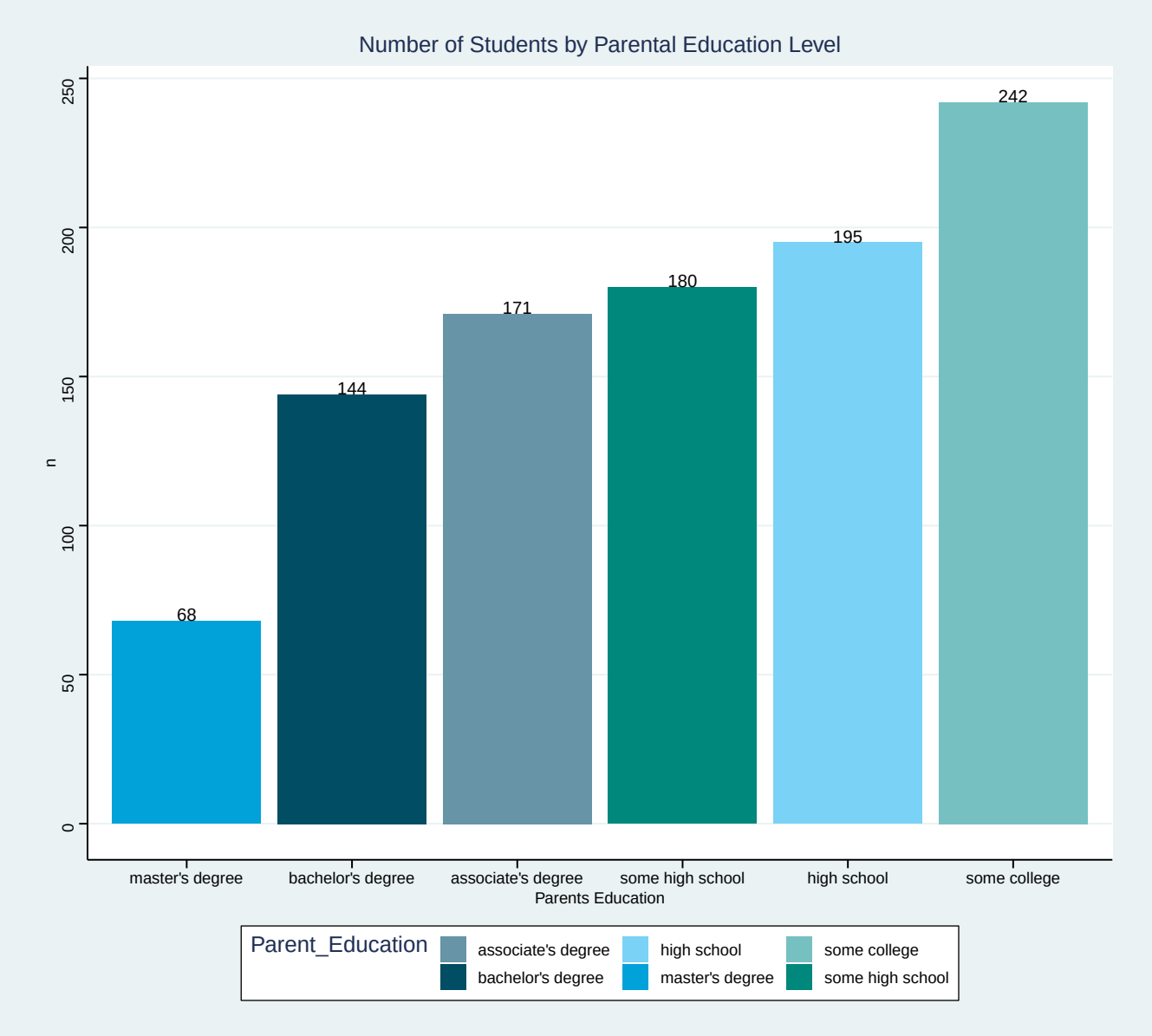


在这通过直方图根据性别显示有多少学生来自不同的社区，

可以明显看出 A 组男女生数量都远小于其他四组，C 组女生最多而 D 组男生最多。

2.3.2 父母受教育程度的统计（直方图）

```
data.st %>% group_by(Parent_Education) %>% summarise(n = n()) %>%
  ggplot(aes(x = reorder(Parent_Education, n), y = n, fill = Parent_Education)) +
  geom_col()  + geom_text(aes(label = n), vjust = 0.01) +
  labs(x = "Parents Education") +
  ggtitle('Number of Students by Parental Education Level') +
  scale_color_economist()+scale_fill_economist() +
  theme_stata()
```

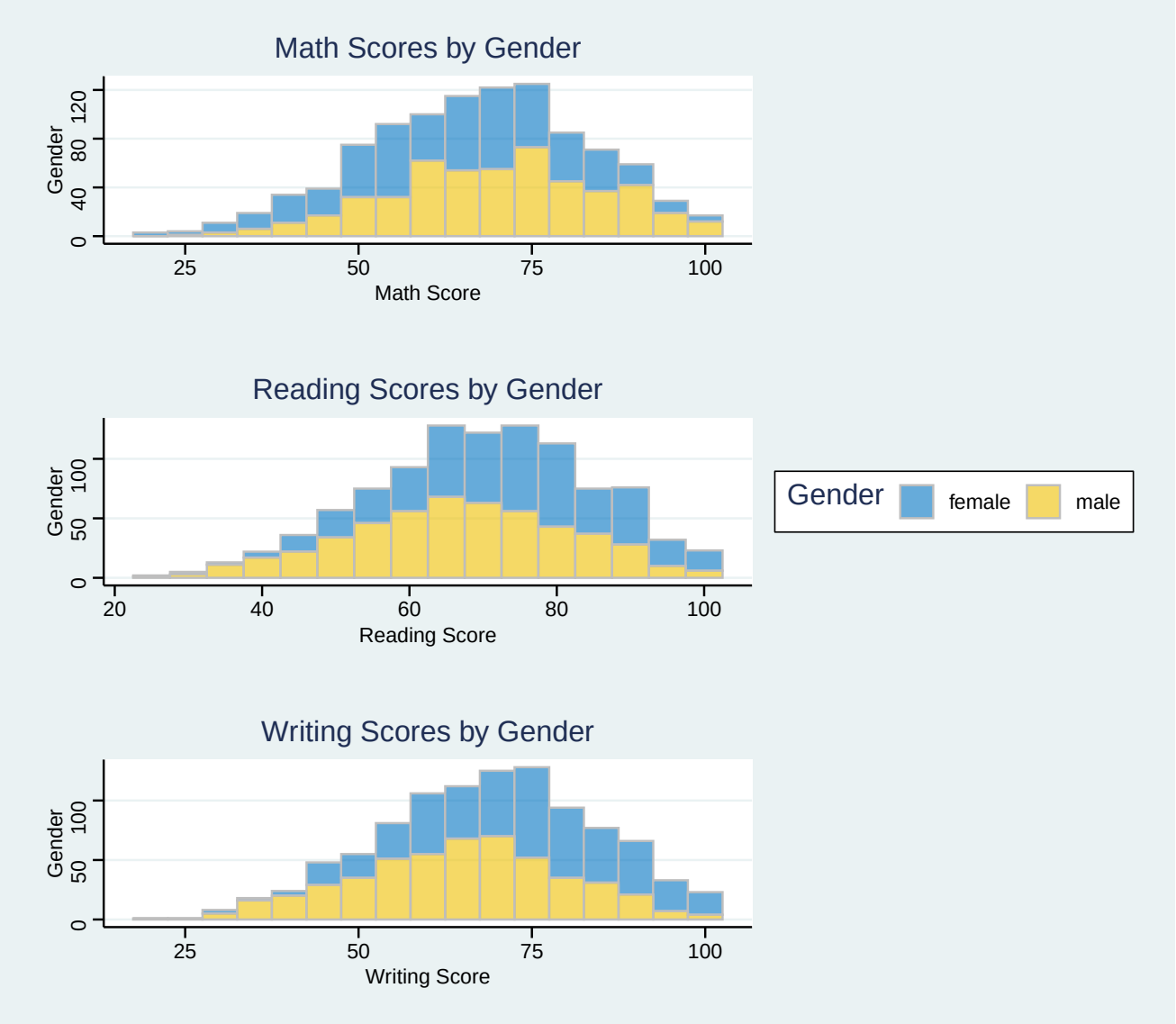


显示了不同受教育程度的父母，他们中的大多数都为高中和一些学院毕业，硕士毕业的父母较少。

2.3.3 按性别划分考试分数（堆积柱状图）

```
# Math scores by Gender plot

p1 <- ggplot(data.st, aes(M_Score)) +
    geom_histogram(binwidth=5, color="gray", aes(fill=Gender)) +
    xlab("Math Score") + ylab("Gender") +
  ggtitle("Math Scores by Gender") + # 坐标轴和标题
  scale_color_jco(alpha=0.6)+scale_fill_jco(alpha=0.6) +
  theme_stata()
```

```r
# Reading scores by Gender

p2 <- ggplot(data.st, aes(R_Score)) +
    geom_histogram(binwidth=5, color="gray", aes(fill=Gender)) +
    xlab("Reading Score") + ylab("Gender") +
  ggtitle("Reading Scores by Gender") +
  scale_color_jco(alpha=0.6)+scale_fill_jco(alpha=0.6) +
  theme_stata()

# Writing scores by Gender plot

p3 <- ggplot(data.st, aes(W_Score)) +
    geom_histogram(binwidth=5, color="gray", aes(fill=Gender)) +
    xlab("Writing Score") + ylab("Gender") +
  ggtitle("Writing Scores by Gender") +
  scale_color_jco(alpha=0.6)+scale_fill_jco(alpha=0.6) +
  theme_stata()

#patchwork 用于拼接

p1/p2/p3 +
  patchwork::plot_layout(guides = "collect") # 只出现一个图例
```



按性别划分考试分数，以确定每个性别是否有不同的分数倾向，

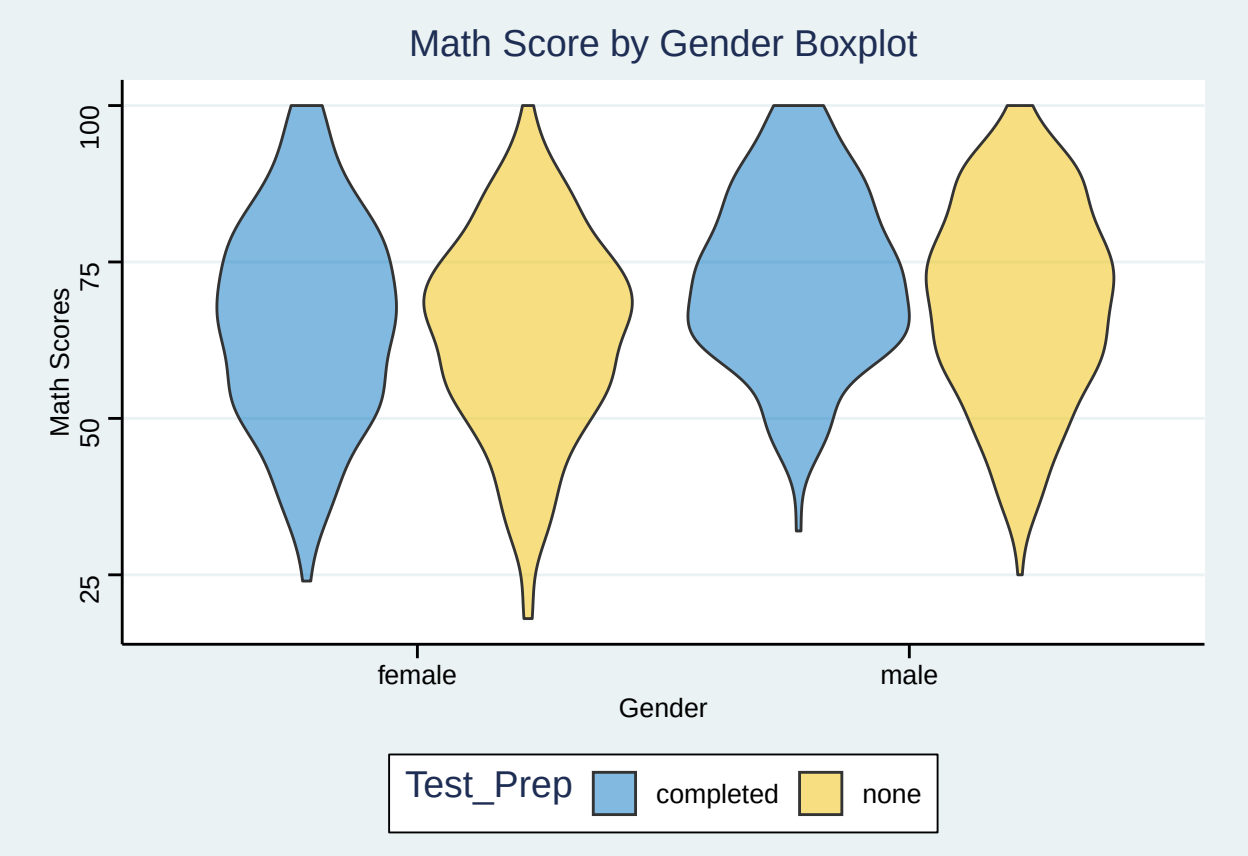
可以明显看出男生的数学成绩明显高于女生，

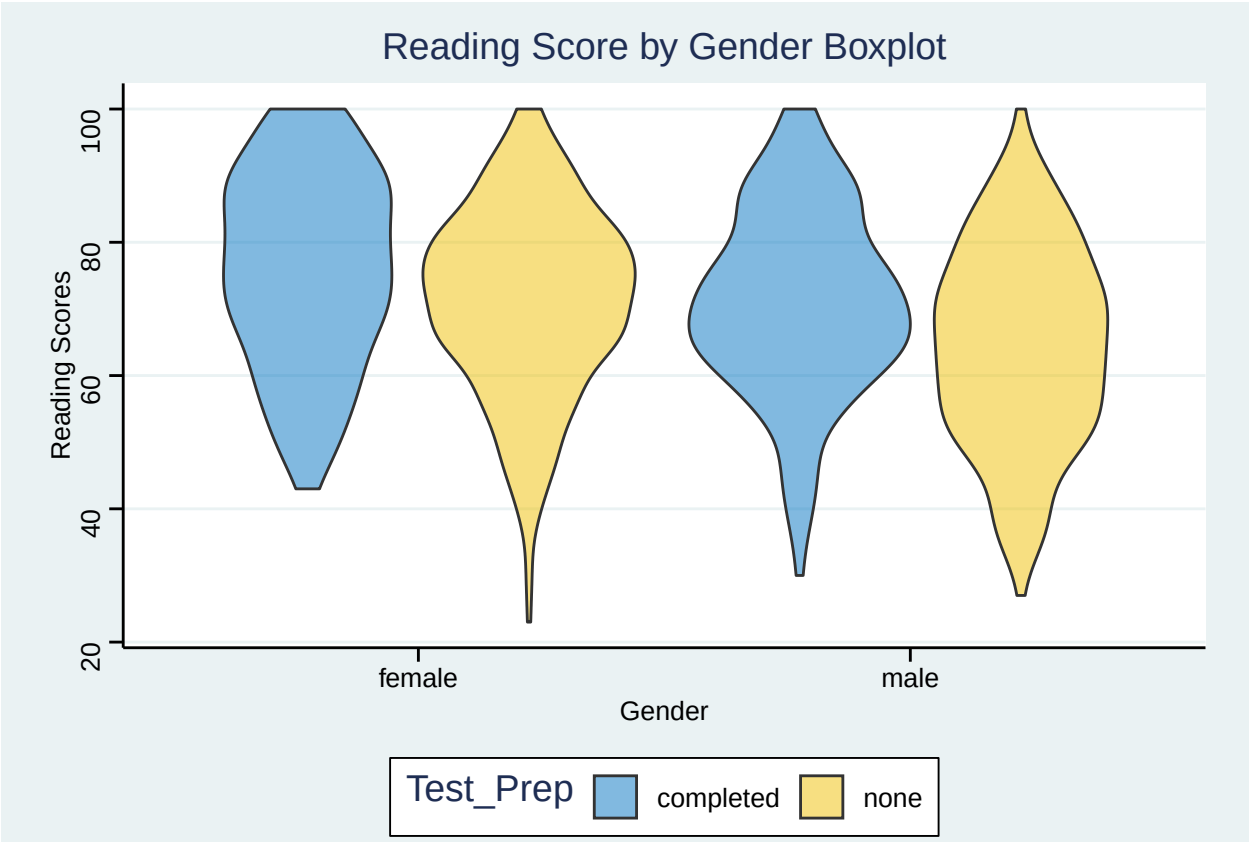
同时女生的阅读和写作成绩也高于男生。

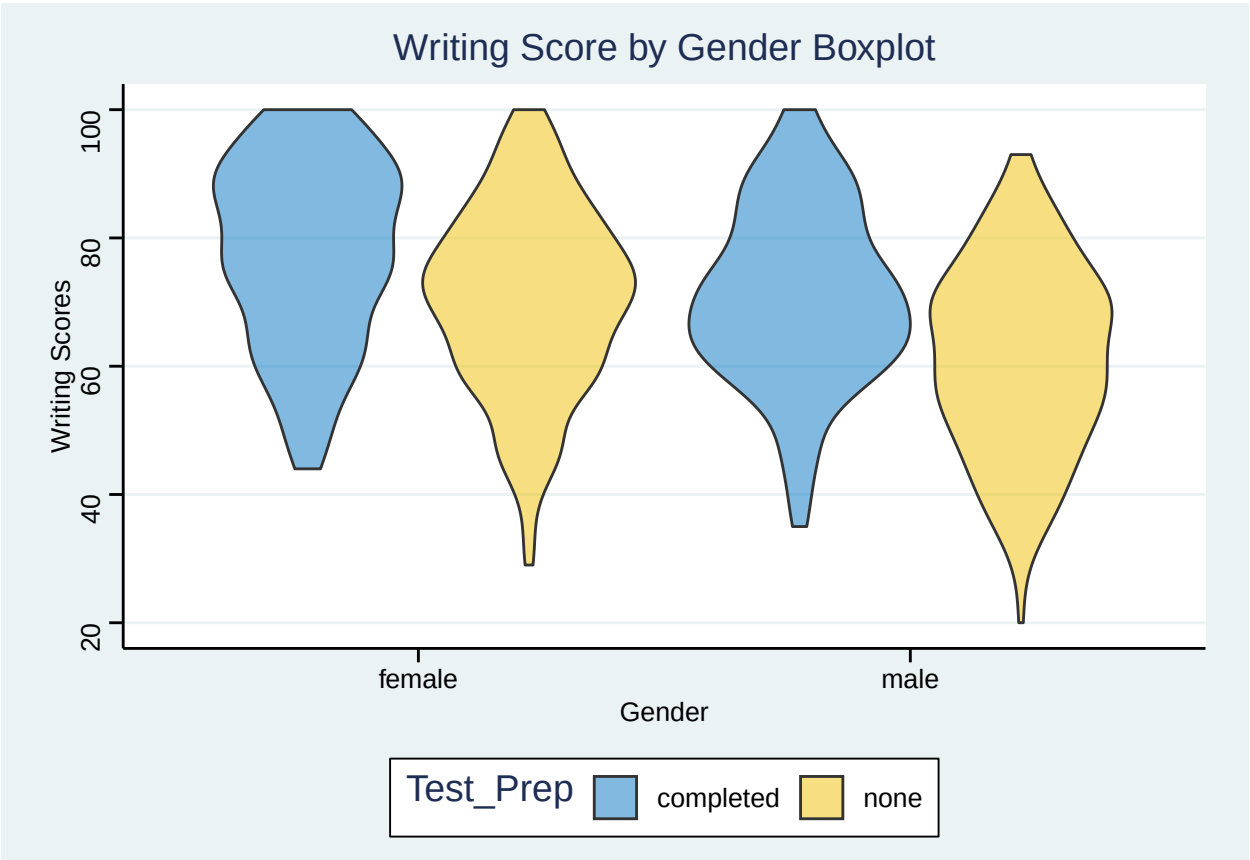
2.3.4 学生成绩与性别和是否做完成预科班课程（小提琴图）

```
# 数学成绩
b1 <- ggplot(data.st, aes(Gender, M_Score, fill = Test_Prep)) +
    geom_violin() + # 绘制小提琴图
    ggtitle("Math Score by Gender Boxplot") +
    xlab("Gender") + ylab("Math Scores") + # 坐标轴和标题
    scale_color_jco(alpha=0.5)+scale_fill_jco(alpha=0.5) + theme_stata()
```

```
# 阅读成绩
b2 <- ggplot(data.st, aes(Gender, R_Score, fill = Test_Prep)) +
    geom_violin() +
    ggtitle("Reading Score by Gender Boxplot") +
    xlab("Gender") + ylab("Reading Scores") +
    scale_color_jco(alpha=0.5)+scale_fill_jco(alpha=0.5) +
    theme_stata()

# 写作分数
b3 <- ggplot(data.st, aes(Gender, W_Score, fill = Test_Prep)) +
    geom_violin() +
    ggtitle("Writing Score by Gender Boxplot") +
    xlab("Gender") + ylab("Writing Scores") +
    scale_color_jco(alpha=0.5)+scale_fill_jco(alpha=0.5) +
    theme_stata()

b1;b2;b3
```







小提琴图相当于密度分布图旋转 90 度，然后再做个对称的镜像，

最宽或者最厚的地方，对应着数据密度最大的地方。

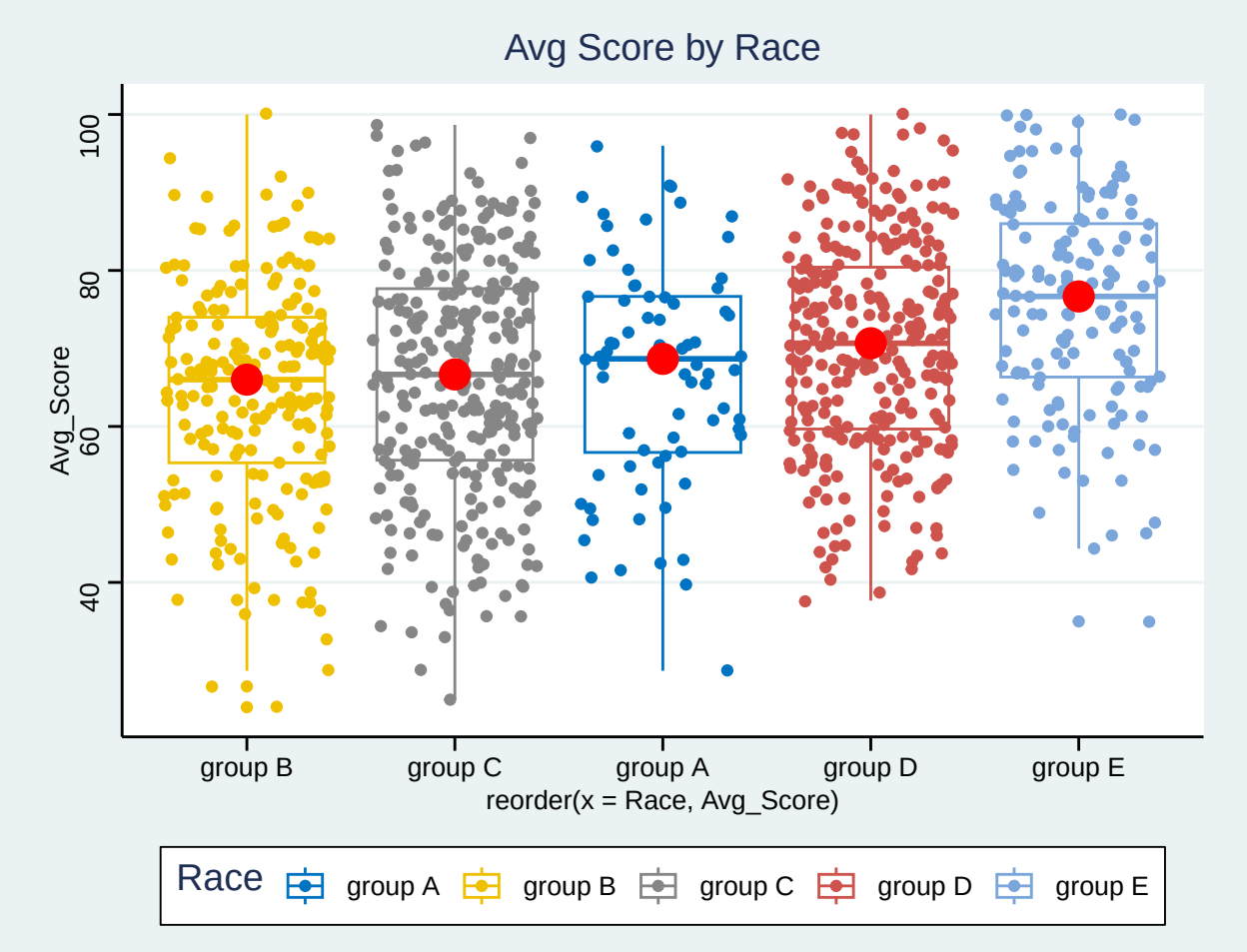
完成预科班学习的学生相比于没有进行预科班学习的学生在所有三项测试中都取得了更好的成绩。

男生的数学成绩较好；女生的阅读和写作成绩好。

在所有三个测试中都存在极端值（分数过高或过低），存在成绩特别好的同时也有成绩非常差的。

2.3.5 不同组学生的平均成绩（散点箱线图）

```
data.st %>%
  #reorder() 按照平均成绩从低到高排序
  ggplot(aes(reorder(x = Race,Avg_Score),y = Avg_Score ,color=Race)) +
  geom_boxplot() + # 箱线图
  geom_jitter()+ # 散点图
  stat_summary(fun.y = median, colour = "red", geom = "point", size = 5) +
  # 添加一个每一组平均值的点显示在图中便于观察
  ggtitle('Avg Score by Race') +
  scale_color_jco()+scale_fill_jco() + theme_stata()
```



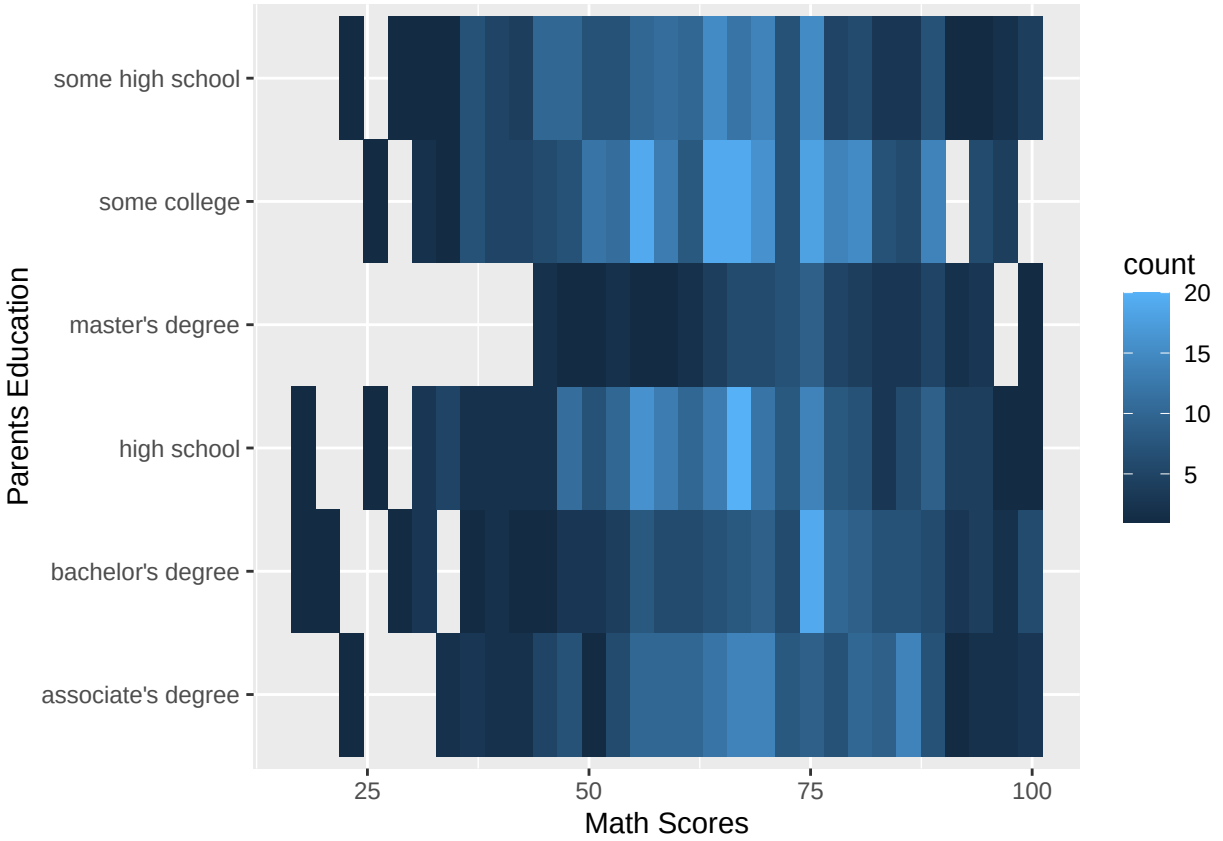
按平均分由低到高排序，可以看出 E 组学生的平均成绩明显高于另外四组，

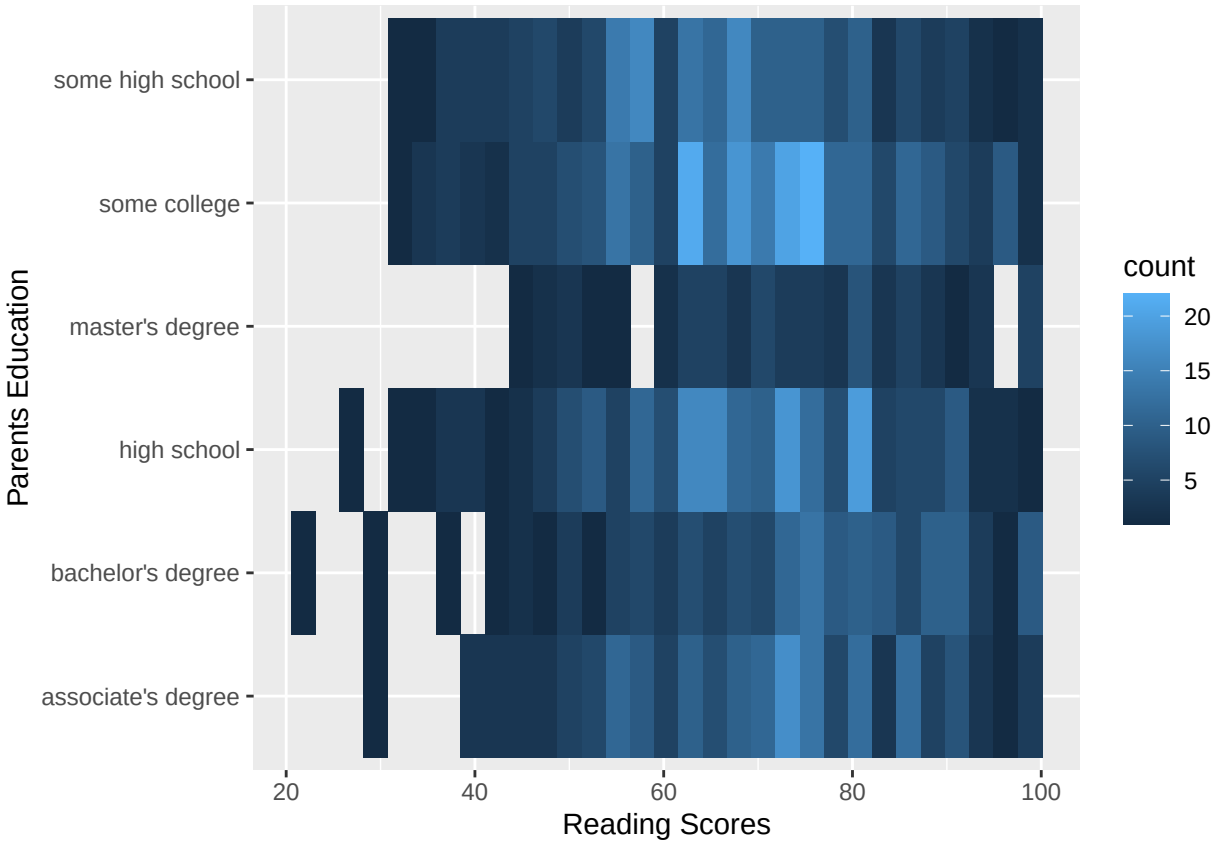
绝大部分 E 组人成绩集中在 70 分以上，

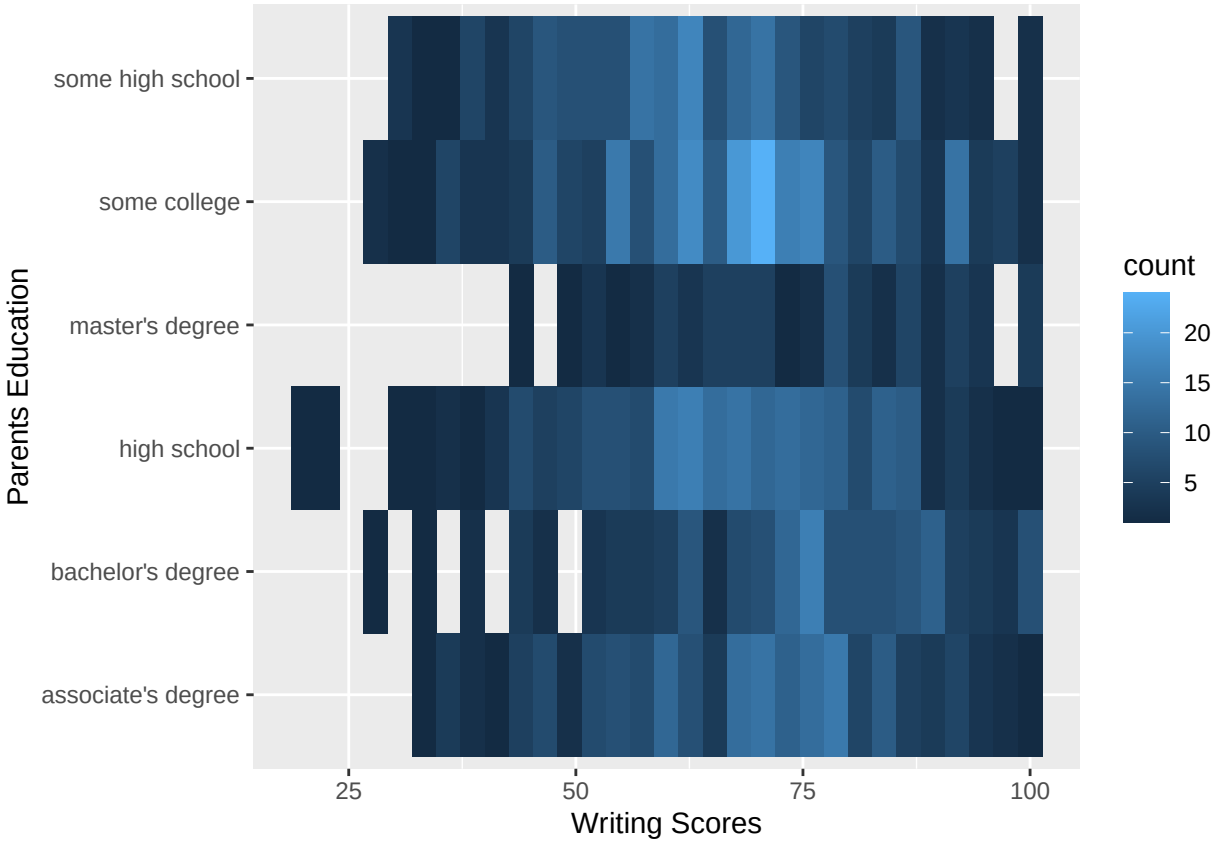
同时 B 组和 C 组有不少极端值（分数过高或过低）。

2.3.6 父母受教育程度和学生成绩（热力图）

```
e1 <- ggplot(data.st) + # 数学成绩
    geom_bin2d(aes(x=M_Score, y=Parent_Education)) +
    xlab("Math Scores") + ylab("Parents Education")

e2 <- ggplot(data.st) + # 阅读成绩
    geom_bin2d(aes(x=R_Score, y=Parent_Education)) +
    xlab("Reading Scores") + ylab("Parents Education")

e3 <- ggplot(data.st) + # 写作成绩
    geom_bin2d(aes(x=W_Score, y=Parent_Education)) +
    xlab("Writing Scores") + ylab("Parents Education")

e1;e2;e3
```



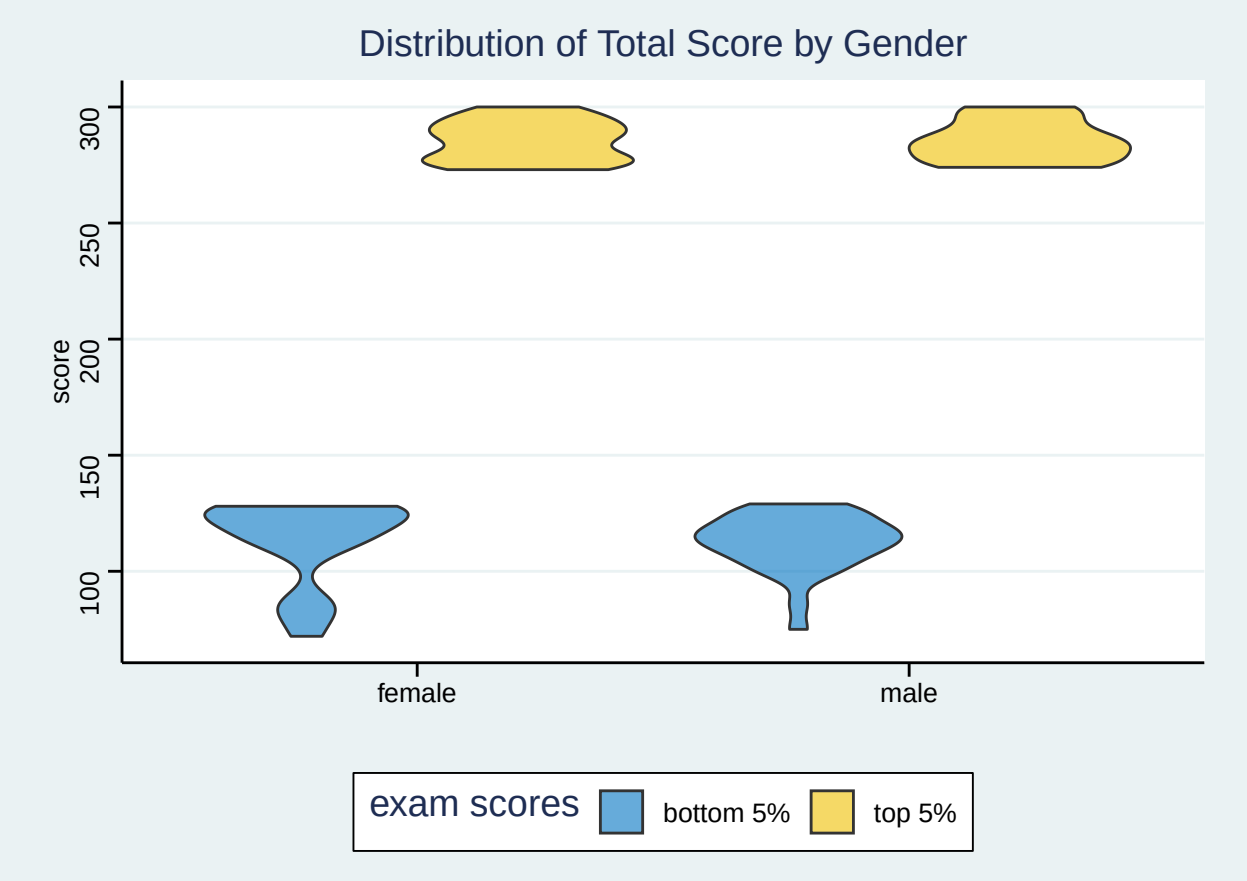


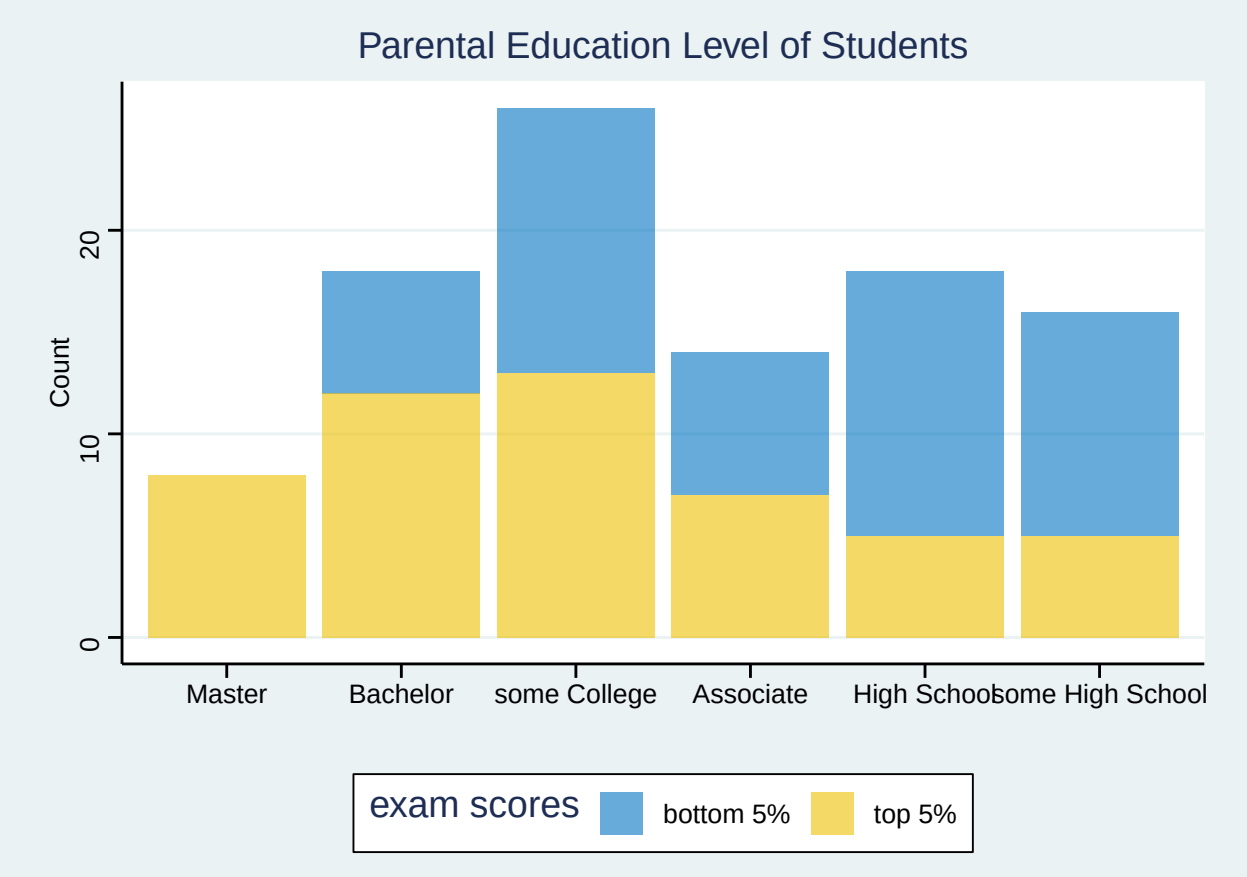


热力图能够很好的生成高质量的矩阵，用聚类算法将结果可视化，颜色越浅表明该分数段人越多。

2.3.7 学生成绩取前 5% 和后 5%（小提琴图和堆积柱状图）

```
# 取出前 5% 和后 5% 的学生
ALL_scores <- data.st %>% # 取出所有学生的成绩
  select(`M_Score`, `R_Score`, `W_Score`)
exam_data <- data.st %>% # 添加总分这一列
  mutate(total = rowSums(ALL_scores, na.rm = FALSE)) %>%
  arrange(desc(total))

top_bottom <- exam_data %>% arrange(desc(total)) %>%
  slice(1:50, 951:1000) %>% # 取前 5% 和后 5% 即前 50 和 951 到 100 名
  select(Gender, `Parent_Education`,total) %>%
  rename(parents = "Parent_Education") %>%
  mutate(group = ifelse(total >= 271, "top", "bottom")) %>%
  mutate(parent_edu_level = ifelse(parents == "master's degree", "1",
  ifelse(parents == "bachelor's degree", "2", ifelse(parents == "some college", "3",
  ifelse(parents == "associate's degree", "4", ifelse(parents == "high school", "5",
  ifelse(parents == "some high school", "6", "0")))))))

# 对数据进行可视化
# 根据性别对学生进行分组，用小提琴图查看其分布
t1 <- ggplot(top_bottom) +
  geom_violin(aes(x = Gender, y = total, fill = group)) + # 小提琴图
  ggtitle(label = "Distribution of Total Score by Gender") +
  labs(fill = "exam scores", y = "score", x = "") +
  scale_color_jco(alpha=0.6) + theme_stata() +
  scale_fill_jco(alpha=0.6,labels = c("bottom 5%", "top 5%"))
# 根据父母受教育程度进行分组
t2 <- ggplot(top_bottom) +
  geom_bar(aes(x = parent_edu_level, fill = group),stat = "count",
           position = "stack") + # 柱状图
  scale_x_discrete(labels=c(
    "1" = "Master", "2" = "Bachelor", "3" = "some College",
    "4" = "Associate","5" = "High School", "6" = "some High School")) +
  theme(axis.text.x = element_text(angle = 15, vjust = 0.7, face = "bold")) +
  ggtitle(label = "Parental Education Level of Students") +
  labs(x = "", fill = "exam scores") + ylab('Count')+
  scale_color_jco(alpha=0.6) + theme_stata() +
  scale_fill_jco(alpha=0.6,labels = c("bottom 5%", "top 5%"))

t1;t2
```


Distribution of Total Score by Gender
score
100
150
200
250
300
female
male
exam scores
bottom 5%
top 5%


Parental Education Level of Students
Count
0
10
20
Master
Bachelor
some College
Associate
High Schoo
ome High School
exam scores
bottom 5%
top 5%

对学生的三门总分进行计算并筛选出所有学生中的前 5% 和后 5%，

不同性别中仅有细微的差别，这表明性别并不会影响一个学生获得高分或是低分

观察每一组的父母受教育水平以及对应学生为前 5% 或者倒数 5% 的数量。

这个图显示，前 5% 的大多数学生的父母都受过大学教育，而后 5% 的大多数学生的父母都没有受过大学教育，

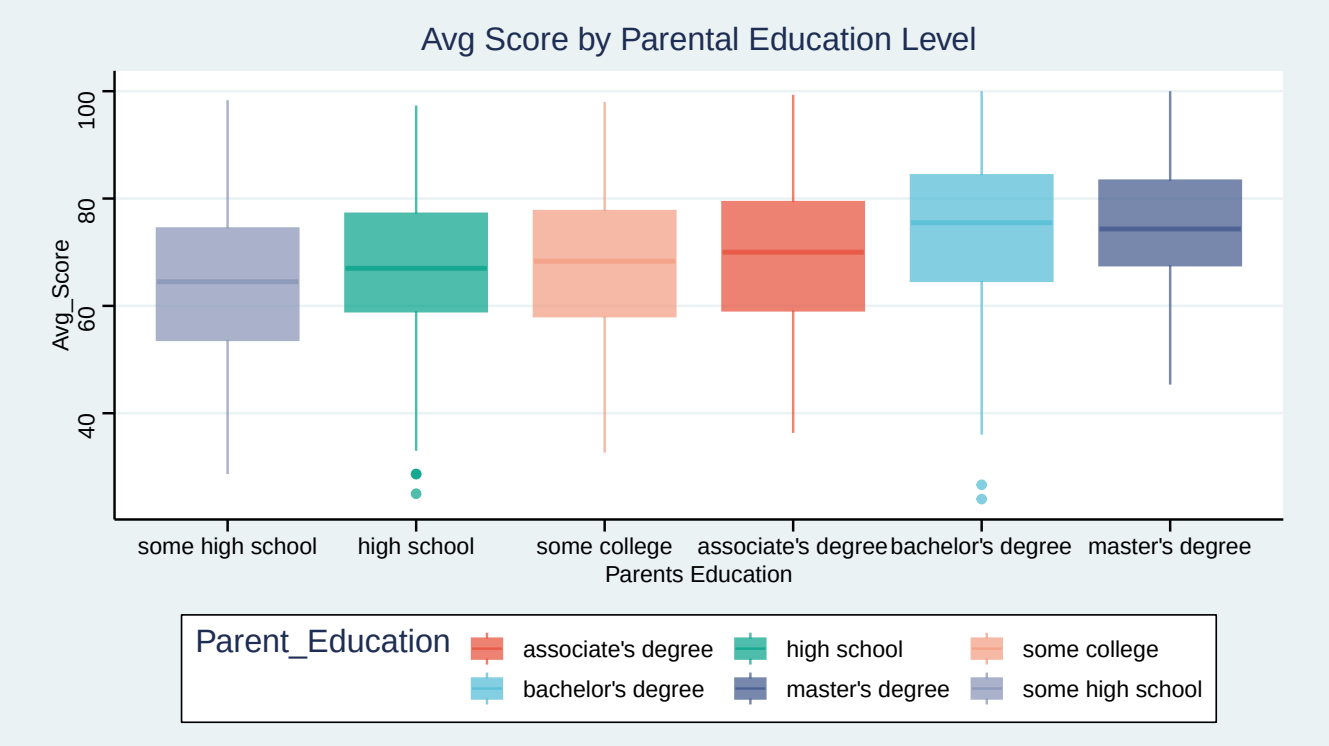
父母硕士毕业的学生在 top5% 中数量比学士毕业的学生要少，这里忽略了每一种受教育水平的父母的总人数并不一致，

从上图中可以发现父母硕士毕业而子女并没有哪怕一个在底部 5%，而最低学历的父母的子女任然有一部分进入了顶部 5%。

父母**受教育程度越大**学生的平均**成绩差距越大**。

根据以上几张图可以看的出**父母受教育程度越高，学生获得高分的概率越高**。### 父母受教育程度与学生平均成绩的回归分析

```
data.st %>%
  ggplot(aes(reorder(x = Parent_Education,Avg_Score),y = Avg_Score,
             color = Parent_Education, fill = Parent_Education)) +
  geom_boxplot() + theme_stata() + labs(x = "Parents Education") +
  ggtitle('Avg Score by Parental Education Level') +
  scale_color_npg(alpha = 0.7 )+scale_fill_npg(alpha = 0.7 )
```



```r
# 分类变量回归分析
mod1 <- lm(Avg_Score ~ Parent_Education, data = data.st)
summary(mod1)
```

```
Call:
lm(formula = Avg_Score ~ Parent_Education, data = data.st)

Residuals:
    Min      1Q  Median      3Q     Max
-49.606  -9.244   0.673  10.287  34.135

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         69.107      1.097  63.017  < 2e-16 ***
Parent_Educationbachelor's degree    4.499      1.622   2.774  0.00564 **
Parent_Educationhigh school         -2.270      1.502  -1.511  0.13120
Parent_Educationmaster's degree      5.687      2.056   2.766  0.00578 **
Parent_Educationsome college        -1.394      1.433  -0.973  0.33087
Parent_Educationsome high school    -4.909      1.531  -3.206  0.00139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.34 on 994 degrees of freedom
Multiple R-squared:  0.04899,   Adjusted R-squared:  0.04421
F-statistic: 10.24 on 5 and 994 DF,  p-value: 1.38e-09
```

输出结果仅有 Intercept 截距和另外五个系数，associate 不见了，而 Parent_Education 变量里有 6 组。

回归时，选择 associate 为基线，bachelor 的系数，可以理解为由 associate 切换到 bachelor，引起 Avg_Score 收入的变化（效应）

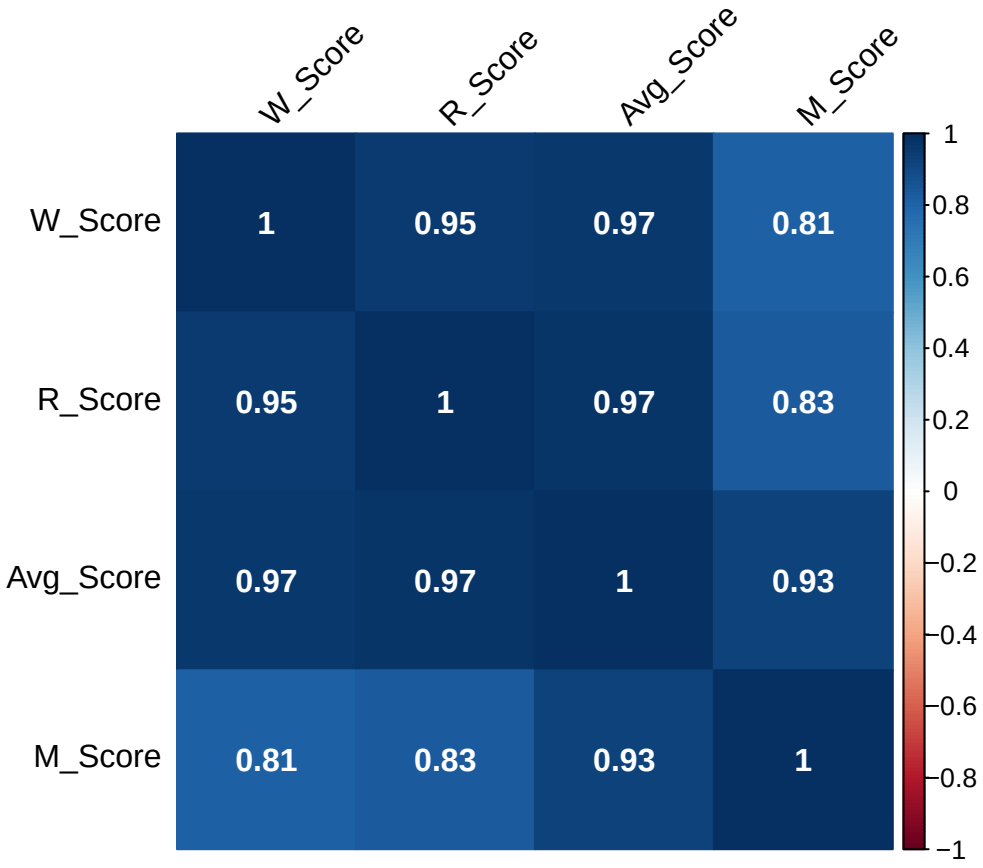
- 对 associate 组的估计，`Avg_Score = 69.107  = 69.107`
- 对 bachelor 组的估计，`Avg_Score = 69.107  + 4.499 = 73.606`
- 对 high school 组的估计，`Avg_Score = 69.107 + -2.270  = 66.837`
- 对 master 组的估计，`Avg_Score = 69.107 + 5.687 = 74.794`
- 对 some college 组的估计，`Avg_Score = 69.107 + -1.394 = 67.713`
- 对 some high school 组的估计，`Avg_Score = 69.107  + -4.909 = 64.198`

仅用父母的受教育程度**无法完全解释**学生平均分（Multiple R-squared 仅有 4.899% 即对学生平均分的解释仅有 4.899%）。

2.4 STEP 4 Regression 回归分析

2.4.1 不同分数间的相关性

```
scores <- c("M_Score", "R_Score", "W_Score","Avg_Score")
score <- data.st[scores]
S <- cor(score)
corrplot(S,method="color",addCoef.col="white",tl.col="black",tl.srt=45,order="AOE")
```



由上图可以看出数学成绩与写作和阅读成绩相关性差不多，而阅读和写作成绩之间的**相关性非常高**，这意味着阅读和写作成绩某一项较高（低）的学生他另一项的成绩也会较高（低）。

2.4.2 学生数学成绩和阅读、写作成绩的回归分析引入分类变量性别

```
# 数学
reg1 <- lm(M_Score~ Gender+R_Score+W_Score, data=data.st)
summary(reg1)
```

```
Call:
lm(formula = M_Score ~ Gender + R_Score + W_Score, data = data.st)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-21.4578  -4.3441  -0.0211   4.3007  18.2481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.55070    1.04058  -6.295 4.59e-10 ***
Gendermale  12.73937    0.42251  30.152  < 2e-16 ***
R_Score      0.39229    0.04584   8.557  < 2e-16 ***
W_Score      0.57939    0.04489  12.906  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.315 on 996 degrees of freedom
Multiple R-squared:  0.8401,    Adjusted R-squared:  0.8396
F-statistic:  1745 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
# 阅读
reg2 <- lm(R_Score~ Gender+M_Score+W_Score, data=data.st)
summary(reg2)
```

```
Call:
lm(formula = R_Score ~ Gender + M_Score + W_Score, data = data.st)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9118  -2.7765   0.1099   2.8840  12.8172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.55066    0.68563   8.096 1.65e-15 ***
Gendermale  -0.60214    0.38934  -1.547    0.122
M_Score      0.17457    0.02040   8.557  < 2e-16 ***
W_Score      0.76936    0.02127  36.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.213 on 996 degrees of freedom
Multiple R-squared:  0.9203,    Adjusted R-squared:   0.92
F-statistic:  3832 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
# 写作
reg3 <- lm(W_Score~ Gender+M_Score+R_Score, data=data.st)
summary(reg3)
```

```
Call:
lm(formula = W_Score ~ Gender + M_Score + R_Score, data = data.st)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7595  -2.6725  -0.0319   2.9250  14.1511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.24629    0.68553   4.735  2.5e-06 ***
Gendermale  -5.24466    0.34368 -15.260  < 2e-16 ***
M_Score      0.24728    0.01916  12.906  < 2e-16 ***
R_Score      0.73788    0.02040  36.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.126 on 996 degrees of freedom
Multiple R-squared:  0.9291,    Adjusted R-squared:  0.9289
F-statistic:  4350 on 3 and 996 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.8401，性别、阅读和写作成绩对数学成绩的解释达到 84%，另外两个分析结果也都达到 92%，说明解释了大部分信息。

以下为三个分析所得出的回归模型：

$$Math\_Score = -6.55 + 12.74 * Gendermale + 0.39 * R\_Score + 0.58 * W\_Score$$

$$Writing\_Score = 3.25 - 5.24 * Gendermale + 0.25 * M\_Score + 0.74 * R\_Score$$

$$Reading\_Score = 5.55 - 0.60 * Gendermale + 0.17 * M\_Score + 0.77 * W\_Score$$

模型解释：

三个模型都以 Genderfemale 为基线，其中 Math_Score 与性别为女的相关性为负，而 Writing_Score 和 Reading_Score 与性别男相关性为负，这与上方所做出的堆积直方图结果吻合，男生的数学成绩优于女生，而女生写作和阅读成绩优于男生，分别呈正相关。

除此之外，虽然数学成绩与男性呈正相关与女性呈负相关但其系数高于阅读和写作与性别关系的系数。

以 Math_Score 为例：

- `Gendermale = 12.74` 当 R_Score 和 W_Score 保持不变时，Gendermale 变化引起的 Math_Score 变化

2.4.3 学生的写作分数和阅读分数的回归分析引入分类变量性别及线性

```
reg <- lm(W_Score~ Gender+R_Score, data=data.st)
summary(reg)
```

```
Call:
lm(formula = W_Score ~ Gender + R_Score, data = data.st)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8988  -3.0291  -0.2005   3.0516  14.1154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.898412   0.731625   2.595   0.0096 **
Gendermale  -2.444703   0.287833  -8.493   <2e-16 ***
R_Score      0.974509   0.009666 100.818   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.455 on 997 degrees of freedom
Multiple R-squared:  0.9172,    Adjusted R-squared:  0.9171
F-statistic:  5524 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
data.st %>%
  ggplot(aes(x = R_Score, y = W_Score, color = Gender)) +
  geom_point(alpha = 0.1) +
  geom_smooth(aes(y = predict(reg))) +
  ggtitle('Writing Score with Gender and Reading Score') +
  scale_color_npg(alpha = 0.7 )+scale_fill_npg(alpha = 0.7 ) +
  theme_stata()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

$$Writing\_Score = 1.90 - 2.44 * Gendermale + 0.97 * R\_Score$$

写作分数和阅读分数**线性正相关**。

3 实验总结

3.1 实验具体内容

1. 男女生数量按组别分：柱状图（横向）
2. 父母的受教育程度的统计：直方图
3. 学生成绩分男女：堆积柱状图
4. 学生成绩分男女分是否准备：小提琴图
5. 学生平均成绩与组别：散点箱线图
6. 学生成绩与家长受教育程度：热力图
7. 学生成绩取前 5% 和后 5% （与性别和家长受教育程度）：小提琴图和堆积柱状图
8. 学生成绩与家长受教育程度的分类回归分析：箱线图
9. 学生成绩间的相关性图
10. 学生成绩与性别的回归分析
11. 学生的写作/阅读分数的回归分析

3.2 数据分析结论

学生成绩分为数学、阅读、写作成绩

数据中还包含学生的性别、父母受教育程度、分组情况、是否完成预科班学习

通过数据可视化和回归分析可以得到以下结论：

1. C 组与 D 组人数最多，而 A 组人数远小于其他组。
2. 高中和一些学院毕业的父母较多而硕士毕业的父母较少。前 5% 的大多数学生的父母都受过大学教育，而后 5% 的大多数学生的父母都没有受过大学教育，**父母受教育程度越高学生的平均成绩差距越大，父母受教育程度越高学生获得高分的概率越高**。
3. 男生的数学成绩明显高于女生，女生的阅读和写作成绩也高于男生，**男生的逻辑思维能力较强而女生的感性和理解能力较强**。
4. 完成预科班学习的学生相比于没有进行预科班学习的学生在所有三项测试中都取得了更好的成绩，**一定的提前预习有助于获取高分**。