notebook

September 14, 2022

1 Analyzing School´ Test Scores

1.1 Background

A school makes every student take year-end math, reading, and writing exams. Includes year-end exam grades for **1,000 students**. The school data, available on this link, It also has the following fields :

1.2 The data

- exam score in math, reading and writing
- gender, race/ethnicity of the student
- the highest education level of either parent
- whether the student took the test preparation course.

The school are interested in knowing -The averages and the influence of taking or not taking the test score. -What are the average scores for the different parental education levels? -The correlation between the results in the different tests.

[100]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
options(warn=-1)
options(message=1)

suppressMessages(library(reshape))
suppressMessages(library(tidyverse))
suppressMessages(install.packages("viridis"))
suppressMessages(library(viridis))
suppressMessages(install.packages("ggthemes"))
suppressMessages(library(ggthemes))
suppressPackageStartupMessages(install.packages("GGally"))
suppressPackageStartupMessages(library(GGally))
suppressPackageStartupMessages(install.packages("ggcharts"))
suppressPackageStartupMessages(library(ggcharts))
suppressPackageStartupMessages(install.packages("ggcorrplot"))
suppressPackageStartupMessages(library(ggcorrplot))
suppressPackageStartupMessages(library(patchwork))
suppressPackageStartupMessages(install.packages("formattable"))
suppressPackageStartupMessages(library(formattable))
suppressPackageStartupMessages(install.packages("flextable"))
suppressPackageStartupMessages(library(flextable))
suppressPackageStartupMessages(install.packages("IRdisplay"))
suppressPackageStartupMessages(library(IRdisplay))
suppressPackageStartupMessages(install.packages("xtable"))
suppressPackageStartupMessages(library(xtable))
suppressPackageStartupMessages(install.packages("ggdark"))
suppressPackageStartupMessages(library(ggdark))
suppressPackageStartupMessages(install.packages("mdthemes"))
suppressPackageStartupMessages(library(mdthemes))
suppressPackageStartupMessages(install.packages("ggExtra"))
suppressPackageStartupMessages(library(ggExtra))
suppressPackageStartupMessages(install.packages("ggtext"))
suppressPackageStartupMessages(library(ggtext))
suppressPackageStartupMessages(install.packages("waffle"))
suppressPackageStartupMessages(library(waffle))

knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

graph <- function(){
    theme(text = element_text(family = "Lato"),
          plot.title = element_text(size = 26, face = "bold"),
          plot.subtitle = element_text(size = 24),
          axis.text.y = element_text(size = 22),
          axis.text.x = element_text(size = 22),
          legend.title = element_blank(),
          legend.text = element_text(size = 20),
          strip.text = element_text(size = 24, angle = 0),
          axis.title.x = element_blank(),
          axis.title.y = element_blank()
          )
}
```

```
Installing viridis [0.6.2] …
	OK [linked cache]
Installing ggthemes [4.2.4] …
	OK [linked cache]
Installing GGally [2.1.2] …
	OK [linked cache]
Installing ggcharts [0.2.1] …
	OK [linked cache]
Installing ggcorrplot [0.1.3] …
	OK [linked cache]
Installing formattable [0.2.1] …
	OK [linked cache]
Installing flextable [0.8.0] …
	OK [linked cache]
Installing IRdisplay [1.1] …
	OK [linked cache]
Installing xtable [1.8-4] …
	OK [linked cache]
Installing ggdark [0.2.1] …
	OK [linked cache]
Installing mdthemes [0.1.0] …
	OK [linked cache]
Installing ggExtra [0.10.0] …
	OK [linked cache]
Installing ggtext [0.1.1] …
	OK [linked cache]
Installing waffle [0.7.0] …
	OK [linked cache]
```

[101]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

df <- readr::read_csv('./data/exams.csv',show_col_types = FALSE)
```

[102]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
# Vamos a preparar la data, son factores las columnas de race/ethnicity, la de␣
 ↪parent education level, lunch, test,prep,course y gender
df<- df %>% mutate(gender = as.factor(gender), parent_educ_lvl=as.
 ↪factor(parent_education_level), lunch= as.factor(lunch), test_prep_course=␣
 ↪as.factor(test_prep_course ), ID= 1:1000)

#lo convertimos en un tibble
df <- as.tibble(df) %>% select(-parent_educ_lvl)
```

```
#hacemos un avg general
df <- df %>% mutate(general_avg = round(((math+reading+writing)/3),2))


head(df)
```

A tibble: 6 × 10

gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math
<fct>	<chr>	<chr>	<fct>	<fct>	<dbl>
female	group B	bachelor's degree	standard	none	72
female	group C	some college	standard	completed	69
female	group B	master's degree	standard	none	90
male	group A	associate's degree	free/reduced	none	47
male	group C	some college	standard	none	76
female	group B	associate's degree	standard	none	71

1.3 Summary Stadistics

[103]:
```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

repr_html.xtable <- function(obj, ...){
    paste(capture.output(print(obj, type = 'html')), collapse="", sep="")
}



df <- df %>%
  rename("race_ethnicity" = `race/ethnicity`) %>%
  mutate(gender = factor(gender),
         race_ethnicity = factor(race_ethnicity,
                                 levels = c("group A","group B","group C","group␣
 ↪D","group E"),
                                 labels = c("group A","group B","group C","group␣
 ↪D","group E")),
         parent_education_level = factor(parent_education_level,
                                         levels = c("some high school","high␣
 ↪school","some college","associate's degree","bachelor's degree","master's␣
 ↪degree"),
                                         labels = c("s/ high school","high␣
 ↪school","some college","associate's","bachelor's","master's")
                                         ),

         lunch = factor(lunch, levels = c("standard","free/reduced"), labels =␣
 ↪c("standard","free/reduced")),
         test_prep_course = factor(test_prep_course, levels =␣
 ↪c("none","completed"), labels = c("none","completed")))
```

```
df %>% summary() %>% xtable()
```

A xtable: 6 × 10

	gender <chr>	race_ethnicity <chr>	parent_education_level <chr>	lunch <chr>	test_prep_co <chr>
X	female:518	group A: 89	s/ high school:179	standard :645	none :642
X.1	male :482	group B:190	high school :196	free/reduced:355	completed:358
X.2	NA	group C:319	some college :226	NA	NA
X.3	NA	group D:262	associate's :222	NA	NA
X.4	NA	group E:140	bachelor's :118	NA	NA
X.5	NA	NA	master's : 59	NA	NA

- Female and male students are more or less equally distributed.
- Only 35% of students have taken the test preparation course.
- The average scores for math, reading, and writing are 66, 69, and 68 respectively.
- In terms of the highest level of education achieved by either parent, there are six categories.

1.4 What are the average reading scores for students with/without the test preparation course?

[104]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

df %>%
  filter(test_prep_course == "completed") %>%
  summarize(avg_score_completed = mean(reading))

df %>%
  filter(test_prep_course == "none") %>%
  summarize(avg_score_none = mean(reading))
```

A tibble: 1 × 1

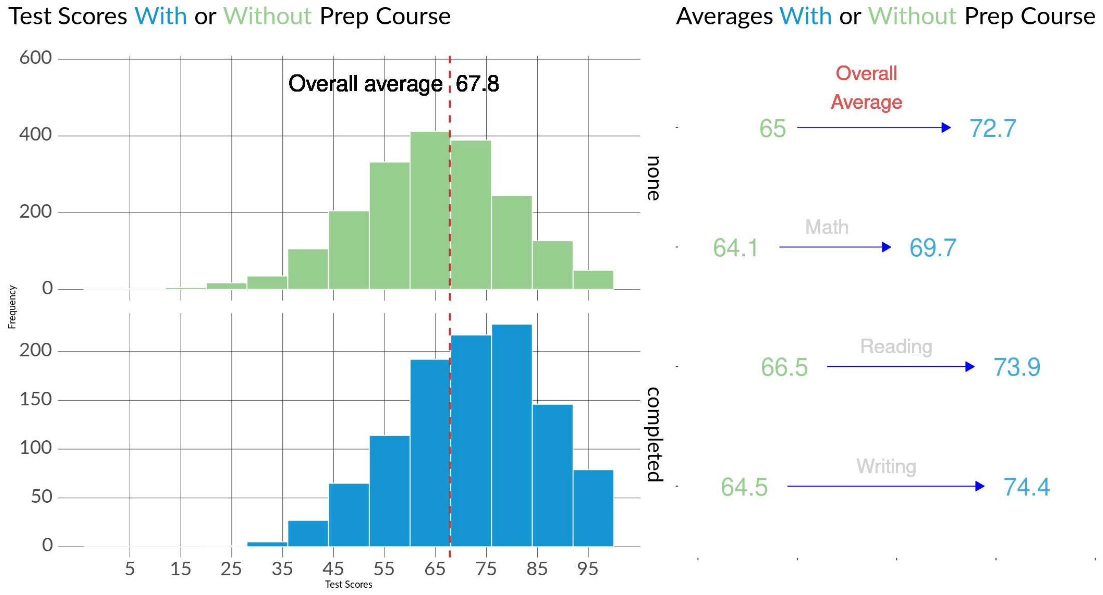
avg_score_completed <dbl>
73.89385

A tibble: 1 × 1

avg_score_none <dbl>
66.53427

[105]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
options(jupyter.plot_mimetypes = "image/png")
options(repr.plot.width = 18, repr.plot.height =10)
options(warn=-1)

df_completed <- df %>% filter(test_prep_course == "none")

testprep1 <- df %>% group_by(test_prep_course) %>%
  select(test_prep_course, math, reading, writing) %>%
  pivot_longer(-test_prep_course,
               names_to = "subject", values_to = "scores") %>%
  mutate(scores = round(scores,2))

# histogram
testplot <- testprep1 %>%
  ggplot(aes(scores, fill = test_prep_course)) +
  geom_histogram(bins = 10, binwidth = 8, color = "white") +
  geom_vline(xintercept = 67.8, color = "#db2b27", linetype = "dashed", size =␣
 ↪1) +
  scale_x_continuous(breaks = c(5,15,25,35,45,55,65,75,85,95)) + #
  scale_fill_manual(values = c("#98cf90","#1696d2")) +
  facet_grid(test_prep_course ~ ., scales = "free_y") +
  geom_text(data = df_completed, aes(x = 56.9, y = 580, label = "\nOverall␣
 ↪average  67.8"), size = 9, color = "black") +
  labs(y = "Frequency", x = "Test Scores", fill = "Test Prep Course", title =
      "<br>Test Scores <span style = 'color:#1696d2'>***With***</span> or <span␣
 ↪style = 'color:#98cf90'>***Without***</span> Prep Course<br/>") +
  as_md_theme(theme_tufte()) +
  as_md_theme(graph()) +
  as_md_theme(theme(
      plot.title.position = "plot",
          plot.title = element_text(size = 28, color = "black"),
          axis.title.y = element_text(color = "black"),
          axis.title.x = element_text(color = "black"),
          panel.spacing = unit(1, "lines"),
          legend.position = "none",
          legend.title = element_text(size = 20),
          axis.line = element_line(color = "white"),
          axis.ticks.x = element_line(color = "white"),
          strip.text.y = element_text(color = "black"),
          strip.background.y = element_blank(),
          panel.background = element_blank(),
          panel.grid.major = element_line(color = "grey30", size = 0.3),
          panel.grid.minor = element_blank(),
          axis.ticks = element_blank()))

tpc <- df %>%  group_by(test_prep_course) %>%
  summarize(Math = round(mean(math),1),
            Reading = round(mean(reading),1),
            Writing = round(mean(writing),1)) %>%
  mutate(Average = (Reading + Math + Writing)/3,  Average = round(Average,1))%>%
  pivot_longer(-test_prep_course,
               names_to = "Subject",
               values_to = "Score") %>%
  ggplot() +
  geom_path(aes(Score, fct_rev(Subject)),
            color = "blue",
            arrow = arrow(angle = 30, length = unit(11, "pt"),ends = "last",␣
 ↪type = "closed")) +
  geom_text(aes(Score, fct_rev(Subject), label = Score, color =␣
 ↪test_prep_course,
            hjust = ifelse(test_prep_course == "none",1.4,-0.4)), size = 10,␣
 ↪face = "bold") +
  geom_text(x = 68.5, y = 4.05, label = "Overall\nAverage", vjust = - 0.3,␣
 ↪color = "#e25552",size = 8) +
  geom_text(x = 66.5, y = 3.05, label = "Math", vjust = - 0.5, color =␣
 ↪"lightgrey", size = 8) +
  geom_text(x = 70, y = 2.05, label = "Reading", vjust = - 0.5, color =␣
 ↪"lightgrey", size = 8) +
  geom_text(x = 69.5, y = 1.05, label = "Writing", vjust = - 0.5, color =␣
 ↪"lightgrey", face = "bold", size = 8) +
  xlim(60,80) +
  theme_tufte() +
  graph() +
  labs(x = " ", y = " ", fill = "Test Prep Course",
      title =
          "<br>Averages <span style = 'color:#1696d2'>***With***</span> or␣
 ↪<span style = 'color:#98cf90'>***Without***</span> Prep Course<br/>") +
  theme(plot.title.position = "plot",
        plot.title = element_textbox(size = 28, color = "black", face =␣
 ↪"plain"),
        legend.position = "none",
        axis.text.x = element_blank(),
        axis.text.y =  element_blank()) +
  removeGrid() +
  scale_color_manual(values =  c("#98cf90","#46ABDB"))


(testplot + tpc) + plot_layout(widths =c(2,1.5))
```



- Students who completed the test preparation course improved their overall scores by an average of 7.5 points, with the greatest improvement seen in Writing (nearly 10 point increase). Those who did not complete the exam preparation course scored below the general average in all subjects.

1.5 What are the average scores for the different parental education levels?

[106]:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
# adding education column with three parental education groups
df1 <- df %>% mutate(education = case_when(parent_education_level %in% c("some␣
 ↪high school", "high school") ~ "didn't go\nto college",
                                      parent_education_level %in% c("some␣
 ↪college", "associate's") ~ "Some college\nor Associate's",
                                      parent_education_level %in% c("bachelor's",␣
 ↪"master's") ~ "Bachelor's\nor Master's"),
             education = factor(education,
                                  levels = c("didn't go\nto college", "Some␣
 ↪college\nor Associate's", "Bachelor's\nor Master's"),
                                  labels = c("didn't go\nto college", "Some␣
 ↪college\nor Associate's", "Bachelor's\nor Master's")))
df1 %>% select(parent_education_level, education) %>%
summary() %>% xtable(NULL= FALSE)
```

A xtable: 6 × 2

	parent_education_level	education
	<chr>	<chr>
X	s/ high school:179	didn't go to college :196
X.1	high school :196	Some college or Associate's:448
X.2	some college :226	Bachelor's or Master's :177
X.3	associate's :222	NA's :179
X.4	bachelor's :118	NA
X.5	master's : 59	NA

[107]:

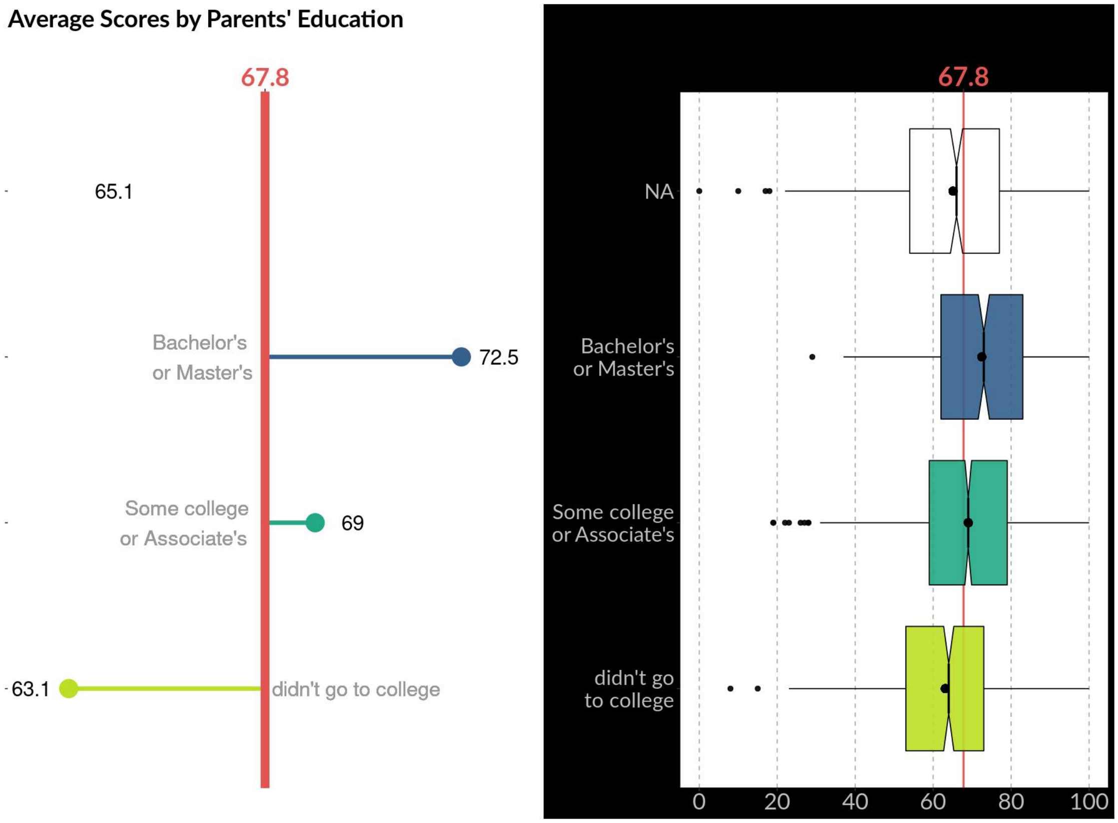
```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
score_parent_educ<- df %>%
  group_by(`parent_education_level`) %>%
  summarize(avg_score = mean(reading))%>%
arrange(desc(avg_score))
score_parent_educ
```

A tibble: 6 × 2

parent_education_level	avg_score
<fct>	<dbl>
master's	75.37288
bachelor's	73.00000
associate's	70.92793
some college	69.46018
s/ high school	66.93855
high school	64.70408

[108]:

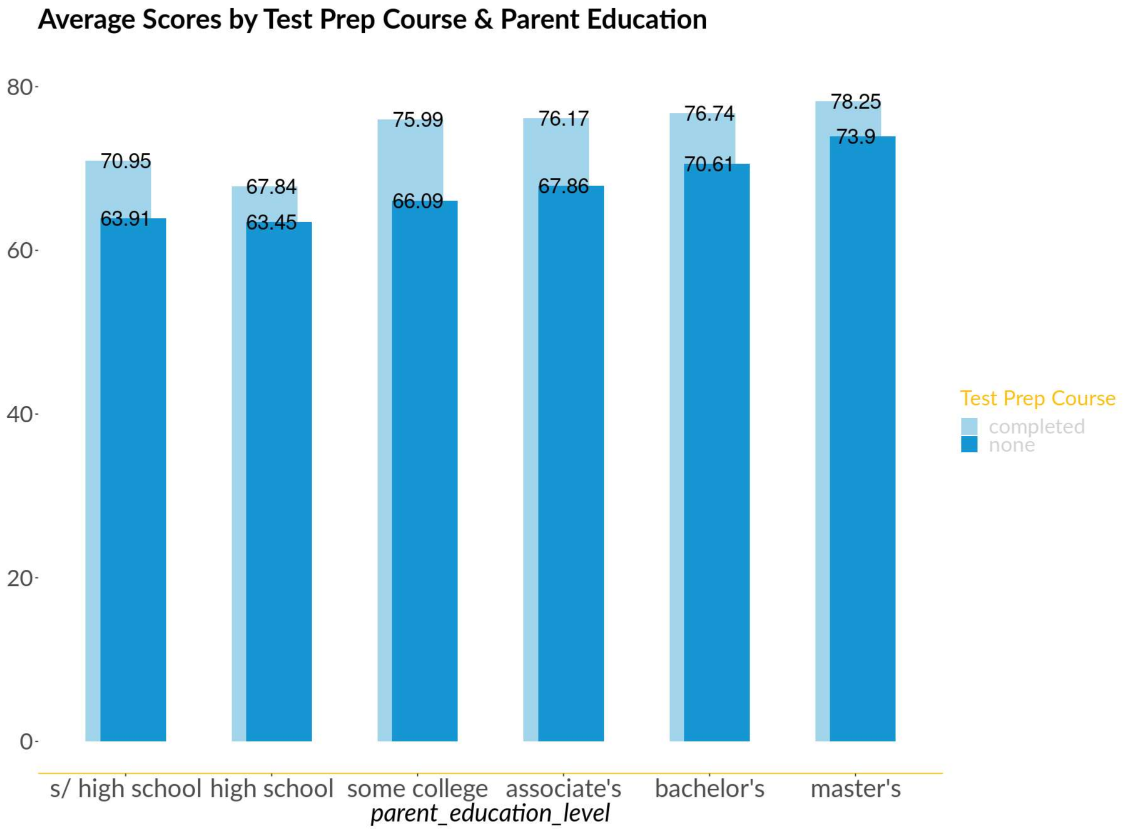
```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
options(jupyter.plot_mimetypes = "image/png")
options(repr.plot.width = 15, repr.plot.height = 11)
options(warn=-1)

edu_boxplot <- df1 %>% group_by(education) %>%
  select(education, math, reading, writing) %>%
  pivot_longer(-education, names_to = "subject", values_to = "scores") %>%
  mutate(scores = round(scores,2)) %>%
  ggplot(aes(scores, fct_reorder(education, scores),
             label = scores,
             fill = fct_reorder(education, scores))) +
  scale_x_continuous(breaks = c(0,20,40,60,80,100),
                     sec.axis = dup_axis(breaks = 67.8, label = "67.8")) +
  coord_cartesian(clip = 'off') +   #This keeps the labels from disappearing
  geom_vline(xintercept = 67.8, color = "#e25552", size = 0.9, alpha = 1) +
  geom_boxplot(notch = TRUE,
        notchwidth = 0.4,
        outlier.colour = "black",
        outlier.fill = "black",
        outlier.size = 2,
        color = "black",
        alpha = 0.9) +
 stat_summary(fun="mean", color = "black", shape = 16, size = 0.9) +
 scale_fill_ordinal(option = "D",begin = 0.3, end = 0.9, direction = -1) +
 theme_tufte() +
 graph() +
 theme(plot.background = element_rect(fill = "black"),
        panel.background = element_rect(fill = "white"),
        legend.position = "none",
        axis.line.x = element_line(),
        axis.title.x.top = element_blank(),
        axis.text.x.bottom = element_text(color = "grey"),
        panel.grid.major = element_line(color = "darkgrey", linetype =␣
 ↪"dashed"),
        axis.text.y = element_text(color = "grey", size = 20),
        axis.text.x.top = element_text(color ="#e25552", face = "bold", size =␣
 ↪25),
        strip.text = element_text(size = 24)) +
removeGridY()


edu_lp_lg <- df1 %>%
group_by(education) %>%
summarize(Math = round(mean(math),1),
          Reading = round(mean(reading),1),
          Writing = round(mean(writing),1)) %>%
mutate(Average = (Reading + Math + Writing)/3,
         Average = round(Average,1)) %>%
ggplot(aes(Average,
           y = fct_reorder(education, Average),
           color = fct_reorder(education, desc(Average)))) +
geom_segment(aes(y = fct_reorder(education, Average),
                   yend = fct_reorder(education, Average), x = 67.8,
                   xend = Average, color = fct_reorder(education,␣
 ↪Average)),size = 2) +
geom_point(size = 8) +
labs(title = "Average Scores by Parents' Education\n") +
coord_cartesian(clip = 'off') +
scale_x_continuous(breaks = c(60,67.8,80,5), position = "top") +
scale_color_viridis_d(option = "D",begin = 0.3, end = 0.9, direction = -1) +
geom_vline(xintercept = 67.8, size = 4, color = "#e25552") +
geom_text(aes(label = Average, x = ifelse(Average > 67.8, Average + 0.9,␣
 ↪Average - 0.9)),
         size = 7, color = "black") +
theme_tufte()+ graph() +
theme(legend.position = "none",
      legend.title = element_blank(),
      axis.text.x = element_text(face = "bold", size = 24, color = "#e25552"),
        panel.grid.major.y = element_blank(),
        panel.grid.major.x = element_line(),
        axis.text.y = element_blank(),
        panel.border = element_blank(),
       plot.title = element_text(size = 23))  +
geom_text(y = 1, x = 68, label = "didn't go to college", color = "gray60",␣
  ↪hjust = 0.01, size = 7)+
geom_text(y = 2, x = 68, label = "Some college\nor Associate's", color =␣
  ↪"gray60", hjust = 1.2, size = 7)+
geom_text(y = 3, x = 68, label = "Bachelor's \nor Master's", color = "gray60",␣
  ↪hjust = 1.2, size = 7)


lp_bp_edu <- (edu_lp_lg | plot_spacer() | edu_boxplot)+ plot_layout(widths =␣
  ↪c(1.2,0.01,1))
lp_bp_edu
```



```
[109]: knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

       score_parent_test<- df %>%
         group_by(`parent_education_level`) %>% filter(test_prep_course ==␣
         ↪"completed") %>%
  summarize(avg= mean(reading) )%>%
arrange(desc(avg)) %>% mutate(completed='completed')

score_parent_test1<- df %>%
  group_by(`parent_education_level`) %>% filter(test_prep_course == "none") %>%
  summarize(avg= mean(reading) )%>%
arrange(desc(avg)) %>% mutate(none='none')

table <- union_all(score_parent_test,score_parent_test1)


table<- table %>% select(-none) %>% rename(test_course= completed) %>%␣
 ↪mutate(test_course = replace_na(test_course, replace = "none"))

table<- table %>% arrange(desc(parent_education_level)) %>%␣
 ↪mutate(avg=round(avg,2))
```

[110]:

```
# grafico de la comparacion
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
table %>%
ggplot(aes(parent_education_level, avg, fill=test_course)) +
geom_col(position =  position_dodge(width = 0.2)) +
scale_fill_manual(values = c("#a2d4ec", "#1696d2")) +
theme_tufte()+ graph() +
labs(title = "Average Scores by Test Prep Course & Parent Education\n",
     fill = "Test Prep Course") +
geom_text(aes(label=avg), size =7, color= "black")+
theme(axis.line.x = element_line(color = "#fdbf11"),
      axis.title.x = element_text(face = "italic", size = 24),
      axis.text.x = element_text(size = 24),
      legend.text = element_text(size = 20, color = "#d2d2d2"),
      legend.title = element_text(size = 20, color = "#fdbf11"))
```

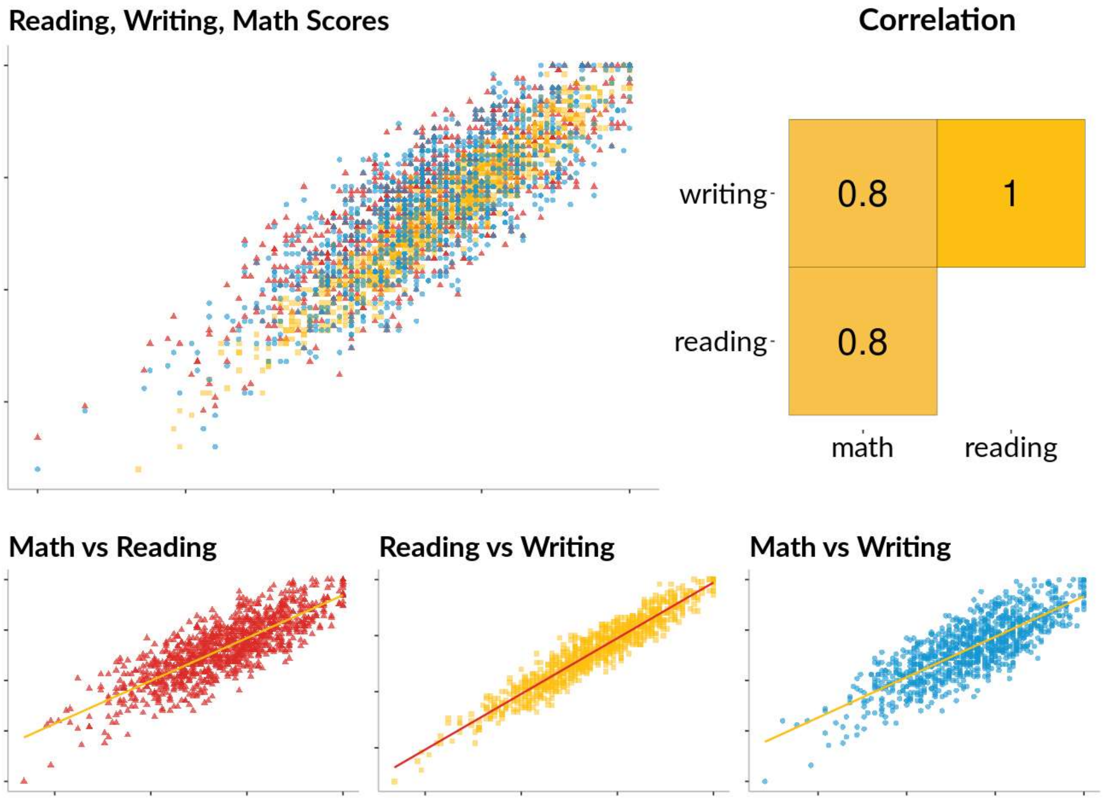


- the higher the educational level of the parents, the higher the results of the kids..
- Students whose parents did not go to college had a low overall average score of 64.1.
- There's an 8.4-point average gap between the lowest scoring group, group A, and the highest scoring group, group E.

1.6 Correlations between scores.

[111]:
```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
options(jupyter.plot_mimetypes = "image/png")
options(repr.plot.width = 11, repr.plot.height = 8)
options(warn=-1)

scores <- df %>% select(math,reading,writing)

plotall <- scores %>%
  ggplot() +
  geom_point(aes(math, reading),shape=17, color="#db2b27", size=1.5,alpha = 0.
↪7) +
  geom_point(aes(reading, writing),shape=15, color="#fdbf11", size=1.5, alpha =␣
↪0.5) +
  geom_point(aes(math, writing),shape=16, color = "#1696d2", size=1.5,alpha = 0.
 ↪6) +
  theme_tufte()+ graph() +
  labs(title = "Reading, Writing, Math Scores") +
  theme(plot.title = element_text(size = 20),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.line = element_line(color = "grey"))


plot1 <- scores %>%
  ggplot(aes(reading, math)) +
  geom_point(shape = 17, color = "#db2b27", size = 1.5, alpha = 0.7) +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE, color = "#fdbf11",␣
 ↪size = 0.7) +
 theme_tufte()+ graph() +
  labs(title = "Math vs Reading") +
  theme(plot.title = element_text(size = 20),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.line = element_line(color = "grey"))

plot2 <- scores %>%
  ggplot(aes(reading, writing)) +
  geom_point(shape=15, color="#fdbf11", size=1.5, alpha = 0.5) +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE, color = "#db2b27",␣
 ↪size = 0.7) +
theme_tufte()+ graph() +
  labs(title = "Reading vs Writing") +
  theme(plot.title = element_text(size = 20),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.line = element_line(color = "grey"))

plot3 <- scores %>%
  ggplot(aes(writing, math)) +
  geom_point(shape=16, color= "#1696d2", size=1.5, alpha = 0.6) +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE, color = "#fdbf11",␣
 ↪size = 0.7) +
theme_tufte()+ graph() +
  labs(title = "Math vs Writing") +
  theme(plot.title = element_text(size = 20),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.line = element_line(color = "grey"))

corr <- round(cor(scores), 1)
```

```
corplot <- ggcorrplot(corr, hc.order = TRUE, type = "upper",
   outline.col = "black",
   lab_size = 9,
   ggtheme = ggplot2::theme_minimal,
   colors = c("#db2b27","#d2d2d2", "#fdbf11"), lab = TRUE) + theme_tufte()+
   labs(title = "Correlation") +
  graph() +
   theme(legend.position = "none",
         text = element_text(size = 20),
         plot.title = element_text(size = 22, hjust = 0.5),
         axis.text.x = element_text(angle = 0, size = 20, color = "black"),
         axis.text.y = element_text(angle = 0, size = 20, color = "black"),
         axis.title.y = element_blank()) +
  guides(colour=guide_legend(keyheight= 16,label.position="bottom")) +
  geom_tile(height=0.6, width=0.6, fill = NA)


all_plots <- (plotall + corplot + plot_layout(widths = c(2,1))) / (plot1 |␣
 ↪plot2 | plot3) + plot_layout(heights = c(8, 4))
suppressMessages(all_plots)
```



- According to the logical hypothesis,there is a strong positive correlation between all three subjects, greater between reading and writing.

1.7 Summary

Test preparation courses are indeed helpful

Overall, test preparation courses helped students increase their overall score by an average of 7.7 points, the biggest improvement was observed in writing test scores.

Students who score well on one subject score well on other subjects

This indicates that a student's propensity to do well in school has less to do with their individual proclivity for certain subjects and more to do with outside factors.

A Higher Education level among parents had a positive influence on students' scores

A higher education among parents seems to have positively influenced their children's school grades. Students whose parents didn't go to college had the lowest scores on average.

1.8 Recommentations

When it comes to lunch status and income disparity, there is little the school can do.

On the positive side, the test preparation course does help students improve their grades, so the obvious and only solution is to make the test prep course mandatory for all students and part of the standard curriculum.

Additionally, the school could consider dedicating more resources to helping students in the minority group A who seem to be struggling more than most.