Final MATH 208 (Question 1)

260869949
Prsa Yadollahi
MATH 208 - Final

```
data("STAR")
STAR <- STAR[c('stark', 'star1', 'readk', 'read1', 'read2')]
STAR <- na.omit(STAR)
STAR <- tibble::rowid_to_column(STAR, "student_ID")
STAR_data = STAR
str(STAR_data)
```

```
## 'data.frame':    3114 obs. of  6 variables:
##  $ student_ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ stark     : Factor w/ 3 levels "regular","small",..: 2 2 1 2 1 1 2 2 1 3 ...
##  $ star1     : Factor w/ 3 levels "regular","small",..: 2 2 1 2 1 1 2 2 1 3 ...
##  $ readk     : int  447 450 448 447 431 451 478 455 430 437 ...
##  $ read1     : int  507 579 651 533 558 548 514 530 490 503 ...
##  $ read2     : int  568 588 614 608 608 596 569 608 622 552 ...
##  - attr(*, "na.action")= 'omit' Named int [1:8484] 1 4 5 6 7 8 9 10 14 15 ...
##   ..- attr(*, "names")= chr [1:8484] "1122" "1160" "1183" "1195" ...
```

```
STAR_data %>% slice(sample(1:n(),5))
```

```
##   student_ID        stark        star1 readk read1 read2
## 1        367        small        small   434   507   618
## 2       2972 regular+aide      regular   472   590   611
## 3       1006 regular+aide regular+aide   518   651   669
## 4       2192 regular+aide regular+aide   425   564   659
## 5        634 regular+aide      regular   425   507   557
```

(A)

```
# plot 1: barplot
# plot 2: mosaicplot
p1<-ggplot(STAR_data,aes(x=star1,fill=stark)) + geom_bar() +
  scale_fill_viridis_d() + ggtitle("Plot 1") + theme_bw()

p2<-ggplot(STAR_data) +   geom_mosaic(aes(x=product(stark,star1),fill=stark))+
  scale_fill_viridis_d() + ggtitle("Plot 2")+ theme_bw()
grid.arrange(grobs=list(p1,p2),nrow=2,ncol=1)
```

Plot 1

Plot 2
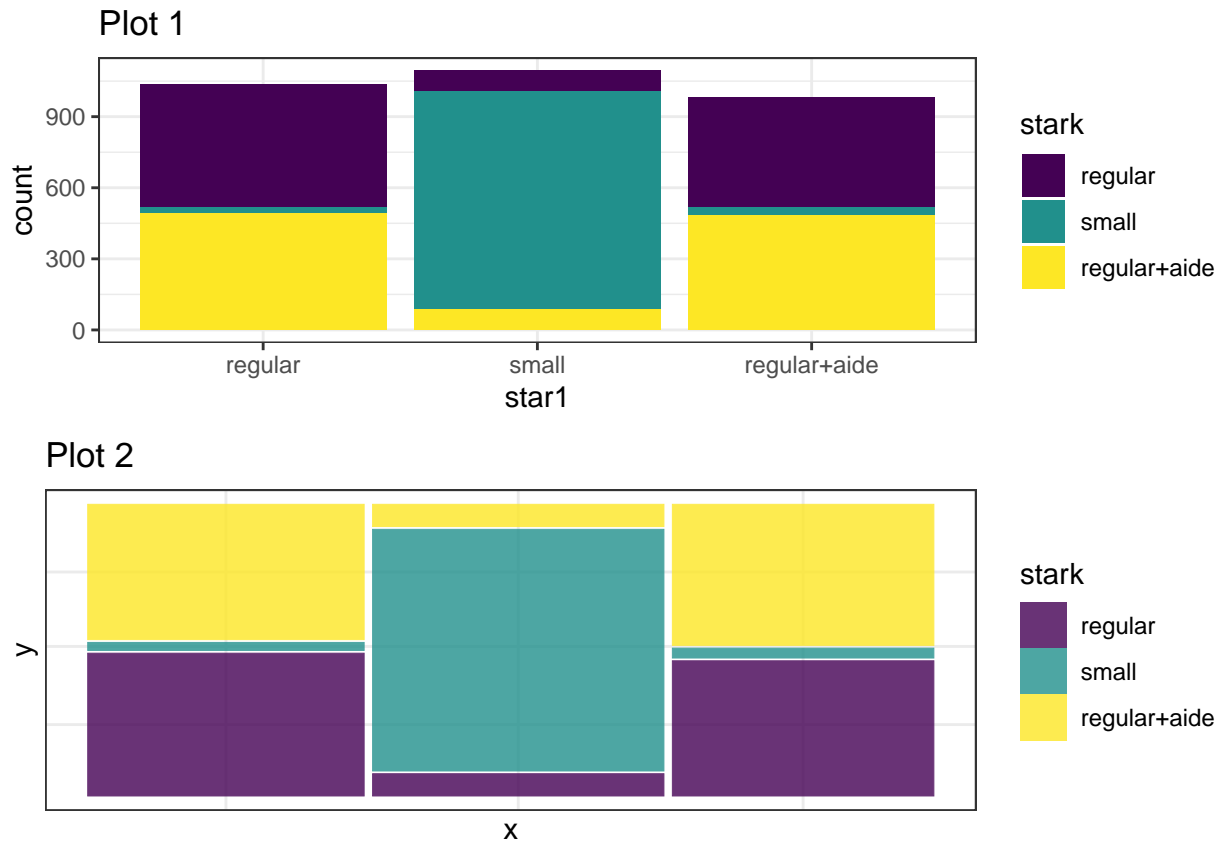
(B) Students are most likely going to have also been in either of the two kindergarten classes, given that they are in either a regular or regular-aide grade 1 class. But, they are most likely to have also been in the small kindergarten class, given that a student is in a small grade 1 class.

(C) The two differences are the following:

- The first being that both the x and y axes in mosaic plot (Plot 2) are qualitative variables whereas the bar plot (plot 1) contains qualitative x axis and quantitative y axis
- The second difference is that the mosaic plot (plot 2) represents values by ratio/proportion. Whereas the bar plot (plot 1) represents values by their counts.

(D)

```r
merge_vec = function(vec) {
  factor(
    vec,
    levels = c("regular", "small", "regular+aide"),
    labels = c("not small", "small", "not small")
  )
}
merged_plot = STAR_data %>%
  mutate(
    STAR_data,
    stark_mod = merge_vec(STAR_data$stark),
    star1_mod = merge_vec(STAR_data$star1),
```
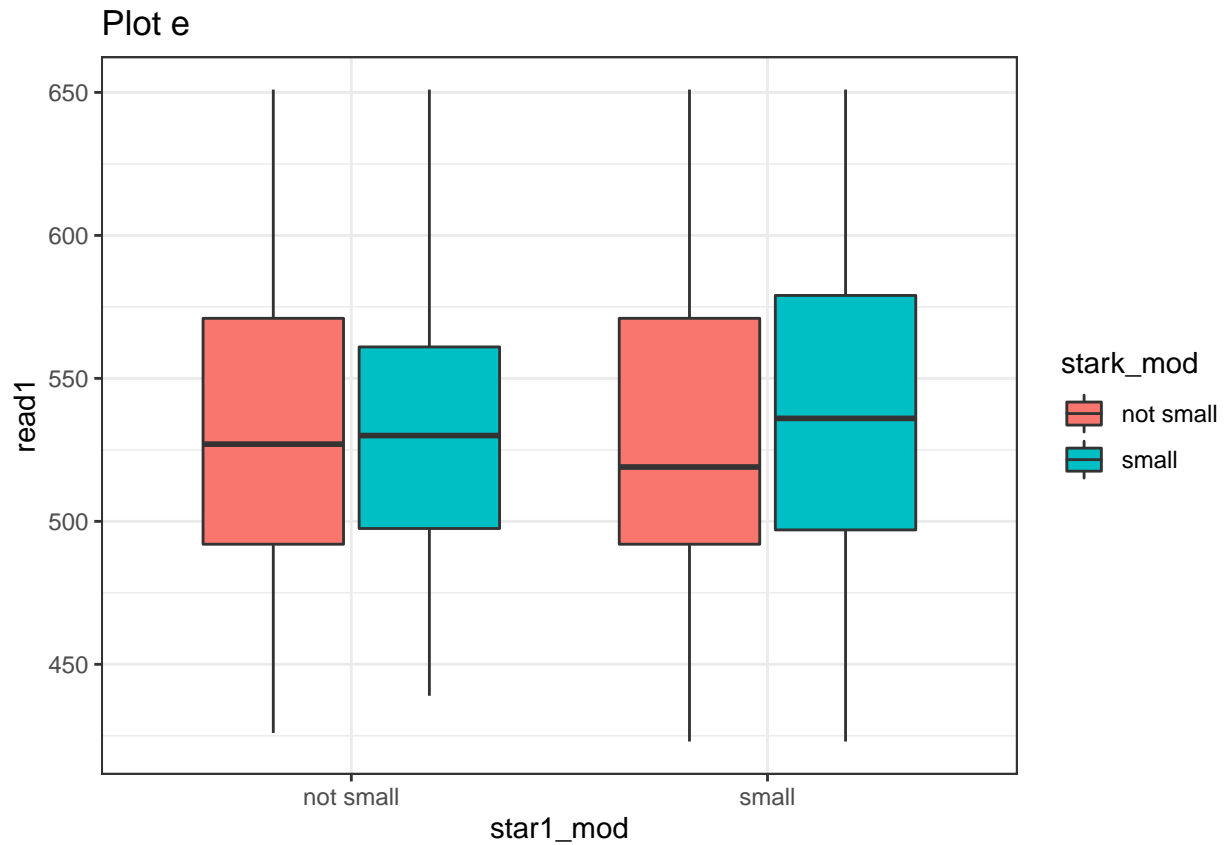
```
    stark = NULL,
    star1 = NULL
  )
```

(E)

```
ggplot(merged_plot,aes(x = star1_mod,fill = stark_mod,y=read1)) +
geom_boxplot() + ggtitle("Plot e") + theme_bw()
```
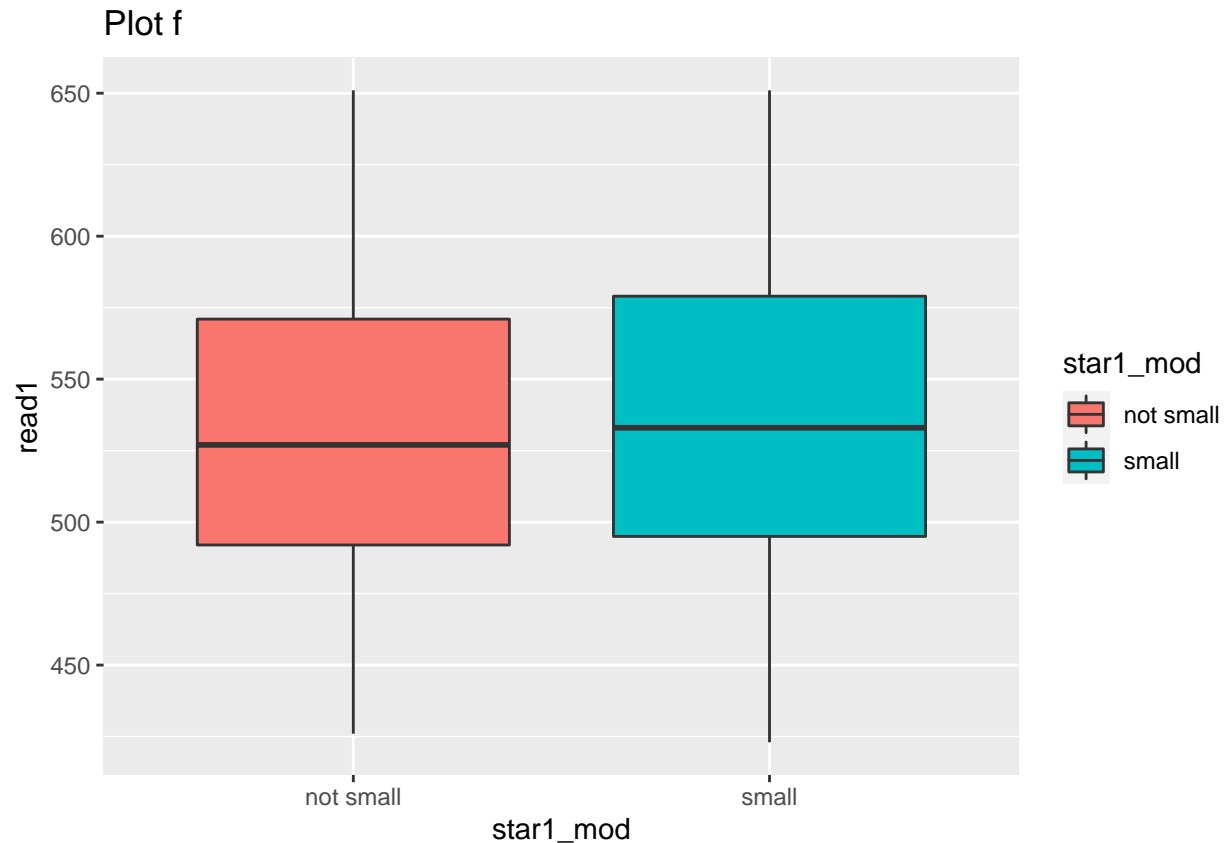

Plot e

(F)

```
plot_f = with(STAR_data, ggplot
            (merged_plot,
              aes(
                x = star1_mod,
                y = read1,
                fill = star1_mod
                )) +
              geom_boxplot() +
              ggtitle("Plot f"))
plot_f
```

Plot f

I created plot F to help visualize the question better. Based on both plots E and F above, I believe that there is a strong correlation between the modified types of class variables and the grade 1 reading test scores. For example, plot F shows a comparison between the class variables and the grade 1 reading scores. We can notice a difference in the mdeian value is bigger in the small class compared to the not small class.

(G)

- Panel g1: geom point plot
- Panel g2: density2d plot

(H) Looking at g1 and g2, we notice that there is a correlation between readk and read1. There is a linear association. For example, as readk increases, so does read1. The association does not vary by much between the modified kindergarten class type variables. In both class types, the line of best fit has a similar slope.

(I)

- A: Line chart
- B: 2-d density plot
- D: 2-d histogram