Question 2 [50 points]

We will re-use the same data that was used in Question 1. The description is repeated below for your convenience. The data for this question comes from the `STAR` dataset from the `AER` library. Below is a summary and five sample rows of a modified version of that dataset containing information from a study examining the effect of reducing class size on student performance in primary school. T

```
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 ...
```

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

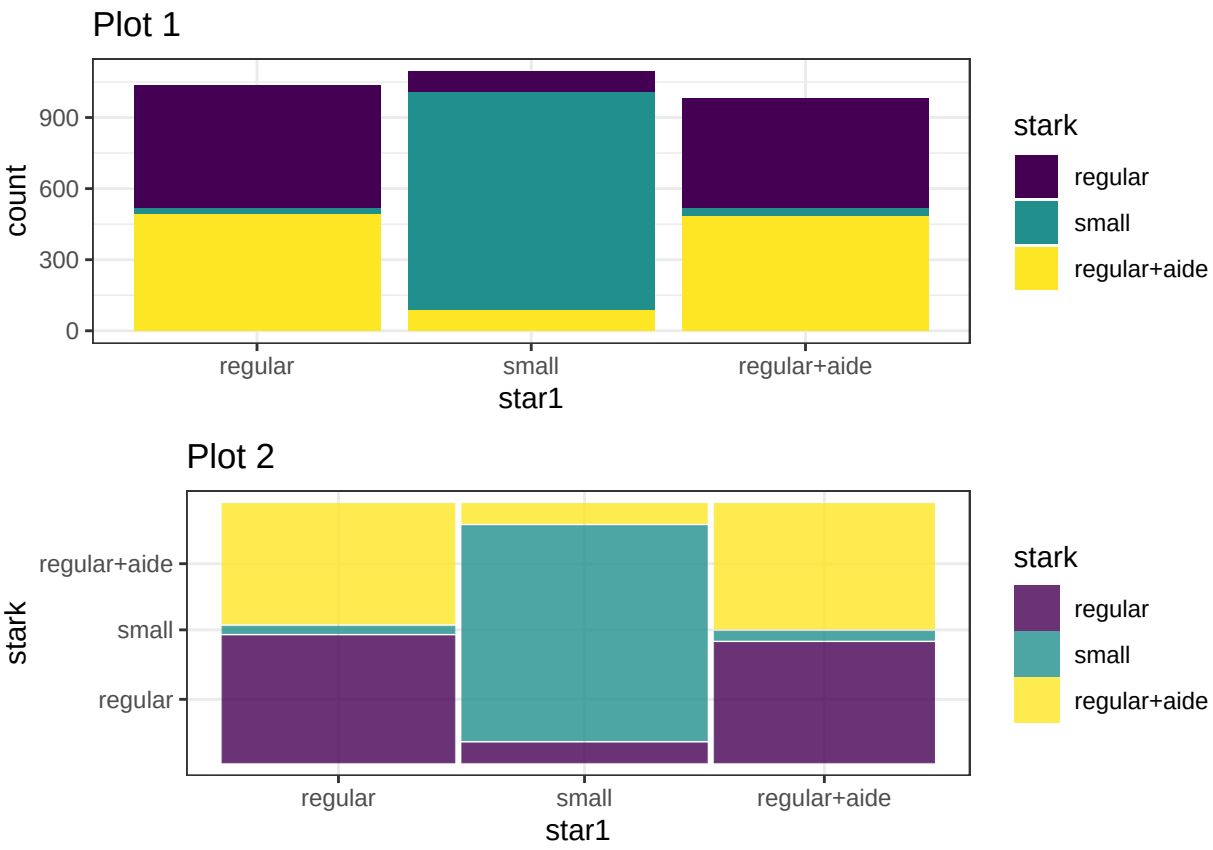
```
  student_ID   stark        star1 readk read1 read2
1       2402   small        small   470   579   599
2       2007   small        small   418   519   503
3       2380 regular regular+aide   445   533   602
4       2020 regular regular+aide   411   475   535
5       2116   small        small   434   530   539
```

Besides the Student ID, we will focus on four other measures from the data: `stark` and `star1`, which indicate the type of class in kindergarten and grade 1, respectively ("regular", "small", or "regular+aide"); and `readk`,`read1`, and `read2` which are reading scores from kindergarten, grade 1 and grade 2 respectively.

(a) **[6 pts]** Below are partially obscured code and two plots of the values of class types for kindergarten and grade 1:

```
p1<-ggplot(STAR_data,aes(x=star1,fill=stark)) + geom_YYYYYYY() +
  scale_fill_viridis_d() + ggtitle("Plot 1") + theme_bw()
p2<-ggplot(STAR_data) + geom_XXXXXXX(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)
```

CONTINUED ON NEXT PAGE



Identify these two plots by name:

Answer:

Plot 1 Plot 2

(b) **[8 pts]** Using these plots, describe the describe the association between `stark` and `star1`. In particular, what does knowing the type of grade 1 class type tell us about the possible kindergartn class type for the students in this sample?

Answer:

CONTINUED ON NEXT PAGE

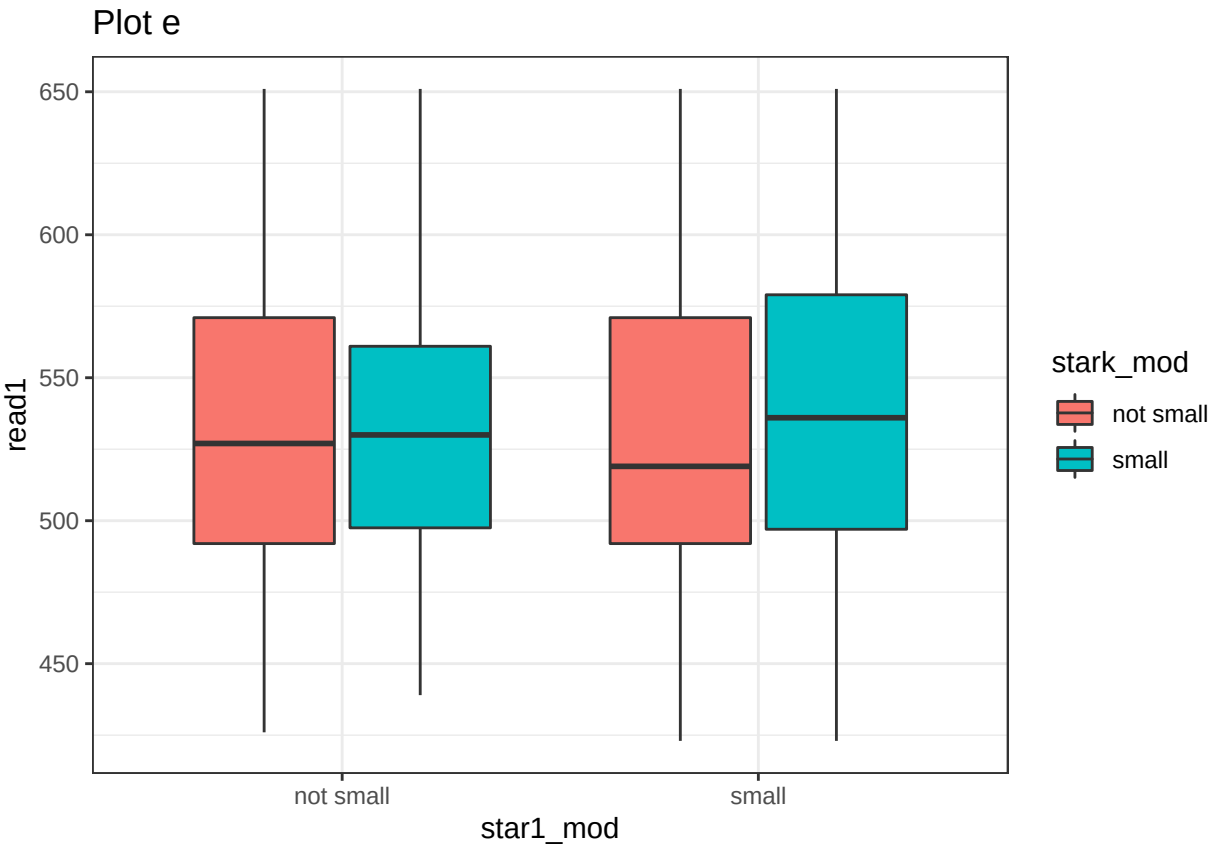
(c) **[6 pts]** Although these plots look similar, they are in fact different. There are two important differences in how these plots were constructed, one which is more obvious than the other. Explain what those two differences are.

Answer:

(d) **[6 pts]** Write a line of code to create new factor variables in `STAR_data` for `stark` and `star1` named `stark_mod` and `star1_mod` which combine the "regular" and "regular+aide" levels into a single level "not small".

Answer:

Below is a figure along with the code (partially obscured) which generated it.



```
ggplot(STAR_data,aes(x=_______,fill=________,y=read1)) +
  geom_______() +  ggtitle("Plot e") + theme_bw()
```

(e) **[4 pts]** What are the missing geometry and aesthetics that generated the figure on the previous page (that is, what are the words that are missing in the code above for Plot e)?

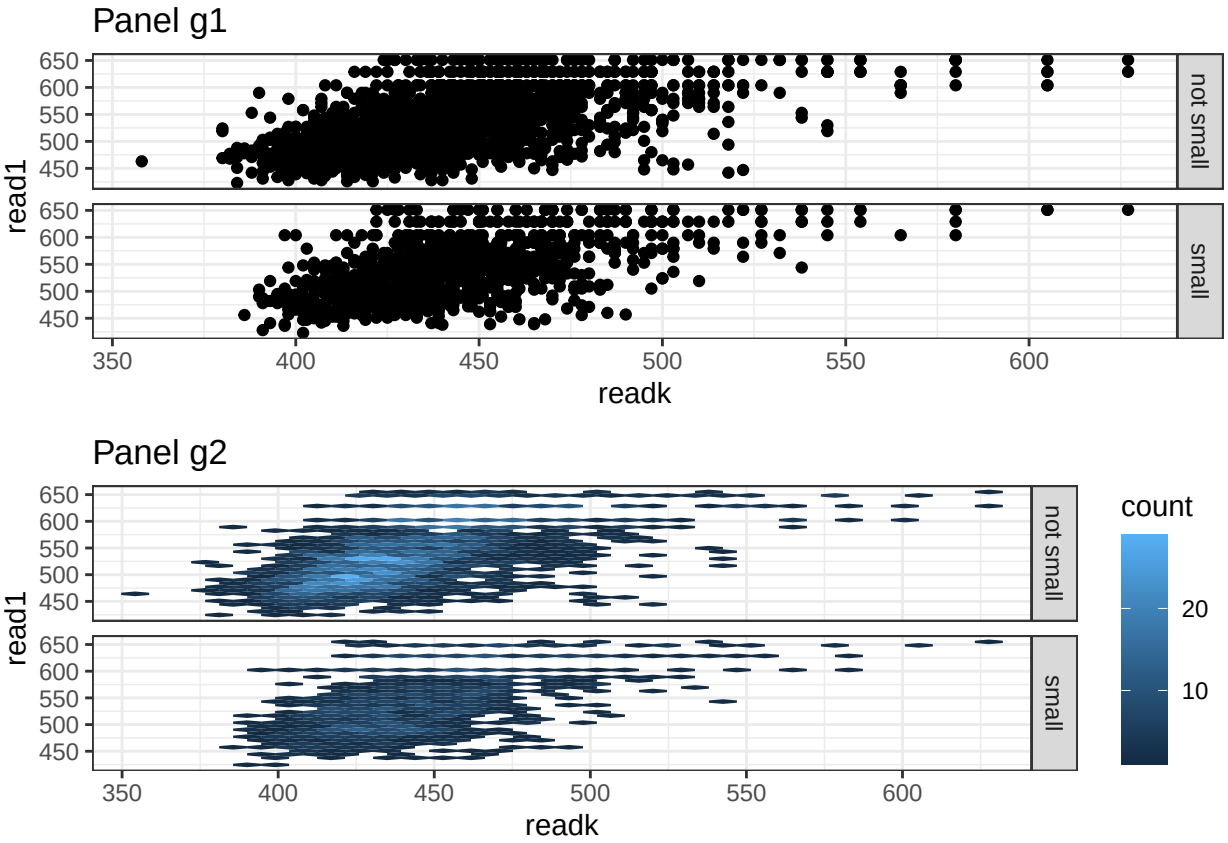
Answer:

(f) **[5 pts]** Based on these plots, do you think there is evidence of an association between the modified type of class variables and the grade 1 reading test score? Explain your answer in 3 sentences or fewer.

Answer:

CONTINUED ON NEXT PAGE

Below is a plot of the reading test scores for kindergarten and grade 1 for the `STAR_data` by levels of the modified kindergarten class type.



(g) **[4 pts]** Identify the two kinds of plots in Panel g1 and g2 by name (note that there are two of the same kind of plot in each panel)

- Panel g1:
- Panel g2:

(h) **[6 pts]** From Panels g1 and g2, would you conclude that there is an association between `readk` and `read1` in either group? Does the association between the two reading test varies seem to vary by levels of the modified kindergarten class type variable? Explain your answers in 4 sentences or fewer.

Answer:

CONTINUED ON NEXT PAGE

(i) **[5 pts]** Which of the following plots could also be used to assess the association between reading scores in kindergarten and grade 1 (assuming that neither variable is transformed)? Circle all that apply.

A. Line chart B. 2-d density plot C. Treemap D. 2-d histogram

Answer:

END OF QUESTION