Data Science in Education: HW #5

In this assignment, you will continue the analysis of the BuCKETS data from Lab 7 by fitting a longitudinal model for the PCK scale scores, conducting a missing data analysis, and considering whether the control and treatment groups are balanced on some potential confounding variables.

Setup

- Create a project in RStudio for this lab.
- Download the **BuCKETS_DATA.xlsx** data file from Canvas and move it to your project folder.
- Create a new R Markdown document with HTML specified as the output format.
- In the setup code chunk:
    - Delete the `include = FALSE` option and add the `message = FALSE` option.
    - Load the packages you will need for this assignment: `readxl`, `janitor`, `tidyverse`, `dataedu`, `nlme`, `emmeans`, `ggmosaic`, and `gridExtra`.
    - As in Lab 7, read in the data using the `read_excel` function and then use the `clean_names` function to create names that are easier to work with. Then process the data by dropping the MSSA subscale scores, converting all demographic information to character variables, re-labeling the demographic variable response options to align with the provided code book, and creating the `race`, `in_state`, and `missing_dp` variables.

Exercises

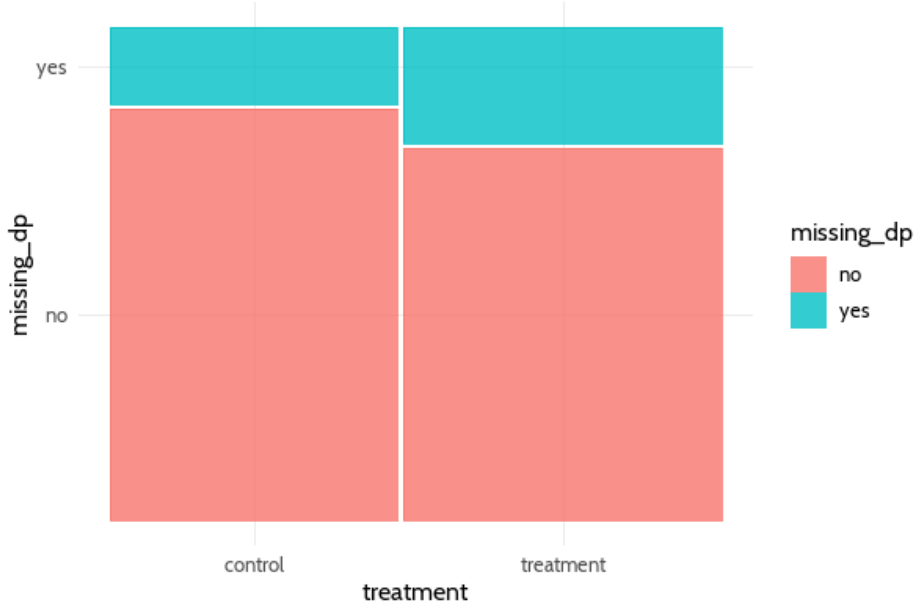
In your R Markdown file, delete the sample code and text. Then add in R code chunks and commentary as necessary to complete the exercises below. Include appropriate headings, hide any messages and warnings, and adjust font sizes in visual displays as necessary to create a polished report. When you have finished working through the exercises, knit your document and upload the resulting HTML file to the HW #5 assignment in Canvas.

1. Conduct a longitudinal analysis of the PCK scores by completing the following steps.
    - Create a data frame where the PCK scores have been pivoted to long format, with one column for time and another column for PCK scores.
    - Construct a line plot to visualize the change in mean PCK score over time separately for each group (control and treatment). Comment on what the pattern in the plot reveals about the change over time for each group.
    - Fit a linear mixed-effects model for the PCK scores with a subject-specific random effect and fixed effects for time, treatment, and their interaction using `na.omit` for the `na.action` argument. Request an ANOVA table for the fitted model.
    - Keeping in mind the principle of hierarchy, refine the model until all remaining terms are statistically significant at the 5% level. Use the `contrast` function of the `emmeans` package to determine which particular levels of the remaining variables are significantly different.
    - Include commentary throughout the model refinement and provide an interpretation of your requested contrast.

2. As noted in Lab 7, there were 8 students who did not have any delayed-post measurements. In all of our longitudinal models, we have chosen to omit the missing values, which will still use all of the available information for these students, rather than remove these students and/or all delayed-post measurements from the analysis. The approach we used to handling the missing values is only warranted if there is no pattern to the missingness; if certain characteristics of these students made them more or less likely to respond at the delayed-post time point, that would be a concern. To check this assumption, you can compare the students who did and did not complete the scales at the delayed-post time point on all other variables in the dataset.

   For example, the following code will create a mosaic plot comparing the distribution of students who were and were not missing delayed-post measurements across the treatment group assignments.

```
ggplot(data = buckets) +
  geom_mosaic(aes(x = product(missing_dp, treatment), fill = missing_dp)) +
  theme_dataedu()
```



   The plot reveals that similar proportions of the students in the control and treatment groups did not complete the delayed-post measurements. To investigate further we can use the `tabyl` function of the `janitor` package to create a two-way table giving the number of students who did and did not complete the delayed-post measurements in each treatment group.

```
trt_miss <- tabyl(buckets, treatment, missing_dp)
trt_miss

##  treatment no yes
##    control 16   3
##  treatment 16   5
```

   We could take this further by conducting a significance test to see if there is evidence of an association between missing delayed-post measurements and treatment group assignment. However, due to the small counts in the table (we need at least 10 in each cell) the theory-based chi-square test of association isn't valid. We can instead use a nonparametric version of this test known as Fisher's exact test. The results of the exact test are interpreted in the same way as those of the chi-square test. The `fisher.test` function from the `janitor` package takes a two-way tabyl as input.

```
fisher.test(trt_miss)

##
##  Fisher's Exact Test for Count Data
##
## data:  trt_miss
## p-value = 0.6984
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.2664948 12.4153174
## sample estimates:
## odds ratio
##   1.645701
```

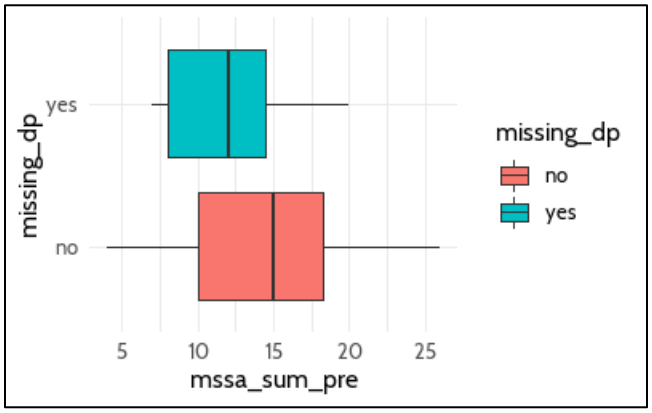
With a *p*-value of .6984, there is not much evidence of an association between treatment group assignment and missing delayed-post measurements, which supports the assumption that there is no pattern to the missingness in the dataset.

Conduct a missing data analysis on demographics by completing the following steps.

- For each of the six remaining demographic variables in the `buckets` data frame – sex, race, year, major, in_state, type_of_hs – create a mosaic plot comparing the distribution of students who were and were not missing delayed-post measurements across the different groups defined by the variable. Use the `grid.arrange` function of the `gridExtra` package to arrange the mosaic plots in 3 rows. Comment on any potential associations.
- For each of these six remaining demographic variables, construct a two-way tabyl giving the number of students who did and did not complete the delayed-post measurements in each group defined by the variable. Based on the results, use either the `chisq.test` or the `fisher.test` function of the `janitor` package to conduct a test of association. Comment on what the *p*-values reveal about the strength of evidence for any associations and what this says about the missing data assumptions that were made in the longitudinal analysis.

3. As an additional check on the missing data assumptions that were made, the students who were and were not missing delayed-post measurements can be compared on each of the scale scores. For example, the following code will create side-by-side boxplots comparing the distribution of MSSA pre-test scores among the students who were and were not missing delayed-post measurements.

```
ggplot(buckets, aes(x = mssa_sum_pre, y = missing_dp, fill = missing_dp)) +
  geom_boxplot() +
  theme_dataedu()
```



The boxplots reveal that the missing data group had a slightly lower median MSSA pre-test score than the group who completed all measurements. We could conduct a significance test to see if the mean difference between the groups is statistically significant. Though the missing data group is small ($n = 8$), the boxplots do not reveal strong skewness such that use of the theory-based two-sample t-test should be permissible.

```
t.test(mssa_sum_pre ~ missing_dp, data = buckets)

##
##  Welch Two Sample t-test
##
## data:  mssa_sum_pre by missing_dp
## t = 1.0148, df = 11.682, p-value = 0.3308
## alternative hypothesis: true difference in means between group no and group y
es is not equal to 0
## 95 percent confidence interval:
##  -2.343078  6.405578
## sample estimates:
##  mean in group no mean in group yes
##          14.40625          12.37500
```

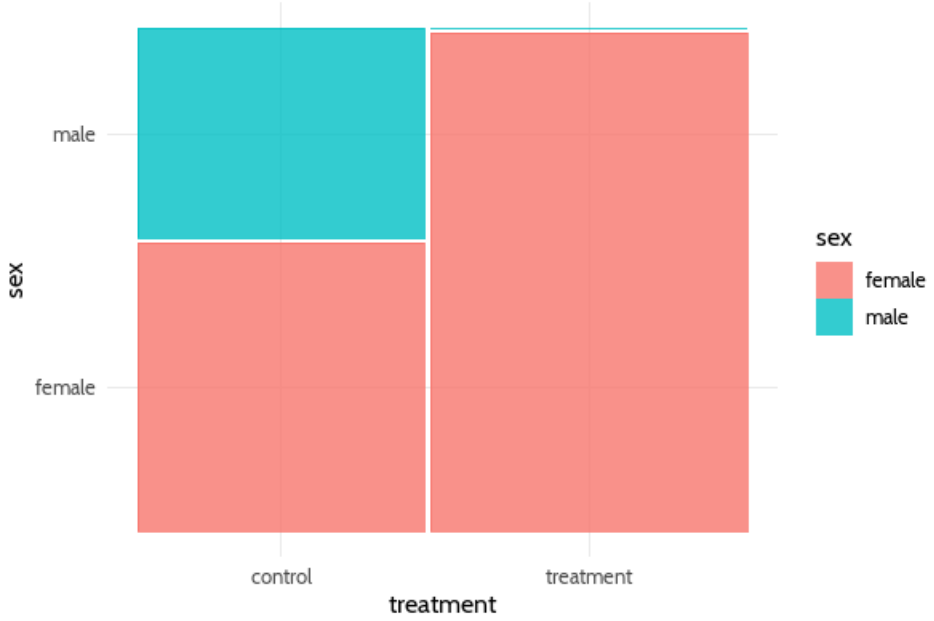
With a p-value of .3308, there is not much evidence of a difference in mean MSSA pre-test scores for the missing data groups, which supports the assumption that there is no pattern to the missingness in the dataset.

Conduct a missing data analysis on scale scores by completing the following steps.

- For each of the five remaining pre-test and post-test scores in the `buckets` data frame – mssa_sum_post, petas_s_sum_pre, petas_s_sum_post, pck_sum_pre, pck_sum_post – create side-by-side boxplots comparing the distribution of scores among students who were and were not missing delayed-post measurements. Use the `grid.arrange` function of the `gridExtra` package to arrange the mosaic plots in 3 rows. Comment on any potential associations.
- For each of the five remaining pre-test and post-test scores, conduct a two-sample t-test to see if the mean difference in scores between the students who were and were not missing delayed-post measurements is statistically significant. Comment on what the p-values reveal about the strength of evidence for any associations and what this says about the missing data assumptions that were made in the longitudinal analysis.

4. Recall that this was a quasi-experiment because even though treatments were imposed, they were not assigned to the students at random. In such cases, we would like to see that the two treatment groups are balanced with respect to all variables that could potentially influence the response. For example, the following code compares the control and treatment groups on their distribution of the sexes.

```
ggplot(data = buckets) +
  geom_mosaic(aes(x = product(sex, treatment), fill = sex)) +
  theme_dataedu()
```



The plot reveals an issue – the treatment group is composed of all females but the control group has more of a balance of males and females. Thus, it is not possible to determine if differences observed in the responses are due to the treatments imposed or the sex of the students. For example, males could be inherently worse at science content knowledge than females, so the significantly lower MSSA scores at post-test for the control group could be due to the fact that this group contains so many males and not because of the treatment.

The following code creates a two-way tabyl and conducts an exact test for the association between sex and treatment group assignment.

```
trt_sex <- tabyl(buckets, treatment, sex)
trt_sex

##  treatment female male
##    control     11    8
##  treatment     21    0

fisher.test(trt_sex)

##
##  Fisher's Exact Test for Count Data
##
## data:  trt_sex
## p-value = 0.0009828
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0000000 0.3900502
## sample estimates:
## odds ratio
##          0
```

Not surprisingly, the *p*-value of .00098 provides strong evidence of an association between sex and treatment group assignment.

Compare the treatment and control groups on the remaining demographic variables as follows.

- For each of the five remaining demographic variables in the `buckets` data frame – race, year, major, in_state, type_of_hs – create a mosaic plot comparing the distribution of the groups defined by the variable between the control and treatment groups. Use the `grid.arrange` function of the `gridExtra` package to arrange the mosaic plots in 3 rows. Comment on any potential associations.
- For each of the five remaining demographic variables, construct a two-way tabyl giving the number of students who were in the treatment and control groups for each group defined by the variable. Based on the results, use either the `chisq.test` or the `fisher.test` function of the `janitor` package to conduct a test of association. Comment on what the *p*-values reveal about the strength of evidence for any associations and what this says about potentially confounding variables.

5. Return to the line plot that you created in exercise 1 to visualize the change in mean PCK score over time separately for the control and treatment groups.
   - What from the comparison of control and treatment groups in exercise 4 might explain the drastic difference in PCK (pedagogical content knowledge) between the two groups at pre-test?
   - What from the missing data analysis in exercise 3 might explain the dramatic drop in PCK scores from post-test to delayed-post for the treatment group?