eda

Kate Colvin

2024-12-11

Data Cleaning

Reading in data, creating summary table

```
score_df <- read_csv("Project_1_data.csv") %>%
  janitor::clean_names() %>%
  mutate(wkly_study_hours =
           case_match(
             wkly_study_hours,
             "< 5" ~ "< 5",
             "> 10" ~ "> 10",
             "10-May" ~ "5-10"),
         wkly_study_hours = factor(wkly_study_hours, c("< 5", "5-10", "> 10")))
```

```
## Rows: 948 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMarita...
## dbl  (4): NrSiblings, MathScore, ReadingScore, WritingScore
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

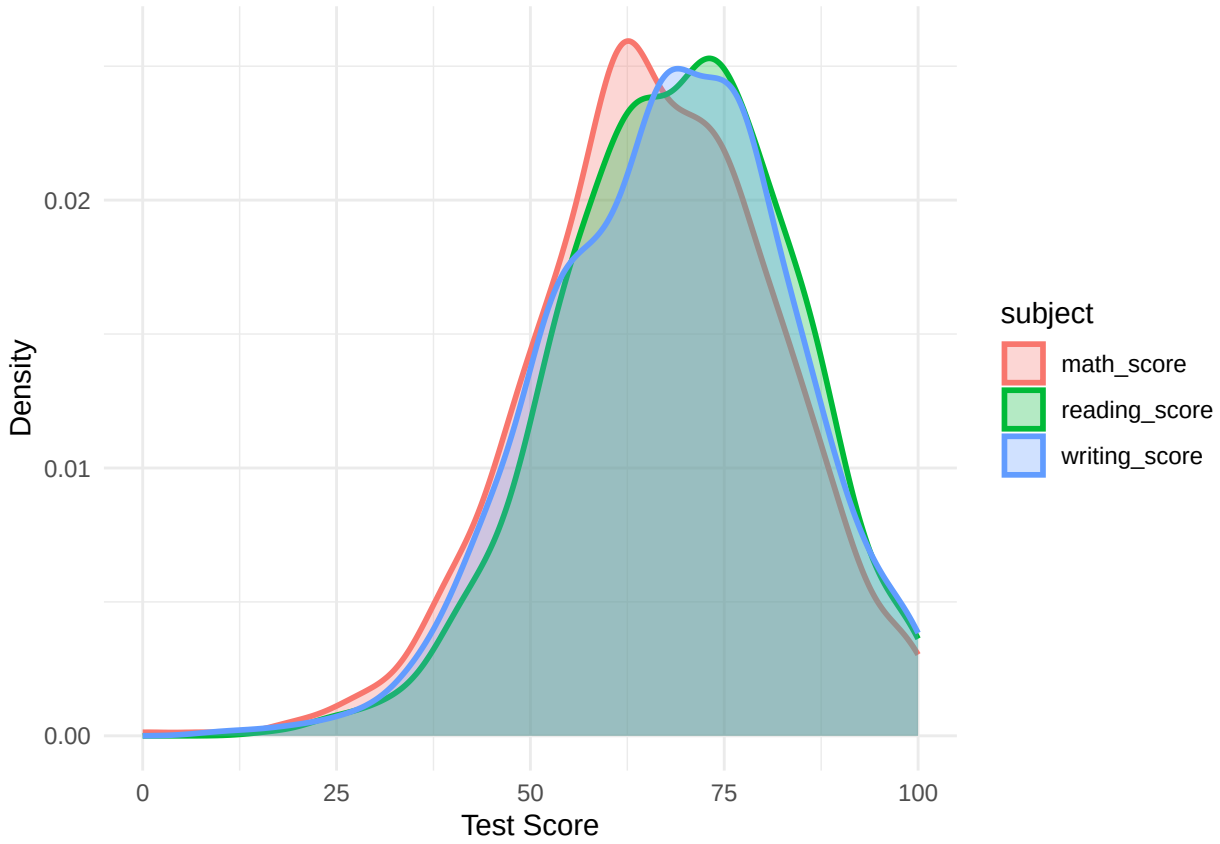
```
table1 <- score_df %>%
  tbl_summary()
table1
```

Visualizations

Plotting the overall distribution of scores, stratified by subject

```
score_long_df <- score_df %>%
  pivot_longer(cols = math_score:writing_score,
               names_to = "subject",
               values_to = "score")

score_dists_plot <- score_long_df %>%
  ggplot(aes(x = score, fill = subject, color = subject)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Test Score", y = "Density") +
  theme_minimal()
score_dists_plot
```



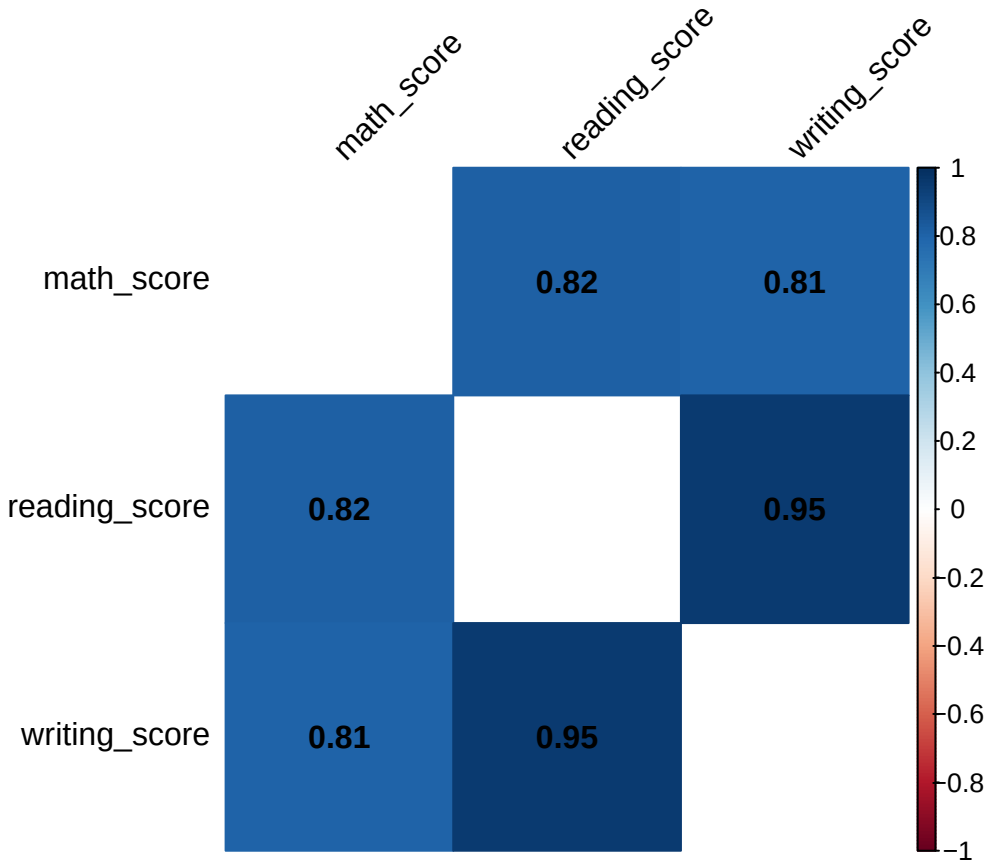

```
# Test for differences

score_test <- score_long_df %>% select(subject, score) %>%
  tbl_summary(by = subject) %>%
  add_overall() %>%
  add_p()
score_test
```

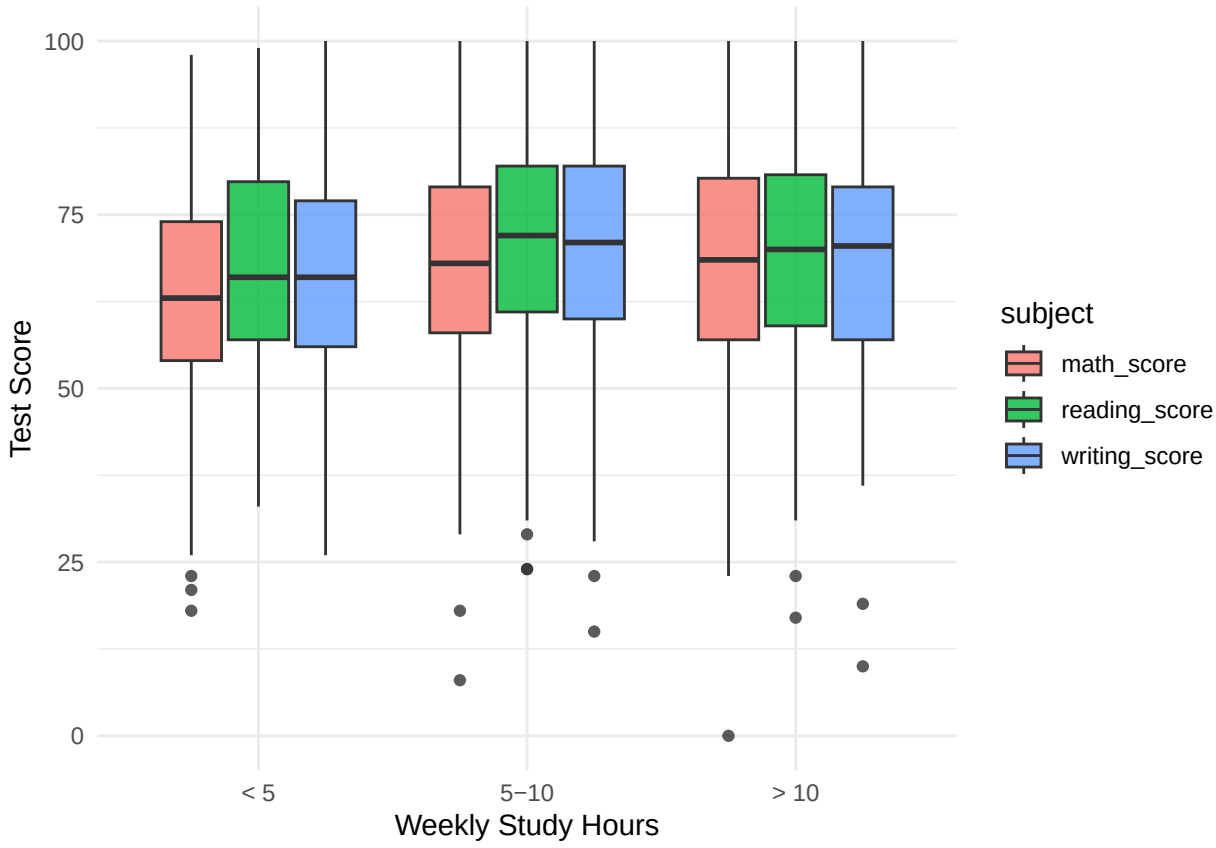
Correlation between test scores

```
cor_matrix <- score_df %>%
  select(math_score, reading_score, writing_score) %>%
  cor()

corrplot(cor_matrix,  method = "color",
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 45,
         order = 'hclust',
         diag = F)
```




```
score_long_df %>% na.omit(wkly_study_hours) %>%
  ggplot(aes(x = wkly_study_hours, y = score, fill = subject)) +
  geom_boxplot(alpha = 0.8) +
  labs(x = "Weekly Study Hours", y = "Test Score") +
  theme_minimal()
```


100
75
50
25
0
Test Score
< 5
5−10
> 10
Weekly Study Hours
subject
math_score
reading_score
writing_score

Characteristic	**N = 948**[1]
gender	
female	488 (51%)
male	460 (49%)
ethnic_group	
group A	80 (9.0%)
group B	171 (19%)
group C	277 (31%)
group D	237 (27%)
group E	124 (14%)
Unknown	59
parent_educ	
associate's degree	198 (22%)
bachelor's degree	104 (12%)
high school	176 (20%)
master's degree	55 (6.1%)
some college	199 (22%)
some high school	163 (18%)
Unknown	53
lunch_type	
free/reduced	331 (35%)
standard	617 (65%)
test_prep	
completed	322 (36%)
none	571 (64%)
Unknown	55
parent_marital_status	
divorced	146 (16%)
married	516 (57%)
single	213 (24%)
widowed	24 (2.7%)
Unknown	49
practice_sport	
never	112 (12%)
regularly	343 (37%)
sometimes	477 (51%)
Unknown	16
is_first_child	604 (66%)
Unknown	30
nr_siblings	
0	101 (11%)
1	245 (27%)
2	213 (24%)
3	198 (22%)
4	76 (8.4%)
5	50 (5.5%)
6	8 (0.9%)
7	11 (1.2%)
Unknown	46
transport_means	

Characteristic	**Overall** N = 2,844[1]	**math_score** N = 948[1]	**reading_score** N = 948[1]	**writing_**
score	68 (57, 78)	66 (56, 76)	70 (59, 80)	6

[1]Median (Q1, Q3)
[2]Kruskal-Wallis rank sum test