ETC1010-5510: Introduction to Data Analysis

Aditya Santosh Rao

Assignment 2

In this assignment, we are going to study test scores from the The Programme for International Student Assessment (PISA) scheme, which tests 15-year-old students across all states in Australia. You can find the data sets and a code book for the assignment in the *Data* folder. Broadly speaking, PISA measures scholastic ability across three categories: science, reading and math.

The goal of this analysis is to understand if differences exist between PISA testing scores across various dimensions, such as income, school type, extra-curricular activities and gender.

Please ensure that the report knits properly into html and all the R code and R outputs are visible in the final knitted report. You will need to save your rendered html document into a pdf file (you can use your internet browser to print your html file into a pdf file) and upload that pdf file into Moodle for submission.

This is an individual assignment and you must use R code to answer all the questions. Make sure that you have your messages and warnings turned off before you submit the assessment (see lines 15-17 of this Rmd file) and echo = FALSE set for the R code chunk where you load your libraries.

Question 1: Read in the pisa data set (1pt) and show the last 5 rows and last three variables in the data frame (1pt).

```
pisa <- read.csv("data/pisa_au.csv")

# Displaying the last 5 rows and last 3 variables
tail(pisa[, (ncol(pisa)-2):ncol(pisa)], 5)
```

```
##          math  science     read
## 14526 578.2601 578.6234 596.3567
## 14527 351.0943 315.8658 381.5746
## 14528 639.1595 664.5759 583.6448
## 14529 707.5041 767.9710 694.8842
## 14530 572.2217 583.7702 534.1746
```

- **The above table displays the last 5 rows and last three variables in the data frame.**

Question 2: Calculate the 75th quantile for `math`, `science` and `read` score across states. (2pts) Create a table where you display the results (1pt). Which

state has the highest score in `read`. (2pts)

```
# Calculate the 75th quantile for math, science, and read scores across states
quantiles_by_state <- pisa |>
  group_by(state) |>
  summarise(
    math_75th = quantile(math, 0.75, na.rm = TRUE),
    science_75th = quantile(science, 0.75, na.rm = TRUE),
    read_75th = quantile(read, 0.75, na.rm = TRUE)
  )

print(quantiles_by_state)
```

```
## # A tibble: 8 × 4
##   state math_75th science_75th read_75th
##   <chr>     <dbl>        <dbl>     <dbl>
## 1 ACT        563.         598.      581.
## 2 NSW        550.         576.      569.
## 3 NT         521.         549.      541.
## 4 QLD        533.         561.      553.
## 5 SA         540.         572.      562.
## 6 TAS        524.         549.      538.
## 7 VIC        553.         575.      567.
## 8 WA         556.         584.      569.
```

```
# State with the highest 75th quantile score in read
state_highest_read <- quantiles_by_state |>
  filter(read_75th == max(read_75th)) %>%
  select(state, read_75th)

print(state_highest_read)
```

```
## # A tibble: 1 × 2
##   state read_75th
##   <chr>     <dbl>
## 1 ACT        581.
```

- **The highest 75th percentile score in 'read' is from the Australian Capital Territory (ACT), with a score of 580.80.**
- **The Australian Capital Territory (ACT) also shows the highest maximum read score at 783.85**

Question 3: For female students born after or on 2000, which type of school had the highest average

performce in `read` (3pts).

```
table1 <- pisa %>% filter(gender == 'female' & birthyr >= 2000) %>% group_by(schtype) %>% summar
ise(high_avg_read = mean(read)) %>% arrange(desc( high_avg_read))
table1
```

```
## # A tibble: 3 × 2
##   schtype  high_avg_read
##   <chr>            <dbl>
## 1 Ind               548.
## 2 Catholic          523.
## 3 Gov               485.
```

- **No female student was born after the year 2000, with a 'read' score of 648.25, the 'Ind' school type has the greatest average performance.**

Question 4: For male and female students born on or after 2000, across the different types of schools, whose scores were more variable in reading? (3pts). Math (1pt)?. Science (1pt)? Place all results in a single table to receive full marks.

```
pisa |> group_by(schtype, gender) |> summarise(across(c(math, science, read), ~IQR(.), .names =
"IQR_{.col}"))
```

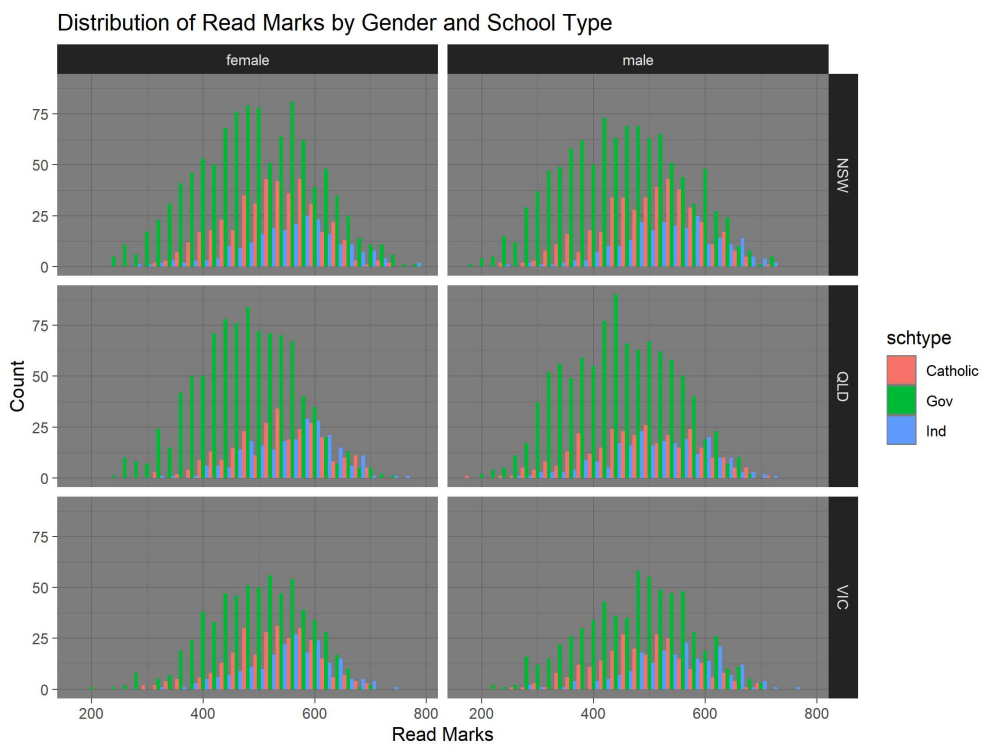
```
## # A tibble: 6 × 5
## # Groups:   schtype [3]
##   schtype  gender IQR_math IQR_science IQR_read
##   <chr>    <chr>     <dbl>       <dbl>    <dbl>
## 1 Catholic female     108.        128.     113.
## 2 Catholic male       124.        139.     124.
## 3 Gov      female     120.        142.     135.
## 4 Gov      male       132.        159.     149.
## 5 Ind      female     103.        106.     104.
## 6 Ind      male       116.        126.     119.
```

- **In all three subjects, males attending 'Gov' school types exhibit the largest interquartile range, reflecting a greater spread and higher variability in their scores.**

Question 5: For the states of VIC, NSW, and QLD, using `geom_histogram` plot the distribution of marks in `read` by gender using a faceted plot with shading

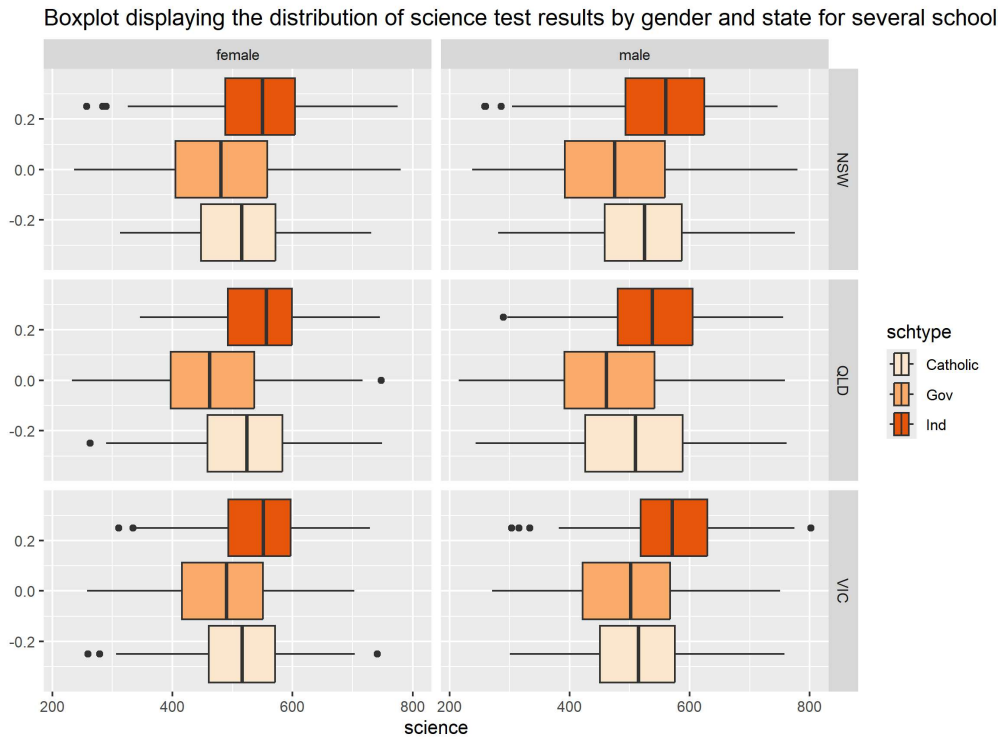
to capture school type (2pt). Which combination of states/schools do female students peform worst in (1pt)? Are the results similar for `science` (1pt)?

```
#Filtering data for the 3 states
filtered_data <- pisa %>%
  filter(state %in% c("VIC", "NSW", "QLD"))

# Plot the distribution of 'read' marks by gender, faceted by state, with shading by school type
ggplot(filtered_data, aes(x = read, fill = schtype)) +
  geom_histogram(binwidth = 20, position = "dodge") +
  facet_grid(state ~ gender) +
  labs(title = "Distribution of Read Marks by Gender and School Type",
       x = "Read Marks",
       y = "Count") +
  theme_dark()
```



- **The distribution of read scores among females throughout the states is represented by the left side of the faceted plots, which are based on the box plot above.**
- **“Gov” school categories had the lowest 50th percentile score in all three states.**
- **For NSW and QLD, the dispersion is reasonably symmetric, while for VIC, it is negatively skewed and has more outliers. When all the “Gov” school kinds are combined in VIC, female performance is often worse.**

```
pisa %>% filter(state %in% c('VIC','NSW','QLD')) %>%
  ggplot(aes(x = science, fill = schtype)) +
  geom_boxplot() +
  facet_grid(state~gender) +
  scale_fill_brewer(palette = "Oranges") +
  ggtitle("Boxplot displaying the distribution of science test results by gender and state for s
everal school types")
```

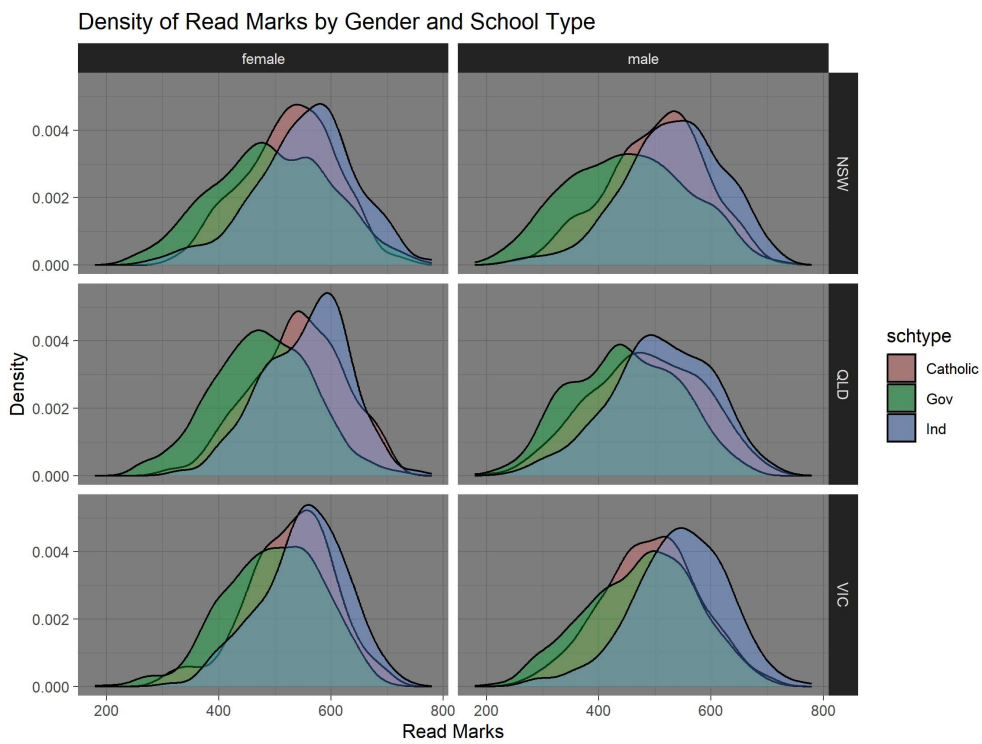


- **The outcomes for science scores are comparable. The 50th percentile ratings for "gov" school categories remain the lowest.**
- **The primary disparities in science and reading results will be found in data distribution, particularly in NSW.**
- **It is interesting to note that QLD has a little positively biased distribution while having the lowest median score.**

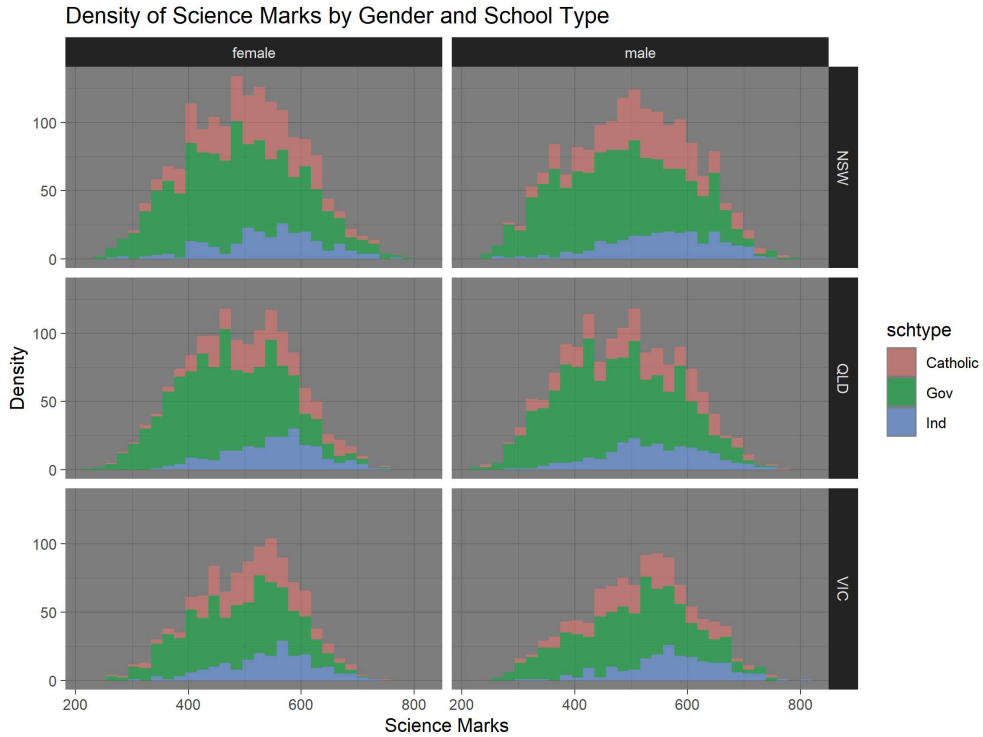
Question 6: Repeat the above exercise using `geom_density` (1pt). Which set of results, those for `geom_density` or `geom_histogram` allows one to more

accurately compare across results (3pt)? Why?

```
# Plot the density of 'read' marks by gender, faceted by state, with shading by school type
ggplot(filtered_data, aes(x = read, fill = schtype)) +
  geom_density(alpha = 0.35) +
  facet_grid(state ~ gender) +
  labs(title = "Density of Read Marks by Gender and School Type",
       x = "Read Marks",
       y = "Density") +
  theme_dark()
```




```
# Plot the density of 'science' marks by gender, faceted by state, with shading by school type
ggplot(filtered_data, aes(x = science, fill = schtype)) +
  geom_histogram(alpha = 0.5) +
  facet_grid(state ~ gender) +
  labs(title = "Density of Science Marks by Gender and School Type",
       x = "Science Marks",
       y = "Density") +
  theme_dark()
```

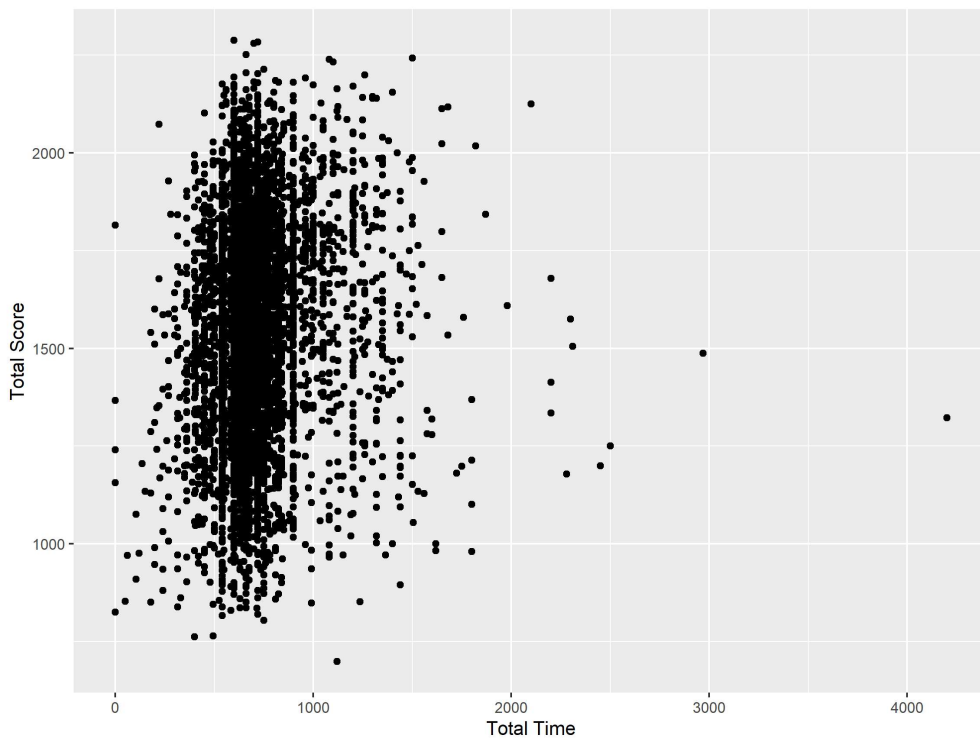


- **Density plots provide a smooth estimate of the distribution, making it easier to see the overall shape and compare peaks across groups.They show where scores are concentrated, offering a clearer view of central tendencies and variability.**
- **The overlapping areas in density plots make it easy to compare different groups, especially when looking at where the densities differ.It's easier to identify shifts in performance between school types and states.**
- **Histograms can be less smooth and harder to interpret when comparing across multiple groups due to the discrete bins.They are useful for understanding the exact distribution of scores within specific ranges.**
- **Hence, Density plots are easier to compare the information from a visual standpoint, but histograms are more accurate.**

Question 7: First, create a data frame called *pisa_filtered* that excludes observations with missing values **among any variable** (2pt). Then, calculate a new variable called *tot_score* that is the sum of the math/science/reading scores and add this to the data frame *pisa_filtered* (1pt); in addition, calculate a new variable *tot_time* as the sum of the math_time/read_time/science_time and add this new

variable to the data frame *pisa_filtered* (1pt). Using a `geom_point()` describe the relationship between *tot_time* (x-axis) and *tot_score* (y-axis). (2pts)

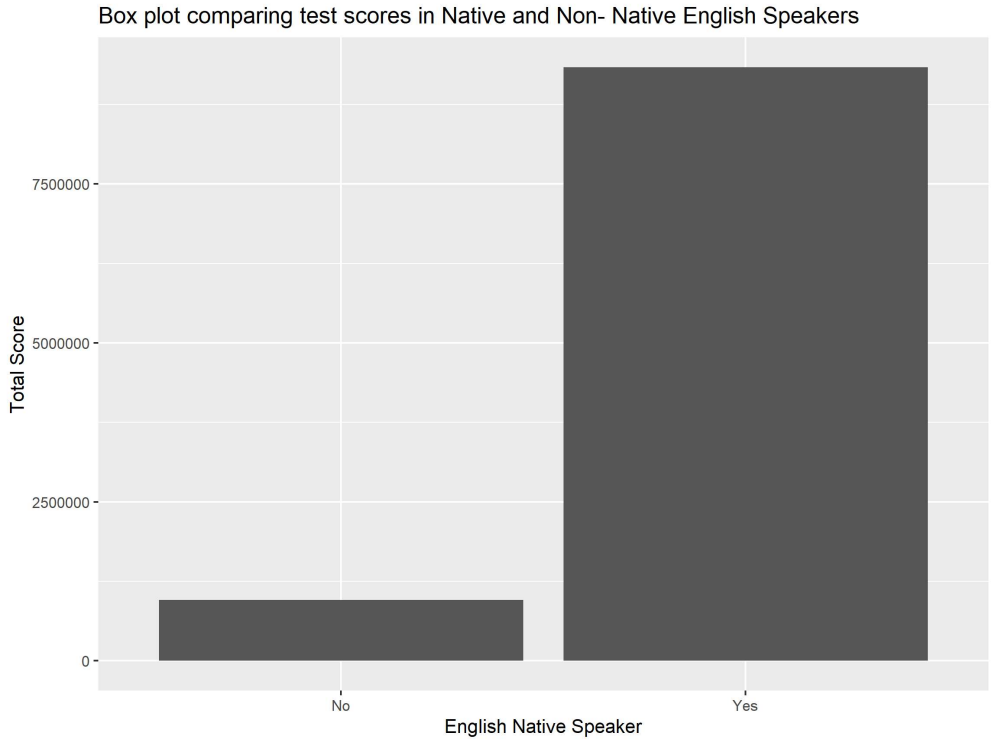
```
pisa_filtered <- na.omit(pisa)

# Calculate tot_score and add to pisa_filtered
pisa_filtered$tot_score <- pisa_filtered$math + pisa_filtered$science + pisa_filtered$read

# Calculate tot_time and add to pisa_filtered
pisa_filtered$tot_time <- pisa_filtered$math_time + pisa_filtered$read_time + pisa_filtered$scie
nce_time

# Plot tot_time vs tot_score
ggplot(pisa_filtered, aes(x = tot_time, y = tot_score)) +
  geom_point() +
  labs(x = "Total Time", y = "Total Score")
```



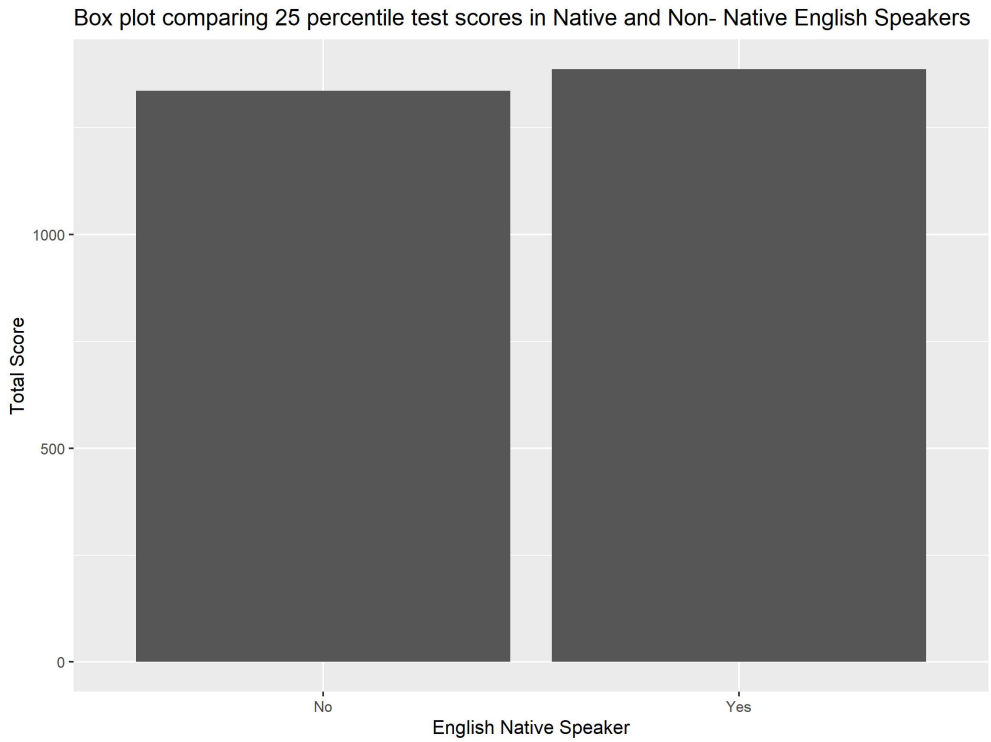
- **Most students have a similar total score around 800, as indicated by the dense clustering of data points horizontally.**
- **There is greater variation in the total time taken by students, as evidenced by the larger vertical spread of data points compared to the horizontal spread.**
- **Some students performed much better despite taking approximately the same amount of time as others, which are visible as outliers from the main cluster.**
- **The lack of a clear pattern or trend between total time and total score suggests that other factors, beyond just time taken, contribute significantly to the variation in scores.**

Question 8. The Australian PISA test is administered in English. It is believed that, on average, students who speak languages other than english at their primary residence may be disadvantaged by having to take the test in English. The variable `language` records the language most often spoken at students homes, with language code 313 referring to `English`. On average, do students who do not speak English in their home perform worse than native english speakers in terms of total scores? Does your answer remain the same when we look at students who perform in the lowest 25th quantile?

```
#Filter for English native speakers with language code 313
pisa_filtered <- pisa_filtered %>% mutate(english_native_speaker = ifelse(language == '313', 'Ye
s', 'No'))
# Box plot for total scores by language group
ggplot(pisa_filtered) +
  geom_bar(aes(x = english_native_speaker, y = `tot_score`),
           stat = 'identity') +
  labs(title = "Box plot comparing test scores in Native and Non- Native English Speakers", x =
"English Native Speaker", y = "Total Score")
```



```
#Filtering for 25th quantile test scores in Native and Non- Native English Speakers
pisa_summary <- pisa_filtered %>% group_by(english_native_speaker) %>% summarise(p25_tot_score =
quantile(`tot_score`, 0.25))

#Box plot comparing 25 percentile test scores in Native and Non- Native English Speakers
ggplot(pisa_summary, aes(x = english_native_speaker, y = p25_tot_score))+ geom_bar(stat = 'ident
ity') + labs(title = "Box plot comparing 25 percentile test scores in Native and Non- Native Eng
lish Speakers", x = "English Native Speaker", y = "Total Score")
```



- **On average, students who speak English at home perform significantly better on the PISA test compared to those who do not speak English at home.**
- **The score differences between native English speakers and non-native speakers are negligible for students in the bottom 25th percentile.**

Question 9. Using `facet_wrap()`, plot the total scores using densities for both males and females. What do the results tell us about the usefulness of looking at

average scores?

```
# Calculate mean total score overall and by gender
mean_tot_score <- mean(pisa_filtered$tot_score, na.rm = TRUE)
mean_tot_score_by_gender <- pisa_filtered %>%
  group_by(gender) %>%
  summarise(mean_tot_score = mean(tot_score, na.rm = TRUE))

ggplot(pisa_filtered, aes(x = tot_score, fill = gender)) +
  geom_density(alpha = 0.75) +
  geom_vline(xintercept = mean_tot_score, linetype = "dashed", color = "black") +
  facet_wrap(~gender) +
  labs(x = "Total Score", y = "Density", title = "Total Score Densities by Gender") +
  geom_vline(data = mean_tot_score_by_gender, aes(xintercept = mean_tot_score), color = "blue",
linetype = "dotted")
```

- **The density plot shows a higher peak (mode) for females, indicating more female students with the same score.**
- **Both male and female score distributions appear approximately normal, with no significant skewness.**
- **Males have a marginally higher average score than females.**
- **The similar distribution shape suggests that both groups have a comparable range and variability in test scores.**
- **The vertical line representing the average score is not particularly useful, as the averages for males and females are very close.**

Question 10. Previous analysis has so yielded several categorical variables that may influence test scores, such as *gender*, *music_instr*, and what type of school the student is from, *schtype*. However, there are many other numerical variables that we may have missed. As a first step in understanding these variables, we can cluster the data to try and see if any patterns emerge beyond those already seen. Use k-means clustering on the following variables with $k = 5$ clusters: *anxtest*, *motivat*, *tot_time*, and *tot_score*.(3pt) Report the number of elements in each cluster (2pt).

```
# K-means clustering
set.seed(6)
km_clusters <- kmeans(pisa_filtered[, c("anxtest", "motivat", "tot_time", "tot_score")], centers
= 5)

# Number of elements in each cluster
table(km_clusters$cluster)
```

```
##
##    1    2    3    4    5
## 1424  453 2567 1968  179
```

- **With a random seed of 6 and using a 5-cluster solution for k-means clustering on the variables anxtest, motivat, tot_time, and tot_score, the observations are distributed across the clusters as follows: Cluster 1 has 1424 observations, Cluster 2 has 453 observations, Cluster 3 has 2567 observations, Cluster 4 has 1968 observations, and Cluster 5 has 179 observations.**

Question 11. To understand the clusters, produce a table that displays the median value of each variable in the cluster, and arrange the values from largest to smallest by *tot_score* (2pt). Describe the relationship between the clusters and the variables *tot_score*,

anxtest, *motivat*; i.e., what do you notice about the similarity between these variables and the assigned cluster? (3pt)

```
# Set the seed for reproducibility
set.seed(6)

# Perform k-means clustering with k=5
km_clusters <- kmeans(pisa_filtered[, c("anxtest", "motivat", "tot_time", "tot_score")], centers
= 5)

# Create a new data frame with cluster assignments
kmeans_5_df <- data.frame(Cluster = km_clusters$cluster, pisa_filtered)

# Define cluster labels
cluster_labels <- c("Anxious/Low Motivation", "High Achiever", "Average", "High Achiever/Efficie
nt", "High Motivation/Inefficient")

# Add cluster labels to the data frame
kmeans_5_df <- kmeans_5_df %>%
  mutate(Label = cluster_labels[Cluster])

# Calculate median values for each cluster and arrange by descending tot_score
median_values <- kmeans_5_df %>%
  group_by(Cluster) %>%
  summarise(across(c("anxtest", "motivat", "tot_time", "tot_score"), median)) %>%
  arrange(desc(tot_score))

# Print the table of median values
median_values
```

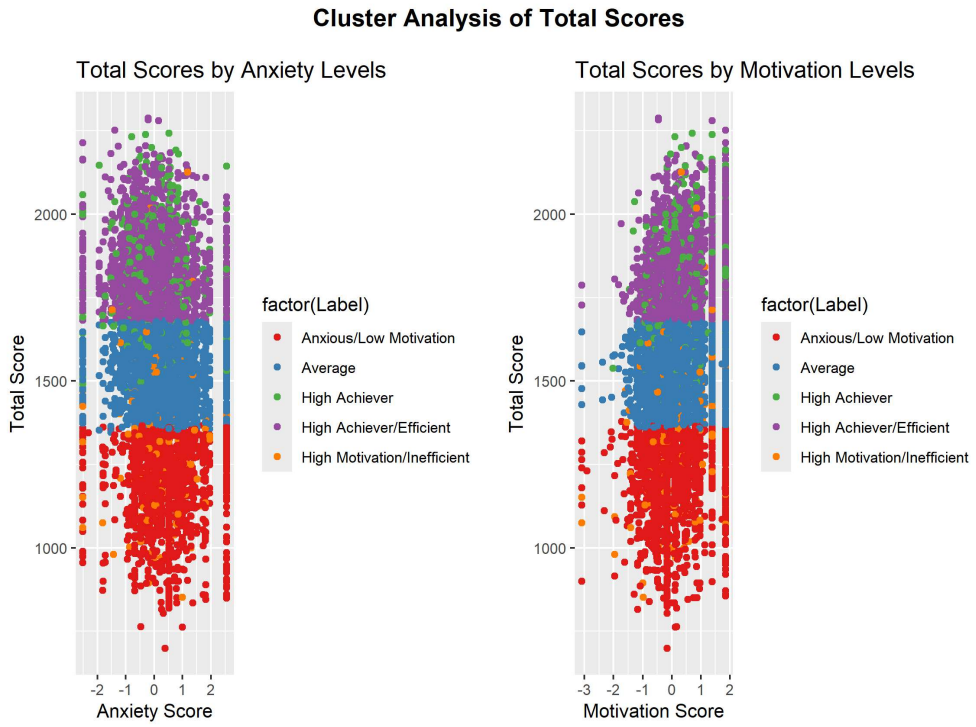
```
## # A tibble: 5 × 5
##   Cluster anxtest motivat tot_time tot_score
##     <int>   <dbl>   <dbl>    <dbl>     <dbl>
## 1       4 -0.0345   0.503      660     1808.
## 2       2  0.0752   0.539     1000     1795.
## 3       3  0.209    0.158      660     1533.
## 4       5  0.156    0.230     1350     1375.
## 5       1  0.259   -0.145      660     1224.
```

- **High anxiety coupled with low motivation leads to the lowest scores, even when spending more time on tasks (Cluster 1).**
- **Low anxiety and high motivation result in the highest scores and efficiency, requiring less time spent (Cluster 4).**
- **Moderate levels of anxiety and motivation produce average scores (Cluster 3 - the largest group).**
- **High motivation can partially offset the negative effects of moderate anxiety, leading to good scores despite spending considerable time (Cluster 2).**

- **Moderate anxiety combined with moderate motivation and spending the most time does not translate to higher scores (Cluster 5).**
- **Low anxiety is a more critical factor than motivation for achieving top performance efficiently.**
- **The majority of students fall into the average cluster with moderate anxiety/motivation levels.**

Question 12. Plot the relationship between `tot_scores` and color the plot by cluster. (3pt). Is there a meaningful difference between the scores across the different clusters? (1pt). What do you think this finding says about the importance of motivational factors in overall test scores? (2pt)

```
kmeans_5_df1 <- kmeans_5_df %>%
 rename(anxiety_score = anxtest, motivation_score = motivat, total_score = tot_score)

# Create the scatter plots
p_anxiety <- kmeans_5_df1 %>%
 ggplot(aes(x = anxiety_score, y = total_score, color = factor(Label))) +
 geom_point() +
 scale_color_brewer(palette = "Set1") +
 labs(title = "Total Scores by Anxiety Levels", x = "Anxiety Score", y = "Total Score")

p_motivation <- kmeans_5_df1 %>%
 ggplot(aes(x = motivation_score, y = total_score, color = factor(Label))) +
 geom_point() +
 scale_color_brewer(palette = "Set1") +
 labs(title = "Total Scores by Motivation Levels", x = "Motivation Score", y = "Total Score")

# Combine the plots
combined_plot <- plot_grid(p_anxiety, p_motivation, ncol = 2)
title <- ggdraw() + draw_label("Cluster Analysis of Total Scores", fontface = "bold")

final_plot <- plot_grid(title, combined_plot, ncol = 1, rel_heights = c(0.1, 1))
final_plot
```



- **Three distinct clusters based on total scores:_ Cluster 4 (High Efficiency) with highest scores, Cluster 1 (Anxious/Low Motivation) with lowest scores, and Cluster 3 (Average) with average scores.**
- **Cluster 2 (Motivated) is small, characterized by high scores with some average outliers.**
- **Cluster 5 (Diligent) shows an even distribution between average and low scores.**
- **Similar total score distributions but significant variability in anxiety levels across clusters.**
- **High scorers can have extremely low or high anxiety levels.**
- **Motivation has a more direct impact on performance than anxiety.**
- **High motivation is associated with high scores, while low motivation typically results in average scores at best.**
- **Overall, the clusters exhibit similar trends in total score distribution, with motivation being a stronger determinant of performance than anxiety.**

Question 14 (ETC5510): Do the findings described in question 11 remain true if we instead analyse the relationship between *read* and cluster assignment?

(5pt)

```
# Set the seed for reproducibility
set.seed(6)

# Select the relevant variables
k_means_data <- pisa_filtered %>%
  select(anxtest, motivat, tot_time, read)

# Perform k-means clustering with k=5
kmeans_result <- kmeans(k_means_data, centers = 5)

# Create a table with cluster sizes and centers
cluster_table <- data.frame(
  Cluster_Size = kmeans_result$size,
  Cluster_Centers = kmeans_result$centers
)

# Print the cluster table
cluster_table
```

```
##   Cluster_Size Cluster_Centers.anxtest Cluster_Centers.motivat
## 1          846               0.2591387               0.2920376
## 2          836               0.1644999               0.4872213
## 3         1976               0.2421139               0.1725094
## 4         2681               0.1511062               0.5517251
## 5          252               0.2022829               0.4666810
##   Cluster_Centers.tot_time Cluster_Centers.read
## 1                 468.3688             491.3349
## 2                 915.4127             543.8466
## 3                 673.5324             436.2662
## 4                 666.4976             588.2725
## 5                1401.9841             531.7736
```

```
# Create a data frame with cluster assignments
kmeans_df <- data.frame(
  Cluster = kmeans_result$cluster,
  pisa_filtered
)

# Define cluster labels
cluster_labels <- c(
  "Average Performers",
  "Anxious/Low Motivation",
  "Calm/High Motivation",
  "High Achievers",
  "Low Achievers"
)

# Add cluster labels to the data frame
kmeans_df <- kmeans_df %>%
  mutate(Label = cluster_labels[Cluster])

# Calculate median values for each cluster and arrange by descending read score
median_values <- kmeans_df %>%
  group_by(Cluster) %>%
  summarise(across(c(anxtest, motivat, tot_time, read), median)) %>%
  arrange(desc(read))

# Print the median values
median_values
```

```
## # A tibble: 5 × 5
##   Cluster anxtest motivat tot_time  read
##     <int>   <dbl>   <dbl>    <dbl> <dbl>
## 1       4  0.0752  0.463       660  582.
## 2       2  0.201   0.463       900  551.
## 3       5  0.0752  0.409      1320  537.
## 4       1  0.209   0.122       495  496.
## 5       3  0.248  -0.0413      675  449.
```

Cluster Profiles:

1. Moderate anxiety/motivation/scores, smaller cluster.
2. High anxiety, low motivation/scores.
3. High motivation, moderate anxiety, high scores, efficient time.
4. Largest, similar to 1 but higher scores.
5. Smallest, moderate anxiety, low scores, inefficient time.

Key Insights:

1. Less total time spent on reading task.
2. Cluster size distribution differs from total scores analysis.
3. High motivation + moderate anxiety = high scores/efficient time.

4. Low motivation + high anxiety = low scores.
5. Inefficient time use = low scores despite moderate anxiety.

```
# Create scatter plot of read scores vs. anxiety levels
p_read_anxiety <- kmeans_df %>%
  ggplot(aes(x = anxtest, y = read, color = factor(Label))) +
  geom_point() +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Read Scores by Anxiety Levels", x = "Anxiety Score", y = "Read Score")

# Create scatter plot of read scores vs. motivation levels
p_read_motivation <- kmeans_df %>%
  ggplot(aes(x = motivat, y = read, color = factor(Label))) +
  geom_point() +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Read Scores by Motivation Levels", x = "Motivation Score", y = "Read Score")

# Combine the scatter plots
combined_plot <- plot_grid(p_read_anxiety, p_read_motivation, ncol = 2)

# Add a title to the combined plot
titled_plot <- ggdraw() +
  draw_label("Cluster Analysis of Read Scores", fontface = "bold") +
  theme(plot.margin = margin(0.5, 0.5, 0.5, 0.5, "cm"))

# Arrange the title and combined plot vertically
final_plot <- plot_grid(titled_plot, combined_plot, ncol = 1, rel_heights = c(0.1, 1))

# Display the final plot
final_plot
```

- **Four distinct anxiety clusters, compared to three in total scores analysis.**
- **Average cluster shows high variability in reading scores.**
- **Motivation clustering similar to total scores analysis.**
- **High anxiety negatively impacts reading performance.**
- **High motivation positively impacts reading performance.**
- **Congruent with total scores analysis findings.**