BUAN 314 - SECTION 01

DATA MINING PROJECT

“ STUDENT PERFORMANCE FACTORS“

Bruklina Noka, Kymberly Patino, Anna Garcia

Executive Summary

Data Collection

For this project we decided to explore a topic that relates to our age group and generation in regards to our academic success. This report provides an analysis and evaluation of student academic performance using a dataset sourced from Kaggle titled *Student Performance Factors*. This dataset consists of 6,607 observations collected at a single point in time, where each observation represents an individual student. The dataset is cross-sectional, meaning that each row captures information about a student at one moment in time rather than tracking changes longitudinally. It contains 20 variables that are grouped in different categories related to academic performance, personal and family background, and learning environment.

The main purpose of the dataset is to utilize it as an educational tool to analyze how factors such as academic behaviors, psychosocial and motivational influences, resource and school environment factors, and background and demographic characteristics relate to academic success, as measured by exam scores.

Variables

Academic Behavior & Effort Hours_Studied Attendance Sleep_Hours Previous_Scores Tutoring_Sessions Physical_Activity	Psychosocial and Motivational Motivational_Level Parental_Involvement Peer_Influence Extracurricular_Activites
Resource and School Environment Access_to_Resources Internet_Access Teacher_Quality School_Type	Background and Demographic Family_Income Parental_Education_Level Distance_from_Home Learning_Disabilities Gender

The variables in this database can be categorized into four different groups that we thought would be appropriate and easy to follow.

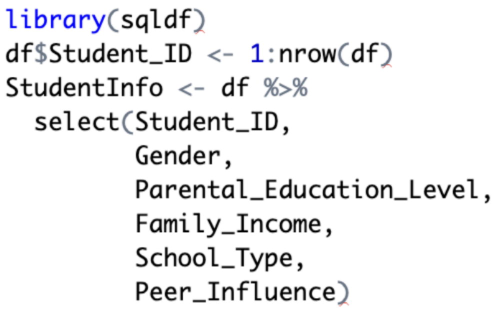
The first category would be Academic Behavior & Effort variables which mostly reflect how much consistent effort students put into their academic work and how they track their time and health. The second category would be Psychosocial and Motivational variables which refer to the social and motivational factors that could impact each student's score. The third category is Resource and School Environment variables which refers to the quality of resources that are provided to students at school. The fourth category is Background and Demographic variables which provide more information on the overall socio-economic and personal development of the student which may influence their exam score directly or indirectly.

Dataset Structure

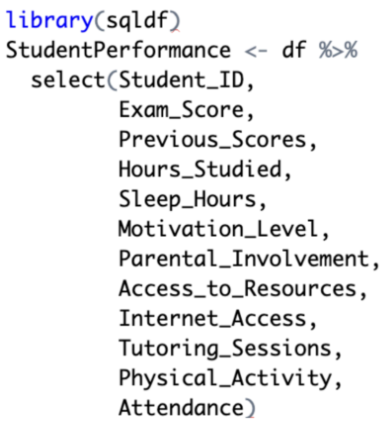
Structurally, the dataset is cross-sectional consisting of 6,607 rows and 20 columns. Each row represents an observation, eg: a student. Each column represents a variable which may be numerical, categorical or boolean. We specifically chose this dataset because it includes a mix of the three and this would produce more promising and realistic results as well. The numerical variables are useful because they can be summarized using arithmetic operations or central tendency and also are appropriate for correlation calculations and linear regression equations. The categorical variables are displayed in a level of “high, medium, low” and display qualitative information and in R these are called factor variables used for statistical plotting and modeling. The boolean variables are the ones which display a 0 or 1, and True or False. These are often called dummy variables and are converted in 0 or 1 so R can incorporate them in the analysis.

Numerical	Categorical	Boolean
Hours_Studied Attendance, Sleep_Hours Previous_Scores Tutoring_Sessions Physical_Activity Exam_Score	Parental_Involvement Motivational_Level Family_Income Parental_Educational_Level School_Type Peer_Influence	Internet_Access Extracurricular_Activities Learning_Disabilities Access_to_ Resources

The dataset is organized in one single table but for the purpose of this project we will create 2 linked tables using primary key (StudentID) and foreign key (Student_Performance). Table 1 is the Student_Info with demographic information about the student and Table 2 is the Study_Performance which includes study habits, resources access and outcomes scores. These will be mentioned in Queries 12 & 14 where the relationship is shown. These table structures will allow us to use SQL to showcase a relationship between tables but also make a distinction between who the student is given his background and how that student performs based on the factors mentioned above.



```
library(sqldf)
df$Student_ID <- 1:nrow(df)
StudentInfo <- df %>%
  select(Student_ID,
         Gender,
         Parental_Education_Level,
         Family_Income,
         School_Type,
         Peer_Influence)
```



```
library(sqldf)
StudentPerformance <- df %>%
  select(Student_ID,
         Exam_Score,
         Previous_Scores,
         Hours_Studied,
         Sleep_Hours,
         Motivation_Level,
         Parental_Involvement,
         Access_to_Resources,
         Internet_Access,
         Tutoring_Sessions,
         Physical_Activity,
         Attendance)
```

Data Cleaning

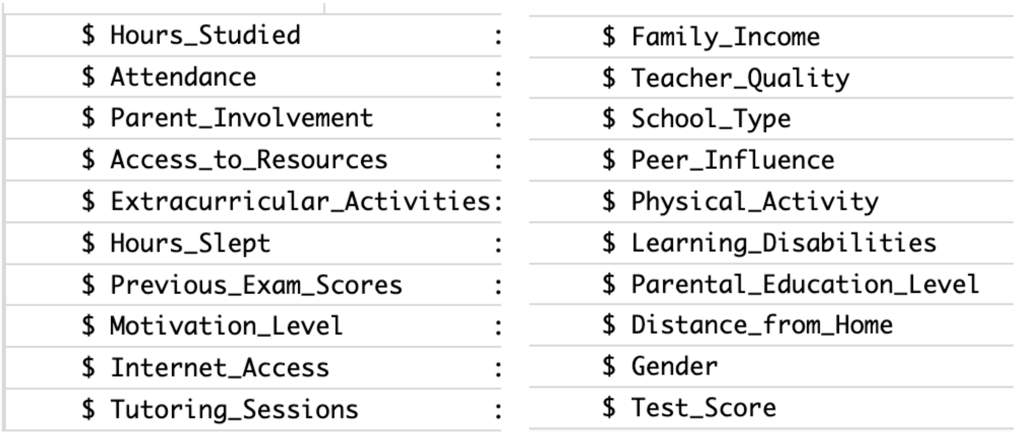
To prepare the dataset for further analysis, we are going to incorporate the following functions that we have learned and apply it to the dataset so we have a clean and synthesized table. Some of the functions we will be using are related to detecting missing or null values, correcting data types, recording categorical variables and creating initial exploratory visualizations to evaluate data quality.

The first step is to import the dataset and check its structure. We used the str(df), summary(df), and dim(df) to give us an overall view of the dataset.

The second step is finding if there are any missing or null values that can impact our analysis. Based on our code: missing_counts <- colSums(is.na(df)), no missing values were detected in the dataset meaning that the data does not need any imputation or any deletion of observations due to null entries. This makes for more accurate and reliable outcomes.

The third step is converting character variables into factor variables for the purpose of statistical modeling. This would allow us to create more meaningful visualizations and regression models. We used the tidyverse package and "mutate function" to apply the conversion and 13 of the variables became factors such as: Parental_Involvement, Access_To_Resources, Extracurricular_Activities, Motivation_Level, Internet_Access, Family_Income, Gender Teacher_Quality, School_Type, Peer_Influence, Learning_Disabilities, Parental_Education_ Level, Distance_From_Home. If we had not done this step, our analysis would not be helpful in identifying which variables affect exam scores.

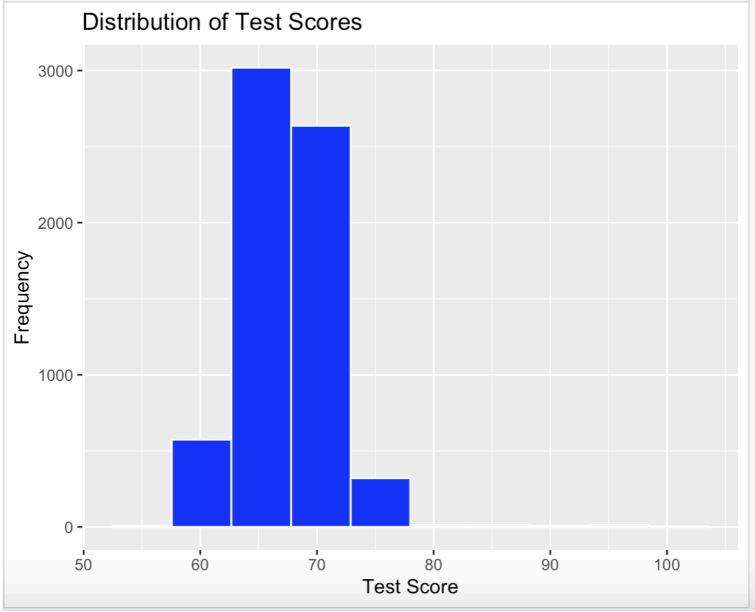
The third step is renaming some columns to ease readability and visibility and fix the aesthetic format. We used the "rename" function to re-write these variables: Exam_Score → Test_Score Parental_Involvement → Parent_Involvement, Sleep_Hours → Hours_Slept, and Previous_Scores → Previous_Exam_Scores. Also, for the Gender variable we used the dplyr package and "recode" function to change male → Male and female → Female to make it uniform. Below is the final table of variables:

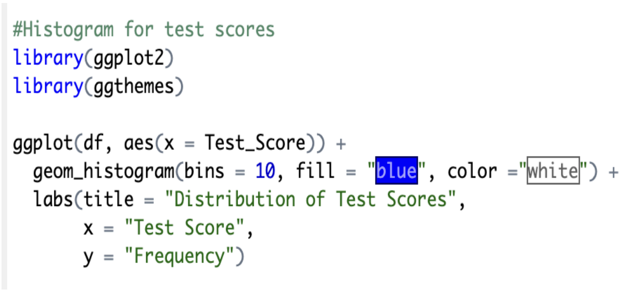


```
$ Hours_Studied             :     $ Family_Income
$ Attendance                :     $ Teacher_Quality
$ Parent_Involvement        :     $ School_Type
$ Access_to_Resources       :     $ Peer_Influence
$ Extracurricular_Activities:     $ Physical_Activity
$ Hours_Slept               :     $ Learning_Disabilities
$ Previous_Exam_Scores      :     $ Parental_Education_Level
$ Motivation_Level          :     $ Distance_from_Home
$ Internet_Access           :     $ Gender
$ Tutoring_Sessions         :     $ Test_Score
```

Data Visualization

Visualization 1- Histogram: Distribution of Test Scores

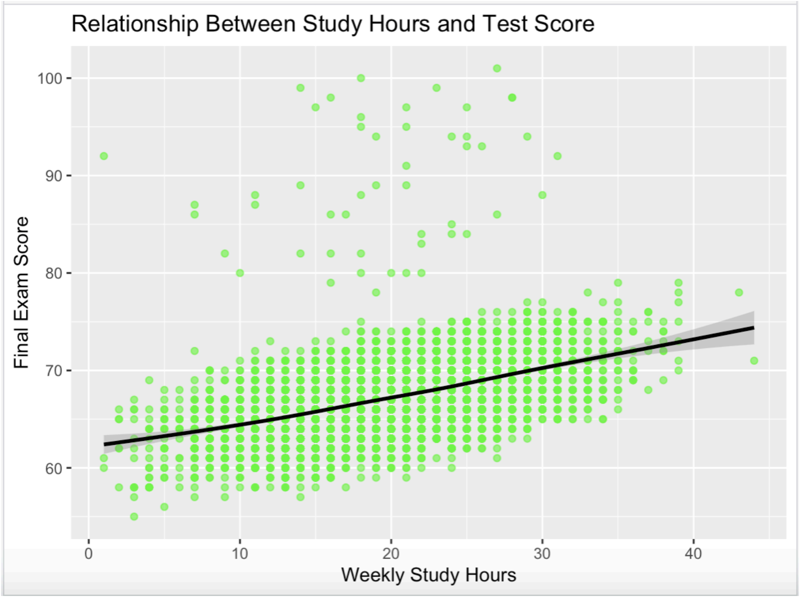


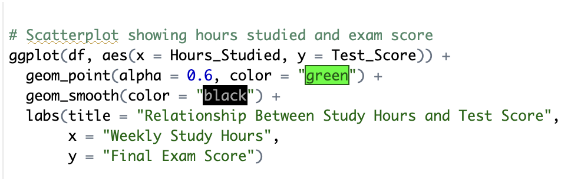


```r
#Histogram for test scores
library(ggplot2)
library(ggthemes)

ggplot(df, aes(x = Test_Score)) +
  geom_histogram(bins = 10, fill = "blue", color ="white") +
  labs(title = "Distribution of Test Scores",
       x = "Test Score",
       y = "Frequency")
```

Using ggplot, we created a histogram to show the distribution of exam scores across all students to get an idea of how they performed overall. The graph shows approximately a bell-shaped distribution showing that most students scored between 65-75. There are some extreme values below 65 and above 75 where students did really well or below average suggesting outliers but overall, there is a consistent consistency.

Visualization 2 - Scatterplot: Study Hours vs Average Exam Score

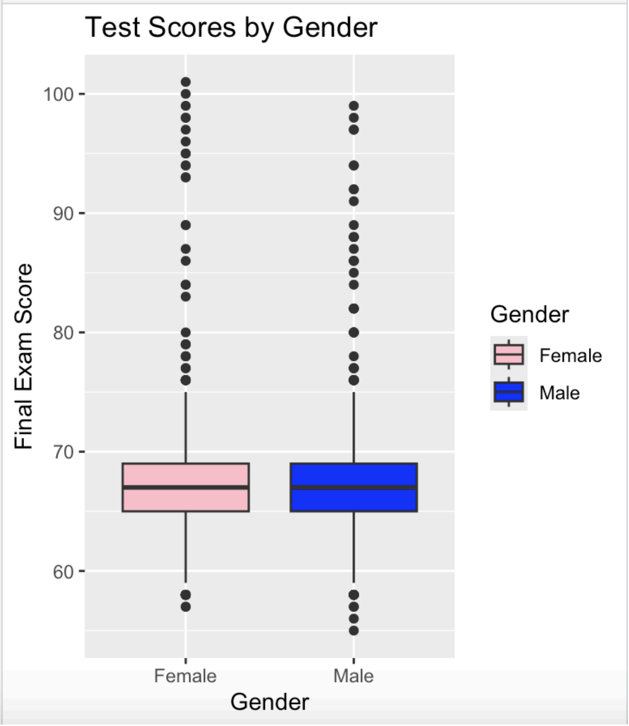


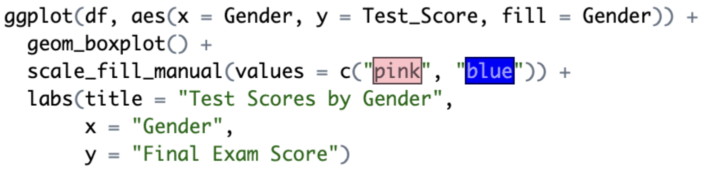


```r
# Scatterplot showing hours studied and exam score
ggplot(df, aes(x = Hours_Studied, y = Test_Score)) +
  geom_point(alpha = 0.6, color = "green") +
  geom_smooth(color = "black") +
  labs(title = "Relationship Between Study Hours and Test Score",
       x = "Weekly Study Hours",
       y = "Final Exam Score")
```

Using ggplot, geom_point and geom_smooth, we wanted to observe if the more hours you study the better you would score on the test and the chart shows a positive correlation between these two variables with an upward slope. The more hours you study, the higher your score will be. However, there is a large spread of variability around the trendline which could imply that hours are not the only determining factor that impacts scores but more like a combination of other factors too. Overall, the relationship is meaningful and proved our hypothesis.

<u>Visualization 3 - Boxplot: Gender vs Exam Score</u>

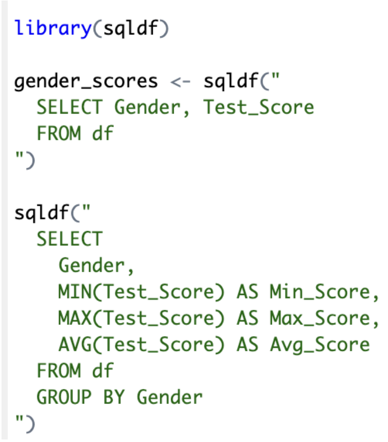




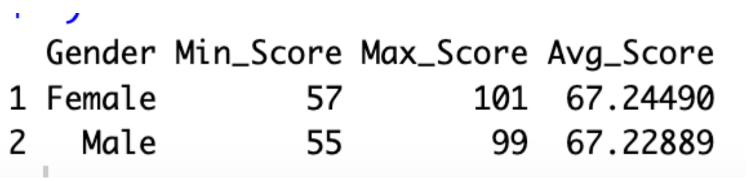
```
ggplot(df, aes(x = Gender, y = Test_Score, fill = Gender)) +
  geom_boxplot() +
  scale_fill_manual(values = c("pink", "blue")) +
  labs(title = "Test Scores by Gender",
       x = "Gender",
       y = "Final Exam Score")
```

Using ggplot, and geom_boxplot we wanted to explore if there was a pattern in scores between females and males so the boxplot shows that females and males almost scored the same overall. The median scores for both genders are nearly the same, and the interquartile ranges largely overlap. However, you can observe that there are more outliers for females suggesting that they performed slightly better.

Query 1: Scores Distribution by Gender



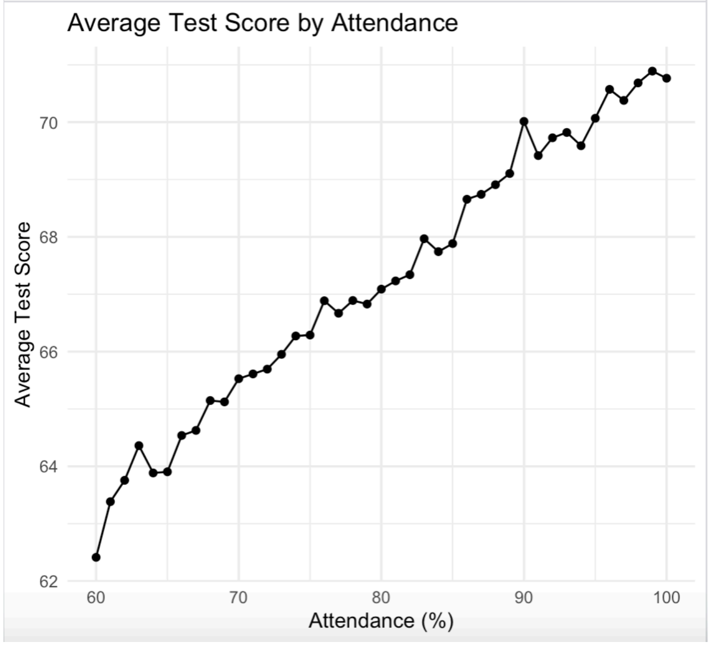
```
library(sqldf)

gender_scores <- sqldf("
  SELECT Gender, Test_Score
  FROM df
")

sqldf("
  SELECT
    Gender,
    MIN(Test_Score) AS Min_Score,
    MAX(Test_Score) AS Max_Score,
    AVG(Test_Score) AS Avg_Score
  FROM df
  GROUP BY Gender
")
```

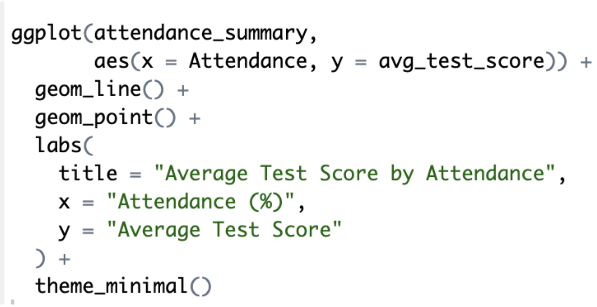


```
  Gender Min_Score Max_Score Avg_Score
1 Female        57       101  67.24490
2   Male        55        99  67.22889
```

In regards to the boxplot, we wanted to see the actual difference in scores between genders. Using the sqldf package, we ran a query to see the minimum, maximum and average score so we can clearly see how the scores are distributed amongst females and males. Females have done better than males as their minimum score is higher by 2 points (57>55), their maximum is higher by 2 points (101>99) and the average is approximately 67 for both. This query was done to help with the visualization and quantify it.

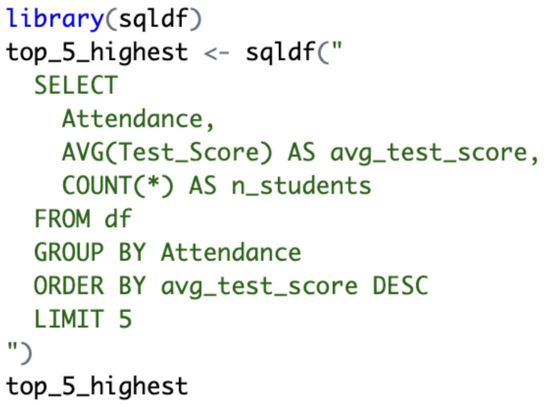
Visualization 4 - Linechart: Attendance vs Average Exam Score

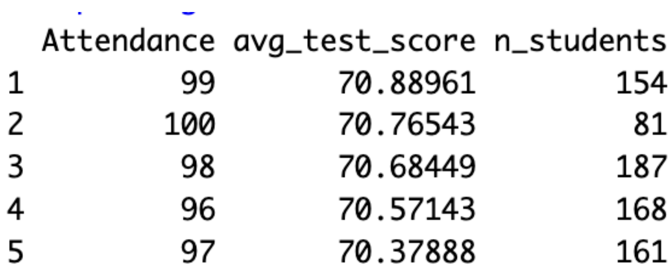


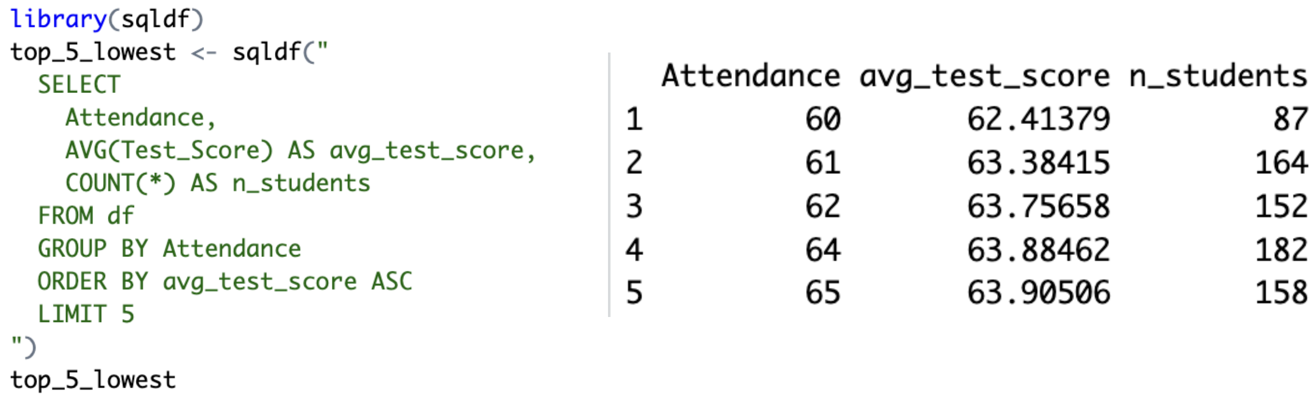

```
ggplot(attendance_summary,
       aes(x = Attendance, y = avg_test_score)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Average Test Score by Attendance",
    x = "Attendance (%)",
    y = "Average Test Score"
  ) +
  theme_minimal()
```

Using ggplot, geom_line and geom_points, we wanted to know if attendance plays a major role in the final exam scores. Our hypothesis is that the more you attend your classes, you would be active and be up to date with everything that is happening and so this would guarantee a higher score. As shown in the line chart, our hypothesis turned out to be true given the strong positive correlation. As attendance increased, there was a sharp increase in average exam scores. Overall, this variable has shown to be the most significant in a student's performance.

Query 2 & 3: Exam Scores by Attendance (Highest & Lowest)


```
library(sqldf)
top_5_highest <- sqldf("
  SELECT
    Attendance,
    AVG(Test_Score) AS avg_test_score,
    COUNT(*) AS n_students
  FROM df
  GROUP BY Attendance
  ORDER BY avg_test_score DESC
  LIMIT 5
")
top_5_highest
```


```
  Attendance avg_test_score n_students
1         99       70.88961        154
2        100       70.76543         81
3         98       70.68449        187
4         96       70.57143        168
5         97       70.37888        161
```

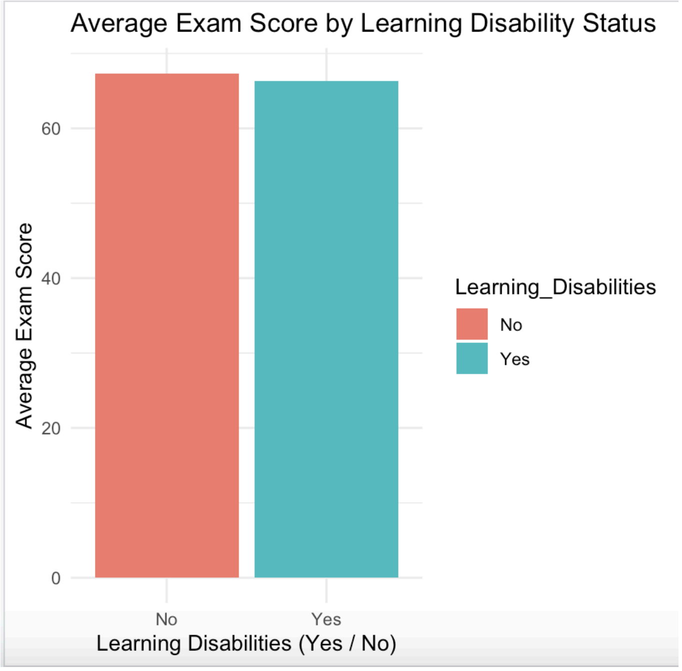


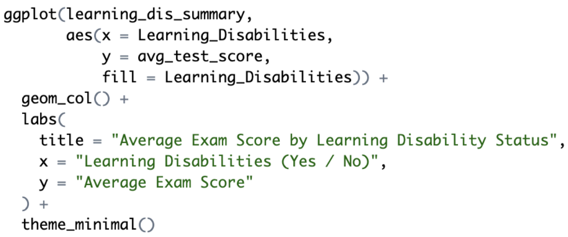
```
library(sqldf)
top_5_lowest <- sqldf("
  SELECT
    Attendance,
    AVG(Test_Score) AS avg_test_score,
    COUNT(*) AS n_students
  FROM df
  GROUP BY Attendance
  ORDER BY avg_test_score ASC
  LIMIT 5
")
top_5_lowest
```

```
  Attendance avg_test_score n_students
1         60       62.41379         87
2         61       63.38415        164
3         62       63.75658        152
4         64       63.88462        182
5         65       63.90506        158
```

Using sqldf, we wanted to see the top 5 highest scores and top 5 lowest scores and explore the difference in points. The top 5 (96-100%) have scored approximately 71 and the lowest 5 (60-65%) have scored approximately 63 and this makes a 7-8 exam point difference which is very meaningful for a student's performance. We can also observe that the number of students per cluster is not disproportionate or varies largely in numbers which reinforces that the set is very consistent and there are no fluctuations.

Visualization 5 - Barchart: Learning Disability vs Average Exam Score

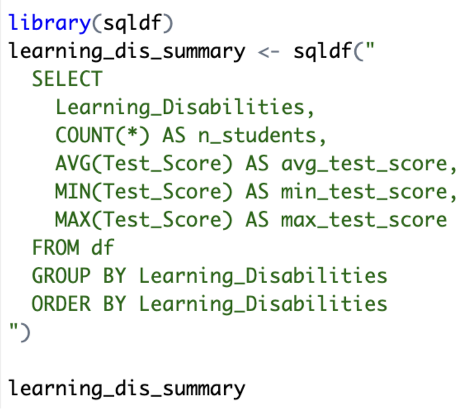




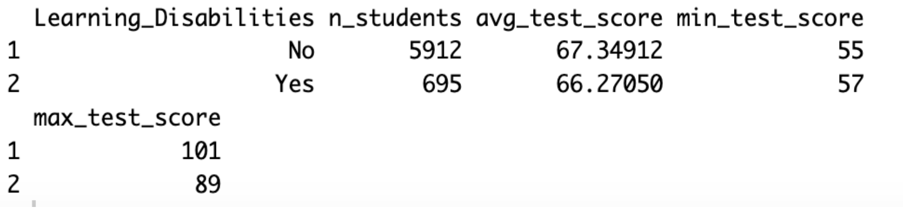
```
ggplot(learning_dis_summary,
       aes(x = Learning_Disabilities,
           y = avg_test_score,
           fill = Learning_Disabilities)) +
  geom_col() +
  labs(
    title = "Average Exam Score by Learning Disability Status",
    x = "Learning Disabilities (Yes / No)",
    y = "Average Exam Score"
  ) +
  theme_minimal()
```

Using ggplot and geom_col, we wanted to create a chart showing the relationship between learning disabilities and average exam scores, because there used to be and still is a stigma around learning disabilities and academic performance and we wanted to see if that is obvious in our dataset. Based on the visual, those students with no disabilities slightly outperformed the students with a disability and to explore that difference we made a query.

Query 4: Scores Distribution in Students with/without disabilities



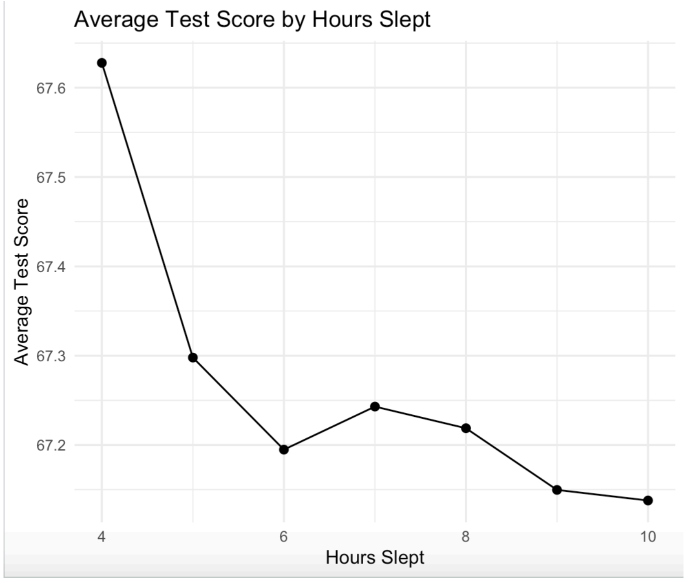
```
library(sqldf)
learning_dis_summary <- sqldf("
  SELECT
    Learning_Disabilities,
    COUNT(*) AS n_students,
    AVG(Test_Score) AS avg_test_score,
    MIN(Test_Score) AS min_test_score,
    MAX(Test_Score) AS max_test_score
  FROM df
  GROUP BY Learning_Disabilities
  ORDER BY Learning_Disabilities
")

learning_dis_summary
```

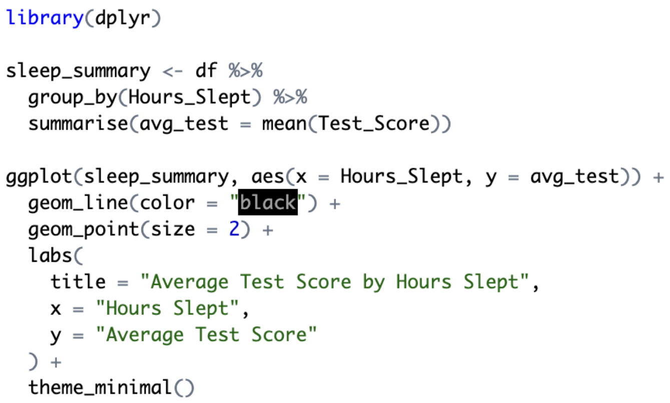


```
  Learning_Disabilities n_students avg_test_score min_test_score
1                    No       5912       67.34912             55
2                   Yes        695       66.27050             57
  max_test_score
1            101
2             89
```

To further explain the barchart, we ran a query to show the minimum, maximum and average score between the two groups of students. Students without learning disabilities have an average exam score of 67.35, while students with learning disabilities hav#Visualization 5 e a slightly lower average score of 66.27, a difference of just over one point. This small gap suggests that learning disability status alone does not substantially affect exam performance.This may indicate that support systems, accommodations, or instructional strategies help level academic outcomes between students with learning disabilities and those without.

Visualization 6 - Linechart: Hours Slept vs Average Test Score

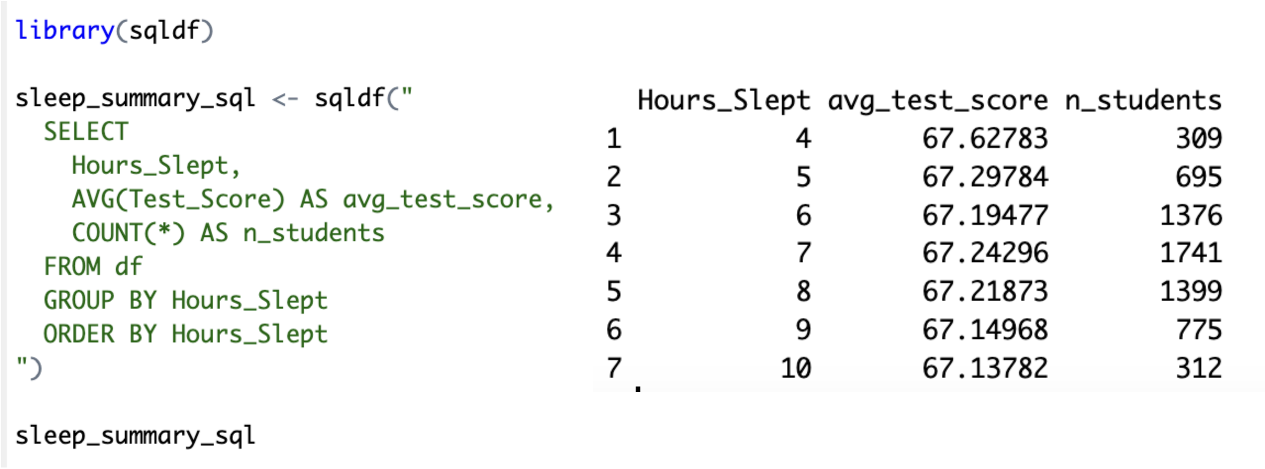




```
library(dplyr)

sleep_summary <- df %>%
  group_by(Hours_Slept) %>%
  summarise(avg_test = mean(Test_Score))

ggplot(sleep_summary, aes(x = Hours_Slept, y = avg_test)) +
  geom_line(color = "black") +
  geom_point(size = 2) +
  labs(
    title = "Average Test Score by Hours Slept",
    x = "Hours Slept",
    y = "Average Test Score"
  ) +
  theme_minimal()
```

Using ggplot, geom_line and geom_point we created a line chart which shows a very weak relationship between hours slept and average test scores. Our belief was that longer sleep hours would allow more restful sleep and prepare students better, but the downward trendline shows the opposite with a correlation of -0.01. While students who slept around 4–5 hours appear to have slightly higher average scores, the overall differences across sleep durations are minimal, with average scores clustered tightly around 67–68 points.

Query 5: Exam Score by Sleep Duration

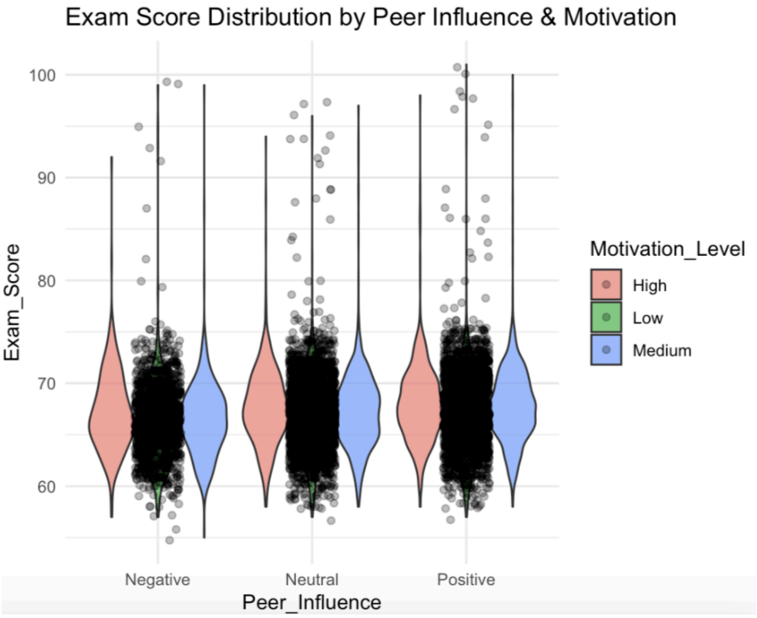


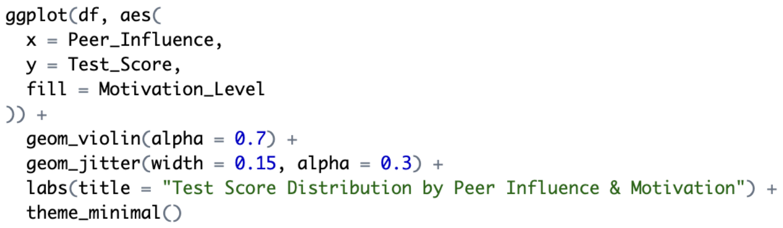
```
library(sqldf)

sleep_summary_sql <- sqldf("
  SELECT
    Hours_Slept,
    AVG(Test_Score) AS avg_test_score,
    COUNT(*) AS n_students
  FROM df
  GROUP BY Hours_Slept
  ORDER BY Hours_Slept
")

sleep_summary_sql
```

```
  Hours_Slept avg_test_score n_students
1           4       67.62783        309
2           5       67.29784        695
3           6       67.19477       1376
4           7       67.24296       1741
5           8       67.21873       1399
6           9       67.14968        775
7          10       67.13782        312
```

Based on the SQL result, most students sleep between 6-8 hours which is an optimal amount but there are 309 students who sleep for 4 hours and score the highest on their exams, 67.67. This could be explained by the fact that they sleep less because they have to study more and the more hours you work, the higher the chance of a better score. On the other hand, sleeping for more than 9-10 hours puts you in the lowest bound of the average score, 67.14.

Visualization 7 - Violin Plot: Exam Scores by Peer Influence and Motivation

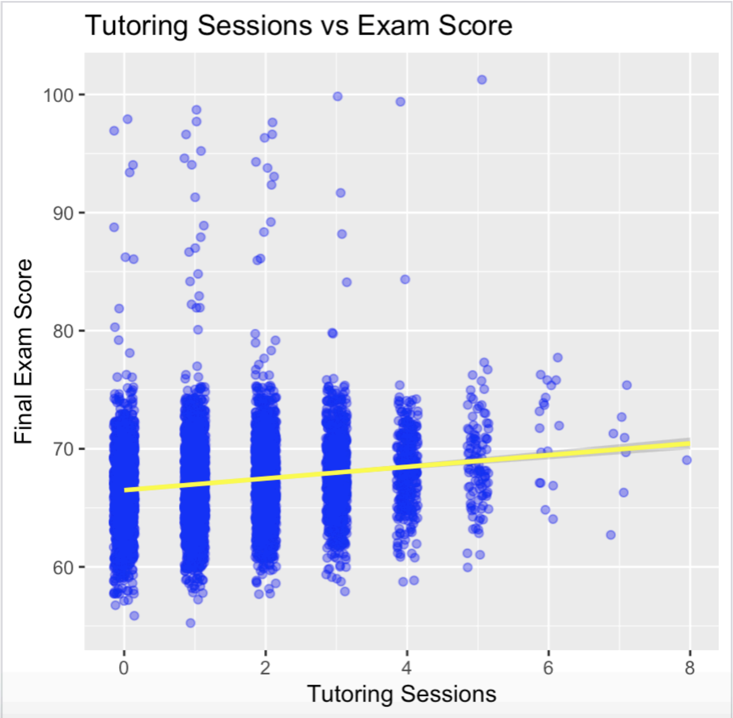


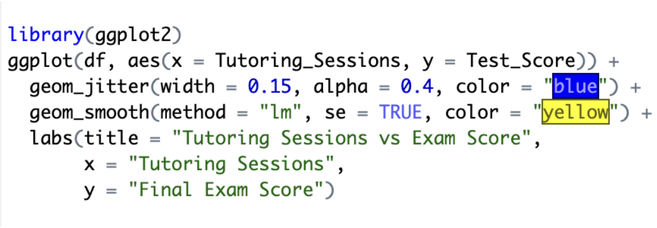


```
ggplot(df, aes(
  x = Peer_Influence,
  y = Test_Score,
  fill = Motivation_Level
)) +
  geom_violin(alpha = 0.7) +
  geom_jitter(width = 0.15, alpha = 0.3) +
  labs(title = "Test Score Distribution by Peer Influence & Motivation") +
  theme_minimal()
```

Using ggplot, geom_violin and geom_jitter, we compared exam scores by both peer influence and motivation level. Motivation clearly has the strongest impact. High-motivation students consistently perform well across all peer groups, while low-motivation students show the lowest and most variable scores. Peer influence helps boost motivation, but motivation itself is the key driver. The points are so dense that they hide some of the green violin shape underneath.

Visualization 8 - Jitter Scatterplot: Exam Score vs Tutoring Sessions

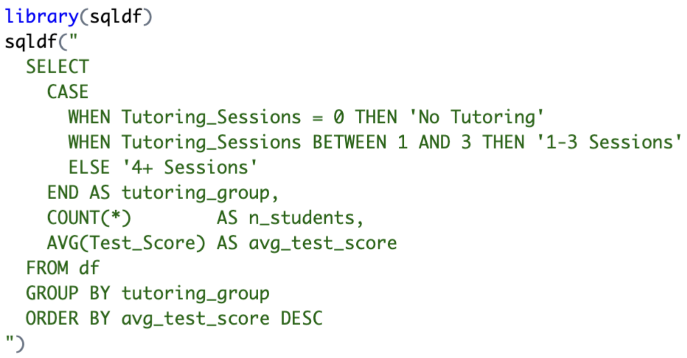




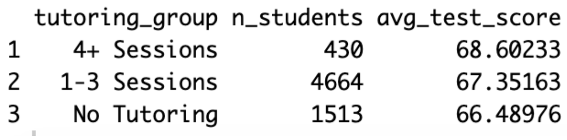
```
library(ggplot2)
ggplot(df, aes(x = Tutoring_Sessions, y = Test_Score)) +
  geom_jitter(width = 0.15, alpha = 0.4, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "yellow") +
  labs(title = "Tutoring Sessions vs Exam Score",
       x = "Tutoring Sessions",
       y = "Final Exam Score")
```

Using ggplot, geom_jitter and geom_smooth, the scatterplot shows a weak positive relationship showing that the more tutoring sessions you attend, the higher your average exam score. This makes sense because the more tutoring you do, and the more time you take to study more and having someone teach you outside of class will boost your academic performance.

Query 6: Exam Distribution and Tutoring Sessions



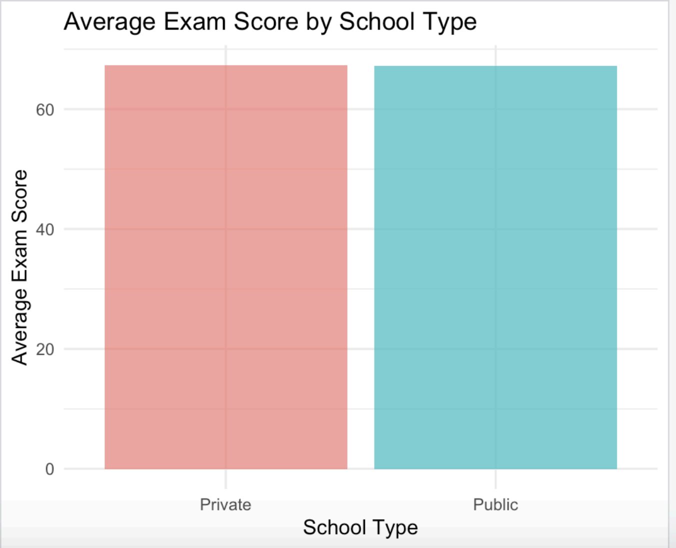
```
library(sqldf)
sqldf("
  SELECT
    CASE
      WHEN Tutoring_Sessions = 0 THEN 'No Tutoring'
      WHEN Tutoring_Sessions BETWEEN 1 AND 3 THEN '1-3 Sessions'
      ELSE '4+ Sessions'
    END AS tutoring_group,
    COUNT(*)        AS n_students,
    AVG(Test_Score) AS avg_test_score
  FROM df
  GROUP BY tutoring_group
  ORDER BY avg_test_score DESC
")
```

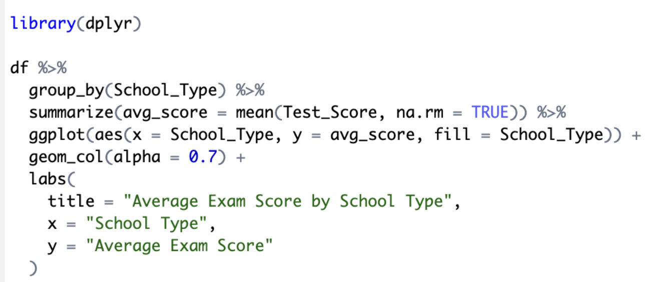


```
  tutoring_group n_students avg_test_score
1    4+ Sessions        430       68.60233
2   1-3 Sessions       4664       67.35163
3    No Tutoring       1513       66.48976
```

To further support our visualization 8, we made a query that shows the number of students per tutoring session and see the spread of data. When writing the code we wanted to do 3 categories of tutoring sessions and divided them in 0, 1-3 and 4+ tutoring sessions to produce more diverse results in our explanations. Although students who attended four or more tutoring sessions achieved the highest average test scores, this group represents the smallest portion of the sample, 430 of them. The majority of students fall within the 1–3 tutoring session category, suggesting that while frequent tutoring is associated with stronger outcomes, access to or engagement in sustained tutoring remains limited for most students.

Visualization 9 - Barchart: Average Exam Score by School Type

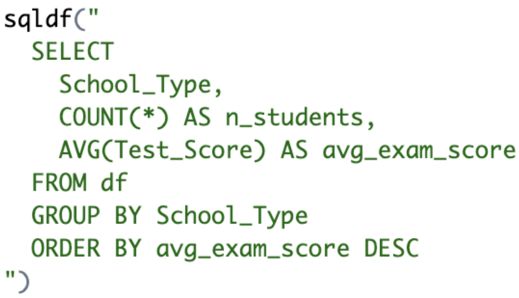




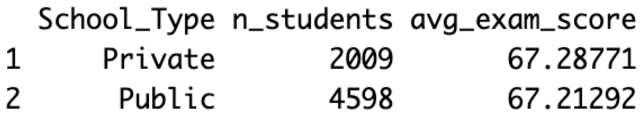
```
library(dplyr)

df %>%
  group_by(School_Type) %>%
  summarize(avg_score = mean(Test_Score, na.rm = TRUE)) %>%
  ggplot(aes(x = School_Type, y = avg_score, fill = School_Type)) +
  geom_col(alpha = 0.7) +
  labs(
    title = "Average Exam Score by School Type",
    x = "School Type",
    y = "Average Exam Score"
  )
```

Using the dplyr package, we wanted to create a barchart illustrating if the type of school that the students go to has an impact on their academic performance. Some people assume that a private education has more benefits than a public one and in this dataset, students who go to private schools do a little bit better but the margin is almost invisible. We wanted to explore this further and made a query on what the actual difference is and how many students attend which kind of institution.

Query 7:



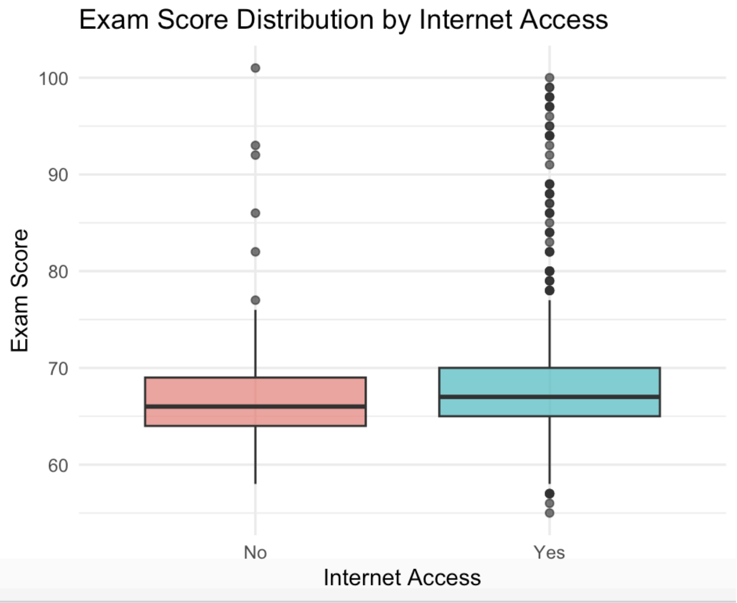
```
sqldf("
  SELECT
    School_Type,
    COUNT(*) AS n_students,
    AVG(Test_Score) AS avg_exam_score
  FROM df
  GROUP BY School_Type
  ORDER BY avg_exam_score DESC
")
```

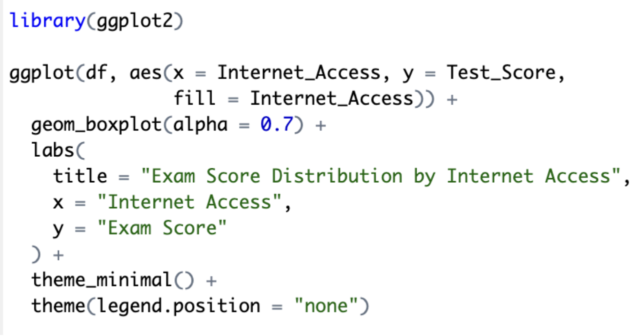


```
  School_Type n_students avg_exam_score
1     Private       2009       67.28771
2      Public       4598       67.21292
```

Based on the query, our results state that 2009 students go to private school and 4598 go to public schools, and this number was expected because a public school is much more affordable and accessible than a private school. However, we were surprised that the average exam scores were almost the exact same, differing by 0.07. This could imply that the type of school is not a very important factor, but it is more about the consistent effort that the students dedicate to their studies regardless which school they go to.

Visualization 10 - Exam Score by Internet Access

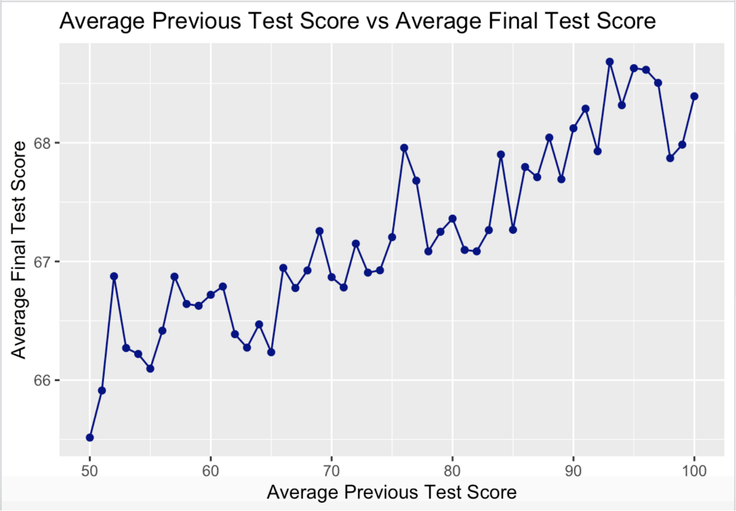


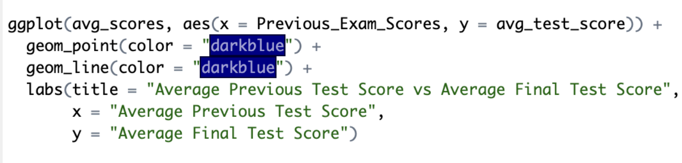


```
library(ggplot2)

ggplot(df, aes(x = Internet_Access, y = Test_Score,
               fill = Internet_Access)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Exam Score Distribution by Internet Access",
    x = "Internet Access",
    y = "Exam Score"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

Using ggplot and geom_boxplot, we made a boxplot showing the difference in exam scores when students have access to the internet. Students who have internet access show a slightly higher median exam score compared to those without access. In addition, the distribution for students with internet access shows a greater number of high-score outliers, indicating a higher potential for top academic performance.Students without internet access display lower median scores and a more compressed score range, suggesting more limited academic outcomes overall.

Visualization 11 - Line Chart: Previous Test Scores vs Test Scores

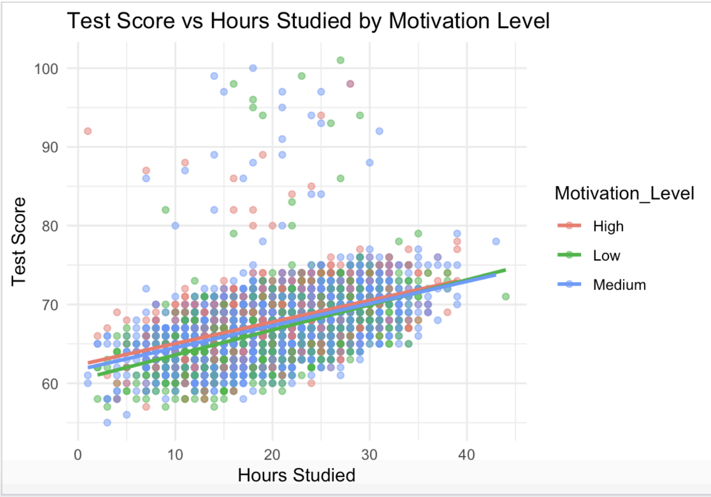


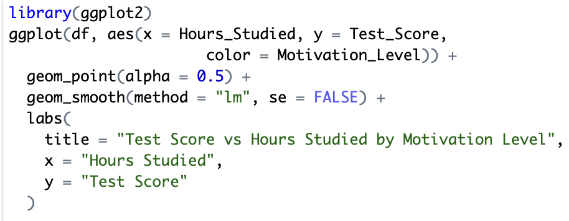


```
ggplot(avg_scores, aes(x = Previous_Exam_Scores, y = avg_test_score)) +
  geom_point(color = "darkblue") +
  geom_line(color = "darkblue") +
  labs(title = "Average Previous Test Score vs Average Final Test Score",
       x = "Average Previous Test Score",
       y = "Average Final Test Score")
```

Using ggplot, geom_point and geom_line, we made a line chart showing a clear positive relationship between average previous test scores and average final test scores. As previous test performance increases, the average final exam score rises steadily, indicating strong academic consistency over time. Students who performed better on earlier exams tend to maintain higher performance on the final exam. Although there are minor fluctuations, the overall upward trend is consistent across the entire score range, suggesting that previous exam scores are a strong predictor of exam performance.

<u>Visualization 12 - Scatterplot: Test Score vs Hours Studied by Motivation Level</u>

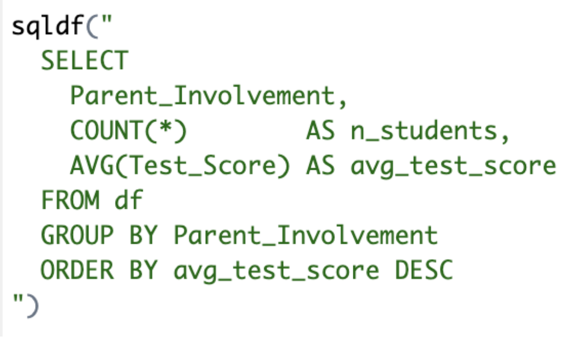




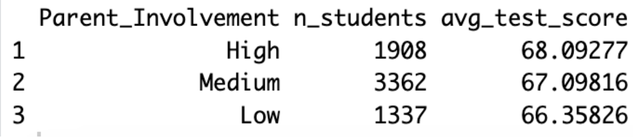
```
library(ggplot2)
ggplot(df, aes(x = Hours_Studied, y = Test_Score,
                         color = Motivation_Level)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Test Score vs Hours Studied by Motivation Level",
    x = "Hours Studied",
    y = "Test Score"
  )
```

Using ggplot, geom_point and geom_smooth, the scatter plot shows a positive relationship between hours studied and test scores across all motivation levels, indicating that increased study time is generally associated with higher exam performance. However, the strength and effectiveness of study time varies by motivation level, as shown by the separate trend lines.Highly motivated students consistently achieve higher test scores at the same study-hour levels compared to students with medium or low motivation. Additionally, the slope of the trend line for highly motivated students is slightly steeper, suggesting that study time yields greater returns when paired with high motivation. Overall motivation acts more as a booster than a single variable that has an actual effect on academic performance.

Query 8:



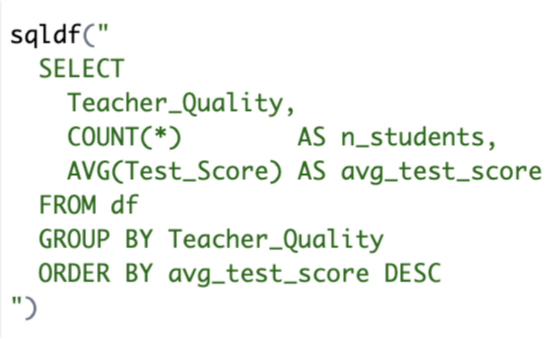
```
sqldf("
  SELECT
    Parent_Involvement,
    COUNT(*)         AS n_students,
    AVG(Test_Score) AS avg_test_score
  FROM df
  GROUP BY Parent_Involvement
  ORDER BY avg_test_score DESC
")
```



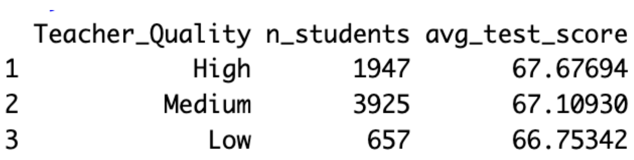
```
  Parent_Involvement n_students avg_test_score
1               High       1908       68.09277
2             Medium       3362       67.09816
3                Low       1337       66.35826
```

With this query we wanted to see if a student's parents are involved in their academics and how that reflects in their exam score. The results show a positive relationship between parent involvement and student exam performance. Students with high parental involvement achieved the highest average exam scores, followed by those with medium involvement, while students with low parental involvement performed the lowest on average. This pattern suggests that increased parental engagement may contribute to improved academic outcomes, potentially through greater academic support, monitoring, and encouragement at home.

Query 9:



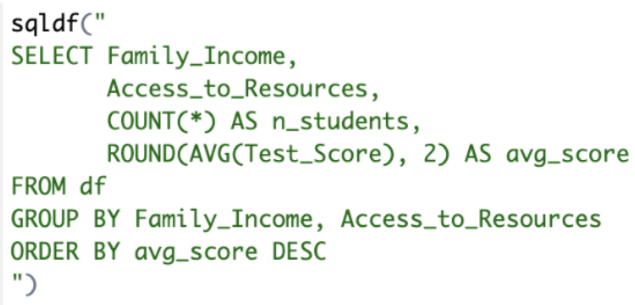
```
sqldf("
  SELECT
    Teacher_Quality,
    COUNT(*)        AS n_students,
    AVG(Test_Score) AS avg_test_score
  FROM df
  GROUP BY Teacher_Quality
  ORDER BY avg_test_score DESC
")
```



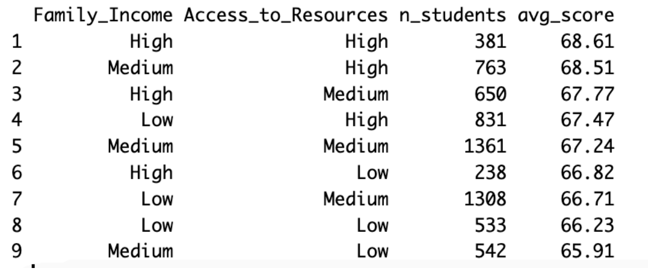
```
  Teacher_Quality n_students avg_test_score
1            High       1947       67.67694
2          Medium       3925       67.10930
3             Low        657       66.75342
```

For this query, we wanted to see the opposite of query 8, how the quality of teachers affect students' academic performance. The results show that higher teacher quality is associated with slightly higher average exam scores. Students taught by high-quality teachers performed better on average than those taught by medium- or low-quality teachers, although the differences across groups are relatively small. This suggests that while teacher quality plays an important role in academic achievement, its impact in this dataset appears smaller compared to other factors such as parental involvement or access to resources.

Query 10:



```
sqldf("
SELECT Family_Income,
       Access_to_Resources,
       COUNT(*) AS n_students,
       ROUND(AVG(Test_Score), 2) AS avg_score
FROM df
GROUP BY Family_Income, Access_to_Resources
ORDER BY avg_score DESC
")
```

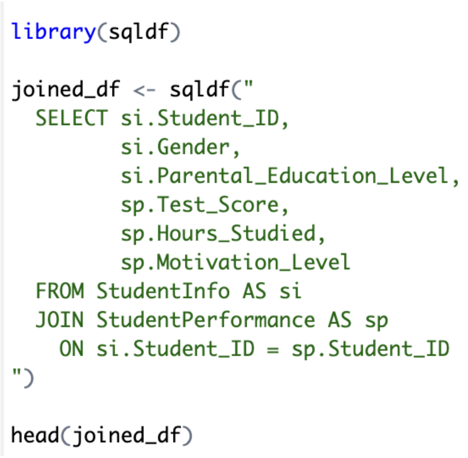


```
  Family_Income Access_to_Resources n_students avg_score
1          High                High        381     68.61
2        Medium                High        763     68.51
3          High              Medium        650     67.77
4           Low                High        831     67.47
5        Medium              Medium       1361     67.24
6          High                 Low        238     66.82
7           Low              Medium       1308     66.71
8           Low                 Low        533     66.23
9        Medium                 Low        542     65.91
```

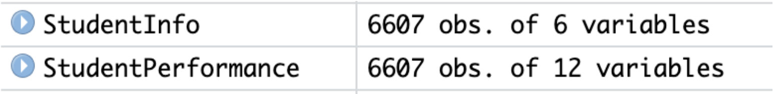
Other variables we wanted to explore is how family income and access to resources affect the student's ability to utilize these two and perform well academically. Family income and access to educational resources shows that access to resources has a strong impact on exam performance across all income levels. Students with high access to resources consistently achieved higher average scores, even when family income was low or medium. In contrast, students with low access to resources exhibited the lowest average scores regardless of income.

These findings suggest that improving access to educational resources may help reduce performance disparities linked to socioeconomic status.

Query 11 & 12: Primary & Foreign Key Relationship

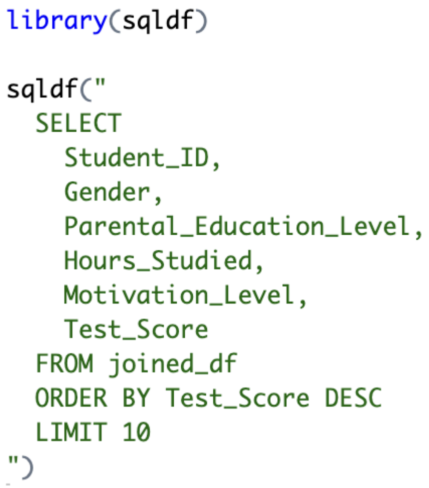


```
library(sqldf)

joined_df <- sqldf("
  SELECT si.Student_ID,
         si.Gender,
         si.Parental_Education_Level,
         sp.Test_Score,
         sp.Hours_Studied,
         sp.Motivation_Level
  FROM StudentInfo AS si
  JOIN StudentPerformance AS sp
    ON si.Student_ID = sp.Student_ID
")

head(joined_df)
```

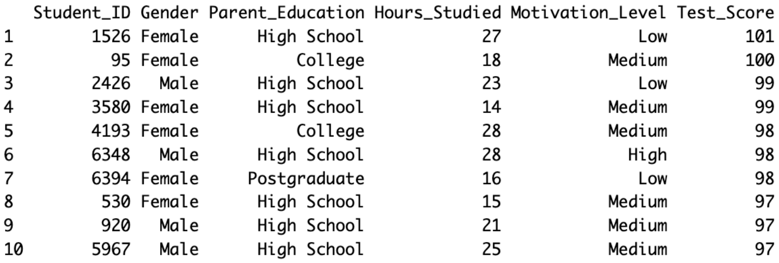


StudentInfo	6607 obs. of 6 variables
StudentPerformance	6607 obs. of 12 variables

For query 11 we wanted to demonstrate the relationship between the primary key which is the Student_Info with all the background and demographic information of the student and the foreign key which is Student_Performance, including all the 12 other variables. We ran a joined query so that we could connect both these tables and utilize it for query 12 which shows the top 10 students by tests score.



```
library(sqldf)

sqldf("
  SELECT
    Student_ID,
    Gender,
    Parental_Education_Level,
    Hours_Studied,
    Motivation_Level,
    Test_Score
  FROM joined_df
  ORDER BY Test_Score DESC
  LIMIT 10
")
```



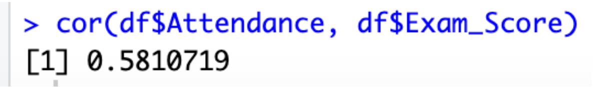
```
   Student_ID Gender Parent_Education Hours_Studied Motivation_Level Test_Score
1        1526 Female      High School            27              Low        101
2          95 Female          College            18           Medium        100
3        2426   Male      High School            23              Low         99
4        3580 Female      High School            14           Medium         99
5        4193 Female          College            28           Medium         98
6        6348   Male      High School            28             High         98
7        6394 Female     Postgraduate            16              Low         98
8         530 Female      High School            15           Medium         97
9         920   Male      High School            21           Medium         97
10       5967   Male      High School            25           Medium         97
```

Using the joined_df as our reference table, we wanted to see what characteristics do the top 10 students with the highest performing scores have. 6 out of 10 are females which also reinforces the point we made in the other charts that females have performed better than males. Another pattern is that most of them study 20+ hours and motivation level is at medium for most. Regarding background characteristics, parent education varies, with many students coming from high-school-educated households, as well as some from college or postgraduate backgrounds.

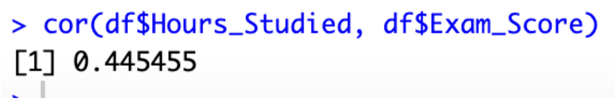
This variation suggests that while parental education may contribute to academic success, it is not a significant predictor for high achievement.

Correlation

To further examine the relationships between our key variables and exam scores, a correlation analysis was performed with *Exam_Score* as the primary variable of interest. The results revealed meaningful correlations that are associated with exam scores.



```
> cor(df$Attendance, df$Exam_Score)
[1] 0.5810719
```



```
> cor(df$Hours_Studied, df$Exam_Score)
[1] 0.445455
```

The strongest positive correlation observed was between *Attendance* and *Exam_Score*, with a correlation coefficient of 0.58. This tells us that there is a moderate strong relationship, suggesting that students who attend class more consistently tend to get higher exam scores. In addition, this tells us that regular class attendance enhances exposure to the course material and professor instructions, which contributes to student comprehension and retention of exam material.

The second strongest correlation was between *Hours_Studied* and *Exam_Score*, with a correlation coefficient of 0.45. This positive correlation indicates that more time spent studying is associated with higher exam scores. This correlation is a bit weaker than *Attendance* but still demonstrates that effort outside of the classroom plays a role in boosting exam success.

Linear Regression Model

Building on the results of the correlation analysis, a regression model was created to further examine the relationship correlation between *Attendance* and *Exam_Score*. *Attendance* was selected as the independent variable for the model because it had the strongest correlation with *Exam_Score* among all the variables analyzed in our dataset.

The estimated regression equation is: *Exam_Score* = 0.196 * *Attendance* + 51.58.

To further break down this equation, it is stating that if a student hypothetically never attended class, they would still score a 51.58 on the exam. In addition, the attendance coefficient is around 0.196, meaning that for every 1% increase in attendance, exam score increases by 0.196 points. This positive coefficient reinforces the previous correlation findings and suggests that consistent attendance does contribute to improved exam outcomes.

The regression model's multiple R-squared value is 0.3376, meaning that approximately 33.76% of the variation in exam scores can be explained by attendance alone. R-squared deals with predictive power, meaning that a higher number means it is very predictive and an accurate value. In this case it highlights how 66.24% of exam scores are influenced by additional factors beyond just attendance. Finally, the p-value associated with *Attendance* was also statistically significant, meaning that attendance is a predictor of exam scores. Overall, this regression model supports the idea of regular class attendance in predicting exam performance. This provides further details on to what degree attendance influences exam performance and captures a more complete understanding.

Conclusion

Throughout this project, we were able to draw several conclusions about the different factors that influence student academic performance on exams and identify meaningful trends and relationships within the dataset. By cleaning the data, creating a wide range of visualizations, and conducting correlation and regression analysis, we gained a comprehensive understanding of how academic behaviors, psychosocial factors, and environmental resources relate to exam scores. While clear relationships were observed between variables such as attendance, hours studied, sleep hours, and tutoring sessions, it is important to recognize that exam performance is influenced by many underlying and interconnected factors. As a result, this analysis should be interpreted as identifying associations rather than establishing causal relationships.

Visualizations played a critical role in supporting and validating our findings. Histograms, scatterplots, boxplots, line charts, and violin plots allowed us to examine both overall trends and group-level differences in exam performance. These visuals demonstrated that effort-based and engagement-driven variables—such as attendance, study hours, motivation level, and parental involvement—consistently showed stronger relationships with exam scores than background characteristics like school type or parental education. Additionally, visual comparisons across groups revealed that access to resources and internet availability can help mitigate performance gaps associated with income differences, emphasizing the importance of equitable learning environments.

One of the most significant findings in this analysis was the relationship between Attendance and Exam_Score. The correlation analysis showed that attendance had the strongest positive correlation with exam scores among all variables examined. This relationship was further reinforced through the linear regression model, which demonstrated that attendance is a statistically significant predictor of exam performance and explains a meaningful portion of the variation in exam scores. These results highlight the critical role that consistent class attendance plays in student success, likely due to increased exposure to course material, instructor guidance, and structured learning environments.

Building on these findings, several practical implications and opportunities for further improvement emerge. Schools and institutions could focus on increasing student motivation through reward systems and goal-setting workshops, strengthening parental involvement through more frequent parent–teacher meetings and academic workshops. Overall, this project demonstrates that student performance is shaped more by engagement, support systems, and access to resources than by fixed background characteristics, suggesting that targeted interventions can meaningfully improve academic success across diverse student populations.