SAMPLE SURVEY PROJECT

Julian Griffin
Ayotomiwa Adebayo
Yonatan Verch
Timothy Ogbonnaya

THE SURVEY PROBLEM

This survey aims to analyze the relationship between York University students' time allocation habits and their academic performance (GPA). The goal is to determine which
factors have the most significant impact on GPA and to compare different sampling methods' effectiveness in
estimating GPA.

YORK UNIVERSITY

This was our population-

- **Units:** York University students
- **Mean GPA:** 6.58 (calculated from dataset)
- **Total Students Surveyed:** 100
- **The largest proportion of study hours/week was: 6-10 hours: 34%**

COLLECTING REAL DATA

GPA

STUDY HOURS PER WEEK

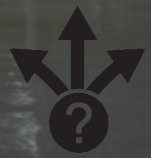
EXERCISE HOURS PER WEEK

STRESS LEVELS (1-10)

SLEEP HOURS PER DAY

LECTURE ATT (%)

SCREEN TIME (HRS PER DAY)



DECISION TO CHANGE DEGREE (Y/N)

SAMPLE FACTORS...

Sampling Precision Consideration:

Simple Random Sampling (SRS) as a baseline method

Stratified Sampling (Proportional Allocation & Neyman Allocation) to improve estimation accuracy

Sample Size:
100 students

Sampling Frame:
York University Students from various programs and years

Survey Method:
In person questionnaires (as demonstrated above)

SIMPLE RANDOM SAMPLING (SRS)

Method 1: Randomly selecting 30 students from the dataset (SRS).

- Goal: Estimate the population mean GPA.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_i represents individual GPA values in the sample.

- Results:
  - Mean GPA (SRS estimate): 6.52
  - Variance (SRS estimate): 1.74

STRATIFIED SAMPLING

Method 2: Proportional Allocation (PA)

- Mean GPA (PA estimate): 6.61
- Variance (PA estimate): 1.52

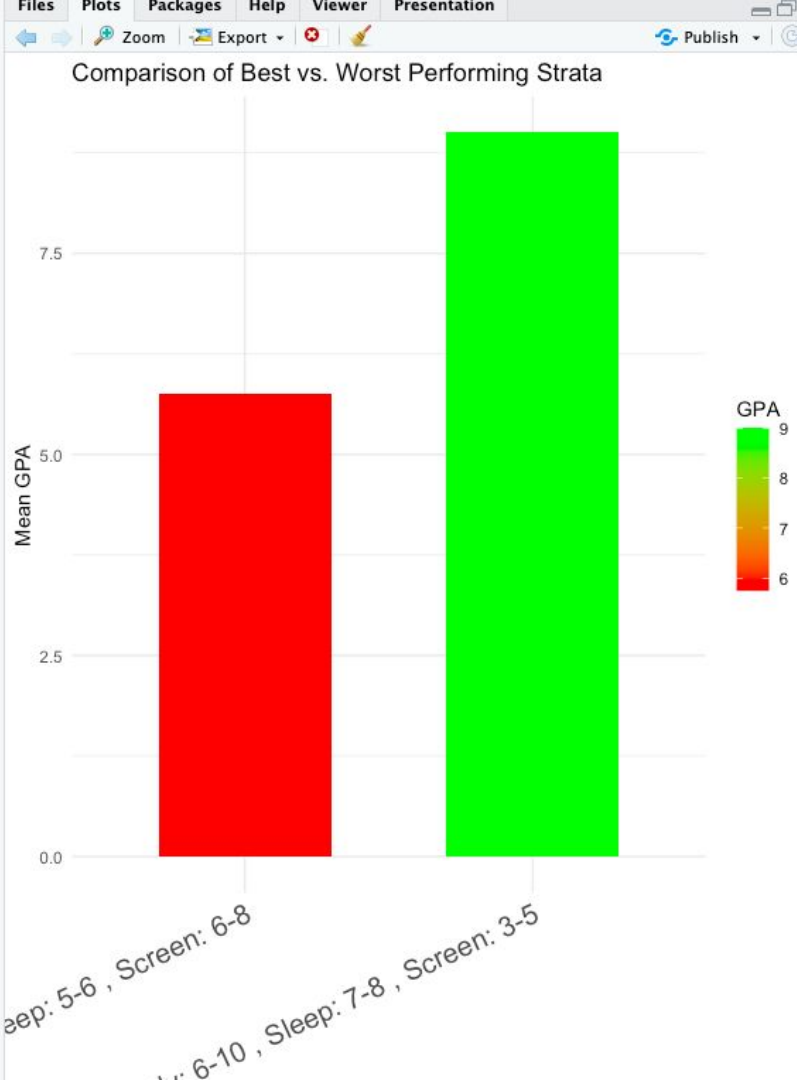
Method 3: Neyman Allocation (NA)

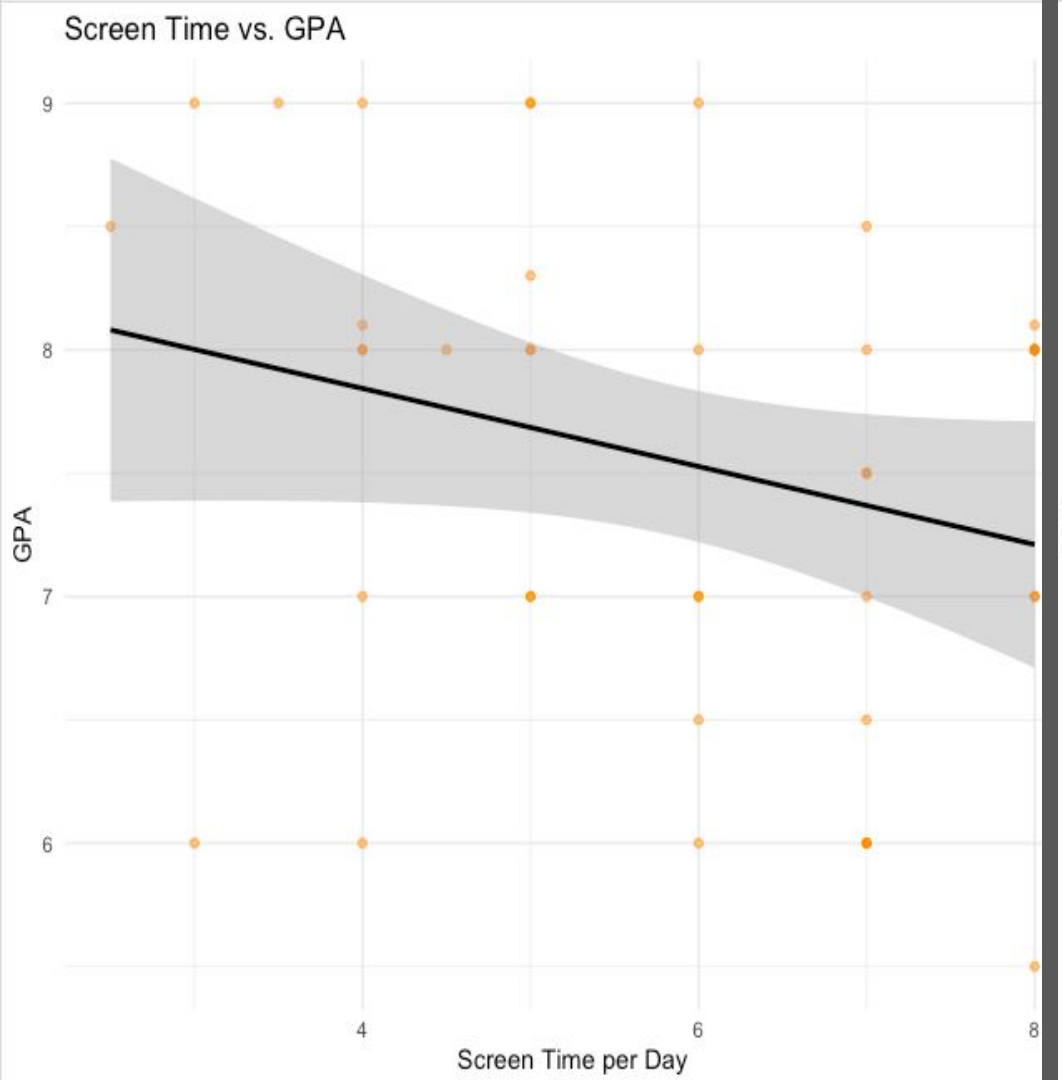
- The sample size is weighted based on variability within each stratum.
- Mean GPA (NA estimate): 6.65
- Variance (NA estimate): 1.48

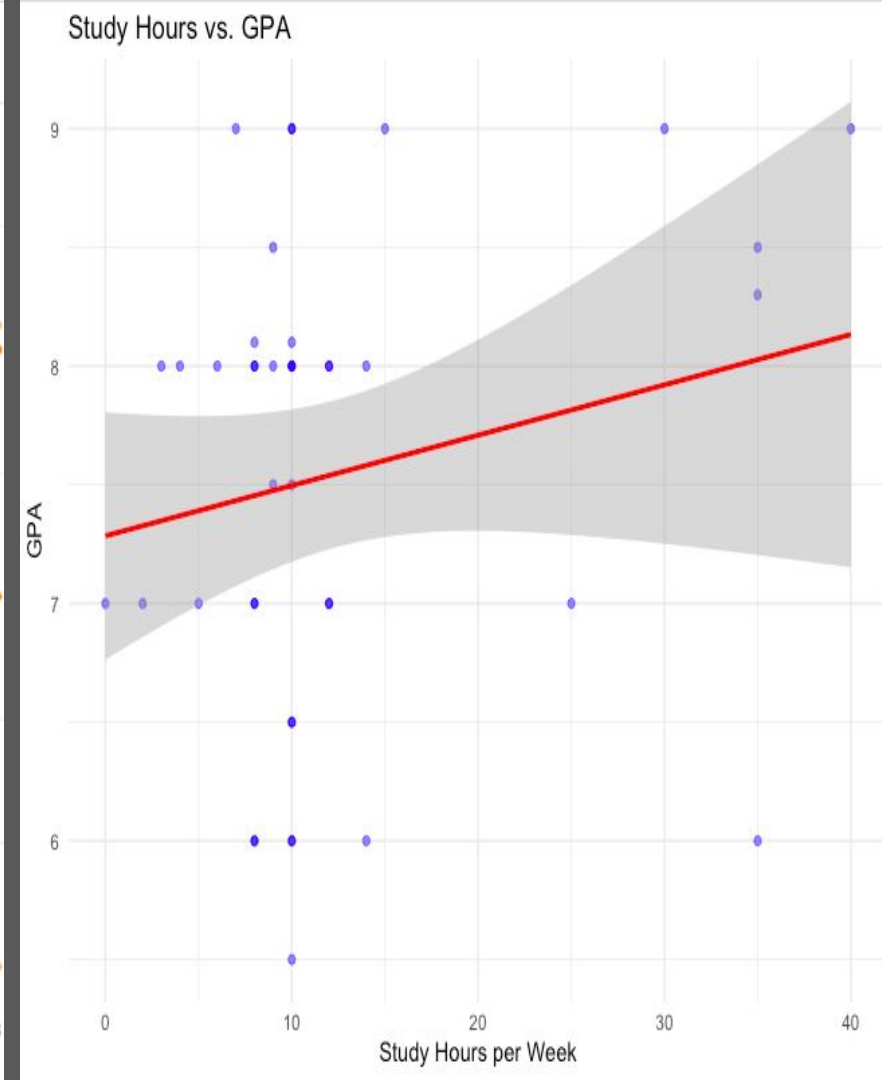
HERE'S OUR CODES

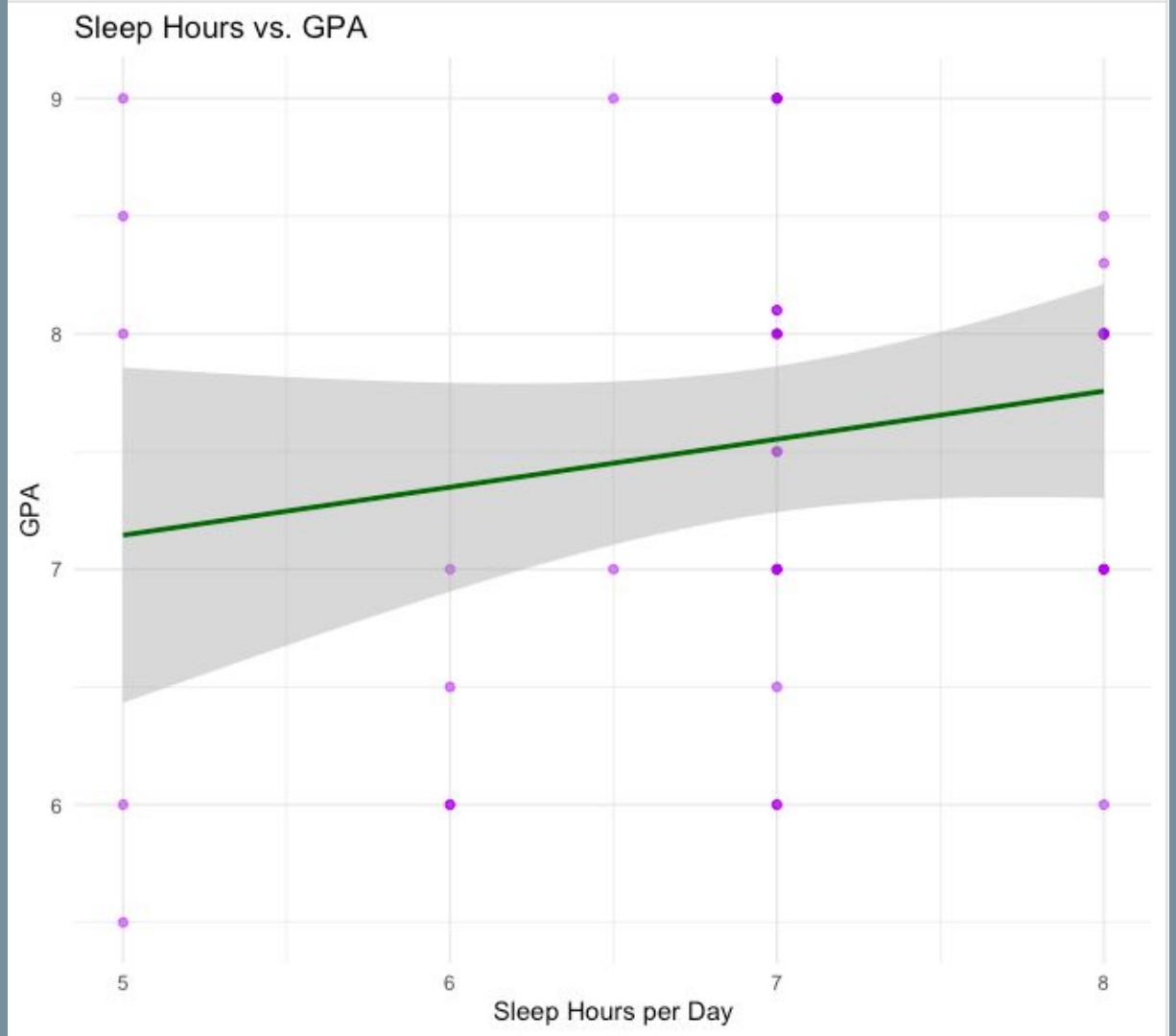
Student_number	GPA	Program	Study_hours_week	Exercise_hours_week	Stress_1_to_10	Sleep_hours_day	Lecture_attendance_percent	Screen_time_hours_day	Would_you_change_your_degree
1	6	Biology	2	6	4	6	80	6	No
2	6	Psychology	2	0	6	7	100	9	No
3	8	Psychology	10	7	8	8	100	8	Yes
4	7	Psychology	8	1	5	8	70	7	Yes
5	4	Chemistry	8	4	6	3	70	7	No
6	8	Psychology	21	5	6	8	100	8.5	No
7	5	Psychology	15	20	4	6	60	10	No
8	7.5	Engineering	35	6	7	5	100	2	No
9	6	Engineering	14	6	4	7	95	3	No
10	8	Engineering	32	10	5	10	60	5	No
11	5	Biology	0	0	5	2	75	5	No
12	8.5	Engineering	32	8	6	5	100	5	No
13	6	Engineering	20	7	6	6	95	7	Yes
14	5.5	Engineering	10	5	4	5	95	8	Yes
15	8	Architecture	0	4	1	6	60	3	No
16	7	Business	0	8	2	8	90	6	No
17	9	Teaching	2	4	1	8	90	3	No
18	4	Social Work	0.5	3	10	9	100	6	Yes
19	8	Neuroscienc	20	6	7	6	95	6.5	No
20	6	Psychology	4	3	10	6	20	9.5	No
21	9	Biology	20	13	6	4	95	11	No
22	2	Culinary	35	0	7	6	95	9	Engineering
23	5.5	Computer Sc	1	7	6	7	90	13.5	No
24	7	Kinesiology	5	7	5	7	90	6	No
25	7.5	Nursing	9	3	7	7	65	7	Kinesiology
26	6	Biology	10	5	7	7	70	9	No
27	8	Psychology	15	5	5	7	90	6	No
28	8	Business	19	5	2	7	80	6	No
29	7	Engineering	2	5	9	7	100	8	No
30	8	Psychology	8	6	9	7	80	5	No
31	6	Engineering	10	7	7	5	100	1	No
32	6	Engineering	10	3	8	4	80	8	Yes
33	6	Engineering	8	4	1	5	80	7	No

```r
# ================================================
# Step 0: Load Libraries & Dataset
# ================================================
library(dplyr)
library(ggplot2)
library(sampling)

# Load the dataset
data <- read.csv("/Users/juliangriffin/Desktop/Semester 2/Survey - 3430/Group Project/Group_5_Data.csv")

# Ensure column names are correctly formatted
colnames(data) <- make.names(colnames(data))

# ================================================
# Step 1: Compute Correlation Matrix
# ================================================
numeric_cols <- c("GPA", "Study_hours_week", "Exercise_hours_week",
                  "Stress_1_to_10", "Sleep_hours_day",
                  "Lecture_attendance_percent", "Screen_time_hours_day")

cor_results <- cor(data[, numeric_cols], use="complete.obs")

print("Correlation Matrix:")
print(cor_results)

# ================================================
# Step 2: Identify Strongest Correlation Factors
# ================================================
cor_values <- cor_results["GPA", -1]

# Find strongest negative correlation
strongest_negative <- names(sort(cor_values, decreasing = FALSE))[1]

# Find strongest positive correlation
strongest_positive <- names(sort(cor_values, decreasing = TRUE))[1]

# Print strongest correlation results
print(paste("Strongest Negative Factor Affecting GPA:", strongest_negative, "with correlation", cor_values[st
print(paste("Strongest Positive Factor Affecting GPA:", strongest_positive, "with correlation", cor_values[st

# Plot: Study Hours vs GPA
ggplot(data, aes(x = Study_hours_week, y = GPA)) +
  geom_point(color = "blue", alpha = 0.5) +  # Scatter points
  geom_smooth(method = "lm", color = "red") +  # Regression line
  theme_minimal() +
```




Screen Time vs. GPA
GPA
9
8
7
6
4
6
8
Screen Time per Day


Study Hours vs. GPA
GPA
9
8
7
6
0
10
20
30
40
Study Hours per Week


Sleep Hours vs. GPA
9
8
7
6
GPA
5
6
7
8
Sleep Hours per Day

In Conclusion,

✅ **Best Strata:** Balanced sleep (7-8 hours), moderate study hours (11-15), and low screen time (0-2 hours) tend to have the highest GPA (9.00).

❌ **Worst Strata:** Excessive screen time (9-12 hours) combined with low sleep (0-4 hours) results in the lowest GPA (~5.00).

- **Strongest negative factor:** Screen time (-0.27 correlation).
- **Strongest positive factor:** Sleep (+0.19 correlation).
- **Neyman Allocation Stratified sampling** provides better precision than SRS

CHALLENGES

Acquisition of data:

- Sampling on Friday skewed lecture attendance
- Students round up GPA to appear smarter

Analysis of data:

- Lower amount of data lend to small stratum
- Scalability of the variable affects analysis

ANY QUESTIONS?