SMDE Laboratory Assignment
First Assignment

,
Aftab Ahmed Choudhry, aftab.ahmed.choudhry@estudiantat.upc.edu
Roger Canal, roger.canal@estudiantat.upc.edu
Miquel Sala Franci, miquel.sala.franci@estudiantat.upc.edu

December 2025
Course 2025/2026

Contents

Chapter 1

QUESTION 1: visualization, Chi Square and T-Test

1.1 Part A: Data Import and Variable Structure

The data was imported from a `.csv` file using the base R function `read.csv()`. The argument `stringsAsFactors = TRUE` was used to correctly interpret textual columns (e.g., `Gender`, `Location`) as `Factor` variables, which is consistent with R lab practices. The structure and types were verified using `str()` and `summary()`.

1.2 Part B: Variable *age* and *stress* Classification

Two new factor variables, `age_cat` and `stress_cat`, were created to categorical levels. The `age` variable was classified into four generational groups: Gen Z (15-28), Gen Y (29-44), Gen X (45-60), and Baby Boomers (61-79). The `stress_level` was classified into two groups: Low (1-5), High (6-10). We used the `cut()` function to create the new factors.

The results were:

Age Category	Frequency
Gen Z	1137
Gen Y	1348
Gen X	1331
Baby Boomers	1184

Stress Category	Frequency
Low	2422
High	2578

1.3 Part C: Analysis of Association

The association between pairs of categorical variables was analyzed using Chi-Square test. This method tests the null hypothesis (H_0) that the variables are independent against the alternative hypothesis (H_1) that they are associated. The significance level used is $p - value = 0.05$.

1.3.1 C.1: Association between Location Type and Age Generation (age_cat)

```
# Create contingency table
location_age_table <- table(data$location_type, data$age_cat)
```

Results

location_type	Gen Z	Gen Y	Gen X	Baby boomers
Rural	225	296	279	243
Suburban	311	396	409	361
Urban	601	656	643	580

Chi-Square Output:

```
 Pearson's Chi-squared test
data:  location_age_table
X-squared = 7.5362, df = 6, p-value = 0.2741
```

Interpretation

Since $p = 0.2741 > 0.05$, the null hypotesis is **NOT** rejected. We conclude that there is **no association** between the participant's location type (location_type) and their age generation (age_cat). The distribution of age generations is similar across the different location types.

1.3.2 C.2: Association between Stress Category (stress_cat) and Age Generation (age_cat)

```
# Create contingency table
stress_age_table <- table(data$stress_cat, data$age_cat)
```

Results

Stress_level	Gen Z	Gen Y	Gen X	Baby boomers
Low	169	694	758	801
High	968	654	573	383

Chi-Square Output:

```
 Pearson's Chi-squared test
data:  stress_age_table
X-squared = 731.8, df = 3, p-value < 2.2e-16
```

Interpretation

Since $p < 0.05$, The null hypothesis is **rejected**. We conclude that there is a **significant association** between the categorical stress level (stress_cat) and the age generation (age_cat). The data suggests that **Low** stress levels are more frequent among older generations (Gen X, Baby Boomers), while **High** stress levels are concentrated in the younger generation (Gen Z).

1.4 Part D: Distribution of numerical variables

Checking the distribution of a numerical variable can be done visualizing Histograms and doing a **statistical check** (Shapiro-Wilk test).
The statistical test utilizes the following hypotheses with a significance level of $p - value = 0.05$:

- H_0: The variable follows a Normal distribution.
- H_1: The variable does **NOT** follow a Normal distribution.

With a large sample size ($N \geq 5000$), the Shapiro-Wilk test is overly sensitive and returned a significant p-value ($p < 0.05$) for 19 out of the 20 numerical variables analyzed. Therefore, we had to manually check the histograms to determine if it follows a normal distribution.

As shown in Appendix A, the only variable that followed a Normal distribution according to the test was `caffeine_intake_mg_per_day` (p-value = 0.1083). However, by observation we conclude that also `daily_screen_time_hours`, `phone_usage_hours`, `entertainment_hours`, `gaming_hours`, `sleep_duration_hours`, `sleep_quality` and `mental_health_score` follow a normal distribution.

1.5 Part E: Test for Effect of Stress Level on Mental Health Score

The final step is to determine if the categorical stress level (`stress_cat`) has an effect on the `mental_health_score` variable. We implemented the **Two-Sample t-Test** for comparing the means of two independent groups.

1.5.1 E.1: Testing Assumptions of the t-Test

Normality Assumption (Group-wise)

The Shapiro-Wilk test was applied separately to the mental health scores for the two stress groups.

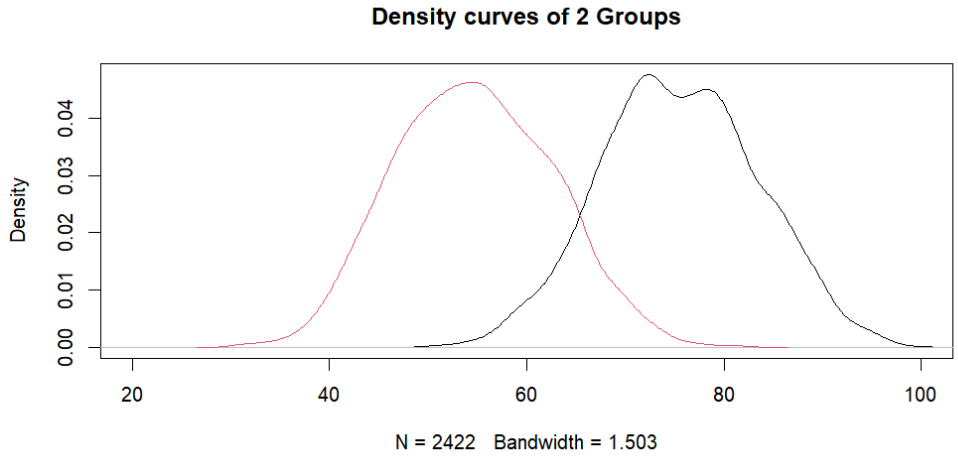


Figure 1.1: Density curves of low stress and high stress groups

Group	Shapiro-Wilk W	p-value	Conclusion ($\alpha = 0.05$)
Low Stress	0.99691	8.20e−05	Reject H_0 (Not Normal)
High Stress	0.99609	3.00e−06	Reject H_0 (Not Normal)

The test rejects the null hypothesis of normality for both groups ($p \ll 0.05$). However, given the very large sample size ($N \geq 5000$), the Central Limit Theorem (CLT) ensures the sampling distribution of the mean remains approximately normal, allowing us to proceed with the t-Test.

Homogeneity of Variances Assumption (F-Test)

The F-test was used to check if the variances of the mental health scores are equal across the two stress groups. The F-test results with p-value of 0.8916. Since $p > 0.05$, we **fail to reject** H_0 and conclude that the variances are statistically equal.

1.5.2 E.2: Visualization (Boxplot)

A boxplot provides a visual comparison of the Mental Health Score distribution across the stress categories.

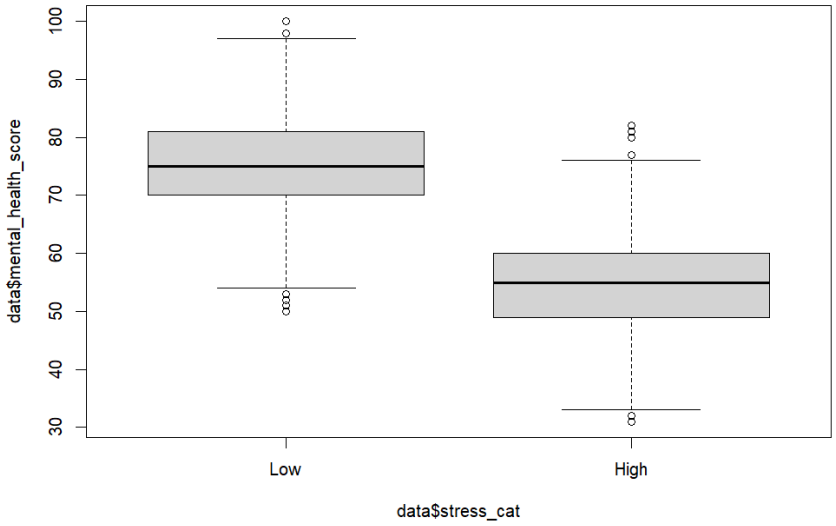


Figure 1.2: Boxplot of Mental Health Score by Stress Category

The boxplot visually demonstrates that the median and overall score distribution for the Low Stress group are higher than those of the High Stress group. This leads us to conclude that there is a *negative* effect between stress level and mental health score.

1.5.3 E.3: Two-Sample t-Test

Given the equal variances, a standard two-sample t-test with `var.equal = TRUE` was performed. We got a t-statistic value of $t = 92.639$ and the p-value $p < 2.2 \times 10^{-16}$. Since the p-value is much less than $\alpha = 0.05$, we **reject** the null hypothesis (H_0) that the means are equal.

In conclusion, there is a significant difference in the mean Mental Health Score between participants with low stress and those with high stress. The difference in means is approximately 20.83, with the **Low Stress** group scoring higher on the mental health than the **High Stress** group. This provides strong evidence that a participant's stress level affects their mental health score.

Chapter 2

QUESTION 2: ANOVA

The second part of the assignment focuses on comparing means across multiple groups using the ANOVA framework.

2.1 Part A: Create Categorical Sleep Variable

Again, the `cut()` function was used to segment the numerical `sleep_quality` variable into the three specified factor levels.

```
data$sleep_cat <- cut(data$sleep_quality, breaks = c(1, 3, 4, 5),
    labels = c("Low", "Medium", "High"), right = TRUE, include.lowest = TRUE)

sleep_cat
   Low Medium   High
  1011   2898   1091
```

2.2 Part B: Distribution Check by Group

We generated Boxplots to visually compare the distributions of the numerical variables across the different groups.

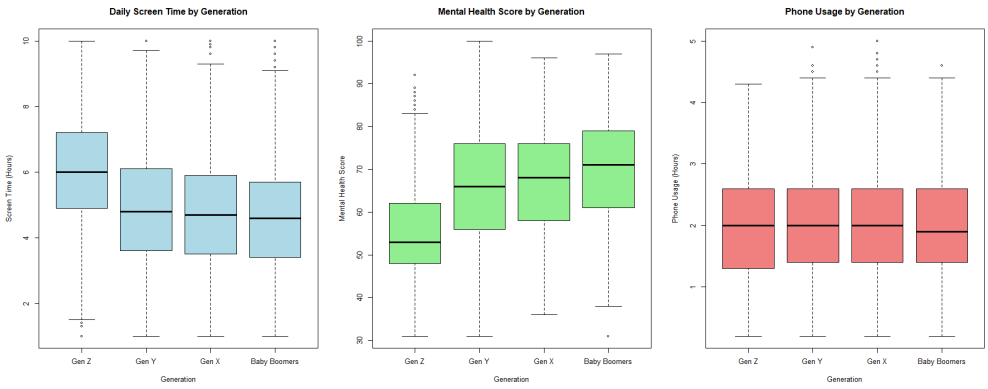


Figure 2.1: Distributions by Age Generation

The boxplots in Figure 2.1 reveal clear trends associated with age.

- **Daily Screen Time:** Shows a clear negative trend. The median screen time is highest for 'Gen Z' and decreases with each subsequent generation.
- **Mental Health Score:** Shows a clear positive trend. The median score is lowest for 'Gen Z' and steadily increases with age, being highest for 'Baby Boomers'.
- **Phone Usage:** Shows a slight negative trend, with 'Gen Z', 'Gen Y', and 'Gen X' having very similar distributions, while 'Baby Boomers' show a slightly lower median usage.

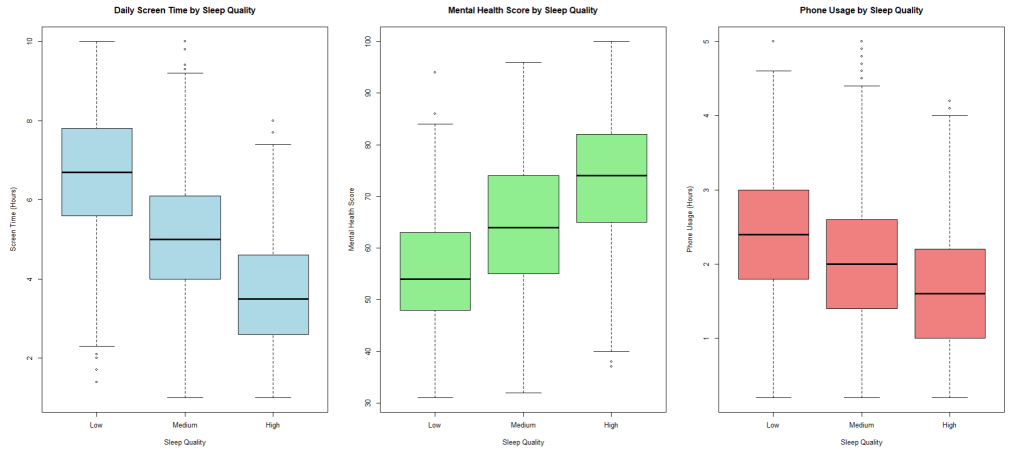


Figure 2.2: Distributions by Sleep Quality

The boxplots in Figure 2.2 reveal strong trends associated with sleep quality.

- **Daily Screen Time:** Shows a strong negative trend. Median screen time is highest for the 'Low' sleep quality group and lowest for the 'High' quality group.
- **Mental Health Score:** Shows a strong positive trend. Median mental health score increases significantly with better sleep quality.
- **Phone Usage:** Shows a strong negative trend, mirroring the screen time results. Better sleep quality is associated with lower median phone usage.

2.3 Part C: One-Way ANOVA Analysis

Four separate One-Way ANOVAs were conducted using `aov()`. For each model, we tested the three key assumptions:

1. Independence of errors (Durbin-Watson test, `dwtest(..., alternative="two.sided")`).
2. Normality of residuals (Shapiro-Wilk test, `shapiro.test()`).
3. Homogeneity of variances (Levene's Test and Breusch-Pagan test).

As per the assignment instructions, multiple comparison post-hoc tests were run for all models found to be significant.

ANOVA 1: `phone_usage_hours ~ age_cat`

- **ANOVA Result:** The model was Not Significant (p-value = 0.298).
- **Assumptions:** All assumptions were met (DW p = 0.7039; Levene's p = 0.9121).
- **Conclusion:** There is no statistically significant difference in `phone_usage_hours` among the different age generations. No post-hoc tests are required.

ANOVA 2: `phone_usage_hours ~ sleep_cat`

- **ANOVA Result:** The model was **Highly Significant** (p-value $< 2e-16$).
- **Assumptions:** All assumptions were met (DW p = 0.8562; Levene's p = 0.3958; BP p = 0.074). The model is valid.
- **Post-Hoc Tests:** All three tests (Tukey HSD, Bonferroni, LSD) were in complete agreement, showing that all three sleep groups ('Low', 'Medium', 'High') are significantly different from one another (LSD groups: 'a', 'b', 'c').

ANOVA 3: `daily_screen_time_hours ~ age_cat`

- **ANOVA Result:** The model was **Highly Significant** (p-value $< 2e-16$).
- **Assumptions:** All assumptions were met (DW p = 0.3474; Levene's p = 0.4423). The model is valid.
- **Post-Hoc Tests:** The tests show that 'Gen Z' (mean 6.07) is significantly different from all other generations. 'Gen Y' (mean 4.85) is also different from 'Baby Boomers' (mean 4.63). 'Gen X' (mean 4.70) and 'Baby Boomers' are not significantly different from each other (LSD Bonf. groups: 'a', 'b', 'bc', 'c').

ANOVA 4: `daily_screen_time_hours ~ sleep_cat`

- **ANOVA Result:** The model was **Highly Significant** (p-value $< 2e-16$).
- **Assumptions:** The independence assumption was met (DW p = 0.7082). However, the assumption of homogeneity of variances was **VIOLATED** (Levene's p = 0.00098; BP p = 1.81e-05).
- **Conclusion:** Although the ANOVA shows a significant p-value, the model itself is statistically invalid because it fails a key assumption. Therefore, the post-hoc test results (which were run) are unreliable and should be interpreted with extreme caution.

Note on Normality: For all four models, the Shapiro-Wilk test returned a $p < 0.05$. This is an expected artifact of the large sample size (N=5000) and is not considered a critical violation, as the ANOVA procedure is robust to slight departures from normality by the Central Limit Theorem. The `qqnorm()` plots confirmed the residuals were approximately normal.

2.4 Part D: Two-Way ANOVA Analysis

A Two-Way ANOVA was fitted.

```
> model5 <- aov(daily_screen_time_hours ~ sleep_cat * age_cat, data=data)
> summary(model5)
                   Df Sum Sq Mean Sq F value Pr(>F)
age_cat             3   1607   535.7 242.597 <2e-16 *
sleep_cat           2   4217  2108.7 955.004 <2e-16 *
age_cat:sleep_cat   6      6     1.0   0.453  0.843
Residuals        4988  11014     2.2
```

Interpretation:

- **Main Effects:** Both `age_cat` ($p < 2e-16$) and `sleep_cat` ($p < 2e-16$) have highly significant main effects on daily screen time.
- **Interaction Effect:** The `age_cat:sleep_cat` interaction is not significant ($p = 0.843$). This indicates that the effect of sleep quality on screen time is the same for all generations, and vice-versa.

All assumptions for the Two-Way ANOVA were met.

- **Homogeneity of Variances (Levene):** $p = 0.2265$ (Passed)
- **Independence of Errors (Durbin-Watson):** $p = 0.7153$ (Passed)
- **Normality of Residuals (Shapiro-Wilk):** $p = 0.009574$. This is considered a pass due to the large N=5000 and visual confirmation from plots.

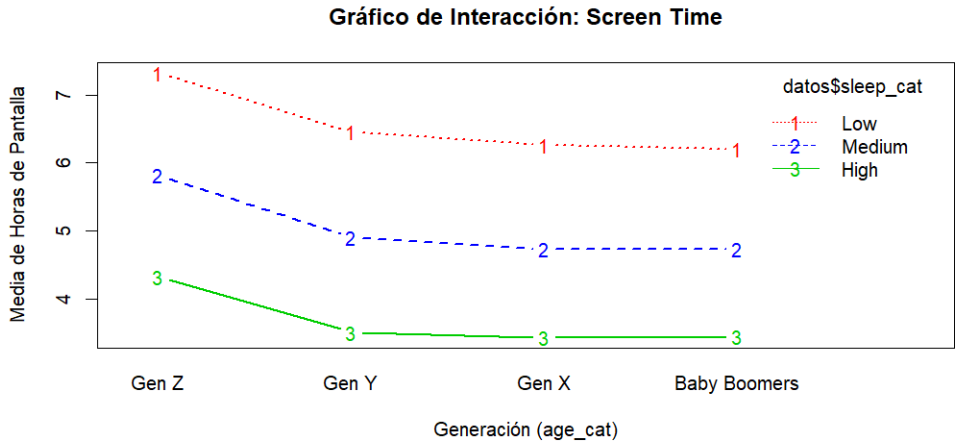


Figure 2.3: Interaction Plot of Screen Time

The plot in Figure 5 visually confirms the ANOVA result. The lines for 'Low', 'Medium', and 'High' sleep quality are almost perfectly parallel, demonstrating the lack of a significant interaction.

Chapter 3

QUESTION 3: REGRESSION ANALYSIS

3.1 Part A: The Best Simple Linear Regression Model

Firstly, we visualized the correlation between the dependent variable `mental_health_score` and all other variables by plotting scatter plots. We also computed the correlation coefficients using the `cor` function. The variable most correlated with `mental_health_score` was `stress_level` with a correlation of $r = -0.9038$. However, since this variable has already been converted into a categorical variable `stress_cat` and it is a discrete variable with only 10 possible values (1–10), we decided to ignore it and instead selected the next most correlated numerical variable, `social_media_hours`, with a correlation of $r = -0.8623$.

A scatter plot was used to visualize the linear relationship between `mental_health_score` and `social_media_hours`. Based on this, the first regression model was fitted as follows:

```
reg1 <- lm(mental_health_score ~ social_media_hours, data = data)
```

The regression results are as follows: Intercept = 95.51, Slope (*social_media_hours*) = -9.38, Residual standard error = 6.63, and $R^2 = 0.7436$.

Since the p-value for `social_media_hours` is $< 2 \times 10^{-16}$, we reject the null hypothesis that the slope is zero, indicating a significant negative relationship between social media usage and mental health.

Next, we checked the assumptions of the linear regression model:

1. **Independence of observations:** Durbin-Watson test yielded a p-value of 0.5871. There is no evidence to reject the null hypothesis, hence the independence assumption is satisfied.
2. **Normality of residuals:** Shapiro-Wilk test gave a p-value of 0.2103, indicating no evidence against normality. Additionally, the Q-Q plot of residuals suggested approximate normality.
3. **Homoscedasticity (equal variance of residuals):** The residual plot showed no apparent pattern, and points were roughly scattered in a rectangular shape. The Breusch-Pagan test produced a p-value of 0.249, confirming no evidence to reject the null hypothesis.

All three assumptions are satisfied for this model, so no transformations were necessary. For every additional hour spent on social media per day, the mental health score decreases by approximately 9.38 points on average.

3.2 Part B: Multiple Linear Regression Model

To construct a multiple linear regression model, we considered adding a second numerical predictor to the simple regression model. The next most correlated variable after `social_media_hours` was `mood_rating`. Thus, we initially fitted the following model:

```
reg2 <- lm(mental_health_score ~ social_media_hours + mood_rating, data = data)
```

Before selecting a final model, all four assumptions of multiple linear regression were examined. The first three assumptions (independence, normality of residuals, and homoscedasticity) were satisfied as in the simple regression model. However, we had to evaluate an additional assumption: multicollinearity. Using the `vif(reg2)` function, both predictors (`social_media_hours` and `mood_rating`) exhibited high VIF values (≈ 6.75). This indicates substantial multicollinearity, suggesting that the two predictors explain overlapping variance in the outcome variable. Since transformations would not solve this issue, we excluded `mood_rating` and chose the next most correlated variable: `physical_activity_hours_per_week`.

We then fitted the third regression model:

```
reg3 <- lm(mental_health_score ~ social_media_hours +
           physical_activity_hours_per_week, data = data)
```

Both predictors were statistically significant, with p-values $< 2 \times 10^{-16}$, indicating that neither slope is zero. Next, we tested all four assumptions for this model:

1. **Independence:** The Durbin–Watson test gave a p-value of 0.7587. We do not reject the null hypothesis, so independence is satisfied.

2. **Normality:** The Shapiro–Wilk test produced a p-value of 0.001868. Although this is below 0.05, the Q–Q plot showed that the residuals were approximately normally distributed. Therefore, the normality assumption is considered acceptable.

3. **Homoscedasticity:** The residual plot showed no clear pattern, with points roughly scattered in a rectangular shape. Although the Breusch–Pagan test returned a p-value of 9.38×10^{-5}, the visual inspection suggests no meaningful violation of homogeneity of variance. Thus, the assumption is considered sufficiently met.

4. **Multicollinearity:** The VIF values for both predictors were approximately 2, which is below the commonly used threshold of 5. Therefore, multicollinearity is not a concern in this model.

All four assumptions were deemed acceptable for the third regression model. The coefficient estimates lead to the following interpretations:

- For each additional hour spent on social media per day, the mental health score decreases by approximately 6.67 points, holding physical activity constant.

- For each additional hour of physical activity per week, the mental health score increases by approximately 2.02 points, holding social media usage constant.

We then compared this model (`reg3`) with the simple linear regression model (`reg1`). First, the adjusted R^2 value increased from 74.35% to 80.58%, indicating a notably better fit. An ANOVA comparison also showed a significantly lower RSS for the multiple regression model and a p-value $< 2.2\times10^{-16}$. Thus, we reject the null hypothesis that the coefficient of `physical_activity_hours_per_week` is zero. This confirms that physical activity provides additional explanatory power beyond social media usage alone.

The multiple regression model (`reg3`) is preferred because it offers a significantly better fit and explains the variability in mental health scores more accurately.

3.3 Part C: Extending the Model with Categorical Predictors

In this section, we examined whether adding categorical predictors could improve the explanatory power of the previously selected multiple regression model. We first added the variable `age_cat` to the model:

```
reg3_age <- lm(mental_health_score ~ social_media_hours +
                physical_activity_hours_per_week + age_cat, data = data)
```

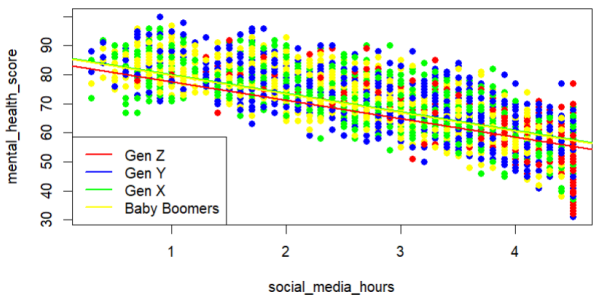
The results from this model indicated the following:

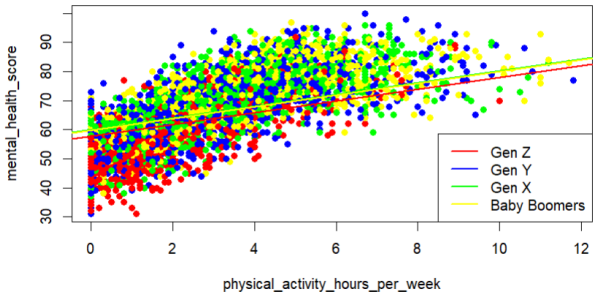
- For every additional hour spent on social media per day, the mental health score decreases by approximately 6.39 points, holding age group and physical activity constant.
- For each additional hour of physical activity per week, the mental health score increases by approximately 2.01 points, holding social media usage and age group constant.
- Compared to the reference group (Gen Z), Gen Y participants score about 2.15 points higher, controlling for social media use and physical activity.
- Gen X participants score approximately 2.29 points higher than Gen Z.
- Baby Boomers score about 2.13 points higher than Gen Z.
- The model explains roughly 81% of the variation in mental health scores.
- The overall regression model is highly significant, with all relevant p-values below 2×10^{-16}.

Since it is not possible to visualize a three-predictor regression model directly in two dimensions, we produced conditional regression plots by holding one predictor constant. Specifically, we generated:

1. Regression lines for all four age categories (*Gen Z*, *Gen Y*, *Gen X*, *Baby Boomers*) while holding `physical_activity_hours_per_week` constant.
2. Regression lines for the same age groups while holding `social_media_hours` constant.

These two figures are presented side by side for comparison:





Next, we fitted a model that included the categorical variable `stress_cat`:

```
reg3_stress <- lm(mental_health_score ~ social_media_hours +
                  physical_activity_hours_per_week + stress_cat, data = data)
```

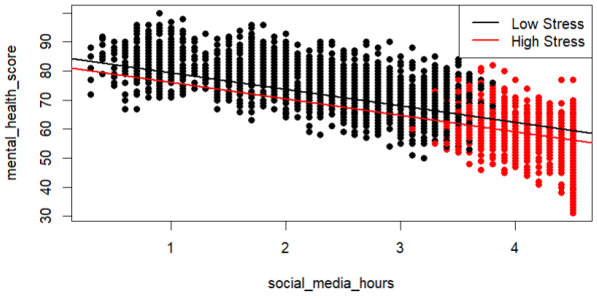
The model results showed:

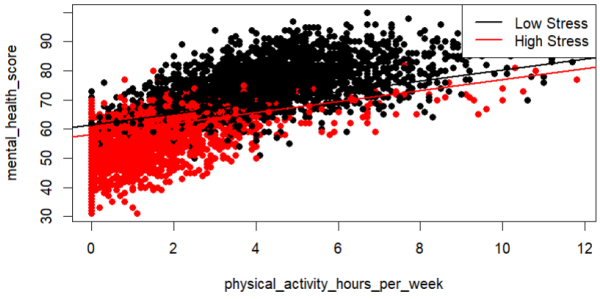
- For every additional hour of social media use per day, the mental health score decreases by approximately 5.67 points, holding physical activity and stress level constant.
- For each additional hour of physical activity per week, the mental health score increases by about 1.88 points, holding social media and stress constant.
- Participants classified as having *High* stress levels score approximately 3.38 points lower than those with *Low* stress, controlling for social media use and physical activity.
- The model explains approximately 81.03% of the variation in mental health scores.
- The overall model is again highly significant, with p-values below 2×10^{-16}.

As before, to visualize the interaction between stress category and the numerical predictors, we produced two sets of conditional regression plots:

1. Regression lines for the two stress groups (*Low* and *High*) while holding `physical_activity_hours_per_week` constant.
2. Regression lines for the same two groups while holding `social_media_hours` constant.

These two figures are also presented side by side for comparison:





3.4 Part D: Choosing the Best Model

To determine which model provided the best overall fit, we compared the adjusted R^2 values of the three candidate models: the baseline multiple regression model (`reg3`), the model including age category (`reg3_age`), and the model including stress category (`reg3_stress`). The adjusted R^2 values obtained were:

- `reg3`: 0.8058
- `reg3_age`: 0.8100
- `reg3_stress`: 0.8103

The baseline model explained approximately 80.6% of the variance in mental health scores. Including `age_cat` increased the adjusted R^2 to 0.8100, although the differences between age groups were relatively small. Incorporating `stress_cat` yielded a slightly higher adjusted R^2 of 0.8103, indicating a marginal improvement beyond the age-extended model.

To further compare the models, we conducted nested model comparisons using the `anova()` function. The comparison between `reg3` and `reg3_age` showed that age category significantly improves the model statistically; however, the magnitude of this improvement is small in practical terms. In contrast, the comparison between `reg3` and `reg3_stress` indicated that stress category improves the model both statistically and practically, with a more substantial contribution relative to age.

Both categorical factors enhance model performance, but stress level provides the most meaningful improvement. Therefore, the model including `stress_cat` is preferred.

3.5 Part E: Final Model Validation

We split the data into a 67% training set and a 33% test set, and then trained the model on the first set and used to make predictions on the second (unseen) set. After, we calculated the Root Mean Square Error (RMSE) was for both.

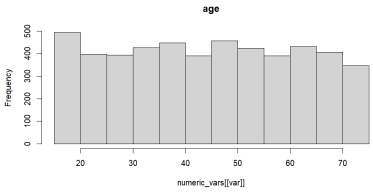
```
> print(paste("RMSE (Train Set):", RMSE_train))
[1] "RMSE (Train Set): 0.0990208680521363"

> print(paste("RMSE (Test Set):", RMSE_test))
[1] "RMSE (Test Set): 0.138923206861641"
```

The RMSE values for both the training (0.099) and test (0.139) sets are extremely low and, more importantly, very similar to each other. This indicates that the model is not overfitting and generalizes well to unseen data.

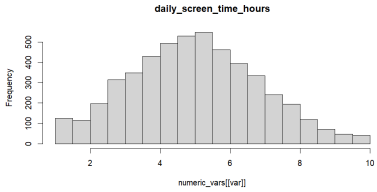
The final model, `model_c_stress`, is validated as both robust and reliable. It is valid for statistical inference (as it meets all assumptions) and valid for prediction (as it demonstrates low error and good generalization).

Appendix A

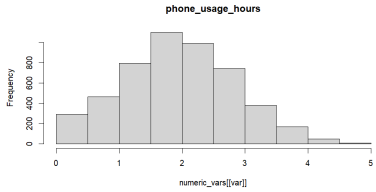
Additional Figures



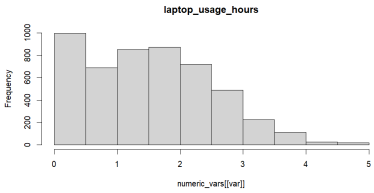
p-value $< 2.2e{-}16$



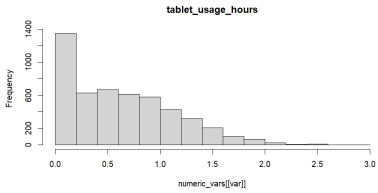
p-value $= 1.877e{-}12$



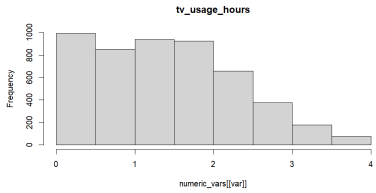
p-value $= 3.068e{-}15$



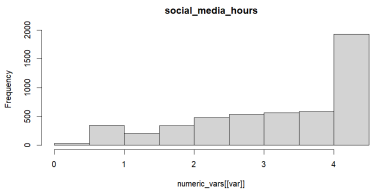
p-value $< 2.2e{-}16$



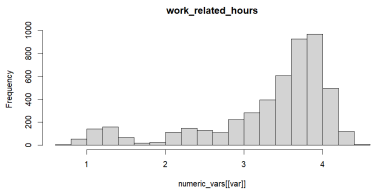
p-value $< 2.2e{-}16$



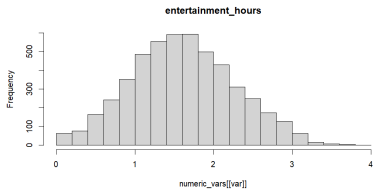
p-value $< 2.2e{-}16$



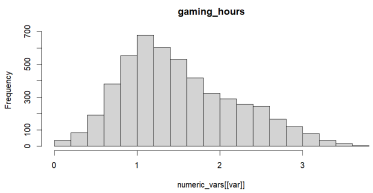
p-value $< 2.2e{-}16$



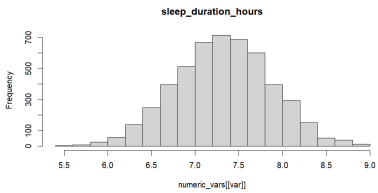
p-value $< 2.2e{-}16$



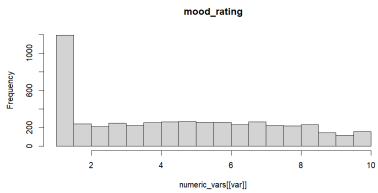
p-value $= 1.144e{-}11$



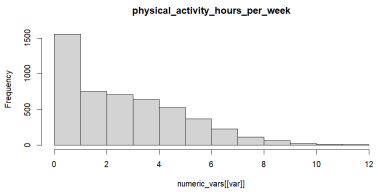
p-value $< 2.2e{-}16$



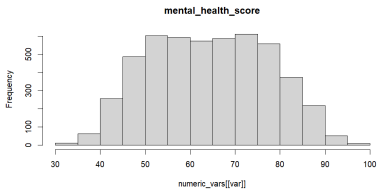
p-value $= 3.413e{-}09$



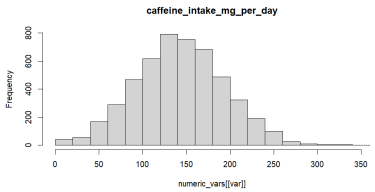
p-value $< 2.2e{-}16$



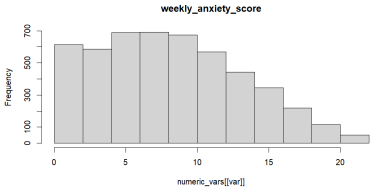
p-value $< 2.2e{-}16$



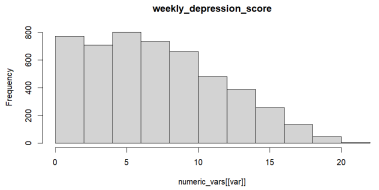
p-value $< 2.2e{-}16$



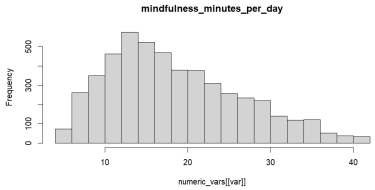
p-value $= 0.1083$



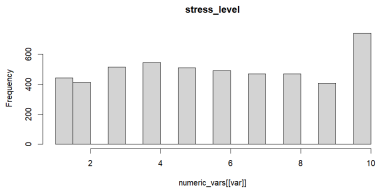
p-value $< 2.2e{-}16$



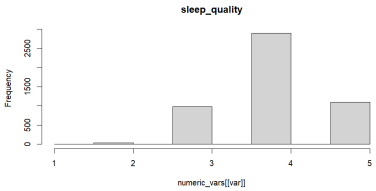
p-value $< 2.2e{-}16$



p-value $< 2.2e{-}16$



p-value $< 2.2e{-}16$



p-value $= 1.877e{-}12$