

Analysis of mHealth App Usage and Medication Adherence

B277873

2025-02-18

1. Introduction

This report presents an analysis of the dataset containing health-related variables. Our objectives are to:

Explore the dataset using descriptive statistics. Test whether age differs across chronic conditions. Evaluate differences in medication adherence between sexes. Investigate the relationship between mHealth app usage and medication adherence. Assess the effect of hospital days on self-reported health (SF-36 score) using regression analysis. All analyses are conducted using tidyverse syntax. Before each statistical test, assumptions (normality and homoscedasticity) are checked via visual methods (histograms and QQ plots), and justifications are provided for the tests chosen.

2. Data Preparation

```
# Load necessary libraries
library(tidyverse)
library(broom)

# Load the dataset (ensure your CSV file is in the working directory)
df <- read.csv("mHealth_assessmentdata_2025.csv")

# Display the first few rows and structure of the dataset
head(df)
```

```
## participant_id sex      age      condition mHealth_app_usage
## 1             1  M 44.39344 Hypertension      Never
## 2             2  M 55.14806 Hypertension      Never
## 3             3  M 71.79078 Hypertension      Never
## 4             4  F 63.95111 Hypertension      Never
## 5             5  M 67.82817 Diabetes        Never
## 6             6  F 63.97655 Hypertension      Rarely
## medication_adherence hospital_days sf36_score
## 1             38.52022             1  36.08940
## 2             47.46258             0  28.13280
## 3             51.13810             2  41.11264
## 4             47.39680             3  41.82250
## 5             48.61806             0  36.36183
## 6             58.00227             1  31.12285
```

```
str(df)
```

```
## 'data.frame':    8954 obs. of  8 variables:
## $ participant_id : int  1 2 3 4 5 6 7 8 9 10 ...
## $ sex            : chr  "M" "M" "M" "F" ...
## $ age            : num  44.4 55.1 71.8 64 67.8 ...
## $ condition      : chr  "Hypertension" "Hypertension" "Hypertension" "Hypertension" ...
## $ mHealth_app_usage : chr  "Never" "Never" "Never" "Never" ...
## $ medication_adherence: num  38.5 47.5 51.1 47.4 48.6 ...
## $ hospital_days   : int   1 0 2 3 0 1 12 0 1 3 ...
## $ sf36_score      : num  36.1 28.1 41.1 41.8 36.4 ...
```

Commentary:

We load the data using `read.csv()` and check its structure with `str()` to understand variable types. Key variables include: age, condition, medication_adherence, hospital_days, sf36_score, and mHealth_app_usage.

#3. Descriptive Statistics ##3.1 Summary Statistics

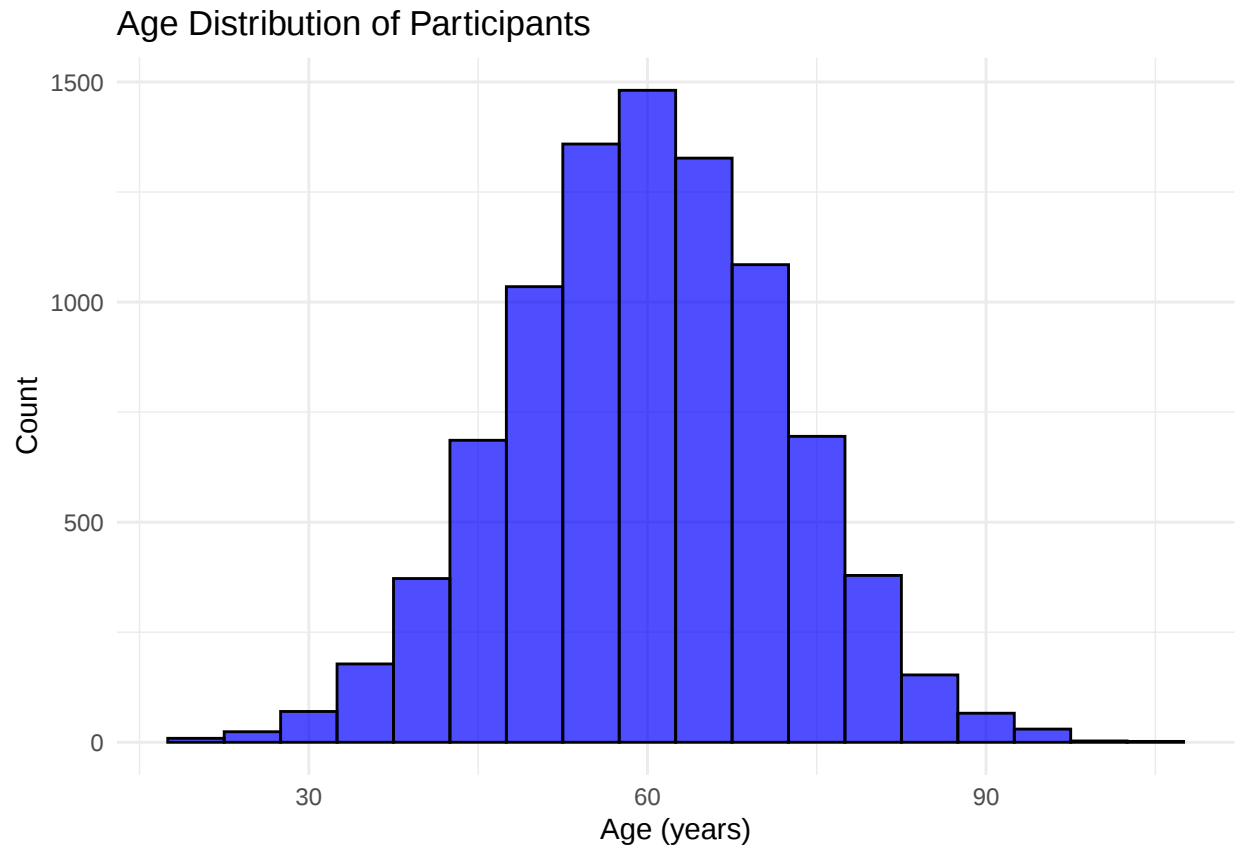
```
# Summary statistics for the entire dataset
summary(df)
```

```
## participant_id      sex            age            condition
## Min.       : 1      Length:8954      Min.       : 18.39      Length:8954
## 1st Qu.:2239      Class :character  1st Qu.: 51.95      Class :character
## Median :4478      Mode  :character  Median : 59.90      Mode  :character
## Mean      :4478                      Mean      : 59.99
## 3rd Qu.:6716                      3rd Qu.: 68.31
## Max.      :8954                      Max.      :106.17
## mHealth_app_usage medication_adherence hospital_days      sf36_score
## Length:8954      Min.       : 24.02      Min.       : 0.000      Min.       :12.84
## Class :character  1st Qu.: 45.22      1st Qu.: 0.000      1st Qu.:39.24
## Mode  :character  Median : 51.84      Median : 1.000      Median :46.02
##                      Mean      : 58.46      Mean      : 1.931      Mean      :46.55
##                      3rd Qu.: 63.81      3rd Qu.: 3.000      3rd Qu.:53.47
##                      Max.      :109.34      Max.      :21.000      Max.      :86.07
```

Commentary: This summary gives an overall view (e.g., mean, median, quartiles) of our variables. It is useful for detecting any anomalies before proceeding to further tests.

3.2 Distribution of Age

```
# Histogram for age distribution
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Age Distribution of Participants", x = "Age (years)", y = "Count") +
  theme_minimal()
```



Commentary:

The histogram Vetter (2017) provides a visual check of the age distribution. A roughly bell-shaped curve would indicate approximate normality.

4. Hypothesis Testing

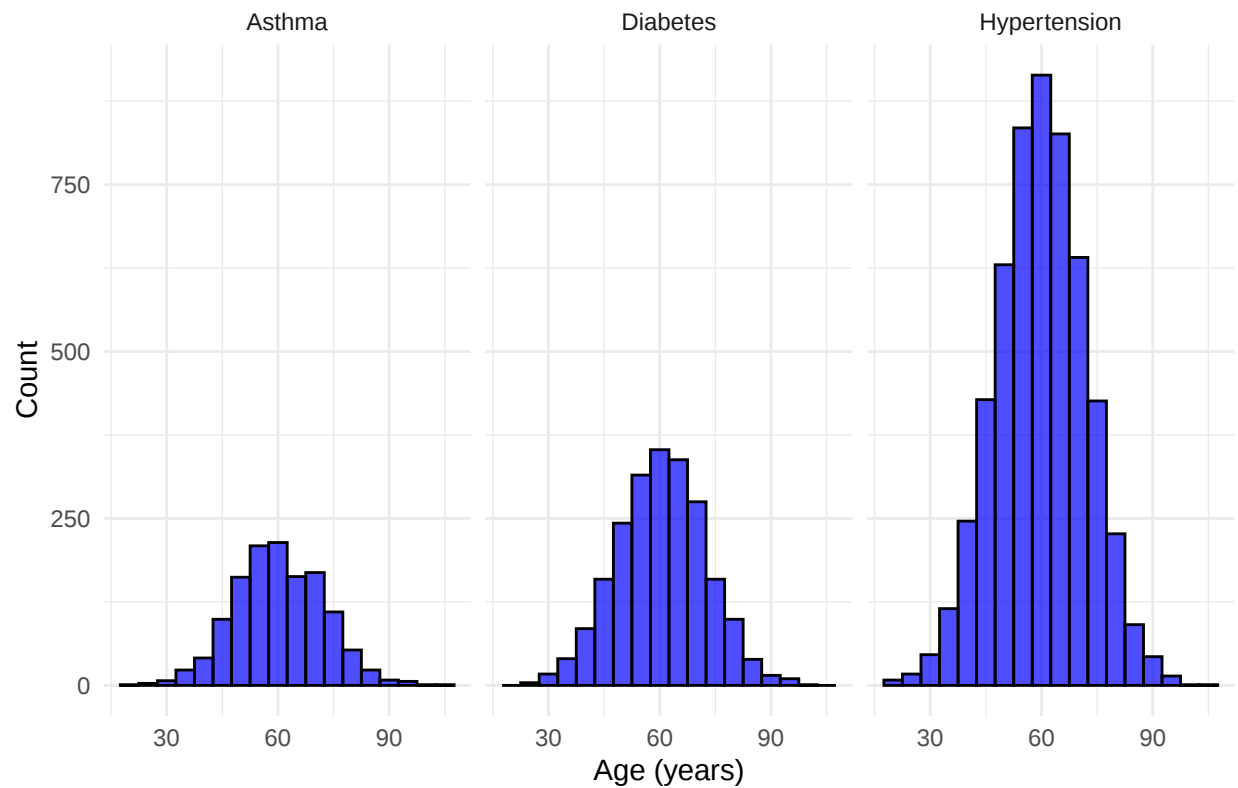
Before performing any tests, we verify key assumptions via visualizations.

4.1 Testing Age Differences Across Conditions

Assumption Check: Normality by Condition

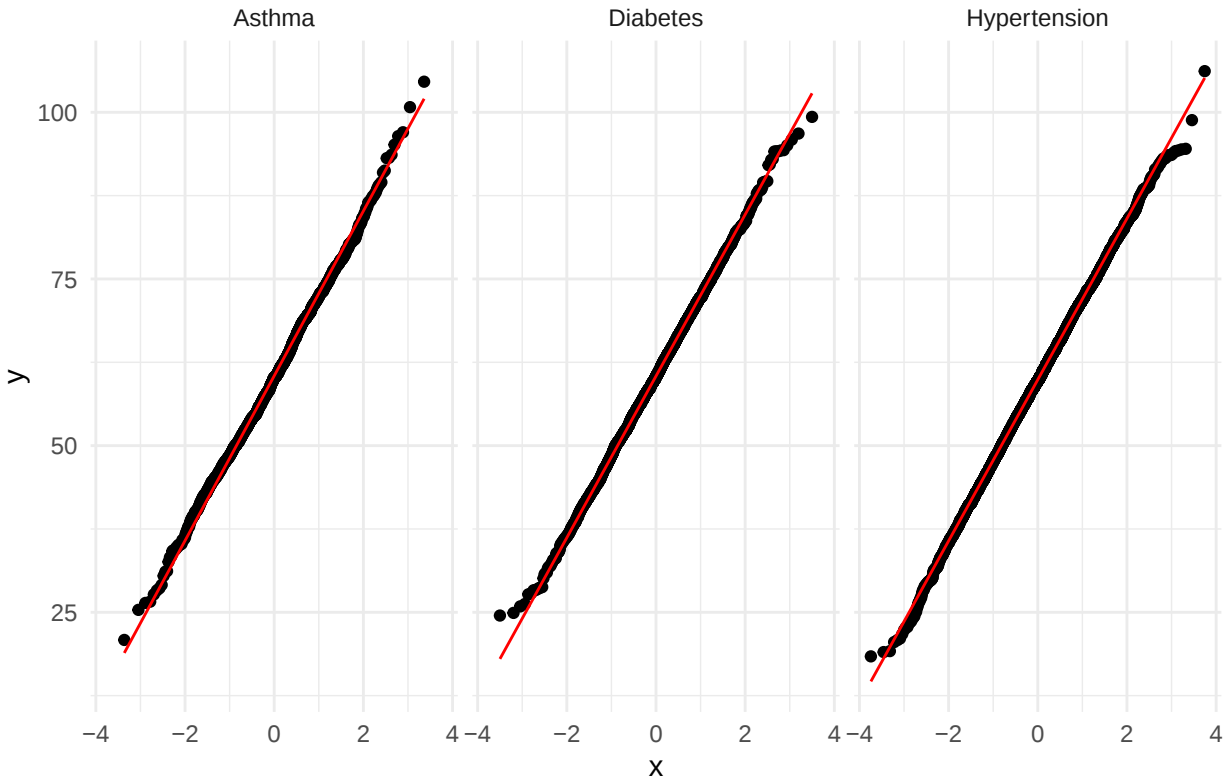
```
# Faceted histograms to inspect age distribution within each condition
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  facet_wrap(~ condition) +
  labs(title = "Age Distribution by Condition", x = "Age (years)", y = "Count") +
  theme_minimal()
```

Age Distribution by Condition



```
# Faceted QQ plots for age within each condition
ggplot(df, aes(sample = age)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  facet_wrap(~ condition) +
  labs(title = "QQ Plot of Age by Condition") +
  theme_minimal()
```

QQ Plot of Age by Condition



Justification: The visual inspection (histograms and QQ plots) suggests that the age distribution for each condition is approximately normal. Hence, a parametric test (ANOVA) is appropriate.

ANOVA Chatzi and Doody (2023) : Age Across Conditions

```
# Conduct a one-way ANOVA to test if age differs across conditions
anova_age <- aov(age ~ condition, data = df)
summary(anova_age)
```

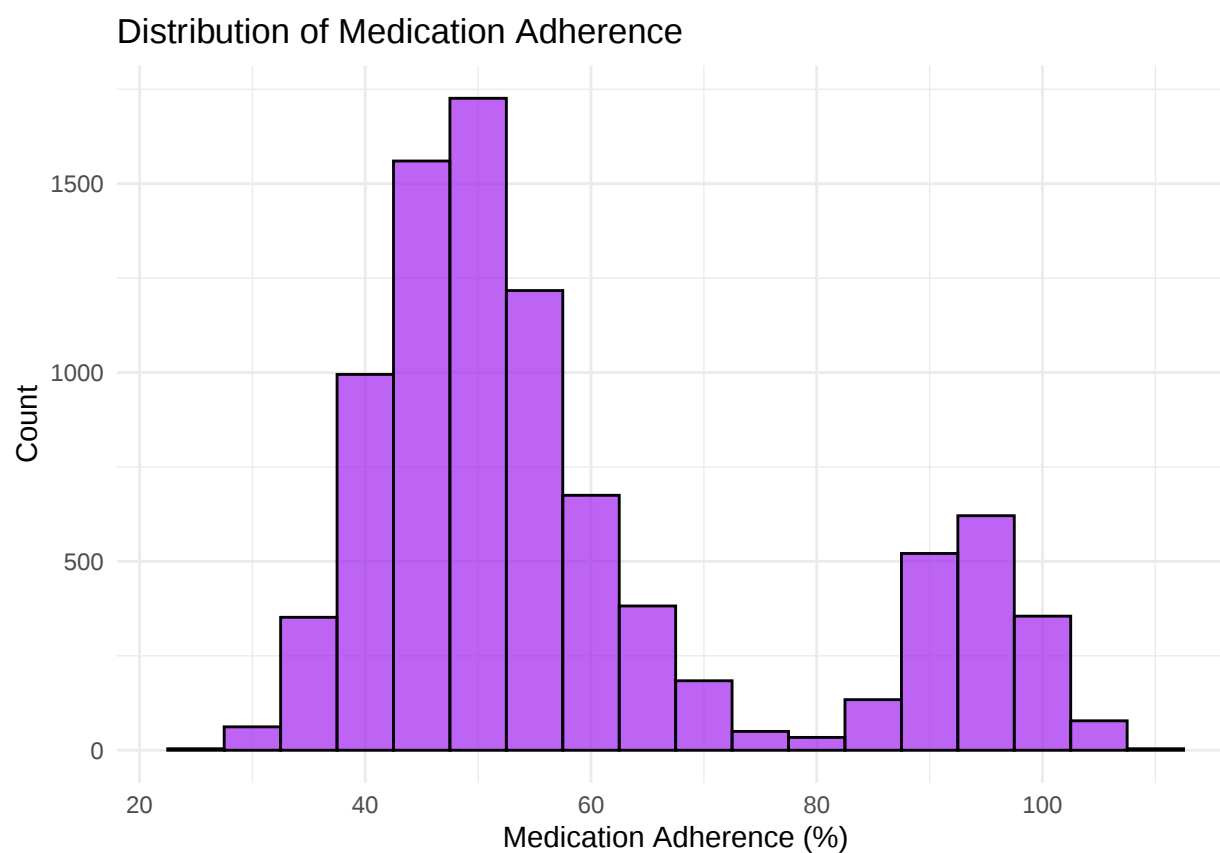
```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## condition      2     860   430.1    2.982 0.0507 .
## Residuals 8951 1291070   144.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reporting & Interpretation:

Output Example: $F(2, 8951) = 2.982$, $p = 0.0507$ Interpretation: The p-value is borderline (approximately 0.051), suggesting a potential trend in age differences across conditions, though not statistically significant at the 0.05 level. Justification: ANOVA Chatzi and Doody (2023) was chosen because the normality assumption appears to hold based on our visual checks. ## 4.2 Testing Medication Adherence Differences Between Sexes Assumption Check: Distribution of Medication Adherence

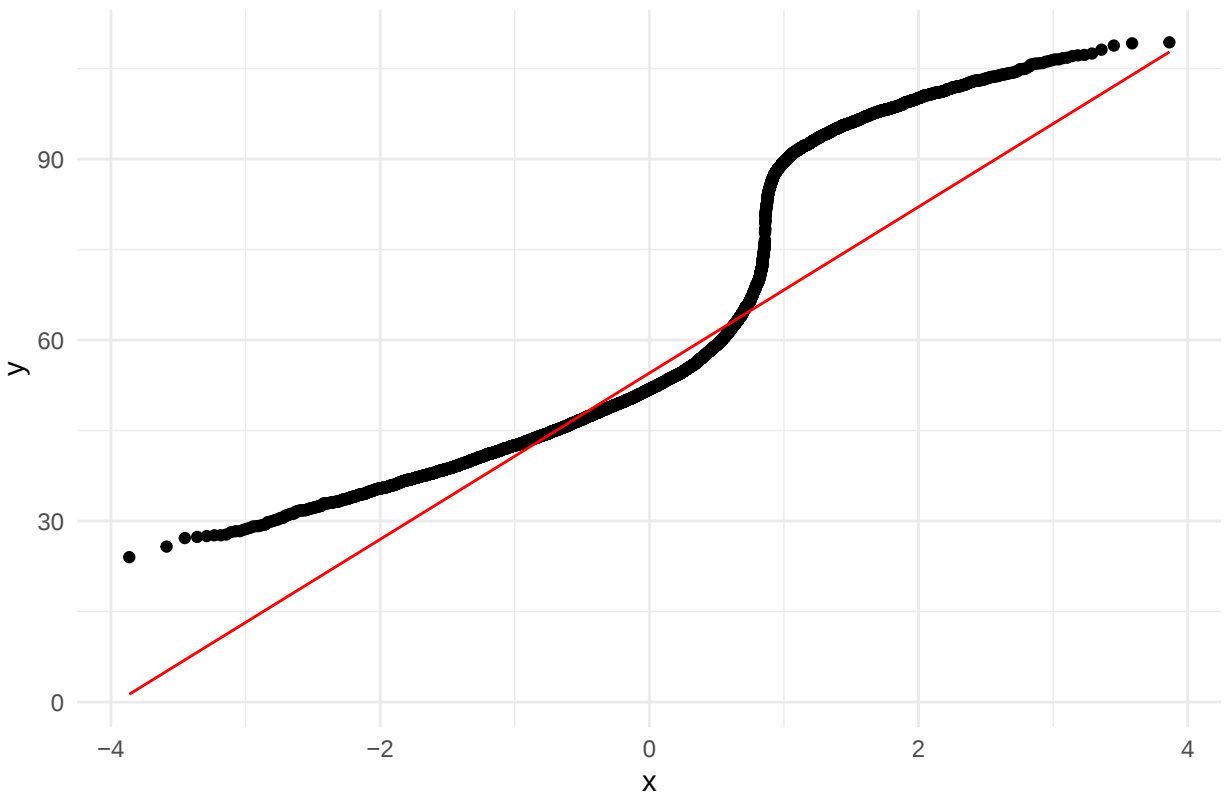
```
# Histogram for medication adherence
ggplot(df, aes(x = medication_adherence)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "black", alpha = 0.7) +
```

```
labs(title = "Distribution of Medication Adherence", x = "Medication Adherence (%)", y = "Count") +  
theme_minimal()
```



```
# QQ Plot for medication adherence  
ggplot(df, aes(sample = medication_adherence)) +  
  stat_qq() +  
  stat_qq_line(color = "red") +  
  labs(title = "QQ Plot of Medication Adherence") +  
  theme_minimal()
```

QQ Plot of Medication Adherence



Justification: Visual checks (histogram and QQ plot Habibzadeh (2024)) indicate that the distribution of medication adherence may not be perfectly normal. To be conservative, we opt for a non-parametric test.

Wilcoxon Rank-Sum Test: Adherence by Sex

```
# Wilcoxon test to compare medication adherence between males and females
wilcox_result <- wilcox.test(medication_adherence ~ sex, data = df)
wilcox_result
```

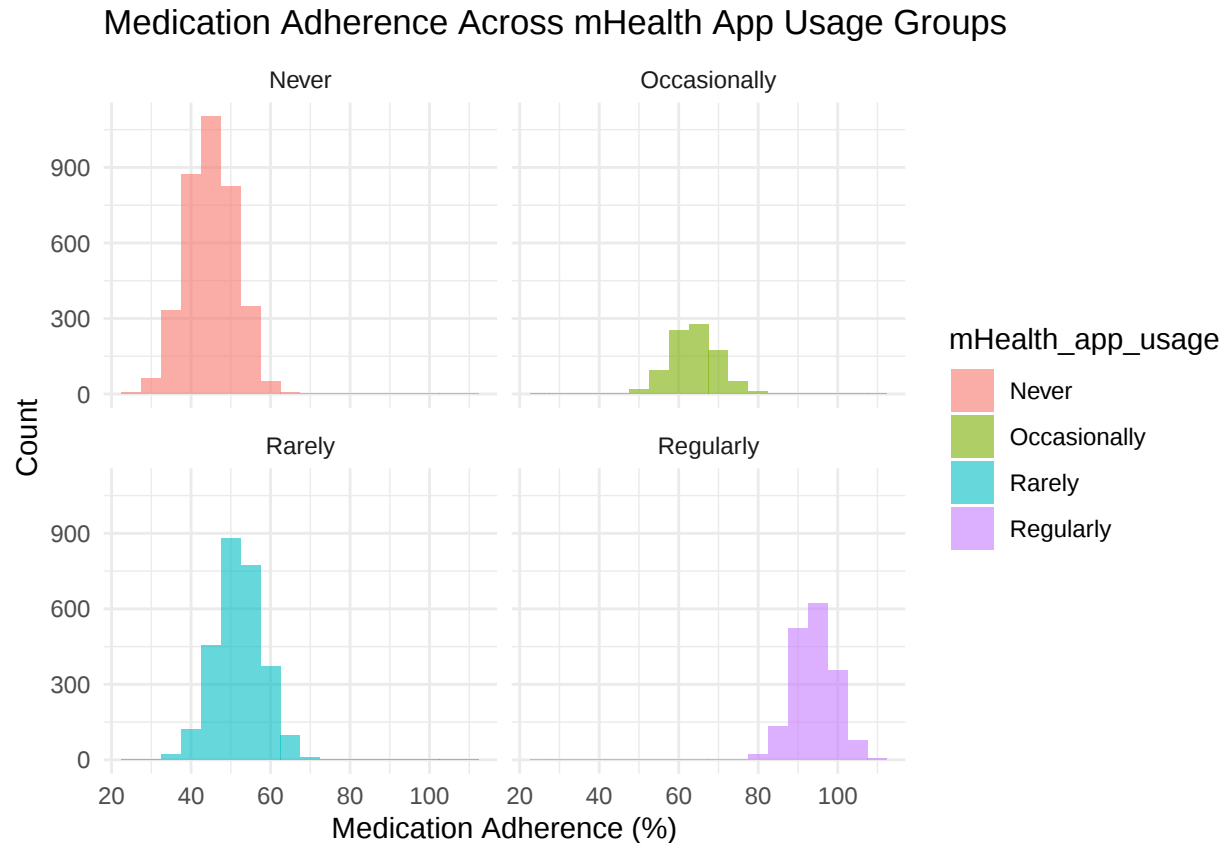
```
##
## Wilcoxon rank sum test with continuity correction
##
## data: medication_adherence by sex
## W = 9885033, p-value = 0.2641
## alternative hypothesis: true location shift is not equal to 0
```

Reporting & Interpretation:

Output Example: p-value = 0.2641 Interpretation: With a p-value > 0.05 , there is no statistically significant difference in medication adherence between males and females. Justification: The non-parametric Wilcoxon test was used due to concerns about normality in the medication adherence variable. ##4.3 Testing the Relationship Between mHealth App Usage and Medication Adherence Assumption Check: Distribution of Adherence by App Usage

```
# Histogram faceted by mHealth app usage
ggplot(df, aes(x = medication_adherence, fill = mHealth_app_usage)) +
```

```
geom_histogram(binwidth = 5, alpha = 0.6, position = "identity") +
facet_wrap(~ mHealth_app_usage) +
labs(title = "Medication Adherence Across mHealth App Usage Groups", x = "Medication Adherence (%)", y = "Count") +
theme_minimal()
```



Justification: The histograms Vetter (2017) suggest that within each app usage group, medication adherence is roughly normally distributed. Therefore, we use ANOVA Chatzi and Doody (2023) to test for differences among groups.

ANOVA: Medication Adherence Across App Usage

```
# ANOVA to test differences in medication adherence among app usage groups
anova_adherence <- aov(medication_adherence ~ mHealth_app_usage, data = df)
summary(anova_adherence)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## mHealth_app_usage    3 2998505   999502   29854 <2e-16 ***
## Residuals          8950   299643     33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reporting & Interpretation:

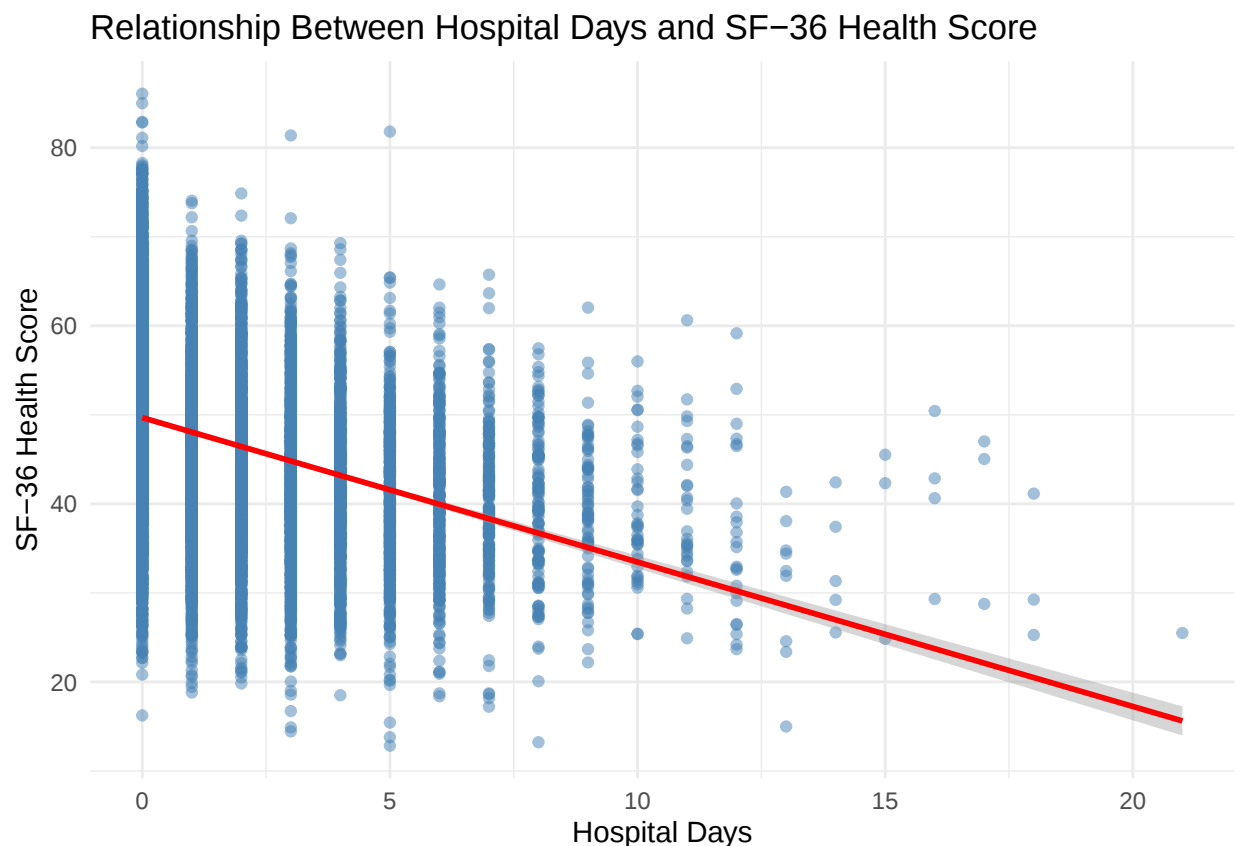
Output Example: $F(3, 8950) = 29854$, $p < 2e-16$ Interpretation: The highly significant p-value indicates that medication adherence significantly differs across the different mHealth app usage categories. Justification: Given the normality assumption was acceptable for these groups, ANOVA Chatzi and Doody (2023) is appropriate.

5. Regression Analysis

Hospital Days & SF-36 Health Score Before running regression, we check assumptions including linearity, normality of residuals, and homoscedasticity.

5.1 Visualizing the Relationship

```
# Scatter plot with regression line: Hospital Days vs. SF-36 Score
ggplot(df, aes(x = hospital_days, y = sf36_score)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Relationship Between Hospital Days and SF-36 Health Score",
       x = "Hospital Days", y = "SF-36 Health Score") +
  theme_minimal()
```



Commentary:

The scatter plot provides an initial view of the relationship. A negative trend would suggest that more hospital days are associated with lower health scores. ## 5.2 Fitting the Linear Regression Model Schober and Vetter (2020)

```
# Fit a simple linear regression model
model_simple <- lm(sf36_score ~ hospital_days, data = df)
summary(model_simple)
```

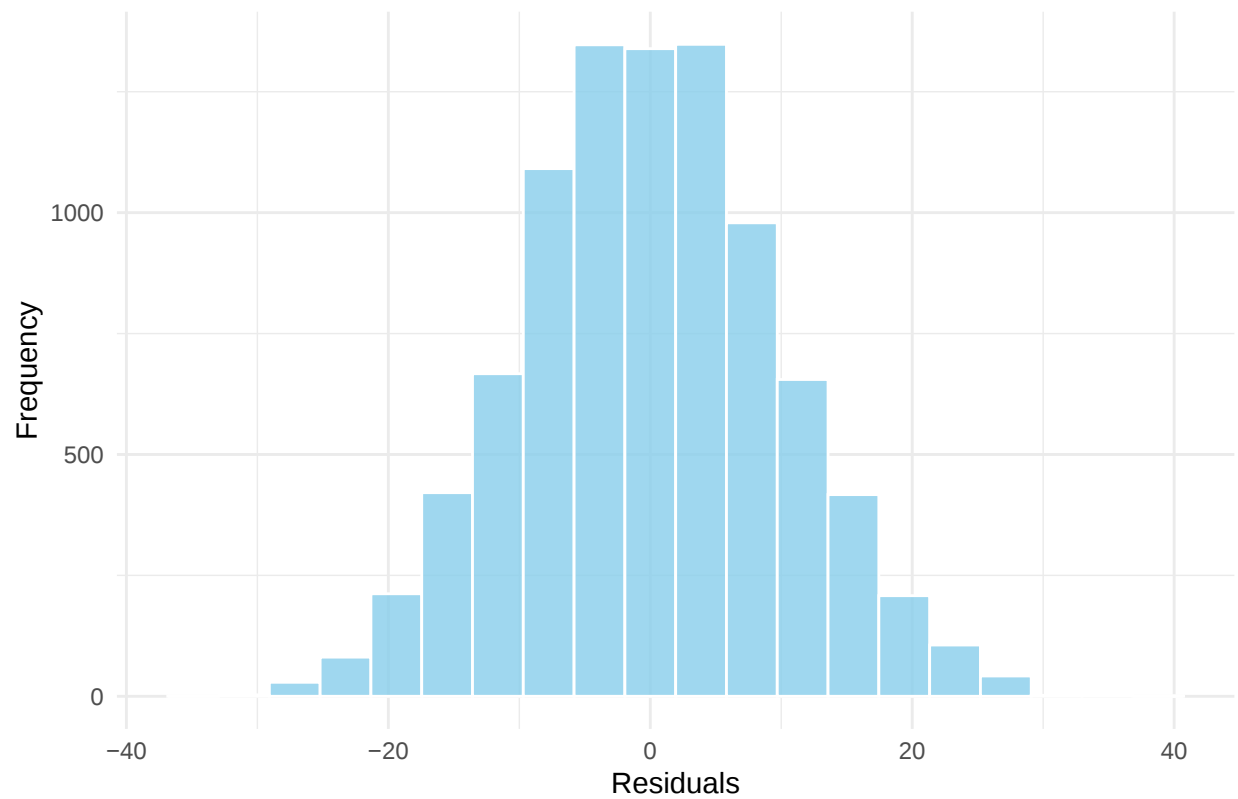
```
##
## Call:
## lm(formula = sf36_score ~ hospital_days, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -33.448  -6.565  -0.118   6.455  40.237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.68567    0.13339  372.49  <2e-16 ***
## hospital_days -1.62207    0.04354  -37.25  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.799 on 8952 degrees of freedom
## Multiple R-squared:  0.1342, Adjusted R-squared:  0.1341
## F-statistic: 1388 on 1 and 8952 DF,  p-value: < 2.2e-16
```

Reporting & Interpretation:

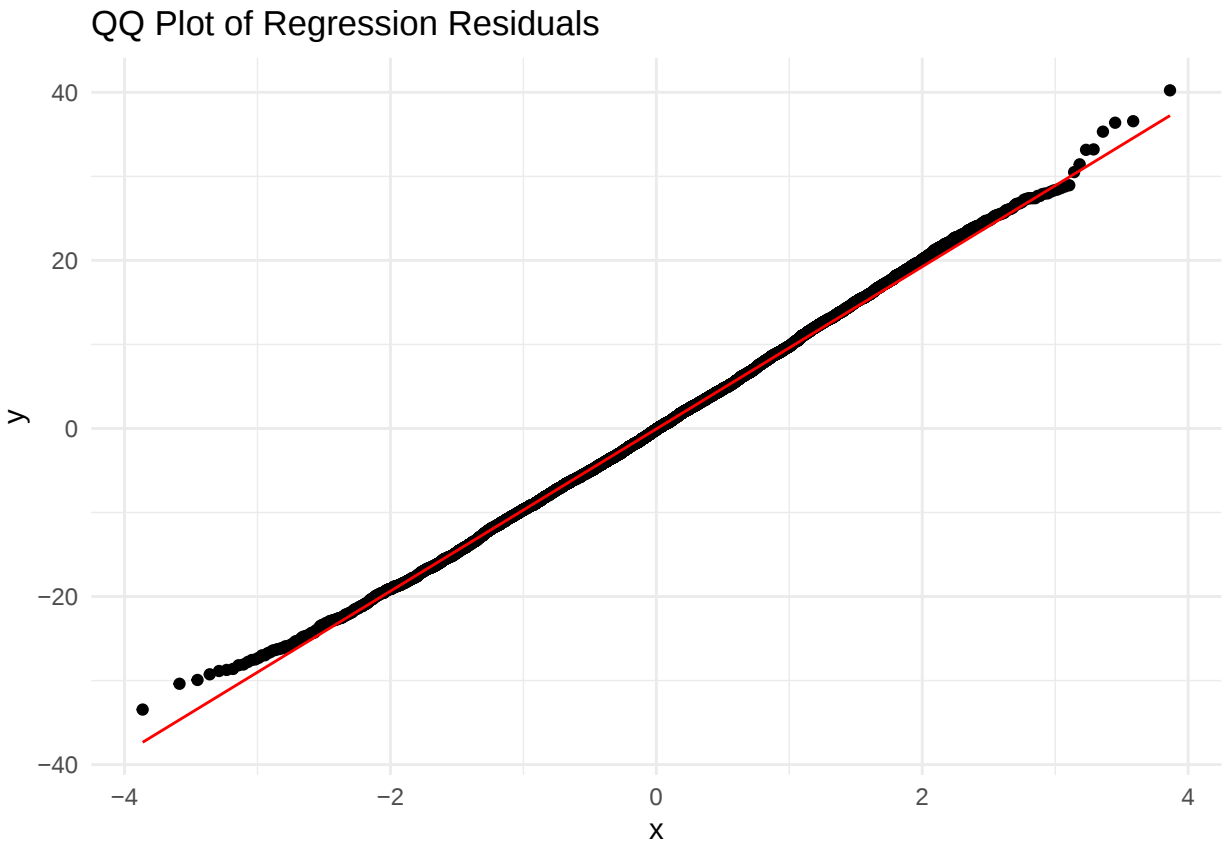
Output Example: Intercept is approximately 49.69 ($p < 2e-16$) Slope is approximately -1.62 ($p < 2e-16$): For each additional hospital day, the SF-36 score decreases by approximately 1.62 points. Justification: Regression is appropriate here because both variables are continuous and the scatter plot suggests a linear relationship. ## 5.3 Checking Regression Assumptions Schober and Vetter (2020) Normality of Residuals

```
# Histogram of residuals
ggplot(data.frame(resid = residuals(model_simple)), aes(x = resid)) +
  geom_histogram(fill = "skyblue", color = "white", bins = 20, alpha = 0.8) +
  labs(title = "Histogram of Regression Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```

Histogram of Regression Residuals



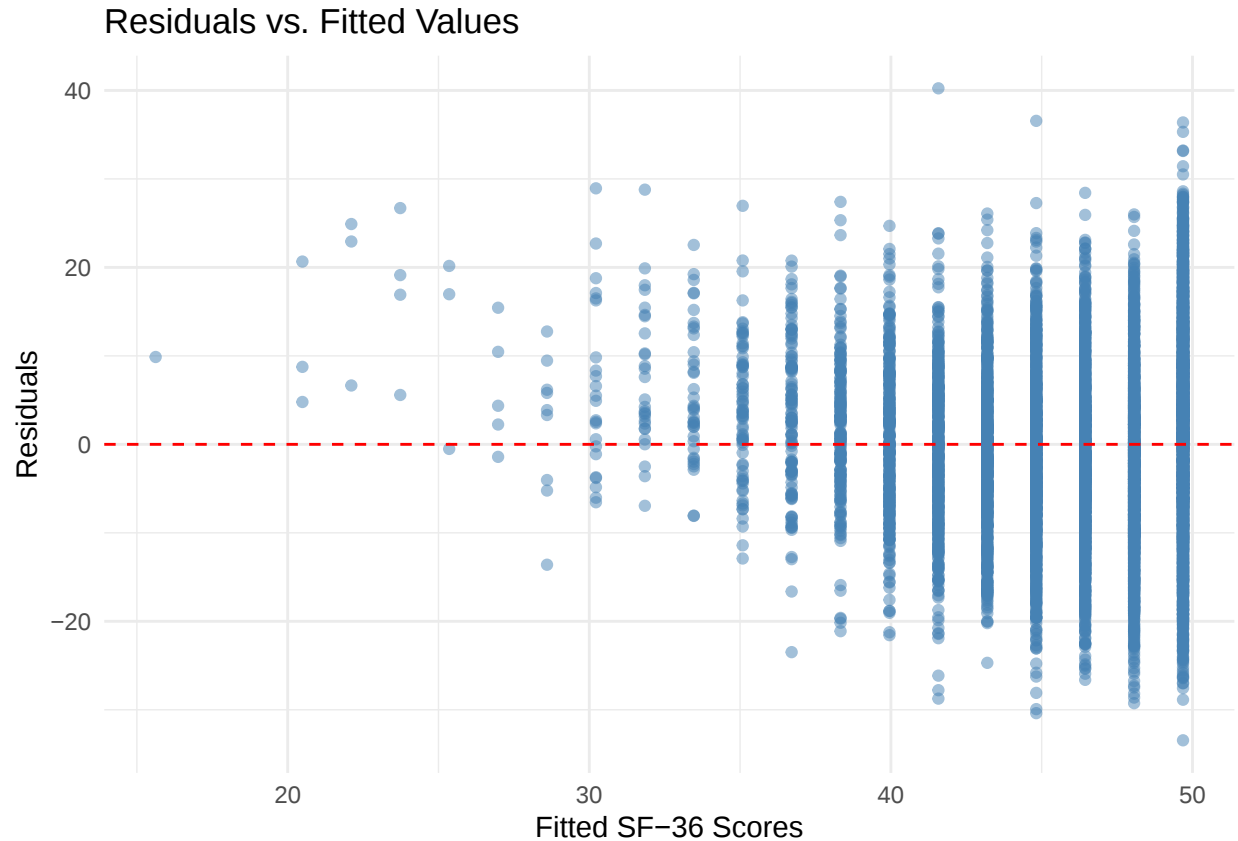
```
# QQ plot of residuals
ggplot(data.frame(resid = residuals(model_simple)), aes(sample = resid)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "QQ Plot of Regression Residuals") +
  theme_minimal()
```



Commentary:

Both plots help assess if the residuals are approximately normally distributed. A roughly bell-shaped histogram Vetter (2017) and points following the diagonal in the QQ plot Habibzadeh (2024) indicate that the normality assumption is met. Homoscedasticity

```
# Residuals vs. Fitted plot to check for homoscedasticity
ggplot(data.frame(fitted = fitted(model_simple), resid = residuals(model_simple)),
  aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted Values", x = "Fitted SF-36 Scores", y = "Residuals") +
  theme_minimal()
```



Commentary:

The residuals vs. fitted plot helps us determine if there is constant variance (homoscedasticity). A random scatter of points without a clear pattern suggests homoscedasticity; any funnel shape would indicate heteroscedasticity.

Conclusion

In this report, we have:

Described the dataset and provided summary statistics. Verified assumptions (normality and homoscedasticity) using histograms and QQ plots. Conducted ANOVA Chatzi and Doody (2023) tests to examine differences in age across conditions and medication adherence across mHealth app usage groups. Employed the Wilcoxon test to assess differences in medication adherence between sexes due to non-normality. Performed a regression analysis Schober and Vetter (2020) showing that increased hospital days are significantly associated with lower SF-36 health scores.

part of this report was created using ChatGPT

References

- Chatzi, Anna, and Owen Doody. 2023. "The One-Way ANOVA Test Explained." *Nurse Researcher* 31 (3): 8–14. <https://doi.org/10.7748/nr.2023.e1885>.
- Habibzadeh, Farrokh. 2024. "Data Distribution: Normal or Abnormal?" *Journal of Korean Medical Science* 39 (3): e35. <https://doi.org/10.3346/jkms.2024.39.e35>.
- Schober, Patrick, and Thomas R. Vetter. 2020. "Linear Regression in Medical Research." *Anesthesia and Analgesia* 132 (1): 108–9. <https://doi.org/10.1213/ANE.0000000000005206>.

Vetter, Thomas R. 2017. “Fundamentals of Research Data and Variables: The Devil Is in the Details.”
Anesthesia and Analgesia 125 (4): 1375–80. <https://doi.org/10.1213/ANE.0000000000002370>.