Data 602 - Project

Josh Brauner, Raahim Salman, Ze Yu

Feburary 14th, 2024

Part 1

Data Wrangling

In the initial phase of our project, the primary goal is to prepare the dataset for analysis. This preparation involves loading the data and then removing data points that could potentially skew or bias our results. Specifically, the dataset encompasses annual home run rates (home runs per at bat) for Barry Bonds, covering several seasons. Per the project's guidelines, it's imperative to exclude the 2001 season—the year Barry Bonds hit a record 73 home runs—from our dataset before proceeding with any further analysis. The rationale behind excluding this particular season is rooted in its outlier status within the scope of our investigation, possibly due to factors such as intentional walks, which are not reflective of the typical performance trends we aim to analyze.

```
bonds_data <- read.csv("bondsdata.csv")
bonds_data_filtered <- subset(bonds_data, season != 2001)
tail(bonds_data_filtered)
```

```
##    season     hrat
## 9    1995 0.065217
## 10   1996 0.081238
## 11   1997 0.075188
## 12   1998 0.067029
## 13   1999 0.095775
## 14   2000 0.102083
```

The removal of the 2001 season data point is a critical step in ensuring a more accurate and unbiased evaluation of Bonds' performance trends over the years leading up to this extraordinary season. By focusing on the years preceding 2001, we aim to construct a statistical model that predicts Bonds' home run rates without the influence of this anomalous year. This approach is intended to provide a clearer perspective on his performance trajectory, allowing us to assess whether there is evidence of an unusual improvement that could be attributed to external factors, such as the speculated use of steroids.

Model Building

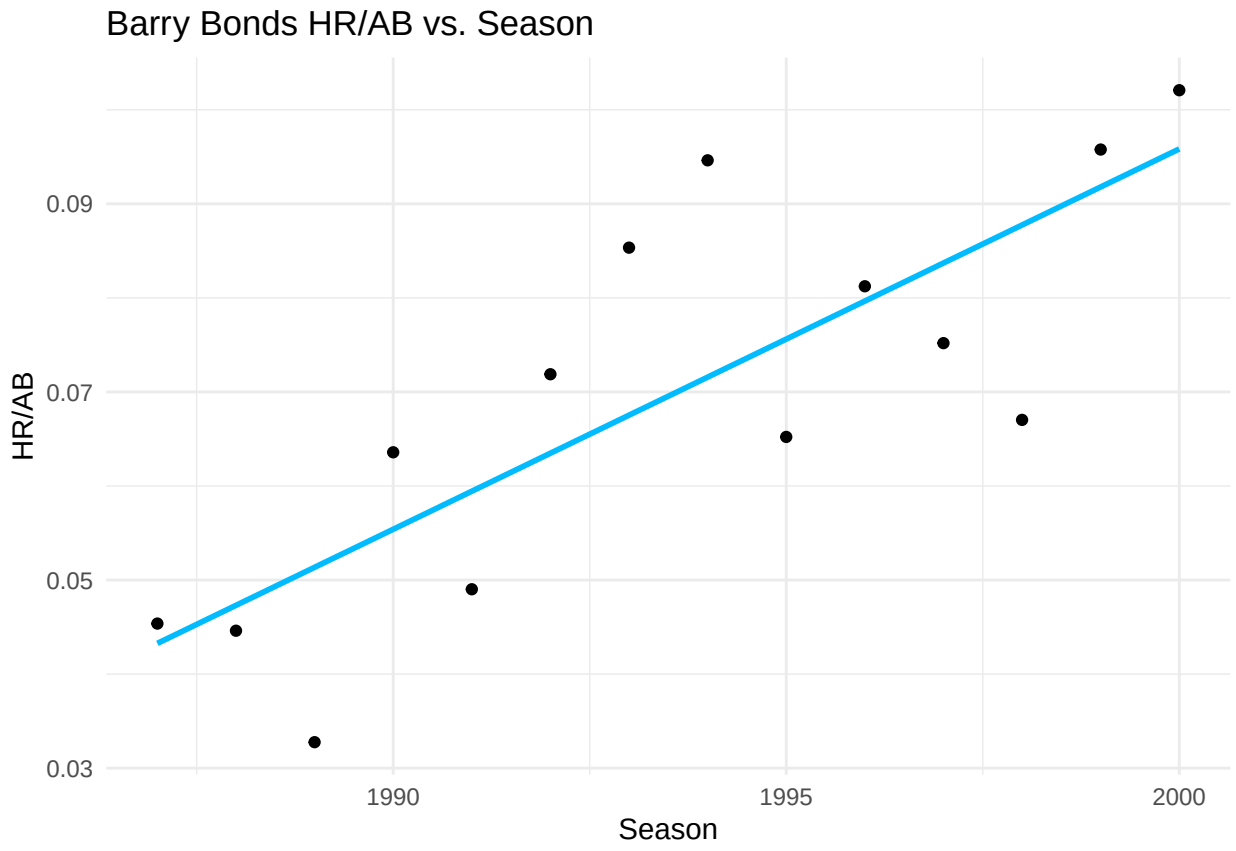
The statistical model we plan to build, $HRAT_i = A + B \times Year_i + e_i$, where $HRAT_i$ represents the home run rate in year i, and $Year_i$ denotes the year of the season, is designed to elucidate the relationship between time (years) and Bonds' home run rates. Through quantifying the trend over the specified years, this linear regression model serves as a tool for analyzing potential significant deviations in performance. Such deviations, if present, could align with steroid use, under the premise that such use would manifest as an atypical increase in home run rates.

```r
model <- lm(hrat ~ season, data = bonds_data_filtered)
model$coef
```

```
##  (Intercept)       season
## -7.992499290  0.004044169
```

The linear regression equation representing the model is $\hat{HRAT_i} = -7.992 + 0.004 \times Year_i$.

```r
ggplot(bonds_data_filtered, aes(x = season, y = hrat)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "#00bbff") +
  labs(title = "Barry Bonds HR/AB vs. Season", x = "Season", y = "HR/AB") +
  theme_minimal()
```



Importantly, to ensure the integrity of the conclusions drawn from this model, we will verify the assumptions underlying linear regression analysis, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of error terms. These checks are fundamental to confirming that our model is accurately specified and that the inferences and predictions derived from it are reliable.

Model Assumption and Validation

Correlation Coefficient Check:

```r
cor(bonds_data_filtered$season, bonds_data_filtered$hrat)
```

```
## [1] 0.7981544
```

A correlation coefficient of 0.7981544 indicates a strong positive linear relationship between the year and the home runs per at bat (HR/AB) for Barry Bonds, excluding the 2001 season.

Significance of Coefficient Estimates

In the context of analyzing the significance of the coefficient estimate for the year in predicting home runs per at bat (HR/AB) for Barry Bonds (excluding the 2001 season), we can formulate the null hypothesis (H_0) and the alternative hypothesis (H_1) as follows, with an alpha level (α) of 0.05:

Null hypothesis: $H_0 : \beta_1 = 0$, $H_0 : \beta_0 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$, $H_A : \beta_0 \neq 0$

```r
summary(model)
```

```
##
## Call:
## lm(formula = hrat ~ season, data = bonds_data_filtered)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020722 -0.009931  0.001841  0.007701  0.023055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.9924993  1.7566775  -4.550 0.000666 ***
## season       0.0040442  0.0008812   4.589 0.000622 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01329 on 12 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.6068
## F-statistic: 21.06 on 1 and 12 DF,  p-value: 0.0006222
```

```r
coef(summary(model))[, "t value"]
```

```
## (Intercept)      season
##   -4.549782    4.589384
```

```r
coef(summary(model))[, "Pr(>|t|)"]
```

```
##  (Intercept)       season
## 0.0006664296 0.0006222474
```

Intercept

- **t-value for Intercept**: -4.549782
- **p-value for Intercept**: 0.0006664296

The intercept's t-value is significantly negative, and the corresponding p-value is much less than the alpha level of 0.05. This statistically significant result suggests that the intercept is significantly different from zero. Therefore we reject the null hypothesis $H_0 : B_0 = 0$ in favor of the alternative hypothesis $H_1 : B_0 \neq 0$\$.

Season

- **t-value for Season**: 4.589384
- **p-value for Season**: 0.0006222474

The t-value for the year coefficient is significantly positive, indicating a positive relationship between the year and HR/AB ratio for Barry Bonds in the dataset analyzed. The p-value associated with this t-value is much less than 0.05, strongly suggesting that we reject the null hypothesis $H_0 : B_1 = 0$ in favor of the alternative hypothesis $H_1 : B_1 \neq 0$.

Implications

The statistical analysis of the year's coefficient reveals that there is a significant linear relationship between the year and the HR/AB ratio. This means that the year significantly predicts the HR/AB ratio for Barry Bonds in the years leading up to 2001, excluding the 2001 season itself. The positive t-value indicates that as the year increases, so does the HR/AB ratio, suggesting an improvement in Bonds' performance in hitting home runs per at-bat attempt over time.

Given the alpha level of 0.05 and the very low p-values obtained for both the intercept and the year coefficient, our analysis provides strong statistical evidence to support the conclusion that there was a significant trend in Barry Bonds' home run rates per at-bat over the years analyzed.

The R-squared value of 0.6371 in your linear regression model indicates that approximately 63.71% of the variance in the dependent variable is explained by the independent variable.

Residual Analysis

Normality of Residuals

In our specific analysis concerning Barry Bonds' home run rates (HR/AB) as a function of the season (year), the normality of residuals implies that the deviations from the predicted home run rates are random and follow a normal distribution. This condition supports the premise that our linear model is appropriately capturing the relationship between the year and HR/AB without systematic bias.

The evidence supporting the normality of residuals in our analysis comes from both graphical and statistical methods:

```
residuals <- resid(model)
residuals_df <- data.frame(Residuals = residuals)

# Generate a Q-Q plot
ggplot(residuals_df, aes(sample = Residuals)) +
  stat_qq() +
  stat_qq_line(colour = "red") +
  ggtitle("Q-Q Plot of Residuals") +
  theme_minimal()
```

In our case, the residuals align closely with the reference line in the Q-Q plot, indicating that their distribution resembles a normal distribution. This evidence suggests that the residuals from our linear model do not deviate significantly from normality.

```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.97132, p-value = 0.8938
```

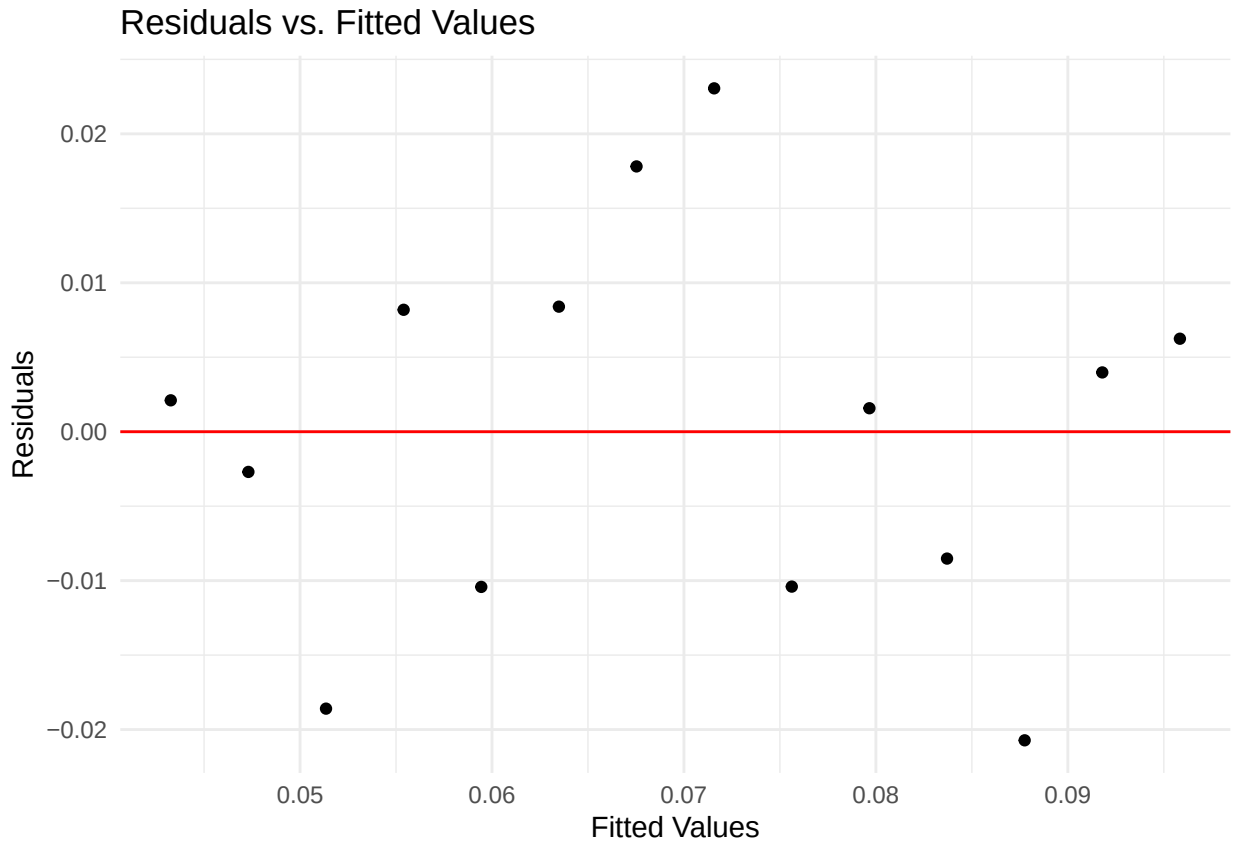
The Shapiro-Wilk test is a statistical test designed to assess the normality of a dataset. In our analysis, the Shapiro-Wilk test for the residuals yields a p-value of 0.8938. Given that this p-value is significantly greater than the conventional threshold of 0.05, we fail to reject the null hypothesis that the residuals are normally distributed. This statistical evidence further substantiates the claim that the residuals of our model are normal.

Homoscedasticity

The examination of homoscedasticity is a critical step in validating the assumptions underlying a linear regression model. Homoscedasticity refers to the condition where the residuals (the differences between observed and predicted values) have constant variance across all levels of the independent variable(s). This assumption ensures that the model's predictive accuracy is uniform across the range of the independent variable, which in this context is the season (year) for Barry Bonds' home run rates (HR/AB).

```r
fitted_values <- fitted(model)
homoscedasticity_df <- data.frame(Fitted = fitted_values, Residuals = residuals)

ggplot(homoscedasticity_df, aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, colour = "red") +
  ggtitle("Residuals vs. Fitted Values") +
  xlab("Fitted Values") +
  ylab("Residuals") +
  theme_minimal()
```




```r
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.04849, df = 1, p-value = 0.8257
```

The Breusch-Pagan (BP) test is a statistical procedure designed to test for heteroscedasticity — the presence of non-constant variance in the error terms of a regression model. In the context of analyzing Barry Bonds' home run rates (HR/AB) as a function of the season (year), the BP test's p-value of 0.8257 significantly exceeds the common alpha level threshold of 0.05. This high p-value indicates that there is insufficient evidence to reject the null hypothesis of the BP test, which states that the error variances are homoscedastic.

The even distribution of residuals across the predicted values, as evidenced by the scatter plot, allows us to conclude that the condition of homoscedasticity holds for our linear regression model. This finding is crucial because it means that our model meets another important assumption of linear regression, reinforcing its validity and the reliability of its predictions and inferences.

Given that our analysis has confirmed both key assumptions of linear regression—normality of residuals and homoscedasticity—we can assert that our linear regression model is valid. This validity implies that the model is well-founded and that the statistical inferences drawn from it, such as confidence intervals and hypothesis tests on the regression coefficients, are based on solid assumptions.

Prediction of Home Run Rate in 2001 Season

```
new_data <- data.frame(season = 2001)
predicted_hrab <- predict(model, newdata = new_data, interval="predict")
print(predicted_hrab)
```

```
##          fit        lwr       upr
## 1 0.09988334 0.06662845 0.1331382
```

The 95% prediction interval for Bonds' HRAT in the 2001 season, ranging from 0.06663 to 0.13314, represents the range within which we would expect Bonds' HRAT to fall, given the trends observed in the data from previous seasons. The fact that Bonds' actual HRAT of 0.153400 falls outside this interval suggests that his performance in 2001 was not only exceptional but also statistically significant in terms of deviation from predicted trends.

While the statistical analysis does not directly address whether Barry Bonds was using steroids in the 2001 season, the significant discrepancy between the predicted and actual HRATs, and the fact that the actual HRAT lies outside the 95% prediction interval, suggest that there were factors influencing Bonds' performance that year beyond what could be expected based on historical trends.

In conclusion, while our linear regression model and subsequent analysis provide valuable insights into Barry Bonds' performance trends leading up to the 2001 season, they also highlight an extraordinary deviation in his 2001 performance that warrants further investigation. The statistical evidence suggests that Bonds' HRAT in 2001 was significantly higher than expected based on past performance, pointing to the presence of additional factors that contributed to this anomaly.

It's important to note that statistical analysis alone cannot prove the use of performance-enhancing substances. Such conclusions would require corroborative evidence beyond the scope of this statistical model. However, our analysis does underscore the exceptional nature of Bonds' 2001 season within the context of his career performance trends.

Part II

Introduction

The ISRUC-Sleep dataset is a resource for researchers interested in the field of sleep studies. This polysomnographic (PSG) dataset was compiled to facilitate a wide range of investigations into sleep patterns, disorders, and the effects of medication on sleep. Comprising data from 100 human adults, including both healthy subjects and those diagnosed with sleep disorders, the dataset is structured to support various research objectives and methodologies.

Dataset Overview

The dataset is organized into three main groups, each tailored to address different research needs:

1. **General Population Data**: It includes one recording session per subject across 100 subjects, offering a broad overview of sleep characteristics in a diverse population.The PSG recordings, associated with each subject, were visually scored by two human experts. The PSG recordings include electrophysiological signals, pneumological signals, and another contextual information of the subjects.
2. **Healthy Subjects Data**: This subset focuses on 10 healthy individuals, allowing for detailed comparisons between healthy sleep patterns and those affected by sleep disorders.
3. **Sleep Stages and Events**: The data contains epoch-by-epoch annotations of sleep stages (Awake, NREM stages N1, N2, N3, and REM) based on the American Academy of Sleep Medicine (AASM) criteria, as well as various sleep-related events and physiological signals (e.g., heart rate, blood-oxygen saturation).

Motivations for Researching Sleep

Researching sleep is crucial for several reasons, reflecting the central role of sleep in human health and well-being:

- **Understanding Sleep Disorders**: With a wide prevalence of sleep disorders across the global population, understanding the nuances of these conditions is vital for developing effective treatments. The ISRUC-Sleep dataset allows researchers to study the specific patterns and anomalies associated with different disorders.
- **Impact on Health**: Sleep has profound effects on physical health, mental health, and cognitive function. Research can uncover how variations in sleep patterns affect these areas, leading to improved guidelines for healthy sleep and interventions for sleep-related health issues.

In summary, the ISRUC-Sleep dataset is a foundational tool for advancing our understanding of sleep and its complex interplay with human health. It enables researchers to explore questions related to sleep physiology, disorders, treatment effects, and beyond, with the ultimate goal of enhancing sleep quality and health outcomes for individuals around the world.

```
summary_data <- read.csv("summary_data.csv")
sleep_stage_output <- read.csv("sleep_stage_output.csv")

dataset <- c(summary_data, sleep_stage_output)
dataset <- data.frame(dataset)
```

```r
dataset <- dataset %>%
  filter(Age > 0, W < 400) %>%
  filter(Sex %in% c("Male", "Female"))
```

```r
summary(dataset)
```

```
##        ID             Expert      Date              Height_in      Height_cm
##  Min.   :  1.00   Min.   :1   Length:73          Min.   :40.0   Min.   :101.0
##  1st Qu.: 28.00   1st Qu.:1   Class :character   1st Qu.:63.0   1st Qu.:160.0
##  Median : 52.00   Median :1   Mode  :character   Median :66.0   Median :167.0
##  Mean   : 51.45   Mean   :1                      Mean   :65.4   Mean   :166.1
##  3rd Qu.: 77.00   3rd Qu.:1                      3rd Qu.:69.0   3rd Qu.:174.0
##  Max.   :100.00   Max.   :1                      Max.   :77.0   Max.   :195.0
##
##    Weight_lbs      Weight_kg           Age            Sex
##  Min.   :  0.0   Min.   : 53.00   Min.   :20.00   Length:73
##  1st Qu.:161.0   1st Qu.: 73.00   1st Qu.:37.00   Class :character
##  Median :176.0   Median : 80.00   Median :50.00   Mode  :character
##  Mean   :175.8   Mean   : 80.90   Mean   :49.23
##  3rd Qu.:196.0   3rd Qu.: 89.25   3rd Qu.:61.00
##  Max.   :254.0   Max.   :115.00   Max.   :85.00
##                  NA's   :1
##   Total_Epoch        Min_SpO2         Min_HR         Max_SpO2
##  Min.   : 341.0   Min.   : 8.00   Min.   :24.00   Min.   : 94.00
##  1st Qu.: 809.0   1st Qu.:59.00   1st Qu.:47.00   1st Qu.: 97.00
##  Median : 872.0   Median :84.00   Median :57.00   Median : 98.00
##  Mean   : 848.2   Mean   :69.23   Mean   :54.21   Mean   : 97.92
##  3rd Qu.: 897.0   3rd Qu.:88.00   3rd Qu.:62.00   3rd Qu.: 99.00
##  Max.   :1032.0   Max.   :97.00   Max.   :74.00   Max.   :100.00
##
##      Max_HR       Average_SpO2     Average_HR    BPOS_Summary
##  Min.   : 76.0   Min.   :86.76   Min.   :47.38   Length:73
##  1st Qu.: 97.0   1st Qu.:93.28   1st Qu.:61.13   Class :character
##  Median :107.0   Median :94.84   Median :67.86   Mode  :character
##  Mean   :126.9   Mean   :94.43   Mean   :68.82
##  3rd Qu.:125.0   3rd Qu.:95.86   3rd Qu.:76.29
##  Max.   :255.0   Max.   :99.30   Max.   :94.03
##
##  Stage_Summary      Events_Summary       Row.Index           W
##  Length:73          Length:73          Min.   :  0.0   Min.   :  5.0
##  Class :character   Class :character   1st Qu.: 54.0   1st Qu.: 95.0
##  Mode  :character   Mode  :character   Median :102.0   Median :149.0
##                                        Mean   :100.9   Mean   :166.9
##                                        3rd Qu.:152.0   3rd Qu.:233.0
##                                        Max.   :198.0   Max.   :382.0
##
##        N3              N2              R              N1            n2
##  Min.   : 42.0   Min.   :139.0   Min.   :  8   Min.   : 11   Min.   : NA
##  1st Qu.:134.0   1st Qu.:208.0   1st Qu.: 84   1st Qu.: 74   1st Qu.: NA
##  Median :171.0   Median :278.0   Median :121   Median :102   Median : NA
##  Mean   :177.4   Mean   :276.4   Mean   :120   Mean   :107   Mean   :NaN
##  3rd Qu.:205.0   3rd Qu.:338.0   3rd Qu.:139   3rd Qu.:137   3rd Qu.: NA
##  Max.   :417.0   Max.   :432.0   Max.   :330   Max.   :245   Max.   : NA
```

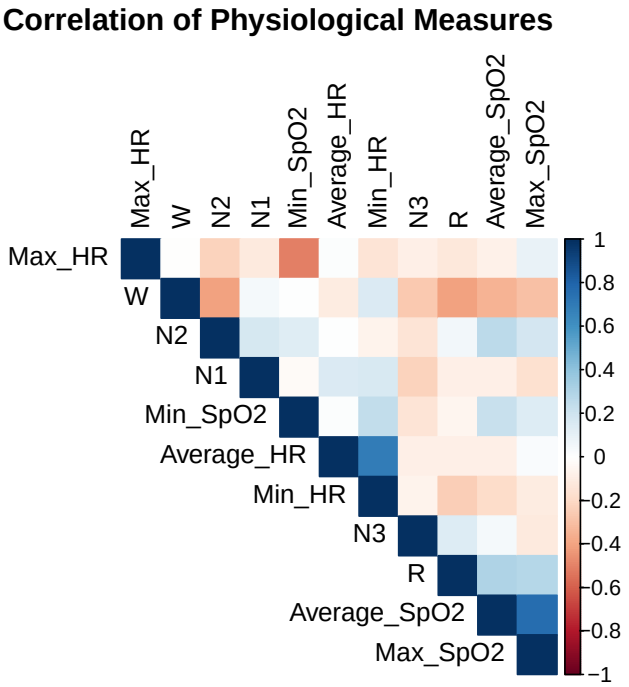
```
##                                                                 NA's   :73
##        U                N
##  Min.   :1.000   Min.   :2
##  1st Qu.:1.000   1st Qu.:2
##  Median :5.000   Median :2
##  Mean   :4.625   Mean   :2
##  3rd Qu.:7.250   3rd Qu.:2
##  Max.   :9.000   Max.   :2
##  NA's   :65      NA's   :72
```

The summary stats show that there are still some incomplete rows in our dataset (for example, there are still some people without a recorded weight), however the data that we are focusing on for the analysis appears to be complete now. These values are Age which goes from 20 to 85 with a median of 50 and a mean of 49.23 as well as Wake Time which is recorded as number of 30 second epochs. The values for Wake Time go from 5 to 382 (2.5 minutes to 191 minutes) with a median of 149 (74.5 minutes) and a mean of 166.9 (83.45 minutes).

For a more comprehensive exploration of the ISRUC-Sleep dataset, let's delve into additional visualizations that can illuminate various aspects of the data. These visualizations will help uncover patterns, trends, and correlations within the dataset, providing deeper insights into sleep behavior, physiological measures, and their interactions with demographic factors like age.

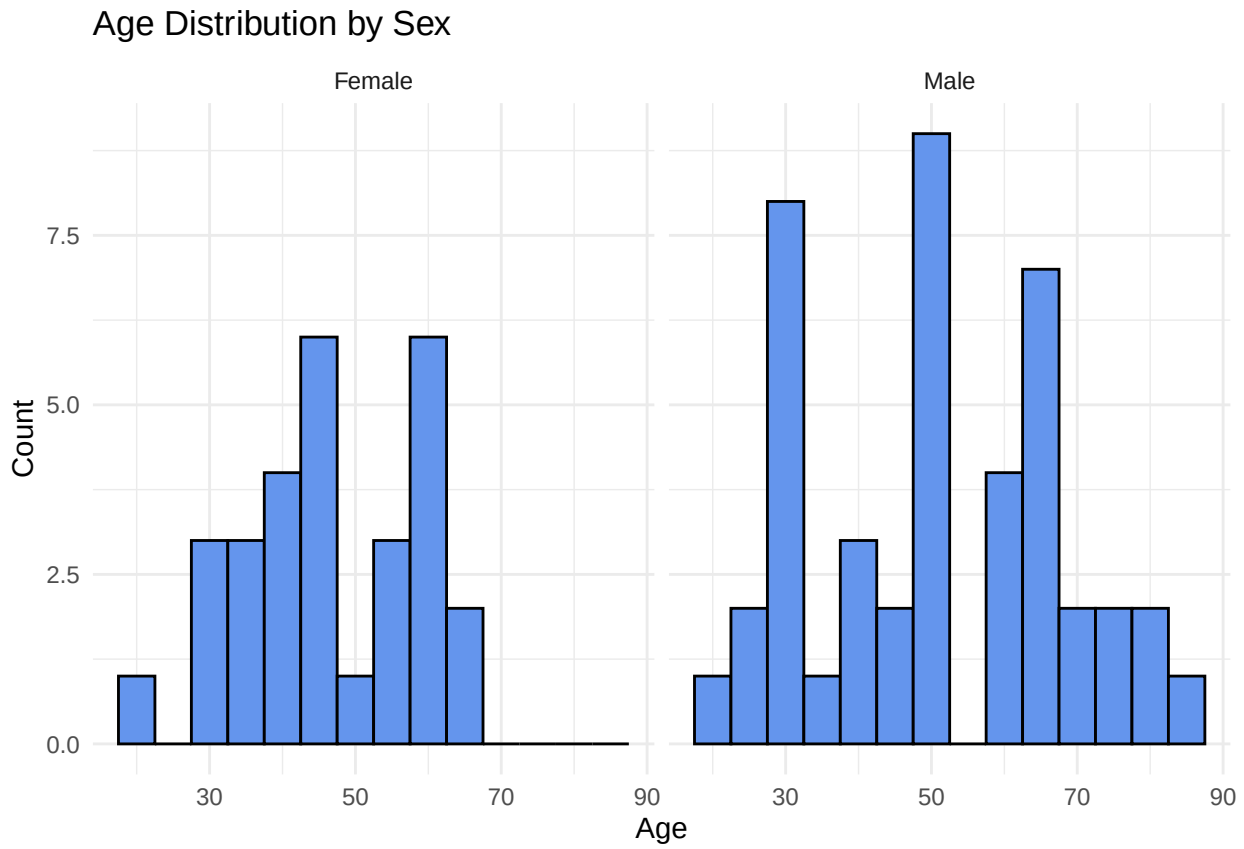
```
physiological_data <- dataset[, c("Average_SpO2", "Average_HR", "Min_SpO2",
                                  "Min_HR", "Max_SpO2", "Max_HR", "W", "N3",
                                  "N2","R","N1")]
cor_matrix <- cor(physiological_data, use = "complete.obs")

corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
         tl.col = "black", mar=c(0,0,2,0),
         title = "Correlation of Physiological Measures")
```

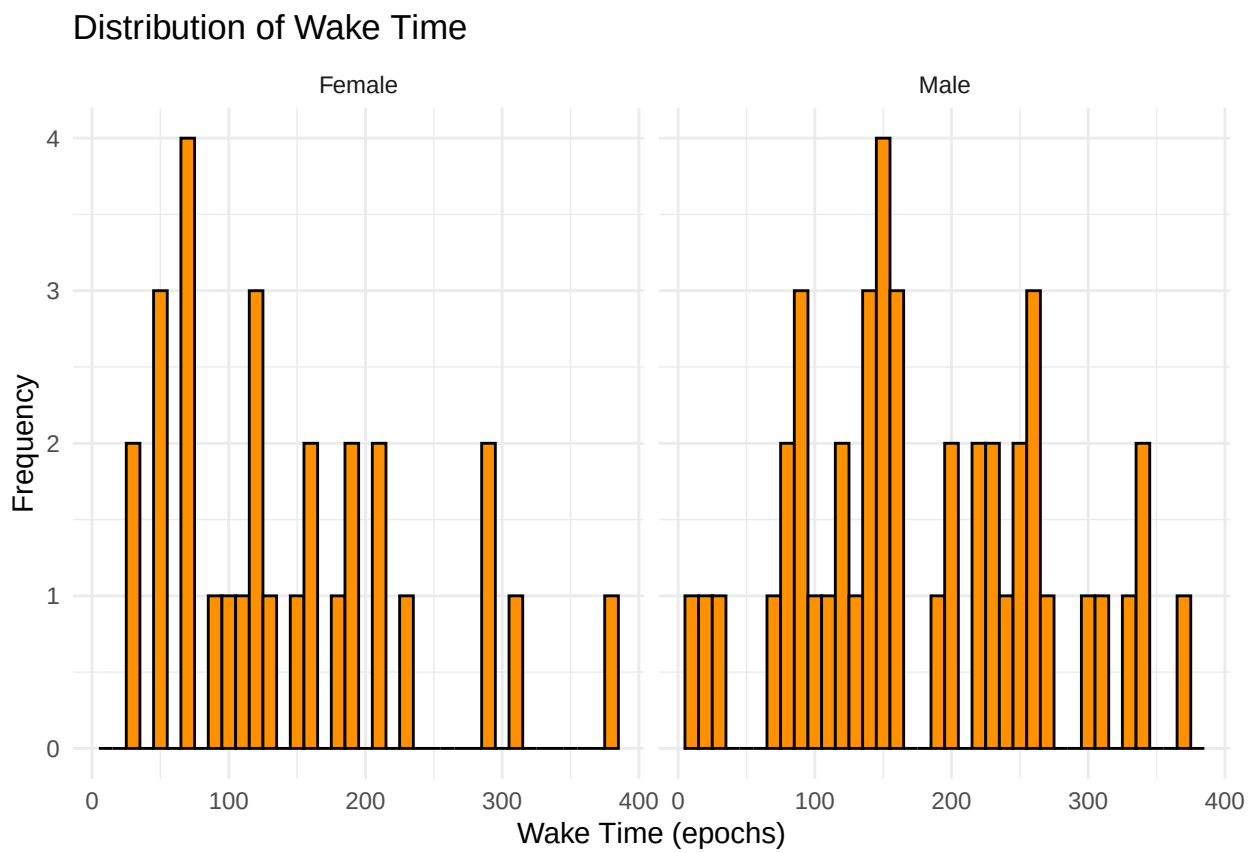


A correlation matrix visualization can help identify relationships between continuous variables such as age, wake time, average heart rate (HR), and average blood-oxygen saturation (SpO2). We can see that there are song strong correlations based on the minimums of various physiological factors.

```
ggplot(dataset, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "cornflowerblue", color = "black") +
  facet_wrap(~Sex) +
  labs(title = "Age Distribution by Sex", x = "Age", y = "Count") +
  theme_minimal()
```

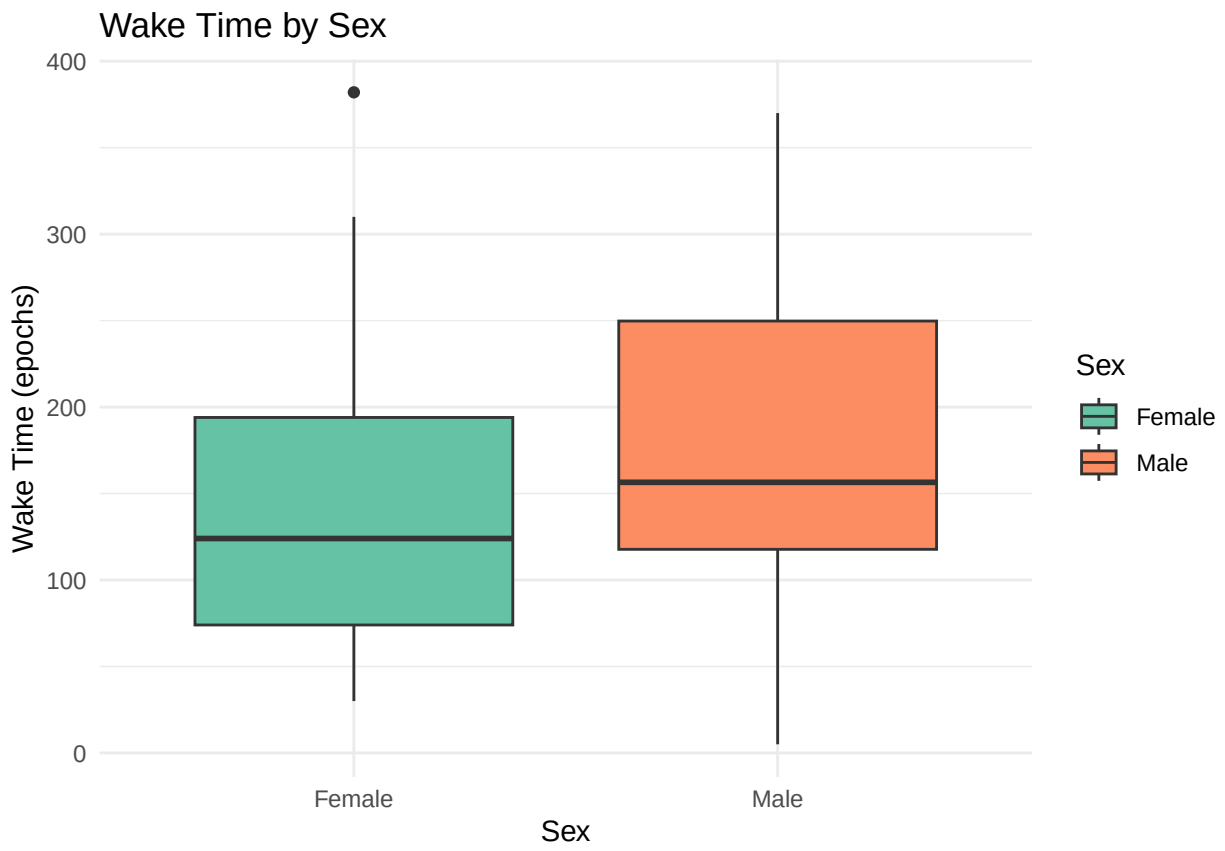



```
ggplot(dataset, aes(x = W)) +
  geom_histogram(binwidth = 10, fill = "#ff9100", color = "black") + facet_wrap(~Sex) +
  labs(title = "Distribution of Wake Time", x = "Wake Time (epochs)",
       y = "Frequency") +
  theme_minimal()
```



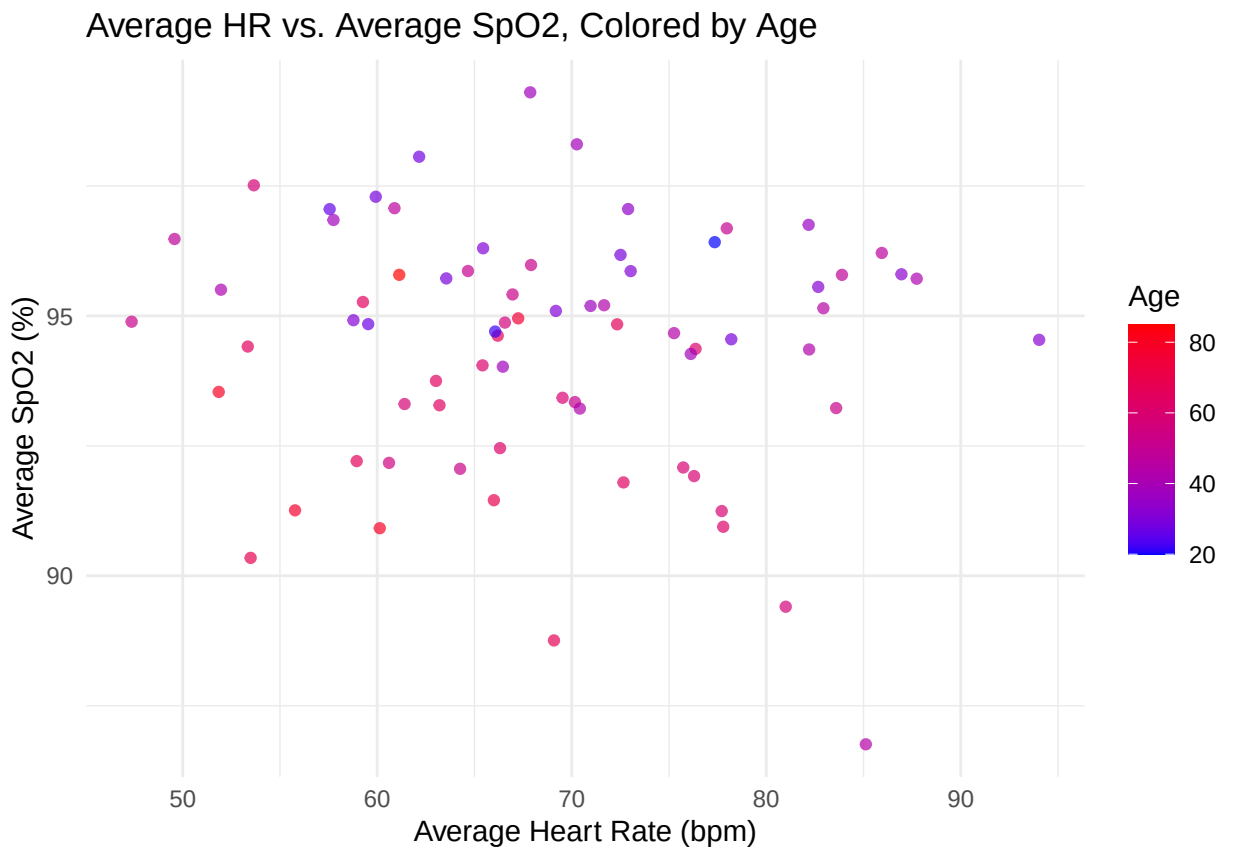
These plots show the distribution of ages and wake times within the dataset, separated by sex, concluding that there are more older males than there are women.

```
ggplot(dataset, aes(x = Sex, y = W, fill = Sex)) +
  geom_boxplot() +
  labs(title = "Wake Time by Sex", x = "Sex", y = "Wake Time (epochs)") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```



A boxplot can illustrate differences in wake time across sexes, highlighting any potential gender differences in sleep patterns, such that on average males have a higher wake time than women.

```
ggplot(dataset, aes(x = Average_HR, y = Average_SpO2, color = Age)) +
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Average HR vs. Average SpO2, Colored by Age",
       x = "Average Heart Rate (bpm)", y = "Average SpO2 (%)") +
  theme_minimal()
```



This scatter plot can help visualize the relationship between average heart rate and average SpO2, with points colored by age to see if there's any age-related pattern, which as we can see the color gets slightly brighter the lower we get on both average hr and average spO2.

Model Building

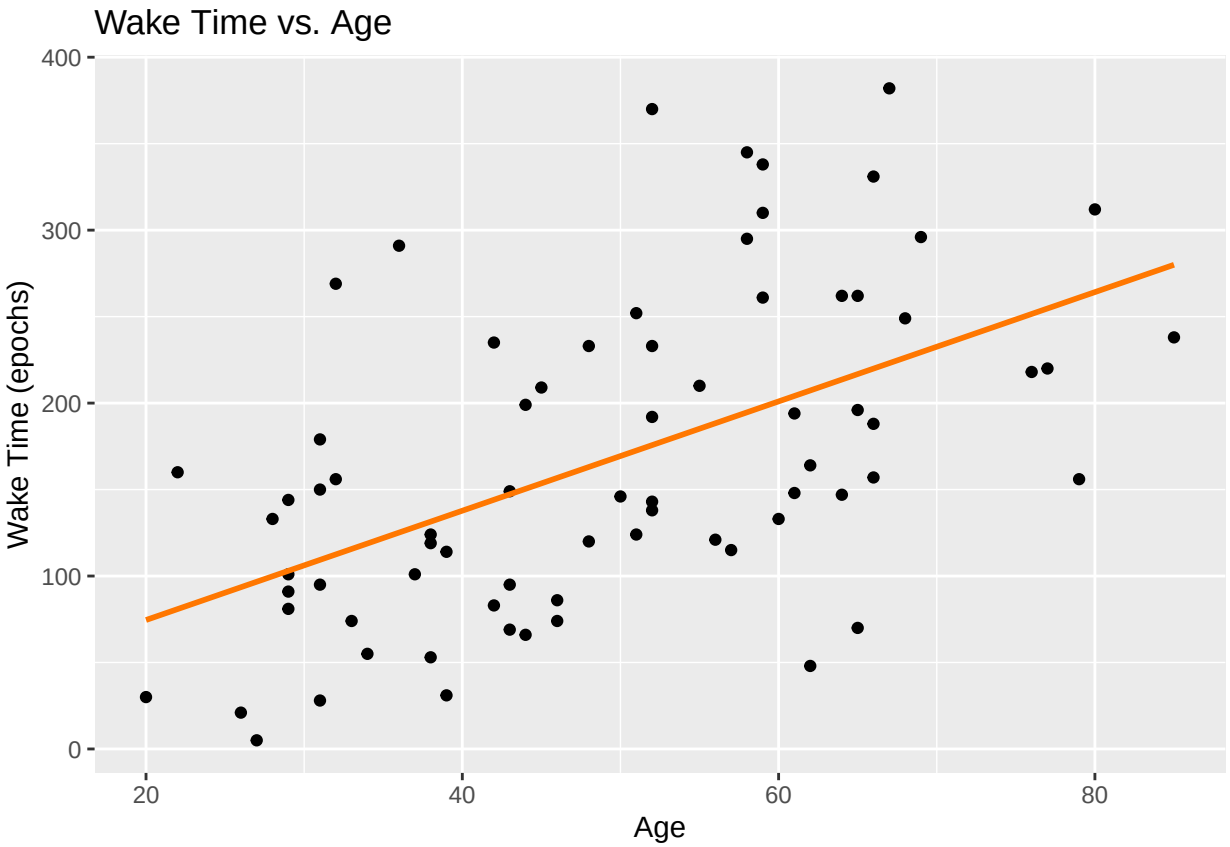
The statistical model we intend to develop, $\hat{W}_i = \alpha + \beta \times Age_i + \epsilon_i$, where $\hat{W}_i$ symbolizes the predicted wake times for the i^{th} individual, and Age_i represents the age of the individual, aims to shed light on the relationship between age and wake times. By quantifying the trend across different ages, this linear regression model acts as a foundational tool for investigating potential significant changes in wake times across the lifespan. Such changes, if detected, could indicate underlying physiological, psychological, or lifestyle shifts associated with aging. This model's ability to highlight deviations from expected patterns provides a quantitative basis for further exploration into the factors influencing sleep patterns and the potential impact of age on sleep quality and duration.

```
model <- lm(W ~ Age, data = dataset)
model$coef
```

```
## (Intercept)         Age
##    11.36195     3.16015
```

The linear regression equation representing the model is $\hat{W}_i = 11.36195 + 3.16015 * Age_i$

```r
ggplot(dataset, aes(x = Age, y = W)) +
  geom_point() +
  geom_smooth(method = "lm", se=FALSE,color = "#ff7700") +
  labs(title = "Wake Time vs. Age", x = "Age", y = "Wake Time (epochs)")
```



This scatter plot, complemented by a linear regression line, explores the relationship between wake time and age. Points represent individual subjects, plotting their wake time against their age. The linear regression line (smoothed line) provides a visual estimate of the trend, showing whether wake time tends to increases as age progresses.

Importantly, to ensure the integrity of the conclusions drawn from this model, we will verify the assumptions underlying linear regression analysis, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of error terms. These checks are fundamental to confirming that our model is accurately specified and that the inferences and predictions derived from it are reliable.

Model Assumption and Validation

Correlation Coefficient Check:

```r
cor(dataset$W, dataset$Age)
```

```
## [1] 0.5353568
```

A correlation coefficient of 0.5353568 indicates a positive linear relationship between the age and amount of time spent awake during sleep.

Significance of Coefficient Estimates

In the context of examining the importance of the coefficient estimate for age in predicting wake times, we can define the null hypothesis (H_0) and the alternative hypothesis (H_1) as follows, with a significance level (α) of 0.05:

Null hypothesis: $H_0 : \beta_1 = 0$, $H_0 : \beta_0 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$, $H_A : \beta_0 \neq 0$

```
summary(model)
```

```
##
## Call:
## lm(formula = W ~ Age, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -159.29  -61.09  -20.61   48.59  194.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3620    30.5293   0.372    0.711
## Age           3.1601     0.5917   5.341 1.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.04 on 71 degrees of freedom
## Multiple R-squared:  0.2866, Adjusted R-squared:  0.2766
## F-statistic: 28.52 on 1 and 71 DF,  p-value: 1.061e-06
```

```
coef(summary(model))[, "t value"]
```

```
## (Intercept)         Age
##   0.3721653   5.3408207
```

```
coef(summary(model))[, "Pr(>|t|)"]
```

```
##  (Intercept)          Age
## 7.108783e-01 1.060644e-06
```

Intercept

- **t-value for Intercept**: 0.3721653
- **p-value for Intercept**: 0.7108783

The intercept's t-value is not significantly far from zero, and the corresponding p-value is much greater than the alpha level of 0.05. This result suggests that the intercept is not statistically significantly different from zero at the conventional significance levels. Therefore, we fail to reject the null hypothesis $H_0 : \beta_0 = 0$ in favor of the alternative hypothesis $H_1 : \beta_0 \neq 0$.

Age

- **t-value for Age**: 5.3408207
- **p-value for Age**: 0.000001060644

The t-value for the age coefficient is significantly positive, indicating a positive relationship between age and wake times in the dataset analyzed. The p-value associated with this t-value is much less than 0.05, strongly suggesting that we reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative hypothesis $H_1 : \beta_1 \neq 0$.

Implications

The statistical analysis of the age coefficient reveals that there is a significant linear relationship between age and wake times. This means that age significantly predicts wake times, with the positive t-value indicating that as age increases, so does the wake time, suggesting an association between older age and longer wake times.

Given the alpha level of 0.05 and the very low p-value obtained for the age coefficient, our analysis provides strong statistical evidence to support the conclusion that age is a significant predictor of wake times. The intercept not being significantly different from zero suggests that the baseline wake time (at age zero, hypothetically speaking) is not statistically distinguishable from zero in this model's context.

The R-squared value of 0.2766 in our linear regression model, derived from real-world data, indicates that approximately 27.66% of the variance in the dependent variable (wake times) is explained by the independent variable (age). This figure, while lower than ideal theoretical models, is not uncommon in studies involving complex human behaviors and physiological responses, where multiple factors beyond age might influence the outcome. The value underscores the inherent complexity and variability in real-world datasets, highlighting that while age is a significant predictor of wake times, a substantial portion of the variance remains unexplained by this model alone.

Residual Analysis

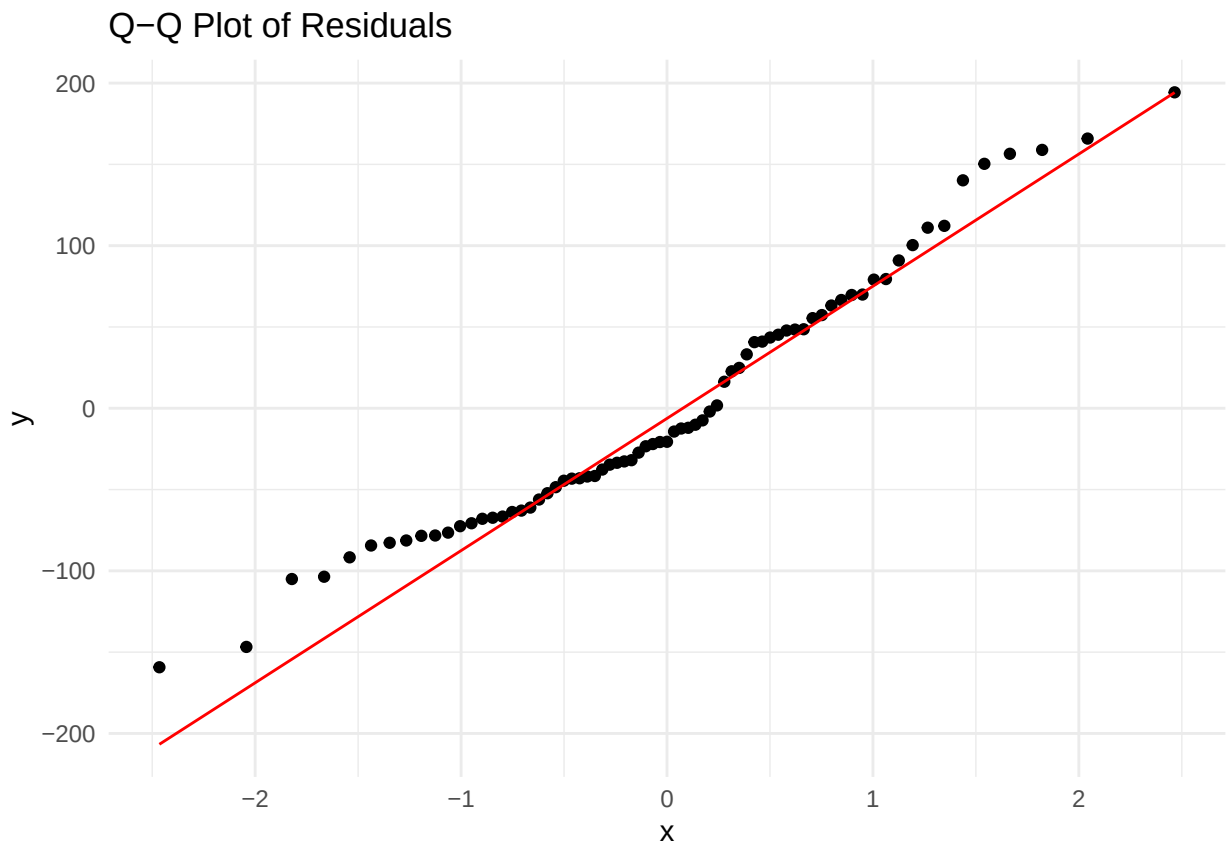
Normality of Residuals

In our specific analysis concerning wake time as a function of age, the normality of residuals implies that the deviations from the predicted wake times are random and follow a normal distribution. This condition supports the premise that our linear model is appropriately capturing the relationship between age and wake time without systematic bias.

The evidence supporting the normality of residuals in our analysis comes from both graphical and statistical methods:

```
residuals <- resid(model)
residuals_df <- data.frame(Residuals = residuals)

# Generate a Q-Q plot
ggplot(residuals_df, aes(sample = Residuals)) +
  stat_qq() +
  stat_qq_line(colour = "red") +
  ggtitle("Q-Q Plot of Residuals") +
  theme_minimal()
```



In our case, the residuals deviate slightly with the reference line in the Q-Q plot, indicating that their distribution might not be a normal distribution. This evidence suggests that the residuals from our linear model do not deviate significantly from normality but to be certain we can perform a Shapiro Wilks test with an alpha of 0.01.

```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.96323, p-value = 0.03179
```

With a p-value of 0.03179 from the Shapiro-Wilk test and an alpha level of 0.01, we do not reject the null hypothesis that the data is normally distributed. The p-value is greater than the alpha level, indicating that there is not enough evidence to conclude the data deviates from normality at the 1% significance level. This result supports the assumption of normality for the residuals in our analysis, suggesting that our statistical model's underlying assumptions are met, enhancing the reliability of the conclusions drawn from it in the context of real-world data variability.

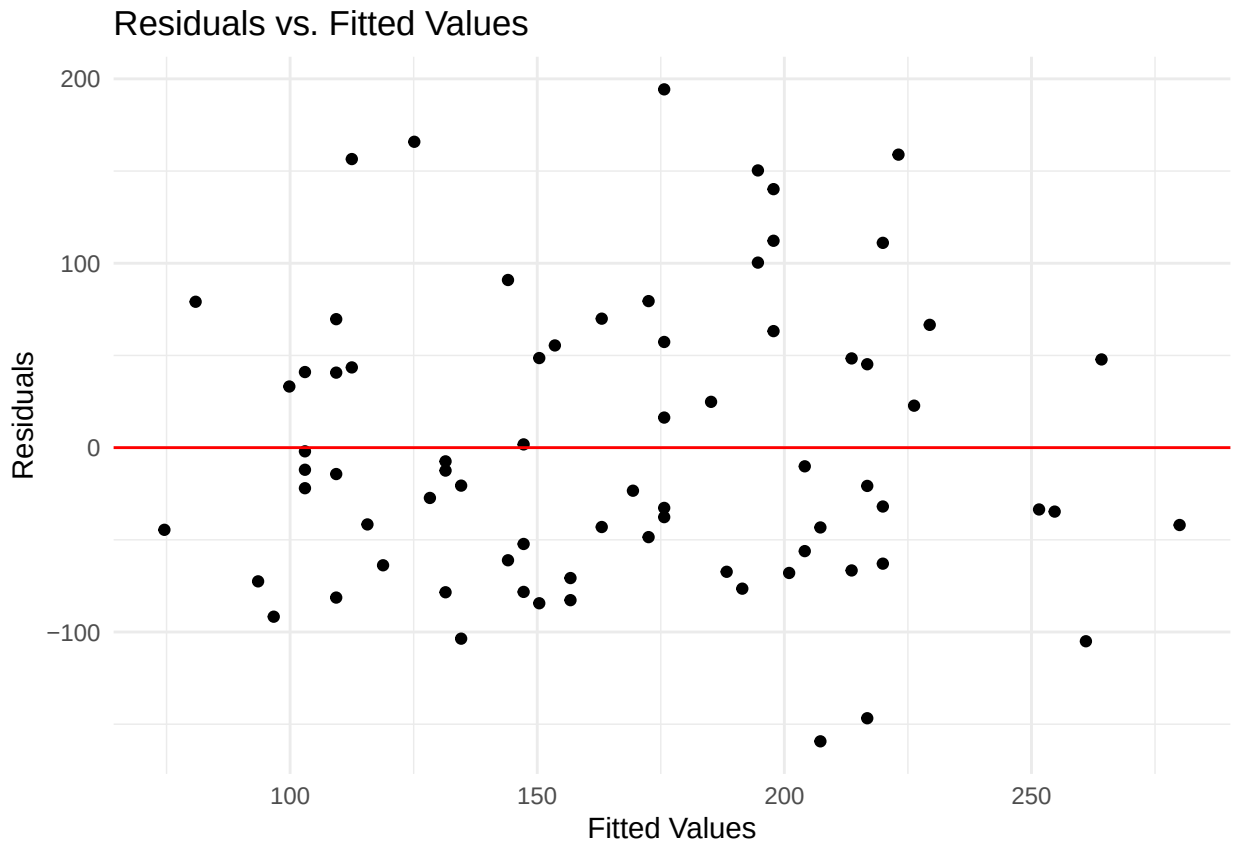
Homoscedasticity

The examination of homoscedasticity is a critical step in validating the assumptions underlying a linear regression model. Homoscedasticity refers to the condition where the residuals (the differences between observed and predicted values) have constant variance across all levels of the independent variable(s). This

assumption ensures that the model's predictive accuracy is uniform across the range of the independent variable.

```
fitted_values <- fitted(model)
homoscedasticity_df <- data.frame(Fitted = fitted_values, Residuals = residuals)

ggplot(homoscedasticity_df, aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, colour = "red") +
  ggtitle("Residuals vs. Fitted Values") +
  xlab("Fitted Values") +
  ylab("Residuals") +
  theme_minimal()
```




```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.83438, df = 1, p-value = 0.361
```

The Breusch-Pagan (BP) test is a statistical procedure designed to test for heteroscedasticity — the presence of non-constant variance in the error terms of a regression model. The BP test's p-value of 0.361 significantly

exceeds the common alpha level threshold of 0.05. This high p-value indicates that there is insufficient evidence to reject the null hypothesis of the BP test, which states that the error variances are homoscedastic.

The even distribution of residuals across the predicted values, as evidenced by the scatter plot, allows us to conclude that the condition of homoscedasticity holds for our linear regression model. This finding is crucial because it means that our model meets another important assumption of linear regression, reinforcing its validity and the reliability of its predictions and inferences.

Given that our analysis has confirmed both key assumptions of linear regression—normality of residuals and homoscedasticity—we can assert that our linear regression model is valid. This validity implies that the model is well-founded and that the statistical inferences drawn from it, such as confidence intervals and hypothesis tests on the regression coefficients, are based on solid assumptions.

References

Khalighi Sirvan, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. "ISRUC-Sleep: A comprehensive public dataset for sleep researchers."Computer methods and programs in biomedicine 124 (2016): 180-192.