Qno_08.R

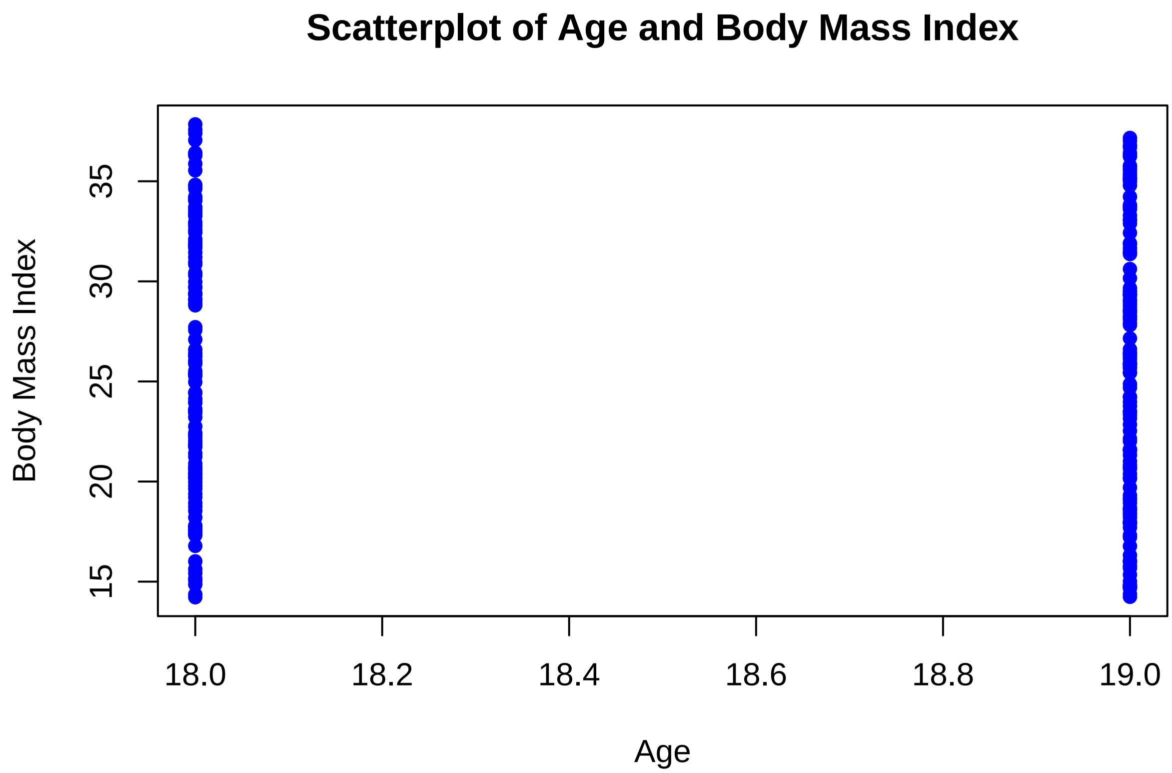
arpan

2023-07-27

```r
# a
set.seed(7916007)

age <- sample(18:19, 250, replace = TRUE)
sex <- sample(c("male", "female"), 250, replace = TRUE)
education <- sample(c("Noeducation", "primary", "secondary", "beyond secondary"), 250, replace = TRUE)
socioeconomic_status <- sample(c("Low", "Middle", "High"), 250, replace = TRUE)
body_mass_index <- runif(250, 14, 38)

dataset <- data.frame(age, sex, education, socioeconomic_status, body_mass_index)

head(dataset)
```

```
##   age    sex        education socioeconomic_status body_mass_index
## 1  19   male          primary                  Low        31.50551
## 2  19 female        secondary               Middle        37.01350
## 3  19 female      Noeducation               Middle        28.28310
## 4  18 female        secondary               Middle        31.44274
## 5  18 female      Noeducation                 High        19.80437
## 6  19 female beyond secondary               Middle        25.84661
```

```r
# b
plot(dataset$age, dataset$body_mass_index, main = "Scatterplot of Age and Body Mass Index",
     xlab = "Age", ylab = "Body Mass Index", pch = 16, col = "blue")
```




```r
# Interpretation:
# The scatterplot shows that there is no any strong relationship between the variables.


# c
# Since we have plotted age and body mass index, we have Pearson correlation coefficient.
# and Spearman Correlation coefficient. Pearson correlation cofficient is used when the
# data is linear where as Spearman is used when the data is not linear. In the
# Scatter plot we got the two only two age 18 and 19 where the BMI data is.
# Since there are only two distinct age values in the dataset (18 and 19),
# the scatter plot will show only two points, and it will not provide much insight
# into the relationship between age and BMI.

# d
correlation_coefficient <- cor(dataset$age, dataset$body_mass_index, method = "pearson")
cat("Correlation Coefficient (Pearson):", correlation_coefficient, "\n")
```

```
## Correlation Coefficient (Pearson): -0.02288328
```

The Pearson correlation coefficient of approximately -0.0229 suggests that there is a very weak and almost negligible linear relationship between the "age" and "body mass index" (BMI) variables in the dataset. The negative sign indicates a small inverse relationship, but the closeness to zero indicates that the correlation is minimal. In practical terms, age and BMI do not appear to have a significant linear association with each other in this dataset.

```r
# e
# we can perform a hypothesis test.
# A common test is to use the t-test for correlation coefficient,
# assuming the null hypothesis that the true correlation is 0.


cor_test_result <- cor.test(dataset$age, dataset$body_mass_index, method = "pearson")

# p-value of the test
p_value <- cor_test_result$p.value
cat("p-value:", p_value, "\n")
```

```
## p-value: 0.7188095
```

```r
if (p_value < 0.05) {
  cat("The correlation coefficient is statistically significant at the 5% level.\n")
} else {
  cat("The correlation coefficient is not statistically significant at the 5% level.\n")
}
```

```
## The correlation coefficient is not statistically significant at the 5% level.
```

Interpretation: The p-value associated with the Pearson correlation coefficient is 0.7188. In hypothesis testing for correlation, the null hypothesis states that there is no significant linear relationship between the two variables (age and body mass index - BMI). The alternative hypothesis suggests that there is a significant linear relationship between the variables.#'