Homework 2 Report

Name

Parul Gupta

Setup

Read in the data with `read_csv()` and store the data as an R object named `dataset`. Check the data to make sure all of the expected observations and variables are there.

```
dataset <- read.csv('maacs.csv.gz')
```

Part 1

We will first consider the relationship between FEV1 and age. In general, it is expected that as children get older (and hence, larger in size), their FEV1 values should get higher.

Consider the statement "FEV1 values in children are higher in older children relative to younger children".

Write a function in R that takes the `dataset` object as an argument and returns `TRUE` if the statement above is true for the dataset and `FALSE` otherwise.

NOTE: In order to write this function, you will need to translate the statement above into something that can be checked with the data. There are many ways in which you can do that translation correctly and you only need to pick one way here.

NOTE: For this part, do not use any plots.

```
check_relationship <- function(dataset) {
  # remove missing values just in case
  valid_data <- dataset[!is.na(dataset$age) & !is.na(dataset$fev1), ]
  corr_val <- cor(valid_data$age, valid_data$fev1)
  return(corr_val > 0)
}
```

Part 2

Fit a linear regression model with FEV1 as the outcome and age as a predictor.

How much does FEV1 change for a 1-year increase in the child's age?

```
# Fit linear regression model
model <- lm(fev1 ~ age, data = dataset)

# View the summary to see coefficients
summary(model)
```

```
##
## Call:
## lm(formula = fev1 ~ age, data = dataset)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.51099 -0.27100  0.00196  0.23616  2.15477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001199   0.134696  -0.009    0.993
## age          0.171162   0.010984  15.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4827 on 131 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6496, Adjusted R-squared:  0.6469
## F-statistic: 242.8 on 1 and 131 DF,  p-value: < 2.2e-16
```

```
# Extract the coefficient for age (the slope)
coef(model)["age"]
```

```
##       age
## 0.1711623
```

Write your data analysis statement interpreting the regression model here: For each additional year of age, FEV1 changes by 0.171 units indicating FEV1 levels rise over time.

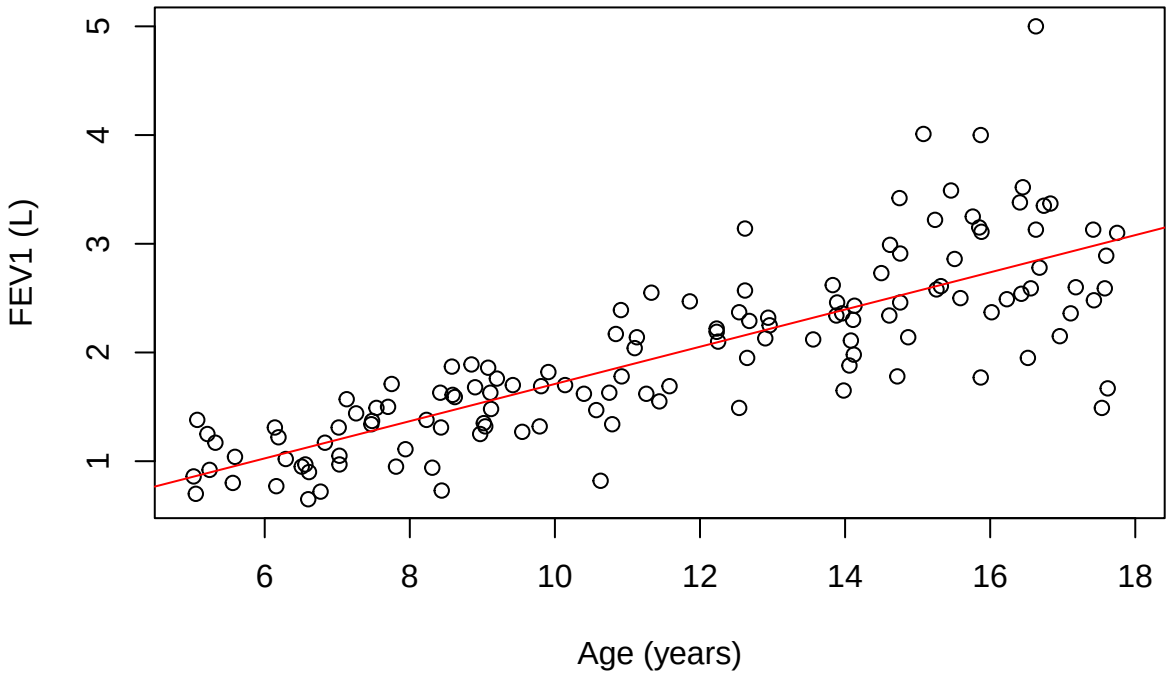
Part 3

Develop **three** supporting premises derived from the data that support the statement you wrote in Part 2. These can be plots, other summary statistics, or model results.

NOTE: At least one supporting premise should use a plot.

```
## Scatter Plot
plot(fev1 ~ age, data = dataset,
     main = "FEV1 vs. Age",
     xlab = "Age (years)",
     ylab = "FEV1 (L)")

abline(model, col = "red")
```




```
# View p value
summary(model)
```

```
##
## Call:
## lm(formula = fev1 ~ age, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51099 -0.27100  0.00196  0.23616  2.15477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001199   0.134696  -0.009    0.993
## age          0.171162   0.010984  15.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4827 on 131 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6496, Adjusted R-squared:  0.6469
## F-statistic: 242.8 on 1 and 131 DF,  p-value: < 2.2e-16
```

Write the three supporting premise statements here:

1. The plot visually indicates a positive trend, where as age increases, the FEV1 levels are increasing.
2. The p value is less than 2e-16, which is also less than 0.05, the typical benchmark for evaluating significance. This provides strong statistical evidence that the relationship between age and FEV1 is not due to random chance.
3. The R-squared value is 0.6496. This means that 64.96% of the variability in FEV1 can be explained by

age alone, demonstrating that age is a strong predictor of FEV1.

Part 4

For each of the supporting premises above, write a function that takes the `dataset` object as an argument and returns `TRUE` if the supporting premise statement above is true for the dataset and `FALSE` otherwise.

For statements involving plots, instead of returning `TRUE` or `FALSE`, you function should do two things:

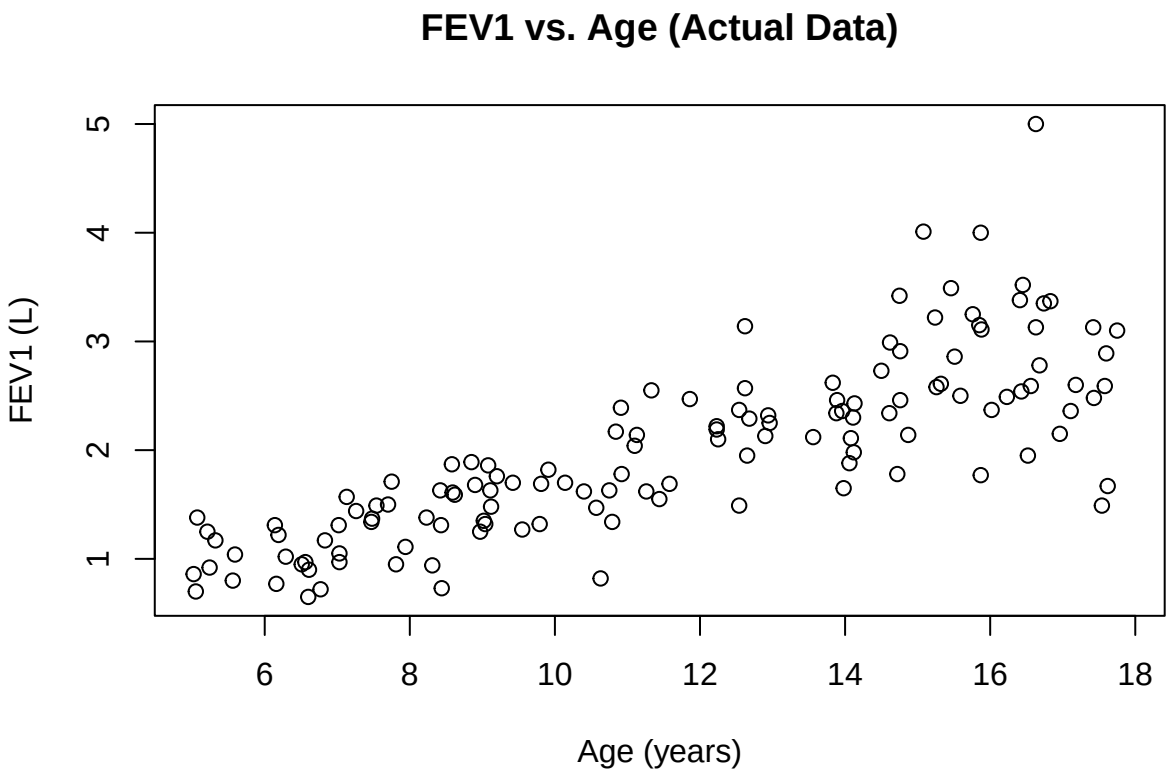
1. Produce the plot that is used in the statement
2. Produce a hypothetical version of the plot in the event that the statement is true. This can be done using simulated data or by simply hand drawing a plot.

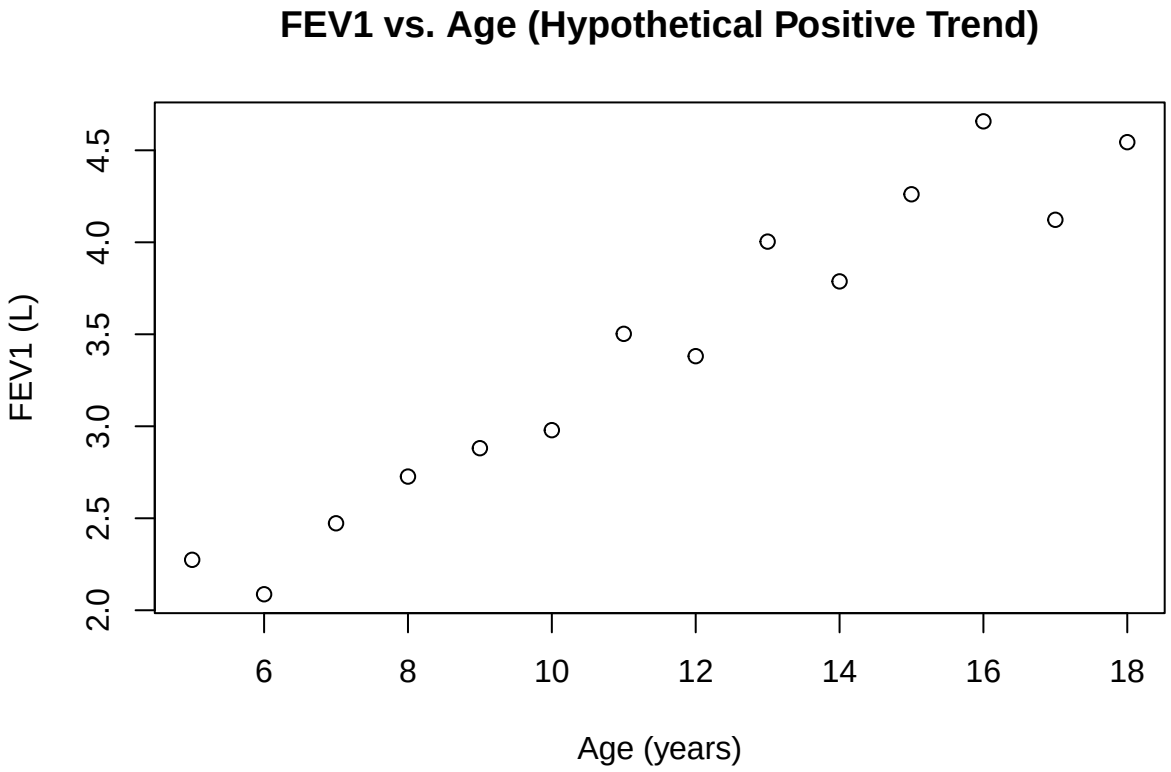
```
## Function for supporting premise statement 1
check_premise1 <- function(dataset) {
  plot(fev1 ~ age, data = dataset,
       main = "FEV1 vs. Age (Actual Data)",
       xlab = "Age (years)",
       ylab = "FEV1 (L)")

  set.seed(42) # for reproducibility
  hypothetical_age <- 5:18
  hypothetical_fev1 <- 1 + 0.2 * hypothetical_age + rnorm(length(hypothetical_age), mean = 0, sd = 0.2)

  plot(hypothetical_fev1 ~ hypothetical_age,
       main = "FEV1 vs. Age (Hypothetical Positive Trend)",
       xlab = "Age (years)",
       ylab = "FEV1 (L)")
}

check_premise1(dataset)
```






```
## Function for supporting premise statement 2
check_premise2 <- function(dataset) {
  model <- lm(fev1 ~ age, data = dataset)

  p_value <- summary(model)$coefficients["age", "Pr(>|t|)"]

  return(p_value < 0.05)
}

check_premise2(dataset)
```

```
## [1] TRUE
```

```
## Function for supporting premise statement 3
check_premise3 <- function(dataset) {
  model <- lm(fev1 ~ age, data = dataset)

  r_squared <- summary(model)$r.squared

  return(r_squared > 0.5)
}

check_premise3(dataset)
```

```
## [1] TRUE
```

Execute each of your function and show that the produce the expected output.

Part 5

Describe one alternative to the primary statement "FEV1 values in children are higher in older children relative to younger children".

There is no linear relationship between FEV1 and age; this means as a child gets older, their FEV1 values do not systematically increase or decrease, but instead remain constant or change in a non-linear, unpredictable way.

Create a fault tree for the alternative outcome describing how the alternative outcome could be realized in the data even if the primary statement were true.

Your fault tree should be created as a separate image and does not need to be created in R. Upload the image of the fault tree to Canvas.