FML_Assignment-1

Sharanya Domakonda

2024-02-04

1. Download a dataset from the web.

File location - https://www.kaggle.com/datasets/moslemcapo/male-vs-female

---

2. Import the dataset into R.

```
#Importing the dataset to R
test <- read.csv("C://Users//91988//Desktop//FML//dataset csv.csv")
```

```
#Summary of the dataset
summary(test)
```

```
##      Date               Gender          Driving.test.result      Bmi
##  Length:354         Length:354         Min.   : 1.000      Min.   :17.39
##  Class :character   Class :character   1st Qu.: 2.000      1st Qu.:26.32
##  Mode  :character   Mode  :character   Median : 6.000      Median :31.82
##                                        Mean   : 5.678      Mean   :30.94
##                                        3rd Qu.: 9.000      3rd Qu.:35.60
##                                        Max.   :10.000      Max.   :42.13
##     Children          Salary         region             smoker
##  Min.   :0.0000   Min.   : 1137   Length:354         Length:354
##  1st Qu.:0.0000   1st Qu.: 3580   Class :character   Class :character
##  Median :1.0000   Median :10602   Mode  :character   Mode  :character
##  Mean   :0.9492   Mean   :15390
##  3rd Qu.:2.0000   3rd Qu.:23568
##  Max.   :5.0000   Max.   :51195
##       age
##  Min.   :18.00
##  1st Qu.:23.00
##  Median :34.00
##  Mean   :37.19
##  3rd Qu.:55.00
##  Max.   :63.00
```

---

3. Print out descriptive statistics for a selection of quantitative and categorical variables

```r
# Descriptive statistics for a quantitative variable
summary(test$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   23.00   34.00   37.19   55.00   63.00
```

```r
# Descriptive statistics for a categorical variable
table(test$region)
```

```
##
## northeast northwest southeast southwest
##        84        84       102        84
```

---

4. Transform at least one variable. It doesn't matter what the transformation is.

```r
#Transforming a variable
test$smoker <- ifelse(test$smoker == "yes", "1", test$smoker)
test$smoker <- ifelse(test$smoker == "no", "2", test$smoker)
```

```r
#summary of the dataset after transformation
summary(test)
```

```
##      Date               Gender           Driving.test.result      Bmi
##  Length:354         Length:354         Min.   : 1.000      Min.   :17.39
##  Class :character   Class :character   1st Qu.: 2.000      1st Qu.:26.32
##  Mode  :character   Mode  :character   Median : 6.000      Median :31.82
##                                        Mean   : 5.678      Mean   :30.94
##                                        3rd Qu.: 9.000      3rd Qu.:35.60
##                                        Max.   :10.000      Max.   :42.13
##     Children          Salary         region             smoker
##  Min.   :0.0000   Min.   : 1137   Length:354         Length:354
##  1st Qu.:0.0000   1st Qu.: 3580   Class :character   Class :character
##  Median :1.0000   Median :10602   Mode  :character   Mode  :character
##  Mean   :0.9492   Mean   :15390
##  3rd Qu.:2.0000   3rd Qu.:23568
##  Max.   :5.0000   Max.   :51195
##       age
##  Min.   :18.00
##  1st Qu.:23.00
##  Median :34.00
##  Mean   :37.19
##  3rd Qu.:55.00
##  Max.   :63.00
```

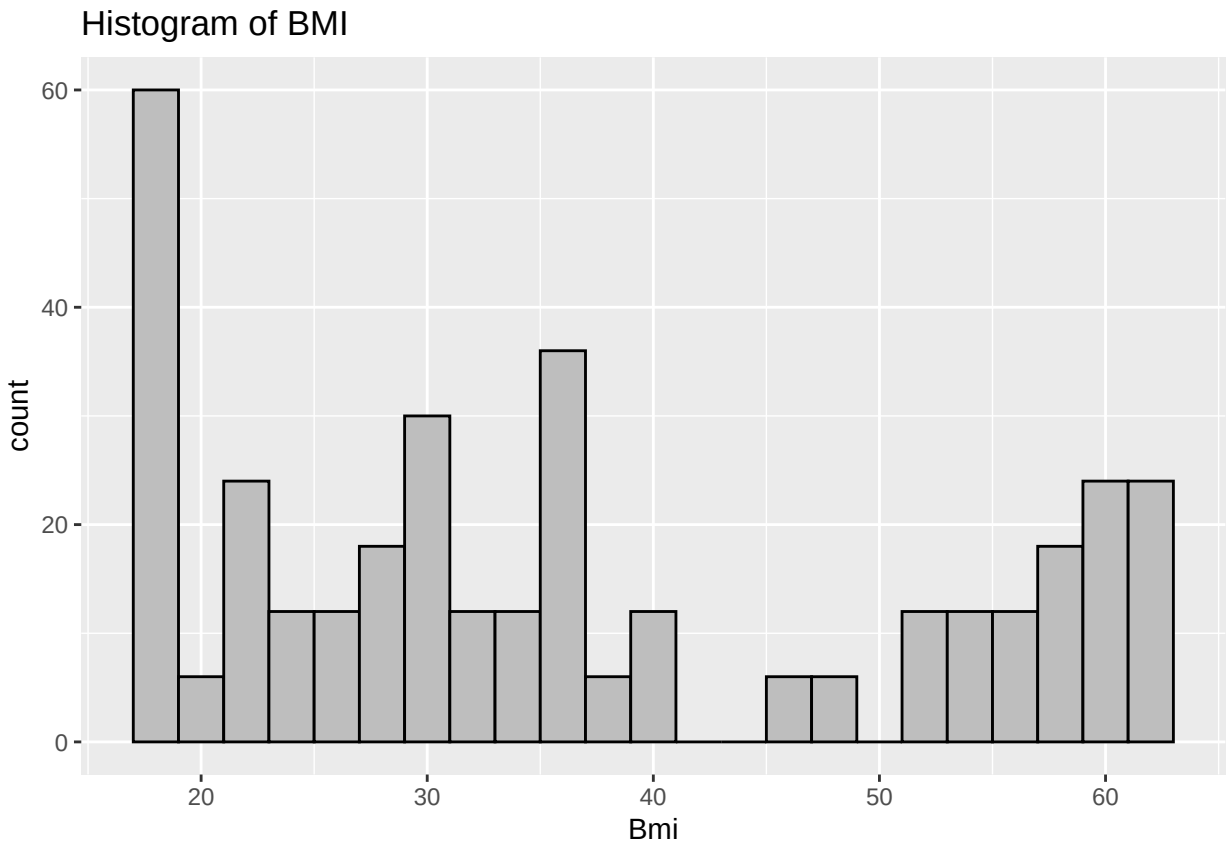
```r
head(test)
```

```
##         Date Gender Driving.test.result    Bmi Children    Salary    region
## 1 01-11-2022 female                   5 27.900        0 16884.924 southwest
## 2 01-11-2022 female                   4 33.770        1  1725.552 southeast
## 3 01-11-2022   male                   8 33.000        3  4449.462 southeast
## 4 01-11-2022   male                   9 22.705        0 21984.471 northwest
## 5 01-11-2022 female                   4 28.880        0  3866.855 northwest
## 6 02-11-2022 female                   2 25.740        0  3756.622 southeast
##   smoker age
## 1      1  19
## 2      2  18
## 3      2  28
## 4      2  33
## 5      2  32
## 6      2  31
```

---

5. Plot at least one quantitative variable, and one scatterplot.

```
#Plotting histogram for one quantitative variable(BMI)
library(ggplot2)
ggplot(test, aes(x = age)) + geom_histogram(binwidth = 2,
      fill = "grey", color = "black") +
   labs(title = "Histogram of BMI", x = "Bmi")
```



```r
#Plotting a scatterplot for two quantitative variables(Age and Gender)
ggplot(test, aes(x = age, y = Gender)) + geom_point(color = "Purple") +
labs(title = "Scatter Plot of Age vs. Gender", x = "Age", y = "Gender")
```