Practice-lab-3.R

acer

2023-10-04

```
# Load necessary libraries (if not already loaded)
library(NHANES)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

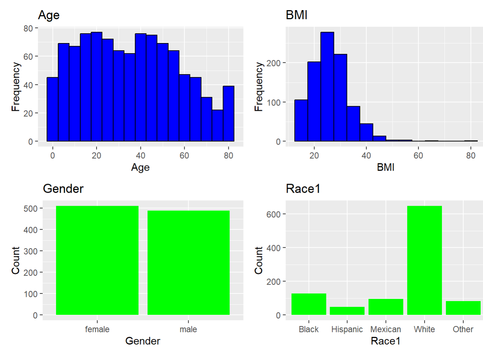
```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Import the NHANES dataset
data("NHANES")

# Choose two continuous and two categorical variables
continuous_vars <- c("Age", "BMI")
categorical_vars <- c("Gender", "Race1")

# Create a subset of 1000 individuals
nhanes_subset <- NHANES %>% sample_n(1000)

# Create a separate dataframe with ID, age, and chosen variables
subset_df <- nhanes_subset %>%
  select(ID, Age, all_of(continuous_vars), all_of(categorical_vars))

# Depict continuous variables as a histogram
histograms <- lapply(continuous_vars, function(var) {
  ggplot(subset_df, aes(x = !!as.name(var))) +
    geom_histogram(binwidth = 5, fill = "blue", color = "black") +
    labs(title = var, x = var, y = "Frequency")
})

# Depict categorical variables as a barplot
barplots <- lapply(categorical_vars, function(var) {
  ggplot(subset_df, aes(x = !!as.name(var))) +
    geom_bar(fill = "green") +
    labs(title = var, x = var, y = "Count")
})

# Show histograms and barplots
gridExtra::grid.arrange(grobs = c(histograms, barplots), ncol = 2)
```

```
## Warning: Removed 38 rows containing non-finite values (`stat_bin()`).
```




```
# Show all data in Table 1
table1 <- summary(subset_df)
print(table1)
```

```
##        ID              Age              BMI           Gender         Race1
##  Min.   :51624   Min.   : 0.00   Min.   :12.90   female:511   Black   :127
##  1st Qu.:56906   1st Qu.:17.00   1st Qu.:21.13   male  :489   Hispanic: 47
##  Median :61697   Median :35.00   Median :25.60                Mexican : 96
##  Mean   :61677   Mean   :35.68   Mean   :26.26                White   :648
##  3rd Qu.:66629   3rd Qu.:52.00   3rd Qu.:30.30                Other   : 82
##  Max.   :71909   Max.   :80.00   Max.   :81.25
##                                  NA's   :38
```

```
# Group into age groups in 10-year increments and aggregate variables
age_groups <- nhanes_subset %>%
  mutate(Age_Group = cut(Age, breaks = seq(0, max(Age), by = 10))) %>%
  group_by(Age_Group) %>%
  summarise_at(vars(continuous_vars), mean, na.rm = TRUE) %>%
  ungroup()
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## ℹ Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(continuous_vars)
##
##   # Now:
##   data %>% select(all_of(continuous_vars))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Calculate proportions of categorical variables within each age group
age_groups <- age_groups %>%
  group_by(Age_Group) %>%
  ungroup()

# Print the age_groups dataframe
print(age_groups)
```

```
## # A tibble: 9 × 3
##   Age_Group   Age   BMI
##   <fct>     <dbl> <dbl>
## 1 (0,10]      5.4  17.3
## 2 (10,20]    15.8  24.0
## 3 (20,30]    25.4  28.2
## 4 (30,40]    35.6  29.1
## 5 (40,50]    45.4  27.8
## 6 (50,60]    55.5  29.3
## 7 (60,70]    65    28.7
## 8 (70,80]    77.0  27.2
## 9 <NA>        0   NaN
```

```
# Show the central tendency and variability of continuous variables as a box plot
boxplot_plot <- subset_df %>%
  mutate(Age_Group = cut(Age, breaks = seq(0, max(Age), by = 10))) %>%
  ggplot(aes(x = Age_Group, y = Age, fill = Age_Group)) +
  geom_boxplot() +
  labs(title = "Box Plot of Age by Age Groups", x = "Age Group", y = "Age")

print(boxplot_plot)
```