Report

Abigail Joseph

Homework 4

CAP2751 - Tools for Data Science

Introduction

The provided R code performs data transformation, EDA, basic statistical analysis, and visualization on a dataset named `caffeine,` which is loaded from a link. The code includes steps for data summarization, visualization, and statistical testing to understand the relationship between variables such as drink type, amount of drink, age, gender, and test scores. Below is a thorough report based on the given code.

Data Loading and Preparation

```
# loading the dataset
url <-
"https://raw.githubusercontent.com/ndphillips/ThePiratesGuideToR/master/data/caffeinestudy.txt"

# loading as 'caffeine' dataframe
caffeine <- read.table(url, header = TRUE, sep = "\t")
```

- The dataset is loaded from a URL using `read.table`.
- The `caffeine` dataframe is created, which contains variables for subsequent analysis.

Descriptive Statistics

 # 2. Write R code to calculate and print:

```
# installing and loading dplyr
install.packages("dplyr")
library(dplyr)

# a. the mean age for each gender.

# pipe operator for readability
caffeine %>%

# grouping based on gender
group_by(gender) %>%

# result
summarise(mean_age = mean(age, na.rm = TRUE))

# b. the mean age for each drink.
caffeine %>%
group_by(drink) %>%
summarise(mean_age = mean(age, na.rm = TRUE))

# c. the mean age for each combined level of both gender and drink
caffeine %>%
group_by(gender, drink) %>%
summarise(mean_age = mean(age, na.rm = TRUE))

# d. the median score for each age
caffeine %>%
```

```
group_by(age) %>%
summarise(median_score = median(score, na.rm = TRUE))
```

- The code calculates the mean age for each gender, drink type, and each combined level of gender and drink, providing insights into the demographics in the dataset.
- The median score for each age is computed, which could help draw several conclusions, such as whether scores tend to increase or decrease with age.

Gender-Specific Analysis

 # 3. For men only, write R code to calculate and print the maximum score for
 # each age.

 # assigning a new variable
max_score_for_men <- caffeine %>%

 # filtering male gender
filter(gender == "male") %>%

 # grouping by age, retrieving the maximum score for each age for men only
group_by(age) %>%

 # result
summarise(max_score = max(score, na.rm = TRUE))

 # printing the result
print(max_score_for_men)

- Maximum test scores for males are calculated for each age, which could reveal whether performance peaks at certain ages for males.

Drink-Level Analysis

```
# 4. Create a dataframe showing, for each level of drink, the mean, median,
# maximum, and standard deviation of scores.

# assigning a new variable
stats_by_drink <- caffeine %>%

# retrieving the 5 number summary for just the drinks
group_by(drink) %>%

# result
summarise(
   mean_score = mean(score, na.rm = TRUE),
   median_score = median(score, na.rm = TRUE),
   max_score = max(score, na.rm = TRUE),
   sd_score = sd(score, na.rm = TRUE)
 )

# Printing the result
print(stats_by_drink)
```

- A new dataframe, `stats_by_drink`, is created that summarizes key statistics for test scores (mean, median, maximum, and standard deviation) for each drink level. This could highlight which drinks are associated with higher or lower performance.

Visualization of Drink-Level Analysis

```
# 5. Write R code to plot the contents of the dataframe created in the previous
# step in a visually pleasant and informative way.

# installing and loading tidyverse
install.packages("tidyverse")
library(tidyverse)

#assigning a new variable
drink_stats_plot <- stats_by_drink %>%

# reshaping the data
pivot_longer(cols = -drink, names_to = "statistic", values_to = "value")

# Create the plot with ggplot2; details to make the plot visually pleasing
ggplot(drink_stats_plot, aes(x = drink, y = value, group = statistic, color = statistic)) +
  geom_line() +
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Score Statistics by Drink Level",
       x = "Drink",
       y = "Score Value",
       color = "Statistic") +
  scale_color_brewer(palette = "Set1") +
  theme(text = element_text(size = 12))
```

- The `tidyverse` package is loaded, which includes `ggplot2`, to create a visually appealing plot from the `stats_by_drink` dataframe.
- The plot shows score statistics by drink level using lines and points, making it easier to compare these statistics visually.

Analysis of Female Participants Over Age 20

```
# 6. Only for females above the age of 20, create a table showing, for each
# combined level of drink and cups, the mean, median, maximum, and standard
# deviation of scores. Also include a column showing how many people were in
# each group.

# assigning a new variable
female_drink_stats <- caffeine %>%

# filtering females above the age of 20
filter(gender == "female" & age > 20) %>%

# groupong by cups
group_by(drink, cups) %>%

# result
summarise(
   mean_score = mean(score, na.rm = TRUE),
   median_score = median(score, na.rm = TRUE),
   max_score = max(score, na.rm = TRUE),
   sd_score = sd(score, na.rm = TRUE),
   n = n(),
   .groups = "drop"
 )

# printing the result
print(female_drink_stats)
```

- A table is generated for female participants over the age of 20, summarizing score statistics by drink type and cups. It also includes the number of participants in each group.
- This could reveal if certain consumption patterns correlate with performance differently in this demographic.

Visualization of Female-Specific Analysis

```
# 7. Write R code to plot the contents of the table created in the previous
# step in a visually pleasant and informative way.

# assigning a new variable
female_drink_stats_long <- female_drink_stats %>%

# reshaping the data to plot it
pivot_longer(cols = c(mean_score, median_score, max_score),
             names_to = "statistic", values_to = "value")

# plotting
ggplot_female_drink_stats <- ggplot(female_drink_stats_long, aes(x = interaction(drink,
cups), y = value, fill = statistic)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  facet_wrap(~ statistic, scales = "free_y") +
  scale_fill_brewer(palette = "Pastel1") +
  theme_minimal() +
  labs(title = "Score Statistics by Drink Type and Cups for Females Over Age 20",
       x = "Drink-Cups",
       y = "Score Value",
       fill = "Statistic") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 45, hjust = 1))

# printing the plot
print(ggplot_female_drink_stats)
```

- The `female_drink_stats` table is reshaped and plotted, resulting into a bar chart that visually separates the statistics.
- This plot can provide a clear visual representation of how score statistics vary with drink type and amount for females over age 20.

Statistical Tests

```
# QUESTION 1: What is the correlation between test scores and type of drink?

# calculating anova test and summary
anova_result <- aov(score ~ drink, data = caffeine)
summary(anova_result)

# QUESTION 2: What is the correlation between test scores and amount of drink?

# calculating correlation coefficient
correlation_score_cups <- cor(caffeine$score, caffeine$cups, use = "complete.obs")
print(correlation_score_cups)

# QUESTION 3: What is the correlation between test scores and age?

# calculating the correlation coefficient
correlation_score_age <- cor(caffeine$score, caffeine$age, use = "complete.obs")
print(correlation_score_age)

# QUESTION 4: What is the correlation between test scores and gender?

# ensuring gender is a factor with two levels
caffeine$gender <- factor(caffeine$gender, levels = c("male", "female"))

# converting gender to a binary variable: 0 for male, 1 for female
caffeine$gender_numeric <- as.numeric(caffeine$gender == "female")

# calculating the correlation between score and gender
correlation_score_gender <- cor(caffeine$score, caffeine$gender_numeric, use =
"complete.obs")
```

 # printing the correlation

print(correlation_score_gender)

- An ANOVA test is conducted to determine if there are statistically significant differences in test scores among different types of drinks.
- Correlation coefficients are calculated between test scores and the amount of drink, as well as between test scores and age.
- The correlation coefficient is used to examine the relationship between test scores and gender, by converting gender to a binary variable.

Scatterplot Analyses

```
# Optional

# 1. Write R code to generate a scatterplot of the performance versus age, for
# different types of drink and comment of the result.

# loading ggplot2
library(ggplot2)

# plotting
ggplot(caffeine, aes(x = age, y = score, color = drink)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Performance vs. Age for Different Types of Drink",
       x = "Age",
       y = "Performance Score",
       color = "Type of Drink") +
  theme(legend.position = "bottom")

# 2. Write R code to generate a scatterplot of the performance versus age,
# for different amounts of drink and comment of the result.

# plotting
scatter_plot <- ggplot(caffeine, aes(x = age, y = score, color = as.factor(cups))) +
  geom_point(alpha = 0.6) +  # Set transparency to see overlapping points
  theme_minimal() +
  labs(title = "Performance vs. Age by Amount of Drink",
       x = "Age",
       y = "Performance Score",
       color = "Amount of Drink (cups)") +
```

```
  theme(legend.position = "bottom")

# printing the scatter plot
print(scatter_plot)

# 3. Write R code to generate a scatterplot of the performance versus age, for
# different genders and comment of the result.

# plotting
scatter_plot_gender <- ggplot(caffeine, aes(x = age, y = score, color = gender)) +
  geom_point(alpha = 0.6) +  # Use alpha to make points slightly transparent
  theme_minimal() +
  labs(title = "Performance vs. Age by Gender",
       x = "Age",
       y = "Performance Score",
       color = "Gender") +
  theme(legend.position = "bottom")

# printing the scatter plot
print(scatter_plot_gender)
```

- Three scatterplots are generated to visualize the relationship between performance scores and age, each considering a different variable: drink types, amount of drink, and gender.
- These plots can help identify trends and potential correlations between age and performance, and how these might differ by drink type, consumption amount, or gender.