```r
#PROBLEM 1
library(tidyverse)
library(e1071)
library(ggplot2)
cereal_data <- read.csv("UScereal1.csv")
max_protein_by_manufacturer <- cereal_data %>%
  group_by(mfr) %>%
  summarize(max_protein = max(protein, na.rm = TRUE))
print(max_protein_by_manufacturer)

#problem 2 a
cereal_data <- read.csv("UScereal1.csv")
print("Summary of missing values:")
print(colSums(is.na(cereal_data)))
replace_missing <- function(column) {
  if (is.numeric(column)) {
    if (any(is.na(column))) {
      if (shapiro.test(column)$p.value > 0.05) {
        return(ifelse(is.na(column), mean(column, na.rm = TRUE), column))
      } else if (skewness(na.omit(column)) < 0) {
        return(ifelse(is.na(column), min(column, na.rm = TRUE), column))
      } else {
        return(ifelse(is.na(column), max(column, na.rm = TRUE), column))
      }
    } else {
      return(column)
    }
  } else {
    return(column)
  }
}
cereal_data_filled <- cereal_data %>%
  mutate_all(replace_missing)
print("Summary of missing values after replacement:")
print(colSums(is.na(cereal_data_filled)))

#problem 2 b
summary(cereal_data_filled)

#problem 3 a
cereal_data_filled <- read.csv("UScereal1.csv")
ggplot(cereal_data_filled, aes(x = mfr, y = fibre, fill = mfr)) +
  geom_boxplot() +
  labs(title = "Spread of Fiber by Manufacturer",
       x = "Manufacturer",
       y = "Fiber") +
  theme_minimal()
ggplot(cereal_data_filled, aes(x = mfr, y = fibre, fill = mfr)) +
  geom_violin() +
  labs(title = "Spread of Fiber by Manufacturer",
       x = "Manufacturer",
       y = "Fiber") +
  theme_minimal()

#problem 3 b
ggplot(cereal_data_filled, aes(x = as.factor(shelf), y = calories, fill =
as.factor(shelf))) +
  geom_boxplot() +
  labs(title = "Outliers in Calories for Each Shelf",
       x = "Shelf",
       y = "Calories") +
  theme_minimal()

#problem 3 c
```

```
cereal_data_filled <- read.csv("UScereal1.csv")
numeric_vars <- select_if(cereal_data_filled, is.numeric)
pairs(numeric_vars, col = as.factor(cereal_data_filled$shelf))

#problem 4 a
cereal_data_filled <- read.csv("UScereal1.csv")
means <- colMeans(select(cereal_data_filled, -c("Name", "mfr", "vitamins")), na.rm = TRUE)
top_four_means <- names(sort(means, decreasing = TRUE)[1:4])
GreaterMeanFour <- cereal_data_filled[, c("Name", "mfr", "vitamins", top_four_means), drop
= FALSE]
print(GreaterMeanFour)

#problem 4 b
cereal_data_filled <- read.csv("UScereal1.csv")
means <- colMeans(select(cereal_data_filled, -c("Name", "mfr", "vitamins")), na.rm = TRUE)
top_four_means <- names(sort(means, decreasing = TRUE)[1:4])
GreaterMeanFour <- cereal_data_filled[, c("Name", "mfr", "vitamins", top_four_means), drop
= FALSE]
correlation_matrix <- cor(select(GreaterMeanFour, -c("Name", "mfr", "vitamins")))
pairs(GreaterMeanFour[, -c(1,2,3)], main = "Pairs Plot of GreaterMeanFour")
print("Correlation Matrix:")
print(correlation_matrix)

#........................................
cereal_data <- read.csv("UScereal1.csv")
corr_matrix <- cor(cereal_data[, c("calories", "protein", "fat", "sodium", "fibre",
"carbo", "sugars", "potassium")], use = "complete.obs")
corr_matrix

# Load the reshape2 package if not loaded
if (!requireNamespace("reshape2", quietly = TRUE)) {
  install.packages("reshape2")
}
library(reshape2)

# Load the ggplot2 package if not loaded
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
library(ggplot2)

# Create the correlation plot
ggplot(melt(corr_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  labs(title = "Correlation Matrix",
       x = "Variable 1",
       y = "Variable 2") +
  scale_fill_gradient2(low = "blue", mid = "white", high = "green", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for
better readability



#program 4_3

cereal_data <- read.csv("UScereal1.csv")

# Remove rows with missing values
cereal_data_clean <- na.omit(cereal_data)

# Fit the linear regression model
fit <- lm(potassium ~ fibre, data = cereal_data_clean)

# Create the scatter plot with regression line
```

```
ggplot(cereal_data_clean, aes(x = fibre, y = potassium)) +
  geom_point(color = "lightgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Simple Linear Regression",
       x = "Fiber Content (grams)",
       y = "Potassium Content (grams)")


#problem 4 c
cereal_data_filled <- read.csv("UScereal1.csv")
numeric_columns <- sapply(cereal_data_filled, is.numeric)
selected_vars <- names(numeric_columns)[numeric_columns][1:2]
cereal_data_filled[, selected_vars] <- lapply(cereal_data_filled[, selected_vars],
function(x) {
  ifelse(is.na(x), mean(x, na.rm = TRUE), x)
})
lm_model <- lm(as.formula(paste(selected_vars[1], "~", selected_vars[2])), data =
cereal_data_filled)
plot(cereal_data_filled[[selected_vars[2]]], cereal_data_filled[[selected_vars[1]]],
     main = "Simple Linear Regression", xlab = selected_vars[2], ylab = selected_vars[1])
abline(lm_model, col = "red")

#problem 4 d
cereal_data_filled <- read.csv("UScereal1.csv")
var1 <- "calories"
var2 <- "protein"
cereal_data_filled[, c(var1, var2)] <- lapply(cereal_data_filled[, c(var1, var2)],
function(x) {
  ifelse(is.na(x), mean(x, na.rm = TRUE), x)
})
lm_model <- lm(as.formula(paste(var1, "~", var2)), data = cereal_data_filled)
predicted_values_before <- predict(lm_model, newdata = cereal_data_filled)
residuals <- residuals(lm_model)
outliers <- which(abs(residuals) > 2 * sd(residuals))  # Adjust the threshold as needed
cereal_data_without_outliers <- cereal_data_filled[-outliers, ]
lm_model_no_outliers <- lm(as.formula(paste(var1, "~", var2)), data =
cereal_data_without_outliers)
predicted_values_after <- predict(lm_model_no_outliers, newdata = cereal_data_filled)
comparison <- data.frame(
  Name = cereal_data_filled$Name,
  Actual = cereal_data_filled[[var1]],
  Predicted_Before = predicted_values_before,
  Predicted_After = predicted_values_after
)

print(comparison)


# Concluding Points:
# 1. The dataset contains information about various cereal products, including nutritional
values and manufacturer details.
# 2. The analysis involved exploring variables like calories, protein, and their
relationships.
# 3. Outliers were identified and removed to enhance the accuracy of predictive models.
# 4. The linear regression model was used to predict values, showing variations before and
after outlier removal.
# 5. Further analysis and insights can be gained by exploring additional variables and
conducting more sophisticated modeling.

# Findings:
# 1. The correlation between calories and protein in cereals indicates a potential
relationship.
# 2. Outliers in the dataset had a notable impact on the accuracy of the predictive model.
# 3. Manufacturers such as General Mills, Kellogg's, and Quaker Oats have a significant
presence in the dataset.
```

```
# 4. The distribution of fiber preference across manufacturers varies, suggesting
different product strategies.
# 5. Imputation of missing values with the mean improved the robustness of the analysis,
but careful consideration is needed.
```