~~~
---
title: "Mobile Data Analytics"
output: pdf_document
date: "2024-03-01"
---

```{r}

# Question 1
mobile_data <- read.csv(file = "/Users/mihuynh/Downloads/Train Data Set/train.csv")
str(mobile_data)

# Turn the variable price range into a factor variable with levels:
# “0” for low, “1” for medium, “2” for high, and “3” for very high.
price_range <- factor(x = mobile_data$price_range, levels = c("0", "1", "2", "3"), labels
= c("Low", "Medium", "High", "Very high"), ordered = is.ordered(c))
str(price_range)

# Make a scatter plot between the variables battery power vs ram.
# Add colors based on price range.
library(ggplot2)
ggplot(data = mobile_data) +
  geom_point(aes(x = ram, y = battery_power, color = price_range)) +
  scale_color_distiller(palette = "Reds", labels = c("Low", "Medium", "High", "Very
high"))

# Find the Pearson correlation between the variables
# ram and battery power.
pearson <- cor(mobile_data$ram, mobile_data$battery_power, method = c("pearson"))
print(pearson)

# Create four separate data sets by sub-setting the “mobile data”
# using the variable price range as
# “priceLow”, “priceMedium”, “priceHigh” and “priceVeryhigh”.
priceLow <- subset(mobile_data, price_range == 0)
priceMedium <- subset(mobile_data, price_range == 1)
priceHigh <- subset(mobile_data, price_range == 2)
priceVeryhigh <- subset(mobile_data, price_range == 3)


# Calculate the Pearson correlation coefficient
# between the variable pair (ram , battery power) separately
# for each price range. Explain any correlations
# you might find in terms of how a cellphone operates.
LowCor <- cor(priceLow$ram, priceLow$battery_power, method = c("pearson"))
MedCor <- cor(priceMedium$ram, priceMedium$battery_power, method = c("pearson"))
HighCor <- cor(priceHigh$ram, priceHigh$battery_power, method = c("pearson"))
VeryhighCor <- cor(priceVeryhigh$ram, priceVeryhigh$battery_power, method = c("pearson"))
print(LowCor)
print(MedCor)
print(HighCor)
print(VeryhighCor)

# Recreate the plot from Part (b), and add the trend lines
# for each price range separately.
ggplot(mobile_data, aes(x = ram, y = battery_power, color = price_range)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, aes(group = price_range), color = "black") +
  scale_color_distiller(palette = "Blues", direction = 1, labels = c("Low", "Medium",
"High", "Very high"))

# Find the average and the medium clock speed of the
# mobile phones which has 4, 6 and 8 cores in their
# processors. Round your answer to two decimal places.
~~~

```
filtered_data <- subset(mobile_data, n_cores %in% c(4, 6, 8))
avg_clock_speed <- mean(filtered_data$clock_speed)
median_clock_speed <- median(filtered_data$clock_speed)
round(avg_clock_speed, digits = 2)
round(median_clock_speed, digits = 2)

# Make density curves of the ram where the 4 price ranges
# are in one plot and describe their shapes respectively.
ggplot(mobile_data, aes(x = ram, group = price_range, fill = price_range)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot", x = "Ram", fill = "Price Range") +
  scale_fill_distiller(palette = "Blues", direction = 1, labels = c("Low", "Medium",
"High", "Very high"))


# Make box plots of the ram where the 4 price ranges
# are in one plot and describe their shapes respectively
ggplot(mobile_data, aes(x = ram, group = price_range, fill = price_range)) +
  geom_boxplot() +
  labs(title = "Box Plot", x = "Ram", fill = "Price Range") +
  scale_fill_distiller(palette = "Blues", labels = c("Low", "Medium", "High", "Very
high"))


# Make a violin plot of the ram where the 4 price ranges
# are in one plot and describe their shapes respectively.
ggplot(mobile_data, aes(x = ram, y = price_range, group = price_range, fill =
price_range)) +
  geom_violin() +
  labs(title = "Violin Plot", x = "Ram", y = "Price Range", fill = "Price Range") +
  scale_fill_distiller(palette = "Blues", direction = 1, labels = c("Low", "Medium",
"High", "Very high"))

# Make a factor variable out of ram by taking the log2 (ram)
# and rounding that value to the nearest whole number.
log_ram <- log2(mobile_data$ram)
round(log_ram)
log_ram_factor <- as.factor(mobile_data$log_ram)

# Make a stacked bar plot to show the relationship between
# price range and log2(ram)
ggplot(mobile_data, aes(x = log_ram, fill = price_range, group = price_range)) +
  geom_bar() +
  labs(title = "Stacked Bar Plot", x = "Log Ram") +
  scale_fill_distiller(palette = "Blues", direction = 1, labels = c("Low", "Medium",
"High", "Very high"))

# MPG DATASET

# Problem 2a
# Turn the variable cyl to an ordered factor variable with levels
# “4”, “5”, “6”, and “8”
library(ggplot2)
data(mpg)
cyl <- factor(x = mpg$cyl, levels = c("4", "5", "6", "8"), ordered = is.ordered(c))
levels(cyl)

# Problem 2b
# Turn the variable trans to a factor variable,
# of which unique values are “auto” and “manu”
trans <- factor(substr(mpg$trans, 1, 4), levels = c("auto", "manu"))
levels(trans)

# Problem 2c
# Turn the variable drv to an ordered factor variable
```

```r
# with levels “f”, “r”, and “4”
drv <- factor(mpg$drv, ordered = TRUE, levels = c("f", "r", "4"))
levels(drv)

# Problem 2d
# Turn the variable fl to a factor variable, of
# which unique values are “gasoline”, “diesel”, and “other”
fl <- factor(ifelse(mpg$fl %in% c("d", "x"), "diesel",
                ifelse(mpg$fl %in% c("e", "c"), "other", "gasoline")))
levels(fl)

# Problem 2e
# Turn the variable class to an ordered factor variable
# with levels “2seater”, “subcompact”, “compact”,
# “midsize”, “suv”, “minivan”, and “pickup”
class <- factor(mpg$class, ordered = TRUE, levels = c("2seater", "subcompact", "compact",
"midsize", "suv", "minivan", "pickup"))
levels(class)

# Problem 2f
# Create a new variable of country to indicate the
# manufacturer base location
country_lookup <- data.frame(manufacturer = c("audi", "chevrolet", "dodge", "ford",
"honda", "hyundai", "jeep", "land rover", "lincoln", "mercury", "nissan", "pontiac",
"subaru", "toyota", "volkswagen"), country = c("Germany", "USA", "USA", "USA", "Japan",
"South Korea", "USA", "UK", "USA", "USA", "Japan", "USA", "Japan", "Japan", "Germany"))
mpg <- merge(mpg, country_lookup, by.x = "manufacturer", by.y = "manufacturer", all.x =
TRUE)
head(mpg)

# Problem 2g
# Draw a bar plot of the variable country and
# arrange the country in decreasing order in terms of the
# number of samples.
library(magrittr)
library(dplyr)
manufacturer_counts <- mpg %>%
  count(manufacturer) %>%
  arrange(desc(n))
mpg$manufacturer <- reorder(mpg$manufacturer, mpg$manufacturer, function(x) sum(x ==
manufacturer_counts$manufacturer))
ggplot(mpg, aes(x = manufacturer)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Number of Samples by Manufacturer", x = "Manufacturer", y = "Number of
Samples") +
  theme_minimal()

# Problem 2h
# Summarize what a typical U.S. car looks like,
# in terms of engine displacement (i.e. displ), number of
# cylinders (i.e. cyl), type of transmission (i.e. trans),
# drive type (i.e. drv), fuel type (i.e. fl), and type
# of car (i.e. class)?
us_cars <- subset(mpg, manufacturer == "ford" | manufacturer == "chevrolet" | manufacturer
== "dodge" | manufacturer == "mercury" | manufacturer == "pontiac" | manufacturer ==
"lincoln")
summary_us_cars <- summary(us_cars[, c("displ", "cyl", "trans", "drv", "fl", "class")])
print(summary_us_cars)

# Problem 2i
# Make a boxplot of the combined miles per gallon
# (i.e. (cty + hwy)/2) of U.S. cars and Japan cars,
# respectively, and report their means, medians,
# standard deviations, and IQRs.
mpg$combined_mpg <- (mpg$cty + mpg$hwy) / 2
```

```
us_cars <- subset(mpg, manufacturer %in% c("ford", "chevrolet", "dodge", "mercury",
"pontiac", "lincoln"))
japan_cars <- subset(mpg, manufacturer %in% c("honda", "toyota", "nissan", "subaru",
"mazda", "mitsubishi"))
ggplot(mapping = aes(x = "U.S. Cars", y = combined_mpg)) +
  geom_boxplot(data = us_cars) +
  labs(title = "Combined Miles Per Gallon of U.S. Cars",
       y = "Combined MPG") +
  theme_minimal()
ggplot(mapping = aes(x = "Japan Cars", y = combined_mpg)) +
  geom_boxplot(data = japan_cars) +
  labs(title = "Combined Miles Per Gallon of Japan Cars",
       y = "Combined MPG") +
  theme_minimal()
us_mean <- mean(us_cars$combined_mpg)
us_median <- median(us_cars$combined_mpg)
us_sd <- sd(us_cars$combined_mpg)
us_iqr <- IQR(us_cars$combined_mpg)
japan_mean <- mean(japan_cars$combined_mpg)
japan_median <- median(japan_cars$combined_mpg)
japan_sd <- sd(japan_cars$combined_mpg)
japan_iqr <- IQR(japan_cars$combined_mpg)
cat("Summary statistics for U.S. cars: \"")
cat("Mean: ", us_mean, "")
cat("Median: ", us_median, "")
cat("Standard Deviation: ", us_sd, "\n")
cat("Interquartile Range (IQR): ", us_iqr, "\n")
cat("Summary statistics for Japan cars: \"")
cat("Mean:", japan_iqr, "\n")

# Problem 2j
# Make a histogram of the engine displacement
# (i.e. displ) of U.S. cars and Japan cars, respectively,
# and describe their shape
us_cars <- subset(mpg, manufacturer %in% c("ford", "chevrolet", "dodge", "mercury",
"pontiac", "lincoln"))
japan_cars <- subset(mpg, manufacturer %in% c("honda", "toyota", "nissan", "subaru",
"mazda", "mitsubishi"))
ggplot(us_cars, aes(x = displ)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  labs(title = "Engine Displacement of U.S. Cars",
       x = "Engine Displacement",
       y = "Frequency")
ggplot(japan_cars, aes(x = displ)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(title = "Engine Displacement of Japan Cars",
       x = "Engine Displacement",
       y = "Frequency")

```