MATH 3070 Lab Project 9

Prachi Aswani

October 31, 2024

Contents

Problem 1 (Verzani problem 3.17) 1
Problem 2 (Verzani problem 3.20) 3
Problem 3 (Verzani problem 3.32) 5
Problem 4 (Verzani problem 3.33) 7

*Remember: I expect to see commentary either in the text, in the code with comments created using `#`, or (preferably) both! **Failing to do so may result in lost points!***

Problem 1 (Verzani problem 3.17)

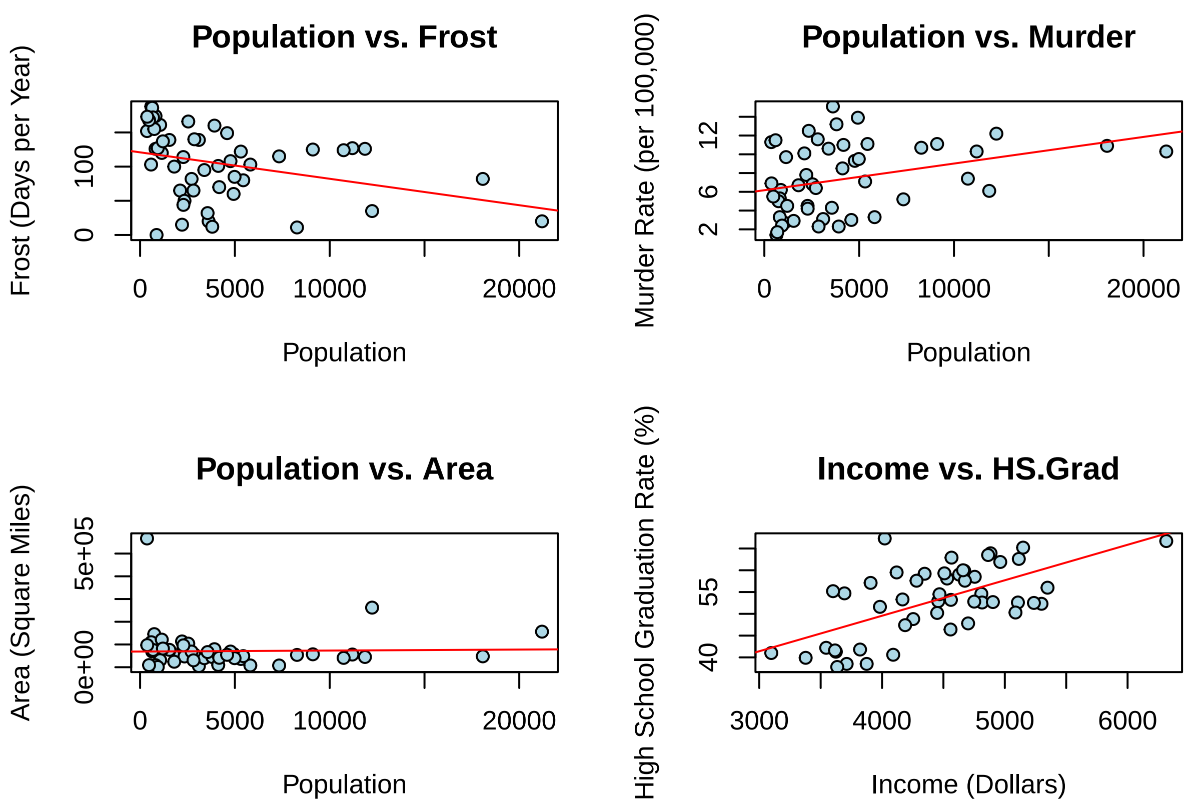
The `state.x77` data set contains various information for each of the fifty United States. We wish to explore possible relationships among the variables. First, we make the data set easier to work with by turning it into a data frame.

```
x77 <- data.frame(state.x77)
```

Now, make scatter plots of `Population` and `Frost`; `Population` and `Murder`; `Population` and `Area`; and `Income` and `HS.Grad`. Do any relationships appear linear? Are there any surprising correlations?

```
# Set up a 2x2 plotting area for better visualization of multiple plots
par(mfrow = c(2, 2))

# Scatter plot of Population vs. Frost
plot(x77$Population, x77$Frost,
     main = "Population vs. Frost",
     xlab = "Population",
     ylab = "Frost (Days per Year)",
     pch = 21, bg = "lightblue")
abline(lm(Frost ~ Population, data = x77), col = "red")  # Add a regression line

# Scatter plot of Population vs. Murder
plot(x77$Population, x77$Murder,
     main = "Population vs. Murder",
     xlab = "Population",
     ylab = "Murder Rate (per 100,000)",
     pch = 21, bg = "lightblue")
```

```r
abline(lm(Murder ~ Population, data = x77), col = "red")

# Scatter plot of Population vs. Area
plot(x77$Population, x77$Area,
     main = "Population vs. Area",
     xlab = "Population",
     ylab = "Area (Square Miles)",
     pch = 21, bg = "lightblue")
abline(lm(Area ~ Population, data = x77), col = "red")

# Scatter plot of Income vs. HS.Grad
plot(x77$Income, x77$HS.Grad,
     main = "Income vs. HS.Grad",
     xlab = "Income (Dollars)",
     ylab = "High School Graduation Rate (%)",
     pch = 21, bg = "lightblue")
abline(lm(HS.Grad ~ Income, data = x77), col = "red")
```




```r
# Reset plotting area to 1x1
par(mfrow = c(1, 1))

# Population vs. Frost: This plot shows a weak negative linear trend.
# As population increases, frost days tend to decrease slightly,
# which could be associated with warmer climates in more populated areas.
# However, the correlation isn't strong.
```

```r
# Population vs. Murder: There is a weak positive trend, suggesting that as
# population increases, the murder rate may slightly increase, but it's
# not a strong linear relationship.

#Population vs. Area: There is virtually no linear trend here, as
# population and area vary independently.

# Income vs. HS.Grad: This is the clearest linear relationship
# of the four plots. There’s a positive correlation between
# income and high school graduation rates, where higher income
# levels are generally associated with higher graduation rates.
```

Problem 2 (Verzani problem 3.20)

The **`batting`** *(**UsingR**) data set contains baseball statistics for the 2002 Major League Baseball season. What is the correlation between the number of strikeouts (`SO`) and the number of home runs (`HR`)? Make a scatter plot to see whether there is any trend. Does the data suggest that in order to hit a lot of home runs one should strike out a lot?*

```r
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 4.3.3
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Warning: package 'HistData' was built under R version 4.3.3
```

```
## Loading required package: Hmisc
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
data("batting")

# Display the structure of the data to understand available variables
str(batting)
```

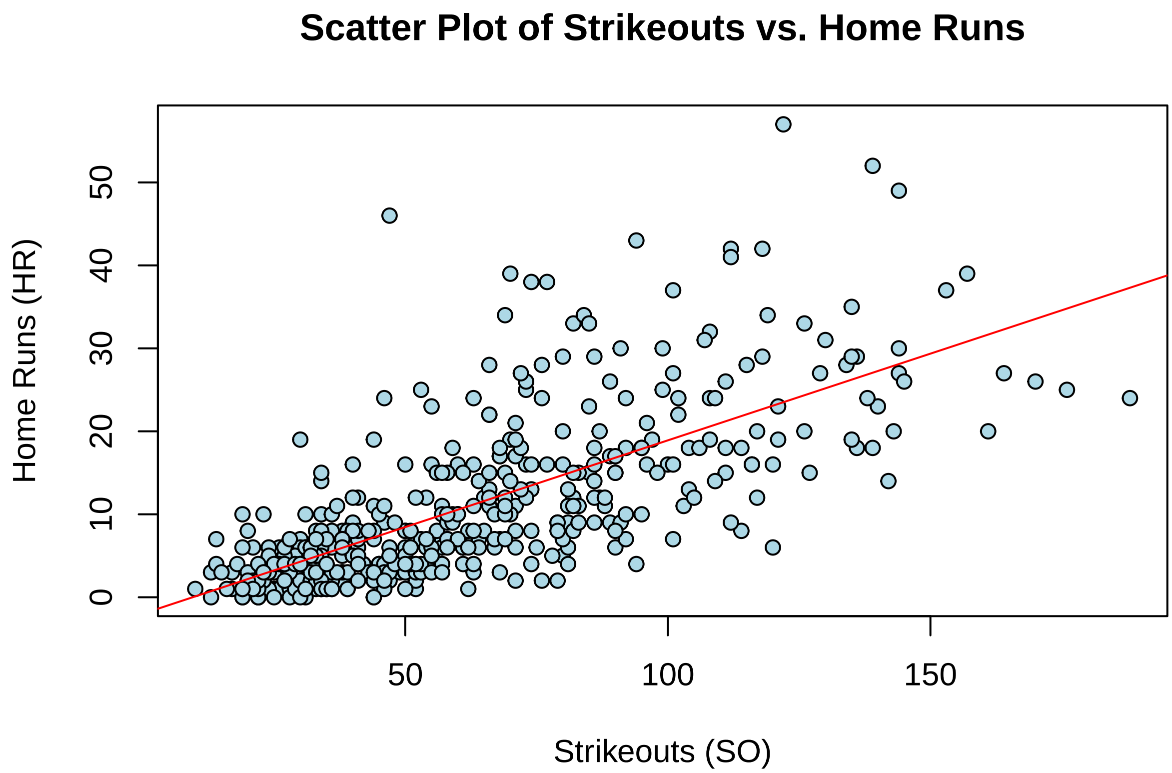
```
## 'data.frame':    438 obs. of  22 variables:
##  $ playerID: Factor w/ 1217 levels "abbotpa01","abernbr01",..: 1207 1125 1107 1086 996 928 679 648 6
##  $ yearID  : num  2002 2002 2002 2002 2002 ...
##  $ stintID : num  1 1 1 1 1 1 1 2 1 2 ...
##  $ teamID  : Factor w/ 30 levels "ANA","ARI","ATL",..: 11 21 28 16 9 29 24 18 11 4 ...
##  $ lgID    : Factor w/ 2 levels "AL","NL": 1 1 1 2 1 1 2 2 1 1 ...
```

```
##  $ G       : num  54 56 44 62 65 69 76 90 33 52 ...
##  $ AB      : num  201 133 168 137 159 158 185 231 107 109 ...
##  $ R       : num  25 22 17 16 15 17 19 33 10 10 ...
##  $ H       : num  57 30 36 34 34 33 44 59 25 23 ...
##  $ DOUBLE  : num  14 8 2 9 7 4 3 17 4 6 ...
##  $ TRIPLE  : num  0 0 1 2 2 0 0 1 0 0 ...
##  $ HR      : num  7 2 0 8 6 4 2 7 0 2 ...
##  $ RBI     : num  27 8 9 24 21 14 19 33 6 9 ...
##  $ SB      : num  2 3 7 1 0 4 1 5 3 1 ...
##  $ CS      : num  0 0 1 0 1 2 1 6 2 0 ...
##  $ BB      : num  12 15 7 7 15 17 9 13 8 3 ...
##  $ SO      : num  39 32 19 38 27 45 33 44 13 20 ...
##  $ IBB     : num  5 1 0 0 2 0 0 0 0 0 ...
##  $ HBP     : num  2 5 1 2 0 5 2 1 1 0 ...
##  $ SH      : num  0 1 3 0 0 4 3 4 4 0 ...
##  $ SF      : num  1 1 1 0 2 2 1 3 1 0 ...
##  $ GIDP    : num  12 4 1 7 4 2 5 2 4 3 ...
```

```r
# Calculate the correlation between strikeouts (SO) and home runs (HR)
cor_SO_HR <- cor(batting$SO, batting$HR, use = "complete.obs")
print(paste("Correlation between strikeouts (SO) and home runs (HR):", cor_SO_HR))
```

```
## [1] "Correlation between strikeouts (SO) and home runs (HR): 0.70846970467262"
```

```r
# Plotting the scatter plot of SO vs. HR
plot(batting$SO, batting$HR,
     main = "Scatter Plot of Strikeouts vs. Home Runs",
     xlab = "Strikeouts (SO)",
     ylab = "Home Runs (HR)",
     pch = 21, bg = "lightblue")

# Adding a regression line to visualize the trend
abline(lm(HR ~ SO, data = batting), col = "red")
```




```
# The correlation of 0.7 suggests that as the number of strikeouts increases,
# the number of home runs also tends to increase.
# This implies that players who strike out more often are
# likely to hit more home runs. While the correlation suggests
# a relationship, it doesn't imply causation. The increase in
# home runs doesn't mean that striking out causes home runs.
# There could be other factors at play, such as player skill
# levels, batting techniques, or overall game strategies.
```

Problem 3 (Verzani problem 3.32)

*The data set `UScereal` (**MASS**) contains information about cereals on a shelf of a United States grocery store. Make a table showing the relationship between manufacturer, `mfr`, and shelf placement, `shelf`. Are there any obvious differences between manufacturers?*

```
# Your code here
# Load the UScereal dataset from the MASS package
library(MASS)
data("UScereal")

# Display the structure of the dataset to understand its variables
str(UScereal)
```

```
## 'data.frame':    65 obs. of  11 variables:
```

```
##  $ mfr      : Factor w/ 6 levels "G","K","N","P",..: 3 2 2 1 2 1 6 4 5 1 ...
##  $ calories : num  212 212 100 147 110 ...
##  $ protein  : num  12.12 12.12 8 2.67 2 ...
##  $ fat      : num  3.03 3.03 0 2.67 0 ...
##  $ sodium   : num  394 788 280 240 125 ...
##  $ fibre    : num  30.3 27.3 28 2 1 ...
##  $ carbo    : num  15.2 21.2 16 14 11 ...
##  $ sugars   : num  18.2 15.2 0 13.3 14 ...
##  $ shelf    : int  3 3 3 1 2 3 1 3 2 1 ...
##  $ potassium: num  848.5 969.7 660 93.3 30 ...
##  $ vitamins : Factor w/ 3 levels "100%","enriched",..: 2 2 2 2 2 2 2 2 2 2 ...
```

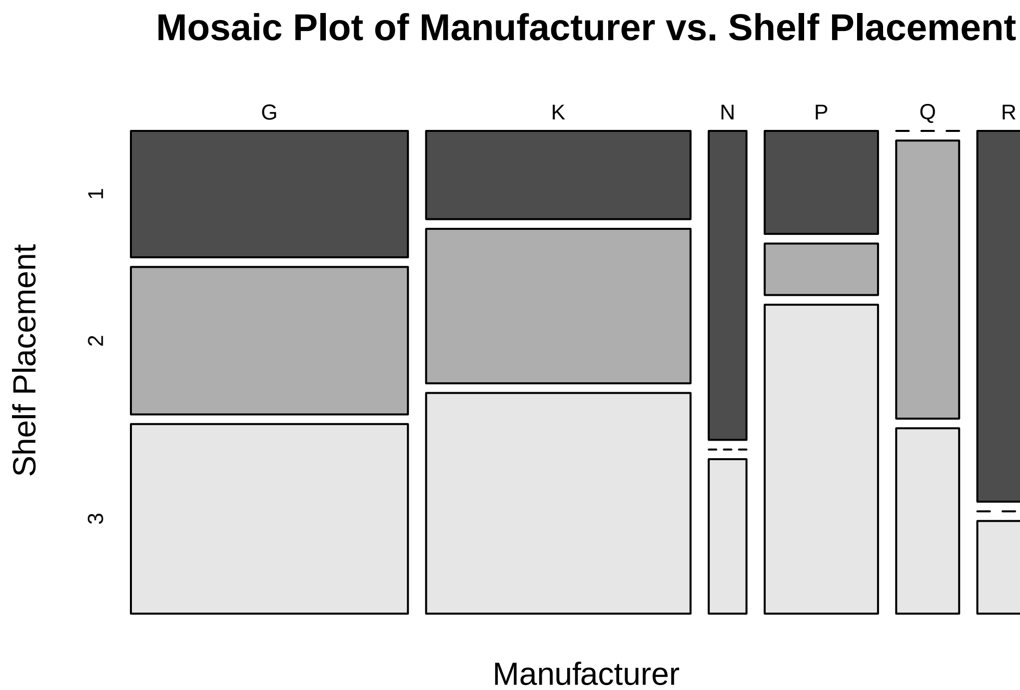
```r
# Create a contingency table to show the relationship between manufacturer (mfr) and shelf placement (s
cereal_table <- table(UScereal$mfr, UScereal$shelf)

# Create a flat contingency table for better readability
flat_cereal_table <- ftable(cereal_table)

# Print the flat contingency table
print("Flat Contingency Table of Manufacturer vs. Shelf Placement:")
```

```
## [1] "Flat Contingency Table of Manufacturer vs. Shelf Placement:"
```

```r
print(flat_cereal_table)
```

```
##     1  2  3
##            
## G   6  7  9
## K   4  7 10
## N   2  0  1
## P   2  1  6
## Q   0  3  2
## R   4  0  1
```

```r
# Plotting the contingency table as a mosaic plot for a visual representation
mosaicplot(cereal_table,
           main = "Mosaic Plot of Manufacturer vs. Shelf Placement",
           xlab = "Manufacturer",
           ylab = "Shelf Placement",
           color = TRUE)
```



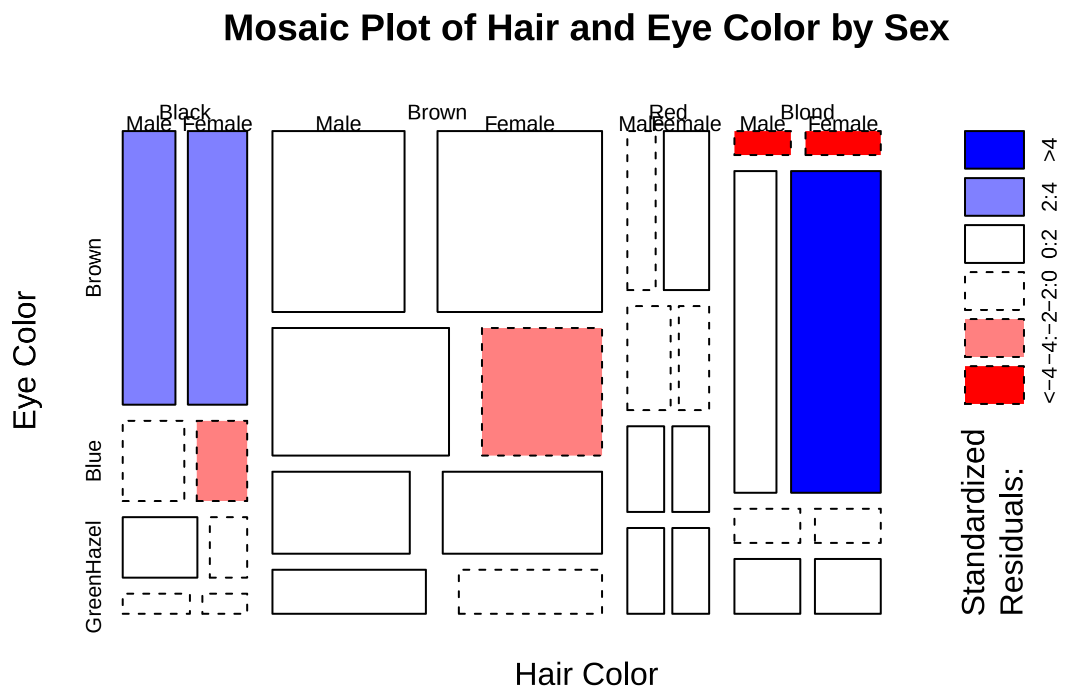

```
# Manufacturers G and K dominate the higher shelf placements,
# particularly Shelf 3, indicating they might prioritize visibility
# for their products.
# Manufacturer N stands out due to its significantly lower numbers,
# particularly on Shelf 2.
#It appears that some manufacturers, like G and K, may be employing
# strategies to maximize visibility by placing more products on higher
# shelves. Conversely, N and R show a lack of presence on certain shelves,
# possibly indicating a different market strategy or a lower demand for
# their cereals.
```

Problem 4 (Verzani problem 3.33)

The help page for `mosaicplot()` demonstrates the data set `HairEyeColor`, which records `sex`, `Hair` color, and `Eye` color for 592 statistics students. The data set comes as a flattened table, so simply passing the object to `mosaicplot()` will create the plot. (Or, as demonstrated, passing `shade = TRUE`, as in `mosaicplot(HairEyeColor, shade = TRUE)`, will produce a colored version.) Make the plot. Why does the help page note, "there are more blue-eyed, blonde females than expected?"

```
# Load the HairEyeColor dataset
data("HairEyeColor")

# Create a mosaic plot with shading
mosaicplot(HairEyeColor, shade = TRUE,
           main = "Mosaic Plot of Hair and Eye Color by Sex",
```

```
         xlab = "Hair Color",
         ylab = "Eye Color")
```




```
# The mosaic plot displays a pronounced blue shading for the
# category of blue-eyed, blonde females, indicating that this
# group is overrepresented in the dataset. This observation
# suggests that there are more blue-eyed, blonde females
# than would be expected based on random distributions of
# hair and eye colors. Factors such as demographic trends
# or cultural preferences could explain this pattern,
# highlighting an interesting aspect of the dataset.
```