MATH 3070 Lab Project 7

Prachi Aswani

October 17, 2024

Contents

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 5.6)

For the `batting` *(**UsingR**) data set, make parallel boxplots of the batting average (*`H/AB`*) for each team. Which team had the greatest median average? (Use* **lattice** *functions for this problem.)*

```
# Load the UsingR package and explore the structure of the batting dataset
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
```

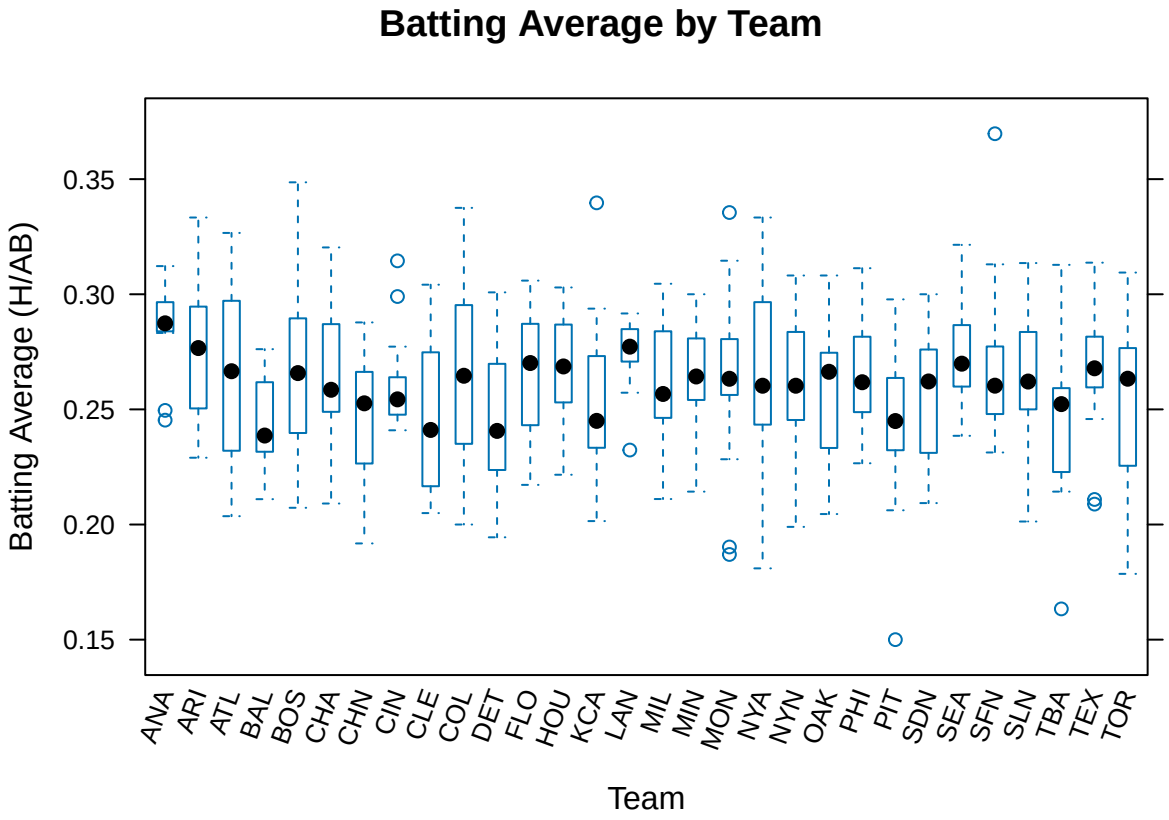
```
## Loading required package: Hmisc
```

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

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

```r
# Load the lattice library for plotting
library(lattice)

# Create a parallel boxplot for the batting average (H/AB) for each team
bwplot(H / AB ~ teamID, data = batting,
       scales = list(x = list(rot = 70)),  # Rotate x-axis labels by 70 degrees for better readability
       xlab = "Team",  # Label for x-axis
       ylab = "Batting Average (H/AB)",  # Label for y-axis
       main = "Batting Average by Team")  # Add a title to the plot
```




```r
# This plot uses lattice to create parallel boxplots for each team's batting average (H/AB).
# The team with the greatest median average is "ANA."
```

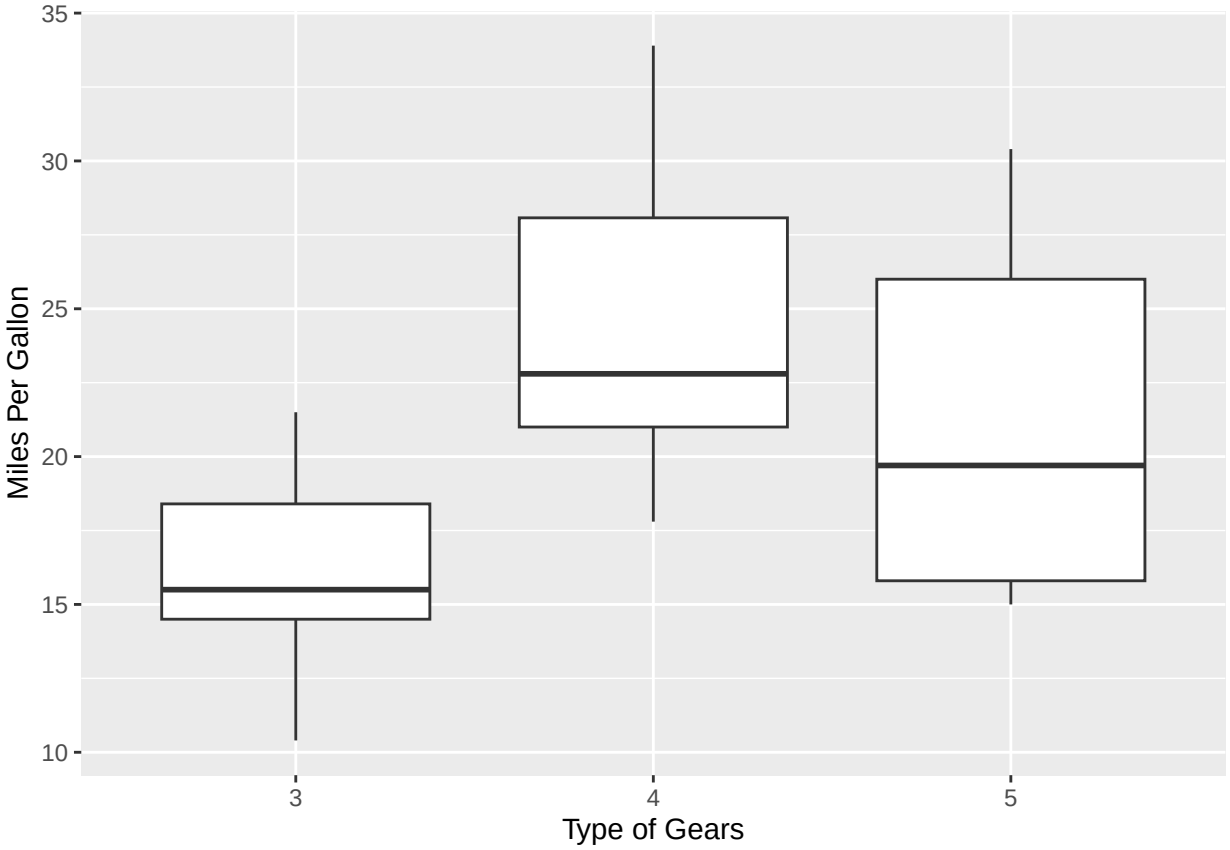
Problem 2 (Verzani problem 5.7)

For the `mtcars` *data set, produce graphics of the following using* ***ggplot2****:*

1. *Boxplots for miles per gallon (*`mpg`*) for groups defined by the number of gears (*`gear`*).*

```r
library(ggplot2)
qplot(as.character(gear), mpg, data = mtcars, geom = "boxplot",
      xlab = "Type of Gears", ylab = "Miles Per Gallon")
```

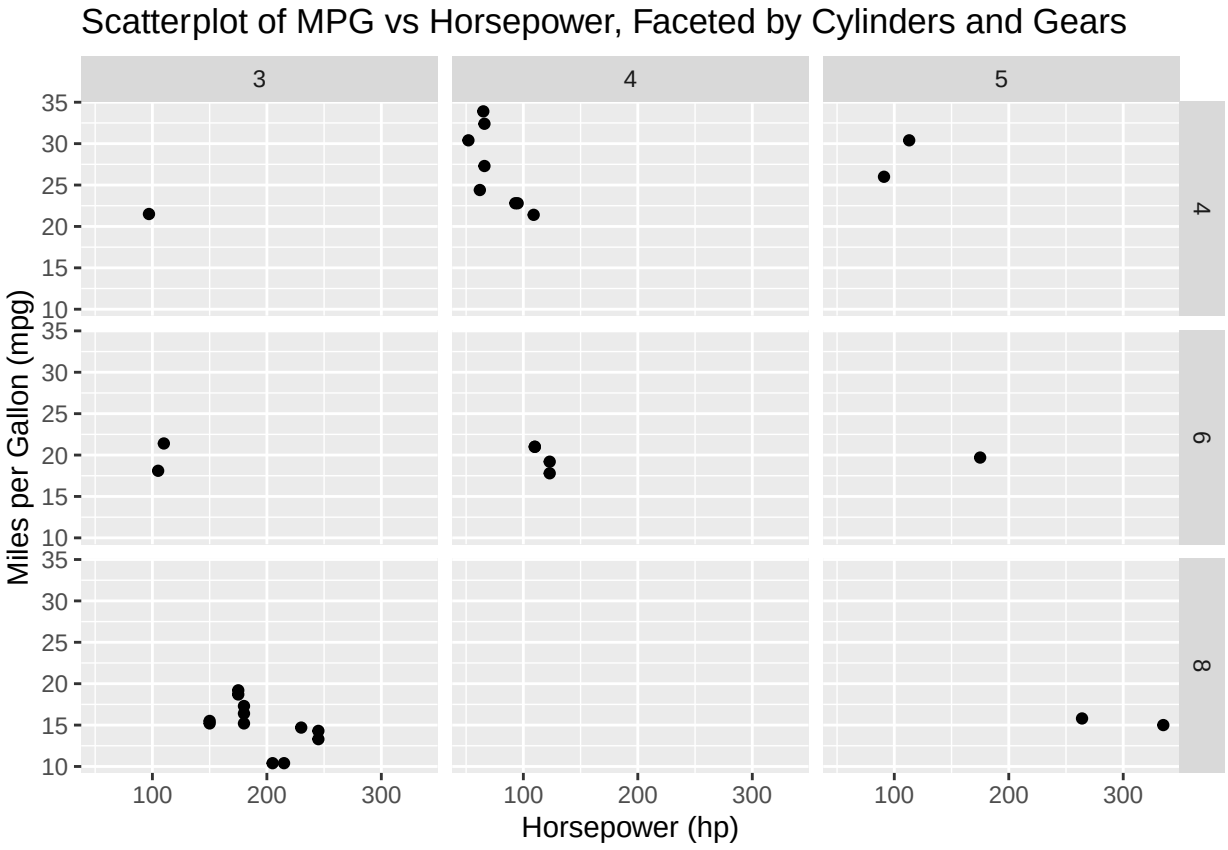
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# This ggplot creates three boxplots defined by the number of gears (3, 4, or 5)
# and shows the miles per gallon (mpg) for cars with each type of gear.
```

3. *A scatterplot of* ***mpg*** *modeled by horsepower (****hp****). Create facets by the number of cylinders (****cyl****) and* ***gear****.*

```
# Create a scatterplot using ggplot2 with horsepower (hp)
# on the x-axis and miles per gallon (mpg) on the y-axis
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point() +  # Add points to represent the relationship between hp and mpg
  facet_grid(cyl ~ gear) +  # Create facets: cylinders (cyl)
  #define rows and gears (gear) define columns
  xlab("Horsepower (hp)") +  # Label for x-axis
  ylab("Miles per Gallon (mpg)") +  # Label for y-axis
  ggtitle("Scatterplot of MPG vs Horsepower, Faceted by Cylinders and Gears")  # Add a title
```




```
#This ggplot2 scatterplot shows miles per gallon (mpg) plotted
# against horsepower (hp) for each car in the mtcars dataset.
# The facets represent different combinations of the number of cylinders (on the y-axis)
# and gears (on the x-axis).
# This helps us see the patterns between mpg and hp for different groups of cars
# based on their number of gears and cylinders.
```

Problem 3

Using the `batting` *data set (UsingR), create a visualization that does the following:*

- *Plots the* rate *of intentional walks (that is, the number of intentional walks divided by the number of times a player was at bat; these are the* `IBB` *and* `AB` *variables in the data set, respectively) against the* rate *of home runs (the* `HR` *variable in the data set) as a scatterplot*
- *Draws a trend line for these variables*
- *Identifies and labels the outlier in the data set in these variables (easily spotted once the scatter plot is drawn)*

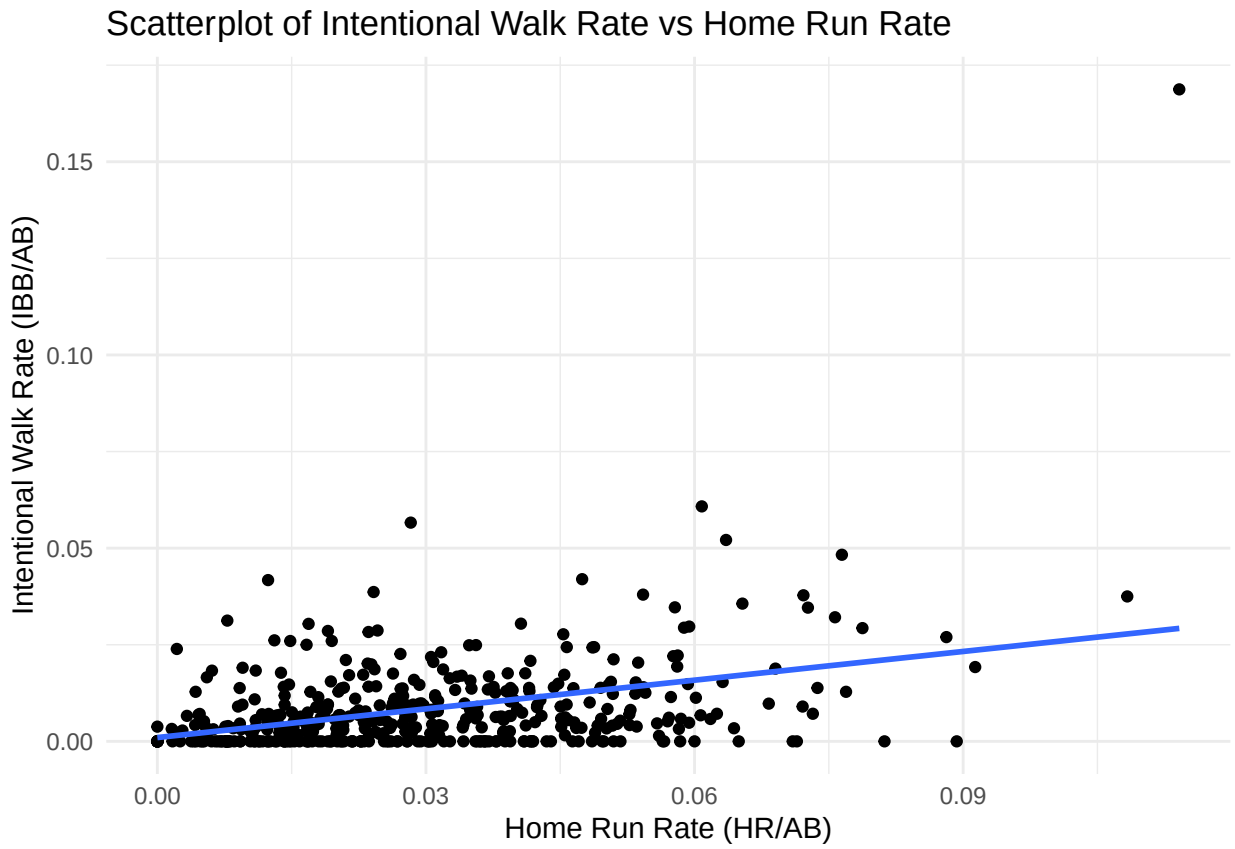
(Hint: `geom`*-type functions can accept data arguments and will use the data set passed rather than the default for the chart. So for the third requirement, consider adding a text layer with* `geom_text(data = ..., aes(...))` *where the argument passed to* `data` *is a subset of the data set consisting of the outlier, and* `aes(...)` *defines how to label that outlier.)*

```r
# Load necessary libraries
library(UsingR)
library(ggplot2)

# Calculate the rate of intentional walks (IBB/AB) and the rate of home runs (HR/AB)
batting$rate_IBB <- batting$IBB / batting$AB  # Intentional walk rate
batting$rate_HR <- batting$HR / batting$AB  # Home run rate

# Create a scatterplot of the rate of intentional walks against the rate of home runs
ggplot(batting, aes(x = rate_HR, y = rate_IBB)) +
  geom_point() +  # Plot points for each player's rates
  geom_smooth(method = "lm", se = FALSE) +  # Add a trend line
  xlab("Home Run Rate (HR/AB)") +  # Label for the x-axis
  ylab("Intentional Walk Rate (IBB/AB)") +  # Label for the y-axis
  ggtitle("Scatterplot of Intentional Walk Rate vs Home Run Rate") +  # Title of the plot
  theme_minimal()  # Use a minimal theme for a cleaner look
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

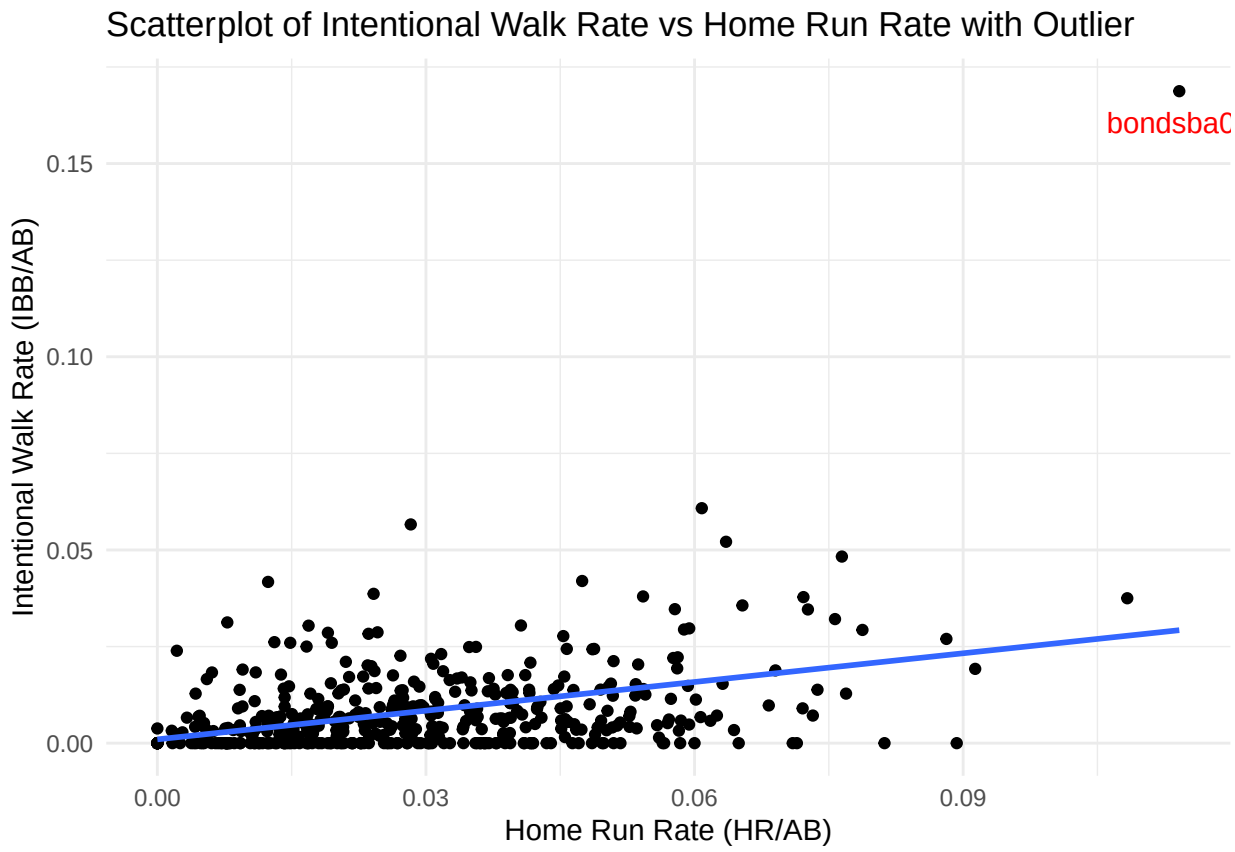



```r
# Identify the outlier (player with the highest intentional walk rate)
outlier <- batting[which.max(batting$rate_IBB), ]

# Add a label to the outlier in the scatterplot using geom_text
ggplot(batting, aes(x = rate_HR, y = rate_IBB)) +
```

```
  geom_point() +  # Plot points for each player's rates
  geom_smooth(method = "lm", se = FALSE) +  # Add a trend line
  geom_text(data = outlier, aes(label = playerID), vjust = 2, color = "red") +
  # Label the outlier with playerID
  xlab("Home Run Rate (HR/AB)") +  # Label for the x-axis
  ylab("Intentional Walk Rate (IBB/AB)") +  # Label for the y-axis
  ggtitle("Scatterplot of Intentional Walk Rate vs Home Run Rate with Outlier") +  # Title of the plot
  theme_minimal()  # Clean theme
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



BONUS Problem

Reconsider the data set from a previous project containing data about the results of 2012 Olympics. I load the data in for you below:

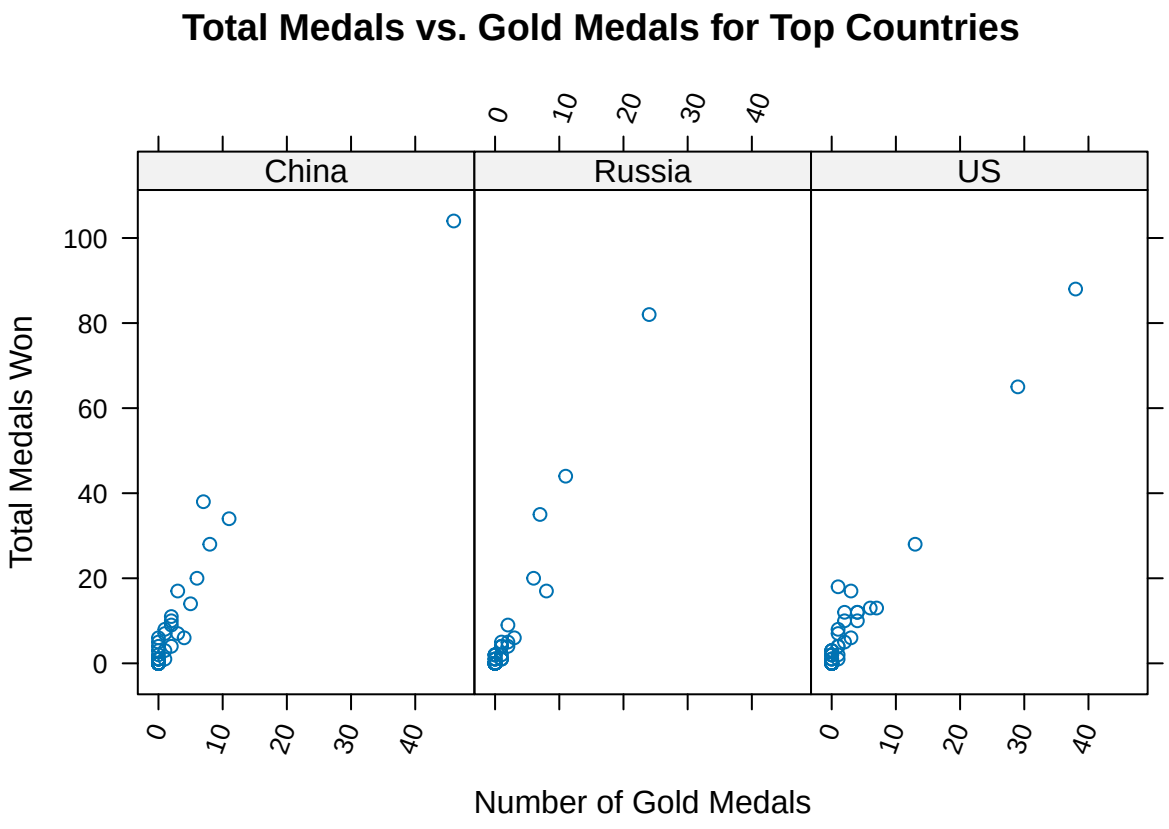
```
setwd("C:/Users/Prachi/OneDrive/Documents/")
olympic2012 <- read.csv("olympic-medals2012.csv")
# Some variables are read in as strings when, in truth, they are numeric (they separate thousands with
# leading to them being read as strings). I fix this below using the transform function,
# which allows for modifying columns in a data frame using methods similar to with.
olympic2012 <- transform(olympic2012, GDP.2011 = as.numeric(gsub(",", "", GDP.2011)),
                         pop.2010 = as.numeric(gsub(",", "", pop.2010)))
```

*Use any plotting system (base R, **lattice**, **ggplot2**) to create plot involving at least three variables in the `olympic2012` data set. Explain the relationship you explored and any interesting findings.* ***Bonus points will be given for plots that I consider exceptionally clean, clear, and insightful, accompanied with good analyses of what you found.***

```
# Load necessary libraries
library(ggplot2)
library(lattice)

# Identify countries that won more than 20 gold, silver, and bronze medals
# This focuses on the most successful countries in the 2012 Olympics
top_countries <- olympic2012$Country.name[olympic2012$Gold > 20 &
                                          olympic2012$Silver > 20 &
                                          olympic2012$Bronze > 20]

# Create a scatter plot using xyplot from the lattice package
# This plot displays the relationship between total medals
# and gold medals for top-performing countries
xyplot(Total ~ Gold | top_countries,  # Total medals as a function of gold medals
       data = olympic2012,  # Use the olympic2012 dataset
       scales = list(x = list(rot = 70)),  # Rotate x-axis labels for better readability
       main = "Total Medals vs. Gold Medals for Top Countries",  # Main title for the plot
       xlab = "Number of Gold Medals",  # Label for the x-axis
       ylab = "Total Medals Won",  # Label for the y-axis
)
```



```
#The scatterplot that I have created below, plots the total medals received
# against the number of gold medals received by the three countries that received
# the most medals, namely, China, Russia and the US.
```