Homework 4

Hannah Marr

II. R

6. Create a boxplot of the distribution of Plain pizza prices over the years using the cleaned dataset pizza data from class. Then, create a histogram displaying the frequency of Plain' pizza prices in 2022. Provide the R code. (2 points)

```
# Load the necessary libraries for data manipulation and visualization
library(tidyverse)  # Collection of R packages for data science
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidyr)      # Specifically for tidying data
library(dplyr)      # For data manipulation (select, filter, group, etc.)
```

```
# Read the pizza data from a CSV file into a dataframe
pizza_raw_data <- read.csv("/Users/hannahmarr/Desktop/Tufts/DATA200/Labs/Pizza_NYC.csv")

# Select specific columns from the dataset for analysis (Name, location, date, price, and style)
pizza_data <- pizza_raw_data %>%
  select(Name, location_lat, location_lng, Date, Year, Price, Style)

# Remove rows with missing values (NAs) to clean the data
pizza_data <- drop_na(pizza_data)

# Check for any remaining missing values in each column after cleaning
colSums(is.na(pizza_data))
```

```
##         Name location_lat location_lng         Date         Year        Price 
##            0            0            0            0            0            0 
##        Style 
##            0
```

```
# Get the dimensions (number of rows and columns) of the cleaned dataset
dim(pizza_data)
```

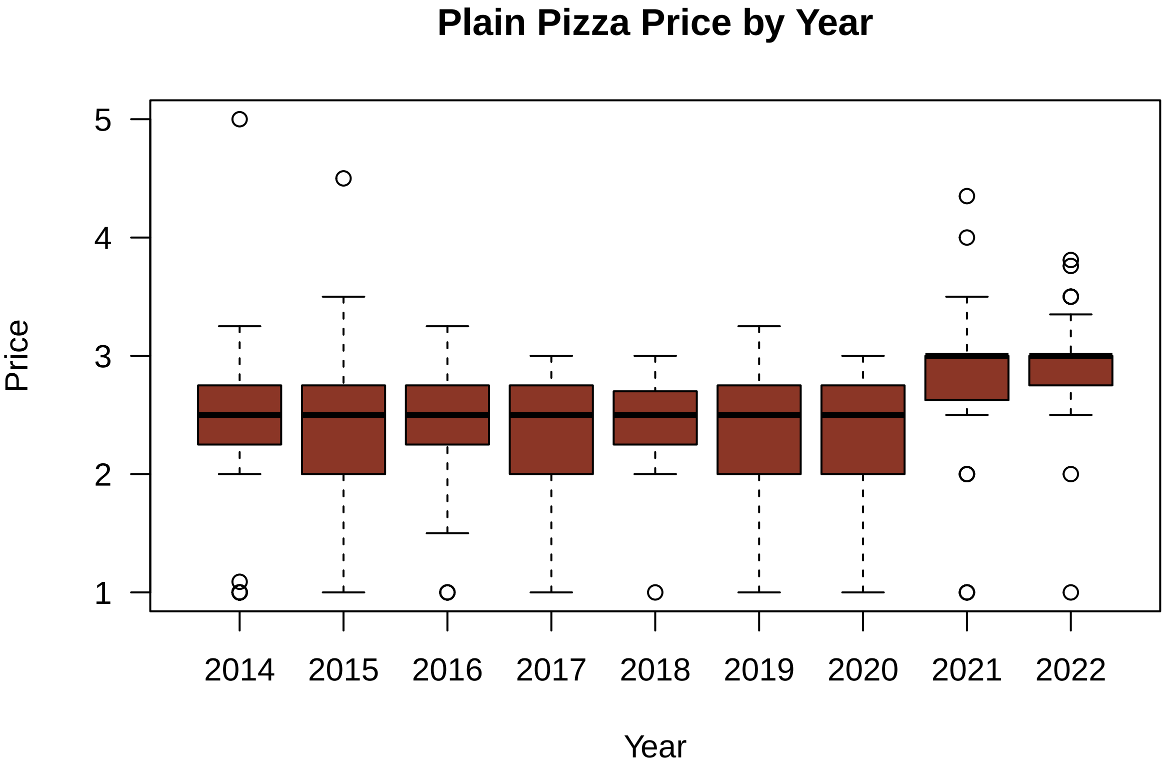
```
## [1] 464   7
```

```
# Filter the pizza_data dataframe into a new dataframe with only Plain pizza
style_plain <- pizza_data %>%
```

```r
  filter(Style == 'Plain')
head(style_plain)
```

```
##                     Name location_lat location_lng      Date Year Price Style
## 1        Angelo’s Pizza     40.62325    -73.93792 2022-1014 2022  3.00 Plain
## 2        Ozone Pizzeria     40.68089    -73.84263 2022-1008 2022  3.00 Plain
## 3            Pino Pizza     40.60001    -73.99946 2022-1003 2022  2.75 Plain
## 4            La Rondine     40.71334    -73.82941 2022-0924 2022  3.25 Plain
## 5    Rony's Fresh Pizza     40.74825    -73.99235 2022-0915 2022  1.00 Plain
## 6 John & Joe's Pizzeria     40.85456    -73.86588 2022-0909 2022  3.50 Plain
```

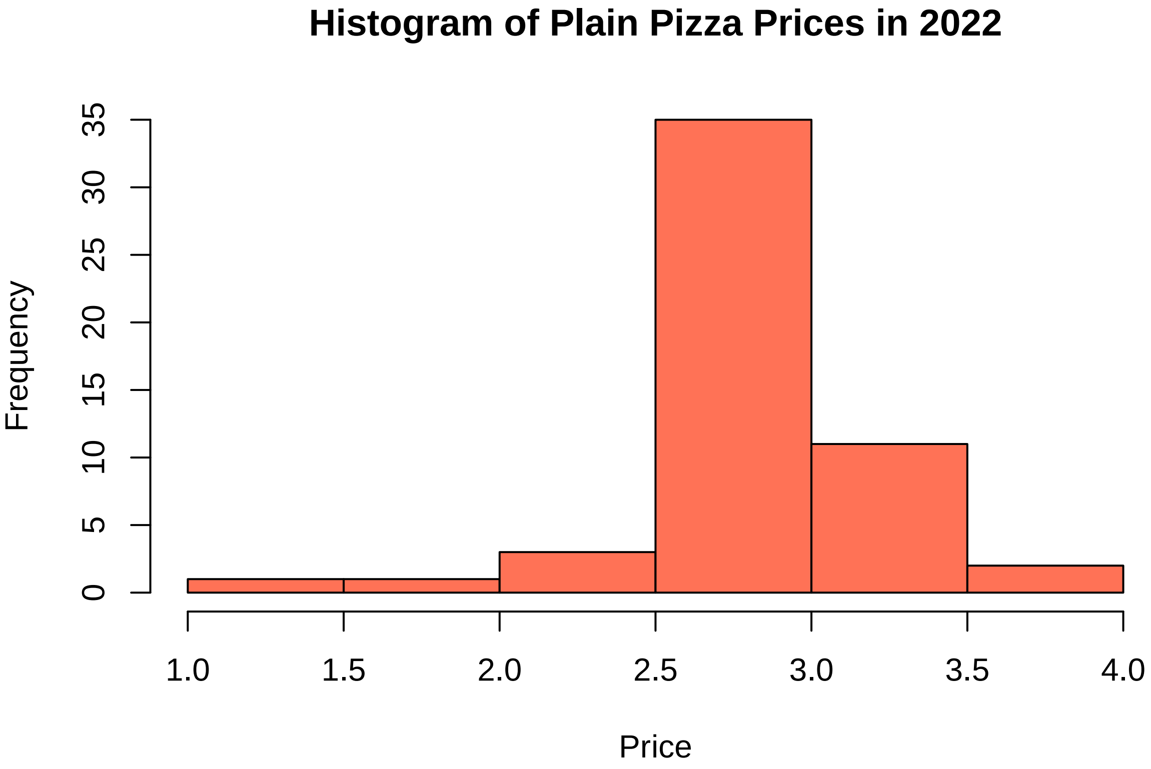
```r
# Create a boxplot to visualize price of plain pizza by year
boxplot(style_plain$Price ~ style_plain$Year,
        main = "Plain Pizza Price by Year",
        xlab = "Year",
        ylab = "Price",
        col = "tomato4",
        las = 1)
```



```r
# Subset the data for the year 2022
plain_pizza_2022 <- subset(style_plain, Year == 2022)
head(plain_pizza_2022)
```

```
##                     Name location_lat location_lng      Date Year Price Style
## 1        Angelo’s Pizza     40.62325    -73.93792 2022-1014 2022  3.00 Plain
## 2        Ozone Pizzeria     40.68089    -73.84263 2022-1008 2022  3.00 Plain
## 3            Pino Pizza     40.60001    -73.99946 2022-1003 2022  2.75 Plain
## 4            La Rondine     40.71334    -73.82941 2022-0924 2022  3.25 Plain
## 5    Rony's Fresh Pizza     40.74825    -73.99235 2022-0915 2022  1.00 Plain
## 6 John & Joe's Pizzeria     40.85456    -73.86588 2022-0909 2022  3.50 Plain
```

```r
# Plot a histogram of the Price column for the year 2022
hist(plain_pizza_2022$Price,
      main = "Histogram of Plain Pizza Prices in 2022",
      xlab = "Price",
      col = "coral1",  # Fill color for the bars
      border = "black")  # Border color for the bars
```



7. Create a subset named pizza pepperoni that only contains Pepperoni pizzas. Group the data by year and calculate the maximum, minimum, and average prices. Then, use a line chart to visualize these trends. You can either create three separate lines for maximum, minimum, and average prices, or combine them into a single chart that includes all three lines. Provide the R code (3 points)

```r
# Subset the data for the style pepperoni
pizza_pepperoni <- subset(pizza_data, Style == 'Pepperoni')
head(pizza_pepperoni)
```

```
##                                       Name location_lat location_lng       Date
## 13                              Pizza Chef     40.88559    -73.91038 2022-0729
## 25                       Valentine's Pizza     40.68753    -73.95443 2022-0513
## 34                          Marinara Pizza     40.72978    -73.98651 2022-0322
## 85                            Pronto Pizza     40.75824    -73.98063 2021-0922
## 86 Artichoke Basille’s Pizza - Times Square     40.75332    -73.98698 2021-0921
## 89                          Champion Pizza     40.73602    -73.99404 2021-0816
##    Year Price     Style
## 13 2022  4.35 Pepperoni
## 25 2022  4.50 Pepperoni
## 34 2022  4.90 Pepperoni
## 85 2021  4.50 Pepperoni
## 86 2021  6.53 Pepperoni
## 89 2021  4.25 Pepperoni
```

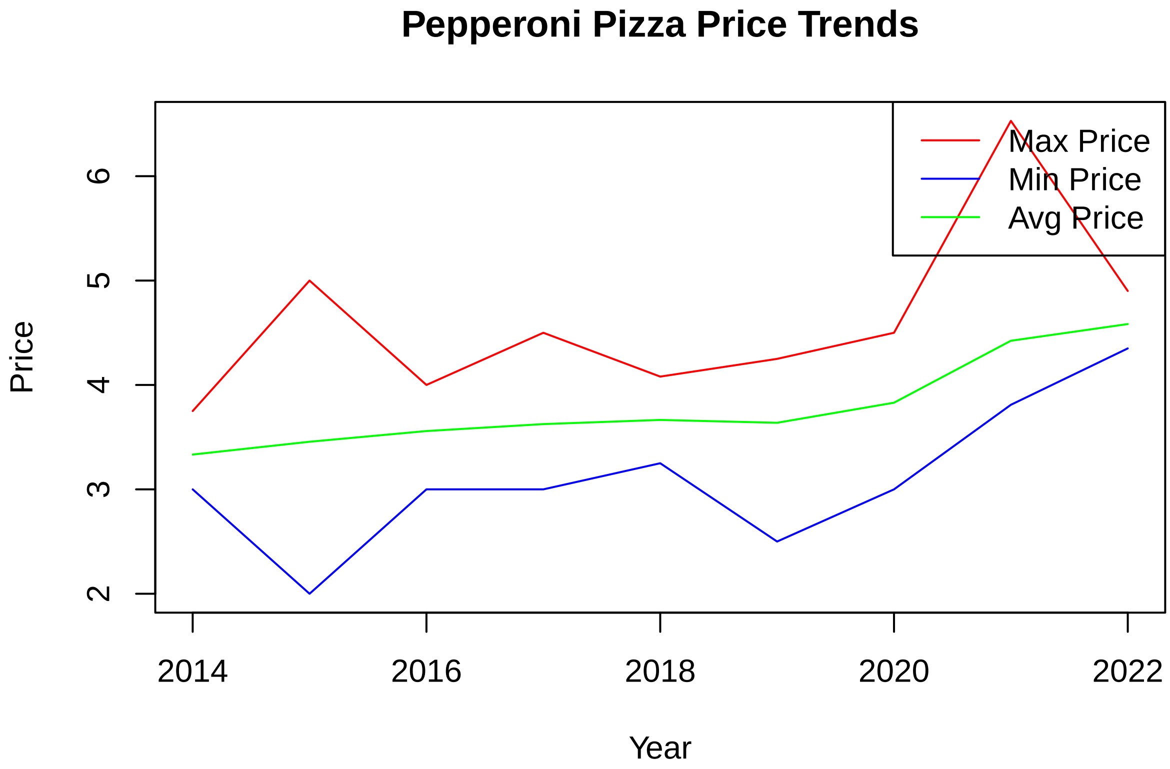
```r
# Group by year and calculate max, min, and average prices
pepperoni_stats <- pizza_pepperoni %>%
  group_by(Year) %>%
  summarise(
    max_price = max(Price, na.rm = TRUE),
    min_price = min(Price, na.rm = TRUE),
    avg_price = mean(Price, na.rm = TRUE)
  )

#View the summarized data
head(pepperoni_stats)
```

```
## # A tibble: 6 x 4
##    Year max_price min_price avg_price
##   <int>     <dbl>     <dbl>     <dbl>
## 1  2014      3.75      3          3.33
## 2  2015      5         2          3.46
## 3  2016      4         3          3.56
## 4  2017      4.5       3          3.62
## 5  2018      4.08      3.25       3.66
## 6  2019      4.25      2.5        3.64
```

```r
# Plot the trends
plot(pepperoni_stats$Year, pepperoni_stats$max_price, type = 'l', col = 'red',
     ylim = c(min(pepperoni_stats$min_price), max(pepperoni_stats$max_price)),
     xlab = 'Year', ylab = 'Price', main = 'Pepperoni Pizza Price Trends')
lines(pepperoni_stats$Year, pepperoni_stats$min_price, type = 'l', col = 'blue')
lines(pepperoni_stats$Year, pepperoni_stats$avg_price, type = 'l', col = 'green')

# Add a legend
legend('topright', legend = c('Max Price', 'Min Price', 'Avg Price'),
       col = c('red', 'blue', 'green'), lty = 1)
```


Pepperoni Pizza Price Trends
Max Price
Min Price
Avg Price
Price
6
5
4
3
2
2014
2016
2018
2020
2022
Year