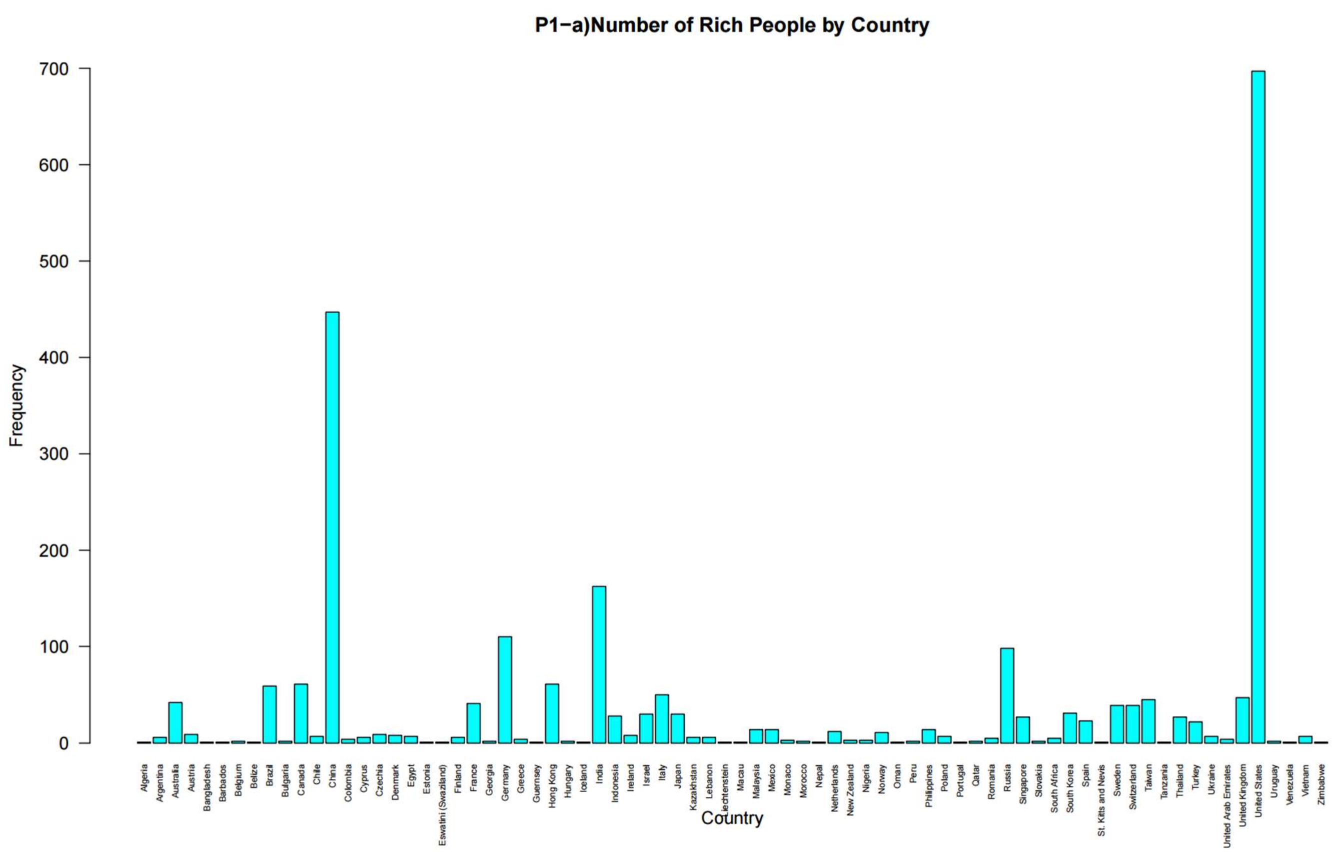
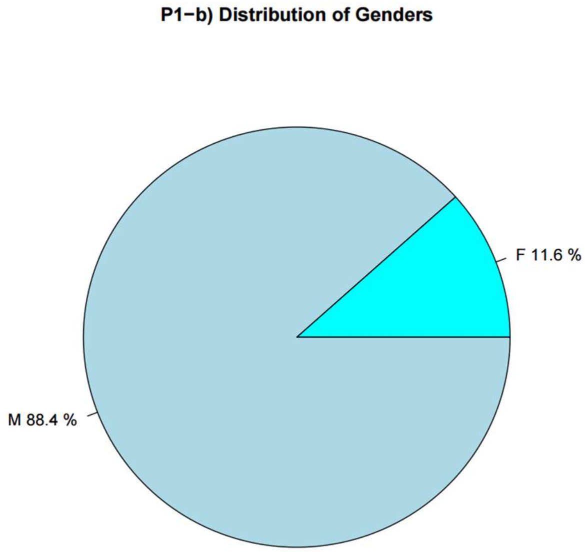
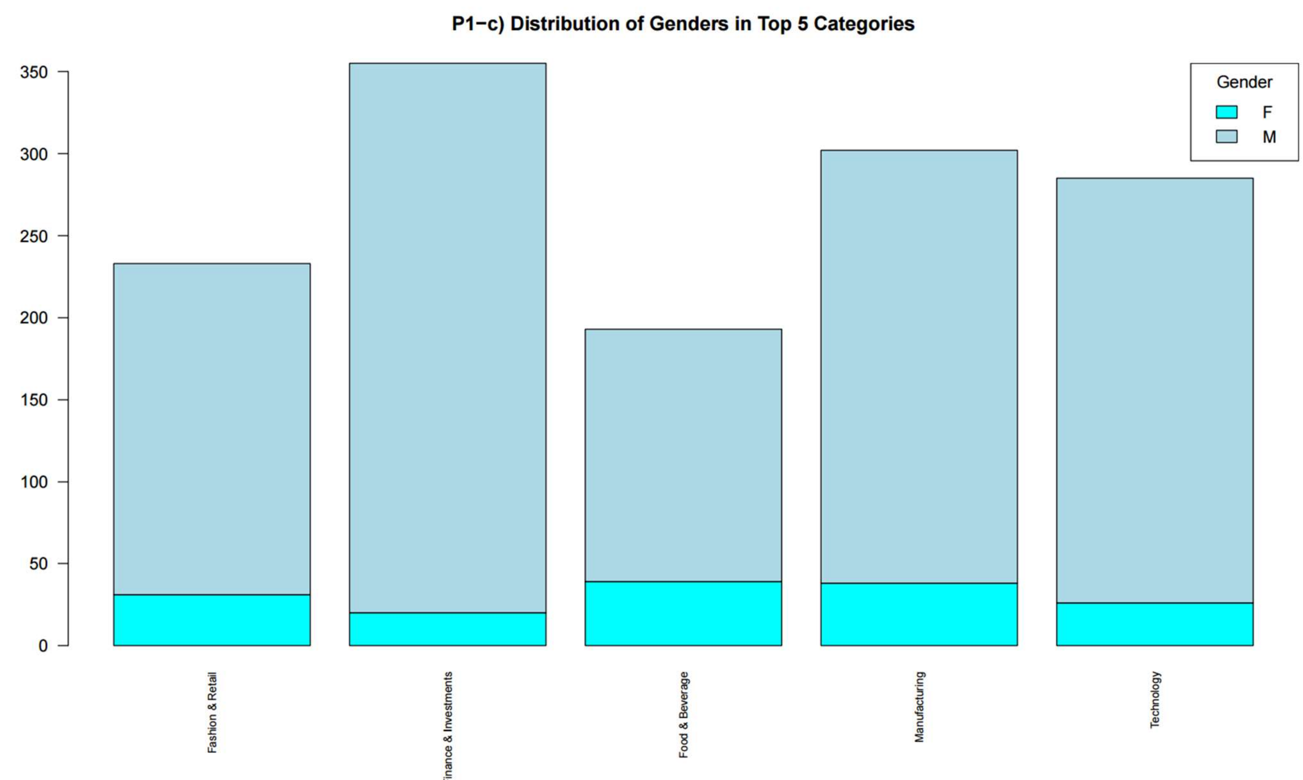
```r
#CS544 01
#Module 3 Assignment
#Laura Won

###Part 1
data_url <- "https://people.bu.edu/kalathur/datasets/forbes.csv"
forbes <- read.csv(data_url)

plot_colors <- c("cyan", "lightblue", "pink")
plot_margins <- c(8, 4, 4, 2)
plot_text_size <- 0.7

#a)
forbes_country_table<-table(forbes$country)
barplot(forbes_country_table,
         col = plot_colors[1],
         ylim = c(0, 700),  #number of frequency
         xlab = "Country",
         ylab = "Frequency",
         main = "P1-a)Number of Rich People by Country",
         las = 2,  #the country names vertical
         cex.names = 0.5)  #country name size can be changed

par(mar = plot_margins)
```



```
#b)
forbes_gender_table <- table(forbes$gender)
percentages <- round(100 * forbes_gender_table / sum(forbes_gender_table), 1)
pie_chart_labels <- paste(names(forbes_gender_table), percentages, "%", sep = " ")
pie(forbes_gender_table, labels = pie_chart_labels, col = plot_colors[1:2], main = "P1-b)
Distribution of Genders")
```




```
#c)
forbes_category_table <- table(forbes$category)
top5categories <- names(sort(forbes_category_table, decreasing = TRUE)[1:5])
top5data <- subset(forbes, category %in% top5categories)
par(mar = plot_margins)  # Reset plot margins
num_genders <- length(unique(top5data$gender))
gender_colors <- plot_colors[seq(1, num_genders)]
barplot(table(top5data$gender, top5data$category),
        main = "P1-c) Distribution of Genders in Top 5 Categories",
        col = gender_colors, las = 2, cex.names = plot_text_size)
legend("topright", legend = rownames(table(top5data$gender, top5data$category)), fill =
plot_colors[1:2], title = "Gender")
```



#d)
What inferences do you draw from the above plots?

United States has the largest number of rich people in the world. China, India, Germany, and Russia are in the top 5 for the number of rich people by country.

By gender ratio, females account for 11.6%, while males make up 88.4%. This indicates a significant difference between the two groups.

Considering the top 5 categories in the dataset, Finance & Investments, Manufacturing, Technology, Fashion & Retail, and Food & Beverage represent the top distributions among the world's richest people's businesses.

There is also a significant gender gap in these top categories, with a much smaller number of females compared to males.

###Part 2

```
us_quarters <- read.csv("https://people.bu.edu/kalathur/datasets/us_quarters.csv")
head(us_quarters)
```

```
          State DenverMint PhillyMint
1      Delaware     401424     373400
2  Pennsylvania     358332     349000
3    New Jersey     299028     363200
4       Georgia     488744     451188
5   Connecticut     657880     688744
6 Massachusetts     535184     628600
```

#a)

```
denver_highest <- us_quarters[which.max(us_quarters$DenverMint), ]
denver_lowest <- us_quarters[which.min(us_quarters$DenverMint), ]
philly_highest <- us_quarters[which.max(us_quarters$PhillyMint), ]
philly_lowest <- us_quarters[which.min(us_quarters$PhillyMint), ]

# Display the results
cat("The highest number of quarters produced by DenverMint is in", denver_highest$State,
"with", denver_highest$DenverMint, "quarters\n")
cat("The lowest number of quarters produced by DenverMint is in", denver_lowest$State,
"with", denver_lowest$DenverMint, "quarters\n")
cat("The highest number of quarters produced by PhillyMint is in", philly_highest$State,
"with", philly_highest$PhillyMint, "quarters\n")
cat("The lowest number of quarters produced by PhillyMint is in", philly_lowest$State,
"with", philly_lowest$PhillyMint, "quarters\n")
```

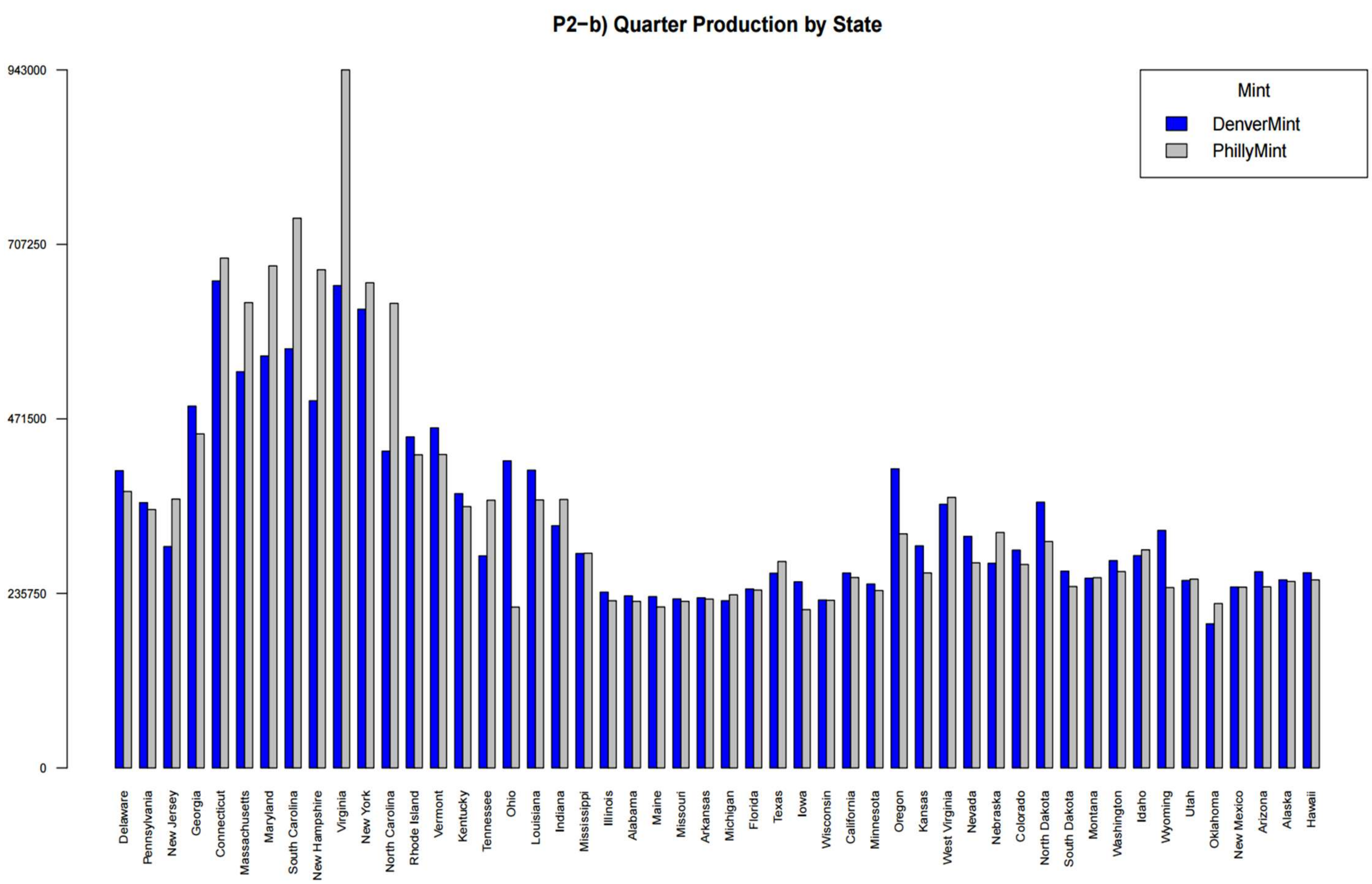
For which state were the highest number of quarters produced by each mint?
The highest number of quarters produced by **DenverMint** is in Connecticut with 657880 quarters

The highest number of quarters produced by **PhillyMint** is in Virginia with 943000 quarters.

For which state were the lowest number of quarters produced by each mint?
The lowest number of quarters produced by **DenverMint** is in Oklahoma with 194600 quarters.

The lowest number of quarters produced by **PhillyMint** is in Iowa with 213800 quarters.

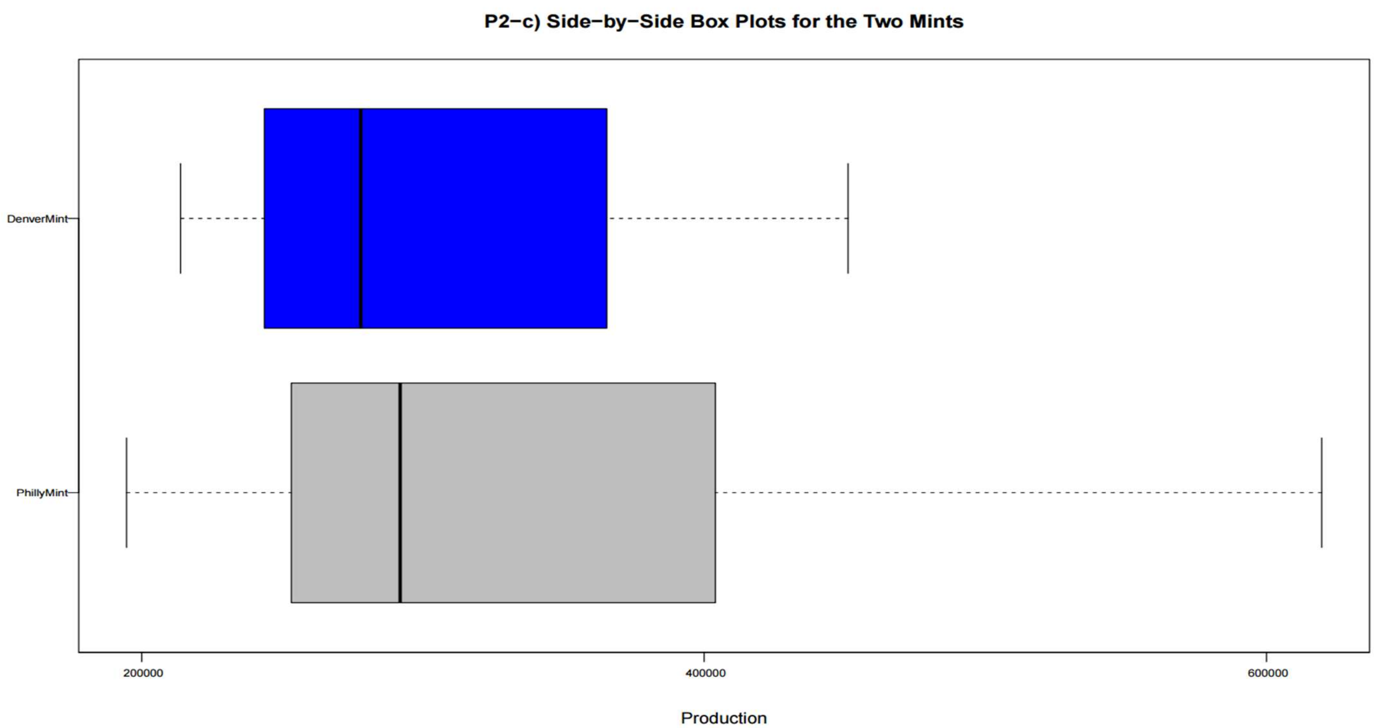
```
#b)
m <- as.matrix(us_quarters[, c("DenverMint", "PhillyMint")])
rownames(m) <- us_quarters$State
plot_colors <- c("blue", "gray")
par(mar = c(6, 3, 4, 1) + 0.1, cex.axis = 0.7)  # Set margins and axis label size

barplot_heights <- barplot(t(m), beside = TRUE, col = plot_colors, ylim = c(0,
ceiling(max(m))), main = "P2-b) Quarter Production by State", las = 2, cex.names = 0.7,
yaxt = 'n')
max_value <- ceiling(max(m))
axis(2, at = seq(0, max_value, length = 5), labels = format(seq(0, max_value, length =
5), scientific = FALSE), las = 1, cex.axis = 0.7)

legend("topright", legend = c("DenverMint", "PhillyMint"), fill = plot_colors, title =
"Mint")
```



```
#c)
denver_mint <- us_quarters$DenverMint
philly_mint <- us_quarters$PhillyMint
par(mar = c(5, 5, 4, 2) + 0.1, mgp = c(3, 0.5, 0))

boxplot(denver_mint, philly_mint, horizontal = TRUE, col = plot_colors[2:1], main = "P2-
c) Side-by-Side Box Plots for the Two Mints", xlab = "Production", las = 1, names =
c("PhillyMint", "DenverMint"), cex.axis = 0.7, outline = FALSE, xaxt = 'n')
axis(1, at = pretty(range(c(denver_mint, philly_mint))), labels =
format(pretty(range(c(denver_mint, philly_mint))), scientific = FALSE), cex.axis = 0.7)
```

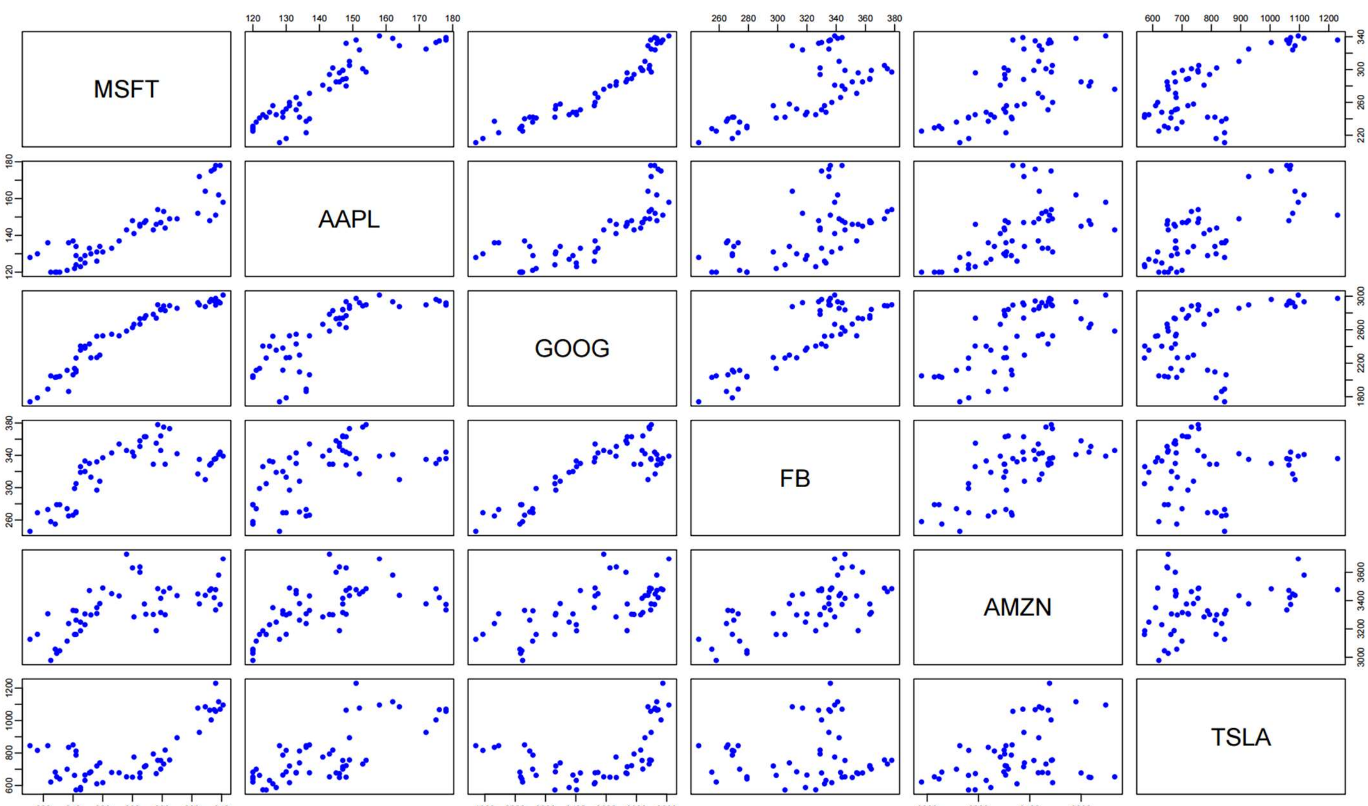
**Show the side-by-side box plots for the two mints. Write any two inferences for each of the box plots.**

Philly Mint shows a higher production level than Denver Mint.
These side-by-side box plots are useful for verifying high and low production levels for both Philly Mint and Denver Mint.
However, it is difficult to identify state-by-state production quantities from this chart.




```
#d)
summary_p <- summary(philly_mint)
philly_iqr <- summary_p[5] - summary_p[2]
philly_lower_fence <- summary_p[2] - 1.5 * philly_iqr
philly_upper_fence <- summary_p[5] + 1.5 * philly_iqr

summary_d <- summary(denver_mint)
denver_iqr <- summary_d[5] - summary_d[2]
denver_lower_fence <- summary_d[2] - 1.5 * denver_iqr
denver_upper_fence <- summary_d[5] + 1.5 * denver_iqr

philly_lower_outliers <- us_quarters$State[philly_mint < philly_lower_fence]
philly_upper_outliers <- us_quarters$State[philly_mint > philly_upper_fence]

denver_lower_outliers <- us_quarters$State[denver_mint < denver_lower_fence]
denver_upper_outliers <- us_quarters$State[denver_mint > denver_upper_fence]
```

```
cat("Philly Mint lower fence outliers (below", philly_lower_fence, "): ",
paste(philly_lower_outliers, collapse = ", "), "\n")
Philly Mint lower fence outliers (below 62100 ):   none
cat("Philly Mint upper fence outliers (above", philly_upper_fence, "): ",
paste(philly_upper_outliers, collapse = ", "), "\n")
Philly Mint upper fence outliers (above 546500 ):  Connecticut, Massachusetts, Maryland,
South Carolina, New Hampshire, Virginia, New York, North Carolina
cat("Denver Mint lower fence outliers (below", denver_lower_fence, "): ",
paste(denver_lower_outliers, collapse = ", "), "\n")
Denver Mint lower fence outliers (below 28173.5 ):   none
cat("Denver Mint upper fence outliers (above", denver_upper_fence, "): ",
paste(denver_upper_outliers, collapse = ", "), "\n")
Denver Mint upper fence outliers (above 628777.5 ):  Connecticut, Virginia

###Part 3
stocks <- read.csv("https://people.bu.edu/kalathur/datasets/stocks.csv")
head(stocks)
        Date MSFT AAPL GOOG  FB AMZN TSLA
1 2021-01-01  216  130 1787 269 3162  816
2 2021-01-08  211  128 1740 246 3127  845
3 2021-01-15  223  136 1891 273 3307  845
4 2021-01-22  237  136 1863 265 3238  835
5 2021-01-29  240  137 2062 266 3331  850
6 2021-02-05  242  134 2096 270 3262  812

#a)
pairs(stocks[, -1], pch = 16, col = "blue")
```




```
#b)
cm <- round(cor(stocks[, -1]), 2)
cm
     MSFT AAPL GOOG   FB AMZN TSLA
MSFT 1.00 0.90 0.95 0.68 0.64 0.71
AAPL 0.90 1.00 0.79 0.54 0.59 0.73
GOOG 0.95 0.79 1.00 0.85 0.67 0.47
FB   0.68 0.54 0.85 1.00 0.66 0.05
AMZN 0.64 0.59 0.67 0.66 1.00 0.34
TSLA 0.71 0.73 0.47 0.05 0.34 1.00
```

#c)
Provide at least 4 interpretations of the results
1. Near perfect correlation (0.9 to 1.0) indicating that the stocks move very closely together: (MSFT & AAPL: 0.90), (MSFT & GOOG: 0.95)

2. Strong correlation (0.7 to 0.9) suggesting a strong tendency to move in the same direction: (GOOG & AAPL: 0.79), (FB & GOOG: 0.85), (TSLA & MSFT: 0.71), (TSLA & AAPL: 0.73)

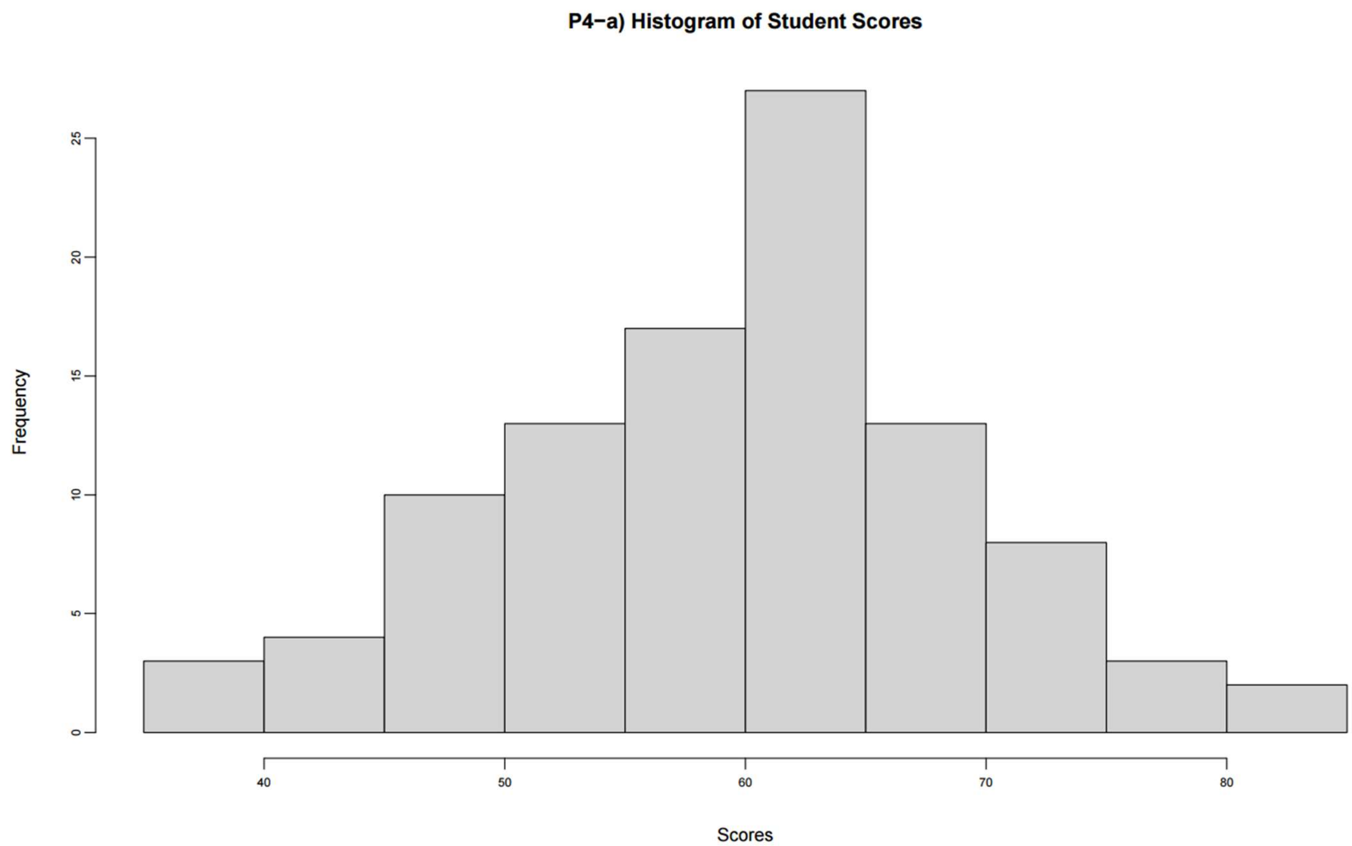
3. Moderate correlation (0.5 to 0.7) indicating significant but not very strong movement together: (MSFT & FB: 0.68), (MSFT & AMZN: 0.64), (AAPL & FB: 0.54), (AAPL & AMZN: 0.59), (GOOG & AMZN: 0.67), (FB & AMZN: 0.66)

4. Low correlation (0 to 0.5) indicating no linear relationship between the stocks: (GOOG & TSLA: 0.47), (FB & TSLA: 0.05), (AMZN & TSLA: 0.34)

#d)
```
for (stock in colnames(cm)) {
  cat("Top 3 correlated stocks for", stock, "\n")
  sorted_correlations <- sort(cm[stock, ], decreasing = TRUE)
  top_3_correlated <- sorted_correlations[2:4] # Exclude the stock itself
  print(top_3_correlated)
  cat("\n")
}
```

```
Top 3 correlated stocks for MSFT
GOOG AAPL TSLA
0.95 0.90 0.71


Top 3 correlated stocks for AAPL
MSFT GOOG TSLA
0.90 0.79 0.73


Top 3 correlated stocks for GOOG
MSFT   FB AAPL
0.95 0.85 0.79


Top 3 correlated stocks for FB
GOOG MSFT AMZN
0.85 0.68 0.66


Top 3 correlated stocks for AMZN
GOOG   FB MSFT
0.67 0.66 0.64


Top 3 correlated stocks for TSLA
AAPL MSFT GOOG
0.73 0.71 0.47
```

```
###Part 4
scores <- read.csv("https://people.bu.edu/kalathur/datasets/scores.csv")

#a)
score_history <- hist(scores$Score, main = "P4-a) Histogram of Student Scores", xlab =
"Scores")
score_counts <- score_history$counts
score_breaks <- score_history$breaks

specific_ranges <- function(score_counts, score_breaks) {
   for (i in 1:length(score_counts)) {
     if (score_breaks[i] >= 35 && score_breaks[i] <= 85) {
       cat(score_counts[i], "students in range (", score_breaks[i], ",", score_breaks[i +
1], "]\n", sep = "")
     }
   }
 }
specific_ranges(score_counts, score_breaks)
3students in range (35,40]
4students in range (40,45]
10students in range (45,50]
13students in range (50,55]
17students in range (55,60]
27students in range (60,65]
13students in range (65,70]
8students in range (70,75]
3students in range (75,80]
2students in range (80,85]
```



```
#b)
custom_breaks <- c(30, 50, 70, 90)
hist_grades <- hist(scores$Score, breaks = custom_breaks, main = "P4-b) Histogram of
Student Grades", xlab = "Scores")
grade_counts <- hist_grades$counts
grade_breaks <- hist_grades$breaks

print_grade_ranges <- function(grade_counts, grade_breaks) {
   grade_labels <- c("C grade", "B grade", "A grade")
   for (i in 1:length(grade_counts)) {
     cat(grade_counts[i], "students in ", grade_labels[i], "range (", grade_breaks[i],
",", grade_breaks[i + 1], "]\n", sep = "")
   }
 }
print_grade_ranges(grade_counts, grade_breaks)
17students in C graderange (30,50]
70students in B graderange (50,70]
13students in A graderange (70,90]
```