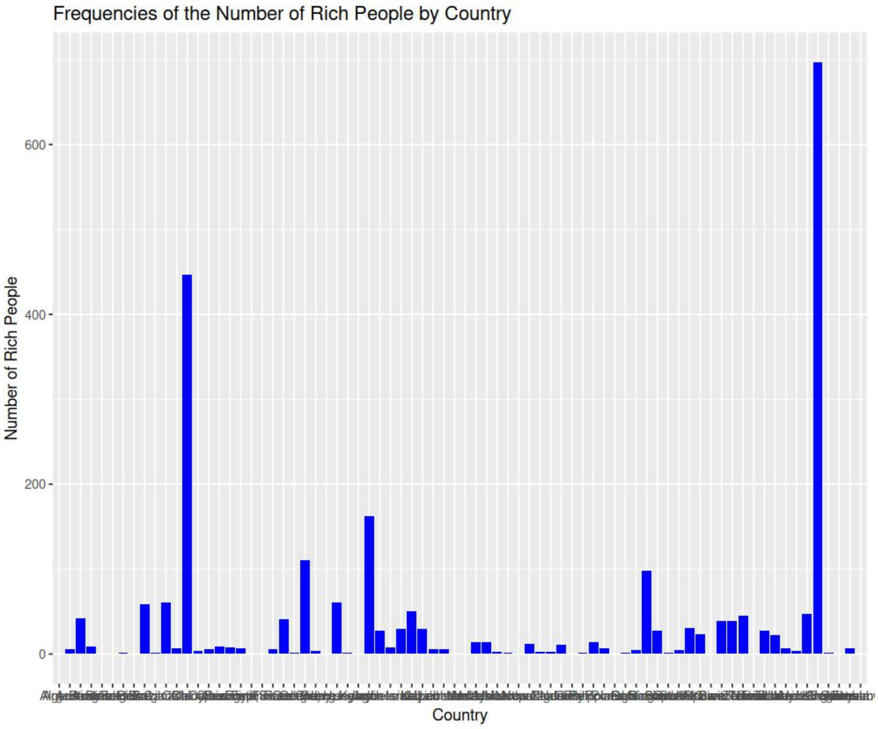
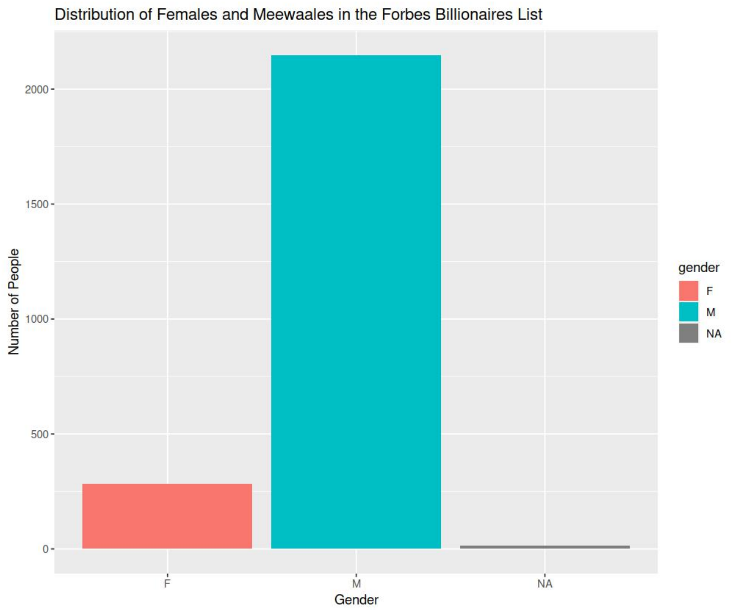
Part1

```r
library(ggplot2)

#part1
forbes <- read.csv("https://people.bu.edu/kalathur/datasets/forbes.csv")
print(forbes)
#a
ggplot(forbes,aes(x = country))+geom_bar(fill = "blue") + xlab("Country") +
  ylab("Number of Rich People") + ggtitle("Frequencies of the Number of Rich People by Country")

#b
ggplot(forbes, aes(x = gender, fill = gender)) + geom_bar() + xlab("Gender") +
  ylab("Number of People") + ggtitle("Distribution of Females and Meewaales in the Forbes Billionaires List")

#c
top5 <- as.data.frame(table(forbes$category))[1:5, ]
ggplot(forbes, aes(x=category, fill=gender)) +
  geom_bar(position="dodge") +
  labs(title="Distribution of Females and Males across Top 5 Categories in Forbes Billionaires List",
       x="Category", y="Count") +
  scale_x_discrete(limits=top5$Var1)
```

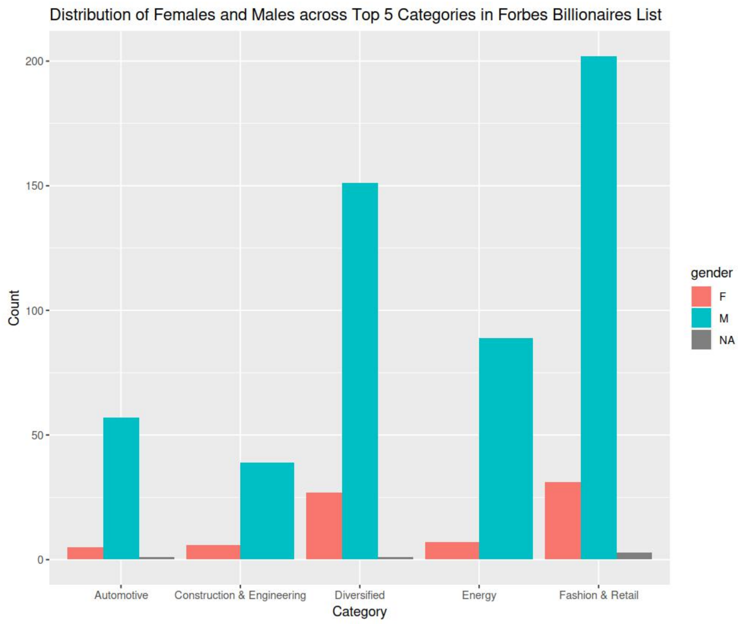
a



b



c



d

1.in first plot, we can know that United States have the highest number of rich people.

2.In second one, we can find that the number of males in Forbes is higher that the number of females

3.In the last one, in all the field, the number of males are all higher than the number of females.

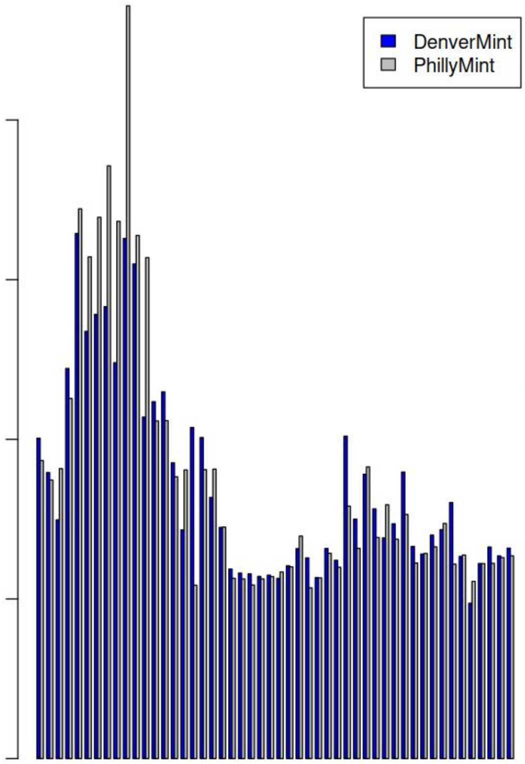
Part2

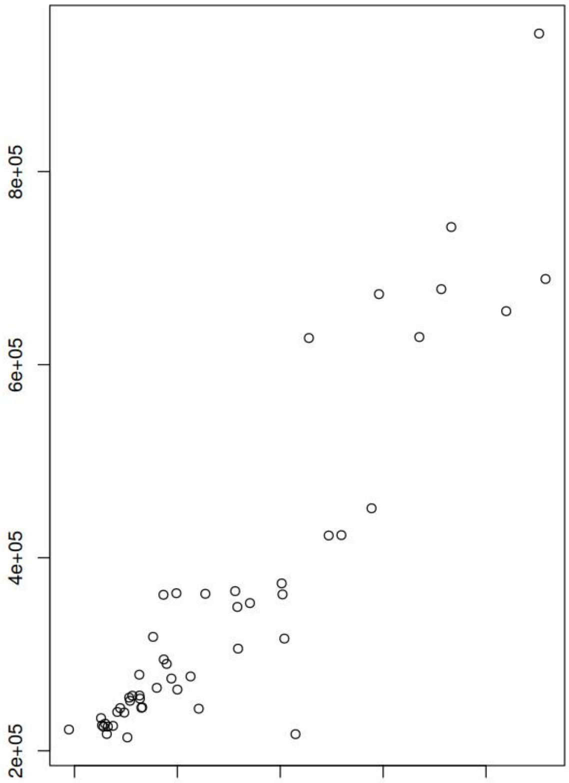
a

```
> us_quarters <- read.csv("https://people.bu.edu/kalathur/datasets/us_quarters.csv")
> us_quarters$State[which.max(us_quarters$DenverMint)]
[1] "Connecticut"
> us_quarters$State[which.max(us_quarters$PhillyMint)]
[1] "Virginia"
> us_quarters$State[which.min(us_quarters$DenverMint)]
[1] "Oklahoma"
> us_quarters$State[which.min(us_quarters$PhillyMint)]
[1] "Iowa"
>
```

b

```
> par(mfrow=c(1,2),mar = c(1, 1, 1, 1))
> barplot(cbind(DenverMint, PhillyMint) ~ State, col = c('blue','grey'),data = us_quarters, beside = T, legend = T)
> plot(us_quarters$DenverMint, us_quarters$PhillyMint)
```

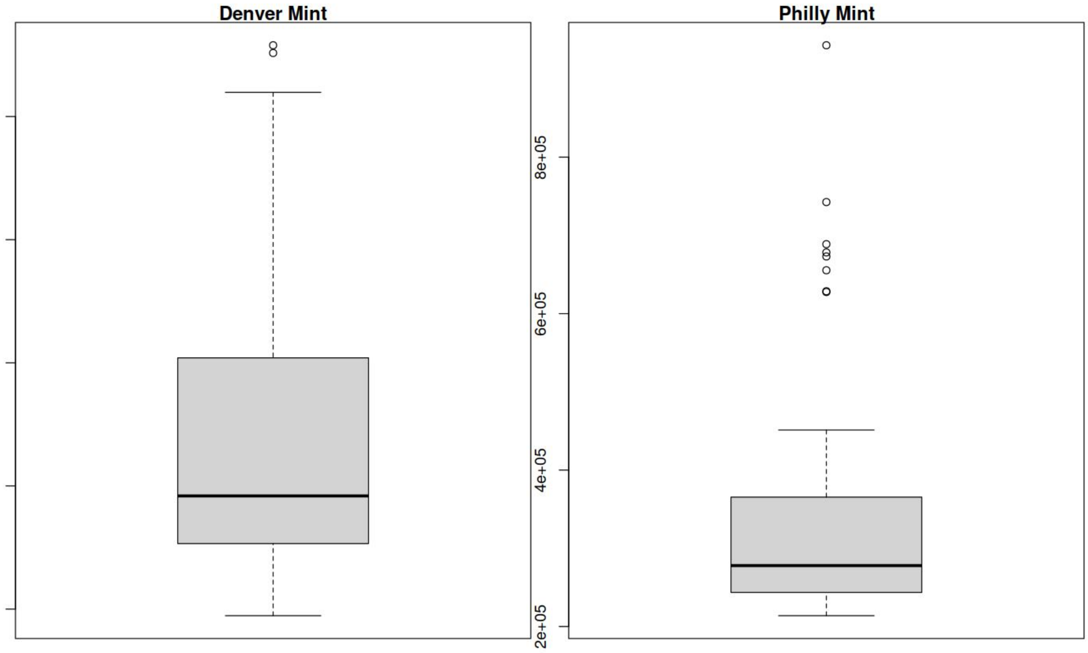




1.From barplot, I find that the two mint productions are positivelt correlated.
2.From scatterplot, I find that there are many outliner states in the dataset.

c

```
par(mfrow=c(1,2),mar = c(1, 1, 1, 1))
boxplot(us_quarters$DenverMint, main="Denver Mint", ylab="Quarters (in thousands)")
boxplot(us_quarters$PhillyMint, main="Philly Mint", ylab="Quarters (in thousands)")
```



1.In Denver Mint, its max value is much higher than the min value, but in Philly Mint, its max value is little higher than its min value.
2.For Philly Mint, it has many outliers.

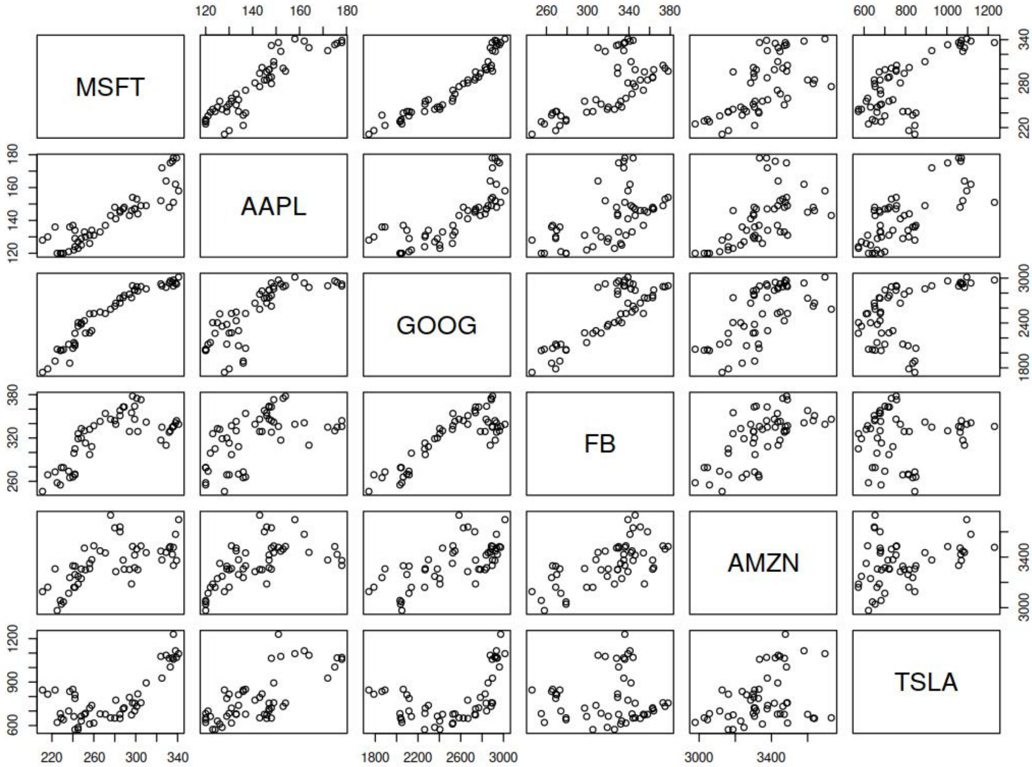
d

```
> f_den = fivenum(us_quarters$DenverMint)
> us_quarters$State[c(which(us_quarters$DenverMint > (f_den[4]+1.5*(f_den[4]-f_den[2]))),
+                     which(us_quarters$DenverMint < (f_den[2]-1.5*(f_den[4]-f_den[2]))))]
[1] "Connecticut" "Virginia"
>
>
> f_ph = fivenum(us_quarters$PhillyMint)
> us_quarters$State[c(which(us_quarters$PhillyMint > (f_ph[4]+1.5*(f_ph[4]-f_ph[2]))),
+                     which(us_quarters$PhillyMint < (f_ph[2]-1.5*(f_ph[4]-f_ph[2]))))]
[1] "Connecticut"    "Massachusetts"  "Maryland"       "South Carolina" "New Hampshire"  "Virginia"
[7] "New York"       "North Carolina"
```

Part3

a

```
> stocks <- read.csv("https://people.bu.edu/kalathur/datasets/stocks.csv")
> pairs(~ MSFT + AAPL + GOOG + FB + AMZN + TSLA, data = stocks)
```



b

```
> #b
> stocks1 <- subset(stocks, select = -c(Date))
> cm <- cor(stocks1)
> round(res, 2)
     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

1.Positive correlations indicate that two stocks tend to move in the same direction, while negative correlations indicate that two stocks tend to move in opposite directions.

2.A correlation value close to 1 indicates strong positive correlation, while a value close to -1 indicates strong negative correlation. A value close to 0 indicates weak or no correlation.

3.The diagonal elements of the correlation matrix are all 1, which indicates that each stock is perfectly positively correlated with itself.

4.The correlation matrix provides a summary of the linear relationships between each pair of stocks, but it does not capture non-linear relationships or other types of dependencies between stocks.

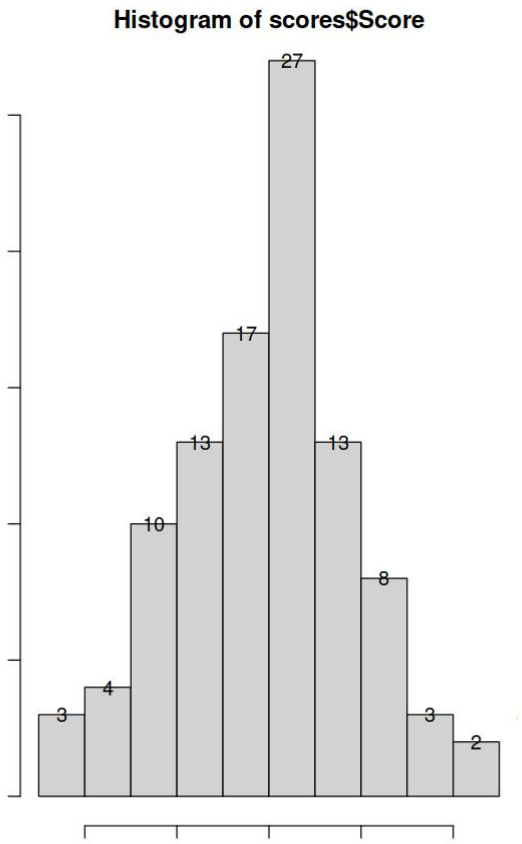
d

```
> #d
> n <- ncol(stocks)
> for (i in 1:n) {
+   stock <- colnames(stocks)[i+1]
+   corr <- cm[i, ]
+   top3 <- names(sort(corr, decreasing = TRUE))[2:(2 + 3)]
+   cat(sprintf("Top 3 for Stock %s\n%s\t%s\t%s\n%0.2f\t%0.2f\t%0.2f\n\n",
+               stock, top3[1], top3[2], top3[3], corr[top3[1]], corr[top3
[2]], corr[top3[3]]))
+ }
Top 3 for Stock MSFT
GOOG    AAPL    TSLA
0.95    0.90    0.71

Top 3 for Stock AAPL
MSFT    GOOG    TSLA
0.90    0.79    0.73

Top 3 for Stock GOOG
MSFT    FB      AAPL
0.95    0.85    0.79

Top 3 for Stock FB
GOOG    MSFT    AMZN
0.85    0.68    0.66

Top 3 for Stock AMZN
GOOG    FB      MSFT
0.67    0.66    0.64

Top 3 for Stock TSLA
AAPL    MSFT    GOOG
0.73    0.71    0.47
```

Part4

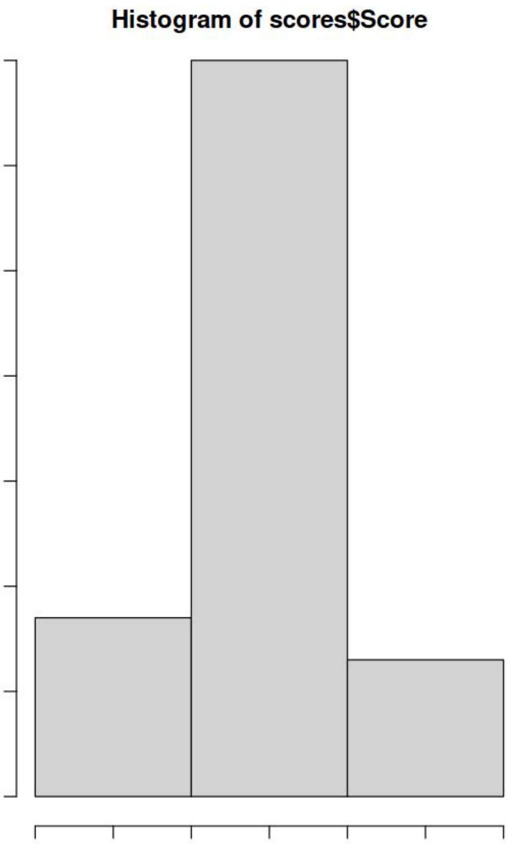
a

```
> scores <- read.csv("https://people.bu.edu/kalathur/datasets/scores.csv")
>
> #a
> h <- hist(scores$Score,breaks=8)
> text(h$breaks+2.5,h$counts,labels=h$counts)
```



b

```
> #b
> g <- hist(scores$Score,breaks=c(30,50,70,90))
> shc <- unlist(g[2])
> shb <- unlist(g[1])
> shg <- c("C","B","A")
> numIter = length(shc)
> for (i in 1:numIter) {
+   st <- sprintf("%d students in %s grade range (%d,%d]",shc[i],shg[i],shb
[i],shb[i+1])
+   print(st)
+ }
[1] "17 students in C grade range (30,50]"
[1] "70 students in B grade range (50,70]"
[1] "13 students in A grade range (70,90]"
```


Histogram of scores$Score