Stores and Stock Return Data Visualization

Khanin Sisaengsuwanchai

2022-07-11

1. Sales of stores dataset

```
# Download the sale data
data = read.csv("store.csv")
dim(data)
```

```
## [1] 2080   10
```

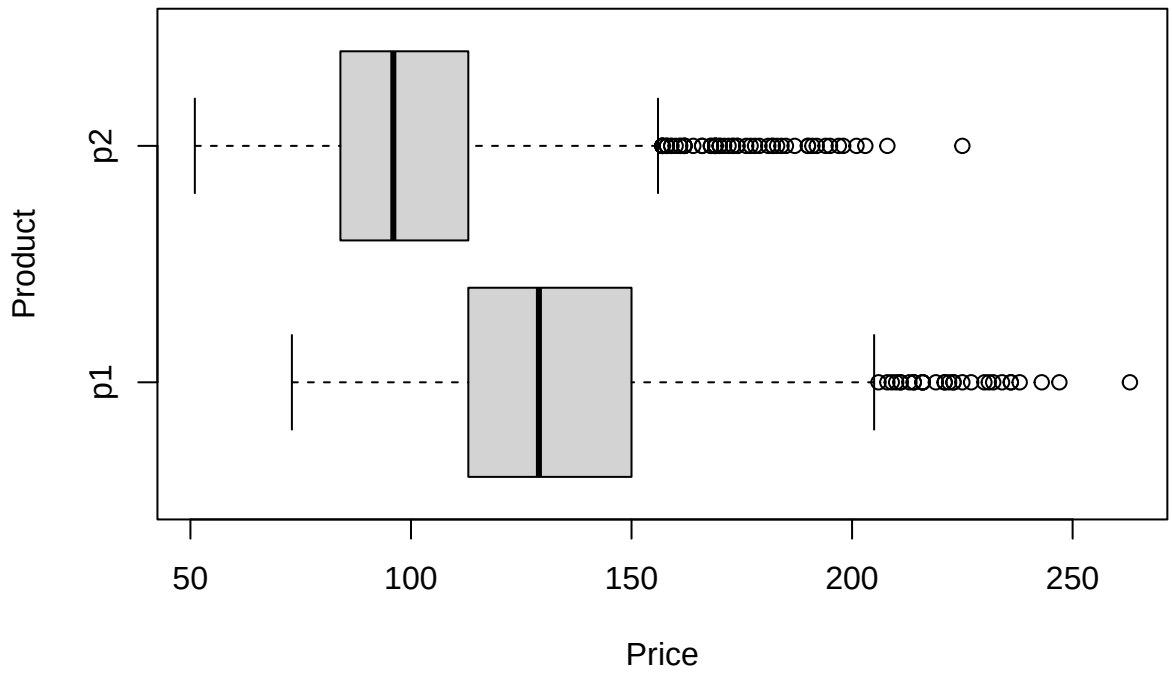
```
colnames(data)
```

```
##  [1] "storeID" "Year"    "Week"    "p1sales" "p2sales" "p1price" "p2price"
##  [8] "p1prom"  "p2prom"  "country"
```

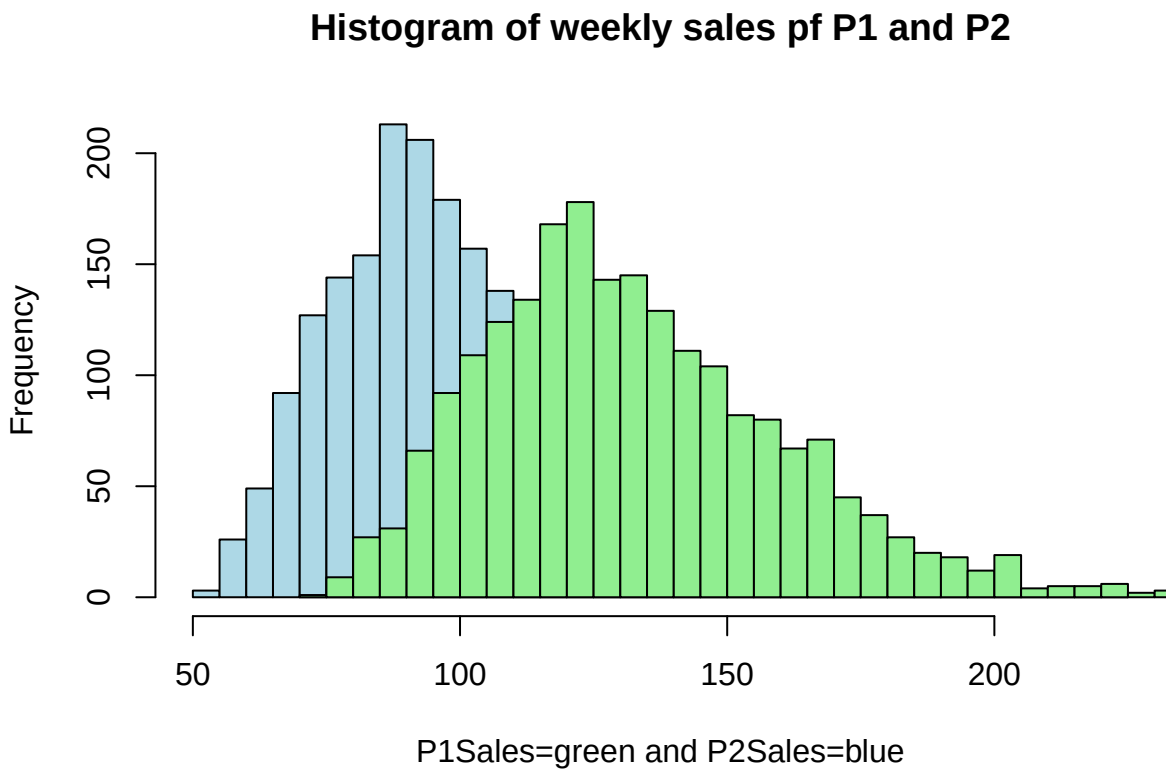
```
## Convert the types of data
data$storeID = as.factor(data$storeID)
data$country = as.factor(data$country)
data$p1prom = as.factor(data$p1prom)
data$p2prom = as.factor(data$p2prom)

# Summarize the data
summary(data)
```

```
##     storeID          Year          Week          p1sales       p2sales
##  101    : 104   Min.   :1.0   Min.   : 1.00   Min.   : 73   Min.   : 51.0
##  102    : 104   1st Qu.:1.0   1st Qu.:13.75   1st Qu.:113   1st Qu.: 84.0
##  103    : 104   Median :1.5   Median :26.50   Median :129   Median : 96.0
##  104    : 104   Mean   :1.5   Mean   :26.50   Mean   :133   Mean   :100.2
##  105    : 104   3rd Qu.:2.0   3rd Qu.:39.25   3rd Qu.:150   3rd Qu.:113.0
##  106    : 104   Max.   :2.0   Max.   :52.00   Max.   :263   Max.   :225.0
##  (Other):1456
##     p1price         p2price     p1prom   p2prom   country
##  Min.   :2.190   Min.   :2.29   0:1872   0:1792   AU:104
##  1st Qu.:2.290   1st Qu.:2.49   1: 208   1: 288   BR:208
##  Median :2.490   Median :2.59                     CN:208
##  Mean   :2.544   Mean   :2.70                     DE:520
##  3rd Qu.:2.790   3rd Qu.:2.99                     GB:312
##  Max.   :2.990   Max.   :3.19                     JP:416
##                                                   US:312
```

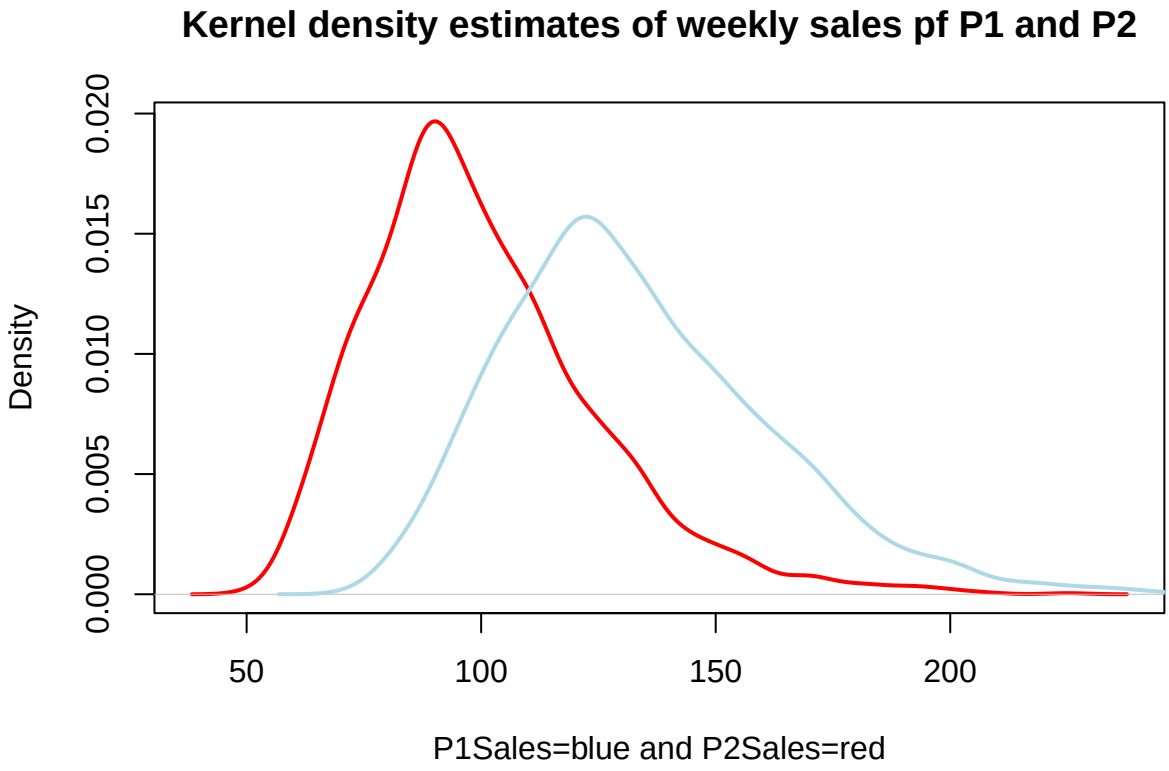
```r
# Compare the weekly sales of P1 (p1sales) to those of P2 by means of

## a) Two Boxplots on same chart
boxplot(list(data$p1sales,data$p2sales),horizontal=T, names=c("p1", "p2"),
        xlab="Price", ylab="Product")
```



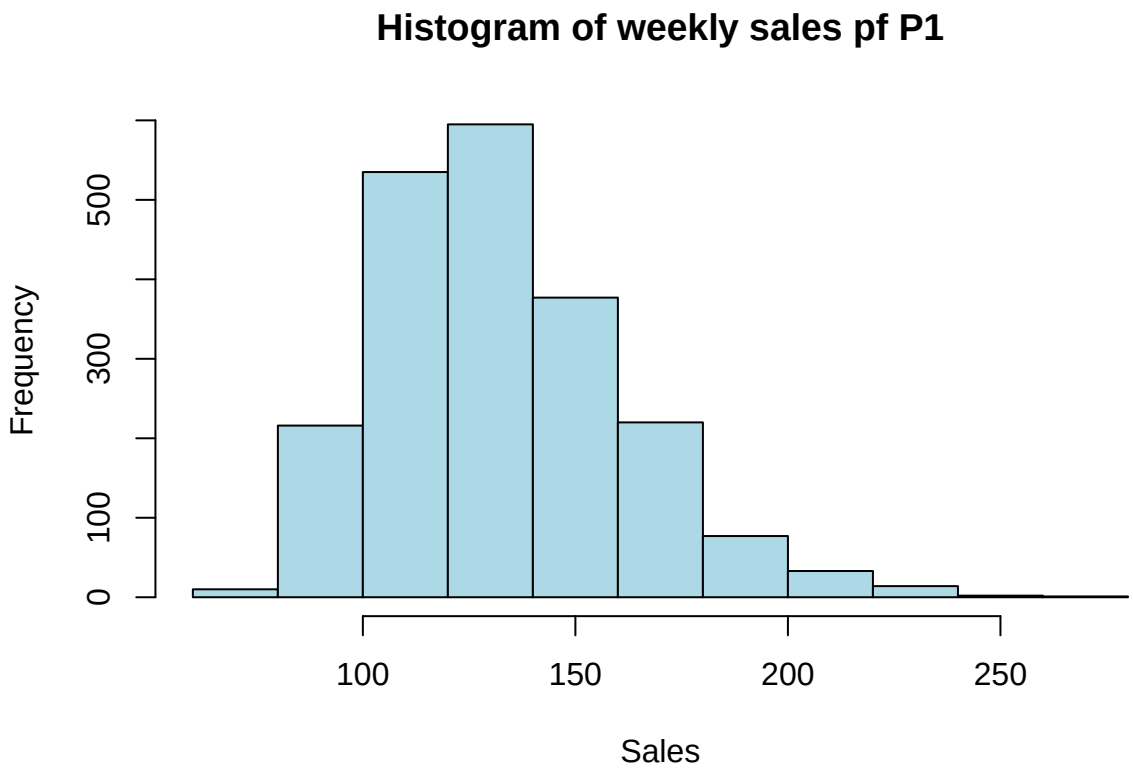

```r
## b) Two overlapping histograms (absolute frequencies) with breaks=30. Use hist(...,add=TRUE,...)
# for the second histogram.
hist(data$p2sales, breaks=30, col="light blue", xlab="P1Sales=green and P2Sales=blue",
     main="Histogram of weekly sales pf P1 and P2")
hist(data$p1sales, breaks=30, col="light green",add=TRUE)
```



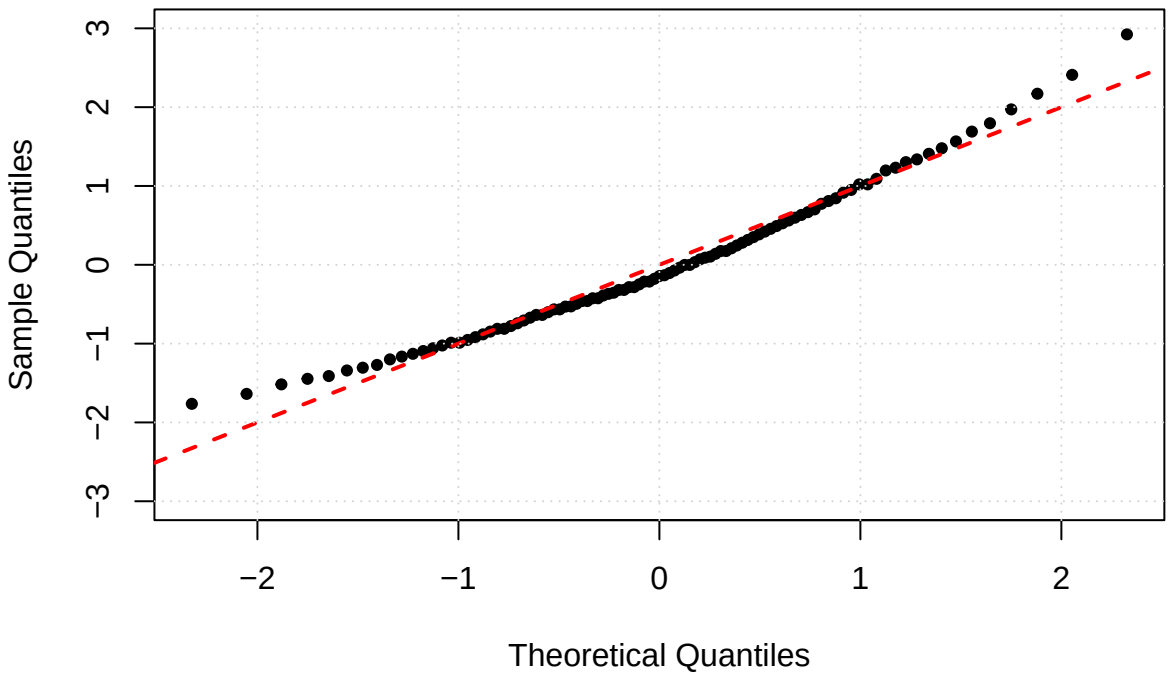

```
## c) Two overlapping Kernel density estimates of weekly sales Use lines(density( ),...) for the
# second kernel density.
plot(density(data$p2sales), col="red", lwd=2, xlab="P1Sales=blue and P2Sales=red",
     main="Kernel density estimates of weekly sales pf P1 and P2")
lines(density(data$p1sales), col="light blue", lwd=2)
```



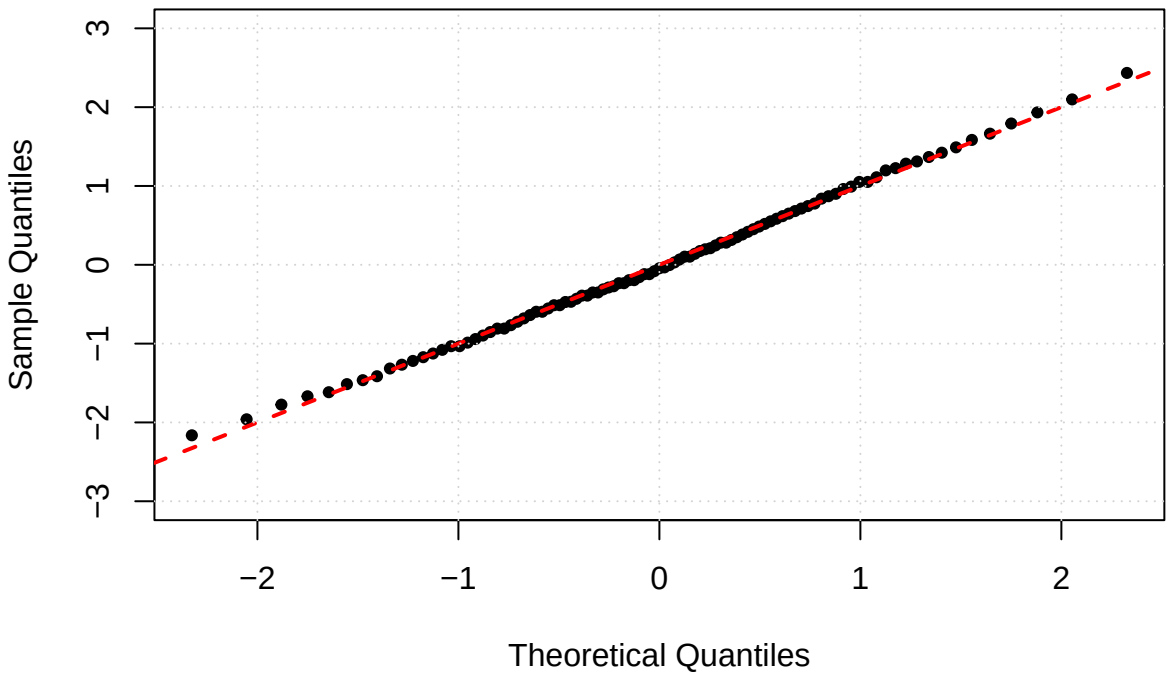

```
# Assess the normality of weekly sales of P1.

## a) Draw a histogram of weekly sales of P1.
hist(data$p1sales, col="light blue", xlab="Sales",
     main="Histogram of weekly sales pf P1")
```




```r
## b) Draw a normal quantile-quantile plot of weekly sales of P1.
a = seq(0, 1, 0.01)
x = scale(data$p1sales)
plot(qnorm(a), quantile(x, a), pch=19, cex=0.7, ylim=c(-3,3),
     ylab="Sample Quantiles", xlab="Theoretical Quantiles")
abline(0,1,lty=2,col="red", lwd=2)
grid()
```




```
## c) Find the natural log of p1sales (call it p1logsales). Draw a quantile plot of p1logsales.
a = seq(0, 1, 0.01)

p1logsales = log(data$p1sales) # Natural log of p1sales
x = scale(p1logsales)
plot(qnorm(a), quantile(x, a), pch=19, cex=0.7, ylim=c(-3,3),
     ylab="Sample Quantiles", xlab="Theoretical Quantiles")
abline(0,1,lty=2,col="red", lwd=2)
grid()
```




```
## After log transformation, the q-q plot happens to be more normal distribution.
```

2. Stock return dataset

```
# Draw a biplot and identify groups of stocks that are highly related (that is, those that have similar

library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# Download the data
stock = read.csv("stockreturns.csv")
summary(stock)
```

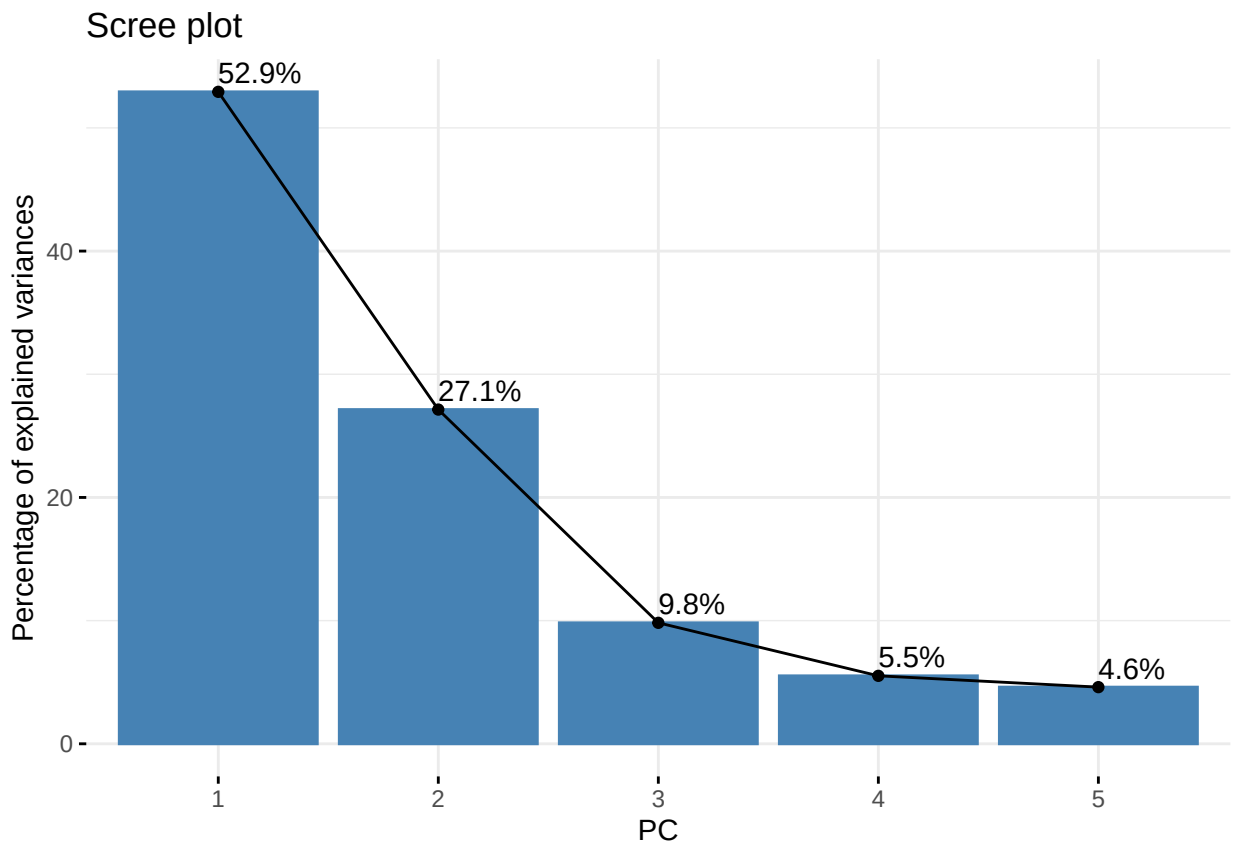
```
##       JPM                Citibank              WFargo
##  Min.   :-0.045867   Min.   :-0.0597924   Min.   :-0.0362141
##  1st Qu.:-0.013564   1st Qu.:-0.0132409   1st Qu.:-0.0080536
##  Median : 0.003363   Median : 0.0017339   Median : 0.0003354
##  Mean   : 0.001063   Mean   : 0.0006554   Mean   : 0.0016261
##  3rd Qu.: 0.016804   3rd Qu.: 0.0140293   3rd Qu.: 0.0100178
```

```
##  Max.    : 0.048480   Max.    : 0.0525266   Max.    : 0.0406957
##       Shell                Exxon
##  Min.    :-0.053948   Min.    :-0.063605
##  1st Qu.:-0.014470   1st Qu.:-0.012539
##  Median : 0.006335   Median : 0.005215
##  Mean    : 0.004049   Mean    : 0.004039
##  3rd Qu.: 0.022237   3rd Qu.: 0.021622
##  Max.    : 0.061994   Max.    : 0.078416
```

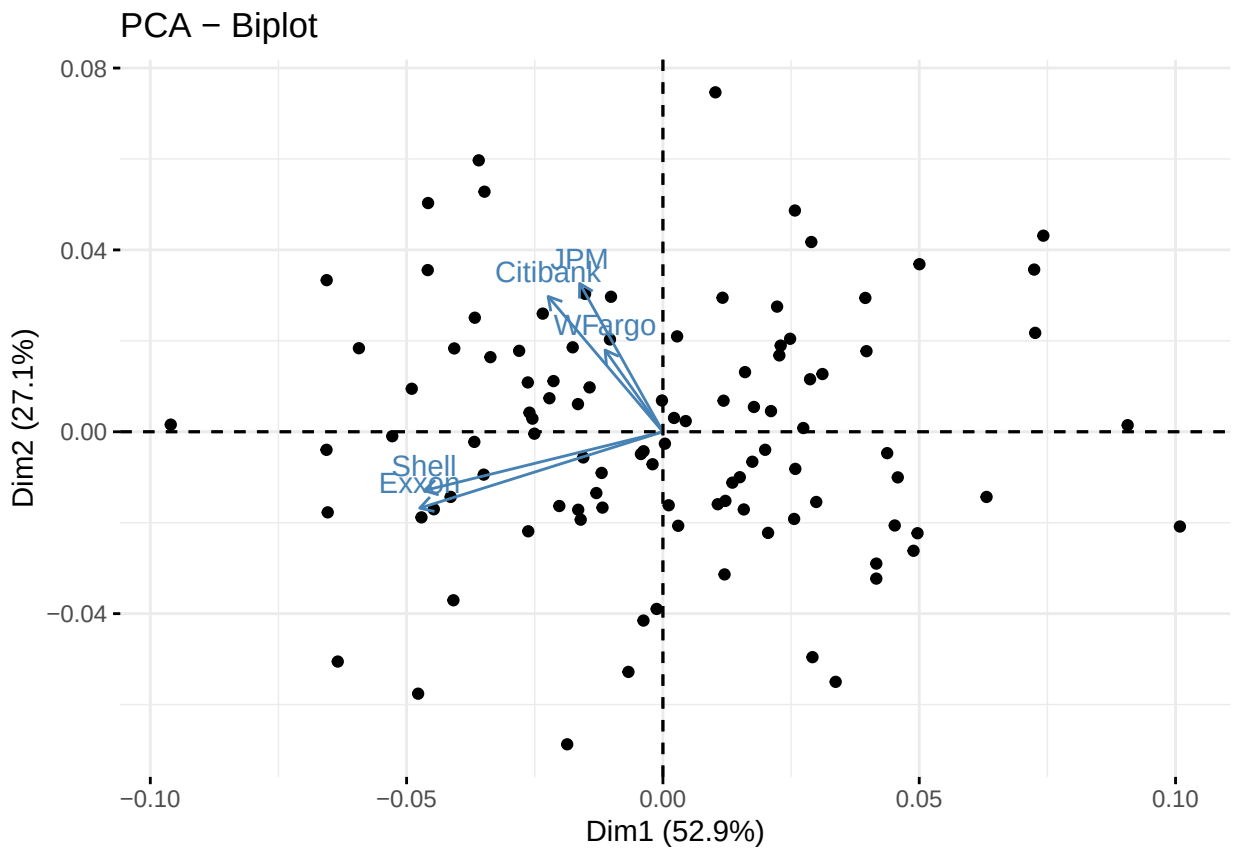
```
str(stock)
```

```
## 'data.frame':    103 obs. of  5 variables:
##  $ JPM     : num  0.01303 0.00849 -0.01792 0.02156 0.01082 ...
##  $ Citibank: num  -0.00784 0.01669 -0.00864 -0.00349 0.00372 ...
##  $ WFargo  : num  -0.00319 -0.00621 0.01004 0.01744 -0.01013 ...
##  $ Shell   : num  -0.0448 0.012 0 -0.0286 0.0292 ...
##  $ Exxon   : num  0.00522 0.01349 -0.00614 -0.00695 0.04098 ...
```

```
# Find PCA of this data
pca = prcomp(stock) # I do no scale because the data has been in the similar range.
fviz_screeplot(pca, addlabels=TRUE, choice="variance", xlab="PC")
```




```
# Draw a biplot
fviz_pca_biplot(pca, label="var")
```

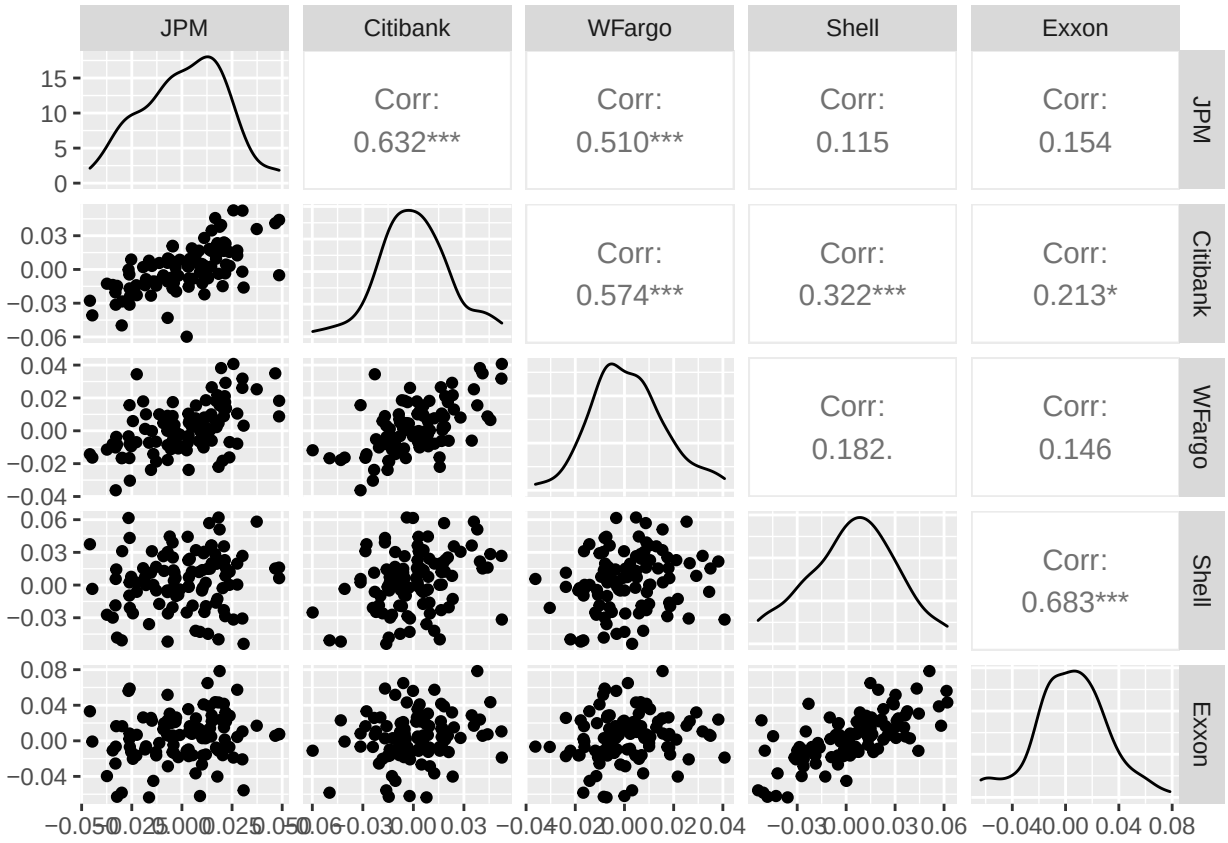



```
# According to the biplot, JPM, Citibank, and WFargo are highly related. Another
# group of stocks is Shell and Exxon, indicating by similar eigenvectors.
```

```
# Find the correlation matrix of the stocks' weekly rates of return.
# How does this correlation matrix confirms the result from the biplot?

library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(stock)
```




```
## According to the correlation matrix, JPM, Citibank, and WFargo are highly related
#  with correlation coefficients more than 0.5, while Shell and Exxon also have
#  high correlation values, which confirms the biplot.


# Correlation between transformed variables PCs and original variables Xs
pcaDat = get_pca(pca)
pcaDat$coord
```

```
##                 Dim.1        Dim.2         Dim.3        Dim.4        Dim.5
## JPM      -0.008240463  0.016555621 -0.0051953513  0.007914425 -0.001282903
## Citibank -0.011364238  0.015103596  0.0039762651 -0.004944953  0.006417901
## WFargo   -0.005725215  0.009122290  0.0005996382 -0.005935588 -0.008508065
## Shell    -0.023630398 -0.006565506  0.0102357410  0.003688399 -0.001618687
## Exxon    -0.024071832 -0.008522342 -0.0102893211 -0.002583880  0.001021854
```

```
# This correlation matrix tells us the eigenvectors used in biplot. Given the
# PC1 and PC2, we can clearly see that the eigenvectors of Shell and Exxon are
# very close. Additionally, JPM, Citivank, and WFargo are also closely related.
# This, as a result, reassure the result in the biplot.
```