Homework 2

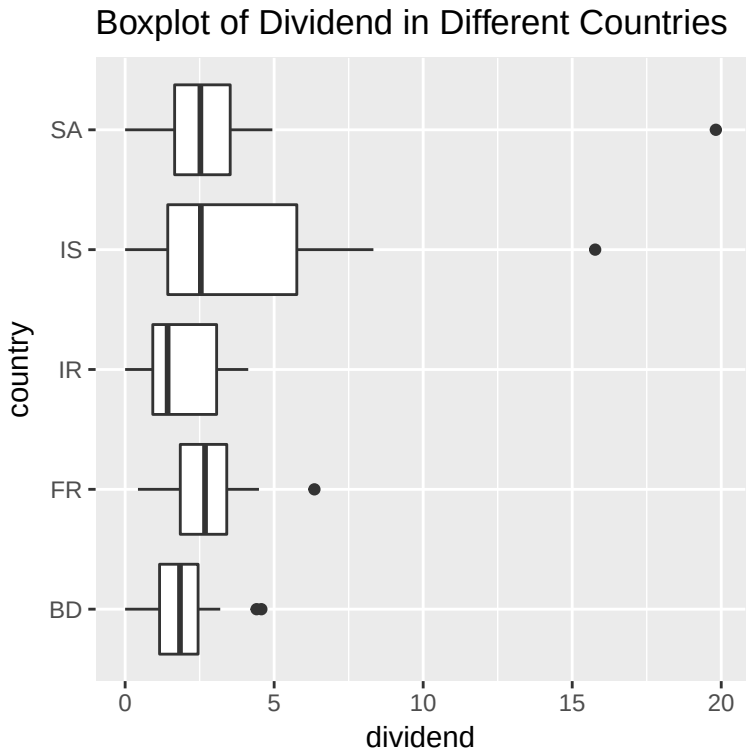
Shun-Lung Chang, Dilip Hiremath

```
library(magrittr) # pipeline operator
library(dplyr) # data manipulation
library(tidyr) # data manipulation
library(ggplot2) # plot
```

```
companies <- read.table("data/Companies.txt", sep = "\t", header = TRUE, stringsAsFactors = FALSE)
```

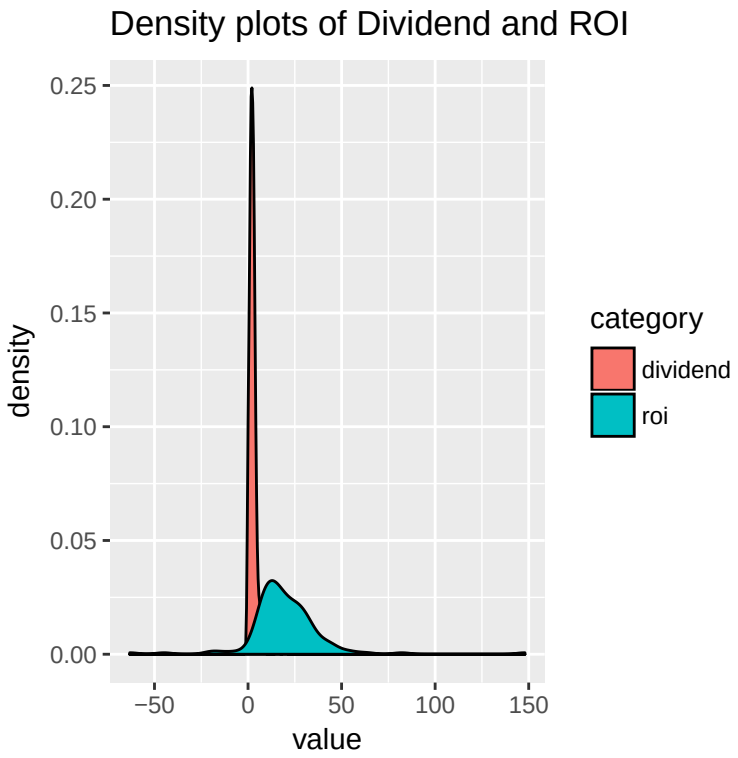
1. (2.5 points) Using a suitable graphical display, investigate whether there are any differences between countries in company dividends. In particular, look at central tendency and spread of the company dividend and provide a short summary of the distributional shape of company dividends for each country as well as the presence of outliers.

```
ggplot(data = companies) +
    geom_boxplot(aes(x = country, y = dividend), na.rm = TRUE) +
    coord_flip() +
    ggtitle("Boxplot of Dividend in Different Countries")
```



2. (2.5 points) Draw density plots for the variables dividend and roi. Put both plots together in one figure for easy comparison. Point out the interesting features of these distributions.

```
companies %>%
    select(dividend, roi) %>%
    gather(category) %>%
    ggplot() +
        geom_density(aes(x = value, fill = category)) +
        ggtitle("Density plots of Dividend and ROI")
```



3. You aim at comparing the empirical distribution for return on investment with a theoretical counterpart.

(a) (half a point) Compute mean and standard deviation for the variable roi and report them.

```
mean(companies$roi, na.rm = TRUE)
```

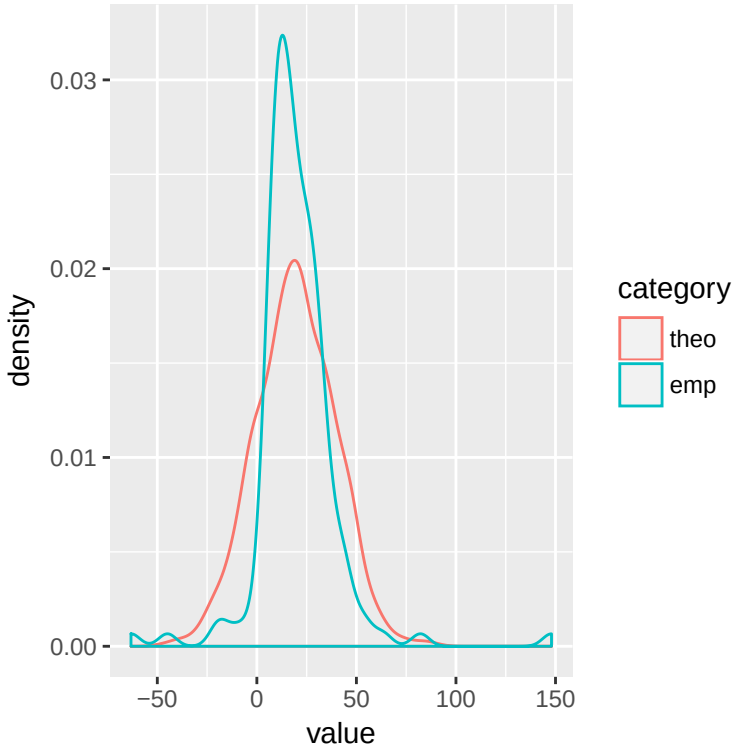
```
[1] 19.79081
```

```
sd(companies$roi, na.rm = TRUE)
```

```
[1] 20.02681
```

(b) (1 point) Use these statistics as parameters for a normal distribution. Plot the probability density function for this normal distribution and compare its shape with the shape of the corresponding density plot obtained in Question 2. Use a suitable range for the scores of this variable.

```
set.seed(42)

normal <- rnorm(n = 1000, mean(companies$roi, na.rm = TRUE), sd(companies$roi, na.rm = TRUE))
df_roi <- rbind(data.frame(category = "theo", value = normal),
      data.frame(category = "emp", value = companies$roi))

ggplot(data = df_roi) +
    geom_density(aes(x = value, color = category))
```



(c) (1 point) Do you find this normal distribution suitable for modelling the corresponding empirical distributions? Why or why not?

4. You aim at comparing the empirical distribution for dividend with a theoretical counterpart.

(a) (half a point) Compute mean and standard deviation for the variable dividend and report them.

```
mean(companies$dividend, na.rm = TRUE)
```

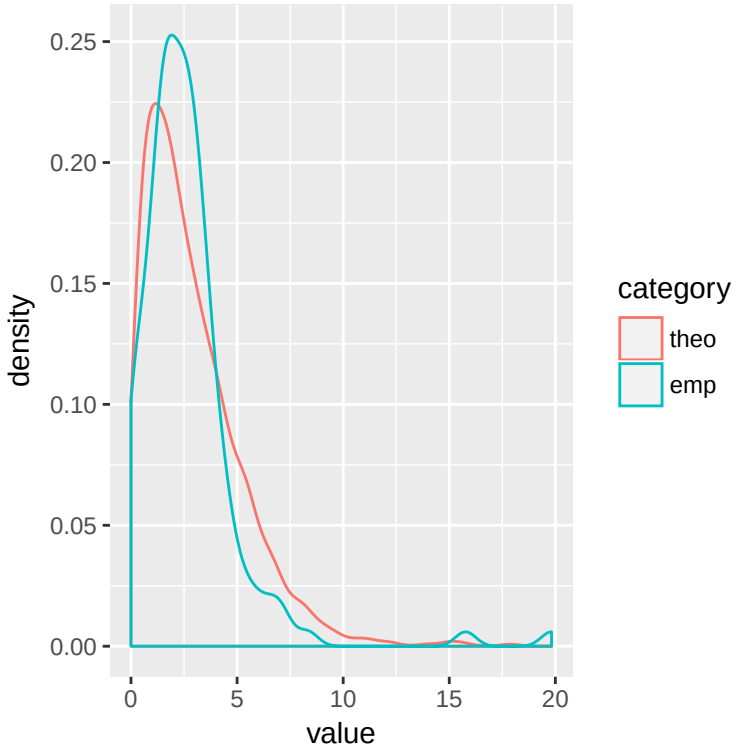
```
[1] 2.648841
```

```r
sd(companies$dividend, na.rm = TRUE)
```

```
[1] 2.459952
```

(b) (1.5 points) Due to the skewness of the empirical distribution you select a χ^2-distribution to model the data. The χ^2-distribution has one parameter, called degree of freedoms, ν. An integer number which is identical to the theoretical mean of the distribution. Round the computed mean to the nearest integer and use this as parameter for your χ^2- distribution. Plot the probability density function for this χ^2-distribution and compare its shape with the shape of the corresponding density plot obtained in Question 2. Use a suitable range for the scores of this variable.

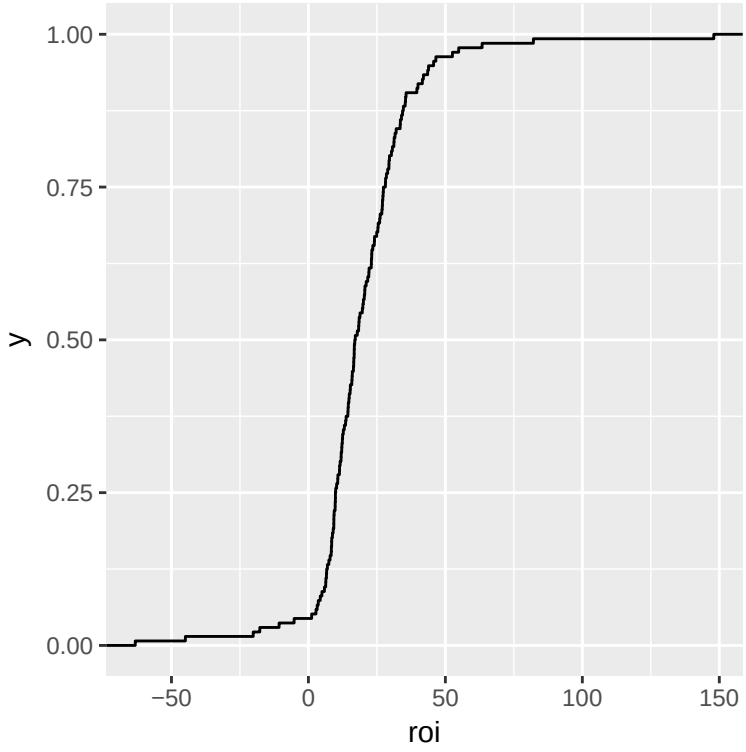
```r
set.seed(42)
chisq <- rchisq(1000, df = round(mean(companies$dividend, na.rm = TRUE)))

df_dividend <- rbind(data.frame(category = "theo", value = chisq),
      data.frame(category = "emp", value = companies$dividend))
ggplot(data = df_dividend) +
    geom_density(aes(x = value, color = category))
```



(c) (half a point) Do you find this χ^2-distribution suitable for modelling the corresponding empirical distributions? Why or why not?

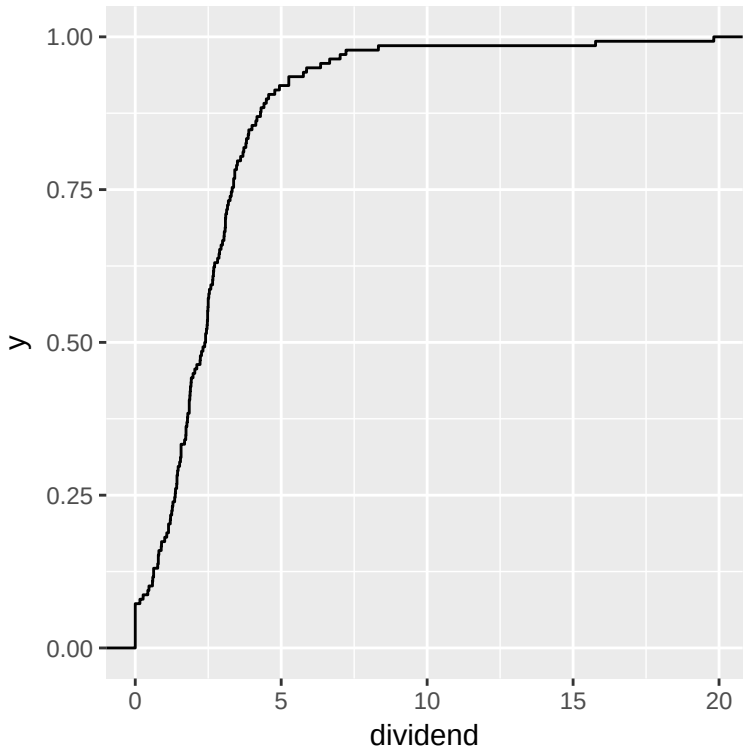
5. (2.5 points) Draw a plot of the empirical cumulative distribution function (ECDF) for the variable roi. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the ECDF plot?

```
ggplot(data = companies) +
    stat_ecdf(aes(x = roi))
```



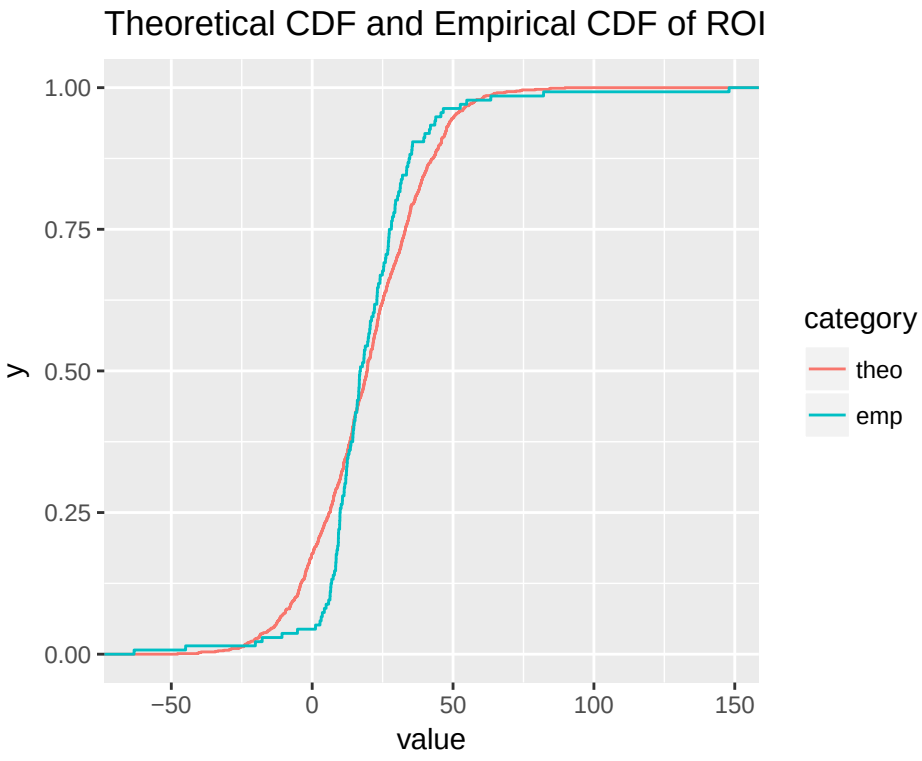
6. (2.5 points) Draw a plot of the empirical cumulative distribution function (ECDF) for the variable dividend. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the ECDF plot?

```
ggplot(data = companies) +
    stat_ecdf(aes(x = dividend))
```

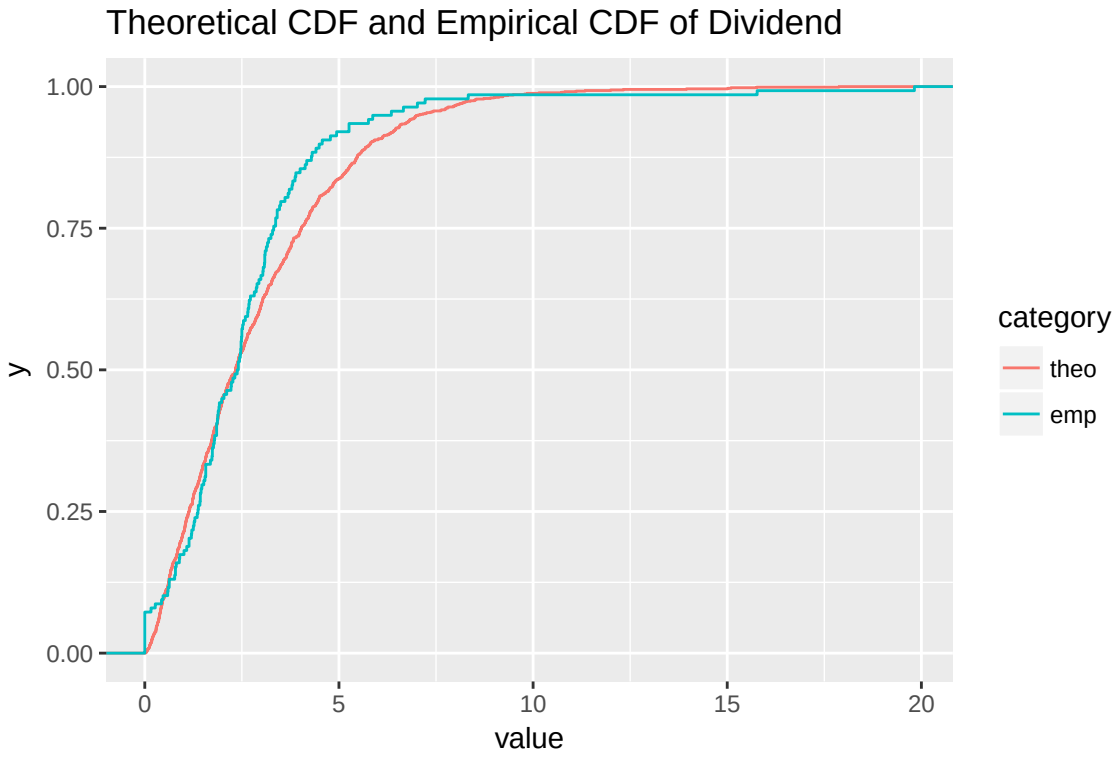


7. (2.5 points) Now, draw the corresponding cumulative distribution functions (CDF) for the densities derived in Question 3 and Question 4. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the CDF plot? Compare the CDFs with the corresponding ECDFs!

```
ggplot(df_roi) +
    stat_ecdf(aes(x = value, color = category)) +
    ggtitle("Theoretical CDF and Empirical CDF of ROI")
```




```
ggplot(df_dividend) +
    stat_ecdf(aes(x = value, color = category)) +
    ggtitle("Theoretical CDF and Empirical CDF of Dividend")
```

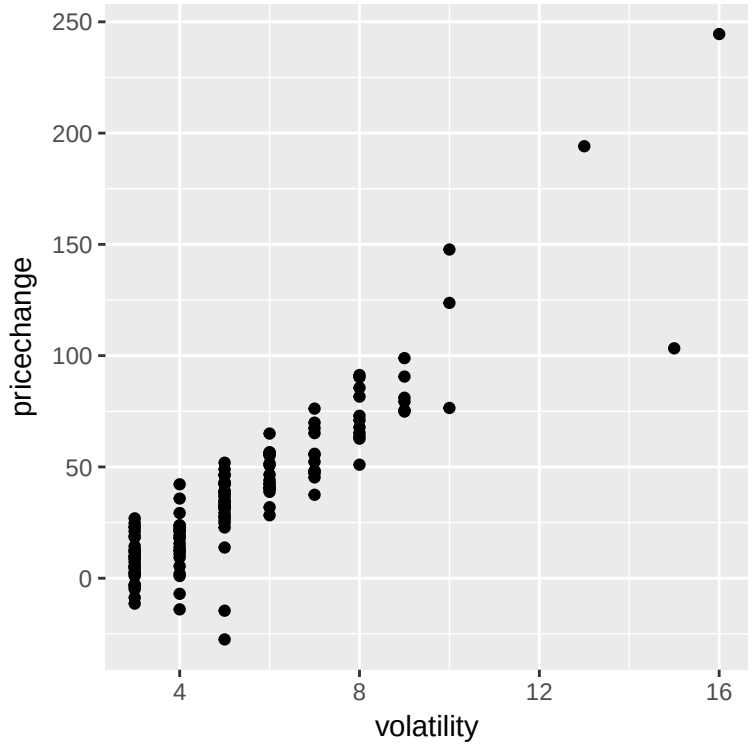


8. Now, you investigate whether there is any relationship between volatility and price change of a company's stock.

(a) (1 point) Would you expect any relationship? If yes, which one? If not, why not?

(b) (1 point) Draw a scatter plot of volatility against pricechange and comment on it.

```
ggplot(data = companies) +
    geom_point(aes(x = volatility, y = pricechange))
```



(c) (half a point) Calculate the Pearson correlation coefficient and comment on it.

```
cor(companies$pricechange, companies$volatility)
```

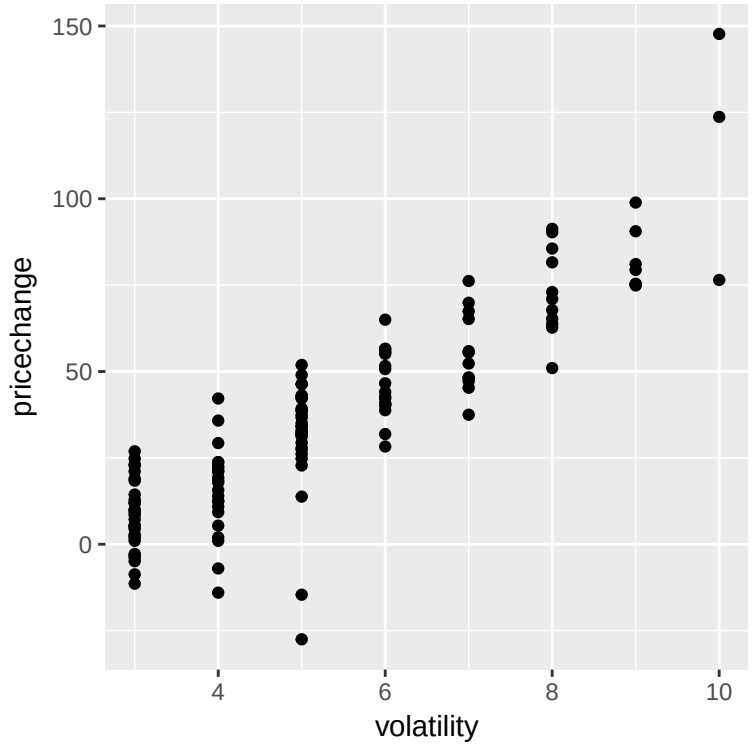
```
[1] 0.9002681
```

9. There seem to be (at least) three observations that fall a bit far from the other points; two having a high price change score and high volatility, the third one only having high volatility.

(a) (1 point) Draw the scatter plot leaving out these three points.

```
# remove points whose volatility is greater than 12
d <- companies[!companies$volatility > 12, ]
```

```
ggplot(data = d) +
    geom_point(aes(x = volatility, y = pricechange))
```



(b) (half a point) Is there now a stronger linear relationship?

(c) (1 point) Compute the Pearson correlation coefficient anew, ignoring these three companies. How has the correlation coefficient changed?

```
cor(d$pricechange, d$volatility)
```

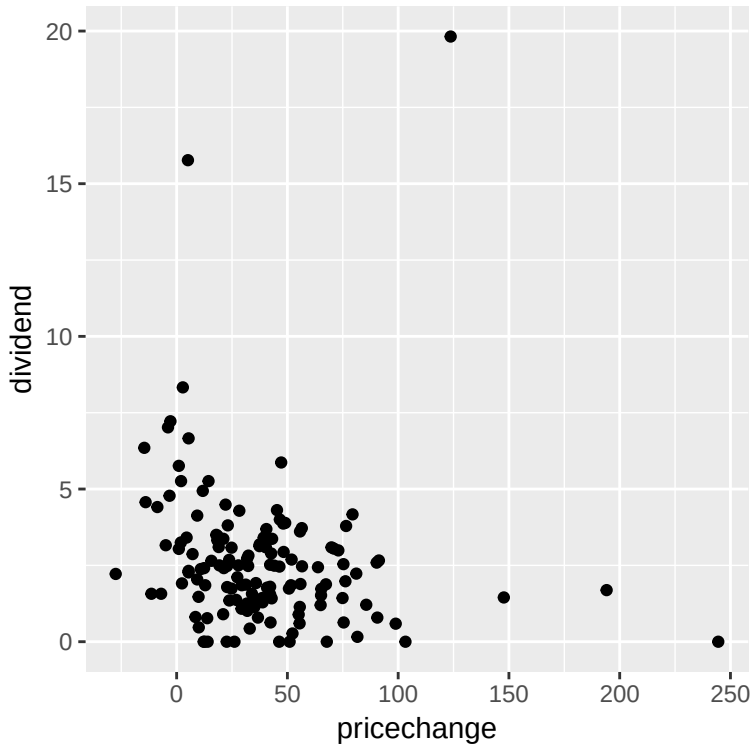
```
[1] 0.8861143
```

10. Now, you investigate whether there is any relationship between price change over the last 12 months and dividend.

(a) (half a point) Would you expect a relationship?

(b) (half a point) Draw a scatter plot of price change against dividend and comment on it.

```
ggplot(data = companies) +
    geom_point(aes(x = pricechange, y = dividend))
```



(c) (half a point) Calculate the Pearson correlation coefficient and comment on it.

```
cor(companies$pricechange, companies$dividend, use = "complete.obs")
```

```
[1] -0.130151
```

(d) (half a point) Calculate Kendall's τ and Spearman's Rank correlation and comment on them.

```
cor(companies$pricechange, companies$dividend, use = "complete.obs", method = "kendall")
```

```
[1] -0.1821568
```

```
cor(companies$pricechange, companies$dividend, use = "complete.obs", method = "spearman")
```

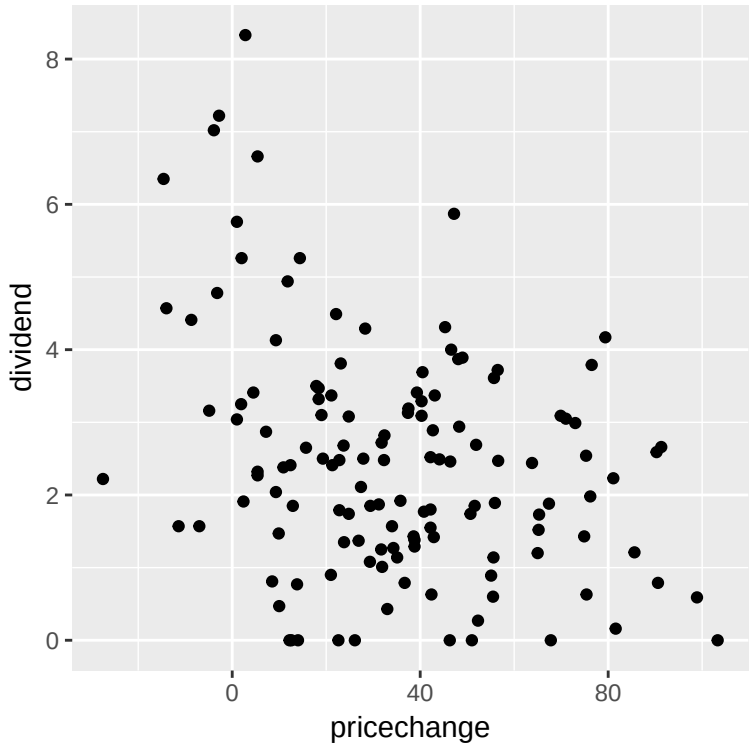
```
[1] -0.2710561
```

(e) (half a point) Which of the three correlation coefficient deems most appropriate to you in this situation?

11. You continue investigating the relationship between price change over the last 12 months and dividend. From the revious question you conclude that there are five outliers: two companies with high dividend, three with high price change.

(a) (half a point) Eliminate these five outliers and draw the plot anew. Has the relationship changed?

```
d <- companies[!(companies$dividend > 15 | companies$pricechange > 140), ]

ggplot(data = d) +
    geom_point(aes(x = pricechange, y = dividend))
```



(b) (1 point) Give the company names for those that you eliminated from your analysis?

```
companies[(companies$dividend > 15 & !is.na(companies$dividend) | companies$pricechange > 140 ), "name"]
```

```
[1] "IDB Holdings"    "Brait SA. (JSE)" "Combined Motor"  "Peregrine"
[5] "PSG Group"
```

(c) (1 point) Calculate the Pearson correlation coefficient for the restricted set. By how much has the correlation coefficient changed.

```
cor(d$pricechange, d$dividend, use = "complete.obs")
```

```
[1] -0.3311798
```