Class 05: Data Visualization with GGPLOT

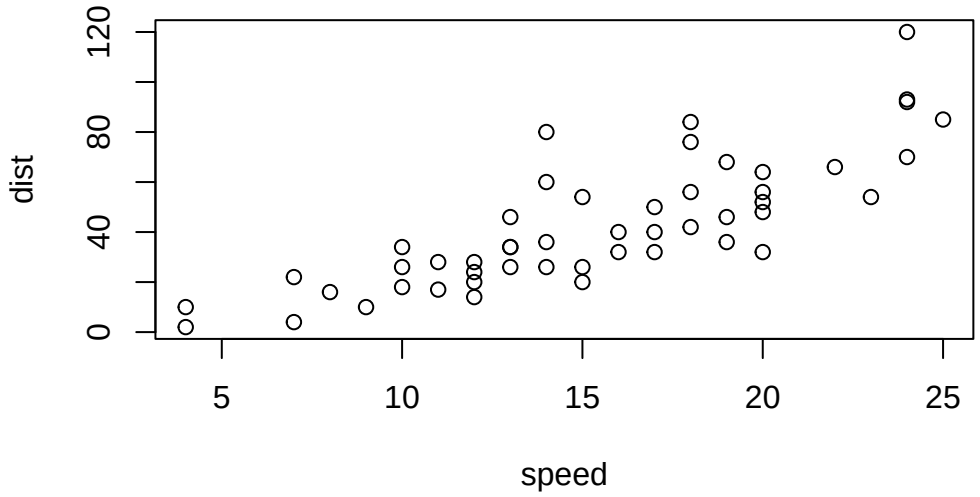
Jessica Diaz-Vigil

4/19/23

GGPLOT

We are going to start by generating the plot of class 04. This code is plotting the **cars** `dataset`.

```
plot(cars)
```



First, we need to install the package. We do this by using the `install.packages` command.

```
# install.packages('ggplot2')
```

After that, we need to load the package.

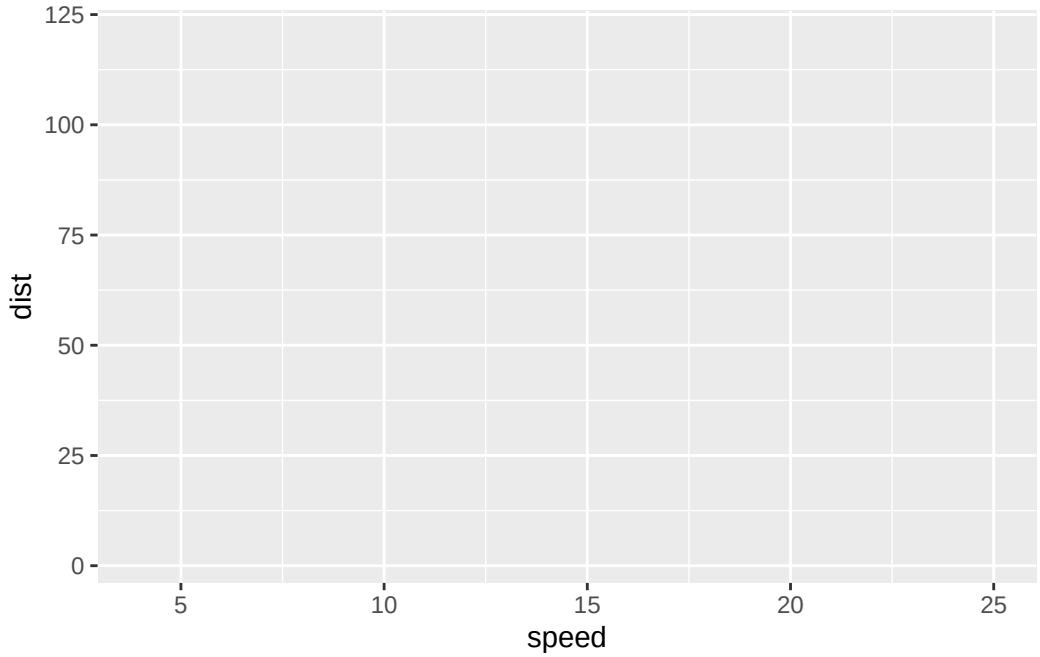
```
library(ggplot2)
```

We are going to build the plot of the cars `dataframe` by using ggplot2.

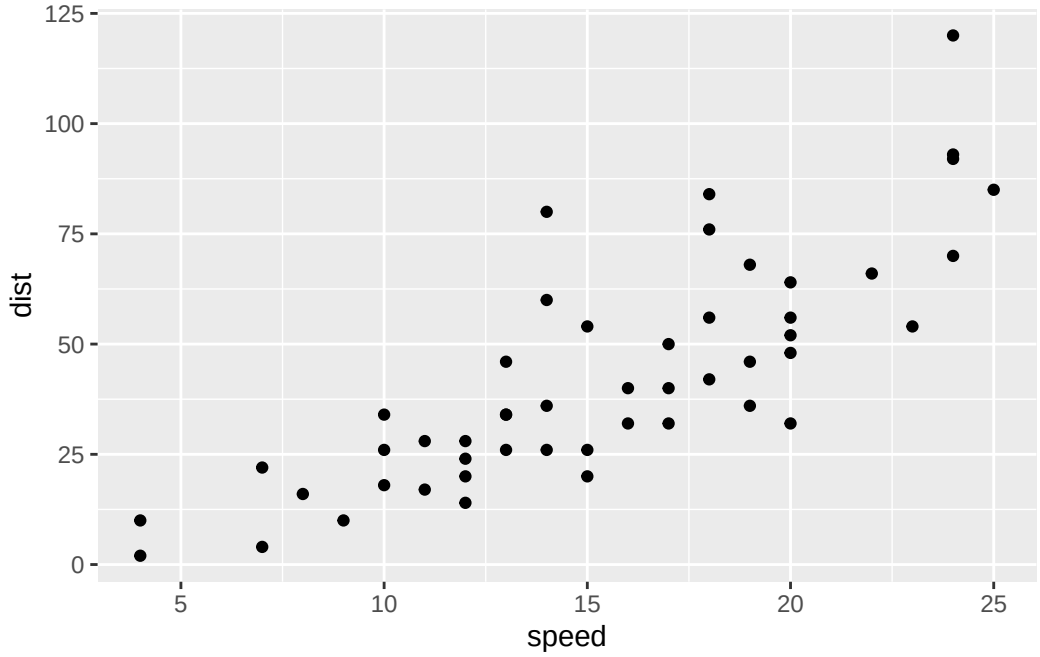
```
ggplot(data = cars)
```



```
ggplot(data = cars) + aes(x=speed, y=dist)
```

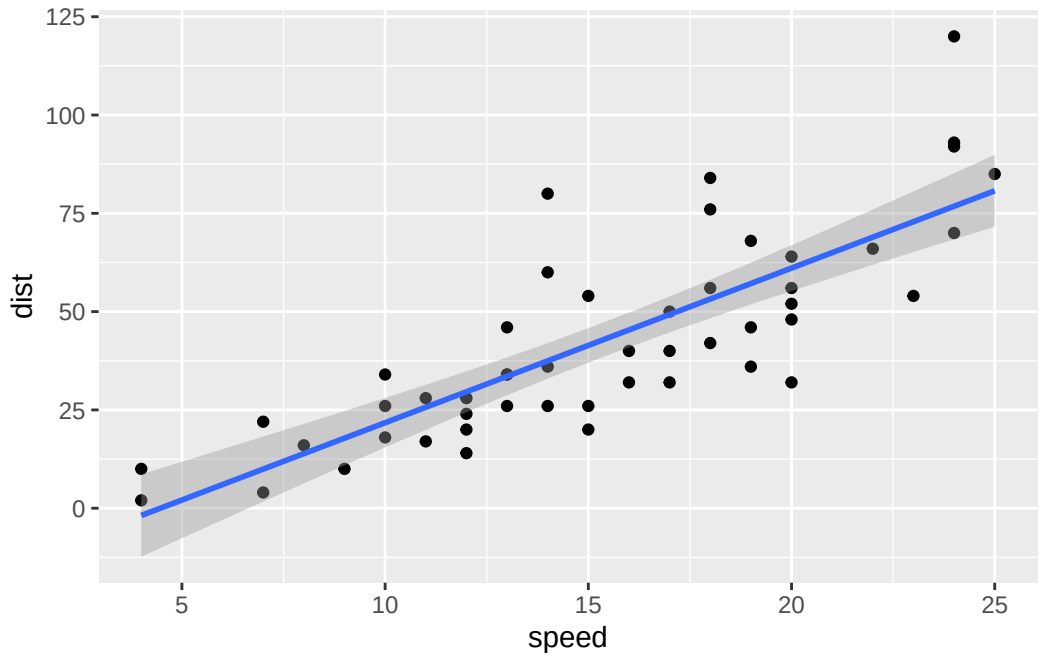



```
ggplot(data = cars) +
  aes(x=speed, y=dist) + geom_point()
```



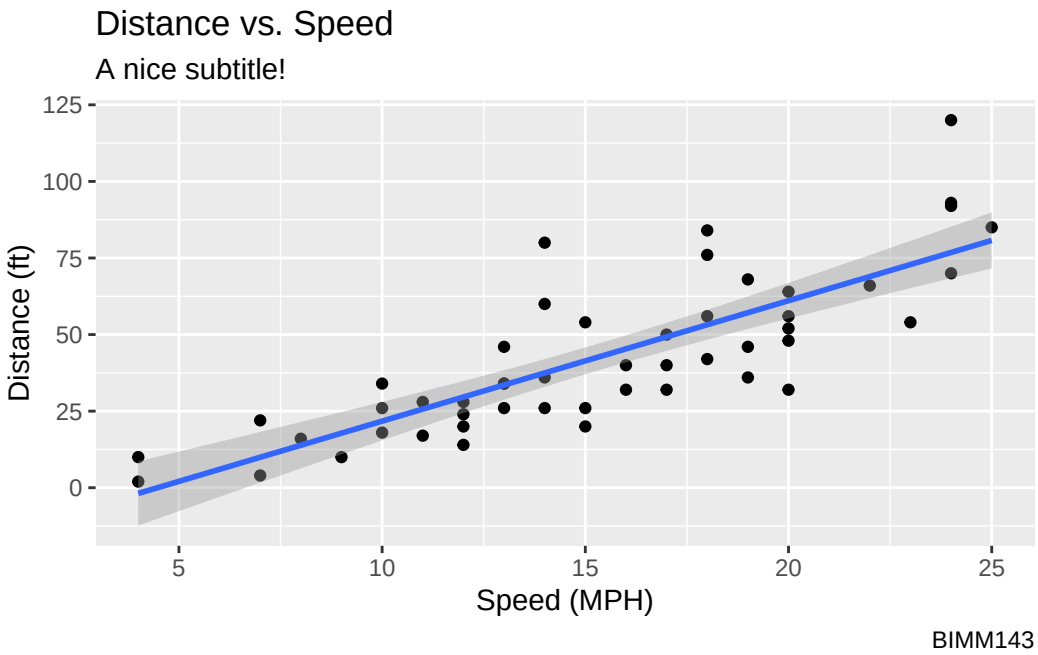
```
  ggplot(data = cars) +
    aes(x=speed, y=dist) + geom_point() +
    geom_smooth(method = 'lm')
```

```
`geom_smooth()` using formula = 'y ~ x'
```



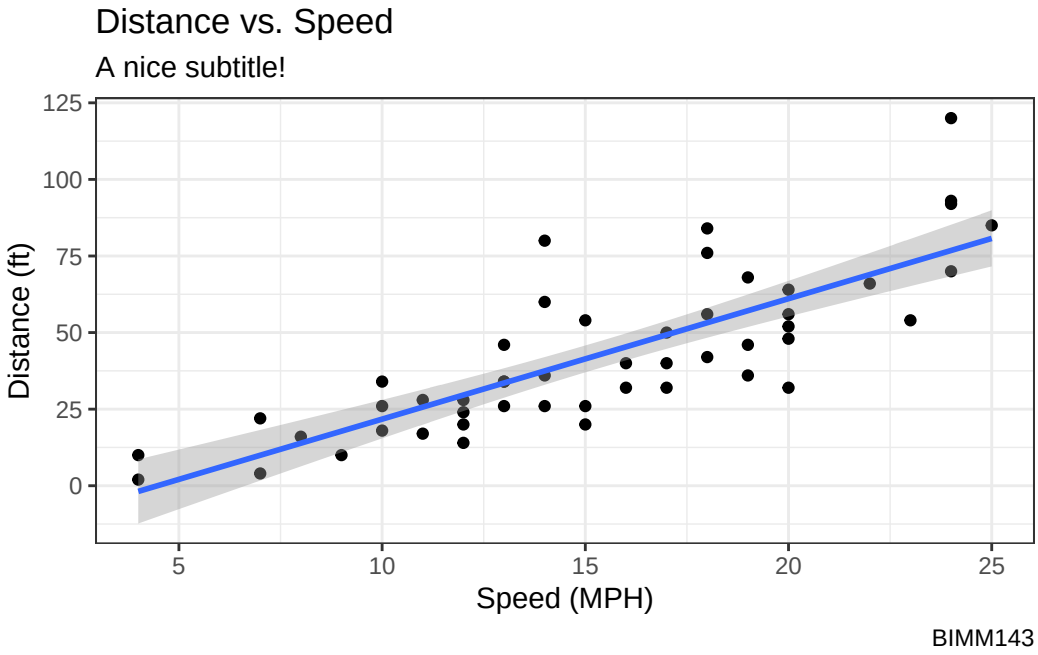

```
  ggplot(data = cars) +
    aes(x=speed, y=dist) + geom_point() +
    geom_smooth(method = 'lm') +
    labs( title = "Distance vs. Speed",
          subtitle = 'A nice subtitle!',
          caption = 'BIMM143',
          x = "Speed (MPH)",
          y = 'Distance (ft)')
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data = cars) +
  aes(x=speed, y=dist) + geom_point() +
  geom_smooth(method = 'lm') +
  labs( title = "Distance vs. Speed",
        subtitle = 'A nice subtitle!',
        caption = 'BIMM143',
        x = "Speed (MPH)",
        y = 'Distance (ft)') + theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



PLOTTING EXPRESSION DATA

Loading the data from the URL

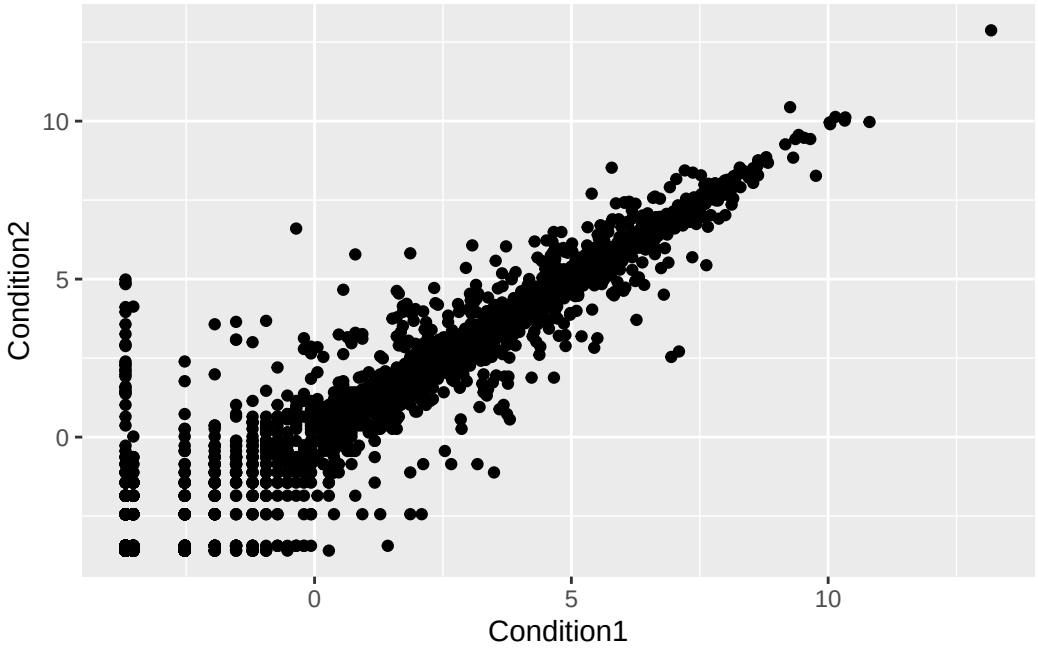
```
url <- "https://bioboot.github.io/bimm143_S20/class-material/up_down_expression.txt"

genes <- read.delim(url)
head(genes)
```

```
        Gene Condition1 Condition2      State
1      A4GNT -3.6808610 -3.4401355 unchanging
2       AAAS  4.5479580  4.3864126 unchanging
3      AASDH  3.7190695  3.4787276 unchanging
4       AATF  5.0784720  5.0151916 unchanging
5       AATK  0.4711421  0.5598642 unchanging
6 AB015752.4 -3.6808610 -3.5921390 unchanging
```

Initial ggplot

```
ggplot(data = genes) +
  aes(x = Condition1, y = Condition2) +
  geom_point()
```




```
nrow(genes)
```

```
[1] 5196
```

```
ncol(genes)
```

```
[1] 4
```

```
colnames(genes)
```

```
[1] "Gene"       "Condition1" "Condition2" "State"
```

```
table(genes['State'])
```

```
State
      down unchanging         up
        72       4997        127
```

```
round( table(genes$State)/nrow(genes) * 100, 2 )
```

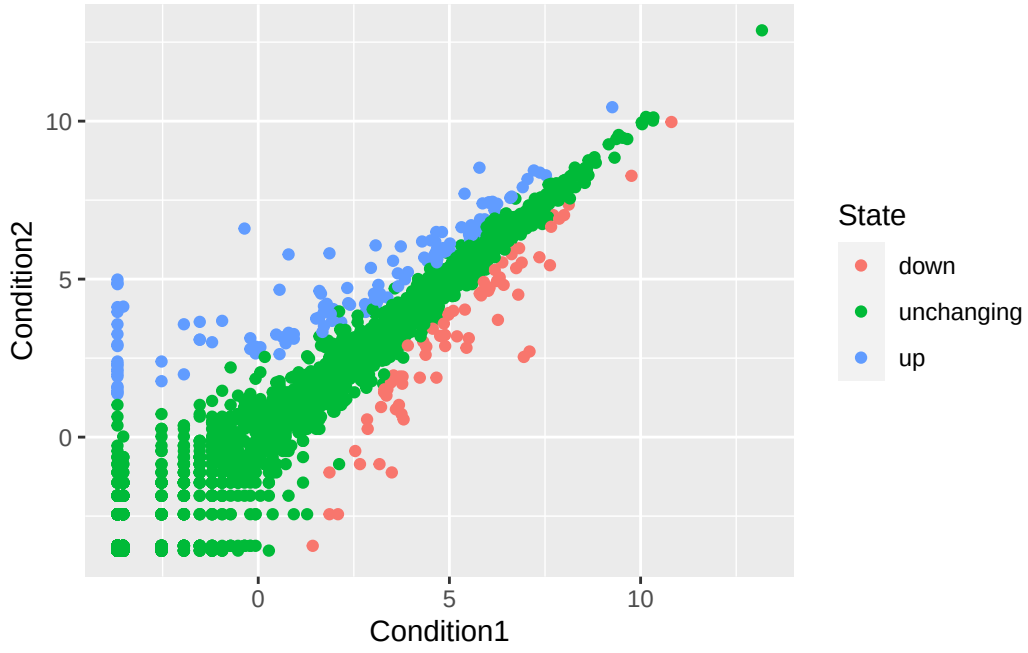
```
      down unchanging         up
      1.39      96.17       2.44
```

Storing the plot

```
p <- ggplot(data = genes) +
  aes(x = Condition1, y = Condition2,
      col = State) + geom_point()
```

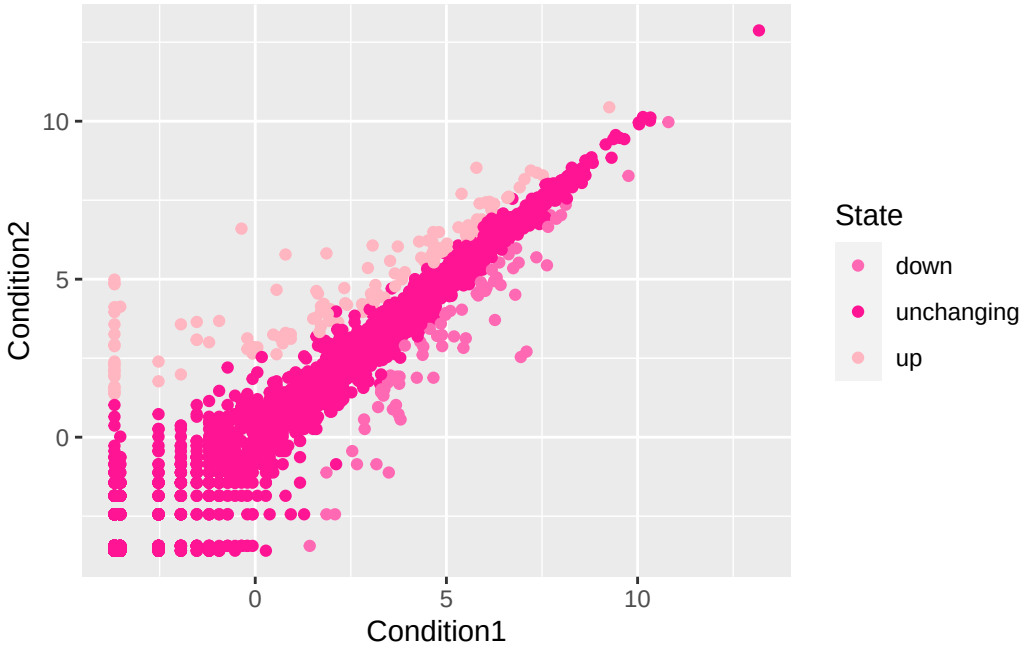
Adding color to the plot

```
ggplot(data = genes) +
  aes(x = Condition1, y = Condition2, col = State) +
  geom_point()
```



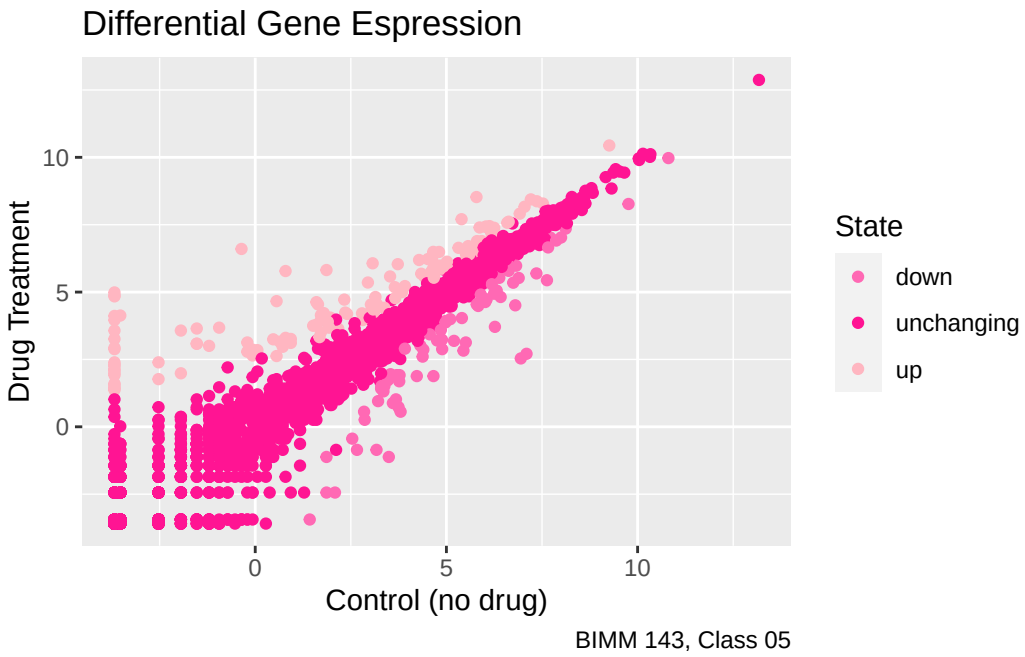
Changing colors (pink!!)

```
p1 <- p + scale_colour_manual( values=c("hotpink","deeppink","lightpink") )
p1
```



Adding Labels

```
p1 + labs(title = "Differential Gene Espression",
          x = "Control (no drug)",
          y = "Drug Treatment",
          caption = "BIMM 143, Class 05")
```



LAB QUESTIONS

Q1. For which phases is data visualization important in our scientific workflows?

All of the above: Communication of results, Exploratory data analysis, Detection of outliers

Q2. True or False? The ggplot2 package comes already installed with R?

False

Q3. Which plot types are typically NOT used to compare distributions of numeric variables?

Network graphs

Q4. Which statement about data visualization with ggplot2 is incorrect?

ggplot2 is the only way to create plots in R

Q5. Which geometric layer should be used to create scatter plots in ggplot2?

geom_point()

Q6. Use the `nrow()` function to find out how many genes are in this `dataset`. What is your answer?

5196

Q7. Use the `colnames()` function and the `ncol()` function on the `genes`data frame to find out what the column names are (we will need these later) and how many columns there are. How many columns did you find?

4

Q8. Use the `table()` function on the `State` column of this data.frame to find out how many 'up' regulated genes there are. What is your answer?

127

Q9. Using your values above and 2 significant figures. What fraction of total genes is up-regulated in this `dataset`?

2.44