Class05: Data Vis with ggplot

Alexis (PID: A17628362)

Graphics systems in R

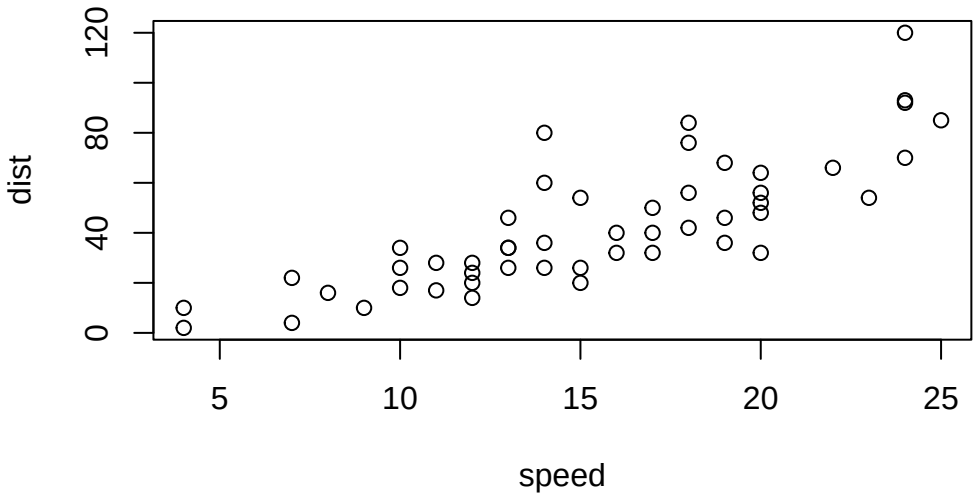
There are many graphics systems in R for making plots and figures.

We have already played a little with **"base R"** graphics and the 'plot()' function.

Today we will start learning about a popular graphics package called 'ggplot2()'.

This is an add on package - i.e. we need to install it. I install it (like I install any package) with the 'install.packages()' function.

```
plot(cars)
```



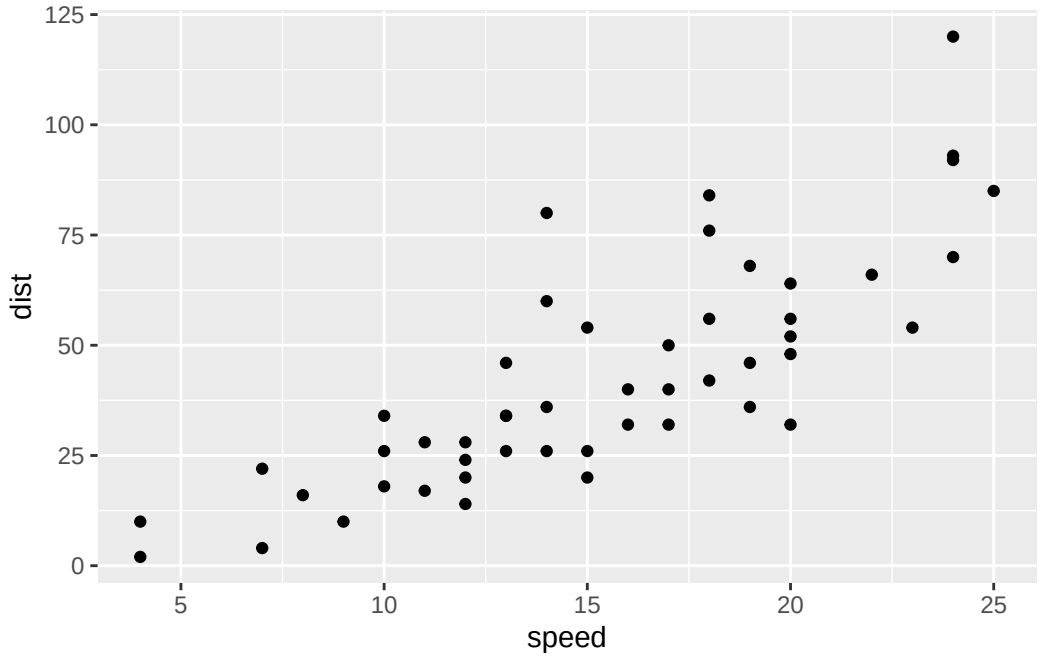
Before I can use the functions from a package I have to load up the package from my "library". We use the 'library(ggplot2)' command to load it up.

```
library(ggplot2)
ggplot(cars)
```


Every ggplot is made up of at least 3 things: - data (the numbers etc. that will go into your plot) - aes (how the columns of data map to the plot aesthetics) - geoms (how the plot actually looks, points, bars, lines, etc.)

ctrl + alt + 1 to open a new chunk

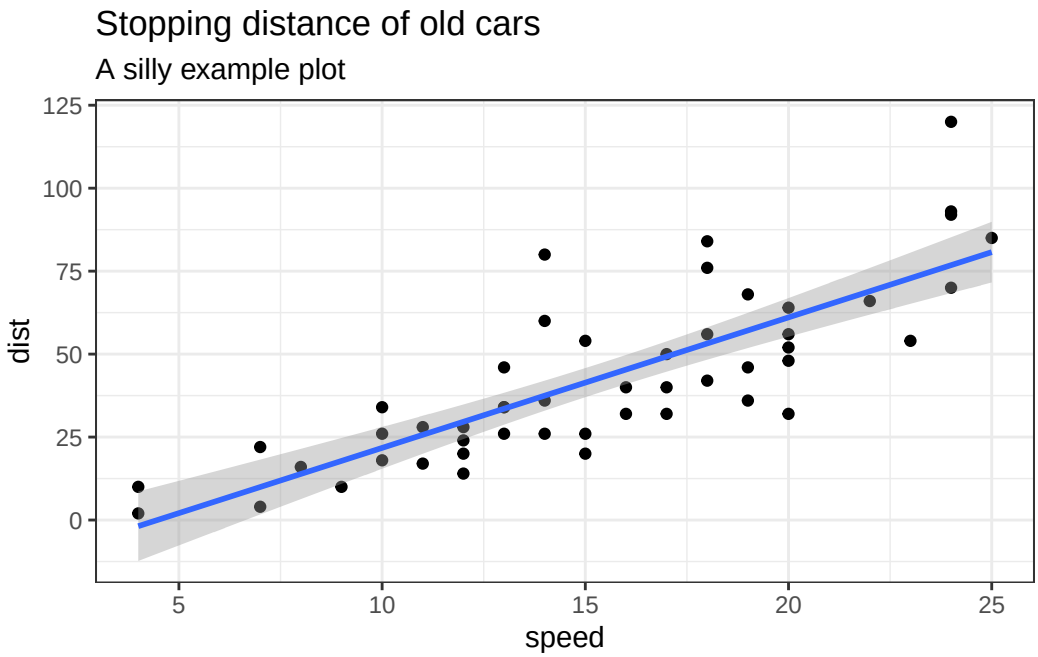
```
ggplot(cars) +
  aes (x=speed, y=dist) +
  geom_point()
```



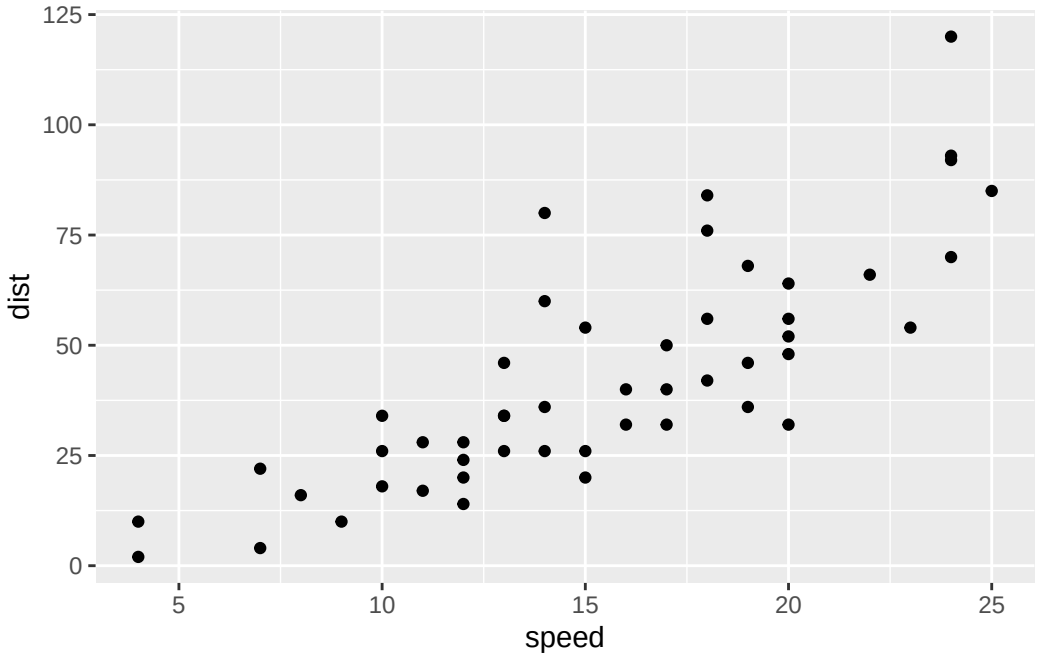
For simple ggplot is more verbose - it takes more code - than base R plot.

```
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title="Stopping distance of old cars",
       subtitle = "A silly example plot") +
  theme_bw()
```

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




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



Q1. For which phases is data visualization important in our scientific workflows? - Communication of Results - Exploaratory Data Analysis (EDA) - Detection of Outliers

Q2. T or F? The ggplot2 package comes already installed with R? - False

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

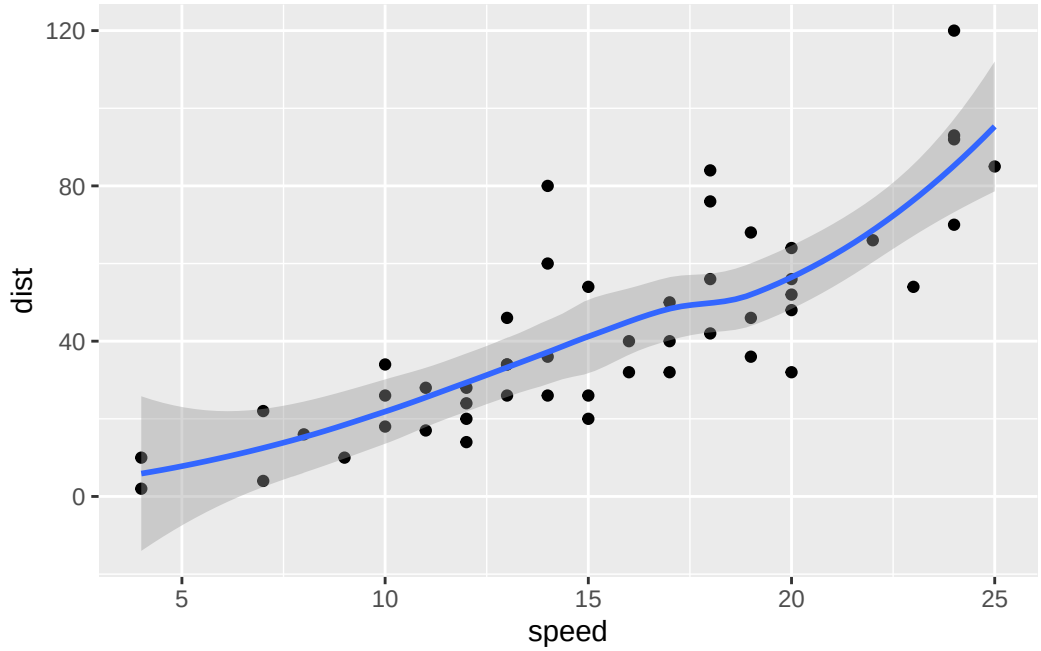
Q4. Which statement about data visualization with ggplot2 is incorrect? - ggplot2 is the only way to create plots in R

Q5. Which geometric layer should be used to create scatter plots in ggplot2? - geom_point()

Q6. In your own RStudio can you add a trend line layer to help show the relationship between the plot variables with the geom_smooth() function?

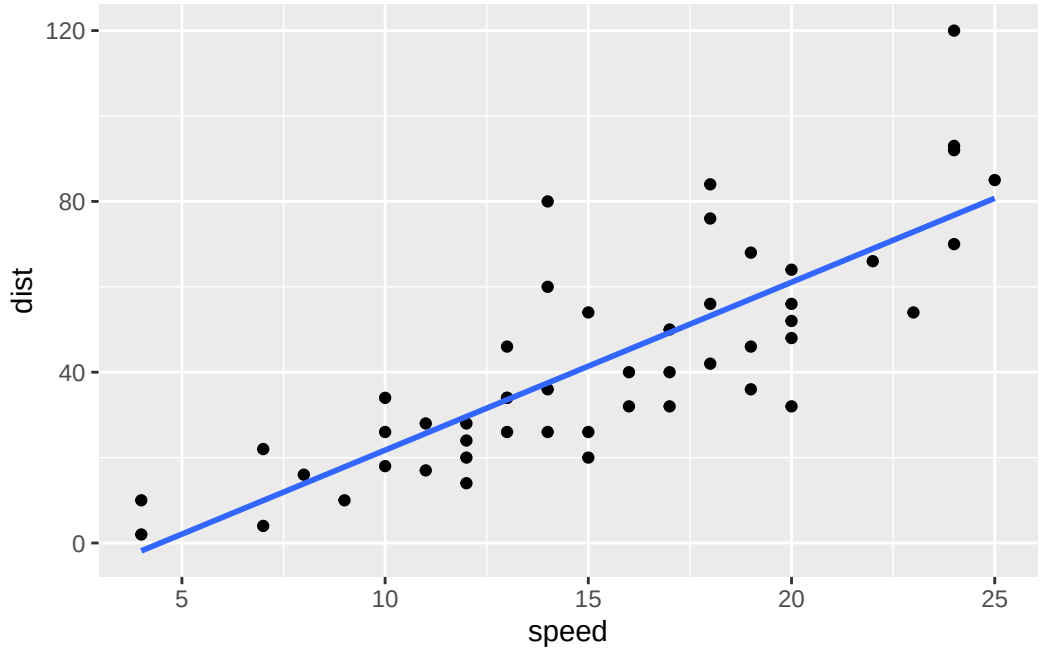
```
ggplot(cars) +
 aes(x=speed, y=dist) +
 geom_point() +
 geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



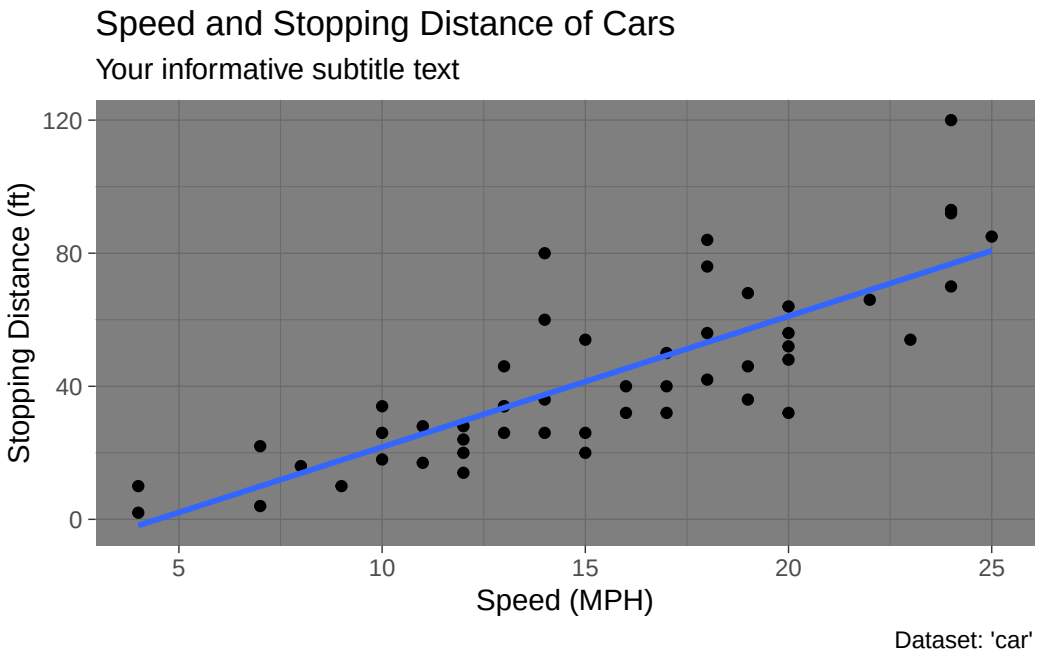

```
  ggplot(cars) +
   aes(x=speed, y=dist) +
   geom_point() +
   geom_smooth(method="lm", se=FALSE)
```

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




```
ggplot(cars) +
 aes(x=speed, y=dist) +
 geom_point() +
 labs(title="Speed and Stopping Distance of Cars",
    x="Speed (MPH)",
    y="Stopping Distance (ft)",
    subtitle= "Your informative subtitle text",
    caption="Dataset: 'car'") +
 geom_smooth(method="lm", se=FALSE) +
 theme_dark()
```

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




```
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
```

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

```
nrow(genes)
```

```
[1] 5196
```

- There are 5196 genes in this dataset using nrow(genes).

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?

```
colnames(genes)
```

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

```
ncol(genes)
```

```
[1] 4
```

- There are 4 columns.

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?

```
table(genes$State)
```

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

- There are 127 'up' regulated genes using the table() function.

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

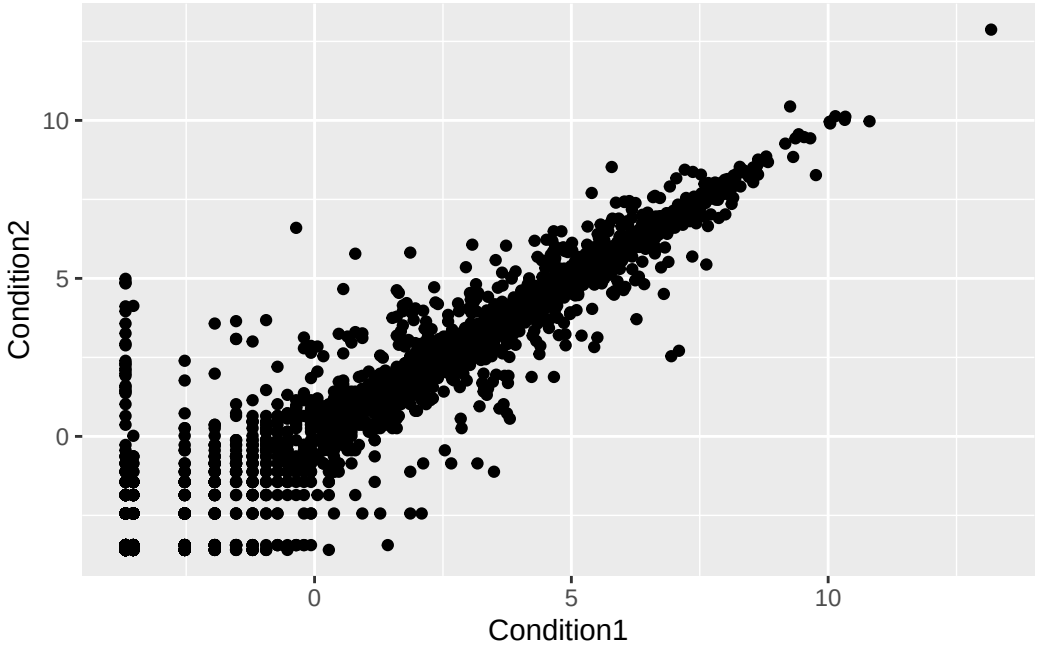
```
round( table(genes$State)/nrow(genes) * 100, 2 )
```

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

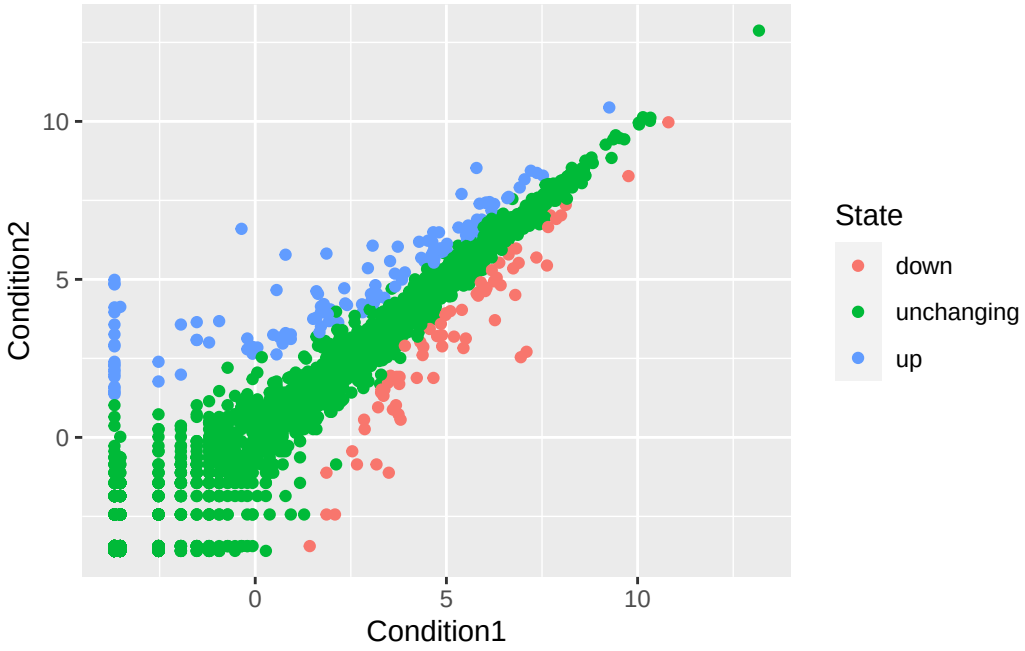
- 2.44/100 of total genes are up-regulated in this dataset.

Q10. Complete the code below to produce the following plot.

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




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


Condition2
10
5
0
0
5
10
Condition1
State
down
unchanging
up