Class 5: Data Viz with ggplot

Thrisha Praveen: A16836270

Intro to ggplot

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

- All of the above

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

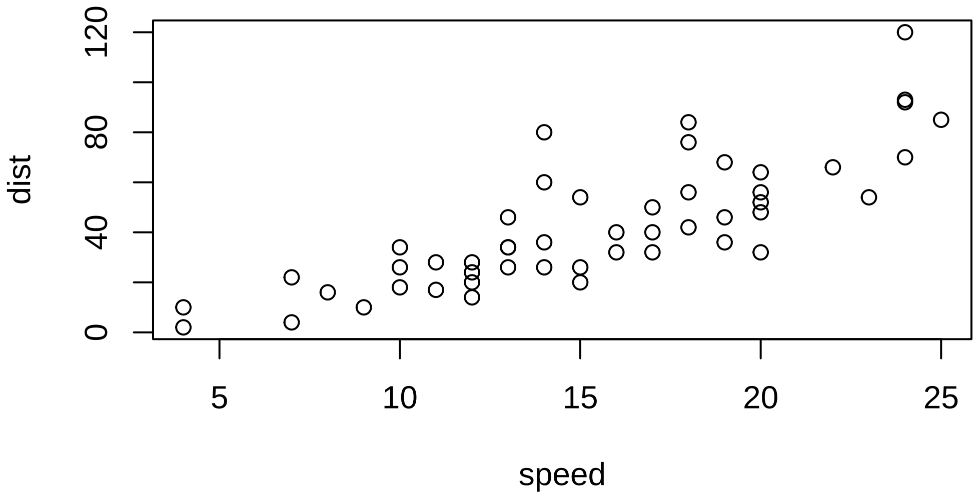
Q4. Which statement about data visualization with ggplot2 is incorrect?

- ggplot2 is the only way to create plots in R

There are many graphics systems in R (ways to make plots and figures). These include "base" R plots. Today we will focus mostly on the **ggplot2** package.

Let's start with a plot of a simple in-built dataset called `cars`.

```
plot(cars)
```

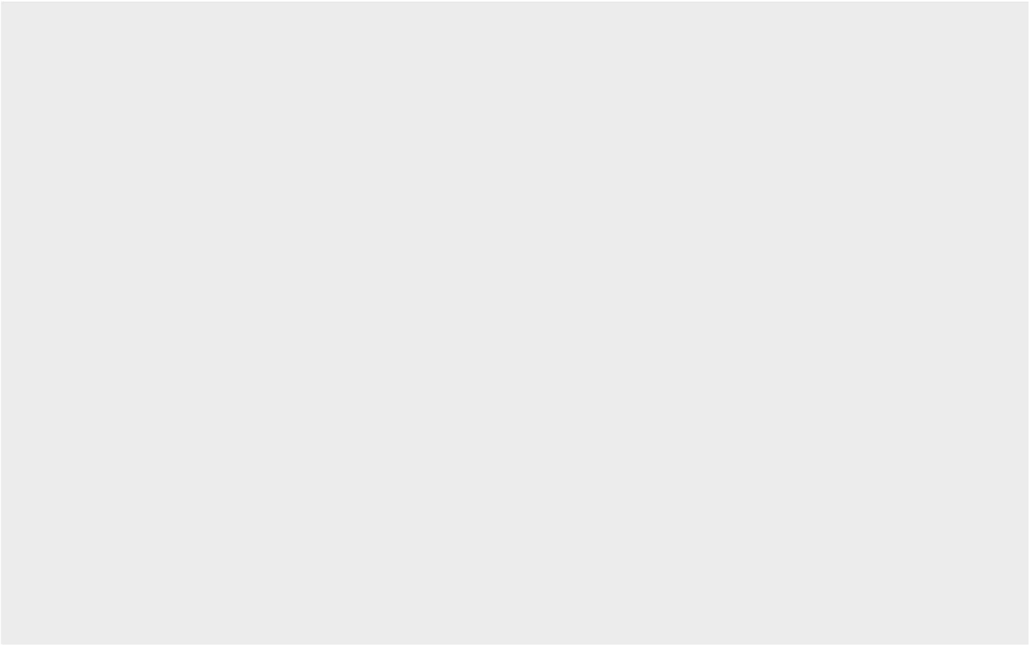


Let’s see how we can make this figure using **ggplot**. First, I need to install this package on my computer. To install any R package, I use the function `install.packages()`.

> I will run `install.packages("ggplot2")` in my R console, not this quarto document!

Before I can use any functions from add-on packages, I need to load the package from my “library()” with the `library(ggplot2)` call.

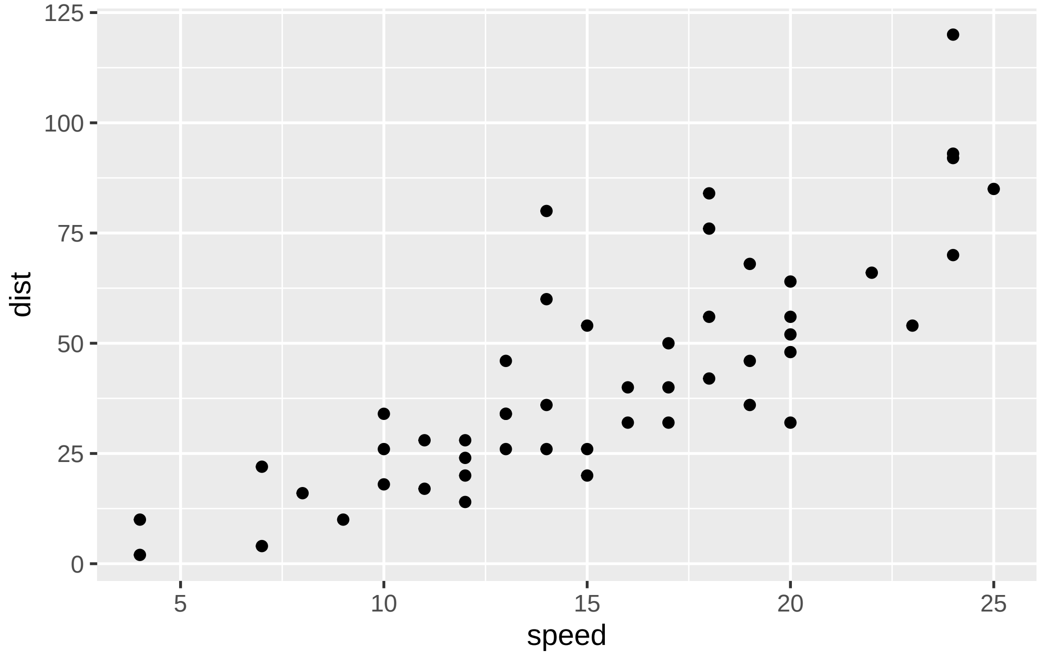
```
library(ggplot2)
ggplot(cars)
```



All ggplot figures have at least 3 layers, which include:

- **data** (the input dataset I want to plot from)
- **aes** (the aesthetic mapping of the data to my plot)
- **geoms** (the geom_point(), geom_line(), etc. that I want to draw)

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



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

- geom_point()

Q. 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?

- shown below

Q. Argue with geom_smooth() to add a straight line from a linear model without the shaded standard error region?

- shown below

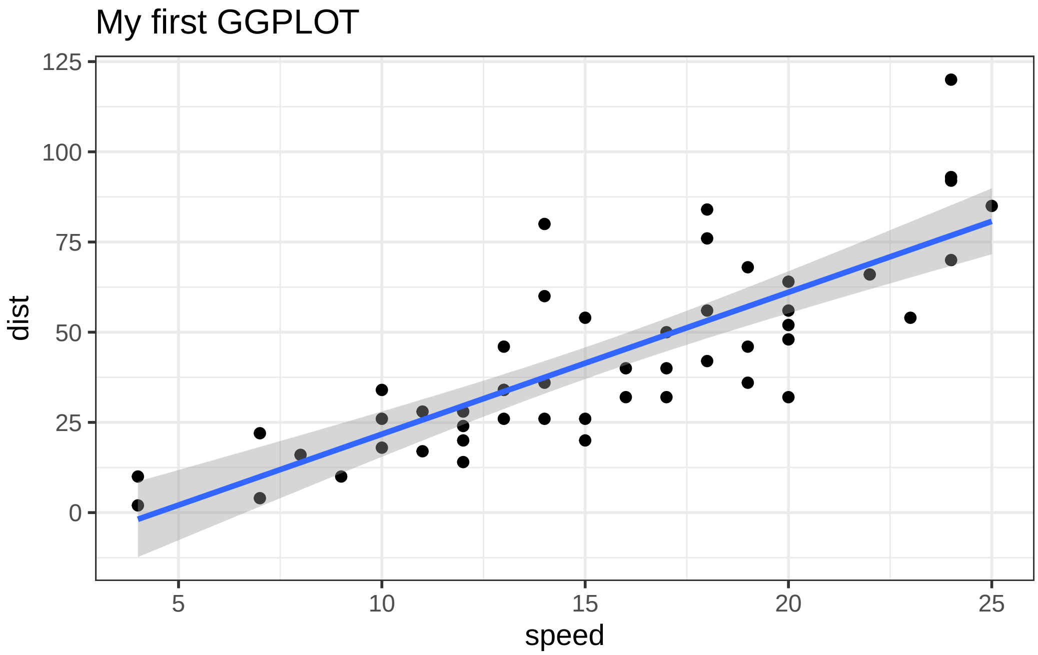
Q. Can you finish this plot by adding various label annotations with the labs() function and changing the plot look to a more conservative "black & white" theme by adding the theme_bw() function:

- shown below

Let's add a line to show the relationship here:

```r
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point() +
  geom_smooth(method="lm") +
  theme_bw() +
  labs(title="My first GGPLOT")
```

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



```r
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
```

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

- 5196

Q. 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

Q. 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

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

- 0.02

```
nrow(genes)
```

```
[1] 5196
```

```
colnames(genes)
```

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

```
ncol(genes)
```

```
[1] 4
```

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

```

      down unchanging         up 
        72       4997        127 
```

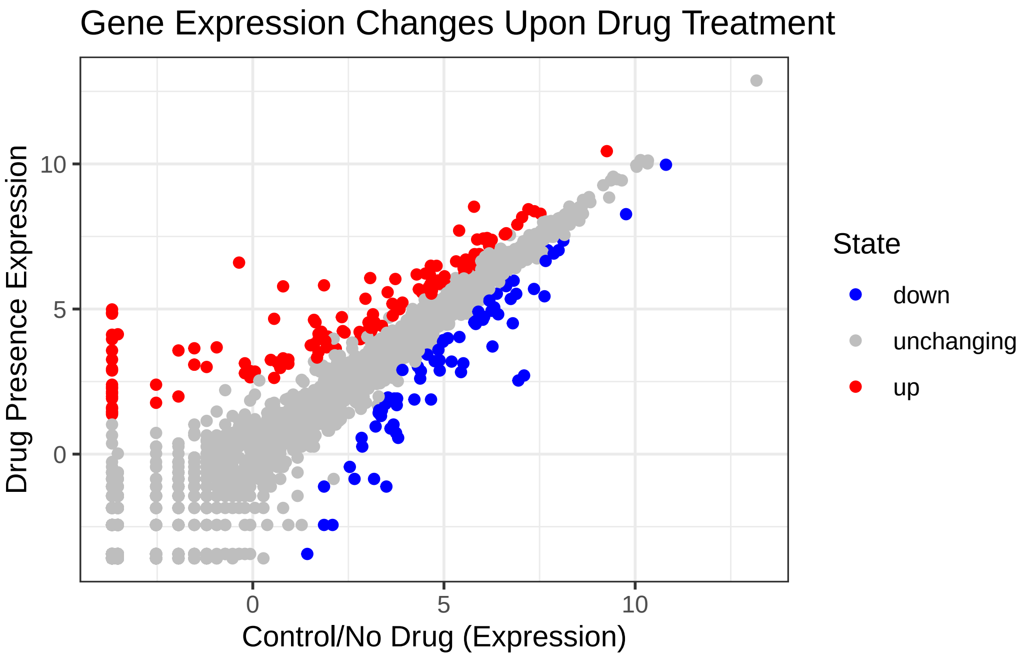
```
round(table(genes$State)/nrow(genes)*100, 2)
```

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

A first plot of this dataset:

Q. Complete the code below to produce the following plot `ggplot(___) + aes(x=Condition1, y=___) _ _____`

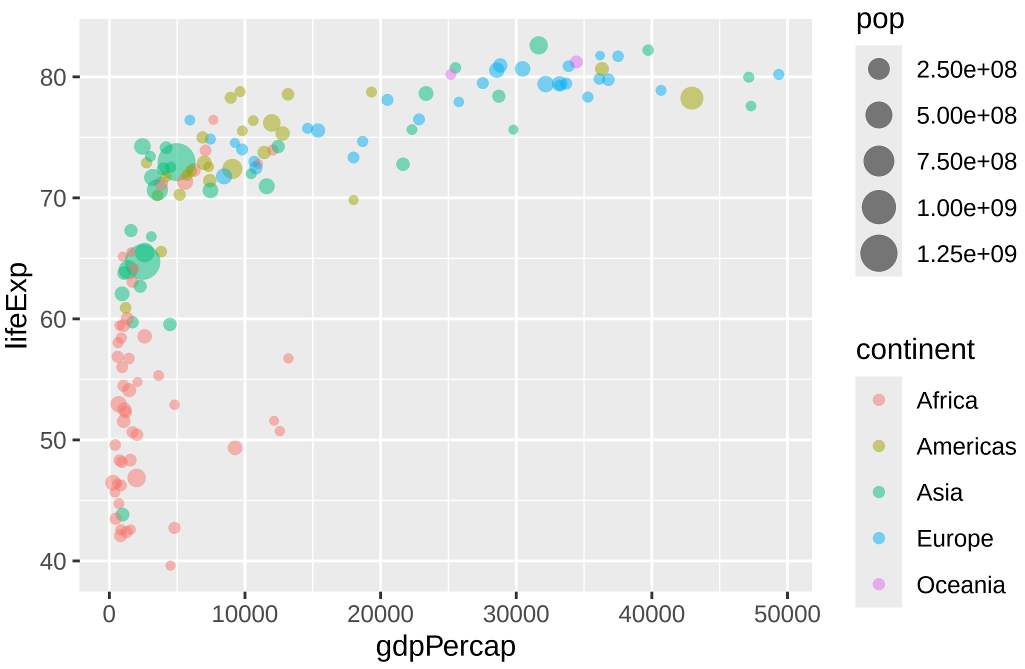
```
p <- ggplot(genes) + 
  aes(x=Condition1, y=Condition2, col=State) +
  geom_point() +
  theme_bw()+
  labs(title="Gene Expression Changes Upon Drug Treatment",
       x = "Control/No Drug (Expression)",
       y = "Drug Presence Expression")

p + scale_color_manual( values=c("blue","gray","red") )
```



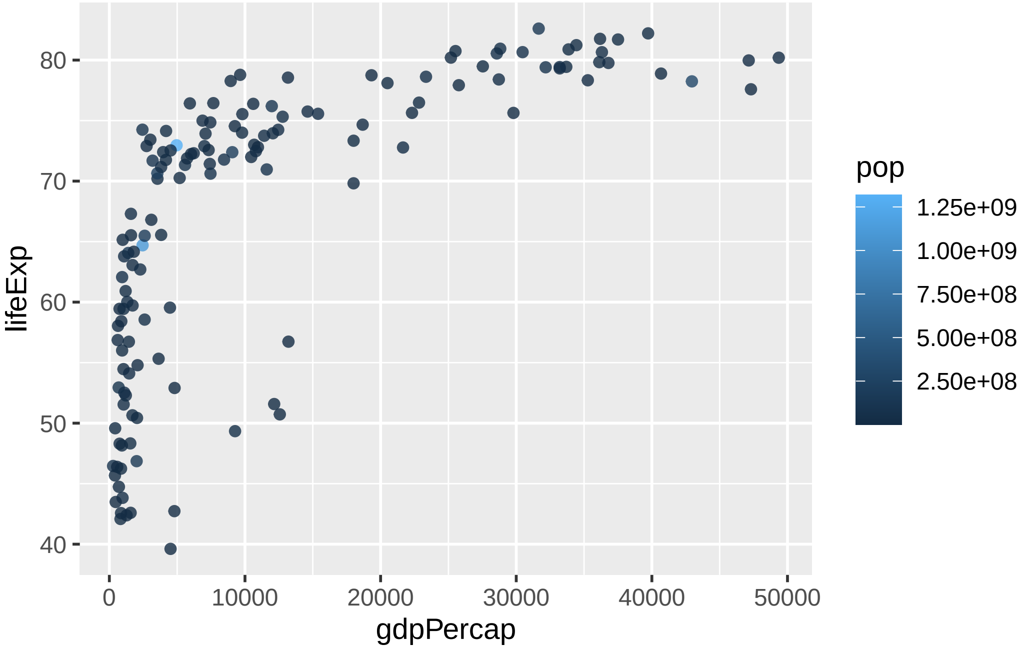
```r
library(gapminder)


# install.packages("dplyr")  ## un-comment to install if needed
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

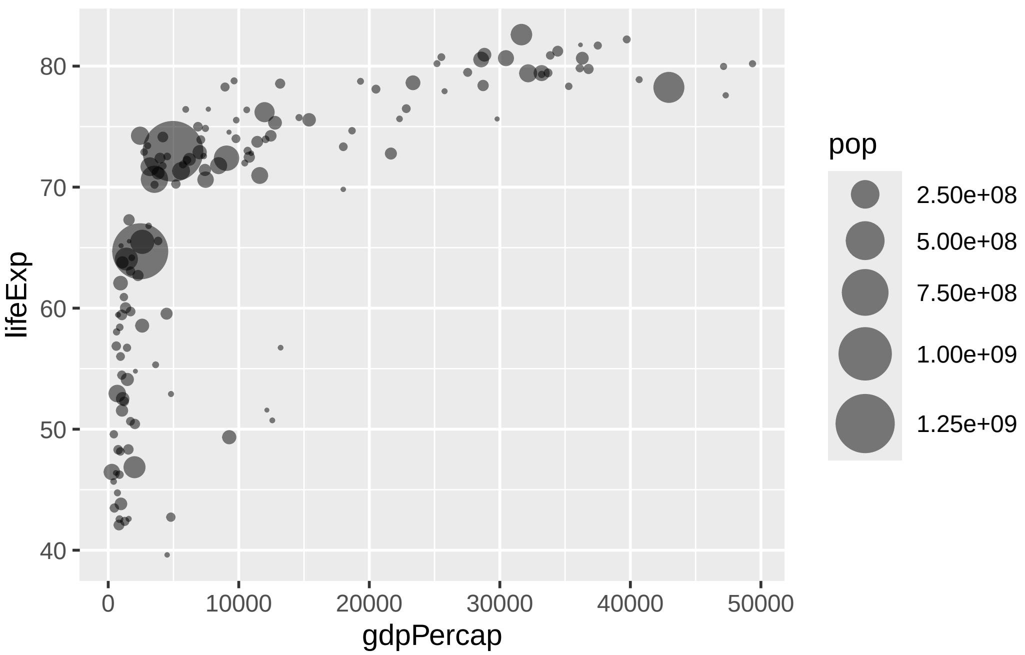
```r
gapminder_2007 <- gapminder %>% filter(year==2007)

ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, color=continent, size=pop) +
  geom_point(alpha=0.5)
```



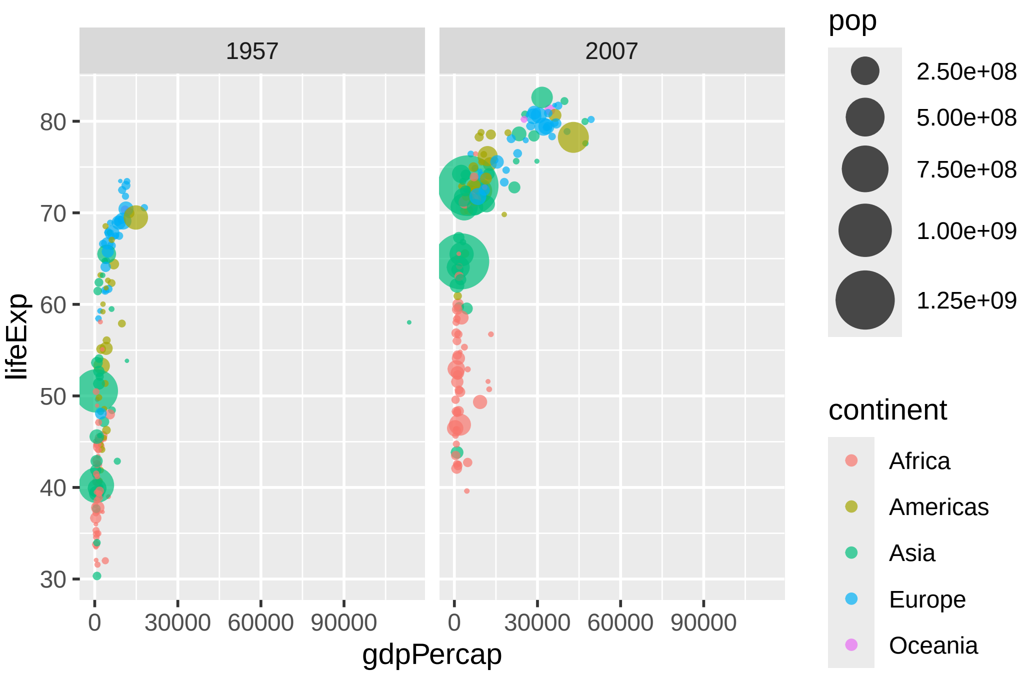
```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = pop) +
  geom_point(alpha=0.8)
```



```
ggplot(gapminder_2007) +
  geom_point(aes(x = gdpPercap, y = lifeExp,
                 size = pop), alpha=0.5) +
  scale_size_area(max_size = 10)
```



```
gapminder_1957 <- gapminder %>% filter(year==1957|year==2007)
ggplot(gapminder_1957) +
  geom_point(aes(x = gdpPercap, y = lifeExp, color=continent,
                 size = pop), alpha=0.7) +
  scale_size_area(max_size = 10) +
  facet_wrap(~year)
```


1957
2007
lifeExp
80
70
60
50
40
30
0
30000
60000
90000
gdpPercap
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
continent
Africa
Americas
Asia
Europe
Oceania