Class 5: Data Viz with ggplot

Jordan Prych(PID: A17080226)

Intro to ggplot

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

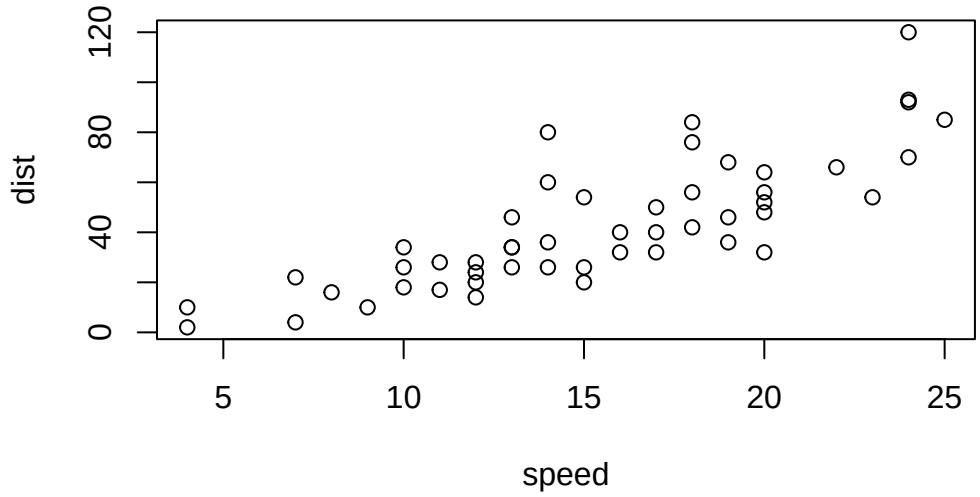
Let's start with a plot of a simple in-built dataset called `cars`

```
cars
```

```
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
11    11   28
12    12   14
13    12   20
14    12   24
15    12   28
16    13   26
17    13   34
18    13   34
19    13   46
20    14   26
21    14   36
22    14   60
```

```
23    14   80
24    15   20
25    15   26
26    15   54
27    16   32
28    16   40
29    17   32
30    17   40
31    17   50
32    18   42
33    18   56
34    18   76
35    18   84
36    19   36
37    19   46
38    19   68
39    20   32
40    20   48
41    20   52
42    20   56
43    20   64
44    22   66
45    23   54
46    24   70
47    24   92
48    24   93
49    24  120
50    25   85
```

```
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)
```

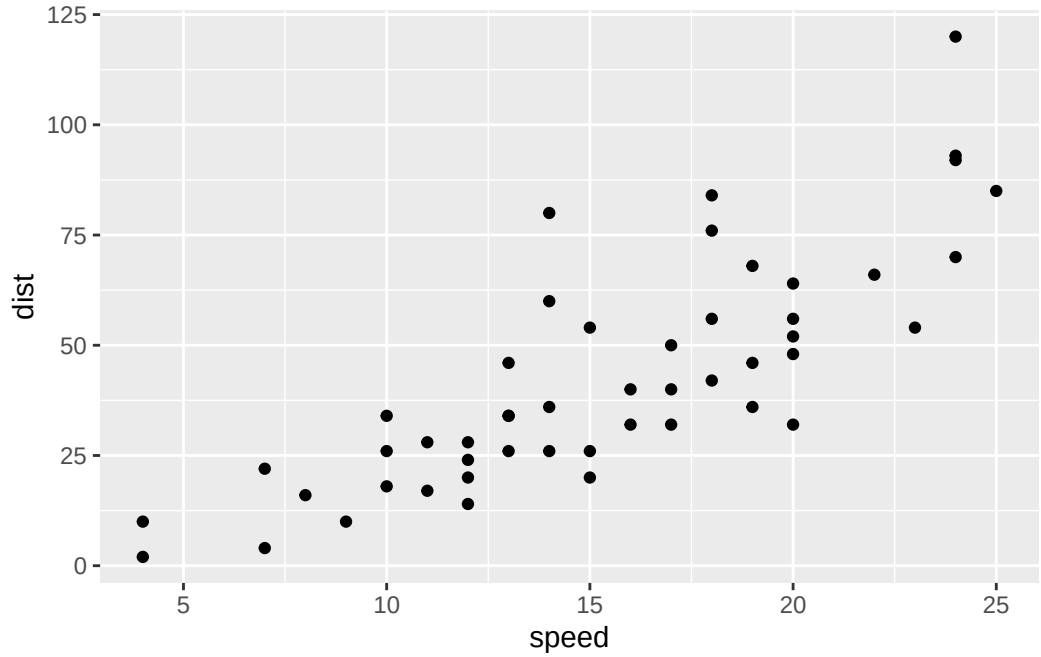
All ggplot figures have at least 3 things (called layers). These include:

-**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)

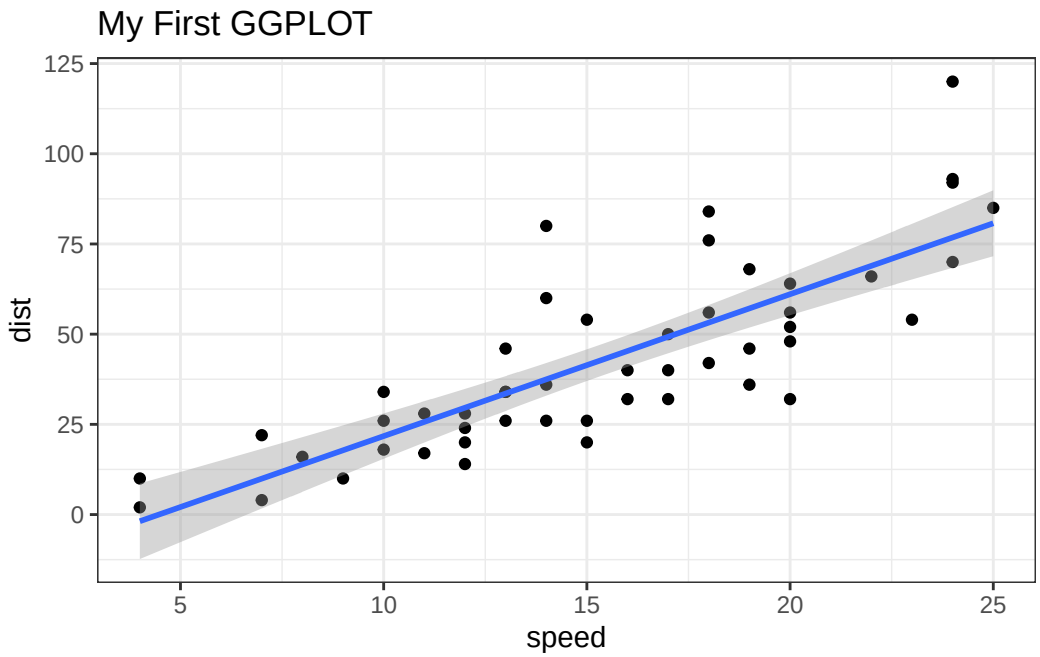
```
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point()
```



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

```
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'
```



Q1 Which geometric layer should be used to create scatter plots in ggplot2?

Geom_point()

Gene Expression Figure

The code to read the dataset

```
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
```

Q2 How many genes are in this dataset?

```
nrow(genes)
```

```
[1] 5196
```

Q3 How many columns are in this dataset?

```
ncol(genes)
```

```
[1] 4
```

Q4 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
```

Q5 Using your values above and 2 significant figures. What fraction of total genes is up-regulated in this dataset?

```
n.tot <- nrow(genes)
vals <- table(genes$State)
vals/n.tot
```

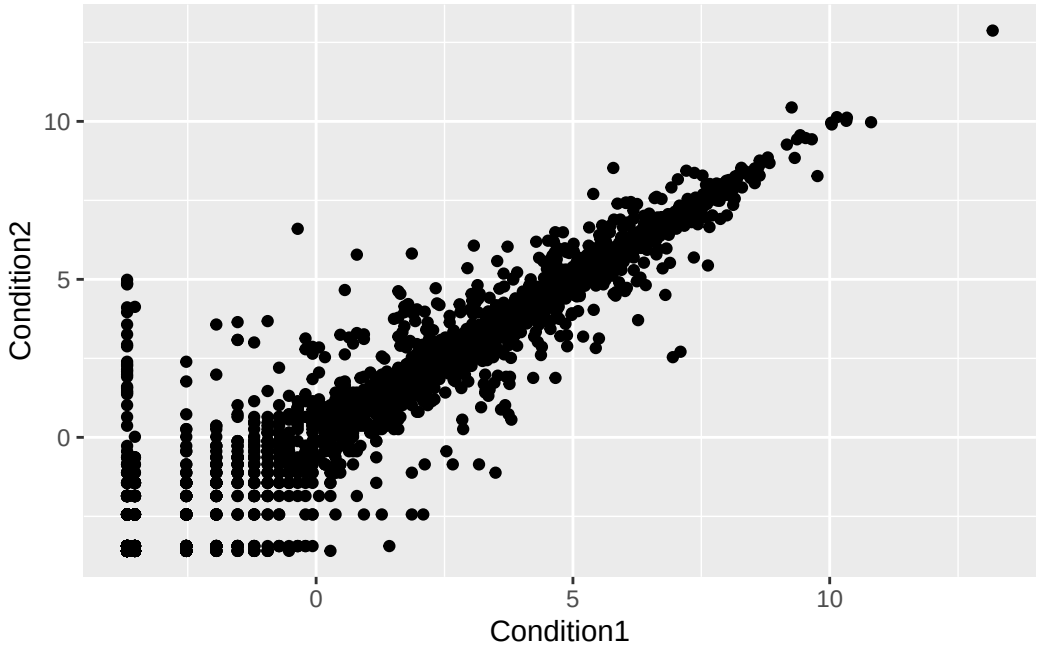
```
      down unchanging         up
0.01385681 0.96170131 0.02444188
```

```
vals.percent<- vals/n.tot * 100
round(vals.percent, 2)
```

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

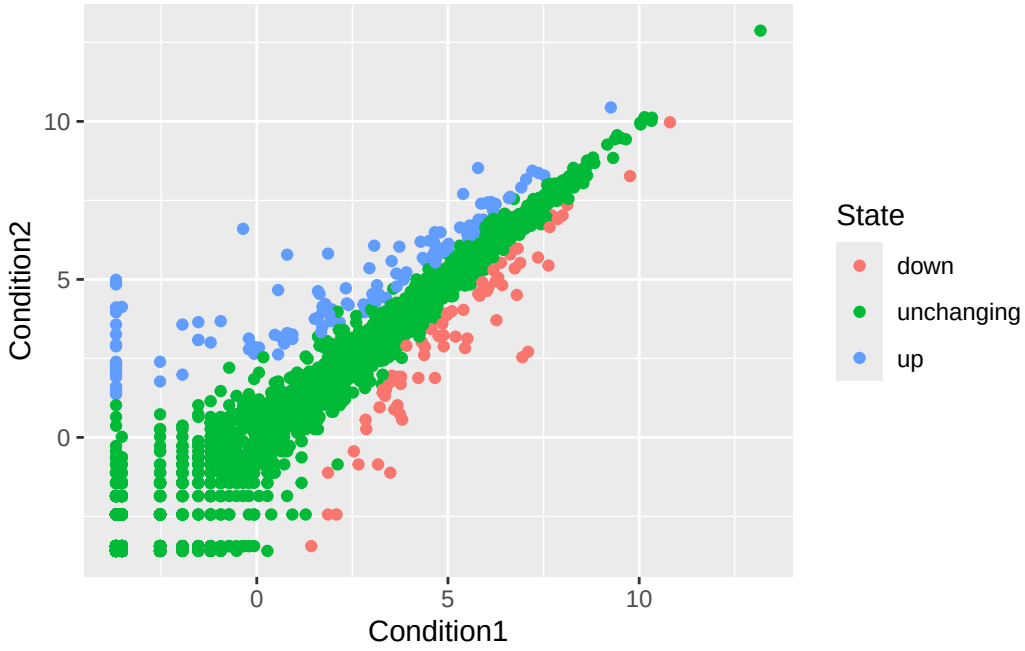
A first plot of this dataset:

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



To add color to this plot:

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

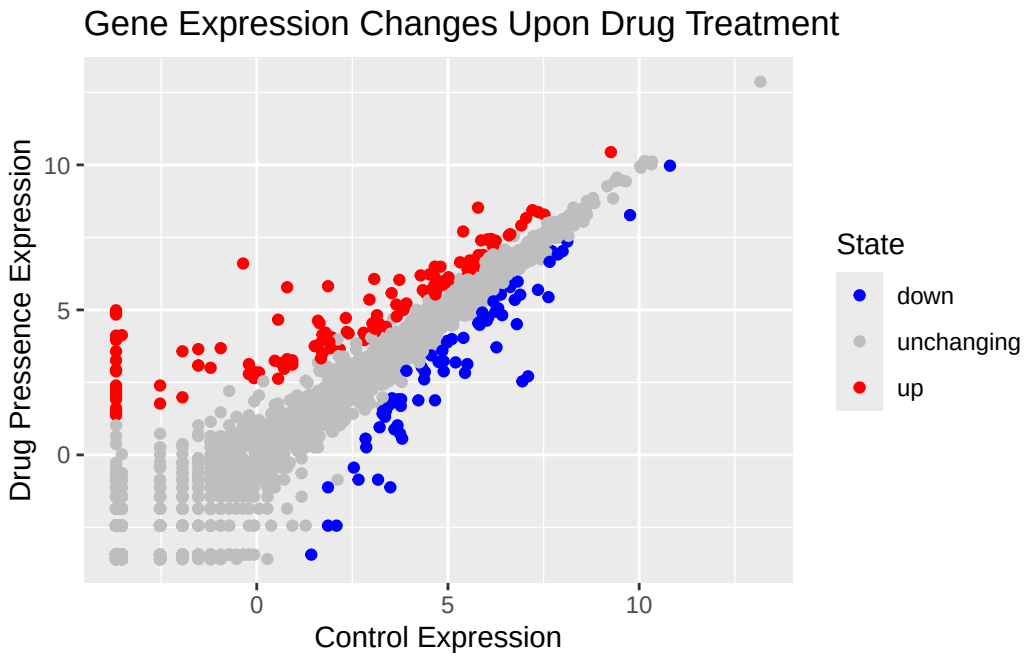


To save plot as p:

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

To specify color scale and add titles:

```
p + scale_colour_manual(values=c("blue", "gray", "red")) + labs(title="Gene Expression Chang
```



GapMinder Figures

The code to read the data

```
# File location online
url <- "https://raw.githubusercontent.com/jennybc/gapminder/master/inst/extdata/gapminder.ts
gapminder <- read.delim(url)
```

First, we need to install the dplyr package with the command `install.packages("dplyr")`

I will run 'install.packages("dplyr") in the Console and not in the quarto document

Before I can use any functions from add-on packages, I need to load the package from my "library()" with the `library(dplyr)` call. We will filter the data for rows with the year value of **2007** and save as `gapminder_2007`

```
library(dplyr)
```

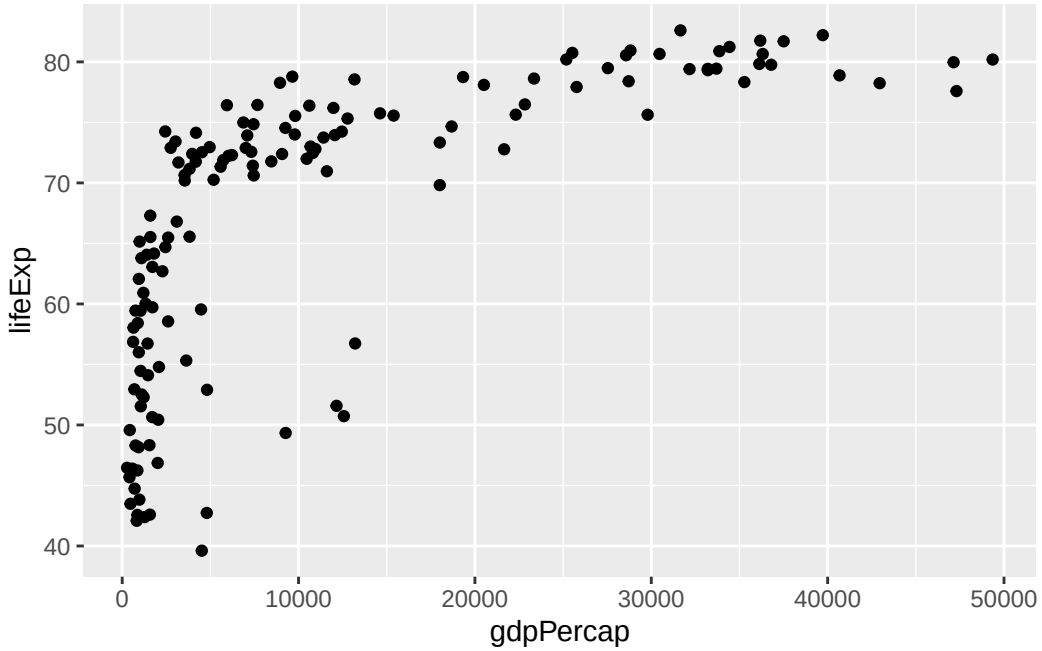
```
Attaching package: 'dplyr'
```

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

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

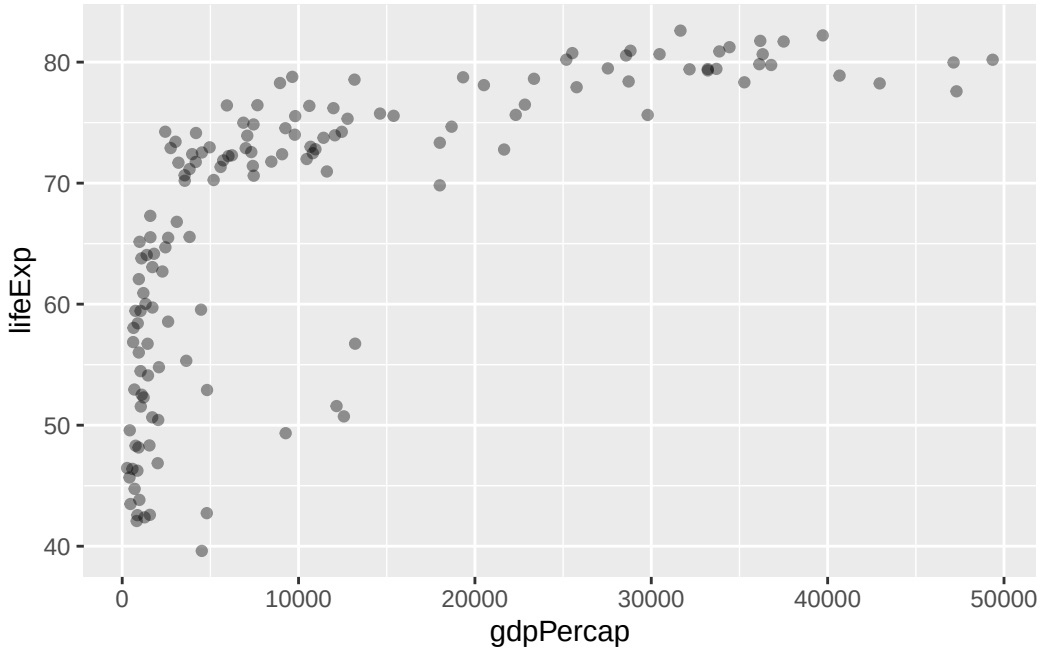
Q1 Complete the code below to produce a first basic scater plot of this gapminder_2007 dataset:

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp) +
  geom_point()
```



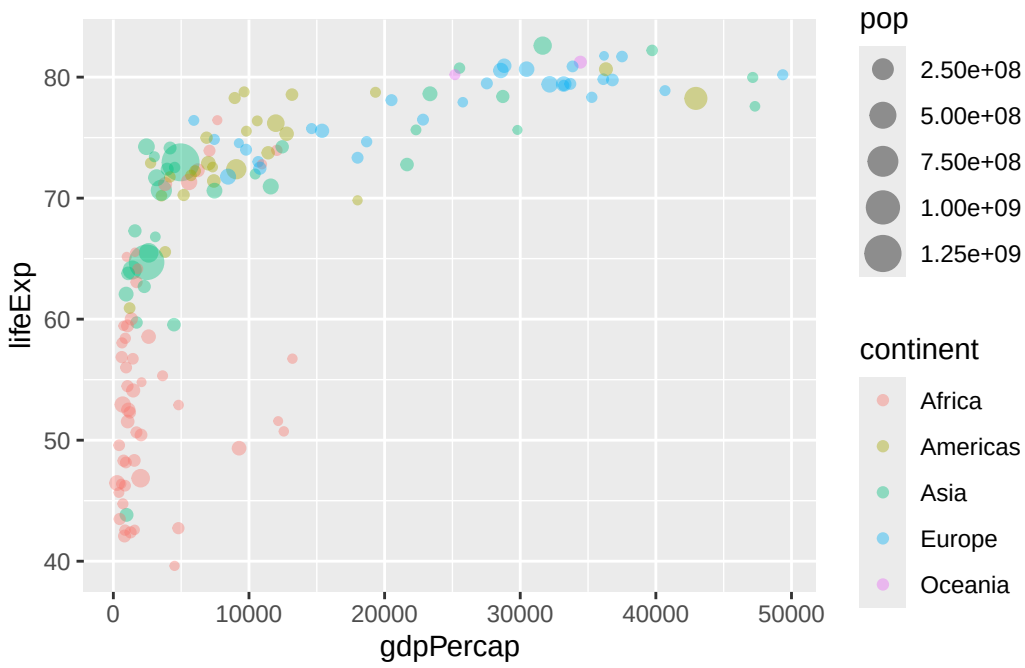
To observe overlapping points, use alpha argument:

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp) +
  geom_point(alpha=0.4)
```



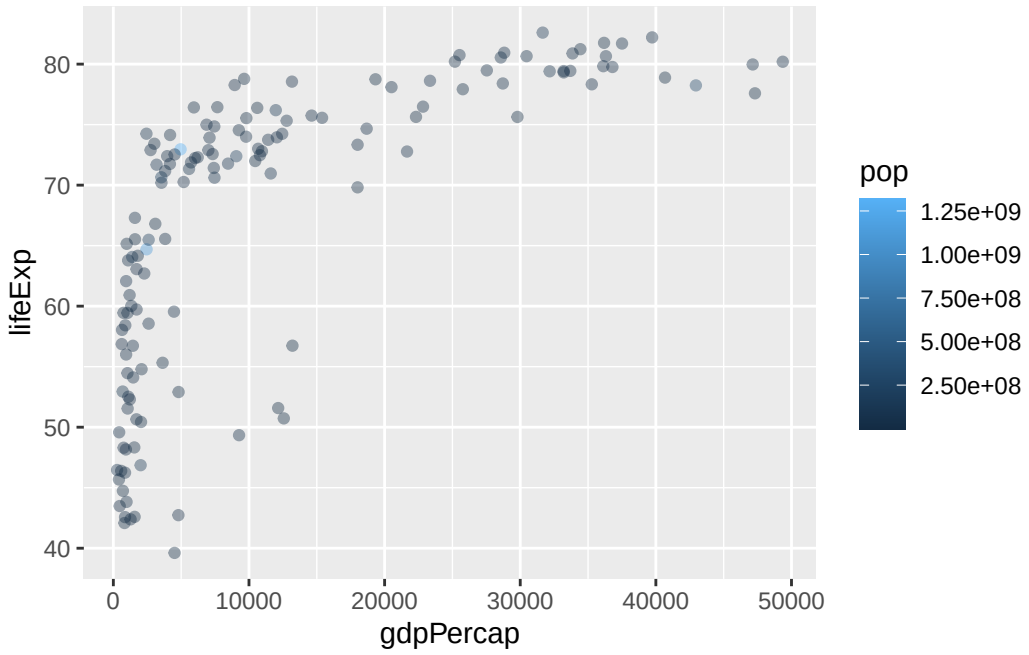
Mapping more variables to the aesthetic to add dimension to the plot:

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, color=continent, size=pop) +
  geom_point(alpha=0.4)
```



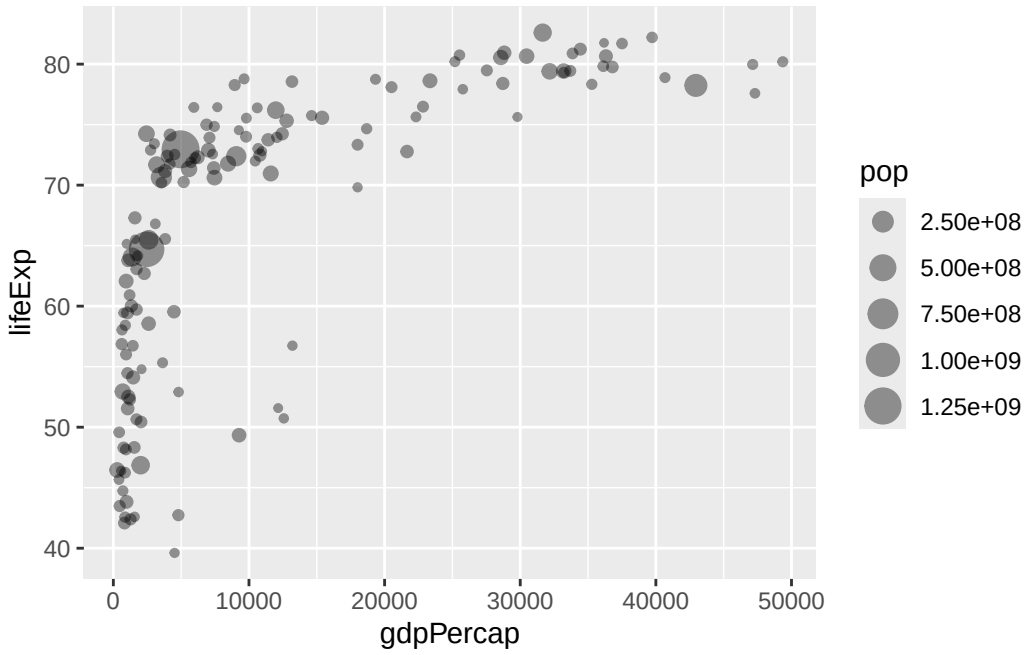
Color by numeric variable population(pop): >This changes the scale to be continuous

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, color=pop) +
  geom_point(alpha=0.4)
```



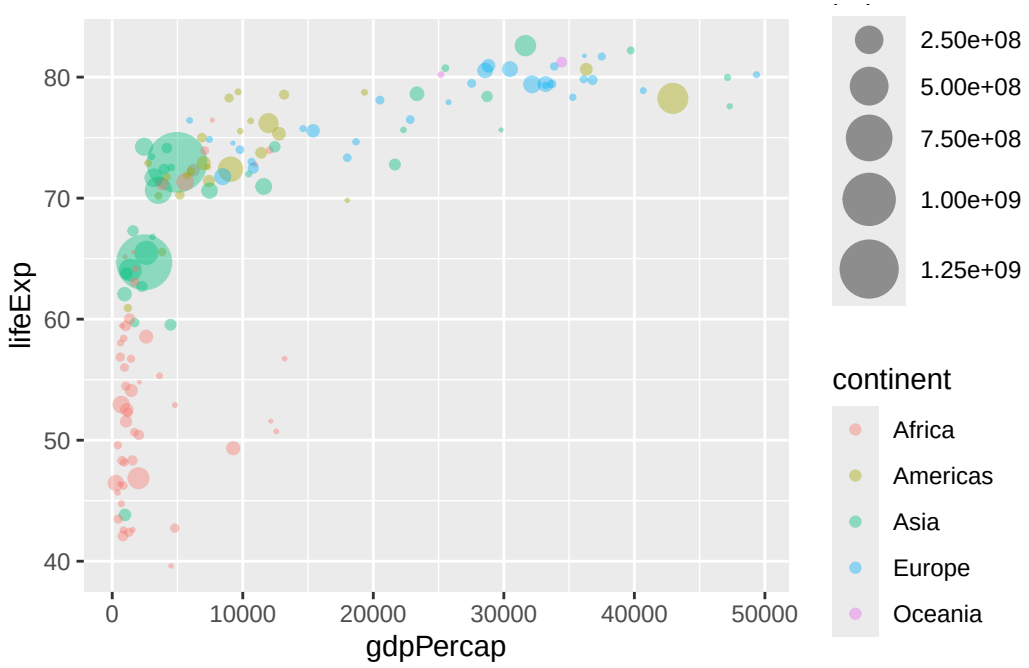
Adjusting Point Size

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, size=pop) +
  geom_point(alpha=0.4)
```



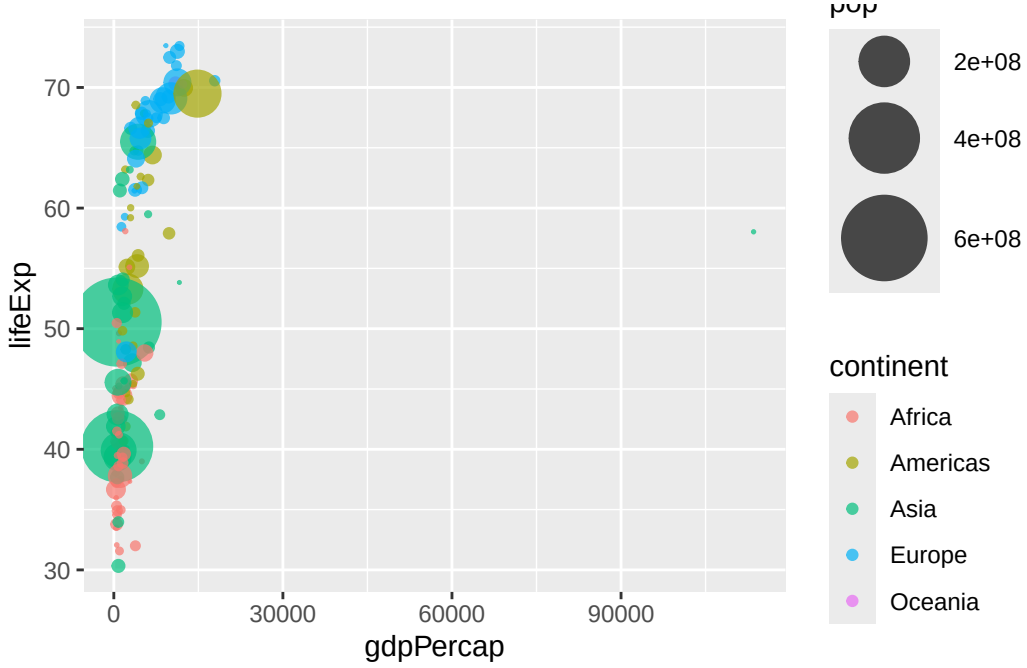
Use `scale_size_area()` to reflect proportional population differences by point size

```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, size=pop, color=continent) +
  geom_point(alpha=0.4) +
  scale_size_area(max_size=10)
```



Q2 Adapt the code you have learned thus far to reproduce our gapminder scatter plot for the year 1957?

```
library(dplyr)

gapminder_1957 <- gapminder %>% filter(year==1957)

ggplot(gapminder_1957) + aes(x=gdpPercap, y=lifeExp, size=pop, color=continent) +
  geom_point(alpha=0.7) +
  scale_size_area(max_size=15)
```

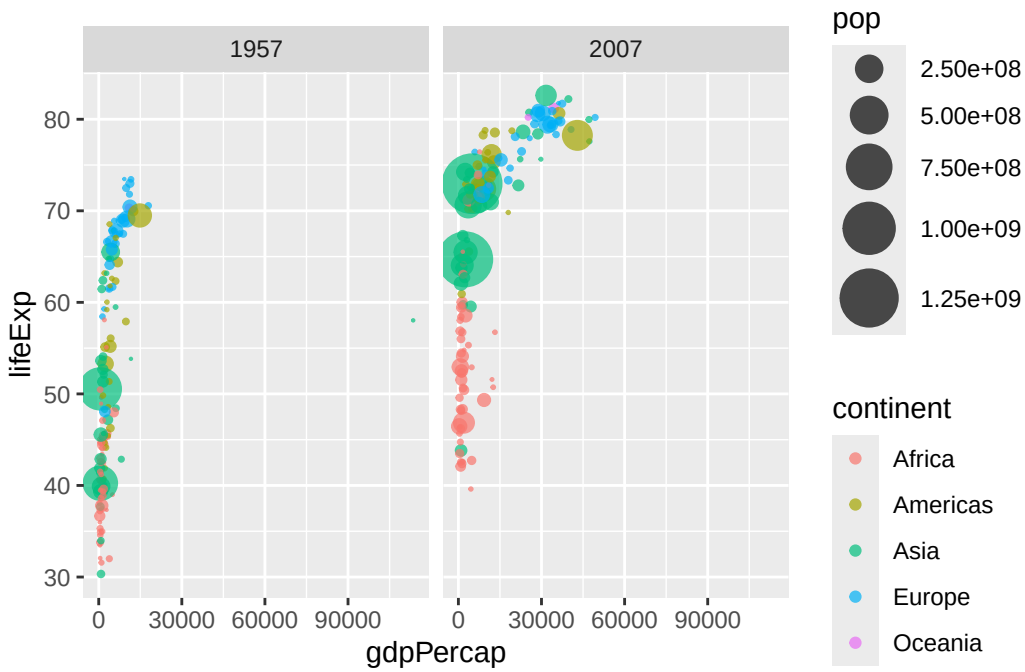


What do you notice about this plot? Is it easy to compare with the one for 2007?

The size of the points are larger in the 1957 plot compared to the 2007 plot, making it easier to compare population size compared to a smaller scale size in the 2007 plot.

Q3 Do the same steps above but include 1957 and 2007 in your input dataset for ggplot(). You should now include the layer facet_wrap(~year) to produce the following plot:

```
gapminder_1957 <- gapminder %>% filter(year==1957 | year==2007)

ggplot(gapminder_1957) +
  aes(x=gdpPercap, y=lifeExp, size=pop, color=continent) +
  geom_point(alpha=0.7) +
  scale_size_area(max_size=10) + facet_wrap(~year)
```



Patchwork Figures

First, we need to install the patchwork package with the command `install.packages("patchwork")`

> I will run 'install.packages("patchwork") in the Console and not in the quarto document

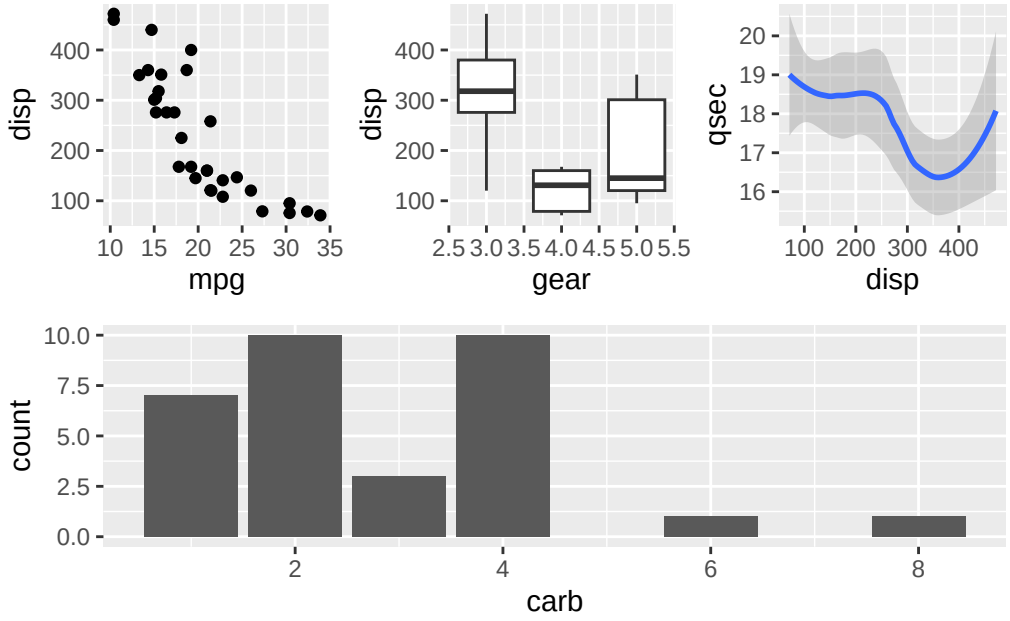
Before I can use any functions from add-on packages, I need to load the package from my "library()" with the `library(patchwork)` call.

```
library(patchwork)
```

From Class 5 Worksheet:

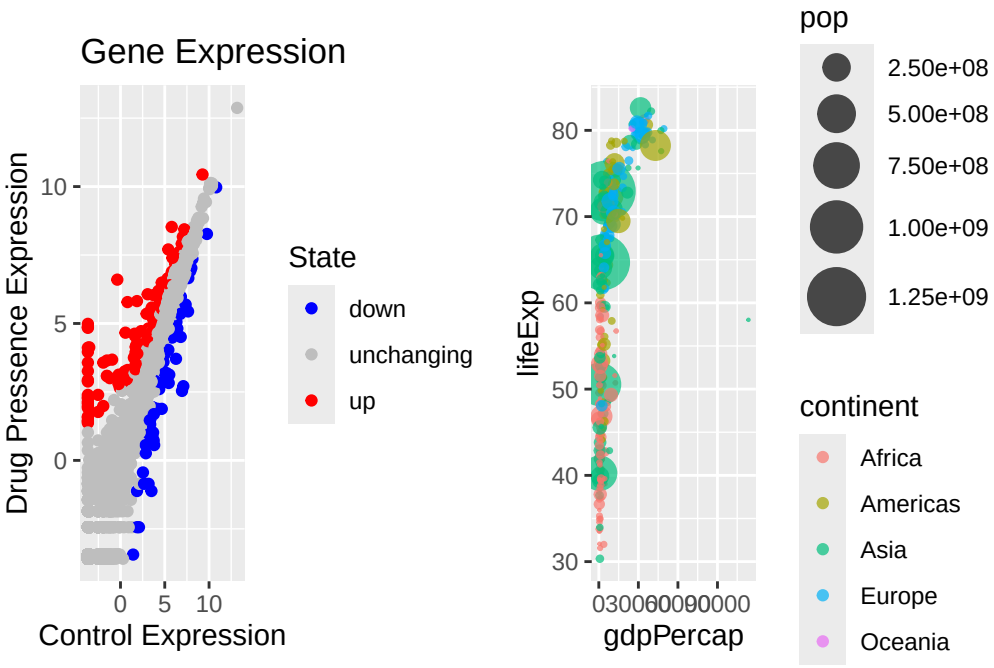
```
# Setup some example plots
p1 <- ggplot(mtcars) + geom_point(aes(mpg, disp))
p2 <- ggplot(mtcars) + geom_boxplot(aes(gear, disp, group = gear))
p3 <- ggplot(mtcars) + geom_smooth(aes(disp, qsec))
p4 <- ggplot(mtcars) + geom_bar(aes(carb))

# Use patchwork to combine them here:
(p1 | p2 | p3) /
      p4
```

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



To combine plots into a multi-panel figure:

```
p1<- p + scale_colour_manual(values=c("blue", "gray", "red")) + labs(title="Gene Expression"
p2<- ggplot(gapminder_1957) +
  aes(x=gdpPercap, y=lifeExp, size=pop, color=continent) +
  geom_point(alpha=0.7) +
  scale_size_area(max_size=10)

#Use patchwork to combine them
(p1|p2)
```


Gene Expression
Drug Pressence Expression
Control Expression
State
down
unchanging
up
lifeExp
gdpPercap
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
continent
Africa
Americas
Asia
Europe
Oceania