Class 05: Data Visualization with ggplot

Hetian Su

Scatter Plots

Every ggplot contain at least 3 layers:

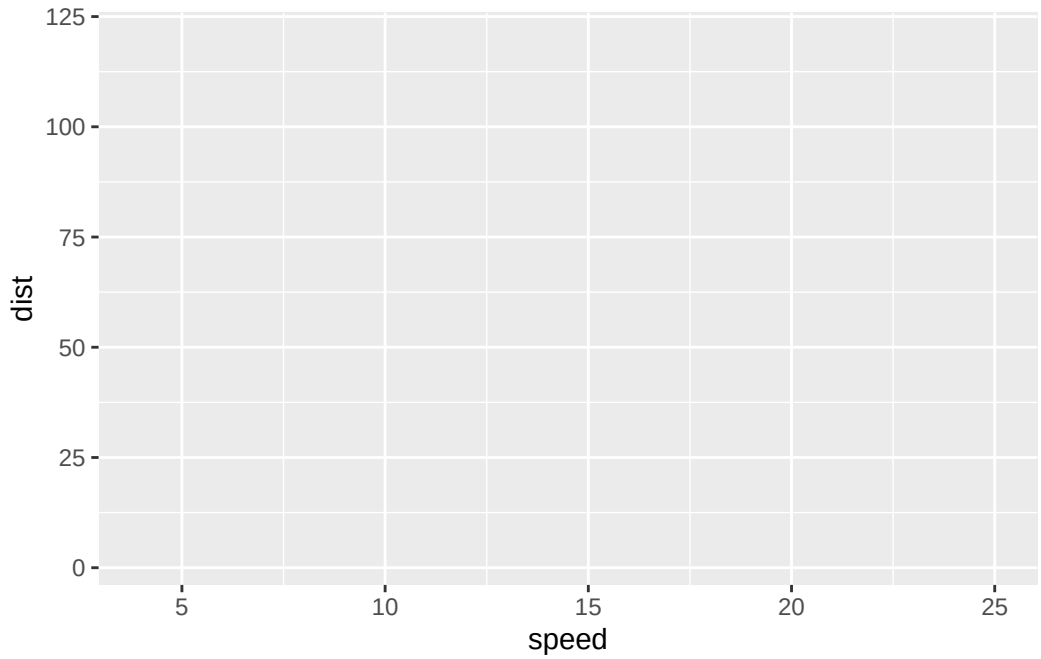
Data the data frame input

Aes the asesthetic features to map to the plot

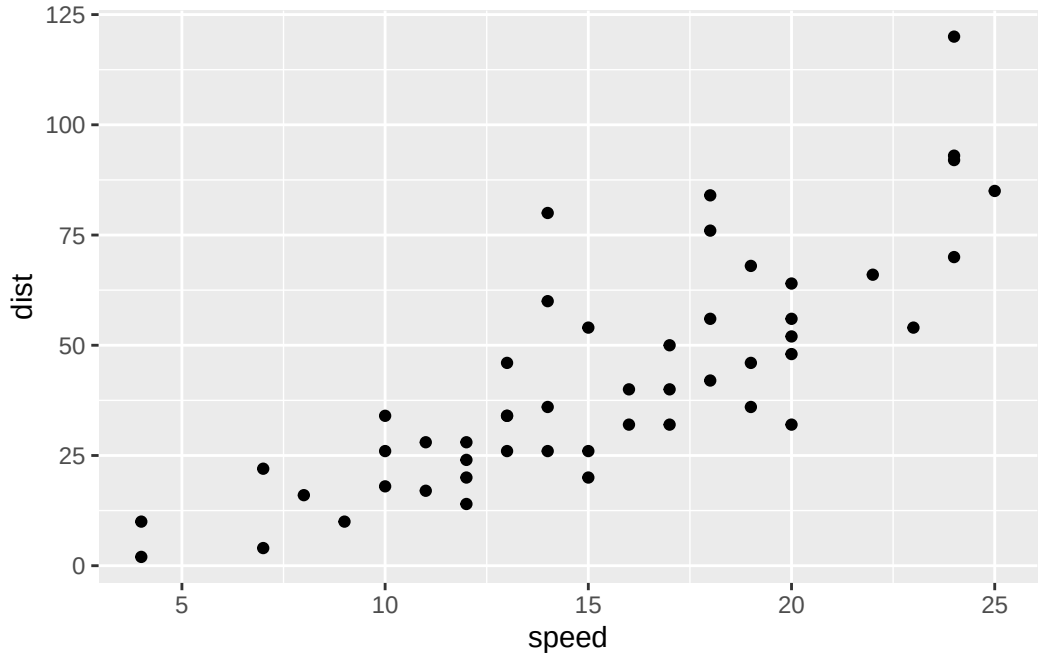
Geom the way of data representation

```
#load ggplot
# install.packages('ggplot2')
library(ggplot2)

ggplot(cars)
```

```
#map distance and speed to aesthetics
ggplot(data=cars)+
  aes(speed, dist)
```

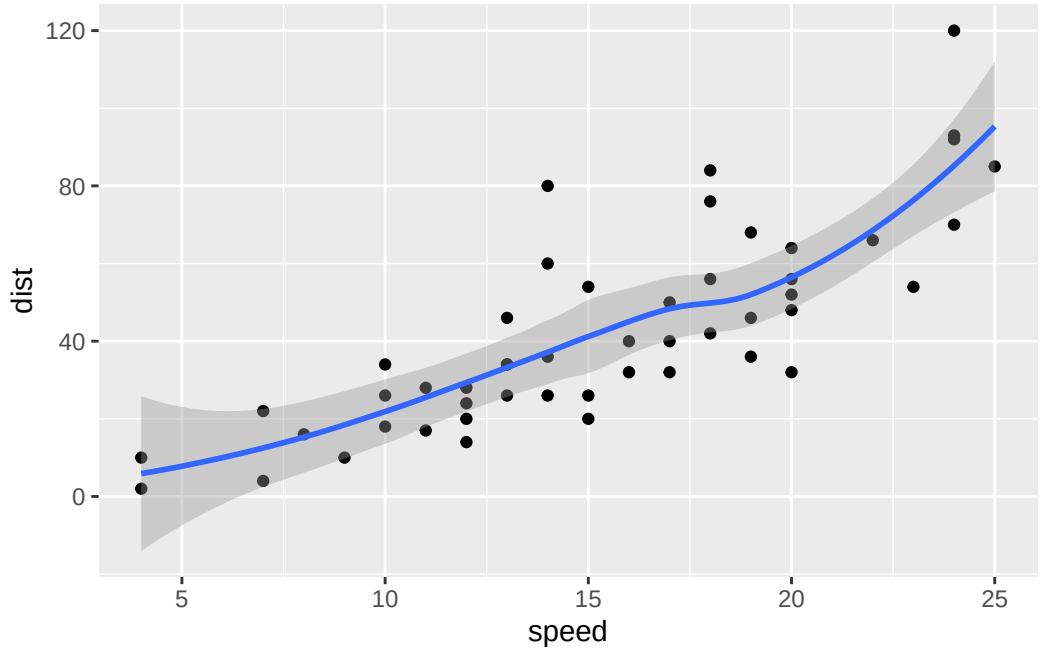



```
#Now specify a goem layer
ggplot(cars)+
  aes(speed, dist)+
  geom_point()
```



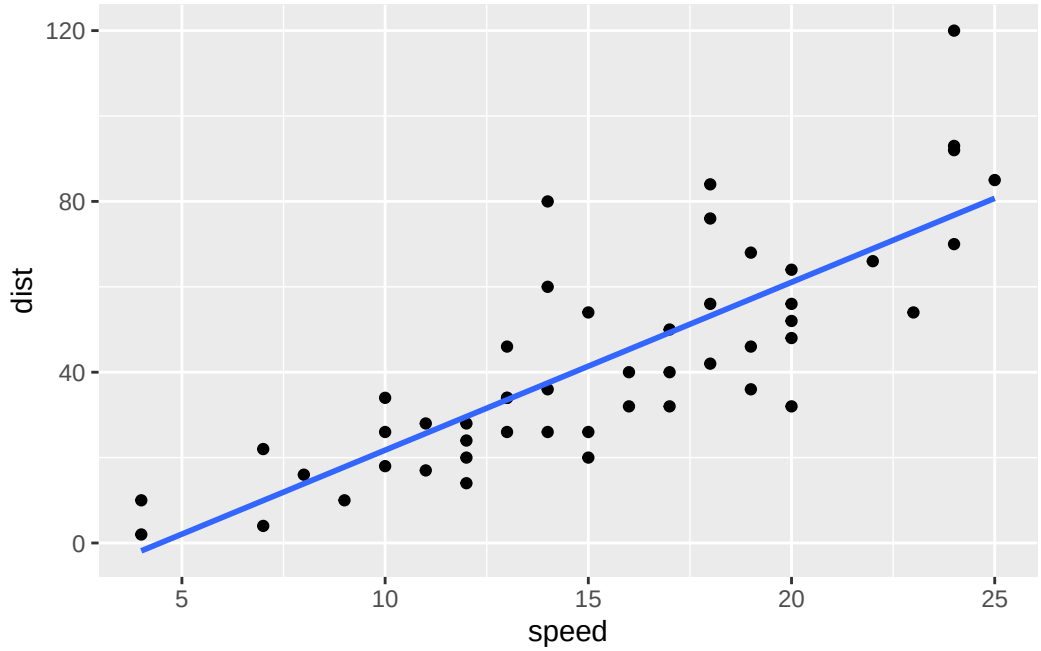

```
#add a trend line layer
ggplot(cars)+
  aes(speed, dist)+
  geom_point()+
  geom_smooth()
```

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



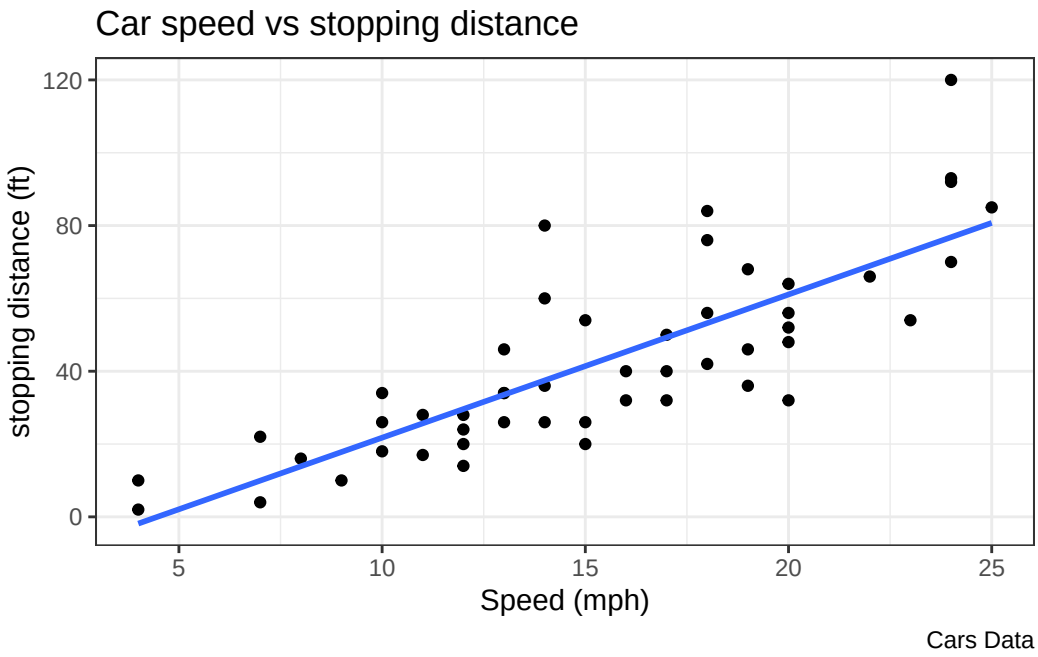

```
#use trend line without standard errors
ggplot(cars)+
  aes(speed, dist)+
  geom_point()+
  geom_smooth(method='lm', se=FALSE)
```

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




```
#add labels and change theme to black/white
ggplot(cars)+
  aes(speed, dist)+
  geom_point()+
  geom_smooth(method='lm', se=FALSE)+
  theme_bw()+
  labs(title='Car speed vs stopping distance', x='Speed (mph)', y='stopping distance (ft)'
```

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



More aesthetics

```
#load drug treatment and gene expression data
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
```

```
#explore the dataset
nrow(genes)
```

```
[1] 5196
```

```
colnames(genes)
```

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

```
ncol(genes)
```

```
[1] 4
```

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

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

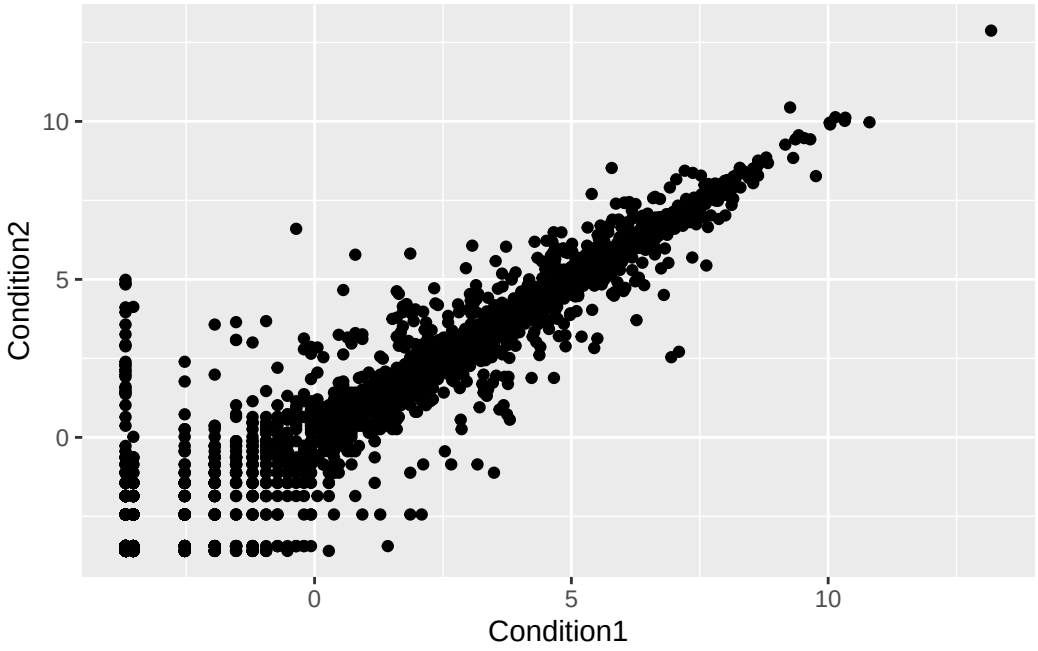
```
upFrac = round(127/5196*100, 2)
upFrac
```

```
[1] 2.44
```

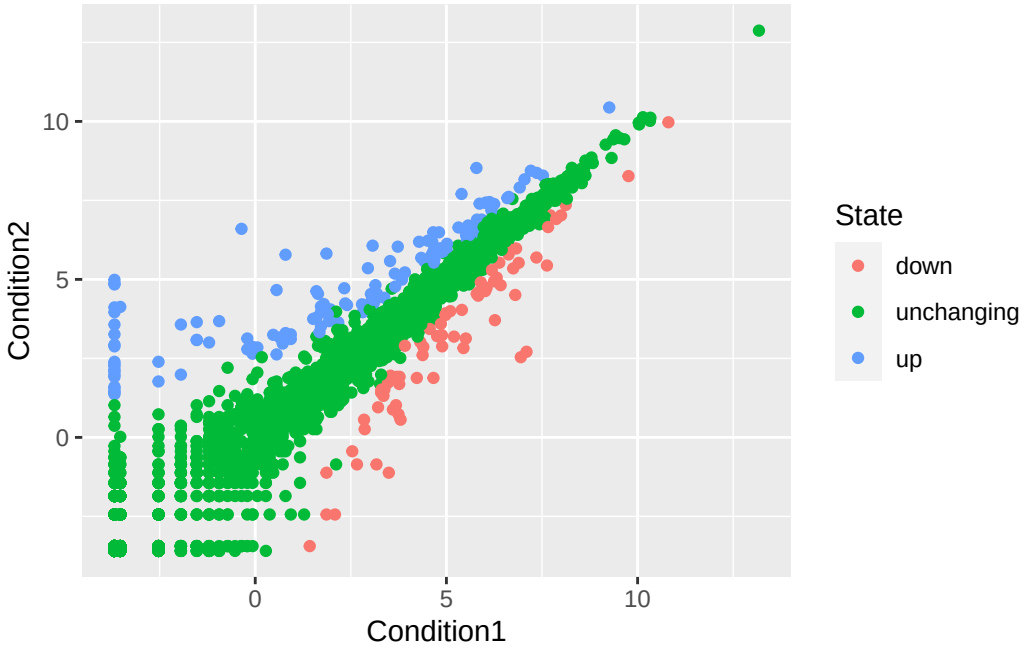
There are 5196 genes in the dataset, 4 columns, 2.44 % genes are upregulated.

Create the simple scatter plot for genes dataset

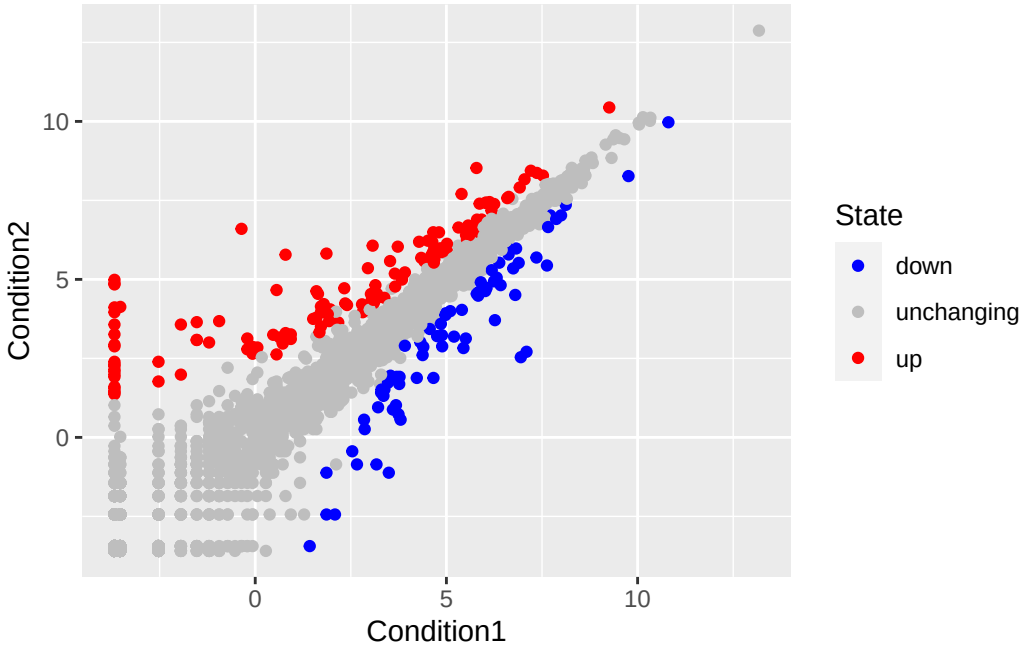
```
ggplot(genes)+
  aes(Condition1, Condition2)+
  geom_point()
```




```
#map color to State and store the plot object to p
p <- ggplot(genes)+
  aes(Condition1, Condition2, col=State)+
  geom_point()
p
```

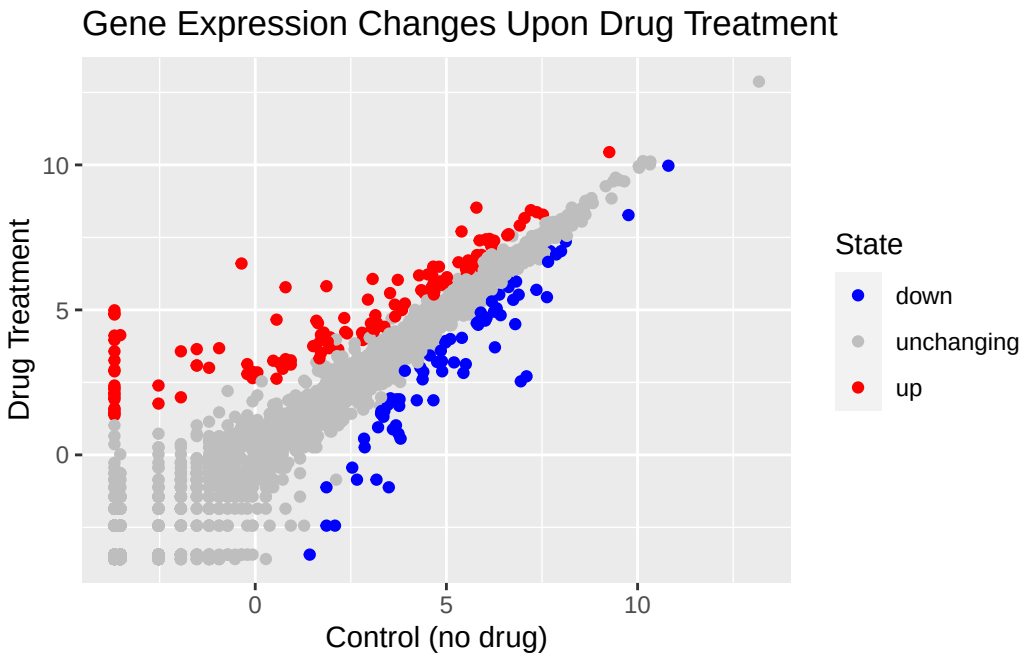



```
#manually define the color scale
p <- p + scale_color_manual(values = c('blue','grey','red'))
p
```



Add labels to the plot

```
p <- p + labs(title='Gene Expression Changes Upon Drug Treatment', x='Control (no drug)',y
p
```



Going Further

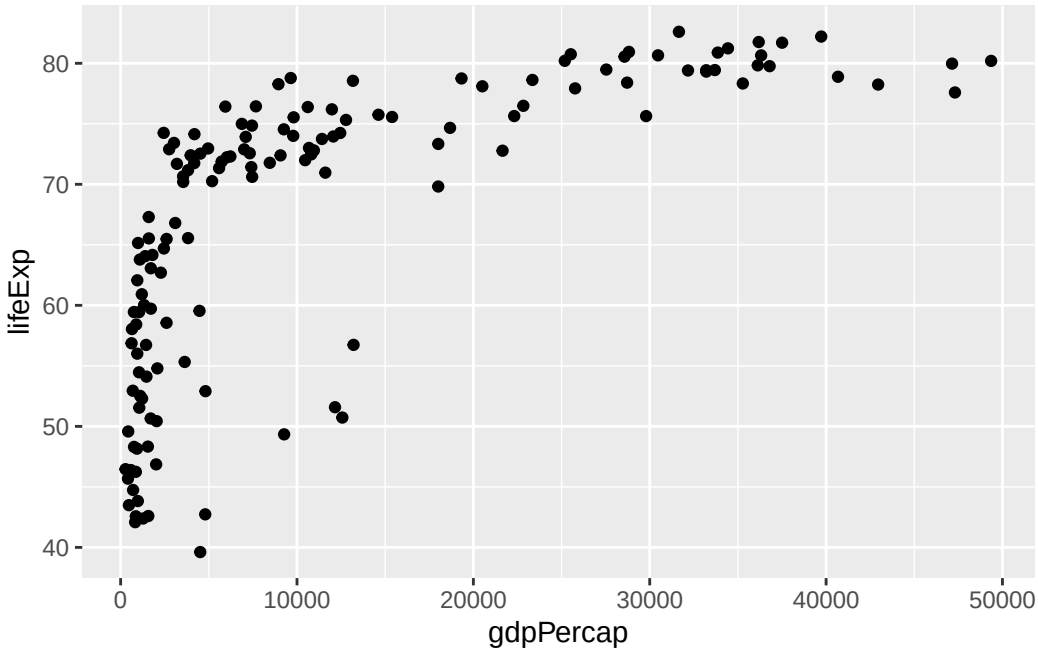
```
#load the gapminder dataset
url <- 'https://raw.githubusercontent.com/jennybc/gapminder/master/inst/extdata/gapminder.
gapminder <- read.delim(url)
```

```
# install.packages('dplyr')
library(dplyr)
```

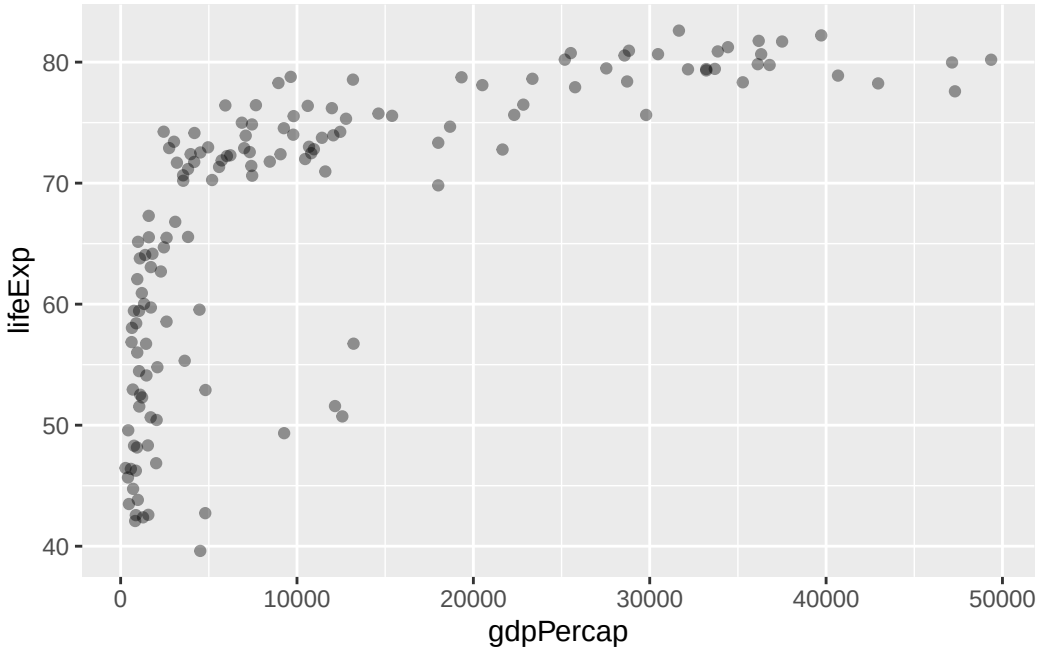
```
Attaching package: 'dplyr'

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)
```

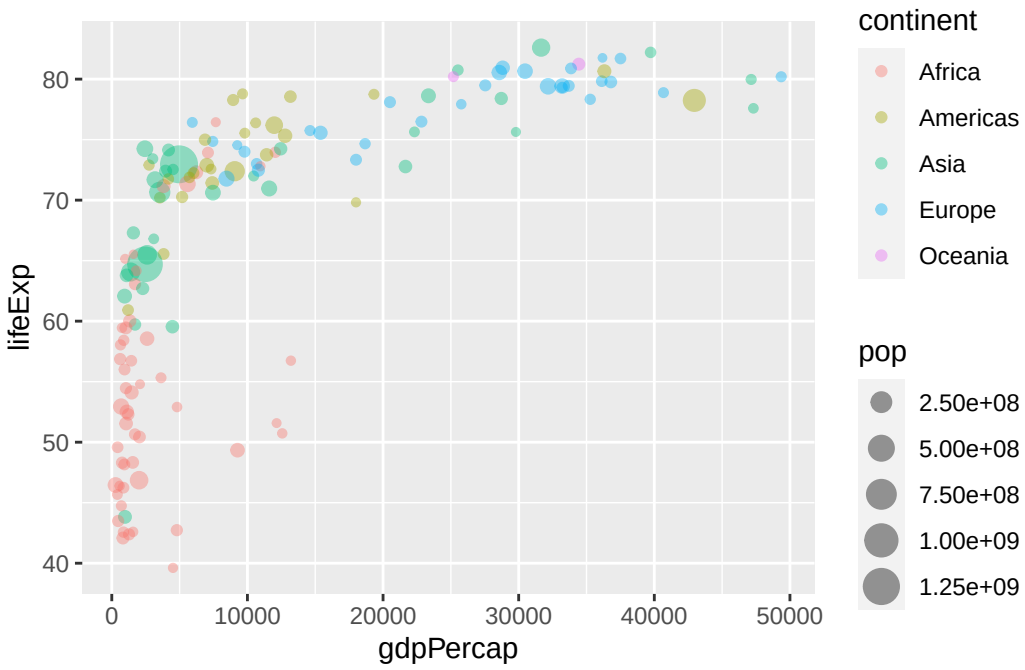
```r
#simple plot for lifeexp vs gdp per captia
ggplot(gapminder_2007)+
  aes(gdpPercap, lifeExp)+
  geom_point()
```



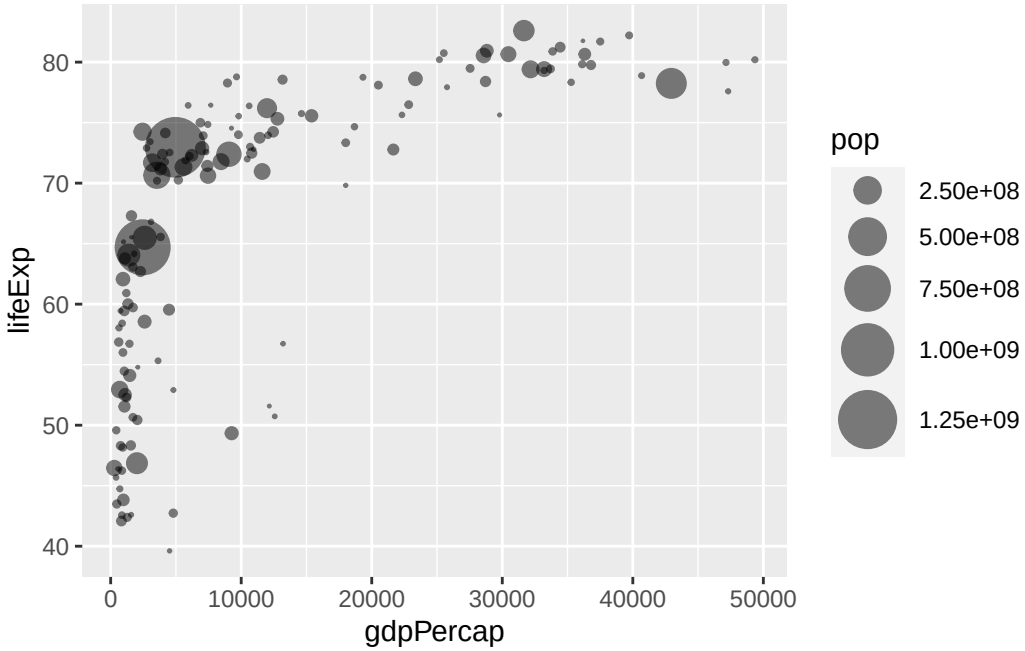

```r
#add transparency to points
ggplot(gapminder_2007)+
  aes(gdpPercap, lifeExp)+
  geom_point(alpha=0.4)
```




```
#map to continent and population size to aesthetics
ggplot(gapminder_2007)+
  aes(gdpPercap, lifeExp, col=continent, size=pop)+
  geom_point(alpha=0.4)
```

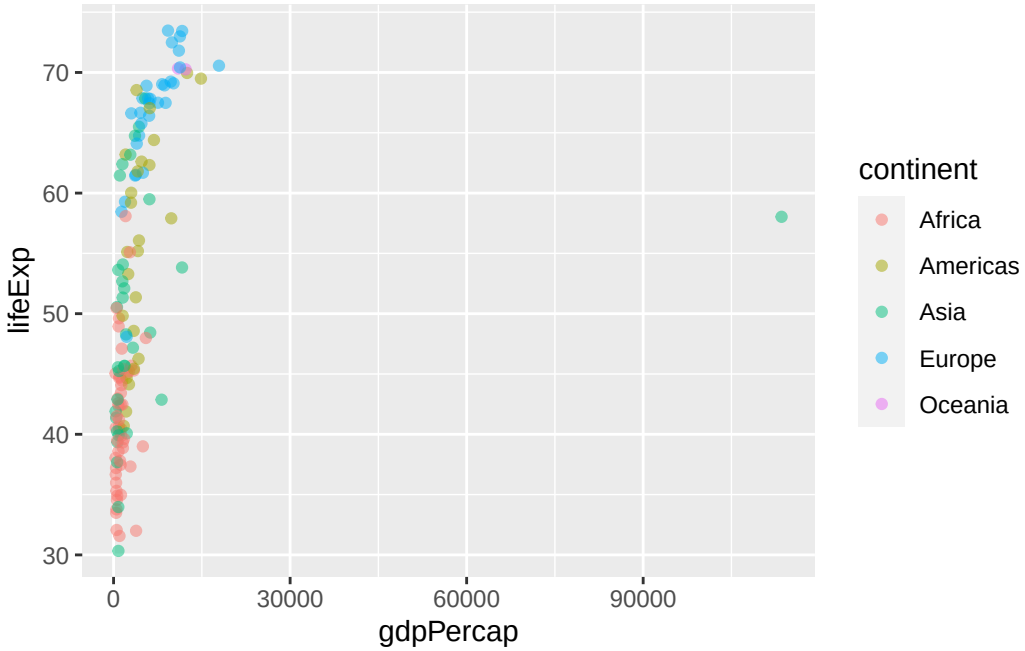



```
#adjust the point sizes to be proportional to actual population sizes
ggplot(gapminder_2007)+
  aes(gdpPercap, lifeExp, size=pop)+
  geom_point(alpha=0.5)+
  scale_size_area(max_size = 10)
```

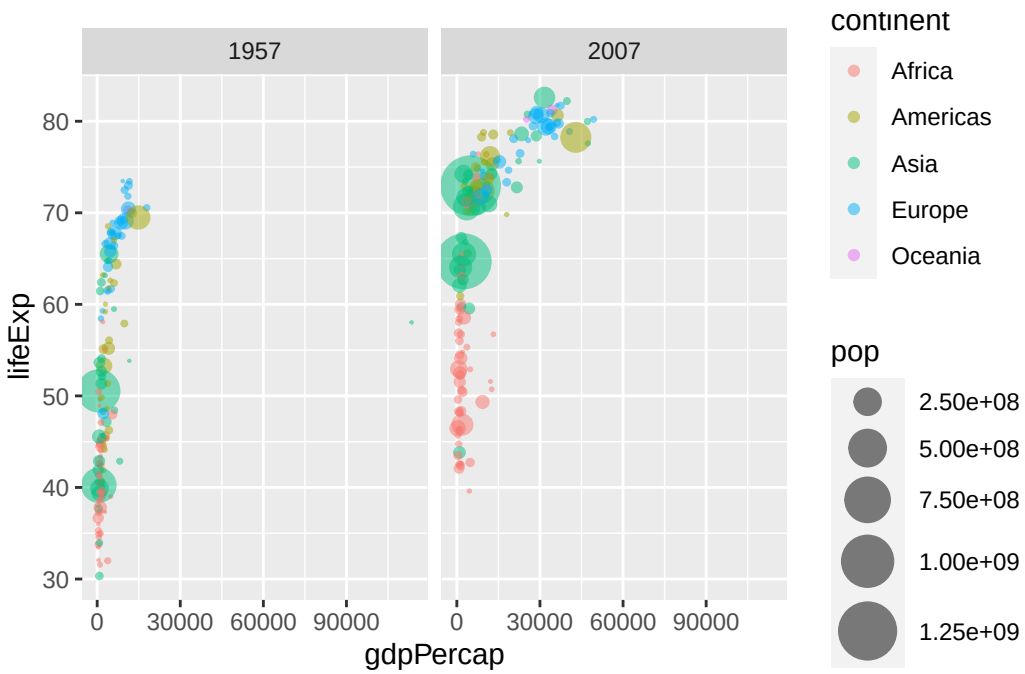


Create plot for gapminder 1957 and compare it to the plot of 2007

```
gapminder_1957 <- gapminder%>%filter(year==1957)
ggplot(gapminder_1957)+
  geom_point(aes(gdpPercap, lifeExp, color=continent), alpha=0.5)+
  scale_size_area(max_size = 10)
```

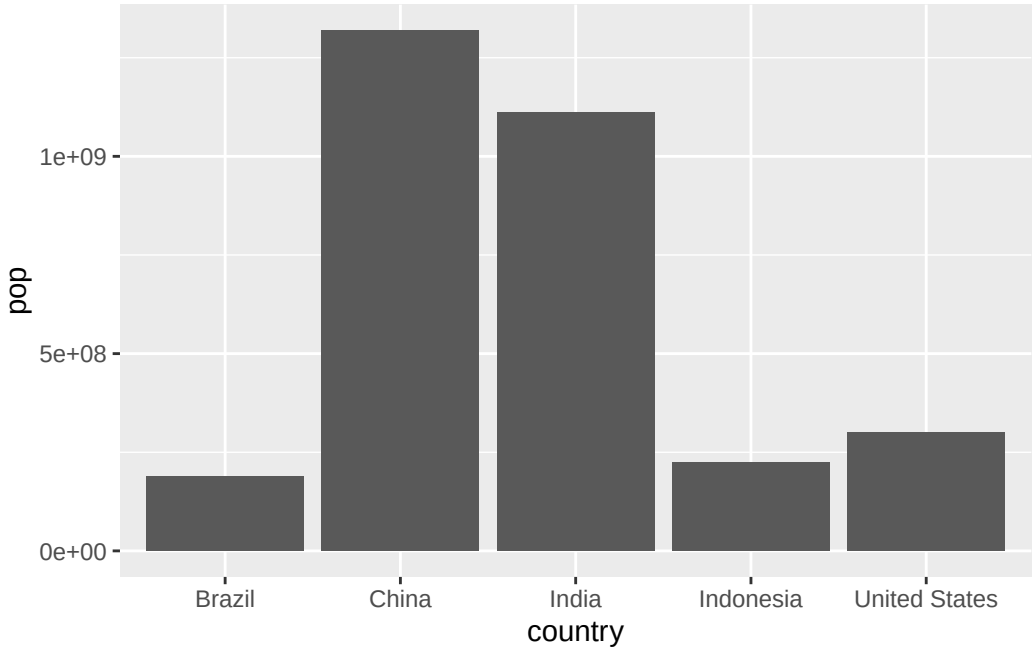



```
#create side by side plot for the 2 filtered datasets
gapminder_comb <- gapminder%>%filter(year==1957 | year==2007)

ggplot(gapminder_comb)+
  geom_point(aes(gdpPercap, lifeExp, color=continent, size=pop), alpha=0.5)+
  scale_size_area(max_size = 10)+
  facet_wrap(~year)
```

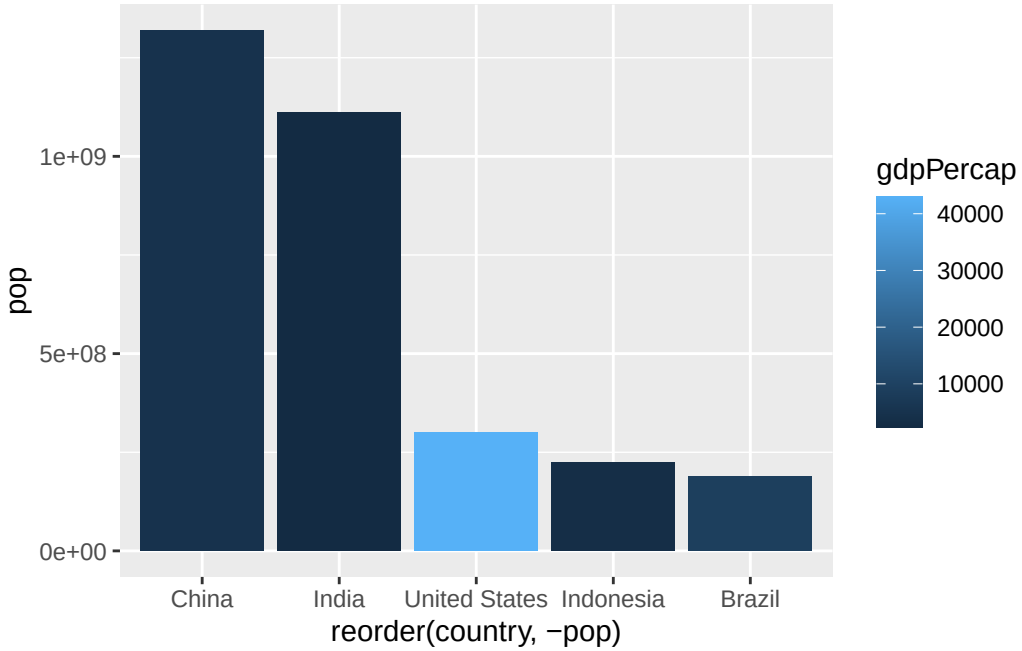


Box Plot

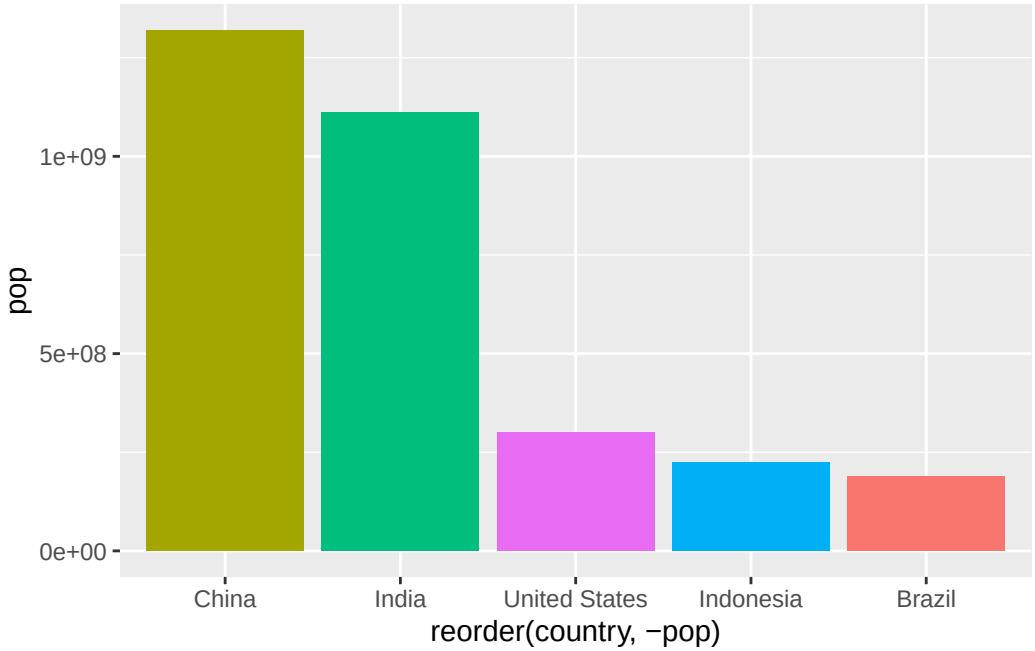
```
#boxplot for gapminder of 5 biggest countries in 2007
gapminder_top5 <- gapminder%>%filter(year==2007)%>%
  arrange(desc(pop))%>%
  top_n(5, pop)

ggplot(gapminder_top5, aes(country, pop))+
  geom_col()
```




```
#plot country vs population and scale color with gdp
ggplot(gapminder_top5)+
  geom_col(aes(reorder(country,-pop), pop, fill=gdpPercap))
```




```
#fill by country
ggplot(gapminder_top5)+
  geom_col(aes(reorder(country,-pop), pop, fill=country))+
  guides(fill='none')
```


pop
1e+09
5e+08
0e+00
China
India
United States
Indonesia
Brazil
reorder(country, −pop)