Data Visualization with ggplot2 (Multivariate)

Learning Objectives

1. Summarize three categorical variables using bar charts
2. Summarize more than two quantitative variables using scatterplots

We will be using the `gapminder` dataset, from the `gapminder` package. Install and load the `gapminder` package. Also load the `tidyverse` package (which automatically loads the `ggplot2` package).

```
library(tidyverse)
library(gapminder)
```

We are going to create two data frames from the original `gapminder` data frame for the examples:

1. `Data` which only contains the year 2007, with an additional binary variable `expectancy`, which is `low` if the country's life expectancy is less than 70 years, and `high` otherwise.
2. `Data.all` which contains data from all years, with the additional binary variable `expectancy`.

```
Data<-gapminder%>%
  mutate(expectancy=ifelse(lifeExp<70,"Low","High"))%>%
  filter(year==2007)

Data.all<-gapminder%>%
  mutate(expectancy=ifelse(lifeExp<70,"Low","High"))
```

1. Summarize three categorical variables using bar charts

Previously, we created a bar chart to look at how `expectancy` varies across the continents.

```
ggplot(Data,aes(x=continent, fill=expectancy))+
  geom_bar(position = "fill")
```

Suppose we want to see how these bar graphs vary across the years

```
ggplot(Data.all,aes(x=continent, fill=expectancy))+
  geom_bar(position = "fill")+
  facet_wrap(~year)
```

Notice that three categorical variables are summarized in this bar chart. Is there something that can be done to improve this bar chart? How would you make this improvement?

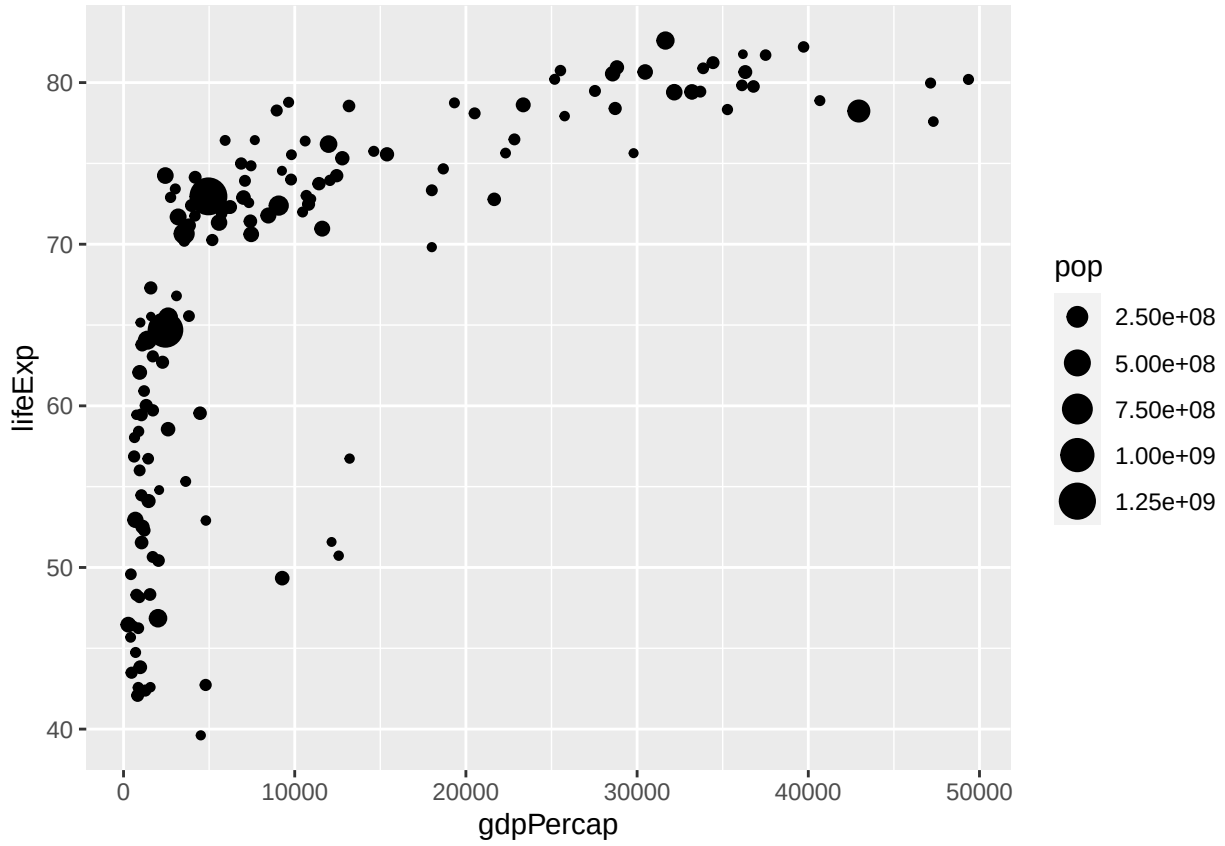
As mentioned earlier, since `Year` is a discrete variable, we can use graphical summaries that are meant for categorical variables with `Year`.

2. Summarize more than two quantitative variables using scatterplots

Previously, we created a scatterplot of life expectancy against GDP per capita. We can include another quantitative variable in the scatterplot, by using the size of the plots. We

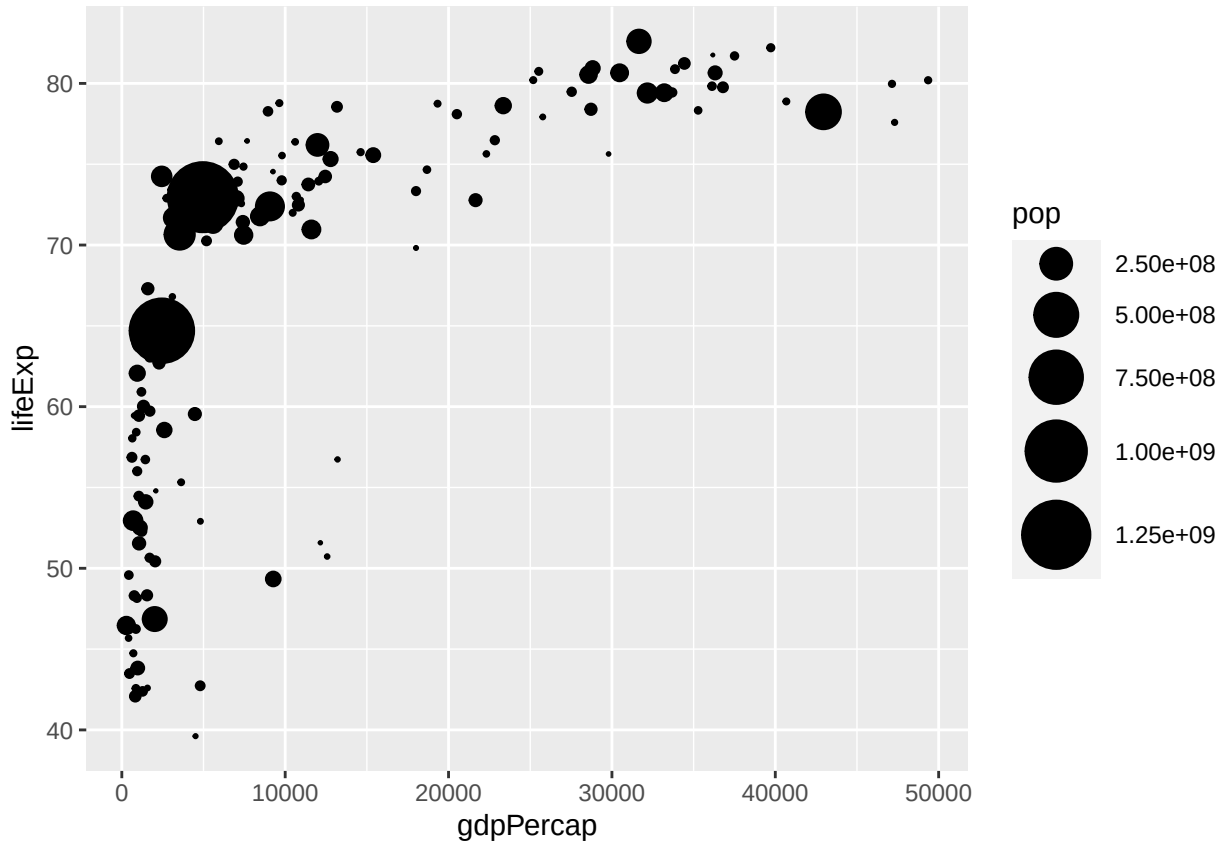
can use the size of the plots to denote the population size of the countries. This is supplied via `size` in `aes()`

```
ggplot(Data, aes(x=gdpPercap, y=lifeExp, size=pop))+
  geom_point()
```



We can adjust the size of the plots by adding a layer `scale_size()`

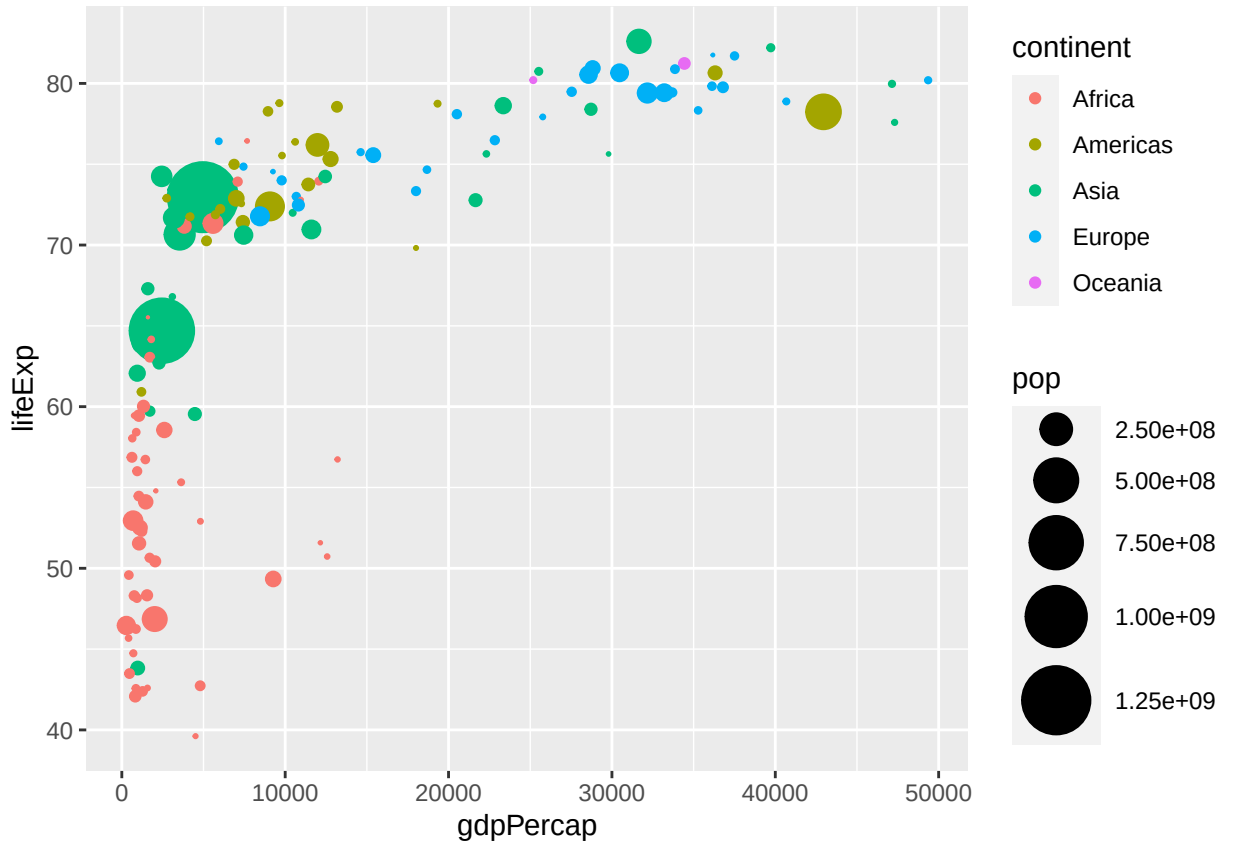
```
ggplot(Data, aes(x=gdpPercap, y=lifeExp, size=pop))+
  geom_point()+
  scale_size(range = c(0.1,12))
```



This scatterplot summarizes three quantitative variable.

We can use different-colored plots to denote which continent each point belongs to

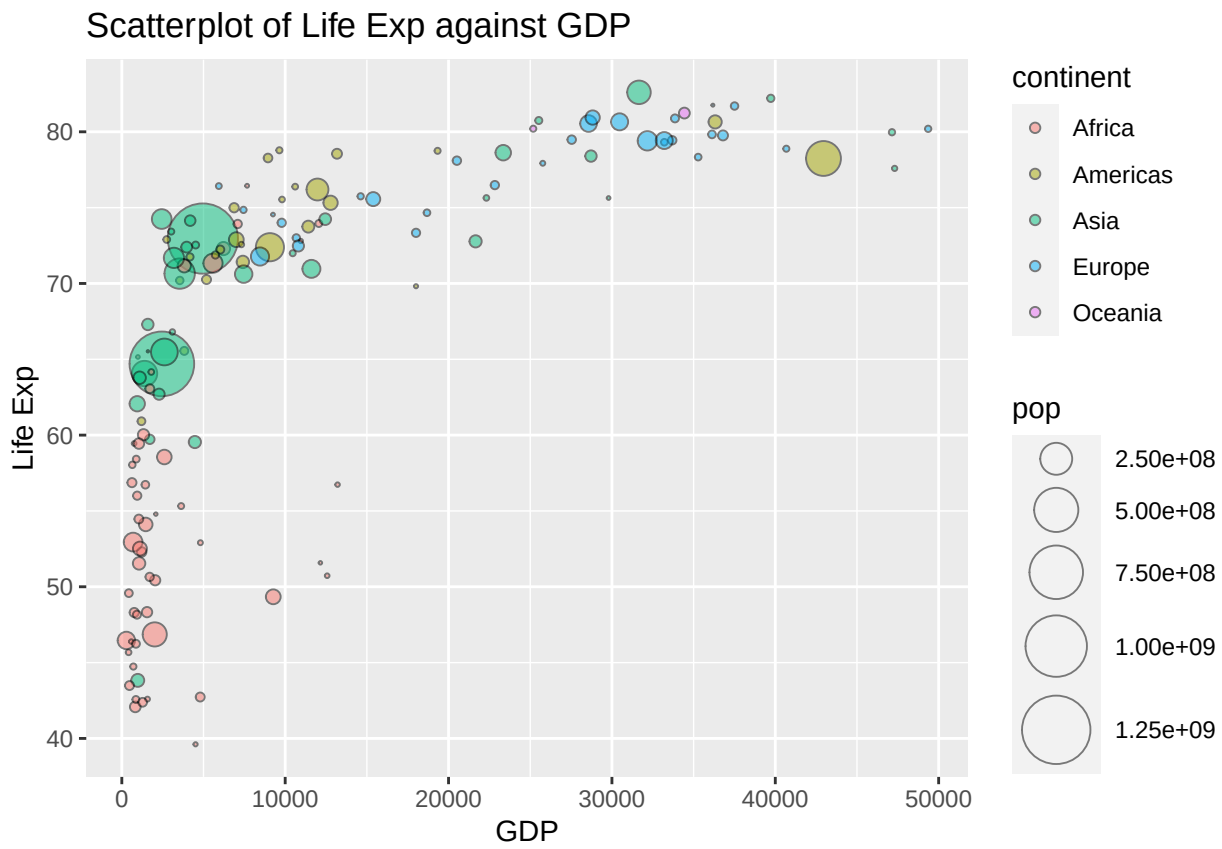
```
ggplot(Data, aes(x=gdpPercap, y=lifeExp, size=pop, color=continent))+
  geom_point()+
  scale_size(range = c(0.1,12))
```



This scatterplot summarizes three quantitative variables and one categorical variable.

We can adjust the plots by changing its shape and making it more translucent via `shape` and `alpha` in `aes()`

```
ggplot(Data, aes(x=gdpPercap, y=lifeExp, size=pop, fill=continent))+
  geom_point(shape=21, alpha=0.5)+
  scale_size(range = c(0.1,12))+
  labs(x="GDP", y="Life Exp", title="Scatterplot of Life Exp against GDP")
```


Scatterplot of Life Exp against GDP
Life Exp
GDP
80
70
60
50
40
0
10000
20000
30000
40000
50000
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09