Data Visualization Project

Tusha Karnani

Dataset 1

For this data visualization project, I imported a data set containing up/downregulation data on 5000+ genes.

```
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
```

```
library(ggplot2)
```

How many genes are in this dataset?

5196

How many upregulated genes are there?

127

```
nrow(genes)
```

```
[1] 5196
```

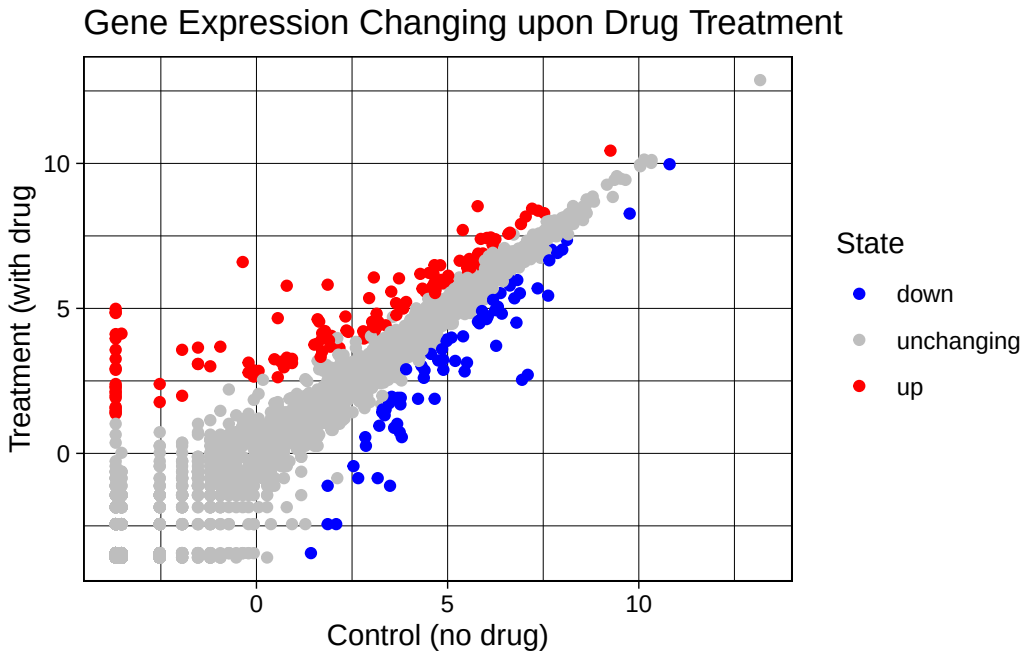
```
sum(genes$State=="up")
```

```
[1] 127
```

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

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

```
ggplot(data=genes) +
  aes(x=Condition1, y=Condition2, col=State) +
  geom_point() +
  scale_colour_manual( values=c("blue","gray","red") ) +
  labs(title="Gene Expression Changing upon Drug Treatment", x="Control (no drug)", y="Treatr
  theme_linedraw()
```



Dataset 2

I then imported a data set containing data on the life expectancy, population, and GDP per capita of 142 countries across 12 years.

```
url <- "https://raw.githubusercontent.com/jennybc/gapminder/master/inst/extdata/gapminder.tsv
gapminder <- read.delim(url)
```

Data on all 142 countries from 2007

```
library(dplyr)
```

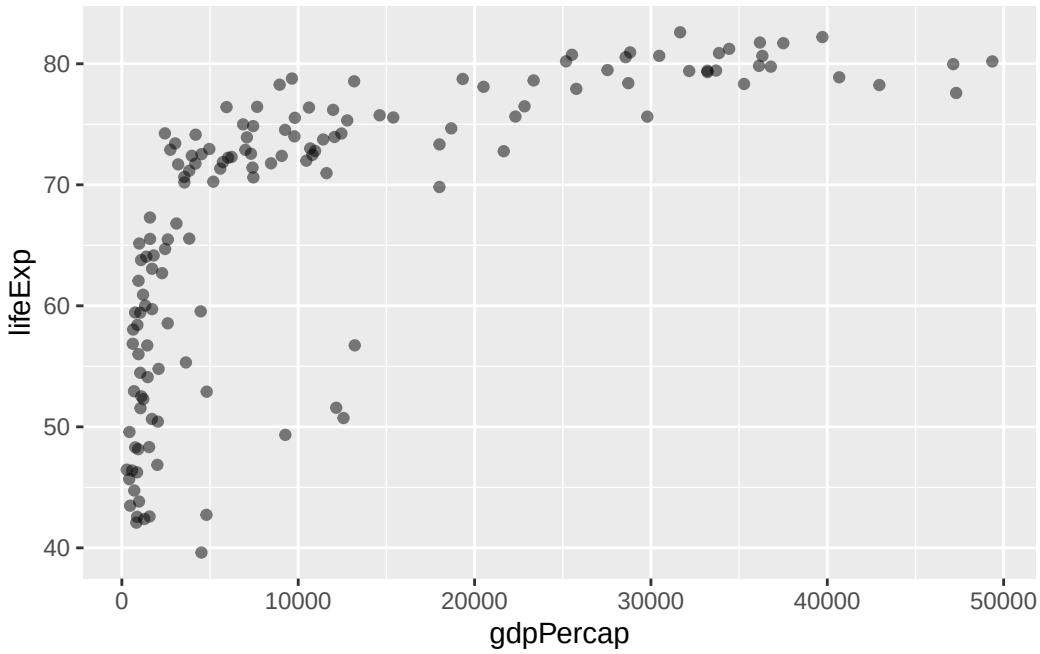
```
Attaching package: 'dplyr'

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)
head(gapminder_2007)
```

```
      country continent year lifeExp      pop  gdpPercap
1 Afghanistan      Asia 2007  43.828 31889923   974.5803
2     Albania    Europe 2007  76.423  3600523  5937.0295
3     Algeria    Africa 2007  72.301 33333216  6223.3675
4      Angola    Africa 2007  42.731 12420476  4797.2313
5   Argentina  Americas 2007  75.320 40301927 12779.3796
6   Australia   Oceania 2007  81.235 20434176 34435.3674
```

Initialization exploratory visualization

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




```
# alpha makes the dots more transparent
```

How many countries are in this dataset?

142

How many continents are in this dataset?

5

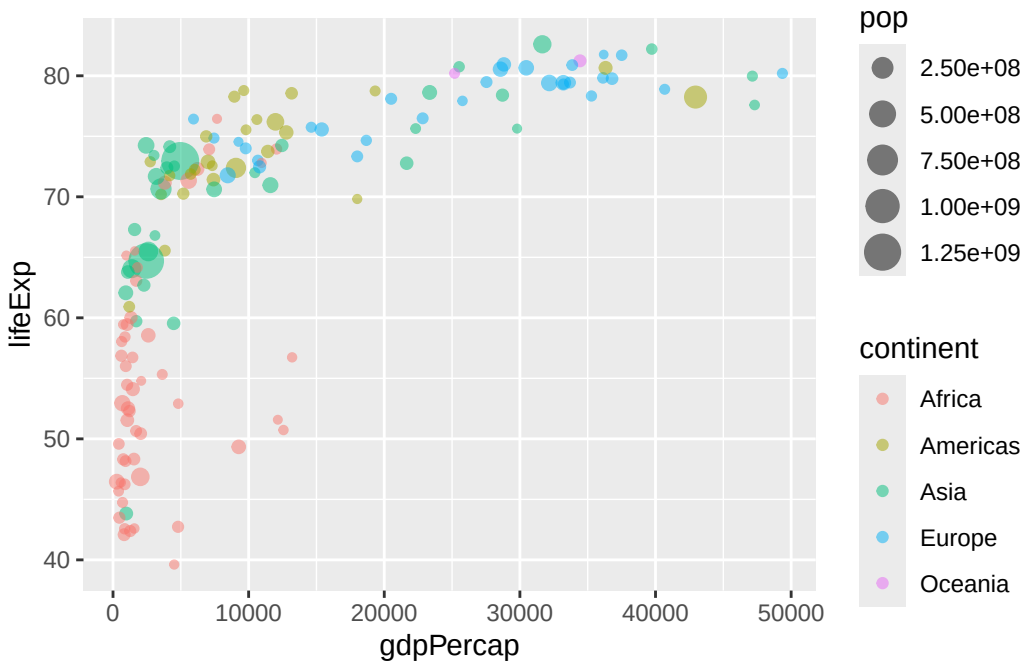
```
length(table(gapminder$country))
```

```
[1] 142
```

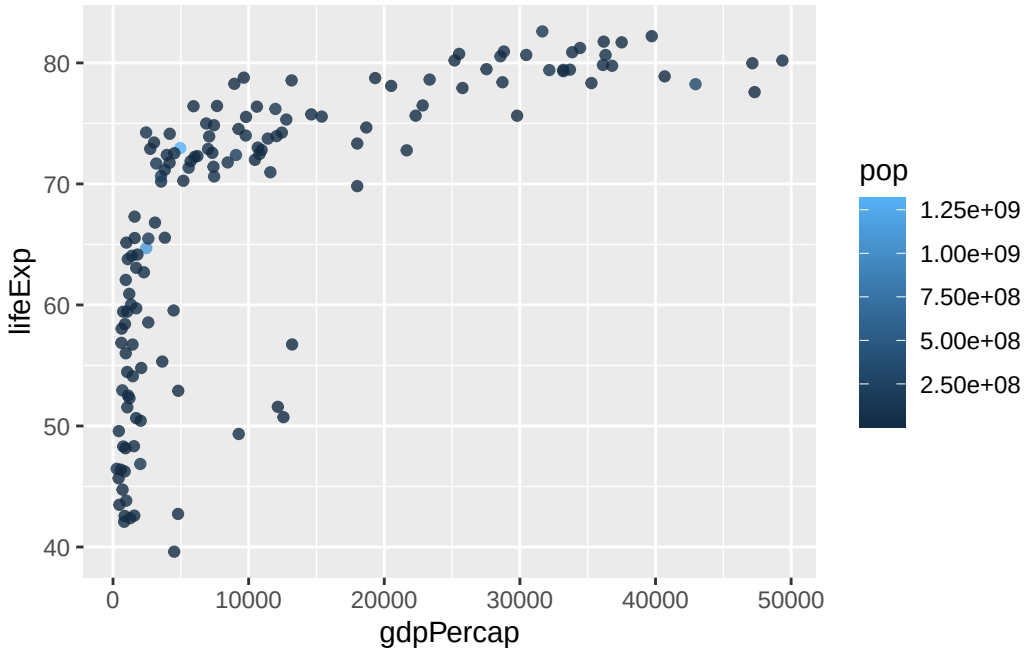
```
length(table(gapminder$continent))
```

```
[1] 5
```

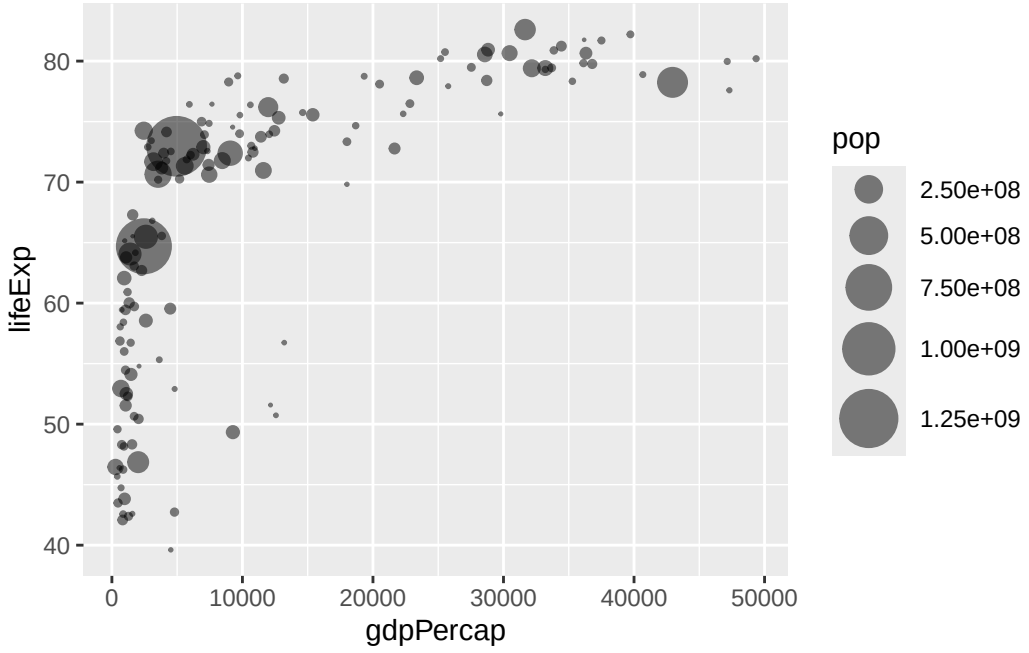
```
ggplot(gapminder_2007) +
  aes(x=gdpPercap, y=lifeExp, color=continent, size=pop) +
  geom_point(alpha=0.5)
```



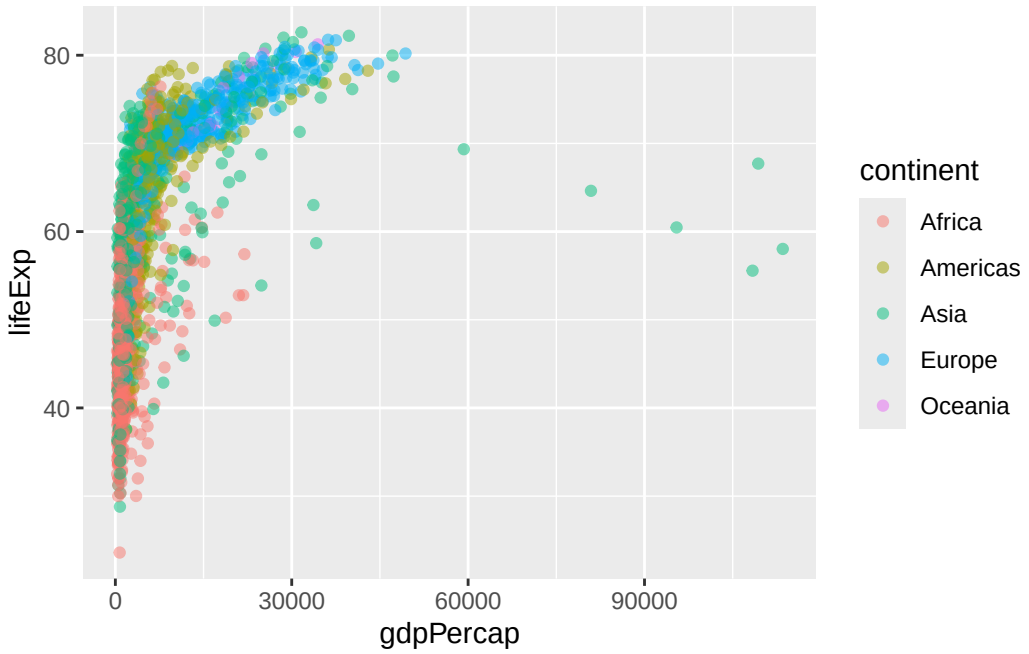

```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = pop) +
  geom_point(alpha=0.8)
```



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




```
ggplot(gapminder) +
  aes(x=gdpPercap, y=lifeExp, col=continent) +
  geom_point(alpha=0.5)
```

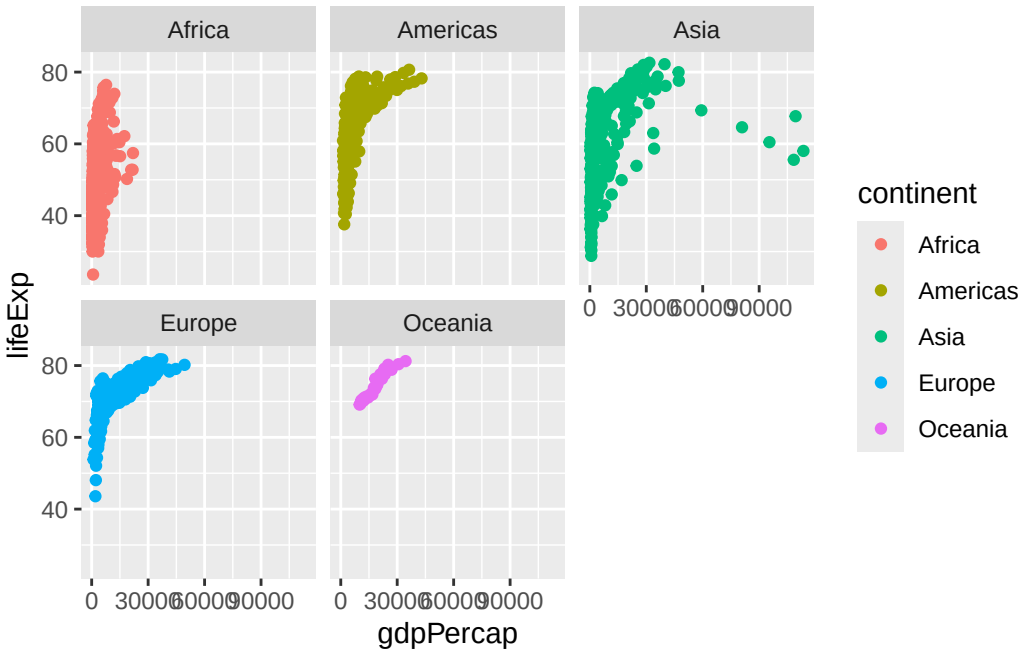


Data across all years for Kuwait

```
kuw <- gapminder %>% filter(country=="Kuwait")
kuw
```

```
   country continent year lifeExp     pop gdpPercap
1   Kuwait      Asia 1952  55.565  160000 108382.35
2   Kuwait      Asia 1957  58.033  212846 113523.13
3   Kuwait      Asia 1962  60.470  358266  95458.11
4   Kuwait      Asia 1967  64.624  575003  80894.88
5   Kuwait      Asia 1972  67.712  841934 109347.87
6   Kuwait      Asia 1977  69.343 1140357  59265.48
7   Kuwait      Asia 1982  71.309 1497494  31354.04
8   Kuwait      Asia 1987  74.174 1891487  28118.43
9   Kuwait      Asia 1992  75.190 1418095  34932.92
10  Kuwait      Asia 1997  76.156 1765345  40300.62
11  Kuwait      Asia 2002  76.904 2111561  35110.11
12  Kuwait      Asia 2007  77.588 2505559  47306.99
```

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp, col = continent, label = country) +
  geom_point() +
  facet_wrap(~continent)
```



Below are some advantages of ggplot over base R for data visualization:

1. **Layered Grammar of Graphics**: ggplot uses a consistent, layered approach. You build plots by adding layers (data, aesthetics, geoms, themes) with the + operator. This makes complex plots easier to construct and modify, compared to base R, where each plot type often requires different functions and arguments [1], [3], [2], [4], [5].
2. **Declarative Syntax**: You specify *what* you want to show (e.g., which variables map to axes, color, shape), not *how* to draw it. This makes code more readable and easier to maintain [1], [3], [2], [4], [5].
3. **Beautiful Defaults**: ggplot produces publication-quality figures with sensible defaults, so your plots look good without extensive tweaking. Base R gives you full control, but making plots look polished can be time-consuming [1], [3], [2], [4], [5].
4. **Faceting and Customization**: ggplot makes it easy to split data into multiple panels (facets) and customize aesthetics, legends, and themes. These features are more cumbersome in base R [3], [2].
5. **Extensibility**: ggplot is part of a large ecosystem of packages for advanced visualizations and customizations, making it more flexible for scientific work [1], [3], [2], [5].