king_dav_vsd_pset_01

Dav King

1/29/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

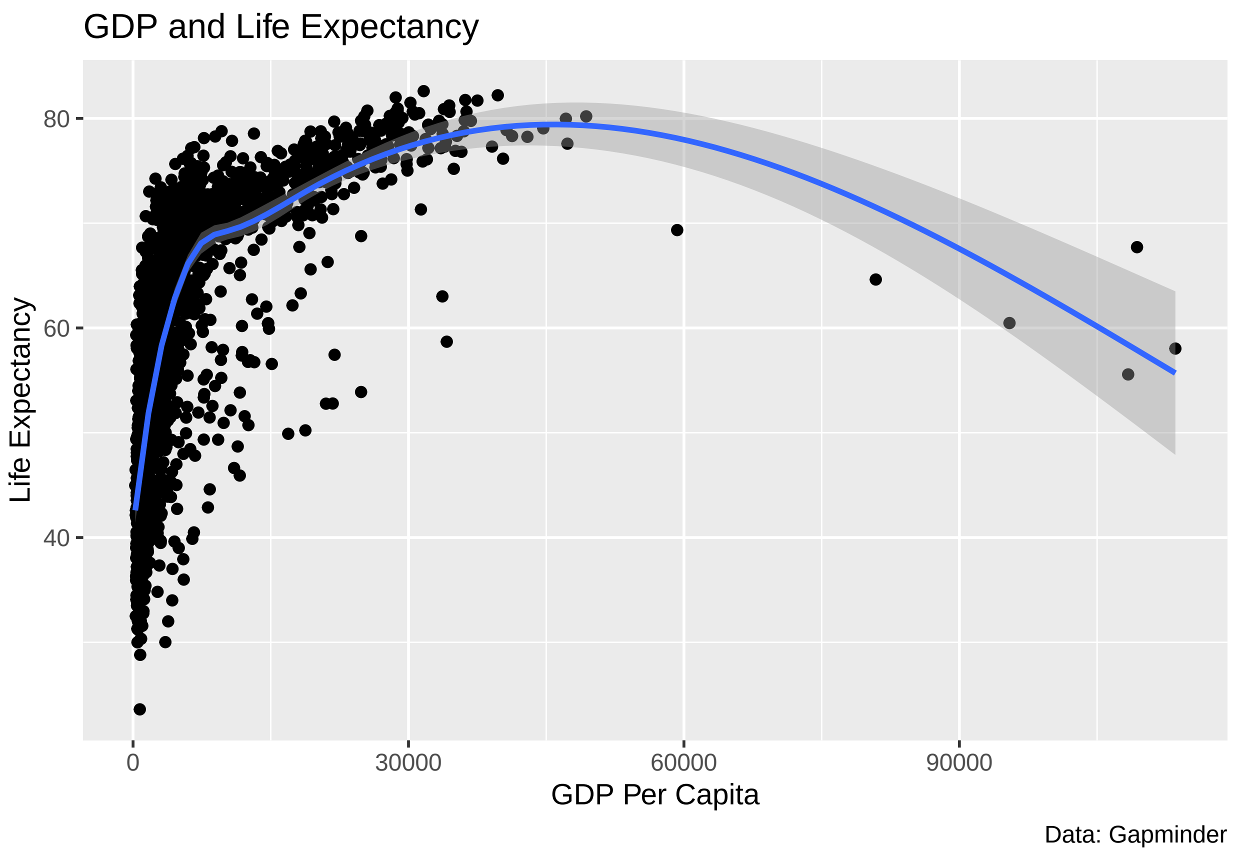
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gapminder)
library(palmerpenguins)
library(socviz)
```

```
p <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp))
```

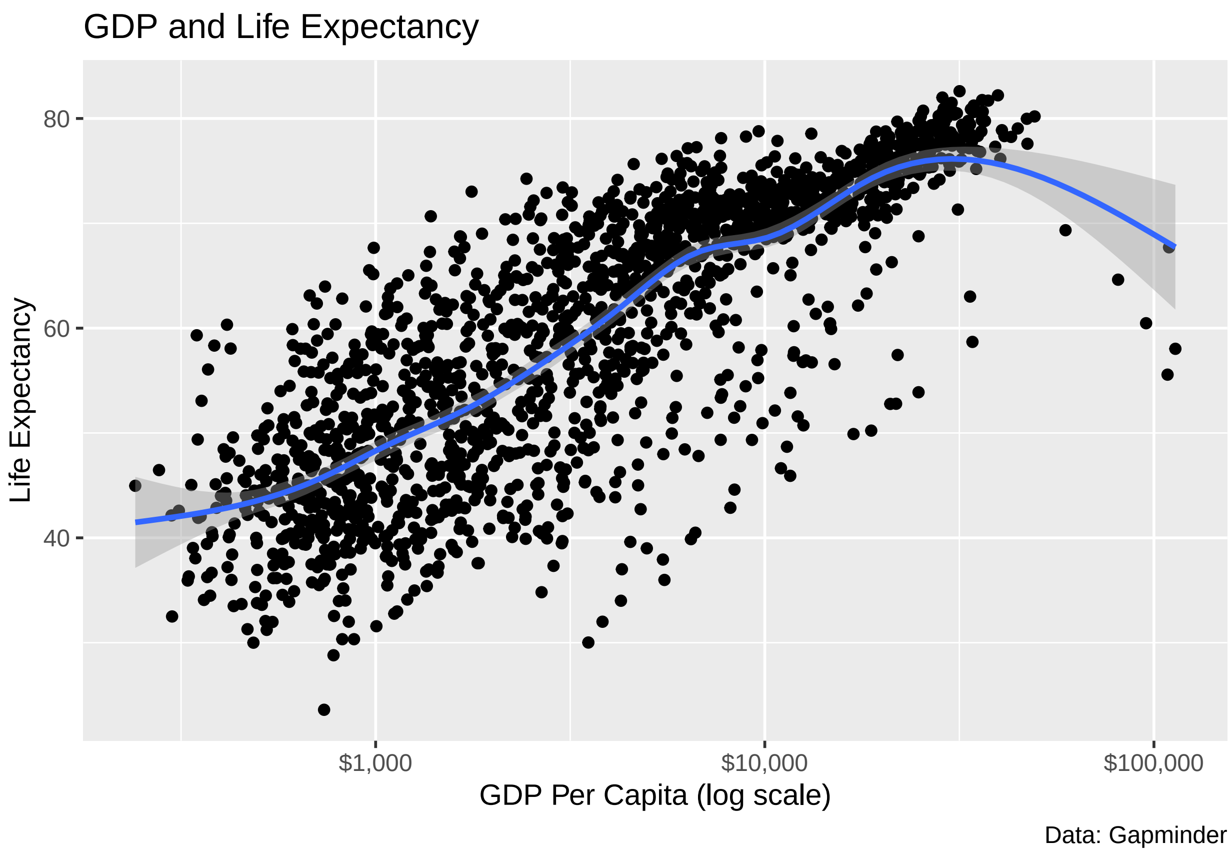
```
p + geom_point() +
  geom_smooth() +
  labs(title = "GDP and Life Expectancy", x = "GDP Per Capita",
       y = "Life Expectancy", caption = "Data: Gapminder")
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



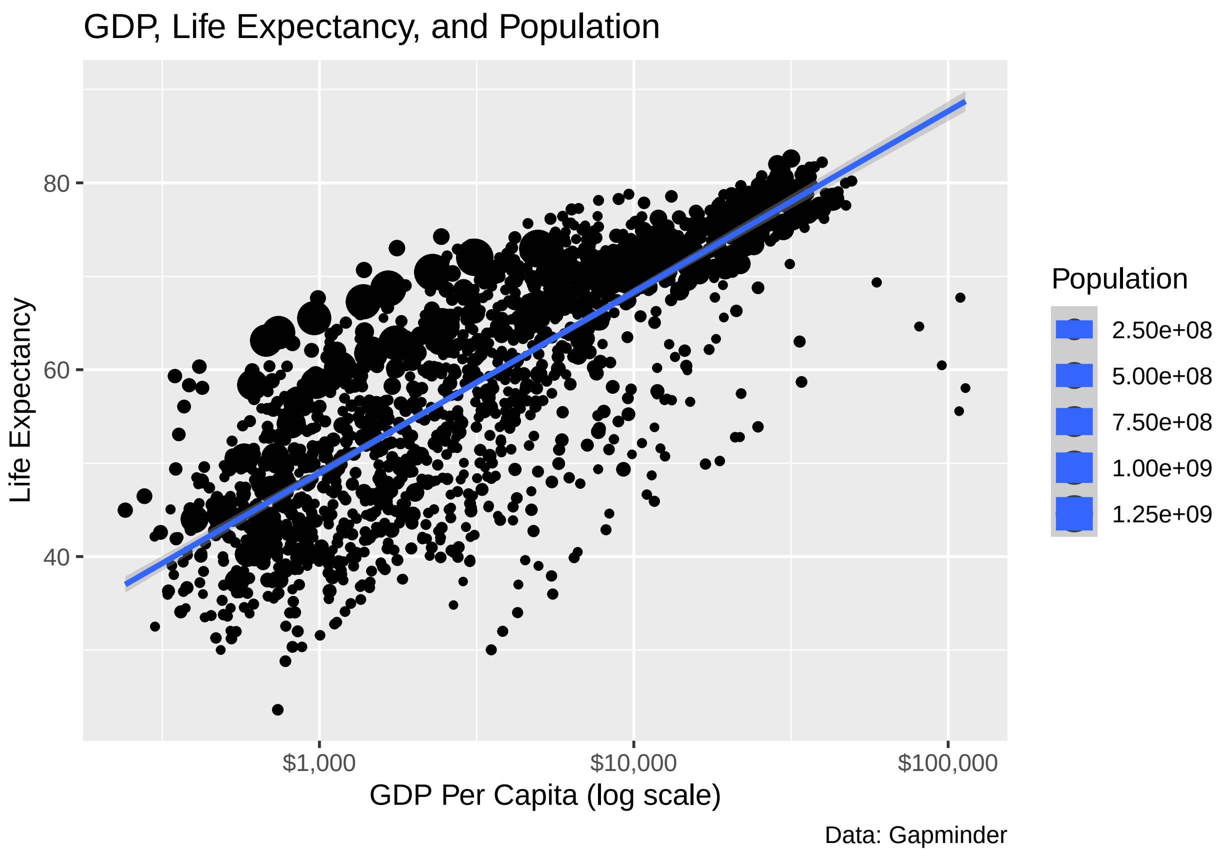
```
p + geom_point() +
  geom_smooth() +
  scale_x_log10(labels = scales::dollar_format()) +
  labs(title = "GDP and Life Expectancy",
       x = "GDP Per Capita (log scale)", y = "Life Expectancy",
       caption = "Data: Gapminder")
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



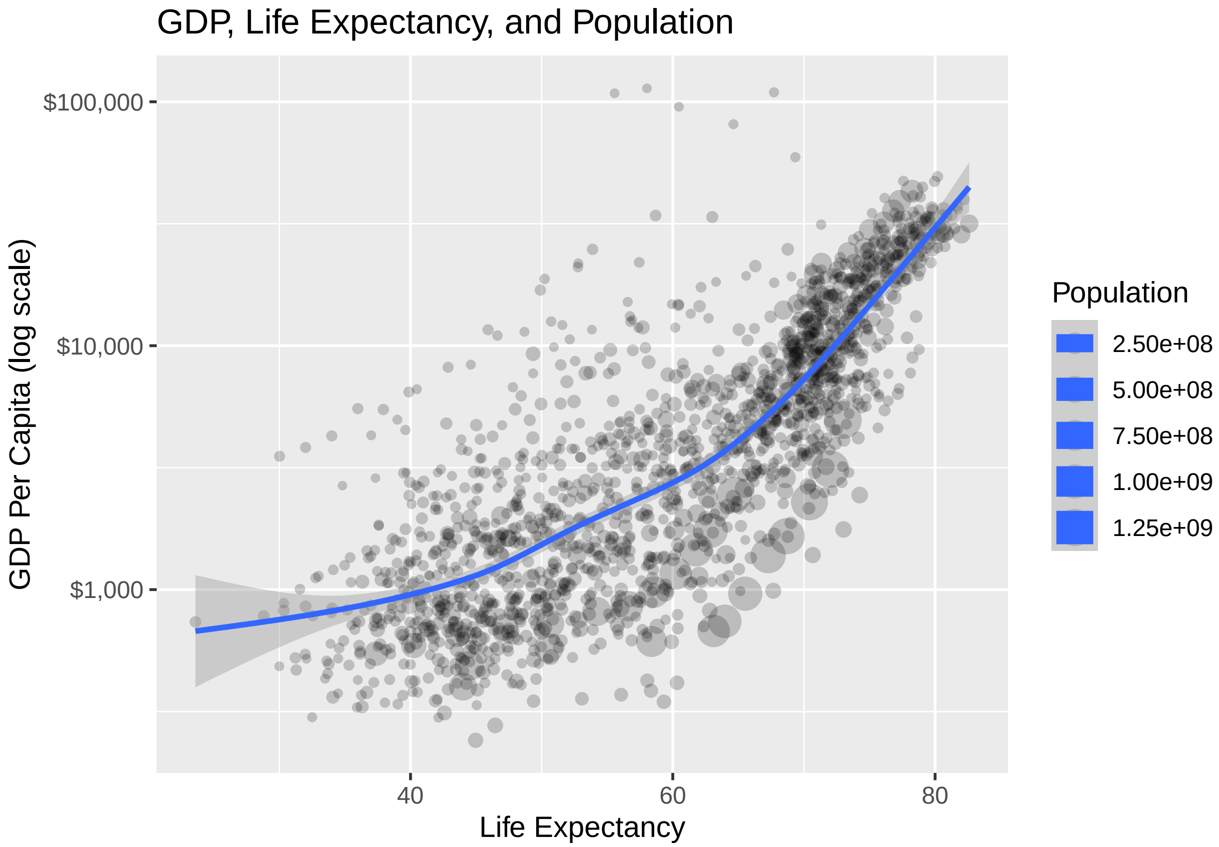
```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, size = pop)) + geom_point() +
  geom_smooth(method = lm) +
  scale_x_log10(labels = scales::dollar_format()) +
  labs(title = "GDP, Life Expectancy, and Population",
       x = "GDP Per Capita (log scale)", y = "Life Expectancy",
       size = "Population", caption = "Data: Gapminder")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
ggplot(gapminder, aes(x = lifeExp, y = gdpPercap, size = pop)) +
  geom_point(alpha = 0.2) +
  geom_smooth() +
  scale_y_log10(labels = scales::dollar_format()) +
  labs(title = "GDP, Life Expectancy, and Population",
       x = "Life Expectancy", y = "GDP Per Capita (log scale)",
       size = "Population")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Discussion

1

These plots draw from almost the same `ggplot(aes())` mappings. The first two use data `p`, which maps `gdpPercap` to x and `lifeExp` to y. The third uses these same two mappings, as well as mapping `pop` to size. The fourth does these same mappings, except the variables mapped to x and y are flipped.

2

To varying degrees, these plots do not meet Tufte's principles. To consider each in turn:

The representation of numbers should be directly proportional to the numerical quantities represented: while I am not directly measuring the area of the circles mapped to size in figures 3 and 4, given that they are a shape scaled radially by a variable, they likely are scaled in terms of area disproportionately to the numerical quantities of the data.

Clear, detailed, and thorough labeling: there is fairly consistent labeling across the plots, and it is especially good that plots for which this is relevant explicitly state that they are using a log scale, as well as denoting the numbers at each point on the scale. However, the scales for size are poor = they suggest five discrete values of population, which is not true, and they are hard to tell apart from one another and colored poorly (size should have been mapped within `geom_point()` and not `ggplot()`). Additionally, there are no explanations of the data on the graphs themselves, including explanations of outliers or trendlines.

Show data variation, not design variation: these graphs are good at that! Scales are consistent and there is a good ink-to-data ratio in these graphs.

Deflated and standardized units of monetary measurement: while these graphs do not explicitly depict time on them, they contain data points at different times - arguably even worse for potentially violating this criterion. Fortunately, these data are inflation-adjusted and denoted in USD, so they meet Tufte's criteria.

No more variable dimensions depicted than data dimensions: once again, we run into the same issue with the circles - scaled radially, but displayed areally. Area is multiple dimensions, but radius is only one - a clear problem. Other than that, there are no issues with dimensions of data in these scatterplots.

Don't quote data out of context: unfortunately, there is no information on any of these graphs that these data points are duplicates for the same country in different years, which could play a clear role in the underlying data structure. They are well labeled, with calls to where the data comes from, but there is a clear problem with directly explaining what the data are.

3

Figure 2 is a good, basic display of what the relationship looks like between the log of GDP Per Capita and Life Expectancy. Figure 3, however, adds a couple of layers to the graph. It takes the loess regression line and makes it linear, which allows us to see the trend in the data much more clearly (even if they are somewhat better modeled by a loess regression line, as all things technically are). It also scales the points by population, which is actually a rather ineffective addition of data - without colors or reduced alpha levels or anything, it's impossible to make sense of this additional data point.

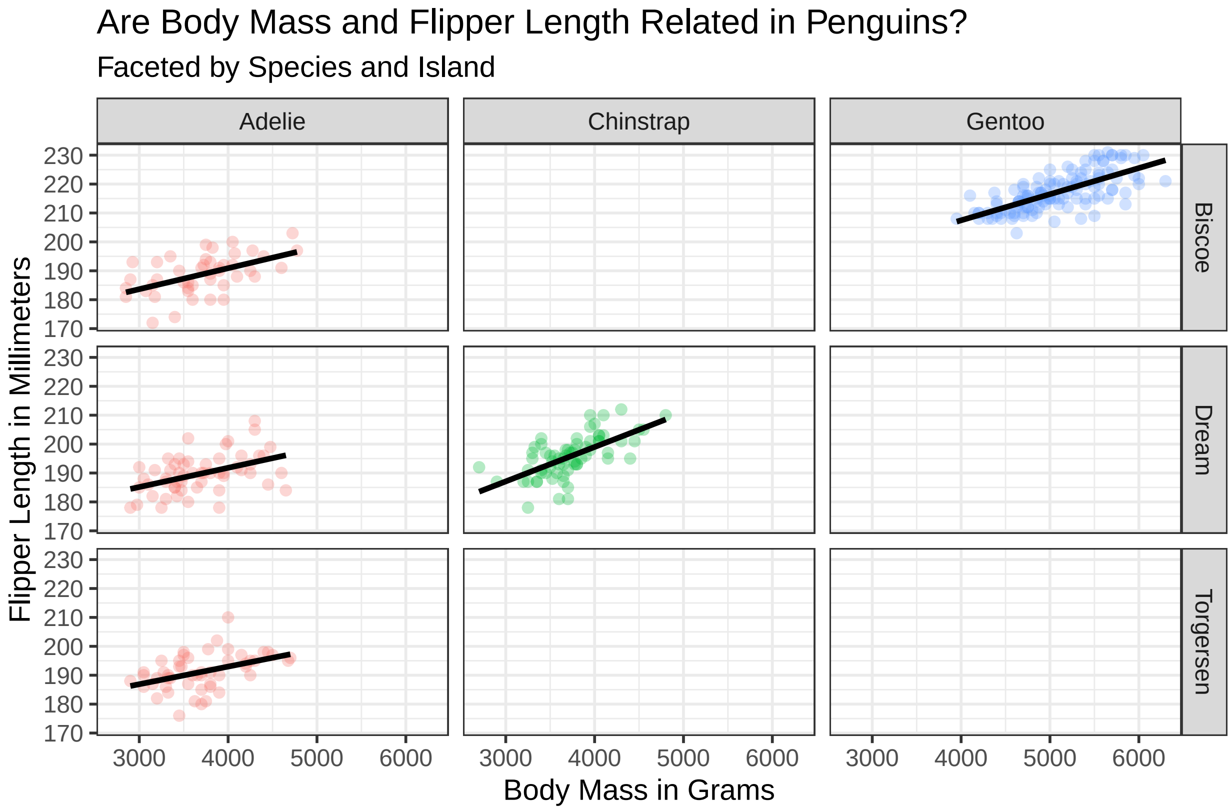
Figure 4 adds a couple of things that improve this graph by a good deal. It maps the variables onto the opposite axes from before, which is neither good nor bad but certainly tells a different story with the data, and reduces the alpha level to allow us to see the points by population better and visualize overlap (even if the scaling of the data is still somewhat ineffective). It certainly suggests that life expectancy is the predictive variable, and also shows just how much data is clustered around a life expectancy of 70-80 (which was obscured with a default alpha = 1).

```r
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = species), alpha = 0.3) +
  facet_grid(island ~ species) +
  geom_smooth(method = lm, se = F, color = "black") +
  labs(title = "Are Body Mass and Flipper Length Related in Penguins?",
       subtitle = "Faceted by Species and Island", x = "Body Mass in Grams",
       y = "Flipper Length in Millimeters",
       caption = "Data: German, Williams, and Fraser (2014)") +
  theme_bw() +
  guides(color = "none")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Data: German, Williams, and Fraser (2014)

This visualization is interesting because it shows a few different data points. First, it shows a fairly clear linear relationship between body mass and flipper length, even if it is somewhat scattered (i.e. large residuals and not ideal r-values). It shows that this relationship is almost exactly the same within species on different islands, but not the same across species even on the same island. It tells us that Adelie penguins live on all three islands, whereas Chinstrap are found only on Dream Island and Gentoo only on Briscoe, and it tells us that Gentoo penguins are almost always larger (i.e. heavier and with longer flippers) than other species of penguin.