Introduction to ggplot

Abhinav Anand

Setup

The following discussion assumes you have donwloaded R and RStudio. Additionally, the package suite `tidyverse()` which includes the package `ggplot2()` needs to be included.

1. For downloading R, visit https://cran.r-project.org/
2. For downloading RStudio visit https://www.rstudio.com/
3. For downloading `ggplot2()`, type `install.packages("ggplot2")` or equivalently for `tidyverse()` type `install.packages("tidyverse")`

Introduction to ggplot

The `gg` of `ggplot` stands for (layered) `grammar of graphics` (Wilkinson 2005), (Wickham 2010). This idea will be further explored by the means of data from the package `gapminder()`. To install, type `install.packages("gapminder")` in the RStudio console.

```
data_gapminder <- gapminder::gapminder
```

Notes

1. Why `<-` as opposed to `=` ?
2. Why `gapminder::gapminder` ?
3. What is a dataframe?

A data frame is a rectangular collection of variables (in the columns) and observations (in the rows). It's different from a 'mere' matrix since the columns have variable names usually.

Questions?

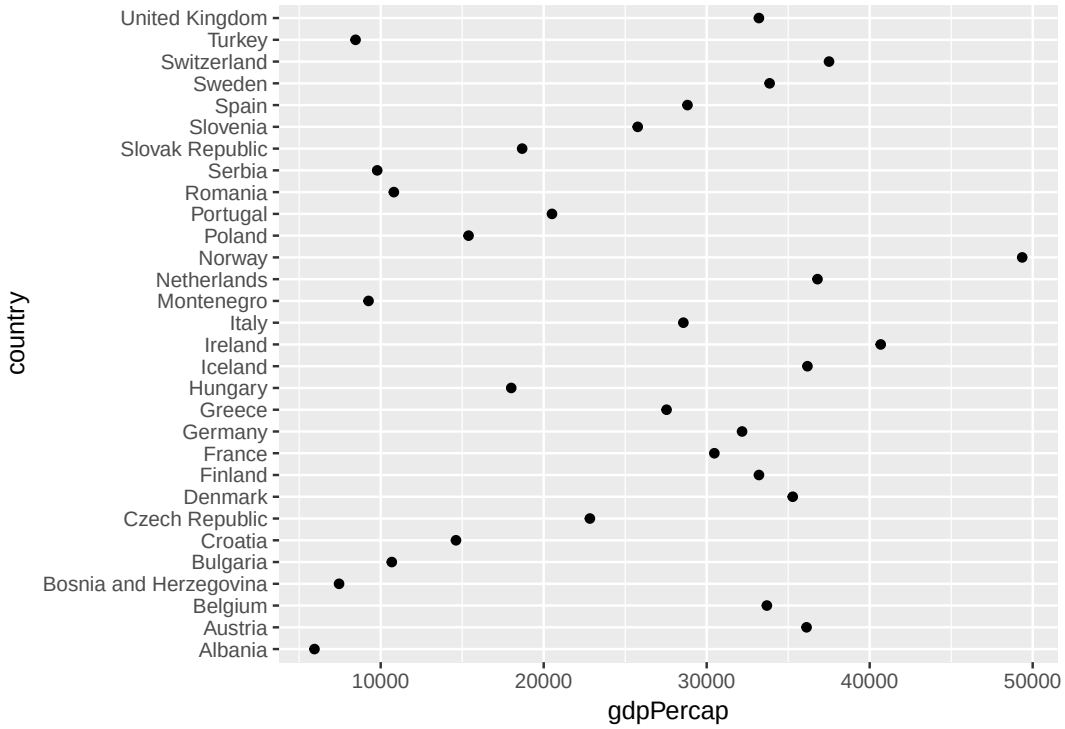
Are "Western" countries richer than "Eastern" countries?

The current state of Europe (in 2007):

```
data_eur_2007 <- data_gapminder %>%
  dplyr::filter(year == 2007) %>% #isolates variables for year 2007
  dplyr::filter(continent == "Europe")

plot_eur <- ggplot(data = data_eur_2007) +
  geom_point(mapping = aes(x = gdpPercap, y = country))

plot_eur
```



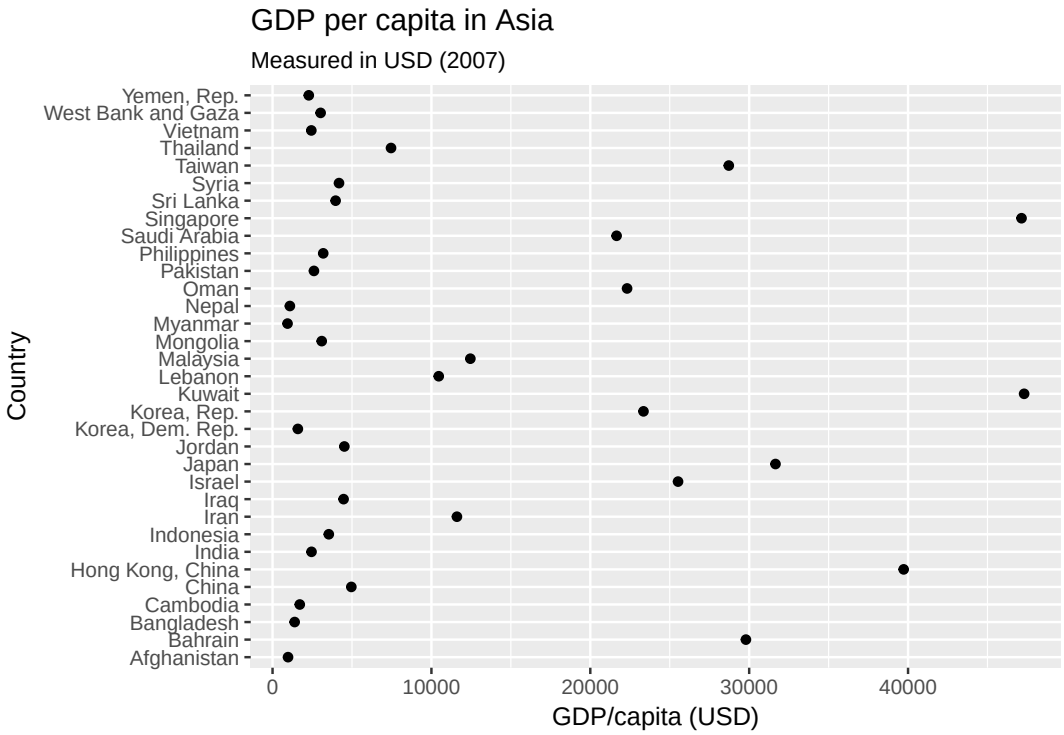
There is high variation—from Albania to Norway.

What about Asian countries in 2007?

```
data_asia_2007 <- data_gapminder %>%
  dplyr::filter(year == 2007) %>% #isolates variables for year 2007
  dplyr::filter(continent == "Asia") #collect only Asian countries
```

```
plot_asia <- ggplot(data = data_asia_2007) +
  geom_point(mapping = aes(x = gdpPercap, y = country)) +
  labs(x = "GDP/capita (USD)",
       y = "Country",
       title = "GDP per capita in Asia",
       subtitle = "Measured in USD (2007)")

plot_asia
```



Again, large variation among Asian countries but what about the bounds? What can we say about the question?

Graphics

We start with the function `ggplot()`. It creates a coordinate system that we will add layers to. The first argument is the dataset to use in the graph.

`ggplot(data = data_eur_2007)` creates an empty graph. The function `geom_point()` adds a layer of points to our plot. Each geom function in ggplot2 takes a mapping argument. This

defines how variables in our dataset are mapped to aesthetics such as axes, colors, shapes etc. The x and y arguments of `aes()` specify which variables to map to the x and y axes.

References

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28.

Wilkinson, Leland. 2005. *The Grammar of Graphics (Statistics and Computing)*. Berlin, Heidelberg: Springer-Verlag.