Problem set 9: Visualizing data

ADS2

2023-11-16

You will need the packages `tidyr, dplyr`, and `ggplot2` (can be obtained by attaching `tidyverse` directly) for the first three problems and `map` for the optional one. You could use the `gather()` and `drop_na()` in `tidyr` and `filter()` in `dplyr`. We recommend you make graphs using `ggplot2`.

Part 1: the GDP in European countries.

Import the `GDP` dataset and clean the data. (The data are obtained from the world bank https://data.worldbank.org/indicator/NY.GDP.MKTP.CD and trimmed.)

1. Import the dataset and reshape the dataset by putting the GDP values from different years into one variable (column), the key is `Year`.
2. Clean the data by removing the incomplete records.
3. Extract the GDPs of "Germany", "France", "Italy" and "Greece" in the years 1960, 1970, 1980, 1990, 2000, 2010 and 2018. Put the data into a new data frame. (**Hint**: `dplyr::filter`)

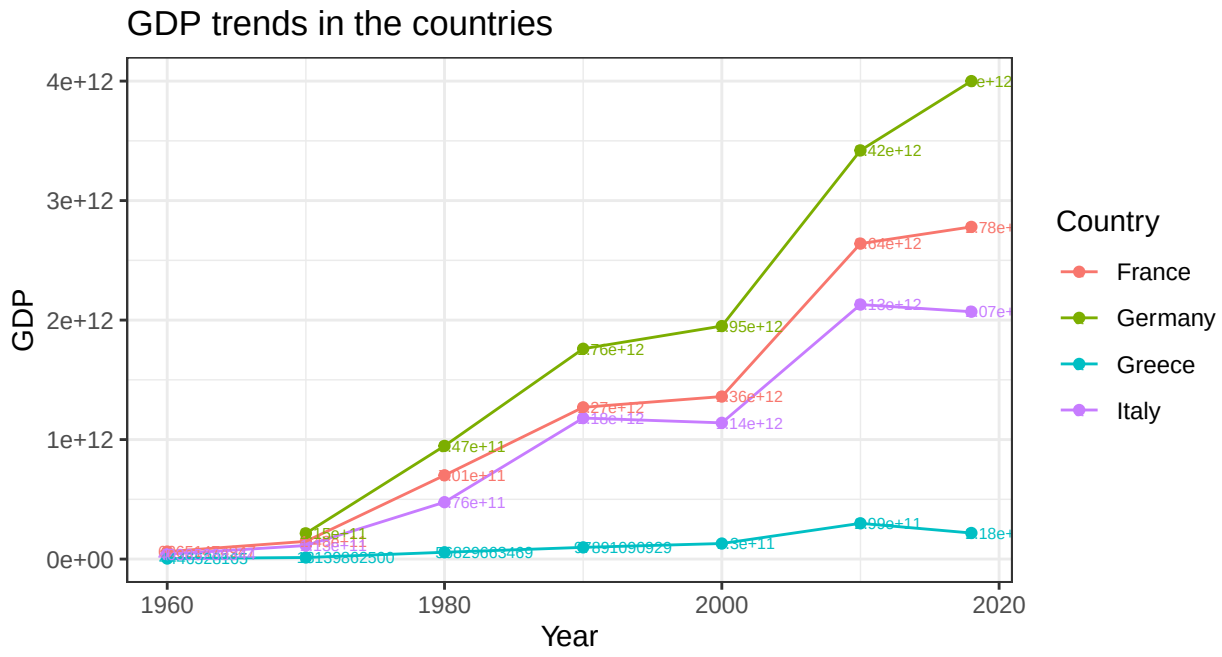
```
library(tidyverse)
GDP <- read.csv(file = "GDP.csv")
colnames(GDP) <-
  gsub(pattern = "[X]",
       replacement = "",
       x = colnames(GDP))
colnames(GDP)[1] <- "Country"
GDP_subset <- GDP %>%
  gather(key = "Year", value = "GDP", 2:60) %>%
  drop_na() %>%
  filter(
    Country %in% c("Greece", "Germany", "France", "Italy") &
      Year %in% c(1960, 1970, 1980, 1990, 2000, 2010, 2018)
  ) %>%
  arrange(Year, Country) %>%
  mutate(Year = as.integer(Year),
         Country = as.factor(Country),
         GDP = as.numeric(GDP))
head(GDP_subset)
```

```
##   Country Year          GDP
## 1  France 1960  62651474947
## 2  Greece 1960   4446528165
## 3   Italy 1960  40385288344
```

```
## 4  France 1970 148000000000
## 5 Germany 1970 215000000000
## 6  Greece 1970  13139862500
```

We want to see the trend of GDP growth in the three countries, please first use point + line plots to present the data, distinguish countries by colours like this. Please also add the title and the GDP numbers at each point. (**Hint**: use `geom_point()` + `geom_line()`)

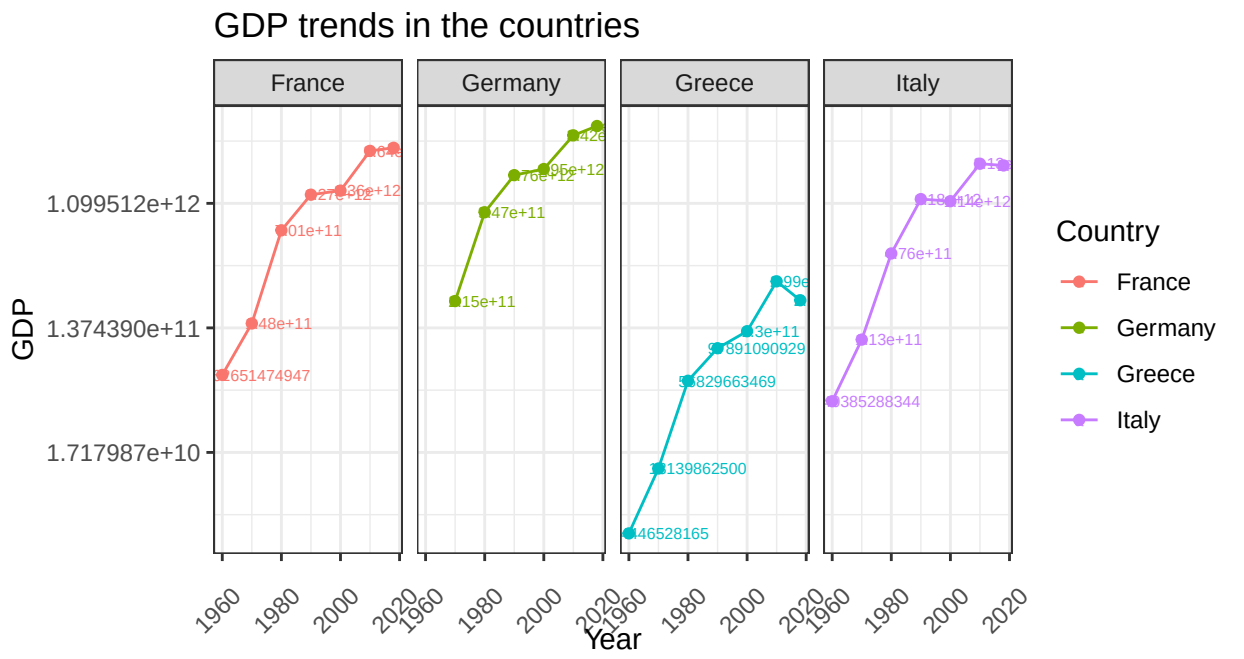
```
g <- ggplot(data = GDP_subset, aes(x = Year, y = GDP, color = Country))

g1 <- g + geom_point() + geom_line(aes(group = Country)) +
# Note that I specified the `group` aesthetic in the `geom_line()`
  geom_text(mapping = aes(label = GDP),
                      hjust = 0.1,
                      nudge_x = 0.05,
                      size = 2) +
  labs(title = "GDP trends in the countries") +
  theme_bw()

g1
```



You can redo the plot using point + fitting with "loess" method. (**Hint**: `geom_point + geom_smooth`)

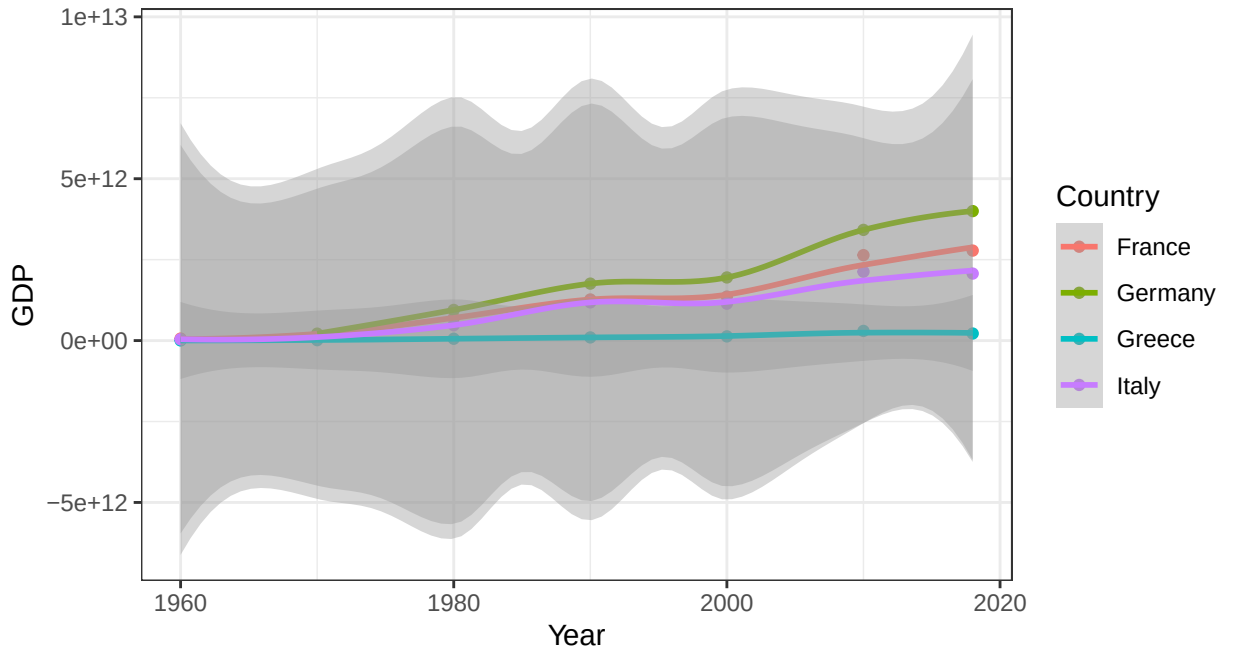
Then try to rescale the GDP by `log2()` and facet the plot by countries.

```
g2 <- g1 + scale_y_continuous(trans = 'log2') + facet_wrap( ~ Country, ncol = 4) +
# Note that I added a statistical transformation in the Y-scale
  theme(axis.text.x = element_text(size = 9, angle = 45, vjust = 0.1))
g2
```



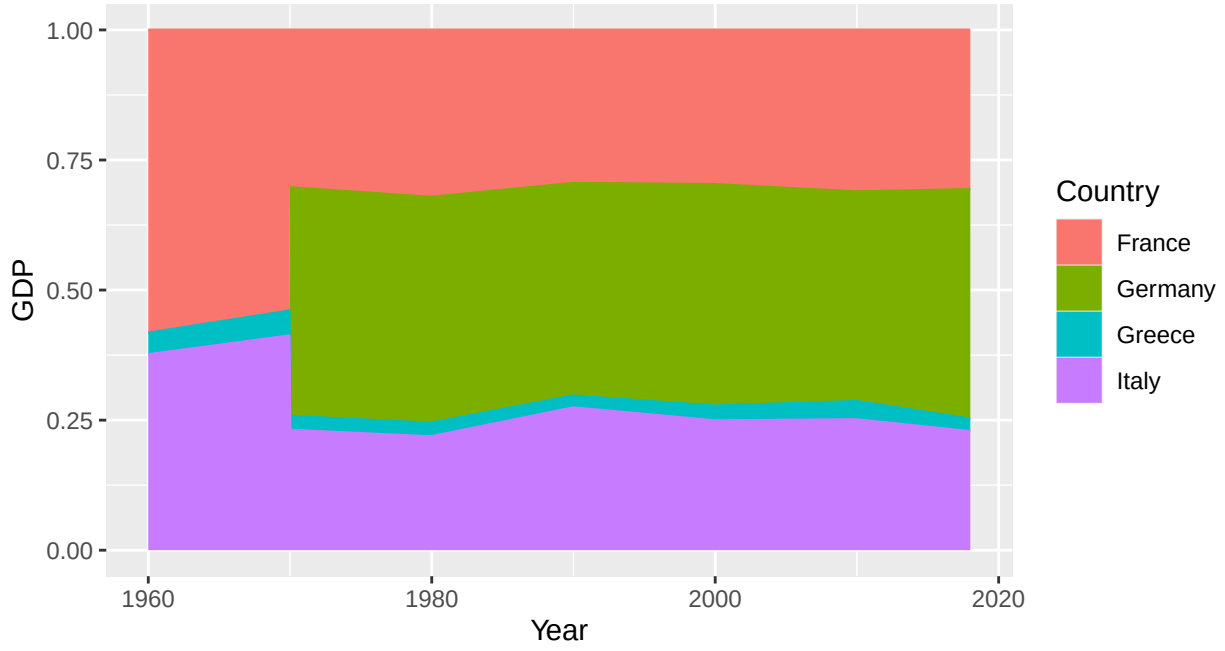
You can redo the plot using point + fitting with "loess" method. (**Hint**: `geom_point() + geom_smooth()`)

```
ggplot(data = GDP_subset, aes(
  x = Year,
  y = GDP,
  color = Country
)) +
  geom_point() +
  geom_smooth(method = "loess") +
  theme_bw()
```



You can also use area plot and `geom_area()` to show the percentiles of the countries' GDP in different years. (**Hint**: `geom_area()`, use the right `position` adjustment).

```
ggplot(data = GDP_subset, aes(
  x = Year,
  y = GDP,
  color = Country)) +
  geom_area(aes(fill = Country), position = "fill")
```



Part 2 (Optional)

We want to show the GDP differences of the three countries in 2018 on the map. First, we need to get the 2018 GDP data out by subsetting the dataset. (**Hint**: `dplyr::filter`)

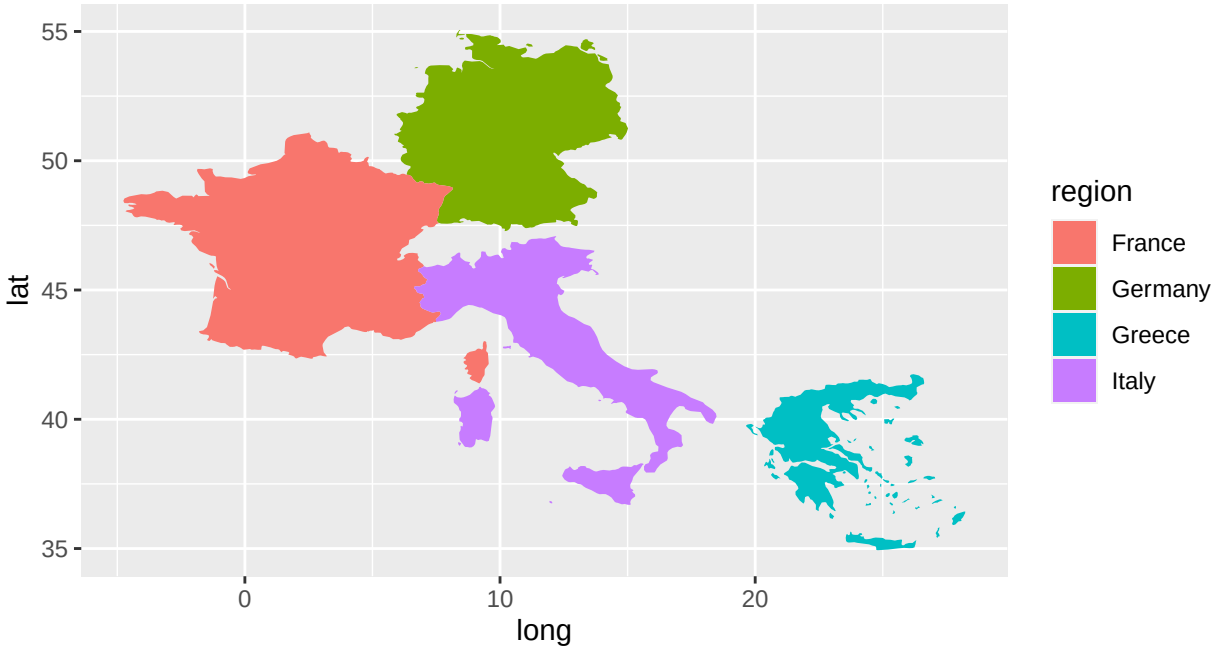
```
GDP_subset_2018 <- filter(GDP_subset, Year == 2018)
```

The map data can be obtained from the package `map` using the following code.

```
library(maps)
eu_map <- map_data("world", region = c("France", "Germany", "Italy", "Greece"))
```

Now you can plot the map using `ggplot2` with `geom_polygon` (Each country is a polygon). Try the code below.

```
ggplot(eu_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = region))
```

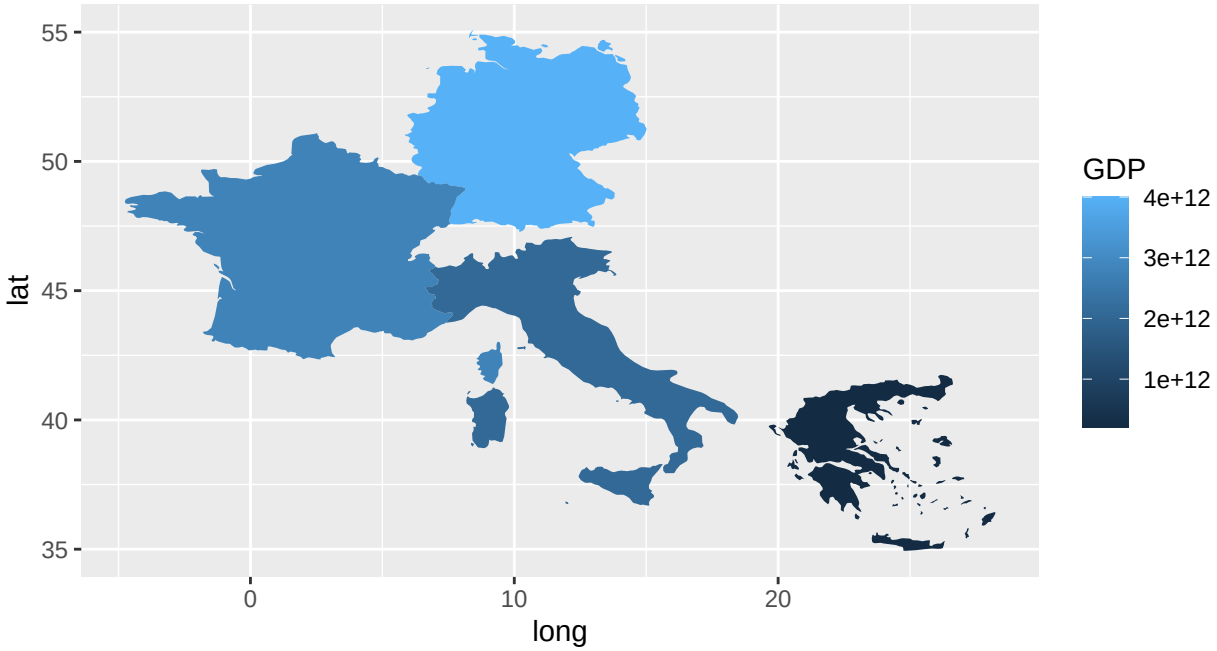


Now we can try to present the GDP. The first step is to merge the GDP data and map data. (**Hint**: use `dplyr::left_join`).

```
gdp.map <- left_join(eu_map, GDP_subset_2018, by = c("region"="Country"))
```

Then use `ggplot2` to colour the countries based on their GDP values.

```
gmap <-
  ggplot(gdp.map, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
                                                                              GDP))
gmap
```



Last update by Dmytro Shytikov in 2023