Week 10: Messy Data

Business Research Methods

Bhaswar Chakma

20 April 2021

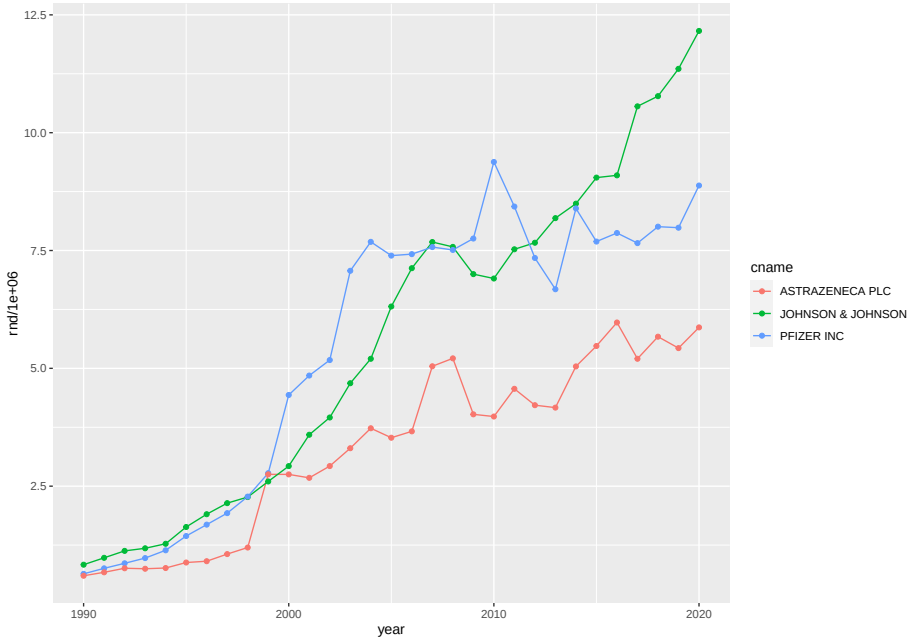
Messy Data

```
library(readxl)
library(tidyverse)
# Exercise 1
df_rnd <- read_excel("./data/w10-datastream.xlsx",
                     sheet = "research&dev",
                     skip = 3)
# Exercise 2
df_ec <- read_excel("data/w10-datastream.xlsx",
                    sheet = "employee-cash",
                    skip = 3)
```

Exercise 1

```
df_rnd_clean <- df_rnd %>%
  # drop unwanted variables
  select(-2, -3) %>%
  # reshape
  pivot_longer(
    cols = -1,
    names_to = "year",
    values_to = "rnd"
  ) %>%
  # year as integer
  mutate(year = as.integer(year)) %>%
  # Extract names
  separate(Name, into = c("cname", NA), sep = "-")
```

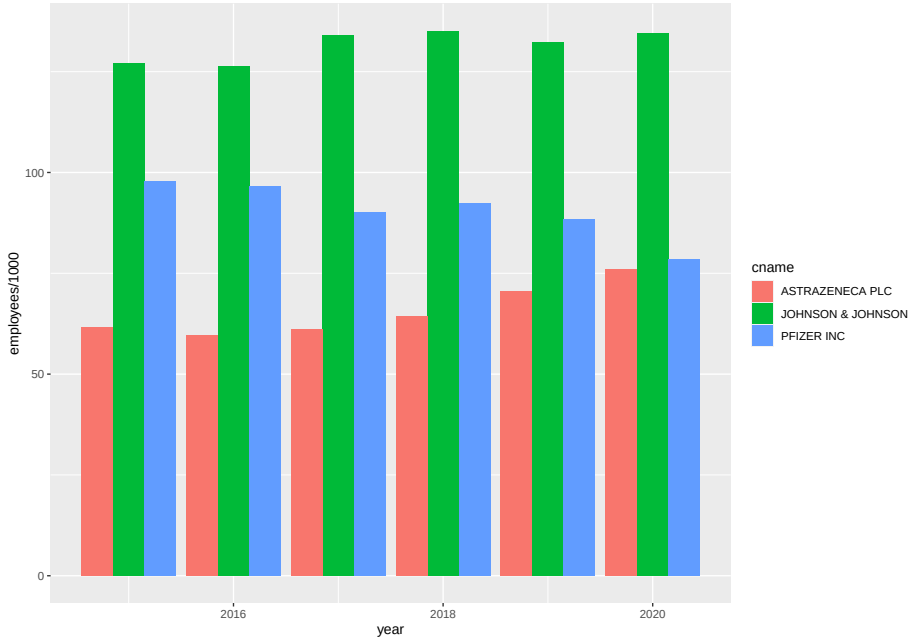
```r
df_rnd_clean %>%
  # values are given in thousand
  ggplot(aes(x = year, y = rnd/1e6, color = cname)) +
  geom_line() +
  geom_point()
```


12.5
10.0
7.5
5.0
2.5
rnd/1e+06
1990
2000
2010
2020
year
cname
ASTRAZENECA PLC
JOHNSON & JOHNSON
PFIZER INC

Exercise 2

```
df_ec_clean <- df_ec %>%
  select(-2, -3) %>%
  pivot_longer(cols = -1,
               names_to = "year",
               values_to = "numbers") %>%
  separate(Name, into = c("cname", "type"), sep = "-") %>%
  pivot_wider(names_from = type,
              values_from = numbers) %>%
  rename(employees = 3,
         cash = 4) %>%
  mutate(year = as.integer(year))
```

```r
# employees
df_ec_clean %>%
  filter(year > 2014) %>%
  ggplot(aes(x = year, y = employees/1000, fill  = cname)) +
  geom_col(position = "dodge")
```


employees/1000
100
50
0
2016
2018
2020
year
cname
ASTRAZENECA PLC
JOHNSON & JOHNSON
PFIZER INC

More Messy Data

```
df <- read_excel("data/Have you met _Ted__(1-38).xlsx")

df_renamed <- df %>%
  select(6:15) %>%
  rename(name = 1,
         id = 2,
         gender = 3,
         date_of_birth = 4,
         struggling = 5,
         reason_struggle = 6,
         practice_r = 7,
         social_media = 8,
         country = 9,
         himym = 10)
```

Exercise 1

1. Fix country names

```
df_renamed2 <- df_renamed %>%
  mutate(country = case_when(country == "BRAZIL" ~ "Brazil",
                             country == "FRANCE" ~ "France",
                             country != "Brazil" |
                               country != "France" ~ country))
```

```
table(df_renamed$country)
```

```
##
##   Brazil   BRAZIL   France   FRANCE  Germany     Iran Portugal
##        2        1        2        1        3        1       28
```

```
table(df_renamed2$country)
```

```
##
##   Brazil   France  Germany     Iran Portugal
##        3        3        3        1       28
```

2. Generate a dummy variable `international`: 1 if international student; otherwise 0

```
df_renamed %>%
  mutate(international = if_else(country != "Portugal", 1, 0))
```

- Fix country names
  - A better solution `str_to_title()`

```
df_renamed %>%
  mutate(country2 = str_to_title(country))
```

Exercise 2

```
df_time <- df_renamed %>%
  select(name, practice_r, social_media, gender) %>%
  mutate(practice_r2 = if_else(practice_r == "3h", 180,
                               as.numeric(practice_r)),
         social_media2 = if_else(social_media == "120min", 120,
                                 as.numeric(social_media)),
         practice_r2 = practice_r2/7)
```

```
df_time %>%
  ggplot(aes(x = practice_r2, y = social_media2, color = gender)) +
  geom_point()
```

```
df_time %>%
  ggplot(aes(x = gender, y = social_media2, fill = gender)) +
  geom_boxplot()
```

Exercise 3

```
df_renamed %>%
  select(name, date_of_birth) %>%
  mutate(date_of_birth = ymd(date_of_birth)) %>%
  arrange(date_of_birth)
```

```
df_renamed %>%
  select(name, date_of_birth) %>%
  mutate(date_of_birth = ymd(date_of_birth),
         month = month(date_of_birth, label = TRUE)) %>%
  ggplot(aes(x = month)) +
  geom_bar() +
  labs(title = "Birthdays")
```

Questions?

bhaswar.chakma@ucp.pt