HW4-Data Analysis

Aylin Mumcular

05 10 2019

install.packages("tidyverse") install.packages("dplyr") install.packages("gapminder")

Q1 Univariate Option 1

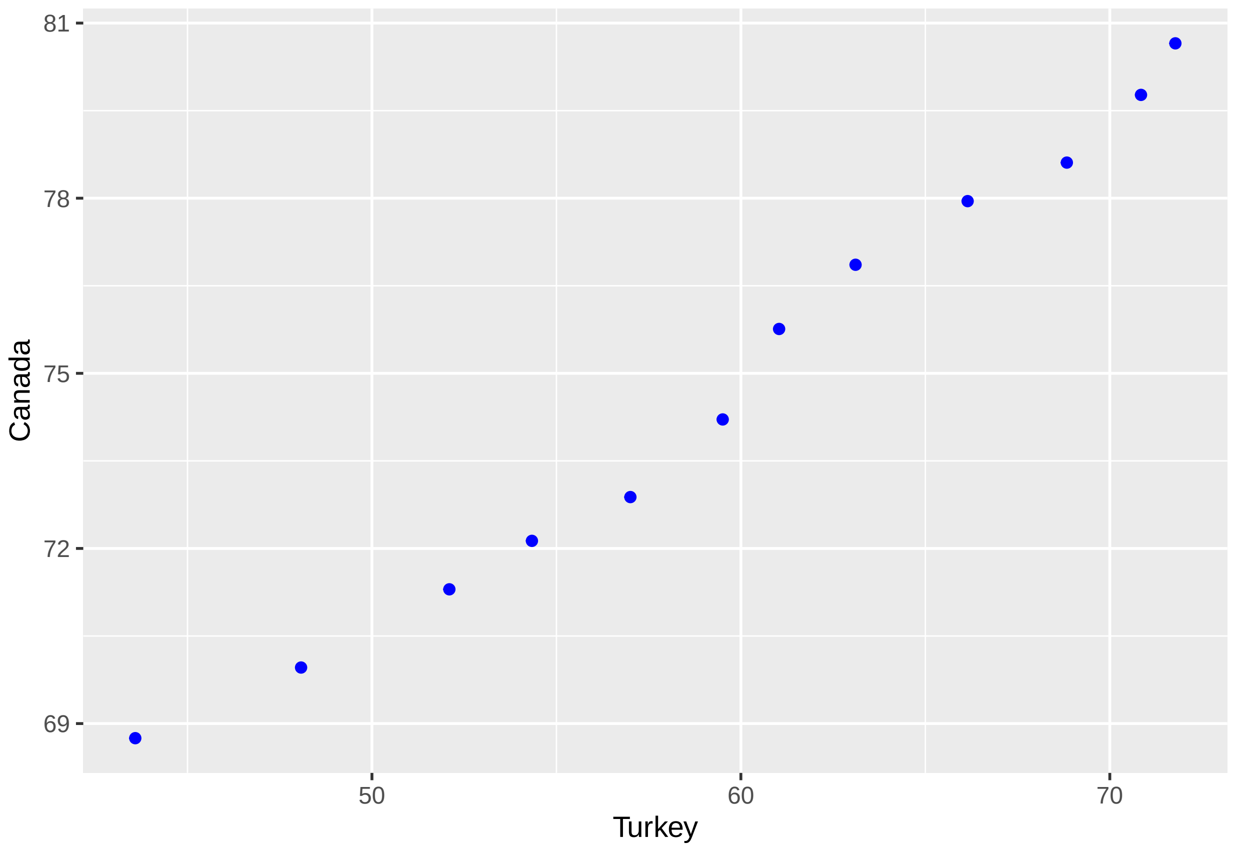
1. Make a tibble with one row per year, and columns for life expectancy for two or more countries.

```
(gap_wider <- gapminder %>%
               filter(country == "Canada" | country == "Turkey") %>%
                  pivot_wider(id_cols = year,
                          names_from = country,
                          values_from = lifeExp))
```

```
## # A tibble: 12 x 3
##     year Canada Turkey
##    <int>  <dbl>  <dbl>
##  1  1952   68.8   43.6
##  2  1957   70.0   48.1
##  3  1962   71.3   52.1
##  4  1967   72.1   54.3
##  5  1972   72.9   57.0
##  6  1977   74.2   59.5
##  7  1982   75.8   61.0
##  8  1987   76.9   63.1
##  9  1992   78.0   66.1
## 10  1997   78.6   68.8
## 11  2002   79.8   70.8
## 12  2007   80.7   71.8
```

2. Take advantage of this new data shape to scatterplot life expectancy for one country against that of another.

```
gap_wider %>%
  ggplot(aes(Turkey, Canada)) +
  geom_point(colour = "blue")
```



3. Re-lengthen the data.

```
(gap_longer <- gap_wider %>%
                 pivot_longer(cols = c("Canada", "Turkey"),
                              names_to = "country",
                              values_to = "lifeExp"))
```

```
## # A tibble: 24 x 3
##     year country lifeExp
##    <int> <chr>     <dbl>
##  1  1952 Canada     68.8
##  2  1952 Turkey     43.6
##  3  1957 Canada     70.0
##  4  1957 Turkey     48.1
##  5  1962 Canada     71.3
##  6  1962 Turkey     52.1
##  7  1967 Canada     72.1
##  8  1967 Turkey     54.3
##  9  1972 Canada     72.9
## 10  1972 Turkey     57.0
## # ... with 14 more rows
```

Q2 Multivariate Option 1

1. Make a tibble with one row per year, and columns for life expectancy and GDP per capita (or two other numeric variables) for two or more countries.

```
(gap_wider_mult <- gapminder %>%
                    filter(country == "Canada" | country == "Turkey") %>%
                      pivot_wider(id_cols     = year,
                                  names_from  = country,
                                  names_sep   = "_",
                                  values_from = c(lifeExp, gdpPercap)))
```

```
## # A tibble: 12 x 5
##     year lifeExp_Canada lifeExp_Turkey gdpPercap_Canada gdpPercap_Turkey
##    <int>          <dbl>          <dbl>            <dbl>            <dbl>
##  1  1952           68.8           43.6           11367.            1969.
##  2  1957           70.0           48.1           12490.            2219.
##  3  1962           71.3           52.1           13462.            2323.
##  4  1967           72.1           54.3           16077.            2826.
##  5  1972           72.9           57.0           18971.            3451.
##  6  1977           74.2           59.5           22091.            4269.
##  7  1982           75.8           61.0           22899.            4241.
##  8  1987           76.9           63.1           26627.            5089.
##  9  1992           78.0           66.1           26343.            5678.
## 10  1997           78.6           68.8           28955.            6601.
## 11  2002           79.8           70.8           33329.            6508.
## 12  2007           80.7           71.8           36319.            8458.
```

2. Re-lengthen the data.

```
(gap_longer_mult <- gap_wider_mult %>%
                      pivot_longer(cols = c(-year),
                          names_to = c(".value", "country"),
                          names_sep = "_"))
```

```
## # A tibble: 24 x 4
##     year country lifeExp gdpPercap
##    <int> <chr>     <dbl>     <dbl>
##  1  1952 Canada     68.8    11367.
##  2  1952 Turkey     43.6     1969.
##  3  1957 Canada     70.0    12490.
##  4  1957 Turkey     48.1     2219.
##  5  1962 Canada     71.3    13462.
##  6  1962 Turkey     52.1     2323.
##  7  1967 Canada     72.1    16077.
##  8  1967 Turkey     54.3     2826.
##  9  1972 Canada     72.9    18971.
## 10  1972 Turkey     57.0     3451.
## # ... with 14 more rows
```

Q3 Table Joins

```
guest <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/attend.c
```

```
## Parsed with column specification:
## cols(
##   party = col_double(),
##   name = col_character(),
##   meal_wedding = col_character(),
##   meal_brunch = col_character(),
##   attendance_wedding = col_character(),
##   attendance_brunch = col_character(),
##   attendance_golf = col_character()
## )
```

```
email <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/emails.c
```

```
## Parsed with column specification:
## cols(
##   guest = col_character(),
##   email = col_character()
## )
```

3.1 For each guest in the guestlist (guest tibble), add a column for email address, which can be found in the email tibble.

```
e_sep <- as_tibble(email) %>%
          rename(name = guest) %>% #Change the name so that it will match in both tables
            separate_rows(name, sep = ", ") #Separate names


guest %>%
  left_join(e_sep, by = "name")
```

```
## # A tibble: 30 x 8
##    party name  meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr> <chr>        <chr>       <chr>            <chr>
##  1     1 Somm~ PENDING      PENDING     PENDING          PENDING
##  2     1 Phil~ vegetarian   Menu C      CONFIRMED        CONFIRMED
##  3     1 Blan~ chicken      Menu A      CONFIRMED        CONFIRMED
##  4     1 Emaa~ PENDING      PENDING     PENDING          PENDING
##  5     2 Blai~ chicken      Menu C      CONFIRMED        CONFIRMED
##  6     2 Nige~ <NA>         <NA>        CANCELLED        CANCELLED
##  7     3 Sine~ PENDING      PENDING     PENDING          PENDING
##  8     4 Ayra~ vegetarian   Menu B      PENDING          PENDING
##  9     5 Atla~ PENDING      PENDING     PENDING          PENDING
## 10     5 Denz~ fish         Menu B      CONFIRMED        CONFIRMED
## # ... with 20 more rows, and 2 more variables: attendance_golf <chr>,
## #   email <chr>
```

3.2 Who do we have emails for, yet are not on the guestlist?

```
e_sep %>%
  anti_join(guest, by = "name")
```

```
## # A tibble: 3 x 2
##   name            email
##   <chr>           <chr>
## 1 Turner Jones    tjjones12@hotmail.ca
## 2 Albert Marshall themarshallfamily1234@gmail.com
## 3 Vivian Marshall themarshallfamily1234@gmail.com
```

3.3 Make a guestlist that includes everyone we have emails for (in addition to those on the original guestlist).

```
guest %>%
  full_join(e_sep, by = "name")
```

```
## # A tibble: 33 x 8
##    party name  meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr> <chr>        <chr>       <chr>            <chr>
##  1     1 Somm~ PENDING      PENDING     PENDING          PENDING
##  2     1 Phil~ vegetarian   Menu C      CONFIRMED        CONFIRMED
##  3     1 Blan~ chicken      Menu A      CONFIRMED        CONFIRMED
##  4     1 Emaa~ PENDING      PENDING     PENDING          PENDING
##  5     2 Blai~ chicken      Menu C      CONFIRMED        CONFIRMED
##  6     2 Nige~ <NA>         <NA>        CANCELLED        CANCELLED
##  7     3 Sine~ PENDING      PENDING     PENDING          PENDING
##  8     4 Ayra~ vegetarian   Menu B      PENDING          PENDING
##  9     5 Atla~ PENDING      PENDING     PENDING          PENDING
## 10     5 Denz~ fish         Menu B      CONFIRMED        CONFIRMED
## # ... with 23 more rows, and 2 more variables: attendance_golf <chr>,
## #   email <chr>
```