Lab 3 - Data tidying and joining

Edmond Niu

```
library(tidyverse)
```

Part 1

```
country_inflation <- read_csv("data/country-inflation.csv")
```

Question 1

a. Country_inflation has 44 rows and each row represents a different country. There are 30 columns and each column represents a different year from 1993-2021 (inclusive).

```
glimpse(country_inflation)
```

```
Rows: 44
Columns: 30
$ country <chr> "Australia", "Austria", "Belgium", "Canada", "Czech Republic",~
$ `1993`  <dbl> 1.753653, 3.631786, 2.754426, 1.865079, 20.813026, 1.257862, 2~
$ `1994`  <dbl> 1.9696348, 2.9534065, 2.3775445, 0.1655629, 10.0394242, 1.9920~
$ `1995`  <dbl> 4.6277666, 2.2433638, 1.4679612, 2.1487603, 8.9905306, 2.08360~
$ `1996`  <dbl> 2.6153846, 1.8609741, 2.0770246, 1.5705311, 8.7587747, 2.12629~
$ `1997`  <dbl> 0.2248876, 1.3059833, 1.6281605, 1.6212164, 8.5961567, 2.18216~
$ `1998`  <dbl> 0.8601346, 0.9224669, 0.9492503, 0.9959425, 10.6983655, 1.8456~
$ `1999`  <dbl> 1.4831294, 0.5689865, 1.1208482, 1.7348430, 2.1354484, 2.49779~
$ `2000`  <dbl> 4.4574351, 2.3448684, 2.5445178, 2.7194400, 3.7753883, 2.90327~
$ `2001`  <dbl> 4.407135, 2.650000, 2.469258, 2.525120, 4.662676, 2.337875, 2.~
$ `2002`  <dbl> 2.981575, 1.810359, 1.645214, 2.258394, 1.902981, 2.424437, 1.~
$ `2003`  <dbl> 2.7325960, 1.3555538, 1.5889640, 2.7585632, 0.1187392, 2.07507~
```

```
$ `2004`  <dbl> 2.3432552, 2.0612068, 2.0972831, 1.8572587, 2.7601078, 1.15435~
$ `2005`  <dbl> 2.6918317, 2.2991377, 2.7814326, 2.2135520, 1.8570979, 1.81781~
$ `2006`  <dbl> 3.555288, 1.441547, 1.791208, 2.002025, 2.533993, 1.924221, 1.~
$ `2007`  <dbl> 2.3276113, 2.1685559, 1.8230563, 2.1383840, 2.8531244, 1.69326~
$ `2008`  <dbl> 4.350299, 3.215951, 4.489444, 2.370271, 6.358664, 3.416268, 4.~
$ `2009`  <dbl> 1.771117e+00, 5.063094e-01, -5.314567e-02, 2.994668e-01, 1.019~
$ `2010`  <dbl> 2.9183400, 1.8135317, 2.1892992, 1.7768715, 1.4727273, 2.31092~
$ `2011`  <dbl> 3.303850, 3.286583, 3.532082, 2.912135, 1.917219, 2.758682, 3.~
$ `2012`  <dbl> 1.7627802, 2.4856751, 2.8396634, 1.5156782, 3.2876231, 2.39791~
$ `2013`  <dbl> 2.44988864, 2.00015749, 1.11309594, 0.93829190, 1.43829787, 0.~
$ `2014`  <dbl> 2.48792271, 1.60580560, 0.34000283, 1.90663591, 0.34398859, 0.~
$ `2015`  <dbl> 1.50836672, 0.89656529, 0.56142915, 1.12524136, 0.30936455, 0.~
$ `2016`  <dbl> 1.276990945, 0.891592367, 1.973852647, 1.428759547, 0.68350420~
$ `2017`  <dbl> 1.9486474, 2.0812686, 2.1259709, 1.5968841, 2.4505340, 1.14713~
$ `2018`  <dbl> 1.9114009, 1.9983819, 2.0531650, 2.2682257, 2.1494949, 0.81360~
$ `2019`  <dbl> 1.6107679, 1.5308955, 1.4368196, 1.9492690, 2.8478760, 0.75813~
$ `2020`  <dbl> 0.84690554, 1.38190955, 0.74079181, 0.71699963, 3.16129528, 0.~
$ `2021`  <dbl> 2.863910, 2.766667, 2.440249, 3.395193, 3.839845, 1.853045, 2.~
```

b.

```
  pull(country_inflation, var="country")
```

```
 [1] "Australia"                  "Austria"
 [3] "Belgium"                    "Canada"
 [5] "Czech Republic"             "Denmark"
 [7] "Finland"                    "France"
 [9] "Germany"                    "Greece"
[11] "Hungary"                    "Iceland"
[13] "Ireland"                    "Italy"
[15] "Japan"                      "Korea"
[17] "Luxembourg"                 "Mexico"
[19] "Netherlands"                "New Zealand"
[21] "Norway"                     "Poland"
[23] "Portugal"                   "Slovak Republic"
[25] "Spain"                      "Sweden"
[27] "Switzerland"                "Türkiye"
[29] "United Kingdom"             "United States"
[31] "Argentina"                  "Brazil"
[33] "Chile"                      "China (People's Republic of)"
[35] "Estonia"                    "India"
[37] "Indonesia"                  "Israel"
```

```
[39] "Russia"                    "Saudi Arabia"
[41] "Slovenia"                  "South Africa"
[43] "Colombia"                  "Costa Rica"
```

Question 2

Argentina, Turkiye, and Brazil had the top 3 highest inflation rates in 2021. US inflation rate (2021): 4.69. These inflation rates for these countries are substantially larger than the US inflation rate (by magnitudes greater). This showcases that the US inflation rate in 2021 was not considered a high inflation rate compared to other countries in that year.

```
country_inflation |>
  arrange(desc(`2021`)) |>
  select(country, `2021`) |>
  slice(0:3)
```

```
# A tibble: 3 x 2
  country   `2021`
  <chr>      <dbl>
1 Argentina  48.4
2 Türkiye    19.6
3 Brazil      8.30
```

Question 3

New Zealand has the largest inflation change over this time period. Inflation increased between 1993 and 2021 in this country by approximately 3-fold.

```
country_inflation |>
  mutate(inf_ratio = `2021` / `1993`) |>
  arrange(desc(inf_ratio)) |>
  select(country, inf_ratio)
```

```
# A tibble: 44 x 2
   country       inf_ratio
   <chr>             <dbl>
 1 New Zealand        3.06
 2 Canada             1.82
 3 Australia          1.63
 4 Ireland            1.60
 5 United States      1.59
 6 Norway             1.52
 7 Denmark            1.47
 8 Iceland            1.10
 9 Netherlands        1.04
10 Finland            1.00
# i 34 more rows
```

Question 4

This pivoted data frame has 1276 rows and 3 columns.

```
pivoted_CI <- country_inflation |>
  pivot_longer(
    cols = !country, names_to = "year", values_to = "annual_inflation",
    names_transform = as.numeric)

pivoted_CI
```

```
# A tibble: 1,276 x 3
   country    year annual_inflation
   <chr>     <dbl>            <dbl>
 1 Australia  1993            1.75
 2 Australia  1994            1.97
 3 Australia  1995            4.63
 4 Australia  1996            2.62
 5 Australia  1997            0.225
 6 Australia  1998            0.860
 7 Australia  1999            1.48
 8 Australia  2000            4.46
 9 Australia  2001            4.41
10 Australia  2002            2.98
# i 1,266 more rows
```

Question 5

a. Brazil in 1994 had the highest inflation rate of 2075.888.

```
pivoted_CI |>
  filter(annual_inflation == max(annual_inflation, na.rm = TRUE))
```

```
# A tibble: 1 x 3
  country  year annual_inflation
  <chr>   <dbl>            <dbl>
1 Brazil   1994            2076.
```

b. Ireland in 2009 had the lowest inflation rate of -4.478.

```
pivoted_CI |>
  filter(annual_inflation == min(annual_inflation, na.rm = TRUE))
```

```
# A tibble: 1 x 3
  country  year annual_inflation
  <chr>   <dbl>            <dbl>
1 Ireland  2009            -4.48
```

c. Brazil in 1994 had the highest inflation rate of 2075.888, while Ireland in 2009 had the lowest inflation rate of -4.478.

```
pivoted_CI |>
  filter(annual_inflation == max(annual_inflation, na.rm = TRUE) |
         annual_inflation == min(annual_inflation, na.rm = TRUE))
```

```
# A tibble: 2 x 3
  country  year annual_inflation
  <chr>   <dbl>            <dbl>
1 Ireland  2009            -4.48
2 Brazil   1994          2076.
```

Question 6

a. I chose these countries because I am a US citizen, and I want to visit Japan (I love sushi) and Italy (I love some good pasta).

```
countries_of_interest <- c("Italy", "United States", "Japan")
```

b.

```
pivoted_COF <- pivoted_CI |>
  filter(country == "Italy" |
         country == "United States" |
         country == "Japan")

distinct(pivoted_COF, country)
```

```
# A tibble: 3 x 1
  country
  <chr>
1 Italy
2 Japan
3 United States
```

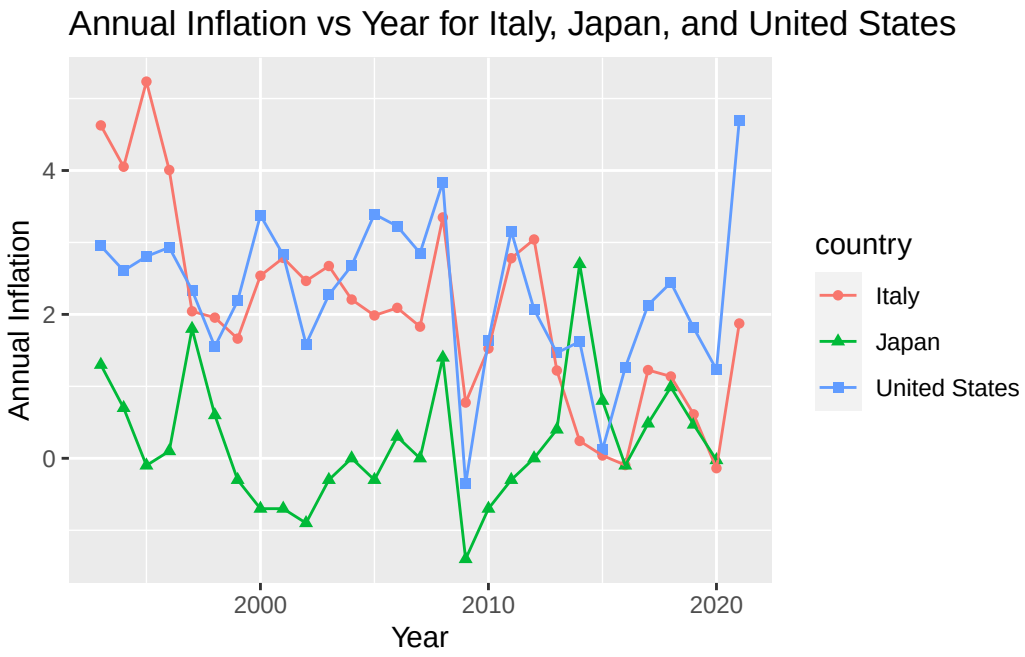
Question 7

Overall, United States inflation rate (blue) is similar to Italy's inflation rate (red) throughout 1993-2021. However, both US and Italy's inflation rates are typically higher than Japan's inflation rate (green) for most of the years between 1993-2014. In 2008, there was a huge drop in inflation rates for all 3 countries, likely due to the Great Recession. US and Italy's inflation rates recovered faster than Japan's inflation rate recovery, which happened on a much slower scale. In the recent year of 2020-2021, Italy and United States' inflation rates surged upward, while Japan does not have a value for 2021. Overall, looking at just starting and ending years (2021 vs 1993), Italy and Japan's inflation rates decreased but the US's inflation rate increased. The US (in 2021) has the highest inflation rate of 4.8% compared to Italy at 2%, and Japan is unknown (but likely to be around 2% or lower).

```
ggplot(data=pivoted_COF, aes(x=year, y=annual_inflation,
                             color=country, shape=country)) +
  geom_point() +
  geom_line() +
  labs(
    title = "Annual Inflation vs Year for Italy, Japan, and United States",
    x = "Year",
    y = "Annual Inflation"
  ) + theme_grey()
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 1 row containing missing values (`geom_line()`).
```


Annual Inflation vs Year for Italy, Japan, and United States
Annual Inflation
4
2
0
2000
2010
2020
Year
country
Italy
Japan
United States

Part 2

```
us_inflation <- read_csv("data/us-inflation.csv")
cpi_divisions <- read_csv("data/cpi-divisions.csv")
```

Question 8

a. The us_inflation dataset has 132 rows and 4 columns. The columns include country, cpi_division_id, year, and annual_inflation. This dataset represents all different entries of annual_inflation based on country, year, and which cpi_division_id this annual_inflation value is representing.

b. The cpi_divisions dataset has 12 rows and 2 columns. The columns include id and description. This dataset is mapping all 12 ids (which are represented by variable: cpi_division_id in the us_inflation dataset) to one of the 12 different categories that the CPI index is broken down into (food, clothing, etc).

c. The joined_USInflation_CPI dataset has 132 rows and 5 columns. The names of the columns are: country, cpi_division_id, year, annual_inflation, and description. This dataset has combined both the us_inflation dataset and cpi_divisions dataset into one dataset, with a mapping of cpi_division_id from us_inflation to id from cpi_divisions. You can think of this joined dataset to be the us_inflation dataset with one additional column from the cpi_divisions dataset, which is (description). The other column in cpi_divisions (id) is not included in this joined dataset because that is the column value that is used for mapping: it is mapped to a column value in us_inflation (cpi_division_id) to create the mapping for the join.

```
joined_USInflation_CPI <- us_inflation |>
  full_join(cpi_divisions, join_by(cpi_division_id==id))

joined_USInflation_CPI
```

```
# A tibble: 132 x 5
   country       cpi_division_id  year annual_inflation description
   <chr>                   <dbl> <dbl>            <dbl> <chr>
 1 United States               1  2011             4.80  Food and non-Alcoholic ~
 2 United States               1  2012             2.45  Food and non-Alcoholic ~
 3 United States               1  2013             0.908 Food and non-Alcoholic ~
 4 United States               1  2014             2.64  Food and non-Alcoholic ~
 5 United States               1  2015             1.17  Food and non-Alcoholic ~
 6 United States               1  2016            -1.33  Food and non-Alcoholic ~
```

```
 7 United States              1  2017          -0.202 Food and non-Alcoholic ~
 8 United States              1  2018           0.456 Food and non-Alcoholic ~
 9 United States              1  2019           0.885 Food and non-Alcoholic ~
10 United States              1  2020           3.51  Food and non-Alcoholic ~
# i 122 more rows
```

Question 9

a. I chose these divisions because they are most relevant to my everyday life. I buy food and non-alcoholic beverages everyday, I buy clothing and footwear everyday, and I often eat at restaurants and stay at hotels a few times a month.

```
divisions_of_interest <- c("Food and non-Alcoholic beverages",
                           "Clothing and footwear",
                           "Restaurants and hotels")
```

```
DOI_joined_USInflation_CPI <- joined_USInflation_CPI |>
  filter(description == "Food and non-Alcoholic beverages" |
         description == "Clothing and footwear" |
         description == "Restaurants and hotels")

distinct(DOI_joined_USInflation_CPI, country)
```

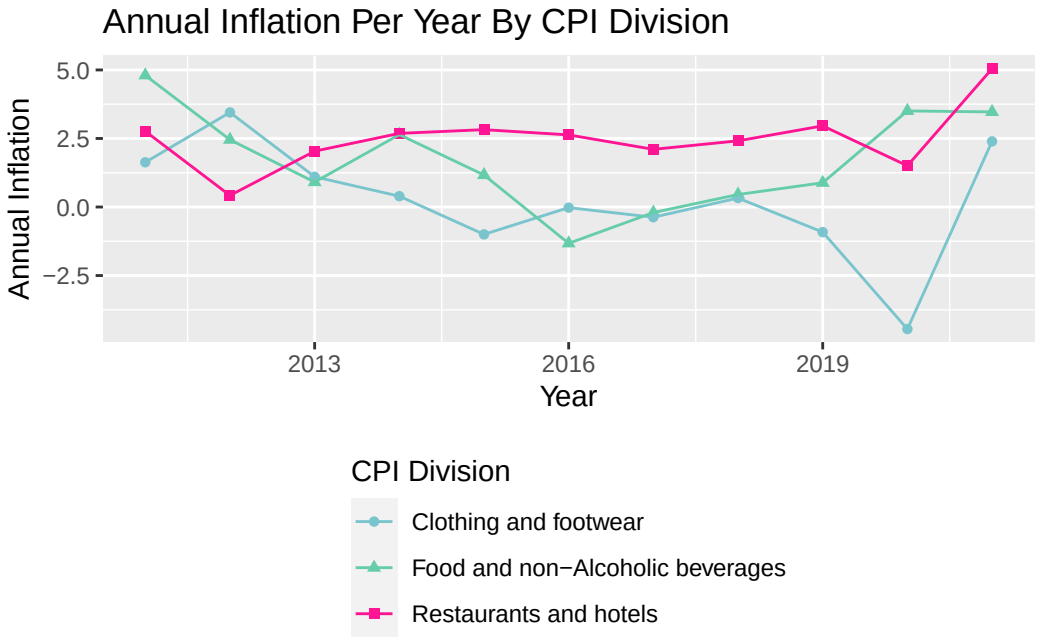
```
# A tibble: 1 x 1
  country
  <chr>
1 United States
```

Question 10

When it comes to restaurants and hotel, this CPI division is the most stable. The inflation rates associated with this division typically have been the highest (when compared to clothing/footwear and food/non-alcoholic beverages. Clothing/footwear and food/non-alcoholic beverages seem to be tied for most years except for after 2019, where clothing and footwear inflation rates dramatically decreased for one year into the negatives and then rebounded to around 2.5%. Overall, year 2020 was when both clothing/footwear and restaurant/hotels inflation rates took a considerable dip. Overall, restaurant and hotels inflation rate is the most stable, with food and non-alcoholic beverages behind the second most stable, and clothing/footwear being the most volatile. As of 2021, the inflation rate of clothing and footwear is the lowest at 2.5%, followed by food/non-alcoholic beverages at ~3.75%, and restaurant and hotels inflation rate is the highest at 5.0%.

```
my_palette2 <- c('cadetblue3', 'aquamarine3', 'deeppink')
```

```
ggplot(data = DOI_joined_USInflation_CPI, aes(x = year, y = annual_inflation,
      shape = description, color = description)) +
  geom_point() +
  geom_line() +
  labs(
    title = "Annual Inflation Per Year By CPI Division",
    x = "Year",
    y = "Annual Inflation",
    color = "CPI Division",
    shape = "CPI Division"
    ) +
  scale_color_manual(values = my_palette2) +
  theme(
        legend.position = "bottom",
        legend.direction = "vertical"
        )
```


Annual Inflation Per Year By CPI Division
Annual Inflation
5.0
2.5
0.0
−2.5
2013
2016
2019
Year
CPI Division
Clothing and footwear
Food and non−Alcoholic beverages
Restaurants and hotels