# Homework 03 - Data tidying

Jemimah Osei (jko35)

2025-02-26

## Setup

Load packages and data:

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ──────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
# your code here
country_inflation <- read_csv("data/country-inflation.csv")
```

```
Rows: 48 Columns: 36
──                              Column                    specification
──────────────────────────────────────────────────
Delimiter: ","
chr  (1): country
dbl (35): 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1992, ...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
Rows: 168 Columns: 4
──                              Column                    specification
──────────────────────────────────────────────────
Delimiter: ","
chr (1): country
dbl (3): cpi_division_id, year, annual_inflation

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
Rows: 12 Columns: 2
──                              Column                    specification
──────────────────────────────────────────────────
Delimiter: ","
chr (1): description
dbl (1): id

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

# Exercises

## Exercise 1

The data set is in the wide format and it contains inflation data from various countries with each row representing a particular country and its inflation rates over different years while the columns represent the names of the countries and specific years

## Exercise 2

```
# your code here

new_inflation <- country_inflation|>
 pivot_longer(cols = 2:last_col(),
               names_to = "year",
               values_to = "annual_inflation") |>
  mutate(
    year = as.integer(year),
    annual_inflation = as.numeric(annual_inflation)
  )

new_inflation
```

```
# A tibble: 1,680 × 3
   country     year annual_inflation
   <chr>      <int>            <dbl>
 1 Argentina   2002               NA
 2 Argentina   2001               NA
 3 Argentina   2000               NA
 4 Argentina   1999               NA
 5 Argentina   1998               NA
 6 Argentina   1997               NA
 7 Argentina   1996               NA
 8 Argentina   1995               NA
 9 Argentina   1994               NA
10 Argentina   1993               NA
# i 1,670 more rows
```

## Exercise 3

```
# your code here
new_inflation |>
  filter(country %in% c("United States", "Japan", "India", "South Africa")) |>
  ggplot(mapping = aes(x = year, y = annual_inflation, color = country, group =
country)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
```

```
  scale_y_continuous(labels = label_percent(scale = 1)) +
  labs(
    title = "Inflation Trends Across Developing and Developed  Economies",
    x = "Year",
    y = "Annual Inflation Rate (%)",
    color = "Country"
  ) +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: Removed 5 rows containing missing values or values outside the scale
range
(`geom_line()`).
```

```
Warning: Removed 5 rows containing missing values or values outside the scale
range
(`geom_point()`).
```
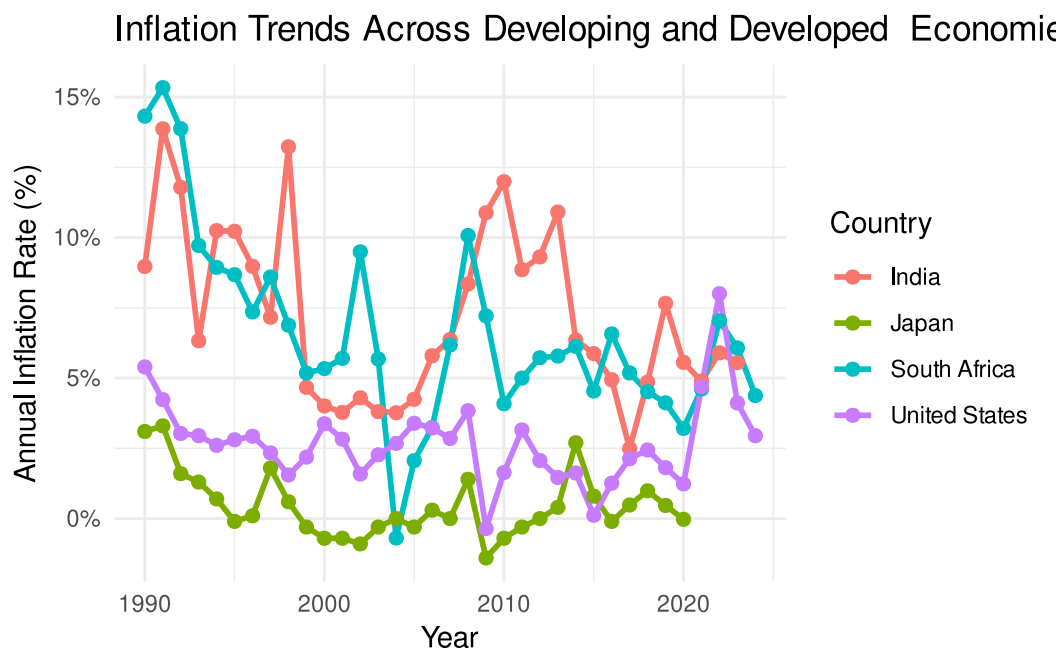


Description

I chose these countries because they gave me a good mix between developing and developed countries/ economies . From the graph the inflation rates of the developing countries like South Africa and India was relatively higher than their developed counterparts. The graph included cer-

tain anomalies occuring. For example, around 2004 South Africa experienced a steep decline in inflation to somewhere below 0%

## Exercise 4

```r
# your code here
us_inflation <- us_inflation |>
  left_join(y=cpi_divisions, by = c("cpi_division_id" = "id"))

us_inflation
```

```
# A tibble: 168 × 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            2.46  Food and non-alcoholic …
 3 United States               1  2013            0.904 Food and non-alcoholic …
 4 United States               1  2014            2.41  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 158 more rows
```

## Exercise 5

```r
# your code here
us_inflation |>
  filter(description %in% c("Food and non-alcoholic beverages", "Housing, water,
electricity, gas and other fuels", "Recreation and culture", "Health", "Alcoholic
beverages, tobacco and narcotics")) |>
  ggplot(mapping = aes(x = year, y = annual_inflation / 100, color = description,
group = description)) +
  geom_line() +
  geom_point() +
  scale_y_continuous(labels = label_percent(accuracy = 1)) +
  labs(
    title = "US Inflation Trends Across A Subset of CPI Divisions",
    subtitle ="Inflation Trends Between Basic Necessities and Luxuries",
    x = "Year",
    y = "Annual Inflation Rate (%)",
    color = "CPI Division"
  ) +
  theme_minimal()+
```

```
theme(
  legend.position = "bottom",
  legend.title = element_text(size = 8),
  legend.text = element_text(size = 6),
) +
guides(color = guide_legend(nrow = 3))
```

## US Inflation Trends Across A Subset of CPI Divisions

### Inflation Trends Between Basic Necessities and Luxuries
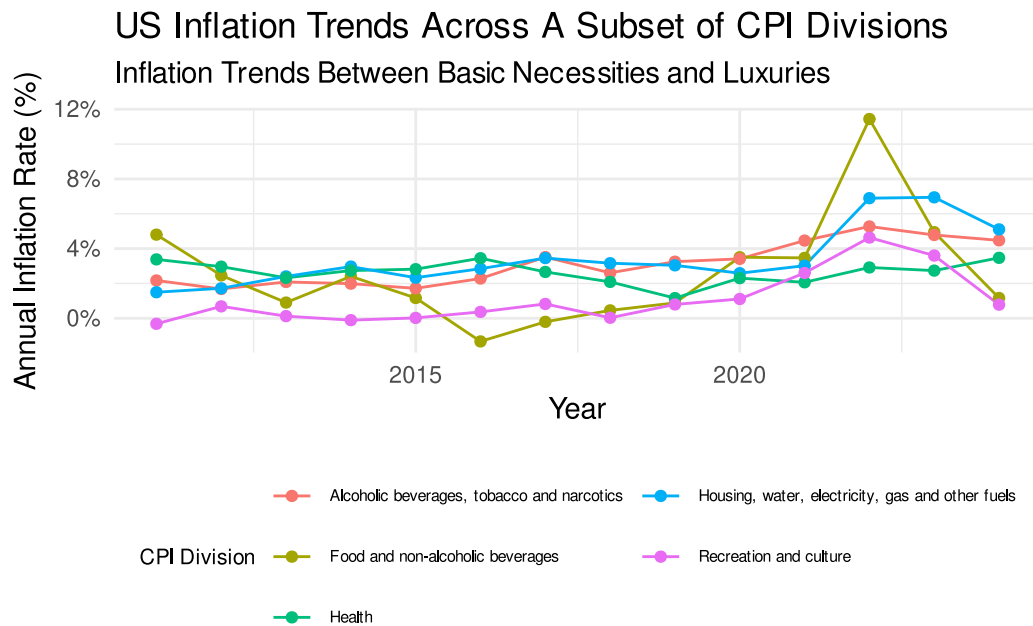


**Description**

I chose this set of cpi divisions because it gave me the opportunities to observer the anual inflation rate over. the years for basic necessities of life life food and health in contrast to some non necessities like alcohol and recreation. From the graph I observed that After 2021 the inflation rates generally rose higher for all of my divisions before dropping down at at around 2024. Also the inflation rate for food dropped significantly from 2010 to 2015 with a brief fluctuation at around 2013 to 2014.