Week2Assignment

Accessing Eurostat Data

The European Union collects extensive data on its member states and their populations. The cheif repository for this data is Eurostat. The `restatapi` package provides an interface to the Eurostat API. The following code demonstrates how to access data on the population of the EU member states and pull it directly into R. In order to get started, we need to load the `tidyverse` and `restatapi` packages.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
restatapi: - version 0.23.1
           - config file with the API version 2 loaded from GitHub (the 'current' API versior
           - 2 from the 8 cores are used for parallel computing, can be changed with 'options
           - 'auto' method will be used for file download, can be changed with 'options(resta
           - the Table of contents (TOC) was not pre-loaded into the deafult cache ('.restata
```

You can find a list of all available datasets on the [Eurostat website]https://ec.europa.eu/eurostat/data/database. The `get_eurostat_data` function allows you to access these datasets directly from R. The function takes the dataset code as an argument. The dataset code for educational attainment of the EU member states by age and gender is `edat_lfse_03`. The following code demonstrates how to access this dataset and store it in a data frame called `EUEduc`. It then generates a summary of the dataset.

```
 sex            age        unit            isced11           geo
 F:41349   Y25-34 :13823   PC:124158   ED0-2   :28355   BE     :  3954
 M:41380   Y30-34 :13803               ED3-8   :28371   EL     :  3954
 T:41429   Y20-24 :13742               ED3_4   :28366   ES     :  3954
           Y15-64 :12297               ED3_4GEN: 5356   IT     :  3954
           Y25-64 :12297               ED3_4VOC: 5406   PT     :  3954
           Y35-44 :12295               ED5-8   :28304   DK     :  3951
           (Other):45901                                (Other):100437
     time           values
 2023   : 6270   Min.   : 0.40
 2022   : 6269   1st Qu.:20.20
 2021   : 6268   Median :37.00
 2016   : 5102   Mean   :41.82
 2017   : 5092   3rd Qu.:61.50
 2018   : 5092   Max.   :98.40
 (Other):90065
```

Note that the summaries of some of the variables in the dataset are portrayed differently than others. Some of the variables, like sex, are are categorical, so the summary shows the number of observations for each category. The only numeric, continuous variable is "values," so the summary shows the mean, median, and other statistics for that variable. This dataset is in long format, meaning that each row represents a unique combination of the categorical variables.

So, in the next chunk, we filter the data to show only the observation where sex is "Total," age is "25-64," isced11 is "ED5-8," geo is "Italy," and time is "2023." The value column therefore shows the percentage of the total population of Italy between 25 and 64 in 2023 that has completed ISCED levels 5-8, or some kind of college degree.

```
EUEduc %>%
  filter(sex=="T", age=="Y25-64", isced11=="ED5-8", geo=="IT", time=="2023")
```

```
Key: <sex, age, unit, isced11, geo>
      sex    age   unit isced11    geo   time values
   <fctr> <fctr> <fctr>  <fctr> <fctr> <fctr>  <num>
1:      T Y25-64     PC   ED5-8     IT   2023   21.6
```

Data visualization is about creating the right visual representation of the data to communicate the information effectively. The `ggplot2` package is a powerful tool for creating visualizations in R. The following code demonstrates how to create a bar plot of the percentage of the population with a college degree in the EU member states in 2023.

```
# Bar Plot of the percentage of the population with a college degree in the EU member states

EUEduc %>%
  filter(sex=="T", age=="Y25-64", isced11=="ED5-8", time=="2023") %>%
  ggplot(aes(x=geo, y=values)) +
  geom_col(stat="identity") +
  coord_flip() +
  labs(title="Percentage of Population with College Degree in EU Member States in 2023",
       x="Country",
       y="Percentage")
```

```
Warning in geom_col(stat = "identity"): Ignoring unknown parameters: `stat`
```