Assignment\_01

Mohammed Al Hor

2022-10-08

### Questions —-

#### Q1 - introduction and descriptives —-

1. Load the tidyverse-package.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
library(ggplot2)

1. Read the data from ‘oecd\_data.csv’.

# Set our working directory  
setwd("~/Data-Science-Business-Analytics/Data")  
# Load our csv file with headers  
oecd\_data <- read.csv("oecd\_data.csv", header = TRUE)

1. Get a first view on the data by getting the dimensions, show the first 5 rows of the data.frame and giving the summary.

To get the dimensions of a dataset we use the dim() function:

dim(oecd\_data)

## [1] 15168 7

# Dimensions are 15168 rows, 7 columns

To show the first 5 rows we use the head() function with n = 5

head(oecd\_data,n=5)

## reg\_id region year country\_code pc\_real\_ppp per real\_ppp  
## 1 ITG27 Cagliari 2000 IT 28821 203200 15650.50  
## 2 KR031 Daegu 2000 KR 13764 NA 34806.70  
## 3 ITG13 Messina 2000 IT 24273 207700 16050.20  
## 4 US09 Connecticut 2000 US 61231 2118200 208907.00  
## 5 UKF12 East Derbyshire 2000 UK 19919 95000 5318.88

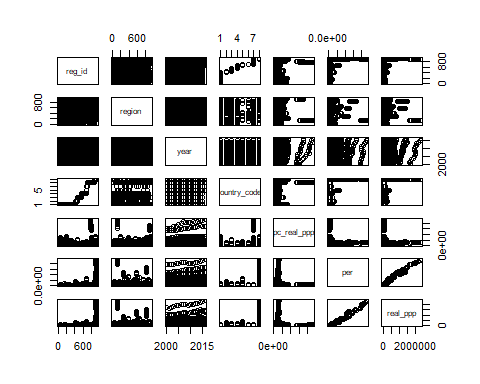
To get summary of the data we use the summary() function

summary(oecd\_data)

## reg\_id region year country\_code   
## Length:15168 Length:15168 Min. :2000 Length:15168   
## Class :character Class :character 1st Qu.:2004 Class :character   
## Mode :character Mode :character Median :2008 Mode :character   
## Mean :2008   
## 3rd Qu.:2012   
## Max. :2016   
##   
## pc\_real\_ppp per real\_ppp   
## Min. : 11364 Min. : 2600 Min. : 175.7   
## 1st Qu.: 26270 1st Qu.: 63388 1st Qu.: 4511.1   
## Median : 31530 Median : 117000 Median : 8551.9   
## Mean : 35383 Mean : 370223 Mean : 30311.6   
## 3rd Qu.: 38684 3rd Qu.: 221000 3rd Qu.: 17171.2   
## Max. :462774 Max. :23265300 Max. :2382750.0   
## NA's :27 NA's :140

1. Use the standard plot-function from R (not ggplot()) that get a first visual view on the data. To get a plot of the whole dataframe we use the plot function and input the dataframe

plot(oecd\_data)



1. How many observations are there by country, by year? Show it in a table. To get the amount of observations by country and year we can use the table() function

table(oecd\_data$country\_code)

##   
## DE ES FR IT KR SE UK US   
## 6432 944 1616 1760 272 336 2941 867

table(oecd\_data$year)

##   
## 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015   
## 933 933 933 933 933 933 933 933 933 933 933 933 933 933 933 933   
## 2016   
## 240