HW7

Homework7

1.How can you check for missing values (NA) in the **fert_cons_data** dataset? Please provide the R code and give a brief explanation of what is happening in the code.

```
library(WDI)
library(tidyr)
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
# Gather fertilizer consumption data from WDI
fert_cons_data <- WDI(indicator = "AG.CON.FERT.ZS")
missing_values <- is.na(fert_cons_data)
missing_count <- colSums(missing_values)
print(missing_count)
```

```
       country          iso2c          iso3c           year AG.CON.FERT.ZS 
             0              0              0              0           4512 
```

```
# This function returns a logical matrix with the same dimensions as fert_cons_data, where
```

2. What is the purpose of reshaping the **`fert_cons_data`** dataset into a wide format?

   The purposes of reshaping may include simplify visualization, summary statistics, comparative analysis, modeling, and data export, but the decision to reshape format should be based on the specific analytical needs.

```
# Reshape fert_cons_data to year wide-format
fert_wide <- tidyr::pivot_wider(fert_cons_data,
names_from = year,values_from = AG.CON.FERT.ZS)
```

3. How can you rename the columns **`Year`** and **`Fert`** in the **`fert_long`** dataset? Please provide the R code and give a brief explanation of what is happening in the code.

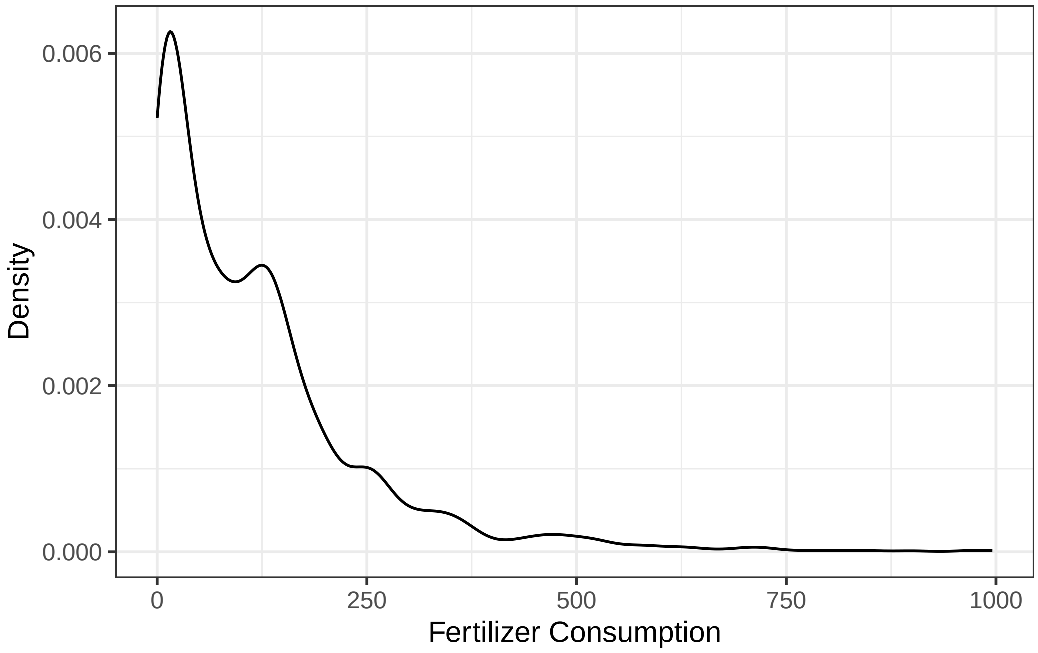
```
# gather a fert_long dataset
fert_long <- tidyr::pivot_longer(fert_wide, cols = `2016`:`2010`, names_to = "Year",
# to rename year and fert_cons columns
fert_long <- dplyr::rename(fert_long,
                           year = Year,
                           fert_cons = Fert)
head(fert_long)
```

```
# A tibble: 6 x 61
  country    iso2c iso3c `2022` `2021` `2020` `2019` `2018` `2017` `2009` `2008`
  <chr>      <chr> <chr>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
2 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
3 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
4 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
5 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
6 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
# i 50 more variables: `2007` <dbl>, `2006` <dbl>, `2005` <dbl>, `2004` <dbl>,
#   `2003` <dbl>, `2002` <dbl>, `2001` <dbl>, `2000` <dbl>, `1999` <dbl>,
#   `1998` <dbl>, `1997` <dbl>, `1996` <dbl>, `1995` <dbl>, `1994` <dbl>,
#   `1993` <dbl>, `1992` <dbl>, `1991` <dbl>, `1990` <dbl>, `1989` <dbl>,
#   `1988` <dbl>, `1987` <dbl>, `1986` <dbl>, `1985` <dbl>, `1984` <dbl>,
#   `1983` <dbl>, `1982` <dbl>, `1981` <dbl>, `1980` <dbl>, `1979` <dbl>,
#   `1978` <dbl>, `1977` <dbl>, `1976` <dbl>, `1975` <dbl>, `1974` <dbl>, ...
```

```
# have the fert_long_sub dataset by drop outliers
fert_long_sub <- subset(x = fert_long, fert_cons <= 1000)
```

4. What function is used to create a density plot of the **fert_cons** variable in the **fert_long** dataset? Please provide the R code and give a brief explanation of what is happening in the code.

```
library(ggplot2)
# Create density plot
ggplot(data = fert_long_sub, aes(fert_cons)) +
  geom_density() +
  xlab("Fertilizer Consumption") +
  ylab("Density") +
  theme_bw()
```



5. How can you recode the country name “Korea, Rep.” to “South Korea” in the **fert_long_sub** dataset? Please provide the R code and give a brief explanation of what is happening in the code.

```
# Recode country == "Korea, Rep." to "South Korea"
fert_long_sub$country[fert_long_sub$country ==
"Korea, Rep."] <- "South Korea"
head(fert_long_sub)
```

```
# A tibble: 6 x 61
  country    iso2c iso3c `2022` `2021` `2020` `2019` `2018` `2017` `2009` `2008`
  <chr>      <chr> <chr>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
2 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
3 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
4 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
5 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
6 Africa Ea~ ZH    AFE       NA   28.8   29.3   24.7   23.7   24.2   17.0   17.2
# i 50 more variables: `2007` <dbl>, `2006` <dbl>, `2005` <dbl>, `2004` <dbl>,
#   `2003` <dbl>, `2002` <dbl>, `2001` <dbl>, `2000` <dbl>, `1999` <dbl>,
#   `1998` <dbl>, `1997` <dbl>, `1996` <dbl>, `1995` <dbl>, `1994` <dbl>,
#   `1993` <dbl>, `1992` <dbl>, `1991` <dbl>, `1990` <dbl>, `1989` <dbl>,
#   `1988` <dbl>, `1987` <dbl>, `1986` <dbl>, `1985` <dbl>, `1984` <dbl>,
#   `1983` <dbl>, `1982` <dbl>, `1981` <dbl>, `1980` <dbl>, `1979` <dbl>,
#   `1978` <dbl>, `1977` <dbl>, `1976` <dbl>, `1975` <dbl>, `1974` <dbl>, ...
```

6. How is the **fert_cons_log** variable created in the **fert_long_sub** dataset? Please provide the R code and give a brief explanation of what is happening in the code.

```
#make a new variable based on the old variable
fert_long_sub$fert_cons_log <- log(fert_long_sub$fert_cons)
summary(fert_long_sub$fert_cons_log)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -Inf   2.899   4.410    -Inf   5.057   6.903
```

7. What is the purpose of creating the **fert_cons_group** variable in the **fert_long_sub** dataset?

```
# The purpose is to create numeric factor levels variable
# Attach fert_long_sub data frame
attach(fert_long_sub)
# Created new fert_cons_group variable based on # fert_cons
fert_long_sub$fert_cons_group[fert_cons < 18] <- 1
```

```
Warning: Unknown or uninitialised column: `fert_cons_group`.
```

```
fert_long_sub$fert_cons_group[fert_cons >= 18 & fert_cons < 81] <- 2
fert_long_sub$fert_cons_group[fert_cons >= 81 & fert_cons < 158] <- 3
fert_long_sub$fert_cons_group[fert_cons >= 158] <- 4
fert_long_sub$fert_cons_group[is.na(fert_cons)] <- NA
# Detach data frame
detach(fert_long_sub)
```

```
summary(fert_long_sub$fert_cons_group)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
1.000   1.750   3.000   2.501   3.000   4.000 
```

8. How can you convert the **fert_cons_group** variable into a factor variable with labels "low", "medium low", "medium high", and "high"? Please provide the R code and give a brief explanation of what is happening in the code.

```
# Create vector of factor level labels
fc_labels <- c("low", "medium low", "medium high", "high")
# Convert fert_cons_group to a factor
fert_long_sub$fert_cons_group <- factor(fert_long_sub$fert_cons_group, labels = fc_la
```

9. What is the purpose of using the **countrycode** package in the **iso2c** variable creation? The **countrycode** package in R is used to convert country names or codes into ISO 3166-1 alpha-2 country codes, providing consistency and compatibility for data analysis across different datasets or systems.

10. How can you merge the **fin_regulator**, **disprop_data**, and **fert_long_sub** datasets based on the **iso2c** variable? Please provide the R code and give a brief explanation of what is happening in the code.

```
library(rio)
# Place the URL into the object fin_url
fin_url <- "https://bit.ly/2xlQ2j5"
# Download data
fin_regulator <- import(fin_url, format = "csv")
# load countrycode
library(countrycode)
fin_regulator$iso2c <- countrycode(fin_regulator$country,
origin = "country.name", destination = "iso2c")
head(fin_regulator)
```

```
  idn     country year reg_4state iso2c
1   1 Afghanistan 1987          1    AF
2   1 Afghanistan 1988          1    AF
3   1 Afghanistan 1989          1    AF
4   1 Afghanistan 1990          1    AF
5   1 Afghanistan 1991          1    AF
6   1 Afghanistan 1992          1    AF
```

```
# Place shortened URLinto url object
url <- "http://bit.ly/14aSjxB"
# Download data
disprop_data <- rio::import(url, format = "csv")
# Show variable names
names(disprop_data)
```

```
[1] "country"            "iso2c"              "year"
[4] "disproportionality"
```

```
# Merge fin_regulator and disprop_data
merged_data_2 <- merge(fin_regulator, disprop_data, union("iso2c", "year"), all = TRU
# Merge combined data frame with fert_long_sub
merged_data_2 <- merge(merged_data_2, fert_long_sub, union("iso2c", "year"), all = TR
names(merged_data_2)
```

```
 [1] "iso2c"              "year"               "idn"
 [4] "country.x"          "reg_4state"         "country.y"
 [7] "disproportionality" "country"            "iso3c"
[10] "2022"               "2021"               "2020"
[13] "2019"               "2018"               "2017"
[16] "2009"               "2008"               "2007"
[19] "2006"               "2005"               "2004"
[22] "2003"               "2002"               "2001"
[25] "2000"               "1999"               "1998"
[28] "1997"               "1996"               "1995"
[31] "1994"               "1993"               "1992"
[34] "1991"               "1990"               "1989"
[37] "1988"               "1987"               "1986"
[40] "1985"               "1984"               "1983"
[43] "1982"               "1981"               "1980"
[46] "1979"               "1978"               "1977"
[49] "1976"               "1975"               "1974"
[52] "1973"               "1972"               "1971"
[55] "1970"               "1969"               "1968"
[58] "1967"               "1966"               "1965"
[61] "1964"               "1963"               "1962"
[64] "1961"               "1960"               "fert_cons"
[67] "fert_cons_log"      "fert_cons_group"
```

The link to Github repository: https://github.com/weiziwu/Week_7_HW.git