Untitled

Anous Ahmed

#Question 1 - EDA

Perform as many of the following tasks as you can. If you cannot perform a task, write a brief explanation of how you would perform it. The questions that are starred (*) require only text responses.

1. Load the dataset and display the first few rows.

```
library("tidyverse")
```

```
Warning: package 'ggplot2' was built under R version 4.4.2

Warning: package 'tidyr' was built under R version 4.4.2

Warning: package 'forcats' was built under R version 4.4.2

-- 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 beco
```

```
library("stevedata")
```

```
Warning: package 'stevedata' was built under R version 4.4.2
```

```r
eq_passengercars = eq_passengercars
```

```r
head(eq_passengercars, 10)
```

```
# A tibble: 10 x 6
   country   ccode category                                    type   year    value
   <chr>     <dbl> <fct>                                       <fct> <dbl>    <dbl>
 1 Australia   900 Export Quality Index                        51. ~  1963    1.05
 2 Australia   900 Export quality 95 percent interval - lo~ 51. ~  1963    0.902
 3 Australia   900 Export quality 95 percent interval - up~ 51. ~  1963    1.05
 4 Australia   900 Unit value of exports                       51. ~  1963    0.771
 5 Australia   900 Unit value 95 percent interval - lower ~ 51. ~  1963    0.527
 6 Australia   900 Unit value 95 percent interval - upper ~ 51. ~  1963    0.829
 7 Australia   900 Trade value of exports                      51. ~  1963 2510.
 8 Cambodia    811 Export Quality Index                        51. ~  1963   NA
 9 Cambodia    811 Export quality 95 percent interval - lo~ 51. ~  1963   NA
10 Cambodia    811 Export quality 95 percent interval - up~ 51. ~  1963   NA
```

2. Display the structure and summary of the dataset.

```r
# check the structure
str(eq_passengercars)
```

```
tibble [60,424 x 6] (S3: tbl_df/tbl/data.frame)
 $ country : chr [1:60424] "Australia" "Australia" "Australia" "Australia" ...
 $ ccode   : num [1:60424] 900 900 900 900 900 900 900 811 811 811 ...
 $ category: Factor w/ 7 levels "Export Quality Index",..: 1 2 3 4 5 6 7 1 2 3 ...
 $ type    : Factor w/ 1 level "51. Transport equipment, Passenger cars": 1 1 1 1 1 1 1 1 1 1
 $ year    : num [1:60424] 1963 1963 1963 1963 1963 ...
 $ value   : num [1:60424] 1.052 0.902 1.054 0.771 0.527 ...
```

```r
# check for rows and columns
dim(eq_passengercars)
```

```
[1] 60424     6
```

```r
# check for rows and columns
summary(eq_passengercars)
```

```
  country              ccode
Length:60424       Min.   :  2.0
Class :character   1st Qu.:235.0
Mode  :character   Median :435.0
                   Mean   :435.5
                   3rd Qu.:630.0
                   Max.   :950.0
                   NA's   :364
                                                  category
Export Quality Index                                  :8632
Export quality 95 percent interval - lower bound:8632
Export quality 95 percent interval - upper bound:8632
Unit value of exports                                 :8632
Unit value 95 percent interval - lower bound          :8632
Unit value 95 percent interval - upper bound          :8632
Trade value of exports                                :8632
                                        type           year
51. Transport equipment, Passenger cars:60424   Min.   :1963
                                                1st Qu.:1976
                                                Median :1988
                                                Mean   :1988
                                                3rd Qu.:2001
                                                Max.   :2014


    value
Min.   :        0
1st Qu.:        1
Median :        1
Mean   :   225616
3rd Qu.:        1
Max.   :143605360
NA's   :18515
```

3. Is the dataset tidy? If not, what makes it messy? If you were to make it tidy, what would be the "unit of observation" for each row? (*)

The dataset is not tidy because the variables, *category* and *type* have multiple pieces of information that must be separated and inserted in different columns. The *category* column contains both a measurement type and and confidence interval descriptors. On the other hand, the *type* column contains information on numerical data, general categories and product types. Therefore, the dataset would need to be tidy in order to carry out further analysis.

The *unit of observation* for each row would be specific measurement (e.g., export quality index, unit value of exports) for a given country, sector, product type, year, and interval.

4. If you think the data is not tidy, tidy the dataset.

```
eq_passengercars_tidy = eq_passengercars %>%
  separate(type, into = c("industry_sector", "product_type"), sep = ",", remove = T) %>%
  separate(industry_sector, into = c("sector_code", "sector_name"), sep = "\\. ", remove = T
  separate(category, into = c("measurement", "interval"), sep = " - ", fill = "right")
```

**