Assignment1-R Programming_Group10

Group 10 - Anumol, Bruno, Clifford, Rushia, Skanda

2023-02-12

#Install Packages install.packages('tidyverse') install.packages('dplyr') install.packages('magrittr') install.packages('ggplot') install.packages('xycorr')

Import Data Set Fuel_Consumption_2000.2022.csv

```
Fuel_Consumption_2000.2022 <- read.csv(
"C:/Users/Rushia/Downloads/Fuel_Consumption_2000-2022.csv/Fuel_Consumption_2000-2022.csv")

fuel=Fuel_Consumption_2000.2022
```

Print the structure of dataset

```
str(fuel)
```

```
## 'data.frame':    22556 obs. of  13 variables:
##  $ YEAR            : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ MAKE            : chr  "ACURA" "ACURA" "ACURA" "ACURA" ...
##  $ MODEL           : chr  "1.6EL" "1.6EL" "3.2TL" "3.5RL" ...
##  $ VEHICLE.CLASS   : chr  "COMPACT" "COMPACT" "MID-SIZE" "MID-SIZE" ...
##  $ ENGINE.SIZE     : num  1.6 1.6 3.2 3.5 1.8 1.8 1.8 3 3.2 1.8 ...
##  $ CYLINDERS       : int  4 4 6 6 4 4 4 6 6 4 ...
##  $ TRANSMISSION    : chr  "A4" "M5" "AS5" "A4" ...
##  $ FUEL            : chr  "X" "X" "Z" "Z" ...
##  $ FUEL.CONSUMPTION: num  9.2 8.5 12.2 13.4 10 9.3 9.4 13.6 13.8 11.4 ...
##  $ HWY..L.100.km.  : num  6.7 6.5 7.4 9.2 7 6.8 7 9.2 9.1 7.2 ...
##  $ COMB..L.100.km. : num  8.1 7.6 10 11.5 8.6 8.2 8.3 11.6 11.7 9.5 ...
##  $ COMB..mpg.      : int  35 37 28 25 33 34 34 24 24 30 ...
##  $ EMISSIONS       : int  186 175 230 264 198 189 191 267 269 218 ...
```

List the variables in your dataset

```
names(fuel)
```

```
##  [1] "YEAR"             "MAKE"             "MODEL"            "VEHICLE.CLASS"
##  [5] "ENGINE.SIZE"      "CYLINDERS"        "TRANSMISSION"     "FUEL"
##  [9] "FUEL.CONSUMPTION" "HWY..L.100.km."   "COMB..L.100.km."  "COMB..mpg."
## [13] "EMISSIONS"
```

Print the top 15 rows of your dataset

```
head(fuel, n=5)
```

```
##   YEAR  MAKE   MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS TRANSMISSION FUEL
## 1 2000 ACURA   1.6EL       COMPACT         1.6         4           A4    X
## 2 2000 ACURA   1.6EL       COMPACT         1.6         4           M5    X
## 3 2000 ACURA   3.2TL      MID-SIZE         3.2         6          AS5    Z
## 4 2000 ACURA   3.5RL      MID-SIZE         3.5         6           A4    Z
## 5 2000 ACURA INTEGRA    SUBCOMPACT         1.8         4           A4    X
##   FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg. EMISSIONS
## 1              9.2            6.7             8.1         35       186
## 2              8.5            6.5             7.6         37       175
## 3             12.2            7.4            10.0         28       230
## 4             13.4            9.2            11.5         25       264
## 5             10.0            7.0             8.6         33       198
```

Write a user defined function using any of the variables from the data set

```
FuelConsumption_Year <- function(FUEL.CONSUMPTION) {
  2021 - FUEL.CONSUMPTION
}
print(head(FuelConsumption_Year(fuel$FUEL.CONSUMPTION), N=10))
```

```
## [1] 2011.8 2012.5 2008.8 2007.6 2011.0 2011.7
```

Use data manipulation techniques and filter rows based on any logical criteria that exist in dataset

```
# Attach tidyverse packages to use data manipulation, reading, transforming and visualizing dataset
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
filterfuel= as.data.frame(filter(fuel,fuel$MAKE=="BMW"))
fuel %>% filter(MAKE=="BMW") %>% slice_head(n = 10)
```

```
##    YEAR MAKE           MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS TRANSMISSION
## 1  2000  BMW 323 CONVERTIBLE       COMPACT         2.5         6           A5
## 2  2000  BMW 323 CONVERTIBLE       COMPACT         2.5         6           M5
```

```
## 3  2000  BMW           323Ci       COMPACT         2.5         6           A5
## 4  2000  BMW           323Ci       COMPACT         2.5         6           M5
## 5  2000  BMW            323i       COMPACT         2.5         6           A5
## 6  2000  BMW            323i       COMPACT         2.5         6           M5
## 7  2000  BMW           328Ci       COMPACT         2.8         6           A5
## 8  2000  BMW           328Ci       COMPACT         2.8         6           M5
## 9  2000  BMW            328i       COMPACT         2.8         6           A5
## 10 2000  BMW            328i       COMPACT         2.8         6           M5
##    FUEL FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg. EMISSIONS
## 1     Z             13.0            8.9            11.2         25       258
## 2     Z             12.4            8.6            10.7         26       246
## 3     Z             12.3            7.8            10.3         27       237
## 4     Z             11.5            7.5             9.7         29       223
## 5     Z             12.3            7.8            10.3         27       237
## 6     Z             11.5            7.5             9.7         29       223
## 7     Z             12.4            7.9            10.4         27       239
## 8     Z             11.5            7.5             9.7         29       223
## 9     Z             12.4            7.9            10.4         27       239
## 10    Z             11.5            7.5             9.7         29       223
```

Identify the dependent & independent variables and use reshaping techniques and create a new data frame by joining those variables from dataset

```r
# Create a new data frame with MAKE and MODEL column
Col_Make_Model <- cbind(fuel$MAKE, fuel$MODEL)
print(head(Col_Make_Model, n = 4))
```

```
##      [,1]    [,2]
## [1,] "ACURA" "1.6EL"
## [2,] "ACURA" "1.6EL"
## [3,] "ACURA" "3.2TL"
## [4,] "ACURA" "3.5RL"
```

```r
# Find vehicles whose make is AUDI
df_make_is_Audi <- fuel %>%
                filter(MAKE=="Audi")

# Find vehicles whose model is A4
df_model_is_A4 <- fuel %>%
                filter(MAKE=="A4")

# Create a new data frame with patients whose age under 30 by merging 2 prepared data frame
df_make_Audi_model_A4 <- rbind(df_make_is_Audi, df_model_is_A4)
print(head(df_make_Audi_model_A4, n = 2))
```

```
##   YEAR MAKE      MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS TRANSMISSION FUEL
## 1 2018 Audi         A3    Subcompact           2         4          AM7    X
## 2 2018 Audi A3 quattro    Subcompact           2         4          AM6    X
##   FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg. EMISSIONS
## 1              9.1            6.8             8.0         35       188
## 2              9.7            7.5             8.7         32       205
```

Remove missing values in your dataset

```
newfuel=fuel
newfuel%>%filter(!is.na(MAKE)) %>% slice_head(n = 5)
```

```
##   YEAR  MAKE   MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS TRANSMISSION FUEL
## 1 2000 ACURA   1.6EL       COMPACT         1.6         4           A4    X
## 2 2000 ACURA   1.6EL       COMPACT         1.6         4           M5    X
## 3 2000 ACURA   3.2TL      MID-SIZE         3.2         6          AS5    Z
## 4 2000 ACURA   3.5RL      MID-SIZE         3.5         6           A4    Z
## 5 2000 ACURA INTEGRA    SUBCOMPACT         1.8         4           A4    X
##   FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg. EMISSIONS
## 1              9.2            6.7             8.1         35       186
## 2              8.5            6.5             7.6         37       175
## 3             12.2            7.4            10.0         28       230
## 4             13.4            9.2            11.5         25       264
## 5             10.0            7.0             8.6         33       198
```

Identify and remove duplicated data in dataset

```
fuel%>%distinct(YEAR, MAKE, MODEL, ENGINE.SIZE,TRANSMISSION,FUEL, keep_all=TRUE) %>% slice_head(n = 10)
```

```
##    YEAR  MAKE             MODEL ENGINE.SIZE TRANSMISSION FUEL keep_all
## 1  2000 ACURA             1.6EL         1.6           A4    X     TRUE
## 2  2000 ACURA             1.6EL         1.6           M5    X     TRUE
## 3  2000 ACURA             3.2TL         3.2          AS5    Z     TRUE
## 4  2000 ACURA             3.5RL         3.5           A4    Z     TRUE
## 5  2000 ACURA           INTEGRA         1.8           A4    X     TRUE
## 6  2000 ACURA           INTEGRA         1.8           M5    X     TRUE
## 7  2000 ACURA INTEGRA GSR/TYPE R         1.8           M5    Z     TRUE
## 8  2000 ACURA               NSX         3.0          AS4    Z     TRUE
## 9  2000 ACURA               NSX         3.2           M6    Z     TRUE
## 10 2000  AUDI                A4         1.8           A5    Z     TRUE
```

Reorder multiple rows in descending order

```
fuel%>%arrange(desc(EMISSIONS)) %>% slice_head(n = 6)
```

```
##   YEAR        MAKE                   MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS
## 1 2021     Bugatti        Chiron Pur Sport    Two-seater         8.0        16
## 2 2022     Bugatti        Chiron Pur Sport    Two-seater         8.0        16
## 3 2022     Bugatti      Chiron Super Sport    Two-seater         8.0        16
## 4 2003     FERRARI                   ENZO    TWO-SEATER         6.0        12
## 5 2021 Lamborghini    Aventador Sian Coupe    Two-seater         6.5        12
## 6 2021 Lamborghini Aventador Sian Roadster    Two-seater         6.5        12
##   TRANSMISSION FUEL FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg.
## 1           AM7    Z              30.3           20.9            26.1         11
## 2           AM7    Z              30.3           20.9            26.1         11
```

```
## 3          AM7    Z             30.3           20.9            26.1         11
## 4          AS6    Z             30.6           17.6            24.8         11
## 5          AM7    Z             28.3           16.8            23.1         12
## 6          AM7    Z             28.3           16.8            23.1         12
##   EMISSIONS
## 1       608
## 2       608
## 3       608
## 4       570
## 5       539
## 6       539
```

Rename some of the column names in dataset

```
names(fuel)[names(fuel) == "YEAR"] <- "Year"
```

Add new variables in data frame by using a mathematical function

```
fuel %>% filter(!is.na(EMISSIONS)) %>% mutate(mathf= 2000 - EMISSIONS) %>% slice_head(n=10)
```

```
##    Year  MAKE              MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS
## 1  2000 ACURA              1.6EL       COMPACT         1.6         4
## 2  2000 ACURA              1.6EL       COMPACT         1.6         4
## 3  2000 ACURA              3.2TL      MID-SIZE         3.2         6
## 4  2000 ACURA              3.5RL      MID-SIZE         3.5         6
## 5  2000 ACURA            INTEGRA    SUBCOMPACT         1.8         4
## 6  2000 ACURA            INTEGRA    SUBCOMPACT         1.8         4
## 7  2000 ACURA INTEGRA GSR/TYPE R    SUBCOMPACT         1.8         4
## 8  2000 ACURA                NSX    SUBCOMPACT         3.0         6
## 9  2000 ACURA                NSX    SUBCOMPACT         3.2         6
## 10 2000  AUDI                 A4       COMPACT         1.8         4
##    TRANSMISSION FUEL FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km. COMB..mpg.
## 1            A4    X              9.2            6.7             8.1         35
## 2            M5    X              8.5            6.5             7.6         37
## 3           AS5    Z             12.2            7.4            10.0         28
## 4            A4    Z             13.4            9.2            11.5         25
## 5            A4    X             10.0            7.0             8.6         33
## 6            M5    X              9.3            6.8             8.2         34
## 7            M5    Z              9.4            7.0             8.3         34
## 8           AS4    Z             13.6            9.2            11.6         24
## 9            M6    Z             13.8            9.1            11.7         24
## 10           A5    Z             11.4            7.2             9.5         30
##    EMISSIONS mathf
## 1        186  1814
## 2        175  1825
## 3        230  1770
## 4        264  1736
## 5        198  1802
## 6        189  1811
## 7        191  1809
```

```
## 8        267  1733
## 9        269  1731
## 10       218  1782
```

Create a training set using random number generator engine

```
#Extract 3 random rows without replacement
fuel %>% sample_n(8, replace=FALSE)
```

```
##   Year          MAKE                 MODEL            VEHICLE.CLASS ENGINE.SIZE
## 1 2012       PORSCHE 911 CARRERA CABRIOLET              MINICOMPACT         3.6
## 2 2016        NISSAN            MURANO AWD STATION WAGON - MID-SIZE         3.5
## 3 2022        Toyota     Corolla Cross AWD               SUV: Small         2.0
## 4 2017        NISSAN              ROGUE AWD              SUV - SMALL         2.5
## 5 2013 MERCEDES-BENZ  S 350 BLUETEC 4MATIC                FULL-SIZE         3.0
## 6 2008           BMW         X6 XDRIVE 35i                      SUV         3.0
## 7 2005           BMW     645Ci CONVERTIBLE               SUBCOMPACT         4.4
## 8 2010          AUDI   TT ROADSTER QUATTRO               TWO-SEATER         2.0
##   CYLINDERS TRANSMISSION FUEL FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km.
## 1         6          AS7    Z             11.3            7.4             9.5
## 2         6          AV7    X             11.2            8.3             9.9
## 3         4         AV10    X              8.1            7.4             7.8
## 4         4           AV    X              9.6            7.4             8.6
## 5         6          AS7    D             10.2            6.3             8.4
## 6         6           A6    Z             14.4           10.0            12.4
## 7         8          AM6    Z             15.6            9.4            12.8
## 8         4          AS6    Z              9.7            7.1             8.5
##   COMB..mpg. EMISSIONS
## 1         30       218
## 2         29       232
## 3         36       182
## 4         33       203
## 5         34       227
## 6         23       285
## 7         22       294
## 8         33       196
```

```
#Select top 2 rows ordered by a variable
fuel %>% top_n(2, FUEL.CONSUMPTION)
```

```
##   Year      MAKE                      MODEL   VEHICLE.CLASS ENGINE.SIZE
## 1 2003   FERRARI                       ENZO      TWO-SEATER           6
## 2 2015 CHEVROLET EXPRESS 3500 PASSENGER FFV VAN - PASSENGER           6
## 3 2015       GMC  SAVANA 3500 PASSENGER FFV VAN - PASSENGER           6
##   CYLINDERS TRANSMISSION FUEL FUEL.CONSUMPTION HWY..L.100.km. COMB..L.100.km.
## 1        12          AS6    Z             30.6           17.6            24.8
## 2         8           A6    E             30.6           20.6            26.1
## 3         8           A6    E             30.6           20.6            26.1
##   COMB..mpg. EMISSIONS
## 1         11       570
## 2         11       418
## 3         11       418
```

Print the summary statistics of dataset

```
fuel %>% group_by(MAKE) %>% summarise(mean(FUEL.CONSUMPTION)) %>% slice_head(n=6)
```

```
## # A tibble: 6 x 2
##   MAKE         `mean(FUEL.CONSUMPTION)`
##   <chr>                           <dbl>
## 1 ACURA                            10.9
## 2 ALFA ROMEO                       10.3
## 3 ASTON MARTIN                     17.6
## 4 AUDI                             12.6
## 5 Acura                            11.0
## 6 Alfa Romeo                       11.3
```

Use any of the numerical variables from the dataset and perform the following statistical functions: Mean, Median, Mode, Range

```
# Find mean of emissions
mean(fuel[["EMISSIONS"]])
```

```
## [1] 250.0685
```

```
# Find median of emissions
median(fuel[["EMISSIONS"]])
```

```
## [1] 243
```

```
# Find mode of emissions
cal_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
mode(fuel[["Emissions"]])
```

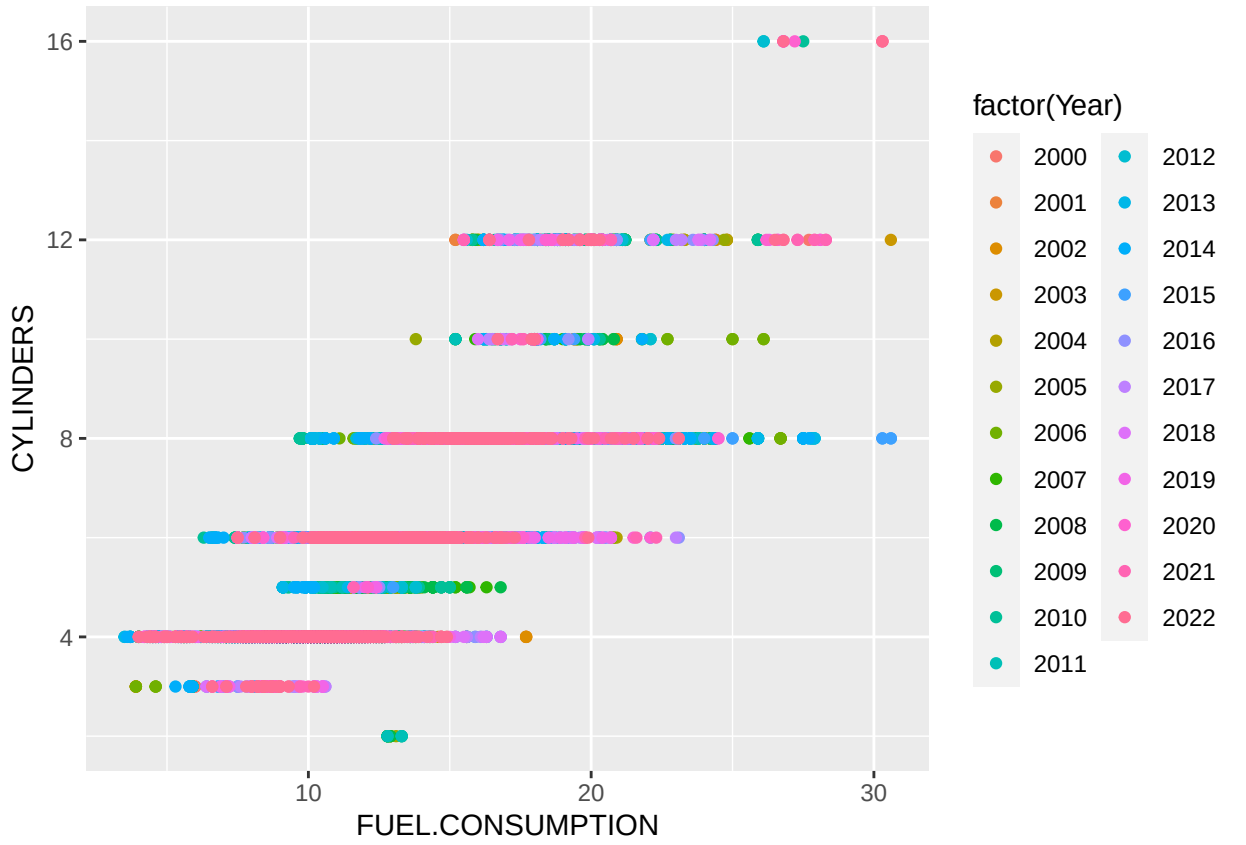
```
## [1] "NULL"
```

```
# Find range of claim
print(range(fuel[["EMISSIONS"]]))
```

```
## [1]  83 608
```

Plot a scatter plot for any 2 variables in dataset

```
ggplot(data = fuel,aes(x = FUEL.CONSUMPTION,y = CYLINDERS ,col = factor(Year)))+geom_point()
```



Plot a bar plot for any 2 variables in dataset

```
ggplot(data = fuel,aes(x = Year,fill = factor(MAKE)))+geom_bar()
```

Find the correlation between any 2 variables by applying least square linear regression model

```
xycorr = cor(fuel$Year, fuel$EMISSIONS, method="pearson")
print(xycorr)
```

```
## [1] -0.04786904
```