AARYAN PUJARA
102599490

ASSIGNMENT 1

Question 1

1.1 Weight versus age of chicks on different diets

Description

The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.

Usage

ChickWeight

Format

An object of class c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame") containing the following columns:

weight

a numeric vector giving the body weight of the chick (gm).

Time

a numeric vector giving the number of days since birth when the measurement was made.

Chick

an ordered factor with levels 18 < ... < 48 giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

Diet

a factor with levels 1, ..., 4 indicating which experimental diet the chick received.

Details

The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

This dataset was originally part of package nlme, and that has methods (including for [, as.data.frame, plot and print) for its grouped-data classes.

1.2 We are going to Find the variance in chick weight by comparing time and weight, we are also going to find when was the last time each chick was weighed to see if any of the readings are missing. We are also going to summarise all the data to see how many readings were taken over time.

```
> tapply(ChickWeight$Time,
+        ChickWeight$Chick, FUN=var)
      18       16       15       13        9       20       10        8       17       19
 2.00000 18.66667 24.00000 50.08333 50.08333 50.08333 50.08333 44.00000 50.08333 50.08333
       4        6       11        3        1       12        2        5       14        7
50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333
      24       30       22       23       27       28       26       25       29       21
50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333
      33       37       36       31       39       38       32       40       34       35
50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333
      44       45       43       41       47       49       46       50       42       48
36.66667 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333 50.08333
> tapply(ChickWeight$Time,
+        ChickWeight$Chick,
+        FUN=function(x)diff(range(x)))
18 16 15 13  9 20 10  8 17 19  4  6 11  3  1 12  2  5 14  7 24 30 22 23 27 28 26 25 29 21 33
 2 12 14 21 21 21 21 20 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21
37 36 31 39 38 32 40 34 35 44 45 43 41 47 49 46 50 42 48
21 21 21 21 21 21 21 21 21 18 21 21 21 21 21 21 21 21 21
> summary(ChickWeight$Chicks)
Length  Class   Mode
     0   NULL   NULL
> summary(ChickWeight$Chick)
18 16 15 13  9 20 10  8 17 19  4  6 11  3  1 12  2  5 14  7 24 30 22 23 27 28 26 25 29 21 33
 2  7  8 12 12 12 12 11 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12
37 36 31 39 38 32 40 34 35 44 45 43 41 47 49 46 50 42 48
12 12 12 12 12 12 12 12 12 10 12 12 12 12 12 12 12 12 12
>
```

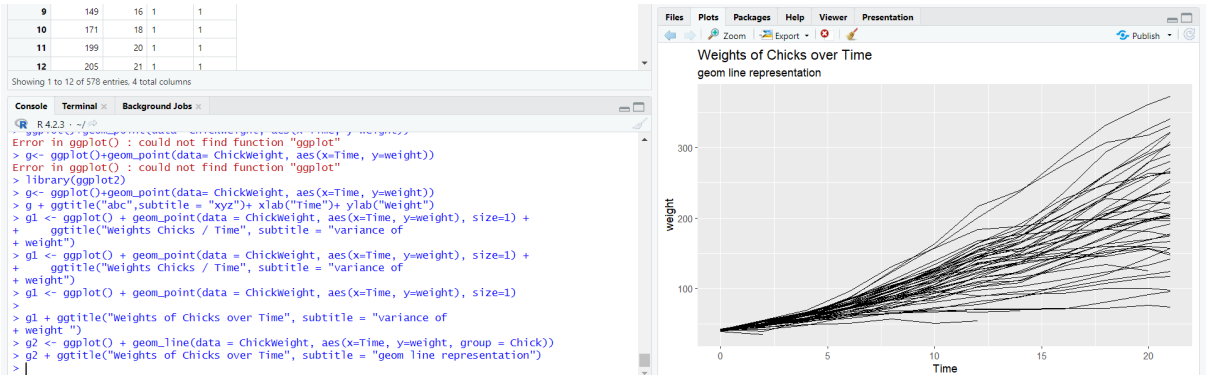
After analysing all this data we find out that chick 8,15,16,18,44 have missing weight records.

1.3

```
                                  (Other):506
> summary(ChickWeight)
     weight           Time           Chick     Diet
 Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
 1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
 Median :103.0   Median :10.00   20     : 12   3:120
 Mean   :121.8   Mean   :10.72   10     : 12   4:118
 3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
 Max.   :373.0   Max.   :21.00   19     : 12
                                 (Other):506
```

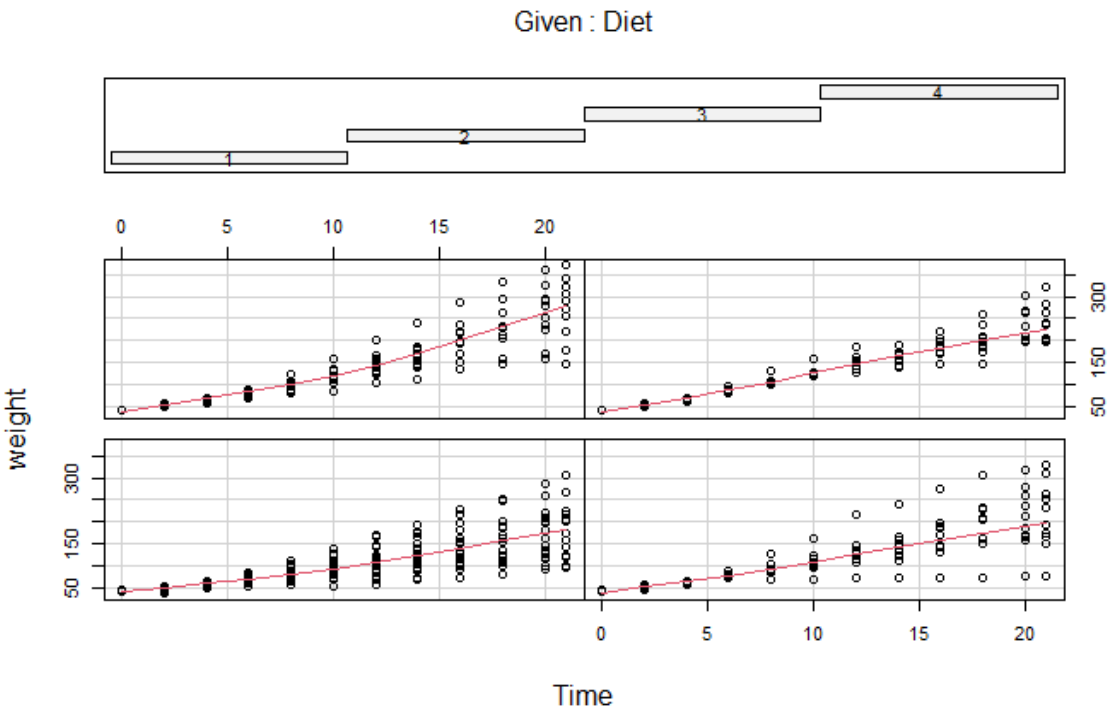
1.4

We can see that there is a regular time interval of 2 days in this plot for recording the weight. As the time passes on the weight of most the chicks increases and so does the variance of weights but there is a decrease in weight for one of the chicks.



There are 4 lines in the plot above where we can conclude that 4 datasets of 4 chicks are incomplete.

```
> g4 <- coplot(weight ~ Time | Diet, data = ChickWeight, panel = panel.smooth)
```



The observations from this plot are:

- Diet 1 gives lightest chicks whereas Diet 3 gives heaviest chicks.
- Some of the chicks die prematurely

Question 2

2.1

```
> str(Cleaning1)
'data.frame':   101 obs. of  16 variables:
 $ ID     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE    : int  18 23 39 24 27 26 26 26 28 24 ...
 $ GENDER : chr  "Male" "Female" "Male" "Male" ...
 $ YRSTUDY: int  3 7 10 6 16 16 10 8 9 6 ...
 $ Q1     : int  8 15 9 10 10 10 14 12 9 10 ...
 $ Q2     : int  14 21 12 15 20 16 18 17 15 13 ...
 $ Q3     : int  16 20 14 15 21 16 16 11 21 23 ...
 $ Q4     : chr  "19" "23" "12" "18" ...
 $ Q5     : chr  "18" "19" "22" "17" ...
 $ Q6     : int  14 16 20 15 23 18 11 10 23 19 ...
 $ Q7     : int  21 20 16 20 29 20 15 18 21 24 ...
 $ Q8     : int  27 22 22 29 26 27 27 21 23 23 ...
 $ Q9     : int  25 24 26 30 27 24 25 27 20 26 ...
 $ Q10    : int  26 26 26 25 25 25 25 25 25 25 ...
 $ Q11    : int  22 18 23 27 21 25 24 24 26 19 ...
 $ Q12    : int  14 17 15 13 24 14 13 10 22 18 ...
```

- There are 4 Identity Categories which are year of study, age ,ID and Gender
- The record contains 101 Students and 16 Variables
- 13 variables are numerical and 3 are nominal

2.2

```
> summary(Cleaning1)
       ID              AGE            GENDER             YRSTUDY             Q1
 Min.   :  1   Min.   :18.00   Length:101         Min.   : 0.000   Min.   : 7.00
 1st Qu.: 26   1st Qu.:19.00   Class :character   1st Qu.: 3.000   1st Qu.: 9.00
 Median : 51   Median :23.00   Mode  :character   Median : 6.000   Median :10.00
 Mean   : 51   Mean   :23.35                      Mean   : 6.634   Mean   :10.94
 3rd Qu.: 76   3rd Qu.:26.00                      3rd Qu.: 9.000   3rd Qu.:12.00
 Max.   :101   Max.   :39.00                      Max.   :20.000   Max.   :19.00

       Q2              Q3             Q4                 Q5                 Q6
 Min.   : 9.00   Min.   :10.00   Length:101         Length:101         Min.   : 9.00
 1st Qu.:13.00   1st Qu.:15.00   Class :character   Class :character   1st Qu.:13.00
 Median :15.00   Median :18.00   Mode  :character   Mode  :character   Median :16.00
 Mean   :15.61   Mean   :18.02                                         Mean   :16.19
 3rd Qu.:19.00   3rd Qu.:21.00                                         3rd Qu.:19.00
 Max.   :28.00   Max.   :30.00                                         Max.   :26.00
 NA's   :1
       Q7              Q8              Q9             Q10             Q11
 Min.   :13.00   Min.   :19.00   Min.   :16.00   Min.   :22.00   Min.   :13.00
 1st Qu.:16.00   1st Qu.:24.00   1st Qu.:24.00   1st Qu.:23.00   1st Qu.:21.00
 Median :20.00   Median :26.00   Median :26.00   Median :24.00   Median :23.00
 Mean   :20.29   Mean   :25.45   Mean   :25.59   Mean   :23.73   Mean   :23.01
 3rd Qu.:22.00   3rd Qu.:28.00   3rd Qu.:27.00   3rd Qu.:25.00   3rd Qu.:25.00
 Max.   :52.00   Max.   :32.00   Max.   :33.00   Max.   :26.00   Max.   :52.00
                                                 NA's   :1       NA's   :1
      Q12
 Min.   :10.00
 1st Qu.:15.00
 Median :17.50
 Mean   :17.61
 3rd Qu.:20.00
 Max.   :26.00
 NA's   :1
```

There are 4 missing values in Q2,10,11,12
The max values of Q5,7 are bigger than 40

2.3

```
> all_na <- colnames(Cleaning1)[apply(Cleaning1,2,anyNA)]
> all_na
[1] "Q2"  "Q10" "Q11" "Q12"
> all_na[is.na(all_na)] = NA
```

Question 2 part 4 not done

Question 3

3.1

```
> str(Fuel_Cons_2022)
'data.frame':   967 obs. of  11 variables:
 $ Make         : chr  "Acura" "Acura" "Acura" "Acura" ...
 $ Model        : chr  "ILX" "MDX SH-AWD" "MDX SH-AWD A-SPEC" "MDX Hybri
 $ Cylinders    : int  4 6 6 6 4 4 6 4 6 6 ...
 $ Transmission: chr  "AM8" "AS9" "AS9" "AM7" ...
 $ Fuel_type    : chr  "Z" "Z" "Z" "Z" ...
 $ City_Fuel    : num  9.9 12.3 12.2 9.1 11 11.3 8.4 10.2 11.4 12 ...
 $ Hwy_Fuel     : num  7 9.2 9.5 9 8.6 9.1 8.2 7.4 7.7 8.2 ...
 $ Comb_Fuel    : num  8.6 10.9 11 9 9.9 10.3 8.4 8.9 9.8 10.3 ...
 $ Emission_co2: int  199 254 258 210 232 241 196 209 228 240 ...
 $ Rating_Co2   : int  6 4 4 5 5 5 6 5 5 5 ...
 $ Smog_Rating  : int  3 3 3 3 6 6 7 3 3 3 ...
>
```

The fuel consumption for each vehicle varies by

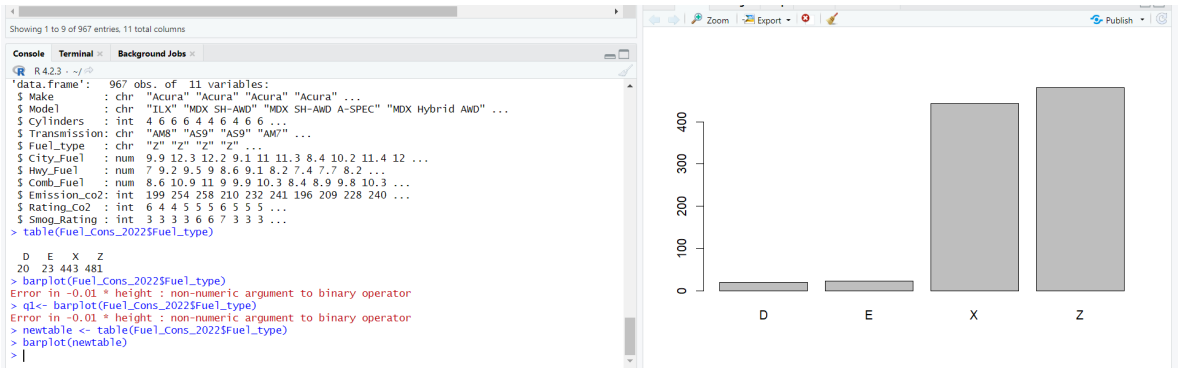
- their Model, Fuel_type
- the location of purchasing the fuel was (City_Fuel, Hwy_Fuel or Comb_Fuel)
- size of Cylinders
- There are 7 Numeric and 4 Nominal Variables

3.2 From the data produced we can say that FUel D is least popular and Fuel Z is the most.

```
> table(Fuel_Cons_2022$Fuel_type)

  D   E   X   Z
 20  23 443 481
>
```

3.3



3.4

```
> library(psych)
> psych::describeBy(Fuel_Cons_2022$City_Fuel, Fuel_Cons_2022$Cylinders)

 Descriptive statistics by group
group: 3
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 12 8.52 0.93    8.6    8.58 0.52 6.6 9.8   3.2 -0.61     -0.5 0.27
------------------------------------------------------------------------
group: 4
   vars   n mean   sd median trimmed  mad min  max range  skew kurtosis   se
X1    1 425 9.89 1.78   10.1   10.03 1.63 4.2 14.3  10.1 -0.85     1.01 0.09
------------------------------------------------------------------------
group: 5
   vars n mean sd median trimmed mad  min  max range skew kurtosis se
X1    1 2 12.1  0   12.1    12.1   0 12.1 12.1     0  NaN      NaN  0
------------------------------------------------------------------------
group: 6
   vars   n  mean  sd median trimmed  mad min  max range skew kurtosis  se
X1    1 295 12.89 1.7   12.8    12.8 1.33 7.5 22.1  14.6 1.17      5.6 0.1
------------------------------------------------------------------------
group: 8
   vars   n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 202 16.25 2.14   15.8   15.96 1.63 12.8 24.5  11.7 1.29     1.66 0.15
------------------------------------------------------------------------
group: 10
   vars n  mean   sd median trimmed mad  min max range  skew kurtosis   se
X1    1 6 17.83 0.26     18   17.83   0 17.5  18   0.5 -0.54    -1.96 0.11
------------------------------------------------------------------------
group: 12
   vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis  se
X1    1 23 20.61 3.37     20   20.32 1.19 15.5 28.1  12.6 1.09     0.24 0.7
------------------------------------------------------------------------
group: 16
   vars n mean   sd median trimmed mad  min  max range skew kurtosis  se
X1    1 2   27 0.28     27      27 0.3 26.8 27.2   0.4    0    -2.75 0.2
> |
```

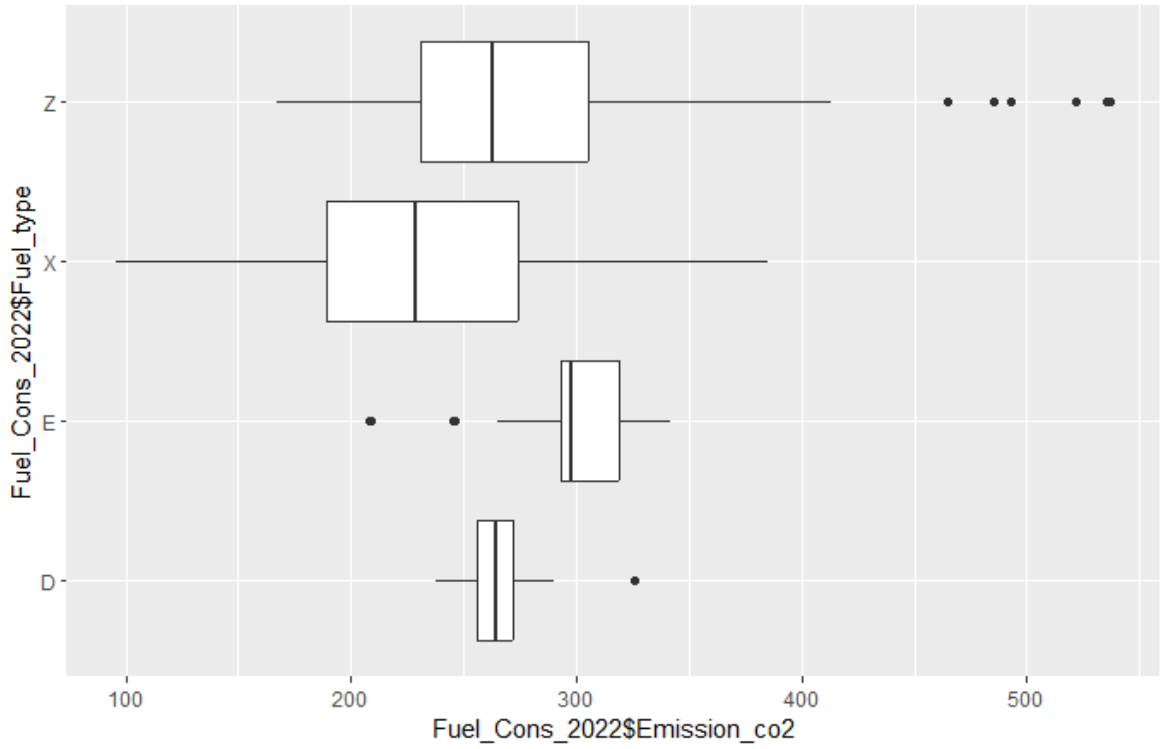
- The more the cylendirs in a car the more is the fuel consumption.
- Std deviation is highest for Group 12 as there is more data there and least for Group 5 because of scarce data.

3.5

```
X1    1 2   27 0.28     27      27 0.3 26.8 27.2   0.4    0    -2.75 0.2
> list <- subset(Fuel_Cons_2022, Fuel_Cons_2022$Smog_Rating == 7 & Fuel_Cons_2022$Transmission ==
"A6" & Fuel_Cons_2022$Fuel_type =="X" )
> list
          Make       Model Cylinders Transmission Fuel_type City_Fuel Hwy_Fuel Comb_Fuel
208 Chevrolet     Equinox         4           A6         X       8.9      7.7       8.4
209 Chevrolet Equinox AWD         4           A6         X       9.3      8.0       8.7
    Emission_co2 Rating_Co2 Smog_Rating
208          196          6           7
209          204          6           7
> |
```

3.6

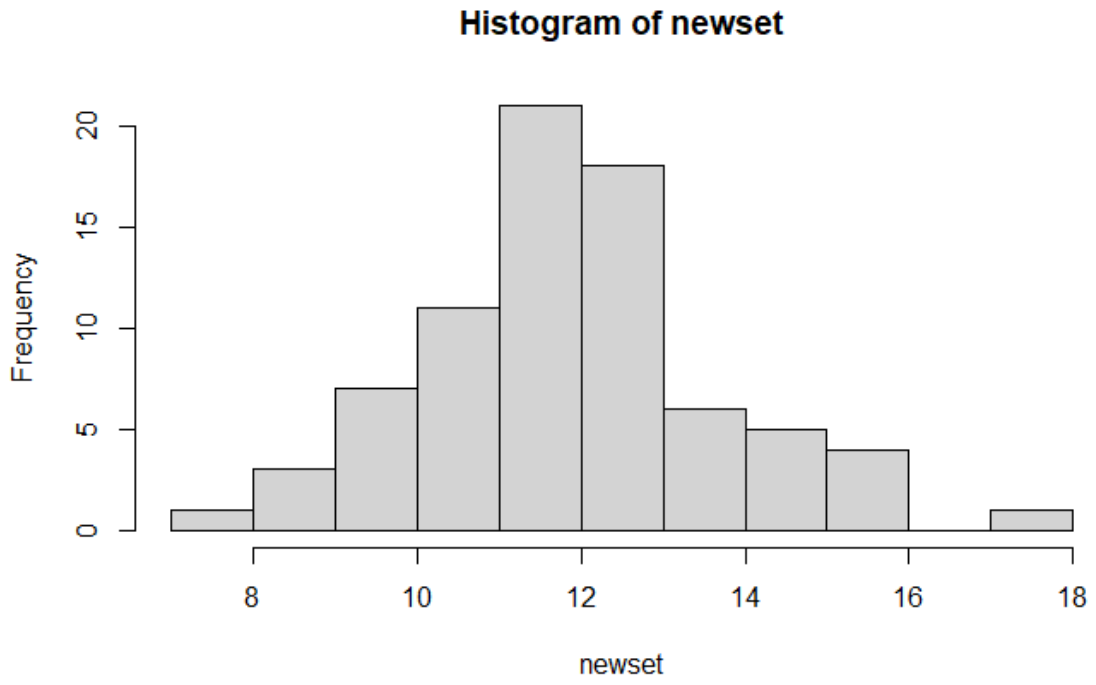
```
ggplot() + geom_boxplot(aes(x=Fuel_Cons_2022$Emission_co2, y=
Fuel_Cons_2022$Fuel_type))
```



- Fuel X is the most efficient fuel as the emission levels of the fuel type is lowest of all and the minimum value of emission lies behind 100.
- Fuel type z is the least efficient fuel with a high range of emission starting after 150 and extending post 400
- Fuel type E has the highest Median and Fuel X has the least.
- All Medians lie between 200 and 300

3.7

```
> newset <- subset(Fuel_Cons_2022$Comb_Fuel, Fuel_Cons_2022$Transmission=="A8")
> hist(newset)
```



3.8

```
> Fuel_Cons_2022$compare <-
((Fuel_Cons_2022$Emission_co2/Fuel_Cons_2022$Cylinders*100))
> Fuel_Cons_2022
```

Cleaning | Cleaning1 | Fuel_Cons_2022

Filter

Make	Model	Cylinders	Transmission	Fuel_type	City_Fuel	Hwy_Fuel	Comb_Fuel	Emission_co2	Rating_Co2	Smog_Rating	compare
Acura	ILX	4	AM8	Z	9.9	7.0	8.6	199	6	3	4975.000
Acura	MDX SH-AWD	6	AS9	Z	12.3	9.2	10.9	254	4	3	4233.333
Acura	MDX SH-AWD A-SPEC	6	AS9	Z	12.2	9.5	11.0	258	4	3	4300.000
Acura	MDX Hybrid AWD	6	AM7	Z	9.1	9.0	9.0	210	5	3	3500.000
Acura	RDX AWD	4	AS10	Z	11.0	8.6	9.9	232	5	6	5800.000
Acura	RDX AWD A-SPEC	4	AS10	Z	11.3	9.1	10.3	241	5	6	6025.000
Acura	RLX Hybrid	6	AM7	Z	8.4	8.2	8.4	196	6	7	3266.667
Acura	TLX A-SPEC	4	AM8	Z	10.2	7.4	8.9	209	5	3	5225.000
Acura	TLX SH-AWD	6	AS9	Z	11.4	7.7	9.8	228	5	3	3800.000
Acura	TLX SH-AWD A-SPEC/Limited Edition	6	AS9	Z	12.0	8.2	10.3	240	5	3	4000.000
Alfa Romeo	4C Spider	4	AM6	Z	9.7	6.9	8.4	197	6	1	4925.000
Alfa Romeo	Giulia	4	A8	Z	10.0	7.2	8.7	205	6	3	5125.000
Alfa Romeo	Giulia AWD	4	A8	Z	10.5	7.7	9.2	217	5	3	5425.000
Alfa Romeo	Giulia Quadrifoglio	6	A8	Z	13.5	9.3	11.6	271	4	3	4516.667
Alfa Romeo	Stelvio	4	A8	Z	10.3	8.1	9.3	218	5	3	5450.000
Alfa Romeo	Stelvio AWD	4	A8	Z	10.8	8.3	9.6	226	5	3	5650.000
Alfa Romeo	Stelvio AWD Quadrifoglio	6	A8	Z	13.9	10.3	12.3	288	3	3	4800.000
Aston Martin	DB11 V8	8	A8	Z	13.0	9.8	11.5	271	4	3	3387.500