Data Mining: Phase I

Rayane Adam, Roudy Bou Francis

2023-10-20

For this phase, the following sub-packages of tidyverse are required:

```r
library(tidyverse) #ggplot and dyplr are loaded with tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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 become errors
```

Table of contents:

This code will perform the following steps:

1. *Data preprocessing*: The code will

- remove rows with missing values
- convert numbers char variables to numerical factors
- rename the variables in a meanigful way
- get rid of the useless columns
- create new ones useful for our model

2. *Data visualization and exploration:* The code will

- create box plots to visualize the distribution of CO2 Emissions across different vehicle sizes, transmission, fuel types

- create scatter plots to visualize the relationship between CO2 Emissions and quantitative variables (with qualitative variables group labeling to spo patterns)
- create a correlation matrix to visualize the relationships between some of the variables in the dataset

3. Hypothesis exploration: The code will highlight the following hypotheses that were either logically assumed or intuitively withdrawn from the previous data exploration and visualization

4. Linear regression: The code will build linear regression model to test the hypotheses including

- simple linear regression
- multiple linear regression
- interaction effect
- polynomial regression

Data

The Dataset used for this project has been downloaded from Government of Canada's Open Data.
It's specifically published by the Natural Resources of Canada, Fuel consumption section - 2023. *To get access to the data, download the csv file via this link*

Accordingly this dataset revolves around **Fuel Consumption** data collected in 2023 for over `833 instances`. Each instance have the following attributes:

Col Number	Name	Description
1	Model	The car model's year
2	Make	Car type - company
3	Model	The model's name
4	Vehicle Class	Class of the vehicle
5	engine_size	in Litres
6	Transmission	Like automatic or manual…
7	Fuel Type	The type of fuel the car uses
8	Fuel Consumption - City (L/100 km)	Fuel consumption rate in liters per 100 kilometers when driven in city conditions.
9	Fuel Consumption - Hwy (L/100 km)	The fuel consumption rate in liters per 100 kilometers when driven on highways.
10	comb (L/100 km)	Combined Fuel Consumption mixed between city and highway (L/100 km)
11	Comb (MPG)	Combined Fuel Consumption in a different unit (MPG)
12	CO2 Emissions	The amount of CO2 emissions produced by the car
13	CO2	rating: 1-10 from worst to best
14	Smog	rating: 1-10 from worst to best

Again, to access the csv file just click here. Now will be loading it in a data frame, with a glimpse on the 1st 10 rows:

```
df <- read.csv("data/MY2023 Fuel Consumption Ratings.csv", header = TRUE)
#head(df, n = 3) #viewing the first 3 rows
```

::: {#data-preprocessing .section .level2} ## Data Preprocessing

Data cleaning: removing NA values

This is dedicated to removing empty rows and columns from the dataset.
steps to follow:

- removes all the columns containing missing values (plenty of empty columns)
- removes all the rows containing missing values (description of the dataset at the ed of the file Practically, there should be 14 columns i the dataset (as per the description above) but heres how many there is. Moreover the data ends at row 834, proof:

```
print(paste("number of columns:", ncol(df) ,"; number of rows:", nrow(df)))
```

```
## [1] "number of columns: 224 ; number of rows: 7357"
```

Checking is there exists any missing values in the dataset, if yes returns TRUE.
(it does return True in many columns, some columns are purely empty, and we need to fix this in our data frame)

```
#head(is.na(df),n=2) # checks if our dataset contains missing values, hid cz output is too large
#summary(df) # checks how many NA we do have
#the output is hidden becuase it's huge
```

Removing empty columns

```
df <- subset(df, select = !apply(is.na(df), 2, any))
print(paste("number of columns after:", ncol(df))) #removes all the columns containing missing values
```

```
## [1] "number of columns after: 14"
```

Removing the useless rows:

- First 2 rows in data combine to mean the header - remove one of them `(check how in the code below)`
- Last row is 833 - remove what follows `(check how in the code below)`

(we have a proof from before that the # of rows is far more greater than 833)

```
head(df, n=2)
```

```
##   Model  Make Model.1 Vehicle.Class Engine.Size Transmission Fuel
## 1  Year                                     (L)              Type
## 2  2023 Acura Integra     Full-size         1.5          AV7    Z
##   Fuel.Consumption             X             X.1        X.2 CO2.Emissions
## 1  City (L/100 km) Hwy (L/100 km) Comb (L/100 km) Comb (mpg)        (g/km)
## 2              7.9           6.3             7.2         39           167
##      CO2   Smog
## 1 Rating Rating
## 2      6      7
```

```r
print(df[832:836,])
```

```
##                        Model  Make     Model.1 Vehicle.Class Engine.Size
## 832                     2023 Volvo XC60 B6 AWD    SUV: Small         2.0
## 833                     2023 Volvo XC90 B5 AWD SUV: Standard         2.0
## 834                     2023 Volvo XC90 B6 AWD SUV: Standard         2.0
## 835
## 836 Understanding the table
##     Transmission Fuel Fuel.Consumption   X  X.1 X.2 CO2.Emissions CO2 Smog
## 832          AS8    Z             11.1 8.7 10.0  28           233   5    7
## 833          AS8    Z             10.5 8.4  9.6  29           223   5    5
## 834          AS8    Z             11.9 9.1 10.6  27           249   5    7
## 835
## 836
```

checking these from data set:

```r
df <- df[2:834,]
print(paste("number of rows after:", nrow(df)))
```

```
## [1] "number of rows after: 833"
```

Checking the presence of duplicate rows that have the same information:

```r
# Check for duplicate rows
has_duplicates <- any(duplicated(df) | duplicated(df, fromLast = TRUE))
if (has_duplicates) {
  print("The dataset contains at least two rows with the same information.")
} else {
  print("The dataset does not contain two rows with the same information. The dataset has no duplicate r
}
```

```
## [1] "The dataset does not contain two rows with the same information. The dataset has no duplicate ro
```

Data cleaning: columns naming conventions

Notice how the header of the data has unconventional names:

```r
colnames(df)
```

```
##  [1] "Model"            "Make"             "Model.1"          "Vehicle.Class"
##  [5] "Engine.Size"      "Transmission"     "Fuel"             "Fuel.Consumption"
##  [9] "X"                "X.1"              "X.2"              "CO2.Emissions"
## [13] "CO2"              "Smog"
```

This is bad :/ We need to fix this:

```r
new_names <- c("model_year", "car_make", "model_name", "vehicle_class", "engine_size", "transmission",

colnames(df) <- new_names
colnames(df)
```

```
##  [1] "model_year"         "car_make"           "model_name"
##  [4] "vehicle_class"      "engine_size"        "transmission"
##  [7] "fuel_type"          "city_consumption"   "hwy_consumption"
## [10] "mix_consumption"    "mix_consumption_2"  "CO2_emission"
## [13] "CO2_rate"           "smog_rate"
```

Data cleaning: data type conversion

Make sure the numbers are converted to the right data type:

- numerical values/ measures –> double
- rank values –> factor

```
glimpse(df) #checks what type of data each feature is
```

```
## Rows: 833
## Columns: 14
## $ model_year        <chr> "2023", "2023", "2023", "2023", "2023", "2023", "202~
## $ car_make          <chr> "Acura", "Acura", "Acura", "Acura", "Acura", "Acura"~
## $ model_name        <chr> "Integra", "Integra A-SPEC", "Integra A-SPEC", "MDX ~
## $ vehicle_class     <chr> "Full-size", "Full-size", "Full-size", "SUV: Small",~
## $ engine_size       <chr> "1.5", "1.5", "1.5", "3.5", "3.0", "2.0", "2.0", "2.~
## $ transmission      <chr> "AV7", "AV7", "M6", "AS10", "AS10", "AS10", "AS10", ~
## $ fuel_type         <chr> "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z~
## $ city_consumption  <chr> "7.9", "8.1", "8.9", "12.6", "13.8", "11.0", "11.3",~
## $ hwy_consumption   <chr> "6.3", "6.5", "6.5", "9.4", "11.2", "8.6", "9.1", "8~
## $ mix_consumption   <chr> "7.2", "7.4", "7.8", "11.2", "12.4", "9.9", "10.3", ~
## $ mix_consumption_2 <chr> "39", "38", "36", "25", "23", "29", "27", "29", "29"~
## $ CO2_emission      <chr> "167", "172", "181", "263", "291", "232", "242", "23~
## $ CO2_rate          <chr> "6", "6", "6", "4", "4", "5", "5", "5", "5", "5", "5~
## $ smog_rate         <chr> "7", "7", "6", "5", "5", "6", "6", "7", "7", "5", "5~
```

```
df <- transform(df,
                      engine_size=as.numeric(engine_size),
                      city_consumption=as.numeric(city_consumption),
                      hwy_consumption=as.numeric(hwy_consumption),
                      mix_consumption=as.numeric(mix_consumption),
                      mix_consumption_2=as.numeric(mix_consumption_2),
                      CO2_emission=as.numeric(CO2_emission),
                      CO2_rate=as.factor(CO2_rate),
                      smog_rate=as.factor(smog_rate)
                    )
glimpse(df) #checks the data type after conversion
```

```
## Rows: 833
## Columns: 14
## $ model_year        <chr> "2023", "2023", "2023", "2023", "2023", "2023", "202~
## $ car_make          <chr> "Acura", "Acura", "Acura", "Acura", "Acura", "Acura"~
## $ model_name        <chr> "Integra", "Integra A-SPEC", "Integra A-SPEC", "MDX ~
## $ vehicle_class     <chr> "Full-size", "Full-size", "Full-size", "SUV: Small",~
## $ engine_size       <dbl> 1.5, 1.5, 1.5, 3.5, 3.0, 2.0, 2.0, 2.0, 2.0, 3.0, 2.~
## $ transmission      <chr> "AV7", "AV7", "M6", "AS10", "AS10", "AS10", "AS10", ~
```

```
## $ fuel_type          <chr> "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z~
## $ city_consumption   <dbl> 7.9, 8.1, 8.9, 12.6, 13.8, 11.0, 11.3, 11.2, 11.3, 1~
## $ hwy_consumption    <dbl> 6.3, 6.5, 6.5, 9.4, 11.2, 8.6, 9.1, 8.0, 8.1, 9.4, 7~
## $ mix_consumption    <dbl> 7.2, 7.4, 7.8, 11.2, 12.4, 9.9, 10.3, 9.8, 9.8, 11.0~
## $ mix_consumption_2 <dbl> 39, 38, 36, 25, 23, 29, 27, 29, 29, 26, 32, 31, 24, ~
## $ CO2_emission       <dbl> 167, 172, 181, 263, 291, 232, 242, 230, 231, 256, 20~
## $ CO2_rate           <fct> 6, 6, 6, 4, 4, 5, 5, 5, 5, 5, 5, 5, 4, 5, 5, 4, 5, 4~
## $ smog_rate          <fct> 7, 7, 6, 5, 5, 6, 6, 7, 7, 5, 5, 5, 3, 5, 5, 3, 7, 5~
```

::: {#data-cleaning-columns-editing .section .level3} ### Data cleaning: Columns editing

Dropping 2 and Mutating 2 :)

1. The `model year` is 2023 for all instances, thus we will drop it.

```
unique(df$model_year)
```

```
## [1] "2023"
```

```
df <- subset(df, select = - model_year)
```

p.s. `mix_consumption` is measured i L/100 km, `mix_consumption_2` i MPG; converting between them is via the following formula: we will only use one of them as predictor - drop the second one.

$$mpg = \frac{235.215}{L/100km}$$

```
df <- subset(df, select = - mix_consumption_2)
```

2. `Transmission` and `Vehicle_class` have too may values that can easily get grouped in new cols:

- Transmission is either Automatic or Manual (if starts with A automatic, M manual)
- Vehicle class gives rise to vehicle_size_category based on its size: small, medium, large or special (*this includes special purpose cars and 2 seaters*)

```
unique(df$vehicle_class)
```

```
##  [1] "Full-size"                "SUV: Small"
##  [3] "SUV: Standard"            "Compact"
##  [5] "Mid-size"                 "Minicompact"
##  [7] "Two-seater"               "Subcompact"
##  [9] "Station wagon: Small"     "Station wagon: Mid-size"
## [11] "Pickup truck: Small"      "Pickup truck: Standard"
## [13] "Minivan"                  "Special purpose vehicle"
```

```
unique(df$transmission)
```

```
##  [1] "AV7"  "M6"   "AS10" "A8"   "A9"   "AM7"  "AS8"  "AM8"  "AV"   "AS9"
## [11] "A10"  "A6"   "M7"   "AV1"  "AM6"  "AS7"  "AV8"  "AV6"  "AS6"  "AV10"
## [21] "M5"   "AS5"  "A7"
```

```r
df <- mutate(df, vehicle_size_category = case_when(
  vehicle_class == "Full-size" ~ "Large",
  vehicle_class == "SUV: Standard" ~ "Medium",
  vehicle_class == "Mid-size" ~ "Medium",
  vehicle_class == "Minicompact"~ "Small",
  vehicle_class == "SUV: Small"~"Small",
  vehicle_class == "Compact"~"Small",
  vehicle_class == "Two-seater"~"Special",
  vehicle_class == "Subcompact"~"Small",
  vehicle_class == "Station wagon: Small"~"Small",
  vehicle_class == "Station wagon: Mid-size"~"Medium",
  vehicle_class == "Pickup truck: Small"~"Small",
  vehicle_class == "Pickup truck: Standard"~"Medium",
  vehicle_class == "Special purpose vehicle"~"Special",
  vehicle_class == "Minivan"~"Small"
))


df <- mutate(df, transmission_type_category = case_when(
  grepl("^A", df$transmission) ~ "Automatic",
  grepl("^M", df$transmission) ~ "Manual",
))

colnames(df)
```

```
##  [1] "car_make"                  "model_name"
##  [3] "vehicle_class"             "engine_size"
##  [5] "transmission"              "fuel_type"
##  [7] "city_consumption"          "hwy_consumption"
##  [9] "mix_consumption"           "CO2_emission"
## [11] "CO2_rate"                  "smog_rate"
## [13] "vehicle_size_category"     "transmission_type_category"
```

3. Lastly, `CO2 rate` and `Smog rate` are concluded based on the CO2 emitted, thus for this regression task of response `yCO2_emissions`, we must drop these values from the tale useless

```r
df <- subset(df, select = - CO2_rate)
df <- subset(df, select = - smog_rate)
```

Final data frame has the following:

```r
colnames(df)
```

```
##  [1] "car_make"                  "model_name"
##  [3] "vehicle_class"             "engine_size"
##  [5] "transmission"              "fuel_type"
##  [7] "city_consumption"          "hwy_consumption"
##  [9] "mix_consumption"           "CO2_emission"
## [11] "vehicle_size_category"     "transmission_type_category"
```

Data exploration: outliers

Extreme outlier cars are the ones that pollute the most (or very low emissions - which is not the case here as you ca see now). They are often older, bigger, and more powerful than most cars. In this project we're interested in studying what makes a car have that irregular level of CO2 emitted. This is why, we need to detect them to later on reflect on their properties.

Two extreme outliers are showing

```
outliers <- vector()
q1 <- quantile(df$CO2_emission, 0.25)
q3 <- quantile(df$CO2_emission, 0.75)

for (i in 1:nrow(df)) {
  lower_bound <- q1 - 3 * IQR(df$CO2_emission) #extreme outliers because why not :p
  upper_bound <- q3 + 3 * IQR(df$CO2_emission)

  if (df$CO2_emission[i] < lower_bound | df$CO2_emission[i] > upper_bound) {
    outliers <- append(outliers, i) #app if its lower or upper than the limits
    print(df[i,])
  }
}
```

```
##     car_make       model_name vehicle_class engine_size transmission fuel_type
## 131  Bugatti Chiron Pur Sport    Two-seater           8          AM7         Z
##     city_consumption hwy_consumption mix_consumption CO2_emission
## 131             30.3            20.9            26.1          608
##     vehicle_size_category transmission_type_category
## 131               Special                  Automatic
##     car_make         model_name vehicle_class engine_size transmission
## 132  Bugatti Chiron Super Sport    Two-seater           8          AM7
##     fuel_type city_consumption hwy_consumption mix_consumption CO2_emission
## 132         Z             30.3            20.9            26.1          608
##     vehicle_size_category transmission_type_category
## 132               Special                  Automatic
```

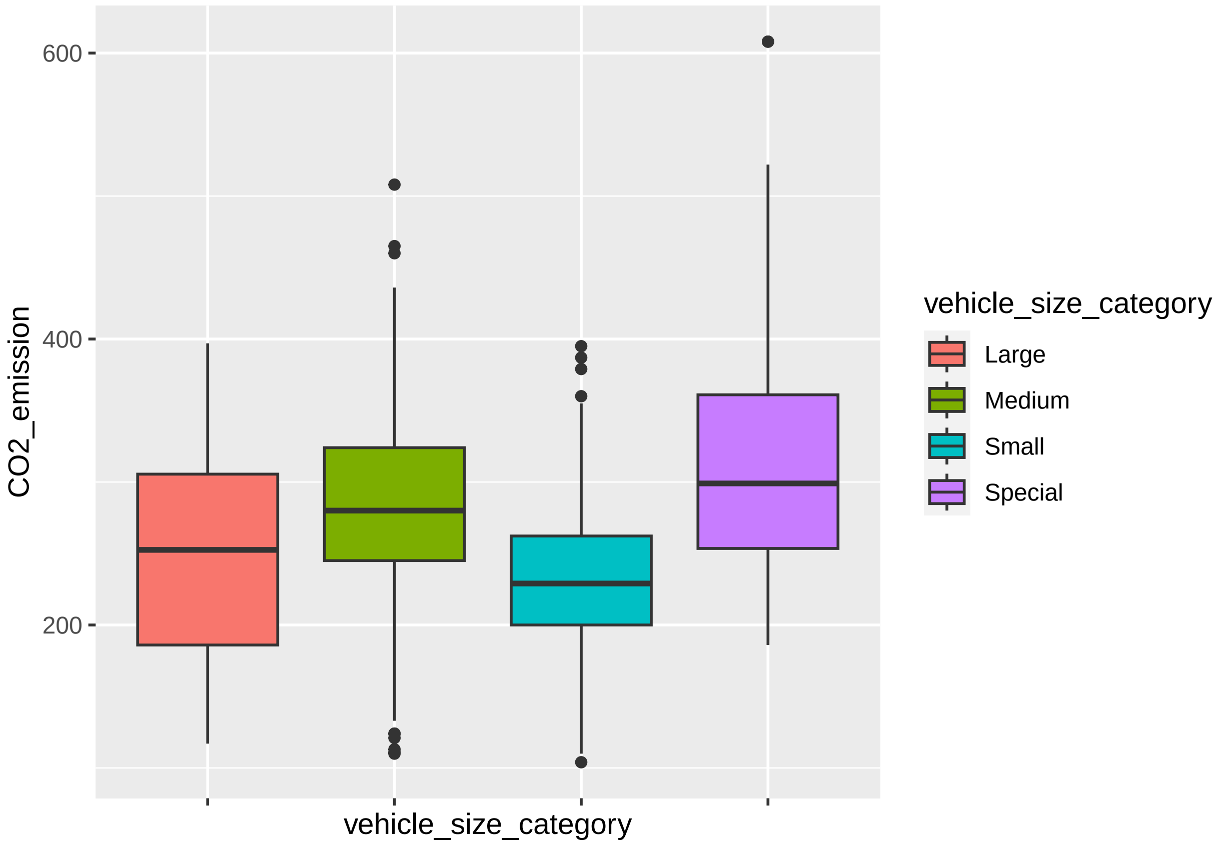
Notice how the two outliers found have a CO2 emission of >600 g/km. They are both Bugattis, large cars with large engines, and use premium fuel... These make sense to any car expert: they are indeed pollutants. However, we're data scientists! We find the pattern with numbers! At the end of this project, the models will actually explain how we can get to such conclusions. Every feature that contributed to this high emission will be broken down into pieces and analyzed.

Data Visualization:

Before coming up with hypotheses to test on the data, we need to look for patterns and relationships. This can be done via plotting (mainly scatterplots for nums, boxplots for categorical and correlation matrices)

Vehicle class & CO2

```
ggplot(data=df) +
  geom_boxplot(mapping = aes(x=vehicle_size_category, y=CO2_emission, fill=vehicle_size_category))+
  theme(axis.text.x = element_blank())
```

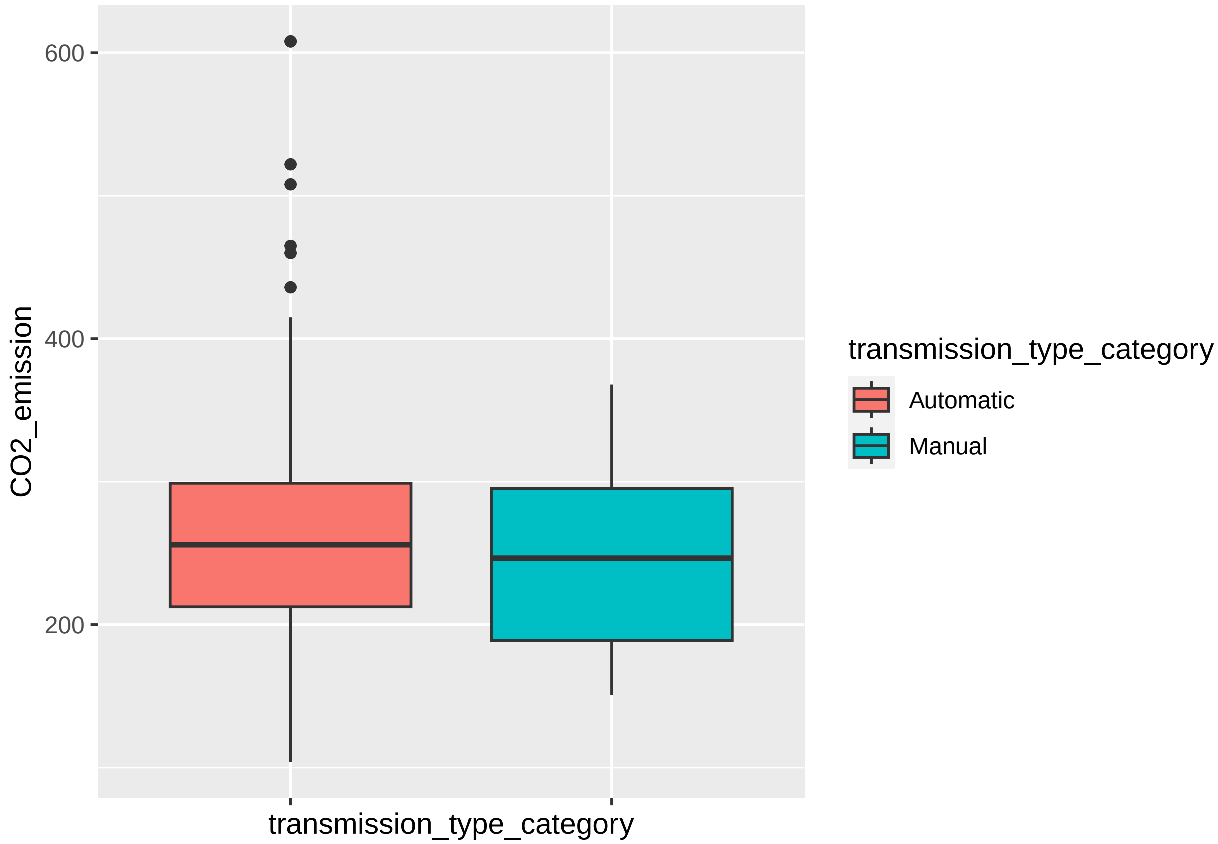


Notice how some vehicle types like `2 seater` *which is a special car have exceptionally a high production of CO2 whereas a* `small minivan` *which is small relatively has ecological levels of emissions. A pattern is present where it is shown that cars that are considered as special have relatively higher CO2 Emissions than other types of cars.*

P.S. Notice that a special car has a suspicious level of CO2 emissions greater than 600 ~maybe outlier or ~maybe not. (Later On)

Transmission and CO2 levels:

```
ggplot(data=df) +
  geom_boxplot(mapping = aes(x=transmission_type_category, y=CO2_emission, fill=transmission_type_categ
  theme(axis.text.x = element_blank())
```

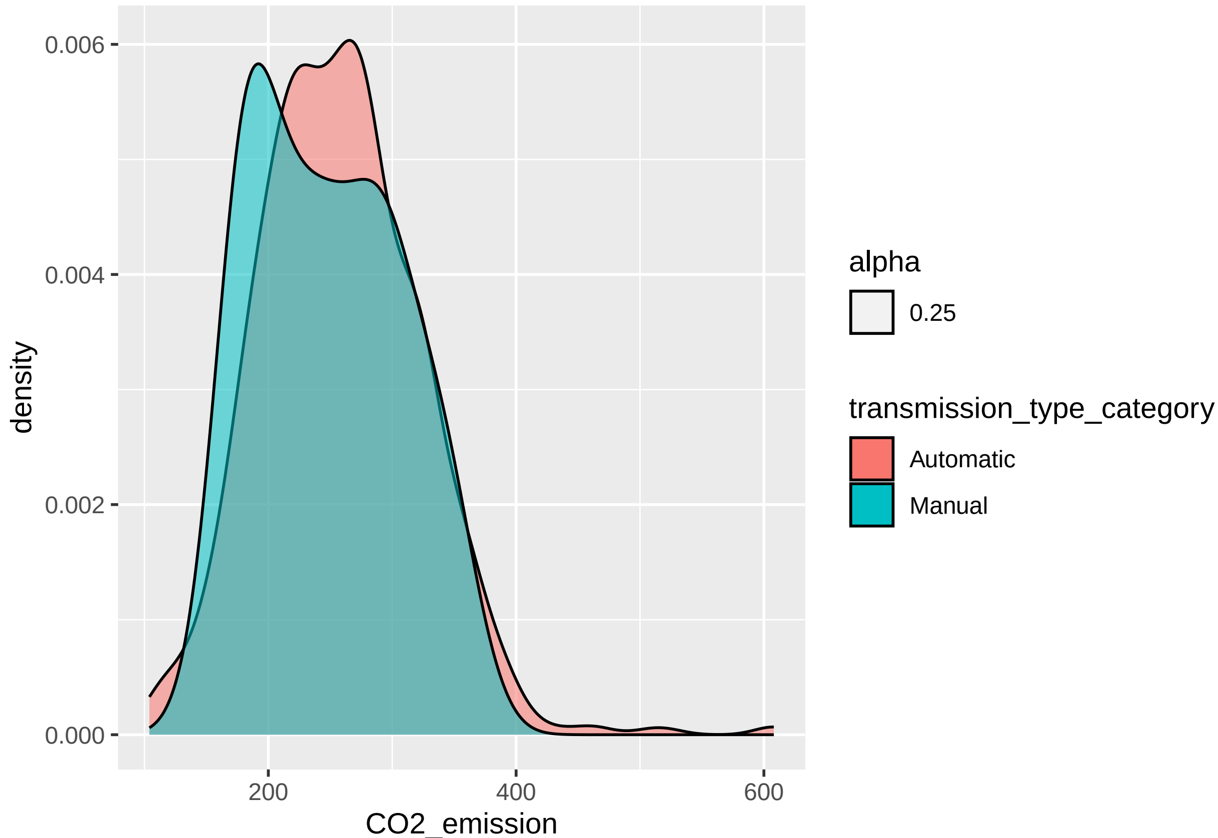


There is no general trend whatsoever between the transmission type and the CO2 emitted by the car.

P.S. Notice that an automatic car has a suspicious level of CO2 emissions greater than 600 ~maybe outlier or ~maybe not. (Later On)

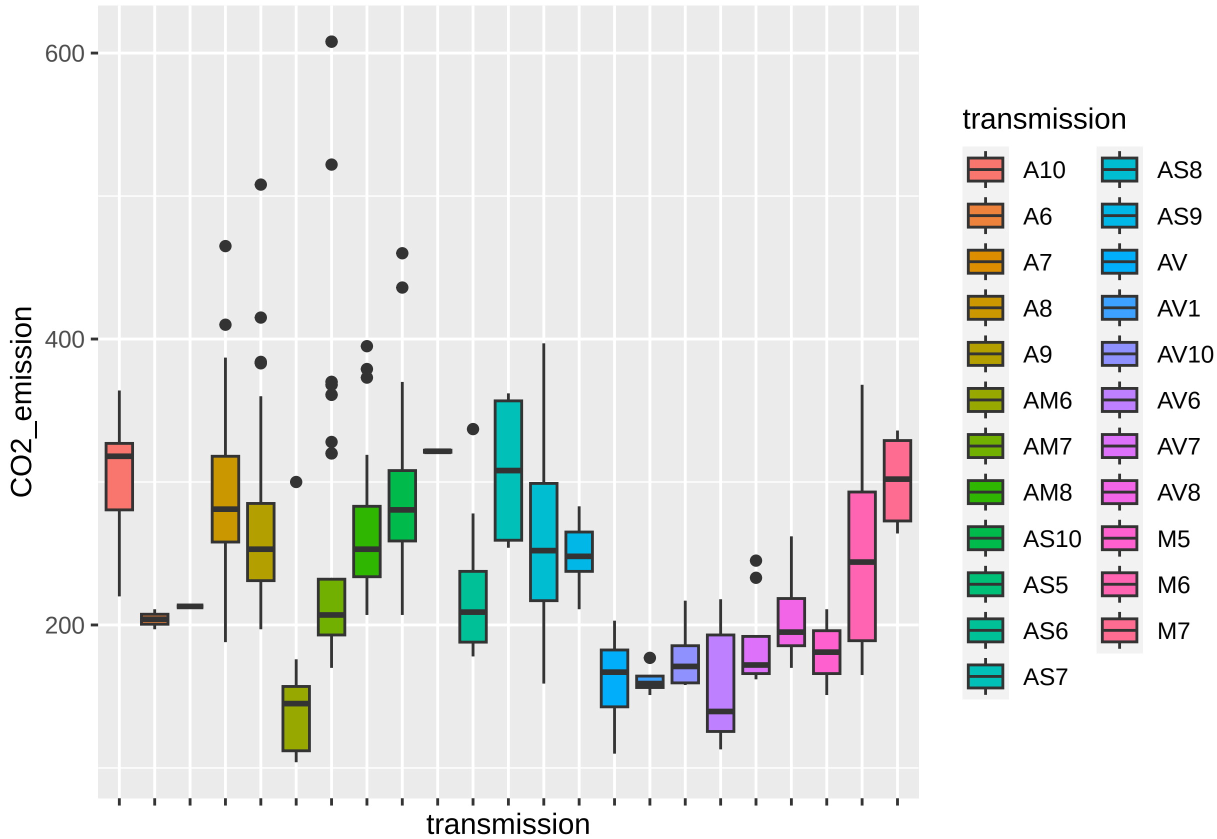
Density Curve of CO2 Emission Based on the General Transmission Type

```
ggplot(data =df) + geom_density(mapping = aes(x = CO2_emission, fill =transmission_type_category, alpha
```



Additional data exploration gives an extra evidence that the CO2 emissions are not affected by the general transmission type of the car. We should look at the individual transmission types

```
ggplot(data=df) +
  geom_boxplot(mapping = aes(x=transmission, y=CO2_emission, fill=transmission))+
  theme(axis.text.x = element_blank())
```



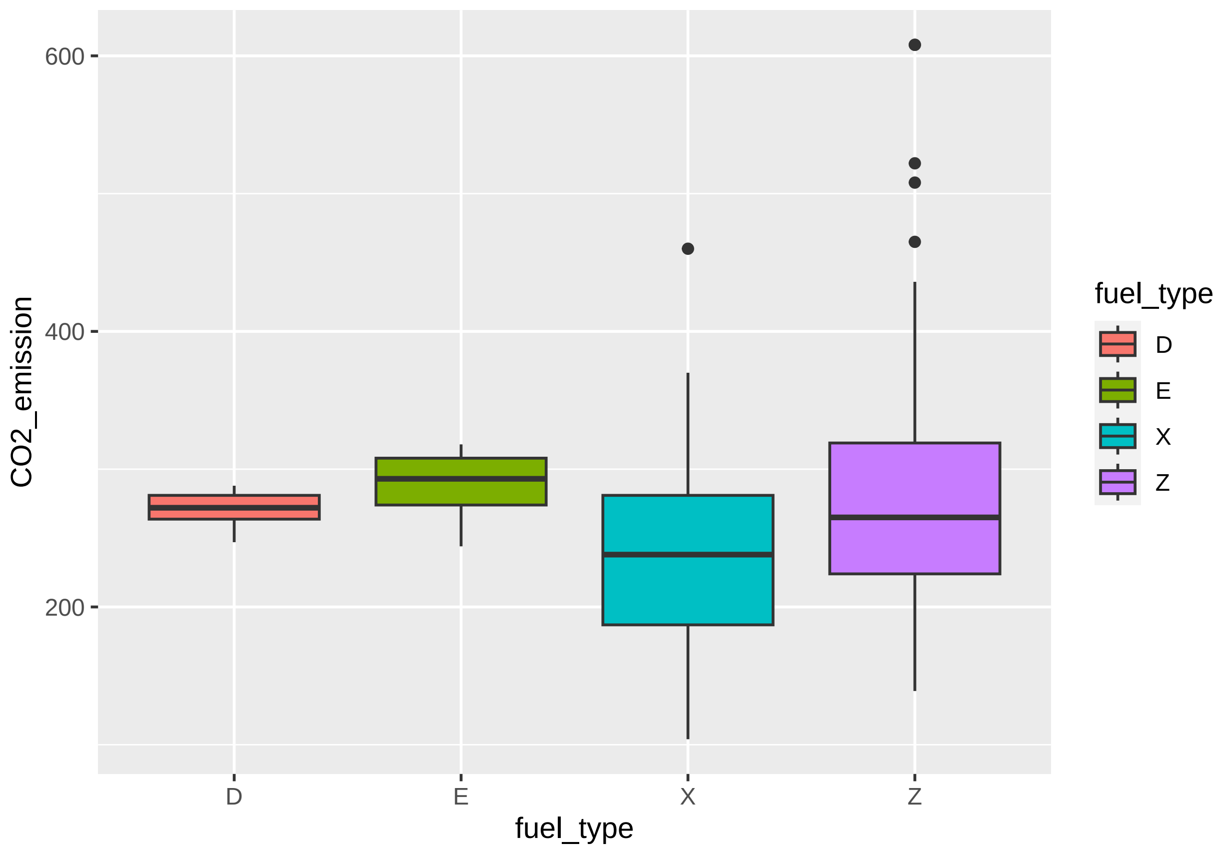
Individual transmission do have a clearer relationship with the amount of CO2 emitted.

i.e. a specific transmission type can affect the repsonse

P.S. Notice that a AM7 car has a suspicious level of CO2 emissions greater than 600 ~maybe outlier or ~maybe not. (Later On)

Fuel type & CO2

```
ggplot(data=df) +
  geom_boxplot(mapping = aes(x=fuel_type, y=CO2_emission, fill=fuel_type))
```



Some fuel types are more notorious than others, we will see how we can incorporate this result later on

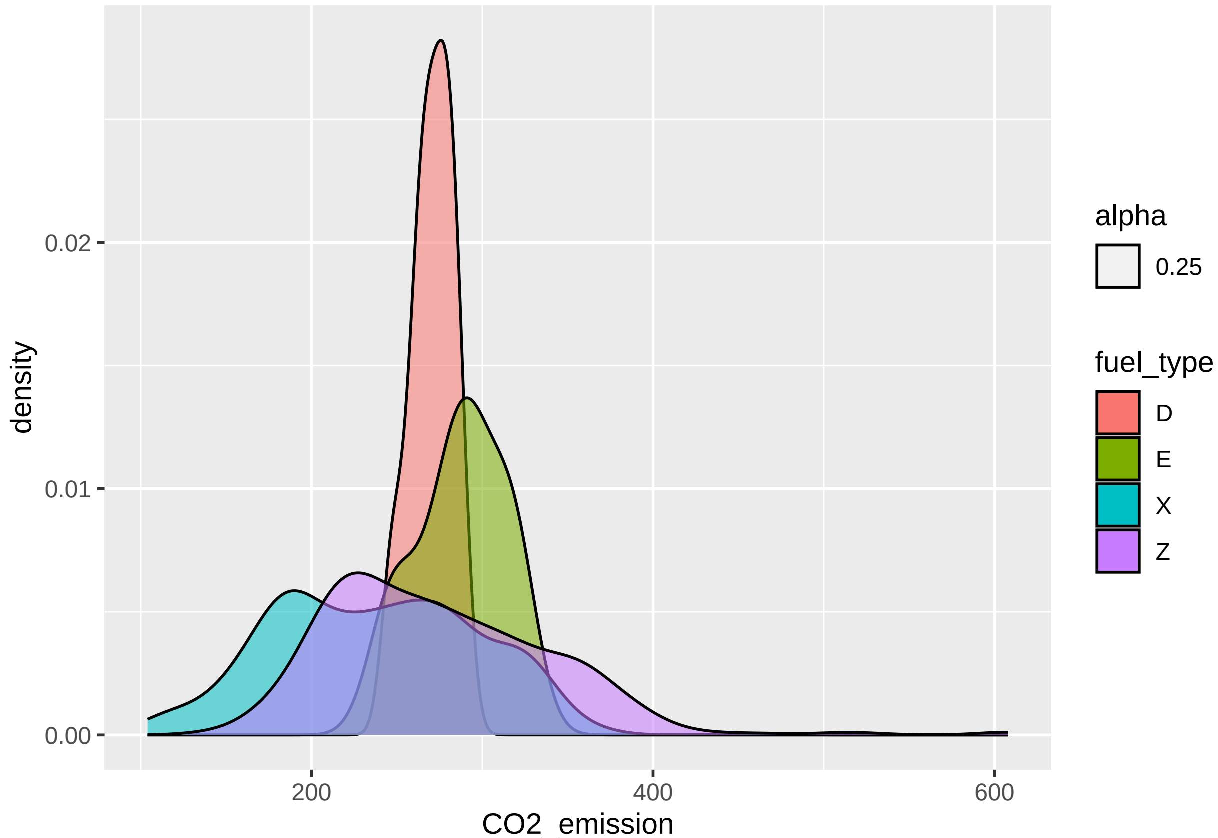
p.s. (source: here)

- X - regular gasoline
- Z - premium gasoline
- D - diesel
- E - ethanol (E85)

P.S. Notice that a car that uses premium gasoline has a suspicious level of CO2 emissions greater than 600 ~maybe outlier or ~maybe not. (Later On)

Density Curve of CO2 Emission Based on the Fuel Type

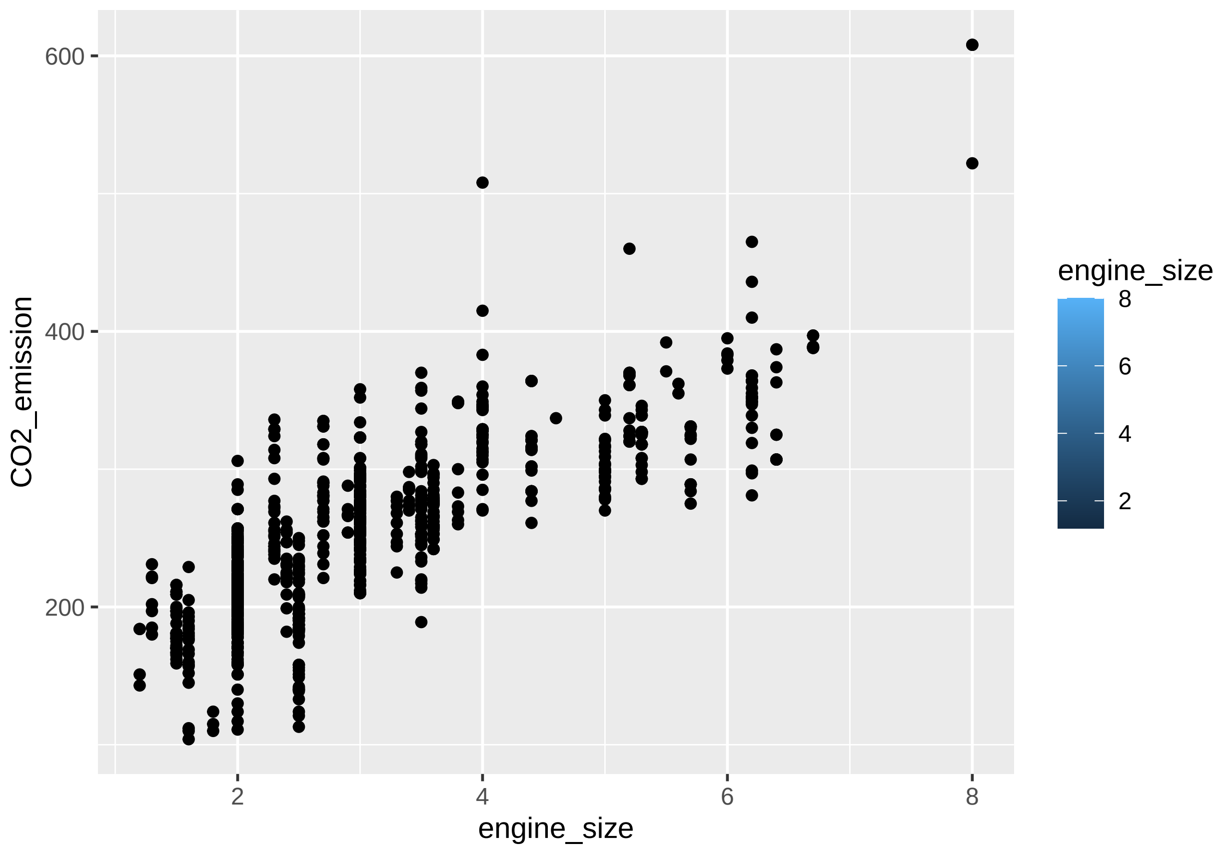
```
ggplot(data =df) + geom_density(mapping = aes(x = CO2_emission, fill =fuel_type, alpha = 0.25))
```



Check the interpretation here

Engine Size and CO2 Emissions

```
ggplot(data=df) +
  geom_point(mapping= aes(x=engine_size, y=CO2_emission, fill=engine_size))
```



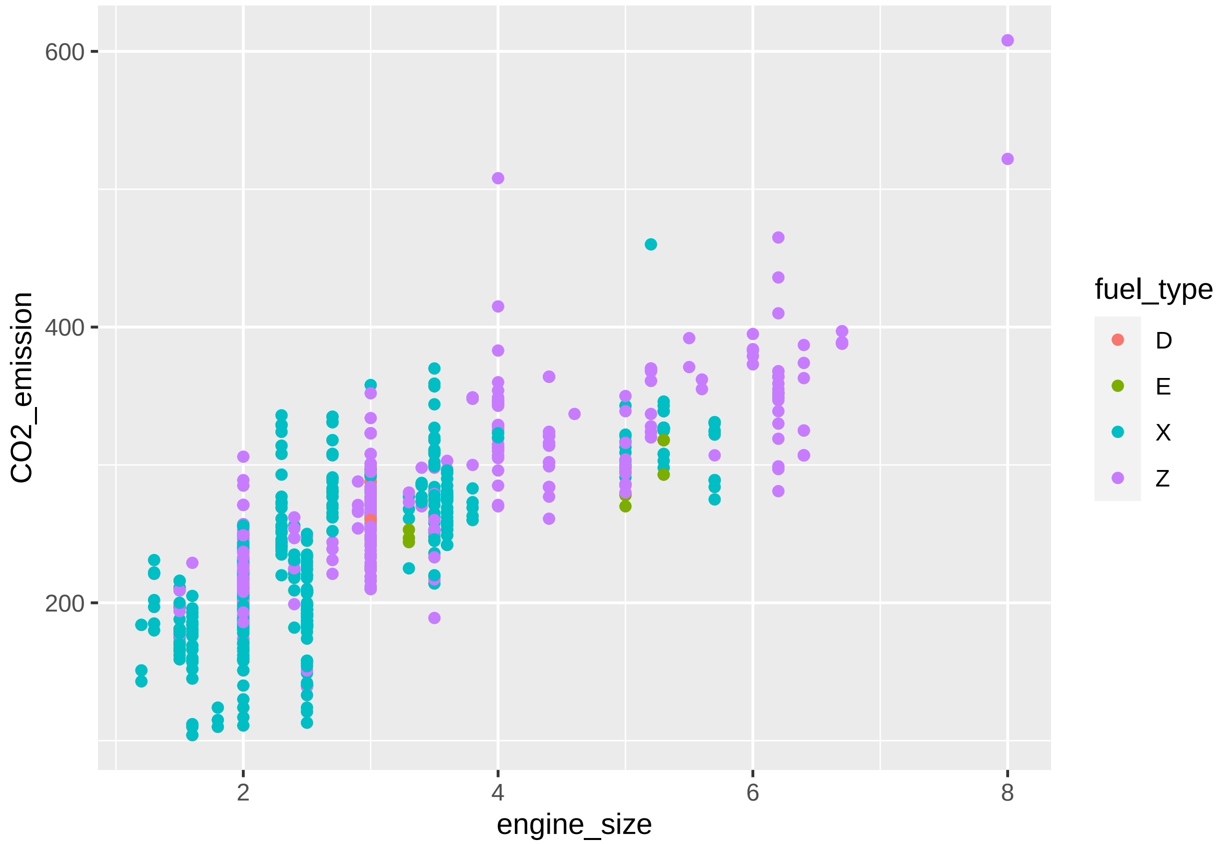
A pattern can be perceived:

- Generally, a larger engine size will result in a higher emission of CO2.

This pattern is perceived if the visualizations were just made on these 2 variables. Fortunately, we can visualize the pattern using additional variables.

Engine Size and Fuel Types

```
ggplot(data=df) +
  geom_point(mapping= aes(x=engine_size, y=CO2_emission, color=fuel_type))
```



A pattern can be perceived (a general one to be tested):

- The larger the engine the more likely it is using premium gasoline

  Z. – purple - & the higher the CO2 emitted

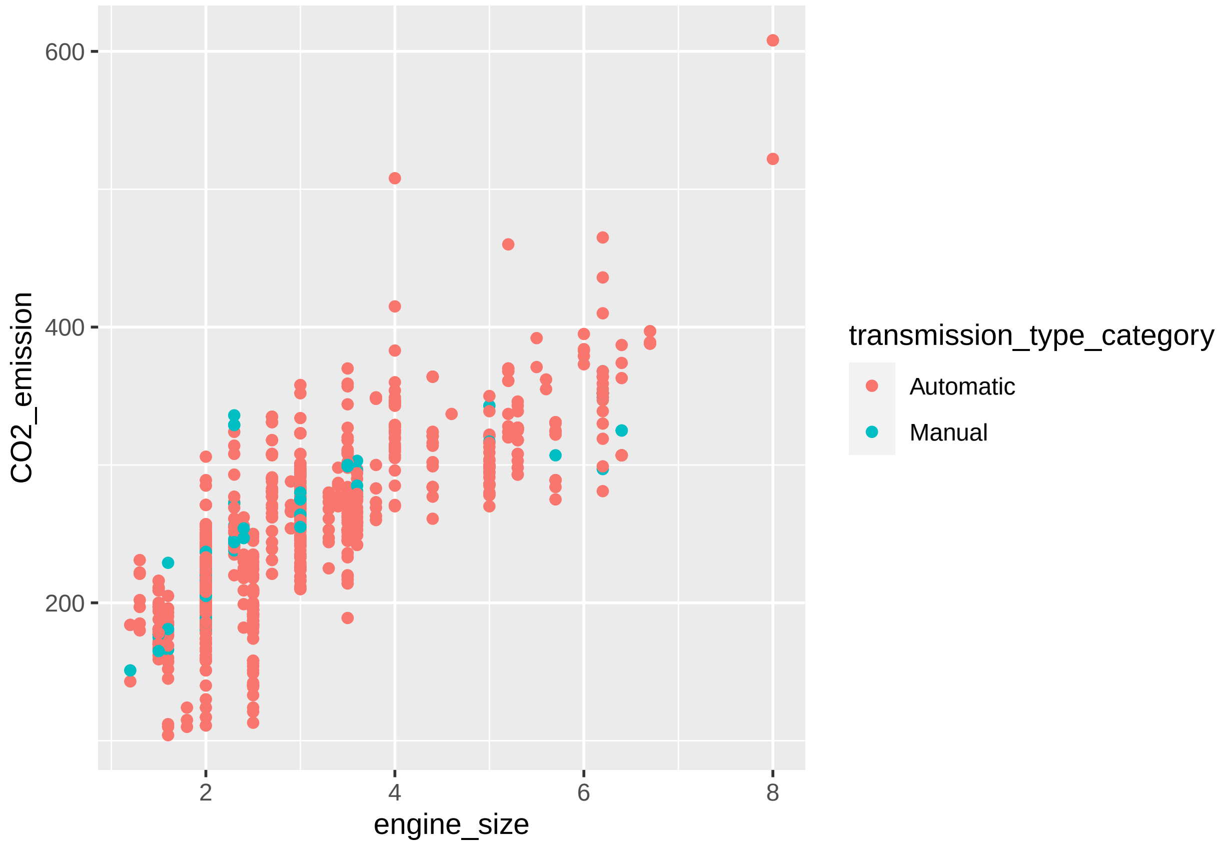
- The smaller the engine the more likely it is using regular gasoline

  X. – blue - & the lower the CO2 emitted

P.S. Notice that a car with engine size = 8 (largest) and that uses premiuim gasoline has a suspicious level of CO2 emissions greater than 600 ~maybe outlier or ~maybe not. (Later On)

Engine size and Transmission types

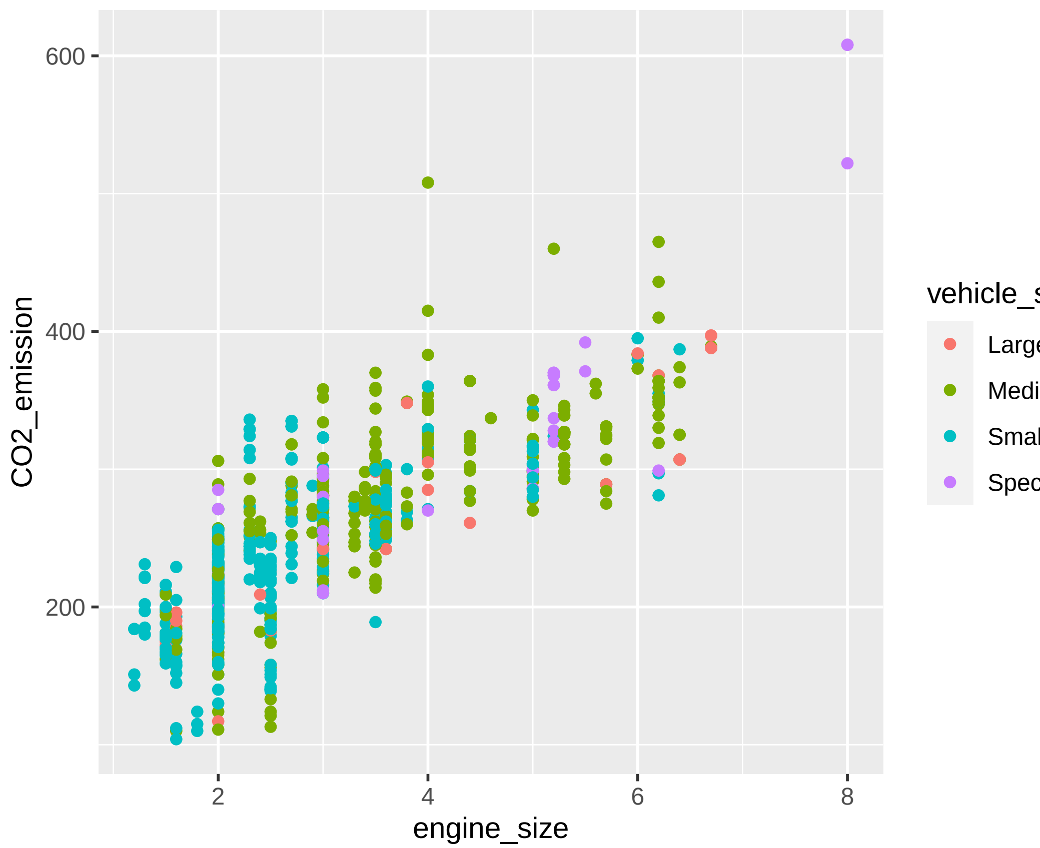
```
ggplot(data=df) +
  geom_point(mapping= aes(x=engine_size, y=CO2_emission, color=transmission_type_category))
```



There is no apparent relationship between engine_size and transmission type specifically when it comes to CO2 emissions; It only seems that automatic is more widely used in comparison to manual. However, mainly looking at engine_size, the larger the engine_size, the larger the emission of CO2

By now, you can see the suspicious value wihtout our p.s.

```
ggplot(data=df) +
  geom_point(mapping= aes(x=engine_size, y=CO2_emission, color=vehicle_size_category))
```



Engine Size and Vehicle Size

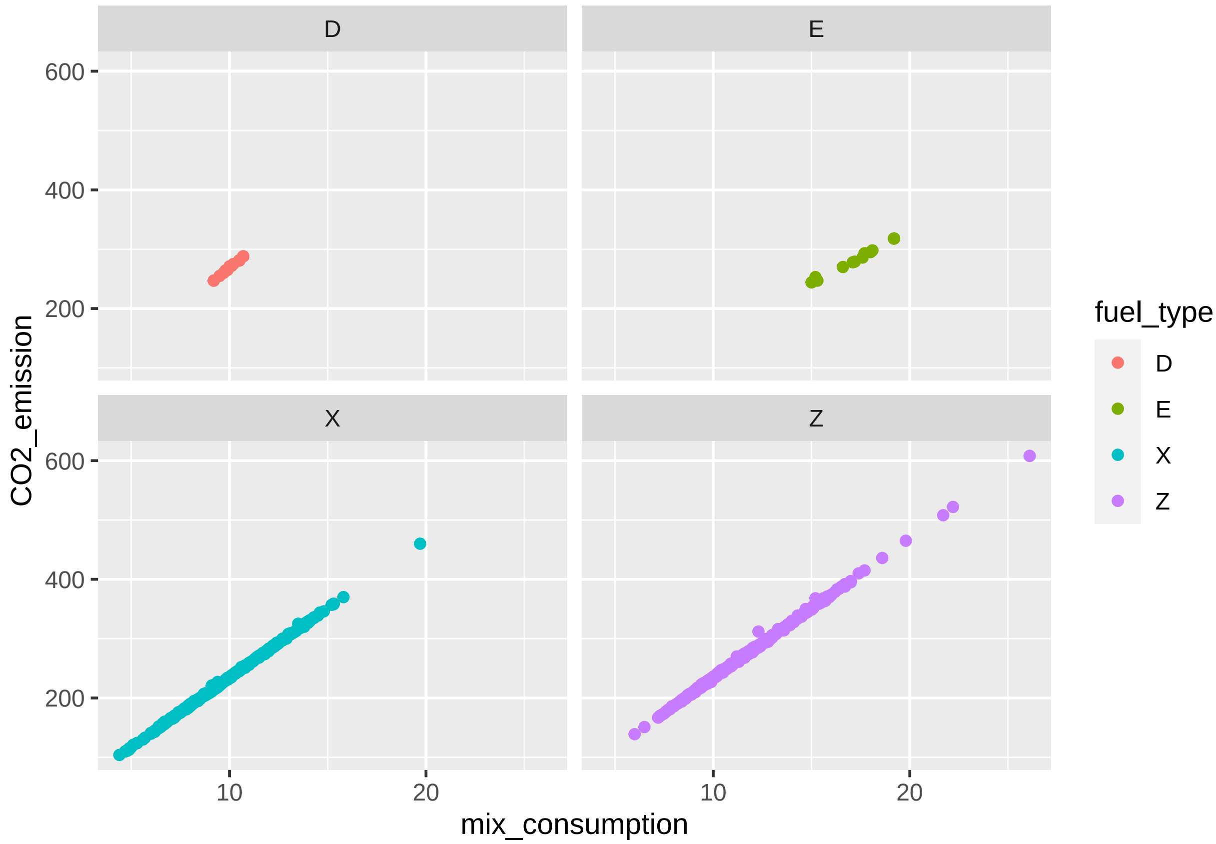
Generally, smaller cars have smaller engines, resulting in a lower emission of CO2. The pattern is interesting to study.

"Well, isn't it just remarkable how you've uncovered the 'suspicious' value all on your own, right?

Fuel consumption, fuel type and CO2

This general pattern persists when looking at the fuel consumption in a specific road (city, highway or both) as seen in the cor matrix. Here we're vizualizing consumption in a mixed road - having it the most tightly correlated with CO2 levels. (see matrix next)

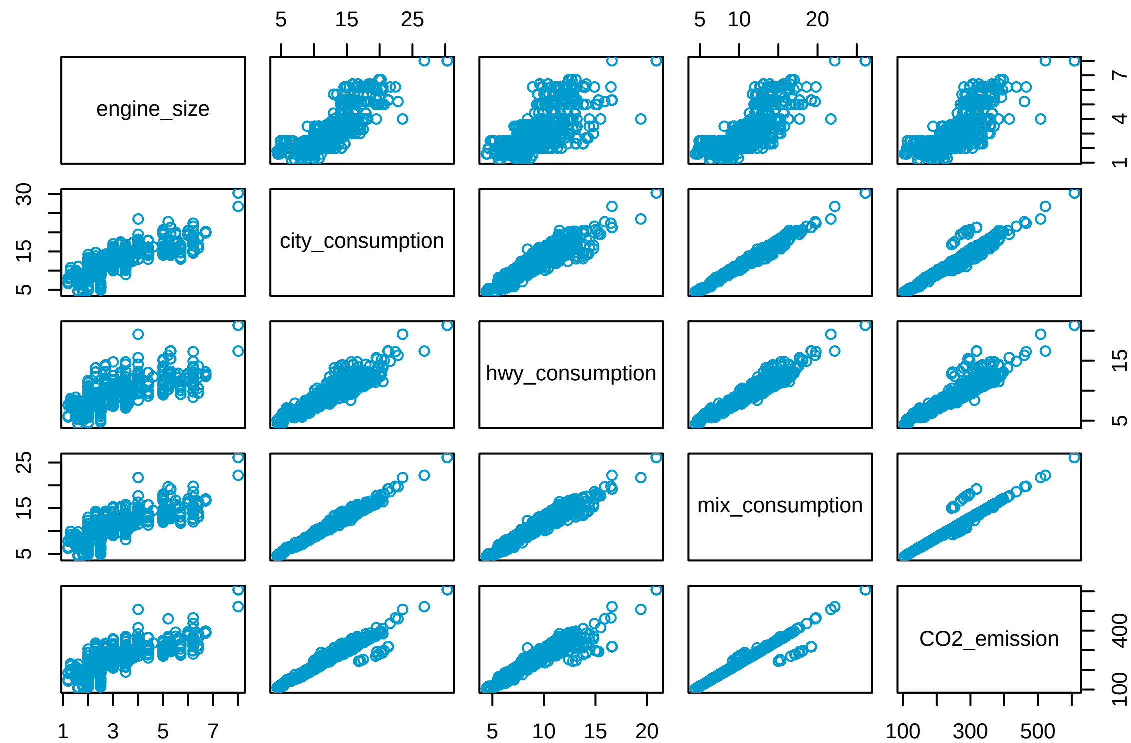
```
ggplot(data=df) +
  geom_point(mapping = aes(x=mix_consumption, y=CO2_emission, color=fuel_type)) +
  facet_wrap(~ fuel_type, nrow=2)
```



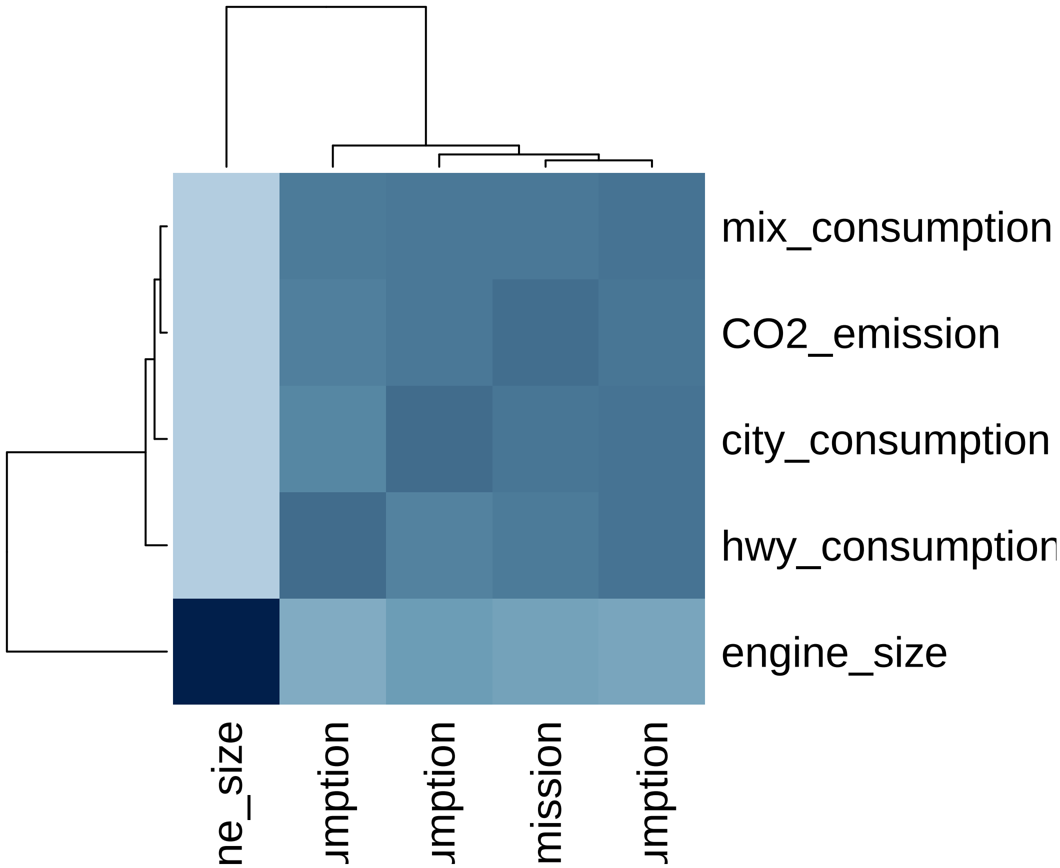
In general type D and X and E show a lower consumption (Z reaches very high values) when driving in combination of highway and city, resulting in lower emissions As per the density curve shown here if we compare the distribution of the D fuel type with the density curve, we can notice that the diesel has relatively a predictable and low CO2 emissions. In other words, we can notice on the density curve, that the red curve represents that the diesel fuel type has relatively a small standard deviation (and this is shown in the correlation matrix). In addition, we can notice that the premium fuel type is the one that has the largest emissions of CO2 (shown in the density curve and in the correlation matrix)

Correlation matrix

```
pairs(~engine_size+city_consumption+hwy_consumption+mix_consumption+CO2_emission, df, col = "#009ACD")
```




```
corr_matrix <- cor(df[, c("engine_size","city_consumption", "hwy_consumption", "mix_consumption", "CO2_
pal <- colorRampPalette(c("#b3cde0", "#6497b1" ,"#011f4b"))(100)
heatmap(cor(corr_matrix), col=pal)
```



Notice how the patterns are very strongly correlated when it comes between fuel consumption in city, highway, mix and the CO2 emission. There exist a mild correlation between engine_size and the rest - both the `correlation between all types of fuel consumption` *and the* `not-very-linear relationship between engine_size and CO2 emission` *shall be studied later*

Hypotheses:

From what we have seen in the visualizations above, boxplots, scatter plots and correlation matrices we can ask the following questions:

Simple Linear Regression:

- **Is there a significant relationship between engine_size and CO2 emission?**
  As per the relationship perceived here, the larger the engine, the higher the emission.

H_0: There is no linear relationship between engine size and CO2 emission of the car

- **Is there a significant relationship between the vehicle type and CO2 emission?** An interesting question that we would like to test is whether each vehicle class has a specific CO2 level emitted to check which are the most ecological and which are the most deleterious on the env.

H_0: There is no linear relationship between vehicle class and CO2 emission of the car

Multiple Linear Regression:

- **Is there a useful linear relationship between CO2 emissions and any of the predictors?** Application of backward elimination

H_0: There is no linear relationship between all predictors and CO2 emission of the car

+ Interaction effect:

- **Is there an interaction between engine size and fuel type? or does the effect of engine size depend on fuel type** Notice there is a pattern here when looking at the different CO2 emissions for different instances of engine_sizes with their corresponding fuel type
- **Is there an interaction between fuel type and fuel consumption?** Notice how, when filtered by fuel types here, Some fuel types show substantially lower consumption than others, which might in turn affect emissions since they are highly correlated.
- **Is there an interaction between fuel type and fuel consumption, & engine size and fuel type?** Checking if the inter actors behave simultaneously and affect the response.

Polynomial Regression:

- **Does a linear model of engine size and a larger polynomial model (of degree 3) fit the data equally well?**

In this correlation matrix, `engine_size` seem like the only feature not to have a strongly positive correlation with the CO2 emissions, it seems like a case worth studying on a higher degree polynomial!!

::: {#regression-models .section .level2} ## Regression Models:

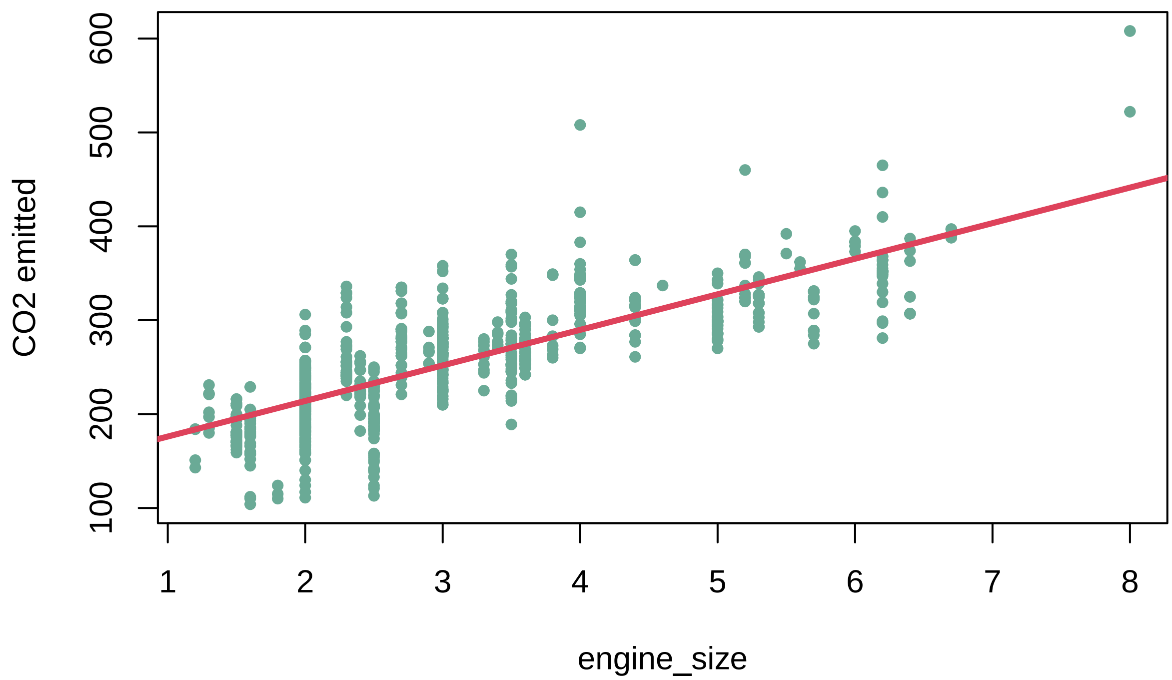
Linear Regression:

```
slr_engine_model <- lm(CO2_emission ~ engine_size, data=df)
summary(slr_engine_model)
```

Is there a significant relationship between engine_size and CO2 emission?

```
##
## Call:
## lm(formula = CO2_emission ~ engine_size, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.933  -23.132   -1.058   22.130  218.255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.2453     3.4057   40.59   <2e-16 ***
## engine_size  37.8749     0.9941   38.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.8 on 831 degrees of freedom
## Multiple R-squared:  0.636,  Adjusted R-squared:  0.6355
## F-statistic:  1452 on 1 and 831 DF,  p-value: < 2.2e-16
```

```r
plot(df$engine_size, df$CO2_emission, xlab = "engine_size", ylab= "CO2 emitted", col= "#6aaa96", pch=20)
abline(slr_engine_model, col="#de425b", lwd=3, lty=1)
```



Interpretation:

- P_value < 0.05, thus we reject H_0 which means that there is a linear relationship between engine size and CO2 emissions.
- The positive coefficient estimates for different engine_sizes indicate the expected increase in CO2 emissions for each level of engine_size compared to the reference level (Intercept).
- The model appears to explain a significant portion of the variation in CO2 emissions, as indicated by the high R-squared value `0.636`.
- The residuals are relatively small (`38.8`) and the coefficients have a low standard error, which is a sign of a good fit.

```r
slr_vehicle_model <- lm(CO2_emission ~ vehicle_class, data=df)
summary(slr_vehicle_model)
```

Is there a significant relationship between the vehicle type and CO2 emission?

```
##
## Call:
## lm(formula = CO2_emission ~ vehicle_class, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.770  -34.287   -5.287   30.713  289.971
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            210.984      6.846  30.818  < 2e-16 ***
## vehicle_classFull-size                  45.724     10.364   4.412 1.16e-05 ***
## vehicle_classMid-size                   18.998      8.533   2.226   0.0263 *
## vehicle_classMinicompact                52.925     13.377   3.956 8.27e-05 ***
## vehicle_classMinivan                    16.016     21.494   0.745   0.4564
## vehicle_classPickup truck: Small        65.428     14.758   4.433 1.05e-05 ***
## vehicle_classPickup truck: Standard     87.820      8.765  10.019  < 2e-16 ***
## vehicle_classSpecial purpose vehicle    41.816     25.061   1.669   0.0956 .
## vehicle_classStation wagon: Mid-size    76.238     19.229   3.965 7.99e-05 ***
## vehicle_classStation wagon: Small      -18.984     15.951  -1.190   0.2343
## vehicle_classSubcompact                 38.509      9.253   4.162 3.49e-05 ***
## vehicle_classSUV: Small                 18.303      7.859   2.329   0.0201 *
## vehicle_classSUV: Standard              93.786      8.270  11.340  < 2e-16 ***
## vehicle_classTwo-seater                107.045     11.397   9.392  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.91 on 819 degrees of freedom
## Multiple R-squared:  0.3073, Adjusted R-squared:  0.2963
## F-statistic: 27.95 on 13 and 819 DF,  p-value: < 2.2e-16
```

```r
#plot(df$vehicle_class, df$CO2_emission, xlab = "vehicle type", ylab= "CO2 emitted", col= "#6aaa96", pc
#abline(slr_vehicle_model, col="#de425b", lwd=3, lty=1)
```

Interpretation:

- P_value < 0.05, thus we reject H_0 which means that there is a linear relationship between vehicle class and CO2 emissions.

- The coefficient estimates for each level of vehicle class represent the expected change in CO2 emissions for vehicles in those categories compared to the reference level(intercept). If the coefficient estimate is positive, this indicates the expected increase in CO2 emissions. If the coefficient estimate, this indicates the expected decrease in CO2 emissions.
- Some levels of vehicle class have coefficients with low p-values, indicating statistical significance. These levels have a significant effect on CO2 emissions. For example, '*' and '***' significance codes suggest highly significant effects (*** > * in significance)
- In our model, R squared is around 0.3073. This means that roughly 30.73% of the variance in CO2 emissions is explained by the vehicle class. This could be an indication of a poor model
- The residual standard error (RSE) is 53.91, which does represent generally a low RSE. This RSE along with R squared indicates that the model is poor. **The interpretation of the metrics indicate that the model of simple linear regression of the CO2 emissions onto the vehice class is a poor fit.**

Multiple Regression: Backward Selection

```
full_model <- lm(CO2_emission ~ ., data = df)
final <- step(full_model, direction = "backward")
```

```
## Start:  AIC=1445.53
## CO2_emission ~ car_make + model_name + vehicle_class + engine_size +
##     transmission + fuel_type + city_consumption + hwy_consumption +
##     mix_consumption + vehicle_size_category + transmission_type_category
##
##
## Step:  AIC=1445.53
## CO2_emission ~ car_make + model_name + vehicle_class + engine_size +
##     transmission + fuel_type + city_consumption + hwy_consumption +
##     mix_consumption + vehicle_size_category
##
##
## Step:  AIC=1445.53
## CO2_emission ~ car_make + model_name + vehicle_class + engine_size +
##     transmission + fuel_type + city_consumption + hwy_consumption +
##     mix_consumption
##
##
## Step:  AIC=1445.53
## CO2_emission ~ car_make + model_name + engine_size + transmission +
##     fuel_type + city_consumption + hwy_consumption + mix_consumption
##
##
## Step:  AIC=1445.53
## CO2_emission ~ model_name + engine_size + transmission + fuel_type +
##     city_consumption + hwy_consumption + mix_consumption
##
##                    Df Sum of Sq     RSS    AIC
## - model_name       649    2349.8  3275.1 1200.4
## - transmission      15      28.0   953.3 1440.4
## <none>                            925.3 1445.5
## - engine_size        1       9.0   934.3 1451.6
## - hwy_consumption    1      12.6   937.9 1454.8
## - mix_consumption    1      20.3   945.6 1461.6
## - city_consumption   1      21.6   946.9 1462.8
## - fuel_type          3   15940.2 16865.5 3857.7
##
## Step:  AIC=1200.44
## CO2_emission ~ engine_size + transmission + fuel_type + city_consumption +
##     hwy_consumption + mix_consumption
##
##                   Df Sum of Sq    RSS    AIC
## - engine_size       1         5   3280 1199.8
## <none>                            3275 1200.4
## - transmission     22       180   3455 1201.0
## - mix_consumption   1        50   3325 1211.1
## - hwy_consumption   1       274   3549 1265.3
## - city_consumption  1       293   3568 1269.8
```

```
## - fuel_type         3    204081 207356 4649.8
##
## Step:  AIC=1199.79
## CO2_emission ~ transmission + fuel_type + city_consumption +
##     hwy_consumption + mix_consumption
##
##                    Df Sum of Sq    RSS    AIC
## - transmission     22       175   3455 1199.0
## <none>                            3280 1199.8
## - mix_consumption   1        49   3330 1210.3
## - hwy_consumption   1       274   3554 1264.5
## - city_consumption  1       301   3582 1271.0
## - fuel_type         3    210972 214253 4675.0
##
## Step:  AIC=1199.03
## CO2_emission ~ fuel_type + city_consumption + hwy_consumption +
##     mix_consumption
##
##                    Df Sum of Sq    RSS    AIC
## <none>                            3455 1199.0
## - mix_consumption   1        45   3500 1207.7
## - hwy_consumption   1       303   3758 1267.1
## - city_consumption  1       328   3783 1272.5
## - fuel_type         3    222504 225959 4675.4
```

```
summary(final)
```

```
##
## Call:
## lm(formula = CO2_emission ~ fuel_type + city_consumption + hwy_consumption +
##     mix_consumption, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1852 -0.9554 -0.0367  0.7333 22.7264
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        35.1691     0.5435   64.711  < 2e-16 ***
## fuel_typeE       -157.1097     0.7270 -216.115  < 2e-16 ***
## fuel_typeX        -34.5121     0.4700  -73.423  < 2e-16 ***
## fuel_typeZ        -35.3360     0.4757  -74.289  < 2e-16 ***
## city_consumption    9.4809     1.0714    8.849  < 2e-16 ***
## hwy_consumption     7.5246     0.8840    8.512  < 2e-16 ***
## mix_consumption     6.3685     1.9491    3.267  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.045 on 826 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 1.368e+05 on 6 and 826 DF,  p-value: < 2.2e-16
```

The main predictors that persisted the elimination:

- fuel type: (3 dummy var expressing 4 different types)

  - any fuel - there is a contribution in ~ 35g/km CO2 emitted (if diesel aloe it will e strictly this value)
  - else if fuel is ethanol (E85), there will e ~ 157 decrease in the qty emitted than diesel
  - else if fuel is regular gas (X) ~ 34 decrease
  - else premium (Z) ~ 35 decrease

  DIESEL IS THE MOST POLLUTANT

- city_consumption: 1u up will results in 9.4 g/km of CO2

- hwy_consumption: 1u up will results in 7.5 g/km of CO2

- mix_consumption: 1u up will results in 6.3 g/km of CO2

Interpretation: - The R-squared value is 0.999, indicating that approximately 99.9% of the variance in CO2 emissions is explained by the predictors in the model. - The F-statistic has an extremely low p-value (< 2.2e-16), which indicates that the model is highly significant, thus indicating that at least one of the predictors or interactions is influential in explaining CO2 emissions. - All the coefficients have low significant p-values. This is an indication of a great model. - The residual standard error (RSE) is 2.045, which represents generally a low RSE. This RSE indicates a very good and significant fit. To sum up, the model generated by backward selection is a very good and significant model.

::: {#interaction-effect .section .level3} ### Interaction Effect:

::: {#is-there-an-interaction-between-engine_size-and-fuel-type .section .level4} #### ***Is there an interaction between engine_size and fuel type?**

```
int_engine_fuel_model <- lm(CO2_emission ~ engine_size*fuel_type, data = df)
summary(int_engine_fuel_model)
```

```
##
## Call:
## lm(formula = CO2_emission ~ engine_size * fuel_type, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.766  -23.261   -2.678   20.313  211.956
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             160.261      9.284  17.262  < 2e-16 ***
## engine_size              36.713      1.261  29.105  < 2e-16 ***
## fuel_typeE               -8.483     63.667  -0.133  0.89403
## fuel_typeX              -30.480     10.619  -2.870  0.00421 **
## fuel_typeZ              -11.069      8.700  -1.272  0.20360
## engine_size:fuel_typeE   -8.382     13.078  -0.641  0.52175
## engine_size:fuel_typeX    1.681      2.121   0.793  0.42825
## engine_size:fuel_typeZ       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.91 on 826 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6519
## F-statistic: 260.7 on 6 and 826 DF,  p-value: < 2.2e-16
```

There is no interaction effect between any of the fuel types and the engine_size!!! We can only conclude the following (based of p value): * CO2 is up 36.7g/km when engine_size is bigger in 1 unit * y default, fuel will increase CO2 160g/km * regular gas X only +130g/km

i.e. so any increase in engine_size will affect the CO2 emitted strictly y its coefficient (same for the use of a particular fuel type) - do not depend/interact with each other

Is there an interaction between fuel type and fuel consumption in a combination of city and highway roads?

```r
int_mix_fuel_model <- lm(CO2_emission ~ mix_consumption*fuel_type, data = df)
summary(int_mix_fuel_model)
```

```
##
## Call:
## lm(formula = CO2_emission ~ mix_consumption * fuel_type, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6316 -0.9598 -0.0687  0.8178 24.1163
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.9307     9.1522   0.320 0.748889
## mix_consumption             26.5742     0.9084  29.253  < 2e-16 ***
## fuel_typeE                 -21.1801    10.8479  -1.952 0.051222 .
## fuel_typeX                  -2.8789     9.1599  -0.314 0.753381
## fuel_typeZ                  -3.3386     9.1605  -0.364 0.715607
## mix_consumption:fuel_typeE  -9.1066     0.9672  -9.416  < 2e-16 ***
## mix_consumption:fuel_typeX  -3.1147     0.9091  -3.426 0.000643 ***
## mix_consumption:fuel_typeZ  -3.1359     0.9090  -3.450 0.000589 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.809 on 825 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 1.499e+05 on 7 and 825 DF,  p-value: < 2.2e-16
```

Overall, this model suggests that the interaction between mix consumption and fuel type significantly influences CO2 emissions, and the model has a very high level of explanatory power.

Interpretation:

- The R-squared value (0.9992) is very high, indicating that a large proportion of the variance in CO2 emissions is explained by the predictors, including the interaction term.
- The F-statistic has an extremely low p-value, this means that the model is highly significant, indicating that at least one of the predictors, including the interaction terms, is influential in explaining CO2 emissions.
- It is true that not all the coefficients are significant (have a low p-value), but the significance codes suggest that most coefficients, especially all the **interaction terms** , are highly significant.

This interpretation suggests that the interaction between mix consumption and fuel type is very significant.

What happens if we perform a multiple regression model and include both interactions terms

```
multi_model <- lm(CO2_emission ~ engine_size*fuel_type + mix_consumption*fuel_type, data = df)

# View the summary of the model
summary(multi_model)
```

```
##
## Call:
## lm(formula = CO2_emission ~ engine_size * fuel_type + mix_consumption *
##     fuel_type, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4319 -0.9989 -0.0633  0.7939 24.0065
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.9295     9.1044   0.212 0.832213
## engine_size                  0.3337     0.1089   3.065 0.002244 **
## fuel_typeE                 -28.4766    11.7166  -2.430 0.015293 *
## fuel_typeX                  -1.8912     9.1122  -0.208 0.835632
## fuel_typeZ                  -1.7286     9.1219  -0.189 0.849749
## mix_consumption             26.5742     0.9031  29.426  < 2e-16 ***
## engine_size:fuel_typeE      -2.7229     1.3200  -2.063 0.039440 *
## engine_size:fuel_typeX      -0.3646     0.1600  -2.279 0.022921 *
## engine_size:fuel_typeZ           NA         NA      NA       NA
## fuel_typeE:mix_consumption  -7.9790     1.1445  -6.971 6.44e-12 ***
## fuel_typeX:mix_consumption  -3.1048     0.9046  -3.432 0.000628 ***
## fuel_typeZ:mix_consumption  -3.2850     0.9050  -3.630 0.000301 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.798 on 822 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 1.062e+05 on 10 and 822 DF,  p-value: < 2.2e-16
```

Overall, whenever we combine the interaction terms, this can lead to a high statistical significance.

Interpretation:

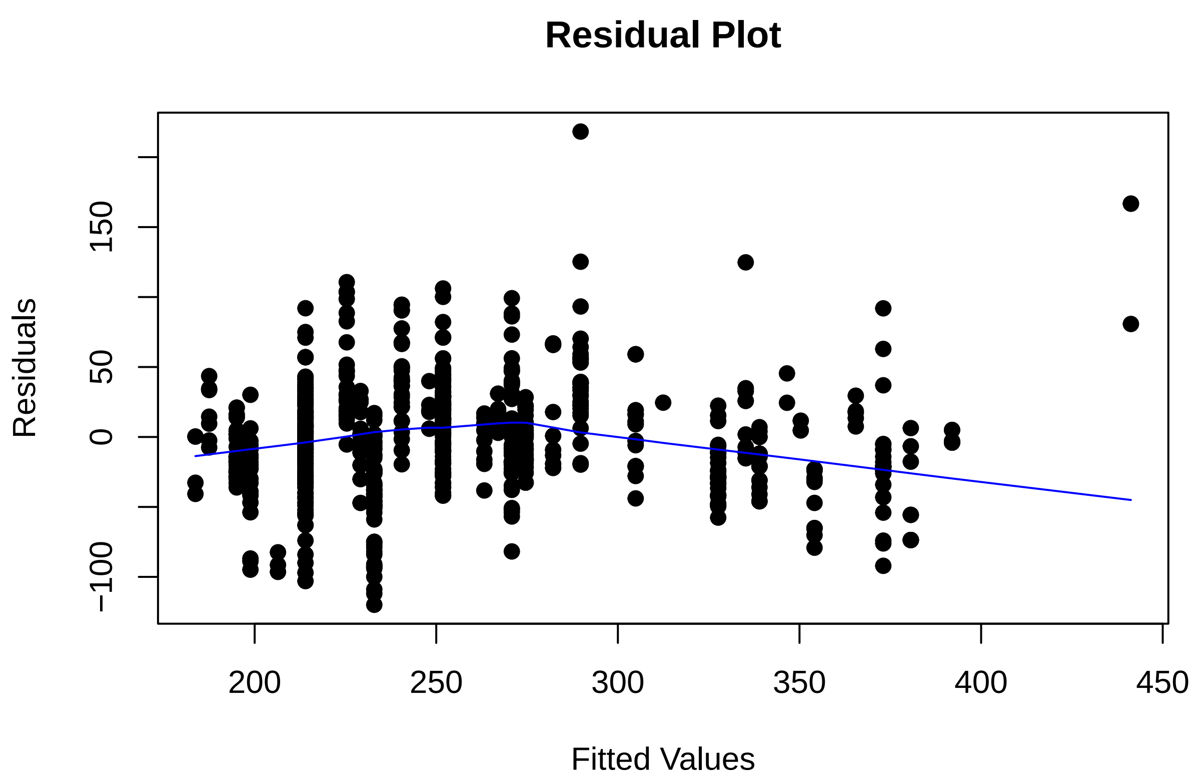
- The R-squared value is 0.9992, indicating that approximately 99.92% of the variance in CO_2 emissions is explained by the predictors and their interactions in the model.
- The F-statistic has an extremely low p-value ($< 2.2e\text{-}16$), which indicates that the model is highly significant, thus indicating that at least one of the predictors or interactions is influential in explaining CO_2 emissions.
- It is true that not all the coefficients are significant (have a low p-value), but the significance codes suggest that most coefficients, especially all the **interaction terms** , are highly significant.
- The residual standard error (RSE) is 1.798, which represents generally a low RSE. This RSE indicates a very good and significant fit.

Residual Plot to Determine Patterns

```r
# Compute the residuals
residuals <- resid(slr_engine_model)

# Create a scatterplot of residuals against the fitted values
plot(fitted(slr_engine_model), residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residual Plot",
     pch = 19)

# Add a Lowess curve to the plot
lines(lowess(fitted(slr_engine_model), residuals), col = "blue")
```



Notice from the residual plot that we can see a pattern. The presence of a pattern in the plots indicate a problem with some aspect of the linear model. There is a little pattern in the residuals, suggesting that applying polynomial regression can improve the fit to the data

Polynomial Regression

note: a polynomial model is still a linear regression model since the coef are linear.

```r
pol_engine_model <- lm(CO2_emission ~ poly(engine_size, 3), data=df)
summary(pol_engine_model)
```

```
##
## Call:
## lm(formula = CO2_emission ~ poly(engine_size, 3), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.561  -20.682   -0.149   20.318  215.557
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             257.472      1.291 199.360  < 2e-16 ***
## poly(engine_size, 3)1 1478.209     37.275  39.657  < 2e-16 ***
## poly(engine_size, 3)2 -137.344     37.275  -3.685 0.000244 ***
## poly(engine_size, 3)3  283.094     37.275   7.595 8.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.27 on 829 degrees of freedom
## Multiple R-squared:  0.6648, Adjusted R-squared:  0.6636
## F-statistic:   548 on 3 and 829 DF,  p-value: < 2.2e-16
```

```r
#plot(df$engine_size, df$CO2_emission, xlab = "engine_size", ylab= "CO2 levels", col= "#B9BDC1", pch=20
```

- R squared is 0.6648, i.e. the model explain 66.48% of the variation in the dependent variable
- F is high enough (548) to conclude that independent variables are jointly etter off in predicting the dependent variable than what's done individually.
- H_0 is rejected (p<0.05), the poly model is superior to the simple one degree model.

What does that mean?

- The coefficient for poly(engine_size, 3)1 is positive and highly significant, indicating that CO2 emissions increase as engine size increases. The coefficients for poly(engine_size, 3)2 is negative and significant, indicating that the rate of increase in CO2 emissions slows down as engine size increases. The poly(engine_size, 3)3 is + again.
- Engine size affects how much CO2 a car emits. The bigger the engine, the more CO2 the car emits. But the rate at which CO2 emissions increase slows down as engine size increases.), the only outliers that are present in our dataset are the ones with CO2 emi ## Conclusion of the Outliers

As it was mentioned before outliers that are present in our dataset are the ones with CO2 emissions = 608, and this was the value that we were constantly mentioning with the visualization of each plot (our famous P.S.). Hence,let's recall a major concept in statistics: Outliers are data points that significantly differ from the majority of the data in a dataset. While outliers are typically considered as unusual or extreme values, they can sometimes have legitimate reasons for their presence. And this is the case of our outliers; they have large CO2 emissions for legitimate reasons. This reason is: Heterogeneity. In certain datasets, heterogeneity or diversity among data points can lead to outliers. For example, in a dataset of income, a few individuals with exceptionally high incomes may be outliers, but they are valid data points.If we go back to our outliers in our dataset, we can see that it is a result of heterogeneity, because if we inspect about our outliers, we discover that they are bugati cars with engine size = 8, they use premium gasoline and they are 2 seated(special). Hence, we notice that the bugati outlier is to the heterogeneity or diversity among data points.

BIF524: Data Mining

Course Project Phase I: Regression

LAU: Fall 2023

Acknowledgments

Unfortunately, Roudy's laptop fell and the screen was broken and the same happened to Rayane's screen. Hence two people must be thanked.

Thanks to Rayane's sister for letting Rayane use her laptop.

Thanks to Roudy's brother for letting Roudy use his laptop.