

SAS-PROJECT CIS-5250



CO2 EMISSION BY VEHICLE

CALIFORNIA STATE UNIVERSITY LOS ANGELES

SUBMITTED TO: DR. SHILPA BALAN

SUBMITTED BY: POOJA MADHUP

CIN: 401977573

TABLE OF CONTENTS

<u>TOPIC</u>	<u>PAGE NO</u>
<u>PART -A</u>	
Dataset: List of dataset(s) URL's -----	3
List of Abbreviations -----	3
<u>PART-B</u>	
Introduction -----	5
<u>PART-C</u>	
Dataset Description -----	6
<u>PART-D</u>	
Data Cleaning -----	8
<u>PART-E</u>	
Data Visualization -----	13
<u>PART-F</u>	
Statistical Summary -----	21
<u>PART-G</u>	
Statistical Tests -----	23
Conclusion -----	32
<u>PART-H</u>	
References -----	33

A. Dataset: List of dataset(s) URL's

Data set link

[CO2 Emission by Vehicles | Kaggle](#)

[Fuel consumption ratings - Open Government Portal \(canada.ca\)](#)

Website link

<https://www.kaggle.com/>

<https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles?select=CO2+Emissions+Canada.csv>

List of Abbreviation's:

This dataset demonstrates how vehicle-released co2 emission might vary depending on different automobile features. This dataset is collected from the Canadian government's official open data websites.[1]. There are 12 columns and 7385 rows overall. The fuel consumption ratings for cities and highways of Canada are displayed in liters per 100 Kilometers(L/100Km), and the combined rating (45% highway, 55% cities) is given both in L/100Km and miles per gallon (mpg). The combined carbon dioxide emissions from city and highway driving are measured in grams per kilometer. These units will be displayed in the dataset and data description as well. The features have been described using a few abbreviations. Here is the list of them. The model, Transmission, and Fuel type columns have been used abbreviations. The data description table (dataset) contains the same information. The complete form of the below abbreviation with column name is as follows

Model

4WD/4X4 = Four-wheel drive

AWD = All-wheel drive

FFV = Flexible-fuel vehicle

SWB = Short wheelbase

LWB = Long wheelbase

EWB = Extended wheelbase

Transmission

A = Automatic

AM = Automated manual

AS = Automatic with select shift

AV = Continuously variable

M = Manual

3 - 10 = Number of gears

Fuel Type

X = Regular gasoline

Z = Premium gasoline

D = Diesel

E = Ethanol (E85)

N = Natural gas

B. INTRODUCTION:

Carbon dioxide (CO₂) emissions are a global issue that impacts climate change. The emissions have increased constantly throughout the years, with some decreasing points of historical world events, such as the pandemic. Since 1940, global carbon dioxide (CO₂) emissions from fossil fuels and industry have increased almost yearly. From 2000, the increase was even more considerable. A few years do not follow the same pattern, however. Some major global events can cause emissions reduction.[2] The CO₂ emissions produced by passenger cars have been steadily rising over the past two decades. Those vehicles produced approximately three billion tons of carbon dioxide emissions worldwide in 2020. These numbers represent almost 10% of the carbon dioxide emissions in the year. Also, 6% reduction compared to 2019 due to the pandemic.[3] The urgency to reduce CO₂ emissions is widely recognized, as they are the chief driver of global climate. The average Canadian car will burn 21 tons of gasoline over its lifespan, equaling 130 barrels. Therefore, emission of CO₂ is a severe problem.

MOTIVATION: The motivation behind this topic is to analyze and contribute to reducing carbon dioxide concerning a vehicle that helps to get an environmentally pollution-free and healthy life. [4] We want to explore the data and see the number of factors that can impact a vehicle's fuel efficiency and CO₂ emission using exploratory data analysis. Data visualization is a vital component of exploratory data analysis since it enables a data analyst to look at their data and become familiar with the variables and relationships among them. During this project, we will clean the data, analyze the data in visualization form and statistical summary, find the correlation between the response variable and other factors. We will then perform linear regression to measure the impact of features on response variable. This data analysis report can be used by the Canadian Government and other researchers to spread the awareness of the CO₂ emission that will affect the environment and human life.

C. DATASET DESCRIPTION:

The Carbon dioxide (CO₂) emission by vehicle dataset originated from the Canadian Government's official data websites, which monitor fuel economy and consumption to become energy efficient in Canada. [5] The dataset contains 12 columns and 7386 rows. [6] The dataset consists of column names, and data descriptions are as follows:

Column Name	Data Description	Examples
Make (String)	It represents the company of the vehicle or vehicle brand (make) used in the Canada country.	Audi, BMW, Honda etc.
Model (String)	It represents the model of the car	Audi - A4 Honda - Accord BMW - 320i
Vehicle class (String)	It represents the vehicle type such as small, compact, or mid-size based on the car weight, horsepower, and engine etc.	Compact Sub-Compact Full-Size Pickup-Truck Standard etc.
Engine Size in L (Integer)	It represents the size of engine for vehicle used in liters	2 liters, 2.4 liters
Cylinders (Integer)	It represents number of cylinders used in the vehicle	3, 4, 8, 10, 12
Transmission (string)	It represents the number of gears used while transmission type.	A = Automatic M = manual
Fuel type (string)	It represents the different type of fuel used	X = Regular gasoline Z= Premium gasoline

Fuel consumption city (Decimal)	It represents the fuel consumed by the vehicle in the city area in the kilometers(L/100Km)	9.9 L/100Km
Fuel consumption hwy (highway) (Decimal)	It represents the fuel consumed by the vehicle in the highway road in the kilometers(L/100Km)	6.7 L/100Km
Fuel consumption comb in Km (Decimal)	It represents the fuel consumed by the vehicle in the highway and city both the road in the kilometers(L/100Km)	8.5 L/100Km
Fuel consumption comb in mpg (Integer)	It represents the combined fuel consumed by the vehicle in the highway and city both the road in the miles per gallon form (mpg)	33 mpg (miles per gallon)
CO2 Emission (Integer)	It represents the total combined CO2 emission (carbon dioxide) by the driving vehicles in the both the area high and city in the grams per kilometers.	196 g/Km

An extract of co2 emission by vehicle dataset are as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Make	Model	Vehicle Class	Engine Size(L Cylinders	Transmission	Fuel Type	Fuel Consumption City	Fuel Consumption Hwy	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)	
2	ACURA	ILX	COMPACT	2	4 AS5	Z	9.9	6.7	8.5	33	196	
3	ACURA	ILX	COMPACT	2.4	4 M6	Z	11.2	7.7	9.6	29	221	
4	ACURA	ILX HYBRID	COMPACT	1.5	4 AV7	Z	6	5.8	5.9	48	136	
5	ACURA	MDX 4WD	SUV - SMALL	3.5	6 AS6	Z	12.7	9.1	11.1	25	255	

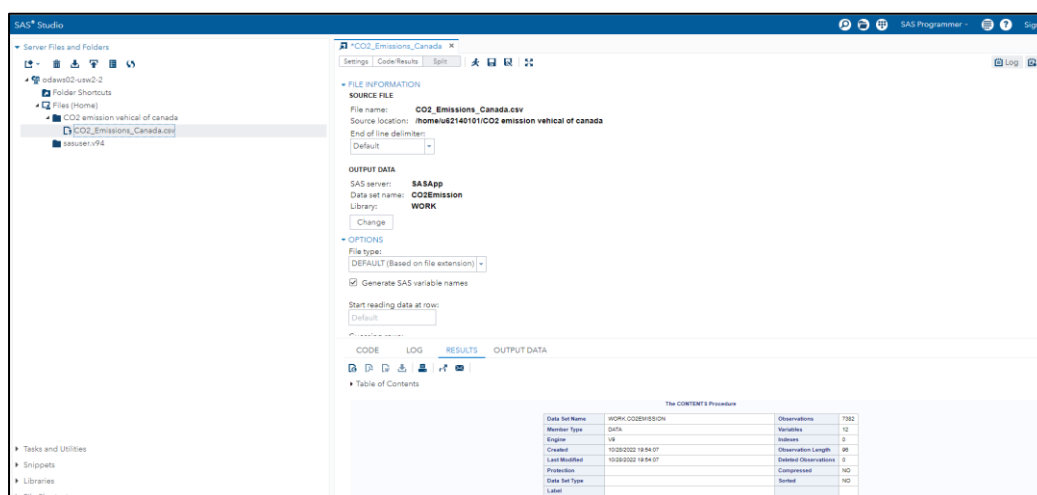
Tools and Technologies Used

Language Used: SAS 9.4

IDE Used: SAS Studio (SAS OnDemand for Academic)

D. DATA CLEANING

The data we obtain from the official government open data portal website is usually raw data of the dataset. This data must be cleaned enough to analyze the more superficial and adequate data. Data cleaning also plays a vital role in data analysis to get high-quality data and increase overall productivity. We need to eliminate redundant, inaccurate data, and irrelevant data. Many techniques are used for data cleaning, and they may differ from dataset to dataset. Some data cleaning approaches that have been used to clean the co2 emission by vehicle dataset are listed below. Microsoft Excel has been used for cleaning the dataset.



CODE LOG RESULTS OUTPUT DATA							
Table: WORK.CO2EMISSION View: Column names Filter: (none)							
Total rows: 7382 Total columns: 12							
Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)
1 ACURA	ILX	COMPACT	2	4	AS5	Z	9.9
2 ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2
3 ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6
4 ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7
5 ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1
6 ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9
7 ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8
8 ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8
9 ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4
10 ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6
11 ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2
12 ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1
13 ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	Z	9.7

1) Missing Values

Missing data is very common in a real-life dataset. The missing value role is essential to improve the dataset's accuracy. In the below dataset, the column 'Fuel Consumption city' have missing values. The

values for BMW cars (make) in the column fuel consumption city are entered by the fuel consumption rating data from the open government website [5]. For instance, the below screenshot is highlighted in yellow for the fuel consumption city for the 'BMW 550i xDRIVE GRAN TURISMO model is 15', and the fuel consumption city for 'The BMW 760Li model is 18.7' respectively.

Before Cleaning:

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)
01	BMW	528i xDRIVE	MID-SIZE	2	4 A8	Z		10.6	7.2
02	BMW	535d xDRIVE	MID-SIZE	3	6 A8	D		9.2	6.4
03	BMW	535i xDRIVE	MID-SIZE	3	6 A8	Z		11.9	8
04	BMW	535i xDRIVE GRAN TURISMO	FULL-SIZE	3	6 A8	Z		12.5	8.6
05	BMW	550i xDRIVE	MID-SIZE	4.4	8 A8	Z		14.4	9.6
06	BMW	550i xDRIVE GRAN TURISMO	FULL-SIZE	4.4	8 A8	Z			9.8
07	BMW	640i xDRIVE GRAN COUPE	COMPACT	3	6 A8	Z		11.9	8
08	BMW	650i xDRIVE CABRIOLET	SUBCOMPACT	4.4	8 A8	Z		15	9.8
09	BMW	650i xDRIVE COUPE	COMPACT	4.4	8 A8	Z		14.4	9.6
10	BMW	650i xDRIVE GRAN COUPE	COMPACT	4.4	8 A8	Z		15	9.8
11	BMW	740Li xDRIVE	FULL-SIZE	3	6 A8	Z		12.5	8.6
12	BMW	750i xDRIVE	FULL-SIZE	4.4	8 A8	Z		15	9.8
13	BMW	750Li xDRIVE	FULL-SIZE	4.4	8 A8	Z		15	9.8
14	BMW	760Li	FULL-SIZE	6	12 A8	Z			11.5
15	BMW	ACTIVEHYBRID 3	COMPACT	3	6 A8	Z		9.2	7.1
16	BMW	ACTIVEHYBRID 5	MID-SIZE	3	6 A8	Z		10.5	7.9
17	BMW	ACTIVEHYBRID 7L	FULL-SIZE	3	6 A8	Z		10.5	7.6

After Cleaning:

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)
101	BMW	528i xDRIVE	MID-SIZE	2	4 A8	Z		10.6	7.2
102	BMW	535d xDRIVE	MID-SIZE	3	6 A8	D		9.2	6.4
103	BMW	535i xDRIVE	MID-SIZE	3	6 A8	Z		11.9	8
104	BMW	535i xDRIVE GRAN TURISMO	FULL-SIZE	3	6 A8	Z		12.5	8.6
105	BMW	550i xDRIVE	MID-SIZE	4.4	8 A8	Z		14.4	9.6
106	BMW	550i xDRIVE GRAN TURISMO	FULL-SIZE	4.4	8 A8	Z		15	9.8
107	BMW	640i xDRIVE GRAN COUPE	COMPACT	3	6 A8	Z		11.9	8
108	BMW	650i xDRIVE CABRIOLET	SUBCOMPACT	4.4	8 A8	Z		15	9.8
109	BMW	650i xDRIVE COUPE	COMPACT	4.4	8 A8	Z		14.4	9.6
110	BMW	650i xDRIVE GRAN COUPE	COMPACT	4.4	8 A8	Z		15	9.8
111	BMW	740Li xDRIVE	FULL-SIZE	3	6 A8	Z		12.5	8.6
112	BMW	750i xDRIVE	FULL-SIZE	4.4	8 A8	Z		15	9.8
113	BMW	750Li xDRIVE	FULL-SIZE	4.4	8 A8	Z		15	9.8
114	BMW	760Li	FULL-SIZE	6	12 A8	Z		18.7	11.5
115	BMW	ACTIVEHYBRID 3	COMPACT	3	6 A8	Z		9.2	7.1
116	BMW	ACTIVEHYBRID 5	MID-SIZE	3	6 A8	Z		10.5	7.9
117	BMW	ACTIVEHYBRID 7L	FULL-SIZE	3	6 A8	Z		10.5	7.6

2) Illegal Values

Fixing the illegal values in the dataset is very crucial to increase the productivity level. The **Fuel consumption comb (mpg)** field for the Hyundai Genesis AWD model contains illegal values, as highlighted below in yellow. The value after cleaning is replaced by '20' based on calculating the value from the fuel consumption rating tool of the Canadian government. To get the fuel consumption comb

value, we are using the fuel consumption rating tool to calculate the value; refer to this link <https://fcc-ccc.nrcan-rncan.gc.ca/en>.

Before Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmissi	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions
1596	HYUNDAI	ELANTRA GT	MID-SIZE	2	4	M6	X	9.8	6.9	8.5	33	196
1597	HYUNDAI	EQUUS	FULL-SIZE	5	8	AS8	Z	15.8	10.2	13.3	21	306
1598	HYUNDAI	GENESIS AWD	FULL-SIZE	3.8	6	AS8	X	14.4	9.4	12.1	23	278
1599	HYUNDAI	GENESIS AWD	FULL-SIZE	5	8	AS8	Z	17.3	10.5	14.2	!#%	327
1600	HYUNDAI	GENESIS COUPE SUBCOMP		3.8	6	AS8	Z	14.6	9.6	12.3	23	283

After Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmissi	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions
1596	HYUNDAI	ELANTRA GT	MID-SIZE	2	4	M6	X	9.8	6.9	8.5	33	196
1597	HYUNDAI	EQUUS	FULL-SIZE	5	8	AS8	Z	15.8	10.2	13.3	21	306
1598	HYUNDAI	GENESIS AWD	FULL-SIZE	3.8	6	AS8	X	14.4	9.4	12.1	23	278
1599	HYUNDAI	GENESIS AWD	FULL-SIZE	5	8	AS8	Z	17.3	10.5	14.2	20	327
1600	HYUNDAI	GENESIS COUPE SUBCOMP		3.8	6	AS8	Z	14.6	9.6	12.3	23	283

3) Duplicate Records

Duplicate records can slow down the data and confuse. As a result, duplicate columns and rows in the dataset should be eliminated. The highlighted rows of the **BMW 740Li xDRIVE SEDAN** car model has two of the exact details. Hence, this comes under the same record case. Rows for the BMW 740Li xDRIVE SEDAN car model row have been discarded after posit-cleaning; it removes redundancy from the analysis.

Before Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmissi	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions
1191	BMW	435i xDRIVE COUPE	COMPACT	3	6	AS8	Z	11.9	7.8	10	28	230
1192	BMW	435i xDRIVE COUPE	COMPACT	3	6	M6	Z	12	8.3	10.3	27	237
1193	BMW	435i xDRIVE GRAN COUP	COMPACT	3	6	AS8	Z	11.9	7.8	10	28	230
1194	BMW	528i SEDAN	MID-SIZE	2	4	AS8	Z	10.4	7.1	8.9	32	205
1195	BMW	528i xDRIVE SEDAN	MID-SIZE	2	4	AS8	Z	10.6	7.2	9.1	31	209
1196	BMW	535d xDRIVE SEDAN	MID-SIZE	3	6	AS8	D	9.2	6.3	7.9	36	213
1197	BMW	535i xDRIVE SEDAN	MID-SIZE	3	6	AS8	Z	12.1	8.1	10.3	27	237
1198	BMW	535i xDRIVE GRAN TURIS	FULL-SIZE	3	6	AS8	Z	12.8	8.9	11	26	253
1199	BMW	550i xDRIVE SEDAN	MID-SIZE	4.4	8	AS8	Z	14.4	9.7	12.3	23	283
1200	BMW	550i xDRIVE GRAN TURIS	FULL-SIZE	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1201	BMW	640i xDRIVE CABRIOLET	SUBCOMPACT	3	6	AS8	Z	12.1	8.1	10.3	27	237
1202	BMW	640i xDRIVE GRAN COUP	COMPACT	3	6	AS8	Z	12.1	8.1	10.3	27	237
1203	BMW	650i xDRIVE CABRIOLET	SUBCOMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1204	BMW	650i xDRIVE COUPE	COMPACT	4.4	8	AS8	Z	14.4	9.7	12.3	23	283
1205	BMW	650i xDRIVE GRAN COUP	COMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1206	BMW	740ld xDRIVE SEDAN	FULL-SIZE	3	6	AS8	D	10.2	7.1	8.8	32	238
1207	BMW	740Li xDRIVE SEDAN	FULL-SIZE	3	6	AS8	Z	12.1	8.1	10.3	27	237
1208	BMW	750i xDRIVE SEDAN	FULL-SIZE	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1209	BMW	750Li xDRIVE SEDAN	FULL-SIZE	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1210	BMW	760Li SEDAN	FULL-SIZE	6	12	AS8	Z	18.7	11.6	15.5	18	356
1211	BMW	ACTIVEHYBRID 3	COMPACT	3	6	AS8	Z	9.7	7.5	8.7	32	200
1212	BMW	ACTIVEHYBRID 5	MID-SIZE	3	6	AS8	Z	10.6	7.9	9.4	30	216
1213	BMW	ACTIVEHYBRID 7L	FULL-SIZE	3	6	AS8	Z	10.6	7.7	9.3	30	214
1214	BMW	ALPINA B6 xDRIVE GRAN	COMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294

After Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmissi	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions
1191	BMW	435i xDRIVE COUPE	COMPACT	3	6	AS8	Z	11.9	7.8	10	28	230
1192	BMW	435i xDRIVE COUPE	COMPACT	3	6	M6	Z	12	8.3	10.3	27	237
1193	BMW	435i xDRIVE GRAN COUP	COMPACT	3	6	AS8	Z	11.9	7.8	10	28	230
1194	BMW	528i SEDAN	MID-SIZE	2	4	AS8	Z	10.4	7.1	8.9	32	205
1195	BMW	528i xDRIVE SEDAN	MID-SIZE	2	4	AS8	Z	10.6	7.2	9.1	31	209
1196	BMW	535d xDRIVE SEDAN	MID-SIZE	3	6	AS8	D	9.2	6.3	7.9	36	213
1197	BMW	535i xDRIVE SEDAN	MID-SIZE	3	6	AS8	Z	12.1	8.1	10.3	27	237
1198	BMW	535i xDRIVE GRAN TURIS	FULL-SIZE	3	6	AS8	Z	12.8	8.9	11	26	253
1199	BMW	550i xDRIVE SEDAN	MID-SIZE	4.4	8	AS8	Z	14.4	9.7	12.3	23	283
1200	BMW	550i xDRIVE GRAN TURIS	FULL-SIZE	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1201	BMW	640i xDRIVE CABRIOLET	SUBCOMPACT	3	6	AS8	Z	12.1	8.1	10.3	27	237
1202	BMW	640i xDRIVE GRAN COUP	COMPACT	3	6	AS8	Z	12.1	8.1	10.3	27	237
1203	BMW	650i xDRIVE CABRIOLET	SUBCOMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1204	BMW	650i xDRIVE COUPE	COMPACT	4.4	8	AS8	Z	14.4	9.7	12.3	23	283
1205	BMW	650i xDRIVE GRAN COUP	COMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1206	BMW	740Li xDRIVE SEDAN	MID-SIZE	3	6	AS8	D	10.2	7.1	8.8	32	238
1207	BMW	740Li xDRIVE SEDAN	FULL-SIZE	3	6	AS8	Z	12.1	8.1	10.3	27	237
1208	BMW	750Li xDRIVE SEDAN	FULL-SIZE	4.4	8	AS8	Z	15.2	9.8	12.8	22	294
1209	BMW	760Li SEDAN	FULL-SIZE	6	12	AS8	Z	18.7	11.6	15.5	18	356
1210	BMW	ACTIVEHYBRID 3	COMPACT	3	6	AS8	Z	9.7	7.5	8.7	32	200
1211	BMW	ACTIVEHYBRID 5	MID-SIZE	3	6	AS8	Z	10.6	7.9	9.4	30	216
1212	BMW	ACTIVEHYBRID 7L	FULL-SIZE	3	6	AS8	Z	10.6	7.7	9.3	30	214
1213	BMW	ALPINA B6 xDRIVE GRAN	COMPACT	4.4	8	AS8	Z	15.2	9.8	12.8	22	294

4) Misspelling Name

Correcting the misspelling of a name in a dataset is essential since it may lead to uncertainty—the highlighted part of the ‘make’ (manufacture) screenshot column is shown below. SCION brand name is misspelled. Therefore, we replaced the incorrect value of **SCION** with to correct value throughout the column ‘Make.’

Before Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinder	Transmissi
2061	ROLLS-ROYCE	PHANTOM COUPE	COMPACT	6.7	12	AS8
2062	ROLLS-ROYCE	PHANTOM DROPHEAD CO	COMPACT	6.7	12	AS8
2063	ROLLS-ROYCE	WRAITH	MID-SIZE	6.6	12	AS8
2064	SCIONIO	FR-S	MINICOMPAC	2	4	AS6
2065	SCIONIO	FR-S	MINICOMPAC	2	4	M6
2066	SCIONIO	iQ	MINICOMPAC	1.3	4	AV
2067	SCIONIO	tC	COMPACT	2.5	4	AS6
2068	SCIONIO	tC	COMPACT	2.5	4	M6
2069	SCIONIO	xB	STATION WAG	2.4	4	AS4
2070	SCIONIO	xB	STATION WAG	2.4	4	M5
2071	SMART	FORTWO CABRIOLET	TWO-SEATER	1	3	AM5

After Cleaning

1	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission
2059	ROLLS-ROYCE	PHANTOM	FULL-SIZE	6.7	12	AS8
2060	ROLLS-ROYCE	PHANTOM EWB	FULL-SIZE	6.7	12	AS8
2061	ROLLS-ROYCE	PHANTOM COUPE	COMPACT	6.7	12	AS8
2062	ROLLS-ROYCE	PHANTOM DROPHEAD COUPE	COMPACT	6.7	12	AS8
2063	ROLLS-ROYCE	WRAITH	MID-SIZE	6.6	12	AS8
2064	SCION	FR-S	MINICOMPACT	2	4	AS6
2065	SCION	FR-S	MINICOMPACT	2	4	M6
2066	SCION	iQ	MINICOMPACT	1.3	4	AV
2067	SCION	tC	COMPACT	2.5	4	AS6
2068	SCION	tC	COMPACT	2.5	4	M6
2069	SCION	xB	STATION WAGON	2.4	4	AS4
2070	SCION	xB	STATION WAGON	2.4	4	M5
2071	SMART	FORTWO CABRIOLET	TWO-SEATER	1	3	AM5

5) Filtering

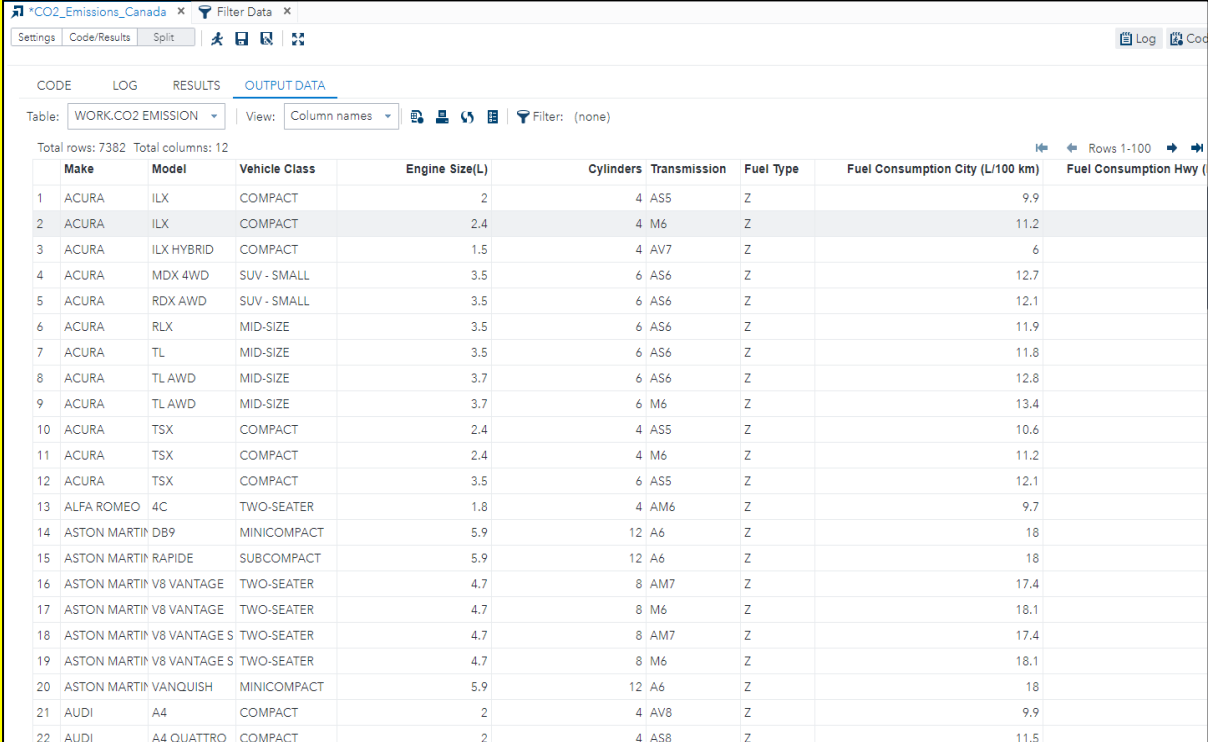
Before filtering the dataset using SAS studio, we first import the dataset in SAS studio and view the dataset co2 emission by vehicle in tabular format. Here, we found that the dataset has vast data, and we need to filter the data to analyze some scenarios.

Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)
1	ACURA	ILX	COMPACT	2	4	AS5	9.9
2	ACURA	ILX	COMPACT	2.4	4	M6	11.2
3	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	6
4	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	12.7
5	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	12.1
6	ACURA	RLX	MID-SIZE	3.5	6	AS6	11.9
7	ACURA	TL	MID-SIZE	3.5	6	AS6	11.8
8	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	12.8
9	ACURA	TL AWD	MID-SIZE	3.7	6	M6	13.4
10	ACURA	TSX	COMPACT	2.4	4	AS5	10.6
11	ACURA	TSX	COMPACT	2.4	4	M6	11.2
12	ACURA	TSX	COMPACT	3.5	6	AS5	12.1
13	ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	9.7
14	ASTON MARTIN	DB9	MINICOMPACT	5.9	12	A6	18

Filtering the data is a standard method to identify specific observations. The below pre-filtered table shows 7382 rows and 12 columns. We need to filter the data for co2 emission column to

get the highest data emission for further visualization. The chart below shows the co2 emission value between 96 to 522 gm/km. In order to analyze the highest rate of co2 emission with the dataset. We use the filter function. After filtering, the dataset of the co2 emission column for greater than 405 value with all the variables. As a result, we get 86 rows and 12 columns of the dataset. This filtered dataset has been used for further data visualization and statistical section. We are only using the filtered dataset for the analysis below. The rest of the analysis has been done on full dataset.

Before Filtering



	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)
1	ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	
2	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	
3	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6	
4	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	
5	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	
6	ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	
7	ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	
8	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	
9	ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	
10	ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	
11	ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2	
12	ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1	
13	ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	Z	9.7	
14	ASTON MARTIN	DB9	MINICOMPACT	5.9	12	A6	Z	18	
15	ASTON MARTIN	RAPIDE	SUBCOMPACT	5.9	12	A6	Z	18	
16	ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	AM7	Z	17.4	
17	ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	M6	Z	18.1	
18	ASTON MARTIN	V8 VANTAGE S	TWO-SEATER	4.7	8	AM7	Z	17.4	
19	ASTON MARTIN	V8 VANTAGE S	TWO-SEATER	4.7	8	M6	Z	18.1	
20	ASTON MARTIN	VANQUISH	MINICOMPACT	5.9	12	A6	Z	18	
21	AUDI	A4	COMPACT	2	4	AV8	Z	9.9	
22	AUDI	A4 QUATTRO	COMPACT	2	4	AS8	Z	11.5	

After Filtering

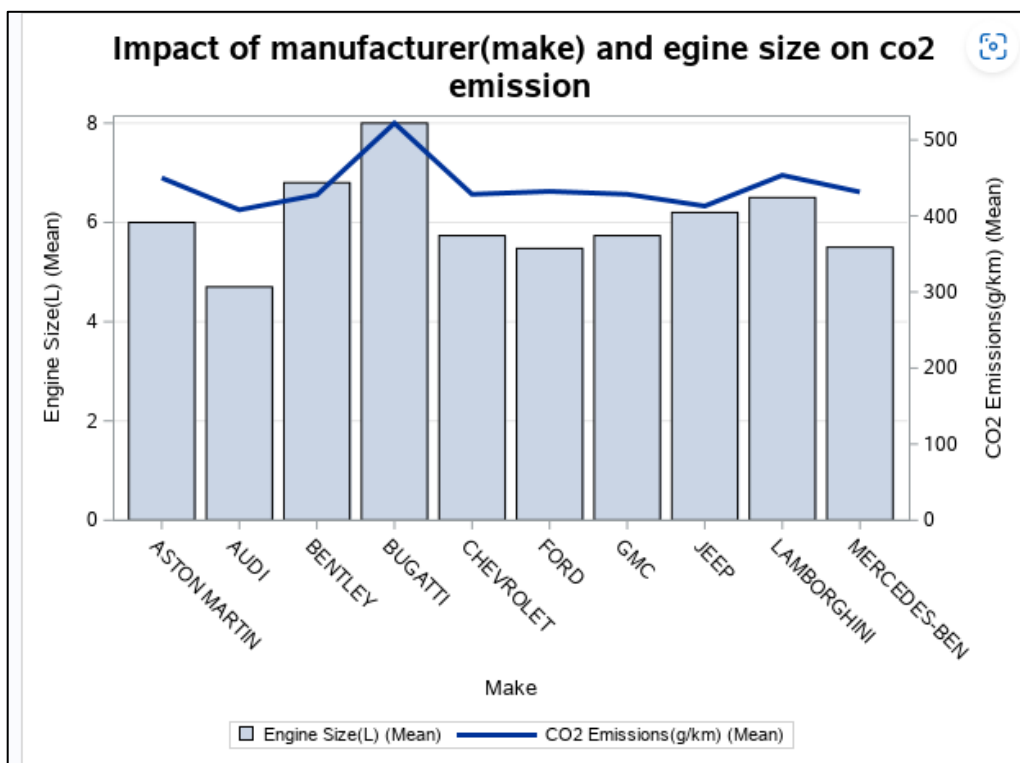
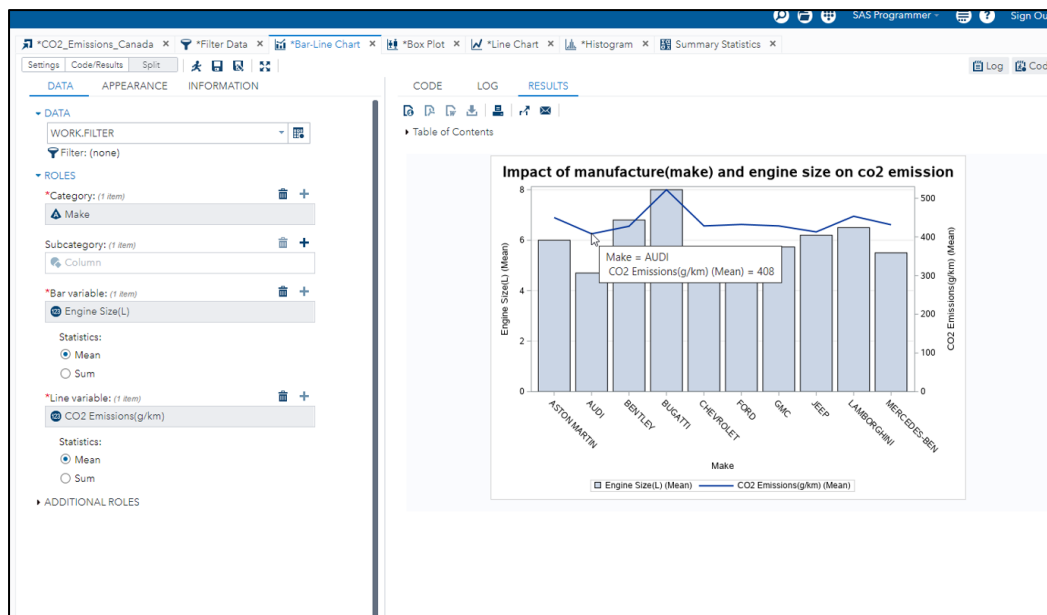
The screenshot displays the SAS Studio interface for a data filtering operation. On the left, the 'DATA' tab is active, showing a filter rule for 'CO2 Emissions(g/km)' with a comparison of 'Greater than' and a value of '405'. The 'OUTPUT DATA' tab on the right shows a table with 86 rows and 6 columns: Make, Model, Vehicle Class, Engine Size(L), Cylinders, and Transmis. The table lists various vehicle models and their specifications.

Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmis
GMC	SAVANA 2500	VAN - PASSE	6	8	A6
GMC	SAVANA 3500	VAN - PASSE	4.8	8	A6
GMC	SAVANA 3500	VAN - PASSE	6	8	A6
GMC	SAVANA 3500	VAN - PASSE	6	8	A6
GMC	SAVANA 3500	VAN - PASSE	6	8	A6
LAMBORGHINI	AVENTADOR C TWO-SEATER		6.5	12	AM7
LAMBORGHINI	AVENTADOR R TWO-SEATER		6.5	12	AM7
LAMBORGHINI	VENENO ROAD TWO-SEATER		6.5	12	AM7
MERCEDES-BEI	G 550	SUV - STAND	5.5	8	AS7
MERCEDES-BEI	G 63 AMG	SUV - STAND	5.5	8	AS7
BENTLEY	MULSANNE	MID-SIZE	6.8	8	AS8
CHEVROLET	EXPRESS 2500	VAN - PASSE	6	8	A6
CHEVROLET	EXPRESS 3500	VAN - PASSE	6	8	A6
GMC	SAVANA 2500	VAN - PASSE	6	8	A6
GMC	SAVANA 3500	VAN - PASSE	6	8	A6
LAMBORGHINI	AVENTADOR C TWO-SEATER		6.5	12	AM7
LAMBORGHINI	AVENTADOR R TWO-SEATER		6.5	12	AM7
MERCEDES-BEI	AMG G 63	SUV - STAND	5.5	8	AS7
MERCEDES-BEI	AMG G 65	SUV - STAND	6	12	AS7
ASTON MARTIN	V12 VANTAGE	TWO-SEATER	6	12	M7
BENTLEY	MULSANNE	MID-SIZE	6.8	8	AS8
BENTLEY	MULSANNE EW	MID-SIZE	6.8	8	AS8
CHEVROLET	EXPRESS 2500	VAN - PASSE	6	8	A6

E. DATA VISUALIZATION

In this section, we have created some visualization in charts and graphical form. This visualization helps us identify the vehicle features impacted most by increasing co2 emissions using a vehicle while driving in Canada. Which car make has affected to increase of co2 emission in the environment and so on. Data visualization tools are helpful for understanding and exploring the co2 emission by vehicle data at scale. In this project, we used the SAS studio tool for data analysis. There are two screenshots displayed to define each visualization. One screenshot suggests the graphical result, and the other gives us information about the graph, such as what column is selected in category variables, sub-categories, and numerical variables are getting used to visualize the data.

1. How does the manufacturer (make) and engine size impact the CO2 emission?



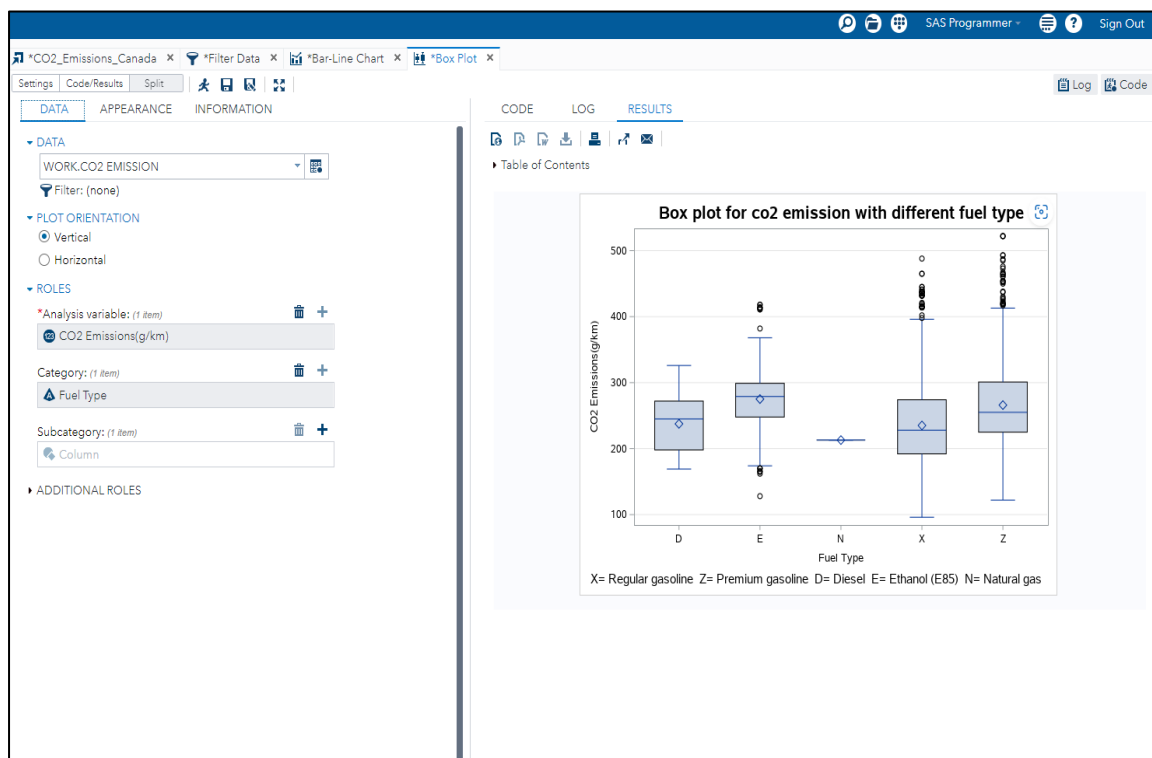
Application used: Filter, Box-Line Chart

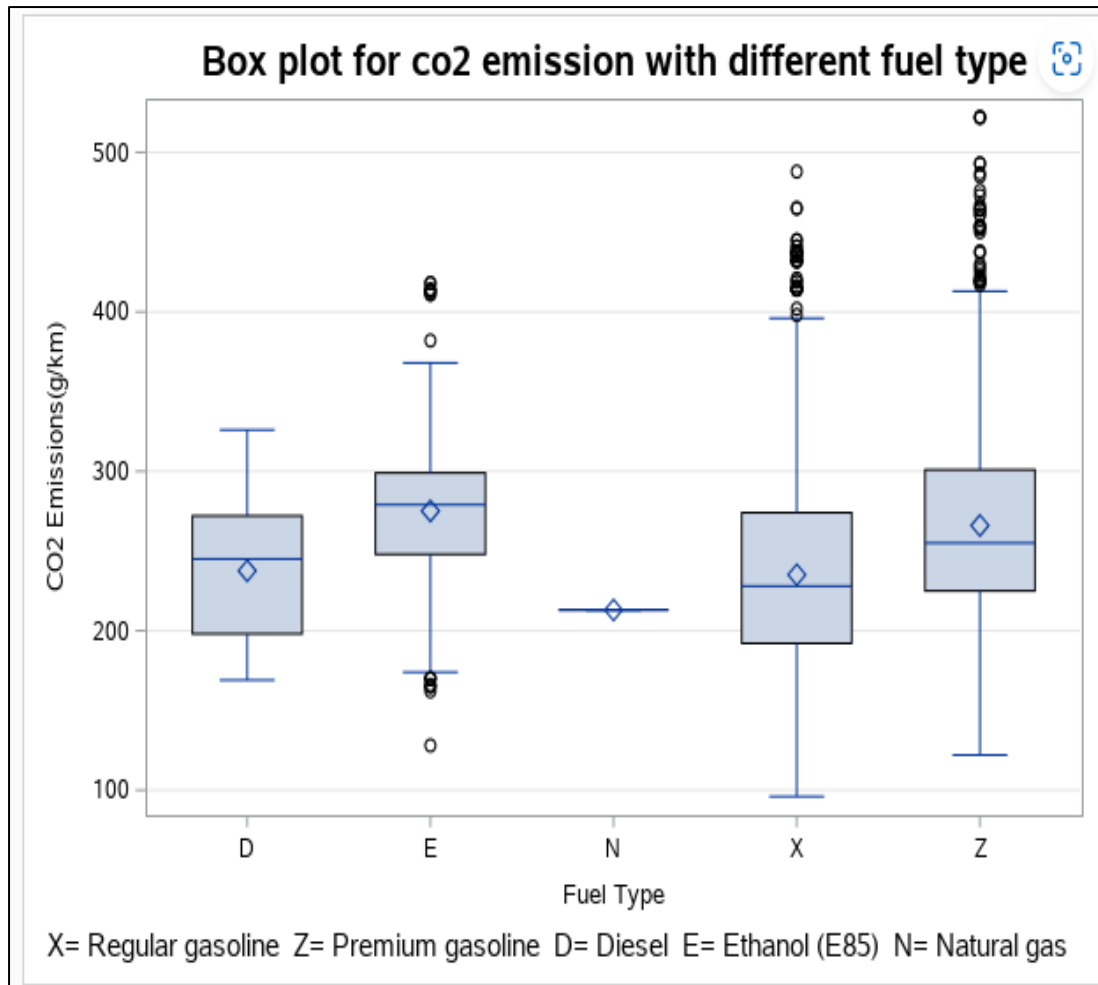
Insights:

This visualization illustrates the average size of the engine and average co2 emission rate changes among the car manufacturers in Canada. This visualization is done using the bar-line

chart. The left axis represents the average engine size in liters, and the right axis shows the average co2 emission in grams per kilometer. We observe that the Bugatti car emitted the highest carbon dioxide, around 552 gm/km, and the highest volume of 8 liters engine size, among others. The bar represents the engine size in a liter of the given car brand list, and the line in the graph indicates the average co2 emissions in g/km concerning the make. We used a filtered dataset for this visualization because we like to display that the topmost car-producing company produces the highest carbon dioxide emissions while driving. In the Bar-Line chart, we can conclude that the higher the engine size released, the more carbon dioxide the manufacturer followed. The top three car brands that polluted the environment easily are Bugatti, Bentley, and Jeep, followed by Aston martin compared to others, while Audi emitted the lowest co2 emission. We believe that if people want to buy a car, they should consider environmental pollution, which has less co2 emission.

2. How much variance in co2 emission present in different fuel type?





Application used: Boxplot

Insights:

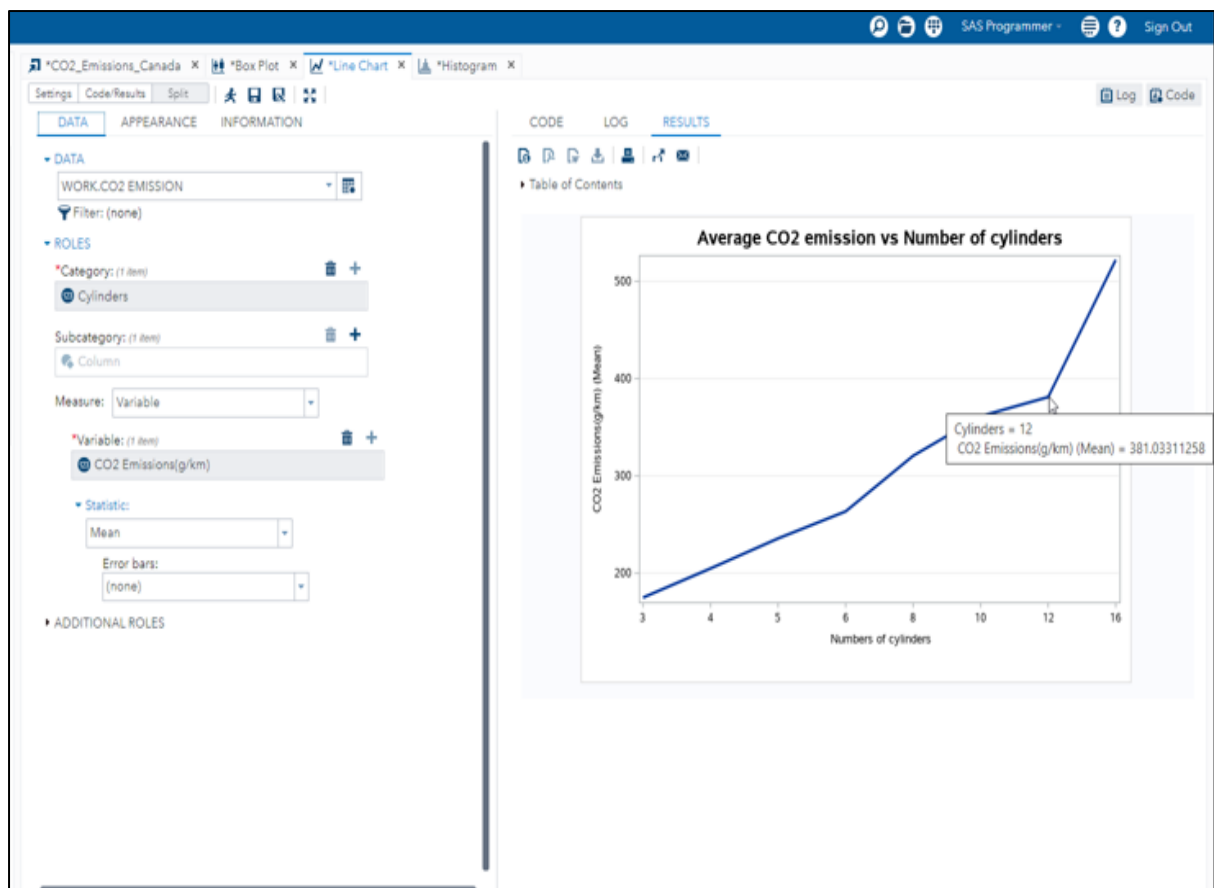
This data visualization displayed the relationship between carbon dioxide emission and fuel types. The abbreviation of fuel type shows in the footnote of the graph. This visualization uses the box plot of co2 emissions by vehicle dataset. We selected co2 emissions data from the drop-down box in the data section, then selected co2 emission as the analysis variable with fuel type as a category. Fuel type plays an essential role for different types of cars and their model. It is clear from the boxplot that the better-quality fuel type increases the co2 emission value while the diesel fuel type decreases the co2 emission. Taking cursor on the boxes of the plot will give details about the mean, median, maximum, minimum values, number of

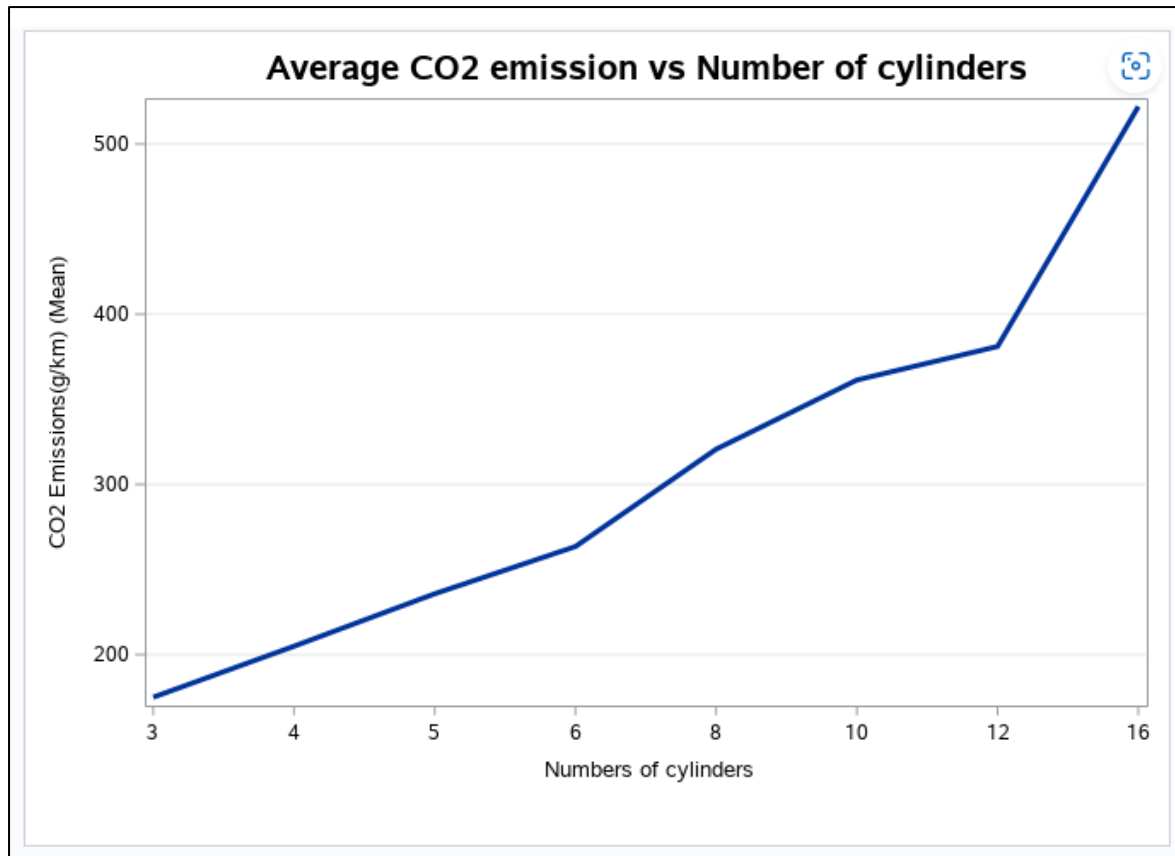
observations, and quartile ranges using a boxplot. For example, the graph shows that fuel type z (premium gasoline) released more than 419 grams of carbon dioxide emissions.

In contrast, natural gas is rarely used in the car and releases less carbon dioxide emissions.

Therefore, we conclude that the broader the size of the box of fuel type will increase the carbon dioxide emissions of a vehicle, and there are some outliers in the above figure. We should always choose the efficient fuel type to get the lower level of carbon dioxide emissions from a vehicle.

3. Does the number of cylinders used in the vehicle have any impact on the average of co2 emission?



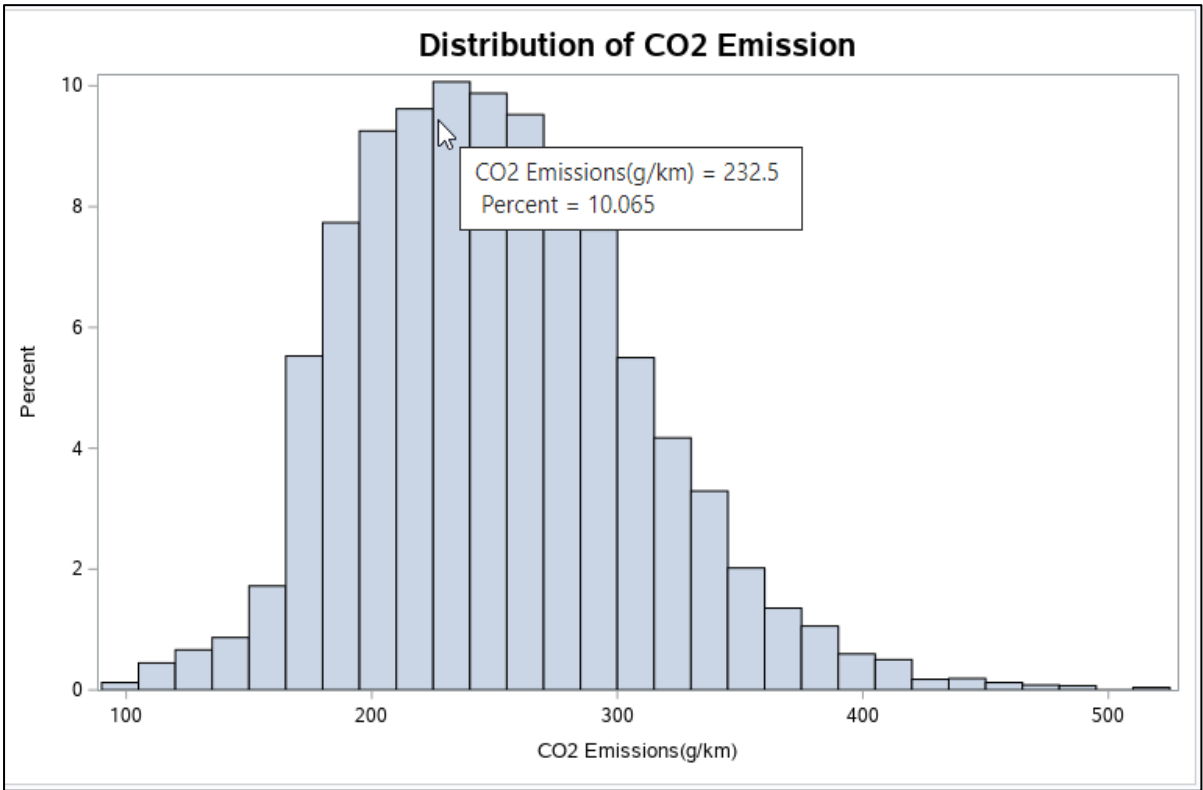
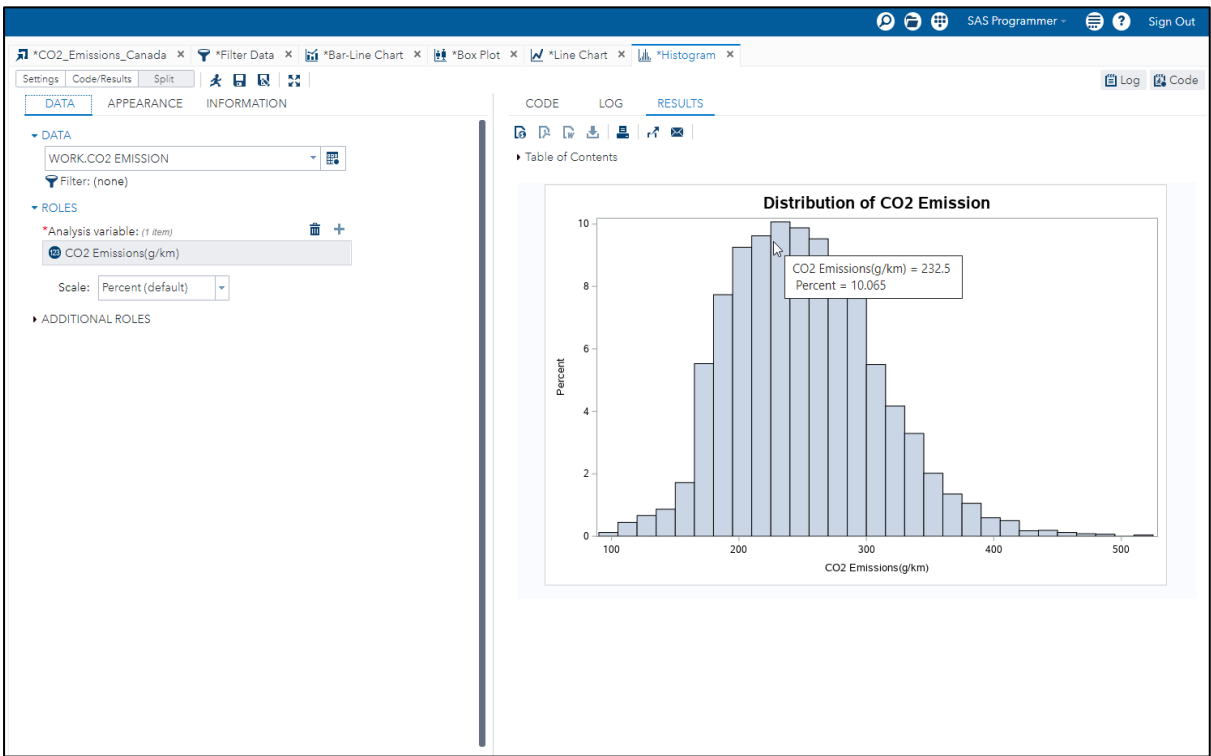


Application used: Line Chart

Insights:

This visualization shows the effect of average carbon dioxide released in grams per kilometer with the number of cylinders used in the vehicle. This analysis is done using a line chart. The above graph shows the approximately linear relationship between the average carbon dioxide emissions and the number of cylinders used for a car. In other words, we can conclude that the number of cylinders for the vehicle is directly proportional to the average carbon dioxide emissions. For example, if the number of cylinders in the car is 8, the average carbon dioxide emissions become 381 grams per kilometer. Similarly, carbon dioxide emissions decrease with fewer cylinders for a vehicle. Based on this analysis, we should use less number of cylinders in our vehicles to get less carbon dioxide emission and make the environment and human life healthy.

4. How we illustrate the distribution of CO2 emissions?



Application used: Histogram

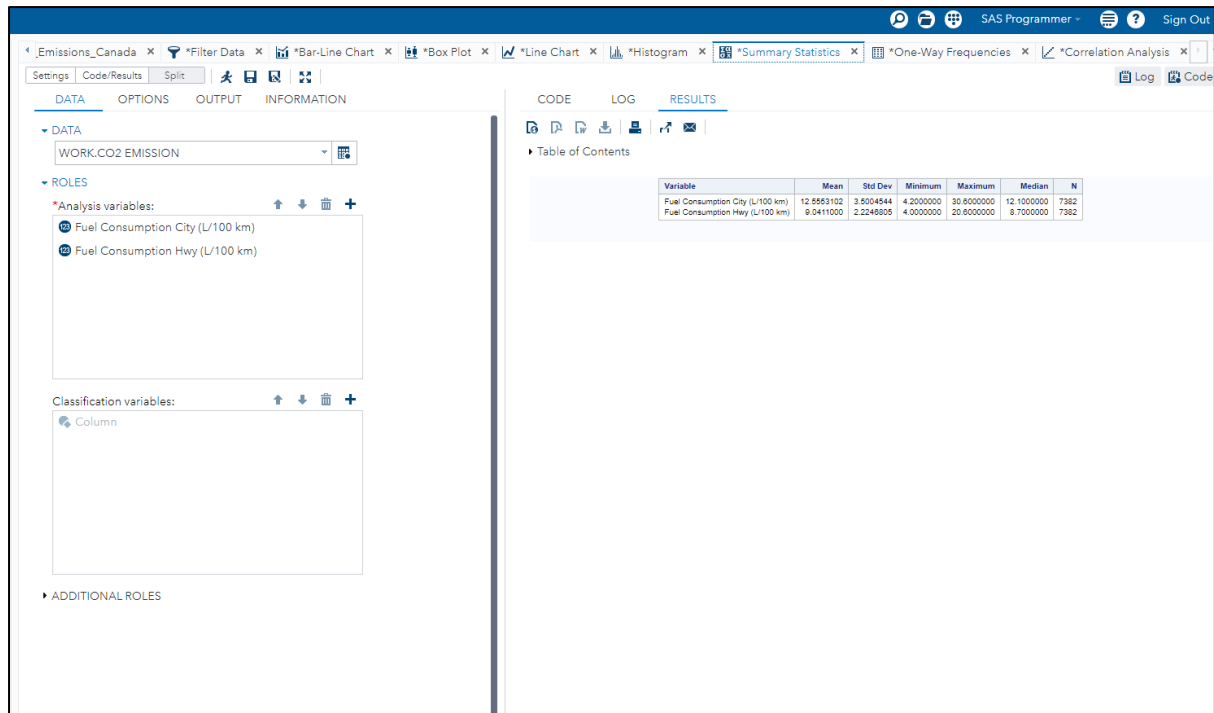
Insights:

This visualization displays the percentage of carbon dioxide emissions in Canada. The histogram is being used for analysis purposes. The histogram is similar to a bar chart. The only difference between them is that histogram is used for continuous variables, whereas a bar chart is used for categorical data. We can say the percentage of carbon dioxide emitted from the given data with the number of grams per kilometer. For example, 10 is the highest percentage released, 232.5 g/km of co₂ in Canada, and less than 0.5 percent emitted more than 397.5 grams of carbon dioxide. On taking cursor in any bar on the figure, we can see the details of co₂ emission in percentage and gram per kilometer. The percentage of emitted carbon dioxide needed when we analyze the carbon dioxide emission not only for a country but also useful for analysis worldwide. Such visualization does make data analysis more accessible and straightforward.

F. STATISTICAL SUMMARY

Summary Statistics for fuel consumption in city and fuel consumption in highway in Canada

Summary statistics are a part of descriptive statistics that summarizes and provides the gist of the information about the selected data. Data summaries usually present the datasets of mean, median, maximum value, minimum value, and standard deviation. We analyze summary statistics of two variables, such as the fuel consumption city and fuel consumption highway column from the dataset as shown in the above tabular form.



Variable	Mean	Std Dev	Minimum	Maximum	Median	N
Fuel Consumption City (L/100 km)	12.5553102	3.5004544	4.2000000	30.6000000	12.1000000	7382
Fuel Consumption Hwy (L/100 km)	9.0411000	2.2246805	4.0000000	20.6000000	8.7000000	7382

Mean: The summary statistics depict that the average fuel consumption in the city used the vehicle across Canada country is 12.55 liters per 100 kilometers with respect to carbon dioxide emission. In comparison, the average fuel consumption on the highway is 9 liters per 100 kilometers. As a result, the population driving in the city consumes more fuel in their vehicle than driving on the highway.

Standard Deviation: The summary statistics show the standard deviation for the fuel consumption in the city is 3.50 liters and fuel consumption on the highway is 2.24 liters. This concludes that the data fuel consumed on the highway has a lower variance than the fuel consumed in the city.

Minimum and Maximum: The summary statistics shows that the minimum value of fuel consumption in the city of a car with respect to carbon dioxide emissions is 4.20 liters, and the

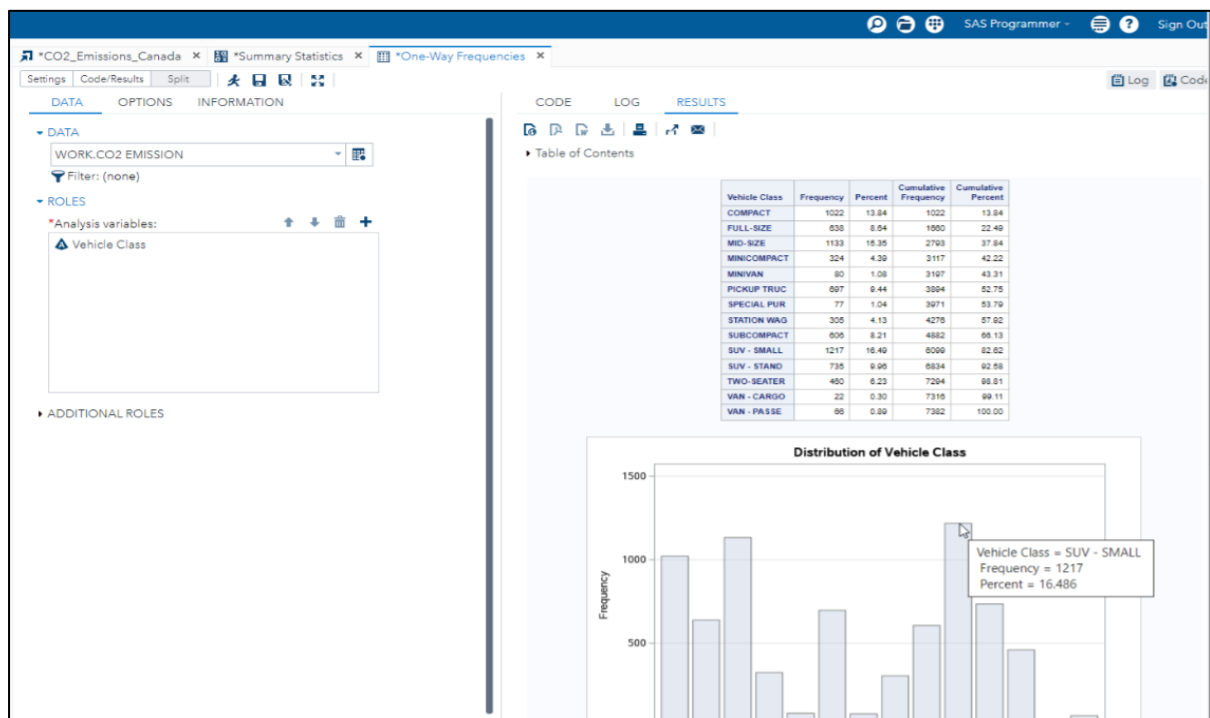
maximum value is 30.60 liters. On the other hand, a vehicle's fuel consumption on the highway is a minimum of 4 liters per 100 kilometers, and the maximum value is 20.60 liters per 100 kilometers. Hence, we conclude that the carbon dioxide emitted is less corresponding to the fuel consumption on the highway than city across the country.

Median: The median value, which is the middle value for the fuel consumption in the city, is 12.1 liters per 100 kilometers, whereas the fuel consumption on the highway of a vehicle is 8.7 liters per 100 kilometers which is lower than the fuel consumed in the city. Hence, we can summarize that less fuel consumption decreases carbon dioxide emissions.

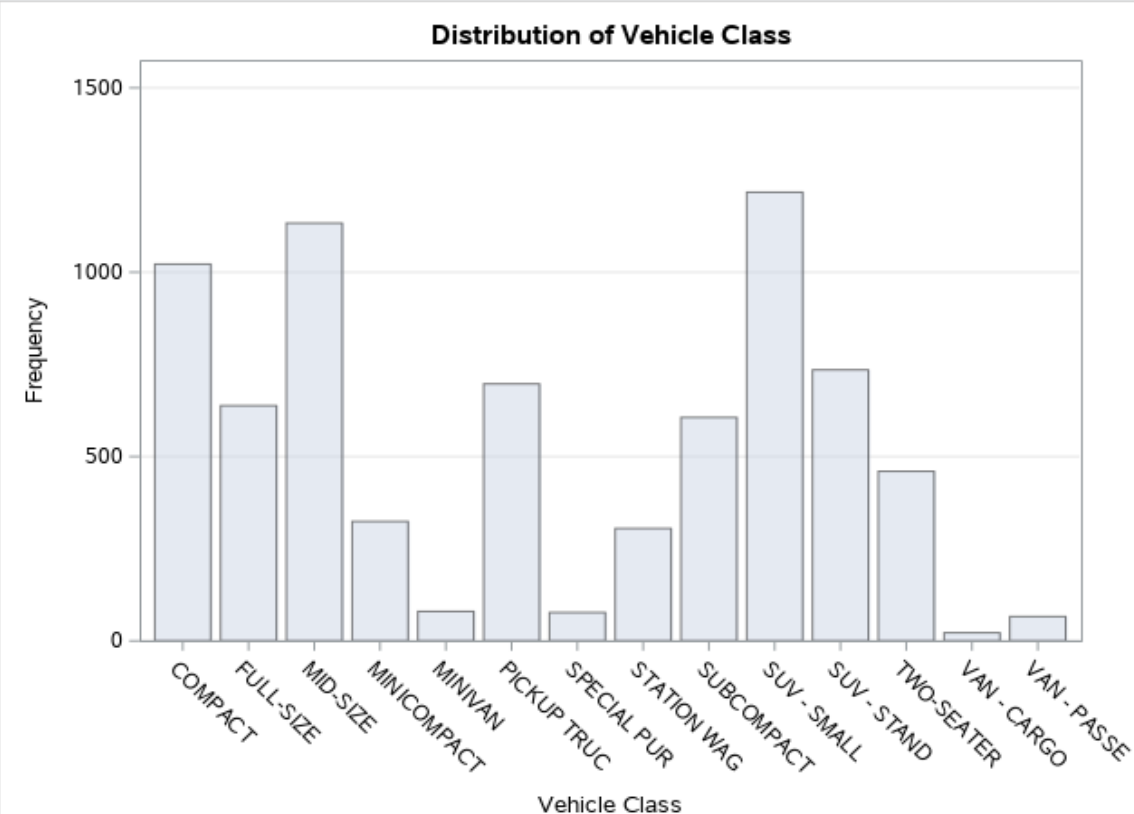
G. STATISTICAL TEST

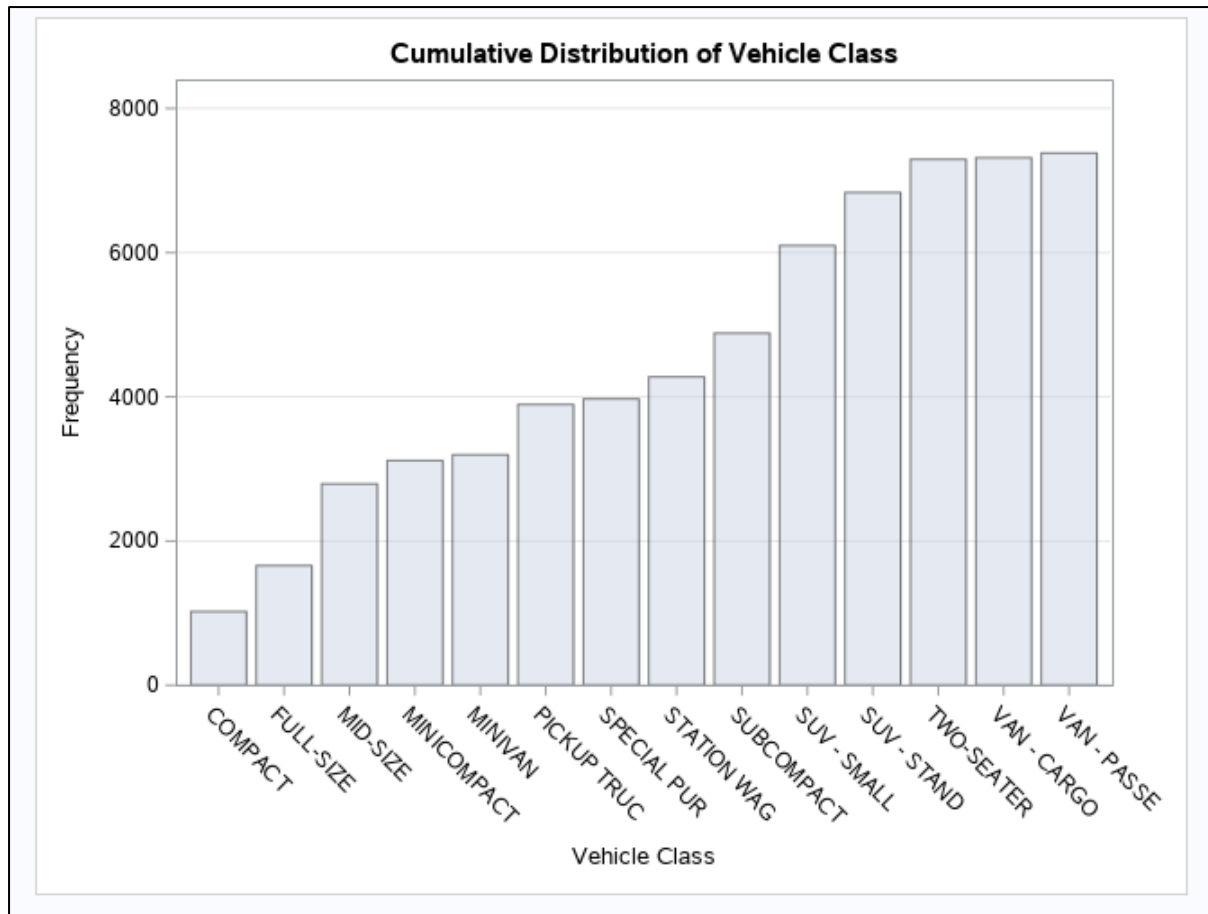
A statistical test is a method to evaluate the null hypothesis by analyzing data from the experiment to determine whether a predicted variable has a relationship with the outcome variable, and it helps us to estimate the difference between two groups.[7] The following section covers one-way frequency, correlation, and linear regression between variables.

i. Identify the frequency distribution of vehicle class column using one-way frequency



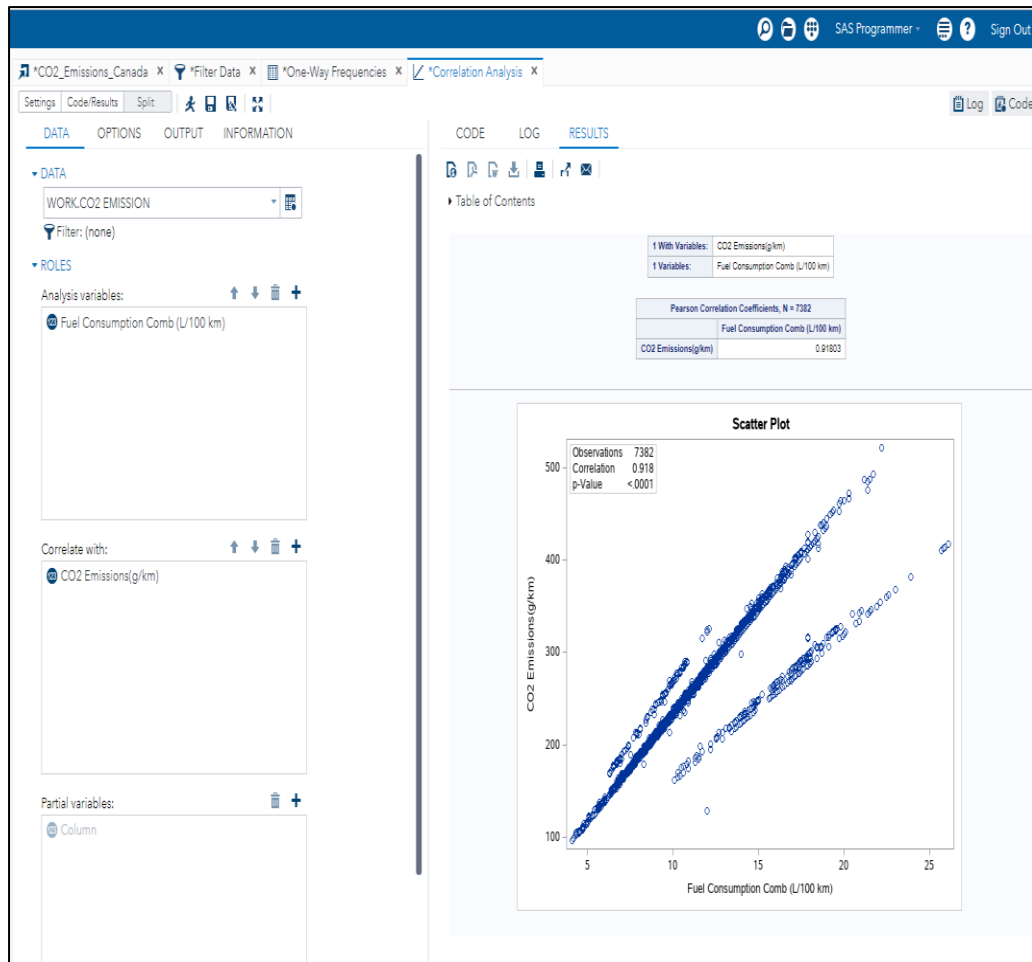
Vehicle Class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
COMPACT	1022	13.84	1022	13.84
FULL-SIZE	638	8.64	1660	22.49
MID-SIZE	1133	15.35	2793	37.84
MINICOMPACT	324	4.39	3117	42.22
MINIVAN	80	1.08	3197	43.31
PICKUP TRUC	697	9.44	3894	52.75
SPECIAL PUR	77	1.04	3971	53.79
STATION WAG	305	4.13	4276	57.92
SUBCOMPACT	606	8.21	4882	66.13
SUV - SMALL	1217	16.49	6099	82.62
SUV - STAND	735	9.96	6834	92.58
TWO-SEATER	460	6.23	7294	98.81
VAN - CARGO	22	0.30	7316	99.11
VAN - PASSE	66	0.89	7382	100.00





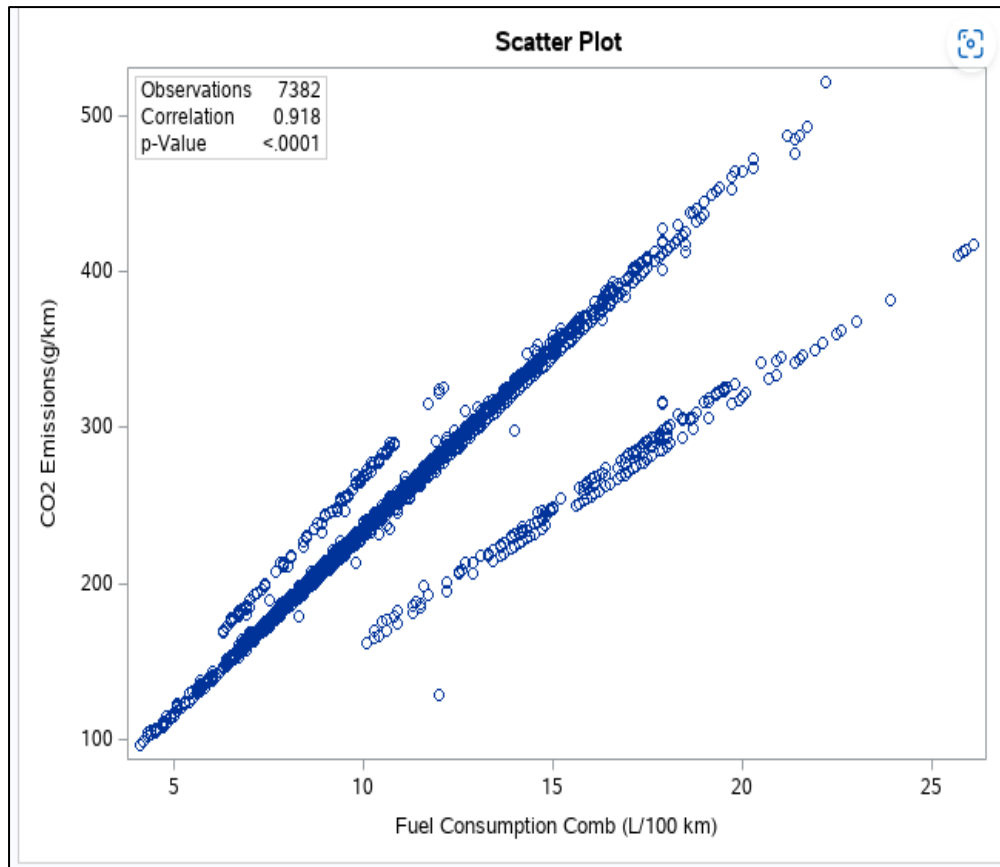
One-way frequency is a data tabulation that only considers one categorical variable at a time. With the help of one-way frequency statistics, the frequency, percent, cumulative frequency, and cumulative percent columns of the selected categorical variable **Vehicle Class** from the co2 emission by vehicle dataset are shown in the above table. There are fourteen types of vehicle classes which are listed in the table. We can identify that the SUV-SMALL vehicle class is the most used car type with a frequency of 1,217, which make up 16.49% distribution. Moreover, the VAN-PASSE vehicle class is rarely used with a frequency of 66, which is a 0.89% distribution. The cumulative frequency of VAN-PASSE vehicle class is the highest, 7,382, which makes up 100% cumulative percentage distribution. In comparison, the lowest is the COMPACT vehicle class type which is 1,022 cumulative frequency which makes 13.84% cumulative distribution respectively.

- ii. **Find out the correlation between fuel consumption combined both (Highway, City) driving and co2 emissions using scatter plot.**



1 With Variables:	CO2 Emissions(g/km)
1 Variables:	Fuel Consumption Comb (L/100 km)

Pearson Correlation Coefficients, N = 7382	
	Fuel Consumption Comb (L/100 km)
CO2 Emissions(g/km)	0.91803



A statistical metric called correlation represented as a number, indicates the strength and direction of a relationship between two or more variables. However, a correlation between variables does not imply that a change in one variable is the direct result of a change in the values of the other variable.[8] There are three types of correlations positive correlation, negative correlation, and no correlation. A positive correlation is a relationship between two variables that occurs when two variables move in the same direction. As a result, one variable increase while the other increases and vice versa. A negative correlation is a relationship between two variables in which an increase in one variable relates to a decrease in another variable. A zero correlation emerges when there is no relationship between two variables.

For the correlation, we analyzed the fuel consumption comb (city and highway) column correlated with co2 emissions column using the given dataset. As a result, the Pearson correlation coefficient (N) of the fuel consumption comb variable with the co2 emissions

response variable is 7382. The correlation between the selected two variables is 0.918, which concludes that there is a positive correlation between the fuel consumption comb variable and carbon dioxide emissions as the response variable. In other words, fuel consumption in a combination of cities and highways increases while carbon dioxide emission increases. The scatter plot shows the linear relationship between the two variables respectively.

iii. Performance linear regression for linear significance using

The screenshot displays the SAS Programmer interface for a linear regression model. The left pane shows the DATA tab with 'WORK.CO2 EMISSION' as the dependent variable and 'Cylinders' as a classification variable. The right pane shows the RESULTS tab with a table of contents and summary statistics.

Table of Contents

Data Set	WORK.CO2 EMISSION
Dependent Variable	CO2 Emissions(g/km)
Selection Method	None

Number of Observations Read	7382
Number of Observations Used	7382

Class Level Information

Class	Levels	Values
Cylinders	8	3 4 5 6 8 10 12 16

Dimensions

Number of Effects	2
Number of Parameters	2

Least Squares Summary

Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	60086.5651
1	Engine Size(L)	2	50582.7571*

* Optimal Value of Criterion

Least Squares Model (No Selection)

Data Set	WORK.CO2 EMISSION
Dependent Variable	CO2 Emissions(g/km)
Selection Method	None

Number of Observations Read	7382
Number of Observations Used	7382

Class Level Information		
Class	Levels	Values
Cylinders	8	3 4 5 6 8 10 12 16

Dimensions	
Number of Effects	2
Number of Parameters	2

Least Squares Summary			
Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	60086.5651
1	Engine Size(L)	2	50582.7571*
* Optimal Value of Criterion			

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18305697	18305697	19393.4	<.0001
Error	7380	6966071	943.91207		
Corrected Total	7381	25271769			

Root MSE	30.72315
Dependent Mean	250.56258
R-Square	0.7244
Adj R-Sq	0.7243
AIC	57953
AICC	57953
SBC	50583

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	134.360993	0.907812	148.01	<.0001
Engine Size(L)	1	36.779625	0.264107	139.26	<.0001

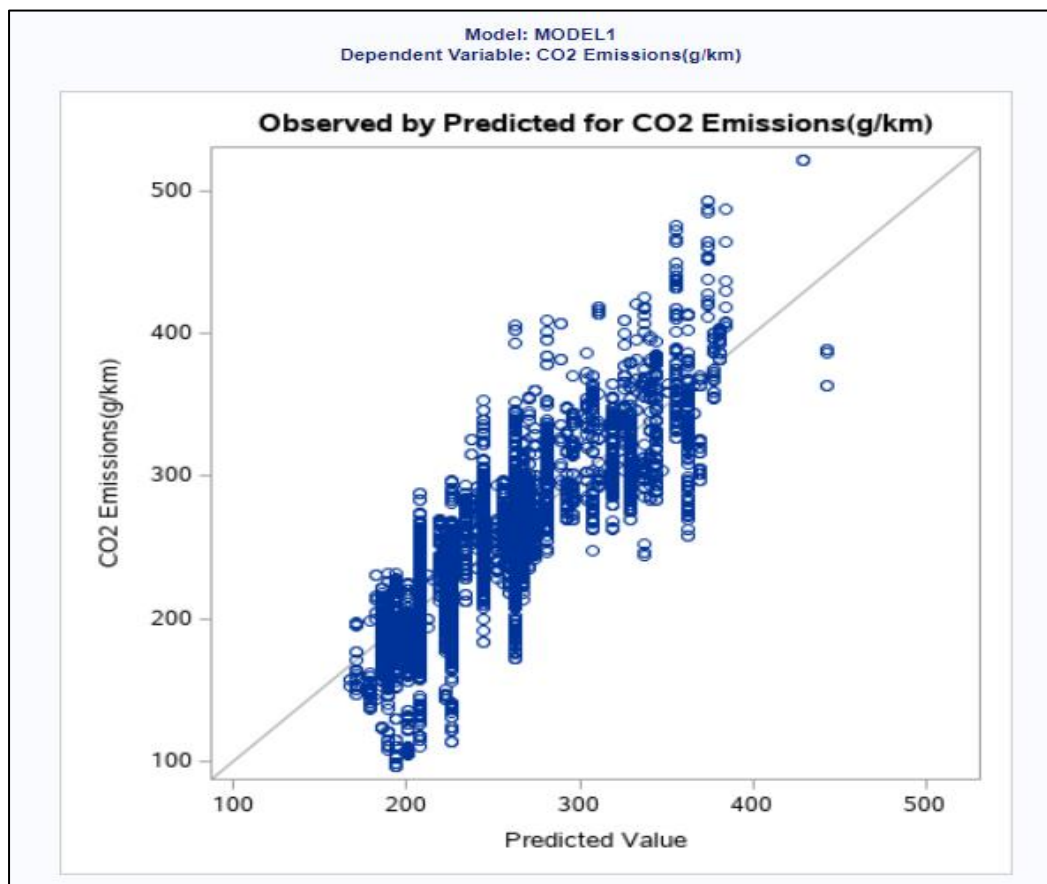
The linear regression model in statistics is used to find the relationship between the continuous (independent or explanatory) and response (dependent) variables. This method analyzes the linear relationship between two quantitative variables (dependent variable and independent variable). There are two types of techniques to identify linear regression. The case with one dependent and one independent variable is known as Simple linear regression. The case with multiple independent variables with a dependent variable is known as Multiple linear regression.

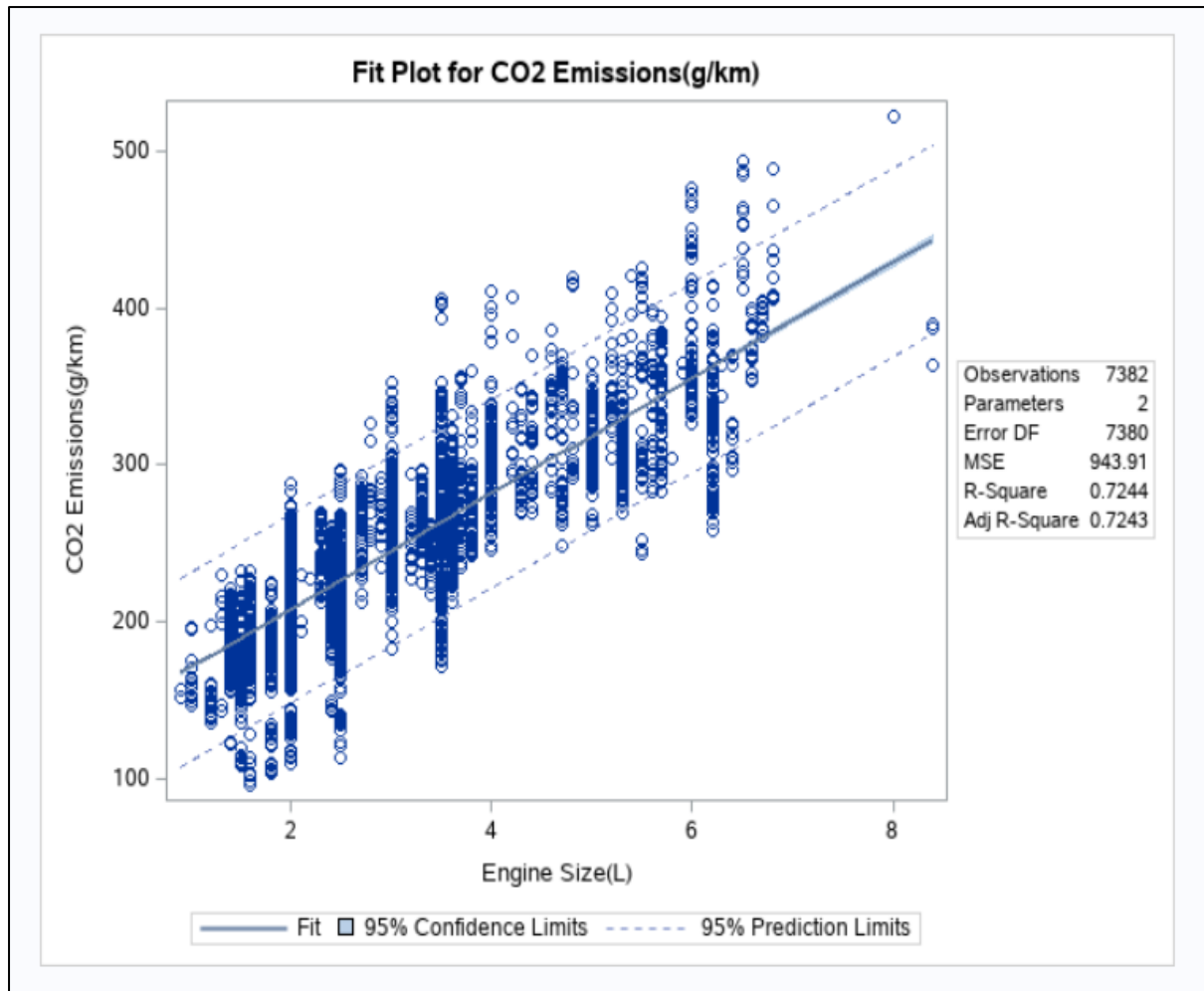
The Simple linear regression equation is as follows:

$$Y = a + b X$$

(Where Y is the dependent variable, X is the explanatory variable, a is the intercept, and b is the slope of the line)

$$Y = 134.36 + 36.77 * \text{Engine Size}$$





The above visualization is a simple linear regression used to understand the relationship between one predictor variable and a response variable. The simple linear regression model uses the engine size in liters to predict the vehicles' overall carbon dioxide (co2) emission. The Parameter Estimate table displays the standard error, t-value, and $\text{Pr} > |t|$ value. The t-value and the associated p-value determine the hypothesis of whether the linear effect on carbon dioxide emissions is zero. In this case, the p-values are $p < 0.001$ and less than 0.05, which means we can conclude that the predictor variables have a statistically significant linear effect on carbon dioxide emission. In this case, 72% of the variation in the predictor variables' performance can explain the carbon dioxide emissions.

CONCLUSION

Carbon dioxide cannot be completely emitted from the environment; however, it can be reduced through using public transport and carpool or bike with friends instead of driving alone. In addition, walk when possible or use ride sharing service, driving more efficiently, and maintain the car to make air less polluted. It is also contributed to one of the main causes of global warming. We as human are responsible for to get a healthy living environment and one of the steps to follow rigorously is to become more carbon dioxide emissions without consideration of impact of post vehicle use. We should always spread the awareness of the carbon footprint to make the people use their vehicle carefully and efficiently.

- The major reason to increases carbon dioxide emissions is by using the luxurious as vast engine size with huge amount of vehicle (manufacturer).
- We should use premium fuel for higher efficiency to reduce the CO₂ emissions
- The lower cylinder vehicles have higher milage per gallon therefore we should use lower cylinder automobiles
- Regular maintenance of car provide better fuel efficiency hence reduces the CO₂ emissions
- Driving on highway provides better fuel efficiency so we should use public transport within the city and drive car on highway at constant speed for better overall milage

H. REFERENCE

- [1] Government of Canada [Greenhouse gas emissions - Canada.ca](#)
- [2] Ian Tiseo (Dec, 2021) [Global CO2 emissions from passenger cars 2020 | Statista](#)
- [3] Hannah Ritchie and Max Roser (May 2020) [CO2 emissions - Our World in Data](#)
- [4] Paul Collins (2021) [Car CO2 emissions: How can you reduce your footprint? \(selectra.com\)](#)
- [5] Data set link - [CO2 Emission by Vehicles | Kaggle](#)
- [6] Fuel Consumption rating [Fuel consumption ratings - Open Government Portal \(canada.ca\)](#)
- [7] Rebecca Bevans(July,2022) [Choosing the Right Statistical Test | Types & Examples \(scribbr.com\)](#)
- [8] Dr. Saul McLeod (2020) [Correlation Definitions, Examples & Interpretation - Simply Psychology](#)