Deliverable 2

Group11:

Nihar Gopidi - 801327998

Akhila Chitturi - 801308961

Akhil Vadlakonda - 801275406

Dinesh Reddy Kankanala - 801274360

Shashank Patlolla – 801328367

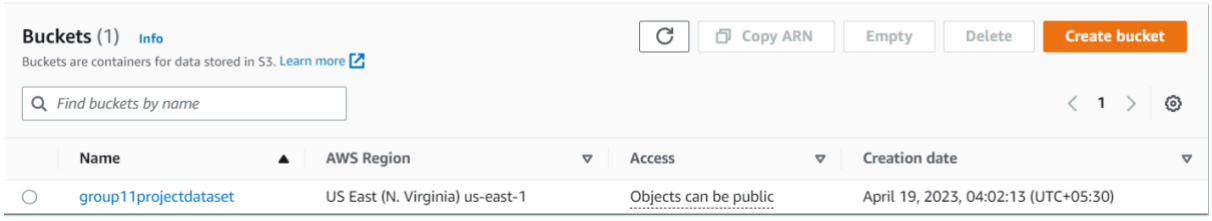
Git Hub link:

https://github.com/akhilachitturi1804/BigDataGroup11

5) Data Understanding:

a) Exploratory Data Analysis:

We uploaded the data to an S3 bucket named it group11project dataset.



b) **Data preparation**

Fetching the dataset with Jupyter Notebook, we used the following code to read the dataset:

```
df_copy=df.copy() # making a copy of orginal dataset
df_copy.head()
```

Python

	brand	name	bodyType	color	fuelType	year	mileage	transmission	power	price	vehicleConfiguration	enginel
0	Toyota	Land Cruiser Prado	jeep 5 doors	blue	Diesel	1995.0	168000.0	AT	130.0	1860000	3.0 SX Wide limited diesel turbo	1
1	Toyota	Land Cruiser	jeep 5 doors	black	Diesel	NaN	260000.0	Automatic	286.0	2300000	NaN	
2	Toyota	Vitz	hatchback 5 doors	blue	Gasoline	2019.0	100000.0	CVT	95.0	1075000	1.3 F Safety Edition III 4WD	1l
3	Toyota	Mark II	sedan	grey	Gasoline	2002.0	239000.0	AT	160.0	480000	2.0 Grande Four	
4	Toyota	RAV4	jeep 5 doors	golden	Gasoline	2010.0	101000.0	AT	170.0	1450000	2.4 AT Long Престиж Плюс	2

```
df.describe()
```

	year	mileage	power	price
count	915699.000000	1.491876e+06	1.484500e+06	1.498740e+06
mean	2005.327732	1.823240e+05	1.588104e+02	1.147137e+06
std	8.206993	1.013326e+05	7.169883e+01	1.370128e+06
min	1953.000000	1.000000e+03	4.500000e+01	6.000000e+03
25%	1999.000000	1.100000e+05	1.070000e+02	3.800000e+05
50%	2006.000000	1.770000e+05	1.400000e+02	7.500000e+05
75%	2012.000000	2.500000e+05	1.850000e+02	1.400000e+06
max	2021.000000	1.000000e+06	6.500000e+02	3.300000e+07

```
df_copy.drop(columns=['year','vehicleConfiguration','engineName','engineDisplacement','link','parse_date',/
                      'date','location'],axis=1,inplace=True)
df_copy
```

Pythor

	brand	name	bodyType	color	fuelType	mileage	transmission	power	price
0	Toyota	Land Cruiser Prado	jeep 5 doors	blue	Diesel	168000.0	AT	130.0	1860000
1	Toyota	Land Cruiser	jeep 5 doors	black	Diesel	260000.0	Automatic	286.0	2300000
2	Toyota	Vitz	hatchback 5 doors	blue	Gasoline	100000.0	CVT	95.0	1075000
3	Toyota	Mark II	sedan	grey	Gasoline	239000.0	AT	160.0	480000
4	Toyota	RAV4	jeep 5 doors	golden	Gasoline	101000.0	AT	170.0	1450000
...	...	...	...	...	...	...	...	...	...
1498735	Toyota	Caldina	station wagon	white	Gasoline	250000.0	AT	260.0	390000
1498736	Honda	HR-V	jeep 3 doors	silver	Gasoline	250000.0	CVT	105.0	370000
1498737	Mazda	CX-7	jeep 5 doors	black	Gasoline	108000.0	AT	244.0	500000
1498738	Mitsubishi	RVR	jeep 5 doors	burgundy	Gasoline	112000.0	CVT	139.0	1100000
1498739	Nissan	Elgrand	minivan	grey	Gasoline	111000.0	AT	240.0	1599999

```
df_copy.fillna(0) #to fill empty values with 0
```

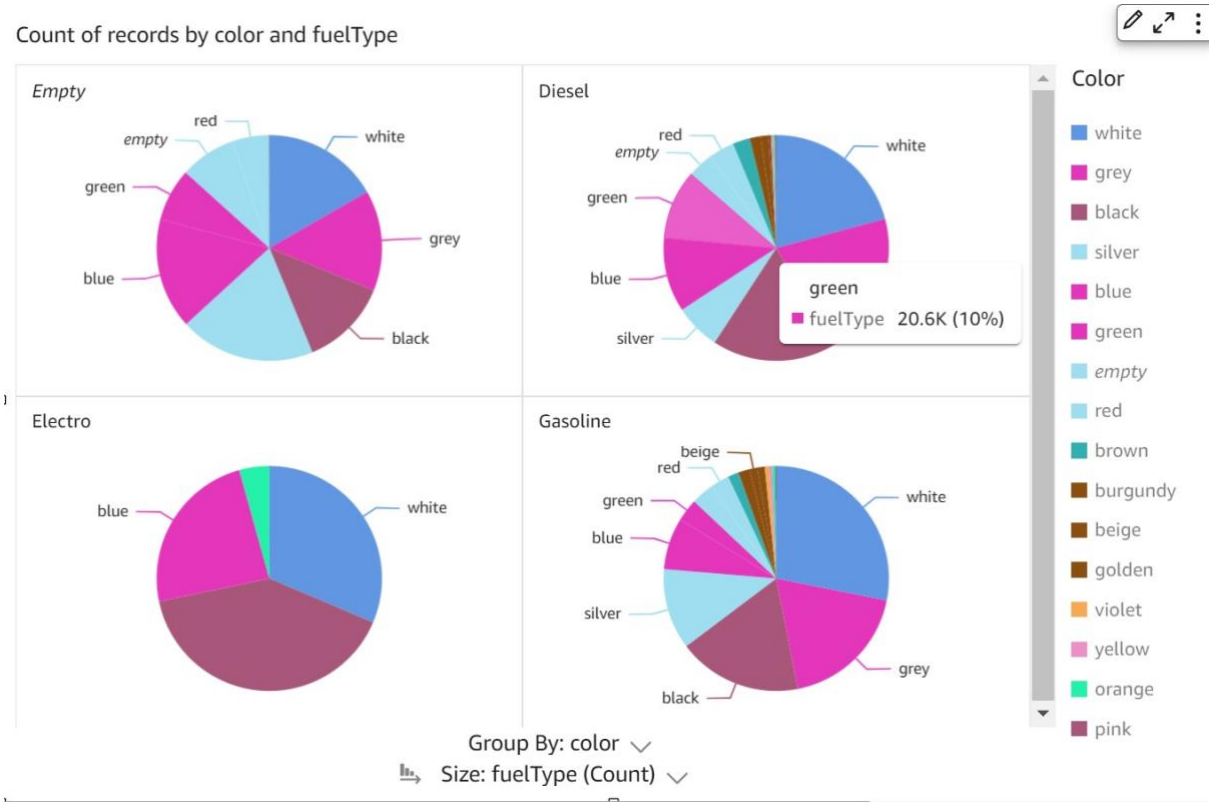
	brand	name	bodyType	color	fuelType	mileage	transmission	power	price
0	Toyota	Land Cruiser Prado	jeep 5 doors	blue	Diesel	168000.0	AT	130.0	1860000
1	Toyota	Land Cruiser	jeep 5 doors	black	Diesel	260000.0	Automatic	286.0	2300000
2	Toyota	Vitz	hatchback 5 doors	blue	Gasoline	100000.0	CVT	95.0	1075000
3	Toyota	Mark II	sedan	grey	Gasoline	239000.0	AT	160.0	480000
4	Toyota	RAV4	jeep 5 doors	golden	Gasoline	101000.0	AT	170.0	1450000
...	...	...	...	...	...	...	...	...	...
1498735	Toyota	Caldina	station wagon	white	Gasoline	250000.0	AT	260.0	390000
1498736	Honda	HR-V	jeep 3 doors	silver	Gasoline	250000.0	CVT	105.0	370000
1498737	Mazda	CX-7	jeep 5 doors	black	Gasoline	108000.0	AT	244.0	500000
1498738	Mitsubishi	RVR	jeep 5 doors	burgundy	Gasoline	112000.0	CVT	139.0	1100000
1498739	Nissan	Elgrand	minivan	grey	Gasoline	111000.0	AT	240.0	1599999

1498740 rows × 9 columns

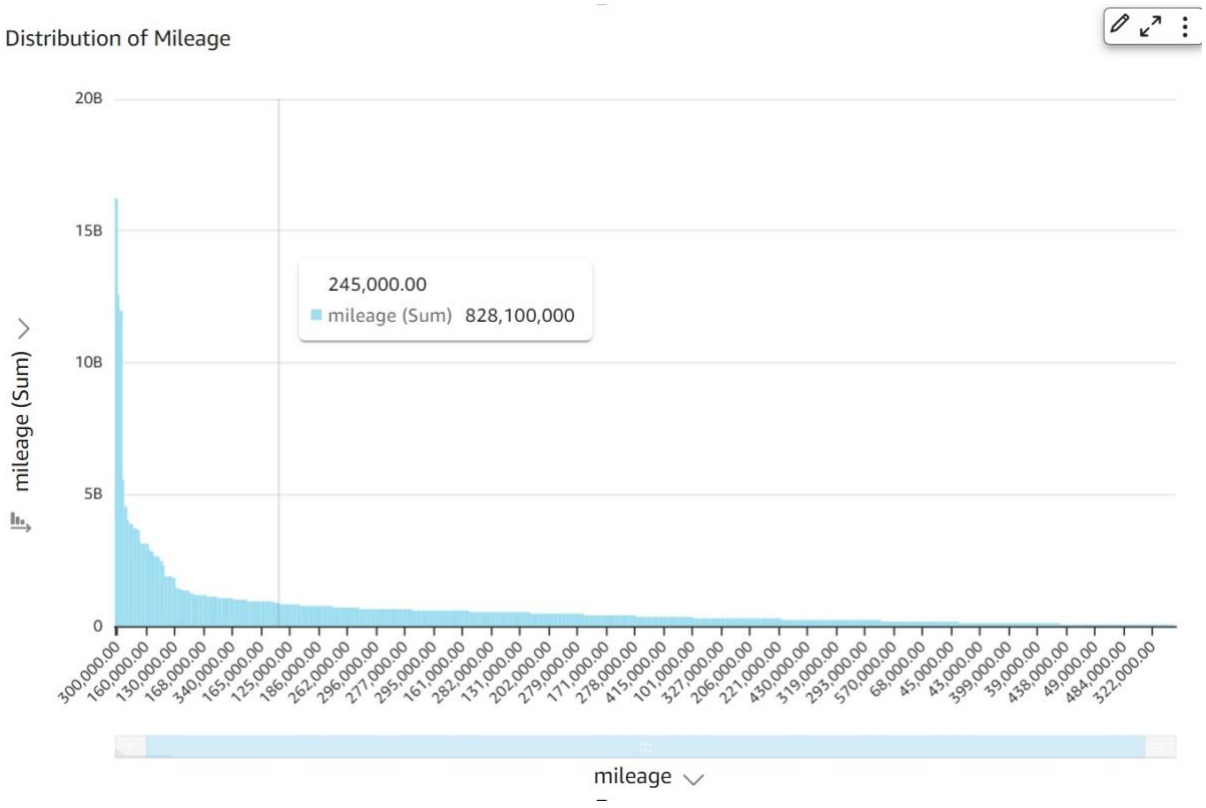
6) Data preparation:

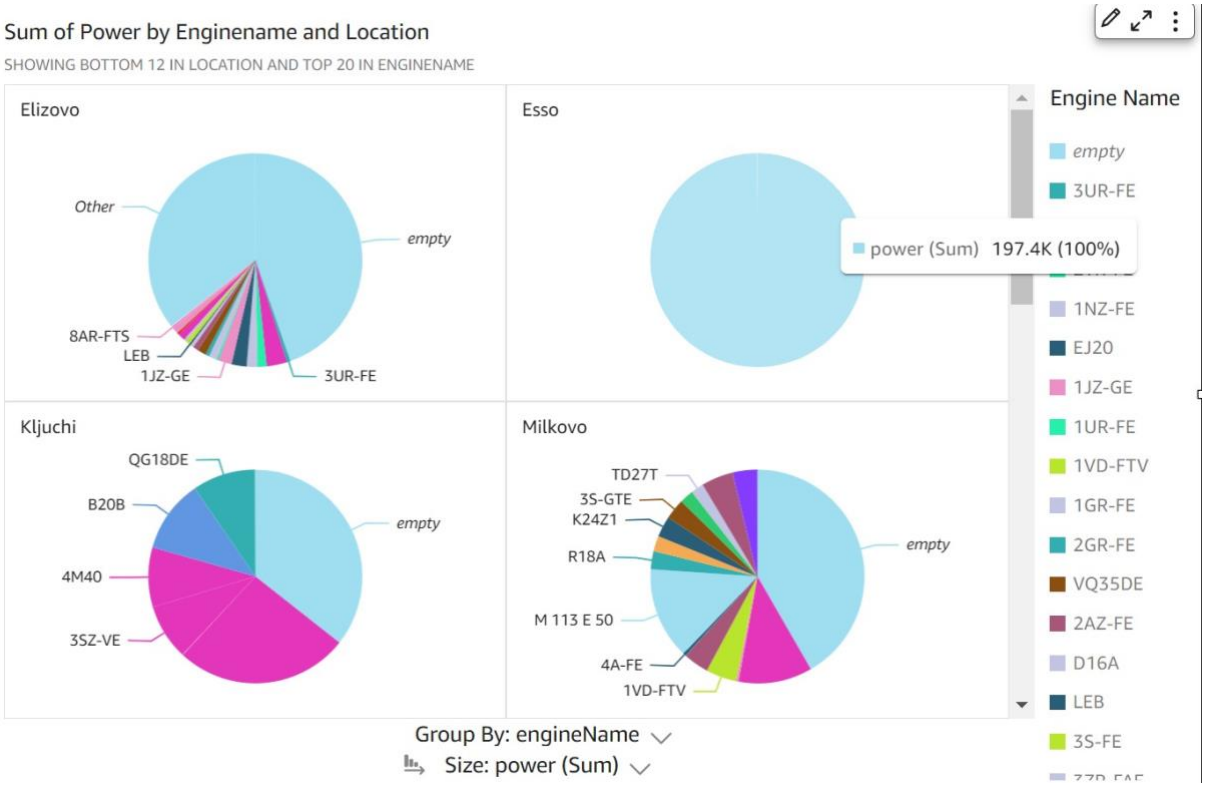
Quick Sight:

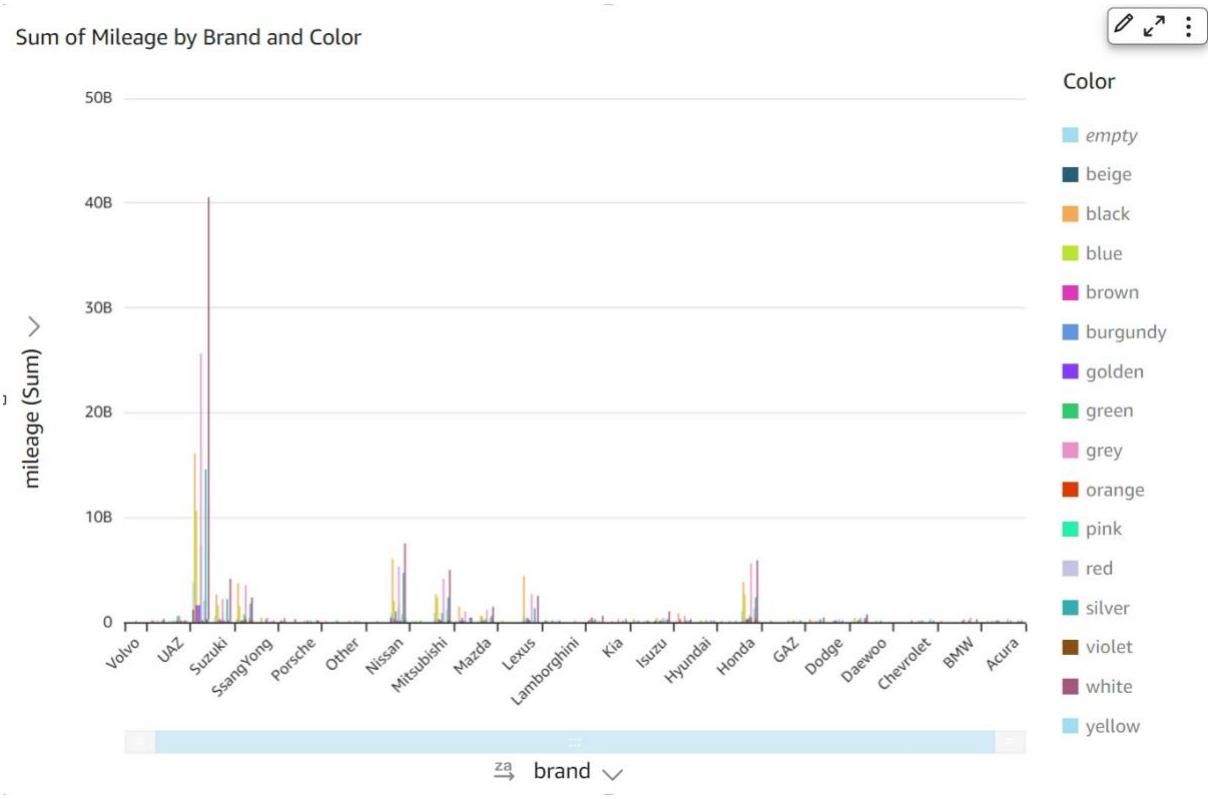
This is a pie chart representation between the car colour and the Fuel used the cars. This classifies and helps us to understand what type of fuel cars and what colour of cars were sold in that respective domain.

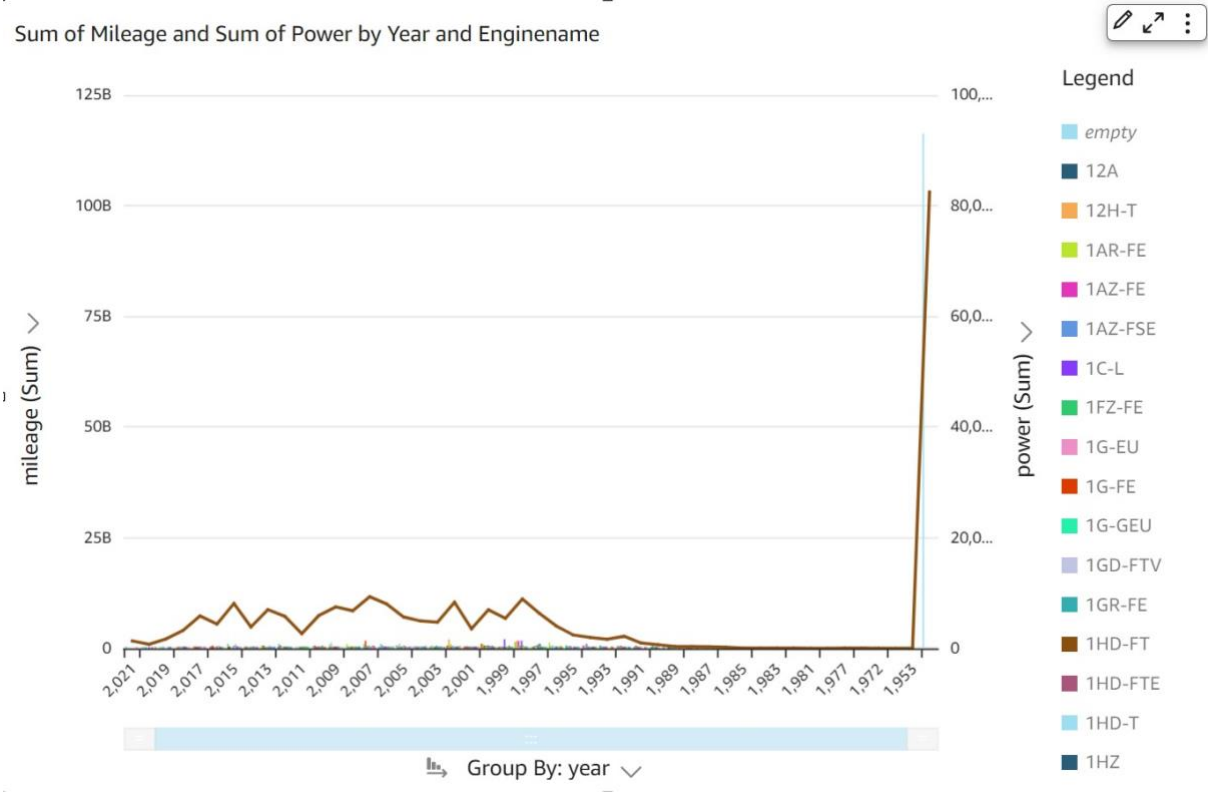


This is a distribution of the mileage of the cars and the sum of total distance travelled by the car.

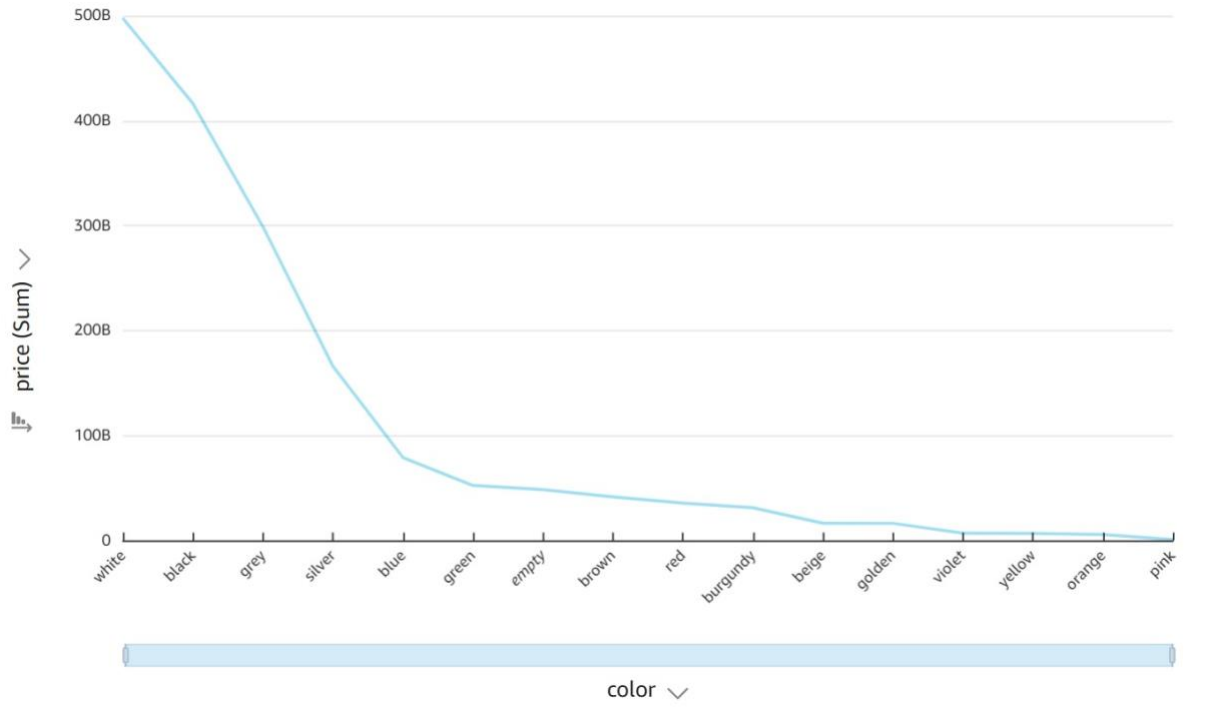




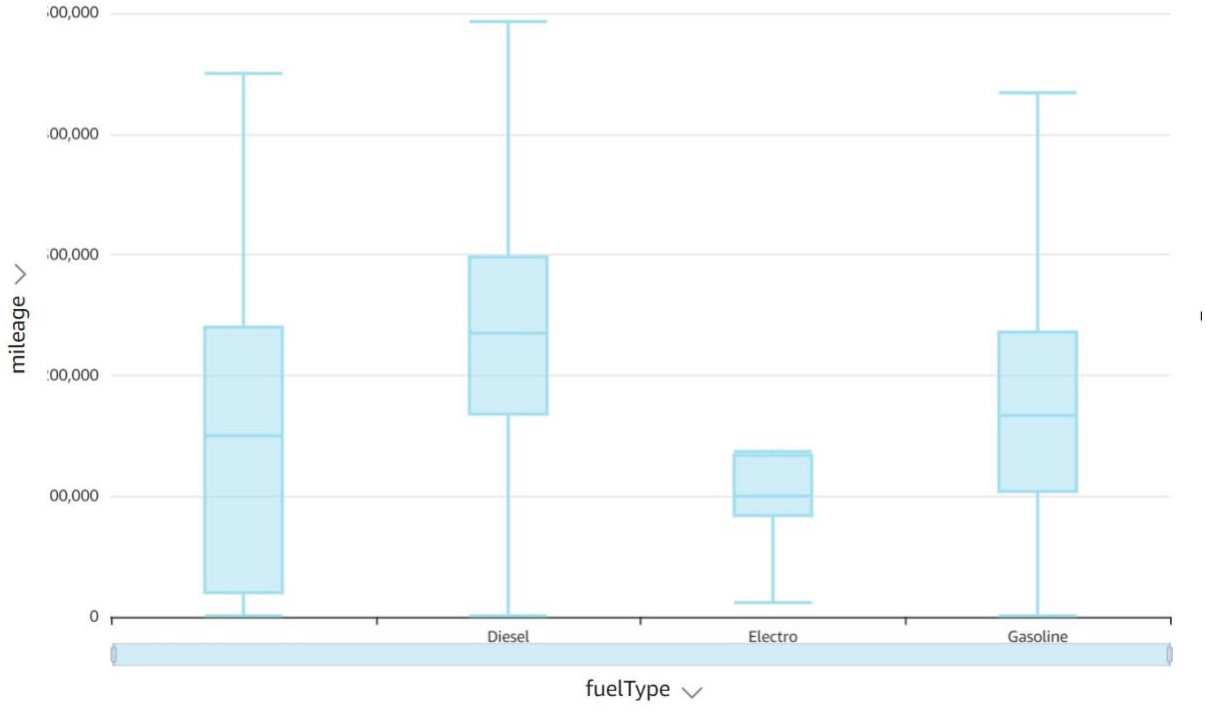

Sum of Mileage by Brand and Color
mileage (Sum)
50B
40B
30B
20B
10B
0
Volvo
UAZ
Suzuki
SsangYong
Porsche
Other
Nissan
Mitsubishi
Mazda
Lexus
Lamborghini
Kia
Isuzu
Hyundai
Honda
GAZ
Dodge
Daewoo
Chevrolet
BMW
Acura
brand
Color
empty
beige
black
blue
brown
burgundy
golden
green
grey
orange
pink
red
silver
violet
white
yellow


Sum of Mileage and Sum of Power by Year and Enginename
mileage (Sum)
125B
100B
75B
50B
25B
0
power (Sum)
100,...
80,0...
60,0...
40,0...
20,0...
0
2,021
2,019
2,017
2,015
2,013
2,011
2,009
2,007
2,005
2,003
2,001
1,999
1,997
1,995
1,993
1,991
1,989
1,987
1,985
1,983
1,981
1,977
1,972
1,953
Group By: year
Legend
empty
12A
12H-T
1AR-FE
1AZ-FE
1AZ-FSE
1C-L
1FZ-FE
1G-EU
1G-FE
1G-GEU
1GD-FTV
1GR-FE
1HD-FT
1HD-FTE
1HD-T
1HZ

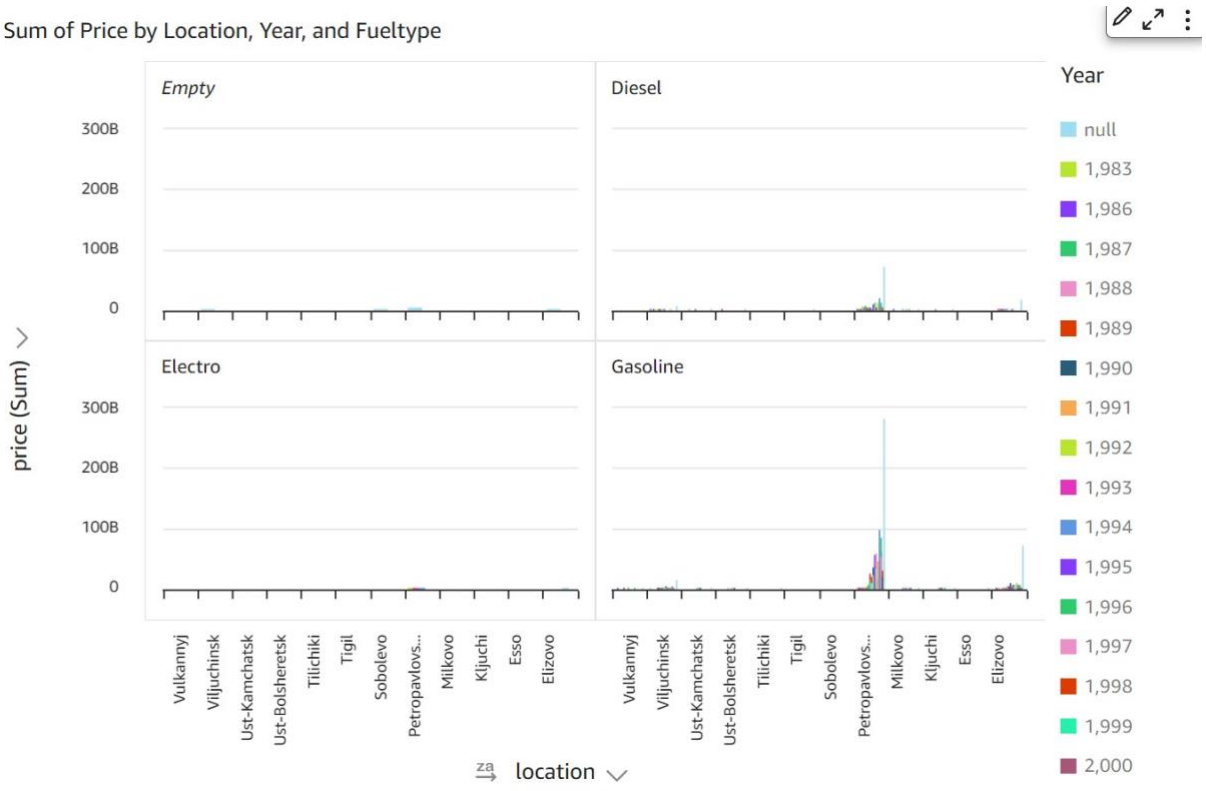
Sum of Price by Color



Mileage by Fueltype



Sum of Price by Location, Year, and Fueltype



Sum of Power by Location and Year