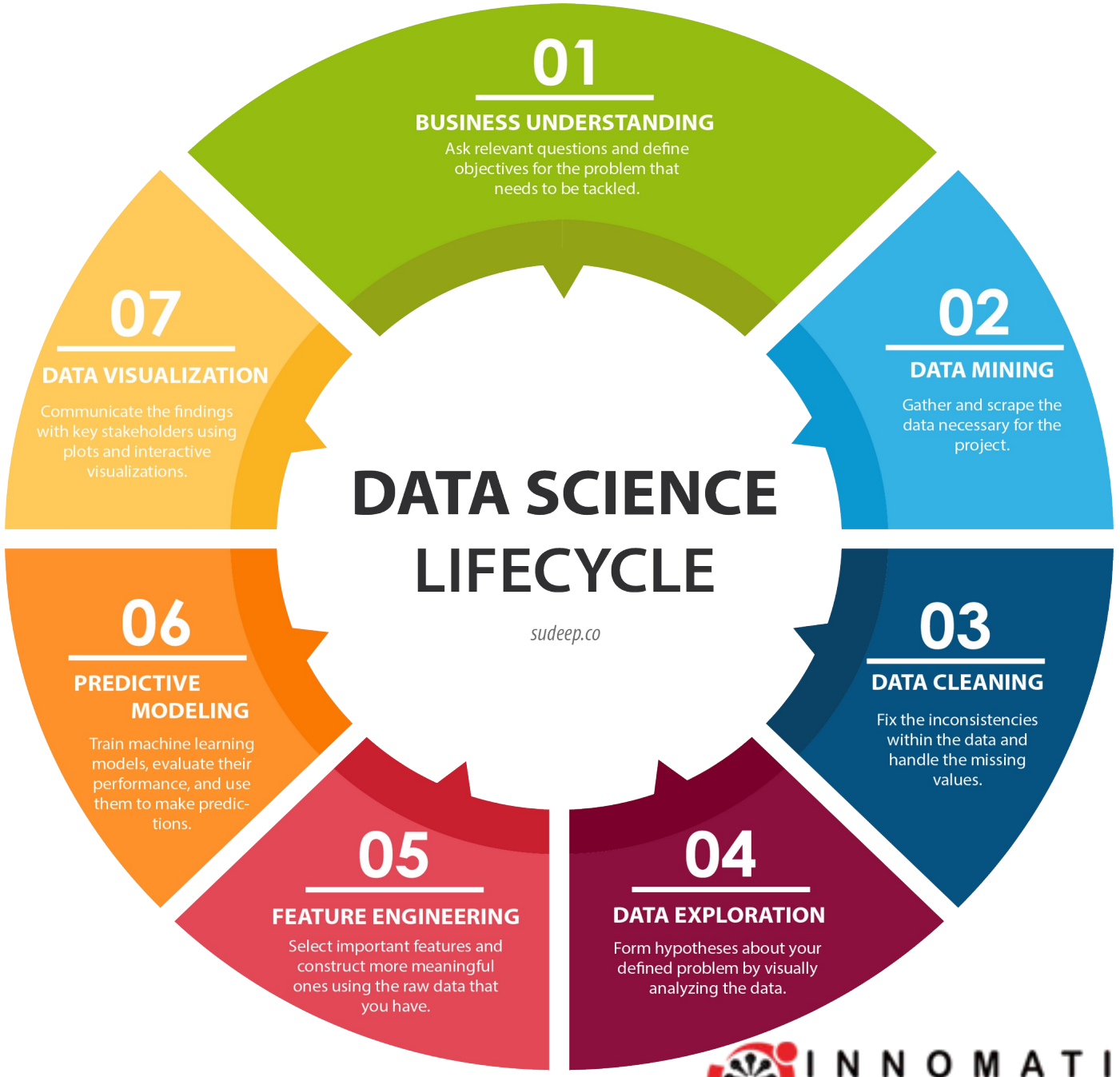
INNOMATICS
RESEARCH LABS

Name: Anuja S. Raktate
Email: anujaraktate2000@gmail.com
Education: Msc(Mathematics)(2022)

INNOMATICS
RESEARCH LABS

Web Scraping And EDA Project



INNOMATICS RESEARCH LABS

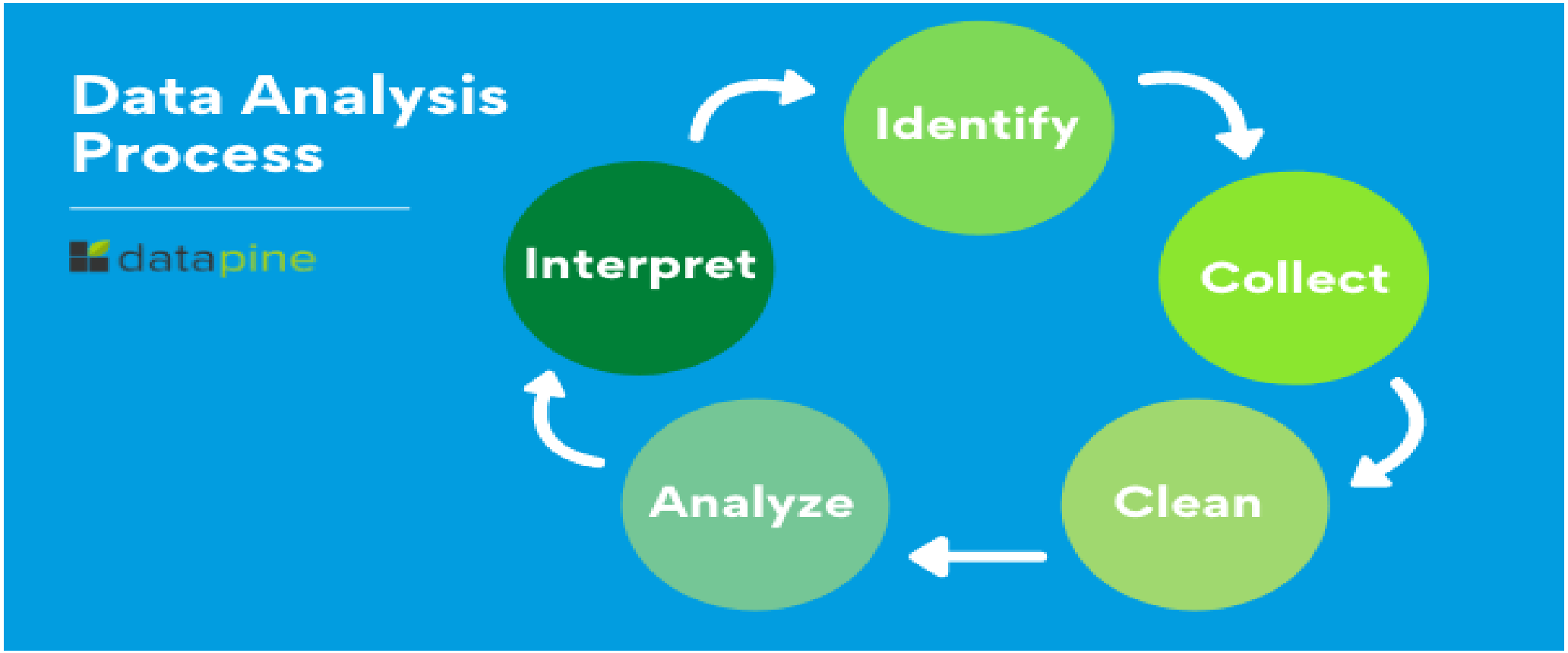
Why we want to learn Data Science?

- With Data Science, one can analyze massive graphical data, temporal data, and geospatial data to draw insights. It also helps in seismic interpretation and reservoir characterization.
- Learning about data science provides an opportunity for you to recreate yourself

What is Web Scrapping?

- Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications.
- Web Scrapping used for: 1. Price Monitoring 2. Market Research 3. News Monitoring

THE DATA ANALYSIS PROCESS



FIRST HAND / SECOND HAND CARS

SUB TITLE :-

- URL
- Problem Statement
- Extract the Data
- Data Frame
- Export into .csv format
- Read CSV File
- Clean the Data
- Data Analysis and Visualization (EDA)
  1) Uni-variate Analysis
  2) Bi-Variate Analysis/Multivariate
- Conclusion

INNOMATICS RESEARCH LABS

DATA COLLECTION

- "What is data?" The abridged answer is, data is various kinds of information formatted in a particular way. Therefore, data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities
- Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences.

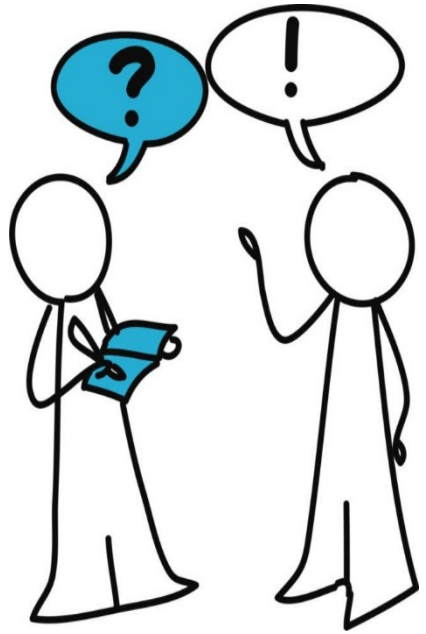
URL=https://www.cartrade.com/buy-used-cars/?gcc=1#city=10&sc=-1&so=-1&pn=1

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" itemscope itemtype="http://schema.org/WebPage" lang="en">
▶<head prefix="og: https://ogp.me/ns# fb: https://ogp.me/ns/fb#">…</head>
▼<body id="idbybody">
  <script type="text/javascript" src="https://stc.aeplcdn.com/staticminv2/javascript/usedcar/header-6521796bd2.js" defer crossorigin="anonymous"></script>
 ▶<style>…</style>
 ▶<style>…</style>
 ▶<div class="ct_hdr_red" id="ct_menunew" itemscope itemtype="https://schema.org/Organization">…</div>
 ▶<script type="text/javascript" id>…</script>
 ▶<noscript>…</noscript>
 ▶<script type="text/javascript" id>…</script>
 ▶<script type="text/javascript" id>…</script>
 ▶<noscript>…</noscript>
 ▶<script type="text/javascript" id>…</script>
  <style> .chattopicon.close { display: none; } </style>
  <link href="https://imgd-ct.aeplcdn.com" rel="preconnect dns-prefetch" crossorigin>
  <link rel="stylesheet" href="https://stc.aeplcdn.com/staticminv2/css/usedcar/desktop/listingtemplate-3d6a6c8d0a.css" type="text/css">
 ▶<script type="text/javascript" id>…</script>
 ▶<noscript>…</noscript>
 ▶<script type="text/javascript" id>…</script>
 ▶<script type="text/javascript" id>…</script>
 ▶<noscript>…</noscript>
 ▶<script type="text/javascript" id>…</script>
  <link rel="stylesheet" href="https://stc.aeplcdn.com/staticminv2/css/usedcar/desktop/absure-banner-792448a5ef.css" type="text/css">
  <link rel="stylesheet" href="https://stc.aeplcdn.com/staticminv2/css/usedcar/desktop/faq-b2767d2eda.css" type="text/css">
  <link rel="stylesheet" href="https://stc.aeplcdn.com/staticminv2/css/usedcar/desktop/nearbycities-138fb8eab9.css" type="text/css">
  <script type="text/javascript" src="https://stc.aeplcdn.com/staticminv2/javascript/usedcar/desktop/jcarouselmin-7119647992.js" defer crossorigin="anonymous">
  </script>
  <script type="text/javascript" src="https://stc.aeplcdn.com/staticminv2/javascript/usedcar/desktop/jcarousel-e56e1520a4.js" defer crossorigin="anonymous">
  </script>
```

INNOMATICS RESEARCH LABS

- # Problem Statement:-

1) Analayzing Selling Price of used Cars.
2) Comparison Of Cost Of Cars Vs EMI Of Car Price.
3) Analayzing Average Car Prize For Data Benchmarking.
4) Extra Features Of Car That Improve Scalability.
5) Specific Models Of Cars In High Demand.



Scraped the Raw Data

```
df
```

	CarNameWithModelYear	Cost	EMI	Features
0	\n\t\t\t\t\t\t\t\t\t\t\t2013 Maruti Suzuki A...	₹2,20,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\nEMI starts at ₹3,653\n	63,484 KMs|Petrol|Noida\n
1	\n\t\t\t\t\t\t\t\t\t\t\t2014 Maruti Suzuki A...	₹2,65,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\n	42,458 KMs|Petrol|Bhopal\n
2	\n\t\t\t\t\t\t\t\t\t\t\t2020 Volkswagen Polo...	₹6,50,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\nEMI starts at ₹10,794\n	64,293 KMs|Petrol|Aurangabad\n
3	\n\t\t\t\t\t\t\t\t\t\t\t2020 BMW X1 sDrive20...	₹37,00,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n...	\nEMI starts at ₹61,444\n	25,000 KMs|Diesel|Pune\n
4	\n\t\t\t\t\t\t\t\t\t\t\t2018 BMW X1 sDrive20...	₹25,00,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n...	\nEMI starts at ₹41,516\n	69,000 KMs|Diesel|Delhi\n
...	...	...	...	...
442	\n\t\t\t\t\t\t\t\t\t\t\t2021 Mahindra Marazz...	₹13,10,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n...	\nEMI starts at ₹21,754\n	28,495 KMs|Diesel|Mumbai\n
443	\n\t\t\t\t\t\t\t\t\t\t\t2017 Maruti Suzuki V...	₹6,85,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\n	61,472 KMs|Diesel|Delhi\n
444	\n\t\t\t\t\t\t\t\t\t\t\t2017 Honda WR-V S MT...	₹6,50,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\nEMI starts at ₹10,794\n	39,693 KMs|Petrol|Chennai\n
445	\n\t\t\t\t\t\t\t\t\t\t\t2018 Volkswagen Vent...	₹8,60,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n\...	\nEMI starts at ₹14,281\n	54,314 KMs|Petrol|Pune\n
446	\n\t\t\t\t\t\t\t\t\t\t\t2019 Volkswagen Vent...	₹10,99,000\n\t\t\t\t\t\t\t\t\t\t\t\tMake\n...	\nEMI starts at ₹18,250\n	38,754 KMs|Petrol|Hyderabad\n

447 rows × 4 columns

INNOMATICS RESEARCH LABS

Data Cleaning

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled

Steps involved in the process of cleaning the Data

1) Remove all special characters

2)Extracting required words from the data

```
import re
regex = r'[\d]+,+[\d]+[\s]+KMs'
df['KMs']=df['Features'].apply(lambda x:re.compile(regex).search(x).group())
```

```
import numpy as np
df['Fuel Type']=np.arange(0,len(df['Features']))
for i in range(0,len(df['Features'])):
    a=df['Features'][i].split('|')
    df['Fuel Type'][i]=a[1]
```

```
import re
regex = r'\n+[\d]+\s'
df['Car Model Year']=df['CarNameWithModelYear'].apply(lambda x:re.compile(regex).search(x).group())
```

```
import re
regex = r'\s+[\w]+\s+[\w]+'
df['Brand Name']=df['CarNameWithModelYear'].apply(lambda x:re.compile(regex).search(x).group())
```

```
import numpy as np
df['City']=np.arange(0,len(df['Features']))
for i in range(0,len(df['Features'])):
    a=df['Features'][i].split('|')
    df['City'][i]=a[2]
```

3) Cleaning / Filling Missing Data Replace Nan with a Scalar Value

```
df["EMI"]=df["EMI"].replace({'':0},regex=True)
```

4) Drop Duplicate Values

```
df.drop_duplicates(keep='first')
```

```
df1=df.drop(['Unnamed: 0','Features'], axis=1)
```

5) Check for Missing Values

```
print(" \nCount total NaN at each column in a DataFrame : \n\n",
      df1.isnull().sum())
```

```
Count total NaN at each column in a DataFrame :

 Unnamed: 0        0
Car Name           0
Cost               0
EMI                0
Features           0
KMs                0
Fuel Type          0
Car Model Year     0
Brand Name         0
City               0
dtype: int64
```

6)Converting Datatypes into Required Types

```
df['Cost']=df['Cost'].astype('float32')
```

```
for i in range(0,len(df['Cost'])):
    if df['Cost'][i]== '1550000\n 15.9 Lakh':
        df['Cost'][i]= 15.9*10**5
```

```
df["EMI"]=df["EMI"].astype(str)
```

```
def convert_l(x):
    for i in range(0,len(x)):
        a=x[i]
        print(a)
        if a!=0:

            for j in range(0,len(a)):
                if a[j]=='L':
                    b=a[0:j-1]
                    print(b)
                    b=float(b)
                    b=b*10**5
                    df['EMI'][i]=b
```

```
convert_l(df['EMI'])
```



```
df['EMI']=df['EMI'].astype(int)
```

```
df['KMs']=df['KMs'].astype(int)
```

```
df['Car Model Year']=df['Car Model Year'].astype(int)
```

Final Data(Cleaned Data)

	Car Name	Cost	EMI	KMs	Fuel Type	Car Model Year	Brand Name	City
0	013 Maruti Suzuki Alto K10 VXi\n	220000.0	3653	63484	Petrol	2013	Maruti Suzuki	Noida
1	014 Maruti Suzuki Alto 800 Lxi\n	265000.0	0	42458	Petrol	2014	Maruti Suzuki	Bhopal
2	020 Volkswagen Polo Comfortline Plus 1.0L MPI\n	650000.0	10794	64293	Petrol	2020	Volkswagen Polo	Aurangabad
3	020 BMW X1 sDrive20d xLine\n	3700000.0	61444	25000	Diesel	2020	BMW X1	Pune
4	018 BMW X1 sDrive20d xLine\n	2500000.0	41516	69000	Diesel	2018	BMW X1	Delhi
...	...	...	...	...	...	...	...	...
442	021 Mahindra Marazzo M6 Plus 7 STR\n	1310000.0	21754	28495	Diesel	2021	Mahindra Marazzo	Mumbai
443	017 Maruti Suzuki Vitara Brezza ZDi\n	685000.0	0	61472	Diesel	2017	Maruti Suzuki	Delhi
444	017 Honda WR-V S MT Petrol\n	650000.0	10794	39693	Petrol	2017	Honda WR	Chennai
445	018 Volkswagen Vento Highline Plus 1.2 (P) AT ...	860000.0	14281	54314	Petrol	2018	Volkswagen Vento	Pune
446	019 Volkswagen Vento Highline Plus 1.2 (P) AT ...	1099000.0	18250	38754	Petrol	2019	Volkswagen Vento	Hyderabad

447 rows × 8 columns

INNOMATICS RESEARCH LABS

Exploratory Data Analysis

- EDA is **applied to investigate the data and summarize the key insights**.
- Statistics, exploratory data analysis (EDA) is **an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods**.
- The primary goal of EDA is **to maximize the analyst's insight into a data set and into the underlying structure of a data set**, while providing all of the specific items that an analyst would want to extract from a data set, such as: a good-fitting, parsimonious model.

INNOMATICS RESEARCH LABS

Univariate Analysis

Continuous variable

Central Tendency

- *Mean*
- *Median*
- *Mode*

```
df1.mean()
```

```
Unnamed: 0        2.230000e+02
Cost              1.699385e+06
EMI               2.460055e+04
KMs               3.648456e+04
Car Model Year    2.017767e+03
dtype: float64
```

```
df1.median()
```

```
Unnamed: 0         223.0
Cost            890000.0
EMI              14115.0
KMs              34216.0
Car Model Year    2018.0
dtype: float64
```

```
df1.mode()
```

	Unnamed: 0	Car Name	Cost	EMI	Features	KMs	Fuel Type	Car Model Year	Brand Name	City
0	0	016 Hyundai Creta SX Plus 1.6 Petrol\n	650000.0	0.0	1,815 KMs|Petrol|Mumbai	1815.0	Petrol	2019.0	Maruti Suzuki	Mumbai
1	1	019 BMW X1 sDrive20d xLine\n	NaN	NaN	11,802 KMs|Petrol|Kanpur	2308.0	NaN	NaN	NaN	NaN
2	2	019 Maruti Suzuki XL6 Zeta AT Petrol\n	NaN	NaN	12,626 KMs|Petrol|Mumbai	11802.0	NaN	NaN	NaN	NaN
3	3	NaN	NaN	NaN	14,676 KMs|Petrol|Mumbai	12626.0	NaN	NaN	NaN	NaN
4	4	NaN	NaN	NaN	18,530 KMs|Diesel|Mumbai	14676.0	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
442	442	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Measures of Dispersion

Standard Deviation

```
df1.std()
```

```
Unnamed: 0        1.291820e+02
Cost              1.841582e+06
EMI               2.865791e+04
KMs               2.091222e+04
Car Model Year    2.421782e+00
dtype: float64
```

Variance

```
df1.var()
```

```
Unnamed: 0        1.668800e+04
Cost              3.391422e+12
EMI               8.212760e+08
KMs               4.373211e+08
Car Model Year    5.865029e+00
dtype: float64
```

Inter-Quartile Range

```
from scipy.stats import iqr
iqr(df1['Cost'])
```

```
1087500.0
```

```
from scipy.stats import iqr
iqr(df1['EMI'])
```

```
13990.5
```

```
from scipy.stats import iqr
iqr(df1['KMs'])
```

```
32657.0
```

```
from scipy.stats import iqr
iqr(df1['Car Model Year'])
```

```
2.0
```

Skewness

```
df1.skew()
```

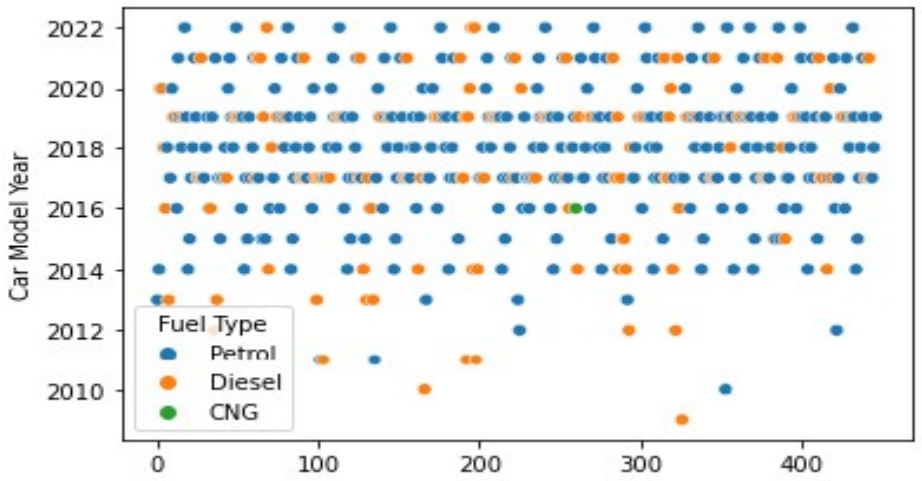
```
Unnamed: 0       0.000000
Cost             1.953329
EMI              1.639939
KMs              0.265595
Car Model Year  -0.587099
dtype: float64
```

Describe()

```
df1.describe()
```

	Cost	EMI	KMs	Car Model Year
count	447.00	447.000000	447.000000	447.000000
mean	1699384.75	24600.550336	36484.559284	2017.767338
std	1841581.50	28657.914236	20912.224851	2.421782
min	206699.00	0.000000	1000.000000	2009.000000
25%	537500.00	7763.500000	22673.000000	2017.000000
50%	890000.00	14115.000000	34216.000000	2018.000000
75%	1625000.00	21754.000000	55330.000000	2019.000000
max	13600000.00	111000.000000	93207.000000	2022.000000

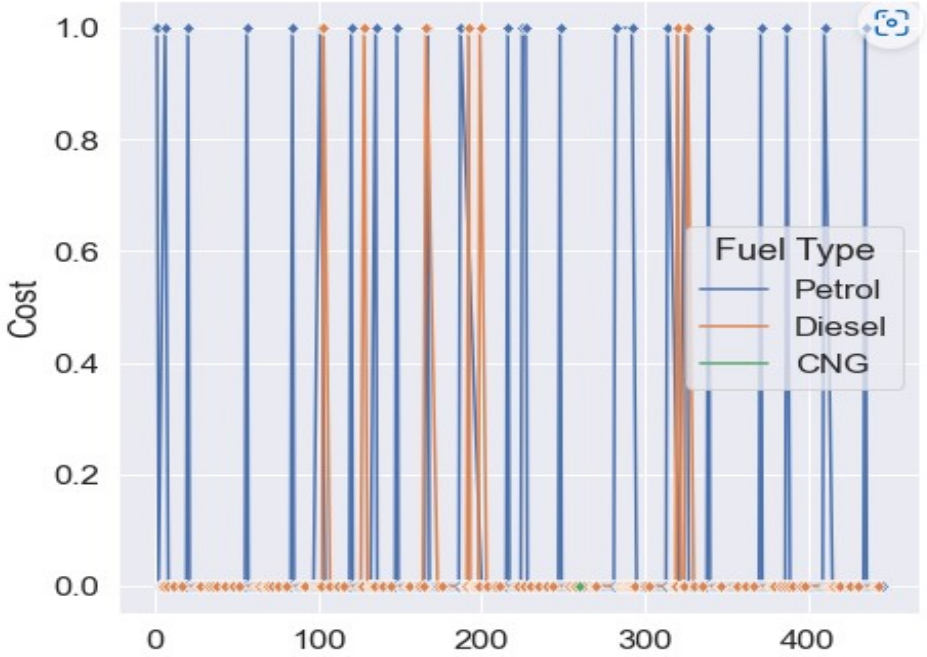
SCATTER PLOT :



.

•In our plot here the blue colors dot show the number of Petrol cars, the orange colour shows number of Disel cars and the green colour shows number of CNG cars used from this we can observe that in our data in year 2010-2023 there is large amount of Petrol cars were used than Disel cars and very few amount of CNG cars are there.

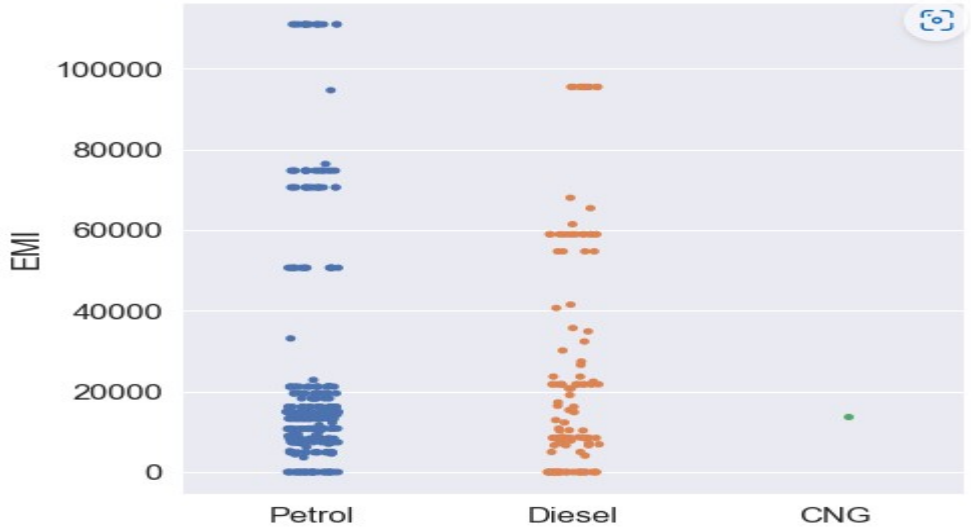
LINE PLOT (with markers) :



•The Line Plot of 'Cost' it has two important factor cost and Index and here from the plot based on our dataset we observe that there are maximum numbers of petrol cars are present which having cost<400000 and there are very few numbers of Disel cars are present which having cost<400000 i.e Cost of most of the disel cars is >400000 . and there is no any CNG Cars which satisfies our condition i.e cost of cars<400000 , that means There is no any CNG cars present under the cost 4lakh.that's the informataion of Cost of cars by using its Fuel type shown here.

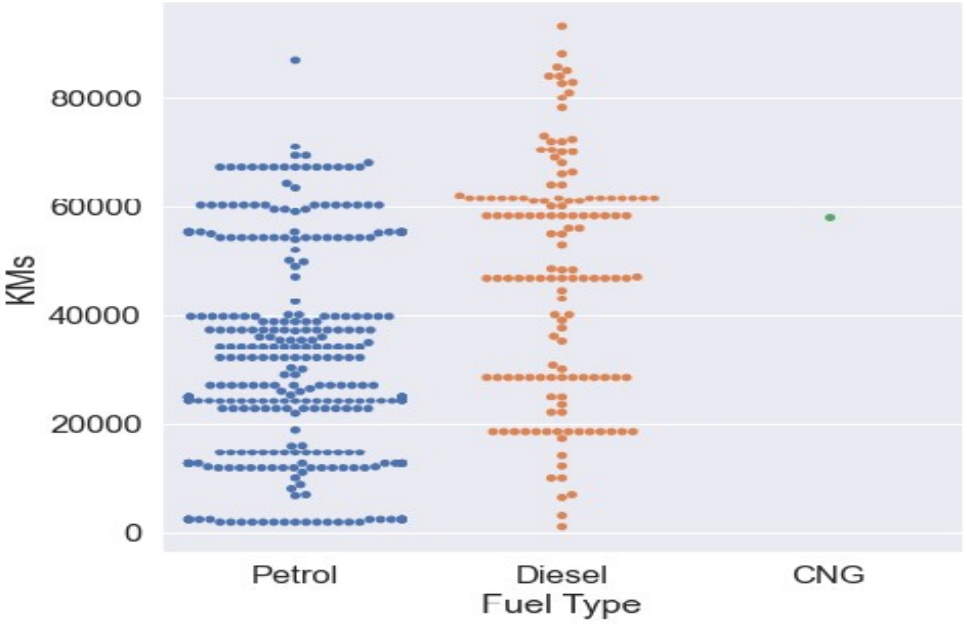
INNOMATICS
RESEARCH LABS

STRIP PLOT :



•so, basically EMI is mostly comes to range between 0 to 20000. And here the petrol & Disel are both the fuel types but mostly the cars having fuel type as petrol are used most and has EMI is also at minimum range of 0-20000 as compare to Petrol Cars, Disel cars having also EMI at range 0-20000 but the use of disel cars is less than petrol. and there is very less amount of CNG cars present in our data set having EMI between the same range. and the cars of EMI cost at Rs.40000 & above there is minumum amount cars present having fuel type petrol and disel.

SWARM PLOT :



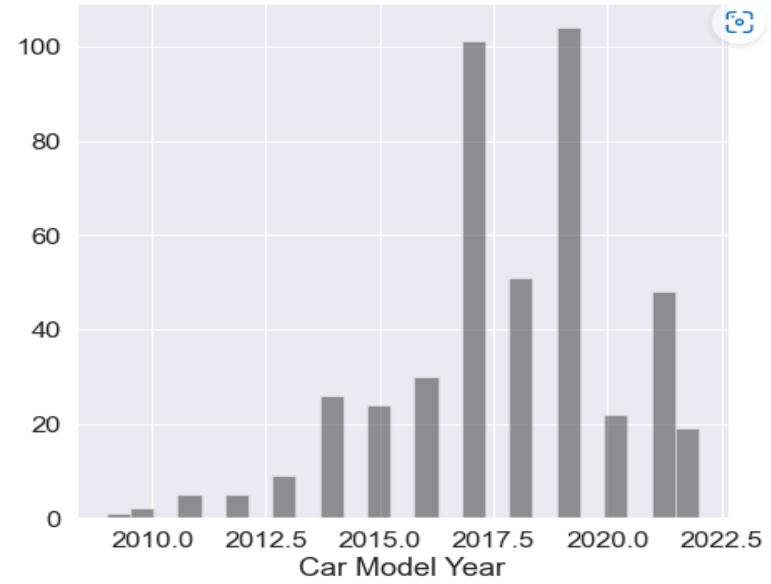
• so, basically KMs is mostly comes to range between 0 to 70000. And here the Petrol & Disel are both the Fuel Types Cars used frequently but mostly the cars having Fuel Type as Petrol are used most.

•As compare to petrol Cars, Disel cars having also KMs at range 0-70000 but the use of disel cars is less than petrol. and there is very less amount of CNG cars present in our data set having KMs range is 60000 aprox.

•And the cars having KMs70000 & above there is very few amount cars present having fuel type petrol and disel.

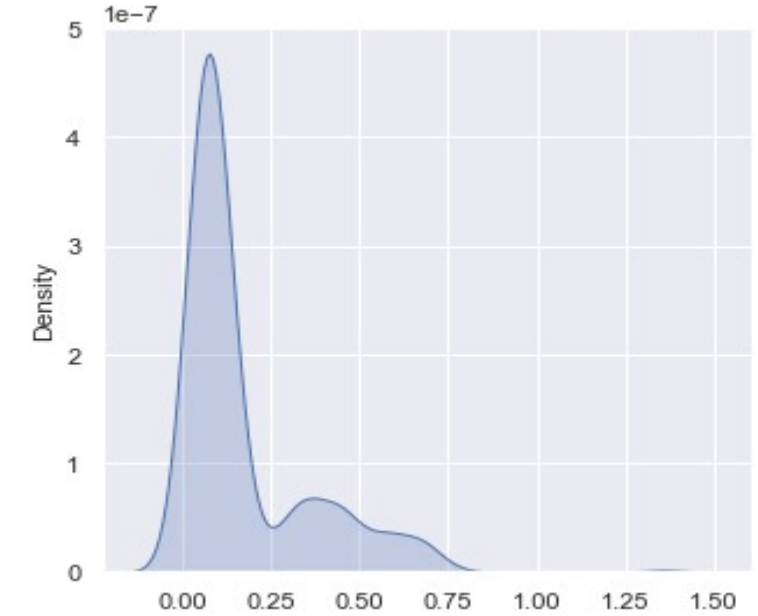
INNOMATICS RESEARCH LABS

HISTOGRAMS :



- So The most of the car's having model year in between 2017-2020 and it's percentage is very high like 80% & above. and Car's of model Year 2010-2016 is present at minimum Percentage structure of car model year.
- Also The car's of model year between 2010-2013 is present in very less amount in our d ataset and 2017-2019 is the very high percentage of the car model year. But 2019 is the very high car model year in percentage wise as compare all of this year.

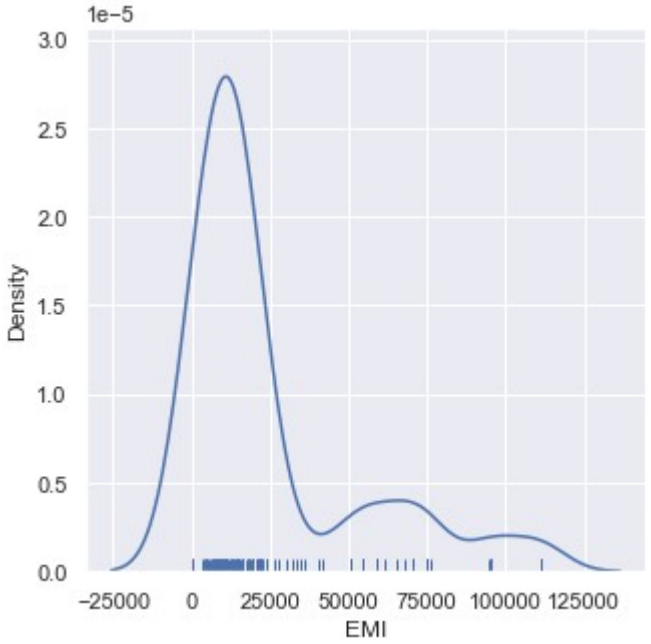
DENSITY PLOT :



- Here cost of some cars is very high which is shown in graph as density of cost i.e it is grater than 4 as density wise. and also the cost of 0.50% cars i.e between 0-1 that having cost in between 0.25 to 0.75. and cost of remaning cars present in our dataset is very low as compare to others.

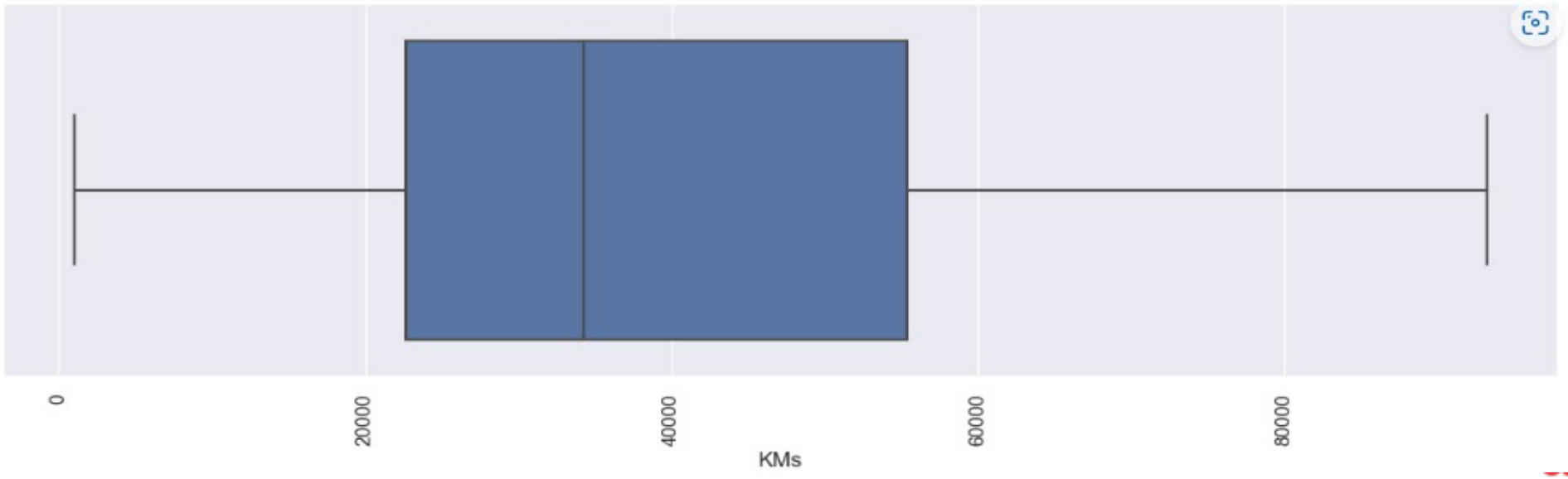
INNOMATICS RESEARCH LABS

RUG PLOTS :



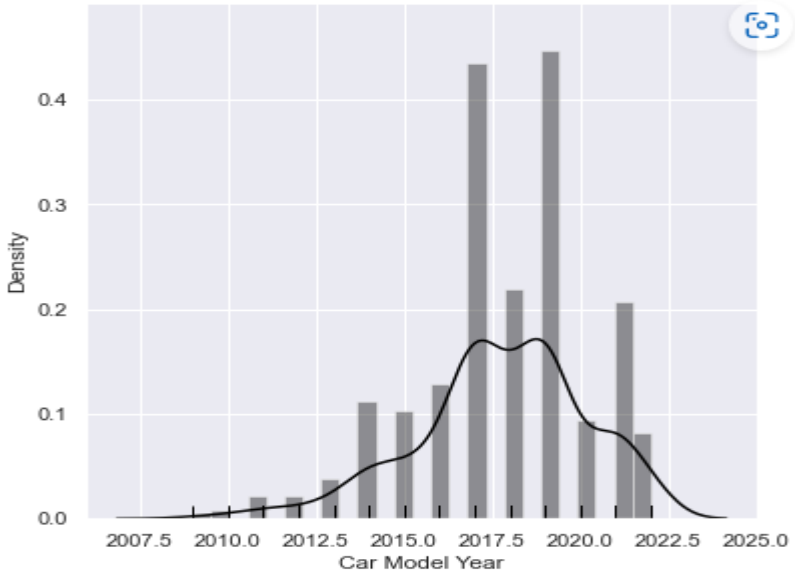
•Here The density of cars which having minumum amount of EMI upto Rs.25000 is very high which is between 2.5-3.0 of the density as shown in graph. and their are very few cars like density in between 0.0-0.5 which has very high amount of EMI cost between Rs.40000-Rs.140000 are shown here.

BOX PLOTS :



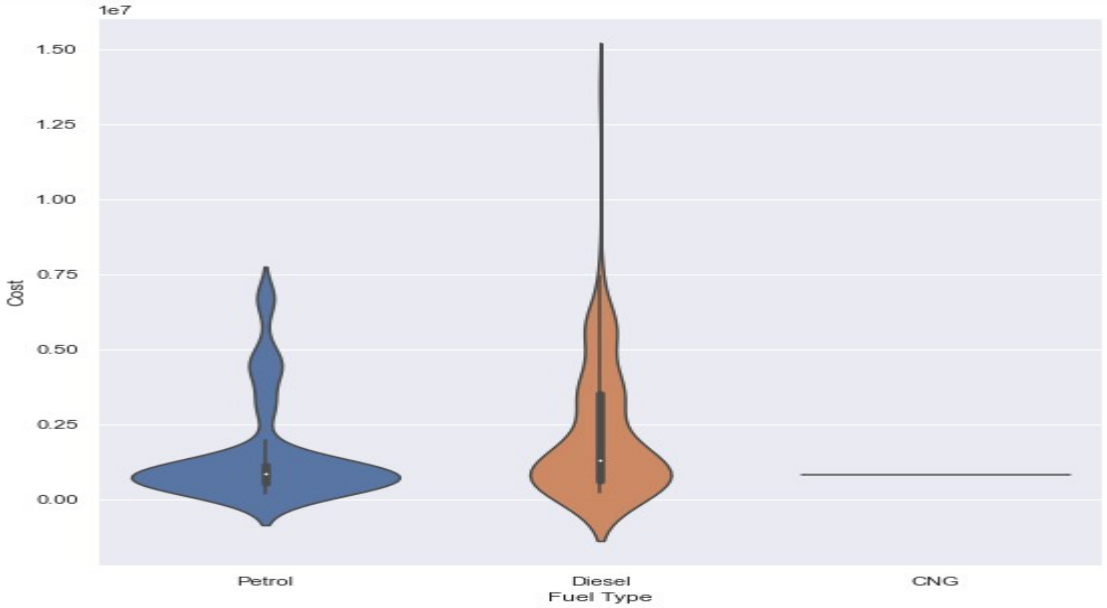
•This graph shows us the range of KMs of cars .The Cars having KMs between the range 20000-40000 is maximum than the cars having KMs between range 40000-60000.

distplot() :



•This plot is Car Model year and their Density of Car's which present in given Year's .So The most of the car's of model year present in year 2017-2021 and it's density is very high i.e 0.4 & above. and Car's of model Year 2010-2015 is present at minimum density structure of car model year.
•Also car's of model year 2010-2012 is present in very less amount and basically 2017-19 is the very high percentage of the car model year. But 2019 is the very high car model year in percentage wise as compare all of this year.

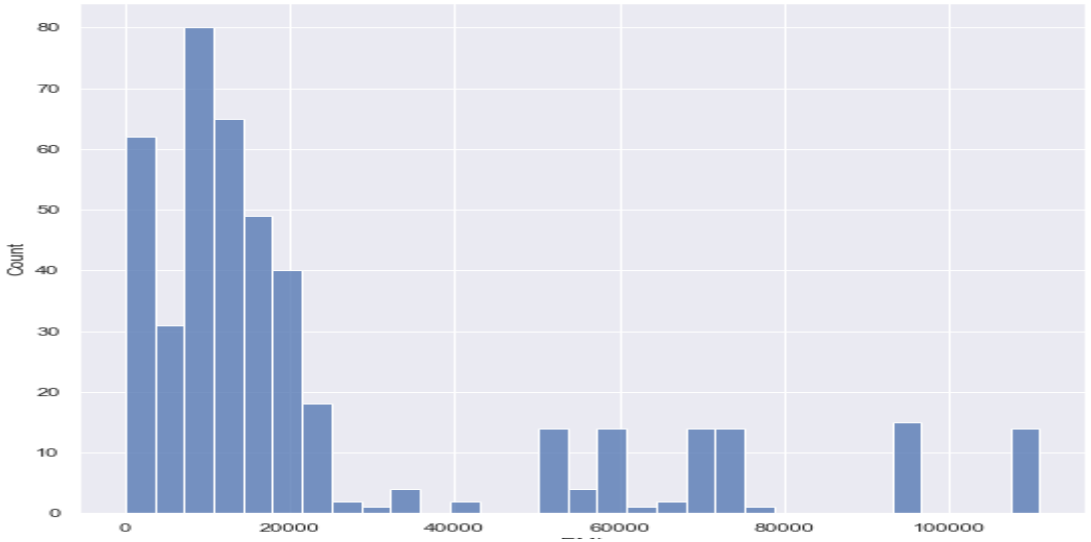
VIOLIN PLOTS :



•The Violin Plot of 'Cost' of Cars considering its Fuel Type like Petrol, Disel, CNG and here from the plot based on our dataset we observe that on the above of 0.75 of cost there are very few vheical range it is near about 1% only but under 0.75 of cost there is maximum range of Cost of vehical. Also it shows in the range of 0.00 to 0.25 of cost there is very maximum range area of cost of vehical depends on Fuel Type and that's the informataion of Cost of cars by using its Fuel type shown here.

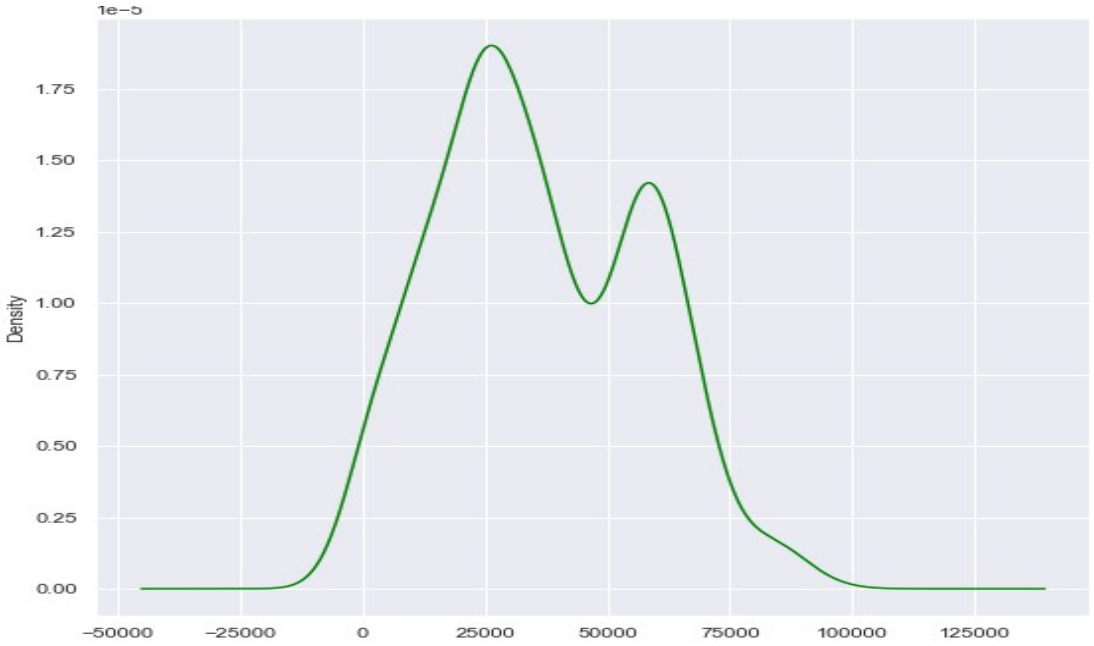
INNOMATICS RESEARCH LABS

Histogram:



•Here The count of cars which has minimum amount of EMI cost between the range 0-25000 is very high which is approx to 80% of the total count .and their are very few cars which has very high amount of EMI cost between Rs.50000-Rs.12500 are shown here which has count between range 0-15 aprox.

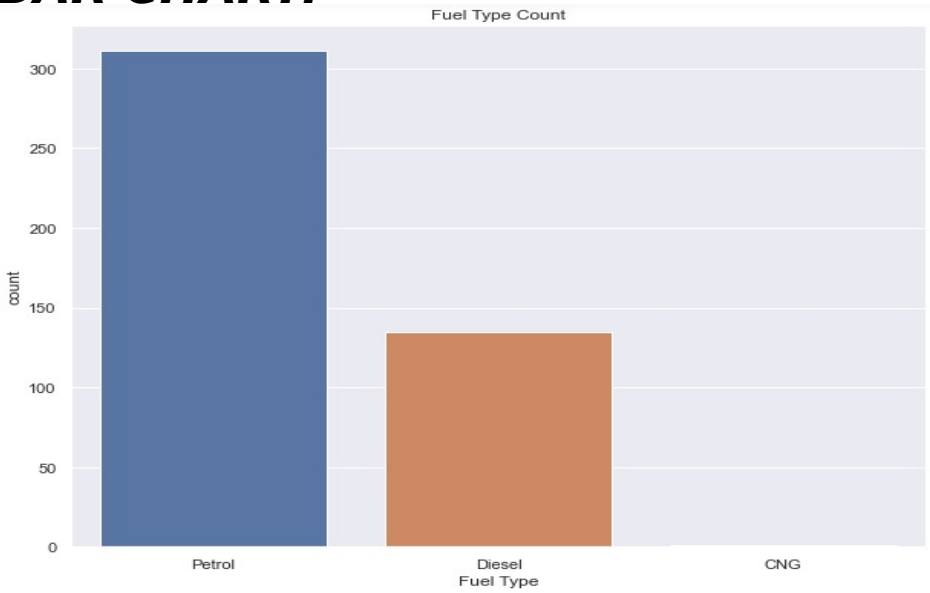
Density Plot:



•The Cars having KMs between the range 20000-40000 is maximum. i.e its density is grater than 1.75 and the cars having KMs between range 40000-60000 have density in between 1.00-1.50.are observed from this graph.

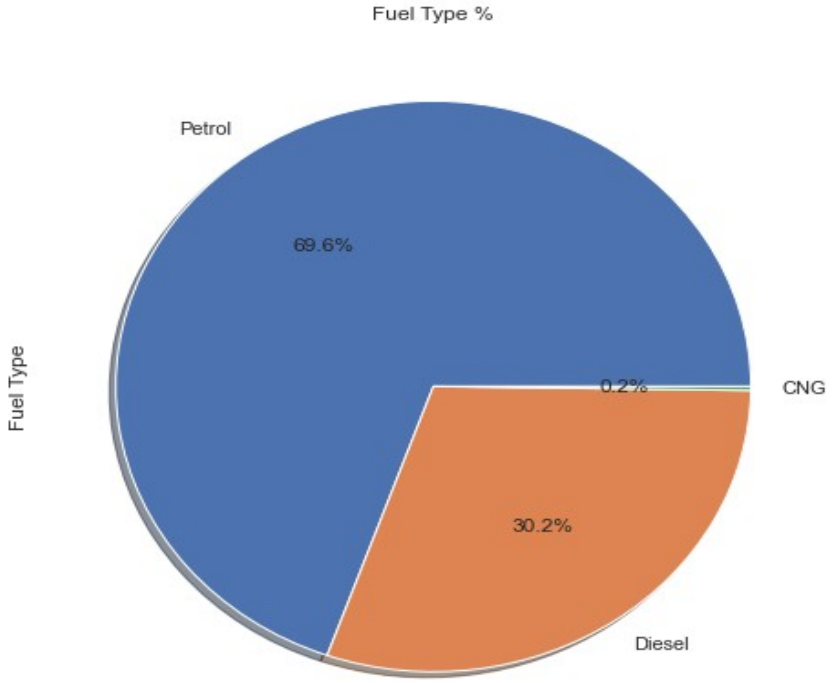
INNOMATICS RESEARCH LABS

BAR CHART:



•Here In our dataset the count of Petrol cars is maximum i.e above 300 of the count of total cars present in our dataset. their are minumum count of Disel cars which is aprox 130 only. and their are very less amount of CNG Cars present in our datset. CNG cars are very rare in our data set.

PIE CHART :

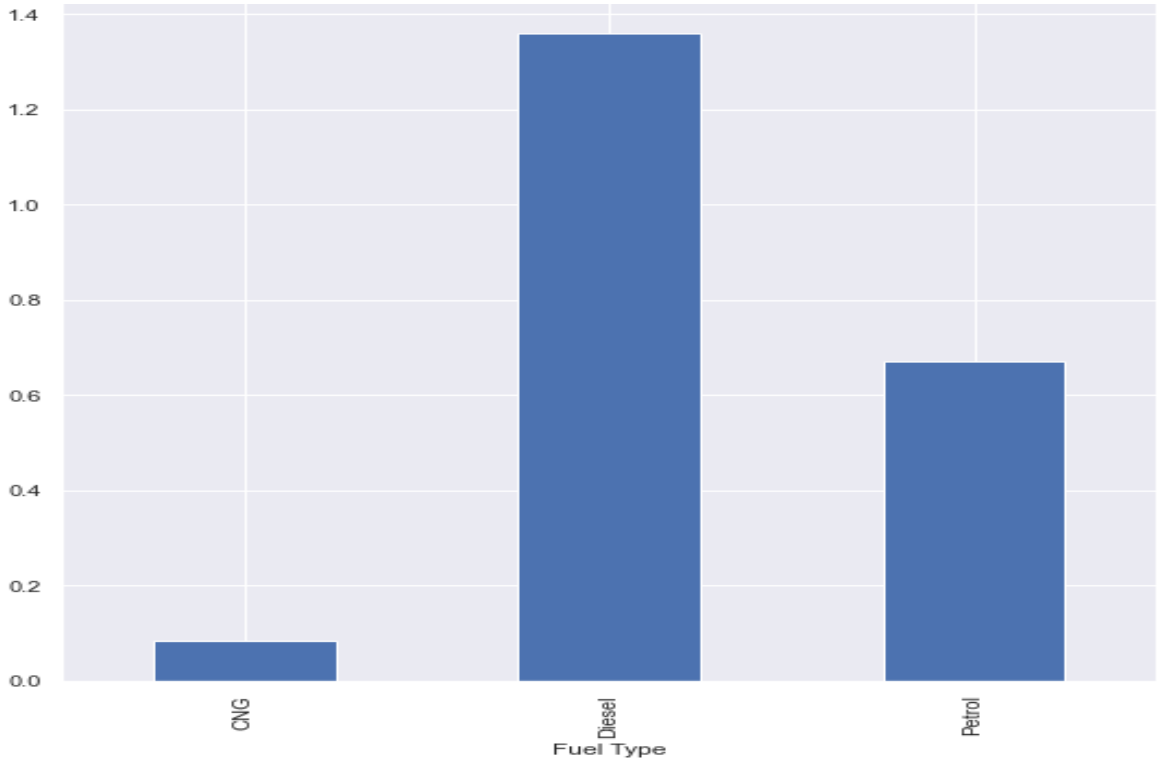


- Also From by this pie diagrame we can observe that their are 69.6% of Petrol Cars, 30.2% of Disel Cars And 0.2% of CNG Cars present in our data set.

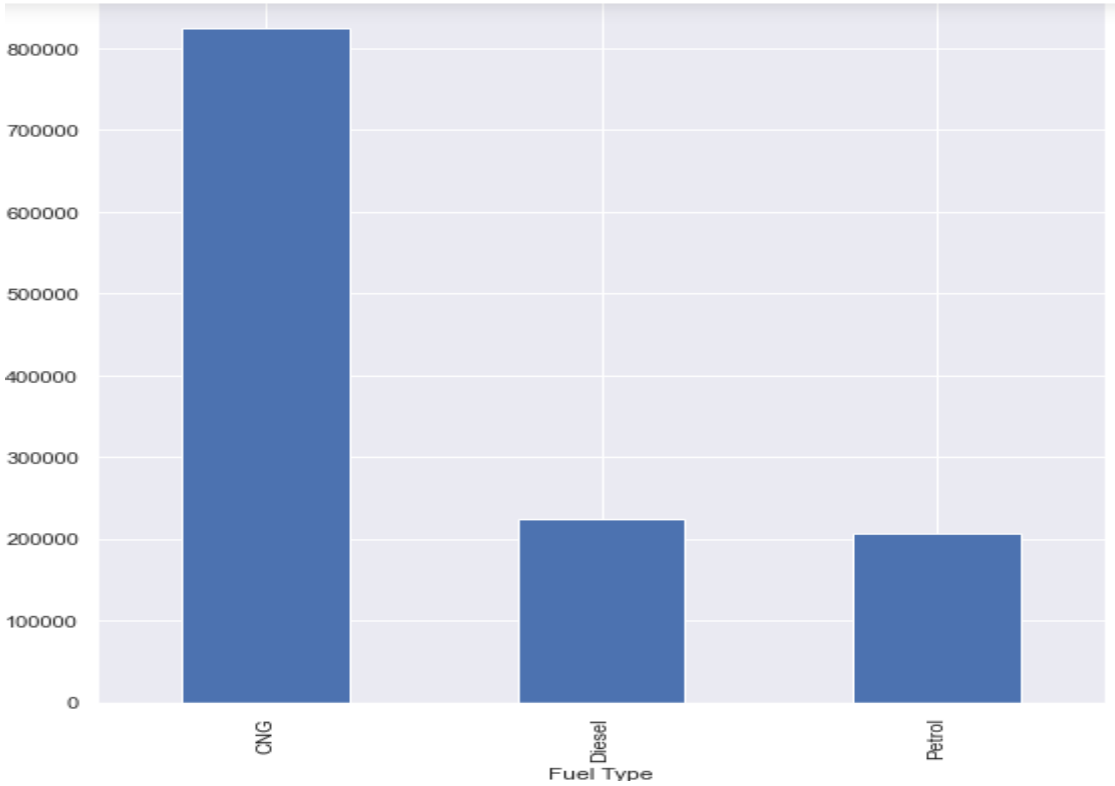
INNOMATICS RESEARCH LABS

variate/ Multivariate Analysis :

groupby():



Groupby function for Fuel Type and its Cost for Maximum.



Groupby function for Fuel Type and its Cost for Minimum.

INNOMATICS RESEARCH LABS

***Pivot Table*:**

Brand Name	Car Model Year	Cost	EMI	KMs	Unnamed: 0
250 Avantgarde	2010	875000.0	0.0	47000.0	353.000000
Audi A4	2014	1399000.0	0.0	81000.0	196.000000
	2016	2700000.0	0.0	82695.0	256.000000
Audi Q3	2015	1425000.0	23664.0	70405.0	336.000000
Audi Q5	2011	1236999.0	0.0	68000.0	198.000000
...		...	...	...	...
Volkswagen Polo	2020	650000.0	10794.0	64293.0	2.000000
Volkswagen Tiguan	2017	2150000.0	35704.0	69986.0	98.000000
Volkswagen Vento	2011	298000.0	0.0	59000.0	135.000000
	2018	860000.0	14281.0	54314.0	226.214286
	2019	1099000.0	18250.0	38754.0	243.833333

[106 rows x 4 columns]

•From this pivot table we get the information about First Hand/ Second Hand cars present in our dataset with its different features like its Brand Name, Car Model Year, Cost of car , EMI , and its KMs.

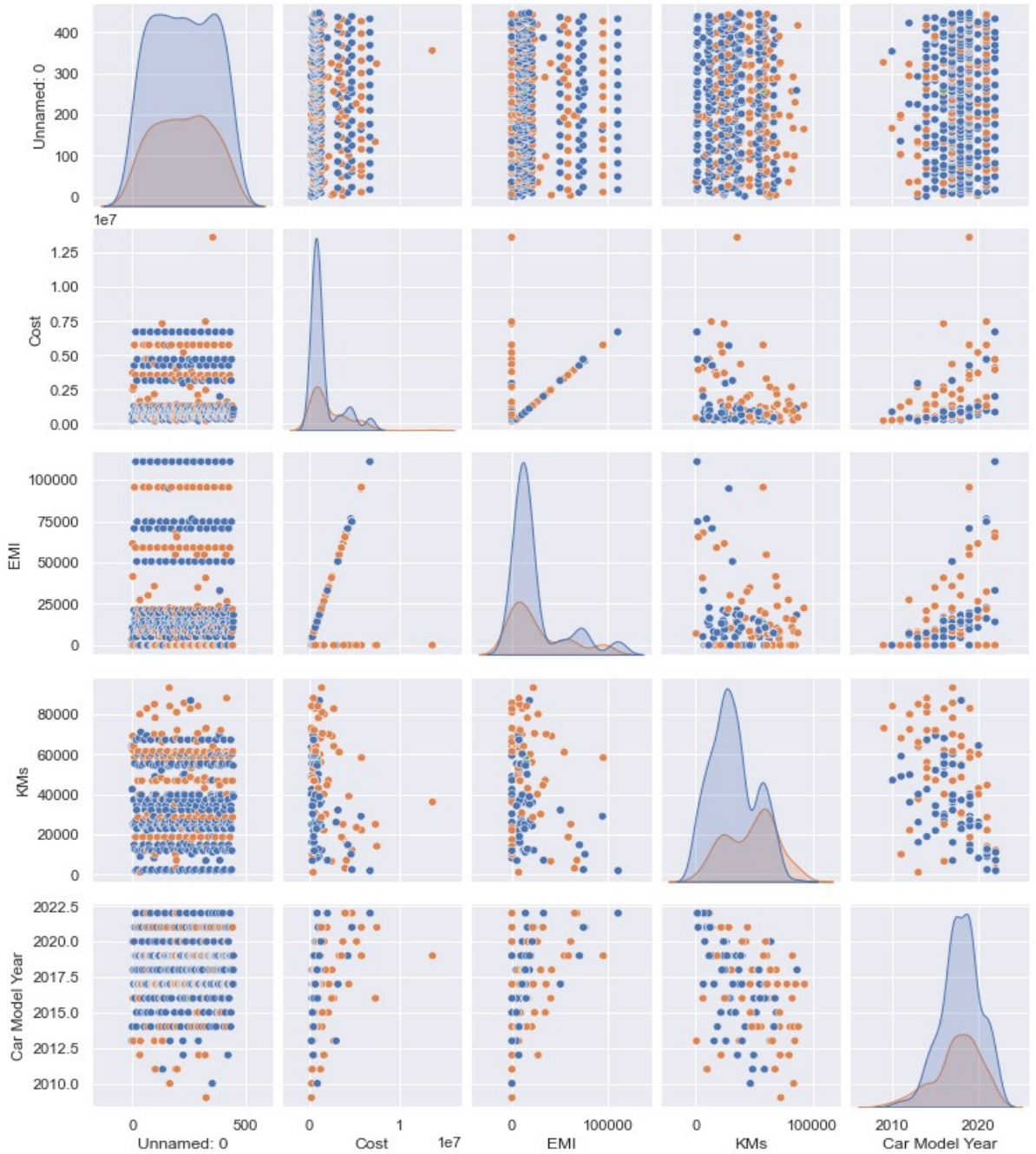
***Crosstab*:**

Fuel Type	CNG	Diesel	Petrol
Brand Name			
250 Avantgarde	0	0	1
Audi A4	0	2	0
Audi Q3	0	2	0
Audi Q5	0	1	0
Audi TT	0	0	1
...	...	...	...
Toyota Innova	0	2	0
Volkswagen Ameo	0	0	1
Volkswagen Polo	0	0	2
Volkswagen Tiguan	0	1	0
Volkswagen Vento	0	0	21

65 rows × 3 columns

Here we can use the crosstab plot for all Cars Brand and considering its Fuel type like Petrol, Disel, CNG present in our dataset.

Pairplot:

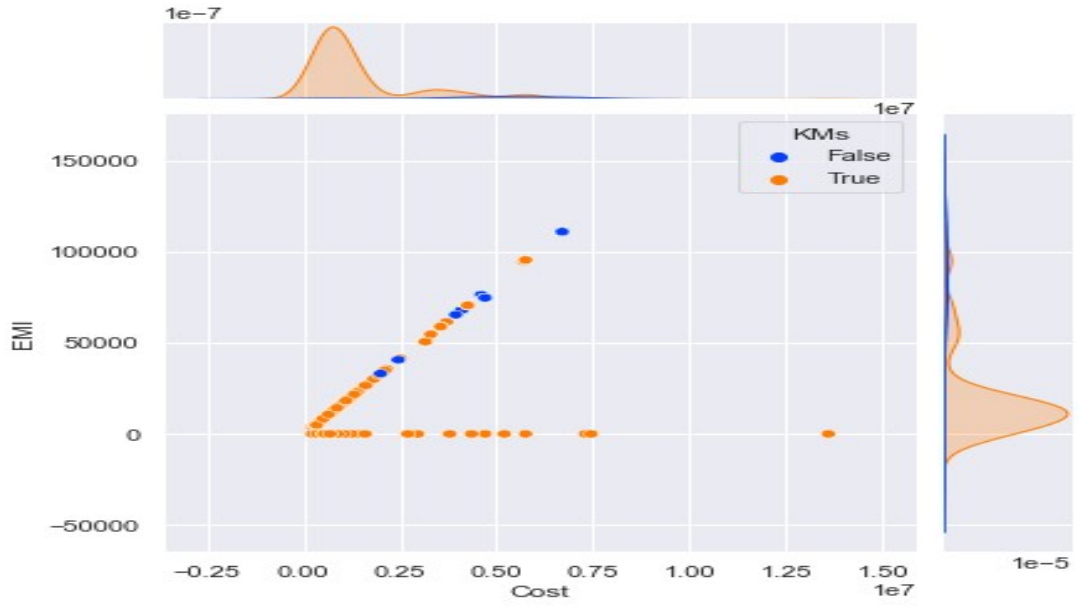


Heatmap:

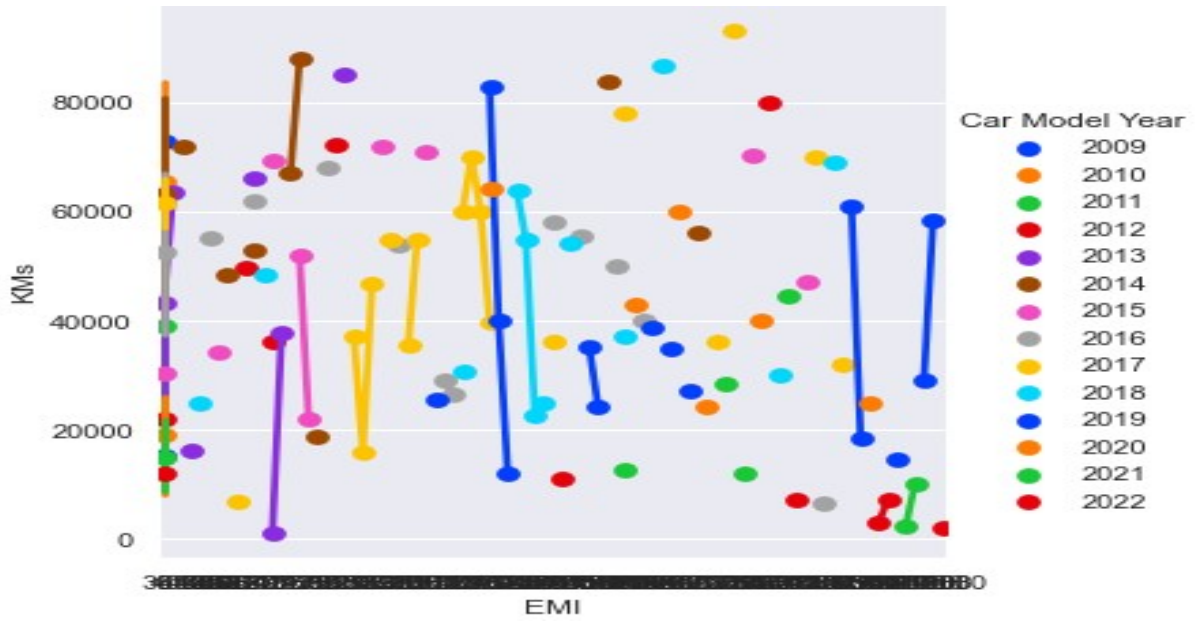
•Heat Maps is to better visualize the volume of different Features of Cars within a dataset and assist in directing viewers towards areas on data visualizations that matter most.

INNOMATICS RESEARCH LABS

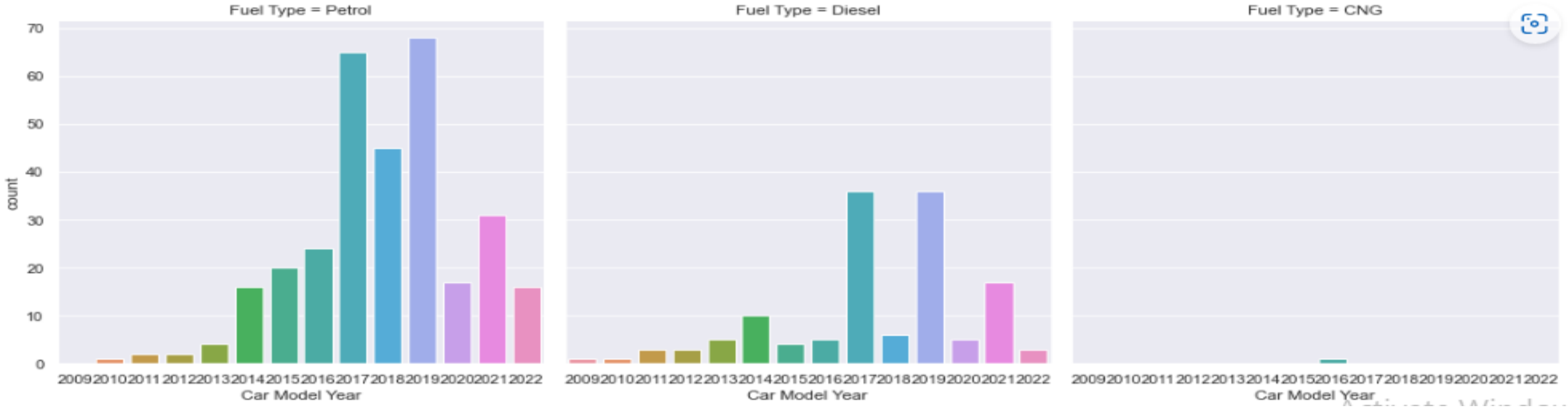
Jointplot:



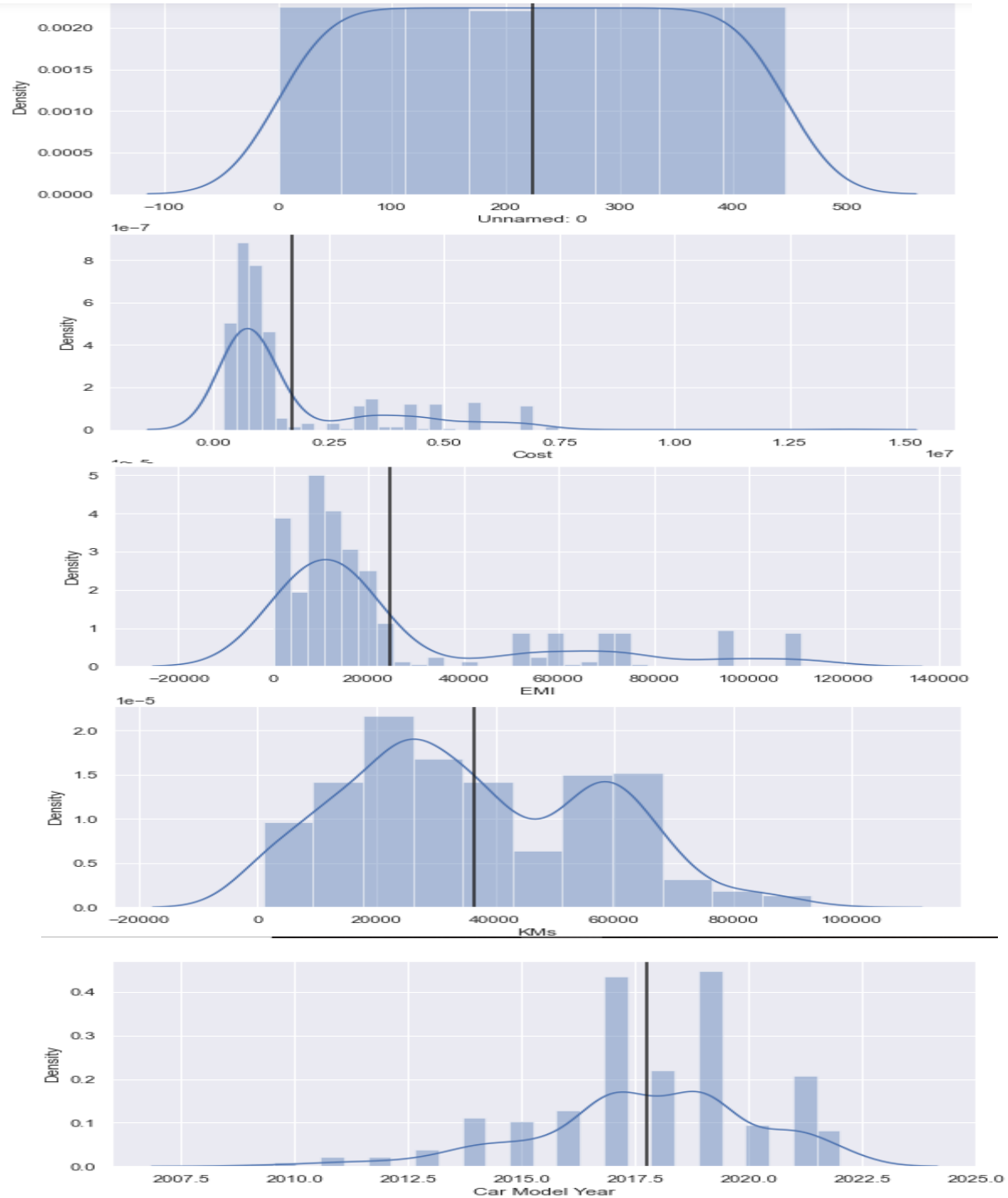
Point plot:



Count Plot:



Distribution Plot:



INNOMATICS RESEARCH LABS

Conclusion:-

•The data that we are going to use in this example is about First Hand/ Second Hand Cars. Specifically containing various information data points about the used cars, like their Car Name, Cost, EMI, Features, KMs, Fuel Type, Car Model Year, Brand Name, City etc.

• By this analysis we can observe that in our data in year 2010-2023 there is large amount of Petrol cars were used than Disel cars and very few amount of CNG cars are there.

•Also There are most of the car's having model year in between 2017-2020. And the count of the Cars having minumum amount of EMI upto Rs.25000 is very high .

•Also From by the pie diagram we can conclude that their are 69.6% of Petrol Cars, 30.2% of Disel Cars And 0.2% of CNG Cars present in our data set. i.e. By considering the fuel type of cars the count of petrol cars is high , than Disel and CNG Cars.

THANK YOU

INNOMATICS
RESEARCH LABS