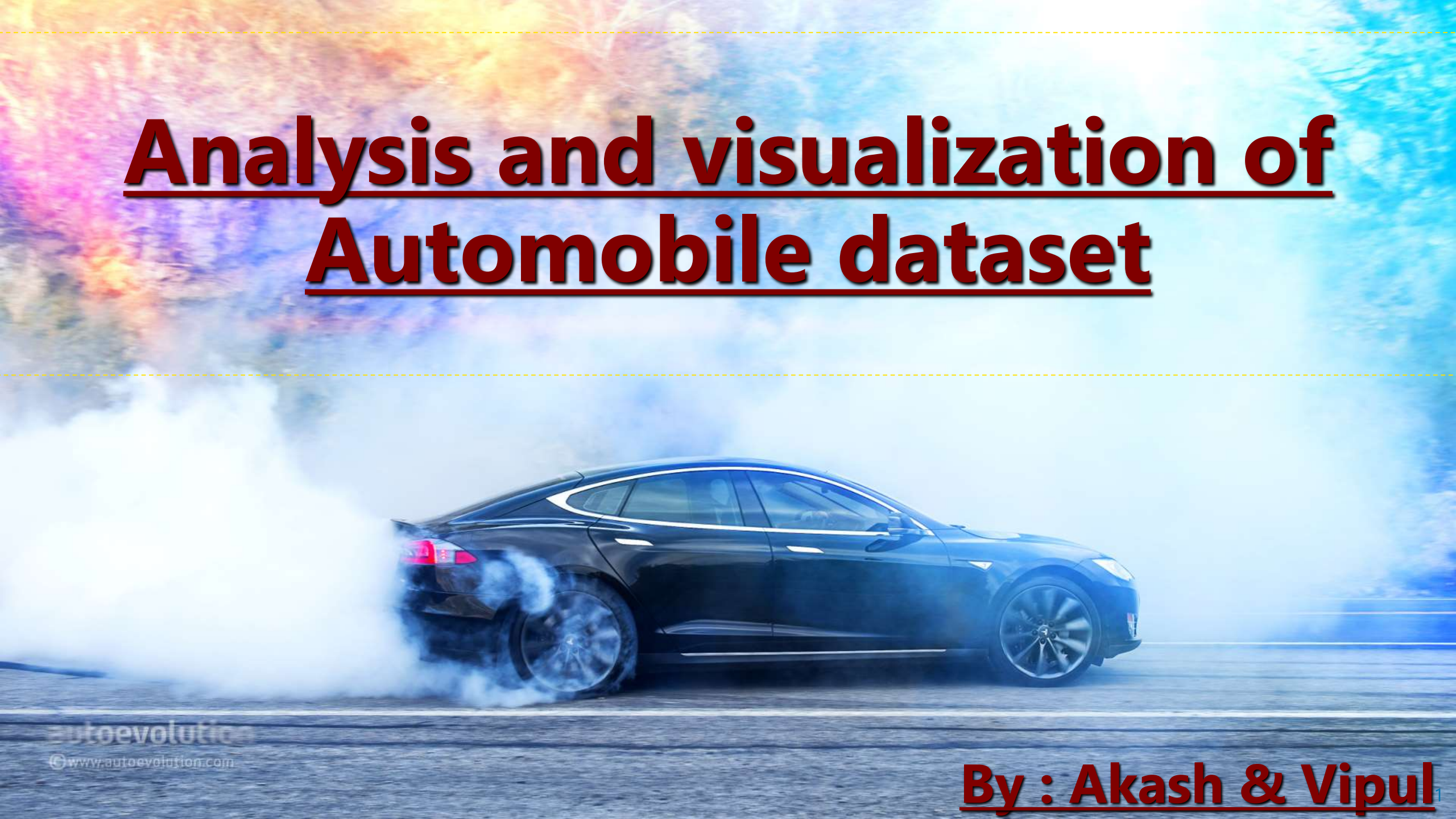
Analysis and visualization of Automobile dataset



By : Akash & Vipul

DATA SCIENCE

“Data science is ultimately about using this data in creative ways to generate business value.”

- **Data science helps to uncover patterns and find insights about data behavior and world trends to an unprecedented extent.**
- Data collection includes data gathering from diverse sources.
- Data munging or data cleaning is method of generalizing, clustering, pattern matching, filtering data.
- Exploratory Data Analysis (EDA) is the process of analyze patterns, trends, outliers and relationship amongst the data variables.
- Data Visualization helps to create graphical displays of high-dimensional data containing many variables and find correlation amongst the data and find insights.

Automobile

Automobile dataset consists of three types of entities:

(a) The specification of an auto(car) in terms of various characteristics.

(b) Use python Jupiter notebook to load the given data.

(c) Than clean the non-numeric values to null ,drop invalid data and fill the missing value .

(d) Also done univariate and bi-variate analysis to see impact on automobile data pricing and other.

Automobile dataset will be clustering the cars by specification. For this, I will non-numeric(?) variables from the data set and work with just the specification variables which are both categorical and numerical.

Automobile

Exploratory Data Analysis Using Python

!

1	2	3	4	5
Domain Info	Data Set Description	Data Analysis using python	Data visualization	Findings

Domain Info

(Car specification)

Automobile Data Analysis

Data set → Data analysis & visualization → Insight

Automobile Data Analysis

The aim here is to explore the data set and find basic relationships of different variables from an automobile dataset. It will be starting with some domain info with brief exploratory analysis involving summary statistics and graphical visualization.

Clean the given dataset and analyze the different attributes and find correlation of different attributes and define statistical summary of the given dataset. Visualization of univariate and bi-variate analysis for better understanding of individual and multiple data attribute's impact on others.

Dataset Description

This Data set contains Info about Automobile (Cars Specification)

No.	Data variables	Type	Description
1	symboling	int	-3 -2 -1 0 1 2 3.
2	normalized-losses	String	Continuous from 65 to 256
3	make	String	Different brands of car
4	fuel-type	String	“diesel”, "gas"
5	aspiration	String	“std” , "turbo"
6	num-of-doors	String	“four”, "two"
7	body-style	String	“Hardtop", "Hatchback” ,“Sedan”, “Wagon"," Convertible”
8	drive-wheels	String	"4wd","fwd","rwd"
9	engine-location	String	“front”, "rear"
10	wheel-base	float	Continuous from 88.6 to 120.4
11	length	float	Continuous from 141.1 to 208.1
12	width	float	Continuous from 60.3 to 72.3
13	height	float	Continuous from 47.8 to 59.8
14	curb-weight	int	Continuous from 1488 to 4066
15	engine-type	String	“dohc”,"dohcv“ ,..:
16	num-of-cylinders	String	“eight” , "five",..:
17	engine-size	int	Continuous from 61 to 326
18	fuel-system	String	"1bbl","2bbl",..:
19	bore	String	Continuous from 2.54 to 3.94
20	stroke	String	Continuous from 2.07 to 4.17
21	ratio	float	Continuous from 7 to 23
22	horsepower	String	Continuous from 48 to 288
23	peak-rpm	String	Continuous from 4150 to 6600
24	city-mpg	int	Continuous from 13 to 49
25	highway-mpg	int	Continuous from 16 to 54
26	price	String	Continuous from 15118 to 45400

I. Importing libraries and load the data set

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
df=pd.read_csv('C:\\Users\\tanis\\OneDrive\\Desktop\\Automobile_data.csv')
df.head()
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	pe
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154	5
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.4	10.0	102	5
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.4	8.0	115	5

5 rows × 26 columns

Data Analysis

Data Cleaning

We'll cover the following:

- Dropping empty rows from dataset.
- Find out data having '?' value for numeric columns and replace it with mean.
- Remove the records which are having the value '?' in non-numeric column.
- Using astype() method to change the type of value.
- Using the DataFrame.applymap() function to clean the entire dataset, element-wise

II. Data Cleaning

```
print('cleaning the data')
df.dropna(how='all',inplace=True)
```

```
cleaning the data
```

```
print(df['normalized-losses'].value_counts())
d=df['normalized-losses'].loc[df['normalized-losses']=='?'].value_counts()
d
```

```
df['num-of-doors'].loc[df['num-of-doors']=='?']
df=df[df['num-of-doors']!='?']
df['num-of-doors'].loc[df['num-of-doors']=='?']
```

```
Series([], Name: num-of-doors, dtype: object)
```

III. Statistical summary of dataset

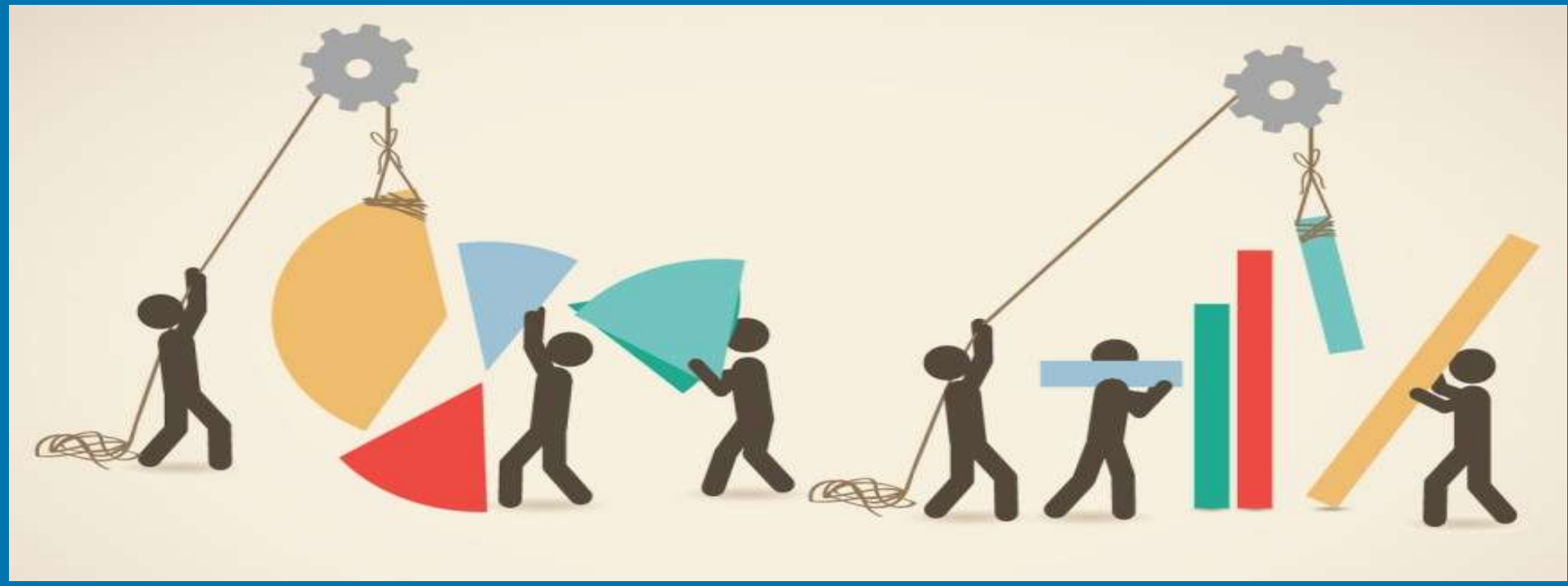
Slide Type Slide

```
print("3. Summary Statistics of different variable")
Automobile.describe()
```

```
3. Summary Statistics of different variable
```

	symboling	normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke	compression-ratio	horsepower	
count	203.000000	203.000000	203.000000	203.00000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	2
mean	0.837438	121.871921	98.781281	174.11330	65.915271	53.731527	2557.916256	127.073892	3.330931	5.593892	10.093202	104.463054	5
std	1.250021	31.784599	6.040994	12.33909	2.150274	2.442526	522.557049	41.797123	0.271327	16.547416	3.888216	39.612384	
min	-2.000000	65.000000	86.600000	141.10000	60.300000	47.800000	1488.000000	61.000000	2.540000	2.070000	7.000000	48.000000	4
25%	0.000000	101.000000	94.500000	166.55000	64.100000	52.000000	2145.000000	97.000000	3.150000	3.110000	8.600000	70.000000	4
50%	1.000000	122.000000	97.000000	173.20000	65.500000	54.100000	2414.000000	120.000000	3.310000	3.290000	9.000000	95.000000	5
75%	2.000000	137.000000	102.400000	183.30000	66.900000	55.500000	2943.500000	143.000000	3.585000	3.435000	9.400000	116.000000	5
max	3.000000	256.000000	120.900000	208.10000	72.300000	59.800000	4066.000000	326.000000	3.940000	122.000000	23.000000	288.000000	6

Visualization



One of the most important benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals.

IV. The impact of other variable on automobile pricing

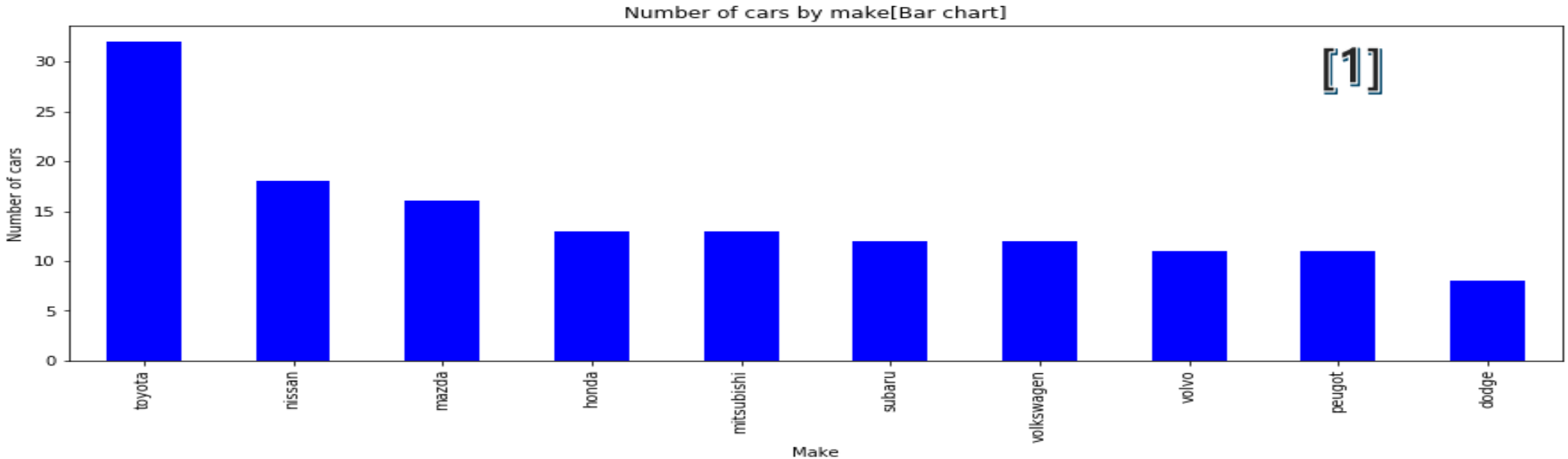
```
a=df["make"].value_counts()
a.plot(kind="bar",figsize=(15,5))
plt.show()
```

```
j=df['drive-wheels'].value_counts()
print(j)
j.plot(kind='pie',autopct='%.2f')
plt.show()
```

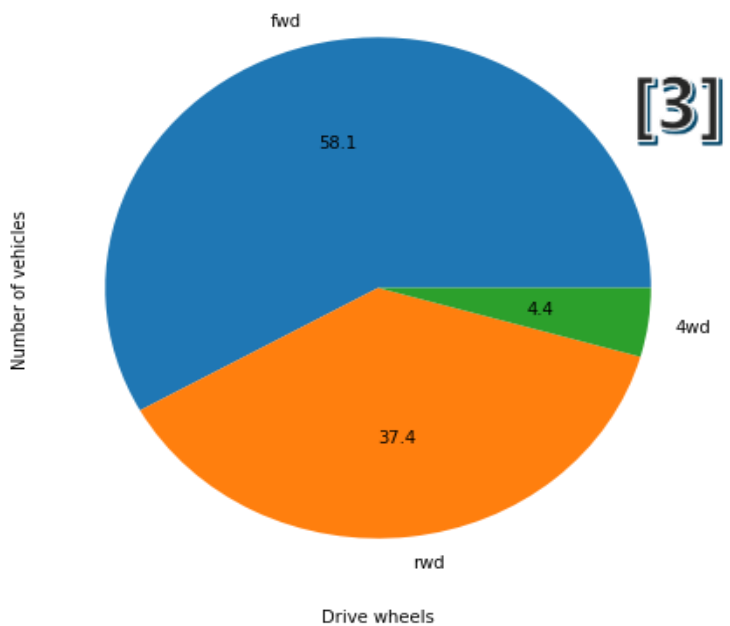
```
drive-wheels
fwd    118
rwd     76
4wd      9
Name: count, dtype: int64
```

```
b=df['aspiration'].value_counts()
print(b)
b.plot(kind='pie',autopct='%.2f')
plt.show()
```

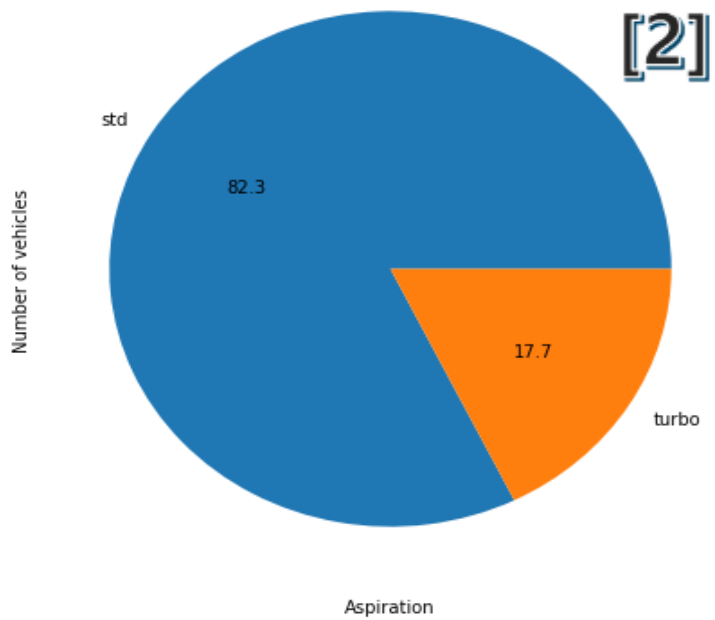
```
aspiration
std      168
turbo     37
Name: count, dtype: int64
```



Drive wheels [Pie chart]



Aspiration frequence diagram(Pie Chart)



Findings :

1. Toyota is the make of the car which has most number of vehicles with more than 40% than the 2nd highest Nissan
2. Most preferred fuel type for the customer is standard vs trubo having more than 80% of the choice
3. For drive wheels, front wheel drive has most number of cars followed by rear wheel and four wheel. There are very less number of cars for four wheel drive.
4. Curb weight of the cars are distributed between 1500 and 4000 approximately
5. Symboling or the insurance risk rating have the ratings between -3 and 3 however for our dataset it starts from -2. There are more cars in the range of 0 and 1.
6. Normalized losses which is the average loss payment per insured vehicle year is has more number of cars in the range between 65 and 150.
7. For No. of doors , car has most preferable is four doors than two doors.

Graph Discription

Scatter plot of

1.Price and engine size : The more the engine size the costlier the price is.

2.Normalized losses and symbolling : From the scattered plot, it's very evident that the lesser the rating lesser the normalized loss. It looks like the negative ratings are better for the car which has lesser losses.

3.City and Highway MPG, Curb weight based on Make of the car :

- It is clear that for both city and highway mileage of the automobile is inversely proportional to the curb weight.

- Heavier the Automobile less is the mileage for both City and Highway.

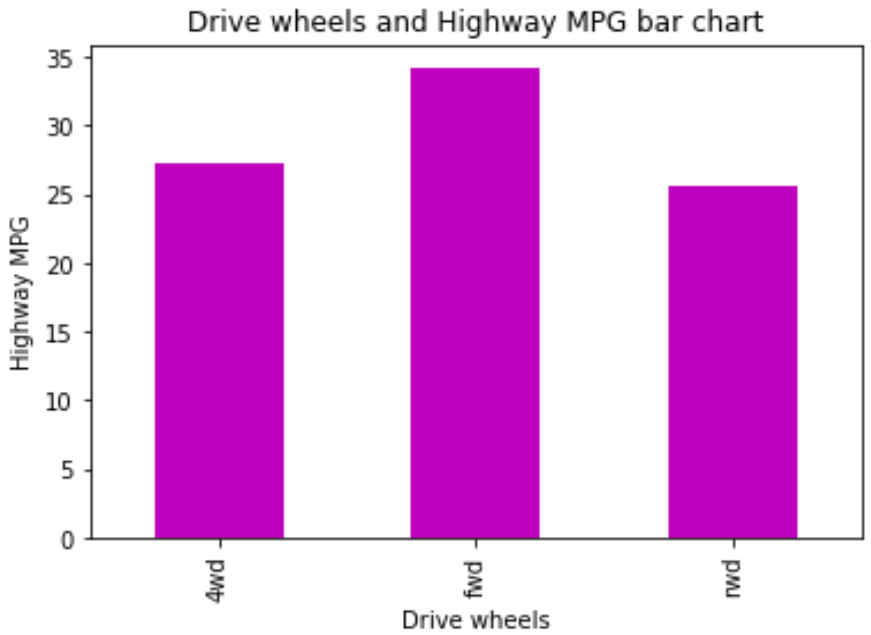
Bar chart of

5.Normalized losses based on body style and no. of doors: Normalized losses are distributed across different body style but the two door cars has more number of losses than the four door cars.

6. Drive wheels based on City and Highway MPG : It's very evident that the front wheel drive cars are most preferable than others.

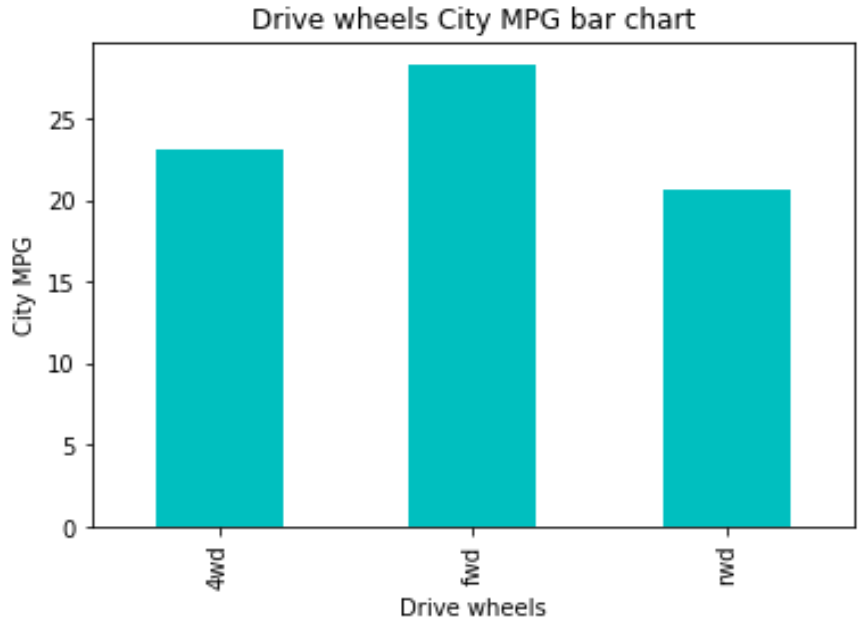
```
#drive wheels and highway mpg bar chart
print("Drive wheels and City MPG bar chart")
print(df.groupby('drive-wheels')['highway-mpg'].mean().plot(kind='bar', color = 'm'))
plt.title("Drive wheels and Highway MPG bar chart")
plt.ylabel('Highway MPG')
plt.xlabel('Drive wheels');
```

```
Drive wheels and City MPG bar chart
Axes(0.125,0.11;0.775x0.77)
```

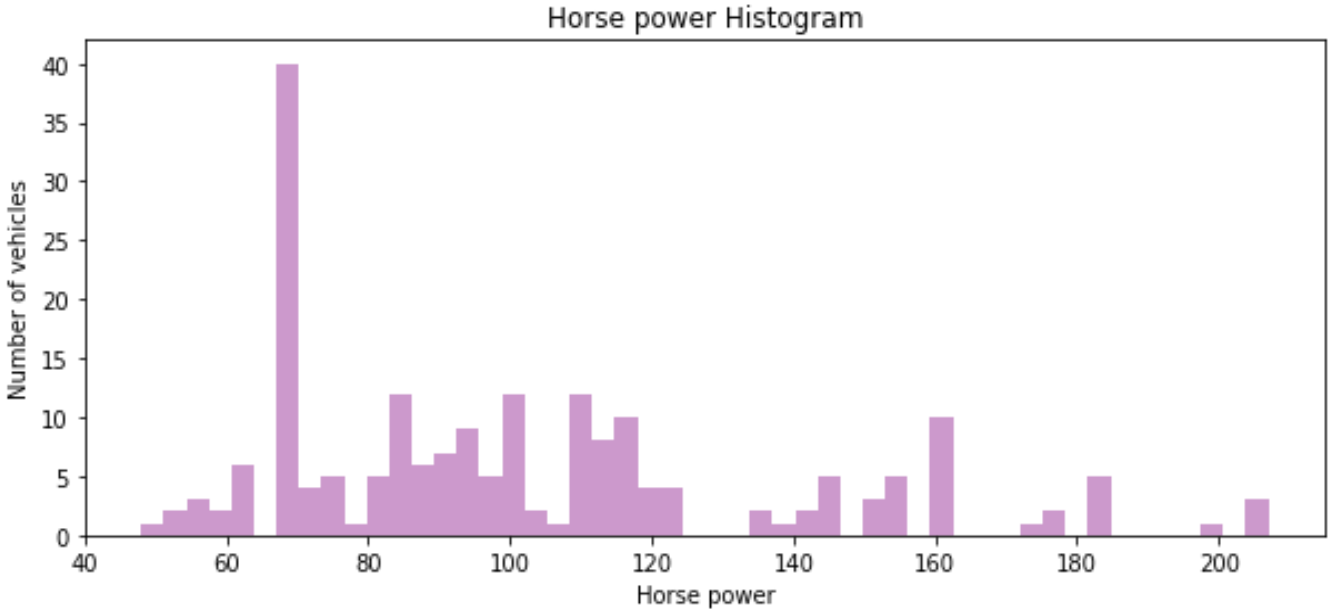


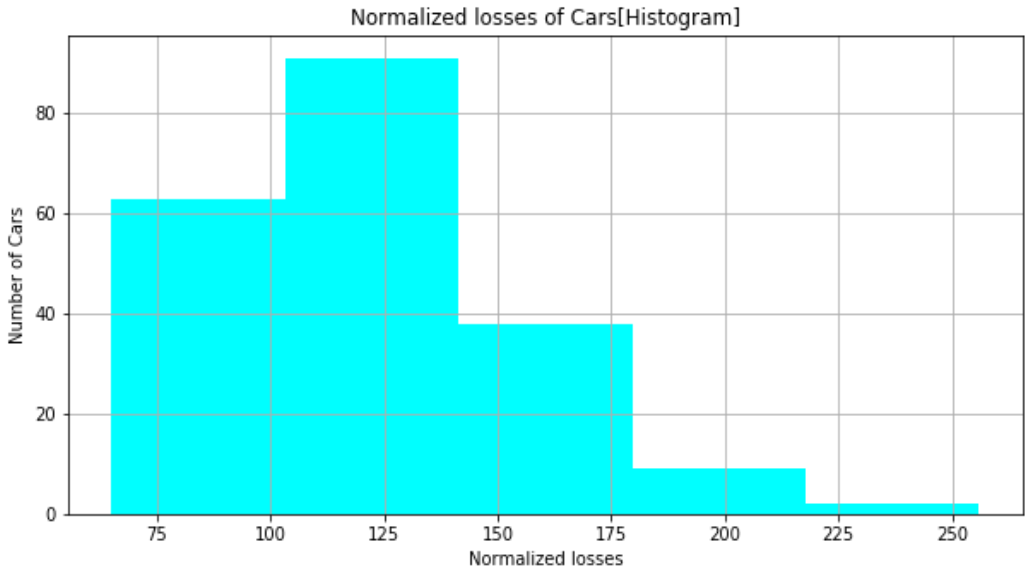

```
#drive wheels with city mpg mean() bar chart
print(df.groupby('drive-wheels')['city-mpg'].mean().plot(kind='bar', color = 'c'))
plt.title("Drive wheels City MPG bar chart")
plt.ylabel('City MPG')
plt.xlabel('Drive wheels');
```

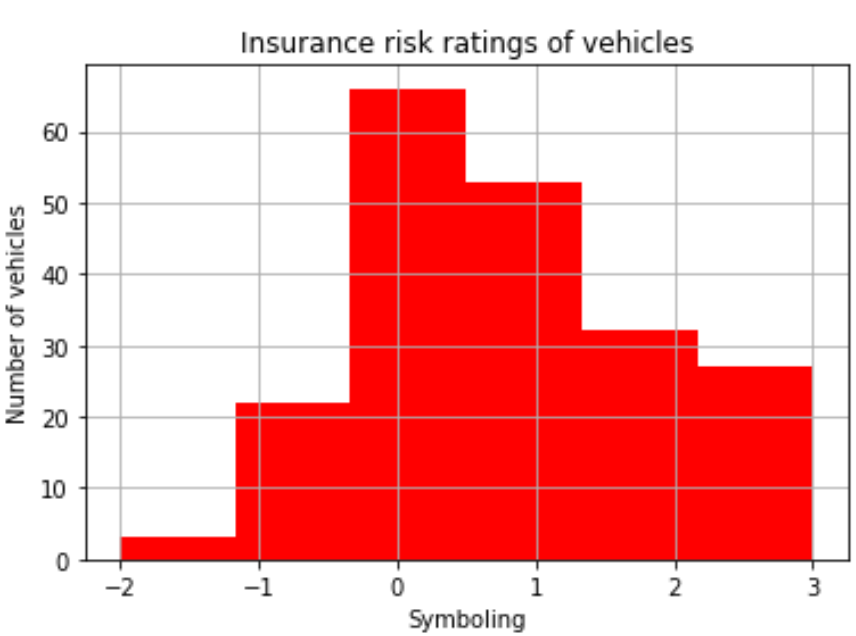
```
Axes(0.125,0.11;0.775x0.77)
```

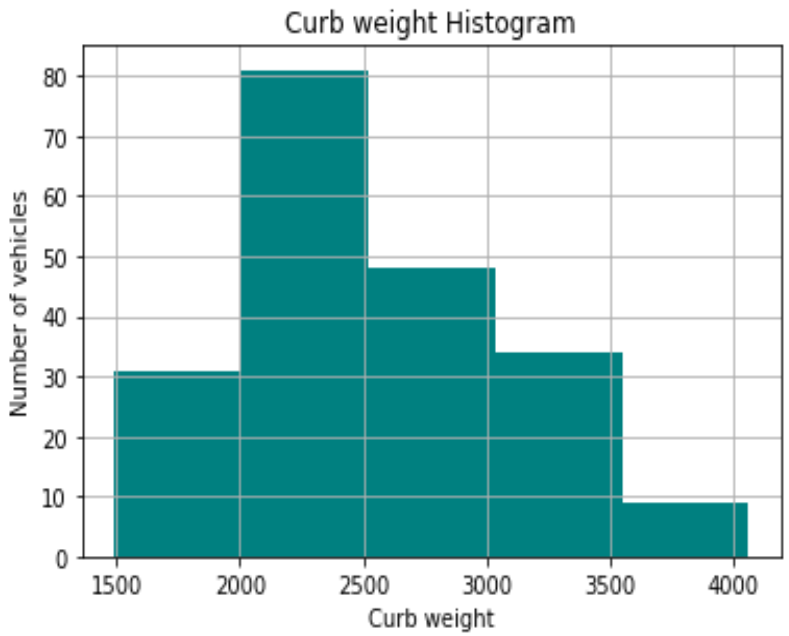


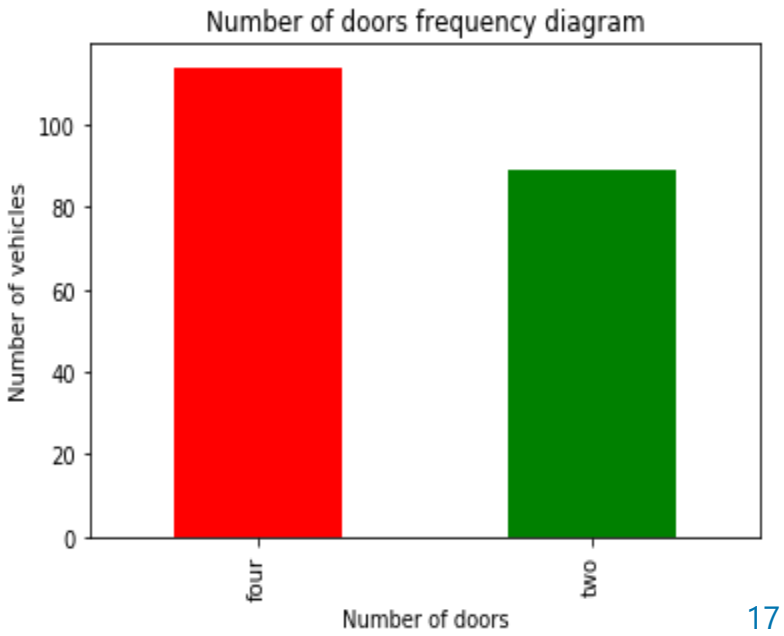
Continue..


Horse power Histogram
Number of vehicles
Horse power


Normalized losses of Cars[Histogram]
Number of Cars
Normalized losses


Insurance risk ratings of vehicles
Number of vehicles
Symboling


Curb weight Histogram
Number of vehicles
Curb weight


Number of doors frequency diagram
Number of vehicles
four
two
Number of doors

Findings :

Below are our findings on the make and price of the car:

1. The most expensive car is manufacture by Mercedes benz and the least ex pensive is Chevrolet.
2. The premium cars costing more than 20000 are BMW, Jaquar, Mercedes be nz and Porsche.
3. Less expensive cars costing less than 10000 are Chevrolet, Dodge, Honda, Mitsubishi, Plymoth and Subaru.
4. Rest of the cars are in the midrange between 10000 and 20000 which has t he highest number of cars.

THANK YOU