Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as smf
```

Importing Dataset

dataset=pd.read_csv('car performance.csv')
dataset

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year
0	18.0	8	307.0	130.0	3504	12.0	70
1	15.0	8	350.0	165.0	3693	11.5	70
2	18.0	8	318.0	150.0	3436	11.0	70
3	16.0	8	304.0	150.0	3433	12.0	70
4	17.0	8	302.0	140.0	3449	10.5	70
393	27.0	4	140.0	86.0	2790	15.6	82

Finding missing data

dataset.isnull().any()

mpg	False
cylinders	False
displacement	False
horsepower	True
weight	False

X

origin False car name False dtype: bool

There are no null characters in the columns but there is a special character '?' in the 'horsepower' column. So we we replaced '?' with nan and replaced nan values with mean of the column.

```
dataset['horsepower']=dataset['horsepower'].replace('?',np.nan)
dataset['horsepower'].isnull().sum()
    6
dataset['horsepower']=dataset['horsepower'].astype('float64')
dataset['horsepower'].fillna((dataset['horsepower'].mean()),inplace=True)
dataset.isnull().any()
                    False
    mpg
    cylinders
                    False
    displacement
                    False
    horsepower
                    False
    weight
                    False
    acceleration
                    False
    model year
                    False
    origin
                    False
    car name
                    False
    dtype: bool
```

dataset.info() #Pandas dataframe.info() function is used to get a quick overview

```
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
#
    Column
                  Non-Null Count
                                 Dtype
    -----
- - -
                                  ----
0
    mpg
                  398 non-null
                                 float64
                398 non-null
 1
    cylinders
                                 int64
 2
    displacement 398 non-null
                                 float64
 3
                  398 non-null
                                 float64
    horsepower
                                 int64
 4
    weight
                  398 non-null
 5
    acceleration 398 non-null
                                 float64
 6
    model year 398 non-null
                                 int64
7
    origin
                  398 non-null
                                 int64
8
    car name
                  398 non-null
                                 object
```

<class 'pandas.core.frame.DataFrame'>

dtypes: float64(4), int64(4), object(1)

memory usage: 28.1+ KB

dataset.describe() #Pandas describe() is used to view some basic statistical det

	mpg	cylinders	displacement	horsepower	weight	accele
count	398.000000	398.000000	398.000000	398.000000	398.000000	398.
mean	23.514573	5.454774	193.425879	104.469388	2970.424623	15.
std	7.815984	1.701004	104.269838	38.199187	846.841774	2.
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.
25%	17.500000	4.000000	104.250000	76.000000	2223.750000	13.
50 %	23.000000	4.000000	148.500000	95.000000	2803.500000	15.
75 %	29.000000	8.000000	262.000000	125.000000	3608.000000	17.
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.

There is no use with car name attribute so drop it

dataset=dataset.drop('car name',axis=1) #dropping the unwanted column.

corr_table=dataset.corr()#Pandas dataframe.corr() is used to find the pairwise c
corr_table

	mpg	cylinders	displacement	horsepower	weight	acce
mpg	1.000000	-0.775396	-0.804203	-0.771437	-0.831741	
cylinders	-0.775396	1.000000	0.950721	0.838939	0.896017	
displacement	-0.804203	0.950721	1.000000	0.893646	0.932824	
horsepower	-0.771437	0.838939	0.893646	1.000000	0.860574	
weight	-0.831741	0.896017	0.932824	0.860574	1.000000	
acceleration	0.420289	-0.505419	-0.543684	-0.684259	-0.417457	
model year	0.579267	-0.348746	-0.370164	-0.411651	-0.306564	
origin	0.563450	-0.562543	-0.609409	-0.453669	-0.581024	

Data Visualizations

[] 😽 14 cells hidden

The P-value is the probability value that the correlation between these two variables is statistically significant.

Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the

- p-value is \$<\$ 0.001: we say there is strong evidence that the correlation is significant.
- the p-value is \$<\$ 0.05: there is moderate evidence that the correlation is significant.
- the p-value is \$<\$ 0.1: there is weak evidence that the correlation is significant.
- the p-value is \$>\$ 0.1: there is no evidence that the correlation is significant.

[]	▶ 25 cells hidden				

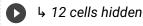
Seperating into	Depender	nt and Inc	depend	lent vari	ab	les
-----------------	----------	------------	--------	-----------	----	-----

[] 4 cells hidden

Splitting into train and test data.

[] 4 3 cells hidden

decision tree regressor



random forest regressor

[] 4 9 cells hidden

linear regression

[] 4 8 cells hidden

Colab paid products - Cancel contracts here