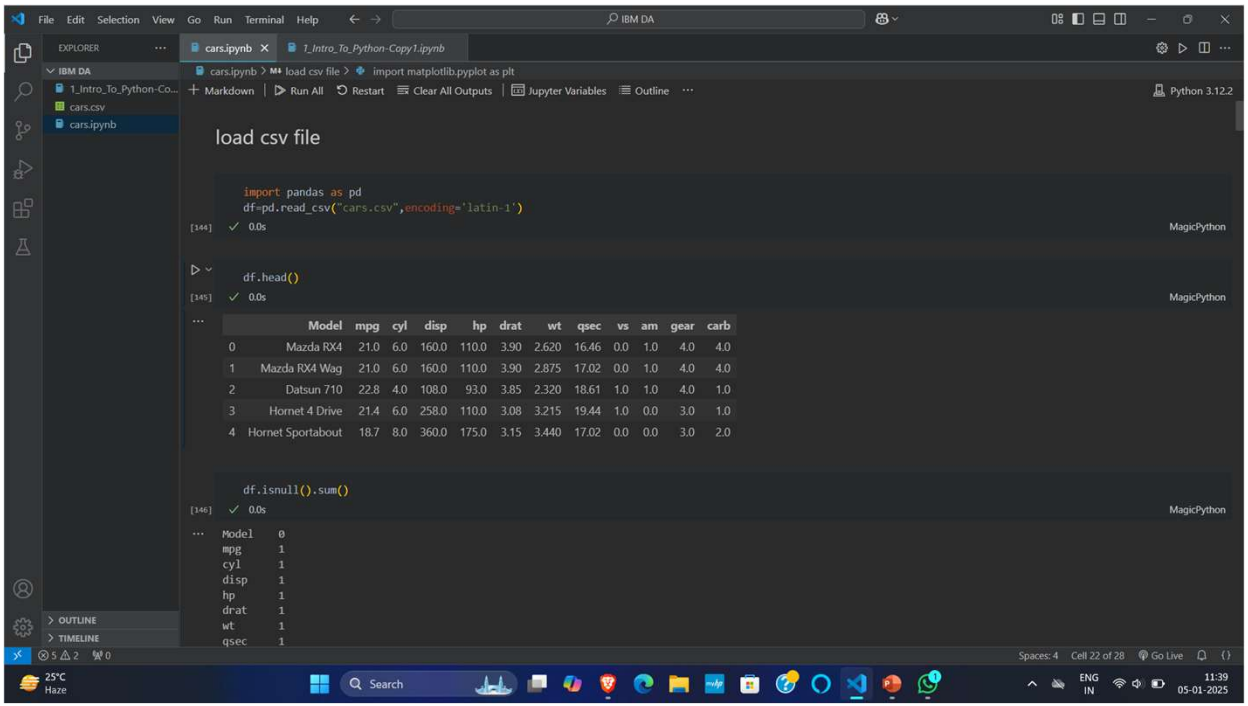
Exploratory Data Analysis of Car Specifications and Performance

MUDGOLWAD RAVI
JNTUH COLLEGE OF ENGINEERING SULTANPUR
SANGAREDDY, TELANGANA.

ASSIGNMENT-1: IDE-VSCODE



ASSIGNMENT-2

1. Load the "cars.csv" data file in Python.

```
import pandas as pd
df=pd.read_csv("cars.csv",encoding='latin-1')
```

[144] ✓ 0.0s

```
df.head()
```

[145] ✓ 0.0s

	Model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6.0	160.0	110.0	3.90	2.620	16.46	0.0	1.0	4.0	4.0
1	Mazda RX4 Wag	21.0	6.0	160.0	110.0	3.90	2.875	17.02	0.0	1.0	4.0	4.0
2	Datsun 710	22.8	4.0	108.0	93.0	3.85	2.320	18.61	1.0	1.0	4.0	1.0
3	Hornet 4 Drive	21.4	6.0	258.0	110.0	3.08	3.215	19.44	1.0	0.0	3.0	1.0
4	Hornet Sportabout	18.7	8.0	360.0	175.0	3.15	3.440	17.02	0.0	0.0	3.0	2.0

2.Display the number of rows and columns in the dataset.

```
print("No. of rows and columns in datset:",df.shape)
```

70] ✓ 0.0s

```
No. of rows and columns in datset: (33, 12)
```

```
print("No. of rows in the dataset:",df.shape[0])
```

71] ✓ 0.0s

```
No. of rows in the dataset: 33
```

```
print("No. of columns in the dataset:",df.shape[1])
```

72] ✓ 0.0s

```
No. of columns in the dataset: 12
```

3. Analyze the data using summary statistics such as Mean, Median, and Standard Deviation.

```
df.describe()
```

[152] ✓ 0.0s

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
count	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.0000
mean	20.090625	6.187500	230.721875	146.687500	3.596563	3.217250	17.848750	0.437500	0.406250	3.687500	2.8125
std	6.026948	1.785922	123.938694	68.562868	0.534679	0.978457	1.786943	0.504016	0.498991	0.737804	1.6152
min	10.400000	4.000000	71.100000	52.000000	2.760000	1.513000	14.500000	0.000000	0.000000	3.000000	1.0000
25%	15.425000	4.000000	120.825000	96.500000	3.080000	2.581250	16.892500	0.000000	0.000000	3.000000	2.0000
50%	19.200000	6.000000	196.300000	123.000000	3.695000	3.325000	17.710000	0.000000	0.000000	4.000000	2.0000
75%	22.800000	8.000000	326.000000	180.000000	3.920000	3.610000	18.900000	1.000000	1.000000	4.000000	4.0000
max	33.900000	8.000000	472.000000	335.000000	4.930000	5.424000	22.900000	1.000000	1.000000	5.000000	8.0000

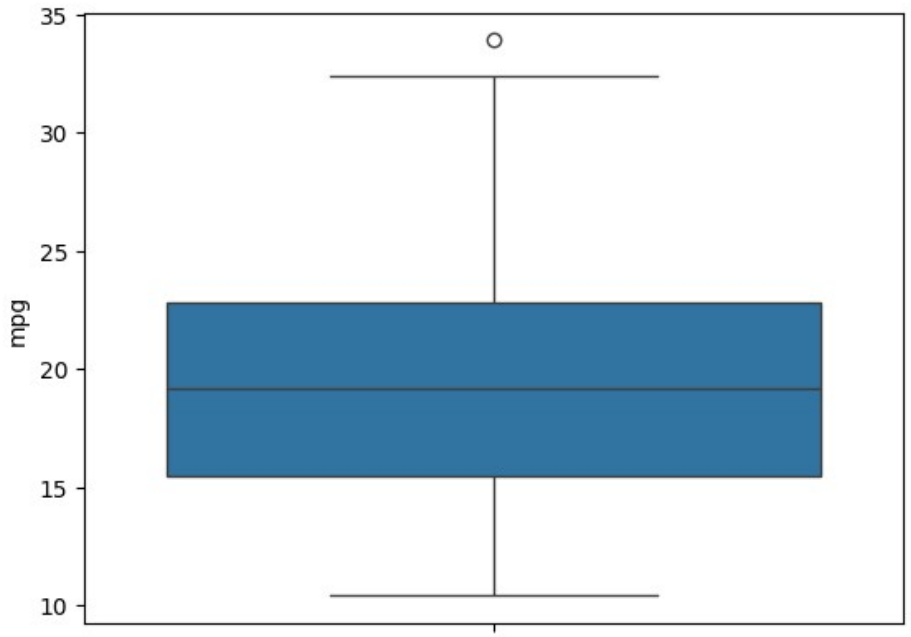
Mean vs. Median:

•The **mean** is sensitive to extreme values (outliers), as it takes into account all data points.
•The **median**, however, is resistant to outliers because it only considers the middle value(s).
•If the mean and median are close, it suggests that the data does not have large skewness or extreme outliers.

Standard Deviation (Std):

•A high standard deviation relative to the mean might indicate variability or outliers.
•In your data, the standard deviations are moderate, which aligns with the lack of outliers.

4. Provide interpretations for the 25th and 75th percentiles




```
def remove_outlier(col_name):
    sorted(col_name)
    Q1,Q3=col_name.quantile([0.25,0.75])
    IQR=Q3-Q1
    lower_bound=Q1-1.5*IQR
    upper_bound=Q3+1.5*IQR
    return lower_bound,upper_bound
```

[159] ✓ 0.0s

```
import numpy as np
low,high=remove_outlier(df1['mpg'])
```

[160] ✓ 0.0s

```
df1['mpg']=np.where(df1['mpg']>high,high,df1['mpg'])
```

[ ]

```
df1['mpg']=np.where(df1['mpg']<low,low,df1['mpg'])
```

[ ]

Range of Values:
•Highlight the actual values of the 25th and 75th percentiles for a specific column.
•This helps to understand the data spread within the lower and upper quartiles.

Interquartile Range (IQR):
•The difference between the 75th percentile (Q3) and 25th percentile (Q1) is called the **IQR**.
•The IQR measures the spread of the middle 50% of the data. You can explain whether the data is tightly clustered or widely spread based on this range.

Outlier Identification:
•Data points that fall below Q1−1.5×IQRQ or above Q3+1.5×IQRQ are often considered outliers.

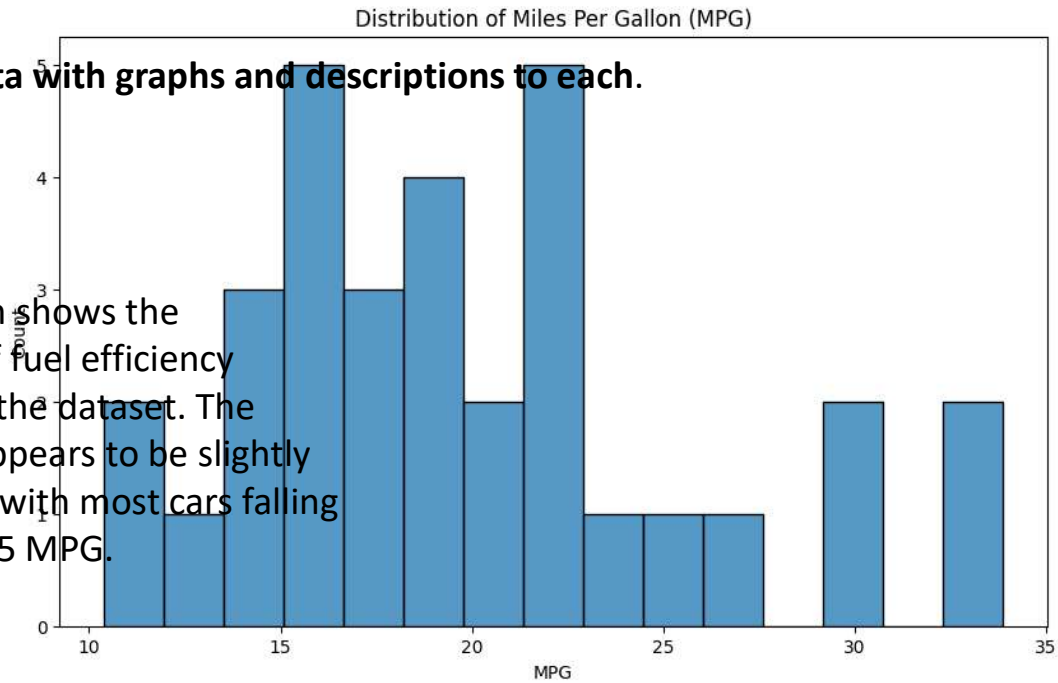
Practical Interpretation:
•Relate the percentiles to the context of the data. For example:
    •If analyzing car prices, the 25th percentile might represent budget models, while the 75th percentile might represent higher-end models.
    •For fuel efficiency, Q1 might represent less efficient cars, while Q3 might highlight more efficient ones.

ASSIGNMENT-3

1. Visualize the data with graphs and descriptions to each.

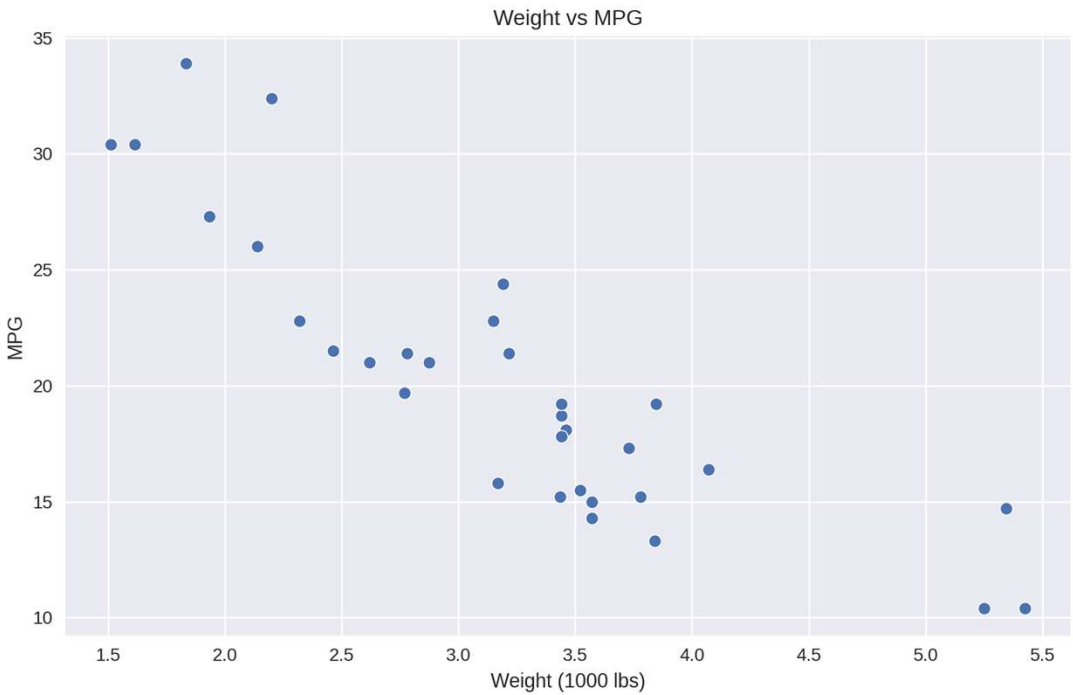
Count vs MPG

- This histogram shows the distribution of fuel efficiency (MPG) across the dataset. The distribution appears to be slightly right-skewed, with most cars falling between 15-25 MPG.



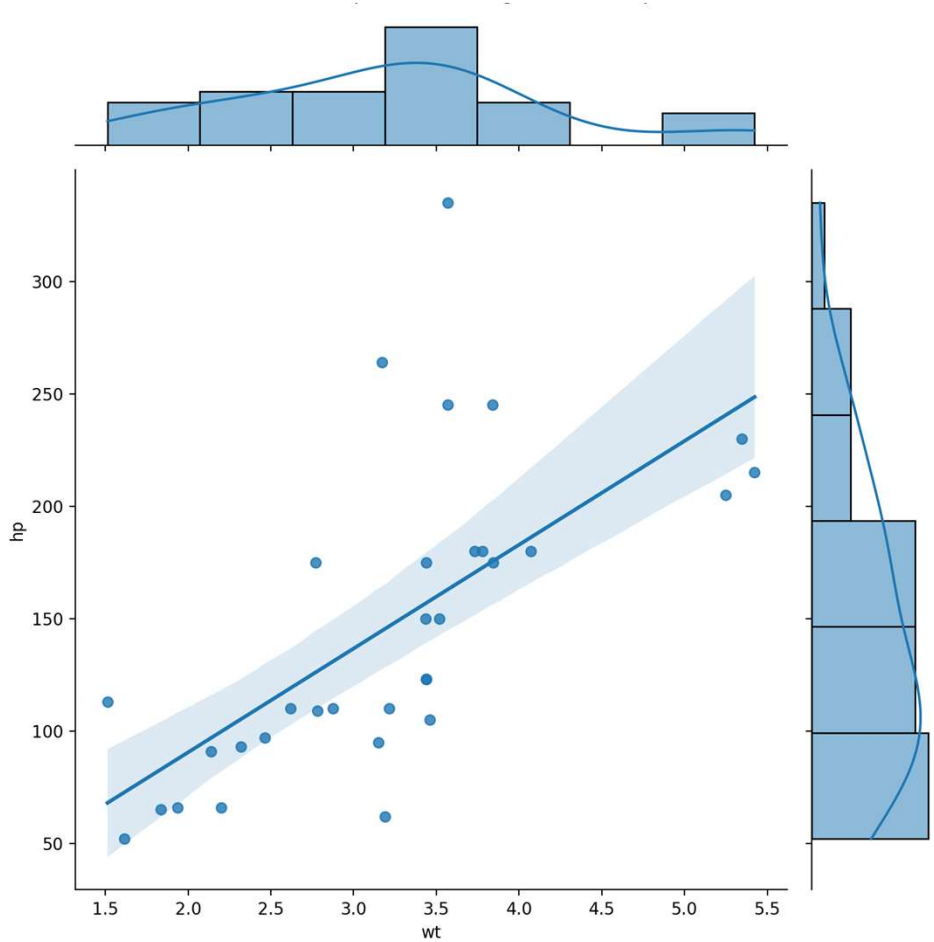
Weight vs MPG Relationship

- This scatter plot shows a strong negative correlation between weight and MPG - as weight increases, fuel efficiency tends to decrease.



Weight vs Horsepower Relationship

- Strong positive correlation between weight and horsepower
- As car weight increases, horsepower tends to increase
- The relationship appears to be roughly linear
- The marginal distributions show that both weight and horsepower have somewhat normal distributions with slight right skew



Distribution of Cylinders

- Most cars in the dataset have either 4, 6, or 8 cylinders
- 8-cylinder cars are the most common, followed by 4-cylinder cars
- 6-cylinder cars are the least common in this dataset