Week3 :- Exploratory Data Analysis (EDA)

• Preliminary step in data anlysis to :

→ EDA is an approach to analyze data in order to summarize main characteristics of the data

→ Gain better understanding of the data set

→ Uncover relationships bth variables

→ Extract important variables

Que : "What are the characteristics that have the most impact on the car price ?"

1) Descriptive statistics :- Which describe basic feature of a data set and obtain a short summary about the sample and measures of the data

2) GroupBy :- Basic grouping and how this can help to transform our data set

3) ANOVA :- the analysis of variance, a statistical method in which the variation in a set of observations is divided into distinct components

4) Correlation :-

5) Advanced correlation - statistics a) Pearson correlation b) correlation

Lect : Descriptive statistics

→ Describe basic features of data

short summaries about the sample and measures of the data

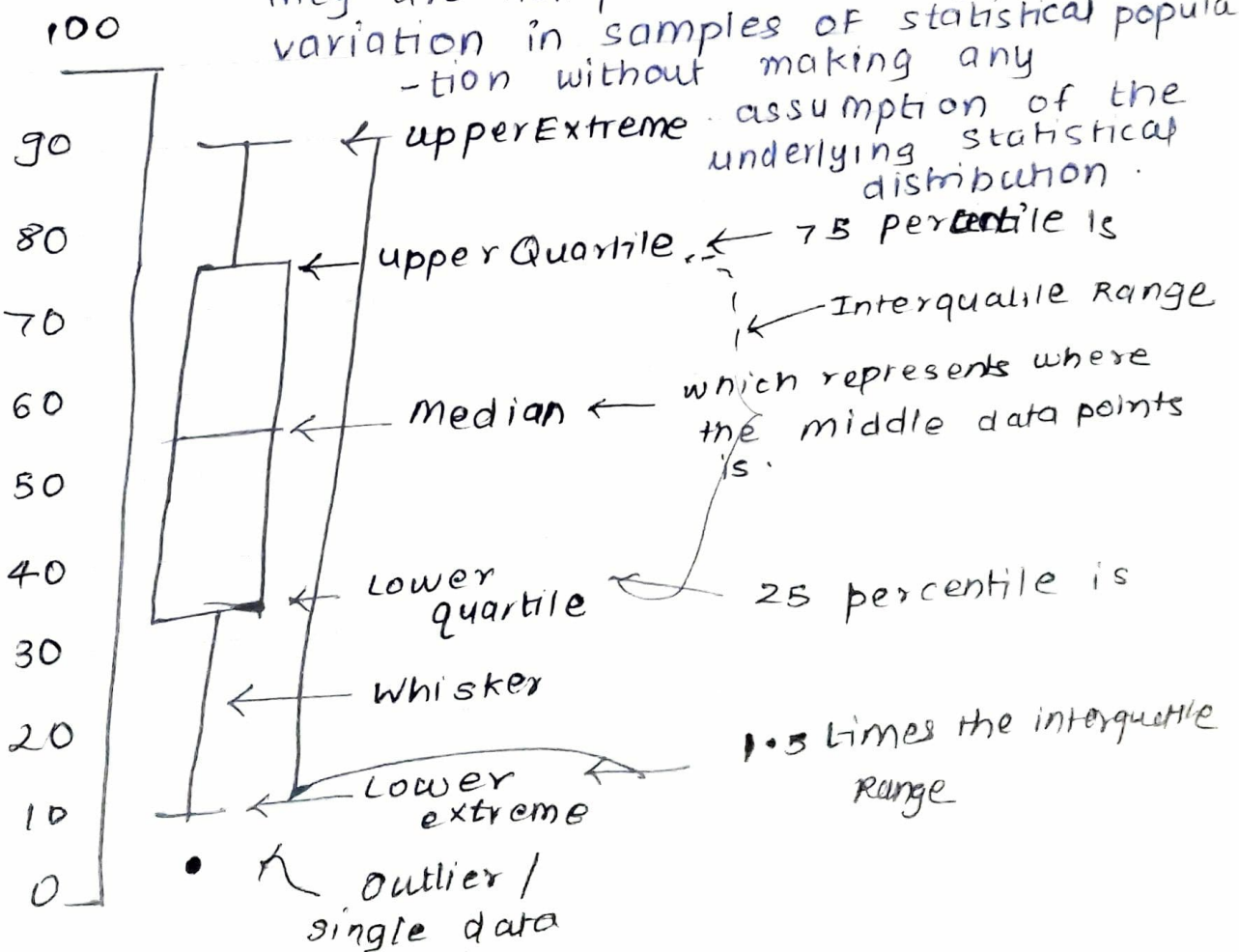
→ Summarize statistics using pandas describe() method, df.describe() → computes basic statics for all numerical variables

→ Summarize the categorical data is by using the value_counts() method

```
drive_wheels_counts = df["drive_wheels"].value_counts()
drive_wheels_counts.rename(columns=
  {'drive-wheels': 'value_counts'}, inplace=True)
drive_wheels_counts.index.name='driver-wheels'
```

→ **Box-plots** → can visualize the various distribution of the data

They are non-parametric; they display variation in samples of statistical population without making any assumption of the underlying statistical distribution.



eg.

```
sns.boxplot(x="drive-wheels", y="price", data=df)
```

c) Scatter Plot

→ Each observation represented as a point

→ Scatter plot show the relationship bt^n two variables

1. Predictor / independent variable - on x-axis
2. Target / dependent variables on y-axis

* matplotlib function

```
y = df["price"]
x = df["engine-size"]
plt.scatter(x,y)
plt.title("Scatterplot of Engine vs Price")
plt.xlabel("Engine size")
plt.ylabel("Price")
```

Outcome:-
^ Linear relationship bt^n Engine & price

Lect : GroupBy in Python

Grouping data

- Use Panda dataframe.Groupby() method:

→ Can be applied on categorical variables

→ Group data into categories

→ Single or multiple variables

groupby() - Example

```
df_test = df [['drive-wheels', 'body-style', 'price']]
df_grp = df_test.groupby (['drive-wheels',
                           'body_style'], as_index= False).
                                                  mean()
```

* Pandas method - Pivot()

- One variable displayed along the columns and other variable displayed along the rows.

```
df_pivot = df_grp.pivot (index='drive_wheels',
                         columns = 'body-style')
```

* Heatmap: It takes a rectangular grid of data and assigns a color intesity based on the data value at grid point.

→ Plot target variables over multiple variables.

```
plt.pcolor (df_pivot, cmap= 'RdBu')
plt.colorbar()
plt.show()
```

Lect : Correlation :- It is a statistical metric for measuring to what extent different variables are interdependent.

For eg:- i) Lung cancer → Smoking
ii) Rain → Umbrella

- Correlation doesn't imply causation

Correlation - Positive Linear Relationship

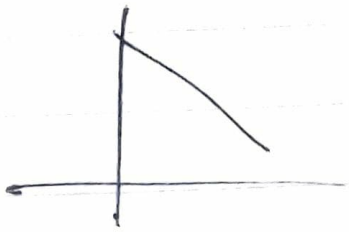
- Correlation bt^n two features (enegine-size & price)

```
sns.regplot (x="engine-size", y="price", data=df)
plt.ylim (0, )
```

Negative Linear Relationship

- Correlation bt^n two features (highway-mpg and price)

```
sns.regplot (x="highway-mpg", y='price",
                    data = df)
plt.ylim(0, )
```



→ Weak correlation bt^n two features

(peak-rpm and price)

```
sns.regplot (x='peak-rpm', y = "price", data=df)
plt.ylim(0, )
```

Lect Correlation - Statistics

Pearson correlation

→ measure the strength of the correlation between two features

- correlation coefficient
- P-value

1) close to +1 : Large positive relationship
close to -1 : Large Negative relationship
close to 0 : No relationship

2) P-value < 0.001 strong certainty in the result
P-value < 0.05 Moderate —11—
P-value < 0.1 weak —11—
P-value > 0.1 NO —11—

• Strong correlation:

→ Correlation coefficient close to 1 (or -1)

→ P value less than 0.001

Pearson Correlation Example

```
pearson_coef, p_value = stats.pearsonr(df['horsepower'], df['price'])
```

→ pearson correlation : 0.81 } strong +ve correlation

P-value : 9.35e-48

Correlation - Heatmap :- The color scheme indicates the pearson correlation coefficient, indicating the strength of the correlation bt^n 2 variables.

* Analysis of Variance (ANOVA)

→ Statistical comparison of groups

→ Eg: average price of different vehicles makes

* Why do we perform ANOVA?

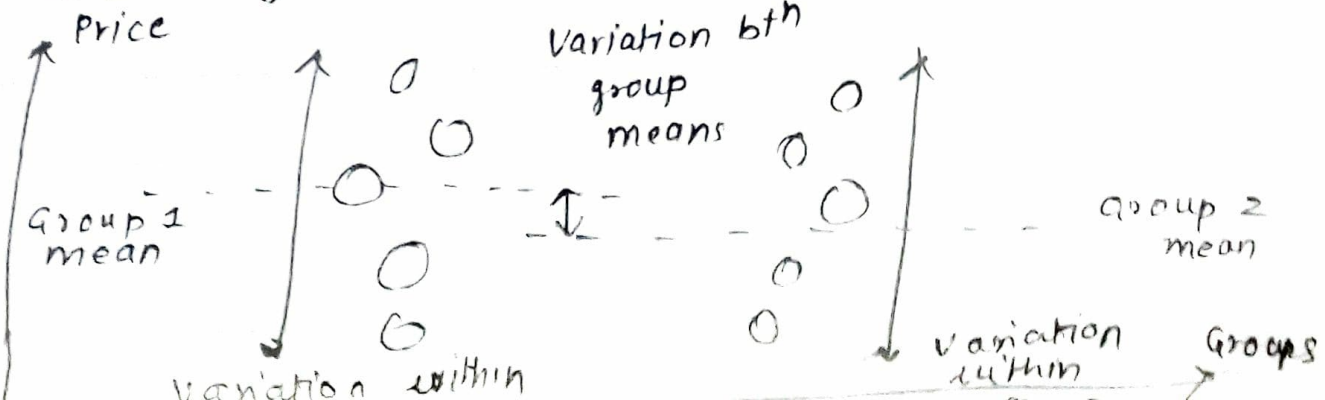
→ Finding correlation bt^n different groups of a categorical variable.

* What we obtain from ANOVA?

→ F-test score : variation bt^n sample group means divided by variation within sample group

→ p-value : confidence degree

• Small F imply poor correlation bt^n variable categories and target variable



→ Large F imply strong correlation btn variable categories and target variable.