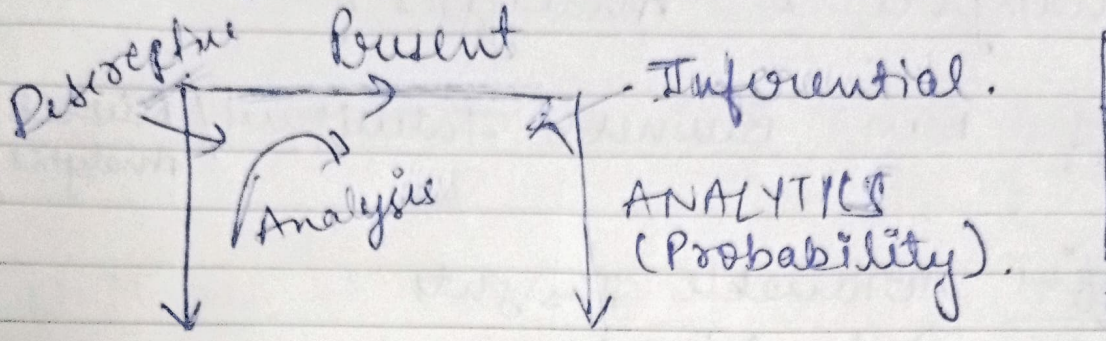
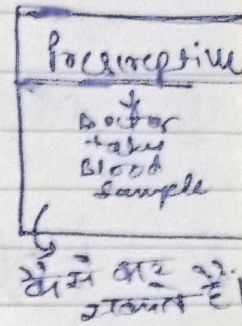
Workshop DAY-1

Statistics

31/10/22

① Inferential & Descriptive Statistics.





Variable → which tells characteristics of ~~data~~ particular item

Q What is Statistics? Different parts of Statistics.

Q How we can use Stats for Business Model?

Q How M-L & Deep Learning is used?

Types of Variables

- Qualitative (Categorical)
  - Object / factors
- Quantitative / Numeric
  - Integers
    - Discrete
  - Continuous (float or Decimal)

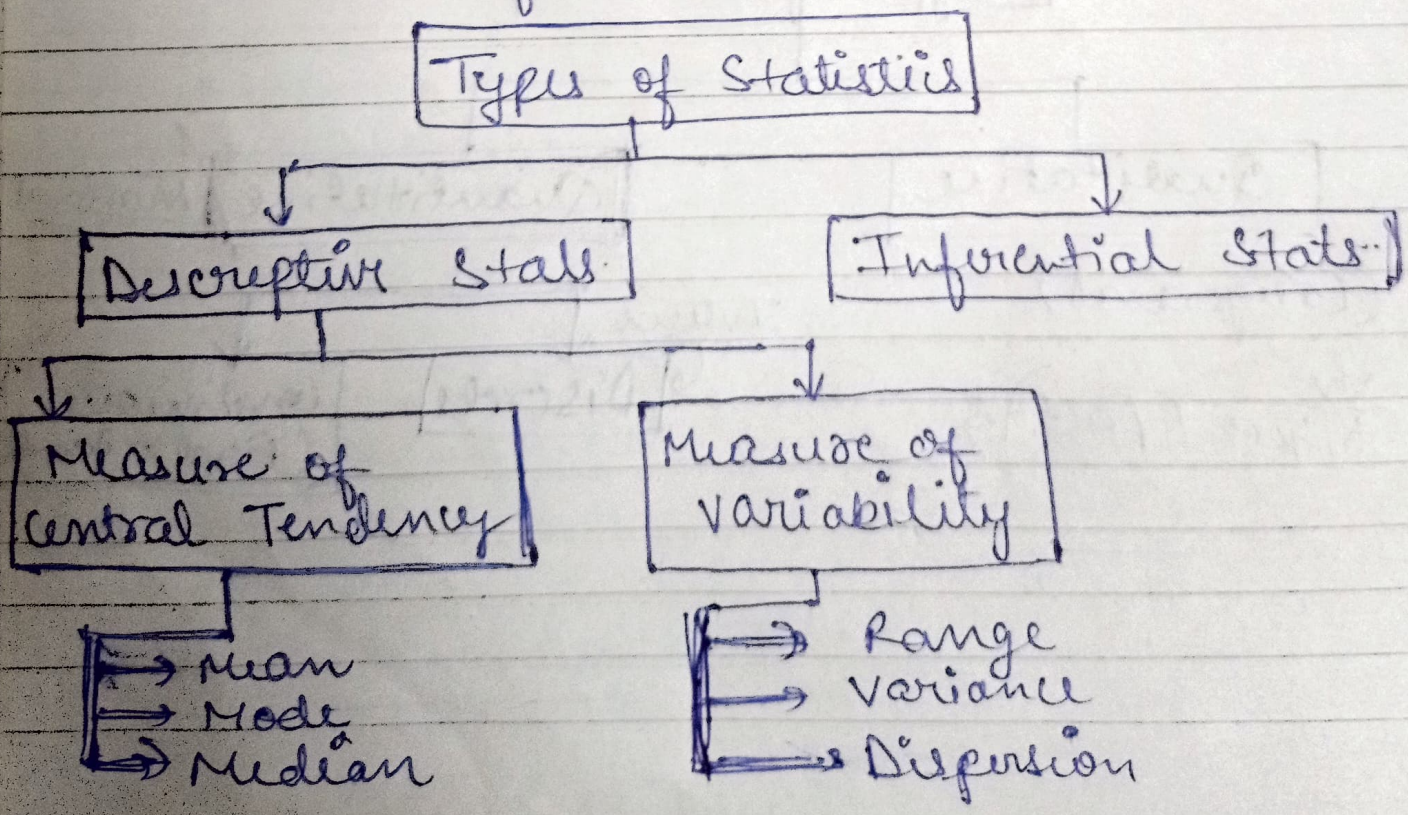
Q What is the difference between Analysis & Analytics?

Q Diff btw Business Intelligence / Business Analytics?

Q Diff Actionable Insights
Data Science
Machine Learning
Central Tendency

Q Different type of Data types?

Ans → Statistics simply means numerical data, or we can say, a field of math that deals with collection of data, tabulation, and Interpretation of numerical data.



(i) **Mean** → Average of all values in Sample Set

$$\text{Mean } (m) = \frac{\text{Sum of all terms}}{\text{Total no. of terms}}$$

(ii) **Median** → Data set is ordered from lowest to highest and then find exact "middle".

$$\text{Median} = a \quad b \quad c \quad d$$
$$= \frac{b+c}{2}$$

(iii) **Mode** → The value repeated most of time in central set.

$$\text{Mode} = a \quad b \quad c \quad a \quad f \quad a$$
$$= a$$

(iv) **Range** → Given measure of how to spread apart values in sample set or data set.

$$\text{Range} = \text{Max} - \text{min}.$$

(v) **Variance** → Describes how much a r.v defers from exp. value also computed as square of deviation.

$$\sigma^2 = \sum_{i=1}^{n} [(x_i - \bar{x})^2 \div n]$$

n → total data points
$\bar{x}$ → mean of " "
x_i → individual data points

(vi) Dispersion → Measure of dispersion of set of data from its mean

$$\sigma = \sqrt{(1 \div n) \sum_{i=1}^{n} (X_i - \mu)^2}$$

✱ Inferential Statistics :-

→ It makes inference & prediction about population based on a sample data taken from population

→ It generalizes a large dataset & applies probabilities to draw a conclusion.

→ Types :-

1. one sample test of difference
2. Confidence Interval
3. Contingency table & chi-Square stat
4. T-test or Anova
5. Pearson correlation.
6. Bi-variate Regression
7. Multi-variate Regression

Ans-2

Mistakes in Business can cost millions of dollars, Statistics can explain why a past event happened and predict what is likely to happen in future.

Ans 4

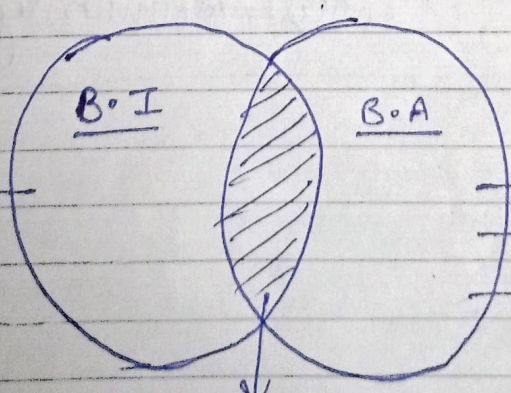
ML can be used as :-

1. Real-time chatbot agents.
2. Decision Support.
3. Image Recognization
4. Speech Recognization
5. NLP
6. Sentiment analysis.
7. Dynamic pricing tactics
8. fraud detection

Ans 5

Analysis	Analytics
1. Collection → Manipulation → Examination of data for deep insights	1. Taking Analyzed data & working on it in a Meaningful way.
→ Open-Refine, Rapid Miner, Node XL, Tableau etc.	→ Python, SAS, Tableau, Excel etc.

Ans 6



Descriptive Analysis

- Diagnostic Analytics
- Predictive Analysis
- Prescriptive Analysis

- Collect, Analyze & visualize data
- Identifies pain Points
- Generate reports.

Ay 7

→ Actionable Insights → Conclusions drawn from data that can be turned directly into an action or response.

→ Central Tendency → the statistical measure that identifies a single value as representation of an entire distribution

Median → Most Imp. component,

Skewness → left or right standard deviation

-ve skew | +ve skew

- row → Observations
- column → Variables / Parameters.

Outliers → which are exceptional with other data sets or values.

- 0 skewness → Normal Distribution (Symetric Distribution)
(Poisson Distribution)

→ Correlation → Multicollinearity (other name).

count
mean
std
min
25%
50%
75%
max

To revise

- Outliers → formulas.
- Quartiles & InterQuartiles
- Correlation & Covariance → Difference.
- Boxplot.
- Scatterplot

Codes :-

1. mt['mpg'].var()
2. mt.head() → first five observations.
3. mt.tail() → Last 5 observations.
4. mt['cyl'].count() → Total.
5. mt['cyl'].value_counts() → no. of counts.
6.

8	14
4	11
6	7

7. Type of Analysis :-

- **Univariate** — Analysis of single variable
  - # Best Representation
    - barplot
    - boxplot
    - Histogram.
- **Bivariate** — Analysis of 2 var.
  - Histogram / Barchart
  - Line Graph
  - Scatterplot
  - # Best Statistical Method
    - correlation
    - chi Square test
- **Multivariate**
  - Scatterplot matrix
  - hyperbox
  - Trellis Display
  - Parallel coordinates
  - Starplots

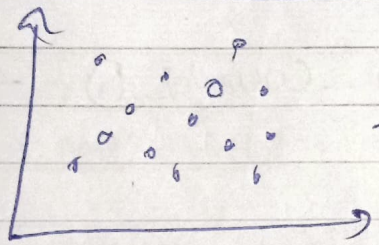
☆ Descriptive Statistics

→ Scatter plot
→ Pearson's correlation coefficient
→ Simple linear.

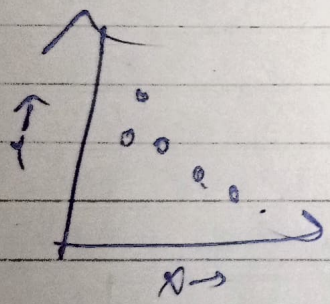
+Correlation — relationship b/w 2 variables, (−1, +1) → close to (1,1) are considered as strong.

(Diagonal Matrix)

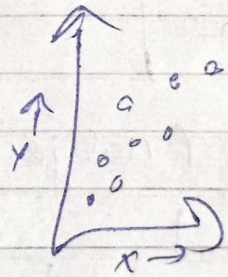
± 0.4 se neeche we consider No-relation.



→ No particular Direction.



−ve correlation



+ve correlation

☆ Pearson's coefficient.

☆ Simple Linear Regression :-

$$Y = \Lambda + \beta X + e$$

or

$$Y = mx + c$$

Dependent variable — slope — constant. — coefficient

I Simple Linear Regression.

II Multivariate.

Best Visualization

#1 `mt['cyl'].value_counts().plot.bar()`.

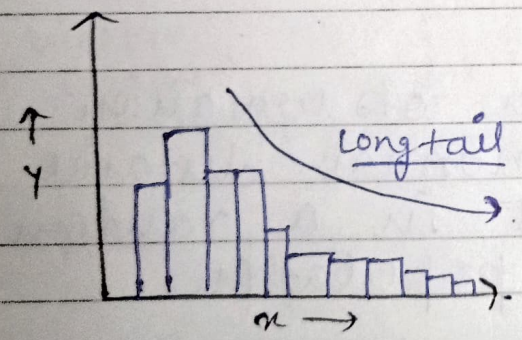
#1 `mt['cyl'].plot.box()`.

Skewness : It is a measure of the asymmetry of a distribution.

can have :-

1. Right (or +ve) Skewness
2. Left (-ve) Skewness
3. **zero Skewness** → Normal distribution or poisson distribution
   - mean = median

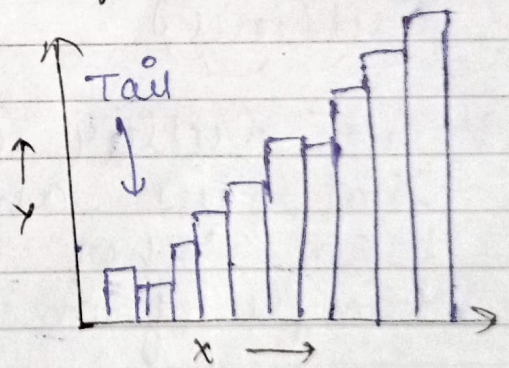
Right Skewed (+ve)



mean > median

→ Longer on the right side of its peak.

Left Skewed (-ve)



mean < median

→ Longer on the Left side of its peak.

Pearson's Median Skewness = $3 \times \frac{(\text{Mean} - \text{Median})}{\text{Stand. deviation}}$

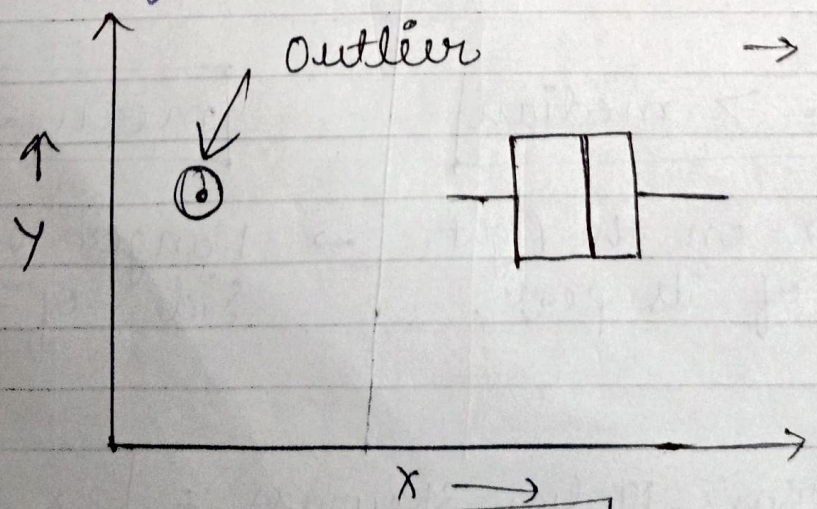
Transformation based on type of skew,

Type of Skew	Intensity of skew	Transformation
Right	Mild	Do not transform
	Moderate	Square root
	Strong	Natural log
	Very strong	Log base 10
Left	Mild	Do not transform
	Moderate	Reflect * Sq. root
	Strong	Reflect * Nat. log
	Very Strong	Reflect * $\log_{10}$

Reflect means, take largest observation k subtract each observation from (k+1)

☆ Outliers

→ An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



→ May contain Imp. Information

BOXPLOT REPRESENTATION

Correlation

→ Aka Diagonal matrix (-1, +1)

→ It is a statistical measure that expresses the extent to which two variables are linearly related.

→ Either they are directly proportional ($A \propto B$) or Indirectly proportional ($A \propto \frac{1}{B}$) or Maybe they don't have Any Relationship.

Q How is it measured?

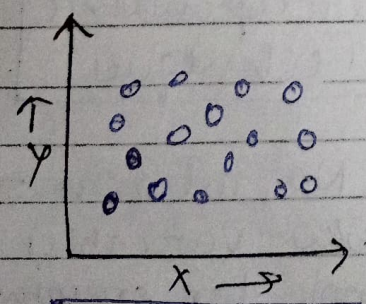
→ The Sample correlation coefficient, r, quantifies the strength of the relationship.

$-1 < r < 0$ | $0 < r < 1$

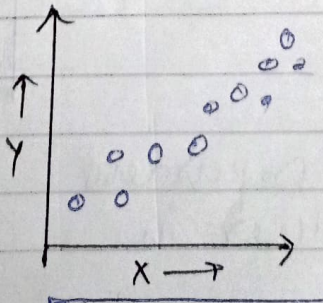
- $r \to 0$ ($r = 0$ or $r \to 0$): Weaker Linear Relationship
- $r \to +ve$: Value of Both Variables Increase Together
- $r \to -ve$: Value of one Variable ↑ & other ↓.

Code :→ `mt.corr()`

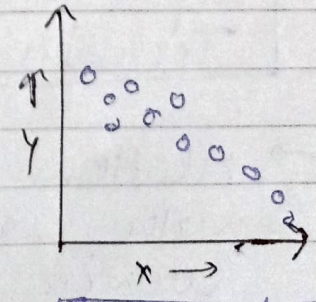
∴ r value below ±0.4 we consider No Relation.



No Relation



+ve correlation



- ve correlation

Pearson Correlation Coefficient →

Definition : The Pearson Corr. Coff. 'r', is the most common way of measuring a linear correlation.

→ It is a Number between -1 & +1 that measures the Strength and direction of the relationship between two variables.

☆ Multi-variate Analysis

→ Includes all Statistical techniques that are used to analyze more than two variables at one.

→ Aim is to find pattern & correlations between several variables Simultaneously.

Analysis Techniques

Dependence Techniques	Interdependence Techniques
→ Relations are dependent with each other in certain way	→ No relation b/w Variables → Look for grouping them.

→ Techniques :-

1 dipnt var → 2 or more Ind vars.

① Multiple Linear Regression

b/w 2 → Spam/Not Spam

② Multiple Logistic Regression

③ Multi-variate analysis of Variance (MANOVA)

④ Factor Analysis

⑤ Cluster Analysis.

Multiple Linear Regression :-

→ It is dependence Method which looks at the relationship b/w one dependent & two or more Independent variables.

Multiple Logistic Regression :-

→ It is used to Calculate (predict) the probability of a binary event occuring.

→ Only two possible Outcomes, either (1) or (0). like Spam & Not Spam.

MANOVA :-

→ MANOVA is used to measure the effect of multiple independent variables on two or more dependent variables.

Factor Analysis :-

→ It is an Interdependence technique which seeks to reduce the Number of variables in a dataset.

→ Model with too many variables can suffer over-fitting in which model fits too closely, making it less distinguishable to future datasets.

Cluster Analysis :-

→ It is an Interdependence technique used to group similar items within a dataset into clusters.

→ Clusters can be classified in 2 types :-

(1) Intracluster distance, Distance b/w data points within One cluster.

(2) Intercluster distance, Distance b/w data points in diff. cluster.

Machine Learning Algo.

✱ Simple Linear Regression :-

→ It is a statistical method that allows us to summarize and study relationships between two continious (quantitative) variables

→ x → 'predictor', 'explanatory' or 'feature' or 'Independent Variable'

→ y → 'response', 'outcome', 'dependent variable'

→ ~~Bad~~ Approach for predicting a response using a single feature.

Steps of doing it :-

1. Analyze the correlation.
2. Keep changing θ_0 & θ_1 or coeffient & interupt.
3. Minimal Learning Rate
4. Obtain Global Minima
5. Have best-fit line.

Residual Errors :- $y_i - h(x_i)$

Equation Basic :- $y = mx + c$

or

$y' = w_1 x_1 + b$

or

$h_\theta^{(x_i)} = \theta_0 x_0 + \theta_1 x_i$

Squarred Error or Cost fun^n

$J(\theta_0, \theta_1) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ ← predicted

★ Quartiles :- Set of values that divide a list of numerical data into three quarters.

$$Q_1 = \left[\frac{(n+1)}{4}\right] \text{th item}$$

$$Q_2 = \left[\frac{n+1}{2}\right] \text{th item}$$

$$Q_3 = \left[3.\frac{n+1}{4}\right] \text{th item}$$

General formula,

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_2 - l_1)$$

$Q_r \rightarrow r^{th}$ Quartile
$l_1 \rightarrow$ lower limit
$l_2 \rightarrow$ upper limit
$f \rightarrow$ frequency
$c \rightarrow$ cumulative freq. of class preceding the quartile class.

$$\text{Quartile deviation} = \frac{(Q_3 - Q_1)}{2}$$

→ Inter Quartile Range (midspread) :- (IQR) is the diff b/w upper & lower quartile.

$$IQR = Q_3 - Q_1$$

PANDAS PROFILING LIBRARY

(11) .. Changing Object dtype into Numeric,

```
lu['Gender1'] = lu['Gender'].map({'male':0,
                                  'female':1})
```

Scale of Measurement

1. Nominal Eg → Eye colour, size.
2. Ordinal Eg → Customer Service (★★★★★)
   - In Proper order
3. Interval
4. Ratio

Q How will you decide that working data-set is Categorical & Numerical.

QUARTILES

→ Quartiles set or split the data into 4 equal parts, first quartile Q1, divides the smallest 25% of the values from the other 75% that are larger.

→ The second Quartile Q2 is the median i.e, 50% of the values are smaller than the median & 50% are larger.

→ The third Quartile Q3 divides the smallest 75% of the values from the largest 25%.

Q → What is Regression

Q → ML & its types

Q → Stand. Error, SI, deviation

Q → Dependent & Independ Variable

We can Minimise the Skewness :- (or balance)

1. Scaling
2. Log Transformation.
3. Pre-processing.

---

* Scales of Measurment, tells us how precisely variables are recorded.

→ Nominal : Data can only be ① 'categorized'.
Eg → (Gender, Car brands)

→ Ordinal : Data can be ① 'categorized' & ② 'Ranked'.
Eg → (Top 5 olympic Medalist)

→ Interval : Data can be :
Eg → (Test scores)
① → Categorized
② → Ranked
③ → Evenly Spaced.

→ Ratio : Data can be :
Eg → (Height, Age, weight)
① → Categorized
② → Ranked
③ → Evenly Spaced
④ → Has a Natural zero.

Q How to decide working data set is Categorical or Numerical ?

Mainly based on 'OBJECTIVE', [illegible]

Ans Categorical data means, data type that can be stored and identified based on the names or labels given to them.

→ A process called 'MATCHING' is done, to draw out the similarities or relations between the data and then they are grouped accordingly.

* Numerical data refers to data that is in the form of numbers and not in any language or description form.

Q. What is Regression ?

Ans It is a technique that attempts to determine the strength & character of the relationship between one dependent variable and a series of other variables. (a.k.a Independent variables).

Q M.L and its types ?

Ans Machine Learning is defined as Study of computer programs that leverage algorithms & statistical models to learn through inference

and patterns without being explicitly programmed.

Types of M.L

- SUPERVISED Learning → (Task-driven)
- UNSUPERVISED Learning → (Data-Driven)
- REINFORCEMENT Learning → (Learn from Errors)

Q What is 'Standard Deviation' & 'Standard Errors'?

Standard Deviation	Standard Error
① Denotes variability within the Sample.	Denotes variability among multiple samples from a population.
② Can Reffered as 'Sample Statistic'	Can be Reffered as 'Population Parameter'
③ Measures dispersion of individual value from their mean value.	measures accuracy of Sample mean as an estimate of the population mean.
④ It is type of ('Descriptive Stats.')	It is type of ('Inferential Stats.')

⑤ Calculated by square rooting the variance.	Calculated as $\frac{SD}{\sqrt{(\text{Sample size})}}$
⑥ (cannot be) Used to calculate confidence interval of population	Used to calculate the same
⑦ Not a function of sample size	depends on the size of sample, S.Z↑, S.E↓

Sample Size $\propto \frac{1}{S.E}$

Q "Dependent" & "Independent" Variables

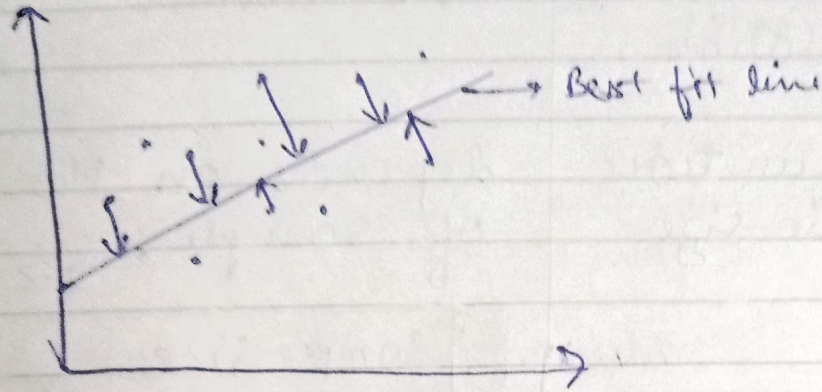
Ans
- "Dependent" → Effect (JOB SATISFACTION)
- "Independent" → Cause (TYPE OF WORK ENVIRONMENT)

Independent variables a.k.a
- → Explanatory variables
- → Predictor variables
- → Right-hand-side variables (appear on right hand side of regression Eq.)

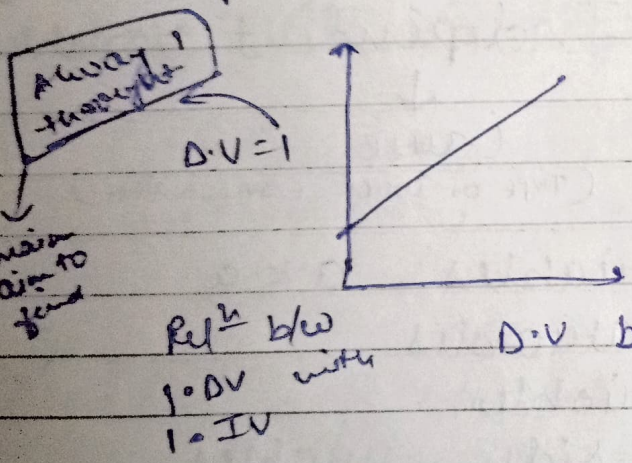
Types of I.V

- Experimental Variables. (can be manipulated Directly)
- Subject Variables (cannot be manipulated Directly).

Dependent Variables a.k.a
→ Response Variables
→ Outcome Variables
→ Left-hand-Side-Variables.



Simple L.R



Rel^{n} b/w 1 D.V with 1 I.V

D.V behaviour → continuous / Decimal / float.

Expression on which SLR works, = OLS
↓
Ordinary ~~ordinary~~ least Square.