FDS & ML

6/8/24

Data Science Process:-

1) Define the problem Statement.
2) Data acquisition.
3) Data preparation - Apply EDA [Exploratory Data Analysis]
4) Model planning (Data finalization)
5) Visual Communication. {Visualization}
6) Model Deployment

- Instance based (future selection)
- Model-based.

- Titanic data

Know your data (Commands)

df - dataframe

- Data size — df.shape {row & column of data}
- Data look like — df.head()
- Detail of data — df.info()
- Calculate null value — df.isNull().sum()
- Look mathematically — df.describe()
- Checking duplicates — df.duplicated().sum()
- Checking correlation b/w columns — df.corr()

Type of Data
- Categorial
- Numerical.

1) Categorial Data → EDA univariate {Exploratory Data Analysis}
2) Numerical Data.

Categorial Data
- Pie chart ⟹ df['column name'].value_counts().plot(kind='pie')
- Count chart ⟹ sns.countplot(df['column name'])

__/__/__

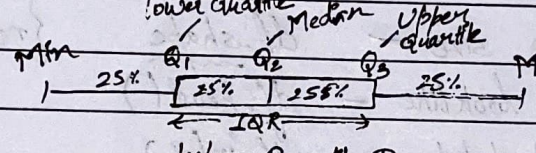
Types of Data

- **Categorial (Qualitative)**
  - Nominal (Gender, colour, city)
  - Ordinal (Rating, Grade)
- **Numerical (Quantitative)**
  - Discreate
  - Continuous
- **Time & Date**
- **Image**
- **Audio**
- **Video**

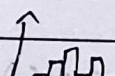
- PassengerId – Nominal
- Survival – nominal
- Survived – ordinal
- pClass – Ordinal
- name – nominal
- sex – nominal
- age – continuous
- SibSp – Discreate
- Parch – discreate
- Ticket – nominal
- Fare – continuous
- Cabin – nominal
- embark – nominal

Numerical univariate

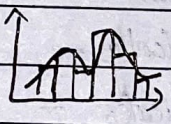
→ Box plot ⇒



→ Histogram ⇒



↳ Dist plot = histogram + PDA (Probability density .....)



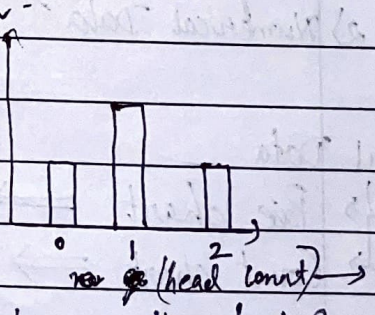
- sns.histplot(df['Age'])
- sns.distplot(df['Age'])

Probability – It is the major of likelihood of an event occurring.

1) • Random variable
2) • Probability distribution – description of possible value of each random variable.

Case: Prob of getting head

Sample space: {(H,H) (H,T) (T,T) (T,H)}



3) Conditional probability – If first event occurs, then prob of second event occurs.

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A/B)}{P(A)}$$

4) Baye's theorem

Statistics : Science of collecting, analysing, interpretating, presenting & organizing data.

Types:-

1) Descriptive statistic – Set of data is defined.
2) Inferential statistic – Sample space is not defined.

Hypothesis ... – statistic also helped in hypothesis.

. Correlation – relationship bpw 2 entities.

∴ Regression – Model defining relationship bpw dependent variable and one or more independent variable.

Bivariate / Multivariate

- heat map → sns.heatmap
  pd.crosstab
  sns.distplot(titanic['Survived']==0 ..., hist = false).
- Scatterplot → sns.scaterplot(x="Survived", y="Age")

Commands :- (Titanic dataset)

1) Age > 29 — df[df['Age'] > 29]
2) Drop a column — df1 = df.drop(columns = ['Age'])
   df2 = df.drop(3) ← Index no/Column no.
3) Delete rows which have null value — df.dropna(inplace = True)
4) Change datatype of column — df['Amount'] = df['Amount'].astype('int')
5) Update value of a column — df[2, 'Age'] = (23) → column value
6) To find null values. — df.isnull().sum()

inplace = True ⇒ will drop column from the current dataframe & store in it.

Tendency (Measures of the center).
- The mean (the average value)
- The median (the mid point value)
- The mode (the most common value)

Spread (Measures of Variability) ← min & max. variance

Moment. Variance → $\Sigma(x-\bar{x})^2/(N-1)$ Mean(μ) = $\Sigma_{i=1}^{N} x_i / N$

Standard Deviation → $\sqrt{\text{variance}}$

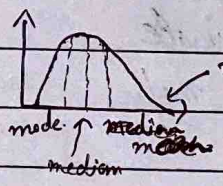
m, 2, 8, 4

Moment

momentum 1.	Mean	Central tendency	
2.	~~Standard Deviation~~ Variance (Volatility)	Dispersion	
3.	Skewness	Symmetry (+ve or -ve)	$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{x_i - x}{\sigma}\right]$ ③
4.	Kurtosis	Shape (Tall or flat)	$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{x_i - \bar{x}}{\sigma}\right]$ ④

Types of Skewness

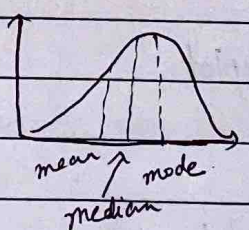
1) Right skewness (+ve skew)



mode < median < mean.

LN RP

2) Left skewness (-ve skew)



mean < median < mode.

Types of Kurtosis

1) Mesokurtosis: An excess kurtosis of 0. Normal distributions are mesokurtosis.
2) Platykurtosis: Neg. excess kurtosis. few outliers
3) Leptokurtosis: Positive excess kurtosis. Many outliers.

__/__/____

Types of Statistics → Descriptive
→ Inferential

→ Analysis of Variance (ANOVA) - involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions. Useful with categorial data.
(how the categorial value is related to the result)
↳ within / between

→ Regression Analysis - Involves modelling the relationship b/w a dependent variable & one or more independent variables. For ex:- predicting the sales of a product based on advertising expenditure.

→ Confidence intervals - involves estimating the range of values that a population parameter could take based on a sample of data. ex. estimating the population mean height

→ Chi - square Test - Test independence or association (correlation) b/w 2 diff categorial variables ex. Testing gender & occupation are indep. or not.

→ Sampling techniques - Involves ensuring that the sample of data is representative of the population. ex. random sampling to select individuals from a population

→ Bayesian Statistics - Alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. for ex:- updating the probability of a disease given a positive test results.

* Weighted mean - sum / weight of product.
* Trimmed mean

Measure of Dispersion -

Coefficient of variation = (Standard Deviation / mean) x 100 %

__/__/__

Types of histogram

Symmetric | Bimodal | Uniform | ...

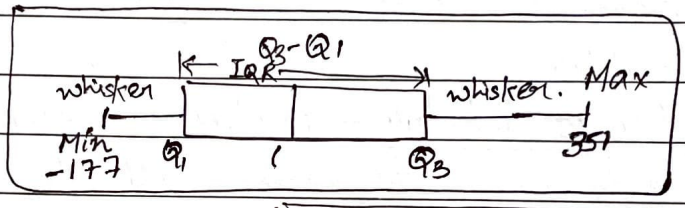
Types of Bivariate.

1) Categorial – categorial → Contingency Table / Crosstab
2) Numerical – numerical → Scatter plot
3) Categorial – Numerical → contingence.

*Distribution Type

Box Plot

(odd Index)	Sorted data in ascending order.
1	6
2	10
3	11 ⇐ $Q_1\left(\frac{11+31}{2}=21\right)$
4	31
5	35
6	50
7	65 ← Q_2
8	75
9	100
10	105
11	201 ⇐ $Q_3\left(\frac{105+201}{2}=153\right)$
12	700
13	1000



$Min = Q_1 - 1.5 \times IQR$

$= -177$

$Max = Q_3 + 1.5 \times IQR = 351$

Outliers = 700, 1000.

$Q_1 = 21$

$Q_2 = 65$

$Q_3 = 153$

$IQR = 153 - 21 = 132$

Even Index	
1	20
2	30
3	33 → $Q_1 = 34$
4	35
5	37
6	40 → 426 → Q_2
7	46
8	50
9	100
10	300 → $Q_3 = 200$
11	800
12	700

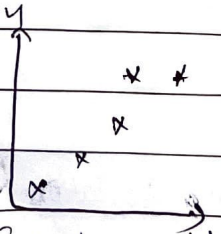
$IQR = 166$ (153 − 27 = 132)

$Min = -215$

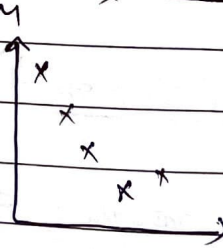
$Max = 449$

Outlier = 500, 700

Correlation:

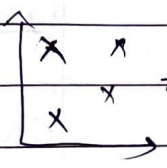


Positively correlated.



negatively correlated

300 − 33 = 267



Data cleansing

- Standardising format
- Removing redundant value
- Removing null values
- Transforming feature.
- Removing structural errors.

__/__/____

Data Cleansing

Handling Noisy Data
- Binning → Clustering
- Regression → Combined computer & human inspection.

Losless & Lossy Compression

TA Syllabus

- Data science process
- Types of data.
- Basic command of python "know your data"
- Importance of Data Preprocessing
- Confusion matrix
- univariant analysis.
- Bivariant analysis.
- Momentum (1,2,3,4)
- Steps in Data Preprocessing
- Types of distribution (uniform, normal etc).
- Python code

Confusion Matrix (Main ko confusion hai)

y ŷ

Actual \ Predicted	1	0	
1	TP	FN	P
0	FP	TN	N
	↑	↑	

Prediction: 1 → Positive, 0 → Negative
Actual: Same → True, Difference → False

P = TP + FN
N = FP + TN

True positive Rate = TP/P ↓ False Positive Rate = FP/P
True negative Rate = TN/N ↓ False Negative Rate = FN/N

Accuracy = TP + TN / (TP + TN + FP + FN)

TPR = Recall ?

P = 4
N = 6

y	ŷ
0	0
1	1
2	2
2	2
1	1
0	1
1	1
2	1
0	1
1	1

Precision = TP / (TP + FP) ?

TP = 3
TN = 3
FN = 3
FP = 1

Actual \ Prediction	0	1	2
0	1	2	0
1	0	4	0
2	0	1	2

	1	0
1	2	3
0	2	3

__/__/____

Define a funct^{n} for following conversion.

x	x.
112233	11-22-33
445566	44-55-66

```
def con(v):
    return x[:2]+'-'+x[2:4]+'-'+x[4:]
```

11-22-33	112233
44-55-66	445566

```
def con(v):
    return x[:2]+x[3:5]+x[6:]
```

Remove all char except number =>

11,223$	11223
@55`667	55667