Notes

Statistics

life Cycle of Datascience

(1) **DATA SCINCE**
- Requirement Gathering

service Based company
TCS, HCL

Product Based company
Google
Youtube, GPay
Gmail, Apple

Business Analyst

Product Manager

SALES

Product Manager (Project Manager)

Domain Expertise

Product Manager

Data Science Project
- Requirement Gathering

DATA ANALYTICS Team

Requirement Gathering	→	Data Analyst Data scientist

Domain Knoweledge
product Manager

Business Analyst

Product Manager

Project Manager

(1) Data Ananys
(2) Data scientis
(3) Big Data Engineer
(4) Cloud Engineer

```
┌─────────┐   ┌─────────┐   ┌─────────┐
│ Internal│   │ 3rd party│   │web scrapin│
│ DataBase│   │ API's    │   │          │
└─────────┘   └─────────┘   └─────────┘
```

mysql, Nos ql

```
┌──────────────┐        ┌─────┐
│  Big  Data   │  ──→   │     │
│    Engineer  │        │     │
└──────────────┘        └─────┘
```
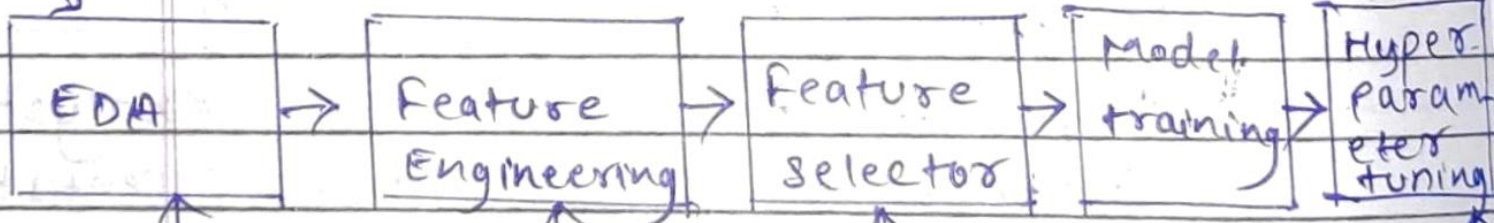
life cycle of DS project

Exploratory Data Analysis

improve the model

```
┌─────┐   ┌──────────┐   ┌──────────┐   ┌─────────┐   ┌─────────┐
│ EDA │ → │ Feature  │ → │ Feature  │ → │ Model   │ → │ Hyper-  │
│     │   │Engineering│  │ selector │   │ training│   │ param   │
│     │   │          │   │          │   │         │   │ eter    │
│     │   │          │   │          │   │         │   │ tuning  │
└─────┘   └──────────┘   └──────────┘   └─────────┘   └─────────┘
```
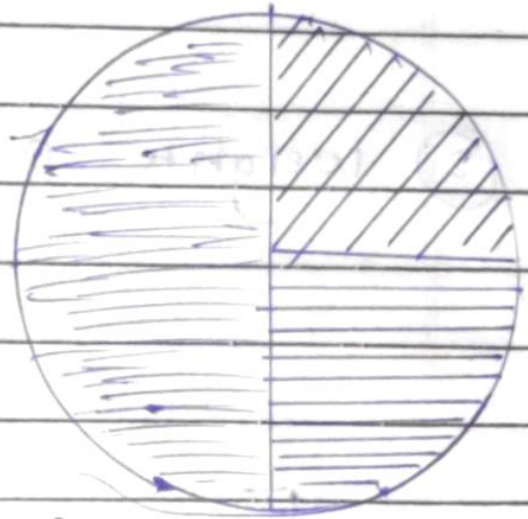
training with
ML Algorithms

statistics

Analysis of DATA
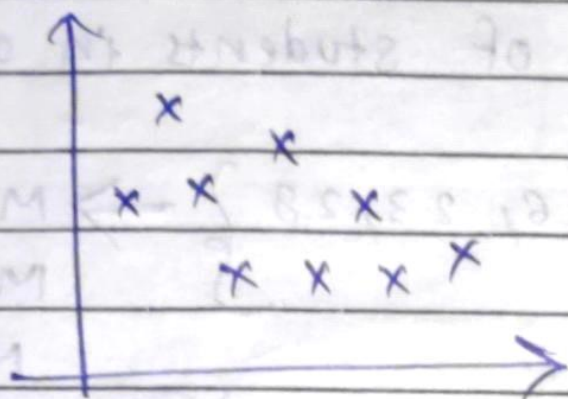
Descriptive stats

↑
summarizing the
Data

Descriptive ←
Stats

$Age = \{12, 13, 14, 18, 20, 25\}$ pie

⇊

Average Age

⇊

Descriptive stats.

Statistics Defination :-

statistics is a science of collecting, organizing and analyzing the data

Data → It is a facts or preces of infor mations.

Example → ① Ages of students in a class.

$$\left\{ 24, 25, 26, 23, 28 \right\} \Rightarrow \text{Mean,} \\ \text{Median,} \\ \text{Mode} \\ \left[\text{standard deviation} \right]$$

② weights of student in class.

statistics

Descriptive stats [Eda +f E]
① It consists of organizing   Inferential
and summarizing the   stats.
data.
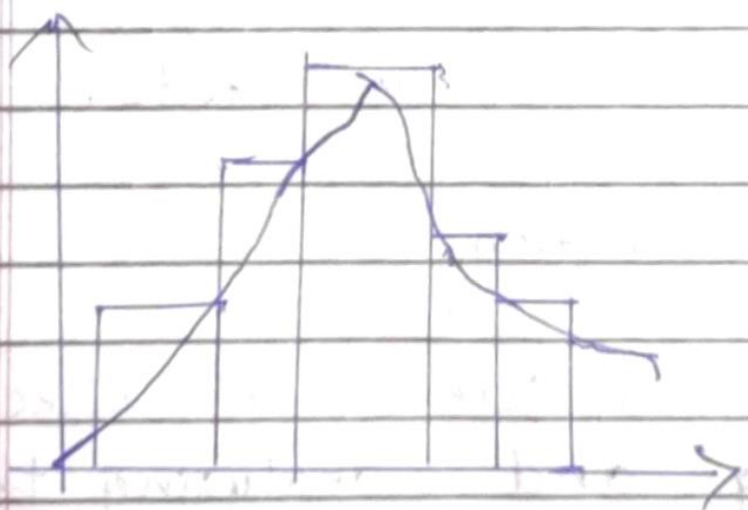
Histogram   Bar chart

Pie → Distribution
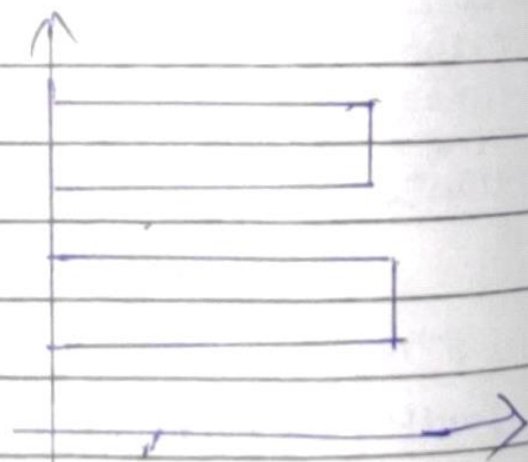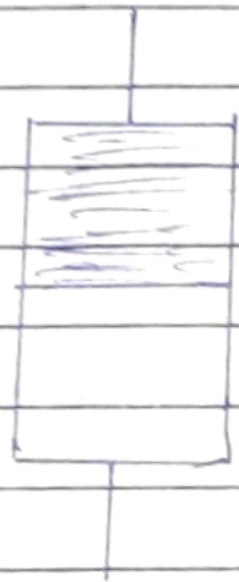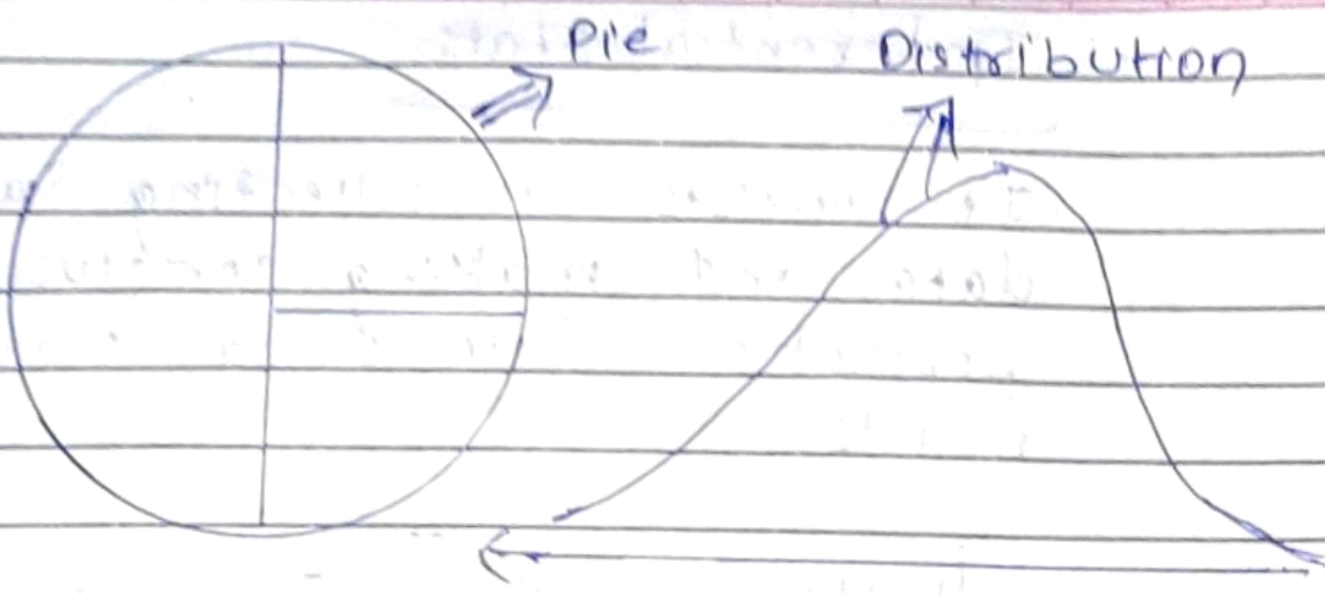
→ candle stick   Box plot
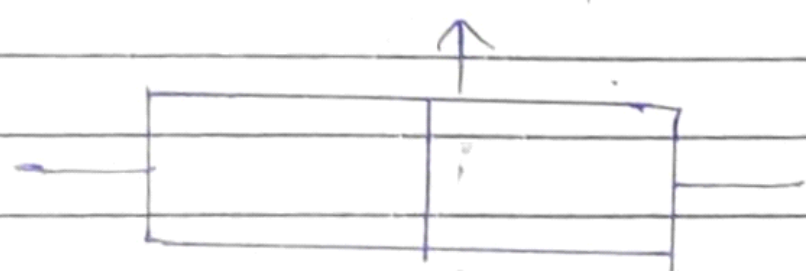
x = ~~Heig~~ Height
y = weight
(Height)   (weight)
—   —
—   —
—   —
—   —
—   —
—   —
—   —

y

x↑   y↑

Inferential stats

It consists of collecting Sample data and making conclusion about population data using some experiments.

[making conclusion] → Hypothesis Testing.

~~univisity~~

university ——→ 500 people

[Class A]
⇓

[sample data] ⇒ [Age] ⇒ Average age of the entire university
⇓

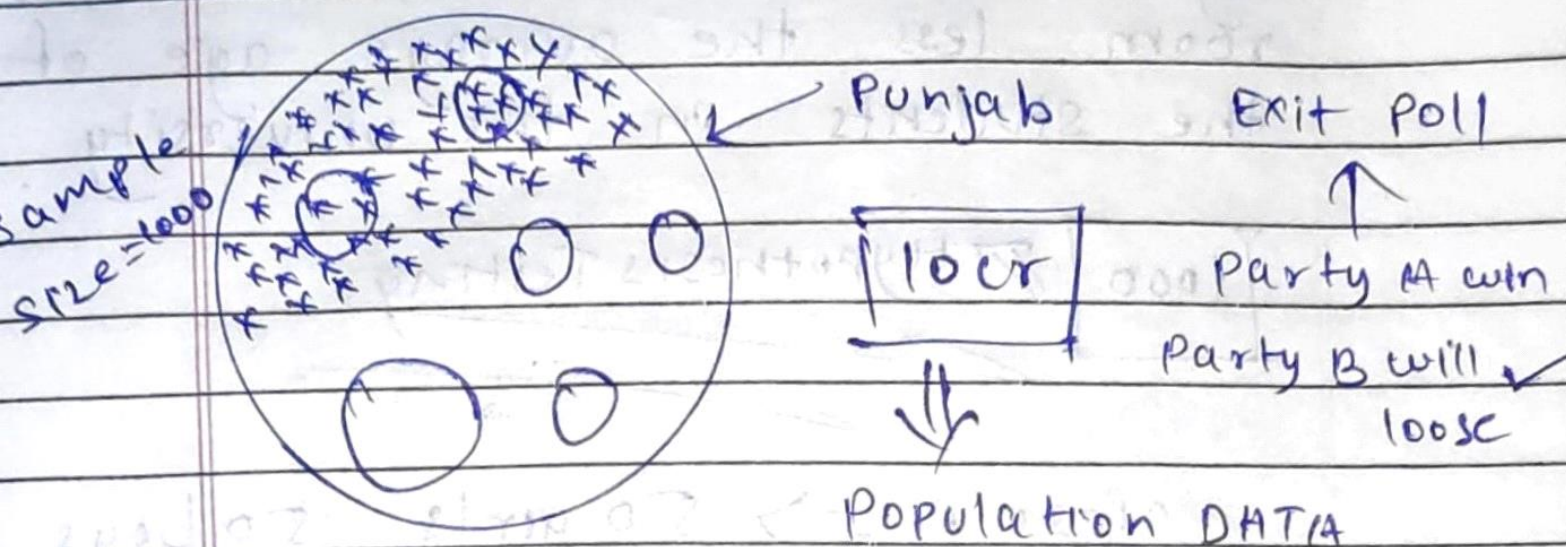Hypothesis Testing

C.I ⇒ confidence Interval

P-value

① z test
② t test
③ chisquare test
④ F test

Sample Data vs population Data

Punjab

10 cr

Population DATA

Exit Poll

↑

Party A win

Party B will
loose

Eg:- lets say there 20 classrooms in a university and you have collected the age of students in one classroom.

Ages = { 21, 20, 18, 34, 17, 24, 25, 26,

23, 22 }

weight { — — — — — ... — — } — ... }

Descriptive stats :-

what is the average age of students in the classroom?

Relation between Age & Gender

Inferential stats := Are the average age of the students in the class-room less the average age of the students in the university

$\boxed{1000}$ $\not\pi$ Hypothesis Testing

class A $\Rightarrow$ 50 girls   50 boys

⇓   ⇓

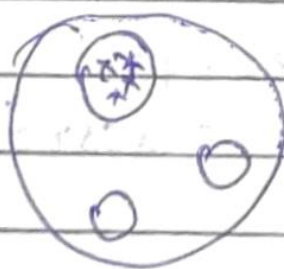choose a sample 95%   92%

↑

Sampling Techniques

Population (N)   Sample (n)

Ⓜ simple Random Sampling :-

Every member of the population (N) has an equal chance of being selected for your sample (n)
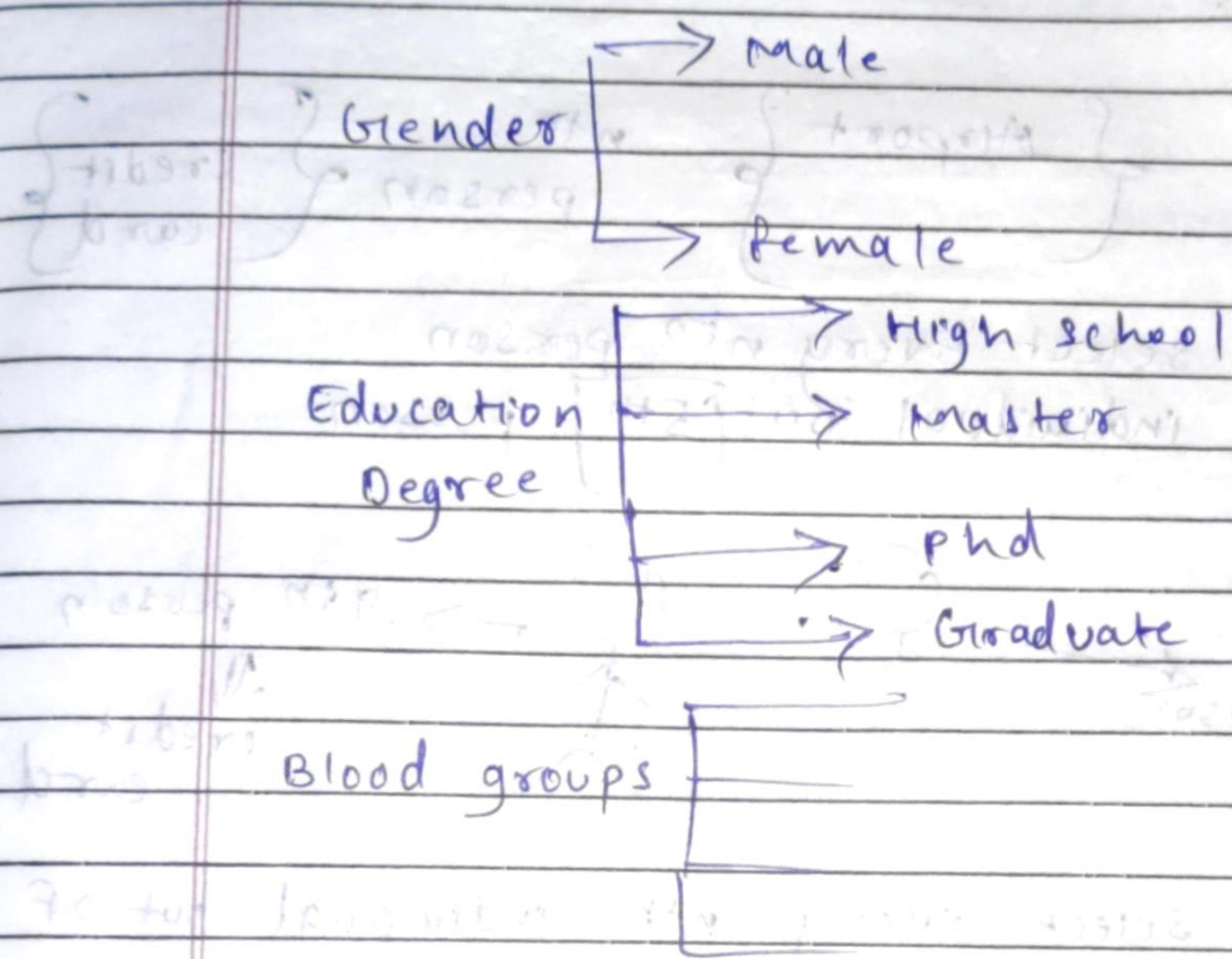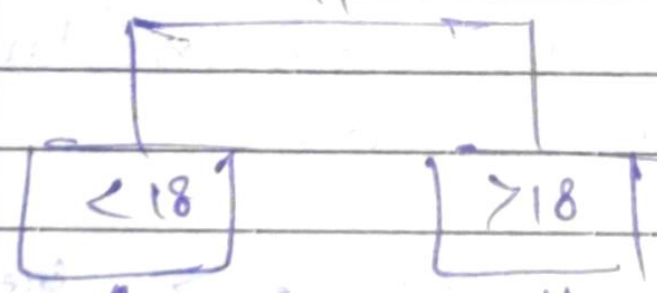
$\boxed{n = 1000}$

$\boxed{<18}$ $\boxed{>18}$ → Random sampling
average

(2) **stratified Sampling :**

Gender → Male
        → female

Education Degree → High school
                → Master
                → phd
                → Graduate

Blood groups

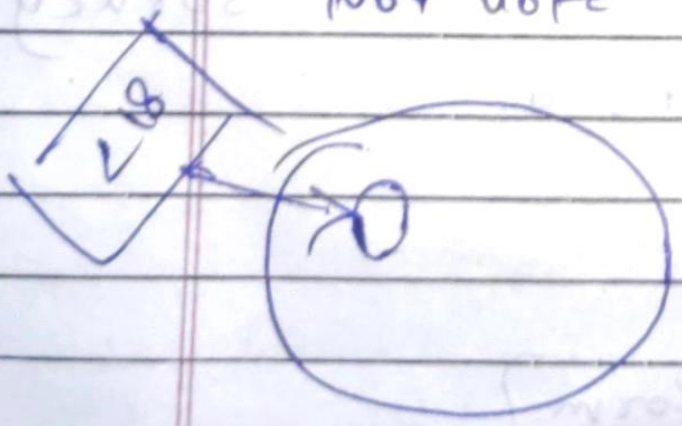Population

{ Exit
  poll }

< 18   > 18

Not vote   Random sampling

(3) Systematic Sampling :

$\{$ Airport $\}$ ⟶ n^{th} person $\{$ Credit Card

select Every n^{th} person
individual but [5th] person

5th person ⟵

⟶ 9th person
↓
credit car
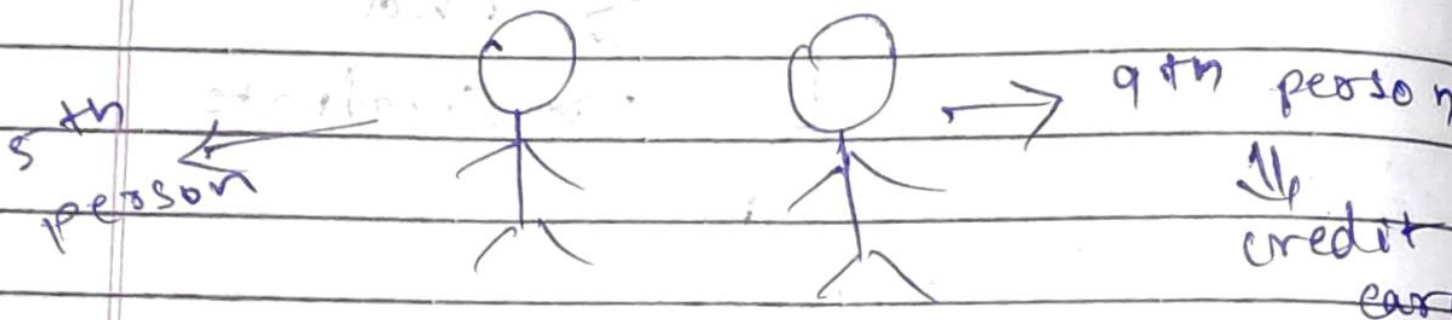
select every n^{th} individual out of
Population (N)

(4) Convenience Sampling - only those
who are interested in the sur
will only participate

$\{$ Data Science Survey ⟶ General
⇑   AI
Ineuron job for a   Survey
Specific
⇓
$\{$ fill the form $\}$

Examples :-

(1) Survey Regarding New Technology
⇓
convenience Sampling

(2) RBI survey ⟹ women only [married women]

(3) credit card :- stratified + Random sampling

(1) Variable :- Variable is a property that can take any values

Eg :- age = 14   Variables
      age = 25
      age = 100   Ages = [24, 25, 26, 27, 28, 29]
                                    ⇓
                              collection

Two Different types of variables :-

(1) Quantitative Variable → Measured Numerically { Mathematical operation }
    Ex :- Age, weight, height, rainfall (cm)

(2) Qualitative Variable :- categorical variables { Based on some characteristics they { are group together }
    Eg :- Gender, Types of flower, Type of movies

Quantitative Variables

① Discrete variable
Eg ÷ [whole number]

↓
fixed

① E.g ÷ No. of Bank Account
$\{1, 2, 3, 4, 5\}$

② No. of children ÷ whole number
③ pincode = fixed

② Continuous variable ÷

Continuous [Continuous] → Decimal value

Eg ÷

① Height, weight, ages, Rainfall, speed

                    → Married
Marital
                    → Not Married
                         ↓
                    categorical variable

Continuous ← continuous variable (top right)

Assessment

(2) what kind of variable is :-

Marital status $\longrightarrow$ categorical variable

Movie duration $\longrightarrow$ continuous variable

Pincode $\longrightarrow$ Discrete variable

IQ $\longrightarrow$ Discrete variable

Pancard $\longrightarrow$ categorical variable