

# **HOW TO GET DATASETS FOR MACHINE LEARNING**

# What is Dataset?

A dataset is a collection of records usually presented in tabular form.

Name	Age	City	Department
Sam	25	Jodhpur	Digital Marketing
Sharon	27	Jaipur	Developer
Jack	32	Pune	SEO
Mark	42	Mumbai	Content Writer
Diana	50	Hyderabad	Trainer

# Types of Datasets

**Numeric Data  
(Quantitative)**

**Categorical Data  
(Qualitative)**

**Ordinal Data**

# NUMERIC DATA (QUANTITATIVE)

Numeric data, also known as quantitative data, is data that you typically present in number form, and it doesn't include any language or descriptive form.

It's always measurable, and we can add it together.

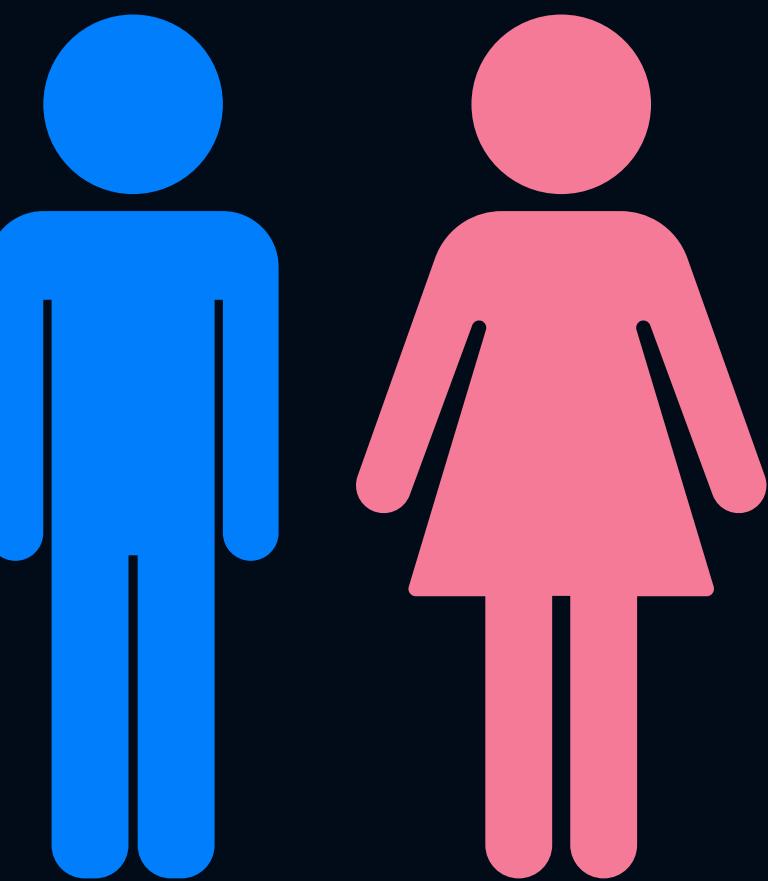
Example: Age, Blood Pressure, Temperature etc

5°C

# CATEGORICAL DATA (QUALITATIVE)

Categorical data is a type of qualitative data that is described using words instead of numbers. It can be grouped into categories, rather than being measured numerically.

Example: Gender etc.



# ORDINAL DATA

Ordinal data is a type of qualitative data that organizes variables into ordered categories.

The categories are ranked based on a hierarchical scale, like high to low.

Example: Food Rating.

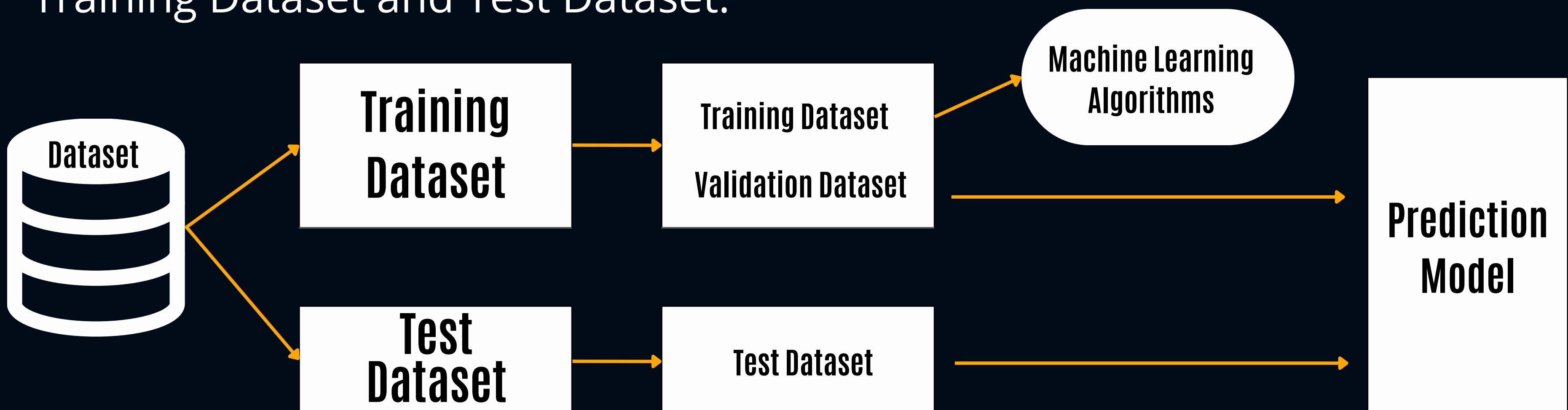
FOOD HYGIENE RATING



VERY GOOD

# Need Of Dataset

In the development phase of ML projects, datasets are classified as Training Dataset and Test Dataset.



kaggle.com/c/dogs-vs-cats/data

# Dogs vs. Cats

Create an algorithm to distinguish dogs from cats

Overview Data Code Models Discussion Leaderboard Rules

Files 3 files

Size 853.96 MB

Type zip, csv

License Subject to Competition Rules

Learn at your own pace

Learn from the best professionals

Meet expert teachers

Share knowledge

www.kaggle.com

Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic.

Learn more. Ok, Got it.

datasetsearch.research.google.com/search?src=0&query=cat%20vs%20dog&docid=L2cvMTF0YzUyd3Rfeg%3D%3D

Google

cat vs dog

Last updated Download format Usage rights Topic Provider Free Saved data sets

100+ data sets found

Cat\_and\_Dog

Cat and Dog

Bingsu/Cat\_and\_Dog

Explore at: [huggingface.co](#) [Kaggle | kaggle.com](#)

5 scholarly articles cite this dataset ([View in Google Scholar](#))

H

Cat\_and\_Dog  
huggingface.co  
kaggle.com  
Updated Oct 2, 2023

Google Dataset

+3more

Updated Sep 21, 2019

Learn at your own pace

Learn from the best prof

Meet expert teachers

on Hugging Face and contributed by the HF Datasets community

Share knowledge

Type here to search

Dataset Search - G...

19°C Haze 19:50 16-01-2024

# Government Dataset

**INDIA**

<https://data.gov.in>

**USA**

<https://data.gov/>

**EUROPEAN  
UNION**

<https://www.opendatani.gov.uk/>

Northern  
Ireland

<https://www.data.europa.eu/euodp/data/dataset>

# What is Data Pre-processing?

Data Pre-processing is a process of converting the raw data in suitable form.

The steps of Data processing are following:

1. Getting Dataset
2. Importing Libraries
3. Importing Datasets
4. Finding missing values
5. Encoding Categorical Data
6. Splitting Dataset into Training Dataset and Test Dataset
7. Feature Scaling

# Features and Labels in Dataset

In ML, Feature means property of Training Data.

In ML, Label means the output we get from our model after training.

1	Person	Height(in Feet)	Weight (in kg)	Foot Size(in inches)
2	Male	6	81	12
3	Male	5.92	86	11
4	Male	5.58	90	12
5	Female	5.92	77	10
6	Female	5	45	6
7	Female	5.5	68	8
8	Female	5.52	58	7
9	Female	5.75	68	9

# Features and Labels in Dataset

Case 1: The prediction (Y) Calories here is a label.

Calories is the column that we want to predict using various features like

X1: Gender

X2: Height

X3: Weight

1	Person	Age	Height(in Feet)	Weight (in kg)	Duration	Heart_Rate	Body_Temp	Calories
2	Male	68	6	94	29	105	40.8	231
3	Male	70	5.92	79	5	88	38.7	26
4	Female	20	5.44	60	14	94	40.3	66

# Features and Labels in Dataset

Case 2: The prediction (Y) Heart\_Rate here is a label.

Heart\_Rate is the column that we want to predict using various features like

X1: Gender

X2: Weight

1	Person	Age	Height(in Feet)	Weight (in kg)	Duration	Heart_Rate	Body_Temp	Calories
2	Male	68	6	94	29	105	40.8	231
3	Male	70	5.92	79	5	88	38.7	26
4	Female	20	5.44	60	14	94	40.3	66

# Identify Features and Labels



# Regression

Regression is a statistical method that helps us to understand and predict the relationship between variables.

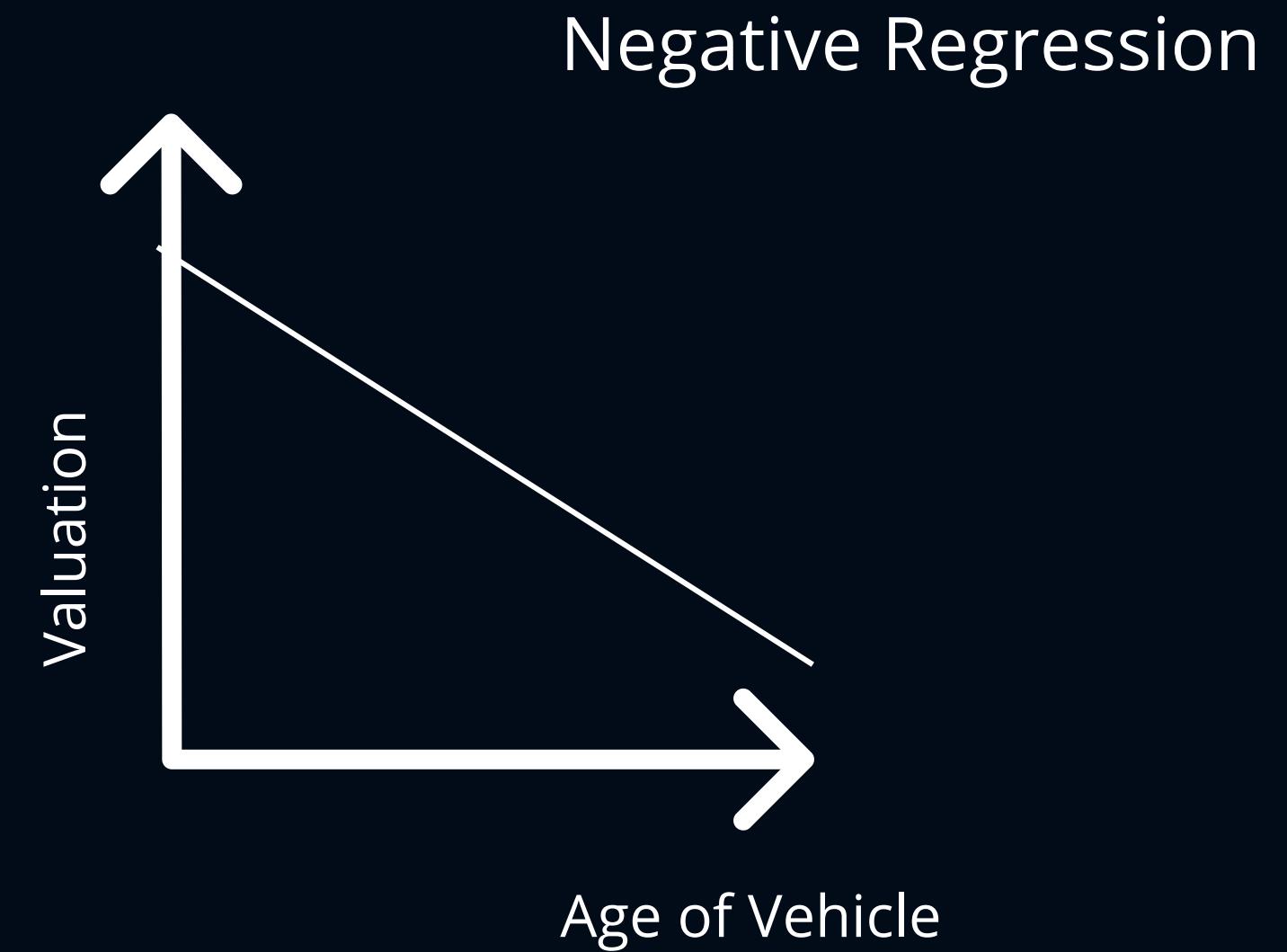
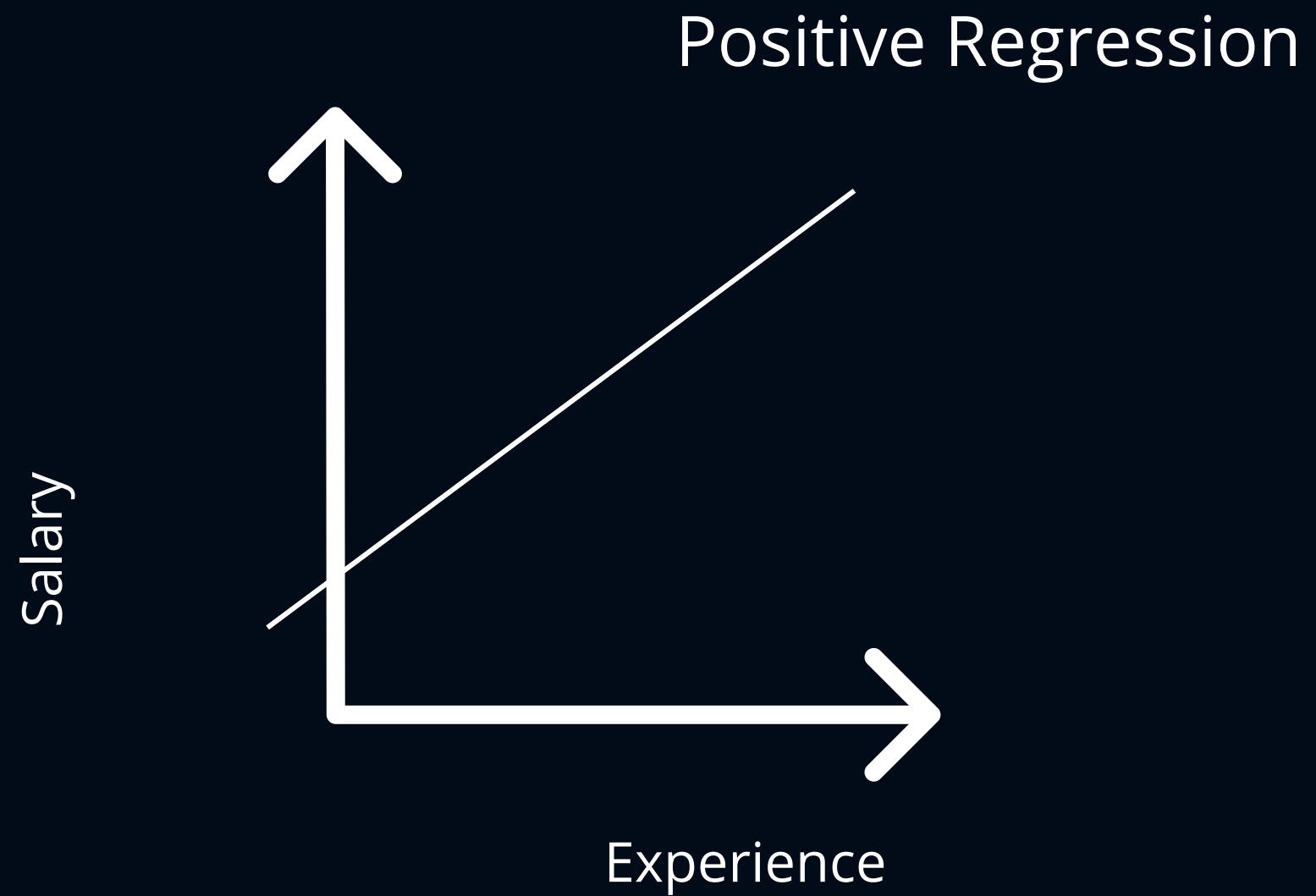
Describe how one variable (dependent variable) changes as another variable (independent variable) changes.

**Dependent variable:** Trying to predict or explain (Y)

**Independent variable:** That are used to predict or explain the changes in the dependent variable.

**Example:** Predict the salary on basis of experience.

# Regression-Positive/ Negative



# Types of Regression

**Linear  
Regression**

**Multi Linear  
Regression**

**Polynomial  
Regression**

# Linear Regression

Equation of Linear Regression :  $Y = mX + b$

Here,  $Y$  represents dependent variable.

Here  $X$  represents independent variable.

Here  $m$  represents the slope of the line ( How much  $Y$  changes for a unit change in  $X$ )

Here  $b$  represents the intercept ( the value of  $Y$  when  $X=0$ )

Problem: Predict the prize of Pizza

Phase 1: Data Collection

Phase 2: Calculations

Phase 3: Prediction

Phase 4: Visualization

# Linear Regression

Diameter (X)	Price (Y)	Mean (X)	Mean (Y)	Deviations (X)	Deviations (Y)	Product of Deviations	Sum of Product of Deviations	Square of Deviations for X
8	10							
10	13							
12	16							

$$m = (\text{sum of product of deviations}) / (\text{sum of square of deviations for } X)$$

$$b = \text{Mean of } Y - (m * \text{Mean of } X)$$

