Data Scientist -> Understands the Business

Data Understanding

Data Prepare

Data Model

Data Evaluation

Deploy

Monitor

Optimize

Data Analyst

Required Information Meta data

Data Collection

Assigned Goal

Pre-Determined

Data Cleaning

Data Prepared

Exploratory data Analyst

Inference

Model Deployment

Data Visualization

Report Writing

1. Data Types
   1. Qualitative
      1. Categories Data
         1. Nominal
         2. Ordinal
   2. Quantitative
      1. Numeric
         1. Discrete
         2. Continuous

Type of Data Analyst

Descrptive -> what Happened what said by data -> Exploring the data

Diagnostic 🡪 why did it Happen -> Pattern

Predictive -> what will happen -> ML DL

Prescriptive -> how can we make it happen -> Data Driven Decision making

Life Cycle

Data Acquire -> Create Capture Gather

Clean -> Organize filter Annotate clean

Use -> Analyze Mining Model Data Driven Data Decision Act Visualization

Publish ->Share Workflow Portal

Store

**Outlier**

Observation in datasets that deviate from rest if data sets

Sometime indicates errors and anomalies in the data

Very important to identify it in the data science project

Mean Median are affected due to Outliers

Variance Standard Deviation movement to left and right

Performance of Machine Learning Models

Types of Outliers

Univariate Outliers Variate 🡪 Variable

Multivariate Outliers

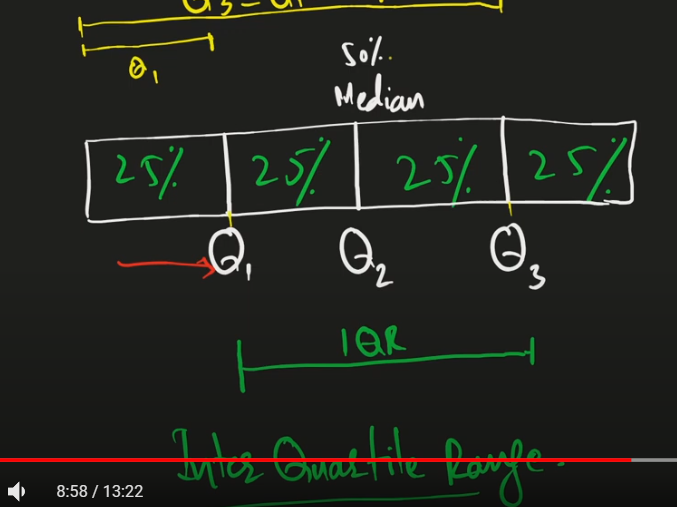
Global Outliers 🡪 a single data point in data that anomalies for complete Data sets

Collective Outliers

Contexture Outliers

Method to Identify the outliers

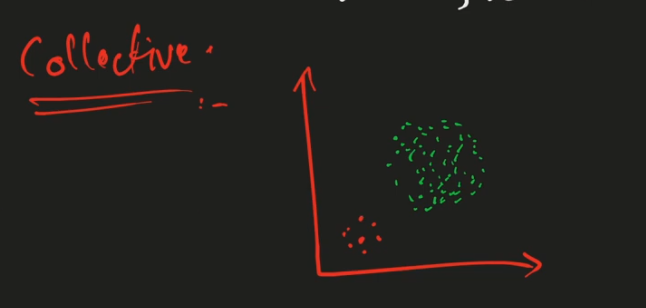
Visually 🡪 Box Plot Histogram

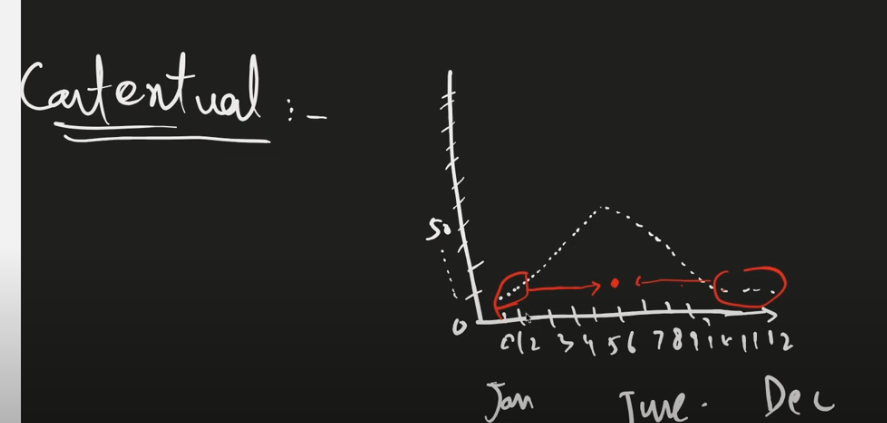


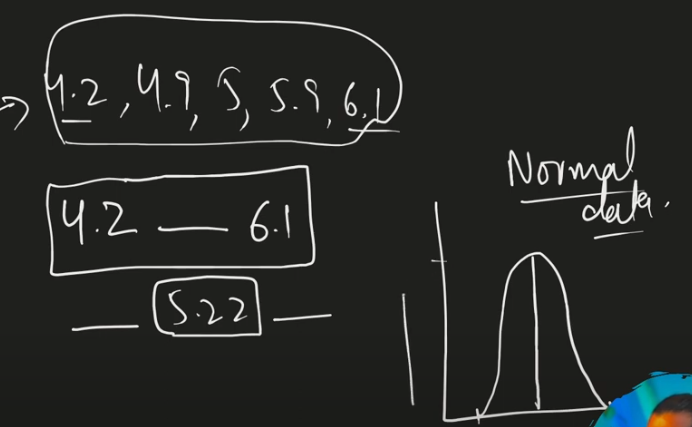
A blackboard with yellow writing

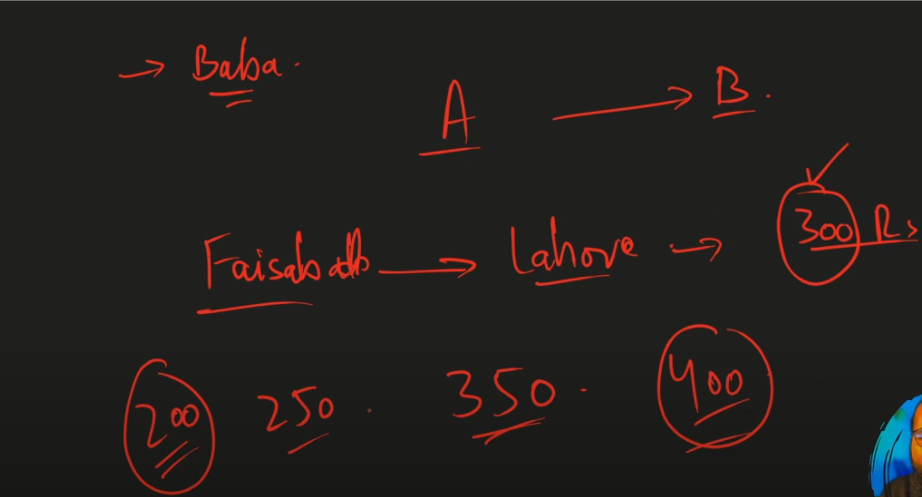
Description automatically generated











A blackboard with white text

Description automatically generated

Deal with Outliers

Remove them

Transform them

Impute 🡪 deal with missing value replace with mean medians

Use Good ML Model

A black and white text

Description automatically generated

https://codanics.com/data-visualization-in-data-science/

Importance

Esay in Presentation

Summarize

Easy for stack holder

Trends

4 Important

Quick insight from data

Informed Decision Making

Engagements with audience investor etc.

Data Exploration

Key principles

Simplicity

Consistency 🡪 same colors and fonts

Accuracy 🡪 Plot your data complete

A blackboard with yellow writing

Description automatically generated

A blackboard with yellow writing

Description automatically generated

A blackboard with a green and yellow circle and numbers

Description automatically generated

A blackboard with white text

Description automatically generated

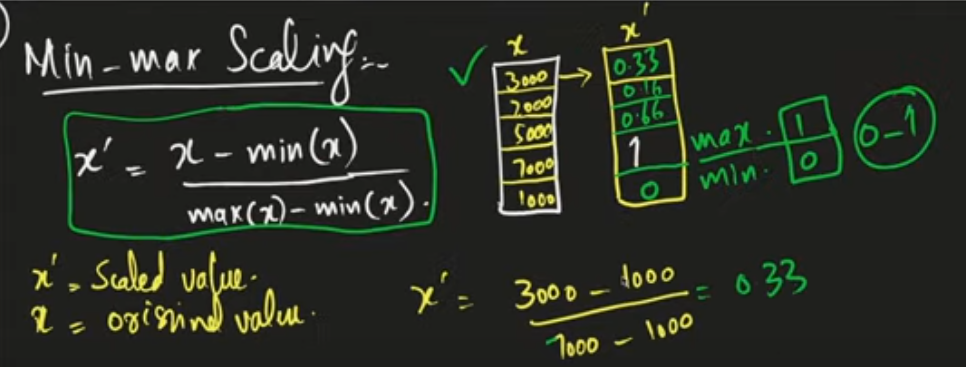
Feature Scaling

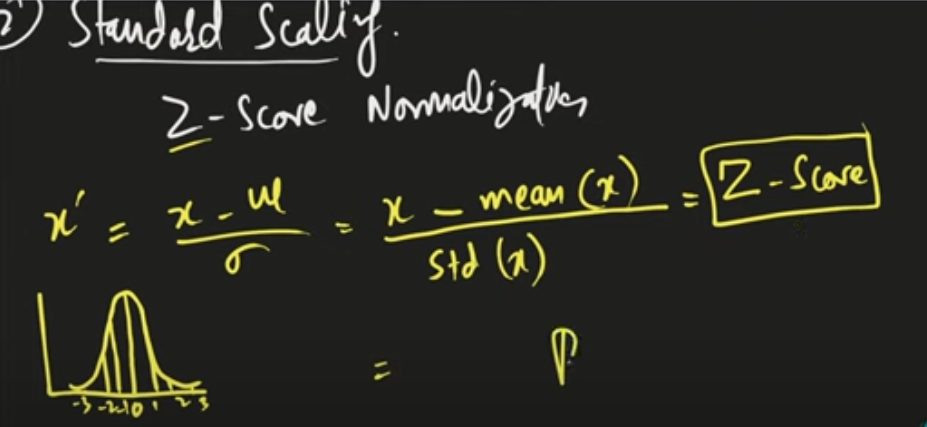
Its is method to limit the range of variable so that they can be compared on same ground

Importance of Feature Scaling

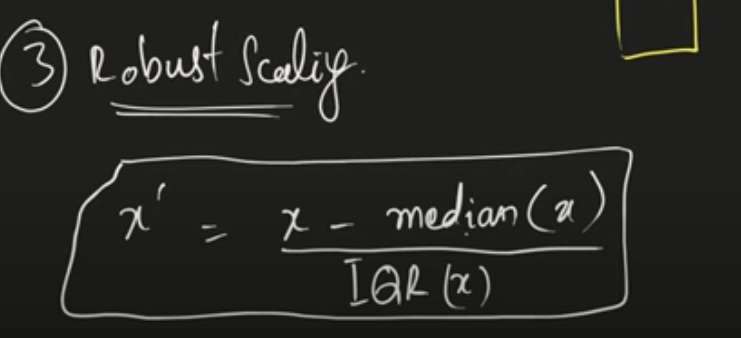
Helps us to perform model faster and better

Especially models like linear regression logistic regression support vector machine



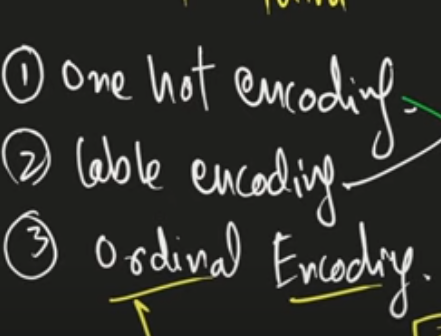


Standard scaling gives us negative values



A black background with white text

Description automatically generated

Feature Encoding   


One Hot Encoding make different column in the form of true and false

Label Encoding convert into numeric in same columns

Ordinary Encodings convert to Meaning full encodes