Data Science Workshop

Mean

- [10,10,10,10,10] mean = 10
- [10,10,10,10,100] mean = 28 (100 is outlier)
- [100,10,10,10,100] mean = 46 (100 is outlier)

- So mean is impacted by outlier (soln: trimmed mean or median)
- [100,10,10,10,100]
- For Trim=1 [100(ignored),10,10,10,100(ignore)] mean=10

Outlier

- An outlier is a data point that is significantly different from other data points in a dataset. Rare data points
- Eg age = [192,5,2,8,9,23] (192 is outlier)

- Outliers has a significant impact while training ml model.
- So it is preferred to remove it before training model.

- Finding outlier: (Box plot or IQR-Inter Quartile range)

Median

- [10,10,10,10,10] Median = 10
- [10,10,10,10,100] Median = 10 (100 is outlier)
- [100,10,10,10,100] Median = 10 (100 is outlier)

- So Median is not impacted by outlier
- It can only be impacted if more than 50% of data points are outliers which contradicts definitions of outlier (rare points)
- Disadvantage : It only take central values so it doesn't truly signifies the data property
- If you want to invest in product don't look at just mean sales instead look both mean and median sales. Because mean can be easily diverted by outlier and median ignores non central values. So look both and take your decision.

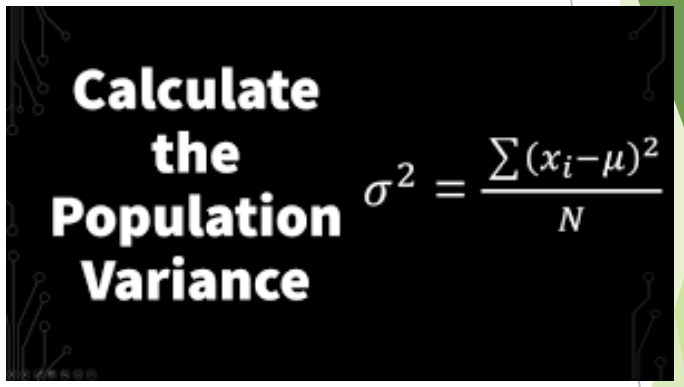
Mode

- [10,20,20,10,30] Mode = 20
- [10,10,10,20,20,30] Mode = 10
- [100,10,20,20,100] Mode = 20 (100 is outlier)

- So Mode is not impacted by outlier.
- It can only be impacted if most frequent value itself is outlier which contradicts definitions of outlier (rare points)
- Disadvantage : It only take central most frequent value so it doesn't truly signifies the data property
- If you want to invest in product you can check mode sales to see what is frequent number of sales.
- https://www.mathsisfun.com/mode.html

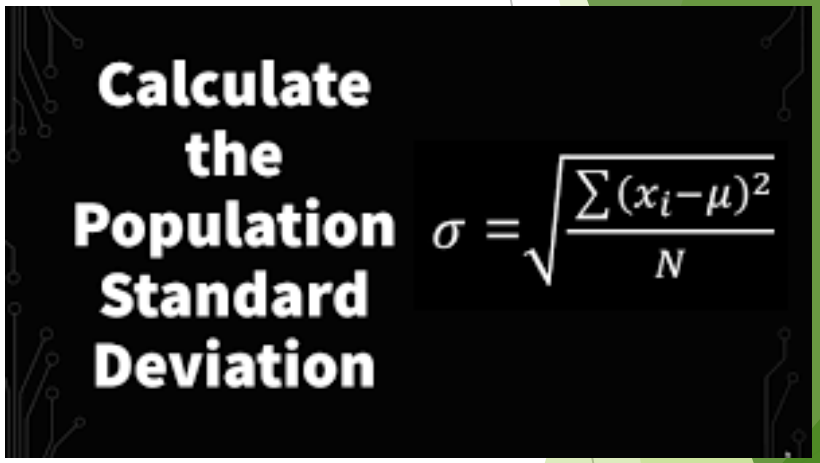
Variance

- Definition: Variance measures how far each data point in a dataset is from the mean.
- Purpose: It quantifies the spread or dispersion of data.
- Calculation: https://www.wikihow.com/Calculate-Variance
- Problem: It exaggerates the deviation from mean (soln: std-dev)



Standard deviation

- Definition: Standard deviation is a statistical measurement that looks at how far individual points in a dataset are dispersed from the mean of that set.
- Purpose: If data points are further from the mean, there is a higher deviation within the data set. It is calculated as the square root of the variance.
- Calculation: https://byjus.com/maths/standard-deviation/
- It is calculated in same scale

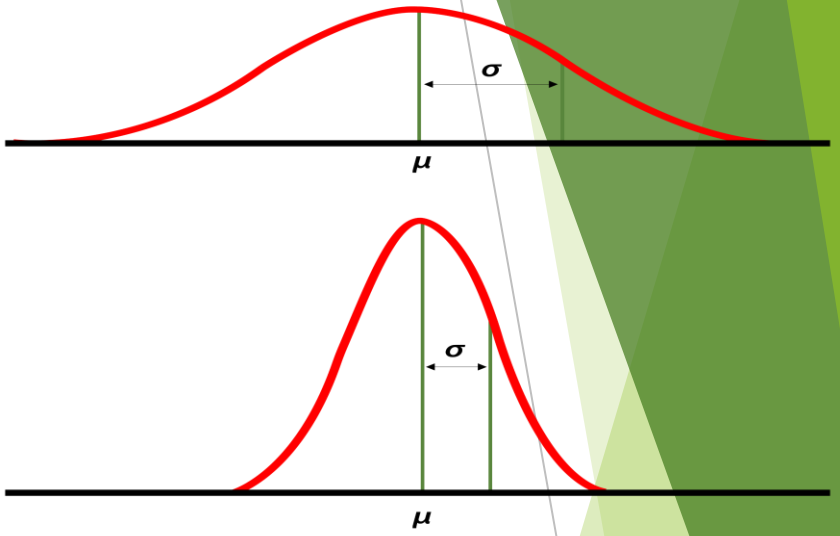


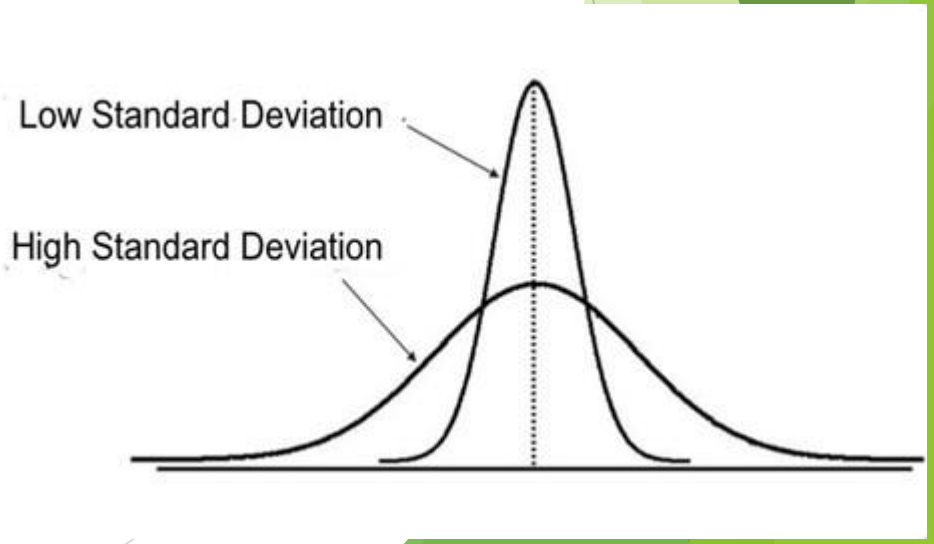
Std-dev significance

- Marks class A: [45,47,48,50,51]
- Mean: 48.2
- Var: 4.56, std-dev = sqrt(var) = 2.14

- Marks class B:[10,20,50,90,100]
- Mean: 54
- Var: 1304, std-dev = sqrt(var) = 36.11

- If you want to invest in any product which would you prefer low std-dev sales or high std-dev sales?





Percentiles

- Definition: A percentile is a measure that indicates the value below which a given percentage of observations in a dataset falls.
- Example: The 25th percentile (or Q1) is the value below which 25% of the data points lie.
- Data: [10,20,30,40,50,60,70,80,90,100]
  - 25th Percentile (Q1): 30
  - 50th Percentile (Median): 50
  - 75th Percentile (Q3): 70

2. Formula for Percentile:

$$P_k = x_i \text{ where } i = \left\lceil \frac{k}{100} \times n \right\rceil$$

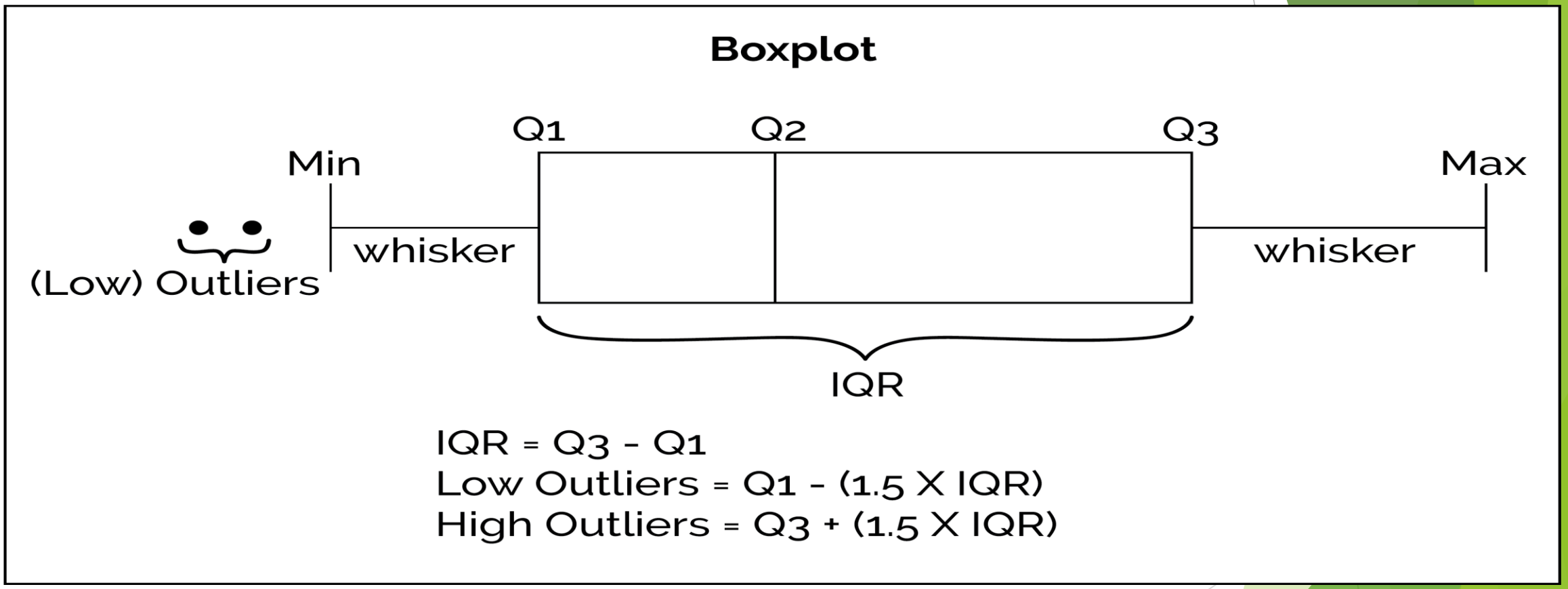
Where:

- P_k: kth percentile
- x_i: Data point at the rank i (sorted dataset)
- n: Total number of data points
- k: Percentile value (e.g., 25, 50, 75)

Percentiles Continued

- **Quartiles:**
- Divide data into **4 equal parts**, each representing 25% of the dataset.
  - **Q1 (25th percentile):** Value below which 25% of the data lies.
  - **Q2 (50th percentile or Median):** Value below which 50% of the data lies.
  - **Q3 (75th percentile):** Value below which 75% of the data lies
- **Deciles:**
- Divide data into **10 equal parts**, each representing 10% of the dataset.
  - **D1 (10th percentile):** Value below which 10% of the data lies.
  - **D5 (50th percentile):** Value below which 50% of the data lies.
  - **D9 (90th percentile):** Value below which 90% of the data lies.

Percentiles Significance



- https://discovery.cs.illinois.edu/learn/Exploratory-Data-Analysis/Quartiles-and-Box-Plots/

Correlation

- Correlation measures the strength and direction of a linear relationship between two variables.
- https://www.mathsisfun.com/data/correlation.html
- **Range:** −1 to +1
  - +1: Perfect positive correlation (as one variable increases, the other increases).
  - 0: No correlation.
  - −1: Perfect negative correlation (as one variable increases, the other decreases).

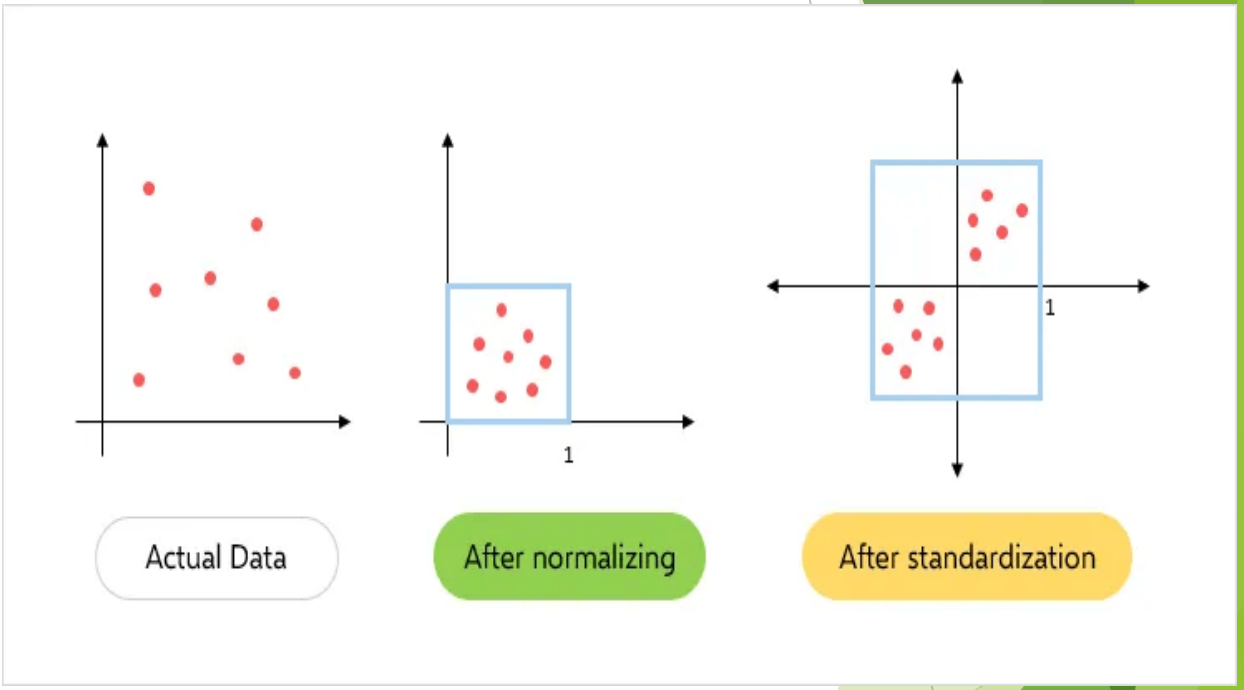
  Read world example:

  1. Positive correlation: temp vs ice cream sales, rent vs time, pandemic vs marks
  2. Negative correlation: Any electronics age vs price, value of money vs time
  3. No correlation: Roll no vs exam marks, exam seat vs exam marks

Data Normalization

- Var1: Age = [22,32,56,11,27,60]
- Var2: salary = [22k,32k,56k,11k,27k,60k]
- Dist (22,22k) and (32,32k) = 10,000 units

- After scaling
- Var1:[0.2245,0.4286,0.9184,0.0000,0.3265,1.0000]
- Var2:[0.2245,0.4286,0.9184,0.0000,0.3265,1.0000]
- Dist (22,22k) and (32,32k) = 0.2888 units



Correlation

- A correlation is a statistical measure of the relationship between two variables/columns/features.
- Using a scatterplot, we can generally assess the relationship between the variables and determine whether they are correlated or not.
- The correlation coefficient is a value that indicates the strength of the relationship between variables. The coefficient can take any values from -1 to 1. The interpretations of the values are:
  - **-1:** Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
  - **0:** No correlation. The variables do not have a relationship with each other.
  - **1:** Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).
- Examples:
- Positive: Experience vs Salary
- Negative: Electronics (TV, Mobile, car, etc) age vs resale value
- No/zero: weight vs marks in exam

Correlation (continue):-

$$r = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2}}$$

- https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html