Week 01_Module 01

Tuesday, November 4, 2025 7:57 PM

Week 01: Statistics and Mathematics for ML

Module 01: Descriptive Statistics and Distributions

- Mean, median, mode: when each is appropriate
- Variance and standard deviation as measures of spread
- Percentiles, quartiles, IQR, and z-score
- Distribution shapes: symmetric vs skewed, long tails; outlier detection with IQR fences
- When median and IQR are preferred over mean and standard deviation

Module 02: Probability Basics for ML
Module 2.5: Worked examples on center, spread, IQR fences, and z-scores
Module 03: Data Quality, Scaling, and Encoding
Module 3.5: Bayes and confusion-rate worksheets
Module 4: Quiz

What is Statistics & Statistics for ML

What is Statistics?

Statistics is the science of working with data — **collecting, summarizing, analyzing, and interpreting** it to understand the real world through numbers.

Example:

Suppose we collect the marks of all students in a class.
The **mean** tells us the class's average performance,
the **median** shows the middle score,
and the **standard deviation** tells us how spread out the marks are.
That's statistics — finding meaning in data.

What is Statistics for Machine Learning?

Statistics for ML helps us **understand the data before we train models**.
A model learns patterns from data — so if we don't understand our data's behavior, we risk building biased or misleading models.
We use statistics in ML to study **the center, spread, shape, and relationships** within data.
It forms the foundation for preprocessing, feature scaling, and model evaluation.

Types of Statistics

There are two main types of statistics:

1. **Descriptive Statistics** — describe or summarize data.
   Example: the average mark of a class is 78.
2. **Inferential Statistics** — draw conclusions about a population from a sample.
   Example: using one class's marks to estimate the average for the whole university.
   In machine learning, we mostly rely on **descriptive statistics** because we focus on **understanding and preparing our dataset**, not on generalizing to an unseen population."

Module 01: Descriptive Statistics and Distributions

Module 01 Overview

Descriptive statistics helps us **summarize and visualize** data using numbers and graphs.
We'll learn how to describe data's center, spread, and shape.

Topics Covered:

- Mean, Median, Mode — and when each is appropriate
- Variance and Standard Deviation — measuring spread
- Percentiles, Quartiles, IQR, and Z-Score
- Distribution Shapes — symmetric vs skewed, outliers via IQR fences
- When Median and IQR are preferred over Mean and SD

Example:

In salary data, if one CEO earns far more than others, the mean shoots up.
The **median**, however, better represents a "typical" salary.
That's why median and IQR are often safer for skewed data.

Module 02: Probability Basics for ML

Module 02 – Probability Basics

Next, we'll study **probability**, the mathematics of uncertainty.
We'll discuss events, sample space, conditional probability, and independence.
We'll intuitively understand **Bayes' Theorem**, the core idea behind Naïve Bayes classification.
Then we'll explore **sensitivity, specificity, false positive and negative rates**, and how **class imbalance** can distort evaluation metrics — critical knowledge for healthcare, fraud, or sentiment-analysis models.

Module 2.5 and 3 Overview

From Theory to Practice

In Module 2.5, we'll do worked examples — computing center, spread, IQR fences, and Z-Scores by hand and in Python.
Then Module 03 takes us into **Data Quality, Scaling, and Encoding** — the practical bridge between statistics and machine learning.
We'll cover:

- Types of missing data (MCAR, MAR, MNAR) and simple imputation
- Standardization, Min-Max, and Robust Scaling — formulas and use cases
- Nominal vs Ordinal Encoding and their geometric impact
- Distance metrics (Euclidean, Manhattan, Cosine)
- Covariance, Correlation, and a conceptual intro to PCA

Module 3.5 and 4 Overview

Finally, in **Module 3.5**, we'll apply what we've learned through **Bayes' and confusion-rate worksheets** — completely hands-on, no coding required.

Then, **Module 4** wraps up this unit with the **Concept Quiz**

These final modules help connect everything, understanding data, applying probability, and evaluating model performance

before we move into coding in later weeks.

Week's Goal

This Week's Goal

Week 1 is about building intuition.
You'll learn how to summarize data, detect outliers, understand distributions, and connect these concepts to machine learning tasks.

Remember: We're not training to be statisticians, but to use statistics as a tool to make machine learning models accurate, robust, and explainable.

Week 01_Module 01_Part 02

Friday, November 07, 2025 3:38 PM

Mean (Average)

Mean is what we usually call the *average*.
You add up all values and divide by the number of observations

Formula:

$$\text{Mean} = \frac{\sum x_i}{n}$$

Example:
For `[10, 12, 14, 16, 18, 100]`
Mean = (10 + 12 + 14 + 16 + 18 + 100) / 6 = 28.3

Median

Median is the middle value when data is sorted.
It splits the dataset into two equal halves.

Example:
Sorted `[10, 12, 14, 16, 18, 100]` → middle two are 14, 16
Median = (14 + 16)/2 = 15

Mode

Mode is the most frequent value in the dataset.

Example:
For `[2, 2, 3, 3, 3, 4, 5]`, mode = 3

Comparison & When to Use Which

Situation	Best Measure	Reason
Symmetric data (e.g., height,	**Mean**	All values contribute

temperature)		equally
Skewed data with outliers (e.g., income, price)	**Median**	Resistant to outliers
Categorical data (e.g., color, gender)	**Mode**	Works with non-numeric data

Common Mistakes

- Using mean with skewed data → misleading.
- Using median for categorical → meaningless.
- Using mode for continuous numeric → rarely helpful.
- Forgetting to check distribution before imputing missing values.

Summary & Wrap-Up

- **Mean** – good for normal, balanced data.
- **Median** – good for skewed or outlier-rich data.
- **Mode** – best for categorical data.

In Machine Learning, choosing the right measure helps you **represent data correctly**, **fill missing values wisely**, and **build more accurate models**.

Week 01_Module 01_Part 03

Friday, November 7, 2025 9:02 PM

Variance:

Variance measures the *average squared deviation* of data points from the mean.
The formula for **population variance** is:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

For a **sample** (a subset of data), we divide by n−1 instead of n:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Difference between Population Variance and Sample Variance

- **Population Variance (σ^2)** measures how spread out the *entire population's* data is from the mean.
- **Sample Variance (s^2)** estimates population variance using a *subset (sample)* of the data.

◇ **Example**

Suppose we have data on exam scores:
[70, 80, 90]

Step 1: Mean
$\bar{x} = (70+80+90)/3 = 80$

Step 2: Differences from mean
70 → $(70 - 80)^2 = 100$
80 → $(80 - 80)^2 = 0$
90 → $(90 - 80)^2 = 100$

Sum of squared differences = 200

Step 3: Variances

- **Population Variance (σ^2)** = 200 / 3 = **66.67**
- **Sample Variance (s^2)** = 200 / (3 − 1) = **100**

→ The **sample variance is slightly higher** because we divide by a smaller number to correct for bias.

Standard Deviation:

Standard deviation is the square root of variance, giving us a spread measure in the *same unit* as the data:

Example: If the average house price is $200 K with SD = $50 K, most houses lie within $150 K–$250 K.

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}$$

Variance & SD in ML Pipelines

Variance and standard deviation appear throughout ML:

• **Feature Scaling:** Standardization uses to make features comparable by ensuring that each feature has mean = 0 and standard deviation = 1.

$$z_i = \frac{x_i - \mu}{\sigma}$$

• **Regularization:** Ridge and Lasso reduce coefficient variance to prevent overfitting.

• **Model Diagnostics:** High variance in model performance across folds suggests instability.

Bias-Variance Trade-off

In model evaluation, **variance** also refers to how much a model's predictions change with different training data.

A high-variance model memorizes the training set (overfitting), while low variance but high bias underfits.

So, variance of data and variance of model parameters are different concepts but share the same intuition — *instability due to spread.*

Comparison Table

Concept	Variance	Standard Deviation
Definition	Average of squared deviations	Square root of variance
Formula	$\Sigma(x-\mu)^2 / n$ or $\Sigma(x-\bar{x})^2 / (n-1)$	$\sqrt{\text{Variance}}$
Unit	Squared unit (e.g., $\$^2$)	Same as data (e.g., $)
Interpretation	Mathematical measure	Intuitive spread
Use	Theoretical/statistical analysis	Communication & scaling

Both quantify variability, but standard deviation is easier to interpret because it's in the same scale as the data.

Example (same data [70, 80, 90])

We already found variance = 66.67 (population).

Then

Standard Deviation (σ)$=\sqrt{66.67}=8.16$

☑ Interpretation:

- Variance = 66.67 (squared units, less intuitive).
- Standard deviation = 8.16 → means scores typically vary ±8 points from the mean.

Summary & Takeaways

☑ Variance measures *how far* data points deviate from the mean on average.

☑ Standard deviation is the square root of variance, in original units.

☑ Both are critical for understanding data distribution, detecting outliers, and standardizing features.

☑ In ML, low variance features may add little value, while extremely high variance may require scaling or transformation.

Week 01_Module 01_Part 04

Saturday, November 8, 2025 12:02 PM

Understanding Percentiles, Quartiles, IQR, and Z-Score in Machine Learning

Outline

1. **Introduction**
2. **Percentiles and Quartiles**
3. **Interquartile Range (IQR)**
4. **Z-Score**
5. **Python Implementation & Outlier Detection**

2. Percentiles and Quartiles

Definition:

"Percentile tells us the value below which a given percentage of observations fall."

Example:

"If your test score is in the 90th percentile, it means you scored better than 90% of the people."

Formula (Conceptually):

Percentile rank = (number of values below x / total number of values) × 100

Quartiles:

"Quartiles are special percentiles that divide data into four parts."

- **Q1 (25th percentile):** 25% of data lies below this value.
- **Q2 (50th percentile):** This is the **median**.
- **Q3 (75th percentile):** 75% of data lies below this value.

ML Connection:

"We often use percentiles to handle outliers in data.
For example, when preprocessing numerical features, we can remove values below the 1st percentile or above the 99th percentile to reduce the effect of extreme data points."

3. Interquartile Range (IQR)

Definition:

"IQR measures the spread of the middle 50% of data — it's calculated as Q3 minus Q1."

Formula:

IQR = Q3 − Q1

Interpretation:

"A smaller IQR means data points are tightly packed.
A larger IQR means data has high variability."

Outlier Detection Rule:

Any point below **Q1 — 1.5 × IQR** or above **Q3 + 1.5 × IQR** is an **outlier**.

ML Connection:

"IQR-based filtering is common before training models to prevent extreme values from skewing results — for example, in **housing price prediction**, or **salary datasets**, where a few very large values can distort the model's understanding of normal behavior."

4. Z-Score

Definition:

"Z-score tells us how many standard deviations a value is away from the mean."

Formula:

$$z = \frac{x-\mu}{\sigma}$$

Where:

- x = individual value
- μ = mean of the dataset
- σ = standard deviation

Interpretation:

- z = 0 → value is exactly at the mean
- z = 1 → 1 standard deviation above mean
- z = -2 → 2 standard deviations below mean

ML Connection:

"Z-score normalization is part of **standard scaling**, which transforms all features to have mean 0 and standard deviation 1.
This is critical for models like **KNN**, **SVM**, or **Gradient Descent-based models**, which are sensitive to feature scale."

Summary

Concept	Formula	Use in ML
Percentile	position measure	Feature understanding
Quartile	Q1, Q2, Q3 (25th, 50th, 75th)	Spread summary
IQR	Q3 – Q1	Outlier detection
Z-score	$(x-\mu)/\sigma$	Standardization & scaling

Conclusion

To summarize:

- **Percentiles and quartiles** help describe data spread.
- **IQR** shows the range of central data and helps detect outliers.
- **Z-score** standardizes data and helps identify extreme values.

Together, these tools are fundamental for **data preprocessing**, **feature scaling**, and **robust ML model performance**.

Week 01_Module 01_Part 05

Saturday, November 8, 2025 2:01 PM

Distribution Shapes for ML: Symmetric vs Skewed, Long Tails, and IQR Outliers

Goal: Understand how distribution shape affects ML models and practice outlier detection using IQR fences.

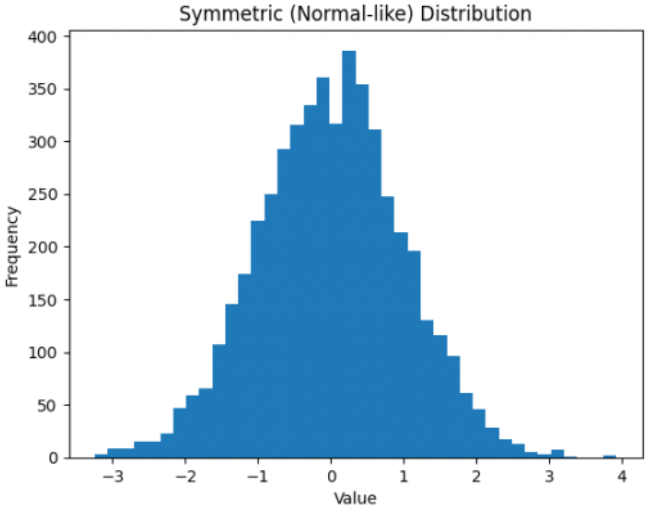
What you'll do:

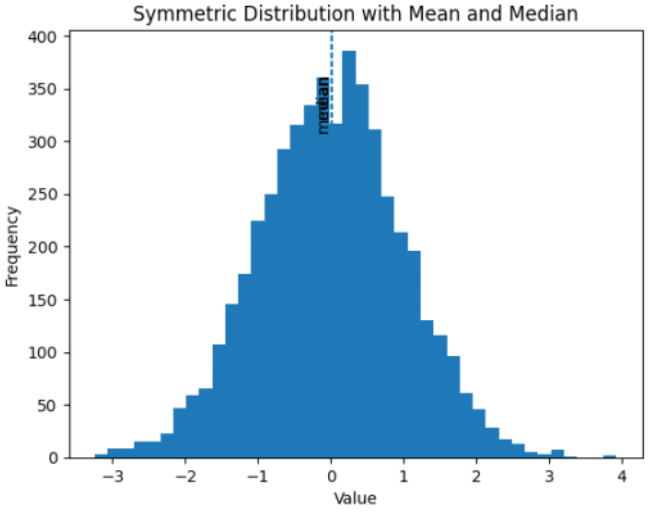
- Visualize symmetric vs. skewed distributions
- See what long tails look like and why they matter
- Detect outliers using IQR fences in code

Symmetric Distributions

"A symmetric distribution looks the same on both sides of its center. The most common example is the **Normal Distribution**. Here, the mean, median, and mode all lie at the center."

Visuals:





ML Focus:

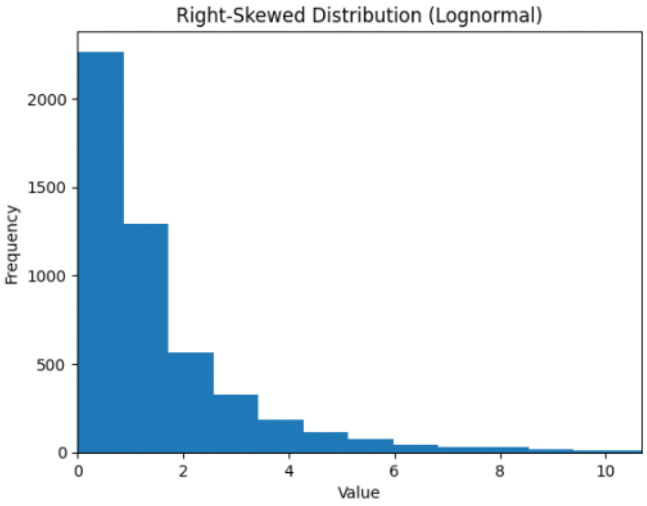
Many ML algorithms, such as **Linear Regression**, **SVMs**, and **KNN**, assume features follow a roughly normal distribution for optimal

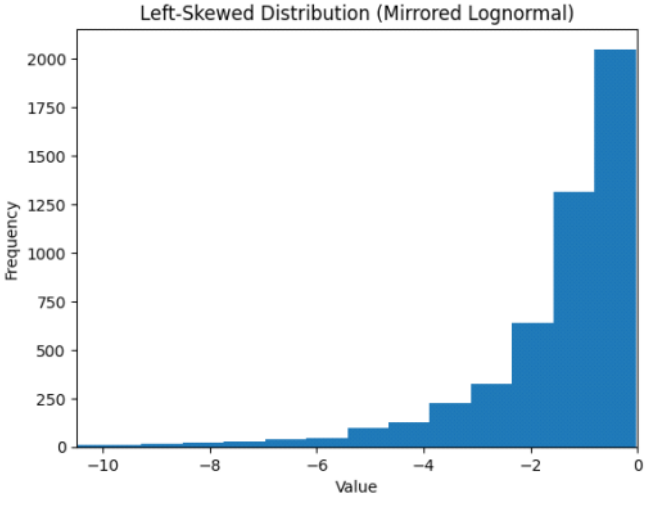
performance. If data is symmetric, we often don't need transformations.

Skewed Distributions

A skewed distribution is not symmetric. It can be **positively skewed (right-skewed)** or **negatively skewed (left-skewed)**.

Visuals:





ML Focus:

Right-skewed data often appears in income, price, or reaction time datasets. For ML models, skewness can reduce model accuracy and cause bias. To fix this, we often apply **log transformation**, **Box–Cox**, or **Yeo–Johnson transformations** to make data more symmetric.

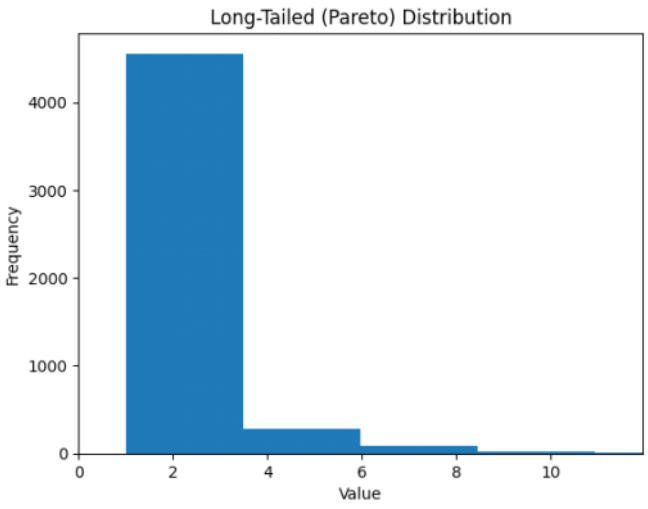
Example:

For instance, in predicting house prices, applying a log transform to the price column can improve model stability and perfomance.

Long Tails and Their Significance

A long tail means extreme values stretch far from the center. Long-tailed distributions are common in real-world datasets — like user engagement or social media followers — where a few points dominate the range.

Visuals:



ML Focus:

Long-tailed data challenges ML models. Models may overfit to frequent cases and ignore rare but important events — like fraud detection or rare disease prediction. Handling long tails often involves **data normalization**, **resampling**, or **outlier-aware models**.

In short:

Concept	Focus	Meaning	Example
Skewness	Asymmetry	Which side the data lean toward	Income distribution (right-skewed)
Long Tail	Tail length and thickness	How extended rare or extreme values are	YouTube video views (long right tail)

Outlier Detection using IQR Fences

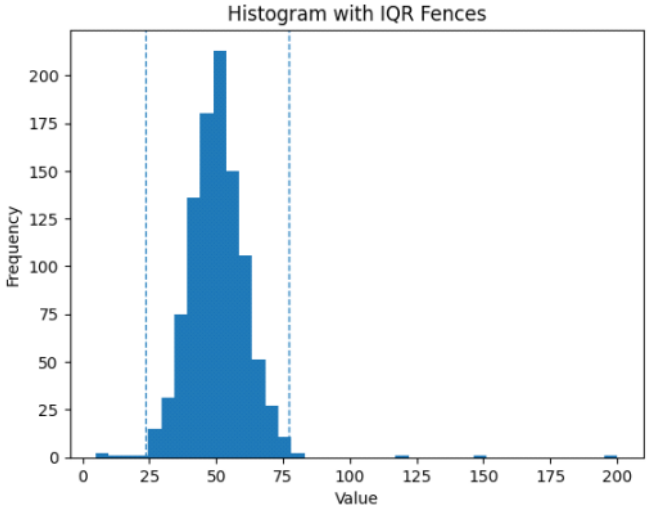
Outliers are data points that differ significantly from most observations. They can distort model training and degrade performance. A simple yet effective method to detect outliers is using the **Interquartile Range (IQR)**.

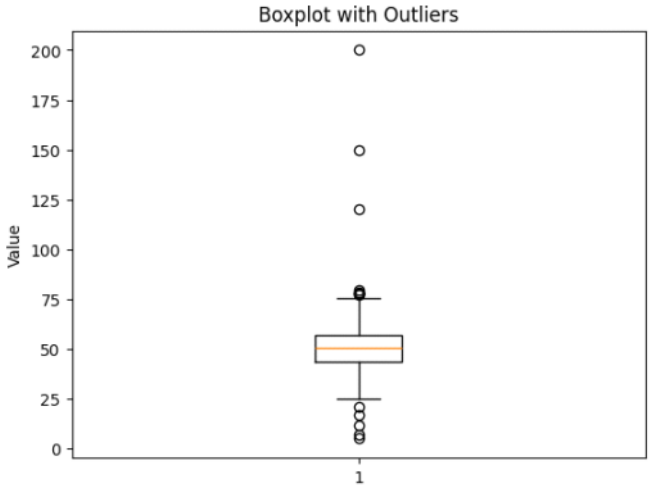
Step-by-step explanation:

1. Q1 = 25th percentile
2. Q3 = 75th percentile
3. IQR = Q3 – Q1
4. Lower Fence = Q1 – 1.5 × IQR
5. Upper Fence = Q3 + 1.5 × IQR

   Any data point outside these fences is considered an outlier.

Visuals:





- Q1 = 30, Q3 = 70 → IQR = 40
- Lower = 30 – 1.5×40 = –30, Upper = 70 + 1.5×40 = 130 → values < –30 or > 130 = outliers.

ML Focus:

In ML preprocessing, outlier removal or capping helps improve model robustness, especially for models sensitive to scale — like Linear Regression or KNN.

When to Keep vs. Remove Outliers

Not all outliers are bad. In some ML problems, like fraud detection or medical diagnosis, outliers represent critical cases. Instead of removing them, we may label or model them separately.

Recap & Summary

Let's summarize what we learned:

☑ Symmetric distributions make ML models more stable.
☑ Skewed or long-tailed data often need transformation.
☑ IQR is a quick and effective outlier detection method.
☑ Always analyze whether outliers represent noise or valuable insight.

Understanding data shape is the first step toward building trustworthy ML models.

Week 01_Module 01_Part 06

Saturday, November 8, 2025 4:33 PM

Quick Recap of Basics

Let's quickly recall what these measures do.

- **Mean** → The arithmetic average; it represents the *center* of the data.
- **Median** → The middle value when data are sorted.
- **Standard Deviation (SD)** → Measures how far data points typically spread from the mean.
- **Interquartile Range (IQR)** → The difference between the 75th and 25th percentiles — in other words, it captures the *middle 50%* of your data.

Both pairs measure **center** and **spread**, but they behave very differently when your data has **outliers** or **skewness**.

Why Mean and SD Fail Sometimes

Mean and standard deviation work great for **symmetric, normally distributed data** — like heights or IQ scores in large populations.
But if you have **outliers** — for example, one billionaire in an income dataset — the mean gets pulled toward the extreme.
Example:

```
import numpy as np
data = [25, 27, 28, 29, 30, 31, 35, 500]  # one big outlier
print("Mean:", np.mean(data))
print("Median:", np.median(data))
```

The mean shoots up, but the median stays stable — that's why we call median a **robust** measure.

When to Use Median and IQR

Use **Median and IQR** when:

1. **Data are skewed or have long tails**
   - Example: *Income, house prices, YouTube views, medical costs.*
   - These are **right-skewed**, and the mean gets dragged upward by a few large values.
   - The median gives a better picture of a "typical" case.
2. **There are outliers or extreme values**
   - Outliers inflate SD and make the spread look huge even if most data are normal.
   - IQR ignores extremes by focusing on the middle 50%.
3. **You need robust statistics for ML preprocessing**
   - Algorithms like **RobustScaler** in sklearn use **median and IQR** instead of mean and SD to scale data safely.
   - It's perfect for datasets where outliers shouldn't dominate scaling.

When Mean and SD Are Still Better

Use **Mean and SD** when:

- Data are roughly **normal (bell-shaped)**.

- You plan to apply algorithms assuming normality — for instance, **Linear Regression**, **Naive Bayes**, or **Z-score scaling**.

Wrap-up and Summary

So to summarize:

- **Mean & SD** → Best for *normal, clean data*.
- **Median & IQR** → Best for *skewed, messy, or outlier-heavy data*.

Always visualize first — a simple histogram or boxplot can tell you whether your data needs a "robust" approach or not.

Week 01_Module 01_Part 07

Saturday, November 8, 2025 5:21 PM

1. Mean, Median, Mode – Finding the Center

We started with **measures of central tendency** — the *mean*, *median*, and *mode*.

- **Mean** is the average — great when the data is clean and balanced.
- **Median** is the middle value — perfect when your data has outliers.
- **Mode** is the most frequent value — often useful for categorical features.

In ML, using the median instead of mean can make your model more robust to extreme values.

2. Variance and Standard Deviation – Measuring the Spread

Then we explored **how spread out the data is**.

- **Variance** tells us how far data points are from the mean, on average.
- **Standard deviation (σ)** is just the square root of variance, making it easier to interpret in the same unit as the data.

In ML, features with large variance can dominate the model, which is why we normalize or standardize data before training.

3. Percentiles, Quartiles, IQR, and Z-score

Next, we moved into **relative positioning and outlier detection**.

- **Percentiles** and **quartiles** divide data into ranks or quarters.
- **Interquartile Range (IQR)** = Q3 − Q1 — helps identify the "middle 50%" of data.
- **Z-score** measures how far a point is from the mean in terms of standard deviations.

ML use case: Z-scores are often used for anomaly detection or data cleaning steps.

4. Distribution Shapes and Outliers

We learned how the **shape of the distribution** tells a story:

- **Symmetric** distributions like the normal curve are balanced.
- **Right-skewed** distributions (like income) have long tails to the right.
- **Left-skewed** distributions tail off to the left.

Outliers stretch the tails — and using the **IQR method**, we can set boundaries to detect them.

In ML, outliers can drastically affect models like Linear Regression, so detecting and handling them is critical.

5. When to Use Median and IQR

Finally, we saw **when median and IQR beat mean and standard deviation** — that's when data is **not symmetric**, or has **outliers**.

For example, in predicting **income**, a few millionaires can ruin the mean, but the median stays honest.

So in ML preprocessing, median and IQR are often used in robust scaling.

🚀 Conclusion

So that wraps up Module 1!

You now understand how to describe your data — its **center, spread, and shape** — and how to detect unusual patterns before any modeling begins.

In the next module, we'll move from describing data to **understanding probability and Bayes' theorem**, which will build the foundation for how ML actually "learns."