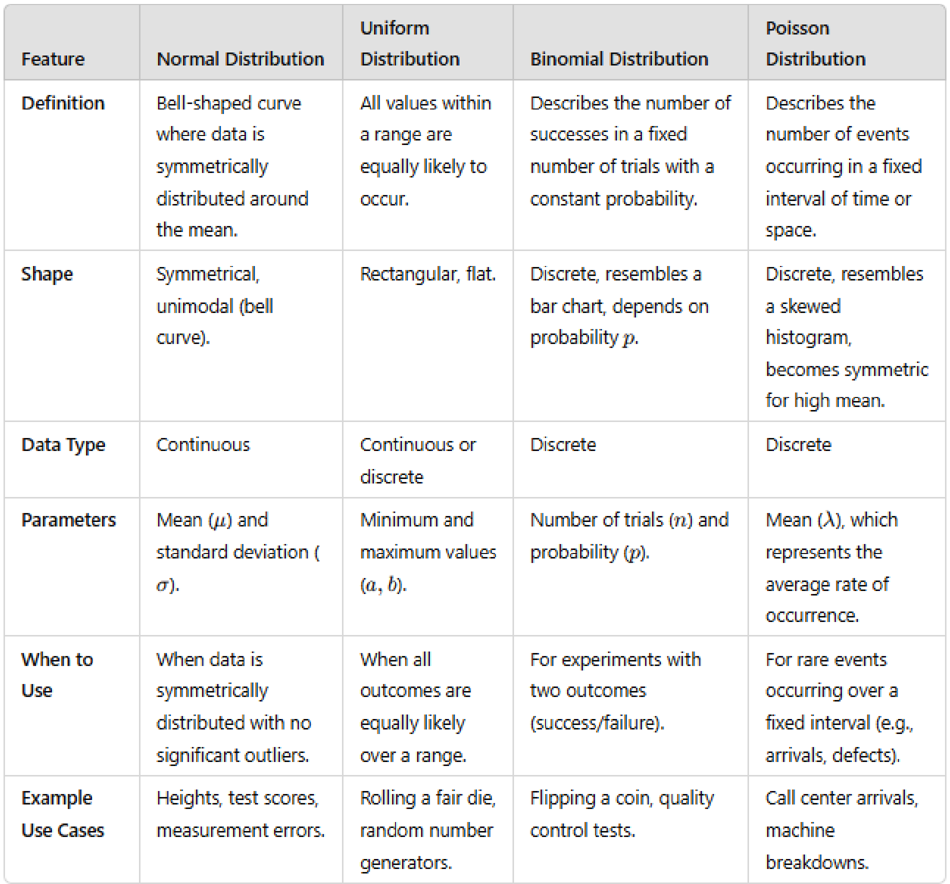
AWS ML - Data Analysis

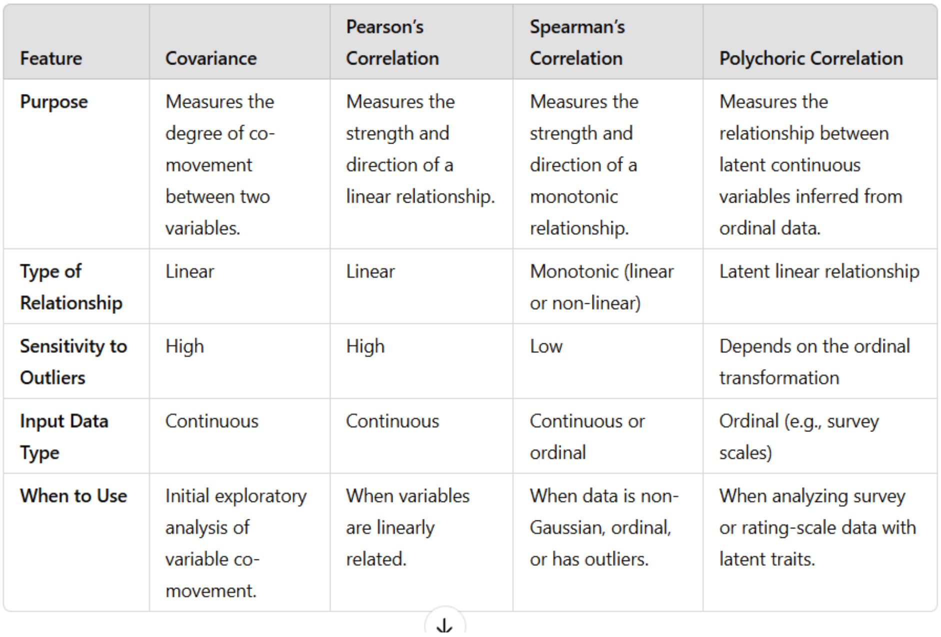
1. Data Distributions



Feature	Normal Distribution	Uniform Distribution	Binomial Distribution	Poisson Distribution
Definition	Bell-shaped curve where data is symmetrically distributed around the mean.	All values within a range are equally likely to occur.	Describes the number of successes in a fixed number of trials with a constant probability.	Describes the number of events occurring in a fixed interval of time or space.
Shape	Symmetrical, unimodal (bell curve).	Rectangular, flat.	Discrete, resembles a bar chart, depends on probability p.	Discrete, resembles a skewed histogram, becomes symmetric for high mean.
Data Type	Continuous	Continuous or discrete	Discrete	Discrete
Parameters	Mean (μ) and standard deviation (σ).	Minimum and maximum values (a, b).	Number of trials (n) and probability (p).	Mean (λ), which represents the average rate of occurrence.
When to Use	When data is symmetrically distributed with no significant outliers.	When all outcomes are equally likely over a range.	For experiments with two outcomes (success/failure).	For rare events occurring over a fixed interval (e.g., arrivals, defects).
Example Use Cases	Heights, test scores, measurement errors.	Rolling a fair die, random number generators.	Flipping a coin, quality control tests.	Call center arrivals, machine breakdowns.

- **Bernouilli Distribution** : is a **special case** of the Binomial distribution where the number of trials (nnn) is equal to 1. Bernoulli models one trial, while Binomial models multiple independent Bernoulli trials.
- **Gamma Distribution** : A two-parameter family of continuous probability distributions often used to model waiting times.
- **Beta Distribution** : A distribution defined on the interval [0, 1], often used in Bayesian statistics and to model distributions of probabilities.
- **Log-Normal Distribution** : Represents a variable whose logarithm is normally distributed, often used in reliability analysis and stock prices.
- **Chi-Squared Distribution** : Used primarily in hypothesis testing and constructing confidence intervals, particularly with variance estimates.
- **t-Distribution** : Used in small sample size settings for estimating population parameters when the standard deviation is unknown.

2.Correlation Coefficients



Feature	Covariance	Pearson's Correlation	Spearman's Correlation	Polychoric Correlation
Purpose	Measures the degree of co-movement between two variables.	Measures the strength and direction of a linear relationship.	Measures the strength and direction of a monotonic relationship.	Measures the relationship between latent continuous variables inferred from ordinal data.
Type of Relationship	Linear	Linear	Monotonic (linear or non-linear)	Latent linear relationship
Sensitivity to Outliers	High	High	Low	Depends on the ordinal transformation
Input Data Type	Continuous	Continuous	Continuous or ordinal	Ordinal (e.g., survey scales)
When to Use	Initial exploratory analysis of variable co-movement.	When variables are linearly related.	When data is non-Gaussian, ordinal, or has outliers.	When analyzing survey or rating-scale data with latent traits.

3. feature Engineering

Dealing with missing values

- Do nothing
- **Most frequent value, Mean and Median replacement :** Median may be a better choice than mean when outliers are present. It's not the best solution : Only works on column level, misses correlations between features and Can't use on categorical features
- **Dropping**
- **MICE** Multiple Imputation by Chained Equations finds relationships between features and is one of the most advanced impututation methods available.
- **Using machine learning techniques** such as KNN, Regression
- **deep learning model** to impute missing values for categorical data

#todo : **Interpolation/Extrapolation, Forward Filling/Backword Filling, Hot deck Imputation**

Dealing with unbalanced data

Handling Outliers

Data points that lie more than one standard deviation (square root of variance) from the mean can be considered as unusual

AWS's Random Cut Forest algorithm for outlier detection. ,Found within QuickSight, Kinesis Analytics, SageMaker, and more

Numerical feature engineering

- **Normalization** : rescales values into [0,1]
- **Standardization** : rescales data to have a mean of 0 and a standard deviation of 1
- **Binning**
  Transform numerical data to categorical data/ ordinal data
  when there is uncertainty in the measurments
  Quantile binning : ensure having the same number of samples in each bin

Transforming , Feature Extraction

create or replace data by doing a trasnformation on it

Encoding Categorical features

- **One-hot encoding** for categorical data : create “buckets” for each category. 1 for a category and 0 for others
- **Binarizer Encoding** : features with binary nature
- **Label Encoding**
  - Ordinal Encoding

Feature Selection

Use feature selection to **filter irrelevant or redundant features**

- requires Normalization
- Removes features :**variance thresholds** (#todo)

Shuffling

Why? some models may learn from residual signals in the training data resulting from the order.

Adding regularizations

- **Lasso**: Prevents overfitting by eliminating irrelevant features (automatic feature selection).
- **Ridge**: Reduces overfitting by shrinking all coefficients, especially useful when features are correlated.

3. Amazon QuickSight

Business analytics and visualisations

Applications:

- Interactive viz
- Dashboards and KPI

Examples of data sources :

- Redshift
- Aurora/RDS
- EC2-hosted databases
- S3
- Snowflake

Data sets are imported into SPICE : Super-fast, Parallel, In-memory, Calculation Engine

Features of Amazon Quicksights ML insights :

- Anomaly detection
- forecasting
- auto-narratives

Quicksight Q

A feature in QuickSight that answers business questions **with natural language** instead of creating sql queries.

You must set up topics associated with datasets

QuickSight Visual Types :

- **AutoGraph** allows you to select the most appropriate visualisations
- **Bar Charts/Histograms** for comparison and distribution on a single dimension
- **Line graphs** / **Area line charts :** for changes over time.
- **Scatter plots/ heatmaps** for correlation, multiple distribution
- **Pie graphs** for aggregation, categorical attributes
- **Donut charts** : Pourcentage of Total amout
- **Gauge Charts** : Compare values in a measure
- **Tree maps** for hierarchical aggregation,
- **Pivot tables** for tabular data
- **KPI's :** compare key value to its target value
- **Geospatial charts**
- **World Clouds** *:* word or phrase frequency
- **Pair plots** are best used for spotting correlations between pairs of attributes.
- **Box & whisker,** or just "box plots", organize your data into quartiles, and display outliers in the outer quartiles, distribution on a multiple dimensions

4. Elastic MapReduce (EMR)

- Managed Hadoop framework on EC2 instances
- Includes Spark, HBase, Presto, Flink, Hive and more
- If you have a massive dataset that you need to preprocess, EMR provides a way to distributing the load of processing that data across an entire cluster of computers
- A cluster in EMR is a collection of nodes (EC2 instances)
    - **Master node:** manages the cluster
      • Single EC2 instance
    - **Core node:** Hosts HDFS data and runs tasks
      • Can be scaled up & down, but with
      some risk
    - **Task node:** Runs tasks, does not host data
      • No risk of data loss when removing
      • Good use of **spot instances**

EMR / AWS Integration

- Amazon EC2 for the instances that comprise the nodes in the cluster
- Amazon VPC to configure the virtual network in which you launch your instances
- Amazon S3 to store input and output data
- Amazon CloudWatch to monitor cluster performance and configure alarms
- AWS IAM to configure permissions
- AWS CloudTrail to audit requests made to the service
- AWS Data Pipeline to schedule and start your clusters

EMR Storage options

- HDFS
- EMRFS: access S3 as if it were HDFS

- Local file system
- EBS (Elastic Block Storage) for HDFS

5. Amazon Mechanical Turk