Getting to Know your Data

I. Data Objects and Attribute Types

Types of Datasets

1. Record-based Datasets.
   - Relational Records: Traditional tables with rows and columns, such as those in databases
   - Data Matrix: A numerical representation where row represent entities and columns represent attributes.
   - Document Data: Text-based datasets represented using term-frequency vectors, where each document is represented by the frequency of terms it contains
   - Transaction Data: Logs of transactional information, such as sales data (e.g. "items bought together")

2. Graph and Network Datasets
   - World Wide Web: Hyperlinked web pages from a direct graph structure
   - Social or Information Networks: Nodes represent entities (e.g. people), and edges represent relationships (e.g. friendship or information flows)
   - Molecular Structures: Chemical compounds where nodes represent atoms, and edges represent bonds

3. Ordered Datasets
- Video Data : Sequences of images forming a video
- Temporal Data : Times-series data representing observations over time (e.g. stock prices, weather patterns.)
- Sequential Data : Ordered sequences, such as clickstream data or logs of transaction events.
- Genetic Sequence Data : DNA, RNA, or protein sequences, represented by nucleotide or amino acid sequences

4. Spatial, Image, and Multimedia Datasets
- Spatial Data : Geographic information such as maps or geospatial coordinates
- Image Data : Visual data represented as pixel matrices (e.g. photos, satellite images)
- Video Data : Combines spatial and temporal aspects, consisting of sequences of images over time.

- What is Data?

Data is a collection of objects and their attributes, used to describe and analyze entities.

Attributes / Features

Tid	Refund	Married ?	Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No

Observation → Objects / Records

Key Concepts

1) Data Objects:
- represent entities or elements on which data is collected
- Also referred to as records, points, cases, samples, or instances

2) Attributes:
- properties or characteristics that describe data objects.
- Examples: eye color, temperature, etc.
- Also known as variables, features, fields, dimensions

3) Observation:
- A set of measurments collected for all attributes of a single data object

By Combining attributes for elements, data is structured for analysis

Types of Attributes:

1. Nominal Attributes
- Categories with no inherent order
- Examples:
  - Eye color {blue, green, brown}
  - Gender {male, female}

2. Ordinal Attributes
→ Ordered categories, but differences between ranks are not meaningful
→ Examples:
. Taste ranking: {poor, fair, good, excellent}
. Education Level: {high school, college, graduate}

3. Interval Attributes
→ Numeric attributes with meaningful intervales but no True zero
→ Examples:
. Temperature in Celcius or Fahrenheit
. Calendar years

4. Ratio Attributes
→ Numeric attributes with meaningful intervales and a True zero
→ Examples:
. Age, weight, Salary
. Distance measurement

Type	Operations	Applications
Nominal	=, ≠	Identification: Zip codes, IDs Grouping or Clustering
Ordinal	<, >	Preference analysis: Products ratings or customer feedback.
Interval	+, −	Time or Trend Analysis: Tracking changes over time.
Ratio	+, −, ×, ÷	Ratio Analysis, Financial Metrics: Analyzing proportions, ratio, or growth

The type of Attributes determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

Discrete vs. Continuous Attributes

1. Discrete Attributes

- Finite or countable values
- Ex: word count, transaction count, zip codes
- Characteristics:
  - Represented by integers
  - Finite or countably infinite set of values
  - Special case: Binary attributes (e.g. yes/no)

2. Continuous Attributes

- Real numbers within a range
- Ex: Height, Weight, Temperature
- Characteristics:
  - Measured and represented with floating-point values.
  - Requires rounding or precision adjustments in real-world use.

II. Descriptive Statistics

1) Summarizing Categorical Data

. **Frequency Distribution:**
Tabular summary of data showing the number (frequency) of items in each of several classes (categories)

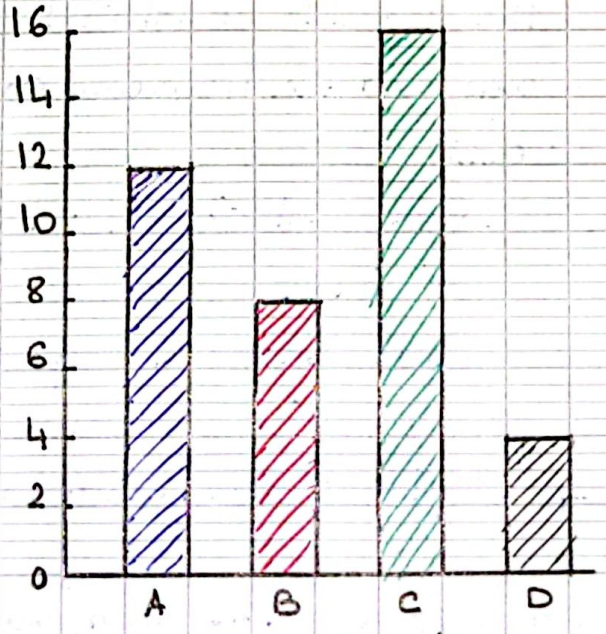
→ Relative Frequency of a class $= \frac{\text{Frequency of the class}}{n}$

n: total nb of observations

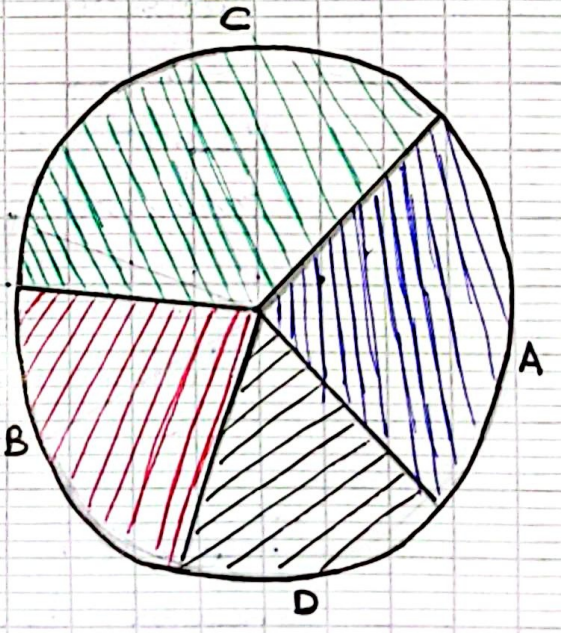
. **Percent Frequency Distribution:**
Summarize the percent frequency of the data in each class

→ Percent Frequency of a class $= \frac{\text{Frequency of the class}}{n} \times 100$

. **Bar and Pie Chart:**
Graphical device for presenting relative frequency and percent frequency distributions for categorical data



Bar Chart



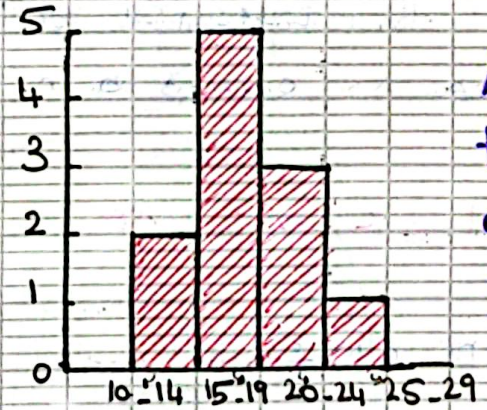
Pie Chart

- In data Exploration, frequency dist. and percent frequency for categorical variables help us understand how the data is distributed among different categories
- These tools and measures highlight patterns, dominance, or underrepresentation among categories.
- They help identify data quality issues (e.g. missing or unbalanced categories)

2) Summarizing Numerical Data

- Relative frequency and percent frequency for numerical data, are defined in the same manner as for qualitative data.

- Histogram is a common graphical representation of numerical data



Histogram

A histogram is constructed by placing the variable of interest on the horizontal axis and the frequency on the verlical axis

↳ Bins: Intervals dividing the range of values

↳ Bars: Represent frequencies within bins

- These tools help identify the distribution type (normal, skewed, etc.)
- They reveal key summary stats like mean, median, and mode (indirectly)
- They help detect anomalies (e.g. outliers, extreme values)

3) Measures of Central Tendency

Central tendency provides a single value that represents the "center" of a dataset.

3.1. Mean (Average)

- What it is: The sum of all values divided by the number of values. → $\bar{x} = \sum_{i=1}^{N} x_i / N$
- Interpretation:
  - Represent the "center" of the data if the distribution is symmetric.
  - Sensitive to outliers; a single extreme value can skew it significantly.
- Use in ML: Common for normally distributed data but less reliable for skewed distributions

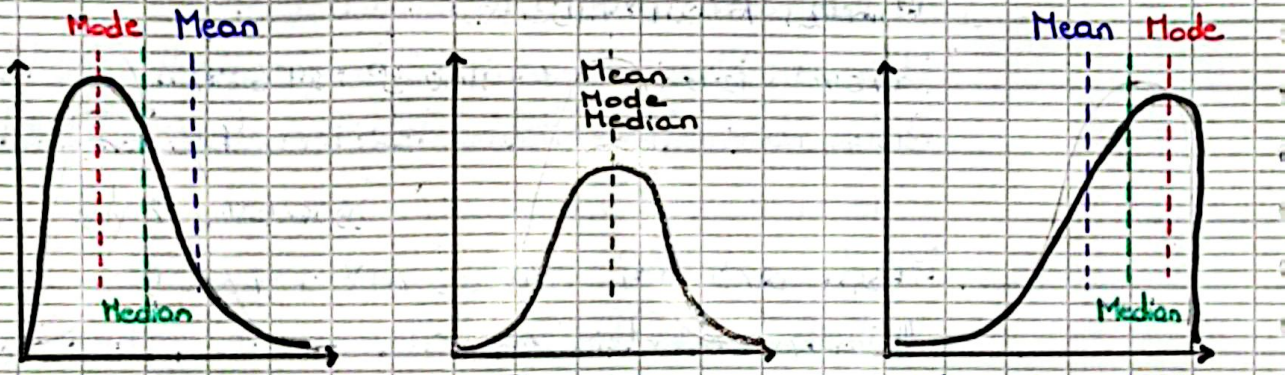
3.2. Median

- What it is: The middle value when the data is sorted (or the average of the 2 middle values for an even nb of observations)
  - if N is odd: $m = N+1/2$ with N: nb of observations
  - if N is even $m = (L + L+1)/2$ with $L = N/2$
- Interpretation:
  - Represent the 50th percentile, unaffected by outliers.
  - Indicates the typical value in skewed dist.
- Use in ML: Often used when the data contains outliers or is skewed.

3.3. Mode

- What is it: The most frequently occurring value(s) in the dataset
- Interpretation:
  - Indicates the most common or dominant value
  - Useful for categorical data or datasets with repeated values.
- Use in ML: Relevant for identifying dominant categories or clusters.

3.4. Symmetric vs. Skewed data



Positively Skewed
mean > median

Symmetric
mode = mean = median

Negatively Skewed
mean < median

- Symmetric Data
  - Indicates a normal distribution (bell-shaped curve)
  - Suitable for models that assume normality (e.g. linear regression).

- Skewed Data
  - Indicates potential outliers or extream values
  - May require Transformations (e.g. log or square root) to normalize for certain models.

3.5. Why Central Tendency matters in ML and DS

→ Data Distribution Analysis:
- It provides a quick summary of where most data points lie.
- It helps assess whether the data is skewed, symmetric, or contains extrem values.

→ Feature Engineering:
- Helps determine if transformations (e.g log, normalization) are necessary
- Skewed data might need normalization to improve model performance

→ Model Assumptions:
- Some algo. (e.g. linear regression, SVMs) assume a normally distributed dataset.
- Understanding Central Tendency helps check how well the data fits these assumptions

→ Anomaly Detection:
- Highlights potential anomalies by showing values that deviate significantly from the "typical" range

→ Benchmarking:
- Helps establish baselines for comparing data before and after preprocessing or transformations

4) Measures of Dispertion

Dispertion describes the spread or variability in a dataset.

4.1. Range

- What it measures: The difference between the maximum and minimum values in the dataset
- Interpretation
  - Shows the total spread of the data
  - A high range might indicate high variability or outliers
  - Doesn't account for data distribution, so it's sensitive to extreme values.

4.2. Variance

- What it measures: The average of the squared differences between each data point and the mean.
  - Variance = $\frac{\sum (x_i - \bar{x})^2}{N}$
- Interpretation
  - Indicates how far data points are from the mean on average
  - A high variance suggests a wide spread, a low variance indicates clustering around the mean

4.3. Standard Deviation (SD)

- What it measures: The square root of the variance representing dispersion in the same units as the data
  - ↳ $SD = \sqrt{Variance}$
- Interpretation
  - A widely used measure of spread
  - Higher SD implies greater variability; lower SD implies consistency
  - Easier to interpret than variance as it's in the same unit as the data

4.4. Percentiles and Quartiles

Measures that describe the distribution of data by dividing it into ranked portions

- ↳ Percentiles
  - What they are: They devide the dataset into 100 equal parts
    - The n-th percentile is the value below which n% of data falls.
    - Example: The 25th percentile (P25) means 25% of the data is below this value, and 75% is above
  - Interpretation:
    - Helps identify thresholds or cutoffs in data (A student scoring at the 90th percentile performed better than 90% of their peers)
    - Useful for spotting outliers (e.g. values below the 1st percentile or above the 99th percentile.

→ Quartiles

- What they are: Divide the dataset into four equal parts based on percentiles
- Q1 (1st Quartile): The 25th percentile - The value below which 25% of the data lies
- Q2 (2nd Quartile): The 50th percentile - Splits the data into two halves.
- Q3 (3rd Quartile): The 75th percentile - The value below which 75% of the data lies

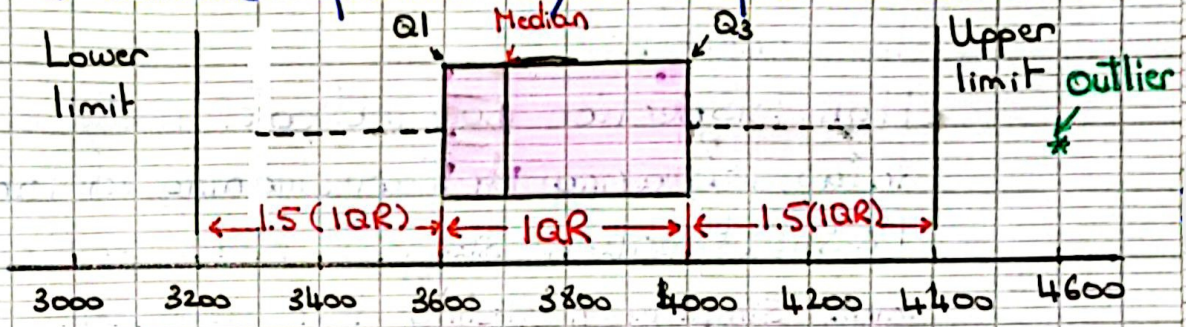
4.5. Interquartile Range (IQR)

- What it measures: The range of the middle 50% of the data
  - → IQR = Q3 - Q1
- Interpretation
  - → Focuses on the central portion of the data
  - → Useful for detecting and handling outliers (Values outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are potential outliers)
- Example: In Income data, if Q1 is $30,000 and Q3 is $70000, the middle 50% of people earn between these amounts.

- The five-number summary of distribution consist of:

Minimum, Q1, Median, Q3, Maximum

. Box Plot:
Popular way of visualizing a distribution.
It incorporates the five-number summary as follow:
→ Ends of the box are at the quartiles so that the box length is the interquartile range.
→ Median is marked by a line within the box.
→ Two lines (called whiskers) outside the box extend to minimum and maximum
→ Outliers: points beyond a specified outlier threshold



4.6. Why Measures of Dispersion are Important in ML and DS
→ Feature Scaling:
. Variability across features can affect model performance, espacially for distance based algorithms(e.g. KNN, SVM). Understanding dispersion helps decide whether to normalize or standardize features.
→ Outlier Detection.
. High dispersion may indicate the presence of outliers. IQR and SD are particularly helpful in identifying such points.

→ Model Selection.

. Models like Linear Regression assume low variability in residuals. High dispersion might require transformations or alternatives models

→ Data Insights.

. Dispersion highlights heterogeneity in data. (diverce or varied characteristics, dist, or patterns) which might indicate the need for stratification or feature engineering.

→ In General:

. High Dispertion may indicate noisy or unpredictable data, requiring further cleaning or analysis

. Low Dispertion suggests more uniformity, which migh simplify modeling but also risk overfitting if there's insufficient variability for learning.

. These insights are suitable for preprocessing and deciding which ML techniques are suitable for the data.

III- Visual Data Representation

1. Box-and-Whisker Plot

. Visualize the distribution of a dataset using summary statistics.

. Components:

- Box: Interquartile Range (IQR: Q3 - Q1)
- Whiskers: Extend to the smallest and largest values within 1.5 x IQR
- Outliers: points beyond whiskers, plotted individually.

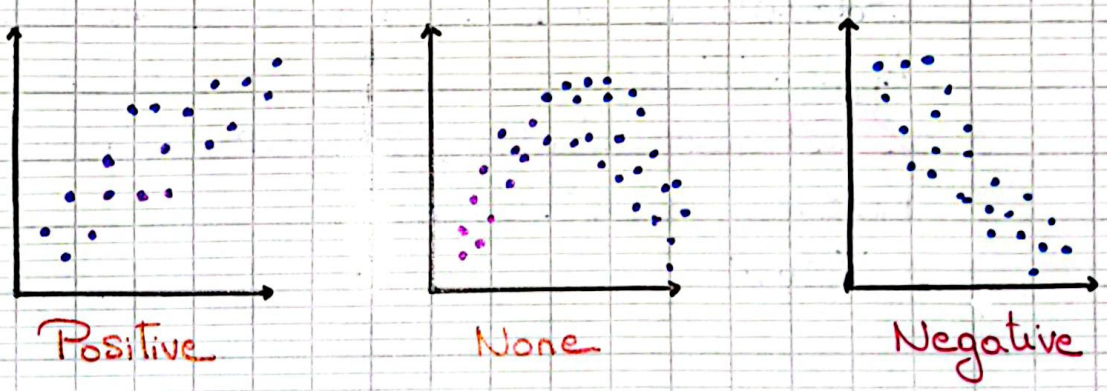
. Interpretation
  - A short box indicates low variability within the middle 50% of data.
  - Long whiskers may suggest higher variability or the presence of outliers

2. Scatter Plot

. Identify relationships or patterns between two numeric variables

. Components:
  - Each point represents a pair of values
  - x-axis : Independent variable
  - y-axis : Dependent variable

. Types of Correlations
  - Positive : x and y increase together
  - Negative : x increases while y decreases
  - None : No discernible pattern



3. Quantile Plot

. Visualize all data points and their cumulative behavior

. Steps:
1. Sort the data in ascending order
2. Plot each data point (x_i) against its quantile rank (p)

. Interpretation: Identify how values are distributed across quantiles

4. Quantile - Quantile plot (q-q plot)

. Compare two distributions by plotting their quantile against each other

. Steps:
1. Calculate quantiles for each dataset
2. Plot quantiles from dataset A (x) vs. dataset B (y)

. Interpretation
- points along a 45-degree line suggest similar distributions
- deviations indicate differences in distributions

. Applications:
- Test normality (e.g. compare dataset to a normal distribution