ASSIGNMENT-2

Q1. What methods are used in 2nd Business of moment also provide its formula.

Ans:

* Methods Used:

1. Variance (σ²): This measures how far a set of numbers is spread out from their average (mean). Variance is the average of the squared differences from the mean.
   * Formula for Population Variance:

σ2=N1​i=1∑N​(xi​−μ)2

* + Formula for Sample Variance:

S2=n−11​i=1∑n​(xi​−xˉ)2

1. Standard Deviation (σ): This is the square root of the variance and provides a measure of the average distance of each data point from the mean.
   * Formula for Population Standard Deviation:

σ=square root of σ2

* + Formula for Sample Standard Deviation:

s=square root of s2​

Use of 2nd Business Moment:

* Understanding Variability: The second moment (variance) helps assess how much individual data points differ from the mean, giving insight into the variability of the data.
* Risk Assessment: In finance and risk management, variance is crucial for understanding the volatility of asset prices or returns.
* Quality Control: In manufacturing, variance helps determine whether a process is stable and within specifications.

Example:

Consider the following data set representing exam scores:

70,75,80,85,90

1. Calculate the Mean (μ):

μ=570+75+80+85+90​/5=80

1. Calculate the Variance (σ2):

σ2=5/1​((70−80)2+(75−80)2+(80−80)2+(85−80)2+(90−80)2)

σ2= 51​(100+25+0+25+100)=250/5​=50

1. Calculate the Standard Deviation (σ\sigmaσ):

σ=50​≈7.07

Q2. What is variance, why do you calculate variance.

Ans:

* This measures how far a set of numbers is spread out from their average (mean).Variance is the average of the squared differences from the mean.
* How away your data from mean.
* Example :
  + Formula for Population Variance:

σ2=N1​i=1∑N​(xi​−μ)2

* + Formula for Sample Variance:

S2=n−11​i=1∑n​(xi​−xˉ)2

* Why do you calculate variance ?

Variance measures how far the data points are spread out from the average (mean). It shows how much the values in a dataset differ from the typical value, helping to understand the overall variability or consistency in the data.

Q3. What is Standard deviation and why do we calculate.

Ans:

* Standard deviation is a measure that tells us how spread out the values in a dataset are from the average (mean). In data analysis, it helps us understand the consistency or variability of the data.
* Why do we calculate it?
  + To see if the data points are close to the average (low variability) or spread out (high variability).
  + It helps detect outliers (unusual values) and assess risk or stability in data.
  + It makes it easier to compare different datasets by showing how much variation exists in each.
* For example, in sales data, a low standard deviation means sales are stable, while a high one means sales are fluctuating widely.

Key Difference:

* Variance gives a broader measure of spread by squaring differences, but it’s in squared units.
* Standard deviation simplifies this by taking the square root, so the result is in the same units as the data, making it easier to understand.

Q4. What do you understand by 3rd Moment and 4th Moment of business.

* The 3rd moment of a distribution is associated with skewness, which measures the asymmetry of the data distribution around its mean.
* Interpretation:
* Positive Skewness: Indicates that the right tail of the distribution is longer or fatter than the left tail, meaning that there are outliers on the higher end.
* Negative Skewness: Indicates that the left tail is longer or fatter than the right tail, meaning that there are outliers on the lower end.
* Zero Skewness: Indicates a symmetric distribution.
* Formula for Skewness:

Skewness = (x-mean)3 /​ σ3

* Use in Business:

Understanding skewness helps businesses identify trends in sales data, customer behavior, or financial metrics. For example, if sales data shows positive skewness, it may indicate that a small number of products are driving high revenues.

* The 4th moment of a distribution is associated with kurtosis, which measures the "tailedness" or the sharpness of the peak of the distribution.
* Interpretation:
  + High Kurtosis: Indicates a distribution with heavy tails and a sharper peak (more outliers). This is known as leptokurtic.
  + Low Kurtosis: Indicates a distribution with lighter tails and a flatter peak (fewer outliers). This is known as platykurtic.
  + Mesokurtic: A normal distribution has a kurtosis of about 3, which serves as a baseline.
* Formula for Kurtosis:

​

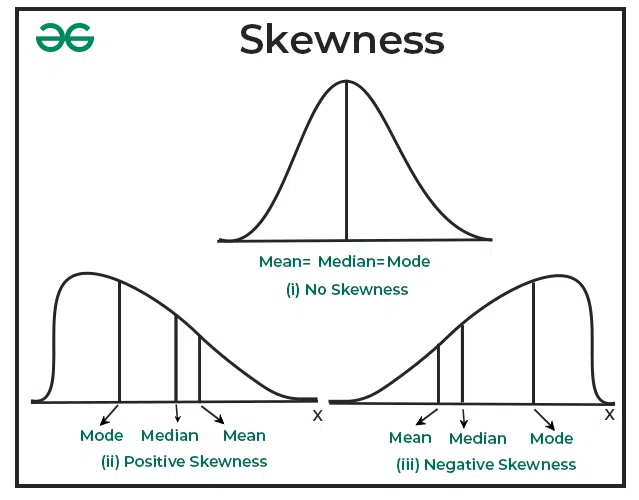
* Use in Business:
  + Kurtosis helps businesses assess risk and volatility. For instance, in finance, a high kurtosis in return distributions might indicate a higher probability of extreme returns (both high and low), which is crucial for risk management and investment strategies.

Summary:

* 3rd Moment (Skewness): Measures the asymmetry of a data distribution, helping businesses understand the direction of outliers and trends.
* 4th Moment (Kurtosis): Measures the tailedness of a data distribution, helping businesses assess the risk of extreme values in their data.

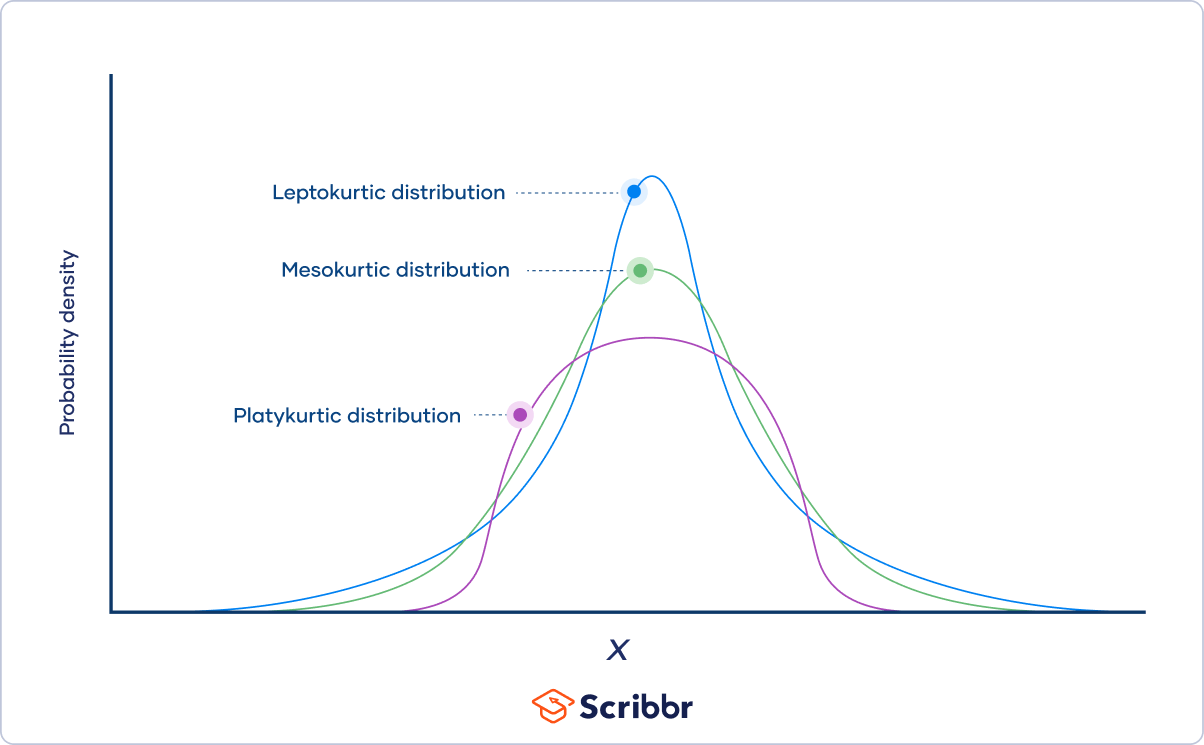
Q5. What are different types of skewness and how to differentiate between them.

* Positive Skewness: Indicates that the right tail of the distribution is longer or fatter than the left tail, meaning that there are outliers on the higher end.
* Negative Skewness: Indicates that the left tail is longer or fatter than the right tail, meaning that there are outliers on the lower end.
* Zero Skewness: Indicates a symmetric distribution



Q6. What are different types of kurtoses and how to differentiate between them.

* High Kurtosis: Indicates a distribution with heavy tails and a sharper peak (more outliers). This is known as leptokurtic.
* Low Kurtosis: Indicates a distribution with lighter tails and a flatter peak (fewer outliers). This is known as platykurtic.
* Mesokurtic: A normal distribution has a kurtosis of about 3, which serves as a baseline.



Q7. Let’s say you have calculated the skewness and kurtosis with ‘+’ve value what relation would you be observing in Mean, Median, Mode. Draw a diagrammatic view and justify your answer.

Ans:

* Positive skewness means Mean > Median > Mode.
* Positive kurtosis means the distribution has more extreme values and a sharper peak, indicating a more concentrated distribution with some outliers.

Q8. What are different steps in data pre-processing name any 5 which you think are crucial.

Q9. What do understand by type casting?

* In data analytics, type casting refers to the process of converting data from one type to another. This is crucial when working with datasets where data types (such as integers, floats, strings, or categorical types) need to be aligned for proper analysis or computation.
* Why Type Casting is Important in Data Analytics:
  + Data Compatibility: Certain operations or functions may require data to be in a specific type. For example, mathematical operations like summing values can only be performed on numeric data types (integers or floats).
  + Avoiding Errors: Mismatched data types can lead to errors or incorrect results. For instance, trying to perform arithmetic operations on strings can cause errors.
  + Efficient Processing: Converting data types can improve the efficiency of computations. For example, converting floating-point numbers to integers can reduce memory usage.
* Common Use Cases:
  + Data Cleaning: To ensure that all columns have the correct data types before performing any analysis or applying machine learning models.
  + Feature Engineering: Converting categorical features to numerical representations for model training.
  + Performance Optimization: Reducing data size by converting data types to more appropriate ones (e.g., using float32 instead of float64).

Q10. How would you justify if data is duplicate in row and in column.

1. Row Duplicates:

* Valid: If the duplicates represent legitimate, repeat data (e.g., multiple orders by the same customer).
* Invalid: If the same transaction or record appears more than once by mistake.
* Tools: Use Conditional Formatting to highlight and Remove Duplicates to clean erroneous rows.

2. Column Duplicates:

* Valid: Repeated values in non-unique columns (e.g., "Country" or "City").
* Invalid: Duplicates in unique identifier columns (e.g., "Order ID") suggest data issues.
* Tools: Use COUNTIF to check occurrences or Remove Duplicates to clean the column.

3. Action:

* Keep valid duplicates if they make sense in context. Remove or correct erroneous duplicates.
* Document your actions for clarity

Q11. Why one should be conscious to check data is balanced or not, how to identify and what method one can apply on it.

Ans :

* Imbalanced dataset is a problem it can lead to biased models and inaccurate predications
* Imbalanced dataset have many more instances of certain classes than of others
* Imbalanced data can lead to biased models, where the algorithm favors the majority class and under-represents the minority class.
* How to Identify Imbalanced Data?
  + Frequency Counts: Check the count of each category (especially for classification tasks). A large difference in class distribution signals imbalance.
  + Visualization: Use bar charts or pie charts to visualize the distribution of categories.
  + Statistical Measures: For regression, calculate measures like variance or spread to detect imbalances in continuous variables.
* To address this issue various techniques are used to balance the dataset including over sampling ,under sampling and SMOTE(synthetic minority over sampling)

Q12. What are different variants of missing values, what are different methods you would apply to overcome the issue. Explain when to choose mean and median for the same.

**Variants of Missing Values:**

1. MCAR (Missing Completely at Random):
   * Missing values occur entirely at random, with no relationship to other data. Example: A survey respondent accidentally skips a question.
2. MAR (Missing at Random):
   * Missing values depend on other observed data but not on the missing data itself. Example: Income data might be missing for certain age groups.
3. MNAR (Missing Not at Random):
   * The reason for the missing data is related to the missing value itself. Example: People with higher incomes may be less likely to report their income.

**Methods to Handle Missing Data:**

1. Deletion Methods:
   * Listwise Deletion: Remove entire rows that have missing data. Used when the proportion of missing data is small and missingness is random (MCAR).
   * Pairwise Deletion: Only exclude missing data from specific analyses, but keep the row for other analyses. Suitable for MAR scenarios with partial missingness.
2. Imputation Methods:
   * Mean Imputation: Replace missing values with the mean of the observed data for that variable. This works well if the data is normally distributed and the missingness is MCAR.
   * Median Imputation: Replace missing values with the median. Preferred if the data is skewed (e.g., income data).
   * Mode Imputation: Replace missing categorical values with the mode (most frequent category). Effective for categorical variables with MCAR.
   * K-Nearest Neighbors (KNN): Replace missing values based on the closest observations in the data. This works when the missing data has a relationship with other variables.
   * Multiple Imputation: Generate several different imputed datasets and combine results. Suitable for MAR and complex datasets.
   * Predictive Models: Use machine learning models like regression or decision trees to predict missing values based on other data.

**When to Choose Mean vs. Median:**

* Mean Imputation:
  + Use when the data is normally distributed or the variance is low.
  + Advantage: Keeps the mean of the dataset unchanged.
  + Drawback: Sensitive to outliers, which can skew the imputed value.
* Median Imputation:
  + Use when the data is skewed (e.g., income, age) or contains outliers.
  + Advantage: Robust to outliers and provides a more accurate central tendency for skewed data.
  + Drawback: May not capture the true nature of the distribution as effectively as the mean in some cases.

**Choosing between mean and median depends on the distribution of the data:**

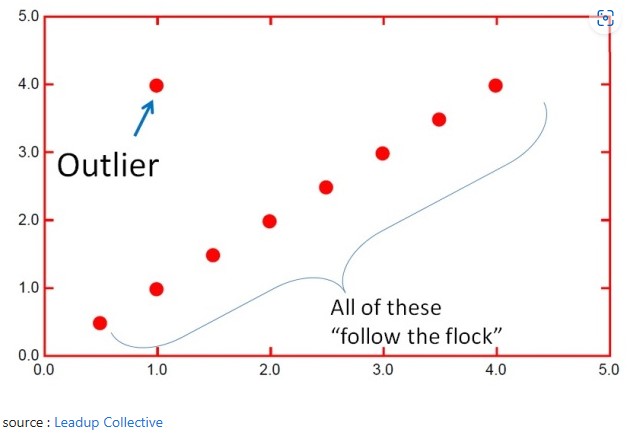
* Mean: Better for symmetric data without outliers.
* Median: Better for skewed data with outliers.

Q13. How Variance is useful while selecting features for analysis. Q14. How to check outlier in data and name the visual used to show.

* Outlier : Values in our data which are unsual for a given context are called outliers

Or

outliers are those specific data points that differ significantly from other data points in the dataset.



* Interquartile Range (IQR): Any data point that falls below Q1 - 1.5*IQR or above Q3 + 1.5*IQR is considered an outlier.
* Visual Methods:
  + Box Plot: The most commonly used visual for detecting outliers. The "whiskers" in the box plot represent the range of normal data, and points outside of these whiskers indicate outliers.
  + Scatter Plot: A useful visual for identifying outliers in bivariate or multivariate data.
  + Histogram: Helps detect outliers by showing irregularly spaced bars or spikes in data distribution.

Q15. How to identify difference between suspected outlier and confirmed outlier.

* Suspected Outliers
  + Suspected outliers are data points that initially appear unusual compared to the rest of the dataset but require further investigation to be confirmed.
  + Use Case 1: Sales Data (E-commerce Sales)
    - Scenario: A retailer is analyzing daily sales data and notices that on a particular day, the sales are much higher than average.
    - Suspected Outlier: Using visual tools like a box plot or statistical methods like Z-scores, the day with unusually high sales is flagged because it is more than three standard deviations away from the mean.
  + Detection Methods:
    - Box Plot: Visualizing data to flag values outside the "whiskers."
    - Z-score: Identifying values that are beyond a Z-score threshold (e.g., > 3 or < -3).
    - IQR Method: Values that fall below Q1 - 1.5*IQR or above Q3 + 1.5*IQR are considered suspected outliers.
    - At this stage, the sales spike could be an outlier, but more investigation is needed to confirm.
* Confirmed Outliers
  + Confirmed outliers are data points that have been examined through statistical methods, domain knowledge, or external investigation and validated as genuine anomalies.
  + Use Case 2: Sensor Data (Temperature Readings)
    - Scenario: A weather station collects hourly temperature data from multiple sensors, and one sensor records a reading of 150°F, which is highly unusual for the region.
    - Confirmed Outlier: After verifying that the sensor is functioning correctly and no extreme weather event has occurred, the reading is confirmed as an outlier caused by a sensor malfunction.
  + Confirmation Methods:
    - Grubbs' Test: A statistical test to identify outliers in normally distributed data.
    - Domain Knowledge: Using expertise to determine if the outlier makes sense. In this case, 150°F is unrealistic, so it's likely an error.
    - Manual Investigation: Checking the context or real-world events that could explain the data point (e.g., system logs showing sensor malfunction).
    - Here, the outlier is confirmed to be an error and should be excluded from further analysis.

Comparison: Suspected vs. Confirmed Outliers

| Aspect | Suspected Outlier | Confirmed Outlier |
| --- | --- | --- |
| Detection | Detected through basic statistics or visual methods | Validated through deeper analysis or domain knowledge |
| Need for Further Review | Requires investigation | Has been thoroughly reviewed and verified |
| Tools | Box plots, Z-scores, IQR | Grubbs’ Test, Dixon’s Q Test, manual validation |
| Example | A day with unusually high sales | A temperature reading of 150°F from a faulty sensor |

Real-World Use Case 4: Financial Transactions (Fraud Detection)

In financial analytics, fraud detection often involves outliers. For example, unusual spending patterns can be flagged as outliers.

Suspected Outlier: A customer suddenly spends $10,000 on electronics in one day, while their usual spending is around $100 per day.

Confirmed Outlier: After confirming that the customer did not authorize the transaction and that it occurred from an unusual location, this outlier is confirmed as fraudulent activity.

Outcome: The transaction is flagged, and fraud prevention steps are taken.

Practical Identification Methods in Data Analytics:

Suspected Outliers:

* Visualization: Use box plots, scatter plots, or histograms to visually detect outliers.
* Summary Statistics: Calculate measures like mean, median, and standard deviation to identify values that fall far outside typical ranges.

Confirmed Outliers:

* Statistical Tests: Apply tests like Grubbs' test, Dixon's Q test, or Mahalanobis Distance to confirm whether a suspected outlier is statistically significant.
* Real-World Validation: Check the context or external factors to determine whether the data point reflects an actual anomaly or error.

Q16. What are different methods to resolve the outlier problems.

Outlier treatment – ‘3R’ technique 3R technique

• Rectify

• Retain

• Remove

Box Plot (Whiskers Plot)

• IQR = Q3-Q1

• LL = Q1-(1.5IQR)

• UL = Q3+(1.5IQR)

Q17. What is discretization and different types of it.

* Definition: Discretization is the process of converting continuous data into discrete categories or intervals (bins).
* Simplifies Continuous Data: Discretization converts continuous data (e.g., age, income) into smaller categories or groups (e.g., age groups like 20-30, 31-40). This makes the data easier to analyze.
* Improves Model Performance: Some algorithms work better with categories rather than continuous values. Discretization helps in cases where relationships between categories are more meaningful.
* Handles Noisy Data: By grouping data into intervals, it reduces the impact of small fluctuations or noise in the data.

**Types:**

* Binarization: Converts data into binary categories (0 or 1) based on a threshold.

Example: Suppose you have a dataset of people’s ages. If you want to classify people as either “young” or “not young,” you can define a threshold (e.g., 30 years). Anyone younger than 30 is assigned a value of 1 (young), and anyone older or equal to 30 is assigned a value of 0 (not young).

* Rounding: Rounds continuous data to a fixed precision to reduce noise or simplify data.

Example: If a dataset contains the height of people in centimeters, you might round each value to the nearest centimeter or nearest 10 centimeters (e.g., 167.5 cm becomes 168 cm).

* Binning: Divides continuous data into categories or intervals, which can be equal-width, equal-frequency, or custom-defined. 

Example: If you have a dataset of exam scores ranging from 0 to 100, you can bin the data into three categories: "low" (0-40), "medium" (41-70), and "high" (71-100).

Q18. What is encoding method, different types and their difference.

* Definition: Encoding involves converting categorical data into numerical values, which can be understood by machine learning algorithms.
* Why Needed:
* Model Compatibility: Most machine learning algorithms (like linear regression, SVM, etc.) require numerical input, so categorical data must be encoded into numbers.
* Preserves Data Information: Encoding methods like One-Hot Encoding and Label Encoding help in preserving the relationships between categories and help models process the data effectively.
* Improves Performance: Encoding allows models to make better predictions by translating qualitative data into quantitative formats, making patterns more identifiable.

Q19. Why is feature scaling used and different methods of applying them. Also mention their formulas.

* Transformation in data analytics refers to the process of converting data from one format, structure, or value set to another to make it suitable for analysis or modeling. This process can involve scaling, normalizing, encoding, or modifying data in various ways.
* Why is Transformation Required?

1. Improves Data Quality:
   * Raw data often contains inconsistencies, missing values, or noise. Transformation helps clean the data, making it more uniform and usable for analysis.
2. Makes Data Compatible:
   * Many algorithms require data to be in a specific format or range (e.g., numeric, scaled). Transforming the data ensures that it fits the requirements of machine learning models.
3. Reduces Skewness:
   * If the data distribution is heavily skewed, transformation (like logarithmic or square root transformation) can make it more symmetrical, improving model performance.
4. Enhances Interpretability:
   * Transformation can make data easier to interpret by simplifying complex relationships or patterns, making the analysis more meaningful.
5. Handles Different Scales:
   * Features with different scales (e.g., age vs income) can distort the model. Scaling transformations like normalization or standardization ensure that all features contribute equally.

Q20. As per your choice which method in scaling is your first choice, justify your answer.

Q21. What is transformation, why is it required.

Key takeaways

* Data transformation is a critical process that involves cleaning, filtering, aggregation, and converting data to facilitate analysis and decision-making, commonly implemented through ETL (Extract, Transform, Load) procedures.
* The process enhances data quality by standardising, reducing redundancy, and ensuring compatibility with analytical tools, playing a pivotal role in data analysis to enable data-driven business insights especially in cloud data warehouse settings.
* Data transformation faces challenges such as the high cost of specialist expertise and resource intensity but offers benefits like improved data quality, compatibility, and innovative data architectures; it involves stages such as discovery, mapping, execution, and review.

What is data transformation?

Data transformation involves converting data from one format or structure into another, often to make it more suitable for analysis or storage. This process can include:

* Data cleaning
* Data filtering
* Data aggregation
* Converting non-numeric features to numeric ones
* Resising inputs to a fixed size
* Adding, copying, and replicating data