Question 1

What are the five assumptions of linear regression ?

1. Linearity

- This assumption states that there is a linear relationship between the independent variables (predictors) and the dependent variable (outcome). In other words, the change in the dependent variable is proportional to the change in the independent variables.

- Example : The relationship between hours studied and exam scores, a linear assumption would imply that each additional hour of study results in a consistent increase in the exam score.

2. Independence of Errors/Residuals

- The residuals (errors) of the model should be independent of each other. This means that the error of one observation should not influence the error of another.

- **Example**: In time series data, if today's error in predicting a stock price is related to yesterday's error, this would violate the independence assumption.

3. Homoscedasticity

- The variance of the residuals should be constant across all levels of the independent variables

- **Example**: If you're predicting house prices based on square footage, the variability in price predictions should be roughly the same for both small and large houses.

4. Normality of Residuals

- The graph of residuals should be approximately normally distributed.

- In a model predicting employee salaries based on years of experience, the differences between the actual and predicted salaries should follow a normal distribution.

5. No Multicollinearity

- The independent variables should not be too much correlated with each other.

- If you have two predictors in your model, "number of rooms" and "house size in square feet," these two might be highly correlated as number of size corelates with house size.

Question 2

What is Bias - Variance tradeoff ?

1. Bias : It is the difference between the average prediction of the model and the actual value.

- High Bias : Model is too simple to capture the underlying patterns in the data (underfitting).

- Low Bias : Model is able to capture patterns in data.

2. Variance : It measures how much the model's predictions would change if it were trained on a different dataset.

- High Variance : Model fits the training data very well but fails on unseen data (overfitting).

- Low Variance : It performs well on training as well as unseen data/test data.

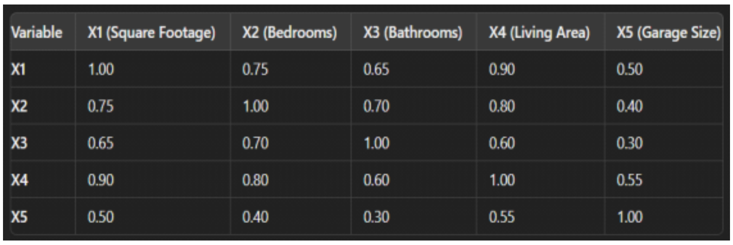
Best Model :
Low Bias and Low Variance

Question 3

In what ways you can handle Multicollinearity ?

1. Using correlation matrix

- Identify pairs (or groups) of independent variables that are highly correlated with each other and remove one of them.
- Use a correlation matrix or a heatmap to visualize correlations between variables.
- Let's explain with an example :



Variable	X1 (Square Footage)	X2 (Bedrooms)	X3 (Bathrooms)	X4 (Living Area)	X5 (Garage Size)
X1	1.00	0.75	0.65	0.90	0.50
X2	0.75	1.00	0.70	0.80	0.40
X3	0.65	0.70	1.00	0.60	0.30
X4	0.90	0.80	0.60	1.00	0.55
X5	0.50	0.40	0.30	0.55	1.00

- **X1 (Square Footage) and X4 (Living Area)**: Correlation = 0.90
- **X1 (Square Footage) and X2 (Bedrooms)**: Correlation = 0.75
- **X2 (Bedrooms) and X4 (Living Area)**: Correlation = 0.80

Here, the correlation between **above independent variables is too high**, indicating that these variables are strongly related. You can remove any one from each pair.

2. Using Variance Inflation Factor (VIF)

VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value greater than 5 or 10 indicates a problematic level of multicollinearity.

- Calculate the VIF for each predictor in your model.
- Remove predictors with high VIF values

3. Use ML models that have less effect of Multicollinearity

- Ridge Regression
- Decision Trees
- Random Forest
- Support Vector Machines with Regularization

Question 4

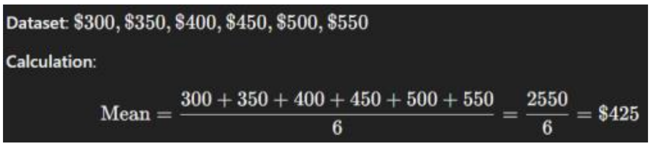
Explain mean , median , mode and standard deviation ?

1. Mean

- It the average of your given data. It's a measure of central tendency, which gives us an idea of where the centre of a dataset lies.

  Example:

  - **Scenario**: Suppose you want to find the average monthly spending on groceries for a family over 6 months.



Dataset: $\$300, \$350, \$400, \$450, \$500, \$550$

Calculation:

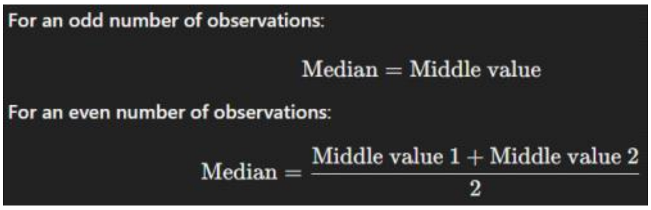
$$\text{Mean} = \frac{300 + 350 + 400 + 450 + 500 + 550}{6} = \frac{2550}{6} = \$425$$

  - **Interpretation**: The average monthly grocery spending for this family is $425.

  - It gives the central point of the data and is useful in various analyses and comparisons. However, its sensitivity to outliers means that it might not always be the best measure of central tendency, especially in skewed distributions.

2. Median

- The median is the middle value of a dataset when it's ordered from smallest to largest. It divides the dataset into two equal halves.
- Formulas :



For an odd number of observations:

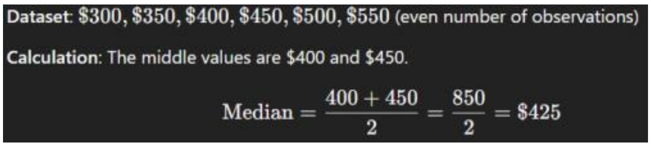
$$\text{Median} = \text{Middle value}$$

For an even number of observations:

$$\text{Median} = \frac{\text{Middle value 1} + \text{Middle value 2}}{2}$$

Example:

  - **Scenario**: Let's continue with the grocery spending example.



Dataset: $\$300, \$350, \$400, \$450, \$500, \$550$ (even number of observations)

Calculation: The middle values are \$400 and \$450.

$$\text{Median} = \frac{400 + 450}{2} = \frac{850}{2} = \$425$$

  - **Interpretation**: The median spending on groceries is $425, which is the same as the mean in this example.

  - It is a statistical measure used to represent the central value of a dataset, particularly when the data is skewed or contains outliers. It is robust, making it useful for imputation, evaluation metrics, and feature engineering

3. Mode

- The mode is the value that appears most frequently in a dataset. A dataset can have more than one mode or no mode at all if all values are unique.

- There is no specific formula for the mode; it's simply the most frequent value.

Example:

  - **Scenario**: Imagine you're analysing the colors of cars in a parking lot to determine the most popular color.

  - **Dataset**: Red, Blue, Red, Green, Blue, Red, Black, Blue, Red

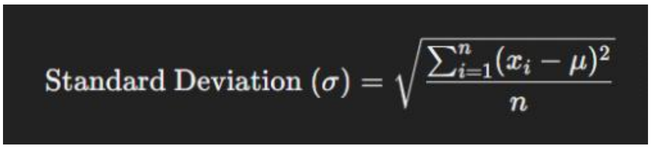
  - **Calculation**: The color "Red" appears 4 times, which is more frequent than any other color.

- **Interpretation**: The mode is "Red," indicating that red is the most popular car color in this parking lot.

- It is particularly useful for categorical and discrete data analysis, missing value imputation, feature engineering, evaluating class imbalance, and summarizing data distributions. While less common in continuous data analysis compared to the mean and median.

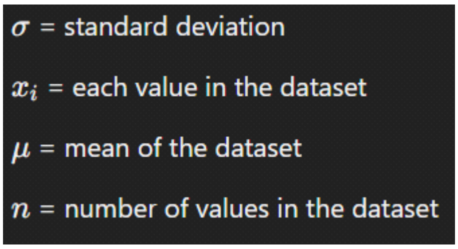
4. Standard Deviation

- It measures the amount of variation or dispersion in a dataset. A low standard deviation indicates that the values are close to the mean, while a high standard deviation indicates that the values are spread out over a wider range from mean.
- Formula :



$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Where



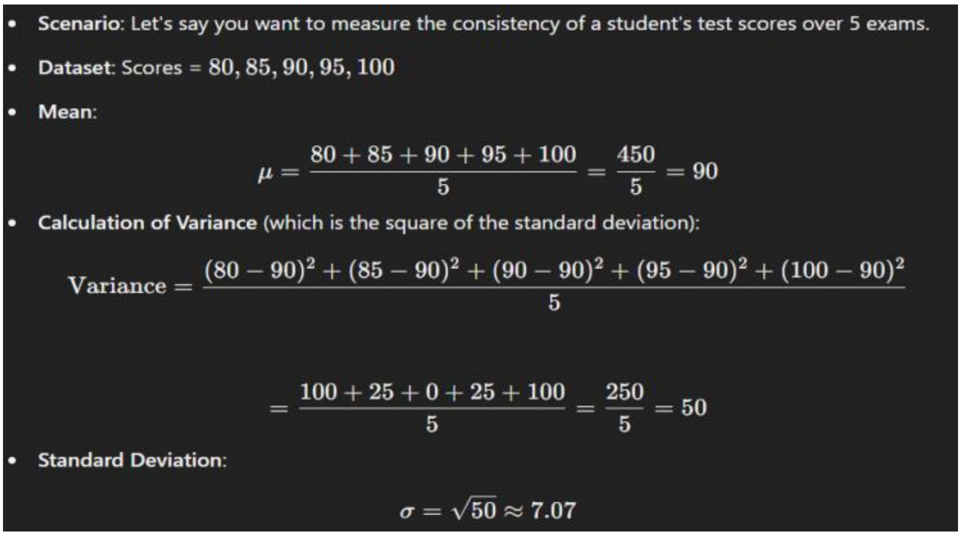
σ = standard deviation

x_i = each value in the dataset

μ = mean of the dataset

n = number of values in the dataset

Example :



- **Scenario:** Let's say you want to measure the consistency of a student's test scores over 5 exams.
- **Dataset:** Scores = $80, 85, 90, 95, 100$
- **Mean:**

$$\mu = \frac{80 + 85 + 90 + 95 + 100}{5} = \frac{450}{5} = 90$$

- **Calculation of Variance** (which is the square of the standard deviation):

$$\text{Variance} = \frac{(80-90)^2 + (85-90)^2 + (90-90)^2 + (95-90)^2 + (100-90)^2}{5}$$

$$= \frac{100 + 25 + 0 + 25 + 100}{5} = \frac{250}{5} = 50$$

- **Standard Deviation:**

$$\sigma = \sqrt{50} \approx 7.07$$

- **Interpretation**: The standard deviation of the student's test scores is approximately 7.07. This means that the scores typically deviate from the mean by about 7.07 points.

Question 5

What is assumption of Naïve Bayes Algorithm ?

Feature Independence (No Multicollinearity):

- **Assumption**: The Naive Bayes algorithm assumes that all features are conditionally independent given the class label. This means that the presence of a feature does not affect the presence of any other feature.
- When the assumption of feature independence is violated, the Naive Bayes model may produce biased or inaccurate probabilities which leads to **Misclassification** and **Reduced Accuracy**.

Question 6

What are outliers ? How we can detect/visualize them ? How we can handle them ?

Outliers are data points that differ significantly from other observations in a dataset. They can be unusually high or low compared to the rest of the data

To Detect/Visualize Outliers we use :

- **Box Plot:** Provides a graphical view of data spread and outliers.
- **Scatter Plot:** Useful for identifying outliers in two-dimensional data.

To Handle Outliers we use :

- **Z-Score:** Identifies data points with Z-scores greater than 3 or less than -3.

- **IQR (Interquartile Range):** Identifies outliers as those falling below Q1−1.5×IQR or above Q3+1.5×IQR.

- **Winsorization :** Outliers are capped at specified percentiles, which means extreme values are replaced with more moderate values

Question 7

What is Curse of Dimensionality ? How we can prevent it ?

The "curse of dimensionality" refers to various problems and challenges that arise when working with high-dimensional data. As the number of dimensions (or features) increases, these issues become more hectic.

1. **Increased Data Sparsity:**
   - **Problem:** In high-dimensional spaces, data points become sparse. Data points are wide spread.

   - **Effect:** This sparsity can make it difficult to find meaningful patterns or relationships in the data.

2. **Distance Metrics Become Less Informative:**
   - **Problem:** Distance-based methods (e.g., KNN) become less effective as dimensionality increases. In high-dimensional spaces, the distance between points becomes less distinguishable.

   - **Effect:** All points may appear to be roughly the same distance from each other, so distance based algorithms are unable to find patterns in data.

3. **Overfitting:**
   - **Problem:** With many features, models can become overly complex resulting in overfitting.

   - **Effect:** This leads to poor generalization on new, unseen data.

4. **Computational Complexity:**
   - **Problem:** The computational cost of processing high-dimensional data can be expensive. Algorithms may require significantly more time and resources.

   - **Effect:** Training models and performing computations become slower and more resource-intensive.

5. **Increased Risk of Multicollinearity:**
   - **Problem:** In high-dimensional spaces, features may become highly correlated with each other.

   - **Effect:** This multicollinearity can lead to various other problems.

How to handle Curse of Dimensionality :

- **Dimensionality Reduction/Feature Extraction:**
  - **Principal Component Analysis (PCA):** Reduces the number of features by transforming them into a smaller set of uncorrelated components while preserving most of the variance.

- **Feature Selection:**
  - **Filter Methods:** Use statistical tests to select the most relevant features (e.g., Chi-Square, ANOVA).

  - **Wrapper Methods:** Use model performance to evaluate feature subsets (e.g., Recursive Feature Elimination).

  - **Embedded Methods:** Perform feature selection as part of the model training process (e.g., Lasso regression).

- **Domain Knowledge:**
  - **Create New Features to replace Old ones :** Combine existing features to create new ones.

  - **Domain Knowledge:** Use domain expertise to identify and select the most meaningful features.

- **Increasing Data Size:**
  - **Acquire More Data:** Having more data points can help mitigate the effects of high dimensionality by providing a denser representation of the feature space.

- **Algorithm Choice:**
    - **Use Algorithms Robust to High Dimensionality:** Some algorithms (e.g., Random Forest , XGBoost) handle high-dimensional data better.

Question 8

What are different types of data ?

There are four types of data:

- Numerical

  a. Continuous
  b. Discrete

- Categorical

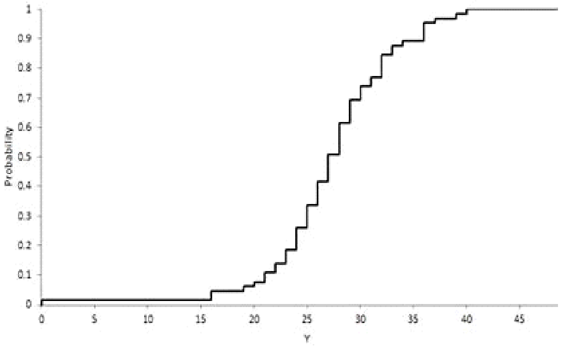
  a. Nominal
  b. Ordinal

Continuous	**Discrete**	**Nominal**	**Ordinal**
Continuous data can take any value within a given range. These are often in decimals too.	Discrete data consists of distinct, separate values. It can only take specific values, often counted in whole numbers.	Nominal data is categorical data without any inherent order.	Ordinal data is categorical data with a clear, meaningful order or ranking.
A person's height can be 165.5 cm, 165.55 cm, or 165.555 cm. The values can keep getting more precise.	A family can have 1, 2, or 3 children but unlike continuous data it cannot have 2.5 children.	**Gender**: Male, Female, Other. Here there is no order. We can't say Men > Women or Women> Men.	**Survey Ratings**: Poor, Good, Very Good, Excellent. Here, Excellent > Very Good > Good > Poor
The temperature could be 22.3°C, 22.33°C, or 22.333°C.	A person can own 1, 2, or 3 cars, but not 2.7 cars.	**Blood Type**: A, B, AB, O.	**Education Level**: High School, Bachelor's, Master's, Ph.D. Here according too importance of degrees, PHD > Masters > Bachelors >High School

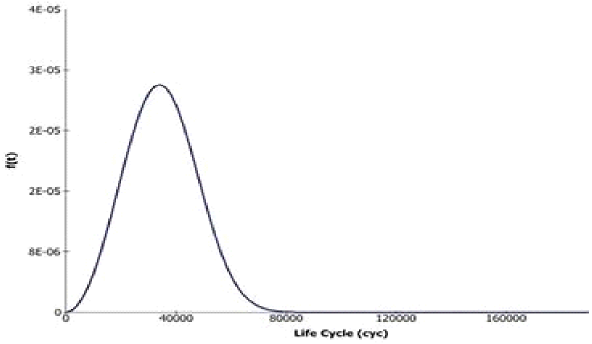
Question 9

Explain CDF(Cu**mulative Distribution Function)** and PDF(Probability Density Function) ?

Aspect	**CDF (Cumulative Distribution Function)**	**PDF (Probability Density Function)**
Definition	Tells you the total probability of getting a value less than or equal to a certain number.	It tells you how likely it is that the variable will fall within a specific range of values.
Example	Imagine you're filling a glass with water from a jug. The water level in the glass represents the height of the water. • **At Half-Full (50%)**: The CDF tells you the probability that the water level is at or below the halfway mark. If you've poured water and filled half the glass, the CDF gives you the total chance that the water is anywhere from the bottom up to halfway. • **At Full (100%)**: When the glass is completely full, the CDF at this point would be 100%, meaning the water level has reached or is below the full capacity of the glass.	Imagine you have a jar full of marbles of different colors. The PDF would show you which color you are most likely to pick if you reach into the jar. For example, if there are more red marbles, the PDF is higher for red, indicating a higher chance of picking a red marble.
Use in ML	Used to calculate percentiles, quantiles, and cumulative probabilities.	Used to model the distribution of data, such as in Gaussian Naive Bayes, KDE (Kernel Density Estimation), etc.
Graphical Appearance	The CDF is a non-decreasing curve that typically starts at 0 and approaches 1 as value increases.	The PDF can have peaks and valleys, showing how the probability density varies across different values.
Used on what type of Data	The CDF is defined for both discrete and continuous data.	The PDF is only defined for generally continuous data.
Unit of Measure	Unitless	Has units. On X-axis the same unit as feature unit and on Y-axis the probability.



CDF GRAPH



PDF GRAPH

Question 10

What is a kernel in Support Vector Machine ? Which type of kernel is used in which condition ?

A kernel function is a mathematical tool used in Support Vector Machines (SVMs) to transform data into a higher-dimensional space. This transformation helps in finding a separating hyperplane that can classify data more effectively, especially when the data is not linearly separable.

Example :

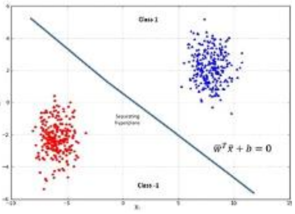
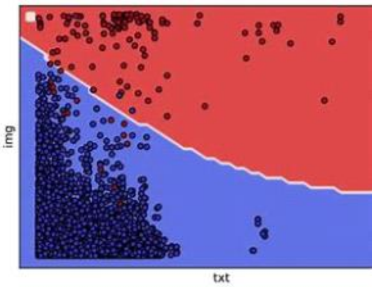
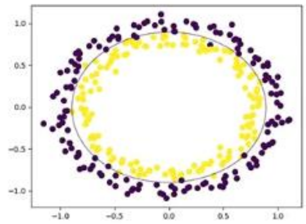
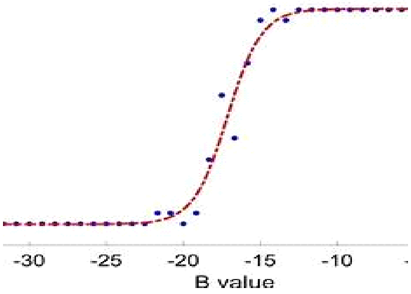
Imagine you have a bunch of fruit that are hard to separate into apples and oranges just by looking at their size and color in a 2D picture.

A kernel function is like adding a new feature (like texture) and changing the picture into a 3D space where apples and oranges are more easily separated by a plane.

This new perspective helps in drawing a clear boundary between the two types of fruit.

There are mainly four types of kernel function in SVM :

1. **Linear**
2. **Polynomial**
3. **RBF (Radial Basis Function)**
4. **Sigmoid**

Linear	**Polynomial**	**RBF**	**Sigmoid**
Use the linear kernel when your data is linearly separable	Use the polynomial kernel when your data requires a non-linear decision boundary that can be represented as a polynomial function. It's useful when you suspect that interactions between features are polynomial in nature.	Use the RBF kernel when you need to handle complex, non-linear relationships in your data. The RBF kernel is effective for data with no clear linear or polynomial boundary.	It is only used for Specific non-linear relationships and are very les used in SVM.
Imagine a 2D plot with two classes (e.g., circles and squares) where a straight line can cleanly separate them.	Imagine you have a graph of the height of a plant over time. If the height increases in a curved way rather than a straight line, this curved increase can be described using a polynomial function.	Imagine a 2D plot with circular clusters of data points. The RBF kernel can create a decision boundary that separates these clusters.	Imagine a decision boundary that resembles the shape of an S-curve.
 Class 1 Separating hyperplane $\vec{w}^T\vec{x}+b=0$ Class -1		 1.0 0.5 0.0 -0.5 -1.0	 -30 -25 -20 -15 -10 B value