Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

There were categorical variable for yr, mnth, season, weekday, holiday, weathersit, workingday

Overall the rentals are increasing year on year indicating a booming market

Fall and summer has highest rentals

Favorable weather condition considerably increase rentals

Weekdays vs Weekends is not significant but good weather conditions improve rental on any given day

Most rentals happen in Summer and fall months; need to make seasonal supply of rentals

---

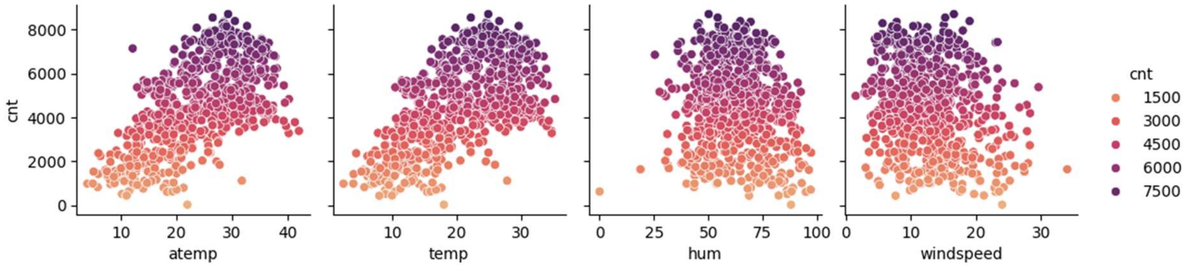
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
Dummy variable technique is used to represent categorical variables as binary variables – [0,1] for modelling. Dummy variable creation will create this binary variable for the category. The binary values will represent presence and absence of the categorical variables, called as indicator variable

The number of dummy variables created will be (n-1) where n is the number of categories in the categorical variables. Now when creating the dummy variables creating (n) dummy variables, where n is the number of categories may introduce multicollinearity issue as information in one category can be predicted from others. For this reason (n-1) rule is followed and python variable use drop_first=True while creating dummy variables to adhere to this rule.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

atemp (0.631) and temp (0.627) are highly correlated with target variable. ‘atemp’ is later deemed redundant with analysis as it is a derived value and is eliminated during model preparation

So ‘temp’ is the variable highly correlated with target variable



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

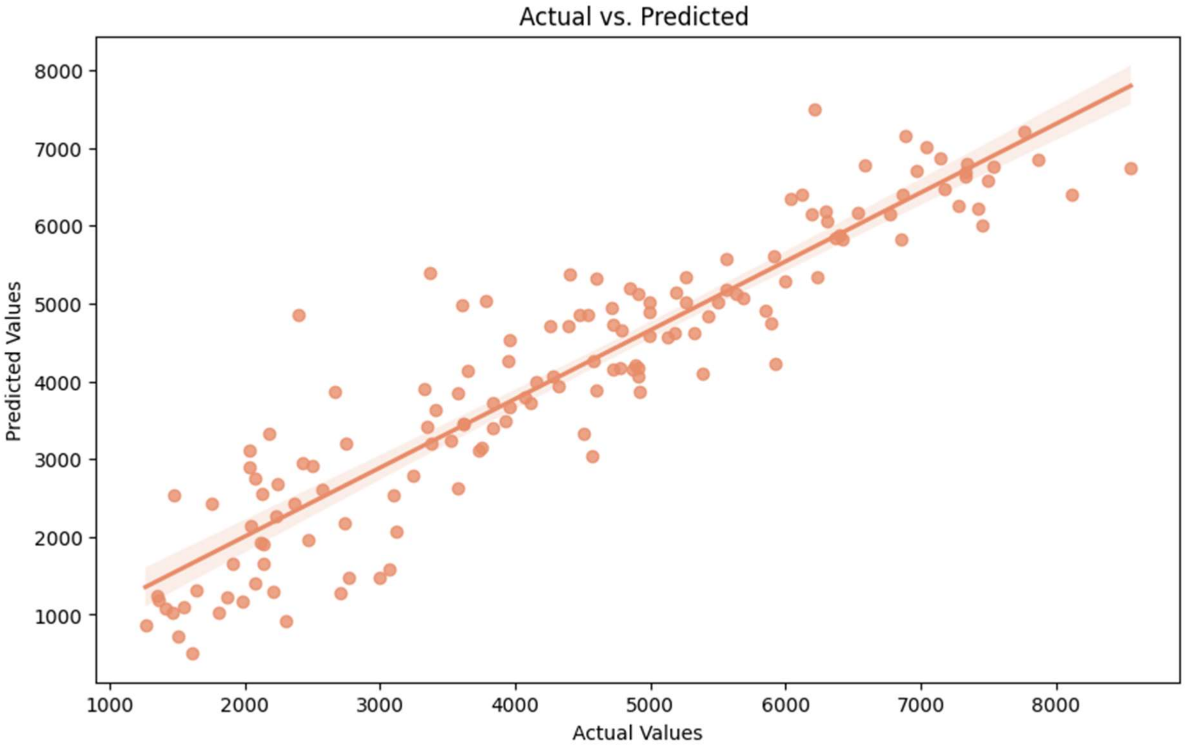
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Based on the analysis, I validated the Linear Regression assumptions as follows:

Linearity Assumption:

Linear relationship between dependent and independent variables:

a) Scatter plots showed linear relationships between variables
b) Actual vs Predicted plot showed points aligning with 45-degree line
c) Residuals scattered randomly around zero line
d) High correlation ($R^2 > 0.82$) confirms linear relationships
e) Slope of actual vs predicted close to 1



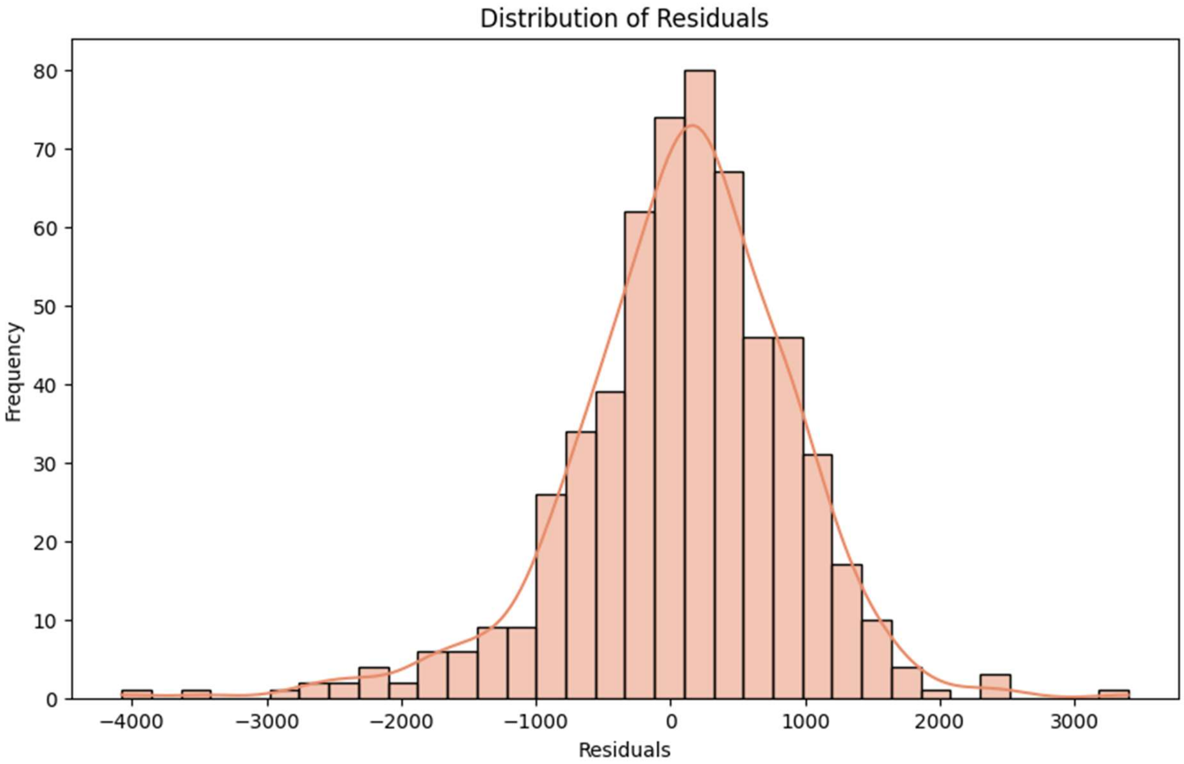
Independence of Residuals:

a) Durbin-Watson statistic = 2.026 (close to 2)
b) Indicates no significant autocorrelation in residuals
c) Validates independence assumption

Normality of Residuals:

Distribution plot of residuals showed:

a) Approximately bell-shaped curve
b) Generally acceptable for large sample size (n=584)

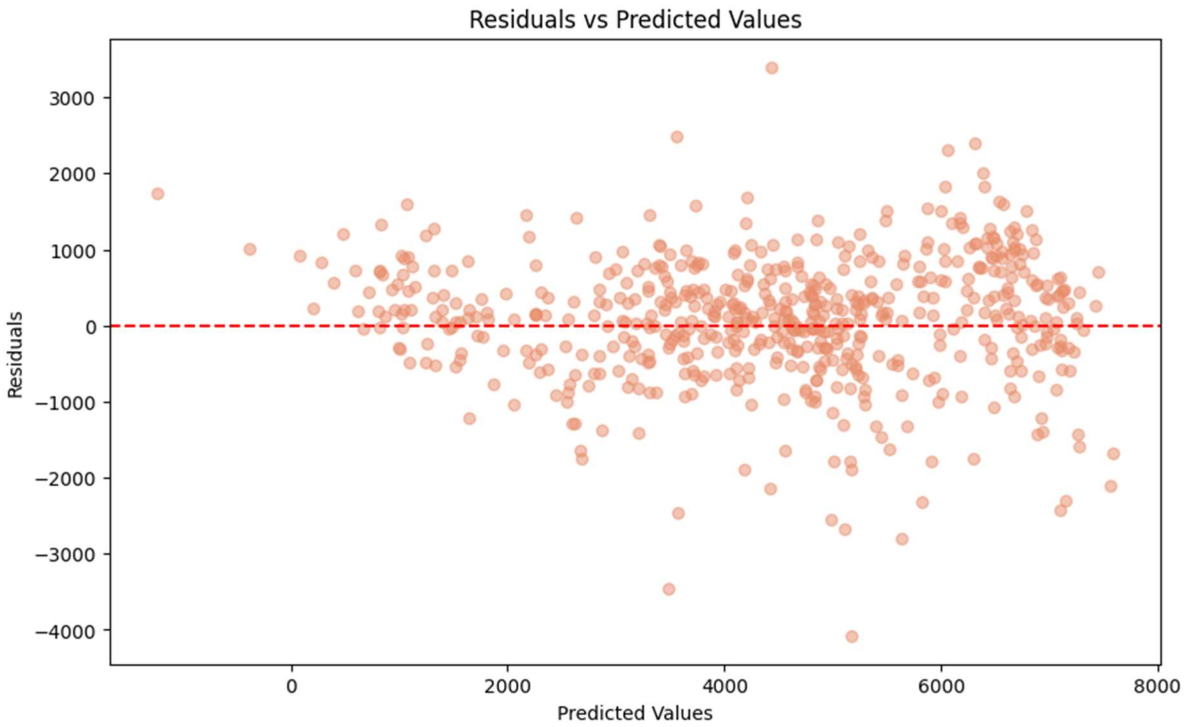


c)

Homoscedasticity (Constant Variance):

Residuals vs Predicted values plot showed:

a) Wider spread at higher predicted values
b) Variance increases with higher rental predictions



Multicollinearity:

a) Checked using VIF values
b) Removed 'atemp' due to high correlation with 'temp'
c) Final model showed acceptable VIF values except temperature
d) Kept temperature despite high VIF (5.001) due to its importance as predictor

Model Performance Validation:

a) R^2 Train = 0.8211, R^2 Test = 0.8296
b) Small difference (0.0085) indicates good generalization
    a. RMSE Train = 826.32, RMSE Test = 763.76
c) Consistent error metrics between train and test sets

These validations confirm the model's reliability

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the final model's coefficients, here are the top 3 features contributing significantly to bike demand:

Temperature (coef: 3520.22):

a) Strongest positive influence
b) Each unit increase in temperature leads to ~3520 more rentals
c) Highly significant (p-value < 0.001)
d) Key factor for demand prediction

Year (coef: 2020.62):

a) Indicates strong year-over-year growth
b) Shows good business potential
c) Highly significant (p-value < 0.001)

Weather_Light_Snow_Rain (coef: -2414.30):

a) Strongest negative impact
b) Decreases rentals by ~2414 bikes
c) Shows weather's crucial role in demand
d) Highly significant (p-value < 0.001)
e) Important for weather-based planning

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It predicts the value of the dependent variable based on the linear combination of the independent variables.

Types of Linear Regression:

Simple Linear Regression: One independent variable.
Multiple Linear Regression: Two or more independent variables.

Mathematical representation:

Linear regression assumes a relationship of the form:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

Where:

y: Dependent variable (target).

β0: Intercept (value of yy when all xi=0xi=0).
β1,β2,…,βnβ1,β2,…,βn: Coefficients of independent variables.
x1,x2,…,xnx1,x2,…,xn: Independent variables (features).
ϵ: Error term or residual.

Assumptions:

a) Linearity: The relationship between predictors and target is linear.
b) Independence: Observations are independent of each other.
c) Homoscedasticity: Constant variance of residuals across all levels of independent variables.
d) Normality: Residuals (errors) are normally distributed.
e) No Multicollinearity: Independent variables should not be highly correlated.

How it works:

Linear regression tries to find the best-fit line (or hyperplane) by minimizing the difference between the actual and predicted values of the dependent variable.

This difference is quantified using a cost function like the Mean Squared Error (MSE).

Cost Function:

$$\mathrm{MSE} = 1/\mathrm{n}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{yi\text{-}y\hat{}i})^2$$

yi: Actual value.
y^i: Predicted value.

Optimization:

Linear regression uses Ordinary Least Squares (OLS) or Gradient Descent to minimize the cost function.

OLS directly computes the optimal β values using:

$$\beta = \left(\mathrm{X}^{\mathrm{T}}X\right)^{-1}\mathrm{X}^{\mathrm{T}}\,\mathrm{y}$$

Prediction:

Once coefficients (β0,β1,…,βnβ0,β1,…,βn) are estimated, predictions are made as:
y^=β0+β1x1+β2x2+⋯+βnxn

Evaluation:

R-squared :

$$\mathrm{R}^{2=\,1\text{-}}\,\frac{\mathrm{SSresiduals}}{\mathrm{SStotal}}$$

Explains the proportion of variance in the target variable explained by the model.

Adjusted R-squared:

Adjusts R^2 for the number of predictors.

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE):

Quantify the error in predictions.

Advantages:

---

a) Simple to Implement: Easy to understand and implement.
b) Interpretability: Coefficients provide insights into the impact of independent variables.
c) Scalability: Works efficiently with small to medium datasets.

Limitations:

a) Assumptions Must Be Met:
- Violating assumptions (e.g., non-linearity, multicollinearity) can lead to unreliable results.

b) Sensitive to Outliers:
- Outliers can disproportionately affect the coefficients.

c) Limited to Linear Relationships:
- Cannot model complex, non-linear relationships without transformations or polynomial regression.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. This illustrates the importance of visualizing data and the limitations of relying solely on summary statistics. They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data during analysis and the effect of outliers on statistical properties.

The Four Datasets
Dataset I:

Shows a typical linear relationship
Data points follow an expected pattern with some random variation
Represents what we might consider a "normal" correlation

Dataset II:

Clearly shows a non-linear (curved) relationship
Demonstrates why Pearson correlation can be misleading for non-linear patterns
Perfect example of why linear regression isn't always appropriate

Dataset III:

Shows a perfect linear relationship except for one outlier
Demonstrates how a single point can significantly affect correlation
Highlights the importance of identifying and investigating outliers

Dataset IV:

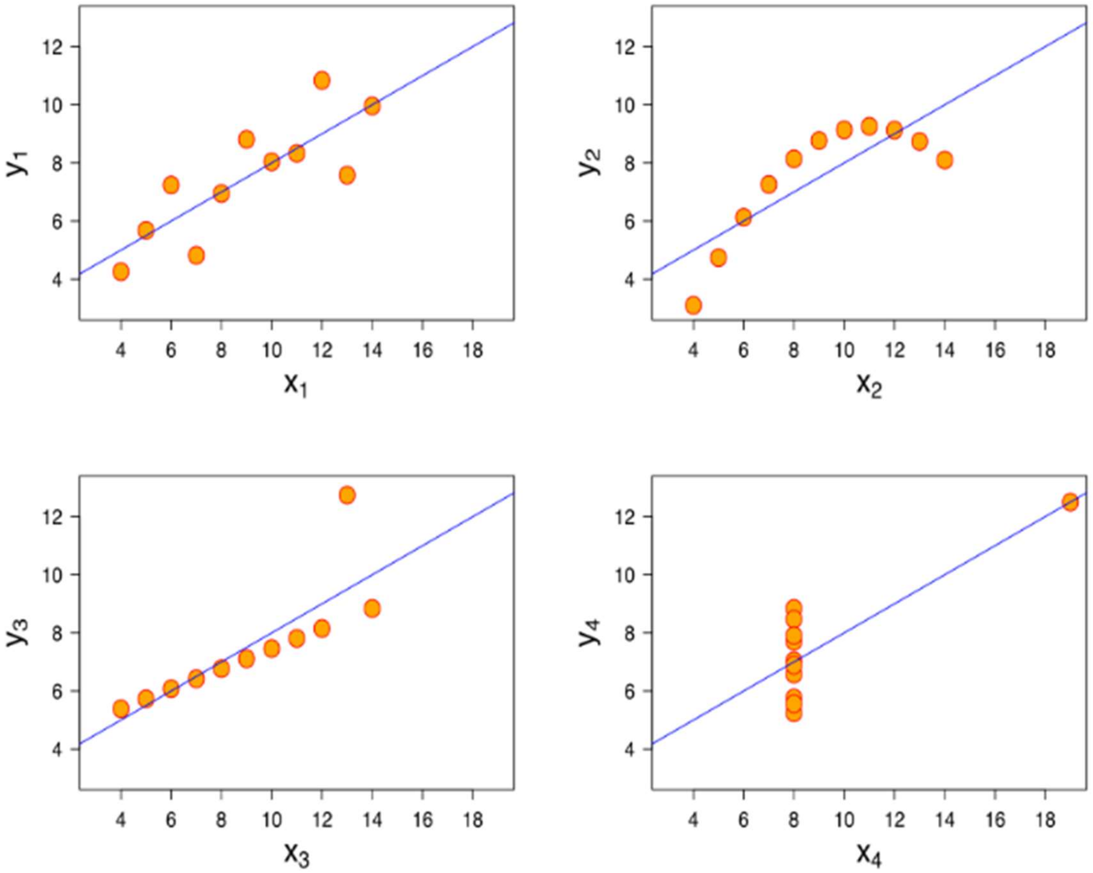
Contains a vertical line of points with one outlier
Shows how a single influential point can create misleading correlations
Demonstrates the concept of leverage in statistical analysis

Statistical Properties (Nearly Identical for All Four Sets)
All four datasets share these properties:

Mean of x = 9.0
Variance of x = 11.0
Mean of y = 7.5
Variance of y = 4.1
Correlation = 0.816
Linear regression line equation: y = 3 + 0.5x

Complete Dataset Values

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Graphical representation of Anscombe's quartet

Key Lessons

Visual Representation:

a) Essential for understanding true data relationships
b) Reveals patterns that statistics alone might miss
c) Helps identify outliers and unusual patterns

Statistical Limitations:

a) Summary statistics can mask important features
b) Correlation coefficients don't tell the whole story
c) Same statistics can represent very different relationships

Data Analysis Best Practices:

a) Always visualize data before analysis
b) Consider multiple analytical approaches
c) Be aware of outliers and their influence
d) Don't rely solely on summary statistics

Modern Relevance

Anscombe's Quartet remains a powerful teaching tool in modern data science:

a) Demonstrates the importance of exploratory data analysis
b) Shows why automated analysis needs human oversight
c) Illustrates the value of data visualization in the age of big data

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's correlation coefficient (r) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson and is sometimes referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or simply the correlation coefficient.

Mathematical Formula

The formula for Pearson's r is:

$$r = \Sigma(x - \mu x)(y - \mu y) / (n * \sigma x * \sigma y)$$

Where:

x, y are the variables
μx, μy are their means
σx, σy are their standard deviations
n is the number of pairs of values

Key Characteristics

Range:

a) Values always fall between -1 and +1
b) -1 indicates perfect negative correlation
c) +1 indicates perfect positive correlation
d) 0 indicates no linear correlation

Interpretation:

e) +0.7 to +1.0: Strong positive correlation
f) +0.3 to +0.7: Moderate positive correlation
g) -0.3 to +0.3: Weak or no correlation
h) -0.7 to -0.3: Moderate negative correlation
i) -1.0 to -0.7: Strong negative correlation

Properties:

a) Symmetrical: correlation of X with Y equals correlation of Y with X
b) Scale-invariant: not affected by changes in measurement units
c) No units of measurement
d) Independent of sample size

Assumptions:

a) Variables are continuous
b) Linear relationship between variables
c) Variables are normally distributed
d) Absence of significant outliers
e) Homoscedasticity (equal variances)

Limitations:

a) Only measures linear relationships
b) Sensitive to outliers
c) Does not imply causation
d) May miss non-linear relationships
e) Requires interval or ratio level data

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a preprocessing technique that adjusts the values of numeric features in a dataset to a similar range. It transforms the data to ensure all features contribute equally to the analysis and prevents features with larger magnitudes from dominating the model.

Why is Scaling Performed?

Algorithm Performance:

a) Improves convergence speed for gradient descent
b) Prevents features with larger values from dominating

Feature Comparison:

a) Makes features comparable regardless of original units
b) Enables fair contribution of all features

c) Prevents bias towards larger-scale features

Model Stability:

a) Reduces numerical instability during computation
b) Improves model convergence
c) Enhances prediction accuracy
d) Makes training more efficient

Types of Scaling

1. ***Normalization (Min-Max Scaling)***
Scales data to a fixed range, typically [0, 1] or [-1, 1]
Formula:
X_normalized = (X - X_min) / (X_max - X_min)

Characteristics:

a) Bounds values between 0 and 1
b) Preserves zero values
c) Preserves shape of original distribution
d) Handles outliers poorly

Use Cases:

---

a) Image processing
b) When data has a bounded range
c) When you need values between 0 and 1

2. Standardization (Z-score Scaling)
Transforms data to have mean=0 and standard deviation=1
Formula:
X_standardized = (X - μ) / σ

Where:

μ is the mean
σ is the standard deviation

Characteristics:

---

a) Centers data around 0
b) No bounded range
c) Better handling of outliers

d) Assumes normal distribution

Use Cases:

a) Linear models
b) When data approximates normal distribution
c) When outliers are present

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A VIF value becomes infinite when there is perfect multicollinearity in the data, meaning one predictor variable can be perfectly predicted from one or more other predictor variables. Here's a detailed explanation:

Mathematical Reason:

VIF = 1 / (1 - R^2)
When R^2 = 1 (perfect correlation), VIF = 1 / (1 - 1) = 1/0
Division by zero results in ***infinity***

Common Scenarios Leading to Infinite VIF:

1. Direct Linear Relationships
2. Dummy Variable Trap
3. Matrix Inversion Issues

Examples: Age, BirthYear, Current Year – Any two can predict third

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform.

The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line

Interpretations

a) Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
b) Y values < X values: If y-values quantiles are lower than x-values quantiles.
c) X values < Y values: If x-values quantiles are lower than y-values quantiles.
d) Different distributions – If all the data points are lying away from the straight line.

Advantages

a) Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be identified from the single plot.
b) The plot has a provision to mention the sample size as well