Assignment-based Subjective Questions

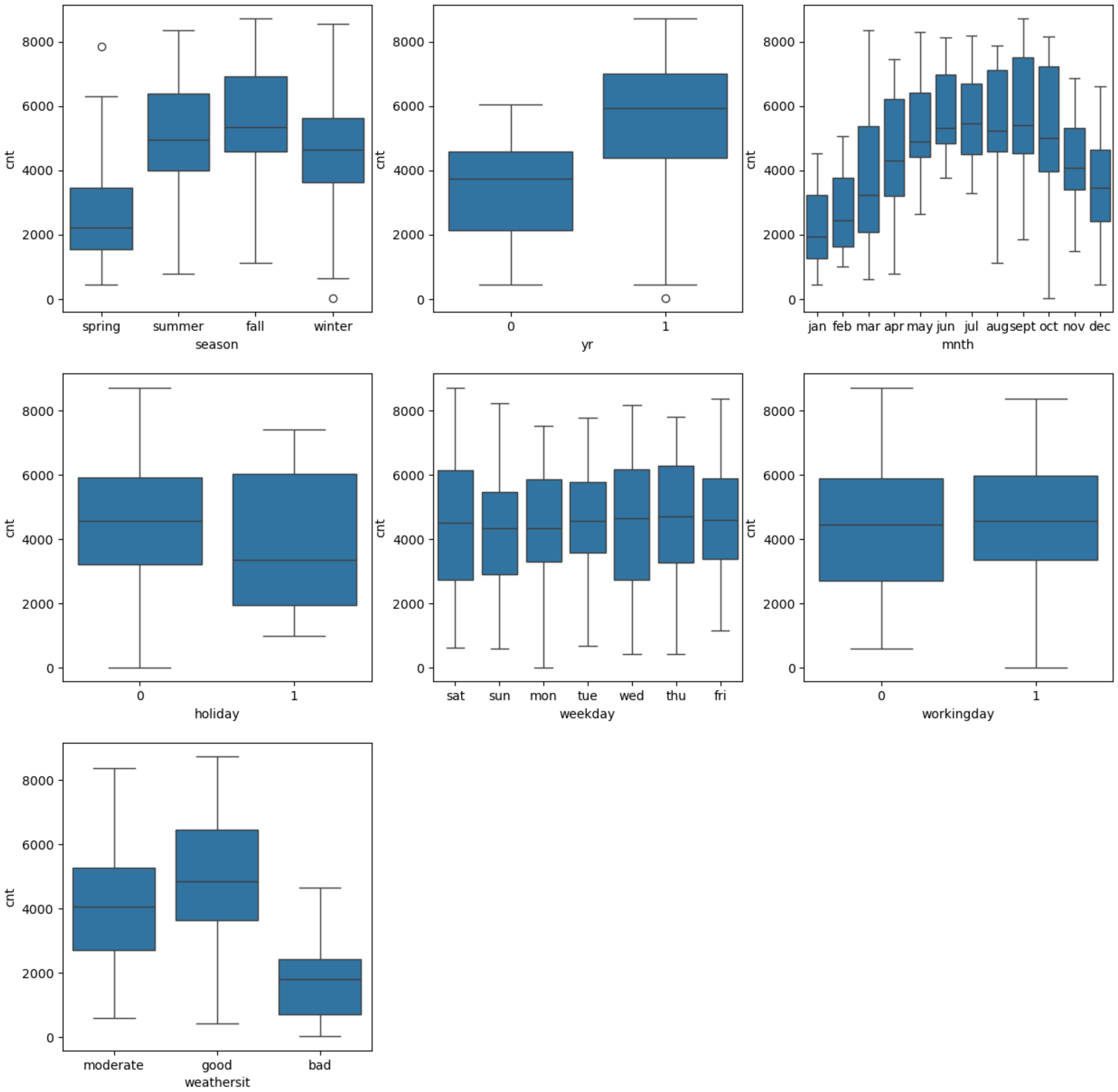
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Analysis of Categorical Variables and Their Effect on the Dependent Variable

I conducted an analysis of the categorical variables using boxplots and bar charts. The key insights derived from the visualizations are as follows:

- **Seasonal Trends**: The fall season recorded the highest number of bookings. Additionally, booking counts increased significantly from 2018 to 2019 across all seasons.
- **Monthly Patterns**: The highest number of bookings occurred between May and October. The trend shows a steady increase from the beginning of the year, peaking around mid-year, followed by a decline towards the year's end.
- **Weather Conditions**: Clear weather conditions were associated with a higher number of bookings, which aligns with expectations.
- **Day of the Week Impact**: Bookings were more frequent on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week.
- **Holiday Effect**: Bookings were lower on non-holiday days, which is reasonable as people may prefer to stay home and spend time with family on holidays.
- **Working vs. Non-Working Days**: There was no significant difference in bookings between working and non-working days.
- **Yearly Growth**: The number of bookings in 2019 was higher than in 2018, indicating positive business growth.

These insights suggest that seasonal, weather, and temporal factors significantly influence booking patterns, with a clear upward trend in demand over time.



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)

Using `drop_first=True` is important as it helps reduce the number of dummy variables created, thereby minimizing redundancy and **preventing multicollinearity** among dummy variables.

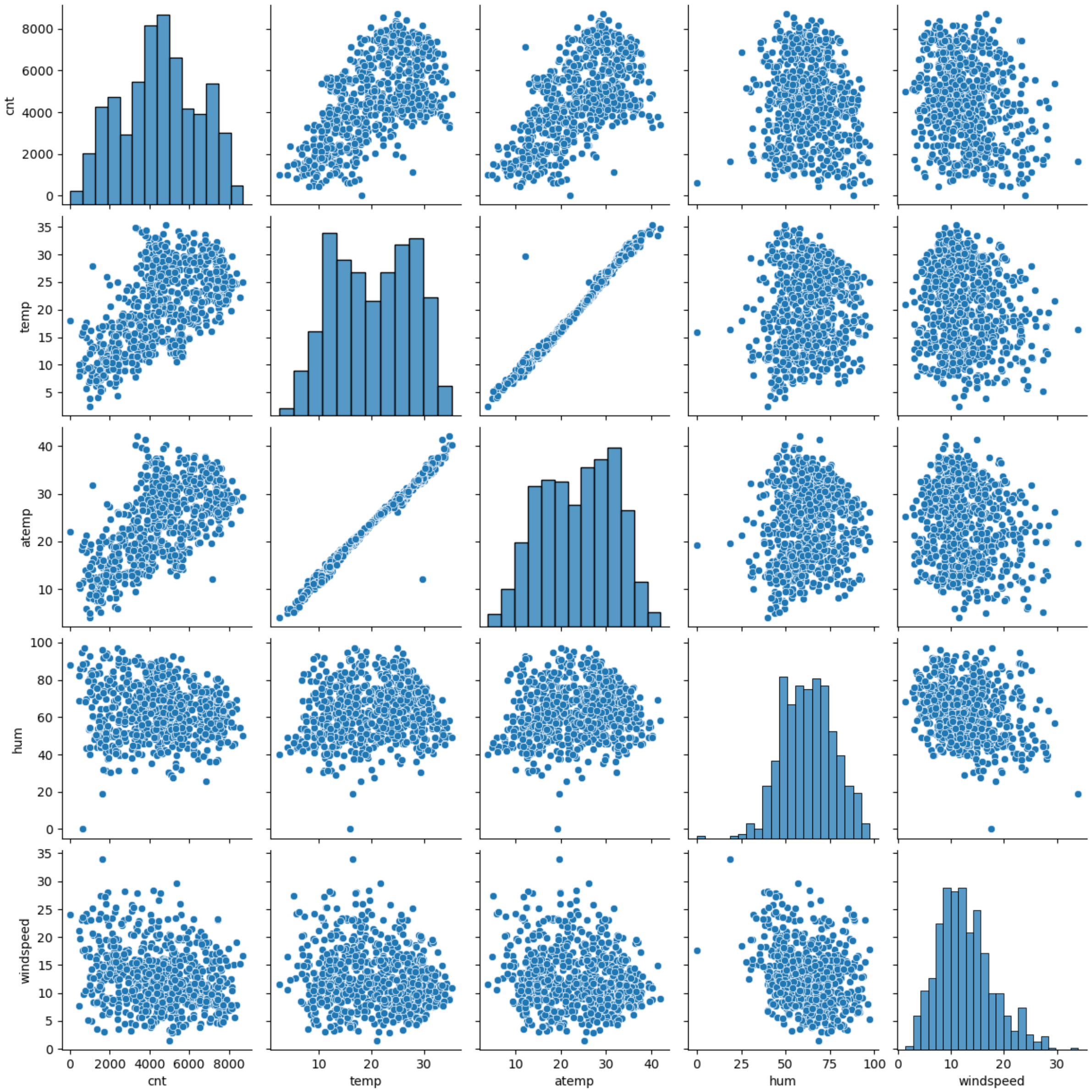
When set to `True`, it generates **k-1** dummy variables for a categorical feature with **k** unique values by removing the first category.

Example:

If a categorical column has three values: **A, B, and C**, creating dummy variables for all three would be unnecessary. If a data point is neither **A** nor **B**, it is implicitly **C**, making the third dummy variable redundant.

This approach ensures a more efficient representation of categorical variables, particularly in regression models.

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)



The 'temp' and 'atemp' variables have the highest correlation when compared to the rest with the target variable as 'cnt'.

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)

I have validated the assumptions of the Linear Regression model based on the following five key criteria:

- **Normality of Error Terms**: The residuals should follow a normal distribution.
- **Multicollinearity Check**: There should be minimal or no multicollinearity among independent variables.

- **Linearity Validation**: A linear relationship should be evident between the independent and dependent variables.
- **Homoscedasticity**: The residuals should exhibit constant variance, with no discernible pattern.
- **Independence of Residuals**: Residuals should be independent, with no signs of autocorrelation.

These validations ensure the reliability and accuracy of the regression model.

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)

The top three features that significantly influence shared bike demand are **temperature, year, and season**. These variables play a crucial role in explaining variations in demand patterns.

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 fundamental supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It establishes a linear relationship between the input variables (predictors) and the target variable by fitting a straight line to the data.

1. Mathematical Representation

For a simple linear regression model with one independent variable (**X**) and one dependent variable (**Y**), the equation is:

Y=β0+β1X+εY = \beta_0 + \beta_1X + \varepsilonY=β0+β1X+ε

Where:

- **Y** = Dependent variable (target)
- **X** = Independent variable (predictor)
- **β_0** = Intercept (value of Y when X = 0)
- **β_1** = Coefficient (slope) that determines the relationship between X and Y
- **ε** = Error term (accounts for the variability not explained by the model)

For multiple linear regression (where there are multiple predictors **X_1, X_2, X_3, ..., Xn**), the equation generalizes to:

Y=β0+β1X1+β2X2+...+βnXn+εY = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilonY=β0+β1X1+β2X2+...+βnXn+ε

2. Model Training Using Ordinary Least Squares (OLS)

Linear regression is trained using the **Ordinary Least Squares (OLS) method**, which minimizes the sum of squared errors (SSE) between actual and predicted values. The cost function is:

SSE=∑(Yi−Y^i)2SSE = \sum (Y_i - \hat{Y}_i)^2SSE=∑(Yi−Y^i)2

Where **Y_i** is the actual value and **$\hat{Y}_i$** is the predicted value. The model finds the best-fit line by optimizing the regression coefficients (β-values) to minimize SSE.

3. Assumptions of Linear Regression

For accurate predictions, linear regression relies on the following assumptions:

1. **Linearity** – The relationship between the independent and dependent variables should be linear.
2. **Independence of Errors** – Residuals (errors) should be independent and not correlated.
3. **Homoscedasticity** – The variance of residuals should be constant across all levels of the independent variable.
4. **Normality of Errors** – The residuals should follow a normal distribution.
5. **No Multicollinearity** – Independent variables should not be highly correlated with each other.

4. Evaluation Metrics

To assess model performance, common evaluation metrics include:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **R^2 Score** (coefficient of determination, indicating model goodness-of-fit)

5. Applications of Linear Regression

- Forecasting sales, stock prices, and demand patterns
- Analyzing trends in economics and finance
- Predicting real estate prices based on features like location and size
- Understanding the impact of independent variables in scientific studies

Conclusion

Linear regression is a simple yet powerful technique for predictive modeling and statistical analysis. While it provides interpretability and efficiency, its effectiveness depends on meeting the key assumptions. If assumptions are violated, advanced techniques such as **polynomial regression, ridge regression, or decision trees** may be used for better accuracy.

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 created by the statistician **Francis Anscombe** in 1973 to demonstrate the importance of data visualization in statistical analysis. These datasets highlight how different data distributions can have identical summary statistics but vastly different underlying structures.

1. Importance of Anscombe's Quartet

Traditional statistical measures like **mean, variance, correlation, and linear regression coefficients** can sometimes be misleading when describing data. The quartet illustrates that datasets with the same statistical properties can have completely different distributions when visualized.

2. The Four Datasets

Each dataset in Anscombe's Quartet consists of **11 (x, y) pairs**, and all four share nearly identical statistical properties:

- **Mean of X and Y**
- **Variance of X and Y**
- **Correlation coefficient (~0.816)**
- **Linear regression line (Y = 3 + 0.5X)**

Despite these similarities, visualizing the datasets reveals stark differences in patterns:

3. Key Observations

Dataset 1: A Typical Linear Relationship

- The points follow a **linear trend**, making the regression line an appropriate fit.
- This dataset behaves as expected in a simple linear regression scenario.

Dataset 2: Nonlinear Relationship

- The data follows a **curved** (quadratic) pattern rather than a straight line.
- Linear regression fails to model this correctly, despite having the same statistical properties as Dataset 1.

Dataset 3: Influence of an Outlier on Regression

- The points mostly align in a straight line **except for one extreme outlier**.
- This single outlier **distorts the regression line**, making it unreliable.

Dataset 4: High Leverage Point Affecting Correlation

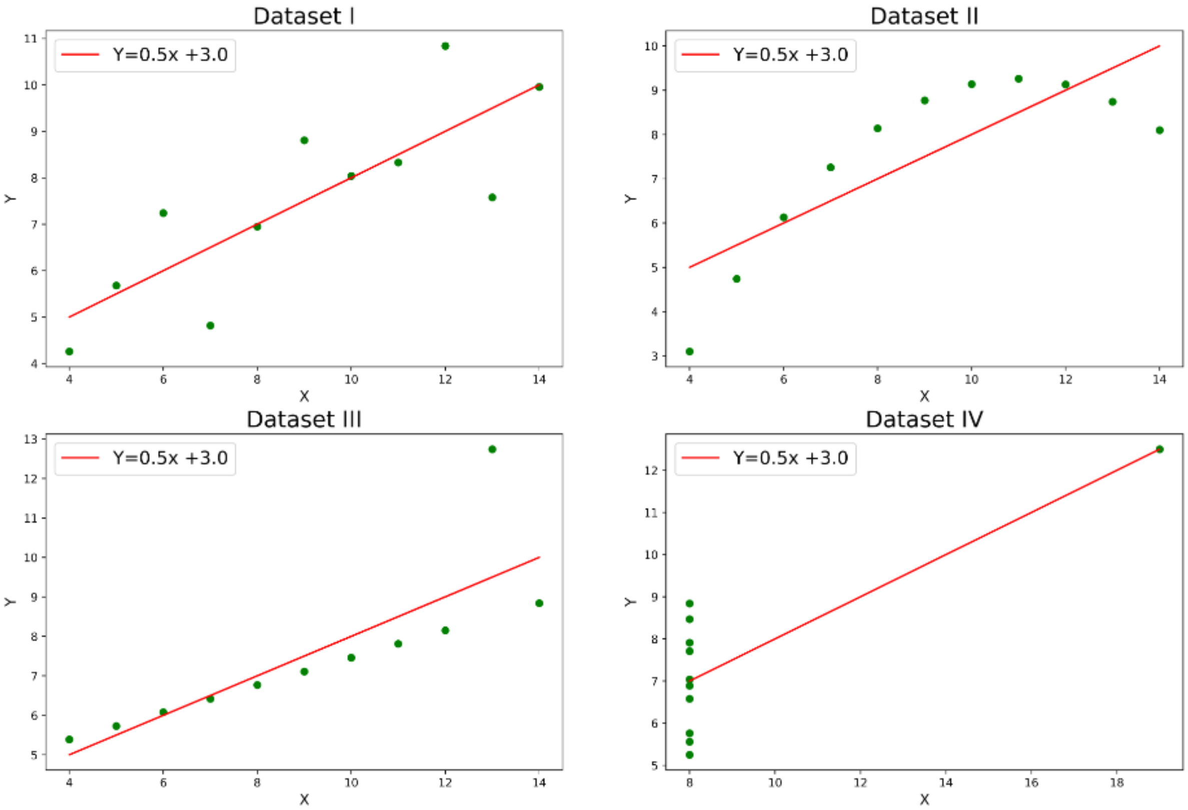
- Most points remain constant except for **one extreme X-value** (a high-leverage point).
- The regression line is heavily influenced by this one point, despite the majority of the data showing

no clear trend.

4. Lessons from Anscombe’s Quartet

- **Always visualize your data** before drawing conclusions.
- **Summary statistics alone are not sufficient** to understand the true nature of a dataset.
- **Outliers and non-linearity can distort regression models**, making it necessary to check residuals and fit more appropriate models.
- **Graphical analysis (scatter plots) should accompany numerical analysis** to gain accurate insights.

5. Visualization of Anscombe’s Quartet



Conclusion

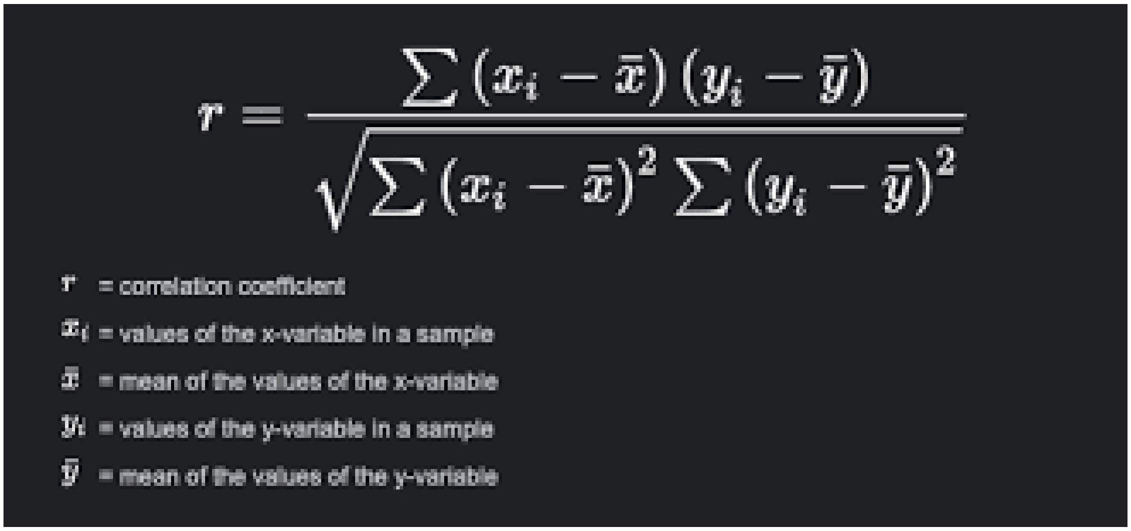
Anscombe’s Quartet is a powerful demonstration of the **limitations of statistical summaries** and the necessity of **visualizing data** before making decisions. It underscores the importance of **data exploration, outlier detection, and choosing the right model** for analysis.

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 R, also known as the **Pearson correlation coefficient (r)**, is a statistical measure that quantifies the **strength and direction of a linear relationship** between two continuous variables. It ranges from **-1 to +1**, where:

- **+1** → Perfect **positive** correlation (as one variable increases, the other also increases).
- **0** → No correlation (no linear relationship between variables).
- **-1** → Perfect **negative** correlation (as one variable increases, the other decreases).

Formula for Pearson's R



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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)

What is Scaling?

Scaling is the process of transforming numerical data into a specific range or distribution to ensure that all features contribute equally to a model. It is essential when working with machine learning algorithms that are sensitive to varying feature magnitudes, such as **linear regression, k-nearest neighbors (KNN), support vector machines (SVM), and gradient descent-based algorithms (e.g., logistic regression, neural networks, etc.)**.

Why is Scaling Performed?

1. **Ensures Equal Contribution of Features** – Prevents variables with large magnitudes from dominating the model.
2. **Improves Model Performance** – Helps gradient-based optimization algorithms converge faster.
3. **Prevents Numerical Instability** – Reduces computation errors caused by large-scale

differences.

4. **Required for Distance-Based Models** – Algorithms like **KNN and K-means clustering** rely on Euclidean distance, which can be biased toward larger values if not scaled properly.

Difference Between Normalized Scaling and Standardized Scaling

Feature	Normalized Scaling	Standardized Scaling
Definition	Rescales values to a fixed range, typically **[0,1]** or **[-1,1]**	Transforms data to have **zero mean** and **unit variance**
Formula	X′=X−XminXmax−XminX' = \frac{X - X_{min}}{X_{max} - X_{min}}X′=Xmax−XminX−Xmin	X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ
Effect on Data	Values are **bounded** within a specific range	Data is centered around **0**, with a standard deviation of **1**
Used When?	Features have **different scales** and are **not normally distributed**	Data follows a **Gaussian (normal) distribution** or when outliers exist
Example Algorithms	**Neural networks, KNN, K-means, Min-Max scaling applications**	**Linear regression, logistic regression, PCA, SVM**
Sensitive to Outliers?	**Yes** (outliers can significantly affect min-max values)	**Less sensitive** (since it standardizes based on mean and standard deviation)

Conclusion

- **Normalization (Min-Max Scaling)** is best when you need values in a fixed range (e.g., [0,1]).
- **Standardization (Z-score Scaling)** is preferred for normally distributed data and when handling outliers.
- **Choice of scaling depends on the dataset and the machine learning algorithm used.**

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)

Variance Inflation Factor (VIF) measures multicollinearity among independent variables in a regression model. A **VIF value becomes infinite** (or extremely high) when **perfect multicollinearity** exists, meaning one predictor is a **perfect linear combination** of one or more other predictors.

Reasons for Infinite VIF

1. **Perfectly Correlated Variables** – If two or more independent variables have a correlation of **1 or -1**, VIF becomes infinite.
2. **Duplicate Features** – Including the same variable twice (e.g., `height` and `height_in_cm`) leads to perfect correlation.
3. **One Feature is a Linear Combination of Others** – If X3=X1+X2X_3 = X_1 + X_2X3=X1+X2, then X3X_3X3 is perfectly predicted by X1X_1X1 and X2X_2X2, causing infinite VIF.
4. **Dummy Variable Trap** – If all dummy variables of a categorical feature are included without using `drop_first=True`, perfect multicollinearity occurs.

Infinite VIF occurs due to perfect multicollinearity, making regression coefficients unreliable. Identifying and removing redundant variables can resolve this issue and improve model stability

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)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the quantiles of a dataset against the quantiles of a reference distribution (typically normal). If the data is normally distributed, the points will align along a straight line; deviations from the line suggest non-normality.

Use and Importance in Linear Regression:

In linear regression, one key assumption is that the residuals are normally distributed. A Q-Q plot of residuals helps assess this assumption:

- Normal residuals: Points align with the 45-degree line, indicating valid regression results.
- Non-normal residuals: Deviations suggest problems with the model, possibly requiring transformations or a different modeling approach.

Key Points:

- Assumption Checking: Ensures normality of residuals, crucial for valid inference.
- Outlier Detection: Outliers appear as points far from the line.
- Improves Model Validity: Non-normal residuals may signal model inadequacy.

Conclusion:

A Q-Q plot is vital for checking the normality assumption in linear regression, ensuring reliable predictions and statistical inferences.