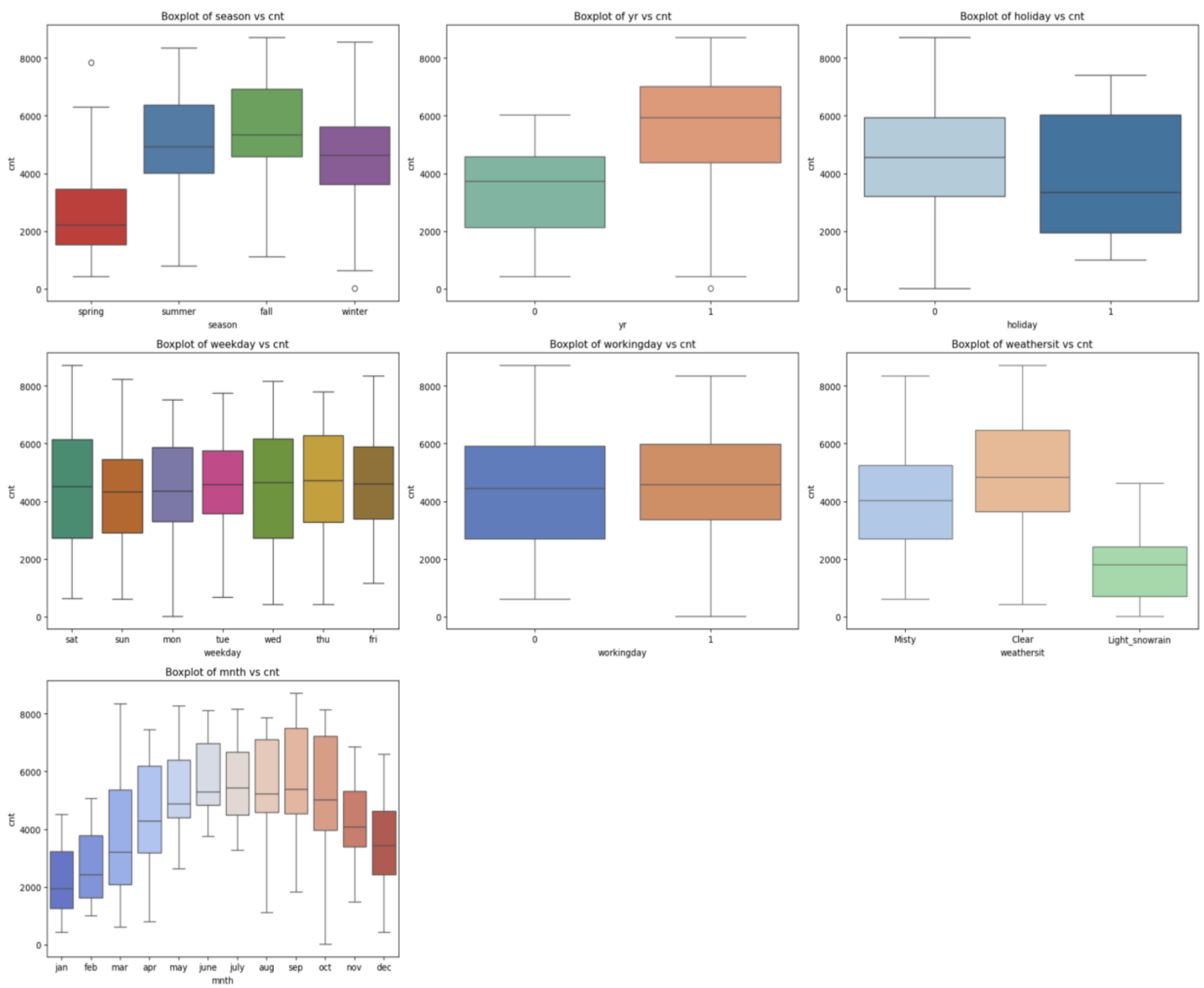
Assignment-based Subjective Questions

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:

The categorical variable in the dataset were season, holiday, weekday, working day, yr and weather sit and month.



- Season: The demand was highest during the Fall season, followed by the Summer season.
- Working day: The counts are similar on both working days and non-working days.
- Holiday: The user counts were lower on holidays compared to regular days.
- Weather sit: Clear weather had the highest median user counts and the widest range.
- Year: The user counts were higher in 2019 compared to 2018.
- Month: Bookings were highest between May and September, with a pattern of increasing from the start to the middle of the month, then declining towards the year's end.
- Weekday: Thursday, Friday, and Saturday had more bookings compared to other days of the week.

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:
Using drop_first=True during dummy variable creation in pandas helps avoid multicollinearity in regression models, particularly when using techniques like Ordinary Least Squares (OLS). Here's why it's important:

1. Dummy Variable Trap:

- When creating dummy variables for a categorical feature with k categories, you end up with k binary columns (one for each category).
- In such cases, these dummy variables are **linearly dependent** because the presence of one category can always be inferred from the others. This is called the **dummy variable trap** and leads to multicollinearity in the model.

Example: For a feature Color with values ['Red', 'Green', 'Blue'], the dummy variables might look like this:

Color_Red	Color_Green	Color_Blue
1	0	0
0	1	0
0	0	1

Here, if Color_Red and Color_Green are known, you can infer Color_Blue. This redundancy causes multicollinearity.

2. Avoiding Redundancy

- Setting drop_first=True removes one dummy variable column (e.g., Color_Blue), leaving only k-1 columns.
- The removed column acts as a **reference category**, and the remaining columns represent deviations from this reference.

Color_Red	Color_Green
1	0
0	1
0	0

In this case, the absence of both Color_Red and Color_Green implies Color_Blue.

3. Improved Model Interpretability

- By dropping the first category, the coefficients of the remaining dummy variables are interpreted relative to the reference category, making the model easier to understand.

When to Use drop_first=False?

- When the downstream model or algorithm can handle multicollinearity (e.g., tree-based models), you might not need to drop the first category.
- For descriptive purposes or certain analyses where all categories are needed explicitly.

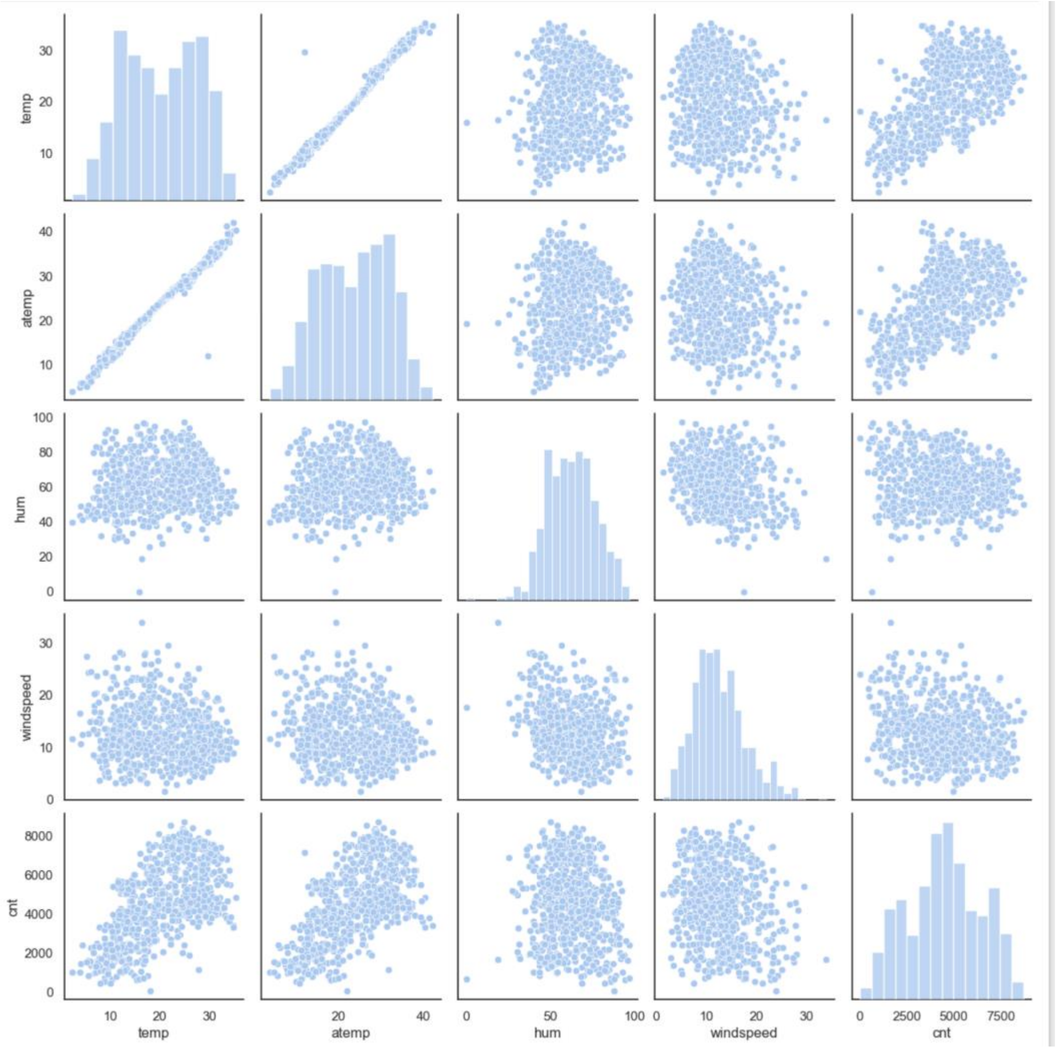
In summary, drop_first=True is important to prevent multicollinearity issues and improve the interpretability of linear models.

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:

Using the below pairplot it can be seen that , “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable cnt. The correlation between temp and atemp is almost 1.

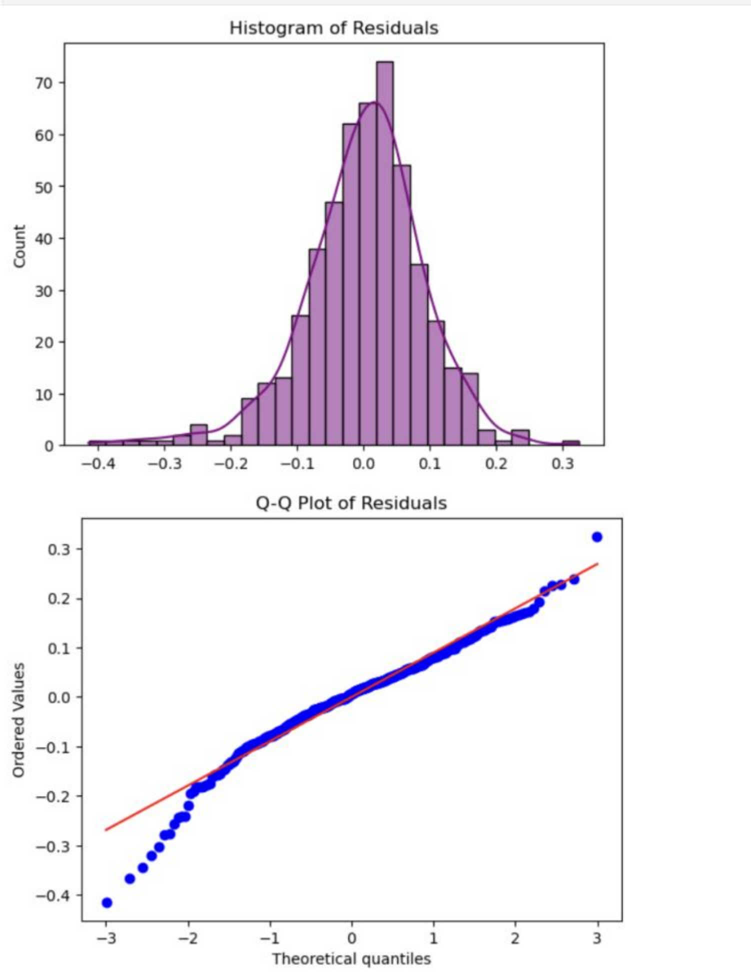


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

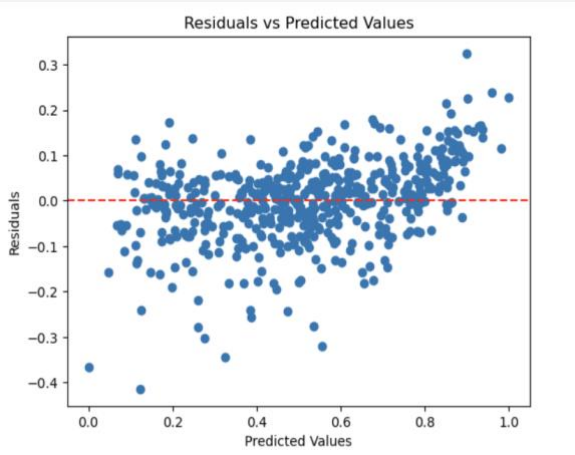
Answer:

1. Normality of Residuals

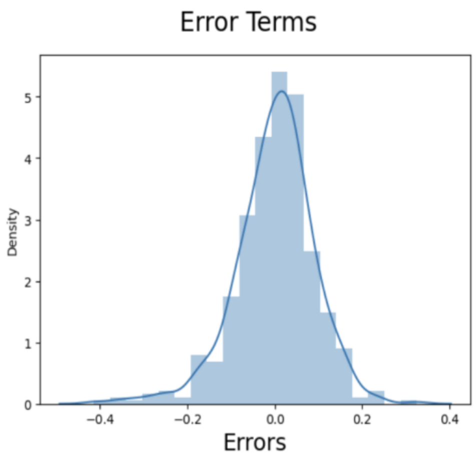


---

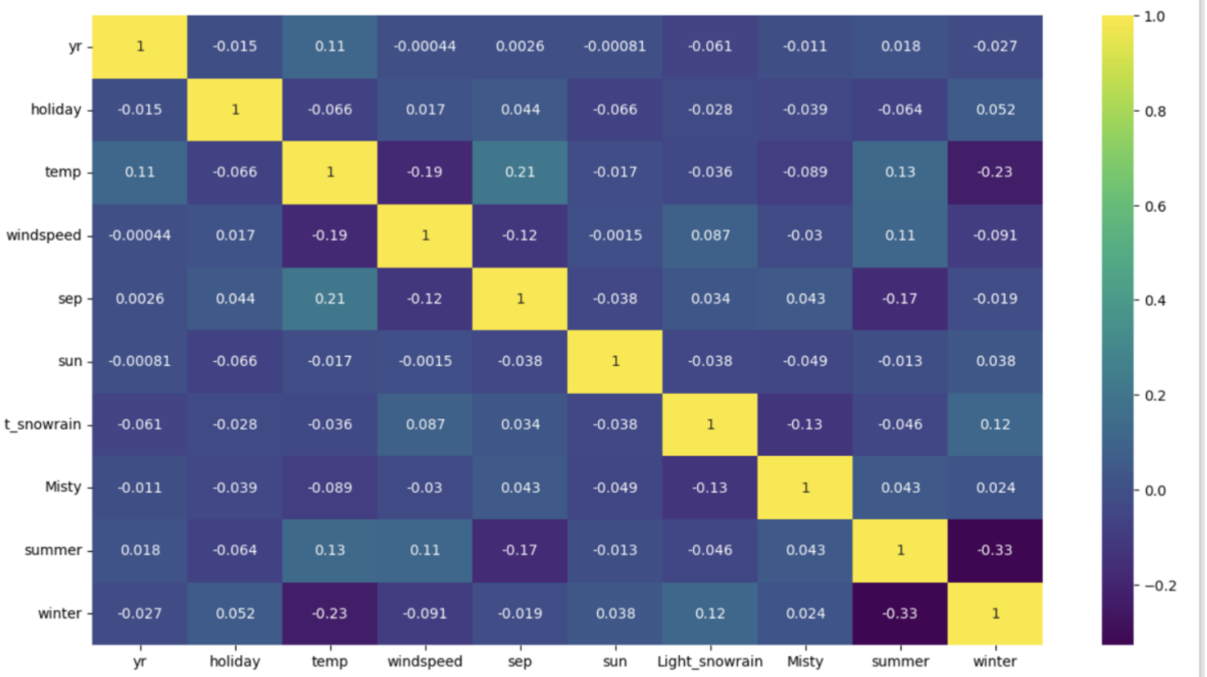
2. Linearity: Check if the relationship between independent variables and the dependent variable is linear.



3. Error term:



4. No Multicollinearity



	yr	holiday	temp	windspeed	sep	sun	Light_snowrain	Misty	summer	winter
yr	1	-0.015	0.11	-0.00044	0.0026	-0.00081	-0.061	-0.011	0.018	-0.027
holiday	-0.015	1	-0.066	0.017	0.044	-0.066	-0.028	-0.039	-0.064	0.052
temp	0.11	-0.066	1	-0.19	0.21	-0.017	-0.036	-0.089	0.13	-0.23
windspeed	-0.00044	0.017	-0.19	1	-0.12	-0.0015	0.087	-0.03	0.11	-0.091
sep	0.0026	0.044	0.21	-0.12	1	-0.038	0.034	0.043	-0.17	-0.019
sun	-0.00081	-0.066	-0.017	-0.0015	-0.038	1	-0.038	-0.049	-0.013	0.038
t_snowrain	-0.061	-0.028	-0.036	0.087	0.034	-0.038	1	-0.13	-0.046	0.12
Misty	-0.011	-0.039	-0.089	-0.03	0.043	-0.049	-0.13	1	0.043	0.024
summer	0.018	-0.064	0.13	0.11	-0.17	-0.013	-0.046	0.043	1	-0.33
winter	-0.027	0.052	-0.23	-0.091	-0.019	0.038	0.12	0.024	-0.33	1

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:

1. **temp (Coefficient: 0.5471):** Temperature has the highest positive coefficient, indicating that higher temperatures strongly correlate with increased bike demand.
2. **yr (Coefficient: 0.2328):** The year variable suggests a substantial positive trend in bike demand over time, with 2019 seeing higher usage than 2018.
3. **Light_snowrain (Coefficient: -0.2883):** This feature has the largest negative impact. Inclement weather (light snow or rain) significantly reduces bike demand, as adverse conditions discourage users.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)

Answer:
1. Linear Regression Algorithm:

Linear Regression is a statistical model that helps in understanding the relationship between a dependent variable with given independent variable through a straight line. This means that when there is a change in the value of one or more independent variables, whether increase or decrease, the value of the dependent variable also changes accordingly.

- Dependent variable, in any given task, is the variable we aim to predict. In linear regression, it is denoted by 'y'.
- Independent variables are the variables that affect the dependent variable and are denoted as X1, X2, ..., Xn.
- The algorithm estimates coefficients (b1, b2, ..., bn) for each independent variable and an intercept (b0).

Therefore, the linear regression equation formed by these terminologies is:

y = b0 + b1X1 + ……… + bnXn

The primary objective in linear regression is to find the values of b0, b1, b2,……., bn that minimize the sum of squared differences between the predicted and actual values, and this is known as Ordinary Least Square(OLS) regression.

2. Types of Linear Regression

There are two types of Linear Regression:

2.1 **Simple Linear Regression** – it explains the relationship between a dependent variable and one independent variable using a straight line.

$$\mathbf{y = b_o + b_1.X}$$

2.2 **Multiple Linear Regression** – it is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).

y = b0 + b1X1 + ……… + bnXn

3. Steps in Linear Regression

3.1 Hypothesis

Linear regression assumes a linear relationship between predictors and the target variable.

3.2 Cost Function

The cost function measures how well the model's predictions fit the data. The most common cost function is the **Mean Squared Error (MSE)**:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.3 Optimization

The algorithm minimizes the cost function by finding the best-fitting line using techniques like **Ordinary Least Squares (OLS)**.

OLS minimizes the sum of squared residuals:

$$\min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.4 Gradient Descent (Alternative Method)

For large datasets, gradient descent is often used. It iteratively adjusts the coefficients (β\betaβ) by following the gradient of the cost function:

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} MSE$$

4. Assumptions of Linear Regression

To ensure reliable results, linear regression makes several key assumptions:

1. **Linearity**: The relationship between predictors and the target is linear.
2. **Independence**: Residuals are independent of each other.
3. **Homoscedasticity**: Residuals have constant variance across all levels of predicted values.
4. **Normality of Residuals**: Residuals are normally distributed.
5. **No Multicollinearity**: Predictors are not highly correlated with each other.

5. Evaluation Metrics

After fitting the model, its performance is evaluated using metrics like:

1. **R-squared (R^2)**: Proportion of variance explained by the model.

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{residual}}}{\mathrm{SS}_{\mathrm{total}}}$$

2. **Adjusted R^2**: Adjusted for the number of predictors in the model.
3. **Root Mean Squared Error (RMSE)**: Square root of MSE.
4. **Mean Absolute Error (MAE)**: Average of absolute residuals.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer:
Anscombe's Quartet, created by the statistician Francis Anscombe in 1973, was designed to highlight the importance of visualizing data before analyzing it or building a model.

- The primary purpose of Anscombe's Quartet is to emphasize that relying solely on summary statistics can be misleading, as it may overlook underlying patterns or structures in the data.
- By visualizing the data, one can gain a deeper understanding of its characteristics and complexity.
- In essence, Anscombe's Quartet illustrates that simple numerical summaries are insufficient for capturing the full nature of the data.
- It serves as a reminder that plotting data is crucial for accurate analysis and helps avoid incorrect assumptions that could arise from focusing only on summary statistics.
- The quartet consists of four datasets that have nearly identical statistical properties, including the same mean, variance, correlation, and linear regression line for both x and y. Despite these similarities, when plotted, the datasets look vastly different from each other.

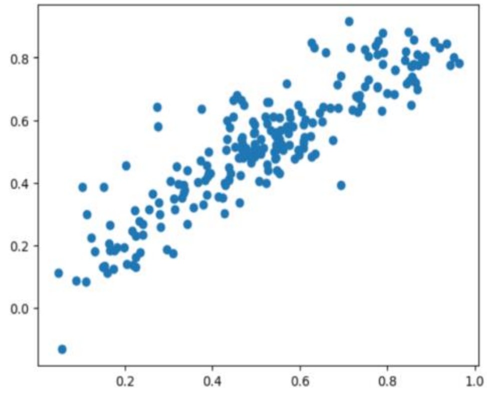
Anscombe's Quartet Four Datasets

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

- The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.
- Pearson coefficients (r) range from +1 to -1:
  - r = 1 representing a positive correlation
  - r = -1 representing a negative correlation
  - r = 0 representing no relationship.

- The Pearson coefficient shows correlation, not causation.
- The sign of (r) indicates the direction of the relationship.



This scatter plot shows Positive Correlation.

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:

What is Scaling?

Scaling refers to the process of transforming the features of your dataset to a specific range or distribution, making them comparable or compatible for use in machine learning algorithms. The purpose of scaling is to adjust the scale of the data to improve the performance and convergence of certain algorithms. Some machine learning models, particularly those based on distance metrics or gradient-based optimization (e.g., k-nearest neighbors, support vector machines, and neural networks), perform better when the input features are on a similar scale.

Why is Scaling Performed?

Scaling is important for several reasons:

1. **Ensure Equal Importance**: Features in a dataset may have different units or magnitudes (e.g., height in centimeters and weight in kilograms). Without scaling, algorithms may give more importance to features with larger magnitudes, leading to biased results.
2. **Improve Model Convergence**: Gradient-based optimization algorithms (like those used in logistic regression, neural networks, or gradient boosting) converge faster when the features are scaled. Features with larger ranges can dominate the gradient and slow down the convergence process.
3. **Improve Algorithm Performance**: Algorithms like k-nearest neighbors (KNN) or support vector machines (SVM), which are distance-based models, are sensitive to the scale of features. Large differences in scale can affect the performance and the results of these algorithms.

4. **Maintain Numerical Stability**: Features with vastly different scales may cause numerical issues during computations, leading to instability, especially in algorithms like linear regression and principal component analysis (PCA).

Types of Scaling

There are two main types of scaling techniques: **Normalization** and **Standardization**. Both methods adjust the scale of the data but in different ways.

1. Normalization (Min-Max Scaling)

Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. This is achieved by subtracting the minimum value of the feature and then dividing by the range (maximum - minimum):

$$x' = \frac{x - \mu}{\max(x) - \min(x)}$$

2. Standardization (Z-Score Scaling)

Standardization (or Z-score normalization) transforms the data such that it has a **mean of 0** and a **standard deviation of 1**. This is done by subtracting the mean of the feature and then dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

Differences Between Normalization and Standardization

Normalized	**Standardized**
It uses Minimum and Maximum of features for scaling.	It uses Mean and Standard Deviation for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values are between [0,1] or [-1,1].	It is not bound to a certain range.
It can be affected by outliers	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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:

The **Variance Inflation Factor (VIF)** is a diagnostic tool used to detect **multicollinearity** in a regression model. High VIF values indicate strong correlations among the independent variables, which complicates the interpretation of each variable’s unique contribution to the dependent variable. Although VIF is useful for identifying multicollinearity, it can encounter issues and potentially become infinite in certain situations.

An **infinite VIF** arises when there is **perfect multicollinearity** among the independent variables. This occurs when one or more variables are exactly predictable as a linear combination of the other variables in the model. In such cases, the regression model faces a situation where the predictors are perfectly correlated, and the matrix inversion required for calculating VIF becomes impossible, resulting in infinite values.

When a variable is a linear combination of others, it introduces **perfect multicollinearity**, which causes the VIF to approach infinity. This makes it impossible to estimate the regression coefficients reliably because the model cannot differentiate the individual effects of correlated variables.

Infinite VIF values lead to several issues:

- **Unreliable coefficient estimates**: The coefficients become unstable and difficult to interpret.
- **Inflated standard errors**: The standard errors of the regression coefficients grow larger, making it challenging to draw valid statistical conclusions and leading to weak t-tests and confidence intervals.

To resolve this issue, you should identify and remove one of the perfectly correlated variables, as retaining both will cause perfect multicollinearity. By doing so, you can restore the reliability of the model's coefficient estimates and improve the model’s interpretability.

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:

A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to assess the **normality** of a dataset by comparing the quantiles of the observed data to the quantiles of a theoretical normal distribution. In a Q-Q plot, if the data points lie approximately along a straight line, it indicates that the data is normally distributed.

Importance of a Q-Q Plot in Linear Regression:

1. **Assessing Normality of Residuals**: In linear regression, the Q-Q plot helps to visually inspect if the **residuals** (the differences between observed and predicted values) follow a normal distribution. If the points deviate from a straight line, it suggests that the residuals are not normally distributed.
2. **Impact on Statistical Tests**: Non-normality in residuals can distort the accuracy and reliability of the statistical tests (e.g., t-tests and F-tests) that are commonly used in linear regression. These tests assume normality for valid inference.
3. **Identifying Outliers**: Outliers or extreme values can affect the normality of residuals. A Q-Q plot can help identify whether these outliers are consistent with a normal distribution or if they indicate a problem in the model.
4. **Assumption Check**: Linear regression assumes that residuals are normally distributed, particularly for the validity of hypothesis testing and confidence intervals. A Q-Q plot provides a way to assess whether this assumption holds true.
5. **Model Adjustments**: By analysing the Q-Q plot, analysts can detect violations of normality, such as skewness or heavy tails. If such violations are identified, adjustments like data transformations or alternative modelling techniques may be necessary.
6. **Diagnostic Tool**: The Q-Q plot is a key diagnostic tool for ensuring the validity of the regression model, helping identify areas where the model might need refinement or improvement.