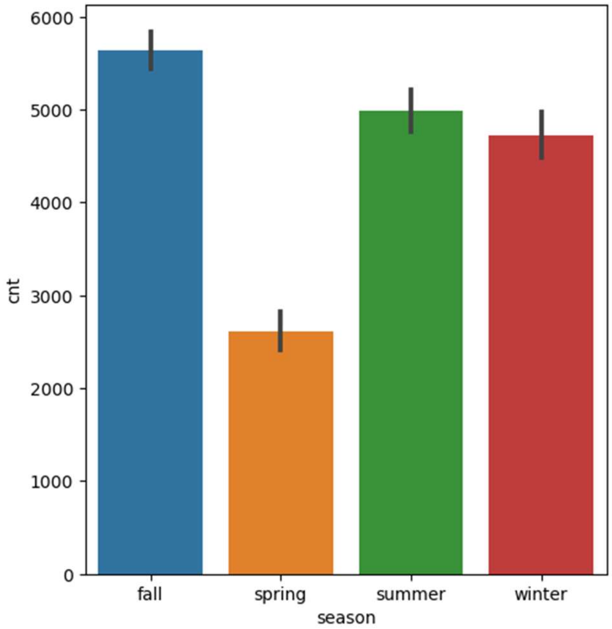
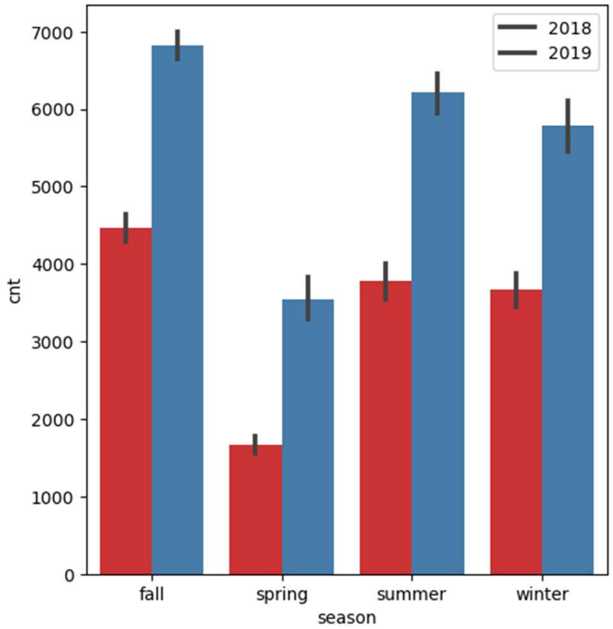
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: Based on the analysis of the categorical variables from the dataset, we can derive the inference as follows:

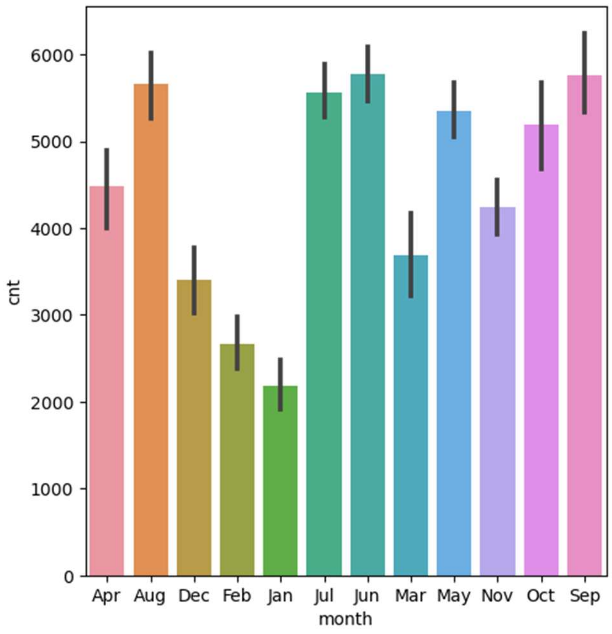
- Season fall has highest demand for rental bikes.
- Demand has increased from 2018 to 2019.
- September month has highest demand. After September, demand is decreasing.
- When there is a holiday, demand has decreased.
- Weekday is not giving clear picture about demand.
- The clear weathersit has highest demand.

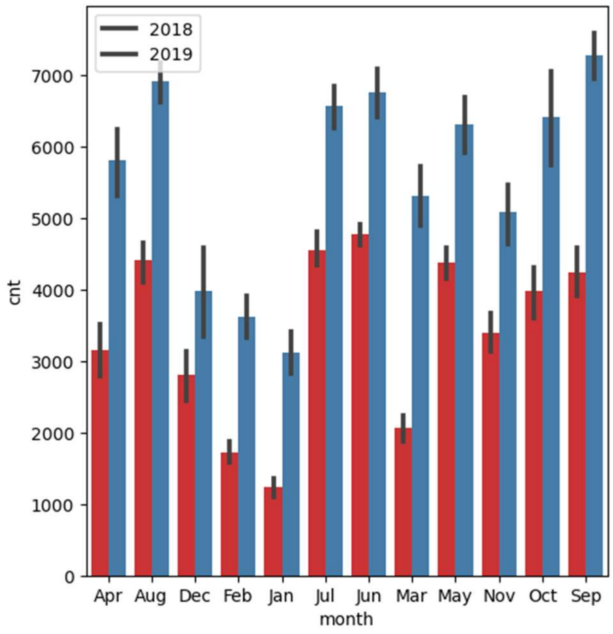




Useful Insights:

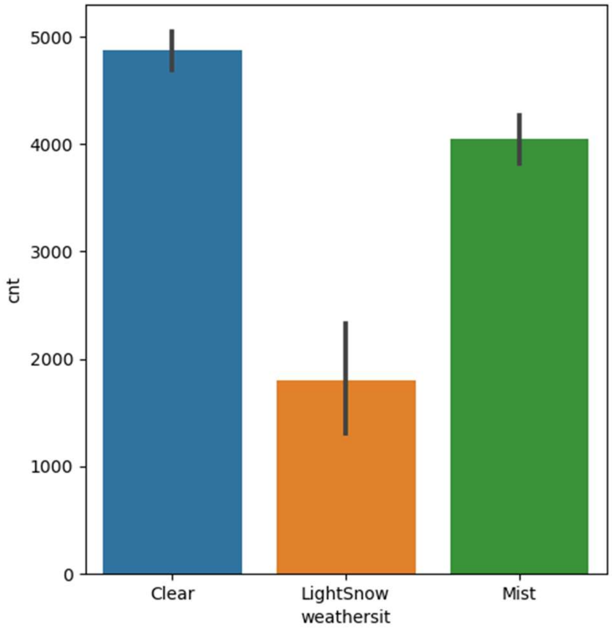
- fall season has highest bookings compare to other seasons
- Booking count has increased from 2018 to 2019

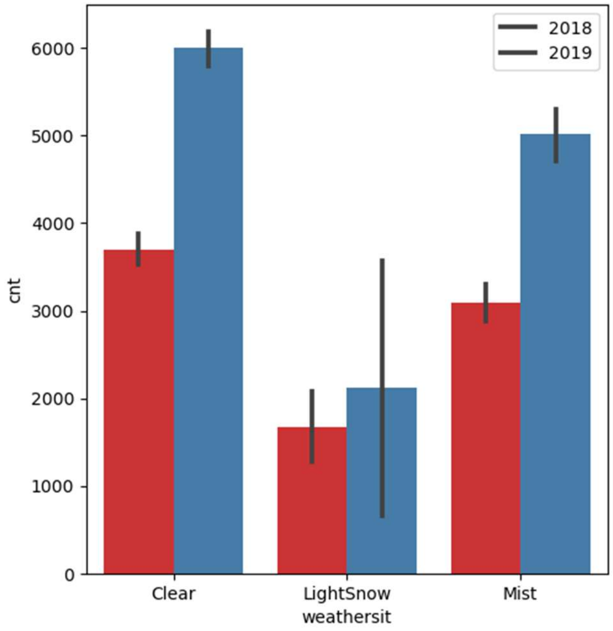




Useful Insights:

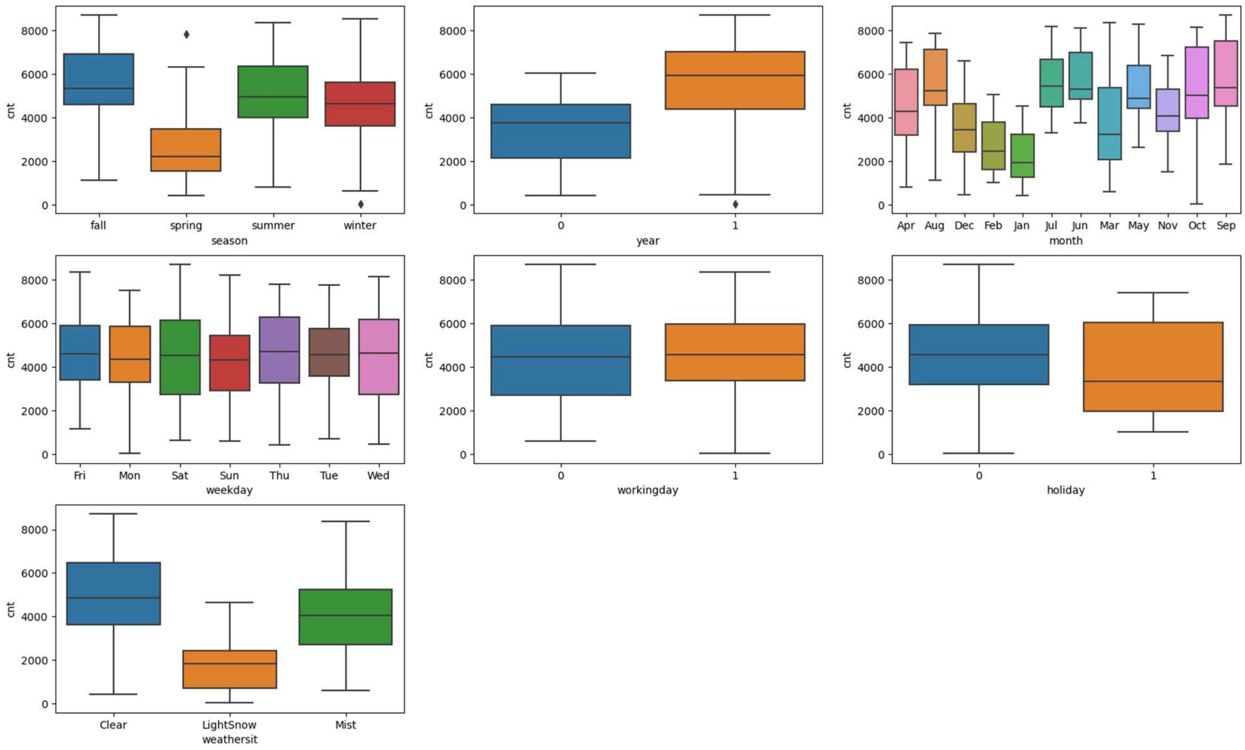
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct
- Number of bookings for each month seems to have increased from 2018 to 2019.





Useful Insights:

- Clear weather has attracted more bookings
- Booking count has increased from 2018 to 2019



2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

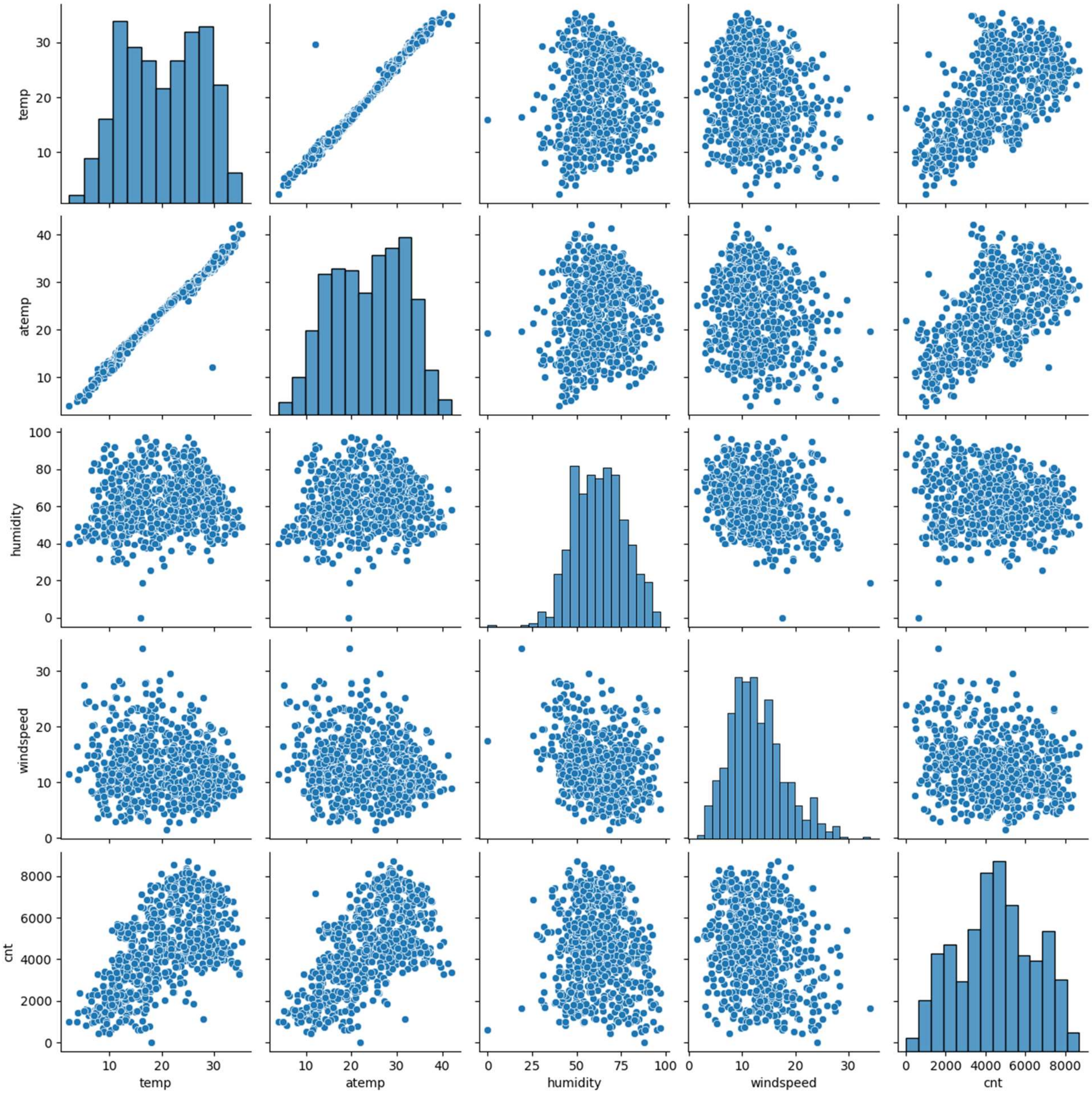
Ans: Using drop_first=True during dummy variable creation is important for avoiding multicollinearity and ensuring that the resulting dummy variables are interpretable and useful in regression models.

When creating dummy variables for categorical features with k categories, k dummy variables can lead to perfect multicollinearity (also known as the dummy variable trap). This occurs because the k dummy variables are linearly dependent; the value of one can be perfectly predicted from the others.

Dropping the first dummy variable establishes a reference category against which the effects of the other categories are measured. This reference category is implicitly represented by the absence of the dummy variables for the other categories in the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

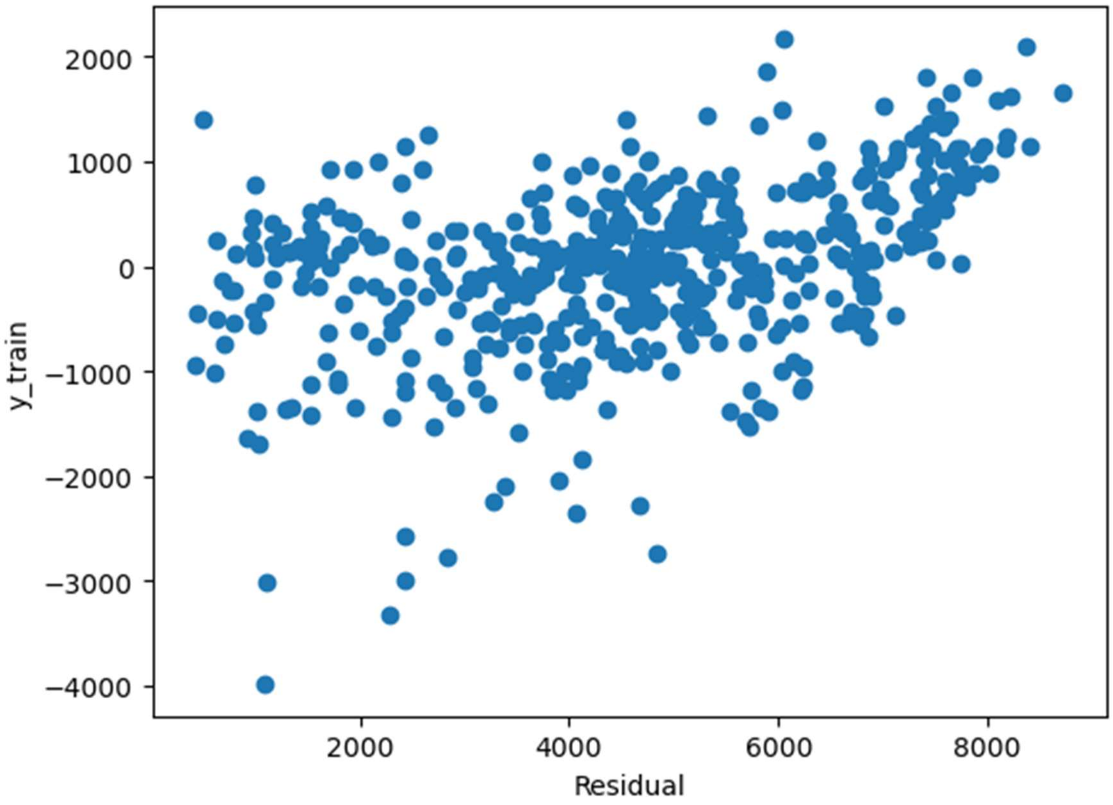
Ans: The numerical variable temp and atemp has the highest correlation with the target variable cnt.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: After building the linear Regression model on the training set, we can validate the assumptions of Linear Regression:

- **Linearity Check:**

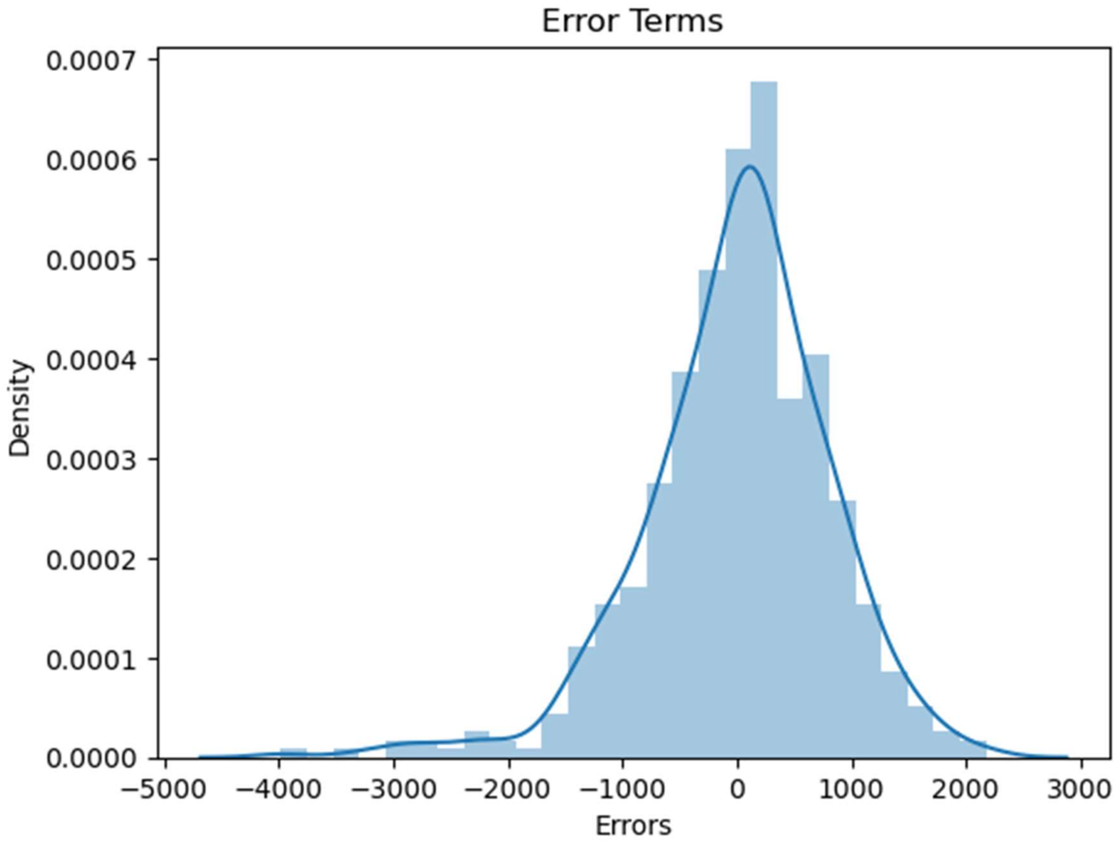


- **No auto-correlation:**
  1. A Durbin-Watson statistic of 2.112 suggests that there is no significant autocorrelation in the residuals of the regression model.

- **No Multicollinearity:**
  1. VIF (Variance Inflation Factor) VIF<=5 implies no multicollinearity between independent variables.

```
2. Features   VIF
3. 2          temp  4.80
4. 1    workingday  4.32
5. 0          year  2.08
6. 6           Sat  1.69
7. 8          Mist  1.55
8. 9        spring  1.55
9. 10       winter  1.51
10.   3          Dec  1.23
11.   4          Mar  1.17
12.   5          Sep  1.16
13.   7    LightSnow  1.07
```

- **Homoscedasticity**
  1. Residuals have constant variance at every level of x.
- **Normal distribution of error terms**



1. Error terms follow normal distribution.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1.temp with Coeff 3511.83

2. year with Coeff 2000.71

3. Sep with Coeff 587.53

```
  OLS Regression Results
==============================================================================
Dep. Variable:                  cnt   R-squared:                       0.824
Model:                          OLS   Adj. R-squared:                  0.820
Method:               Least Squares   F-statistic:                     212.3
Date:              Mon, 29 Jul 2024   Prob (F-statistic):          3.09e-180
Time:                      13:57:03   Log-Likelihood:                -4148.8
No. Observations:               511   AIC:                             8322.
Df Residuals:                   499   BIC:                             8373.
Df Model:                      11
Covariance Type:            nonrobust
==============================================================================
              coef    std err        t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      1607.2610    209.988     7.654      0.000    1194.692    2019.830
year       2000.7151     73.229     27.321      0.000    1856.841    2144.590
workingday   418.0908     97.550      4.286      0.000     226.431     609.751
temp       3511.8316    262.801     13.363      0.000    2995.498    4028.165
Dec       -325.5894    148.410     -2.194      0.029    -617.176     -34.003
Mar        526.1006    157.527      3.340      0.001     216.602     835.599
Sep        587.5322    130.776      4.493      0.000     330.593     844.472
Sat        462.5344    127.735      3.621      0.000     211.569     713.500
LightSnow  -2370.2325    220.801    -10.735      0.000   -2804.047   -1936.418
Mist       -660.1303     77.964     -8.467      0.000    -813.309    -506.952
spring    -1245.8149    143.049     -8.709      0.000   -1526.867    -964.762
winter      576.5599    110.610      5.213      0.000     359.242     793.878
==============================================================================
Omnibus:                     77.085   Durbin-Watson:                  2.112
Prob(Omnibus):                 0.000   Jarque-Bera (JB):              167.989
Skew:                        -0.817   Prob(JB):                    3.32e-37
Kurtosis:                      5.285   Cond. No.                        15.0
==============================================================================
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The primary goal is to predict the value of the dependent variable based on the values of the independent variables.

Linear regression aims to fit a linear equation to observed data. The simplest form is simple linear regression, which involves a single independent variable x and a dependent variable y. The relationship is modeled by the equation:

$y=\beta 0+\beta 1x+\epsilon$

where:

- y is the dependent variable.
- x is the independent variable.
- β0 is the y-intercept (the value of y when x=0).
- β1 is the slope of the line (the change in y for a one-unit change in x).
- ϵ is the error term (the difference between the observed and predicted values of y).

2. Assumptions:

Linear regression relies on several key assumptions:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The variance of error terms is constant across all levels of the independent variable.
- **Normality**: For inference purposes, the error terms are normally distributed.

3. Estimation of Parameters:

The parameters β0 and β1 are estimated using the least squares method, which minimizes the sum of the squared differences between the observed and predicted values.

4. Model Evaluation:

The fit of the linear regression model is evaluated using various metrics:

- **R-squared (R^2)**: Represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, with higher values indicating a better fit.
- **Residuals Analysis**: The residuals (errors) should be randomly scattered without patterns, indicating a good model fit.
- **p-values**: Used to determine the statistical significance of the coefficients. Low p-values (< 0.05) suggest that the independent variables are significant predictors of the dependent variable.

Linear regression is a foundational technique in statistics and machine learning, providing a straightforward way to model relationships between variables. Its simplicity and interpretability make it widely used, though it requires careful consideration of assumptions and proper evaluation to ensure valid results.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans: Anscombe’s quartet is a collection of four datasets that have nearly identical simple statistical properties, yet they appear very different when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how summary statistics alone can be misleading. Here’s a detailed explanation:

1. Description of the Datasets:

Each of the four datasets in Anscombe's quartet consists of 11 (x, y) points. Despite their different distributions, these datasets share several statistical properties:

- The mean of the x values is 9.
- The mean of the y values is approximately 7.5.
- The variance of the x values is 11.
- The variance of the y values is approximately 4.12.
- The correlation between x and y is around 0.816.
- The linear regression line has the equation y=3+0.5xy = 3 + 0.5xy=3+0.5x.

2. Importance of Graphical Analysis:

While the summary statistics are identical or nearly identical for all four datasets, their graphical representations reveal significant differences:

1. **Dataset I**: The points form a roughly linear pattern, making the linear regression model a good fit.
2. **Dataset II**: The points form a clear curve, indicating a non-linear relationship where a linear regression model is not appropriate.
3. **Dataset III**: All points except one are perfectly aligned in a linear pattern. The outlier significantly affects the regression line, showcasing the impact of outliers on statistical measures.
4. **Dataset IV**: Most points have the same x value, except one. This outlier drastically changes the appearance and the regression line, showing how a single data point can skew results.

3. Lessons from Anscombe's Quartet:

Anscombe's quartet teaches several important lessons:

- **Graphing Data**: Always visualize your data. Graphs can reveal patterns, trends, and outliers that summary statistics might obscure.
- **Context Matters**: Understanding the context of data is crucial. Summary statistics should not be interpreted in isolation.
- **Impact of Outliers**: Outliers can heavily influence statistical measures, including means, variances, and regression coefficients. Identifying and understanding outliers is essential.
- **Misleading Statistics**: Similar statistical properties do not guarantee similar data distributions. Graphs can help avoid misleading conclusions drawn from summary statistics alone.

Anscombe's quartet highlights the limitations of relying solely on summary statistics for data analysis and emphasizes the importance of visualizing data. By examining graphs alongside statistical measures, analysts can gain a more comprehensive understanding of the data and make more informed decisions.

3. What is Pearson’s R? (3 marks)

Ans: Pearson’s R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which a pair of variables are linearly related, indicating both the strength and direction of the relationship. Here's a detailed explanation:

1. Definition and Calculation:
Pearson's R is a statistic that ranges from -1 to 1. It is calculated using the formula:

2. Interpretation:
The value of Pearson's R indicates both the strength and direction of the linear relationship between two variables:

- **r=1**: Perfect positive linear relationship. As one variable increases, the other also increases proportionally.
- **r=−1**: Perfect negative linear relationship. As one variable increases, the other decreases proportionally.
- **r=0**: No linear relationship. Changes in one variable do not predict changes in the other. Values between -1 and 1 indicate the degree of linear correlation:
- **0 < |r| < 0.3**: Weak linear relationship.
- **0.3 ≤ |r| < 0.7**: Moderate linear relationship.
- **0.7 ≤ |r| < 1**: Strong linear relationship.

3. Assumptions and Limitations:
Pearson's R relies on certain assumptions and has limitations that must be considered:

- **Linearity**: Pearson's R measures only linear relationships. Non-linear relationships can lead to misleading interpretations.
- **Normality**: The variables should be approximately normally distributed for valid inferences, especially in smaller sample sizes.
- **Homogeneity of Variance (Homoscedasticity)**: The spread of data points should be consistent across the range of values.
- **Independence**: Observations should be independent of each other.
Limitations include:
- **Sensitivity to Outliers**: Outliers can significantly influence the correlation coefficient, leading to misleading conclusions.
- **Restricted Range**: If the data range is restricted, the correlation may underestimate the true relationship.
- **Causation**: Correlation does not imply causation. A high correlation between two variables does not mean one causes the other.

Pearson's R is a widely used measure of the linear relationship between two variables, providing insight into the strength and direction of the relationship. While it is a powerful tool, its proper application requires consideration of underlying assumptions and potential limitations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is a preprocessing step in data analysis and machine learning where the range of data is adjusted to fit within a specific scale. This is particularly important when the data consists of features with different units or magnitudes, which can negatively impact the performance of certain algorithms.

Scaling is performed for several reasons:

- **Improves Algorithm Performance**: Many machine learning algorithms, such as gradient descent-based methods (e.g., linear regression, logistic regression, neural networks), and distance-based methods (e.g., K-nearest neighbors, K-means clustering), perform better and converge faster when the features are on a similar scale.
- **Prevents Dominance of Larger Magnitude Features**: Features with larger magnitudes can dominate the learning process, leading to biased results. Scaling ensures that each feature contributes equally to the model.
- **Facilitates Comparisons**: Scaling allows for meaningful comparisons between features by bringing them to a common scale.

Difference Between Normalized Scaling and Standardized Scaling**:**

Normalized Scaling (Min-Max Scaling):

- **Definition**: This method scales the data to a fixed range, usually between 0 and 1.
- **Use Cases**: Normalized scaling is useful when you want to preserve the relationships between the original data values and scale them proportionately. It's often used when the data does not have outliers or when the algorithm does not assume normally distributed data.

Standardized Scaling (Z-score Scaling):

- **Definition**: This method scales the data such that it has a mean of 0 and a standard deviation of 1.
- **Use Cases**: Standardized scaling is useful when the data has outliers or when the algorithm assumes normally distributed data, as it centers the data around the mean and adjusts for variance. It is commonly used in algorithms like support vector machines, principal component analysis, and linear regression.

Scaling is a crucial preprocessing step to ensure that features contribute equally to the analysis and to enhance the performance of various algorithms. Normalized scaling adjusts the range of data to a fixed interval, usually [0, 1], while standardized scaling adjusts the data to have a mean of 0 and a standard deviation of 1. Both methods serve different purposes and are chosen based on the specific requirements of the analysis or algorithm.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity. A high VIF indicates high multicollinearity between the predictor variables, which can make the estimates of the regression coefficients unstable and difficult to interpret.

The value of VIF can become infinite when there is perfect multicollinearity among the predictor variables. This means that one predictor variable is an exact linear combination of one or more other predictor variables.

Perfect Multicollinearity:

- Perfect multicollinearity occurs when R2=1 which means that X can be perfectly predicted by a linear combination of the other predictors.

Causes of Perfect Multicollinearity

- **Duplicate Variables**: Including the same variable multiple times in different forms.
- **Linear Dependencies**: One variable is a perfect linear function of other variables
- **Dummy Variable Trap**: Including all categories of a categorical variable as dummy variables without dropping one category to serve as the reference group.

Implications of Infinite VIF

- **Unstable Coefficient Estimates**: Coefficients become very sensitive to small changes in the model or the data.
- **Inaccurate Statistical Tests**: Standard errors of the coefficients become inflated, leading to unreliable t-tests and p-values.
- **Model Interpretation**: The presence of perfect multicollinearity makes it impossible to determine the unique contribution of each predictor.

Addressing Infinite VIF

- **Remove Redundant Variables**: Identify and remove or combine variables that are linearly dependent.
- **Principal Component Analysis (PCA)**: Reduce the dimensionality of the data while retaining most of the variance.
- **Regularization Techniques**: Use methods like Ridge Regression or Lasso that can handle multicollinearity.

Infinite VIF occurs due to perfect multicollinearity, where a predictor variable is an exact linear combination of other predictor variables. This situation leads to unreliable and unstable regression coefficients, making the model difficult to interpret. Addressing multicollinearity involves identifying and modifying the relationships between predictor variables to ensure a more stable and interpretable model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Ans: A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.

Construction of a Q-Q Plot

1. **Quantiles Calculation**: Calculate the quantiles of the observed data and the theoretical distribution.
2. **Plotting**: Plot the quantiles of the observed data on the y-axis and the quantiles of the theoretical distribution on the x-axis.
3. **Interpretation**:
    1. If the data follows the theoretical distribution, the points should form a roughly straight line.
    2. Deviations from the line indicate departures from the theoretical distribution.

Use and Importance of a Q-Q Plot in Linear Regression

Q-Q plots are particularly important in the context of linear regression for diagnosing the assumptions about the residuals (errors). Linear regression relies on several key assumptions, including the normality of residuals, homoscedasticity, and independence. Here's how Q-Q plots are used and why they are important:

1. Assessing Normality of Residuals

- **Purpose**: One of the assumptions of linear regression is that the residuals (errors) are normally distributed. This assumption is crucial for making valid inferences about the regression coefficients and for the accuracy of confidence intervals and hypothesis tests.
- **Usage**: After fitting a linear regression model, plot the residuals on a Q-Q plot against a normal distribution.
- **Interpretation**: If the residuals are normally distributed, the points will lie on or near the straight line. Significant deviations suggest that the residuals are not normally distributed, which could affect the validity of the regression results.

2. Identifying Outliers

- **Purpose**: Outliers can have a disproportionate impact on the regression model, affecting the estimates of coefficients and the overall fit.
- **Usage**: In a Q-Q plot, outliers will appear as points that deviate significantly from the straight line.
- **Interpretation**: Identifying these outliers allows the analyst to investigate and potentially address these anomalies, either by transformation, removal, or understanding their influence on the model.

3. Detecting Skewness and Kurtosis

- **Purpose**: Skewness (asymmetry) and kurtosis (tailedness) of the residuals can indicate violations of normality.
- **Usage**: A Q-Q plot can reveal these characteristics:
  - **Skewness**: Points will curve away from the line in one direction.
  - **Kurtosis**: Heavy-tailed distributions will show points deviating more at the ends, while light-tailed distributions will show points clustering near the line.
- **Interpretation**: Understanding skewness and kurtosis helps in deciding if transformations (e.g., log, square root) are needed to normalize the residuals.

A Q-Q plot is a diagnostic tool used to compare the distribution of a dataset to a theoretical distribution, often the normal distribution. In linear regression, Q-Q plots are essential for assessing the normality of residuals, identifying outliers, and detecting skewness and kurtosis. Ensuring that the residuals meet the normality assumption is critical for the validity of the regression model's inferences, confidence intervals, and hypothesis tests.