Bike Sharing Case Study

Dibya

Assignment-based Subjective Questions

Question #1

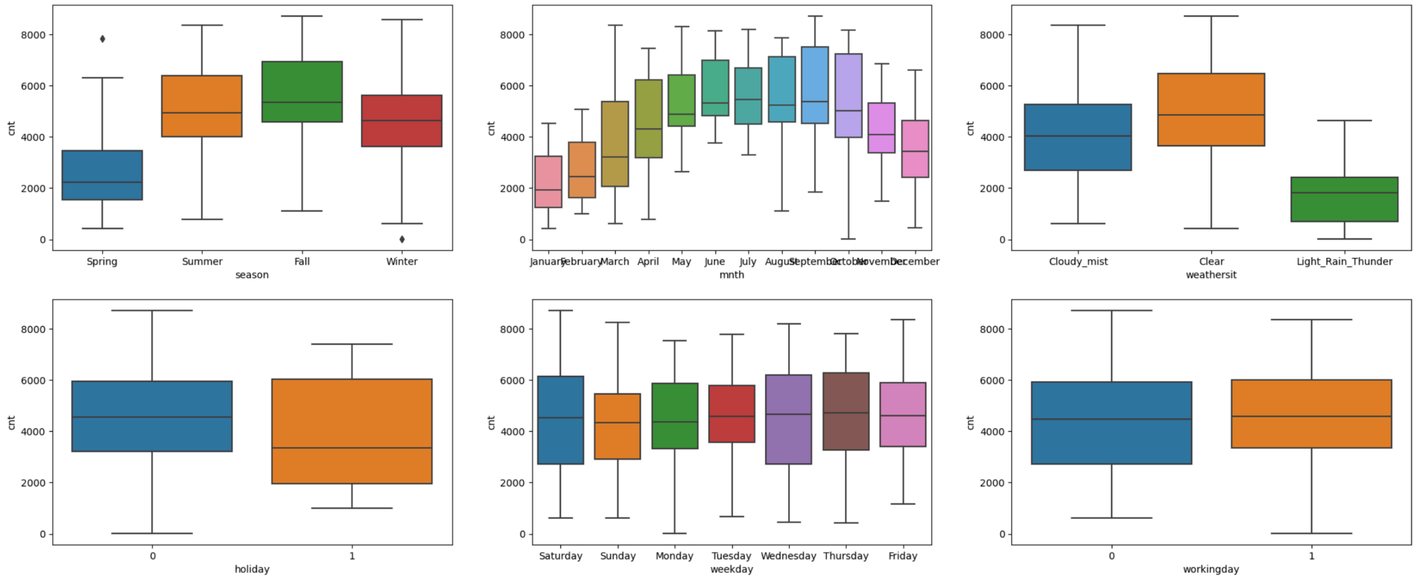
From your analysis of the categorical variables from the dataset, what could you infer about

- Season , mnth(month), weathersit , holiday , weekday and working day are categorical variables in the data set.
- Fall season the number of bike booking is more by looking at the distribution. Summer and Winter also has significant amount of booking. This can be a good predictor. Booking are happening with a median of more than 5000 for summer and fall.
- Month can be a good predictor as each month has a significant amount of booking.May , Jun , July ,Aug , Sept bookings are happening with a median of more than 4000 .
- Clear weather sit has more visitor due to which the bike hiring is happening with a median of more than 4000.
- Thunderstorm does not attract the bikers so it has very less booking compared to others

Question #1

From your analysis of the categorical variables from the dataset, what could you infer about

- Weekdays booking are consistent across the medians
- Working days and non working days both have booking which is consistent across the medians.



Question #2

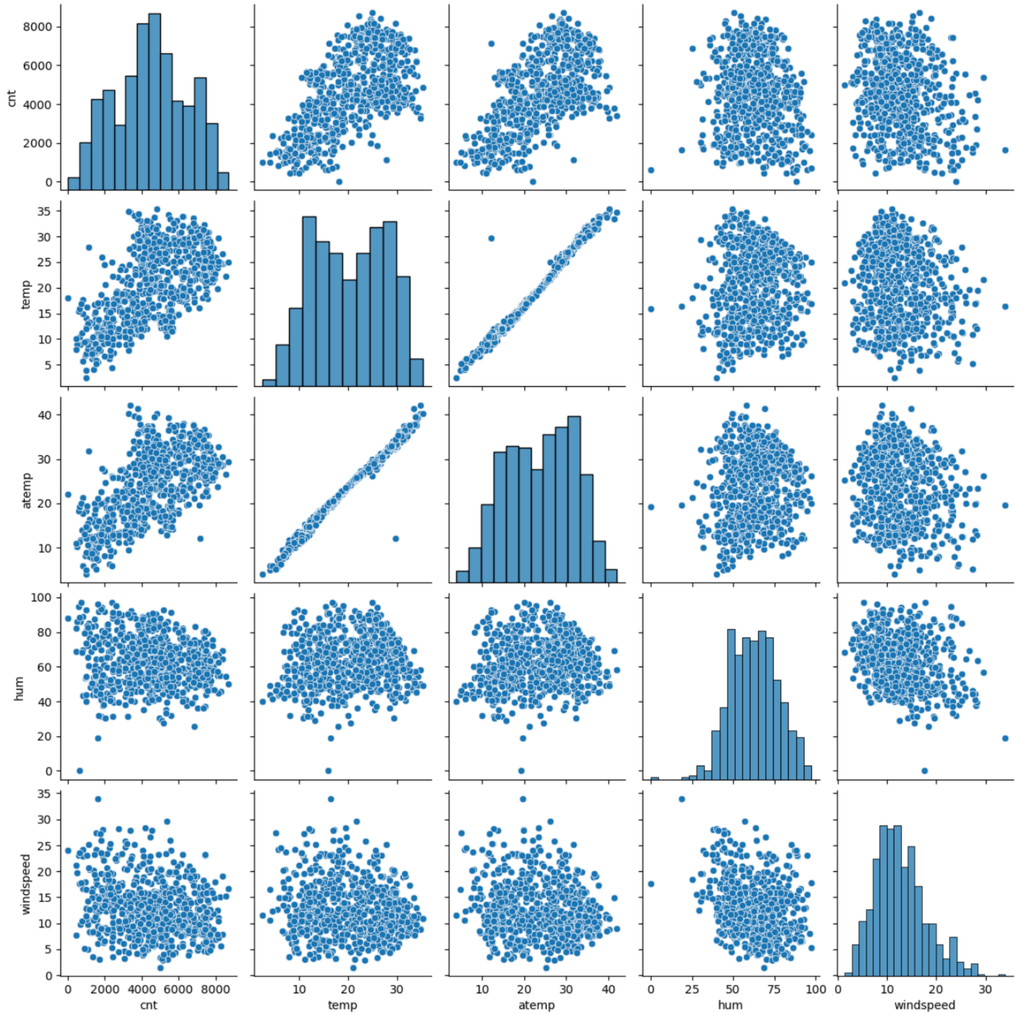
Why is it important to use drop_first=True during dummy variable creation?

- It avoids Multicollinearity
- If we don’t set this flag , then dummy column and the from the column which it derived will coexist in the data set. They can be predicated from other variables.
- Dropping the first dummy variable, its avoid this redundancy and the associated problems.

Question #3

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

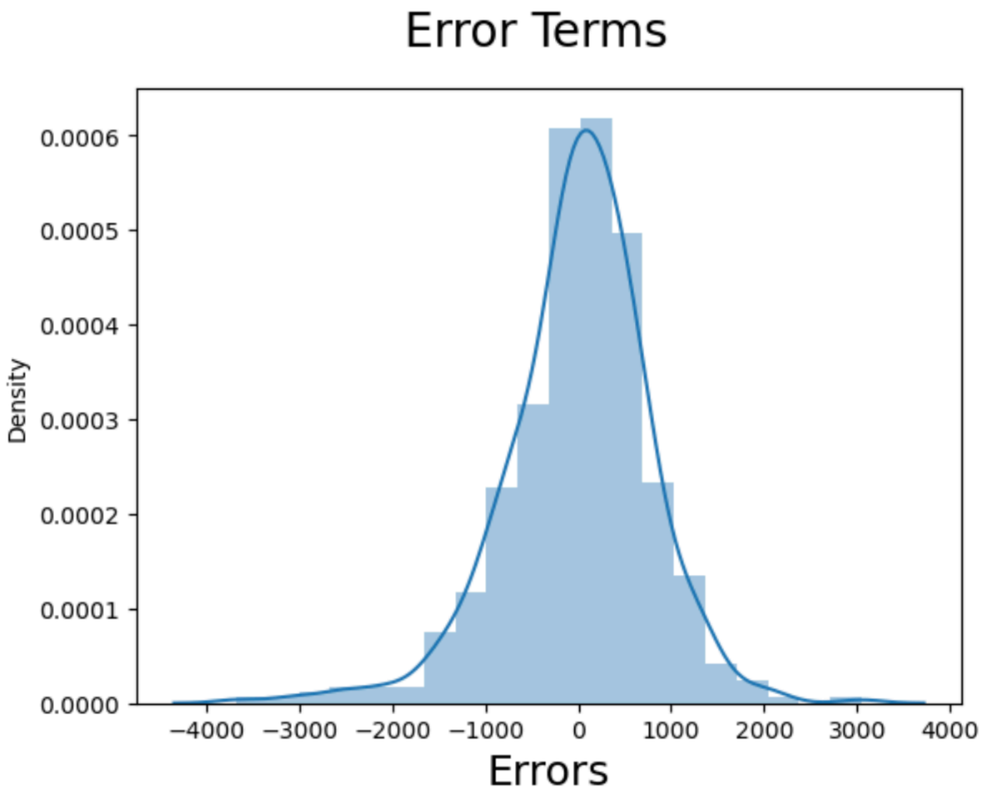
- temp , atemp both have highest correlation



Question #4

How did you validate the assumptions of Linear Regression after building the model on the Training set

- Error terms are normally distributed as per the LR assumption
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)
- Variance Inflation Factor (VIF) for each predictor was under desired limit.
- By ensuring no multicollinearity by using drop_first = true for dummy variables
- It quite evident from the plot that the above assumption holds good



Question #5

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

- Major Predictor
  - yr(Year)
    - Coefficient : 2029.34
      - Increase in year increase bike hiring by 2029.34
  - temp(Temperature):
    - Coefficient : 4265.03
      - Increase in temp increase bike hiring by 4265.03
  - windspeed
    - Coefficient : -1285.15
      - Increase in temp decreases bike hiring by -1285.15

	0
const	1812.85
yr	2029.34
holiday	-914.41
temp	4265.03
windspeed	-1285.15
Spring	-569.51
Summer	413.18
Winter	737.14
July	-427.95
September	660.70
Sunday	-420.22
Cloudy_mist	-714.46
Light_Rain_Thunder	-2516.27

General Subjective Questions

Question #1

Explain the linear regression algorithm in detail.

- Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting linear relationship, which can be used for prediction or understanding the underlying patterns.
- Basic Concept The linear regression model assumes that the relationship between the dependent variable y y and the independent variable(s) X X can be represented as a linear equation: $y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + \cdots + \beta p x p + \epsilon$ y=β 0 +β 1 x 1 +β 2 x 2 +⋯+β p x p +ϵ Where: y y is the dependent variable. $x 1 , x 2 , \ldots , x p$ x 1 ,x 2 ,...,x p are the independent variables. $\beta 0$ β 0 is the intercept. $\beta 1 , \beta 2 , \ldots , \beta p$ β 1 ,β 2 ,...,β p are the coefficients of the independent variables. ϵ ϵ is the error term (residual), representing the difference between the observed and predicted values.
- Steps in Linear Regression
  - Data Preparation:
  - Model Specification
  - Estimation of Coefficient
  - Model Fitting
  - Assumption Check
    - Linearity: The relationship between the dependent and independent variables is linear.
    - Independence: The residuals (errors) are independent.
    - Homoscedasticity: The residuals have constant variance at every level of the independent variable(s).
    - Normality: The residuals are normally distributed. No Multicollinearity: The independent variables are not highly correlated with each other.

Question #2

Explain the Anscombe’s quartet in detail.

- One of the primary uses of Anscombe's quartet is to demonstrate the critical role of data visualization in statistical analysis. Despite the datasets having nearly identical summary statistics(mean, variance, correlation, and linear regression coefficients)), their graphical representations reveal significant differences

- Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

- It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

- It serves as a reminder to validate statistical models and check assumptions,The quartet shows that a linear regression model might not be suitable for all datasets, even if the summary statistics suggest otherwise.

Question #3

What is Pearson's R?

- Pearson's R is a fundamental tool in statistics for measuring the strength and direction of the linear relationship between two continuous variables.
- Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.
- it is widely used in statistics and data analysis to assess whether two variables are related and how they change together.
- The value of Pearson's R ranges from -1 to 1: r = 1 r=1: Perfect positive linear relationship (as one variable increases, the other increases proportionally). r = − 1 r=−1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally). r = 0 r=0: No linear relationship (the variables do not move together in a linear fashion).
- Values between -1 and 1 indicate the degree of linear relationship:
  - 0.1 ≤ ∣ r ∣ < 0.3 0.1≤∣r∣<0.3: Weak correlation.
  - 0.3 ≤ ∣ r ∣ < 0.5 0.3≤∣r∣<0.5: Moderate correlation.
  - 0.5 ≤ ∣ r ∣ ≤ 1 0.5≤∣r∣≤1: Strong correlation.

Formula

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

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 #4

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range and distribution of numerical features. It ensures that the features contribute equally to the model's performance by bringing them to a common scale without distorting differences in the ranges of values.
- Scaling performed to
  - Improved Model Performance:
    - Many machine learning algorithms, especially those that rely on distance metrics (e.g., k-Nearest Neighbors, Support Vector Machines, and clustering algorithms), are sensitive to the scale of input features. Features with larger scales can dominate the learning process.
  - Faster Convergence:
    - Gradient-based algorithms like Gradient Descent used in linear regression, logistic regression, and neural networks can converge faster when features are scaled, as it helps avoid steep or flat gradients.
  - Equal Contribution
    - Ensures that all features contribute equally to the model's decisions, preventing features with larger magnitudes from disproportionately influencing the model.

Question #4

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Two type of scaling
  - Normalized Scaling
    - When you want to bound your features within a specific range.
    - When you know the distribution of the data is not Gaussian or you do not know the distribution.
    - Useful in algorithms that require bounded inputs, such as neural networks and k-Nearest Neighbors.
  - Standardized Scaling:
    - When the data follows a Gaussian (normal) distribution.
    - When you need to handle features with different units and scales.
    - Commonly used in algorithms like linear regression, logistic regression, and principal component analysis (PCA).

Question #5

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

- VIF becomes infinite when there is perfect multicollinearity in the data.
- Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of one or more other predictor variables.
- This means that one predictor can be exactly predicted from the others, leading to a situation where the denominator in the VIF calculation formula becomes zero.

Question #6

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

- A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given sample of data follows a particular distribution, typically the normal distribution. It compares the quantiles of the sample data to the quantiles of a specified theoretical distribution, such as the standard normal distribution.

- Importance

  - Assumption Checking: Q-Q plots are commonly used in linear regression to assess the assumption of normality of residuals. Residuals are the differences between the observed and predicted values in regression analysis. A Q-Q plot of residuals helps to visually inspect whether the residuals follow a normal distribution. If the residuals are normally distributed, the Q-Q plot will show points approximately along the diagonal line.

  - Model Validation: Q-Q plots are part of the diagnostic tools used to validate linear regression models. By examining the Q-Q plot, you can identify any departures from normality in the residuals, which may indicate potential issues with the model.

  - Decision Making: The information from the Q-Q plot can guide decisions about whether transformations or adjustments are needed to improve the model's performance. For example, if the Q-Q plot reveals significant deviations from normality, it may indicate the need for a different model specification or data transformation.

Question #6

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

- Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals. By visually comparing the quantiles of residuals to those expected under a normal distribution, analysts can gain insights into the adequacy of their regression model and make informed decisions about model validation and improvement.