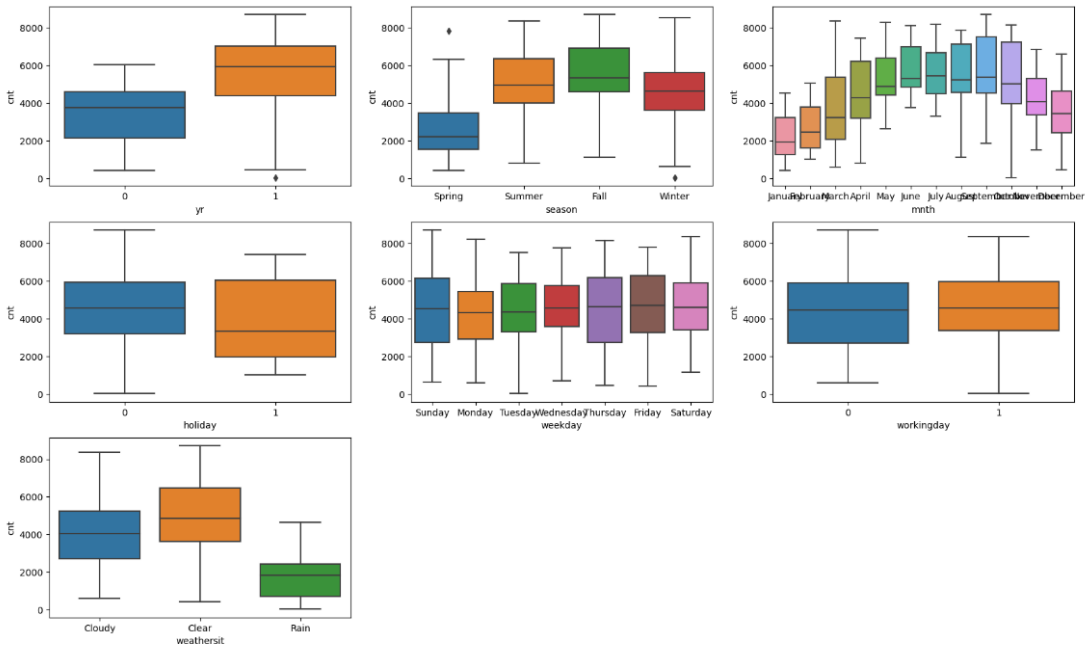
Assignment based Subjective questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: From the below boxplots of the categorical variables:



I found that the variables ‘yr’, ‘season’, ‘mnth’ and ‘weathersit’ have a significant amount of impact on the dependent variable. And the other categorical variables does not impact the dependent variable ‘cnt’ significantly as we can see that, medians of all the categories in each of the remaining categorical variable are similar to each other, which implies that those categories not does not explain the variance in the dependent variable significantly.

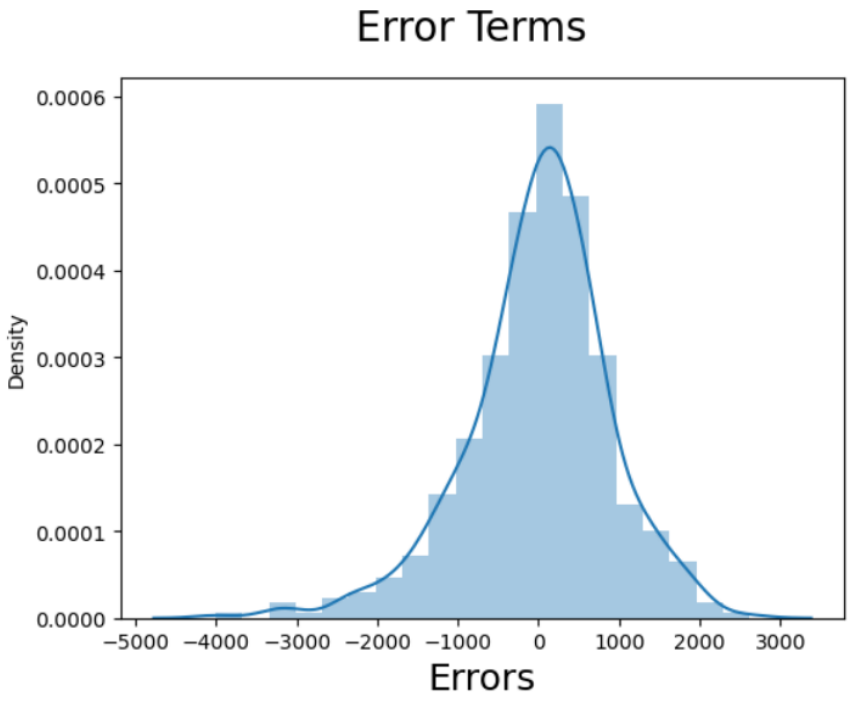
Q2: Why is it important to use drop_first=True during dummy variable creation?
Ans: Setting drop_first=True during dummy variable creation automatically drops the first level of each categorical variable. This means that if you have a categorical variable with n levels, you'll create (n-1) dummy variables. By dropping one level, you eliminate the multicollinearity issue, as the information from the dropped level is already captured by the other dummy variables. This helps in improving the stability and interpretability of the model.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: ‘temp’ variable has the highest correlation with the target variable.

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: Validating the assumptions of linear regression is crucial to ensure the reliability of your model.

>>Linearity:
Assumption: The relationship between the independent variables and the dependent variable is linear.
Validation: I have already used scatterplots to visually inspect the relationship between each independent variable and the dependent variable. You can also use residual plots to check for linearity.

>>Error terms are normally distributed with mean zero(Residual Analysis)
Assumption: Residuals should be normally distributed.
Validation: I have plot a distplot to validate this, as shown below:



Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Based on the final model, the top 3 features that significantly contribute towards explaining the demand of the shared bikes are - 'temp', 'windspeed' and 'yr'.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.
Ans: Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The core idea of linear regression is to model the relationship between the independent variables and the dependent variable as a linear equation.

>> Linear Equation
The linear regression model assumes a linear relationship between the independent variables (features) and the dependent variable (target). The basic form of a linear equation for a simple linear regression with one independent variable is:

$$\mathbf{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \mathrm{X} + \in$$

Where,
Y is the dependent variable, **X** is the independent variable, $\boldsymbol{\beta_0}$ is the y-intercept(constant), $\boldsymbol{\beta_1}$ is the slope of the line, and $\in$ is the error term.

>>Objective function
The objective in linear regression is to find the values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ that minimize the sum of squared differences between the predicted values ($\hat{Y}$) and the actual values (Y) in the training data. The objective function (also called the loss or cost function) is often represented as:

$$\textbf{Minimize} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \widehat{\mathbf{Y}}_\mathbf{i})^2$$

Various optimization techniques, such as gradient descent, are used to find the optimal values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$.

>>Training the model
During the training phase, the model is fitted to the training data. The algorithm adjusts the values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ iteratively to minimize the loss function.
Gradient descent is commonly used for optimization. It calculates the gradient of the loss function with respect to the model parameters and updates the parameters in the direction that minimizes the loss.

>>Predictions
Once the model is trained, it can be used to make predictions on new, unseen data. For a given set of independent variables, the model predicts the corresponding dependent variable using the learned coefficients.

>>Assumptions of Linear Regression
Linear regression makes several assumptions, including linearity (relationship is linear), independence of errors, homoscedasticity (constant variance of errors), and normality of errors.

>>Types of Linear Regression
Simple Linear Regression: Involves one independent variable.
Multiple Linear Regression: Involves more than one independent variable.

>>Evaluation
Model performance is often evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared.

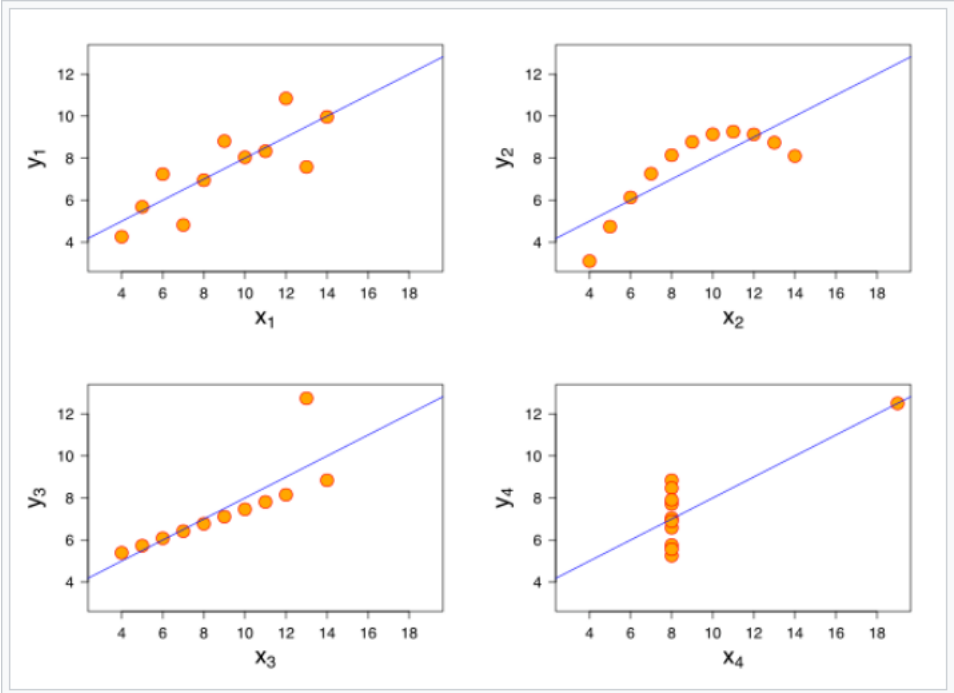
>>Regularization (optional)
In cases where multicollinearity is an issue, regularization techniques like Ridge or Lasso regression can be applied to penalize large coefficients.

>>Applications: Linear regression is widely used in various fields for tasks like predicting sales, housing prices, stock prices, and more.

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but exhibit very different characteristics when graphically visualized or analyzed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphically exploring and understanding data before relying solely on summary statistics.
The four datasets in Anscombe's quartet share the same mean, variance, correlation coefficient, and linear regression line parameters, but they differ significantly in their distributions and patterns. This highlights the limitations of relying solely on summary statistics and underscores the importance of visualizing data to gain a more comprehensive understanding.

For all four datasets:
Mean of x: 9.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.12
Correlation coefficient: 0.816
Linear regression:y=3.0+0.5x



Despite the similar summary statistics, when you graphically represent these datasets, you'll notice that they have distinct shapes and patterns. One might have a linear relationship, another could be quadratic, and so

on. Anscombe's quartet serves as a cautionary example against relying solely on summary statistics and emphasizes the importance of data visualization in statistical analysis. It reminds analysts and researchers to explore their data visually and to look for patterns and relationships that might not be apparent in summary measures alone.

Q3: What is Pearson's R?
Ans: Pearson's correlation coefficient, often denoted as "r" or Pearson's "r," is a measure of the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson, a British statistician, in the late 19th century.

The Pearson correlation coefficient ranges from -1 to 1. The value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally. A value of -1 indicates a perfect negative linear relationship, where as one variable increases, the other decreases proportionally. A value of 0 suggests no linear correlation between the two variables.

The formula for Pearson's correlation coefficient (r) is:

$$\mathbf{r} = \frac{\sum(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{Y_i} - \overline{\mathbf{Y}})}{\sqrt{\sum(\mathbf{X_i} - \overline{\mathbf{X}})^2 \cdot \sum(\mathbf{Y_i} - \overline{\mathbf{Y}})^2}}$$

Where:
$\mathbf{X_i}$ and $\mathbf{Y_i}$ are the individual data points of the two variables.
$\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are the means of the two variables.

Pearson's correlation coefficient is widely used in statistics and research to quantify the strength and direction of relationships between variables. It is important to note that Pearson's correlation measures only linear relationships and may not accurately capture non-linear associations between variables.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a preprocessing technique in data analysis and machine learning that involves adjusting the scale of the input features. The goal of scaling is to bring all the features to a similar range or scale to ensure

that no single feature dominates the others. This is particularly important when working with algorithms that are sensitive to the magnitude of the input features, such as gradient-based optimization algorithms in machine learning.

Scaling is performed for several reasons:

>>Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of the input features. For example, distance-based algorithms like k-nearest neighbors or clustering algorithms may be affected by the scale of the features.

>>Convergence Speed: Gradient descent-based optimization algorithms converge faster when the features are on a similar scale. This is crucial for iterative optimization processes.

>>Regularization: Regularization techniques, like L1(lasso regression) or L2(ridge regression) regularization, assume that all features are on the same scale. Scaling helps in achieving this assumption.

Two common methods for scaling are normalized scaling and standardized scaling:
A) Normalized Scaling (Min-Max Scaling):
-formula: **Xnorm = X-Xmin/Xmax-Xmin**
-Normalization scales the features to a specific range, usually between 0 and 1. It is useful when the features have different ranges and the algorithm requires them to be within a specific interval.

B) Standardized Scaling (Z-score normalization):
-formula: **Xstd = X-μ/σ**
-Standardization transforms the features to have a mean (μ) of 0 and a standard deviation (σ) of 1. This method assumes that the data follows a normal distribution and is less sensitive to outliers compared to Min-Max Scaling.

Both normalized scaling and standardized scaling aim to make features comparable in scale, but they use different formulas and have different effects on the distribution of the data. The choice between them depends on the specific requirements of the algorithm being used and the characteristics of the data.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the severity of multicollinearity in a set of independent variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to isolate the individual effect of each variable on the dependent variable.

The formula for calculating the VIF for a variable is:

$$\mathbf{VIF_i} = \frac{\mathbf{1}}{\mathbf{1} - \mathbf{R_i^2}}$$

Where,
R_i^2 is the R-squared value obtained by regressing the i^{th} independent variable against all other independent variables.
If the R_i^2 value is very close to 1, it indicates that the i^{th} variable is highly correlated with the other variables, leading to a situation where the denominator in the VIF formula becomes very close to zero. In mathematical terms, dividing by a number close to zero results in a very large number, and when this happens, the VIF can become extremely high, leading to values that are practically infinite.

So, the VIF becomes infinite when there is perfect multicollinearity, meaning one or more independent variables in the regression model can be exactly predicted by a linear combination of the others. In practical terms, this indicates a severe problem with the model, as it becomes challenging to discern the individual impact of correlated variables.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It is commonly used in statistics to compare the distribution of a sample with a known or assumed theoretical distribution, such as a normal distribution.

The Q-Q plot is constructed by plotting the quantiles of the observed data against the quantiles of the theoretical distribution. If the points on the plot fall along a straight line, it suggests that the data follows the theoretical distribution. Deviations from the straight line indicate departures from the assumed distribution.
In the context of linear regression, Q-Q plots are often used to check the normality of residuals. Residuals are the differences between the observed values and the values predicted by the regression model. The normality assumption of residuals is crucial for valid statistical inference and hypothesis testing in linear regression.

Here's how the Q-Q plot is used in the context of linear regression:

>>Residuals Normality Check:
After fitting a linear regression model, you obtain residuals by subtracting the predicted values from the observed values.
A Q-Q plot of the residuals is created, with the quantiles of the residuals on the y-axis and the quantiles of the normal distribution on the x-axis.

>>Interpretation:
If the residuals follow a normal distribution, the points on the Q-Q plot will roughly form a straight line.
Deviations from the straight line may indicate non-normality in the residuals.

>>Importance:
Assumptions of normality are important for valid hypothesis testing and confidence interval estimation in linear regression.
Violations of the normality assumption can affect the accuracy and reliability of statistical inferences based on the regression model.

>>Identification of Outliers:
Q-Q plots can also be helpful in identifying outliers in the residuals. Outliers may appear as points deviating significantly from the expected line.

Therefore, the Q-Q plot is a valuable diagnostic tool in linear regression analysis. It helps assess the normality of residuals, which is a critical assumption for making valid statistical inferences based on regression models.