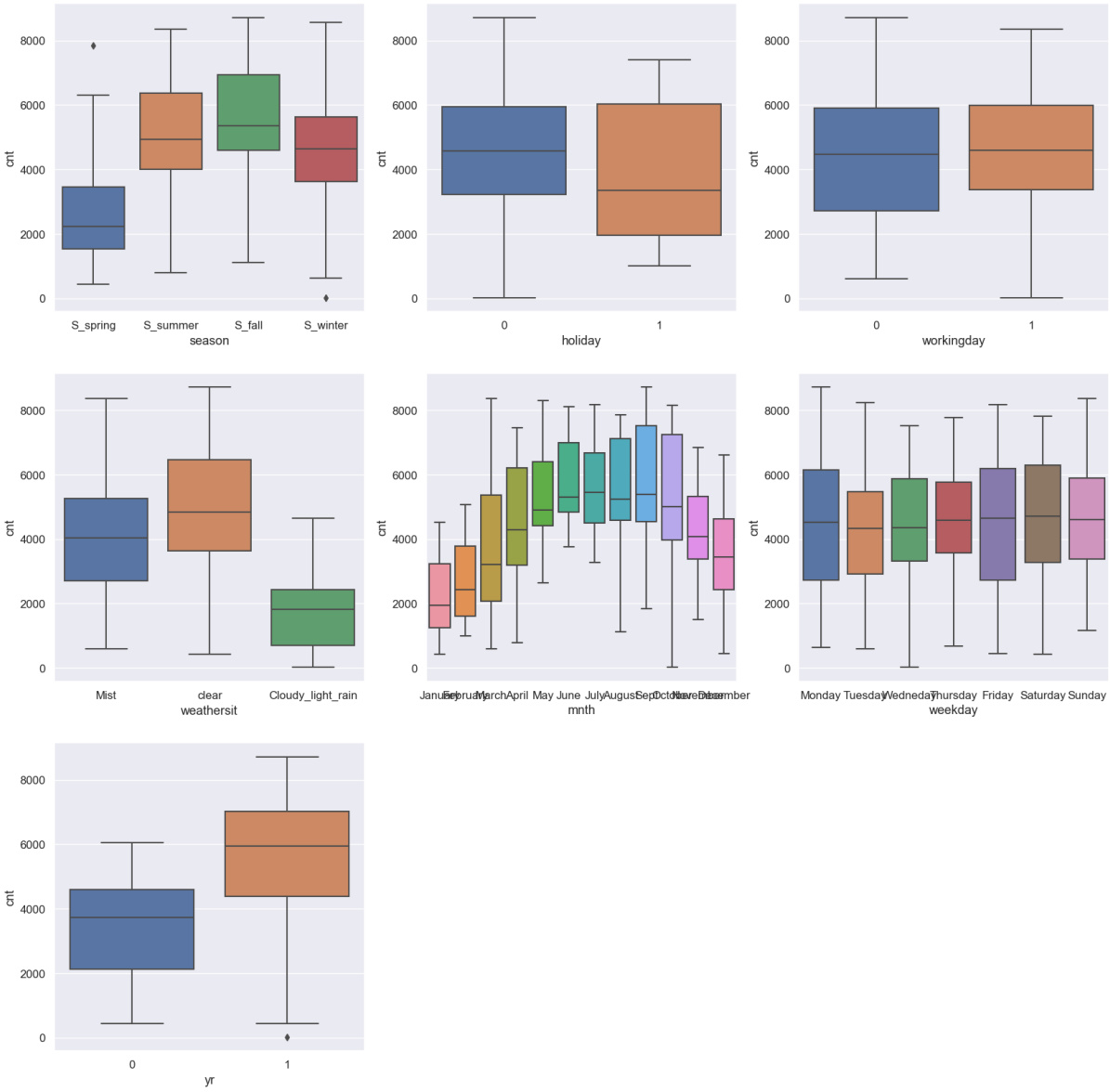
Assignment-based Subjective Questions

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

Multiple categorical features are present in dataset like season, mnth, holiday, weathersit etc.

The effect of these on dependant variable is plotted with boxplot



We will get huge information from above box plot .

- Cnt is higher in summer and fall season
- Holiday is not affecting more on cnt
- Working is also not affecting significantly
- Clear weather has higher cnt, mist and cloudy environment has lesser cnt
- Cnt distribution is variable with mnth
- Even distribution during workday
-

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

When creating dummy variables from categorical variables, it is common to use the "one-hot encoding" technique. This technique involves creating a binary variable for each category of the original categorical variable, where the value of the variable is 1 if the observation belongs to that category, and 0 otherwise.

However, when creating dummy variables, it is important to use the drop_first=True parameter. This parameter specifies that one of the categories of the original categorical variable should be dropped, and not included as a dummy variable.

The reason for dropping one category is to avoid the problem of "dummy variable trap", which can occur when all the categories of the original categorical variable are included as dummy variables. The dummy variable trap occurs when the variables are linearly dependent, meaning that one variable can be expressed as a linear combination of the other variables. This can lead to issues such as overfitting, multicollinearity, and unstable coefficient estimates.

By dropping one category and using drop_first=True, we can avoid the dummy variable trap and ensure that the variables are linearly independent. In addition, it can also help to reduce the dimensionality of the dataset, which can be beneficial for computational efficiency.

In summary, it is important to use drop_first=True when creating dummy variables to avoid the dummy variable trap and ensure that the variables are linearly independent.

Suppose we have a dataset of bike rental counts, where one of the categorical variables is "season", which can take the values of 1, 2, 3, or 4. These values represent the four seasons of the year, where 1 is spring, 2 is summer, 3 is fall, and 4 is winter.

To create dummy variables from this categorical variable, we can use the pd.get_dummies() function in Python.

the drop_first=True parameter to drop the first category (spring).

The resulting output will be a new dataframe with additional columns for each of the remaining categories (summer, fall, and winter), where each column will have a value of 1 if the observation belongs to that category, and 0 otherwise.

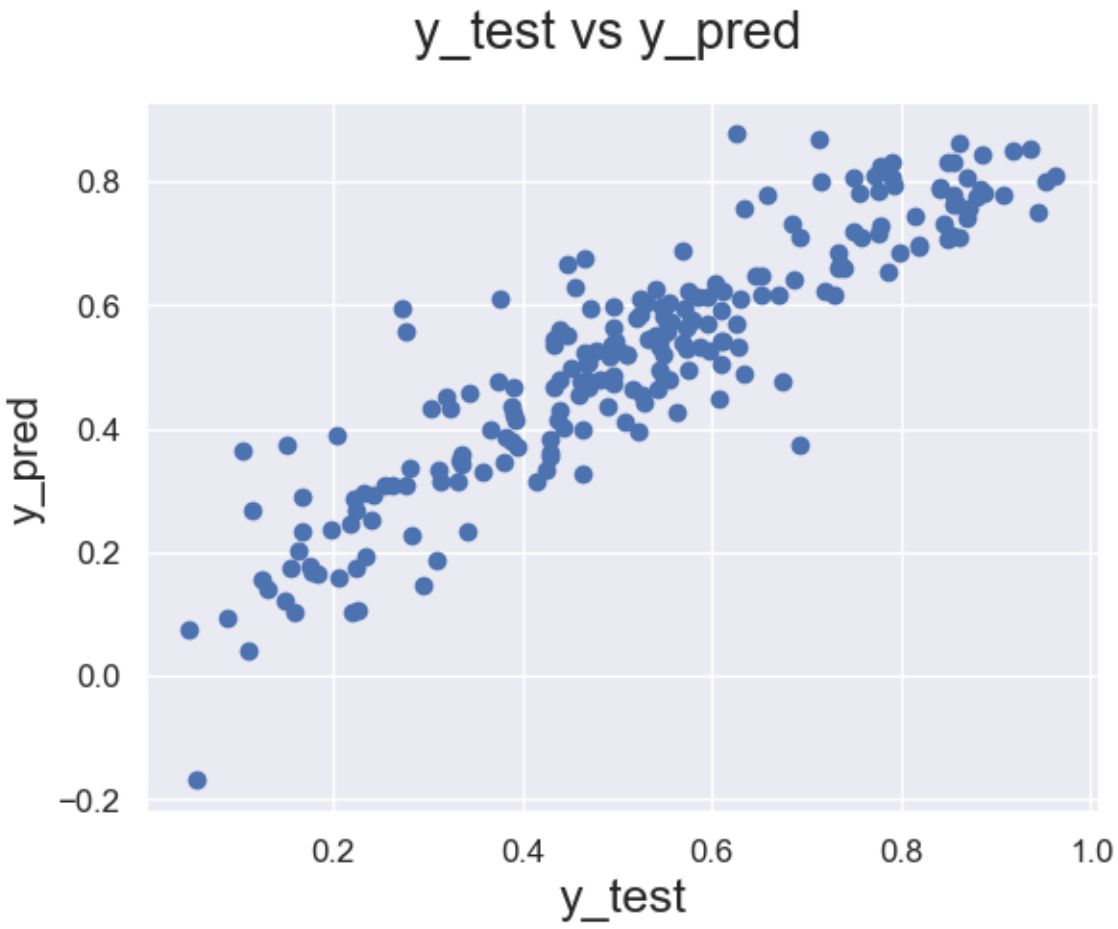
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

To determine which numerical variable has the highest correlation with the target variable based on the pair-plot, we can look for the scatter plot that shows the strongest linear relationship with the target variable.

And **‘temp’ Vs ‘cnt’** showed the strongest / highest correlation with the target variable which is considered for model building.

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

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target



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

- the top 3 features contributing significantly towards target variable is

1. temp

2. windspeed

3. Clear_weathersit

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm that is used for modelling the relationship between a dependent variable and one or more independent variables. It is a simple but powerful algorithm that has many practical applications in various fields, including economics, finance, marketing, and social sciences.

The main idea behind linear regression is to find the best line that fits the data points in a given dataset. This line is known as the regression line, and it represents the relationship between the independent variables (also known as predictors or features) and the dependent variable (also known as the response or target).

The regression line is defined by the equation:

$y = b0 + b1x1 + b2x2 + ... + bn*xn$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients of the regression line.

The goal of the linear regression algorithm is to find the values of the coefficients that minimize the sum of the squared errors (SSE) between the predicted values of the dependent variable and the actual values. The SSE is defined as:

$SSE = \Sigma(yi - \hat{y}i)2$

where yi is the actual value of the dependent variable for the i-th observation, and $\hat{y}i$ is the predicted value of the dependent variable for the i-th observation.

The linear regression algorithm uses a technique called ordinary least squares (OLS) to estimate the coefficients of the regression line. OLS works by minimizing the SSE with respect to the coefficients. This is done by taking the partial derivative of the SSE with respect to each coefficient, setting it equal to zero, and solving for the coefficient.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have identical statistical properties but exhibit markedly different graphical representations. These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of relying solely on summary statistics.

Each of the datasets in Anscombe's quartet consists of 11 pairs of x and y values. The summary statistics for each dataset are as follows:

Dataset I: x = 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5; y = 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68;

Mean of x = 9.0, Mean of y = 7.5, Variance of x = 11.0, Variance of y = 4.12, Correlation coefficient between x and y = 0.82

Dataset II: x = 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5; y = 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74;

Mean of x = 9.0, Mean of y = 7.5, Variance of x = 11.0, Variance of y = 4.12, Correlation coefficient between x and y = 0.82

Dataset III: x = 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5; y = 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73;

Mean of x = 9.0, Mean of y = 7.5, Variance of x = 11.0, Variance of y = 4.12, Correlation coefficient between x and y = 0.82

Dataset IV: x = 8, 8, 8, 8, 8, 8, 8, 19, 19, 19, 19; y = 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 10.84, 9.13, 7.26;

Mean of x = 9.0, Mean of y = 7.5, Variance of x = 11.0, Variance of y = 4.12, Correlation coefficient between x and y = 0.82

As can be seen from the summary statistics, each of the datasets has the same mean and variance for both x and y, as well as the same correlation coefficient between x and y. However, when plotted, each dataset exhibits a unique pattern.

Dataset I forms a roughly linear relationship between x and y, with a slight positive slope. Dataset II also shows a linear relationship, but with a steeper slope and a different intercept. Dataset III has a non-linear relationship, with an apparent parabolic shape. Dataset IV has a very weak linear relationship, but is heavily influenced by an outlier point.

The significance of Anscombe's quartet lies in its ability to highlight the limitations of relying solely on summary statistics to describe a dataset. Even though each dataset has identical summary statistics, they have different patterns when plotted. This underscores the importance of visualizing data to gain a more complete understanding of its structure and behaviour.

Additionally, Anscombe's quartet can be used to illustrate the importance of data cleaning and outlier detection. Dataset IV, for example, has an outlier point that heavily influences the relationship between x and y. Removing this point would result in a completely different pattern.

In summary, Anscombe's quartet is a powerful demonstration of the importance of visualizing data and the limitations of relying solely on summary statistics. It highlights the need for careful data cleaning and outlier detection, as well as the importance of choosing appropriate data visualization techniques to gain a deeper understanding of the underlying patterns in a dataset.

3.What is Pearson’s R?

Pearson's r, also known as the Pearson correlation coefficient or Pearson's product-moment correlation coefficient, is a measure of the linear correlation between two continuous variables. It was developed by the British mathematician Karl Pearson in the late 19th century.

The Pearson correlation coefficient is a number between -1 and +1, where -1 indicates a perfectly negative correlation, 0 indicates no correlation, and +1 indicates a perfectly positive correlation. The formula for calculating Pearson's r is:

$$r = (\Sigma[(x_i - \bar{x})(y_i - \bar{y})]) / \text{sqrt}(\Sigma[(x_i - \bar{x})^2] * \Sigma[(y_i - \bar{y})^2])$$

where xi and yi are the values of the two variables, $\bar{x}$ and $\bar{y}$ are their means, and Σ denotes the sum of the terms over all the data points.

Pearson's r is widely used in statistics and data analysis to quantify the strength and direction of the linear relationship between two variables. It is particularly useful when analysing bivariate data, where the values of two variables are measured for each individual or observation.

However, it is important to note that Pearson's r only measures linear correlation, and may not be appropriate when the relationship between two variables is non-linear. Additionally, Pearson's r can be influenced by outliers and may not be robust in the presence of extreme values. In such cases, alternative correlation measures such as Spearman's rank correlation coefficient or Kendall's tau may be more appropriate.

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

Scaling refers to the process of transforming numerical data so that it falls within a specific range or distribution. The purpose of scaling is to standardize the range of features or variables in a dataset, so that they can be compared and analysed more effectively.

There are several reasons why scaling is performed:

To improve the performance of machine learning algorithms: Some machine learning algorithms, such as K-nearest neighbours and gradient descent, can be sensitive to the scale of the input features. Scaling the features to a similar range can improve the performance of these algorithms.

To enable comparison of variables: Variables with different ranges or units cannot be compared directly. Scaling enables the comparison of variables on the same scale, making it easier to identify patterns and relationships between variables.

To reduce the influence of outliers: Scaling can reduce the impact of extreme values, making the analysis more robust and accurate.

There are several methods for scaling data, but two of the most common methods are normalized scaling and standardized scaling:

Normalized scaling: Normalization scales the data to a range of 0 to 1, so that the minimum value in the data is scaled to 0, and the maximum value is scaled to 1. Normalization preserves the shape of the distribution of the data, but may not be appropriate when there are outliers in the data.

Standardized scaling: Standardization scales the data to have a mean of 0 and a standard deviation of 1. Standardization preserves the shape of the distribution of the data and is robust to outliers. Standardized scaling is often preferred over normalization in machine learning applications.

In summary, scaling is an important data pre-processing step that helps to standardize the range of variables in a dataset. Normalized scaling and standardized scaling are two common methods for scaling data, with standardized scaling being preferred in machine learning applications.

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

Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables in a multiple regression analysis. It measures how much the variance of the estimated regression coefficient is inflated due to collinearity among the predictor variables.

In some cases, the value of VIF can be infinite. This happens when one or more of the predictor variables in the model is perfectly correlated with a linear combination of the other predictor variables. In other words, there is complete multicollinearity among the predictor variables.

Complete multicollinearity occurs when one or more of the predictor variables can be expressed as a linear combination of the other predictor variables. For example, if we have two predictor variables, x1 and x2, and x2 = 2*x1, then x2 is completely determined by x1, and there is perfect multicollinearity between the two variables.

In such cases, when the correlation between the predictor variables is perfect, the coefficient estimates cannot be determined uniquely. This leads to an infinite value of VIF for the variable that is involved in the linear combination.

It is important to detect and address multicollinearity in a multiple regression analysis, as it can lead to inaccurate and unstable coefficient estimates, and can make it difficult to interpret the effects of the individual predictor variables. One way to address multicollinearity is to remove one or more of the correlated predictor variables from the model. Another approach is to use regularization techniques, such as ridge regression or lasso regression, which can help to reduce the impact of multicollinearity on the coefficient estimates.

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 technique used to determine if a set of data comes from a known distribution, such as a normal distribution. The Q-Q plot compares the observed data to the expected values of a theoretical distribution, and plots the quantiles of the observed data against the corresponding quantiles of the theoretical distribution.

The use and importance of a Q-Q plot in linear regression can be explained as follows:

Checking the normality assumption: One of the important assumptions of linear regression is that the residuals (the difference between the predicted values and the actual values) should follow a normal distribution. The Q-Q plot can be used to visually check if the residuals are normally distributed or not. If the residuals follow a straight line on the Q-Q plot, then it suggests that the residuals are normally distributed.

Detecting outliers: The Q-Q plot can also be used to detect outliers in the data. Outliers are points that do not follow the expected pattern of the data, and they can affect the results of a linear regression analysis. If there are outliers in the data, they will appear as points that are far away from the straight line on the Q-Q plot.

Assessing goodness of fit: The Q-Q plot can be used to assess the goodness of fit of a linear regression model. If the residuals follow a straight line on the Q-Q plot, then it suggests that the model fits the data well. If the residuals do not follow a straight line, then it suggests that there may be a problem with the model, such as nonlinearity or heteroscedasticity.

In summary, the Q-Q plot is a useful tool for checking the normality assumption, detecting outliers, and assessing the goodness of fit of a linear regression model. It provides a visual representation of how well the observed data fits a theoretical distribution, and can help to identify any issues that may affect the validity of the linear regression analysis.