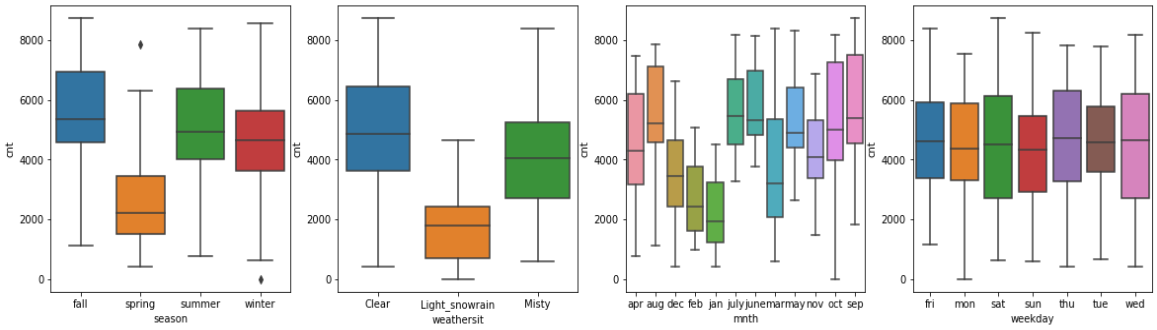
Assignment-based Subjective Questions

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

   Ans:
   The categorical variables like season, month, weathersit and weekday had clear impact on the dependent variable. This can be seen clearly through visualization.



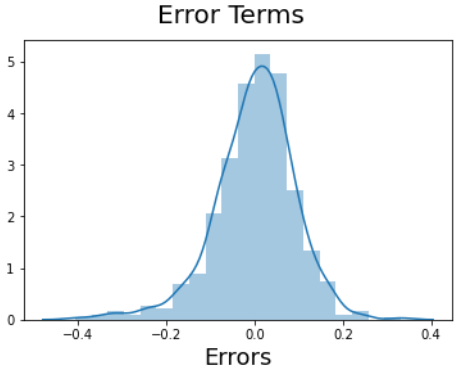
   - Weathersit and season clearly indicates the influence of its each class on cnt.
   - Fall season has more demand than others.
   - Clear weather also indicates more demand and Light_snowrain has least demand of bikes.
   - Aug, June, July, oct months shows more demand compared to others**.**
   - All Weekdays have somewhere similar influence on cnt.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Ans:
   It is important to use drop_first = true because while creating dummies it creates n dummy variables for n level. And according to the theory we can represent n classes with n-1 dummy variables.

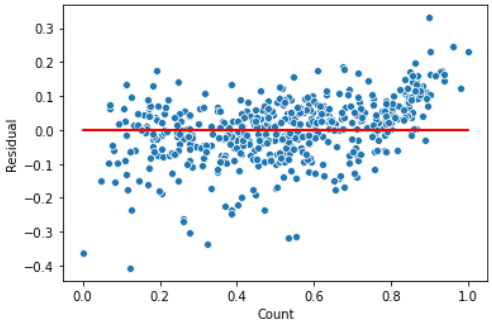
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans:
   Among the numerical variables, temp and atemp has highest correlation of around 0.63 with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Ans:
   I have validated the assumptions of Linear Regression by the following analysis-

   - Normality of error terms.



   - Homoscedasticity
     There are no visible patterns in residuals.



- Multicollinearity
  There should not be any significant correlation among independent variables.
- Independence of residuals
  No auto -correlation.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans:
   Based on final model parameters, below 3 are the top features contributing significantly towards explaining the demand of shared bikes -
   - temp
   - year
   - light_snowrain

General Subjective Questions.

1. **Explain the linear regression algorithm in detail.**
   Ans:
   Linear Regression is a machine learning algorithm based on supervised learning.
   Linear regression is a statistical model which analyses the linear relationship between independent and dependent variable. Linear relationship between variables means if independent variable changes(increase/decrease) then dependent variable also changes with respect to the independent variable.

   Mathematically the relationship can be defined by following equation:

   $$Y = mX + c$$

   Here, Y is dependent variable to predict
   X is the dependent variable
   m is the slope of the regression line
   c is a constant, known as the Y intercept. If X = 0, Y would be equal to c.

   Linear regression is of two types:
   - Simple Linear Regression
   - Multiple Linear Regression

- Assumptions
  - No Multi-collinearity
  - No Auto-correlation
  - Relationship between response and feature variables should be linear.
  - Normality of error terms.

2. **Explain the Anscombe's quartet in detail.**

Ans:
Anscombe's quartet was constructed in 1973 by statistician Francis to illustrate the importance of plotting data before analysing it and building model. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. **What is Pearson's R?**

Ans :

There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Between 0 and 1 indicates Positive correlation. 0 indicates no correlation and between 0 and -1 indicates Negative correlation. When the value of the correlation coefficient is exactly 1.0, it is said to be a perfect positive correlation. A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all.

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

Ans :

Feature Scaling is a technique to bring the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Difference between normalized scaling and standardized scaling:

Normalized/Min-Max Scaling : - This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardized Scaling :- It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

Standardization may be used when data represent Gaussian Distribution, while Normalization is great with Non-Gaussian Distribution. Impact of Outliers is very high in Normalization.

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

Ans:

If there is a perfect correlation then VIF is infinite. A large value of VIF indicates that there is a high correlation between variables. In case of perfect correlation we get r-squared =1 which makes 1/(1-r-squared) infinity.

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

Ans:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

In summary, A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.