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?

   - During the fall season, the rental count is high.
   - From June to September people seemed to use bikes
   - In the year 2019, there were large rentings of bikes.
   - Working days i.e. Wednesday, Thursday have a high count of bike rentals.
   - If the weather is clear, the rentals of bikes increase and as the weather sit is moderate or high, the bike rentals decrease.

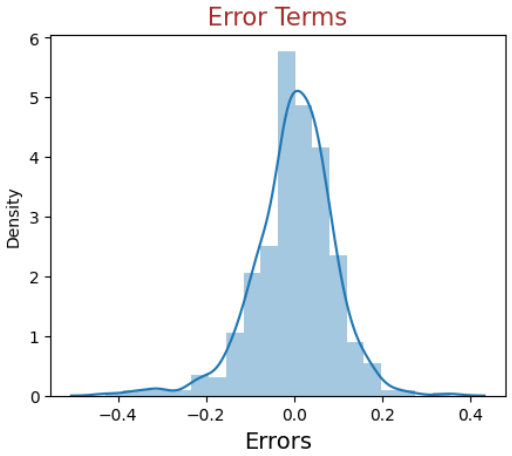
2. Why is it important to use drop_first=True during dummy variable creation?

   It is important to use drop_first=True while creating dummy variables, as it helps to remove the redundant or extra column which reduces the high correlations i.e. Multi-collinearity among the dummy variables.

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

   The variables ‘temp’ and ‘atemp’ have 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?



   - The distribution of residuals is centered at 0.
   - The residuals are scattered around the mean i.e. at 0.
   - The residual plot is a normal distribution.

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

   According to the final model, the three features contributing significantly to the demand for shared bikes are,

   1. **Temperature (temp)**: As the temperature increases by 1 unit, the count increases by 0.4509. It is positively correlated.
   2. **Year (yr)**: It is positively correlated with a coefficient of 0.2344.
   3. **Weather Sit (weathersit_high)**: As the weathersit_high increases by 1 unit, the count decreases by -0.2905. It is negatively correlated.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   - Linear Regression is a Supervised Machine Learning model. It is an algorithm that shows a linear relationship between an independent and a dependent variable to predict the new datasets.
   - The equation of Linear Regression is,

     $y = mx + c$

     where, m is the slope and c is the intercept.
   - There are two types of Linear Regression:
     1. Simple Linear Regression:
        This regression method takes one dependent and one independent variable i.e. x and y.
        The equation for Simple Linear regression is,

        $y = \beta 0 + \beta 1$

        where, β0 is the intercept and β1 is the slope

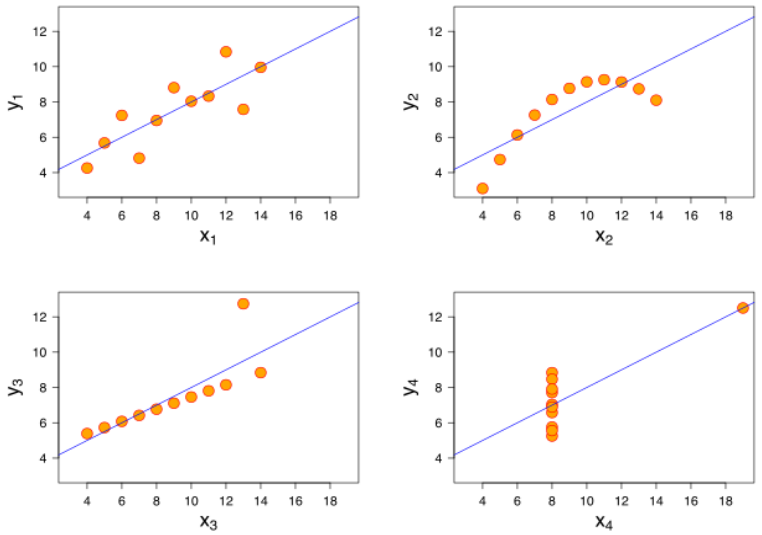
     2. Multiple Linear Regression:
        This regression method takes more than one independent variable and one dependent variable i.e. X1, X2,..,Xn and Y.
        The equation for Multiple Linear regression is,

        $y = \beta 0 + \beta 1X + \beta 2X + \ldots + \beta nX$

        where,
        X1, X2, …, Xn are the independent variables
        Y is the dependent variable
        β0 is the intercept
        β1, β2, …, βn are the slopes

   - The objective of using linear regression is to find the best-fit line, which implies that the error between the predicted and actual values should be minimal.
   - The best Fit Line provides a straight line that represents the linear relationship between the dependent and independent variables.

2. Explain the Anscombe's quartet in detail. (3 marks)
   - Anscombe's quartet is a set of four datasets that have identical statistical properties such as mean, variance, correlation, and regression line.
   - In the below example, the datasets are identical. However, when the data is plotted on a graph, it is clear that the datasets are very different. This demonstrates the importance of looking at data graphically, and not just relying on statistical properties.



   - This shows that it is important to look at data statistically and graphically, rather than only checking statistical properties. A graph can show you the data distribution, and any outliers or other influential observations.

3. What is Pearson's R?
   Pearson's correlation coefficient (r) is a statistical formula that measures the linear correlation between two variables. It describes the strength and direction of the linear relationship between two quantitative variables.
   Pearson's r is a number between –1 and 1.
   If r=0, indicates no association between the two variables.
   If r>0, it indicates a positive association between the two variables.
   If r<0, it indicates a negative correlation between the two variables.

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

   - Scaling is a technique that normalizes the range of features in a dataset. It is performed during the data pre-processing to handle highly varying values or units to increase the algorithm's effectiveness and model processing.
   - Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalization.

- When the features are scaled, several machine learning methods perform better or converge more quickly. The algorithm's performance can be enhanced by scaling the features.
- Scaling features ensure that each features are given the same consideration during the process. Without scaling, features with high scales will increase weightage in the learning, producing skewed outcomes. This bias is removed through scaling, which ensures that each feature contributes equally to model predictions.

- The difference between normalized scaling and standardized scaling is,

1. Standardized scaling:
   - In the Standardisation scaling, all of the data is converted into a standard normal distribution with mean zero and standard deviation one.
   - The equation for Standardisation,

$$x = \frac{x - mean(x)}{sd(x)}$$

2. Normalized or MinMax Scaling:
   - In Normalization scaling, the data is scaled in the range of 0 and 1.
   - The equation for Standardisation,

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.
- Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.
- The formula for VIF is:

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

  where, R^2 is the R-squared value of the model.

- VIF conditions:
  - VIF = 1: Variables are not correlated and multicollinearity does not exist.
  - VIF between 1 to 5: Variables are moderately correlated. It can also be called a good VIF value.
  - VIF > 5: Variables are highly correlated and there is a high possibility of multicollinearity existing.

- When the R-squared score is 1, the VIF values become Infinite. It means that the independent variables are highly correlated with each other and there is a perfect correlation.

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

- A Q-Q plot, or Quantile-Quantile plot, is a graphical method for comparing two probability distributions. It plots the quantiles of one data set against the quantiles of another data set.
- It helps determine if a dataset follows a particular type of probability distribution, such as normal, uniform, or exponential. If both sets of quantiles came from the same distribution, the points should form a roughly straight line.
- It is a scatterplot created by plotting two sets of quantiles against one another.
- A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed.
- Q-Q plots are graphical tools that help you assess the validity of some assumptions in regression models, such as normality, linearity, and homoscedasticity.