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?

➔ The impact of each level of the categorical variables can be inferred when we encode them and use it in our model.

In our model, all the levels of season and weathersit variables are statistically significant. Ex, we saw by how much the rain affects the bike rental demand.

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

➔ When we use drop_first, we are reducing one additional column which is anyway explained by other levels of the categorical variable. Thereby reducing the correlations amongst the levels, so that they become linearly independent.

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

➔ The 'temp' variable has the highest correlation of 0.63 with the target variable and the strongest linear relationship.

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

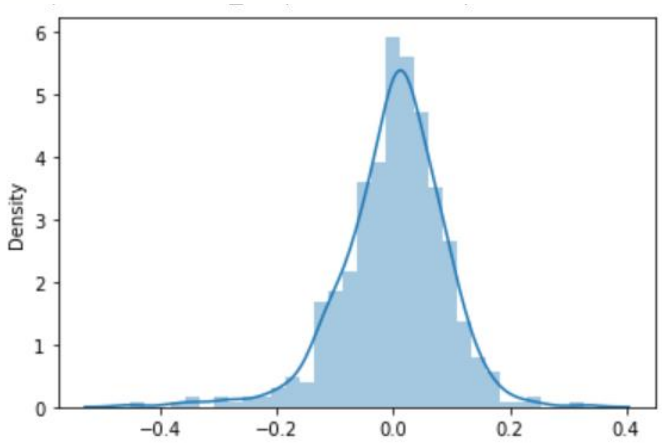
➔ One of the major fundamental assumptions is that the 'Distribution of the error terms (or residuals) are normally distributed which is centered around the mean 0'.

This can be validated by plotting a histogram of residuals, which is difference between y_train and y_train_pred.

```
y_train_pred = lr6.predict(X_train_sm6) #predicts y values
for train dataset for given X values

res = y_train - y_train_pred

#plotting the residuals pattern
sns.distplot(res)
```



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

➔ Top 3 features would be,

- Temp
- Light rain/snow level of weathersit (value 3 in original dataset)
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

➔ Regression is a supervised learning technique where output variable to be predicted is a continuous variable. The output variable, also known as dependent or target variable(y), shows linear relationship with one or more independent (X) variables, hence called as linear regression.

Linear regression performs the task to predict a dependent variable value (y) based on given independent variable (X). So, this regression technique finds out a linear relationship between X(input) and y(output).

Our goal is to find the best fit line by training the model with significant variables.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

where,

y= Dependent Variable.

x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient.

- We use Ordinary Least Squares (OLS) method to obtain the best-fit line, as in obtain the best coefficients.
- The error between this predicted y value and the actual y value is called error term or residual.
- Our goal now is to minimize this error to best predict the output. We do this by using Gradient Descent algorithm. The idea is to start with random a0 and a1 values and then iteratively updating the values, reaching minimum cost.

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

2. Explain the Anscombe's quartet in detail.

➔ Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

These datapoints are analyzed using only descriptive statistics to find the mean, standard deviation, and correlation between x and y.

3. What is Pearson's R?

➔ Pearson's correlation (also known as Pearson's R) is used to measure how strong a relationship is between two variables in linear regression.

The correlation coefficient formula returns a value between 1 and -1. Here,

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

Formula is given by, $R = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$

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

➔ Feature Scaling is a technique to standardize the independent features present in the dataset 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 as higher and consider smaller values as lower values, regardless of the unit of the values.

Example,

If an algorithm is not using the feature scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's not true and, in this case, the algorithm will give wrong predictions.

So, we use Feature Scaling to bring all values of multiple features to the same magnitudes and thus, tackle this issue.

Scaling is performed for,

1. **Ease of interpretation**
2. **Faster convergence of gradient descent method**

Types of Scaling methods,

1. **Min-Max normalization**

   In statistics, normalization is the method of rescaling data where we try to fit all the data points between the range of 0 to 1 so that the data points can become closer to each other.

In this method of scaling the data, the minimum value of any feature gets converted into 0 and the maximum value of the feature gets converted into 1.

We can represent the normalization as follows,

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

2. Standardization

The basic concept behind the standardization function is to make data points centered about the mean of all the data points presented in a feature with a unit standard deviation. This means the mean of the data point will be zero and the standard deviation will be 1.

So, in standardization, the data points are rescaled by ensuring that after scaling they will be in a curve shape. Mathematically we can represent it as follow,

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

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

➔ VIF is infinite when there is a perfect correlation between 2 independent variables.

➔ We get R2 =1, which lead to 1/(1-R2) infinity.

➔ To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

➔ 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. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

➔ If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

➔ For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve.