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?

   Answer:

   Based on the below observations for every unit increase in Y(Cnt) the variables will increase/decrease by

Variable [X]	**Coefficient [B_0]**
Winter	0.0964
Monday	0.0616
Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	-0.2471

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

   Answer:

   Suppose if there are n levels in a variable, then it is enough to have (n-1) columns for dummy variable creation to save space and time

   Example: If the column has yes/no/maybe then it is enough to have “no” and “maybe”. If both are 0, then it is evident that “yes” is 1

   So that is why drop_first = True is important to skip first column.

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

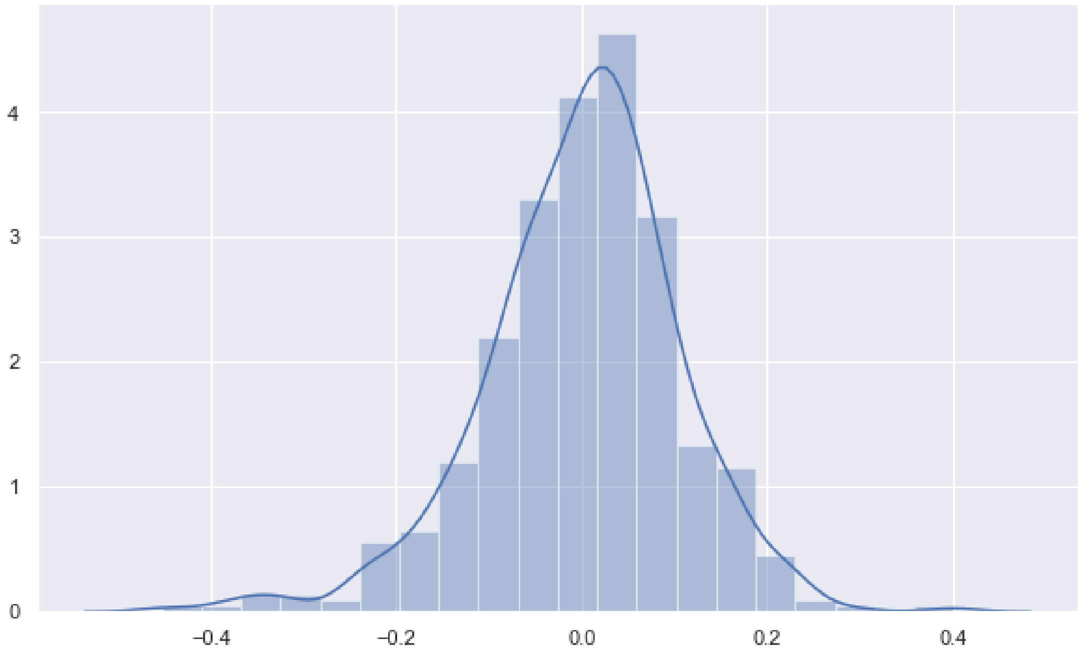
   Answer:

   atemp has the highest correlation with the target variable. The value is **0.65**.

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

   Answer:

   I have checked the distplot for y_train and y_train_pred to check if the plot is normally distributed with mean = 0



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

   Answer:

   The top 3 features are

Feature	**Coefficient [B_0]**
Temperature [atemp]	0.6357
Year [yr]	0.2341
Weather situation [weathersit]	-0.2471

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression algorithm is used to train a model based on the existing data.

This model then can be used for future predictions.

Suppose we want to predict sales based on marketing.

There are 2 variables

a. Sales [Y: Target variable] – Dependent Variable
b. Marketing Budget [X] – Independent Variable

Sale is a dependent variable since the value of sales change based on the marketing budget

So we can plot a scatter plot for the above scenario and check if both variables are linearly dependent.

Which means if the increase in Marketing Budget also increase Sales.

If there is a linear relationship, then we can do linear regression.

Now we must Fit a line which passes through the scatterplot.

Since this a straight line we know the straight line equation is

$Y = B_1 + B_0.X$

Where X and Y are our variables

B_1 is the y-intercept

B_0 is the slope B_1

Now we need to find B_0 and B_1 to get the best fit line

Ordinary Least Squares:

In order to get the best fit line we should minimise the difference between actual y value and y predicted value

We can do that by minimizing Residual Sum of Squares [RSS]

$RSS = \sum(Y_i - Y_{pred})^2$

Cost Function:

For linear regression the cost function is achieved by minimizing the RSS value

Gradient Descent:

In order to get B_0 and B_1 we need to start with random B_0 and B_1 values and then iteratively updating the values, reaching the minimum cost

2. Explain the Anscombe’s quartet in detail.

   Answer:

   Suppose we got same mean, variance and correlation coefficients values for different datasets.

   When plotting the graph the data might look very different when it comes to curvature and outliers.

   This is called Anscombe’s quartet.

   So it is important to check graphs to visualize rather than just go by statistical numbers.

3. What is Pearson’s R?

   Answer:

   Pearsons’s R is a correlation coefficient commonly used in linear regression.

   Correlation Coefficients are used to measure how strong a relationship between two variables

   The measure can lie between -1 and 1.

   -1 mean strong negative correlation

   0 means no correlation

   1 means strong positive correlation

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

   Answer:

   What is Scaling?

   Scaling is a technique to standardize the independent features present in the data in a fixed range.

   Why Scaling?

   Suppose the values of one variable is in 1000’s and others in 10’s. Then in this case it wont be accurate to measure Linear Regression using these two variables. So we have to bring them to the same scale to measure. So we have to use **Scaling**.

   Difference between normalized and standardized scaling

   Normalization

   The values are rescaled into a range of [0, 1]

   Standardization

   Rescaling data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

   Answer:

   VIF becomes infinite when R^2 becomes 1

   which means, VIF = 1/1- R^2 becomes infinite

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

   Answer:

   Definition: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

   Use and importance: Q-Q plot helps us to check

   a. Populations came from common distribution
   b. Have common scale