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 categorical variable in the dataset were season, weathersit, holiday, mnth, yr and weekday.

   These features were visualized using boxpot

   1. Season - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt.
   2. Weathersit -Demand was very low when it was light_rain, light_snow whereas demand increases when it is clear or few clouds demand was moderate in mist_cloudy weather.
   3. Holiday - Demand reduces on holiday.
   4. Mnth - September has highest no of demands while during initial 3 months demand was low and during the last 2 months of November and December demand was moderate, rest of the year it as almost high
   5. Yr - The number of demands in 2019 was more than 2018

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

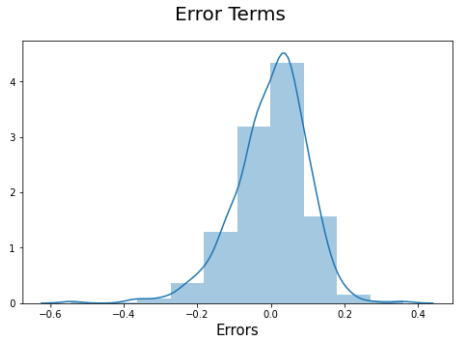
   To explain/show k different categories k-1 columns are sufficient hence to reduce redundancy we drop first column.
   Another reason is to keep multicollinearity under control.

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

   'temp' and 'atemp' are the two numerical variables which are highly correlated with the target variable 'cnt'

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

   We validated the assumption by plotting the distplot of error terms and check if the error terms are following a normal distribution.



We validated another assumption of multicollinearity by ensuring that the VIF for all the features is less than 5.

	Features	VIF
2	windspeed	3.72
3	spring	3.64
12	workingday	2.73
6	jan	2.14
11	feb	1.85
0	yr	1.82
4	summer	1.76
10	mist-cloudy	1.48
7	nov	1.28
5	dec	1.20
8	sept	1.19
1	holiday	1.11
9	light_snow-light_rain-thunderstorm	1.07

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

The top3 features are:

light_snow-light_rain-thunderstorm : -0.315603
yr : 0.245175
spring : -0.193847

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a supervised Machine Learning algorithm that is used in prediction of numeric values. Linear Regression is the most basic form of regression analysis.
Regression is the most commonly used predictive analysis model.
Linear regression is based on the equation of straight line y = mx + c.
It assumes that there is a linear relationship between the dependent variable(y) and the independent variable(x). Linear Regression is performed when the dependent variable is of continuous in nature and independent variables could be continuous, categorical in nature.
Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.
Linear Regression is divided into simple linear regression and multiple linear regression.
1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.
2. Multiple Linear Regression : MLR is used when the dependent variable is predicted using multiple independent variables.

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.
They have very different distributions and appear differently when plotted on scatter plots.
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
This tells us about the importance of visualizing the data before applying various algorithms out there to build models.

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

Pearson's R is a measure of the strength of the linear association between the variables. It ranges from -1 to +1. It shows the linear relationship between two sets of data
1 indicates strong positive relationship
-1 indicates strong negative relationship
0 indicates no relationship at all

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where
x = value of x-variable
x bar = mean of x-variables
y = value of x-variable
y bar = mean of y-variables

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

   Scaling is performed to standardize or normalize the values of independent variables in the dataset.
   Scaling is performed during data preparation step before splitting dataset into train and test.
   Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. After normalizing the feature min value we get is 0 and max value is 1. Hence it is also called MinMax Scaling.
   Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. Unlike Normalization, Standardization does not have any boundary limits.

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

   A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

   I there is a perfect correlation then VIF = infinity as R-square will be 1, therefore

   1/1-1= 1/0 = infinity

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

   A Q–Q plot is a plot of the quantiles of two distributions against each other. The pattern of points in the plot is used to compare the two distributions.
   If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
   The q-q plot is used to answer the following questions:
   Do two data sets come from populations with a common distribution?
   Do two data sets have common location and scale?
   Do two data sets have similar distributional shapes?
   Do two data sets have similar tail behavior?