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:
   - Count of bikes rented is maximum in the fall season, followed by summer and winter, with spring being the season with the least rented bikes for 2018 and 2019.
   - In 2019 the maximum number of bikes were rented in September followed by August, whereas in 2018 the maximum number of bikes were rented in June, followed by July. The last bikes were rented in the month of January and February for both 2018 and 2019.
   - In 2019, the maximum number of bikes were rented on Fridays and the least on Tuesdays. Whereas, in 2018, the maximum number of bikes were rented on Sundays, and the least number of bikes were rented on Fridays.
   - Bike rentals are maximum in case of clear to partly cloudy days followed by misty to cloudy days with a huge difference. Rentals are the least on days of light snow rain to scattered clouds.
   - Bikes demand increased significantly from 2018 to 2019, with an approximate increase of 2205 more bikes rented in 2019 compared to 2018.
   - Lesser bikes are rented on holidays, whereas it is very high on normal days.
   - On non-working days slightly more bikes were rented than on working days.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans:
   In preparation for dummy variables, we get exactly the number of columns equal to the number of categories of a specific variable, for which the dummy variables are made. Now, we can work with 1-number_of_categories_of_a_variable for better feature selection and the least correlation between features for building an efficient linear regression model.
   So, we drop the first column using **drop_first=True**, which helps us drop the first column of the dummies created and finally reduce the number of columns helping in better feature selection, and also helping us understand each distinct category using binary 1 or 0 values.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: ‘temp’ has the 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?
   Ans:
   - Checking the distribution of error terms, where we saw that the distribution was normal, and the mean was 0.

- Testing for Homoscedasticity, where the plot between y_train and y_train_pred shows an almost constant variance of predictions
- Testing correlation between error terms, where we see that there is no correlation between the error terms.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans:
   The top 3 features are:
   - yr
   - holiday
   - temp

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans:
   Linear regression is a supervised form of machine learning, which we use to build a model based on historical data, to predict a dependent feature of the testing data.
   When the linear regression model is made of one independent and one dependent variable/feature, it is called a simple linear regression model, whereas, when multiple independent variables are used to make a model, to predict the target variable, it is called a multiple linear regression model.

   The equation of a simple linear model is :
   $y = \beta 0 + \beta 1 * x$
   Here,
   - y is the target variable.
   - x is the predictor variable.
   - B0 is the intercept.
   - B1 is the slope of the best-fitting line or the coefficient of the predictor variable.

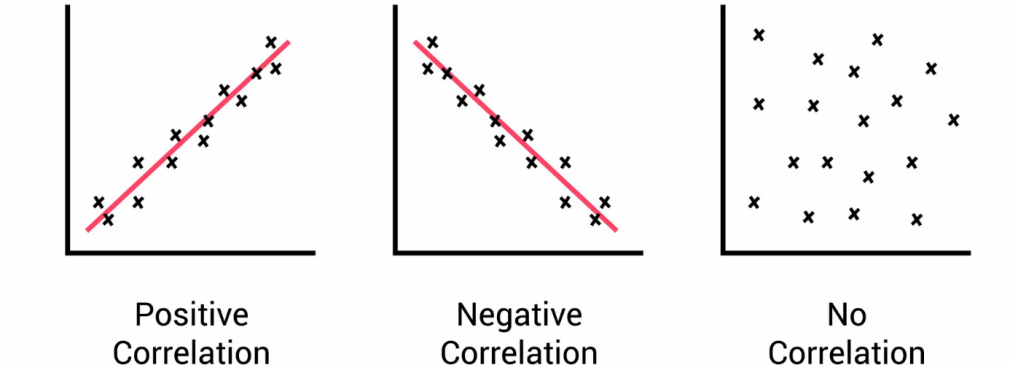
   The equation of a multiple linear regression model is :
   $y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \beta 3 * x3 + \beta 4 * x4 + \ldots + \beta n * xn$
   Here, there are n number of independent or predictor variables, that are used to predict the values of the target y variable.

   First, we start by splitting the entire dataset into train and test datasets.
   In case of more than one feature being used in the linear regression model building, we select the variables/features that can be used to predict the values of the dependent variable and we train our model accordingly.

The model's performance is evaluated using a cost function, typically Mean Squared Error (MSE). This measures the difference between the predicted values and the actual values. Reduction of this cost function helps us to find the best-fit line for the model.
Then we estimate the coefficients β0 and β1 that give us the best-fit line.
When we are done with finding the coefficients and building the model, we use this model to make predictions on the target variable.
The performance of this linear model is then evaluated using r-squared metrics, to understand how efficient the model is, and how much variance the model can explain.

2. Explain the Anscombe's quartet in detail.
   Ans:
   Anscombe's quartet was introduced by the statistician Francis Anscombe in 1973, to emphasize the importance of graphically representing and exploring data before making final conclusions based on the summary statistics.
   Anscombe's quartet talk about a set of four datasets where each of these datasets have almost similar to identical statistics properties, such as - mean, variance, correlation coefficient and linear regression coefficient. But, though they have the same statistical properties, these datasets have very different patters and relationships when they are plotted and visualized over a graph to draw conclusions. This highlights the importance of data visualization, to explore the data in a vast way in order of understanding the underlying structure of data.

3. What is Pearson's R?
   Ans:
   Pearson's R, also called Pearson's coefficient, is the numerical summary of the strength of the linear association between the variables.
   If the variables tend to go up or down together, then the correlation between the variables will be positive. If the variables tend to go up or down but in an opposing manner, i.e., for up of one variable, the one goes down, then there is a negative correlation between the variables.
   The correlation coefficient can take a value from -1 to +1. A positive value denotes a positive correlation, whereas a negative value denotes a negative correlation. A 0 correlation value means there is no association between the two variables.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans:
   The transformation of numerical features of a dataset into a standard range or scale is called scaling.
   Scaling is performed to make sure that all the features of the dataset are on the same scale so that every feature used for building a model shares the same magnitude. It is important to perform feature scaling because without scaling the machine-learning algorithm tends to weigh greater values as higher, and lower values as smaller.

Normalized scaling	**Standardized scaling**
Also known as Min-Max scaling.	Also known as z-score normalization or standardization.
Scales the features to a fixed range, typically between 0 and 1.	Scales the features to have zero mean and unit variance.
Affected by outliers.	Less affected by outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans:
   The Variance Inflation Factor (VIF) is used to understand the correlation between the independent variables. A large VIF value indicates that a specific variable has a high correlation with other variables, similarly, a low VIF means the opposite of it. We consider the significance of a variable based on if the VIF of the variable is below 5.
   But when the value of VIF is infinite, it means a perfect correlation between variables. To handle this, we need to drop the specific variable with such a VIF value and check for the VIF for the existing variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans:
   A Q-Q plot, also called a Quantile-Quantile plot, is a graphical tool that is used to understand if a dataset follows a particular theoretical distribution. It compares the quantiles of a dataset with the quantiles of a known theoretical distribution. By plotting this data we can understand if the data is distributed normally or if there is any deviation from the expected distribution.