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?**

➔ Inference drawn from analysing the categorical variables are:

(1) Season vs Count :Most bikes are being rented at fall season and it is consistent for every year.

(2) Month vs Count: Most bikes are being rented on the month of September thus confirming our observation for the season variable that rentals increase at Fall and the bike rentals dip on the month of July suggesting that bike rentals are less in monsoon.

(3) Weather Situation vs Count: Most number of bikes are rented on a clear day and there is a huge dip on bike rentals for a drizzling or overcast day suggesting that bike rentals are heavily effected because of bad weather.

(4) Weekday vs Count: Most number of bikes are being rented on a Saturday.

(5) Holiday vs Count: More bikes are rented on a working day rather than a holiday.

(6) Workingday vs Count: More bikes are rented on a working day rather than a holiday.

(7) Year vs Count: The number of bike rentals is increasing every year as we can see that the number of bike rentals has significantly increased in 2019 from 2018.

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

➔ 'drop_first=True' is a parameter used with pandas.getdummies() which is used to create dummy variable from a variable to encode it in a proper way. Here 'drop_first=True' is used to get n-1 dummies for n dummy variables by removing the 1st variable. If we do not use 'drop_first=True' then it will lead to Dummy variable trap which means n dummy variables for n variables will be created which themselves are collinear thus turning into a fallacy.

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

➔ By observing the pair-plot among the numerical variables we can conclude that temp has the highest correlation with the target variable. As its scatterplot is showing a linear spread among the two variables.

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

➔ I validated the linear regression model by checking the following things:

(1) Normality of Error terms:Checking if the error follows a normal distribution or not by plotting a distplot of error terms.

(2) Multi Collinearity:Checking if the variables are strongly correlated by plotting a heatmap of correlation coefficients and checking the VIF value of each variables.

(3) Linearity:Checking the linearity of Residual vs Residual +component by plotting a CCPR plot.

(4) Homoscedasticity:Checking if the scatterplot of predicted values vs residual shows any clear pattern.

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

➔ Based on the final Linear regression model 'temp','september','yr', these three variables appear to have contributed the most on the demand of shared bikes positively.

General Subjective Questions

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

➔ Linear Regression is defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change,the value of dependent variable will also change. Mathematically it is represented with the straight line equation

$$Y=mX+C$$

Where Y is the dependent variable and X is the independent variable,m is the slope of the regression line and c is a constant known as intercept. We make some assumptions about the dataset for the Linear Regression Model:

1) Multi Collinearity: The Linear Regression model assumes that there is almost no multi collinearity between the variables.
2) Auto-Correlation:Linear regression Model assumes that there is almost no auto correlation between the variables.
3) Linearity: Linear Regression Model assumes that there is a linear relationship between response and feature variables.

4) Normality of Error terms:Linear Regression Model assumes that the error terms would be normally distributed.
5) Homoscedasticity:There should be no visible pattern in residual values.

2) Explain the Anscombe's quartet in detail.

➔ Anscombe's Quartet was developed by statistician Francis Anscombe.It comprises four datasets,each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics.But things change completely,when they are plotted. The summary statistics show that the mean and variance of x and y are identical across all groups.i.e, when we plot the regression line for each group we see the scatter pattern is completely different even though the regression line is being fitted.

3) What is Pearson's R?

➔ Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together,the correlation coefficient will be positive. If the variables tend to go up and down in opposition to each other the correlation coefficient will be negative. The Pearson correlation coefficient 'R' can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association and a value less than 0 indicates a negative association.

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 data 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,higher and consider smaller values as lower regardless of their unit.

The key differences between normalized and standardized scaling are:

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

➔ When the value of VIF is infinite it shows a perfect correlation between two independent variables.In the case of perfect correlation,we get R-squared R2=1,which in turn leads the 1/1-R2 value to infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

➔ The quantile-quantile plot is a graphical technique for determining if two datasets come from populations with a common distribution.
The Q-Q plot is a plot of the quantiles of the first dataset against the quantiles of the second dataset. Quantile means the fraction of the points below the given value.i.e, 0.3 quantile is the point at which the 30% of the data fall below and 70% fall above that value. A 45 degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line,the greater the evidence for the conclusion that the two datasets have come from populations with different distributions.