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? (3 marks)
   Ans: Observations made after analysing the relationship of the categorical variables with the target variables are as follows:
   - Number of rentals were more in 2019 than 2018
   - Rentals are less on holidays ('holiday' variable)
   - Rentals are increasing from Jan-June, with maximum in Sep and the demand drops after that
   - Number of rentals are more in clear weather and are least in snowy conditions
   - Rentals are max during the Fall season, followed by summer
   - Weekday is inconclusive for rental demand

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Ans: Using drop_first=True is needed to get n-1 dummy variables when there are n categories in a variable. It is important to drop one level as n-1 dummy variables can fully capture the whole information.
   For example, if we have 4 categories (Spring, summer, fall and winter) in the 'season' variable, we can capture the data using just 3 dummy variables. If value of summer is 0, fall is 0 and winter is 0, then value of spring will always be 1 and it can be conveyed with just the first 3 dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: Looking at the pairplot, the 'temp variable 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? (3 marks)
   Ans: Following steps were followed
   - For the little to no multicollinearity assumption, VIF values were calculated and variables with VIF>5 were dropped
   - For the linear relation assumption, a heatmap of correlation matrix was plotted and correlation values were checked. Variables like 'temp', 'atemp'have a strong correlation to the target variable 'cnt'
   - For the homoscedasticity assumption, scatter plot was built for y-test vs test_residual as well as y-train vs train_residual, for both the plots the error terms are not varying much, thus validating the assumption
   - The residuals should be normally distributed which is validated by plotting a histogram on the residual on the train dataset

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Ans: The top 3 features significant variables are – 'temp', 'yr' and 'Weather Light Snow with coefficients 0.4237, 0.2646 and -0.2687 respectively.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans Linear Regression is a supervised machine learning algorithm which performs a regression task. The predictions obtained from linear regression are continuous values such as predicting the price of a house. These predictions are obtained with the help of independent variables which have a statistically significant linear relationship with the target variable (variable whose values we are trying to predict) and are able to explain the variance in the values of the target variables.
   The linear regression algorithm tries to reduce the errors in predictions using the Ordinary Least Squares (OLS) method. Key assumptions that should be satisfied in order to be able to implement the linear regression algorithm are
      1. Linear relationship between independent and dependent variables
      2. Multivariate normality for all variables
      3. Little to no multicollinearity in the data
      4. Little or no autocorrelation in the data
      5. Homoscedasticity in the data

   If a single independent variable is used to predict the dependent variable, it's called Simple Linear Regression and if more than one independent variables are used, it's called Multiple Linear Regression.
   Equation of Simple Linear Regression - y = B0 + B1x1 (where y - dependent variable, B0 - intercept, B1 - coefficient, x1 - independent variable)
   Equation of Multiple Linear Regression - y = B0 + B1x1 + B2x2 + .. + Bnxn (where y - dependent variable, B0 - intercept, B1 - coefficient of x1, x1 - first independent variables)

2. Explain the Anscombe's quartet in detail. (3 marks)
   Ans The experiment with Anscombe's quartet highlights the importance of data visualisation. It contains 4 datasets with nearly similar summary statistics, but appear very different when plotted on a graph. It was developed to demonstrate both the importance of graphing data before analyzing it and the impact of outliers on the statistical summary of a dataset.

3. What is Pearson's R? (3 marks)
   Ans Pearson's R is commonly used in Linear Regression to find out the correlation among pairs of variables.
   The values of Pearson's R range from -1 to 1, with: -1 indicating strong negative linear relation, 0 indicating no relationship 1 indicating strong positive linear relation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Ans Scaling is a technique used to standardize the independent variables and convert them into comparable values or bring them all into a fixed range in order to improve the interpretability of a model. It is important in case variables have units and magnitudes with huge differences. Since the ML algorithm won't understand the units of an independent variable and might assign more weight to greater values, it is important to scale them.
   Normalized scaling - this technique rescales the values of a variable to range between 0 and 1 Standardized scaling - this technique rescales the values to have mean = 0 and standard deviation = 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans If the value of VIF for an independent variable is infinity, i.e. R squared = 1 (VIF = 1/(1-R2)), it means that the variable is highly collinear with other independent variables and we should definitely drop this variable as it will hamper the interpretability of the model results.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans Q-Q plot is a plot of the quantiles of one dataset against the quantiles of another dataset, used to understand if two datasets came from populations with the same distribution such as Normal distribution, Uniform etc. If all points of quantiles lie on or close to straight line at a 45 degree angle from x-axis, the datasets are said to have the same distribution. Otherwise, if all points of quantiles lie away from the straight line at an angle of 45 degree from x-axis, they are said to have a different distribution. In case of linear regression, say we get the training data and test data separately, we can make Q-Q plots to understand if both are from populations with same distributions.