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 final equation given by the model is

cnt= 0.235*yr-0.0862 * holiday+0.4758 * temp - 0.1325 * windspeed - 0.1032 * Spring+0.0504 * Winter-0.0616 * July+0.0498 * September-0.2562 * Light Snow and Rain

As per the analysis carried out on the data set, following categorical variables impact the bike rentals for the company.

- Season – There is a negative impact of the Spring season on the bike rental count, having coefficient value as 0.1032, while there is a positive impact due to Winter season
- Month – There is a negative correlation with respect to July (0.062) and positive correlation with September (0.05)
- Day – no impact seen in the final model
- Weather Situation – Light rain and Snow have a negative impact on the bike rentals (with coefficient value as 0. 2562)

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

When we do the dummy encoding of the column having categorical values, one of the groups can be represented in the form of zeros and act as a reference for the other category values and if it not dropped it is redundant in our encoded values, hence we need n-1 dummy columns for a given n values in the category column. The option Drop_first=True helps us to drop the first level and giving us the n-1 values for the given categorical feature.

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

Temperature(temp) is the variable having the high correlation value of 0.64 with target variable.

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

The assumptions of the linear Regression were validated after plotting a scatter plot with the fitted regression line and carrying out the Residual analysis of the train data. In the residual analysis we check if the error terms are also normally distributed (which is actually, one of the major assumptions of linear regression). This is done by plotting the histogram on the training data actual values and predicted values from model.

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 model and the coefficient values for the variables, below are the top 3 features contributing significantly towards the demand of shared bikes:

Variable	Coefficient Value
temp	0. 4758
Weather Situation - Light snow and Rain	-0. 2562
Year (yr)	0. 235

The final equation given by the model is

cnt= 0.235*yr-0.0862 * holiday+0.4758 * temp - 0.1325 * windspeed - 0.1032 * Spring+0.0504 * Winter-0.0616 * July+0.0498 * September-0.2562 * Light Snow and Rain

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the regression analysis which is used for the prediction\analysis of the target variable with respect to the independent variables available in the data set and there is a linear relationship between the dependent and independent variables. In linear regression based on the given set of data we plot the best fit line for the data points and equation of the lines is denoted using the linear equation, as mentioned below:

$y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 +$.. and so on

Here the target variable has the linear relationship on the variables X1,X3,X3. The coefficients in the above equation is find out by minimizing the cost function, using methods like differentiation or Gradient descent method.

Based on the number of variables the Linear Regression models can be classified into:
Simple Linear Regression – 1 independent variable
Multiple Linear Regression – No of independent variables are more than 1

The strength of a simple linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$, while in case of Multiple Linear regression there are few considerations which needs to be kept in mind, for instance – Multicollinearity, where the independent variables may have high correlation with other independent variables and strength of the model is measured using Adjusted R^2.

Adjusted $R2 = 1 - (1 - R2)(N - 1)N - p - 1$

Usage - Linear Regression analysis can be used to evaluate trends or make estimates and forecasts.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is defined as a group of four data sets having identical simple descriptive statistics but have a very different distributions and these appear differently when plotted using scatter plots. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of graphing data before carrying out the analysis. There are four data set plots which have nearly same statistical observations, involving **variance**, and **mean** of all x,y points in all four datasets.
This tells us that it is important to plot the data set to see the various anomalies and this should be done before applying any model on the dataset. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Reference - https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. What is Pearson's R?

Pearson's R is the statistic that measures the linear correlation between the two given variables and has the value range between +1 and -1 . The value of +1 denotes the positive correlation, the value of 0 means no correlation, whereas the value of -1 denotes negative linear correlation between the variables. It also indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Pearson correlation coefficient is the covariance of the two variables divided by the product of t heir standard deviations.

Reference - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

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

Scaling is the process of reducing the range and bring all the different features present in the dataset to same scale. This is carried out to ensure that our model\algorithm is not biased towards one of the features having high magnitude of the scale. The regression models which uses Gradient descent algorithm as an optimization technique requires the data to be scaled, scaling ensures the step size used in the gradient descent method is small and consistent, thereby ensuring the method moves smoothly towards the minima.

There are two ways to scale the features present in the data:

1) Normalization
2) Standardization

Normalization	**Standardization**
Technique where the values are scaled in such a way that they are between 0 and 1	Technique where the values are centered around the mean with a unit standard deviation.
Scaled values are between 0 and 1	There is no particular range on the values
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

Reference - https://en.wikipedia.org/wiki/Feature_scaling

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. It is calculated using the below formula:

$$VIF = 1/1 - R^2$$

VIF calculates how well one independent variable is explained by all the other independent variables combined.
Now if there is perfect correlation between the variables i.e R square will have the value as 1, the value for VIF will be 1/(1-1) i.e infinity. This will only happen when there is perfect correlation between two independent features and in order to resolve this we need to drop the feature causing this perfect correlation.

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

Quantile-Quantile plot is a probability plot, in which uses graphical representation for comparing the two probability distributions by plotting the quantities against each other. The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted.
The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
It can be used with the sample sizes and detect the presence of the outliers in the data.

Reference - https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot