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:

Below mentioned are the categorical Variables and the effect on count:

1. Season – Fall season got maximum active.
2. yr – 2019 has more count than 2018.
3. mnth – September has more count.
4. holiday – Counts drop on holidays
5. weekday – Counts increase on weekdays
6. weathersit – count decreases during rain (partly cloudy). Clear sky has more counts.

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

Ans: After removing drop_first = True below is the observation: Dummy variables became correlated to each other. This was disrupting the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: temp and atemp has the highest correlation with the target variable. Hence atemp was dropped.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Error Term should form a normal curve on histogram. This is the sign to validate the assumptions of linear regression model. Same can be observed on Cell no 74.

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: Below are the top influencers and the direction of influence

1. temp – Positive Influence
2. Light rain_Light snow_Thunderstorm – Negative Influence
3. Yr – Positive Influence

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s), is linear regression.
Here are the steps for linear regression:

Load data: By using pandas we load raw dataset into pandas data frame for further processing

Cleaning data: This involves trimming the columns, taking care of null and duplicate rows.

EDA: This involves exploring data by visualization tool and grabbing insights from it.

Train and test split: We split data into two groups in the ratio of 70:30.

RFE & VIF and R value fixation: To get model into good shape we need to preserve r score and high as possible and VIF should be less than 5. This is a reiterative step.

Check Error curve: This should be normally distributed.

Match R scores for test and train. Both should be approximately equal.

Derive the final equation and predict more data using trained model.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

A regression model can be fooled by using (smartly) arranged data. In some cases of multiple datasets which are different but after training the regression model appears same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is Anscombe's Quartet.

3. What is Pearson's R? (3 marks)

Ans: Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is used in linear regression. The Pearson's R value always lie between -1 and +1. The latter indicating a perfectly positive and linear correlation and former indicating perfectly linear negative regression. The values laying in between represent the relative collinearity of two variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is necessary for a model to be functional with the appropriate range of coefficients. E.g. If there were two independent variables names price and months on which the sale of car depended the price range would be far too high because there are only 12 months in a year. In such case scaling the variable price appropriately won't allow decimal errors in the model.

Below are the two types of Scaling:

- Normalized scaling: Values are rearranged in the scale of 0 to 1. Also known as min-max scaling. Typically used in Neural networks broadly.
- Standardized scaling: In this scaling values are rearrange in positive and negative range in such a way mean will be approximately or close to 0.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: In case of perfect correlation between dependent variable and independent variable the R squared value comes out to be 1. Hence VIF, turn to become infinity. (VIF = (1/(1-R^2)).

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 graphical tool to assess if sets of data come from the same statistical distribution. It is helpful in regression when testing and training datasets differently. It is important to check whether both data come from the same background to maintain the sanity of model.