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)**
   - ➔ windspeed and humidity doesn’t seems to have linear dependency with the count
   - ➔ temp and atemp has almost similar impact to count
   - ➔ weathersit 1(Clear, Few clouds, Partly cloudy, Partly cloudy) has more preference/denser distribution to contribute toward count
   - ➔ humidity,windspeed,weekday,season/month these data seems to be less contributing to the target output(count)
   - ➔ rising temperature and season type 3 (oct,nov,sept) seems to have more impact on count.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - ➔ To reduce the complexity of the dataset without losing information and also avoid the unnecessary computation.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   - ➔ “ temp and atemp“ given that we did not considered casual and registered column for our analysis otherwise registered has the highest.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - ➔ Check the normality of the error
   - ➔ check the heteroskedasticity
   - ➔ check the dependence among the output variable
   - ➔ verifying the multicollinearity among the selected final feature.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   - ➔ temperature(positive corr.),year(2019 has more sales than 2018),season/month(oct,nov,sept)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - ➔ This is the supervised learning algorithm which is used to predict the output variable also called as dependent variable based on the N numbers of input variable also called as independent variable.
   - ➔ Here the algorithm tries to derive the best fit line which will pass through the data point plotted using the different independent variables’ value of the training set.

   The process of deriving the best fit line is:
   1. initialize the random line with the equation y=C+ B1X1 + B2X2+…. BnXn
   2. select the cost function to analyze the error

c. use the optimizer like Gradient Descent to minimize the cost value in each iteration by deriving new coefficient value (C,B1,B2,…Bn) in each iteration of optimization

d. Finalize the coefficients which gives the lowest MSE or RMSE ,etc

various types of linear regression algorithm as well like simple linear regression and multiple Linear regression

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

- ➔ This is the plotting for the 4 dataset whose main intention is to show that even the description statistics like mean, SD, Correlation, Best fit line equation, etc of the data seems to be similar we might get different trends in the data if we plot the data visually.
- ➔ The main intension here is to show why its always better to plot the data rather than only relying on the descriptive value because sometimes those value might have been impacted by the outliers and other data property.
- ➔ It tries to explain its always see the validity of the derived model graphically rather than relying on the numbers only.

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

- ➔ This is the correlation coefficient index which is used to explain the relationship between two variales. The value for this always lies between -1 to 1 where -1 indicates that the two variable are Perfect negatively corelated, 0 indicates the no correlation between those and 1 indicates the Perfect positive correlation.
- ➔ Anywhere between 0.5 to 1 and -0.5 to -1 indicates relatively strong positive and negative correlation respectively
- ➔ This value can also used to compute the R2 value for the ML models like linear regression which will explain how much variance of the data is explained by these two variable.

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

- ➔ Scaling is the technique used to resize/rearrange the wider variability among data to a defined range without losing the underlying information depicted by the datapoints.
- ➔ scaling is performed to have uniformed impact on the feature value so that higher magnitude value doesn't get any special priority in interpretation by the Machine Learning algorithm compared to smaller value even though their unit of measurement are different.
- ➔ normalized scaling is a scaling strategy where the value are shrinked value between the range of 1 and -1 where as standard scaling is used to shrink value such that the mean of the distribution will be 0 and standard deviation will be 1. Being minimum and maximum value being used normalized scaling seems to be impacted when there is presence of outliers in the data where as standardized scaling is less vulnerable to outliers. Example of Normalized scaling is MinMax Scaler and that of standardized scaling is Standard Scaler

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

- this will only happen if there is the prefect correlation between the variable because as we know that VIF = 1/1-R2, and R2 in simple term is squared of correlation coefficient R which will be 1 when the R is 1 indicating the perfect positive correlation.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- Q-Q plot stands for quantile quantile plots which is used to determine what kind of distribution the dataset follow. To determine the type of distribution following steps are followed:
  a. dataset is divided into the N quantiles based on N datapoints
  b. the probability distribution is taken like normal distribution or uniform distribution or right skewed or left skewed distribution, etc and divided into same N quantiles
  c. Now the quantiles of the dataset and the divided quantiles of the probability distribution are mapped against each other and the line is fitted at each intersecting points. Better the fit more confidence on the type of distribution against which the dataset was mapped.

In linear regression it can be used to see if the data samples are normally distributed or not in case the distribution is unknow. Based on the type of distribution identified we can make decision on what type of scaling strategy to be used. Also it can be used in the residual analysis to see if the error terms are normally distributed or not .