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?

   Answer :

   - separate column and use dummies to eliminate columns which are not necessary.
   - After assigning the dummy values we can drop these categorical variables.
   - The segregation of categorical variables helps in giving more granularity on impact /influence of these variables on Target variables.
   - Bike demand is increased in 2019 compared to 2018 with high during the June – Sept period and low in Jan period
   - Bike demand has been high during the week days as compared with the week ends and additionally when it comes to Weather is sunny demand is high as compared to cloudy and rainy with seasonally low demand of spring.

   A) Out of 6 categorical variables , we can use lambda function to segregate the variables into

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

   Answer :

   - Basically, to avoid the unnecessary column, which is not adding value , we drop the True first case.
   - In the given case study for each of 4 categorical variables True first is applied , except weather sit to drop the redundant variable not contributing or impacting on target variables and its significance in imbibed into the variables.

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

   - On High level , Temperature has highest correlation with the target variable cnt .
   - Additionally , there is multi colleanearity between temp and tempa , where one of the variable can be dropped .

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

    - Ensure the distribution of errors terms are as per normal distributions
    - Error terms are independent in the chart and constant.

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

    - Year
    - Atemp
    - Season (Winter) are contributing to the shares bike demand.

General Subjective - Questions

1. Explain the linear regression algorithm in detail.

    - Linear regression uses scatter plots etc to predict how different variables predict the outcome of specific variable which is Target variables.
    - A line passing between the plotted variables which is predictive line defines how close and far the variables are from the predictive values.
    - The three type of regression logistic , linear and polynomial with mostly user linear regression in current case study where if dependent variables expand on Target which is positive linear regression
    - Predictive model formula = Y = C+mx with C is constant and m is slope and x is dependent variable and Y is target variable and under multi variables linear regression Y = Beta 0+ beta1X1+beta2X2+beta3X3…..betaNXn

2. Explain the Anscombe's quarter in detail.

    - This shows the data visualization in unique way .
    - This shows the four data sets having identical properties may appear to be different on graphs and visualize in different way.

3. What is Pearson's R?

    - This is statistical method to derived coefficient of variables and ranges between -1.
    - This method cannot differentiate between dependent and independent variables and it cannot identify the non-linear relationship between variables.
    - Additionally coefficient is derived by coefficient of co-variance between variables/(Product of SD))

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    - Scaling refers to alignment of units of measurement or converting all numerical variables into single uniform scale for performing model build , train and test etc.
    - Scaling is used when dependent and target values values are not aligned and this impacts the model built where coefficients are not comparable.(Example Sales = 500 crores and locations are 100 , unless we uniform their scale we can build model but not accurate.
    - MaxMinscaling converts everthing into range of scale between of 0 -1 and standardized converts every scaling into normal distribution based on mean as 0 and SD
    - Difference between two scaling is normalization loses some outliers vs standardization. **(In the case study wee lost 1 row as outlier) Total 729 rows.**

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

    - This is because of perfect correlation between independent and target variables and R2 values will be 1 .
    - There exists multicollinearity between different continuous and Categorical variables and need to drop those variables

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

    - This plot helps in defining the shapes of graphs for linear regressions.
    - Additionally, Q-Q plots represents data plotted with certain values falling above the median and certain falling below median with
    - This plot helps to identify whether two or three sets of data falls or comes within the same distribution