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)

   Season - The variables indicating season , that is summer & winter seem to be having impact on the count of bikes hired for every 1 unit , it seem to be adding 910 and 1175 bikes approximately if other parameters are kept zero

   Month – It is seems months Aug , Sep and Oct seems to be the highly influencing the count of bikes hired

   weathersit – Mist and Light snow seems to have a negative impact on the count

   Holiday – seems to be impacting the count negatively

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

   More the dummy features , the model might have the chance of overfitting or hard to fit . Hence drop_first =True or dropping the first columns helps in reducing an extra dummy variable column .It also helps in reducing any co-relation between variables.

   If the categorical value is with n-levels , we need only n-1 columns to represent dummy variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   atemp (input variable) variable seems to be having highest co-relation with cnt (output variable).

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

   The following validations helps in meeting the assumptions as set by linear regression model ;
   1)Error terms form a normal distribution curve
   2)Center of mean of this curve is Zero
   3)Error terms have constant variance
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   temp , yr and winter season seems to be impacting the bike hiring most .

General Subjective Questions

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

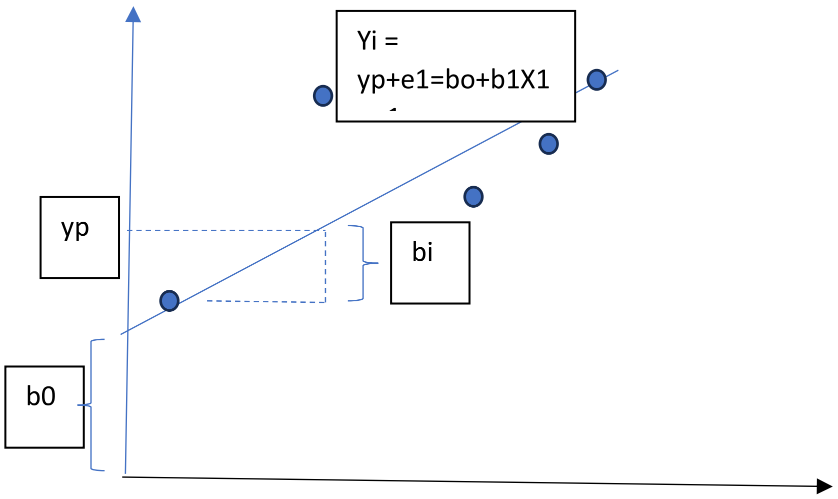
   In Machine learning, Linear regression algorithm is applied to supervised machine learning. It is used to predict behaviour of a dependent or output variable based on independent variables related to it. The relation of output (y) and input variable (X) can be fit into a straight line that best fits the different data points. It would statistically allow one to do a predictive analysis. The relation of the output and input variables can be denoted as below;

   $y= b0+ b1\ X1+ b2X2 +b3X3 .....biXi +.....+ bnXn$

   Where y is the output variable and Xi denotes the independent variable.
   b0 – stands for constant which is value taken by y when all X variables are zero
   b1..bn - denotes the slope or regression coefficient or scale factor

   The above equation is used to denote the best fit line , model is used to evaluate different weights/coefficient (bi) to establish strong relationship between variables. Cost function is used to optimise the coefficients and it is denoted by MSE(mean squared error)

   $MSE = 1/N \sum (yi –(mXi -b)) (yi –(mXi -b))$
   N – total observation points
   yi – Actual datapoint or value observed



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

Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data to analyze it before building the model. The group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Note : The emphasize is to visualise the data before building model; since datasets even it is from same domain might relate differently , though apparently it might look similar. Visualisation helps in understanding the data distribution, patterns, co-relation, outliers and anomalies.

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

Pearson's co-relation co-efficient R , is a way of interpreting the linear co-relation of variable This number falls in the range -1 and 1 that measures the strength and direction of the relationship between two variables.
0-1 : positive co-relation; eg:size of pizza and pizza price
0 : no co-relation ; eg: price of a meal and meal serving plate price

-1- 0 : negative co-relation ; eg: more the expenditure ,less is the saving.

Alternatively it is known by other names like corelation coefficient or bivariate co-relation.

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

Scaling in machine learning is a step taken prior to data modelling to normalise the independent variables or input variables. In reality the data varies in magnitude, range and unit ; scaling enables the model to look at these features on a same scale/range. This would enable an apple-to-apple comparison of features; thus improving the prediction accuracy. The data range is brought to fit within a scale, let us say 0 to 1 ; so that the feature values are the same range ; this is specifically applied for linear regression to improve the efficacy of the model .

<u>Difference between normalised and standardized scaling</u>

Standardization	**Normalization**
The data is not bound in a range	Data is bound in a range 0-1 or -1 to 1
Doesn't affect outliers much	Affect outliers
Assumes that data is in Gaussian distribution	Used when data is not in Gaussian distribution or normal distribution
X(stand) = X- Xmean/Std deviation	X(new) = X-Xmin/Xmax-Xmin

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

When there is a perfect co-relation between variables VIF would be infinte .The variation inflation factor or VIF is calculated as below ;

VIF = 1/ 1-R square

When RSquare = 1 , VIF would infinite . RSquare = 1 means , there is no error and the variables are in 100% relationship
(RSquare = 1 – RSS/TSS , RSS- sum off squares of residuals , TSS- Total sum of squares.RSS denotes the difference between predicted value and actual value , TSS – is the sum of squares wrt mean)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   The purpose Q-Q plot or Quantile -Quantile plot is to verify whether datasets come from same distribution. It is constructed by plotting the datasets quantiles long x axis and other on y axis.

   This becomes important in linear regression when train and test sets are received as two different lots.Using Q-Q plot , it is verified whether they are coming population with same distribution.

   It helps in verifying

   1) Whether data coming from same population
   2) Have same scale
   3) Have same distribution
   4) Have similar behavior

      Similar distribution – Data lies closer or on a straight line at an angle of 45 degress from x axis
      Different distribution – if data points lie away from straight line at an angle of 45 degrees

      Python – provides statmodels.api to plot Q-Q grapgh.