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)

   Bike rental counts show a positive correlation with temp due to which rental count could increase with high temperature and there might be no takers during a low temp days

   Bike rental count show a negative correlation with humidity due to which rental count are less with hight humidity and increase on a low humid days

   Bike rental counts shows a negative correlation with windspeed due to which rental counts are less on high windspeed days and low on low windspeed days
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

   it helps in reducing the extra column created **during dummy variable creation**
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
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Will validate the model using mixed approach where we will select 15 variables of RFE using automated approach and remove variable with higher pvalues and VIF using manual methods

   Will use Sickit learn for it compatibility with rfe

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   We can see the demand for bikes depends mainly on :
   -yr , holiday ,Spring, Mist_Cloudy, Light rain_Light snow_Thunderstorm,3 ,5 ,6, 8, 9, sunday, 7, 10
   -Demands increases in the month of 3, 5, 6, 8 ,9, 7 , 10 and yr
   -Demand decreases if it is holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy, Sunday
   -Final recommendations for the company:
   -Demand is higher in month of 3, 5 , 6, 8, 9 ,7 and 10

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between

dependent and independent variables, they are considering and the number of independent variables being used

2. Explain the Anscombe's quartet in detail. (3 marks)
   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
   It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
3. What is Pearson's R? (3 marks)
   In statistics, the Pearson correlation coefficient , also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation,[1] is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).
4. What is scaling? Why is scaling performed?
   Scaling (geometry), a linear transformation that enlarges or diminishes objects. Scale invariance, a feature of objects or laws that do not change if scales of length, energy, or other variables are multiplied by a common factor
   Standard Scaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. Standard Scaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature

What is the difference between normalized scaling and standardized scaling? (3 marks)

Normalization	standardization.
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called Min Max Scaler for Normalization.	Scikit-Learn provides a transformer called Standard Scaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization

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

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