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)

   For suppose if they are categorical variables, converting them to integers and then adding to the dependent variables set, helps in improving the model performance. We can achieve getting a good R-Squared value for the model.

   The way to handle categorical variables is by creating dummy/indicator variables. We can have (N-1) dummy variables (new columns) to describe categorical variable with N levels.

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

   drop_first = True helps in reducing the number of dummy variables created.
   For example, if I had dummy variables, a,b and c created with 3 different columns;

a	b	c
1	0	0
0	1	0
0	0	1

   The columns can be reduced to two by eliminating one. The variable values are calculated by considered the rest variables like, a = 00, b = 10, c = 01

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

   Highest correlation was with registered column. Hence, it is not considered under dependent variable, I say the highest correlation with target variable was temp variable.

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

   On having proper analysis with the excel data provided, I assumed ***cnt*** was going high with some features using a pivot table. On proper analysis of the pivot table, I understood there are multiple features for predicting the target variable.
   Applying the scaling of variables in the training data gives rsquared and pvalue (OLS).
   Checking with the pvalue of individual feature, the insignificant variables get eliminated for good model score. Once all the pvalues were < 0.05, then the model seems to be working fine and VIF comes into picture.
   On calculating VIF and pvalue we can decide which features are insignificant later.
   - If pvalue is high(>0.05) and VIF is low(<5), the feature can be eliminated.
   - If pvalue is low and VIF is high, the feature may stay in the model if it shows a good rsquared value.
   - If pvalue is low and VIF is also low, the feature can stay in model undoubtedly as it is significant.

The plot taken from the model is **normally distributed,** this is a usual assumption for linear regression.

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

   - season
   - yr
   - temp

   (based on pvalue calculated and variable selection with RFE)

General Subjective Questions

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

   It is supervised machine learning algorithm which predicts the target variable based on previous experiences. We provide previous data to machine and helps machine to learn from it. The data will be analyzed and predicted by following the steps-

   a. Reading and understanding the data

      a. Data format is checked thoroughly here whether the data has null or not-null values.

      b. If data holds some categorical values, they have to be converted to required format (numbers).

   b. Visualizing of data and Data Preparation

      a. The pair-plots are created to understand how well the features are correlated with target variable.

      b. Data Preparation includes splitting the dataset in two categories – Train and Test. The entire model building starts from train dataset and then transformed to test dataset.

      c. The assumed significant features has to be rescaled so that we can build a good model.

   c. Training the model (Model building)

      a. This step involves checking the correlation between features for building the model.

      b. pvalue and VIF value decides if the model is good to go.

      c. Model building can happen in 2 ways –

         i. Consider one feature first, find the pvalue and rsquared value. Next, keep adding the other features to the model to know if the pvalue and rsquared value get better. Addition of features should happen until there is a

rsquared value improvement found.

ii. Find the rsquared and pvalue by considering all the features at a time. Consider the high values and start eliminating the respective feature to see the model improvement.

d. Residual Analysis

a. Check for the normally distributed plot

e. Evaluation and Prediction of Test set

a. This is the last step which transforms all the analysis done on train set to test set. Hence we do prediction at this stage assuming any extra data value should run with the model created.

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.

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

The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

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

Scaling indicates type/category of the variable. Few variables may hold integer values and few float. On calculating the OLS, there are chances of getting unexpected coefficient values which leads to wrong model. Hence feature scaling has to be done, if many dependent variables are found and if most have different scales. The main reason for scaling is for ease of interpretation and faster convergence for gradient descent methods.
There are two methods of scaling –

a. Normalization – Also known as minmax scaling. This method focuses such that the value of variable lies in between 0 and 1 as minimum and maximum value.
b. Standardization – This method focuses such that mean is zero and standard deviation is one.

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

We may have value of VIF as infinite if the rsquared value is 1,i.e, for a perfect correlation.
Lets think about the formula for VIF – 1/(1-rsquared)

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 mean Quantile-Quantile plot. This plot comes from common distribution of the datasets. This plot helps in understanding if the data is common if we received any training set and test set received separately.

Q-Q plot can be used for sample size datasets and in detecting the changes in scales/symmetrics, presence of outliers etc.