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)

Ans. There is a significant effect of categorical variables on the dependant variable, for e.g., 'season' variable shows that the demand was high during 'fall' season, 'month' also indicates the same. The demand of bikes was maximum in August(08) which lies in 'fall' season only. And if we pick 4 months when the demand was high, we see June to September (06 to 09) shows the highest peak, which again lies in the 'Fall' season.

Similarly, the demand is zero when weather situation is 'High Snow/Rain' and highest when the weather is 'clear'.

Also, people tend to hire more bikes on 'working days' than on a 'holiday'. It may be possible that people prefer to commute using bikes while going to work.

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

Ans. Using **drop_first=True** during dummy variable creation means to drop the first column of dummies created for any categorical variable. This implies if we have a categorical variable with 'm' levels, we should create 'm-1' dummy variables. It helps in reducing the extra variable creation because that variable can be explained by rest of the dummy variables all together. This will reduce the redundancy and chances of Multicollinearity.

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

Ans. After having a look on the pair-plot while performing EDA, we see that among all the numeric variables 'temp' (temperature) seems to have the highest correlation with the target variable which is 'cnt'. And this correlation is showing a positive trend.

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

Ans. After building the final model, to validate the assumptions of Linear Regression we did Residual Analysis:

1. Errors were normally distributed; we plotted a graph with residuals to validate this assumption and found errors were following a normal distribution where the mean was lying nearly at 0.
2. Error terms were homoscedastic in nature, which means they had constant variance.
3. Errors were independent of each other, in the graph we see if errors were showing any trend or not. We analysed there was no pattern, and the distribution was totally random which validates the independency.

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

Ans. The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

i. 'Temp' : Temperature contributes the most in explaining the demand of bikes, by showing a positive trend.
ii. 'Light Snow/Rain' : If weather situation is Light Snow or Rain, then people tend to avoid taking the bikes. Hence, it shows a strong negative relation with the demand of bikes.
iii. 'yr' : We see that with each passing year people tend to hire more shared bikes due to popularity among people.

General Subjective Questions

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

Ans. After the data cleaning/handling and EDA, Linear Regression Algorithm can be executed which consists of mainly 3 stages –

i. Data Preparation

- After analysing the correlation between the variables, we move forward for Data preparation. Linear Regression model deals with numeric data, so we change the categorical data to binary format by creating dummies.
- We then divide the data into test and train set, where we create our model using train set and evaluate the model on test set.
- Next comes the most important step of data preparation, i.e., to perform scaling on the variables. We can use any of the techniques either normalisation or standardisation to perform scaling.
- Then we divide the train set into X and y, where X are the independent variables and y is the dependant variable.

ii. Data Modelling

After the data is ready, we build a model by using automated approach(RFE) or manual approach(Adding variables one by one and building the model).

Generally mixed approach is preferred where the features are selected using automated approach(usually RFE) and then the variables with pvalue higher than 0.05 or VIF higher than 5 are eliminated one by one. Once the model is rebuild with all the variables which satisfies the pvalue and VIF, we move forward to evaluate the model and validates the assumptions of Linear Regression.

iii. Data Evaluation

Once we are ready with the model, we validate the assumptions of Linear Regression by doing residual analysis. Assumptions say; Zero mean, independent, Normally distributed error terms that have constant variance. We check all these in Residual Analysis.

Evaluate the model on Test set and check the Adjusted R Square for both test and train data. The difference in the R2score of these two should not be greater than 5%, otherwise this will be a case of overfitting.

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

Ans. Anscombe's quartet comprises of four different data sets wherein each data sets contain 11 (x,y) data points. These data sets have almost identical summary statistics, (*mean of x and y will be*

the same in all the four datasets, variances are also approximately same, correlation between x and y are almost the same) but shows different distributions when we plot graphs on these datasets.
In a nutshell, the visualisation of the datasets are completely different wherein they have exactly the same summary statistics. This shows the importance of visualising the data during data analysis.

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

Ans. Pearson's R is a statistic that measures linear correlation between two variables let say X and Y. It has a value between +1 and −1.

A value of **1** implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of **–1** implies that all data points lie on a line for which Y decreases as X increases. A value of **0** implies that there is no linear correlation between the variables at all.

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

Ans. Scaling is a technique to standardize the independent variables/features present in the data in a fixed range. It is performed during the Data Preparation to handle highly varying values or units. It also takes care of the outliers/extreme values. If scaling is not done, then our model ends up with very weird coefficients that might be difficult to interpret.

The reason to perform scaling in regression is for easy interpretation and faster convergence for gradient descent methods.

Feature scaling can be done using two popular methods which are Normalized scaling and Standardized scaling. The main difference between the two is, **Normalization** scales a variable to have a values between 0 and 1, while **Standardization** transforms data to have a mean of 0 and a standard deviation of 1.

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

Ans. Variance Inflation Factor (VIF) is a technique used to detect multicollinearity in a model. It can be calculated as **$1/(1-R^2)$**, where R^2 is the square of correlation between the variables, which measures the proportion of variation in the variables.

As the formula states, higher the value of R^2, lower will be the value of denominator hence, higher will be the value of VIF. In the case, when R^2 is 1 (*which shows the perfect correlation between the variables*), the denominator for the above formula becomes 0, and any number divided by zero is Infinity. Thus, the VIF comes out to be Infinity is there exists perfect correlation (i.e., R^2 is 1)

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

Ans. Q-Q Plots, also known as Quantile-Quantile plots are plots of two quantiles against each other. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot, if the two data sets come from a common distribution, the points will fall on that reference line.

This helps in linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. This plot can be useful in detecting the presence of outliers, many distributional aspects like shifts in location, shifts in scale or changes in symmetry.