**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 demand in year 2019 is higher as compared to 2018.
* Bike demand takes goes down in spring and when there is Light Snow & Rain
* All the positive coefficients like yr, mnth\_Sep indicate that an increase in these values will lead to an increase in the value of cnt.

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

* To reduce the collinearity between dummy variables
* To delete extra column while creating dummy variable

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

* Variables atemp, temp show positive correlation with target variable

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

* Error terms are centred around 0 and follows a normal distribution

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

Top three features contributing significantly towards explaining the demand are:

1. yr (0.247820),
2. weathersit\_Light Snow & Rain (-0.256802)
3. season\_Spring (-0.303522)

**General Subjective Questions**

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

* Linear regression algorithm is a technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s).
* After looking into the data and cleaning it with exploratory data analysis, we split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output).
* After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.
* According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset.
* The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model

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

* In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

3. What is Pearson’s R? (3 marks)

* Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables.
* It is commonly used in linear regression.
* The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression.
* The values in between denotes the relative collinearity of two variables.

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

* When there are a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.
* So we need to scale features because of two reasons: 1. Ease of interpretation 2. Faster convergence for gradient descent methods
* We can scale the features using two very popular method:
* 1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
* 2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc
* Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.
* Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

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

* There is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1.
* Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

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 is a graphical tool to assess if sets of data come from the same statistical distribution.
* It is helpful in linear regression when we are given testing and training datasets differently.
* Hence, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.