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 mark)

    - The various categorical values as defined below had an effect on the dependent variables.
        - season (1:spring, 2:summer, 3:fall, 4:winter)
        - yr : (1:spring, 2:summer, 3:fall, 4:winter)
        - mnth : month (1 to 12)
        - holiday
        - weekday: day of the week
        - workingday: working day
        - weathersit: weather situation
    - The correlation between these categorical columns also had a high multicollinearity between themselves. As we could see in the model that the season and weather situation are closely linked as light snow situation is linked to winter month
    - Similar would be for weekday, holiday, working day as these are also highly correlated.
    - Their effect on the dependent variable can be seen as per the model based on the final rfe3 model where holiday, season and weathersit are part of the coefficient for the dependent variable.
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

    - The "drop_first = True" plays of role to make sure that one of column is dropped off.
    - For example, if we have 3 values for categorical columns – A, B, C. Now if we define three columns colA, colB, ColC, then to know which value is A or B or C can be defined by the two columns colB and ColC. B would be defined as [1, 0], C would be defined as [0, 1] and A would be defined as [0,0] where the first value is colB and second is in colC. This was we can reduce columns and only keep two which is what the function does.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

    - Temperature (temp and atemp) has the highest correlation.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

    - We can see from the final model that here is a linear relationship between the X and y.
    - From the residual analysis, we can see that the normally distributed, having constant variance (Homoscedasticity) and are independent.

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

    - The following X variables are contributing highly to the y variable.
        - temp (positive)
        - hum (negative)
        - yr (positive)

General Subjective Questions

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

    - Linear Regression is one of the models used in machine learning where we train a model to predict the behavior of the data based on some variables.
    - Linear Regression is a machine learning algorithm based on supervised learning and it performs a regression task based on the mathematical and statistical equations for linear regression.
    - The Linear Regression assumes that there is a linear correlation between the independent (X) and dependent (y) variables that means the two variables.
    - This means that x-axis and y-axis are linearly correlated. This can be defined mathematically by y = a +bx where a (intercept constant), b (slope), x = Independent variable from dataset, y = Dependent variable from dataset
    - This definition defines the relationship between the independent and the dependent variable.
    - This equation only shows a single independent variable, but the mathematical equation can be expanded to manage multiple independent variables.
    - Linear Regression approach is used to model which will predict the value for target variable based on independent variables. It is mostly used for finding out the relationship between variables. It is used extensively in forecasting simulations.
2. Explain the Anscombe’s quartet in detail. (3 marks)

    - Anscombe's quartet is a set of four datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different as defined by the by the statistician Francis Anscombe in 1973
    - These datasets that have almost identical simple statistical properties yet have very different distributions and appear very different when graphed.

- This shows the importance of Data Visualization and using graphs to understand data.
- This becomes more important in Machine Learning as it can show the understanding of the data and bring out anomalies which could impact the final model or algorithm.
- This was also illustrated in the UpGrad statistical course.

3. What is Pearson's R?
    - In statistics, the Pearson correlation coefficient (r) is the one of the way of measuring a linear correlation and simply defines the linear correlation between two sets of data
    - The Pearson is a descriptive statistic, and it summarizes the characteristics of the dataset.
    - It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
    - This concept was developed by Karl Pearson
    - Pearson's R also follows the assumptions of the Linear Regression – Independent variables, Linear relationship, Homoscedasticity.
    - The degree of correlation is defined by the values which lie between +1 and -1. For example, a value between +0.3 and +0.49 will be moderate degree correlation in positive direction.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - Scaling is a pre-processing step used in Machine learning models.
    - The scaling is applied to the independent variables to normalize a data within a set of range which allows all the dataset to have a common range.
    - The data which is normally collected contains features with various magnitude and units of data, and the models will only consider the magnitude but not the unit. Thus, it makes sense to normalize the data so all features from a numerical range perspective are treated similarly.
    - There are two types of scaling normally used.
        - Normalization method rescales the values into a range of [0,1]. The normalization process defines as the maximum value equal to 1 and minimum value as 0 and the rest of the values between these two values as per the ratios.
        - Standardization is another scaling method where the values are centered around the mean with a unit standard deviation. This

means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

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

    - The VIF become infinite when there is a perfect correlation in the dataset variables.
    - This basically means that two variables are perfectly correlated with R squared equal to 1. Hence as per the formula of VIF, the denominator will become zero and the resultant will be infinite.

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

    - The Q-Q plots are plots of two quantiles between sample dataset and a theoretical dataset to understand which probability distribution is being followed by the data. The dataset could have a normal, uniform, or exponential.
    - This approach can help to plot to understand if the distribution of the data across the quantile are similar or different.
    - Probability distributions are essential in data analysis and decision-making. This is one of the key attributes for defining the Machine Learning models.
    - The linear regression machine learning models work best under some distribution assumptions. Knowing which distribution, we are working with can help us select the best model.