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.

Effect of independent variable on dependent variables are:

- The most important factor was clear weather, that implement a great optimal for sharing bikes as renting, as temperature is optimal, the humidity is less and because of that the weather is clear.
- Majorly the bikes are on rent on non-holidays. On holiday's, people spend the time with their families. They prefer to have a rent a bike on working days.
- On Sunday and Wednesday, the booking is on high. Busy schedule of the working peoples might be reason.
- Working day's & non-working days the spread for business have almost the same median.
- Fall months have higher median by the overall spread during reflection of season plot. Which is due to weather condition are most optimal to ride bikes for following spread.
- Overall, business increases or decreases by analysis.

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

Ans. A variable with n level can be represented by n-1 dummy variables. drop_first=True is important to use, as it helps in reducing the extra columns created during dummy variables creation. The correlation created among dummy variables hence reduces.

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

Ans. The numerical variable which has highest correlation among target variables are "temp" and "cnt".

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

Ans. It is the most important part as in data science perspectives. It relies on some basic factors like linearity, homoscedasticity, absence of multi-collinearity, independence, and normality of errors etc.

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. temp (Positive correlation)
Weathersit_Light_Snow (Negative correlation)
Yr_2019 (Positive correlation)

General Subjective Questions

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

Ans. It is known as machine learning algorithm based on supervised learning of data. The linear regression mainly performs a regression model with target driven predictive values which is based on independent variables.

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

Ans. A group of four data sets in simple descriptive statistics are known as Anscombe's quartet. The data set can be distracted by peculiarities with variables that fools regression model. When these models are visualised by scatter plot, they have very different distributions and appearance.

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

Ans. Pearson's correlation also known as Pearson's R is a measure of linear correlation between two data sets. It is the product of their standard deviation and the ration between covariance of two variables. In simple words we can say that it measures the strength between two variables and has a value between -1 to 1.

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

Ans. Scaling is the technique to standardise the independent feature present in the data set in a fixed range. It is performed during the data is pre-processing to handle highly varying magnitude or values. Standardization or Z-score normalization is the transformation of features by dividing by standard deviation and subtracting from mean. Normalized means, rescaling the value from [0,1].

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

Ans. The VIF = Infinity occurs at perfect correlation. It is showing perfect correlation between two independent variables. We often get R2 = 1 in case of perfect correlation, which leads to 1/(1-R2) infinity.

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 plot is known as Quantile-Quantile plot in linear regression, which is a graphical tool to confirm, where the data came from like from normal, exponential, or uniform distribution. In sample data set, it determines that the data sets are from population with a common distribution along with testing datasets.