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: I have used the boxplot and bar plot to analyse categorical columns. The few insights we may get from the visualisation are listed below.

- Bookings appear to have increased over the 'Fall' season. And, from 2018 to 2019, the number of bookings in each season dramatically increased.
- Demand is increasing month after month till June. The demand is at its peak in September. Demand has decreased since September.
- Demand is less during holidays.
- Both working days and non-working days looked to have about the same number of bookings.
- Bookings for 2019 were higher than those for 2018, indicating positive business growth.
- It appears obvious that clear weather led to more bookings.

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

Ans: It is crucial to use drop_first = True since it aids in eliminating the excess column formed when a dummy variable is created. Thus, it lessens the correlations that are produced among dummy variables.

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

Ans: 'temp' variable has the highest correlation with the target variable.

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

Ans: Based on the following five assumptions, I have validated the linear regression model's assumption.

- Normality of error terms - Error terms should be normally distributed
- Multicollinearity check - There should not be any significant multicollinearity among variables.
- Linear relationship validation – Actual and predicted values should follow same pattern
- Homoscedasticity - There should be no visible pattern in residual values.
- Independence of residuals – Error terms should be independent of each other

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 as follows –

- temp
- weathersit_Light_snowrain
- windspeed

General Subjective Questions

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

Ans: An analytical statistical model that examines the linear correlation between a dependent variable and a set of independent variables is known as a linear regression. According to the linear relationship between variables, the value of the dependent variable will vary proportionally (increase or decrease) when the value of one or more independent variables changes. Mathematical equation of Linear regression is represented by

$$y = mX + c$$

where, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y. c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

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

Ans: Anscombe's Quartet is a set of four data sets that, while they appear to be almost equal in simple descriptive statistics, have certain anomalies that, if a regression model is developed, would deceive it. When displayed on scatter plots, they have significantly different distributions and show up differently.

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

Ans: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

Ans: Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units. In the absence of feature scaling, a machine

learning algorithm would often prioritise larger values over smaller ones, regardless of the unit of measurement.

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

Ans: VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove one of the variables from the dataset that is the source of this ideal multicollinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical method for identifying whether two data sets are from populations with a similar distribution.

Use of Q-Q plot: The quantiles of the first data set are plotted against the quantiles of the second dataset in a q-q figure. A quantile is the percentage of points that fall below the specified number. In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is plotted. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The farther the two data sets deviate from this reference line, the more evidence there is that they came from populations with different distributions.

Importance of Q-Q plot: It is frequently desirable to determine whether the assumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.