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)**
   Answer:
   - I have noticed that, there is a huge hike in 2019 compare to 2018.
   - Maximum bookings are observed in the month of May, June, July, August and October.
   - On working days booking appears to be average. It’s almost the same either if its holiday or working day.
   - To conclude, we can see that there is a good progress in business comparatively.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:
   When creating dummy variables using `drop_first=True` is important because it helps in reducing the dimensionality of the data and avoids the dummy variable trap. it helps in reducing the extra column created during dummy variable creation. This reduces the correlations created 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)**
   Answer:
   ‘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)**
   Answer:
   - Linearity should be visible among variables.
   - Insignificant multicollinearity among variables.
   - Error terms should be normally distributed.
   - There should be no visible pattern in residual values.
   - There should be no auto-correlation.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Answer:
   - temp
   - winter
   - month

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer:
   Linear regression is a type of supervised machine learning algorithm that is used to model the relationship between a dependent variable and one or more independent variables. It is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. It is a statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
   Mathematically the relationship can be represented as Y = mX + c.
   Here, 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.
Furthermore, the linear relationship can be positive or negative in nature as explained below−

Linear regression is of the following two types −
Simple Linear Regression
Multiple Linear Regression

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

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. It was developed by statistician Francis Anscombe. It comprises four
datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely.

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

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

(3 marks)

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.