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 done analysis on categorical columns using the bar plot. Below are the few points we can infer from the visualization –

   - Fall season attracts higher bookings than other seasons, and there's notable surge in bookings from 2018 to 2019 across all seasons.
   - The bike rentals in 2019 exhibited a significant 65% increase compared to the previous year, 2018.
   - Most bikes are booked on Friday and Thursday
   - Most bikes are booked on working day as compared to holiday

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

Ans- By setting drop_first=True when creating a dummy variable, you avoid including one of the hierarchy levels as a separate dummy variable. This helps to prevent complete multicollinearity between the dummy variables. When a category is dropped, the reference category is the starting point for comparison with the other categories. This simplifies the interpretation of the model and simplifies the effect of individual categories on the target variable.

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

Ans- The column temp is highly correlated.

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

   Ans- I have validated the assumption of Linear Regression Model based on below 4 assumptions –

   - **Normality of error terms** : Error terms should be normally distributed
   - **Multicollinearity check** : here should be insignificant multicollinearity among variables.
   - **Linear relationship validation**: Linearity should be visible among variables
   - **Independence of residuals** : 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)

Ans-

- year
- Temp
- mnth_sep

General Subjective Questions

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

Answer-

Linear regression may be defined as the 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 with the help of following equation −

$Y = mX + c$
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−

- Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.
- Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.

Linear regression is of the following two types-

- Simple Linear Regression
- Multiple Linear Regression

2. What is Anscombe's quartet?

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet they exhibit significantly different patterns when graphed and analyzed. Anscombe's quartet consists of four distinct datasets, each comprising a set of (x, y) data points. Despite having similar statistical properties, these datasets have vastly different patterns when graphed and analyzed. The quartet was designed to emphasize the importance of data visualization and the limitations of relying solely on summary statistics.

3. What is Pearson's R?

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?

Answer:

Scaling is a pre-processing step in data analysis that involves transforming the features of a dataset to a specific range or distribution. The goal of scaling is to bring all features to a common scale, which can help improve the performance of certain algorithms, ensure fairness in feature contributions, and make the optimization process more efficient.

Normalized Scaling (Min-Max Scaling): Normalized scaling, also known as min-max scaling, transforms features to a specified range, usually between 0 and 1.

Standardized scaling centres the data at zero and adjusts it based on its spread (standard deviation). This transformation maintains the relative distances between data points and is useful when the features have different units or distributions.

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

Answer:

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.**

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.