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?

Answer:

The analysis has been performed using the boxplot. Below are few of the observations from the visualization –

- Fall season seems to have more bookings than other seasons and spring season had very least no of bookings.
- The booking count had drastically increased from 2018 to 2019.
- The pattern of booking has been in increasing order from the month of Jan to June and it followed decreasing pattern from Sep to Dec.
- The month of Sep had maximum number of bookings and the month of Jan had least number of bookings.
- Clear weather attracted the bookers.
- Booking is equal on both working and non-working days.
- Booking seems to be more on the holiday.
- Thur, Fri, Sat and Sun had more no of booking when compared to other days of the week.

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

Answer:

drop_first=True flag helps in reducing the extra column creating during dummy variable creation which also helps in reducing the correlations created among dummy variables.

For Example, let's say we have three types of values in categorical column, and we want to create dummy variable for that column. In this case, if one variable is not Var-1 or Var-2 then it's obvious that it will be Var-3. So we do not need 3rd variable to identify the Var-3.

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

Answer:

'tmp' variable has highest correlation with the target variable 'cnt'

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

Answer:

The assumption of Linear Regression can be validated using the following assumptions:

- The error terms should be normally distributed.
- Multicollinearity validation.
- There should be independence of residuals.
- The variables should have linearity visibility.
- The p-value must be dropped down.

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

Answer:

'tmp', 'winter', 'sep' are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions:

1.Explain the linear regression algorithm in detail.

Answer:

Linear Regression may be defined as the statistical model which analyses the linear relationship between a set of variables i.e., a dependent variable with a set of independent variables. Here linear relationship mean that the when the value of one or more independent variable increase or decreases then the dependent variable will also change accordingly.

The mathematical equation for the linear relationship can be given as:

$Y = mX + c$

Where,
Y → dependent variable to be predicted
m → slope of the regression
c → constant (Y-intercept)
X → independent variable used to make predictions.

Linear regression is of two types:

1. Simple Linear Regression
2. Multiple Linear Regression

Linear Regression itself can be positive if both the dependent and independent variables increases linearly and the linear regression can be negative if both the dependent and independent variables decreases.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet 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, and we must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar 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.

$r = 0$ → No association between the two variables
$r > 0$ → Positive association between the two variables, if one variable increases the other variable should also increase

r < 0 → Negative association between the two variables, if one variable increases the other variable tend to decrease and vice versa

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

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 especially to handle outliers. 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 mg to be higher than 5 kg 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.

S.No	Normalized Scaling	Standardized Scaling
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	. It is not bounded to a certain range.

4	It is really affected by outliers.	It is much less affected by outliers.

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

Answer:

VIF = infinity. Indicates the perfect correlation. 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.

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.