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:
Insights shows the relationship between categorical variables and a Target variable.
Bike Rentals are more :

- In the year 2019 compared to 2018
- During the month of September
- During the Fall season and then in summer
- On Saturday, Wednesday and Thursday

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

Answer:
It helps in reducing the extra column created during dummy variable creation. Hence it 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:
Bike rentals are highly correlated to temperature

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

Answer:
The Errors formed a normal distribution which helped in validating the assumptions of Linear Regression after building the model on the training set

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:
Working days
Temperature
Season
Year and Month

General Subjective Questions

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

Answer:

Linear regression simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.

- If there is a single input variable (x), such linear regression is called simple linear regression.
- If there is more than one input variable, such linear regression is called multiple linear regression.
- The linear regression model gives a sloped straight line describing the relationship within the variables.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

Linear Regression equation

$$y = b0 + b1x$$

y= Dependent Variable.
x= Independent Variable.
b0= intercept of the line.
b1 = Linear regression coefficient.

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

Answer:

Anscombe's Quartet is defined as a group of 4 data sets which are nearly identical in simple descriptive statistics, but there are some unusual features in the dataset that impacts the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these 4 data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all 4 datasets.

The 4 datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R

Answer:

Pearson's Correlation Coefficient (r):

Is defined as the (sample) covariance of the variables divided by the product of their (sample) standard deviations, measures the strength of a linear relationship between two quantitative variables. The results will be between -1 and 1. We get a number somewhere in between the above values. The closer the value of r gets to zero, the greater the variation the data points are around the straight line of best fit. Positive values indicate direct relationships (as one variable increases, the other increases as well). Negative values indicate inverse relationships.

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

Answer:

What:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization Scaling:

It brings all of the data in the range of 0 and 1.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Answer:
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:
Q-Q Plots (Quantile-Quantile plots)
These are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.