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:

After analysing categorical variables against the target variable (dependent variable), we can infer following things:

•Season Fall has the highest cnt across all season.

• May June July August September seem to show uptrend towards bike demands (cnt variable). Then trend decrease by end of the year continue till January.

• There is drop in the mean value of the bike demand (cnt) on holidays. So people stay at home.

• No major change in bike demand (cnt) for weekday. As the mean value is almost same for all days. There is minor increase in demand on Friday Saturday Sunday.

• No major change in bike demand (cnt) for working day. Doesn't impact the bike demand

• During Clear weather there is a highest bike demand (cnt), people prefer clear weather to ride bike.

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

Answer:

the drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding. By setting it as true we reduce the correlation between the encoded variable.

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

Answer:

'tmp' variable has the highest correlation with target variable (cnt).

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

Answer:

Assumptions of the linear regression :

1 Error terms are normally distributed with mean zero

By plotting the displot on the residue we are able to visualize mean is zero and error terms are normally distributed.

2. linear relationship between X and Y
By plotting pairplot graph we can see the linear relationship between the cnt and independent variables.

3. homoscedasticity
by plotting scatterplot among the residue and cnt variable. There shouldnt be any pattern formed between them.

4. Error terms are independent of each other
Using Durbin-watson statistic we can identify there is no auto correlation if value is between 2 and -2.

5. Multicollinearity
By using VIF (variance inflation factor) formula we are able to perform multicollinearity check.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
1. temp
2. weather_snow
3. yr

General Subjective Question

1. **Explain the Linear regression algorithm in detail. (4marks)**
Answer:
Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
The linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis) is called linear regression.

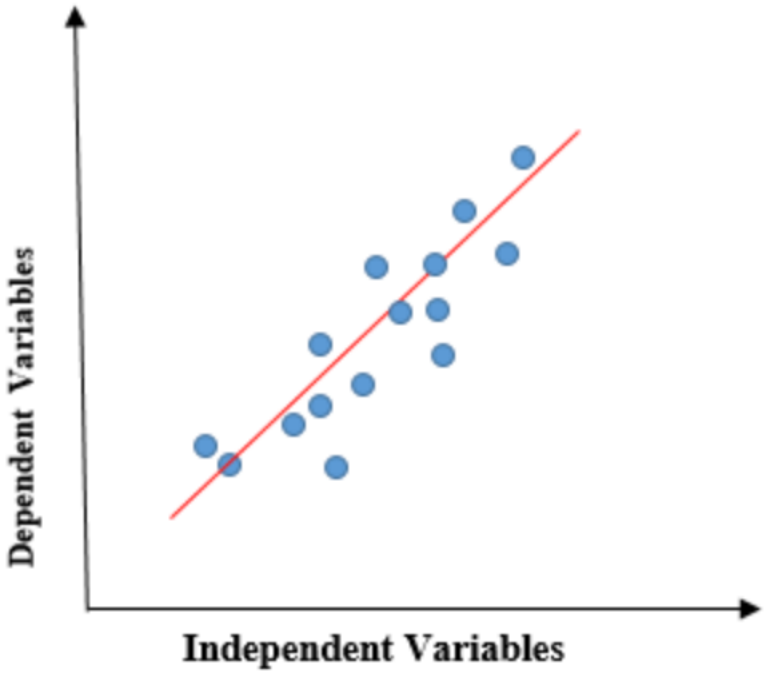
If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression.** The linear regression model gives a sloped straight line describing the relationship within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

Mathematically the relationship can be represented with following equations-

$$Y = mX + c$$

Y = dependent variable to predict
X = independent variable we are making use to predict

c = constant (Y intercept), if X =0 , then Y =c



Linear Relationship can be positive and negative.

1. Positive when both independent and dependent variable increase.
2. Negative when independent decreases with increase in dependent variable.

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

Answer:

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to

see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

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

Answer:

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.

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 in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

ML algorithm just sees number, if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model.

Normalized Scaling:

1. Minimum and maximum value of features are used for scaling.
2. It is used when features are of different scale.
3. Scale values between 0-1 or -1 to 1.
4. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

Standardized Scaling

1. Mean and standard deviation is used for scaling.
2. It is used when we want to ensure zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer :

VIF stands for Variance Inflation Factor.
When features are perfectly correlated, then VIF become infinity.
VIF = 1 / 1-R^2

In case of perfect correlation R^2 become 1.

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.
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.