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:

In the data, there are total 7 categorical variables. i.e season, year, month, holiday, weekday, working day, and weather sit.

Out of which there are few variables which took place in the final model. Like Year, month, season, weather sit and weekday.

- Year has a positive influence on the dependent variable (count) and also the demand in 2019 is on the higher side.
- July and September months have taken place in the final model. Where July has a negative impact on the dependant variable means there is a negative correlation between the variables, whereas the September month is positively correlated with dependant variable.
- Season summer and winter have a positive correlation with the dependant variable and also have a high beta associated with them. There is a high demand in the season.
- On the non-working day (Sunday) the dependant variable has a negatively correlated relation which means the demand on holidays are low for the bikes.
- Weather situations like Light Snow and mist + cloudy have a negative impact on the demand of bikes.

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

Answer:

Whenever we create dummy variable for any categorical variable, the function gives us the n creation, but as the requirement we need only n-1 output. So at the time of creating the dummies we drop the first variable.

For example:

If there a categorical variable (temperature) with three outcomes: high, medium and low.

High	Medium	Low
0	1	0
1	0	0
0	0	1

So in this situation only two variable can explain the situation for the third one itself.

Like 0 0 can easily say that the temperature is **low.**

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

Answer:

The 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?**

Answer:

We have three assumptions of linear regression:

- **Linearity :** With the help of pairplot we are able to check the correlation between the independants variables and the dependant variable.
- **Multicollinearity:** By using the VIF (Variance inflation factor) we are ab;e the remove the correlated independent variables.
- **Normality of errors / residuals:** With the help of distplot we can see that the error are normally distributed.

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

Answer:

The top 3 features contributing significantly towards the demand of the shared bikes are temperature, year and season variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a Machine Learning algorithm which is used for supervised learning. It helps in predicting a dependent variable (target) based on the given independent/feature variable(s). The regression technique tends to establish a linear relationship between a target variable and the other given feature variables.

There are two types of linear regression- simple linear regression and multiple linear regression. **Simple linear regression** is used when a single independent variable is used to predict the value of the target variable.

Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable.

A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

3. What is Pearson's R?

Answer:

Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

Where -1 indicates the perfectly negatively correlated, and +1 shows perfectly positively correlated.

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

Answer:

Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units.

If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

Scaling helps in understanding the data better as it makes the variable comparable.

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

Answer:

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity a**s VIF = 1/(1-R2)**. This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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 are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.