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:

season, workingday, weathersit, weekday, yr, holiday, and mnth categorical variables are used to perform analysis. Following are the analysis results

Categorical Variables	Analysis
season	• Bike demand likely to be higher in summer and fall season. • Marketing campaigns can be planned around this period
yr	• Bookings increased in Year 2019 compared to previous year
mnth	• Booking is higher between June to October months.
holiday	• Ad-hoc (casual) customers prefers biking during holiday whereas registered users less biking during holiday
weekday	• bike usage is higher on working days by registered users whereas Ad-hoc (casual) users book bike less on weekday
workingday	• There is no much difference in the booking over weekend or weekdays. • Bike rental is high during weekday by registered users • There are bookings on working days by registered users and ad-hoc booking over weekend hence this information does not give much to identify potential customers.
weathersit	• Usage of the biked by Registered users is higher irrespective to weather condition. • Usage can be generalised as commuting to the workplace / school etc • Typically, common weather if high usages is clean/few clouds days

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

Answer:

During modelling, dummy variables are created to prepare data to represent categorical variable with 1 and 0 value. These dummy variables can cause multi-correlations hence to drop the base/reference category while creating dummy variables drop_first=True is used.

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. i.e. 0.63

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

Answer:

Validation of the assumptions by checking following

a) Linear relationship – It should be visible among independent and dependent variables.
b) Error distribution - Error terms are independent of each other, it should be normally distributed
c) Multicollinearity Check - There should be insignificant multicollinearity among variables.

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 the demand are:

a) Temperature ("weathersit") – This affects the Business positively e.g. Raining, Humidity, Windspeed and Cloudy impacts the Business.
b) Year ("Yr") – higher bookings compared to previous years
c) Season ("season") – this has major contribution in the demand of the bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

In summary, Linear Regression algorithm is a technique used for supervised learning to find linear relationship among variables. Algorithm is used if the labels are continuous, like the daily bike booking daily. The representation of linear regression is y = bx + c

It helps in predicting a dependent variable(target) based on the given independent variable(s). In Regression, graph between the variables which best fit the given data points is plotted. The machine learning model delivers predictions based on this data.

Regression graph shows a line or curve that connects through all the data points in such a way that the vertical distance between the data points and the regression line is minimum. Establish a linear relationship between a dependent and other independent variable.

Linear Regression - Simple linear regression and multiple linear regression

a) **Simple linear regression** - Single independent variable to predict the value of the target variable.
b) **Multiple Linear Regression** - Multiple independent variables to predict the numerical value of the target variable.

Regression line - the relationship between the dependent and independent variables.

Positive linear relationship: Dependent variable on the Y-axis along with the independent variable in the X-axis.

Negative linear relationship - Dependent variables value decreases with increase in independent variable value increase in X-axis

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four data sets several identical statistical properties to illustrate the fact. 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 like:
- Outliers should be removed while analysing the data.
- Descriptive statistics do not fully depict the data set in its entirety.

3. What is Pearson's R?

Answer:
Also known as Pearson's correlation coefficients, it measures the strength between the different variables and the relation with each other 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.

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

Answer:

It is a technique performed to bring all values to same magnitude to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

The scaling normalizes these varied datatypes to a particular data range 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 interpretation of variables and units of those variables are kept open collect as much as possible. 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 The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

Answer:

Infinite indicates a perfect correlation between two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2).

It shows 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:

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 is called Q-Q (quantile-quantile) plots. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions