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 :
From my analysis of various categorical variables using the box plots and bar plots I could infer the following

- The bike demand is high when weather is clear and few clouds. However demand is less in case of light snow and light rainfall.
- The booking (demand) increased drastically from 2018 to 2019. Demand was more in fall season
- From the months June (Jun) to October (Oct) is the period when bike demand is high. January (jan) is the lowest demand month.
- Bike demand (booking) is almost similar on working and non-working days
- Spring season seems to have attracted least bookings and fall season saw the highest bookings
- There is not much change in demand based on working or non-working days
- Demand (bookings) is comparatively less on holidays than on non-holidays

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

Answer:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the 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:
Variable **"temp"** 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? (3 marks)

Answer:
To validate the assumptions of Linear regression I have used the below 5 ways

- **Normality of error terms** : The error terms should be normally distributed. I have plotted a graph and found that it followed a normal distribution
- **Multicollinearity check** : There should be no or insignificant multicollinearity among the variables
- **Linear relationship** : visible linearity among the variables
- **Homoscedasticity** : There should be no visible pattern in residual values. This was validated by plotting a graph
- **Independence of residuals** : No auto-correlation. Durbin Watson value was 2.08 which signifies no auto correlation

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

Answer:

Top 3 features contributing significantly towards explaining the demand are

1. Temperature (temp)
2. weathersit(Light Snow, Mist Cloudy)
3. Year (yr)

General Subjective Questions

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

Answer :

Linear regression is a statistical method for modelling the linear relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the values of the independent variables.

The linear regression model assumes that the relationship between the dependent and independent variables is linear, which means that the change in the dependent variable is directly proportional to the change in the independent variable. This is represented by the following equation:

y = b0 + b1x1 + b2x2 + ... + bn*xn

Where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients or weights of the model. The coefficient b0 is the intercept, which is the value of y when all the independent variables are equal to 0. The coefficients b1, b2, ..., bn represent the slope of the line for each independent variable, which measures the effect of each independent variable on the dependent variable.

To fit a linear regression model, we need to find the optimal values for the coefficients b0, b1, b2, ..., bn that minimize the error between the predicted values and the actual values of the dependent variable. There are several methods for estimating the coefficients, such as ordinary least squares, gradient descent, and stochastic gradient descent.

Once the coefficients are estimated, we can use the fitted model to predict the value of the dependent variable for a given set of values of the independent variables. For example, if we have a model with one independent variable (x1) and we want to predict the value of y for a given value of x1, we can plug the value of x1 into the equation and solve for y.

Linear regression is a simple and widely used method for predicting the value of a continuous variable. It has several assumptions, such as the linearity of the relationship between the dependent and independent variables, the independence of the errors, and the homoscedasticity of the errors, which need to be satisfied in order to obtain accurate predictions.

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

Answer :

Anscombe's quartet is a set of four small datasets that were created in 1973 by the statistician Francis Anscombe to demonstrate the importance of visualizing data before analyzing it. The datasets are designed to have nearly identical statistical properties, but when they are plotted on a graph, they look very different from one another. This illustrates the fact that statistical measures, such as the mean and variance, do not always capture the underlying structure of a dataset and that it is important to visualize the data to get a better understanding of it.

The four datasets in Anscombe's quartet are:

1. A simple linear regression with slope 3 and intercept 2.
2. A simple linear regression with slope 3 and intercept 2, but with one outlier.
3. A quadratic curve with a single outlier.
4. A dataset that has a strong relationship between two variables, but also has two outliers that follow a different pattern.

3. What is Pearson’s R? (3 marks)

Answer:

Pearson's R (also known as Pearson's correlation coefficient, Pearson's product-moment correlation coefficient, or simply the correlation coefficient) is a measure of the linear correlation between two variables. It is a value between -1 and 1, where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.

The formula for Pearson's R is:

$$r = \sum(x - \bar{x})(y - \bar{y}) / \surd[\sum(x - \bar{x})^2 \sum(y - \bar{y})^2]$$

where x and y are the two variables being correlated, $\bar{x}$ is the mean of x, and $\bar{y}$ is the mean of y. The numerator of the formula calculates the covariance between x and y, and the denominator standardizes this value by dividing by the standard deviations of x and y.

Pearson's R is a widely used measure of correlation and is often used in statistical analysis to determine the strength and direction of the relationship between two variables. It is particularly useful when the relationship between the variables is linear, but it can also be used to identify nonlinear relationships.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is the process of transforming variables so that they can be compared on the same scale. It is often performed to improve the performance of machine learning models by putting the variables on the same scale.

There are two main types of scaling: normalization and standardization.

Normalization scales the variables so that they have values between 0 and 1. It is useful when the variables have different units of measurement or different scales, but you want to compare them. To normalize a variable, you subtract the minimum value from the variable and then divide the result by the range (maximum value minus minimum value).

Standardization scales the variables so that they have a mean of 0 and a standard deviation of 1. It is useful when you want to compare the variables to a normal distribution or when the variables have the same units of measurement but different means and standard deviations. To standardize a variable, you subtract the mean from the variable and then divide the result by the standard deviation.

In general, normalization is used when the data is skewed or the data has a large standard deviation, while standardization is used when the data is approximately normally distributed. However, both techniques can be used in different situations and it is often a matter of preference or what works best for a particular model.

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

(3 marks)
Answer :

The variance inflation factor (VIF) is a measure of multicollinearity in a multiple regression model. It is calculated by taking the ratio of the variance of the model with all the predictors to the variance of the model with a single predictor. A VIF value of 1 indicates no multicollinearity, while a value greater than 1 indicates increasing multicollinearity.

A VIF value of infinite indicates perfect multicollinearity, which occurs when there is a linear relationship between two or more of the predictors in the model. This can cause problems because it can lead to unstable and unreliable coefficient estimates, as well as make it difficult to assess the contribution of each predictor to the model.

Perfect multicollinearity usually occurs when one predictor can be exactly expressed as a linear combination of the other predictors. For example, if two predictors are perfectly correlated, then one of them can be expressed as a linear combination of the other. In this case, the VIF for one of the predictors would be infinite.

To avoid perfect multicollinearity, it is important to carefully examine the correlations between the predictors in your model and consider removing one of the correlated predictors if necessary. It is also important to keep in mind that multicollinearity can occur even when the VIF values are not infinite, so it is always a good idea to check the VIF values for all the predictors in your model.

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

(3 marks)

Answer :

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to check whether a sample of data comes from a specified distribution. It can be used to compare the distribution of a sample to a theoretical distribution, such as the normal distribution, or to compare the distributions of two different samples.

In linear regression, a Q-Q plot is often used to check whether the residuals (errors) of the model are normally distributed. This is important because many statistical tests and assumptions of linear regression assume that the residuals are normally distributed. If the residuals are not normally distributed, it may indicate that the model is not appropriate for the data or that there are underlying patterns in the data that the model is not capturing.

To create a Q-Q plot, the sample is sorted in ascending order and divided into equal quantiles. The quantiles of the sample are then plotted against the quantiles of the theoretical distribution. If the sample comes from the theoretical distribution, the points on the Q-Q plot will lie on a straight line. Deviations from this line can indicate that the sample does not come from the specified distribution.

In summary, a Q-Q plot is a useful tool for checking the normality of the residuals in a linear regression model, which is important for evaluating the assumptions and reliability of the model.