Assignment-based Subjective Questions Answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Categorical variables are variables that represent groupings of some kind. They are sometimes recorded as numbers, but the numbers represent categories rather than actual amounts of things.

In order to determine the effect of categorical variables on the dependent variable, you can use regression analysis. Regression analysis is a statistical technique that helps to identify the relationship between a dependent variable and one or more independent variables.

To perform regression, we will need to convert our categorical variables into quantitative data. Once we have done this, we can use regression to determine the effect of our categorical variables on the dependent variable.

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

It is important to use drop_first=True during dummy variable creation to avoid the dummy variable trap .

The dummy variable trap occurs when we include all the dummy variables in a regression model, which can lead to multicollinearity and inaccurate predictions .

Multicollinearity is a phenomenon where two or more independent variables in a regression model are highly correlated with each other.

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

Registered has the highest correlation wit the target variable.

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

training set?

After building a linear regression model on the training set, it is important to validate the assumptions of linear regression to ensure that the model is appropriate for the data and that the predictions are accurate . Here are some common assumptions of linear regression:

Linearity: The relationship between the dependent and independent variables should be linear. This can be checked by examining a scatter plot of the data.

Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables. This can be checked by examining a residual plot of the data.

Normality: The errors should be normally distributed. This can be checked by examining a histogram or Q-Q plot of the residuals.

Independence: The errors should be independent of each other. This can be checked by examining a residual plot of the data.

To validate these assumptions, we can use various statistical tests and visualizations such as scatter plots, residual plots, histograms, and Q-Q plots . If any of these assumptions are violated, we may need to transform the data or use a different type of regression model .

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

Month, working day and registered explaining the demand of the shared bikes because most of the user are registered and using bike by time month and working day.

General Subjective Questions Answers

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

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best linear relationship between the dependent variable and the independent variables that can be used to make predictions about the dependent variable.

The linear regression algorithm works by fitting a line to the data that minimizes the sum of the squared errors between the predicted values and the actual values of the dependent variable. The line is defined by an intercept and a slope, which are estimated from the data using a method called ordinary least squares (OLS).

OLS estimates the intercept and slope of the line by minimizing the sum of the squared errors between the predicted values and the actual values of the dependent variable. The squared errors are calculated as the difference between the predicted value and the actual value of the dependent variable, squared.

Once the intercept and slope have been estimated, they can be used to make predictions about the dependent variable for new values of the independent variables. The quality of these predictions can be evaluated using statistical measures such as R-squared, which represents the proportion of variance in the dependent variable that is explained by the independent variables.

Linear regression can be used for both simple linear regression, which involves only one independent variable, and multiple linear regression, which involves two or more independent variables. In addition, linear regression can be extended to include non-linear relationships between the dependent and independent variables by using polynomial regression or other non-linear regression techniques.

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

Anscombe’s quartet is a set of four datasets that have nearly identical statistical properties, but are visually distinct. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and checking assumptions before performing statistical analyses .

Each dataset in the quartet consists of 11 (x,y) pairs, and has the same mean, variance, correlation coefficient, and linear regression line as the other datasets. However, when plotted, each dataset has a different pattern of points that highlights the importance of visualizing data before drawing conclusions .

The first dataset is a simple linear relationship between x and y. The second dataset is a non-linear relationship between x and y. The third dataset is a perfect example of how outliers can affect the correlation coefficient. The fourth dataset is an example of how a single outlier can have a significant impact on the linear regression line .

The quartet demonstrates that statistical measures such as mean, variance, correlation coefficient, and linear regression line can be misleading if we do not visualize the data first. Therefore, it is important to always visualize data before performing statistical analyses to ensure that our conclusions are accurate.

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

Pearson's R is a statistical measure that represents the strength and direction of the linear relationship between two variables . It is also known as the Pearson correlation coefficient or Pearson product-moment correlation coefficient.

Pearson's R ranges from -1 to 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation . A positive correlation means that as one variable increases, the other variable also increases, while a negative correlation means that as one variable increases, the other variable decreases .

Pearson's R is commonly used in linear regression to evaluate the relationship between the dependent and independent variables. It can also be used to identify outliers and influential observations in a dataset .

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

and standardized scaling? (3 marks)

Scaling is a data preprocessing technique used to standardize the range of features in a dataset. It involves transforming the values of the features so that they have a similar scale, which can help improve the performance of machine learning models.

Scaling is performed because machine learning models are sensitive to the scale of the features in the dataset. If the features have different scales, then some features may dominate others, leading to poor model performance. Scaling helps to standardize the range of features so that they have a similar scale and can be compared on an equal footing.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized scaling scales the values of the features to a fixed range between 0 and 1. This can be done using the formula:

x_norm = (x - min(x)) / (max(x) - min(x)

where x is the original value of the feature, min(x) is the minimum value of the feature, and max(x) is the maximum value of the feature.

Standardized scaling scales the values of the features so that they have a mean of zero and a standard deviation of one. This can be done using the formula:

x_std = (x - mean(x)) / std(x)

where x is the original value of the feature, mean(x) is the mean value of the feature, and std(x) is the standard deviation of the feature.

The main difference between normalized scaling and standardized scaling is that normalized scaling scales the values of the features to a fixed range between 0 and 1, while standardized scaling scales the values of the features so that they have a mean of zero and a standard deviation of one.

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

(3 marks)

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression model. It measures how much the variance of the estimated regression coefficient is increased due to multicollinearity in the model .

The VIF can be calculated for each independent variable in the model, and a value greater than 1 indicates that there is some degree of multicollinearity between that variable and the other independent variables in the model .

If the VIF is infinite, it means that there is perfect multicollinearity between the independent variables in the model. This occurs when one or more independent variables can be expressed as a linear combination of the other independent variables in the model .

Perfect multicollinearity can cause problems in a regression model because it makes it impossible to estimate the regression coefficients using ordinary least squares (OLS) . Therefore, it is important to identify and remove any independent variables that exhibit perfect multicollinearity before fitting a regression model.

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

A Q-Q plot (quantile-quantile plot) is a graphical technique used to compare the distribution of a sample to a known distribution, such as the normal distribution. The Q-Q plot is created by plotting the quantiles of the sample against the quantiles of the known distribution. If the sample is normally distributed, the points on the Q-Q plot will fall along a straight line.

In linear regression, Q-Q plots are used to check the normality assumption of the residuals. The residuals are the differences between the predicted values and the actual values of the dependent variable. If the residuals are normally distributed, then the Q-Q plot of the residuals will be approximately linear.

The normality assumption is important in linear regression because it affects the accuracy of the statistical tests used to evaluate the significance of the regression coefficients. If the residuals are not normally distributed, then the statistical tests may be inaccurate and lead to incorrect conclusions about the significance of the regression coefficients.

Therefore, Q-Q plots are an important tool for checking the normality assumption in linear regression and ensuring that our conclusions are accurate.