

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: The demand for bike is less in the month of spring when compared to other seasons.
The demand for bike increased in the year 2019 when compared to the year 2018.

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

Answer: When dummy variables are created, drop_first helps to reduce the extra column created. If there are K columns, only K-1 columns are required for analysis.

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

Answer: Looking at the pair plot 'temp' has the highest correlation of 0.63

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

Answer: Validating assumptions using the following.

1. **Linearity:** By checking if the relationship between X and Y is Linear.
 2. **Homoscedastic:** By checking if the variance of residual is the same as value X
 3. **Independence:** By checking if the observations are independent of each other and that they do not follow any pattern.
 4. **Normality:** And for any value of X, Y is normally distributed.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top 3 features contributing significantly are:

1. weathersit_cloudy
2. yr_2019
3. temp

General Subjective Questions

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

Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between input and the target variable.

Regression analysis is used for three types of applications:

1. Finding out the effect of Input variables on Target variable.
2. Finding out the change in Target variable with respect to one or more input variable.
3. To find out upcoming trends

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

An example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot it over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(\text{slope}) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$a(\text{intercept}) = \frac{n \sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y -intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

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

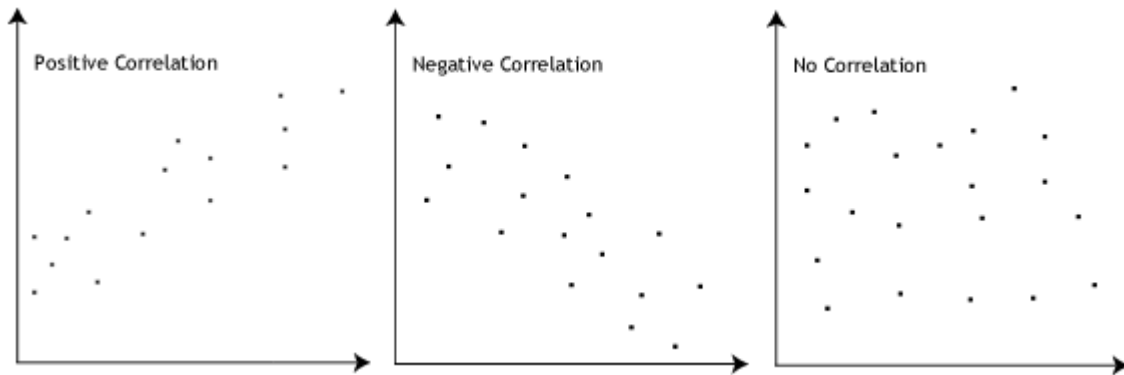
Answer: Anscombe's Quarter is defined as a group of data sets which are identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have different distributions and appear differently when plotted on scatter plots.

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

Answer: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1 .

The Pearson's correlation coefficient varies between -1 and $+1$ where:

- $r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- $r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- $r = 0$ means there is no linear association
- $r > 0 < 0.5$ means there is a weak association
- $r > 0.5 < 0.8$ means there is a moderate association
- $r > 0.8$ means there is a strong association



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r = correlation coefficient
- x_i = values of the x-variable in a sample
- \bar{x} = mean of the values of the x-variable
- y_i = values of the y-variable in a sample
- \bar{y} = mean of the values of the y-variable

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

Answer: Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling used? Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

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

Answer: If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2) = \infty$. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.