

## 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?

Ans-

- a. Spring session is having impact from the categorical value.
  - b. Working days is having the impact on the model.
  - c. Mist is also affecting Bike rental.
2. Why is it important to use `drop_first=True` during dummy variable creation?

Ans- it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans- temp and atemp are highly correlated.

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

Ans-

- a. based on Variance Inflation Factor method evaluate the result set. All the variables should be less than 5.
  - b. Residual analysis – The error terms are normally distributed.
  - c. The training and testing accuracy should be nearly equal hence there is no Overfit/Underfit situation.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
- a. Spring season: -0.6760
  - b. Temperature: 0.4029
  - c. Mist: -0.3619
  - d. Sun: 0.4451

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans-

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior 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.

Example for that can be 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 behavior 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.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

## 3. What is Pearson's R?

the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation,[1] is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and -1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

- $r = 1$  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
  - $r > 0 < 5$  means there is a weak association.
  - $r > 5 < 8$  means there is a moderate association.
  - $r > 8$  means there is a strong association.
- ## 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization usually means to scale a variable to have a value between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. In this way, If there is perfect correlation, then  $VIF = \infty$ . A large value of VIF indicates that there is a correlation between the variables.

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

It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.