

Assignment-based Subjective Questions

- 1. Why is it important to use `drop_first=True` during dummy variable creation?**

Ans. `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.

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

Ans. Looking at the pair plot we observe that `atemp` and `temp` have the highest correlation i.e. 0.99 which is almost 1.

- 3. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans. From visualizing categorical variables we were able to identify the dummy variables. Which are : `mnth`, `season`, `weathersit` and `weekday`

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

Ans. Linear Regression is used as a base model here despite knowing that Linear Regression wouldn't work here. We fit the model to use 'error terms' of this model as a reference to check how good the other models are performing.

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

Ans. We can see the demand for bikes depends mainly on below variables:

yr , holiday ,Spring, Mist_Cloudy, Light rain_Light
snow_Thunderstorm, 3 ,5 ,6, 8, 9, sunday, 7, 10

- Demands increases in the month of 3, 5, 6, 8 ,9, 7 , 10 and yr
- Demand decreases if it is holiday , Spring, Light rain_Light
snow_Thunderstorm, Mist_cloudy, Sunday

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 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.

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

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R?

Ans. 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

Pearson r 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?

Ans. 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.

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.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called <code>MinMaxScaler</code> for Normalization.	Scikit-Learn provides a transformer called <code>StandardScaler</code> for standardization.
6.	This transformation squishes the n-dimensional data	It translates the data to the mean vector of original data to the origin

S.NO.	Normalisation	Standardisation
	into an n-dimensional unit hypercube.	and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Ans. 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 lead to $1/(1-R^2)$ infinity. 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.

Ans. 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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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.