

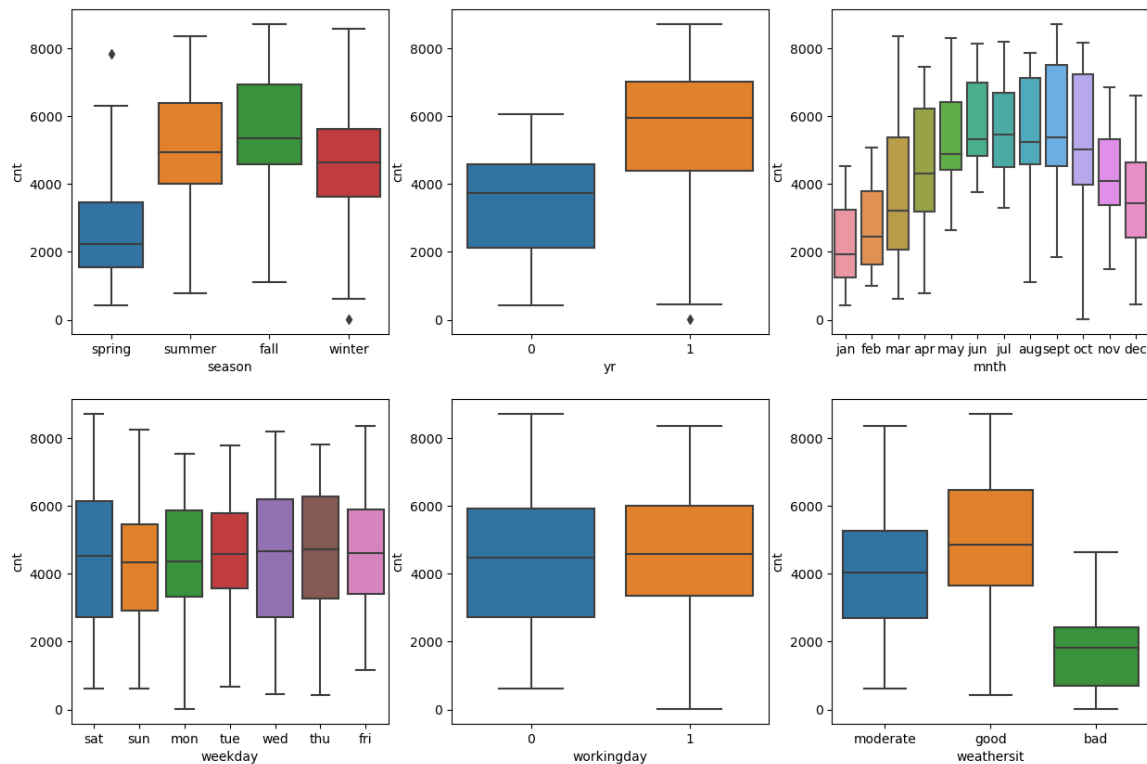
LINEAR REGRESSION ASSIGNMENT

Assignment-based Subjective Questions:

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

Answer: The demand of bike is less in the month of spring when compared with other seasons.

There are a couple of categorical variables namely season, mnth, yr, weekday, workingday and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same. These variables are visualized using bar plot and Box plot both.



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

Answer: The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence `drop_first=True` is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

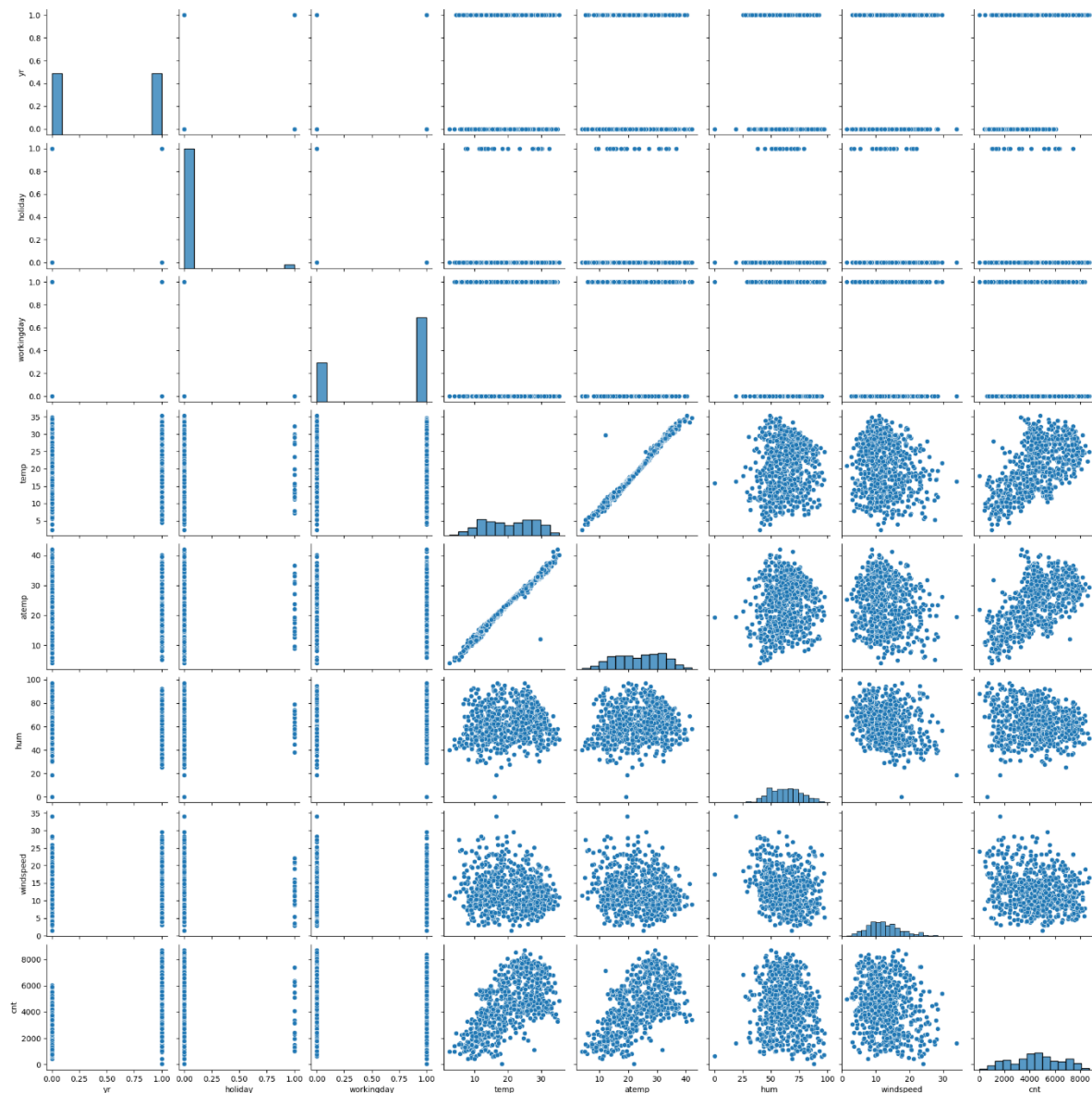
E.g.: If there are 3 levels, the `drop_first` will drop the first column.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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 A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

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

Answer: The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.



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

Answer: I have validated the assumption of Linear Regression Model based on below 5 assumptions

–

- Normality of error terms 'o' Error terms should be normally distributed.
- Multicollinearity check 'o' There should be insignificant multicollinearity among variables.
- Linear relationship validation 'o' Linearity should be visible among variables
- Homoscedasticity 'o' There should be no visible pattern in residual values.
- Independence of residuals 'o' No auto-correlation.

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

Answer: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- Temperature
- Season
- Year

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer: 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

Question 2. Explain the Anscombe's quartet in detail.

Answer: 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 analysing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Question 3. What is Pearson's R?

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)

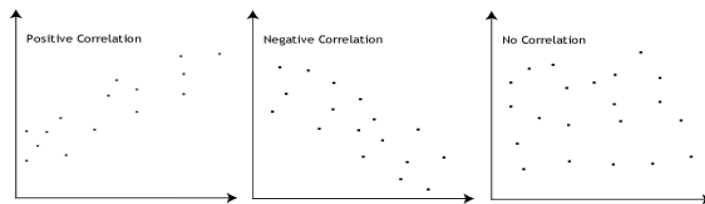
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 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

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

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

Answer: 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/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.

sklearn. preprocessing. MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

sklearn. preprocessing. scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

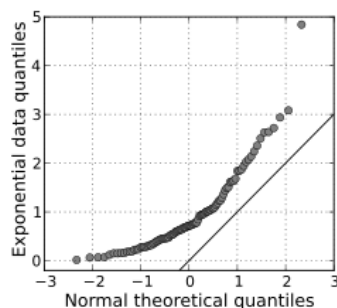
Answer: If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which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

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

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 showing the 45-degree 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.