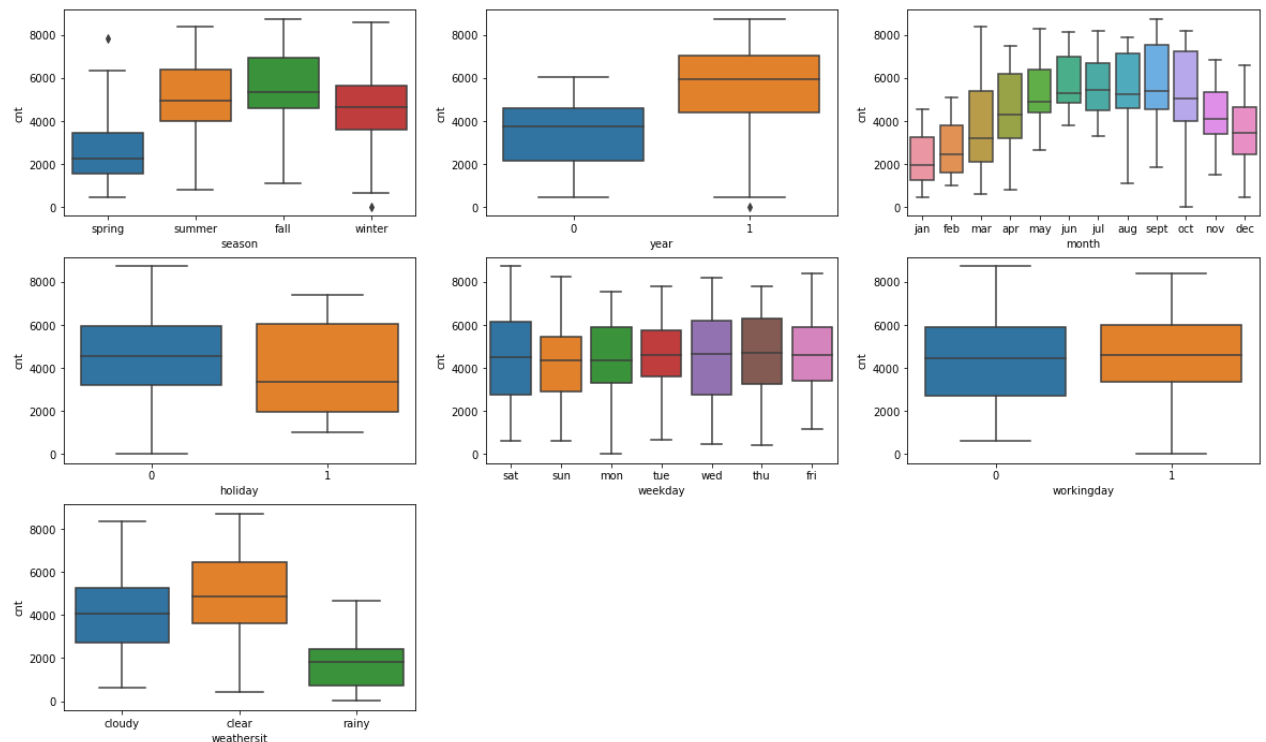


Assignment-based Subjective Questions

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

Ans) Categorical variables like Season, year, month, weathersit, weekday and holiday has impact on dependent variable cnt. Below figure shows the correlation among them.



2. 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. Also, it reduces the correlations created among dummy variables.

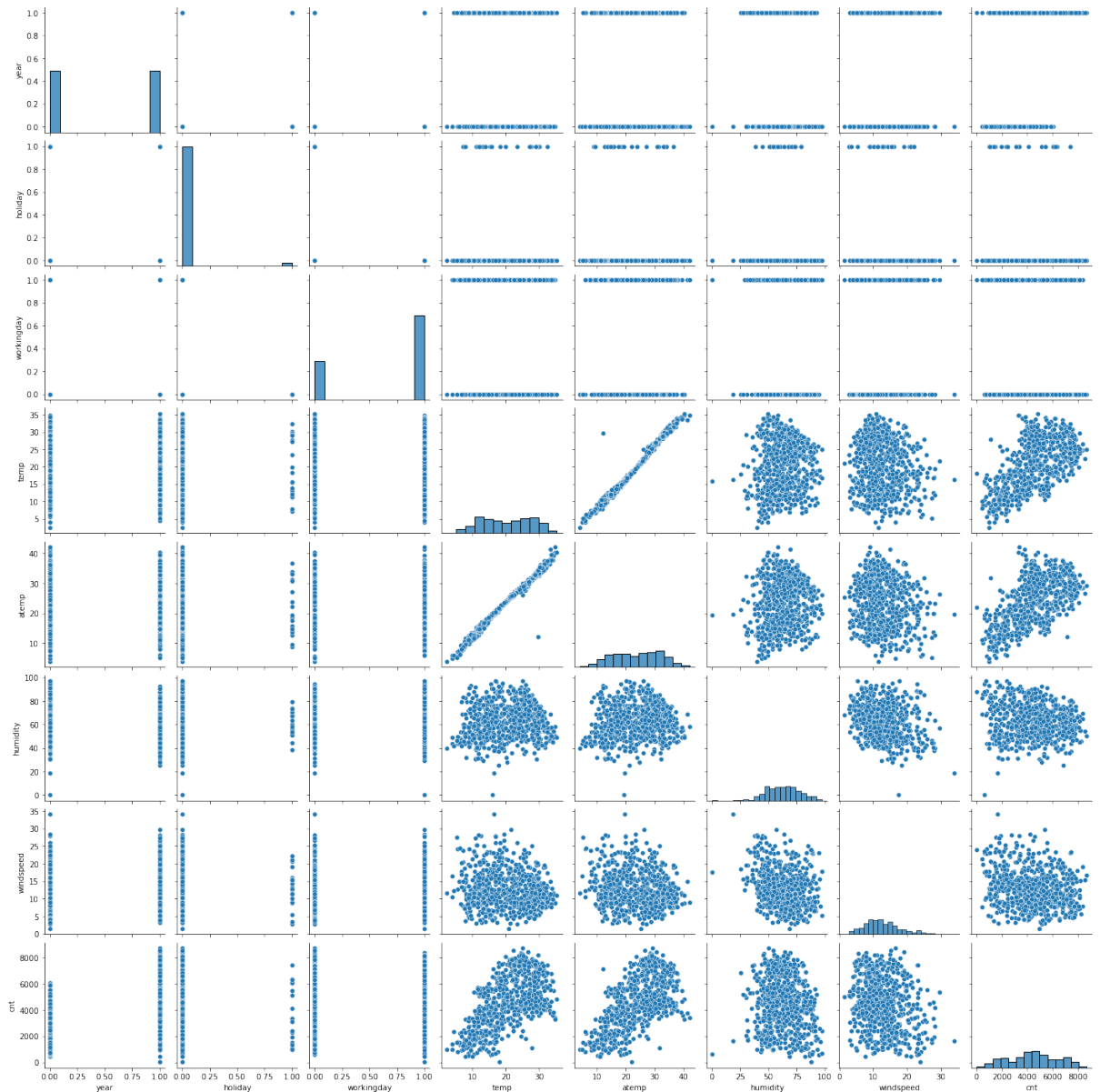
Let's say we have 3 types of fruit as below and we want to create dummy variables for the column.

- 1.Apple
- 2.Banana
- 3.Orange

So for this example suppose a variable is not Apple and Banana that means it's an Orange . so with 2 variables we are able to identify .

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

Ans) Based on the graph its temp and atemp.

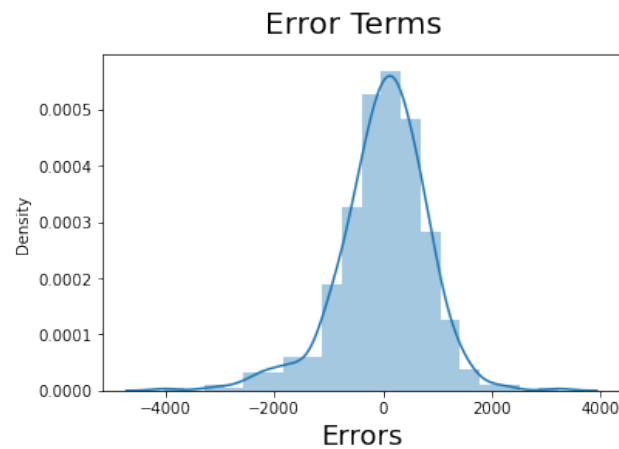


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans) I have validated the assumption of Linear Regression Model based on below assumptions –

- Normality of error terms

- Error terms should be normally distributed and mean as zero.



- Linear relationship between Predictor variable and target variable
 - All predictor variables p-value is less than 0.05 and R square value on train data set is 0.826 and adjusted R square is 0.823 .Hence all variance in data is being explained by these predictor variables.
- Error terms are independent of each other:
 - Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

Ans) Top 3 features significantly contributing towards demand of shared bikes are:

1. Temp
2. Year
3. weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans) Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase

or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

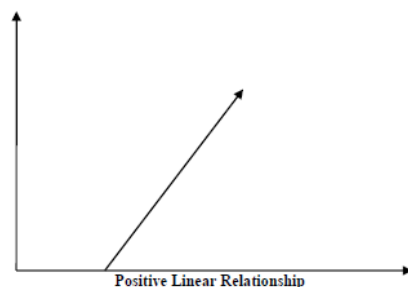
m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. (If $X = 0$, Y would be equal to c.)

Linear relationship can be positive or negative in nature as explained below–

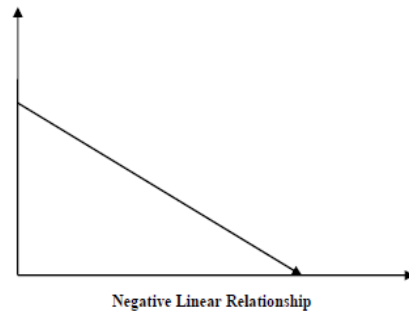
- Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



- Negative Linear relationship:

A linear relationship will be called negative if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Linear regression can be divided into two types –

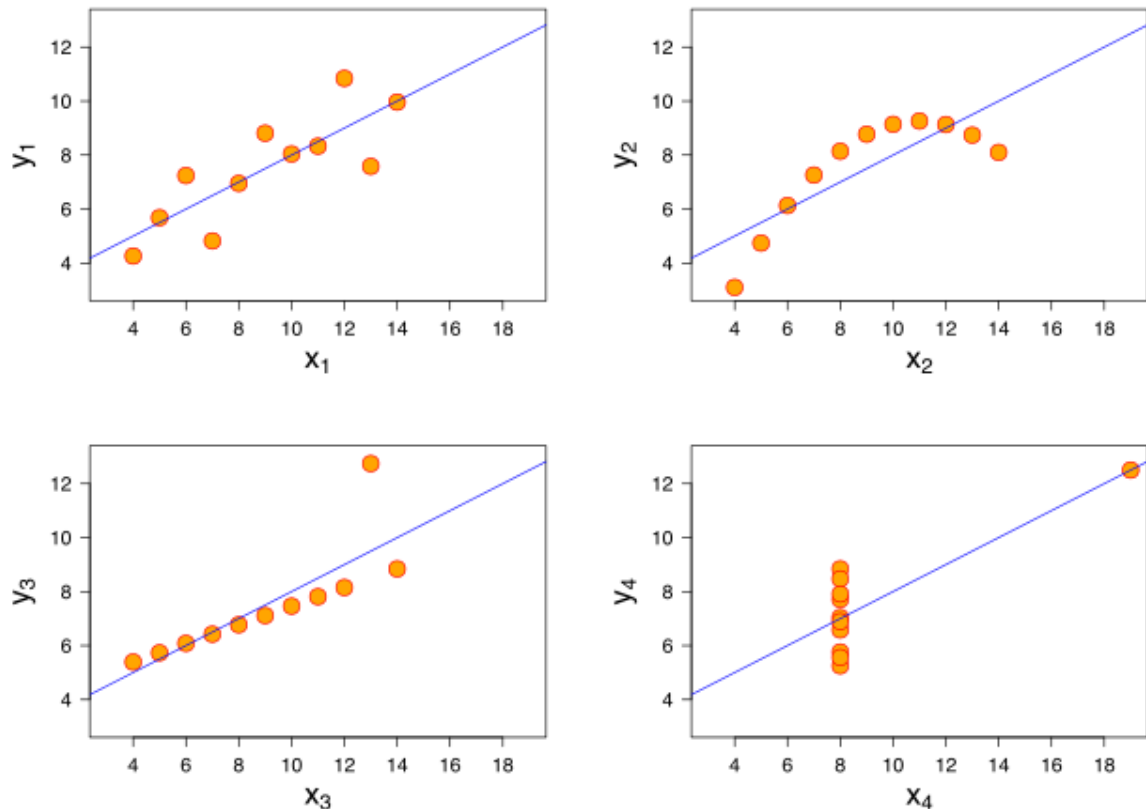
- Simple Linear Regression
- Multiple Linear Regression

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model

1. Multi-collinearity –
Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2. Auto-correlation –
Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
3. Relationship between variables –
Linear regression model assumes that the relationship between response and feature variables must be linear.
4. Normality of error terms –
Error terms should be normally distributed
5. Homoscedasticity –
There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

Ans) 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.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

- 1) 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 .
- 2) 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.
- 3) 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.
- 4) 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.

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of $x : s_x^2$	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of $y : s_y^2$	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

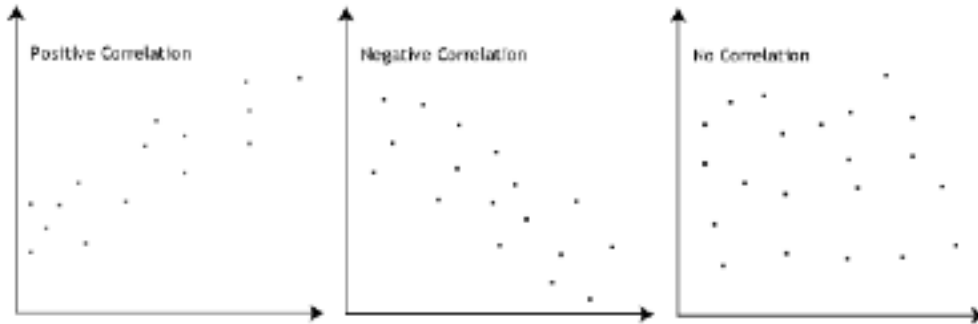
Anscombe's quartet

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

3. What is Pearson's R?

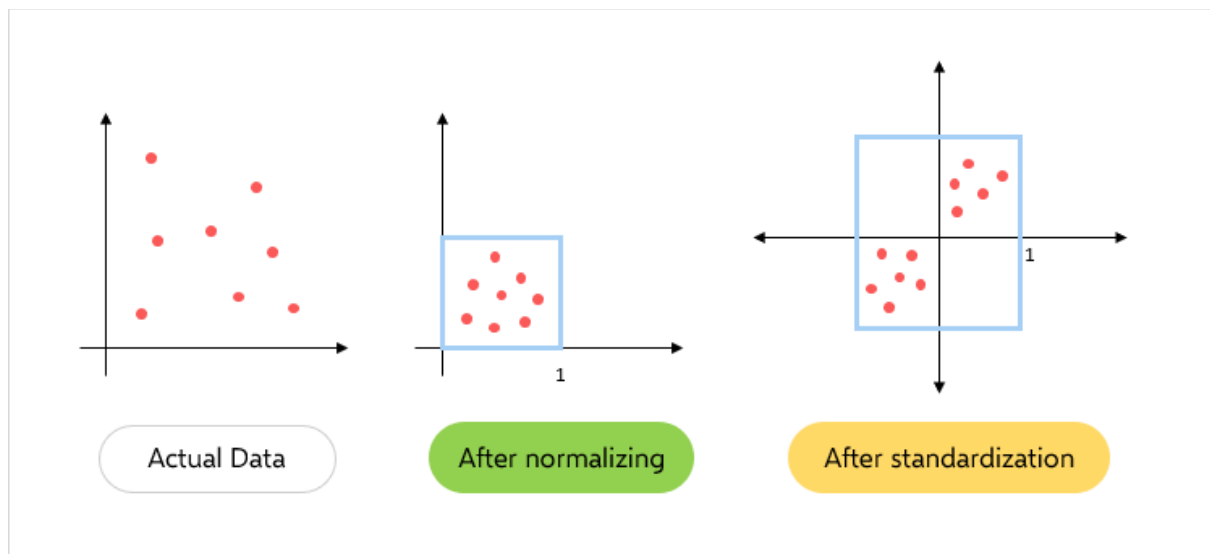
Ans) The Pearson correlation coefficient (**R**) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Ans) 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.



Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between $[0, 1]$ or $[-1, 1]$.	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
--	--

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 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).

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

Ans) Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots

Importance of QQ Plot in Linear Regression :

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also

- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution

- If both datasets have common location and common scale

- If both datasets have similar type of distribution shape

- If both datasets have tail behavior