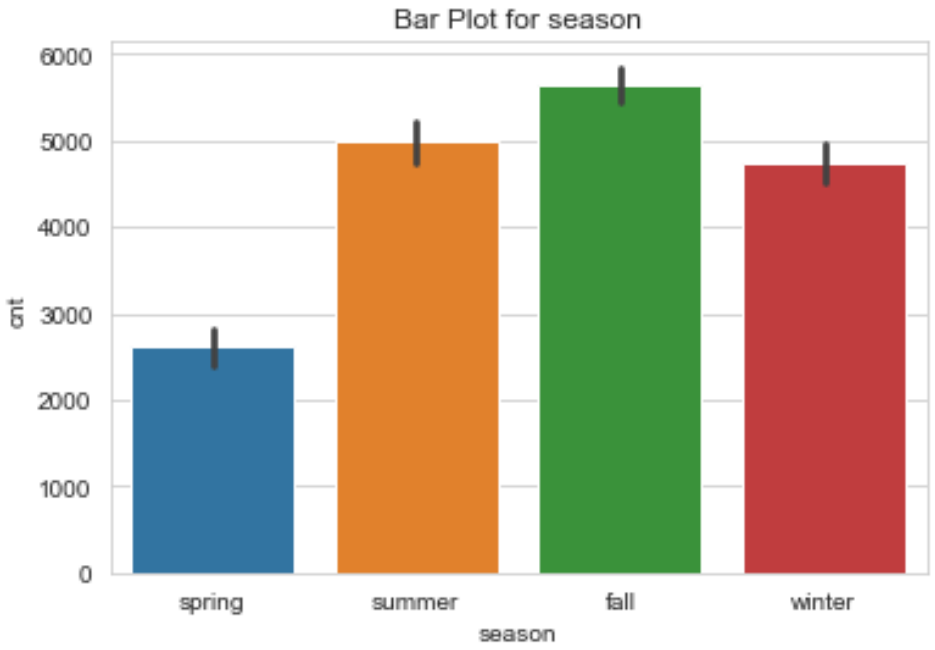
SOLUTIONS TO ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

By Vibhav Mann

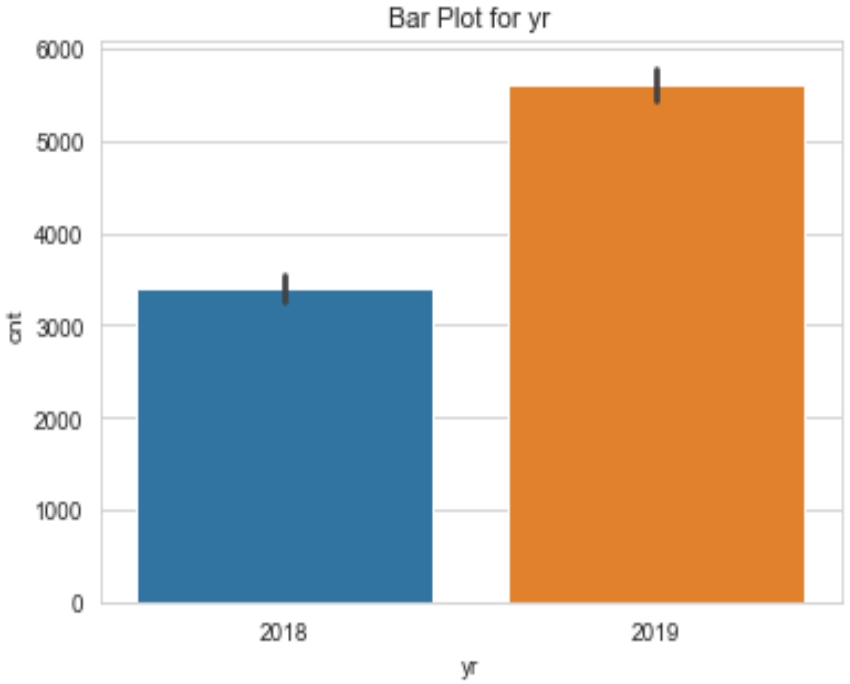
1. FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOU INFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE?

SEASON COLUMN



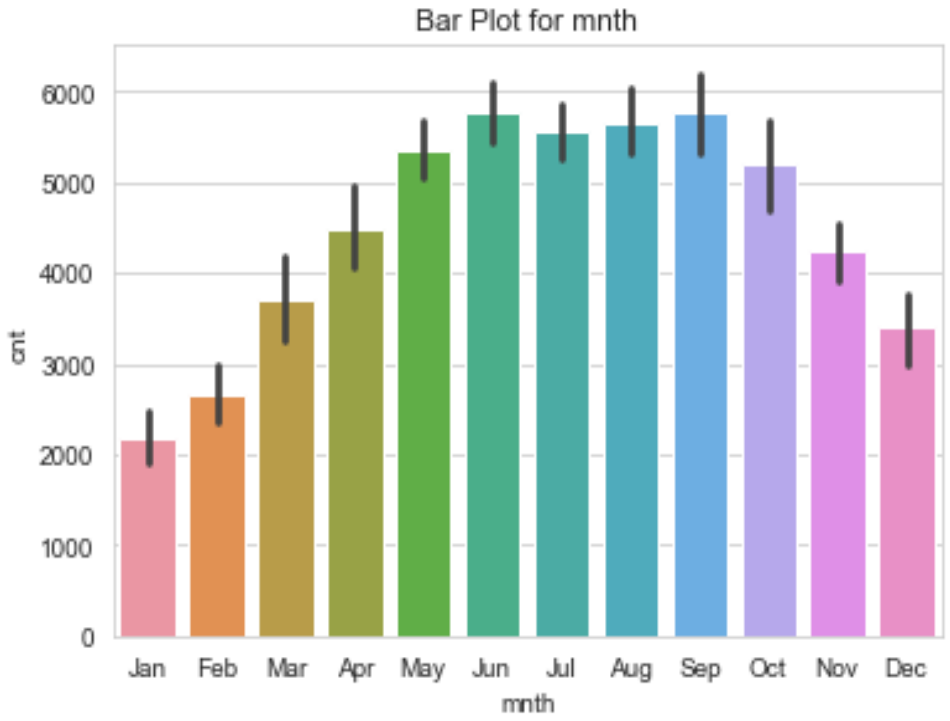
1. Looking at the bar plot for the season variable, we can say that the maximum rental count of the bikes was observed during the fall season followed by the summer season, then winter and lastly by the spring season.

YR COLUMN



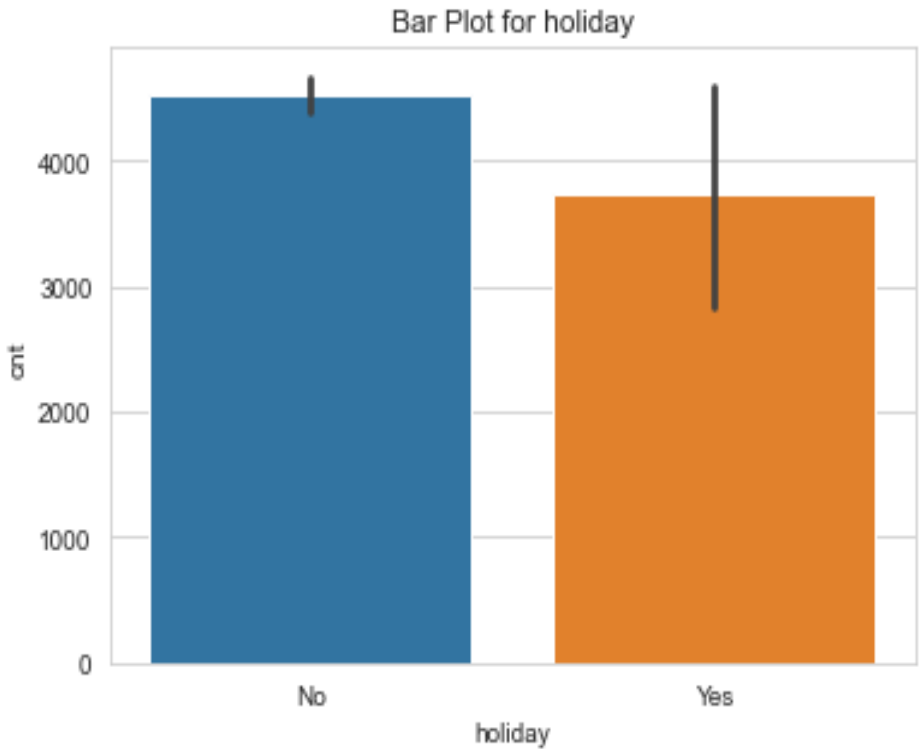
2. Looking at the bar plot for the yr column, we can see that more number of bikes were rented in the year 2019 as compared to 2018 and hence, the strategies were working in favour of the company. The rental count increased from almost 3500 to 5500 which is a good increase in the count.

MTNH COLUMN



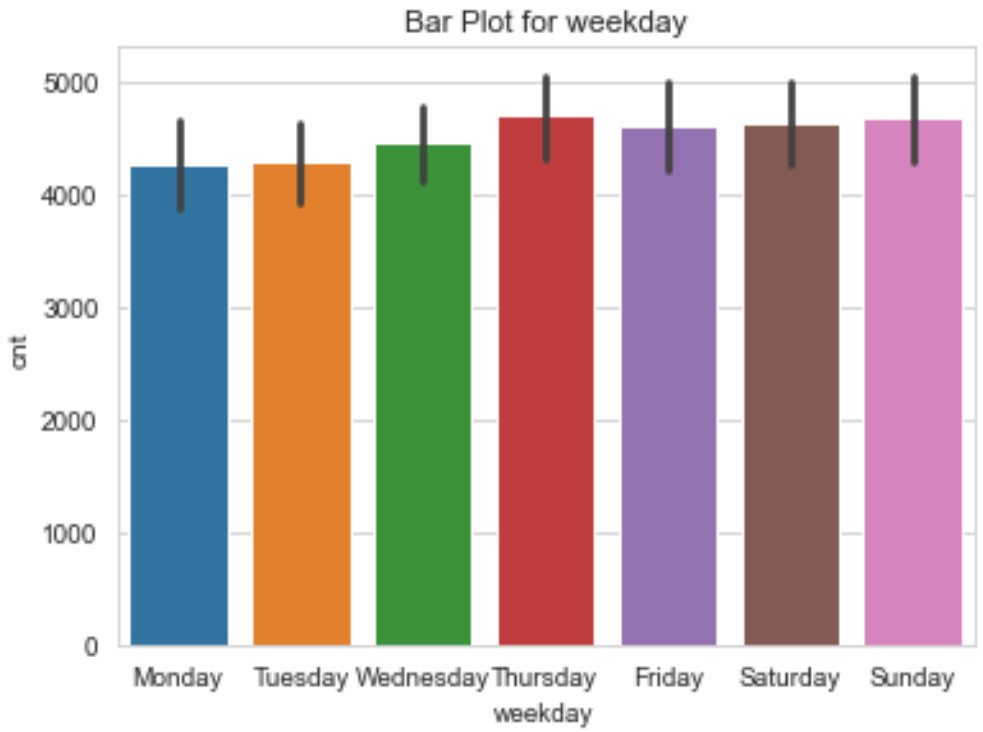
3. Looking at the bar plot for the mnth column, the sales of the bikes seem to be highest in the months of June and September. This is followed by August, July, May, October, April, November, March, December, February. The sales seem to be lowest in the month of January which is expected as winters are at their peak during the months of January and February.

HOLIDAY COLUMN



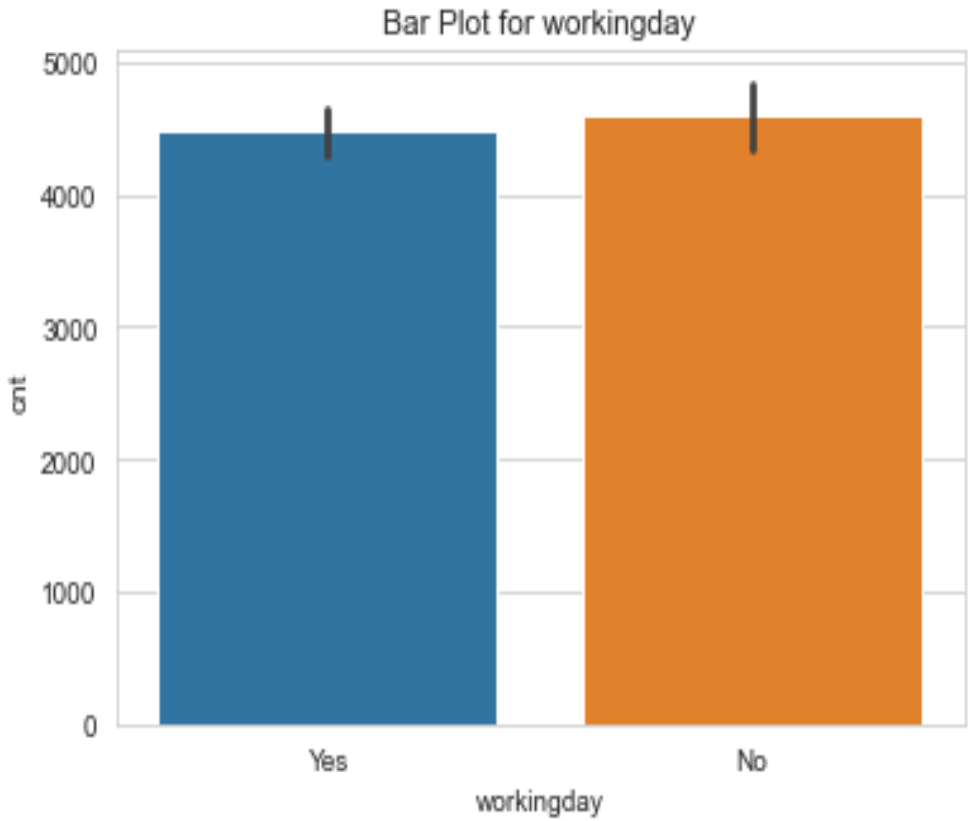
4. Looking at the bar plot for the holiday column, we can say that the rental counts for the bikes were lesser when the day was a holiday and more for day that was not a holiday, which is also generally expected.

WEEKDAY COLUMN



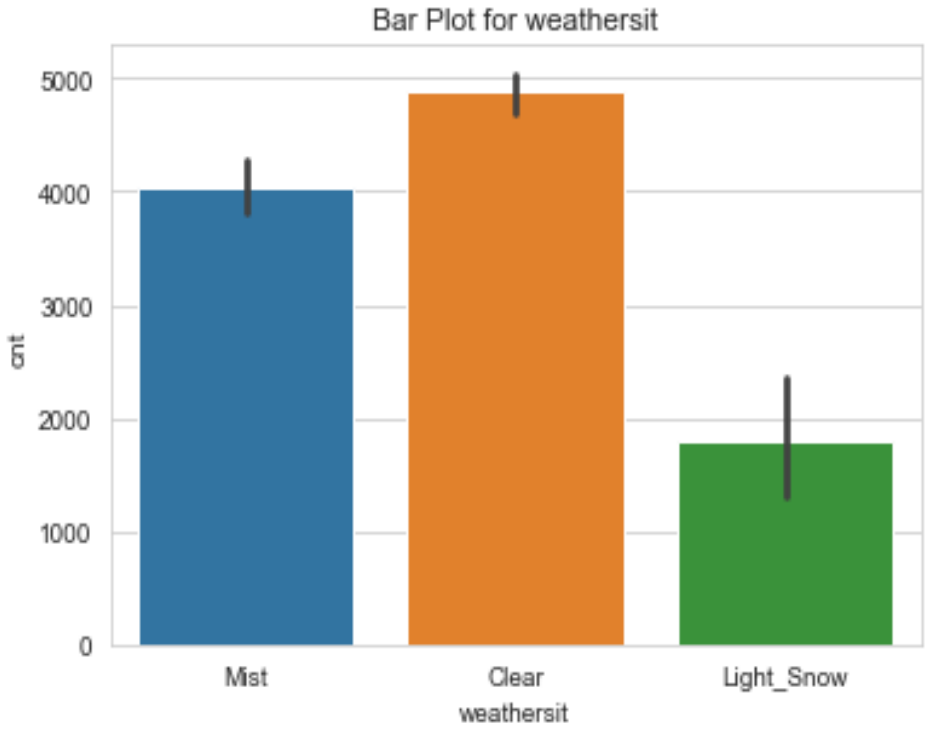
5. Looking at the sales for the weekday column, we can say that the rental count of the bikes seemed to be the most on Thursdays, followed closely by Sundays, Saturdays and Fridays, then by Wednesdays and the least on Mondays and Tuesdays.

WORKINGDAY COLUMN



6. Looking at the bar plot for the workingday column, we can say that the rental count seem to be almost the same for the day being a working day or not but slightly more for day not being a working day.

WEATHERSIT COLUMN



7. Looking at the weathersit column, the first thing to notice is that there in no bike sale during the Heavy_Rain weather which is as expected. There could also be very minimal sales but in the dataset provided, there are none. The second thing that we can notice from the bar plot of the weathersit column is that the maximum sales is during the Clear weather, followed by the Mist weather and lastly, minimum in the Light_Snow weather.

2. WHY IS IT IMPORTANT TO USE **DROP_FIRST=TRUE** DURING DUMMY VARIABLE CREATION?

The function of drop_first=True is basically to drop the first column. It is used while creating dummy variables. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted. The reason why we use drop_first is because during the creation of dummy variables, one of the dummy variables can be explained by the other. Let me explain this better using an example. Say, we have genders as an attribute in a data set and we need to create dummy variables for the gender column. On creating we will assign 0 to either male or female and 1 to the one not selected earlier. Let us suppose we assign 0 to male and 1 to female. This means that all the rows having the gender attribute as female would be 1 and all the rows having gender as male would be 0. Hence, intuitively thinking, do we need the male dummies column? Since, the female attributes are 1, wherever the attribute is absent/0 we can automatically assume that the gender for that entry is male. This is why we drop the first column using the drop_first method for the purpose of achieving optimality.

3. LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

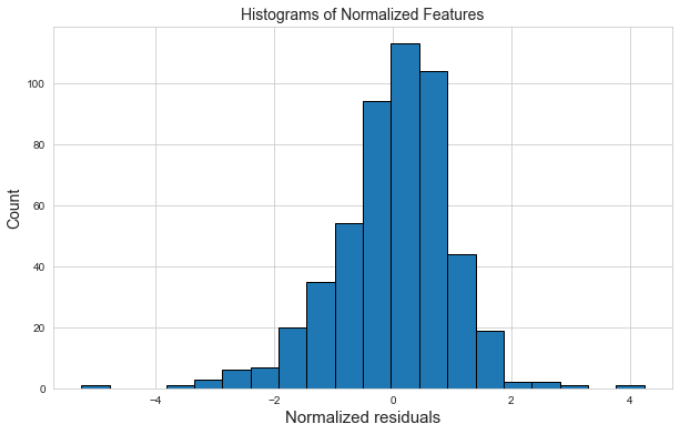
	temp	atemp	hum	windspeed	casual	registered	cnt
temp	1.000000	0.991721	0.123278	-0.155395	0.542052	0.538877	0.626791
atemp	0.991721	1.000000	0.136447	-0.182099	0.542862	0.543094	0.630475
hum	0.123278	0.136447	1.000000	-0.229375	-0.088424	-0.112580	-0.122119
windspeed	-0.155395	-0.182099	-0.229375	1.000000	-0.161548	-0.214596	-0.230287
casual	0.542052	0.542862	-0.088424	-0.161548	1.000000	0.392534	0.671825
registered	0.538877	0.543094	-0.112580	-0.214596	0.392534	1.000000	0.944973
cnt	0.626791	0.630475	-0.122119	-0.230287	0.671825	0.944973	1.000000

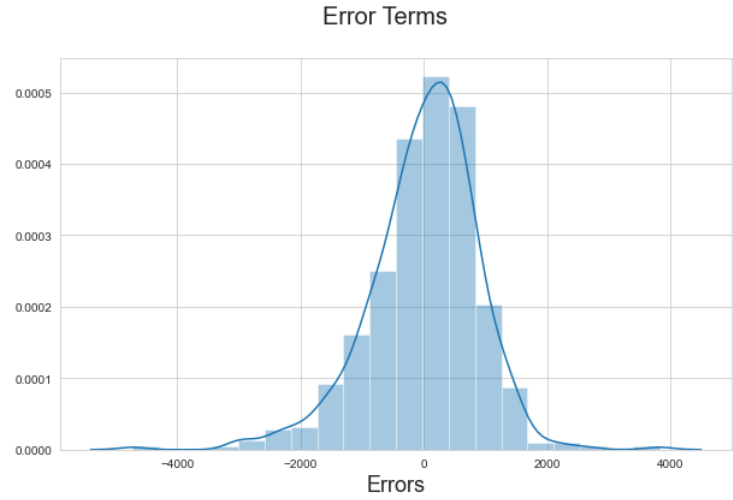
The ‘atemp’ variable showed the maximum correlation with the ‘cnt’ variable having a correlation score of 0.99. Since, the pair plots cannot be posted in here, I have added the correlation table to compliment my answer.

4. HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

There exist four assumptions to Linear Regression. These are as follows:

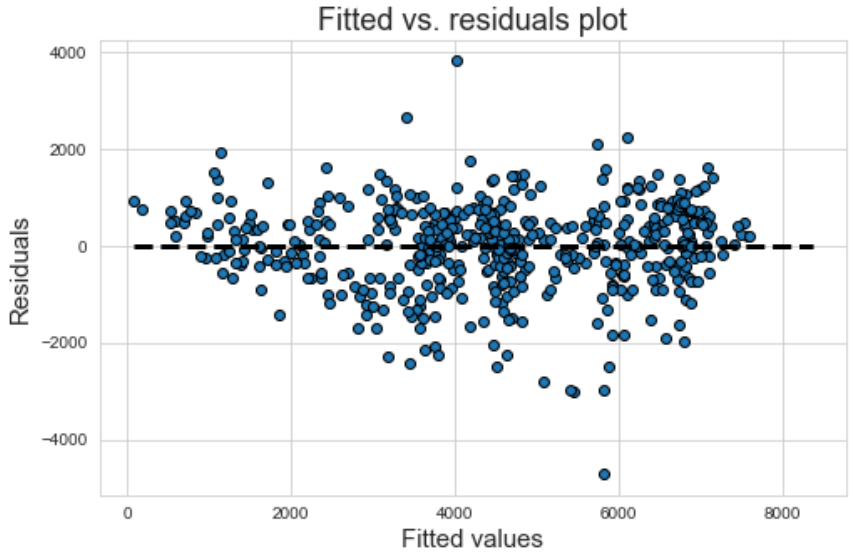
- Linear Relationship between dependent and independent variable. This assumption is more valid in the case of Simple Linear Regression and not in Multiple Linear Regression. This step can be checked by plotting a pair plot of all the variables in the given data set.
- The second assumption is that error terms are normally distributed. For this I have created two histograms in my Jupyter Notebook and posted them here as well. This shows that the error terms are normally distributed.





- The third assumption is that the error terms are independent of each other. To check this I have plotted a **Residuals vs predicting variables Plots** in my Jupyter notebook in line 127. The plot is too long and hence, cannot be fitted here. Residual plots show some bit of clustering but overall the assumptions linearity and independence seem to hold because the distribution seem random around the 0 axis.

- The final assumption is that Error terms have a constant variance. To check this I have plotted a Fitted vs Residuals scatter plot in the Jupter notebook which I'm attaching below as well. When we plot the fitted response values (as per the model) vs. the residuals, we clearly observe that the variance of the residuals increases with response variable magnitude. Therefore, the problem does not respect homoscedasticity and some kind of variable transformation may be needed to improve model quality.



5. BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

Looking at the coefficient values of the final model ie., model no 19 in my Jupyter notebook, the following three variables have been arranged as per their importance in an order of highest to lowest :

1. yr _2019
2. mnth_Sep
3. weathersit_Light_Snow

6. EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

There are mainly three types of Machine Learning Algorithms :

1. Classification
2. Clustering
3. Regression

These three types of algorithms are further classified into 2 types of methods, namely:

1. Supervised Learning (Consists of Classification and Regression)
2. Unsupervised Learning (Consists of Clustering)

Linear Regression Algorithm is, basically, a machine learning algorithm based on supervised learning methods. In Linear Regression(LR), labels are present for the data and they are continuous in nature. A very basic example of an LR would be predicting the score of a particular student using predictive analytics. Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.There are two types of Linear Regression :

- Simple Linear Regression (SLR) – Simple linear regression is used to estimate the relationship between two quantitative variables. SLR is a type of linear regression having a single independent variable. Example is the relationship between rainfall and soil erosion. The formula for simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

B_0 is the **intercept,** the predicted value of **y** when the **x** is 0.

B_1 is the regression coefficient – how much we expect **y** to change as **x** increases.

x is the independent variable (the variable we expect is influencing **y**).

e is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Multiple Linear Regression – Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. MLR is a type of linear regression having multiple independent variables. An example would be how rainfall, temperature, and amount of fertilizer added affect crop growth. The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

y = the predicted value of the dependent variable

$\mathbf{B_0}$ = the y-intercept (value of y when all other parameters are set to 0)

$\mathbf{B_1X_1}$= the regression coefficient (B_1) of the first independent variable ($\mathbf{X_1}$) (a.k.a. the effect that increasing the value of the independent variable has on the predicted **y** value)

… = do the same for however many independent variables you are testing

$\mathbf{B_nX_n}$ = the regression coefficient of the last independent variable

e = model error (a.k.a. how much variation there is in our estimate of **y**)

7. 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. 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 and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets. It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

8. WHAT IS PEARSON'S R?

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 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. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844. It is commonly represented by the Greek letter ρ (rho). Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient r.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

9. WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Scaling, as the name suggests, is a method of scaling the features in a dataset for the purpose of making them comparable against one another. Scaling only affects the coefficients and none of the other parameters like the t-statistic or the F-statistic etc.

Scaling is performed for the sole reason of comparison. For example in a dataset having two features as distance travelled in kilometres and another feature as distance travelled in miles are not comparable. To make them comparable, we have to either convert the features having kilometres as unit to miles or miles as unit to kilometres.

There are mainly two types of scaling methods:

- Standardisation – This method brings all the data into a standard normal distribution with mean 0 and standard deviation 1. The formula used for Standardisation is $x = (x - x(\text{mean}))/(\text{std. deviation}(x))$

- Normalisation – It is a method that scales all the data into the range of 0 to 1. The formula for Normalisation is $x = (x - x(min))/(x(max) - x(min))$. Normalisation is also referred to as MinMax Scaling.

The major difference between scaling and normalisation is that Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

10. YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN?

Multicollinearity refers to the problem when the independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables. Two variables are perfectly collinear if there is an exact relationship between the two variables. If the independent variables are perfectly collinear, then our model becomes singular and it would not be possible to uniquely identify the model coefficients mathematically. Hence, collinearity is a problem that needs to be addressed when we are building a multiple regression model. There exist two ways to deal with multicollinearity :

Looking at pairwise correlations of different pairs of independent variables. The drawback of this method is that if there are more than 50 such variables, then plotting them against each other would become a very tedious task and inefficient at the same time. There is also a possibility that instead of just one variable, the independent variable may depend upon a combination of other independent variables.

This is where the Variance Inflation Factor comes in. Variance Inflation Factor (VIF) indicates how well one independent variable is explained by all the other independent variables. Generally, as a thumb rule we consider a VIF greater than 5 to be a high VIF and hence, such a feature is not a very desirable feature for the modelling process. The formula for calculating the VIF is

$$VIF_i = \frac{1}{1 - R_i^2}$$

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).

11. WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

The q-q plot is used to check if the error terms are normally distributed.

The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.