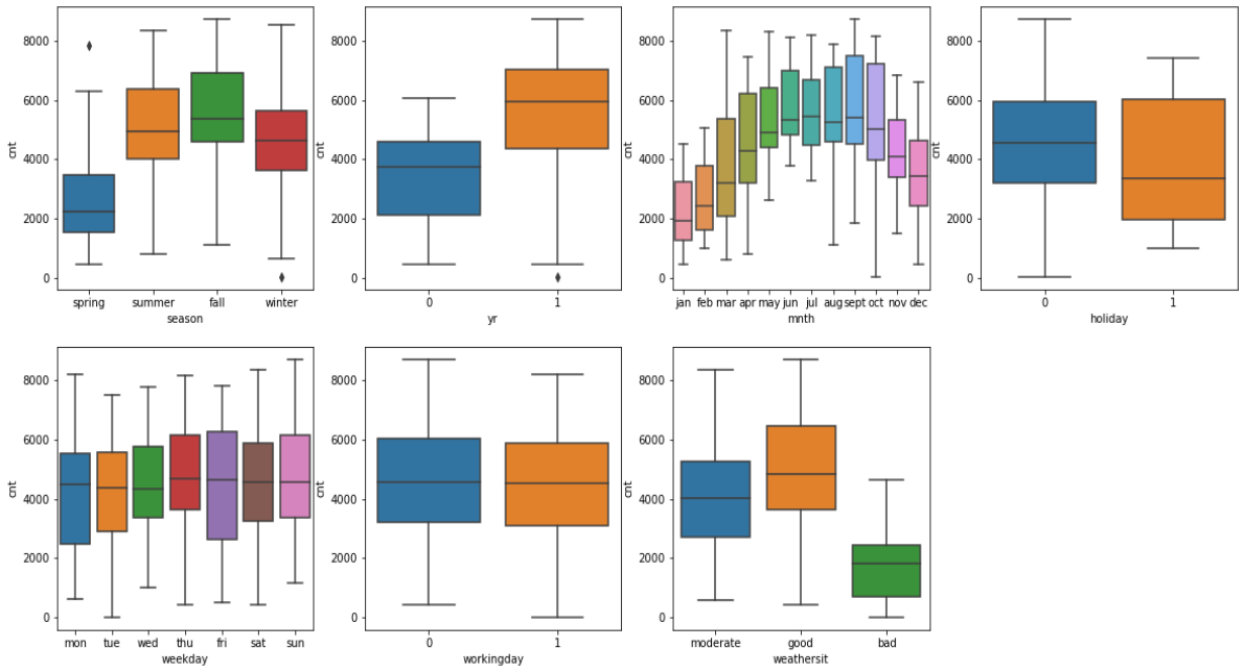
Assignment-based Subjective Questions.

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

Ans 1) After analyzing the categorical variables from the dataset a few inferences about its effect on the dependent variables are given below: -

- In the season attribute the 'fall' season has the highest demand when it comes to renting bikes from BoomBikes.
- In the 'yr' attribute we can see that the demand for BoomBikes have grown in the 2019 compared to 2018.
- If we see the 'mnth' attribute it shows that the demand for BoomBikes is gradually increasing till June, post which we can see some irregularity and we can see the demand is highest in the month of September.
- We can see in the 'holiday' attribute that if there is a holiday then the demand has decreased.
- On observing the 'weekday' attribute we cannot find a definite relation with the demand of BoomBikes.
- The good label (with clear cloud or partial cloud) in the 'weathersit' attribute has the highest demand.
- We can say that the Bike demand is more during the month of September, we can say that the demand is less in the beginning and end of the year due to the extreme weather conditions.



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

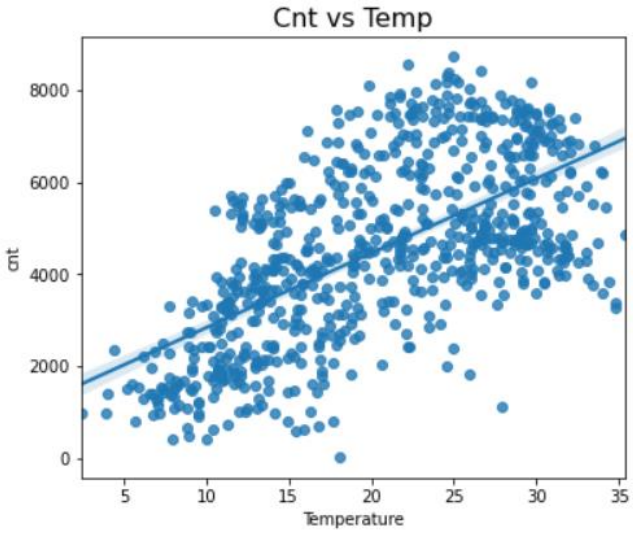
Ans 2) The drop_first=True is used in the dummy variable creation because it helps in reducing the extra columns that were created during the dummy variable creation.

So, for Example: Let's say we gave three variables: Furnished, Semi-furnished and unfurnished. We can only take two variables as furnished will be 1-0, semi-furnished will be 0-1, so we can remove unfurnished as it will be indicated by 0-0.

It is also used to reduce the correlation created among the dummy variables.

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

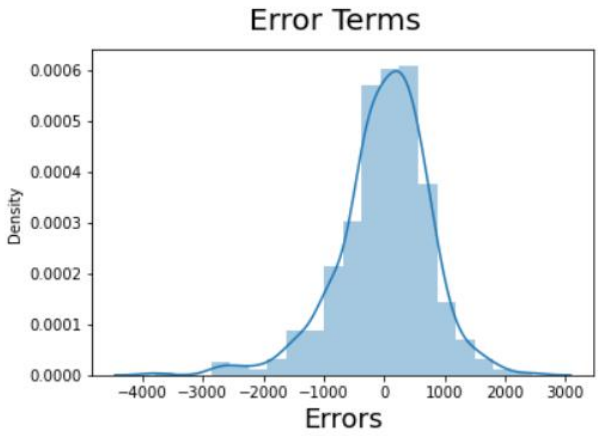
Ans 3) After doing a quick analysis using the pair-plot among the numerical variables, we can see that 'temp' (temperature) has the highest correlation with the target variable.



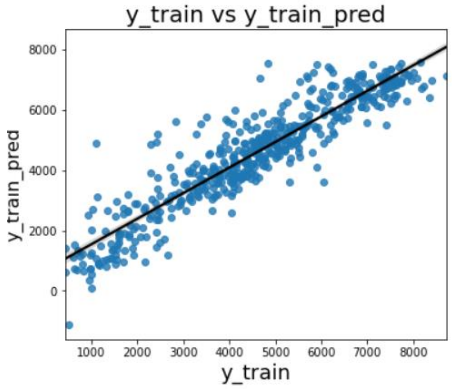
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans 4) The assumptions required to validate the Linear Regression after building the model on the training set are as follows: -

- We first used the Residual analysis to find the difference between the observed value and the predicted value of the dependent variable and we have observed that we have achieved a normally distributed curve with both the sum and the mean of the error approximately near to zero.



- We have also achieved homoscedasticity with the variance being constant with respect to the line.



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

Ans 5) As per the final model, the top three features contributing significantly towards the demand of the shared bikes for BoomBikes are as follows: -

- Temperature could play a very big role for the bike rentals. This is because of the positive correlation of temperature with the target variable cnt. As the temperature increases so does the need of shared bikes.
- We can observe that the year has positive impact in shared biked. The year 2019 shows greater demand of shared bikes compared to the previous year that is 2018
- The month of September and the winter season also has positive effect on the shared bike demand in the market.

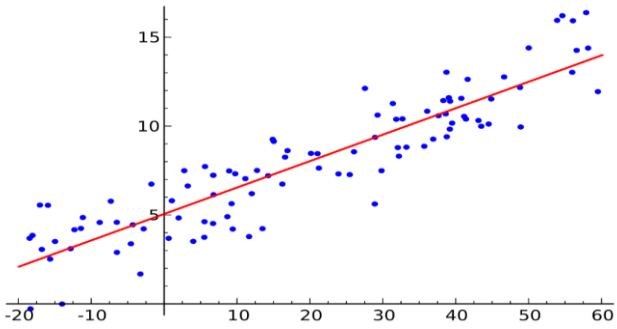
General Subjective Questions.

1. Explain the linear regression algorithm in detail?

Ans 1) Linear regression is a Machine Learning technique which is used to predict the values of one variable based on the value of the other variable. The variable that we need to predict are known as dependent variable and the variable that are used to predict the other variables are known as independent variables. This form of analysis is used to estimate the coefficients of the linear equation. The analysis includes one or more independent variables that can help to predict the dependent variable in the best possible way.

The Linear Regression is used to fit a straight line or the surface that minimizes the difference between the actual and the predicted values. Some examples of Linear Regression application and use cases in real life scenarios are Evaluating the sales of an employee, analyzing the pricing change as per market growth, sports analysis etc.

The graph below is used to explain how linear regression is used: -



The above graph has a red line which is referred as the best fit straight line. Depending upon the given data points, we try to plot a line that can model the points in the best way possible. The line is modeled based on the below provided equation.

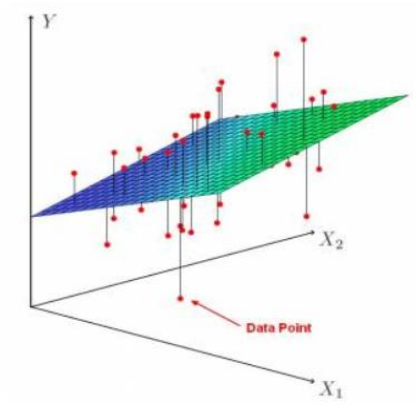
The following equation is for Simple Linear Regression with one independent variable.

y= B0 + B1X

The above equation has terms defined below:-
y- It is the output variable. It is also known as dependent variable or target variable.
X- It is the input variable. It is also described as the feature in machine learning or is also known as independent variable.
B0- y-intercept at time zero.
B1- It is the coefficient of the independent variable.

There is another form of regression also known as Multiple Linear Regression, it assumes there is a relationship between two or more independent variables and one dependent variable.

The Multiple Linear Regression model can be represented as a plane (in 2-Dimensions) or in the form of hyperplane (in higher dimension).



y= B0 + B1X1+ B2X2+……. BnXn + e

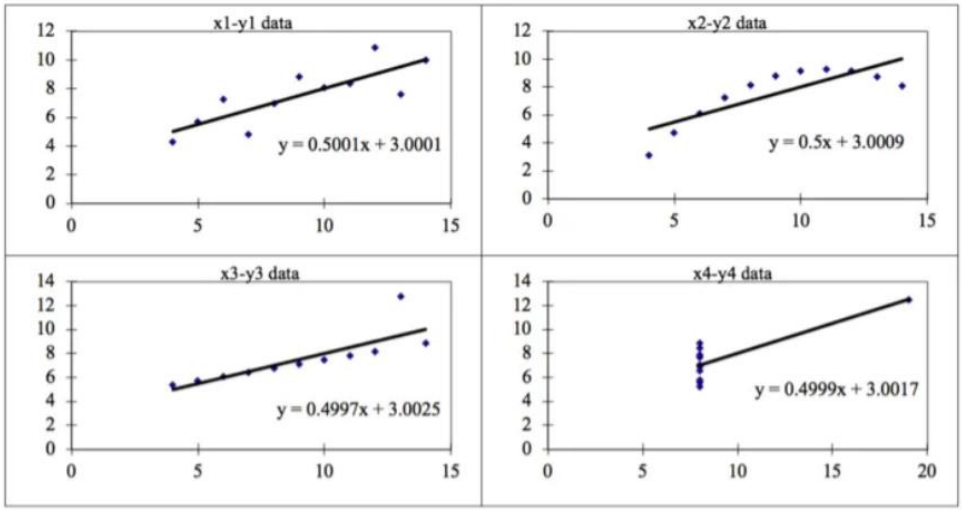
y- It is the output variable. It is also known as dependent variable or target varaiable.
B0- y-intercept at time zero.
B1X1- The regression coefficient (B1) of the first independent variable (X1) (a.k.a the effect that increasing the value of the independent variable has in the predicted y values)
B2X2+…. – Similar step as above is repeated for however many independent variable that the dataset contains.
BnXn- The regression coefficient of the last independent variable.
e- Model error.

2. Explain the Anscombe's quartet in detail.

Ans 2) Anscombe's quartet consists of four datasets that have nearly identical simple statistical properties, yet they appear very different when they are graphed. Each of this datasets consist of eleven (x,y) points that needs to be analyzed. The model was constructed in the year 1973 by Francis Anscombe to observe and demonstrate the importance of graphing data before analyzing it. It also helps to depict the effect of outliners on statistical properties.

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

On plotting the above given datasets on a scatter plot, all the datasets generate different kind of plots that are not able to be interpreted by any linear regression algorithm which is then been fooled by these peculiarities and can be represented as follows.



The above four datasets can be represented as the following: -

Dataset 1: It fits the linear regression model quite well.

Dataset 2: The data are not linear, and the model couldn't be fit in a straight line.

Dataset 3: The plots depict the outliners involved in the dataset and hence it cannot be handled by a simple linear regression model.

Dataset 4: It shows the outliners involved in the dataset which similarly cannot be handled by the linear regression model.

After this the set of council then started analyzing the above datasets using the descriptive statistics and hence found the mean, standard deviation and correlation between x and y.

Hence, we can now conclude that all the important features in the dataset must be visualized before implementing any machine learning algorithm on those data points. Therefore, all this will be helping to prepare a good model and make the right analysis.

3. What is Pearson's R?

Ans 3) The Persons Correlation Coefficient is also known as Pearson's r, the Pearson productmoment correlation coefficient (PPMCC), or bivariate correlation. It is a part of statistics which is used to measure the linear correlation between the two variables. Like all the correlation, it also has numerical value which lies in between the range of -1.0 to +1.0.

Moreover, whenever we discuss about correlation it generally is related to the Pearson's r. However, it cannot identify and analyze the nonlinear relationship between two variables and cannot differentiate between the dependent and independent variables.
We can also define it as the covariance of the two variables divided by the product of their standard deviation. The Pearson's Correlation Coefficient is named after Karl Pearson.

The Pearson's Correlation Coefficient Formula: -

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Ans 4) Scaling is a step of data pre-processing which is applied to independent variables to normalize the data which are present in a particular range. It plays avital role in increasing speeding up the speed in the algorithm.

The reason behind performing scaling is because most of the times the collected data set contains features which are highly varying in magnitude, units and range. If scaling is not performed, then the algorithm only takes the magnitude in account and not the units hence the modeling becomes incorrect. To correct this, we need to perform scaling to bring all the variables to the same level of magnitude so that the analysis can be made efficiently.

It is important to understand that scaling affects the coefficients and not the other parameters like t-statistics, F-statistic, p-value, R-squared etc.

The difference between Normalized and Standardized scaling are as follows: -
Normalized Scaling- It brings all the data in the range of 0 and 1. We can also use the package *sklearn.preprocessing.MinMaxScaler* helps to implement normalization in Python.

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

Standardized Scaling- It replaces the value of there Z-scores. It brings all of the data into a standard normal distribution which has a mean zero and a standard deviation as 1. We can use the sklearn.preprocessing.scale package to implement the standardization in Python.

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

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

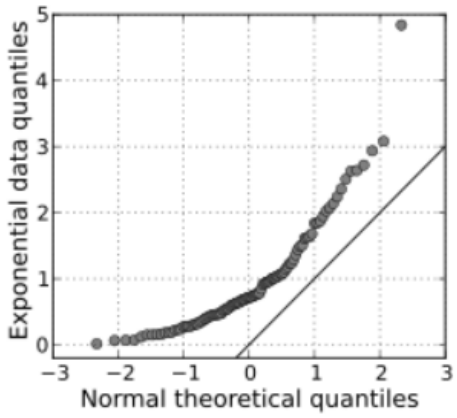
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans 5) The VIF is infinite only when there is a perfect correlation. This shows a perfect correlation between two independent variables. If we get a perfect correlation, we get R2=1 , which ultimately leads to 1/(1-R2) infinity. The only way to sought this problem is to drop one of the variables from the provided dataset which is leading to perfect collinearity.

An infinite VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which depicts a 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 6) The Q-Q plots are known as Quantile-Quantile plots they are plots of two quantiles which are against each other. When we say quantile, we mean that it is a fraction where certain values fall below the quantile. For example, the median is a quantile where 50% of the data fall below that and 50% lie above it. The reason for using the Q-Q plots is to understand and find out if the two sets of data come from the same distribution. Let's say that a 45 degree angle is plotted on the Q-Q plot; if the two datasets come from a common distribution, the points will fall on a reference line.



The above plot shows the quantiles from a theoretical normal demonstration on the horizontal axis. Here it's being compared to the set of data on y-axis. This particular type of Q-Q is known as normal quantile-quantile. Also here the points are not on 45 degree line, and follows a curve path suggesting that the data's are not normally distributed.