Answers to 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?

A. The relationship between categorical variables and a Target variable shows that:

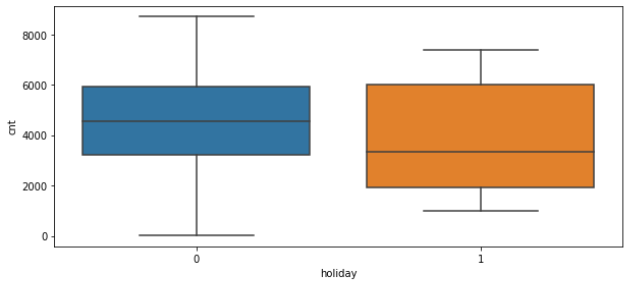
Demand of bike is more if it is not a holiday.

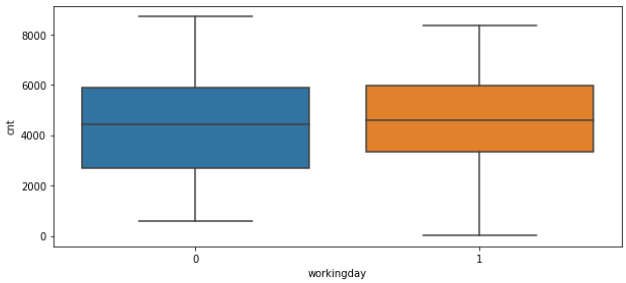
Demand of bikes is more in month of September.

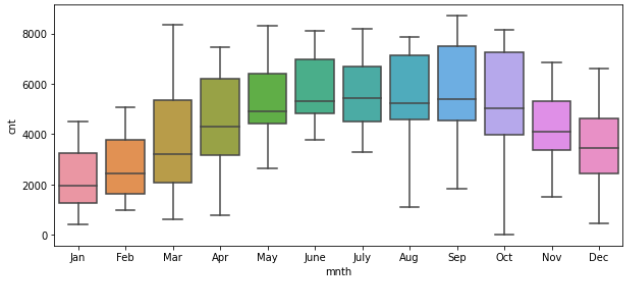
In clear weather sit bikes demands are more compared to mist cloudy.

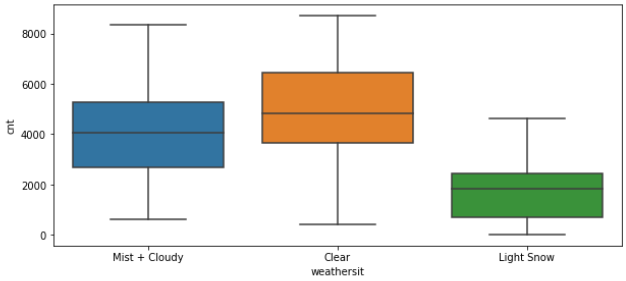
In fall season demand of bike is more.

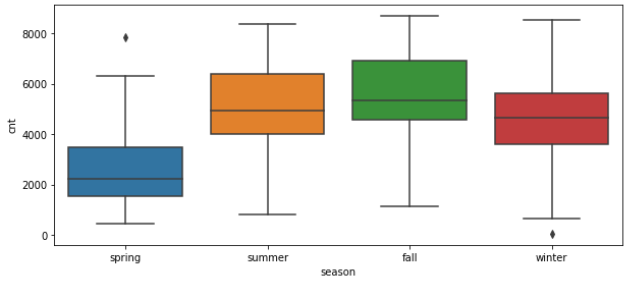
In year 2019 demand of bike increases as compared to 2018.

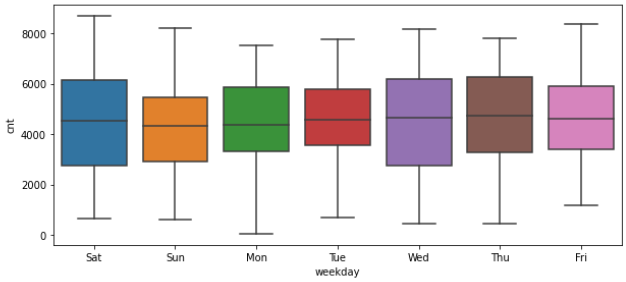


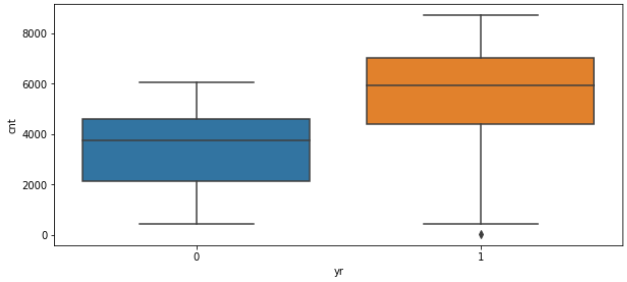












2.Why is it important to use drop_first=True during dummy variable creation?

A. 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. For example :

Relationship status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

Now, you don't need three columns. You can drop the Single column, as the type of Single can be identified with just the last two columns the table looks like,

Relationship status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

A. Looking at the pair-plot among the numerical variables, atemp and temp are 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?

A. After building the model on the training set the assumptions of Linear Regression are:

-Checking whether the error terms are normally distributed or not.
-The probability distribution of the error has constant variance or not.
-The error values are statistically independent or not.

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

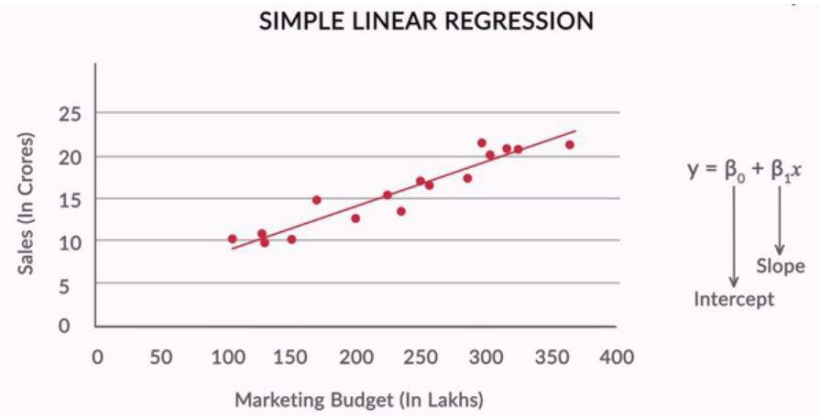
A. temp(temperature) ,yr (year),Sep(September month) are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

Answer to General Subjective Questions:

1. Explain the linear regression algorithm in detail.

A .Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$



The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable

The strength of the linear regression model can be assessed using 2 metrics:
1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

Where $R^2 = 1 - (RSS / TSS)$

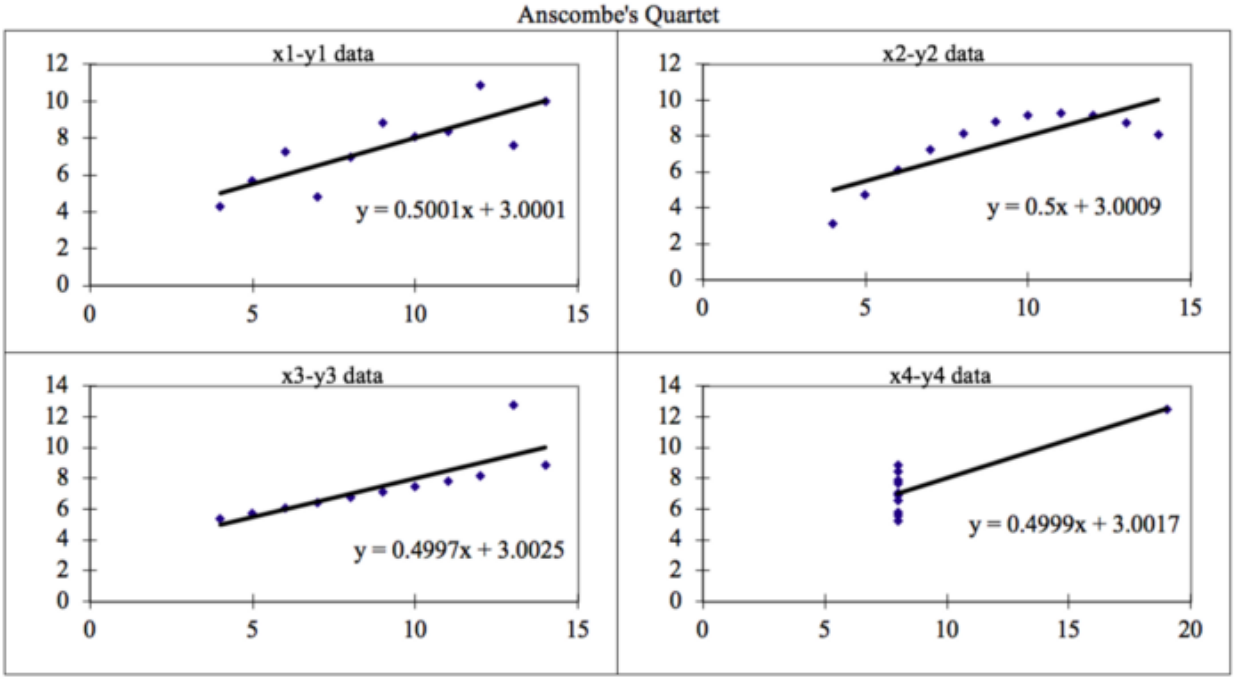
The assumptions of simple linear regression are:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

2. Explain the Anscombe's quartet in detail.

A. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

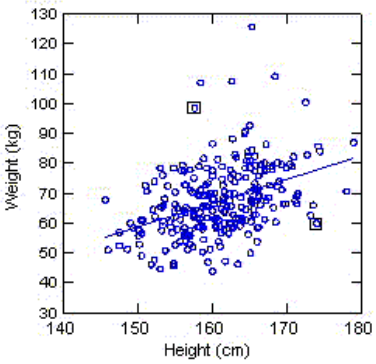
Four datasets that fools the Linear Regression model if built.



3. What is Pearson's R?

A. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



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

A Scaling is is a important feature while building a model. When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

There are two types of Scaling methods they are

Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

A. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

A 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.