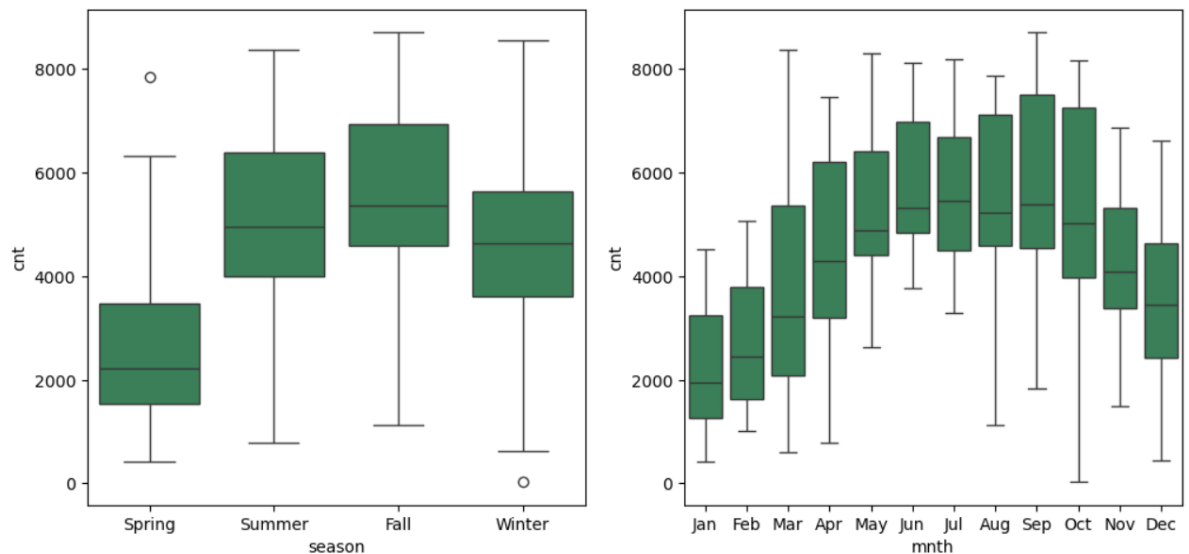


## Assignment based Subjective Questions

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

Boxplots were plotted using categorical variables and the dependent variable (cnt) as the value. From the boxplots it could be seen that certain categories lead to an overall higher cnt value. For example – The Fall season had a higher median of count of rentals. Similarly, the months of June and July contributed to higher number of rentals when compared with other months.



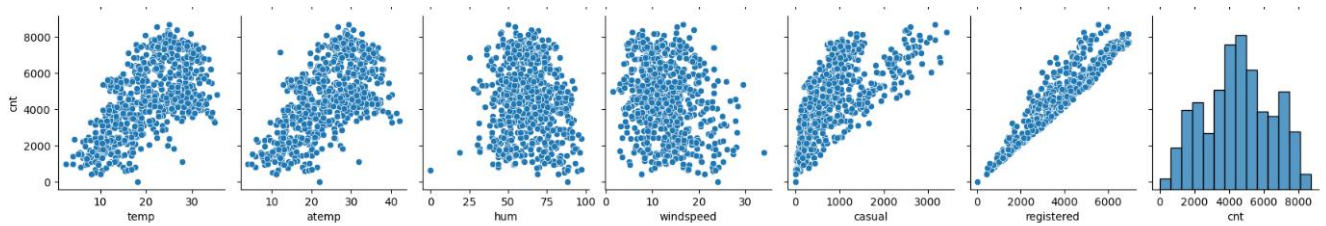
After arriving at a good model, it can be seen that the model is dependent on only few of the categorical variables for a good amount. It can be seen that **Spring** season and **Cloudy Misty** weather have a stronger negative effect on the cnt value. This means that generally if it is the spring Season and the weather is cloudy misty, there are lesser chances that bikes will be taken on rental

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

When we use `pd.get_dummies()` to create dummy variables, if there are 'n' unique values, it created 'n' columns so that each column will get value 1 if the value represents that. But n-1 columns are enough to represent the same data. It is because if there is 0 in all the other columns, it means that it represents this last column. If this column is not dropped, it will add multicollinearity. This is because this last column can be easily derived from the rest of the dummy columns. Multicollinearity is not good for a model hence should be avoided.

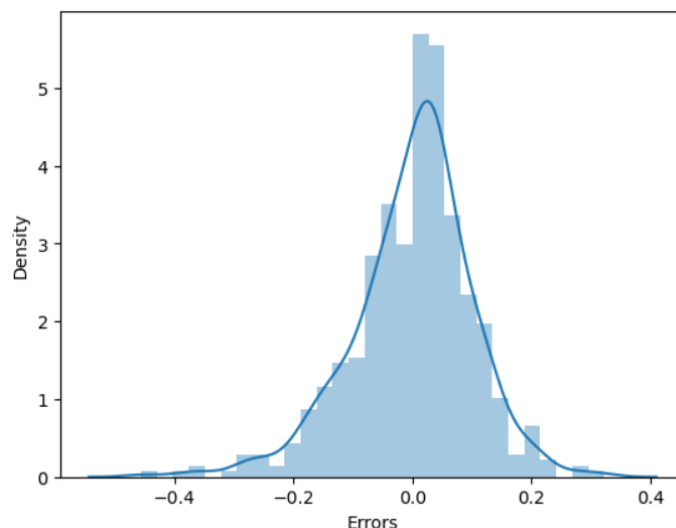
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From the pair plot it is clearly visible that the 'registered' column has the highest correlation with the target 'cnt' variable.



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

An analysis was performed on the residuals formed after using the model to predict the values for the dataset present in the training set. The residuals (error terms) seem to follow a normal distribution with the mean centred around 0. This is one of the key assumptions we take while doing linear regression.



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

Since the training dataset is scaled to the same range, we can directly relate the magnitude of the coefficients to get the amount of contribution. The top 3 variables in that order are 'temp', 'weathersit\_Light Rain' and 'yr'. It is positively correlated with temp and yr and negatively correlated with Light Rain. This means that with higher temperature it is more chance for people to take rental. There is a higher chance of rental if the year is 2019. There is lesser chance that there will be rentals when the weather is of type Light Rain.

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression is a type of Supervised Learning model. This means that there are a set of independent variables which provide an effect on the dependent variable. The regression model helps in understanding how much can these independent variables affect the dependent variable and in what ways (positive or negative) in a mathematical manner. All

the input and output variables are numerical in nature when it comes to Regression Models. Linear Regression algorithm tries to fit a line which best explains the given input and helps in predicting the output dependent variable.

There are 2 types of linear regression models -

- a. Simple Linear Regression -> Single input variable used to predict single output variable
- b. Multiple Linear Regression -> Multiple independent input variables together predict a single dependent output variable.

General equation of a linear regression model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where  $\beta_0$  is y-intercept,  $\beta_1$  is slope of the equation,  $\epsilon$  is the error term, y is dependent variable and x is the independent variable.

This equation can be extended for the multiple linear regression algorithm by adding x2, x3 and so on. There instead of fitting a line, we will fit a hyperplane.

There are a few assumptions of linear regression :

- a. Linearity -> Linear relationship between X and y.
  - b. Independence -> The observations should be independent
  - c. Normality -> The error terms (difference between predicted and actual) follow a normal distribution centered at 0.
  - d. Homoscedasticity -> The error terms have a constant variance and do not follow any patterns
3. Multicollinearity -> There is no high correlation in between the independent variables used to predict.

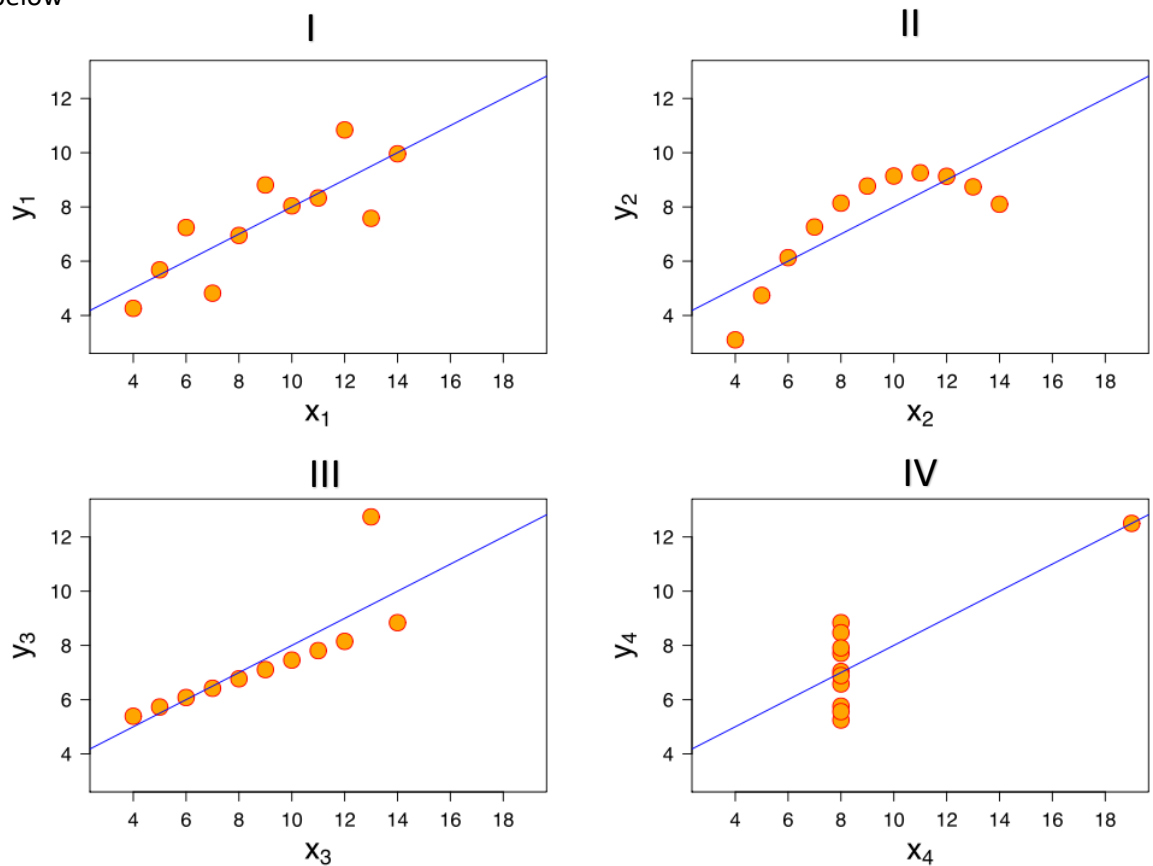
The main objective of the linear regression model is to find the best coefficients so that the error terms are minimized. This will help us arrive at the best line to fit the data. It is done using the Least squares method – Which means minimizing the sum of squares of the residuals. To find the coefficients, we use the gradient descent method. This allows us to quickly arrive at the best coefficients. This method is faster when the datapoints are all scaled to the same level.

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

Anscombe's quartet helps us understand the importance of visualization before building a regression model as statistics alone can sometimes make us blind to the actual observations spread. There are 4 datasets, each having x and y columns. All these 4 datasets have the same statistical properties -> mean of x = 9, mean of y = 7.5, SD of x = 3.16, SD of y = 1.94 and also all of them have a similar regression line. But there are differences in the values as can be observed by the subsequent plots. The dataset is 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

When these 4 different datasets are plotted on a graph, the variations can be observed below



Model I -> This is the general dataset that we imagine when it comes to simple regression, the points are near the regression line and slope seems right

Model II -> The points here do not follow a linear trend at all. Still the regression line is similar.

Model III -> Except for the one outlier all the other points follow a linear trend but their slope is different. If the outlier was removed the linear regression fit would have changed and could have been perfect for the rest of the points

Model IV -> All points are in a straight line and one point on the extreme end influences the regression line heavily and the line seems to only pass through that outlier.



This shows us the importance of first visualizing the points before going ahead with modelling.

### 3. What is Pearson's R?

The Pearson's R coefficient also referred to as Pearson's correlation coefficient gives the amount of correlation between 2 variables. It shows whether the 2 variables have a strong relationship with one another – whether one increases with the other or decreases or does not have any effect with the other.

**Pearson Correlation Coefficient**

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



[Pearson's correlation calculation](#)

$r$  = Pearson Coefficient  
 $n$  = number of pairs of the stock  
 $\sum xy$  = sum of products of the paired stocks  
 $\sum x$  = sum of the x scores  
 $\sum y$  = sum of the y scores  
 $\sum x^2$  = sum of the squared x scores  
 $\sum y^2$  = sum of the squared y

The value lies between -1 and 1.

1 -> Indicates strong positive relationship. If one increases other also increases and vice-versa

-1 -> Indicates strong negative relationship. If one increases other decreases and vice-versa

0 -> Indicates no relationship. Change in one variable doesn't affect the other.

Disadvantages:

- a. Can only work for linear relationships.
- b. Does not show which variable is dependent on the other.

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

Scaling is the process of transforming the input dataset to make sure all columns are in comparable terms. It involves standardizing or normalizing the independent variables.

If the independent variables are all in different scales (some in 0.1s and some in 1000s) then the coefficient for each of these variables will also be in different scales. The issue then becomes how to interpret the coefficients. We can't say that a higher coefficient means this variable has a higher say on the dependent variable. If we whereas change all values to a similar scale, the interpretation of coefficients becomes more easy.

In algorithms that use gradient descent to arrive at the best coefficients (Most regression models) it allows for a faster convergence of the algorithm.

There are 2 very popular scaling methods – Normalization and Standardization.

Normalized Scaling	Standardized Scaling
Referred to as Min Max Scaling. Brings to the range of [0, 1] or [-1,1]	Transforms to a normal distribution with mean as 0 and Standard Deviation as 1
Sensitive to Outliers	Not sensitive to Outliers
$\frac{x - \min(x)}{\max(x) - \min(x)}$	$\frac{x - \text{mean}(x)}{sd(x)}$
Use when distribution is not Gaussian and when data doesn't have outliers (KNN). Also in neural networks since there the values should strictly be <1.	Use for Gaussian distribution even in presence of outliers. Clustering algorithms and PCA generally prefer this scaling

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

The formula for Variance Inflation Factor is as follows

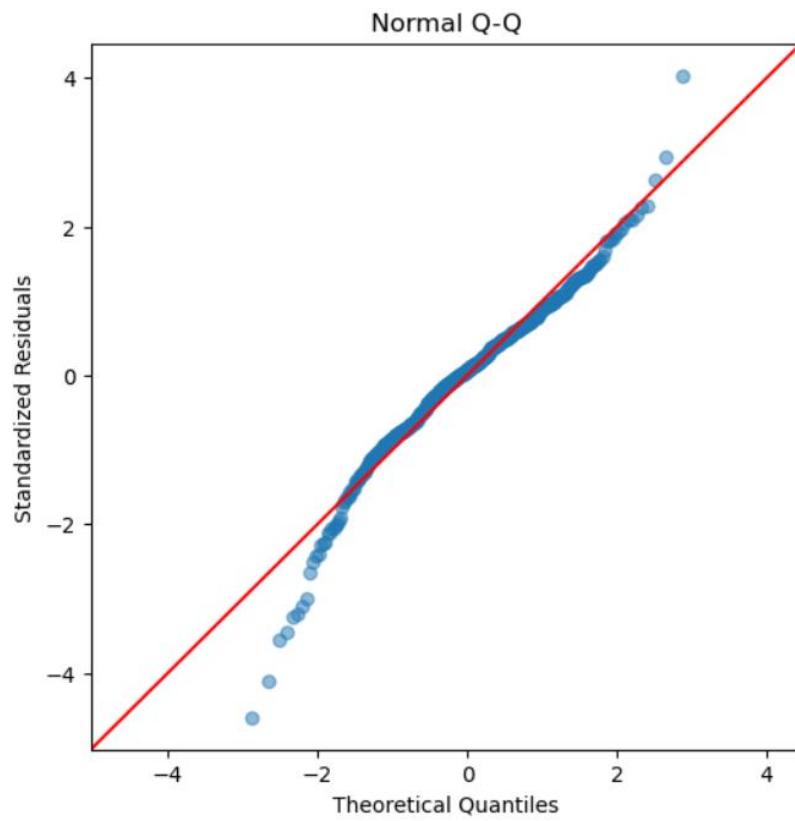
$$VIF_i = \frac{1}{1 - R_i^2}$$

The VIF value can become infinite if the denominator becomes 0. This will happen if the value of  $R^2$  becomes 1. This means that there are 2 variables which are perfectly correlated with one another. Which also means that one of the independent variables can be represented in a linear equation using another independent variable. This also means that the multicollinearity is very high for a particular independent variable. Like how in our Boom bikes example, the cnt variable can be expressed as a sum of registered and casual variables. Hence using both of them directly to predict the cnt variable would have led to very high VIF. Hence those 2 variables are not supposed to be used in the final model.

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

A Quantile Quantile Plot (Q-Q Plot) is a graphical plot which can help us test our assumptions of whether a particular dataset follows some distribution. It is mostly used to check whether the data follows a normal distribution.

In linear regression a Q-Q Plot can be used to evaluate the assumptions that the residuals follow a normal distribution. To make the plot, the theoretical quantiles are calculated for the distribution to be checked and the sample quantiles are also calculated. Then they are plotted like a scatter plot on one against one another. If the residuals follow the given theoretical distribution, then the points in Q-Q plot would lie on a straight line for the residuals. In Python the method ProbPlot can be used to get the Q-Q Plot. From the Q-Q Plot attached below it can be observed that the points are spread almost on a straight line. So we can say that the assumption that the residuals follow a normal distribution can be visually validated.



QUESTIONS ABOVE ANSWERED BY – MAALOLAN K

DATE - 29/07/2024