Linear Regression Subjective Questions and Answers

Assignment-based Subjective Questions

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

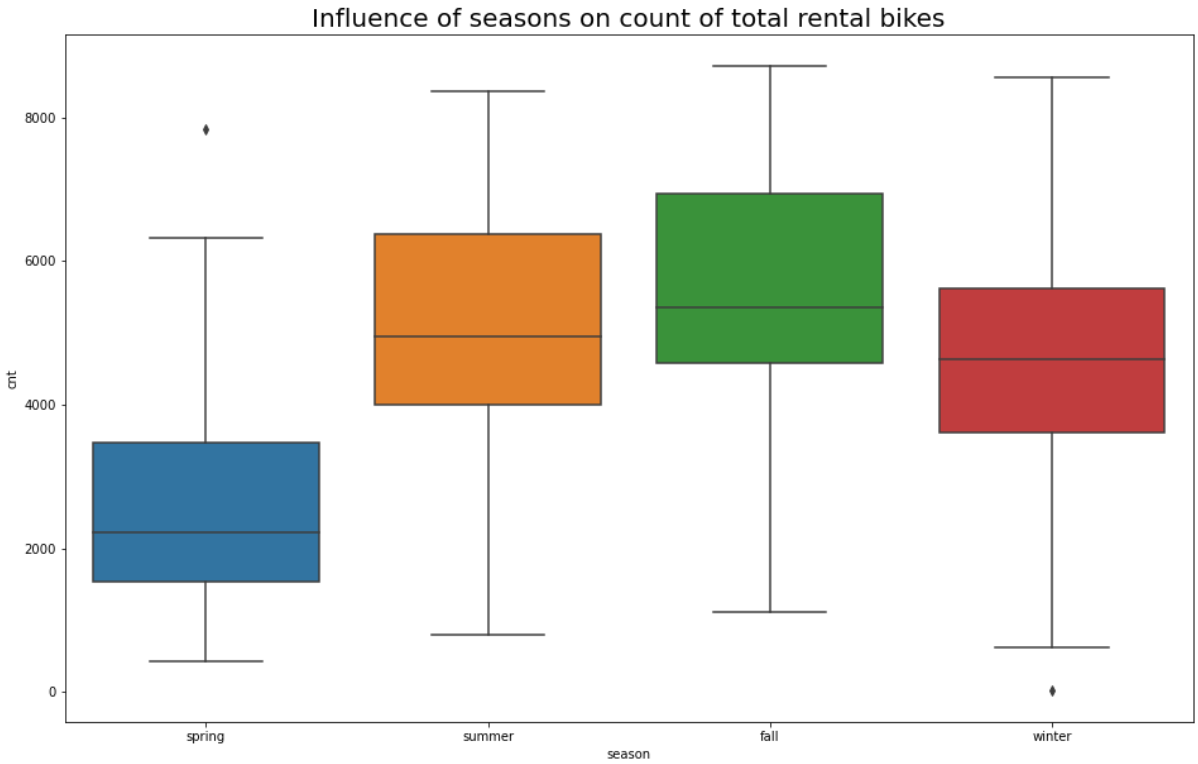
Answer: There are 4 categorical variables in the dataset viz.:

1. Season
2. Month
3. Weekday
4. Weathersit

Effects of each of these variables on the dependent variable is explained below:

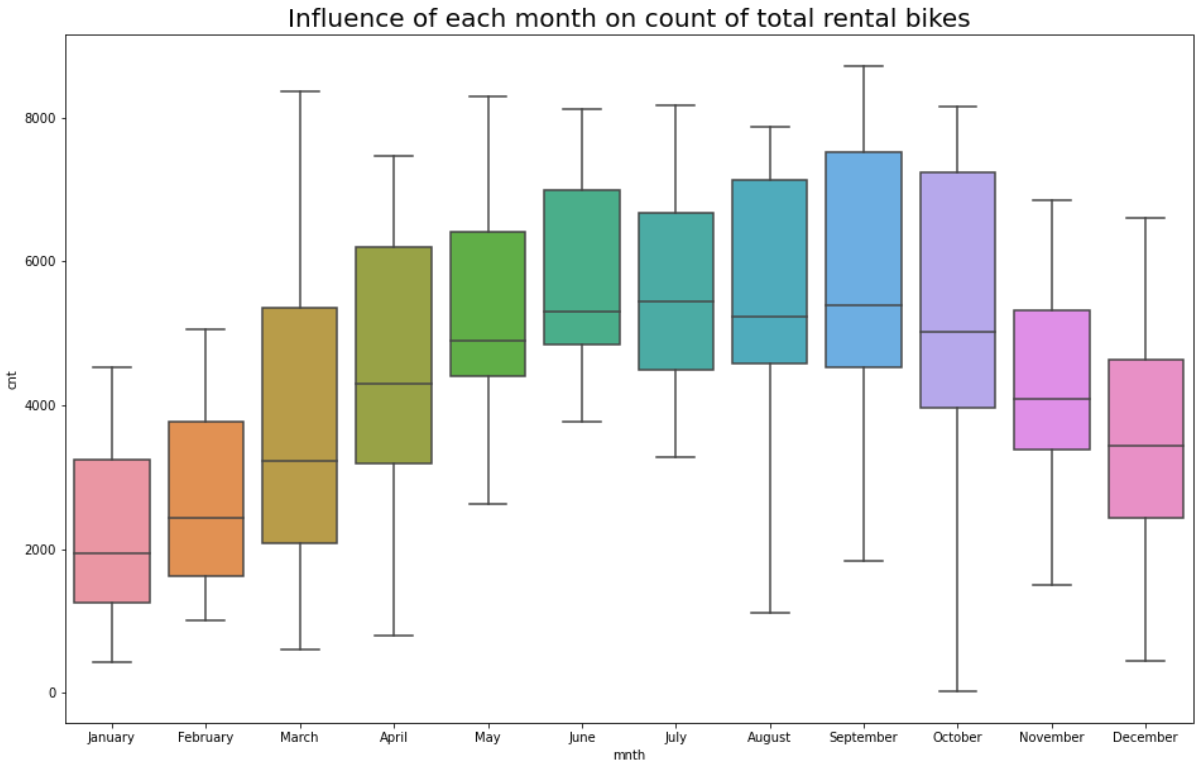
Effect of Season on the dependent variable

During the spring season demand of bikes ar at their lowest, it gradually increases and reaches its peak at the fall season then again tapers off in the winter.



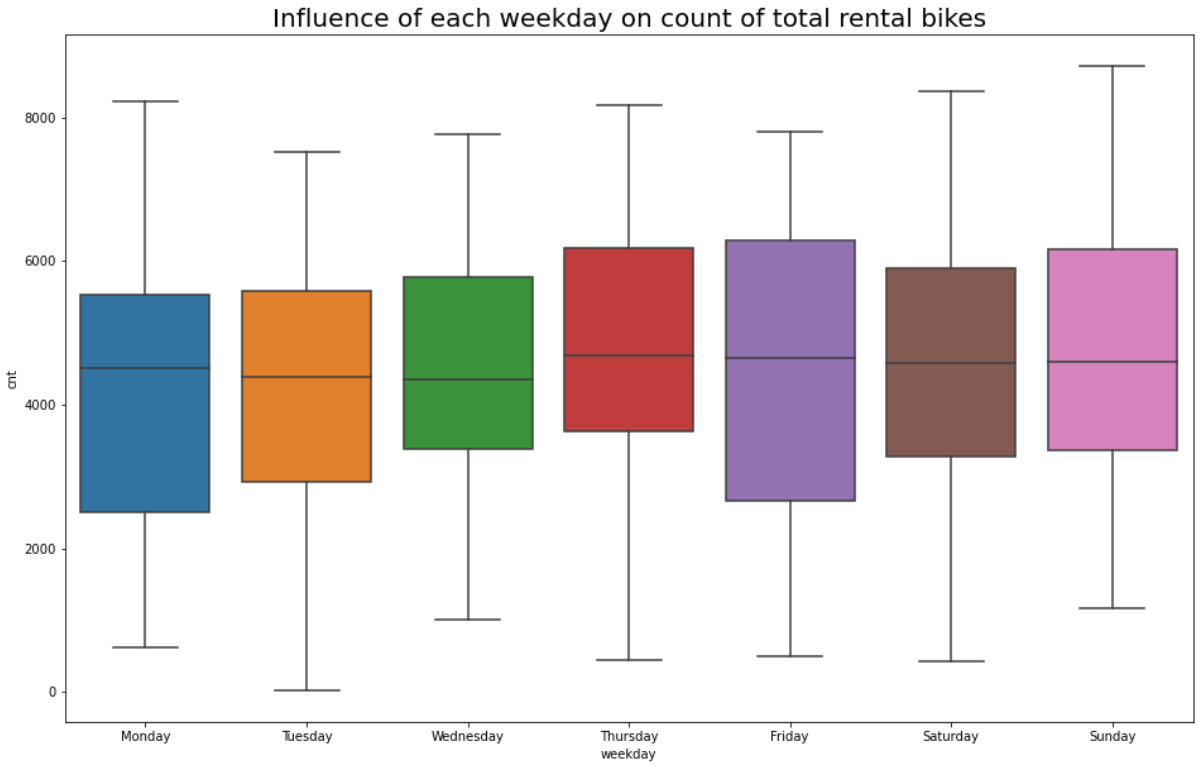
Effect of Month on the dependent variable

Starting from January the count of rented bikes increases steadily till July where it dips a bit and again picks up pace till September where it registers maximum rentals and then the rental couts decreases.



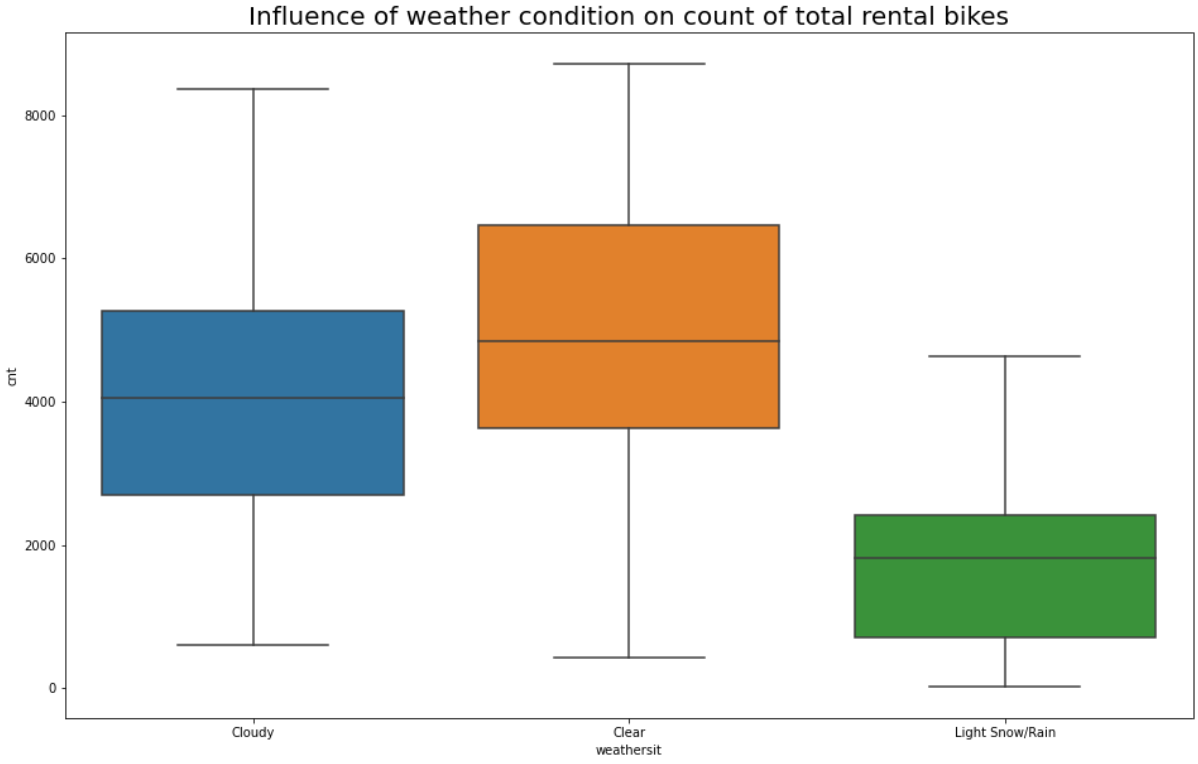
Effect of Weekday on the dependent variable

No discernible trend is observed here.



Effect of weathersit on the dependent variable

Total rental count of bikes seems to be higher when the weather is clear.

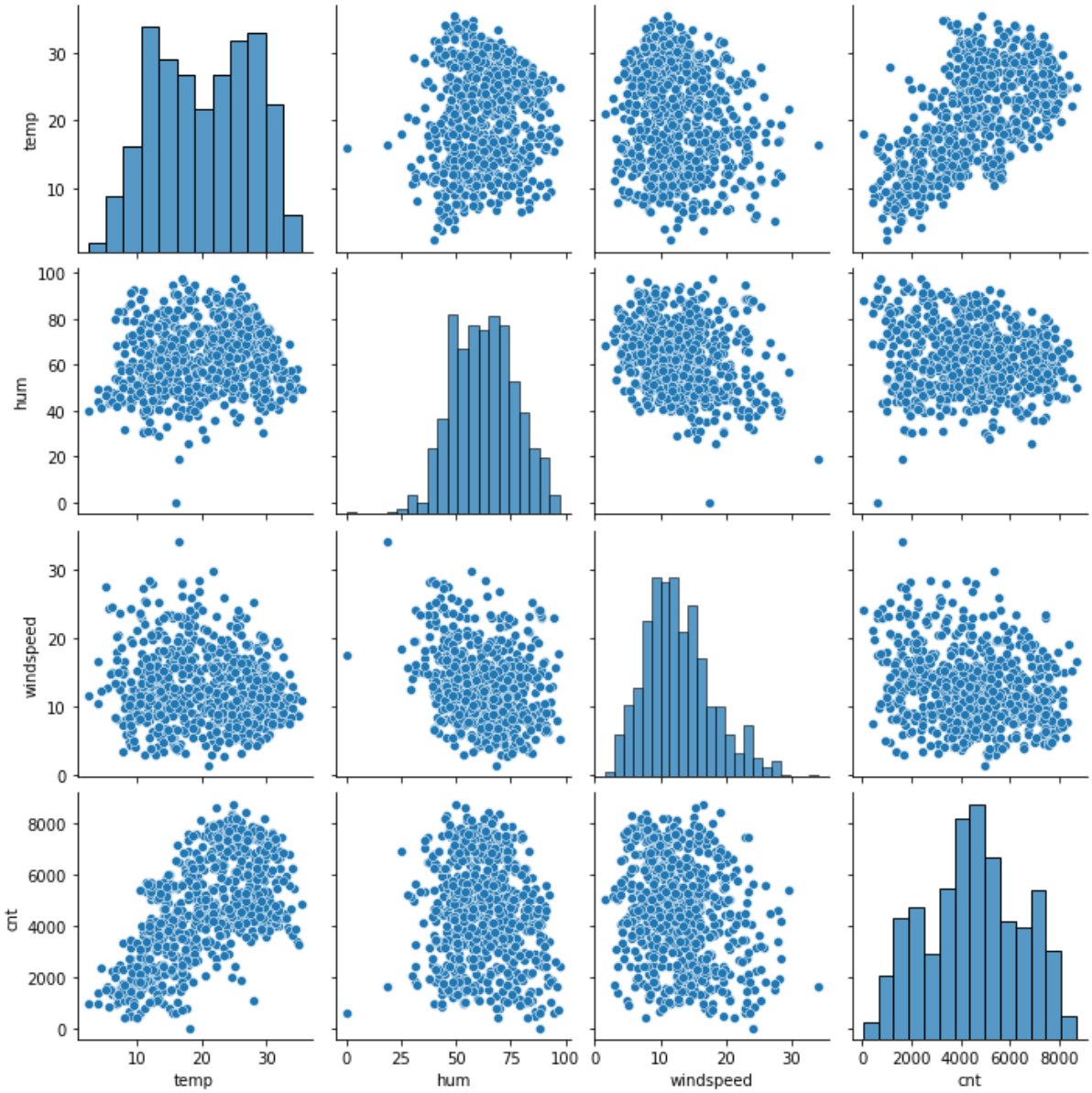


Question 2 : Why is it important to use drop_first = True during dummy variable creation?

Answer: The reason to use drop_first = True during dummy variable creation is because it helps us create N-1 categorical dummy variable from N categorical features which in turn helps in reducing the correlations created among dummy variables.

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

Answer: Numerical variable temp has the highest correlation with the target variable.



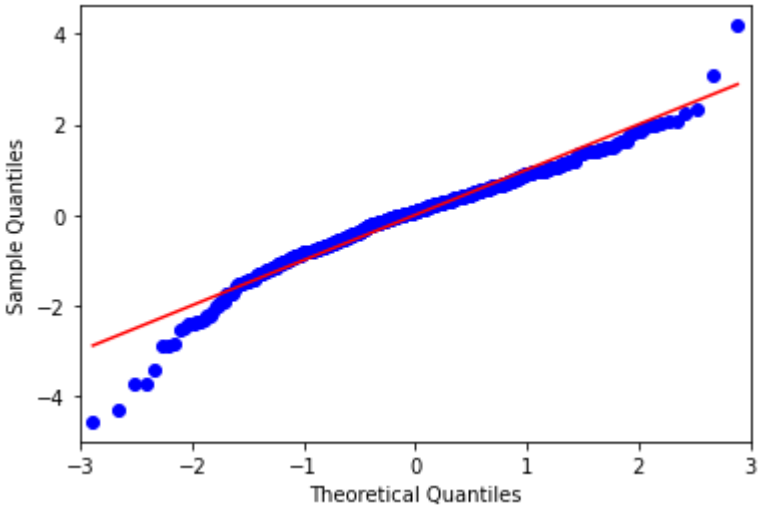
Question 4 : How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: The assumptions of Linear Regression were validated after building the model on the training set by plotting 3 graphs:

1. QQ-plot - This plot shows if the residuals are normally distributed or not.
2. Residuals vs Fitted - This graph shows if there are any nonlinear patterns in the residuals, and thus in the data as well.

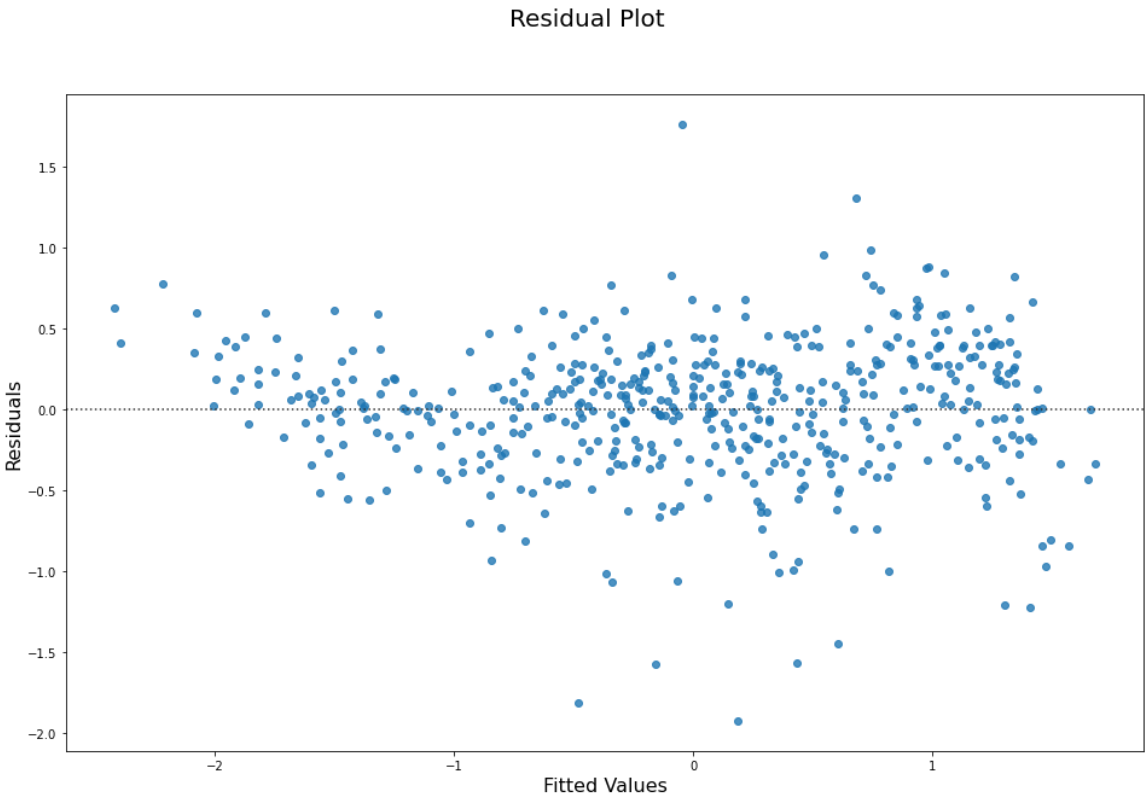
3. Scale-Location plot - This plot is a way to check if the residuals suffer from non-constant variance i.e. heteroscedasticity.

QQ plot of our dataset



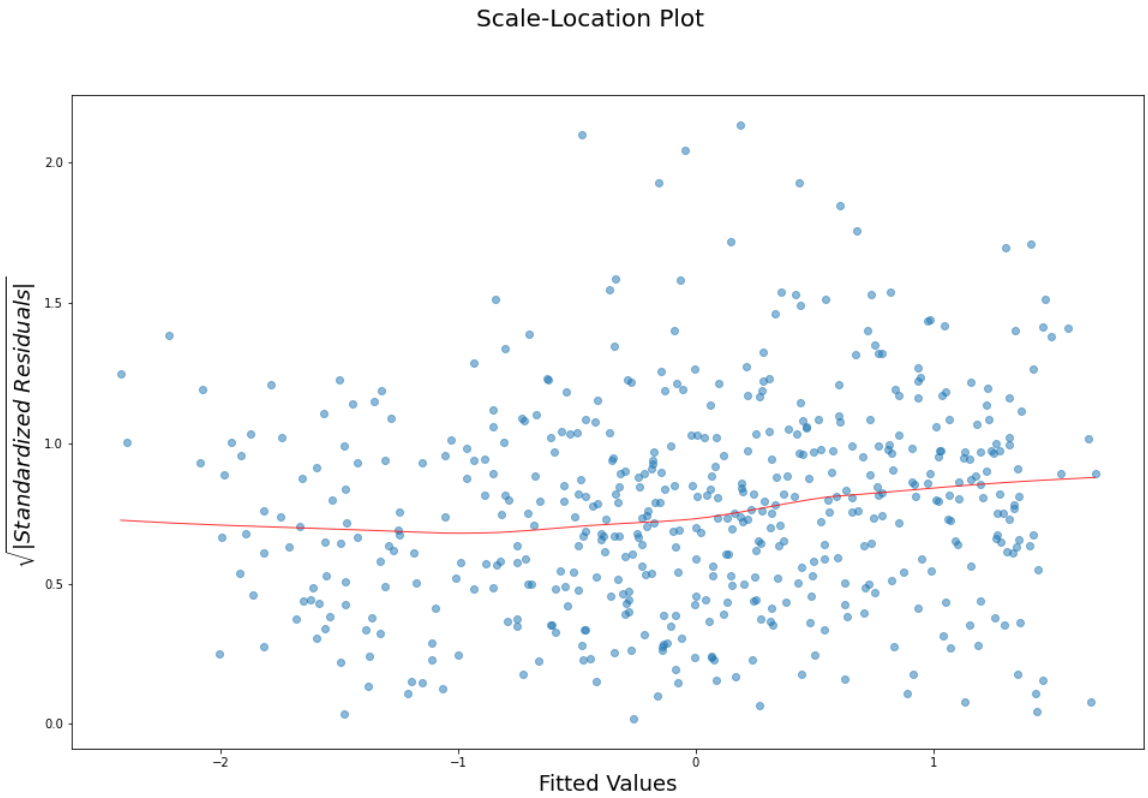
The QQ plot shows that the **residuals are normally distributed**.

Residual vs Fitted plot of our dataset



The equally spread residuals around a horizontal line without distinct patterns are a good indication of **having the linear relationships**.

Scale-Location plot of our dataset



The near horizontal red line proves that the **residuals are Homoscedastic**.

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

Answer: The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:

1. weathersit
2. temp
3. seasons

General Subjective Questions

Question 1 : Explain the linear regression algorithm in detail.

Answer: 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.

This regression technique finds out a linear relationship between the independent and dependent variables. Hence, the name is Linear Regression.

In mathematical notation the linear regression model is written as:

$y_i = b_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + \epsilon_i$

Where, y_i is the predicted response of the model

b_0 is the constant/intercept term i.e. the response of the model if all feature variables are zeroes.

β is the co-efficeient of the variable X

X is the independent/feature variable of the model

ϵ_i is the inherent error because in general, data doesn't fall exactly on a line.

Linear regression model tells us:

1. For a unit change in X the output y will change by β units when, all the other independent variables (X's) are held constant.
2. If all of the independent variables (X's) are driven to zero, the output of the model will still be equal to b_0

Question 2 : Explain the the Anscombe's quartet in detail.

Answer: 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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc..

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

Image has been taken from Wikipedia

Question 3 : What is Pearson's R?

Answer: Pearson's R is a correlation coefficient. Correlation coefficients are used to measure how strong a relationship is between two variables:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.

- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

Answer: Scaling is step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Scaling transform the data so that all the variables are at the same level of magnitude. In absence of scaling during modelling on the data, the model(s) more often takes magnitude of the features along with their units also, which is not the desired outcome. We only want the models to take the magnitude of the data into account and scaling solves this problem.

Normalised scaling brigs all of the data in the range of 0 and 1. Whereas in standardized scaling all of the data is transformed into a standard normal distribution which has mean 0 and standard deviation 1. Also, in normalised scaling data loses some of the information contained in it, especially about outliers.

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

Answer: VIF is expressed as $1/(1-R^2)$, so if a feature has perfect co-relation with some other feature in the data their R^2 value would be 1 and as such VIF would be ∞.

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

Answer: Q-Q plots are a quick and visual way to assess whether two sets of data came from the same distribution or not. We can use QQ plots to check our data against any distribution, not just the normal distribution (for which it's majorly used).

In linear regression context it helps us to visualize whether the residuals are normally distributed or not.