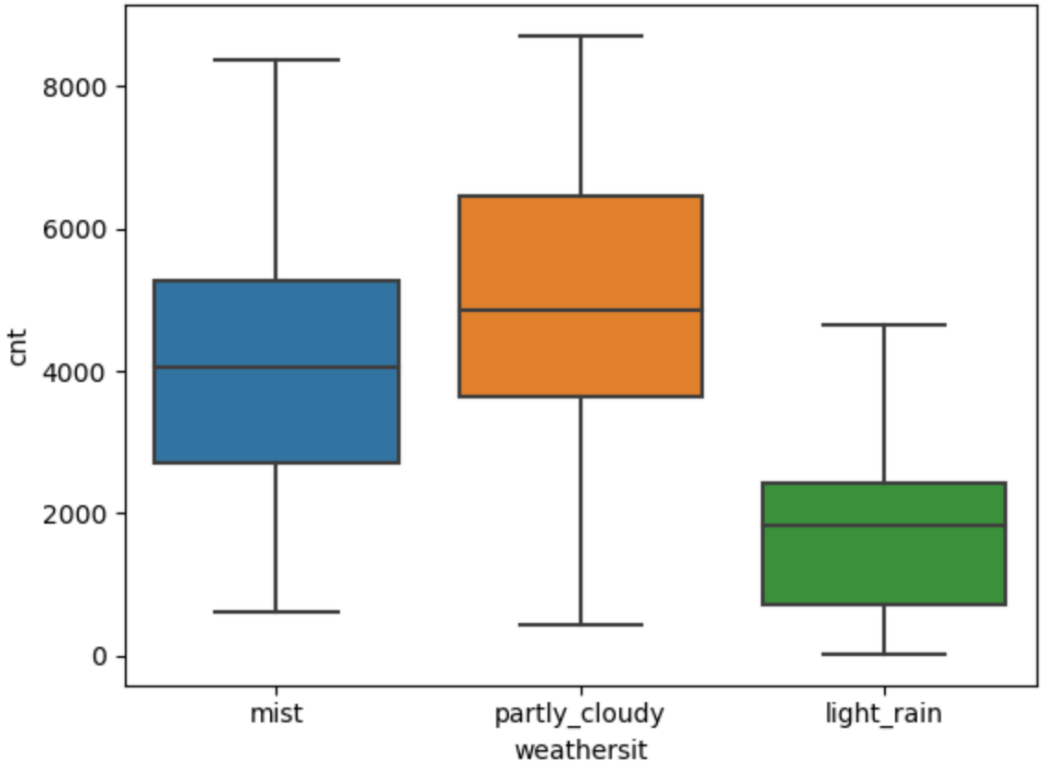
Assignment-based Subjective Questions

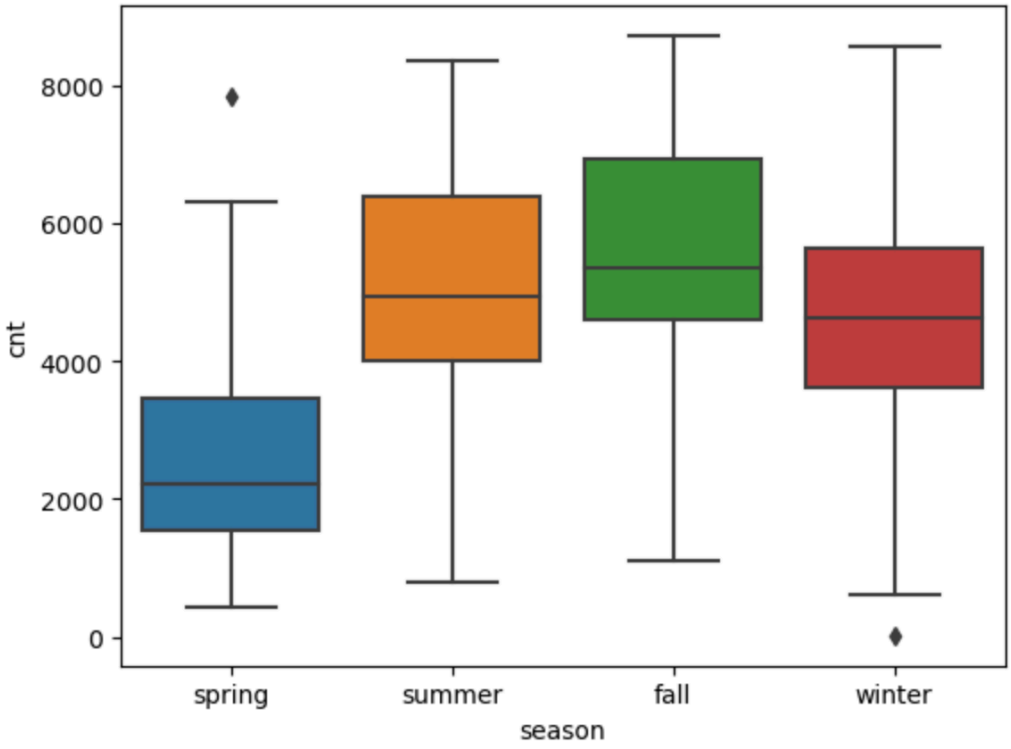
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. From my analysis of the categorical variables “season” and “weathersit” from the dataset, I could infer the following effects on the dependent variable:

Median cnt decreases for weathersit in the order : partly_cloudy, mist, light_rain as can be seen from the boxplot



Median cnt decreases for season in the order : fall, summer, winter, spring as can be seen from the boxplot



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

Ans. It is important to use drop_first = True during dummy variable creation for redundancy removal. As even after dropping first variable, all others being 0 represents the first one being true.

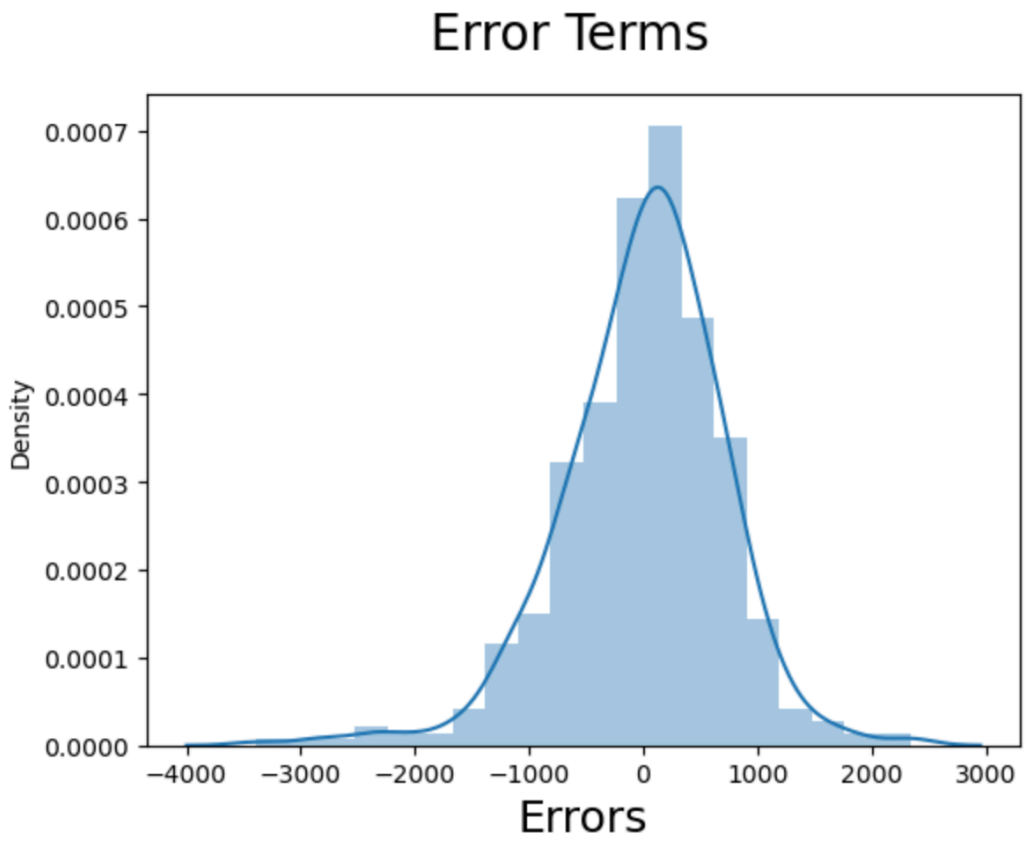
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Looking at the pair-plot among “independent” numerical variables, temp and atemp have the highest correlation with the target variable cnt. Please note : we have not considered the numerical variable registered as it is also a dependent variable.

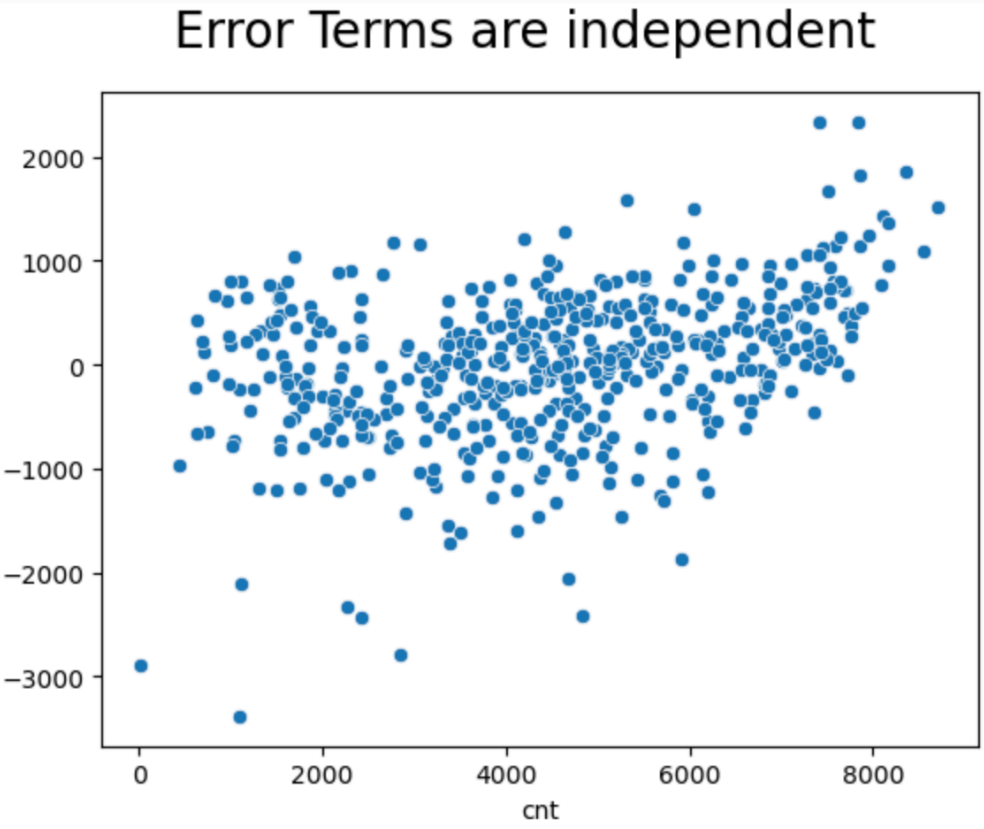
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. We validate the assumptions of Linear Regression after building the model on the training set as follows:

1) There is a linear relationship between independent variables and the target dependent variable as linear model fits it.
2) Error terms are normally distributed with mean 0.



3) Error terms are independent of each other.



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

Ans. Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature, humidity and windspeed with coefficients of 3285.78, -1309.58, -1594.24 respectively in the multiple linear regression model.

The adjusted r-squared of the train split is 0.849 and r-squared of the test split is 0.808 which are close to each other.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Broadly speaking, linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).
Two types of linear regression:

- Simple linear regression – Number of independent variables is 1
- Multiple linear regression – Multiple independent variables

Linear regression algorithm is an algorithm to predict a linear relationship between dependent (target variable) and independent variables (predictors). It uses gradient descent technique to minimize the error term associated with the linear model.

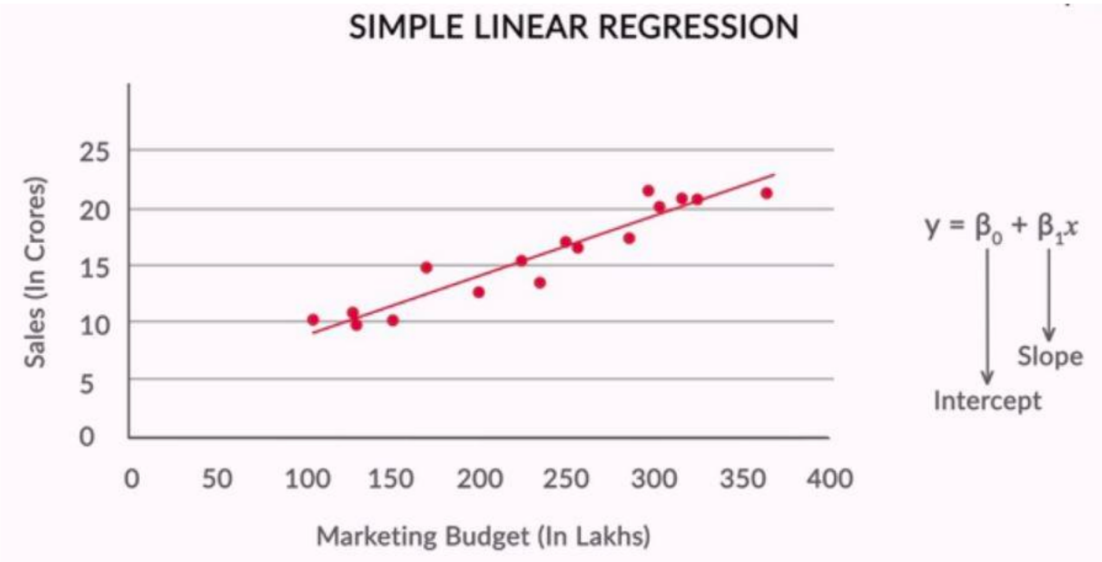
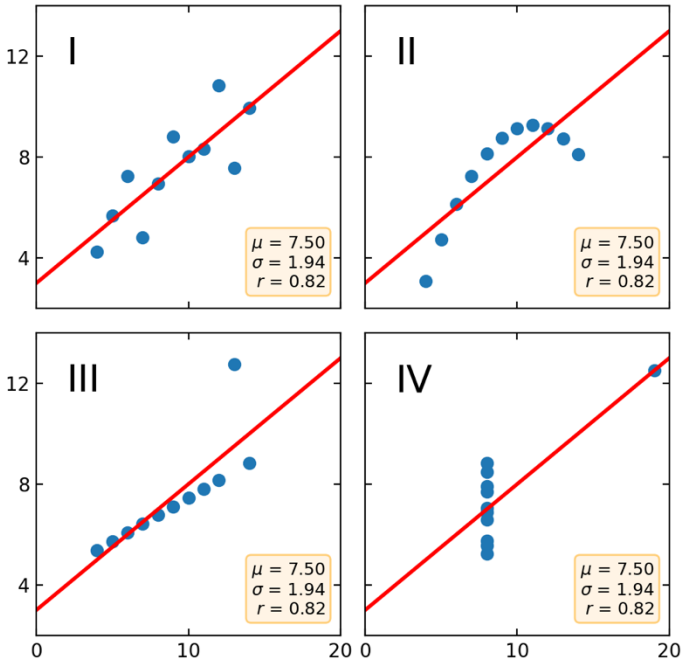


Figure 3 - Regression Line

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



3. What is Pearson's R?

Ans. In statistics the Pearson's R or the Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

Ans. Scaling is the changing the scale of a feature to a more easily interpretable one e.g. changing to a scale of 0 to 1.

Scaling is performed for the following reasons:

1. Ease of Interpretation
2. Faster convergence of gradient descent methods

Difference between normalized scaling and standardized scaling is that scale of normalized scaling is between 0 and 1 while it is not true for standardized scaling. In standardized scaling the mean is 0 and standard deviation is 1.

Standardized scaling : xnew = (x – mean(x))/sd(x)
Normalized scaling : xnew = (x – min(x))/(max(x)-min(x))

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

Ans. This happens because the Ri^2 = 1 thus making VIF = 1 / (1-Ri^2) infinite. This happens when the independent variable whose VIF is to be calculated by measuring the relationship against the remaining independent variable has an R square of 1.

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

Ans. In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical methid for comparing two probability distribution by plotting their *quantiles* against each other. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate).

The use and importance of a Q-Q plot in linear regression is that we can conclude that x and y come from similar distribution if all points of quantiles lies on or close to a straight line at an angle of 45 degree from x-axis.