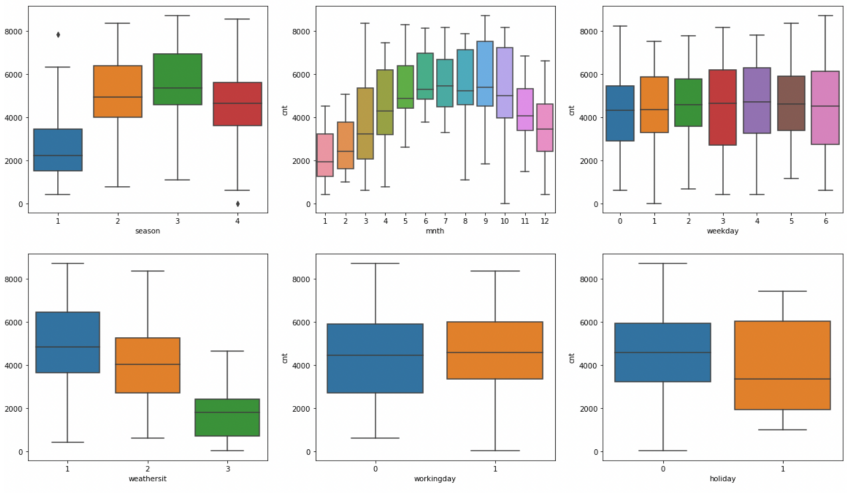
Assignment-based Subjective Questions

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

Answer: Dependent variable Cnt is dependent on categorical variable as below:

- Count Seems to be in maximum in Fall (Autumn) followed by Summer, Spring & Winter respectively.
- Count is more on Clear & Misty Days as compared to Light Snow / Rainfall.
- Count is more in months of 5,6,7,8 & 9.
- Count on Working Day / Non-Working Day has almost same median.



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

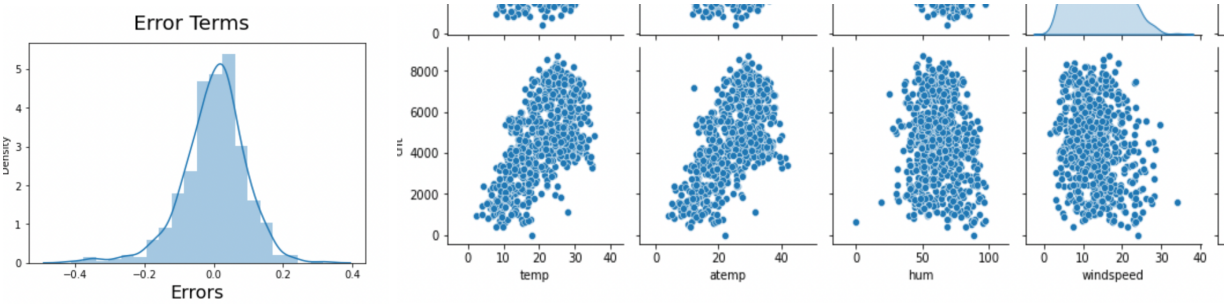
Answer: Basically, the idea is to create n-1 dummy variables if we have n levels in a categorical variable. We drop the first one to reduce the redundancy among the dummy variables.

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

Answer: Temp and Atemp have correlation with count of around 0.63

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

Answer: We validated with distribution plot that error terms are normally distributed. And pair plots to determine linear relationship.



We validated VIF values are within permissible range to verify that there is no multi collinearity. We also validated that error terms have constant variance.

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 for explaining demand of the shared bikes are Temperature, Season and Weathersit as Cloudy, Mist, Light snow and rain.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression algo falls under the Supervised learning method in machine Learning. The basic purpose of Linear Regression is to predict the dependent variable(y) based on a given variable(x). In this module we have studied 2 types of Linear regression:

a. Simple Linear Regression
b. Multiple Linear Regression

Simple Linear Regression: In simple linear regression variable y and x can be represented with the help of a Straight Line and they have a linear relationship with each other. y is dependent on one x variable. It can be represented by below:

$$Y = \beta 0 + \beta 1X$$

Multiple Linear Regression: In multiple linear regression variable y is dependent on two or more x variables. It can be represented as below:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn$$

We find the best fit regression line by minimizing the Cost function or the Residual sum of squares. We can also measure the best model by calculating R2 and Adjusted R2.

Also, before predicting a test variable through model, we have different assumptions in Linear Regression which are supposed to be verified. Once all the assumptions are fulfilled, we predict the values of dependent variable.

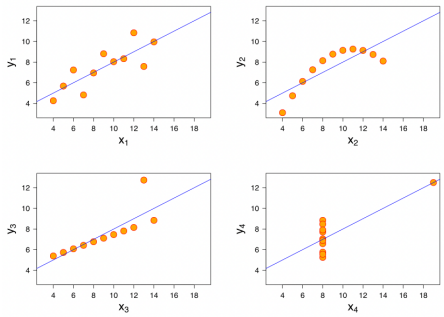
2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet was introduced by Statistician Francis Anscombe to explain the importance of graph over statistical calculations. It consists of 4 datasets which has almost same mean, sample variance, correlation and R square but when plotted over a graph demonstrate different characteristics. Below are the four datasets.

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

Below the graph which they display when plotted:



3. What is Pearson's R?

Answer: Pearson's R is a method of calculating linear association between 2 continuous variables. It is given by below formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The value of r ranges from -1 to 1. If the value is positive that means the variables move together in positive direction otherwise, they move together in opposite direction. If the value is 0 then the 2 variables are not associated with each other.

The limitation of Pearson's R is that it cannot determine non-linear relationship and does not differentiate between dependent and independent variables. The following guidelines have been defined for Pearson's R:

Strength of Association	Coefficient, *r*	
	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

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

Answer: Scaling is basically changing the magnitude of data values within a desired range. If we have a lot of independent variables with different scales of values then without scaling the coefficients calculated might be very difficult to predict. So scaling is performed on variables. It also helps in easy interpretation. After scaling Gradient Descent becomes much faster.

Difference between normalized scaling and standardized scaling is Normalization rescales the value into the range of [0,1]. This is also known as min/max scaling. Whereas, Standardization rescales data to have mean of Zero and standard deviation of 1.

The important to note is that Scaling affects just the coefficients not any of the statistical values like t-value, p value etc.

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

Answer: Infinite VIF represents that the two variables are perfectly correlated. The R2(R-square) value is 1. In such case we need to remove one of the variables. In this case the assumption of Multicollinearity fails for Multiple regression model.

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

Answer: Q-Q plot or Quantile- Quantile plot is a plot of quantiles of first data set against the second data set. It helps to determine if the 2 different data set came from the same distribution or not. If they are from the distribution then points will lie on the reference line. If the 2 distributions plotted are similar the points on Q-Q plot lie on the line y=x, if they are linear then they will follow a line but not necessarily y=x.

The use of Q-Q plot is: If let's say client provides the Train and Test data separately then using Q-Q plot we can confirm if this data is from the same population or distribution.