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 Visualizing Season, Month, Year, weather sit and weekday

   - Spring season has the Minimum count, Fall has Maximum count, Summer and Winter has Moderate count
   - There were no values for heavy rain or snowfall but highest count was recorded in Clear and Partly Cloudy
   - Count reduced During holidays
   - September month has highest number of count and December has least count
   - Count is increasing each year.

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

   Ans:

   Drop first will drop first column after creating dummies for all categories. As other categories combined can explain the dropped categories. In a way its redundant value. And if the column is considered for model building, there will be multi-collinearity among the independent variables.

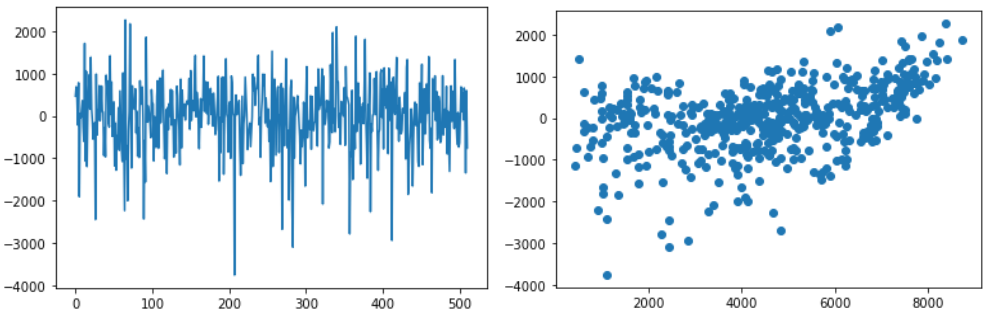
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

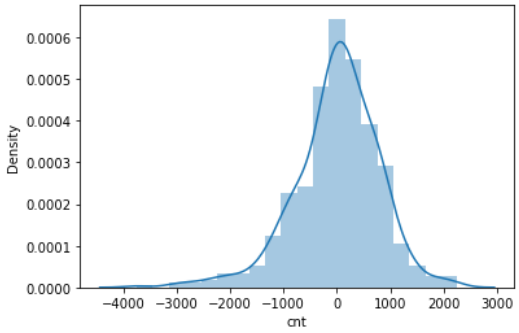
   Ans:

   “temp” and “atemp” are highly correlated to the target variable (cnt)

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

   Ans:





The Residuals doesn't follow any pattern in the plot. And the distribution of the residual is normally distributed with mean value as 0. Thus, we can validate the assumptions of Linear Regression.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans:
   Top 3 features:
   - Yr
   - Temp
   - weathersit_bad

General Subjective Questions:

1. Explain the linear regression algorithm in detail?
   Ans:
   Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear regression is based on the popular equation "y=mx+c".

   There are 3 assumptions associated with a linear regression model:

   1. The relationship between X and the mean of Y is linear.
   2. The variance of residual is the same for any value of X.
   3. Observations are independent of each other.

   In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

   Regression algorithm is classified as simple linear regression and multiple linear regression.

   1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

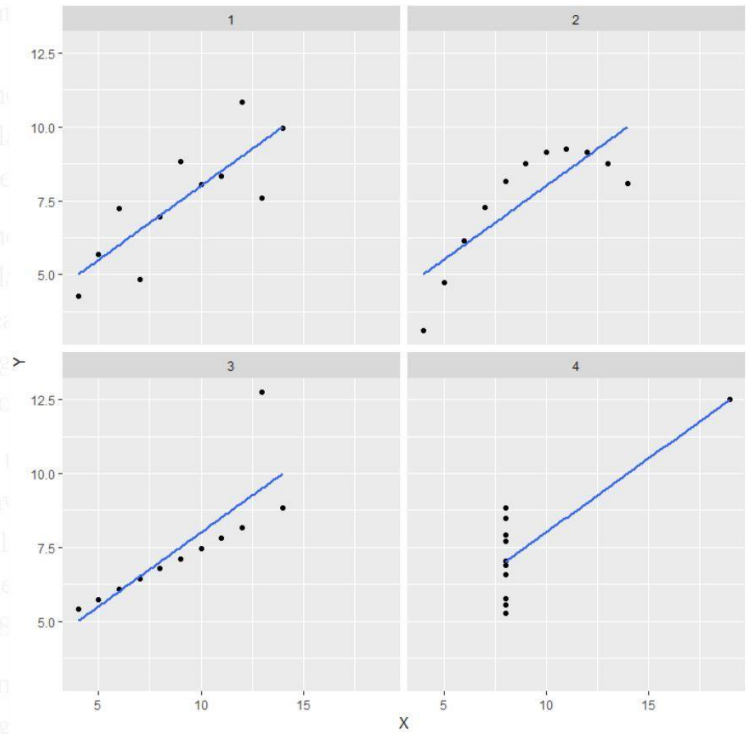
The equation for MLR will be:
Y = B0 + B1 X1 +B2 X2 + B3 X3

B1 = coefficient for X1 variable
B2= coefficient for X2 variable
B3 = coefficient for X3 variable and etc..
B0 is the intercept.

2. Explain the Anscombe's quartet in detail?

Ans:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. It demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



- In the first one, the scatter plot seems to be a linear relationship between x and y.
- In the second one, that there is a non-linear relationship between x and y.
- In the third one, there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one has high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

   Ans:

   Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data.

   If, r = 1 means the data is perfectly linear with a positive slope

   r = -1 means the data is perfectly linear with a negative slope

   r=0 means there is no linear association.

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

   Ans:

   Scaling is a method to bring the independent variables dataset under same range of values.

   Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

   Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

   Ans:

   The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF)=1/(1-R_1^2). If there is perfect correlation, then VIF infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

   Ans:

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

   The q-q plot is used to answer the following questions:

   Do two data sets come from populations with a common distribution?

Do two data sets have common location and scale?
Do two data sets have similar distributional shapes?
Do two data sets have similar tail behaviour?