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?**

   The categorical variables that were present in the dataset were:

   - Season: The bike rental count was high during the Fall season, with a good increase during the Fall of 2019 compared to the previous year. Followed by summer, winter and spring in total.
   - Month: The months June to September has the highest number of bike rentals, with June and September being equally highest, compared to other months and hence this duration is to be monitored. Also the least being January followed by February.
   - Weather: It is clear that when the weather is good or clear, the bike rental count is the highest followed by misty. Hence the weather condition is a good predictor to be noted.
   - Weekday: We see that the bike rentals on Thursday is the highest almost the same towards the weekend, but with this we can't conclude whether this could be good predictor or not as there is no drastic fall or hike in the number for any day. Hence, we need further analysis.
   - Holiday: The number of bikes rented during a non-holiday is more compared to that of a holiday.
   - Workingday: The number of bike rental is more during a working day compared to a holiday. But we need to validate it's significance in model building.
   - Year: From the previous observations and from the above graph we can say that, with reference to all the other variables, the number of bike rentals is higher in 2019 than in 2018.

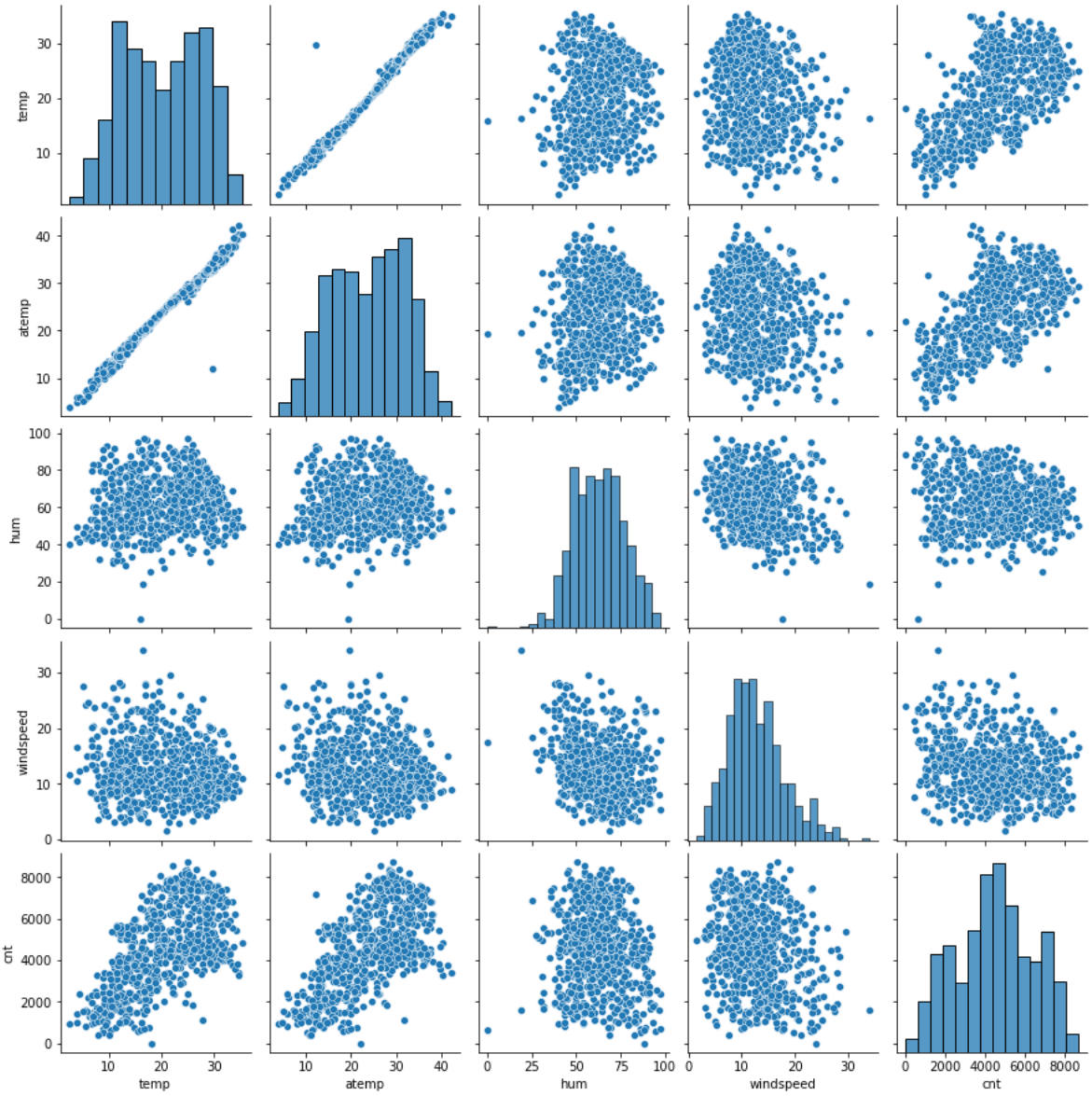
2. **Why is it important to use drop_first=True during dummy variable creation?**

   While encoding categorical variables, for example seasons, we have 4 different levels such as spring, summer, fall and winter. When we encode this data manually, we use 4 digits. But it is always understood that if it is not summer, fall or winter, by default is spring and hence we just need 3 columns to describe them rather than 4.

   With this reduced number of columns, we can reduce redundancy and in turn reduce autocorrelation. If we don’t drop a column, this may lead to an inefficient model as there can exist mulit-collinearity between features and also a best fit model will be tedious to achieve.

   Hence, for variables with n- levels, n-1 columns are enough to create the dummy variables.

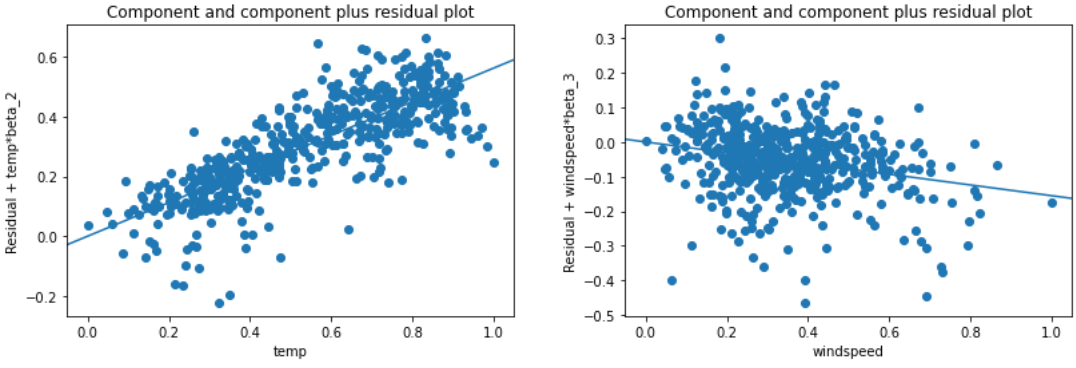
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**



There is a strong correlation between temp and cnt and also between atemp and cnt. Hence, temp and atemp are having a high correlation with the target variable, cnt.

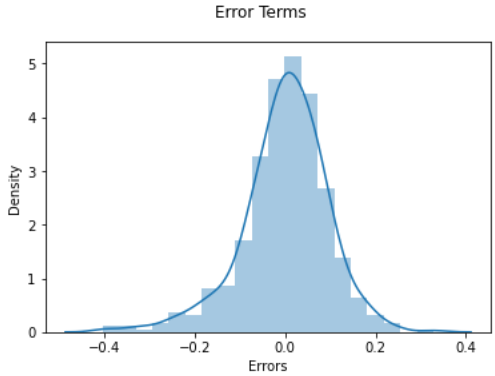
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

a. Linear Relationship:



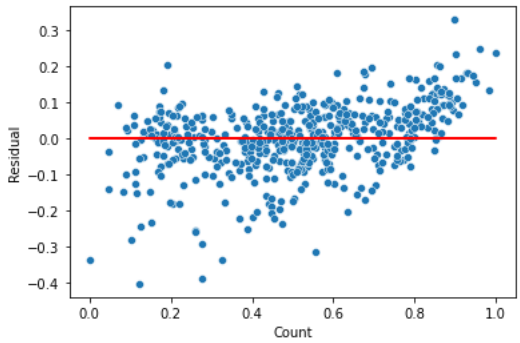
From the above graphs we can see that the linearity model and the predictor variable is preserved.

b. Residual analysis of the train data:



It can be seen that the error terms are normally distributed with mean=0.

c. Independent error terms and homoscedasticity:



From the above graph it can be seen that the error doesn't follow any pattern and hence independent of each other and is homoscedastic in nature.

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

Based on the final model, the top 3 predictor variables that is significant in BoomBikes bike bookings are:

a. Temperature (Temp): A coefficient value of 0.561862 indicates that a temperature has significant impact on bike rentals

b. Light Rain & Snow (weathersit =3): A coefficient value of -0.302220 indicates that the light snow and rain stops people from renting the bikes which can be because they might get affected by the weather condition like rain or snow.

c. Year (yr): A coefficient value of 0.230534 indicates that there has been a substantial increase in the rentals in the subsequent year compared to the previous year.

General Subjective Questions

1. **Explain the linear regression algorithm in detail**.

   Linear Regression is a supervised machine learning algorithm where it mainly used for prediction purpose. It has a target variable which needs to be predicted either one or various independent variables. In Linear regression, the main assumption is that when the independent variables (X) taken one at a time, has a linear relationship with the target variable (y).

   Plotting a best fit line between X and y is called the regression line. The equation for this line is

   y=mX+c

   where y is the dependent variable, X= the independent variable, m=slope of the line and c=y-intercept.

   In ML, the linear regression is divided into 2 categories:

   a. Simple linear regression: This is the simplest algorithm where it has one dependent and one independent variable. The straight line is plotted which is represented as below:

      $$y = \beta_0 + \beta_1 X$$

      Where β_0 is the constant (y intercept)

      β_1 is the coefficient of the independent variable to be found as part of the model building

   The best-fit line is found by minimizing the RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

   $$e_i = y_i - y_{pred}$$

   $$RSS = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_i X_i)^2$$

The strength of a LR model can be found using the:
R-squared method: it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. It always lies between 0 and 1. Higher the value, better the fit.

   b. Multiple linear regression: Similar to SLR, the MLR is used to predict one target variable with multiple independent variables.
      The equation of MLR is

      $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \ldots \beta_n X_n$$

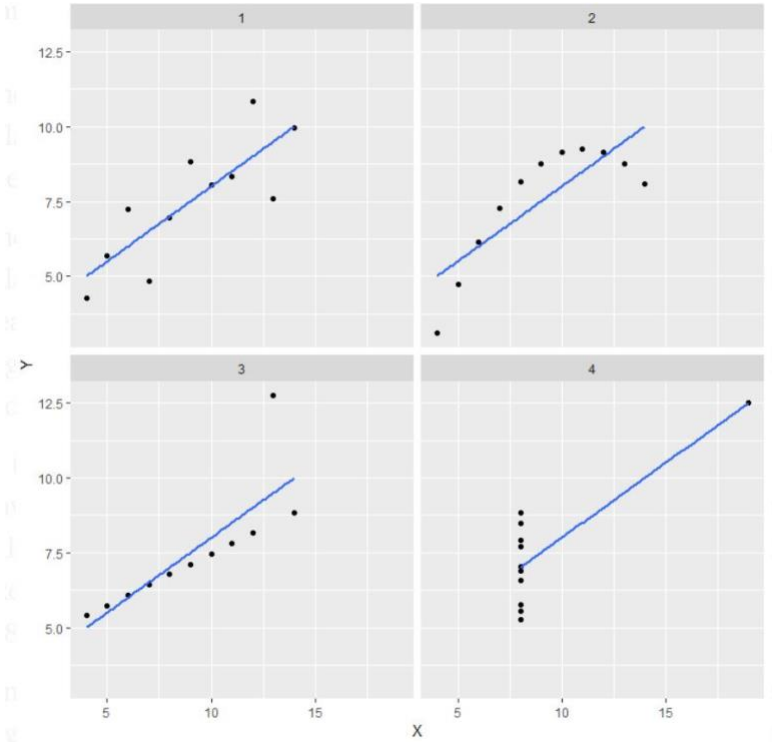
The assumptions to be made before LR are :

- Linear relationship between dependent and independent variable
- Error terms are independent to each other and doesn't follow any pattern

- Error terms are normally distributed with mean=0
- Follows Homoscedasticity.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a quartet consisting of 4 datasets. These 4 datasets have identical basic statistical properties like mean, standard deviation. But when these datasets plotted, look entirely different. Francis John "Frank" Anscombe, a statistician, found 4 datasets of 11 data points which had identical statistical parameters, but when plotted they looked very different from each other.



The above graph is a simple example of the Anscombe's quartet.
In fig 1, (top left), the graph appears more like a linear regression
In fig 2, (top right), there is a non-linear relation between the variables
In fig 3, (bottom left), there lies a outlier which makes the line not a best fit for the dataset
In fig 4, (bottom right), there is a high leverage point which produces high correlation coefficient.

The main reason for the emergence of this concept was to understand the importance of graphical interpretation of a dataset before statistically analysing them as simple statistical parameters are quite deceiving.

3. What is Pearson's R?

The Pearson's R is a statistical measure of linear relationship between 2 variables. It is a normalized measure of covariance between 2 variables. The value of Pearson's R lies between 1 and -1.
1=perfect positive relationship
-1=perfect negative relationship
0=no linear relationship

The Pearson's R between 2 variables is symmetric in nature. The Pearson's R value is independent of units of measurement.

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

Scaling is the process of normalizing the data within the range of an independent variable. There will be columns in a dataset which higher integer values while some have smaller integer values. Eg, in the housing data set, area had a higher integer value while other variables had smaller values. If scaling is not done, there might be chances that the final model equation weighs the variables smaller if it has smaller integer values and weigh higher if the other way. Hence, by scaling we bring all the variables in the same comparable scale.

- **Normalized scaling**: It is also called as Min-max scaling. It is calculated as:

```
X_new = (X - X_min)/(X_max - X_min)
```

The values range between [1,-1] or sometimes [0,1].
This technique is chosen when there are no outliers in the data as they can't handle outliers. This is used when features are of different scale. This is useful when we don't know the distribution of the dataset.

- **Standardized scaling**: This technique is useful when the data follows a normal of Gaussian distribution (not necessarily true). The formula is:

X_new = (X - mean)/Std

It is not bounded by a certain range and much less affected by outliers.

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

Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating VIF is :

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared =1, which lead to 1/(1-R1) =infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

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 shaped of the distribution.

A q-q plot is a scatterplot created by plotting 2 sets of quantiles against one another. If both sets come from the same distribution, we should see the points forming a line that's almost a straight.

This plot useful to know

- The source of the datasets (from same or different population)
- Common location and scale of the data sets
- Similar distributional shapes
- Similar tail behavior of the data sets.