INFERENCES

From the dataset given, we observe that there are 4 categorical variables contributing to the bike demand. These are:

1. Season
2. Month
3. Weekday
4. Weather Situation

On observing there trends w.r.t. cnt (dependent variable), we come across the following graph

Machine generated alternative text:
Mar May "n Nuv 

**From this graph, we make following conclusions:**

1. **Season:** We see that the sales of bike is highest in fall and lowest in spring. It is more in summer than winter. Indicating that this could be one of the driving factors for the target variable.
2. **Month:** All the months have different variation for cnt of the bikes. It being highest for the month of September, followed by October to being lowest for January.
3. **Weekday:** There is some trend of target variable even with weekday, but yes, the cnt is more for weekdays than weekends. This is an interesting insight and should be considered will model creation.
4. **Weather Situation**: We observe that there is absolute no count for severe weather condition. Good substantial number or Good weather situation and then decreases for Bad weather situation.

**drop\_first=True** is nothing but **dummy encoding** and **drop\_first=False** is **one hot-encoding.**

When we create dummy variables we are actually replicating the existing variables with dummy identity using get\_dummies function. In this process of creating dummy variables we get all distinct values as variables for that column whereas we need only n-1 variables to explain n variables.

Say for example, in the dataset given, we had a feature named **season, with 4 possibilities i.e. Spring, Summer, Fall and Winter.** Once I create dummies for season, I get all these 4 distinct variables. Though if we see logically, we can actually drop one, as the values of three can very accurately explain all 4 possibilities. Say, if spring, summer, winter all are 0 this implies that Fall is 1.

So, **drop\_first=True**

* helps us in achieving this standard format.
* Moreover, if we don’t drop the first column then the **dummy variable will be correlated** (redundant). This may affect the model adversely and the effect would be more stronger when the cardinality is smaller.

Pair plot shows the **highest correlation of temperature (temp) and feeling temperature** (atemp) with the count. (target variable). But since, temp and atemp are collinear and follow the same trend with target variable. We drop atemp and use temp for the final model creation.

Correlation Image to explain the result above

Machine generated alternative text:
0.0082 
-0.0029 
0.049 
0.047 
-0.11 
-0.012 
0.57 
0.0082 
-0.25 
-0.029 
-0.033 
-0.016 
0.0063 
-0.069 
-0.0029 
-0.25 
0.053 
0.053 
0.023 
-0.019 
0.063 
0.049 
-0.029 
0.053 
0.99 
0.13 
-0.16 
bmp 
0.047 
-0.033 
0.053 
0.99 
O. 14 
-0.18 
ümp 
-0.11 
-0.016 
0.023 
0.13 
O. 14 
-0.25 
-0.099 
-0.012 
0.0063 
-0.019 
-0.16 
-0.18 
-0.25 
-0.24 
0.57 
-0.069 
0.063 
-0.099 
-0.24 

In order to validate the assumptions of linear regression on the model I created, I did the following:

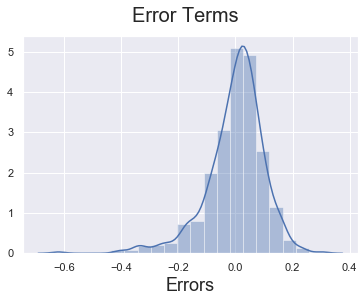
1. **Model creation**
   1. Firstly, talking about the model creation, we concluded the final best fit line equation to be:

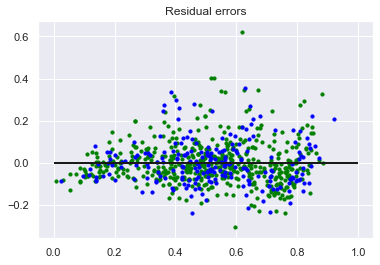
**cnt = 0.0788 + 0.2389 \* yr + 0.0482 \* workingday + 0.5515 \* temp - 0.1839 \* windspeed + 0.0875 \* summer + 0.1174 \* winter + 0.0874 \* Sep + 0.0593 \* Sat - 0.0677 \* Fair**

Here, we have 9 variables and the model has 79.1% R2 value for training set.

1. **Validation of linear regression assumptions:**
   1. We calculated the residual that is nothing but y\_train-y\_train\_pred
   2. Plotted the residual to see the graph characteristics.

A picture containing object, antenna

Description automatically generated



1. From the above plots we concluded the following;
   1. **There is a linear relationship between X and Y** and hence the multi-variate linear model fitted on the data gives a good R2 value and explains the variance in the data to a great extent.
   2. **Error terms are normally distributed** with mean zero (not X, Y). This is clearly evident from the second figure, where error terms follow a normal distribution.
   3. **Error terms are independent of each other,** as we can see form the graph above, we do not see any visible pattern for error terms
   4. **Error terms have constant variance (homoscedasticity),** the variance is not following any pattern as error term changes. Hence, we see a constant variance for the error terms.

Based on the final model created, I get set of 9 features significant enough to contribute to the model and having considerable VIF values.

The final model equation Is:

**cnt = 0.0788 + 0.2389 \* yr + 0.0482 \* workingday + 0.5515 \* temp - 0.1839 \* windspeed + 0.0875 \* summer + 0.1174 \* winter + 0.0874 \* Sep + 0.0593 \* Sat - 0.0677 \* Fair**

Out of these 9 features,

* **the top 3 contributing** to the model are **temp, yr and windspeed**.
* They have high correlation with cnt (dependent variable) and this also evident from their coefficient values.
* For**:**
  + **yr it is 0.2389 (r = 0.57)**,
  + **temp is 0.5515 (r=0.63)** and
  + **windspeed is -0.1839 (r = -0.24)**.
  + Yr and temp are positively correlated and windspeed is negatively correlated with the cnt.