Linear Regression Assignment Subjective questions

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 available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.
- “Season” –
  - As per the provided dataset, the best season for bike riding are summer and fall.
    - Hence, these two seasons should be first preference for planning.
- “workingday” –
  - Count of total rental bikes is slighty higher on `Weekdays` as compared to `weekends`.
- “weathersit” –
  - Count of total rental bikes is considerably higher on Clear weather days.
  - Count of total rental bikes is very low during Rainy days..
- “weekday” –
  - count of total rental bikes is slighty higher on `Thrusday` as compared to other days.
- “yr” –
  - This shows that count of the total rental bikes increases in 2019 when compared to previous year
- “mnth” –
  - Count of total rental bikes is low in months of Nov to Apr
  - Count of total rental bikes is observed maximum during May to Oct months

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

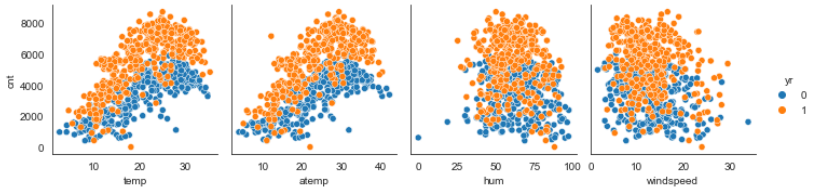
- Dummy variables are used to categorize or classify the values of categorical variable.
- pandas.get_dummies() has an argument drop_first which can be set to True or False, depending on if we want to drop the first level and get j – 1 dummies from j values of a categorical variable
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

- With the pair plots of the numeric variables and the target column, the variable temp and atemp seems to have the highest correlation with the target variable

```
In [25]:
# Plotting Pairplot to find Linear relationship between the cnt variable and other numeric variables

sns.set_style("white")
x =sns.pairplot(bikes_df, palette='tab10', x_vars=['temp', 'atemp', 'hum', 'windspeed'], y_vars=['cnt'] , hue='yr')
plt.show()
```

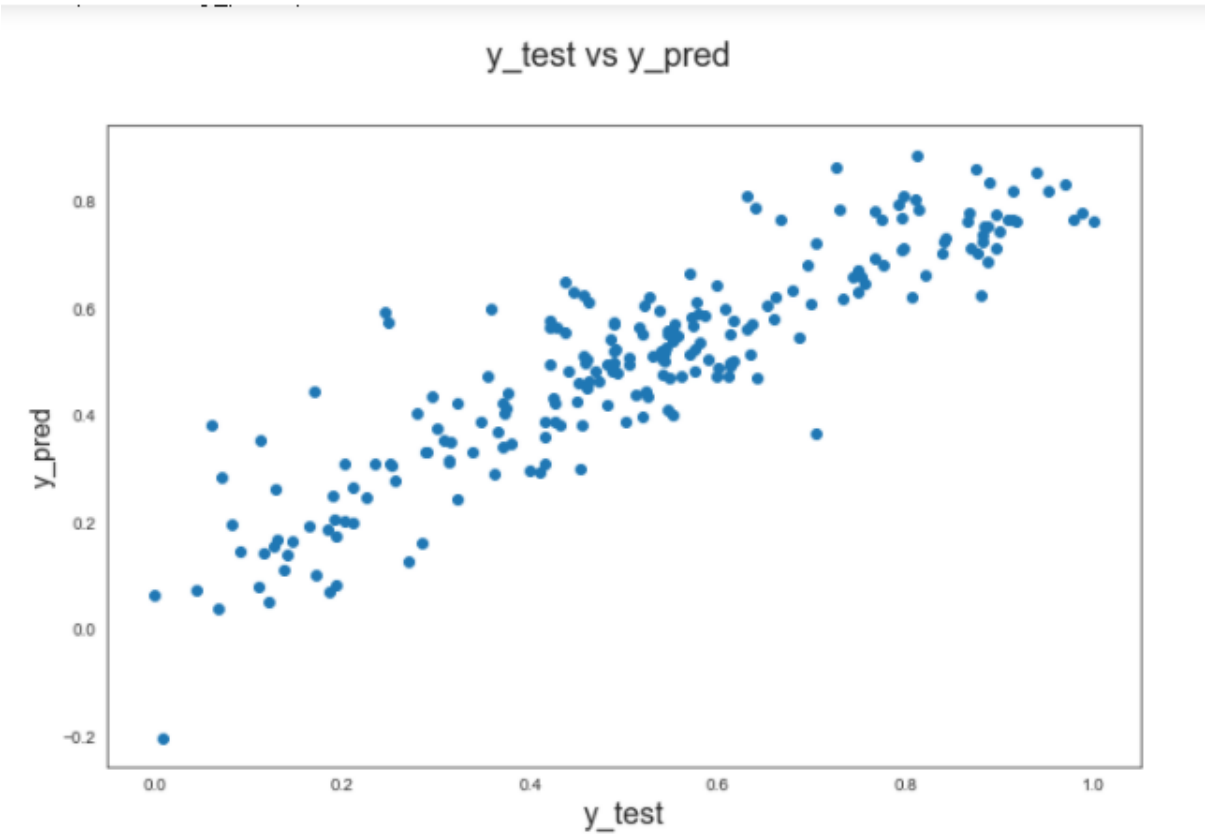


- Hence.as temp and atemp variables are highly co-related with each-other they follow similar relation pattern with `cnt` variable.

---

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

    1. **Linear relationship**– The linearity is verifed by plotting the graph, which symmetrically of the actual vs predicted plot.

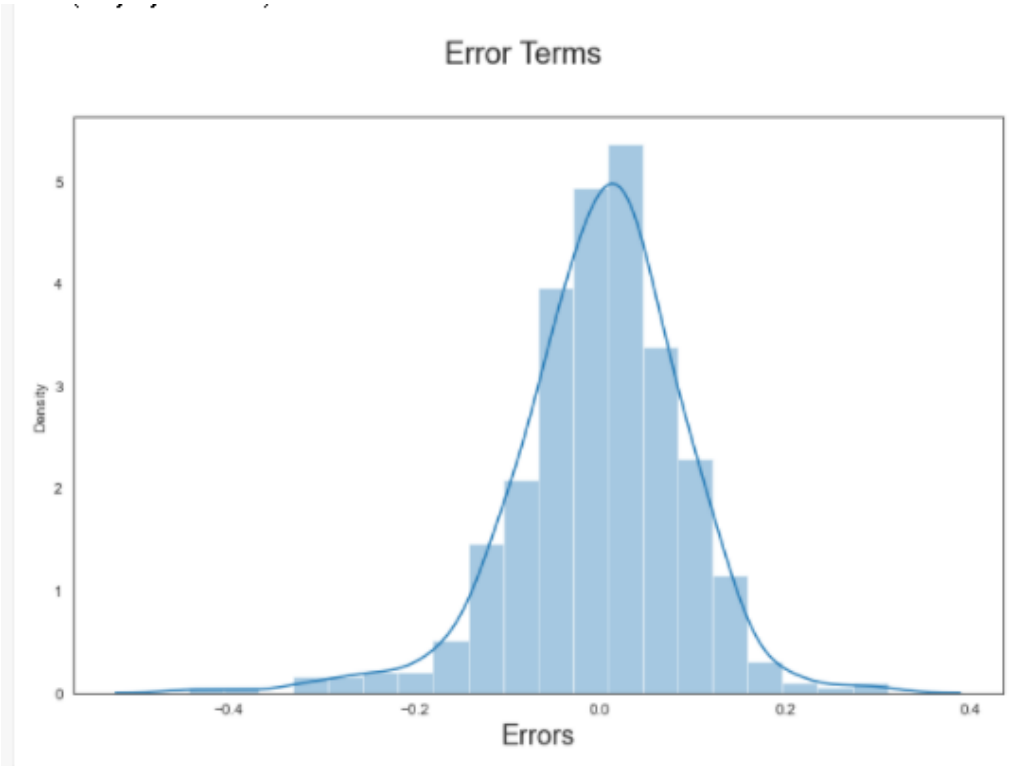


2. Multi-collinearity – We verified that there is little or no Multi-collinearity between the features by using VIF

In [55]:
```
#Printing VIF summary for the model
print(lm2_vif)
```

	Features	VIF
2	atemp	4.99
3	windspeed	3.83
5	season_winter	2.62
0	yr	2.06
4	season_summer	2.04
10	mnth_Nov	1.81
6	mnth_Aug	1.59
13	weathersit_Mist_cloudy	1.57
7	mnth_Dec	1.41
11	mnth_Sep	1.35
9	mnth_Jan	1.29
8	mnth_Feb	1.26
12	weathersit_Light rain_Thunderstorm	1.09
1	holiday	1.06

3. Error terms are normally distributed: As per assumptions of liner regression, Error terms should be normally distributed. Below histogram shows the same results.



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

The top 3 features that are contributing significantly towards the target variable are

- **yr,** count of total rental bikes is considerably higher in `2019` as compared to `2018`.Hence there is relationship between data of previous year and current year.
- **atemp** Count of total rental bikes is directly proportional to atemp
- **weathersit: -** count of total rental bikes is considerably higher on `Clear` weather days and decreases during `Rainy` days.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.
  - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
    - The straight-line equation is $Y = \beta_0 + \beta_1 X$
    - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i.
    - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum(y_i - \beta_1 x_i - \beta_0)^2$
  - The unconstrained minimization are solved using 2 methods
    - Closed form
    - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.

- $\text{RSS} = = \sum_{i=1}^{n}(y_i - y_{pred})^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

---

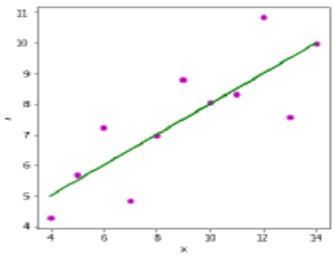
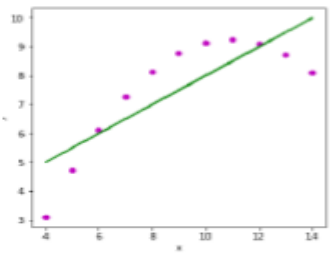
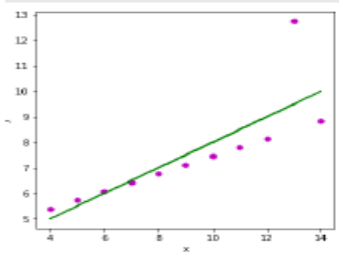
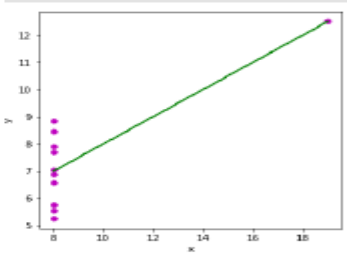
2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprise of four data sets that have identical simple statistical properties but appear very different when graphed.
- Each data set consist of 11 (x,y) points.
- They were constructed in 1973 by statistician Francis Anscombe to demonstrate the importance of graphing data before realizing it and the effect of outliers on statistical properties
- Once Anscombe found 4 sets of 11 data points in his dream and requested the council as his last wish to plot those points. Following are the sets of 11 data points

```
In [3]: df
```

Out[3]:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Explanation of the above graphs

- For 1st graph, there is linear relationship between x and y
- For 2nd graph, there is a non-linear relationship between x and y
- For 3nd graph,, there is a perfect linear relationship between x and y
- For 4th graph,one high data point exists which gives a high correlation coefficient

---

3. What is Pearson's R?

The Pearson's R is also known as Pearson's correlation coefficients. It is used to find the strength between the different variables and the relation with each other. Its value is always between -1 and 1. The interpretation of the coefficients are:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

$Σx^2$ = the sum of squared x scores

$Σy^2$ = the sum of squared y scores

---

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

- It is the data preparation step for model building. It normalizes these varied datatypes to one common data range.
- Most of the times the feature data have mismatch between variables and units of the variables. This results in to the high variance in units and ranges of data. Without scaling if processing the data without the appropriate unit conversion is done then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. Model prediction and accuracy of prediction stays constant after scaling.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

The reason for VIF to be infinite can only be in one condition, when R^2 is 1 then the VIF value also tends to infinite as denominator of below equation is zero.

$$VIF = \frac{1}{1 - R^2}$$

The situation arises when there is a perfect correlation between 2 independent variables.

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

- Q Q Plots (Quantile – Quantile plots) are plots of two quantiles. The range of values is divided into quantiles eg. 25%, 50%, 75%.
- The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. Interpretations
- If all the data points of quantile are at the straight line at an angle of 45 degree from x-axis.
    - Y values is less than X values: If y-values quantiles are less than x-values quantiles.
    - X values is less than Y values: If x-values quantiles are less than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - We can mention sample size in QQ plot.