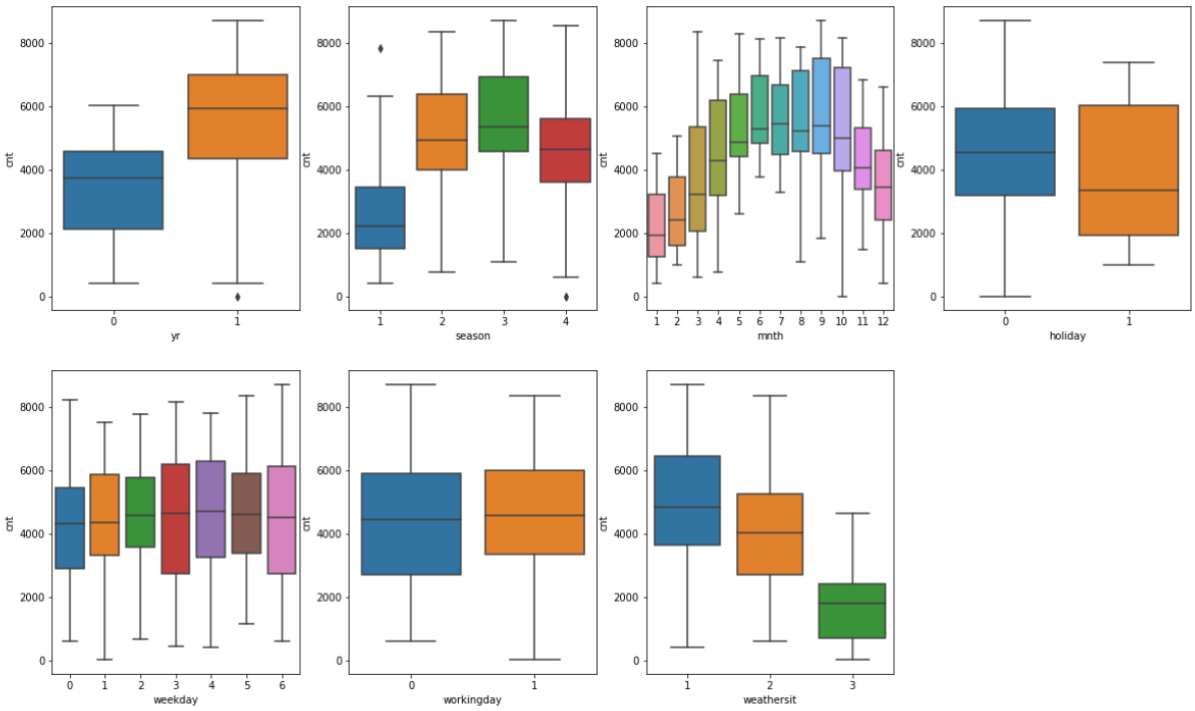
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

ANSWER:



From the above BoxPlots, we conclude below analysis:

yr and cnt:

Second year has more increase in customers count as it's median is on higher side in comparison to the first year.

season and cnt:

Bike sharing is more used in Season 2 and 3. It seems seasons 2 and 3 are have better climate to go out using Bikes. Season 3 has highest number of Bike sharing customers.

mnth and cnt:

month 6-9 has more bike sharing customers in comparison to the other month. It seems people are liking using bikes in summers. month 1 and 2 has very low numbers of bike sharing customers.

holiday and cnt:

People are using bikes more when it is not holiday. So, people are using bike for the work purpose more.

weekday and cnt:

weekday 4 has highest number of bike sharing customers. weekday 0 has lowest number of customers.

workingday and cnt:

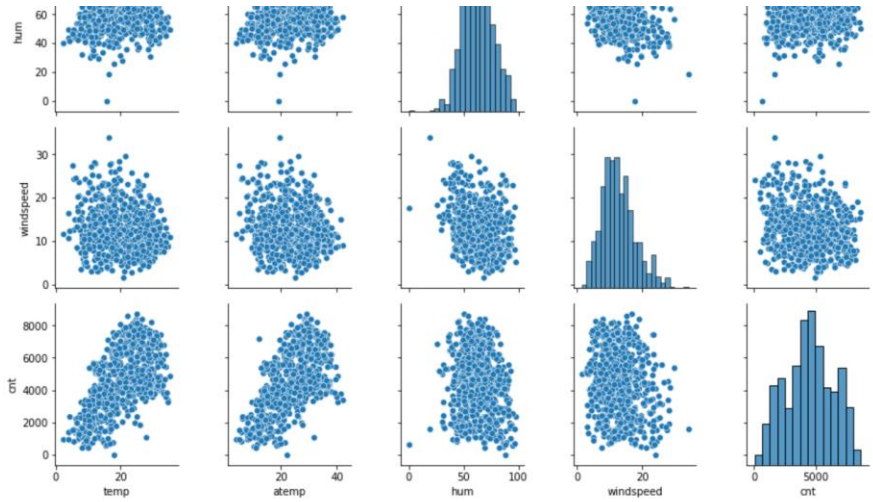
working day is inverse to the holiday. If it's working day, then the number of customers are more.

weathersit and cnt:

When the weather is Clear, Few clouds, Partly cloudy, Partly cloudy, then there is increase in the customer count. People like going out on bike in clear weather. People don't like using bike when weather has Light Snow, Light Rain, Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
ANSWER:
Drop_first=True, is required to drop extra column in you are doing categorization using the dummy variables. With drop_first = True will always create n-I columns when you have n categories of a variable. It reduces one extra column in the model and also help reduce the correlation among the variable while building the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
ANSWER:



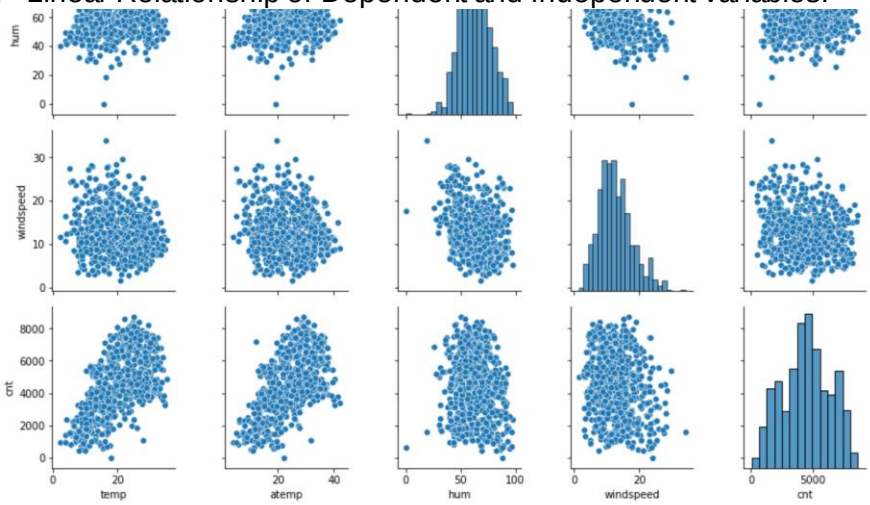
Looking in above screenshot from the notebook file, we can say that temp and atemp has highest correlation with the target variable 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANSWER:

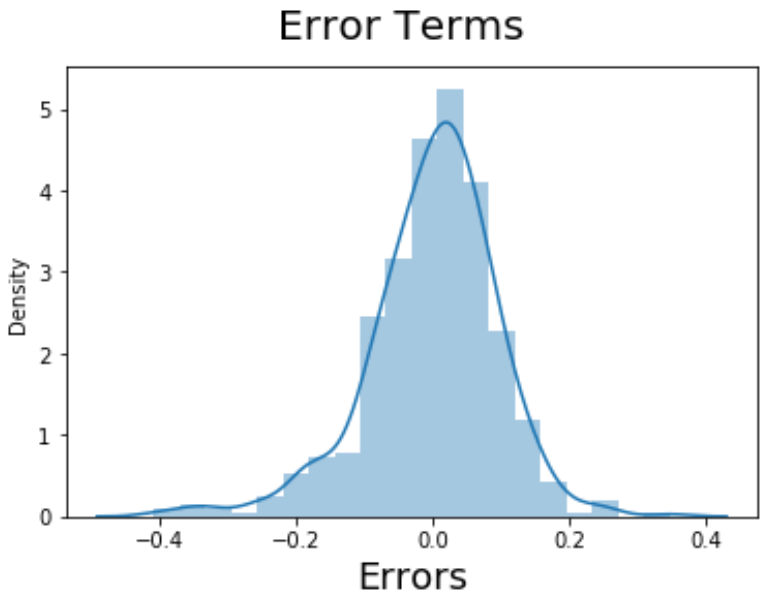
Below are few Linear Regression Assumptions and their validation in my assignment. For details you can refer the assignment Residual analysis section.

1. Linear Relationship of Dependent and Independent variables:



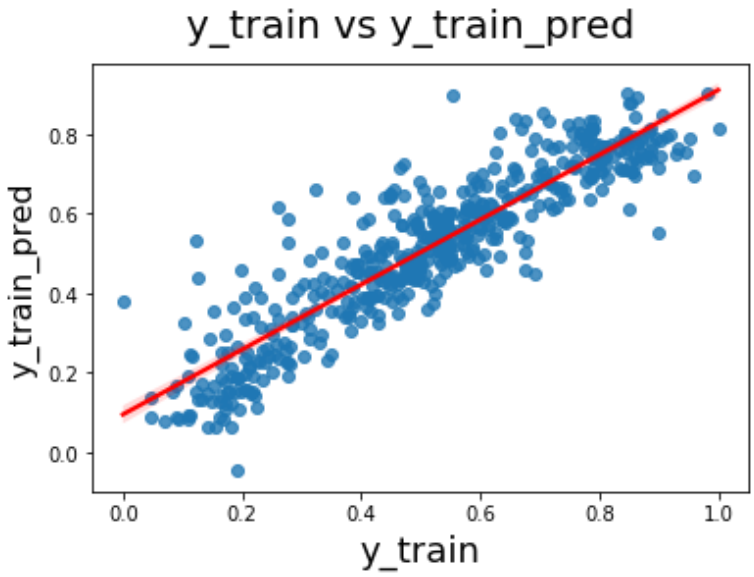
We can refer Bike sharing example pair plot, where we can see that 'atemp' and 'temp' variable are linearly related to the target variable 'cnt'. So, we have linear relationship in our data.

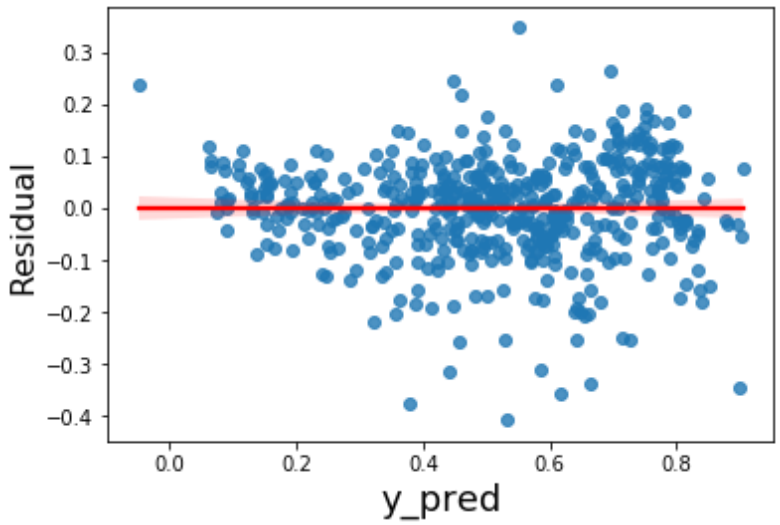
2. Normal Distribution of the Error terms



The above graph show that residual error terms are normally distributed. The mean is around 0.

3. Homoscedasticity (equal/constant variance) of residuals





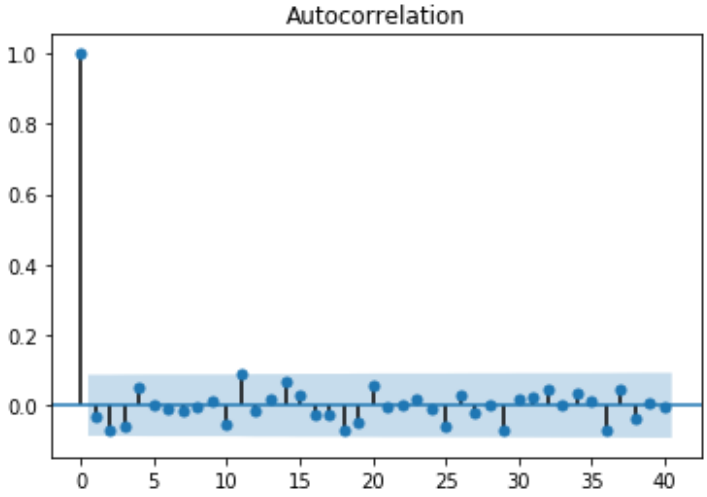
The outcome of the above plotting is that points are symmetrically distributed around a diagonal line in the former plot or around a horizontal line in the latter one. Based on both the above graphs we can roughly say that there is constant variance of error terms.

4. Checking the mean of the residual

lr_14.resid.mean() --- 6.16391469456104e-16

The mean of the residual is very close to 0. This passed the one of the assumptions of Linear regression that residual mean should be Zero.

5. No autocorrelation of residuals



6. No Multicollinearity between the variables.

ANSWER:

The VIF Score of the final model is below:

	Feature	VIF
0	const	33.95
4	hum	1.74
2	workingday	1.63
8	weekday_6	1.63
9	weathersit_2	1.54
3	temp	1.24
7	season_4	1.23
6	season_2	1.18
10	weathersit_3	1.17
5	mnth_9	1.11
1	yr	1.03

None of the variable has VIF more than 2. This indicates that Multicollinearity factor is very low among the independent variables.

Since there is no point outside the significant level of the graph, we can infer that there is no autocorrelation of residuals.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANSWER:

The Linear model build will have below equation:

cnt = 0.07535 X const + 0.225079 X yr + 0.043816 X workingday + 0.604548 X temp - 0.119122 X hum + 0.105277 X mnth_9 + 0.081059 X season_2 + 0.146614 X season_4 + 0.054996 X weekday_6 - 0.052457 X weathersit_2 - 0.291491 X weathersit_3

The Top 3 feature contributing based on above equation:

temp -- 0.604548 – Temperature is highest contributor for the model. It impacts the demand.

yr -- 0.225079 – Year Change has also good impact on the model.

weathersit_3 -- - 0.291491 --- When weather situation is bad, then it reduces the demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

ANSWER:

Basic Understanding Linear Regression

Linear Regression is the supervised Machine Learning model in which the **model finds the best fit linear line between the independent and dependent variable** i.e it finds the linear relationship between the dependent and independent variable.

Linear Regression is of two types: **Simple and Multiple**.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Multiple Linear Regression: Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

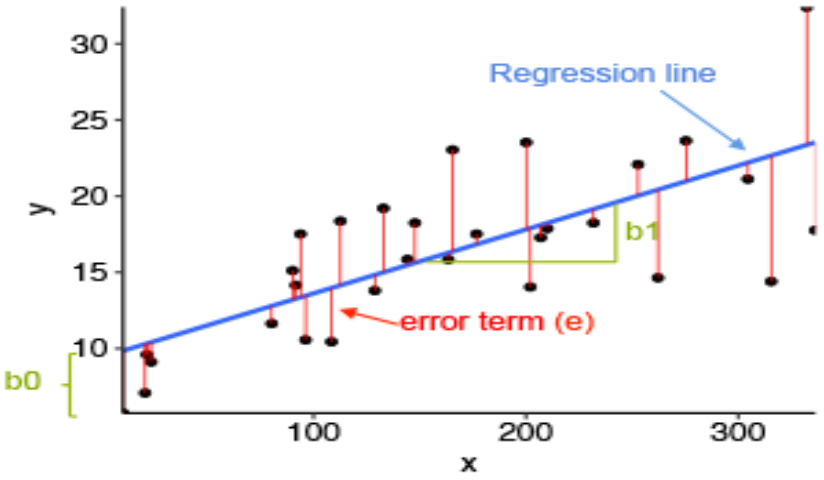
Equation of Simple Linear Regression, where b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_0 + b_1 x_1$$

Equation of Multiple Linear Regression, where b_0 is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.
Let's understand this with the help of a diagram.



In the above diagram,

- x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_0 is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.

The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as **the Sum of Residuals/Errors.**

Mathematical Approach:

Residual/Error = Actual values – Predicted Values
Sum of Residuals/Errors = Sum(Actual- Predicted Values)
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2
i.e

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

The strength of the linear regression model can be assessed using 2 metrics:

1. **R^2 or Coefficient of Determination**
2. **Residual Standard Error (RSE) R^2 or Coefficient of Determination**

You also learnt an alternative way of checking the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, 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. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as:

$R^2 = 1 - (RSS / TSS)$

RSS(Residual Sum of Squares): Instatistics,itisdefinedas thetotal sumoferroracross the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

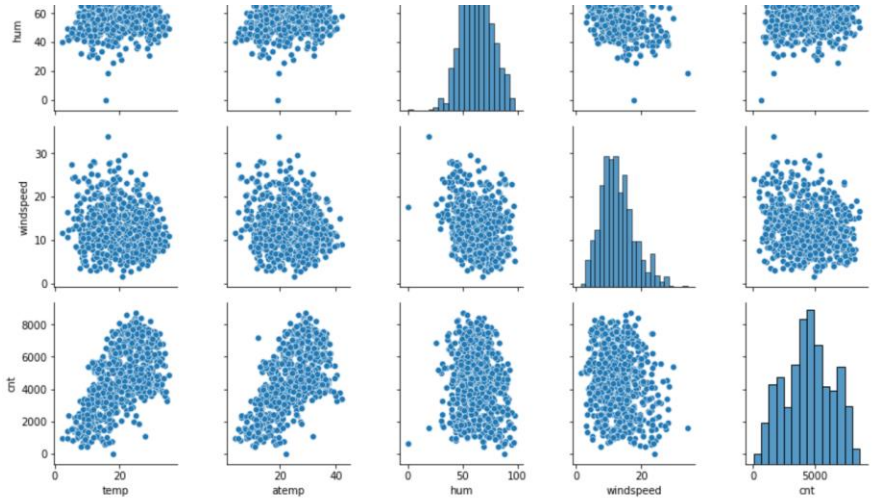
TSS(Total sum of squares): Itis thesumoferrorsofthedatapoints from meanof response variable.

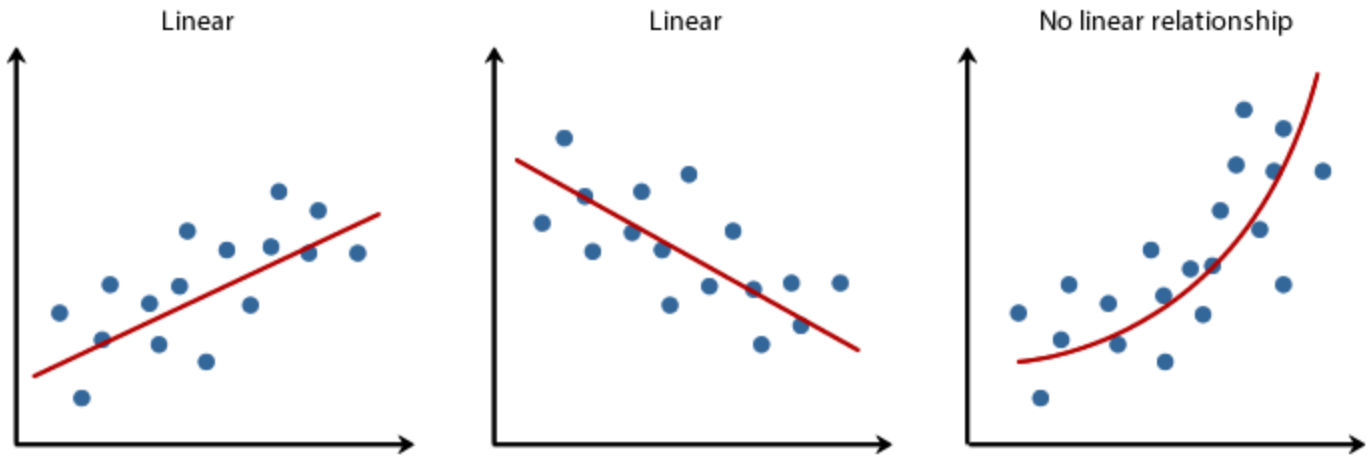
Assumptions of Linear Regression

The basic assumptions of Linear Regression are as follows:

Linear Relationship of Dependent and Independent variables:

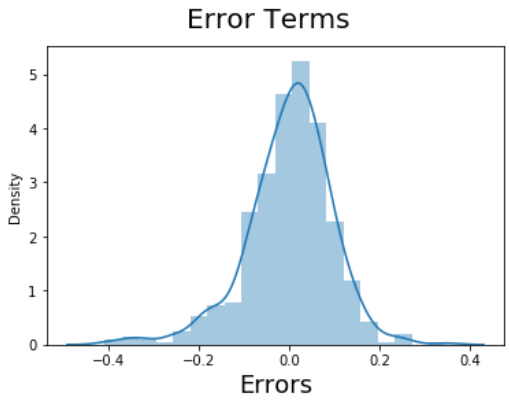
It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.





We can refer Bike sharing example pair plot, where we can see that 'atemp' and 'temp' variable are linearly related to the target variable 'cnt'. So, we have linear relationship in our data.

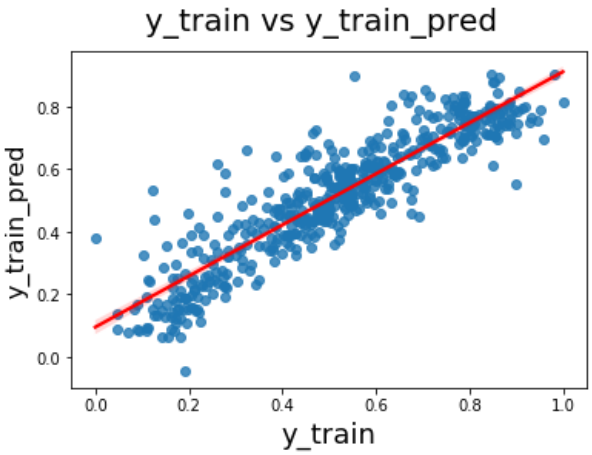
Normal Distribution of the Error terms: The error terms should be normally distributed.

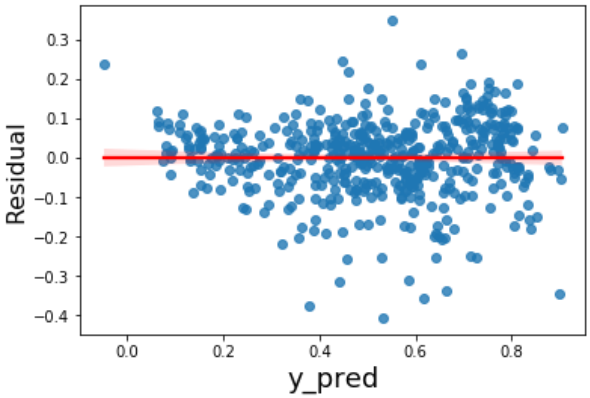


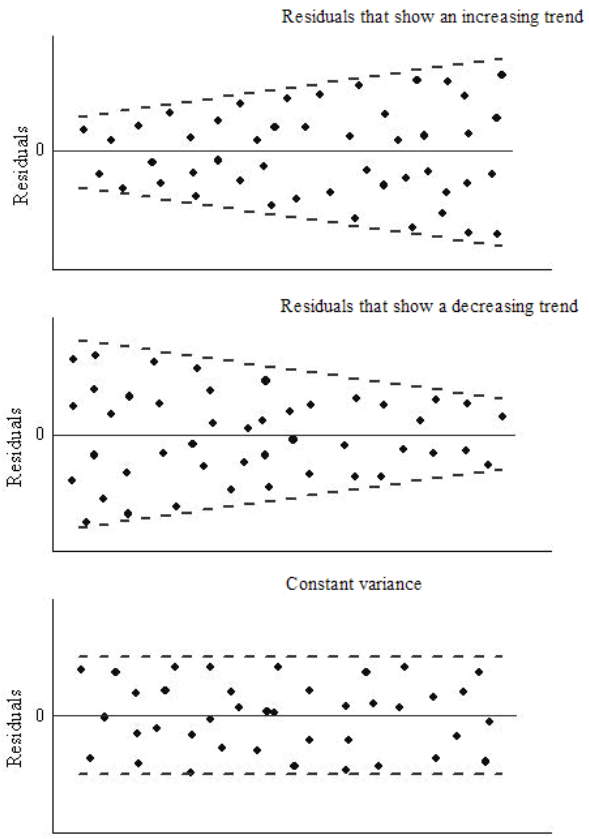
The above graph show that residual error terms are normally distributed. The mean is around 0.

Homoscedasticity (equal/constant variance) of residuals:

The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.


y_train vs y_train_pred
y_train_pred
y_train
0.0
0.2
0.4
0.6
0.8
1.0

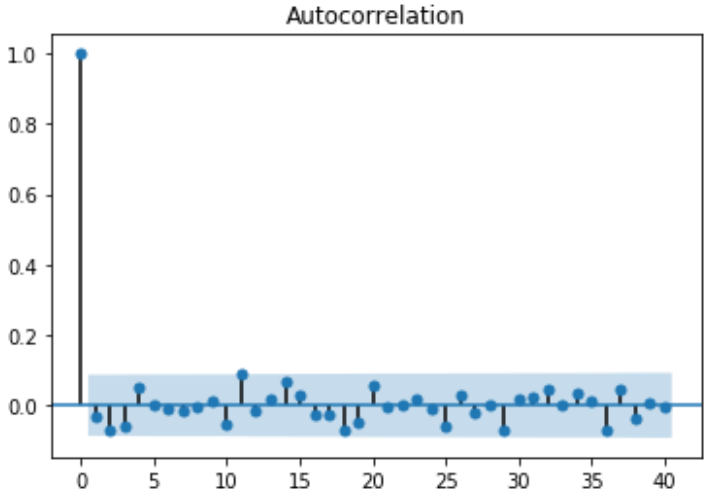

Residual
y_pred
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0.0
0.2
0.4
0.6
0.8



The outcome of the above plotting is that points are symmetrically distributed around a diagonal line in the former plot or around a horizontal line in the latter one. Based on both the above graphs we can roughly say that there is constant variance of error terms.

No autocorrelation of residuals:

The error terms should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.



No Multicollinearity:

The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.

The VIF Score of a model is below:

	Feature	VIF
0	const	33.95
4	hum	1.74
2	workingday	1.63
8	weekday_6	1.63
9	weathersit_2	1.54
3	temp	1.24
7	season_4	1.23
6	season_2	1.18
10	weathersit_3	1.17
5	mnth_9	1.11
1	yr	1.03

None of the variable has VIF more than 2. This indicates that Multicollinearity factor is very low among the independent variables.

2.Explain the Anscombe's quartet in detail. (3 marks)

ANSWER:

Wikipedia Definition:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

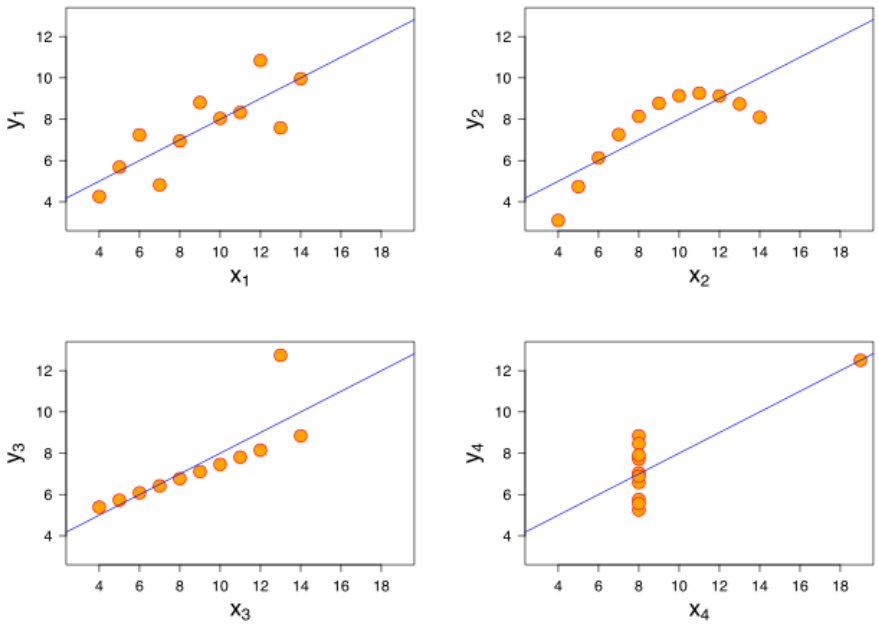
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R? (3 marks)
ANSWER:

Pearson's Correlation Coefficient ®

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

How is the Correlation coefficient calculated?

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula given is:

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

Where,

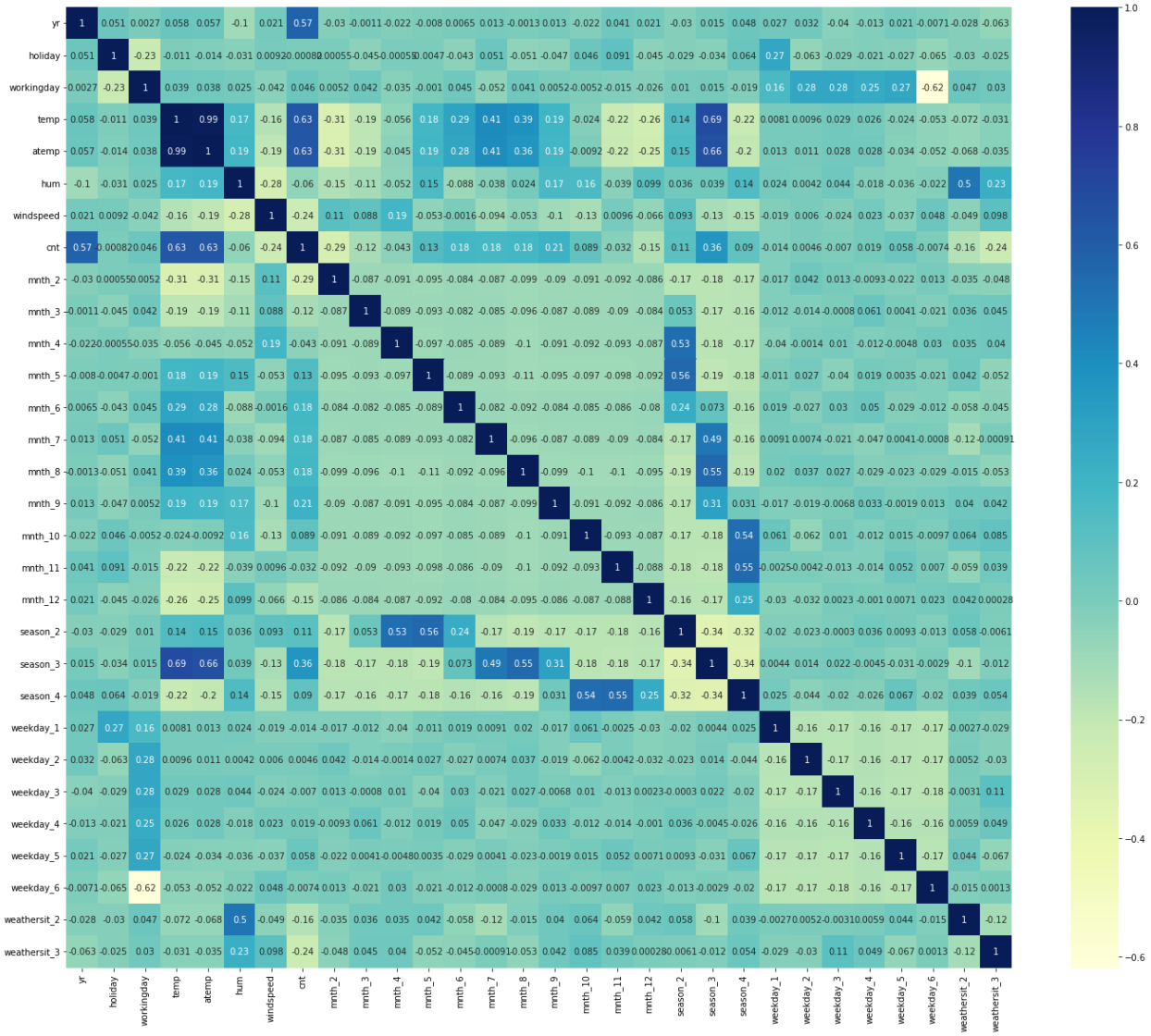
N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Let's see below Figure where I have calculated the correlation matrix and then plotted it with help of Heatmap for Bike sharing problem.



We can the highest correlation (0.99) is between **temp** and **atemp** variable. We have negative correlation as well. **workingday_6** has high negative correlation with **workingday**.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

ANSWER:

Feature Scaling: When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

Example: If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

So, we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

2. **MinMax Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

You can refer below figure to get a view about the data which needs scaling:

	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	0
3	Spain	27	48000	1
4	Germany	30	54000	0
5	Spain	38	61000	0
6	Germany	40	1000	1
7	France	35	58000	1
8	Spain	78	52000	0
9	France	48	79000	1
10	Germany	50	83000	0
11	France	37	67000	1

When I applied Minmax Scaling on the Bike sharing numerical features, then below is the snippet outcome of data table. All the columns under the red box are scaled. Their value is between 0 and 1.

```
# Apply scaler() to all the columns except the '0-1' and 'dummy' variables
num_vars = ['temp', 'atemp', 'hum', 'windspeed','cnt']

df_train[num_vars] = scaler.fit_transform(df_train[num_vars])
df_train.head()
```

	yr	holiday	workingday	temp	atemp	hum	windspeed	cnt	mnth_2	mnth_3	...	season_3
483	1	0	0	0.497426	0.487055	0.609956	0.194850	0.722734	0	0	...	0
650	1	0	0	0.416433	0.409971	0.513852	0.255118	0.815347	0	0	...	0
212	0	0	1	0.887856	0.819376	0.572294	0.276919	0.488265	0	0	...	1
714	1	0	0	0.378013	0.381804	0.871429	0.161523	0.433042	0	0	...	0
8	0	0	0	0.098690	0.048706	0.451083	0.700017	0.092039	0	0	...	0

5 rows × 30 columns

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANSWER:
If there is a perfect correlation between two independent variables, then VIF becomes infinity. The perfect correlation between two independent variables means R-Square = 1. VIF = 1/(1-R-Square).

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

When I was building a multi linear regression model using bike sharing dataset and I build a model with all the available columns of the train dataset, then I got a model where multiple feature where having VIF = infinity

See the below screenshot for reference:

In [141]:
```
# Check for the VIF values of the feature variables.
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

In [142]:
```
# Create a dataframe that will contain the names of all the feature variables and their respective VIFs
def get_vif(dataset):
    vif = pd.DataFrame()
    vif['Features'] = dataset.columns
    vif['VIF'] = [variance_inflation_factor(dataset.values, i) for i in range(dataset.shape[1])]
    vif['VIF'] = round(vif['VIF'], 2)
    vif = vif.sort_values(by = "VIF", ascending = False)
    return vif
```

In [143]:
```
vif = get_vif(X_train)
print(vif)
```

```
        Features     VIF
16       mnth_10     inf
17       mnth_11     inf
31  weathersit_2     inf
30  weathersit_1     inf
29     weekday_6     inf
28     weekday_5     inf
27     weekday_4     inf
26     weekday_3     inf
25     weekday_2     inf
24     weekday_1     inf
23     weekday_0     inf
22      season_4     inf
```

As assumption of linear regression, we should not have multicollinearity in the model. We should start dropping columns which are highly correlated and having high value of VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWER:

Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

How Q-Q plots can help us identify the distribution types?

The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In case of linear Regression Q-Q plot can help us doing the residual analysis of the error terms. One of the assumption of the linear regression is the error terms are normally distributed. We can do such analysis using Q-Q plot. Let's take example of the Bike Sharing model I have built. Now I have to do normal distribution check of the error terms. Below The sample screen shot from my analysis:

We can see that the error terms are terms are almost linear post the value of the -2. This indicates that the error terms are normally distributed.

In [1064]:
```
import scipy.stats as stats
sm.qqplot((y_train - y_train_pred),line='45',fit=True,dist=stats.norm)
```

Out[1064]: