General Subjective Questions & Answers

Prepared by:
Mr. Akshay Bugade

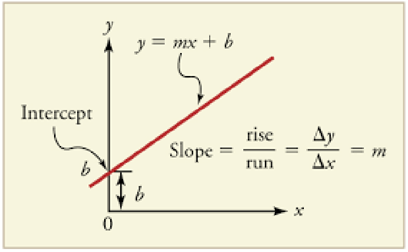
Date: 05/01/2022

1. Explain the linear regression algorithm in detail.

- Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

- Linear regression is based on the popular equation **"y = mx + b".**

  Y= Dependent/Target Variable, X= Independent Variable
  M= Slope or Coefficient of line
  B= Intercept


y
$y = mx + b$
Intercept
b
Slope $= \frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x} = m$
b
0
x

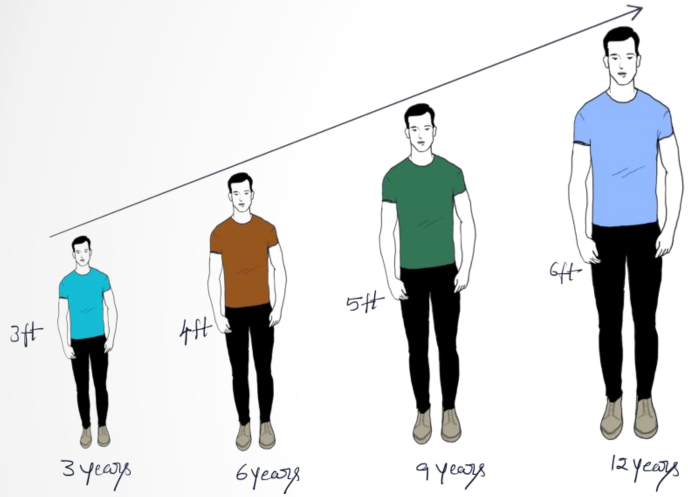
- It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

- Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

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

Regression is broadly divided into simple linear regression and multiple linear regression.

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



In **Linear Regression** we try to fit this straight line to the training data to predict a dependent variable (*y*) which has a linear (or) close to linear relationship with an independent variable (*x*).

In the above picture, consider the height of the person as the dependent variable (*y*) and the age of the person as independent variable (*x*).

Let's consider two points are (3 years, 3 feet) and (9 years, 5 feet). We've studied this problem in high school mathematics. Given two points on a straight line *(x1, y1)* and *(x2,y2)*,

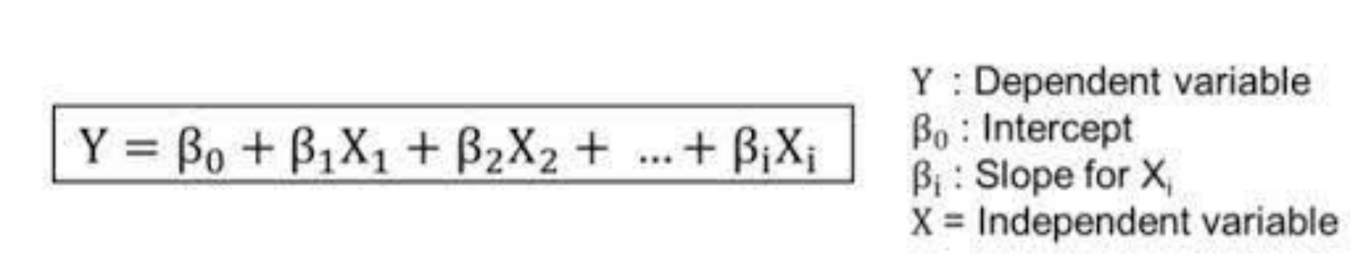
$$\text{Slope of the line, } m = \frac{y2 - y1}{x2 - x1} = \frac{5-3}{9-3} = \frac{2}{6} = \frac{1}{3}$$

$$\text{and intercept, } \quad 3 = 3(\frac{1}{3}) + c \Rightarrow c = 2$$

So when we train a linear regression model on the above Age v/s Height dataset, the model will have values *m = 1/3* and *c = 2* as model weights. Based on these values the prediction happens for the unknown data.

For example we want know how tall a person would measure when he is **7 years of age?** Pretty much simple! Substitute the values in equation *y = mx + c*. **Which gives the answer as 4.33 feet.**

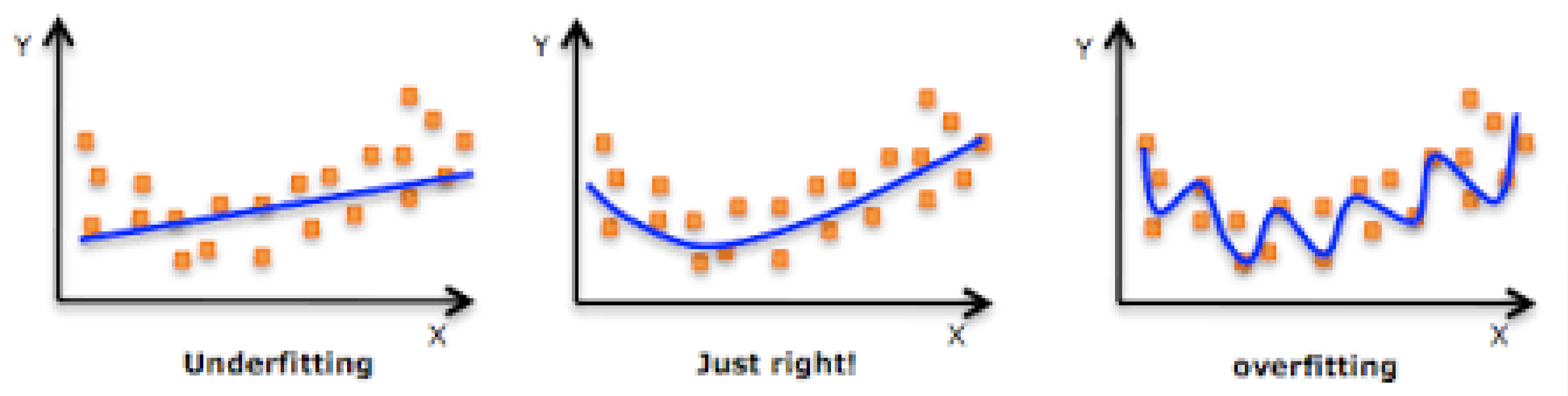
- **Multiple Linear Regression :MLR i**s used when the dependent variable is predicted using multiple independent variables. Which form a hyper plane instead of Straight Line



$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i}$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

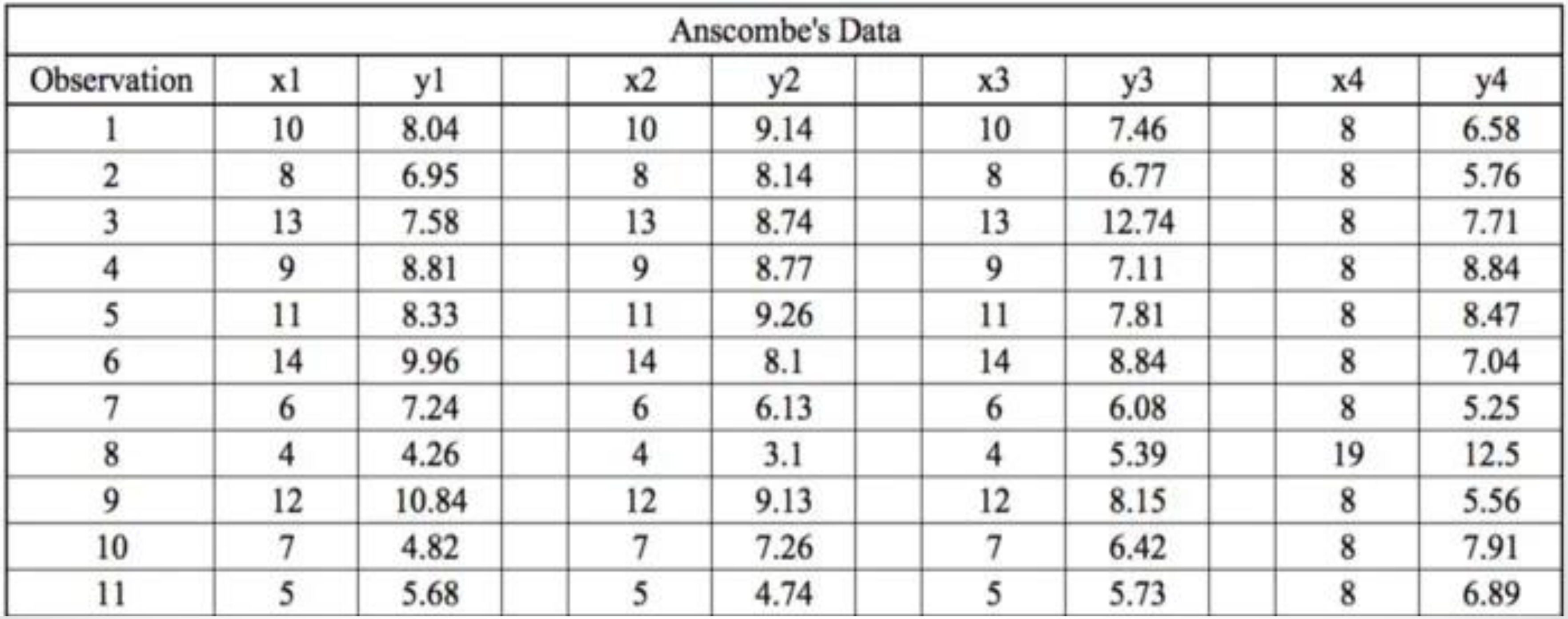
- **Regulation Method (Ridge,Lasso,elastic net etc.)** to overcome with Under fit & Over fit of Model, to make a Generalized Model, which give better accuracy when we implement on unknown or future data.



2. **Explain Anscombe's quartet in detail.**

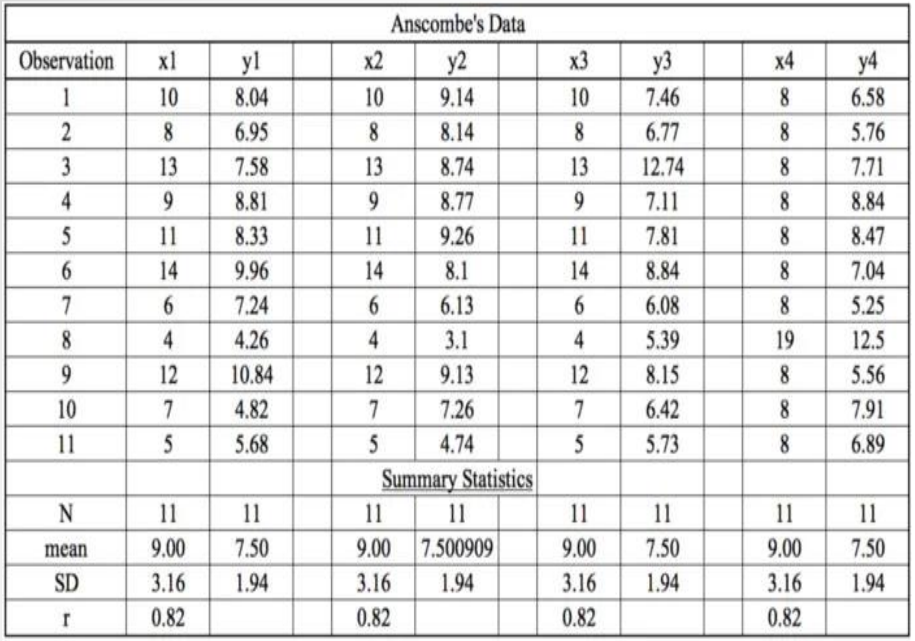
Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Lets Work on below data:

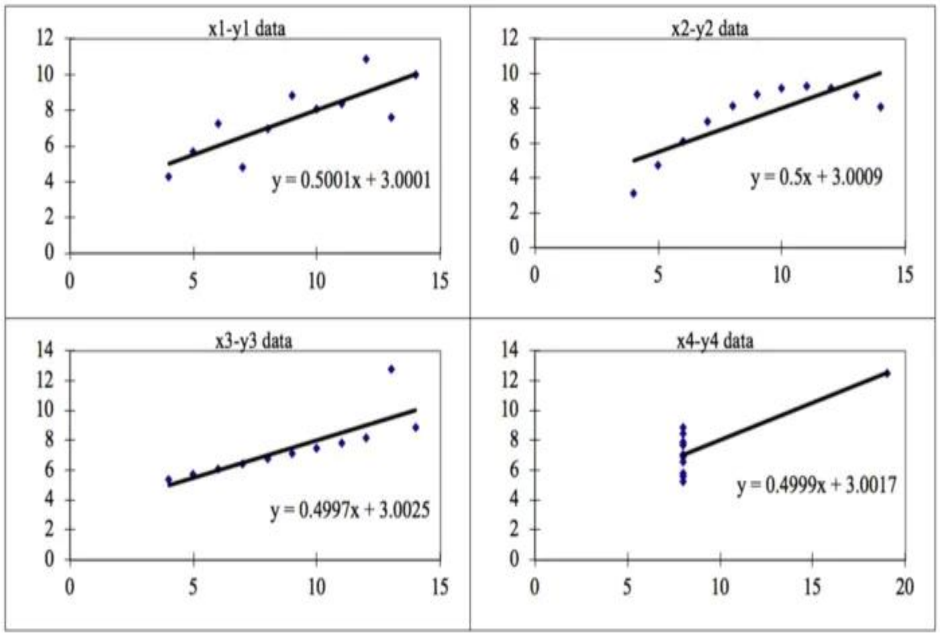


Anscombe's Data												
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



Observations:

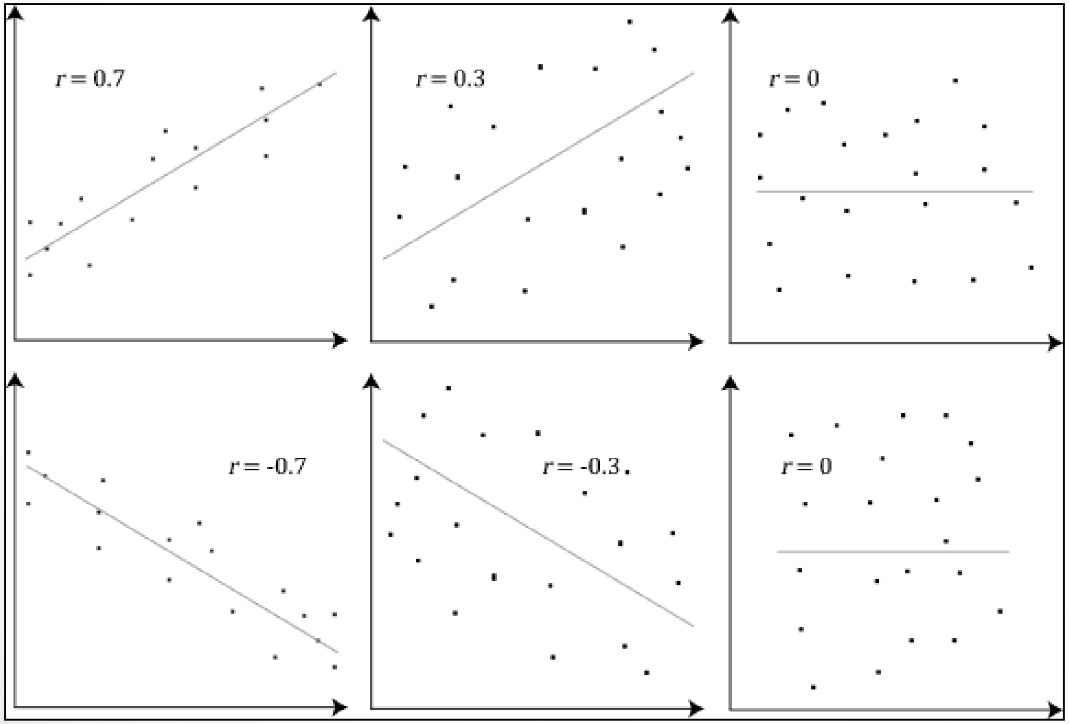
- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Conclusion:

- We have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson’s R?

- 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?
- 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



Variable Distribution & Pearson’s Correlation (r)

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

Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

1) Normalization -

- 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 Neighbors and Neural Networks.
- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as **Min-Max scaling.**
- Formula: $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2) Standardization-

- 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.
- Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.
- Formula: $$X' = \frac{X - \mu}{\sigma}$$

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

VIF - the variance inflation factor

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

- VIF = $1/(1-R^2)$

- If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity(i.e $R^2 = 1$)

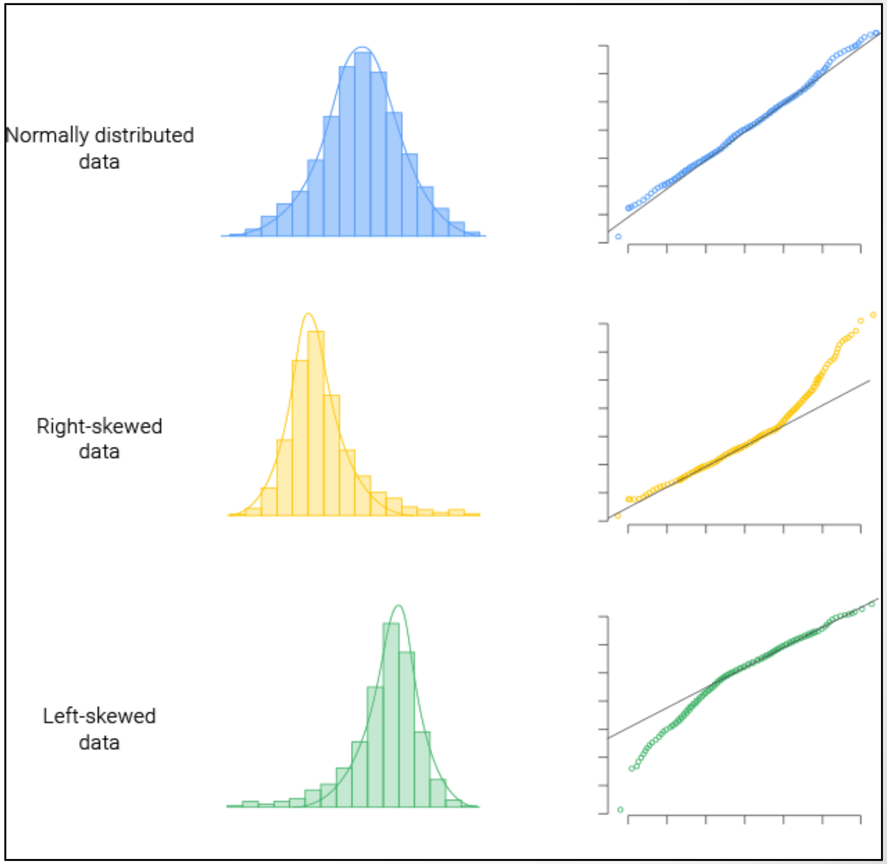
- A large value of VIF indicates that there is a correlation between the variables.

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

- The purpose of Q -Q plots is to find out if two sets of data come from the same distribution.
- A 45 degree angle is plotted on the Q -Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- This particular type of Q -Q plot is called a normal quantile-quantile (Q-Q) plot.

Q-Q 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 Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.
- If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.



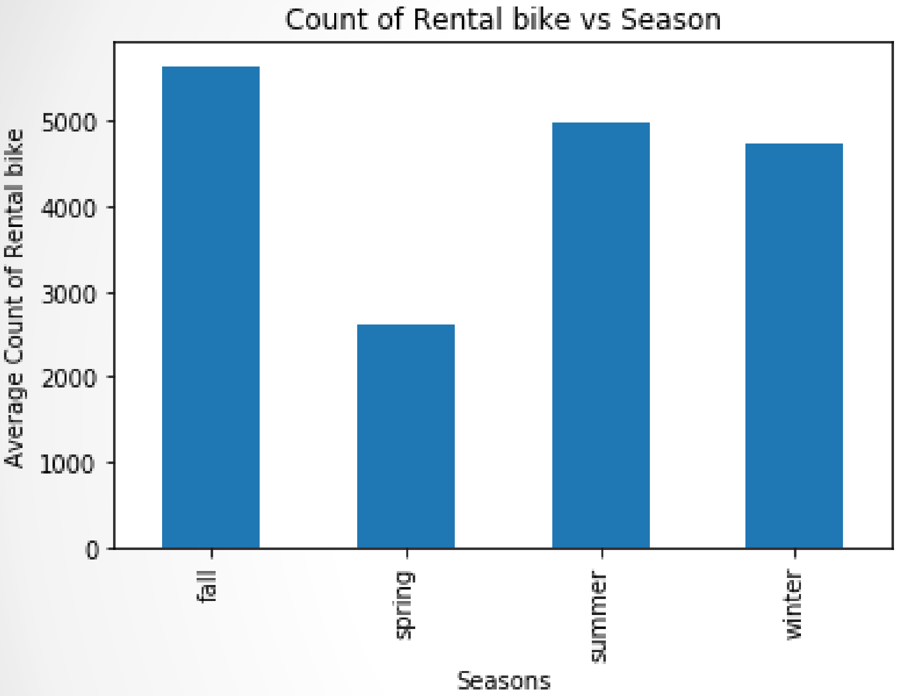
Q-Q Plot Vs. Distribution

Assignment-based Subjective Questions & Answers

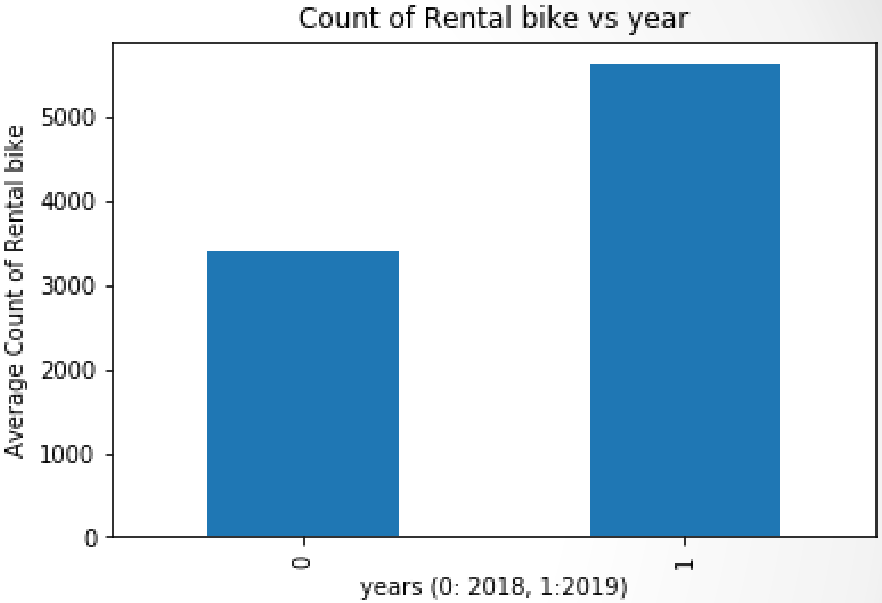
Prepared by:
Mr. Akshay Bugade

Date: 05/01/2022

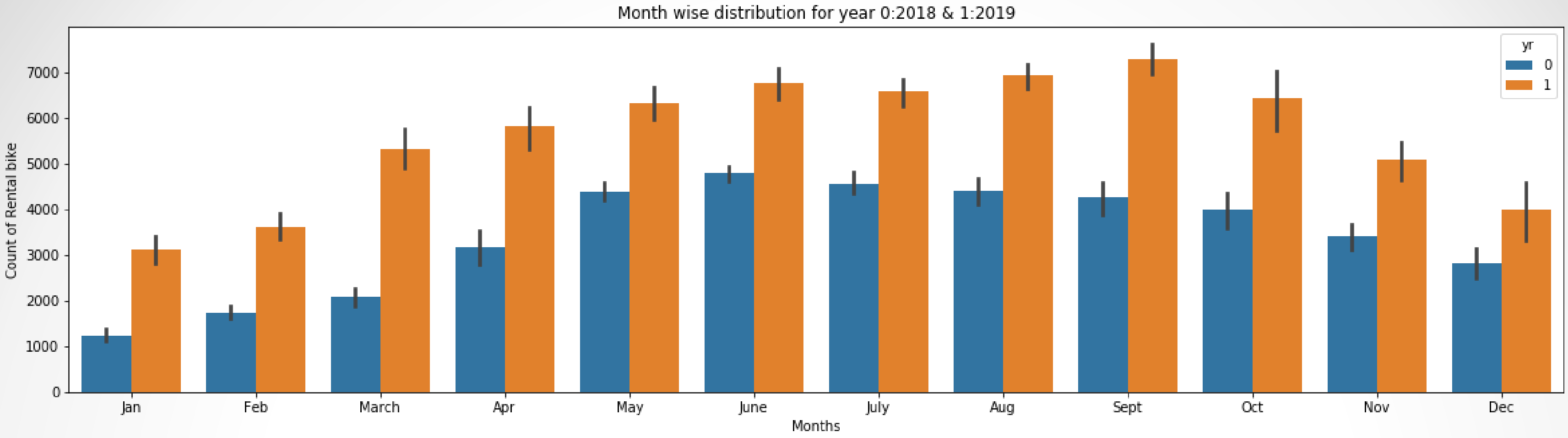
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



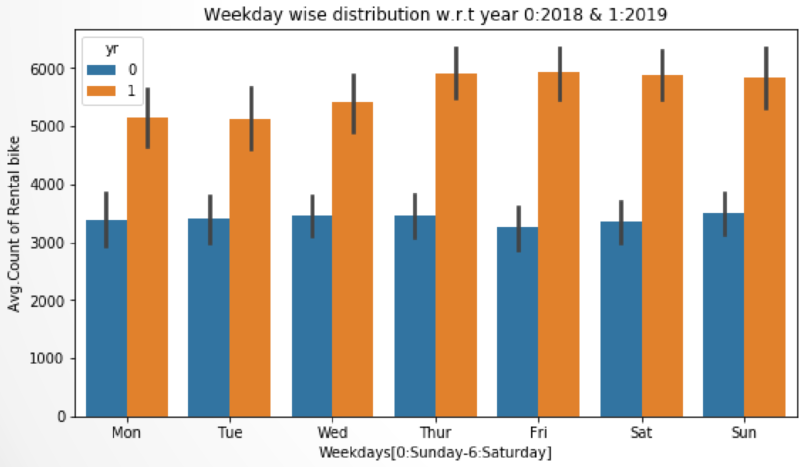
Avg.count of total rental bikes in Fall season is more & less in Spring

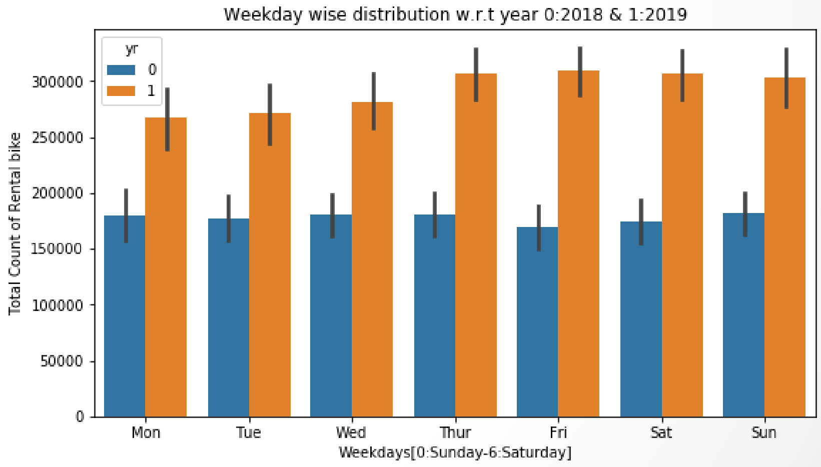


Avg.count of total rental bikes in year:2019 is more & less in year:2019

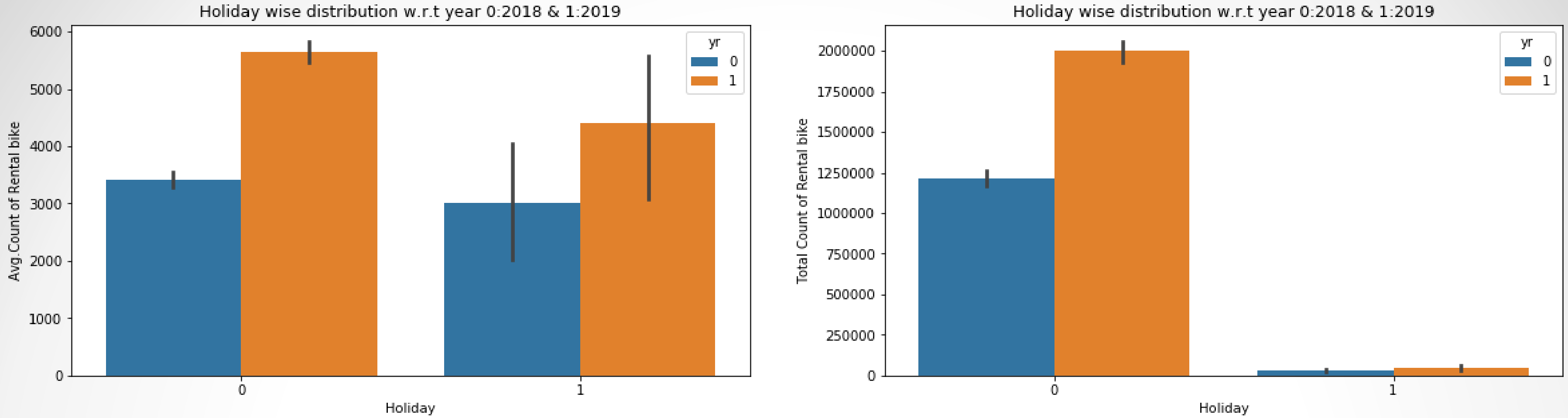


Count of total rental bikes in year:2019 is more compare to year:2018 for each months

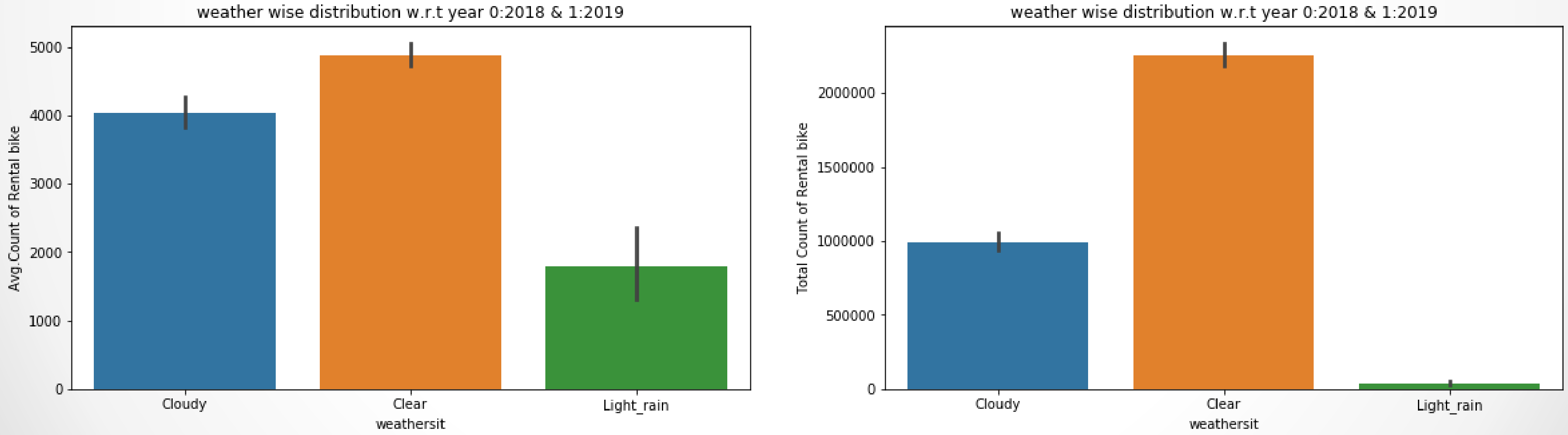




From Above Graphs we can see there are no such significant difference for Rental count in year 2018 for all days but there is slight increase in Rental count on Thursday to Sunday than other days



Avg.counts of rental bike doesn't show any major difference for holiday & non holiday but if we can see from second graph Total count of Rental bike is high in non holiday as compare to holiday w.r.t Year



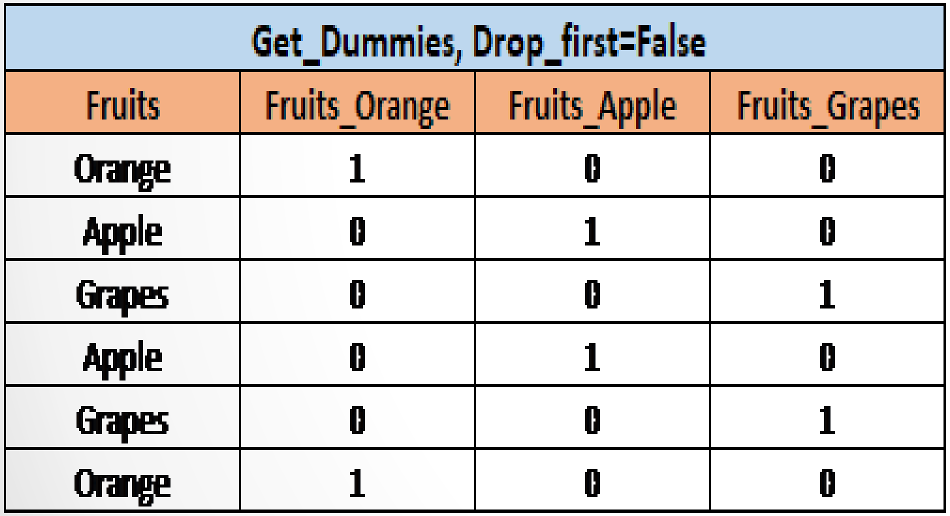
Total & Average count of Rental Bikes in clear weather sit is high & less in Light rain, Scattered Cloudy days

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

- Dummy Variable is use to create new columns of different distinct categories in Binary classification i.e. (0 or 1) & drop_first is use to reduce the columns, for better explanation please see below images of drop_first = False Vs. drop_first =True

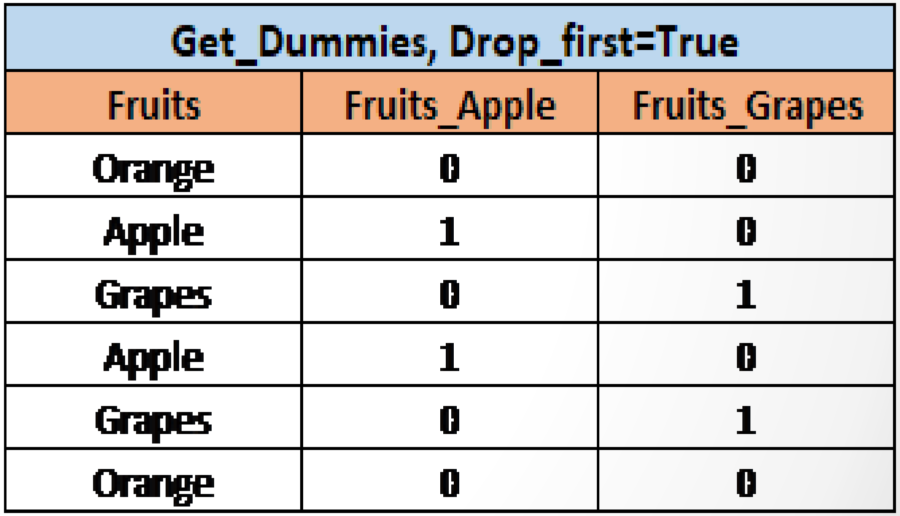
Example for better understanding:

Table 1



Get_Dummies, Drop_first=False			
Fruits	Fruits_Orange	Fruits_Apple	Fruits_Grapes
Orange	1	0	0
Apple	0	1	0
Grapes	0	0	1
Apple	0	1	0
Grapes	0	0	1
Orange	1	0	0

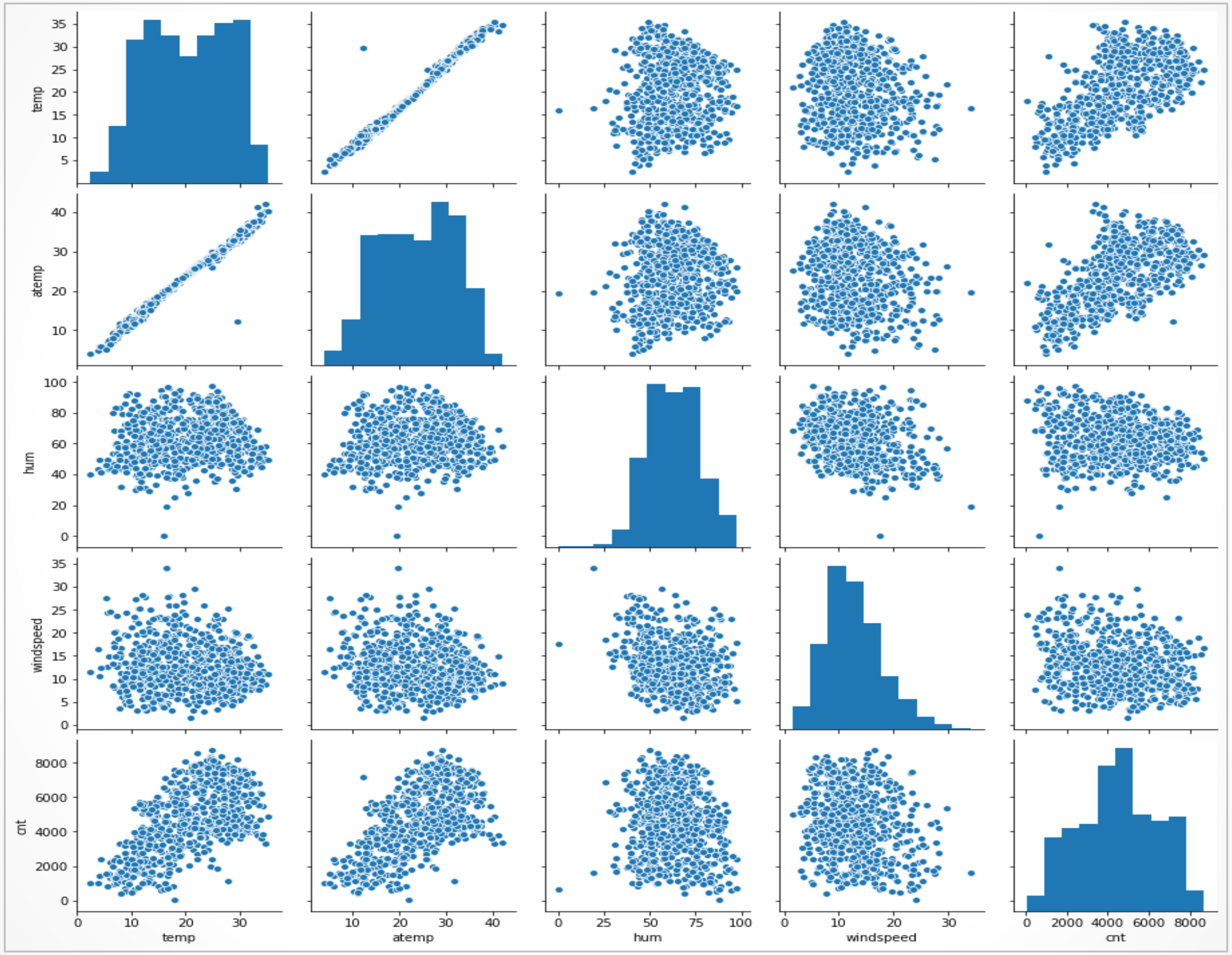
Table 2



Get_Dummies, Drop_first=True		
Fruits	Fruits_Apple	Fruits_Grapes
Orange	0	0
Apple	1	0
Grapes	0	1
Apple	1	0
Grapes	0	1
Orange	0	0

In Table 2 : if Fruits_Apple = 0 & Fruits_Grapes = 0 it will consider as Fruits_Orange =1

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



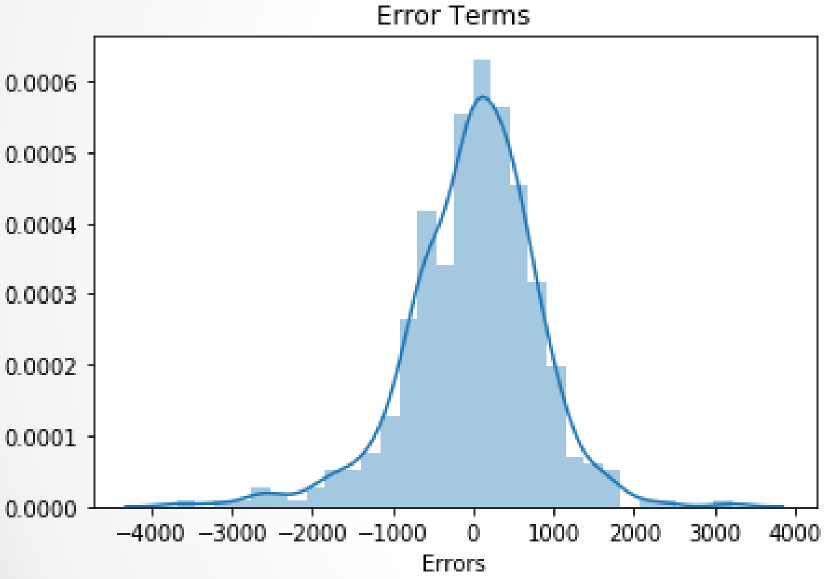
As we can see ‘atemp’ & ‘temp’ has some what positive correlation with Target Variable 'cnt' & 'atemp' and 'temp' has positive correlation between them

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

1) There is Linear Relationship between Independent & Target Variable (R^2 should be high)

2) Residual terms are normally distributed with mean = 0 (Ref.Graph-1)

3) No or little Multicollinearity (i.e VIF < 5)



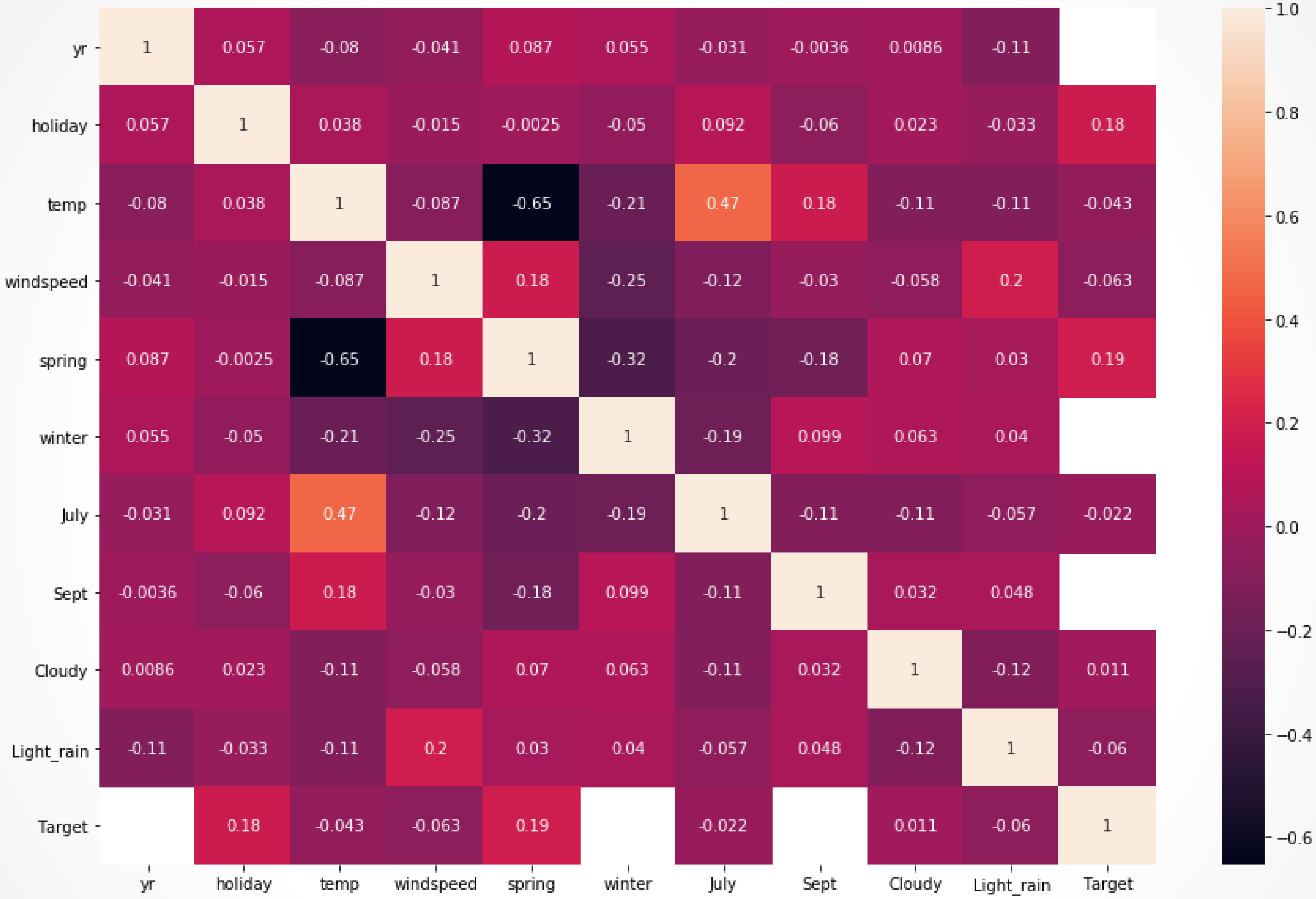
Graph-1

Models	No.of Features	R^2	p-Value	VIF
Model 1	15	0.844	p > 0	VIF > 30
Model 2	14	0.840	p > 0	VIF > 5
Model 3	13	0.838	p > 0	VIF < 5
Model 4	12	0.836	p > 0	VIF < 5
Model 5	11	0.834	p > 0	VIF < 5
Model 6	10	0.833	p = 0	VIF < 5

Final Model:

Features- 10 Nos
p-values=0
VIF < 5
R^2 on Train dataset=0.833
R^2 on Test Dataset = 0.807

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



Year, Season-**Winter** & Month-**Sept** has higher contribution on Target variable

Thank You