Bike Sharing Assignment – Subjective Questions

Prashant Ramdas Patil

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 temperature, past year demands, weather of the day and season are prime factors driving demand of shared bikes. It has been observed between the range of 25 and 35 degrees, the demand is on higher side while spring season and weather with light rain /snow has negative impact. The demand increased on year on year basis for the years 2018 and 2019 indicating some form of law of attraction.

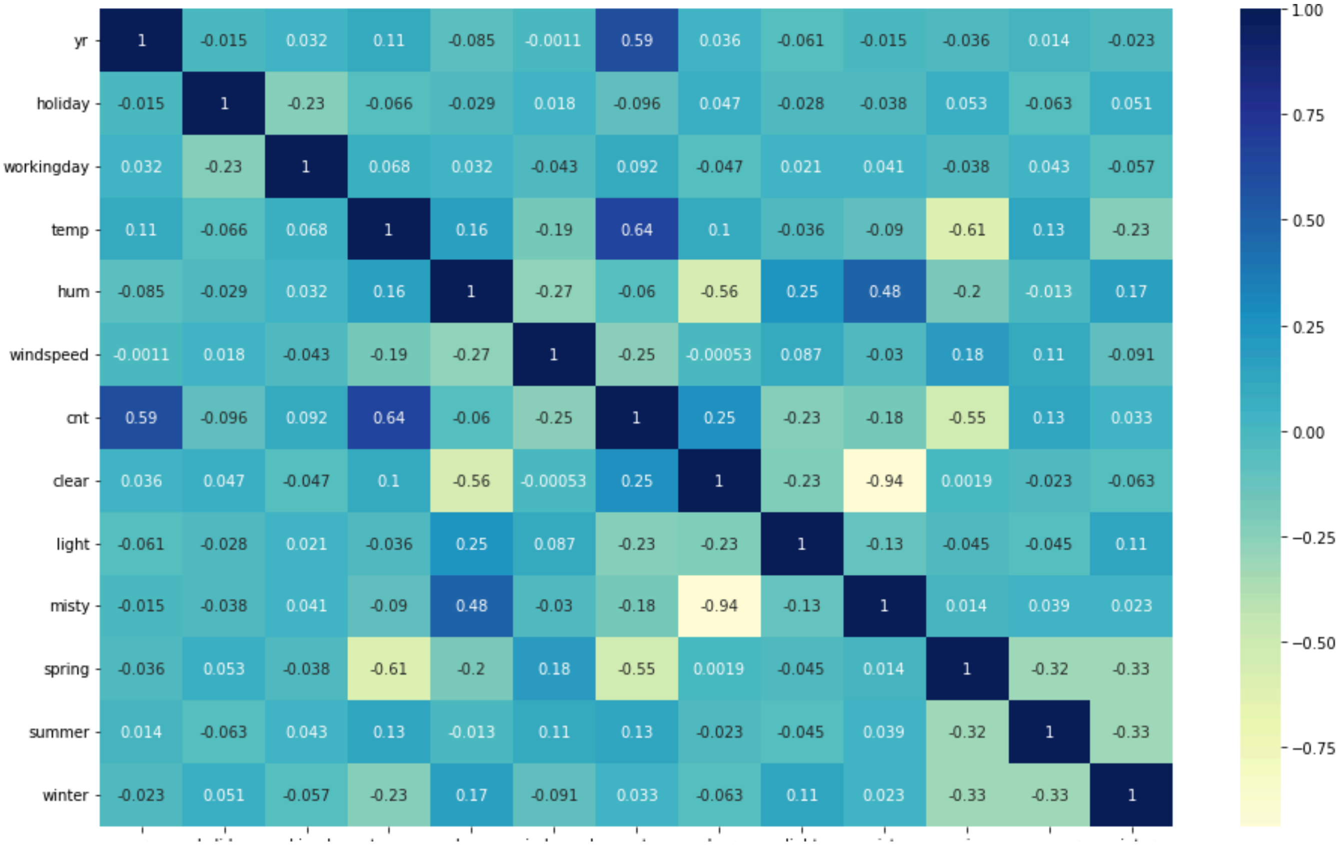
Overall the higher marketing spend can be targeted for Summer and Fall season that enables demand of shared bikes.

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

Typically in linear regression, we want to optimize number of independent variables as less as possible. This helps in assessing and ease of understanding on impact of these variables on the final out come. When we are assessing categorical variables into the linear regression, each category is converted to binary variables. However presence of one variable can be assessed due to absence of other defined categories and hence can be avoided. This also improves efficiency during model generation.

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

From the pair plot, temperature is having the highest coefficient.



yr	1	-0.015	0.032	0.11	-0.085	-0.0011	0.59	0.036	-0.061	-0.015	-0.036	0.014	-0.023
holiday	-0.015	1	-0.23	-0.066	-0.029	0.018	-0.096	0.047	-0.028	-0.038	0.053	-0.063	0.051
workingday	0.032	-0.23	1	0.068	0.032	-0.043	0.092	-0.047	0.021	0.041	-0.038	0.043	-0.057
temp	0.11	-0.066	0.068	1	0.16	-0.19	0.64	0.1	-0.036	-0.09	-0.61	0.13	-0.23
hum	-0.085	-0.029	0.032	0.16	1	-0.27	-0.06	-0.56	0.25	0.48	-0.2	-0.013	0.17
windspeed	-0.0011	0.018	-0.043	-0.19	-0.27	1	-0.25	-0.00053	0.087	-0.03	0.18	0.11	-0.091
cnt	0.59	-0.096	0.092	0.64	-0.06	-0.25	1	0.25	-0.23	-0.18	-0.55	0.13	0.033
clear	0.036	0.047	-0.047	0.1	-0.56	-0.00053	0.25	1	-0.23	-0.94	0.0019	-0.023	-0.063
light	-0.061	-0.028	0.021	-0.036	0.25	0.087	-0.23	-0.23	1	-0.13	-0.045	-0.045	0.11
misty	-0.015	-0.038	0.041	-0.09	0.48	-0.03	-0.18	-0.94	-0.13	1	0.014	0.039	0.023
spring	-0.036	0.053	-0.038	-0.61	-0.2	0.18	-0.55	0.0019	-0.045	0.014	1	-0.32	-0.33
summer	0.014	-0.063	0.043	0.13	-0.013	0.11	0.13	-0.023	-0.045	0.039	-0.32	1	-0.33
winter	-0.023	0.051	-0.057	-0.23	0.17	-0.091	0.033	-0.063	0.11	0.023	-0.33	-0.33	1

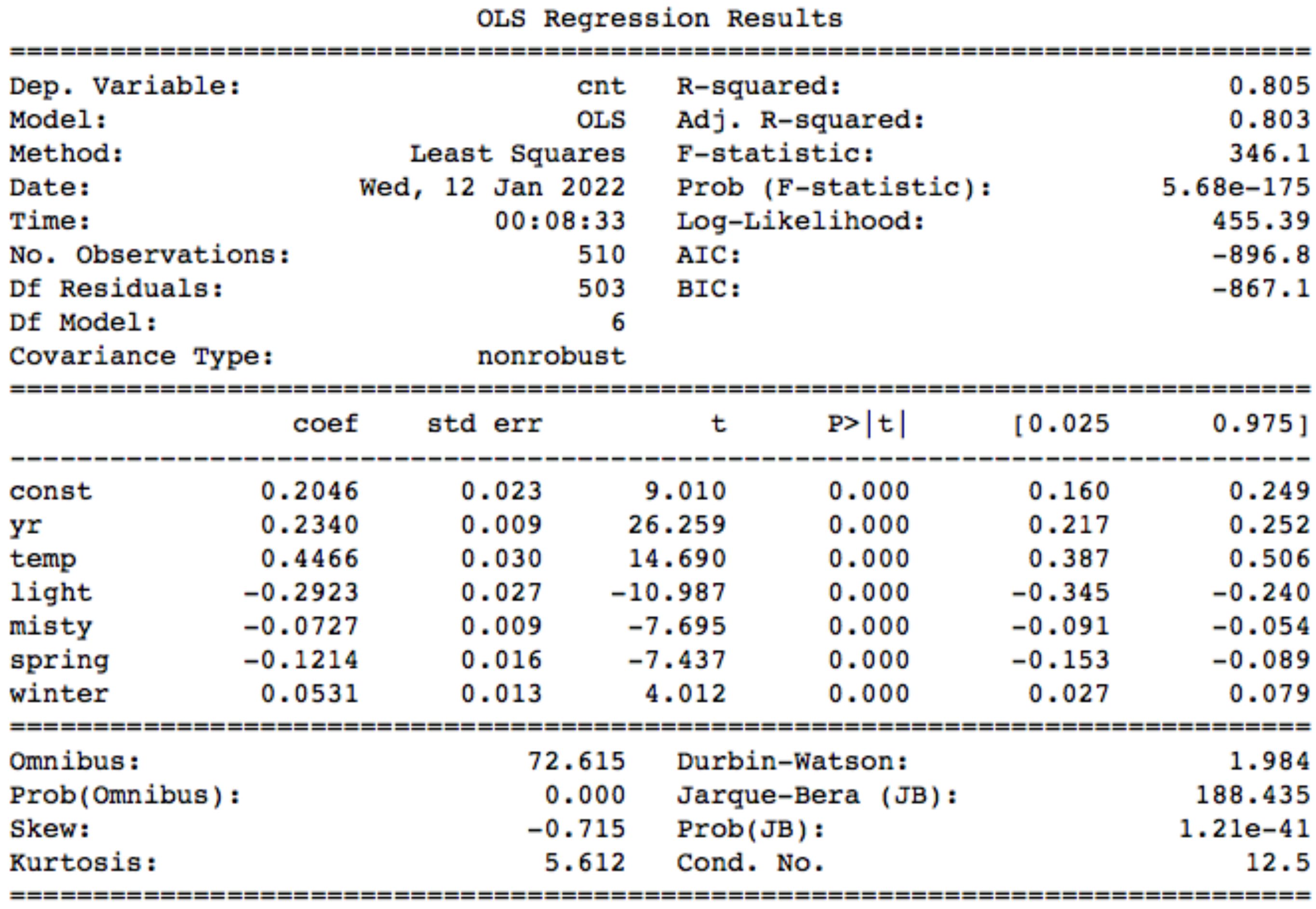
4. How did you validate the assumptions of Linear Regression after

building the model on the training set?

The first step is checking the value of R-squared where it should be as close to 1

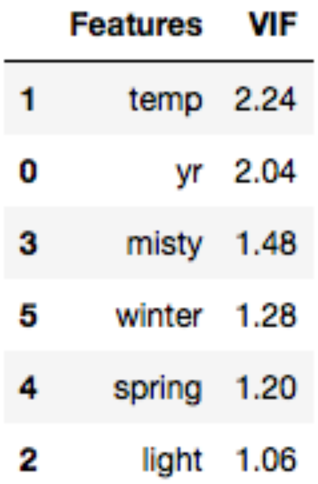
The second step is checking the probability of F-static which should be close to zero.

The third is checking the probability of t value of independent variables that should 0 or nearly 0.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     346.1
Date:                Wed, 12 Jan 2022   Prob (F-statistic):          5.68e-175
Time:                        00:08:33   Log-Likelihood:                 455.39
No. Observations:                 510   AIC:                            -896.8
Df Residuals:                     503   BIC:                            -867.1
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2046      0.023      9.010      0.000       0.160       0.249
yr             0.2340      0.009     26.259      0.000       0.217       0.252
temp           0.4466      0.030     14.690      0.000       0.387       0.506
light         -0.2923      0.027    -10.987      0.000      -0.345      -0.240
misty         -0.0727      0.009     -7.695      0.000      -0.091      -0.054
spring        -0.1214      0.016     -7.437      0.000      -0.153      -0.089
winter         0.0531      0.013      4.012      0.000       0.027       0.079
==============================================================================
Omnibus:                       72.615   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              188.435
Skew:                          -0.715   Prob(JB):                     1.21e-41
Kurtosis:                       5.612   Cond. No.                         12.5
==============================================================================
```

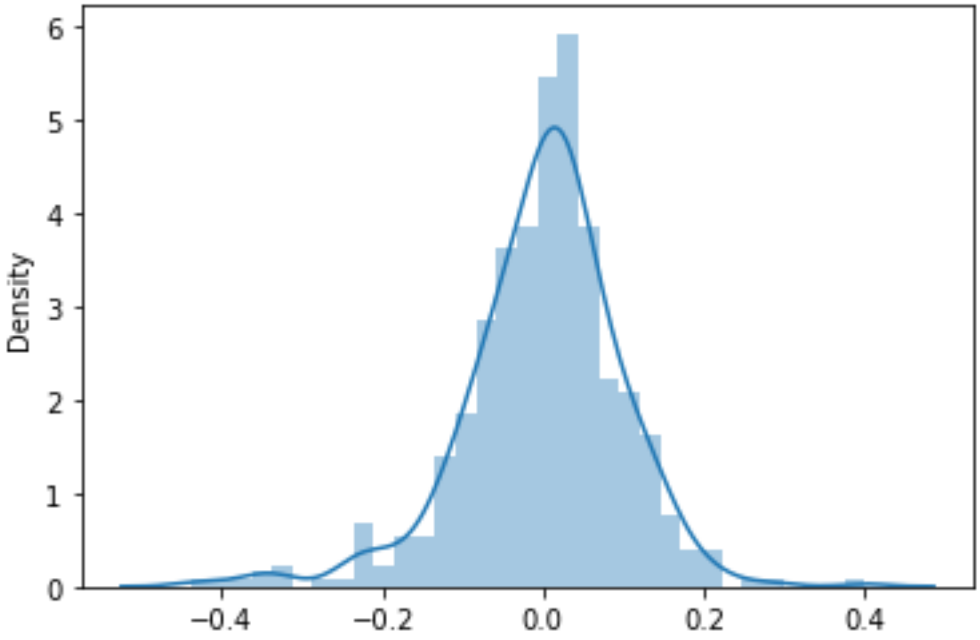
The fourth step is checking the VIF factor for multicollinearity where the VIF score is required to be less than 5.



	Features	VIF
1	temp	2.24
0	yr	2.04
3	misty	1.48
5	winter	1.28
4	spring	1.20
2	light	1.06

The fifth step is check the residual error and its distribution

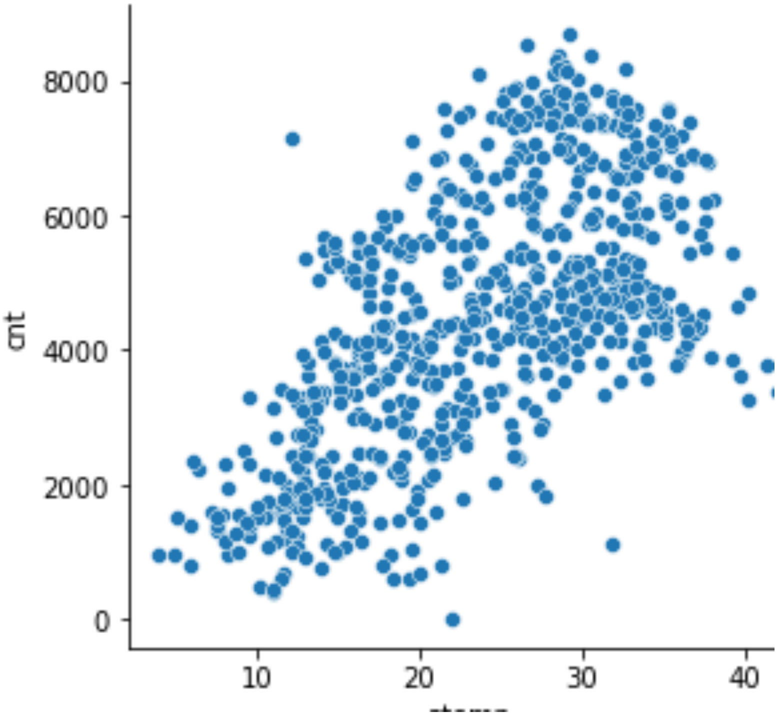
The residual error distribution should be normal distribution across mean 0.



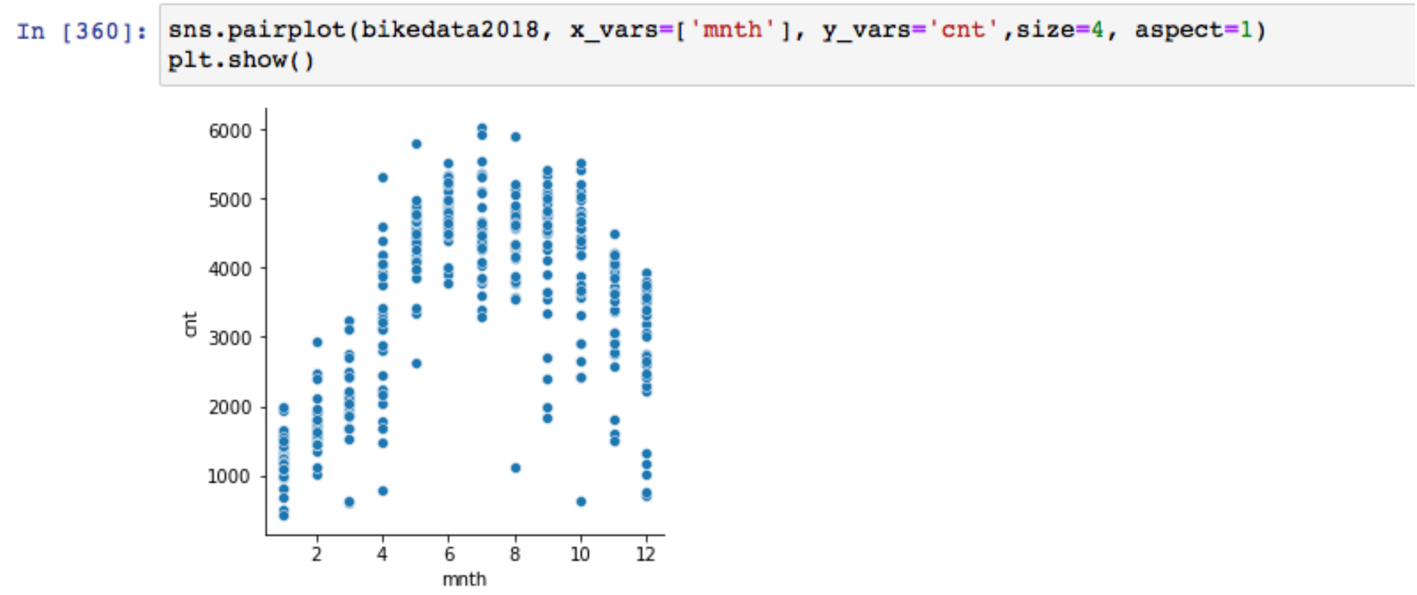
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The tope 3 features impacting the demand are –

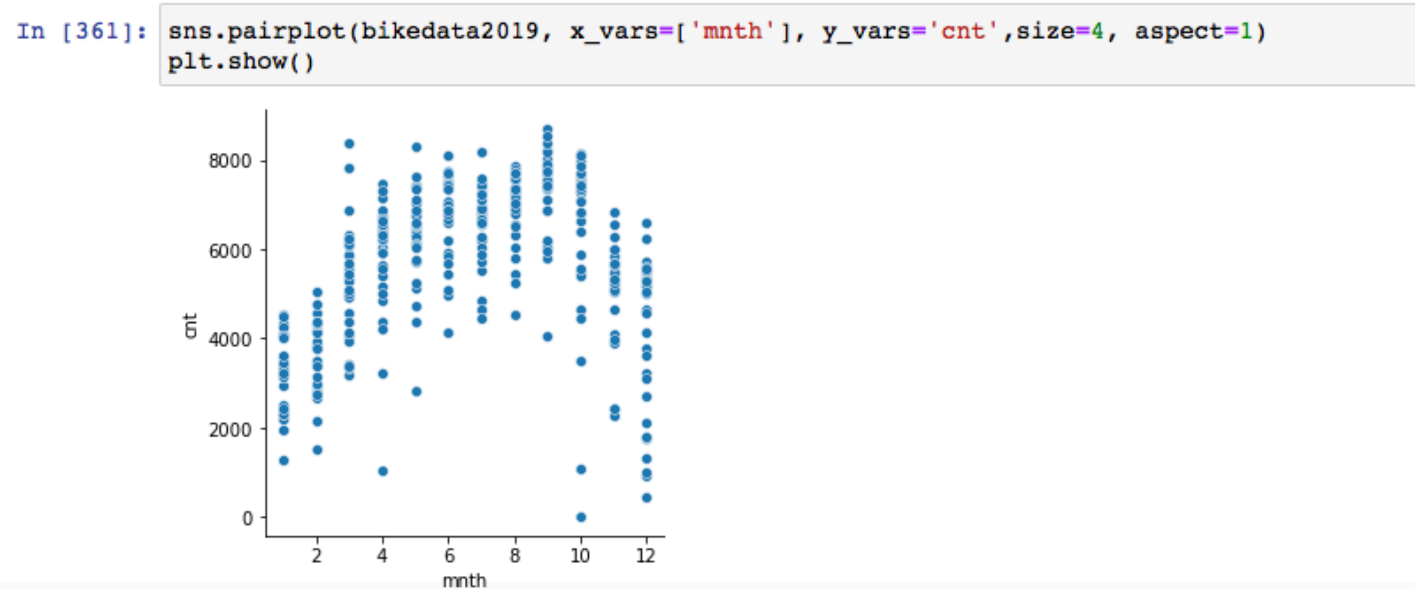
1. ***Temperature :We see a positive coefficient with demand of share bikes especially between range of 25-35Degrees***



2. ***Previous year demand - I think similar to law of attraction(not just the same), when more people rent bikes for fun /exercise/ or for other reasons – it may inspire otherts o do the same.***



```
In [360]: sns.pairplot(bikedata2018, x_vars=['mnth'], y_vars='cnt',size=4, aspect=1)
          plt.show()
```



```
In [361]: sns.pairplot(bikedata2019, x_vars=['mnth'], y_vars='cnt',size=4, aspect=1)
          plt.show()
```

3. ***The weather with light rain/snow impacts negatively. For obvious reason no one like riding being wet and especially on lower temperatures.***

General Subjective Questions

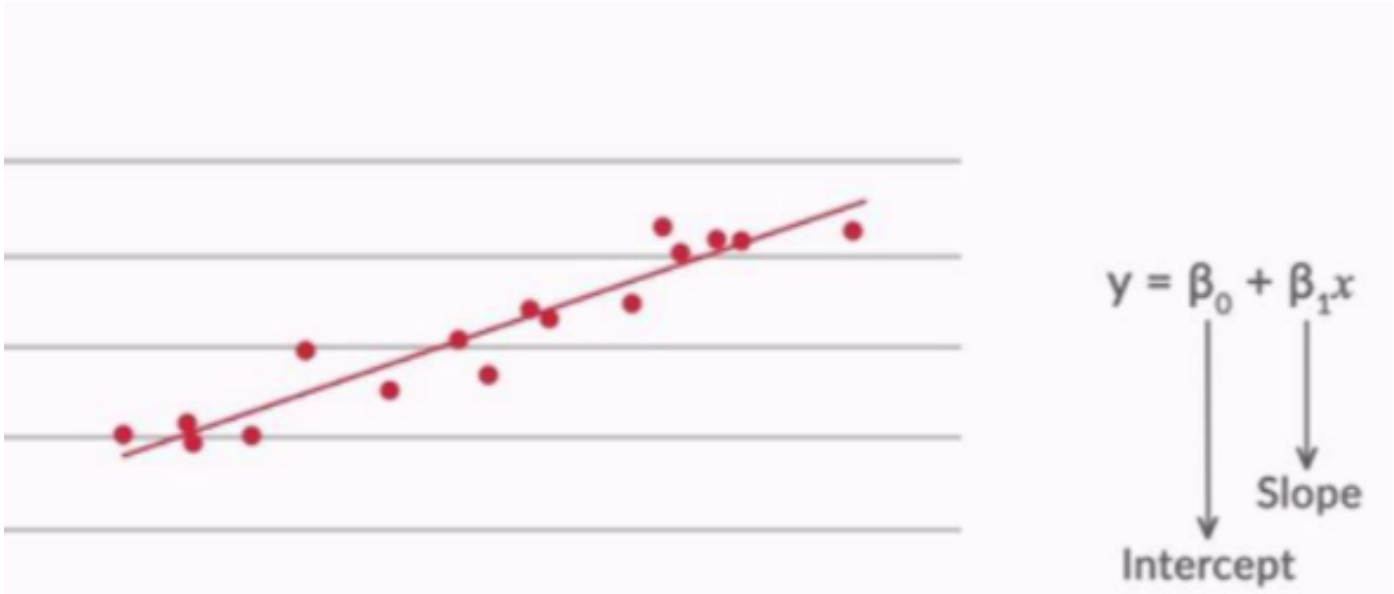
6. *Explain the linear regression algorithm in detail.*

The linear regression algorithm determines direct relation between dependent variables with one or more independent variables. For example it may amount of sales generated with respect to marketing spend.

This comes under the category of Supervised learning model and the output predicted is a continuous variable.

Linear regression can be further classified into two types –

a. ***Simple linear regression where output variable is dependent on single independent variable.***
Equation of the regression line is given by the following expression: $Y = \beta 0 + \beta 1X$



b. ***Multiple linear regression where the dependent variable is directly impacted by more than one independent variables.***

Equation of multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Important assumptions of linear regression algorithm ➔

- It is assumed that there is a linear relationship between the dependent and independent variables
- It is assumed that the error terms are normally distributed.
- It is assumed that the residuals have a mean value of zero
- It is assumed that the residual terms have the same variance
- It is assumed that the residual terms are independent of each other
- The independent variables are measured without error.
- The independent variables are linearly independent of each other

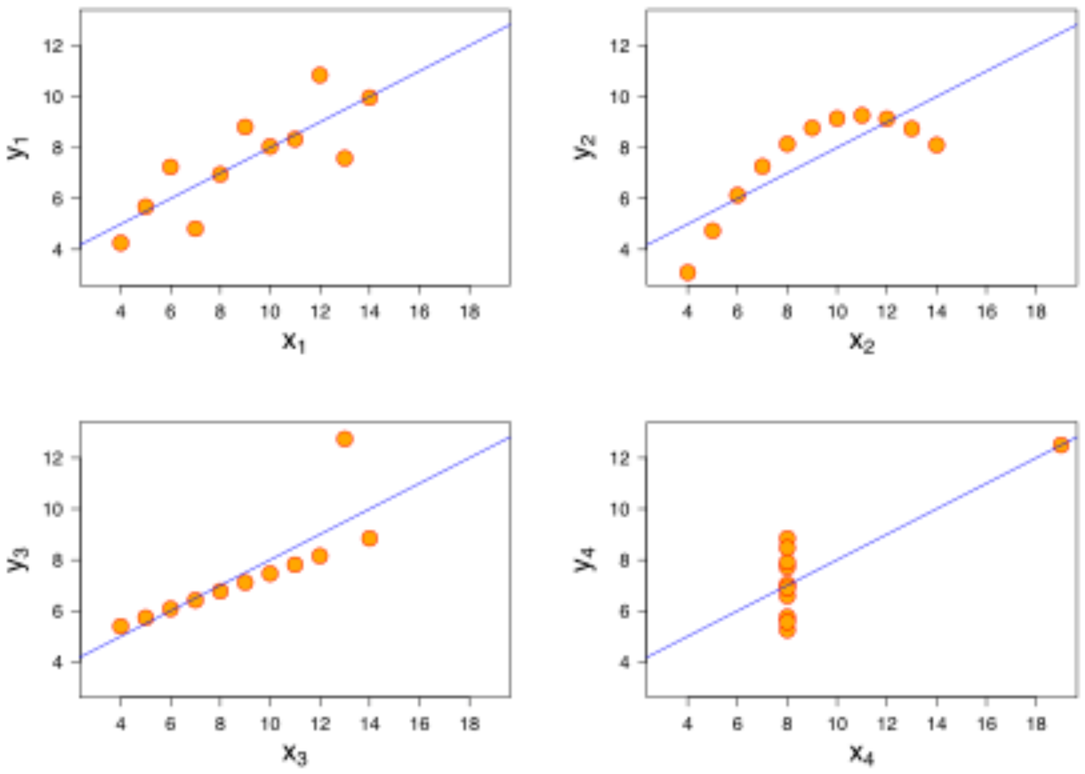
7. Explain the Anscombe’s quartet in detail.

The distribution of dependent variables with respect to independent cannot be judged based on statistical numbers such as mean, variance . It is essential to visualize the distribution especially while determining linearity between the variables.

Anscombe’s quartet, which shows such relation between 4 datasets having similar statistical numbers, highlights the important of visualizing such distribution.

In brief one show always rely on plotting the data rather than relying on summary statistics alone.

The first dataset shows linear equation while others are non linear.

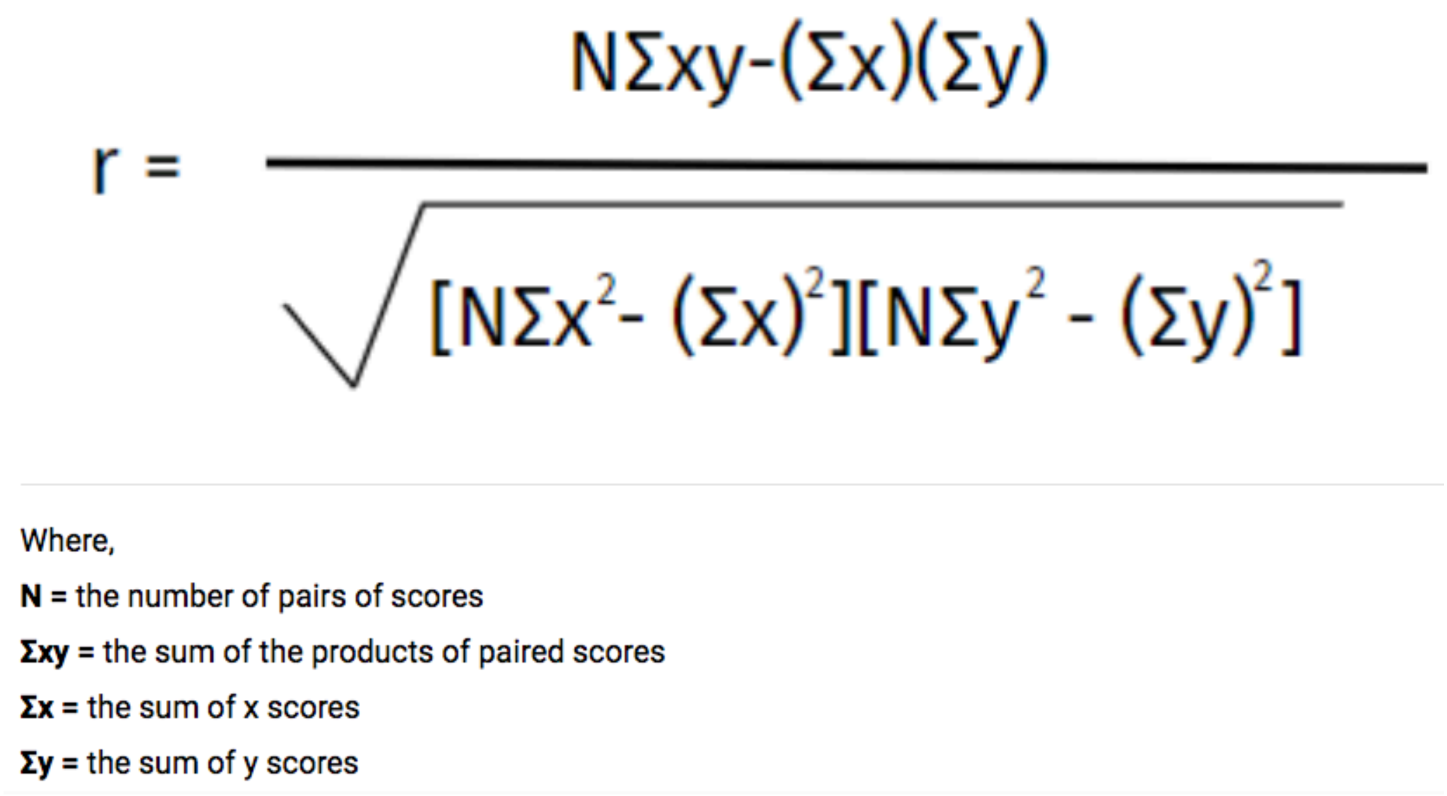


8. *What is Pearson’s R?*

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

This is used to calculate a linear relationship between the two given variables.

It is calculated using below formula



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

Where,

N = the number of pairs of scores

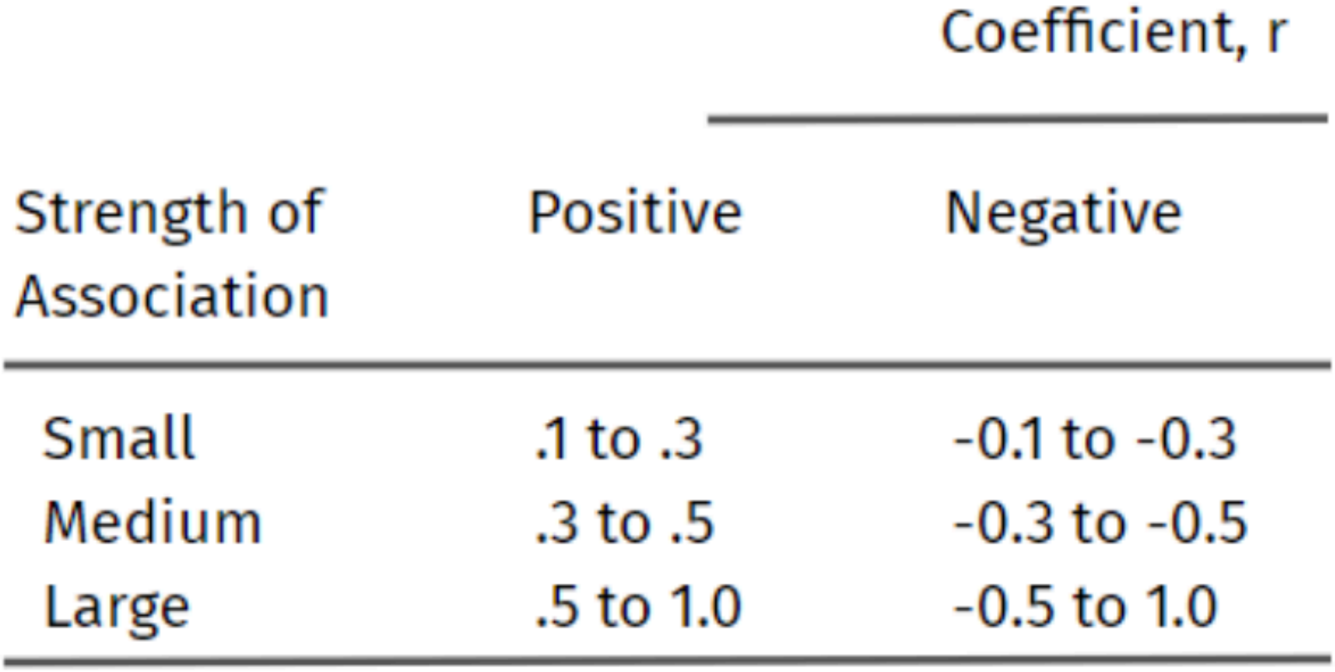
Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

Guidelines to interpret the Pearson coefficient correlation ➔



	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

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

While deriving model of multiple linear regressions the independent variables will be on different scale. So without scaling, scale of coefficients will of different magnitude and hence difficult to interpret the model. Hence scaling is applied on continuous variables while deriving linear regression model. It also helps in relative comparison of coefficients of all independent variables involved.

In standardized scaling , the variables are scared in such a way that their mean is zero and stand deviation is 1. Formual for deriving this scale is

X = x-mean(x) / std(x)

In Normalized scaling model the variables are scaled in such a way that their value lies between zero and one. It is derived from the formula –

X = x =min(x) / max(x) – min(x)

This model is often used especially when we don't know about the distribution.

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

Variance Inflation Factor determines relationship of one independent variable with other independent variables.

This is calculated as VIF = 1 /1- Ri^2

Now if the relation between concerned independent variable with respect to other independent variables is highly linear, it means Ri^2 will be close to one making VIF score as infinitive.

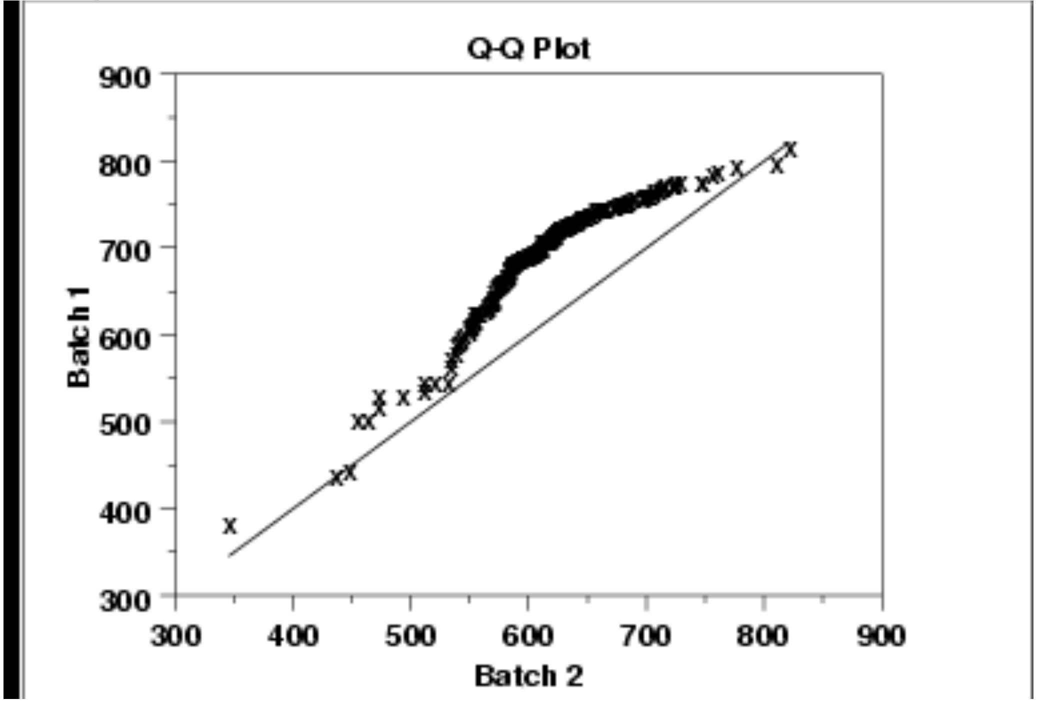
11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are used to find the type of distribution for a random variable
It is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

The q-q plot is formed by:
Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.



The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.