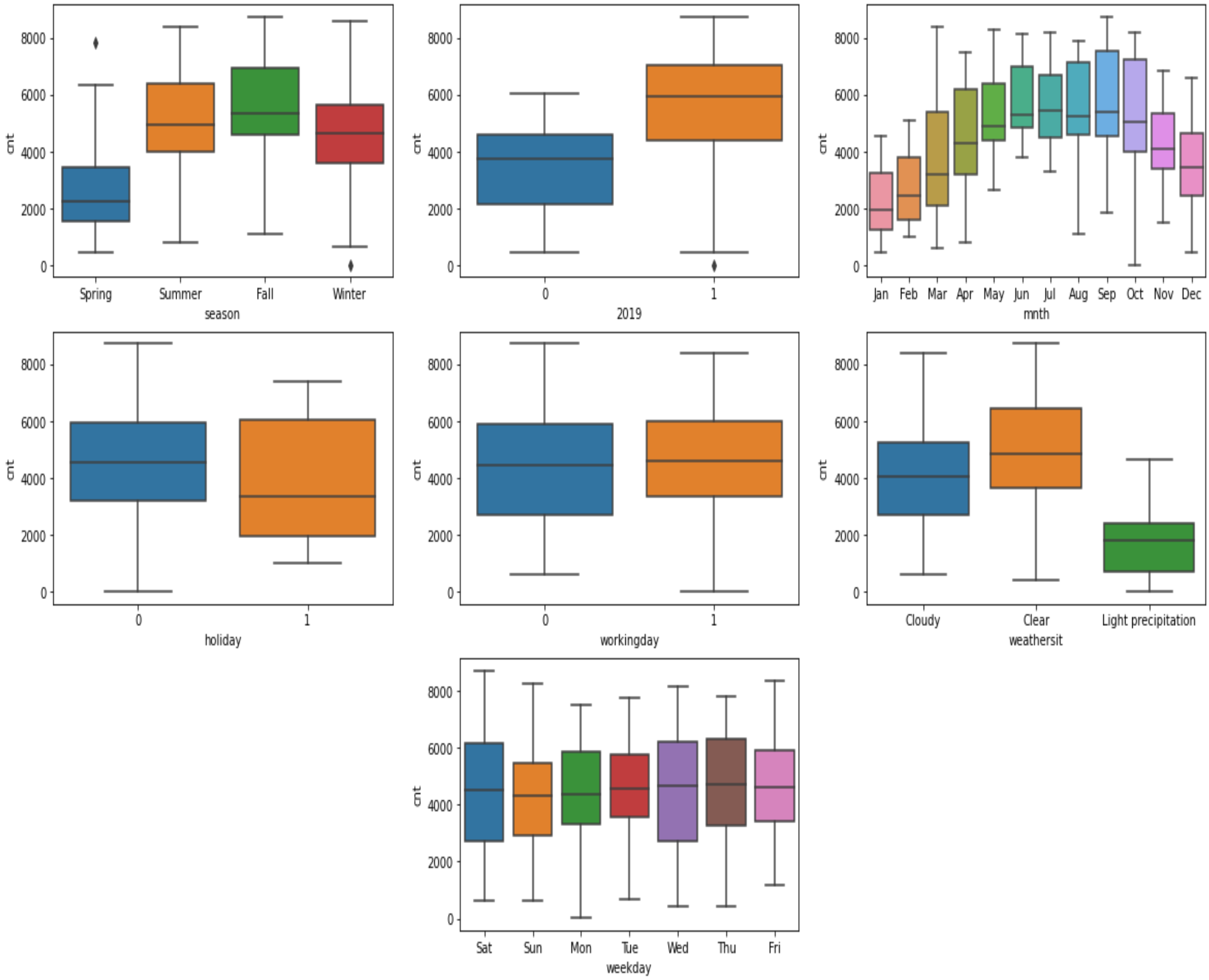
Assignment-based Subjective Questions

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

Ans. Categorical variables have significant effects on the dependent variables.



As could be seen from the set of box plots, ‘workingday’ and ‘workday’ do not have any significant difference in their medians and hence, do not play any important role. Therefore, instead of taking ‘workingday’, ‘holiday’ has beeen considered as a relevant feature in the model.

Whereas, ‘year’, ‘months’, ‘seasons’ have drastic effects on the target variable. In ‘weathersit’, it can be clearly seen that there is no rental during heavy precipitation, i.e., Heavy Rain, Ice Pallets, Thunderstorm, Mist, Snow, or Fog, Most of the rentals have occurred during clear weather condition followed by clody and light precipitation weather conditions.

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

Ans. During dummy variable creation, as many numbers of variables are there (value counts), that many columns generated.

The following table displays that the first row suggests red, second row suggests blue whereas the thir row suggests blue.

Red	Green	Blue
1	0	0
0	0	1
0	1	0

But in actual practice we can work out with one column less than the columns generated. If there are n number of variables, (n-1) numbers of columns are practically required.

Green	Blue
0	0
0	1
1	0

Now, the first row suggests that it is neither green nor blue, so it must be red given that there are only three colours available. Second row suggests blue and third row suggests green.

Hence it is important to use drop_first=True during dummy variable creation since it increases efficiency and lets us work with lesser number of features.

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

Ans. It is obvious for ‘registered’ to have the highest correlation but since ‘registered’ and ‘casual’ are directly involved in the target variable (‘casual’ + ‘registered’ = ‘cnt’), ‘temp’ should be considered rather to have the highest correlation with the target variable.

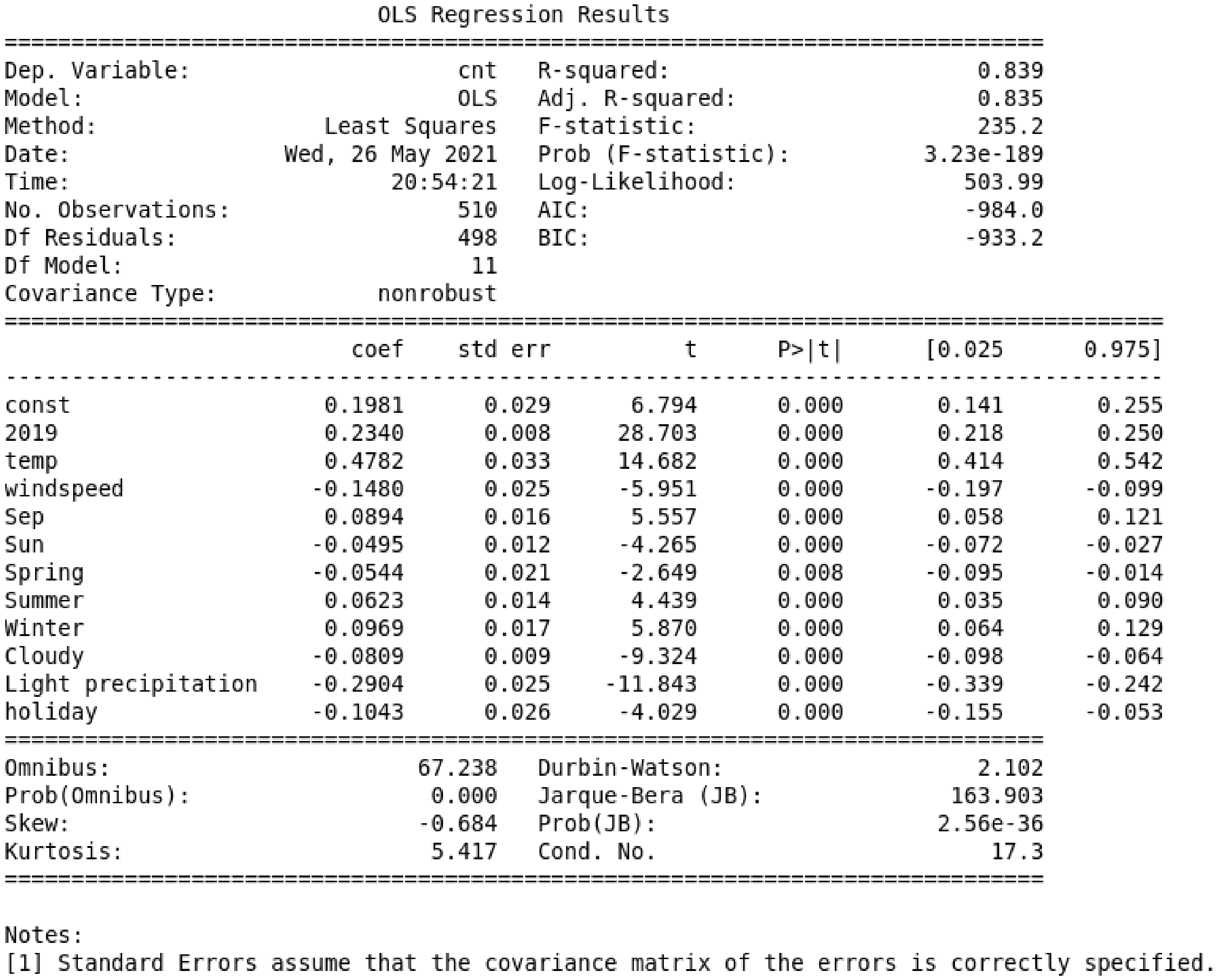
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. The assumptions are validated in the following ways :-

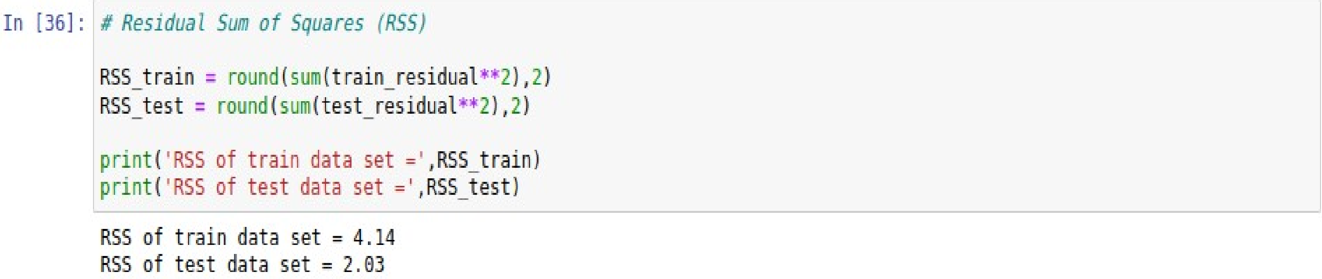
i. Linear relationship between X and Y : Linear Regression refers to a group of techniques for fitting and studying the straight-line relationship between two variables. Linear regression estimates the regression coefficients β_0 and β_1, β_2 ,………, β_p in the equation

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

where X_1, X_2 ,......., X_p are the independent variable, Y is the dependent variable, β_0 is the Y intercept, β_1, β_2 ,…, β_p are the slopes, and ε is the error.
All these factors is clearly visible in the following model summary and the Residual Sum of Errors calculated (provided below respectively) :

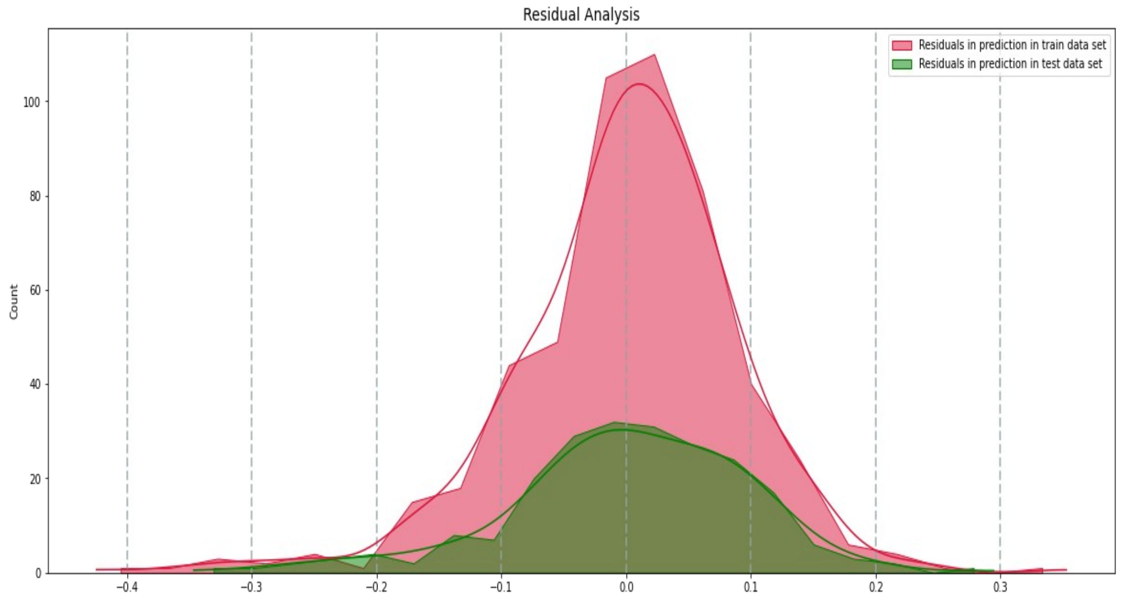


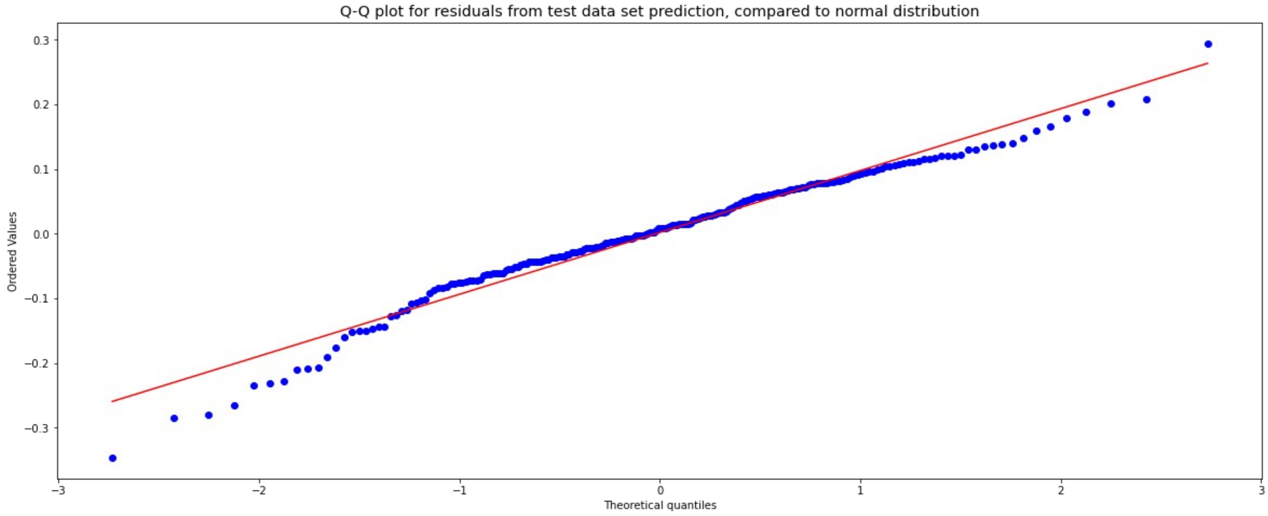
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.839
Model:                            OLS   Adj. R-squared:                  0.835
Method:                 Least Squares   F-statistic:                     235.2
Date:                Wed, 26 May 2021   Prob (F-statistic):          3.23e-189
Time:                        20:54:21   Log-Likelihood:                 503.99
No. Observations:                 510   AIC:                            -984.0
Df Residuals:                     498   BIC:                            -933.2
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.1981      0.029      6.794      0.000       0.141       0.255
2019                    0.2340      0.008     28.703      0.000       0.218       0.250
temp                    0.4782      0.033     14.682      0.000       0.414       0.542
windspeed              -0.1480      0.025     -5.951      0.000      -0.197      -0.099
Sep                     0.0894      0.016      5.557      0.000       0.058       0.121
Sun                    -0.0495      0.012     -4.265      0.000      -0.072      -0.027
Spring                 -0.0544      0.021     -2.649      0.008      -0.095      -0.014
Summer                  0.0623      0.014      4.439      0.000       0.035       0.090
Winter                  0.0969      0.017      5.870      0.000       0.064       0.129
Cloudy                 -0.0809      0.009     -9.324      0.000      -0.098      -0.064
Light precipitation    -0.2904      0.025    -11.843      0.000      -0.339      -0.242
holiday                -0.1043      0.026     -4.029      0.000      -0.155      -0.053
==============================================================================
Omnibus:                       67.238   Durbin-Watson:                   2.102
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              163.903
Skew:                          -0.684   Prob(JB):                     2.56e-36
Kurtosis:                       5.417   Cond. No.                         17.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



```
In [36]: # Residual Sum of Squares (RSS)

         RSS_train = round(sum(train_residual**2),2)
         RSS_test = round(sum(test_residual**2),2)

         print('RSS of train data set =',RSS_train)
         print('RSS of test data set =',RSS_test)

         RSS of train data set = 4.14
         RSS of test data set = 2.03
```

ii. Error terms are normally distributed (not X, Y) : It can be observed by plotting the frequency of occurrence of error terms in both train and test data set. Moreover, plotting a Q-Q plot of test data set residuals is also helpful.

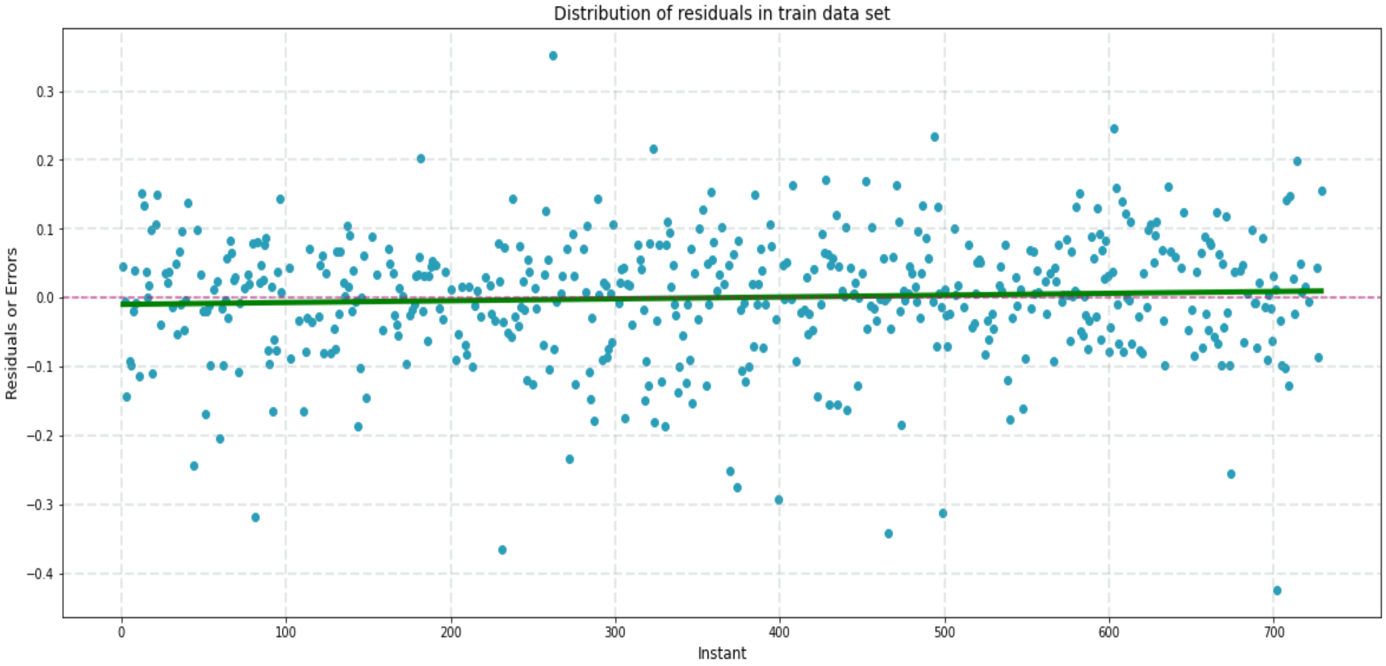




iii. Zero mean assumption,

iv. Error terms are independent of each other,

v. Error terms have constant variance (homoscedasticity) : All of these assumptions can be validated by plotting the distribution of residuals. The error points are entirely random and hardly is there any best fit line passing through the error points in both train and test data sets apart from y=0.



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

Ans. The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows (not in any sorting order) :-

i. Temperature – It is named as ‘temp’ in the data set which has a positive coefficient of 0.4782 and 14.682 t-score. This feature provides temperature in Celsius.

ii. Light Snow, Light Rain, Thunderstorm , Scattered clouds, Light Rain, Scattered clouds weather situation – It is named as ‘Light precipitation’ in the data set which has a negative coefficient of 0.2904 and -11.843 t-score. This feature classifies an instant with the corresponding weather condition.

iii. The year 2019 – It is named as ‘2019’ in the data set which has a positive coefficient of 0.2340 and 28.703 t-score. This feature classifies an instant with the corresponding year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression refers to a group of techniques for fitting and studying the straight-line relationship between two variables. Linear regression estimates the regression coefficients β_0 and β_1, β_2 ,………, β_p in the equation

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

where X_1, X_2 ,……., X_p are the independent variable, Y is the dependent variable, β_0 is the Y intercept, β_1, β_2 ,…, β_p are the slopes, and ε is the error.

Let,

Y_i = Actual target value
Y_{pred} = Predicted target value

In case of Simple Linear Regression,

$\varepsilon = Y_i - Y_{pred}$

$\Rightarrow \varepsilon = Y_i - (\beta_0 + \beta_1 X_1)$

$\Rightarrow \varepsilon^2 = [Y_i - (\beta_0 + \beta_1 X_1)]^2$

Similarly, in case of Multiple Linear Regression,

$$\therefore \ \text{RSS} = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_i X_i)]^2$$

The RSS is minimized using gradient descent. The suitable RSS value then leads to the building of a model with lowest residuals and errors.

Let, $\bar{y}$ = mean of all data points in the actual target values

$$\text{TSS} = \sum_{i=1}^{n} [Y_i - \bar{y}]^2$$

R^2 is the value that is calculated to determine the coverage of data variance. The quality of the model is directly proportional to R^2.

$$R^2 = 1 - (\text{RSS} / \text{TSS})$$

For any feature, the null hypothesis is taken equal to 0 and alternate hypothesis not equal to 0.

$$\left.\begin{array}{l} H_o = 0 \\ H_1 \neq 0 \end{array}\right\}$$

Now, the P-value is calculated based on the t-score,

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

If P-value < 0.5, the null hypothesis can be rejected and hence the feature is significant.

In the prepared model, there could be multicollinearity which refers to how a predictor variable is related to other variables excluding the target variable. Multicollinearity is measured using Variance Inflation Factor (VIF).

$$\text{VIF}_\text{i} = \frac{1}{1 - R_i^2}$$

Features with VIF values greater than 5 are generally ignored.

Based on the above factors, it is possible for multiple models to have been created each with different combinations and numbers of features, R-squared values and feature VIFs. Among all the models, the best one is chosen based on Adjusted R-squared term rather than R-squared.

This is because R-squared always increases with addition of features. It is obvious to cover more data variance with increase in features but it does not consider the number of predictor variables involved in the model.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N-1)}{N-p-1}$$

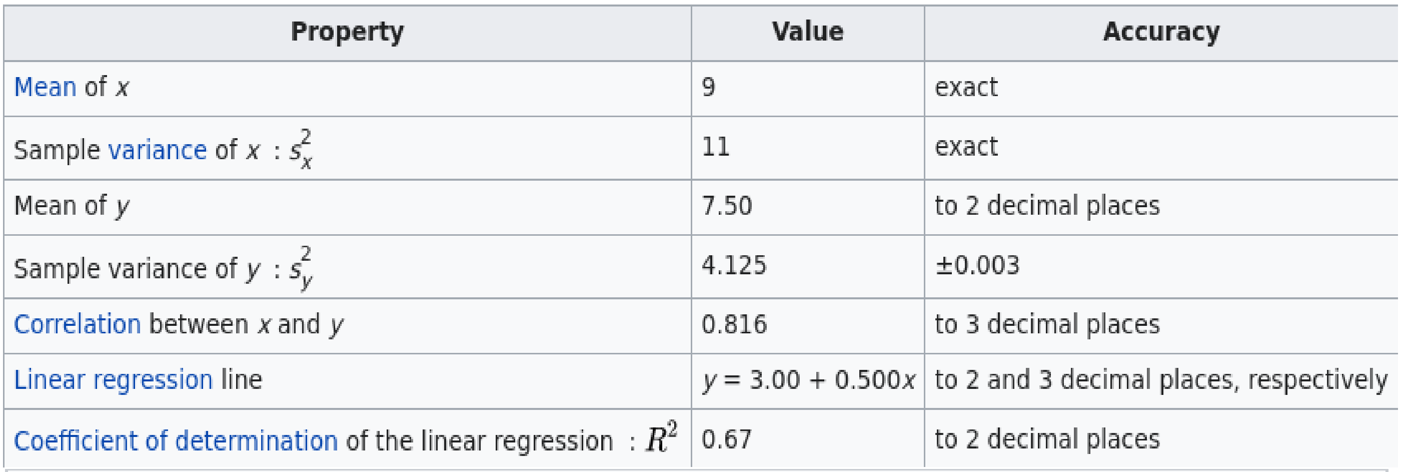
Based on all the above factors, a suitable and efficient Linear Regression model is built.

2. Explain the Anscombe's quartet in detail.

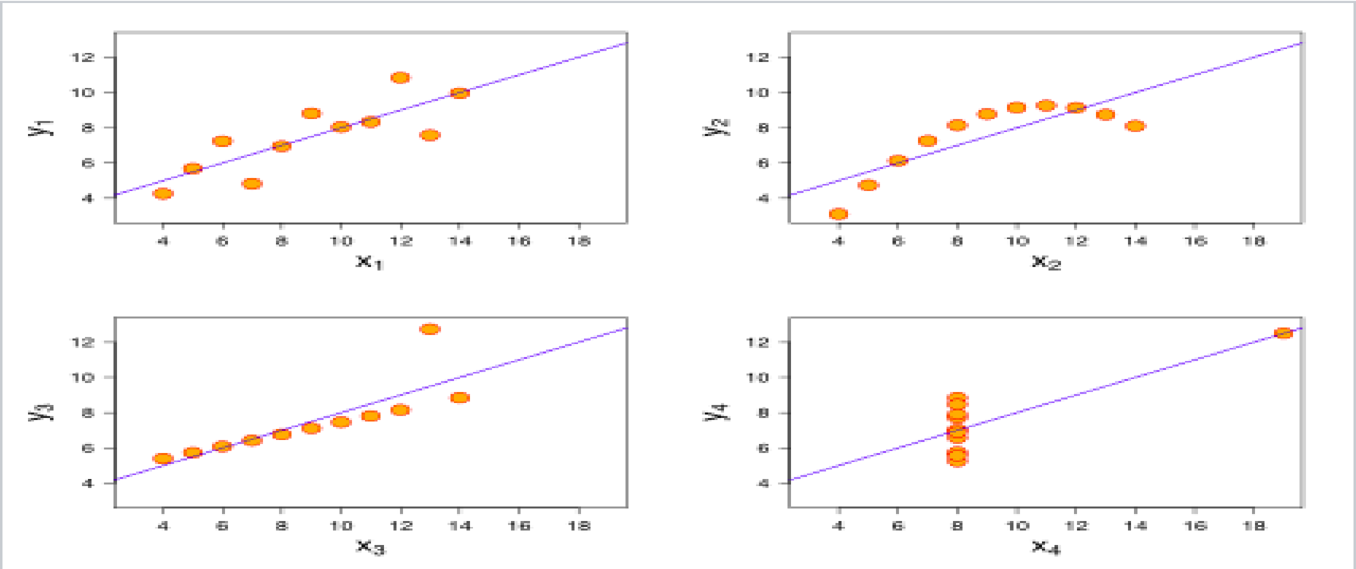
Ans. In 1973, an English statistician, Francis Anscombe came up with the concept that even though the statistical properties like mean, variance, correlation, etc. of different data sets are similar, the look very different when plotted during visualization.

Till then it was a common misconception among statisticians that once the data has been crunched into properties, it is useless to plot them anymore since they assumed them to be similar.

Refer to the following two images (Image source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet) for Francis Anscombe's work on four data sets used by him to prove his postulate :-



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places



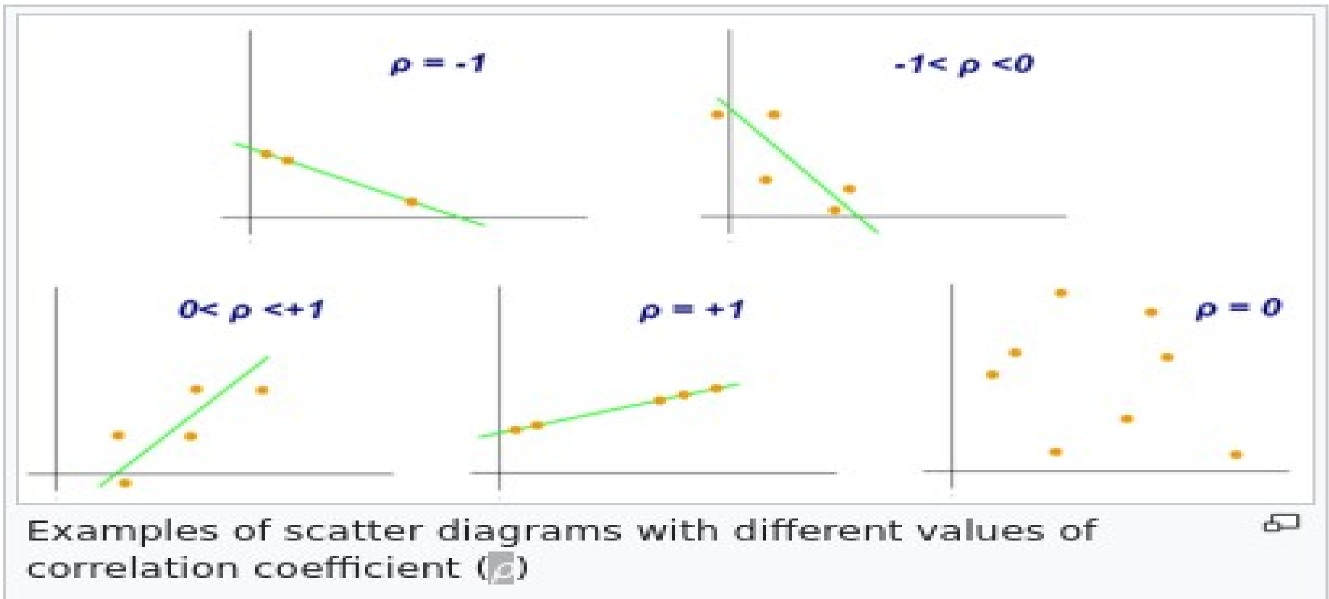
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson’s R?

Ans. Pearson’s r measures the strength of the linear relationship between two variables. It is calculated using the following formula.

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{(\sum x_i^2 - \frac{(\sum x_i)^2}{n})}\sqrt{(\sum y_i^2 - \frac{(\sum y_i)^2}{n})}}$$

The maximum and minimum value of pearson’s r value could be 1 and -1 respectively. It is also referred to as Pearson correlation coefficient (PCC), denoted by *ρ*. Refer to the following image (Image source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient) for detailed understanding :-



Examples of scatter diagrams with different values of correlation coefficient (ρ)

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

Ans. Scaling is the process of converting the variables in same comparable order. In other words, it is the process of increasing or decreasing the values of variables maintaining their ratios.

It is done in order to make the each variable values comparable to each other. Usually it may happen that in a data set, one column is in order of millions whereas others in order of hundreds or ones. In such cases the co-efficients will be large and will be complicated for those variables to work with. Therefore, scaling is performed on all variables. There are other reasons too for scaling which include ease of interpretation and faster convergence for gradient descent methods.

There are basically two types of scaling. Difference between them is summarized as below :

Normalized Scaling	Standardized 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. This is also called MinMax scaling. $x = \frac{x - min(x)}{max(x) - min(x)}$	The variables are scaled in such a way that their mean is zero and standard deviation is one. $x = \frac{x - mean(x)}{sd(x)}$

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

Ans. VIF is the measure of multicollinearity. Multicollinearity is the relation between two independent variables. Value of VIF is directly proportional to the multicollinearity.

We know,

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

R-squared is the factor of VIF of a certain feature, i.e., the R-squared value while considering that particular feature to be target variable and other features to be predictor variable.

From the equation it is clear that VIF can be zero only when the denominator is zero on the left hand side, that is,

$$1 - R_i^2 = 0$$

$$R_i^2 = 1$$

$$\text{or,} \quad R_i = 1$$

Therefore for VIF to be infinite, R-squared has to be equal to 1.

Again,

$$R^2 = 1 – (RSS / TSS)$$

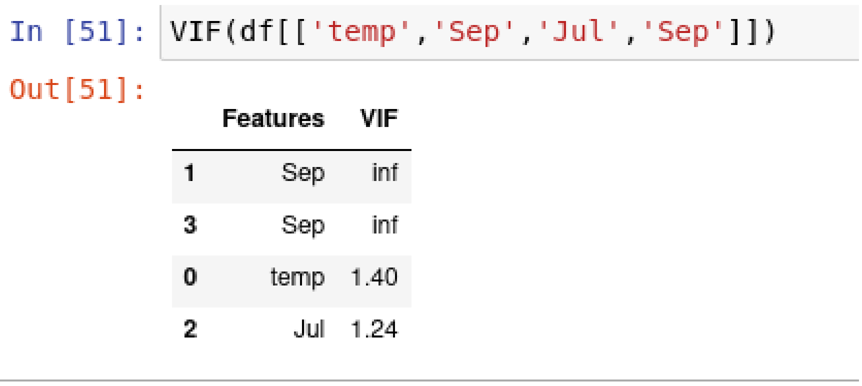
According to the above equation, for R-squared to be 1, RSS has to be equal to 0.

$$RSS = 0$$

This is plausible only if there is no residual error or no difference between actual and predicted values. This happens when there are columns with specific mathematical relationship between them. In simple terms, if one feature is a mathematically transformed or deduced value of another feature.

It can happen if two values have exact same values, i.e., one column is multiple of 1 of the other column.

Refer to the following image where VIF is calculated for some columns that inclued two similar columns ‘Sep’ with exact same values. They have infinite VIF values.



```
In [51]: VIF(df[['temp','Sep','Jul','Sep']])
```

Out[51]:

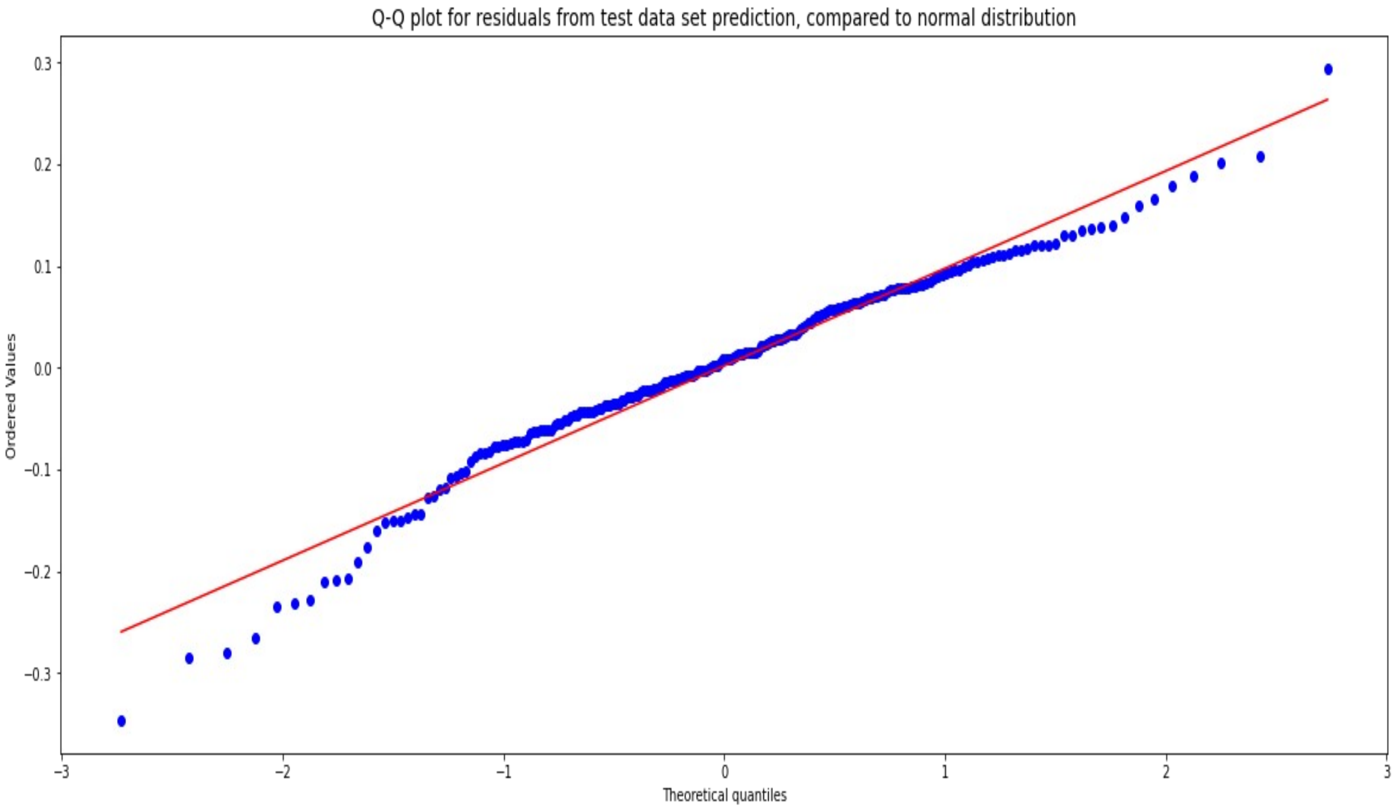
	Features	VIF
1	Sep	inf
3	Sep	inf
0	temp	1.40
2	Jul	1.24

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

Ans. A Q-Q plot, also known as Quantile-Quantile plot is a graphical technique for determining if a data set belongs to a population with any common distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The Q-Q plot is prepared by marking the quantiles in a data set and marking them with respective quantiles of another population from any specific distribution. This gives some data points in scattered format with a best fit line along with. More accurate the fit, more the target data set is comparable to the distribution chosen.

Following is an example of Q-Q plot used to validate if the error terms are normally distributed which is one of the assumption of linear regression. In fact, it is one of the major importance of a Q-Q plot in linear regression.


Q-Q plot for residuals from test data set prediction, compared to normal distribution
Ordered Values
Theoretical quantiles