Linear Regression Subjective Questions

Ashutosh Kulkarni

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?**

   - Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Most of the bookings has been done during the month of May, June, July, August, September and October. The trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
   - Pleasant wheathersit attracted more bookings.
   - Thursday, Friday, Saturday and Sunday have more number of bookings as compared to the start of the week.
   - On holidays the demand for bike rentals decreases.
   - The demand for bike rentals remains almost equal either on working days or non-working days.
   - Year 2019 attracted more number of booking from the previous year 2018, which shows good progress in terms of business.

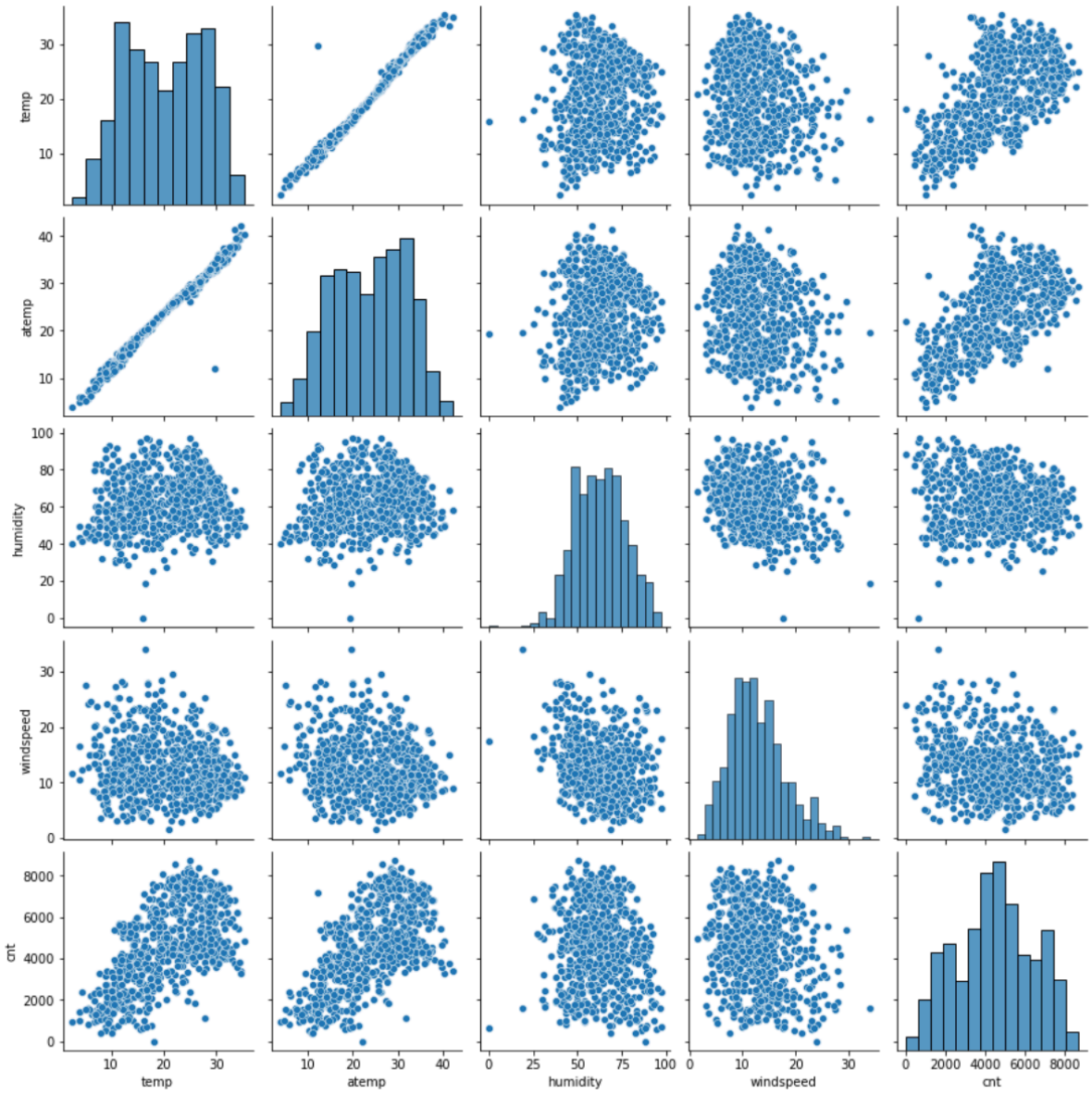
2. **Why is it important to use drop_first=True during dummy variable creation?**

   If there are n levels in a categorical variable, we can create n dummy columns (or variables) to represent each level with a distinct column. However it is important to consider the multicollinearity issues which can arise if we consider all n levels of dummy variable creation.

   If a ML model has two or more variables has the similar values or has same information, then it can result in Multicollinearity. It is observed that multicollinearity affects the interpretability of the model. If we create n dummy variables for n levels of categorical data, the nth variable contains no different information.

   Hence if we use drop_first = True then the only n-1 dummy variables will be created. This helps to avoid multicollinearity issues and maintain an interpretable model.

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

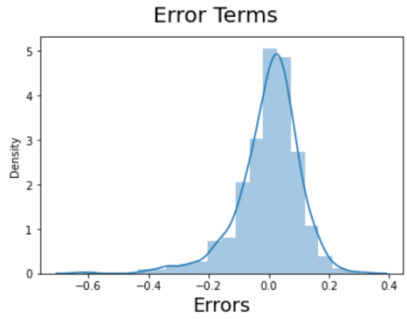


As seen from above pair-plot, it is quite evident that there is a clear trend between the `'temp'` and `'atemp'` variables and the target variable, `'cnt'`.

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

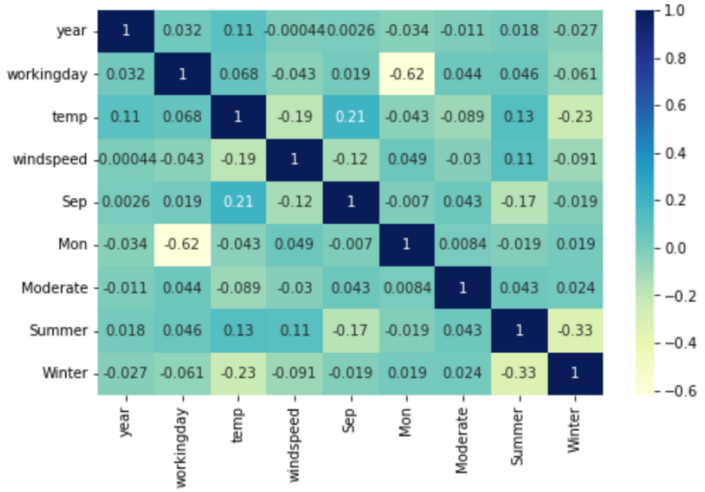
Assumption of Linear Regression Model are validated based on below 5 points -

- Normality of error terms - Error terms should be normally distributed

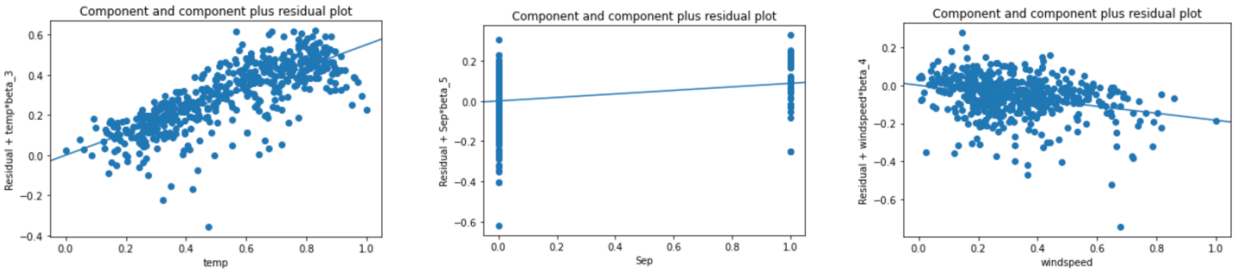


- Multicollinearity check - There should be insignificant multicollinearity among variables.

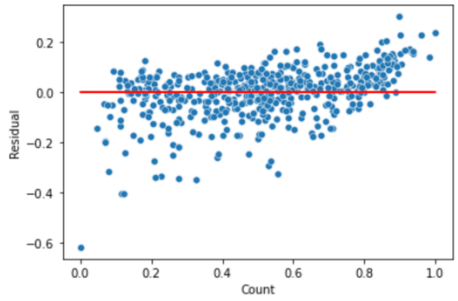
	Features	VIF
2	temp	4.76
1	workingday	4.02
3	windspeed	3.40
0	year	2.01
5	Mon	1.68
7	Summer	1.57
6	Moderate	1.50
8	Winter	1.38
4	Sep	1.20



- Linear relationship validation - Linearity should be visible among variables



- Homoscedasticity - There should be no visible pattern in residual values.



- Independence of residuals - No auto-correlation

Durbin-Watson value of final model is 2.067, which signifies there is no autocorrelation.

Durbin-Watson:	2.067
Jarque-Bera (JB):	452.216
Prob(JB):	6.35e-99
Cond. No.	11.5

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

The demand for shared bikes was inferred by looking at the coefficients of the variables in the final qualifying model. The coefficient for a specific variable estimated the amount of rise in demand when the variable value is increased by one unit while all the other variables are held constant. The coefficients of the categorical variable in the qualifying model.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	210.0
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	1.38e-163
Time:	12:04:51	Log-Likelihood:	437.81
No. Observations:	510	AIC:	-855.6
Df Residuals:	500	BIC:	-813.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0786	0.021	3.751	0.000	0.037	0.120
year	0.2388	0.009	25.858	0.000	0.221	0.257
workingday	0.0479	0.013	3.818	0.000	0.023	0.073
temp	0.5514	0.022	24.814	0.000	0.508	0.595
windspeed	-0.1838	0.028	-6.550	0.000	-0.239	-0.129
Sep	0.0876	0.018	4.948	0.000	0.053	0.122
Mon	0.0591	0.016	3.655	0.000	0.027	0.091
Moderate	-0.0663	0.010	-6.794	0.000	-0.086	-0.047
Summer	0.0886	0.012	7.646	0.000	0.066	0.111
Winter	0.1161	0.012	10.019	0.000	0.093	0.139

Omnibus:	136.515	Durbin-Watson:	2.067
Prob(Omnibus):	0.000	Jarque-Bera (JB):	452.216
Skew:	-1.225	Prob(JB):	6.35e-99
Kurtosis:	6.909	Cond. No.	11.5

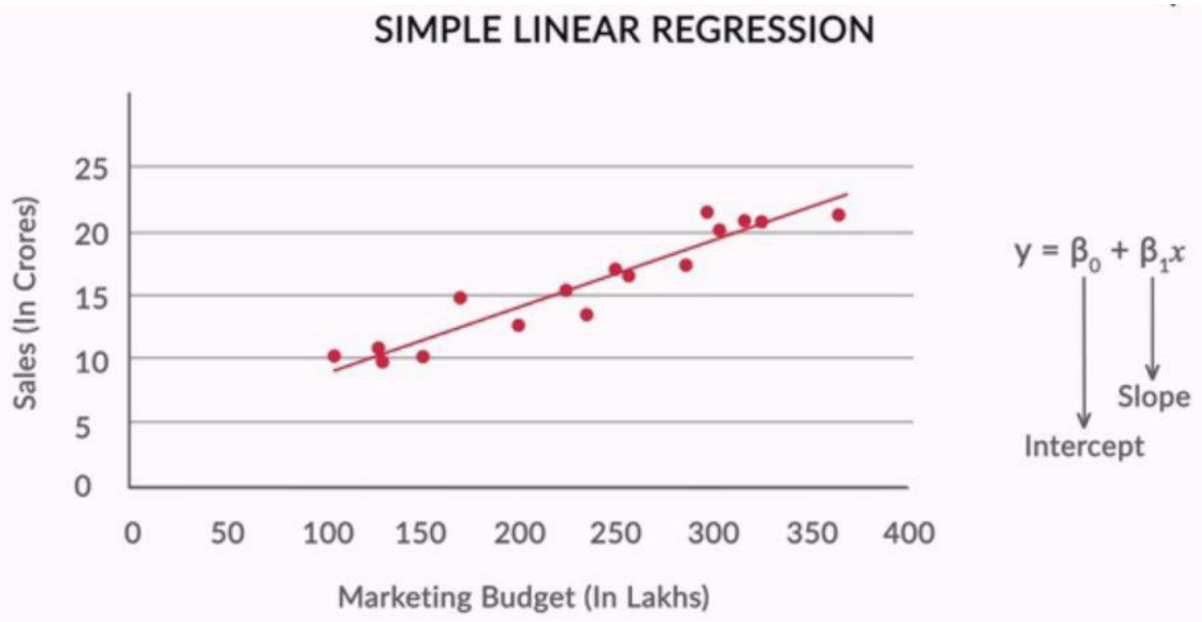
- Temperature (`temp`) - A coefficient value of ‘0.5514’ indicated that a unit increase in `temp` variable increases the bike hire numbers by ‘0.5514’ units.
- Year (`year`) - A coefficient value of ‘0.2388’ indicated that a unit increase in `year` variable increases the bike hire numbers by 0.2388 units.
- Wind Speed (`windspeed`) - A coefficient value of ‘-0.1838’ indicated that, w.r.t `windspeed`, a unit increase in `windspeed` variable decreases the bike hire numbers by 0.1838 units.

General Subjective Questions

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

Linear regression is a the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
The standard equation of the regression line is given by the following expression:

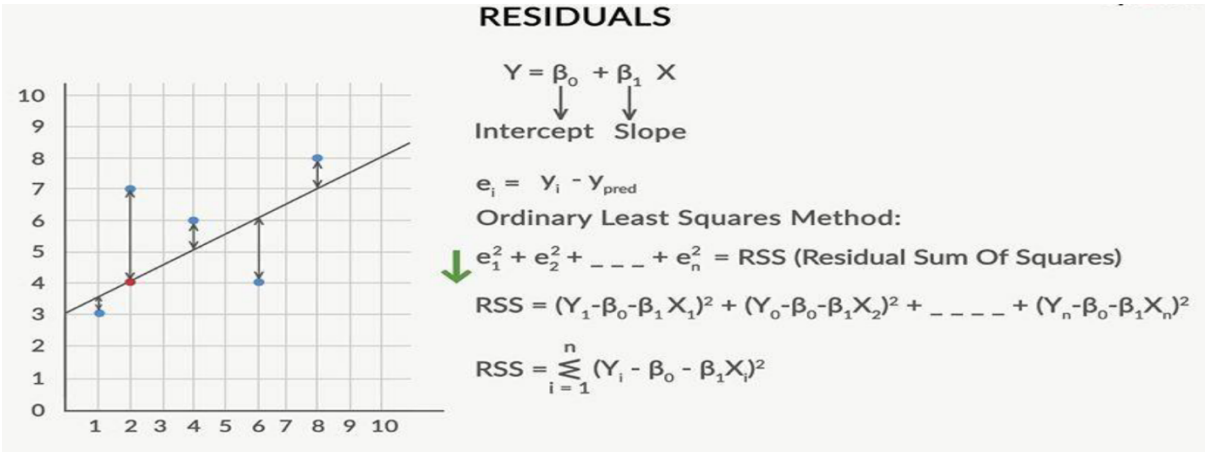
$$Y = \beta_0 + \beta_1 X$$



In the above example we can observe the linear relationship between Marketing Budget (independent) variable and the Sales (dependent) variable.

Once the linear trend observed, we can try to fit a line through the data points such that the error between the actual data points and the predicted values on the line, for a given value of x is minimum.

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



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

- R^2 or Coefficient of Determination
- Residual Standard Error (RSE)

We can also perform linear Regression with more than one independent variable but the dependent variable has to be only one.

In a nutshell, Linear Regression would be a simple and fairly accurate model if,

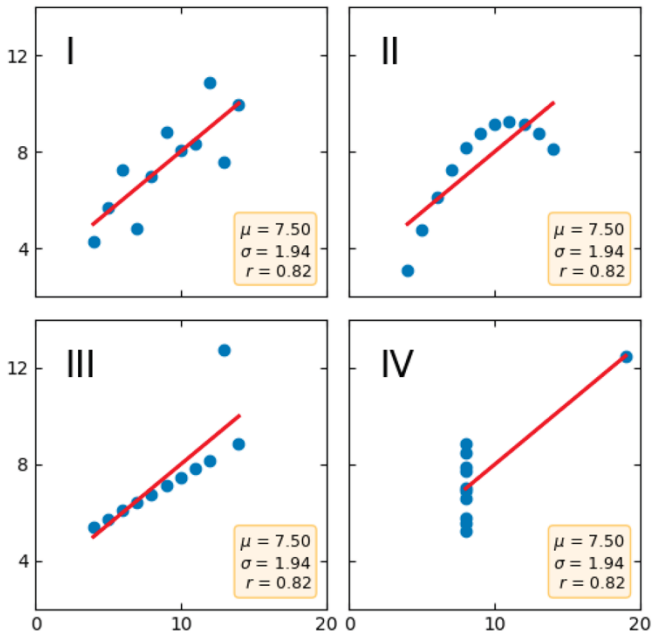
- There is a linear relationship between the independent variables and the dependent variable
- The fit of the line with the data points provides a fairly accurate predictions, given the values of the dependent variables (which is measured using the R-squared score)
- The error terms are normally distributed.

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

Anscombe's Quartet is a concept in Statistics which emphasizes the importance of data visualization data along with the the summary statistics.

The summary statistics consists of calculating values like the Average (mean), Median, Mode (if categorical), and the spread of data — minimum value, 25th percentile, 50th percentile (same as median), 75th percentile and maximum value — if it is a numerical column. The summary statistics of any given data provides the big picture instead of values of each datapoint and to intuitively estimate the ranges and values of the data. Although summary statistics provides a good idea about the data, it doesn't provide the visuals for the distribution of data.

Anscombe's Quartet is a collection of 4 different datasets with different individual data points having the same average values for the data yet widely different distributions.



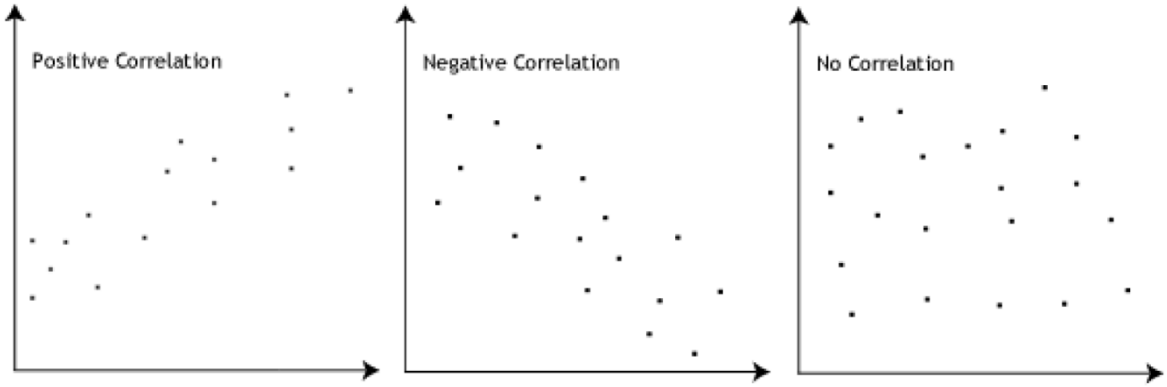
In the above picture, the summary statistics of the 4 distinct dataset have identical values - 7.50 for mean of the data, 1.94 for Standard Deviation and 0.82 for the Correlation Coefficient - yet they have vastly different distributions.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data it reveals a lot about structure of the dataset.

3. **What is Pearson's R?**

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables increases or decreases together, the correlation coefficient will be positive. If the variables increases or decreases in opposition with low values of one variable associated with high values of the other variable, then the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can have values ranging from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; so that the value of one variable increases, the value of other variable also increases. A value less than 0 indicates a negative association; so that the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Scaling is a process of converting a feature variable or multiple feature variables into a standard scale or a common scale.

This process is required because as multiple feature variables come with their own scales of data and their distribution. The ML model which is using the features to predict the response variable would implicitly take the feature variable with a higher scale to be more important than a feature variable with a lower scale.

Another reason the scaling is recommended is it makes the interpretation of the model really complex. Due to variations in the scales, the coefficients for different variables would have extremely high and low values. Scaling helps to prevents the above issues and makes the data much more interpret-friendly. It also helps in improvising the performance of the process of finding the coefficients. This performance improvement of gradient descent algorithm is because the algorithm responds better to data in a restricted scale than in a low and high variable scales combinations.

The two common types of scalers used are

- MinMax Scaler (Normalized Scaler)
- Standard Scaler

S.NO.	MinMax scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation
3.	Scales values between [-1, 1].	Scale values are not bounded to a certain range
4.	Scaling is affected by presence of outliers	Scaling is not affected by presence of outliers
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization

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

When the value of VIF is infinite it implies a perfect correlation between two independent variables. In the case of perfect correlation, R-squared (R2) is equals to 1, which lead to 1/ (1-R2) as infinity.

In order to avoid perfect multicollinearity, we should identify and drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (Q-Q) plot is a graphical technique to determine if two data sets are derived from populations with a common distribution.

A 'Q-Q' plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. If the two sets derived from a population with the same distribution, the points should fall approximately along the 45 degree reference line. If the points fall distance apart from the reference line then the we can consider that the two data sets are derived from populations with different distributions.

Importance of Q-Q plot:

A Q-Q plot is useful to determine if two populations are of the same distribution. We can verify if the residuals follow the normal distribution which is one of the assumption in regression. A Q-Q plot helps to compare the sample distribution of the variable against any other possible distributions graphically.