Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. Summer and winter seasons (against fall) have a positive relation with the bike rentals too - implying higher demand. However, the spring season implies lower demand compared to Fall.
2. Year 2019 saw significant increase in the demand over 2018 possibly due to increasing popularity/visibility
3. We see almost a quadratic trend for Rentals by months **- the Rentals increase as the year progresses, peak at June/July and then gradually decline as we approach the end of the year**
4. As against April, the month of September shows a positive demand seasonality while the month of January sees a cyclically lower seasonal demand
5. The adverse weather conditions of Light Rain and thunderstorms as well as Mist/Cloudy leads to less people opting for bike rentals. Clearly, the most favorable weather for Shared Bike rentals is Clear / Less Cloudy, while the adverse weather condition like Light Rain/Thunderstorms see less people using shared bikes
6. There isn't significant difference in Combined Rentals between Working and Non-working Day, however, the holidays within non-working day see a considerable dip in Rentals
7. As expected, Casual Bike Rentals are significantly higher on non-working days, however, registered Rentals are higher on working days
8. Casual Rentals are much higher on weekends with weekdays being consistently lower. However, registered rentals are higher on weekdays, Wednesday being the highest

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: drop_first=True ensures we do not fall for the concept called “Dummy trap”. If there are n levels within a categorical variable, then n-1 number of dummies (dropping one of the level’s dummy) ensures we do not create a case of perfect multicollinearity because a linear combination of n-1 dummies will generate the dropped dummy. Hence, it is a very important part of dummy creation

---

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: “temp” (the variable Temperature) has the highest correlation with the target. Given that “temp” and “atemp” signify very similar concept, they are highly correlated with each other and both of them have almost same correlation with the target

---

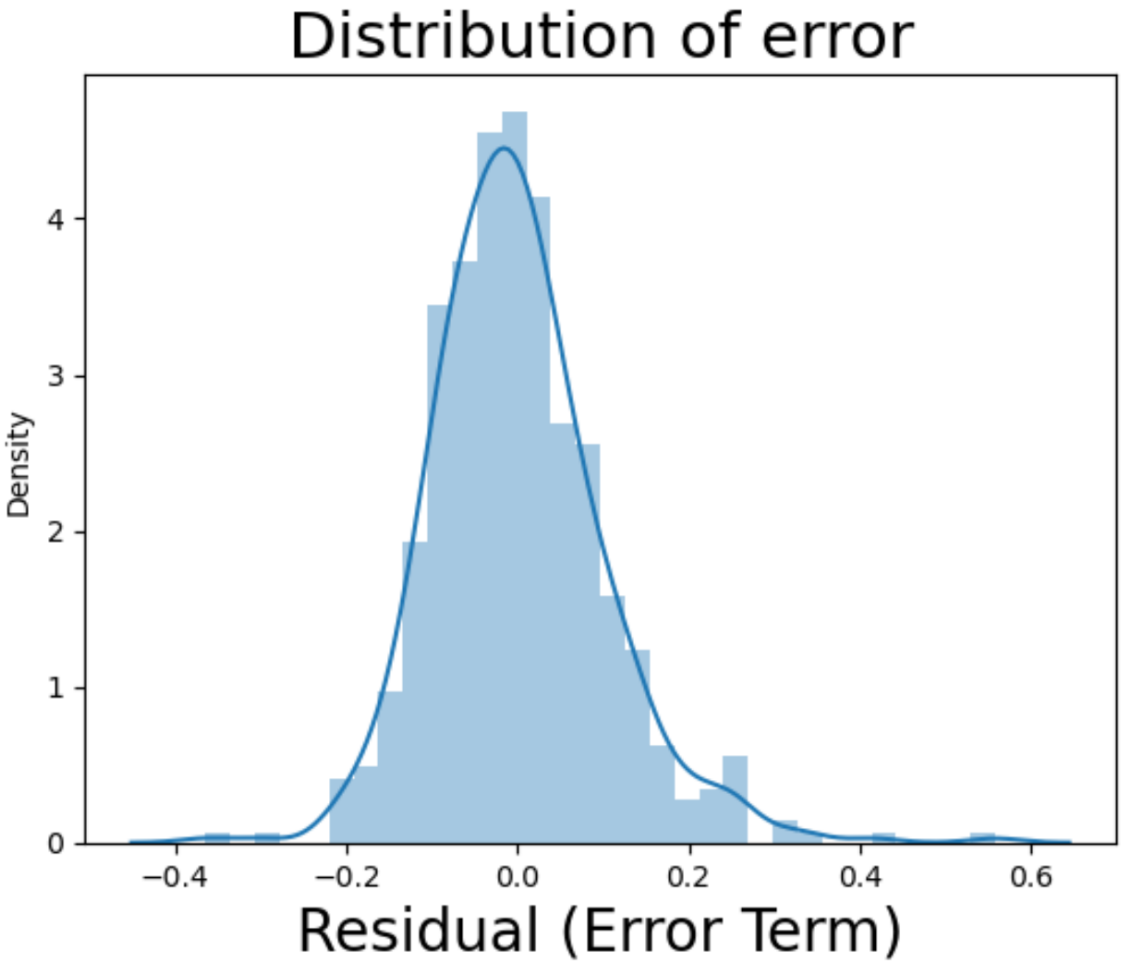
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

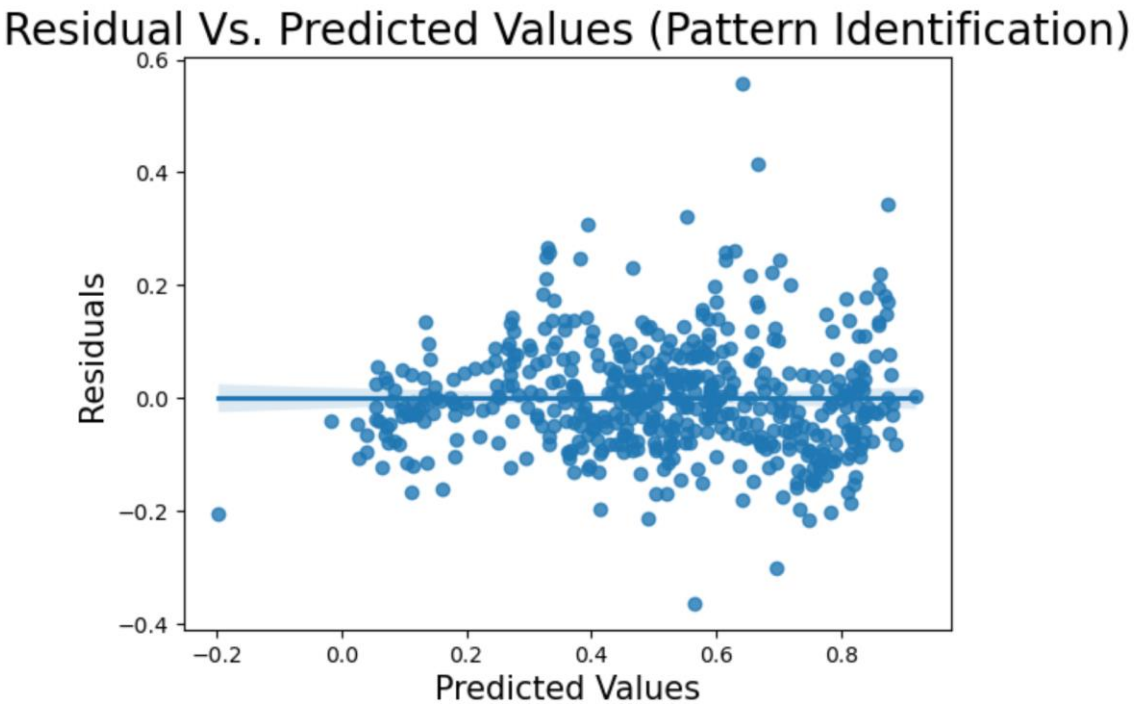
Answer:

1.Linearity in parameters: None of the beta coefficients are non linear

2 Normally distributed errors: Plotted the residuals and they seem normally distributed, centered around 0



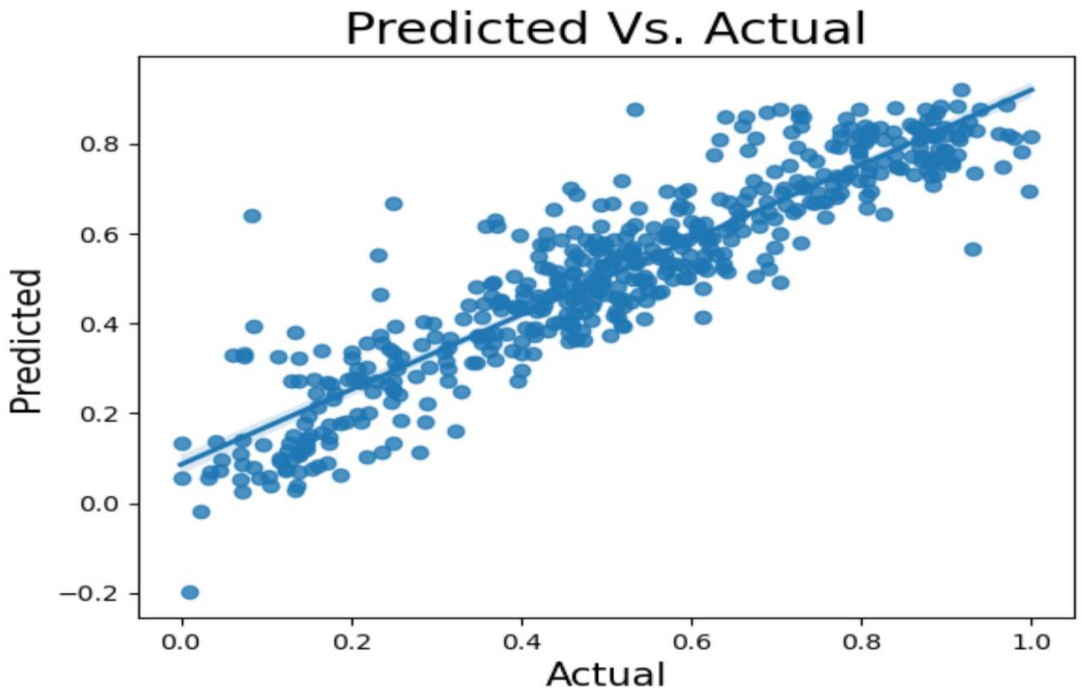
3 Independence of error term: The plot of Residual v/s Predicted doesn't show any specific pattern being followed, no specific concentration. The correlation between the residuals and the predicted values is almost negligible



4 Multicollinearity – The correlation matrix of the feature vector and the range of VIF values confirm that the problem of multicollinearity doesn't exist in the final model

	Features	VIF
0	windspeed	4.90
1	season_spring	4.57
2	temp	4.18
3	mnth_February	2.47
4	season_winter	2.45
5	mnth_January	2.38
6	yr_2019	2.11
7	mnth_November	1.91
8	weathersit_Mist & Cloudy	1.54
9	mnth_December	1.54
10	mnth_September	1.20
11	weathersit_Light Snow & Rain	1.10

5 Heteroscedasticity check – The scatter plot shows an evenly distributed pattern; hence the assumption of Homoscedasticity holds!



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer:

Basis the table below, the variables “temp” (Temperature), “weathersit_Light Snow & Rain” (Indicator of Light Rain/Snow + Thunderstorms weather) and “yr_2019” (Indicator of year 2019 over year 2018) are the top 3 features of the model that contribute significantly in explaining the demand of shared bikes

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2956	0.027	10.823	0.000	0.242	0.349
temp	0.3486	0.032	10.869	0.000	0.286	0.412
windspeed	-0.1266	0.030	-4.231	0.000	-0.185	-0.068
weathersit_Light Snow & Rain	-0.2950	0.028	-10.380	0.000	-0.351	-0.239
weathersit_Mist & Cloudy	-0.0801	0.010	-8.139	0.000	-0.099	-0.061
season_spring	-0.0912	0.021	-4.407	0.000	-0.132	-0.051
season_winter	0.1025	0.016	6.302	0.000	0.071	0.134
yr_2019	0.2706	0.009	29.455	0.000	0.253	0.289
mnth_December	-0.1142	0.022	-5.263	0.000	-0.157	-0.072
mnth_February	-0.0661	0.024	-2.741	0.006	-0.114	-0.019
mnth_January	-0.0918	0.025	-3.679	0.000	-0.141	-0.043
mnth_November	-0.1077	0.021	-5.034	0.000	-0.150	-0.066
mnth_September	0.0600	0.016	3.658	0.000	0.028	0.092

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. It could be univariate (with one independent variable) as well as multivariate (more than one independent variables).
Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. We use multiple linear regression when you want to know:
1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how age, education level, and profession affect income)
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected income at certain levels of age, education level, and profession)

$$h = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots$$

Here, 'h' is called the hypothesis. This is the predicted output variable or the dependent variable. Theta0 is the bias term or the intercept and all the other theta values are variable coefficients. They
are initiated randomly in the beginning, then optimized with the algorithm so that the formula can predict the dependent variable as accurately as possible.
Setting up the cost function :
The most commonly used cost function is "Squared errors" – which is the sum of squared differences between the actual dependent variable and the predicted. It can be represented as

$$J(\theta_0, \theta_1, \theta_2, ..) = \frac{1}{2m} \sum (h_i - y_i)^2$$

As we can see above, J, with Beta parameters is a function of actual and predicted values. This means we need to get the set of Beta values which minimize the sum of squared errors, which brings
us to the next step of the Multivariate Linear regression algorithm – Optimization

Optimization :
The Optimization algorithm is needed to minimize the cost function we've denoted above. The idea is to start with the initial values of Beta coefficients (could be 0 or randomized) and then iteratively update the Beta values so that the cost function (SSE) keeps on reducing with each step. One of the ways of doing it is Gradient Descent algorithm.
In Gradient Descent algorithm, we take the partial differential of the cost function with respect to each Beta (Theta) value and deduct the value from each Beta (Theta) value. Alpha is a learning rate and it is usually constant which determines the magnitude of each step taken to reach the global minima.

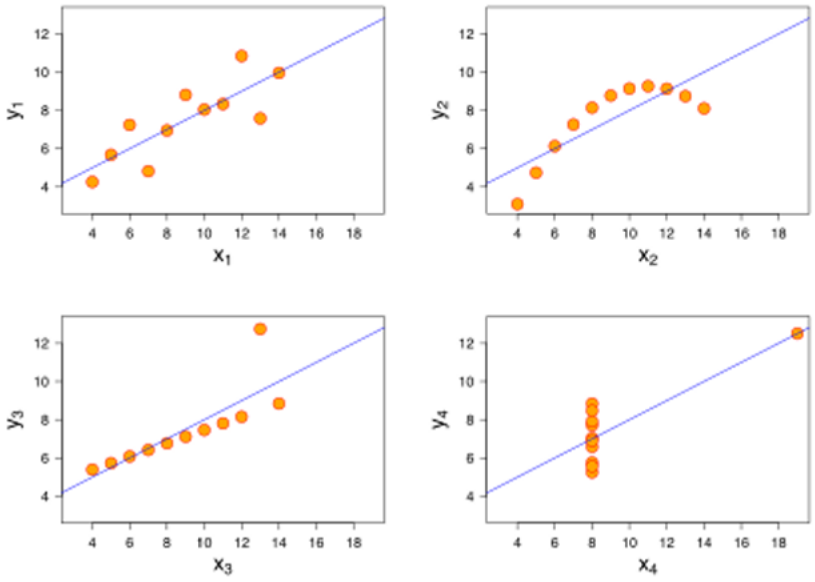
$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \Sigma (h_i - y_i)$$

$$\theta_i = \theta_i - \alpha \frac{1}{m} \Sigma (h_i - y_i) X_i$$

We usually scale the variables before we run the optimization algorithm as It leads to faster convergence.
PS – There are multiple optimization algorithms available to do this, e.g., Newton Raphson, Fisher's etc, however, due to the ease of scaling Gradient Descent, it is the most widely used one.
We need to validate the assumptions of linear regression to ensure that the estimates are BLUE (Best linear unbiased estimate).
We also need to check the performance of the model on the training set as well as the test set. Some of the commonly used metrics are R Squared, Adjusted R Squared, RMSE, Actual v/s Predicted
plots, Residuals plot. The metrics should not vary drastically across training and test sets as it may very well indicate the problem of overfitting (We randomly split the overall sample into Test and Training sets, the model is fit on the training set and then the test set is scored using the coefficients
obtained from training the model)

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The theory of Anscombe's quartet reinforces the importance of visualizing the data in any analytic model building process as against relying only on statistical measures.
Statistics are great for describing general trends and aspects of data, but statistics alone cannot fully
depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet," are shown below.

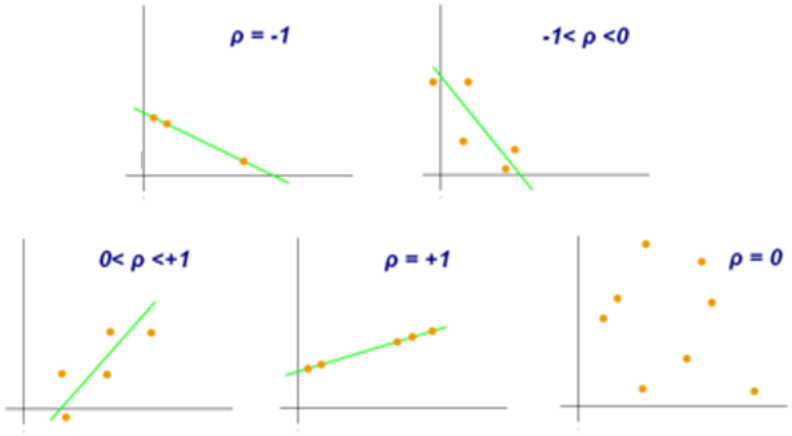


All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear
regression. But, as we can clearly see, they are all quite different from one another.
Anscombe's Quartet is a great demonstration of the importance of graphing data to analyse it.
Given simply variance values, means, and even linear regression cannot accurately portray data in its
native form. Anscombe's Quartet shows that multiple data sets with many similar statistical
properties can still be vastly different from one another when graphed. Anscombe's Quartet focuses
on the dangers of outliers in data sets. If the bottom two graphs did not have that one point that
strayed so far from all the other points, their statistical properties would no longer be identical to
the two top graphs. In fact, their statistical properties would more accurately resemble the lines that
the graphs seem to depict.
For instance, while all four data sets have the same linear regression line, it is obvious that the top
right graph should not be analysed with a linear regression at all because it is a curvature.
Conversely, the top left graph probably should be analysed with a linear regression because it is
a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in
graphing your data before analysing it.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R is the Coefficient of Correlation (coined by Pearson). It measures the strength and
direction of linear relationship between 2 variables. The values can lie between -1 and +1, -1
indicating perfect negative correlation and +1 indicating perfect positive correlation. 0, on the

other hand, denotes no correlation or independence.



The formula is illustrated below :

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}}.$$

where:
n is sample size
x_{i},y_{i} are the individual sample points indexed with i
There are countless applications and uses of Pearson's R, for instance, in the context of linear regression, if we look at the correlation between independent variables, it helps us understand the degree of multicollinearity in the data. If we check the correlation between the dependent variable and the independent variables, it helps us get a view of how strongly a predictor is related (linearly) with the target, even before building the model. For Univariate linear regression, the R squared value is square of Pearson's R.
The only drawback is that, Pearson's R can be calculated only for numeric variables (unless we create
dummies for the categorical variables). If we want to understand the relationship between 2 categorical variables, we can do so by calculating Cramer's V (a modification of ChiSquare value). However, unlike Pearson's R, cramer's V only tell us the strength of relationship and not the direction, ranging between 0 and 1

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Scaling implies transforming a variable in order to bring the range of values on a specific scale or within a specified interval.

If we bring the data within the range of 0 - 1, we can do it using Min-max scaling which is also called normalized scaling :

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

If the variables are scaled in such a way that their mean is zero and standard deviation is one, it is called standardized scaling.

Scaling is done to bring the variables to the same scale, it helps in :

1. Ease of interpretation
2. Faster convergence of optimization algorithms, e.g., Gradient Descent

Scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p values, R-square, etc. The advantage of Standardisation over the other is that it does not compress

the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier). Standardised scaling will affect the values of dummy variables but Min

Max scaling will not.

With respect to model building, Scaling should always be done after the test-train split since we do

not want the test dataset to learn anything from the train data. So, if we are performing the test train split earlier, the test data will then have information regarding the data like the minimum and

maximum values, etc.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Variance inflation factor (VIF) is one of the indicators of linear association between the various predictors. To answer the question, let us look at how it is calculated –

Assume that there are 3 independent variables x1, x2 and x3.

To calculate the VIF of x1, we regress x1 on x2 and x3, calculate the R Squared value of the fitted model and the VIF value of x1 is calculated as : 1 / (1 – Rsquared). Imagine, if the R Squared value of

the regression of x1 on a2 and x3 was 1, it would make the denominator 0 and the RHS value (VIF) equal to infinity.

This means that if a given independent variable is a linear combination of one or more than one independent variables, the R Squared value for that variable's regression on other independent variables will be 1 making VIF as infinity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

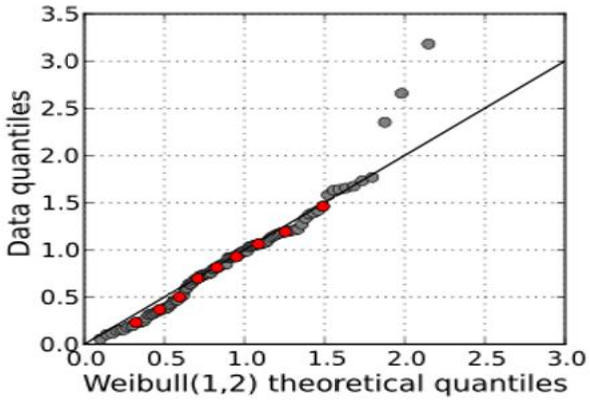
Q-Q plot or Quantile-Quantile plot is the graphical representation of the quantiles of two variables across 2 axes.

It helps us understand the proximity of the distribution of 2 variables.

It can also be used to understand if a given variable follows certain distribution or not. For example,

in the plot below, the quantiles of a variable have been plotted on the Y axis. On the X axis, however,

the theoretical quantiles of Weibull distribution are plotted.



When we are looking at the Q-Q plot, if the points fall near the 45 degree line, it means the probability distributions of the 2 variables are similar. A deviation from 45 degree line suggests differences in the distribution of the variables being compared.

It is incredibly useful even in linear regression, e.g., we can check if the dependent variable or residuals are following normal distribution or not (which is one of the assumptions of Linear Regression / OLS). The below chart from my assignment is an illustration –