Assignment-based Subjective Questions

Q1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans We can see effect of following categorical variables on dependent variable:

- season: For season=summer/fall, higher rides observed as compared to spring/winter
- yr: For yr=2019, higher rides as compared to 2018, same effect is observed for mnth, in 2019, each month have higher rides as compared to 2018
- mnth: For mnth=5->10, have higher rides as compared to other months
- weathersit: If weathersit=clear, then more rides as compared to Mist/Light Snow
- workingday: workingday has slightly more rides as compared to non-working day but different is not significant
- weekday: weekday=3 or 4 observe slightly more rides as compared to other days
- holiday: On holiday median of rides is significantly less as compared to non-holiday
- If we group, the continuous variables to categories then following are the observations
    - temp range [25,30] -> more rides
    - hum range [50, 70] -> more rides
    - windspeed range [5, 15] -> more rides

Q2 Why is it important to use drop_first=True during dummy variable creation?

Ans

- We cannot use categorical variables directly in machine learning models. They must be converted into meaningful numerical representations. This process is called encoding. The pandas.get_dummies() function converts categorical variables into dummy or indicator variables.
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it **reduces the correlations created among dummy variables**. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables, it removes the first column which is created for the first unique value of a column.
- Dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1. What you remove in redundancy, you gain confusion.
- Multicollinearity occurs when independent variables in a regression model are correlated. So why is correlation a problem? A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant. If all the variables are correlated, it will become difficult for the model to tell how strongly a particular variable affects the target since all the variables are related. In such a case, the coefficient of a regression model will not convey the correct information.

Q3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans Following variables have highest correlation with the target variable:

- registered (0.95)
- casual (0.67)
- atemp (0.63)
- temp (0.63)
- yr (0.57)
- season (0.4)

Q4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans

- Linear regression relies on a few key assumptions: linearity, homoscedasticity, absence of multicollinearity, independence, and normality of errors.
- Linearity
    - Linear regression assumes that there exists a linear relationship between the dependent variable and the predictors.
    - How can it be verified?
        - Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.
        - In addition, a partial residual plot that represents the relationship between a predictor and the dependent variable while taking into account all the other variables may help visualize the “true nature of the relationship” between variables.
    - What could it mean for the model if it is not respected?
        - If linearity is not respected, the regression will underfit and will not accurately model the relationship between the dependent and the independent variables.
    - What could be done?
        - Independent variables and the dependent variables could be transformed so that the relationship between them is linear. For instance, you could find that the relationship is linear between the *log* of the dependent variables and some of the independent variables squared.
- Homoscedasticity
    - Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.
    - How can it be verified?
        - To verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.
    - What could it mean for the model if it is not respected?
        - In the case of heteroscedasticity, the model will not fit all parts of the model equally and it will lead to biased predictions. It also often means that confounding variables, important predictors, have been omitted.
    - What could be done?

        - As heteroscedasticity generally reflects the absence of confounding variables, it can be tackled by reviewing the predictors and providing additional independent variables
- Absence of multicollinearity
    - Multicollinearity refers to the fact that two or more independent variables are highly correlated
    - How can it be verified?
        - Pairwise correlations could be the first step to identify potential relationships between various independent variables.
        - A more thorough method, however, would be to look at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others and calculating a score as follows: VIF = 1/ (1-R^2)
        - Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large *R-squared* for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multicollinearity.
    - What could it mean for the model if it is not respected?
        - The model may be producing inaccurate coefficient estimates that could thus not be interpreted. It may thus hurt inference power and possibly predictive performance.
        - In the presence of multicollinearity, the regression's results may also become unstable and vary tremendously depending on the training data.
    - What could be done?
        - Multicollinearity can be fixed by performing feature selection: deleting one or more independent variables.
        - A common approach is to use backward subset-regression: start by building a regression with all the potential independent variables and iteratively remove variables with high VIF and using domain-specific knowledge.
        - Another method could be to isolate and keep only the interaction effects between multiple independent variables (using intuition or regularization generally).
        - As multicollinearity is reduced, the model will become more stable, and the coefficients' interpretability will be improved.
- Independence of residuals (absence of autocorrelation)
    - Autocorrelation refers to the fact that observations' errors are correlated.
    - How can it be verified?
        - To verify that the observations are not auto-correlated, we can use the Durbin-Watson test. The test will output values between 0 and 4.
        - The closer it is to 2, the less autocorrelation there is between the various variables (0–2: positive autocorrelation, 2–4: negative auto-correlation).
    - What could it mean for the model if it is not respected?

        - Auto-correlation could mean that the linearity of the relationship is not respected or that variables may have been omitted.
        - Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable.
    - What could be done?
        - For time-series, one could add a lag variable. Another potential way to tackle this is to modify the variables from absolute value to relative.
        - More generally, variables should be further fine-tuned and added to the model.
- Normality of errors
    - If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.
    - How can it be verified?
        - To verify the normality of error, an easy way is to draw the distribution of residuals against levels of the dependent variable.
        - One can use a QQ-plot and measure the divergence of the residuals from a normal distribution. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.
    - What could it mean for the model if it is not respected?
        - If it is not respected, it may highlight the presence of large outliers or highlight other assumptions being violated (i.e. linearity, homoscedasticity). As a result, calculating t-statistics and confidence intervals with the standard methodologies will become biased.
    - What could be done?
        - In the case where errors are not normally distributed, one could verify that the other assumptions are respected (i.e. homoscedasticity, linearity), as it may often be a tell-tale sign of such a violation, and fine-tune the model accordingly.
        - Otherwise, one should also attempt to treat the large outliers in the data and check if the data could not be separate subsets using different models.

Q5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans Top three features contributing significantly towards explaining the demand of the shared bikes are:

1. atemp
2. weathersit
3. yr
4. windspeed
5. season

General Subjective Questions

Q1 Explain the linear regression algorithm in detail

Ans

- In simple terms, linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.
- In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.
- General equation of a straight line which is fitted during simple linear regression.

  $$y=\beta0+\beta1*x$$

  For this regression line, β0 is the intercept and β1 is the slope. For a unit increase in the quantity of x, y increases by β1*1 = β1 units.
- Residuals are defined as the difference between the y-coordinates of actual data and the y-coordinates of predicted data.

  $$e(i) = y(i) - y(pred), \text{ where } y(pred) = \beta0+\beta1*x$$
- The Ordinary Least Squares method has the criterion of the minimization of the sum of squares of residuals.

  Ordinary Least Squares Method: e1^2 + e2^2 + ... + en^2 = RSS (Residual Sum of Squares).

  $$RSS = \text{Summation } (i=1 \rightarrow n) (Y(i) – (\beta0 + \beta1*X(i)))^2$$

  The residual sum of squares tells you how much of the dependent variable's variation your model did not explain.

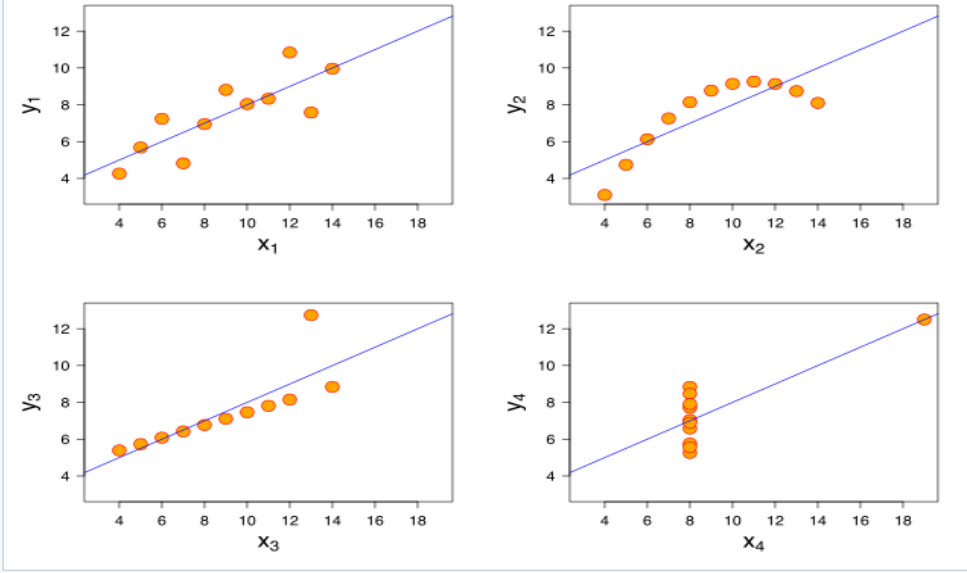
  The coefficients of the least squares' regression line are determined by the Ordinary Least Squares method — which basically means minimizing the sum of the squares of the: (y-coordinates of actual data - y-coordinates of predicted data)
- Ways to minimize cost function (RSS)
  - Differentiation
    - Differentiate the cost function and equate it to 0
    - You get 2 equations, solve those 2 equations and then you will get β0 and β1
  - Gradient Descent
    - Gradient Descent is an optimisation algorithm which optimises the objective function (for linear regression it's cost function) to reach to the optimal solution.
    - Start with some initial value of β0 and β1
    - Then iteratively move to better β0 and β1 such that cost function is minimized
- After determining the best fit line, there are a few critical questions you need to answer, such as:
  - How well does the best-fit line represent the scatterplot?
  - How well does the best-fit line predict the new data?
- TSS - Total Sum of Squares = Σ (Y(i) – mean of Y)^2

- The coefficient of determination, or R^2 = R^2 is a measure that provides information about the goodness of fit of a model.
    - R^2= 1 - RSS/TSS = 1 - (Σ (Y(i) – Y(pred))^2) / (Σ (Y(i) – mean of Y)^2)

Q2 Explain the Anscombe's quartet in detail.

Ans

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.
- Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.
- He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



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

- For all four datasets:

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

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.
- The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet

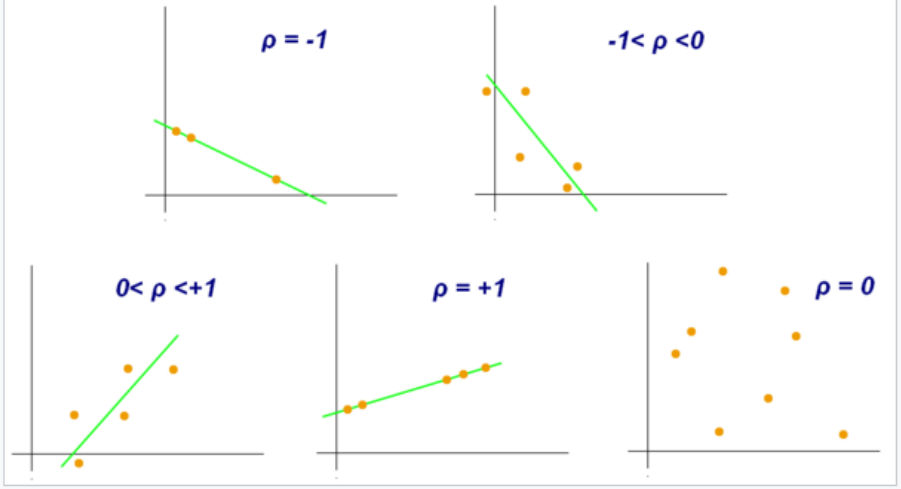
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Q3 What is Pearson's R?

Ans

- In statistics, the Pearson's R (Pearson correlation coefficient (PCC) / Pearson product-moment correlation coefficient (PPMCC) / Bivariate correlation / correlation coefficient) is a measure of linear correlation between two sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

- As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).
- Example of scatter plot with different values of correlation coefficient.



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

- Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables {\displaystyle (X,Y)}(X, Y) the formula for ρ is:

$$\rho(X, Y) = \text{cov}(X, Y) / (\sigma(x) * \sigma(y))$$

where cov(X, Y) is the covariance

σ(x) is the standard deviation of X

σ(y) is the standard deviation of Y

- Pearson's correlation coefficient, when applied to a sample, is commonly represented by r(xy) and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for r(xy) by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data {(x1, y1), (x2, y2),…, (xn, yn)} consisting of n pairs, r(xy) is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n is sample size

x(i), y(i) are the individual sample points indexed with i

x_bar = 1/n ∑i(i->n) x(i)

Rearranging gives us this formula for r(xy):

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

- Mathematical properties:
  - The absolute values of both the sample and population Pearson correlation coefficients are on or between −1 and 1. Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation).
  - The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).
  - A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient.
- Interpretation
  - The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

  More generally, note that (Xi – X_bar)(Yi – Y_bar) is positive if and only if Xi and Yi lie on the same side of their respective means. Thus the correlation coefficient is positive if Xi and Yi tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative (anti-correlation) if Xi and Yi tend to lie on opposite sides of their respective means. Moreover, the stronger is either tendency, the larger is the absolute value of the correlation coefficient.

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans

- What is scaling?
  - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- Why is scaling performed?
  - 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, regardless of the unit of the values.
  - Example: If an algorithm is not using the feature scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

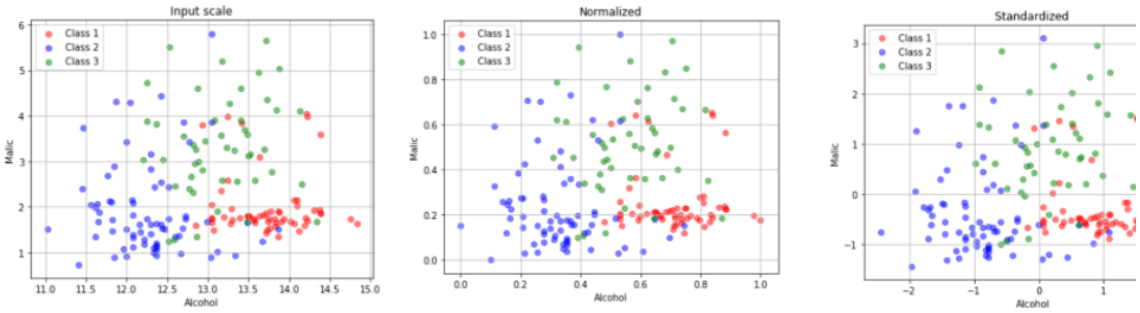
- What is the difference between normalized scaling and standardized scaling?
    - Normalized scaling
        - This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

    - Standardized scaling
        - It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

- Effect of normalization and standardization can be understood with following scatter plot:



Scatter plot of Raw data, Normalized data, Standardized data

    - In the raw data, feature alcohol lies in [11,15] and, feature malic lies in [0,6]
    - In the normalized data, feature alcohol lies in [0,1] and, feature malic lies in [0,1]
    - In the standardized data, feature alcohol and malic are centered at 0
- When to use what?
    - “Normalization or Standardization?” — There is no obvious answer to this question: it really depends on the application.
    - For example, in clustering analyses, standardization may be especially crucial to compare similarities between features based on certain distance measures. Another prominent example is the Principal Component Analysis, where we usually prefer standardization over normalization since we are interested in the components that maximize the variance.
    - However, this doesn’t mean that normalization is not useful at all! A popular application is image processing, where pixel intensities must be normalized to fit within a certain range (i.e., 0 to 255 for the RGB color range). Also, a typical neural network algorithm requires data on a 0–1 scale

Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans

- Multicollinearity (or collinearity) occurs when one independent variable in a regression model is linearly correlated with another independent variable.
- The presence of multicollinearity results in several problems:
  - The fitted regression coefficients will change substantially if one of the values of one of the x variables is changed only a bit.
  - The variance of the estimated coefficients will be inflated, which means that it will be hard to detect statistical significance. Furthermore, it's possible that the F statistic is significant but the individual t statistics are not.
  - Ultimately, multicollinearity makes prediction less accurate. For a given model, the underlying assumption is that the relationships among the predicting variables, as well as their relationship with the target variable, will be the same. However, when multicollinearity is present, this is less likely to be the case
- How to detect and eliminate multicollinearity
  - A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable.

$$VIF_j = \frac{1}{1-R_j^2}$$

  - VIF measures the ratio between the variance for a given regression coefficient with only that variable in the model versus the variance for a given regression coefficient with all variables in the model.
  - A VIF of 1 (the minimum possible VIF) means the tested predictor is not correlated with the other predictors.
  - The higher the VIF: 1) The more correlated a predictor is with the other predictors, 2) The more the standard error is inflated, 3) The larger the confidence interval, 4) The less likely it is that a coefficient will be evaluated as statistically significant
  - **If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.**

Q6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come 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 data set. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
- The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are

uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

- Interpretation
    - The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right.
    - If the two distributions being compared are identical, the Q–Q plot follows the 45° line y = x.
    - If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y = x.
    - If the general trend of the Q–Q plot is flatter than the line y = x, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis.
    - Conversely, if the general trend of the Q–Q plot is steeper than the line y = x, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis.
    - Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.
- The q-q plot is used to answer the following questions:
    - Do two data sets come from populations with a common distribution?
    - Do two data sets have common location and scale?
    - Do two data sets have similar distributional shapes?
    - Do two data sets have similar tail behavior?
- Importance
    - When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.
    - If two samples do differ, it is also useful to gain some understanding of the differences.
    - The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.