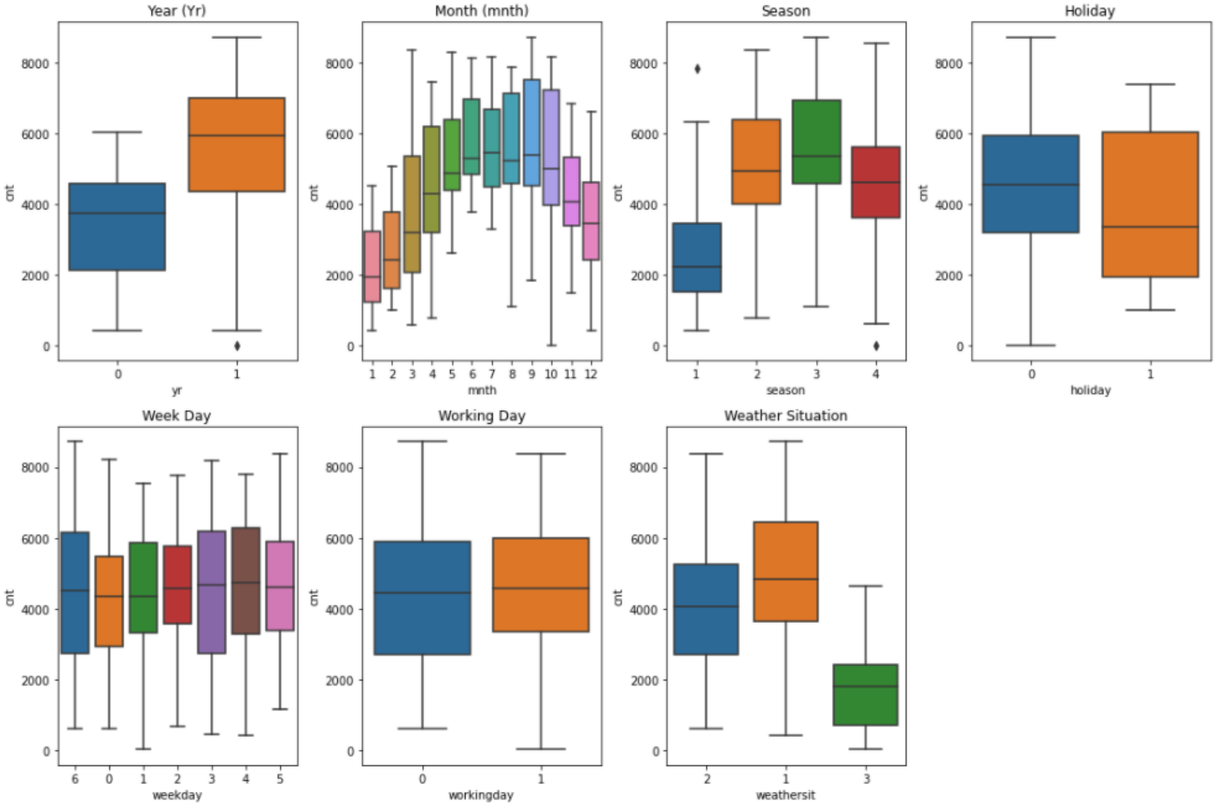
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Below are the categorical variables:

1. Year (yr)
2. Month (mnth)
3. Season
4. Holiday
5. Weekday
6. Working Day
7. Weather Situation (weathersit)

Their relationship with count of total rental bikes is as shown:



Year:

These is an upward trend every year. The median in 2019 is almost equal to the maximum of rentals in 2018.

Month:

When compared with median, we observe that starting month, January records lowest rentals but it continues to peak till June, stays there for almost four months, till September, reduces a little in October and starts to drop in November and December.

September is the month where the rentals peak both by 75 and 100 percentiles.

Season:

Rentals shows the best season is Fall and the next best season is Summer. They gets reduced in winter when compared to summer but the worst season for Rentals is Spring. 75% rentals in Spring are far lesser than the 25% of Winter, which is the next worst month after Spring.

Holiday:

There is a lot of variance on holidays whereas not much shown on non-holidays. Going by median (50 percentile), non-holidays fare better than the holidays but their 75 percentiles is almost same

Weekday:

The rentals are almost similar on each weekday if we go with the medians.

Working Day:

The rentals are almost similar on both working day and a non-working day, only a very little more variance is observed on a non-working day.

Weather Situation:

Weather Situation largely impacts the sales. Rentals peak when the weather is is clear or with Few clouds or if it is partly cloudy. Rentals drop when it is misty with few, broken or no clouds. Rentals drop to an alarming level where the 75% is also less than 25% of other weather situations when it is Raining/snowing with or without scattered clouds.

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

Any categorical variable with n-levels can be indicated with (n-1) dummy variables. For example the 4 seasons can be explained with three dummy variables.

Season	Spring	Summer	Fall
Spring	1	0	0
Summer	0	1	0
Fall	0	0	1
Winter	0	0	0

Here four seasons are represented by three variables, Spring, Summer and Fall. The seasons, Spring Summer and Fall are represented by 1 under their column but the absence of 1 in all the three columns i.e., 0 in every season column shows Winter.

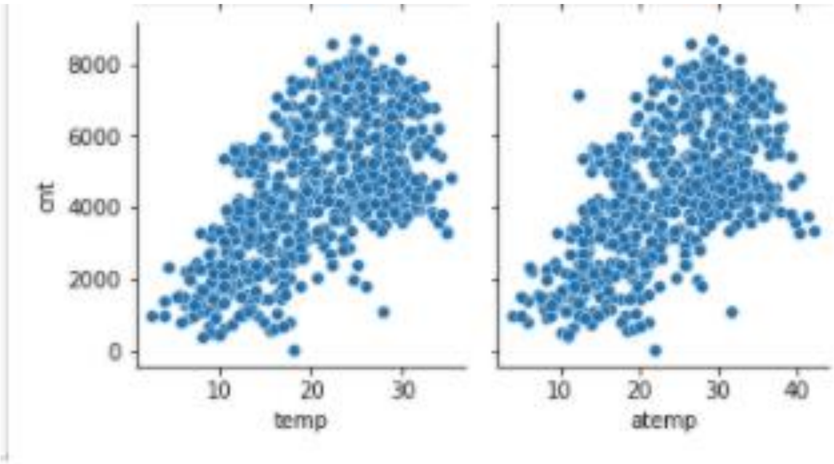
Here **drop_first=True** fulfils this behavior.
If **drop_first** is not specified or is false, then n dummy variables are created and it creates multicollinearity issue. When **drop_first=True**, then (n-1) variables are created and thereby mitigating the dummy variable trap which causes multicollinearity.

Specifying **drop_first=True** drops the first dummy variable. In our example, Spring will be dropped if it is the first occurring season in the data i.e., spring will be shown with all zeros.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temperature Felt (atemp) shows the maximum correlation with the target variable. Next comes the original temperature which is a fraction less.

	Actual Temperature (Temp)	**Temperature Felt (aTemp)**
Rentals Count Correlation	0.627	0.63

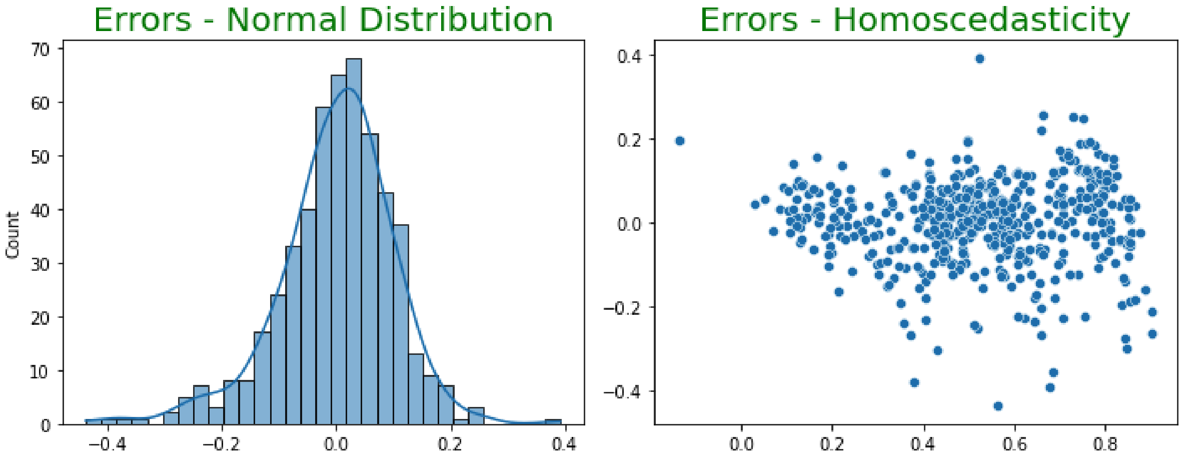


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The model obtained can be validated with the following

1. Errors are still normally distributed with zero mean.
2. Errors are independent of each other
3. Errors have constant variance i.e, homoscedasticity

These rules are true for the errors obtained for the model built. Errors were normally distributed with zero mean and they are independent of each other with almost constant variance.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

My equation for count is :

count = 0.3433 + (0.2457 * yr) + (0.3885 * atemp) + (-0.1584 * windspeed) + (-0.1725 * spring) + (-0.0379 * dec) + (-0.0722 * jul) + (-0.0419 * nov) + (-0.0810 * mist) + (-0.2842 * rain)

So top 3 features are

1. Temperature Felt (atemp) at 38.85%
2. Year (yr) at 24.57%
3. Weather situation, if its rainy, then almost 28.42% less rentals compared to clear weather and if its misty, then rentals are down by 8.1% compared to clear weather.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a method to predict the dependent variable (Y) based on values of independent variables (X). It can be used for the cases where we want to predict some continuous quantity.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point of the regression line.

Simple linear regression is an approach for predicting a quantitative response using a single feature (or “predictor” or “input variable”). It takes the following form called a Linear Regression Equation:

$$\mathbf{Y = \beta0 + \beta1 * x}$$

Y is the response
x is the feature
β0 is the intercept
β1 is the Multiplier / Unit change Together β0 and β1 are called the model coefficients

How to calculate the model coefficient

The following are the formulas to calculate the linear model coefficient.

$$\boldsymbol{\beta1 = \sum (X - \bar{x})(Y - \bar{y}) / \sum (X - \bar{x})2}$$

$$\boldsymbol{\beta o = \bar{y} - \beta1 \bar{x}}$$

$\bar{x}$ – mean of X
$\bar{y}$ – mean of Y

Estimating (Learning) Model Coefficients: Generally speaking, coefficients are estimated using the least squares criterion, which means we find the line (mathematically) which minimizes the sum of squared residuals (or “sum of squared errors”).

Linear Regression Line: While doing linear regression our objective is to fit a line through the distribution which is nearest to most of the points. Hence reducing the distance (error term) of data points from the fitted line.

Evaluation of Algorithm

MAE – This is mean absolute error. It is robust against the effect of outliers. Using the previous example, the resultant MAE would be (30-10) = 20

MSE – This is mean squared error. It tends to amplify the impact of outliers on the model’s accuracy. For example, suppose the actual y is 10 and predictive y is 30, the resultant MSE would be $(30-10)^2 = 400$.

RMSE – This is the root mean square error. It is interpreted as how far on average; the residuals are from zero. It nullifies the squared effect of MSE by square root and provides the result in original units as data. Here, the resultant RMSE would be $\surd\ (30-10)^2 = 20$. Don't get baffled when you see the same value of MAE and RMSE. Usually, we calculate these numbers after summing overall values (actual – predicted) from the data

Coefficient of Determination (R Square):-

It suggests the proportion of variation in Y which can be explained with the independent variables.

Mathematically: R2 = SSR/SST

or R2 = Explained variation / Total variation

or R2 = 1 – (Unexplained variation / Total variation) = 1-(RSS/TSS)

It explains the proportion of variation in the dependent variable that is explained by the independent variables.

Explained variation is the sum of the squared of the differences between each predicted y-value and the mean of y.

Explained variation = $\Sigma(\hat{y} - \bar{y})2$

Unexplained variation is the sum of the squared of the differences between the y value of each ordered pair and each corresponding predicted y-value.

Unexplained variation = $\Sigma(y - \hat{y})2$

Total variation about a regression line is the sum of the squares of the differences between the y-value of each ordered pair and the mean of y.

Total variation = $\Sigma(y - \bar{y})2$

Range of R Square from 0 to 1, R Square of 0 means that the dependent variable cannot be predicted from the independent variable, R2 of 1 means the dependent variable can be predicted without error from the independent variable, If the value of R Square is 0.912 then this suggests that 91.2% of the variation in Y can be explained with the help of given explanatory variables in that model. In other words, it explains the proportion of variation in the dependent variable that is explained by the independent variables.

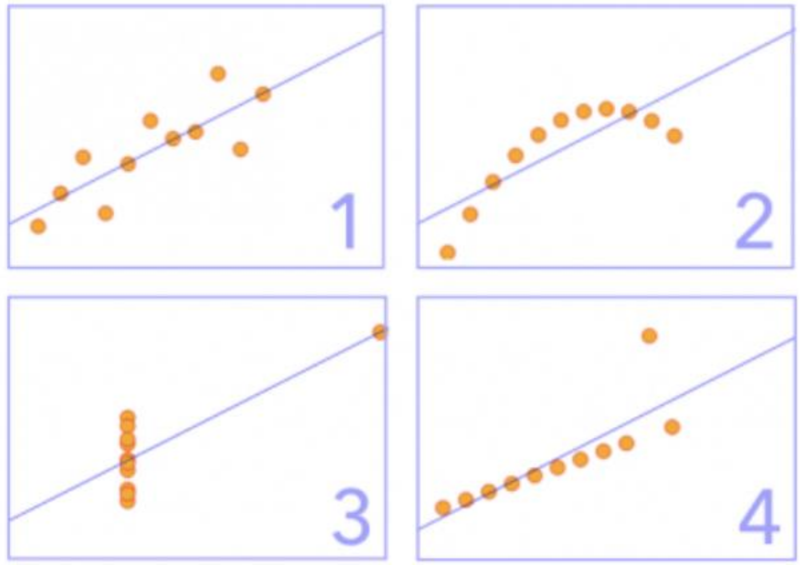
2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's quartet shows four different graphs with same statistical properties.

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

For all these four sets of data, statistical properties are same:

S.No	Property	Value
1	Mean(x)	9
2	Standard_deviation(x)	3.32
3	Mean(y)	7.5
4	Standard_deviation(y)	2.03
5	correlation between x and	0.81
6	Linear equation line	y = 3 + 0.5x
7	Coefficient of determination (r-square)	0.67



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

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.

3. What is Pearson’s R? (3 marks)

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression.

It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data?
It can range from -1 to +1.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn’t a positive or negative increase. The two just aren’t related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.75| = .75, which has a stronger relationship than .65.
One of the most commonly used formulas is Pearson’s correlation coefficient formula. If you’re taking a basic stats class, this is the one you’ll probably use:

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

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Sx and sy are the sample standard deviations, and sxy is the sample covariance.

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The population correlation coefficient uses σx and σy as the population standard deviations, and σxy as the population covariance.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

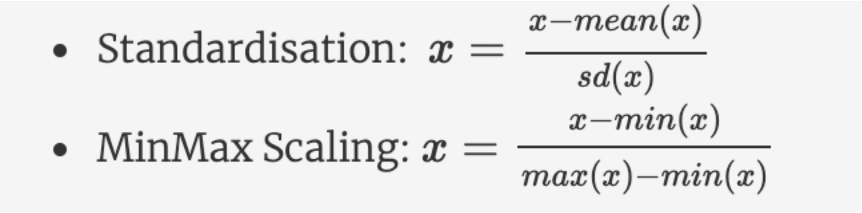
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are two major methods to scale the variables, i.e. standardization and Minmax scaling.

Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. Use Standard Scaler if you know the data distribution is normal.

MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1. It will transform each value in the column proportionally within the range [0,1]. Use this as the first scaler choice to transform a feature, as it will preserve the shape of the dataset (no distortion).

The formulae in the background used for each of these methods are as given below:



- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

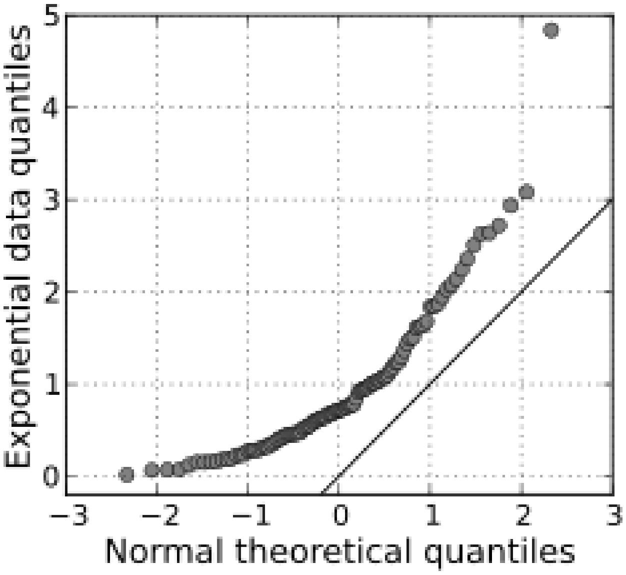
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

   A large value of VIF indicates that there is a correlation between the variables.
   If there is perfect correlation, then r-square becomes 1. Thereby VIF becomes infinity.

   1/(1-r_square). = 1 / 0 = infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It’s being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a normal quantile-quantile plot. The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

How to Make a Q-Q Plot?

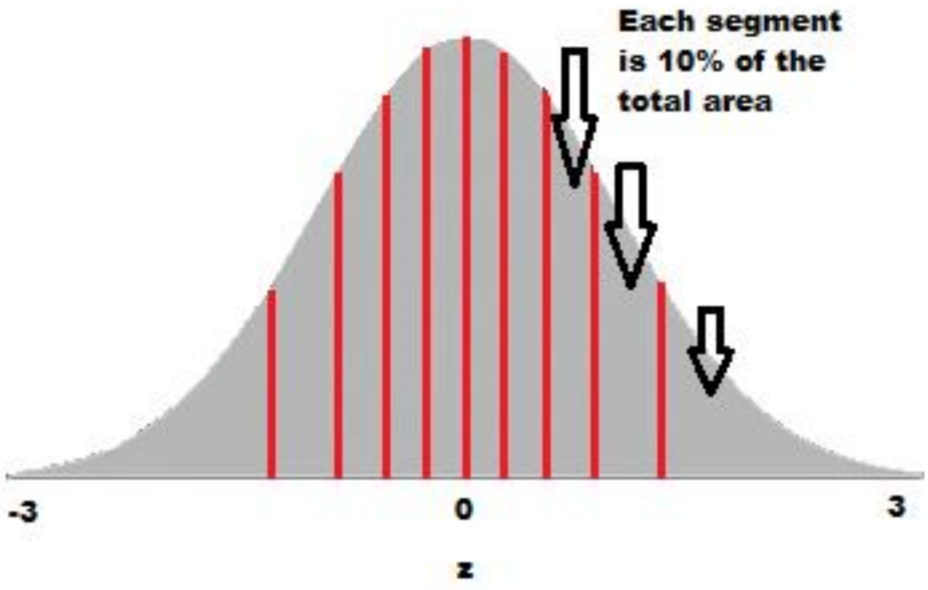
Do the following values come from a normal distribution?

7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: **Order the items from smallest to largest**.

- 3.77
- 4.25
- 4.50
- 5.19
- 5.89
- 5.79
- 6.31
- 6.79
- 7.19

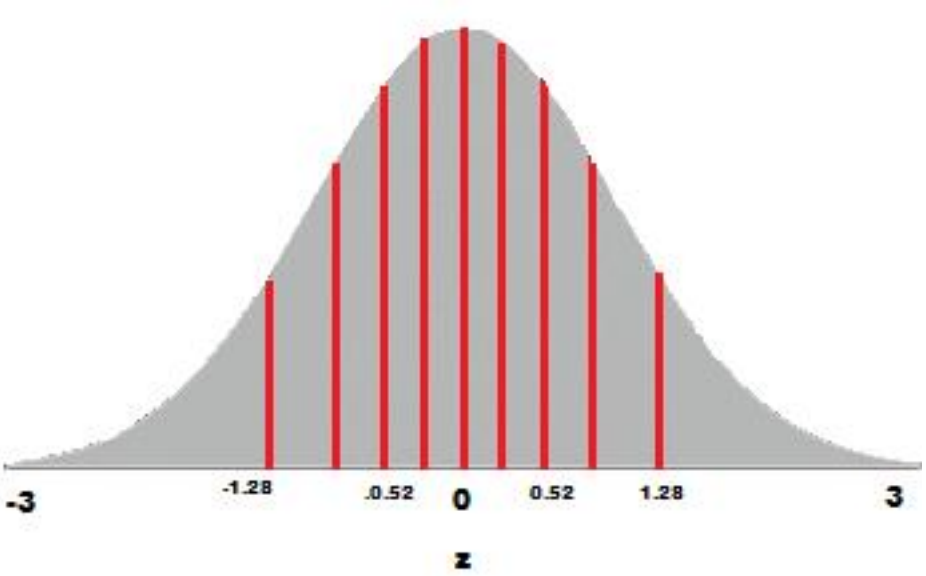
Step 2: **Draw a normal distribution curve.** Divide the curve into n+1 segments. We have 9 values, so divide the curve into 10 equally sized areas. For this example, each segment is 10% of the area (because 100% / 10 = 10%).



Step 3: Find the z-value (cut-off point) for each segment in Step 3. These segments are areas, so refer to a z-table (or use software) to get a z-value for each segment.

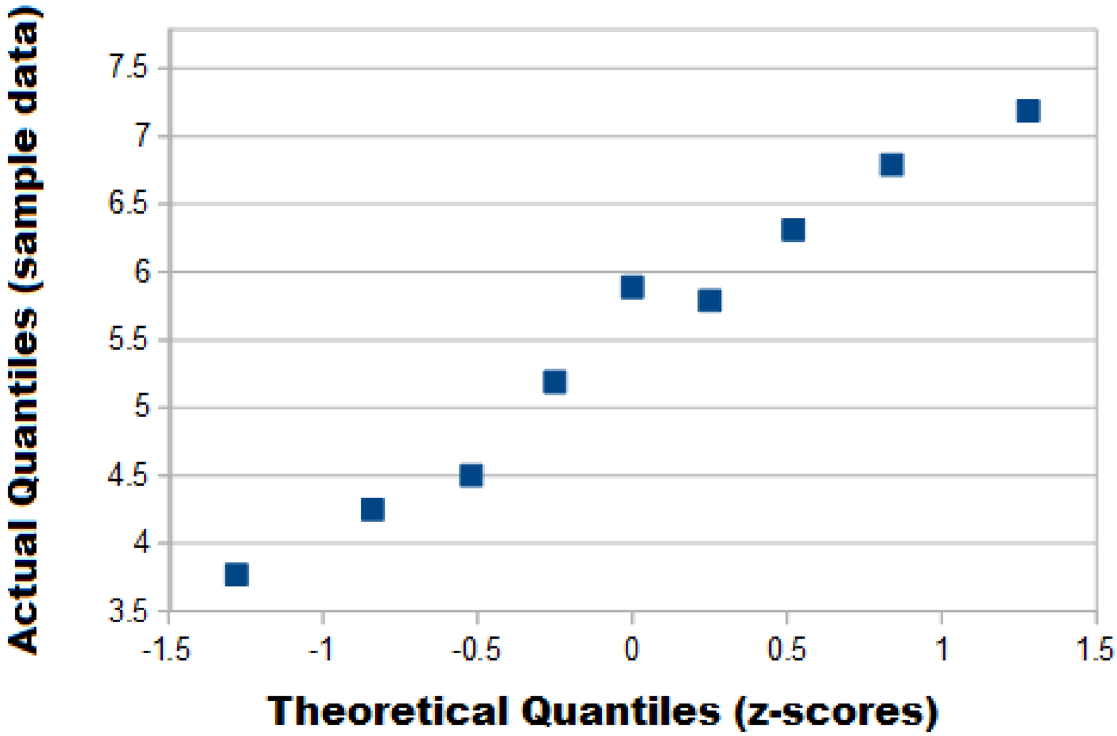
The z-values are:

- 10% = -1.28
- 20% = -0.84
- 30% = -0.52
- 40% = -0.25
- 50% = 0
- 60% = 0.25
- 70% = 0.52
- 80% = 0.84
- 90% = 1.28
- 100% = 3.0



A few of the z-values plotted on the graph.

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3). I used Open Office for this chart:



The (almost) straight line on this q q plot indicates the data is approximately normal.