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?***

The effect of categorical variables on the dependent variable depends upon the categorical variable correlation with target variable. The higher the correlation value towards -1 or 1, the stronger the effect.

For Example, from the dataset

The boxplot graph of *'weathersit'* shows that people tends to rent more bike in clear weather and 'season' show the fall is the top season to rent bike.

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

Simply put because one level of your categorical feature (here location) become the reference group during dummy encoding for regression and is redundant. A categorical variable of K categories, or levels, usually enters a regression as a sequence of K-1 dummy variables. This amounts to a linear hypothesis on the level means.

Note that if you using pandas.get_dummies, there is a parameter i.e. drop_first so that whether to get k-1 dummies out of k categorical levels by removing the first level. Please note default = False, meaning that the reference is not dropped and k dummies created out of k categorical levels!

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

'*temp*' is the numerical variable having the highest correlation with target variable

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

These Assumptions which when satisfied while building a linear regression model produces a best fit model for the given set of data.
There are 5 basic assumptions of Linear Regression Algorithm:

Linear Relationship between the features and target:
Can be validated by plotting a scatter plot between the features and the target.

Little or no Multicollinearity between the features:

Pair plots and heatmaps (correlation matrix) can be used for identifying highly correlated features.

Homoscedasticity Assumption:
A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity

Normal distribution of error terms:
Normal distribution of the residuals can be validated by plotting a q-q plot.

Little or No autocorrelation in the residuals:
Autocorrelation can be tested with the help of Durbin-Watson test. The null hypothesis of the test is that there is no serial correlation. The Durbin-Watson test statistics is defined as:

$$\sum_{t=2}^{T}((e_t - e_{t-1})^2) / \sum_{t=1}^{T} e_t^2$$

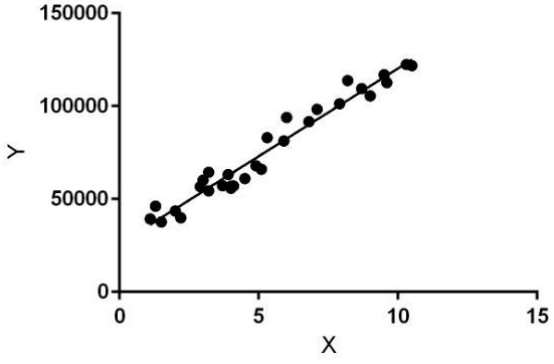
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. 'Light Snow' Weather
2. 'September' Month
3. 'Sunday' Weekday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.
Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2.x$$

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update θ1 and θ2 values to get the best fit line?

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

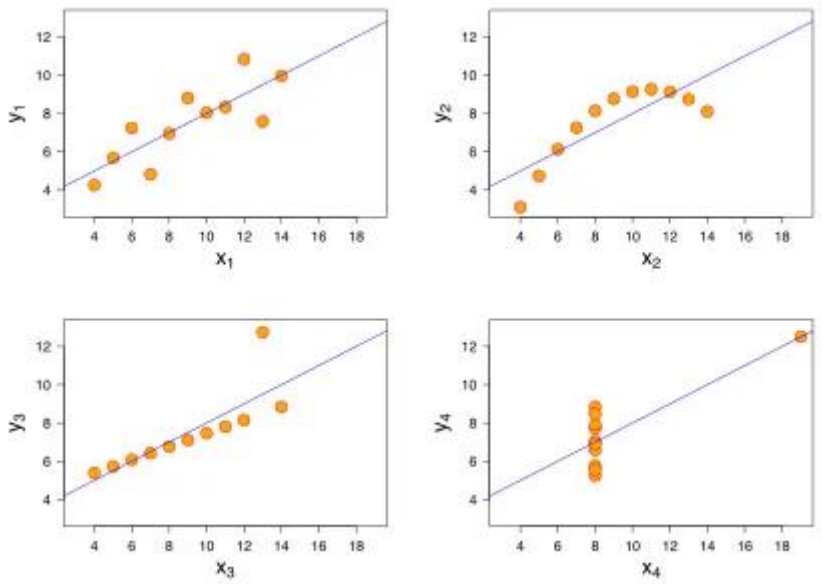
$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

Gradient Descent:
To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

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



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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analysing it and the effect of outliers on statistical properties.

For all four datasets:

Property	Value
Mean of x in each case:	9 (exact)
Variance of x in each case:	11 (exact)
Mean of y in each case:	7.50 (to 2 decimal places)
Variance of y in each case:	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case:	0.816 (to 3 decimal places)
Linear regression line in each case:	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse 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	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

A procedure to generate similar data sets with identical statistics and dissimilar graphics has since been developed.

Note: A computer should make both calculations and graph. Both sorts of output should be studied; each will contribute to understanding.

3. *What is Pearson's R?*

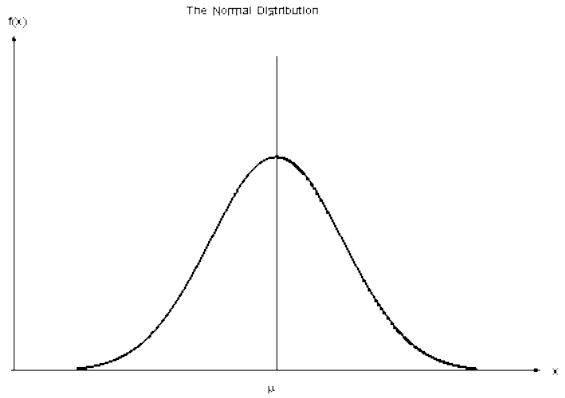
Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r

Questions a Pearson correlation answers

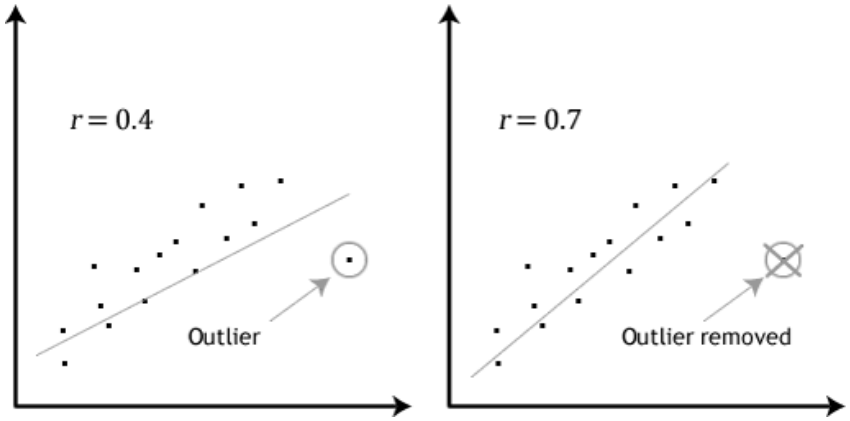
a. Is there a statistically significant relationship between age and height?
b. Is there a relationship between temperature and ice cream sales?
c. Is there a relationship among job satisfaction, productivity, and income?
d. Which two variables have the strongest co-relation between age, height, weight, size of family and family income?

Assumptions

1. For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.

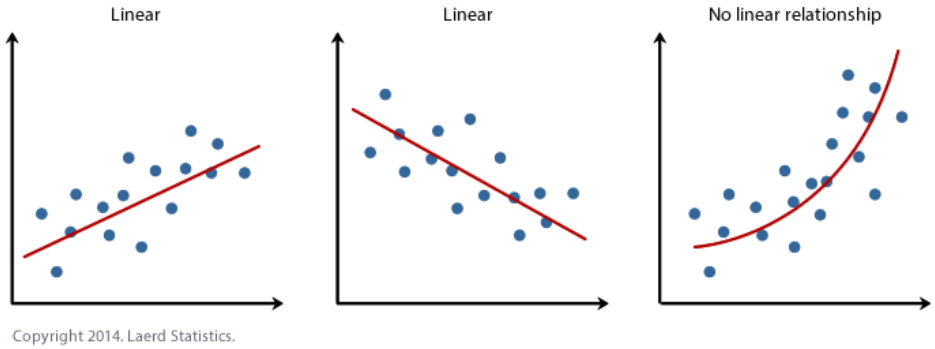


2. There should be no significant outliers. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.



3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

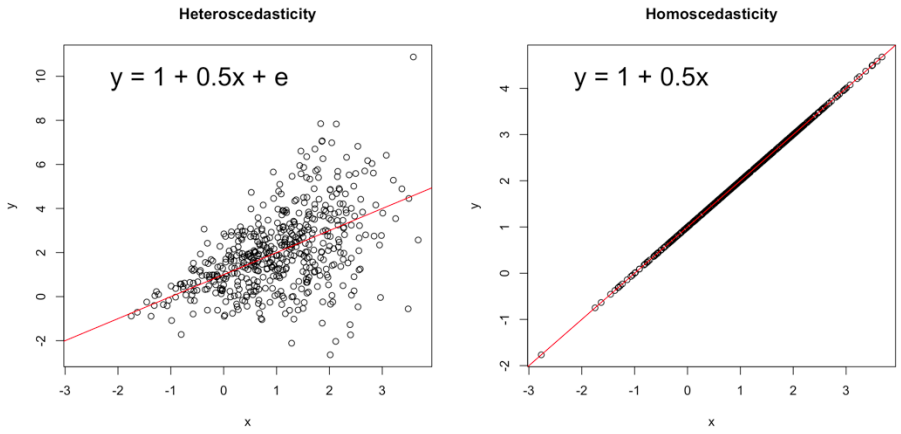
4. The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric.



5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example, if

you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

6. Homoscedascity. I've saved best for last. The hard is hard to pronounce but the concept is simple. Homoscedascity simply refers to 'equal variances. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.



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

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.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally.

Formally, Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally.

Formally, if a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

Examples of Algorithms where Feature Scaling matters
1. K-Means uses the Euclidean distance measure here feature scaling matters.
2. K-Nearest-Neighbours also require feature scaling.
3. Principal Component Analysis (PCA): Tries to get the feature with maximum variance, here too feature scaling is required.
4. Gradient Descent: Calculation speed increase as Theta calculation becomes faster after feature scaling. If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

Examples of Algorithms where Feature Scaling matters
1. K-Means uses the Euclidean distance measure here feature scaling matters.
2. K-Nearest-Neighbours also require feature scaling.
3. Principal Component Analysis (PCA): Tries to get the feature with maximum variance, here too feature scaling is required.
4. Gradient Descent: Calculation speed increase as Theta calculation becomes faster after feature scaling.

Normalization vs. standardization is an eternal question among machine learning newcomers. Let me elaborate on the answer in this section.

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

It is a good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+\cdots$

$〖VIF〗_1=1/ (1-R_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$

$〖VIF〗_2=1/ (1-R_2^2)$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:
One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.

The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.

The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

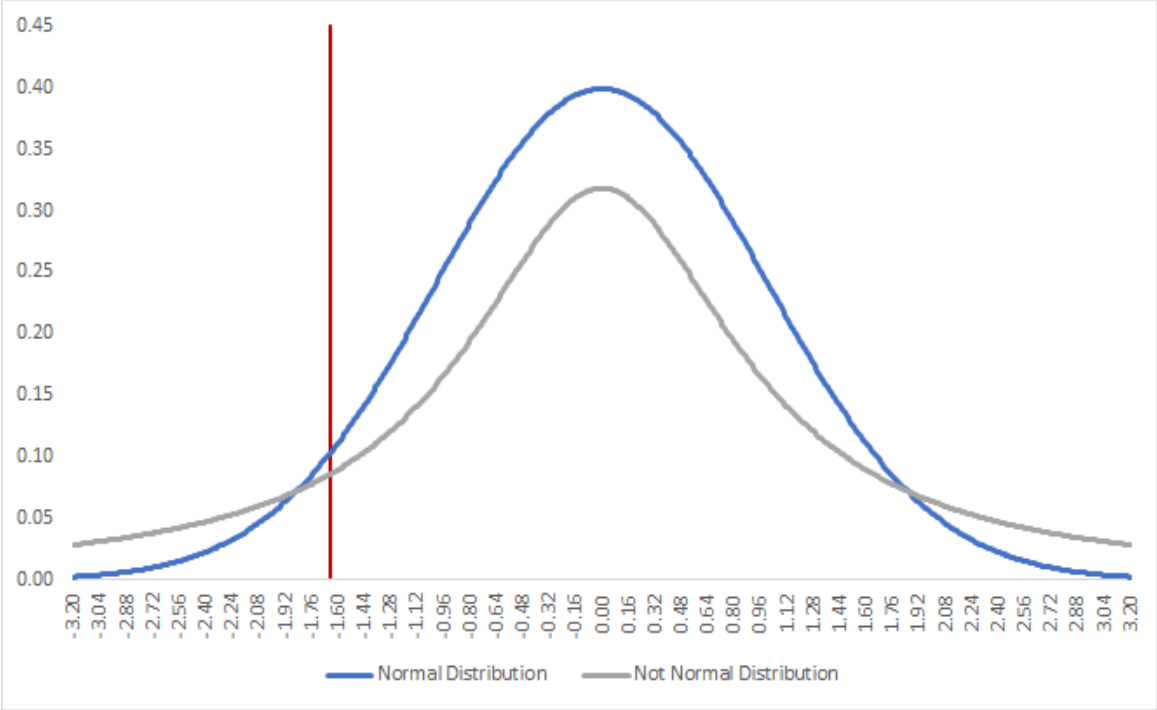
6. *What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.*

Things that are normally distributed are great. Knowing that something conforms to the normal distribution (and knowing its mean and standard deviation) allows us to make all kinds of useful inferences about it. For example, we can be reasonably sure where its value will fall say 95% of the time (between -1.96 and +1.96 standard deviations of the mean).
But if our variable is actually not normally distributed, then our inferences will be wrong, sometimes very wrong. And depending on the application, the consequences of our inaccurate inferences can range from being merely inconvenient to even dangerous.
That's where QQ plots come in. They're a quick and visual way to assess whether a variable is normal or not (we can use QQ plots to check our data against any distribution, not just the normal distribution).

The "QQ" in QQ plot means quantile-quantile — that is, the QQ plot compares the quantiles of our data against the quantiles of the desired distribution (defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).
Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets. For example, you've probably heard of percentiles before — percentiles are quantiles that divide our data into 100 buckets (that are ordered by value), with each bucket containing 1% of observations. Quartiles are quantiles that divide our data into 4 buckets (0–25%, 25–50%, 50–75%, 75–100%). Even our old friend, the median is a quantile — it divides our data into two buckets where half our observations are lower than the median and half our higher than it.

So, what does it mean to compare quantiles? Let's step back from QQ plots for a moment and think about a simpler way to compare 2 distributions, histograms. How would we figure out whether two distributions are the same? Well a decent first pass would be to overlay the distributions one on the other and stare really hard. But what should we be staring for? One simple test would be to pick a point on the X axis and see what proportion of each distribution lies to each side of it. For example, in finance we are often concerned with downside risk (the left tail of the distribution) — or in other words, what happens to our portfolio when things go bad.

Let's say we are concerned with really terrible events so we decide to look at outcomes that lie more than 1.65 standard deviations to the left of (in other words, below) the mean — we will call this point our threshold. If the distribution of our data were normal, then approximately 5% of our observations would lie to the left of our threshold:



Normal distribution (blue) and -1.65 SD threshold (red)

But what if our data were not normal? We can do the same analysis as above and see how many observations lie to the left of our threshold:

Normal and Not Normal Distribution visual comparison

Visually we can see that a lot more of the Not Normal distribution (the grey line — it's a Student's T-distribution with 1 degree of freedom) lies to the left of the threshold. So, if the distribution of our portfolio is actually the grey line, but we model it with the blue line, we will be significantly understating the frequency of a terrible outcome (terrible outcomes are ones to the left of our threshold, the red line). We would be assuming that there is only a 5% chance of a terrible outcome, when in reality 17% of the area under the grey line (its cumulative density function) lies to the left of our terrible outcome's threshold.

So, we would be understating the risk of a terrible outcome by a factor of 3!

That's why it's important to check that something is normal. And that's where QQ plots really shine. In essence, QQ plots do what we just did with our overlaid histograms (and threshold), but it does it for every observation in our data.