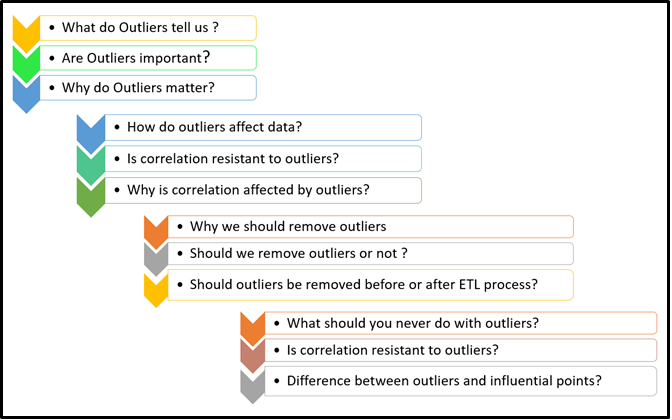
HANDLING OUTLIERS

In statistics, an outlier is an observation point that is distant from other observations. If they are the result of a mistake, then we can ignore them, but if it is just a variance in the data we would need think a bit further.

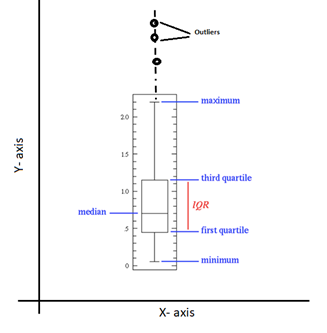


Generally, Outliers affect statistical results while doing the EDA process, we could say a quick example is the MEAN and MODE of a given set of data set, which will be misleading that the data values would be higher than they really are.

The CORRELATION COEFFICIENT is highly sensitive to outliers. Since it measures the strength of a linear relationship between two variables. the relationship dependent of the data. correlation is a non-resistant measure and r (correlation coefficient) is strongly affected by outliers.

BOX PLOT

In descriptive statistics, a box plot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.



If you have a categorical value then you can use that with any continuous variable and do multivariate outlier analysis.

SCATTER PLOT

A scatter plot , is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Z-SCORE

The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured.

The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution.

While calculating the Z-score we re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

|  |
| --- |
| from scipy import stats  import numpy as np  z = np.abs(stats.zscore(boston\_df))  print(z)  #Let’s try and define a threshold to identify an outlier.  threshold = 3  print(np.where(z > 3))  >>>https://miro.medium.com/max/875/1*cpxxI5Ny7DmRbnn6zWECfQ.png  #The first array contains the list of row numbers and second #array respective column numbers, which mean z[55][1] have a #Z-score higher than 3.  print(z[55][1])  >>>3.375038763517309 |

IQR score -

Box plot use the IQR method to display data and outliers(shape of the data) but in order to be get a list of identified outlier, we will need to use the mathematical formula and retrieve the outlier data.

The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

|  |
| --- |
| Q1 = boston\_df\_o1.quantile(0.25)  Q3 = boston\_df\_o1.quantile(0.75)  IQR = Q3 - Q1  print(IQR)  >>>IQR for each column  print(boston\_df\_o1 < (Q1 - 1.5 \* IQR)) |(boston\_df\_o1 > (Q3 + 1.5 \* IQR))  >>>Boolean T/F is returned |

Most difficult decision could be how one should deal with the outlier: Remove / Correct?

To summarize their explanation- bad data, wrong calculation, these can be identified as Outliers and should be dropped but at the same time you might want to correct them too, as they change the level of data i.e. mean which cause issues when you model your data. For ex- 5 people get salary of 10K, 20K, 30K, 40K and 50K and suddenly one of the person start getting salary of 100K. Consider this situation as, you are the employer, the new salary update might be seen as biased and you might need to increase other employee’s salary too, to keep the balance. So, there can be multiple reasons you want to understand and correct the outliers.

DO or DO NOT (Drop Outlier)

* Before dropping the Outliers, we must analyze the dataset with and without outliers and understand better the impact of the results.
* If you observed that it is obvious due to incorrectly entered or measured, certainly you can drop the outlier. No issues on that case.
* If you find that your assumptions are getting affected, you may drop the outlier straight away, provided that no changes in the results.
* If the outlier affects your assumptions and results. No questions simply drop the outlier and proceed with your further steps.

|  |
| --- |
| #For removing outlier using z score  boston\_df\_o = boston\_df\_o[(z < 3).all(axis=1)] |

|  |
| --- |
| #For removing outlier using IQR  boston\_df\_out = boston\_df\_o1[~((boston\_df\_o1 < (Q1 - 1.5 \* IQR)) |(boston\_df\_o1 > (Q3 + 1.5 \* IQR))).any(axis=1)] |