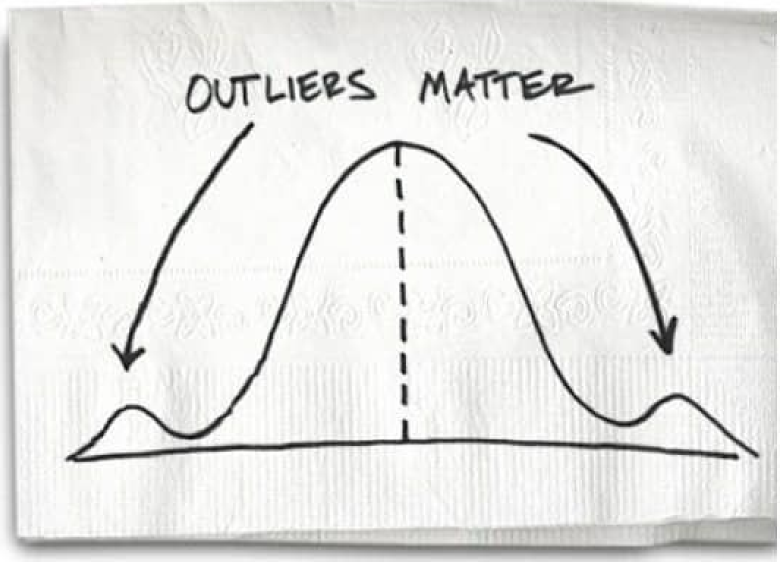
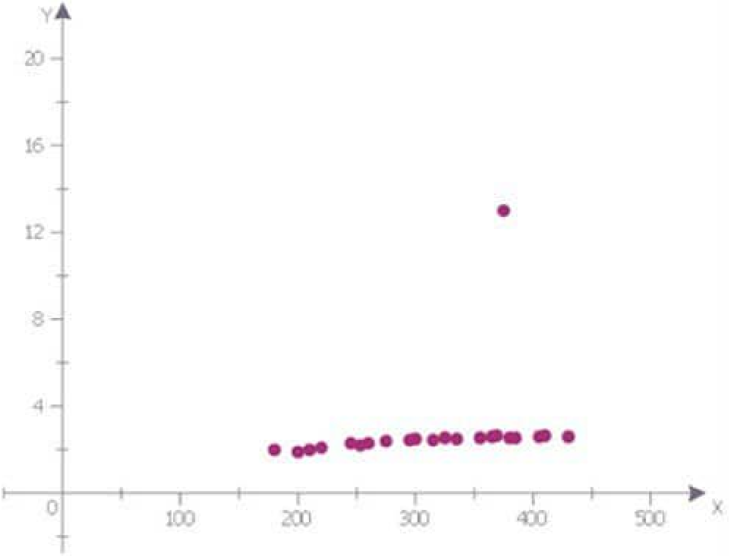
Outliers





- **Outliers are points which are like introverts who never mingle with others, who stay away from other points or group of points(distribution) like me.**
- **Outliers are extreme values that deviate from other observations on data, an outlier is an observation that diverges from an overall pattern on a sample.**
- **If you don't take care of the outliers you will feel sad with your results.**

Outliers

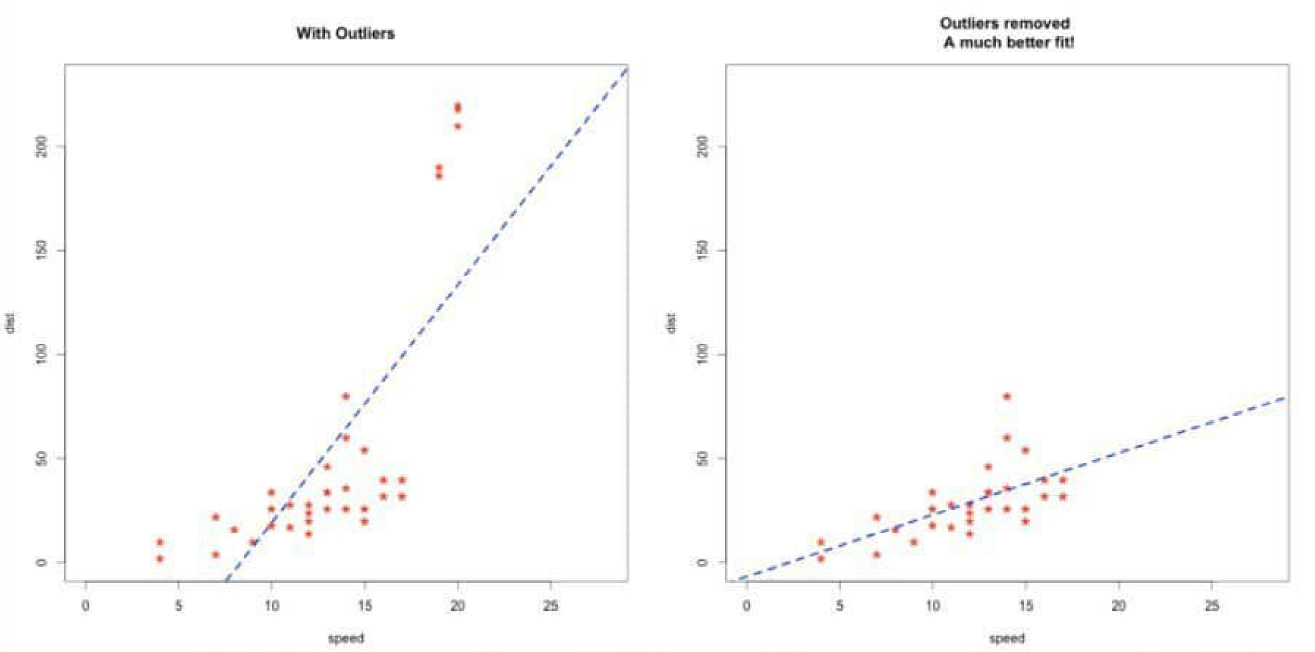
- There are two types of outliers, **univariate and multivariate.**
- **Univariate Outlier:** A univariate outlier is a data point that consists of extreme value on one variable.
- **Multivariate Outlier:** A multivariate outlier is a combination of unusual scores on at least two variables.
- Outliers can also come in different flavours depending on the environment: point outliers, contextual outliers, or collective outliers(We focus these mainly in Anomaly detection).
- **Point outliers** are single data points that lay far from the rest of the distribution.
- **Contextual outliers:** if it deviates significantly with respect to a specific context of the object(noise).
- **Collective outlier:** A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers.

Outliers

- So what are the causes of outliers?
- The best way to tackle outlier is to find out the reason why it occurred. The method to deal with them would then depend on the reason for their occurrence.
- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

Outliers

- **Why do we need to detect outliers?**
- **Outliers can impact the results of our analysis and statistical modelling in a drastic way.**
- **We also discussed the outlier impact in logistic regression, if you don't remember just go back and read again.**



outliers aren't always a bad thing. It's very important to understand this. Simply removing outliers from your data without considering how they'll impact the results is a recipe for disaster.

@learn.machinelearning

Outliers

> “Outliers are not necessarily a bad thing. These are just observations that are not following the same pattern as the other ones. But it can be the case that an outlier is very interesting. For example, if in a biological experiment, a rat is not dead whereas all others are, then it would be very interesting to understand why. This could lead to new scientific discoveries. So, it is important to detect outliers.
>
> – Pierre Lafaye de Micheaux, Author

@learn.machinelearning

Outliers

- how to detect outliers?
- There are many outlier detection methods divided into supervised methods, semi-supervised methods, and unsupervised methods.
- Various visualization methods, like Box-plot, Histogram and Scatter Plot can also be used.
- some thumb rules like
- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR(Interquartile range)
- Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as an outlier
- Data points, three or more standard deviation away from mean are considered outlier
- Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding.

Outliers

- There are many other algorithms like KNN, DBscan, Isolation forest, LOF etc..
  We will discuss more these algorithms when Their time comes.

Outliers

How to deal with outliers?

- **Deleting observations:** We delete outlier values if it is due to a data entry error, data processing error or outlier observations are very small in numbers. When there is no importance for these points in the model building can also be deleted.
- **Transforming and binning values:** Transforming variables can also eliminate outliers. Natural log of value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows dealing with outliers well due to binning of a variable. We can also use the process of assigning weights to different observations and you can also use a sigmoid function to squash values.

Outliers

How to deal with outliers?

- **Imputing:** We can also impute outliers. We can use the mean, median, mode imputation methods. Before imputing values, we should analyse if it is a natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use a statistical model to predict values of outlier observation and after that, we can impute it with predicted values.
- **Treat Outliers separately:** If there are a significant number of outliers, we should treat them separately in the statistical model. One of the approaches is to treat both groups as two different groups and build an individual model for both groups and then combine the output.