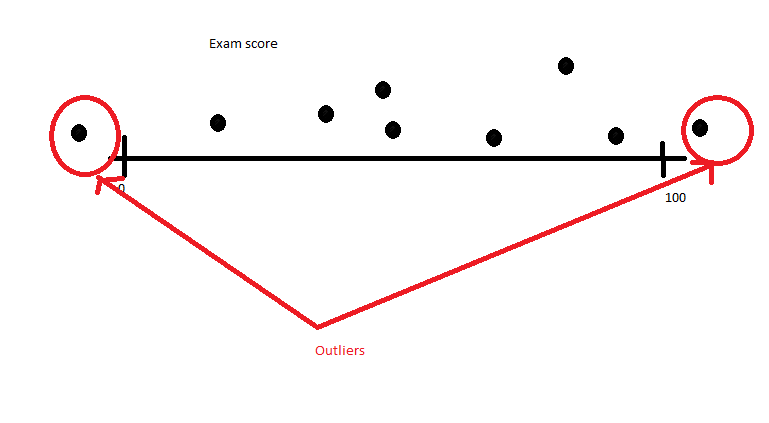
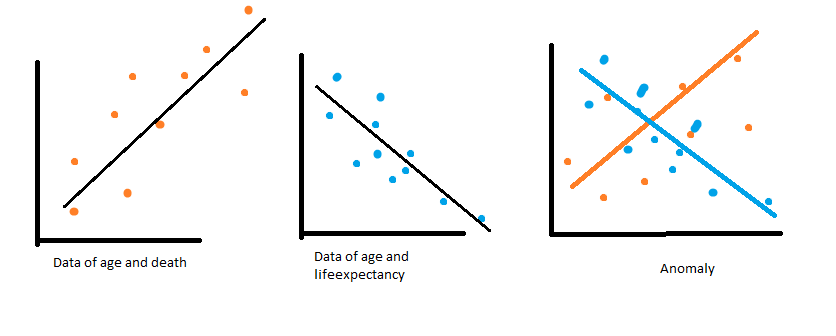
1. **Difference between outlier and Anomaly**

Outlier is an extreme data point. If you consider the exam score of students in a class where assume that the exam for English is out of hundred and children could only get scores between zero to hundred. As you can see in the below picture there are two points marked in red has distributed beyond the limit which is ‘zero to hundred’. As mentioned above these are not possible outcomes of exam results and, they can be considered as extreme points or outliers in the dataset with respect to the observation setup we have taken.



*figure 1.1.*

Anomalies are the points taken by a separate process. Another example as given in the following figure 1.2.

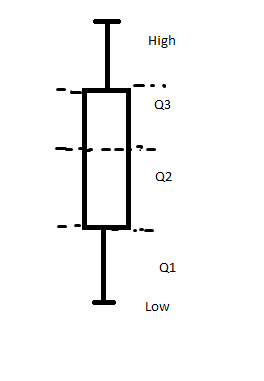


*figure 1.2.*

In the first graph, it shows the as the age increases number of death in each state in a country also increases. On the other hand, second graph shows as age increases the life expectancy of the population in the same state decreases. If both data have been mixed up for some reason and unable to distinguish specific features this situation is called an Anomaly. This means the two different data created an anomalous pattern to the dataset making anomaly. So in conclusion we can say that outlier is based on the range of possible values of a variable meantime anomaly arises when anomalous patterns are seen in the data.

1. **What is relation between IQR and Boxplot**

In exploratory data analysis, a boxplot is often used for descriptive statistics. For finding outliers one of the best visualisation methods used is called Box Plot sometimes also called Whisker Plots. In this method, we can clearly see the outliers that lie beyond the low and high limits of the whisker plot. Univariate analysis is the most efficient way of determining outliers in a dataset. IQR plots can visualize the first quartile, second quartile and third quartile of the data distribution making it easy for outlier visualization.



*figure 2.2.*

Using these quartile points is used Q3 is the 75th Percentile of the data. Q2 is the 50th percentile of the data. Also, Q1 is the 25th percentile of the data. Similarly using this IQR is found by taking the difference between Q3 and Q1. In IQR it is always assumed that data is distributed normally.

IQR = Q3-Q1

Low = Q1 – 1.5 \* IQR

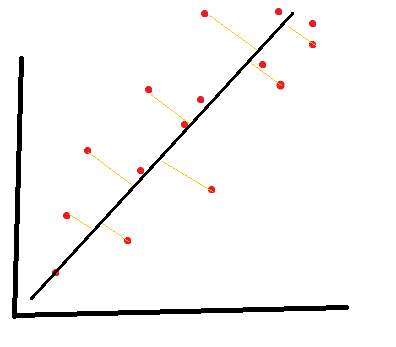
High = Q3 – 1.5 \* IQR

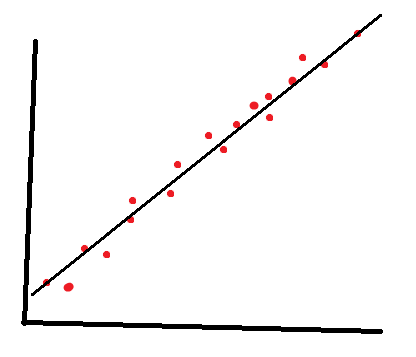
So any point that comes below low or above high as per the above statistical equation will be considered as an outlier or extreme point.

1. **DB Scan algorithm implementation**

(Answer attached as ipynb file with the submission)

1. **Can a noisy Dataset mislead the model generating Algorithm to generate an Inaccurate model? Explain How?**

Noise is really for the data analysis. Noise gives generality to the dataset. Variance is required in the data which could give knowledge in the dataset.

Fig 4.1 fig 4.2

Some points that are not following the exact trend of the model are called noise. Noise is different from outliers. Outliers are not good for the model. A noisy dataset will never make inaccurate models in fact noise is only adding generality to the model. Noise can be considered as knowledge for the model. Also, it makes the model efficient by being the knowledge of the normal data points. Noise points are part of the dataset but it does not follow the perfect trend of the model. If you look at figure 4.1 data points are distributed along the trend of the distribution which does not contain any noise. But in Fig 4.2 you can see that there is enough variance between the trend and points that are not following the perfect trend these points are called as noise in the data. If the number of points is way from the mean those points will be having noisy behavior. In the above two examples, the orange line represents the variance of datapoints that are away from the trend of the model(does not follow the perfect trend of the model). Noisy points are away from the mean and having some variance associ ated with it making it noisy point.

Outliers are points that are not acceptable as a possible outcome for example marks of students shows more than 100 in an exam, are not possible or acceptable values with respect to the outcome of the experience. Unlike outliers, noise points can be possible values but have more variance with respect to the trend of the model. An anomaly represents a pattern in the data that does not follow the expected behavior.

Like cocktail party problem, if you are in a cocktail party trying to record a person voice, you could also get some other signals like music or other people speaking in the background. The main signal will be the voice of the person you are targeting to record, but the other noises you also get along with the person’s voise is called as noise because this signal it does not follow the perfect trend of the main signal.

If the point in the model is following the model following perfect trend there would not be any variability in the data and it will not be able to interpret any general knowledge in the data. Whenever you increase the variance, high there will also be too much knowledge in the model which could also make some confusion to the model. So Large number of points with a huge variance will also lead to confusion to the model. If the model is getting less knowledge also it would not be enough knowledge for the model to make good results.