INTERVIEW DAY - EDA (04) Outlier Detection & treatment

Q) What is outlier and give an example?

Ans - Outlier are the data points that differs significantly from other observations.

Example → Suppose the Age of the Customer is 110, then he/she is an outlier from the average/normal population.

Q) What are the outlier detection techniques & treatment techniques?

Ans - i) Outlier detection techniques - Percentile, Box plot, Z score

ii) Remove outlier techniques - Capping based on upper and Lower range.

Q) What do you mean by by percentile and box plot?

Ans - i) Percentile → Nth percentile of an observation variable is the value of first N elements of the data values when it is sorted in ascending order.

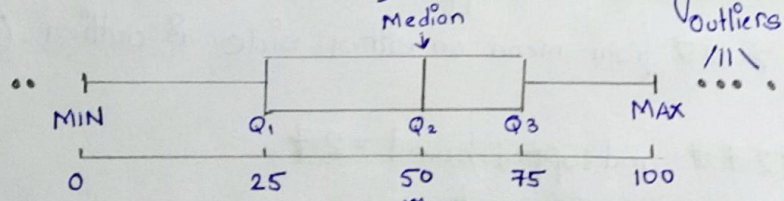
Example - 50% percentile, till 50% percentile what is the number.

Consider a list → 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

50 percentile will be 500.

ii) Boxplot → It measures how far apart the entire data is in terms of values.

Graphical representation of i) Three quantiles (First, Second, Third)
ii) Smallest and largest value.



Example → -10, 1, 2 | 3, 4, 5 | 6, 7, 8 | 9, 10, 20

Q1 (2.5) Q2 (5.5) Q3 (8.5) Percentile

IQR = Q3 - Q1
= 8.5 - 2.5
= 6

Finding Outliers, $Q_1 - 1.5(IQR)$ & $Q_3 + 1.5(IQR)$

$= 2.5 - 1.5(6)$ $= 8.5 + 1.5(6)$

$= 2.5 - 9 = -6.5$ $= 8.5 + 9 = 17.5$.

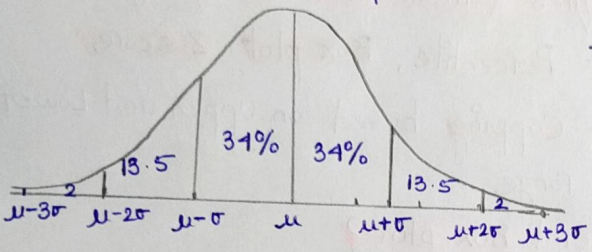
So any data outsider $[-6.5, 17.5]$ are outlier. So, -10 and 20 are outliers.

Therefore, minimum → -6.5, maximum → 17.5, $Q_1 \rightarrow 2.5$, $Q_2 \rightarrow 5.5$, $Q_3 \rightarrow 8.5$

outliers → -10, 20.

9) Why 1.5 in IQR method of Outlier Detection? why not 1 and 2?

Ans - 1.5 because it clearly controls the sensitivity of the range.
- A bigger scale would make the outlier(s) to be considered as data point(s) while a smaller one would make some of the data point(s) to be preceived as outliers.
- Lets say data follows normal distribution.



→ With one standard deviation of mean, μ+σ cover 68% data (68.26%)
→ μ+2σ cover 95% data (95.44%)
→ μ+3σ cover 99% data (99.72%)
Rest 0.28% of whole data lies outside μ+3σ and this part consider as outliers.

- First and third quartiles, Q1 and Q3 lies at -0.675σ and +0.675σ

Lets calculate IQR decision in terms of σ.

Scale = 1 :

Lower bound = $Q_1 - 1 * IQR = Q_1 - 1(Q_3 - Q_1) = -0.675\sigma - 1*(0.675-(-0.675))\sigma = -2.025\sigma$

Upper bound = $Q_3 + 1 * IQR = Q_3 + 1(Q_3 - Q_1) = 2.025\sigma$

So when scale = 1, any data beyond 2.025σ from mean on either side consider to be outlier. (small range)
But as we know, upto 3σ data is useful. So we need to increase scale.

Scale = 2.
Lower bound = -3.375σ and upper bound = 3.375σ.
Data lies beyond 3.375 from mean on either side is outlier. (Big range)

So scale = 1.5.
Lower bound = -2.7σ and upper bound = 2.7.
2.7σ is near to 3σ but not exact.
To get exact 3σ, scale = 1.7 but then 1.5 is more symmetrical & approx

9) How through z score, we can find outliers?

Ans - Z score finds the distribution of Normal Data.
- In Normal Distribution, Mean is 0 and Standard deviation is 1.
- In Z score, we will rescale the data to the center and check for the data which are too far from center will be treated as outlier.
- In most cases we take the value upto 3 (3 SD of mean, 99.7% of values within) so z score which are more than 3 will be treated as Outliers.