TE Course - DSBDA
Outlier handling

Sonal Gore, PCCOE, Pune

Outlier

- Deviates significantly from normal data points
- Example: set of random numbers: 1, 99, 100, 101, 103, 109, 110, 201

1 and 201 are outliers.

- Outliers can be detected using visualization, implementing mathematical formulas on the dataset, or using the statistical approach
- Outliers aren’t always that obvious: for example - paychecks last month as $225, $250, $25, $235.

Visualization – using box plot / scatterplot

- Box plot for univariate outliers
  - captures the summary of the data effectively and efficiently with only a simple box and whiskers.
  - Boxplot summarizes sample data using 25th, 50th, and 75th percentiles.
  - some boxplots may not show outliers.
  - box and whiskers charts can be a useful tool to display outliers after calculated what your outliers actually are
- Scatterplot for multivariate outliers

Outlier detection using IQR

- most effective way to find outliers is by using interquartile range (IQR). The IQR contains the middle bulk of your data
- An outlier is then a data point x_i that lies outside the interquartile range. That is:

$$\boxed{\begin{array}{c} x_i > Q3 + k(IQR) \bigvee x_i < Q1 - k(IQR), \\ \text{where } IQR = Q3 - Q1 \text{ and } k \geq 0\,. \end{array}}$$

- Using the interquartile multiplier value k=1.5, the range limits are the typical upper and lower whiskers of a box plot.
- An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.
  High = (Q3) + 1.5 IQR
  Low = (Q1) – 1.5 IQR
- More suitable for skewed distributions
- Find the outliers for the following data set: 3, 10, 14, 22, 19, 29, 70, 49, 36, 32.

Outlier detection - Using z-score

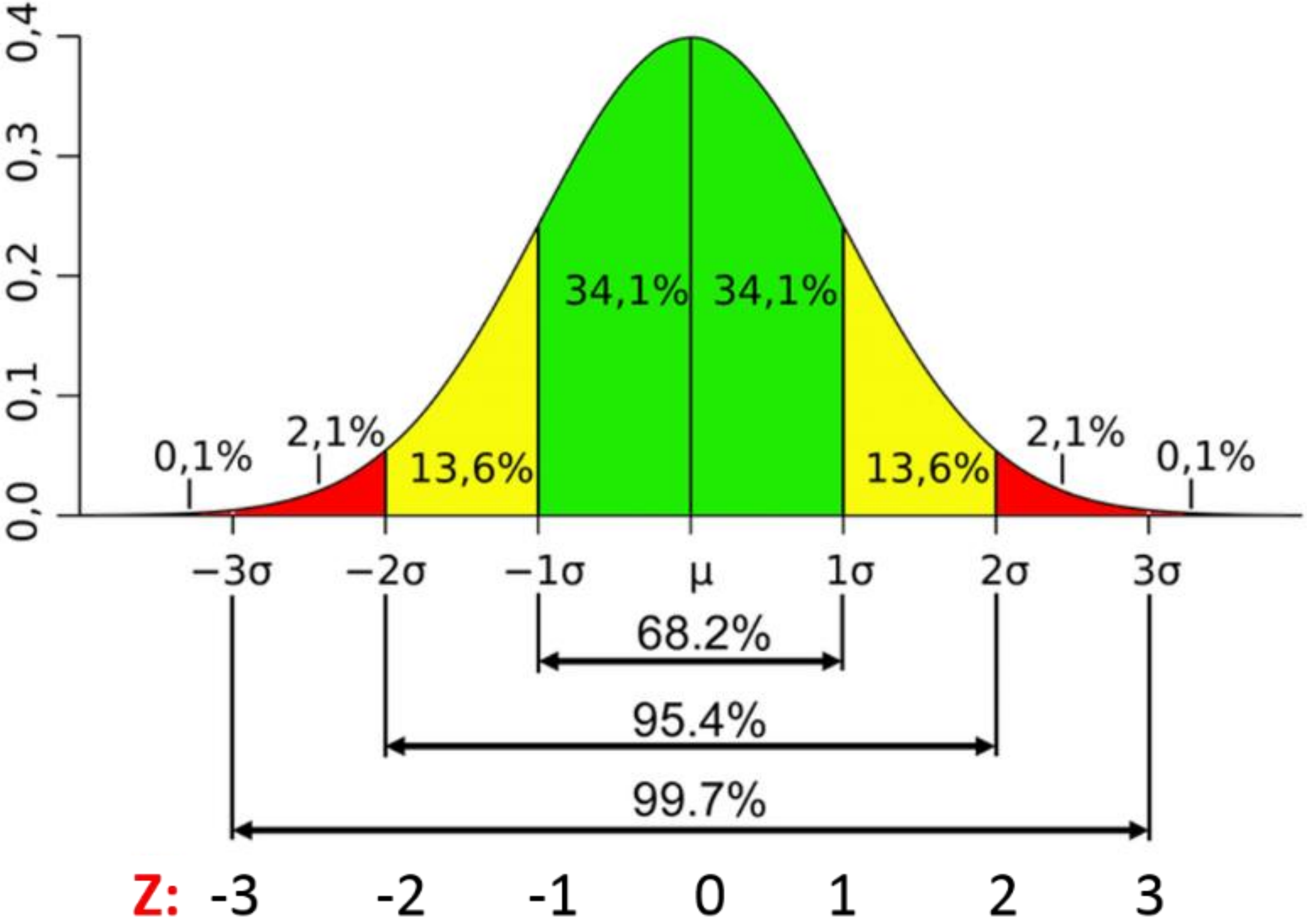
- Z-score is a parametric outlier detection method in a one dimensional feature space
- assumes a Gaussian distribution of data (z-score – standard score)
- outliers are data points that are in the tails of distribution and therefore far from mean
- How far depends on a set threshold z_{thr} for the normalized data points z_i calculated with the formula:

$$z_i = \frac{x_i - \mu}{\sigma} ,$$

- where x_i is a data point, μ is the mean of all x_i and is the standard deviation of all x_i.
- An outlier is then a normalized data point which has an absolute value greater than z_{thr}. That is:

$$|z_i| > z_{thr}$$

- Commonly used z_{thr} values are 2.5, 3.0 and 3.5.


0,4
0,3
0,2
0,1
0,0
34,1%
34,1%
13,6%
13,6%
2,1%
2,1%
0,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ
68.2%
95.4%
99.7%
Z: -3 -2 -1 0 1 2 3

References

- https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/
- https://www.kdnuggets.com/2018/12/four-techniques-outlier-detection.html
- https://www.statisticshowto.com/statistics-basics/find-outliers/
- https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/