FORMERLY Industries Helping Hands

THE iSCALE

Data Science | 30 Days of Machine Learning | Day - 20

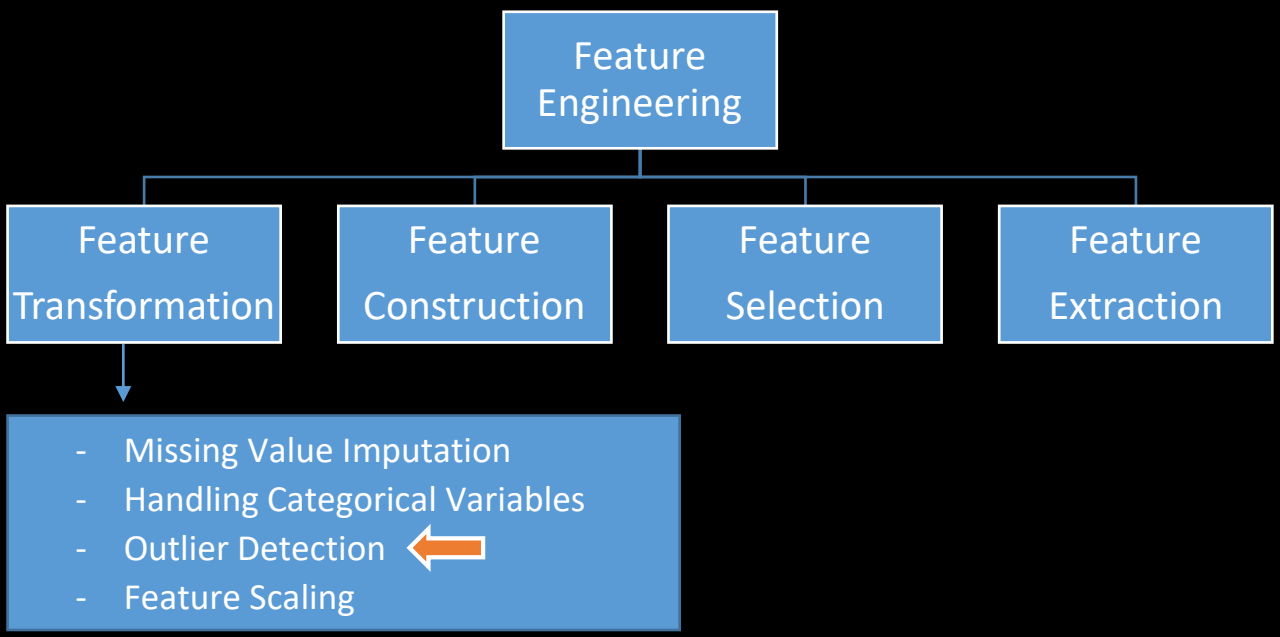
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 20----

Outliers: Percentile method technique

- **What is a percentile method?**
- **Outliers example with percentile**
- **Trimming with percentile outlier's method**
- **What is the Winsorization in outliers?**
- **Capping with percentile outlier's method**
- **What is the difference between trimming and winsorizing outliers?**

- Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



Techniques to detect & remove outliers: -

Z-score treatment: - This technique assumes that the column follows a normal distribution.

IQR (Interquartile Range) based filtering: - The IQR method involves calculating the range between the first quartile (Q1) and the third quartile (Q3).

Percentile method: - In this approach, a threshold is set based on percentiles. For example, if the threshold is set at 5%, any data point above the 95th percentile or below the 5th percentile is considered an outlier. These outliers can be removed or handled accordingly.

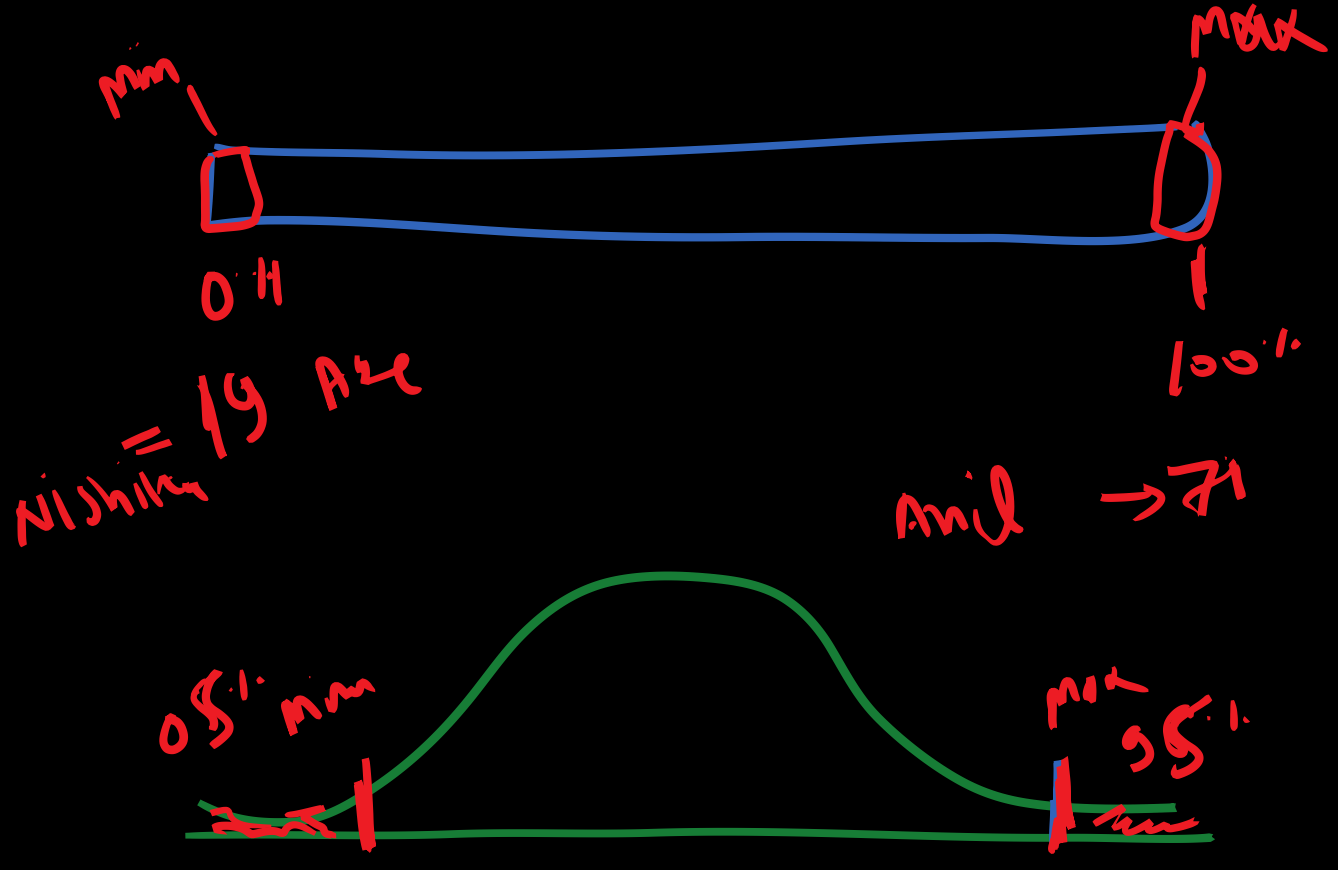
Winsorization:- Winsorization involves replacing outliers with values at a certain percentile, rather than removing them completely.

What is a percentile method?

Percentile method: - In this approach, a threshold is set based on percentiles. For example, if the threshold is set at 5%, any data point above the 95th percentile or below the 5th percentile is considered an outlier. These outliers can be removed or handled accordingly

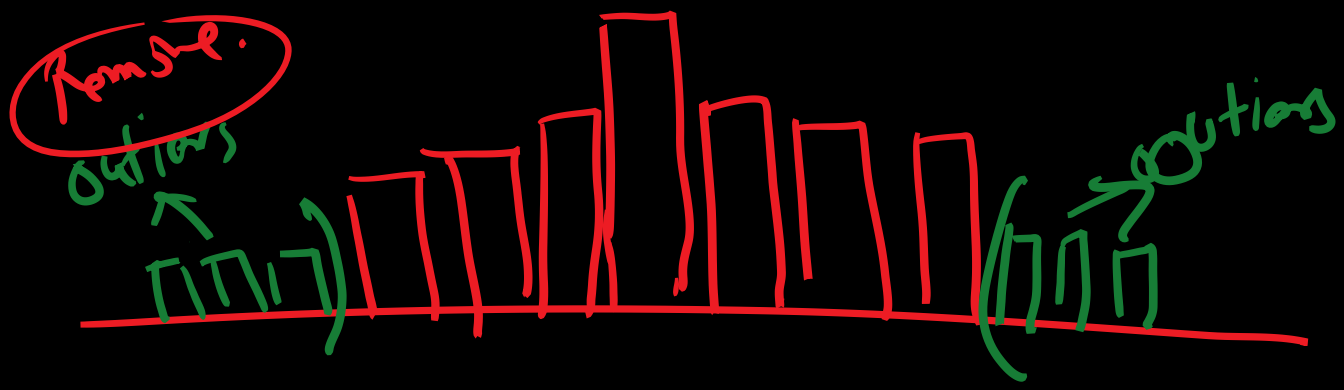
- Example Outliers:

S.No	Name	Age
1	**Nishant**	**30**
2	**Priyesh**	**28**
3	**Nishika**	**19**
4	**Anil**	**71**
5	**Avantika**	**29**
6	**Ragini**	**37**



- Trimming with percentile outlier's method

Trimming excludes the outlier values from our analysis. By applying this technique, our data becomes thin when more outliers are present in the dataset. Its main advantage is its fastest nature.



- What is the Winsorization in outliers?

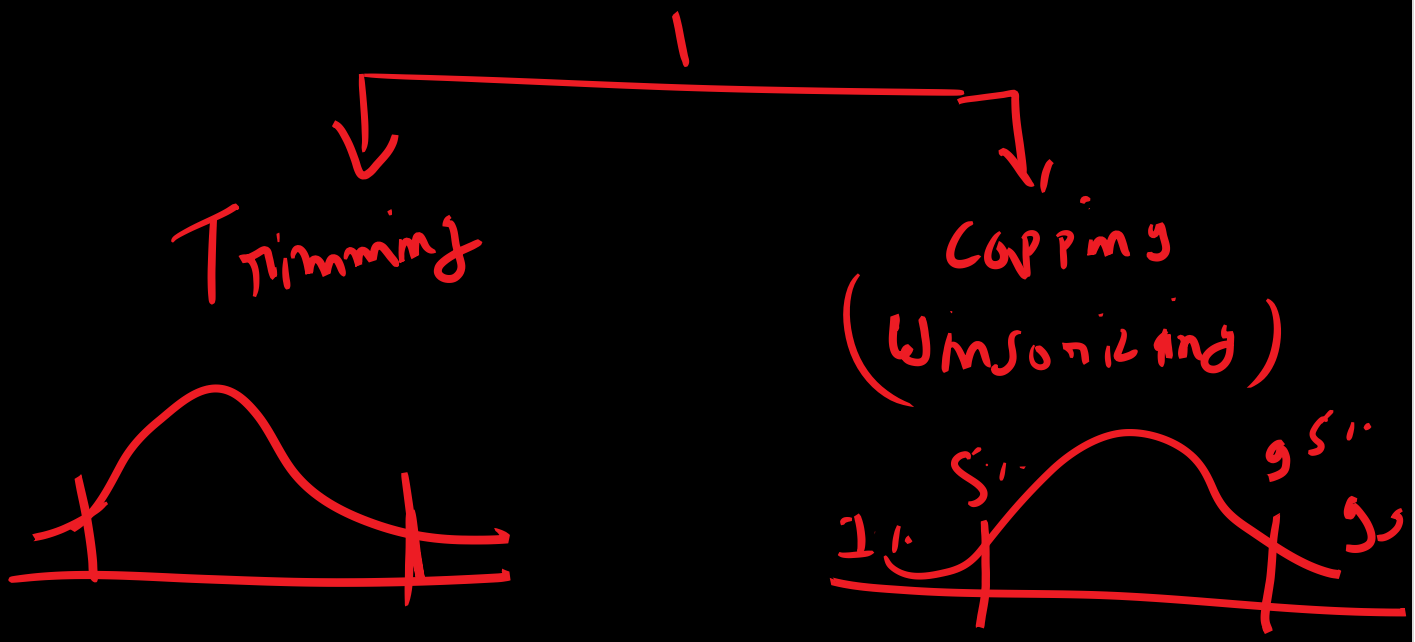
Winsorizing is an alternative to capping. The process of replacing the extreme values of statistical data in order to limit the effect of the outliers on the calculations or the results obtained by using that data. The mean value calculated after such replacement of the extreme values is called winsorized mean.

What is the difference between trimming and winsorizing outliers?

Winsorizing data means to replace the extreme values of a data set with a certain percentile value from each end, while Trimming or Truncating involves removing those extreme values.

https://ndgigliotti.medium.com/trimming-vs-winsorizing-outliers-e5cae0bf22cb

Two standard approaches are trimming and Winsorizing. Trimming amounts to simply removing the outliers from the dataset. Winsorizing, on the other hand, amounts to changing the value of each outlier to that of the nearest inlier.



<Start Coding>

```
#Import Library

import numpy as np
import pandas as pd

#Import Dataset

df = pd.read_csv('weight-height.csv')

----
df.head(

----
df.shape

#Describe Height

df['Height'].describe()


#Import Seaborn and Plot

import seaborn as sns
----
sns.distplot(df['Height'])
----
sns.boxplot(df['Height'])
```

```
#Define Upper Limit
upper_limit = df['Height'].quantile(0.99)
upper_limit

#Define Lower Limit
lower_limit = df['Height'].quantile(0.01)
lower_limit

#Set range : If Criteria is not full fill
df[(df['Height'] >= 74.78) | (df['Height'] <=
58.13)]

#Define New Data Frame With Limit
(Trimming through percentile method)
new_df = df[(df['Height'] <= 74.78) & (df['Height']
>= 58.13)]
----
new_df['Height'].describe()
----
df['Height'].describe()

#Plot Draw After Trimming
sns.distplot(new_df['Height'])
----
sns.boxplot(new_df['Height'])
```

```
#Capping With Winsorization
df['Height'] = np.where(df['Height'] >= upper_limit,
        upper_limit,
        np.where(df['Height'] <= lower_limit,
        lower_limit,
        df['Height']))


----
df.shape
----
df['Height'].describe()
----
sns.distplot(df['Height'])
----
sns.boxplot(df['Height'])
```