# Tasks:

Many real-world datasets exhibit outliers or extreme values. Discuss the potential impact of outliers on measures of central tendency and variability, and propose at least two methods for identifying and handling outliers in data analysis. How can Python be used to implement these methods? Provide examples from real-world datasets to illustrate your points.

### What is an outlier?

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

### What are the criteria to identify an outlier?

1. Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile
2. Data point that falls outside of 3 standard deviations. we can use a z score and If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.Typically, Z-score greater than 3 is considered extreme.

### What is the reason for an outlier to exists in a dataset?

1. Variability in the data
2. An experimental measurement error

### What are the impacts of having outliers in a dataset?

1. It causes various problems during our statistical analysis
2. It may cause a significant impact on the mean and the standard deviation

### Various ways of finding the outlier.

1. Visual Methods:

- Using scatter plots
- Box plot

1. Calculation:

- using z score
- using the IQR interquantile range

Comment: On my graph after the Z score and IQR, we notice that there are still some outlier points, which is due to the fact that the distance of the outliers was very high from the other points.

**<u>Datasets:</u>** https://www.kaggle.com/datasets/krishnaraj30/finance-loan-approval-prediction-data

## References:

Goyal, C. (2021). Why you shouldn’t just delete outliers. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/

Codebasic.(2020). Outlier detection and removal: z score, standard deviation | Feature engineering tutorial python. https://www.youtube.com/watch?v=KFuEAGR3HS4

Coursera. https://www.coursera.org/learn/foundations-of-data-science/lecture/2vmAG/welcome-to-module-2

Machinelearningplus. How to detect outliers with z-score. https://www.machinelearningplus.com/machine-learning/how-to-detect-outliers-with-z-score/