Missing Values & Outliers

EMPOWERING High Performance Technology Teams

AGENDA

- Business Example
- Missing values: Detection and Treatment
- Outlier vs Anomaly
- Outlier Detection and Treatment

BUSINESS APPLICATION

Credit risk assessment is critical for loan approval and risk management of banks. However, the problem of missing credit risk data may greatly reduce the effectiveness of the assessment model.

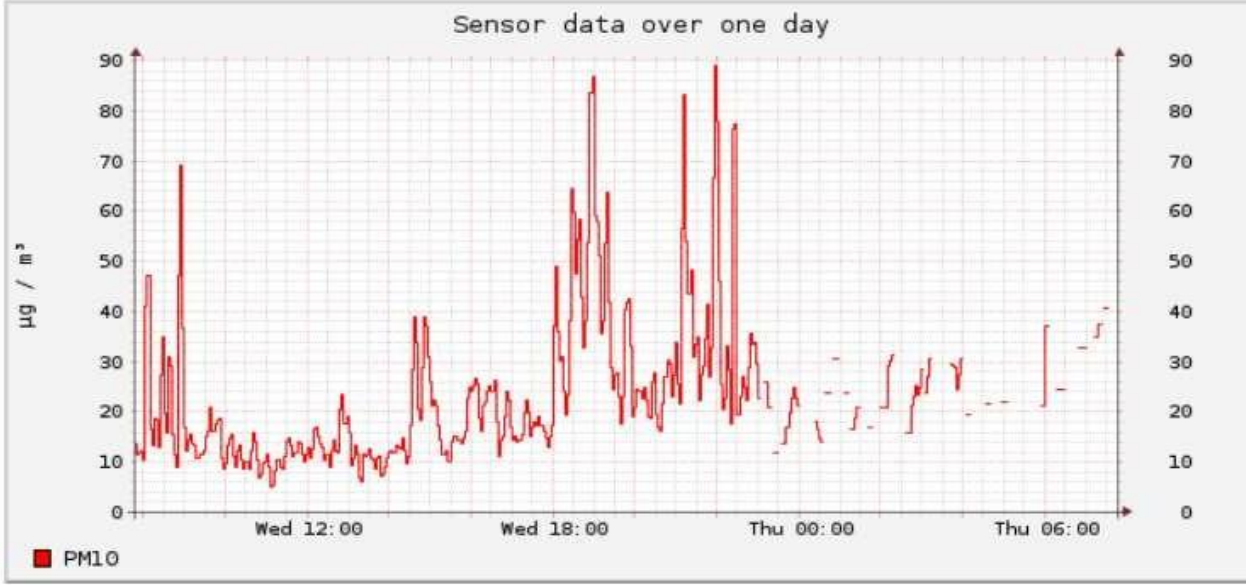
The credit card industry has utilized the concept of outliers in detecting credit card fraud in the past and now. Outliers may be considered as gems which bring to light activities and information that have significant implications on a subject matter or an organization.

MISSING VALUES: DETECTION

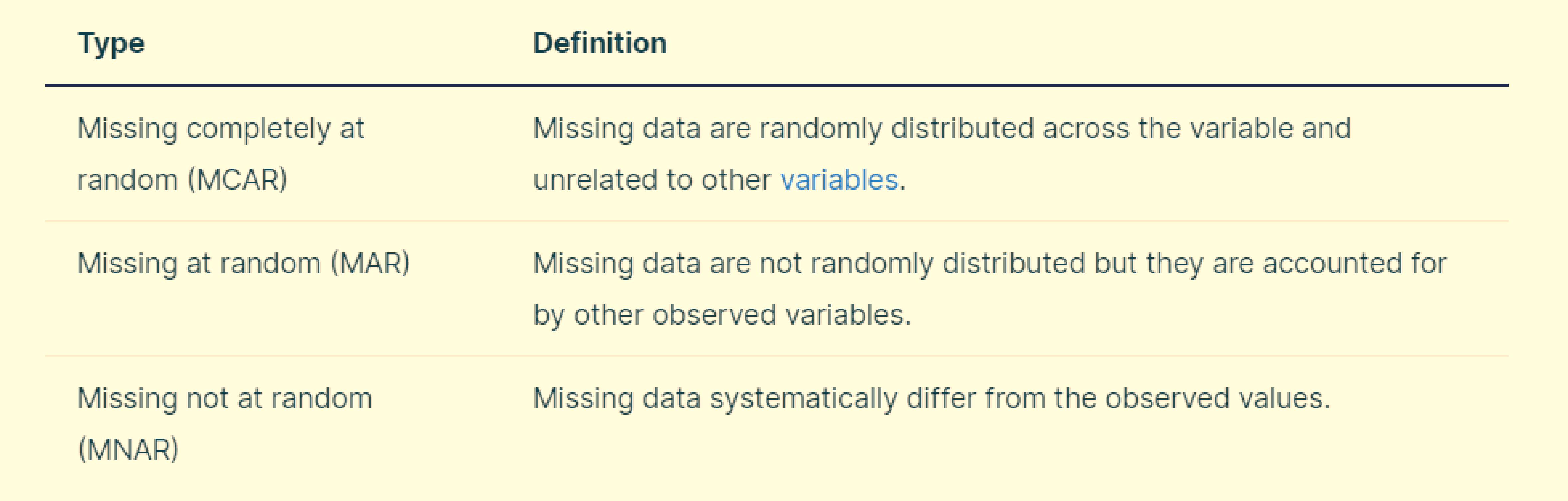
- Missing values are common when working with real-world datasets

- Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons

- Missing data may be problematic since they can sometimes cause sampling bias; your results may not be generalizable outside of your study because your data come from an unrepresentative sample



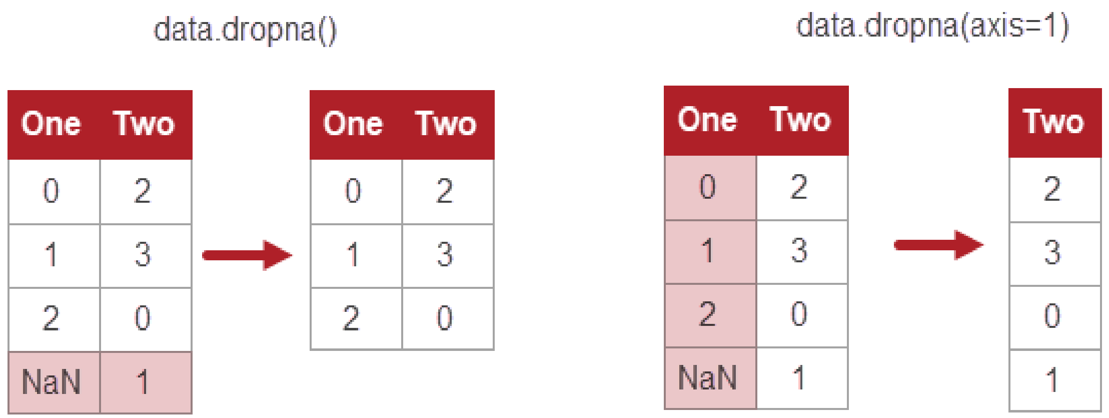
MISSING VALUES: TYPES



Type	Definition
Missing completely at random (MCAR)	Missing data are randomly distributed across the variable and unrelated to other variables.
Missing at random (MAR)	Missing data are not randomly distributed but they are accounted for by other observed variables.
Missing not at random (MNAR)	Missing data systematically differ from the observed values.

TREAT MISSING VALUES

- Drop the Observation:
  - In statistics, this method is called the listwise deletion technique. In this solution, we drop the entire observation if it contains a missing value.
  - Only if we are sure that the missing data is not informative, we perform this. Otherwise, we should consider other solutions



data.dropna()

One	Two
0	2
1	3
2	0
NaN	1

One	Two
0	2
1	3
2	0

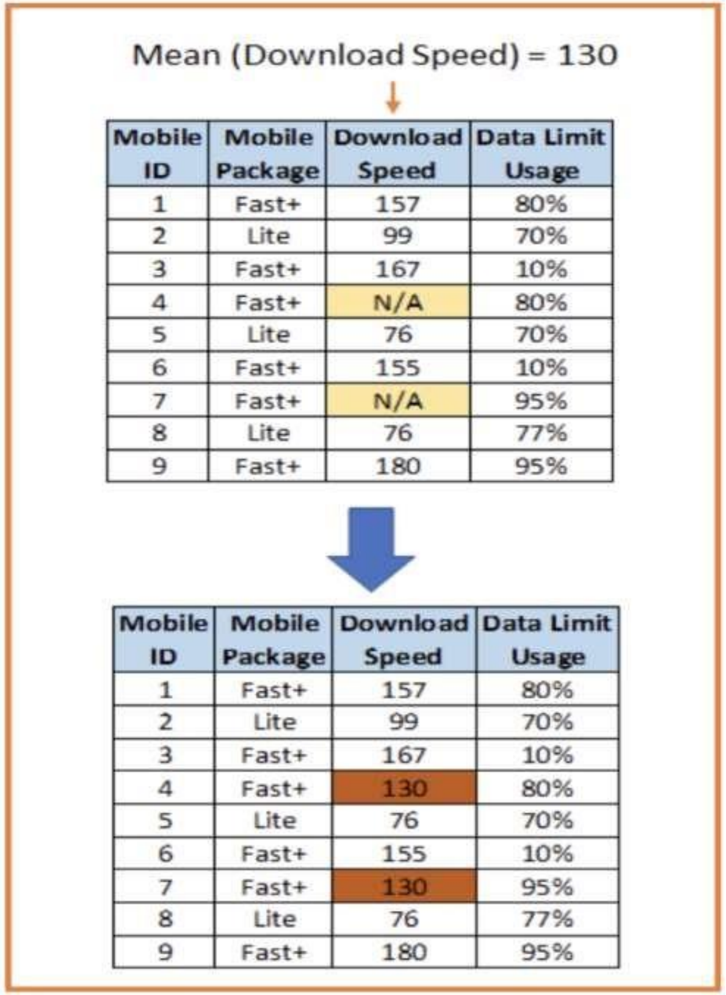
data.dropna(axis=1)

One	Two
0	2
1	3
2	0
NaN	1

Two
2
3
0
1

TREAT MISSING VALUES

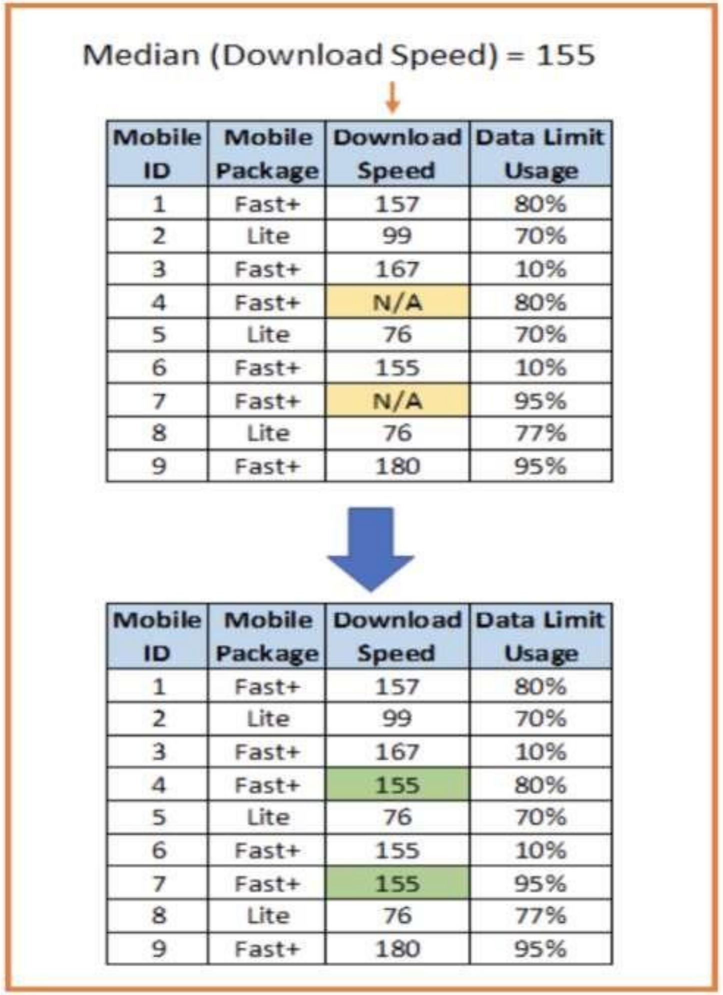
- Impute the missing values:
  - When the feature is a numeric variable, we can conduct missing data imputation
  - We replace the missing values with the average or median value of the same feature



Mean (Download Speed) = 130

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%



Median (Download Speed) = 155

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%

TREAT MISSING VALUES

method	**description**
dropna()	Drop missing observations
dropna(how='all')	Drop observations where all cells is NA
dropna(axis=1, how='all')	Drop column if all the values are missing
dropna(thresh = 5)	Drop rows that contain less than 5 non-missing values
fillna(0)	Replace missing values with zeros
isnull()	returns True if the value is missing
notnull()	Returns True for non-missing values

PYTHON IMPLEMENTATION HANDLING MISSING VALUES

OUTLIERS VS ANOMALY

- **Outlier:** A value that you predictably find in your data that indicates your model does not work properly. Outliers are often indicators that the model does not describe the data properly and thus we should question the results of our model or quality of our data.

- **Anomaly**: A value that against all odds you find in your data that indicates your model does work properly. The concept of anomalies starts outside the theoretic world and inside the applied world: we want to look for unusual behavior in our data, sometimes motivated by the fact that we are interested in finding behavior that someone is trying to hide (like a virus in an email).

OUTLIERS VS ANOMALY

- To summarize, the two concepts are very similar in terms of the statistics behind them (i.e., unusual values given your fitted model) but come at the idea from different angles

- When we talk about outliers, we typically mean an unusual data point in the data used to fit our model; whereas an anomaly is usually meant as an unusual data point outside of the data used to fit our model

OUTLIER DETECTION

- A data object that deviates significantly from the normal objects as if it were generated by a different mechanism

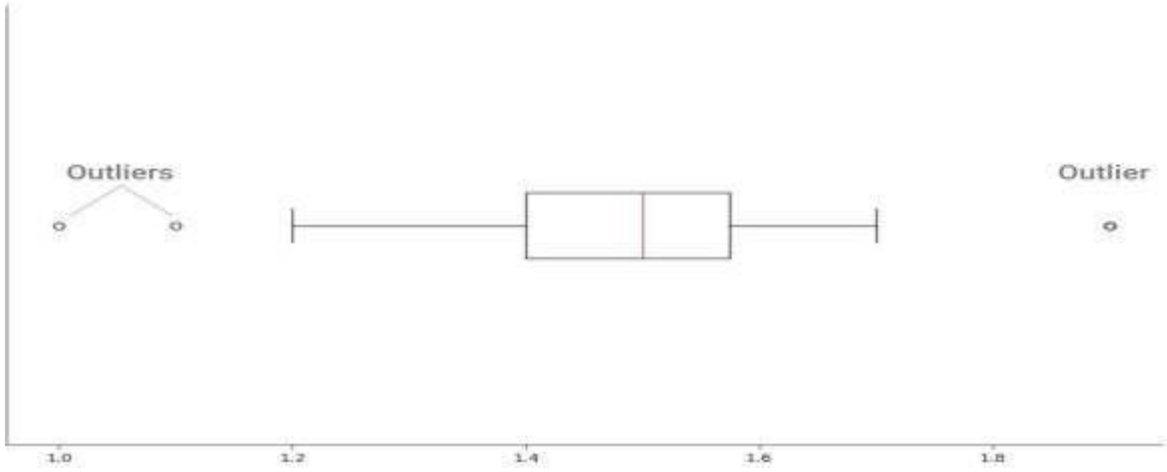
- **Outliers are interesting**: It violates the mechanism that generates the normal data

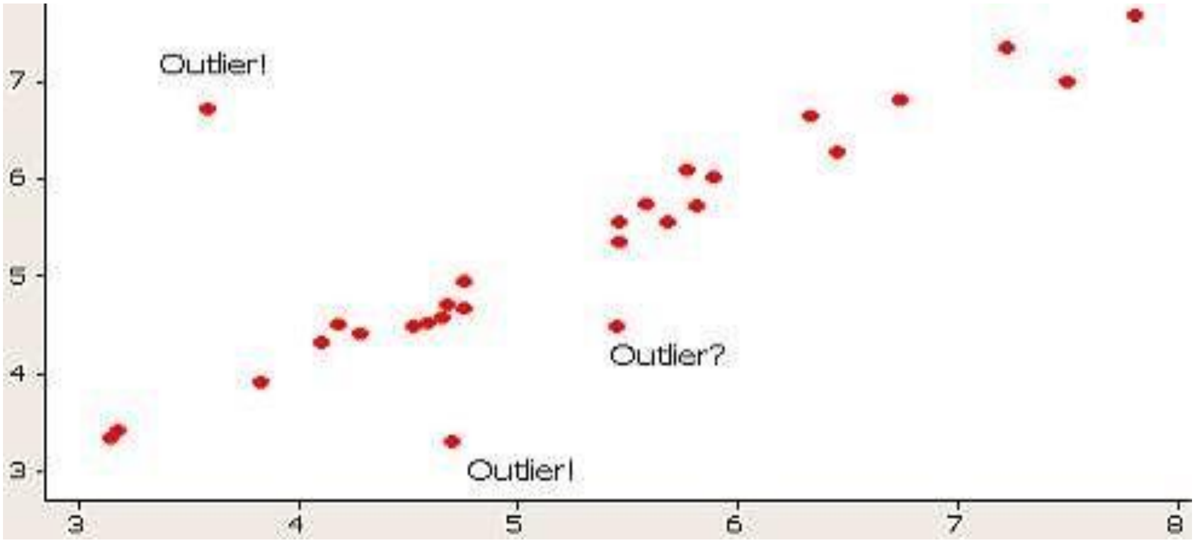
- **Applications**: Credit card fraud detection, Telecom fraud detection, Segmentation medical analysis

OUTLIER DETECTION

- To ease the discovery of outliers, we have plenty of methods in statistics

- Discover outliers with visualization tools or statistical methodologies
  - Box plot
  - Scatter plot
  - Z-score
  - IQR score





OUTLIER TREATMENT

- Once you’ve detected outliers now it’s time to treat them. Here are some methods you can follow to treat outliers
  - Drop outliers
  - Replace them with central tendency
  - Capping

PYTHON IMPLEMENTATION HANDLING OUTLIERS

QUIZ QUESTION 1

- **How do you handle missing or corrupted data in a dataset?**
  1. Drop missing rows or columns
  2. Replace missing values with mean/median/mode
  3. Assign a unique category to missing values
  4. All of the above

QUIZ QUESTION 2

- **How does pandas represent missing values when displaying a DataFrame or Series object?**
  1. NaN
  2. Null
  3. None
  4. "Missing Value"

QUIZ QUESTION 3

- **What will the following code snippet return?**

  some_dataframe.isna()

  1. A count of the missing values in each column
  2. A count of the missing values in each row
  3. A single boolean value True if the DataFrame contains one or more missing values, or False if there are none.
  4. A corresponding dataframe of boolean values, with True in cells that contain missing values, and False in cells that contain valid data

QUIZ QUESTION 4

- **Which of the following code snippets returns a count of the missing values in each column of a DataFrame?**
  1. some_dataframe.isna().sum()
  2. some_dataframe.dropna()
  3. some_dataframe.is_na().sum()
  4. some_dataframe.value_counts.is_na()

QUIZ QUESTION 5

- **Select all the correct code snippets that returns the unique values for a given column.**
  1. some_dataframe.get_unique_values()
  2. some_dataframe.unique()
  3. some_dataframe['some_column'].unique()
  4. some_dataframe.some_column.unique()

QUIZ QUESTION 6

- **A value that is much higher or much lower than the other values in a set of data.**
  1. Outlier
  2. Box Plot
  3. Range
  4. Histogram

QUIZ QUESTION 7

- **Find the outliers in the given data set below.**

  28, 26, 29, 30, 81, 32, 37

  A - 29

  B - 26

  C - 37

  D - 81

QUIZ QUESTION 8

- **Which is not the technique to detect outlier**
  1. One hot encoding
  2. Boxplots
  3. Z-score
  4. Inter Quantile Range(IQR)

SUMMARY

- Data values may be missing completely at random, missing at random or missing not at random
- Missing data may be problematic (depending on the type) since they can sometimes cause sampling bias
- Outlier is a data object that deviates significantly from the normal objects as if it were generated by a different mechanism
- Outliers bring to light activities and information that have significant implications on a subject matter or an organization

THANK
YOU !