Data Science Workshop

Data Preprocessing

- Steps:
  1. Data Cleaning.
  2. Data Fusion.
  3. Data Reduction.
  4. Data Transformation.

Data Cleaning

- Importance of Data Cleaning:

GIGO (Garbage in, Garbage out).

Data Cleaning

+ Data Cleaning Process:
- Data cleaning process can be divided into three levels:

Level 1, Level 2 and Level3

Data Cleaning

1

Level 1: Clean up the **table.**

2

Level 2: Unpack, restructure, and reformulate the table.

3

Level 3: Evaluate and correct the values.

Data Cleaning: Level 1

- This step is considered a **general cleaning step**, essential to start the next two levels of cleaning.

- To surpass this level, you need to make sure that your dataset:
  - Is in standard and preferred data structure.
  - Has codable and intuitive columns' titles.
  - Each row has a unique identifier.

Data Cleaning: Level 2

- It is more about preparing the dataset for analysis and the tools for this process.

- A level 2 cases that tends to happen frequently:
  - Unpacking Columns,
  - Reformulating the table and,
  - Restructuring the table.

Data Cleaning: Level 3

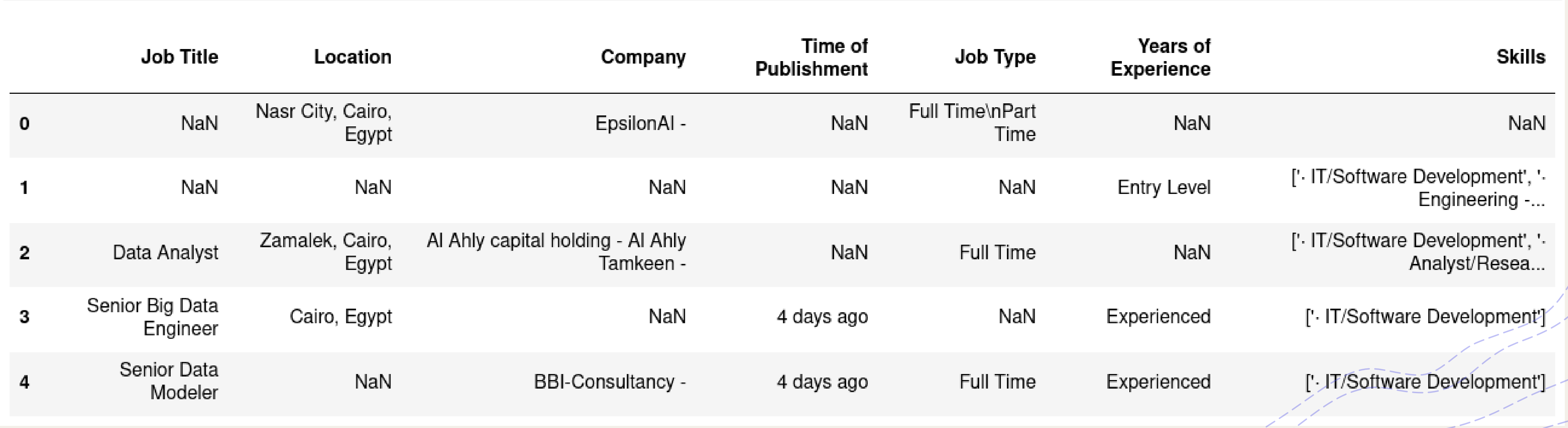
- Steps of Level 3 data cleaning:

1. Handling missing values.
2. Handling Errors in data.
3. Detecting extreme data points.

Data Cleaning: Missing Values

- **Missing values**: are values aren't available in dataset for unknown reasons.
- Python represents missing values in pandas data frame as via **NaN,** before we start, we need to make sure all missing values are in the same representation

(transform all other representations ex. None, N/A, 0, 9999, … to NaN).



	Job Title	Location	Company	Time of Publishment	Job Type	Years of Experience	Skills
0	NaN	Nasr City, Cairo, Egypt	EpsilonAI -	NaN	Full Time\nPart Time	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	Entry Level	['· IT/Software Development', '· Engineering -...
2	Data Analyst	Zamalek, Cairo, Egypt	Al Ahly capital holding - Al Ahly Tamkeen -	NaN	Full Time	NaN	['· IT/Software Development', '· Analyst/Resea...
3	Senior Big Data Engineer	Cairo, Egypt	NaN	4 days ago	NaN	Experienced	['· IT/Software Development']
4	Senior Data Modeler	NaN	BBI-Consultancy -	4 days ago	Full Time	Experienced	['· IT/Software Development']

Data Cleaning: Missing Values

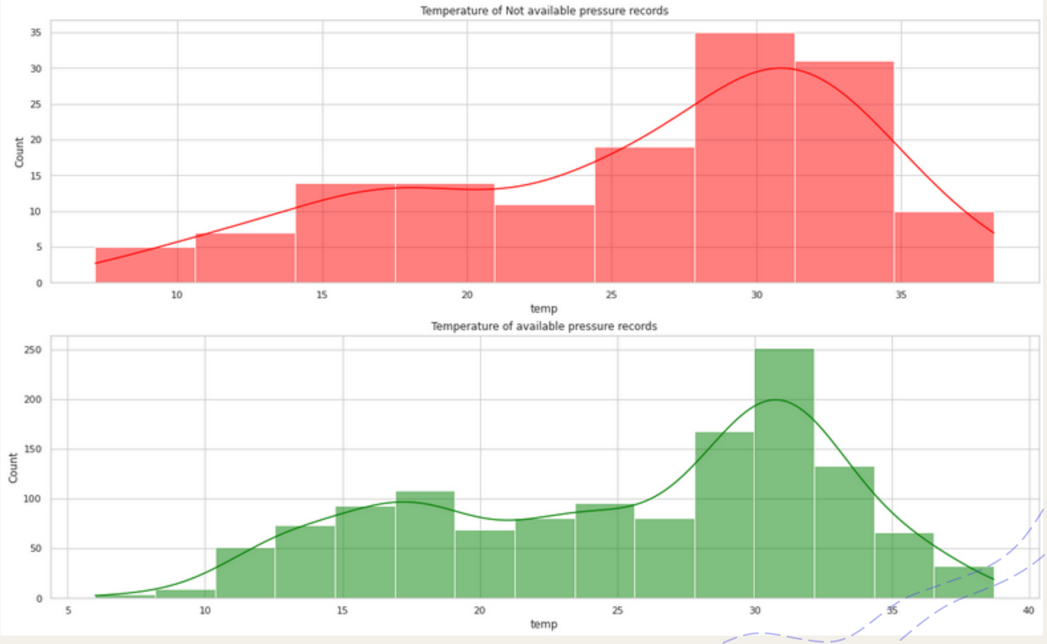
- Causes of missing values:
  - Human error.
  - Incomplete surveys.
  - Lost records from databases.
  - Programming bugs.
  - Sensor malfunctions.

Data Cleaning: Missing values

+ Types of missing values:
  - MCAR: Missing Completely at Random.
    - There is no systematic reasons for missing values.
  - MAR: Missing at Random.
    - There is systematic reason that causes missing values but doesn't always cause missing values.
  - MNAR: Missing not at Random.
    - The most problematic type, and we must figure out its reason and stop it.

Data Cleaning: Missing values

+ **MCAR(missing completely at random)**: in this type missing values are randomly and doesn't depend on other variables, So distribution of other variables is the same for available and non-available rows.

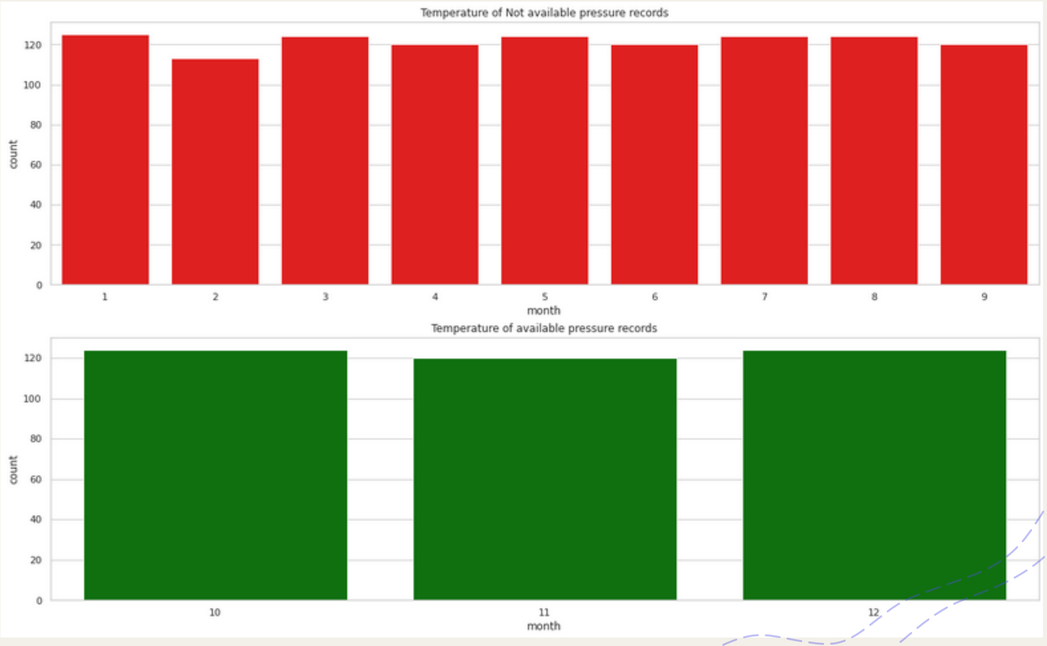


Data Cleaning: Missing values

MAR(Missing at Random): in this type missing values are not totally random, but it may have systematic reason that causes missing values, but this reason doesn't have to lead to missing values. (ex. Sensors of distance at senor's boundaries)

Data Cleaning: Missing values

- **MNAR(missing not at random):** The pattern of missingness is related to other variables in the dataset, but in addition, the values of the missing data are not random.



Data cleaning: Missing values

+ Dealing with missing values:
  - Keeping them as is

    use techniques that handle missing values ex. KNN.
  - Deleting rows with missing values.

    should be careful and avoid removing MAR, MNAR as it will lead to high bias.
  - Deleting columns with missing values.
  - Estimate and Impute values

Data Cleaning: Missing values

There are several ways to estimate missing values:

- General Central tendency (mean, median, mode) . better for MCAR.
- Group central tendency. (better for MAR).
- Regression analysis.
- Interpolation.

Thank You