IoT Data Cleaning and Visualisation

By

Richard Sserunjogi & Lillian Muyama
Department of Computer Science,Makerere University, Uganda
richard.sserunjogi@airqo.net & lillian@airqo.net

Data Preparation Process

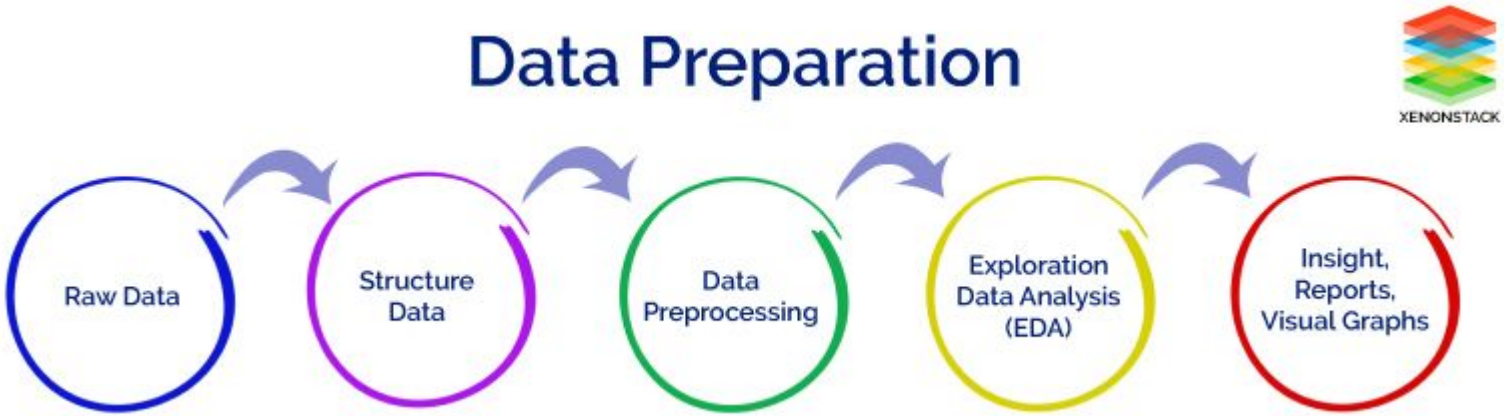
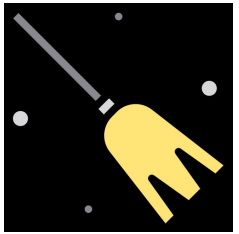


Figure 1: Data preparation process. source:xenonstack.com

Data Cleaning



Data cleaning is the process of **detecting** and **correcting** (or removing) corrupt or inaccurate records from a data set. The incomplete, inaccurate or irrelevant parts of the data can be replaced, modified or deleted.

Why is it important?

- GIGO Concept: How well you clean your data has a direct impact on the quality of your results.

What we are looking at:

- Validity e.g. data types, range constraints etc.
- Accuracy i.e. degree to which the data is close to the true values.
- Completeness i.e. degree to which all required data is known.
- Consistency i.e. the degree to which the data is consistent.
- Uniformity i.e. the degree to which the data is specified using the same unit of measure.

The Cleaning Process

1. **Inspect** the data for inaccuracies, anomalies, inconsistencies and gaps.

   With Python's Pandas library, we can use functions such as dataframe.describe(), dataframe.info(), dataframe.shape, dataframe.isnull().sum() etc. Also visualisations

2. **Clean** the data.
3. **Verify** that the data has been cleaned using some of the methods in 1 and several additional methods dataframe.head(), dataframe.tail().
4. Always **record** what has been done to the data.

Cleaning IoT data - ex. AirQo

- Dealing with datatypes
- Dealing with null values
- Dealing with outlier values
- Dealing with units of measure
- Dropping unwanted columns

The main columns we shall deal with are the **2 PM 2.5 columns**, **the 2 PM 10 columns** and the **latitude** and **longitude** coordinates. **Temperature** and **humidity** will also feature at a later time.

Data Visualisation

Process of interpreting data and presenting it in a pictorial or graphical format.

Examples(Air quality related visuals)

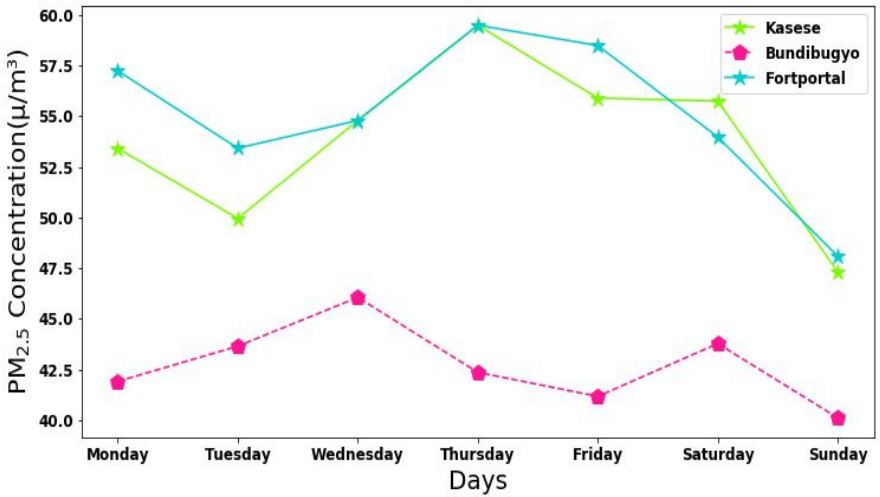


Figure 2. Average $PM_{2.5}$ for each day of the week for three monitoring stations in Western Uganda

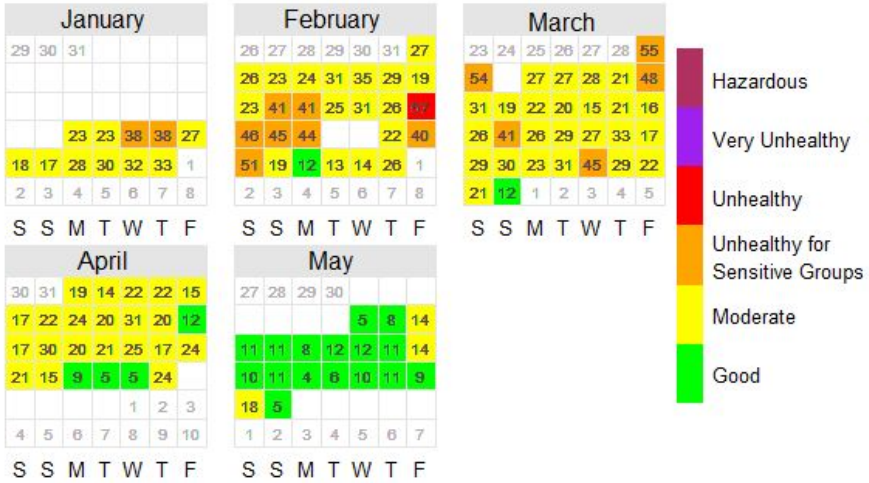


Figure 3: Daily average $PM_{2.5}$ for a certain location Uganda from January to May 2019

Examples cont..

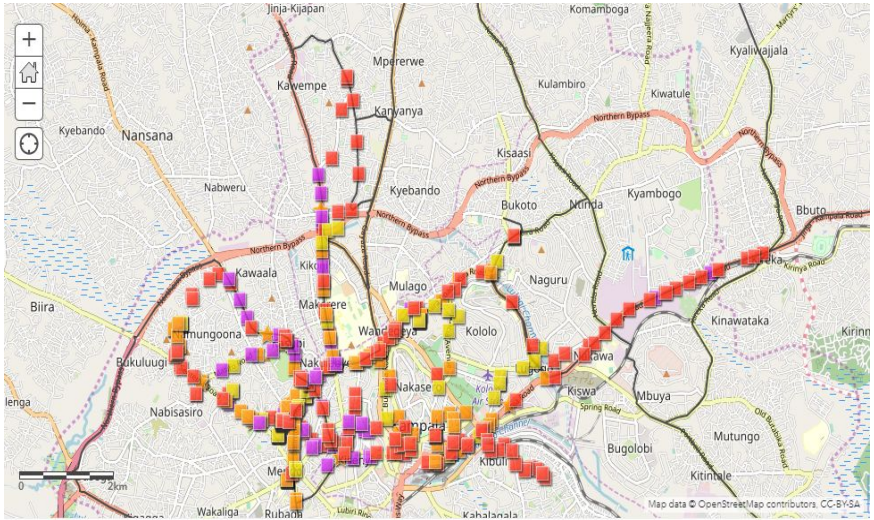


Figure 4: Map showing $PM_{2.5}$ concentrations for one of the mobile air quality monitoring units in Kampala on 23-Feb-2019

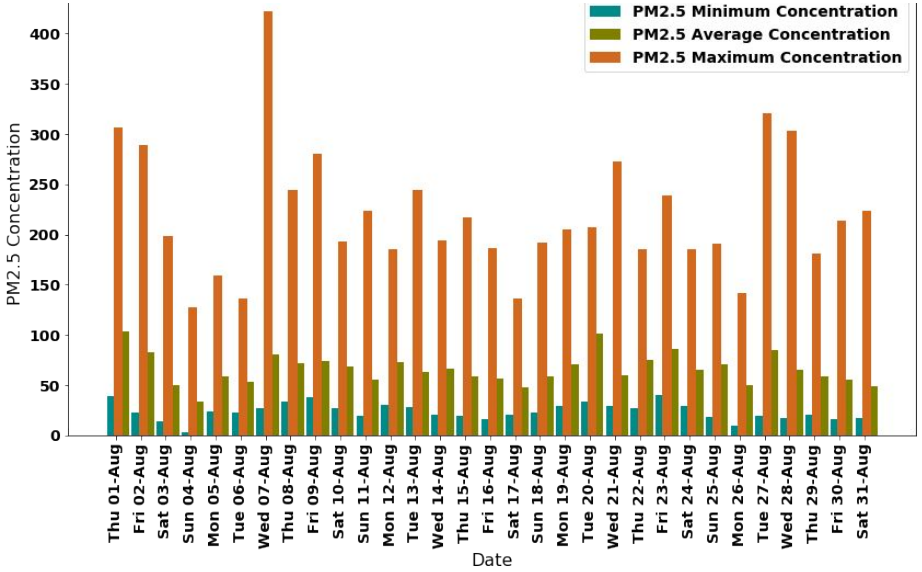


Figure 5: Daily average $PM_{2.5}$ for a certain location Uganda from January to May 2019

Why is data visualisation important?

- Easier trend & pattern visualisation within data sets
- Simplifies data by giving a full picture of the scoped parameters leading to efficient decision making
- Unifies interpretation through the use of graphics and charts

Common python visualisation packages

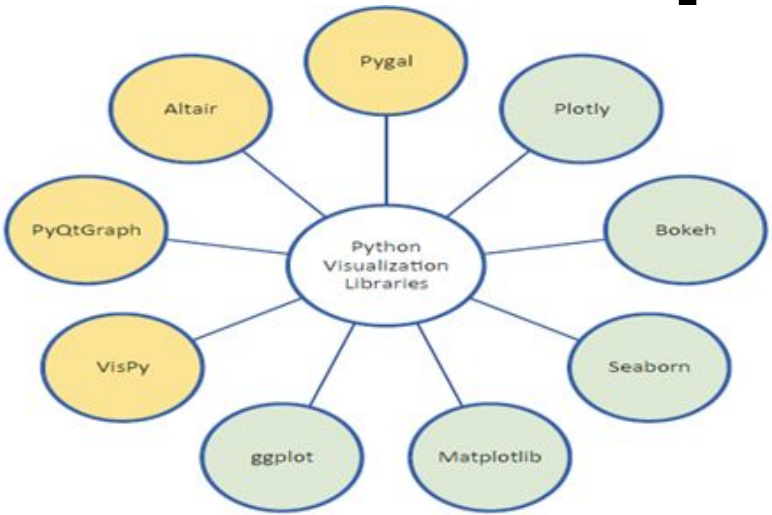


Figure 6: Libraries for python data visualisation (Src: Data Analysis and visualisation Using Python book)

Practical Session

Prerequisite

- Knowledge of python programming language
- Have anaconda/conda installed

Required Packages

- Numpy
- Pandas
- Matplotlib
- Seaborn