Data Science Workshop Session 3

Ikenna Ivenso

27 June 2020

www.tqstem.org

Objectives

1. Getting Data
2. Cleaning Data
3. Transforming Data
4. Exploring Data
5. Working with Unstructured Data

Getting Data

Getting Data from a Database:

- SQL is the most popular database querying tool/language
- To get started with SQL, try this interactive tutorial
- Python also has libraries for connecting to and interacting with databases (e.g. pyodbc)

Data can also be retrieved from CSV, TXT, EXCEL and many other file formats

- In the case of EXCEL, individual sheets can be retrieved

Cleaning Data

- Sometimes data can come in formats that are not easy to work with
- This can happen in so many different ways
  - Wrong types
  - Inconsistent formats
  - Missing characters
  - Extra characters
  - etc.
- The goal of data cleaning is to make the data usable
- *Example*: reported RAM capacity of computers (inconsistent format)

Transforming Data

- Data transformation is the process of adjusting or modifying data to make it easier to use
- Transformation may involve modifying the original data
- Transformation can occur in many different ways depending on the goal
- We'll look at 2 examples :
    1. Tagging our data to identify high-memory machines
    2. Working with date and time data

Exploring Data

- Exploring data helps us gain some understanding of it
- Data exploration can help expose problems in data
  - Missing data, wrong data and spurious data
  - Mixed data types in the same fields
  - Outliers
  - etc.
- Data exploration can also help point us towards the best approach to solve a given problem
- Visualization is a very useful tool for data exploration

Unstructured Data

- Most data is unstructured at its source
- If we're lucky, the data will be preprocessed into tabular form
- But sometimes, we may need to work with data in its unstructured form
- Unstructured data can come in countless different formats
- Let's look at an example from book reviews

THANK YOU!
Any Questions

www.tqstem.org