Open Discussion

Asessing Data

Cleaning Data

Asessing Data

Types of Unclean Data

1. Dirty/ low-quality data [content issues]
2. Messy/ untidy data [structural issues]

Types of Assessment

1. **Visual assessment**
   - Opening it and looking through the data in its entirety, in pandas, a text editor or a spreadsheet application.
   - Great for getting acquainted with the dataset.

2. **Programmatic assessment**
   - Uses code to view specific parts of the data. You can even plot the data, but plotting isn't done very often when wrangling. This is more for exploratory data analysis.

Types of Visual Assessment

1. **Directed**
   - Driven by the problem you want to solve, like checking the values in the columns and rows you plan to use in your analysis.

2. **Non-Directed**
   - Just scrolling aimlessly and stumbling upon issues.

Types of Programmatic Assessment

1. **Directed**
   - Driven by the problem you want to solve, like checking the values in the columns and rows you plan to use in your analysis.

2. **Non-Directed**
   - Randomly typing in programmatic assessments without any directed goal in mind. The ***.sample()*** method pandas is the on data frames, displays a random sample of entries.

Steps of Assessment

1. **Issue Detection**

   Either visually or programmatically

2. **Issue documentation**

   Writing down your findings

Data Quality Dimensions

1. **Completeness**: do we have all of the records that we should? Do we have missing records or not? Are there specific rows, columns, or cells missing?

2. **Validity**: we have the records, but they're not valid, i.e., they don't conform to a defined schema (rules). These rules can be real-world constraints (e.g. negative height is impossible) and table-specific constraints (e.g. unique key constraints in tables).

Ps. These are listed in decreasing order of severity, meaning that the dimension listed first, completeness, is the most important.

Data Quality Dimensions

3. **Accuracy**: inaccurate data is wrong data that is valid. It adheres to the defined schema, but it is still incorrect. Example: a patient's weight that is 5 lbs too heavy because the scale was faulty.

4. **Consistency**: inconsistent data is both valid and accurate, but there are multiple *correct* ways of referring to the same thing. Consistency, i.e., a standard format, in columns that represent the same data across tables and/or within tables is desired.

Ps. These are listed in decreasing order of severity, meaning that the dimension listed first, completeness, is the most important.

Cleaning Data

Open Discussion

Types of Data Cleaning

1. **Manual Data Cleaning** includes:
   - Retyping incorrect data
   - Copying and pasting columns and rows
   - However, manual cleaning is inefficient, error-prone, and demoralizing. So never clean manually.

2. **Programmatic Data Cleaning** uses code to:
   - Automate cleaning tasks
   - Minimize repetition
   - Save time

Data Cleaning Steps

While there are ways to clean data manually using spreadsheet programs and text editors, the best way to clean data is to code it yourself.

This requires three steps:

1. ***Define:*** how you will clean the issue in words. Use verbs to define actions

2. ***Code:*** convert your definitions into executable code

3. ***Test:*** your data to ensure your code was implemented correctly

Be in Demand