Data Cleaning

An Expanding the Data Analytics Technician Pipeline Module

Data Cleaning

Dirty data is data that has erroneous or missing information. It's data that is not quite ready to be analyzed, or contains the 3 "I"s: inaccurate, incomplete, or inconsistent data. Dirty data is unreliable data, and untrustworthy as being accurate and original. The meaning conveyed by the data may change as it is transformed into models, graphs and into other software applications. It is important to identify dirty data, and intentionally correct or "clean" it so it can be trusted to produce reliable predictions and results.

Examples of dirty data include, but are not limited to:

- Missing data or gaps in columns
- outliers that may skew the data too much to be useful to analysis
- duplicated values where there shouldn't be
- Merged values that shouldn't be
- syntax or data type conversion errors (too much or too little white space, integers where decimals should be, or numbers being treated as letters, formatting errors)
- unwanted or irrelevant values
- inconsistent structure preventing accurate mapping of important data points
- data that does not contain standardized information (St. instead of street, or differing units of measurement)
- cross-set errors which come from combining data from two different sources

Data can be stored in many different formats, but the three main categories are

- Cleaning data with Excel https://www.simplilearn.com/tutorials/excel-tutorial/excel-data-cleaning
- Cleaning with Python https://towardsdatascience.com/your-ultimate-data-manipulation-cleaning-cheat-sheet-731f3b14a0be
- Other open source tools: https://careerfoundry.com/en/blog/data-analytics/best-data-cleaning-tools/

Vocabulary

- fidelity - data which retains its accuracy and meaning as it is transformed during data analysis.
- integrity - the assurance that information is unchanged from its original source and has not been accidentally or maliciously modified, altered, or destroyed.
- outliers - data points that are outside the trend or pattern
- nested functions - a function inside a function
- right() taking part of a data point from the end to the beginning (right to left)
- left() - taking part of a data point from the beginning to the end (left to right)

- Mid() - taking the middle part of a data point starting from a specific position and going for a specified number of digits/characters
- Len() - finding the size of the data point
- Trim() - cleaning a data point by removing spaces that may surround the data point
- drop_duplicates - cleaning data by removing duplicate rows of data
- split() - breaking combined data points into separate columns based upon spacing or punctuation. Also known as parsing
- pivot table -a statistic table that reorganize data columns and rows into aggregate or grouped points for analysis

Cleaning Using Excel

Excel offers many great ways to clean data, including:

- removing duplicates
- parsing data from text to columns
- being able to correct the number type
- clearing formatting
- spelling checking
- adjusting the case (upper/lower, sentence case)
- highlighting errors using conditional formatting
- trimming white space
- finding and replacing values

You will have the opportunity to practice these in the Activity included below.

Cleaning Using Python

If you have experience with basic Python, you may want to use pandas and numpy Python libraries to clean your data. You will probably be using pandas dataframe format for a lot of the cleaning.

Python offers the following commands to help clean data

- drop()
- drop_duplicates()
- split()
- upper()
- lower()
- capitalize()
- title()
- swapcase()

- head()
- isnull()
- NaN
- try/except
- fillna()

You will have the opportunity to practice these in the Activity included below.

Activity 1: Excel

Data Cleaning with Excel part 1:
https://www.youtube.com/watch?v=WRk9t5yo5Zs&ab_channel=Analytics4All Be sure to download the sample file provided.

Data Cleaning with Excel part 2:
https://www.youtube.com/watch?v=Jt76Q0dVEm4&ab_channel=Analytics4All

(Optional additional resource if needed) Top 30 data cleaning tricks in Excel
https://www.youtube.com/watch?v=K2bbUwDxAxk&ab_channel=YodaLearningAcademy

Steps

1. Download the sample files, and follow along with the tutorial
2. Complete the worksheet below:

Data Cleaning with Excel Worksheet

What steps did the author use to find and remove duplicate, unwanted data? Can you think of another situation or two where this skill would be useful in analyzing the data? Explore the menu options for this to help answer the question.	
To fill in the missing values in the MECH column, the author entered "99" into a separate column, then used copy/paste/autofill to correct the data. List the steps.	
After numbers are converted to the appropriate text in the video, Google "Excel Data Type Conversion Functions". Create a list of the functions and the definition to the right.	
What is the purpose of VLookup? What dirty data problem(s) was solved using VLookup?	

Why is it important to press F4 in Excel after highlighting the area you wish to work with?	
List the steps to using Find/Replace to fix spelling error “AOILI” with “OIL”	
Use your own works to explain the nested function used to correct the license plate number.	
What was the key combo of shift+ctrl+enter used for?	
Create a pivot table to average the prices of the car. Paste it to the right.	

Activity 2: (Python)

Resources

Data Cleaning with Numpy and Pandas: https://realpython.com/python-data-cleaning-numpy-pandas/
Be sure to download the sample file provided.

https://github.com/realpython/python-data-cleaning/blob/master/Datasets/BL-Flickr-Images-Book.csv

https://github.com/realpython/python-data-cleaning/blob/master/Datasets/university_towns.txt

https://github.com/realpython/python-data-cleaning/blob/master/Datasets/olympics.csv

Cleaning missing values: **https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b**

Steps

1. Download the sample files, and follow along with the tutorial

2. After completing both tutorials, try cleaning the sample files from the Excel activity by yourself using Numpy and Pandas

Extension:

Have students research open source data cleaning tools such as Kinme or Open Refine, Knn Imputing (Python tool), Outlier/Anomaly Detection (Python Tool), and present cleaning with that data tool, again using the same data set from the Excel or the Python tutorial.

This material is based upon work supported by the National Science Foundation under Grant DUE 2055411.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.