03/29/2021

Data Analysis Lifecycle:

Data Discovery – Understanding the field to be analyzed

Includes:

- Defining the objective of the analysis (goal of the collection and analysis of the data)
- What question do you want to answer?
- What is the kind of data would you need to answer it?
- Are there proxies for the kind of data you need?
- Define the scope of the analysis or objective
- What approach is going to be used? Can impact the kind of data you need to collect; it may be influenced by the data that is available
- What specialist information about the domain in question do you need to know before you begin?
- Have any other analyses/models addressed this question before? How would your approach differ?
- Identify the tools to be use: Excel? Python? Tableau? R? Do you need things like big data tool? Parallel processing in Spark or another big data technique?
- Identify the types of analyses: dashboard? Machine learning?
- Can you break the problem down into smaller pieces?

Some people consider "data discovery" to be part of the data exploration phase
Some use "discovery" to mean more like understanding and learning – including understanding the domain of study, and the specific data set you are using.

Many datasets come with information about the variables in the form of metadata, or a dictionary/glossary

All of this is part of the planning process.

Web scraping

In Excel: set up to scrape tabular data off the web (HTML) pages
Give it the URL and Excel brings up a list of tables that are on that page, and you can import one or all of those tables into Excel.

In Python:
Requests
BeautifulSoup

API – some companies allow you to sign up for an API key that allows you to scrape data off their website
Some will provide info in HTML format, but some in JSON, or data formats

Census:
American Community Survey (ACS)
Data.census.gov
IPUMS – microdata