4/1/2024

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?
- 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 or models addressed the question before? How would your approach differ?
- Identify the tools to use: Excel? Python? Tableau? R? Do you need things like a big data tool? Parallel processing? Spark or another big data package?
- Identify the types of processing/visualizations: dashboard? Machine learning?
- Can you break the problem down into smaller pieces?

Some people consider "data discovery" to be 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 data dictionary/glossary.

All of this is part of planning process.

Webscraping

In Excel:
Set up to retrieve tabular data (data in tables) in HTML pages very easily.
Give Excel the URL and Excel brings up a list of tables that are on that page, and you can import one or all of the tables into Excel.

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 other formats

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

Resources:

1. https://www.dataquest.io/blog/free-datasets-for-projects/
2. https://www.nature.com/sdata/policies/repositories
3. https://data.gov/
4. https://github.com/dsc/colorbrewer-python
5. https://realpython.com/beautiful-soup-web-scraper-python/
6. https://www.dataquest.io/blog/python-api-tutorial/
7. https://rapidapi.com/blog/how-to-use-an-api-with-python/
8. https://www.tibco.com/reference-center/what-is-data-discovery
9. https://www.netsuite.com/portal/resource/articles/erp/data-discovery.shtml
10. https://bi-survey.com/data-discovery
11. https://www.xenonstack.com/insights/data-discovery-tools
12. https://www.sisense.com/glossary/data-discovery/
13. https://coresignal.com/blog/data-discovery/

Extended commentary:

Data discovery is an essential step in the data analysis lifecycle, where the analyst or the data scientist looks for relevant data that is required to answer the research question or to address the business problem at hand. It involves gathering data from multiple sources, such as databases, spreadsheets, data warehouses, data lakes, social media, and other external data sources.

The process of data discovery typically involves defining the research question or the business problem, identifying the key variables, and determining the data sources that contain the relevant information. The analyst then performs data profiling to assess the quality of the data, such as the completeness, accuracy, consistency, and integrity of the data.

Data discovery also involves exploratory data analysis, which is an iterative process of exploring the data to identify patterns, trends, relationships, and outliers. This process may involve visualizations, statistical analysis, or machine learning techniques to gain insights into the data.

Overall, data discovery is an essential step in the data analysis lifecycle as it helps to identify the data sources that are relevant for the analysis and ensure that the data is of high quality and integrity, which is necessary for accurate and reliable analysis results.

Excel can be used to scrape data from the web using the "From Web" feature. Here's how to do it:

Open Excel and create a new workbook.

Click on the "Data" tab in the ribbon and select "From Web".

In the "New Web Query" dialog box, enter the URL of the web page from which you want to scrape data and click "Go".

Use the mouse to select the data you want to scrape on the web page. This will highlight the data in yellow in the "Web Query" dialog box.

Click "Import" to import the selected data into Excel.

Excel will create a table with the scraped data. You can then manipulate and analyze the data as needed. Note that this method may not work for all websites, particularly those that require authentication or use complex JavaScript to display data. In such cases, other web scraping tools may be more appropriate.

In Python, you can use various libraries to scrape data off the web, such as Beautiful Soup, Scrapy, and Selenium. Here is a brief overview of each approach:

Beautiful Soup: Beautiful Soup is a Python library that allows you to extract data from HTML and XML documents. You can use it to parse the HTML content of a web page and extract the information you need. It has a simple and intuitive API, making it easy to use even for beginners.

Scrapy: Scrapy is a more powerful and comprehensive web scraping framework in Python. It provides a more structured and efficient way to extract data from websites, with advanced features like automatic throttling, asynchronous requests, and built-in support for handling common web scraping tasks like following links and dealing with pagination.

Selenium: Selenium is a web testing tool that can also be used for web scraping. It allows you to automate web browsers and simulate user interactions, making it possible to scrape data from websites that require JavaScript execution or user input.

All of these approaches can be used to scrape data off the web in Python, depending on your specific needs and the complexity of the task at hand.