Data Planning

Goal:

- Clearly define your goal (write it down)
  - Measures of success, and plans on how to achieve that

Deliverable:

- Documentation of your goal
- If you haven't defined success, you won't know when you have achieved it

How You Get There: By answering questions about the final product & formulating or identifying any initial hypotheses.

Acquisition

- AKA: Data gathering
  Data Import
  Data Wrangling (Acquisition + Prep)

Goal: • create a path from original data sources to the enviornment in which you will work w/ the data

Deliverable: A file, acquire.py, that contains the function(s) needed to reproduce the acquisition of data

How to get There: SQL: clean-up, integration, aggregation or other manipulation of data in the SQL Enviornment

Pylib: pandas

- May use Spark and/or Hive when acquiring data from a distributed enviornment such as HDFS.
- Examples of source types:
  - RDBMS
  - HDFS
  - Static local flat files (csv, txt, xls

My SQL (Structured Query Language)

- RDBMS - Relational Database Management system
  - Stores data in tables + creates relationships between the data in different tables.
  - Most common way to permanantly Store data.
  - Manages the data

- Database (DBMS) - Actual location of the data stored on a disk

- Database Client - Program used to connect to a databas

- Database Server - Computer that runs the DMBS + stores the data
  - Either on premises or in the cloud

- DDL (Data Definition Language) - Commands that change the structure of the database or the DMBS
  - Changes structure of tables in database

- DML (retreival) (Data Manipulation Language) - Insert, update, delete, retrieve information from databases

Queries - Statements sent to the server individually, w/ results sent back to the client
- ex) how many clients does this seller have?
- Asks questions

mySQL: Used to interact w/ database server run