***Udacity Data Analyst Track***

**III. Intro to Data Analysis**

1. Data Analysis Process

* 1) Figure out a question or a problem to solve
* What are characteristic of students who pass project, how to stock a store w/ what people want to buy?
* 2) Data Wrangling
* 2a) Data acquisition
* 2b) Data Cleaning (investigating it 1st)
* 3) Data Exploration
* Get familiar w/ the data, build intuition about it, + try to find patterns in it
* 4) Draw conclusions or make predictions
* Predict: which movies users will like
* Conclude: users are less likely to click certain articles
* This phase usually requires statistics + machine learning
* 5) Communicate Findings
* Reports, dashboards, presentations, blog posts, papers, emails, etc.
* This process does not follow a straight line and it is iterative
* May have to go back to the question and refine it as we become more familiar w/ the dataset
* Sometimes a dataset is acquired before coming up w/ a question + we investigate the data to think of one

*Data Wrangling*

* Can get data from a CSV, a database, an API, scraping a webpage, and then combine data from various sources that are in different formats
* CSV files have no formulas and are much easier to process w/ code, unlike Excel spreadsheets
* Can represent a CSV in python where each row is
* A list, so the entire dataset is a list of lists
* A dictionary, which works well if the CSV has a header 🡪 keys = col names, values = fields
* Can use Python’s **unicodecsv** library to load in CSV files, or use open to open a file connection and read it in