1. Chapter 1: Intro, skills and the value of a research question
   1. We may not need to sell analytics
   2. Lay out challenges (problems we solve in the book)
   3. Go over skills of a security data scientist
   4. The value of a research question (a.k.a. “so what?”)
   5. a little deeper overview of the use cases
   6. “how to use this book”
2. Chapter 2: Basic Data Manipulation
   1. Sometimes data loves company (data merging)
   2. Use Case: Bad IPs (Alienvault? perhaps vz iocdb)
      1. Write an auto-retrieval script
      2. Write a script to do geo-location (I thought there was an open geo IP project, but can’t seem to find it)
      3. Or ASN or whatever
      4. pull in other data sources to provide value
      5. descriptive statistics per country (over time?)
         1. 5 - number summary
         2. blox plots, distributions, or something
   3. Use Case: Honeypot data?
3. ata Munging

Possible sources for use cases:

1. Netflow
   1. Take a 24 hour period (or hourly sampling), compare current snapshot to previous/similar views, look for clustering
   2. System classification (unsupervised)
      1. MDS (distance plotting)
      2. k-nearest neighbors
         1. note, these two were taken directly from ML for hackers
   3. System classification (supervised)
      1. random forest
         1. I can collect a handful of “known bad” IP addresses and their function and we could pull/create a handful of profiles to train on.
2. Mod\_security logs
3. port scans / honeypot data
   1. Descriptive visualziations (re-do box plot?)
   2. frequency of scans, describe the scans, (like my blog series last year)
4. list of bad IPs (or IP addresses)
   1. ASN
      1. ASN “badness” from alienvault
      2. as a proportion of the ASN entity population over time
   2. Geolocation
      1. Write a script to pull geo-location from one or more services
      2. Write a web-scraper to pull population (like we did for zero access)
      3. do some linear regression to see how they map
   3. DNS / Passive DNS
      1. malware C&C algorithms (may be a bit tough to touch on)
5. Nessus (or other) Vuln scan
6. Spam (general)
7. Bots
   1. Malware C&C
   2. netflow analysis of
8. Open DNS
9. Password lists
   1. Character analysis, etc
   2. so-what?
10. Meta-analysis
    1. internal vs external actor
11. Firewall logs
12. Patch coverage
13. Logins / LDAP queries
14. SNORT (or other) IDS
15. Phishing
16. VERIS data
    1. Conversion tools - public HHS data dump (we can walk through converting to VERIS)
    2. Screen scraping for victim demographics
    3. call to other frameworks
    4. Completeness of data collection (as a component of time)
    5. VISUALizing with tableau, R