IDS 521 Data Warehousing Project Prof. Ali Tafti
Spring 2015

Example data sources

This document provides some ideas on example datasets to consider using for your course project. You are encouraged to search further and consider any datasets not listed here.

1) **University of Arkansas Datasets, access to Microsoft SQL Server 2012 BI Cubes Department Stores:** http://enterprise.waltoncollege.uark.edu/IBM.asp
    a. Note that Sam's Club Data and some of the other datasets have serious data integrity problems. I have found the Dillard's data to be the best among others on that server.
2) **Research-quality datasets: https://bitly.com/bundles/hmason/1**
    a. A hub for many different datasets. Links to data sources in many other websites used by researchers in many different fields. Site created by Hilary Mason: http://www.businessweek.com/articles/2012-04-26/hilary-mason-from-tiny-links-big-insights.
3) Public datasets on Amazon Web Services: http://aws.amazon.com/publicdatasets/
4) **City Forward Datasets:** This is a repository of data from cities around the word. Data comes from vetted sources: http://cityforward.org/wps/wcm/connect/CityForward_en_US/City+Forward/Home
5) **City of Chicago Data Portal Datasets:** The City of Chicago has made over 273 datasets available on topics such as energy, crime, public transportation, and more: https://data.cityofchicago.org/
6) **Kaggle** crowd-sourcing competitions provide interesting and rich datasets: **http://www.kaggle.com/competitions**
7) Google Insights: www.**google**.com/**insights**/search/
    a. Offers rich historical and regional data on Google search data over time. Statistics can be gathered on categories of search terms, or specific search terms. It can be used as a basis for predictive models for a wide variety of consumer behavior such as sales in any industry, automobile accidents, investor sentiment, air quality, etc.. You might also use Google Correlate to explore correlations as you search for a topic (to do some 'unsupervised learning'): http://www.google.com/trends/correlate/
8) The UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) contains many interesting data sets that can be used for your term projects.
9) Government datasets:
    a. Open Government Directive Agency: http://www.data.gov/ogd
    b. National Bureau of Economic Research: http://www.nber.org/data/
        i. Other data collections: http://www.nber.org/links/data.html
    c. Bureau of Labor Statistics: http://www.bls.gov/
    d. FDIC data on U.S. banks: http://www.fdic.gov/bank/statistical/index.html