Project Dataset List

The following datasets located at **www.kaggle.com (http://www.kaggle.com)** are candidates for the group project. I have looked at the layout and contents of the source data and feel that they are capable of supporting the data warehouse to be built by a group.

(NOTE: Some of these datasets may require tweaking to fully meet the project requirements.)

Kaggle requires a free account to be created to access the site.

The URLs and a short description of each dataset follows.

**https://www.kaggle.com/manasgarg/ipl (https://www.kaggle.com/manasgarg/ipl)**

All Indian Premier League Cricket matches between 2008 and 2016.

This is the ball by ball data of all the IPL cricket matches till season 9.

The dataset contains 2 files: deliveries.csv and matches.csv.

matches.csv contains details related to the match such as location, contesting teams, umpires, results, etc.

deliveries.csv is the ball-­by-­ball data of all the IPL matches including data of the batting team, batsman, bowler, non­striker, runs scored, etc.

**https://www.kaggle.com/abecklas/fifa­world­cup (https://www.kaggle.com/abecklas/fifa­world­cup)**

The FIFA World Cup is a global football competition contested by the various football-­playing nations of the world. It is contested every four years and is the most prestigious and important trophy in the sport of football.

The World Cups dataset show all information about all the World Cups in history, while the World Cup Matches dataset shows all the results from the matches contested as part of the cups.

**https://www.kaggle.com/aparnashastry/building­permit­applications­data (https://www.kaggle.com/aparnashastry/building­permit­applications­data)**

A building permit is an official approval document issued by a governmental agency that allows you or your contractor to proceed with a construction or remodeling project on one's property.

The data was downloaded for the dates ranging from Jan 1st, 2013 to Feb 25th, 2018 using the filter in San Francisco open data portal. There are 43 columns and close to 200k records in the downloaded version (kept here). Description is separately uploaded as a dictionary.

**https://www.kaggle.com/meganrisdal/la­county­restaurant­inspections­and­violations (https://www.kaggle.com/meganrisdal/la­county­restaurant­inspections­and­violations)**

Restaurants and markets in Los Angeles county are regularly inspected for health code violations. The county makes this data publicly available and accessible, enabling a transparent look into this public health information.

The data covers health code inspections (inspections.csv) and health code violations (violation.csv).

**https://www.kaggle.com/cyaris/2016­mlb­season (https://www.kaggle.com/cyaris/2016­mlb­season)**

2016 MLB Season

Scraped data from [https://www.baseball­reference.com/](about:blank)

**https://www.kaggle.com/jameslko/gun­violence­data (https://www.kaggle.com/jameslko/gun­**

**violence­data)**

Gun Violence Data

A record of more than 260K gun violence incidents with detailed information about each incident, available in CSV format. We hope that this will make it easier for data scientists and statisticians to study gun violence and make informed predictions about future trends.

The CSV file contains data for all recorded gun violence incidents in the US between January 2013 and March 2018, inclusive.

**https://www.kaggle.com/cites/cites­wildlife­trade­database (https://www.kaggle.com/cites/cites­**

**wildlife­trade­database)**

A year in the international wildlife trade

The **Convention on International Trade in Endangered Species of Wild Fauna and Flora**, or **CITES** for short, is an international treaty organization tasked with monitoring, reporting, and providing recommendations on the international species trade. CITES is a division of the IUCN, which is one of the principal international organisation focused on wildlife conversation at large. It is not a part of the UN (though its reports are read closely by the UN).

This dataset contains records on every international import or export conducted with species from the CITES lists in 2016. It contains columns identifying the species, the import and export countries, and the amount and characteristics of the goods being traded (which range from live animals to skins and cadavers).

**https://www.kaggle.com/new­york­city/nyc­citywide­payroll­data (https://www.kaggle.com/new­**

**york­city/nyc­citywide­payroll­data)**

This dataset contains the salary, pay rate, and total compensation of every New York City employee. In this dataset this information is provided for the 2014, 2015, 2016, and 2017 fiscal years and provides a transparent lens into who gets paid how much and for what.

Note that fiscal years in the New York City budget cycle start on July 1st and end on June 30th (see here: (http://www1.nyc.gov/site/omb/about/new­york­city­budget­cycle.page) ). That means that this dataset contains, in its sum, compensation information for all City of New York employees for the period July 1, 2014 to June 30, 2017.

**https://www.kaggle.com/jonasalmeida/2015­deidentified­ny­inpatient­discharge­sparcs (https://www.kaggle.com/jonasalmeida/2015­deidentified­ny­inpatient­discharge­sparcs)**

2015 de­identified NY inpatient discharge (SPARCS)

This is the public dataset made available at https://health.data.ny.gov/Health/Hospital­Inpatient­ Discharges­SPARCS­De­Identified/82xm­y6g8 (https://health.data.ny.gov/Health/Hospital­Inpatient­ Discharges­SPARCS­De­Identified/82xm­y6g8) by the Dept of Health of New York state. The following description can be found at that page:

*The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De­identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. The de­identified data file does not contain data that is protected health information (PHI) under HIPAA. The health information is not individually identifiable; all data elements considered identifiable have been redacted. For example, the direct identifiers regarding a date have the day and month portion of the date removed.*

**https://www.kaggle.com/lantanacamara/hong­kong­horse­racing (https://www.kaggle.com/lantanacamara/hong­kong­horse­racing)**

The dataset contains the race result of 1561 local races throughout the Hong Kong racing seasons 2014­-2016 and more information will be added into the dataset. The dataset is divided into two tables (which can be joined by race\_id).

So, can you find any pattern for the winner under some condition? Did you spot out a winning strategy? (FYI, betting on all horses equally will bring a loss of ~17.5% on average) Which jockey/trainer is worth following?

**https://www.kaggle.com/cdc/mortality (https://www.kaggle.com/cdc/mortality)**

Every year the **CDC (http://www.cdc.gov/)** releases the country’s most detailed report on death in the United States under the **National Vital Statistics Systems (http://www.cdc.gov/nchs/nvss.htm)** . This **mortality dataset (http://www.cdc.gov/nchs/deaths.htm)** is a record of every death in the country for the year 2014, which includes detailed information about the causes of death and demographic background of the deceased.

It's been said that "statistics are human beings with the tears wiped off." This is especially true with this dataset. Each death record represents somebody's loved one, often connected with a lifetime of memories and sometimes tragically too short.

Putting the sensitive nature of the topic aside, analysing mortality data is essential to understanding the complex circumstances of death across the country. The US Government uses this data to determine life expectancy and understand how death in the U.S. differs from the rest of the world. Whether you’re looking for macro trends or analyzing unique circumstances, we challenge you to use this dataset to find your own answers to one of life’s great mysteries.

**https://www.kaggle.com/open­source­sports/professional­hockey­database (https://www.kaggle.com/open­source­sports/professional­hockey­database)**

The Hockey Database is a collection of historical statistics from men's professional hockey teams in North America.

**https://www.kaggle.com/faa/wildlife­strikes (https://www.kaggle.com/faa/wildlife­strikes)**

The dataset contains a record of each reported wildlife strike of a military, commercial, or civil aircraft between 1990 and 2015. Each row contains the incident date, aircraft operator, aircraft make and model, engine make and model, airport name and location, species name and quantity, and aircraft damage.