1. **1000 movies data.**

About this file

Here's a data set of 1,000 most popular movies on IMDB in the last 10 years. The data fields included are:

Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascrore

Rank

Movie rank order

Title

The title of the film

Genre

A comma-separated list of genres used to classify the film

Description

Brief one-sentence movie summary

Director

The name of the film's director

Actors

A comma-separated list of the main stars of the film

Year

The year that the film released as an integer.

Runtime (Minutes)

The duration of the film in minutes.

Rating

User rating for the movie 0-10

Votes

Number of votes

Revenue (Millions)

Movie revenue in millions

Metascore

An aggregated average of critic scores. Values are between 0 and 100. Higher scores represent positive reviews.

**2. Breast Cancer survival**

**Data Set Information:**

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

**Attribute Information:**

1. Age of patient at time of operation (numerical)   
2. Patient's year of operation (year - 1900, numerical)   
3. Number of positive axillary nodes detected (numerical)   
4. Survival status (class attribute)   
-- 1 = the patient survived 5 years or longer   
-- 2 = the patient died within 5 year

**3. Fast Food Restaurants**

This dataset is a list of 10,000 fast food restaurants from Datafiniti's Business Database updated between April 2018 and June 2018. Each business listing includes a variation of the phrase Fast Food within the Category field. All fields within this dataset have been flattened to streamline your data analysis. This version is a sample of a large dataset. The full dataset is available through Datafiniti.

Columns

id

dateAdded

dateUpdated

address

categories

city

country

keys

latitude

longitude

name

postalCode

province

sourceURLs

websites

**4. boxing matches**

Various information about a lot of **boxing matches**

**5. deliveries.csv**

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, non-striker, runs scored, etc.

Columns

match\_id

inning

Tells if the first set of batting was going on or second. 1: First Innings 2: Second Innings

batting\_team

The team name which is currently batting.

bowling\_team

The team name which is currently bowling.

over

Describe the current over number.

ball

Describe the current bowl no of the current over.

batsman

Name of the batsman on striking end.

non\_striker

Name of the batsman on non-striking end.

bowler

is\_super\_over

wide\_runs

bye\_runs

legbye\_runs

noball\_runs

penalty\_runs

batsman\_runs

extra\_runs

total\_runs

player\_dismissed

dismissal\_kind

fielder

**6. matches.csv**

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, non-striker, runs scored, etc.

Columns

id

season

city

date

team1

team2

toss\_winner

toss\_decision

result

dl\_applied

Duckworth Lewis method

winner

win\_by\_runs

win\_by\_wickets

player\_of\_match

venue

umpire1

umpire2

umpire3

**7. car sales.csv**

# Context

This dataset was collected by me from car sale advertisements for study/practice purposes in 2016. Though there is couple well known car features datasets they seems quite simple and outdated. Car topic is really interesting. But I wanted to practice with real raw data which has all inconvenient moments (as NA’s for example).

This dataset contains data for more than 9.5K cars sale in Ukraine. Most of them are used cars so it opens the possibility to analyze features related to car operation. At the end of the day I look at this data as a subset from all Ukrainian car fleet.

# Content

Dataset contains 9576 rows and 10 variables with essential meanings:

* car: manufacturer brand
* price: seller’s price in advertisement (in USD)
* body: car body type
* mileage: as mentioned in advertisement (‘000 Km)
* engV: rounded engine volume (‘000 cubic cm)
* engType: type of fuel (“Other” in this case should be treated as NA)
* registration: whether car registered in Ukraine or not
* year: year of production
* model: specific model name
* drive: drive type

Data has gaps, so be careful and check for NA’s. I tried to check and drop repeated offers, but theoretically duplications are possible.

**8. summar olympics**

Data Dictoinary is included in the same github repo with the name “olympics dictionary. csv”