**Kindly ignore the file DecisonTree.csv**

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”

**9. NYC\_Flight\_Data.csv**

On-time data for all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013

**year,month,day**

Date of departure

**dep\_time,arr\_time**

Actual departure and arrival times (format HHMM or HMM), local tz.

**sched\_dep\_time,sched\_arr\_time**

Scheduled departure and arrival times (format HHMM or HMM),local tz.

**dep\_delay,arr\_delay**

Departure and arrival delays, in minutes. Negative times represent early

departures/arrivals.

**hour,minute**

Time of scheduled departure broken into hour and minutes.

**carrier**

Two letter carrier abbreviation

**tailnum**

Plane tail number

**flight**

Flight number

**origin,dest**

Origin and destination.

**air\_time**

Amount of time spent in the air, in minutes

**distance**

Distance between airports, in miles

**time\_hour**

Scheduled date and hour of the flight as a POSIXct date

**10. European\_Premier\_League.csv**

This dataset contains results from every Premier League match from 1993-1994 to 2017-2018. It also includes half time results, but only from 1995-96 to 2017-18. Columns include Division (denoted as E0), HomeTeam, AwayTeam, FTHG (final time home goals), FTAG (final time away goals), FTR (full time result), HTHG (half time home goals), HTAG (half time away goals), HTR (half time result), and season.

Div

The division the match was played in. Because these matches are all from the same division, they are denoted as E0.

Date

The date the match was played.

HomeTeam

The name of the home team.

AwayTeam

The name of the away team.

FTHG

The total number of goals scored by the home team during the match at full time.

FTAG

The total number of goals scored by the away team during the match at half time.

FTR

The full time result, denoted as 'H' for home team win, 'A' for away team win, or 'D' for draw,

HTHG

The total number of goals scored by the home team at half time.

HTAG

The total number of goals scored by the away team at half time.

HTR

The half time result, denoted 'H' for home team advantage, 'A' for away team advantage, or 'D' for draw.

Season

The season in which the match was played.

**11. winequality.csv**

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine.

fixed acidity

most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

volatile acidity

the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

citric acid

found in small quantities, citric acid can add 'freshness' and flavor to wines

residual sugar

the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

chlorides

the amount of salt in the wine

free sulfur dioxide

the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide

amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

density

the density of water is close to that of water depending on the percent alcohol and sugar content

pH

describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

sulphates

a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

alcohol

the percent alcohol content of the wine

quality

output variable (based on sensory data, score between 0 and 10)

**12. Automobile\_data.csv**

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

Note: Several of the attributes in the database could be used as a "class" attribute.

The variables are self-explanatory.

symboling

rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling"

* 1. symboling: -3, -2, -1, 0, 1, 2, 3.   
     2. normalized-losses: continuous from 65 to 256.   
     3. make:   
     alfa-romero, audi, bmw, chevrolet, dodge, honda,   
     isuzu, jaguar, mazda, mercedes-benz, mercury,   
     mitsubishi, nissan, peugot, plymouth, porsche,   
     renault, saab, subaru, toyota, volkswagen, volvo   
       
     4. fuel-type: diesel, gas.   
     5. aspiration: std, turbo.   
     6. num-of-doors: four, two.   
     7. body-style: hardtop, wagon, sedan, hatchback, convertible.   
     8. drive-wheels: 4wd, fwd, rwd.   
     9. engine-location: front, rear.   
     10. wheel-base: continuous from 86.6 120.9.   
     11. length: continuous from 141.1 to 208.1.   
     12. width: continuous from 60.3 to 72.3.   
     13. height: continuous from 47.8 to 59.8.   
     14. curb-weight: continuous from 1488 to 4066.   
     15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.   
     16. num-of-cylinders: eight, five, four, six, three, twelve, two.   
     17. engine-size: continuous from 61 to 326.   
     18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.   
     19. bore: continuous from 2.54 to 3.94.   
     20. stroke: continuous from 2.07 to 4.17.   
     21. compression-ratio: continuous from 7 to 23.   
     22. horsepower: continuous from 48 to 288.   
     23. peak-rpm: continuous from 4150 to 6600.   
     24. city-mpg: continuous from 13 to 49.   
     25. highway-mpg: continuous from 16 to 54.   
     26. price: continuous from 5118 to 45400.

**13. facebook\_data.csv**

This dataset contains 99903 entries with 15 columns. Column names are well defined so that everyone can interpret easily.