

DATA CLEANING & DATA MANIPULATION

PETRA ISENBERG

VISUAL ANALYTICS

WHAT IS “DIRTY DATA”?

BEFORE WE CAN TALK ABOUT CLEANING, WE NEED TO KNOW ABOUT TYPES OF ERROR AND WHERE THEY COME FROM

SOURCES OF ERROR

DATA ENTRY ERRORS

MEASUREMENT ERRORS

DISTILLATION ERRORS

DATA INTEGRATION ERRORS

DATA ENTRY ERROR

LOTS OF DATA IS
ENTERED BY HAND

TYPOGRAPHIC ERRORS

MISUNDERSTANDING
DATA OR CONVENTIONS

“SPURIOUS INTEGRITY”

“SPURIOUS INTEGRITY”

ENTERING BAD DATA IN RESPONSE TO (OFTEN WELL-INTENTIONED) INTERFACE CONSTRAINTS

“SPURIOUS INTEGRITY”

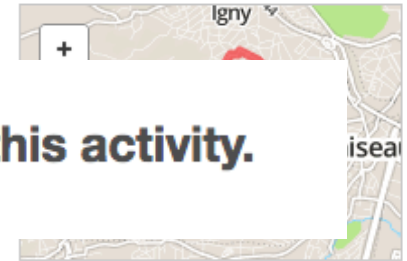
Step 1: Activity/Equipment Type

Step 2: Add a Map

Step 3: Additional Details

Add An Activity

Activity Details



Activity Type:	Running
Equipment Type:	None
Route:	None
Distance:	5.62 mi.
Duration:	--:--

Date of Activity:

Duration:

< September 2014 >

Su	M	T	W	Th	F	Sa
7						
14	1					
21	22	23	24	25	26	27
28	29	30				

00 : 00 : 00



Oops! You forgot to enter a duration for this activity.

5.62 mi

Training Plan:

None

Average Heart Rate (optional):

bpm

MEASUREMENT ERRORS

SENSOR ISSUES

MALFUNCTIONS

PLACEMENT

INTERFERENCE

MISCALIBRATION



DISTILLATION ERRORS

SOME DATA MAY BE LOST OR COMPRESSED
BEFORE IT ENTERS
THE DATABASE

0.345413 → 0.35

National Price Index → NPI

1985, \$2, Apples

1985, \$2, Oranges → 1985, \$2, "Apples, Oranges, Cucumbers"

1985, \$2, Cucumbers

DATA INTEGRATION ERRORS

DATA OFTEN COMES FROM MULTIPLE SOURCES

SCHEMAS CHANGE OVER TIME

DATA IS OFTEN COERCED FROM
ONE TYPE TO ANOTHER

**CAN LEAD TO DATA LOSS,
DUPLICATION, AND OTHER**

WHY IS THIS IMPORTANT?

**MOST OF THE TIME IN THE DATA
ANALYSIS PROCESS IS ACTUALLY
SPENT HERE!**

“I spend more than half my time integrating, cleansing, and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

[Kandel 2012]

SOME DATA QUALITY ISSUES

MISSING DATA

MISSED MEASUREMENTS, REDACTED ITEMS, INCOMPLETE FORMS, ETC.

ERRONEOUS VALUES

MISPELLINGS, OUTLIERS, "SPURIOUS INTEGRITY", ETC.

ENTITY RESOLUTION

DIFFERENT VALUES, ABBREVS., 2+ ENTRIES FOR THE SAME THING?

TYPE CONVERSION

E.G., ZIP CODE OR PLACE NAME TO LAT-LON

DATA INTEGRATION

MISMATCHES AND INCONSISTENCIES WHEN COMBINING DATA

DETECTING ERRORS

LOOK FOR OUTLIERS / ANOMALIES

EXAMINE DATA TYPES

SCHEMA CHECKING

VALIDATE WITH OTHER DATA

OTHER HEURISTICS

HISTORICALLY – MORE FOCUS ON AUTOMATED APPROACHES

DETECTION METHODS

+ CAN IDENTIFY POTENTIAL ANOMALIES

- HARD TO KNOW IF THEY'RE REALLY ANOMALOUS OR HOW TO CORRECT THEM

Type	Issue	Detection Method(s)
Missing	Missing record	Outlier Detection Residuals then Moving Average w/ Hampel X84
		Frequency Outlier Detection Hampel X84
Inconsistent	Missing value	Find NULL/empty values
	Measurement units	Clustering Euclidean Distance
		Outlier Detection z-score, Hampel X84
	Misspelling	Clustering Levenshtein Distance
	Ordering	Clustering Atomic Strings
	Representation	Clustering Structure Extraction
	Special characters	Clustering Structure Extraction
Incorrect	Erroneous entry	Outlier Detection z-score, Hampel X84
	Extraneous data	Type Verification Function
	Misfielded	Type Verification Function
	Wrong physical data type	Type Verification Function
Extreme	Numeric outliers	Outlier Detection z-score, Hampel X84, Mahalanobis distance
	Time-series outliers	Outlier Detection Residuals vs. Moving Average then Hampel X84
Schema	Primary key violation	Frequency Outlier Detection Unique Value Ratio

MISSING AND IMPOSSIBLE VALUES

1. LOOK AT EMPTY/MISSING VALUES
2. LOOK AT IMPOSSIBLE VALUES

Gender = 3

Heart Rate = 0

Unlikely Dates (e.g. "01/01/0001")

JUST SORTING THE DATA CAN HELP HIGHLIGHT ISSUES LIKE THESE

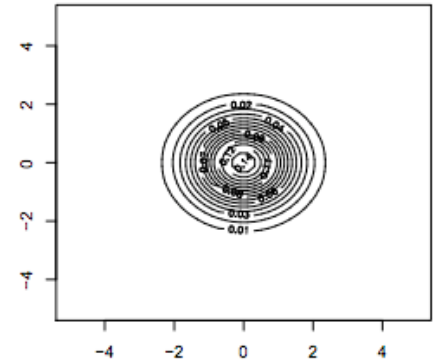
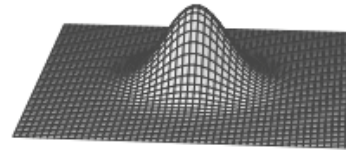
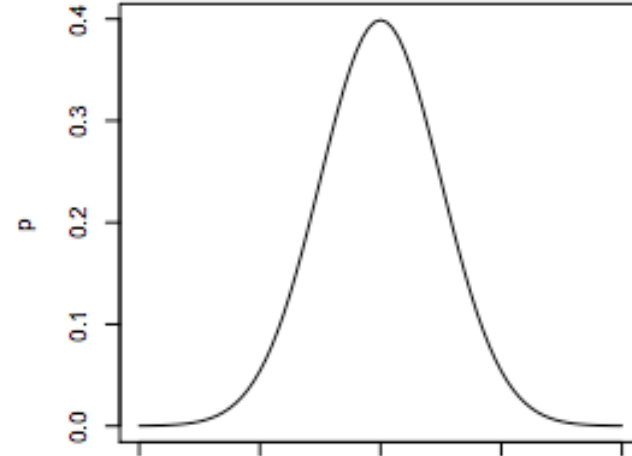
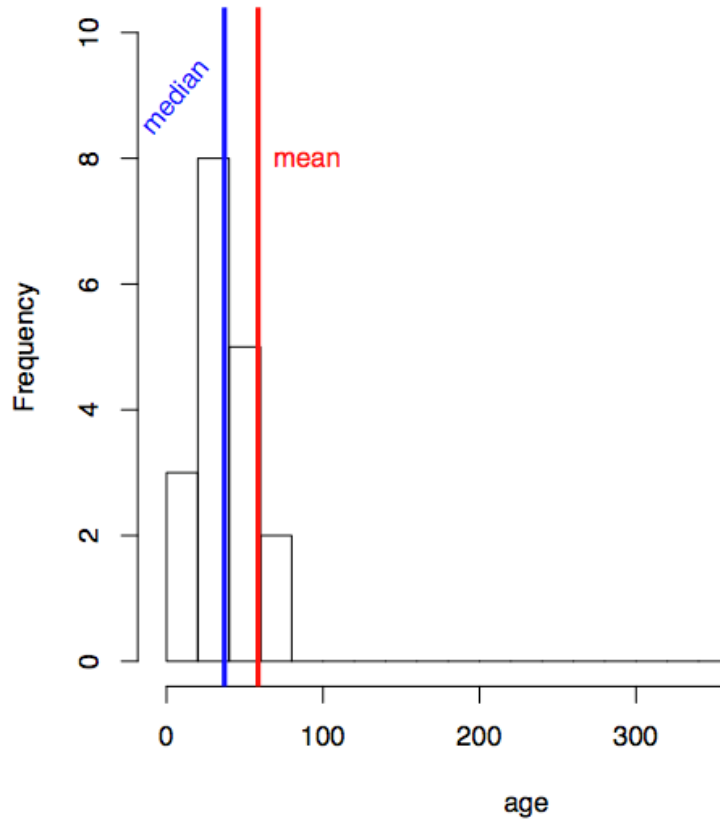
OUTLIER DETECTION

1. EXAMINE DISTRIBUTIONS
2. MODEL DATA AND LOOK FOR RESIDUALS
3. PARTITION DATA

FOR **ONE DATA DIMENSION** OR **MULTIPLE DIMENSIONS**

EXAMINE DISTRIBUTIONS

Histogram of age



DETECTING DUPLICATES

Title

Ben-Hur

Ben Hur

BEN-HUR

Ben-Hur (1959 film)

Name

Anand Vaskar

Anand Vaskkar

A. Vaskar

Vaskar, Anand

THESE MIGHT ALL BE THE SAME

SOME USEFUL DISTANCE METRICS

LEVENSHTEIN (“STRING-EDIT”) DISTANCE

How many edits do I need to change one value into another?

Ben-IHur

Ben Hur

DISTANCE = 1

Anand Vaskar

Anand Vaskkar

DISTANCE = 1

SOME USEFUL DISTANCE METRICS

LEVENSHTEIN (“STRING-EDIT”) DISTANCE

How many edits do I need to change one value into another?

Ben-Hur

Ben-Hur (1959 film)

DISTANCE = 12

Anand Vaskar

Vaskar, Anand

DISTANCE = 12

SOME USEFUL DISTANCE METRICS

SOUNDEX / METAPHONE

How similar do they sound?

Ben-Hur

Ben-Hurr

Been Her

Anand Vaskar

Anand Vaskkar

Ahnund Vachkar

SOME USEFUL DISTANCE METRICS

“FINGERPRINTING” METHODS

Strip away unimportant details.

(e.g., remove punctuation, capitals, and sort)

Anand Vaskar → anand vaskar

Vaskar, Anand → anand vaskar

AND MANY MORE

STRING/KEY COMPARISONS

DISTANCE METRICS FOR NUMERIC DATA

e.g., HAMPEL X84 (UNIVARIATE), MAHALANOBIS (MULTIVARIATE)

“Quantitative Data Cleaning for Large Databases”

Hellerstein (2008)

Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein^{*}
EECS Computer Science Division
UC Berkeley
<http://dfti.cs.berkeley.edu/jmh>

February 27, 2008

1 Introduction

Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the raison d’être of entire agencies or firms.

Despite the importance of data collection and analysis, data quality remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or incomplete data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decade on various aspects of data cleaning: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in quantitative attributes of large databases, though we also provide references to data cleaning methods for other types of attributes. The discussion is targeted at computer practitioners who manage large databases of quantitative information, and designers developing data entry and auditing tools for end users. Because of our focus on quantitative data, we take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in robust statistics (Rousseeuw and Leroy 1987; Hampel et al. 1986; Huber 1981). In addition, we stress algorithms and implementations that can be easily and efficiently implemented in very large databases, and which are easy to understand and visualize graphically. The discussion covers statistical notations and methods, algorithmic building blocks, efficient relational database implementation strategies, and user interface considerations. Throughout the discussion, references are provided for deeper reading on all of these issues.

1.1 Sources of Error in Data

Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate

^{*}This survey was written under contract to the United States Economic Commission for Europe (UNICEE), which holds the copyright on this version.

DECIDING HOW TO FIX PROBLEMS

YOU CAN DO ALMOST ALL OF
THIS IN **SQL** ... BUT IT'S A LOT OF WORK

DECIDING HOW TO FIX PROBLEMS

WHICH DUPLICATE TO KEEP?

OUTLIERS: KEEP, REMOVE, OR REPAIR?

BADLY-STORED DATES, ADDRESSES, OR KEYS MAY
NEED TO BE PARSED MANUALLY

DECIDING HOW TO FIX PROBLEMS

FUZZY MATCHING SYSTEMS

MACHINE LEARNING TO DETECT/RESOLVE
ERRORS

**USUALLY REQUIRES HUMAN JUDGMENT
(ESPECIALLY FOR NEW DATA)**

INTERACTIVE PROFILING

Schema Browser

- Creative Type
- Distributor
- IMDB Rating
- IMDB Votes
- MPAA Rating
- Major Genre
- Production Budget

Related Views:

Anomalies

Anomaly Browser

Missing (6)

MPAA Rating

Creative Type

Source

Major Genre

Distributor

Release Location

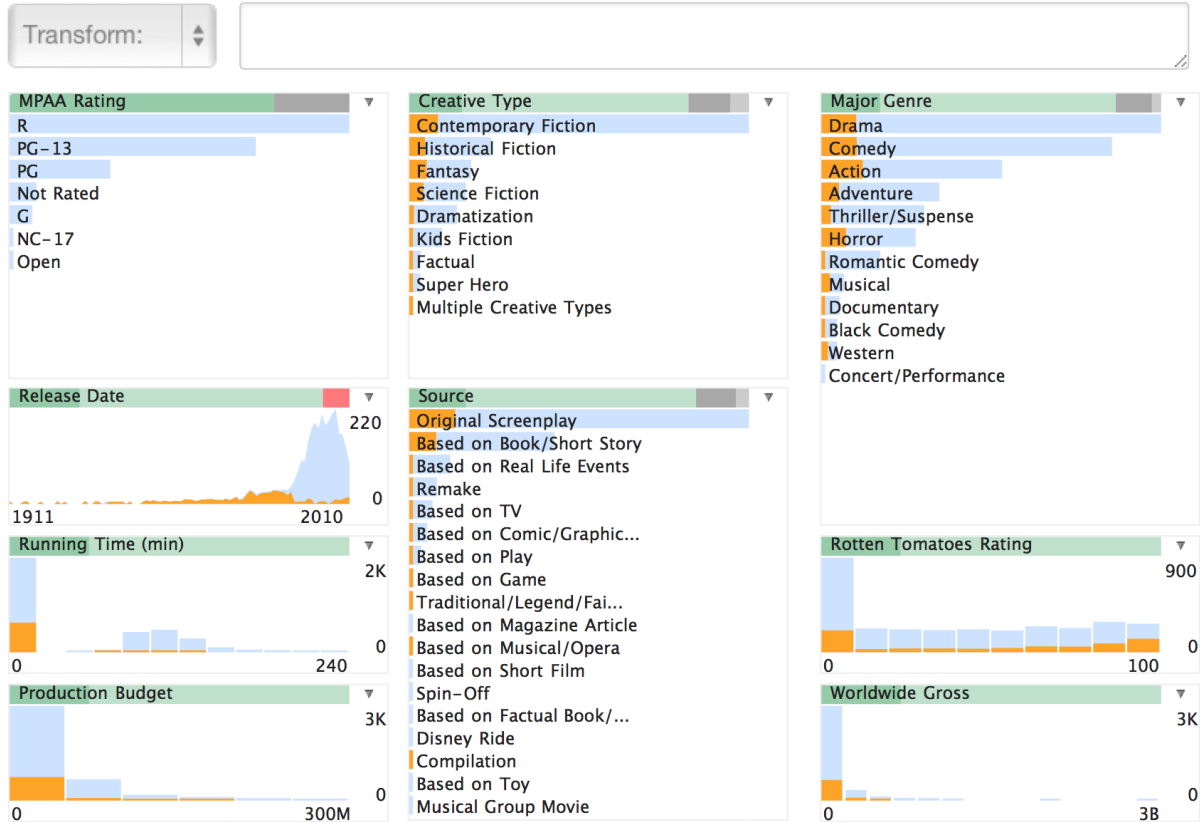
Error (2)

Extreme (7)

Inconsistent (3)

Distributor (Levenshtein)

Source (Levenshtein)



“PROFILING” DATA

UNDERSTANDING WHAT ASSUMPTIONS YOU CAN
MAKE ABOUT DATA

INTERACTIVELY IDENTIFYING
DATA QUALITY ISSUES

AN EXAMPLE

IMDb Find Movies, TV shows, Celebrities and more

Movies TV News Videos

Rotten Tomatoes by Flixster

Search movies, actors, critics

MOVIES

Now Playing

In 6 theaters near San Francisco, CA. [Change location](#)

The Hunger Games

142 min - Action

Your rating: **7.6**

Ratings: 7.6/10 from 1,170 reviews

Set in a future where the twelve districts to fight Katniss Everdeen volunteered place for the latest match

Director: [Gary Ross](#)

Writers: [Gary Ross](#) (screenplay), and [2 more](#)

Stars: [Jennifer Lawrence](#), [Hemsworth](#)

[Watch Trailer](#) + [W](#)

96 photos | 23 videos | 9081 news articles | full cast & crew

7 nominations [See more awards](#)

Related Videos

Music Video: [The Hunger Games](#)

Trailer: [The Hunger Games](#)

[See all 23](#)

People who liked this also liked...

[Help](#)

[The Ides of March](#)

[The Way, Way Back](#)

[Help](#)

The Hunger Games (2012)

TOMATOMETER **84%** (All Critics | Top Critics)

Average Rating: 7.2/10
Reviews Counted: 257
Fresh: 217 | Rotten: 40

Thinking and superbly acted, The Hunger Games captures the dire violence, raw emotion, and amoral scope of its source novel.

MY RATING

WANT TO SEE IT | NOT INTERESTED

Add a Review (Optional)

MOVIE INFO

Every year in the ruins of what was once North America, 1 Panem forces each of its twelve districts to send a teenage Hunger Games. A twisted punishment for a past uprising, it is a national television spectacle, where one survivor remains. Pitted against one another until only one survivor remains. Pitted against one another until only one survivor remains. Pitted against one another until only one survivor remains.

PG-13, 2 hr. 22 min. In Theaters
Drama, Mystery & Suspense, Science Fiction | Box Office & Fantasy | Lionsgate

Directed By: Gary Ross
Written By: Suzanne Collins, Gary Ross, Billy Ray

Friend Ratings

March 27, 2012

Jon Whetstone

The Hunger Games Trailer & Photos

More Photos (39) | More Trailers (4)

The Hunger Games
Aug 29, 2011

Cast

Jennifer Lawrence
Katniss Everdeen

Liam Hemsworth

Josh Hutcherson
Peeta Mellark

Woody Harris

THE NUMBERS

BOX OFFICE DATA, MOVIE STARS, IDLE SPECULATION

Learn About Our Research and Data Services

Wednesday, May 16, 2012

Search [Go](#)

News

Latest News
Coming Soon
Trailers
Theater Counts
Oscars

The Movies

Daily Chart
Weekly Sales Chart
Weekend Chart
Archive
Records
Top Rated
Popular
Budgets
Franchises
Keywords

Home Market

DVD Sales Chart
Blu-ray Sales Chart
2012 DVD Chart
2011 DVD Chart
2010 DVD Chart
Coming Soon
Archive

Market Analysis

Overview
2012
2011
2010
Distributors
Genres
MPAA Ratings
Sources
Prod'n Methods
Creative Types

International

World Chart
News
Records

The People

Star Archive
Records
Directors
Popular

The Speculation

HSX Analysis

The Site

Forums
Site Map
About Us
The Crunch
Data Feeds
Home Page

amazon.com

Movies & TV on DVD and Blu-ray
Save up to 40% on Bestsellers

[Show now](#)

save May in sales event! It's going on now!

Great deals available at your Toyota dealer.

TOYOTA moving forward

Ready to Buy

The Hunger Games

The Numbers Rating: 6.88 (24 votes) [Rate It](#) - [Rating Details](#)

Rotten Tomatoes Rating: **84%** - [Fresh!](#)

Theatrical Performance	
Domestic Box Office	\$387,007,048
International Box Office	\$131,600,000
Worldwide Box Office	\$518,607,048

[For full financial breakdown, please contact our research team.](#)

Released March 23, 2012 (Wide)

Production Budget \$80,000,000

MPAA Rating **PG-13** for intense violent thematic material and disturbing images - all involving teens.

Domestic Marketing Budget Source: \$45 million ([N.Y. Times](#))

Highest Combined Star Gross 139 ([see full chart](#))

Keywords [Lionsgate](#)

Distributed by Source [Based on Book/Short Story](#)

Major Genre [Thriller/Suspense](#)

Production Method [Live Action](#)

Creative Type [Science Fiction](#)

News (See All...)

2012-05-15 Weekend Wrap-Up: Avengers Begin New Century Club

2012-05-10 Weekend Predictions: Avengers Overshadows New Releases

2012-05-07 Weekend Wrap-up: Avengers Assemble a New Record Book

2012-05-03 Weekend Predictions: Will Box Office Records Be Avenged?

2012-05-03 International Box Office: Avengers are Marvelous

2012-04-30 Weekend Wrap-Up: The Box Office Will Be Avenged

2012-04-29 Weekend Estimates: Think Like a Man Rises Above the Pack

2012-04-26 Weekend Predictions: Seven-Day Engagement

2012-04-26 International Box Office: Battle on the High Seas

2012-04-23 Weekend Wrap-Up: Moviegoers were Very Thoughtful

[Submit news for this movie](#)

Trailer

[More trailers...](#)

THE 2012 Mazda3

Starting at \$15,200*

IF IT'S NOT WORTH DRIVING, IT'S NOT WORTH BUILDING.

[EXPLORE NOW](#) [MazdaUSA.com](#)

Title	Release Date	MPAA Rating	Distributor	Rotten Tomatoes Rating	IMDB Rating
The Land Girls	Jun 12, 1998	R	Gramercy		6.1
First Love, Last Rites	Aug 7, 1998	R	Strand		6.9
I Married a Strange Person	Aug 28, 1998		Lionsgate		6.8
Slam	Oct 9, 1998	R	Trimark	62	3.4
Mississippi Mermaid	Jan 15, 1999		MGM		
Following	Apr 4, 1999	R	Zeitgeist		7.7
Foolish	Apr 9, 1999	R	Artisan		3.8
Pirates	Jul 1, 1986	R		25	5.8
Duel in the Sun	Dec 31, 2046			86	7
Tom Jones	Oct 7, 1963			81	7
Oliver!	Dec 11, 1968		Sony Pictures	84	7.5
To Kill A Mockingbird	Dec 25, 1962		Universal	97	8.4
Tora, Tora, Tora	Sep 23, 1970				
Hollywood Shuffle	Mar 1, 1987			87	6.8
Over the Hill to the Poorhouse	Sep 17, 2020				
Wilson	Aug 1, 2044				7
Darling Lili	Jan 1, 1970				6.1
The Ten Commandments	Oct 5, 1956			90	2.5
12 Angry Men	Apr 13, 1957		United Artists		8.9
Twelve Monkeys	Dec 27, 1995	R	Universal		8.1
1776	Nov 9, 1972	PG	Sony/ Columbia	57	7

Title	Release Date	MPAA Rating	Distributor	Rotten Tomatoes Rating	IMDB Rating
The Land Girls	Jun 12, 1998	R	Gramercy		6.1
First Love, Last Rites	Aug 7, 1998	R	Strand		6.9
I Married a Strange Person	Aug 28, 1998		Lionsgate		6.8
Slam	Oct 9, 1998	R	Trimark	62	3.4
Mississippi Mermaid	Jan 15, 1999		MGM		
Following	Apr 4, 1999	R	Zeitgeist		7.7
Foolish	Apr 9, 1999	R	Artisan		3.8
Pirates	Jul 1, 1986	R		25	5.8
Duel in the Sun	Dec 31, 2046			86	7
Tom Jones	Oct 7, 1963			81	7
Oliver!	Dec 11, 1968		Sony Pictures	84	7.5
To Kill A Mockingbird	Dec 25, 1962		Universal	97	8.4
Tora, Tora, Tora	Sep 23, 1970				
Hollywood Shuffle	Mar 1, 1987			87	6.8
Over the Hill to the Poorhouse	Sep 17, 2020				
Wilson	Aug 1, 2044				7
Darling Lili	Jan 1, 1970				6.1
The Ten Commandments	Oct 5, 1956			90	2.5
12 Angry Men	Apr 13, 1957		United Artists		8.9
Twelve Monkeys	Dec 27, 1995	R	Universal		8.1
1776	Nov 9, 1972	PG	Sony/ Columbia	57	7

Arnolds Park	Oct 19, 2007	PG-13	The Movie Partners
Sweet Sweetback's Baad Asssss Song	Jan 1, 1971		
And Then Came Love	Jun 1, 2007	Not Rated	Fox Meadow
Around the World in 80 Days	Oct 17, 1956	PG	United Artists
Barbarella	Oct 10, 1968		Paramount Pictures
Barry Lyndon	1975		Warner Bros.
Barbarians, The	March, 1987		
Babe	Aug 4, 1995	G	Universal
Boynton Beach Club	Mar 24, 2006	R	Wingate Distribution
Baby's Day Out	Jul 1, 1994	PG	20th Century

Bad Boys	Apr 7, 1995	6.6	53929
Body Double	Oct 26, 1984	6.4	9738
The Beast from 20,000 Fathoms	Jun 13, 1953		
Beastmaster 2: Through the Portal of Time	Aug 30, 1991	3.3	1327
The Beastmaster	Aug 20, 1982	5.7	5734
Ben-Hur	Dec 30, 2025	8.2	58510
Ben-Hur	Nov 18, 1959	8.2	58510
Benji	Nov 15, 1974	5.8	1801
Before Sunrise	Jan 27, 1995	8	39705

PROFILING IN OPEN REFINE

Movies Analysis - Google Refine

127.0.0.1:3333/project?project=1615121211153

Google refine Movies Analysis Permalink

Open... Export Help

Facet / Filter Undo / Redo 7

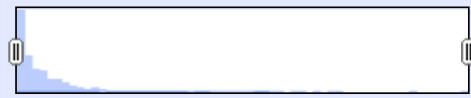
Refresh Reset All Remove All

69 matching records (2448 total) Extensions: Freebase

Show as: rows records Show: 5 10 25 50 records « first < previous 1 - 10 next > last »

All	Title	ReleaseDate	USGross	MPAARating	WorldwideGross	USI
6.	Doogal	2006-02-24T00:00:00Z	7578946	G	26942802	
116.	Beauty and the Beast	1991-11-13T00:00:00Z	171340294	G	403476931	
142.	Aladdin	1992-11-11T00:00:00Z	217350219	G	504050219	
200.	The Lion King	1994-06-15T00:00:00Z	328539505	G	783839505	
255.	Pocahontas	1995-06-10T00:00:00Z	141579773	G	347100000	
268.	Babe	1995-08-04T00:00:00Z	63658910	G	246100000	
273.	The	1995-08-	669276	G	669276	

USGross change reset

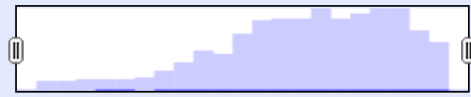


0.00 — 610,000,000.00

Numeric Non-numeric Blank Error

69 0 0 0

ReleaseDate change reset



1987-02-20 00:00:00 — 00:00:00

INTERACTIVE DATA CLEANING



Trifacta Wrangler

<https://www.trifacta.com/>



Wrangler (Stanford HCI Group)

<http://vis.stanford.edu/wrangler/>



OpenRefine (formerly Google Refine)

<http://openrefine.org/>

DATA CLEANING IN GOOGLE REFINE

The screenshot displays the Google Refine interface for a project named "Movies Analysis". The browser address bar shows "127.0.0.1:3333/project?". The main interface includes a "Facet / Filter" section on the left with a histogram for "USC" and a "ReleaseDate" facet. The central area shows "69 matching records" (448 total) in a table view. A yellow box labeled "FILTER" is overlaid on the "USC" facet, and another yellow box labeled "TRANSFORM" is overlaid on the table. Two large black curved arrows indicate a workflow from the filter to the transform step. The table contains columns for movie titles, release dates, and other identifiers.

Rank	Title	Release Date	Box Office	Genre	Other ID
6.	Doc...	24 100:00:00Z	171340294	G	403476931
116.	Beauty and the Beast	1991-11-30 00:00:00Z	217350219	G	504050219
149.		1994-06-15 00:00:00Z	328539505	G	783839505
255.	Pocahontas	1995-06-10 00:00:00Z	141579773	G	347100000
268.	Babe	1995-08-04 00:00:00Z	63658910	G	246100000
273.	The	1995-08-04 00:00:00Z	669276	G	669276

[Google Refine Intro Video](#)

REFERENCES

“Quantitative Data Cleaning for Large Databases”

Hellerstein (2008)

Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein*
EECS Computer Science Division
UC Berkeley
<http://db.cs.berkeley.edu/jmh>

February 27, 2008

1 Introduction

Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the *raison d’être* of entire agencies or firms.

Despite the importance of data collection and analysis, data *quality* remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decades on various aspects of *data cleaning*: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in *quantitative* attributes of large databases, though we also provide references to data cleaning methods for other types of attributes. The discussion is targeted at computer practitioners who manage large databases of quantitative information, and designers developing data entry and auditing tools for end users. Because of our focus on quantitative data, we take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in *robust statistics* [Rousseeuw and Leroy, 1987, Hampel et al., 1986, Huber, 1981]. In addition, we stress algorithms and implementations that can be easily and efficiently implemented in very large databases, and which are easy to understand and visualize graphically. The discussion mixes statistical intuitions and methods, algorithmic building blocks, efficient relational database implementation strategies, and user interface considerations. Throughout the discussion, references are provided for deeper reading on all of these issues.

1.1 Sources of Error in Data

Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate

*This survey was written under contract to the United Nations Economic Commission for Europe (UNECE), which holds the copyright on this version.

CSVKIT

🏠 csvkit

1.0.3

Search docs

Tutorial

Reference

Tips and Troubleshooting

Contributing to csvkit

Release process

License

Changelog



SMS API for web & mobile applications

Make and receive SMS messages in your applications with just a few lines of code.

Docs » csvkit 1.0.3

[Edit on GitHub](#)

csvkit 1.0.3

About

build **passing** **FIXME** **Migrate to GitLab** coverage **87%** pypi **v1.0.3** license **MIT**

python **2.7** | 3.3 | 3.4 | 3.5 | 3.6

csvkit is a suite of command-line tools for converting to and working with CSV, the king of tabular file formats.

It is inspired by pdftk, gdal and the original csvcut tool by Joe Germuska and Aaron Bycoffe.

If you need to do more complex data analysis than csvkit can handle, use [agate](#).

Important links:

- Repository: <https://github.com/wireservice/csvkit>