## cse 512 - Data Visualization Data Transformation



#### Jeffrey Heer University of Washington

#### **Session Outline**

Data Models

Data Tables & Transformations Data Wrangling & Profiling Visualizing Distributions Dimensionality Reduction



# **Data Models**

#### Data Models / Conceptual Models

**Data models** are formal descriptions Math: sets with operations on them Example: integers with + and x operators

**Conceptual models** are mental constructions Include semantics and support reasoning

Examples (data vs. conceptual)1D floats vs. temperatures3D vector of floats vs. spatial location

## **Types of Variables**

#### **Physical Types**

Characterized by storage format Characterized by machine operations *Example*: bool, int32, float, double, string, ...

#### **Abstract Types**

Provide descriptions of the data May be characterized by methods / attributes May be organized into a hierarchy *Example*: plants, animals, metazoans, ...

## Taxonomy of Data Types (?)

1D (sets and sequences) Temporal 2D (maps) 3D (shapes) nD (relational) Trees (hierarchies) Networks (graphs)

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

N - Nominal (labels or categories)

• Fruits: apples, oranges, ...

- N Nominal (labels or categories)
  - Fruits: apples, oranges, ...
- O Ordered
  - $\cdot\,$  Quality of meat: Grade A, AA, AAA

- N Nominal (labels or categories)
  - Fruits: apples, oranges, ...
- O Ordered
  - $\cdot$  Quality of meat: Grade A, AA, AAA
- Q Interval (location of zero arbitrary)
  - Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
  - $\cdot$  Only differences (i.e., intervals) may be compared

- N Nominal (labels or categories)
  - Fruits: apples, oranges, ...
- O Ordered
  - $\cdot$  Quality of meat: Grade A, AA, AAA
- Q Interval (location of zero arbitrary)
  - Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
  - $\cdot$  Only differences (i.e., intervals) may be compared
- Q Ratio (zero fixed)
  - Physical measurement: Length, Mass, Time duration, ...
  - $\cdot\,$  Counts and amounts

- N Nominal (labels or categories)
  - Operations: =,  $\neq$
- O Ordered
  - Operations: =,  $\neq$ , <, >
- Q Interval (location of zero arbitrary)
  - Operations: =,  $\neq$ , <, >, -
  - $\cdot$  Can measure distances or spans
- Q Ratio (zero fixed)
  - Operations: =,  $\neq$ , <, >, -, %
  - $\cdot\,$  Can measure ratios or proportions

#### From Data Model to N, O, Q

Data Model 32.5, 54.0, -17.3, ... Floating point numbers

**Conceptual Model** Temperature (°C)

**Data Type** Burned vs. Not-Burned (N) Hot, Warm, Cold (O) Temperature Value (Q-interval)

#### **Dimensions & Measures**

**Dimensions** (~ independent variables) Often discrete variables describing data (N, O) Categories, dates, binned quantities

Measures (~ dependent variables) Data values that can be aggregated (Q) Numbers to be analyzed Aggregate as sum, count, avg, std. dev...

Not a strict distinction. The same variable may be treated either way depending on the task.

# Example: U.S. Census Data

#### **Example: U.S. Census Data**

People Count: # of people in group
Year: 1850 – 2000 (every decade)
Age: 0 – 90+
Sex: Male, Female
Marital Status: Single, Married, Divorced, ...

## Example: U.S. Census

People Count Year Age Sex Marital Status

2,348 data points

	А	В	С	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

### Census: N, O, Q-Interval, Q-Ratio?

People Count	Q-Ratio
Year	Q-Interval (O)
Age	Q-Ratio (O)
Sex	Ν
Marital Status	Ν

#### **Census: Dimension or Measure?**

People Count Year Age Sex Marital Status Measure Dimension Depends! Dimension Dimension

# Census Data Demo

demo link: us-population-1850-2000

# Data Tables & Transformations

#### **Relational Data Model**

Represent data as a **table** (or *relation*) Each **row** (or *tuple*) represents a record Each record is a fixed-length tuple Each **column** (or *field*) represents a variable Each field has a *name* and a *data type* A table's **schema** is the set of names and types A **database** is a collection of tables (relations)

**Operations on Data Tables: table(s) in, table out** 

**Operations on Data Tables: table(s) in, table out** Project (select): select a set of columns Filter (where): remove unwanted rows Sort (order by): order records Aggregate (group by, sum, min, max, ...): partition rows into groups + summarize Combine (join, union, ...): integrate data from multiple tables

Project (select): select a set of columns
select day, stock

day	stock	price	day	stock
10/3	AMZN	957.10	10/3	AMZN
10/3	MSFT	74.26	10/3	MSFT
10/4	AMZN	965.45	10/4	AMZN
10/4	MSFT	74.69	10/4	MSFT

Filter (where): remove unwanted rows
select \* where price > 100

day	stock	price			
10/3	AMZN	957.10	day	stock	price
10/3	MSFT	74.26	10/3	AMZN	957.10
10/4	AMZN	965.45	10/4	AMZN	965.45
10/4	MSFT	74.69			

Sort (order by): order records
select \* order by stock

day	stock	price	day	stock	price
10/3	AMZN	957.10	10/3	AMZN	957.10
10/3	MSFT	74.26	10/4	AMZN	965.45
10/4	AMZN	965.45	10/3	MSFT	74.26
10/4	MSFT	74.69	10/4	MSFT	74.69

Aggregate (group by, sum, min, max, ...): select stock, min(price) group by stock

day	stock	price		
10/3	AMZN	957.10	stock	min(price)
10/3	MSFT	74.26	AMZN	957.10
10/4	AMZN	965.45	MSFT	74.26
10/4	MSFT	74.69		

#### Join (join) multiple tables together

		•				
day	stock	price	day	stock	price	min
10/3		957 10			•	
10/5		/3/.10	10/3	AMZN	957.10	957.10
10/2	NACET	71 74				
10/3		/4.20	10/3	MSFT	74.26	74.26
10/1						
10/4	AIVIZIN	705.45	10/4	AMZN	965.45	957.10
10/1						
10/4	IVISE I	/4.69	10/4	MSFT	74.69	74.26

stock	min
AMZN	957.10
MSFT	74.26

select t.day,t.stock,t.price,a.min
from table as t, aggregate as a
where t.stock = a.stock

## **Roll-Up and Drill-Down**

Want to examine population by year and age? **Roll-up** the data along the desired dimensions



## **Roll-Up and Drill-Down**

Want to see the breakdown by marital status? **Drill-down** into additional dimensions

SELECT year, age, marst, sum(people) FROM census GROUP BY year, age, marst





#### ORIGINAL

YEAR	AGE	MARST	SEX	PEOPLE
1850	0	0	1	1,483,789
1850	5	0	1	1,411,067
1860	0	0	1	2,120,846
1860	5	0	1	1,804,467

• • •

• • •

#### PIVOTED (or CROSS-TABULATION)

AGE	MARST	SEX	1850	1860	
0	0	1	1,483,789	2,120,846	
5	0	1	1,411,067	1,804,467	

Which format might we prefer? Why?

#### Tidy Data [Wickham 2014]

- How do rows, columns, and tables match up with observations, variables, and types? In "tidy" data:
- 1. Each variable forms a column.
- 2. Each observation forms a row.
- 3. Each type of observational unit forms a table.
- The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.
- Our pivoted table variant was not "tidy"!
- (This is a variant of <u>normalized forms</u> in DB theory)

#### **Common Data Formats**

#### CSV: Comma-Separated Values (d3.csv)

year,age,marst,sex,people 1850,0,0,1,1483789 1850,5,0,1,1411067

• • •
# **Common Data Formats**

### CSV: Comma-Separated Values (d3.csv)

year,age,marst,sex,people 1850,0,0,1,1483789 1850,5,0,1,1411067

. . .

### JSON: JavaScript Object Notation (d3.json)

L {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789}, {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067}, ...

# **Common Data Formats**

### CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people
1850,0,0,1,1483789
1850,5,0,1,1411067
```

. . .

### JSON: JavaScript Object Notation (d3.json)

```
L
{"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},
{"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},
...
]
```

### Binary Formats: Arrow, Parquet, ...

# **Data Wrangling**

I spend more than half of 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.

Anonymous Data Scientist from our 2012 interview study







# In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

★ 17 ★ ···

Bureau http://l	of Justice Stat bjs.ojp.usdoj.go	istics – Data Online DV/			
Reporte	d crime in Alaba	ama			
Year 2004 2005 2006 2007 2008	Population 4525375 4029.3 4548327 3900 4599030 3937 4627851 3974.9 4661900 4081.9	Property crime rate 987 2732.4 309.9 955.8 2656 289 968.9 2645.1 322.9 980.2 2687 307.7 1080.7 2712.6 288.6	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
Reporte	d crime in Alask	a			
Year 2004 2005 2006 2007 2008	Population 657755 3370.9 663253 3615 670053 3582 683478 3373.9 686293 2928.3	Property crime rate 573.6 2456.7 340.6 622.8 2601 391 615.2 2588.5 378.3 538.9 2480 355.1 470.9 2219.9 237.5	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
Reporte	d crime in Arizo	ona			
Year 2004 2005 2006 2007 2008	Population 5739879 5073.3 5953007 4827 6166318 4741.6 6338755 4502.6 6500180 4087.3	Property crime rate 991 3118.7 963.5 946.2 2958 922 953 2874.1 914.4 935.4 2780.5 786.7 894.2 2605.3 587.8	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
Reporte	d crime in Arkar	nsas			
Year 2004 2005 2006 2007 2008	Population 2750000 4033.1 2775708 4068 2810872 4021.6 2834797 3945.5 2855390 3843.7	Property crime rate 1096.4 2699.7 237 1085.1 2720 262 1154.4 2596.7 270.4 1124.4 2574.6 246.5 1182.7 2433.4 227.6	Burglary rate	Larceny-theft rate	Motor vehicle theft rate

### **Data**Wrangler

Suggestions	rows: 408 prev next	
	H Year	♦ 📆 Property_crime_rate ♦
Delete rows 8,10	1 Reported crime in Alabama	
Delete empty rows	3 2004	4029.3
Delete rows where Property_crime_rate	5 2006	3937
is null	6 2007 7 2008	3974.9 4081.9
Delete rows where Year is null	8 9 Reported crime in Alaska	
Script Export	10	
Split data repeatedly on newline into	11 2004	3370.9
rows	12 2005	3615
Split data repeatedly on ','	13 2006	3582
	14 2007	3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts Kandel et al. [CHI 2011]

⊳	Split	Cut	Extract	Edit	Fill	Tra	nslate	Drop	Merge	Wrap	Delete	Promote	e Fold	Pivot	Transpo	se				
Tra	insform	Suggesti	ons																	
							#	spl	it	#	sp	lit1	#	s	olit2	#	split3		#	sp
						1	Report	ed crime	in Alabama											
						2														
						3	Year			Popu	lation		Prope	erty crim	e rate	Burgl	ary rate	1	Larceny-t	heft
						4	2004			4525	375		4029.	. 3		987		:	2732.4	
						5	2005			4548	327		3900			955.8		1	2656	
						6	2006			4599	030		3937			968.9		:	2645.1	
						7	2007			4627	851		3974.	. 9		980.2		1	2687	
						8	2008			4661	900		4081.	. 9		1080.	7	:	2712.6	
						9														
						10	Report	ed crime	in Alaska											
						11														
						12	Year			Popu	lation		Prope	erty crim	e rate	Burgl	ary rate	1	Larceny-t	heft
						13	2004			6577	55		3370.	. 9		573.6		:	2456.7	
_						14	2005			6632	53		3615			622.8		1	2601	
Tra	Insform 3	Script		E	xport	15	2006			6700	53		3582			615.2		1	2588.5	
⊳ s	plit data	repeated	lly on newlin	ne into		16	2007			6834	78		3373.	. 9		538.9		:	2480	
r	ows					17	2008			6862	93		2928.	.3		470.9		1	2219.9	
⊳ s	plit data	repeated	lly on 'tab'			18														
						19	Report	ed crime	in Arizona											
						20														
						ROW	S: 458													

▶ Split Cut Extract Edit Fill	Tra	nslate Drop	Merge V	Vrap	Delete	Promote	Fold	Pivot	Transpose				
Transform Suggestions													
Delete row 2		# sp	lit	#	spl	it1	#	sp	lit2	#	split3	#	sp
	1	Reported crime	in Alabama										
Delete empty rows	2												
	3	Year		Popula	ation		Prope	rty crime	e rate	Burglary r	ate	Larceny-	cheft
Delete rows where split is null	4	2004		45253	75		4029.	3		987		2732.4	
Delete rows where split1 is null	5	2005		45483	27		3900			955.8		2656	
	6	2006		45990	30		3937			968.9		2645.1	
Delete rows where split2 is null	7	2007		46278	51		3974.	9		980.2		2687	
	8	2008		46619	0 0		4081.	9		1080.7		2712.6	
Delete rows where split3 is null	9												
Fold using 2 as a key	10	Reported crime	in Alaska										
	11												
	12	Year		Popula	ation		Prope	rty crime	e rate	Burglary r	ate	Larceny-	theft
	13	2004		65775	5		3370.	9		573.6		2456.7	
	14	2005		66325	3		3615			622.8		2601	
Transform Script Export	15	2006		67005	3		3582			615.2		2588.5	
Split data repeatedly on newline into	16	2007		68347	8		3373.	9		538.9		2480	
rows	17	2008		68629	3		2928.	3		470.9		2219.9	
Split data repeatedly on 'tab'	18												
	19	Reported crime	in Arizona										
	20												
	ROW	S: 458											

▶ Split Cut Extract Edit Fill	Tra	nslate Dro	p Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose				
Transform Suggestions													
Delete row 2		#	split	*	sp	lit1	#	sp	lit2	#	split3	#	sp
	1	Reported cr	ime in Alabam	a									
Delete empty rows	2												
Data and the second second second	3	Year		Po	pulation		Prope	erty crime	e rate	Burglary rat	ce	Larceny	-theft
Delete rows where split is hull	4	2004		45	25375		4029	.3		987		2732.4	
Delete rows where split1 is null	5	2005		45	48327		3900			955.8		2656	
	6	2006		45	99030		3937			968.9		2645.1	
Delete rows where split2 is null	7	2007		46	27851		3974	.9		980.2		2687	
	8	2008		46	61900		4081	.9		1080.7		2712.6	
Delete rows where split3 is null	9												
Fold using 2 as a key	10	Reported cr	ime in Alaska										
	11												
	12	Year		Po	pulation		Prope	erty crime	e rate	Burglary rat	te	Larceny	-theft
	13	2004		65	7755		3370	. 9		573.6		2456.7	
	14	2005		66	3253		3615			622.8		2601	
Transform Script Export	15	2006		67	0053		3582			615.2		2588.5	
Split data repeatedly on newline into	16	2007		68	3478		3373	.9		538.9		2480	
rows	17	2008		68	6293		2928	.3		470.9		2219.9	
Split data repeatedly on 'tab'	18												
	19	Reported cr	ime in Arizon	a									
	20												
	ROW	S: 458											

▶ Split Cut Extract Edit F	ll Tra	inslate	Drop Merge	Wrap	Delete Promo	te Fold	Pivot Transpo	se		
Transform Suggestions										
		*	split	#	split1	#	split2	# \$	aplit3 #	sp
	1	Reported	crime in Alabar	na						
	2	Year		Popula	tion	Prope	rty crime rate	Burglary rat	e Larceny-th	eft
	3	2004		452537	5	4029.	3	987	2732.4	
	4	2005		454832	27	3900		955.8	2656	
	5	2006		459903	0	3937		968.9	2645.1	
	6	2007		462785	1	3974.	9	980.2	2687	
	7	2008		466190	0	4081.	9	1080.7	2712.6	
	8	Reported	crime in Alaska	a						
	9	Year		Popula	tion	Prope	rty crime rate	Burglary rat	e Larceny-th	eft
	10	2004		657755	i	3370.	9	573.6	2456.7	
	11	2005		663253	1	3615		622.8	2601	
	12	2006		670053	}	3582		615.2	2588.5	
	13	2007		683478	}	3373.	9	538.9	2480	
	14	2008		686293	}	2928.	3	470.9	2219.9	
Transform Script Ex	port 15	Reported	crime in Arizon	na						
Split data repeatedly on newline into	16	Year		Popula	tion	Prope	rty crime rate	Burglary rat	e Larceny-th	eft
rows	17	2004		573987	9	5073.	3	991	3118.7	
Split data repeatedly on 'tab'	18	2005		595300	7	4827		946.2	2958	
spin call repeatenty on the	19	2006		616631	8	4741.	6	953	2874.1	
Delete empty rows	20	2007		633875	5	4502.	6	935.4	2780.5	
	ROW	VS: 357								

▶ Split Cut Extract Edit Fill	Tra	nslate Drop Merge	Wra	ap Delete	Promote	Fold	Pivot	Transpose					
Transform Suggestions													
Delete row 2		# Year		# Popu	lation	#	Property_	crime_rate	#	Burglary_rate	#	Larceny-	
	1	Reported crime in Alabama											
Delete rows where split = 'Year'	2	Year	P	opulation		Prop	perty crime	e rate	Burglar	y rate	Larc	eny-theft	
	3	2004	4	525375		402	9.3		987		2732	. 4	
Delete rows where split1 = 'Population'	4	2005	4	1548327		390	0		955.8			2656	
Delete rows where split2 = 'Property	5	2006	4	1599030		393	7		968.9		2645	.1	
crime rate'	6	2007	4	627851		397	4.9		980.2		2687		
	7	2008	4	661900		408	1.9		1080.7		2712	.6	
Delete rows where split3 = 'Burglary	8	Reported crime in Alaska											
rate	9	Year	P	Population			perty crime	e rate	Burglar	y rate	Larc	eny-theft	
Delete rows where split4 = 'Larceny-	10	2004	6	57755		337	0.9		573.6		2456	.7	
theft rate'	11	2005	6	63253		361	5		622.8		2601		
Promoto row 2 to header	12	2006	6	570053		358	2		615.2		2588	.5	
	13	2007	6	83478		337	3.9		538.9		2480		
	14	2008	6	86293		292	8.3		470.9		2219	.9	
Transform Script Export	15	Reported crime in Arizona											
Split data repeatedly on newline into	16	Year	P	opulation		Pro	perty crime	rate	Burglar	y rate	Larc	eny-theft	
rows	17	2004	5	739879		507	3.3		991		3118	.7	
Split data repeatedly on 'tab'	18	2005	5	953007		482	7		946.2		2958		
	19	2006	6	5166318		474	1.6		953		2874	.1	
Delete empty rows	20	2007	6	338755		450	2.6		935.4		2780	.5	
	ROW	/S: 357											

▶ Split Cut Extract Edit Fi	ר ו	ranslate	Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose					
Transform Suggestions															
		#	Year		#	Popu	lation	#	Property	_crime_rate	#	Burglary_rate		#	Larceny-
	1	Report	ted crime i	n Alabama											
	2	2004			4525	375		402	9.3		987		2	732.4	
	3	2005			4548	327		390	0		955.8		2	656	
	4	2006			4599	030		393	7		968.9		2	645.1	
	5	2007			4627	851		397	4.9		980.2		2	687	
	6	2008			4661	900		408	1.9		1080.7		2	712.6	5
	7	Report	ted crime i	n Alaska											
	8	Year			Popu	lation		Pro	perty cri	me rate	Burgla	ry rate	I	arcen	y-theft
	9	2004			6577	55		337	0.9		573.6		2	456.7	
	10	2005			6632	53		361	5		622.8		2	601	
	11	2006			6700	53		358	2		615.2		2	588.5	5
	12	2007			6834	78		337	3.9		538.9		2	480	
•	13	2008			6862	93		292	8.3		470.9		2	219.9	)
	14	Report	ted crime i	n Arizona											
Transform Script Ex	ort 15	Year			Popu	lation		Pro	perty cri	me rate	Burgla	ry rate	I	arcen	y-theft
Split data repeatedly on newline into	16	2004			5739	879		507	3.3		991		З	118.7	
rows	17	2005			5953	007		482	7		946.2		2	958	
Split data repeatedly on 'tab'	18	2006			6166	318		474	1.6		953		2	874.1	
	19	2007			6338	755		450	2.6		935.4		2	780.5	
Delete empty rows	20	2008			6500	180		408	7.3		894.2		2	605.3	3
Promote row 2 to header	R	OWS: 356													

▶ Split Cut Extract Edit Fill	Tra	nslate Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose				
Transform Suggestions													
Delete row 8		# Y	ear	#	Popula	ation	#	Property	_crime_rate	#	Burglary_rate	#	Larceny-
	1	Reported crim	e in Alabama										
Delete rows where Year = 'Year'	2	2004		45253	75		4029	.3		987		2732	.4
	3	2005		45483	27		3900	)		955.8		2656	j.
Delete rows where Population =	4	2006		45990	30		3937	1		968.9		2645	.1
Population	5	2007		46278	51		3974	1.9		980.2		2687	
Delete rows where Property_crime_rate	6	2008		46619	00		4081	.9		1080.7		2712	.6
= 'Property crime	7	Reported crim	e in Alaska										
Delete rows where Burglary rate -	8	Year		Popula	ation		Prop	erty crim	le rate	Burgla	ry rate	Larc	eny-theft
'Burglary rate'	9	2004		65775	5		3370	).9		573.6		2456	.7
	10	2005		66325	3		3615	j		622.8		2601	
Delete rows where Larceny-theft_rate	11	2006		67005	3		3582	2		615.2		2588	.5
= 'Larceny-theft ra	12	2007		68347	8		3373	3.9		538.9		2480	J
Fill row 8 with values from the left	13	2008		68629	3		2928	3.3		470.9		2219	.9
	14	Reported crim	e in Arizona										
Transform Script Export	15	Year		Popul	ation		Prop	erty crim	e rate	Burgla	ry rate	Larc	eny-theft
Split data repeatedly on newline into	16	2004		57398	79		5073	3.3		991		3118	.7
rows	17	2005		59530	07		4827	1		946.2		2958	ļ
Split data repeatedly on 'tab'	18	2006		61663	18		4741	.6		953		2874	.1
	19	2007		63387	55		4502	2.6		935.4		2780	.5
Delete empty rows	20	2008		65001	80		4087	1.3		894.2		2605	.3
Promote row 2 to header	ROW	/S: 356											

▶ Split Cut Extract Edit Fill	Tra	nslate Drop	Merge W	rap D	Promote Promote	Fold	Pivot	Transpose				
Transform Suggestions												
Delete row 8		# Year		#	Population	#	Property_	crime_rate	#	Burglary_rate	*	Larceny-
	1	Reported crime in	Alabama									
Delete rows where Year = 'Year'	2	2004		4525375		4029	9.3		987		2732	. 4
	3	2005		4548327		3900	)		955.8		2656	
Delete rows where Population =	4	2006		4599030		3937	7		968.9		2645	.1
ropulation	5	2007		4627851		3974	1.9		980.2		2687	
Delete rows where Property_crime_rate	6	2008		4661900		4081	1.9		1080.7		2712	. 6
= 'Property crime	7	Reported crime in	Alaska									
Delete rows where Burglary rate -	8	Year		Populati	ion	Prop	perty crime	e rate	Burglary	y rate	Larc	eny-theft
'Burglary rate'	9	2004		657755		3370	).9		573.6		2456	.7
	10	2005		663253		3615	5		622.8		2601	
Delete rows where Larceny-theft_rate	11	2006		670053		3582	2		615.2		2588	.5
= Larceny-thett ra	12	2007		683478		3373	3.9		538.9		2480	
Fill row 8 with values from the left	13	2008		686293		2928	3.3		470.9		2219	. 9
	14	Reported crime in	Arizona									
Transform Script Export	15	Year		Populati	ion	Prop	perty crime	e rate	Burglar	y rate	Larc	eny-theft
Split data repeatedly on newline into	16	2004		5739879		5073	3.3		991		3118	.7
rows	17	2005		5953007		4827	7		946.2		2958	
Split data repeatedly on 'tab'	18	2006		6166318		4741	1.6		953		2874	.1
	19	2007		6338755		4502	2.6		935.4		2780	.5
Delete empty rows	20	2008		6500180		408	7.3		894.2		2605	.3
Promote row 2 to header	ROW	S: 356										

▶ Split Cut Extract Edit Fi	ll Tra	anslate Drop Merge	Wrap Delete Promote	Fold Pivot Transpose		
Transform Suggestions						
		# Year	# Population	# Property_crime_rate	# Burglary_rate	# Larceny
	1	Reported crime in Alabama				
	2	2004	4525375	4029.3	987	2732.4
	3	2005	4548327	3900	955.8	2656
	4	2006	4599030	3937	968.9	2645.1
	5	2007	4627851	3974.9	980.2	2687
	6	2008	4661900	4081.9	1080.7	2712.6
	7	Reported crime in Alaska				
	8	2004	657755	3370.9	573.6	2456.7
	9	2005	663253	3615	622.8	2601
	10	2006	670053	3582	615.2	2588.5
	11	2007	683478	3373.9	538.9	2480
	12	2008	686293	2928.3	470.9	2219.9
	13	Reported crime in Arizona				
	14	2004	5739879	5073.3	991	3118.7
Transform Script Ex	port 15	2005	5953007	4827	946.2	2958
Split data repeatedly on newline into	16	2006	6166318	4741.6	953	2874.1
rows	17	2007	6338755	4502.6	935.4	2780.5
Solit data repeatedly on 'tab'	18	2008	6500180	4087.3	894.2	2605.3
opin data repeatedry on tab	19	Reported crime in Arkansa	5			
Delete empty rows	20	2004	2750000	4033.1	1096.4	2699.7
b. Dermote and Data bander	ROV	VS: 306				
Promote row Z to header						

▶ Split Cut Extract Edit Fill	Tra	inslate Drop Merge	Wrap Delete Promote	Fold Pivot Transpose		
Transform Suggestions						
Extract from Year between positions 18, 🛛 🍕		# Year	Abo extract	# Population	# Property_crime_rate	# Burgla
25	1	Reported crime in Alabama	Alabama			
Extract from Very on Michamol	2	2004		4525375	4029.3	987
Extract from fear on Alabama	3	2005		4548327	3900	955.8
Extract from Year after 'in '	4	2006		4599030	3937	968.9
	5	2007		4627851	3974.9	980.2
Extract from Year after ' in '	6	2008		4661900	4081.9	1080.7
	7	Reported crime in Alaska				
Extract from Year after 'crime in '	8	2004		657755	3370.9	573.6
Extract from Year after ' any word in '	9	2005		663253	3615	622.8
Extract non-real and any word in	10	2006		670053	3582	615.2
Cut from Year between positions 18, 25	11	2007		683478	3373.9	538.9
	12	2008		686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004		5739879	5073.3	991
Transform Script Expor	15	2005		5953007	4827	946.2
Split data repeatedly on newline into	16	2006		6166318	4741.6	953
rows	17	2007		6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008		6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansa			
Delete empty rows	20	2004		2750000	4033.1	1096.4
Promote row 2 to header	ROW	VS: 306				

Split	Cut	Extract	Edit	Fill	Translate	Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose
-------	-----	---------	------	------	-----------	------	-------	------	--------	---------	------	-------	-----------

Transform Suggestions						
		# Year	Abo extract	# Population	# Property_crime_rate	# Burgla
	1	Reported crime in Alabama	Alabama			
	2	2004		4525375	4029.3	987
	3	2005		4548327	3900	955.8
	4	2006		4599030	3937	968.9
	5	2007		4627851	3974.9	980.2
	6	2008		4661900	4081.9	1080.7
	7	Reported crime in Alaska	Alaska			
	8	2004		657755	3370.9	573.6
	9	2005		663253	3615	622.8
	10	2006		670053	3582	615.2
	11	2007		683478	3373.9	538.9
	12	2008		686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004		5739879	5073.3	991
Transform Script Ex	port 15	2005		5953007	4827	946.2
Split data repeatedly on newline into	16	2006		6166318	4741.6	953
rows	17	2007		6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008		6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansas			
Delete empty rows	20	2004		2750000	4033.1	1096.4
Promote row 2 to header	ROW	/S: 306				

Split	Cut	Extract	Edit	Fill	Translate	Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose
-------	-----	---------	------	------	-----------	------	-------	------	--------	---------	------	-------	-----------

Transform Suggestions						
		# Year	Abo extract	# Population	# Property_crime_rate	# Burgla
	1	Reported crime in Alabama	Alabama			
	2	2004		4525375	4029.3	987
	3	2005		4548327	3900	955.8
	4	2006		4599030	3937	968.9
	5	2007		4627851	3974.9	980.2
	6	2008		4661900	4081.9	1080.7
	7	Reported crime in Alaska	Alaska			
	8	2004		657755	3370.9	573.6
	9	2005		663253	3615	622.8
	10	2006		670053	3582	615.2
	11	2007		683478	3373.9	538.9
	12	2008		686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004		5739879	5073.3	991
Transform Script Ex	port 15	2005		5953007	4827	946.2
Split data repeatedly on newline into	16	2006		6166318	4741.6	953
rows	17	2007		6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008		6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansas			
Delete empty rows	20	2004		2750000	4033.1	1096.4
Promote row 2 to header	ROW	/S: 306				

Split	Cut	Extract	Edit	Fill	Translate	Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose
-------	-----	---------	------	------	-----------	------	-------	------	--------	---------	------	-------	-----------

Transform Suggestions						
		# Year	Abo extract	# Population	# Property_crime_rate	# Burgla
	1	Reported crime in Alabama	Alabama			
	2	2004		4525375	4029.3	987
	3	2005		4548327	3900	955.8
	4	2006		4599030	3937	968.9
	5	2007		4627851	3974.9	980.2
	6	2008		4661900	4081.9	1080.7
	7	Reported crime in Alaska	Alaska			
	8	2004		657755	3370.9	573.6
	9	2005		663253	3615	622.8
	10	2006		670053	3582	615.2
	11	2007		683478	3373.9	538.9
	12	2008		686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004		5739879	5073.3	991
Transform Script Ex	port 15	2005		5953007	4827	946.2
Split data repeatedly on newline into	16	2006		6166318	4741.6	953
rows	17	2007		6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008		6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansas			
Delete empty rows	20	2004		2750000	4033.1	1096.4
Promote row 2 to header	ROW	/S: 306				

▶ Split Cut Extract Edit Fill	Tra	anslate Drop Merge N	Vrap Delete Promote	Fold Pivot Transpose		
Transform Suggestions						
Fill extract with values from above		# Year	Abo extract	# Population	# Property_crime_rate	# Burgla
	1	Reported crime in Alabama	Alabama			
Fill extract with values from below	2	2004	Alabama	4525375	4029.3	987
	3	2005	Alabama	4548327	3900	955.8
Drop extract	4	2006	Alabama	4599030	3937	968.9
Fold extract using header as a key	5	2007	Alabama	4627851	3974.9	980.2
	6	2008	Alabama	4661900	4081.9	1080.7
Fold <b>extract</b> using <b>1</b> as a key	7	Reported crime in Alaska	Alaska			
	8	2004	Alaska	657755	3370.9	573.6
Fold extract using 1, 2 as keys	9	2005	Alaska	663253	3615	622.8
Fold extract using 1, 2, 3 as keys	10	2006	Alaska	670053	3582	615.2
Tord extract using 1, 2, 5 as keys	11	2007	Alaska	683478	3373.9	538.9
	12	2008	Alaska	686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004	Arizona	5739879	5073.3	991
Transform Script Export	15	2005	Arizona	5953007	4827	946.2
Split data repeatedly on newline into	16	2006	Arizona	6166318	4741.6	953
rows	17	2007	Arizona	6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008	Arizona	6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansas			
Delete empty rows	20	2004	Arkansas	2750000	4033.1	1096.4
Promote row 2 to header	ROW	WS: 306				

▶ Split Cut Extract Edit Fi	ll Tra	anslate Drop Merge	Wrap Delete Promote	Fold Pivot Transpose		
Transform Suggestions		H Year	Abs owned	# Demulation	# Duranti arima rata	# Dunal
	1	# Year	Alabama	# Population	# Property_crime_rate	* Burgia
	1	Reported Crime in Alabama	Alabama	4505075	4000.0	007
	2	2004	Alabama	4525375	4029.3	987
	3	2005	Alabama	4548327	3900	955.8
	4	2006	Alabama	4599030	3937	968.9
	5	2007	Alabama	4627851	3974.9	980.2
	6	2008	Alabama	4661900	4081.9	1080.7
	7	Reported crime in Alaska	Alaska			
	8	2004	Alaska	657755	3370.9	573.6
	9	2005	Alaska	663253	3615	622.8
	10	2006	Alaska	670053	3582	615.2
	11	2007	Alaska	683478	3373.9	538.9
	12	2008	Alaska	686293	2928.3	470.9
	13	Reported crime in Arizona	Arizona			
	14	2004	Arizona	5739879	5073.3	991
Transform Script Exp	port 15	2005	Arizona	5953007	4827	946.2
Split data repeatedly on newline into	16	2006	Arizona	6166318	4741.6	953
rows	17	2007	Arizona	6338755	4502.6	935.4
Split data repeatedly on 'tab'	18	2008	Arizona	6500180	4087.3	894.2
	19	Reported crime in Arkansas	Arkansas			
Delete empty rows	20	2004	Arkansas	2750000	4033.1	1096.4
Promote row 2 to header	ROW	VS: 306				
FIDITULE TOW Z TO HEADER						

Extract from Year after 'in '

Fill extract with values from above

► Split Cut Extract Edit Fill	Tra	unslate Drop Merge W	/rap Delete Promote	Fold Pivot Transpose		
Transform Suggestions						
Extract from Year between positions 0.		# Year	Abo extract1	Abo extract	# Population	# Property_
17	1	Reported crime if Alabama	Reported crime in	Alabama		
	2	2004		Alabama	4525375	4029.3
Extract from Year on Reported crime in	3	2005		Alabama	4548327	3900
Extract from Year on 'Reported crime	4	2006		Alabama	4599030	3937
any word '	5	2007		Alabama	4627851	3974.9
	6	2008		Alabama	4661900	4081.9
Extract from Year on 'Reported crime	7	Reported crime in Alaska	Reported crime in	Alaska		
any lowercase word	8	2004		Alaska	657755	3370.9
Extract from Year on 'Reported any	9	2005		Alaska	663253	3615
word in'	10	2006		Alaska	670053	3582
Extract from Year on 'Reported any	11	2007		Alaska	683478	3373.9
word any word '	12	2008		Alaska	686293	2928.3
	13	Reported crime in Arizona	Reported crime in	Arizona		
Cut from Year between positions 0, 17	14	2004		Arizona	5739879	5073.3
Transform Script Export	15	2005		Arizona	5953007	4827
Split data repeatedly on newline into	16	2006		Arizona	6166318	4741.6
rows	17	2007		Arizona	6338755	4502.6
Split data repeatedly on 'tab'	18	2008		Arizona	6500180	4087.3
	19	Reported crime in Arkansas	Reported crime in	Arkansas		
Delete empty rows	20	2004		Arkansas	2750000	4033.1
Promote row 2 to header	ROW	VS: 306				

Extract from Year after 'in '

Fill extract with values from above

▶ Split Cut Extract Edit Fill	Tra	nslate Drop	Merge V	Vrap	Delete	Promote	Fold	Pivot	Transpose				
Transform Suggestions													
Delete rows where Year starts with		# Ye	ar	Abo	extr	act	#	Pop	ulation	#	Property_crime_rate	#	Burgla
'Reported crime in'	1	Reported crime	in Alabama	Alaba	ama								
	2	2004		Alaba	ıma		4525	375		402	9.3	987	
Delete rows where Year contains 'Reported crime in'	3	2005		Alaba	ama		4548	327		390	0	955.8	
Reported crime in	4	2006		Alaba	ıma		4599	030		393	7	968.9	
Extract from Year between positions 0,	5	2007		Alaba	ama		4627	851		397	4.9	980.2	
17	6	2008		Alaba	ama		4661	900		408	1.9	1080.7	
Extract from Year on 'Reported crime in'	7	Reported crime	in Alaska	Alask	(a								
Extract from real on Reported crime in	8	2004		Alask	a		6577	55		337	0.9	573.6	
Extract from Year on 'Reported crime	9	2005		Alask	a		6632	53		361	5	622.8	
any word '	10	2006		Alask	a		6700	53		358	2	615.2	
Extract from Vession 'Reported crime	11	2007		Alask	a		6834	78		337	3.9	538.9	
any lowercase word '	12	2008		Alask	a		6862	93		292	8.3	470.9	
	13	Reported crime	in Arizona	Arizo	ona								
Extract from Year on 'Reported any	14	2004		Arizo	ona		5739	879		507	3.3	991	
Transform Script Export	15	2005		Arizo	ona		5953	007		482	7	946.2	
Split data repeatedly on newline into	16	2006		Arizo	ona		6166	318		474	1.6	953	
rows	17	2007		Arizo	ona		6338	755		450	2.6	935.4	
Split data repeatedly on 'tab'	18	2008		Arizo	ona		6500	180		408	7.3	894.2	
	19	Reported crime	in Arkansas	Arkan	isas								
Delete empty rows	20	2004		Arkan	isas		2750	000		403	3.1	1096.4	
Promote row 2 to header	ROW	S: 306											

Extract from Year after 'in '

Fill extract with values from above

▶ Split Cut Extract Edit	Fill	Tra	nslate Dro	p Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose					
Transform Suggestions															
			#	Year	Abo	ext	ract	#	Pop	ulation	#	Property_crime_rate	#	ł	Burgla
		1	2004		Alak	oama		4525	375		402	29.3	98	37	
		2	2005		Alak	oama		4548	327		390	0.0	95	5.8	
		3	2006		Alak	oama		4599	030		393	37	96	8.9	
		4	2007		Alak	oama		4627	851		391	74.9	98	0.2	
		5	2008		Alak	oama		4661	900		408	81.9	10	80.7	
		6	2004		Alas	ska		6577	55		337	70.9	57	3.6	
		7	2005		Alas	ska		6632	53		361	15	62	2.8	
		8	2006		Alas	ska		6700	53		358	82	61	5.2	
		9	2007		Alas	ska		6834	78		331	73.9	53	8.9	
		10	2008		Alas	ska		6862	93		292	28.3	47	0.9	
		11	2004		Ariz	zona		5739	879		507	73.3	99	1	
		12	2005		Ariz	zona		5953	007		482	27	94	6.2	
		13	2006		Ariz	zona		6166	318		474	41.6	95	3	
		14	2007		Ariz	zona		6338	755		450	02.6	93	5.4	
Transform Script	Export	15	2008		Ariz	zona		6500	180		408	87.3	89	4.2	
Split data repeatedly on newline into		16	2004		Arka	ansas		2750	000		403	33.1	10	96.4	
rows		17	2005		Arka	ansas		2775	708		406	68	10	85.1	
Split data repeatedly on 'tab'		18	2006		Arka	ansas		2810	872		402	21.6	11	54.4	
		19	2007		Arka	ansas		2834	797		394	45.5	11	24.4	
Delete empty rows		20	2008		Arka	ansas		2855	390		384	43.7	11	82.7	
Promote row 2 to header		ROW	/S: 255												

Extract from Year after 'in '

Fill extract with values from above

Delete rows where Year starts with 'Reported crime in'

Split Cut Extract Edit	Fill Translate Drop Merge Wrap Delete Promote Fold Pivot Transpose
Transform Suggestions	💿 Data
	○ Script
	Comma-Separated Values (CSV) Back to Wrangling
	Year, extract, Population, Property crime rate, Burglary rate, Larceny-
	theft_rate,Motor_vehicle_theft_rate
	2004, Alabama, 4525375, 4029.3, 987, 2732.4, 309.9
	2005, Alabama, 4548327, 3900, 955.8, 2656, 289
	2006, Alabama, 4599030, 3937, 968.9, 2645.1, 322.9
	2007, Alabama, 4627851, 3974.9, 980.2, 2687, 307.7
	2004, Alabama, 4661900, 4081.9, 1080.7, 2712.6, 288.6
	2004, Alaska, 65/153, 33/0.9, 5/3.6, 2456.1, 340.6
	2005 Alaska,005253,5013,022.0,2001,531
	2007 Alaska, 683478, 3373, 9, 538, 9, 2480, 355, 1
	2008. Alaska, 686293. 2928. 3. 470. 9. 2219. 9. 237. 5
	2004, Arizona, 5739879, 5073.3, 991, 3118.7, 963.5
	2005, Arizona, 5953007, 4827, 946.2, 2958, 922
	2006, Arizona, 6166318, 4741.6, 953, 2874.1, 914.4
Transform Script	Exert 2007, Arizona, 6338755, 4502.6, 935.4, 2780.5, 786.7
	2008,Arizona,6500180,4087.3,894.2,2605.3,587.8
Split data repeatedly on newline into	2004, Arkansas, 2750000, 4033.1, 1096.4, 2699.7, 237
rows	2005, Arkansas, 2775708, 4068, 1085.1, 2720, 262
b. Calls data associated as a listed.	2006, Arkansas, 2810872, 4021.6, 1154.4, 2596.7, 270.4
Split data repeatedly on tab	2009, Arkansas, 2854/9/, 3945.5,1124.4,25/4.6,246.5
Delete empty rows	2000 ( alifornia 359/2038 3/323 0 666 1 2033 1 704 8
· Delete empty rows	2005. California, 3612030, 5425. 7, 602. 9, 1915. 712
Promote row 2 to header	2006. California, 36457549, 3175, 2, 676, 9, 1831, 5, 666, 8
	2007, California, 36553215, 3032.6, 648.4, 1784.1, 600.2
Delete rows where Year = 'Year'	2008, California, 36756666, 2940.3, 646.8, 1769.8, 523.8
	2004, Colorado, 4601821, 3918.5, 717.3, 2679.5, 521.6
Extract from Year after 'in '	2005,Colorado,4663295,4041,745.1,2736,560
	2006, Colorado, 4753377, 3441.8, 682, 2325.1, 434.8
Fill extract with values from above	2007, Colorado, 4861515, 2991.3, 588.5, 2061.1, 341.7
b Balan and a Manadari and	2008, Colorado, 4939456, 2856.7, 571.4, 2013.7, 271.6
Delete rows where Year starts with "Reported grime in"	2004, Connecticut, 3498966, 2684.9, 456.1, 1908.3, 320.5
Reported crime in	2005, Connecticut, 3500/01, 25/9, 435.5, 1840, 303
	(AUDITED FOR A DUADUS, ATA A A A A A A A A A A A A A A A A A

### **Data**Wrangler

Suggestions	rows: 408 prev next	
	H Year	♦ ∰ Property_crime_rate ♦
Delete rows 8,10	1 Reported crime in Alabama	
Delete empty rows	3 2004	4029.3
Delete rows where Property_crime_rate	5 2006	3937
is null Delete rows where Year is null	6 2007 7 2008	3974.9 4081.9
	8 9Reported crime in Alaska	
Script Export	10	
Split data repeatedly on newline into rows	11 2004	3370.9
	12 2005	3615
Split data repeatedly on ','	13 2006	3582
	14 2007	3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts Kandel et al. [CHI 2011] The first sign that a visualization is good is that it shows you a problem in your data. Every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.

Martin Wattenberg [ACM Queue '09]





### 000

Graph Viewer



1

2.2

### 0 0 **Graph Viewer** Graph Viewer 2.7Roll-up by: 6224 and the second second 1.1 1.1.1 والأقبار العرجاجير والمحالة لحجر متقصبها وترجعو فارورا ರ್ಷ ತಿಳಿಸಲು 90 ve 8.00 10 化酸盐海绵酸盐酸盐和盐油油和盐油油盐和盐油油的 计推进分析 的复数 101 1.18 + All $(X_{ij}^{-1})_{ij}^{-1} \in \mathbb{N}^{n} \to \mathbb{N}^$ Visualization: 1.16 and the $\sim 10$ . . . -18 a a la segura de la 137 + 法法律的 网络名 Matrix 122 A 2019 19 法的现在分词 法无法 1 des 含く湯 Leves et Sort by: -imi+ None 1.2.2 Edge centrality filters: 2126 ( S. 1993) 16 B. 1 an an ta and the second states of $N_{\rm e} = 0$ 10 A 44 2 A R. استبدع فالمراجع والمعتك ساس nesegi وأحجوه والمراجع والمتحود المحاج والاسترا n in faith ann an Anna Anna Anna Thair ann an Anna Anna Anna Anna Anna 1.19.1.29.00 11 A A A A 1.1.1 a cipitatini B 5. S. S. ್ರಮ 1.11 a an ann an 1943. An 1959 Ann an 1969 Ann an 22 A 44 1.2.7.4 - A. 192 4.45 an ann a' i 1.116 2日表 5 1.5 100 요구 같이

100.00 the second second and Bar Jamman 한다다 아이들이 한 승규는 동안을 받는 것 같아.

# Visualize Friends by School?

Berkeley Cornell Harvard Harvard University Stanford Stanford University UC Berkeley UC Davis University of California at Berkeley University of California, Berkeley University of California, Davis



# Data Quality Hurdles

Missing Data Erroneous Values Type Conversion Entity Resolution Data Integration

no measurements, redacted, ...? misspelling, outliers, ...? e.g., zip code to lat-lon diff. values for the same thing? effort/errors when combining data

Anticipate problems with your data!

# Data Wrangling Tools

*Libraries* JavaScript: Arquero Python: Pandas, Polars R: dplyr

*Databases* DuckDB + SQL queries

*Graphical Tools* We'll look at some of these next!
### Trifacta Wrangler (now part of Alteryx)

Q	֎ Campaign Finance	2016 > 🔟 cn16 ~			$1 \rightarrow \square \rightarrow 0$ Generate Resu	ilts 🚨 📀	
Grid	Columns Full Da	ntaset - 461.78kB ~ 17 Colum	nns 4,864 Rows 3 Data Ty	pes	Columns: V All Transformed - 3 Columns Rows: V All Transformed - 4,859 Row	s Q. Filter in grid	50 #
		Source to be dropped					1
	ABC CAND_ID ~	ABC CAND_NAME ~	ABC CAND_NAME1	ABC CAND_NAME2	ABC CAND_PARTY_AFFILIATION ~	CAND_ELECTION_YEAR	- ABC
	4864 Cotoporios	4760 Categories		2.677 Cotonorion	Té Colonaiae	1986, 2082	5) 57 Cotr
	4,004 Categories	4,760 Categories	5,410 Categories	IOUN D	DED	2014	S/ Cate He
	H04L02097		PORV	мартна	REP	2014	AL 50
	H0AL 02095	JOHN BOBERT F. JR	JOHN	ROBERT · E · JR	TND	2016	AL
	HØAL05049	CRAMER, ROBERT E "BUD" JR	CRAMER	ROBERT E 'BUD' JR	DEM	2008	AL
	H0AL05163	BROOKS, MO	BROOKS	МО	REP	2016	AL
	HØAL06088	COOKE, STANLEY KYLE	COOKE	STANLEY KYLE	REP	2010	AL
	HØAL07086	SEWELL, TERRIA.	SEWELL	TERRI A.	DEM	2016	AL
	HØAL07094	HILLIARD, EARL FREDERICK JR	HILLIARD	EARL · FREDERICK · JR	DEM	2010	AL
	H0AL07177	CHAMBERLAIN, DON	CHAMBERLAIN	DON	REP	2012	AL
	H0AR01083	CRAWFORD, ERIC ALAN RICK	CRAWFORD	ERIC · ALAN · RICK	REP	2016	AR
	H0AR01091	GREGORY, JAMES CHRISTOPHER	GREGORY	JAMES CHRISTOPHER	DEM	2010	AR
	H0AR01109	CAUSEY, CHAD	CAUSEY	CHAD	DEM	2010	AR
	H0AR01125	SMITH, PRINCELLA D	SMITH	PRINCELLA	REP	2010	AR
	HØAR02107	GRIFFIN, JOHN TIMOTHY	GRIFFIN	JOHN TIMOTHY	REP	2014	AR
	HØAR02131	ELLIOTT, JOYCE ANN	ELLIOTT	JOYCE · ANN	DEM	2010	AR
	HØAR03022	SKOCH, BERNARD KURT 'BERNIE'	SKOCH	BERNARD · KURT · ' BERNIE '	REP	2010	AR
	HØAR03030	WHITAKER, DAVID JEFFREY	WHITAKER	DAVID JEFFREY	DEM	2010	AR
	H0AR03055	WOMACK, STEVE	WOMACK	STEVE	REP	2016	AR
	H0AS00018	FALEOMAVAEGA, ENI	FALEOMAVAEGA	ENI	DEM	2014	AS
	H0AZ01184	FLAKE, JEFF MR.	FLAKE	JEFF MR.	REP	2012	AZ
	H0AZ01259	GOSAR, PAUL ANTHONY	GOSAR	PAUL ANTHONY	REP	2016	AZ
	101701000	NEUTA OTEVE	MENTA	OTOIC	050	0010	

#### ♀ SUGGESTIONS

ABC CAND_NAME	ABC CAND_NAME1	ABC CAND_NAME2
COX, JOHN R.	COX	JOHN R.
ROBY, MARTHA	ROBY	MARTHA
JOHN, ROBERT E JR	JOHN	ROBERT · E · JR
Affects 1 column, 4859 rows	Creates 2 columns	

ABC	CAND_NAME	ABC	CAND_NAME1
cox, J	DHN·R.	1.1	
ROBY, I	MARTHA	12	
JOHN, ·	ROBERT · E · JR	12	
Affects 1	column, 4859 rows	Creates	s 1 column

#### Cancel Modify Add to Recipe

ABC	CAND_NAME
COX,	· JOHN · R.
ROBY	, MARTHA
JOHN	, ROBERT · E · JR

### AWS Glue DataBrew

≡	nyccitibikes No job runs, no job runs scheduled P Run job 2010 rows)							
DATASETS				APPING ENCODE				
PROJECTS	Ø Viewing 21 columns ▼ 500 rows □ View highlig Source	hted	GRID	🛄 SCHEMA 🛍 PROFILE	Merge columns ×			
_	# start station latitude	# start station longitude	ABC lationg	# end station id				
RECIPES	Total     Sol     Unique     334     Total     Sol       6     1.2%     1	Unique 314 Total 500	Unique <b>334</b> Tetal <b>500</b> 40.72210379, -73.59724301 6 1.2% 40.74177605, -74.00143746	Unique 310	Merge columns Info Merge columns and create a new column			
JOBS	5 11% Min Medan Maan Mode Max 5 11% 40.46 40.74 40.74 40.72 40.85 84 96.8%	Min Median Mean Node Max -74.02 -73.96 -71.96 -74 -73.9	40.75401143, -74.00233877 5 1% All other values 484 96.8%	Min Median Mean Hode 79 3.11 K 2.12 K 325; 3.7	Source column Select two or more columns in the order to merge			
	40.819241	-73.941057	40.819241, -73.941057	3966	ii start station latitude 🛛 🗙			
UMMUNITY	40.68691865	-73.976682	40.68691865, -73.976682	3668	start station longitude ×			
	40.7689738	-73.95482273	40.7689738, -73.95482273	3164	Add a column 🛛 🔻			
	40.7919557	-73.968087	40.7919557, -73.968087	3906				
	40.71638032	-73.94821286	40.71638032, -73.94821286	128	Separator - Optional			
	40.764508	-73.9351	40.704508, -73.9351	3774	Concatenated values are separated by this			
	40.74177603	-74.00149746	40.74177603, -74.00149746	462	,			
	40.72110063	-73.9919254	40.72110063, -73.9919254	470	New column mmo			
	40.75038009	-73.98338988	40.75038009, -73.98338988	312	Name of the target column to merge into			
	40.7668	-73.9347774	40.7668, -73.9347774	372	latlong			
	40.72362738	-73.99949601	40.72362738, -73.99949601	400	Valid characters are alphanumeric, underscore, and space			
	40.773763	-73.96222088	40.773763, -73.96222088	405				
	40.825125	-/5.941616	40.825125, -73.941616	5629	• Preview shown			
	40.72714350	75.57418625	40.70870368, -75.9448625	5070				
	40.75514255	72.070.49	40.73214233, -75.3723681	407	Cancel Apply			
	40.65520077	74.01062797	M0.7 1002, -7.3.33340 A0.65520077 .74.01062797	2041				
	40.72124	72 05161	40 73124 _73 95151	3110				
	40.72210239	-72.00724001	40.72210379 -73.99724901	3113				
	40.78414472	77.98762492	40.78414472 -73.98362492	3160				
	40.7652654	-73.98192338	40.765265473.98192338	468				
	40.72706363	-73.00562137	40.72706363, -73.99662137	3812				
	40.7937704	-73.971888	40.7937704, -73.971888	500				
	Zoom 100% 🛪							

### **Tableau Prep**



### Deepnote

df.head(50)							[13	3]
DEPENDENTS bool	TECHSUPPORT ob	CONTRACT object	PAPERLESSBILL	MONTHLYCHAR	TOTALCHARGES f	CHURNVALUE flo	TENUREMONTHS 1	F
	No	Month-to-m 66%		19.35 - 110.15	44.7 - 7998.8	0.0 - 1.0	1 - 72	
False     82%       True     18%	Yes 26% No interne. 12%	Two year 18% One year 16%	true	L			Ba	Y
false	Yes	Month-to-month	true	83.4	83.4	0	1	Y
false	No	One year	true	100.05	6254.2	1	64	Y
false	No	Month-to-month	true	69.1	69.1	1	1	Y
false	No	Month-to-month	true	85.35	1375.15	1	16	Y
true	No	Month-to-month	true	79.25	1111.65	1	13	Y
false	No	Month-to-month	false	74.4	434.1	1	6	Y
false	No internet service	Two year	false	19.35	1099.6	1	59	Y

### **Observable Data Table Cells**

Ce	ell 1072 =	chinook • custon	ners			59 rows D R	un
Ŧ	Filter 📿 Columns 13	Sort 😽 Slice [	[0, 100] 🗢 SQL				
~	CustomerId number	FirstName string	LastName string	Company string 83% NULL /EMPT	Address string	City string	St str
	0 60	57 categories		11 categories		53 categories	26
0	1	Luís	Gonçalves	Embraer - Empresa Brasileira de Aeronáutica S.A.	Av. Brigadeiro Faria Lima, 2170	São José dos Campos	SF
1	2	Leonie	Köhler	NULL	Theodor-Heuss-Straße 34	Stuttgart	NI
2	3	François	Tremblay	NULL	1498 rue Bélanger	Montréal	Q
3	4	Bjørn	Hansen	NULL	Ullevålsveien 14	Oslo	NU
4	5	František	Wichterlová	JetBrains s.r.o.	Klanova 9/506	Prague	NI
5	6	Helena	Holý	NULL	Rilská 3174/6	Prague	NI
6	7	Astrid	Gruber	NULL	Rotenturmstraße 4, 1010 Innere Stadt	Vienne	NI
7	8	Daan	Peeters	NULL	Grétrystraat 63	Brussels	NI
8	9	Kara	Nielsen	NULL	Sønder Boulevard 51	Copenhagen	NI
9	10	Eduardo	Martins	Woodstock Discos	Rua Dr. Falcão Filho, 155	São Paulo	SF
10	ner nage ¥					page 1 of 6	,

### Quak widget usable in Jupyter Notebooks

	name	nationality	sex	height	weight	sport	gold	silver
	utf8	utf8	utf8	float64	int64	utf8	int64	int64
	unique	<b>undu</b> ndanti	male female	,		athar		1 .
	22 categories	207 categories	female	Ø1.2 2.3	Ø20 180	28 categories	0 5.5	0
0	A Lam Shin	KOR	female	1.68	56	fencing	0	
1	Aauri Lorena Bokesa	ESP	female	1.8	62	athletics	0	
2	Abbey Weitzeil	USA	female	1.78	68	aquatics	1	
3	Abbie Brown	GBR	female	1.76	71	rugby sevens	0	
4	Abby Erceg	NZL	female	1.75	68	football	0	
5	Abdoulkarim Fawzi	CMR	female	1.8	67	volleyball	0	
6	Abigel Joo	HUN	female	1.83	76	judo	0	

Reset 3,420 of 11,538 rows

### **Pandas Profiling**



### VisiData

• • •		d	data — vd itpas2.txt — 96×25
<u>F</u> ile <u>E</u> dit	<u>V</u> iew <u>C</u> olumn	<u>R</u> ow <u>D</u> ata	a <u>P</u> lot <u>S</u> ystem <u>H</u> elp Ctrl+H for help menu
text_re14	<b>↓</b> count#	percent%	histogram ~
1000	65203	12.59	***************************************
2500	40959	7.91	*****
25	32918	6.36	*****
0	31084	6.00	*****
5000	25790	4.98	*****
50	23538	4.54	*****
10	21228	4.10	*****
-25	18460	3.56	*****
2000	15857	3.06	*****
100	12716	2.45	****
-10	12362	2.39	*****
5	12130	2.34	*****
1	10621	2.05	*****
1500	10011	1.93	*****
35	8747	1.69	****
-5	8162	1.58	*****
-50	7676	1.48	****
500	6865	1.33	****
20	6189	1.19	****
-1	5365	1.04	****
30	5153	0.99	***
2	4765	0.92	*** • • • • • • • • • • • • • • • • • •
<pre>2&gt; itpas2_tex</pre>	t_re14_freq		((base) jheer@dreadnought data % ls -1 total 203288
			-rw-rr 1 jheer staff 2336013 Oct 12 15:24 cm.txt -rw-rr 1 jheer staff 814306 Oct 12 15:24 cm.txt -rw-rr0 1 jheer staff 90781486 Oct 12 15:25 itpas2.txt

(base) jheer@dreadnought data % vd cn.txt

. .

# **Visualizing Distributions**

# **Distribution Visualizations**

Strip Plot Jittered Plot Box Plot









# **Distribution Visualizations**

Histogram bin size = 2



Density Plot kde,  $\sigma = 0.5$ 



Violin Plot kde,  $\sigma = 0.5$ 



### Identical boxplots, different distributions

Boxplots are great. They show medians and ranges and enable comparison of different groups. However, boxplots can be misleading. Different datasets can have the same descriptive statistics (left), but quite different underlying distributions (middle). Therefore, it is crucial to visualize the distribution in addition to descriptive statistics. Violin plots with integrated boxplots are great for this.



## Now in 2D! Heatmaps, Contours



### **Kernel Density Estimation (KDE)**

Enables violin plots, heat maps, contour plots...



For a set of input data points...

• •

Represent each point with a "kernel" distribution



Sum the kernels to form a density estimate



Sized by bandwidth (standard deviation)



# **Dimensionality Reduction**

# **Dimensionality Reduction (DR)**

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

In contrast, multidimensional scaling (MDS) attempts to **preserve pairwise distances**.

# **Reduction Techniques**

LINEAR - PRESERVE GLOBAL STRUCTURE **Principal Components Analysis (PCA)** Linear transformation of basis vectors, ordered by amount of data variance they explain.

NON-LINEAR - PRESERVE LOCAL TOPOLOGY t-Dist. Stochastic Neighbor Embedding (t-SNE) Probabilistically model distance, optimize positions.

**Uniform Manifold Approx. & Projection (UMAP)** Identify local manifolds, then stitch them together.

# Mapping Emoji Images



t-SNE

UMAP

PCA

# Principal Components Analysis



1. Mean-center the data. 2. Find  $\perp$  basis vectors that maximize the data variance. 3. Plot the data using the top vectors.

# Principal Components Analysis



Linear transform: scale and rotate original space.

Lines (vectors) project to lines.

Preserves global distances.

## PCA of Genomes [Demiralp et al. '13]



### Vector Space Word Embeddings (word2vec, GloVe)





Male-Female

Verb tense

Country-Capital

### Mapping Machine-Learned Latent Spaces [Liu et al. 2019]



# **Non-Linear Techniques**

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods: **t-SNE** - probabilistic interpretation of distance **UMAP** - tries to balance local/global trade-off

### t-SNE [Maaten & Hinton 2008]

 Model probability P of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

### t-SNE [Maaten & Hinton 2008]

Define a similar probability Q in the low-dimensional (2D or 3D) embedding space, using a Student's t distribution (hence the "t-" in "t-SNE"!). The t-distribution is heavy-tailed, allowing distant points to be even further apart.



### t-SNE [Maaten & Hinton 2008]

- Model probability P of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
- Define a similar probability Q in the low-dimensional (2D or 3D) embedding space, using a Student's t distribution (hence the "t-" in "t-SNE"!). The t-distribution is heavy-tailed, allowing distant points to be even further apart.
- 3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the **P** and **Q** distributions: *KL(P* || *Q*)

## Visualizing t-SNE [Wattenberg et al. '16]



Results can be highly sensitive to the algorithm parameters! Are you seeing real structures, or algorithmic hallucinations?

### Hyperparameters matter!



## Cluster sizes mean nothing...



### **Cluster distances are illusive**



### Random noise may not look like it


### You can see shapes, sometimes



### Multi-Lingual Word Embedding [Google 2016]



### **UMAP** [McInnes et al. 2018]

Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

"Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance." - McInnes et al. 2018



Figure 1: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional colorcube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of n spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.

# **User Engagement with Interactive Articles**

Provide an overview of usage patterns of interactive features.

Identify variations in usage

Represent reader sessions as a feature vector with:

- time spent in each section
- count of variable changes





isInTur

playRif

clear

autotuneGuita

playReference

playNotes

nlavBrats

beatDiff

7.0 guiarState currentProgram

204.0



Bars show time spent in each



204.0

7.0 guitarState 77.0

targetString

currentFrequency





204.0

7.0 guitarState

playBeats

beatDiff

playNotes

77.0

targetString

currentFrequency

... and the count of times each variable changed

playReference

clear





# Mapping Emoji Images



t-SNE

UMAP

PCA



### **Dimensionality Reduction Issues**

#### Reproducible?

Projections are *data-dependent*. Fitting a new projection with different data can give rise to different results.

#### **Reusable?**

PCA and UMAP provide reusable projection functions that can map new points from high-D to low-D. t-SNE (and others, like MDS) do not provide this.

#### Interpretable?

DR plots are hard to interpret! Try multiple methods and hyperparameter settings. Inspect via interaction!



### Time Curves [Bach et al. '16]



(a) Folding time

U.S. Precipitation over 1 Year



# Rover Telemetry [Guy '16]

How to track high-dimensional state?

t Number 7	Timestamp	Event Description
0	0s	Rover begins traveling forward along smooth terrain.
1	188s	Rover begins descending into crater.
2	223s	Rover loses line of sight with lander and packet drops begin.
2	287s	Rover enters shade, causing temp, comms, and power drops.
3	300s	Rover begins traversing smooth bottom of crater.
3	330s	Rover begins climbing out of crater.
3	343s	Rover exits shade; continues uphill.
5	534s	Rover emerges from crater and enters smooth terrain.
5	594s	Rover enters choppy terrain.
6	643s	Rover wheel has fault; rover stops moving.
2 3 3 3 5 5 5 6	287s 300s 330s 343s 534s 594s 643s	Rover enters shade, causing temp, comms, and power drop Rover begins traversing smooth bottom of crater. Rover begins climbing out of crater. Rover exits shade; continues uphill. Rover emerges from crater and enters smooth terrain. Rover enters choppy terrain. Rover wheel has fault; rover stops moving.



#### Using Spearman Correlation Matrix