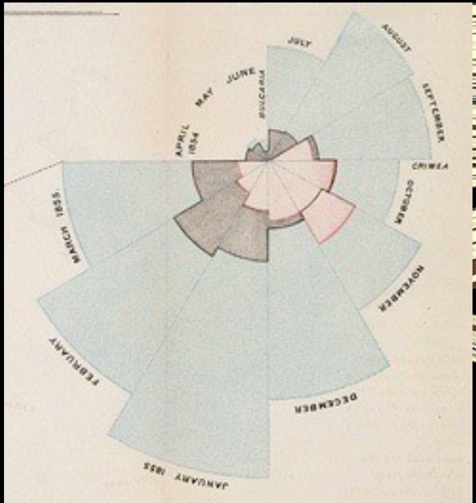
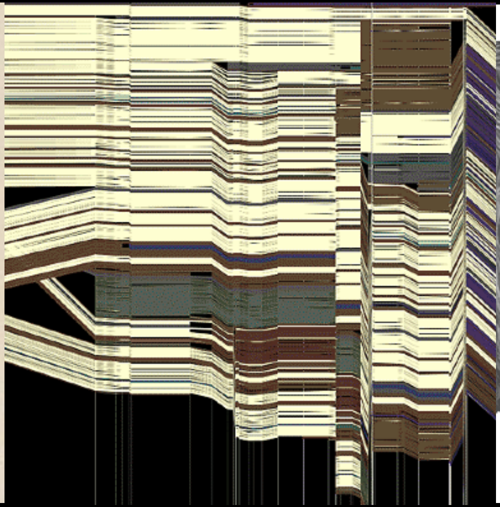
CSE 442 - Data Visualization

Data Transformation





Jeffrey Heer University of Washington

Session Outline

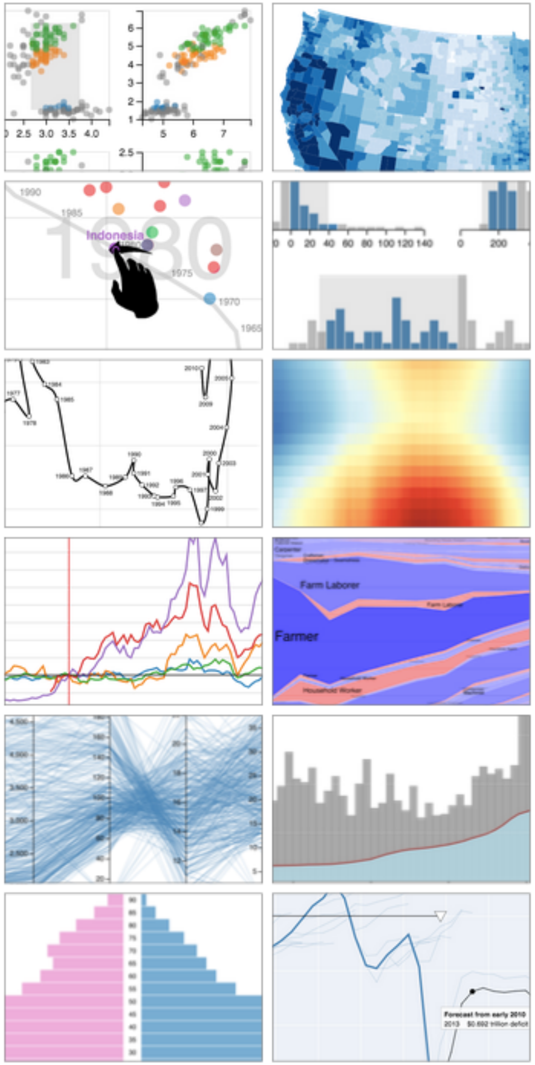
Data Models

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. temperatures
3D 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]

Nominal, Ordinal & Quantitative

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Time duration, …
- Counts and amounts

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Operations: =, ≠

O - Ordered

- Operations: =, ≠, <, >

Q - Interval (location of zero arbitrary)

- Operations: =, ≠, <, >, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, ≠, <, >, -, %
- 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

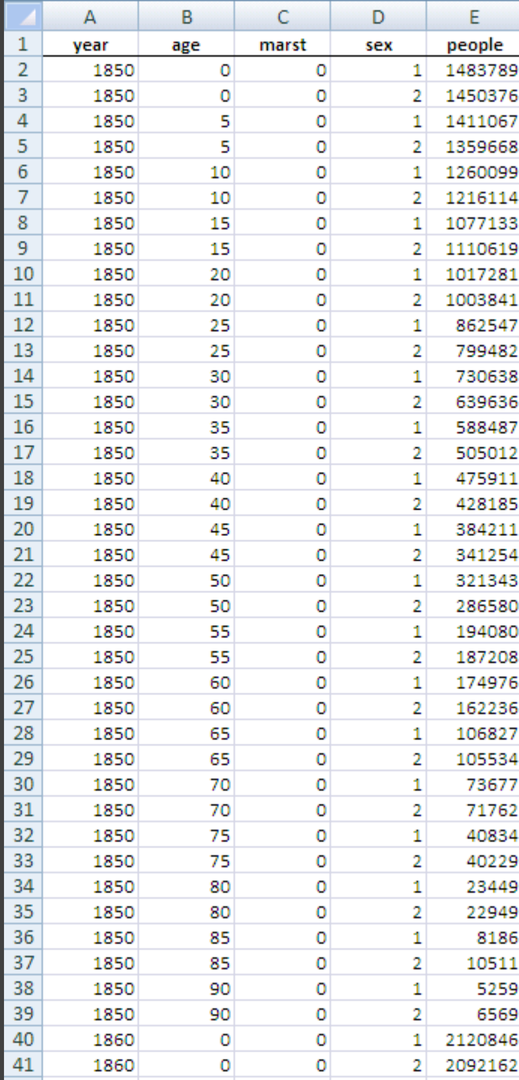
Year

Age

Sex

Marital Status

2,348 data points



	A	B	C	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	N
Marital Status	N

Census: Dimension or Measure?

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

Relational Algebra [Codd ’70] / SQL

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

Relational Algebra [Codd '70] / SQL

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

Relational Algebra [Codd ’70] / SQL

Project (`select`): select a set of columns

```
select day, stock
```

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

→

day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT

Relational Algebra [Codd '70] / SQL

Filter (`where`): remove unwanted rows

```
select * where price > 100
```

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

→

day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45

Relational Algebra [Codd ’70] / SQL

Sort (`order by`): order records

```
select * order by stock
```

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

→

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

Relational Algebra [Codd '70] / SQL

Aggregate (`group by`, `sum`, `min`, `max`, ...):

```
select stock, min(price) group by stock
```

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

→

stock	min(price)
AMZN	957.10
MSFT	74.26

Relational Algebra [Codd '70] / SQL

Join (join) multiple tables together

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

stock	min
AMZN	957.10
MSFT	74.26

→

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

SELECT $\overbrace{\text{year, age,}}^{\textbf{Dimensions}}$ $\overbrace{\text{sum(people)}}^{\textbf{Measure}}$

FROM census

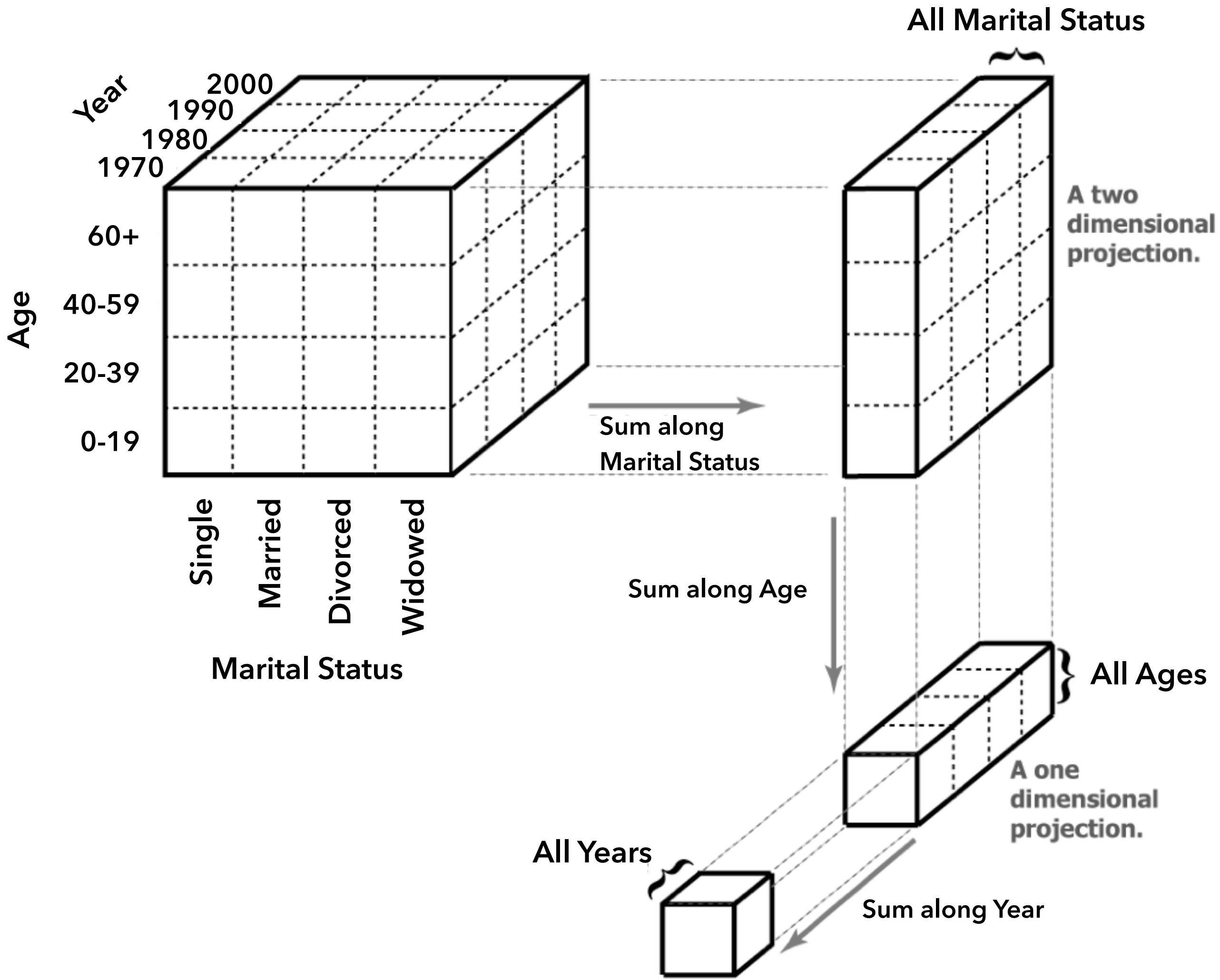
GROUP BY $\underbrace{\text{year, age}}_{\textbf{Dimensions}}$

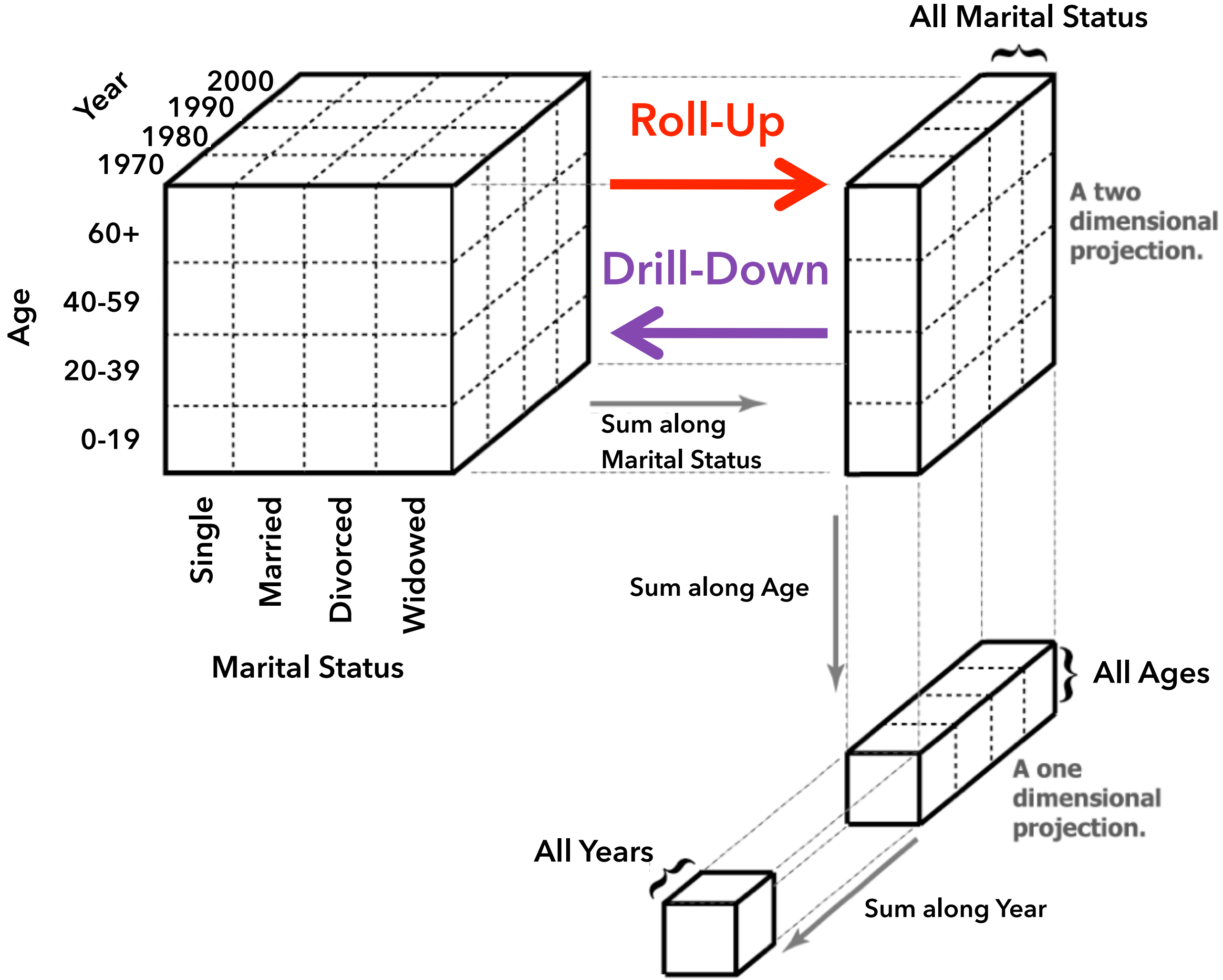
Roll-Up and Drill-Down

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
```


All Marital Status
Year
2000
1990
1980
1970
Age
60+
40-59
20-39
0-19
Single
Married
Divorced
Widowed
Marital Status
Sum along Marital Status
A two dimensional projection.
Sum along Age
All Ages
A one dimensional projection.
All Years
Sum along Year



Roll-Up

Drill-Down

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 normalized forms 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)

```
[
 {"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)

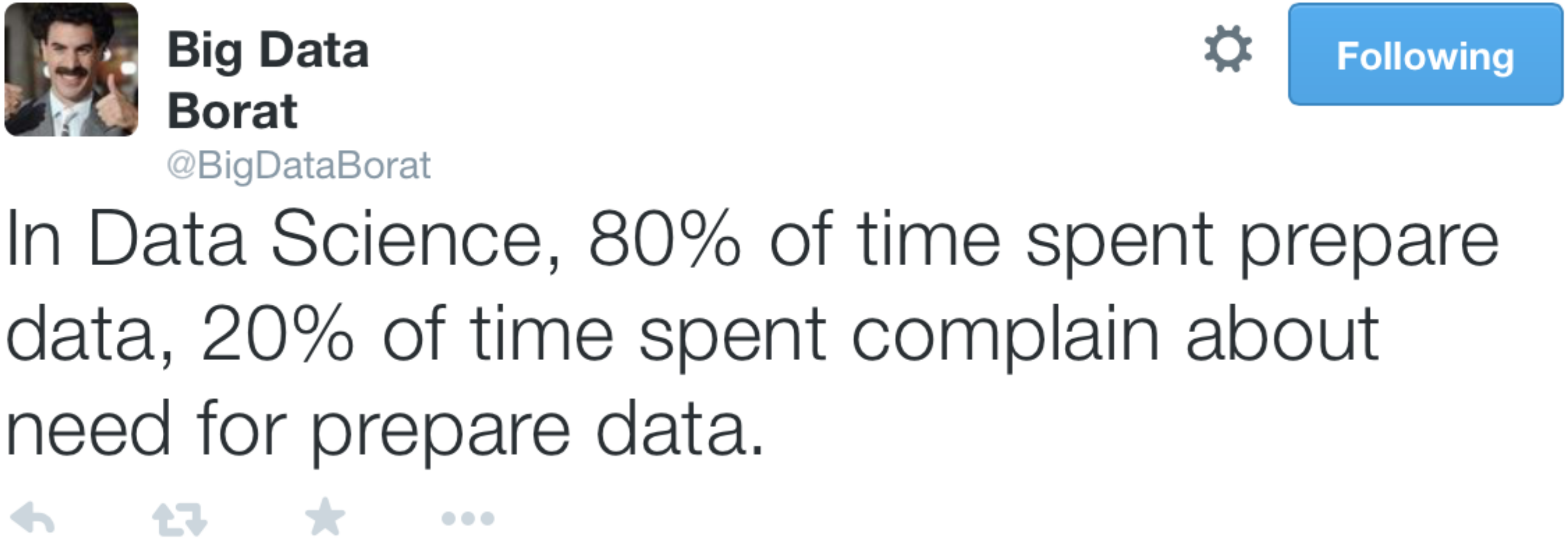
```
[
 {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},
 {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},
 ...
]
```

Binary Formats: Arrow, Parquet, …

Data Wrangling

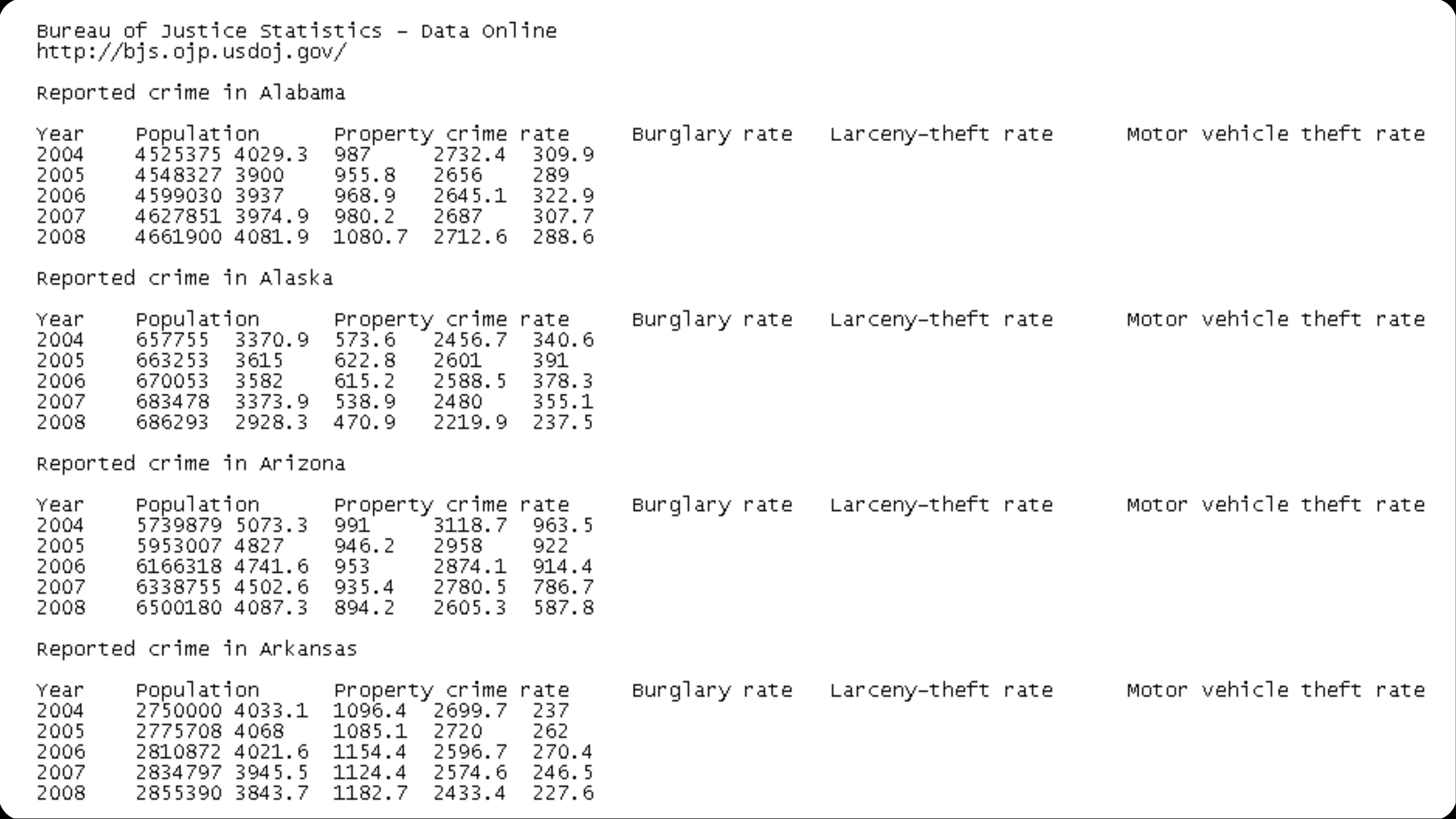
I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any “analysis” at all.

Anonymous Data Scientist
from our 2012 interview study



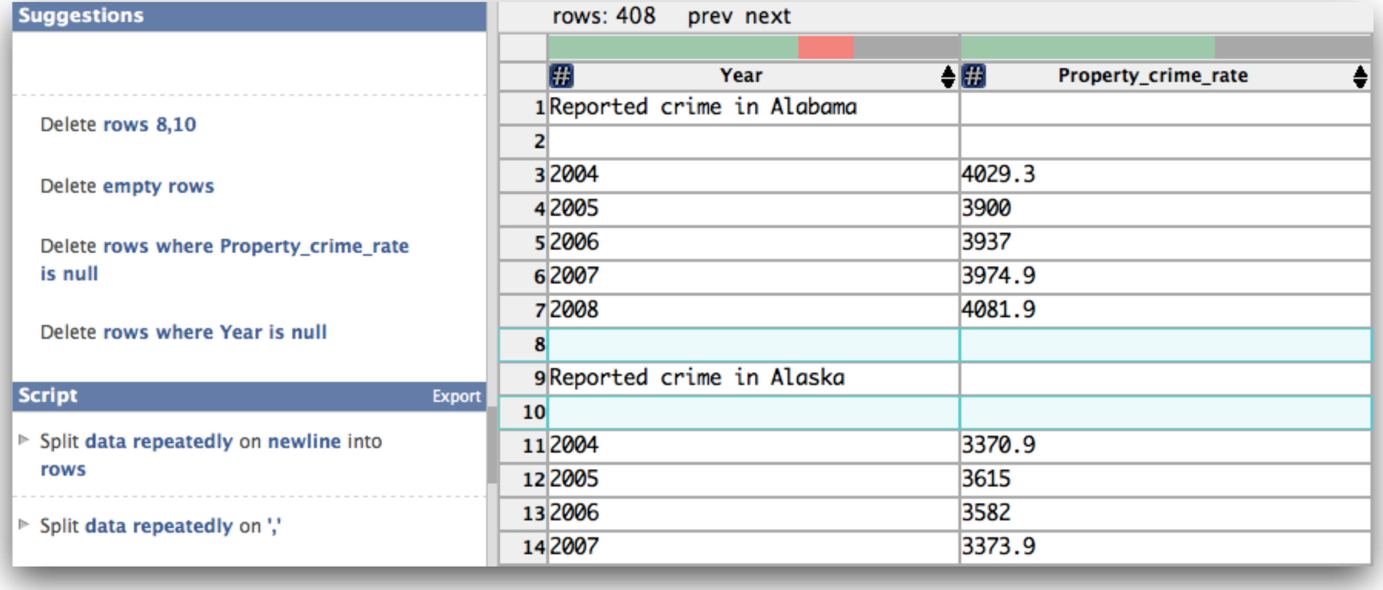
Big Data Borat
@BigDataBorat

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



```
Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year     Population       Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004     4525375 4029.3   987      2732.4   309.9
2005     4548327 3900     955.8    2656     289
2006     4599030 3937     968.9    2645.1   322.9
2007     4627851 3974.9   980.2    2687     307.7
2008     4661900 4081.9   1080.7   2712.6   288.6

Reported crime in Alaska

Year     Population       Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004     657755   3370.9  573.6    2456.7   340.6
2005     663253   3615    622.8    2601     391
2006     670053   3582    615.2    2588.5   378.3
2007     683478   3373.9  538.9    2480     355.1
2008     686293   2928.3  470.9    2219.9   237.5

Reported crime in Arizona

Year     Population       Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004     5739879 5073.3   991      3118.7   963.5
2005     5953007 4827     946.2    2958     922
2006     6166318 4741.6   953      2874.1   914.4
2007     6338755 4502.6   935.4    2780.5   786.7
2008     6500180 4087.3   894.2    2605.3   587.8

Reported crime in Arkansas

Year     Population       Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004     2750000 4033.1   1096.4   2699.7   237
2005     2775708 4068     1085.1   2720     262
2006     2810872 4021.6   1154.4   2596.7   270.4
2007     2834797 3945.5   1124.4   2574.6   246.5
2008     2855390 3843.7   1182.7   2433.4   227.6
```

DataWrangler



Suggestions

Delete rows 8,10

Delete empty rows

Delete rows where Property_crime_rate is null

Delete rows where Year is null

Script Export

▸ Split data repeatedly on newline into rows

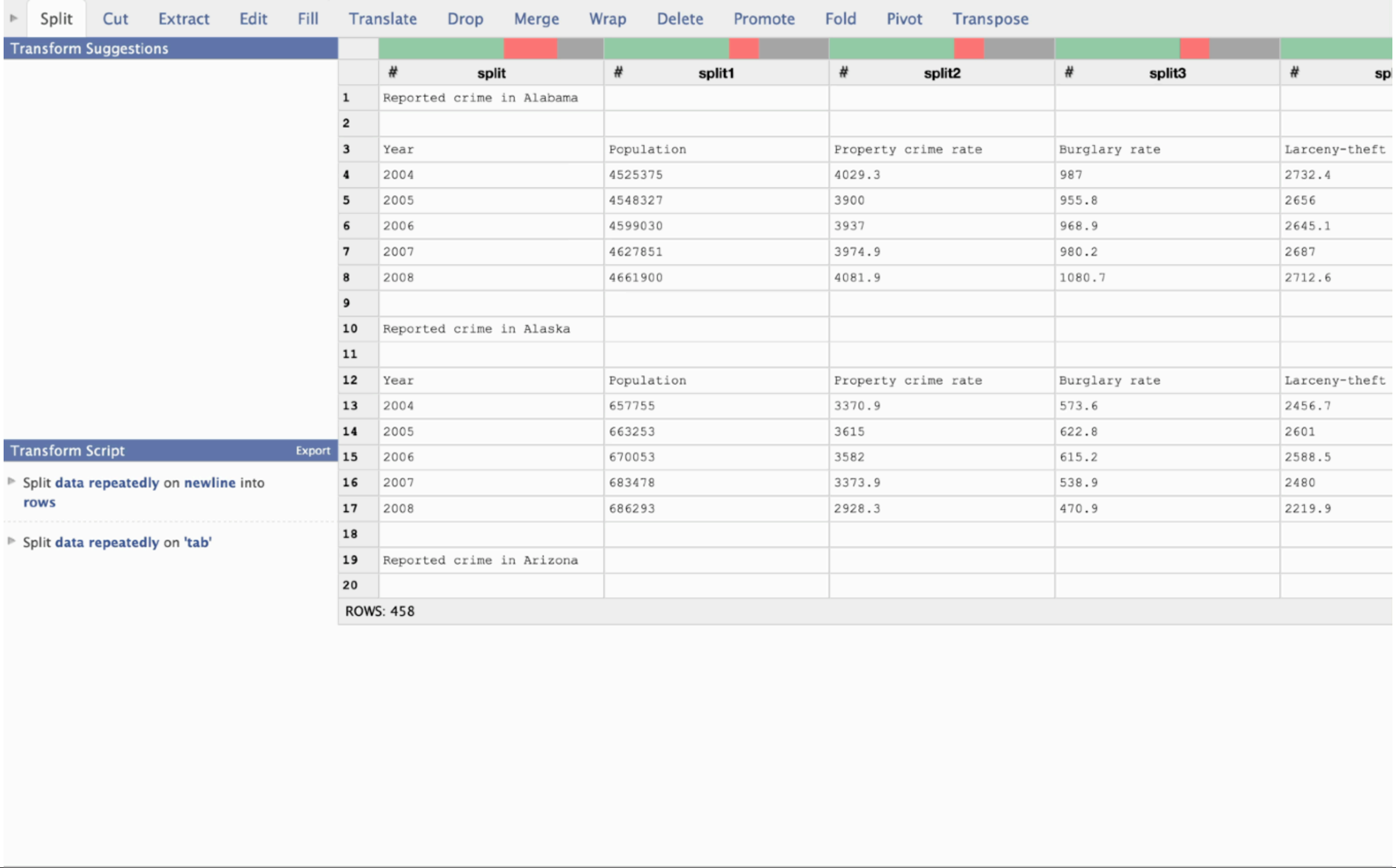
▸ Split data repeatedly on ','

rows: 408 prev next

	Year	Property_crime_rate
1	Reported crime in Alabama	
2		
3	2004	4029.3
4	2005	3900
5	2006	3937
6	2007	3974.9
7	2008	4081.9
8		
9	Reported crime in Alaska	
10		
11	2004	3370.9
12	2005	3615
13	2006	3582
14	2007	3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts

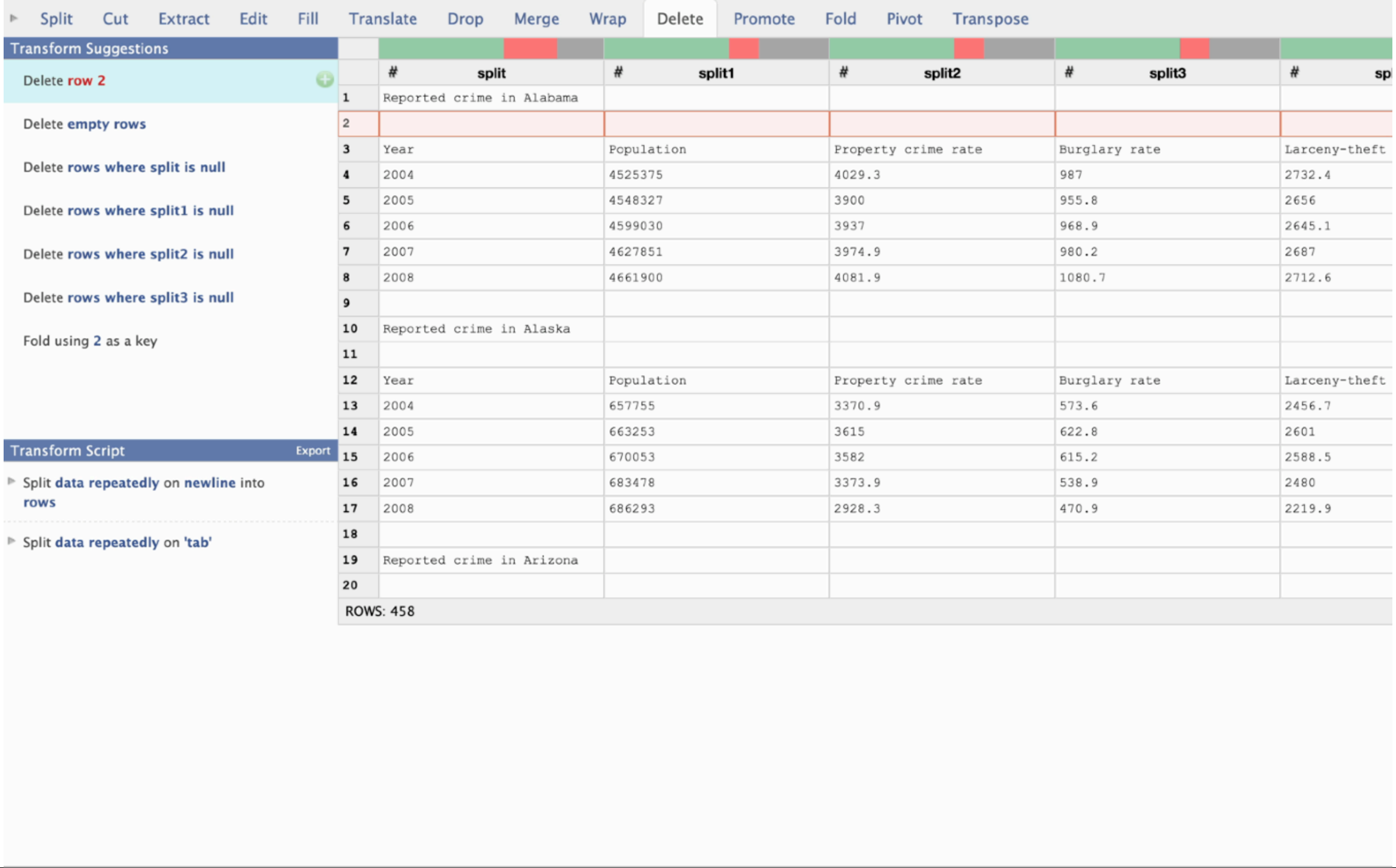
Kandel et al. [CHI 2011]



	# split	# split1	# split2	# split3	# sp
1	Reported crime in Alabama				
2					
3	Year	Population	Property crime rate	Burglary rate	Larceny-theft
4	2004	4525375	4029.3	987	2732.4
5	2005	4548327	3900	955.8	2656
6	2006	4599030	3937	968.9	2645.1
7	2007	4627851	3974.9	980.2	2687
8	2008	4661900	4081.9	1080.7	2712.6
9					
10	Reported crime in Alaska				
11					
12	Year	Population	Property crime rate	Burglary rate	Larceny-theft
13	2004	657755	3370.9	573.6	2456.7
14	2005	663253	3615	622.8	2601
15	2006	670053	3582	615.2	2588.5
16	2007	683478	3373.9	538.9	2480
17	2008	686293	2928.3	470.9	2219.9
18					
19	Reported crime in Arizona				
20					

ROWS: 458

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'



Transform Suggestions

Delete row 2

Delete empty rows

Delete rows where split is null

Delete rows where split1 is null

Delete rows where split2 is null

Delete rows where split3 is null

Fold using 2 as a key

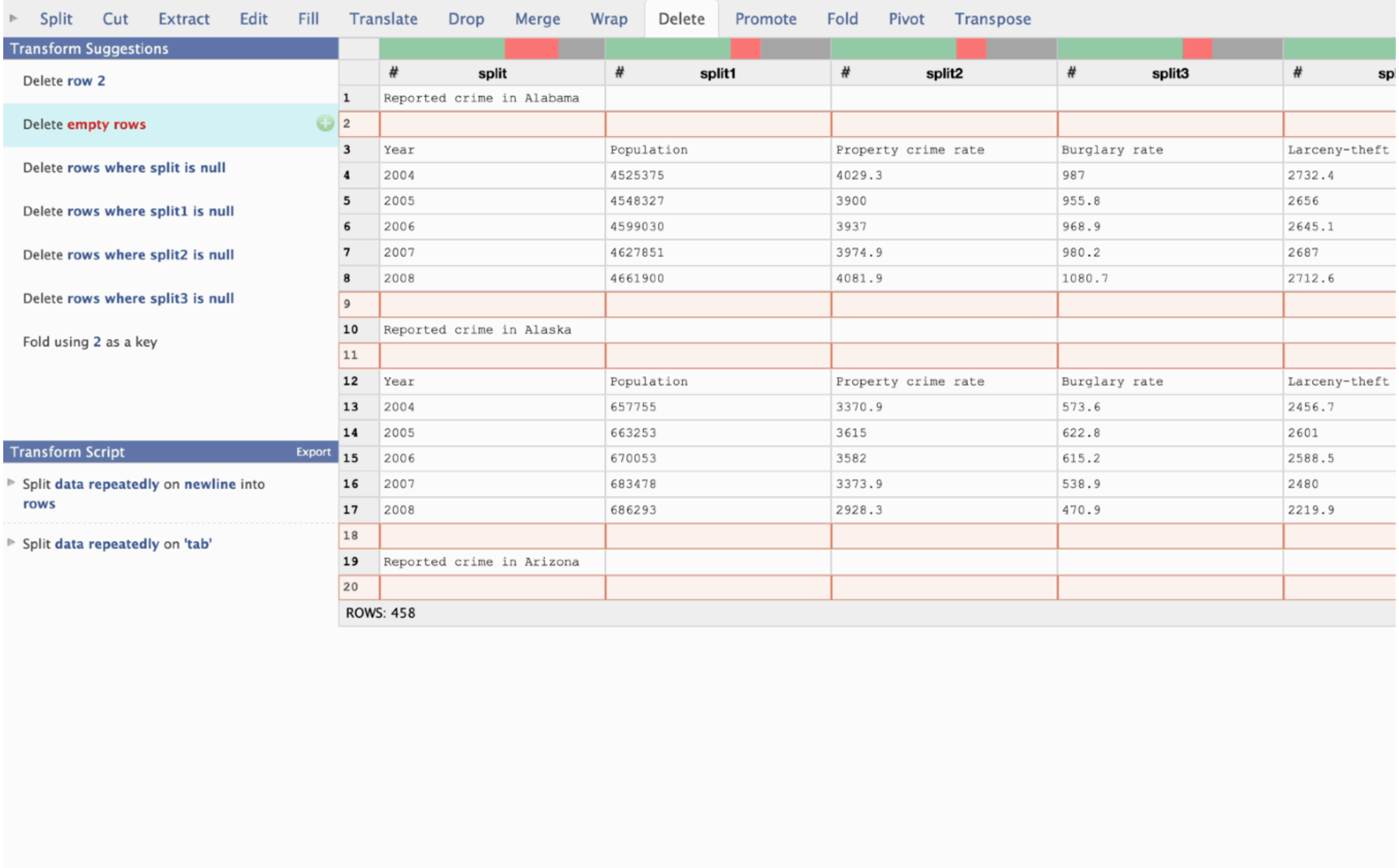
Transform Script

Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

	# split	# split1	# split2	# split3	# sp
1	Reported crime in Alabama				
2					
3	Year	Population	Property crime rate	Burglary rate	Larceny-theft
4	2004	4525375	4029.3	987	2732.4
5	2005	4548327	3900	955.8	2656
6	2006	4599030	3937	968.9	2645.1
7	2007	4627851	3974.9	980.2	2687
8	2008	4661900	4081.9	1080.7	2712.6
9					
10	Reported crime in Alaska				
11					
12	Year	Population	Property crime rate	Burglary rate	Larceny-theft
13	2004	657755	3370.9	573.6	2456.7
14	2005	663253	3615	622.8	2601
15	2006	670053	3582	615.2	2588.5
16	2007	683478	3373.9	538.9	2480
17	2008	686293	2928.3	470.9	2219.9
18					
19	Reported crime in Arizona				
20					

ROWS: 458



Transform Suggestions

Delete row 2

Delete empty rows

Delete rows where split is null

Delete rows where split1 is null

Delete rows where split2 is null

Delete rows where split3 is null

Fold using 2 as a key

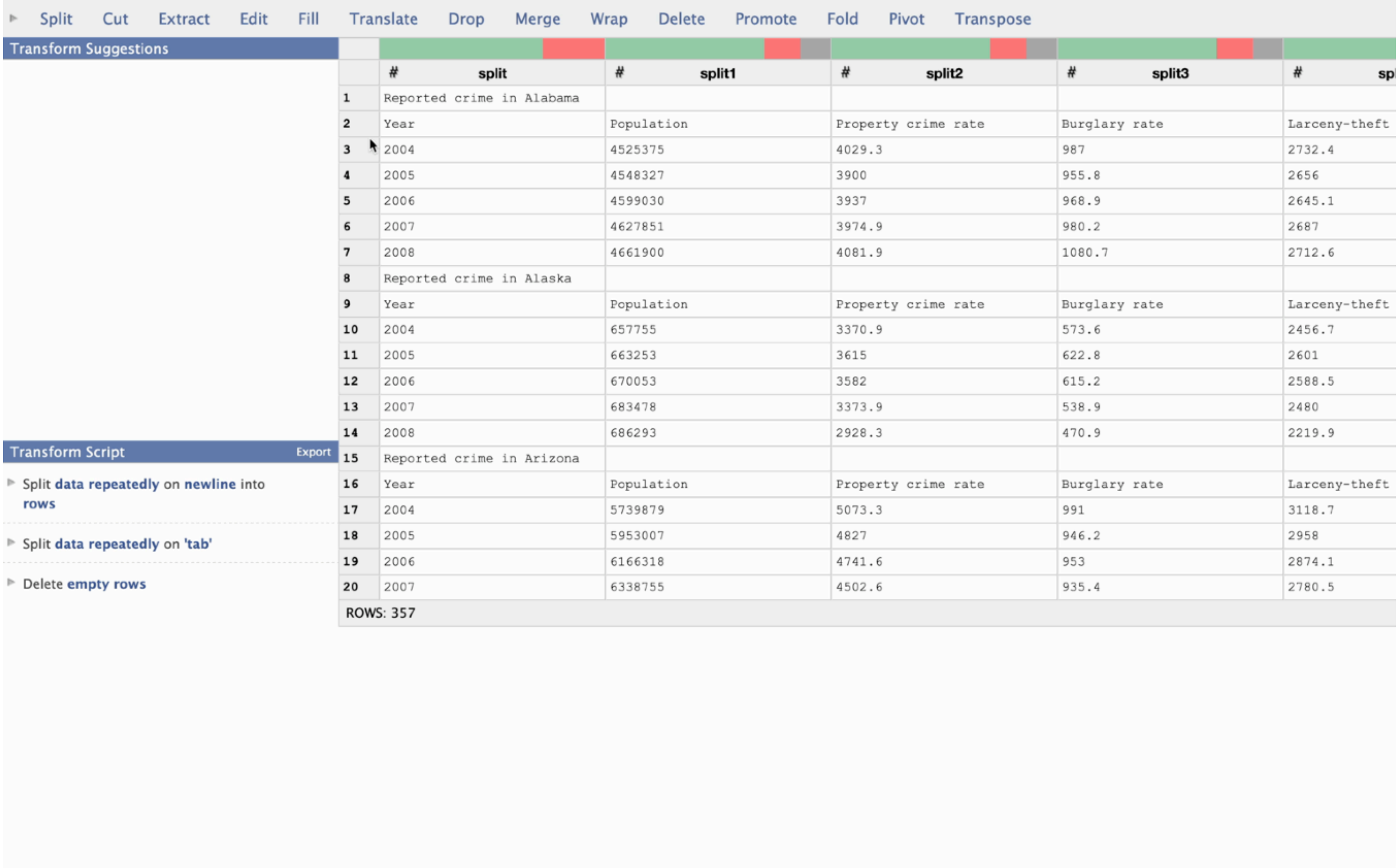
Transform Script

Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

	# split	# split1	# split2	# split3	# sp
1	Reported crime in Alabama				
2					
3	Year	Population	Property crime rate	Burglary rate	Larceny-theft
4	2004	4525375	4029.3	987	2732.4
5	2005	4548327	3900	955.8	2656
6	2006	4599030	3937	968.9	2645.1
7	2007	4627851	3974.9	980.2	2687
8	2008	4661900	4081.9	1080.7	2712.6
9					
10	Reported crime in Alaska				
11					
12	Year	Population	Property crime rate	Burglary rate	Larceny-theft
13	2004	657755	3370.9	573.6	2456.7
14	2005	663253	3615	622.8	2601
15	2006	670053	3582	615.2	2588.5
16	2007	683478	3373.9	538.9	2480
17	2008	686293	2928.3	470.9	2219.9
18					
19	Reported crime in Arizona				
20					

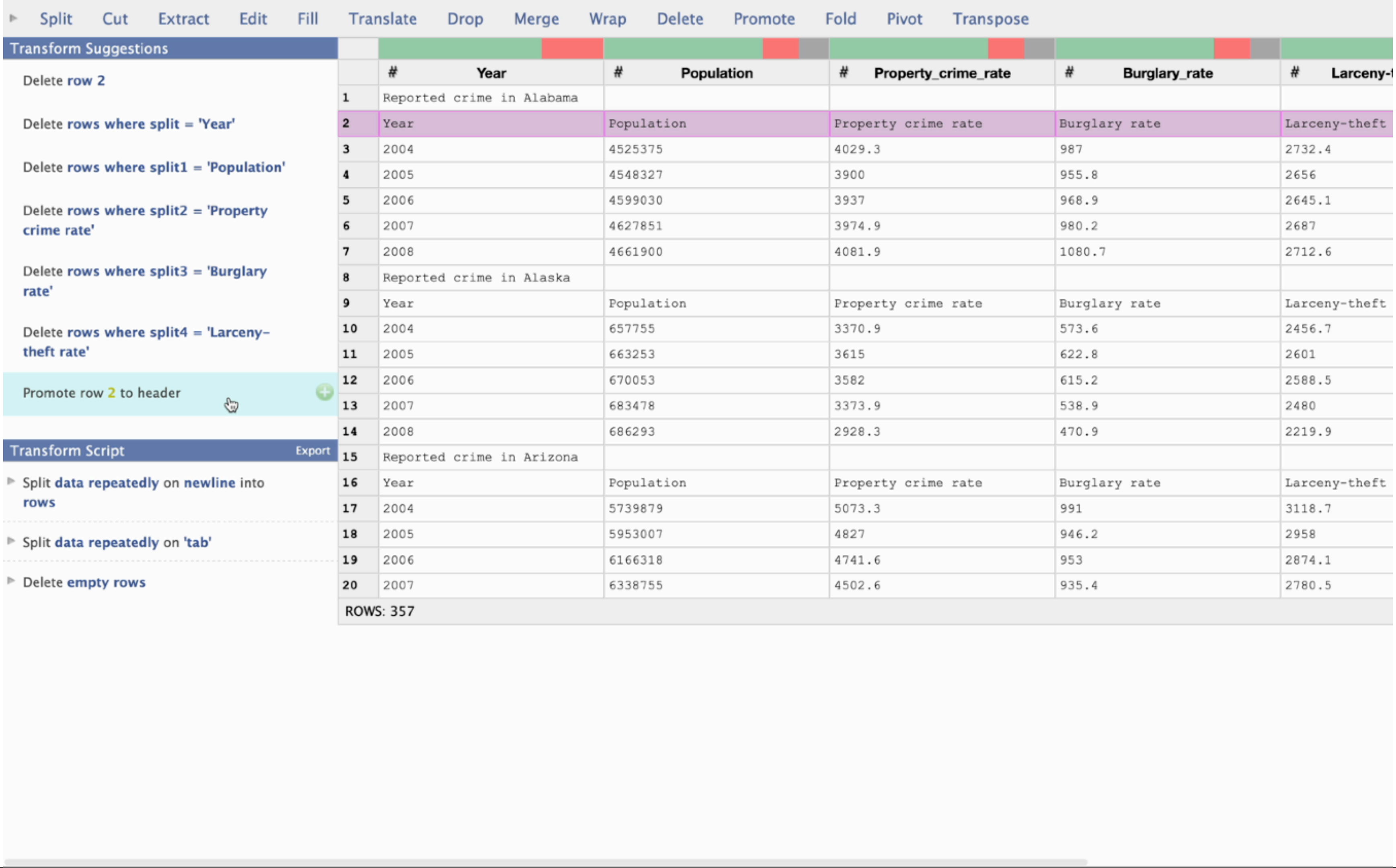
ROWS: 458



- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows

	# split	# split1	# split2	# split3	# sp
1	Reported crime in Alabama				
2	Year	Population	Property crime rate	Burglary rate	Larceny-theft
3	2004	4525375	4029.3	987	2732.4
4	2005	4548327	3900	955.8	2656
5	2006	4599030	3937	968.9	2645.1
6	2007	4627851	3974.9	980.2	2687
7	2008	4661900	4081.9	1080.7	2712.6
8	Reported crime in Alaska				
9	Year	Population	Property crime rate	Burglary rate	Larceny-theft
10	2004	657755	3370.9	573.6	2456.7
11	2005	663253	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
15	Reported crime in Arizona				
16	Year	Population	Property crime rate	Burglary rate	Larceny-theft
17	2004	5739879	5073.3	991	3118.7
18	2005	5953007	4827	946.2	2958
19	2006	6166318	4741.6	953	2874.1
20	2007	6338755	4502.6	935.4	2780.5

ROWS: 357



Transform Suggestions

Delete row 2

Delete rows where split = 'Year'

Delete rows where split1 = 'Population'

Delete rows where split2 = 'Property crime rate'

Delete rows where split3 = 'Burglary rate'

Delete rows where split4 = 'Larceny-theft rate'

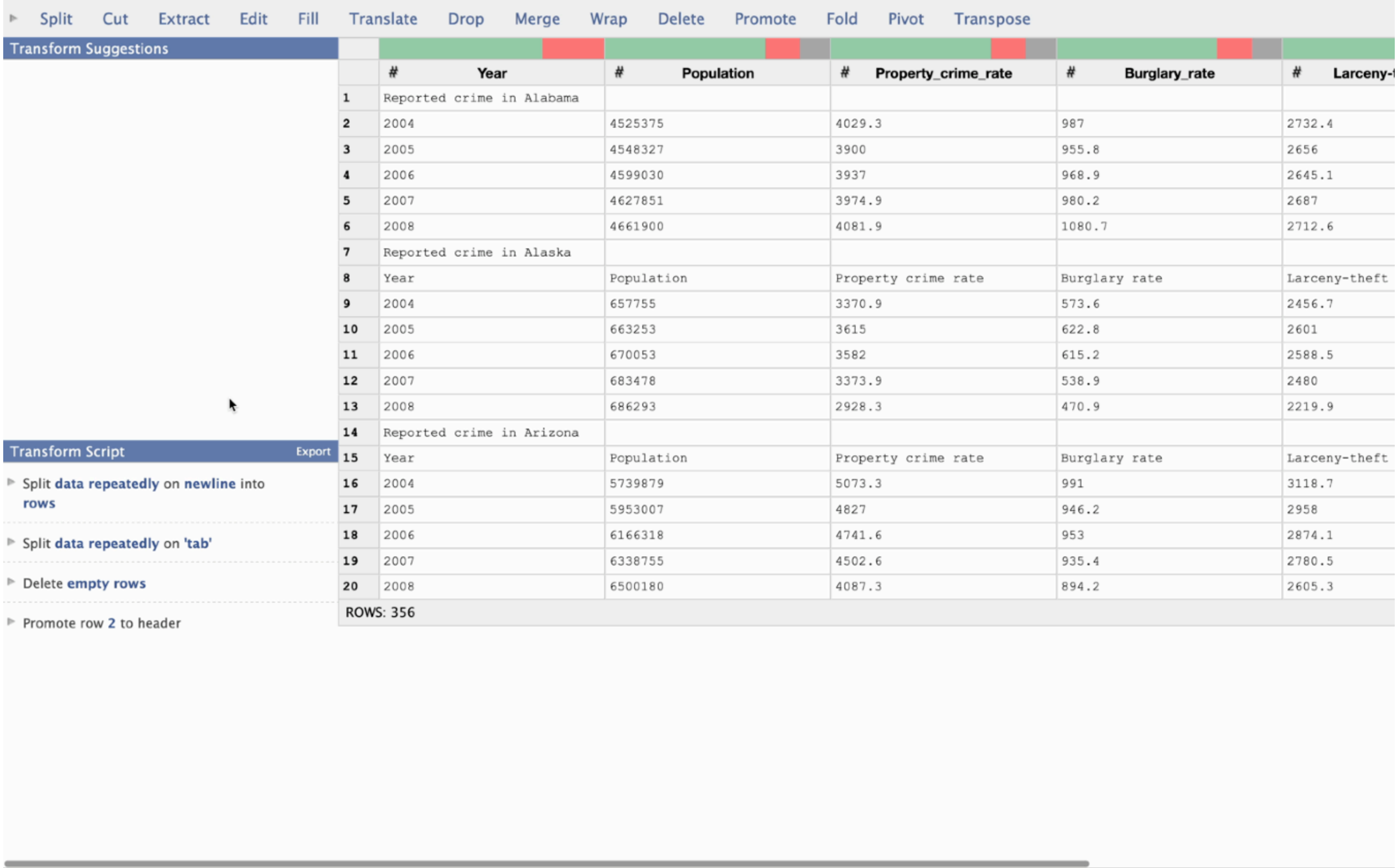
Promote row 2 to header

Transform Script

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows

	# Year	# Population	# Property_crime_rate	# Burglary_rate	# Larceny-
1	Reported crime in Alabama				
2	Year	Population	Property crime rate	Burglary rate	Larceny-theft
3	2004	4525375	4029.3	987	2732.4
4	2005	4548327	3900	955.8	2656
5	2006	4599030	3937	968.9	2645.1
6	2007	4627851	3974.9	980.2	2687
7	2008	4661900	4081.9	1080.7	2712.6
8	Reported crime in Alaska				
9	Year	Population	Property crime rate	Burglary rate	Larceny-theft
10	2004	657755	3370.9	573.6	2456.7
11	2005	663253	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
15	Reported crime in Arizona				
16	Year	Population	Property crime rate	Burglary rate	Larceny-theft
17	2004	5739879	5073.3	991	3118.7
18	2005	5953007	4827	946.2	2958
19	2006	6166318	4741.6	953	2874.1
20	2007	6338755	4502.6	935.4	2780.5

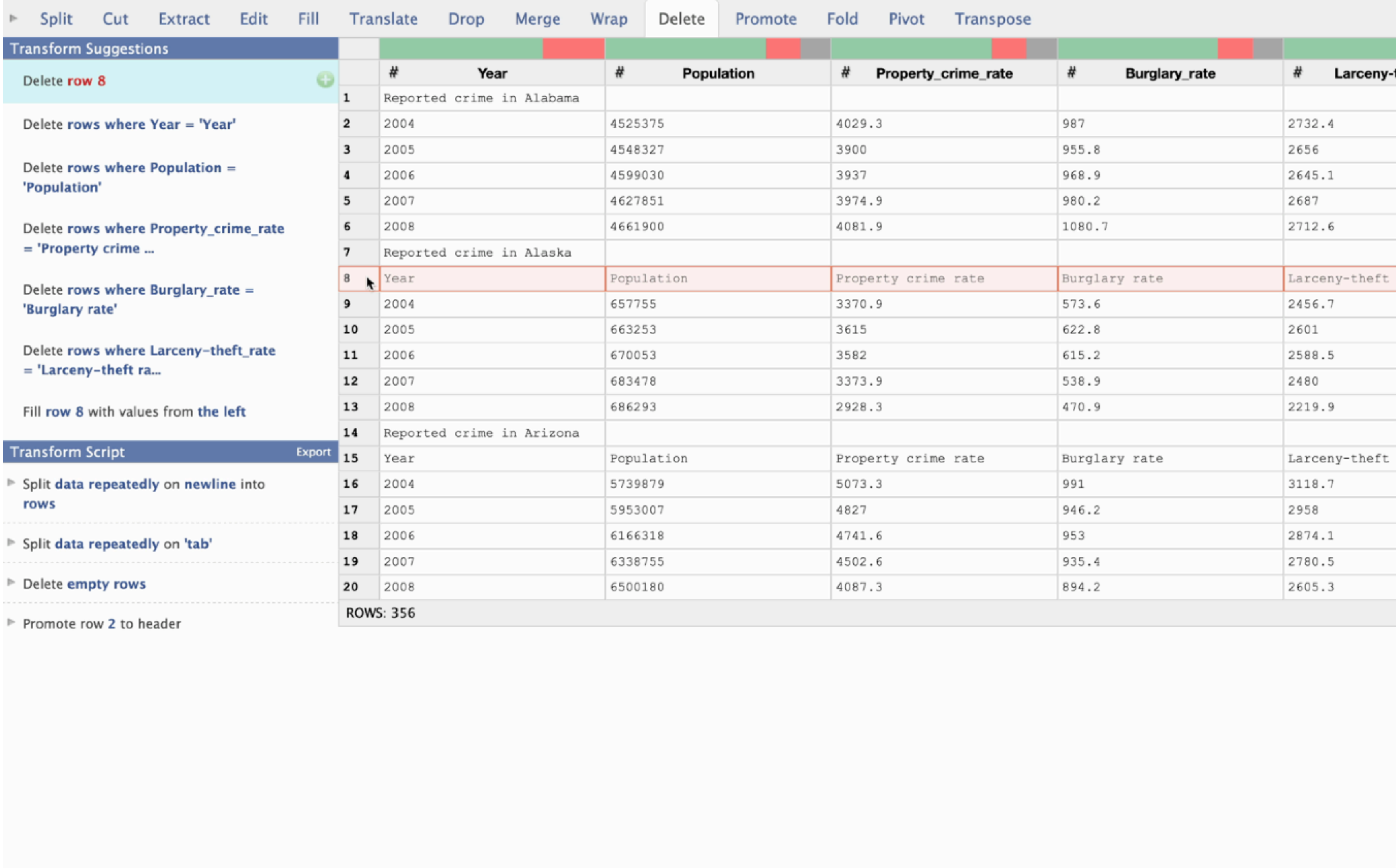
ROWS: 357



	# 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	Year	Population	Property crime rate	Burglary rate	Larceny-theft
9	2004	657755	3370.9	573.6	2456.7
10	2005	663253	3615	622.8	2601
11	2006	670053	3582	615.2	2588.5
12	2007	683478	3373.9	538.9	2480
13	2008	686293	2928.3	470.9	2219.9
14	Reported crime in Arizona				
15	Year	Population	Property crime rate	Burglary rate	Larceny-theft
16	2004	5739879	5073.3	991	3118.7
17	2005	5953007	4827	946.2	2958
18	2006	6166318	4741.6	953	2874.1
19	2007	6338755	4502.6	935.4	2780.5
20	2008	6500180	4087.3	894.2	2605.3

ROWS: 356

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header



Delete row 8

Delete rows where Year = 'Year'

Delete rows where Population = 'Population'

Delete rows where Property_crime_rate = 'Property crime ...

Delete rows where Burglary_rate = 'Burglary rate'

Delete rows where Larceny-theft_rate = 'Larceny-theft ra...

Fill row 8 with values from the left

Split data repeatedly on newline into rows

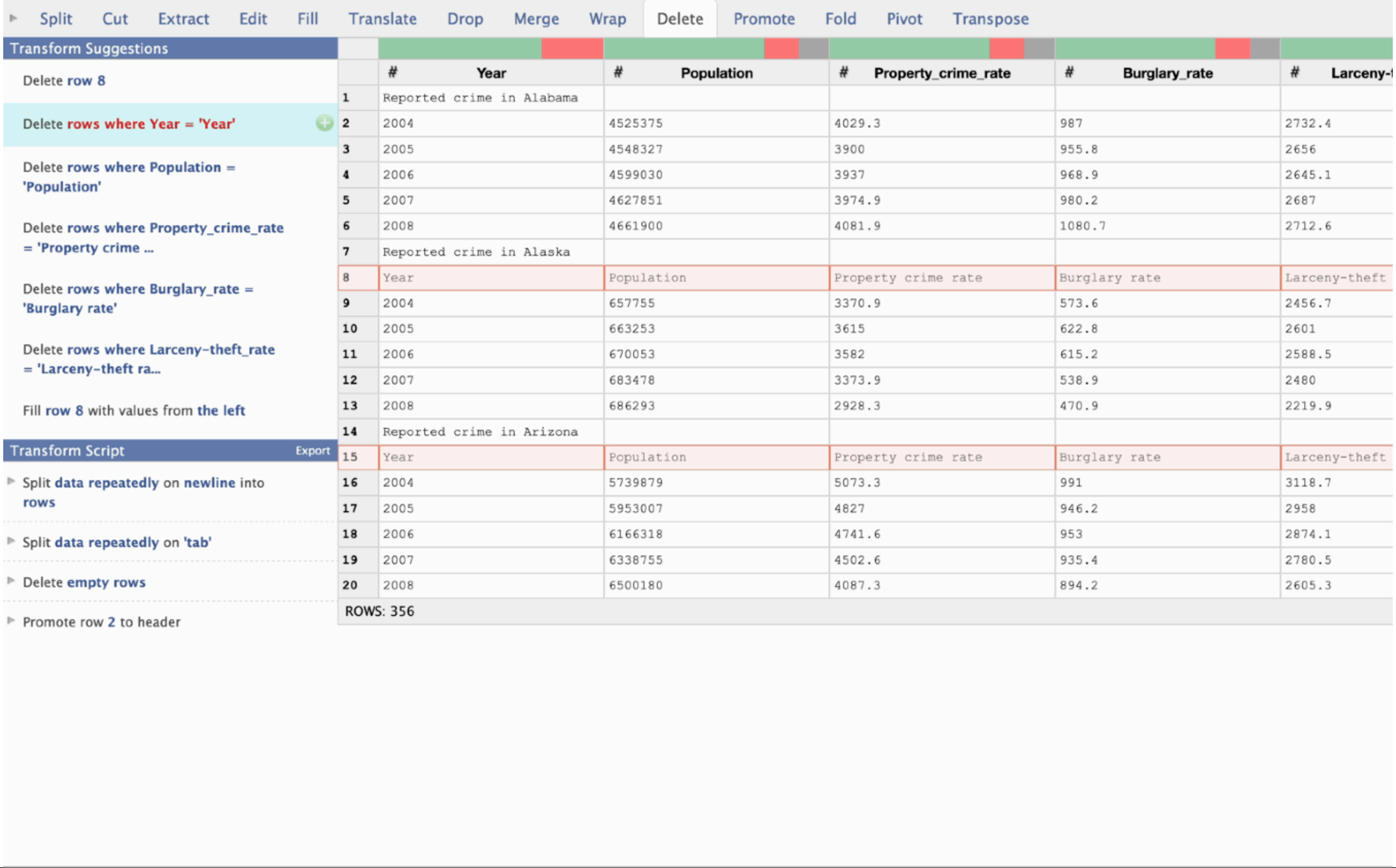
Split data repeatedly on 'tab'

Delete empty rows

Promote row 2 to header

	# Year	# Population	# Property_crime_rate	# Burglary_rate	# Larceny-t
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	Year	Population	Property crime rate	Burglary rate	Larceny-theft
9	2004	657755	3370.9	573.6	2456.7
10	2005	663253	3615	622.8	2601
11	2006	670053	3582	615.2	2588.5
12	2007	683478	3373.9	538.9	2480
13	2008	686293	2928.3	470.9	2219.9
14	Reported crime in Arizona				
15	Year	Population	Property crime rate	Burglary rate	Larceny-theft
16	2004	5739879	5073.3	991	3118.7
17	2005	5953007	4827	946.2	2958
18	2006	6166318	4741.6	953	2874.1
19	2007	6338755	4502.6	935.4	2780.5
20	2008	6500180	4087.3	894.2	2605.3

ROWS: 356



Delete row 8

Delete rows where Year = 'Year'

Delete rows where Population = 'Population'

Delete rows where Property_crime_rate = 'Property crime ...

Delete rows where Burglary_rate = 'Burglary rate'

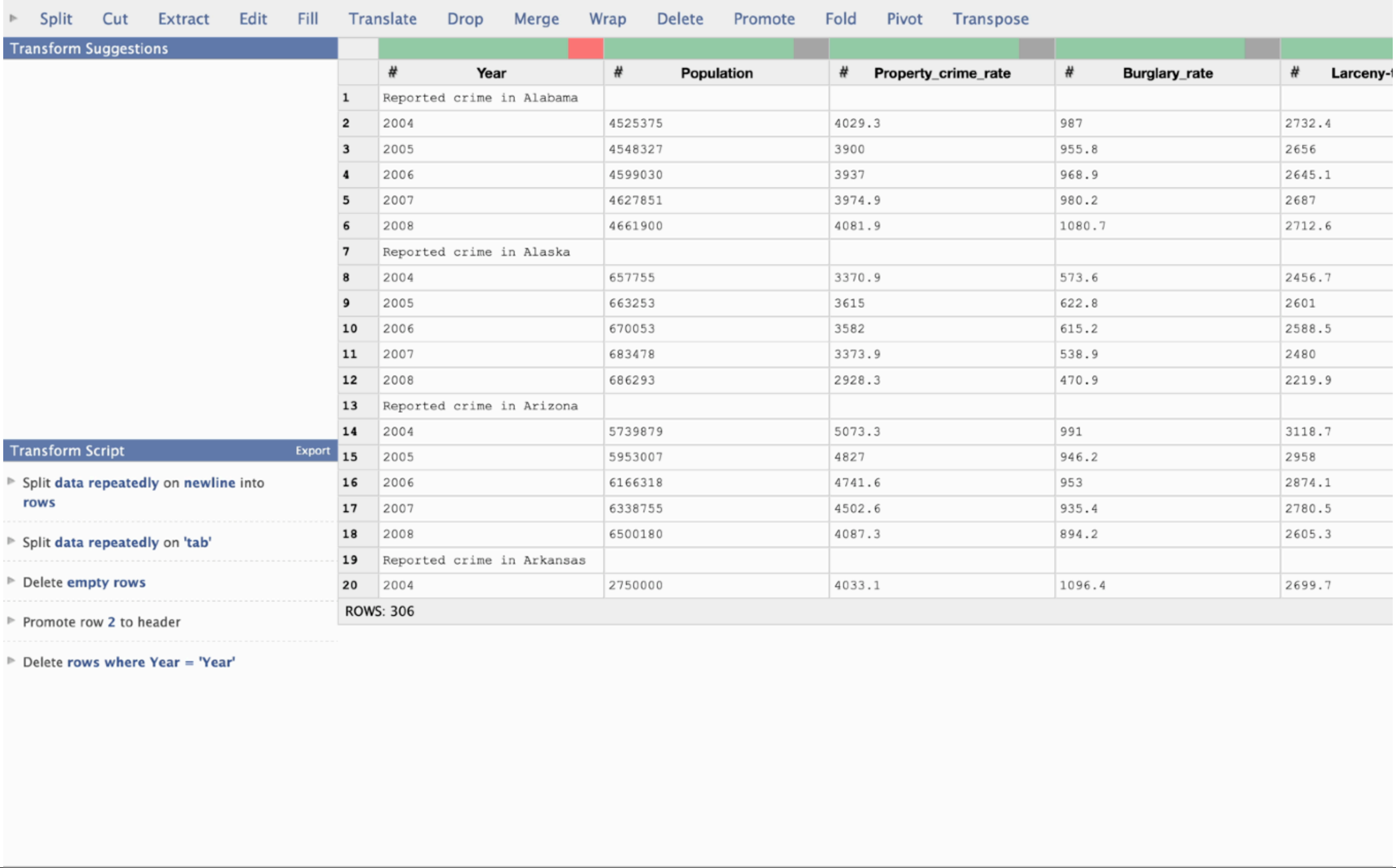
Delete rows where Larceny-theft_rate = 'Larceny-theft ra...

Fill row 8 with values from the left

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header

	# 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	Year	Population	Property crime rate	Burglary rate	Larceny-theft
9	2004	657755	3370.9	573.6	2456.7
10	2005	663253	3615	622.8	2601
11	2006	670053	3582	615.2	2588.5
12	2007	683478	3373.9	538.9	2480
13	2008	686293	2928.3	470.9	2219.9
14	Reported crime in Arizona				
15	Year	Population	Property crime rate	Burglary rate	Larceny-theft
16	2004	5739879	5073.3	991	3118.7
17	2005	5953007	4827	946.2	2958
18	2006	6166318	4741.6	953	2874.1
19	2007	6338755	4502.6	935.4	2780.5
20	2008	6500180	4087.3	894.2	2605.3

ROWS: 356



	# Year	# Population	# Property_crime_rate	# Burglary_rate	# Larceny-t
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
15	2005	5953007	4827	946.2	2958
16	2006	6166318	4741.6	953	2874.1
17	2007	6338755	4502.6	935.4	2780.5
18	2008	6500180	4087.3	894.2	2605.3
19	Reported crime in Arkansas				
20	2004	2750000	4033.1	1096.4	2699.7

ROWS: 306

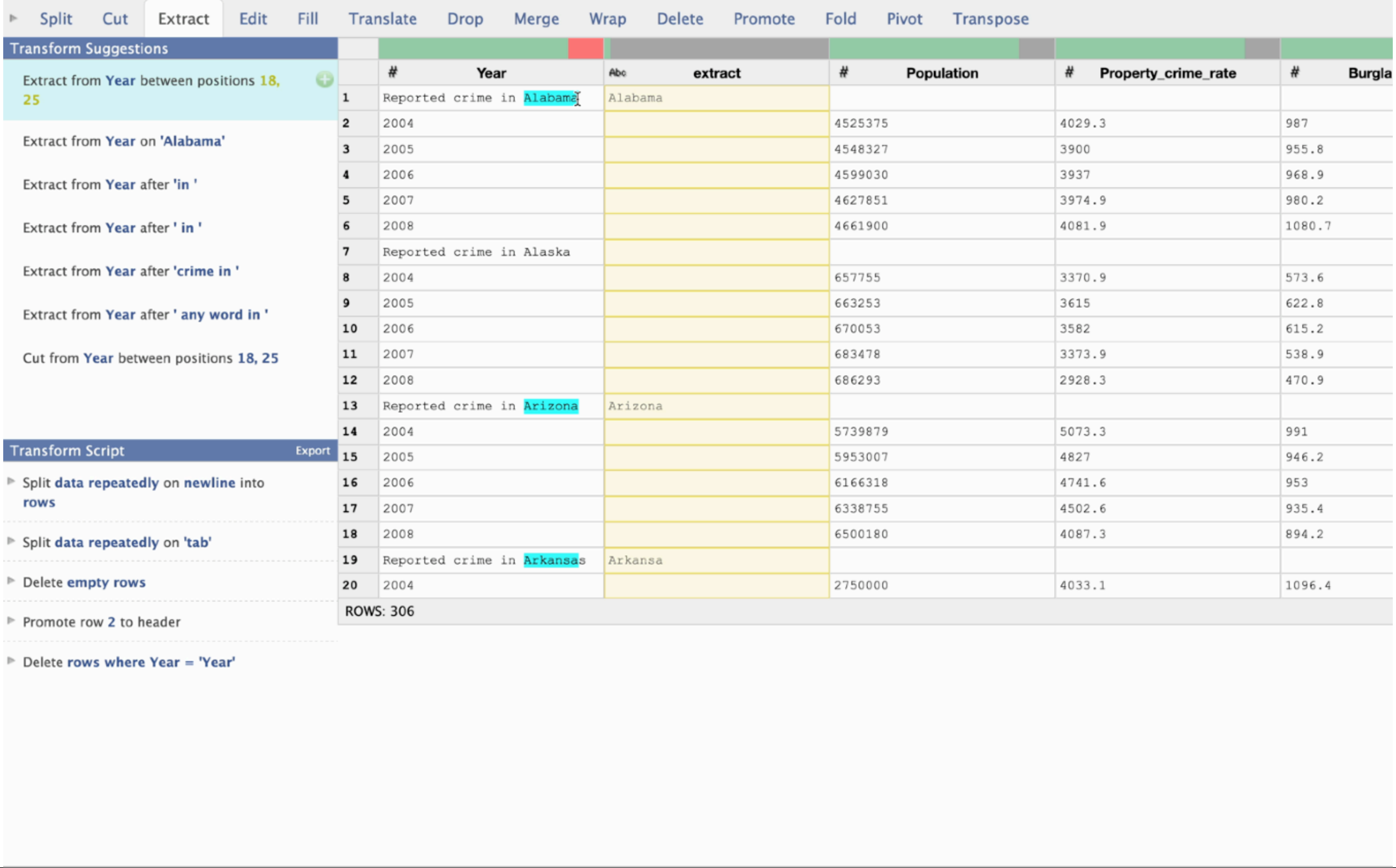
Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

Delete empty rows

Promote row 2 to header

Delete rows where Year = 'Year'



Transform Suggestions

Extract from Year between positions 18, 25

Extract from Year on 'Alabama'

Extract from Year after 'in '

Extract from Year after ' in '

Extract from Year after 'crime in '

Extract from Year after ' any word in '

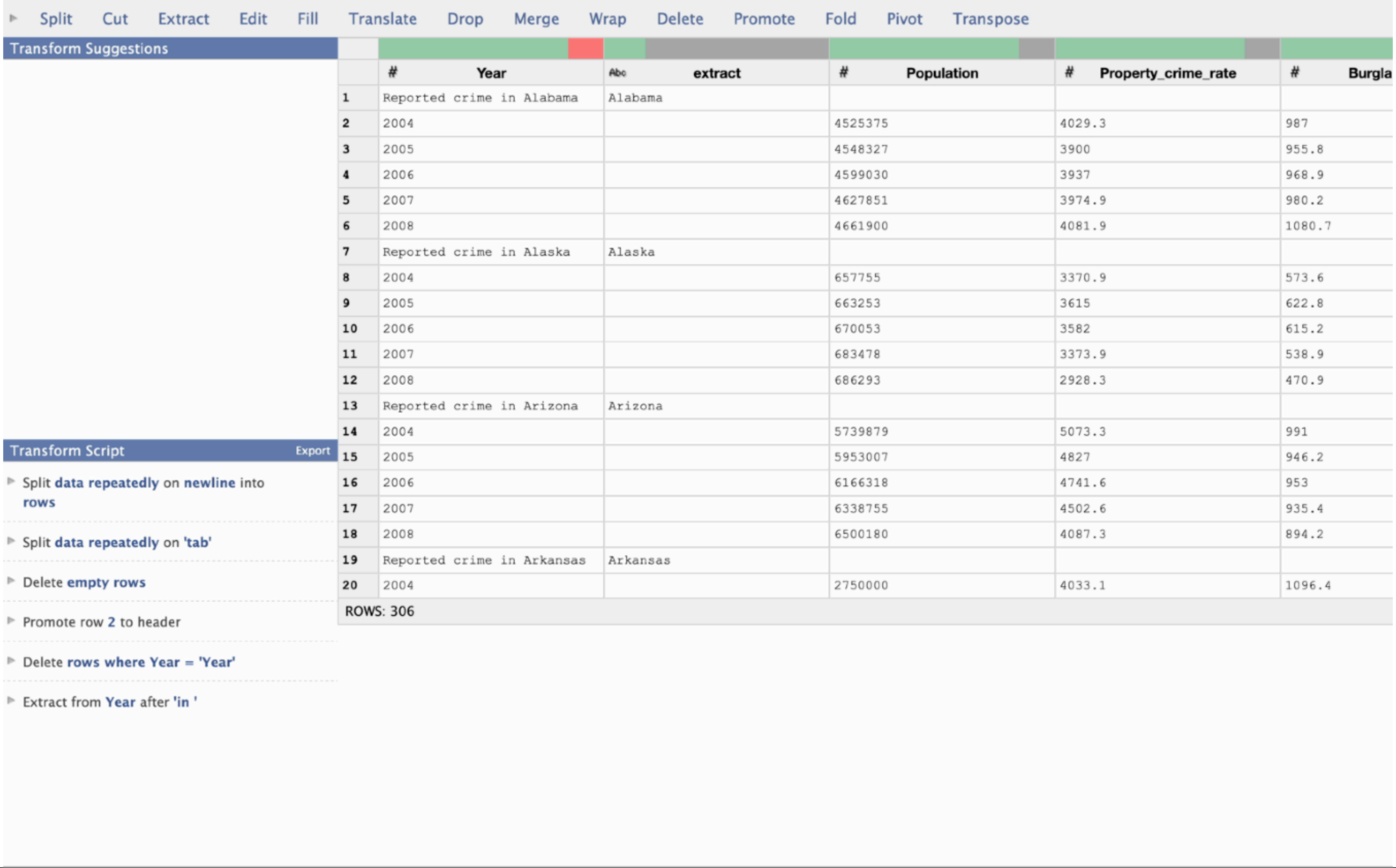
Cut from Year between positions 18, 25

Transform Script

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'

	# Year	Abc 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				
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
15	2005		5953007	4827	946.2
16	2006		6166318	4741.6	953
17	2007		6338755	4502.6	935.4
18	2008		6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansa			
20	2004		2750000	4033.1	1096.4

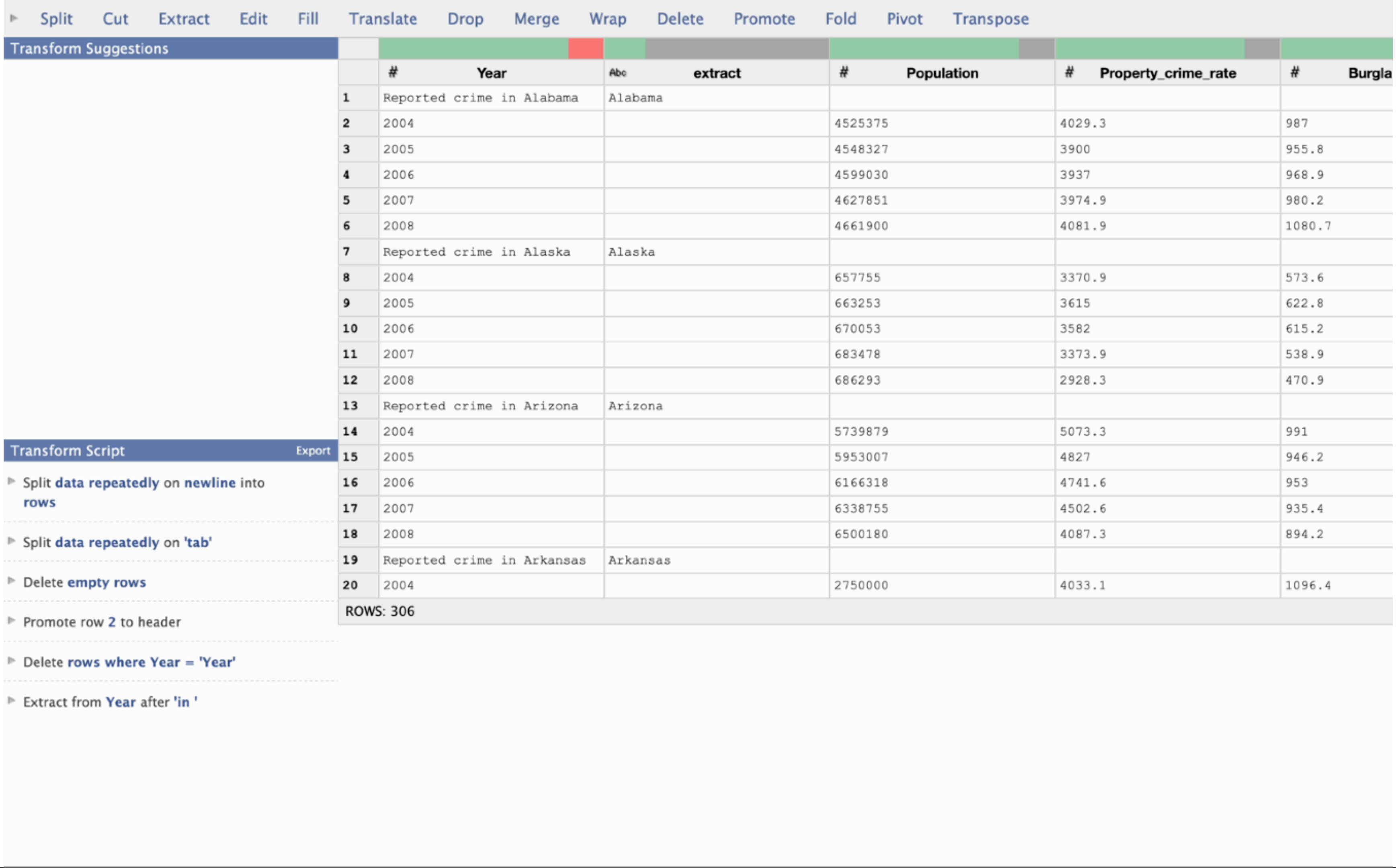
ROWS: 306



- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '

	# Year	Abc 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
15	2005		5953007	4827	946.2
16	2006		6166318	4741.6	953
17	2007		6338755	4502.6	935.4
18	2008		6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004		2750000	4033.1	1096.4

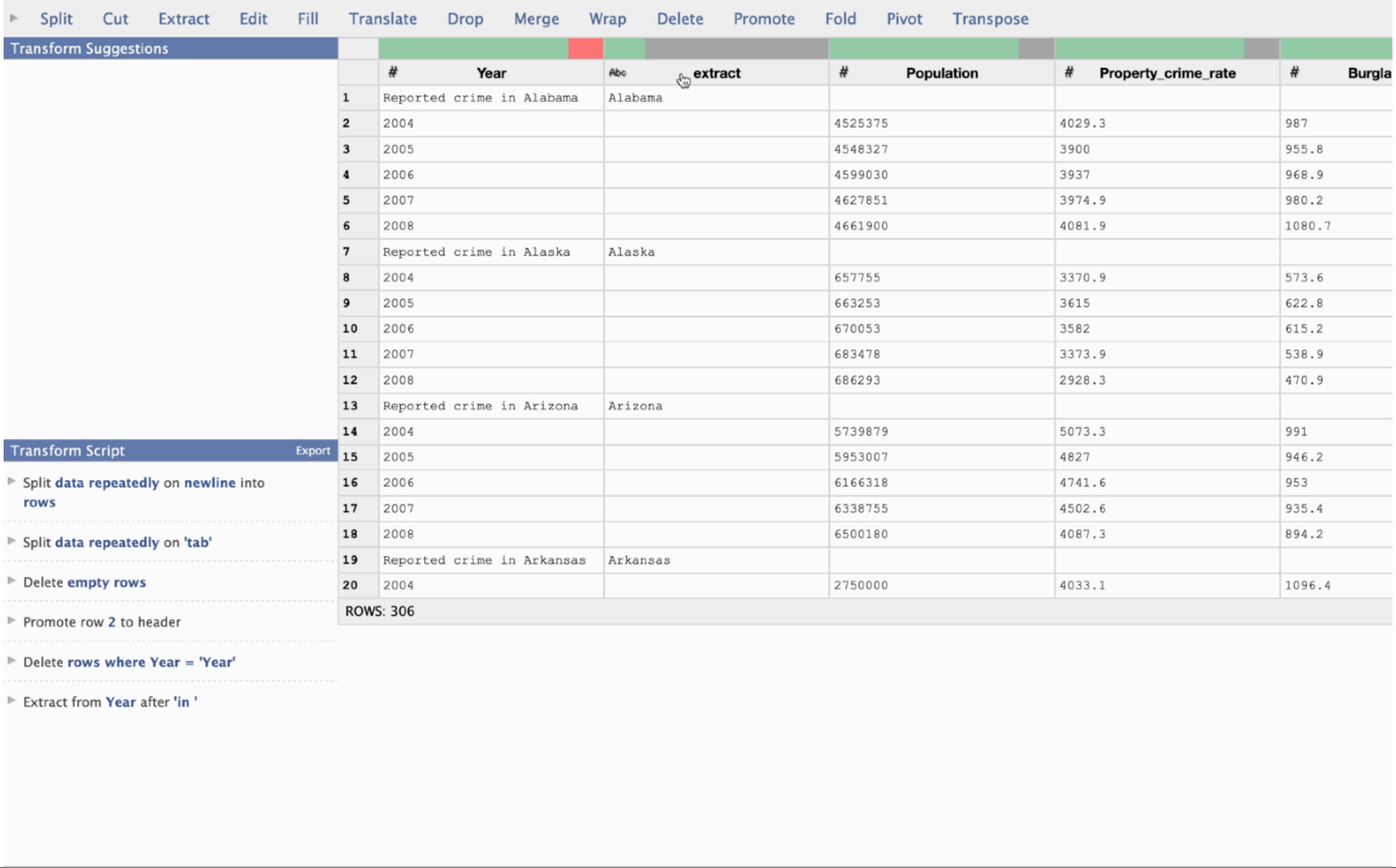
ROWS: 306



- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '

	# Year	Abc 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
15	2005		5953007	4827	946.2
16	2006		6166318	4741.6	953
17	2007		6338755	4502.6	935.4
18	2008		6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004		2750000	4033.1	1096.4

ROWS: 306

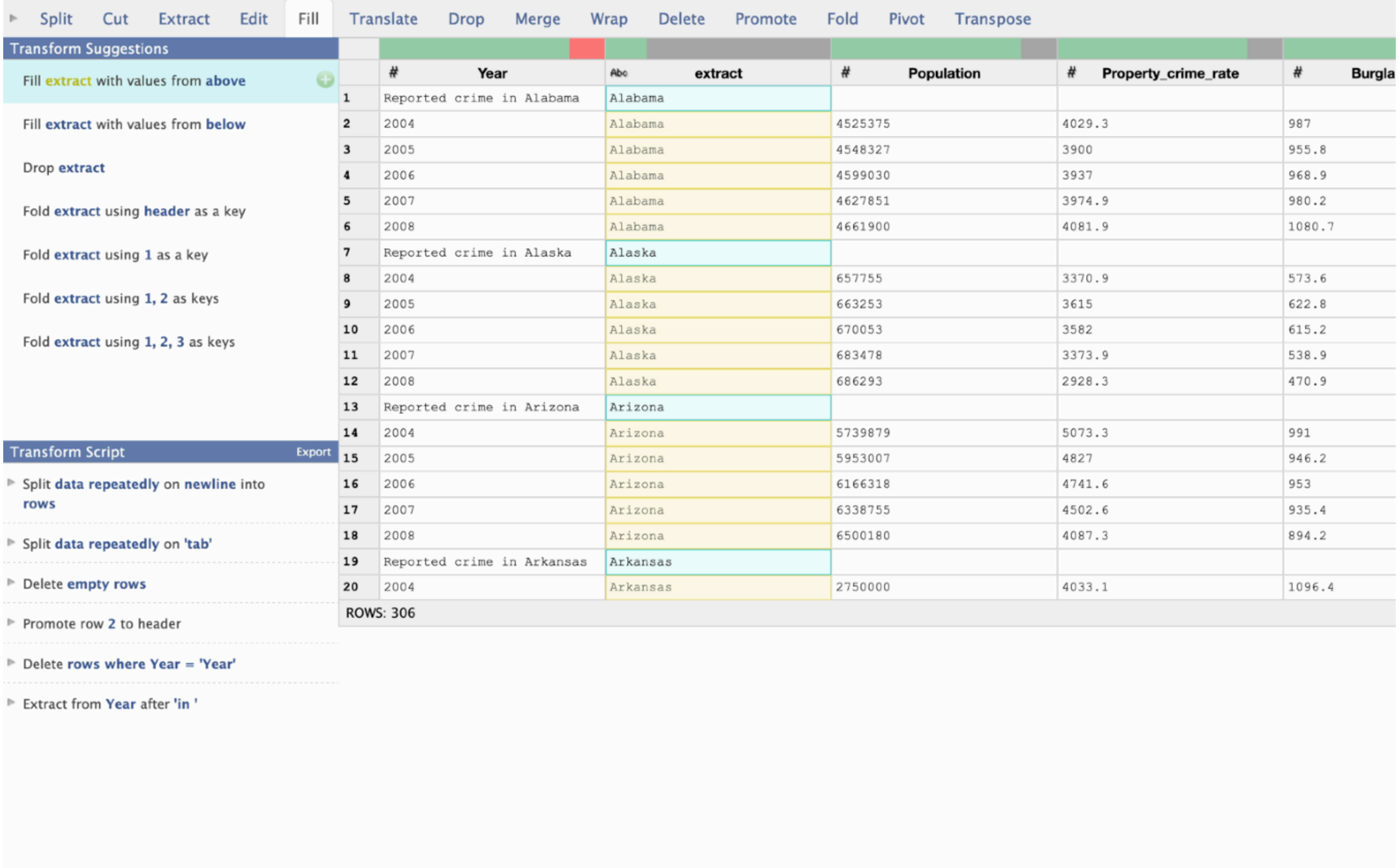


	# Year	Abc extract	# Population	# Property_crime_rate	# Burgla
1	Reported crime in Alabama	Alabama			
2	2004		4525375	4029.3	987
3	2005		4548327	3900	955.8
4	2006		4599030	3937	968.9
5	2007		4627851	3974.9	980.2
6	2008		4661900	4081.9	1080.7
7	Reported crime in Alaska	Alaska			
8	2004		657755	3370.9	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
15	2005		5953007	4827	946.2
16	2006		6166318	4741.6	953
17	2007		6338755	4502.6	935.4
18	2008		6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004		2750000	4033.1	1096.4

ROWS: 306

Transform Script

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '



Transform Suggestions

Fill extract with values from above

Fill extract with values from below

Drop extract

Fold extract using header as a key

Fold extract using 1 as a key

Fold extract using 1, 2 as keys

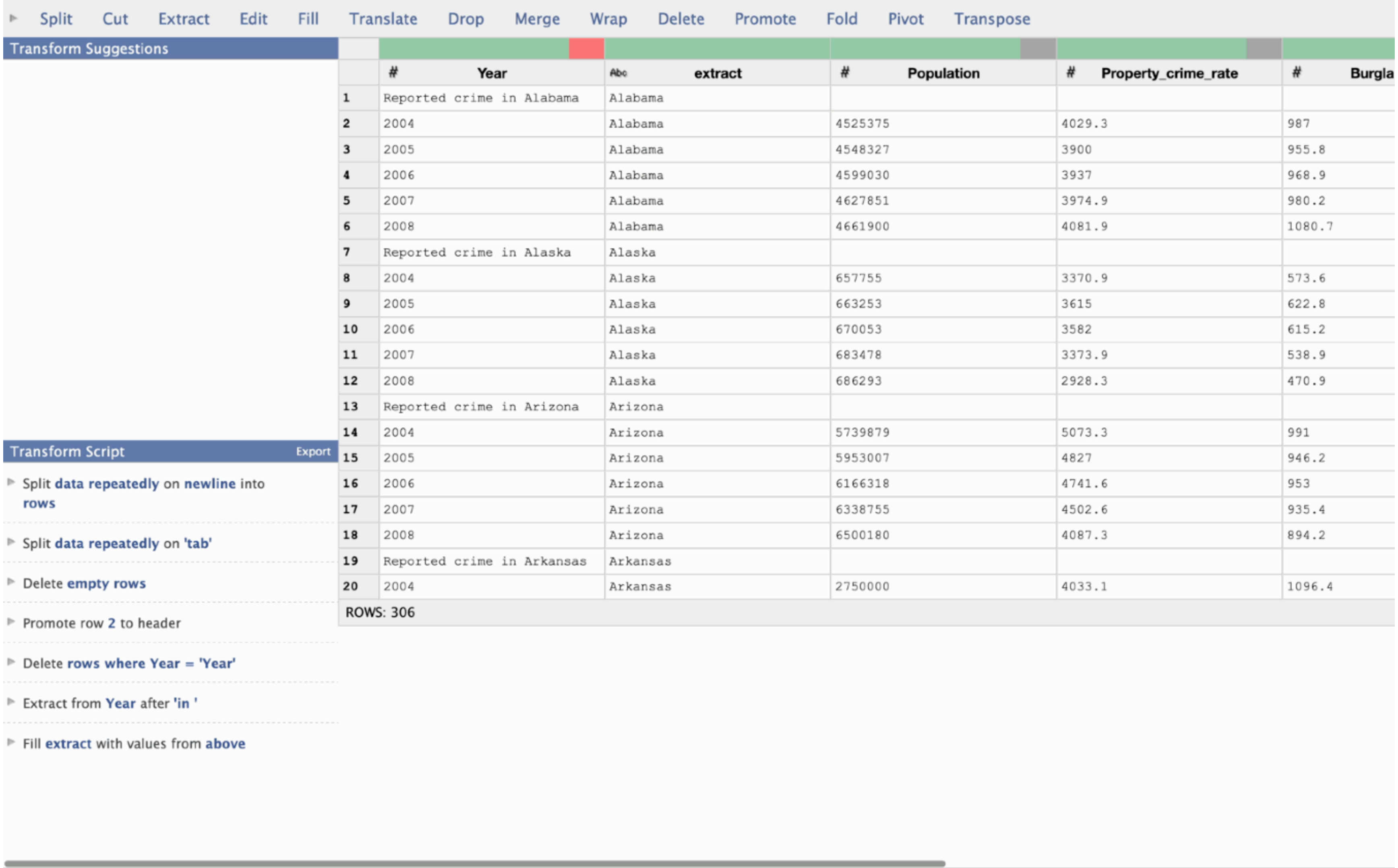
Fold extract using 1, 2, 3 as keys

Transform Script

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '

	# Year	Abc extract	# Population	# Property_crime_rate	# Burgla
1	Reported crime in Alabama	Alabama			
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
15	2005	Arizona	5953007	4827	946.2
16	2006	Arizona	6166318	4741.6	953
17	2007	Arizona	6338755	4502.6	935.4
18	2008	Arizona	6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004	Arkansas	2750000	4033.1	1096.4

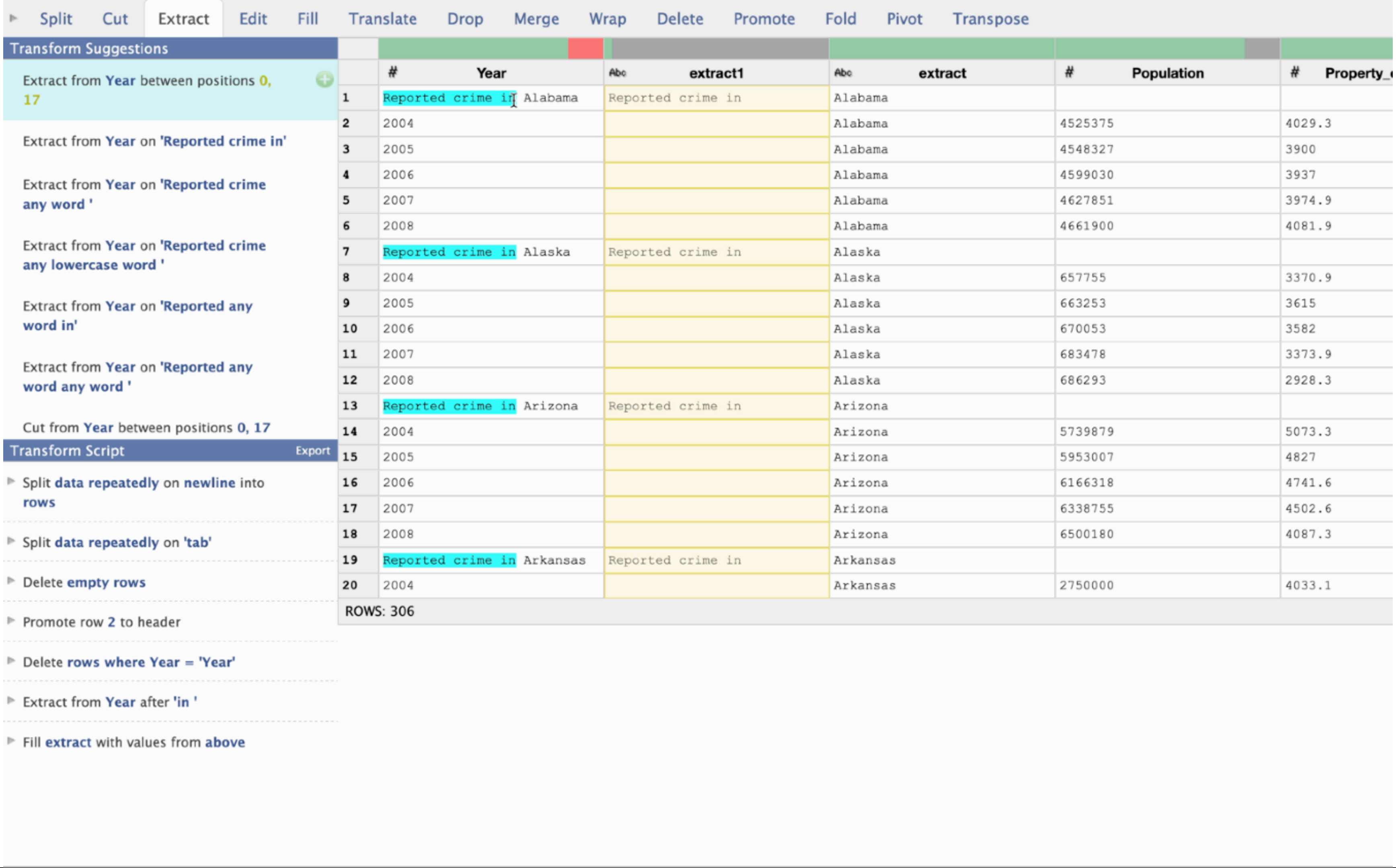
ROWS: 306



	# Year	Abc extract	# Population	# Property_crime_rate	# Burgla
1	Reported crime in Alabama	Alabama			
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
15	2005	Arizona	5953007	4827	946.2
16	2006	Arizona	6166318	4741.6	953
17	2007	Arizona	6338755	4502.6	935.4
18	2008	Arizona	6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004	Arkansas	2750000	4033.1	1096.4

ROWS: 306

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '
- Fill extract with values from above



Transform Suggestions

Extract from **Year** between positions **0, 17**

Extract from **Year** on **'Reported crime in'**

Extract from **Year** on **'Reported crime any word '**

Extract from **Year** on **'Reported crime any lowercase word '**

Extract from **Year** on **'Reported any word in'**

Extract from **Year** on **'Reported any word any word '**

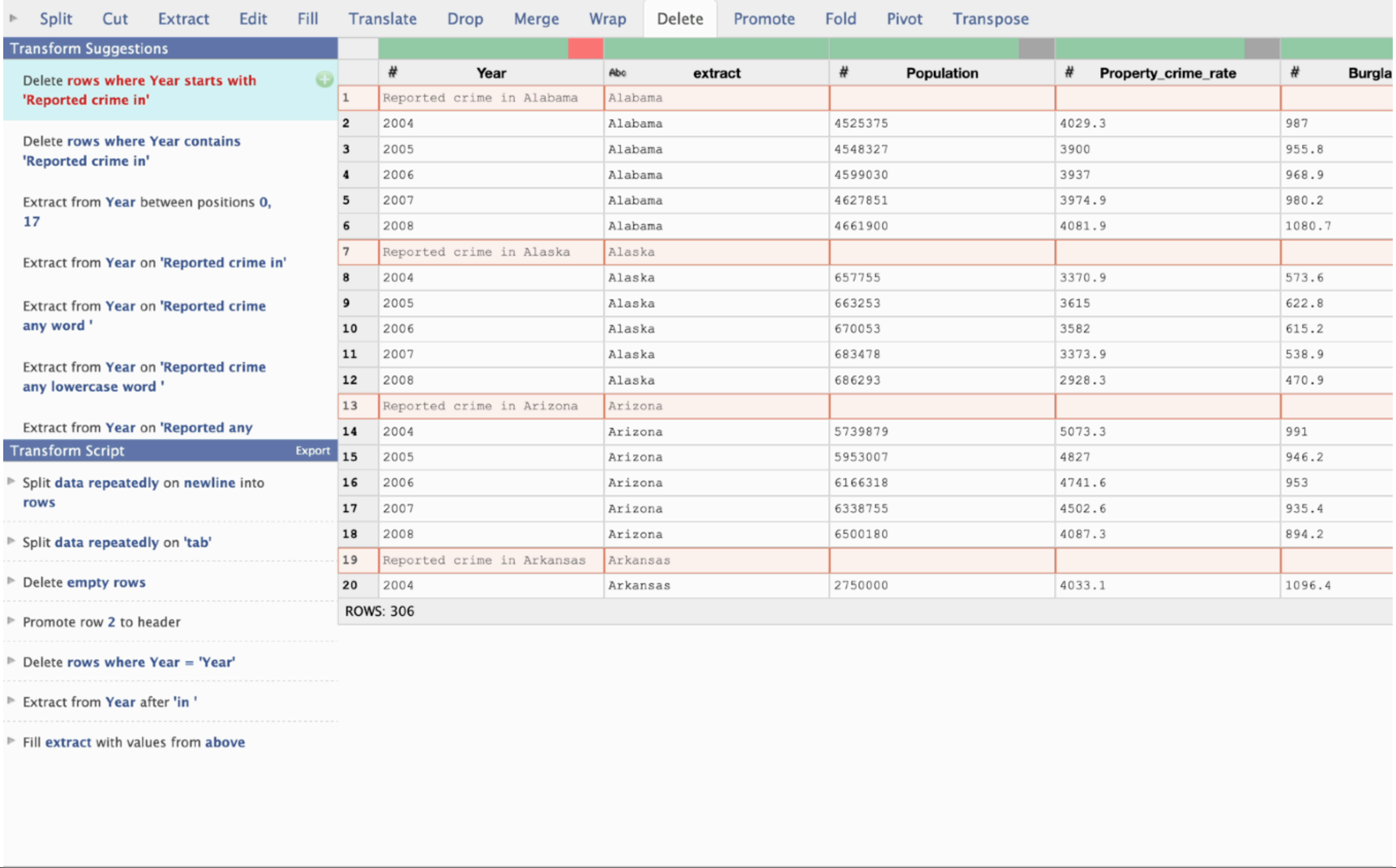
Cut from **Year** between positions **0, 17**

Transform Script

- Split **data repeatedly** on **newline** into **rows**
- Split **data repeatedly** on **'tab'**
- Delete **empty rows**
- Promote row **2** to header
- Delete **rows where Year = 'Year'**
- Extract from **Year** after **'in '**
- Fill **extract** with values from **above**

	# Year	Abc extract1	Abc extract	# Population	# Property_
1	Reported crime in Alabama	Reported crime in	Alabama		
2	2004		Alabama	4525375	4029.3
3	2005		Alabama	4548327	3900
4	2006		Alabama	4599030	3937
5	2007		Alabama	4627851	3974.9
6	2008		Alabama	4661900	4081.9
7	Reported crime in Alaska	Reported crime in	Alaska		
8	2004		Alaska	657755	3370.9
9	2005		Alaska	663253	3615
10	2006		Alaska	670053	3582
11	2007		Alaska	683478	3373.9
12	2008		Alaska	686293	2928.3
13	Reported crime in Arizona	Reported crime in	Arizona		
14	2004		Arizona	5739879	5073.3
15	2005		Arizona	5953007	4827
16	2006		Arizona	6166318	4741.6
17	2007		Arizona	6338755	4502.6
18	2008		Arizona	6500180	4087.3
19	Reported crime in Arkansas	Reported crime in	Arkansas		
20	2004		Arkansas	2750000	4033.1

ROWS: 306



Transform Suggestions

Delete rows where Year starts with 'Reported crime in'

Delete rows where Year contains 'Reported crime in'

Extract from Year between positions 0, 17

Extract from Year on 'Reported crime in'

Extract from Year on 'Reported crime any word '

Extract from Year on 'Reported crime any lowercase word '

Extract from Year on 'Reported any

Transform Script

Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

Delete empty rows

Promote row 2 to header

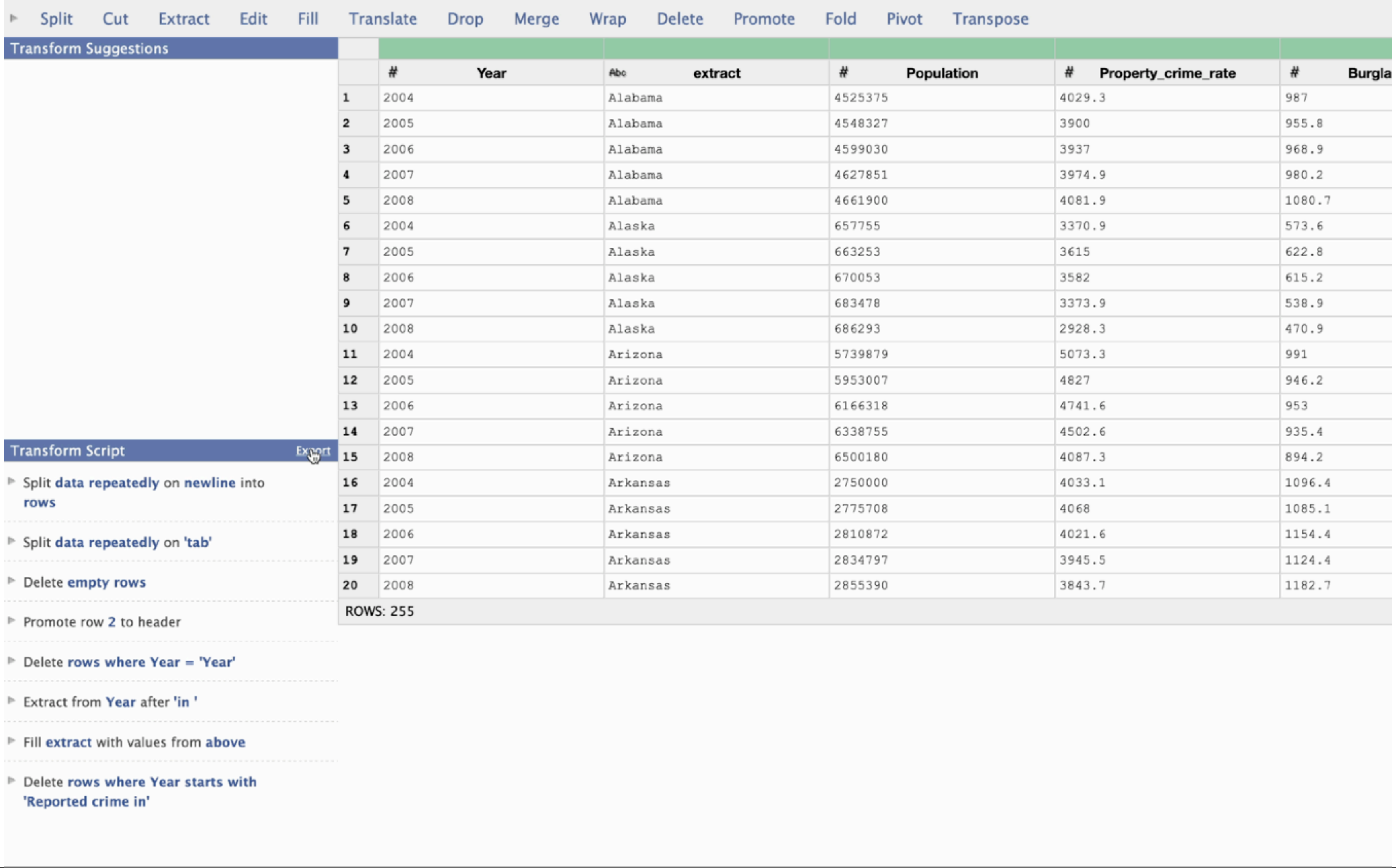
Delete rows where Year = 'Year'

Extract from Year after 'in '

Fill extract with values from above

	# Year	Abc extract	# Population	# Property_crime_rate	# Burgla
1	Reported crime in Alabama	Alabama			
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
15	2005	Arizona	5953007	4827	946.2
16	2006	Arizona	6166318	4741.6	953
17	2007	Arizona	6338755	4502.6	935.4
18	2008	Arizona	6500180	4087.3	894.2
19	Reported crime in Arkansas	Arkansas			
20	2004	Arkansas	2750000	4033.1	1096.4

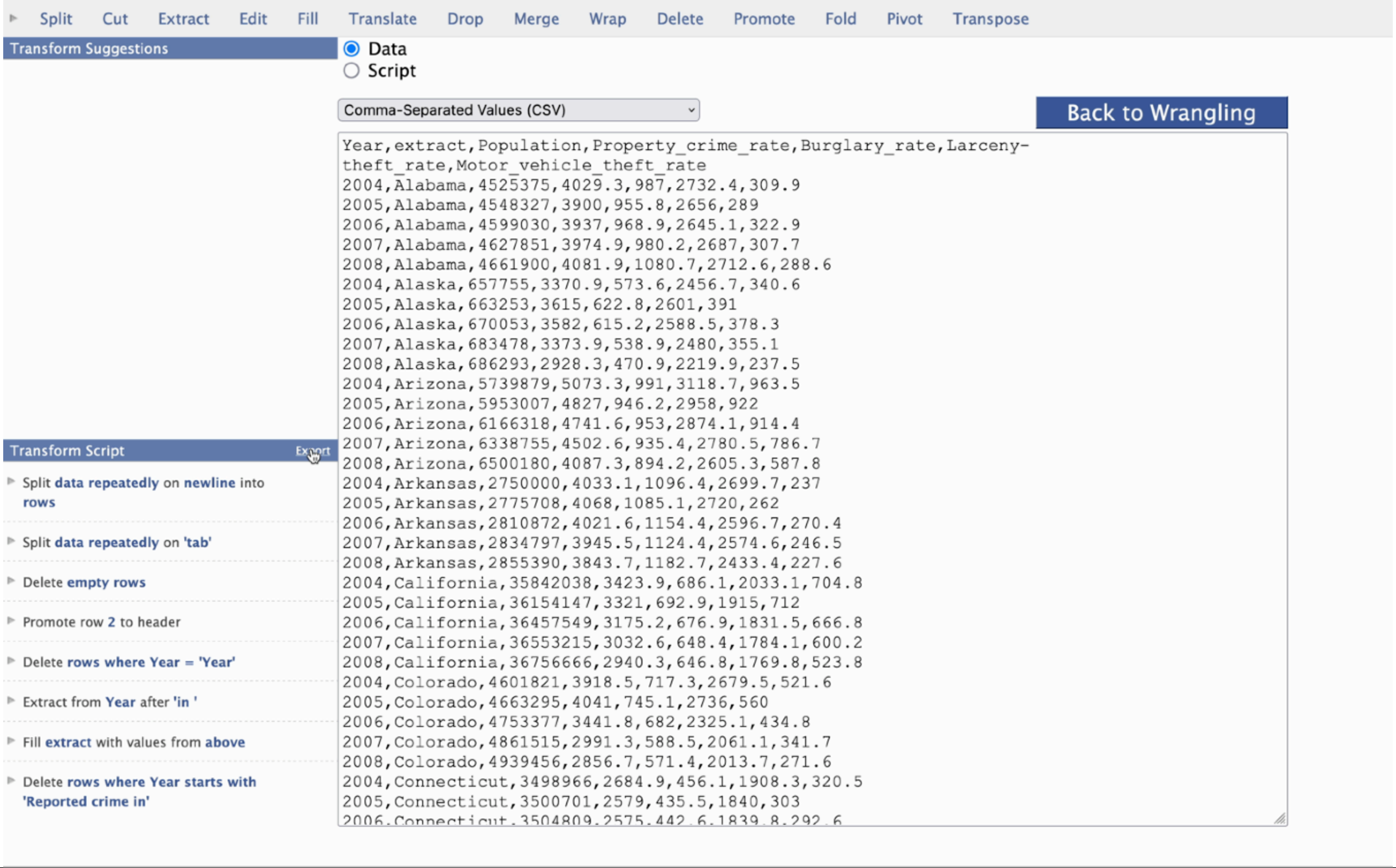
ROWS: 306



	# Year	Abc extract	# Population	# Property_crime_rate	# Burgla
1	2004	Alabama	4525375	4029.3	987
2	2005	Alabama	4548327	3900	955.8
3	2006	Alabama	4599030	3937	968.9
4	2007	Alabama	4627851	3974.9	980.2
5	2008	Alabama	4661900	4081.9	1080.7
6	2004	Alaska	657755	3370.9	573.6
7	2005	Alaska	663253	3615	622.8
8	2006	Alaska	670053	3582	615.2
9	2007	Alaska	683478	3373.9	538.9
10	2008	Alaska	686293	2928.3	470.9
11	2004	Arizona	5739879	5073.3	991
12	2005	Arizona	5953007	4827	946.2
13	2006	Arizona	6166318	4741.6	953
14	2007	Arizona	6338755	4502.6	935.4
15	2008	Arizona	6500180	4087.3	894.2
16	2004	Arkansas	2750000	4033.1	1096.4
17	2005	Arkansas	2775708	4068	1085.1
18	2006	Arkansas	2810872	4021.6	1154.4
19	2007	Arkansas	2834797	3945.5	1124.4
20	2008	Arkansas	2855390	3843.7	1182.7

ROWS: 255

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- 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
Transform Script
Export

```
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
2008,Alabama,4661900,4081.9,1080.7,2712.6,288.6
2004,Alaska,657755,3370.9,573.6,2456.7,340.6
2005,Alaska,663253,3615,622.8,2601,391
2006,Alaska,670053,3582,615.2,2588.5,378.3
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
2007,Arizona,6338755,4502.6,935.4,2780.5,786.7
2008,Arizona,6500180,4087.3,894.2,2605.3,587.8
2004,Arkansas,2750000,4033.1,1096.4,2699.7,237
2005,Arkansas,2775708,4068,1085.1,2720,262
2006,Arkansas,2810872,4021.6,1154.4,2596.7,270.4
2007,Arkansas,2834797,3945.5,1124.4,2574.6,246.5
2008,Arkansas,2855390,3843.7,1182.7,2433.4,227.6
2004,California,35842038,3423.9,686.1,2033.1,704.8
2005,California,36154147,3321,692.9,1915,712
2006,California,36457549,3175.2,676.9,1831.5,666.8
2007,California,36553215,3032.6,648.4,1784.1,600.2
2008,California,36756666,2940.3,646.8,1769.8,523.8
2004,Colorado,4601821,3918.5,717.3,2679.5,521.6
2005,Colorado,4663295,4041,745.1,2736,560
2006,Colorado,4753377,3441.8,682,2325.1,434.8
2007,Colorado,4861515,2991.3,588.5,2061.1,341.7
2008,Colorado,4939456,2856.7,571.4,2013.7,271.6
2004,Connecticut,3498966,2684.9,456.1,1908.3,320.5
2005,Connecticut,3500701,2579,435.5,1840,303
2006,Connecticut,3504809,2575,442.6,1839.8,292.6
```

▸ Split data repeatedly on newline into rows

▸ Split data repeatedly on 'tab'

▸ Delete empty rows

▸ Promote row 2 to header

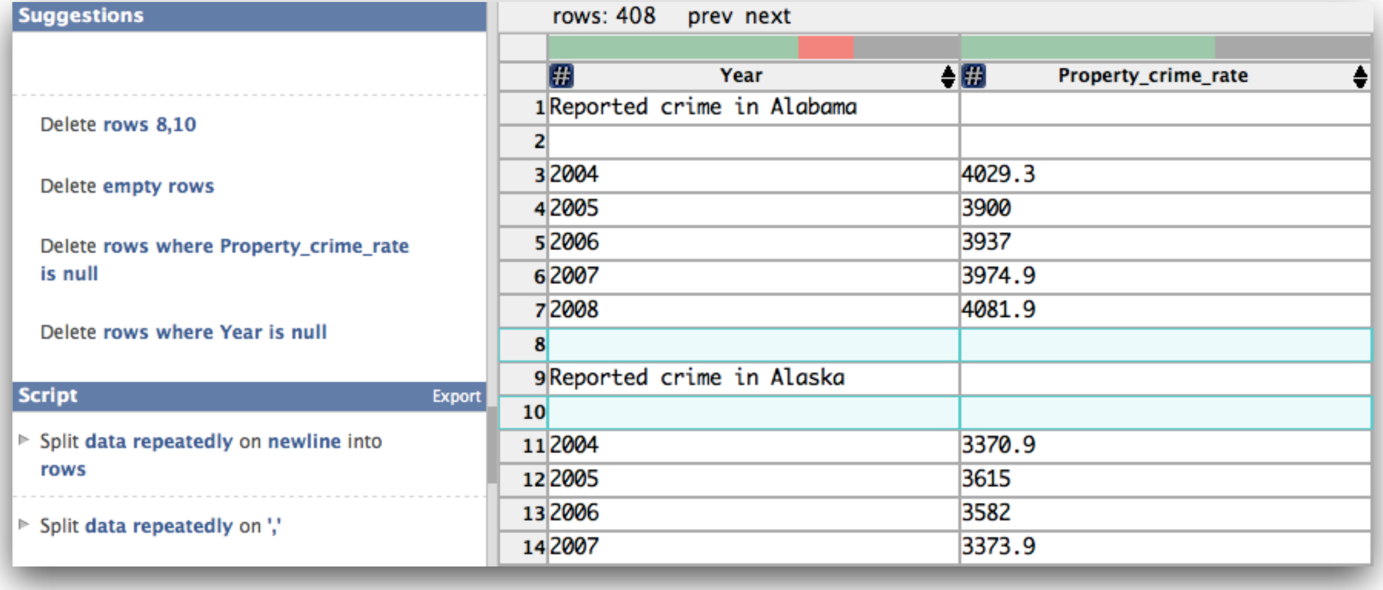
▸ Delete rows where Year = 'Year'

▸ Extract from Year after 'in '

▸ Fill extract with values from above

▸ Delete rows where Year starts with 'Reported crime in'

DataWrangler



Suggestions

Delete rows 8,10

Delete empty rows

Delete rows where Property_crime_rate is null

Delete rows where Year is null

Script Export

▸ Split data repeatedly on newline into rows

▸ Split data repeatedly on ','

rows: 408 prev next

	# Year	# Property_crime_rate
1	Reported crime in Alabama	
2		
3	2004	4029.3
4	2005	3900
5	2006	3937
6	2007	3974.9
7	2008	4081.9
8		
9	Reported crime in Alaska	
10		
11	2004	3370.9
12	2005	3615
13	2006	3582
14	2007	3373.9

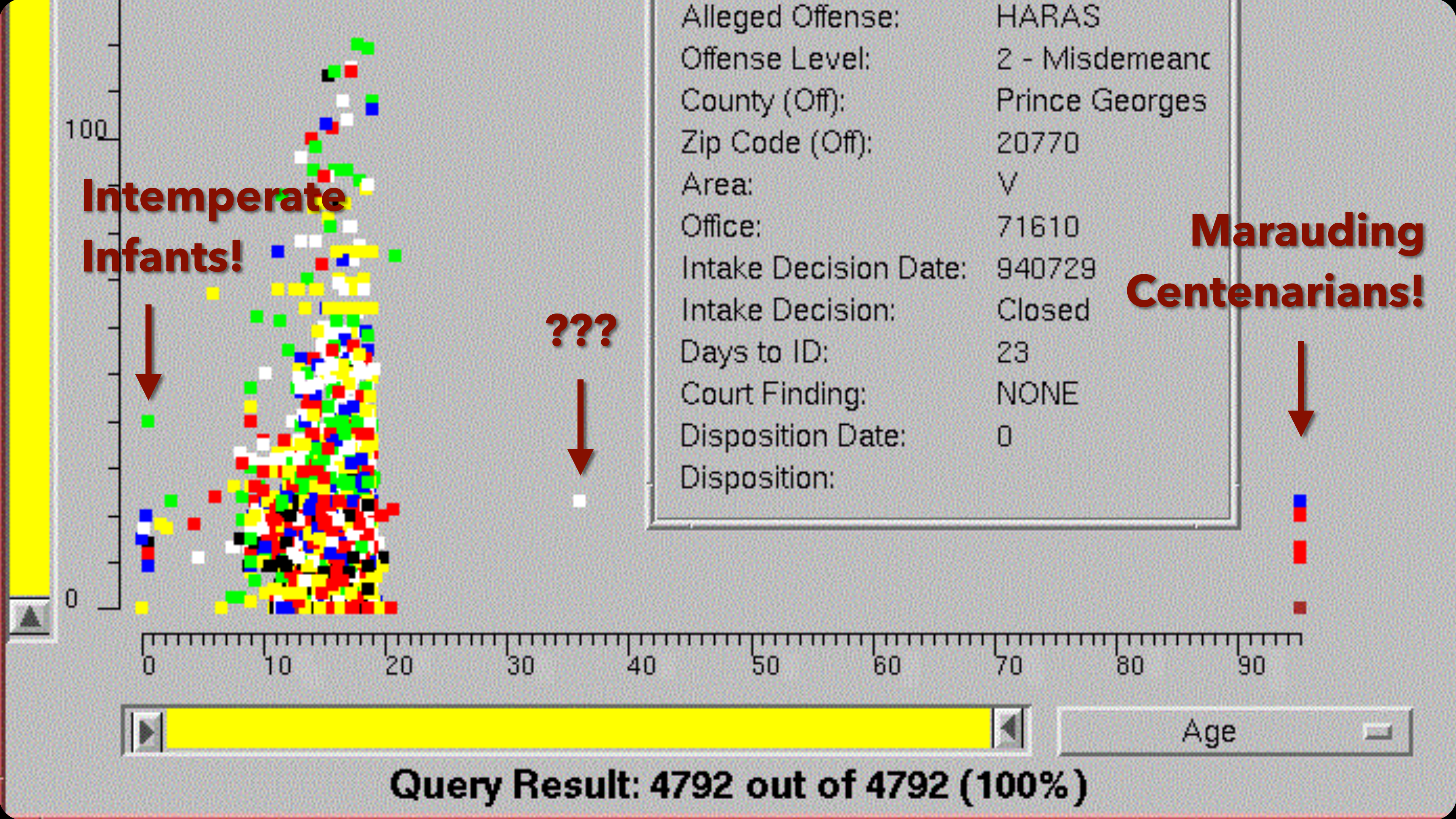
Wrangler: Interactive Visual Specification of Data Transformation Scripts

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]



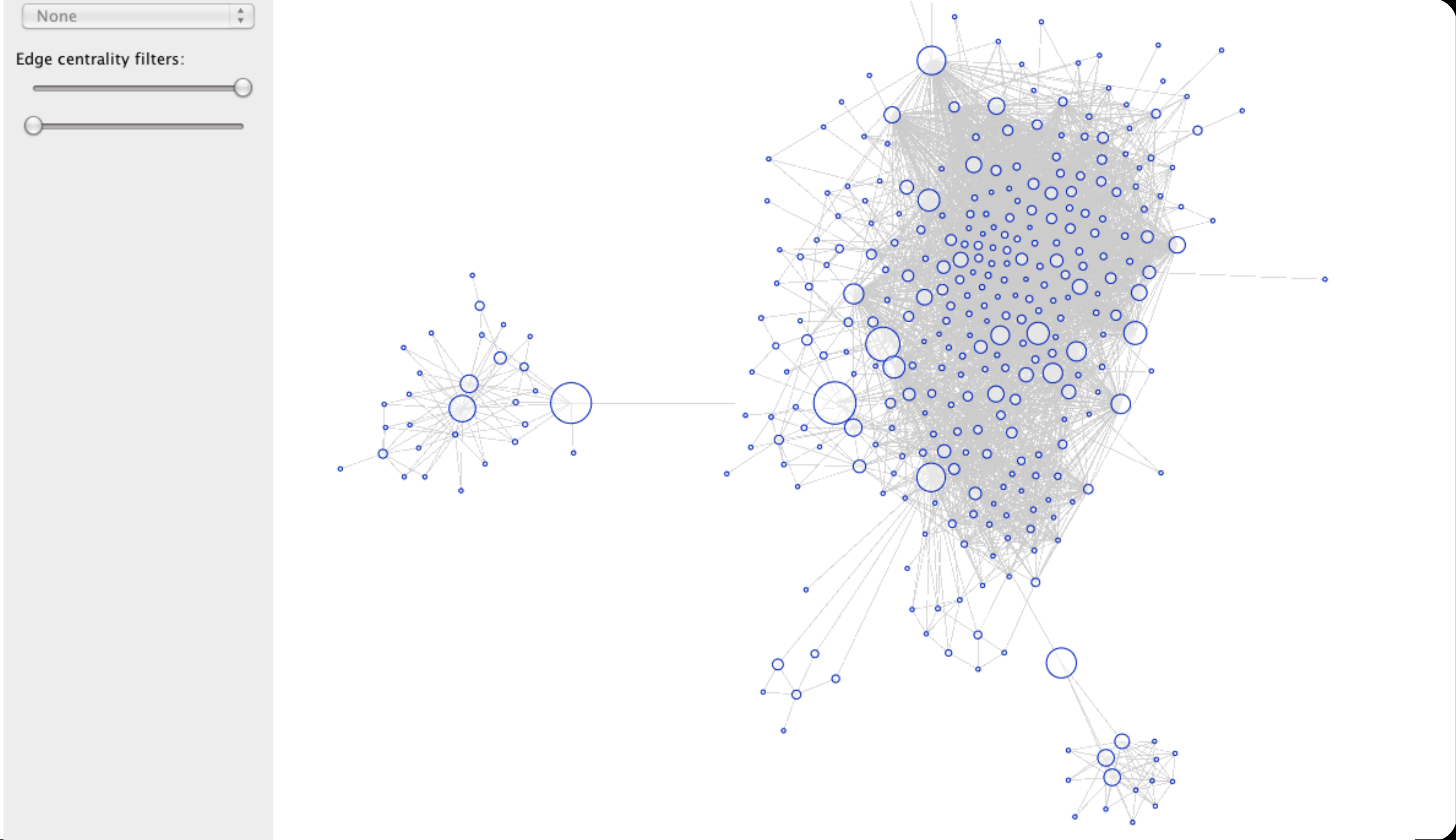
Alleged Offense:	HARAS
Offense Level:	2 - Misdemeanc
County (Off):	Prince Georges
Zip Code (Off):	20770
Area:	V
Office:	71610
Intake Decision Date:	940729
Intake Decision:	Closed
Days to ID:	23
Court Finding:	NONE
Disposition Date:	0
Disposition:	

Query Result: 4792 out of 4792 (100%)

Intemperate Infants!

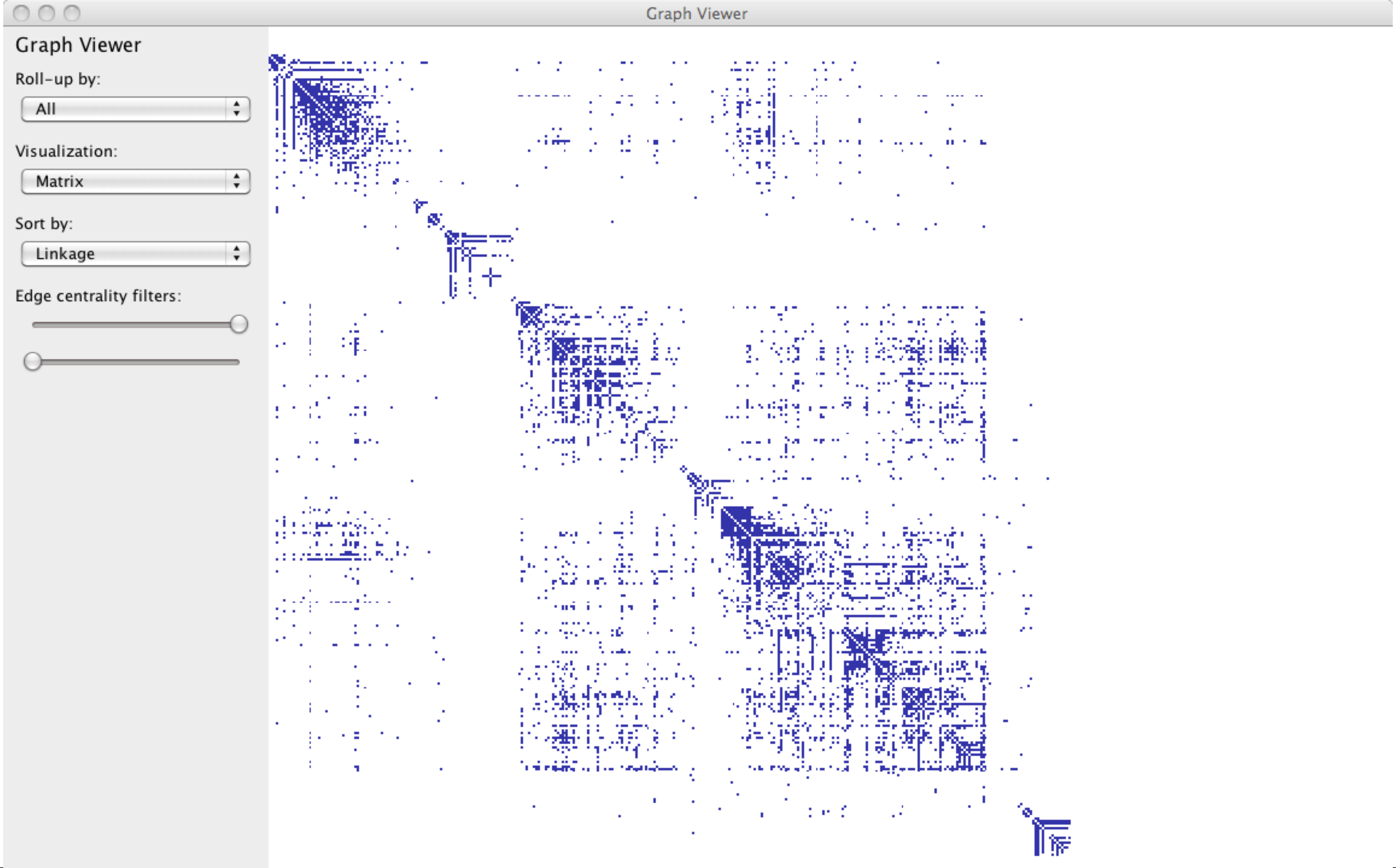
???

Marauding Centenarians!



None

Edge centrality filters:



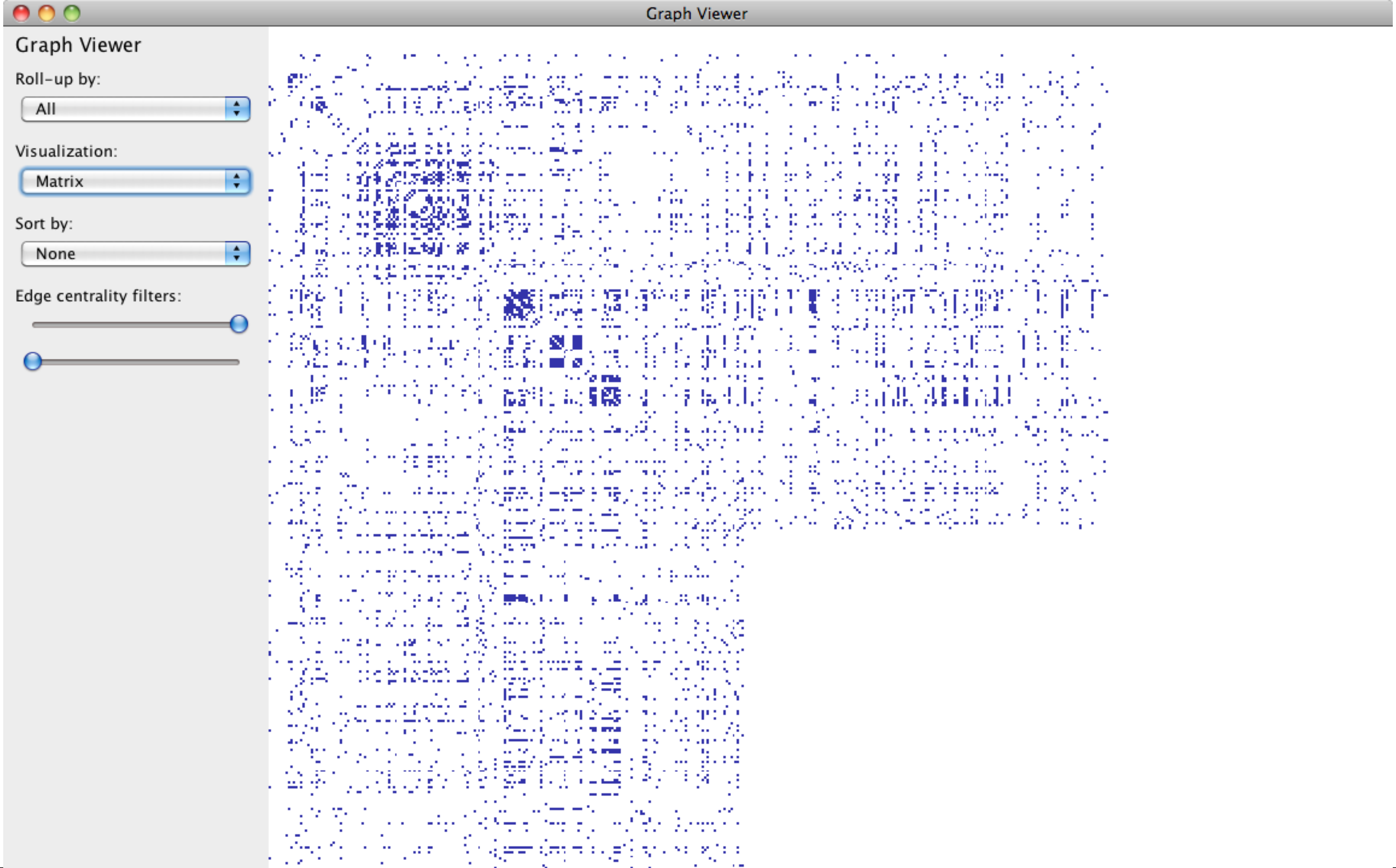
Graph Viewer

Roll-up by: All

Visualization: Matrix

Sort by: Linkage

Edge centrality filters:



Graph Viewer

Graph Viewer

Roll-up by:

All

Visualization:

Matrix

Sort by:

None

Edge centrality filters:

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	no measurements, redacted, …?
Erroneous Values	misspelling, outliers, …?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

Anticipate problems with your data!

Data Wrangling Tools

Libraries

JavaScript: Arquero

Python: Pandas, Polars

R: dplyr

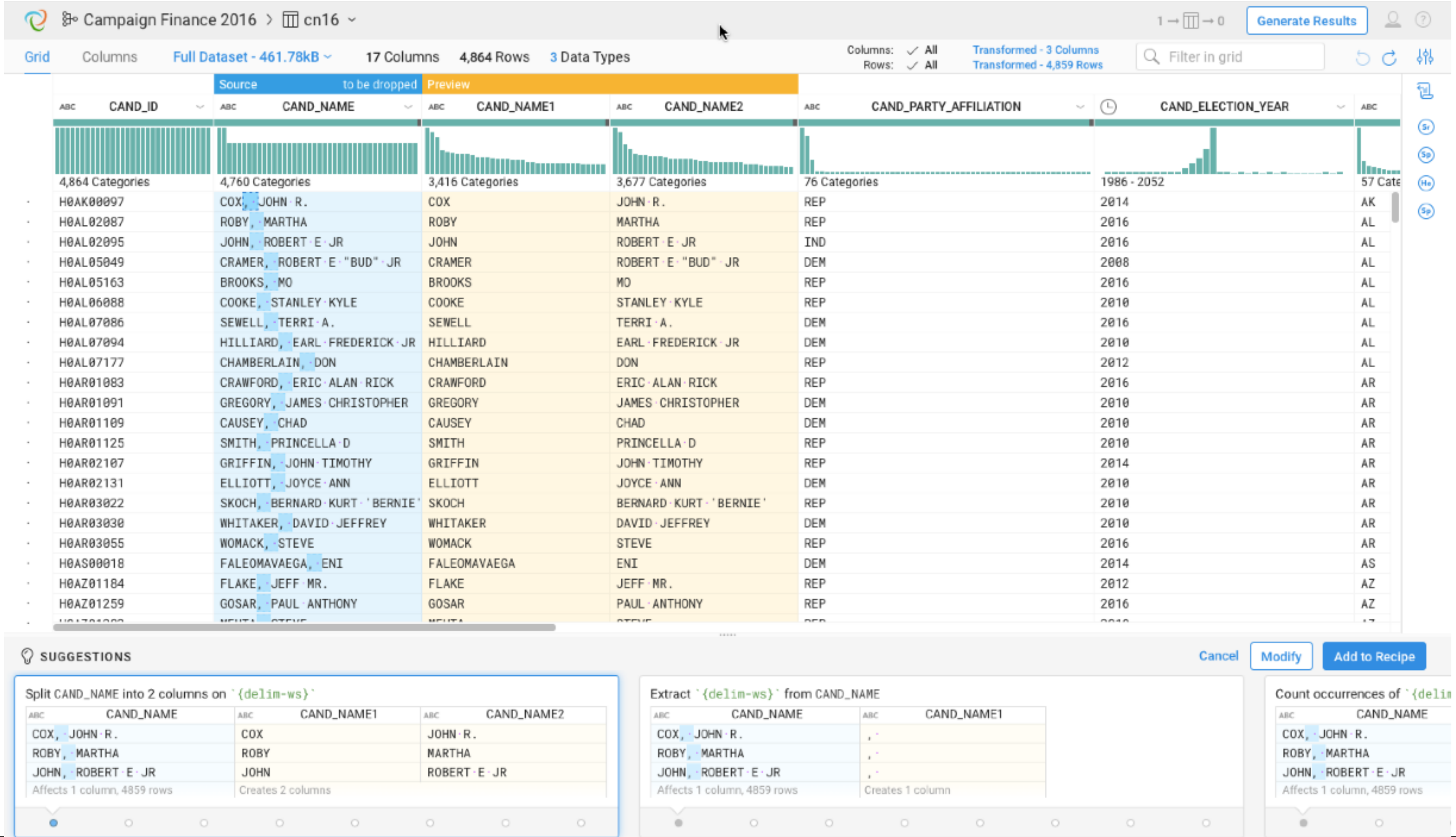
Databases

DuckDB + SQL queries

Graphical Tools

We'll look at some of these next!

Trifacta Wrangler (now part of Alteryx)



	Source to be dropped	Preview				
CAND_ID	CAND_NAME	CAND_NAME1	CAND_NAME2	CAND_PARTY_AFFILIATION	CAND_ELECTION_YEAR	
4,864 Categories	4,760 Categories	3,416 Categories	3,677 Categories	76 Categories	1986 - 2052	57 Cate
H0AK00097	COX, JOHN R.	COX	JOHN R.	REP	2014	AK
H0AL02087	ROBY, MARTHA	ROBY	MARTHA	REP	2016	AL
H0AL02095	JOHN, ROBERT E JR	JOHN	ROBERT E JR	IND	2016	AL
H0AL05049	CRAMER, ROBERT E "BUD" JR	CRAMER	ROBERT E "BUD" JR	DEM	2008	AL
H0AL05163	BROOKS, MO	BROOKS	MO	REP	2016	AL
H0AL06088	COOKE, STANLEY KYLE	COOKE	STANLEY KYLE	REP	2010	AL
H0AL07086	SEWELL, TERRI A.	SEWELL	TERRI A.	DEM	2016	AL
H0AL07094	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 D	REP	2010	AR
H0AR02107	GRIFFIN, JOHN TIMOTHY	GRIFFIN	JOHN TIMOTHY	REP	2014	AR
H0AR02131	ELLIOTT, JOYCE ANN	ELLIOTT	JOYCE ANN	DEM	2010	AR
H0AR03022	SKOCH, BERNARD KURT 'BERNIE'	SKOCH	BERNARD KURT 'BERNIE'	REP	2010	AR
H0AR03030	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

SUGGESTIONS

Split CAND_NAME into 2 columns on `{delim-ws}`

CAND_NAME	CAND_NAME1	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	

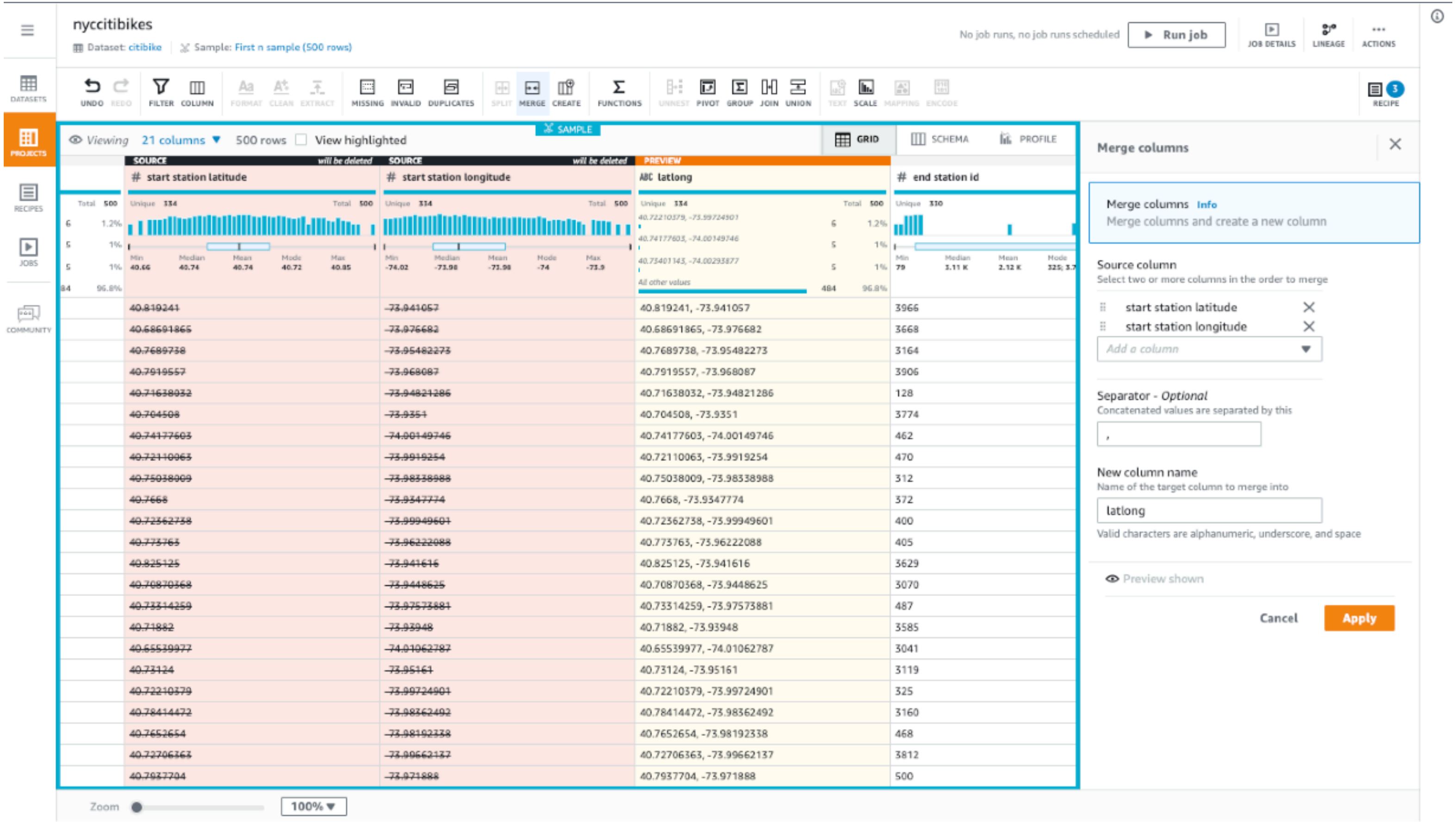
Extract `{delim-ws}` from CAND_NAME

CAND_NAME	CAND_NAME1
COX, JOHN R.	,
ROBY, MARTHA	,
JOHN, ROBERT E JR	,
Affects 1 column, 4859 rows	Creates 1 column

Count occurrences of `{delim

CAND_NAME
COX, JOHN R.
ROBY, MARTHA
JOHN, ROBERT E JR
Affects 1 column, 4859 rows

AWS Glue DataBrew



nyccitibikes

Dataset: citibike | Sample: First n sample (500 rows)

Viewing 21 columns ▼ 500 rows □ View highlighted

SOURCE *will be deleted* # start station latitude	SOURCE *will be deleted* # start station longitude	PREVIEW ABC latlong	# end station id
Unique 334 Total 500 Min 40.66 Median 40.74 Mean 40.74 Mode 40.72 Max 40.85	Unique 334 Total 500 Min -74.02 Median -73.98 Mean -73.98 Mode -74 Max -73.9	Unique 334 Total 500 *40.72210379, -73.99724901* 6 1.2% *40.74177603, -74.00149746* 5 1% *40.73401143, -74.00293877* 5 1% *All other values* 484 96.8%	Unique 330 Min 79 Median 3.11 K Mean 2.12 K Mode 325; 3.7
~~40.819241~~	~~-73.941057~~	40.819241, -73.941057	3966
~~40.68691865~~	~~-73.976682~~	40.68691865, -73.976682	3668
~~40.7689738~~	~~-73.95482273~~	40.7689738, -73.95482273	3164
~~40.7919557~~	~~-73.968087~~	40.7919557, -73.968087	3906
~~40.71638032~~	~~-73.94821286~~	40.71638032, -73.94821286	128
~~40.704508~~	~~-73.9351~~	40.704508, -73.9351	3774
~~40.74177603~~	~~-74.00149746~~	40.74177603, -74.00149746	462
~~40.72110063~~	~~-73.9919254~~	40.72110063, -73.9919254	470
~~40.75038009~~	~~-73.98338988~~	40.75038009, -73.98338988	312
~~40.7668~~	~~-73.9347774~~	40.7668, -73.9347774	372
~~40.72362738~~	~~-73.99949601~~	40.72362738, -73.99949601	400
~~40.773763~~	~~-73.96222088~~	40.773763, -73.96222088	405
~~40.825125~~	~~-73.941616~~	40.825125, -73.941616	3629
~~40.70870368~~	~~-73.9448625~~	40.70870368, -73.9448625	3070
~~40.73314259~~	~~-73.97573881~~	40.73314259, -73.97573881	487
~~40.71882~~	~~-73.93948~~	40.71882, -73.93948	3585
~~40.65539977~~	~~-74.01062787~~	40.65539977, -74.01062787	3041
~~40.73124~~	~~-73.95161~~	40.73124, -73.95161	3119
~~40.72210379~~	~~-73.99724901~~	40.72210379, -73.99724901	325
~~40.78414472~~	~~-73.98362492~~	40.78414472, -73.98362492	3160
~~40.7652654~~	~~-73.98192338~~	40.7652654, -73.98192338	468
~~40.72706363~~	~~-73.99662137~~	40.72706363, -73.99662137	3812
~~40.7937704~~	~~-73.971888~~	40.7937704, -73.971888	500

Merge columns ✕

Merge columns Info
Merge columns and create a new column

Source column
Select two or more columns in the order to merge

- start station latitude ✕
- start station longitude ✕

Add a column ▼

Separator - *Optional*
Concatenated values are separated by this

,

New column name
Name of the target column to merge into

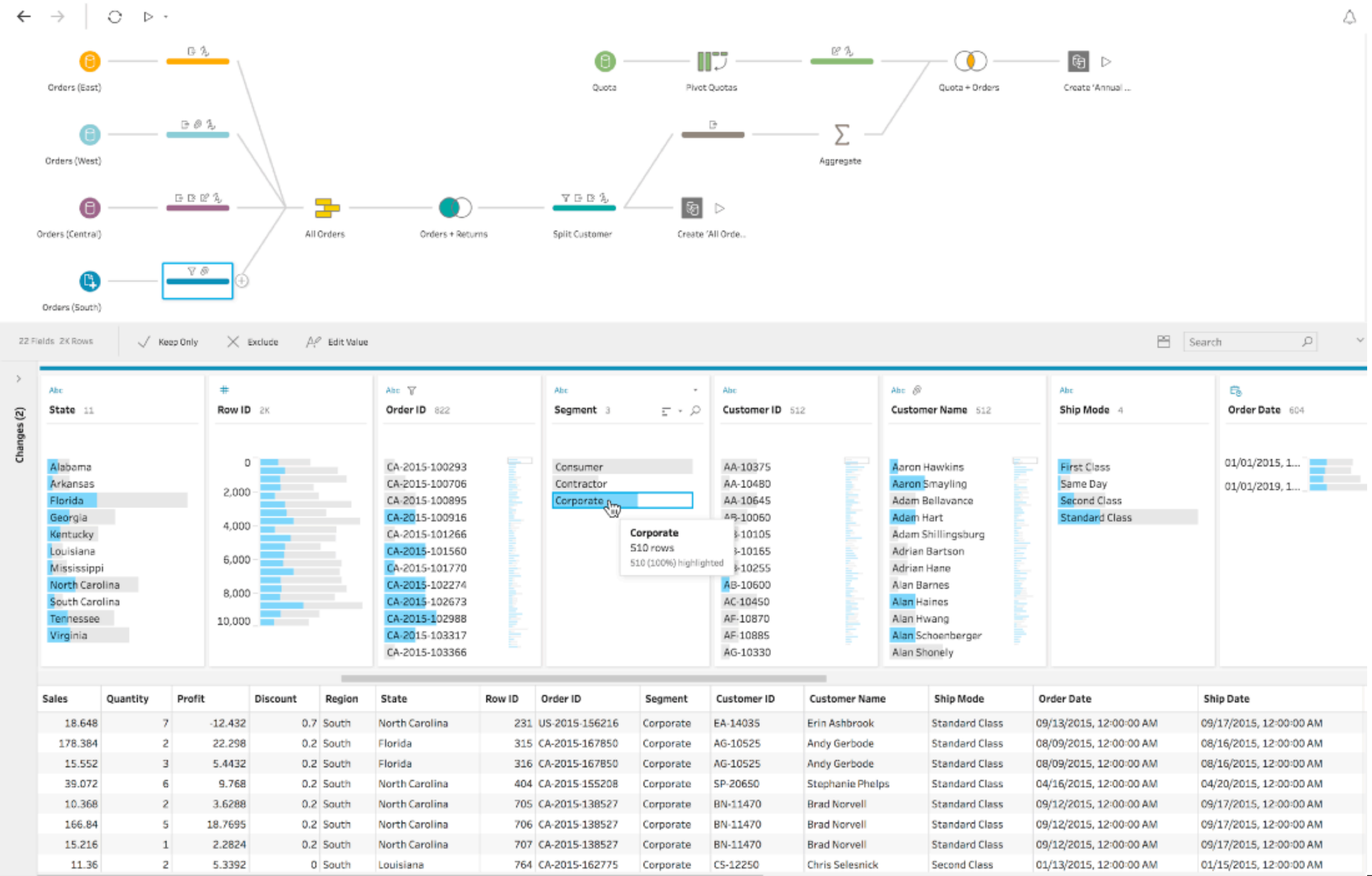
latlong

Valid characters are alphanumeric, underscore, and space

Preview shown

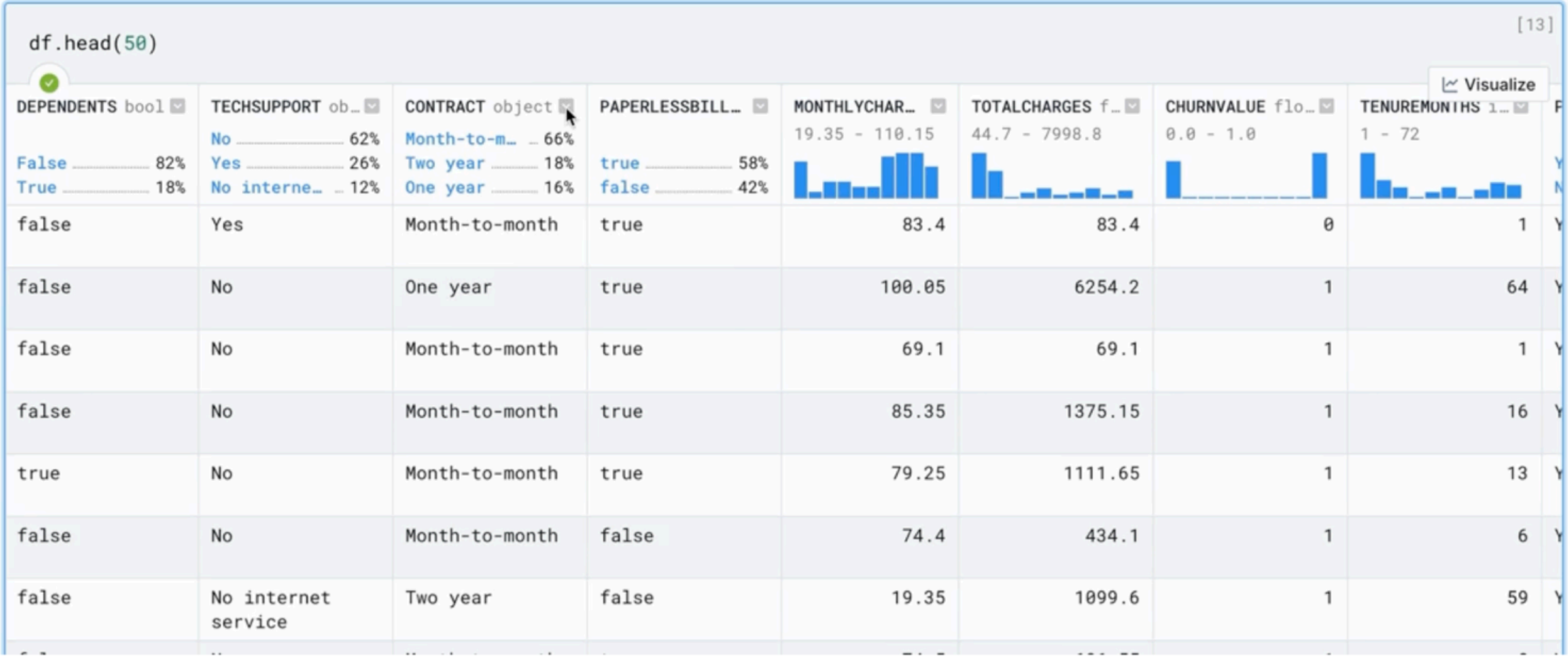
Cancel Apply

Tableau Prep



Sales	Quantity	Profit	Discount	Region	State	Row ID	Order ID	Segment	Customer ID	Customer Name	Ship Mode	Order Date	Ship Date
18.648	7	-12.432	0.7	South	North Carolina	231	US-2015-156216	Corporate	EA-14035	Erin Ashbrook	Standard Class	09/13/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
178.384	2	22.298	0.2	South	Florida	315	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
15.552	3	5.4432	0.2	South	Florida	316	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
39.072	6	9.768	0.2	South	North Carolina	404	CA-2015-155208	Corporate	SP-20650	Stephanie Phelps	Standard Class	04/16/2015, 12:00:00 AM	04/20/2015, 12:00:00 AM
10.368	2	3.6288	0.2	South	North Carolina	705	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
166.84	5	18.7695	0.2	South	North Carolina	706	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
15.216	1	2.2824	0.2	South	North Carolina	707	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
11.36	2	5.3392	0	South	Louisiana	764	CA-2015-162775	Corporate	CS-12250	Chris Selesnick	Second Class	01/13/2015, 12:00:00 AM	01/15/2015, 12:00:00 AM

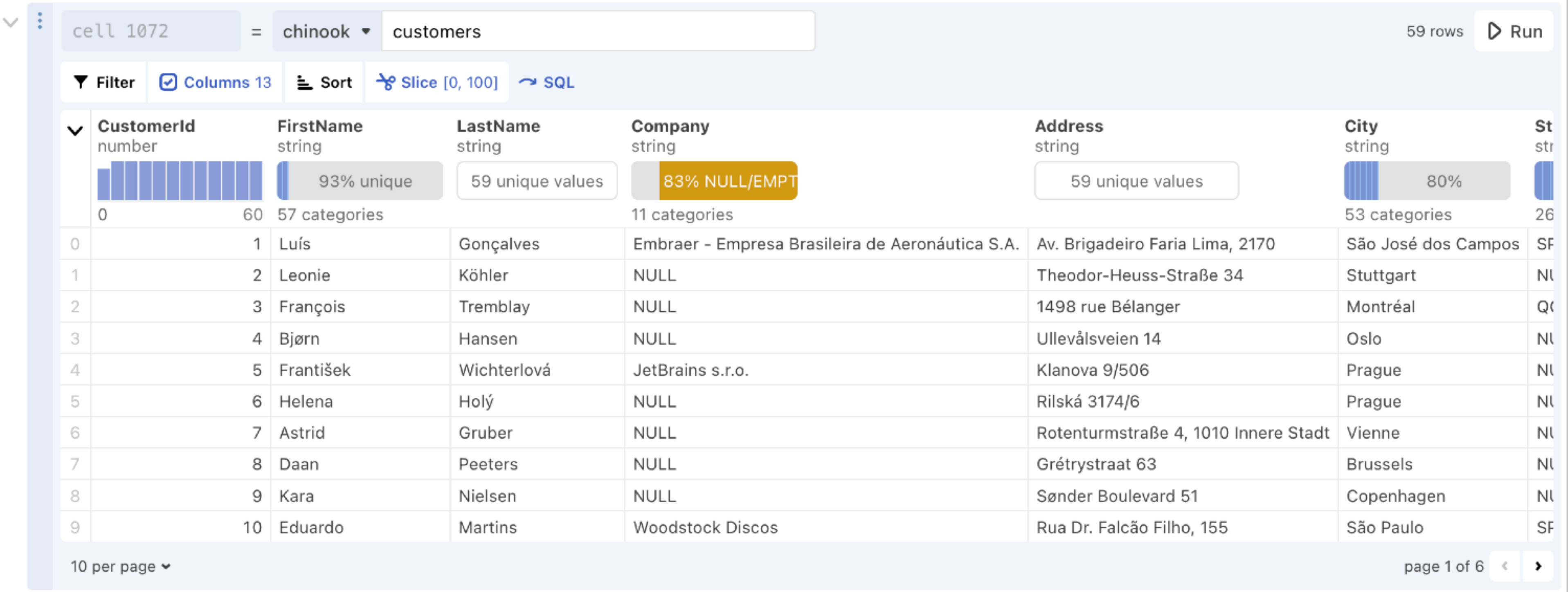
Deepnote



```
df.head(50)
```

DEPENDENTS bool	TECHSUPPORT ob…	CONTRACT object	PAPERLESSBILL…	MONTHLYCHAR…	TOTALCHARGES f…	CHURNVALUE flo…	TENUREMONTHS i…	P
False 82% True 18%	No 62% Yes 26% No interne… 12%	Month-to-m… 66% Two year 18% One year 16%	true 58% false 42%	19.35 - 110.15	44.7 - 7998.8	0.0 - 1.0	1 - 72	Y N
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



cell 1072 = chinook ▾ customers

59 rows ▷ Run

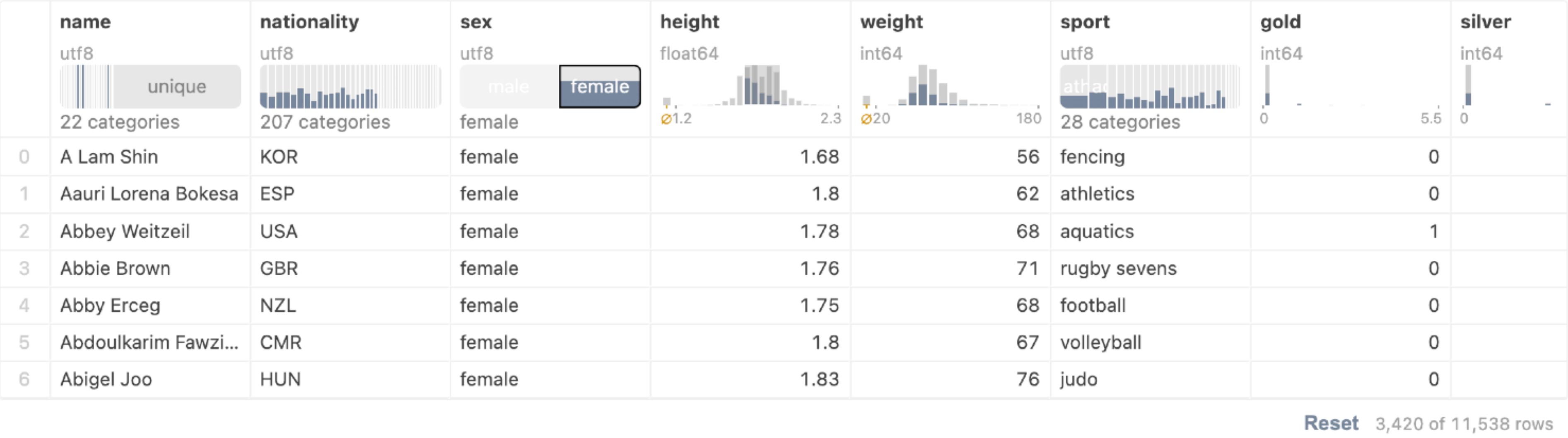
Filter ☑ Columns 13 Sort Slice [0, 100] SQL

	CustomerId	FirstName	LastName	Company	Address	City	St
	number	string	string	string	string	string	str
		93% unique	59 unique values	83% NULL/EMPT	59 unique values	80%	
	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	SP
1	2	Leonie	Köhler	NULL	Theodor-Heuss-Straße 34	Stuttgart	NU
2	3	François	Tremblay	NULL	1498 rue Bélanger	Montréal	QC
3	4	Bjørn	Hansen	NULL	Ullevålsveien 14	Oslo	NU
4	5	František	Wichterlová	JetBrains s.r.o.	Klanova 9/506	Prague	NU
5	6	Helena	Holý	NULL	Rilská 3174/6	Prague	NU
6	7	Astrid	Gruber	NULL	Rotenturmstraße 4, 1010 Innere Stadt	Vienne	NU
7	8	Daan	Peeters	NULL	Grétrystraat 63	Brussels	NU
8	9	Kara	Nielsen	NULL	Sønder Boulevard 51	Copenhagen	NU
9	10	Eduardo	Martins	Woodstock Discos	Rua Dr. Falcão Filho, 155	São Paulo	SP

10 per page ▾

page 1 of 6 ‹ ›

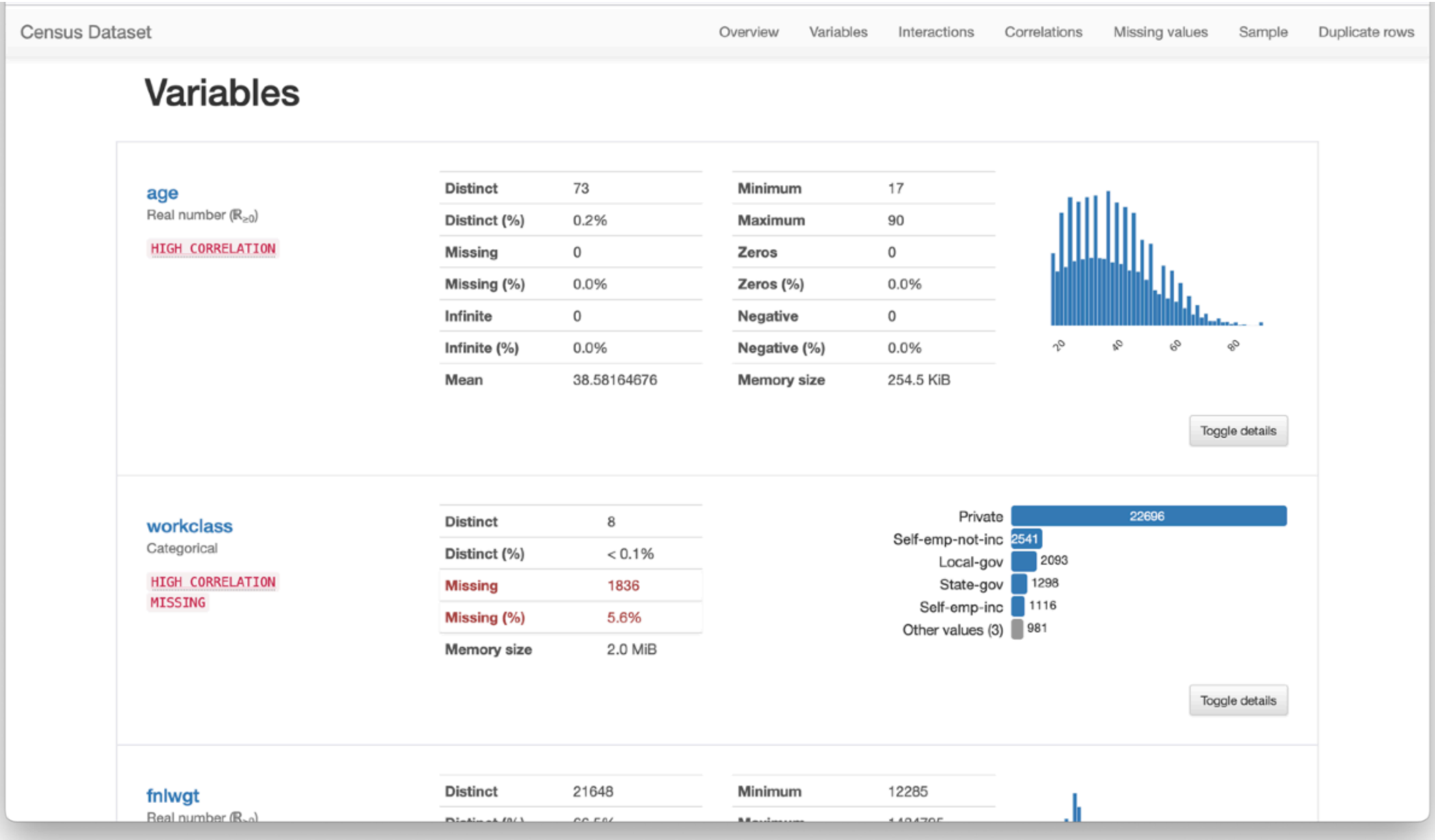
Quak widget usable in Jupyter Notebooks



	name	nationality	sex	height	weight	sport	gold	silver
	utf8	utf8	utf8	float64	int64	utf8	int64	int64
	22 categories	207 categories	female			28 categories		
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



Variables

age

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	73	**Minimum**	17
Distinct (%)	0.2%	**Maximum**	90
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	38.58164676	**Memory size**	254.5 KiB

workclass

Categorical

HIGH CORRELATION
MISSING

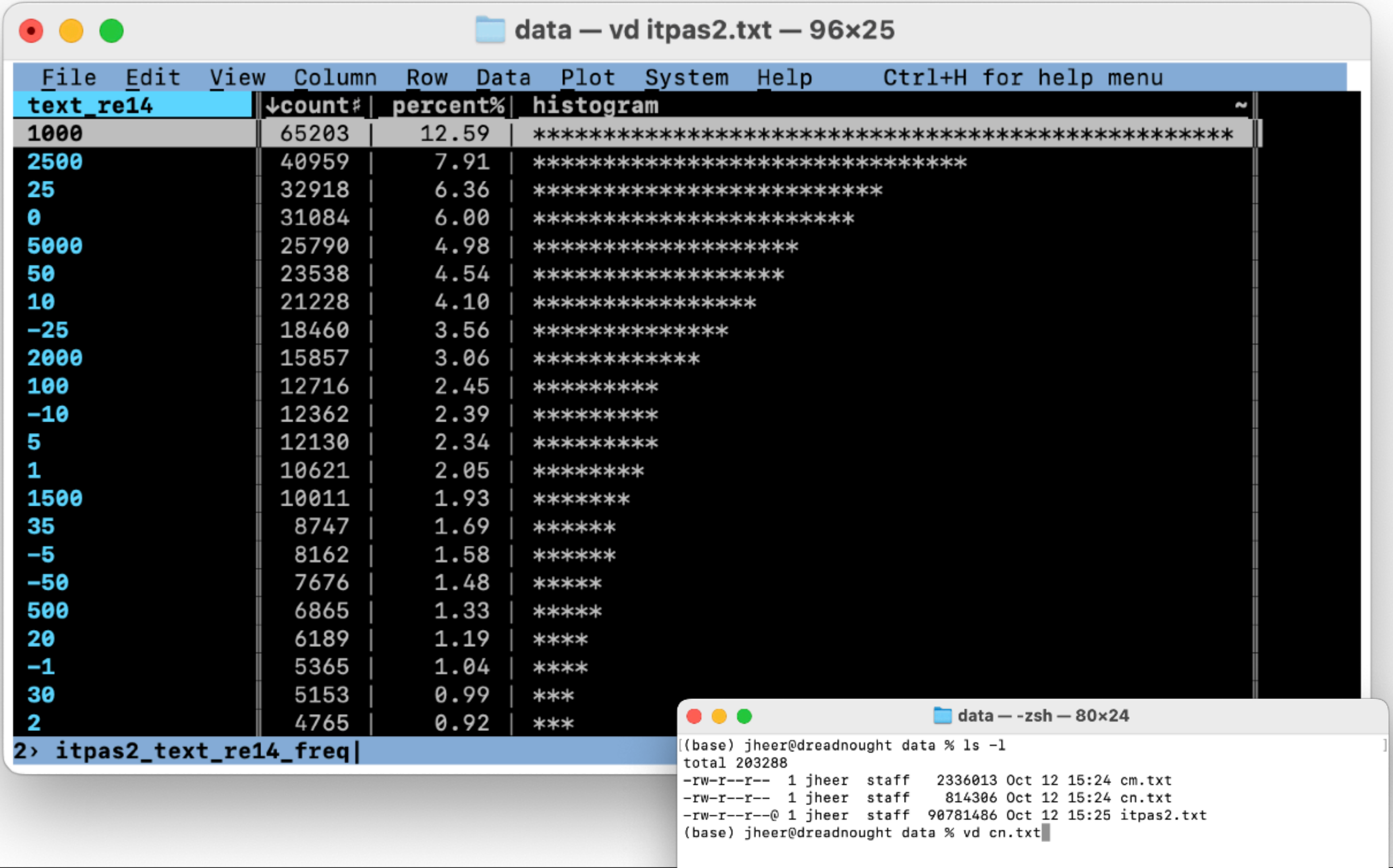
Distinct	8
Distinct (%)	< 0.1%
Missing	1836
Missing (%)	5.6%
Memory size	2.0 MiB

fnlwgt

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	21648	**Minimum**	12285

VisiData



data — vd itpas2.txt — 96×25
File Edit View Column Row Data Plot System Help Ctrl+H for help menu
data — -zsh — 80×24

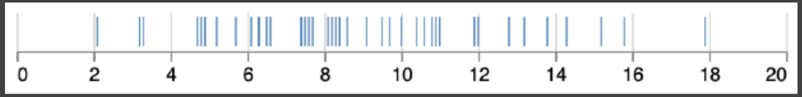
```
text_re14       ║↓count#│ percent%│ histogram                                            ~
1000            ║ 65203 │   12.59 │ **************************************************
2500            ║ 40959 │    7.91 │ *******************************
25              ║ 32918 │    6.36 │ *************************
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 │ ***
2› itpas2_text_re14_freq|
```

```
[(base) jheer@dreadnought data % ls -l
total 203288
-rw-r--r--  1 jheer  staff   2336013 Oct 12 15:24 cm.txt
-rw-r--r--  1 jheer  staff    814306 Oct 12 15:24 cn.txt
-rw-r--r--@ 1 jheer  staff  90781486 Oct 12 15:25 itpas2.txt
(base) jheer@dreadnought data % vd cn.txt
```

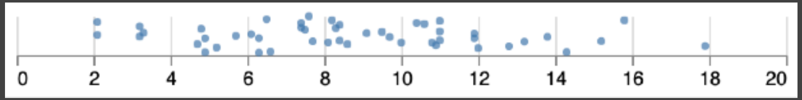
Visualizing Distributions

Distribution Visualizations

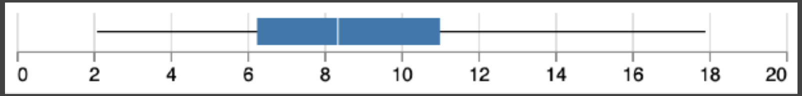
Strip Plot



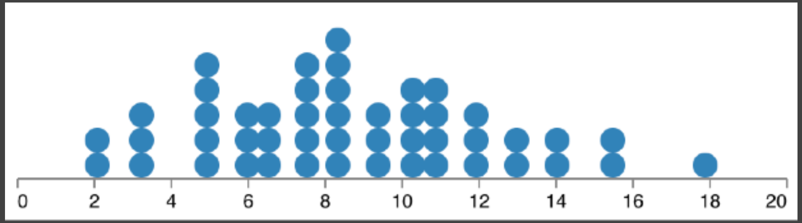
Jittered Plot



Box Plot



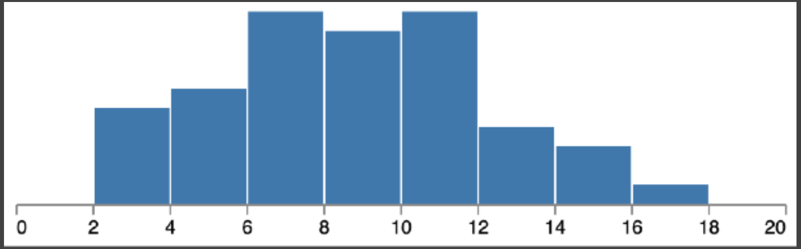
Dot Plot



Distribution Visualizations

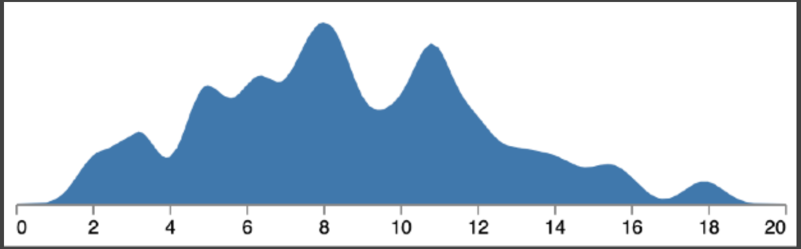
Histogram

bin size = 2



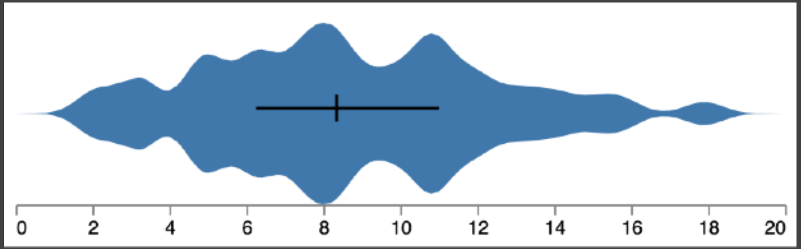
Density Plot

kde, σ = 0.5



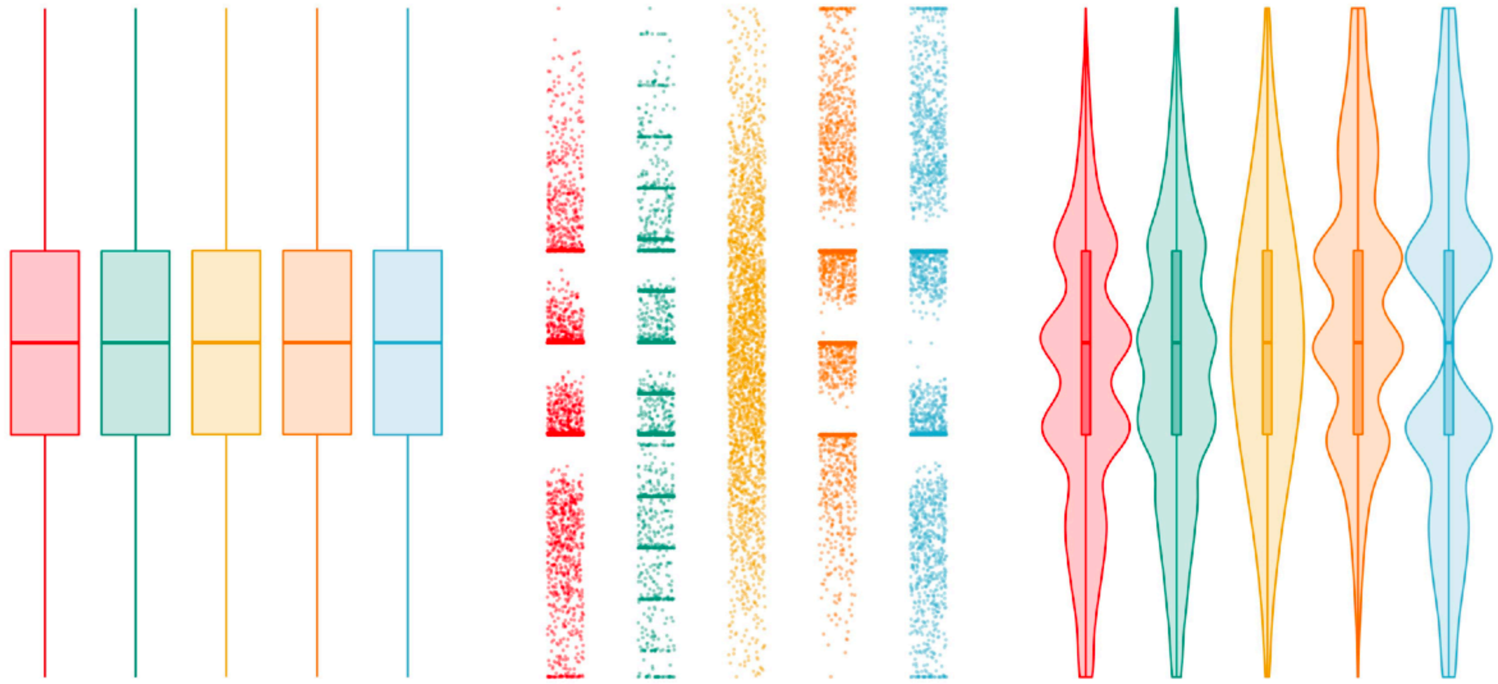
Violin Plot

kde, σ = 0.5



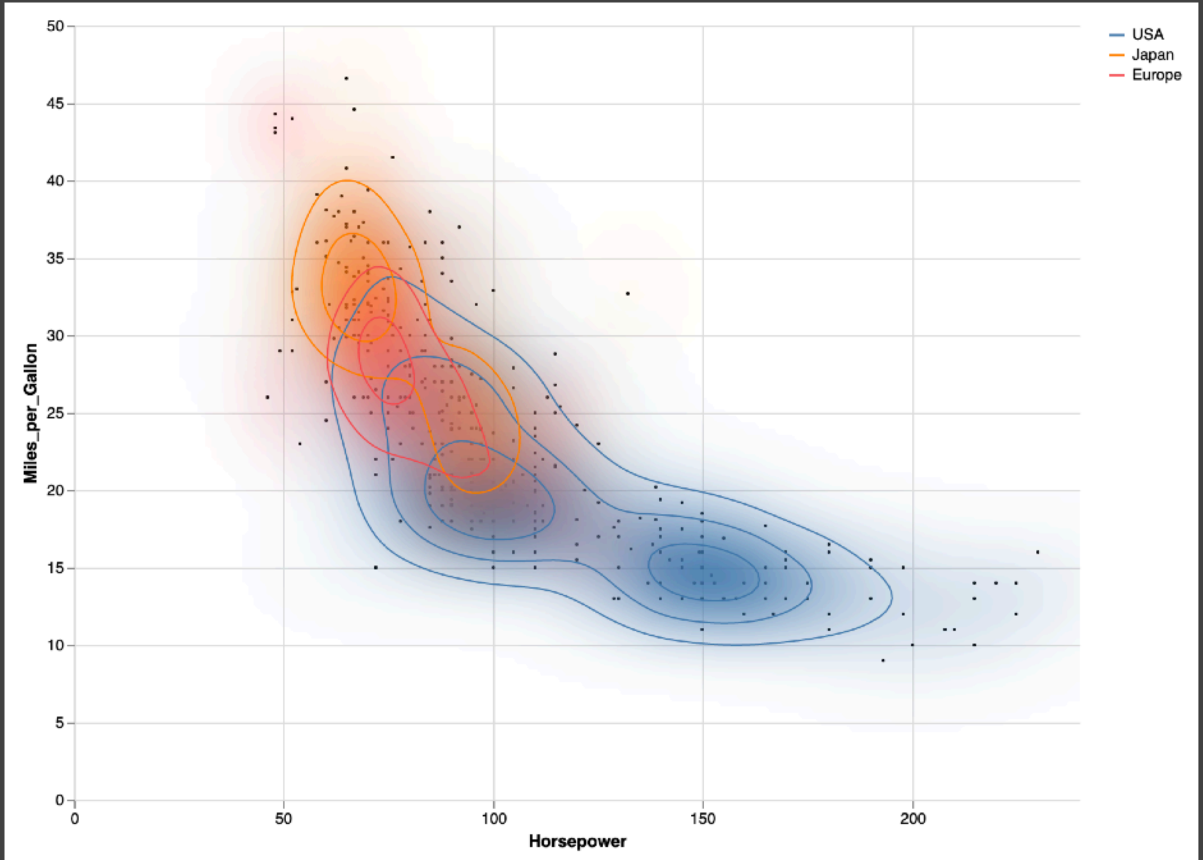
Identical boxplots, different distributions

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.



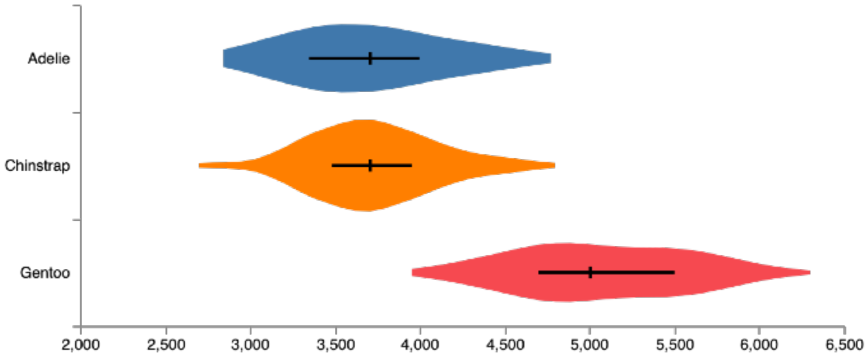
@MBozukova | #30DayChartChallenge | Day 27 - educational
Data source: {datasauRus} R package

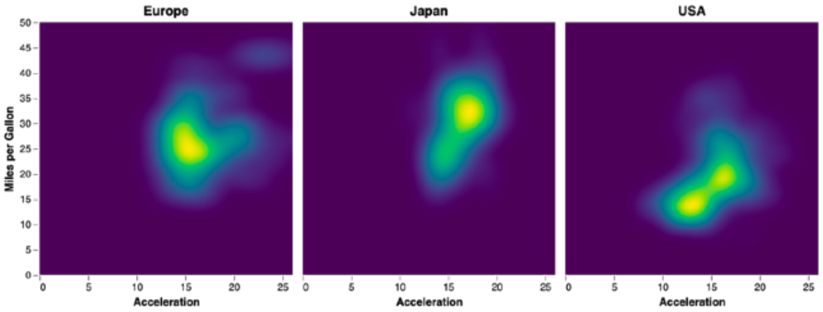
Now in 2D! Heatmaps, Contours

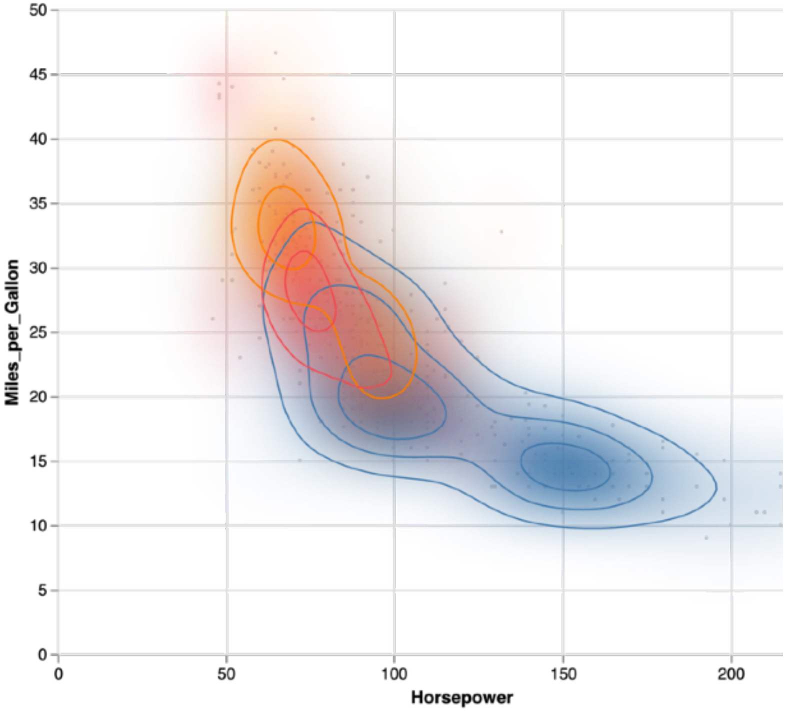


Kernel Density Estimation (KDE)

Enables violin plots, heat maps, contour plots…

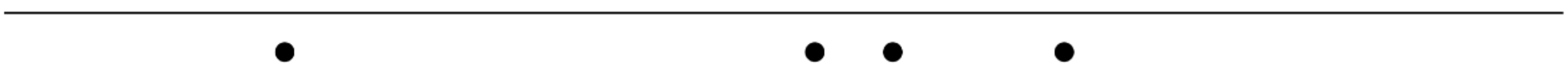






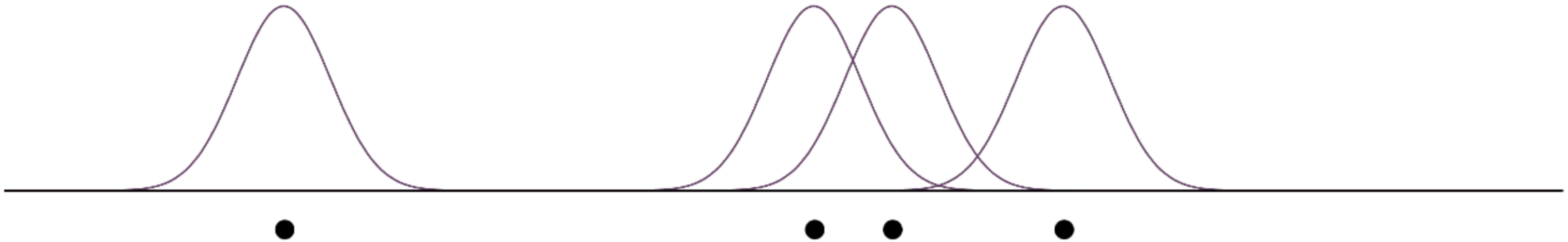
Kernel Density Estimation

For a set of input data points…



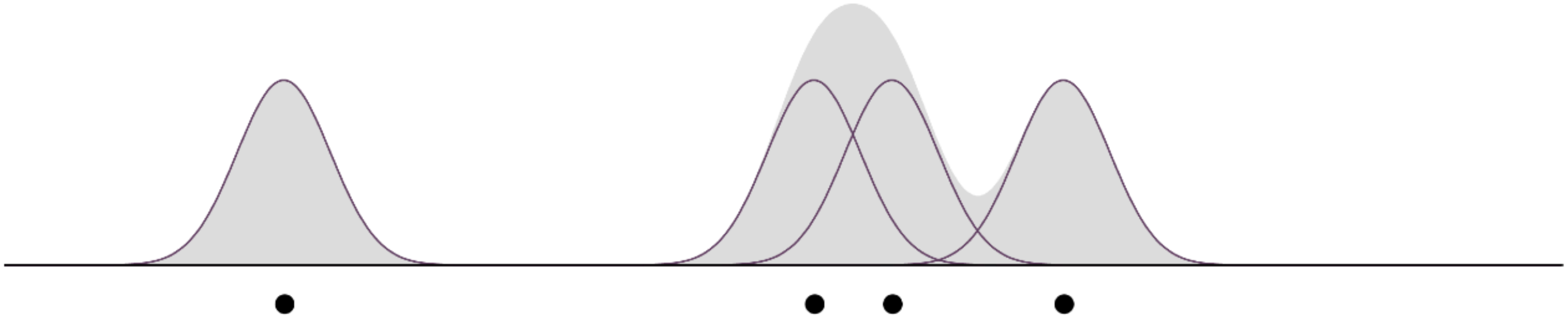
Kernel Density Estimation

Represent each point with a “kernel” distribution



Kernel Density Estimation

Sum the kernels to form a density estimate



Kernel Density Estimation

Sized by bandwidth (standard deviation)