Auto Body Refinishing

Presentation: Shane Cone
Delaware Division of Air Quality (DAQ)

*Note**

- The State of Delaware and DNREC does not endorse or recommend any commercial products, processes, or services. The views and opinions of authors expressed do not necessarily state or reflect those of the State of Delaware or any entity thereof, and they may not be used for advertising or product endorsement purposes.

Category Description

- Emissions Inventory Improvement Project (EIIP) Volume III, Chapter 13,
- “Auto body refinishing is the repairing of worn or damaged automobiles, light trucks, and other vehicles, and refers to any coating applications that occur subsequent to those at original equipment manufacturer (OEM) assembly plants. (Coating of new cars is not included in this category.) The majority of these operations occur at small body shops that repair and refinish automobiles. This category covers solvent emissions from the refinishing of automobiles, including paint solvents, thinning solvents, and solvents used for surface preparation and cleanup.”

Federal Regulation Content

Coating Category	Limit (lb/gal)*
Pretreatment Wash Primer	6.5
Primer/Primer Surfacer	4.8
Primer Sealer	4.6
Single/2-Stage Topcoats	5
Topcoats of 3 or more stages	5.2
Multicolored Topcoats	5.7
Specialty Coatings	7

*Federal Register, 1998

Delaware Air Quality Regulations

Auto Specialty	lb VOC/gal
Vacuum metalizing basecoats	5.5
Texture coatings	5.5
Reflective argent coatings	5.9
Soft specialty coatings	5.9
Gloss Flatteners	6.4
Vacuum metalizing topcoats	6.4
Texture topcoats	6.4
Stencil Coatings	6.8
Adhesion primers	6.8
Ink pad printing coatings	6.8
Electrostatic prep coats	6.8
Resist coatings	6.8

Delaware admin code, title 7, 1124 sec. 12

Old Method

- EIIP Volume III, Chapter 13
- Estimates derived from:
  - 1997 population
  - 1998 Solvent data from Connecticut
  - 1999 Coatings data from Texas
- Updated by Environ report for TX Natural Resource Conservation Commission
  - Final Report AREA AND MOBILE SOURCE EMISSIONS INVENTORY TECHNICAL SUPPORT PROJECT 1990-2010 EMISSION INVENTORY TRENDS AND PROJECTIONS

Environ 2001 report

Table 2.6-2. Emission factors used for estimating auto refinishing coating emissions.

	Facility Size Classes					
	Very Small	**Small**	**Medium**	**Large**	**Very Large**	**Mega**
Annual Revenue ($)	**<200k**	**200k - 400k**	**400k - 600k**	**600k - 1000k**	**$1.0 to 2.4 MM**	**$2.5 to 4.9 MM**
No. of employees ($100k/employee)	**1**	**2 - 3**	**4 - 6**	**7 - 9**	**10 - 24**	**> 24**
Types of Coatings (SCC Assignment)	**VOC lbs/yr**					
PreCoat Primer (2401005600)	60	130	175	305	648	1411
Primer (2401005600)	115	255	310	755	1604	3492
Sealer (2401005600)	65	145	290	315	669	1457
Base Coat (2401005700)	125	290	485	735	1562	3399
Clear Coat (2401005700)	145	300	425	815	1732	3769
Other Products (2401005700)	100	240	340	605	1286	2798
Totals	**610**	**1360**	**2025**	**3530**	**7501**	**16326**

Solvent Tool

- EPA and contractor created Access database tool
- Run using default settings
- Solvent Tool Emission Factor: 75.58 lb/Employee

DE Detailed Shop Data

- Area Sources group convinced each shop to send records of purchased painting materials
  - Only a couple of suppliers in the state
  - Supplier provided printouts of monthly purchases, which were forwarded to DAQ
- Delaware conducted statewide surveys of every auto body shop in the state to follow up/ initiate data collection
- Records include all painting supplies/products that contain VOC, including cleaning solvents

DE Detailed Shop Data Cont.

- Delaware DAQ has a nearly complete record of all paint used in category for the 2014 year
  - Out of 99 shops in the state, we have complete monthly records for 85 shops
    - 7 incomplete records (part of year reported)
    - 7 no records
  - 81 reported their number of painters
- DAQ also asked each shop to report the number of painters at the facility
  - More on this later…

Let's Compare Methodologies

Solvent Tool vs Delaware Surveys for 2014

Method	VOC Tons/year	% change from Survey
Delaware Surveys	33.4	100%
EPA Solvent Tool	226.8	679%

Remember, Delaware's 2014 data is surveys of nearly every auto shop in the state

Solvent Tool Results (defaults for DE) - 2017

County	VOC Tons/year	"Each" – Employees	State DOL Employees
Kent	60.5	1601	1524
New Castle	149.4	3952	3548
Sussex	38.2	1011	1107
Statewide	248.1	6563	6179

Tool (run 7/9/2019 on 6-10-2019 version)

Number of Painters

- Goal: Find a metric than can be easily assessed (e.g. by future surveys) and that predicts paint used

Combine our data with DOL

- Delaware DAQ has access to The Delaware Department of Labor business records
- Includes fields:
  - DOL ID number
  - Business name
  - DBA – doing business as
  - Business address
  - Place address
  - North American Industrial Classification Code (NAICS) code
  - Number of employees (each month)

Problem: How to combine

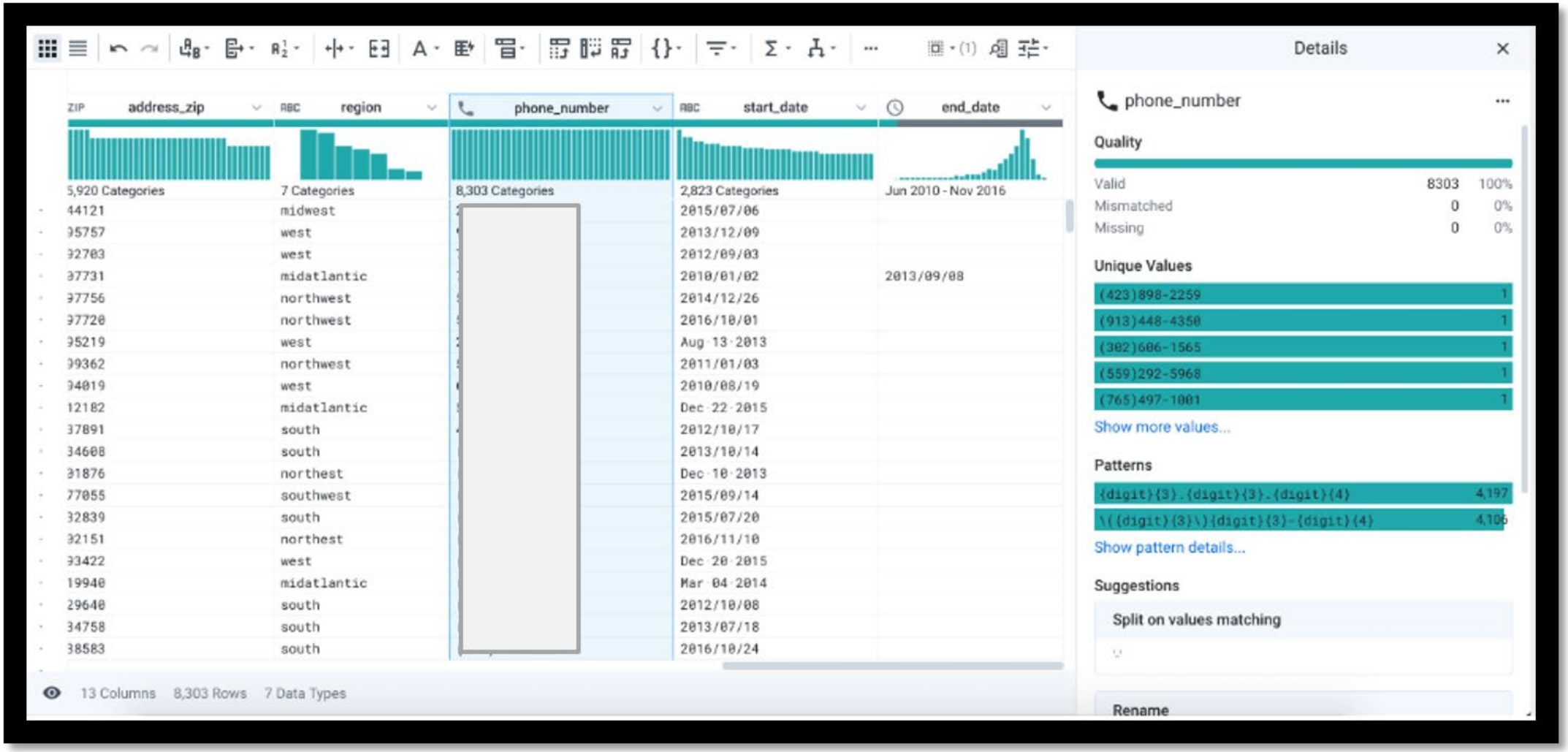
- Business name is a very poor match (<20% of records)
- Business address works for some (~40-50%, using matching algorithms and both business and place address)

Trifacta Wrangler

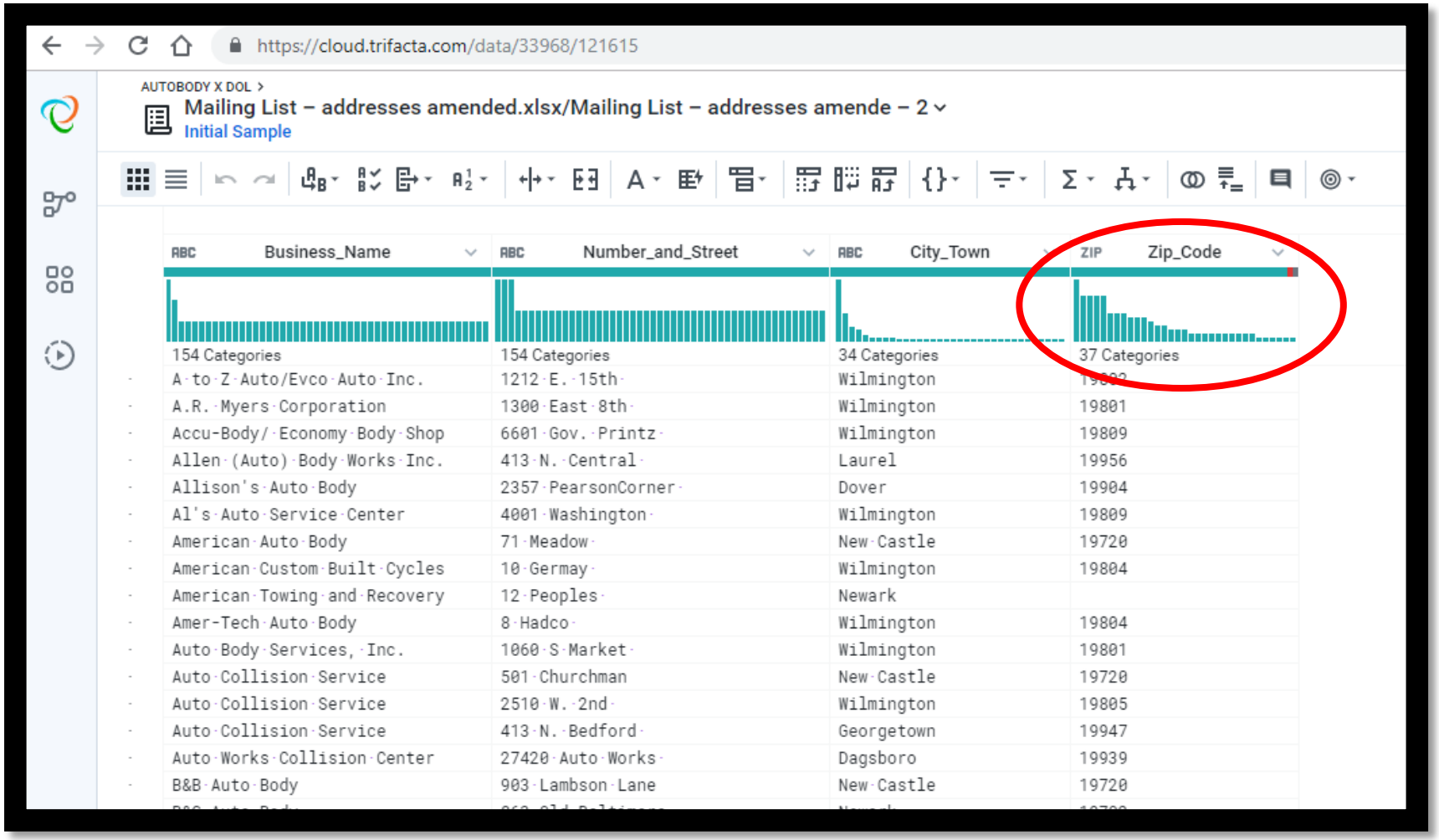
- Tool used to clean and combine data (among other things)
- Saves a ton of time
- Made it possible to combine dirty, entered-by-hand dataset from Area Sources with large (~30,000 rows) DOL dataset
- https://www.trifacta.com/products/wrangler-editions/

Trifacta "Flows"

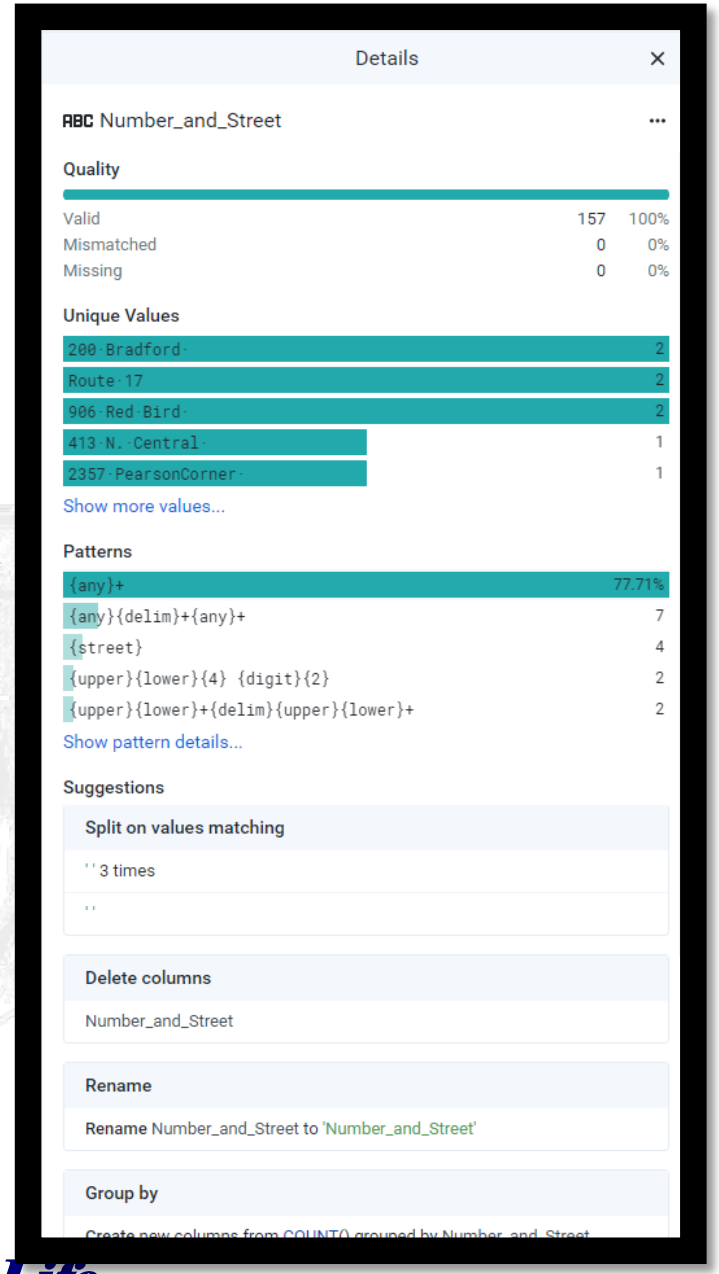
Wrangler


address_zip
region
phone_number
start_date
end_date
5,920 Categories
7 Categories
8,303 Categories
2,823 Categories
Jun 2010 - Nov 2016
44121 midwest 2015/07/06
95757 west 2013/12/09
92703 west 2012/09/03
97731 midatlantic 2010/01/02 2013/09/08
97756 northwest 2014/12/26
97720 northwest 2016/10/01
95219 west Aug 13 2013
99362 northwest 2011/01/03
94019 west 2010/08/19
12182 midatlantic Dec 22 2015
37891 south 2012/10/17
34608 south 2013/10/14
91876 northest Dec 10 2013
77055 southwest 2015/09/14
32839 south 2015/07/20
92151 northest 2016/11/10
93422 west Dec 20 2015
19940 midatlantic Mar 04 2014
29640 south 2012/10/08
34758 south 2013/07/18
38583 south 2016/10/24
13 Columns
8,303 Rows
7 Data Types
Details
phone_number
Quality
Valid 8303 100%
Mismatched 0 0%
Missing 0 0%
Unique Values
(423)898-2259 1
(913)448-4350 1
(302)606-1565 1
(559)292-5968 1
(765)497-1001 1
Show more values...
Patterns
{digit}{3}.{digit}{3}.{digit}{4} 4,197
\({digit}{3}\){digit}{3}-{digit}{4} 4,106
Show pattern details...
Suggestions
Split on values matching
Rename

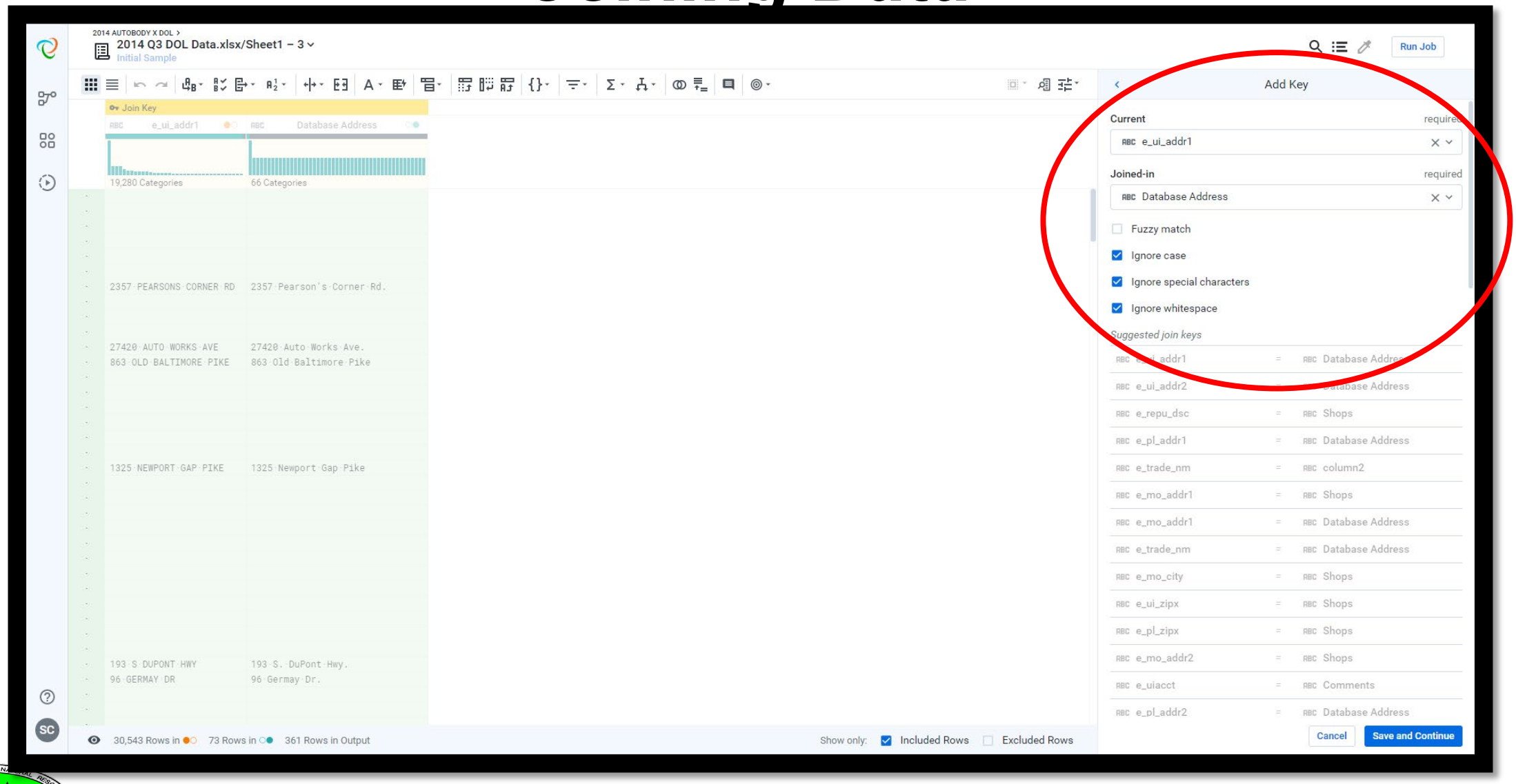
Column Histograms and Data Completeness



- More detailed column info
- Shows info regarding data formats and patterns in the data
- Great data exploration tool



Joining Data



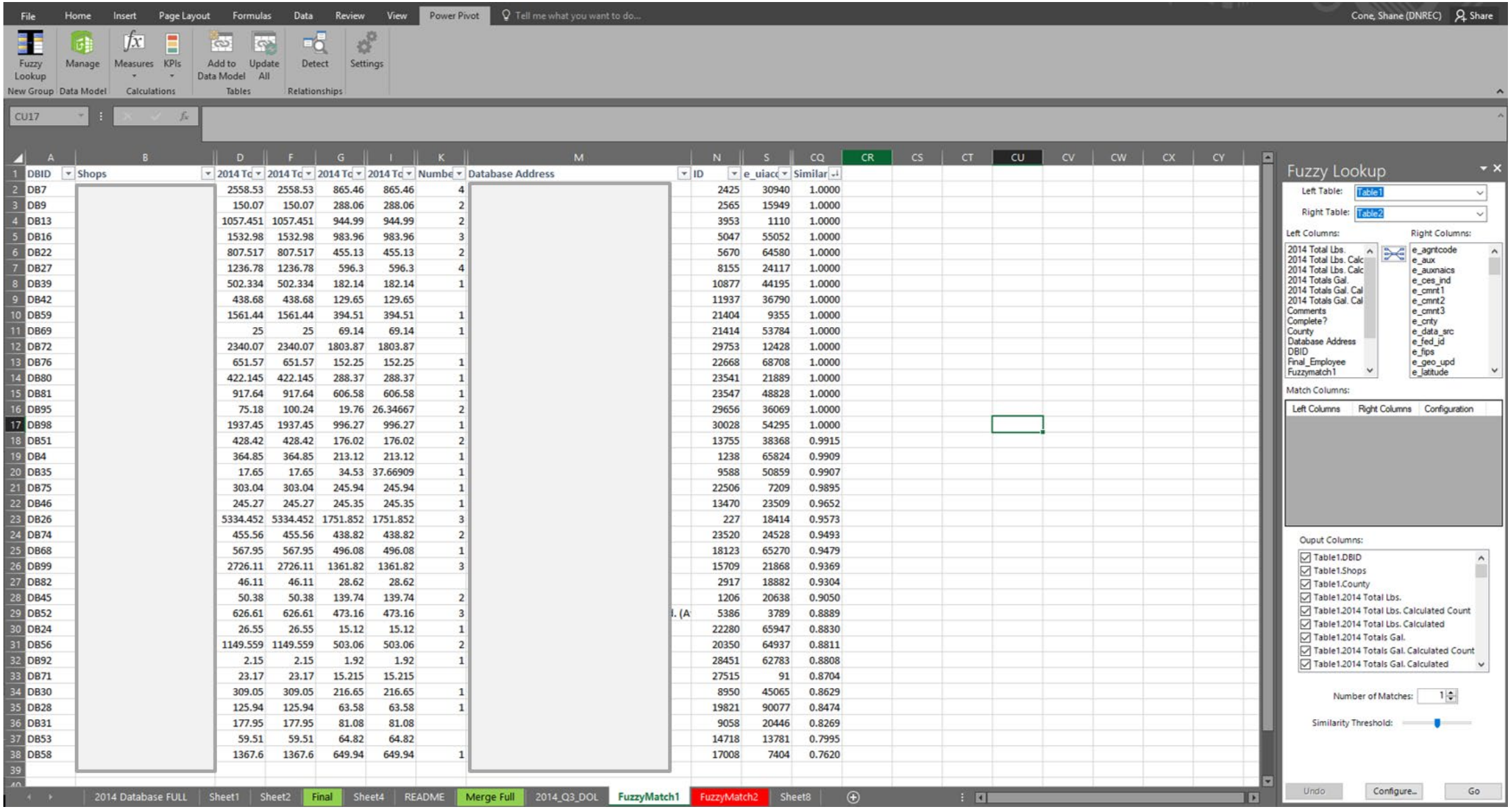
"Fuzzy Matching"

- Trifacta now uses a double the Double Metaphone matching algorithm
- Great for common words
- Not useful for integer values

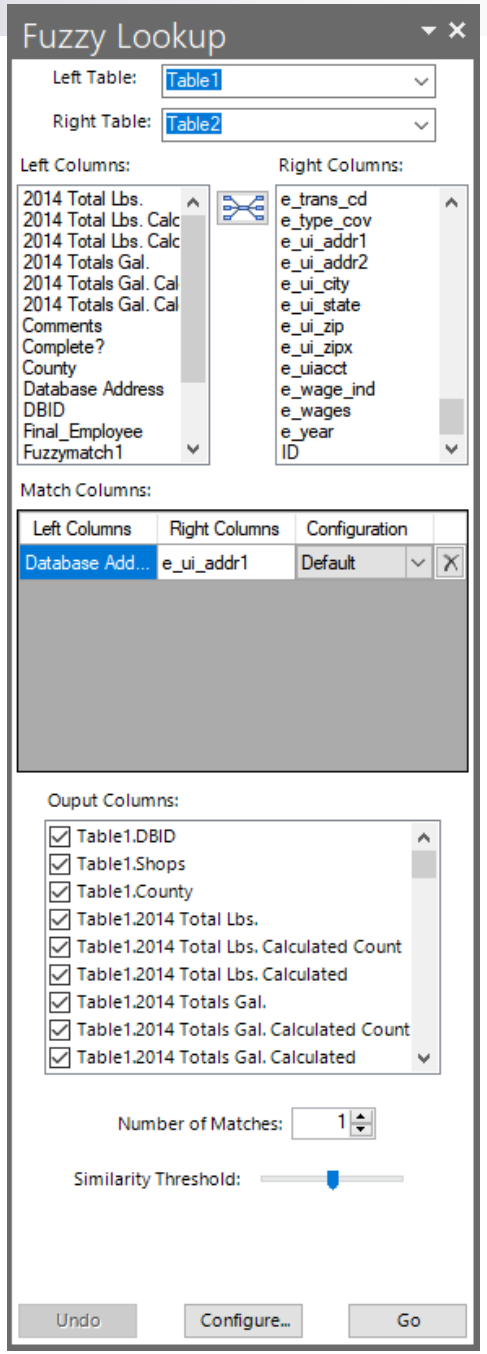
Excel "Fuzzy Lookup" add-in

- Free Add-in for Excel
  - Created and distributed by Microsoft
- Uses Jaccard Similarity
  - Set intersection divided by set union
  - Ex: {a, b, c} and {a, c, d} have Jaccard similarity of 2/4 = 0.5
- Works quite well
  - Cut off around 0.6
  - Still need to manually verify low (and high) matches

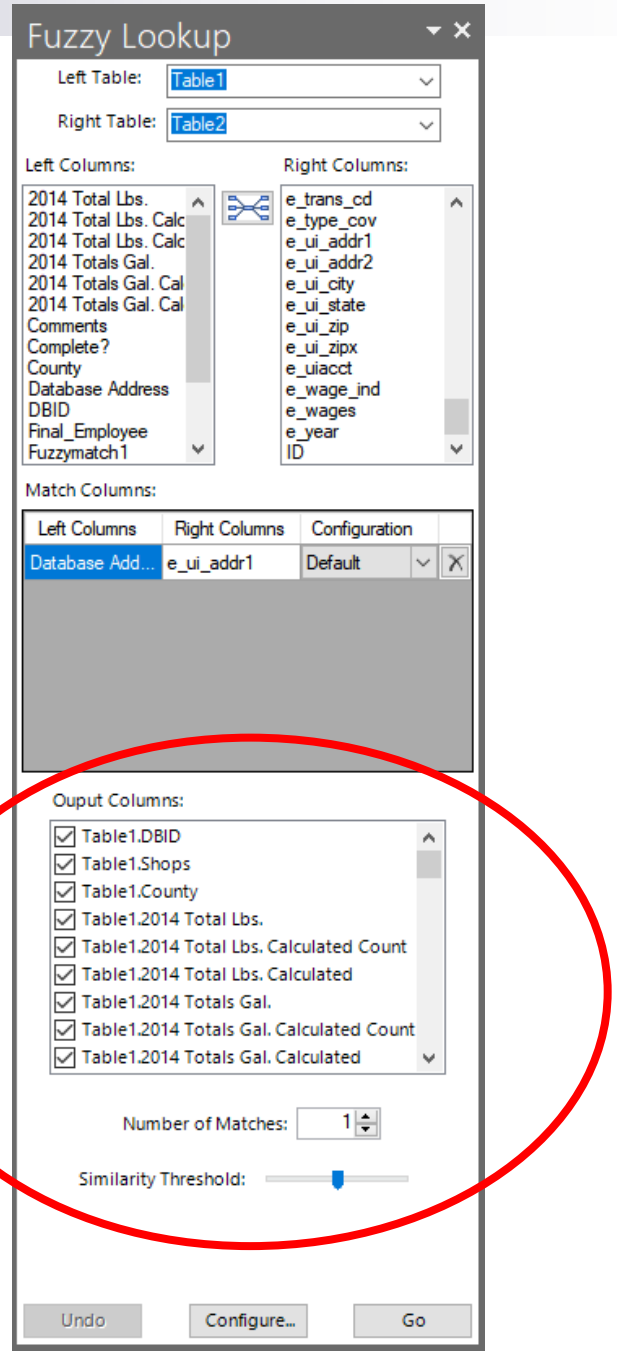
https://www.microsoft.com/en-us/download/details.aspx?id=15011


File Home Insert Page Layout Formulas Data Review View Power Pivot
Tell me what you want to do...
Cone, Shane (DNREC) Share
Fuzzy Lookup
Manage
Measures
KPIs
Add to Data Model
Update All
Detect
Settings
New Group Data Model Calculations Tables Relationships
CU17
DBID Shops 2014 Tc 2014 Tc 2014 Tc 2014 Tc Numbe Database Address ID e_uiacc Similar
DB7 2558.53 2558.53 865.46 865.46 4 2425 30940 1.0000
DB9 150.07 150.07 288.06 288.06 2 2565 15949 1.0000
DB13 1057.451 1057.451 944.99 944.99 2 3953 1110 1.0000
DB16 1532.98 1532.98 983.96 983.96 3 5047 55052 1.0000
DB22 807.517 807.517 455.13 455.13 2 5670 64580 1.0000
DB27 1236.78 1236.78 596.3 596.3 4 8155 24117 1.0000
DB39 502.334 502.334 182.14 182.14 1 10877 44195 1.0000
DB42 438.68 438.68 129.65 129.65 11937 36790 1.0000
DB59 1561.44 1561.44 394.51 394.51 1 21404 9355 1.0000
DB69 25 25 69.14 69.14 1 21414 53784 1.0000
DB72 2340.07 2340.07 1803.87 1803.87 29753 12428 1.0000
DB76 651.57 651.57 152.25 152.25 1 22668 68708 1.0000
DB80 422.145 422.145 288.37 288.37 1 23541 21889 1.0000
DB81 917.64 917.64 606.58 606.58 1 23547 48828 1.0000
DB95 75.18 100.24 19.76 26.34667 2 29656 36069 1.0000
DB98 1937.45 1937.45 996.27 996.27 1 30028 54295 1.0000
DB51 428.42 428.42 176.02 176.02 2 13755 38368 0.9915
DB4 364.85 364.85 213.12 213.12 1 1238 65824 0.9909
DB35 17.65 17.65 34.53 37.66909 1 9588 50859 0.9907
DB75 303.04 303.04 245.94 245.94 1 22506 7209 0.9895
DB46 245.27 245.27 245.35 245.35 1 13470 23509 0.9652
DB26 5334.452 5334.452 1751.852 1751.852 3 227 18414 0.9573
DB74 455.56 455.56 438.82 438.82 2 23520 24528 0.9493
DB68 567.95 567.95 496.08 496.08 1 18123 65270 0.9479
DB99 2726.11 2726.11 1361.82 1361.82 3 15709 21868 0.9369
DB82 46.11 46.11 28.62 28.62 2917 18882 0.9304
DB45 50.38 50.38 139.74 139.74 2 1206 20638 0.9050
DB52 626.61 626.61 473.16 473.16 3 5386 3789 0.8889
DB24 26.55 26.55 15.12 15.12 1 22280 65947 0.8830
DB56 1149.559 1149.559 503.06 503.06 2 20350 64937 0.8811
DB92 2.15 2.15 1.92 1.92 1 28451 62783 0.8808
DB71 23.17 23.17 15.215 15.215 27515 91 0.8704
DB30 309.05 309.05 216.65 216.65 1 8950 45065 0.8629
DB28 125.94 125.94 63.58 63.58 1 19821 90077 0.8474
DB31 177.95 177.95 81.08 81.08 9058 20446 0.8269
DB53 59.51 59.51 64.82 64.82 14718 13781 0.7995
DB58 1367.6 1367.6 649.94 649.94 1 17008 7404 0.7620
2014 Database FULL Sheet1 Sheet2 Final Sheet4 README Merge Full 2014_Q3_DOL FuzzyMatch1 FuzzyMatch2 Sheet8
Fuzzy Lookup
Left Table: Table1
Right Table: Table2
Left Columns:
2014 Total Lbs.
2014 Total Lbs. Calc
2014 Total Lbs. Calc
2014 Totals Gal.
2014 Totals Gal. Cal
2014 Totals Gal. Cal
Comments
Complete?
County
Database Address
DBID
Final_Employee
Fuzzymatch1
Right Columns:
e_agntcode
e_aux
e_auxnaics
e_ces_ind
e_cmnt1
e_cmnt2
e_cmnt3
e_cnty
e_data_src
e_fed_id
e_fips
e_geo_upd
e_latitude
Match Columns:
Left Columns Right Columns Configuration
Ouput Columns:
Table1.DBID
Table1.Shops
Table1.County
Table1.2014 Total Lbs.
Table1.2014 Total Lbs. Calculated Count
Table1.2014 Total Lbs. Calculated
Table1.2014 Totals Gal.
Table1.2014 Totals Gal. Calculated Count
Table1.2014 Totals Gal. Calculated
Number of Matches: 1
Similarity Threshold:
Undo Configure... Go

- Like any join, select columns to use as join key
- Data sources MUST BE named tables (names ranges)



- Like any join, select columns to use as join key
- Data sources MUST BE named tables (names ranges)
- Bottom half of panel
  - Choose columns to output
  - Choose number of matches (left join vs right join)
  - Choose Jaccard Similarity Threshold



Carefully Check Matches!

Shops	Database Address	DOL Business Name	DOL Address_addr1	Jaccard Similarity
Delaware Cadillac Body Shop	3408 Lancaster Pike	A DOOR OF HOPE PREGNANCY CENTER INC	3407 LANCASTER PIKE	0.9573

Creating an Emission Factor

Cleaning the data

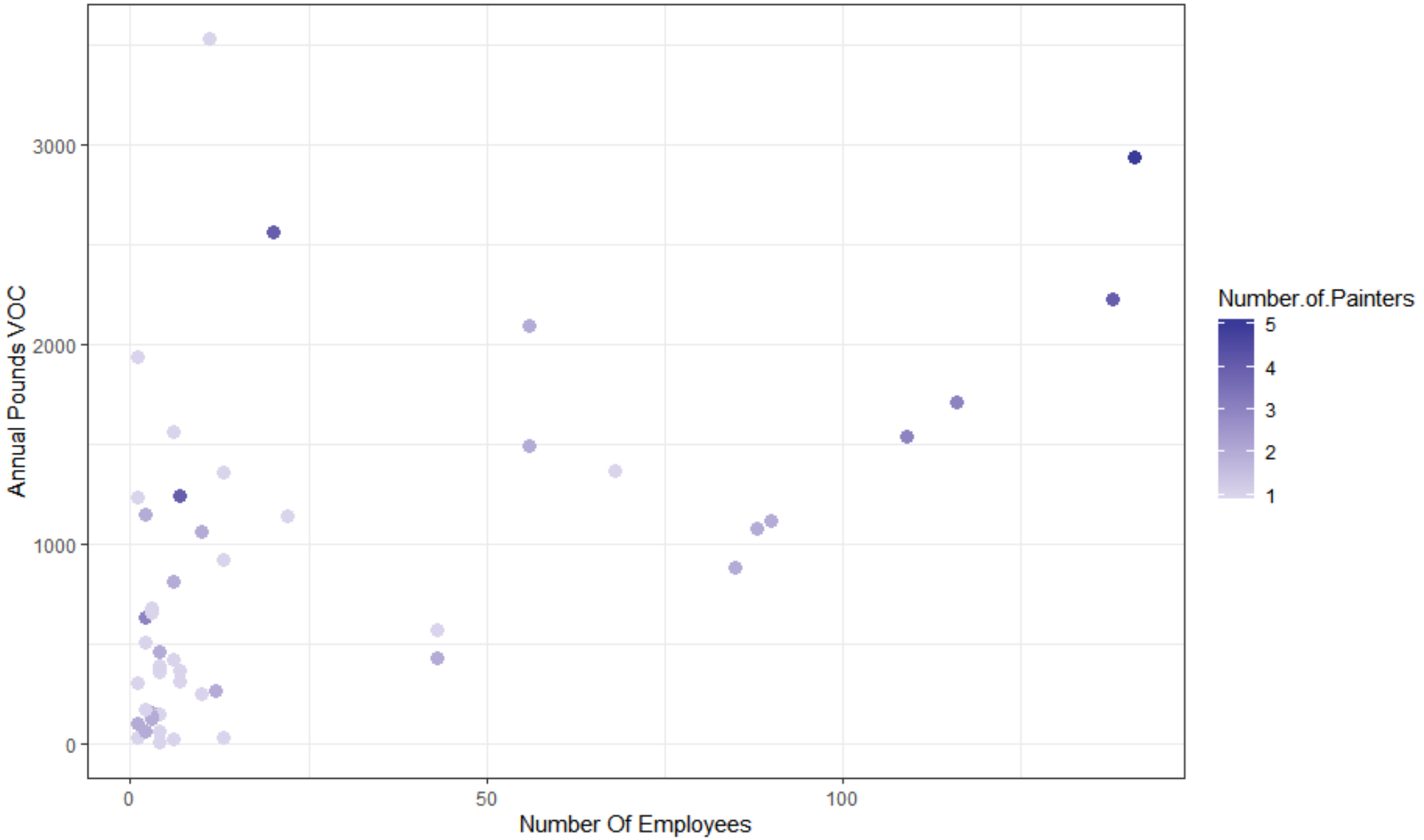
- In some instances (less than half), not all months had reported data, for either LBS VOC or GAL PAINT
- When looking at both fields, if more than 12 reports were filed (i.e. 12 months of LBS VOC, or 6 months of each, etc.), a total annual value was calculated
  - Summed field * 12 / (# of reported months)
- Significant number of missing / unmatchable businesses from our database to DOL-supplied database

Cleaning the data Pt. 2

- Any shop with paint data reported, but 0 employees
  - Changed employees to # of painters
  - Likely a one-man operation, and does nor report self in all months as “employee” – ex. “Gene’s auto body shop”
  - In 2014-2016, 6 instances of DOL data as “0”

Outliers

- Some car dealerships have huge employment
  - They also tend to have high VOC emissions
- Largest outliers – Vocational Technical High Schools
  - They tend to have very few emissions
  - Employment of ~700
  - Set Employment equal to number of painters


Annual Pounds VOC
3000
2000
1000
0
0
50
100
Number Of Employees
Number.of.Painters
5
4
3
2
1

Descriptive Statistics

	Annual Pounds VOC	Annual Gallons Paint	Number of Painters	Number of Employees
Min	2.2	1.9	1	1
1st Qu	255.2	161.9	1	3
Median	568	300.2	1	6
Mean	847	458	1.7	24.8
3rd Qu	1235	636.6	2	21
Max	3525	1741	5	141

Model Fitting

- Partition data to Training and Test
- Remove Null observations
- Use Base R “lm” function to train model on Training data
- n = 98

Generalized Linear Model with Painters

Measure	Value
MSE	657,262
RMSE	810.7
Mean Absolute Error	661
Median Absolute Error	531.6
R^2	**0.12**

Generalized Linear Model with Painters

Measure	Value
MSE	657,262
RMSE	810.7
Mean Absolute Error	661
Median Absolute Error	531.6
R^2	**0.12**

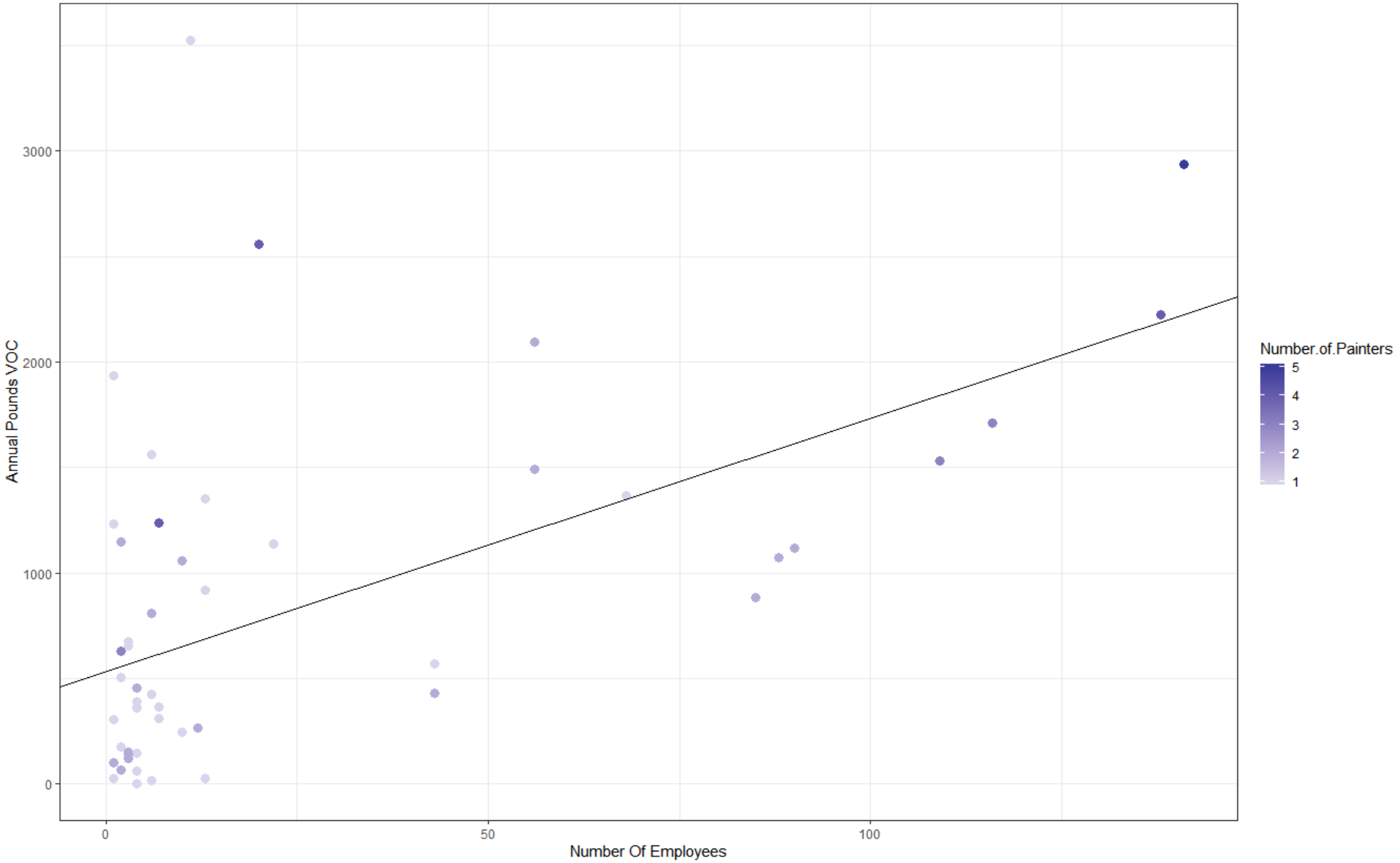
Turns out that modeling emissions based on painters alone is not a good option

Generalized Linear Model with Employees

Measure	Value
MSE	476,445
RMSE	690.2
Mean Absolute Error	550.78
Median Absolute Error	518.28
R^2	**0.33**

Generalized Linear Model with Employees and Painters

Measure	Value
MSE	483,250
RMSE	695
Mean Absolute Error	554
Median Absolute Error	520
R^2	**0.33**


Annual Pounds VOC
3000
2000
1000
0
0
50
100
Number Of Employees
Number.of.Painters
5
4
3
2
1

Model Results (employees only)

	Estimate	std. error
(Intercept)	530.2489	133.6039
Final_Employee	12.0031	2.805941

Linear Model Code

```
# Import packages
library(ggplot2)
library(repr)
library(dplyr)
library(caret)

options(repr.plot.width=4, repr.plot.height=4) # Set the initial plot area dimensions
AutoBody14_16 <- read.csv("Autobody Database Modeling ready 2014_2016.csv")
AutoBody14_16Trim <- AutoBody14_16[ , c("X2014.Total.Lbs..Calculated", "X2014.Totals.Gal..Calculated",
"Number.of.Painters", "Final_Employee")]
AutoBody14_16Model <-  na.omit(AutoBody14_16Trim)
View(AutoBody14_16Model)
set.seed(1955)
## Randomly sample cases to create independent training and test data
partition = createDataPartition(AutoBody14_16Model[,"X2014.Total.Lbs..Calculated"], times = 1, p = 0.75, list =
FALSE)
training = AutoBody14_16Model[partition,] # Create the training sample
dim(training)
test = AutoBody14_16Model[-partition,] # Create the test sample
dim(test)
## define and fit the linear regression model
lin_mod = lm(X2014.Total.Lbs..Calculated ~ Number.of.Painters + Final_Employee, data = training)
summary(lin_mod)$coefficients
print_metrics(lin_mod, test, score, label = 'X2014.Total.Lbs..Calculated')
print_metrics = function(lin_mod, df, score, label){
resids = df[,label] - score
resids2 = resids**2
N = length(score)
r2 = as.character(round(summary(lin_mod)$r.squared, 4))
adj_r2 = as.character(round(summary(lin_mod)$adj.r.squared, 4))
cat(paste('Mean Square Error      = ', as.character(round(sum(resids2)/N, 4)), '\n'))
cat(paste('Root Mean Square Error = ', as.character(round(sqrt(sum(resids2)/N), 4)), '\n'))
cat(paste('Mean Absolute Error    = ', as.character(round(sum(abs(resids))/N, 4)), '\n'))
cat(paste('Median Absolute Error  = ', as.character(round(median(abs(resids)), 4)), '\n'))
cat(paste('R^2                    = ', r2, '\n'))
cat(paste('Adjusted R^2           = ', adj_r2, '\n'))
}
score = predict(lin_mod, newdata = test)
print_metrics(lin_mod, test, score, label = 'X2014.Total.Lbs..Calculated')
hist_resids = function(df, score, label, bins = 10){
options(repr.plot.width=4, repr.plot.height=3) # Set the initial plot area dimensions
df$resids = df[,label] - score
bw = (max(df$resids) - min(df$resids))/(bins + 1)
ggplot(df, aes(resids)) +
geom_histogram(binwidth = bw, aes(y=..density..), alpha = 0.5) +
geom_density(aes(y=..density..), color = 'blue') +
xlab('Residual value') + ggtitle('Histogram of residuals')
}
hist_resids(test, score, label = 'X2014.Total.Lbs..Calculated')
#QQ plot of the residuals. A 1:1 line would indicate perfectly normally distributed residuals
resids_qq = function(df, score, label){
    options(repr.plot.width=4, repr.plot.height=3.5) # Set the initial plot area dimensions
    df$resids = df[,label] - score
    ggplot() +
    geom_qq(data = df, aes(sample = resids)) +
    ylab('Quantiles of residuals') + xlab('Quantiles of standard Normal') +
    ggtitle('QQ plot of residual values')
}

resids_qq(test, score, label = 'X2014.Total.Lbs..Calculated')
```

Code also available on Github: https://github.com/microscone/Autobody-Emission-Factor

Lessons Learned

- This was an exercise in the importance of using a relational database to keep track of data
  - Each shop is entered once, and given a primary key
  - If name changes, primary key stays same
  - If addresses change, ….
    - Probably a new shop
  - Align datasets once – keep as foreign key
- Some shops were not in DOL data in some years…
  - Probably a case of mismatched names, addresses, etc.

Going Forward

- This work shows that predicting any one auto refinishing shop's emissions is quite difficult
- Delaware is happy with outcome, as 2014 response was nearly entire population of shops
- May have limited use outside of Delaware, especially for non-Eastern region large states
  - Delaware has relatively few large shops
  - Our largest shops may not even qualify as "Large" in other states

Going Forward

- Employee Numbers
  - DOL Filter based on NAICS in Solvent Tool for 2014
    - ~6,000
  - DOL Employee numbers for all facilities in DAQ 2014 Database
    - ~3,000