CS412: Introduction to Data Mining **Fall 2015**

Assignment 1

Li Miao *September 22, 2015*

Question 1

Answers

a. (6') $Q_1 = 82$, median = 89 , $Q_3 = 95$.

b. (3') Mean = 87.011.

c. (3') Mode = 95.

d. (3') The distribution of students' final scores is negatively skewed. The mean is less than the median, and both of them are less than the mode.

Question 2 (15 points)

Answers

a.

$$sim(obj1, obj2) = \frac{21}{21 + 28 + 39} = 0.2386$$

		Obj 2	
		1	0
Obj 1	1	21	28
	0	39	112

Table 1: Contingency Table for *Obj1* and *Obj 2*

b. $A = (3, 1, 2)$ and $B = (-1, 0, 8)$.

1. *Euclidean* distance.

$$d(A, B) = \sqrt{(3+1)^2 + (1)^2 + (2-8)^2} = 7.2801$$

2. *Manhattan* distance.

$$d(A, B) = |3 + 1| + |1 - 0| + |2 - 8| = 11$$

3. *Minkowski* distance where $h = \infty$.

$$d(A, B) = max\{|3 + 1|, |1 - 0|, |2 - 8|\} = 6$$

c. The *Euclidean* distance between A and B is always shorter than (or equal to) the *Manhattan* distance because of the triangle inequality. *Manhattan* distance is the sum of two legs, and *Euclidean* distance is the hypotenuse.

d. 1. *Minkowski* distance where $h = 2$ is 412.941.

2. *Minkowski* distance where $h = 3$ is 216.448.

Question 3

Answers

a. Before normalization:
mean = 76.814, variance = 171.396
After normalization:
mean = 0, variance = 1

b. For original score of 90, $z = 1.007$

Question 4

Answers
Consider 10 data points in 2-D space as specified in the table below.

X	0.69	-1.31	0.39	0.05	1.29	0.49	0.19	-0.81	-0.31	0.71
Y	0.89	-1.11	0.59	0.45	1.19	0.69	0.25	-0.71	-0.21	0.71

a.

$$r_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{n\sigma_x\sigma_y} = \frac{5.3858 - 10 * 0.138 * 0.274}{10 * 0.7712 * 0.7327} = 0.283$$

$r_{x,y} > 0$, so X and Y are positively correlated. X's values increase as Y's.

b. PCA may help to reduce the data size.Because we can minimize the co-variance between X and Y.

c.

$$C_{x,y} = \begin{bmatrix} E[(X - \mu_x)(X - \mu_x)] & E[(X - \mu_x)(Y - \mu_y)] \\ E[(Y - \mu_y)(X - \mu_x)] & E[(Y - \mu_y)(Y - \mu_y)] \end{bmatrix}$$

$$= \frac{1}{10} * \begin{bmatrix} (X - \mu_x)(X - \mu_x)^T & (X - \mu_x)(Y - \mu_y)^T \\ (Y - \mu_y)(X - \mu_x)^T & (Y - \mu_y)(Y - \mu_y)^T \end{bmatrix}$$

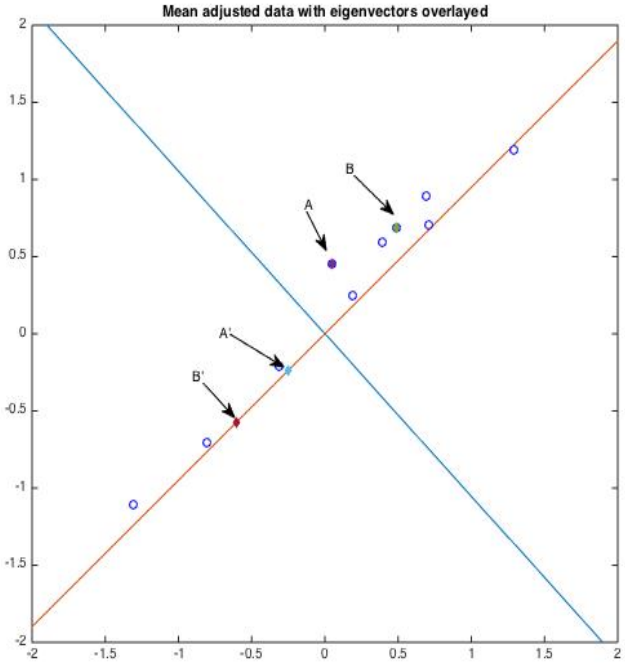
$$= \begin{bmatrix} 0.5353 & 0.5008 \\ 0.5008 & 0.4831 \end{bmatrix}$$

d. Use Matlab, we can find eigenvectors of $C_{x,y}$

$$P = \begin{bmatrix} 0.6885 & -0.7253 \\ -0.7253 & -0.6885 \end{bmatrix}$$

Each row in P is a eigenvector.

There are two principal components. [0.6885,-0.7253] and [-0.7253,-0.6885]. The first principal component is [-0.7253,-0.6885] because its eigenvalue 1.0106, is larger than 0.0078, the eigenvalue of [0.6885,-0.7253].



e.

f. $A = (0.05, 0.45), B = (0.49, 0.69)$

$A\prime = [-0.7253, -0.6885] * A = -0.3461$

$B\prime = [-0.7253, -0.6885] * B = -0.8304$

Mini Machine Problem 1

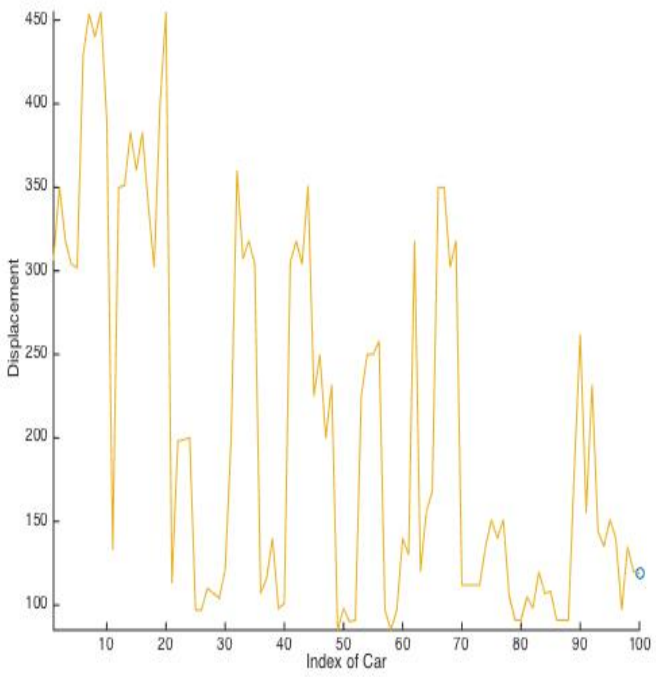
Answers

1. Load the data **carsmall** in Matlab using the following code.

```
load carsmall
X = [MPG,Acceleration,Displacement,Weight,Horsepower];
varNames = {'MPG'; 'Acceleration'; 'Displacement'; 'Weight'; 'Horsepower'};
```

2. (Comet graph is an animated graph. To trace the data points on the screen for the `Displacement` attribute, we use the following code to visualize the `Displacement` attribute. Show the **final comet graph** in the PDF file you will submit by running the following code on Matlab.

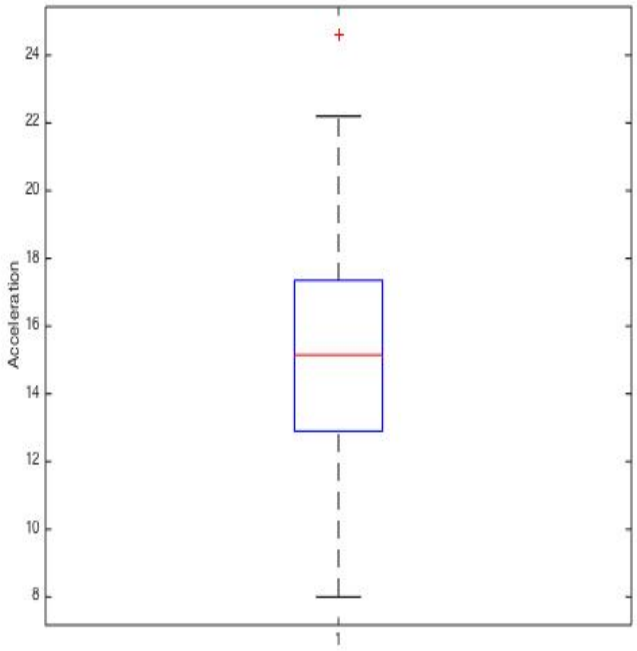
```
comet(Displacement)
xlabel('Index of Car')
ylabel('Displacement')
```



3. (5') Drawing boxlpot is a popular way to visualize a distribution. The two whiskers show the Min observation and the Max observation. The central line shows the median. The edges of the box are the first quantile and the third quantile.

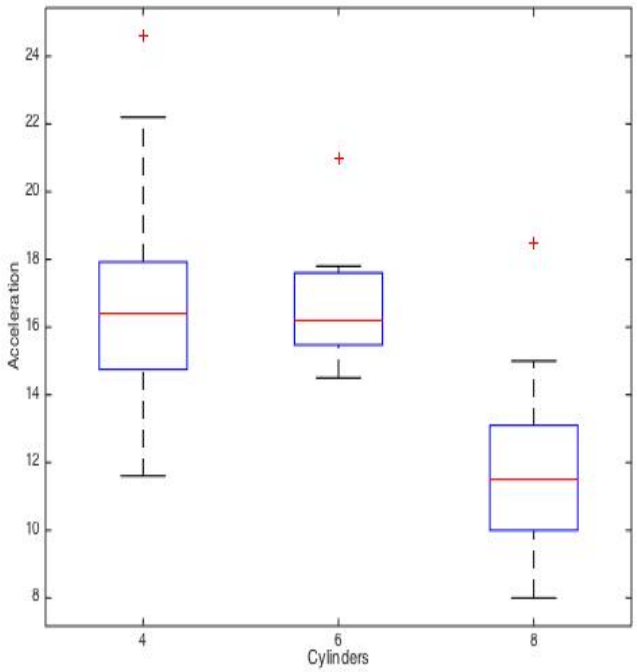
   a. (1') Run the following code on your Matlab to draw a boxplot for the `Acceleration` attribute. Show the **boxplot** in the PDF file you will submit.

```
boxplot(Acceleration)
ylabel('Acceleration')
```



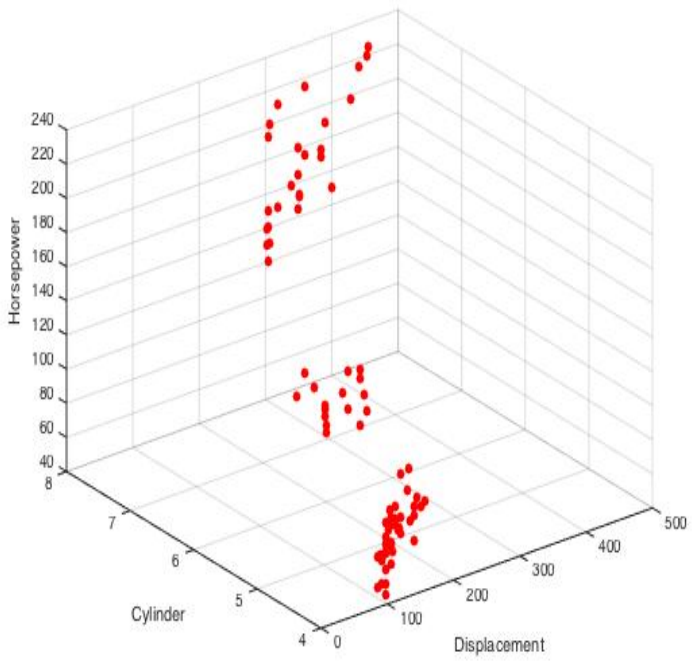
b.

```
boxplot(Acceleration,Cylinders)
xlabel('Cylinders')
ylabel('Acceleration')
```



4. (4') 3-D scatter plots are popularly used to visualize 3 attributes at the same time.
   a. (2') Run the following code to draw a 3-D scatter plot. Show the **3-D plot** in the PDF file you will submit.

```
scatter3(Displacement,Cylinders,Horsepower,'filled','r')
xlabel('Displacement')
ylabel('Cylinders')
zlabel('Horsepower')
```

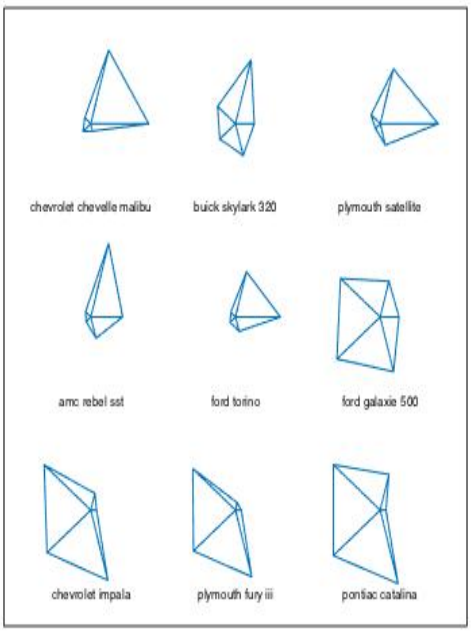


b. Horsepower and displacement a pair of positively correlated attributes. Displacement is the total volume of all the cylindets in an engine. Larger engines tend to produce more power, more turque.

5. (4') Interactive star plots are used to show the values of attributes for each observation. In each star (observation), the spoke length is proportional to the value of that attribute for that observation.

   a. (2') Run the following code. Show the **graph** in PDF file you will submit.

```
h = glyphplot(X(1:9,:), 'glyph','star', 'varLabels',varNames,...
'obslabels',Model(1:9,:));
set(h(:,3),'FontSize',8);
```

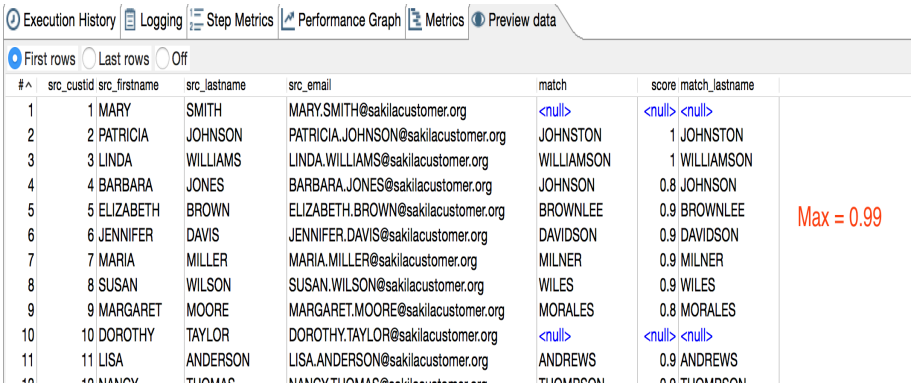


b. Observation: Chevrolet chevelle malibu
MPG: 18
Acceleration: 12
Displacement: 307
Weight: 3504
Horsepower: 130

Mini Machine Problem 2

Answers

1. When $max = 1$, what we do is just to do the exact match. That means each name will match to itself. And the match will always be one. When $max = 0.99$, we are tring to match those names that are similar but not exactly the same.

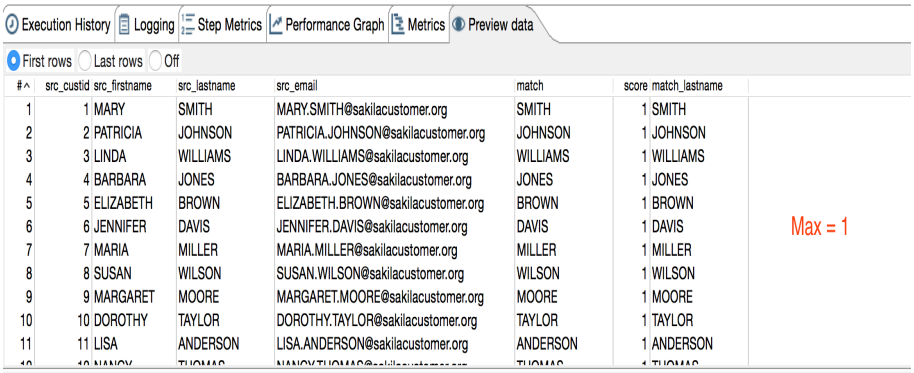


Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	src_custid	src_firstname	src_lastname	src_email	match	score	match_lastname
1	1	MARY	SMITH	MARY.SMITH@sakilacustomer.org	<null>	<null>	<null>
2	2	PATRICIA	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSTON	1	JOHNSTON
3	3	LINDA	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMSON	1	WILLIAMSON
4	4	BARBARA	JONES	BARBARA.JONES@sakilacustomer.org	JOHNSON	0.8	JOHNSON
5	5	ELIZABETH	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWNLEE	0.9	BROWNLEE
6	6	JENNIFER	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIDSON	0.9	DAVIDSON
7	7	MARIA	MILLER	MARIA.MILLER@sakilacustomer.org	MILNER	0.9	MILNER
8	8	SUSAN	WILSON	SUSAN.WILSON@sakilacustomer.org	WILES	0.9	WILES
9	9	MARGARET	MOORE	MARGARET.MOORE@sakilacustomer.org	MORALES	0.8	MORALES
10	10	DOROTHY	TAYLOR	DOROTHY.TAYLOR@sakilacustomer.org	<null>	<null>	<null>
11	11	LISA	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDREWS	0.9	ANDREWS

Max = 0.99

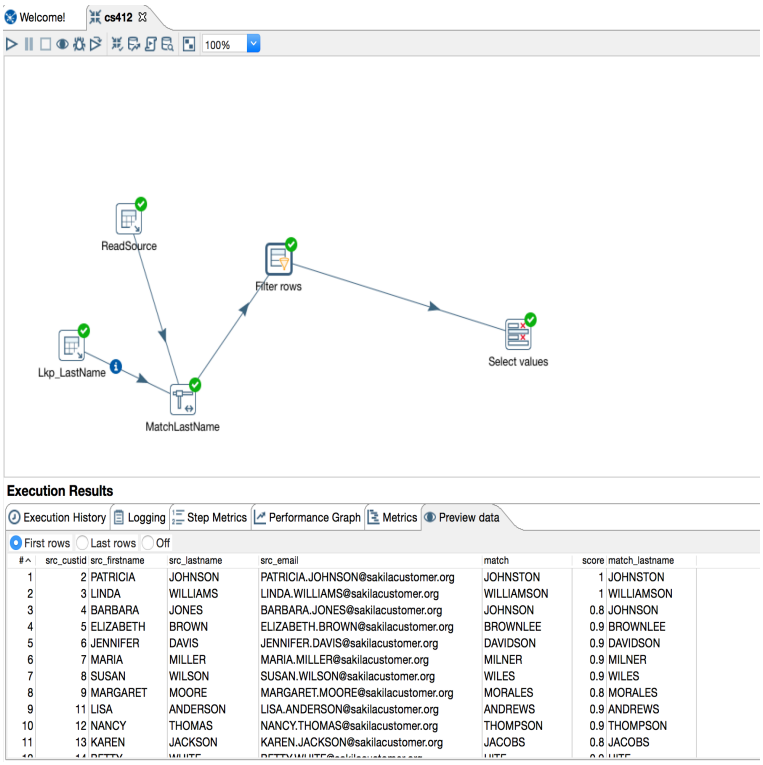


Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

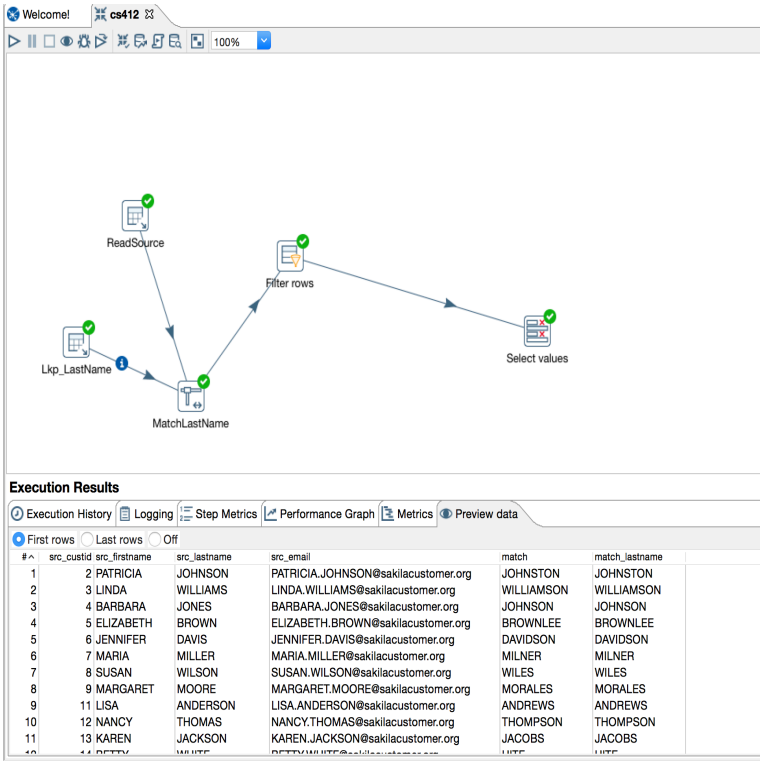
#	src_custid	src_firstname	src_lastname	src_email	match	score	match_lastname
1	1	MARY	SMITH	MARY.SMITH@sakilacustomer.org	SMITH	1	SMITH
2	2	PATRICIA	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSON	1	JOHNSON
3	3	LINDA	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMS	1	WILLIAMS
4	4	BARBARA	JONES	BARBARA.JONES@sakilacustomer.org	JONES	1	JONES
5	5	ELIZABETH	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWN	1	BROWN
6	6	JENNIFER	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIS	1	DAVIS
7	7	MARIA	MILLER	MARIA.MILLER@sakilacustomer.org	MILLER	1	MILLER
8	8	SUSAN	WILSON	SUSAN.WILSON@sakilacustomer.org	WILSON	1	WILSON
9	9	MARGARET	MOORE	MARGARET.MOORE@sakilacustomer.org	MOORE	1	MOORE
10	10	DOROTHY	TAYLOR	DOROTHY.TAYLOR@sakilacustomer.org	TAYLOR	1	TAYLOR
11	11	LISA	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDERSON	1	ANDERSON

Max = 1



#	src_custid	src_firstname	src_lastname	src_email	match	score	match_lastname
1	2	PATRICIA	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSTON	1	JOHNSTON
2	3	LINDA	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMSON	1	WILLIAMSON
3	4	BARBARA	JONES	BARBARA.JONES@sakilacustomer.org	JOHNSON	0.8	JOHNSON
4	5	ELIZABETH	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWNLEE	0.9	BROWNLEE
5	6	JENNIFER	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIDSON	0.9	DAVIDSON
6	7	MARIA	MILLER	MARIA.MILLER@sakilacustomer.org	MILNER	0.9	MILNER
7	8	SUSAN	WILSON	SUSAN.WILSON@sakilacustomer.org	WILES	0.9	WILES
8	9	MARGARET	MOORE	MARGARET.MOORE@sakilacustomer.org	MORALES	0.8	MORALES
9	11	LISA	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDREWS	0.9	ANDREWS
10	12	NANCY	THOMAS	NANCY.THOMAS@sakilacustomer.org	THOMPSON	0.9	THOMPSON
11	13	KAREN	JACKSON	KAREN.JACKSON@sakilacustomer.org	JACOBS	0.8	JACOBS

2.



#	src_custid	src_firstname	src_lastname	src_email	match	match_lastname
1	2	PATRICIA	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSTON	JOHNSTON
2	3	LINDA	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMSON	WILLIAMSON
3	4	BARBARA	JONES	BARBARA.JONES@sakilacustomer.org	JOHNSON	JOHNSON
4	5	ELIZABETH	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWNLEE	BROWNLEE
5	6	JENNIFER	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIDSON	DAVIDSON
6	7	MARIA	MILLER	MARIA.MILLER@sakilacustomer.org	MILNER	MILNER
7	8	SUSAN	WILSON	SUSAN.WILSON@sakilacustomer.org	WILES	WILES
8	9	MARGARET	MOORE	MARGARET.MOORE@sakilacustomer.org	MORALES	MORALES
9	11	LISA	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDREWS	ANDREWS
10	12	NANCY	THOMAS	NANCY.THOMAS@sakilacustomer.org	THOMPSON	THOMPSON
11	13	KAREN	JACKSON	KAREN.JACKSON@sakilacustomer.org	JACOBS	JACOBS

3.