1a) Q1 = 82, Median = 89, Q3 = 95
1b) Mean = 87.011
1c) Mode = 95
1d) The data is positively skewed. This is because the mean < median < mode. 87.011 < 89 < 95.

2a)

$$Jaccard_{Coeff} = \frac{q}{q+r+s} = \frac{21}{21+28+39} = \frac{21}{88} = .239$$

2b) A = (3, 1, 2) B = (-1, 0, 8)
2b1)

$$Euclidean_{Dist} = \sqrt{(|A_1 - B_1|^2 + |A_2 - B_2|^2 + |A_3 - B_3|^2)}$$

S

$$Euclidean_{Dist} = \sqrt{(|3 - (-1)|^2 + |1 - 0|^2 + |2 - 8|^2)}$$

$$Euclidean_{Dist} = \sqrt{(|4|^2 + |1|^2 + |-6|^2)}$$

$$Euclidean_{Dist} = \sqrt{(16 + 1 + 36)}$$

$$Euclidean_{Dist} = \sqrt{53} = 7.280$$

2b2)

$$Manhattan_{Dist} = |A_1 - B_1| + |A_2 - B_2| + |A_3 - B_3|$$

$$Manhattan_{Dist} = |3 - (-1)| + |1 - 0| + |2 - 8|$$

$$Manhattan_{Dist} = |4| + |1| + |-6|$$

$$Manhattan_{Dist} = 11$$

2b3)

$$Minkowski_{h->\infty} = max_f^p |A_{if} - B_{if}|$$

$$|A_1 - B_1| = |3 - (-1)| = 4$$

$$|A_2 - B_2| = |1 - (0)| = 1$$

$$|A_3 - B_3| = |2 - 8| = 6$$

$$Minkowski_{h->\infty} = |A_3 - B_3| = |2 - 8| = 6$$

2c) The Euclidean distance is a direct line between 2 points in space. The Manhattan distance is the distance between 2 points in space if the path traveled from A to B is taken at right angles. If we look at these two distances in a 2-d plane, the Manhattan distance will always be the 2 shorter sides of a

right triangle, whereas the Euclidean distance will be the hypotenuse. This is all assuming the points are not already on a 90 degree axis with each other, which would mean the Manhattan and Euclidean distances would be equal.

2d1) h = 2: 412.941
2d2) h = 3: 216.448

3a)
Before
Mean = 76.814
Variance = 171.396
After
Mean = 0
Variance = 1
3b)
Original Value = 90
Z-Score = 1.007

4a)

$$Mean = \frac{1}{n}\sum_{i}^{n} A_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Correlation Coefficient = .985. The 2 vectors in the data set are positively correlated, meaning as X changes, Y changes in the same direction with a similar magnitude.

4b) PCA will help to reduce the data size because the correlation coefficient is high. The higher the correlation coefficient, the more redundancy exists in a data set. Because there is so much redundancy, PCA will be effective in reducing the data set size.

4c)
Step 1: Zero mean X and Y using eq: **xZM(i) = x(i) – xMean for i in X**. Use same for Y.

```
0.5520   -1.4480    0.2520   -0.0880    1.1520    0.3520    0.0520   -0.9480   -0.4480    0.5720
0.6160   -1.3840    0.3160    0.1760    0.9160    0.4160   -0.0240   -0.9840   -0.4840    0.4360
```

Step 2: Covariance Matrix = $\frac{1}{xNum - 1} * xyM * xyM^T$

$$\begin{bmatrix} .595 & .556 \\ .556 & .537 \end{bmatrix}$$

4d) Calculations in script below:

2 Principal components because matrix is 2x10 (MxN) where the # of PC's are M.

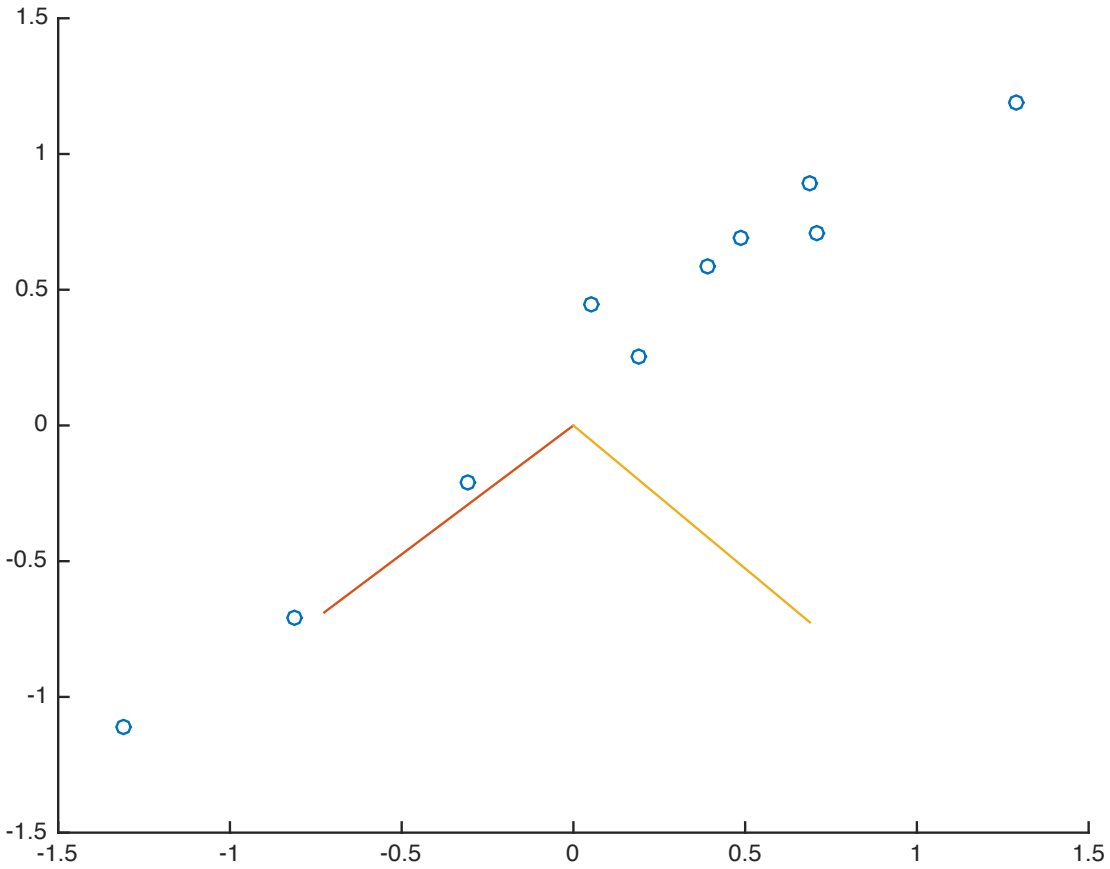
Principal Components:
0.6885 -0.7253 -> P1
-0.7253 -0.6885 -> P2

Most important Principal Component:
-0.7253 -0.6885 -> P2
This is because the covariance matrix looks like this…
0.0086 -0.0000
-0.0000 1.1229 -> Variance is much higher on P2 than P1 meaning more information is retained by using P2

Code:

```
x = [.69, -1.31, .39, .05, 1.29, .49, .19, -.81, -.31, .71];
y = [.89, -1.11, .59, .45, 1.19, .69, .25, -.71, -.21, .71];

xMean = mean(x);
yMean = mean(y);
xNum = length(x);

% Make X and Y zero mean
for i = 1:numel(x)
    xM(i) = (x(i) - xMean); % Zero Mean
    yM(i) = (y(i) - yMean); % Zero Mean
end

% Combine them together to form an 2x10 (MxN) matrix
xyM = [xM; yM]
% Find the covariance matrix
xyCov = (1/(xNum-1))*xyM*xyM'

% Find the Eigen vectors for the covariance matrix
[e_vec,e_val] = eig(xyCov);
e_vec = e_vec
% Find the new principal component matrix
eY = e_vec * xyM
% Ensure Covariance of eY is diagonal matrix
cY1 = (1/(xNum-1))*eY*eY'
% Variance of the second vector is greater so that will be the principal
% component
```

4e)



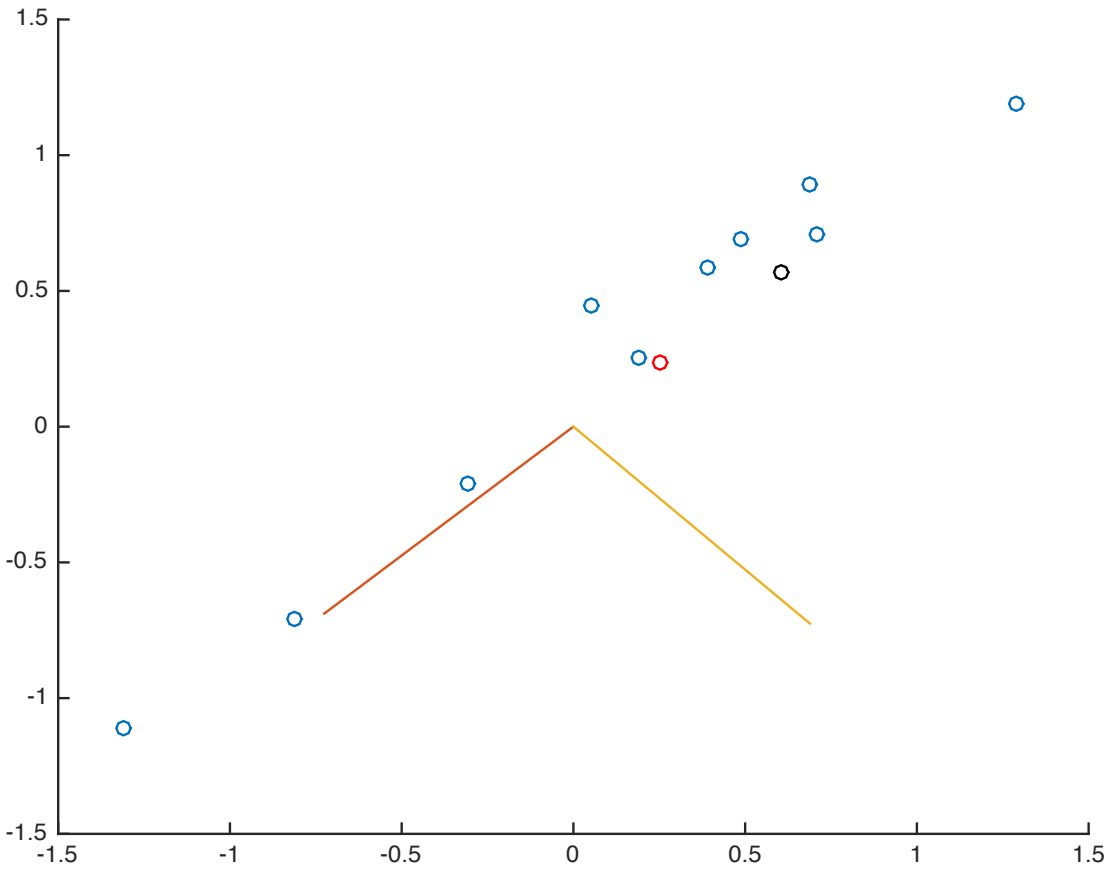
Red line = Largest principal component direction
Orange line = Smaller principal component direction

4f)
Ap = P1 * A = -.346
Bp = P1 * B = -.830

For projecting A onto the graph I used

$$\mathbf{a}_1 = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\mathbf{b} = \frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}}\mathbf{b}.$$

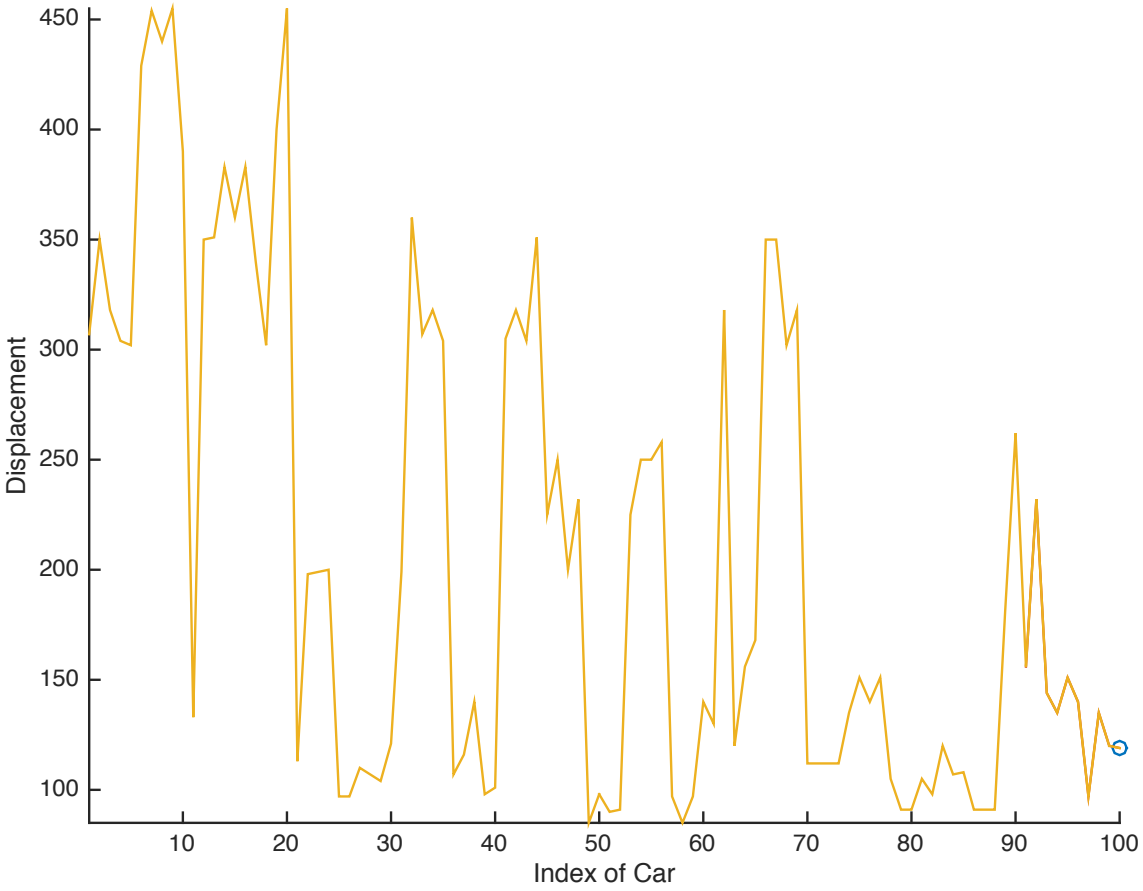
Where a = A (Original data point), b = P1 (Primary component of PCA)



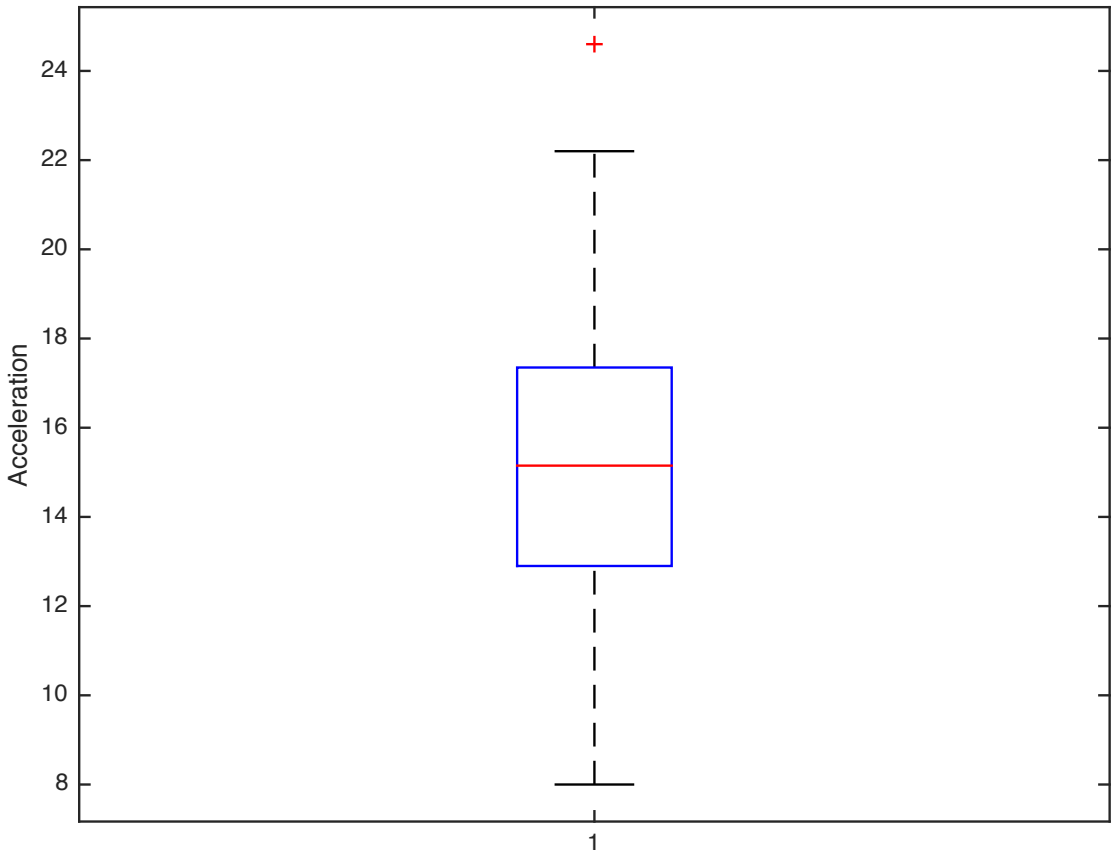
Red circle denotes projection of A
Black circle denotes projection of B

Mini MP 1)

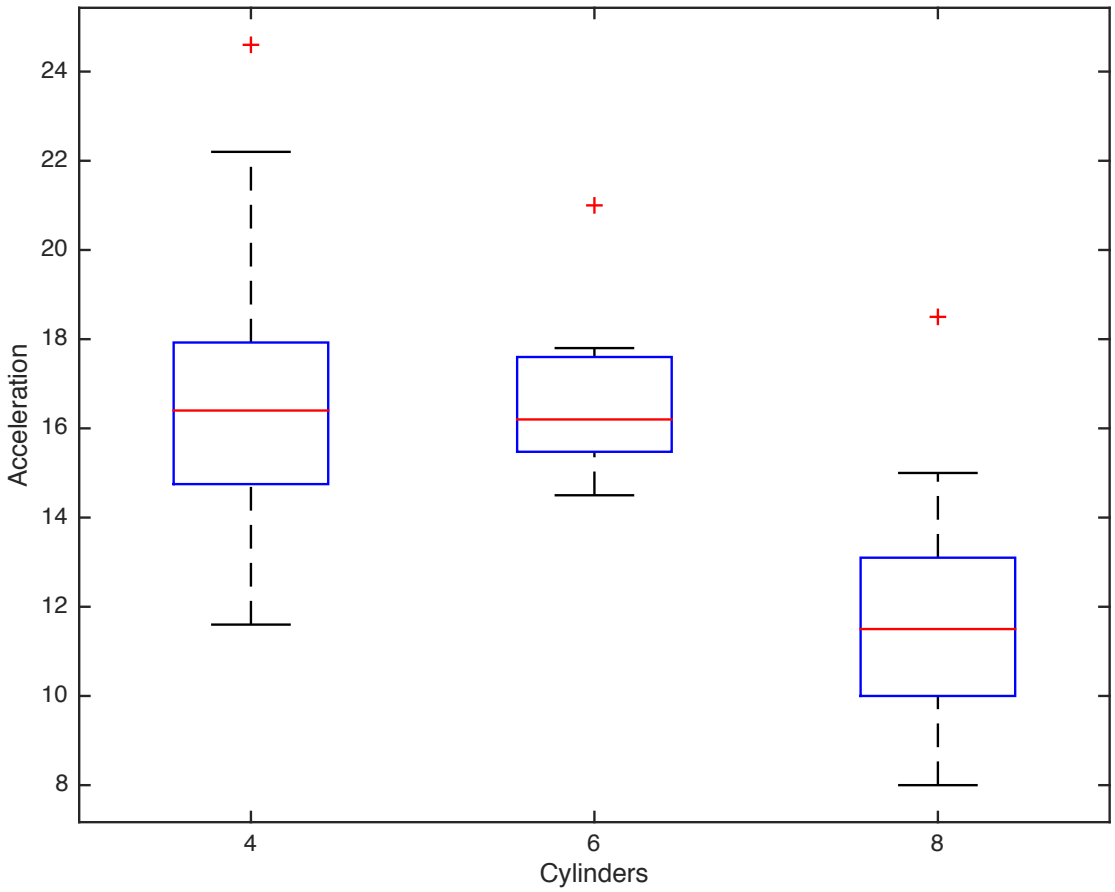
```
load carsmall;
X = [MPG,Acceleration,Displacement,Weight,Horsepower];
varName = {'MPG','Acceleration','Displacement','Weight','Horsepower'};

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



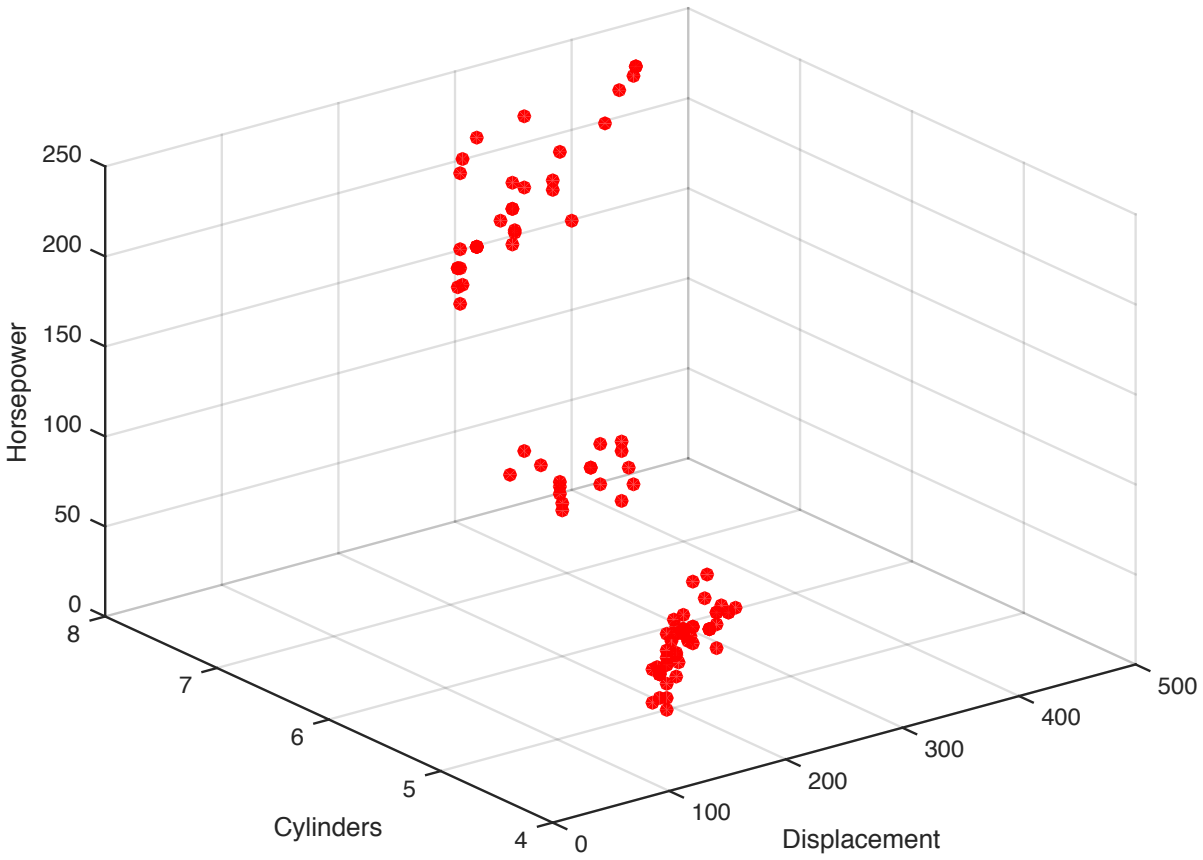
```
boxplot(Acceleration);
ylabel('Acceleration');
```



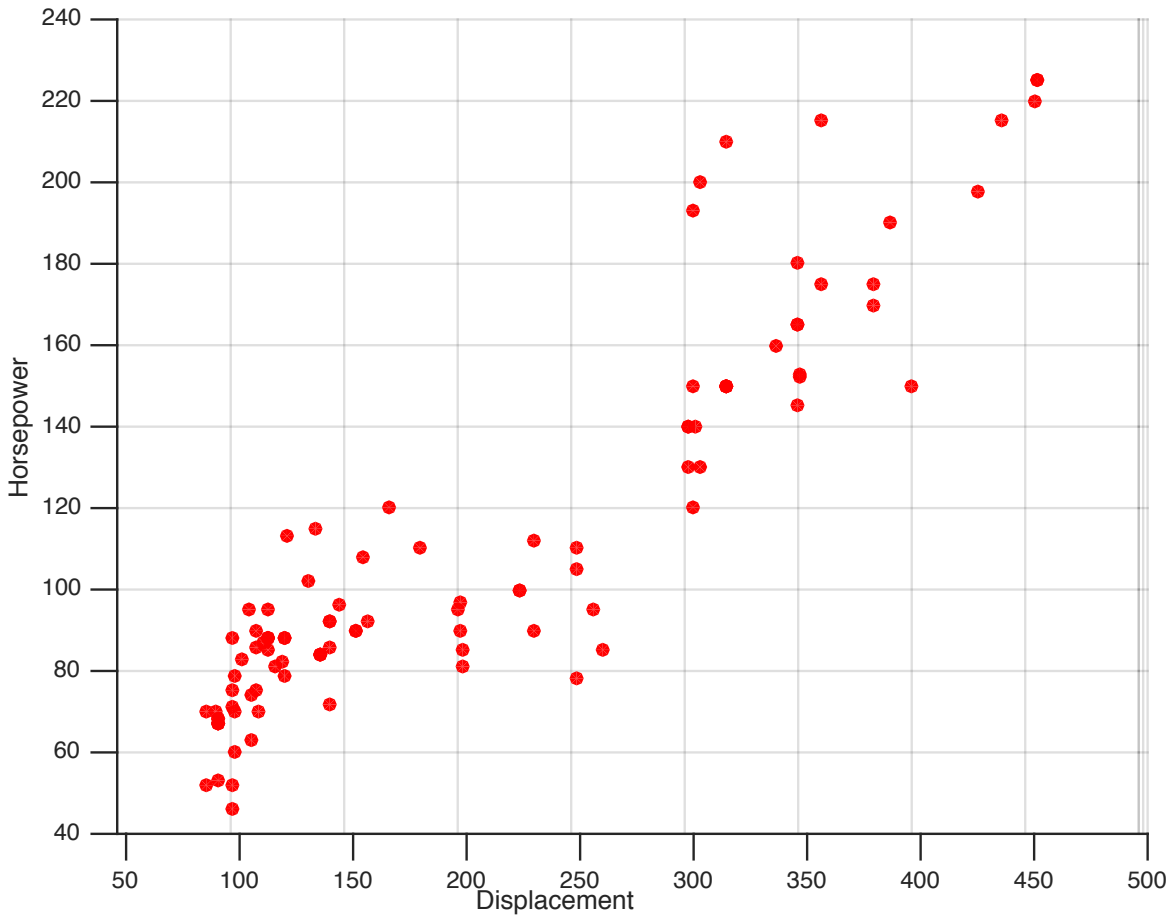
```
boxplot(Acceleration,Cylinders);
xlabel('Cylinders');
ylabel('Acceleration');
```

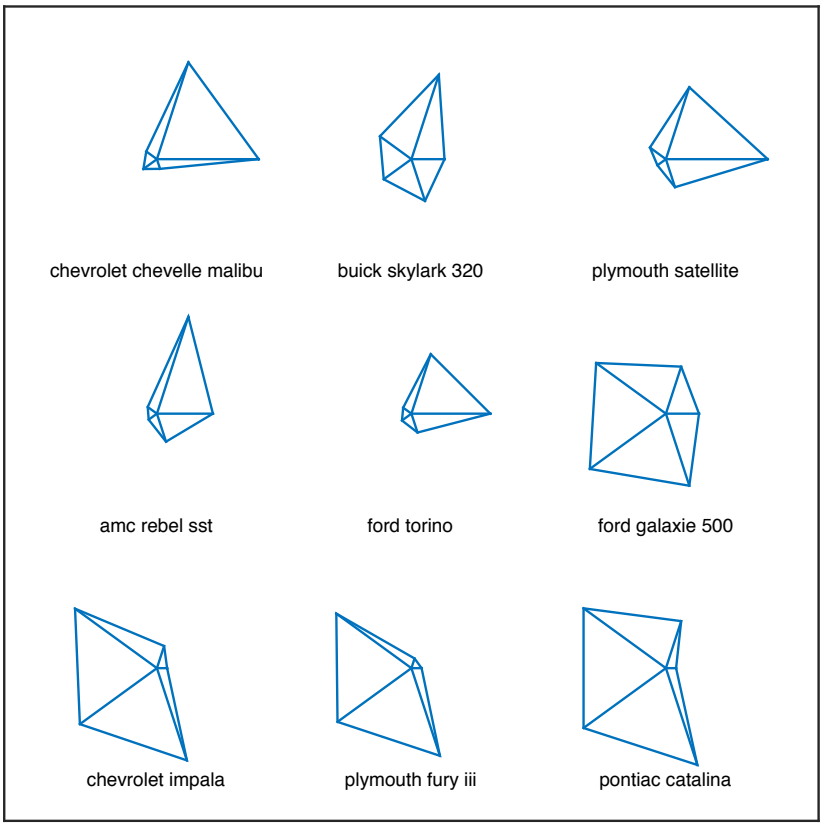


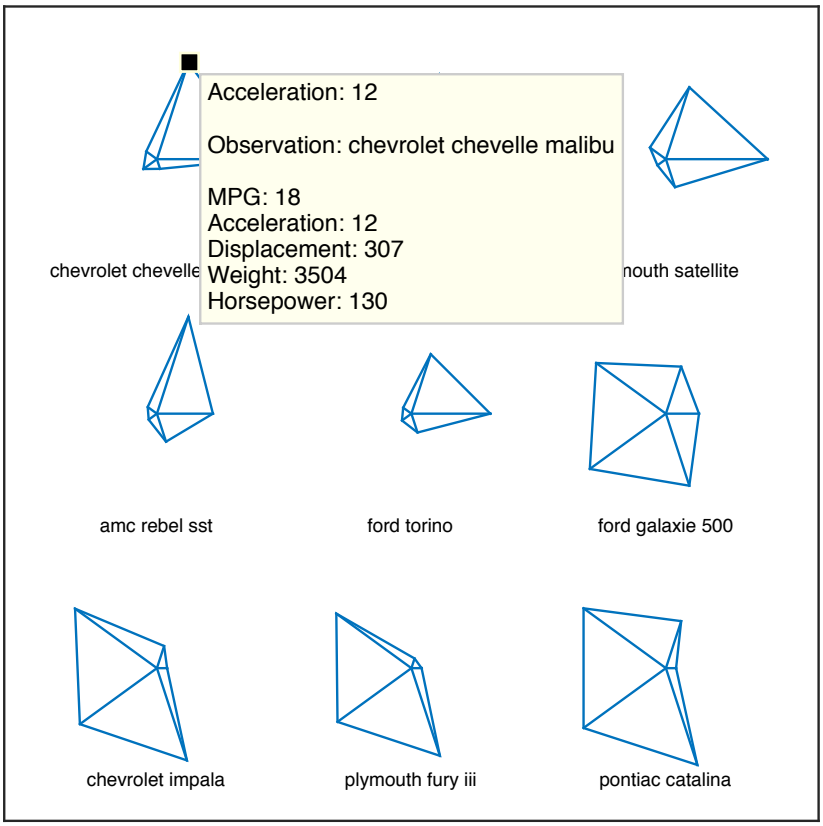
```
scatter3(Displacement,Cylinders,Horsepower,'filled','r');
xlabel('Displacement');
ylabel('Cylinders');
zlabel('Horsepower');
```



There is a positive correlation between displacement and horsepower. As displacement increases, horsepower also increases.


Horsepower
240
220
200
180
160
140
120
100
80
60
40
50
100
150
200
250
300
350
400
450
500
Displacement

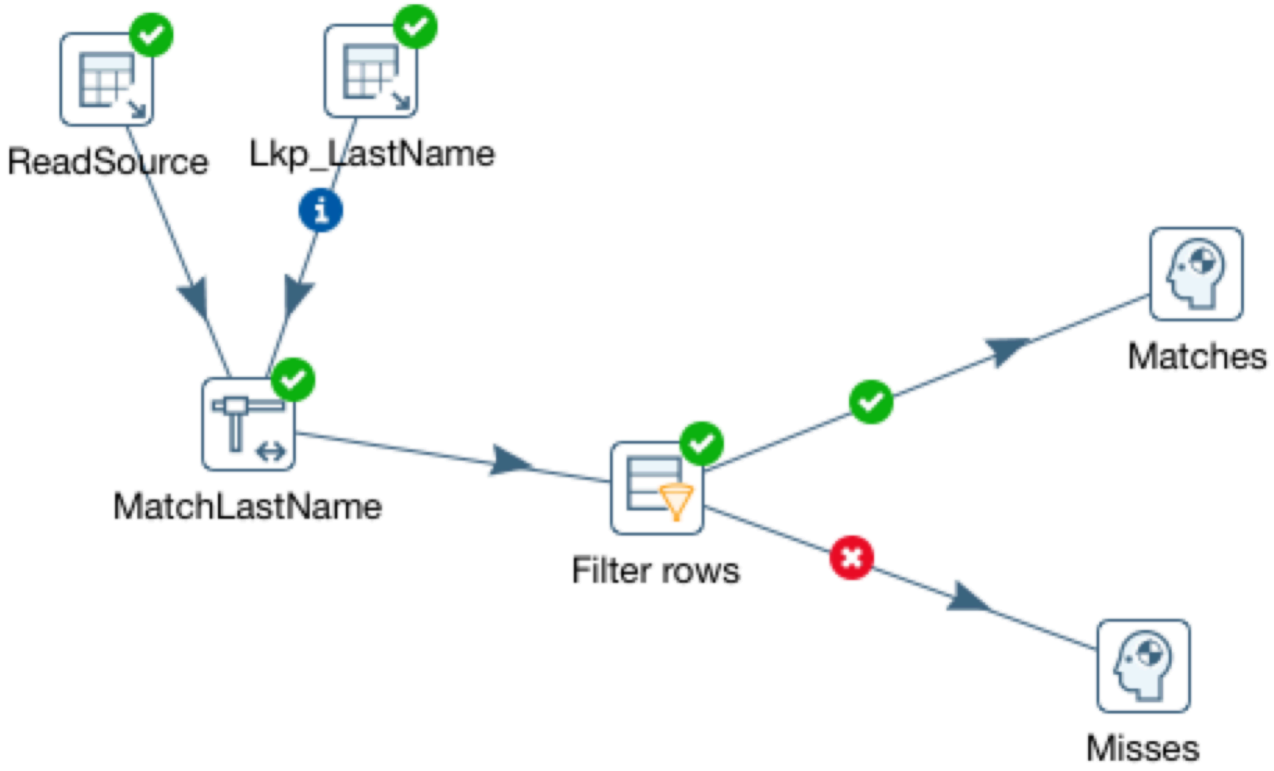

chevrolet chevelle malibu
buick skylark 320
plymouth satellite
amc rebel sst
ford torino
ford galaxie 500
chevrolet impala
plymouth fury iii
pontiac catalina

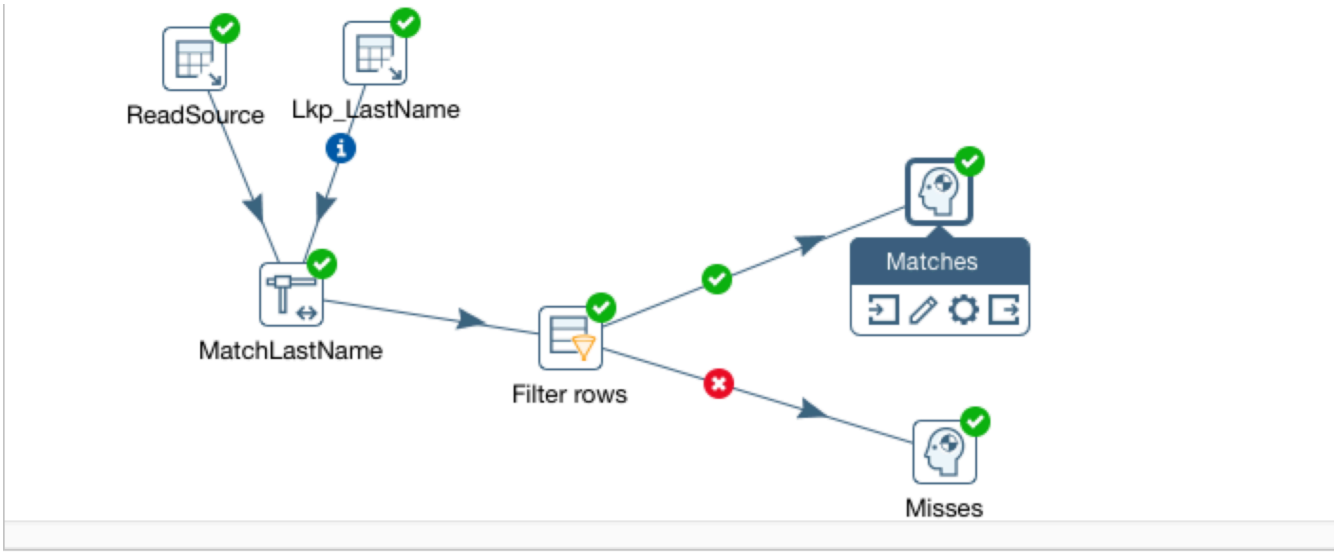

Acceleration: 12
Observation: chevrolet chevelle malibu
MPG: 18
Acceleration: 12
Displacement: 307
Weight: 3504
Horsepower: 130
chevrolet chevelle
nouth satellite
amc rebel sst
ford torino
ford galaxie 500
chevrolet impala
plymouth fury iii
pontiac catalina

Mini MP 2)

1) The max value must be less than 1 so that it does not only find perfect matches between the 2 data sets. I'm not entirely sure why that is the goal here. I suppose by keeping the max value to less than 1, you can find last names that are as close to the same as possible without being identical.

2)



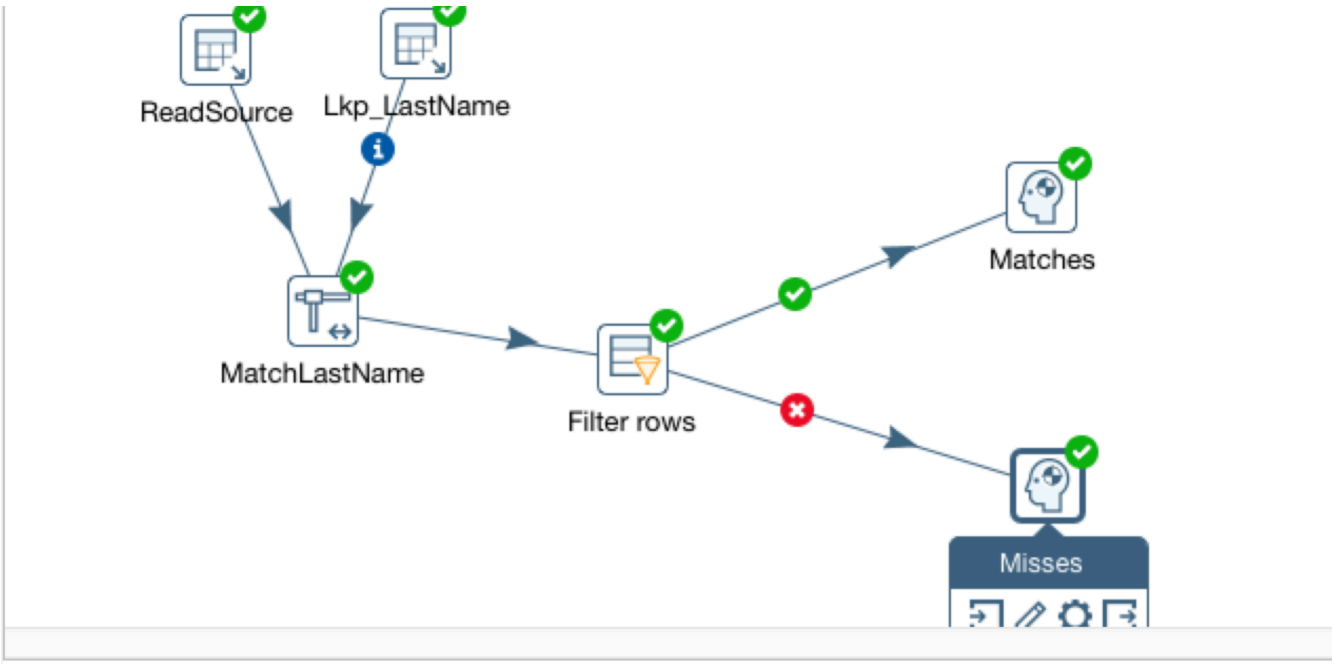


Execution Results

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

First rows | Last rows | Off

#	src_custid	src_lastname	src_email	match	score
1	2	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSTON	1
2	3	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMSON	1
3	4	JONES	BARBARA.JONES@sakilacustomer.org	JOHNSON	0.8
4	5	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWNLEE	0.9
5	6	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIDSON	0.9
6	7	MILLER	MARIA.MILLER@sakilacustomer.org	MILNER	0.9
7	8	WILSON	SUSAN.WILSON@sakilacustomer.org	WILES	0.9
8	9	MOORE	MARGARET.MOORE@sakilacustomer.org	MORALES	0.8
9	11	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDREWS	0.9
10	12	THOMAS	NANCY.THOMAS@sakilacustomer.org	THOMPSON	0.9
11	13	JACKSON	KAREN.JACKSON@sakilacustomer.org	JACOBS	0.8
12	14	WHITE	BETTY.WHITE@sakilacustomer.org	HITE	0.9
13	15	HARRIS	HELEN.HARRIS@sakilacustomer.org	HARRISON	0.9
14	16	MARTIN	SANDRA.MARTIN@sakilacustomer.org	MARTINO	1
15	17	THOMPSON	DONNA.THOMPSON@sakilacustomer.org	THOMAS	0.9
16	18	GARCIA	CAROL.GARCIA@sakilacustomer.org	GARZA	0.9
17	19	MARTINEZ	RUTH.MARTINEZ@sakilacustomer.org	MARTIN	0.9
18	20	ROBINSON	SHARON.ROBINSON@sakilacustomer.org	ROBINS	0.9



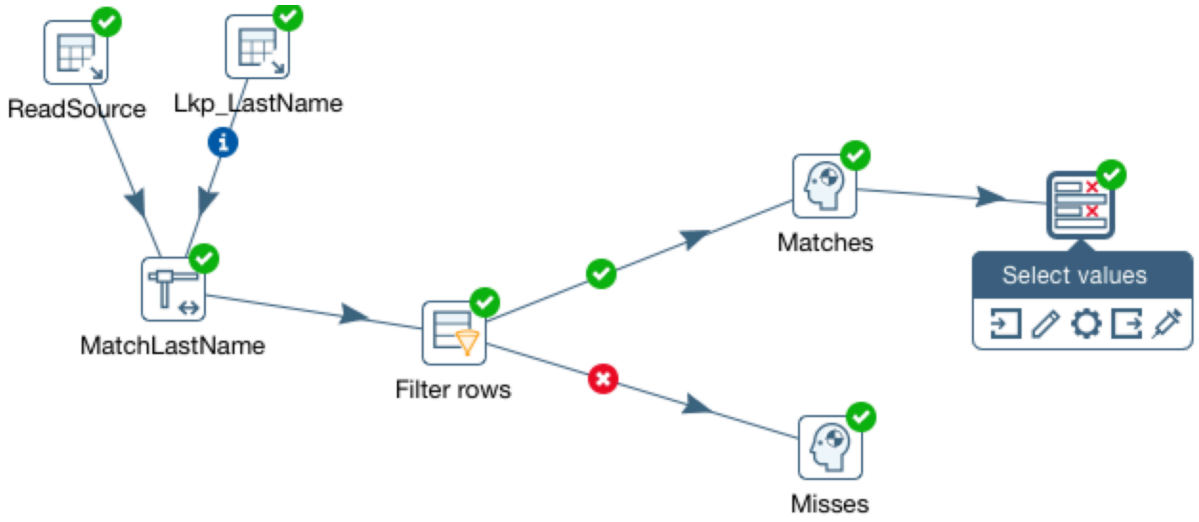
Execution Results

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

First rows | Last rows | Off

#	src_custid	src_lastname	src_email	match	score
1	1	SMITH	MARY.SMITH@sakilacustomer.org	<null>	<null>
2	10	TAYLOR	DOROTHY.TAYLOR@sakilacustomer.org	<null>	<null>
3	24	LEE	KIMBERLY.LEE@sakilacustomer.org	<null>	<null>
4	28	YOUNG	CYNTHIA.YOUNG@sakilacustomer.org	<null>	<null>
5	31	WRIGHT	BRENDA.WRIGHT@sakilacustomer.org	<null>	<null>
6	41	MITCHELL	STEPHANIE.MITCHELL@sakilacustomer.org	<null>	<null>
7	46	CAMPBELL	CATHERINE.CAMPBELL@sakilacustomer.org	<null>	<null>
8	71	JAMES	KATHY.JAMES@sakilacustomer.org	<null>	<null>
9	90	WASHINGTON	RUBY.WASHINGTON@sakilacustomer.org	<null>	<null>
10	96	ALEXANDER	DIANA.ALEXANDER@sakilacustomer.org	<null>	<null>
11	97	RUSSELL	ANNIE.RUSSELL@sakilacustomer.org	<null>	<null>
12	98	GRIFFIN	LILLIAN.GRIFFIN@sakilacustomer.org	<null>	<null>
13	99	DIAZ	EMILY.DIAZ@sakilacustomer.org	<null>	<null>
14	105	SULLIVAN	DAWN.SULLIVAN@sakilacustomer.org	<null>	<null>
15	116	GIBSON	VICTORIA.GIBSON@sakilacustomer.org	<null>	<null>
16	117	MCDONALD	EDITH.MCDONALD@sakilacustomer.org	<null>	<null>
17	120	ORTIZ	SYLVIA.ORTIZ@sakilacustomer.org	<null>	<null>
18	125	WEBB	ETHEL.WEBB@sakilacustomer.org	<null>	<null>
19	128	TUCKER	MARJORIE.TUCKER@sakilacustomer.org	<null>	<null>
20	131	HICKS	MONICA.HICKS@sakilacustomer.org	<null>	<null>
21	139	DIXON	AMBER.DIXON@sakilacustomer.org	<null>	<null>
22	151	PALMER	MEGAN.PALMER@sakilacustomer.org	<null>	<null>

3)



Execution Results

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

First rows | Last rows | Off

#	src_custid	src_lastname	src_email	match
1	2	JOHNSON	PATRICIA.JOHNSON@sakilacustomer.org	JOHNSTON
2	3	WILLIAMS	LINDA.WILLIAMS@sakilacustomer.org	WILLIAMSON
3	4	JONES	BARBARA.JONES@sakilacustomer.org	JOHNSON
4	5	BROWN	ELIZABETH.BROWN@sakilacustomer.org	BROWNLEE
5	6	DAVIS	JENNIFER.DAVIS@sakilacustomer.org	DAVIDSON
6	7	MILLER	MARIA.MILLER@sakilacustomer.org	MILNER
7	8	WILSON	SUSAN.WILSON@sakilacustomer.org	WILES
8	9	MOORE	MARGARET.MOORE@sakilacustomer.org	MORRELL
9	11	ANDERSON	LISA.ANDERSON@sakilacustomer.org	ANDREWS
10	12	THOMAS	NANCY.THOMAS@sakilacustomer.org	THOMPSON
11	13	JACKSON	KAREN.JACKSON@sakilacustomer.org	JACOBS
12	14	WHITE	BETTY.WHITE@sakilacustomer.org	HITE
13	15	HARRIS	HELEN.HARRIS@sakilacustomer.org	HARRISON
14	16	MARTIN	SANDRA.MARTIN@sakilacustomer.org	MARTINO
15	17	THOMPSON	DONNA.THOMPSON@sakilacustomer.org	THOMAS
16	18	GARCIA	CAROL.GARCIA@sakilacustomer.org	GARZA
17	19	MARTINEZ	RUTH.MARTINEZ@sakilacustomer.org	MARTIN
18	20	ROBINSON	SHARON.ROBINSON@sakilacustomer.org	ROBINS
19	21	CLARK	MICHELLE.CLARK@sakilacustomer.org	CLARY
20	22	RODRIGUEZ	LAURA.RODRIGUEZ@sakilacustomer.org	RODRIQUEZ
21	23	LEWIS	SARAH.LEWIS@sakilacustomer.org	WILES