

Question 1

a.

First quantile Q1: 82

Median: 89

Third quantile Q3: 95

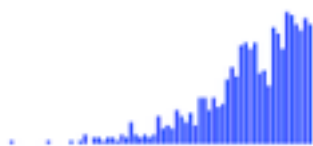
b.

Mean: 87.011

c.

Mode: 95

d.



The data is negative skewed. The mean is smaller than the mode. I printed out a graph, and it also shows that the data is negative skewed.

Question 2

a.

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

$$21/(28+39+21) = 21/88$$

b.

Euclidean distance

$$d(i,j)= \text{sqrt}(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)$$

$$\text{sqrt}((3 - -1)^2 + (1 - 0)^2 + (2 - 8)^2) = 7.28$$

Manhattan distance

$$d(i,j)= |x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

$$|3 - -1| + |1 - 0| + |2 - 8| = 11$$

Minkowski distance where $h = \infty$.

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

$$|3 - -1| = 4$$

$$|1 - 0| = 0$$

$$|2 - 8| = 6$$

$$\max(4,0,6) = 6$$

c.

Euclidean distance calculates the distance between the two point. Manhattan distance calculate the total of distance in each axis. Therefor, when two points are on the same line, they have the same Euclidean distance and Manhattan distance. Otherwise, Manhattan distance is always bigger that Euclidean distance.

d.

$$h2 = 412.941$$

$$h3 = 216.448$$

Question 3

a.

Original Mean: 76.81375

Original Empirical Variance: 171.395805694619

$$\frac{\sum_i (x_i - \text{mean})^2}{n - 1}$$

Mean: 0

Original Empirical Variance: 1

b.

$$Z = \frac{x - \mu}{\sigma}$$

For 90, Z-score :1.00721273991535

Question 4

a.

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

The script used to calculate is attached.

R is: 0.9847

The distribution is closed to a linear function with gradient approximately 0.9847.

b.

I guess it will help reduce the data size. Principal Component Analysis is desirable for reducing the dimensionality of the data, while keeping the data that contribute the most to the standard deviation. It retrieve a portion of the samples as principal components of the data. The number of principal components is usually less than the number of original variables.

c.

Process the data to turn the mean to zero.

Covariance Matrix

```
1/10 * mat * mat.transpose =
```

```
[0.5353  0.5008]  
[0.5008  0.4831]
```

d.

```
1/10 * newMat * newMat.transpose
```

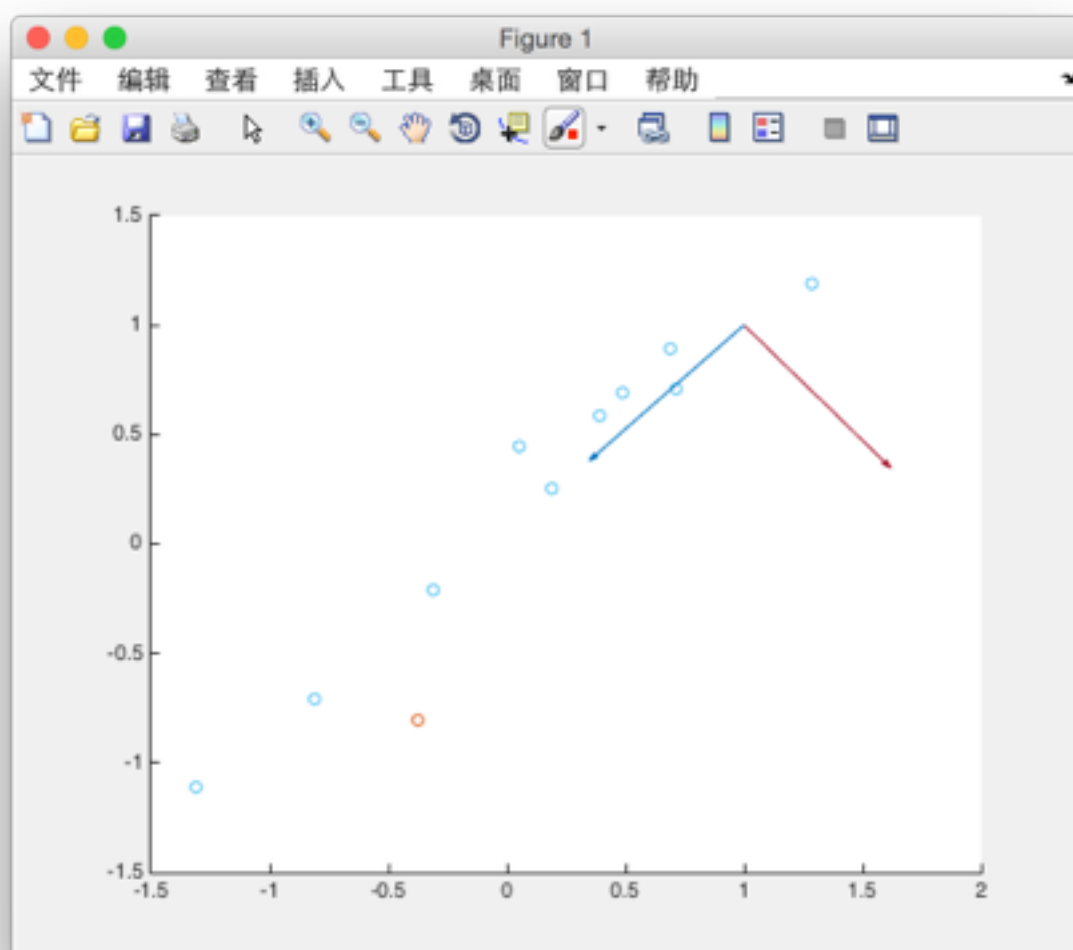
```
[V,D] = eig(cov(A,B)) get  
[ 0.6885 -0.7253]  
[-0.7253 -0.6885]
```

The principal components are $\begin{bmatrix} 0.6885 & -0.7253 \\ -0.7253 & -0.6885 \end{bmatrix}$

The eigenValues are 0.0078 and 1.0107

There are two components. The component that is significant bigger is the most important dimension. the second dimension is the first principal value.

e.



f.

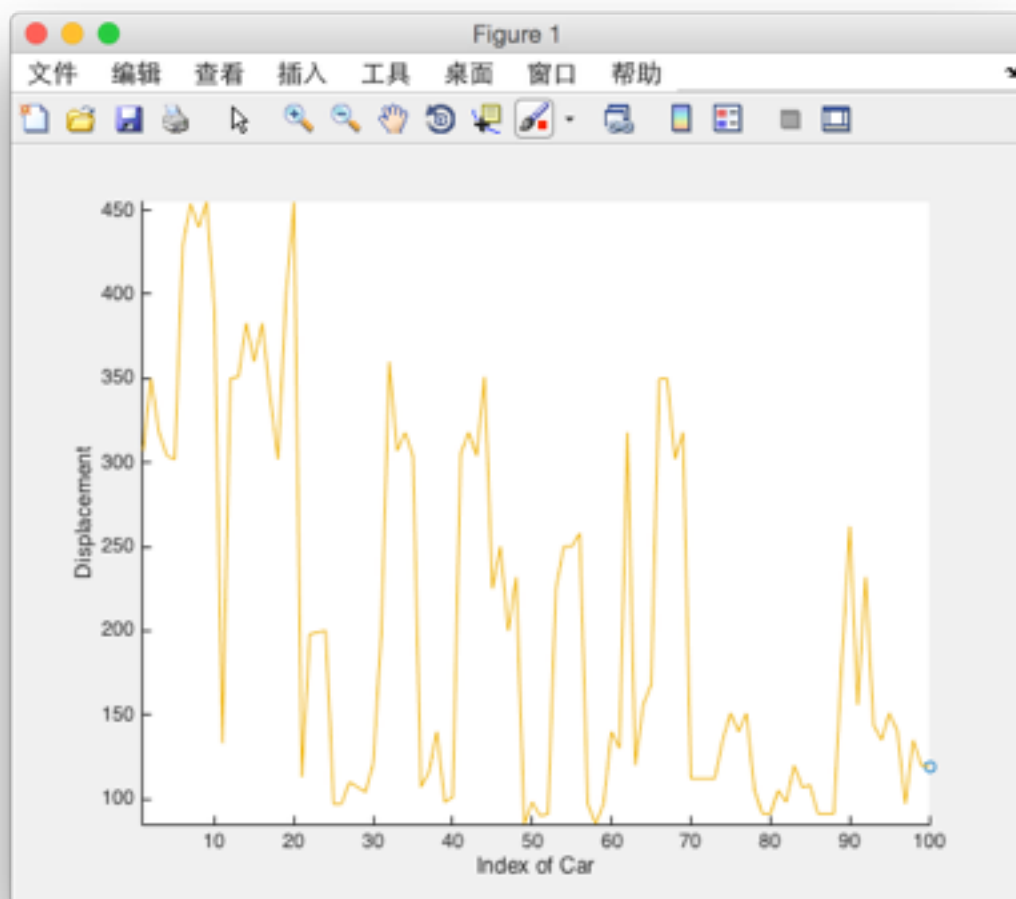
$$C = [-0.7253 \ -0.6885];$$

$$D = \begin{bmatrix} 0.05 & 0.49 \\ 0.45 & 0.69 \end{bmatrix}$$

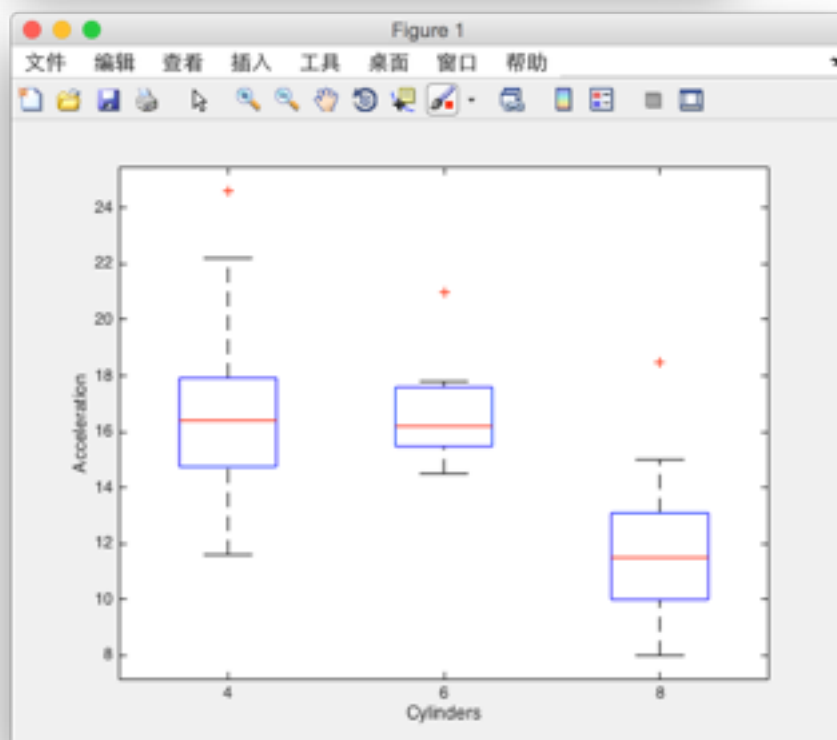
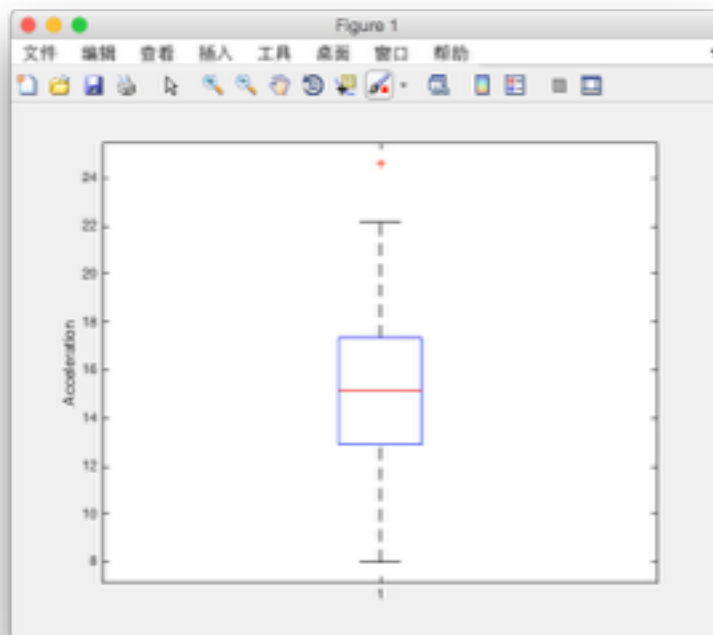
$$C * D = \begin{bmatrix} -0.3736 \\ -0.8014 \end{bmatrix}$$

Mini Machine Problem 1

2.



3.

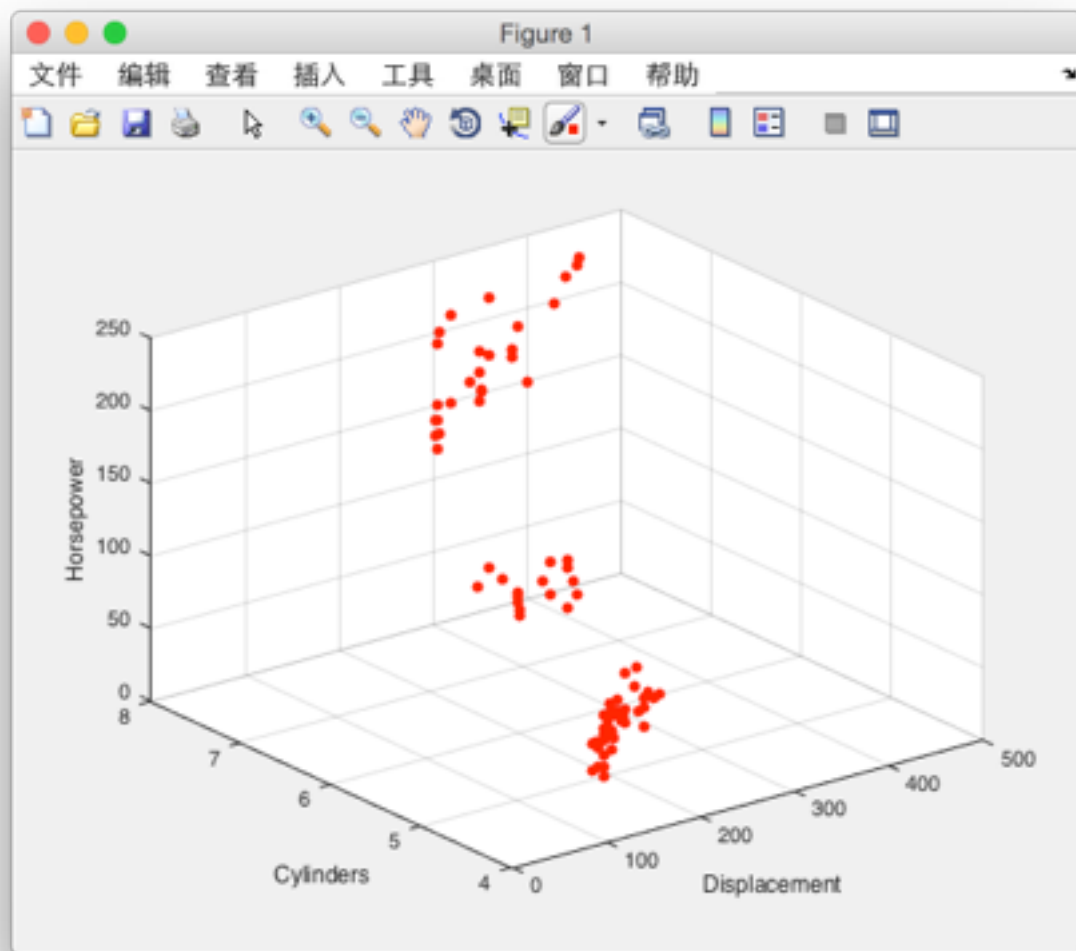


Code:

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

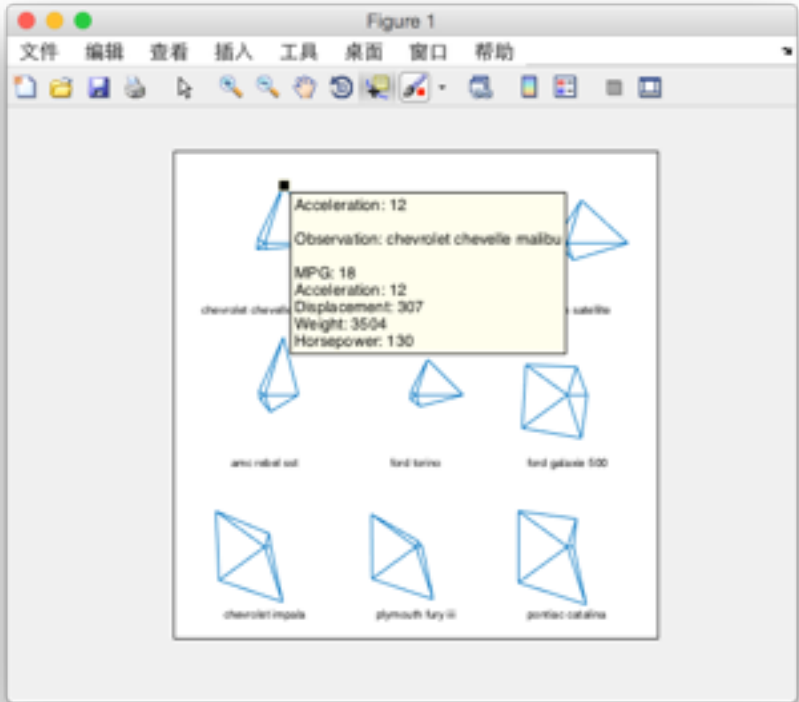
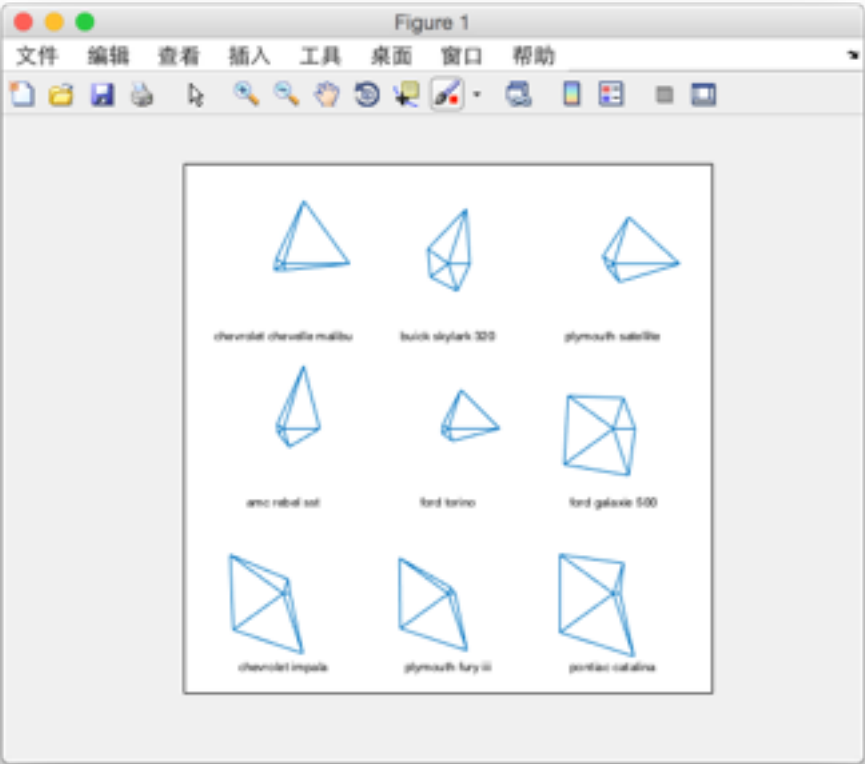
boxplot(Acceleration,Cylinders)
ylabel('Acceleration')
xlabel('Cylinders')
```

4.



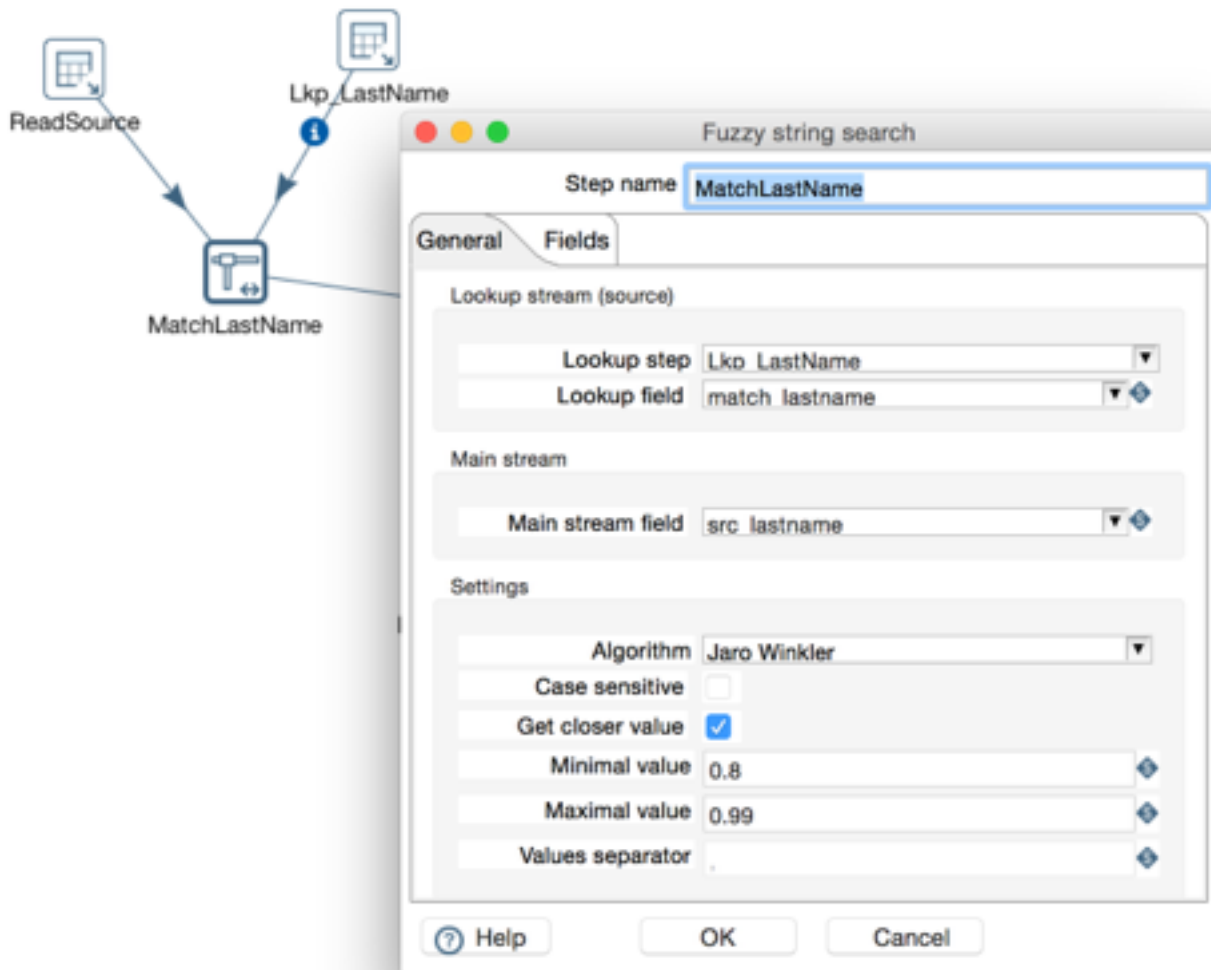
Horsepower and Displacement also have a negative correlation. They make sense because as the Displacement value increases, the Horsepower value decreases.

5.



Mini Machine Problem 2

1.



If I set the max value to 1.00, all the result will have the value 1. Therefore, we should set the distance to 0.99.

2.

The screenshot shows the SAP Data Integration Studio interface. The left pane displays the 'Steps' catalog with categories like Input, Output, Transform, and Flow. The main workspace shows a data flow diagram with steps: ReadSource, Upd_LastName, MatchLastName, and Select. The 'Execution Results' pane at the bottom shows a table with columns: #, src_suid, src_lastname, src_email, match, and score. The table contains 11 rows of data.

#	src_suid	src_lastname	src_email	match	score
1		SMITH	MARY.SMITH@skilacustomer.org	<null>	<null>
2		JOHNSON	PATRICIA.JOHNSON@skilacustomer.org	JOHNSTON	1
3		WILLIAMS	LINDA.WILLIAMS@skilacustomer.org	WILLIAMSON	1
4		TAYLOR	DOROTHY.TAYLOR@skilacustomer.org	<null>	<null>
5		HARRIS	HELEN.HARRIS@skilacustomer.org	HARRISON	1
6		MARTIN	SANDRA.MARTIN@skilacustomer.org	MARTINO	1
7		MARTINEZ	RUTH.MARTINEZ@skilacustomer.org	MARTIN	1
8		ROBINSON	SHARON.ROBINSON@skilacustomer.org	ROBINS	1
9		RODRIGUEZ	LAURA.RODRIGUEZ@skilacustomer.org	RODRIQUEZ	1
10		LEE	KIMBERLY.LEE@skilacustomer.org	<null>	<null>
11		YOUNG	CYNTHIA.YOUNG@skilacustomer.org	<null>	<null>

3.

The screenshot shows the SAP Data Integration Studio interface. The left pane displays the 'Steps' catalog. The main workspace shows a data flow diagram with steps: ReadSource, Upd_LastName, MatchLastName, and Select values. The 'Execution Results' pane at the bottom shows a table with columns: #, src_suid, src_lastname, src_email, match, and score. The table contains 11 rows of data.

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

