Question 1

a.

First quantile Q1: 82
Median: 89
Third quantile Q3: 95

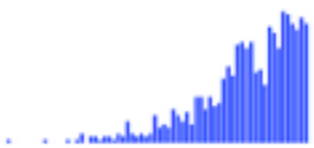
b.

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

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

sqrt((3 - -1)^2 + (1 - 0)^2 + (2 - 8)^2) = 7.28

Manhattan distance

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

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

Minkowski distance where h = ∞.

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

|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

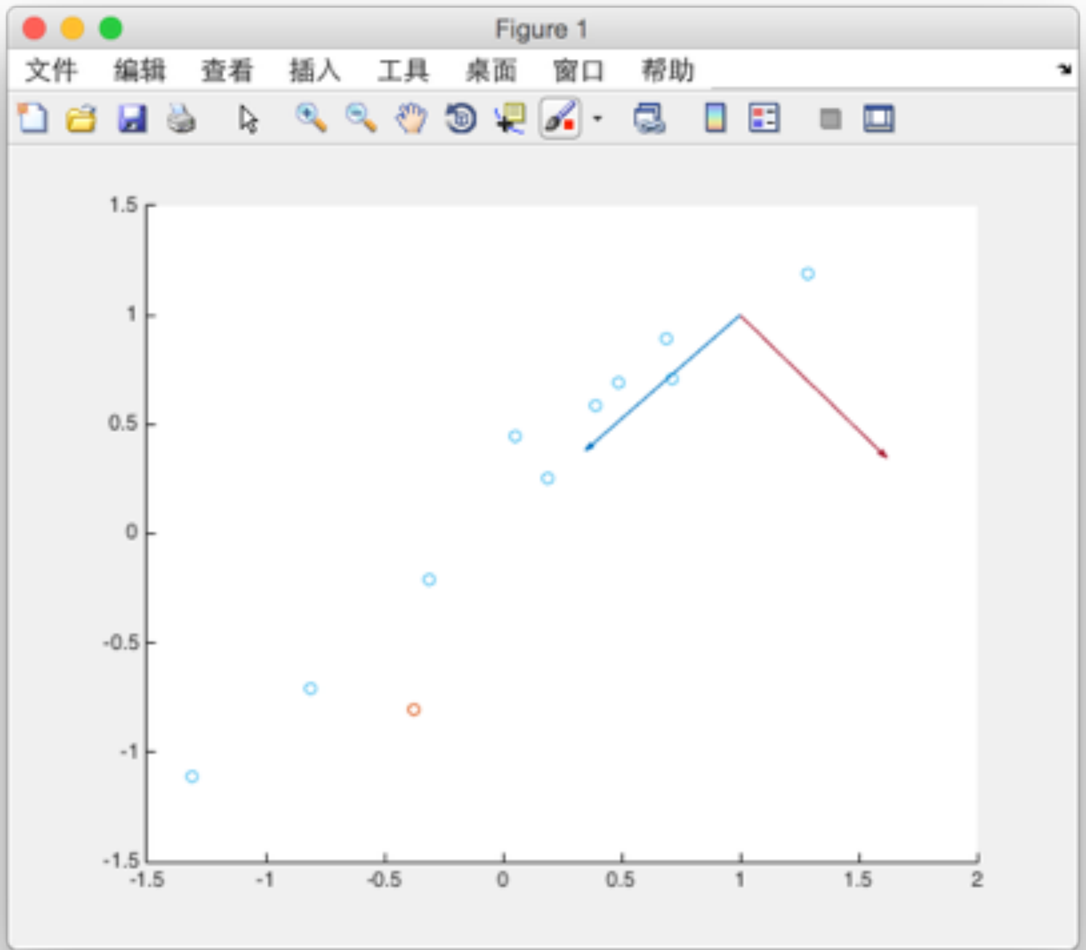
[V,D] = eig(cov(A,B)) get
[0.6885   -0.7253]
[-0.7253   -0.6885]

The principal components are   [0.6885   -0.7253]
[-0.7253   -0.6885]

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.

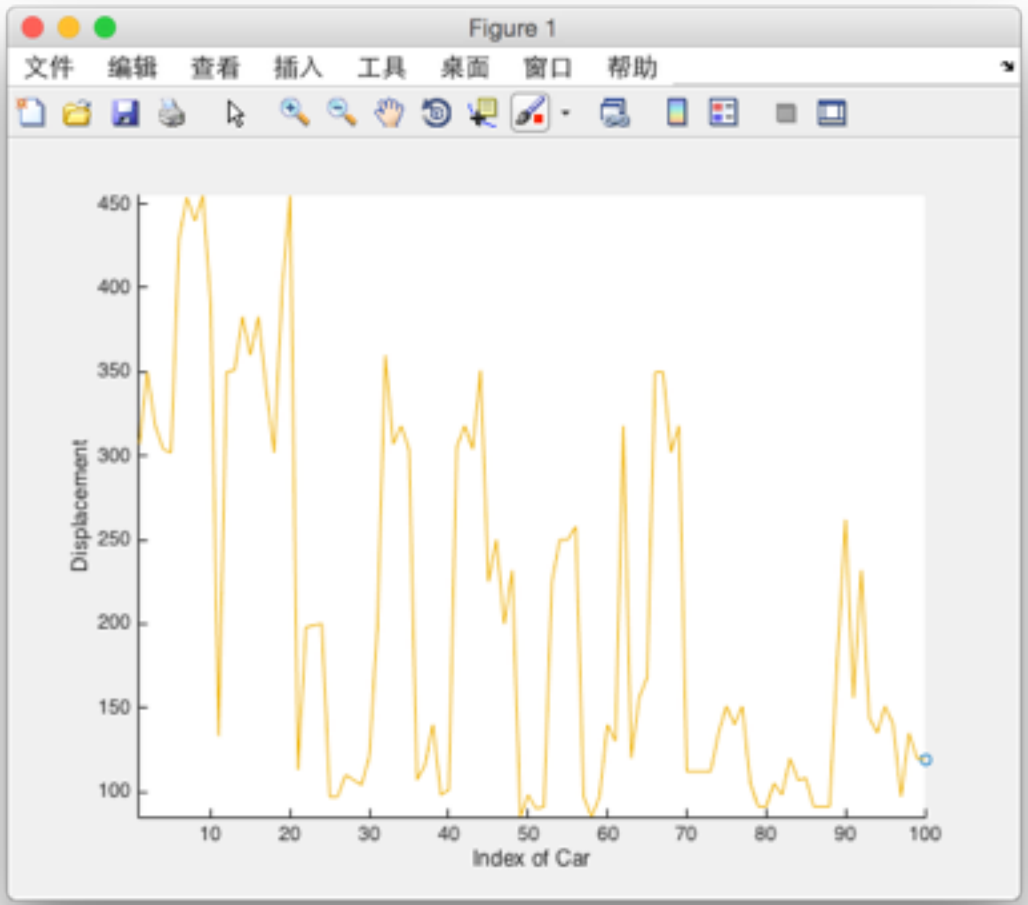
C = [-0.7253 -0.6885];

D = [0.05 0.49]
[0.45 0.69]

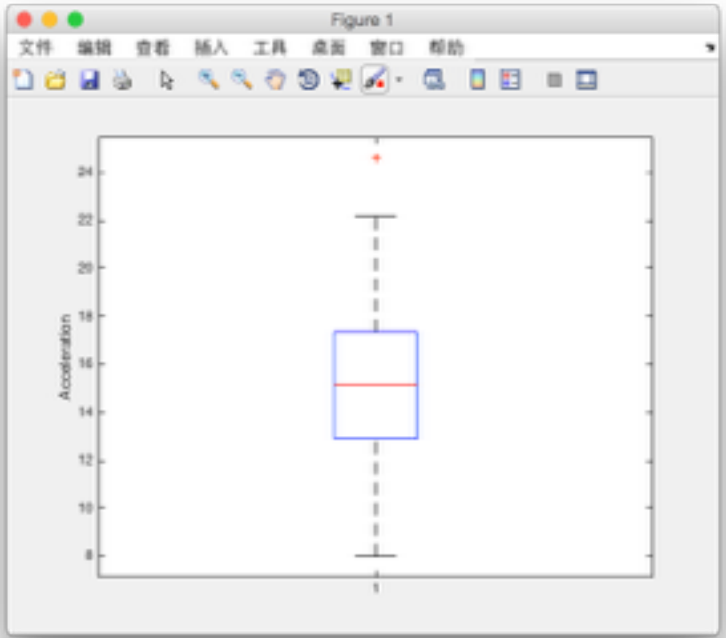
C * D = [-0.3736
-0.8014]

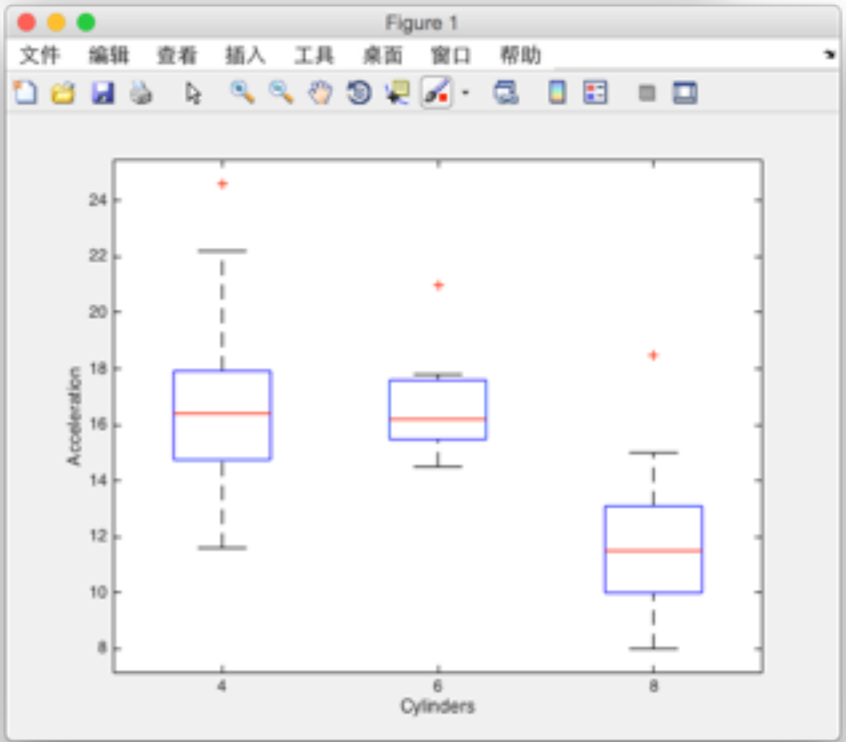
Mini Machine Problem 1

2.



3.



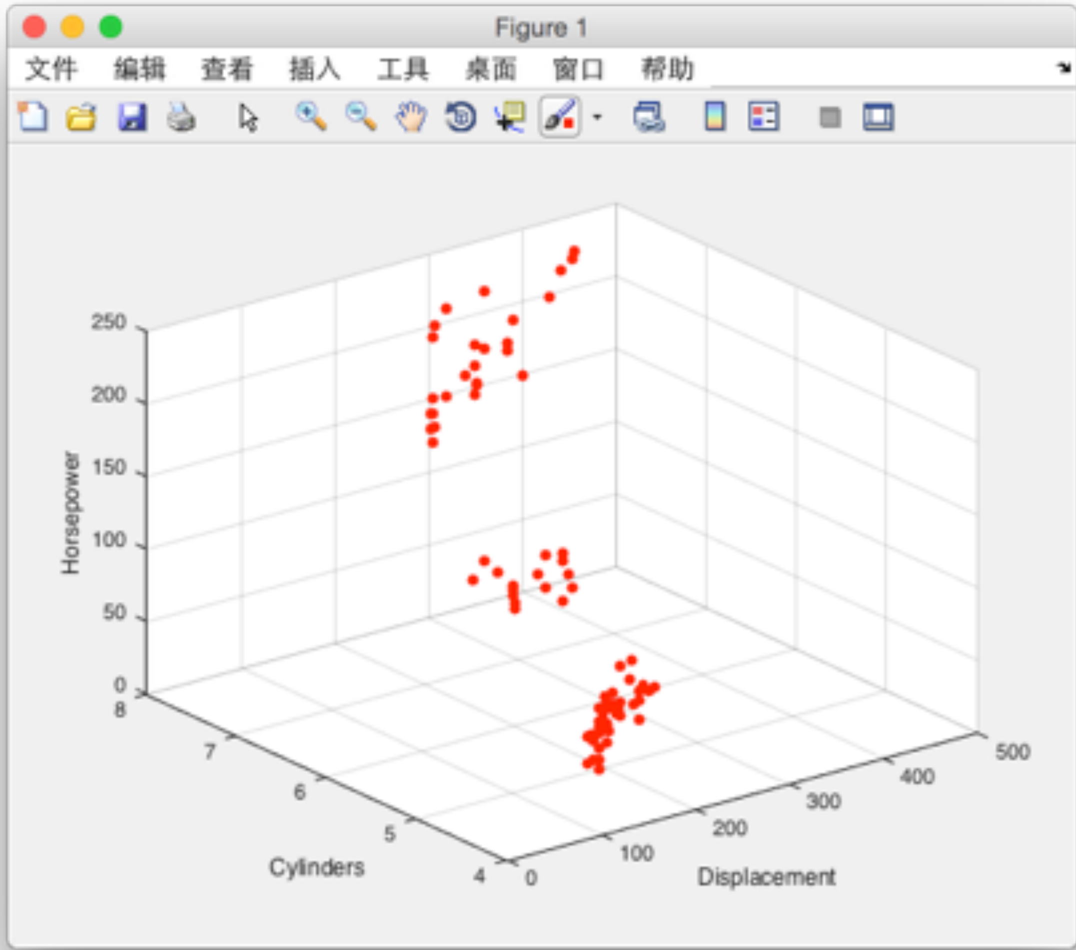


Code:

```
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.

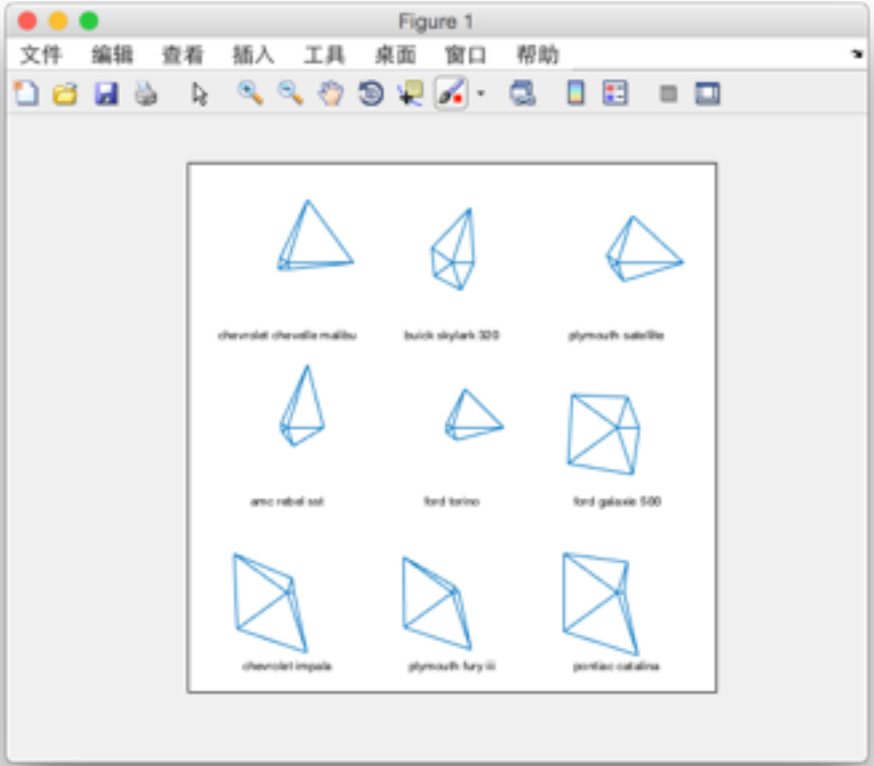


Figure 1
文件 编辑 查看 插入 工具 桌面 窗口 帮助
chevrolet chevelle malibu
buick skylark 320
plymouth satellite
amc rebel sst
ford torino
ford galaxie 500
chevrolet impala
plymouth fury iii
pontiac catalina

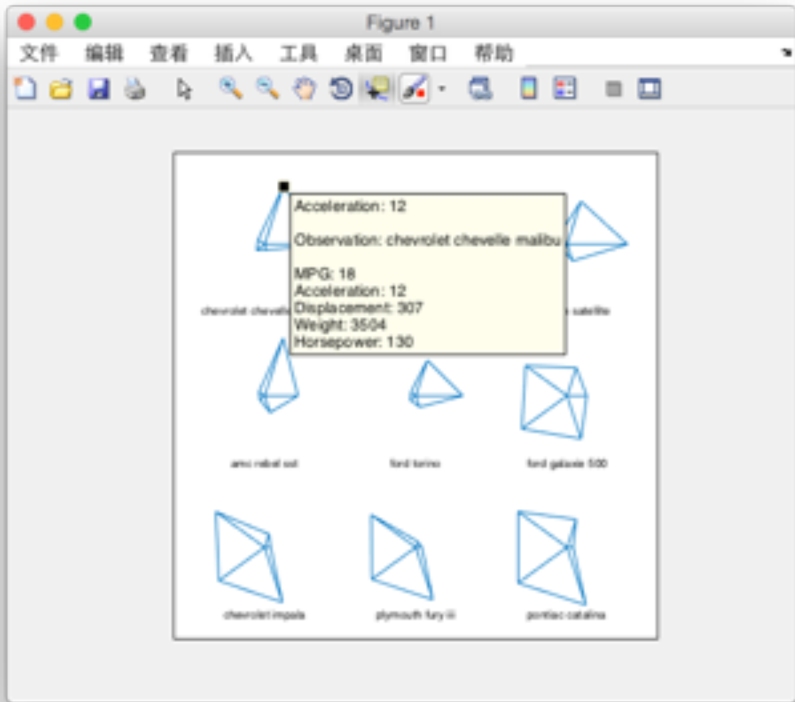
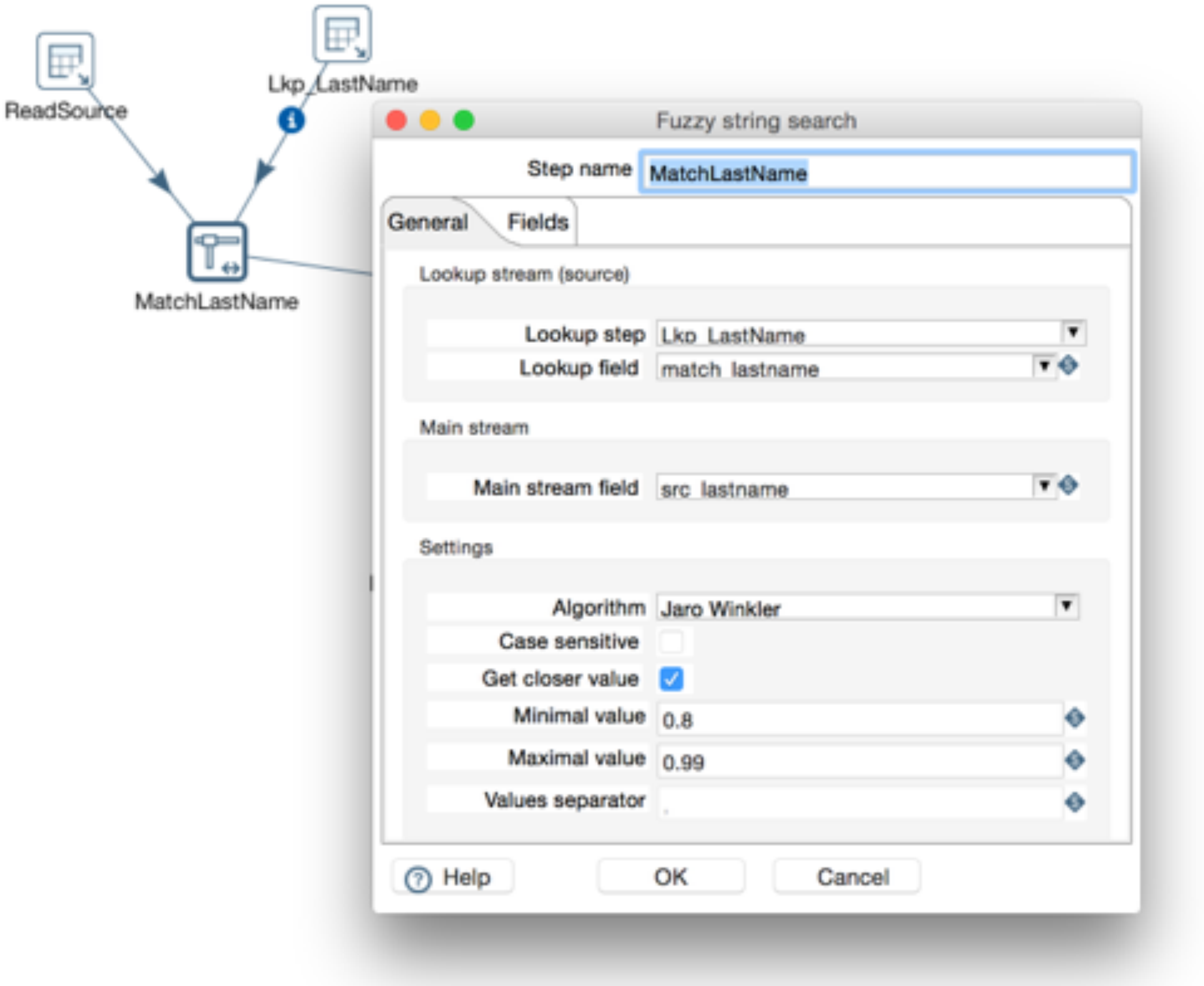


Figure 1
文件 编辑 查看 插入 工具 桌面 窗口 帮助
Acceleration: 12
Observation: chevrolet chevelle malibu
MPG: 18
Acceleration: 12
Displacement: 307
Weight: 3504
Horsepower: 130
amc rebel sst
ford torino
ford galaxie 500
chevrolet impala
plymouth fury iii
pontiac catalina

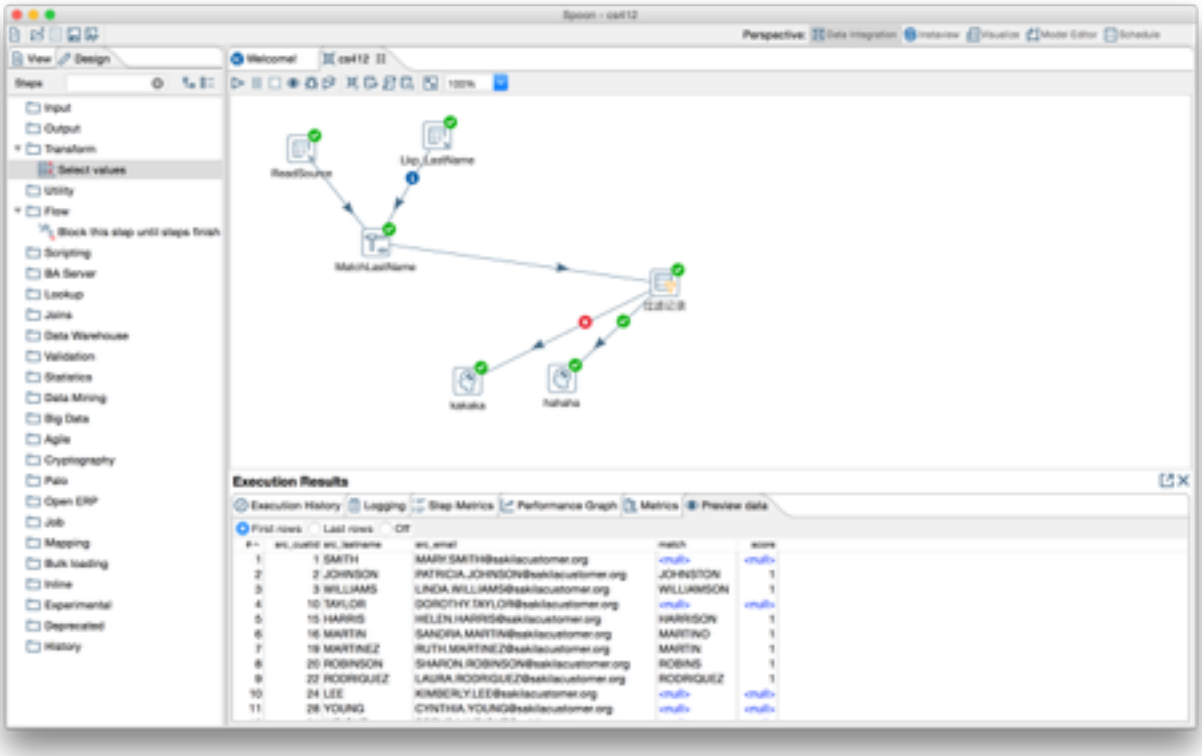
Mini Machine Problem 2

1.



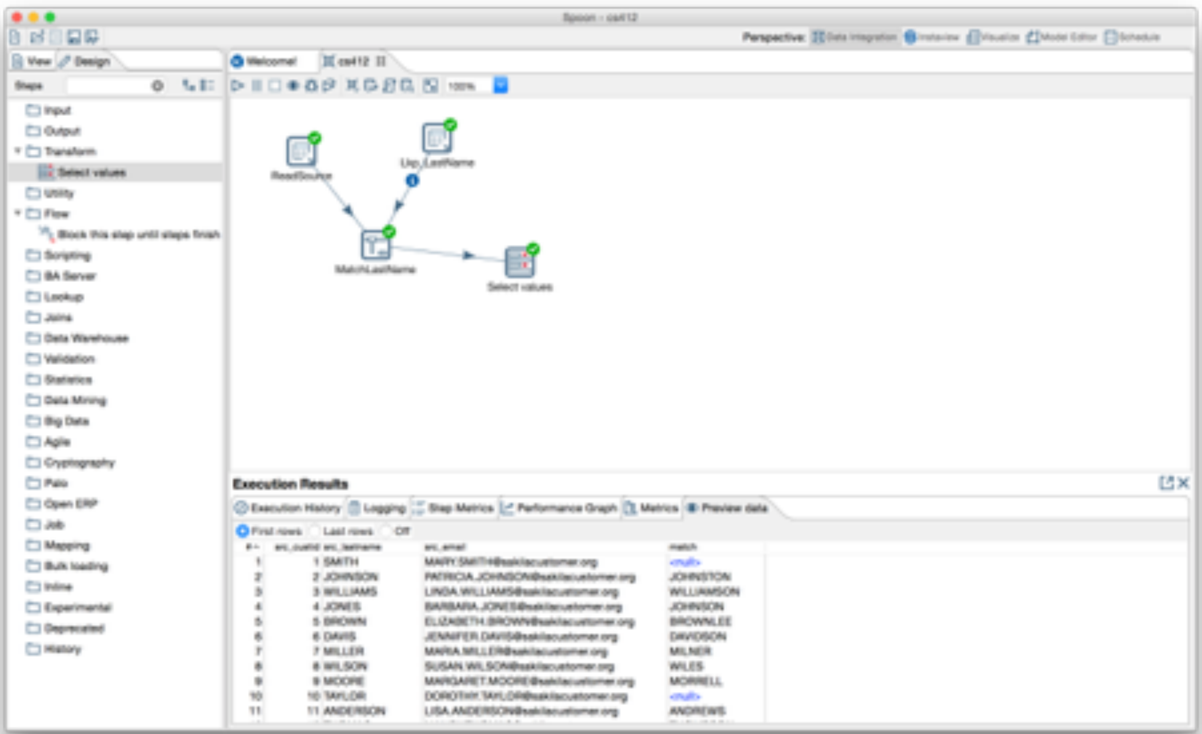
If I set the max value to 1.00, all the result will have the value 1. Therefore, we should set the distance to 0.99.

2.



Spoon - cs412
Perspective:
View
Design
Welcome!
cs412
ReadSource
Up_LastName
MatchLastName
hahaha
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Metrics
Preview data

3.



Spoon - cs412
Perspective:
View
Design
Welcome!
cs412
ReadSource
Up_LastName
MatchLastName
Select values
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Metrics
Preview data