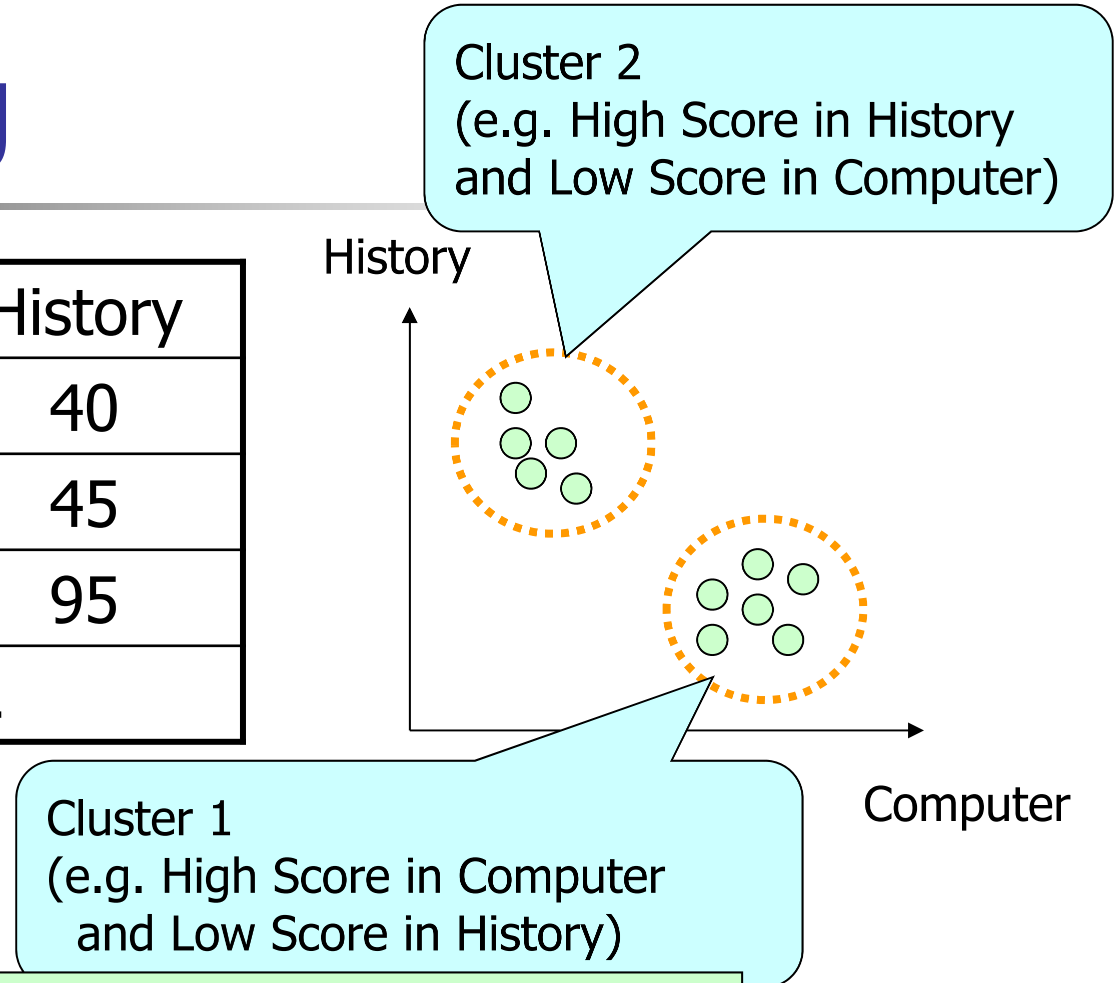
COMP1942

Clustering 1
(Introduction and kmean)

Prepared by Raymond Wong
Some parts of this notes are borrowed from LW Chan's notes
Screenshots of XLMiner Captured by Hao Liu
Presented by Raymond Wong
raywong@cse

Clustering

	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
…	…	…



Problem: to find all clusters

Why Clustering?

- Clustering for Utility
  - Summarization
  - Compression

Why Clustering?

- Clustering for Understanding
  - Applications
    - Biology
      - Group different species
    - Psychology and Medicine
      - Group medicine
    - Business
      - Group different customers for marketing
    - Network
      - Group different types of traffic patterns
    - Software
      - Group different programs for data analysis

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
  - How to use the data mining tool
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach
  - How to use the data mining tool

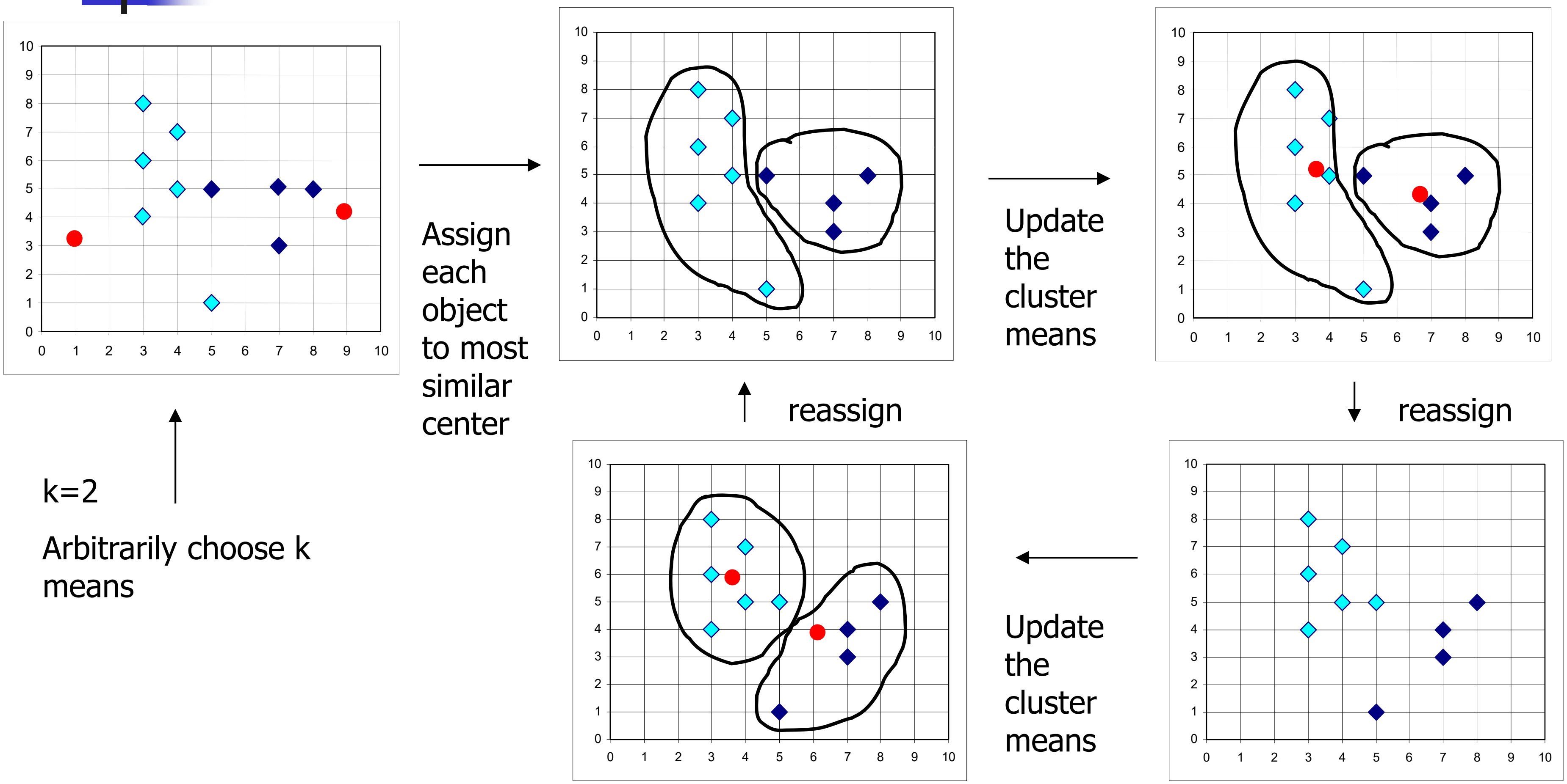
K-mean Clustering

- Suppose that we have n example feature vectors $x_1, x_2, \ldots, x_n$, and we know that they fall into k compact clusters, $k < n$
- Let m_i be the mean of the vectors in cluster i.
- we can say that x is in cluster i if distance from x to m_i is the minimum of all the k distances.

Procedure for finding k-means

- Make initial guesses for the means m_1, m_2, .., m_k
- Until there is no change in any mean
  - Assign each data point to the cluster whose mean is the nearest
  - Calculate the mean of each cluster
  - For i from 1 to k
    - Replace m_i with the mean of all examples for cluster i

Procedure for finding k-means



Initialization of k-means

- The way to initialize the means was not specified. One popular way to start is to randomly choose k of the examples
- The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points

Disadvantages of k-means

- Disadvantages
  - In a “bad” initial guess, there are no points assigned to the cluster with the initial mean m_i.
  - The value of k is not user-friendly. This is because we do not know the number of clusters before we want to find clusters.

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
  - How to use the data mining tool
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach
  - How to use the data mining tool

Sequential k-Means Clustering

- Another way to modify the k-means procedure is to update the means one example at a time, rather than all at once.
- This is particularly attractive when we acquire the examples over a period of time, and we want to start clustering before we have seen all of the examples
- Here is a modification of the k-means procedure that operates sequentially

Sequential k-Means Clustering

- Make initial guesses for the means m_1, m_2, …, m_k
- Set the counts n_1, n_2, .., n_k to zero
- Until interrupted
  - Acquire the next example, x
  - If m_i is closest to x
    - Increment n_i
    - Replace m_i by $m_i + (1/n_i) \cdot (x - m_i)$

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
  - How to use the data mining tool
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach
  - How to use the data mining tool

Forgetful Sequential k-means

- This also suggests another alternative in which we replace the counts by constants. In particular, suppose that a is a constant between 0 and 1, and consider the following variation:

- Make initial guesses for the means $m_1, m_2, \ldots, m_k$
- Until interrupted
  - Acquire the next example x
  - If m_i is closest to x, replace m_i by $m_i+a(x-m_i)$

Forgetful Sequential k-means

- The result is called the "forgetful" sequential k-means procedure.
- It is not hard to show that m_i is a weighted average of the examples that were closest to m_i, where the weight decreases exponentially with the "age" to the example.
- That is, if m_0 is the initial value of the mean vector and if x_j is the j-th example out of n examples that were used to form m_i, then it is not hard to show that

$$m_n = (1-a)^n m_0 + a \sum_{k=1}^{n} (1-a)^{n-k} x_k$$

Forgetful Sequential k-means

- Thus, the initial value m_0 is eventually forgotten, and recent examples receive more weight than ancient examples.
- This variation of k-means is particularly simple to implement, and it is attractive when the nature of the problem changes over time and the cluster centres “drift”.

Clustering Methods

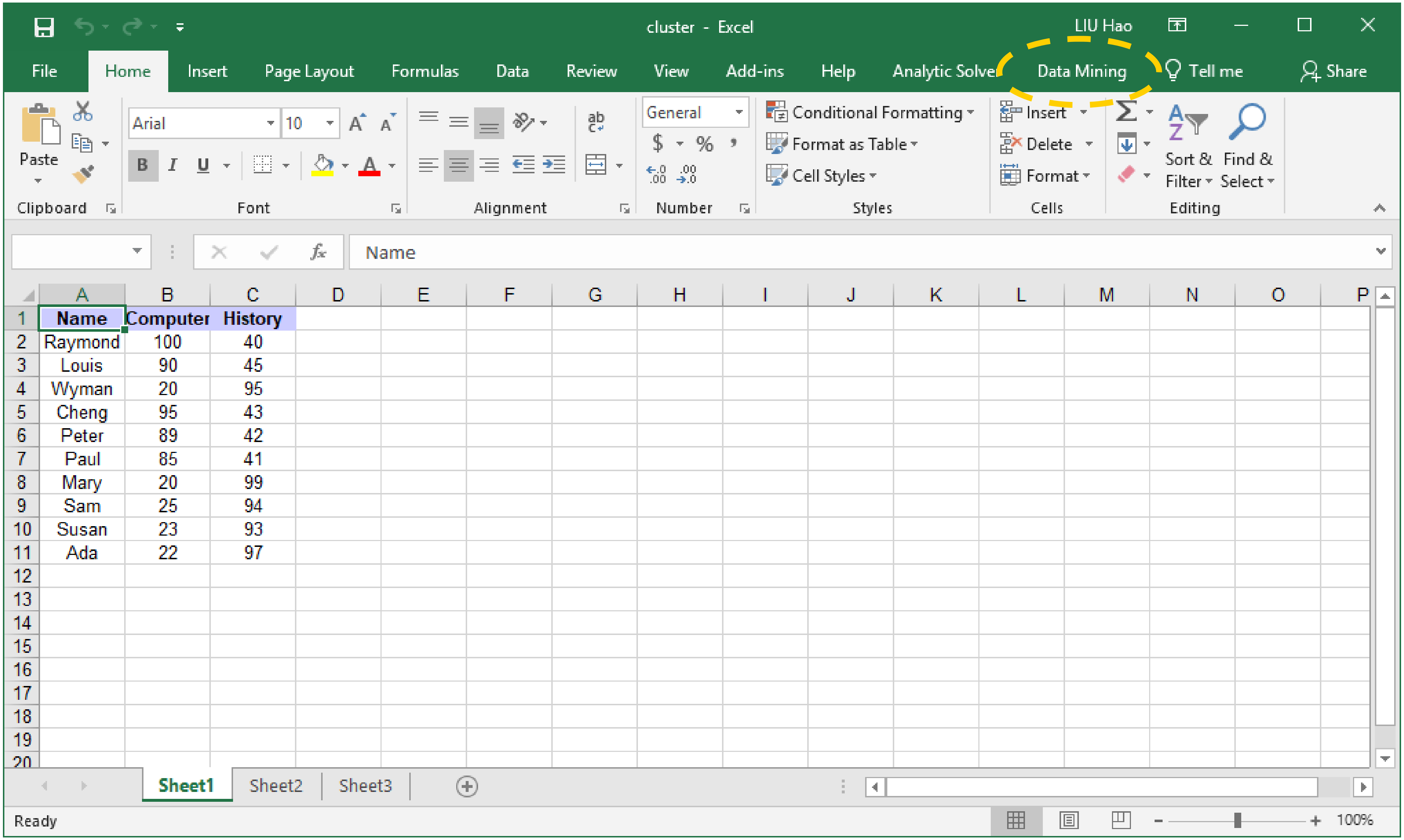
- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
  - How to use the data mining tool
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach
  - How to use the data mining tool

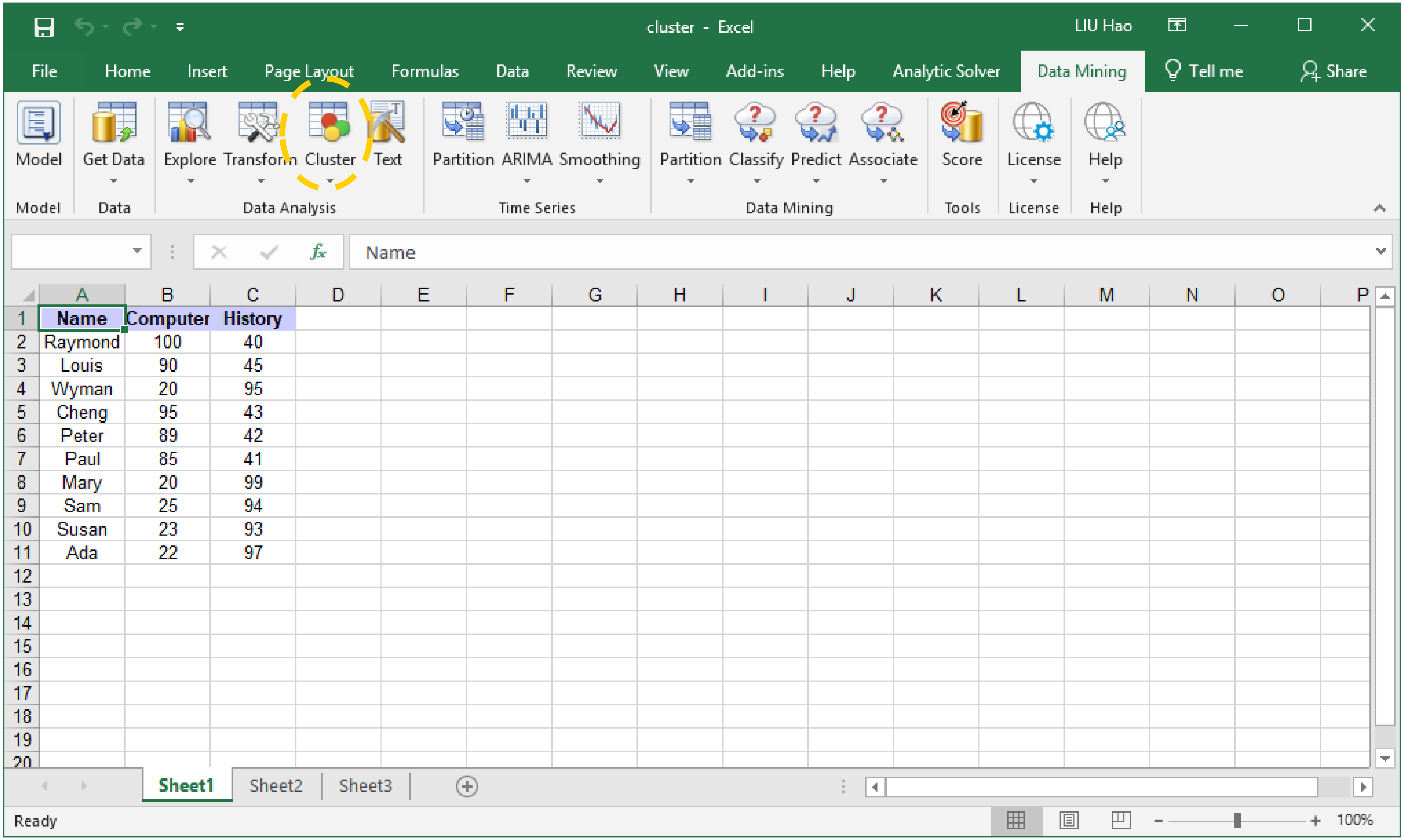
How to use the data mining tool

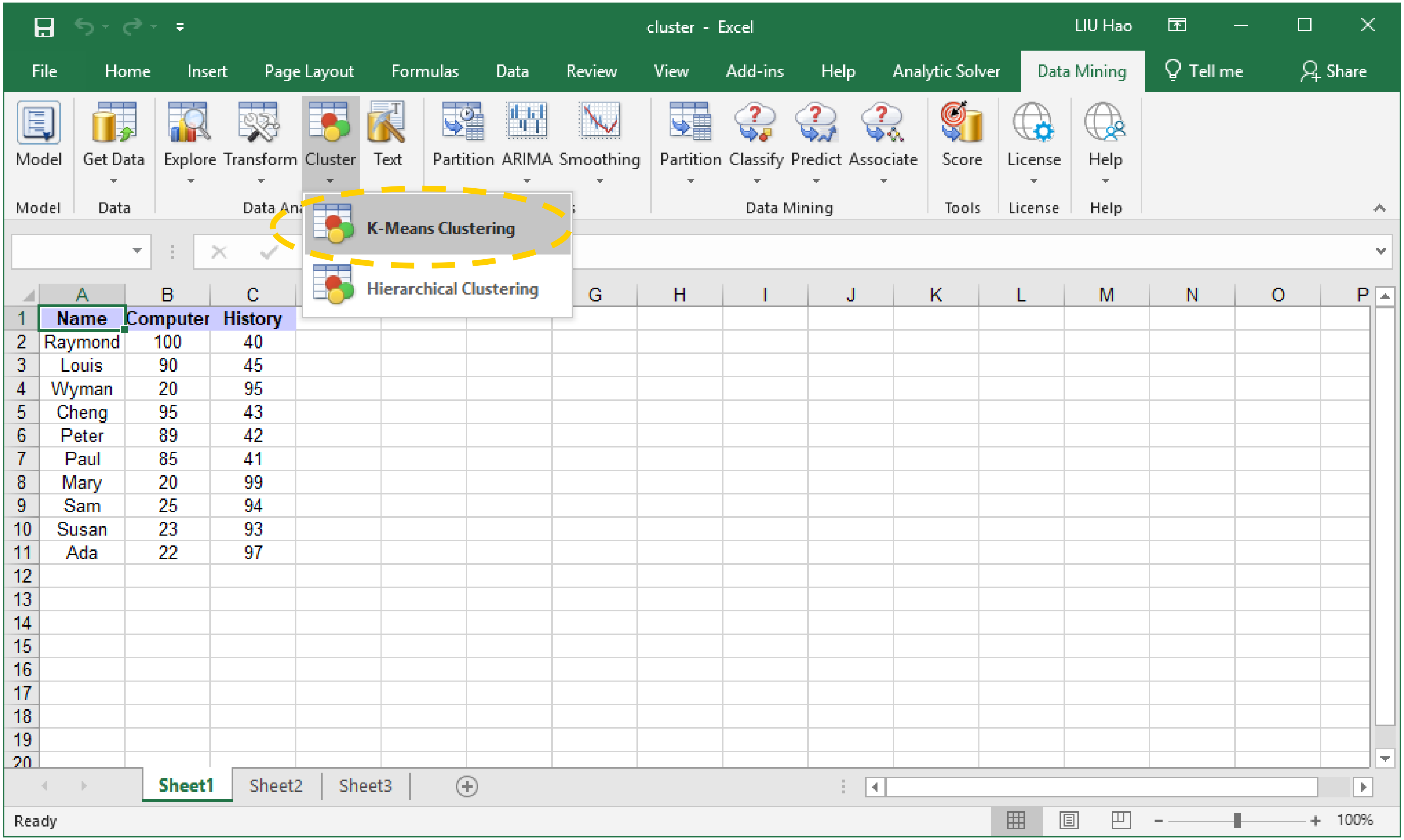
- We have the following 2 versions.
  - XLMiner Desktop (installed in either the CSE lab machine or your computer)
  - XLMiner Cloud (installed as a plugin in your Office 365 Excel)

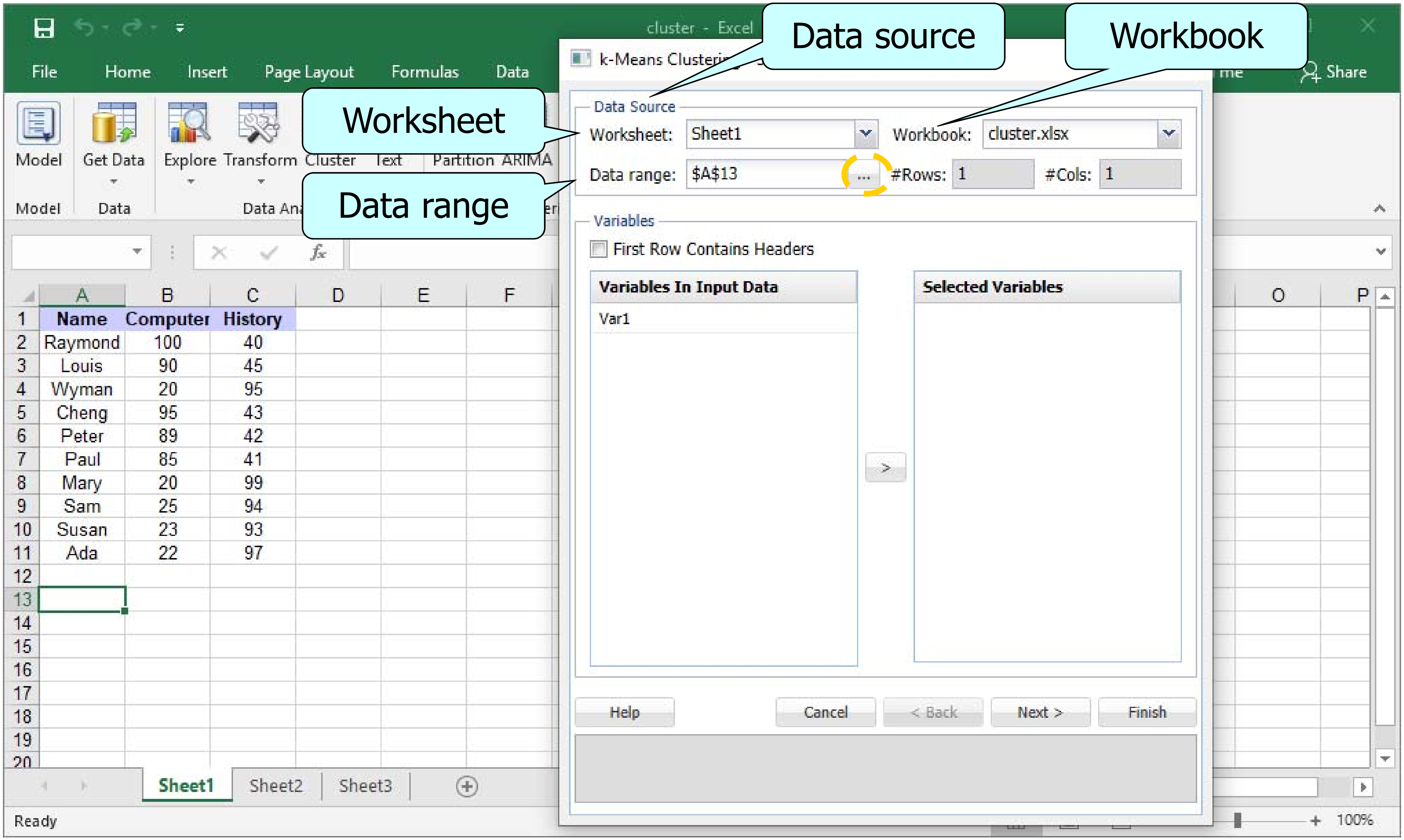
How to use the data mining tool (XLMiner Desktop)

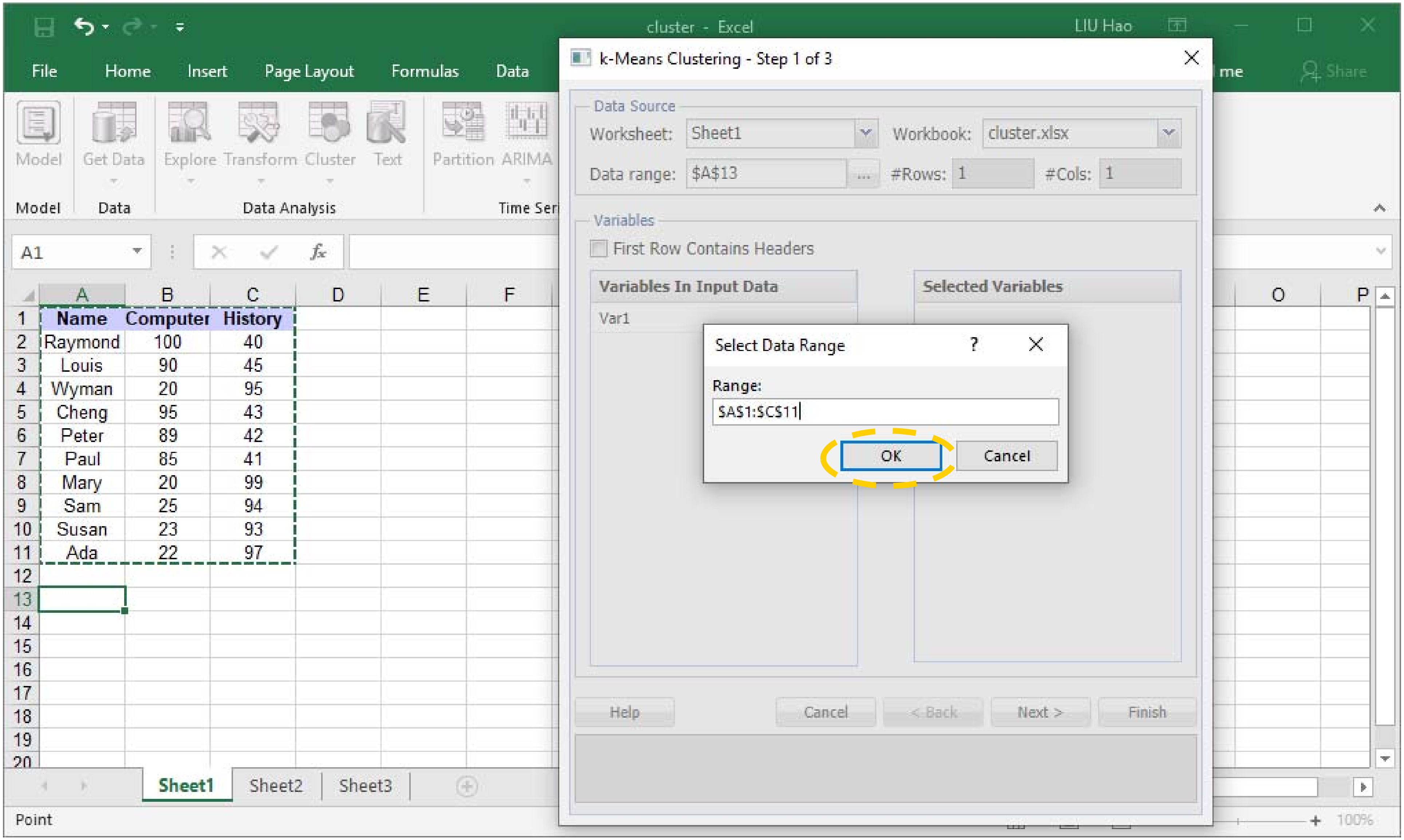
- We can use XLMiner for performing k-mean clustering
- Open “cluster.xlsx” in MS Excel in a CSE lab machine

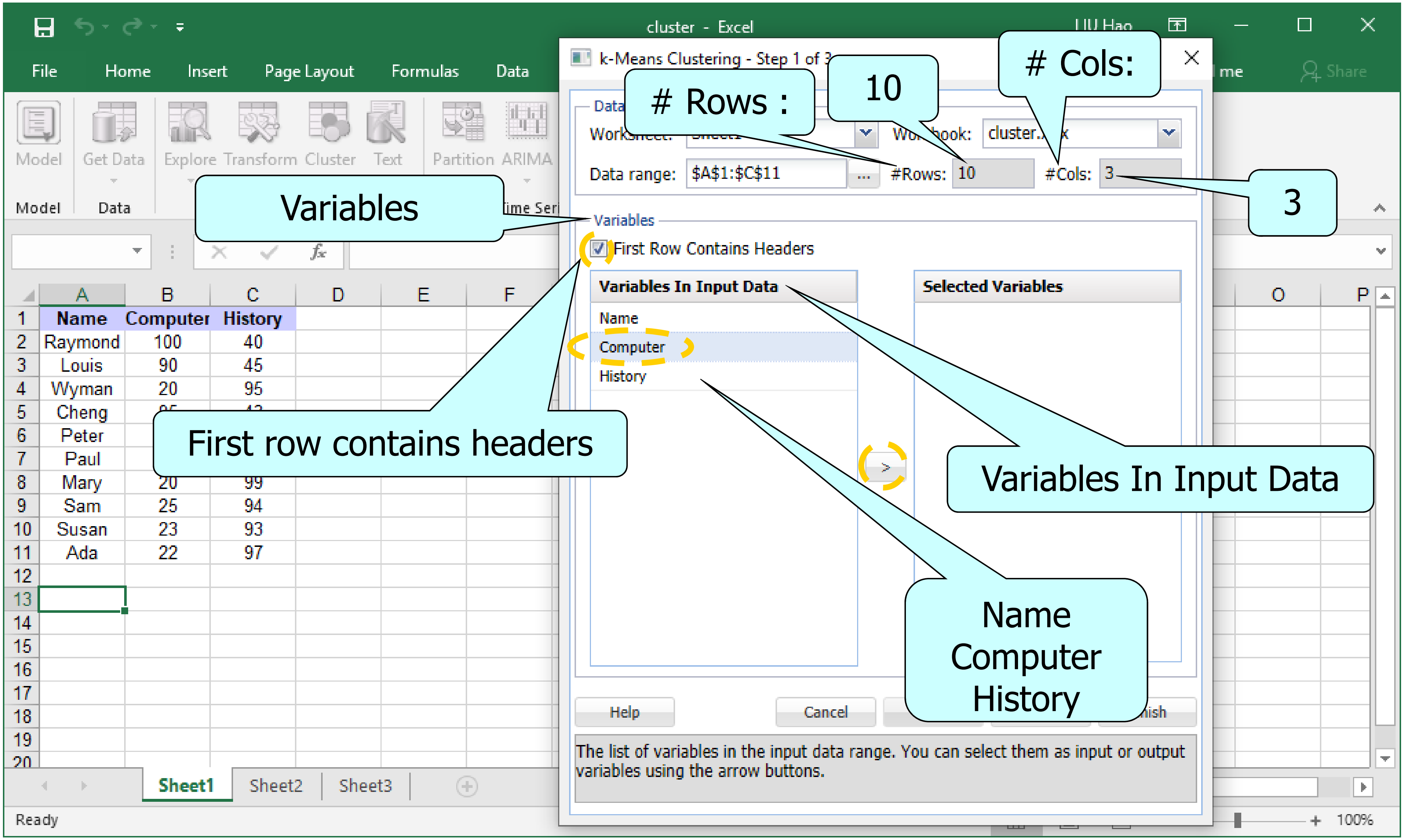

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Paste
Clipboard
Arial
10
Font
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Find & Select
Editing
Name
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

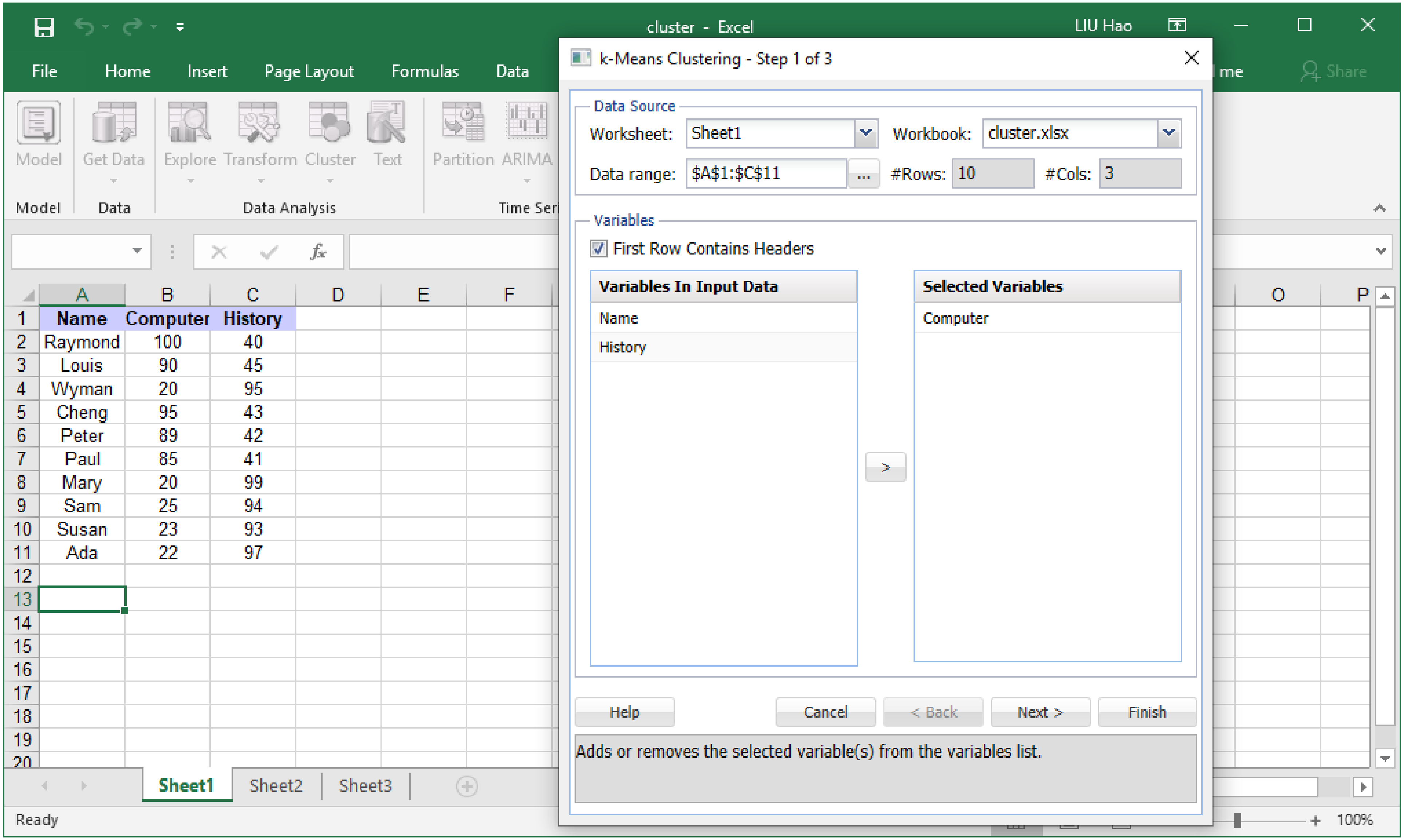

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
Name
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

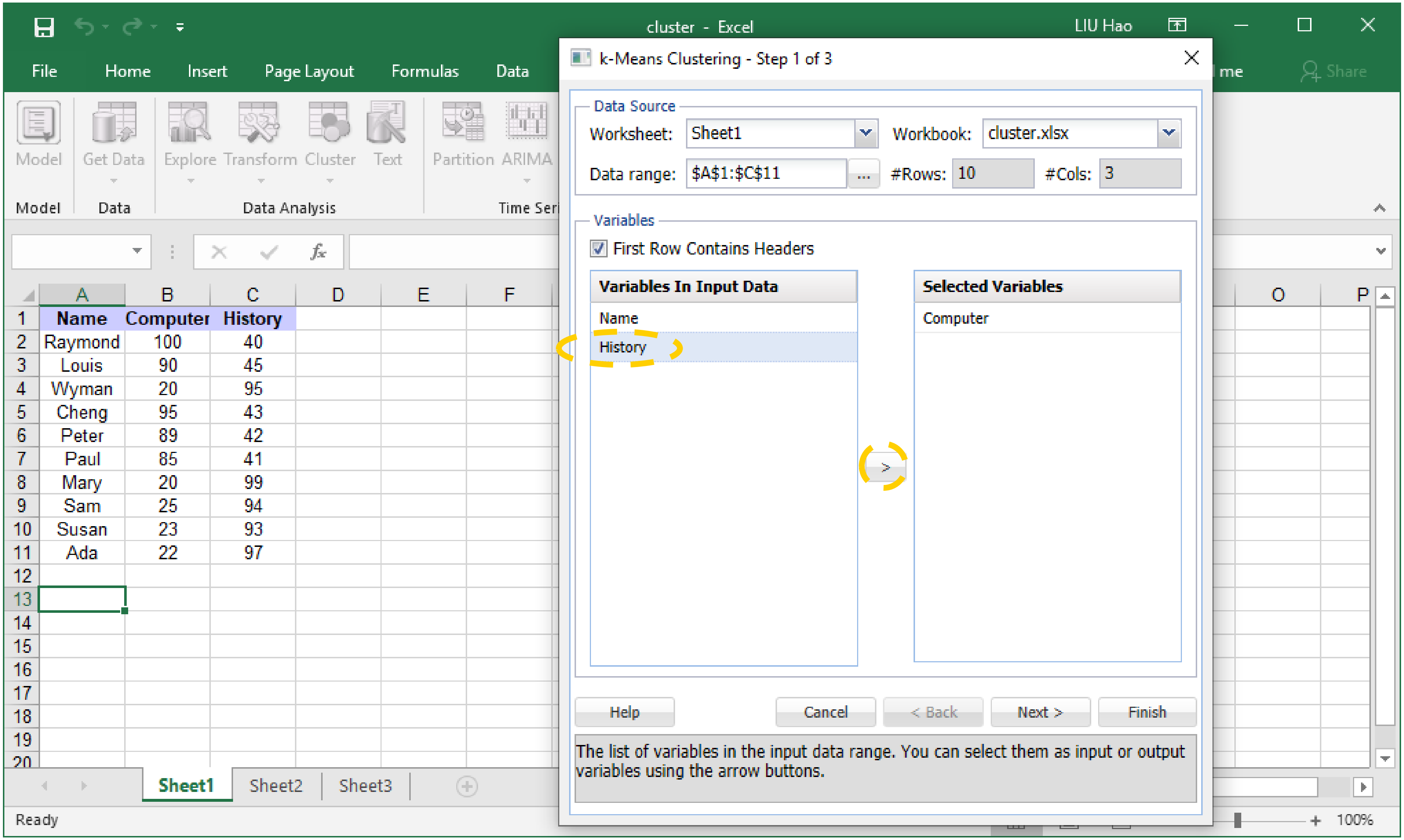

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Mining
Tools
License
Help
K-Means Clustering
Hierarchical Clustering
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

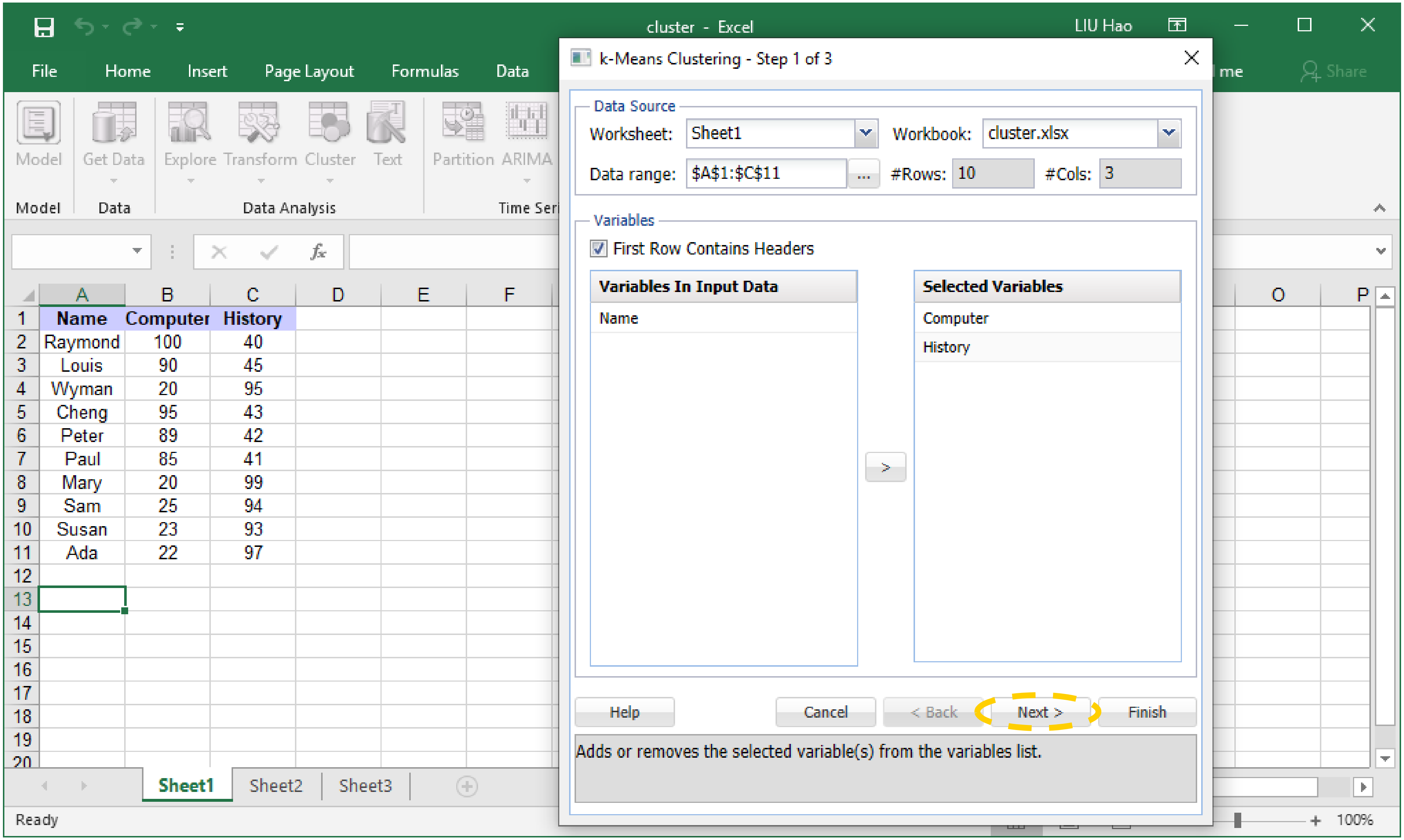

Data source
Workbook
Worksheet
Data range
cluster - Excel
k-Means Clustering
File
Home
Insert
Page Layout
Formulas
Data
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data Source
Worksheet: Sheet1
Workbook: cluster.xlsx
Data range: $A$13
#Rows: 1
#Cols: 1
Variables
First Row Contains Headers
Variables In Input Data
Var1
Selected Variables
Help
Cancel
< Back
Next >
Finish
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

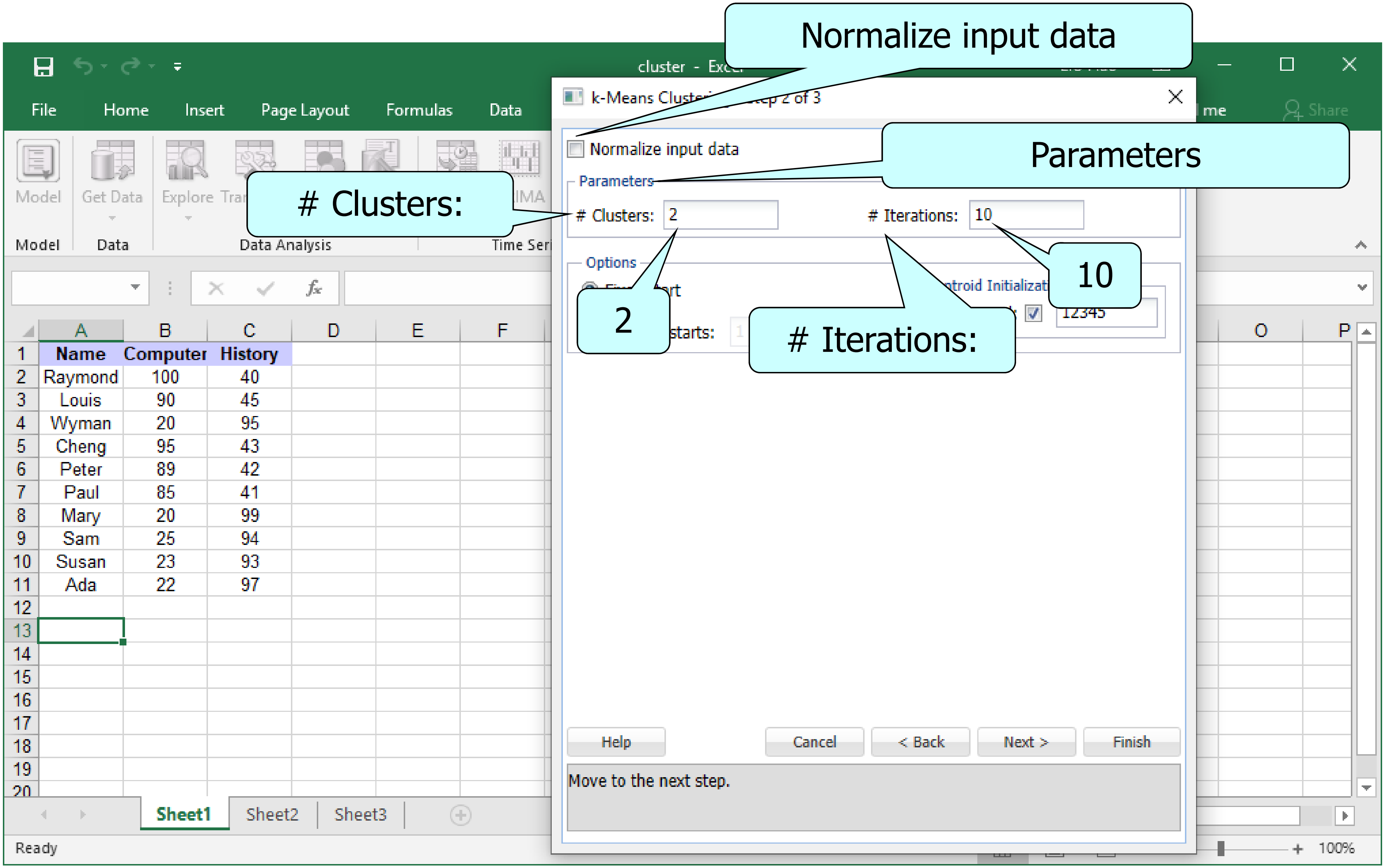

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data
Data Analysis
A1
Name
Computer
History
Raymond
100
40
Louis
90
45
Wyman
20
95
Cheng
95
43
Peter
89
42
Paul
85
41
Mary
20
99
Sam
25
94
Susan
23
93
Ada
22
97
Sheet1
Sheet2
Sheet3
Point
k-Means Clustering - Step 1 of 3
Data Source
Worksheet:
Sheet1
Workbook:
cluster.xlsx
Data range:
$A$13
#Rows:
1
#Cols:
1
Variables
First Row Contains Headers
Variables In Input Data
Selected Variables
Var1
Select Data Range
Range:
$A$1:$C$11
OK
Cancel
Help
Cancel
< Back
Next >
Finish

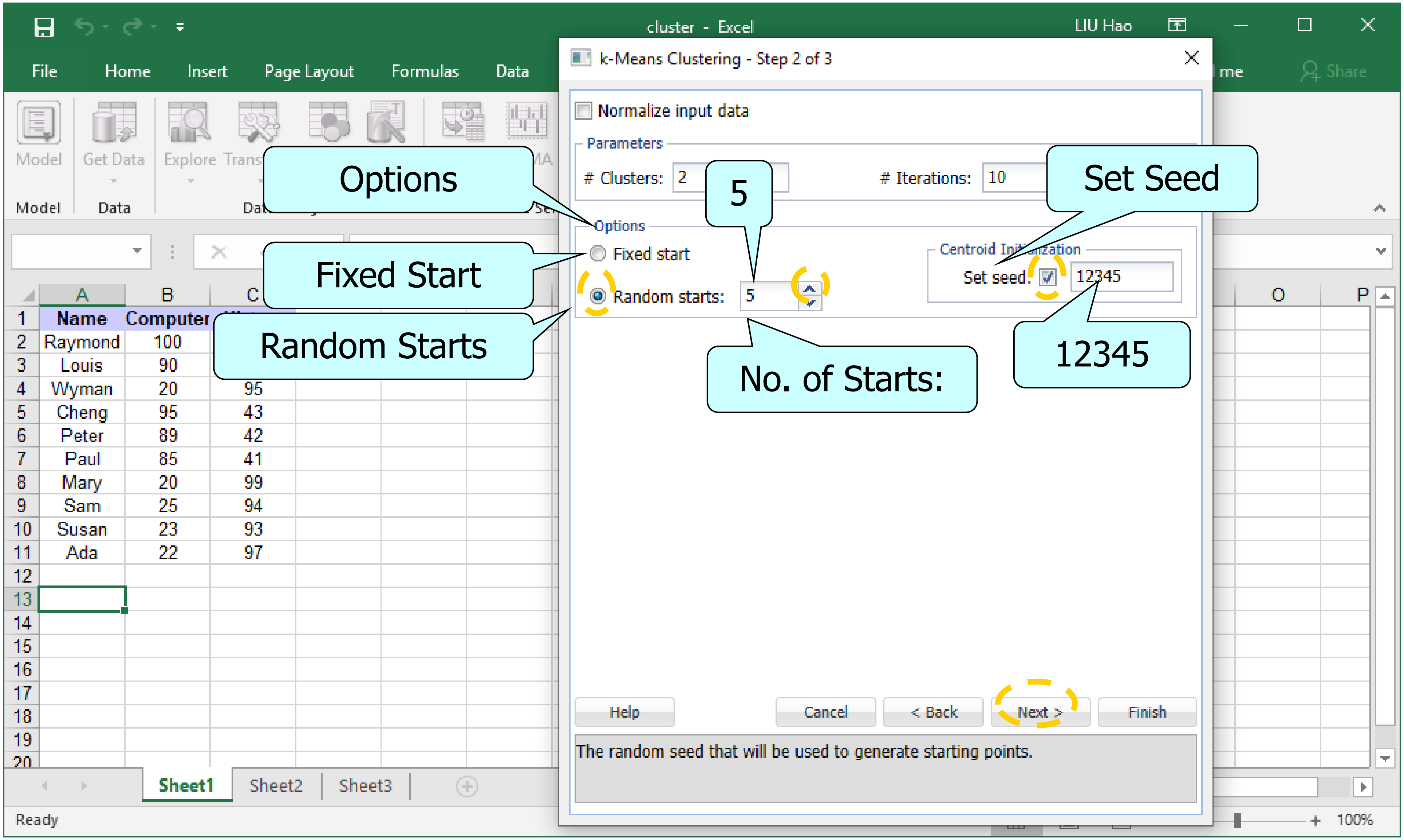

k-Means Clustering - Step 1 of 3
Rows :
10
Cols:
3
Data range: $A$1:$C$11
#Rows: 10
#Cols: 3
Variables
First Row Contains Headers
First row contains headers
Variables In Input Data
Selected Variables
Name
Computer
History
Variables In Input Data
Name
Computer
History
Help
Cancel
The list of variables in the input data range. You can select them as input or output variables using the arrow buttons.
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng
Peter
Paul
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready

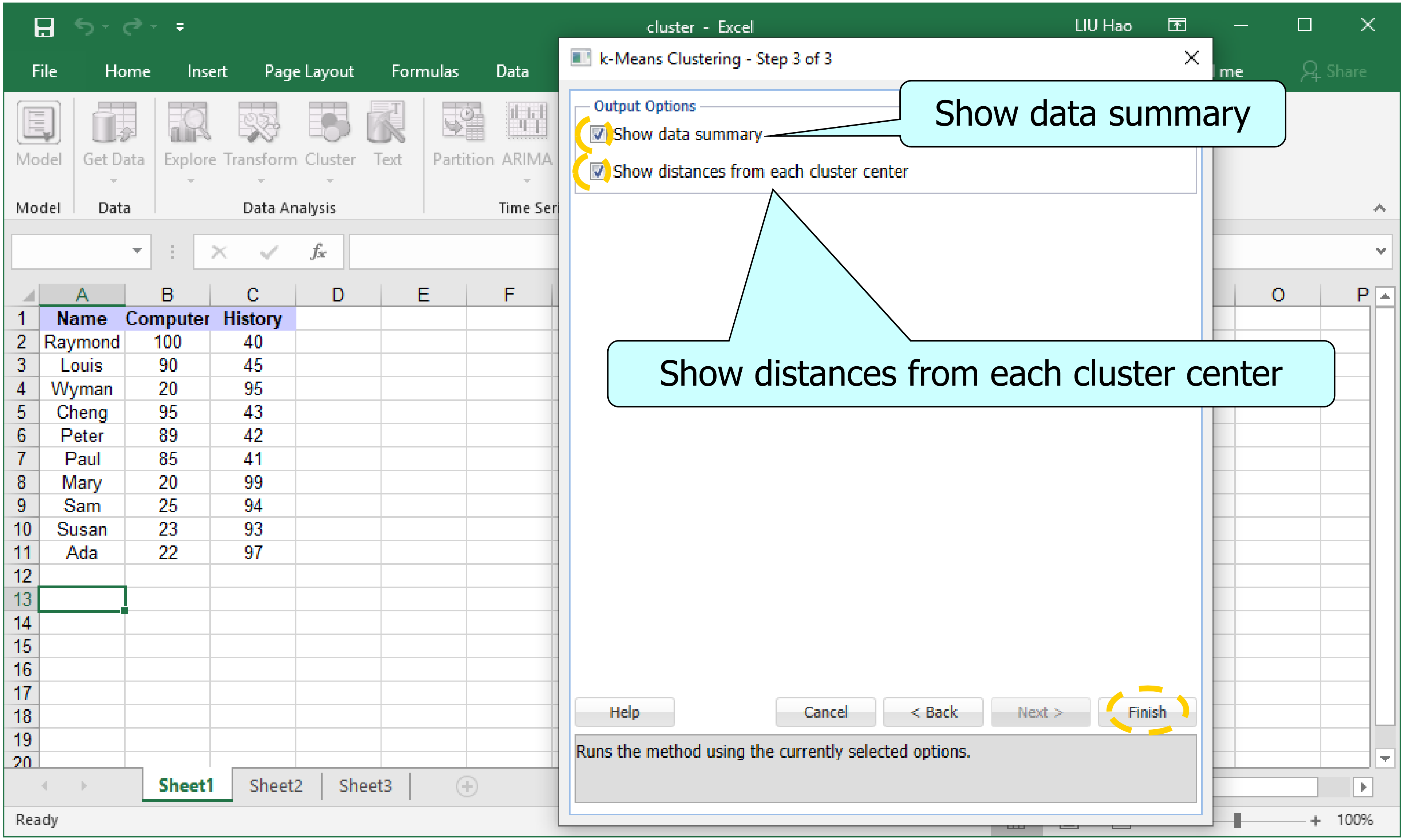

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data Analysis
k-Means Clustering - Step 1 of 3
Data Source
Worksheet: Sheet1
Workbook: cluster.xlsx
Data range: $A$1:$C$11
#Rows: 10
#Cols: 3
Variables
First Row Contains Headers
Variables In Input Data
Name
History
Selected Variables
Computer
Help
Cancel
< Back
Next >
Finish
Adds or removes the selected variable(s) from the variables list.
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

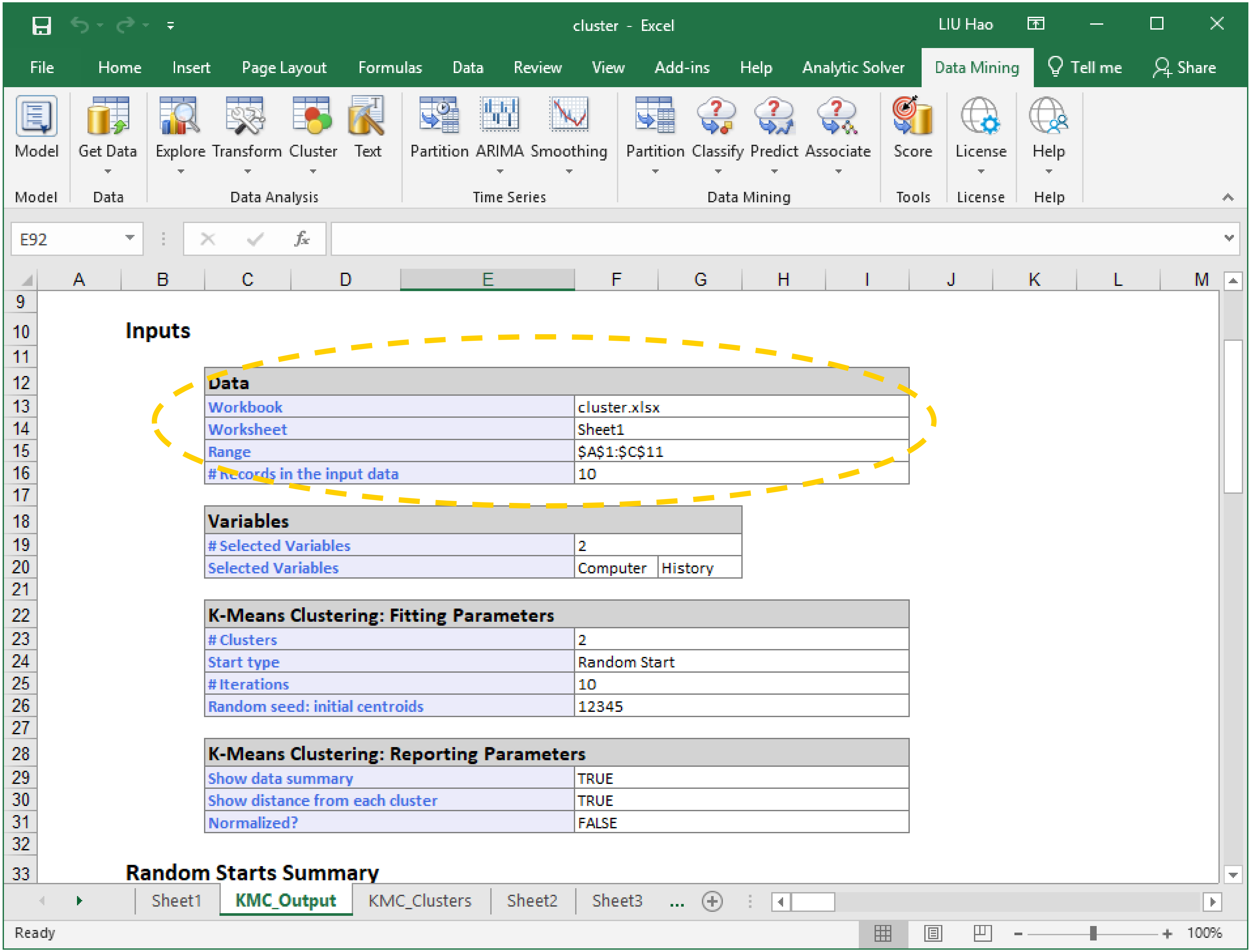

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Model
Data
Data Analysis
k-Means Clustering - Step 1 of 3
Data Source
Worksheet: Sheet1
Workbook: cluster.xlsx
Data range: $A$1:$C$11
#Rows: 10
#Cols: 3
Variables
First Row Contains Headers
Variables In Input Data
Name
History
Selected Variables
Computer
Help
Cancel
< Back
Next >
Finish
The list of variables in the input data range. You can select them as input or output variables using the arrow buttons.
Name
Computer
History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

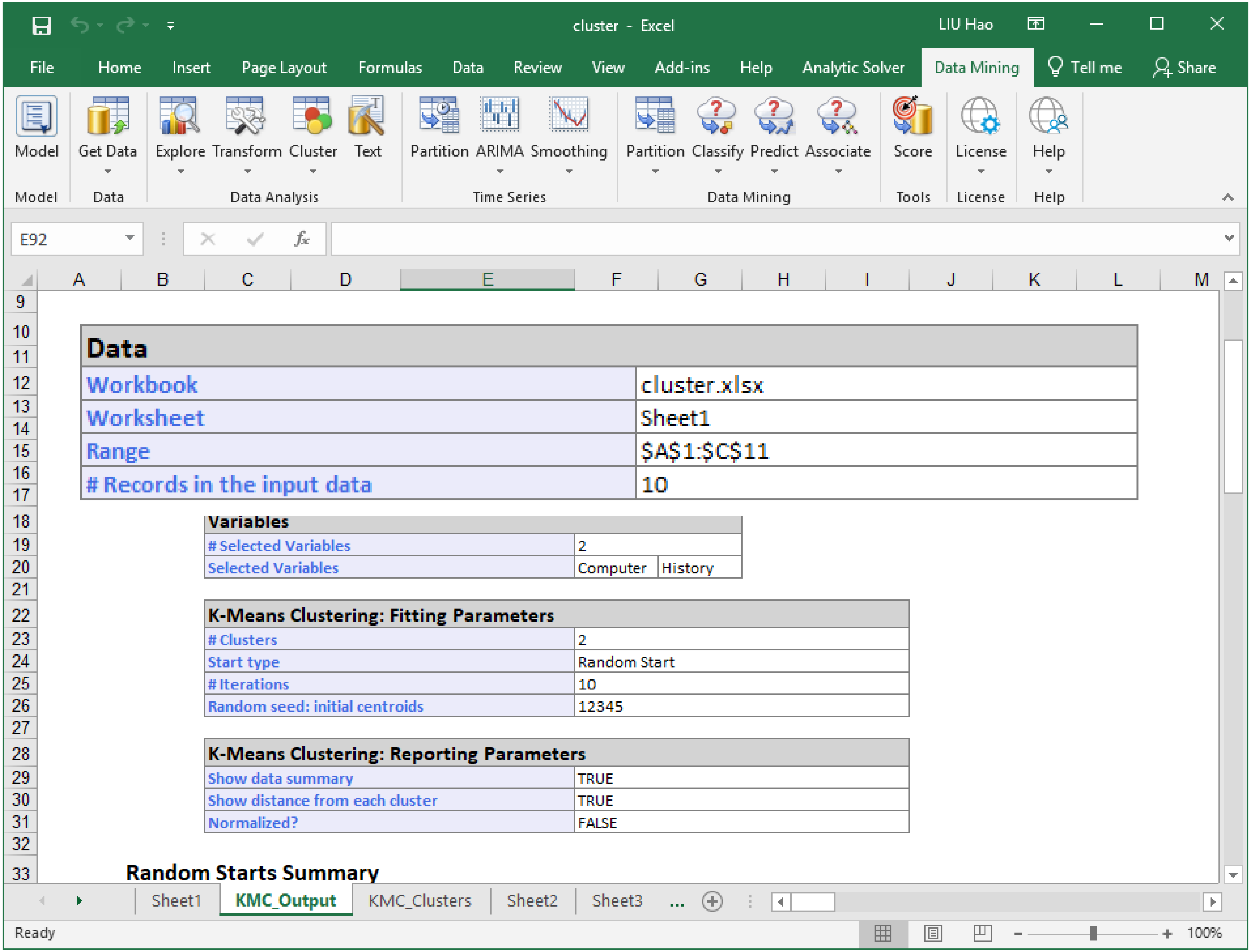

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Model
Data
Data Analysis
k-Means Clustering - Step 1 of 3
Data Source
Worksheet: Sheet1
Workbook: cluster.xlsx
Data range: $A$1:$C$11
#Rows: 10
#Cols: 3
Variables
First Row Contains Headers
Variables In Input Data
Name
Selected Variables
Computer
History
Help
Cancel
< Back
Next >
Finish
Adds or removes the selected variable(s) from the variables list.
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%

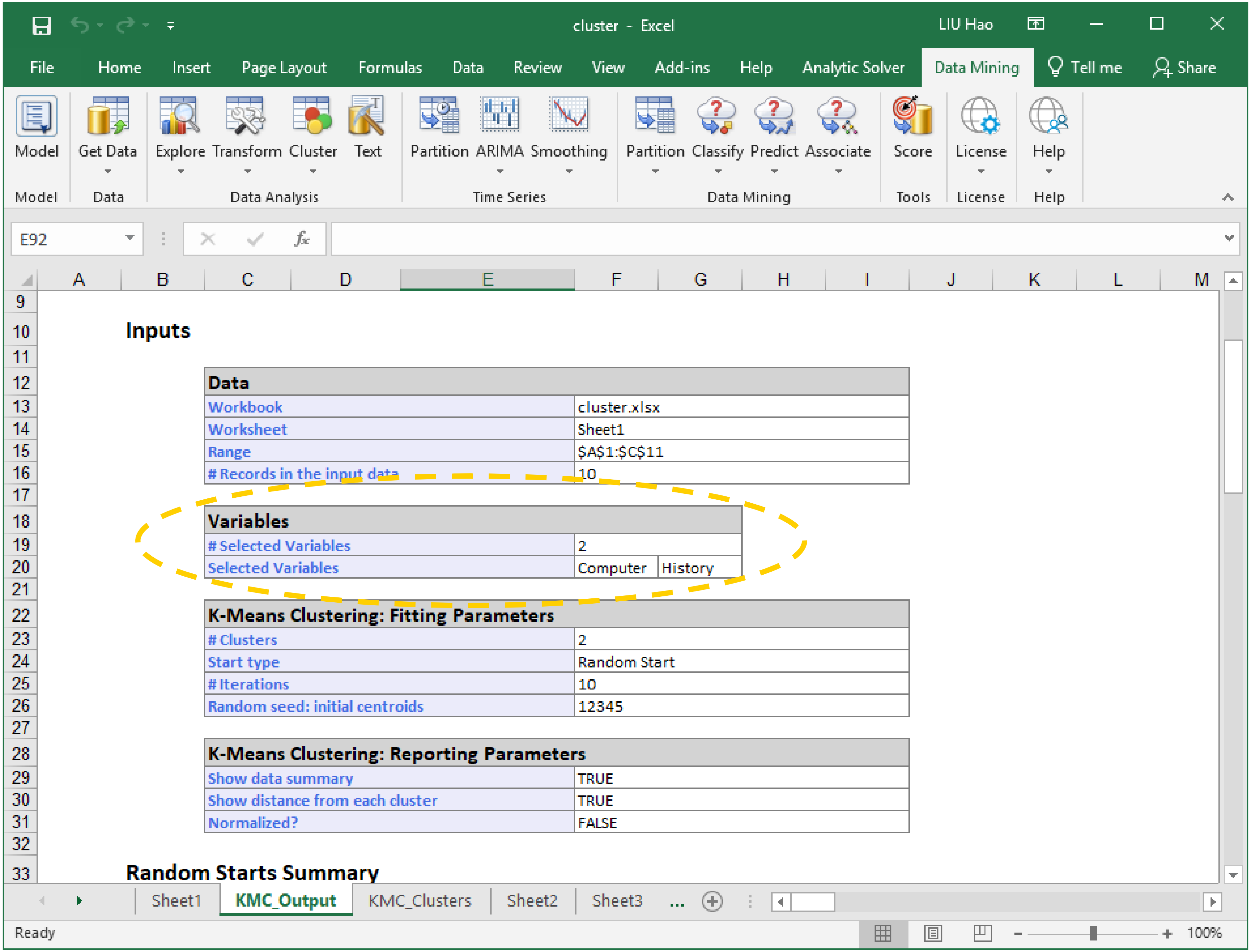

Normalize input data
Parameters
Clusters:
2
10
Iterations:
k-Means Clustering - Step 2 of 3
Normalize input data
Parameters
Clusters: 2
Iterations: 10
Options
Help
Cancel
< Back
Next >
Finish
Move to the next step.
Name Computer History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready

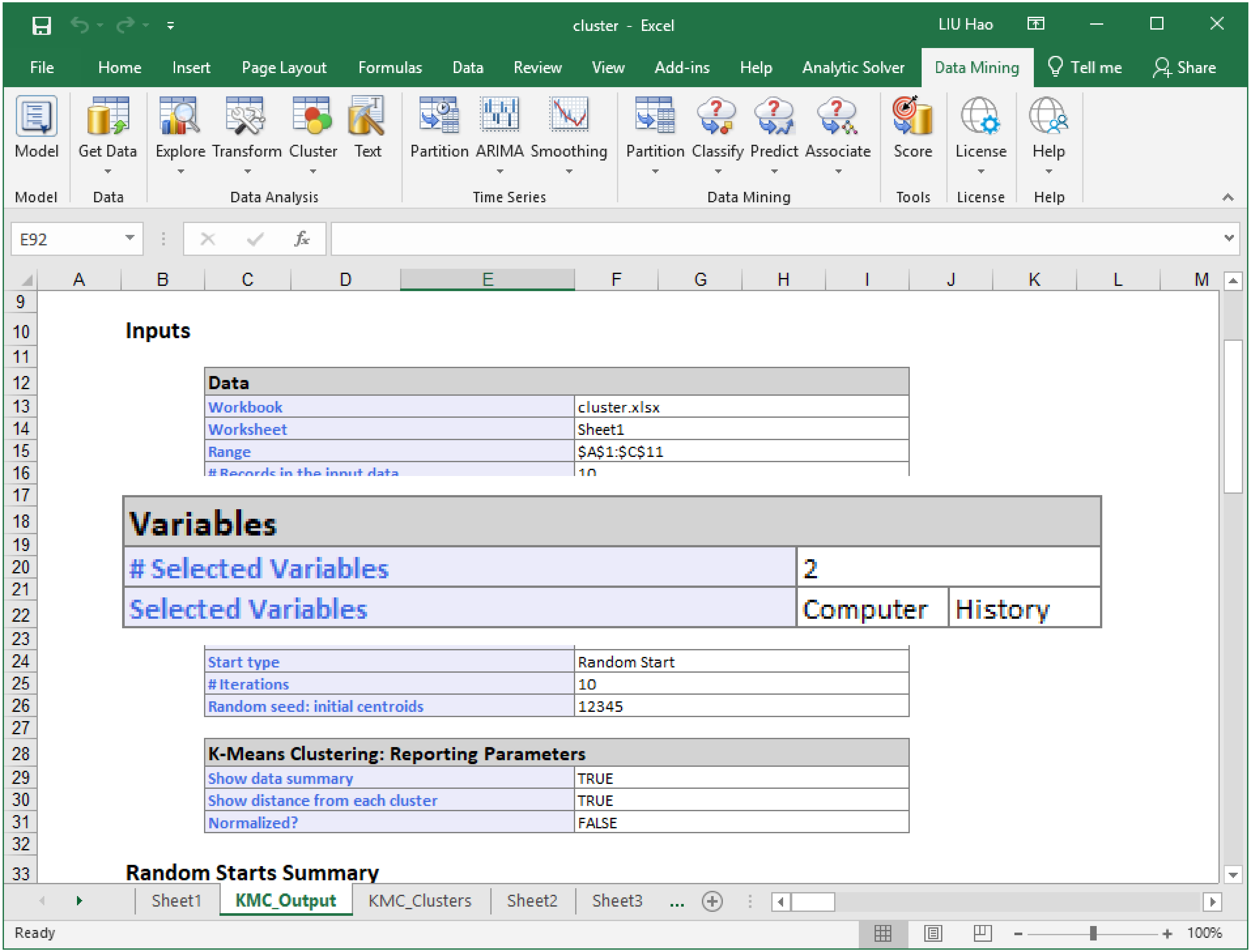

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Model
Get Data
Explore
Trans
Name
Computer
Raymond
100
Louis
90
Wyman
20
95
Cheng
95
43
Peter
89
42
Paul
85
41
Mary
20
99
Sam
25
94
Susan
23
93
Ada
22
97
Sheet1
Sheet2
Sheet3
Ready
k-Means Clustering - Step 2 of 3
Normalize input data
Parameters
Clusters: 2
Iterations: 10
Options
Fixed start
Random starts: 5
Centroid Initialization
Set seed: 12345
Help
Cancel
< Back
Next >
Finish
The random seed that will be used to generate starting points.
Options
Fixed Start
Random Starts
5
No. of Starts:
Set Seed
12345

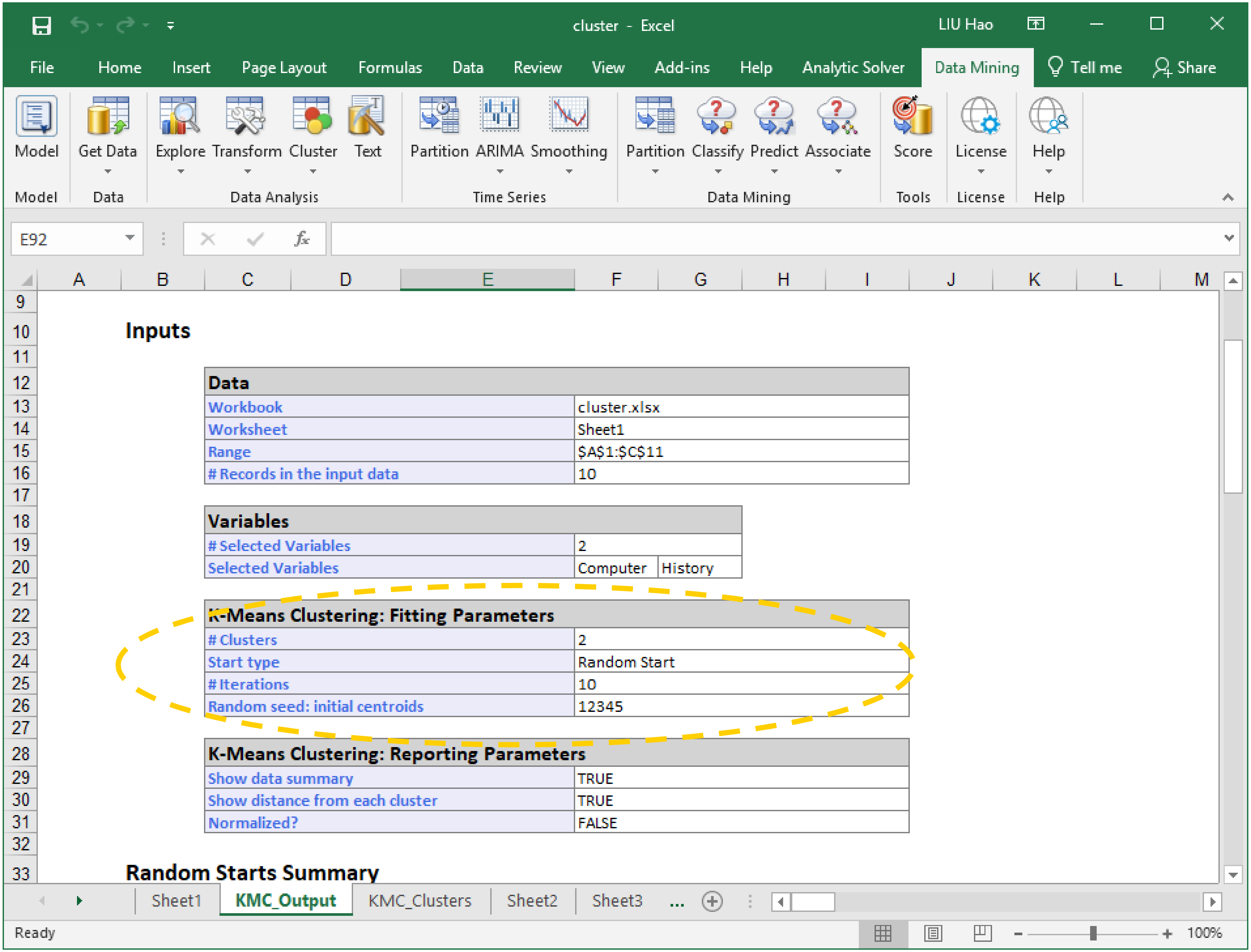

cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data Analysis
Time Seri
k-Means Clustering - Step 3 of 3
Output Options
Show data summary
Show distances from each cluster center
Help
Cancel
< Back
Next >
Finish
Runs the method using the currently selected options.
Show data summary
Show distances from each cluster center
Name
Computer
History
Raymond 100 40
Louis 90 45
Wyman 20 95
Cheng 95 43
Peter 89 42
Paul 85 41
Mary 20 99
Sam 25 94
Susan 23 93
Ada 22 97
Sheet1
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Inputs
Data
Workbook
cluster.xlsx
Worksheet
Sheet1
Range
$A$1:$C$11
Records in the input data
10
Variables
Selected Variables
2
Selected Variables
Computer
History
K-Means Clustering: Fitting Parameters
Clusters
2
Start type
Random Start
Iterations
10
Random seed: initial centroids
12345
K-Means Clustering: Reporting Parameters
Show data summary
TRUE
Show distance from each cluster
TRUE
Normalized?
FALSE
Random Starts Summary
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Data
Workbook
cluster.xlsx
Worksheet
Sheet1
Range
$A$1:$C$11
Records in the input data
10
Variables
Selected Variables
2
Selected Variables
Computer
History
K-Means Clustering: Fitting Parameters
Clusters
2
Start type
Random Start
Iterations
10
Random seed: initial centroids
12345
K-Means Clustering: Reporting Parameters
Show data summary
TRUE
Show distance from each cluster
TRUE
Normalized?
FALSE
Random Starts Summary
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Inputs
Data
Workbook cluster.xlsx
Worksheet Sheet1
Range $A$1:$C$11
Records in the input data 10
Variables
Selected Variables 2
Selected Variables Computer History
K-Means Clustering: Fitting Parameters
Clusters 2
Start type Random Start
Iterations 10
Random seed: initial centroids 12345
K-Means Clustering: Reporting Parameters
Show data summary TRUE
Show distance from each cluster TRUE
Normalized? FALSE
Random Starts Summary
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Inputs
Data
Workbook
cluster.xlsx
Worksheet
Sheet1
Range
$A$1:$C$11
Records in the input data
10
Variables
Selected Variables
2
Selected Variables
Computer
History
Start type
Random Start
Iterations
10
Random seed: initial centroids
12345
K-Means Clustering: Reporting Parameters
Show data summary
TRUE
Show distance from each cluster
TRUE
Normalized?
FALSE
Random Starts Summary
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



Inputs

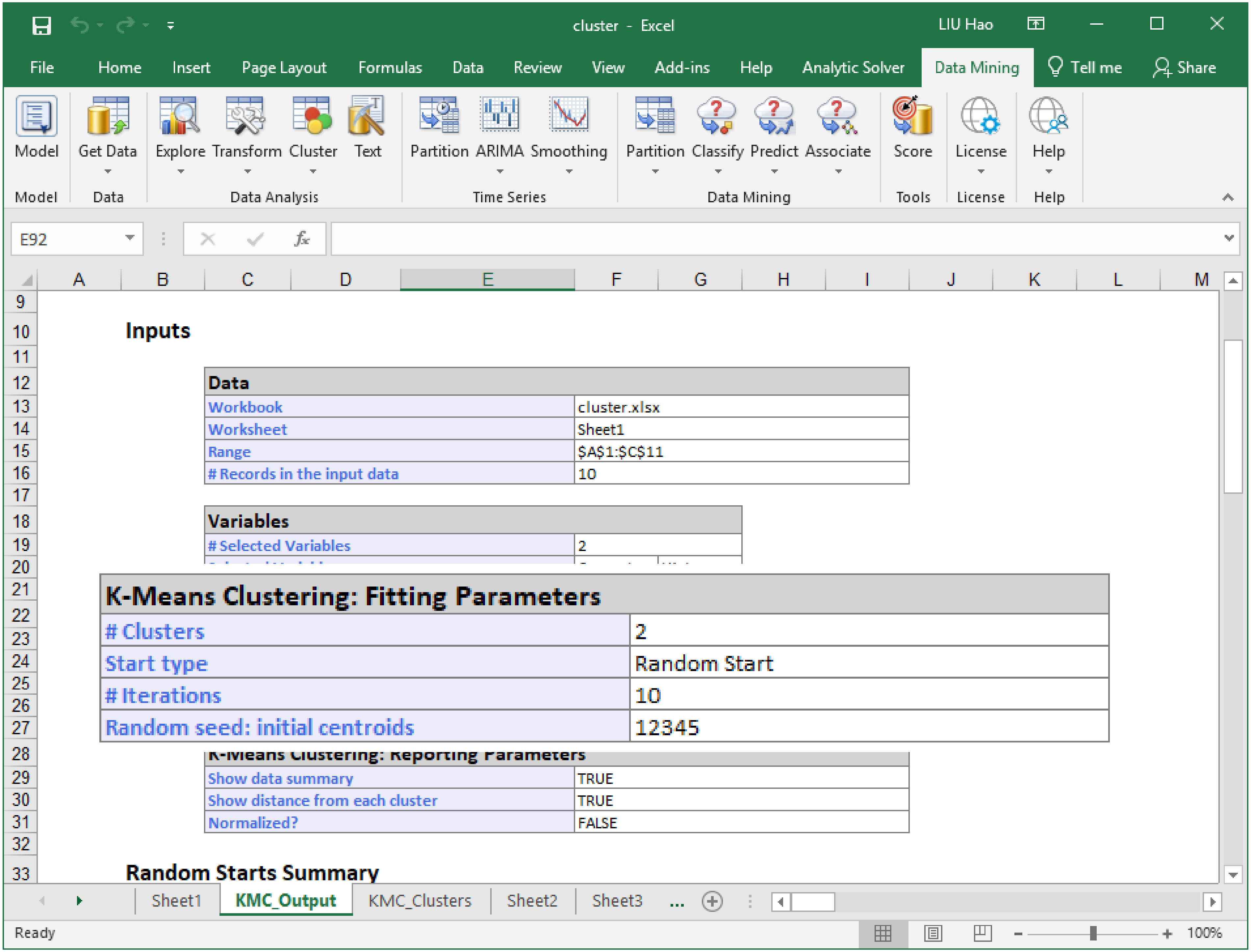
Data	
Workbook	cluster.xlsx
Worksheet	Sheet1
Range	$A$1:$C$11
# Records in the input data	10

Variables		
# Selected Variables	2	
Selected Variables	Computer	History

K-Means Clustering: Fitting Parameters	
# Clusters	2
Start type	Random Start
# Iterations	10
Random seed: initial centroids	12345

K-Means Clustering: Reporting Parameters	
Show data summary	TRUE
Show distance from each cluster	TRUE
Normalized?	FALSE

Random Starts Summary



Inputs

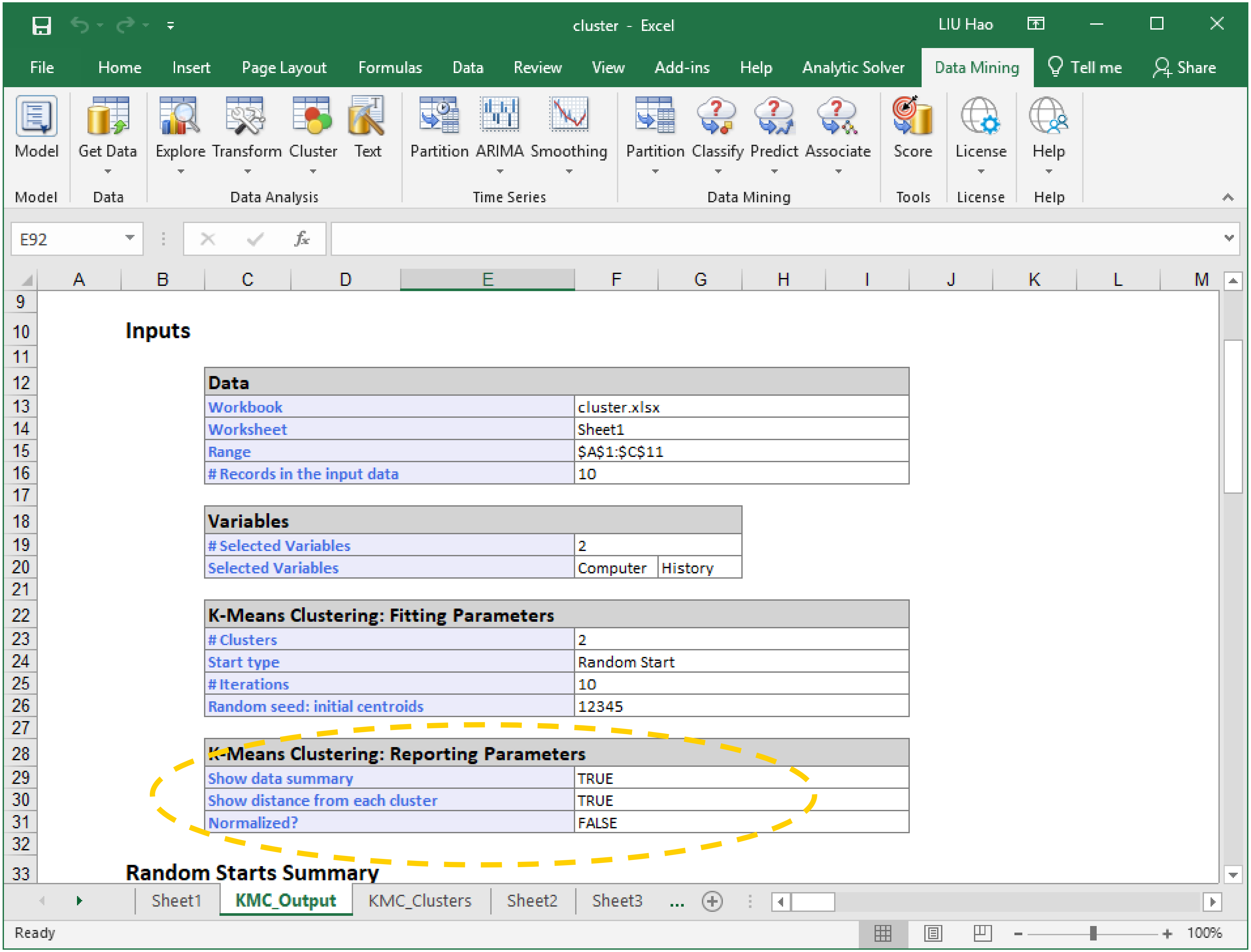
Data	
Workbook	cluster.xlsx
Worksheet	Sheet1
Range	$A$1:$C$11
# Records in the input data	10

Variables	
# Selected Variables	2

K-Means Clustering: Fitting Parameters	
# Clusters	2
Start type	Random Start
# Iterations	10
Random seed: initial centroids	12345

K-Means Clustering: Reporting Parameters	
Show data summary	TRUE
Show distance from each cluster	TRUE
Normalized?	FALSE

Random Starts Summary



Inputs

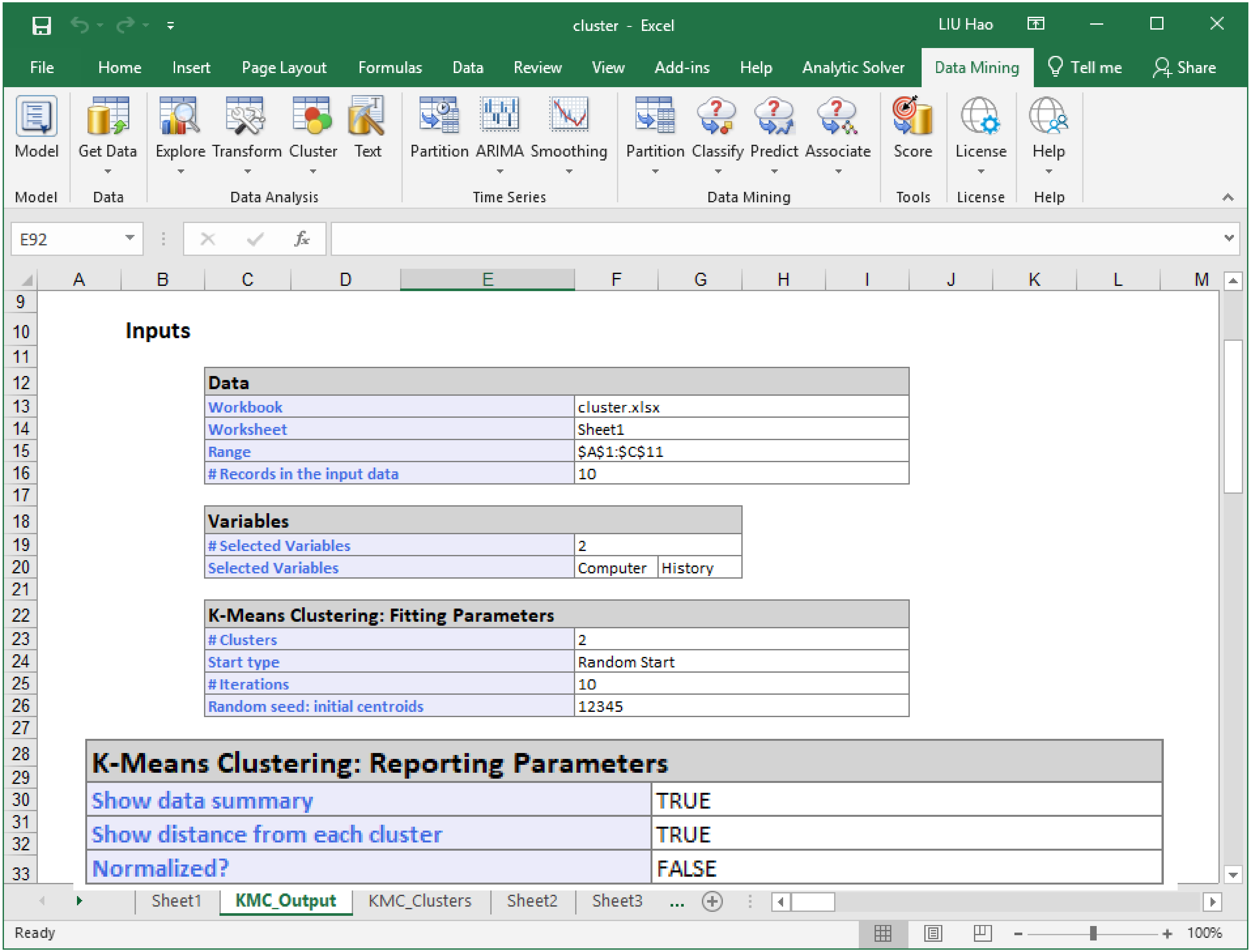
Data	
Workbook	cluster.xlsx
Worksheet	Sheet1
Range	$A$1:$C$11
# Records in the input data	10

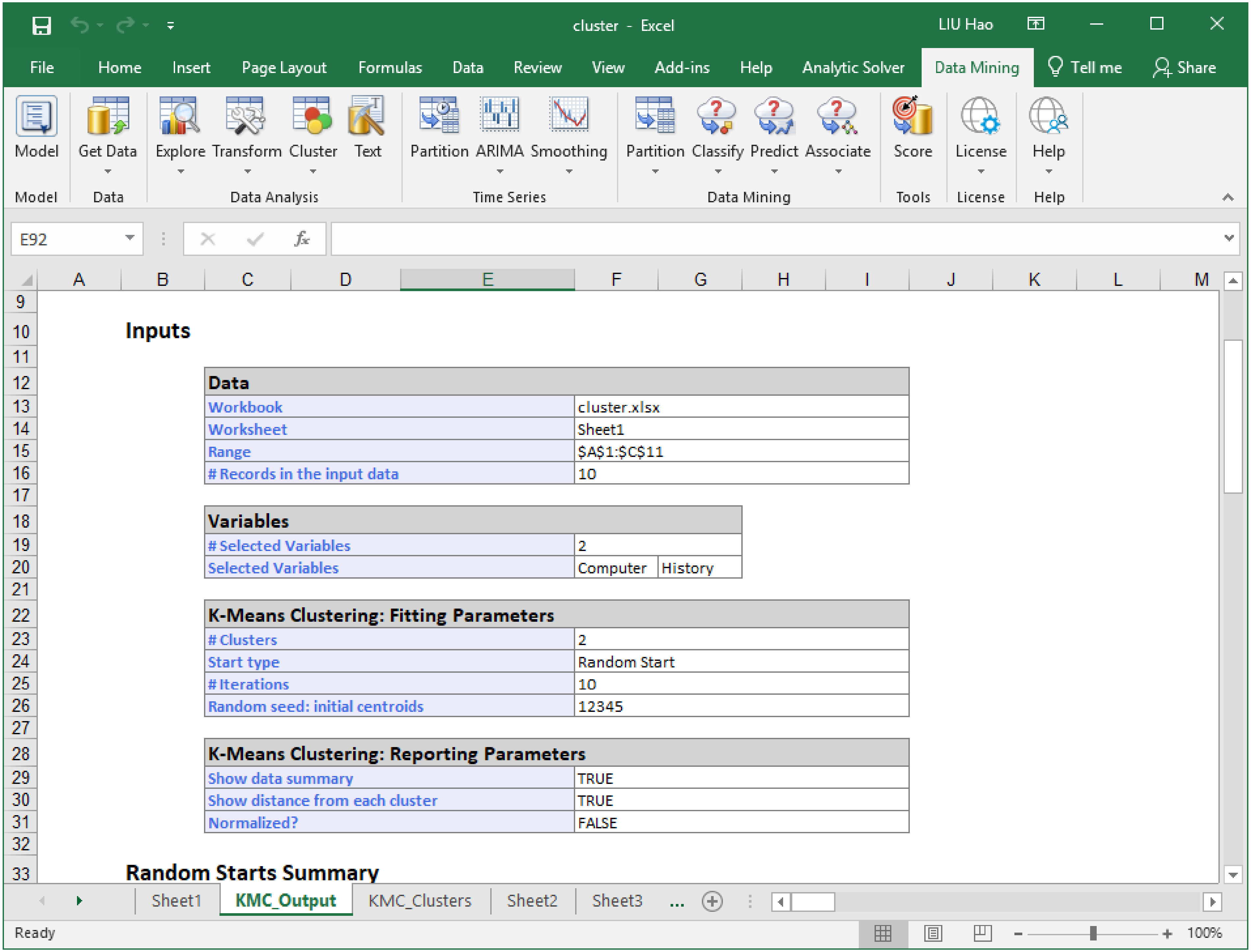
Variables		
# Selected Variables	2	
Selected Variables	Computer	History

K-Means Clustering: Fitting Parameters	
# Clusters	2
Start type	Random Start
# Iterations	10
Random seed: initial centroids	12345

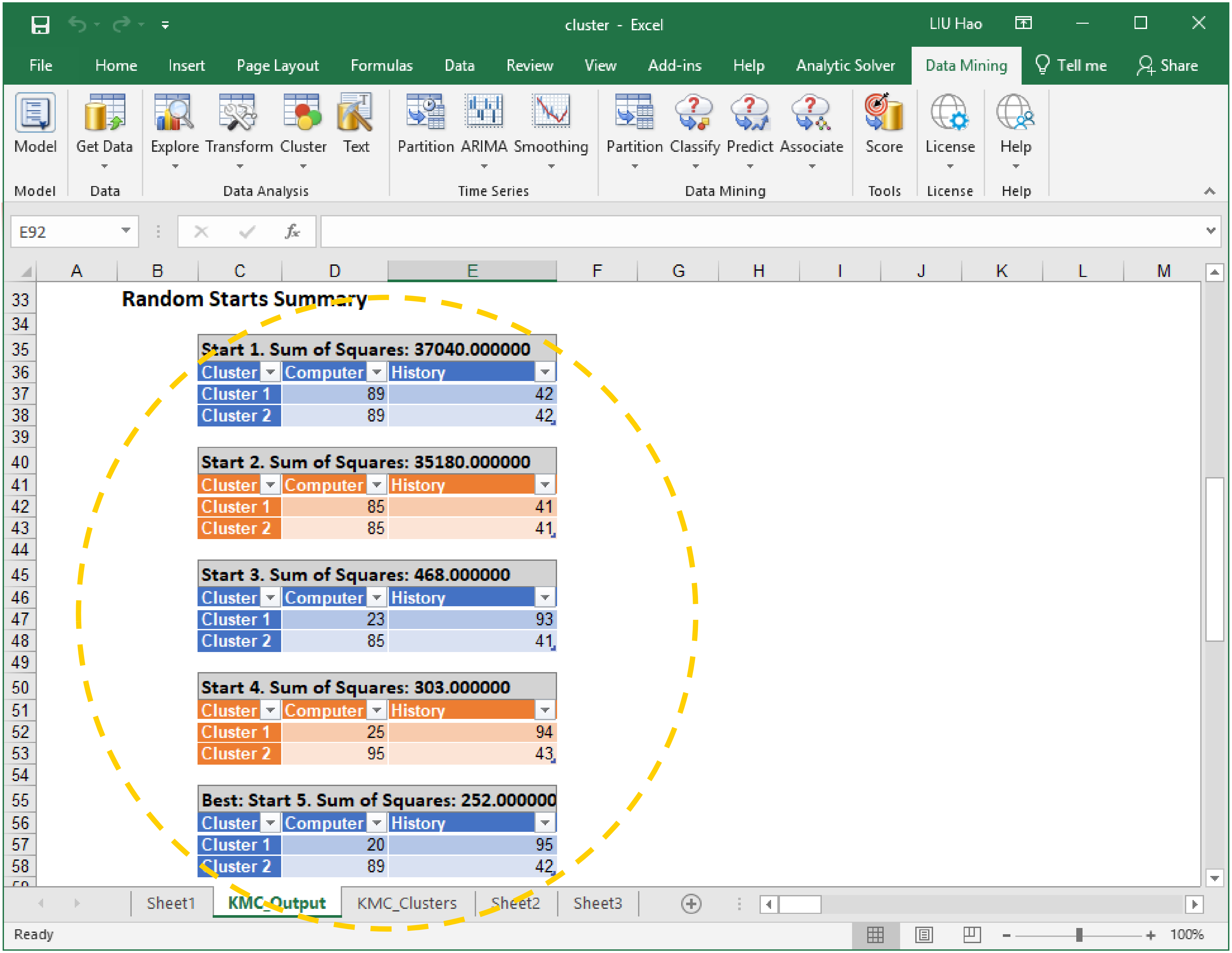
K-Means Clustering: Reporting Parameters	
Show data summary	TRUE
Show distance from each cluster	TRUE
Normalized?	FALSE

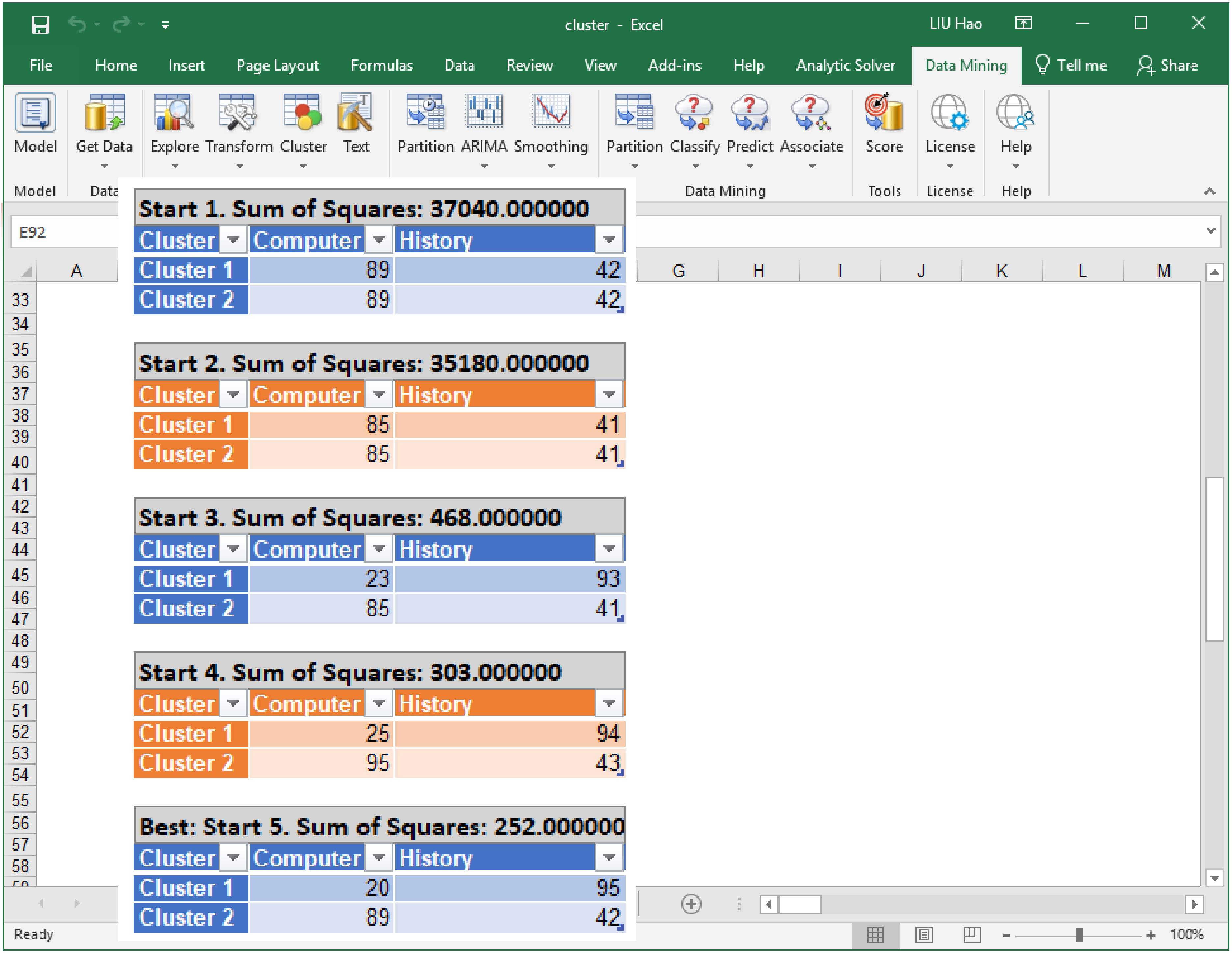
Random Starts Summary


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Inputs
Data
Workbook
cluster.xlsx
Worksheet
Sheet1
Range
$A$1:$C$11
Records in the input data
10
Variables
Selected Variables
2
Selected Variables
Computer
History
K-Means Clustering: Fitting Parameters
Clusters
2
Start type
Random Start
Iterations
10
Random seed: initial centroids
12345
K-Means Clustering: Reporting Parameters
Show data summary
TRUE
Show distance from each cluster
TRUE
Normalized?
FALSE
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Inputs
Data
Workbook
cluster.xlsx
Worksheet
Sheet1
Range
$A$1:$C$11
Records in the input data
10
Variables
Selected Variables
2
Selected Variables
Computer
History
K-Means Clustering: Fitting Parameters
Clusters
2
Start type
Random Start
Iterations
10
Random seed: initial centroids
12345
K-Means Clustering: Reporting Parameters
Show data summary
TRUE
Show distance from each cluster
TRUE
Normalized?
FALSE
Random Starts Summary
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Random Starts Summary
Start 1. Sum of Squares: 37040.000000
Cluster Computer History
Cluster 1 89 42
Cluster 2 89 42
Start 2. Sum of Squares: 35180.000000
Cluster Computer History
Cluster 1 85 41
Cluster 2 85 41
Start 3. Sum of Squares: 468.000000
Cluster Computer History
Cluster 1 23 93
Cluster 2 85 41
Start 4. Sum of Squares: 303.000000
Cluster Computer History
Cluster 1 25 94
Cluster 2 95 43
Best: Start 5. Sum of Squares: 252.000000
Cluster Computer History
Cluster 1 20 95
Cluster 2 89 42
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



Start 1. Sum of Squares: 37040.000000

Cluster	Computer	History
Cluster 1	89	42
Cluster 2	89	42

Start 2. Sum of Squares: 35180.000000

Cluster	Computer	History
Cluster 1	85	41
Cluster 2	85	41

Start 3. Sum of Squares: 468.000000

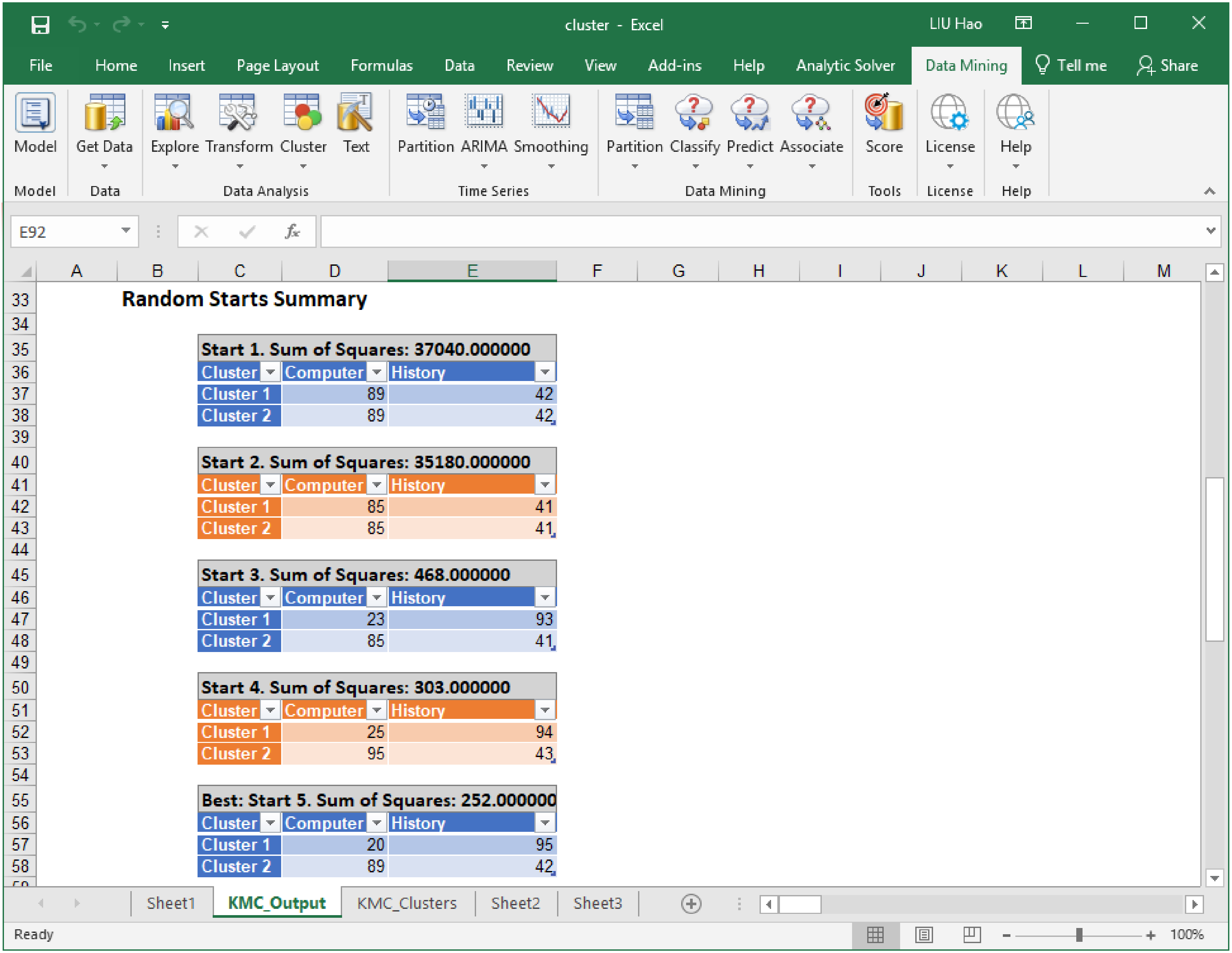
Cluster	Computer	History
Cluster 1	23	93
Cluster 2	85	41

Start 4. Sum of Squares: 303.000000

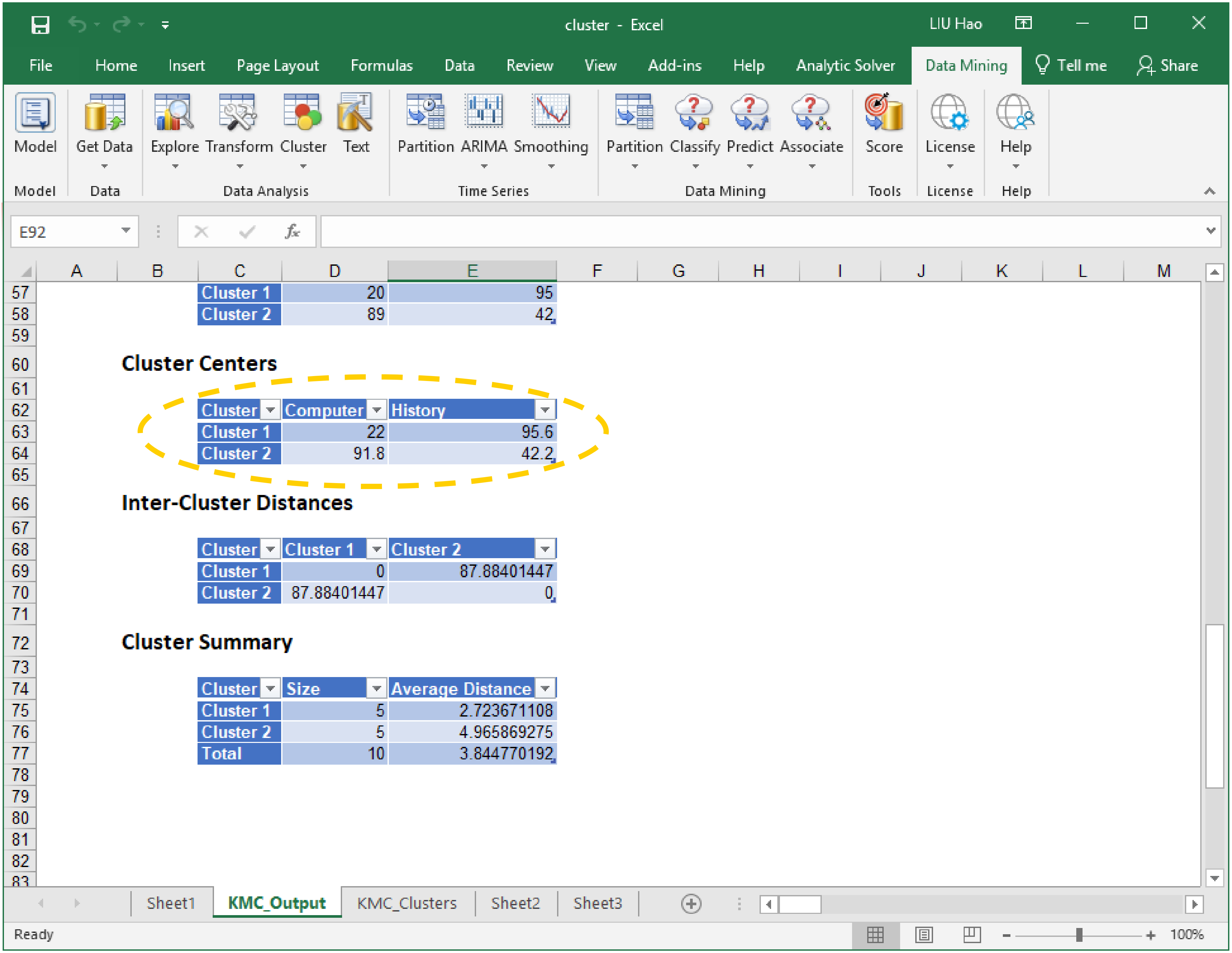
Cluster	Computer	History
Cluster 1	25	94
Cluster 2	95	43

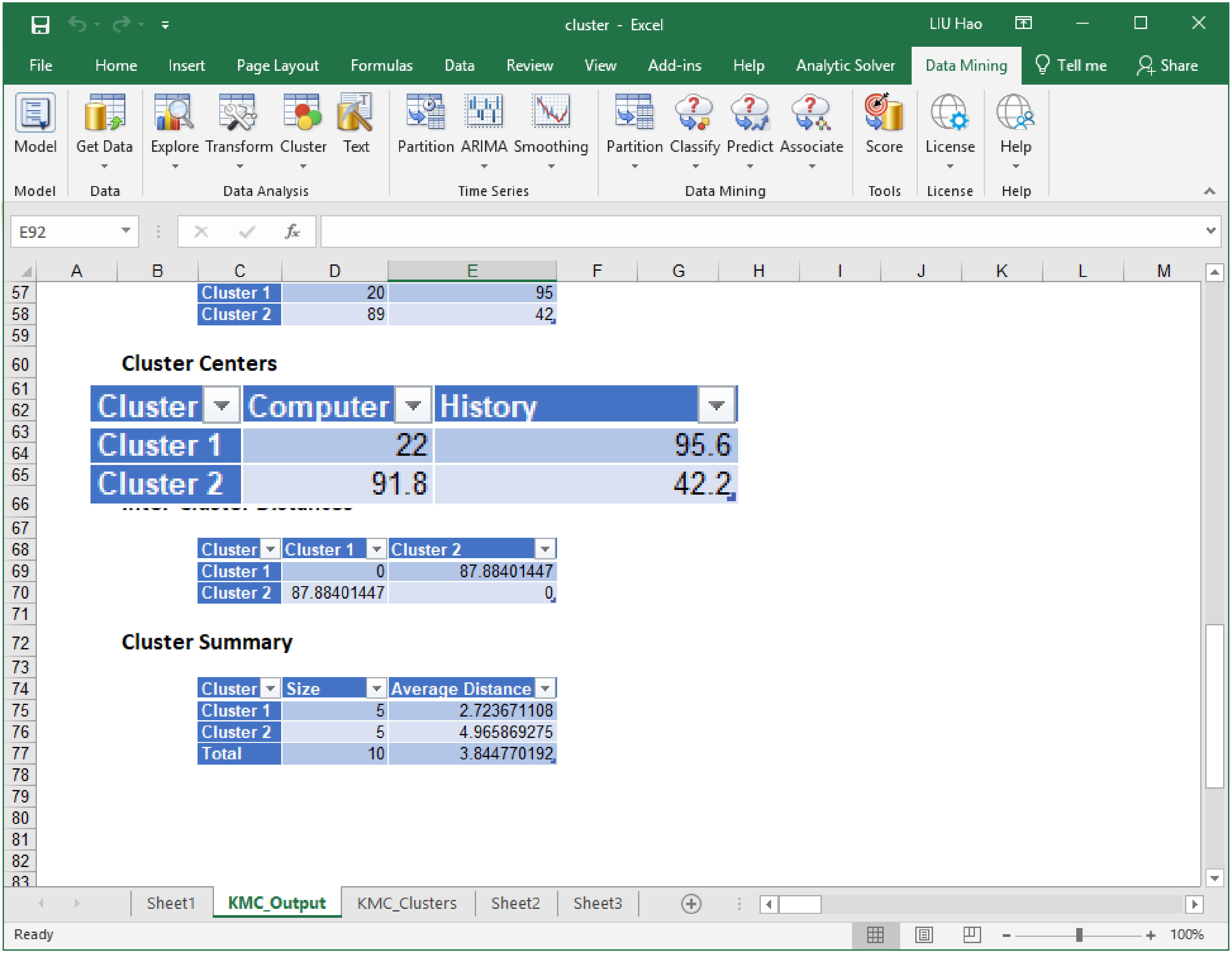
Best: Start 5. Sum of Squares: 252.000000

Cluster	Computer	History
Cluster 1	20	95
Cluster 2	89	42



cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Random Starts Summary
Start 1. Sum of Squares: 37040.000000
Cluster Computer History
Cluster 1 89 42
Cluster 2 89 42
Start 2. Sum of Squares: 35180.000000
Cluster Computer History
Cluster 1 85 41
Cluster 2 85 41
Start 3. Sum of Squares: 468.000000
Cluster Computer History
Cluster 1 23 93
Cluster 2 85 41
Start 4. Sum of Squares: 303.000000
Cluster Computer History
Cluster 1 25 94
Cluster 2 95 43
Best: Start 5. Sum of Squares: 252.000000
Cluster Computer History
Cluster 1 20 95
Cluster 2 89 42
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Cluster 1 20 95
Cluster 2 89 42
Cluster Centers
Cluster Computer History
Cluster 1 22 95.6
Cluster 2 91.8 42.2
Inter-Cluster Distances
Cluster Cluster 1 Cluster 2
Cluster 1 0 87.88401447
Cluster 2 87.88401447 0
Cluster Summary
Cluster Size Average Distance
Cluster 1 5 2.723671108
Cluster 2 5 4.965869275
Total 10 3.844770192
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



Cluster 1	20	95
Cluster 2	89	42

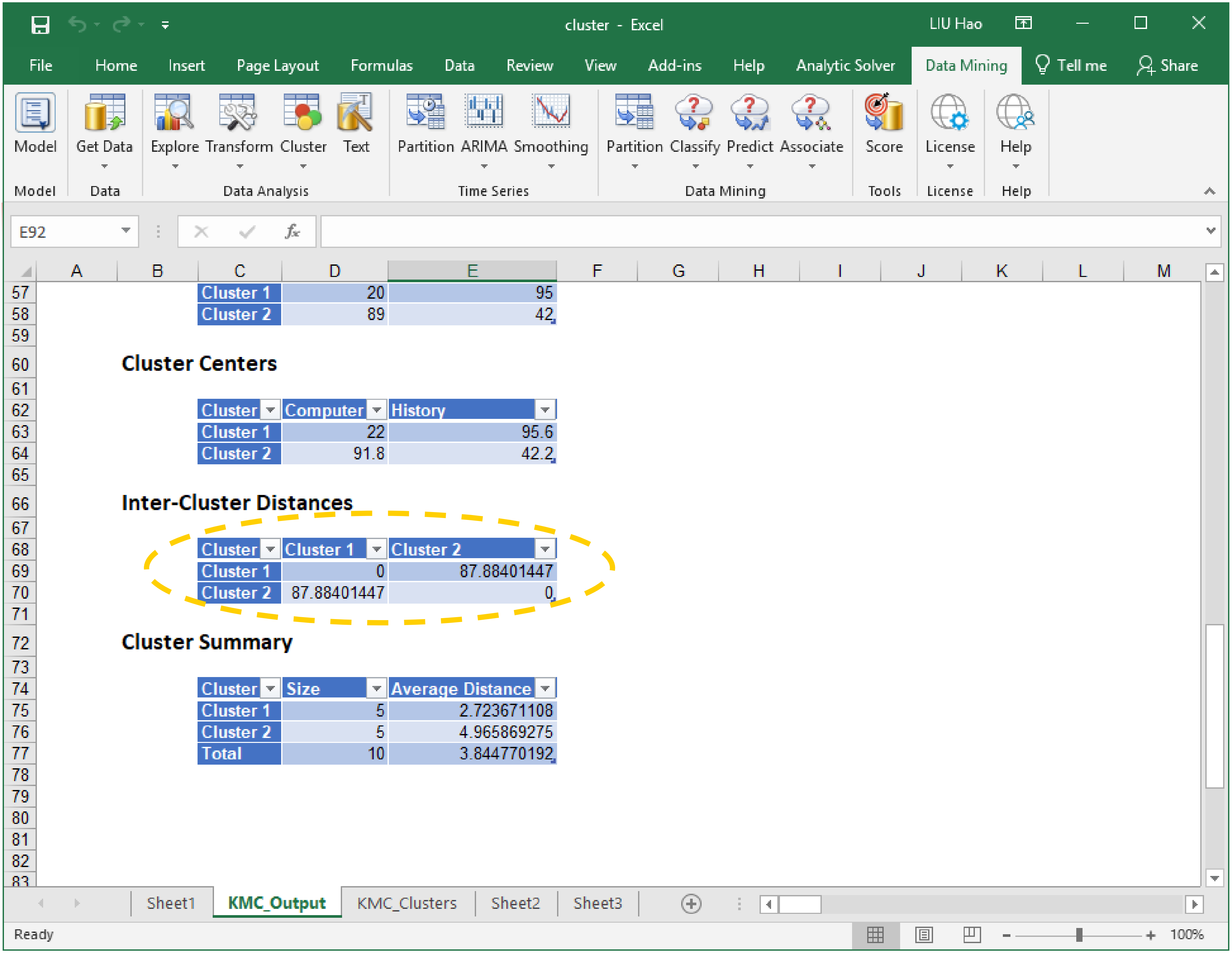
Cluster Centers

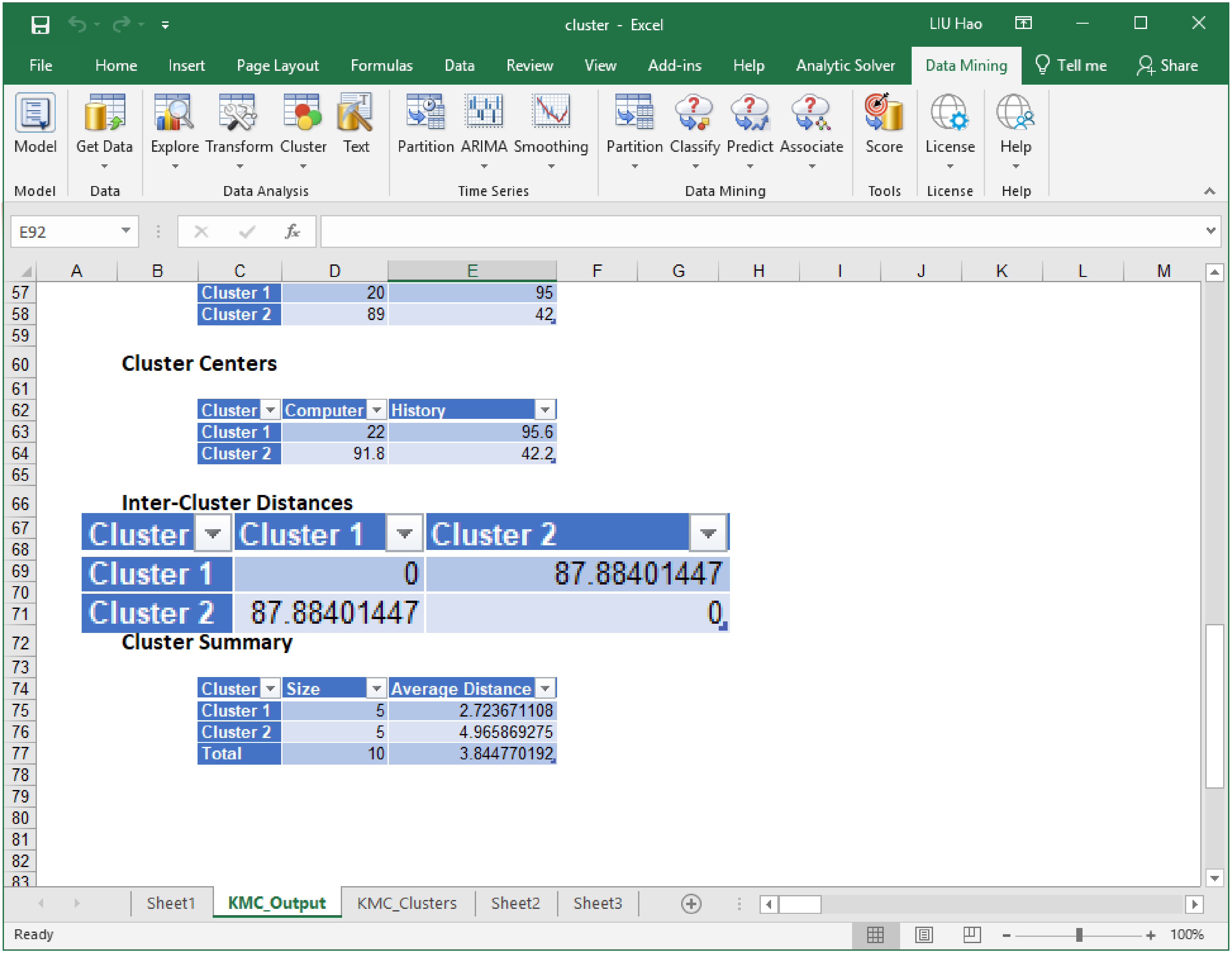
Cluster	Computer	History
Cluster 1	22	95.6
Cluster 2	91.8	42.2

Cluster	Cluster 1	Cluster 2
Cluster 1	0	87.88401447
Cluster 2	87.88401447	0

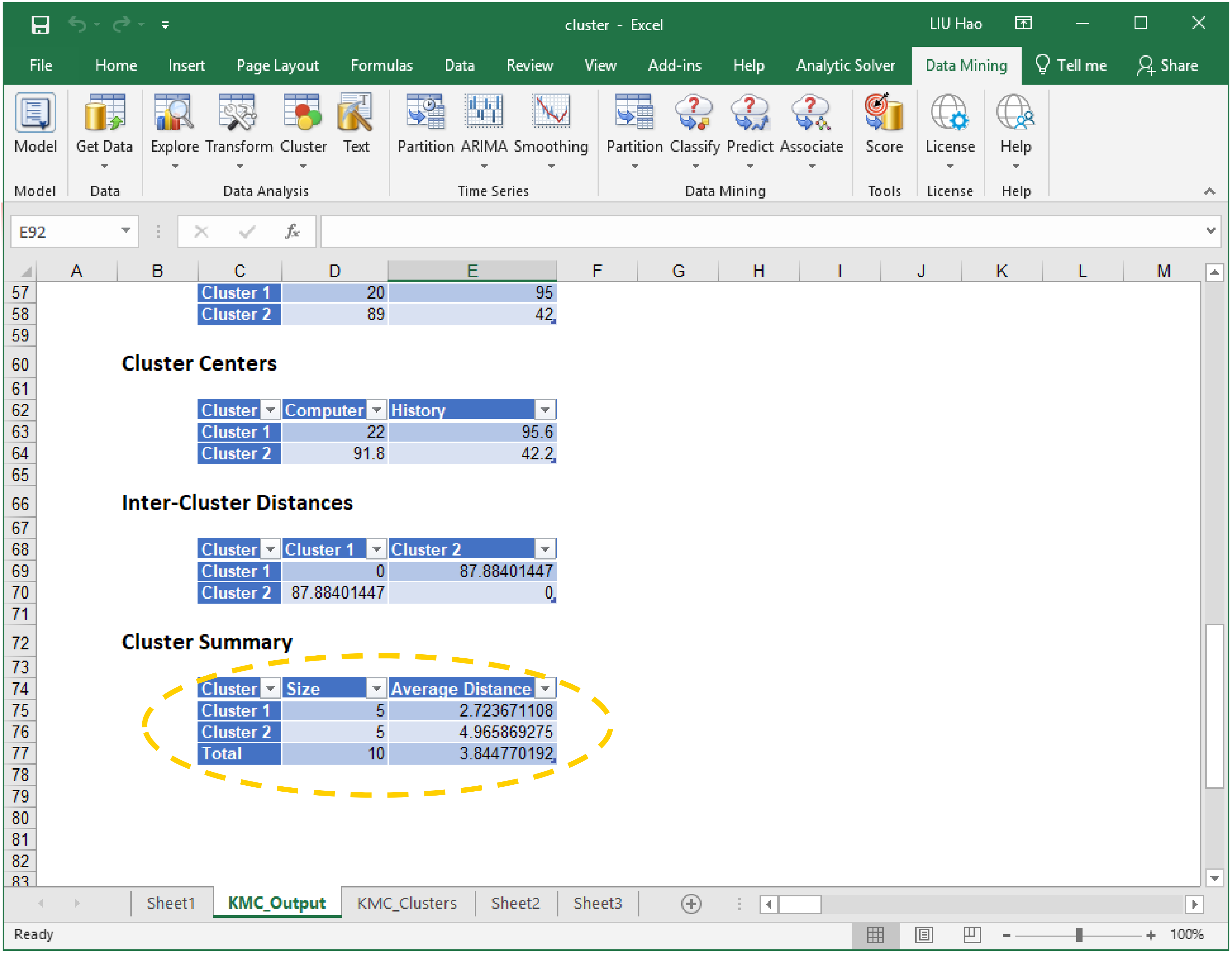
Cluster Summary

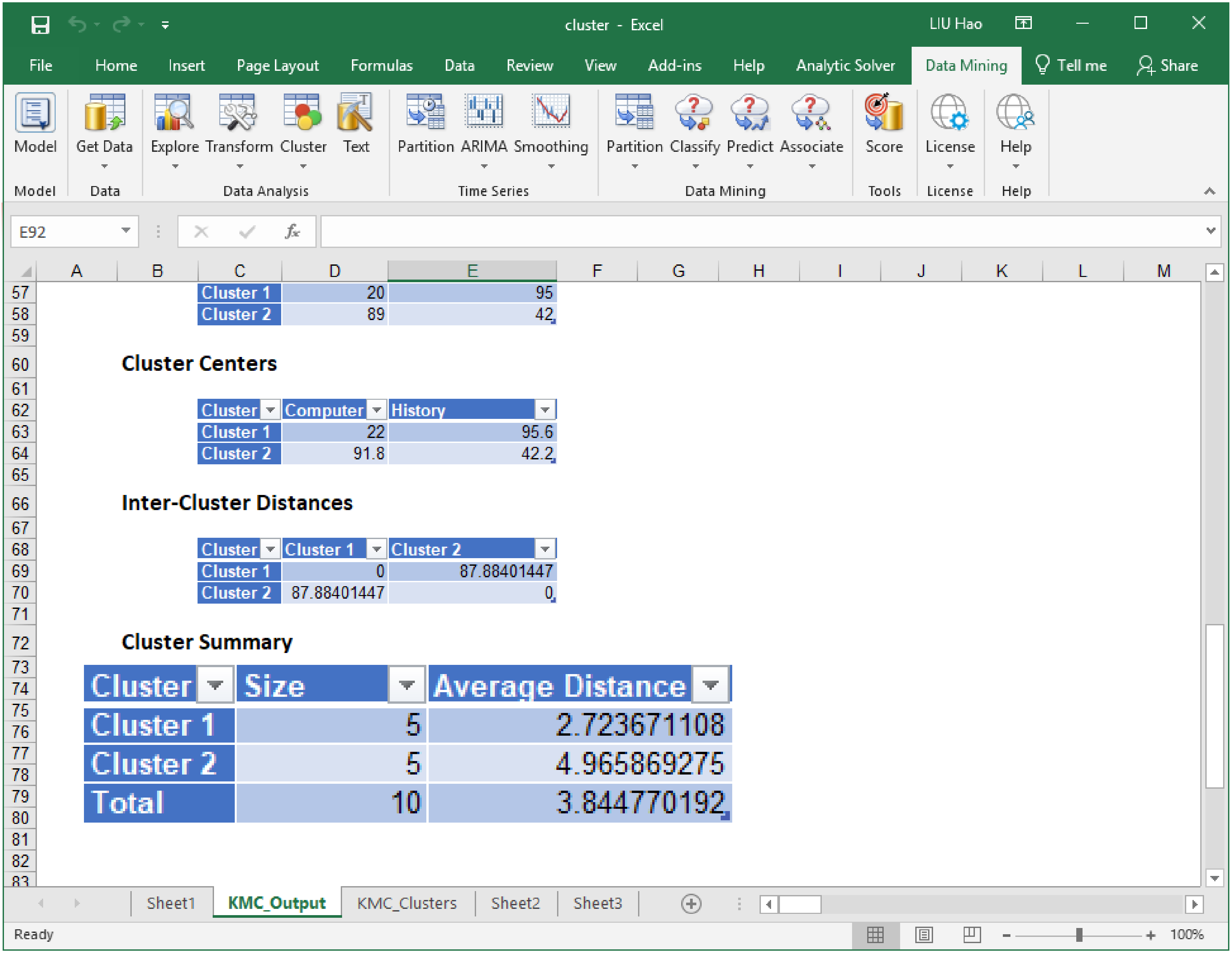
Cluster	Size	Average Distance
Cluster 1	5	2.723671108
Cluster 2	5	4.965869275
Total	10	3.844770192


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Cluster 1 20 95
Cluster 2 89 42
Cluster Centers
Cluster Computer History
Cluster 1 22 95.6
Cluster 2 91.8 42.2
Inter-Cluster Distances
Cluster Cluster 1 Cluster 2
Cluster 1 0 87.88401447
Cluster 2 87.88401447 0
Cluster Summary
Cluster Size Average Distance
Cluster 1 5 2.723671108
Cluster 2 5 4.965869275
Total 10 3.844770192
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Cluster 1 20 95
Cluster 2 89 42
Cluster Centers
Cluster Computer History
Cluster 1 22 95.6
Cluster 2 91.8 42.2
Inter-Cluster Distances
Cluster Cluster 1 Cluster 2
Cluster 1 0 87.88401447
Cluster 2 87.88401447 0
Cluster Summary
Cluster Size Average Distance
Cluster 1 5 2.723671108
Cluster 2 5 4.965869275
Total 10 3.844770192
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Cluster 1 20 95
Cluster 2 89 42
Cluster Centers
Cluster Computer History
Cluster 1 22 95.6
Cluster 2 91.8 42.2
Inter-Cluster Distances
Cluster Cluster 1 Cluster 2
Cluster 1 0 87.88401447
Cluster 2 87.88401447 0
Cluster Summary
Cluster Size Average Distance
Cluster 1 5 2.723671108
Cluster 2 5 4.965869275
Total 10 3.844770192
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



Cluster 1	20	95
Cluster 2	89	42

Cluster Centers

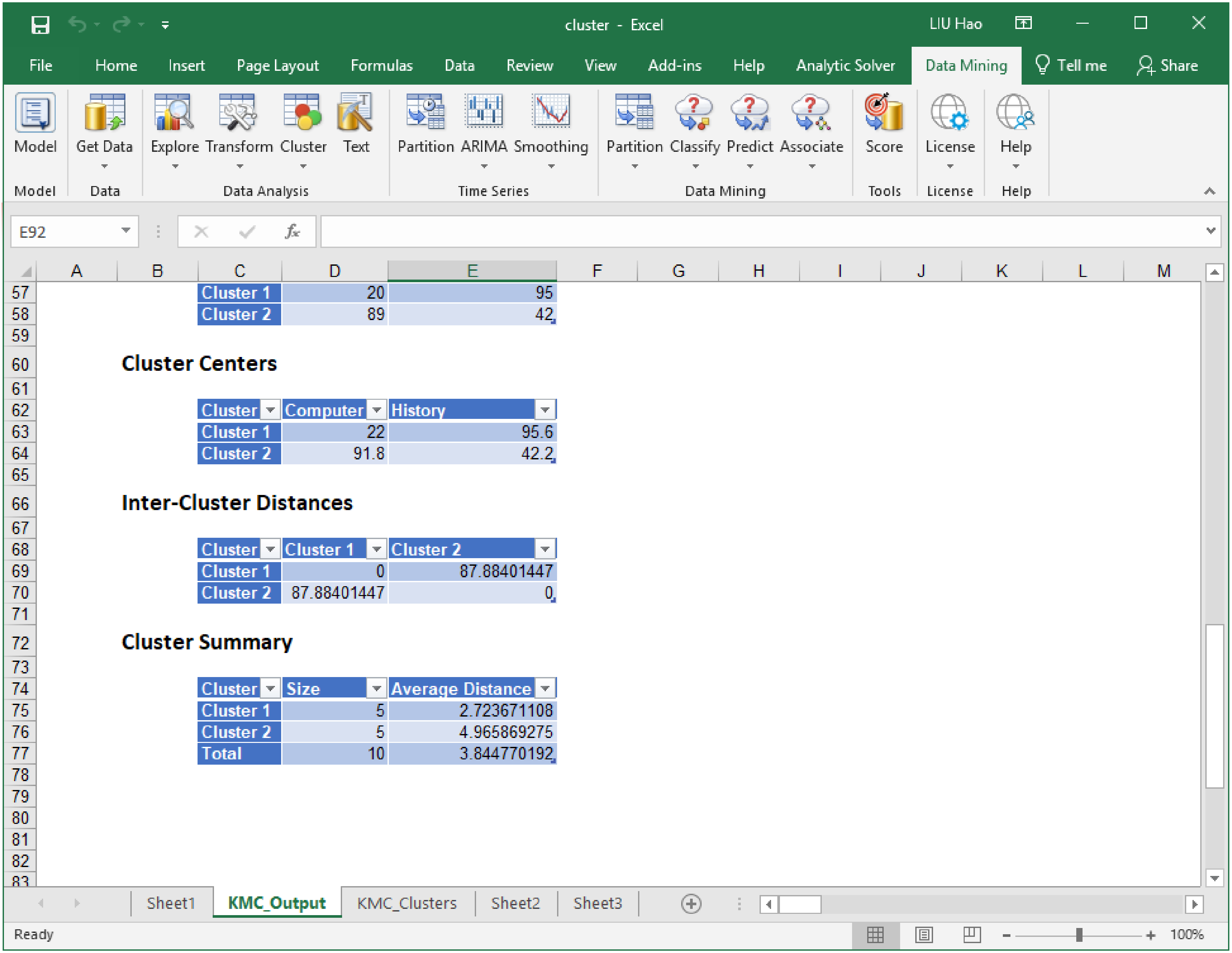
Cluster	Computer	History
Cluster 1	22	95.6
Cluster 2	91.8	42.2

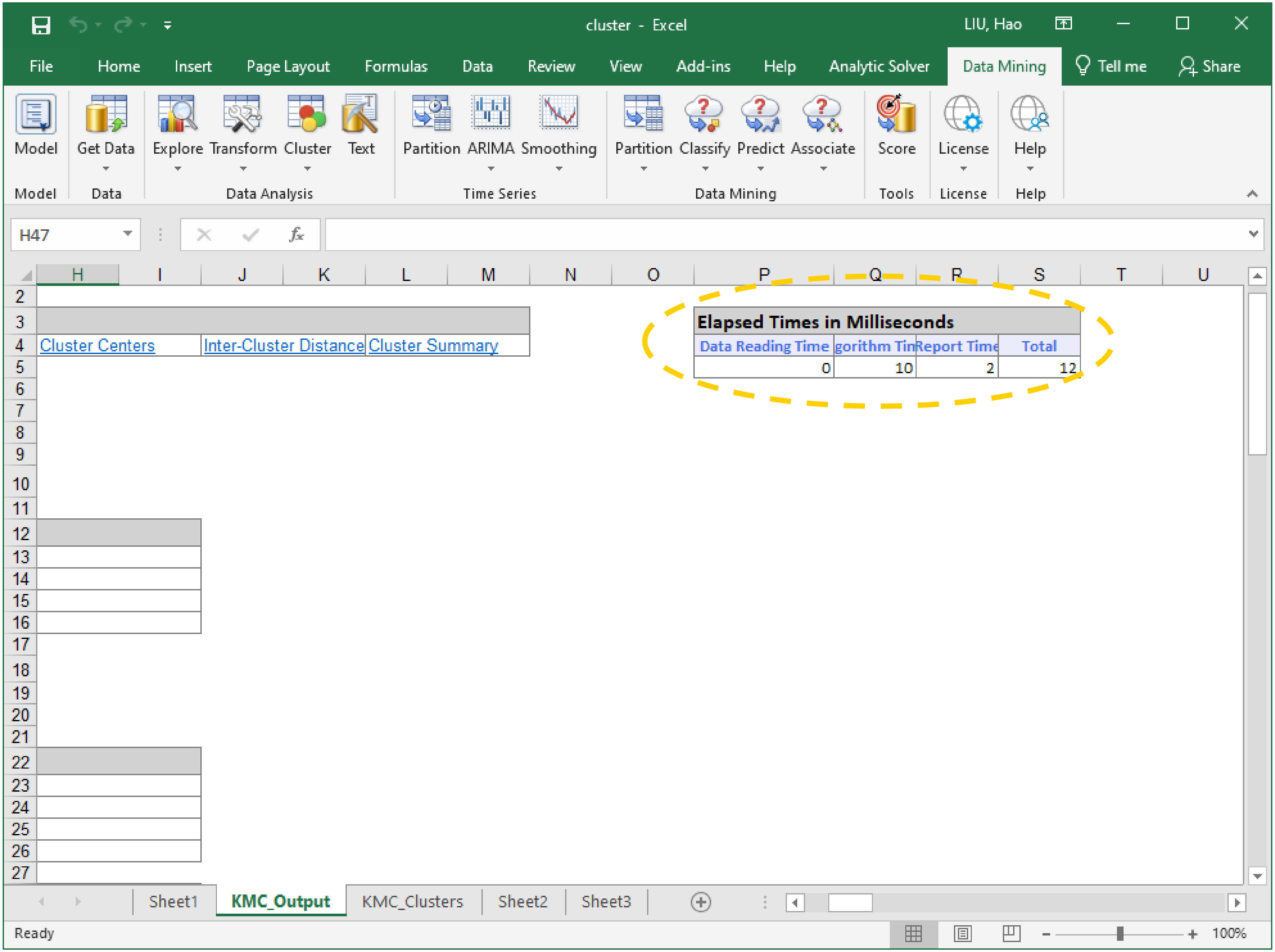
Inter-Cluster Distances

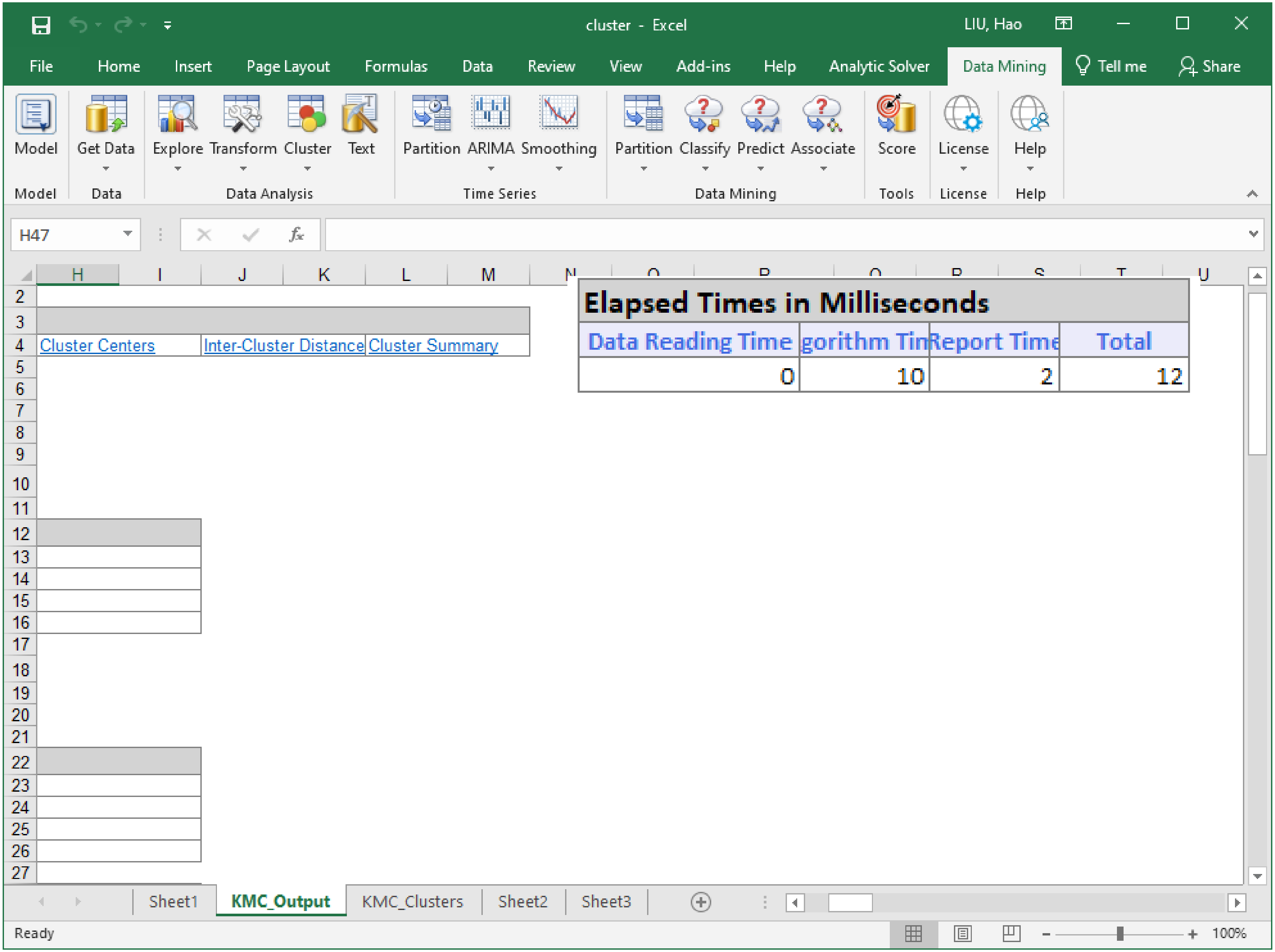
Cluster	Cluster 1	Cluster 2
Cluster 1	0	87.88401447
Cluster 2	87.88401447	0

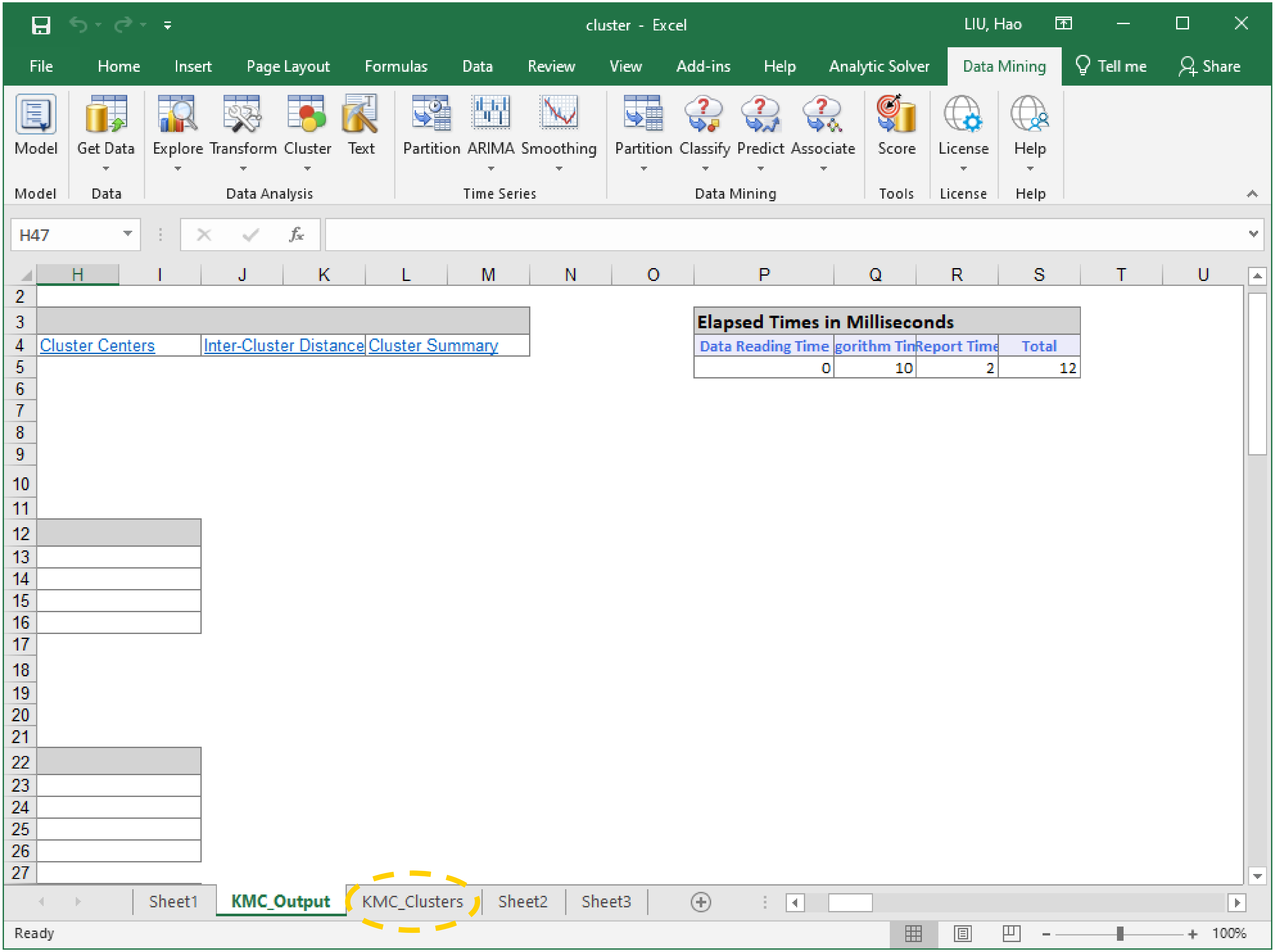
Cluster Summary

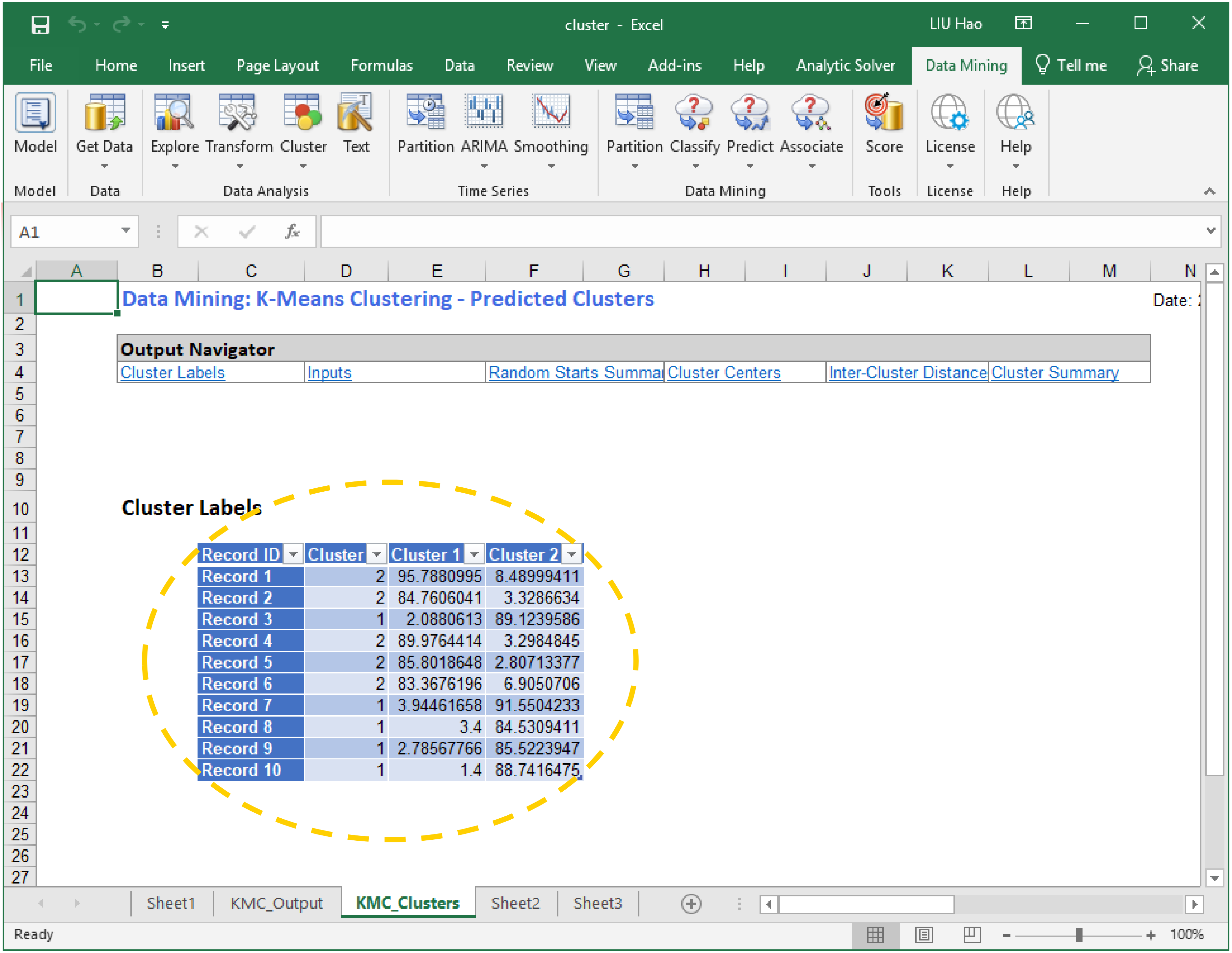
Cluster	Size	Average Distance
Cluster 1	5	2.723671108
Cluster 2	5	4.965869275
Total	10	3.844770192


cluster - Excel
LIU Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
E92
Cluster 1 20 95
Cluster 2 89 42
Cluster Centers
Cluster Computer History
Cluster 1 22 95.6
Cluster 2 91.8 42.2
Inter-Cluster Distances
Cluster Cluster 1 Cluster 2
Cluster 1 0 87.88401447
Cluster 2 87.88401447 0
Cluster Summary
Cluster Size Average Distance
Cluster 1 5 2.723671108
Cluster 2 5 4.965869275
Total 10 3.844770192
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%

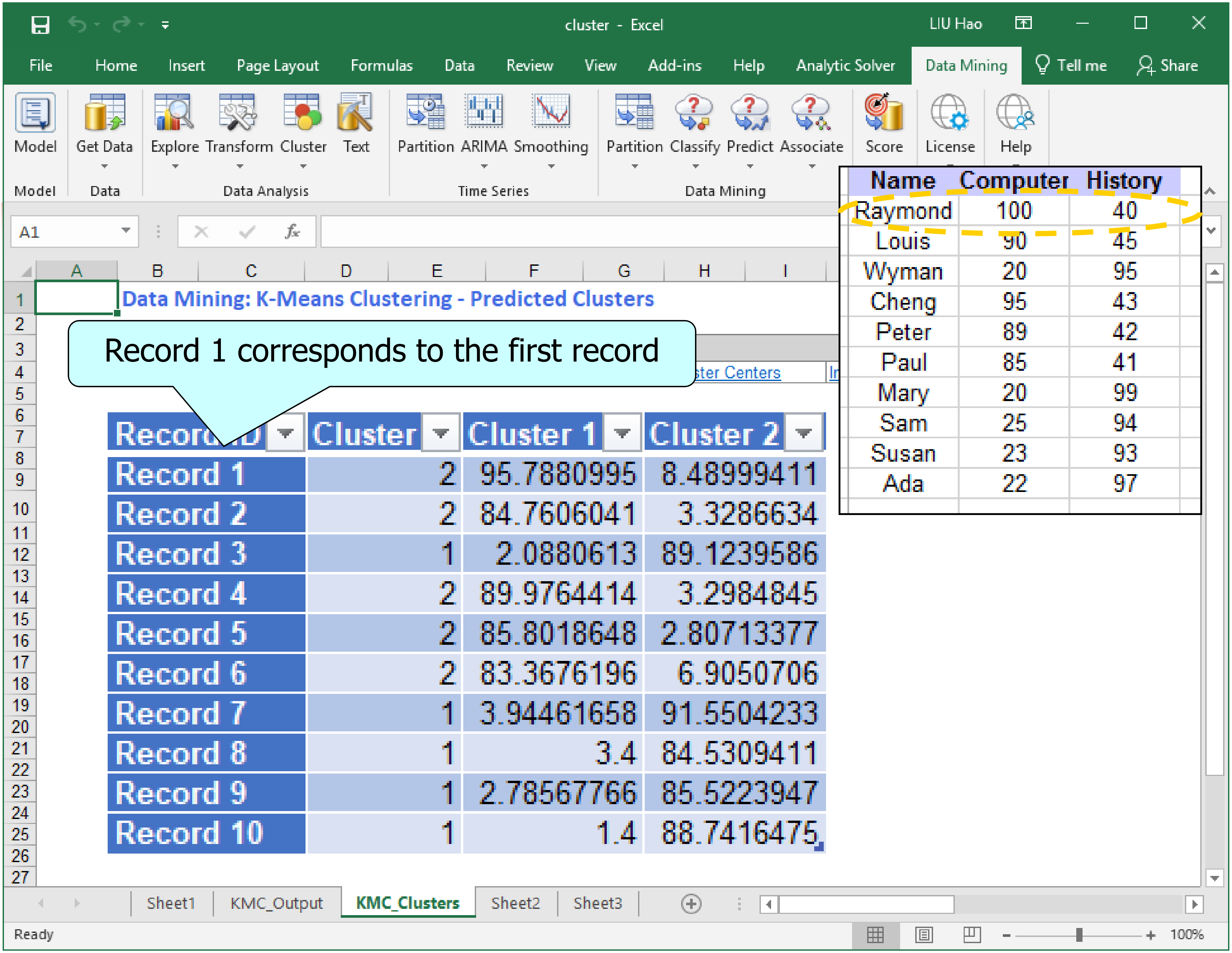

cluster - Excel
LIU, Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
H47
Cluster Centers
Inter-Cluster Distance
Cluster Summary
Elapsed Times in Milliseconds
Data Reading Time
gorithm Tin
Report Time
Total
0
10
2
12
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU, Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
H47
Cluster Centers
Inter-Cluster Distance
Cluster Summary
Elapsed Times in Milliseconds
Data Reading Time
gorithm Tin
Report Time
Total
0
10
2
12
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%


cluster - Excel
LIU, Hao
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
H47
Cluster Centers
Inter-Cluster Distance
Cluster Summary
Elapsed Times in Milliseconds
Data Reading Time
gorithm Tin
Report Time
Total
0
10
2
12
Sheet1
KMC_Output
KMC_Clusters
Sheet2
Sheet3
Ready
100%



Record ID	Cluster	Cluster 1	Cluster 2
Record 1	2	95.7880995	8.48999411
Record 2	2	84.7606041	3.3286634
Record 3	1	2.0880613	89.1239586
Record 4	2	89.9764414	3.2984845
Record 5	2	85.8018648	2.80713377
Record 6	2	83.3676196	6.9050706
Record 7	1	3.94461658	91.5504233
Record 8	1	3.4	84.5309411
Record 9	1	2.78567766	85.5223947
Record 10	1	1.4	88.7416475



Record ID	Cluster	Cluster 1	Cluster 2
Record 1	2	95.7880995	8.48999411
Record 2	2	84.7606041	3.3286634
Record 3	1	2.0880613	89.1239586
Record 4	2	89.9764414	3.2984845
Record 5	2	85.8018648	2.80713377
Record 6	2	83.3676196	6.9050706
Record 7	1	3.94461658	91.5504233
Record 8	1	3.4	84.5309411
Record 9	1	2.78567766	85.5223947
Record 10	1	1.4	88.7416475

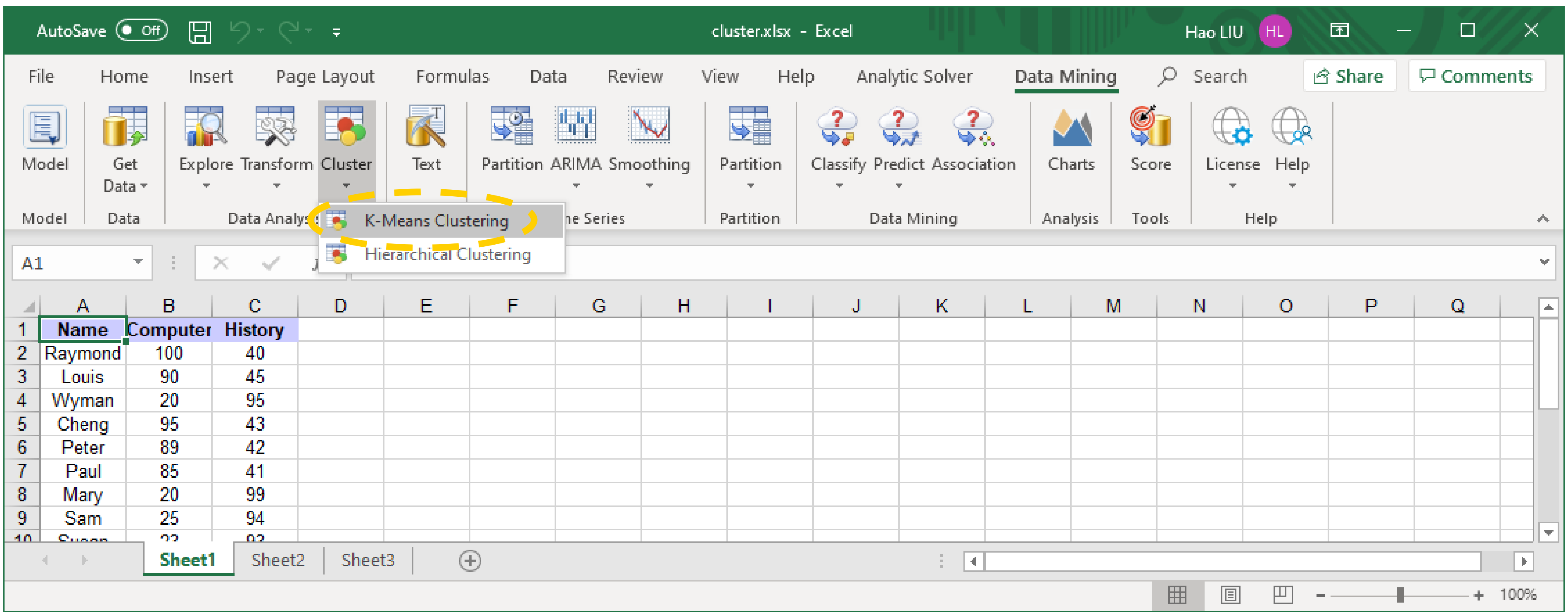
Name	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
Cheng	95	43
Peter	89	42
Paul	85	41
Mary	20	99
Sam	25	94
Susan	23	93
Ada	22	97

How to use the data mining tool

- We have the following 2 versions.
  - XLMiner Desktop (installed in either the CSE lab machine or your computer)
  - XLMiner Cloud (installed as a plugin in your Office 365 Excel)

How to use the data mining tool (XLMiner Cloud)

- The way of opening K-means Clustering in XLMiner Cloud plugin in your Office 365 Excel
  - “Data Mining” Tag → Cluster → K-Means Clustering



How to use the data mining tool (XLMiner Cloud)

- The steps of performing “k-means clustering” in XLMiner Cloud is similar to the steps in XLMiner Desktop.
- The output format and the clustering result of XLMiner Cloud are the same as that from XLMiner Desktop.