# Ryan Timbrook

Data Science 450, Spring 2017
Date: 05/10/2017
Assignment 3

## Description: Clustering

Use K-means clustering algorithm to cluster user sessions of an online shopping site into segments.

Try different clustering runs with various numbers of clusters (e.g., between 4 and 8), and select the result set(s) that seem to best answer as many of the following questions as possible

**Question 1**:
If a new user is observed to access the following pages: Home => Search => Prod_B

Q1.a: According to your clusters, what other product should be recommended to this user?
**Answer Results**:

- Prod_A should be offered. It was selected observed 7 times in conjunction with Prod_B, 4 of which lead to purchases. See table 1.1.a,1.2.a,1.3.a below for data table information.

Q1.b: What if the new user has accessed the following sequence instead: Products => Prod_C?
**Answer Results:**

- Prod_A should be offered, 5 of the 16 page visits lead to purchases
- Prod_B should be offered, 12 of the 35 page visits lead to purchases
  See table 1.1.b, 1.2.b, 1.3.b below for data table information.

**Question 2:**
Can clustering help us identify:

- casual browsers ("window shoppers")
- focused browsers (those who seem to know what products they are looking for)
- searchers (those using the search function to find items they want)?

If so, are any of these groups show a higher or lower propensity to make a purchase?

Table 1.1.a: Cluster 1; Optimal Cluster – k = 3

'Q1 Path Observations:5 of 24 Q1 Path Frequencey: 20.83%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | FALSE | 1 |
| 10 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE | 1 |
| 16 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 34 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | FALSE | 1 |

'Prod_A Observations of Q1 Path: 3 of 5'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE | 1 |
| 16 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |

'Prod_C Observations of Q1 Path: 3 of 5'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 34 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | FALSE | 1 |

Table 1.2.a: Cluster 2; Optimal Cluster – k = 3

'Q1 Path Observations:6 of 45 Q1 Path Frequencey: 13.33%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | TRUE | 2 |
| 8 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | TRUE | 2 |
| 9 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | TRUE | 2 |
| 11 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 15 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |

'Prod_A Observations of Q1 Path: 4 of 6'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | TRUE | 2 |
| 11 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 15 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |

'Prod_C Observations of Q1 Path: 3 of 6'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 15 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |

Table 1.3.a: Cluster 3; Optimal Cluster – k = 3

'Q1 Path Observations:1 of 31 Q1 Path Frequencey: 3.23%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | TRUE | 3 |

'Prod_A Observations of Q1 Path: 0 of 1'

| Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|

'Prod_C Observations of Q1 Path: 1 of 1'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | TRUE | 3 |

Table 1.1.b:

'Q1.b Path Observations:7 of 24 Q1.b Path Frequencey: 29.17%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | FALSE | 1 |
| 25 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 27 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 32 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | FALSE | 1 |
| 34 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 51 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE | 1 |

'Prod_A Observations of Q1.b Path: 5 of 7'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 27 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 32 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | FALSE | 1 |
| 51 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE | 1 |

'Prod_B Observations of Q1.b Path: 4 of 7'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 27 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | FALSE | 1 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | FALSE | 1 |
| 34 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | FALSE | 1 |

Table 1.2.b:

'Q1.b Path Observations:11 of 45 Q1.b Path Frequencey: 24.44%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 30 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 31 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 35 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 52 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 53 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 64 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 69 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | TRUE | 2 |
| 71 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 78 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

'Prod_A Observations of Q1.b Path: 9 of 11'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 31 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 35 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 52 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 53 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 64 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 71 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 78 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

'Prod_C Observations of Q1.b Path: 11 of 11'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 30 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| 31 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 35 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| 52 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| 53 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 64 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 69 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | TRUE | 2 |
| 71 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| 78 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

Table 1.3.b:

'Q1.b Path Observations:11 of 45 Q1.b Path Frequencey: 24.44%'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| **14** | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **30** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| **31** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **35** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **36** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| **52** | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **53** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **64** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **69** | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | TRUE | 2 |
| **71** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **78** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

'Prod_A Observations of Q1.b Path: 9 of 11'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| **30** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| **31** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **35** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **36** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| **52** | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **53** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **64** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **71** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **78** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

'Prod_C Observations of Q1.b Path: 11 of 11'

| | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | train | k3.cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| **14** | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **30** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | TRUE | 2 |
| **31** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **35** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **36** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | TRUE | 2 |
| **52** | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | TRUE | 2 |
| **53** | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **64** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **69** | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | TRUE | 2 |
| **71** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |
| **78** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | TRUE | 2 |

Table 2.1: Optimal number of clusters – k = 3 Plot

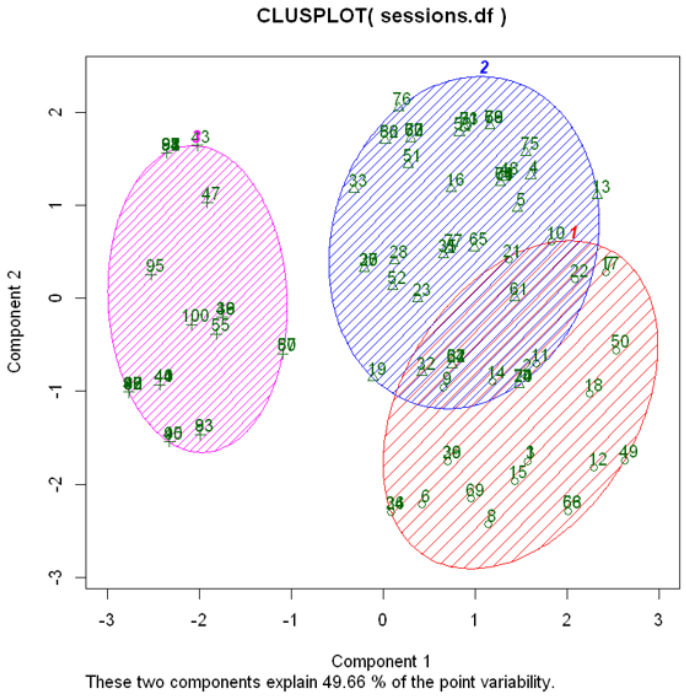



Table 2.2: Optimal number of clusters – k = 3 Data Summary

```
K-means clustering with 3 clusters of sizes 31, 11, 58

Cluster means:
      Home  Products     Search    Prod_A    Prod_B    Prod_C      Cart
1 0.32258065 0.8709677 0.54838710 0.2258065 0.9677419 0.5161290 0.1612903
2 0.09090909 1.0000000 0.09090909 0.0000000 1.0000000 0.8181818 1.0000000
3 0.84482759 0.5862069 0.43103448 0.7931034 0.2413793 0.3448276 0.7758621
   Purchase
1 0.0000000
2 0.9090909
3 0.5000000

Clustering vector:
  [1] 3 3 3 3 3 1 3 1 1 3 3 3 3 3 3 1 3 3 3 1 3 3 3 3 3 3 3 1 3 3 1 3 3 3 2 1 3 1 3
 [38] 1 1 1 1 1 2 1 1 1 2 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3
 [75] 3 3 3 3 3 1 3 3 1 2 1 1 1 2 1 1 2 2 1 2 2 1 2 2 1 1

Within cluster sum of squares by cluster:
[1] 36.258065  4.363636 93.724138
 (between_SS / total_SS =  29.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

| Group.1 | Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.32258065 | 0.8709677 | 0.54838710 | 0.2258065 | 0.9677419 | 0.5161290 | 0.1612903 | 0.0000000 |
| 2 | 0.09090909 | 1.0000000 | 0.09090909 | 0.0000000 | 1.0000000 | 0.8181818 | 1.0000000 | 0.9090909 |
| 3 | 0.84482759 | 0.5862069 | 0.43103448 | 0.7931034 | 0.2413793 | 0.3448276 | 0.7758621 | 0.5000000 |

| Home | Products | Search | Prod_A | Prod_B | Prod_C | Cart | Purchase | k3.cluster |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 3 |
| 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 3 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

Table 2.3: Obtimal number of clusters – k = 3

```
Among all indices:
==================
* 2 proposed  0 as the best number of clusters
* 1 proposed  1 as the best number of clusters
* 1 proposed  2 as the best number of clusters
* 10 proposed  3 as the best number of clusters
* 2 proposed  4 as the best number of clusters
* 1 proposed  7 as the best number of clusters
* 2 proposed  9 as the best number of clusters
* 7 proposed  10 as the best number of clusters

Conclusion
==========================
* According to the majority rule, the best number of clusters is  3 .
```

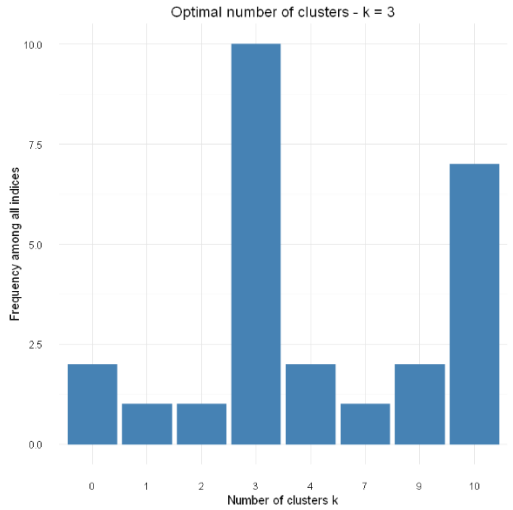


Table 3.1: Exploratory - Silhouette plot and Partion around mediods

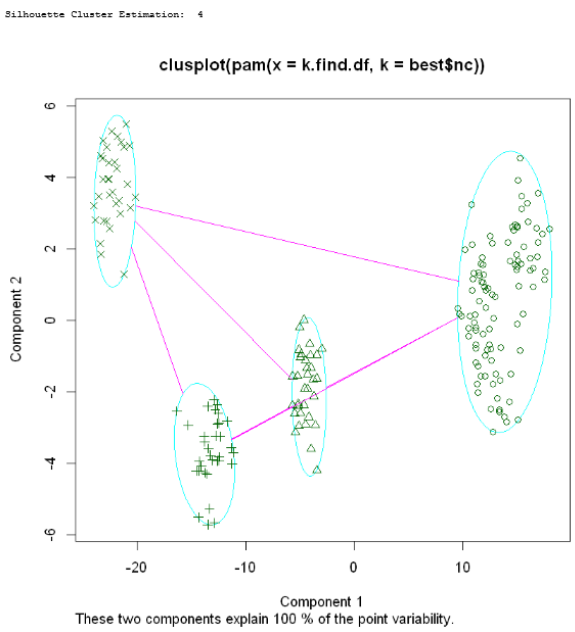



Table 3.2:

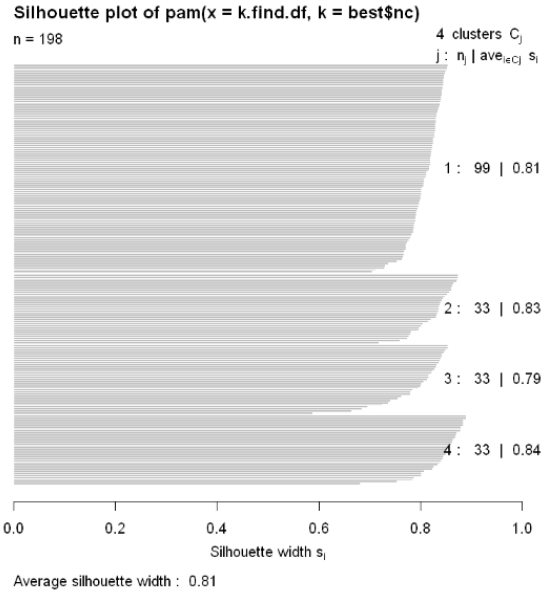



Table 3.3: Exploratory K-means Finding Elbow Plot

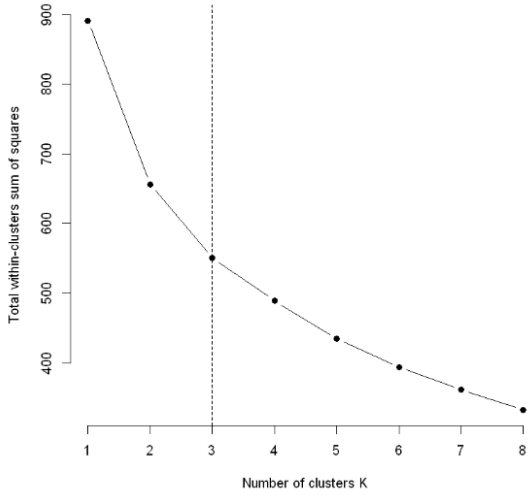


Table 4.1: Exploratory K-means Partitioning - k = 4

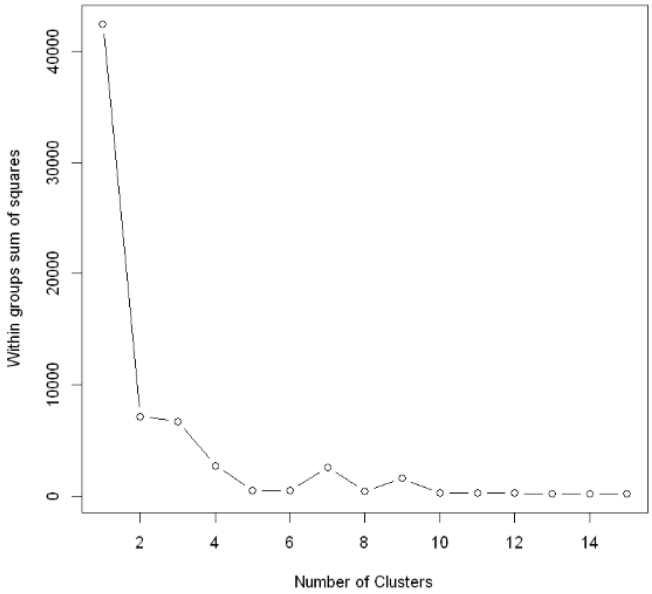
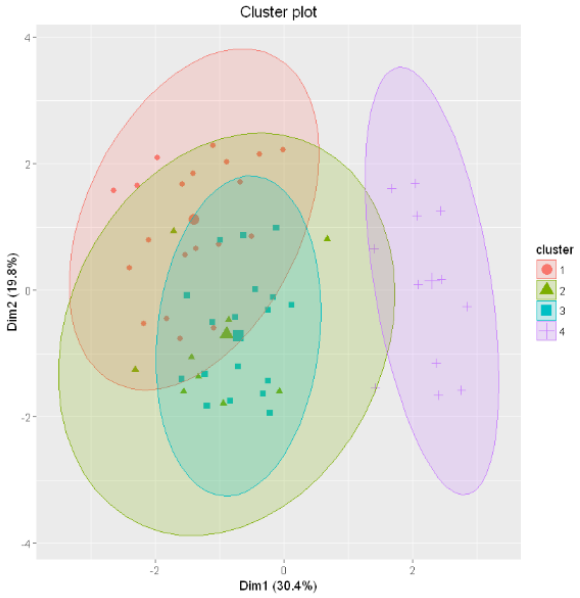



Table 4.2: Exploratory K-means Plot Cluster – K = 4



| | |
|---|---|
| | |