**ITM 6285 Data Mining Lab – Clustering**

**Surabhi Asati**

**Va7892**

**Oct 17th, 2017**

**Task 1: Import the dataset**

> summary(Car\_Identification\_1\_)

Acceleration BodyRoll DownForce Braking Category

Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 Toyota :50

1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 Jaguar :50

Median :5.800 Median :3.000 Median :4.350 Median :1.300 Aston\_Martin:50

Mean :5.843 Mean :3.054 Mean :3.759 Mean :1.199

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800

Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

**Median**

Acceleration: 5.800

body roll: 3.000

down force: 4.350

braking: 1.300

There are 3 categories namely, Toyota, Jaguar and Aston\_Martin with 50 cars in each category.

**Task 2: Standardize the data**

> Car\_Identification$Category <- as.factor(Car\_Identification$Category)

> Car\_Identification$Category=NULL

> CarStandardized=scale(Car\_Identification)

**Why we need to standardize the data?**

We need to have all the data features in the same scale. The reason for this is because otherwise the feature with the highest range will have more weight on the clustering process. For example, if we have a feature with range (0,100) and another with range (0,1), the last will have no effect on the clustering. Since clustering relies on distances it is evident how the feature with the smallest range contributes almost nothing when a distance is calculated.

**Task 3: K-Means clustering**

> kmc=kmeans(CarStandardized,3)

> kmc

K-means clustering with 3 clusters of sizes 53, 50, 47

Cluster means:

Acceleration BodyRoll DownForce Braking

1 -0.05005221 -0.8773526 0.3463713 0.2811215

2 -1.01119138 0.8394944 -1.3005215 -1.2509379

3 1.13217737 0.0962759 0.9929445 1.0137756

Clustering vector:

[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

[49] 2 2 3 3 3 1 1 1 3 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 1 1 1 3 3 3 1 1 1 1 1 1 1 3 3 1 1 1 1 1 1 1 1 1

[97] 1 1 1 1 3 1 3 3 3 3 1 3 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 3 3 3 3 3 3 1 1 3 3 3 1 3 3 3 1 3

[145] 3 3 1 3 3 1

Within cluster sum of squares by cluster:

[1] 44.25778 48.15831 47.60995

(between\_SS / total\_SS = 76.5 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

> kmc$centers

Acceleration BodyRoll DownForce Braking

1 -0.05005221 -0.8773526 0.3463713 0.2811215

2 -1.01119138 0.8394944 -1.3005215 -1.2509379

3 1.13217737 0.0962759 0.9929445 1.0137756

> kmc=kmeans(CarStandardized,3)

> kmc

K-means clustering with 3 clusters of sizes 47, 53, 50

Cluster means:

Acceleration BodyRoll DownForce Braking

1 1.13217737 0.0962759 0.9929445 1.0137756

2 -0.05005221 -0.8773526 0.3463713 0.2811215

3 -1.01119138 0.8394944 -1.3005215 -1.2509379

Clustering vector:

[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

[49] 3 3 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 1 1 1 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2

[97] 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 1 1 1 1 1 1 2 2 1 1 1 2 1 1 1 2 1

[145] 1 1 2 1 1 2

Within cluster sum of squares by cluster:

[1] 47.60995 44.25778 48.15831

(between\_SS / total\_SS = 76.5 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

> kmc$centers

Acceleration BodyRoll DownForce Braking

1 1.13217737 0.0962759 0.9929445 1.0137756

2 -0.05005221 -0.8773526 0.3463713 0.2811215

3 -1.01119138 0.8394944 -1.3005215 -1.2509379

Based on the latest run, average acceleration for each of the three clusters:

[1.13217737 + (-0.05005221) + (-1.01119138)]/3 = (0.07093378)/3 = 0.02364459333

Also, the last data point (150th car) belong to 2nd cluster