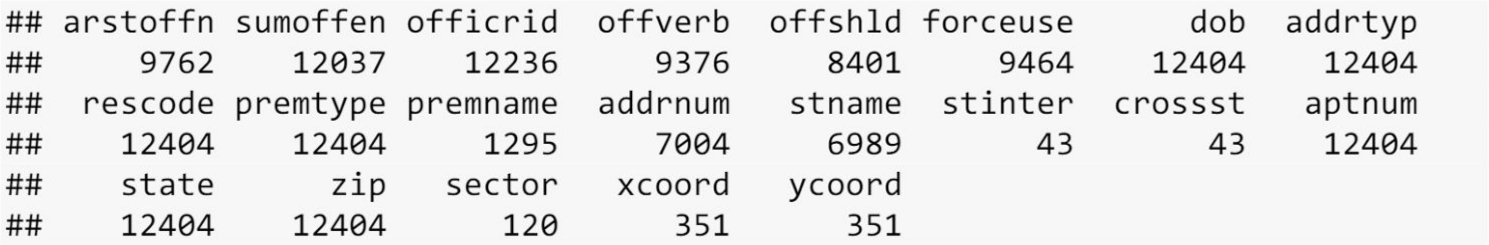
Data Mining

Paichana Kularb
57090015

Stop and Frisk

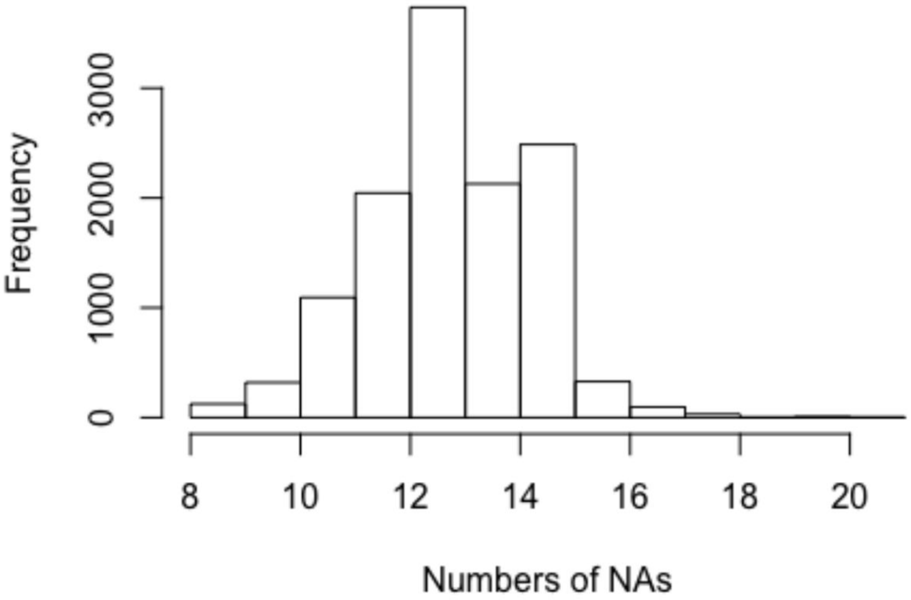
Project 1

Exploring the Data



```
## arstoffn sumoffen officrid  offverb  offshld forceuse      dob  addrtyp
##     9762    12037    12236     9376     8401     9464    12404    12404
##  rescode premtype premname  addrnum   stname  stinter  crossst   aptnum
##    12404    12404     1295     7004     6989       43       43    12404
##    state      zip   sector   xcoord   ycoord
##    12404    12404      120      351      351
```

Exploring the Data



Remove NAs

- Fill NAs with negative cases for example:
  - arstoffn if NA filled with ‘NOARREST’
  - sumoffen if NA filled with ‘NOSUMMON’
- Fill NAs with mean for numerical datas:
  - xcoord and ycord filled with average coordinates of each city
- Fill NAs with mode for categorical datas:
  - Premname, sector, forceuse

New variables

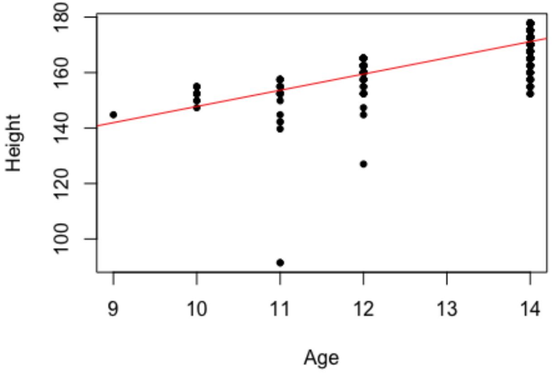
- isforceuse - indicate whether any type of force is used.
- weaponFound - indicate whether any type of weapon is found.
- day - day of the week(monday - sunday)
- Height - in meters combination of ht_feet and ht_inch
- bmi - Body mass index
- Hours - time in hour with fractions = Hours + Minute/60

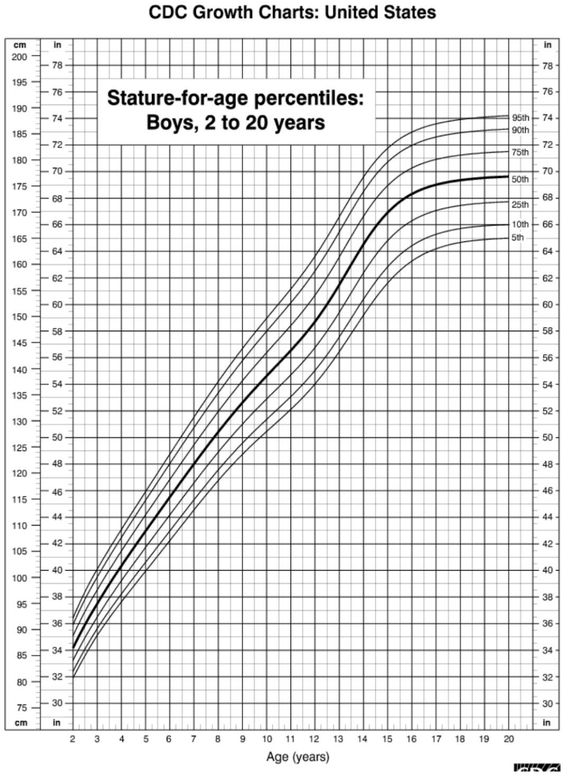
Remove duplicate

- Duplicate condition:
  - If age, height, datestop, weight and race is the same the row is consider as duplicate.
- Total of 117 rows and is removed

Outliers

- Remove age < 5
- Compare data to CDC Growth Chart:
  - Unreasonable age-height combination is replaced with mean.





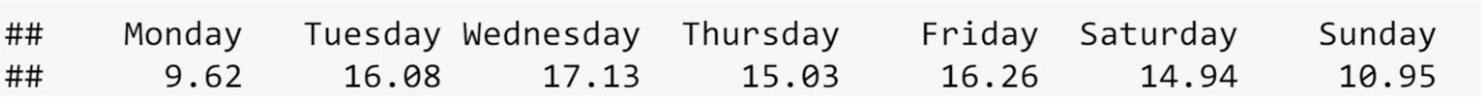
Outliers

- Remove weight <30 kg and weight > 30kgs
- Remove height < 100 cm and height >200 cm
- Remove bmi < 15 and bmi > 60

Interesting Findings

Day of the week

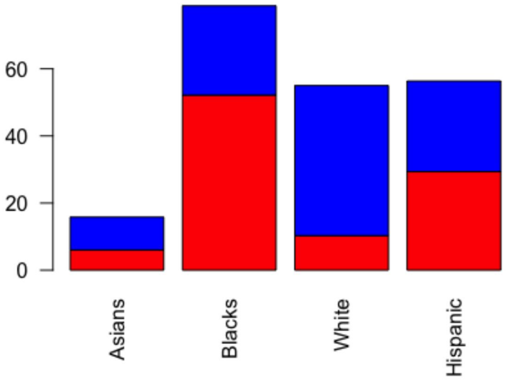
- Least occurrence on Sunday and Monday
- Most occurrence on Wednesday

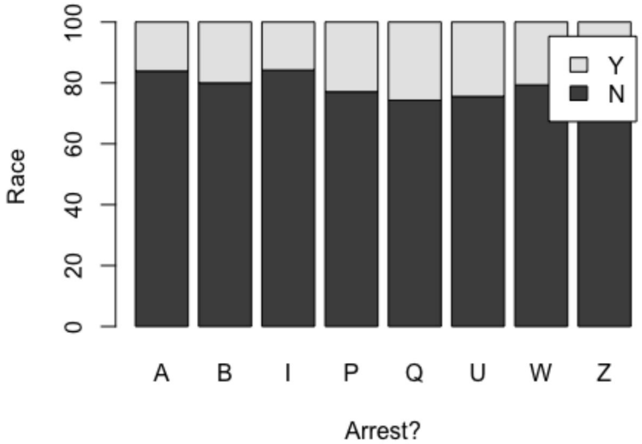


```
##     Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
##       9.62     16.08     17.13     15.03     16.26     14.94     10.95
```

Race

	Asian	Black	Indian	B His	W His	W	Other + Unknown
Data	6.02	52.15	0.31	7.10	22.23	10.30	1.89
Wikipedia	11.8	25.1	-	27.5		44.6	-

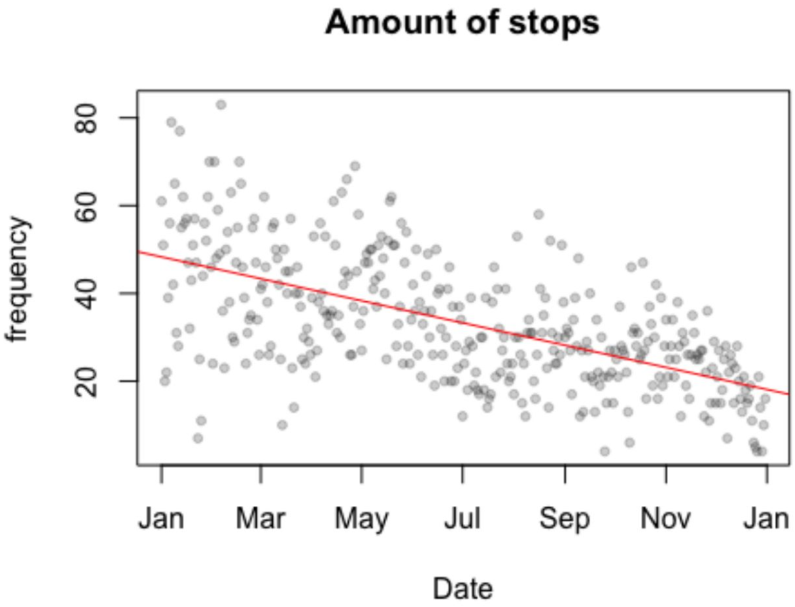




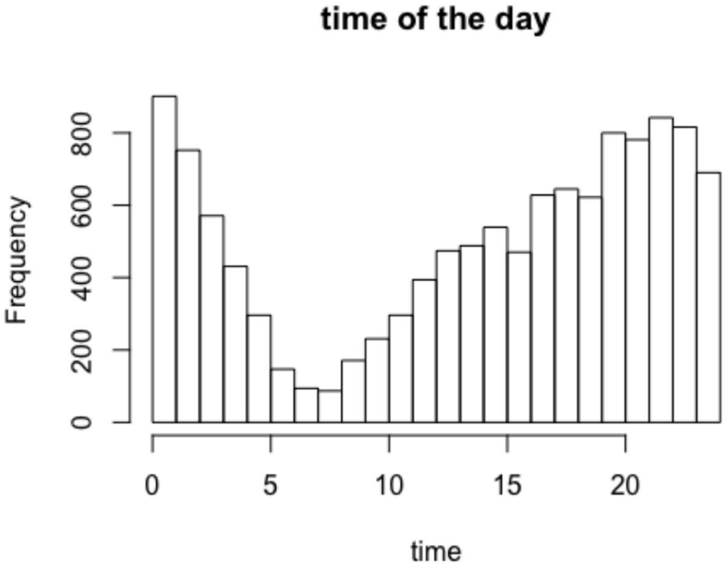
Weapon found - Arrest - is force used

	No	Yes
Weapon found	94.25	5.75
Arrest made	78.71	21.29
Is force used	69.33	30.67

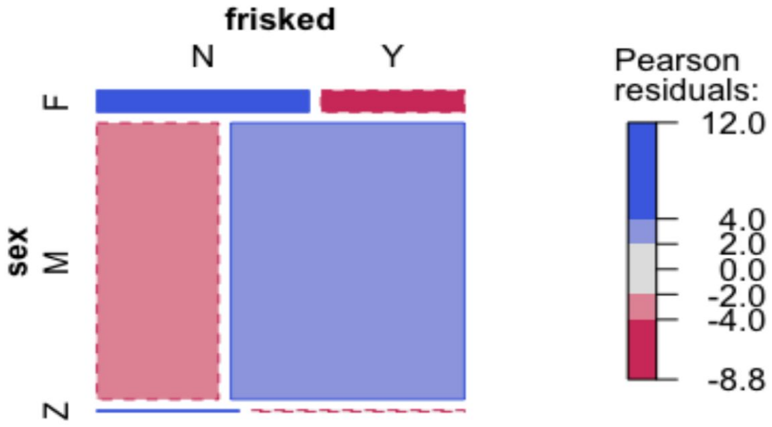
Amount of stop vs time



Time of the day



Sex VS Frisked



Stop and Frisk

Project 2

Machine Learning Libraries

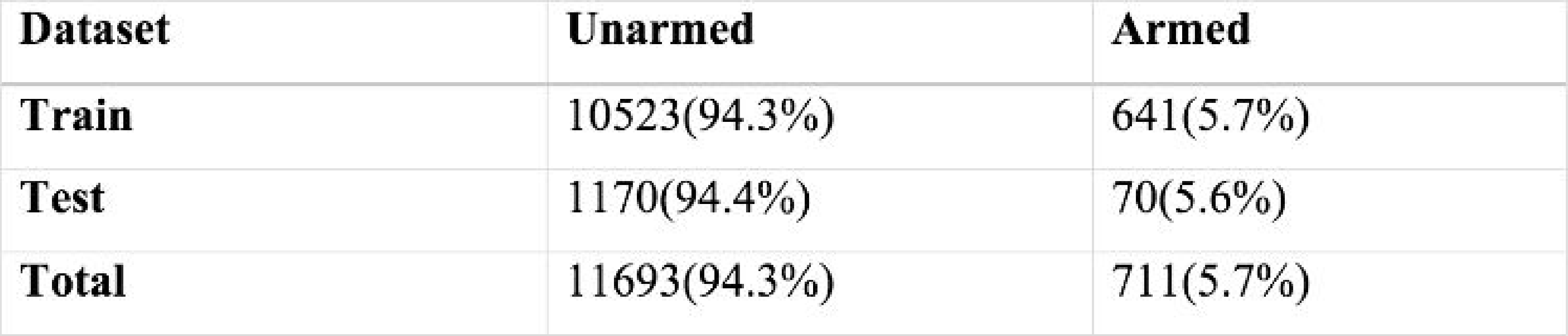
Libraries	Model
e1071	SVM and Naive Bayes
rpart	CART
C5.0	C5.0 tree
randomForest	Random Forest
nnet	Neural Network

Attributes	Description	Derived from
day	Day of the week(Monday-Sunday)	Datestop
hours	Time of the day (hours)	timestop
age	Age of the suspect	
sex	Sex of suspect	
race	Race of suspect	
frisked	Was suspect frisked	
searched	Was suspect searched	
perobs	Period of obersvation(minutes)	
perstop	Period of stop(Minutes)	
typeofid	Suspect's identification type	
weight	Suspect's weight(Kg)	
height	Suspect's height(Cm)	ht_feet, ht_inch
bmi	Suspect's BMI	height,weight
stopReason	Reason for stop	cs_objcs,cs_descr, etc
crimsup	Crime suspected	

Person is Armed?

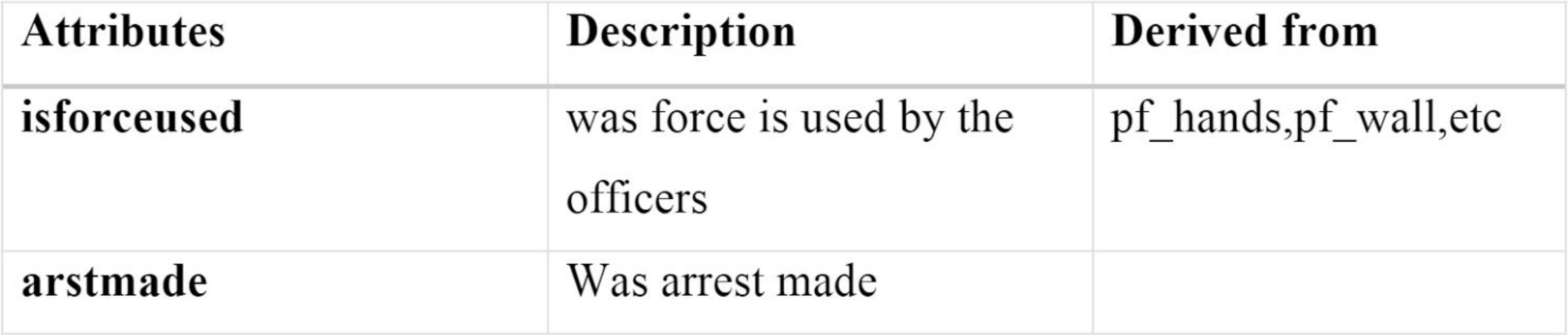
Project 2

Dataset



Dataset	Unarmed	Armed
Train	10523(94.3%)	641(5.7%)
Test	1170(94.4%)	70(5.6%)
Total	11693(94.3%)	711(5.7%)

Additional Attributes



Attributes	**Description**	**Derived from**
isforceused	was force is used by the officers	pf_hands,pf_wall,etc
arstmade	Was arrest made	

Performance Measures

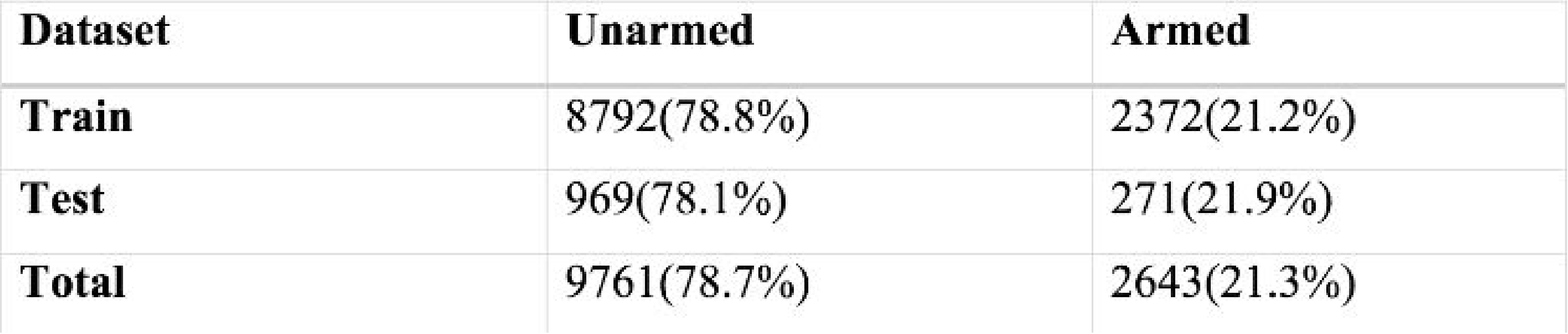
	Precision	Recall	Accuracy
SVM armed	94.4%	100%	94.4%
SVM unarmed	0%	0%	
CART armed	95.5%	98.8%	94.4%
CART unarmed	51.7%	21.4%	
NN armed	96.1%	98.4%	94.7%
NN unarmed	54.8%	32.8%	

Neural Network is the best at predicting whether a person is armed

Arrest Made?

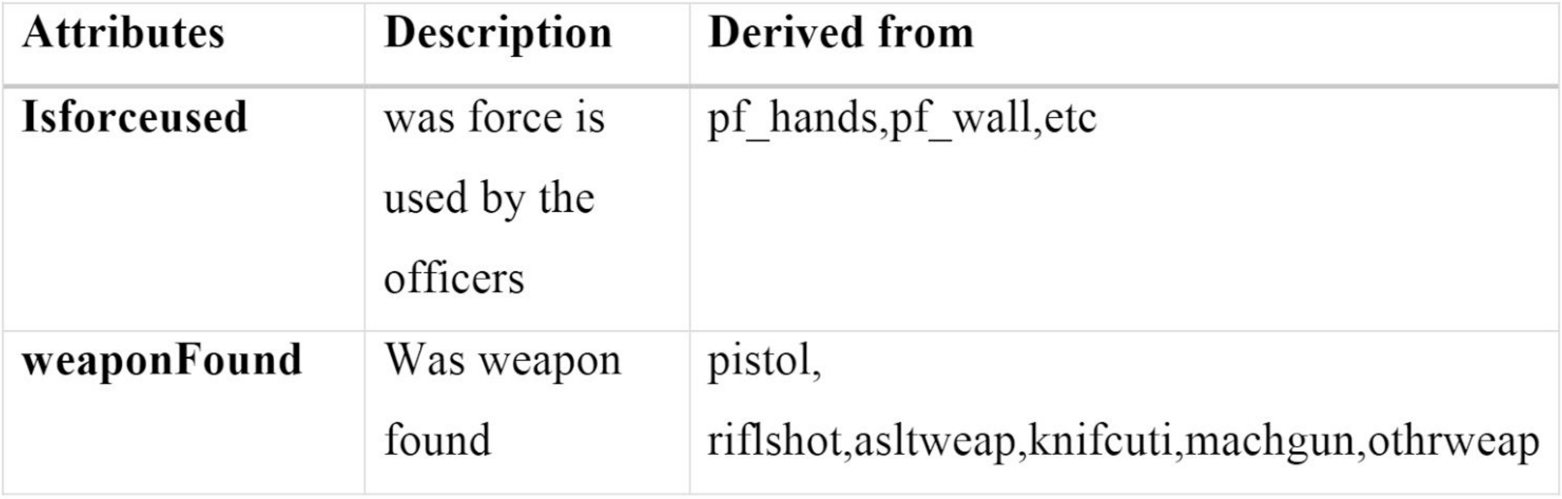
Project 2

Data



Dataset	Unarmed	Armed
Train	8792(78.8%)	2372(21.2%)
Test	969(78.1%)	271(21.9%)
Total	9761(78.7%)	2643(21.3%)

Additional Attributes



Attributes	**Description**	**Derived from**
Isforceused	was force is used by the officers	pf_hands,pf_wall,etc
weaponFound	Was weapon found	pistol, riflshot,asltweap,knifcuti,machgun,othrweap

Performance Measures

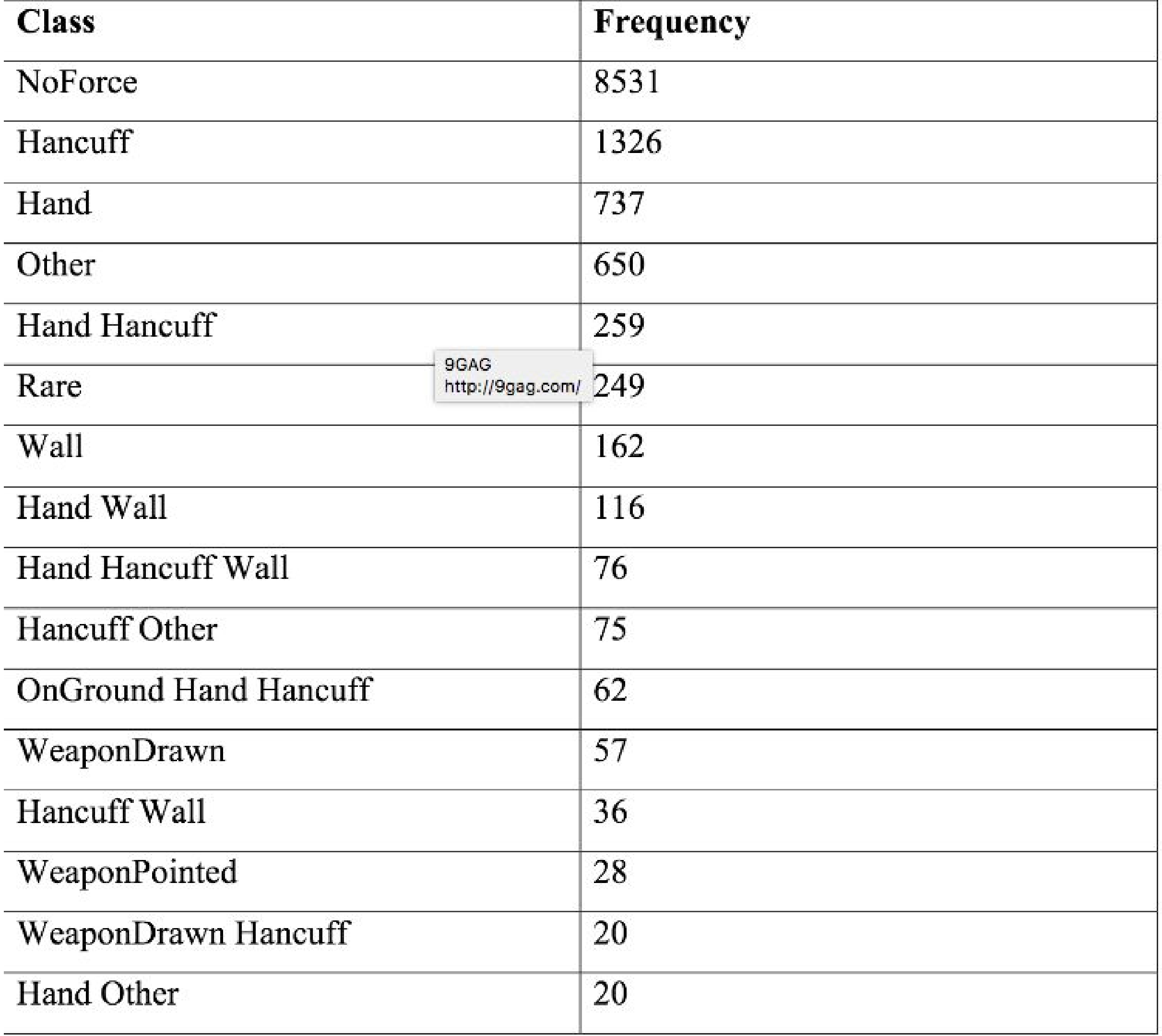
	Precision	Recall	Accuracy
RF arrested	88.9%	94.7%	86.6%
RF not arrested	75.4%	57.6%	
NN arrested	90.5%	93.2%	85.2%
NN not arrested	64.9%	72.7%	
Stacked arrested	91.2%	91.8%	86.7%
Stacked not arrested	70.1%	68.3%	

The performance is very similar no matter which algorithm is used to predict.

Type of Force used?

Project 2

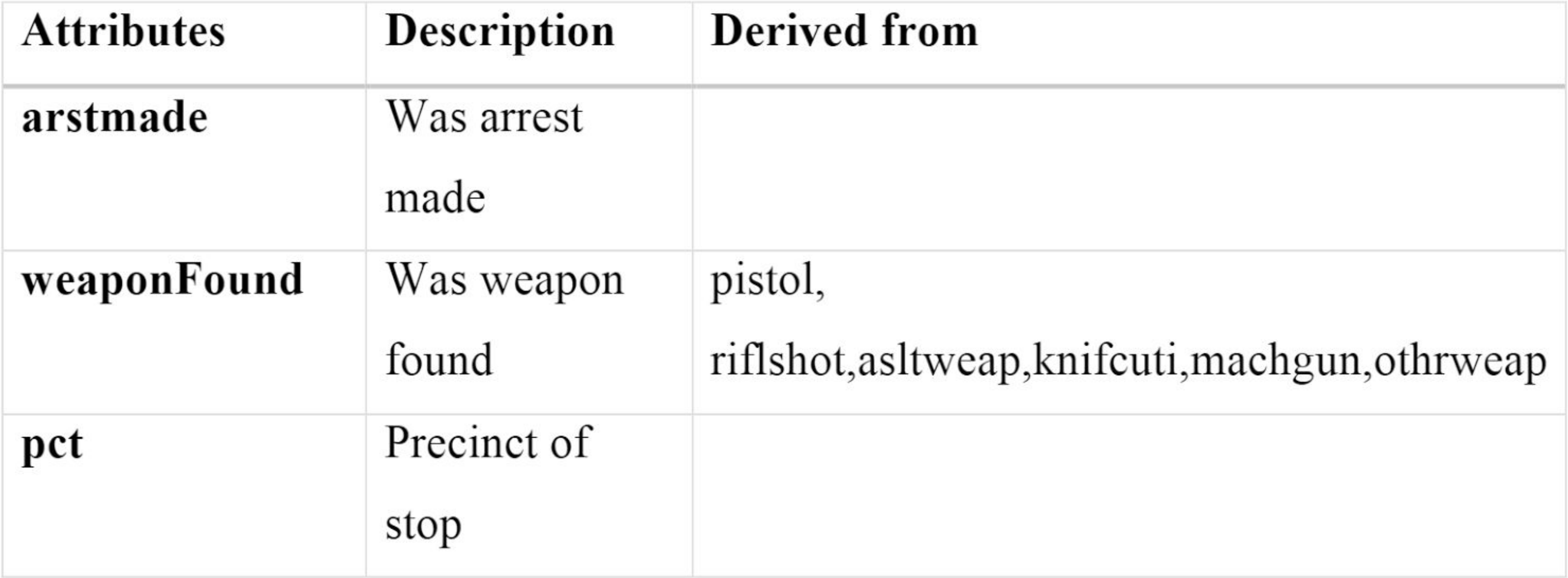
Data



Class	Frequency
NoForce	8531
Hancuff	1326
Hand	737
Other	650
Hand Hancuff	259
Rare	249
Wall	162
Hand Wall	116
Hand Hancuff Wall	76
Hancuff Other	75
OnGround Hand Hancuff	62
WeaponDrawn	57
Hancuff Wall	36
WeaponPointed	28
WeaponDrawn Hancuff	20
Hand Other	20

9GAG
http://9gag.com/

Additional Attributes



Attributes	**Description**	**Derived from**
arstmade	Was arrest made	
weaponFound	Was weapon found	pistol, riflshot,asltweap,knifcuti,machgun,othrweap
pct	Precinct of stop	

Performance Measures

	Accuracy
Naive Bayes	54.6%
C5.0	67.2%
SVM	67.9%

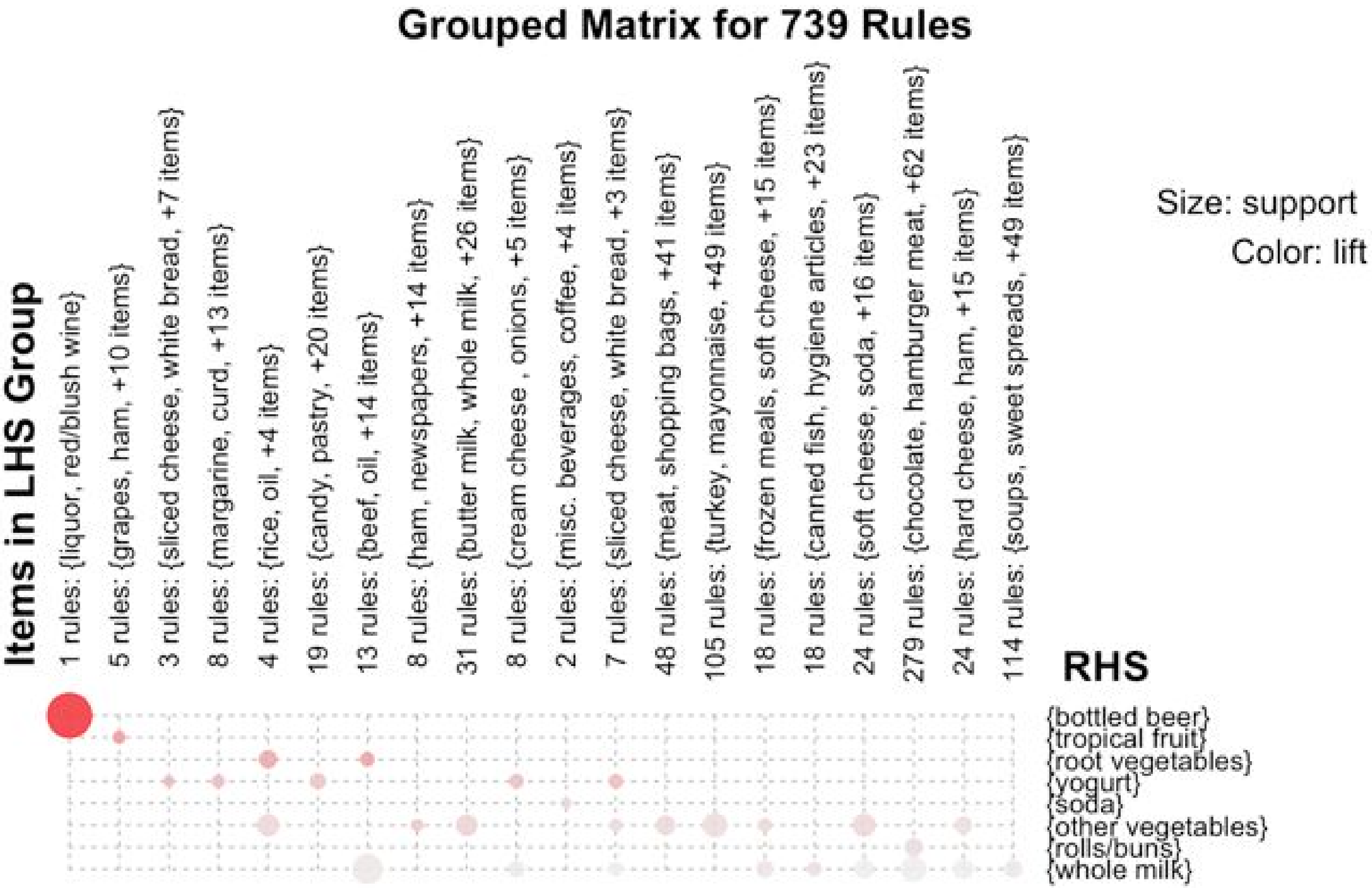
Naive Bayes performs the worst between the three

Groceries

Project 3

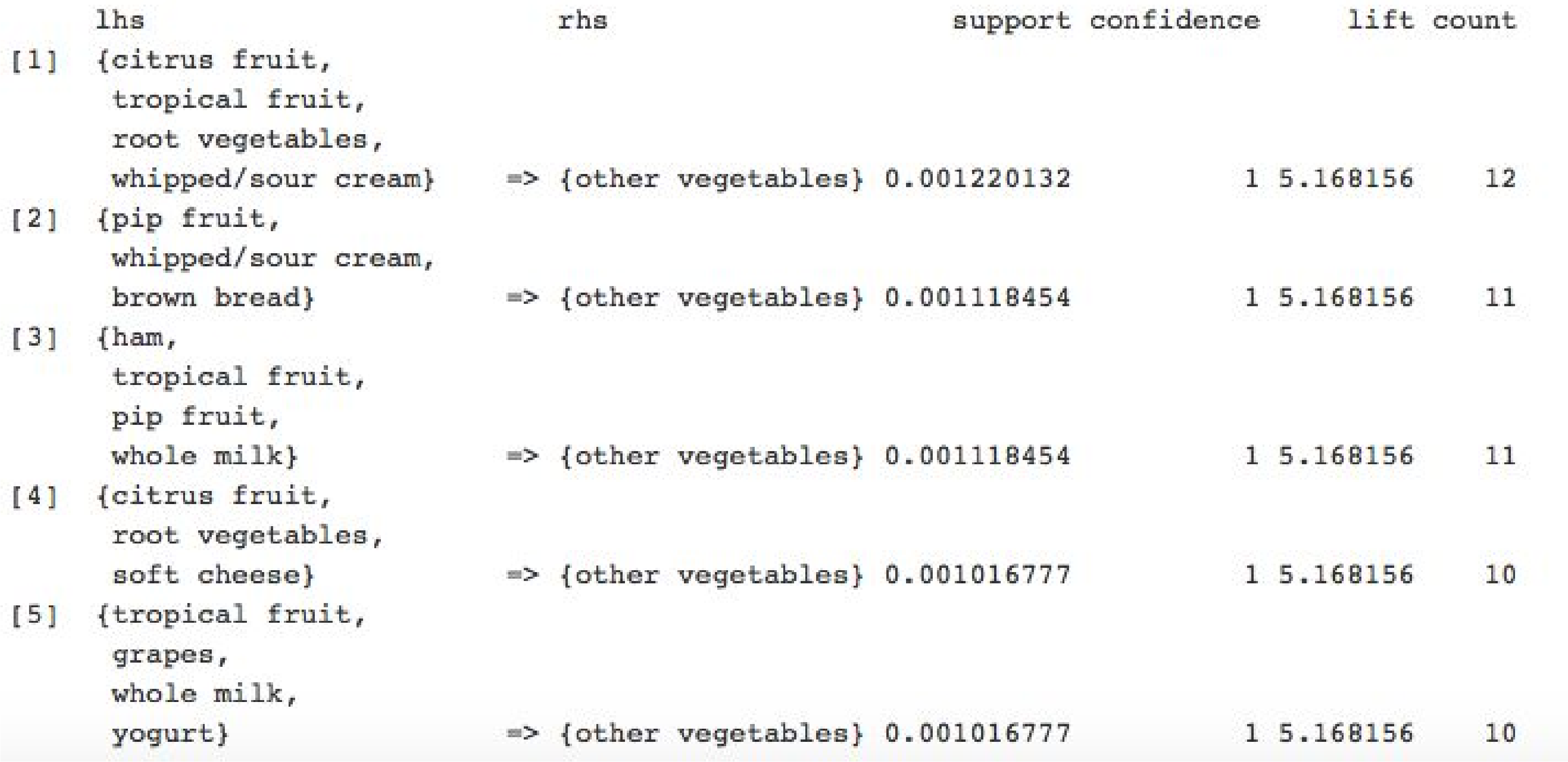
Association Rules

- Minimum Support: 0.001
- Confidence: 0.75
- Apriori Algorithm
- 777 rules



Grouped Matrix for 739 Rules

Top 5 rules

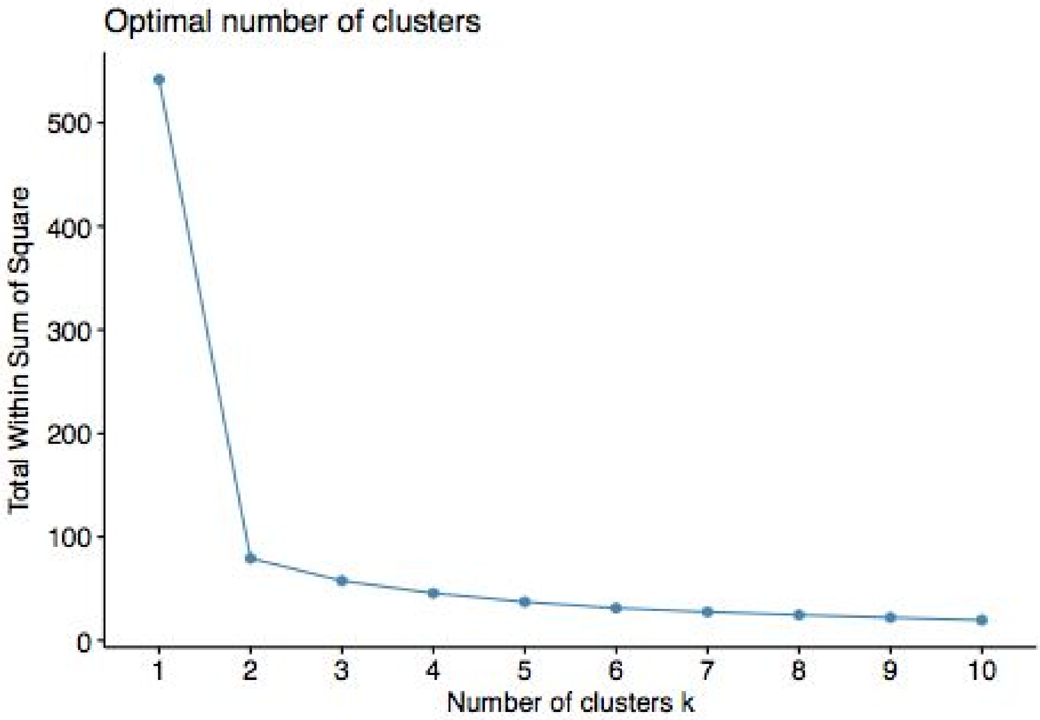


```
    lhs                          rhs                    support confidence     lift count
[1] {citrus fruit,
     tropical fruit,
     root vegetables,
     whipped/sour cream}     => {other vegetables} 0.001220132          1 5.168156    12
[2] {pip fruit,
     whipped/sour cream,
     brown bread}            => {other vegetables} 0.001118454          1 5.168156    11
[3] {ham,
     tropical fruit,
     pip fruit,
     whole milk}             => {other vegetables} 0.001118454          1 5.168156    11
[4] {citrus fruit,
     root vegetables,
     soft cheese}            => {other vegetables} 0.001016777          1 5.168156    10
[5] {tropical fruit,
     grapes,
     whole milk,
     yogurt}                 => {other vegetables} 0.001016777          1 5.168156    10
```

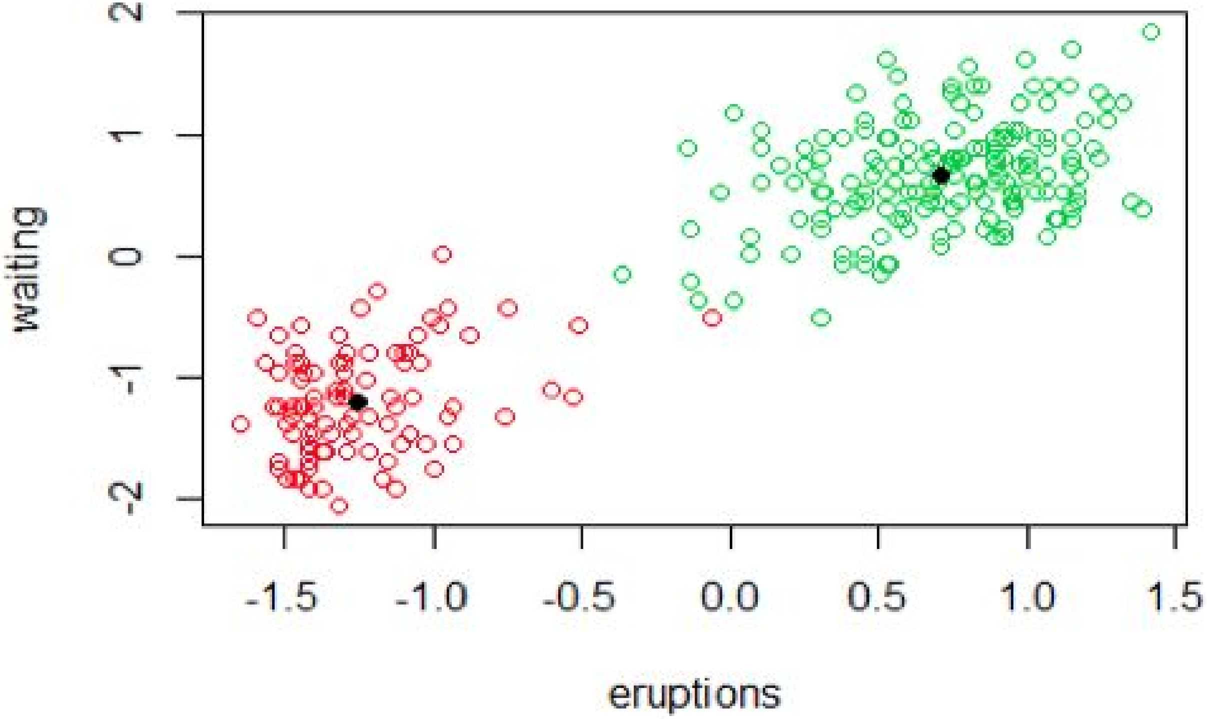
Old Faithful

Project 4

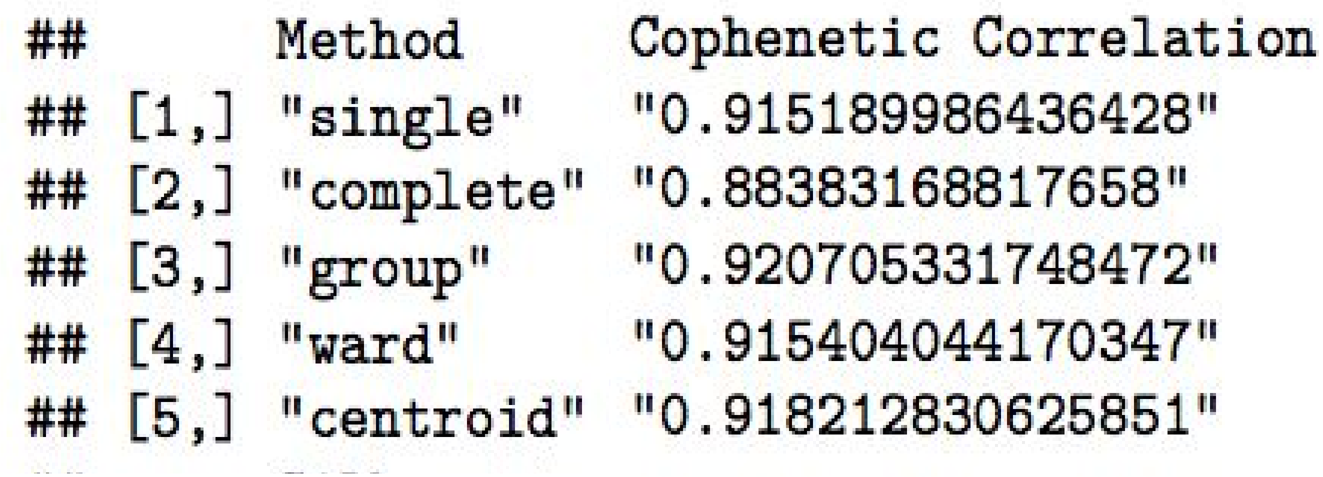
Number of clusters in K-Mean



K = 2



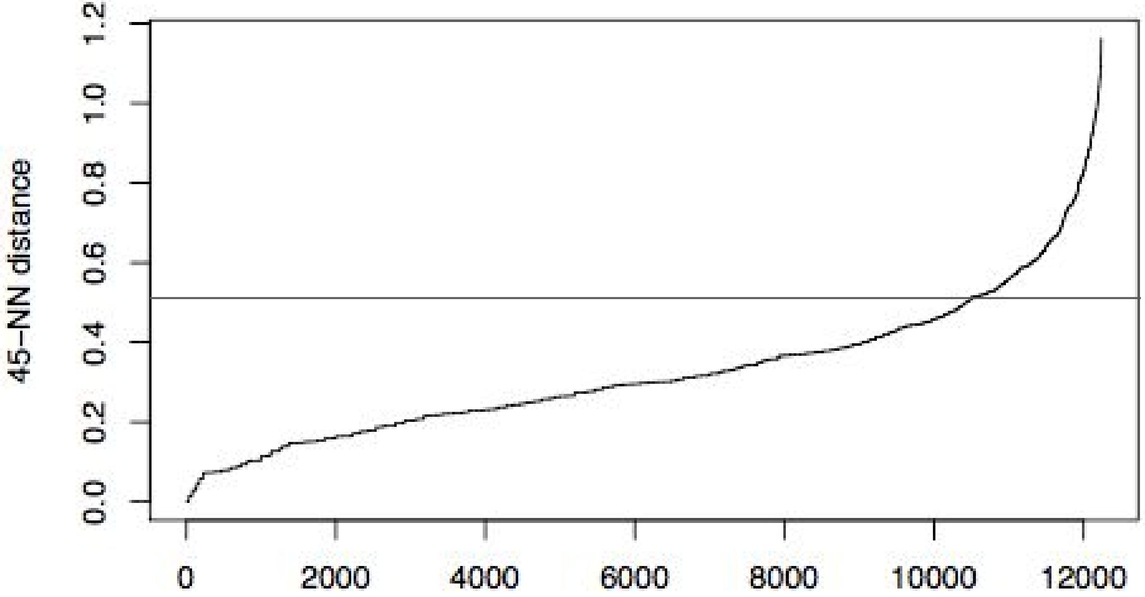
Hierarchical Clustering



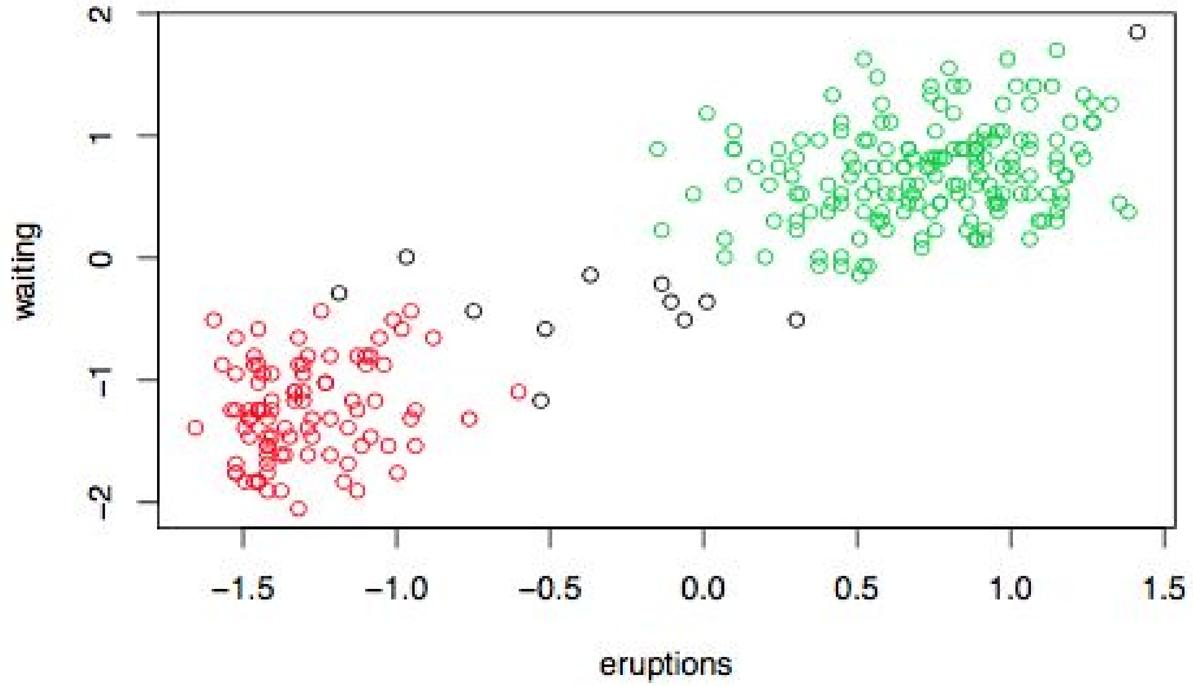
```
##        Method     Cophenetic Correlation
## [1,] "single"   "0.915189986436428"
## [2,] "complete" "0.88383168817658"
## [3,] "group"    "0.920705331748472"
## [4,] "ward"     "0.915404044170347"
## [5,] "centroid" "0.918212830625851"
```

DBSCAN

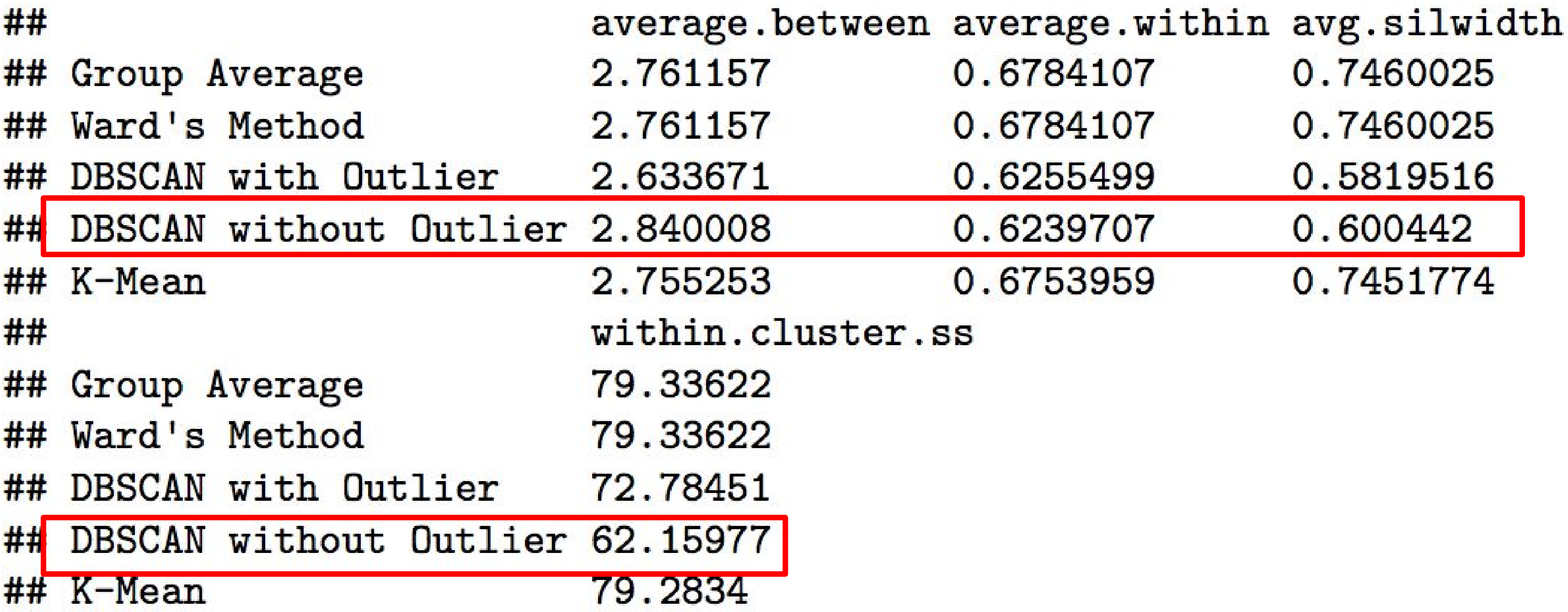
MinPts = 45 and Eps = 0.5



Plot Clusters

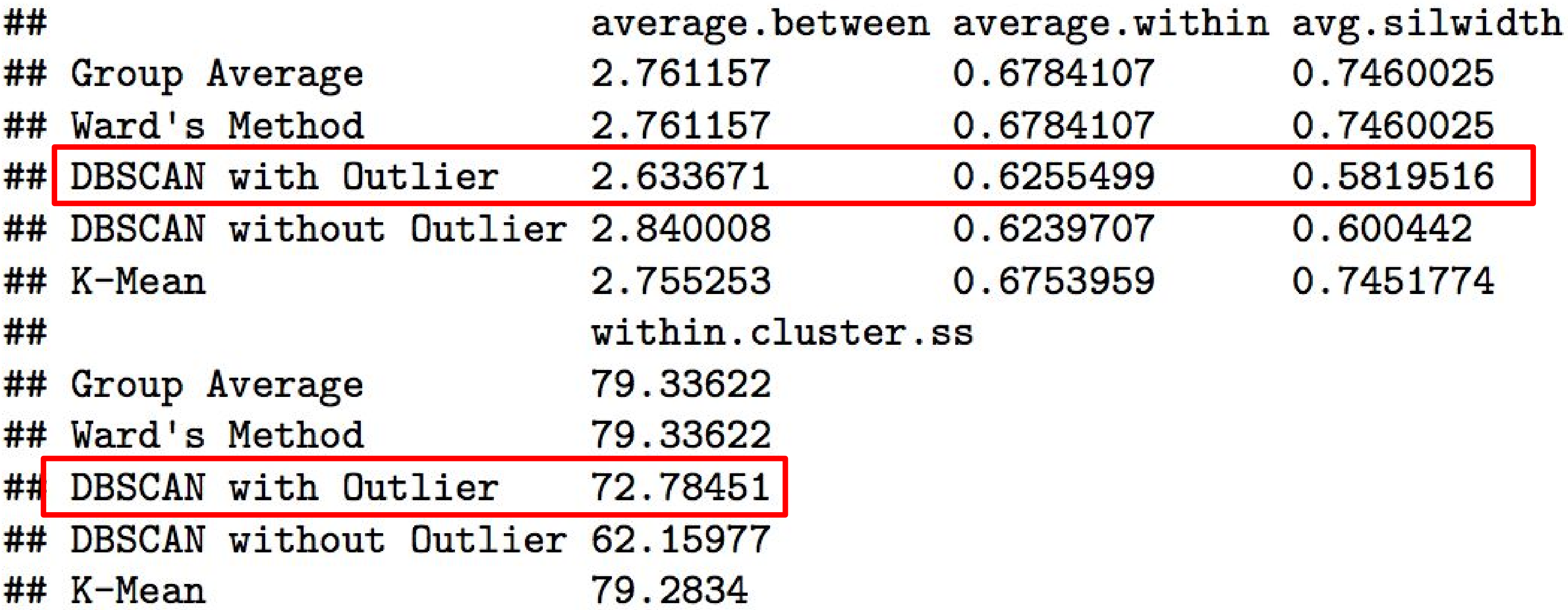


Comparing different methods



```
##                         average.between average.within avg.silwidth
## Group Average           2.761157        0.6784107      0.7460025
## Ward's Method           2.761157        0.6784107      0.7460025
## DBSCAN with Outlier     2.633671        0.6255499      0.5819516
## DBSCAN without Outlier  2.840008        0.6239707      0.600442
## K-Mean                  2.755253        0.6753959      0.7451774
##                         within.cluster.ss
## Group Average           79.33622
## Ward's Method           79.33622
## DBSCAN with Outlier     72.78451
## DBSCAN without Outlier  62.15977
## K-Mean                  79.2834
```

Comparing different methods



```
##                        average.between average.within avg.silwidth
## Group Average          2.761157        0.6784107      0.7460025
## Ward's Method          2.761157        0.6784107      0.7460025
## DBSCAN with Outlier    2.633671        0.6255499      0.5819516
## DBSCAN without Outlier 2.840008        0.6239707      0.600442
## K-Mean                 2.755253        0.6753959      0.7451774
##                        within.cluster.ss
## Group Average          79.33622
## Ward's Method          79.33622
## DBSCAN with Outlier    72.78451
## DBSCAN without Outlier 62.15977
## K-Mean                 79.2834
```