Data Mining - Project 3
Association Rule Mining

- Vivek Sanghvi Jain

Under Guidance of
Prof. Michael Hahsler

Table of Contents

1. Executive Summary & Business Understanding

This report is a continuation of an earlier report which dealt with Data and Visualisation and Predictive Analysis based on the data collected from the Dallas Police Department, on the open data online portal "*Dallas Open Data"*. Using this data, an in-depth analysis was performed using RStudio, which reveals some important associations and correlations and and sheds new light on how important they are and how can they be effectively used to reduce crime.

These rules can prove to be very informative for all the stakeholders and help them understand the frequent item sets in the data transaction. A data transaction is a rule based transaction, which finds the relation between different data attributes.

The report has been complied according to the guidelines of CRISP-DM Framework

2. Data Understanding & Data Preparation

For this report we have used a small part of the data from the Dallas Open Data, as explained in the earlier project. Furthermore, we also use some new arrest data[1] on the assailant or the perpetrator. Which, we merge with the existing data as a left inner join using the following code:

```
data <- merge(x = data, y = NewArrestData, by = "IncidentNum", all.x = TRUE)
```

Where NewArrestData is a subset of the ArrestData containing the following attributes:

1. IncidentNum
2. PerpAgeAtArrestTime (Originally Age)
3. PerpHeight (Originally Height)
4. PerpWeight (Originally Weight)
5. PerpHair (Originally Hair)
6. PerpEyes (Originally Eyes)
7. PerpRace (Originally Race)
8. PerpSex (Originally Sex)

Col #	Column Name	Description, Example	Scale Type
1	IncidentNum	RMS generated number	Nominal
2	PCClass	Severity level of the Arrest Incident & Classification of Offence, according to the Penal Code, either a misdemeanour of type M*, MA, MB, MC or a felony of Type F*, F1, F2, F3, FS, FX	Ordinal
3	Premise	Location type where arrest incident took place For example, Apartment Parking, Residence	Nominal
4	Division	The Divisions in which Tx is divided. They can be Central, North-Central, North-East,North-West South-Central SouthEast, SouthWest	Nominal
5	Month1	Month of the offense	Interval
6	Day1	Day of the offense	Interval
7	Time1	Time of the offense	Ratio
8	CompRace	Complainant's Race(Asian, Black, White,NativeIndian,Italian, Latino)	Nominal
9	CompSex	Complainant's Sex(M or F)	Nominal
10	CompAge	Complainant's Age which is the same as the time of the offense. Numeric	Ratio
11	UCROffDesc	UCR Offense description. A list of several descriptions	Nominal
12	Status	One of the following: Clear by Arrest, Clear by Exceptional Arrest, Closed/Cleared, Open,Returned for Correction or Suspended	Nominal
13	Watch	Time for Policing is divided into 3- hour shifts or watches. They can be 1/2/3 signifying Midnight/Day/Evening Shift	Nominal
14	PerpAgeAtArrestTime	Perpetrator Age which is the same as the time of the offense. Numeric	Ratio
15	PerpHeight	Height of Perpetrator in Feet and inches	Ratio
16	PerpWeight	Weight of Perpetrator	Ratio
17	PerpHair	Color of hair was perpetrator, One of the following: Black, Blonde, Brown, Red, White, Other	Nominal
18	PerpSex	Perpetrator's Sex(M or F)	Nominal
19	PerpRace	Complainant's Race(Asian, Black, White,NativeIndian,Italian, Latino)	Nominal
20	PerpEyes	Color of Eyes of Perpetrator, one of the following: Black, Blue,Brown,Green,Hazel,Other	Nominal

Table 2.1: Data Description for Selected Data Variables

We clean the new arrest data in a similar fashion like we cleaned earlier data, as it had similar demographic inconsistencies, about the perpetrator like the victim.

2.1 Discretization

Here, comes a very important part, that is, preparing the data for transactions. For which, we have to discrete all continues variables and aggregate some other variables because a transaction only understands data which is true and false. If you have a variable which can take upto 5 different values, the transaction will make 5 logical variables, one for each value.

Discretezation means breaking down the continuous range of values into certain values. It can be manually ordered and decided. For, eg: We can split age into three value range, based on a adulthood age and retirement age, into non-adult, adult or pre-retirement and post-retirement. If the adulthood age is 18, retirement age is 65, any age from 0-17 would be a non-adult, anyone from 18-64 would fall under the pre-retirement or working age and anyone of age 65 and over would fall into the retirement age category.

But, there are other mathematical ways to discretize a range.

In this project, the following variables have been disacretized, along with a brief explanation of the reason for choosing the method of discretisation.

1. Time1: This variable represents the time when the 911 call came in. Because, its time, and we inherently understand it to be divided into 24 hours, the same was done over here. It can also be divided in 8 hour shifts, but the original data already had a variable for that called watch, and dividing into 24 hours gives us a more actual and precise understanding of When Crime occurs and the call came in. incident has a time stamp in it and the hour value is extracted from the timestamp using Regular Expression or simple Substring substitution and then assigned to the same variable Time1 getting rid of other Timestamp data which we don't need anymore.

```
data$Time1 <- as.factor(as.numeric(gsub(":\\d\\d", "", data$Time1)))
```

This creates 24 factors for Time1, 1 for each hour.

2. CompAge:

CompAge is Continuous, Discretised into range

Age Group	Number of People in Age Group
[17, 31)	26293
[31, 46)	26485
[46,100]	25285

```
data$CompAge <- discretize(data$CompAge, method = "frequency")
```

In this case, discretization was done using the function discretize and the method chosen was discretization via frequency. There are other methods in the discretization function as well, namingly: "interval" (equal interval width), "frequency" (equal frequency), "cluster" (k-means clustering) and "fixed" (categories specifies interval boundaries) [2]

After trying all of them, the frequency method divided it into ranges with similar observation in each range, which avoids the problem of skewing the data, if we have multiple records of a particular age group for analysis.

3. PerpAge is also continues, discretised using a similar logic as CompAge

Age Group	Number of People in Age Group
[17, 26)	2219
[26,38)	2020
[38,85]	2042

data$PerpAge <- discretize(data$PerpAge, method = “frequency")

4. Perp Weight <- In, this case we specify that we want only 2 categories, but still use the frequency discretize method. Interval gives us 6081 readings for less than 168 pounds and 200 readings for greater than 168, which doesn't produce any good analytical result, which is why we stick to frequency for discretising all the above 3 continuous variables.

data_3$PerpWeigh <- discretize(data_3$PerpWeight, method = "frequency", categories = 2)

Weight Group, Label	Number of People in Weight Group
[0,168)	3141
[168,516]	3140

Apart from discretization, we also have to create binary features to indicate the presence or absence of an item or an event. The following Binary variables are created:

1. is_felony : FALSE:63650 TRUE :39790
   data$is_felony <- data$PCClass == "F*" | data$PCClass == "F1" | data$PCClass == "F2" | data$PCClass == "F3" | data$PCClass == “FS" | data$PCClass ==‘FX’

True for all instances for PCClass F*,F1,F2,F3,FS,FX. Helps separate all Felony incidents for further analysis and also gets rid of any NA’s in the data.

2. is_misdemeanor : FALSE:39815 TRUE :63625
data$is_misdemeanor <- data$PCClass == "M*" | data$PCClass == "MA" | data$PCClass == "MB" | data$PCClass == "MC"
True for all instances for PCClass M*,MA,MB,MC. Helps separate all Misdemeanour incidents for further analysis.

3.arrest FALSE:106960 TRUE : 12203
data$arrest <- data$Status == "Clear by Arrest" | data$Status == "Clear by Exceptional Arrest" | data$Status == "Returned for Correction"
data$arrest <- as.factor(data_sub$arrest)

Helps identify all incidents that led to an arrest for further analysis

4. childRelated FALSE:118514 TRUE : 1028

data$childRelated <- data$UCROffDesc == "CHILD (OFFENSES AGAINST)"
data$childRelated <- as.factor(data$childRelated)
Helps identify all incidents involving a child victim for further analysis.

2.2 Construction of Transaction Datasets

6 Transaction datasets were constructed and operated upon. Transactions are created on these data sets in Section 4. All transactions are created from a single cleaned dataset, "**data**"

1. data_sub_misdemeanor

```
data_sub_misdemeanor <- data[ which(data$is_misdemeanor==TRUE),]

data_sub_misdemeanor<-data_sub_misdemeanor[,c("Premise","Division", "Month1",
"Day1","Time1","CompRace","CompAge","CompSex","UCROffDesc","arrest")]
```

For, when the logical variable is_misdemeanor is true, the incident is subsetted to form this dataset out of which, the transaction is created.
It has 63,625 observations for 10 variables

2. data_sub_felony

```
data_sub_felony <- data[ which(data$is_felony==TRUE),]

data_sub_felony <-data_sub_felony[,c("Premise","Division", "Month1",
"Day1","Time1","CompRace","CompAge","CompSex","UCROffDesc","arrest")]
```

For, when the logical variable is_felony is true, the incident is subsetted to form this dataset out of which, the transaction is created.
It has 39,790 observations for 10 variables

3. data_sub_femalevictim_young

```
data_sub_femalevictim_young <- data[ which(data$CompSex=="F" & data$CompAge=="[17, 31)"),]

data_sub_femalevictim_young <- data_sub_femalevictim_young[,c("Premise","Division", "Month1",
"Day1","Time1","CompRace","UCROffDesc","PCClass","Status")]
```

This is a case, for all female complainants. And I have divided the age into 2 categories, Under 31, which is the young age, and over 31, which is the mature or an older age. And this is the case for all young female complainants under the age of 31. There were 13,539 observations over 8 variables.

4. data_sub_femalevictim_old

```
data_sub_femalevictim_old <- data[ which(data$CompSex=="F" & (data_sub$CompAge=="[31, 46)") |
data_sub$CompAge=="[46,100]") ,]

data_sub_femalevictim_old <- data_sub_femalevictim_old[,c("Premise","Division", "Month1",
“Day1”,"Time1","CompRace","UCROffDesc","PCClass","Status")]
```

This is a also a case, for all female complainants. And I have divided the age into 2 categories, Under 31, which is the young age, and over 31, which is the mature or an older age. And this is the case for all young female complainants over the age of 31. There were 37,662 observations over 8 variables.

5. data_sub_malevictim_young

```
data_sub_malevictim_young <- data[ which(data$CompSex=="M" & data$CompAge=="[17, 31)"),]

data_sub_malevictim_young <- data_sub_malevictim_young[,c("Premise","Division", "Month1",
“Day1”,"Time1","CompRace","UCROffDesc","PCClass","Status")]
```

This is a case, for all female complainants. And I have divided the age into 2 categories, Under 31, which is the young age, and over 31, which is the mature or an older age. And this is the case for all young female complainants under the age of 31. There were 12,748 observations over 8 variables.

6. data_sub_malevictim_old

```
data_sub_malevictim_old <- data[ which(data$CompSex=="M" & (data_sub$CompAge=="[31, 46)") |
data_sub$CompAge=="[46,100]") ,]

data_sub_malevictim_old <- data_sub_malevictim_old[,c("Premise","Division", "Month1",
“Day1”,"Time1","CompRace","UCROffDesc","PCClass","Status")]
```

This is a also a case, for all female complainants. And I have divided the age into 2 categories, Under 31, which is the young age, and over 31, which is the mature or an older age. And this is the case for all young female complainants over the age of 31. There were 39,367 observations over 8 variables.

7. data_perpdata_merged

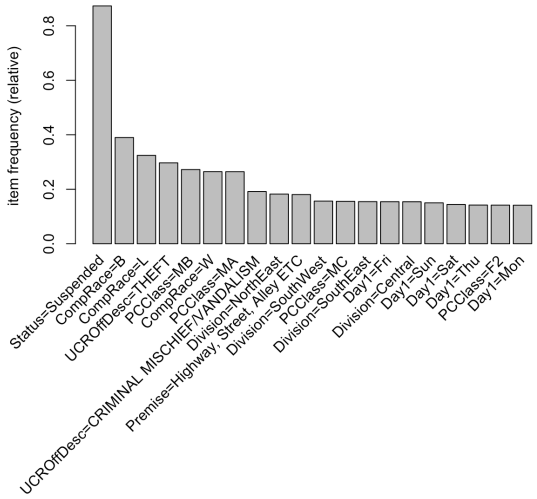
```
data_perpdata_merged <- data[,c("Premise","Division", "Month1",
"Day1","Time1","CompRace","UCROffDesc","PCClass","Status","PerpSex","PerpAgeAtArrestTime","Pe
rpRace","PerpWeight","PerpHeight")]
```

This is a simple case of considering all the new arrest data after performing a join and removing any record with a single Na in it. These are 1252 observations over 14 variables. These might not look that much, but with the current arrest data and existing data after merging is still filled with NA. And analysis I intend to perform needs to get rid of all NAs.

3. Modeling - ItemSets and Transactions

Case 1:Dataset used: data_sub_femalevictim_young

Figure 4.1: Top 20 Frequency Items



TRANSACTION:

Using the above transaction Female Victim Young

```
d_s <- data_sub_femalevictim_young
trans <- as(d_s, "transactions")
d_s <- data_sub_femalevictim_young[,c("Premise","Division", "Month1",
        "Day1","Time1","CompRace","UCROffDesc","PCClass","Status")]
```

Detailed Analysis with emphasis on usefulness of obtained patterns explained in section 5.

FREQUENT ITEMSETS

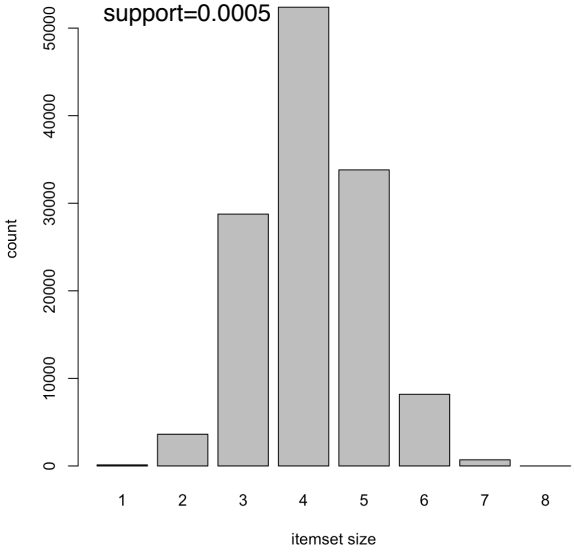
Item sets created for Support = 0.01 and also for support =0.0005 to cover both ends of the support spectrum. Reducing the support from 0.01 to 0.0005 produced a lot of interesting item-sets and rules which would have been missed otherwise. Set of 2631 itemsets was generated with support 0.01 and a Set of 127599 itemsets was generated with support 0.00005

```
is <- apriori(trans, parameter=list(target="frequent", support=0.0005))
Apriori
is <- sort(is, by="support")
inspect(head(is, n=10))
```

Note, all these item sets, have 2 more items with them from the original data, which are, CompSex=Female and CompAge =[17,31)

Frequent Itemsets using both the supports are generated and sorted and some interesting ones are compiled as shown:

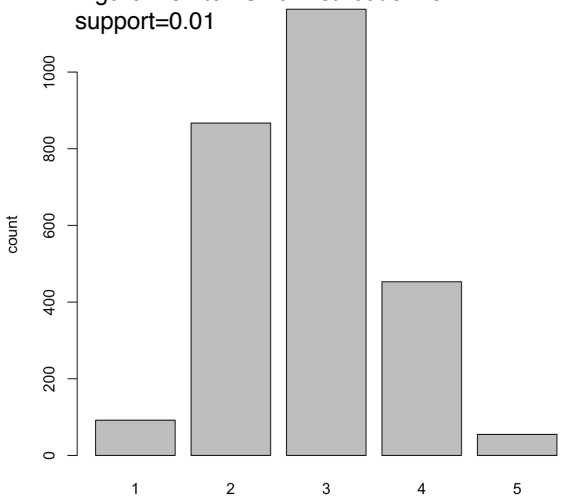
Figure 4.2: ItemSize Distribution for support=0.0005



FREQUENT ITEMSETS

items	support
{CompRace=B,Status=Suspended}	0.327
{UCROffDesc=THEFT}	0.297
{CompRace=L,Status=Suspended}	0.286
{UCROffDesc=THEFT,Status=Suspended}	0.285
{Premise=Highway, Street, Alley ETC,UCROffDesc=MOTOR VEHICLE ACCIDENT,PCClass=MB,Status=Suspended}	0.0277
{Premise=Apartment Residence,CompRace=B,UCROffDesc=BURGLARY,PCClass=F2,Status=Suspended}	0.026
{Premise=Highway, Street, Alley ETC,UCROffDesc=MOTOR VEHICLE ACCIDENT,PCClass=MB,Status=Suspended}	0.0277
{Premise=Apartment Parking Lot,Division=SouthWest,Month1=September,UCROffDesc=THEFT,PCClass=MA,Status=Suspended}	0.000517

Figure 4.3: ItemSize Distribution for support=0.01



MAXIMAL ITEMSETS

is_max <- is[is.maximal(is)]
inspect(head(sort(is_max, by="support")))

items	support
{CompRace=L,UCROffDesc=CRIMINAL MISCHIEF/ VANDALISM,PCClass=MB,Status=Suspended}	0.0321
Division=Central,CompRace=W,UCROffDesc=THEFT,PCClass=MA,Status=Suspended}	0.311
Premise=Highway, Street, Alley ETC,CompRace=L,UCROffDesc=MOTOR VEHICLE ACCIDENT,PCClass=MB,Status=Suspended	0.0277
{Month1=July,UCROffDesc=THEFT,PCClass=MA,Status=Suspended}	0.0261
Premise=Apartment Residence,CompRace=B,UCROffDesc=BURGLARY,PCClass=F2,Status=Suspended	0.0260
Division=NorthWest, CompRace=L, UCROffDesc=ASSAULT, PCClass=MC, Status=Suspended	0.00207
{CompRace=W, UCROffDesc=ROBBERY, PCClass=F2, Status=Suspended}	0.00207

CLOSED ITEMSETS

is_closed <- is[is.closed(is)]
inspect(head(sort(is_closed, by="support")))

items	support
{Status=Suspended}	0.873
{CompRace=B}	0.39
{CompRace=B,Status=Suspended}	0.327
{CompRace=L}	0.324
CompRace=L,Status=Suspended}	0.286

barplot(c(frequent=length(is),closed=length(is_closed), maximal=length(is_max)), ylab="count", xlab="itemsets")
The above barplot has been shown in Figure 4.4

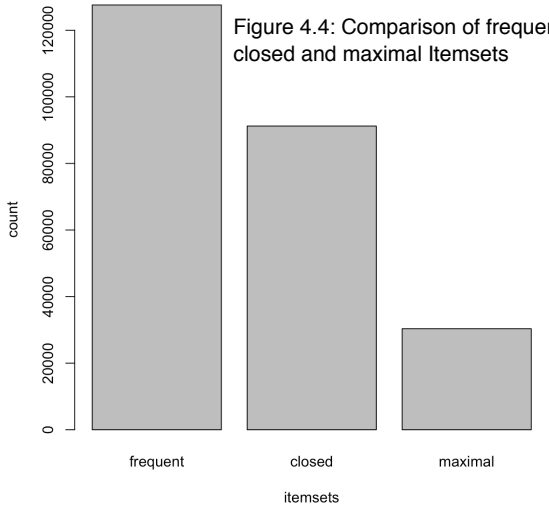


Figure 4.4: Comparison of frequent, closed and maximal Itemsets

ASSOCIATION RULES for Transaction Case1

Note: All Rules also have items CompSex=Female and CompAge=[17,31)

Filtering in the rules have been done by choosing a particular variable to be on the lhs, at times a particular variable on the rhs , at times discarding a particular entry if it is eating up other rules, limiting the number of results in a sorted order according to lift or support or phi or gini index.

rules <- apriori(trans, parameter = list(supp = .005, conf = .9))

Sort by = Lift	Min Support = 0.005	min conf=0.9	No. of Rules = 2543

lhs	rhs	support	conf	lift
CompRace=L,PCClass=F1,Status=Suspended	{UCROffDesc=ROBBERY}	0.00583	0.963	39.1
{PCClass=F1,Status=Suspended}	{UCROffDesc=ROBBERY}	0.01440	0.942	38.2
{CompRace=L,PCClass=F1}	{UCROffDesc=ROBBERY}	.00650	0.936	37.9
{Time1=18,UCROffDesc=CHILD (OFFENSES AGAINST)}	{PCClass=FS}	0.00524	0.973	14.2
{Premise=Single Family Residence - Occupied,PCClass=MC,Status=Clear by Arrest}	{UCROffDesc=ASSAULT}	0.005	1	11.9
{Month1=October,PCClass=MC,Status=Clear by Arrest}	{UCROffDesc=ASSAULT}	0.00539	0.986	11.7

Sort by = Confidence	Min Support = 0.005	min conf=0.9	No. of Rules = 2543

lhs	rhs	support	confidence	lift
{Premise=Apartment Residence,Time1=7,CompRace=L,PCClass=F2}	{UCROffDesc=BURGLARY}	0.00547	1	7.8
{Time1=16,CompRace=L,PCClass=MB}	{Status=Suspended}	0.00539	1	1.15
{Premise=Single Family Residence - Occupied,PCClass=MC,Status=Clear by Arrest}	{UCROffDesc=ASSAULT}	0.00510	1	11.86
{Division=SouthEast,PCClass=MC,Status=Clear by Arrest}	{UCROffDesc=ASSAULT}	0.00643	1	11.86
{Month1=August,CompRace=L,UCROffDesc=MOTOR VEHICLE ACCIDENT}	{Status=Suspended}	0.00591	1	1.15

We also try sorting with respect to Support, but this time, we Fix the RHS of the rule to be those particular incidents whose Status=Suspended.

Sort by = Support	Min Support = 0.005	min conf=0.9	No. of Rules = 2543

lhs	rhs	support	confidence	lift
{UCROffDesc=THEFT}	{Status=Suspended}	0.285	0.961	1.1
{PCClass=MB}	{Status=Suspended}	.25	.918	1.05
{UCROffDesc=THEFT,PCClass=MA}	{Status=Suspended}	0.203	0.973	1.12
{UCROffDesc=CRIMINAL MISCHIEF/VANDALISM	{Status=Suspended}	.176	0.917	1.05
{UCROffDesc=BURGLARY}	{Status=Suspended}	0.118	0.923	1.06

Now Removing Suspended, Robbbery, Burglary by this code: trans2 <- trans[, colnames(trans)!="Status=Suspended"] And sorting by phi and selecting 2 cases

Sort by = Phi	Min Support = 0.005	min conf=0.9	No. of Rules = 2543

lhs	rhs	support	confidence	lift	phi	gini
{CompRace=B,PCClass=MC,Status=Clear by Arrest}	{UCROffDesc=ASSAULT}	0.02046	0.986	11.69	0.472	0.0344
{CompRace=B,UCROffDesc=ASSAULT,Status=Clear by Arrest}	{PCClass=MC}	0.02046	0.979	6.29	0.332	0.0289

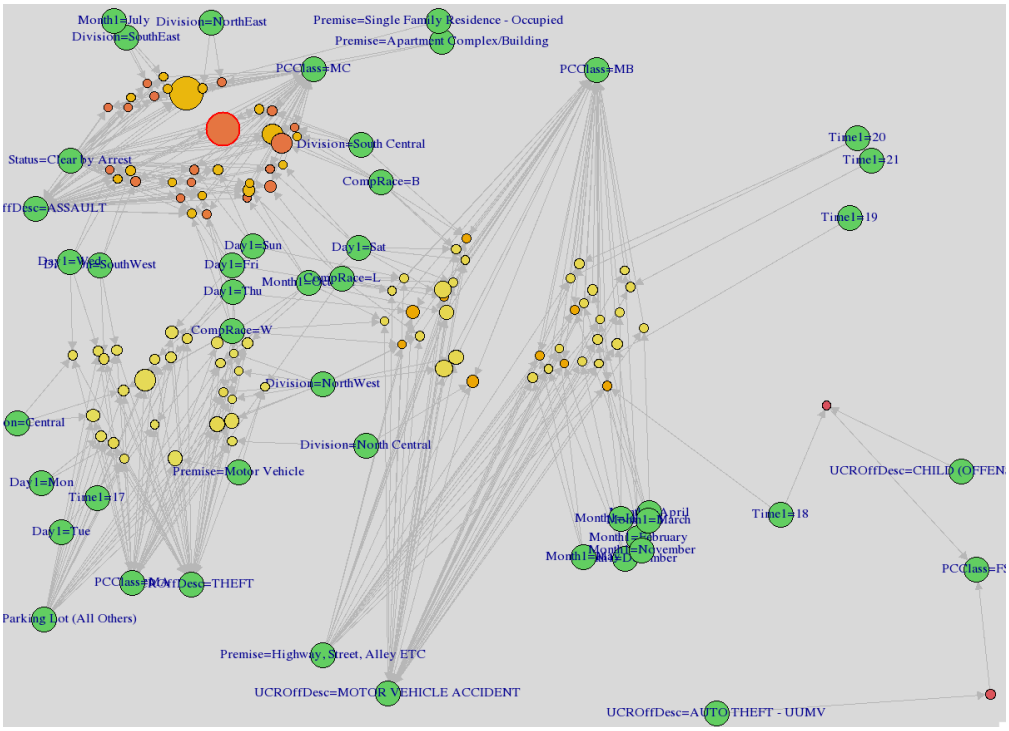


Figure 4.5: Rules Plotted Interactively for Transaction on Young Female Victim dataset

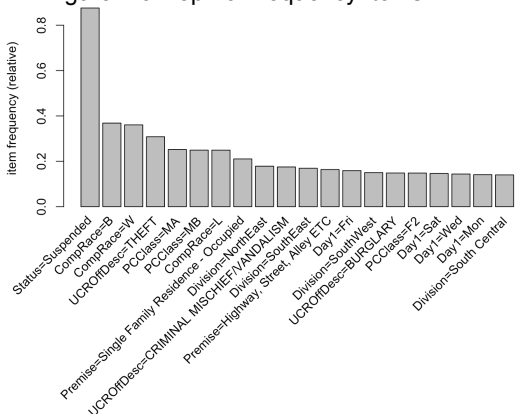
Case 2:Dataset used: data_sub_femalevictim_old

FREQUENT ITEMSETS

Item sets created for Support = 0.001. Set of 50,046 itemsets was generated.

items	support
{Status=Suspended}	0.876
{CompRace=B}	0.369
{CompRace=W}	0.361
{UCROffDesc=THEFT}	0.308

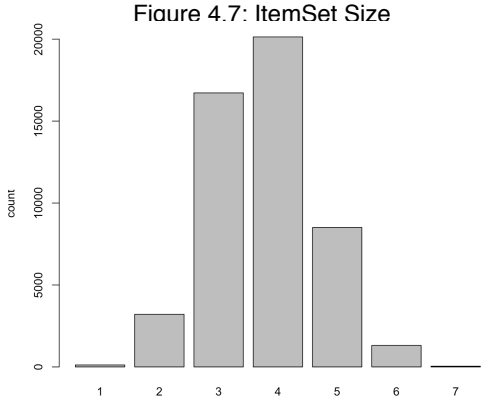
Figure 4.6: Top 20 Frequency Items



MAXIMAL ITEMSETS

items	support
{Premise=Parking Lot (All Others),Division=Central,CompRace=W,UCROffDesc=THEFT,PCClass=MA,Status=Suspended}	0.00558
{Premise=Highway, Street, Alley ETC,Division=NorthWest,CompRace=W,UCROffDesc=MOTOR VEHICLE ACCIDENT,PCClass=MB,Status=Suspended}	0.00552
{Premise=Parking Lot (All Others),Division=NorthEast,CompRace=W,UCROffDesc=THEFT,PCClass=MA,Status=Suspended}	0.00510
{Premise=Apartment Residence,Division=NorthEast,CompRace=B,UCROffDesc=BURGLARY,PCClass=F2,Status=Suspended}	0.00489
{Division=Central,CompRace=W,UCROffDesc=THEFT,PCClass=MB,Status=Suspended}	0.00473
{CompRace=L,UCROffDesc=THEFT,PCClass=MC,Status=Suspended}	0.00454

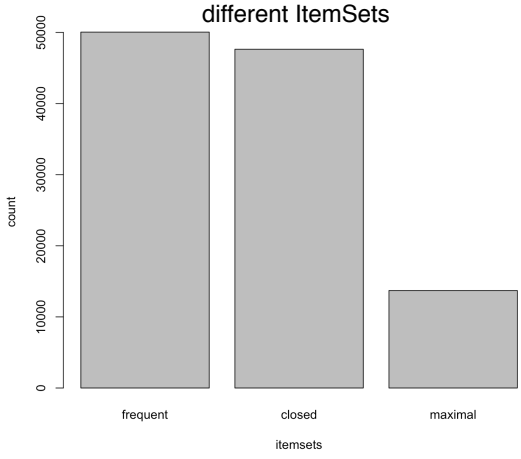
Figure 4.7: ItemSet Size



CLOSED ITEMSETS

items	support
{Status=Suspended}	0.876
{CompRace=W,Status=Suspended}	0.316
{CompRace=B,Status=Suspended}	0.316

Figure 4.8: Comparison between different ItemSets



ASSOCIATION RULES for Transaction Case2

Note: All Rules also have items CompSex=Female and CompAge=[31,100)

RULES

Sort by = Lift	Min Support = 0.001	Min Confidence=0.9	No. of Rules = 20,152

lhs	rhs	support	confidence	lift
{Time1=18,PCClass=FS,Status=Clear by Exceptional Arrest}	{UCROffDesc=CHILD (OFFENSES AGAINST)}	0.00218	1	120.3
{CompRace=L,PCClass=FS,Status=Clear by Exceptional Arrest}	UCROffDesc=CHILD (OFFENSES AGAINST)}	0.00207	0.918	110.4
{Premise=Highway, Street, Alley ETC,PCClass=F*}	{UCROffDesc=OTHER OFFENSES}	0.00305	1.000	95.3
{Premise=Single Family Residence, Occupied,Month1=December,UCROffDesc=SUDDEN DEATH}	{Status=Closed/Cleared}	0.00125	0.979	41.8
{Premise=Apartment Parking Lot,PCClass=F1,Status=Suspended}	{UCROffDesc=ROBBERY}	0.00289	1.000	36.8

Sort by = Confidence	Min Support = 0.001	Min Confidence=0.9	No. of Rules = 20,152

lhs	rhs	support	confidence	lift
{PCClass=F*}	{UCROffDesc=OTHER OFFENSES}	0.00353	1	95.35
{Time1=0,UCROffDesc=LOST PROPERTY}	{Status=Suspended}	0.00114	1	1.14
{Premise=Apartment Parking Lot,CompRace=A}	{Status=Suspended}	0.00122	1	1.14
{Premise=Single Family Residence - Vacant,PCClass=F2}	{UCROffDesc=BURGLARY}	0.01054	1	6.73
{Premise=Single Family Residence -	{PCClass=FS}	0.00106	1	10.60

Sort by = Support	Min Support = 0.001	Min Confidence=0.9	No. of Rules = 20,152

lhs	rhs	support	confidence	lift
{UCROffDesc=THEFT}	{Status=Suspended}	0.2954	0.958	1.09
{UCROffDesc=CRIMINAL MISCHIEF/VANDALISM}	{Status=Suspended}	0.1670	0.953	1.09
{Premise=Single Family Residence - Occupied,UCROffDesc=BURGLARY}	{PCClass=F2}	0.0669	0.945	6.37
{UCROffDesc=ASSAULT,Status=Clear by Arrest}	{PCClass=MC}	0.0208	0.950	7.69
{UCROffDesc=MOTOR VEHICLE ACCIDENT}	{Status=Suspended}	0.1047	0.950	1.08

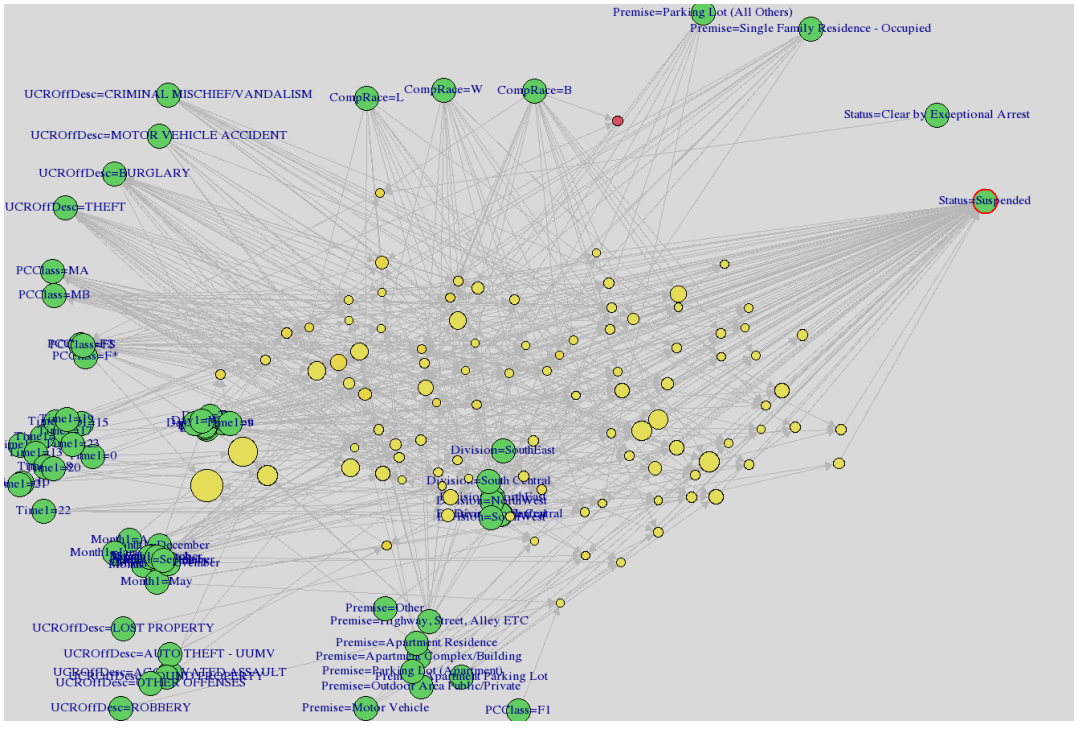


Figure 4.9: Rules Plotted Interactively for Transaction on Young Female Victim dataset

Case 3:Dataset used: data_perpdata_merged

Rules

Sort by = Lift	Min Support = 0.005	Min Confidence=0.7	No. of Rules = 104,787

lhs	rhs	support	confidence	lift
{Premise=Single Family Residence - Occupied,CompRace=B,PCClass=MA,Status=Clear by Arrest,PerpRace=Black}	{UCROffDesc=CHILD (OFFENSES AGAINST)}	0.005591054	0.7000000	30.22069
{Premise=Highway, Street, Alley ETC,PCClass=MB,PerpAgeAtArrestTime=[17,32),PerpWeight=[172,510]	{UCROffDesc=DWI}	0.005591054	0.8750000	26.08333
{Premise=Highway, Street, Alley ETC,PCClass=MB,Status=Clear by Arrest,PerpAgeAtArrestTime=[17,32),PerpWeight=[172,510]}	{UCROffDesc=DWI}	0.005591054	0.8750000	26.08333
{Premise=Highway, Street, Alley ETC,Time1=20,Status=Clear by Arrest}	{UCROffDesc=DWI}	0.005591054	0.7777778	23.18519
{Day1=Tue,CompRace=B,UCROffDesc=AGGRAVATED ASSAULT,PCClass=F2,PerpWeight=[172,510]}	{Status=Clear by Exceptional Arrest}	0.005591054	0.7777778	5.974097
{Day1=Tue,CompRace=B,UCROffDesc=AGGRAVATED ASSAULT,PCClass=F2,PerpRace=Black,PerpWeight=[172,510]}	{Status=Clear by Exceptional Arrest}	0.005591054	0.7777778	5.974097
{Premise=Apartment Residence,PCClass=F2,PerpSex=Female,PerpAgeAtArrestTime=[17,32)}	{UCROffDesc=BURGLARY}	0.005591054	1.0000000	9.858268

4. Evaluation and Analysis of Patterns found in 4

The transaction set is created by providing a minimum support, which can be any value satisfying the required analysis. I tried to stay from 0.0001 which is a low support to 0.01 which is a high support for large dataset as to encompass as many rules and itemsets as possible. Minimum support is a basic requirement for an item to be a part of a frequent item list. This list can then be filtered for specific items, to perform exact analysis, or refined to remove an item which is occupying many observations and which is not that important. This list is then sorted depending on several measures, like, lift, confidence, support, gini, phi.

High Lift implies very high correlation between the itemsets. And high confidence is also preferred, but eventually a tradeoff between both lift and confidence is selected.

Frequent items in a transaction tells us about the association between the data.

To determine association rules, from these items, we use the Apriori algorithm. Apriori algorithm requires some parameters like support and confidence and the generated rules can then be sorted according to different metrics like the lift or the confidence or the support or gini or phi.

How are these patterns useful?
Well, i few have information about how commonly do these attributes occur together in patterns. And if the data is about crime, we can find patterns as accurate as the time frames when certain types of crimes happen in particular areas over a certain demographic of victim. This is very precise information which is not that difficult to generate from large dataset, and these patters, if occur very frequently, can be of immediate use for the law enforcement to reduce crime.

Analysis for Case 1 and Interesting Findings:

For the first dataset in 4. i.e. data_sub_femalevictim_young

From Figure 4.1, we notice that Most of the incidents were suspended, but there were several other common items like the race of the complainant being black, Latino and White, which is true in the case of our dataset as mentioned in the earlier project. There were also several cases of Theft, and Vandalism and a very common premise of crime was the Highway.

Case 1 ItemSets:

Notes on Frequent Itemset table:

Some Frequent Item sets involve young women under the age of 30 being victims of theft and the case was suspended.

Some involve young women under the age of 30 being victims in a Motor Vehicle Accident on a Highway and the case not being solved and suspended.

Some involve young Black women under the age of 30 being victims of Burglary in their Residence Apartments and the case being suspended.

In all, there were several cases of Theft, Robbery and Burglary being suspended

Some Frequent Item sets involve young women under the age of 30 being victims of theft.

Notes on Maximal Itemset table:

Some Maximal Itemsets involve young Latino women under the age of 30 being victims of Criminal Mischief and Vandalism and the case was suspended

Some Maximal ItemSets show White Women under the age of 30 in the Central Division being victims of Theft in Misdemeanour class A and the case was suspended

Others show Latino Women in the NorthWest Divison being victims of Assault with the Penal Code Class: Misdemeanour Class C and Status of case report as suspended.

Notes on Closed Itemset table:

Some Closed Itemsets involve young black women under the age of 30 being victims of crime which couldn't be solved and was suspended. A similar observation is made about several Latino women under the age of 30.

Case 1 Rules:

There are several metrics which can be used to sort and order the rules generated. I have used Lift Confidene Support and Phi for the 2543 rules generated .

Some interesting rules found are:

Latino Women under the age of 31 were more susceptible to being victims of Robbery which often goes unsolved, not only robbery, but also Burglary in the premise Apartment Residence.

There are several cases involving a child victim called in by young women, particularly between 6pm and 7pm. This is in conjunction with my earlier findings on ChildVictims and is incredibly important information for Child Protective Services

Young women were assaulted in the premise Single Family Residence, and the assault led to an arrest, thus the police is doing a good job in such cases. Important information for young women, in case an attempt to assault them is made they can rely on the law enforcement agencies.

Several instances of theft and burglary involving younger women go unsolved and end up being suspended. An earlier analysis on the response time also highlighted the fact that for cases involving theft and burglary, police takes hours to respond, but for cases of violent crimes, like murder or assault, they respond much faster, which makes sense, as the law enforcement has limited resources and they have to prioritise. This is in conjunction with an earlier analysis showing that women were targeted more for theft and burglary and assault.

The Figure 4.5 shows these same rules in a graph format.

It also tells us that there the premise Highway, Street, Alley, etc is very likely to be a location for UCROOffDesc=Motor Vehicle Accident which is also a misdemeanour in most cases.

It informs us that most cases of assault leaded were cleared by arrest, which is good news.

Analysis for Case 2 and Interesting Findings :

From Figure 4.6, we notice that Most of the incidents were suspended, but there were several other common items like the race of the complainant being black, Latino and White, which is true in the case of our dataset as mentioned in the earlier project. There were also several cases of Theft, and Vandalism and a very common premise of crime was an occupied single family residence followed by the Highway.

Notes on Frequent Itemset table:

Most Frequent Item sets involve older women over the age of 30 being victims of a crime, status of which was suspended.

There were several in which older women fell victim to Theft

From the data set, Black and White older women were more susceptible to being the victim of a crime. This rule doesn't give a great deal of insight as the dataset is skewed with Black White and

Notes on Maximal Itemset table:

Older White Women are more susceptible to theft in Parking Lots, and the cases go unsolved. So, this is an important piece of information which can help older White women stay vigilant when traveling in and out of a parking lot.

Older White women are also more likely to be victims of Motor Vehicle Accident, which is classified as a misdemeanour and left unsolved and suspended most times.

Some Maximal Itemsets involve young Latino women under the age of 30 being victims of Criminal Mischief and Vandalism and the case was suspended

Some Maximal ItemSets show White Women under the age of 30 in the Central Division being victims of Theft in Misdemeanour class A and the case was suspended

Others show Latino Women in the NorthWest Divison being victims of Assault with the Penal Code Class: Misdemeanour Class C and Status of case report as suspended.

Older Black women are more susceptible to fall victim to a burglary in their apartments

Notes on Closed Itemset table:

Some Closed Itemsets involve older white and black women over the age of 30 being victims of crime which couldn't be solved and was suspended. A similar observation is made about several Latino women under the age of 30.

Case 2 Rules:

Some interesting rules found are:

A similar result, there were several child offences which resulted in an arrest is noticed to be between 6pm-7pm

There were several cases of older women living in Single Family Residence, the crime report for which specifies, Sudden Death. These women most probably died of natural causes due to old age

There were many instances of Felony F1, Robbery in Apartment Parking Lot , which wee left unsolved.

Three were several burglaries reported in single family residences.

And some similar results like case 1.

Figure 4.9 also shows that most cases involved White, Black, Latino Women, shows most cases being suspended involving theft/burglaries. It shows single family residence and parking lot as most common places of being attacked

Comparison between the 2 cases:

Both these cases involved crimes committed against women, first one against young women and the second one against an older one. The first case involving young women were more prone to theft, robbery, assault and motor vehicle accidents.

The second case involved more missing items, criminal mischief, vandalism, burglaries, child offences and most importantly, the second case is filled with instances of Case Clearing due to Sudden death of an old woman.

Analysis Case 3:

Black man arrested for crime against a black child at single family residence. Sounds like a black father hit his child.

There were lots of DWI which resulted in arrests on highways for perpetrator for age group [17,32) and PerpWeight group [172,510] which is that of a young adult male. Law enforcement doing a good job keeping these people behind bars.

Black male arrested for aggravated assault in the weight range of [172,510]

Apartment Residences more prone to burglary with perpetrators being young women between the age of 17 and 32.

Additional Recommendations for Stakeholders:

The analysis shown above is important from a stakeholders point of view.

From the law enforcement stakeholder point of view, statistics about assault incidents resulting in arrest is an important piece of information, as it tells them that they are ensuring women safety.

From a female citizen stakeholder point of view, they know that police responds well and fast to assault and other violent crime based on response time analysis and these incidents result in arrest.

For Older Women, two important things were noticed, first, they were victims of theft in Parking Lots, so important safety information. Furthermore, several incidents of sudden death recorded, which if at rise compared to previous years, they can start monitoring their health in a better fashion.

5. Deployment

This algorithm can easily run on most data sets. The IT department at Dallas Police HQ can have it setup.

They can use apriori algorithm for association analysis, which is a very straightforward mechanism.

As explained, first data has to be prepared, cleaned, discretized. And logical variables can be made for specific purposes.

Various subsets can be made of the data, for targeted analysis using code available on [3]

Case 1 incorporates code, and earlier projects explain how to perform cleaning.

Once, the data is prepared, a transaction can be generated for the data. This transaction can then give rules based on a user defined minimum support and minimum confidence.

Support and Confidence can be changed to broaden or narrow down the scope.

Then the generated transaction contains things on the left hand side(lhs) which ideally lead to things on th right hand side(rhs). You can also pin specific variables on the right hand side, like if you pin arrest on right hand side, you come across all data in which lhs leads to arrest information.

6. References

[1] https://www.dallasopendata.com/Police/Dallas-Police-Public-Data-RMS-Arrest/r4wm-ig9m).

[2] http://www.inside-r.org/packages/cran/arules/docs/discretize

[3] http://michael.hahsler.net/SMU/EMIS7332/