Association Rule Mining
Data Warehousing and Data Mining
Project Phase 2

Thapaswini Chowdary - 20161066
Jayitha .C - 20171401

September 18, 2018

Abstract

In the second phase, we have attempted to study the homicide dataset by finding frequently occurring item sets, association rules and their support and confidence. The algorithms we used are FP-Growth for finding the frequent item sets and Apriori for mining association rules. The programming environments used are Python and R.

1 Data Set Description

We have chosen to use the homicide data set which is what we used for the Phase 1 submission as well. We chose this data set because it is easy to understand, the number of attributes aren't overwhelming but just enough to enable us to run some fun queries to yield interesting results some of which are historically relevant.

1.1 List of Items

Table 1 shows the attributes/fields and their corresponding domains, the union of all these domains will make the overall item set or the total list of items.

Therefore the total number of items in the dataset are

$$11 + 12 + 31 + 6 + 11 + 3 + 50 + 27 + 854 + 1262 + 3 = 2270$$

1.2 Transaction Description

Every transaction in this dataset selects one value from the domain of each attribute. So every transaction is a record of a single person's death details like the person's name, age, race, city and state or more accurately the location of death, date of death, exact coordinates at which the body was found at and the state of the case.

Table 1: Homicide Dataset

Year	Month	Day	Race	Age	Sex	City	State	Latitude	Longitude	Disposition
Y2007	M1	D1	Asian	A_10(0-9)	Female	Albuquerque	AL	LAT00.00	LON000.00	Closed by Arrest
Y2008	M2	D2	Black	A_20(10-19)	Male	Atlanta	AZ	LAT25.73	LON-122.51	Closed without Arrest
Y2009	M3	D3	Hispanic	A_30(20-29)	S_Unknown	Baltimore	CA	LAT25.74	LON-122.5	Open/ No Arrest
Y2010	M4	.	White	A_40(30-39)		.	CO	LAT25.75	LON-122.49	
Y2011	M5	.	Other	A_50(40-49)		.	DC	.	.	
Y2012	M6	.	R_Unknown	A_60(50-59)		.	.	.	.	
Y2013	M7	.		A_70(60-69)		.	.	.	.	
Y2014	M8	.		A_80(70-79)		.	.	.	.	
Y2015	M9	.		A_90(80-89)		Stockton	.	LAT45.03	.	
Y2016	M10	D29		A_100(90-99)		Tampa	TX	LAT45.04	LON-71.05	
Y2017	M11	D30		A_110(100-102)		Tulsa	VA	LAT45.05	LON-71.04	
	M12	D31				Washington	WI	LAT45.06	LON-71.02	
11	12	31	6	11	3	50	27	854	1262	3

Table 2: Example Transaction

Year	Month	Day	FirstName	LastName	Race	Age	Sex	City	State	Lat	Long	Disposition
2010	5	4	Garcia	Juan	Hispanic	78	Male	Albuquerque	NM	35.1	-106.54	Closed without arrest

Consider the transaction in Table 2. This record tells us that a person named Garcia Juan, a 78 year old Hispanic male died at Albuquerque, NM, specifically at Lat 35.1 and Long -106.54 on 4^{th} May, 2010. The case is reported be solved without any suspect being arrested.

In the dataset provided there are 52179 records (transactions) available and our study will be limited to these records only.

1.3 Data Set Characteristics

The dataset provided has the following characteristics:

- Number of Records: 52179
- Number of Fields: 13 (Out of which we exclude names, only 11 considered)
- Number of Items: 2270

The Data Set Characteristics are summarized in Table 1

1.4 Data Cleaning

To get better results we have performed the following procedures on the dataset

- Date - the date which was originally a composite attribute, has not been broken down to three new attributes - year, month and date. This will help us get aggregated results that are a grouping of one of the three attributes
- Latitude and Longitude - We have rounded off the latitude and longitudes upto two places after the decimal point to yield better results

- Age - instead of using the exact age, we have quantized the age further to take on range values, so if an age value says 70, it means that the actual age of the person is between 60 and 69 (inclusive)
- Unknowns - we have not cleaned up the unknowns as we feel that cleaning them up by any known procedures such as replacing the unknown by mode or median might falsely affect the data. Hence the assumption we are making is that the distribution within the unknowns is the same as that of the distribution we obtain by removing all transaction with the unknown values
- Names - We have stripped the names from the dataset. We might be missing out on interesting results such as a series of death of people with the same name, but this pattern is highly unlikely.
- Ambiguity - To avoid ambiguity when frequent item sets or rules are found in cases where the domains of attributes overlap, we've appended item values with indicators which tell us which field it belongs to as shown in Table 3
  Note: No unknown values in Date, Disposition, City or State

Table 3: Augmented Data

Attribute	Value	Augmented Value
Year	****	Y****
Month	**	M**
Date	**	D**
Age	***	A***
Latitude	**.**	LAT**.**
Longitude	***.**	LON***.**
Race	Unknown	R_Unknown
Age	Unknown	A_Unknown
Sex	Unknown	S_Unknown
Latitude	Unknown	LAT00.00
Longitude	Unknown	LON000.00

2 Algorithms and Tools Used

We do not claim to have written the packages or created the tools that we have used in this assignment.

2.1 Frequent Item Sets

2.1.1 Algorithm - FP-Growth

The algorithm we used to find the frequent item sets is the FP-Tree algorithm. The pseudo code for FP-Growth is given in Fig 1, this is post creation of the the FP-Tree.

The FP-Growth Algorithm is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for

```
Procedure FP-growth (Tree, α)
{
(1) if Tree contains a single path P
(2) then for each combination (denoted as β)
    of the nodes in the path P do
(3)     generate pattern β ∪ α with support =
            minimum support of nodes in β;
(4) else for each a_i in the header of Tree do {
(5)     generate pattern β = a_i ∪ α with
            support = a_i.support;
(6)     construct β's conditional pattern base and
            then β's conditional FP-tree Tree_β;
(7)     if Tree_β ≠ ∅
(8)     then call FP-growth (Tree_β, β)          }
}
```

Figure 1: FP-Growth Algorithm

storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree).

2.1.2 Tools

We used code already available at this link - https://github.com/enaeseth/python-fp-growth We do not claim to have written this code. Given the data set and a minimum support the code returns a list of frequent item sets and their support.

2.2 Association Rule Mining

2.2.1 Algorithm - Apriori Algorithm

The algorithm we used to find association rules within the dataset is the Apriori algorithms. The pseudo code for the algorithm is given in Fig 2.

```
Apriori (T, ε)
    L_1 ← { large 1-itemsets that appear in more than ε transactions }
    k ← 2
        while L_{k-1} ≠ ∅
            C_k ←Generate(L_{k-1})
            for transactions t ∈ T
                C_t ←Subset(C_k, t)
                for candidates c ∈ C_t
                    count[c] ← count[c] + 1
            L_k ← {c ∈ C_k | count[c] ≥ ε}
            k ← k + 1
    return ⋃_k L_k
```

Figure 2: Apriori Algorithm

The algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

2.2.2 Tools

We used the **arules** package that is made available with the programming language **R**. When using this package we get a very good summarization of the results including support, confidence, lift and several other parameters that we've used to study the data

3 Results

3.1 Aggregate Count of each Item in Data Set

Listed below are the frequencies(support) of selected single item sets in the data set. These aggregates have been found using both FP-Growth and Apriori. They have been summarized in tables 4 , 5 and 6

Table 4: Date Supports

Date	Support	Month	Support	Year	Support
D1	1953	M7	5034	Y2016	6290
D5	1807	M8	4789	Y2015	5336
D23	1779	M6	4654	Y2017	5048
D26	1732	M5	4620	Y2012	4908
D21	1731	M9	4553	Y2010	4610
D18	1727	M10	4528	Y2014	4608
D17	1722	M12	4265	Y2013	4599
D8	1717	M11	4237	Y2011	4562
D10	1716	M4	4202	Y2007	4268
D20	1710	M1	4037	Y2008	4215
D30	1709	M3	4031	Y2009	3735
D7	1706	M2	3229		
D14	1706				
D16	1700				
D19	1699				
D28	1686				
D24	1686				
D2	1682				
D25	1678				
D11	1671				
D29	1669				
D4	1662				
D22	1662				
D12	1656				
D9	1650				
D13	1635				
D15	1632				
D27	1626				
D3	1613				
D6	1599				
D31	1257				

The FP-Growth algorithm was forced to work on the dataset whilst grouping the date and Location attribute, this yielded the results in Table 7. This table is an indicator of mass homicides that occurred.

Table 5: Location Supports

City	Support	State	Support
Chicago	5535	CA	6288
Philadelphia	3037	TX	5891
Houston	2942	IL	5535
Baltimore	2827	PA	3668
Detroit	2519	MO	2867
Los Angeles	2257	MD	2827
St. Louis	1677	MI	2519
Dallas	1567	TN	2281
Memphis	1514	FL	2120
New Orleans	1434	LA	1858
Las Vegas	1381	OH	1778
Washington	1345	NV	1381
Indianapolis	1322	DC	1345
Kansas City	1190	IN	1322
.	.	OK	1255
.	.	GA	1219
.	.	NY	1148
.	.	wI	1115
Omaha	409	NC	963
Albuquerque	378	AZ	914
Long Beach	378	AL	801
Sacramento	376	MA	614
Minneapolis	366	KY	576
Denver	312	VA	429
Durham	276	NE	409
San Bernardino	275	NM	378
Savannah	246	MN	366
Tampa	208	CO	312

Table 6: Other Supports

Race	Support	Age	Support	Sex	Support	Disposition	Support
Black	33361	A30	18074	Male	40739	Closed by arrest	25674
Hispanic	6901	A40	9872	Female	7209	Open/No arrest	23583
White	6333	A20	8228	S_Unknown	4231	Closed without arrest	2922
R_Unknown	4199	A50	5953				
Other	700	A60	3539				
Asian	685	A_Unknown	2999				
		A70	1357				
		A10	986				
		A80	504				
		A0	385				
		A90	240				
		A100	40				
		A110	2				

Table 7: Historical Events Detected by FP-Growth

Year	Month	Date	State	City	Support
2017	10	1	NV	Las Vegas	60
2015	12	2	CA	San Bernardino	14
2013	9	16	DC	Washington	13
2015	9	2	IL	Chicago	10
2017	10	31	NY	New York	9
2016	9	5	IL	Chicago	9
2016	2	4	IL	Chicago	9
2011	1	19	PA	Philadelphia	9
2015	4	30	AZ	Phoenix	9
2014	12	31	AZ	Phoenix	9

- The 2017 Las Vegas shooting was a mass shooting on the night of October 1, 2017, when a gunman opened fire on a crowd of concertgoers at the Route 91 Harvest music festival on the Las Vegas Strip in Nevada.
- on December 2, 2015, 14 people were killed in a terrorist attack consisting of a mass shooting and an attempted bombing at the Inland Regional Center in San Bernardino, California.
- The Washington Navy Yard shooting occurred on September 16, 2013, when a lone gunman, 34-year-old Aaron Alexis, fatally shot 12 people in a mass shooting at the headquarters of the Naval Sea Systems Command (NAVSEA) inside the Washington Navy Yard in Southeast Washington, D.C.
- On October 31, 2017, an Islamist terrorist drove a rented pickup truck into cyclists and runners for about one mile (1.6 kilometers) of the Hudson River Park's bike path.
- Chicago displays a sudden surge in homicide deaths, reasons are yet to be known but it has in general been attributed to the slight levy the laws provide there where punishments for owning firearms isn't as harsh as it is in other cities
- ...

3.2 Interesting Results

- **{} => {Male}**
  Support - 0.780 Confidence - 0.780

  According to Fig 3 and Fig 4, the male to female ratio in the USA between the years 2007-2017 has always been less than one i.e. the number of females has always been more than the number of males but according to this data set the number of male deaths are higher than the number of female ones and hence this leads us to believe that the likelihood of a female dying by homicide is a lot less than that for a male

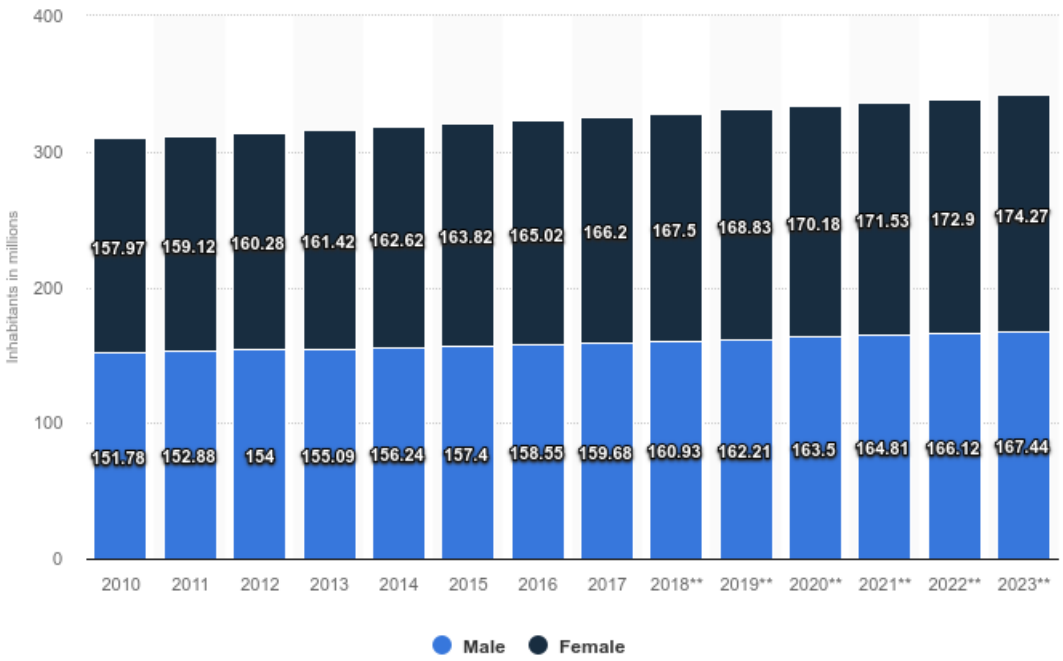


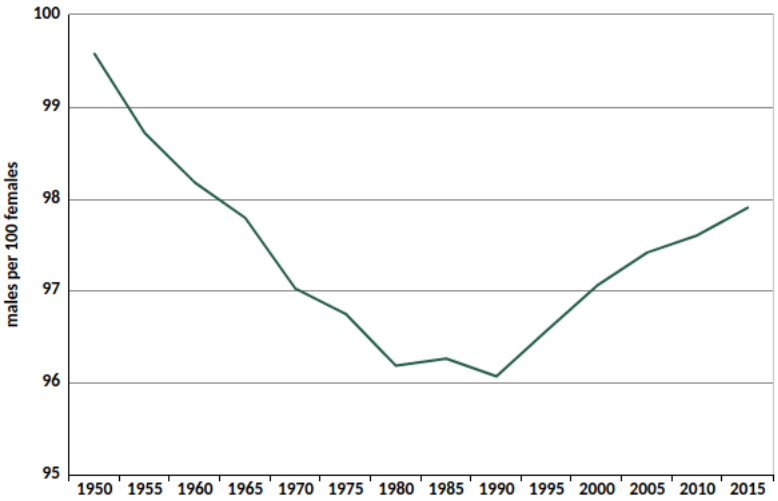
Figure 3: The statistic shows the total population in the United States by gender from 2010 to 2016, with projections up until 2023. In 2017, 166.2 million of the total population of the U.S. were female.

- **{} => {Closed by Arrest}**
  Support - 0.492 Confidence - 0.492
  This result indicates that most cases are closed by arrest, which is good news relatively

- **{Male} => {Open/ No Arrest}**
  Support - 0.381 Confidence - 0.488
  The confidence tells us that 50% of all Male homicide cases are left open, this is interesting because we get this result after removing the bias that the deaths of males by homicide is more

- **{} => {A30}**
  Support - 0.346 Confidence - 0.346
  This result tells us that most death occur among people between the ages of 20-29. One reason could be because most homicides occur in public places and the population within this age range tend to be more out going and hence the probability of finding them at these locations is higher - like the concert shooting in LA

- **{Male} => {A30}**
  Support - 0.297 Confidence - 0.380
  This result tells us that young males between the ages of 20-29 are highly comparatively more likely to get killed in a homicide



Date	Value	Change, %
2015	97.90	0.31 %
2010	97.60	0.19 %
2005	97.41	0.37 %
2000	97.05	0.51 %
1995	96.57	0.52 %
1990	96.07	-0.20 %
1985	96.26	0.08 %
1980	96.18	-0.58 %
1975	96.75	-0.28 %
1970	97.02	-0.79 %
1965	97.79	-0.39 %
1960	98.17	

Figure 4: Male to female ratio of the total population of the United States of America

- **{} => {Black}**
  Support - 0.639 Confidence - 0.639
  This result is very interesting as according to Fig 5 the number of white people in the USA exceeds the number of black people, so the surge in homicide deaths among black people leads us to believe that these homicides occur in certain neighborhoods that are primary residents of the black population

- **{Male} => {Black}**
  Support - 0.56 Confidence - 0.718
  This result is by far the most interesting as it contradicts the population distribution we've found so far. The population of females and white people is supposed to be high, but there seem to be more black male deaths. This result cements the results we've gotten so far.

- **{Black} => {No Arrest}**
  Support - 0.314 Confidence - 0.491
  This result tells us that 50% of all black cases are closed either without proper investigation or that they get "settled" without the interference of the police department

- **{} => {California}**
  Support - 0.120 Confidence - 0.120
  This result that we obtained from both frequent item set generation and association rule mining tells us that either there is an outbreak of homicide deaths in California over other states

- **{Dallas} => {A_Unknown}**
  {Dallas} => {R_Unknown}
  {Dallas} => {S_Unknown}
  {Dallas} => {L_Unknown}

Race and Hispanic/Latino origin	Census 2010, population	Percent of population	Census 2000, population	Percent of population
Total Population	308,745,538	100.0%	281,421,906	100.0%
Single race				
White	196,817,552	63.7	211,460,626	75.1
Black or African American	37,685,848	12.2	34,658,190	12.3
American Indian and Alaska Native	2,247,098	.7	2,475,956	0.9
Asian	14,465,124	4.7	10,242,998	3.6
Native Hawaiian and other Pacific Islander	481,576	0.15	398,835	0.1
Two or more races	5,966,481	1.9	6,826,228	2.4
Some other race	604,265	.2	15,359,073	5.5
Hispanic or Latino	50,477,594	16.3	35,305,818	12.5

Figure 5: Population of the United States by Race and Hispanic/Latino Origin, Census 2000 and 2010

{Dallas} => {F_Unknown}

This result tells us that either the reports from Dallas are flawed or this is a quite an anomaly and better investigation measures need to be taken.

Similar results were found for Los Angeles
{Los Angeles} => {L_Unknown}
{Los Angeles} => {F_Unknown}

- **{} => {L_Johnson}**
  This fun result just lets us know that Johnson is a frequently occurring last name, no pattern was found that indicated that the name had anything to do with the homicide death patterns

- Table 8 and Table 9 gives us a summarized view of what the homicide death rates in Chicago look like

  These results showed that there was a huge surge in homicide deaths in Chicago in 2016. Upon further research, it has been found that the cause of this sudden surge is yet not known and also that murder rates in Chicago vary greatly depending on the neighborhood in question. Many neighborhoods on the South Side tend to be poorer, less educated, predominantly African American, and infested with street gangs.

Table 8: Chicago Homicide Death Patterns as per Year and Month

Year	Month	Support
2016	8	95
2017	6	85
2016	10	80
2016	11	78
2017	7	77
2016	6	75
2013	2	13
2013	3	16
2015	2	19
2014	1	20
2014	2	20
2008	2	20

Table 9: Chicago Homicide Death Patterns as per Year

Year	Support
2016	765
2017	654
2008	510
2012	505
2015	487
2009	457
2007	445
2011	437
2010	434
2013	423
2014	418

4 Support and Confidence

4.1 FP-Growth

- The support for the FP-Growth algorithm was calculated as the number of times the frequent item set occurs in the given transaction set (item set)

4.2 Apriori Algorithm

- The support is calculated as the number of times the LHS and RHS item sets occurred together divided by the total number of transactions
- The Confidence is calculated as the number of times LHS and RHS item sets occurred together divided by the number of times the LHS item set occurred in the transaction set

It was observed that when the support was set to some high value then most of the rules generated were some form of splits of the frequent item sets whereas when the confidence was

set high but support set low, we were able to find association rules which "very strongly held true"

5 Conclusion

In conclusion, we were able to use FP-Growth and Apriori to mine frequent item sets and association rules that yielded very interesting results. Apart from making new observations we were also able to mine information relevant to events that have already occurred.