**Observation for Project 1**

**Association Rule Mining**

We have considered the crime report dataset to perform association rule mining for analyzing the dataset and extract the nature and frequency of crimes in various districts.

**Dataset**

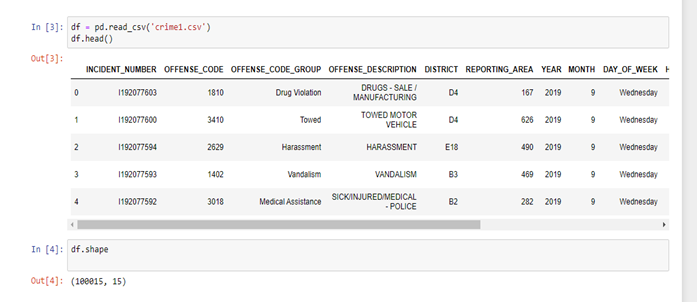
1. This data set is obtained from Kaggle

-<https://www.kaggle.com/AnalyzeBoston/crimes-in-boston>

2. The dataset contains 100k unique rows with 15 attributes indicating the crime report filed in various districts.

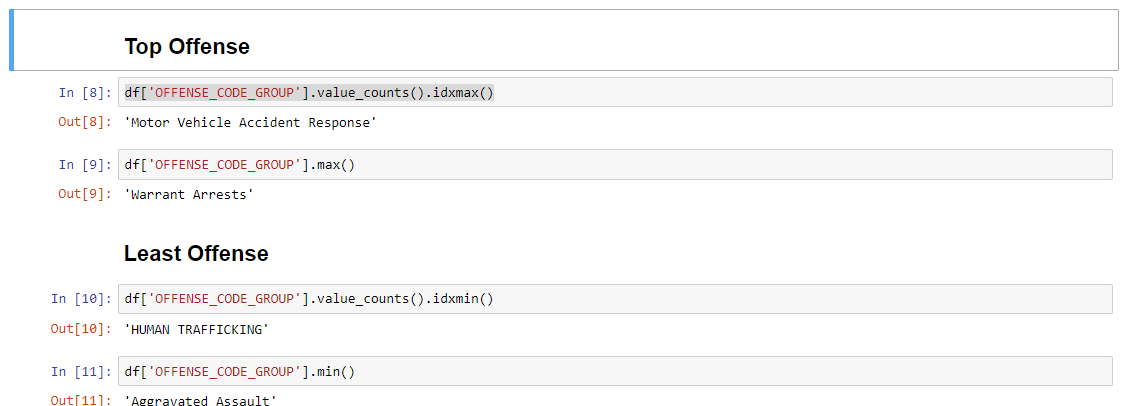
3. The association rule here is performed for the attribute DISTRICT based on the unique OFFENSE\_CODE\_GROUP attribute.

4. There are 12 DISTRICT values and 64 OFFENSE\_CODE\_GROUP values.



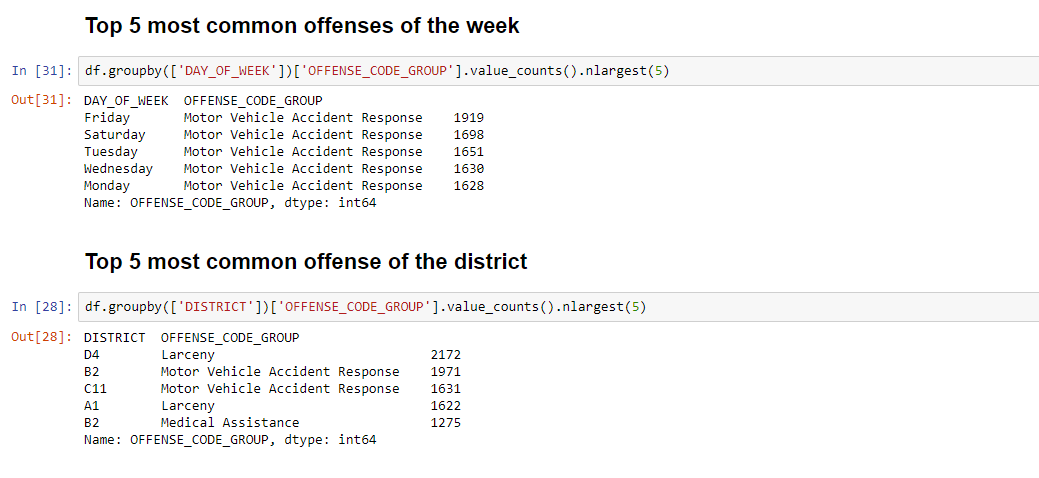
**Pre-Processing and Understanding our Dataset**

1. Using the .*value\_counts()* and *.idxmax() and idxmin()* function and applying it to the columns offense\_code\_group and district we are able to see visually which itemset where amongst the most popular offenses and most least occuring offense



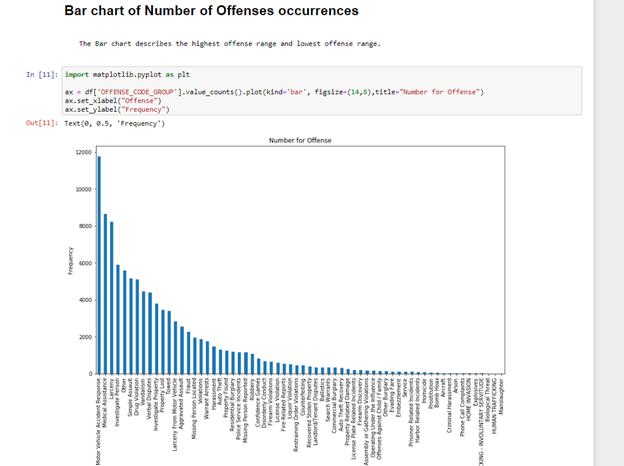
* Identifying the top most common offenses of the district using the function .*nlargest()* setting the parameter to a 5 for top 5 most common offenses.

1. The day of the week which is most popular is Friday with an increasing occurrence of Motor Vehicle Accident Response. This result demonstrates that vehicle accidents are amongst the most popular occuring.
2. The association between district and offense code group we can see that D4 is on top with Larceny. We can also see an association between the number of Motor vehicle accidents also implying there are medical assistance for those involved in the offense.
3. A table version of the bar graph depicted above can be found through our .value\_counts() function ran against the offense\_code\_group in which we can see visual numeric counts
4. Similar to the bar chart we can depict Motor vehicles stands as the highest occurring offense amongst the crime dataset.



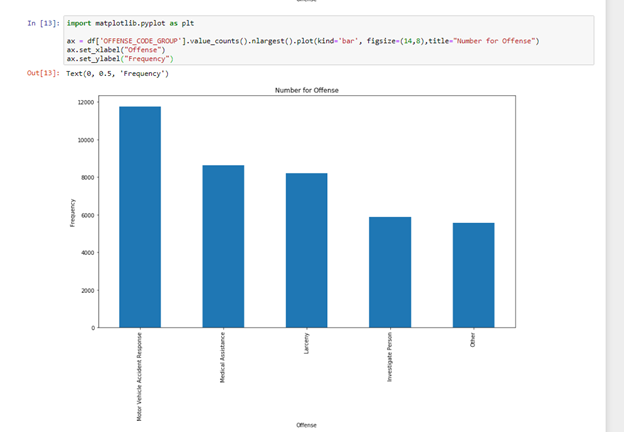
* **Visualization of the Offences**

1. Through the *.plot()* function we are able to display visually how our top offenses stack against other data. This allows us to compare and contrast later in our research the accuracy of our frequent itemset once applying apriori algorithm
2. Motor Vehicles Accident Response has the largest amount of occurrences in our dataset. Even amongst other districts it seems to be the highest documented incident occurring alongside medical assistance.



* **Bar chart with a parameter of nlargest(5) top 5.**

1. The chart below is a visual output of the top 5 most frequent occurrences of offenses occurring across the dataset.
2. Motor Vehicle accident response (10000 - 12000) , Medical assistance (8000- 10000) , Larceny (8000- 10000), Investigate person (6000) are amongst the most popular occurrences.
3. Correlation amongst all 5 offense occurrences have a strong association to one another in terms of offense description. Those who are involved in a motor vehicle accident are more likely to require medical assistance

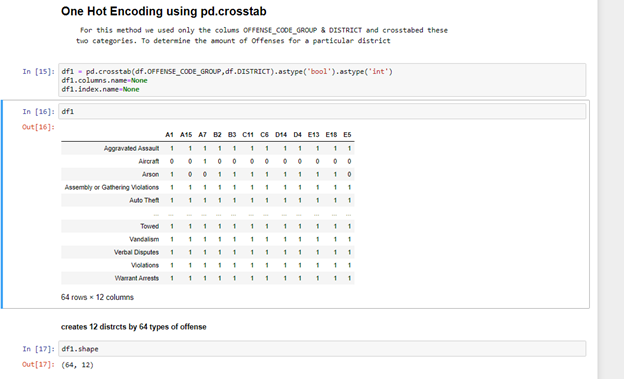


* **Additional Pre-Processing and One-Hot Encoding**

1. To apply the apriori algorithm it states that the dataset to be used must be in all numeric values. Using the DISTRICT and OFFENSE\_CODE\_GROUP column we apply a one-hot encoding method using pd.crosstab and *.astype()* function to set our values to just true and false values and then setting those string values to just integers of 1 and 0.
2. **Dropping Itemsets**: Before applying the algorithm we also needed to **drop** a few columns from our encoded dataset .

* **One Hot Encoding**

1. One Hot encoding is a technique of converting the dataset into binary format.
2. For the occurrence of offence in districts the value is 1 and if no occurrence then 0.
3. Using crosstab the attributes OFFENSE\_CODE\_GROUP and DISTRICT are tabularized.
4. Processing our dataset as numeric values of 1’s and 0’s will allow for as a sufficient argument to be passed through the apriori algorithm and obtaining more concrete results for frequent itemsets and rules passed. Allowing for percentage values when support and confidence is acquired.

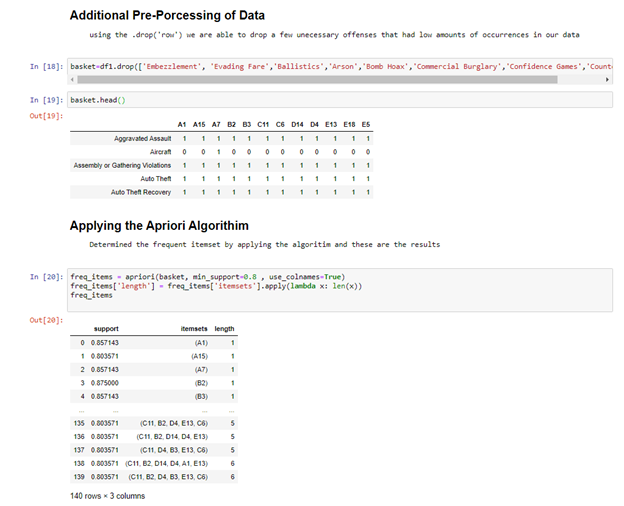


**Apriori Algorithm**

1. It is an [algorithm](https://en.wikipedia.org/wiki/Algorithm) for frequent item set mining and [association rule learning](https://en.wikipedia.org/wiki/Association_rule_learning) over relational [databases](https://en.wikipedia.org/wiki/Databases). It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine [association rules](https://en.wikipedia.org/wiki/Association_rules) which highlight general trends in the [databas](https://en.wikipedia.org/wiki/Database)e.
2. In apriori, all non-empty subset of frequent itemset must be frequent. Using this approach, there is a reduction in the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count.
3. This algorithm derives frequently occured items which are paired to form itemset. Using the support and confidence the most important itemsets are categorized.

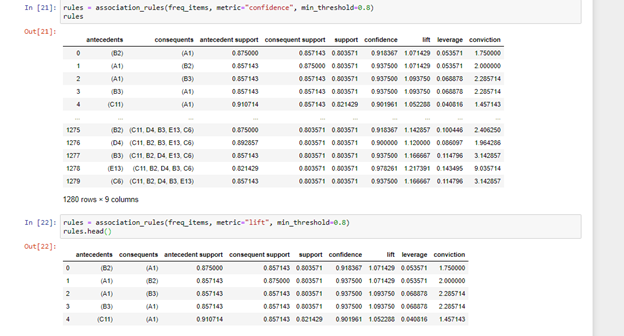
* **Support**

1. The highest support value for the itemsets that occured is 0.8 and highest confidence value is 0.9.
2. The length of the itemset with highest support values does not exceed 6.
3. So as the itemset length increases, the support decreases.



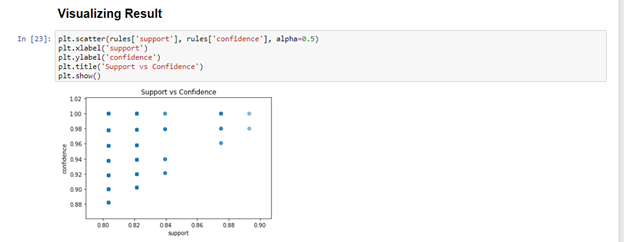
* **Confidence and Lift metric**

1. The minimum support threshold is 0.8
2. The lift metric threshold is 0.8



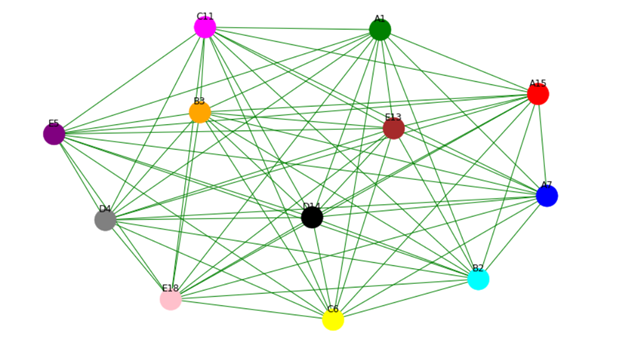
* **Support vs Confidence**

1. The plot indicates that as the support increases the confidence also increases but the frequency of occurrence decreases.



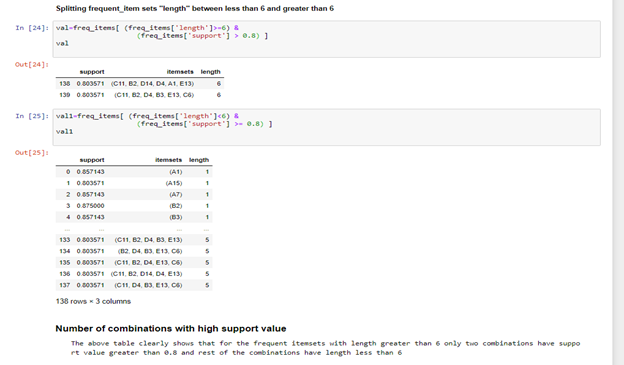
* **A Network Graph Between Two Frequencies**

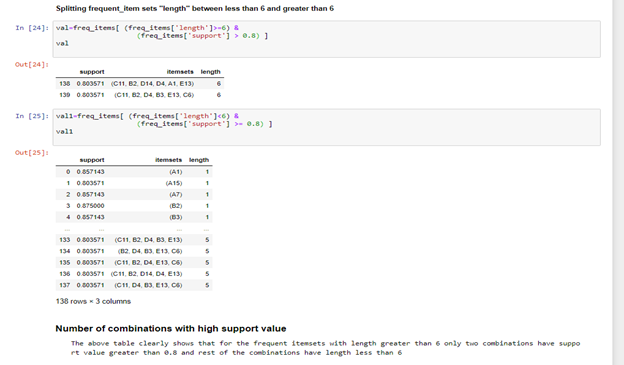
1. Using the information acquired for frequent\_items we are able to depict a network graph
2. From this network graph, we can predict that almost every DISTRICT is connected to every other.



* **Splitting Frequent\_item sets length between less than 6 and greater than 6**

1. The above table clearly shows that for the frequent itemsets with length greater than 6 only two combinations have support value greater than 0.8 and rest of the combinations have length less than 6.

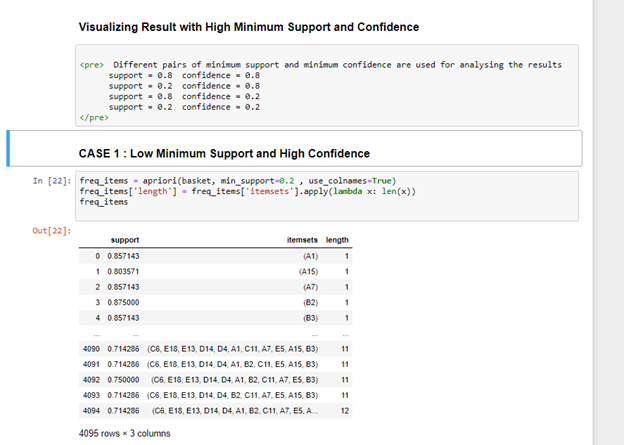




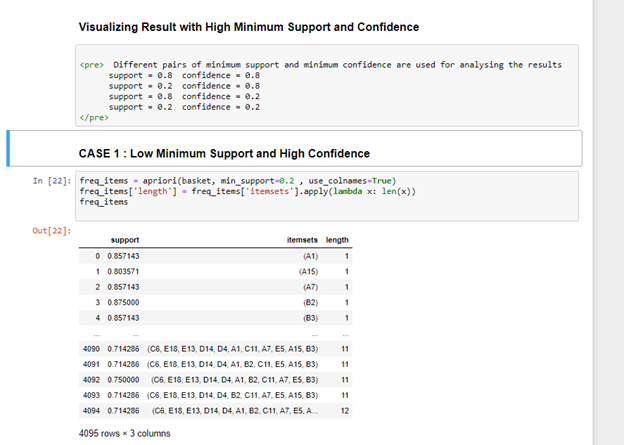
* **Results for different pairs of minimum support and minimum confidence**

For the different pairs the result we obtained was that

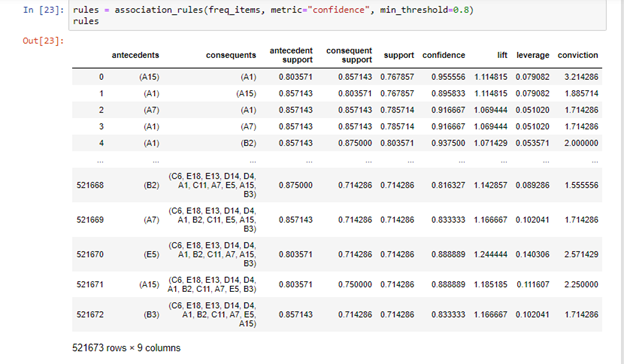
1. The association rule that satisfies both a minimum support and minimum confidence threshold are the strong association rules



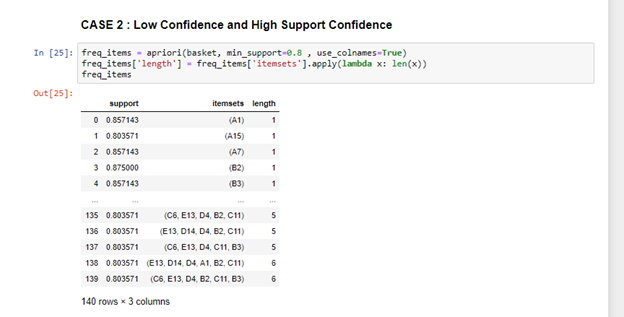
* **Case study 1: Low minimum support and high confidence.**
* Changing the min\_support parameter to 0.2



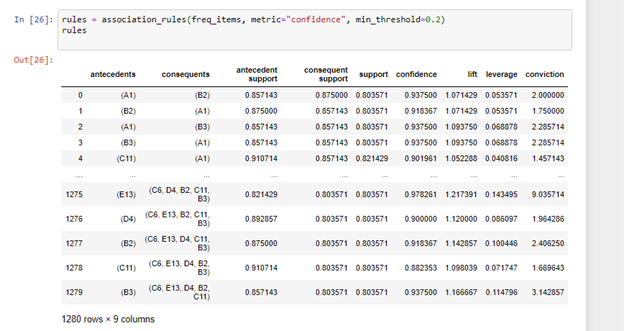
* **Continued Case Study 1 high confidence.**
* Changing the min\_confidence parameter to 0.8



* **Case Study 2: Low confidence and High Support Confidence**
* Min\_support = 0.8



* Confidence min\_threshold = 0.2



* **MAP PLOT**
* **A1 A15 A7 B2**

"Downtown" "Charlestown" "East Boston" "Roxbury"

* **B3 C6 C11 D4**

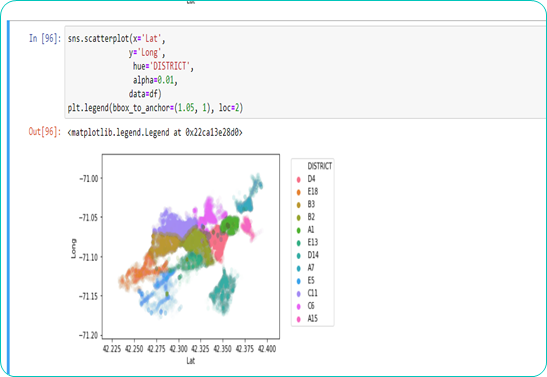
"Mattapan" "South Boston" "Dorchester" "South End"

* **D14 E5 E13 E18**

"Brighton" "West Roxbury" "Jamaica Plain" "Hyde Park“

The plot shows which district is having high crime report

Districts **B2 and C11 .**

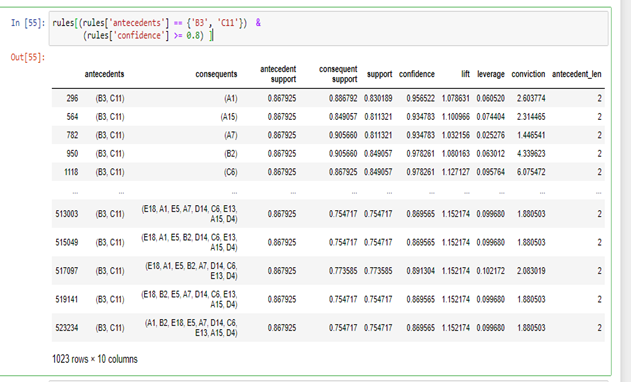
****

* **how will you decide which result sets are meaningful?**
* The above plot and the apriori together proves that the districts B3,C11 have high rate of crime.
* This shows that the result sets are meaningful.
* **ANALYSING FROM THE COMBINATION**

**Antecendents = B3,C11**

**With Confidence >= 0.8**

So with this analysis we can tell that above two mentioned districts have high crime report. Thus the Crime in Boston can be reduced by reducing crime in these two districts.

****

* + With this model we can easily find that districts that have high crime. With further Machine Learning models we can also find what is the reason for the high crime rate in these districts.
  + Our analysis on this dataset can be a huge asset for crime prevention in Boston.
* **Conclusion**

The data set search had been a challenge but eventually we came up with a unique set of criminal investigation data. Inclusion of the network graph was an extra effort that projected the frequency of items in a way more interesting manner.

We have successfully applied the association mining technique using apriori algorithm on to the data and have produced the data sets that follow the necessary rules.