DATA MINING PROJECT:

NYPD Arrests

PREPARED FOR:
Professor Huanying Gu

PREPARED BY:

Neha Bala	1133176
Fernanda Tovar	1090913
Arpit Battu	1129990

TABLE OF CONTENTS

INTRODUCTION

In our project, we will be focusing on a dataset called NYPD arrest. Our goal is to analyze the arrest dataset to determine the correlation between the type of crime, borough, precise location, season, sex and age group of the offenders. The purpose of this is to determine which type of crime is more likely to occur during which time of the year, and which area of a borough.

For this project, we used the software tool called RStudio. Using RStudio, we were able to preprocess the data by cleaning it and clustering. We were also able to analyze the dataset in terms of association rule mining using the apriori algorithm as well as visualize the dataset.

Through this we will be able to determine which areas in NYC are likely to have a particular crime happen. This will allow us to know which location may need to be improved with more security, during certain times of the year, in order to prevent certain crimes from taking place.

DATA SET

https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc

The NYPD Arrest dataset was downloaded from NYC OpenData. This dataset includes all of the arrests that were made by the NYPD in NYC during the 2019 year. Each arrest has information about the type of crime and location of the occurrence. The dataset also included the demographic information of the suspect. This information will tell us the most types of crimes committed, when they are committed and where they were committed.

Furthermore, the dataset contains 215k rows or arrests as well as 18 columns of details of the arrest such as *ARREST_DATE, ARREST_KEY, PD_CD* (classification code) , *ARREST_BORO (Borough) , PD_DESC* (Description via PD code),KY_CD (general classification code), *OFNS_DESC* (Description via KY code), *LAW_CODE, LAW_CAT_CD* (Level of Offense), *AGE_GROUP, PERP_SEX*, *PERP_RACE, ARREST_PRECINET, JURISDICTION_CODE, LATITUDE, LONGITUDE, X_COORD_CD* and *Y_COORD_CD* . In Figure 1, we are able to see the overall breakdown of the data, in the dataset in terms of borough.

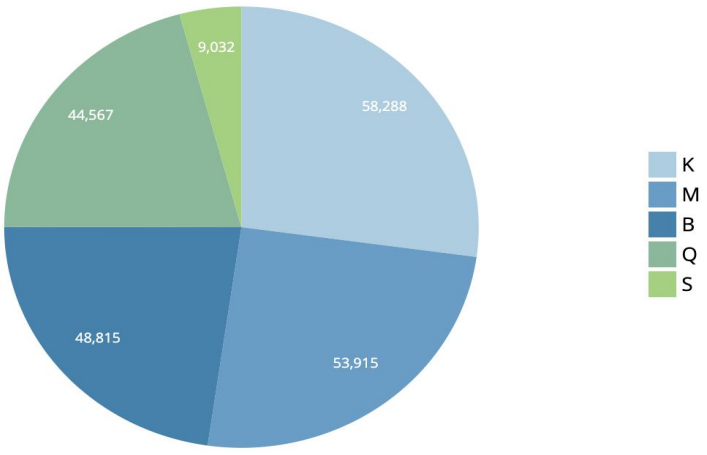


Figure 1: Overall breakdown of arrests in different boroughs in NYC

METHODS

PREPROCESSING

Step 1: Cleaning the Data

In order to transform the data, we first clean up the data to narrow down the attributes or columns. This allows us to remove any attributes that may not be essential to our analysis, such as any codes. This will leave us with attributes that we intend to work with such as ***MONTH, ARREST_BORO, PD_DESC, LATITUDE, LONGITUDE, AGE_GROUP, PERP_SEX, PERP_RACE***.

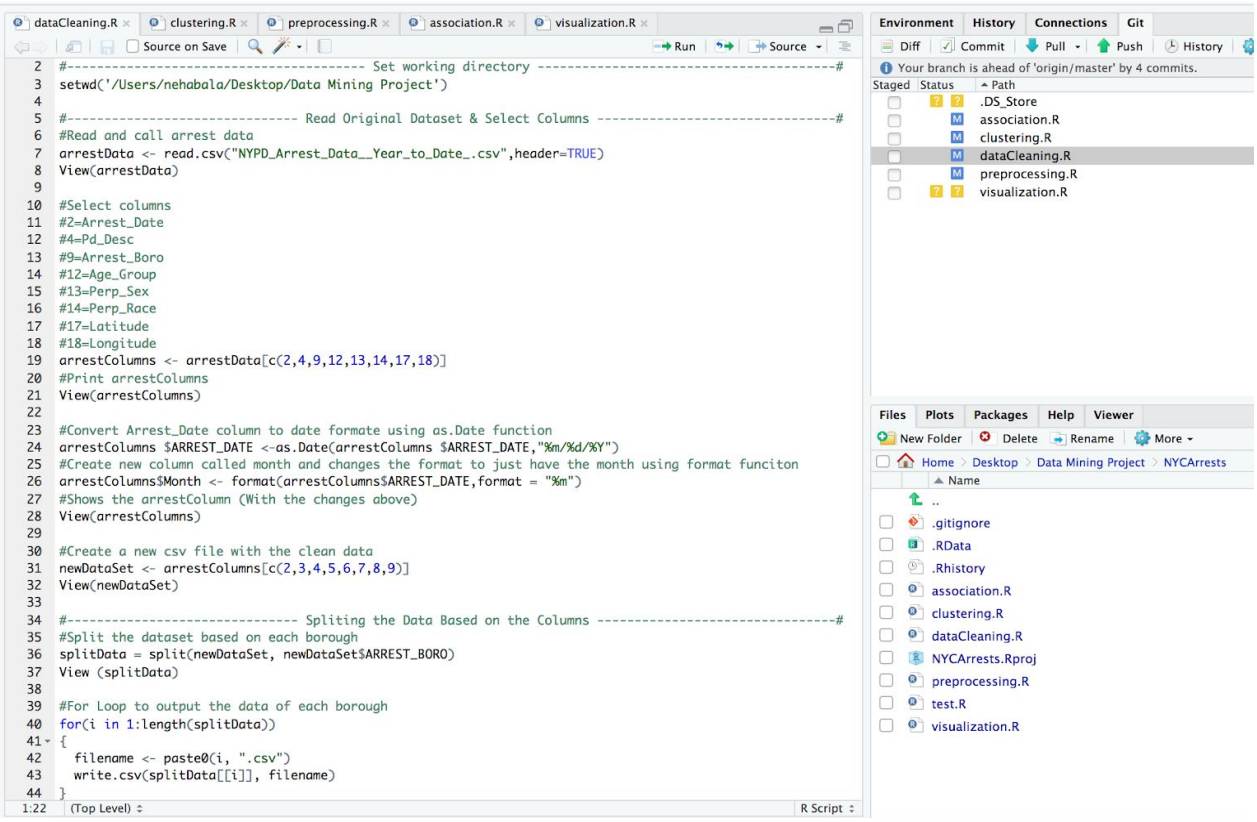


Figure 2: Code for cleaning up the dataset

Our approach for cleaning up the data is demonstrated in Figure 2. We first set the working directory to the appropriate folder where the dataset is located and where we will be outputting our preprocessed file. Then, we read the main dataset file called **NYPD_Arrest_Data_Year_to_Date_.csv** and store

it into a variable called **arrestData.** In Figure 3, we can see the original dataset before the cleaning process.

dataCleaning.R | arrestData | newDataSet | clustering.R | preprocessing.R | association.R | visualization.R

Filter

	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO
1	206892169	12/31/2019	907	IMPAIRED DRIVING,DRUG	347	INTOXICATED & IMPAIRED DRIVING	VTL11920U4	M	M
2	206888084	12/31/2019	739	FRAUD,UNCLASSIFIED-FELONY	112	THEFT-FRAUD	PL 1908301	F	Q
3	206890433	12/31/2019	122	HOMICIDE, NEGLIGENT, VEHICLE,	102	HOMICIDE-NEGLIGENT-VEHICLE	PL 1251201	F	M
4	206890441	12/31/2019	268	CRIMINAL MIS 2 & 3	121	CRIMINAL MISCHIEF & RELATED OF	PL 1450502	F	S
5	206890973	12/31/2019	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	M
6	206891076	12/31/2019	175	SEXUAL ABUSE 3,2	233	SEX CRIMES	PL 13052A1	M	K
7	206891079	12/31/2019	503	CONTROLLED SUBSTANCE,INTENT TO	117	DANGEROUS DRUGS	PL 2201601	F	M
8	206891600	12/31/2019	759	PUBLIC ADMINISTATION,UNCLASS M	359	OFFENSES AGAINST PUBLIC ADMINI	PL 1950001	M	K
9	206882208	12/31/2019	112	MENACING 1ST DEGREE (VICT NOT	126	MISCELLANEOUS PENAL LAW	PL 120141H	F	Q
10	206884271	12/31/2019	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	B
11	206885156	12/31/2019	244	BURGLARY,UNCLASSIFIED,UNKNOWN	107	BURGLARY	PL 1402000	F	Q
12	206887261	12/31/2019	339	LARCENY,PETIT FROM OPEN AREAS,	341	PETIT LARCENY	PL 1552500	M	M
13	206887263	12/31/2019	114	OBSTR BREATH/CIRCUL	344	ASSAULT 3 & RELATED OFFENSES	PL 1211100	M	S
14	206887603	12/31/2019	339	LARCENY,PETIT FROM OPEN AREAS,	341	PETIT LARCENY	PL 1552500	M	K
15	206888897	12/31/2019	339	LARCENY,PETIT FROM OPEN AREAS,	341	PETIT LARCENY	PL 1552500	M	B
16	206889275	12/31/2019	259	CRIMINAL MISCHIEF,UNCLASSIFIED 4	351	CRIMINAL MISCHIEF & RELATED OF	PL 1450001	M	M
17	206888902	12/31/2019	922	TRAFFIC,UNCLASSIFIED MISDEMEAN	348	VEHICLE AND TRAFFIC LAWS	VTL0511001	M	M
18	206889585	12/31/2019	397	ROBBERY,OPEN AREA UNCLASSIFIED	105	ROBBERY	PL 1601503	F	K
19	206890198	12/31/2019	268	CRIMINAL MIS 2 & 3	121	CRIMINAL MISCHIEF & RELATED OF	PL 1450502	F	K
20	206890199	12/31/2019	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F	K
21	206890428	12/31/2019	113	MENACING,UNCLASSIFIED	344	ASSAULT 3 & RELATED OFFENSES	PL 1201500	M	K
22	206890915	12/31/2019	114	OBSTR BREATH/CIRCUL	344	ASSAULT 3 & RELATED OFFENSES	PL 1211100	M	B
23	206890916	12/31/2019	269	MISCHIEF,CRIMINAL, UNCL 2ND	121	CRIMINAL MISCHIEF & RELATED OF	PL 145004A	M	K
24	206890922	12/31/2019	779	PUBLIC ADMINISTRATION,UNCLASSI	126	MISCELLANEOUS PENAL LAW	PL 215510B	F	M
25	206890976	12/31/2019	782	WEAPONS, POSSESSION, ETC	236	DANGEROUS WEAPONS	PL 2650101	M	Q

preprocessing.R | association.R | visualization.R

ARREST_PRECINCT	JURISDICTION_CODE	AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD	Y_COORD_CD	Latitude	Longitude
33	0	25-44	M	WHITE	1000916	245710	40.84108	-73.93977
101	0	25-44	M	BLACK HISPANIC	1055868	156522	40.59601	-73.74212
23	0	18-24	M	BLACK	997462	227024	40.78980	-73.95229
120	0	18-24	M	BLACK	962822	174282	40.64502	-74.07722
25	97	25-44	M	BLACK	1004138	226326	40.78787	-73.92818
60	0	18-24	M	WHITE HISPANIC	990922	149184	40.57616	-73.97598
28	0	25-44	M	BLACK	997020	230319	40.79885	-73.95388
61	0	18-24	M	WHITE	999084	153345	40.58757	-73.94659
106	0	45-64	M	UNKNOWN	1028843	188006	40.68261	-73.83922
40	0	18-24	M	BLACK	1005041	234533	40.81040	-73.92490
109	0	25-44	M	BLACK	1032084	216954	40.76204	-73.82733
25	0	45-64	M	WHITE HISPANIC	1002996	229133	40.79558	-73.93230
120	0	18-24	F	WHITE HISPANIC	960040	174186	40.64475	-74.08724
75	0	45-64	F	BLACK	1020754	176733	40.65170	-73.86845
40	0	45-64	M	WHITE HISPANIC	1009673	235504	40.81305	-73.90816
5	0	45-64	M	ASIAN / PACIFIC ISLANDER	986531	197880	40.70982	-73.99177
34	0	25-44	M	BLACK	1007190	253954	40.86370	-73.91707
75	0	25-44	M	BLACK HISPANIC	1016920	188994	40.68537	-73.88220
63	0	45-64	F	WHITE	1009348	167036	40.62512	-73.90959
75	0	45-64	M	BLACK	1016798	184321	40.67254	-73.88267
67	0	25-44	M	BLACK	1004990	173542	40.64299	-73.92527
47	0	25-44	M	WHITE HISPANIC	1023532	257388	40.87307	-73.85796
70	0	25-44	M	BLACK	991591	169049	40.63068	-73.97355
25	0	45-64	M	BLACK	999997	230635	40.79971	-73.94313
103	0	25-44	M	BLACK	1042592	197006	40.70723	-73.78957

Figure 3: Original Dataset File

From the main dataset, we determine which columns will be selected to be stored in the **arrestColumns** variable. Furthermore, from the **arrestColumns,** we use the **Arrest_Date** column to format it to create a new column called **Month**. Using the month we will be able to determine which season or time

of the year a particular arrest is likely to take place. After this, we create a new variable called **newDataSet**, where we select all columns, except for date, from **arrestColumns**. This is because we will be working with only the month column not the date column. In Figure 3, we can see the final output of the cleaned data.

dataCleaning.R | newDataSet | clustering.R | preprocessing.R | association.R | visualization.R

Filter

	PD_DESC	ARREST_BORO	AGE_GROUP	PERP_SEX	PERP_RACE	Latitude	Longitude	Month
1	IMPAIRED DRIVING,DRUG	M	25-44	M	WHITE	40.84108	-73.93977	12
2	FRAUD,UNCLASSIFIED-FELONY	Q	25-44	M	BLACK HISPANIC	40.59601	-73.74212	12
3	HOMICIDE, NEGLIGENT, VEHICLE,	M	18-24	M	BLACK	40.78980	-73.95229	12
4	CRIMINAL MIS 2 & 3	S	18-24	M	BLACK	40.64502	-74.07722	12
5	ASSAULT 3	M	25-44	M	BLACK	40.78787	-73.92818	12
6	SEXUAL ABUSE 3,2	K	18-24	M	WHITE HISPANIC	40.57616	-73.97598	12
7	CONTROLLED SUBSTANCE,INTENT TO	M	25-44	M	BLACK	40.79885	-73.95388	12
8	PUBLIC ADMINISTATION,UNCLASS M	K	18-24	M	WHITE	40.58757	-73.94659	12
9	MENACING 1ST DEGREE (VICT NOT	Q	45-64	M	UNKNOWN	40.68261	-73.83922	12
10	ASSAULT 3	B	18-24	M	BLACK	40.81040	-73.92490	12
11	BURGLARY,UNCLASSIFIED,UNKNOWN	Q	25-44	M	BLACK	40.76204	-73.82733	12
12	LARCENY,PETIT FROM OPEN AREAS,	M	45-64	M	WHITE HISPANIC	40.79558	-73.93230	12
13	OBSTR BREATH/CIRCUL	S	18-24	F	WHITE HISPANIC	40.64475	-74.08724	12
14	LARCENY,PETIT FROM OPEN AREAS,	K	45-64	F	BLACK	40.65170	-73.86845	12
15	LARCENY,PETIT FROM OPEN AREAS,	B	45-64	M	WHITE HISPANIC	40.81305	-73.90816	12
16	CRIMINAL MISCHIEF,UNCLASSIFIED 4	M	45-64	M	ASIAN / PACIFIC ISLANDER	40.70982	-73.99177	12
17	TRAFFIC,UNCLASSIFIED MISDEMEAN	M	25-44	M	BLACK	40.86370	-73.91707	12
18	ROBBERY,OPEN AREA UNCLASSIFIED	K	25-44	M	BLACK HISPANIC	40.68537	-73.88220	12
19	CRIMINAL MIS 2 & 3	K	45-64	F	WHITE	40.62512	-73.90959	12
20	STRANGULATION 1ST	K	45-64	M	BLACK	40.67254	-73.88267	12
21	MENACING,UNCLASSIFIED	K	25-44	M	BLACK	40.64299	-73.92527	12
22	OBSTR BREATH/CIRCUL	B	25-44	M	WHITE HISPANIC	40.87307	-73.85796	12
23	MISCHIEF,CRIMINAL, UNCL 2ND	K	25-44	M	BLACK	40.63068	-73.97355	12
24	PUBLIC ADMINISTRATION,UNCLASSI	M	45-64	M	BLACK	40.79971	-73.94313	12
25	WEAPONS, POSSESSION, ETC	Q	25-44	M	BLACK	40.70723	-73.78957	12

Showing 1 to 26 of 214,617 entries, 8 total columns

Figure 4: Cleaned Data File

After transforming the data, we further clean the data by splitting the data into five files based on the column borough, one for each borough.

Step 2: Clustering

For the clustering aspect of preprocessing, we used the five files we split during data cleaning. For each file, we created four clusters of the latitude and longitude for each borough. These clusters represent the different sections of each borough in terms of north, east, south, and west.

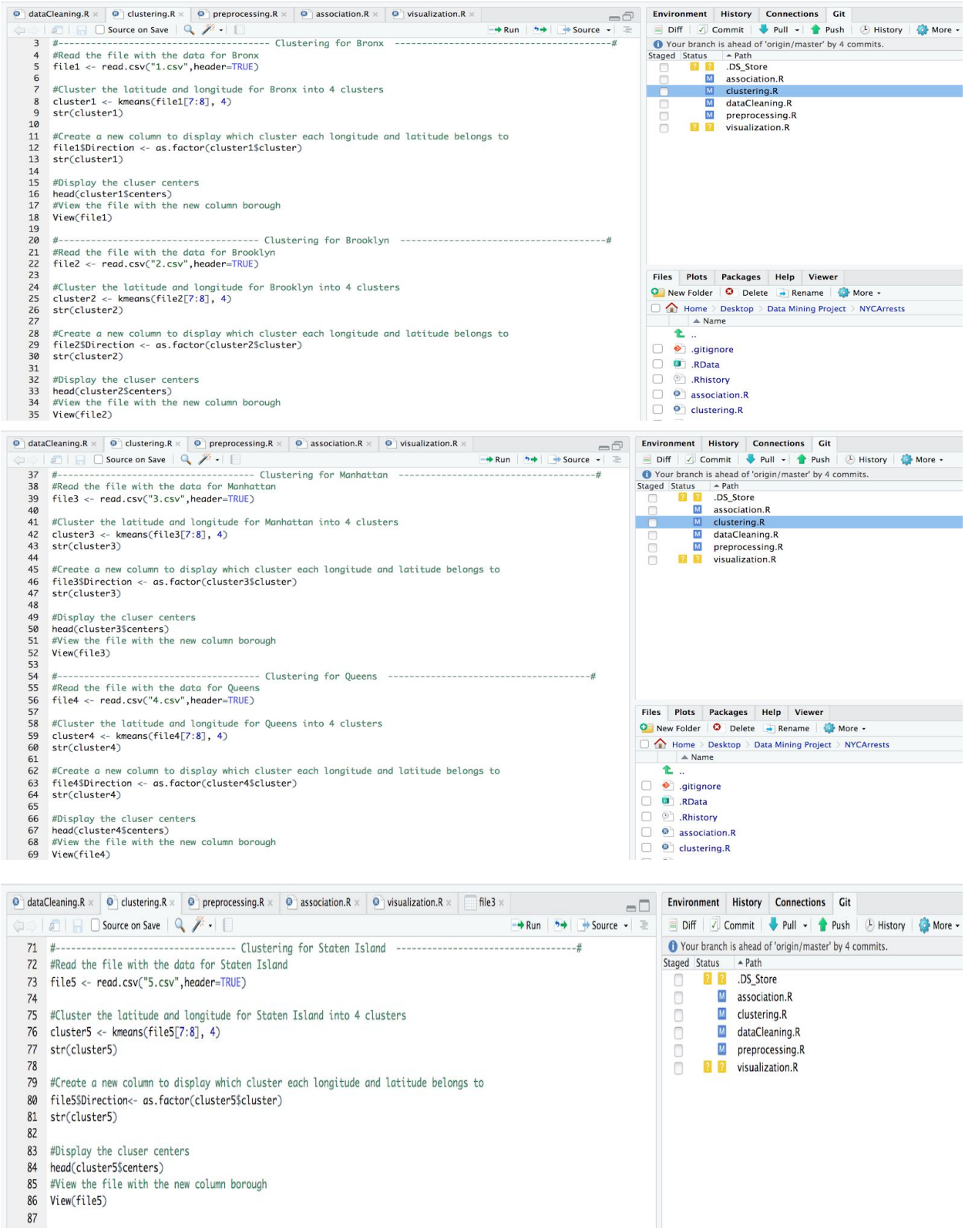


Figure 5: Clustering for each Borough

We first read the csv file of a borough and store the data into a variable. For example, for the Bronx, we are reading the file **1.csv** and storing the data in the variable **file1**. For Brooklyn, we are reading the file **2.csv** and storing the data in the variable **file2**. This applies for all boroughs, as shown in Figure 5.

Furthermore, we use the K-Means algorithm to select **columns [7:8],** which represent the latitude and longitude, to create four clusters for each borough. We store the clusters into variables. For example, for Queens, we store its clusters into **cluster4** and for Staten Island, we store its clusters into **cluster5**. Next, for each borough, we create a new column **Direction** to display which cluster each latitude and longitude belongs to. The **Direction** column will allow us to determine the north, east, south, west sections of each borough.. We can display the cluster centers to see which range of latitude and longitude got clustered and view it for each borough file. In Figure 7, we see an example of this in the Manhattan file.

dataCleaning.R | clustering.R | file3 | preprocessing.R | association.R | visualization.R

Filter

	X	PD_DESC	ARREST_BORO	AGE_GROUP	PERP_SEX	PERP_RACE	Latitude	Longitude	Month	Direction
1	1	IMPAIRED DRIVING,DRUG	M	25-44	M	WHITE	40.84108	-73.93977	12	3
2	3	HOMICIDE, NEGLIGENT, VEHICLE,	M	18-24	M	BLACK	40.78980	-73.95229	12	2
3	5	ASSAULT 3	M	25-44	M	BLACK	40.78787	-73.92818	12	2
4	7	CONTROLLED SUBSTANCE,INTENT TO	M	25-44	M	BLACK	40.79885	-73.95388	12	2
5	12	LARCENY,PETIT FROM OPEN AREAS,	M	45-64	M	WHITE HISPANIC	40.79558	-73.93230	12	2
6	16	CRIMINAL MISCHIEF,UNCLASSIFIED 4	M	45-64	M	ASIAN / PACIFIC ISLANDER	40.70982	-73.99177	12	1
7	17	TRAFFIC,UNCLASSIFIED MISDEMEAN	M	25-44	M	BLACK	40.86370	-73.91707	12	3
8	24	PUBLIC ADMINISTRATION,UNCLASSI	M	45-64	M	BLACK	40.79971	-73.94313	12	2
9	28	TRAFFIC,UNCLASSIFIED MISDEMEAN	M	45-64	M	BLACK	40.75952	-73.99926	12	4
10	30	LARCENY,PETIT FROM OPEN AREAS,	M	25-44	M	WHITE HISPANIC	40.75043	-73.98928	12	4
11	33	CONTEMPT,CRIMINAL	M	25-44	M	BLACK	40.74464	-74.00454	12	1
12	40	ASSAULT 2,1,UNCLASSIFIED	M	45-64	F	WHITE	40.78935	-73.94735	12	2
13	43	ASSAULT 3	M	25-44	M	BLACK	40.73331	-73.99332	12	1
14	45	LARCENY,GRAND FROM OPEN AREAS, UNATTENDED	M	25-44	M	ASIAN / PACIFIC ISLANDER	40.75043	-73.98928	12	4
15	46	LARCENY,PETIT FROM OPEN AREAS,	M	45-64	M	WHITE HISPANIC	40.74134	-73.97839	12	4
16	47	MENACING,UNCLASSIFIED	M	25-44	M	BLACK	40.82565	-73.94359	12	3
17	50	TRAFFIC,UNCLASSIFIED INFRACTIO	M	25-44	M	BLACK	40.73542	-73.97514	12	1
18	52	SEXUAL ABUSE 3,2	M	25-44	M	BLACK	40.80069	-73.94111	12	2
19	55	ASSAULT 2,1,UNCLASSIFIED	M	45-64	F	WHITE HISPANIC	40.77304	-73.98220	12	4
20	62	MURDER,UNCLASSIFIED	M	25-44	M	BLACK	40.85111	-73.93526	12	3
21	69	ROBBERY,OPEN AREA UNCLASSIFIED	M	25-44	M	BLACK HISPANIC	40.71596	-73.99735	12	1

Figure 7: Direction Column for Arrest Borough Manhattan

After we cluster the data for each file, we bind all the files for every borough into one variable called **clustering**. As shown in Figure 8, we first called the **library(dplyr)**, this allows us to use the function **bind_rows**. Furthermore, we select the columns from the data in the variable **clustering**. We do this because we are using the clusters to determine the location, therefore, we remove the columns longitude and latitude. We store the new data into the variable **clusteredData** and write data into a csv file called **clustering.csv**. In Figure 9, we can see the output for the combined data with their clusters indicated.

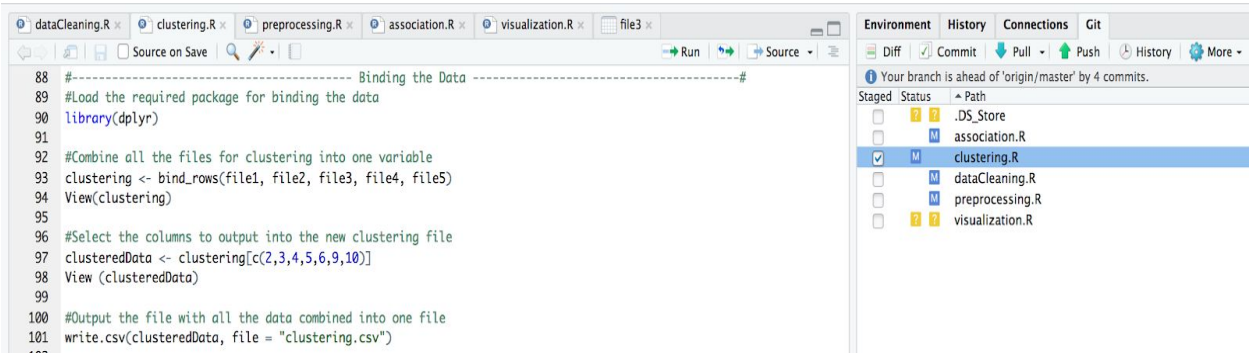
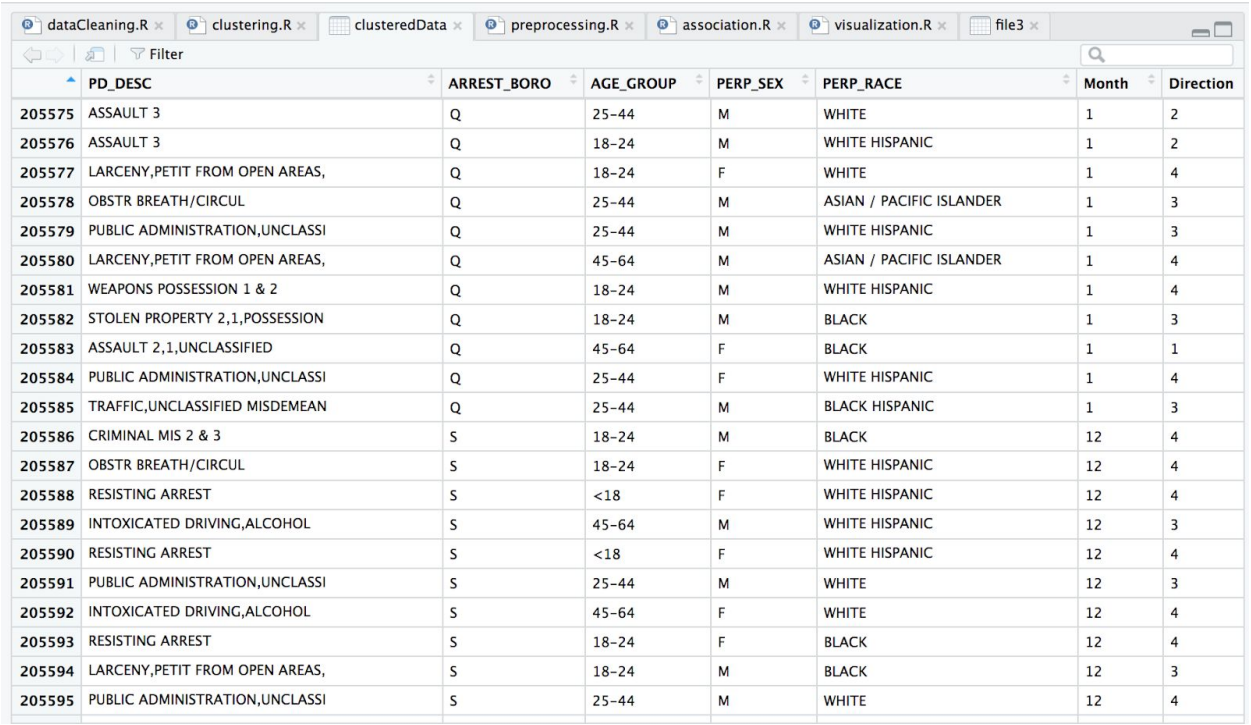


Figure 8: Binding the data



	PD_DESC	ARREST_BORO	AGE_GROUP	PERP_SEX	PERP_RACE	Month	Direction
205575	ASSAULT 3	Q	25-44	M	WHITE	1	2
205576	ASSAULT 3	Q	18-24	M	WHITE HISPANIC	1	2
205577	LARCENY,PETIT FROM OPEN AREAS,	Q	18-24	F	WHITE	1	4
205578	OBSTR BREATH/CIRCUL	Q	25-44	M	ASIAN / PACIFIC ISLANDER	1	3
205579	PUBLIC ADMINISTRATION,UNCLASSI	Q	25-44	M	WHITE HISPANIC	1	3
205580	LARCENY,PETIT FROM OPEN AREAS,	Q	45-64	M	ASIAN / PACIFIC ISLANDER	1	4
205581	WEAPONS POSSESSION 1 & 2	Q	18-24	M	WHITE HISPANIC	1	4
205582	STOLEN PROPERTY 2,1,POSSESSION	Q	18-24	M	BLACK	1	3
205583	ASSAULT 2,1,UNCLASSIFIED	Q	45-64	F	BLACK	1	1
205584	PUBLIC ADMINISTRATION,UNCLASSI	Q	25-44	F	WHITE HISPANIC	1	4
205585	TRAFFIC,UNCLASSIFIED MISDEMEAN	Q	25-44	M	BLACK HISPANIC	1	3
205586	CRIMINAL MIS 2 & 3	S	18-24	M	BLACK	12	4
205587	OBSTR BREATH/CIRCUL	S	18-24	F	WHITE HISPANIC	12	4
205588	RESISTING ARREST	S	<18	F	WHITE HISPANIC	12	4
205589	INTOXICATED DRIVING,ALCOHOL	S	45-64	M	WHITE HISPANIC	12	3
205590	RESISTING ARREST	S	<18	F	WHITE HISPANIC	12	4
205591	PUBLIC ADMINISTRATION,UNCLASSI	S	25-44	M	WHITE	12	3
205592	INTOXICATED DRIVING,ALCOHOL	S	45-64	F	WHITE	12	4
205593	RESISTING ARREST	S	18-24	F	BLACK	12	4
205594	LARCENY,PETIT FROM OPEN AREAS,	S	18-24	M	BLACK	12	3
205595	PUBLIC ADMINISTRATION,UNCLASSI	S	25-44	M	WHITE	12	4

Figure 9: Combined file with all the boroughs & their clusters

Step 3: Converting Data Types

Our final step in preprocessing the data is to convert the data types to **Factor**. This will prepare the data for the apriori algorithm for association rule mining. In order to convert the data types, we first read the data **clustering.csv** file into the variable **clusterFile,** which contains the combined data for each borough.

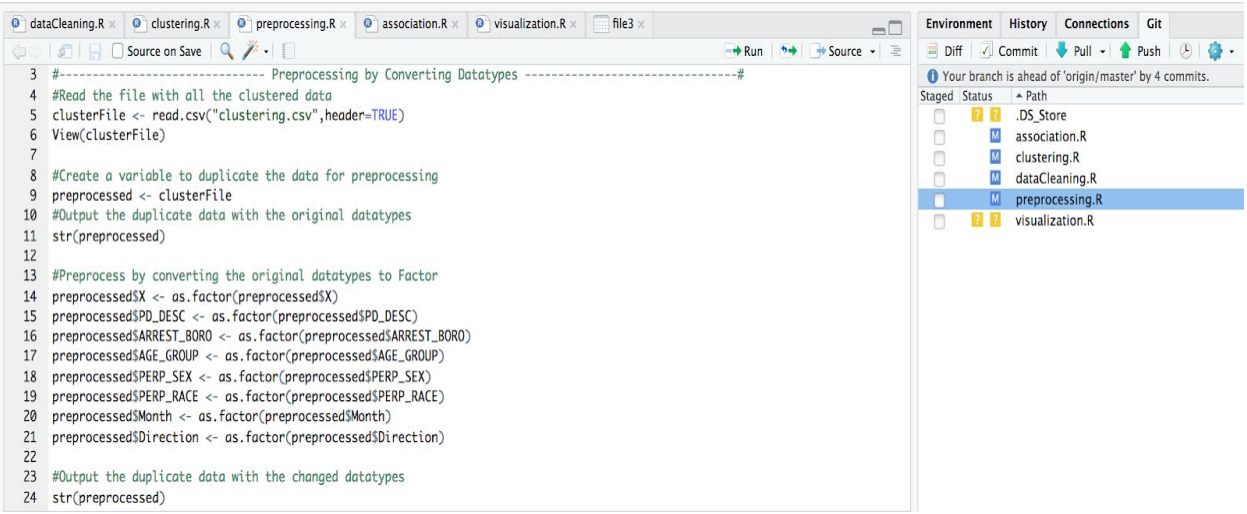


Figure 10: Code for converting data types

As shown in Figure 10, we create a variable called **preprocessed**. This variable will duplicate the data in **clusterFile**, in order to prevent any changes on the original file. Furthermore, we use the variable **preprocessed** to select each column and convert it into the data type **Factor**. This is done by the following: preprocessed$COLUMN_NAME = as.factor (preprocessed$COLUMN_NAME). In Figure 11, we can see the before and after converting the data types of each column.

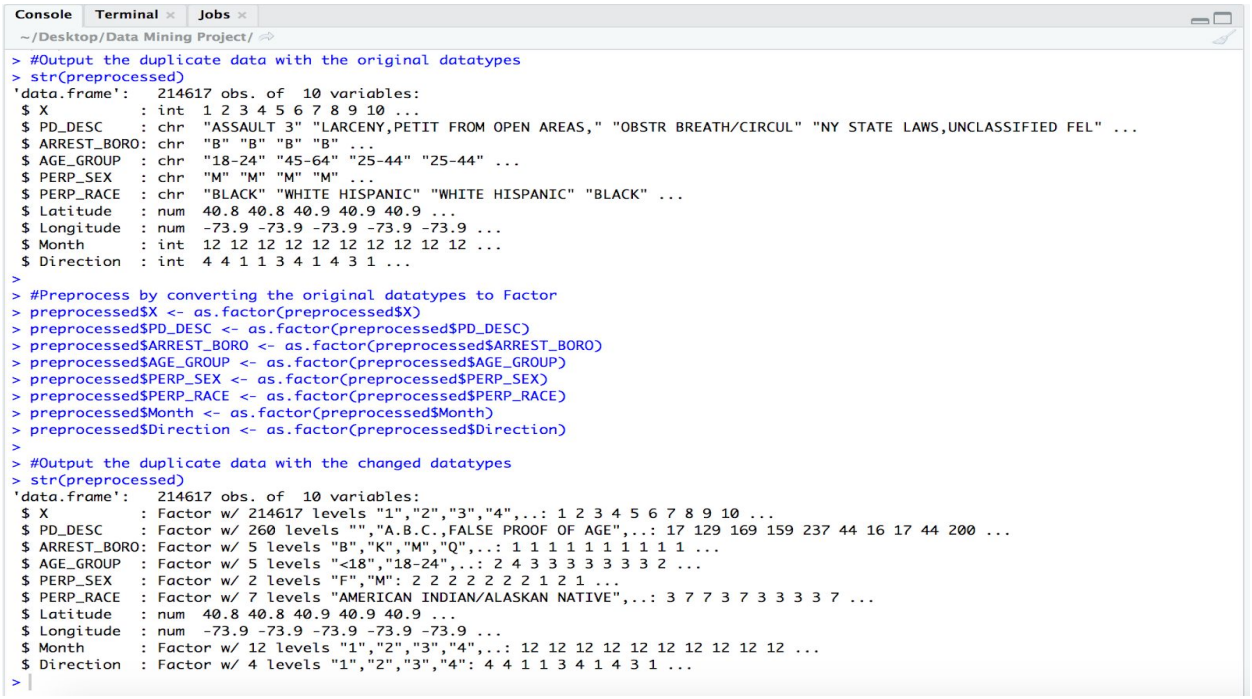


Figure 11: Before and after converting the data types

ASSOCIATION RULE MINING

In order to determine the correlation between the type of arrest, borough, location, month, age group, sex, and race we use association rule mining in particular we use the apriori algorithm. This will allow us to obtain results for analysis to determine which areas in NYC may need to be improved, which months are more likely to have certain crimes, and which sex, race, and age group are likely to commit a certain crime and which location in each borough are likely to face certain crimes.

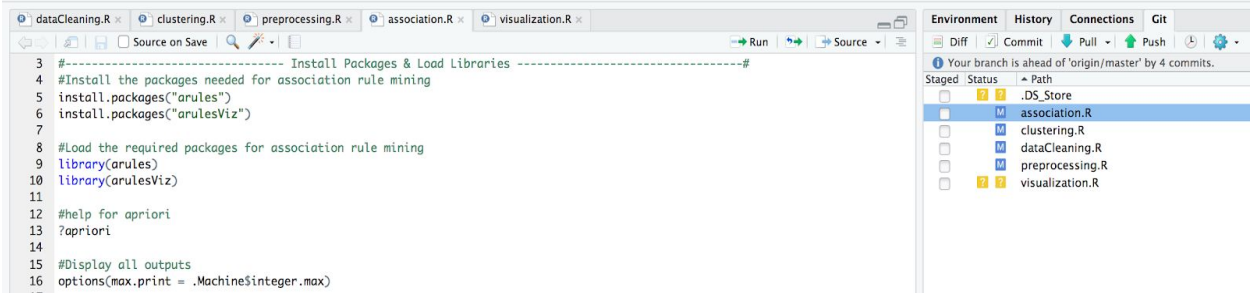


Figure 12: Installing and Loading Libraries

In order to run the apriori algorithm, we first install two packages called **arules** and **arulesViz**. Next, we use the libraries containing the two packages, as shown in Figure 12. Furthermore, in order to run the algorithm, we first use the **options** function in order to make sure that all of the rules are outputted. Next, we set three different support and confidence to run the apriori algorithm, as shown in Figure 13.

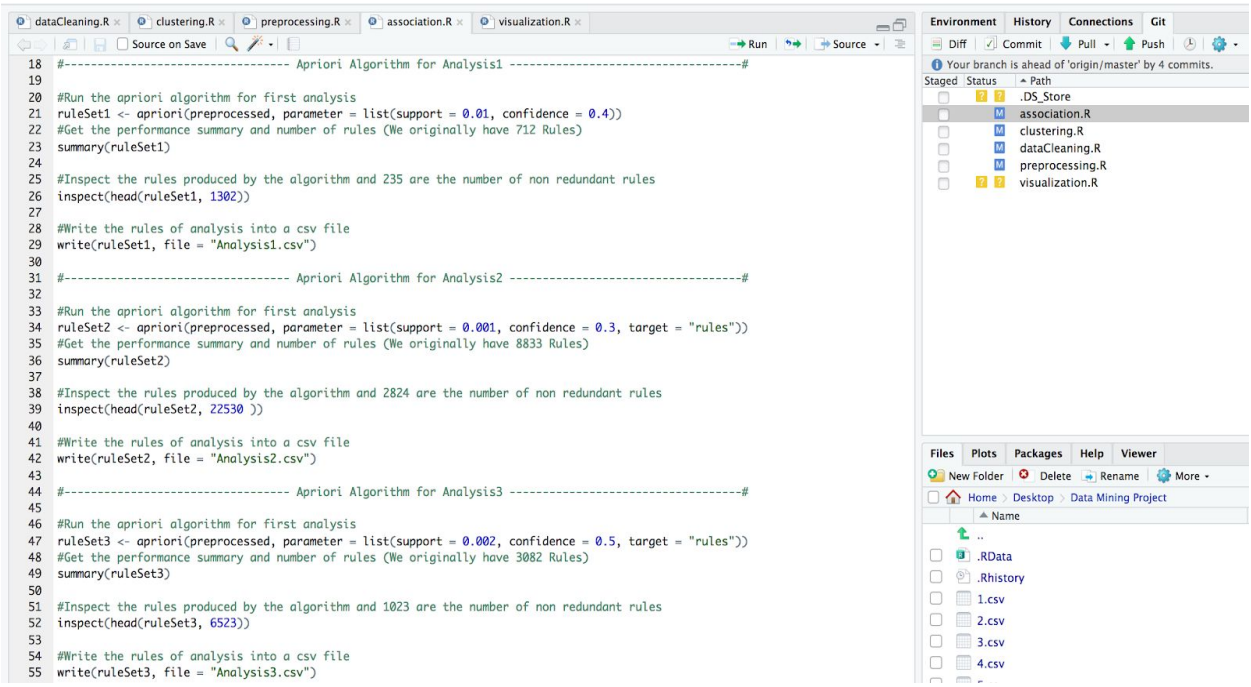
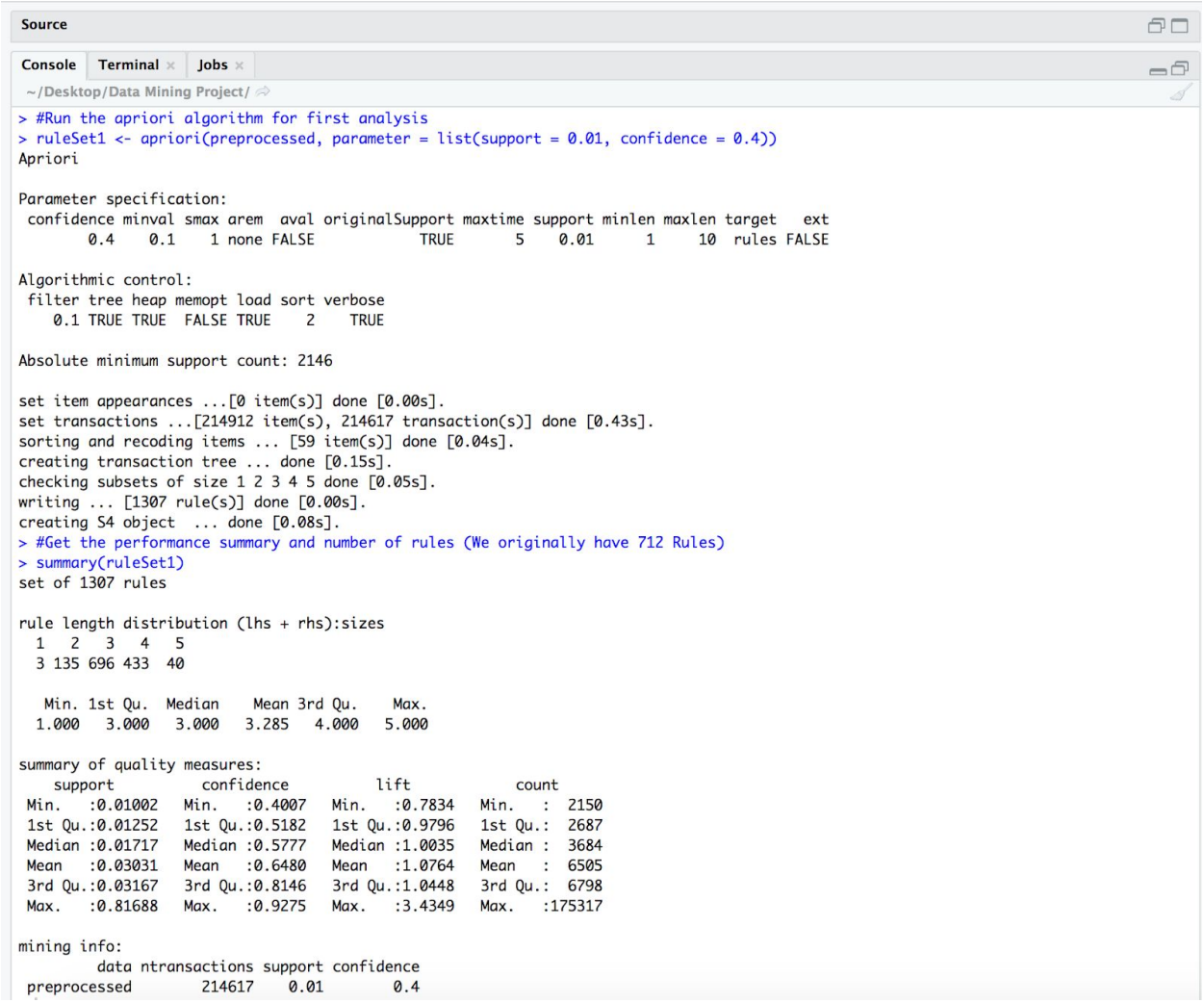


Figure 13: Code for running Apriori Algorithm

In order to run the apriori algorithm, we use the **apriori** function. We indicate the file **preprocessed** to run the algorithm on it. For the first analysis, we set the support to 0.01 and the confidence to 0.4. The conditions for the apriori algorithm, for the first analysis, is stored in the variable **ruleSet1**. This will find the association rules between all the attributes we are working with. In Figure 14, we run the algorithm and use the **summary** function to obtain the summary of the rules generated for **ruleSet1**.



```
Source
Console  Terminal  Jobs
~/Desktop/Data Mining Project/
> #Run the apriori algorithm for first analysis
> ruleSet1 <- apriori(preprocessed, parameter = list(support = 0.01, confidence = 0.4))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.4    0.1    1 none FALSE            TRUE       5    0.01      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2146

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[214912 item(s), 214617 transaction(s)] done [0.43s].
sorting and recoding items ... [59 item(s)] done [0.04s].
creating transaction tree ... done [0.15s].
checking subsets of size 1 2 3 4 5 done [0.05s].
writing ... [1307 rule(s)] done [0.00s].
creating S4 object  ... done [0.08s].
> #Get the performance summary and number of rules (We originally have 712 Rules)
> summary(ruleSet1)
set of 1307 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4   5
  3 135 696 433  40

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.285   4.000   5.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.01002   Min.   :0.4007   Min.   :0.7834   Min.   :  2150
 1st Qu.:0.01252   1st Qu.:0.5182   1st Qu.:0.9796   1st Qu.:  2687
 Median :0.01717   Median :0.5777   Median :1.0035   Median :  3684
 Mean   :0.03031   Mean   :0.6480   Mean   :1.0764   Mean   :  6505
 3rd Qu.:0.03167   3rd Qu.:0.8146   3rd Qu.:1.0448   3rd Qu.:  6798
 Max.   :0.81688   Max.   :0.9275   Max.   :3.4349   Max.   :175317

mining info:
         data ntransactions support confidence
 preprocessed        214617    0.01        0.4
```

Figure 14: Running Apriori Algorithm and Getting Summary of rules

Finally, we use the **inspect** function to inspect the rules by indicating how many rules we would like to view or analyze based on the output of the algorithm. For each performance analysis, we output the inspected rules into three different csv files. In Figure 15, we can see the rules generated by the apriori algorithm for analysis 1.

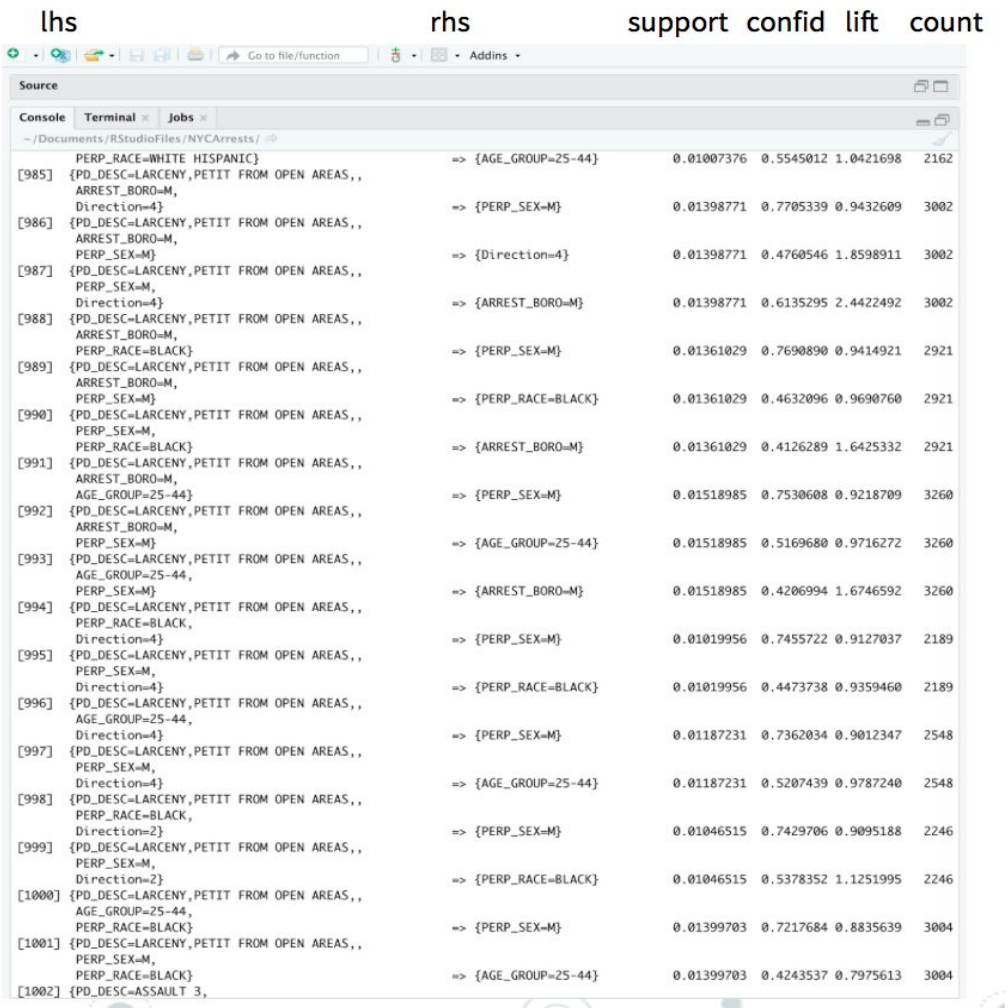


Figure 15: Output of the rules generated by the apriori algorithm for analysis 1

RESULTS

APRIORI ALGORITHM RESULTS

The following are the results of association rule mining obtained by the apriori algorithm. Each analysis contains 6 rules that were generated. The support indicates the probability that an event contains X U Y. The confidence indicates the conditional probability that an event having X also contains Y. The lift indicates the simple correlation between itemset X and itemset Y. Note: The direction column analysis was done with the visualizations done for clustering. Therefore, that's how we can conclude what area of the borough the rule specifically focuses on.

Result 1

Support = 0.01, Confidence = 0.4

ANALYSIS 1

Support: 0.01, Confidence: 0.4, Rules Generated: 1302

X	Y	SUPPORT	CONFIDENCE	LIFT	COUNT
{PD_DESC=STRANGULATION 1ST}	{PERP_SEX=M}	0.01094508	0.9266272	1.1343450	2349
{ARREST_BORO=K, AGE_GROUP=25-44, Month=1}	{PERP_SEX=M}	0.01133181	0.8300341	1.0160990	2432
{PERP_RACE=ASIAN / PACIFIC ISLANDER, Direction=1}	{ARREST_BORO=Q}	0.01508734	0.6972438	3.3576495	3238
{PERP_SEX=M, PERP_RACE=WHITE, Direction=2}	{ARREST_BORO=K}	0.01524110	0.5987553	2.2046229	3271
{PD_DESC=LARCENY,PETIT FROM OPEN AREAS,, PERP_SEX=M, Direction=1}	{ARREST_BORO=M}	0.01398771	0.5331202	2.1221676	3002
{PERP_SEX=M, PERP_RACE=WHITE, Direction=1}	{ARREST_BORO=M}	0.01313503	0.4906876	1.9532577	2819

In analysis 1, we will look at rule #5 as an example. The algorithm generated the arrest of **larceny petit from open areas**, the sex **male**, and **direction 1** being associated with the borough **Queens**. In this case, **direction 1** represents Rockaway Park, Queens.

Result 2

Support = 0.001, Confidence = 0.3

ANALYSIS 2

Support: 0.001, Confidence: 0.3, Rules Generated: 22816

X	Y	SUPPORT	CONFIDENCE	LIFT	COUNT
{PD_DESC=PROSTITUTION}	{PERP_SEX=F}	0.001705364	0.9812332	5.3585073	366
{ARREST_BORO=M, PERP_RACE=WHITE HISPANIC, Month=5,Direction=4}	{PERP_SEX=M}	0.0010623	0.820143	1.003991	228
{PD_DESC=BAIL JUMPING 3}	{ARREST_BORO=Q}	0.001141568	0.7632399	3.6754606	245
{PD_DESC=ASSAULT 3, ARREST_BORO=Q, PERP_SEX=M, Direction=1}	{AGE_GROUP=25-44}	0.0028655	0.527896	0.992168	615
{PD_DESC=AGGRAVATED HARASSMENT 2, Direction=2}	{ARREST_BORO=M}	0.0013885	0.41971830	1.6707536	298
{PD_DESC=ASSAULT 3, PERP_SEX=M, PERP_RACE=BLACK, Month=1}	{ARREST_BORO=K}	0.001029741	0.3226277	1.1879186	221

In analysis 2, we will look at rule #5 as an example. The algorithm generated the arrest of **assault 3**, the sex **male**, the race as **black**, and **month 1** being associated with the borough **Brooklyn**. In this case **month 1** represents January or the time around winter when this crime is likely to take place.

Result 3

Support = 0.002, Confidence = 0.5

ANALYSIS 3

Support: 0.002, Confidence: 0.5, NumOfRules: 6635

X	Y	SUPPORT	CONFIDENCE	LIFT	COUNT
{PD_DESC=DISORDERLY CONDUCT}	{ARREST_BORO=M}	0.002818975	0.8497191	3.382438	605
{ARREST_BORO=Q, AGE_GROUP=25-44, Month=2}	{PERP_SEX=M}	0.007166254	0.8167817	0.9998759	1538
{ARREST_BORO=K, Month=6, Direction=3}	{PERP_SEX=M}	0.0029960	0.7928483	0.970577	643
{ARREST_BORO=M, AGE_GROUP=25-44, Month=12, Direction=2}	{PERP_SEX=M}	0.0020594	0.77003	0.94264	442
{PD_DESC=ASSAULT 3, ARREST_BORO=Q, AGE_GROUP=25-44, PERP_SEX=M, PERP_RACE=WHITE HISPANIC}	{Direction=3}	0.002390305	0.6279070	2.6145066	513
{PD_DESC=FORGERY,ETC.,UNCLASSIFIED-FELO, PERP_SEX=M, PERP_RACE=BLACK, Direction=3}	{ARREST_BORO=K}	0.00206880	0.620979	2.28645	444

In analysis 3, we will look at rule #4 as an example. The algorithm generated the arrest borough as **Manhattan**, age group of **25-44**, the **month 12**, and **direction 2** being associated with the sex as **male**. In this case **direction 2** represents Lower Manhattan and **month 12** represents the month of December or season of winter, where crimes are likely to happen.

VISUALIZATION & ANALYSIS

CLUSTERED MAPS

In order to do visualization of the clustered data for each borough, we needed to install and load all the necessary packages as well as obtain our Google Maps API key, as shown in Figure 16. This allowed us to display all the clusters for each borough. In Figure 17, we are able to see use the function **get_map** to get the map for each borough and indicate the zoom. We use the function **ggmap** and **geom_point** to obtain and display the graph.

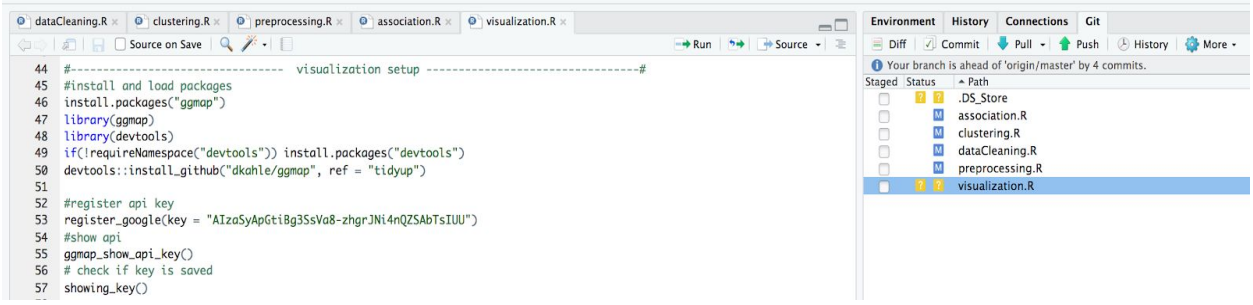


Figure 16: Load packages and test our API key to display location graphs

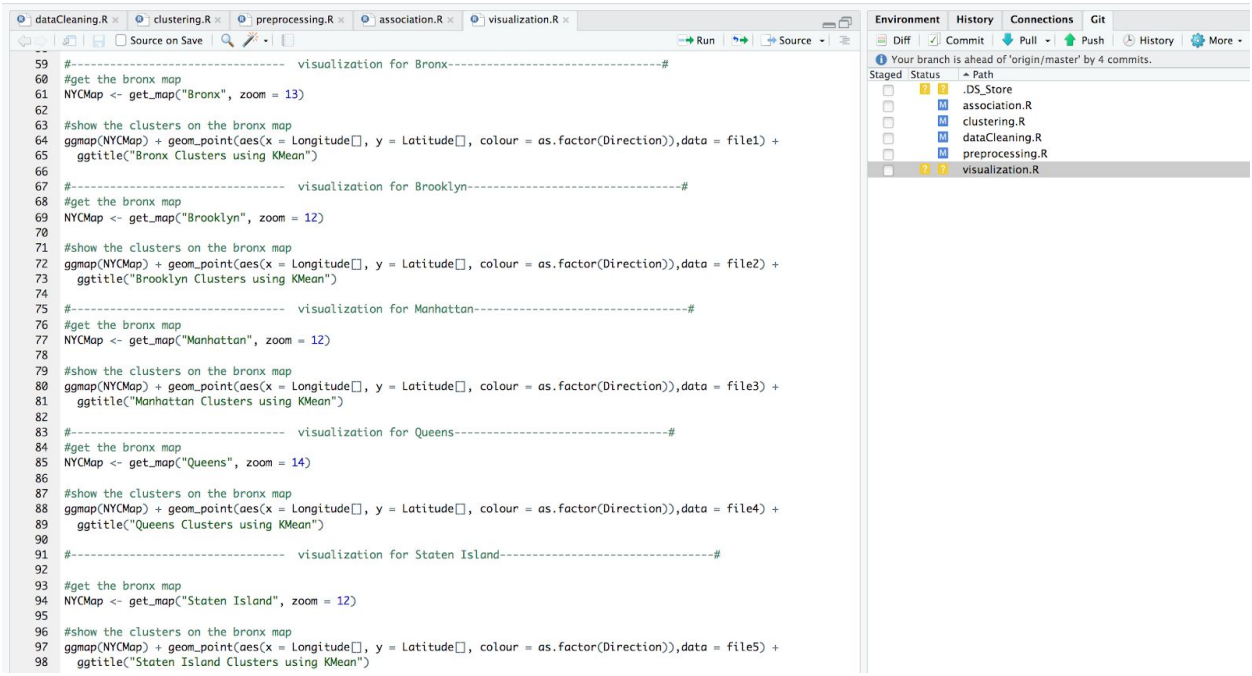


Figure 17: Plot the graphs for each Borough in NYC using the API key

The following graphs display the clusters for each borough as well as which area each of the four clusters represent.

1 =Northwest Bronx , 2 =Southwest Bronx , 3 =Northeast Bronx , 4 = Southeast Bronx

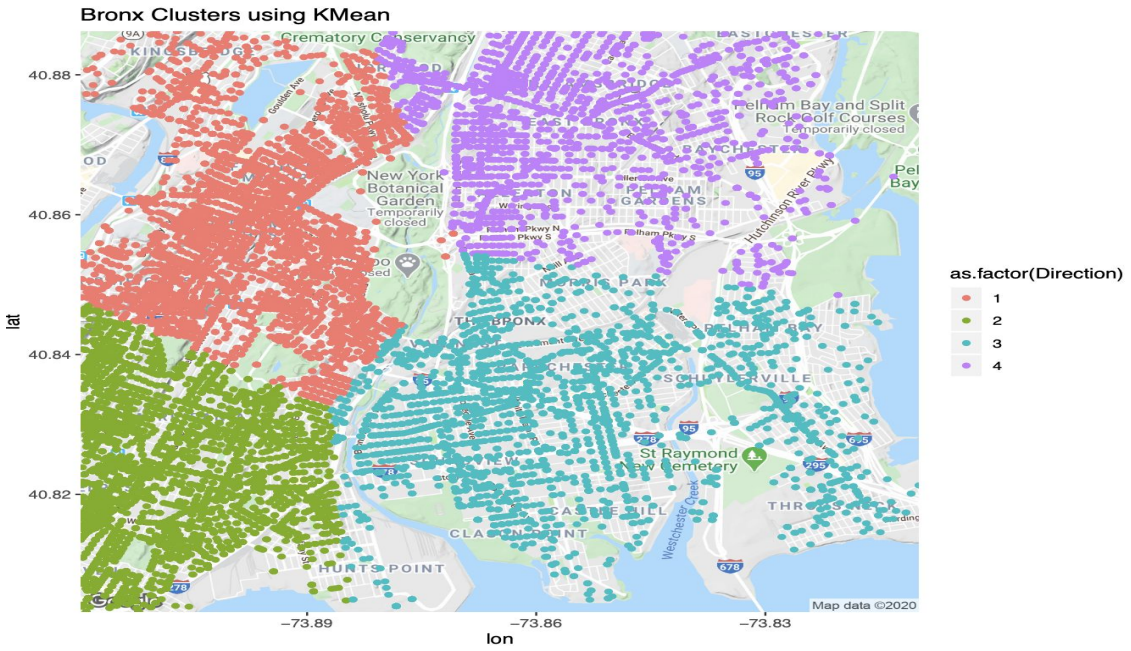


Figure 17: Bronx Clusters

1 = Southeast Brooklyn, 2 =Northeast Brooklyn , 3 =Southwest Brooklyn , 4 =Northwest Brooklyn

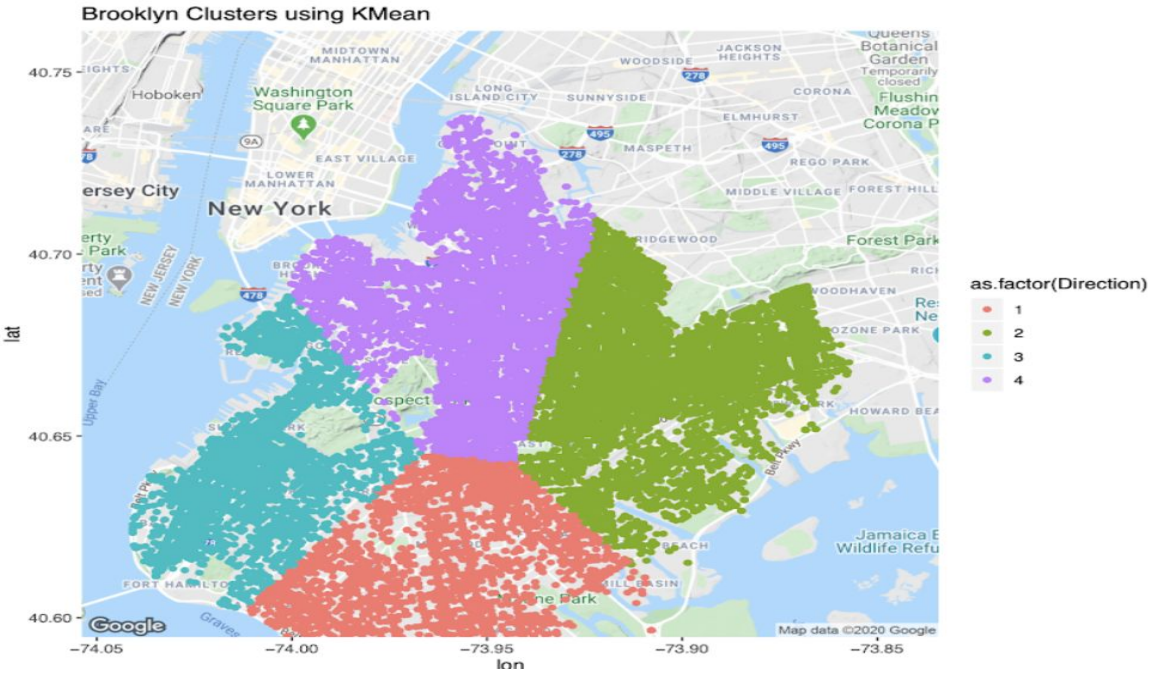


Figure 18: Brooklyn Clusters

1 = Upper Manhattan, 2 = Lower Manhattan, 3 = Washington Heights, 4 =Midtown Manhattan

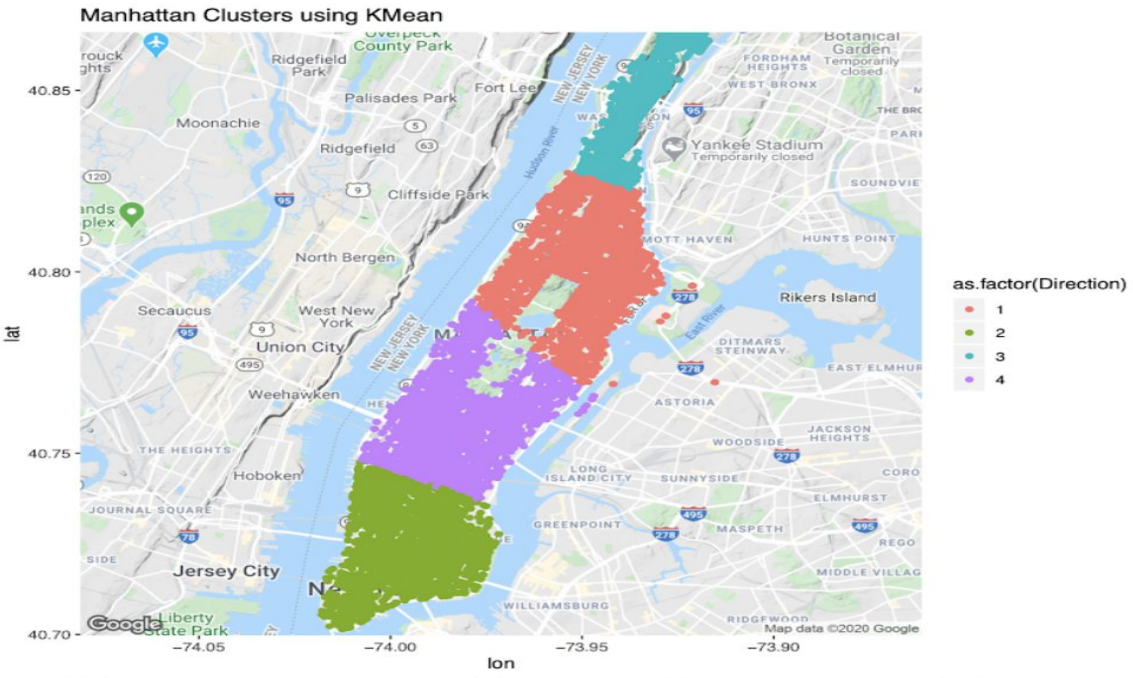


Figure 19: Manhattan Clusters

1 =Rockaway Park , 2 = Astoria, 3=Jamaica , 4 =Flushing

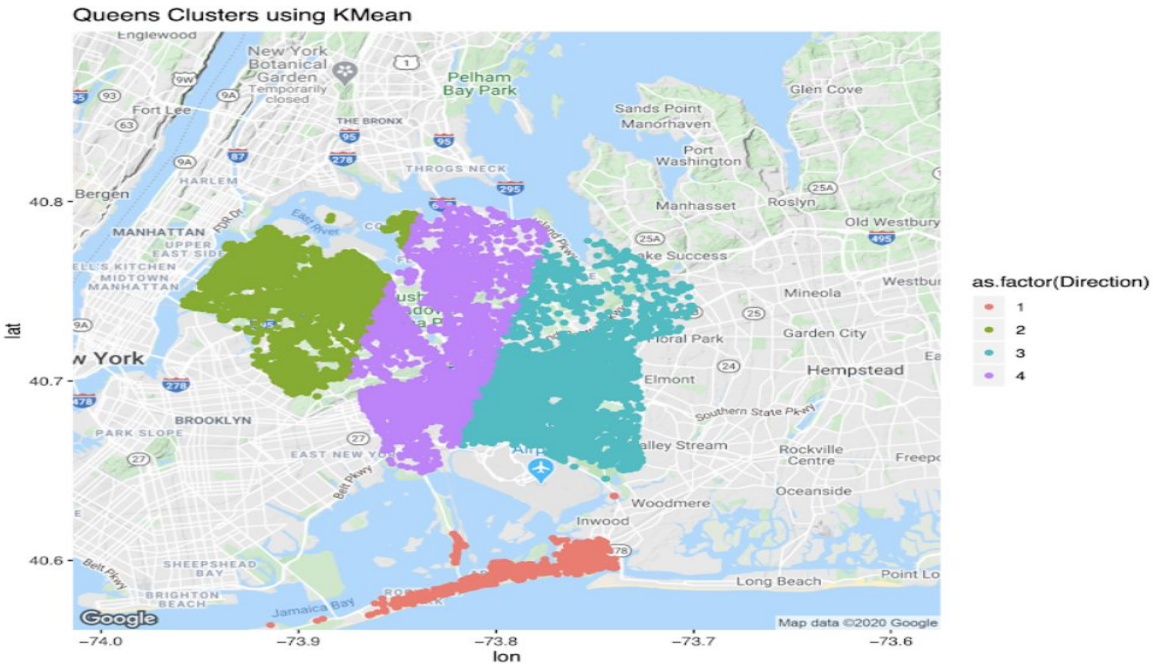


Figure 20: Queens Clusters

1 =North Shore , 2 =Greater St.George , 3 = South Shore, 4 =Gateway

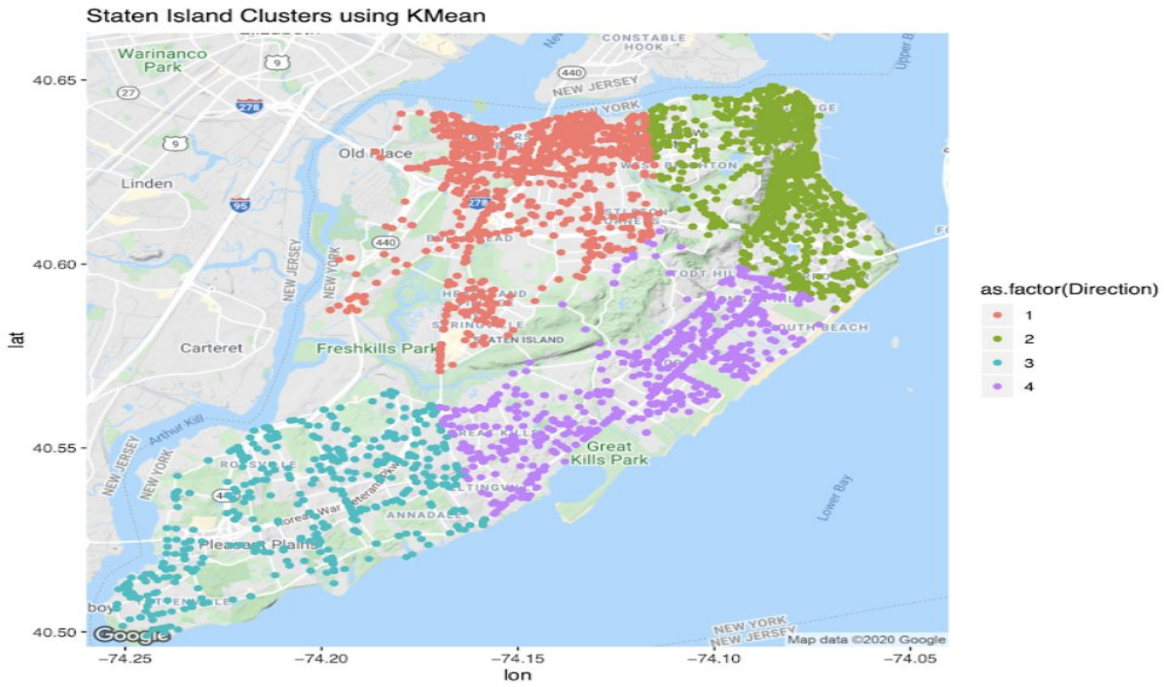


Figure 21: Staten Island Clusters

ITEM FREQUENCY PLOT

One way we visualized the association rules, that were generated by the apriori algorithm, is through an item frequency plot. We first install the package **RColorBrewer** to visualize the data through a bar graph. We plot three graphs using the function **itemFrequencyPlot**, where we specify our items are the rules generated by the apriori algorithm. We chose to display the top 29 absolute rules, as shown in Figure 22.

Figure 22: Item Frequency Plot Code

The following graphs display the 29 best rules based on which attribute is the most likely to have the most crimes or commit the most crimes. From Figures 23,24, and 25 we can further see the results. We can conclude that the sex that is more likely to commit crimes is male. The age group that is likely to commit crimes are between 25-44. The location that is more likely to have crimes is in Brooklyn in direction 2 which represents northeast Brooklyn, as shown in Figure 18. We are also able to see that the most common crime is Assault 3, and crimes are more likely to take place in the month of January.

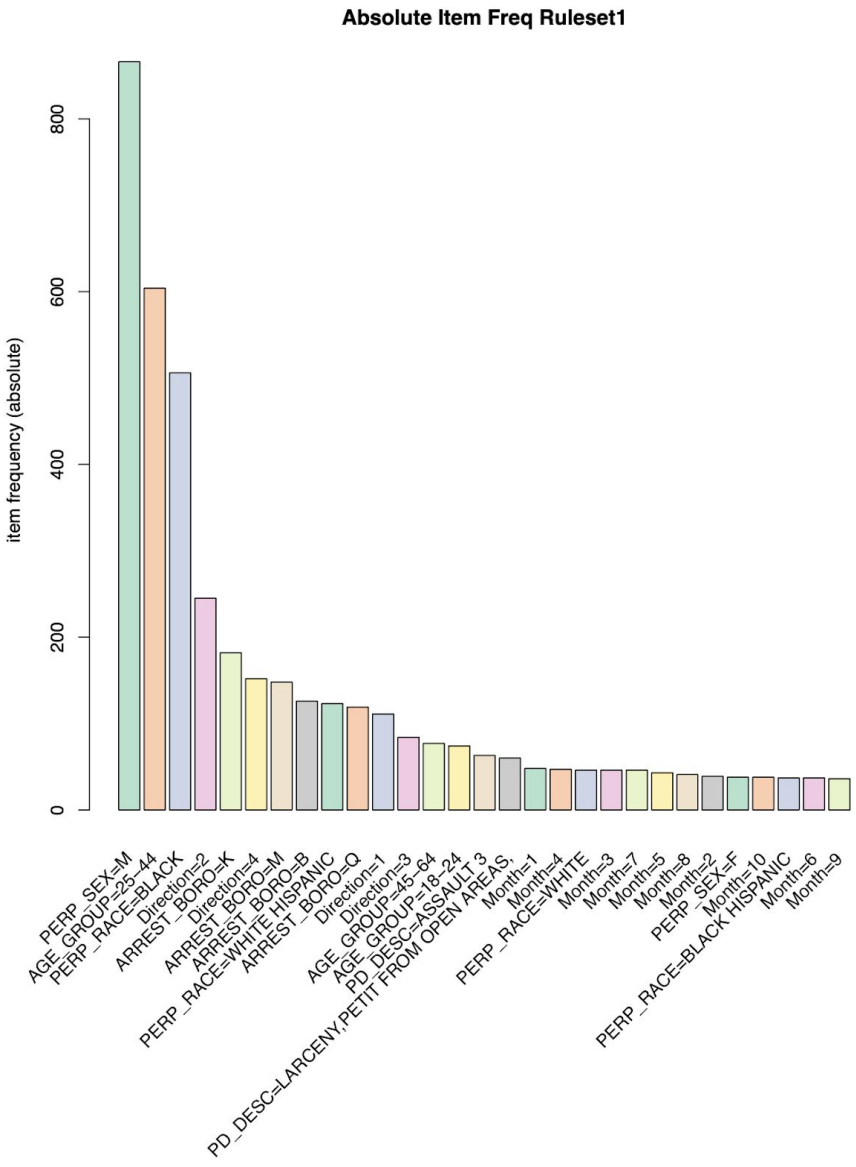


Figure 23: Item Frequency Plot for Ruleset 1

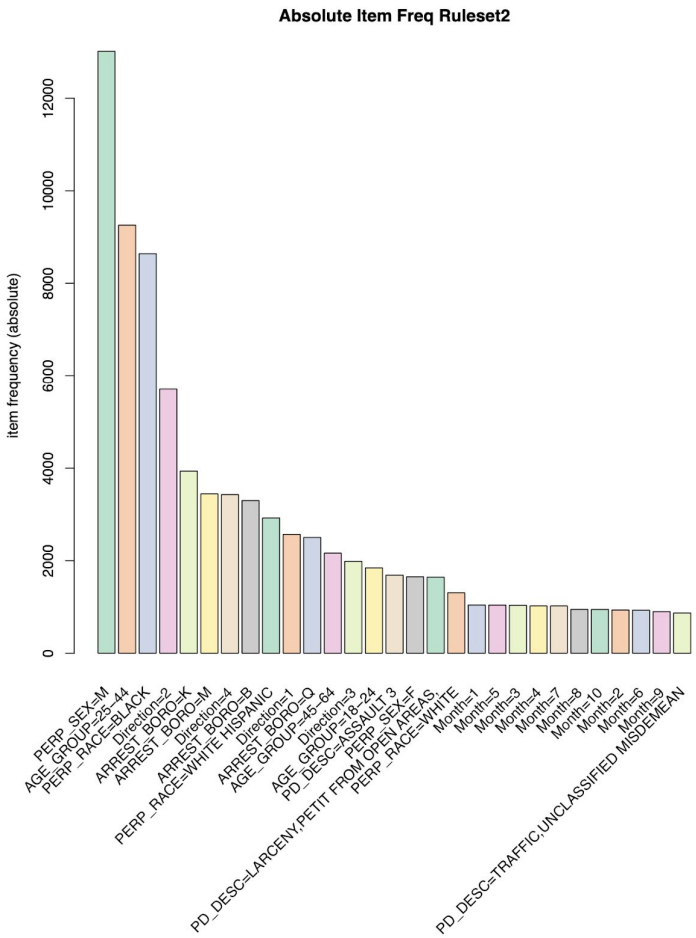


Figure 24: Item Frequency Plot for Ruleset 2

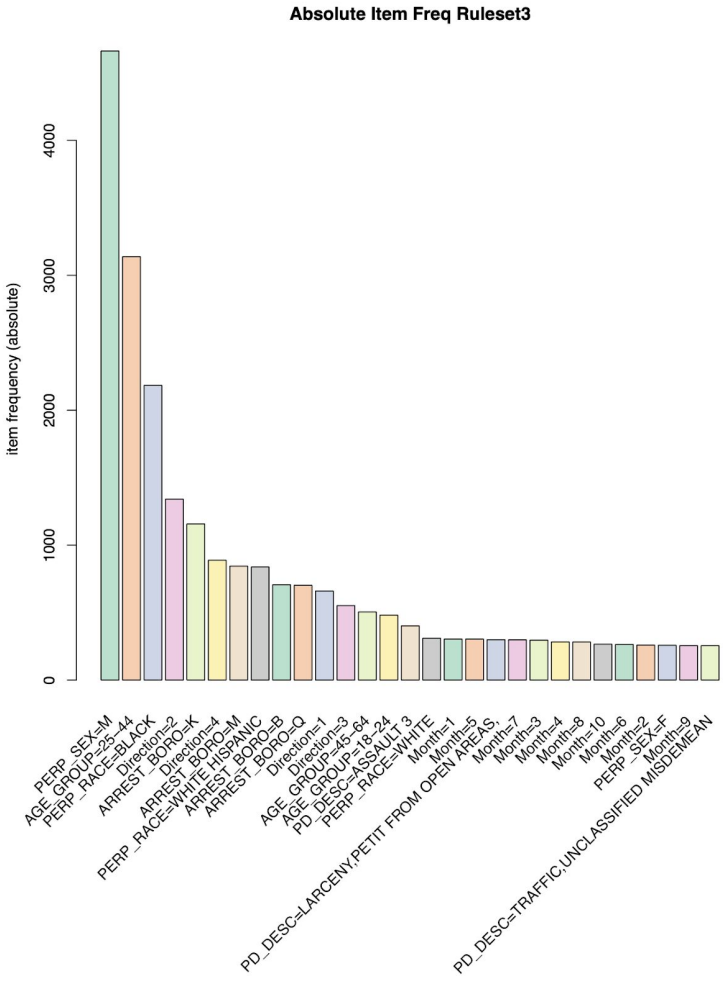


Figure 25: Item Frequency Plot for Ruleset 3

INTERACTIVE SCATTER PLOTS

Another way we visualized the association rules, that were generated by the apriori algorithm, is through an interactive scatter plot. We first, read the ruleSets for each analysis and indicate 10 as our top ten rules. We plot the graph by using the function **plotly_arules** as shown in figure 25. To view the interactive scatter plot, we have provided the following three links for each analysis to a published scatter plot:

Rule Set 1: https://rpubs.com/ftovar/611036
Rule Set 2: https://rpubs.com/ftovar/611057
Rule Set 3: https://rpubs.com/ftovar/611059

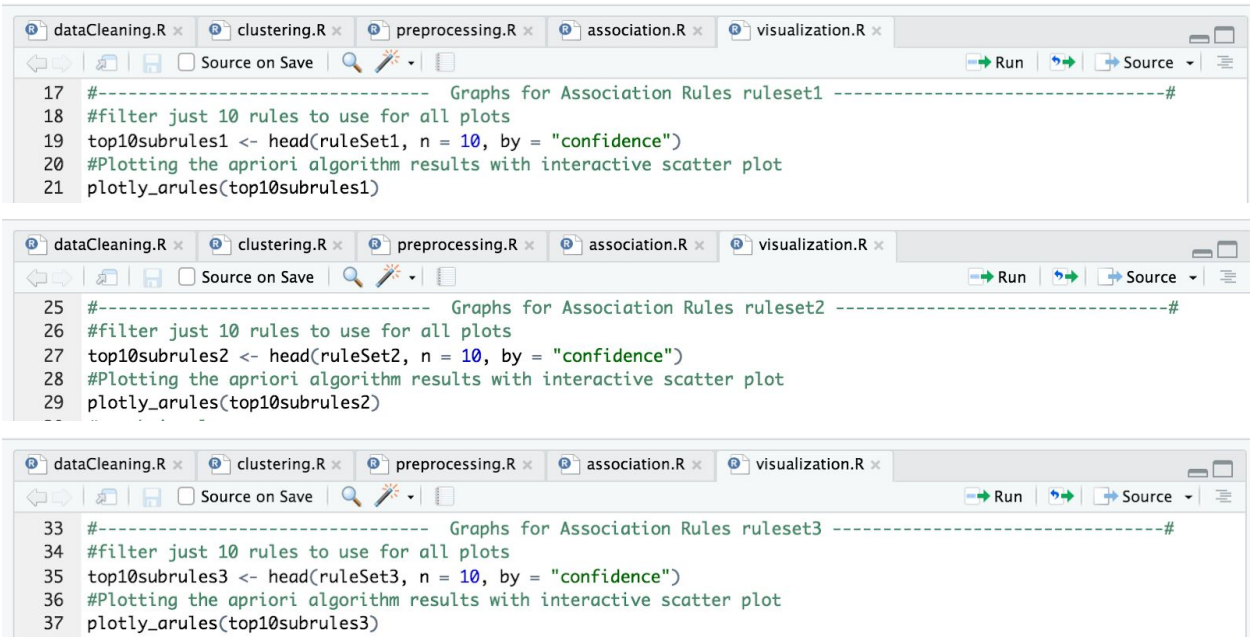


Figure 25: Interactive Scatter Plot Code for all three Rulesets

In Figure 26, we are able to analyze the data based on the support, confidence, and lift. This graph is an interactive scatter plot, therefore, we are able to see the details of each plot, in terms of which association rule is the top 10.

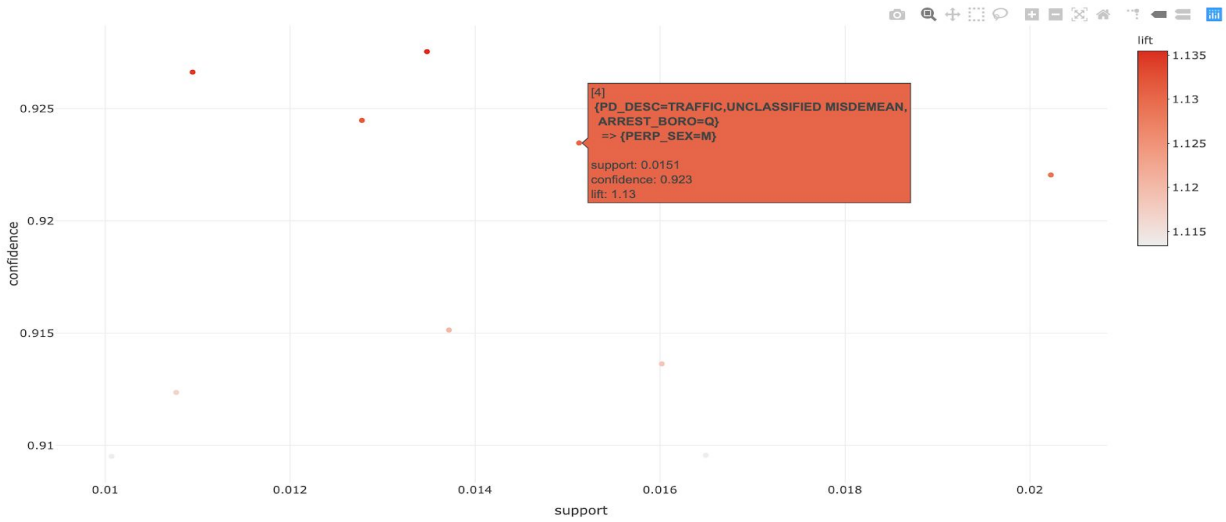


Figure 26: Interactive Scatter Plot Example

As shown in Figure 26, we are able to see an example that indicates that the arrest crime of **traffic unclassified misdemeanor**, and borough of **Queens** is associated with the sex **male**. The support is 0.0151 and the confidence is 0.923 and the lift is 1.13.

VISUALIZATION WIDGETS

Third way we visualized the association rules, that were generated by the apriori algorithm, is through visualization widgets. We use the function **plot** and indicate the number of rules as well the method which is indicated as graph and the engine as **htmlwidget** as shown in figure 27. To view the visualization widgets, we have provided the following three links for each analysis to a published visualization widget:

Rule Set 1: https://rpubs.com/ftovar/611060
Rule Set 2: https://rpubs.com/ftovar/611061
Rule Set 3: https://rpubs.com/ftovar/611062

dataCleaning.R | clustering.R | preprocessing.R | association.R | visualization.R

Source on Save | Run | Source

```
17 #--------------------------------  Graphs for Association Rules ruleset1 --------------------------------#
18 #graphvisual
19 plot (top10subrules1, method = "graph", engine = "htmlwidget")
20
```

dataCleaning.R | clustering.R | preprocessing.R | association.R | visualization.R

Source on Save | Run | Source

```
26 #--------------------------------  Graphs for Association Rules ruleset2 --------------------------------#
27 #graphvisual
28 plot (top10subrules2, method = "graph", engine = "htmlwidget")
29
```

dataCleaning.R | clustering.R | preprocessing.R | association.R | visualization.R

Source on Save | Run | Source

```
36 #--------------------------------  Graphs for Association Rules ruleset3 --------------------------------#
37 #graphvisual
38 plot (top10subrules3, method = "graph", engine = "htmlwidget")
39
```

Figure 27: Visualization Widget Code for all three Rulesets

In Figure 28, we are able to analyze the data based on the top 10 rules. This graph is an interactive graph where we can select the ID and analyze the graph, therefore, we are able to see the details of each plot.

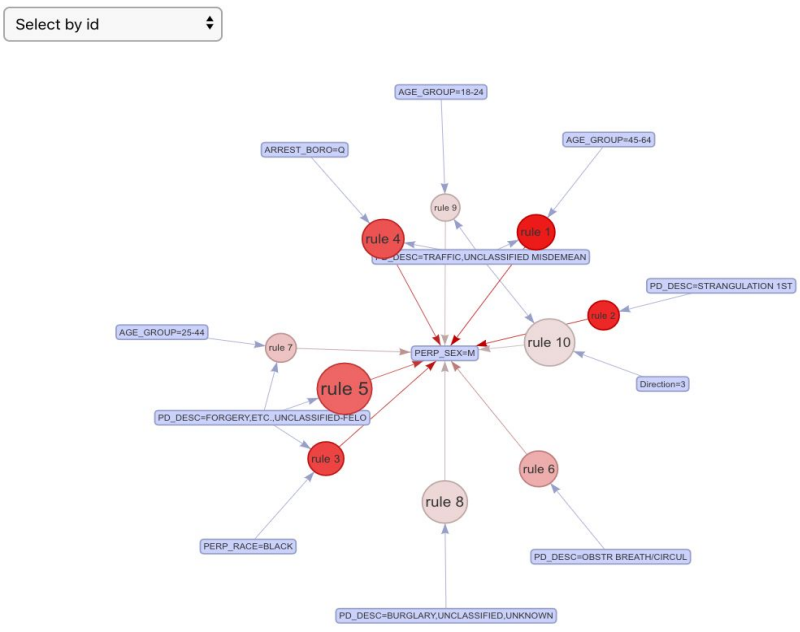


Figure 28: Visualization Widgets

As shown in Figure 29, we are able to see an example that indicates that the arrest crime of **traffic unclassified misdemeanor**, is associated with age group 18-24, men, the borough of Queens and involves rules 1, 4, 9, 10.

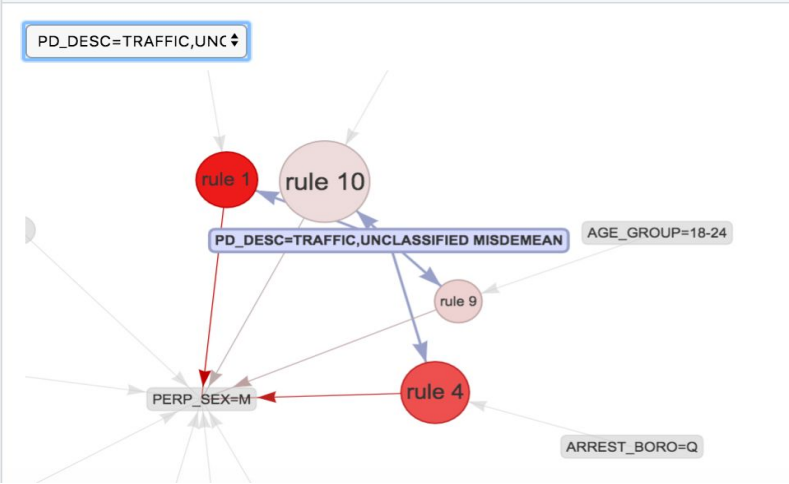


Figure 29: Visualization Widgets Rule 10

CONCLUSIONS

In conclusion, we were able to analyze the dataset for NYPD Arrests for 2019. Through preprocessing we were able to clean the data, cluster, and convert data types. We used the apriori algorithm to find different rule sets with different support and confidence. Through the support we were able to see how frequently rules were likely to occur. Through the confidence we were able to see how likely it is that if x is occurring that y is also occurring. Through the lift we were able to see the performance of the rules and it's likeliness.

Overall, we were able to determine that the sex that is more likely to commit crimes is male. The age group that is likely to commit crimes are between 25-44. The location that is more likely to have crimes is in Brooklyn in direction 2 which represents northeast Brooklyn. We are also able to see that the most common crime is Assault 3, and crimes are more likely to take place in the month of January.

TEAM CONTRIBUTIONS

Neha Bala:

1. Preprocessing methods: Clustering
2. Generate the Apriori Algorithm & Analyze Rules

Fernanda Tovar:

1. Preprocessing methods: Cleaning the data, Converting data types
2. Plotting Multiple Graphs & Visual Analysis

Arpit Battu:

1. Preprocessing methods: Cleaning the data, Clustering
2. Plotting Multiple Graphs & Visual Analysis

Github Code link : https://github.com/tovfer17/NYCArrests