SCHOOL SAFETY REPORT 2013-2014

GROUP 07

JYOTSNA NAMDEO NAKTE
RAHUL CHAKRABORTY
JONATHAN GABLER

TABLE OF CONTENTS

SUMMARY

New York City Police Department (NYPD) has been tasked with the collection and maintenance of crime data for incidents that occur in New York City public schools. The NYPD has provided this data to the New York City Department of Education (DOE). The DOE has compiled this data by schools and locations for the information of parents and students, teachers and staff, and the general-public. The Database motivates towards analyzing number of major, violent, property crimes taking place. Database has crime report according to the locations it takes place. Analysis will help further the general-public to choose safe schools for their children growth and development. Even take various action towards the crimes occurring in the school. The dataset even displays location of the schools where no crimes making it simple for parents to choose preferable environment for their children.

INTRODUCTION

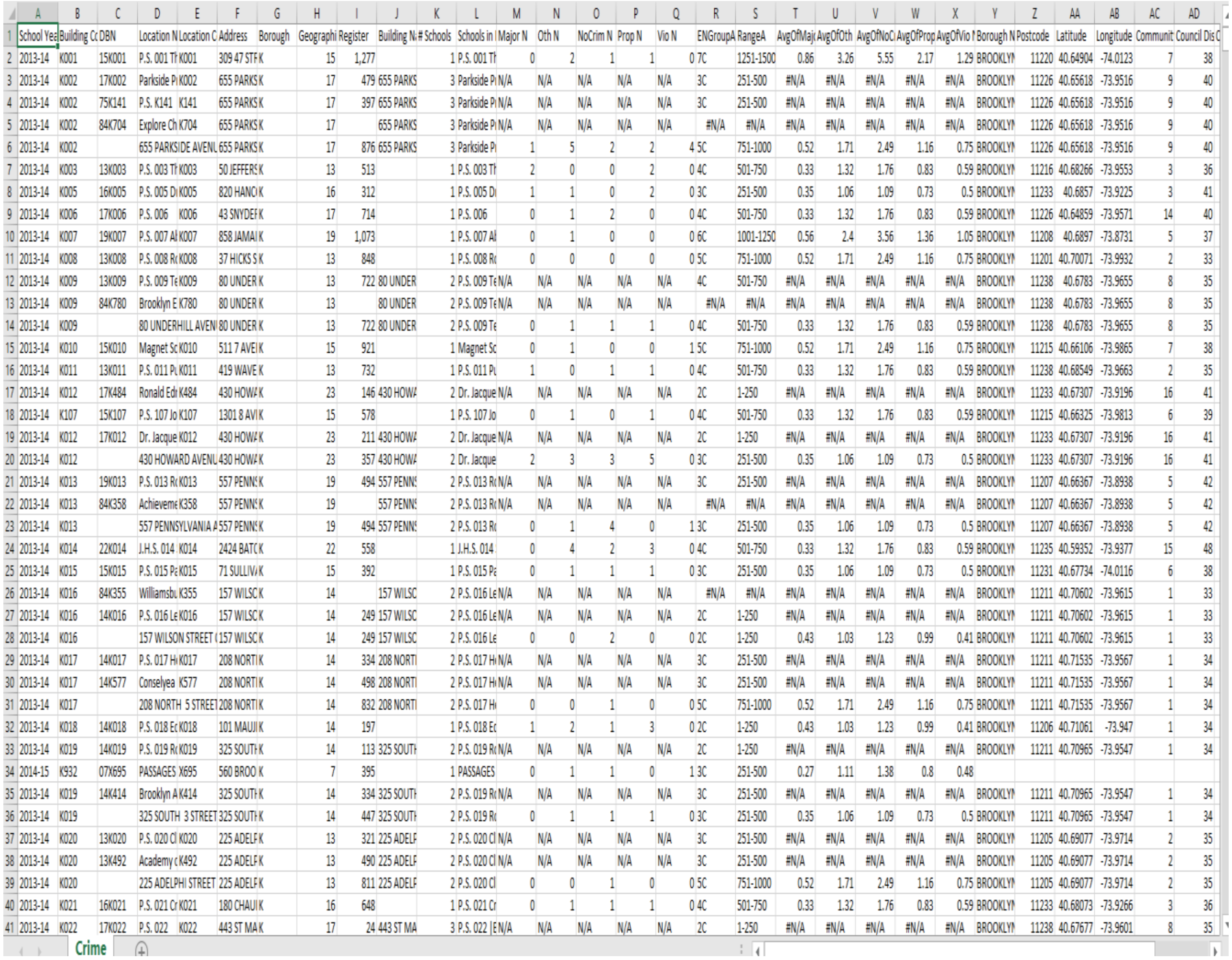
Schools are meant to be a safe environment for teaching and learning and it is of prime importance to keep students safe. Unfortunately, every year, there is a substantial reporting of criminal activities in several schools across the country. Since 1998, the New York Police Department (NYPD) has been collecting and maintaining crime data for incidents that have been occurring in public schools in New York city. This data has been provided to the Department of Education(DOE). The DOE has sorted the data by schools and locations and has made it open and available to the general-public at the NYC Open Data website.

For this project, we have downloaded the 2010-2016 School Safety Report dataset from the NYC Open Data website. Since in some instances, many Department of Education learning communities co-exist within the same building, the data presented is building-specific rather than school-specific. The dataset comprises of only one .csv flat file. This means we have to perform operations on the dataset like splitting, cleaning, normalization, etc. to make it more meaningful and understandable.

We have used Microsoft SQL Server for structuring, cleaning and normalizing the database and R for analysis and visualization. In this report, we have described the steps taken and the subsequent challenges faced in performing the required tasks.

DATASET DESIGN

The Dataset chosen was flat file 2010 - 2016 School Safety Report. The Dataset Information is imported from Agency Department of Education (DOE) having 34 columns 6310 Rows with many duplicated values, missing values, wrong values placed in the dataset. [1]



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	School Yea	Building Cc	DBN	Location N	Location C	Address	Borough	Geographi	Register	Building Na	# Schools	Schools in	Major N	Oth N	NoCrim N	Prop N	Vio N	ENGroupA	RangeA	AvgOfMajc	AvgOfOth	AvgOfNoC	AvgOfProp	AvgOfVio N	Borough N	Postcode	Latitude	Longitude	Communit	Council Dis C
2	2013-14	K001	15K001	P.S. 001 Th	K001	309 47 STF	K	15	1,277		1	P.S. 001 Th	0	2	1	1	0	7C	1251-1500	0.86	3.26	5.55	2.17	1.29	BROOKLYN	11220	40.64904	-74.0123	7	38
3	2013-14	K002	17K002	Parkside Pr	K002	655 PARKS	K	17	479	655 PARKS	3	Parkside Pr	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11226	40.65618	-73.9516	9	40
4	2013-14	K002	75K141	P.S. K141	K141	655 PARKS	K	17	397	655 PARKS	3	Parkside Pr	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11226	40.65618	-73.9516	9	40
5	2013-14	K002	84K704	Explore Ch	K704	655 PARKS	K	17		655 PARKS	3	Parkside Pr	N/A	N/A	N/A	N/A	N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11226	40.65618	-73.9516	9	40
6	2013-14	K002		655 PARKSIDE AVENU		655 PARKS	K	17	876	655 PARKS	3	Parkside Pr	1	5	2	2	4	5C	751-1000	0.52	1.71	2.49	1.16	0.75	BROOKLYN	11226	40.65618	-73.9516	9	40
7	2013-14	K003	13K003	P.S. 003 Th	K003	50 JEFFERS	K	13	513		1	P.S. 003 Th	2	0	0	2	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11216	40.68266	-73.9553	3	36
8	2013-14	K005	16K005	P.S. 005 Di	K005	820 HANC	K	16	312		1	P.S. 005 Di	1	1	0	2	0	3C	251-500	0.35	1.06	1.09	0.73	0.5	BROOKLYN	11233	40.6857	-73.9225	3	41
9	2013-14	K006	17K006	P.S. 006	K006	43 SNYDEF	K	17	714		1	P.S. 006	0	1	2	0	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11226	40.64859	-73.9571	14	40
10	2013-14	K007	19K007	P.S. 007 Ab	K007	858 JAMAI	K	19	1,073		1	P.S. 007 Ab	0	1	0	0	0	6C	1001-1250	0.56	2.4	3.56	1.36	1.05	BROOKLYN	11208	40.6897	-73.8731	5	37
11	2013-14	K008	13K008	P.S. 008 Ro	K008	37 HICKS S	K	13	848		1	P.S. 008 Ro	0	0	0	0	0	5C	751-1000	0.52	1.71	2.49	1.16	0.75	BROOKLYN	11201	40.70071	-73.9932	2	33
12	2013-14	K009	13K009	P.S. 009 Te	K009	80 UNDER	K	13	722	80 UNDER	2	P.S. 009 Te	N/A	N/A	N/A	N/A	N/A	4C	501-750	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11238	40.6783	-73.9655	8	35
13	2013-14	K009	84K780	Brooklyn E	K780	80 UNDER	K	13		80 UNDER	2	P.S. 009 Te	N/A	N/A	N/A	N/A	N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11238	40.6783	-73.9655	8	35
14	2013-14	K009		80 UNDERHILL AVEN		80 UNDER	K	13	722	80 UNDER	2	P.S. 009 Te	0	1	1	1	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11238	40.6783	-73.9655	8	35
15	2013-14	K010	15K010	Magnet Sc	K010	511 7 AVE	K	15	921		1	Magnet Sc	0	1	0	0	1	5C	751-1000	0.52	1.71	2.49	1.16	0.75	BROOKLYN	11215	40.66106	-73.9865	7	38
16	2013-14	K011	13K011	P.S. 011 Pu	K011	419 WAVE	K	13	732		1	P.S. 011 Pu	1	0	1	1	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11238	40.68549	-73.9663	2	35
17	2013-14	K012	17K484	Ronald Edr	K484	430 HOWA	K	23	146	430 HOWA	2	Dr. Jacque	N/A	N/A	N/A	N/A	N/A	2C	1-250	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11233	40.67307	-73.9196	16	41
18	2013-14	K107	15K107	P.S. 107 Jo	K107	1301 8 AVI	K	15	578		1	P.S. 107 Jo	0	1	0	1	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11215	40.66325	-73.9813	6	39
19	2013-14	K012	17K012	Dr. Jacque	K012	430 HOWA	K	23	211	430 HOWA	2	Dr. Jacque	N/A	N/A	N/A	N/A	N/A	2C	1-250	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11233	40.67307	-73.9196	16	41
20	2013-14	K012		430 HOWARD AVENU		430 HOWA	K	23	357	430 HOWA	2	Dr. Jacque	2	3	3	5	0	3C	251-500	0.35	1.06	1.09	0.73	0.5	BROOKLYN	11233	40.67307	-73.9196	16	41
21	2013-14	K013	19K013	P.S. 013 Ro	K013	557 PENNS	K	19	494	557 PENNS	2	P.S. 013 Ro	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11207	40.66367	-73.8938	5	42
22	2013-14	K013	84K358	Achieveme	K358	557 PENNS	K	19		557 PENNS	2	P.S. 013 Ro	N/A	N/A	N/A	N/A	N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11207	40.66367	-73.8938	5	42
23	2013-14	K013		557 PENNSYLVANIA A		557 PENNS	K	19	494	557 PENNS	2	P.S. 013 Ro	0	1	4	0	1	3C	251-500	0.35	1.06	1.09	0.73	0.5	BROOKLYN	11207	40.66367	-73.8938	5	42
24	2013-14	K014	22K014	J.H.S. 014	K014	2424 BATC	K	22	558		1	J.H.S. 014	0	4	2	3	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11235	40.59352	-73.9377	15	48
25	2013-14	K015	15K015	P.S. 015 Pa	K015	71 SULLIVA	K	15	392		1	P.S. 015 Pa	0	1	1	1	0	3C	251-500	0.35	1.06	1.09	0.73	0.5	BROOKLYN	11231	40.67734	-74.0116	6	38
26	2013-14	K016	84K355	Williamsbu	K355	157 WILSC	K	14		157 WILSC	2	P.S. 016 Le	N/A	N/A	N/A	N/A	N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.70602	-73.9615	1	33
27	2013-14	K016	14K016	P.S. 016 Le	K016	157 WILSC	K	14	249	157 WILSC	2	P.S. 016 Le	N/A	N/A	N/A	N/A	N/A	2C	1-250	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.70602	-73.9615	1	33
28	2013-14	K016		157 WILSON STREET (		157 WILSC	K	14	249	157 WILSC	2	P.S. 016 Le	0	0	2	0	0	2C	1-250	0.43	1.03	1.23	0.99	0.41	BROOKLYN	11211	40.70602	-73.9615	1	33
29	2013-14	K017	14K017	P.S. 017 He	K017	208 NORTH	K	14	334	208 NORTH	2	P.S. 017 He	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.71535	-73.9567	1	34
30	2013-14	K017	14K577	Conselyea	K577	208 NORTH	K	14	498	208 NORTH	2	P.S. 017 He	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.71535	-73.9567	1	34
31	2013-14	K017		208 NORTH 5 STREET		208 NORTH	K	14	832	208 NORTH	2	P.S. 017 He	0	0	1	0	0	5C	751-1000	0.52	1.71	2.49	1.16	0.75	BROOKLYN	11211	40.71535	-73.9567	1	34
32	2013-14	K018	14K018	P.S. 018 Ed	K018	101 MAUJI	K	14	197		1	P.S. 018 Ed	1	2	1	3	0	2C	1-250	0.43	1.03	1.23	0.99	0.41	BROOKLYN	11206	40.71061	-73.947	1	34
33	2013-14	K019	14K019	P.S. 019 Ro	K019	325 SOUTH	K	14	113	325 SOUTH	2	P.S. 019 Ro	N/A	N/A	N/A	N/A	N/A	2C	1-250	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.70965	-73.9547	1	34
34	2014-15	K932	07X695	PASSAGES	X695	560 BROO	K	7	395		1	PASSAGES	0	1	1	0	1	3C	251-500	0.27	1.11	1.38	0.8	0.48						
35	2013-14	K019	14K414	Brooklyn A	K414	325 SOUTH	K	14	334	325 SOUTH	2	P.S. 019 Ro	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11211	40.70965	-73.9547	1	34
36	2013-14	K019		325 SOUTH 3 STREET		325 SOUTH	K	14	447	325 SOUTH	2	P.S. 019 Ro	0	1	1	1	0	3C	251-500	0.35	1.06	1.09	0.73	0.5	BROOKLYN	11211	40.70965	-73.9547	1	34
37	2013-14	K020	13K020	P.S. 020 Cl	K020	225 ADELF	K	13	321	225 ADELF	2	P.S. 020 Cl	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11205	40.69077	-73.9714	2	35
38	2013-14	K020	13K492	Academy c	K492	225 ADELF	K	13	490	225 ADELF	2	P.S. 020 Cl	N/A	N/A	N/A	N/A	N/A	3C	251-500	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11205	40.69077	-73.9714	2	35
39	2013-14	K020		225 ADELPHI STREET		225 ADELF	K	13	811	225 ADELF	2	P.S. 020 Cl	0	0	1	0	0	5C	751-1000	0.52	1.71	2.49	1.16	0.75	BROOKLYN	11205	40.69077	-73.9714	2	35
40	2013-14	K021	16K021	P.S. 021 Cr	K021	180 CHAUI	K	16	648		1	P.S. 021 Cr	0	1	1	1	0	4C	501-750	0.33	1.32	1.76	0.83	0.59	BROOKLYN	11233	40.68073	-73.9266	3	36
41	2013-14	K022	17K022	P.S. 022	K022	443 ST MA	K	17	24	443 ST MA	3	P.S. 022 (E	N/A	N/A	N/A	N/A	N/A	2C	1-250	#N/A	#N/A	#N/A	#N/A	#N/A	BROOKLYN	11238	40.67677	-73.9601	8	35

Fig 1: Flat File of the Dataset

We further decided to normalize the dataset thus by converting into database in normalized form to do analysis through SQL and Visualizations through R.

Following were the attributes in the dataset with description: [1]

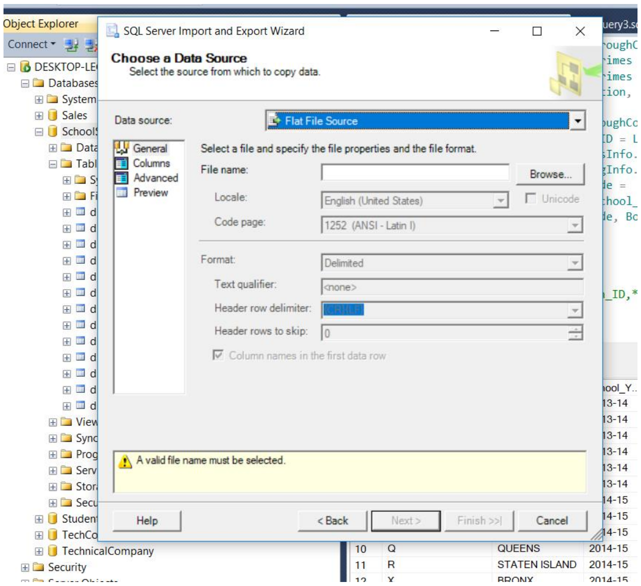
ATTRIBUTES	DESCRIPTION
School Year	Year of the School
Building Code	Unique Code associated to buildings
Location Code	a unique identifier that can include schools, administrative offices, learning communities, etc. When the Learning_Community_Name = ‘School’, the Location_Code is a combination of the borough code and the school number.
Location Name	the name by which the organization is known. For a learning community, it is the official title of the school
Address	Address of the Buildings
Borough	NYC Boro the location is situated
Geographical District Code	the school’s geographical district as defined by the NYC Department of Education.
Register	Number of students on register
Building Name	the official name of the building a school is located
Schools	number of schools in in the building
Schools in Building	names of the schools in the buildings
Major N	number of major crimes
Oth N	number of other crimes
NoCrim N	number of No Crimes crimes
Prop N	number of Property crimes
Vio N	number of violent crimes
ENGroup	group name that the building population falls under
Range	Building population
AvgofMajor N	Average major crimes according to groups
AvgofOth N	Average other crimes according to groups
AvgofNoCrim N	Average no crimes according to groups
AvgofPropN	Average property crimes according to groups
AvgofVio N	Average violent crimes according to groups
Borough Name	Borough Name
Postcode	Postcode of Address
Latitude	Latitude of the Building location
Longitude	Longitude of the Building location
Community Board	Board which looks after the school
Council District	District Council belongs to which school
NTA	Location of the Council

Table 1: Dataset Description

IMPROVISING DATABASE DESIGN AND NORMALISATION

We first started with studying about each attribute and analyzing it individually to understand how to build the database. With the help of the description and according to the goal of the project we drew a rough ERD and separated the dimension into various CSV files and further imported into SSMS (Microsoft SQL SERVER) to create our database.

With the help of SQL Import and Export Wizard we created our database SchoolSafetyReport[9]



SQL Server Import and Export Wizard

Choose a Data Source

Select the source from which to copy data.

Data source: Flat File Source

Select a file and specify the file properties and the file format.

File name:

Locale: English (United States)

☐ Unicode

Code page: 1252 (ANSI - Latin I)

Format: Delimited

Text qualifier: <none>

Header row delimiter: {CR}{LF}

Header rows to skip: 0

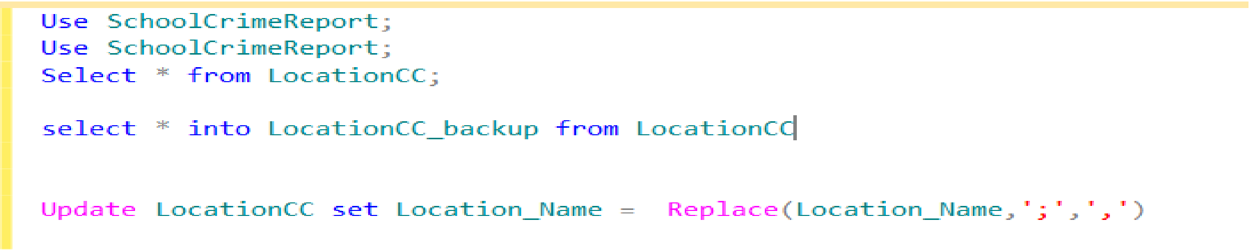
☑ Column names in the first data row

A valid file name must be selected.

10	Q	QUEENS	2014-15
11	R	STATEN ISLAND	2014-15

We faced many challenges in collecting data and creating database, the SchoolSafetyReport Database had many duplicated values, missing values, wrong values entered we certainly studied the data and did further cleaning to maintain high integrity and data quality. We normalized our data further to get best results while analyzing the data and developing conclusions.

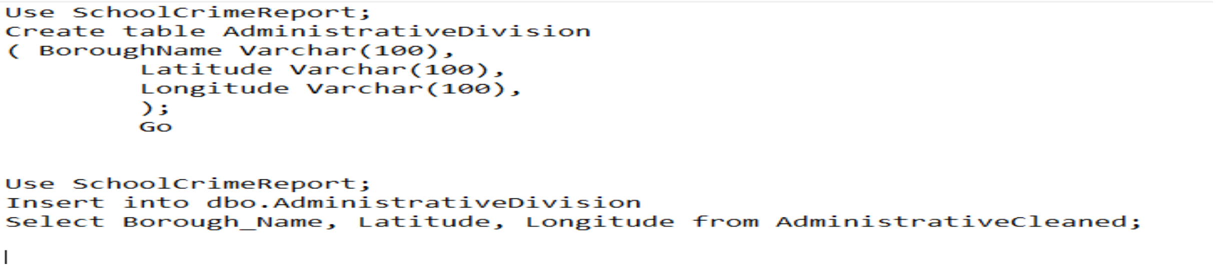
- The attribute Location had many address with comma quotation because of which while reading in csv file the attribute value shifted to other values we further replaced comma punctuation with semi-colon and later imported used the update and replace Query in SQL to further maintain the original Data.



```
Use SchoolCrimeReport;
Use SchoolCrimeReport;
Select * from LocationCC;

select * into LocationCC_backup from LocationCC


Update LocationCC set Location_Name =  Replace(Location_Name,';',',')
```

Fig 2: Query of Replacing ';' with ',' from Location

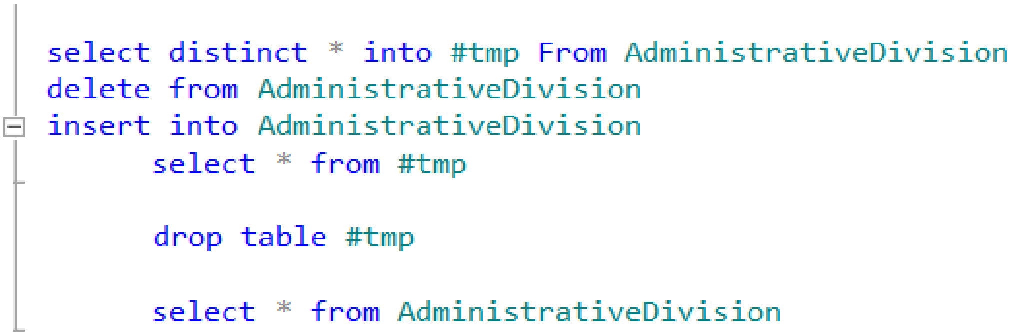
- We maintained our original file as Backup table and further created tables and applied our data cleaning to further normalize the data.



```
Use SchoolCrimeReport;
Create table AdministrativeDivision
( BoroughName Varchar(100),
        Latitude Varchar(100),
        Longitude Varchar(100),
        );
        Go


Use SchoolCrimeReport;
Insert into dbo.AdministrativeDivision
Select Borough_Name, Latitude, Longitude from AdministrativeCleaned;
```

Fig 3: Backup Tables

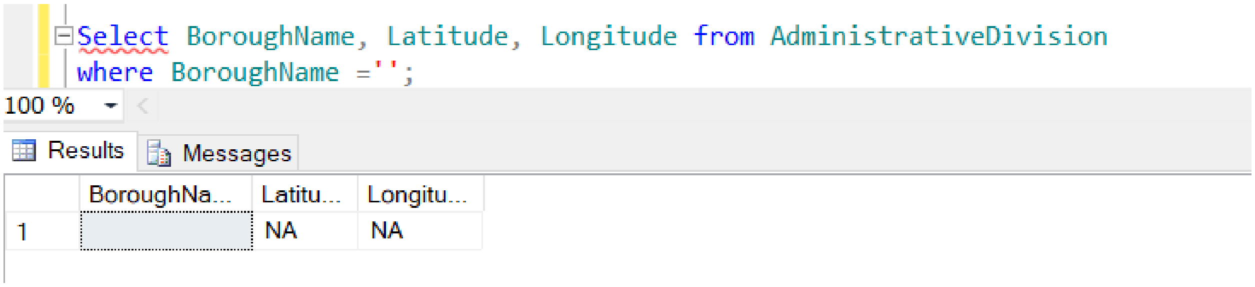
- We first started normalizing the database to find if it is 1NF, we found many duplicate values with repeating values. We further used the distinct SQL query to delete the duplicate values



```
select distinct * into #tmp From AdministrativeDivision
delete from AdministrativeDivision
insert into AdministrativeDivision
      select * from #tmp

      drop table #tmp

      select * from AdministrativeDivision
```

Fig 4: Finding Duplicates Value

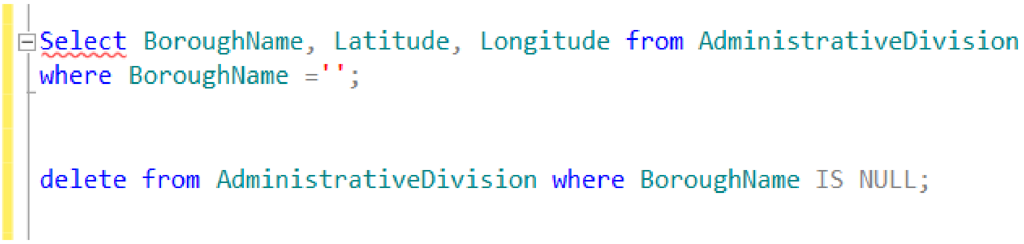
- The data had many null values which we further found and deleted using the where command and delete IS NULL Query.



```
Select BoroughName, Latitude, Longitude from AdministrativeDivision
where BoroughName ='';
```

	BoroughNa...	Latitu...	Longitu...
1		NA	NA

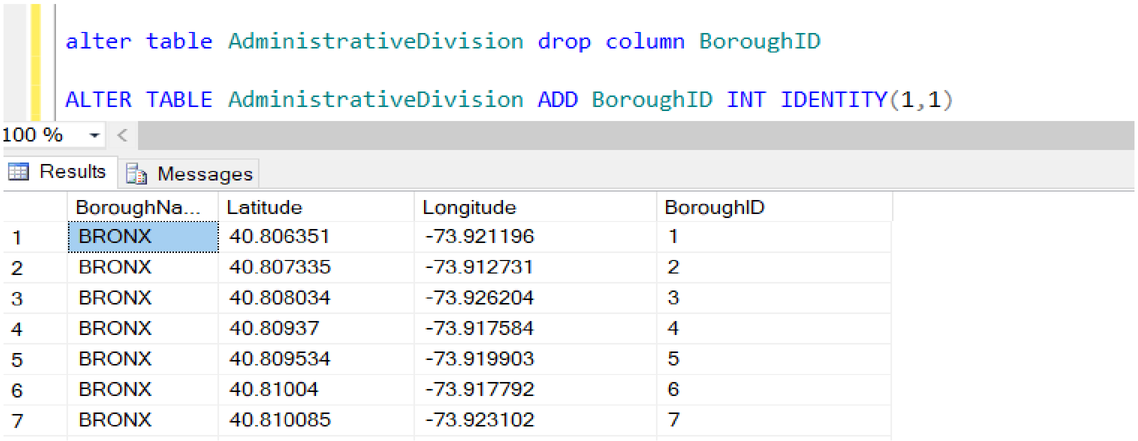
Fig 5: Null Value in Database



```
Select BoroughName, Latitude, Longitude from AdministrativeDivision
where BoroughName ='';


delete from AdministrativeDivision where BoroughName IS NULL;
```

Fig 6: Removing Null Value in Database

- Further to complete the 1NF form we kept all rows unique by introducing unique value primary key to the tables which did not had unique identification by following query.

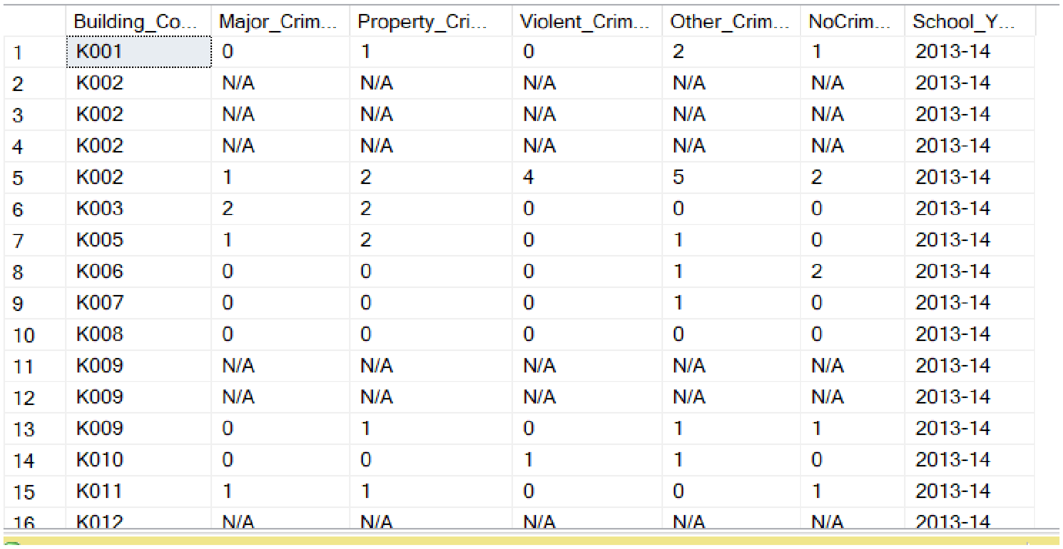


```
alter table AdministrativeDivision drop column BoroughID

ALTER TABLE AdministrativeDivision ADD BoroughID INT IDENTITY(1,1)
```

	BoroughNa...	Latitude	Longitude	BoroughID
1	BRONX	40.806351	-73.921196	1
2	BRONX	40.807335	-73.912731	2
3	BRONX	40.808034	-73.926204	3
4	BRONX	40.80937	-73.917584	4
5	BRONX	40.809534	-73.919903	5
6	BRONX	40.81004	-73.917792	6
7	BRONX	40.810085	-73.923102	7

Fig 7: Adding Primary Key

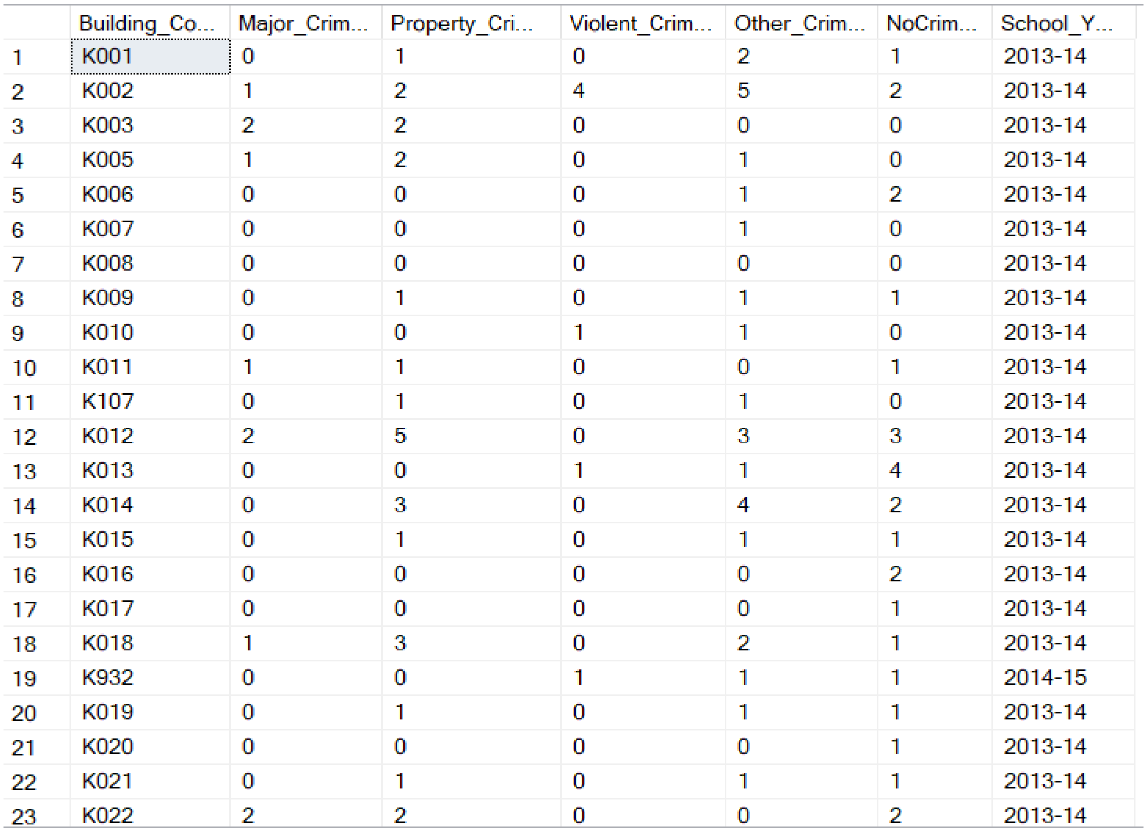
- In Crime Types we had many N/A values and duplicate redundant data as shown following



	Building_Co...	Major_Crim...	Property_Cri...	Violent_Crim...	Other_Crim...	NoCrim...	School_Y...
1	K001	0	1	0	2	1	2013-14
2	K002	N/A	N/A	N/A	N/A	N/A	2013-14
3	K002	N/A	N/A	N/A	N/A	N/A	2013-14
4	K002	N/A	N/A	N/A	N/A	N/A	2013-14
5	K002	1	2	4	5	2	2013-14
6	K003	2	2	0	0	0	2013-14
7	K005	1	2	0	1	0	2013-14
8	K006	0	0	0	1	2	2013-14
9	K007	0	0	0	1	0	2013-14
10	K008	0	0	0	0	0	2013-14
11	K009	N/A	N/A	N/A	N/A	N/A	2013-14
12	K009	N/A	N/A	N/A	N/A	N/A	2013-14
13	K009	0	1	0	1	1	2013-14
14	K010	0	0	1	1	0	2013-14
15	K011	1	1	0	0	1	2013-14
16	K012	N/A	N/A	N/A	N/A	N/A	2013-14

Fig 8: NA in Database

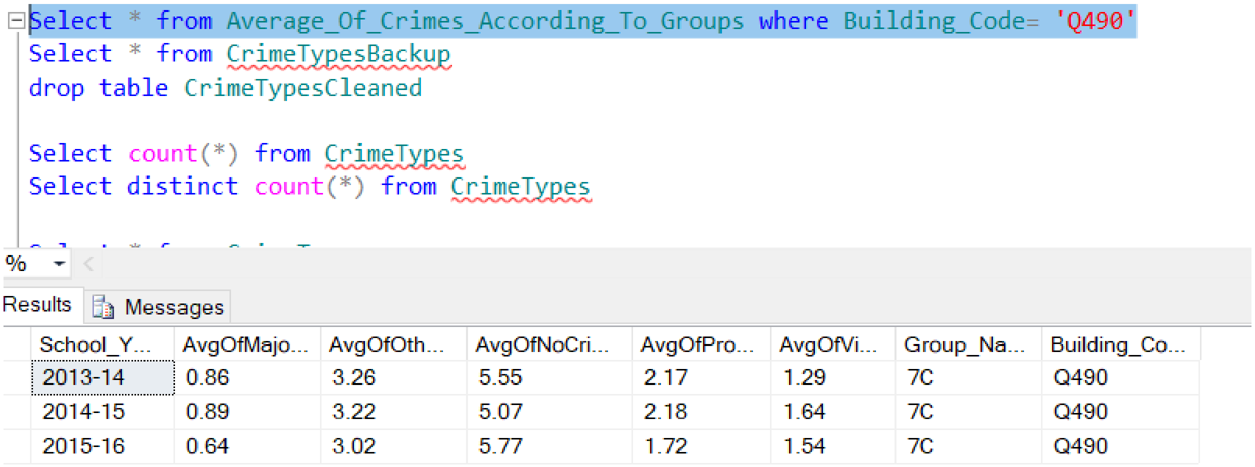
We further used the delete and distinct Query thus getting the clean data as follows:



	Building_Co...	Major_Crim...	Property_Cri...	Violent_Crim...	Other_Crim...	NoCrim...	School_Y...
1	K001	0	1	0	2	1	2013-14
2	K002	1	2	4	5	2	2013-14
3	K003	2	2	0	0	0	2013-14
4	K005	1	2	0	1	0	2013-14
5	K006	0	0	0	1	2	2013-14
6	K007	0	0	0	1	0	2013-14
7	K008	0	0	0	0	0	2013-14
8	K009	0	1	0	1	1	2013-14
9	K010	0	0	1	1	0	2013-14
10	K011	1	1	0	0	1	2013-14
11	K107	0	1	0	1	0	2013-14
12	K012	2	5	0	3	3	2013-14
13	K013	0	0	1	1	4	2013-14
14	K014	0	3	0	4	2	2013-14
15	K015	0	1	0	1	1	2013-14
16	K016	0	0	0	0	2	2013-14
17	K017	0	0	0	0	1	2013-14
18	K018	1	3	0	2	1	2013-14
19	K932	0	0	1	1	1	2014-15
20	K019	0	1	0	1	1	2013-14
21	K020	0	0	0	0	1	2013-14
22	K021	0	1	0	1	1	2013-14
23	K022	2	2	0	0	2	2013-14

Fig 9: Cleaning NA Value in database

- After cleaning we get the clean database table according to the crimes happening each year at the Location

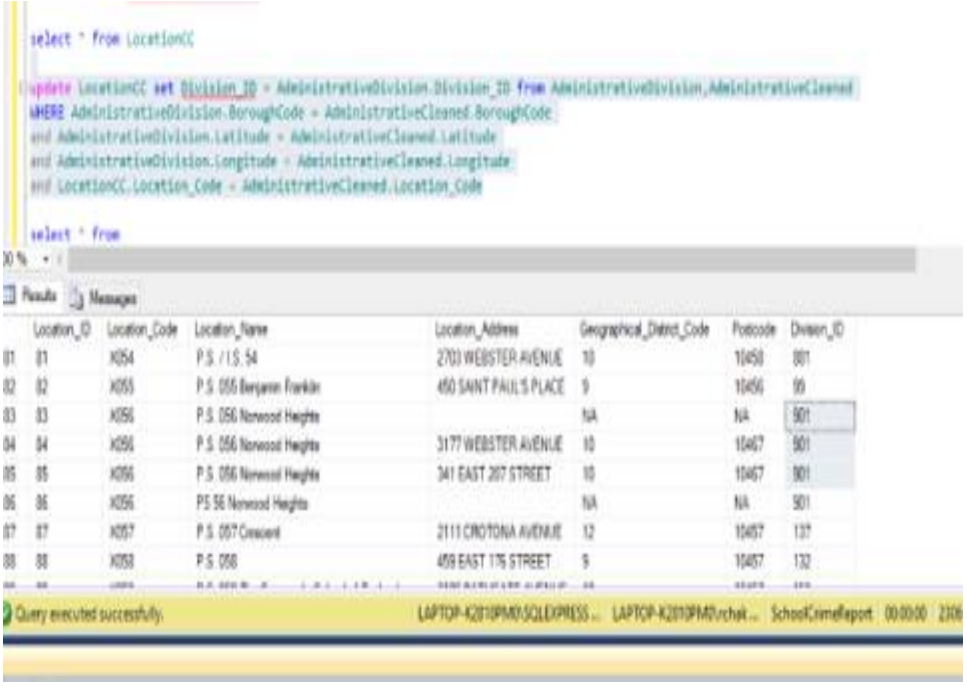


```
Select * from Average_Of_Crimes_According_To_Groups where Building_Code= 'Q490'
Select * from CrimeTypesBackup
drop table CrimeTypesCleaned

Select count(*) from CrimeTypes
Select distinct count(*) from CrimeTypes
```

School_Y...	AvgOfMajo...	AvgOfOth...	AvgOfNoCri...	AvgOfPro...	AvgOfVi...	Group_Na...	Building_Co...
2013-14	0.86	3.26	5.55	2.17	1.29	7C	Q490
2014-15	0.89	3.22	5.07	2.18	1.64	7C	Q490
2015-16	0.64	3.02	5.77	1.72	1.54	7C	Q490

Fig 10: Clean Database

- Further to develop the database to convert into 2NF we assigned unique values primary Key to ever rows where it was not present and we found out all partial dependencies with association.
- In order to set values of the primary keys as foreign Keys in the tables we further run the following SQL Query using UPDATE, SET, JOINS



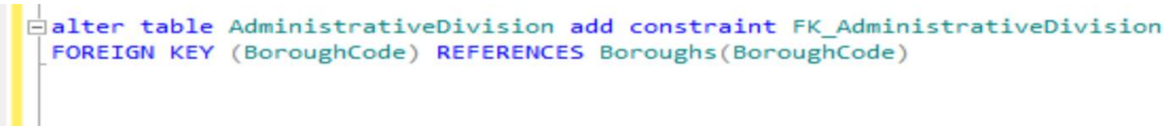
```
select * from LocationCC

update LocationCC set Division_ID = AdministrativeDivision.Division_ID from AdministrativeDivision,AdministrativeCleaned
WHERE AdministrativeDivision.BoroughCode = AdministrativeCleaned.BoroughCode
and AdministrativeDivision.Latitude = AdministrativeCleaned.Latitude
and AdministrativeDivision.Longitude = AdministrativeCleaned.Longitude
and LocationCC.Location_Code = AdministrativeCleaned.Location_Code

select * from
```

	Location_ID	Location_Code	Location_Name	Location_Address	Geographical_District_Code	Postcode	Division_ID
81	81	X054	P.S. / I.S. 54	2703 WEBSTER AVENUE	10	10458	801
82	82	X055	P.S. 055 Benjamin Franklin	450 SAINT PAUL'S PLACE	9	10456	95
83	83	X056	P.S. 056 Norwood Heights		NA	NA	901
84	84	X056	P.S. 056 Norwood Heights	3177 WEBSTER AVENUE	10	10467	901
85	85	X056	P.S. 056 Norwood Heights	341 EAST 207 STREET	10	10467	901
86	86	X056	PS 56 Norwood Heights		NA	NA	901
87	87	X057	P.S. 057 Crescent	2111 CROTONA AVENUE	12	10457	137
88	88	X058	P.S. 058	459 EAST 176 STREET	9	10457	132

Query executed successfully. LAPTOP-K2010PM0\SQLEXPRESS ... LAPTOP-K2010PM0\rchak... SchoolCrimeReport 00:00:00 230

Fig 11: Updating Foreign Key

- We further created a Reference table in the database to maintain the backup history and established the foreign keys using below Query

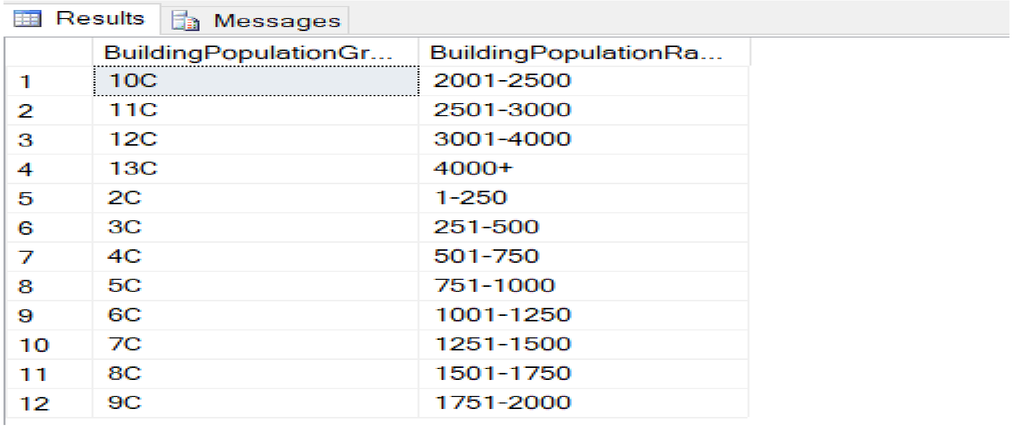


```
alter table AdministrativeDivision add constraint FK_AdministrativeDivision
FOREIGN KEY (BoroughCode) REFERENCES Boroughs(BoroughCode)
```

Fig 12: Adding Foreign Key Constraint

- After the database was linked with the foreign Keys associations we further tried finding out transitive dependencies to normalize further.
  In the attributes BuildingPopulation their Groups and range had many duplicated data .

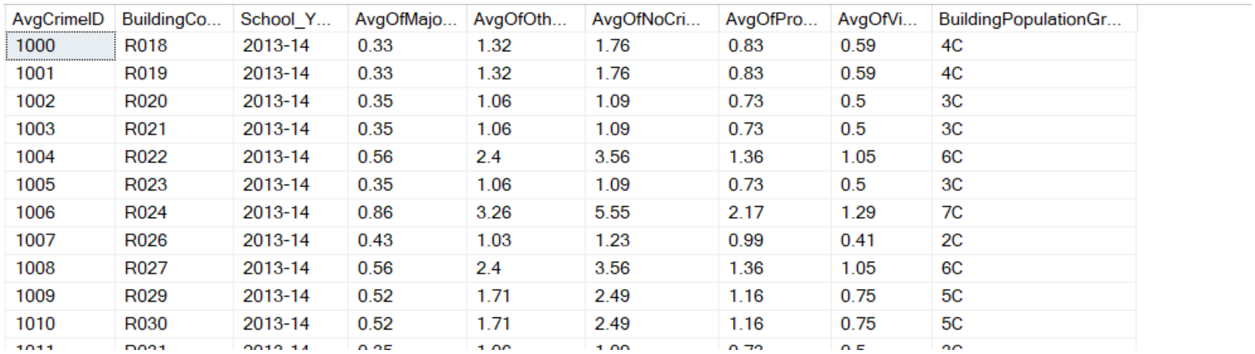
  We further normalized the data by creating separate group of BuildingPopulationGroup thus removing the transitive dependencies [7]



	BuildingPopulationGr...	BuildingPopulationRa...
1	10C	2001-2500
2	11C	2501-3000
3	12C	3001-4000
4	13C	4000+
5	2C	1-250
6	3C	251-500
7	4C	501-750
8	5C	751-1000
9	6C	1001-1250
10	7C	1251-1500
11	8C	1501-1750
12	9C	1751-2000

Fig 13 : Database After 3NF

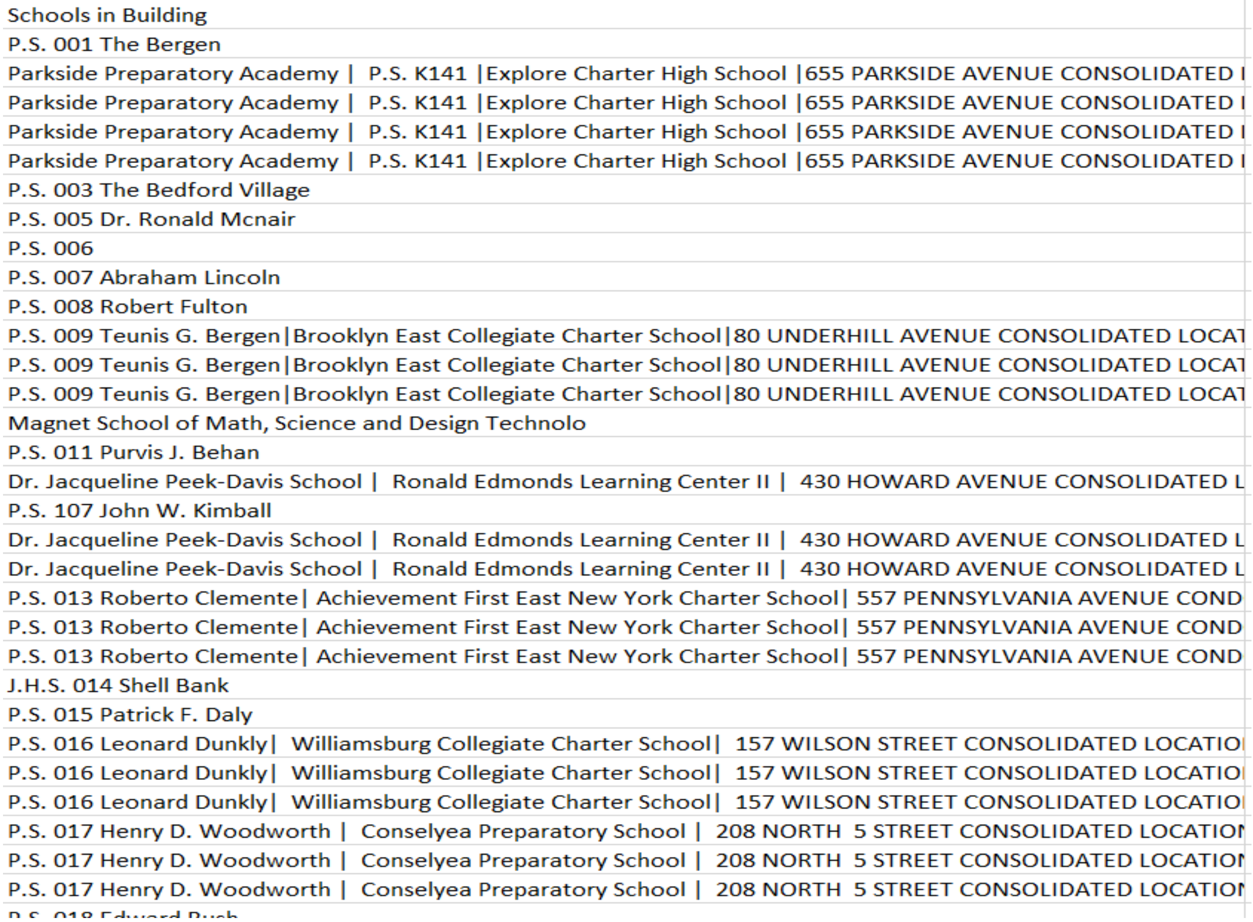
Further adding it as a foreign key Constraint thus removing the transitive dependencies. We further normalized all the tables analyzing and removing the transitive dependencies thus further to reduce the duplicate values columns and maintain data integrity to normalized form.



AvgCrimeID	BuildingCo...	School_Y...	AvgOfMajo...	AvgOfOth...	AvgOfNoCri...	AvgOfPro...	AvgOfVi...	BuildingPopulationGr...
1000	R018	2013-14	0.33	1.32	1.76	0.83	0.59	4C
1001	R019	2013-14	0.33	1.32	1.76	0.83	0.59	4C
1002	R020	2013-14	0.35	1.06	1.09	0.73	0.5	3C
1003	R021	2013-14	0.35	1.06	1.09	0.73	0.5	3C
1004	R022	2013-14	0.56	2.4	3.56	1.36	1.05	6C
1005	R023	2013-14	0.35	1.06	1.09	0.73	0.5	3C
1006	R024	2013-14	0.86	3.26	5.55	2.17	1.29	7C
1007	R026	2013-14	0.43	1.03	1.23	0.99	0.41	2C
1008	R027	2013-14	0.56	2.4	3.56	1.36	1.05	6C
1009	R029	2013-14	0.52	1.71	2.49	1.16	0.75	5C
1010	R030	2013-14	0.52	1.71	2.49	1.16	0.75	5C

Fig 14: Database After 3NF

- We further had many tables like Council, Location with transitive dependencies we further with observations created separate tables to remove deduplication.
- Our Database was in 3NF further we had one of the attribute Schools in Building with many attribute values. Thus, it was a multi-valued attribute in the Building Info table

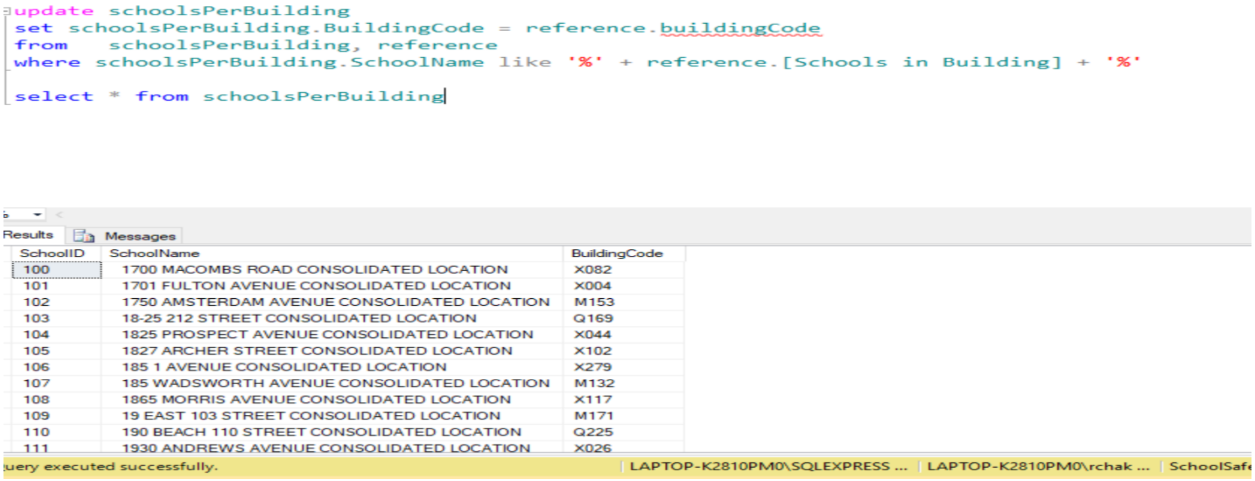


Schools in Building
P.S. 001 The Bergen
Parkside Preparatory Academy | P.S. K141 |Explore Charter High School |655 PARKSIDE AVENUE CONSOLIDATED I
Parkside Preparatory Academy | P.S. K141 |Explore Charter High School |655 PARKSIDE AVENUE CONSOLIDATED I
Parkside Preparatory Academy | P.S. K141 |Explore Charter High School |655 PARKSIDE AVENUE CONSOLIDATED I
Parkside Preparatory Academy | P.S. K141 |Explore Charter High School |655 PARKSIDE AVENUE CONSOLIDATED I
P.S. 003 The Bedford Village
P.S. 005 Dr. Ronald Mcnair
P.S. 006
P.S. 007 Abraham Lincoln
P.S. 008 Robert Fulton
P.S. 009 Teunis G. Bergen|Brooklyn East Collegiate Charter School|80 UNDERHILL AVENUE CONSOLIDATED LOCAT
P.S. 009 Teunis G. Bergen|Brooklyn East Collegiate Charter School|80 UNDERHILL AVENUE CONSOLIDATED LOCAT
P.S. 009 Teunis G. Bergen|Brooklyn East Collegiate Charter School|80 UNDERHILL AVENUE CONSOLIDATED LOCAT
Magnet School of Math, Science and Design Technolo
P.S. 011 Purvis J. Behan
Dr. Jacqueline Peek-Davis School | Ronald Edmonds Learning Center II | 430 HOWARD AVENUE CONSOLIDATED L
P.S. 107 John W. Kimball
Dr. Jacqueline Peek-Davis School | Ronald Edmonds Learning Center II | 430 HOWARD AVENUE CONSOLIDATED L
Dr. Jacqueline Peek-Davis School | Ronald Edmonds Learning Center II | 430 HOWARD AVENUE CONSOLIDATED L
P.S. 013 Roberto Clemente| Achievement First East New York Charter School| 557 PENNSYLVANIA AVENUE COND
P.S. 013 Roberto Clemente| Achievement First East New York Charter School| 557 PENNSYLVANIA AVENUE COND
P.S. 013 Roberto Clemente| Achievement First East New York Charter School| 557 PENNSYLVANIA AVENUE COND
J.H.S. 014 Shell Bank
P.S. 015 Patrick F. Daly
P.S. 016 Leonard Dunkly| Williamsburg Collegiate Charter School| 157 WILSON STREET CONSOLIDATED LOCATIO
P.S. 016 Leonard Dunkly| Williamsburg Collegiate Charter School| 157 WILSON STREET CONSOLIDATED LOCATIO
P.S. 016 Leonard Dunkly| Williamsburg Collegiate Charter School| 157 WILSON STREET CONSOLIDATED LOCATIO
P.S. 017 Henry D. Woodworth | Conselyea Preparatory School | 208 NORTH 5 STREET CONSOLIDATED LOCATION
P.S. 017 Henry D. Woodworth | Conselyea Preparatory School | 208 NORTH 5 STREET CONSOLIDATED LOCATION
P.S. 017 Henry D. Woodworth | Conselyea Preparatory School | 208 NORTH 5 STREET CONSOLIDATED LOCATION
P.S. 018 Edward Bush

Fig 15: Multivalued Attribute in DB

- We further created the Multivalued-attribute Different table as Schools Per Building with its own identification number as Primary Key. To convert into 4NF we further normalize removing multivalued attributes [7]
  Steps we followed to convert it into table:
  We first copy pasted the column in text files, then create separate spilt fields columns using DELIMETER as | further we then did stacking of the attribute values into one column multiple rows.
  Further deleting the duplicate Values and introducing Primary Key Associating with the foreign Key Building Code.

- Updating foreign keys from reference table using like search query:



```
update schoolsPerBuilding
set schoolsPerBuilding.BuildingCode = reference.buildingCode
from   schoolsPerBuilding, reference
where schoolsPerBuilding.SchoolName like '%' + reference.[Schools in Building] + '%'

select * from schoolsPerBuilding
```

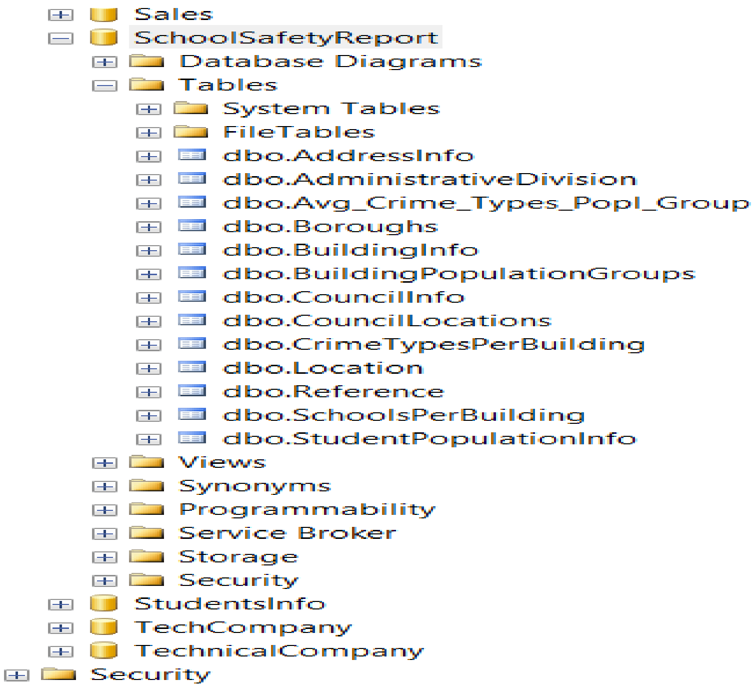
Results | Messages

SchoolID	SchoolName	BuildingCode
100	1700 MACOMBS ROAD CONSOLIDATED LOCATION	X082
101	1701 FULTON AVENUE CONSOLIDATED LOCATION	X004
102	1750 AMSTERDAM AVENUE CONSOLIDATED LOCATION	M153
103	18-25 212 STREET CONSOLIDATED LOCATION	Q169
104	1825 PROSPECT AVENUE CONSOLIDATED LOCATION	X044
105	1827 ARCHER STREET CONSOLIDATED LOCATION	X102
106	185 1 AVENUE CONSOLIDATED LOCATION	X279
107	185 WADSWORTH AVENUE CONSOLIDATED LOCATION	M132
108	1865 MORRIS AVENUE CONSOLIDATED LOCATION	X117
109	19 EAST 103 STREET CONSOLIDATED LOCATION	M171
110	190 BEACH 110 STREET CONSOLIDATED LOCATION	Q225
111	1930 ANDREWS AVENUE CONSOLIDATED LOCATION	X026

```
uery executed successfully.    | LAPTOP-K2810PM0\SQLEXPRESS ... | LAPTOP-K2810PM0\rchak ... | SchoolSafe
```

Fig 16: Updating Foreign Key

After completely normalizing the data to maintain its integrity we quickly structured our data in proper format to do further analysis for conclusions. We built our database with 12 tables linked to each other in one-to- many, many-to-many, one-to-one relation with reference the 13th table as the backup that is the original data.



```
⊞ Sales
⊟ SchoolSafetyReport
    ⊞ Database Diagrams
    ⊟ Tables
        ⊞ System Tables
        ⊞ FileTables
        ⊞ dbo.AddressInfo
        ⊞ dbo.AdministrativeDivision
        ⊞ dbo.Avg_Crime_Types_Popl_Group
        ⊞ dbo.Boroughs
        ⊞ dbo.BuildingInfo
        ⊞ dbo.BuildingPopulationGroups
        ⊞ dbo.CouncilInfo
        ⊞ dbo.CouncilLocations
        ⊞ dbo.CrimeTypesPerBuilding
        ⊞ dbo.Location
        ⊞ dbo.Reference
        ⊞ dbo.SchoolsPerBuilding
        ⊞ dbo.StudentPopulationInfo
    ⊞ Views
    ⊞ Synonyms
    ⊞ Programmability
    ⊞ Service Broker
    ⊞ Storage
    ⊞ Security
⊞ StudentsInfo
⊞ TechCompany
⊞ TechnicalCompany
⊞ Security
```

Fig 17: Final DB Schema

ENTITY-RELATIONSHIP DIAGRAM:

The following is our database ER Diagram for proper understanding of our structure: [5][6]

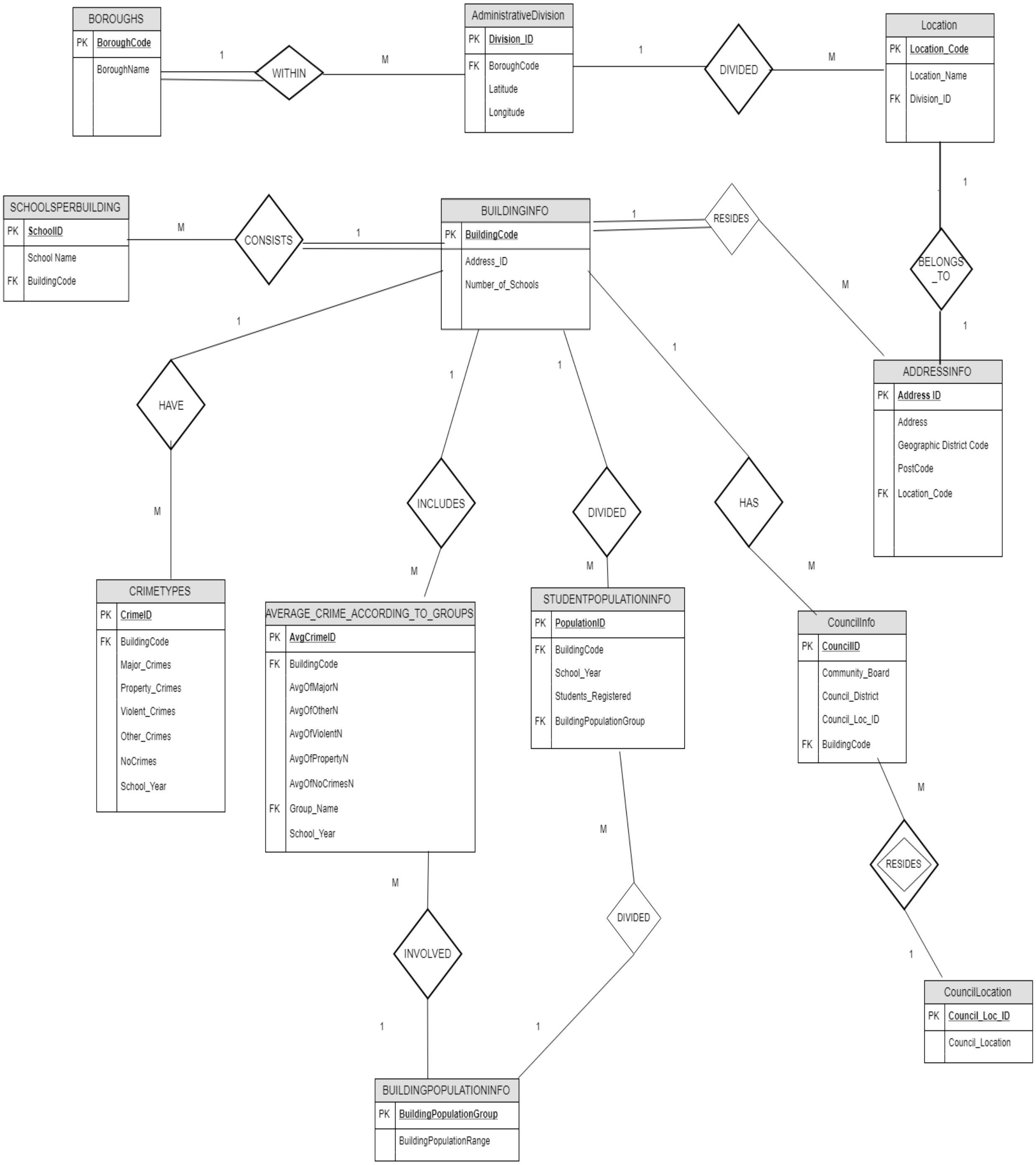
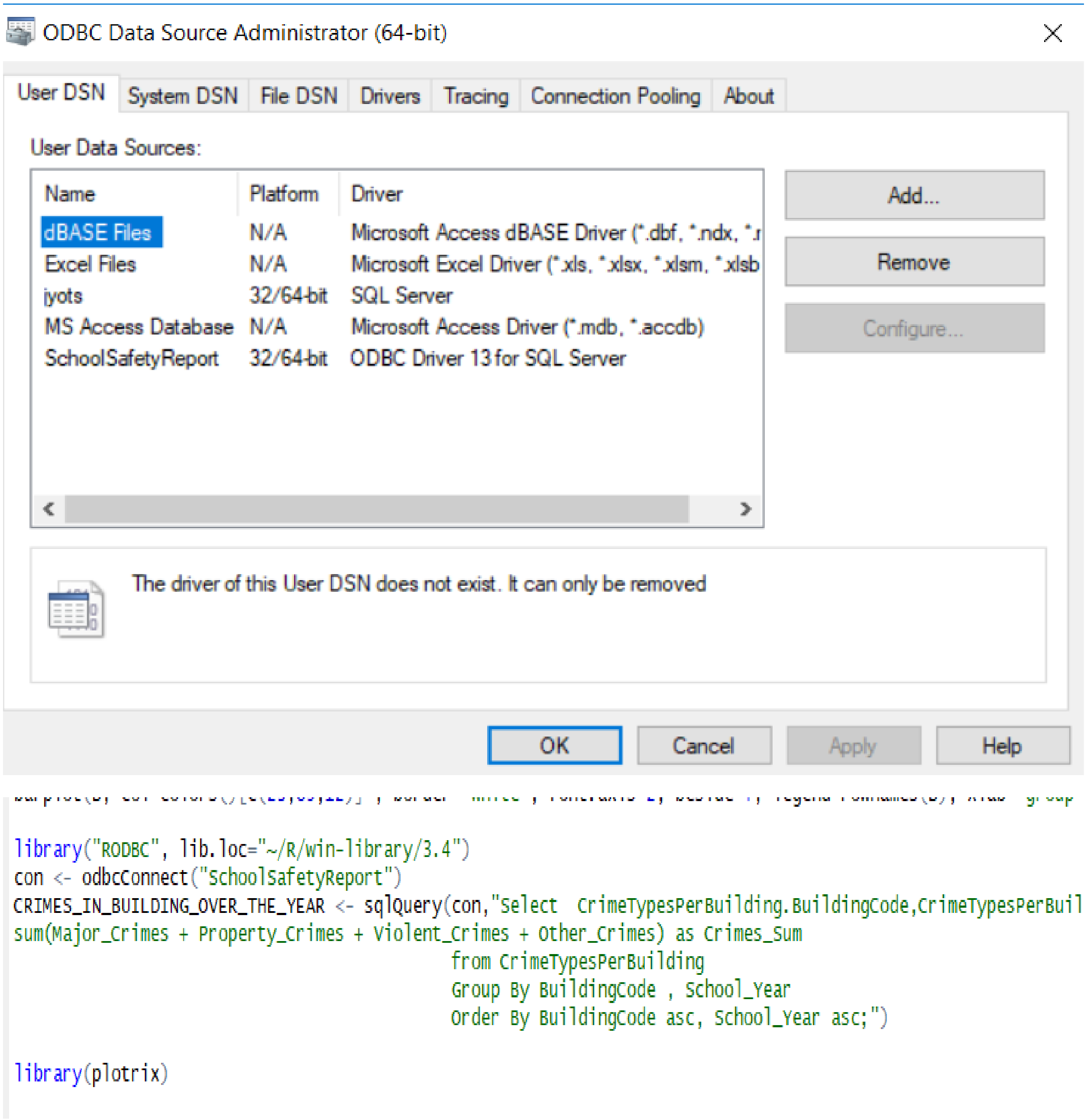


Fig 18: ER Diagram

Meaningful Analysis and Visualizations:

Analyzing the database, we came across various analysis supported with R studio visualization using the ODBC-driver connection. [8]



Name	Platform	Driver
dBASE Files	N/A	Microsoft Access dBASE Driver (*.dbf, *.ndx, *.r
Excel Files	N/A	Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm, *.xlsb
iyots	32/64-bit	SQL Server
MS Access Database	N/A	Microsoft Access Driver (*.mdb, *.accdb)
SchoolSafetyReport	32/64-bit	ODBC Driver 13 for SQL Server

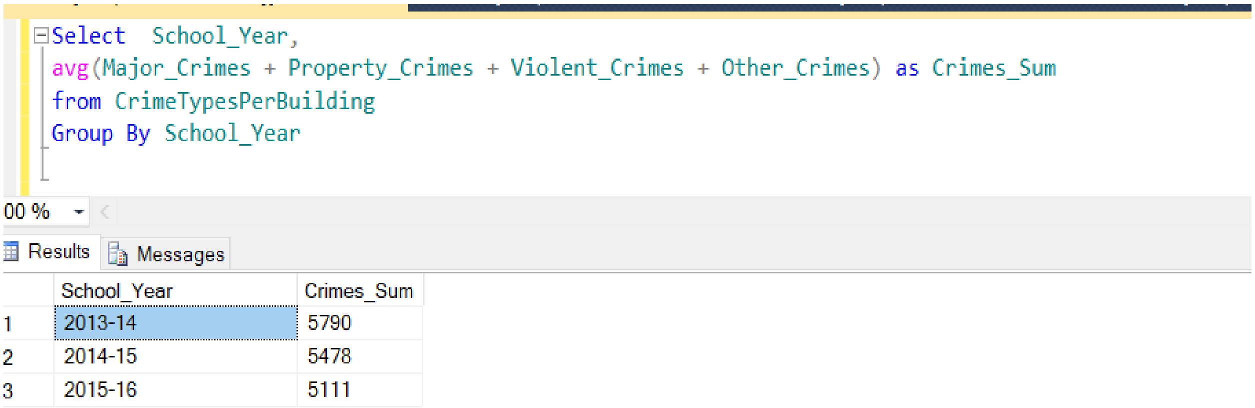
The driver of this User DSN does not exist. It can only be removed

```
barplot(B, col=colors()[c(23,89,12)] , border="white", font.axis=2, beside=T, legend=rownames(B), xlab="group

library("RODBC", lib.loc="~/R/win-library/3.4")
con <- odbcConnect("SchoolSafetyReport")
CRIMES_IN_BUILDING_OVER_THE_YEAR <- sqlQuery(con,"Select  CrimeTypesPerBuilding.BuildingCode,CrimeTypesPerBuil
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
                                        from CrimeTypesPerBuilding
                                        Group By BuildingCode , School_Year
                                        Order By BuildingCode asc, School_Year asc;")

library(plotrix)
```

Fig 19: ODBC Driver Connection

Goal: Analyzing the crime rate over the year

Query:

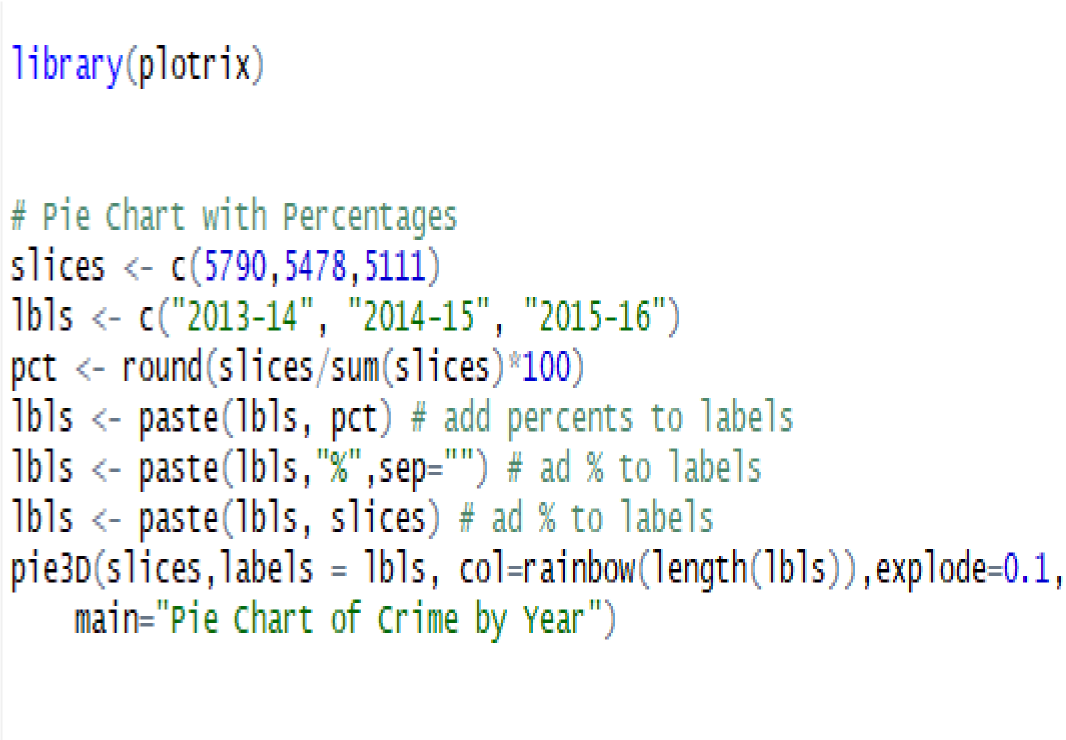


```
Select  School_Year,
avg(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding
Group By School_Year
```

	School_Year	Crimes_Sum
1	2013-14	5790
2	2014-15	5478
3	2015-16	5111

According to the analysis we see that as the years passed the number of crimes rate reduced slightly over all. Further visualized using R Studio Pie-Chart [2]

R-Code:



```
library(plotrix)


# Pie Chart with Percentages
slices <- c(5790,5478,5111)
lbls <- c("2013-14", "2014-15", "2015-16")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
lbls <- paste(lbls, slices) # ad % to labels
pie3D(slices,labels = lbls, col=rainbow(length(lbls)),explode=0.1,
    main="Pie Chart of Crime by Year")
```

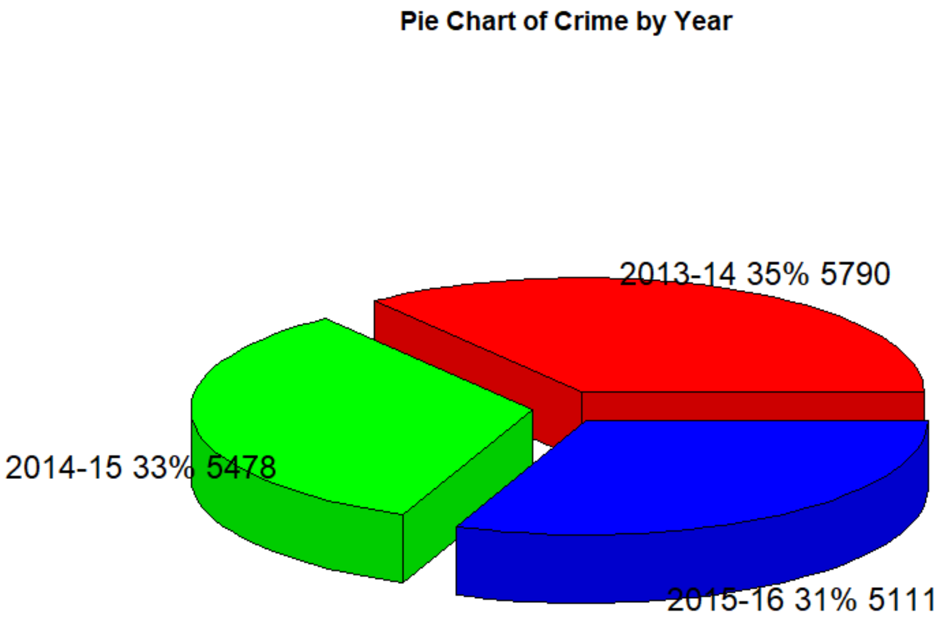
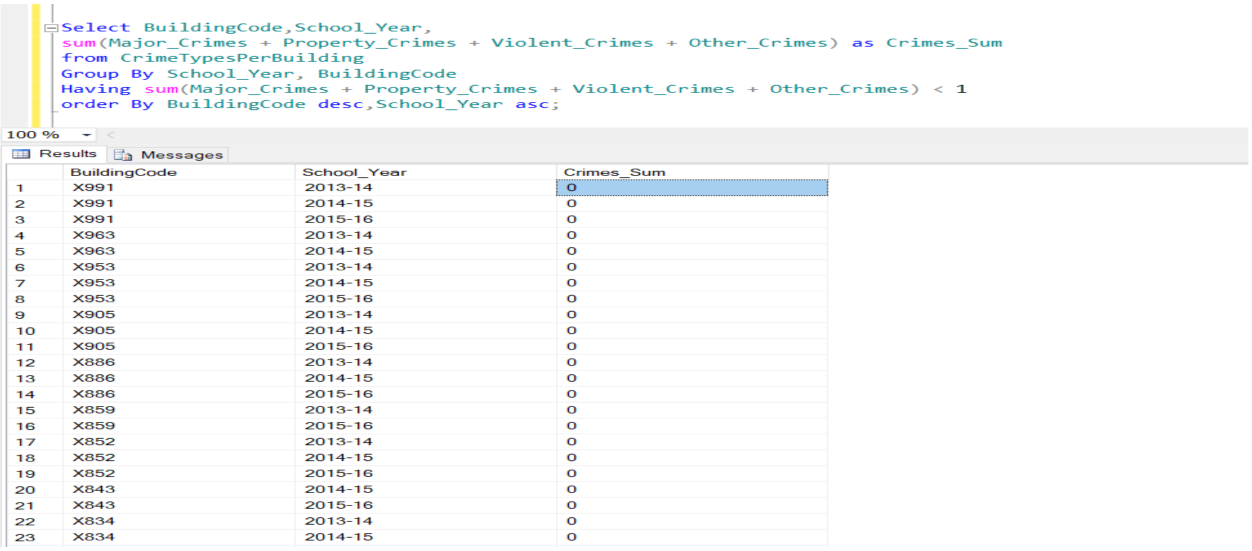


Fig 20: PIE Chart of Crime by Year

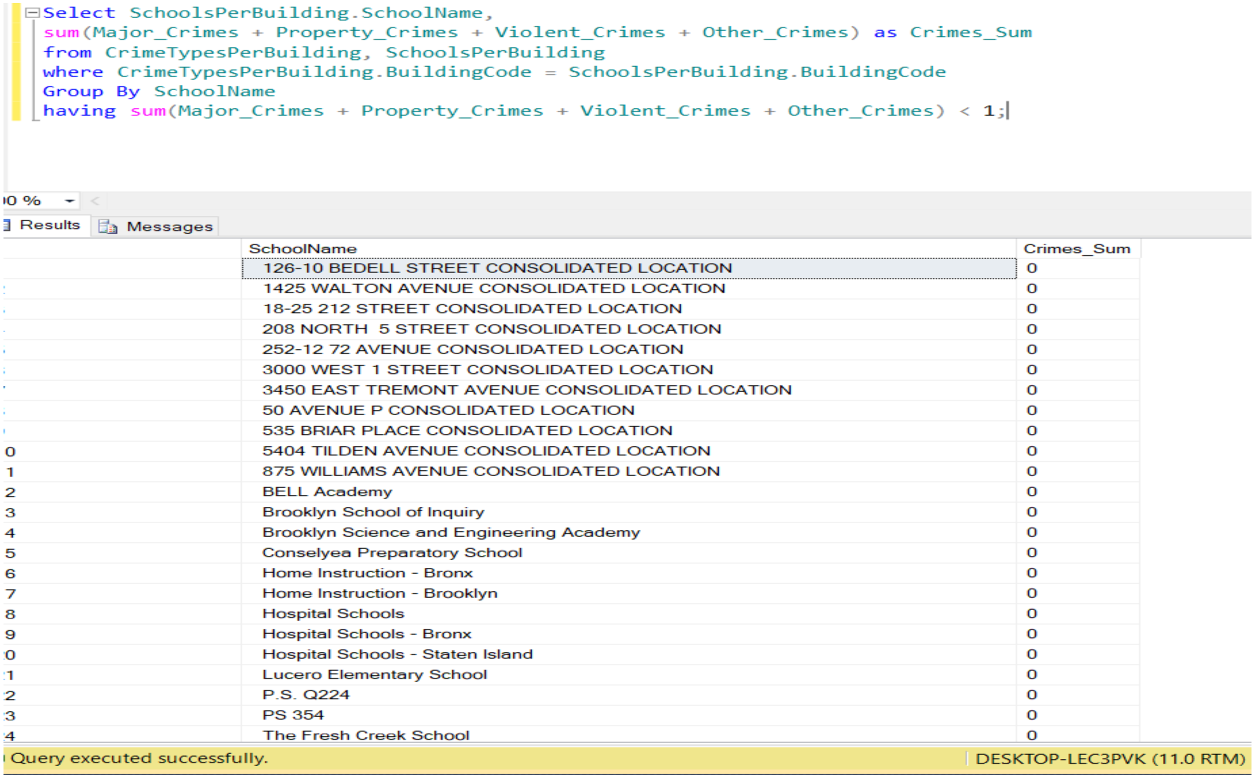
Goal: Finding the Building Code and School Names which are safest to attend which have zero Crimes on the campus.

Query:



```
Select BuildingCode,School_Year,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding
Group By School_Year, BuildingCode
Having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) < 1
order By BuildingCode desc,School_Year asc;
```

	BuildingCode	School_Year	Crimes_Sum
1	X991	2013-14	0
2	X991	2014-15	0
3	X991	2015-16	0
4	X963	2013-14	0
5	X963	2014-15	0
6	X953	2013-14	0
7	X953	2014-15	0
8	X953	2015-16	0
9	X905	2013-14	0
10	X905	2014-15	0
11	X905	2015-16	0
12	X886	2013-14	0
13	X886	2014-15	0
14	X886	2015-16	0
15	X859	2013-14	0
16	X859	2015-16	0
17	X852	2013-14	0
18	X852	2014-15	0
19	X852	2015-16	0
20	X843	2014-15	0
21	X843	2015-16	0
22	X834	2013-14	0
23	X834	2014-15	0



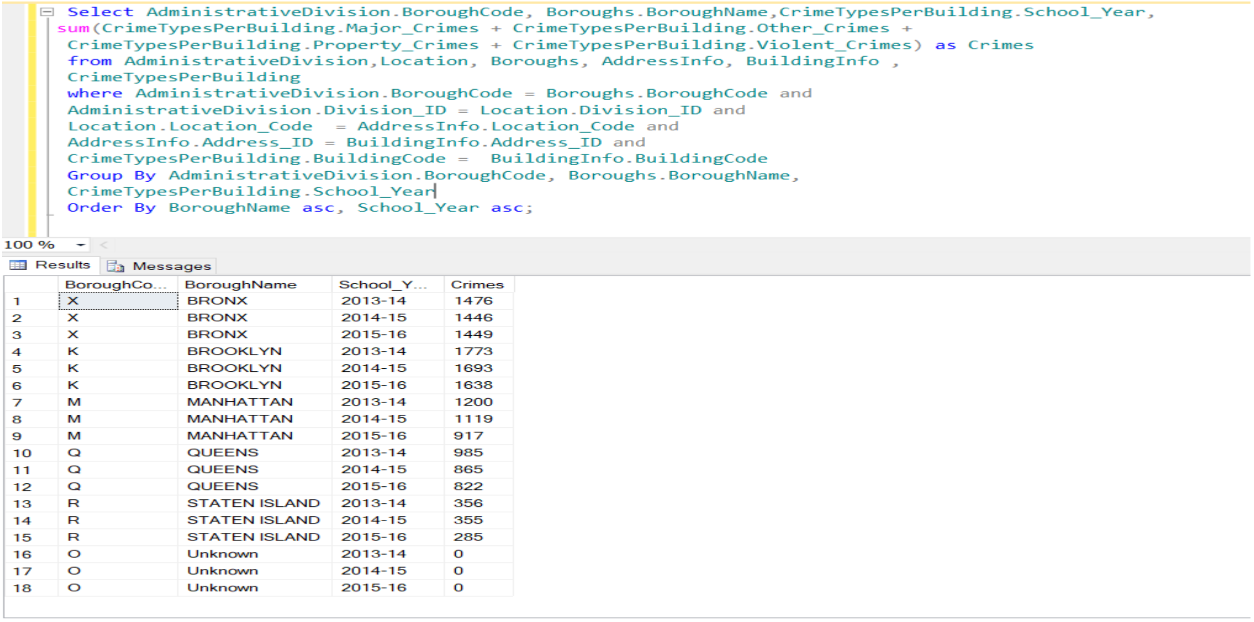
```
Select SchoolsPerBuilding.SchoolName,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding, SchoolsPerBuilding
where CrimeTypesPerBuilding.BuildingCode = SchoolsPerBuilding.BuildingCode
Group By SchoolName
having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) < 1;
```

	SchoolName	Crimes_Sum
	126-10 BEDELL STREET CONSOLIDATED LOCATION	0
	1425 WALTON AVENUE CONSOLIDATED LOCATION	0
	18-25 212 STREET CONSOLIDATED LOCATION	0
	208 NORTH 5 STREET CONSOLIDATED LOCATION	0
	252-12 72 AVENUE CONSOLIDATED LOCATION	0
	3000 WEST 1 STREET CONSOLIDATED LOCATION	0
	3450 EAST TREMONT AVENUE CONSOLIDATED LOCATION	0
	50 AVENUE P CONSOLIDATED LOCATION	0
	535 BRIAR PLACE CONSOLIDATED LOCATION	0
0	5404 TILDEN AVENUE CONSOLIDATED LOCATION	0
1	875 WILLIAMS AVENUE CONSOLIDATED LOCATION	0
2	BELL Academy	0
3	Brooklyn School of Inquiry	0
4	Brooklyn Science and Engineering Academy	0
5	Conselyea Preparatory School	0
6	Home Instruction - Bronx	0
7	Home Instruction - Brooklyn	0
8	Hospital Schools	0
9	Hospital Schools - Bronx	0
0	Hospital Schools - Staten Island	0
1	Lucero Elementary School	0
2	P.S. Q224	0
3	PS 354	0
4	The Fresh Creek School	0

Query executed successfully. DESKTOP-LEC3PVK (11.0 RTM)

Goal: Analyzing the crimes rates across various Boroughs of New York according to the Years.

Query:



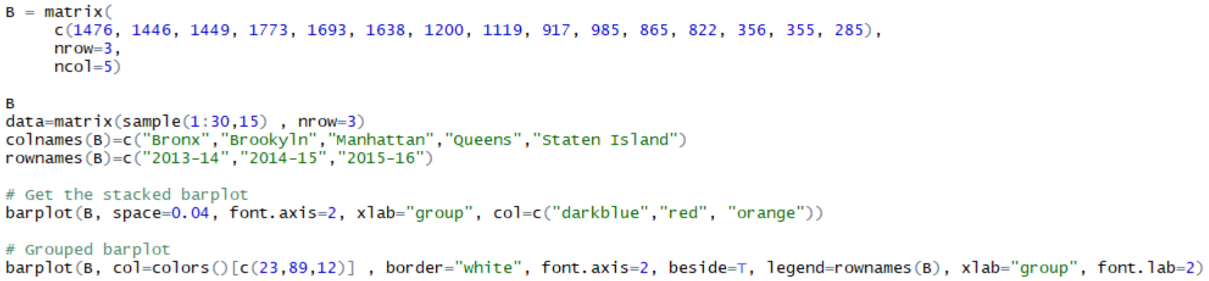
```
Select AdministrativeDivision.BoroughCode, Boroughs.BoroughName,CrimeTypesPerBuilding.School_Year,
sum(CrimeTypesPerBuilding.Major_Crimes + CrimeTypesPerBuilding.Other_Crimes +
CrimeTypesPerBuilding.Property_Crimes + CrimeTypesPerBuilding.Violent_Crimes) as Crimes
from AdministrativeDivision,Location, Boroughs, AddressInfo, BuildingInfo ,
CrimeTypesPerBuilding
where AdministrativeDivision.BoroughCode = Boroughs.BoroughCode and
AdministrativeDivision.Division_ID = Location.Division_ID and
Location.Location_Code  = AddressInfo.Location_Code and
AddressInfo.Address_ID = BuildingInfo.Address_ID and
CrimeTypesPerBuilding.BuildingCode =  BuildingInfo.BuildingCode
Group By AdministrativeDivision.BoroughCode, Boroughs.BoroughName,
CrimeTypesPerBuilding.School_Year
Order By BoroughName asc, School_Year asc;
```

	BoroughCo...	BoroughName	School_Y...	Crimes
1	X	BRONX	2013-14	1476
2	X	BRONX	2014-15	1446
3	X	BRONX	2015-16	1449
4	K	BROOKLYN	2013-14	1773
5	K	BROOKLYN	2014-15	1693
6	K	BROOKLYN	2015-16	1638
7	M	MANHATTAN	2013-14	1200
8	M	MANHATTAN	2014-15	1119
9	M	MANHATTAN	2015-16	917
10	Q	QUEENS	2013-14	985
11	Q	QUEENS	2014-15	865
12	Q	QUEENS	2015-16	822
13	R	STATEN ISLAND	2013-14	356
14	R	STATEN ISLAND	2014-15	355
15	R	STATEN ISLAND	2015-16	285
16	O	Unknown	2013-14	0
17	O	Unknown	2014-15	0
18	O	Unknown	2015-16	0

Analyzing the result, we find that Brooklyn Schools have highest number of crime rates followed by Bronx, then Manhattan, Queens with Staten Island having the lowest number of crime rates.

The crime rates have decreased over years this indicates that the Council are taking measures towards reducing the crime rates in the Schools of New York. Visualization done with the histogram to support analysis. [4]

R-Code:



```
B = matrix(
     c(1476, 1446, 1449, 1773, 1693, 1638, 1200, 1119, 917, 985, 865, 822, 356, 355, 285),
     nrow=3,
     ncol=5)

B
data=matrix(sample(1:30,15) , nrow=3)
colnames(B)=c("Bronx","Brookyln","Manhattan","Queens","Staten Island")
rownames(B)=c("2013-14","2014-15","2015-16")

# Get the stacked barplot
barplot(B, space=0.04, font.axis=2, xlab="group", col=c("darkblue","red", "orange"))

# Grouped barplot
barplot(B, col=colors()[c(23,89,12)] , border="white", font.axis=2, beside=T, legend=rownames(B), xlab="group", font.lab=2)
```

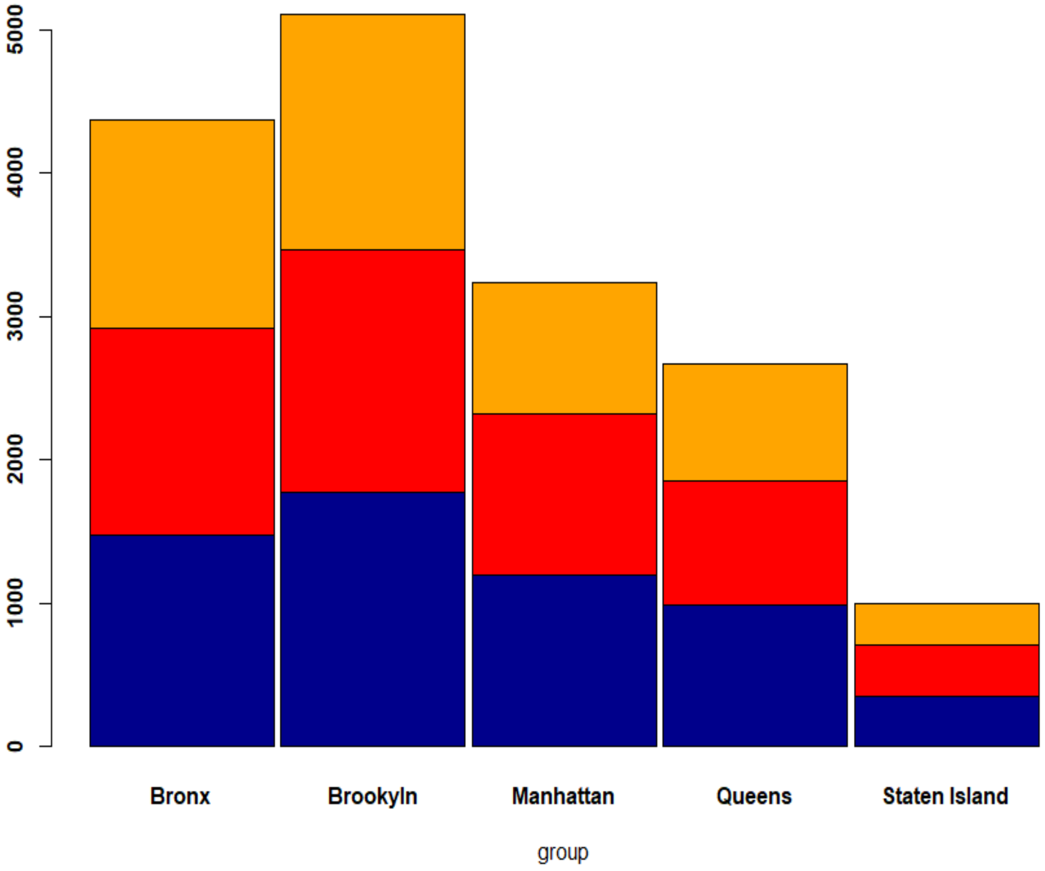
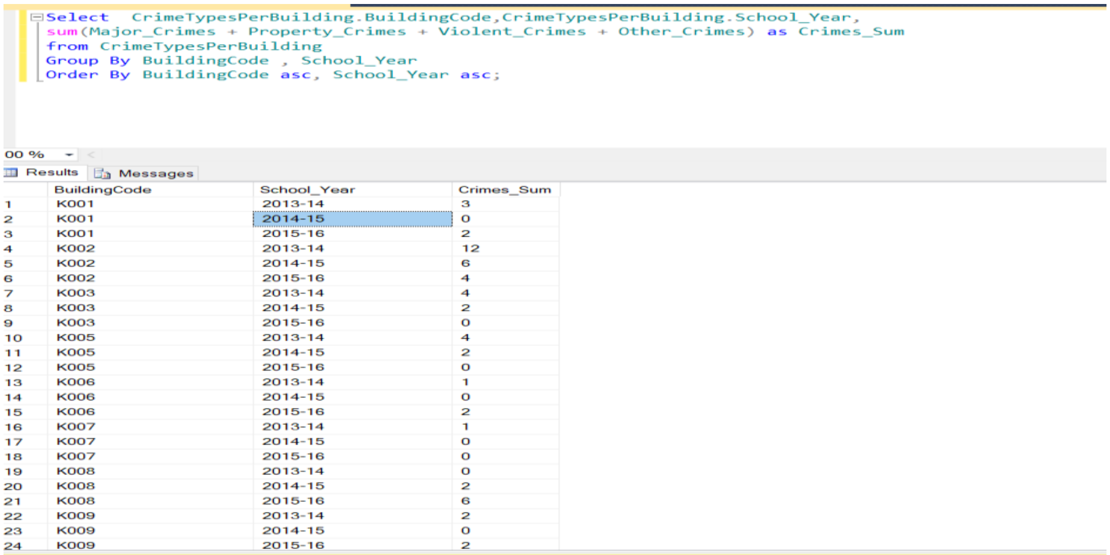


Fig 21 : Crime By Areas

Goal: Analyzing various Buildings and Schools about the crime rates increased or decreased over the years.

Query:



```
Select  CrimeTypesPerBuilding.BuildingCode,CrimeTypesPerBuilding.School_Year,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding
Group By BuildingCode , School_Year
Order By BuildingCode asc, School_Year asc;
```

	BuildingCode	School_Year	Crimes_Sum
1	K001	2013-14	3
2	K001	2014-15	0
3	K001	2015-16	2
4	K002	2013-14	12
5	K002	2014-15	6
6	K002	2015-16	4
7	K003	2013-14	4
8	K003	2014-15	2
9	K003	2015-16	0
10	K005	2013-14	4
11	K005	2014-15	2
12	K005	2015-16	0
13	K006	2013-14	1
14	K006	2014-15	0
15	K006	2015-16	2
16	K007	2013-14	1
17	K007	2014-15	0
18	K007	2015-16	0
19	K008	2013-14	0
20	K008	2014-15	2
21	K008	2015-16	6
22	K009	2013-14	2
23	K009	2014-15	0
24	K009	2015-16	2

Analyzing the result, we find some building schools crimes rates have decreased drastically (Building Code K142) whereas the crime rates at sum places increased sum schools over the time (Building Code K013) sum schools it went from decrease to no crimes to high crimes (Building Code K006). Visualizing over thousand rows is difficult therefore we visualized set of data frames to support results. [4]

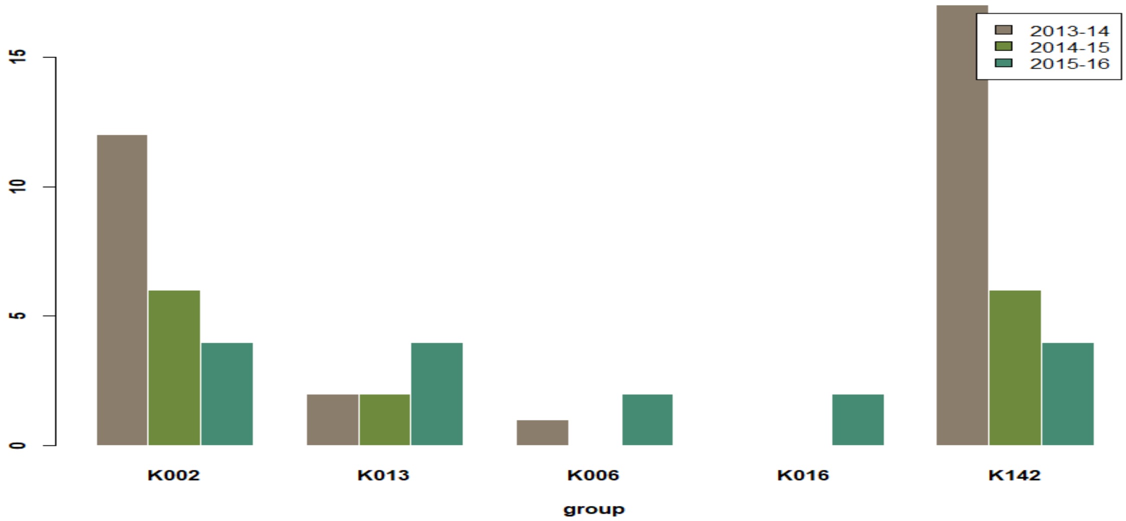
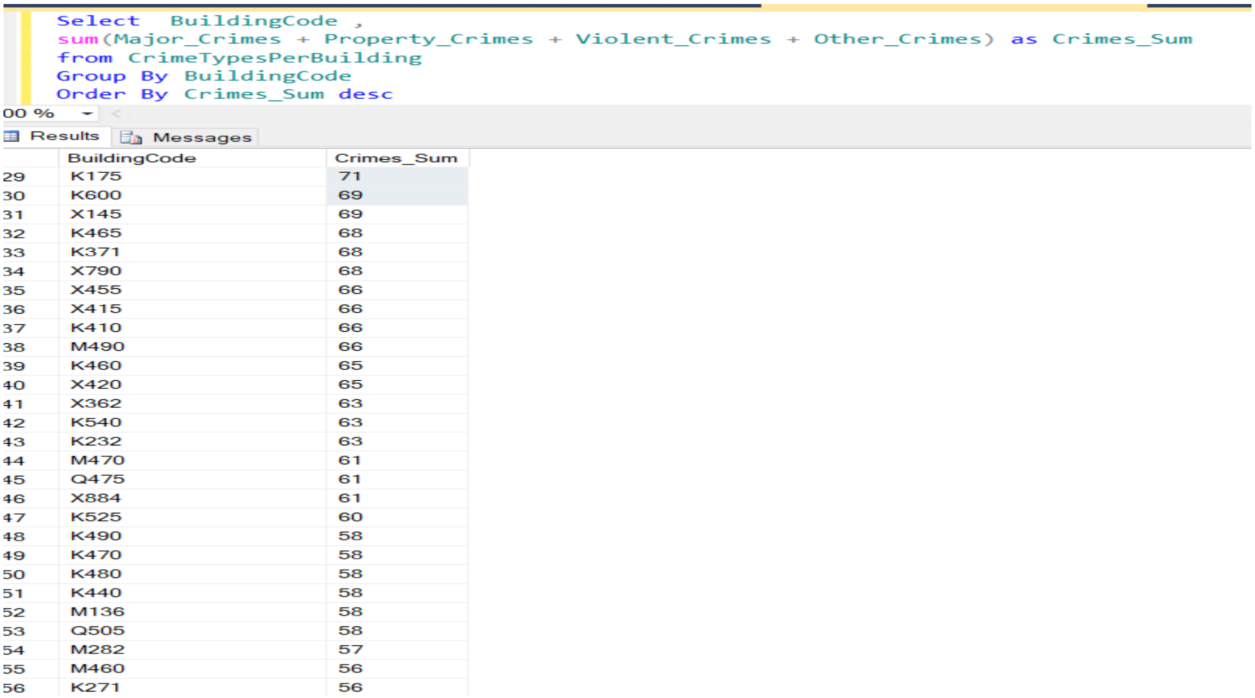


Fig 22: Crime Rate by Building

Goal: Analyzing the highest number of crimes over time across the Building Code

Query:

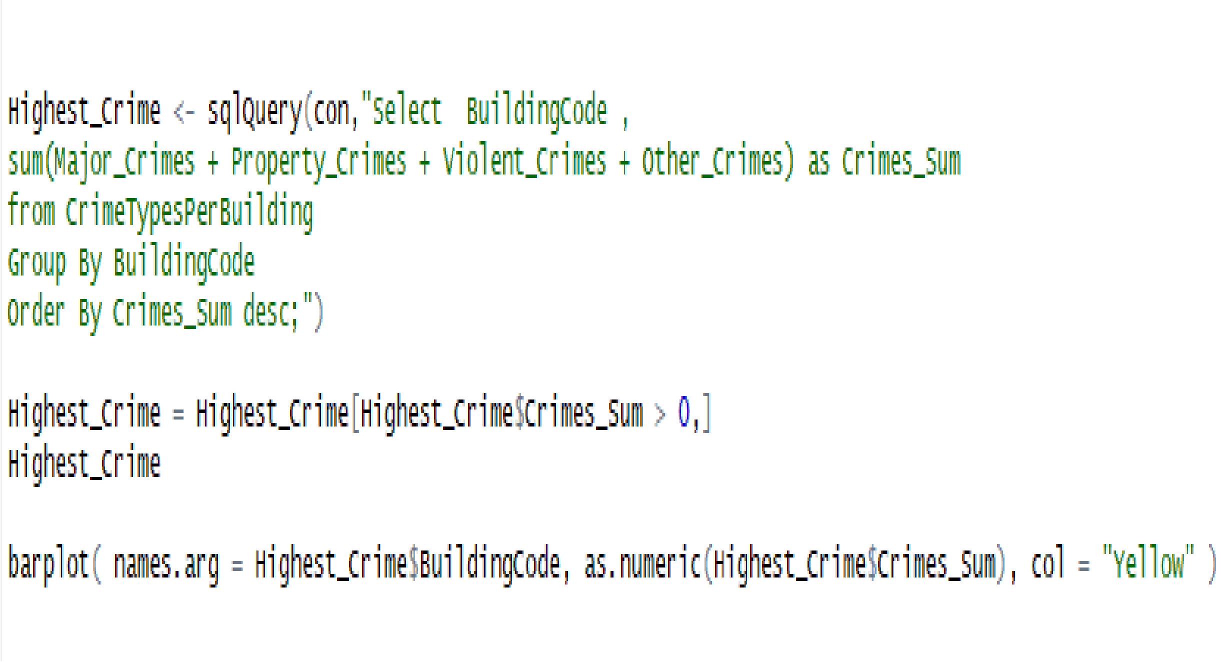


```
Select  BuildingCode ,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding
Group By BuildingCode
Order By Crimes_Sum desc
```

	BuildingCode	Crimes_Sum
29	K175	71
30	K600	69
31	X145	69
32	K465	68
33	K371	68
34	X790	68
35	X455	66
36	X415	66
37	K410	66
38	M490	66
39	K460	65
40	X420	65
41	X362	63
42	K540	63
43	K232	63
44	M470	61
45	Q475	61
46	X884	61
47	K525	60
48	K490	58
49	K470	58
50	K480	58
51	K440	58
52	M136	58
53	Q505	58
54	M282	57
55	M460	56
56	K271	56

We find that the number of crimes according to building Code that are not safe for the students to attend across the decreasing rate visualized histogram. [2]

R-Code:



```
Highest_Crime <- sqlQuery(con,"Select  BuildingCode ,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding
Group By BuildingCode
Order By Crimes_Sum desc;")

Highest_Crime = Highest_Crime[Highest_Crime$Crimes_Sum > 0,]
Highest_Crime

barplot( names.arg = Highest_Crime$BuildingCode, as.numeric(Highest_Crime$Crimes_Sum), col = "Yellow" )
```

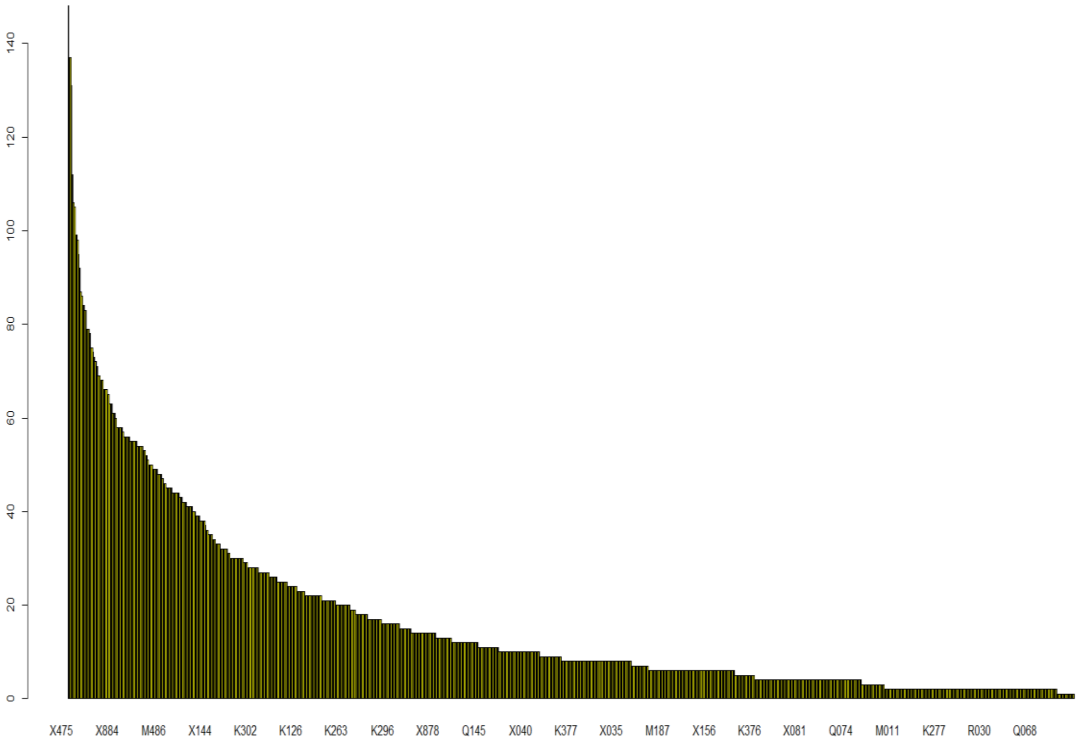
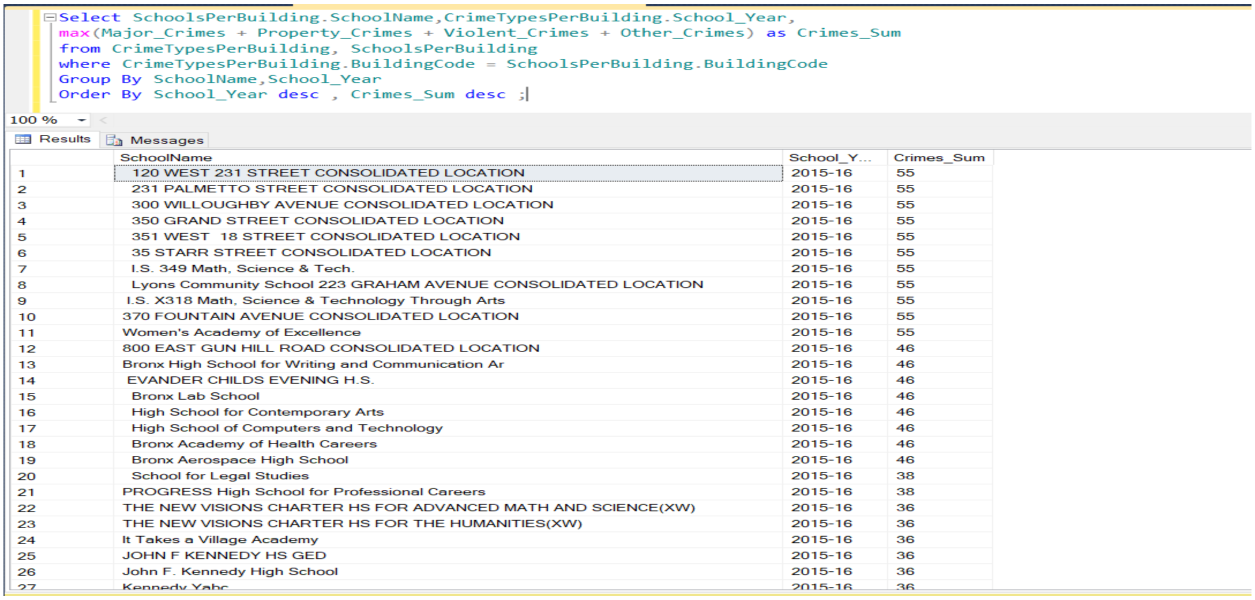


Fig 23: Histogram of Crime Rate From Highest to Lowest by Building Code

Goal: Finding the names of Schools having highest crime rates based on our previous visualization.

Query:

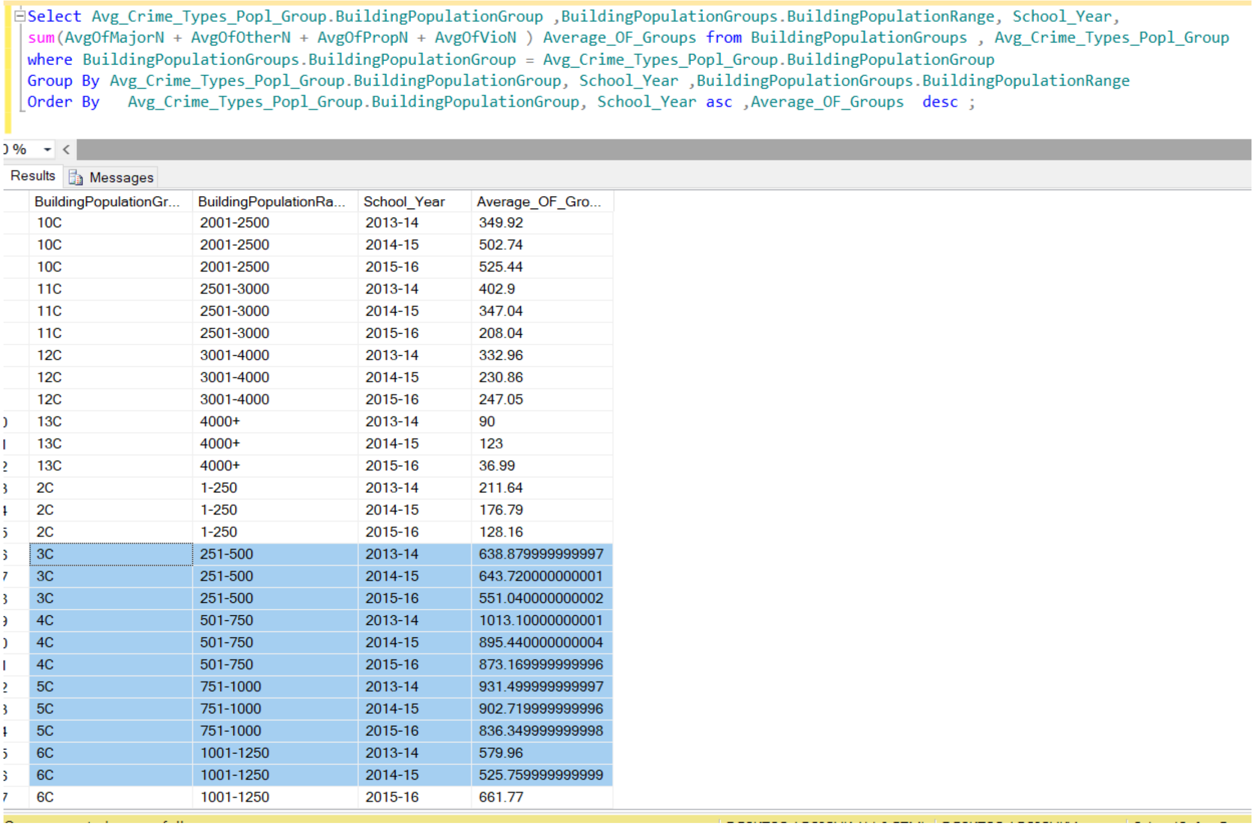


```
Select SchoolsPerBuilding.SchoolName,CrimeTypesPerBuilding.School_Year,
max(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding, SchoolsPerBuilding
where CrimeTypesPerBuilding.BuildingCode = SchoolsPerBuilding.BuildingCode
Group By SchoolName,School_Year
Order By School_Year desc , Crimes_Sum desc ;
```

	SchoolName	School_Y...	Crimes_Sum
1	120 WEST 231 STREET CONSOLIDATED LOCATION	2015-16	55
2	231 PALMETTO STREET CONSOLIDATED LOCATION	2015-16	55
3	300 WILLOUGHBY AVENUE CONSOLIDATED LOCATION	2015-16	55
4	350 GRAND STREET CONSOLIDATED LOCATION	2015-16	55
5	351 WEST 18 STREET CONSOLIDATED LOCATION	2015-16	55
6	35 STARR STREET CONSOLIDATED LOCATION	2015-16	55
7	I.S. 349 Math, Science & Tech.	2015-16	55
8	Lyons Community School 223 GRAHAM AVENUE CONSOLIDATED LOCATION	2015-16	55
9	I.S. X318 Math, Science & Technology Through Arts	2015-16	55
10	370 FOUNTAIN AVENUE CONSOLIDATED LOCATION	2015-16	55
11	Women's Academy of Excellence	2015-16	55
12	800 EAST GUN HILL ROAD CONSOLIDATED LOCATION	2015-16	46
13	Bronx High School for Writing and Communication Ar	2015-16	46
14	EVANDER CHILDS EVENING H.S.	2015-16	46
15	Bronx Lab School	2015-16	46
16	High School for Contemporary Arts	2015-16	46
17	High School of Computers and Technology	2015-16	46
18	Bronx Academy of Health Careers	2015-16	46
19	Bronx Aerospace High School	2015-16	46
20	School for Legal Studies	2015-16	38
21	PROGRESS High School for Professional Careers	2015-16	38
22	THE NEW VISIONS CHARTER HS FOR ADVANCED MATH AND SCIENCE(XW)	2015-16	36
23	THE NEW VISIONS CHARTER HS FOR THE HUMANITIES(XW)	2015-16	36
24	It Takes a Village Academy	2015-16	36
25	JOHN F KENNEDY HS GED	2015-16	36
26	John F. Kennedy High School	2015-16	36
27	Kennedy Yabc	2015-16	36

Goal: Analyzing the crimes according to the groups and between the range of students registered across years.

Query:

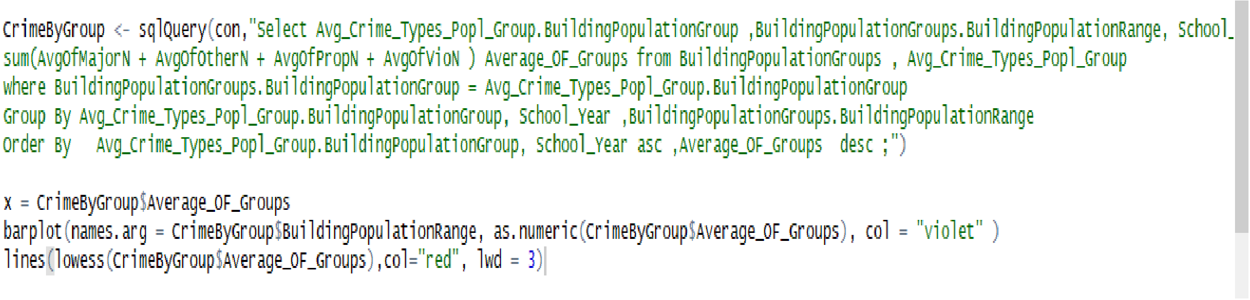


```
Select Avg_Crime_Types_Popl_Group.BuildingPopulationGroup ,BuildingPopulationGroups.BuildingPopulationRange, School_Year,
sum(AvgOfMajorN + AvgOfOtherN + AvgOfPropN + AvgOfVioN ) Average_OF_Groups from BuildingPopulationGroups , Avg_Crime_Types_Popl_Group
where BuildingPopulationGroups.BuildingPopulationGroup = Avg_Crime_Types_Popl_Group.BuildingPopulationGroup
Group By Avg_Crime_Types_Popl_Group.BuildingPopulationGroup, School_Year ,BuildingPopulationGroups.BuildingPopulationRange
Order By   Avg_Crime_Types_Popl_Group.BuildingPopulationGroup, School_Year asc ,Average_OF_Groups  desc ;
```

BuildingPopulationGr...	BuildingPopulationRa...	School_Year	Average_OF_Gro...
10C	2001-2500	2013-14	349.92
10C	2001-2500	2014-15	502.74
10C	2001-2500	2015-16	525.44
11C	2501-3000	2013-14	402.9
11C	2501-3000	2014-15	347.04
11C	2501-3000	2015-16	208.04
12C	3001-4000	2013-14	332.96
12C	3001-4000	2014-15	230.86
12C	3001-4000	2015-16	247.05
13C	4000+	2013-14	90
13C	4000+	2014-15	123
13C	4000+	2015-16	36.99
2C	1-250	2013-14	211.64
2C	1-250	2014-15	176.79
2C	1-250	2015-16	128.16
3C	251-500	2013-14	638.879999999997
3C	251-500	2014-15	643.720000000001
3C	251-500	2015-16	551.040000000002
4C	501-750	2013-14	1013.10000000001
4C	501-750	2014-15	895.440000000004
4C	501-750	2015-16	873.169999999996
5C	751-1000	2013-14	931.499999999997
5C	751-1000	2014-15	902.719999999996
5C	751-1000	2015-16	836.349999999998
6C	1001-1250	2013-14	579.96
6C	1001-1250	2014-15	525.759999999999
6C	1001-1250	2015-16	661.77

Analyzing the result, we find that the group 4C range 501-750 number of students are the highest crime rates occurring with groups like 3C, 4C, 5C range over group 13C with range 4000 group having lowest rate followed by 9C we thus come to conclusion in the range of students 1- 4000 range around 250-1500 have the highest number of crimes. Moreover, in some groups like 13C, 12C crimes have decreased over year, and group 6C it is increased over years. More attention should be given groups where crime rates increase. We conclude that where there more students the crime rates are less. [4]

R-Code:



```
CrimeByGroup <- sqlQuery(con,"Select Avg_Crime_Types_Popl_Group.BuildingPopulationGroup ,BuildingPopulationGroups.BuildingPopulationRange, School_
sum(AvgOfMajorN + AvgOfOtherN + AvgOfPropN + AvgOfVioN ) Average_OF_Groups from BuildingPopulationGroups , Avg_Crime_Types_Popl_Group
where BuildingPopulationGroups.BuildingPopulationGroup = Avg_Crime_Types_Popl_Group.BuildingPopulationGroup
Group By Avg_Crime_Types_Popl_Group.BuildingPopulationGroup, School_Year ,BuildingPopulationGroups.BuildingPopulationRange
Order By   Avg_Crime_Types_Popl_Group.BuildingPopulationGroup, School_Year asc ,Average_OF_Groups  desc ;")

x = CrimeByGroup$Average_OF_Groups
barplot(names.arg = CrimeByGroup$BuildingPopulationRange, as.numeric(CrimeByGroup$Average_OF_Groups), col = "violet" )
lines(lowess(CrimeByGroup$Average_OF_Groups),col="red", lwd = 3)
```

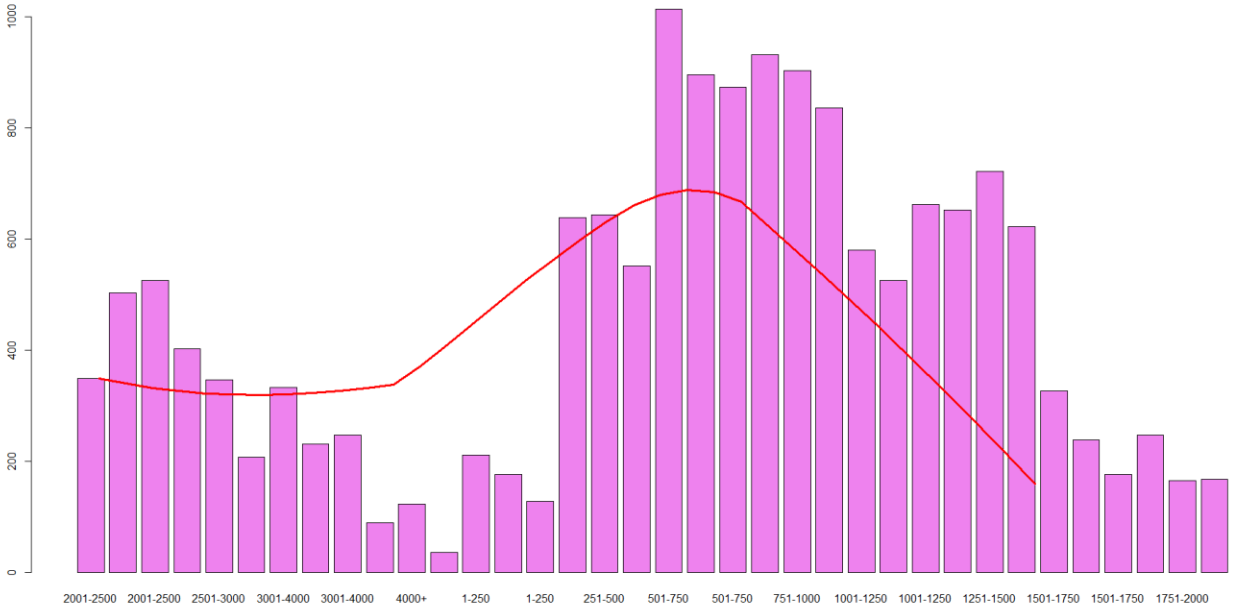
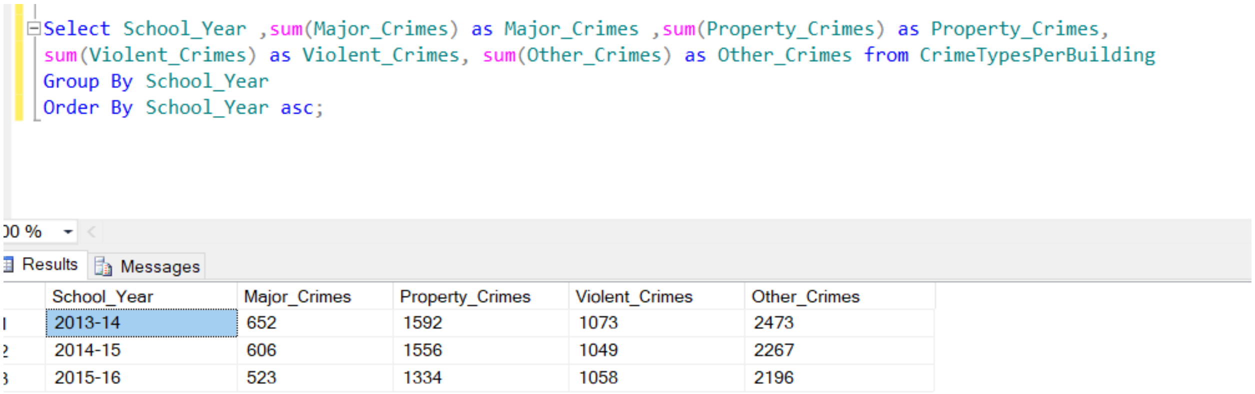


Fig 24: Crime Rate by population Range

Goal: Analyzing different types of Crimes over years

Query:

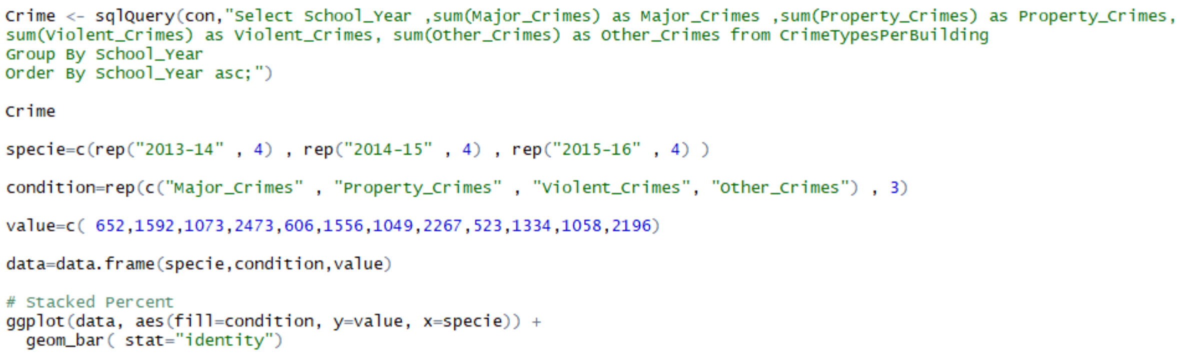


```
Select School_Year ,sum(Major_Crimes) as Major_Crimes ,sum(Property_Crimes) as Property_Crimes,
sum(Violent_Crimes) as Violent_Crimes, sum(Other_Crimes) as Other_Crimes from CrimeTypesPerBuilding
Group By School_Year
Order By School_Year asc;
```

	School_Year	Major_Crimes	Property_Crimes	Violent_Crimes	Other_Crimes
1	2013-14	652	1592	1073	2473
2	2014-15	606	1556	1049	2267
3	2015-16	523	1334	1058	2196

Analyzing the results, we conclude the crimes are decreased over years with other crimes being the highest followed by the Property Crimes with Major Crimes being the lowest which is a good sign that the schools have no major mishaps. Visualized using Mosaic Plot. [3]

R-Code:



```
Crime <- sqlQuery(con,"Select School_Year ,sum(Major_Crimes) as Major_Crimes ,sum(Property_Crimes) as Property_Crimes,
sum(Violent_Crimes) as Violent_Crimes, sum(Other_Crimes) as Other_Crimes from CrimeTypesPerBuilding
Group By School_Year
Order By School_Year asc;")

Crime

specie=c(rep("2013-14" , 4) , rep("2014-15" , 4) , rep("2015-16" , 4) )

condition=rep(c("Major_Crimes" , "Property_Crimes" , "Violent_Crimes", "Other_Crimes") , 3)

value=c( 652,1592,1073,2473,606,1556,1049,2267,523,1334,1058,2196)

data=data.frame(specie,condition,value)

# Stacked Percent
ggplot(data, aes(fill=condition, y=value, x=specie)) +
  geom_bar( stat="identity")
```

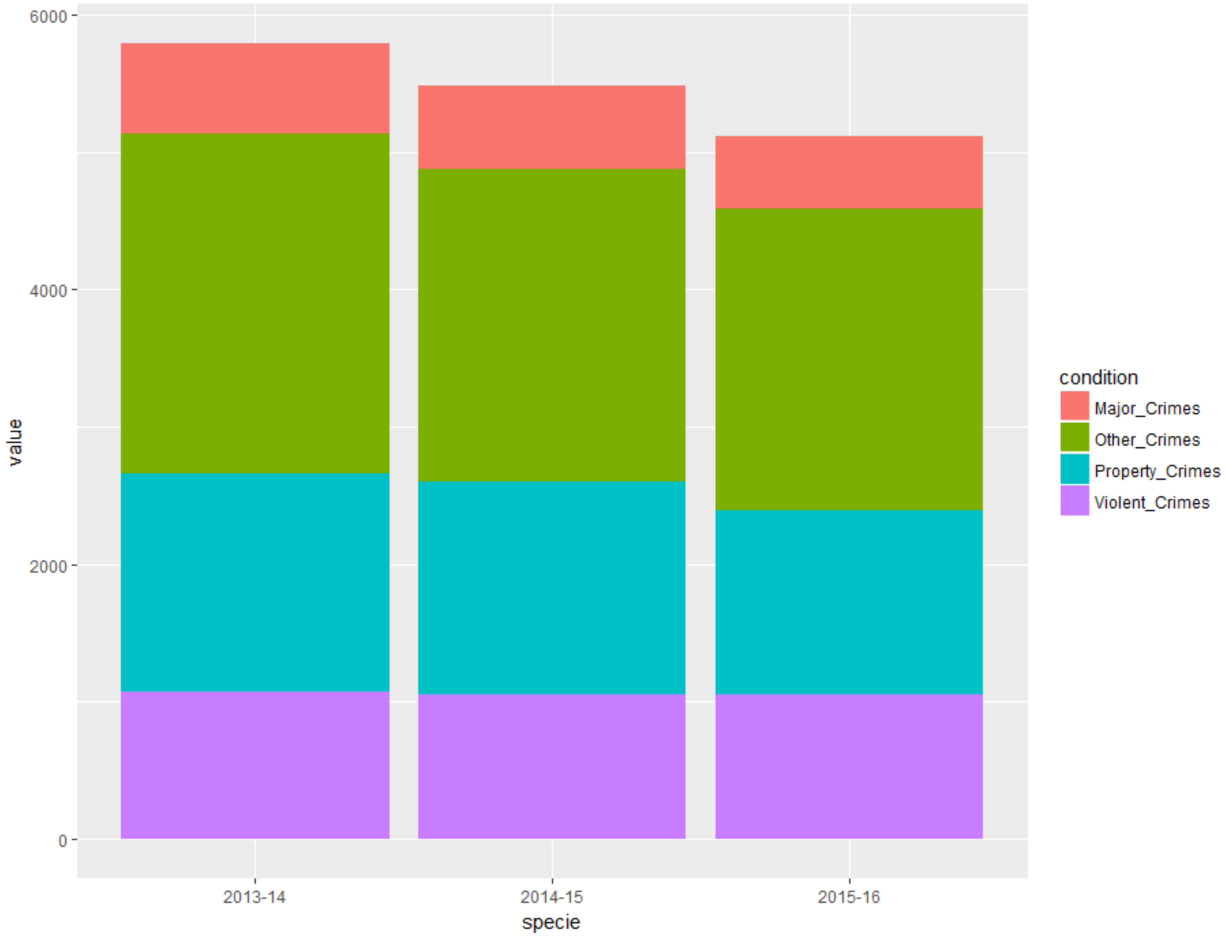
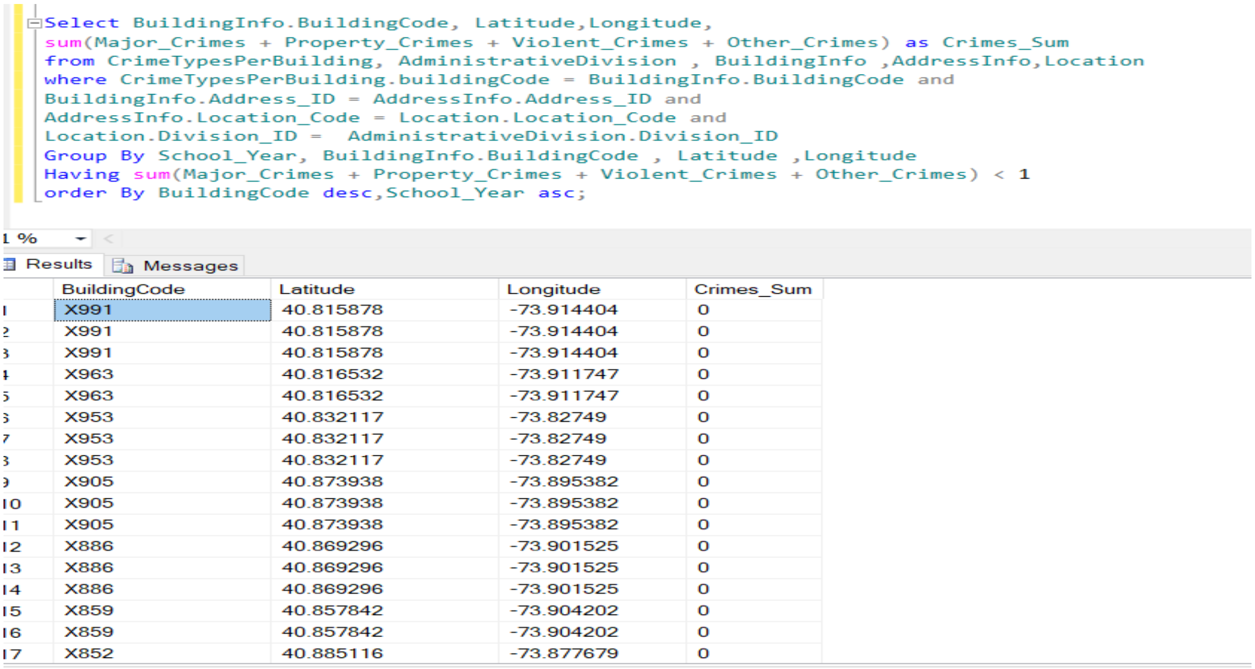


Fig 25: Types of Crime yearly

Goal: Visualizing the Results Mapping the safest schools to attend in New York [10]

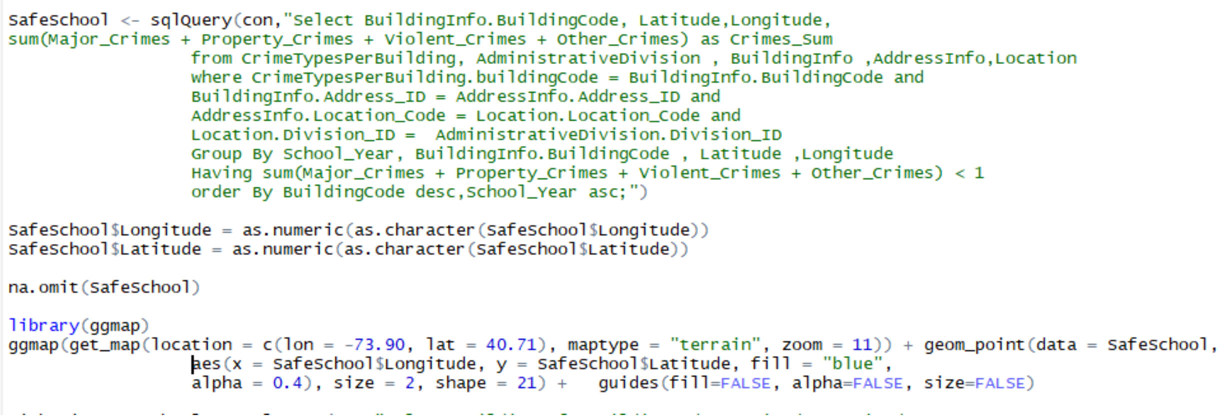
Query:



```
Select BuildingInfo.BuildingCode, Latitude,Longitude,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding, AdministrativeDivision , BuildingInfo ,AddressInfo,Location
where CrimeTypesPerBuilding.buildingCode = BuildingInfo.BuildingCode and
BuildingInfo.Address_ID = AddressInfo.Address_ID and
AddressInfo.Location_Code = Location.Location_Code and
Location.Division_ID =  AdministrativeDivision.Division_ID
Group By School_Year, BuildingInfo.BuildingCode , Latitude ,Longitude
Having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) < 1
order By BuildingCode desc,School_Year asc;
```

	BuildingCode	Latitude	Longitude	Crimes_Sum
1	X991	40.815878	-73.914404	0
2	X991	40.815878	-73.914404	0
3	X991	40.815878	-73.914404	0
4	X963	40.816532	-73.911747	0
5	X963	40.816532	-73.911747	0
6	X953	40.832117	-73.82749	0
7	X953	40.832117	-73.82749	0
8	X953	40.832117	-73.82749	0
9	X905	40.873938	-73.895382	0
10	X905	40.873938	-73.895382	0
11	X905	40.873938	-73.895382	0
12	X886	40.869296	-73.901525	0
13	X886	40.869296	-73.901525	0
14	X886	40.869296	-73.901525	0
15	X859	40.857842	-73.904202	0
16	X859	40.857842	-73.904202	0
17	X852	40.885116	-73.877679	0

R-Code:



```
SafeSchool <- sqlQuery(con,"Select BuildingInfo.BuildingCode, Latitude,Longitude,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
                  from CrimeTypesPerBuilding, AdministrativeDivision , BuildingInfo ,AddressInfo,Location
                  where CrimeTypesPerBuilding.buildingCode = BuildingInfo.BuildingCode and
                  BuildingInfo.Address_ID = AddressInfo.Address_ID and
                  AddressInfo.Location_Code = Location.Location_Code and
                  Location.Division_ID =  AdministrativeDivision.Division_ID
                  Group By School_Year, BuildingInfo.BuildingCode , Latitude ,Longitude
                  Having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) < 1
                  order By BuildingCode desc,School_Year asc;")

SafeSchool$Longitude = as.numeric(as.character(SafeSchool$Longitude))
SafeSchool$Latitude = as.numeric(as.character(SafeSchool$Latitude))

na.omit(SafeSchool)

library(ggmap)
ggmap(get_map(location = c(lon = -73.90, lat = 40.71), maptype = "terrain", zoom = 11)) + geom_point(data = SafeSchool,
                  aes(x = SafeSchool$Longitude, y = SafeSchool$Latitude, fill = "blue",
                  alpha = 0.4), size = 2, shape = 21) +   guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

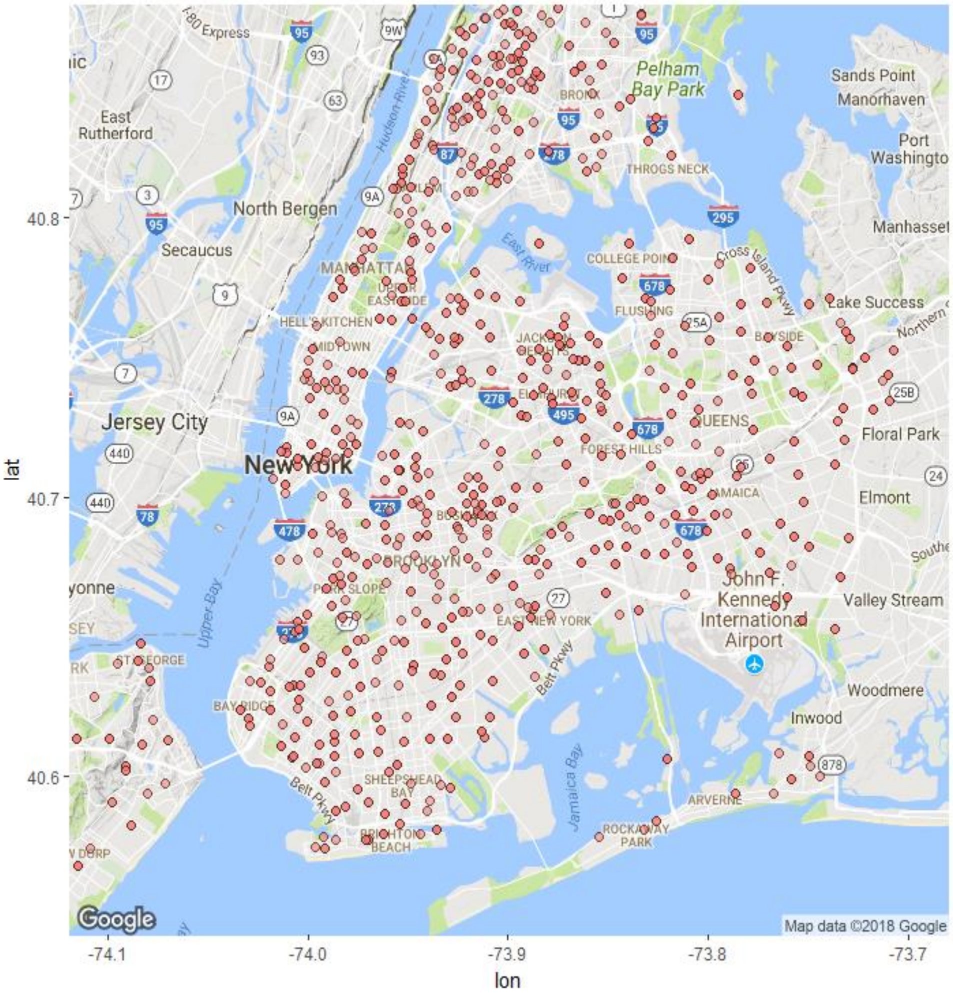
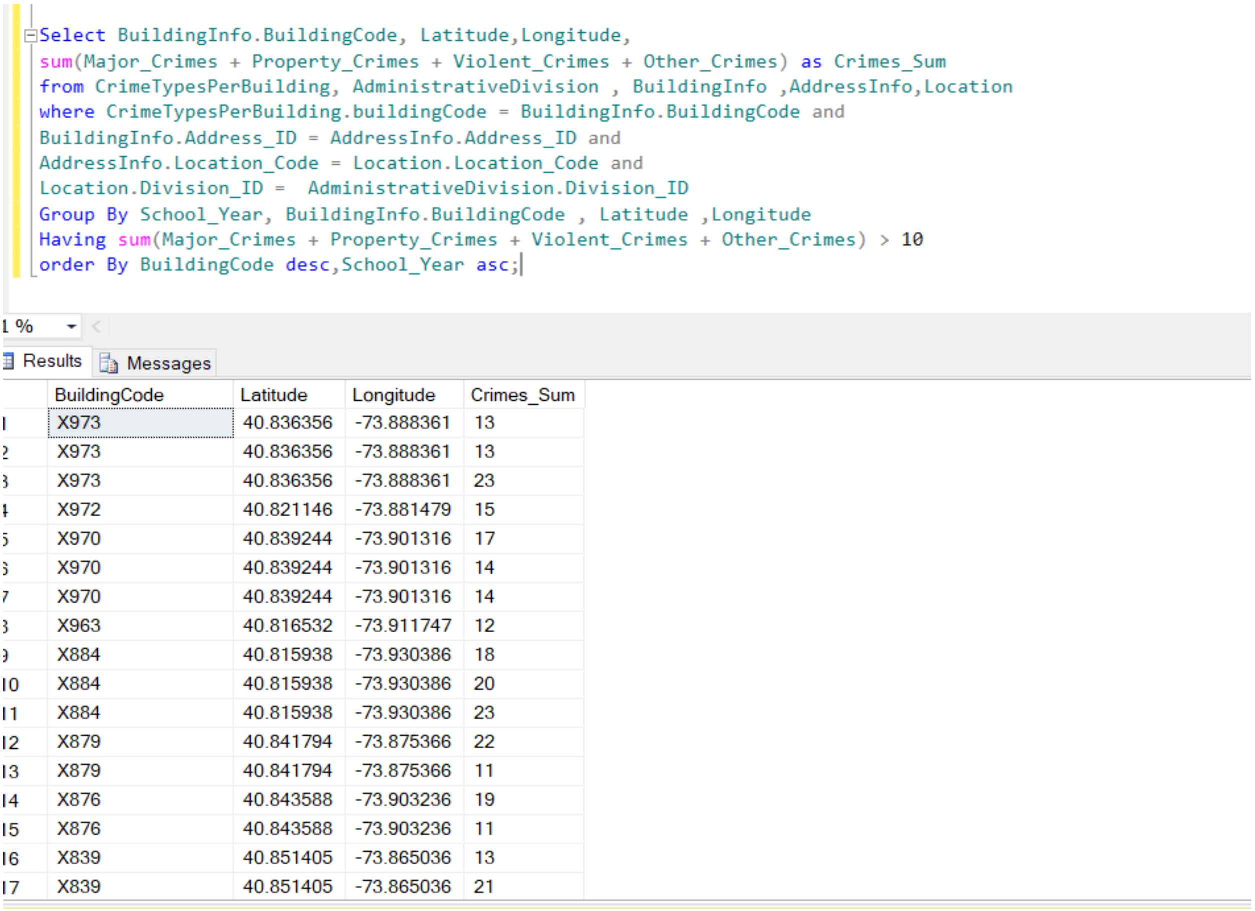


Fig 26: Safe School Location

Goal: Visualizing the Results Mapping the high-risk schools to attend in New York [10]

Query:

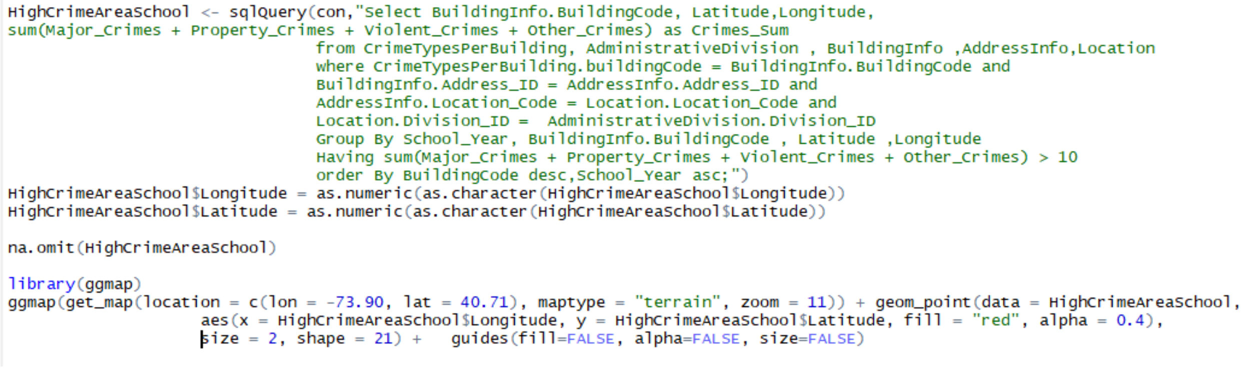


```
Select BuildingInfo.BuildingCode, Latitude,Longitude,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
from CrimeTypesPerBuilding, AdministrativeDivision , BuildingInfo ,AddressInfo,Location
where CrimeTypesPerBuilding.buildingCode = BuildingInfo.BuildingCode and
BuildingInfo.Address_ID = AddressInfo.Address_ID and
AddressInfo.Location_Code = Location.Location_Code and
Location.Division_ID =  AdministrativeDivision.Division_ID
Group By School_Year, BuildingInfo.BuildingCode , Latitude ,Longitude
Having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) > 10
order By BuildingCode desc,School_Year asc;
```

1 %

	BuildingCode	Latitude	Longitude	Crimes_Sum
1	X973	40.836356	-73.888361	13
2	X973	40.836356	-73.888361	13
3	X973	40.836356	-73.888361	23
4	X972	40.821146	-73.881479	15
5	X970	40.839244	-73.901316	17
6	X970	40.839244	-73.901316	14
7	X970	40.839244	-73.901316	14
8	X963	40.816532	-73.911747	12
9	X884	40.815938	-73.930386	18
10	X884	40.815938	-73.930386	20
11	X884	40.815938	-73.930386	23
12	X879	40.841794	-73.875366	22
13	X879	40.841794	-73.875366	11
14	X876	40.843588	-73.903236	19
15	X876	40.843588	-73.903236	11
16	X839	40.851405	-73.865036	13
17	X839	40.851405	-73.865036	21

R-Code:



```
HighCrimeAreaSchool <- sqlQuery(con,"Select BuildingInfo.BuildingCode, Latitude,Longitude,
sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) as Crimes_Sum
                                from CrimeTypesPerBuilding, AdministrativeDivision , BuildingInfo ,AddressInfo,Location
                                where CrimeTypesPerBuilding.buildingCode = BuildingInfo.BuildingCode and
                                BuildingInfo.Address_ID = AddressInfo.Address_ID and
                                AddressInfo.Location_Code = Location.Location_Code and
                                Location.Division_ID =  AdministrativeDivision.Division_ID
                                Group By School_Year, BuildingInfo.BuildingCode , Latitude ,Longitude
                                Having sum(Major_Crimes + Property_Crimes + Violent_Crimes + Other_Crimes) > 10
                                order By BuildingCode desc,School_Year asc;")
HighCrimeAreaSchool$Longitude = as.numeric(as.character(HighCrimeAreaSchool$Longitude))
HighCrimeAreaSchool$Latitude = as.numeric(as.character(HighCrimeAreaSchool$Latitude))

na.omit(HighCrimeAreaSchool)

library(ggmap)
ggmap(get_map(location = c(lon = -73.90, lat = 40.71), maptype = "terrain", zoom = 11)) + geom_point(data = HighCrimeAreaSchool,
                    aes(x = HighCrimeAreaSchool$Longitude, y = HighCrimeAreaSchool$Latitude, fill = "red", alpha = 0.4),
                    size = 2, shape = 21) +   guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

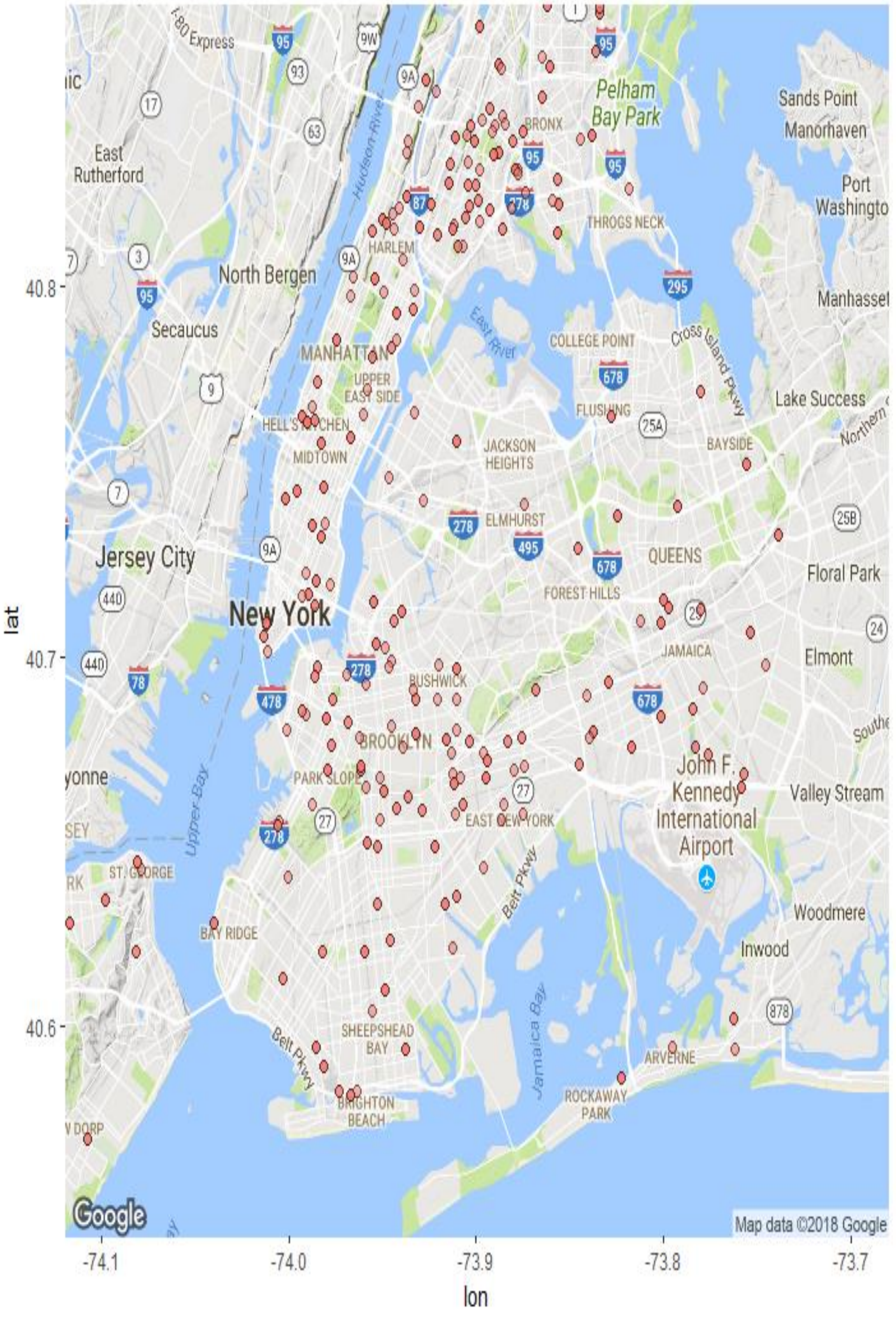


Fig 27: High Risk School Location

CONCLUSION

After getting the normalized database ready, we have come up with some interesting findings after analysis:

- The overall school crime rate has reduced by 4% from the school year 2013-2014 to the school year 2015-2016.
- Comparing the five boroughs, the school crime rate is highest in Brooklyn and lowest in Staten Island.
- Surprisingly the crime rate is lower in buildings with larger number of students.
- Property crime reports are more than violent crime reports.
- Moreover, we found for students which are safe to attend and which schools are with high risk to attend.

Just like this, a lot of useful information can be obtained. It can be used by concerned parents who want to send their children to the safest schools. Security can be increased in places with higher crime rates. There are a lot of external environmental factors that can determine higher crime rates. Schools situated in poor neighborhoods are likely to have more crime incidents. Brooklyn's high crime rate could be due to it being the most populated borough. Similarly, Staten Island is the least populated. We are also able to determine the most popular criminal activities.

By utilizing different data mining and analysis techniques, we have extracted a lot useful information about school crime information. This kind of information would have been much harder to obtained had we not cleaned and normalized the data. Hence, data cleaning and normalization are an absolute necessity if we want to obtain meaningful data, regardless of what kind of data it is.

REFERENCES

[1] https://data.cityofnewyork.us/Education/2010-2016-School-Safety-Report/qybk-bjjc

[2] https://www.r-graph-gallery.com/

[3] https://www.tutorialgateway.org/mosaic-plot-in-r/

[4] http://www.r-tutor.com/r-introduction/

[5]https://www.tutorialspoint.com/dbms/er_diagram_representation.htmL

[6] https://www.draw.io/

[7] https://mycourses.rit.edu/d2l/le/content/686249/Home

[8] https://www.youtube.com/watch?v=2xQX76nEdvo

[9] https://docs.microsoft.com/en-us/sql/ssms/tutorials/tutorial-sql-server-management-studio?view=sql-server-2017

[10] https://cran.r-project.org/web/packages/ggmap/ggmap.pdf