

## 1: Datasets

- All of the datasets are available at: <https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page>
- This is the data about stop, question, and frisk.
- We will be using datasets between years: 2013 – 2016
- Every year has a different dataset file (in total 4 datasets).
- Datasets between 2013 and 2016 are in .csv format and have similar column layout.

Year:	2013	2014	2015	2016	TOTAL:
Rows:	191851	45787	22563	12401	272602
Size:	61,3MB	15MB	7,2MB	4MB	87,5MB

In the first step, I am going to merge data from 2013 to 2016 into one file.

```

ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ cat *.csv > 2013-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ ls
2013-2016.csv sqf-2013.csv sqf-2014.csv sqf-2015.csv sqf-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ hadoop fs -rm Project/*.csv
Deleted Project/sqf-2013.csv
Deleted Project/sqf-2014.csv
Deleted Project/sqf-2015.csv
Deleted Project/sqf-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~$ hadoop fs -put SubwayProject/2013-2016/data.csv Project
ss16249_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls Project
Found 1 items
-rw-r--r-- 1 ss16249_nyu_edu ss16249_nyu_edu 87701814 2022-11-29 09:26 Project/data.csv

```

This merged dataset has now 272, 602 rows and 112 columns in total. Then I put it into HDFS cluster.

- I have manually deleted the 1 row, which is description of the columns in each of the four dataset files (sqf-2013.csv - sqf-2016.csv).
- As the data have 112 columns, I will not provide the data here in the report. (The short sample of the data will be provided in the .zip file).

## 2: Data Cleaning

- The most important columns that I need from the dates are:
  - DATE OF STOP
  - TIME OF STOP
  - CRIME DESCRIPTION
  - X COORDINATES OF STOP
  - Y COORDINATES OF STOP
- I have filtered the data on the following criteria:
  - DATE: I checked if the date is not empty, if it is I dropped the row.
  - TIME: I checked if the time is in correct format (HHmm). If the time, was in format (Hmm), I added “0” before it. Otherwise, I dropped the row.
  - X&Y COORDINATES: I checked if the X Coordinate and Y Coordinate is not empty. If it was, I dropped the row.
  - VIOLENT CRIME:
    - I have considered something violent crime based on the following criteria:
      - If any weapon was found on the person.
      - If there was necessary physical force used by NYPD.
      - If the REASON FOR FRISK - VIOLENT CRIME SUSPECTED was “Y”.

- Based on the chosen HashMap of Violent Crimes discussed with my teammates.
- If the CRIME DESCRIPTION was empty, I assumed it violent. (We agreed this way with my teammates.)
- Then, I run a MapReduce job.

```

ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_cleaning$ hadoop jar cleaning1.jar DataCleaning1 Project/data.csv Project/output_cleaning
2022-12-01 00:45:05,411 INFO client.RMPProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-12-01 00:45:05,609 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-12-01 00:45:05,978 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-12-01 00:45:05,997 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ss16249_nyu_edu/.staging/job_1669608701841_0844
2022-12-01 00:45:06,317 INFO input.FileInputFormat: Total input files to process : 1
2022-12-01 00:45:06,471 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-01 00:45:06,664 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0844
2022-12-01 00:45:06,666 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-01 00:45:06,867 INFO conf.Configuration: resource-types.xml not found
2022-12-01 00:45:06,867 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-01 00:45:07,011 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0844
2022-12-01 00:45:07,051 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0844/
2022-12-01 00:45:07,051 INFO mapreduce.Job: Running job: job_1669608701841_0844
2022-12-01 00:45:14,217 INFO mapreduce.Job: Job job_1669608701841_0844 running in uber mode : false
2022-12-01 00:45:14,219 INFO mapreduce.Job: map 0% reduce 0%
2022-12-01 00:45:31,455 INFO mapreduce.Job: map 5% reduce 0%
2022-12-01 00:45:37,490 INFO mapreduce.Job: map 7% reduce 0%
2022-12-01 00:45:43,523 INFO mapreduce.Job: map 9% reduce 0%
2022-12-01 00:45:49,558 INFO mapreduce.Job: map 12% reduce 0%
2022-12-01 00:45:55,592 INFO mapreduce.Job: map 14% reduce 0%
2022-12-01 00:46:01,626 INFO mapreduce.Job: map 16% reduce 0%
2022-12-01 00:46:07,658 INFO mapreduce.Job: map 19% reduce 0%
2022-12-01 00:46:13,691 INFO mapreduce.Job: map 21% reduce 0%
2022-12-01 00:46:19,724 INFO mapreduce.Job: map 24% reduce 0%
2022-12-01 00:46:25,752 INFO mapreduce.Job: map 26% reduce 0%
2022-12-01 00:46:31,782 INFO mapreduce.Job: map 28% reduce 0%
2022-12-01 00:46:37,812 INFO mapreduce.Job: map 31% reduce 0%
2022-12-01 00:46:43,841 INFO mapreduce.Job: map 33% reduce 0%
2022-12-01 00:46:49,869 INFO mapreduce.Job: map 35% reduce 0%
2022-12-01 00:46:55,896 INFO mapreduce.Job: map 38% reduce 0%
2022-12-01 00:47:01,925 INFO mapreduce.Job: map 40% reduce 0%
2022-12-01 00:47:07,952 INFO mapreduce.Job: map 43% reduce 0%
2022-12-01 00:47:13,977 INFO mapreduce.Job: map 45% reduce 0%
2022-12-01 00:47:20,003 INFO mapreduce.Job: map 47% reduce 0%
2022-12-01 00:47:26,030 INFO mapreduce.Job: map 50% reduce 0%
2022-12-01 00:47:32,054 INFO mapreduce.Job: map 52% reduce 0%
2022-12-01 00:47:38,077 INFO mapreduce.Job: map 55% reduce 0%
2022-12-01 00:47:44,100 INFO mapreduce.Job: map 57% reduce 0%
2022-12-01 00:47:50,126 INFO mapreduce.Job: map 59% reduce 0%
2022-12-01 00:47:56,149 INFO mapreduce.Job: map 62% reduce 0%
2022-12-01 00:48:02,170 INFO mapreduce.Job: map 64% reduce 0%
2022-12-01 00:48:08,205 INFO mapreduce.Job: map 100% reduce 0%
2022-12-01 00:48:14,228 INFO mapreduce.Job: map 100% reduce 100%
2022-12-01 00:48:15,238 INFO mapreduce.Job: Job job_1669608701841_0844 completed successfully
2022-12-01 00:48:15,342 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=5857041
  FILE: Number of bytes written=12206109
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=87701938
  HDFS: Number of bytes written=5594693
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=3
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Rack-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=684836
  Total time spent by all reduces in occupied slots (ms)=13256
  Total time spent by all map tasks (ms)=171209
  Total time spent by all reduce tasks (ms)=3314
  Total vcore-milliseconds taken by all map tasks=171209
  Total vcore-milliseconds taken by all reduce tasks=3314
  Total megabyte-milliseconds taken by all map tasks=701272064
  Total megabyte-milliseconds taken by all reduce tasks=13574144
Map-Reduce Framework
  Map input records=272602
  Map output records=131171
  Map output bytes=5594693
  Map output materialized bytes=5857041
  Input split bytes=124
  Combine input records=0
  Combine output records=0
  Reduce input groups=1
  Reduce shuffle bytes=5857041
  Reduce input records=131171
  Reduce output records=131171
  Spilled Records=262342
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=178
  CPU time spent (ms)=176900
  Physical memory (bytes) snapshot=1225826304
  Virtual memory (bytes) snapshot=9622306816
  Total committed heap usage (bytes)=1373634560
  Peak Map Physical memory (bytes)=775372800
  Peak Map Virtual memory (bytes)=4811550720
  Peak Reduce Physical memory (bytes)=485167104
  Peak Reduce Virtual memory (bytes)=4811968512

```

```

Map-Reduce Framework
  Map input records=272602
  Map output records=131171
  Map output bytes=5594693
  Map output materialized bytes=5857041
  Input split bytes=124
  Combine input records=0
  Combine output records=0
  Reduce input groups=1
  Reduce shuffle bytes=5857041
  Reduce input records=131171
  Reduce output records=131171
  Spilled Records=262342
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=178
  CPU time spent (ms)=176900
  Physical memory (bytes) snapshot=1225826304
  Virtual memory (bytes) snapshot=9622306816
  Total committed heap usage (bytes)=1373634560
  Peak Map Physical memory (bytes)=775372800
  Peak Map Virtual memory (bytes)=4811550720
  Peak Reduce Physical memory (bytes)=485167104
  Peak Reduce Virtual memory (bytes)=4811968512
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=87701814
File Output Format Counters
  Bytes Written=5594693
ss16249_nyu_edu@nyu-dataproc-1:~/SubwayProject/2013-2016/code_cleaning$

```

- This is sample cleaned data:

```

SubwayProject > 2013-2016 > code_cleaning > output > ≡ part-r-00000
1 11222016,2212,CPW,939042,132823
2 12022016,1847,ASSAULT,935054,136263
3 11032016,1725,VIOLENT_CRIME_SUSPECTED,922098,135797
4 11032016,1035,ASSAULT,917675,129284
5 10162016,1951,CPW,918014,123544
6 10162016,1951,CPW,918014,123544
7 10162016,1951,CPW,918014,123544
8 10132016,1320,CPW,923947,129507
9 10112016,0120,PHYSICAL_FORCE_USED,925065,140569
10 12242016,1935,PHYSICAL_FORCE_USED,961179,159021
11 12222016,0304,VIOLENT_CRIME_SUSPECTED,958629,154949
12 12222016,0304,VIOLENT_CRIME_SUSPECTED,958629,154949
13 12172016,1404,CRIMINAL SALE OF CONTROLLED SUBSTANCE,942326,140590
14 12172016,1404,CRIMINAL SALE OF CONTROLLED SUBSTANCE,942326,140590
15 12112016,1454,PHYSICAL_FORCE_USED,951527,146987
16 12112016,0425,VIOLENT_CRIME_SUSPECTED,954598,143146
17 11242016,0315,PHYSICAL_FORCE_USED,938281,137671
18 11242016,0910,PHYSICAL_FORCE_USED,965461,154802
19 11242016,0910,PHYSICAL_FORCE_USED,965461,154802
20 11222016,0530,PHYSICAL_FORCE_USED,957236,161624

```

### 3: Data Profiling

- In the profiling, I wrote 2 MapReduce jobs. One for summarizing what are the total crimes per each hour of the day and another one, which summarizes the frequency of each crime description.
- Here is the MapReduce job log for the total crimes per each hour:

```

ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_hours$ hadoop jar hours.jar HoursApp Project/output_cleaning/part-r-00000 Project/output_hours2022
-12-01 00:52:32,756 INFO client.RMPProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-12-01 00:52:32,940 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-12-01 00:52:33,121 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your appl
ication with ToolRunner to remedy this.
2022-12-01 00:52:33,187 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ss16249_nyu_edu/.staging/job_1669608701841_0846
41.0846
2022-12-01 00:52:33,504 INFO input.FileInputFormat: Total input files to process : 1
2022-12-01 00:52:33,639 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-01 00:52:33,829 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0846
2022-12-01 00:52:33,831 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-01 00:52:34,022 INFO conf.Configuration: resource-types.xml not found
2022-12-01 00:52:34,023 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-01 00:52:34,285 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0846
2022-12-01 00:52:34,322 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0846/
2022-12-01 00:52:34,322 INFO mapreduce.Job: Running job: job_1669608701841_0846
2022-12-01 00:52:41,411 INFO mapreduce.Job: Job job_1669608701841_0846 running in uber mode : false
2022-12-01 00:52:41,412 INFO mapreduce.Job: map 0% reduce 0%
2022-12-01 00:52:47,481 INFO mapreduce.Job: map 100% reduce 0%
2022-12-01 00:52:52,514 INFO mapreduce.Job: map 100% reduce 100%
2022-12-01 00:52:53,529 INFO mapreduce.Job: Job job_1669608701841_0846 completed successfully
2022-12-01 00:52:53,626 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=918203
    FILE: Number of bytes written=2328409
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=5594837
    HDFS: Number of bytes written=347
    HDFS: Number of read operations=8
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Rack-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=13216
    Total time spent by all reduces in occupied slots (ms)=12284
    Total time spent by all map tasks (ms)=3304
    Total time spent by all reduce tasks (ms)=3071
    Total vcore-milliseconds taken by all map tasks=3304
    Total vcore-milliseconds taken by all reduce tasks=3071
    Total megabyte-milliseconds taken by all map tasks=13533184
    Total megabyte-milliseconds taken by all reduce tasks=12578816
  Map-Reduce Framework
    Map input records=131171
    Map output records=131171
    Map output bytes=655855
    Map output materialized bytes=918203
    Input split bytes=144
    Combine input records=0
    Combine output records=0
    Reduce input groups=23
    Reduce shuffle bytes=918203
    Reduce input records=131171
    Reduce output records=23
    Spilled Records=262342
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=99
    CPU time spent (ms)=4570
    Physical memory (bytes) snapshot=1225314304
    Virtual memory (bytes) snapshot=9640701952
    Total committed heap usage (bytes)=1434976256
    Peak Map Physical memory (bytes)=742440960
    Peak Map Virtual memory (bytes)=4811042816
    Peak Reduce Physical memory (bytes)=482873344
    Peak Reduce Virtual memory (bytes)=4829659136
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=5594693
  File Output Format Counters
    Bytes Written=347

```



```

Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=5594693
File Output Format Counters
    Bytes Written=347
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_hours$ hadoop fs -cat Project/output_hours/part-r-00000
01 Count: 9387
02 Count: 6716
03 Count: 4402
04 Count: 2840
05 Count: 1375
06 Count: 802
07 Count: 658
08 Count: 1222
09 Count: 1012
10 Count: 2326
11 Count: 3498
12 Count: 4379
13 Count: 5319
14 Count: 6157
15 Count: 6360
16 Count: 7250
17 Count: 6789
18 Count: 6686
19 Count: 8475
20 Count: 10124
21 Count: 12094
22 Count: 11693
23 Count: 10799
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_hours$

```

- From the results, we can see that the crime in the NYC is highest from 20:00 to 01:00 in the night. The least violent crime is between 06:00 to 07:00.
- Let's move to the second MapReduce job, where we are finding the violent crime frequency. Here is the log of the MapReduce job:

```

ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_crimes$ hadoop jar crimes.jar CrimesApp Project/output_cleaning/part-r-00000 Project/output_crimes
2022-12-01 00:56:56,650 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-12-01 00:56:56,839 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-12-01 00:56:57,156 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-12-01 00:56:57,214 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ss16249_nyu_edu/.staging/job_1669608701841_0847
2022-12-01 00:56:57,555 INFO input.FileInputFormat: Total input files to process : 1
2022-12-01 00:56:57,703 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-01 00:56:57,929 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0847
2022-12-01 00:56:57,931 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-01 00:56:58,144 INFO conf.Configuration: resource-types.xml not found
2022-12-01 00:56:58,145 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-01 00:56:58,302 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0847
2022-12-01 00:56:58,345 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0847/
2022-12-01 00:56:58,346 INFO mapreduce.Job: Running job: job_1669608701841_0847
2022-12-01 00:57:07,464 INFO mapreduce.Job: Job job_1669608701841_0847 running in uber mode : false
2022-12-01 00:57:07,465 INFO mapreduce.Job: map 0% reduce 0%
2022-12-01 00:57:13,544 INFO mapreduce.Job: map 100% reduce 0%
2022-12-01 00:57:18,580 INFO mapreduce.Job: map 100% reduce 100%
2022-12-01 00:57:19,604 INFO mapreduce.Job: Job job_1669608701841_0847 completed successfully
2022-12-01 00:57:19,725 INFO mapreduce.Job: Counters: 54
File System Counters
    FILE: Number of bytes read=2472043
    FILE: Number of bytes written=5436097
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=5594837
    HDFS: Number of bytes written=1076
    HDFS: Number of read operations=8
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
    HDFS: Number of bytes read erasure-coded=0
Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Rack-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=13816
    Total time spent by all reduces in occupied slots (ms)=12472
    Total time spent by all map tasks (ms)=3454
    Total time spent by all reduce tasks (ms)=3118
    Total vcore-milliseconds taken by all map tasks=3454
    Total vcore-milliseconds taken by all reduce tasks=3118
    Total megabyte-milliseconds taken by all map tasks=14147584
    Total megabyte-milliseconds taken by all reduce tasks=12771328

```

```

Map-Reduce Framework
  Map input records=131171
  Map output records=131171
  Map output bytes=2209695
  Map output materialized bytes=2472043
  Input split bytes=144
  Combine input records=0
  Combine output records=0
  Reduce input groups=38
  Reduce shuffle bytes=2472043
  Reduce input records=131171
  Reduce output records=38
  Spilled Records=262342
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=99
  CPU time spent (ms)=4810
  Physical memory (bytes) snapshot=1218142208
  Virtual memory (bytes) snapshot=9607217152
  Total committed heap usage (bytes)=1437597696
  Peak Map Physical memory (bytes)=740171776
  Peak Map Virtual memory (bytes)=4798590976
  Peak Reduce Physical memory (bytes)=477970432
  Peak Reduce Virtual memory (bytes)=4808626176

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=5594693
File Output Format Counters
  Bytes Written=1076
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_crimes$ hadoop fs -cat Project/output_crimes/part-r-00000
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_crimes$ hadoop fs -cat Project/output_crimes/part-r-00000
AGGRAVATED ASSAULT          Count: 11
AGGRAVATED HARASSMENT       Count: 78
AGGRAVATED SEXUAL ABUSE     Count: 2
ARSON                        Count: 52
ASSAULT                      Count: 9923
COERCION                     Count: 27
COURSE OF SEXUAL CONDUCT     Count: 7
CPW                          Count: 53798
CRIMINAL POSSESSION OF CONTROLLED SUBSTANCE Count: 4697
CRIMINAL POSSESSION OF FORGED INSTRUMENT Count: 374
CRIMINAL SALE OF CONTROLLED SUBSTANCE Count: 4108
ENDANGER THE WELFARE OF A CHILD Count: 49
ESCAPE                       Count: 20
FORCIBLE TOUCHING           Count: 285
FRAUDULENT ACCOSTING        Count: 200
HARASSMENT                   Count: 89
HAZING                       Count: 6
HINDERING PROSECUTION       Count: 4
INCEST                       Count: 1
KIDNAPPING                  Count: 76
MENACING                     Count: 391
MURDER                       Count: 108
OBSCENITY                   Count: 5
PHYSICAL FORCE USED          Count: 21205
PROHIBITED USE OF WEAPON    Count: 12
RAPE                         Count: 197
RECKLESS ENDANGERMENT       Count: 281
RESISTING ARREST            Count: 13
RIOT                         Count: 10
SEXUAL ABUSE                 Count: 227
SEXUAL MISCONDUCT           Count: 42
SEXUAL PERFORMANCE BY A CHILD Count: 3
TERRORISM                   Count: 358
UNLAWFULL IMPRISONMENT      Count: 19
UNLAWFULLY DEALING WITH FIREWORKS Count: 11
VEHICULAR ASSAULT           Count: 10
VIOLENT CRIME SUSPECTED    Count: 33728
WEAPON_FOUND                Count: 744
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_crimes$

```

- As we can see the most crimes were with concealed prohibited weapon (CPW).

#### 4: Future Work

- As there is still more data to clean and profile, it is obvious that I would like to add more data to this dataset.
- To do this, I would have to write another DataCleaning Mapper program, as the column layout and column data type changed from year to year before 2013 and after 2016. This will require to write a new DataCleaning program for every year separately and concatenate the result to the results that I have already produced.