Data Profiling, Cleaning and Ingestion using MapReduce

Anirudh Garg
Courant Institute of
Mathematical Sciences
New York University
ag9563@nyu.edu

Objective

To perform data profiling, cleaning, and ingestion on a dataset containing incident-related information. Specifically, we aimed to:

- Filter and clean data based on specific criteria.
- Compute profiling statistics to understand data quality and characteristics.
- Save the cleaned dataset and profiling statistics for further analysis.

Dataset Overview

Fire Department and Emergency Medical Services Dispatched Calls for Service

- **Link**: Dataset Link
- **Ownership**: Publicly accessible data from the City of San Francisco.
- **Data Size**: 2.7 GB

These dataset is periodic, and being updated regularly.

The dataset consisted of the following columns:

- **Call Number:** A unique 9-digit number assigned by the 911 Dispatch Center (DEM) to this call. These number are used for both Police and Fire calls.
- **Unit ID:** Unit Identifier. For example E01 for Engine 1 or T01 for Truck 1.
- **Unit ID:** Unit Identifier. For example E01 for Engine 1 or T01 for Truck 1.
- **Incident Number:** A unique 8-digit number assigned by DEM to this Fire incident.
- **Call Type:** Type of call the incident falls into.
- **Call Date:** Date the call is received at the 911 Dispatch Center. Used for reporting purposes.
- **Watch Date:** Watch date when the call is received. Watch date starts at 0800 each morning and ends at 0800 the next day.
- **Received DtTm:** Date and time of call is received at the 911 Dispatch Center.
- **Entry DtTm:** Date and time the 911 operator submits the entry of the initial call information into the CAD system
- **Dispatch DtTm:** Date and time the 911 operator dispatches this unit to the call.
- **Response DtTm:** Date and time this unit acknowledges the dispatch and records that the unit is en route to the location of the call.
- **On Scene DtTm:** Date and time the unit records arriving to the location of the incident

- **Transport DtTm:** If this unit is an ambulance, date and time the unit begins the transport to the hospital
- **Hospital DtTm:** If this unit is an ambulance, date and time the unit arrives to the hospital.
- **Call Final Disposition:** Disposition of the call (Code). For example TH2: Transport to Hospital - Code 2, FIR: Resolved by Fire Department
- **Available DtTm:** Date and time this unit is not longer assigned to this call and it is available for another dispatch.
- **Address:** Address of intersection or call box point associated with incident (obfuscated address to protect caller privacy)
- **City:** City of incident
- **Zipcode of Incident:** Zip code of incident
- **Battalion:** Emergency Response District (There are 10 Fire Emergency Response Districts).
- **Station Area:** Fire Station First Response Area associated with the address of the incident.
- **Box:** Fire box associated with the address of the incident. A box is the smallest area used to divide the City. Each box is associated with a unique unit dispatch order. The City is divided into more than 2,400 boxes.
- **Original Priority:** Initial call priority (Code 2: Non-Emergency or Code 3:Emergency).
- **Priority:** Call priority once all information has been asssessed (Code 2: Non-Emergency or Code 3:Emergency).
- **Final Priority:** Final call priority (Code 2: Non-Emergency or Code 3:Emergency).
- **ALS Unit:** Does this unit includes ALS (Advance Life Support) resources? Is there a paramedic in this unit?
- **Call Type Group:** Call types are divided into four main groups: Fire, Alarm, Potential Life Threatening and Non Life Threatening.
- **Number of Alarms:** There are five levels of fire alarms (1-5). The number of alarms indicates the number of resources required in an incident. This number is a combination of engines, trucks, rescue squads, chiefs and EMS units.
- **Unit Type:** Type of unit responding
- **Unit sequence in call dispatch:** A number that indicates the order this unit was assigned to this call.
- **Fire Prevention District:** Bureau of Fire Prevention District associated with this address
- **Supervisor District:** Supervisor District number
- **Neighborhooods - Analysis Boundaries:** San Francisco Neighborhood associated with the incident address.
- **RowID:** Unique identifier used for managing data updates. It is the concatenation of Call Number and Unit ID separated by a dash.
- **case_location:** Latitude and Longitude for the call
- **data_as_of:** Timestamp when the record (row) was last updated in the source system.
- **data_loaded_at:** time the data was loaded into the Open Data Portal

Data Acquisition and Access

- **Storage**: Data will be stored on distributed file system, HDFS, to accommodate storage needs and allow team-wide accessibility.
- **Permissions**: As these datasets are public, no additional permissions are required beyond accessing and downloading from the respective government data portals.
- **Access Time**: Instant access; no approval needed for these publicly hosted datasets.

Data Profiling and Cleaning Approach

We focused on profiling and cleaning the dataset, ensuring that data is structured, consistent, and free of erroneous entries. MapReduce and other technolgies taught in the course were used to profile, clean, and transform the data.

Profiling

- **Purpose**: Characterize each data column to understand its content, distribution, and structure. This includes identifying column data types, ranges, unique values, and possible data issues.
- **Tools**: MapReduce jobs will be written to process large datasets and output profiles on each column.

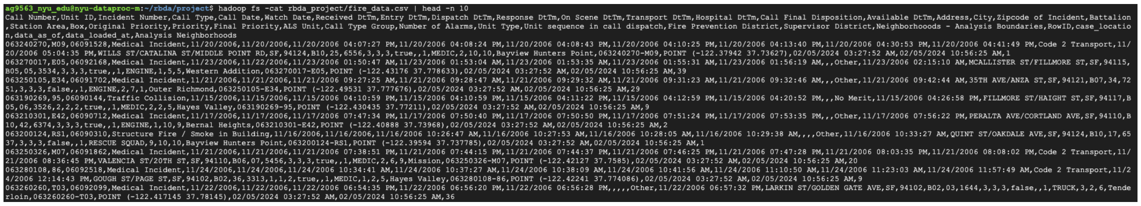
Steps Taken for Data Filtering, Transformation, and Analysis

- **Filter Columns:** Created the `FilterColumns.java` program, including a Mapper (`FilterColumnsMapper.java`) and a Reducer (`FilterColumnsReducer.java`) to filter specific columns from the input data. The columns filtered for further analysis are as follows:

  `Call Number`, `Incident Number`, `Call Type`, `Call Date`, `Received DtTm`, `Response DtTm`, `On Scene DtTm`, `Transport DtTm`, `Hospital DtTm`, `Call Final Disposition`, `Address`, `City`, `Zipcode of Incident`, `Battalion`, `Station Area`, `Box`, `Final Priority`, `ALS Unit`, `Call Type Group`, `Number of Alarms`, `Unit Type`, `Neighborhooods - Analysis Boundaries`
- **Filter and Transform:** Created the `FilterAndTransform.java` program with a Mapper (`FilterAndTransformMapper.java`) and a Reducer (`FilterAndTransformReducer.java`) to apply transformations to the filtered data. The transformation applied was to convert `mmddyyyy` to `yyyymmdd` and then taking only a subset of the dates, Jan 1st 2014 to October 31st 2024 (20140101 to 20241031). This was done using the `Call Date` column. Also, in the `city` column, we find there are multiple names referring to San Francisco, (such as SFO, SAN FRANCISCO, San Francisco), we combine them to refer to San Francisco.
- **Value Counts:** To get to know the value counts of each of the selected columns, I wrote the `ValueCounts.java` program along with the corresponding Mapper (`ValueCountsMapper.java`) and Reducer (`ValueCountsReducer.java`) to calculate the frequency of unique values in a specified column. The function was written such that apart from the 2 standard arguments of input path and output path, another argument was added which took in the parameter of column name for which the value count is required.

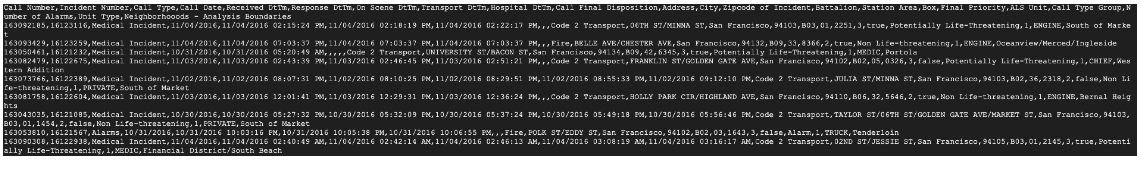
Outputs

Original Data:



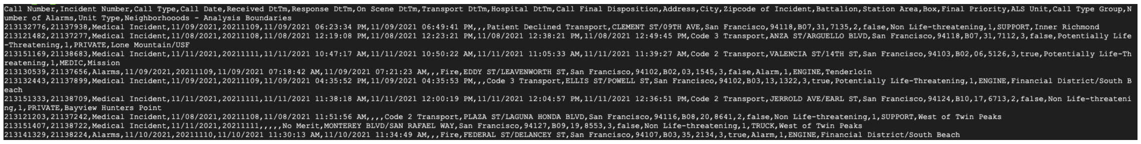
```
ag9563_nyu_edu@nyu-dataproc-m:~/rbda/project$ hadoop fs -cat rbda_project/fire_data.csv | head -n 10
Call Number,Unit ID,Incident Number,Call Type,Call Date,Watch Date,Received DtTm,Entry DtTm,Dispatch DtTm,Response DtTm,On Scene DtTm,Transport DtTm,Hospital DtTm,Call Final Disposition,Available DtTm,Address,City,Zipcode of Incident,Battalion
,Station Area,Box,Original Priority,Priority,Final Priority,ALS Unit,Call Type Group,Number of Alarms,Unit Type,Unit sequence in call dispatch,Fire Prevention District,Supervisor District,Neighborhooods - Analysis Boundaries,RowID,case_locatio
n,data_as_of,data_loaded_at,Analysis Neighborhoods
063240270,M09,06091528,Medical Incident,11/20/2006,11/20/2006,11/20/2006 04:07:27 PM,11/20/2006 04:08:24 PM,11/20/2006 04:08:43 PM,11/20/2006 04:10:25 PM,11/20/2006 04:13:40 PM,11/20/2006 04:30:53 PM,11/20/2006 04:41:49 PM,Code 2 Transport,11/
20/2006 05:04:35 PM,WILLS ST/CATALINA ST/MIDDLE POINT RD,SF,94124,B10,25,6556,3,3,3,true,,1,MEDIC,2,10,10,Bayview Hunters Point,063240270-M09,POINT (-122.37942 37.73627),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,1
063270017,E05,06092168,Medical Incident,11/23/2006,11/22/2006,11/23/2006 01:50:47 AM,11/23/2006 01:53:04 AM,11/23/2006 01:53:35 AM,11/23/2006 01:55:31 AM,11/23/2006 01:56:19 AM,,,Other,11/23/2006 02:15:10 AM,MCALLISTER ST/FILLMORE ST,SF,94115,
B05,05,3534,3,3,3,true,,1,ENGINE,1,5,5,Western Addition,063270017-E05,POINT (-122.43176 37.778633),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,39
063250105,E34,06091702,Medical Incident,11/21/2006,11/21/2006,11/21/2006 09:27:25 AM,11/21/2006 09:28:47 AM,11/21/2006 09:29:32 AM,11/21/2006 09:31:23 AM,11/21/2006 09:32:46 AM,,,Other,11/21/2006 09:42:44 AM,35TH AVE/ANZA ST,SF,94121,B07,34,72
51,3,3,3,false,,1,ENGINE,2,7,1,Outer Richmond,063250105-E34,POINT (-122.49531 37.777676),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,29
063190269,95,06090144,Traffic Collision,11/15/2006,11/15/2006,11/15/2006 04:10:59 PM,11/15/2006 04:10:59 PM,11/15/2006 04:11:22 PM,11/15/2006 04:12:59 PM,11/15/2006 04:20:52 PM,,,No Merit,11/15/2006 04:26:58 PM,FILLMORE ST/HAIGHT ST,SF,94117,B
05,06,3526,2,2,2,true,,1,MEDIC,2,2,5,Hayes Valley,063190269-95,POINT (-122.430435 37.77211),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,9
063210301,E42,06090712,Medical Incident,11/17/2006,11/17/2006,11/17/2006 07:47:34 PM,11/17/2006 07:50:40 PM,11/17/2006 07:50:50 PM,11/17/2006 07:51:24 PM,11/17/2006 07:53:35 PM,,,Other,11/17/2006 07:56:22 PM,PERALTA AVE/CORTLAND AVE,SF,94110,B
10,42,6374,3,3,3,true,,1,ENGINE,1,10,9,Bernal Heights,063210301-E42,POINT (-122.40888 37.73968),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,2
063200124,RS1,06090310,Structure Fire / Smoke in Building,11/16/2006,11/16/2006,11/16/2006 10:26:47 AM,11/16/2006 10:27:53 AM,11/16/2006 10:28:05 AM,11/16/2006 10:29:38 AM,,,,Other,11/16/2006 10:33:27 AM,QUINT ST/OAKDALE AVE,SF,94124,B10,17,65
37,3,3,3,false,,1,RESCUE SQUAD,9,10,10,Bayview Hunters Point,063200124-RS1,POINT (-122.39594 37.737785),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,1
063250326,M07,06091862,Medical Incident,11/21/2006,11/21/2006,11/21/2006 07:38:51 PM,11/21/2006 07:44:15 PM,11/21/2006 07:44:37 PM,11/21/2006 07:46:25 PM,11/21/2006 07:47:28 PM,11/21/2006 08:03:35 PM,11/21/2006 08:08:02 PM,Code 2 Transport,11/
21/2006 08:36:45 PM,VALENCIA ST/20TH ST,SF,94110,B06,07,5456,3,3,3,true,,1,MEDIC,2,6,9,Mission,063250326-M07,POINT (-122.42127 37.7585),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,20
063280108,86,06092518,Medical Incident,11/24/2006,11/24/2006,11/24/2006 10:34:41 AM,11/24/2006 10:37:27 AM,11/24/2006 10:38:09 AM,11/24/2006 10:41:56 AM,11/24/2006 11:10:50 AM,11/24/2006 11:23:03 AM,11/24/2006 11:57:49 AM,Code 2 Transport,11/2
4/2006 12:14:43 PM,GOUGH ST/PAGE ST,SF,94102,B02,36,3313,1,1,2,true,,1,MEDIC,1,2,5,Hayes Valley,063280108-86,POINT (-122.42241 37.774086),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,9
063260260,T03,06092099,Medical Incident,11/22/2006,11/22/2006,11/22/2006 06:54:35 PM,11/22/2006 06:56:20 PM,11/22/2006 06:56:28 PM,,,,,Other,11/22/2006 06:57:32 PM,LARKIN ST/GOLDEN GATE AVE,SF,94102,B02,03,1644,3,3,3,false,,1,TRUCK,3,2,6,Tende
rloin,063260260-T03,POINT (-122.417145 37.78145),02/05/2024 03:27:52 AM,02/05/2024 10:56:25 AM,36
```

Filter Columns:



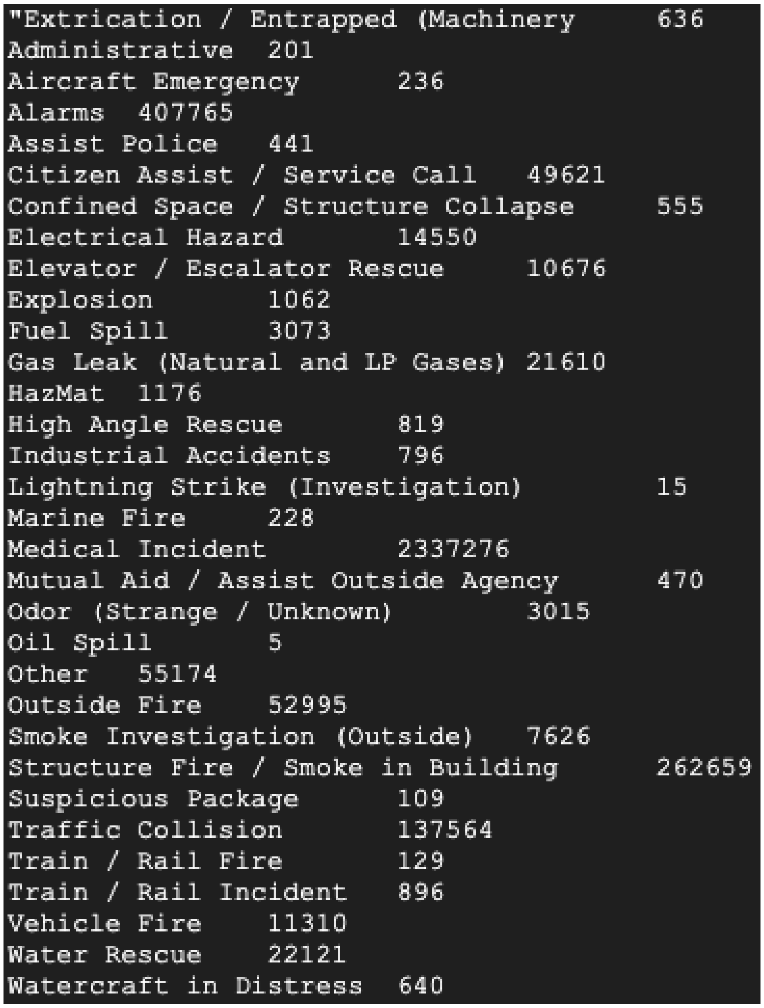
```
Call Number,Incident Number,Call Type,Call Date,Received DtTm,Response DtTm,On Scene DtTm,Transport DtTm,Hospital DtTm,Call Final Disposition,Address,City,Zipcode of Incident,Battalion,Station Area,Box,Final Priority,ALS Unit,Call Type Group,N
umber of Alarms,Unit Type,Neighborhooods - Analysis Boundaries
163092185,16123116,Medical Incident,11/04/2016,11/04/2016 02:15:24 PM,11/04/2016 02:18:19 PM,11/04/2016 02:22:17 PM,,,Code 2 Transport,06TH ST/MINNA ST,San Francisco,94103,B03,01,2251,3,true,Potentially Life-Threatening,1,ENGINE,South of Marke
t
163093429,16123259,Medical Incident,11/04/2016,11/04/2016 07:03:37 PM,11/04/2016 07:03:37 PM,11/04/2016 07:03:37 PM,,,Fire,BELLE AVE/CHESTER AVE,San Francisco,94132,B09,33,8366,2,true,Non Life-threatening,1,ENGINE,Oceanview/Merced/Ingleside
163050461,16121232,Medical Incident,10/31/2016,10/31/2016 05:20:49 AM,,,,,Code 2 Transport,UNIVERSITY ST/BACON ST,San Francisco,94134,B09,42,6345,3,true,Potentially Life-Threatening,1,MEDIC,Portola
163082479,16122675,Medical Incident,11/03/2016,11/03/2016 02:43:39 PM,11/03/2016 02:46:45 PM,11/03/2016 02:51:21 PM,,,Code 2 Transport,FRANKLIN ST/GOLDEN GATE AVE,San Francisco,94102,B02,05,0326,3,false,Potentially Life-Threatening,1,CHIEF,Wes
tern Addition
163073766,16122389,Medical Incident,11/02/2016,11/02/2016 08:07:31 PM,11/02/2016 08:10:25 PM,11/02/2016 08:29:51 PM,11/02/2016 08:55:33 PM,11/02/2016 09:12:10 PM,Code 2 Transport,JULIA ST/MINNA ST,San Francisco,94103,B02,36,2318,2,false,Non Li
fe-threatening,1,PRIVATE,South of Market
163081758,16122604,Medical Incident,11/03/2016,11/03/2016 12:01:41 PM,11/03/2016 12:29:31 PM,11/03/2016 12:36:24 PM,,,Code 2 Transport,HOLLY PARK CIR/HIGHLAND AVE,San Francisco,94110,B06,32,5646,2,true,Non Life-threatening,1,ENGINE,Bernal Heig
hts
163043035,16121085,Medical Incident,10/30/2016,10/30/2016 05:27:32 PM,10/30/2016 05:32:09 PM,10/30/2016 05:37:24 PM,10/30/2016 05:49:18 PM,10/30/2016 05:56:46 PM,Code 2 Transport,TAYLOR ST/06TH ST/GOLDEN GATE AVE/MARKET ST,San Francisco,94103,
B03,01,1454,2,false,Non Life-threatening,1,PRIVATE,South of Market
163053810,16121567,Alarms,10/31/2016,10/31/2016 10:03:16 PM,10/31/2016 10:05:38 PM,10/31/2016 10:06:55 PM,,,Fire,POLK ST/EDDY ST,San Francisco,94102,B02,03,1643,3,false,Alarm,1,TRUCK,Tenderloin
163090308,16122938,Medical Incident,11/04/2016,11/04/2016 02:40:49 AM,11/04/2016 02:42:14 AM,11/04/2016 02:46:13 AM,11/04/2016 03:08:19 AM,11/04/2016 03:16:17 AM,Code 2 Transport,02ND ST/JESSIE ST,San Francisco,94105,B03,01,2145,3,true,Potenti
ally Life-Threatening,1,MEDIC,Financial District/South Beach
```

Filter and Transform Columns:



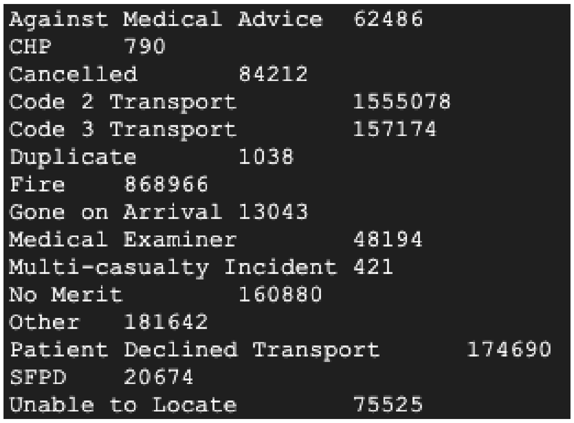
```
Call Number,Incident Number,Call Type,Call Date,Received DtTm,Response DtTm,On Scene DtTm,Transport DtTm,Hospital DtTm,Call Final Disposition,Address,City,Zipcode of Incident,Battalion,Station Area,Box,Final Priority,ALS Unit,Call Type Group,N
umber of Alarms,Unit Type,Neighborhooods - Analysis Boundaries
213132776,21137938,Medical Incident,11/09/2021,20211109,11/09/2021 06:23:34 PM,11/09/2021 06:49:41 PM,,,Patient Declined Transport,CLEMENT ST/09TH AVE,San Francisco,94118,B07,31,7135,2,false,Non Life-threatening,1,SUPPORT,Inner Richmond
213121482,21137277,Medical Incident,11/08/2021,20211108,11/08/2021 12:19:08 PM,11/08/2021 12:23:21 PM,11/08/2021 12:38:21 PM,11/08/2021 12:49:45 PM,Code 3 Transport,ANZA ST/ARGUELLO BLVD,San Francisco,94118,B07,31,7112,3,false,Potentially Life
-Threatening,1,PRIVATE,Lone Mountain/USF
213151169,21138683,Medical Incident,11/11/2021,20211111,11/11/2021 10:47:17 AM,11/11/2021 10:50:22 AM,11/11/2021 11:05:33 AM,11/11/2021 11:39:27 AM,Code 2 Transport,VALENCIA ST/14TH ST,San Francisco,94103,B02,06,5126,3,true,Potentially Life-Th
reatening,1,MEDIC,Mission
213130539,21137656,Alarms,11/09/2021,20211109,11/09/2021 07:18:42 AM,11/09/2021 07:21:23 AM,,,Fire,EDDY ST/LEAVENWORTH ST,San Francisco,94102,B02,03,1545,3,false,Alarm,1,ENGINE,Tenderloin
213132443,21137899,Medical Incident,11/09/2021,20211109,11/09/2021 04:35:52 PM,11/09/2021 04:35:53 PM,,,Code 3 Transport,ELLIS ST/POWELL ST,San Francisco,94102,B03,13,1322,3,true,Potentially Life-Threatening,1,ENGINE,Financial District/South B
each
213151333,21138709,Medical Incident,11/11/2021,20211111,11/11/2021 11:38:18 AM,11/11/2021 12:00:19 PM,11/11/2021 12:04:57 PM,11/11/2021 12:36:51 PM,Code 2 Transport,JERROLD AVE/EARL ST,San Francisco,94124,B10,17,6713,2,false,Non Life-threateni
ng,1,PRIVATE,Bayview Hunters Point
213121203,21137242,Medical Incident,11/08/2021,20211108,11/08/2021 11:51:56 AM,,,,Code 2 Transport,PLAZA ST/LAGUNA HONDA BLVD,San Francisco,94116,B08,20,8641,2,false,Non Life-threatening,1,SUPPORT,West of Twin Peaks
213151407,21138722,Medical Incident,11/11/2021,20211111,,,,,No Merit,MONTEREY BLVD/SAN RAFAEL WAY,San Francisco,94127,B09,19,8553,3,false,Non Life-threatening,1,TRUCK,West of Twin Peaks
213141329,21138224,Alarms,11/10/2021,20211110,11/10/2021 11:30:13 AM,11/10/2021 11:34:49 AM,,,Fire,FEDERAL ST/DELANCEY ST,San Francisco,94107,B03,35,2134,3,true,Alarm,1,ENGINE,Financial District/South Beach
```

Value Counts: Call Type



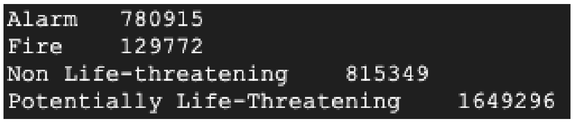
```
"Extrication / Entrapped (Machinery     636
Administrative  201
Aircraft Emergency      236
Alarms  407765
Assist Police   441
Citizen Assist / Service Call   49621
Confined Space / Structure Collapse     555
Electrical Hazard       14550
Elevator / Escalator Rescue     10676
Explosion       1062
Fuel Spill      3073
Gas Leak (Natural and LP Gases) 21610
HazMat  1176
High Angle Rescue       819
Industrial Accidents    796
Lightning Strike (Investigation)        15
Marine Fire     228
Medical Incident        2337276
Mutual Aid / Assist Outside Agency      470
Odor (Strange / Unknown)        3015
Oil Spill       5
Other   55174
Outside Fire    52995
Smoke Investigation (Outside)   7626
Structure Fire / Smoke in Building      262659
Suspicious Package      109
Traffic Collision       137564
Train / Rail Fire       129
Train / Rail Incident   896
Vehicle Fire    11310
Water Rescue    22121
Watercraft in Distress  640
```

Value Counts: Call Final Disposition



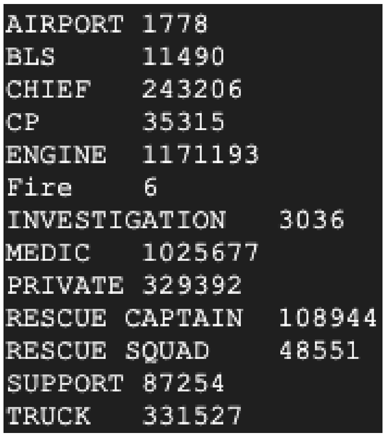
```
Against Medical Advice  62486
CHP     790
Cancelled       84212
Code 2 Transport        1555078
Code 3 Transport        157174
Duplicate       1038
Fire    868966
Gone on Arrival 13043
Medical Examiner        48194
Multi-casualty Incident 421
No Merit        160880
Other   181642
Patient Declined Transport      174690
SFPD    20674
Unable to Locate        75525
```

Value Counts: Call Type Group



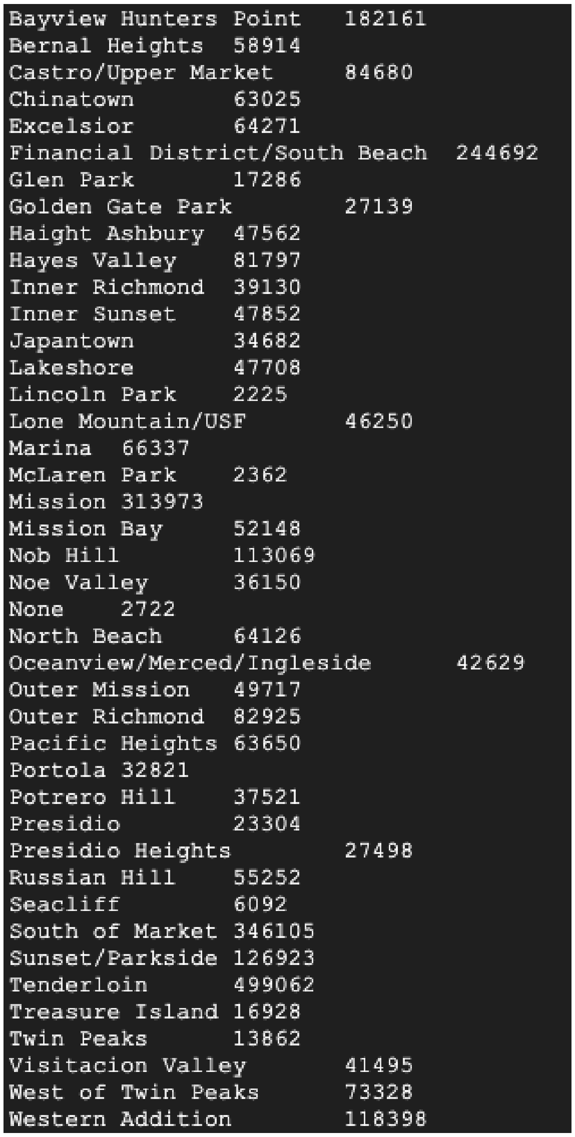
```
Alarm   780915
Fire    129772
Non Life-threatening    815349
Potentially Life-Threatening    1649296
```

Value Counts: Unit Type



```
AIRPORT 1778
BLS     11490
CHIEF   243206
CP      35315
ENGINE  1171193
Fire    6
INVESTIGATION   3036
MEDIC   1025677
PRIVATE 329392
RESCUE CAPTAIN  108944
RESCUE SQUAD    48551
SUPPORT 87254
TRUCK   331527
```

Value Counts: Neighborhooods - Analysis Boundaries



```
Bayview Hunters Point   182161
Bernal Heights  58914
Castro/Upper Market     84680
Chinatown       63025
Excelsior       64271
Financial District/South Beach  244692
Glen Park       17286
Golden Gate Park        27139
Haight Ashbury  47562
Hayes Valley    81797
Inner Richmond  39130
Inner Sunset    47852
Japantown       34682
Lakeshore       47708
Lincoln Park    2225
Lone Mountain/USF       46250
Marina  66337
McLaren Park    2362
Mission 313973
Mission Bay     52148
Nob Hill        113069
Noe Valley      36150
None    2722
North Beach     64126
Oceanview/Merced/Ingleside      42629
Outer Mission   49717
Outer Richmond  82925
Pacific Heights 63650
Portola 32821
Potrero Hill    37521
Presidio        23304
Presidio Heights        27498
Russian Hill    55252
Seacliff        6092
South of Market 346105
Sunset/Parkside 126923
Tenderloin      499062
Treasure Island 16928
Twin Peaks      13862
Visitacion Valley       41495
West of Twin Peaks      73328
Western Addition        118398
```