**Data Cleaning Documentation**

**INFM600 Section 0101**

**JOLLLY JAGUARS**

Akash Udani

Divya Kannan

Samarjith Jawaharlal Sathyanarayan

11/3/2016

**Data Cleaning Documentation**

**Data Set Characteristics:**

**Name:** Calls for Service

**Citation:**

Baltimore Police Department, (2016). Calls for Service [Calls\_for\_Service.csv]. Retrieved from <https://data.baltimorecity.gov/Public-Safety/Calls-for-Service/xviu-ezkt>, 10/10/2016.

**License:**

This data set is under the [Creative Commons Attribution 3.0 Unported License](http://creativecommons.org/licenses/by/3.0/). Under this license, the users can use, share, and modify the data without restraint, and must maintain the same for others as well. Additionally, the user must give appropriate credit to the dataset owner, and list any changes that were made from the original dataset, and include a link to the license.

License Link: <https://creativecommons.org/licenses/by/3.0/legalcode>

**Description and Metadata:**

This data set contains a list of all the 911 emergency calls made in the city of Baltimore from the year 2015, and is updated every day. This data is gathered by ‘Open Baltimore’ and the city of Baltimore, and contains almost 2 million records! For the purposes of our project, we have taken a random sample of 50,000 records for this data set.

The dataset contains no metadata, other than what is provided in the column names. We used our best judgement to find what column names represent. Here are the exact column names, and our interpretation of them:

* callDateTime (Date and time when the call was made)
* priority (Priority of the call, ex. Low, High)
* district (District codes from where the calls were made)
* description (Description of the call)
* callNumber (Unique reference ID for each call)
* incidentLocation (Street from where the call originated)
* location (latitude and longitude coordinates from where the call originated)

**Issues encountered with the data:**

1. There are missing values for a few columns, namely ‘location’ (latitude/longitude) and ‘priority’.
2. The ‘callDateTime’ column values were not standardized (ex. Some values were in 24 hour formats, some values were in AM/PM format).
3. The values for the ‘callNumber’ column were not possible to identify conclusively. (ex. ‘P152310750’).
4. The ‘callNumber’ column, (unique reference ID) skips numbers in the original data set.
5. Many of the values in the ‘description’ column were meaningless (ex. ‘\*\’ and ‘\*56’)
6. Many of the values in the ‘description’ column were misspelled (ex. ‘BURGLARPY’)
7. The values for the ‘district’ column were vague (ex. ‘SW’, ‘NE’)

**Data Cleaning Process:**

We did not remediate all the data issues we found, only the data issues that pertained to the questions we wanted to answer from this data set. We did not want to unnecessarily modify the data set.

**General Data Cleaning:**

1. The original data set consists of almost 2 million records. We downloaded the data in CSV format, and then imported it into R.

*mydata=read.csv('Final Subset Data CSV.csv')*

1. Samarjith took a randomized sample of 50,000 rows.

*subsetcalls <- mydata[sample(1:nrow(mydata), 50000, replace=FALSE),]*

‘subsetcalls’ is the name of the 50,000 subset of the original data, that our team is using for this data analysis project.

1. Only one team member (Samarjith) took a randomized sample, so that we would all have the exact same randomized sample. Then he converted the 50,000 rows sample back to xlsx format, added an index column and sent that subset to rest of the team.

Load the package ‘xlxs’. Then convert the ‘subsetcalls’ data set into xlsx format.

*library(xlsx) #load the package*

*write.xlsx(x = subsetcalls.dataframe, file = "test.excelfile.xlsx",sheetName = "TestSheet", row.names = FALSE)*

1. All 3 teammates now have the same subset of 50,000 rows, heading into the data cleaning process. From this subset, we will create different subsets to answer each question.

**Data Cleaning for Question 1:** Done by Divya

**Question 1:** Are more 911 calls related to burglary made during the night? Or is that just a preconceived notion?

To answer this question, I need to use the columns ‘callDateTime’ and ‘description’.

**‘callDateTime’ column:**

The values in the callDateTime column are not standardized. For example, some values are in 24 hour format whereas other values are in AM/PM format. Also, this question asks if more burglaries occur in the night. Hence to perform analysis based on the time, I need the time data to be separated from the date data. I formatted this column in Excel, and converted all the values into 24 hour time format. As a bonus, this formatting also removes the date, so that the fine values are only the time. The following steps were performed in Excel in CSV type data format, on the randomized 50,000 rows of data.

Steps:

1. Click on the column header (ex. ‘B’) for the column ‘callDateTime’, so that the entire column and all its rows are highlighted.
2. Right click and select ‘Format Cells’. In the pop up that opens, go to the ‘Number tab’, and select the Category as ‘Time’.
3. Select the time format ’13:30:55’ and click on ‘Ok’.
4. Now the date values should disappear, and all the time data should be in 24 hour hh:mm:ss format

**‘description’ column:**

By quickly scanning the ‘description’ column, I could see that there were many description values that came under the burglary category (ex. BURGLARY, Burglary, Robbery Armed, POSS STOLEN). The following steps were performed in R Studio, after I had imported the CSV file of the randomized 50,000 rows of data into R.

1. I installed and loaded the package ‘sqldf’, so that I can write SQL queries in R.
2. To find all the values that will come under the burglary category, I got a list of all the distinct values in the ‘description’ column.

*distinctDescriptionValues <- sqldf('SELECT DISTINCT subsetcalls.Description FROM subsetcalls')*

‘subsetcalls’ is the name of the randomized 50,000 subset that we are all using for the project.

1. Then I read through all the distinct ‘description’ values, and found 10 burglary related ones. I wanted to get the count for each of those ‘description’ values.

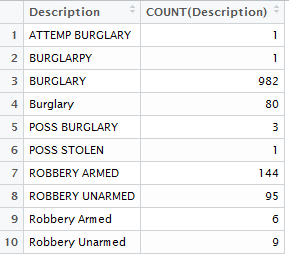
*allburglary<-sqldf('SELECT DISTINCT subsetcalls.Description, COUNT(Description)*

*FROM subsetcalls*

*WHERE Description in*

*("ATTEMP BURGLARY", "BURGLARPY", "Burglary", "BURGLARY", "POSS BURGLARY", "POSS STOLEN", "Robbery Armed", "ROBBERY ARMED", "Robbery Unarmed", "ROBBERY UNARMED")*

*GROUP BY Description')*



I calculated the total of all the burglary related categories to be 1322.

\*I also found another category for ‘Forced Entry’ and considered including it with the burglary cases, because they could be possible burglaries. However, I decided against it, because it is also possible that a forced entry is just a forced entry, and I did not want to distort the data any further.

1. Next I replaced the ‘description’ value all those 9 burglary related values to read ‘BURGLARY’. One of the values was already ‘BURGLARY’, so I did not need to replace that.

*subsetcalls$Description[subsetcalls$Description="ATTEMP BURGLARY"]<-"BURGLARY"*

*subsetcalls$Description[subsetcalls$Description="BURGLARPY"]<-"BURGLARY"*

Likewise, I modified and repeated for the rest of the 10 ‘description’ values

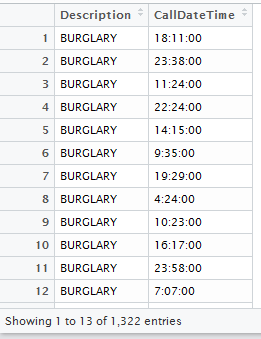
1. Now I took a count of all the ‘description’ values that are ‘BURGLARY’, and found it to be 1322, which matched with my initial calculation.

*finalburglary<-sqldf('SELECT subsetcalls.Description, COUNT(Description) FROM subsetcalls WHERE Description = "BURGLARY"')*



1. I displayed the final subset of the data that I would work with to answer Question 1:

*finalburglary<-sqldf('SELECT subsetcalls.Description, subsetcalls.CallDateTime FROM subsetcalls WHERE Description = "BURGLARY"')*



**Data Cleaning for Question 2:** Done by Samarjith

**Question 2:** Whether the number of blank calls increase/decrease with the arrival of the weekend?

To answer this question, I need to use the columns ‘callDateTime’ and ‘description’.

**‘description’ column:**

By quickly scanning the ‘description’ column, I could see that there were many description values that came under the ‘NO\_VOICE’ category (ex. ‘911/NO VOICE’, ‘911/No Voice’). The following steps were performed in R Studio, after I had imported the CSV file of the randomized 50,000 rows of data into R.

1. I installed and loaded the package ‘sqldf’, so that I can write SQL queries in R.
2. To find all the values that will come under the blank calls category, I got a list of all the distinct values in the ‘description’ column.

*distinctDescriptionValues<-sqldf('SELECT DISTINCT subsetcalls.Description FROM subsetcalls')*

1. Then I read through all the distinct ‘description’ values, and found 5 blank call related ones. I wanted to get the count for each of those ‘description’ values.

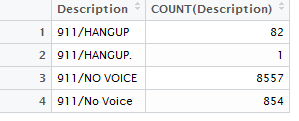
*allblankcalls<-sqldf('SELECT DISTINCT subsetcalls.Description, COUNT(Description)*

*FROM subsetcalls*

*WHERE Description in*

*(“911/HANGUP”, “911/HANGUP.”, “911/NO VOICE”, “911/NO Voice”)*

*GROUP BY Description')*



1. Next I replaced the ‘description’ value all those 4 blank call related values to read ‘NO\_VOICE’.

*subsetcalls$Description[subsetcalls$Description="911/HANGUP"]<-"NO\_VOICE"*

Likewise, I modified and repeated for the rest of the 3 ‘description’ values

1. Now I took a count of all the ‘description’ values that are ‘NO\_VOICE’, and found it to be 9494.

*sqldf('SELECT subsetcalls.Description, COUNT(Description) FROM subsetcalls WHERE Description = "NO\_VOICE"')*



**‘callDateTime’ column:**

The values in the callDateTime column are not standardized. Also, this question asks if more blank calls occur during the weekend. Hence I need only the date data from the ‘callDateTime’ column.

The following steps were performed in R Studio, after I had imported the CSV file of the randomized 50,000 rows of data into R and standardized the dates.

1. I stripped the date from the ‘callDateTime’ column using the command given below. The resulting dates were in factor format.

*subsetcalls$description = format(as.POSIXct(subsetcalls$description,format='%m/%d/%Y'),format='%m/%d/%Y')*

1. I converted the dates from factor format to date format.

*subsetcalls$description =as.Date(subsetcalls$description, format = '%m/%d/%Y')*

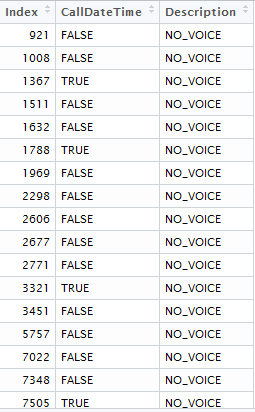
1. I wanted to separate the weekends from the weekdays, so I imported the ‘timeDate’ package.

*library(timeDate)*

1. I singled out the weekends, and by using the below R command

*subsetcalls$description =isWeekend(subsetcalls$description)*

1. Now, the ‘callDateTime’ column contains only ‘True’ (for weekends) or ‘False’ (for weekdays) for all the 50,000 rows.
2. I then created a subset of only the ‘callDateTime’ and ‘description’ columns. I then furthered trimmed it by considering only the rows that had a description of ‘NO\_VOICE’, thus making it easier to address this question



**Data Cleaning for Question 3:** Done by Akash

**Question 3:** Find the top 5 dangerous reasons for calls (e.g. Aggravated assault). Plot them based on the locations.

* From which location (street) did most of these calls originate? Can we find the most dangerous locality in Baltimore based on the 911 calls?
* Is there a pattern for each of these calls with respect to their proximity to the coast? Do these calls increase in frequency when you get closer to the coast?

While planning the data cleaning process, we realized that this question will be hard to implement. We initially planned to use the ‘location’ column with the coordinates, but we realized that this column had many rows without any coordinates. We could not fill in the values, and we could not delete those rows either because there were around 1700 rows out of the 50,000 row subset that had no coordinates.

Hence, instead of trying to identify neighborhoods using coordinates, we changed it to identifying districts using the ‘district’ column. The district column does not have any empty rows.

To answer this question, I need to use the columns ‘district’ and ‘description’.

**‘district’ column:**

This column does not need any data cleaning. The 9 distinct values of the ‘district’ column represent the district values represent the 9 districts of Baltimore. A Google search helped us match the district codes with their names.

For example, (NE – North East, NW- North West, ND – North District, SE -South East, SW – South West, SD – South District, ED – East District, WD – West District, CD – Central District).

**‘description’ column:**

1. Installed and loaded the package ‘sqldf’, so that I can write SQL queries in R.
2. Then we went through each distinct description manually to select the top five dangerous reasons for calls.

*distinctDescriptionValues<-sqldf('SELECT DISTINCT subsetcalls.Description FROM subsetcalls')*

We found the top 5 dangerous types of calls to be: Assault, Destruction of Property, Discharge Firearm, Narcotics, and Armed Robbery.

1. For each selected reason, we found different variations of the same description label (different cases and spellings). We merged description labels as shown below.

|  |  |  |
| --- | --- | --- |
| **Description Labels** | **Count** | **Common Label** |
| COMMON ASSAULT | 2105 | Assault  Total: 2635 |
| Common Assault | 160 |
| AGGRAV ASSAULT | 349 |
| Aggrav Assault | 21 |
|  |  |  |
| DESTRUCT PROP | 143 | Destruction of property  Total: 795 |
| DESTRUCT PROPTY | 597 |
| Destruct Propty | 55 |
|  |  |  |
| DISCHRG FIREARM | 128 | Discharge Firearm  Total: 241 |
| Dischrg Firearm | 9 |
| SHOOTING | 101 |
| Shooting | 3 |
|  |  |  |
| NARCOTICS INSIDE | 242 | Narcotics  Total: 2086 |
| Narcotics Inside | 34 |
| NARCOTICSOutside | 1517 |
| NarcoticsOutside | 106 |
| NARCOTICS ONVIEW | 167 |
| Narcotics OnView | 20 |
|  |  |  |
| ROBBERY ARMED | 144 | Robbery  Total: 150 |
| Robbery Armed | 6 |

1. Next, I standardized all the misspellings and different labels for each top 5 dangerous type of call.

*subsetcalls$Description[subsetcalls$Description="COMMON ASSAULT"]<-“ASSAULT"*

Likewise I standardized the labels for the rest of the ‘description’ values in the chart.

Word Count: 2067