Analysis of Emergency - 911 Calls from Montgomery County, PA

Dhruv Jinesh [1]

[1]Department of Computer Science, Rochetser Institute Of Technology

ABSTRACT

9-1-1 is an USA emergency helpline number for public safety. The 911 system helps the citizen connect to an operator that can relay the emergency to the dispatch center. The operator identifies the address of the caller through the phone company database and connects the caller to the public safety service in their area. Over 240 million calls are made to 911 in the U.S. each year [3]. The 911 calls should be triaged, and responded to based on the urgency of the caller situation. By analyzing the 911 database we seek to find patterns that will help efficiently process calls. Here we work with the data for Mongomery County, PA. The model built using this technique helps the 911 operators anticipate the occurrence of emergencies and effectively respond to calls with appropriate actions.

Keywords

Pre-processing, Feature selection, Decision Tree, Data Visualization, 911 Dataset

1. INTRODUCTION

The 911 calls are the people's last hope of survival during emergency. Before 911 system was introduced, each public safety organization had a different help line number that citizen had to call during emergency. It was very hard to remember what number should be used for what type of emergency. Many people did not have enough knowledge about the different public safety services. It would take an significant amount of time for citizens to get in touch with all these safety agencies, thus causing in delay which could cost their life during emergency. The present 911 call system overcomes the drawback of traditional system. With the 911 system, the citizen has to remember and use only one number during all type of emergency. This system has been proven very effective in the past. However, with more and more people using this system, it has becoming difficult to effectively process each citizen call with minimum response time. Also, the 911 system requires more resources to process this calls.

Local police departments have limited number of patrol cars and other resources are allocated to respond to the emergency calls every day for each area. Some resources are over-utilized and under-utilized in different areas. If the police department distribute more resources to the area which has maximum calls then many calls can be saved. For instance, areas that have many medical emergency calls should have adequate emergency medical service resource. This is only possible when they would have some prior knowledge about what calls they can expect from this area, what is peak time for this calls, type of calls, etc. This type of information can be found out by performing data mining techniques on the past 911 call history data of all areas. In this paper we use 911 emergency call data compiled from Montgomery County in Pennsylvania. The results obtained from data mining techniques will help the public safety services to effectively distribute the resources among different areas based on their needs.

The paper is divided into several parts: a literature survey, data exploration and data visualization, a model developed to predict category of 911 call, results and conclusions.

2. MOTIVATION

The importance of the 911 calling system is not only in the movies. The 911 call center is an important resource for every community in America, and other places around the world with similar call centers. The 911 call center database for Montgomery County was selected as an example of real-world data. Data analytics of this would be helpful and beneficial for communities [1].

A good understanding of expected 911 calls might help with resource allocation for 911 call centers. This not helps local communities and also might help reduce the stress of the 911 operators, who suffer from significant employee turnover.

3. RELATED WORK

The three-digit telephone number 9-1-1 is USA emergency helpline number for public safety which was introduced in 1968. This number can be used to call fire-department, the police or an ambulance. The calls are placed by a Voice over Internet Protocol ("VoIP") End-User to Carrier's Internet Protocol ("IP") services[6]. It is difficult to handle all types of emergency calls due to lack of resources and lack of data. Recently, a lot of analysis is done on 911 emergency calls to find patterns which can help to process the user's request with minimum response time. Data mining techniques on 911 data can help to identify areas with high volume of accidents, health issues, etc. With this knowledge, resources can be allocated accordingly. In one study [5], the focus was to improve the communication system of 911 by installing enhanced technology, improving call-handling performance, improving civilian pay equity and so on. It did not try to understand the reasons of many 911 from an area but focused on how to process those calls. So, it failed to provide

*jad6566@g.rit.edu

a solution for all situations. In another study [4], a better technique was used which used linear regression model to predict number of emergency calls over a period. In this technique, when total number of emergency calls exceeds the threshold (predicted number) an alarm is raised that indicates more resources are needed to process emergency calls. This technique does not consider the location spatial context. The paper[6], used k-means for clustering the dataset. Hot-spot analysis was used for data visualization and recommendation for reallocation of emergency service force. The main challenges all the papers faced was to find what factors like location, time, type of emergency contribute the most to efficiently process 911 calls. The collected data also had missing values for location and reason of emergency call.

4. SYSTEM ARCHITECTURE

The 911 dataset was collected form kaggle[2]. The dataset consists of 188752 entries. This data is collected between 2015-12-10 to 2017-04-13. The data consist of latitude, longitude, description, zip code, title, timestamp, township, address and dummy variable. The attribute name and data type is [2]: lat : String variable, Latitude - lng: String variable, Longitude - desc: String variable, Description of the Emergency Call - zip: String variable, Zipcode - title: String variable, Title - timeStamp: String variable, YYYY-MM-DD HH:MM:SS - twp: String variable, Township - addr: String variable, Address - e: String variable, Dummy variable (always 1). Refer to the table 1 for attributes of the dataset.

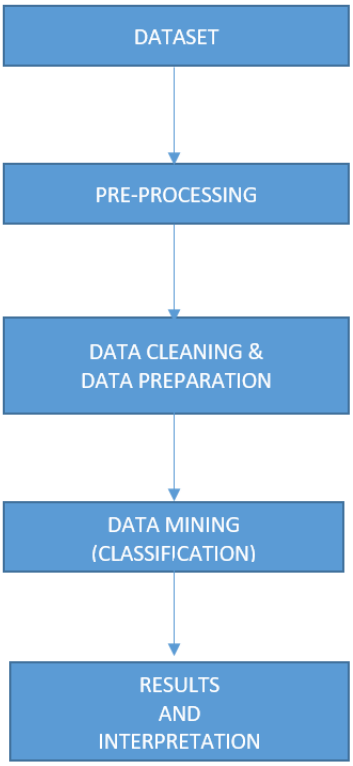


Figure 1: Steps Taken During Data Mining

Table 1: Attributes of Dataset

Attribute Name	Description
Latitude	The Latitude of the call location
Longitude	The Longitude of the call location
Description	Description of the call
Zip	Zip code of the call location
Call Type	Type of 911 call
Time Stamp	Time stamp of the call
Township	Area where the call originated
Address	Address of the call

There is no ethical consideration with this dataset. The collected dataset does not consist of any information that can be used to uniquely identify the user who made that emergency call. Also, the dataset contains only information about reason of emergency call, timestamp, etc. and no information about who made that call. However, if such information can be found then it is a violation of the user's privacy.

In pre-processing stage, dummy variable were removed and time-stamp attribute was converted into proper format. the major challenges involved were categorizing emergency calls based on generic types and cleaning the description of emergency call attribute to extract information. The other challenges include: Dealing with missing values for zip code, removing dummy variable, find city information from latitude longitude, handling the outliers and extract time and data from time stamp. In data cleaning preparation step, missing values were filled, time & date extraction, categorizing the type of call and other issues were handled.

The detail description of the above two steps is as follows: I extracted time and date from the time stamp attribute and created columns of Hours, Min, Day, Month and Year for detailed time series analysis at granular level. Further, I did text analysis on description attribute to find different categories to classify the data based on category. I also removed dummy variable present in the dataset. The data set consist of missing values for attribute zip code.The zip code attribute had 22000 missing values. To fill missing values, 1. I created a dictionary where key is township attribute and value is a dictionary of zipcodes associated with that township area. The zip code dictionary consists of zipcode as key and value as frequency of zipcodes. 2. Find the maximum frequency of zipcode for each township area. 3. For each empty zipcode, check its township area and use the township dictionary to get the zipcode. 4. If township area is null, use google map web service to get the zip code using latitude and longitude attribute.

Lastly, I correctly categorized the type of emergency call (reason for call). So, I have created a clean version of the data and now this file will be used for analysis.

Once the data is cleaned and prepared, I plotted graphs for visualization of hour of call versus day of call, hour of call versus place of call and so on. The trends and patterns in 911 calls related to time, month, reason for call and location of the call using data mining techniques were discovered from graphs. I used different classification algorithms on the dataset. The data was divided into training, testing and validating dataset. The classification algorithm which gives the best accuracy and minimum node impurity measures on all the 3 datasets was selected. I used Naive-Bayes, random-forest and decision stump classification algorithms to predict the type of emergency call based on area, hour, month day attribute. I also found, what type of emergency calls dominate each area, what is the prime time for emergency calls for each area, which type of emergency calls dominate the most on daily basis, how different are emergency calls from 2015 to 2016 and so on. One objective was to predict the type of emergency call based on the area and time at which most emergency calls can be expected with respect to an area.

5. DATA MINING MODEL

Classification is supervised learning with a goal to accurately predict the target class for each case in the data. For this dataset, target variable is the category of emergency call. Classification yields good results if relevant set of attributes are selected to predict target variable. Correlation based Feature Selection technique was used to find relevant attributes. The most relevant features found using this technique were attribute Hour and attribute township. Further, there was a problem of imbalance class value distribution. Due to which I was getting low accuracy using Naive Bayes. So I used re-sampling technique which uniformly distributes the class values. Also, Township attribute contained missing values. Google maps API was used to find missing values using latitude and longitude attribute. Further, I divided the dataset into 60% training set, 25% testing set and 15% validating set. Then Naive Bayes, Decision Table and Random Forest technique were used to predict the target variable. Accuracy of Naive Bayes was 50%, Decision Table was 56 % and Random Forest was 59%. The result also suggested that target variable depends majorily on Hour attribute and then on township attribute. Below is the results of Random Forest.

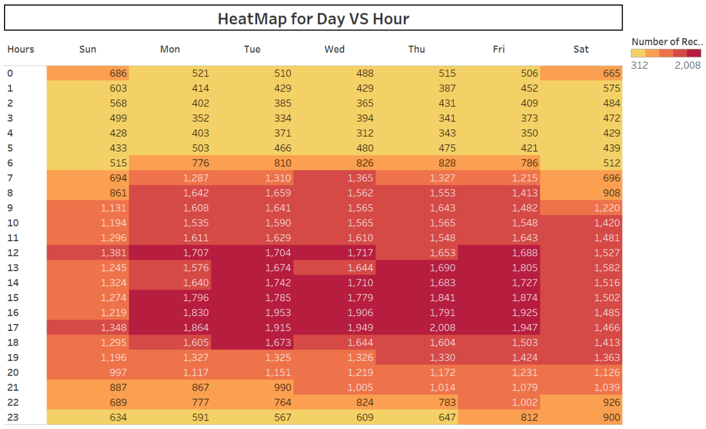
```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 11.23 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      111956               59.3515 %
Incorrectly Classified Instances     76676               40.6485 %
Kappa statistic                          0.2937
Mean absolute error                      0.3355
Root mean squared error                  0.413
Relative absolute error                 83.0309 %
Root relative squared error             91.8745 %
Total Number of Instances           188632

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.703    0.398    0.610      0.703    0.653      0.305    0.724     0.708     EMS
                 0.151    0.022    0.527      0.151    0.234      0.229    0.756     0.359     Fire
                 0.618    0.291    0.579      0.618    0.598      0.323    0.739     0.648     Traffic
Weighted Avg.    0.594    0.304    0.586      0.594    0.574      0.302    0.735     0.637

=== Confusion Matrix ===

     a     b     c   <-- classified as
 62282  2083 24253 |     a = EMS
 12942  3901  9066 |     b = Fire
 26911  1421 45773 |     c = Traffic
```

Figure 2: Random Forest

6. RESULTS & INTERPRETATION

This section contains different types of visualization done to get better understanding of the dataset, slice dice of dataset to find hidden patterns and detail interpretation of all the graphs.

The figure 3 is a heatmap plotted of day versus hour. From the graph, we can see that most of the 911 calls occur during weekdays between 3pm to 5pm. Also, calls are highest between 10am to 5pm and call rate goes down. We also see that there is an increase in calls past 11:00 PM on Friday and Saturday nights.


HeatMap for Day VS Hour

Hours	Sun	Mon	Tue	Wed	Thu	Fri	Sat
0	686	521	510	488	515	506	665
1	603	414	429	429	387	452	575
2	568	402	385	365	431	409	484
3	499	352	334	394	341	373	472
4	428	403	371	312	343	350	429
5	433	503	466	480	475	421	439
6	515	776	810	826	828	786	512
7	694	1,287	1,310	1,365	1,327	1,215	696
8	861	1,642	1,659	1,562	1,553	1,413	908
9	1,131	1,608	1,641	1,565	1,643	1,482	1,220
10	1,194	1,535	1,590	1,565	1,565	1,548	1,420
11	1,296	1,611	1,629	1,610	1,548	1,643	1,481
12	1,381	1,707	1,704	1,717	1,653	1,688	1,527
13	1,245	1,576	1,674	1,644	1,690	1,805	1,582
14	1,324	1,640	1,742	1,710	1,683	1,727	1,516
15	1,274	1,796	1,785	1,779	1,841	1,874	1,502
16	1,219	1,830	1,953	1,906	1,791	1,925	1,485
17	1,348	1,864	1,915	1,949	2,008	1,947	1,466
18	1,295	1,605	1,673	1,644	1,604	1,503	1,413
19	1,196	1,327	1,325	1,326	1,330	1,424	1,363
20	997	1,117	1,151	1,219	1,172	1,231	1,126
21	887	867	990	1,005	1,014	1,079	1,039
22	689	777	764	824	783	1,002	926
23	634	591	567	609	647	812	900

Figure 3: Heatmap of Day versus Hour

The figure 4 shows graph of Hours versus total no. of calls. From the graph, we can see that most of call occurs between 10 am to 5pm. This could be because lot of people travel and work at this time.

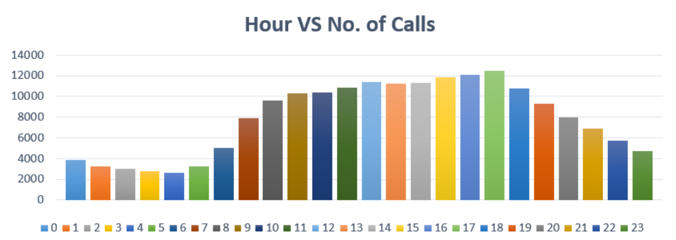
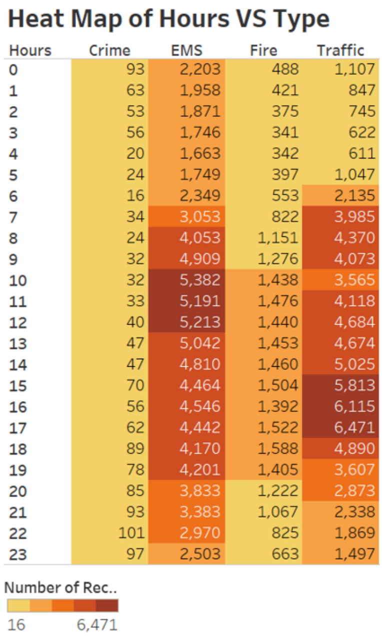


Figure 4: Heatmap of Hour versus No. of calls

Now, let us understand what type of calls are made during this peak time. The figure 5 shows supports our previous idea that lot of people travel and work between 10am to 5pm. So, there are more chances of having traffic accidents and medical emergency compared to other time. This can be seen from figure 5.

In previous graph we talk about 911 calls made during which hour, day and type of call. Now, lets have a look 911 call from monthly point of view. This graph will help us understand whether is their any specific month when 911 calls are highest or lowest. The figure 6 shows graph for year 2016 of month versus number of calls. This graph helps us understand monthly variations. From the graph, we can see that January month had the maximum number of calls. It was found from external source that on Jan 22nd to Jan 24th, 2016, there was an exceptionally heavy snow storm.



Heat Map of Hours VS Type

Hours	Crime	EMS	Fire	Traffic
0	93	2,203	488	1,107
1	63	1,958	421	847
2	53	1,871	375	745
3	56	1,746	341	622
4	20	1,663	342	611
5	24	1,749	397	1,047
6	16	2,349	553	2,135
7	34	3,053	822	3,985
8	24	4,053	1,151	4,370
9	32	4,909	1,276	4,073
10	32	5,382	1,438	3,565
11	33	5,191	1,476	4,118
12	40	5,213	1,440	4,684
13	47	5,042	1,453	4,674
14	47	4,810	1,460	5,025
15	70	4,464	1,504	5,813
16	56	4,546	1,392	6,115
17	62	4,442	1,522	6,471
18	89	4,170	1,588	4,890
19	78	4,201	1,405	3,607
20	85	3,833	1,222	2,873
21	93	3,383	1,067	2,338
22	101	2,970	825	1,869
23	97	2,503	663	1,497

Number of Rec..
16 6,471

Figure 5: Heatmap of Hour versus Type of Call

This was one of the main reason for many calls during those days.

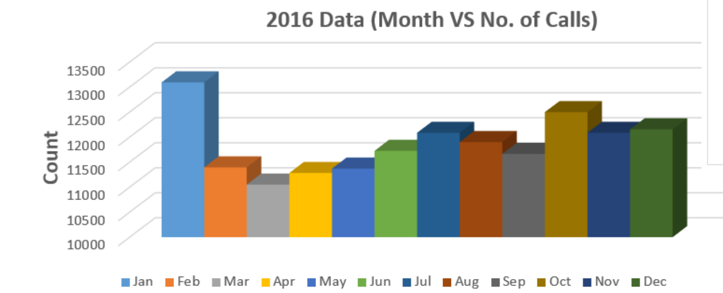


Figure 6: 2016 (Month versus Number of Call)

The figure 7 shows type of calls versus number of calls for Jan 2016. This graph supports the idea that due to snow storm lot of traffic accidents injuries took place which resulted in more 911 calls.

The figure 8 shows the comparison of top 911 call reason for 2016 versus 2017 for the first 3 months. This graph tells us that the calls in 2017 have been reduced. One of the reason could be that there was no natural calamity like snowstorm in 2017.

We break down the 911 calls by locale in the interest of data exploration. Now we will have have look at the total number of 911 calls distributed among different areas in Montgomery County, PA and what type of calls were made in those areas. The figure 9 shows top three township versus type of the 911 call. This will help us to understand which area has maximum number of calls, what is the type of those calls, etc.

Figure 10 shows top 10 township call type distribution. This is another way of looking at the graph to understand

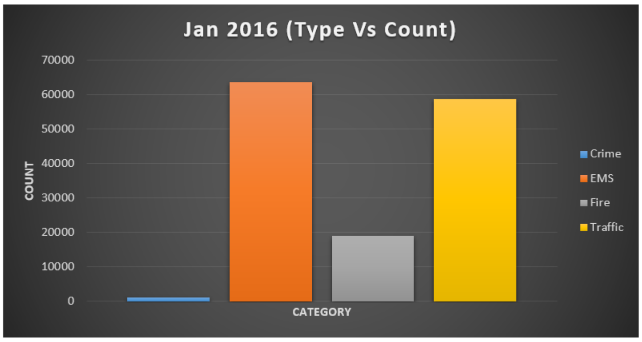


Figure 7: 2016(Type of call versus Number of call

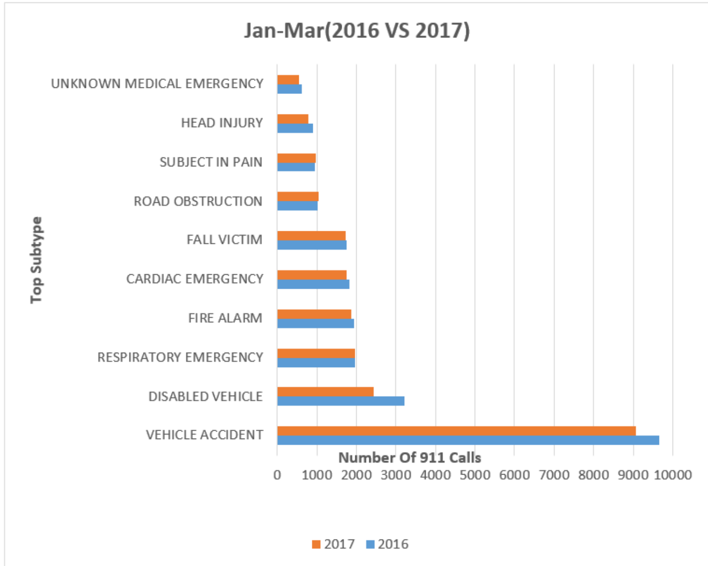
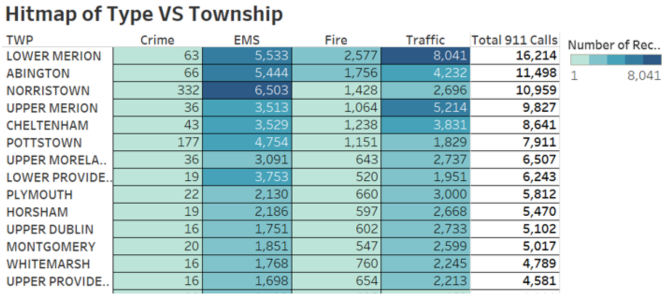


Figure 8: Jan-Mar(2016 versus 2017)

which area has what type of majority calls. From figure 9 and 10 we can interpret that majority number of 911 calls are made from Abington, Lower Merion and Norristown area.

The above graph showed 911 call distribution based on area and what area had maximum number of 911 calls. To get more deeper understanding of this area specific calls I selected top 3 township areas which had maximum number of 911 calls. The figure 11 shows type distribution of top 3 township. The top three township considered are Abington, Lower Merion and Norristown. To compare the 3 township based on type of call, we first have to normalize them based on population density. Normalization formula: Normalized value = ((count of each type / population density) / total count of all type) * 100. Based on this normalization we can now compare this 3 townships. From the graph, we can see that Norristown had approximate 60% of EMS calls which is the highest percentage among all 3 township. Lower Merion had 50% of traffic calls. This helps us understand which type of calls predominate in which area.

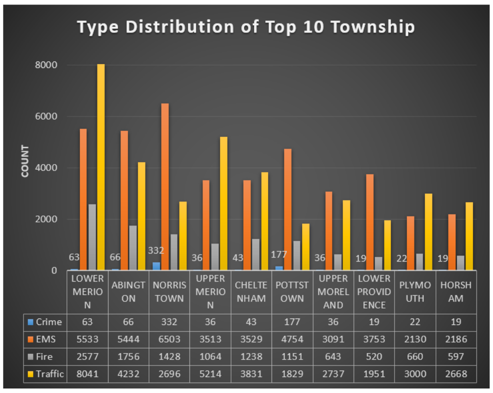
The next two graphs will go deeper on these top 3 township to find more insights. The figure 12 shows EMS: top 17 reason versus call count for the same top 3 townships. Similar normalization technique is used like in previous step.



Hitmap of Type VS Township

TWP	Crime	EMS	Fire	Traffic	Total 911 Calls
LOWER MERION	63	5,533	2,577	8,041	16,214
ABINGTON	66	5,444	1,756	4,232	11,498
NORRISTOWN	332	6,503	1,428	2,696	10,959
UPPER MERION	36	3,513	1,064	5,214	9,827
CHELTENHAM	43	3,529	1,238	3,831	8,641
POTTSTOWN	177	4,754	1,151	1,829	7,911
UPPER MORELA..	36	3,091	643	2,737	6,507
LOWER PROVIDE..	19	3,753	520	1,951	6,243
PLYMOUTH	22	2,130	660	3,000	5,812
HORSHAM	19	2,186	597	2,668	5,470
UPPER DUBLIN	16	1,751	602	2,733	5,102
MONTGOMERY	20	1,851	547	2,599	5,017
WHITEMARSH	16	1,768	760	2,245	4,789
UPPER PROVIDE..	16	1,698	654	2,213	4,581

Figure 9: Top township versus Type of Call



	LOWER MERION	ABINGTON	NORRISTOWN	UPPER MERION	CHELTENHAM	POTTSTOWN	UPPER MORELAND	LOWER PROVIDENCE	PLYMOUTH	HORSHAM
Crime	63	66	332	36	43	177	36	19	22	19
EMS	5533	5444	6503	3513	3529	4754	3091	3753	2130	2186
Fire	2577	1756	1428	1064	1238	1151	643	520	660	597
Traffic	8041	4232	2696	5214	3831	1829	2737	1951	3000	2668

Figure 10: Type distribution of top 10 Township)

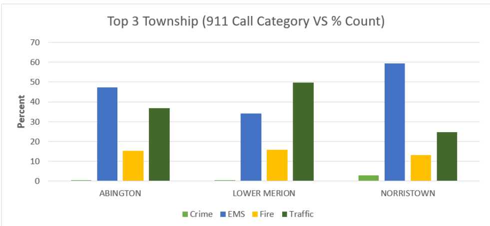


Figure 11: Type distribution of top 3 township

This graph tells us that what was the exact EMS reason for each call and which reason dominates. Norristown and Abington has highest number of respiratory emergency calls. This could be due to pollution, smoking prone area, old age or some other reason. This insight can help to understand that more measures need to be taken to overcome respiratory problems. More insights can be found out from this graph.

Now, let's have a look at top call reasons for this 3 areas. The figure 13 shows top 10 reason for calls for this 3 town. The reason count are again normalized based on population density and percentage ratio. It is clear from the graph that vehicle accident dominates the reason for call. The second most reason is fire alarm in Abington and Lower Merion which is strange. Again, more insights can be found by interpreting this graphs and by comparing 3 townships results.

Since traffic calls are very high among all township. Lets have a look at what time this traffics calls happens in Abington, Lower Merion and Norristown. The figure 13 shows top

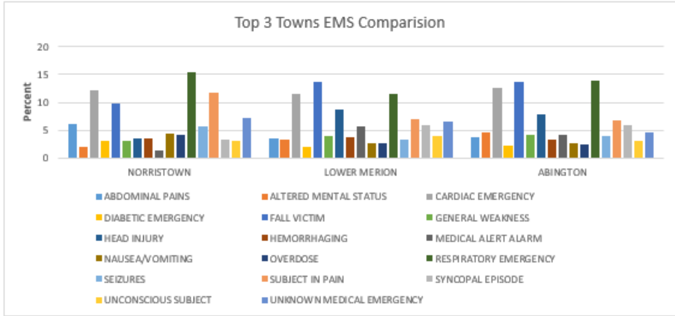


Figure 12: Top 17 EMS reason of top 3 township

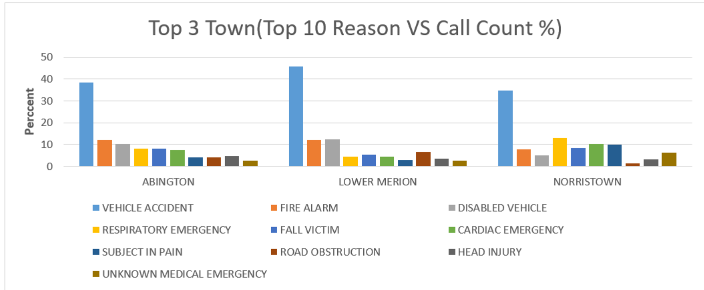


Figure 13: Top 10 call reason of top 3 township

3 township accident call comparison analysis by hour(24-hour format). The y-axis indicates the normalized percentage of total number of traffic calls received at any given time. The total number of calls are normalized based on population density of the region where the call was made and converted into percentage by dividing it by total normalized traffic calls made in that region multiplied by 100.

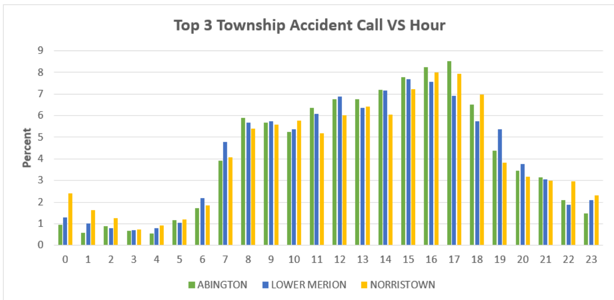


Figure 14: Accident calls made in every hour in top 3 township

Also EMS calls are very high among all township. Lets have a look at what time this EMS calls happens in Abington, Lower Merion and Norristown. The figure 14 shows top 3 township EMS call comparison analysis by hour(24-hour format). The y-axis indicates the normalized percentage of total number of traffic calls received at any given time. The total number of calls are normalized based on population density of the region where the call was made and converted into percentage by dividing it by total normalized EMS calls made in that region multiplied by 100.

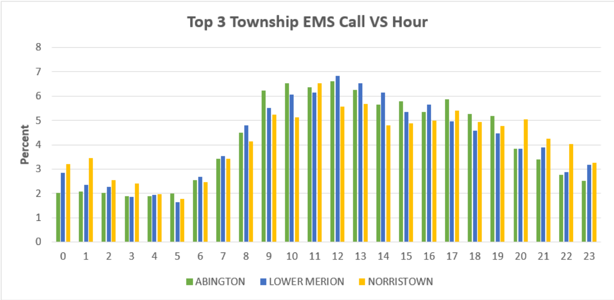


Figure 15: EMS calls made in every hour in top 3 Township

7. FUTURE WORK

In the future, we would perform clustering techniques and plot geographical graphs to get more insights.

8. CONCLUSION

This paper provides a classification model which analyzes the 911 dataset. The various graph plots help to understand the patterns of 911 calls. The heatmaps and other visualization graphs helped to understand that majority of 911 calls are made during 3pm to 5pm, calls are more on weekday compare to weekends, Traffic accidents are maximum during 3pm to 5pm; Abington, Lower Merion and Norristown are the area from which maximum 911 calls are received, EMS calls are the most during 9am to 2pm for this towships. The results of this paper can help the public safety services to make strategic decisions about how to fairly allocate fire, medical and police resources. This will result in shorter response time of public safety services to help their citizens thereby saving many lives. Further, the 911 data-set majorly contained 2016 data, if more past year data would have been shared then better insights could have been found. The main limitation of this paper is that it does not perform clustering on categorical and numerical values for more detail analysis.

9. REFERENCES

[1] http://newsok.com/article/2823546.

[2] https://www.kaggle.com/mchirico/montcoalert/version/40.

[3] https://www.nena.org/?page=911statistics.

[4] H. Jasso, T. Fountain, C. Baru, W. Hodgkiss, D. Reich, and K. Warner. Prediction of 9-1-1 call volumes for emergency event detection. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, pages 148–154. Digital Government Society of North America, 2007.

[5] P. Kuhn and T. P. Hoey. Improving police 911 operations in washington, d.c. *National Productivity Review*, 6(2):125–133, 1987.

[6] A. Selvam, B. Thiagarajan, and T. T. K. Mining patterns from 9-1-1 calls dataset. *International Journal of Applied Information Systems*, 11(7):35–40, Dec 2016.