horizontal line

INFO 208 - BIG DATA TECHNOLOGIES

**MINI - PROJECT**

ANALYSIS OF 911 CALLS DATA

**Rajasree Rajendran**

horizontal line

**TABLE OF CONTENTS**

Overview…………………………………………………………………………………………………. 3

Why the subject matter is important to me……………………………………………………………….... 3

What am I hoping to achieve……………………………………………………………………………….4

Dataset Studied…………………………………………………………………………………………….4

Description and structure of Data………………………………………………………………………….5

Notes on exploration of the dataset………………………………………………………………………...6

Questions to ask of the dataset…………………………………………………………………………….6

Summary………………………………………………………………………………………………….10

What I learned about this dataset………………………………………………………………………….10

What I learned from this mini-project…………………………………………………………………….12

Being a Data Scientist & Skills I might need for the future……………………………………………….13

References…………………………………………………………………………………………………15

**Overview**

“It is a capital mistake to theorize before one has data”, says Arthur Conan Doyle, the writer of Sherlock Holmes. Data Science have been around since a long time, even though the name was coined only recently. The uses and applications of Data Science are beyond imagination, and the world cannot function without data. Data is used in almost all fields and in our day-to-day lives. Life moves along, sometimes smooth and sometimes taking twists and turns. There are instances where there are sudden emergencies occur in lives, where people are desperate for help. Those instances are mostly met by 911 calls, an amazing system in place in the United States.

**Why the subject matter is important to me:**

## Being in America, the one thing which continues to fascinate me is the emergency response here. Coming from a country where emergency responses are mostly not as efficient, the speed, convenience and efficiency of the 911 emergency calls are remarkable. Having personally suffered a loss of a dear one due to the lack of a proper emergency system in my country, the 911 has always piqued my interest. I have always wondered about the amount of calls the system gets in a day and the way the network has been organized in order to attain maximum competence. According to Wikipedia [1], 911 is an emergency telephone number for the North American Numbering Plan (NANP) and is used only for emergency purposes. 911 service can be accessed from over 98% of locations of United States and Canada, and the calls are linked to an emergency dispatch office. For this project, I decided to find open source data about 911 calls and analyze the data, in order to get an idea of the system.

## What I am hoping to achieve:

## I hope to understand the amount and volume of calls made to 911 in a year, and mostly in 2018, and also to understand the nature of the calls. I would also like to understand if there are any patterns hidden in the data, and thus try to understand whether the emergency response team is as efficient as I believe from what very little I know. Being in such a vast ocean of data, the initial step according to Brian Godsey in a Data Science project, is preparing and gathering data and knowledge. According to him, a Data Scientist should be an explorer, since the availability of data everywhere makes Data Science an entity worth of study and exploration. I explored myself and came across an open source data from Kaggle.

## Dataset Studied

In order to find open source 911 call data, I checked Kaggle.com, and came across the following dataset:

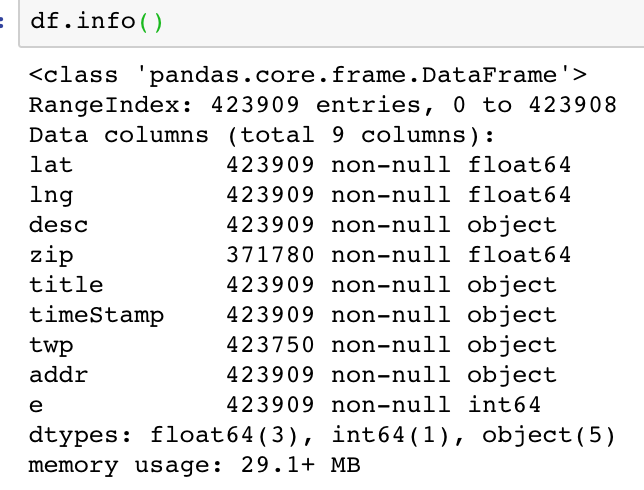
<https://www.kaggle.com/mchirico/montcoalert>

This dataset includes the 911 call data of Montgomery County in Commonwealth of Pennsylvania. As of July 2017, Montgomery County (MontCo) has an estimated population of 826,075 and is the third most populous County in the United States and Commonwealth of Pennsylvania. By choosing this dataset, it helps me identify the pattern of emergency calls made in a very populous place.

**Description & Structure of the Data:**

The data is in *csv* format and the size of the file is 74.87 MB. The dataset has a total of *10 columns*, and *423909 entries*. In order to get a very vivid idea of the dataset, after downloading, I created a new Jupyter notebook and used Python commands to get the structure of the data. The following are the 10 columns of the dataset:

* ***lat***: Latitude value of the caller - data type: Float64
* ***lng***: Longitude value of the caller - data type: Float64
* ***desc***: Description of emergency - data type: object
* ***zip***: Zip code of the caller - data type: Float64
* ***title***: Title of emergency - data type: object
* ***timeStamp***: Date and time of the call - data type: datetime64
* ***twp***: Town of the caller - data type: object
* ***addr***: Address of the caller - data type: object
* ***e:*** Dummy variable which is always 1 - data type: int64
* ***year***: year of call - data type: int64



**Notes on Exploration of the Dataset:**

According to ‘Think like a Data Scientist’, “Uncertainty is one of the principle characteristics of data science work. If there were no uncertainty, there would be no exploration and there would be no problem solving” (Godsey,1998, p.7). Here, the availability of dataset of the 911 calls concurs with the first step of data science process - preparing the data. Now that the data is obtained, it is time to explore and assess the data and to see whether the data is relevant and sufficient. In order to understand this, I applied my findings from Godsey’s book: understand how the data can be used to answer my questions/ attain my goals of this project. My initial goal is to understand the volume of calls in a year, specifically 2018 and also assess the nature of calls in specific regions. This would help to analyze whether the responders are equipped well enough to deal with the emergencies. For instance, if there have been more calls from a particular zip code for vehicle accidents, it means the area should be well equipped with ambulance services. In order to answer these questions, I first decided to do some descriptive statistics of the available data.

**Questions to Ask of the Dataset:**

Keeping in mind the goals of my project, I decided to work on the following specific statistics:

* What is the total number of 911 calls in the year of 2018?
* What are the categories of calls which has the greatest number of entries in 2018?
* Which zip code has the largest number of calls made to 911?
* Which day and time has the greatest number of calls?
* Has there been a decrease or increase of the emergency calls since the previous year?

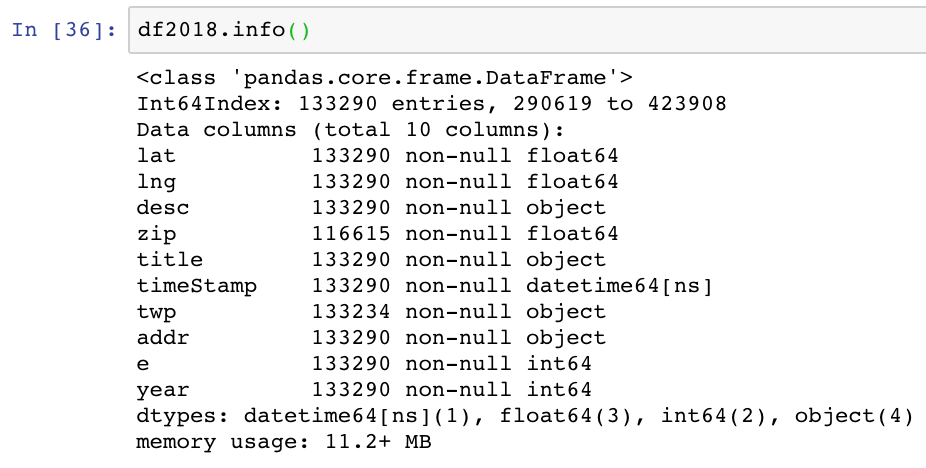
For finding out the answers, I initially used Python in Jupyter Notebook to extract the data from only year 2018. The dataset originally has data from years 2015 to 2018. In order to extract the year 2018, I converted timeStamp format to datetime format using the *pd.to\_datetime ()* function, extracted year to a column and created a new dataframe for 2018, using the following commands:

**df['timeStamp'] = pd.to\_datetime(df['timeStamp'])**

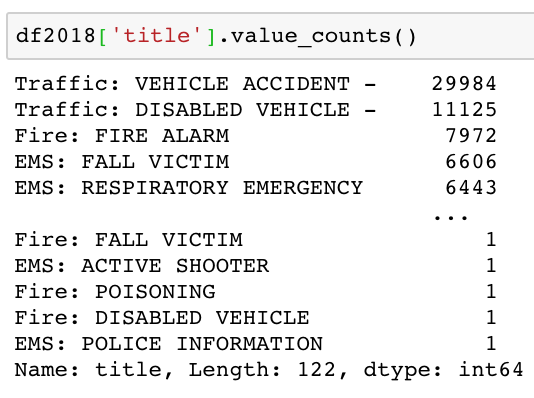
**df['year'] = df['timeStamp']. apply (lambda x: x.year)**

**df2018 = df[df['year'] == 2018]. copy ()**

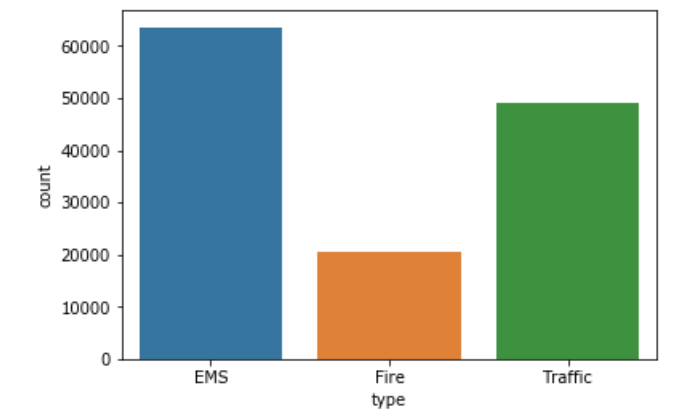
* In order to confirm that everything went as expected, I used *df2018.info ()* function and found that year 2018 has **133290** entries. For further clarity, *df2018.head()* was used to display the first five entries of dataframe. The last five entries were also displayed using *df2018.tail()* to confirm the entries are only from 2018.

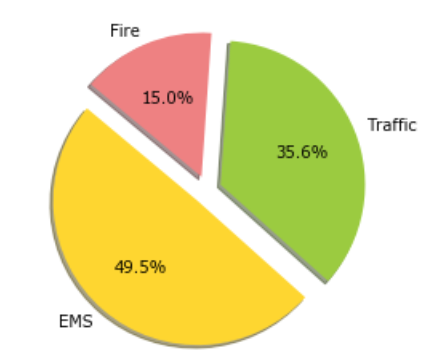


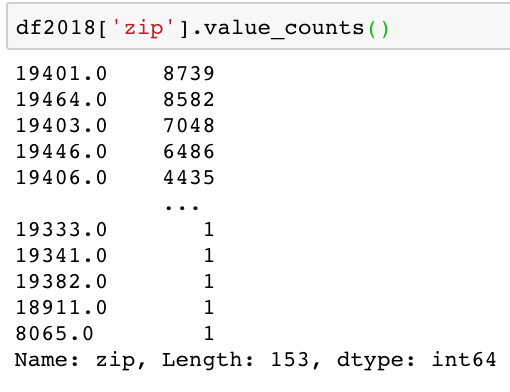
* Now that the first question has been answered, I moved on the explore about the categories of calls in 2018. Since the column ‘title’ has category/title of the emergency, using *value\_counts ()* function, I extracted the types of emergencies in 2018 dataframe. From the results, it was seen that Vehicle Accidents have the greatest number of entries and least number of entries were for *Active shooter, Fall victim, Poisoning, Disabled vehicle* and *Police information.*



* Moving on to the next goal of finding the zip code with largest number of calls made in 2018, I used the *value\_counts ()* function again on column *Zip* and realized that the Zip code **19401** has made the largest number of calls (**8738 calls**). For finding out the day and time that has the greatest number of calls, the timeStamp was split into date, month, hour and day of the week. Based on the split results, a plot was performed to identify the day of the week and time with the greatest number of 911 calls, they were Monday and times were 11.00 am and the time between 2.00 to 4.00 pm. Analysis was made easy by using *Bar charts* and *Pie Charts* for visualization.







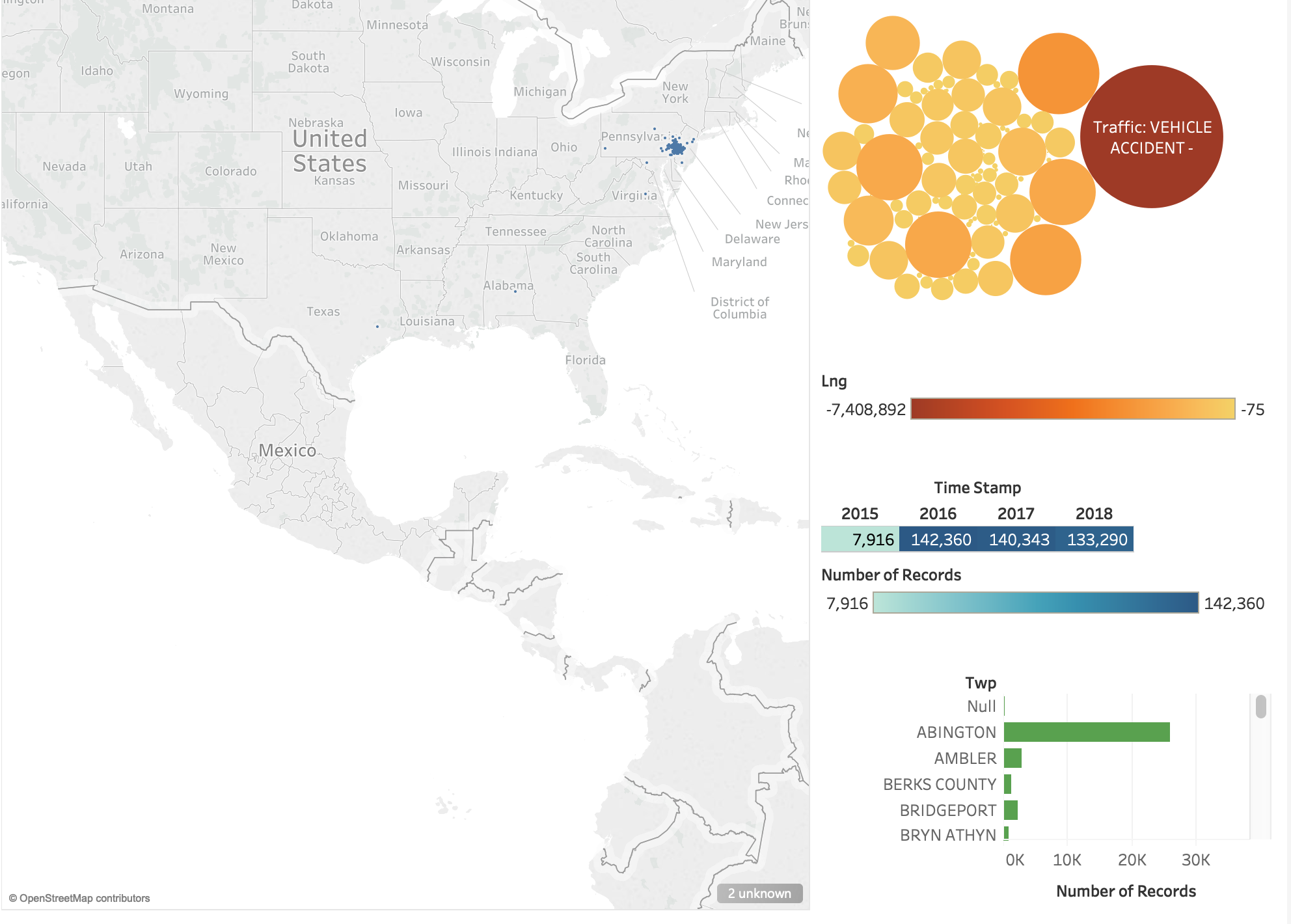
* For checking whether there has been an increase of calls from the previous year, I found the number of calls made in 2017 by splitting data from 2017 to a new dataframe called *df2017* and using *df2017.info ()* command. 2017 had a total of **140343** calls. This indicates a lesser number of calls than year 2018. But the issue here is that, since 2018 is the current year and since there is one more month to go in 2018 to get the data for the entire year, it cannot be concluded that 2017 had a greater number of calls; the number of calls in December 2018 should also be counted in order to conclude that.

**Summary**

**What I learned about this dataset:**

In thismini-project I worked with the 911 call dataset of Montgomery County, and was able to understand the various reasons for which an emergency call is made. It was interesting to understand the volume of calls made in each day, which gave a rough idea of the number of accidents happening in the County. The total number of calls made from the year of 2015 to 2018 is 423909 and in the year of 2018 alone, the number of calls made were 133290. From whatever I studied from this dataset, it can be concluded that *Vehicle accidents* are the major cause of 911 calls for most of the time. It should be analyzed whether these higher amount of road accidents are due to bad road conditions or traffic regulation failures. A huge relief is that *Active shooter* has the least number of calls recorded. There has been only 1 call made about an active shooter in the year of 2018. Also, there has been only 1 call made about poisoning. There has been a considerable amount of calls made about fire alarms. Similarly, respiratory emergency calls are also higher, which could indicate the presence of harmful air pollutants in the air in the particular area. Further detailed analysis of these data could help in determining the air quality of the areas where respiratory emergencies are on the higher number. I could also observe that Mondays had the highest number of emergency calls, and Sundays had the least number of calls. As per common assumption, Mondays are usually stressful, and Sundays are the least stressful, which can be concurred from this data as well. While looking at the time of the day, it could be seen that EMS calls are more during the period of 11.00 am and during the period of 3.00 pm to 5.00 pm, which could indicate that these are the most stressful times of the day. The least number of calls were recorded during the month of November, but it cannot be finalized yet, since the year of 2018 has one more month to go. But based on the existing data, it can be safely predicted that the year of 2018 has comparatively lesser emergencies than the previous years.

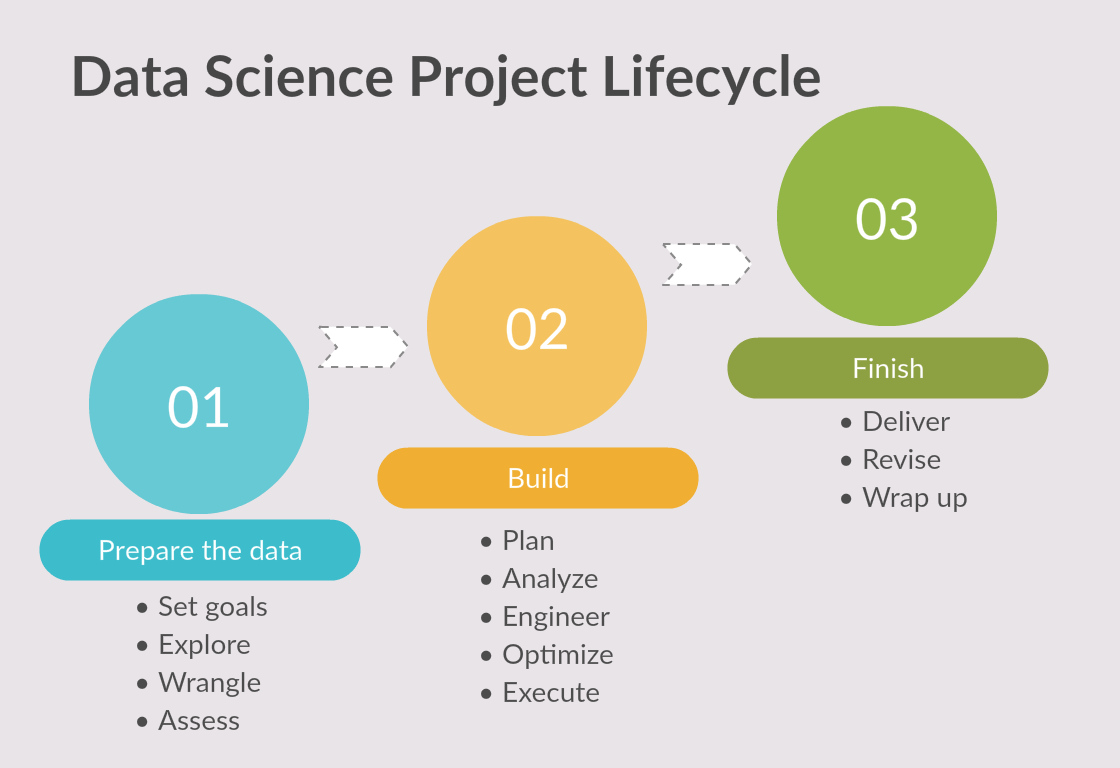
I tried to do data visualization of the overall results using *Tableau software* and the result is attached below:



Data visualization was done by extracting the *csv file* of 911 call data, and constructing various sheets of geo symbol maps, bubble maps, lines and bar charts. These sheets were combined to create this single *dashboard*, which helps in the overall visualization of data of this project.

**What I learned from this mini-project:**

This mini-project had been a wonderful experience. Reading the book “*Think like a Data Scientist*” was an eye-opener towards the lifecycle of a data science project. I could understand the various steps involved in the life cycle, and this mini-project gave me a hands-on experience on most of those steps. As said in the book, the first step of Data preparation is the most time-consuming step, and I found that deciding on a topic from this vast availability of interesting topics was the biggest challenge I faced during this project. There were many projects which piqued my interest and which I wanted to pursue but deciding on one most interesting topic was the biggest decision I had to make. After deciding on the topic, gathering the required data was an interesting step, where I got to read a lot about 911 call data and the various possibilities of the data.



According to Brian Godsey, the main object of any data science project is to produce something that help solve problems and achieve goals. Here, the goals were to understand the amount of emergency calls and the type of the calls made from a particular area. It is alarming that even a small area has these many calls being made each day, indicating that the volume of calls all over the country would be a much bigger number. But it is a relief to realize that help is just a call away, and the security that relief gives is indescribable.

**Being a Data Scientist & Skills I might need for the future:**

For being a Data Scientist, one must be knowledgeable enough to deal with even uncertainties. It would be safe to say that the job of a Data Scientist is much more challenging than the job of a Software Developer. Even though there are a lot of common traits among those two, a Data Scientist has to deal with a lot of uncertainties and probable data. A data scientist applies scientific methods towards solving problems. They should be able to look at data and ask the right questions about it, do the accurate analysis and provide with a useful solution. A data scientist must know how to interact with data and evaluate the data. He/She should know how to research independently and tackle various problems by finding correct solutions for them. A data scientist’s job involves 90% of data mining and cleaning, and 10% data modelling. This concurs with Brian Godsey’s words that the first stage of data preparation takes the longest time to complete.

To become a good Data Scientist, I might need to have a variety of sophisticated skills, to handle a huge volume of data. The data could be originating for multiple sources, each of different type. In order to deal with all the humongous data, I must be proficient with programming languages such as Python or R, and with database querying languages such as SQL. I should have a very good understanding of high-level Statistics, multivariable calculus and linear algebra, in order to do predictive analysis of data. I am expected to have a good understanding about Machine Learning while working with a large volume of data. A very thorough understanding of machine learning algorithms can be very useful in jobs in huge companies dealing with a large amount of data. Data wrangling, data visualization and communication are also very vital skills which a Data Scientist must possess. No data will be perfect, in order to organize unstructured data, data wrangling is essential. Data visualization is very vital, since this helps in understanding trends, patterns and helps in correlating various aspects. I should possess strong knowledge in Hadoop, Apache Spark and NoSQL in order to manage large data and run various statistics. Luckily, in this paper Big Data Technologies, I got to learn a lot about Hadoop, MapReduce, Hive, Spark, Pig, Scala and a lot more. Due to the very competent nature of this course, a lot of learning has happened in the field of Big Data since the beginning of this course. As far as non-technical skills are concerned, a Data Scientist should be highly inquisitive and should possess very good communication skills. The curiosity to discover and research upon data should be present in order to achieve long term goals. I believe I can prove myself to be a very good Data Scientist in future, by acquiring all these skills.

### 

## References

[1] Wikipedia contributors. (2018, November 29). Montgomery County, Pennsylvania. In Wikipedia, The Free Encyclopedia. Retrieved 06:56, December 2, 2018, from <https://en.wikipedia.org/w/index.php?title=Montgomery_County,_Pennsylvania&oldid=871198479>

[2] Godsey, B. (2017). Think like a Data Scientist: Tackle the data science process step-by-step. Shelter Island, NY: Manning

[3] Data.gov (2018). <https://catalog.data.gov/dataset?tags=911>

[4] 911 response time analysis. <http://datadesk.latimes.com/posts/2012/10/lafd-border-analysis/>

[5] Creately, to create lifecycle diagram. <https://creately.com/>

[6] Python data visualization. <https://python-graph-gallery.com/>

[7] CFS Analytics for 911 data study. <https://cfsanalytics.rti.org/news-events/>