**911 Calls Exploring Data and Analysis**

For this Exploring Data Analysis project, we will be analyzing some 911 call data from [Kaggle](https://www.kaggle.com/mchirico/montcoalert). The data contains the following fields:

* 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)

**Exploring Data Analysis report contents:**

* Data overview
* Data cleaning and Feature Engineering: Categorical Data
* Data cleaning and Feature Engineering: Numerical Data
* Hypothesis Testing (3 of them)
* Conducting a formal significance test for one of the hypotheses and discuss the results
* Suggestions for next steps in analyzing this data
* Summary of the quality of this data set and a request for additional data if needed

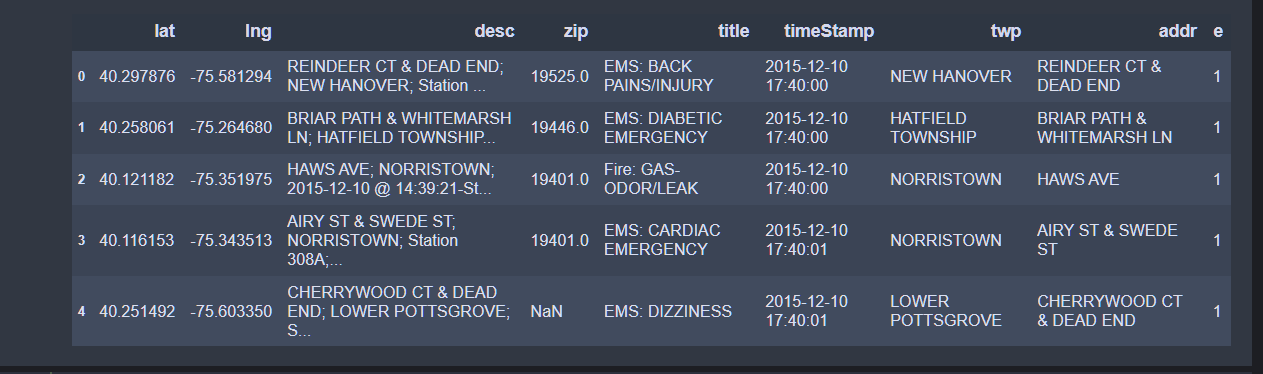
**Exploring Data Analysis Initial Plans:**

* Read data as dataset
* present the summary of characteristics of dataset and find out the data types of features of our dataset.
* Check Nan values, null values, outlier and perform data cleaning
* EDA -> Check makes statistical analysis like determining mean, median, correlation, standard deviation and do visualization while identifying features to use to predict.
* Feature engineering and Hypothesis analysis

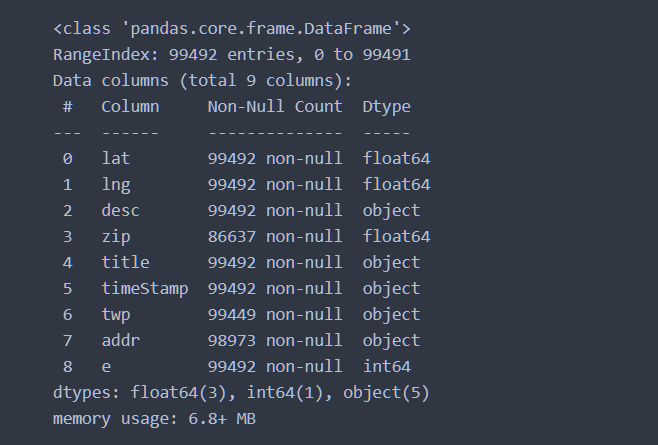
**EMS: Emergency Medical Services**

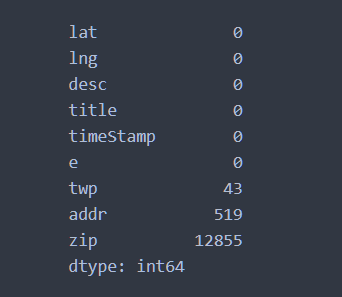
### Read Data as dataset

reading data and view what kind of data it is and look for columns and rows we have.



##### **We need to check characteristics of our dataset and find the summary of each features data types we have**





Through dataset, our data shows that we have the following characteristics:

* Number of rows (entries): 99492
* Number of columns: 9
* Dataset features contains different data types like (floats, objects)
* Dataset features shows missing data:
  + there are 12855 missing data from zip (zip code) column
  + there are 43 missing data from twp (township) column
  + there are 519 missing data addr (address) column
* Most of our data is type of Object (String).

###### 1. We need to find out most consecutive emergency that most frequently use 911 call.

###### 2. find out the place where there are many emergencies between 2015-2016.

###### 3. We will use hypothesis testing to find out time of day that has effect on volume of 911 calls.

###### **The results of above steps will help us to find which models that can be used to track down the uses of volume of 911 calls and days of most emergency.**

After adding other features of hour, month, and days of weeks of emergency calls, the results is the following



## Data Cleaning and Feature Engineering

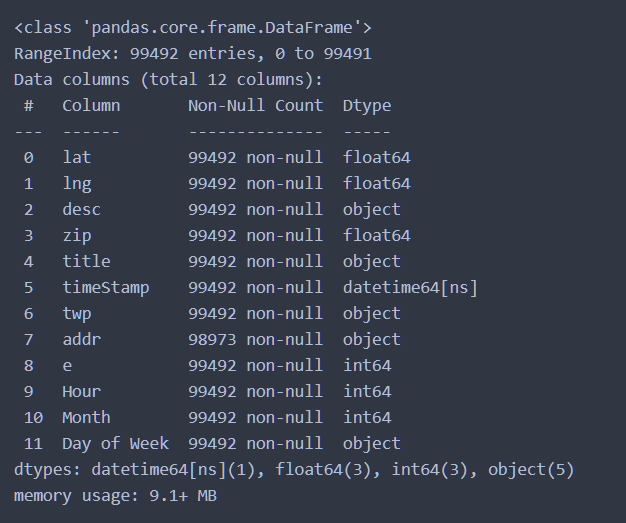
### **1. We will start with Data cleaning**

we found that we have the following data that we need to be careful working with them:

* there are 12855 missing data from zip (zip code) column
* there are 43 missing data from twp (township) column
* there are 519 missing data addr (address) column

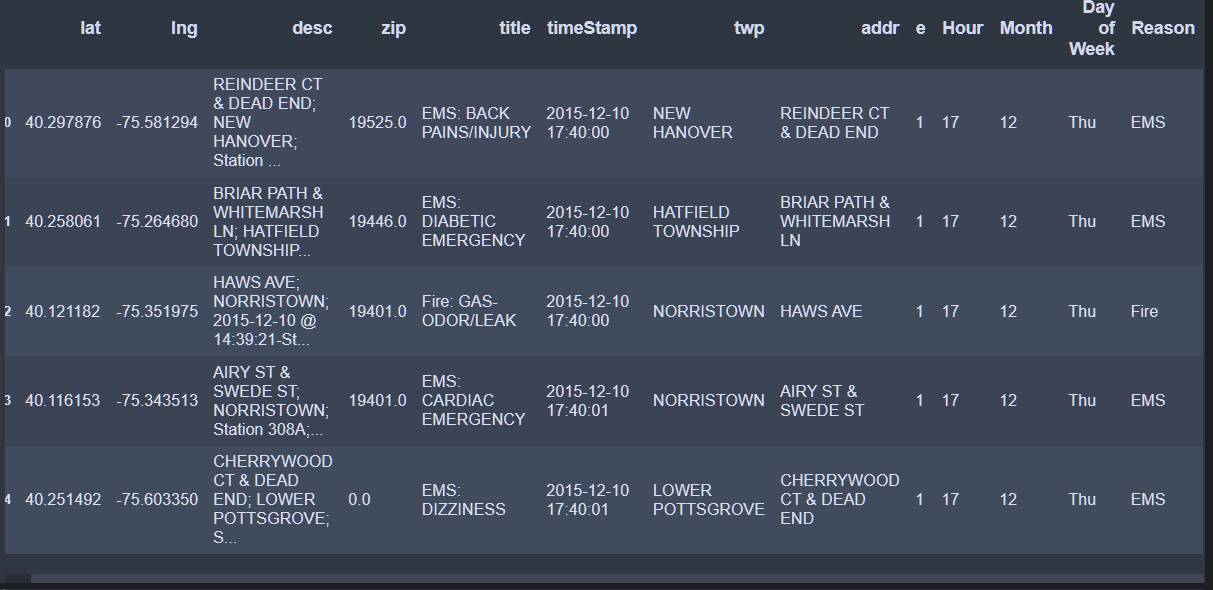
Thus:

* we will replace missing values of zip code into 0
* we will replace missing input data of township into noTown Provided
* we will replace missing values of address to its township



Now we can see that data cleaning have taken place.

To proceed with our goal of finding out the reason of most consecutive emergencies that most frequently use 911 call. we need, here we are go:

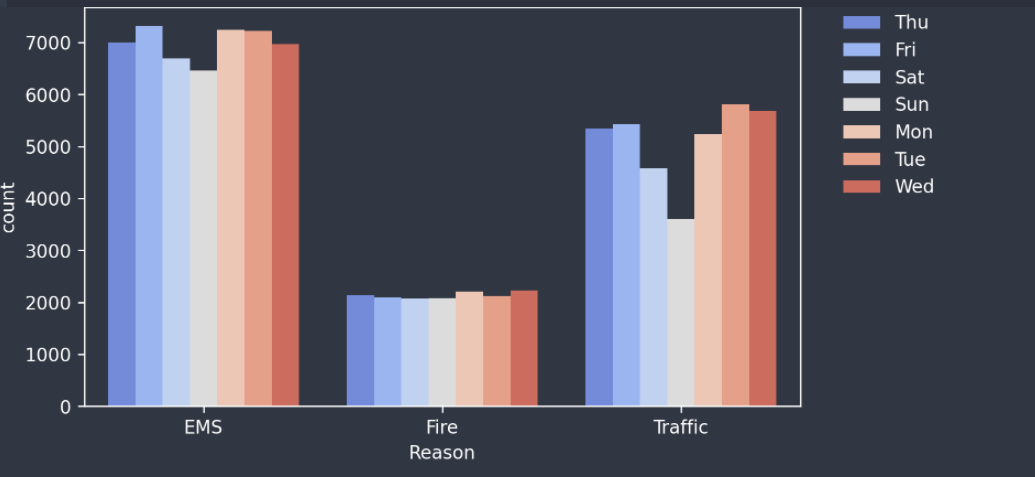


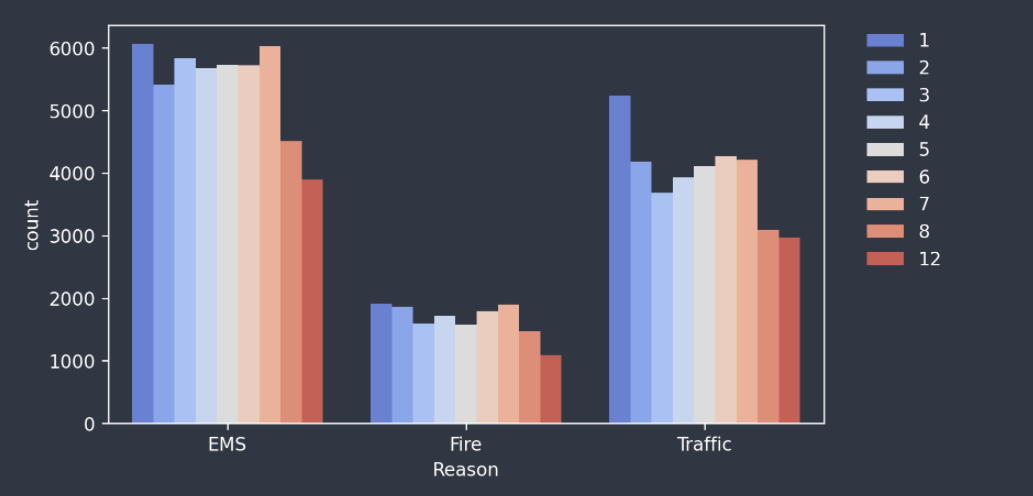
###### **Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner**

###### **1. We need to see general statistical values of results from cleaning and feature engineering to** describe our dataset



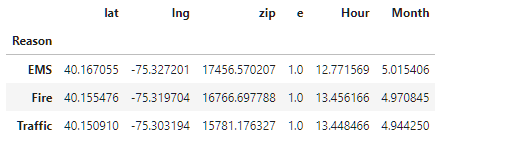
The above results shows that there are no outliers in our data. Then, and the table provide mean, standard deviation, min max, median, quartiles for better visualization

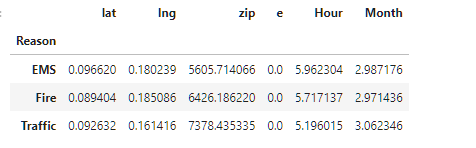




###### From the above charts of reasons of call due to Days of week and month, the visualization shows that there is imbalanced data provided which might cause an issue while model classification

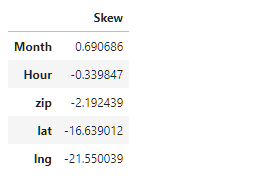
###### **2. Then, we can start to find mean, median, correlation, standard deviation for each group to see the possibilities of model selection.**



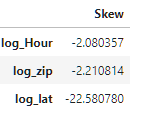


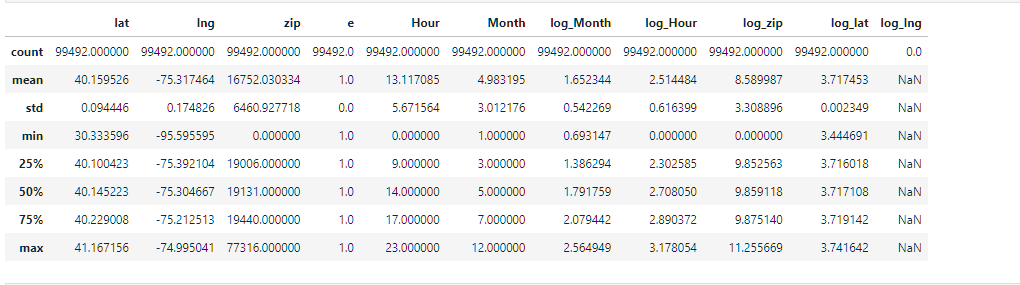
* We can observe from the above results that emergency of Fire has much effect in volume of 911 calls as the hour it takes has greater mean and low standard deviation.
* EMS has least effect in volume of 911 calls as the hour it takes has low mean compared to other reasons.

###### we can observe that, we need to continue with applying transformation to our data to improve them for next step of model fit.



**applying log transformation to skewed features from dataset**

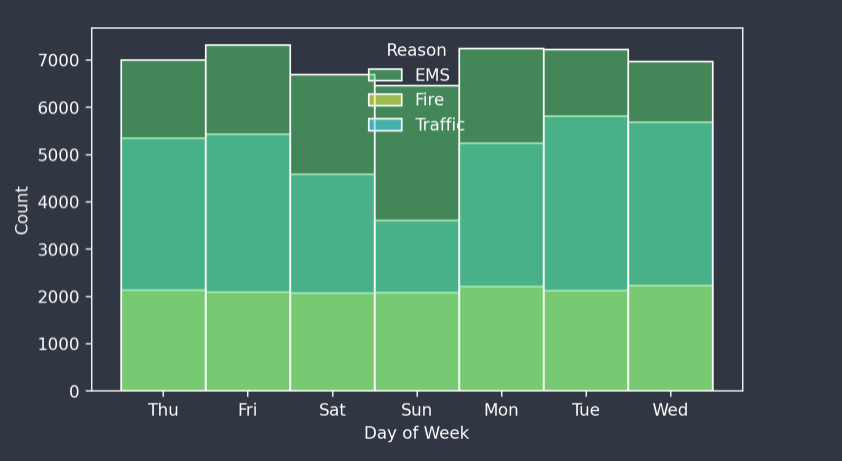


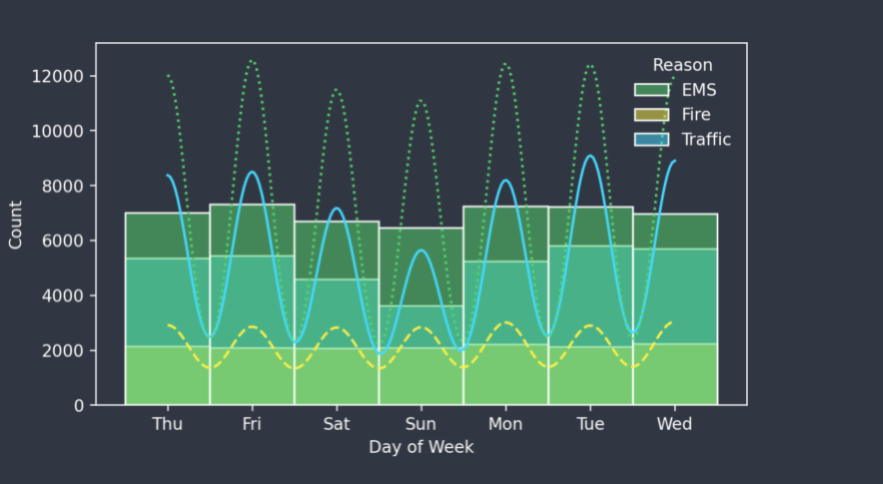




## Hypothesis Testing of our dataset

###### We will use hypothesis testing to find out time of day that has effect on volume of 911 calls.





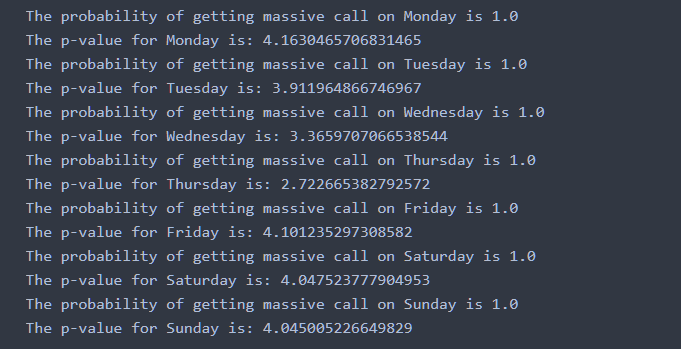
###### **From the above histogram, we can formulate our hypothesis in this way:**

1. time of day has some effect on volume of 911 calls
2. month has some effect on volume of 911 calls
3. EMS has great effect on volume of 911 calls

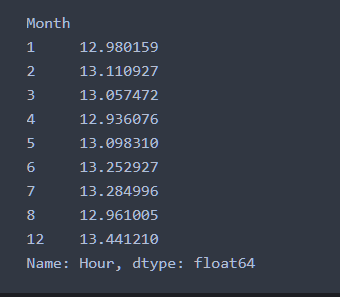
###### **1. Time of day has some effect on volume of 911 calls**

* our null hypothesis will be "time of day has no effect on volume of 911 calls"
* our alternative hypothesis will be "time of day has some effect on volume of 911 calls"

We need to find p-value. as p-value for a statistical model is the probability that when the null hypothesis is true, the statistical summary is equal to or greater than the actual observed results.



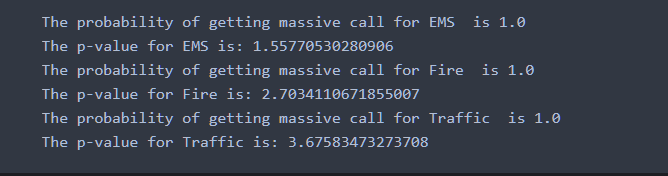
The above results shows that probability of getting massive 911 calls of emergency is 1 which is strange but somehow seems to be true. It means that our p-value is not explaining the observation fairly. but as the p-value is to low compared to significance level, we reject our null hypothesis and we can conclude that most time of day has some effect on volume of 911 calls.



###### **2. month has some effect on volume of 911 calls**

from the above data observations, there are missing data of September, October, and November. for simply, we can conclude that those months has no effect on volume of 911 calls which doesn't make sense since there is problem of missing data.

###### **3. EMS has great effect on volume of 911 calls.**



The above results shows that probability of getting massive 911 calls for different reasons of emergency provided is 1 which is strange but somehow seems to be true. It means that our p-values are not explaining the observation fairly except for p-value for traffic which is slightly greater than our significance level. but as the p-value is to low compared to significance level, we reject our null hypothesis for EMS and Fire, but we accept our null hypothesis for Traffic.

###### **Next steps**

**The following steps need to considered:**

* Missing data of months needs to be taken care of as it tends to reduce the confidences of model selection.
* It is highly recommended to request addition information to this data as its quality is not quite good. increase of features of data will be great.
* Dealing with known reasons we have, and the outcome is categorial data. this means that classification method is quite useful.
* I will suggest k-nearest neighbors to be used as classification method.

###### **Summary of quality of data.**

* The quality of data provides features that are quite useful but addition of other features is highly needed to make predictions and other analysis.
* We have seen that some provided data provided probability of 1, which is not quite true but its accurate is not bad.
* The selection process of model to be used so that the prediction is quite useful with our 911 calls dataset requires more observations which doesn't have missing data as we have seen.