**A Case Study**

**Data Source**:

Lets perform some EDA on real world data. The data we will consider is Baltimore Police Department Arrest data. The data is hosted on: [Data set Source](https://data.baltimorecity.gov/Public-Safety/BPD-Arrests/3i3v-ibrt)  Baltimore Police Depratment’s website: [Baltimore Police Department](http://www.baltimorepolice.org/) .  Data consists of around 155,000 arrests made by the Baltimore Police Department.

This data represents the top arrest charge of those processed at Baltimore’s Central Booking & Intake Facility. This data does not contain those who have been processed through Juvenile Booking. The data set was originally created on October 18, 2011. The data set was last updated on November 18, 2016. It is updated on a monthly basis.

Metadata

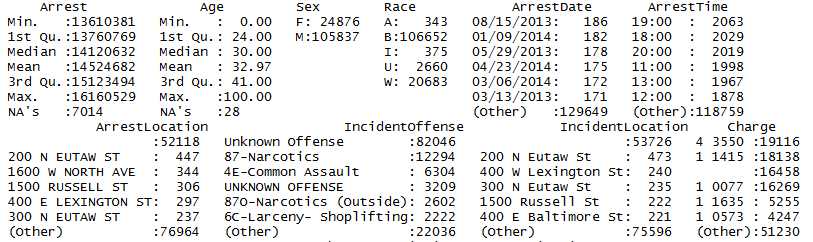
* Arrest-ID
* Age
* Sex
* Race
* ArrestDate
* ArrestTime
* ArrestLocation
* IncidentOffense
* IncidentLocation
* Charge
* ChargeDescription
* District
* Post
* Neighborhood
* Location1(Location Coordinates)

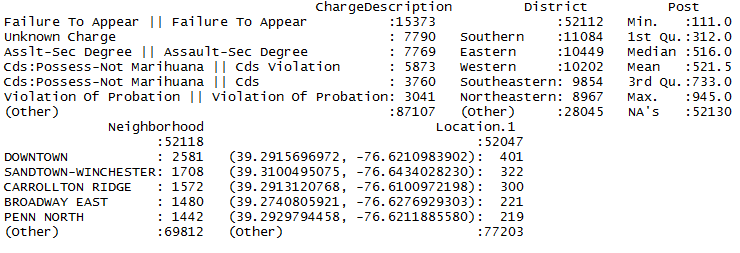
We will go through the analysis tool agnostic.  However, open source tools like R / Python can be used for analysis.

*First we will load the data and understand its dimension.*

Data have 130713 observations and 15 variables.

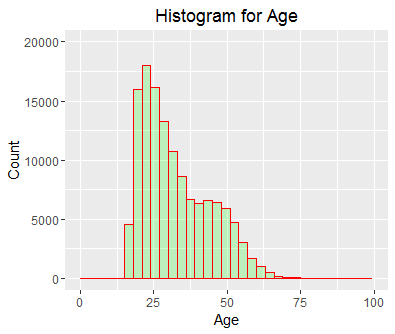
Let’s have a sneak peekof data before we start our analysis.





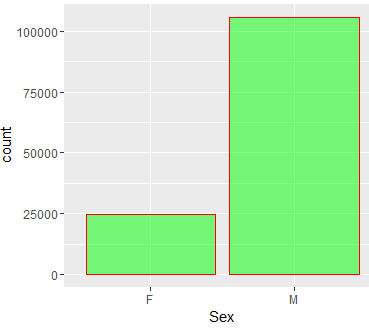
We have NA values and missing labels in some variables. This is quite intuitive in real world data. We have to prepare our data taking care of all hindrances.

Lets’ understand the Age distribution.



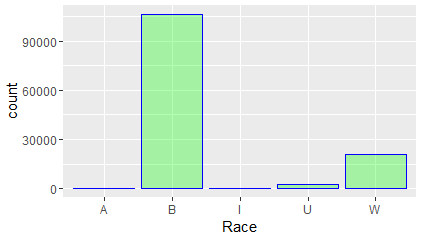
As we see most of arrests are in the age group of 20 -27.

A cursory view of gender (Sex) distribution reveals the following:



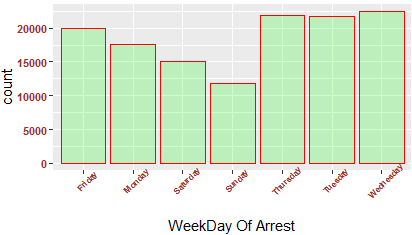
So, number of arrests is dominated by male.

Race is distributed in the following way.

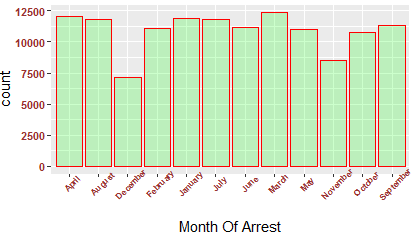


So, we see from data people from B Race group are mostly arrested by police.

We have variables recording the Arrest Date and Arrest Time. We can use these variables to reveal some interesting patterns.



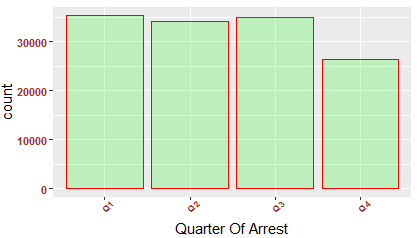
We can see that weekends have fewer number of arrests.



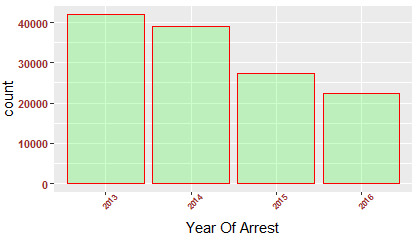
Arrest in different months have some interesting patterns. Can you hypothesize, the data pattern?

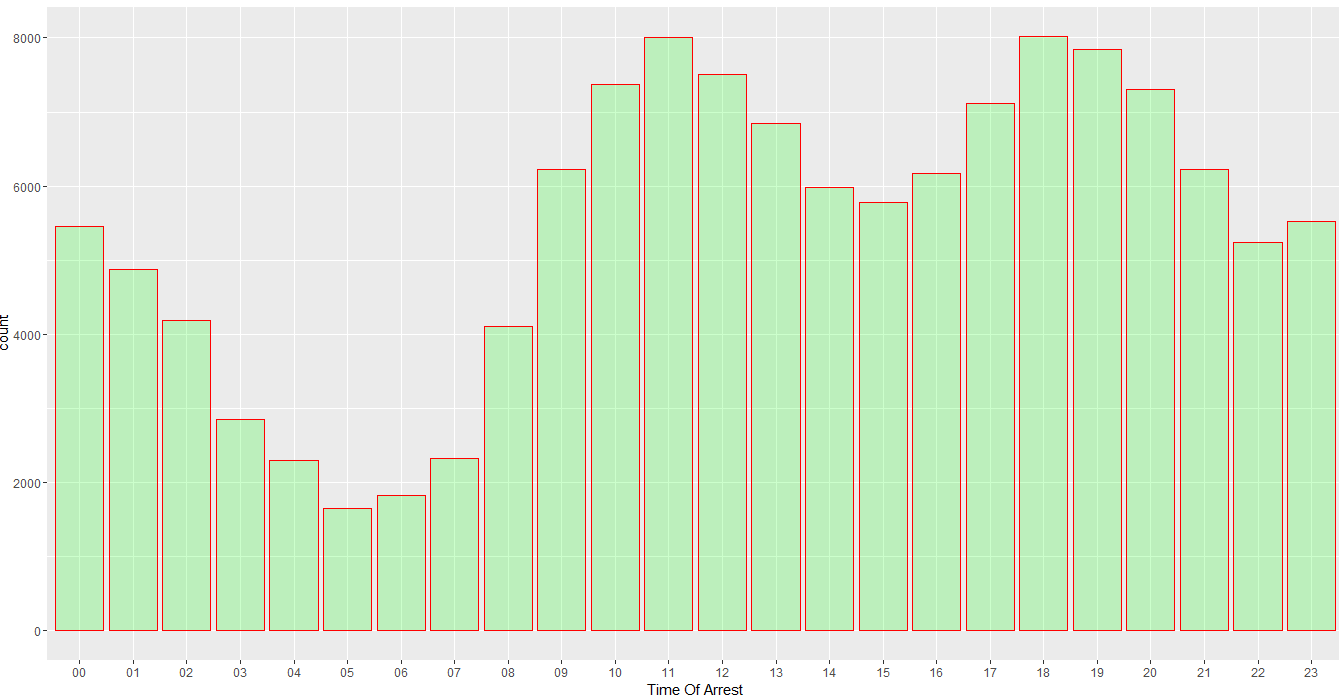
I will leave it as an exercise for you. Please do post your hypothesis in the comments below.

Distribution of data in various quarters are as follows:



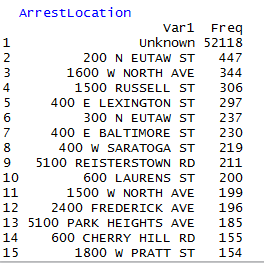
Number of people arrested show a declining trend year-on-year (2013-2016).



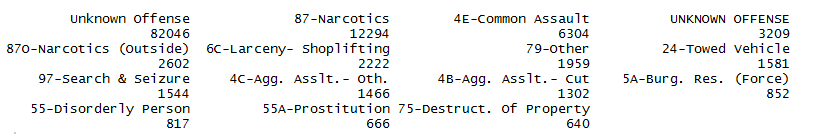


Now, we can see when most of the arrests occur in a day. There are fewer arrests in early morning.

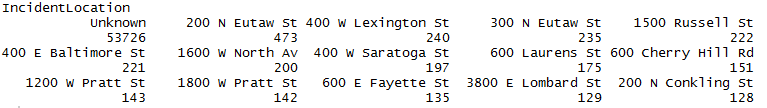
**Top 15 arrest locations are following:**



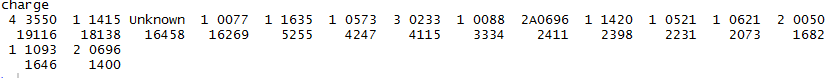
**Top 15 Incident Offense:**



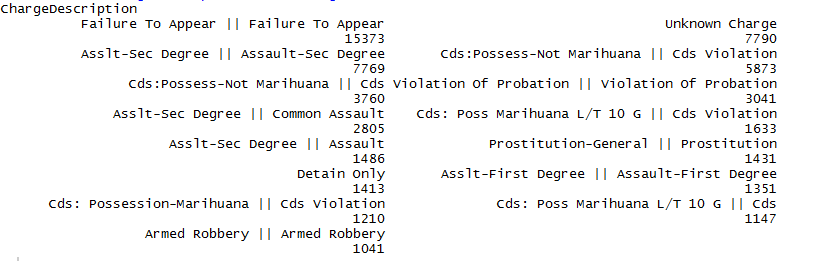
**Top 15 Incident Location**



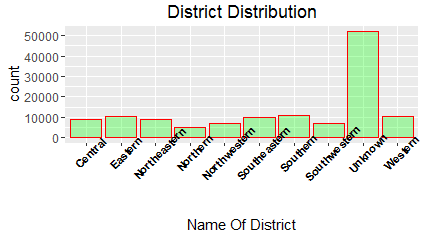
**Top 15 charge**



**Top 15 Charge Description**



Lets’ view the distribution of people arrested across different districts.

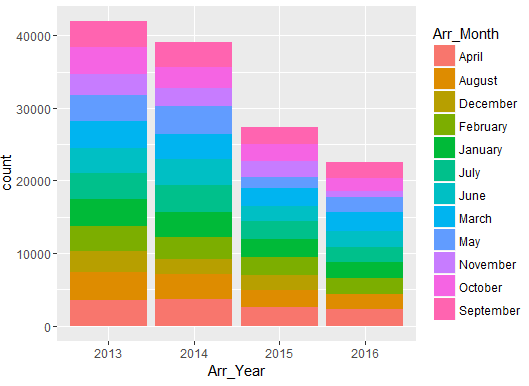


Comparatively, Southern district has more number of arrests.

**Top 15 Posts**

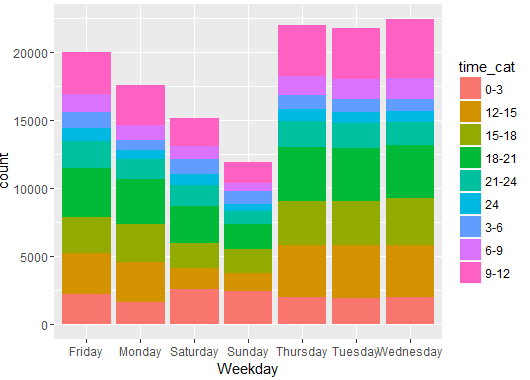
top-15-posts-fusion-analytics-world

**Distribution of Arrest Year & Arrest month**

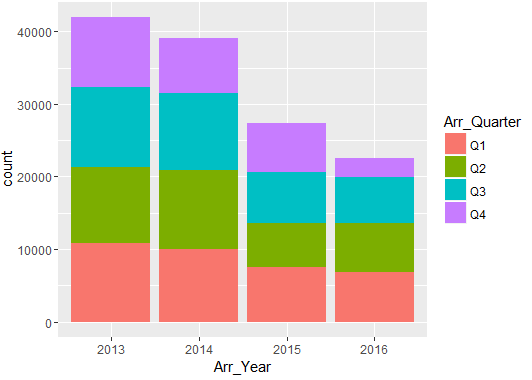


This is self-explanatory.

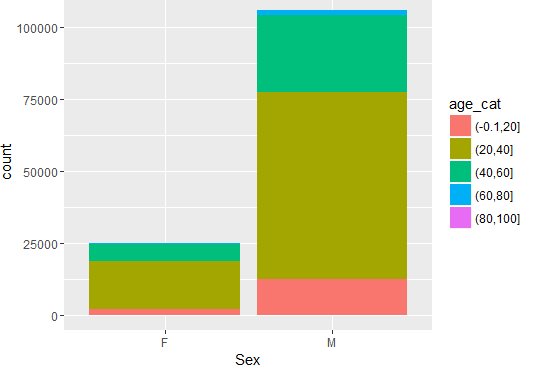
**Distribution of arrests Weekdays & Time**



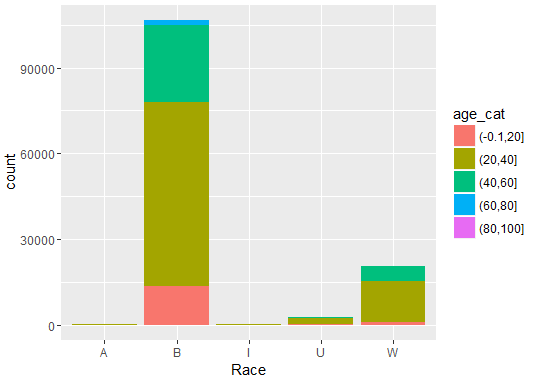
**Distribution of arrests Year & Quarter**



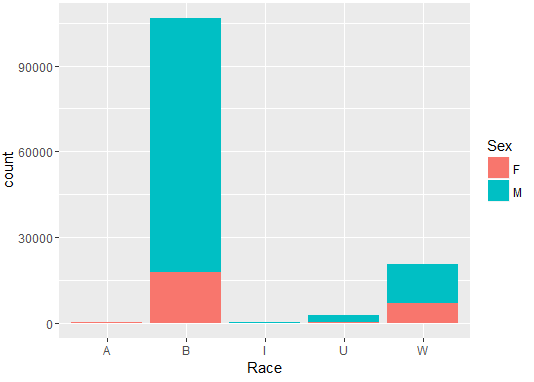
**Distribution of Arrests Age & Sex**



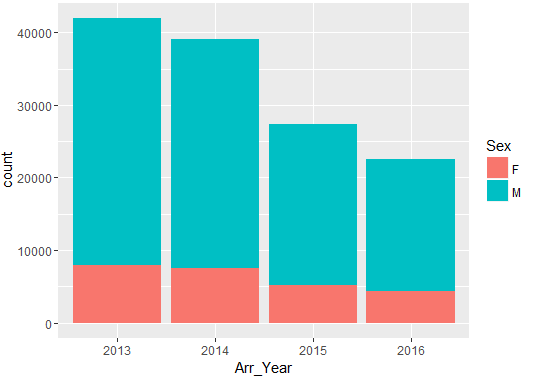
**Distribution of Arrests Age & Race**



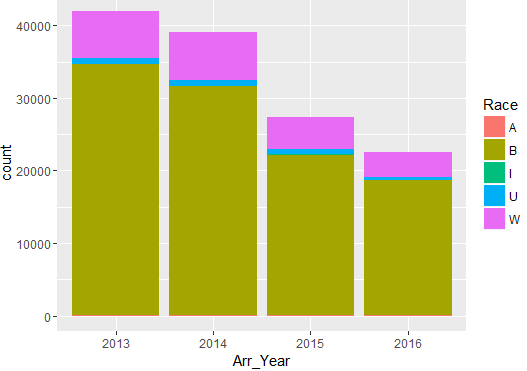
**Distribution of Arrests Race & Sex**



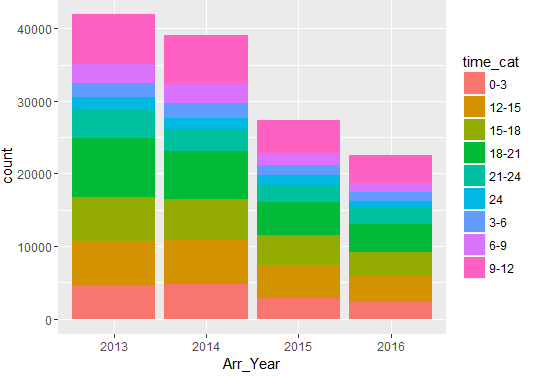
**Distribution of Arrests Year & Sex**



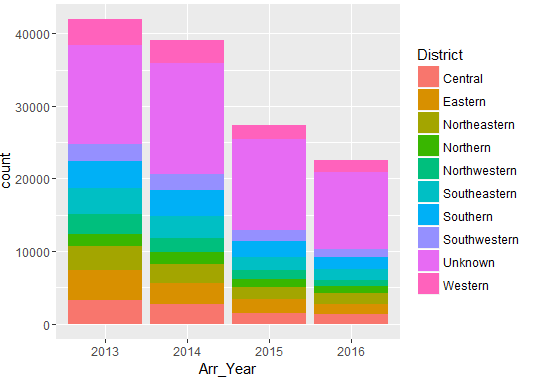
**Distribution of Arrests Year & Race**



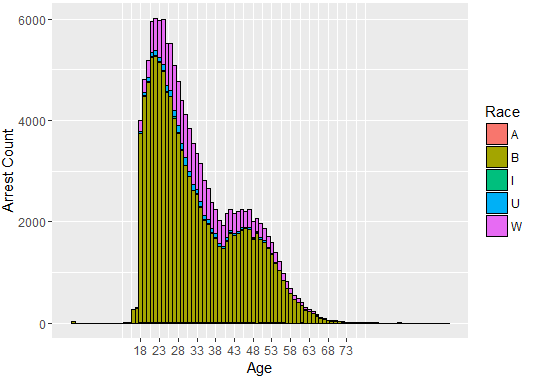
**Distribution of Arrests Year & Time**



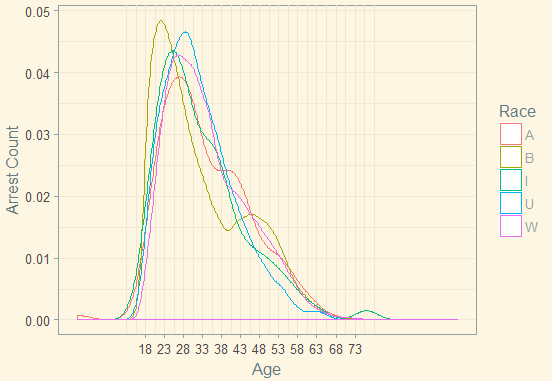
**Distribution of Arrests Year & District**



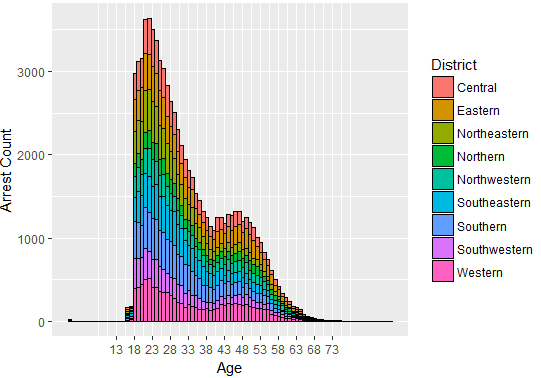
**Distribution of arrests age and race in a histogram**



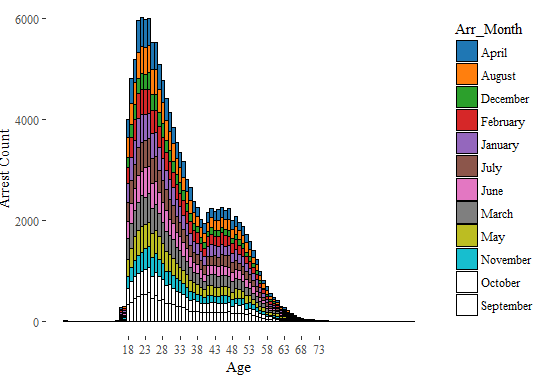
**Density Plot:**



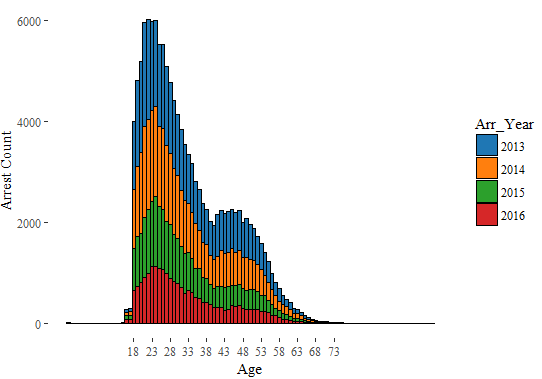
**Arrests across district and age**



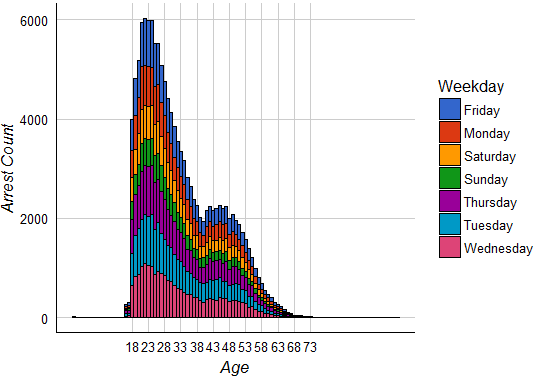
**Arrests Age Month**



**Arrests Age Year**

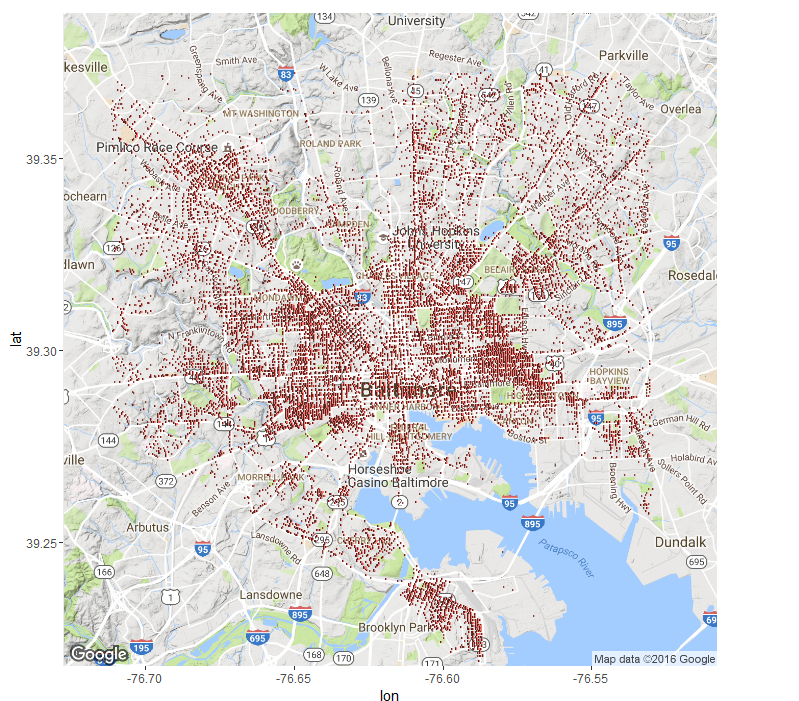


**Distribution of Arrests on age and Weekdays**

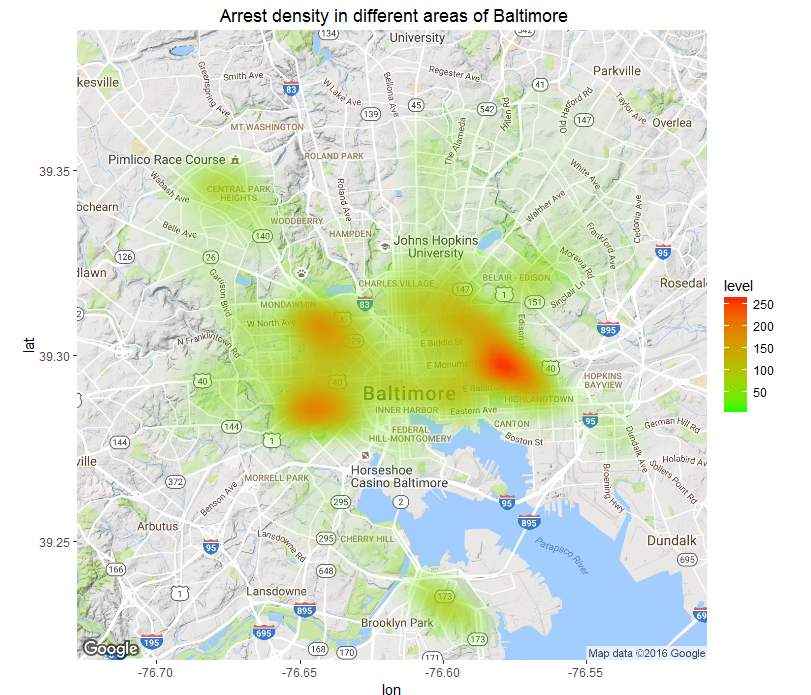


We have seen the distributions of number of arrests across various variables and the interactions among them.

But, if we want to brief stakeholders regarding the distribution of arrests across locations some visual geographic presentations serve the purpose. It give a clear picture.

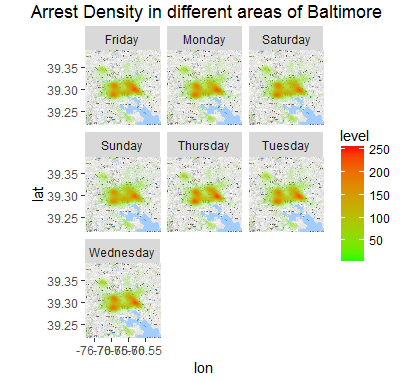


We can also present a picture regarding areas which are most susceptible to crimes and arrests.



The red areas are where majority of arrests occurred and the number decrease down to green areas.

Lets’ view the concentrations of arrests across different weekdays in different locations.



**End Notes:**

We can appreciate the role of exploratory data analysis in understanding the patterns of data and preparation of hypothesis. These serve as building blocks for further analysis. We have seen various plots to understand the data.  For better understanding we have analyzed a real world data set. Stakeholders may take few of the decisions on prevention of crime and deployment of personnel after having a glance on the plots.  A resourceful analysis can help taking complex decisions. We can use Predictive analytics and various optimization techniques for taking such decisions.

**Author**: [Santanu Dutta](https://in.linkedin.com/in/santanu-dutta-02991333)

**Source** : <http://fusionanalyticsworld.com/exploratory-data-analysis-case-study/>