**CHICAGO CRIME PATTERN ANALYSIS USING MACHINE LEARNING & APACHE SPARK**

**CONTENTS**

1. **ABSTRACT**

1. **PROBLEM DEFINITION**

1. **DATA DESCRIPTION**
2. **TECHNIQUES**
3. **RESULT**
4. **ROLE OF TEAM MEMBER**

**ABSTRACT**

The security status of the location where we live is one of the most critical

Important issues for every citizen. As most individuals will remember, In order to determine the security status, it is difficult to ignore the crime rate. Several variables can affect the incidence of crimes, including time, places, population, level of education, income rate. Hence the final project for Big Data Analytics was implemented to provide predictions and exploratory analysis for Chicago Crime Rate Pattern. The data was collected from the Chicago data portal spanning from 2017-2020. Our main goal was set to find the crime trend through the years from which we can know whether the security status in the city is getting better or worse. Secondly, what time is relatively more secure and which time period is more dangerous are also important for us. Besides, as we all know, different crime types have different levels of harm, we are interested in the proportion of different types of crime.

**PROBLEM SETTING**

dentified a few questions to get a perspective on the safety in the city of Chicago, which we answered during our data analytical project.

1. Merged the community area data frame with our existing data to get a better understanding about the crime rate in 77 communities in Chicago
2. Finding the top 20 crime type and its count for the past 7 years of data
3. How the arrest pattern has evolved over the years
4. Finding the peak time and duration at which the criminals commit any crime
5. Types of location crime occurs the most at indoor/outdoor location
6. Which month records the highest number of crimes in Chicago
7. Which day of the week records the highest number of crimes in Chicago

**DATA DESCRIPTION**

**Data Exploration:**

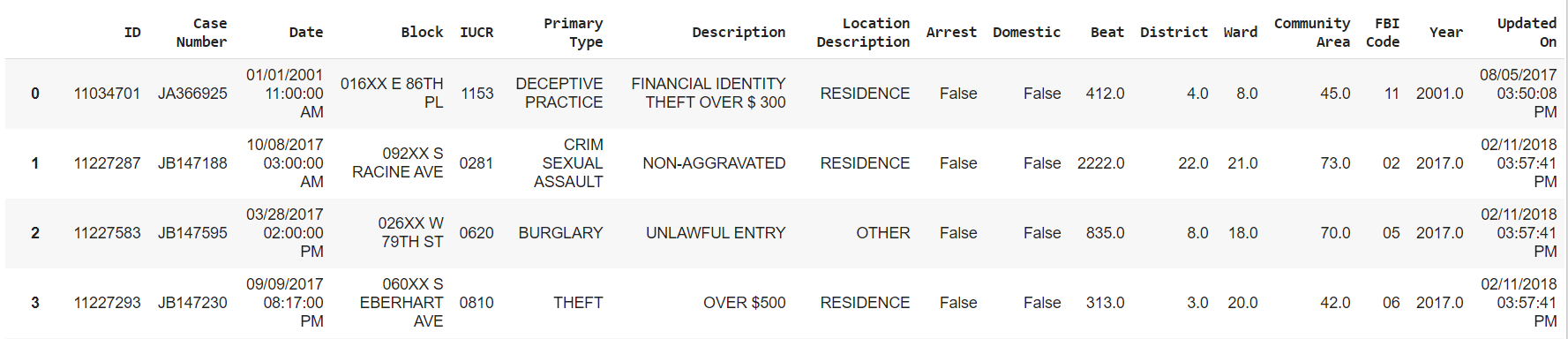
Data was collected from the Chicago Data Portal spanning from 2001 to November 2020, In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. In general, the data contained data such as the date/time that the crime occurred, the block where the crime occurred, type of crime, definition of location, whether an arrest occurred, and coordinates of location.

The data shown also includes more features such as latitude, longitude and location coordinates, but these features do not have strong correlation with respective selected features hence were not used for analysis purpose.

linked the names of community areas obtained from <https://www.chicago.gov/city/en/depts/cpd/dataset/police_stations.html> based on the community area code, Chicago has 77 communities the PDF for the same is attached with the submission.

**SIZE OF DATA:**

The dataset contains 1056780 number of records and 24 columns out of which the 17 columns are important for our analysis. The list of columns name are as following:



ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, ward, Community Area, FBI code, Year, Updated on.

**DATA EXTRACTION:**

1. Downloaded the folder in CSV format and Used Python Pandas
2. Data preprocessing procedure which is cleaning was implemented by:
   1. Removing duplicate rows
   2. Removing missing values (etc. Null/NA values) in the dataset
   3. Filtering out all the features from the dataset that are not relevant

**TECHNIQUES**

We are dealing with a big data set in our project, so the need to use technology to deal with the big data set wasn't an issue. It was however a bit of a struggle to decide between MapReduce and Spark. These two structures have their own qualities and advantages.

The main difference between them in fact, lies in the processing approach: Spark can do it in memory, while Hadoop MapReduce needs to read from and write to a disk. As a consequence, the processing speed varies significantly. But we decided on choosing Spark for the data analysis, as Spark in-memory computation is much faster and efficient in today’s world, we used Spark as our main framework.

**IMPORTANT LIBRARIES USED FOR EXPLORATORY ANALYSIS**

***import pandas as pd***

***import re***

***from pyspark import SparkContext***

***import seaborn as sns***

***import matplotlib.pyplot as plt***

**PREDICTIVE ANALYSIS**

Library Functions:

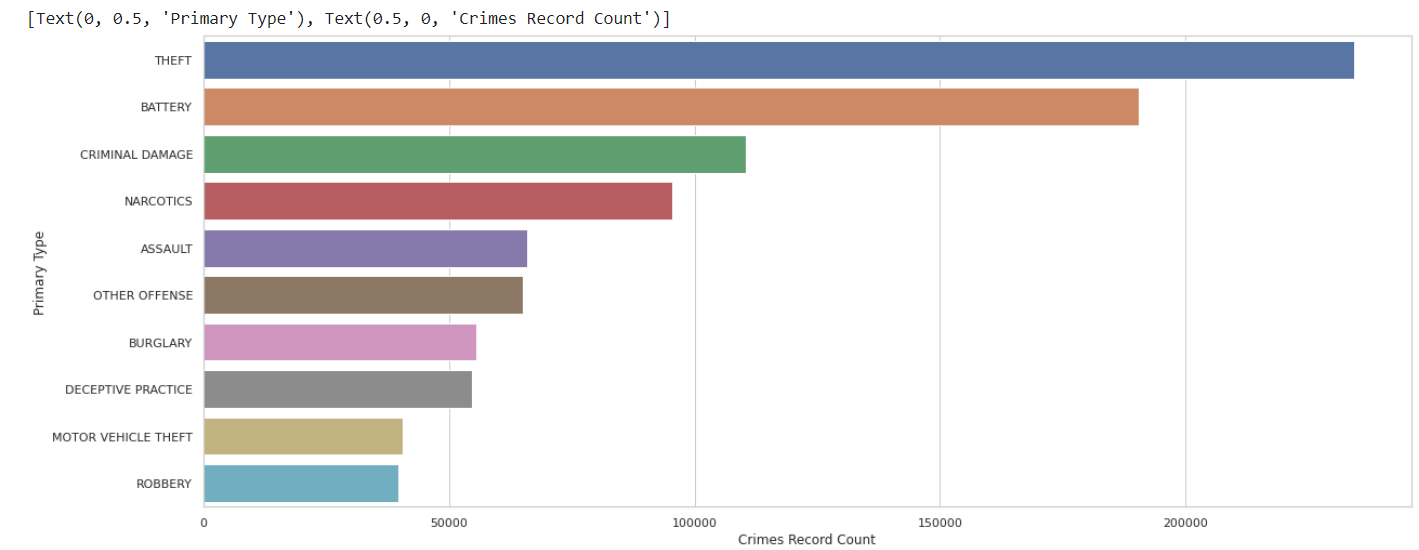
***from pyspark.ml.feature import StringIndexer, VectorAssembler***

***from pyspark.ml.classification import LogisticRegression***

1. For our predictive crime analysis, we used the logistic regression model from the apache spark machine learning library.
2. The correlation among attributes is not high which is an obvious indication of obtaining a less accuracy score. Therefore, to make shift in obtained accuracy we converted the required string type feature into vector index.

**RESULT**

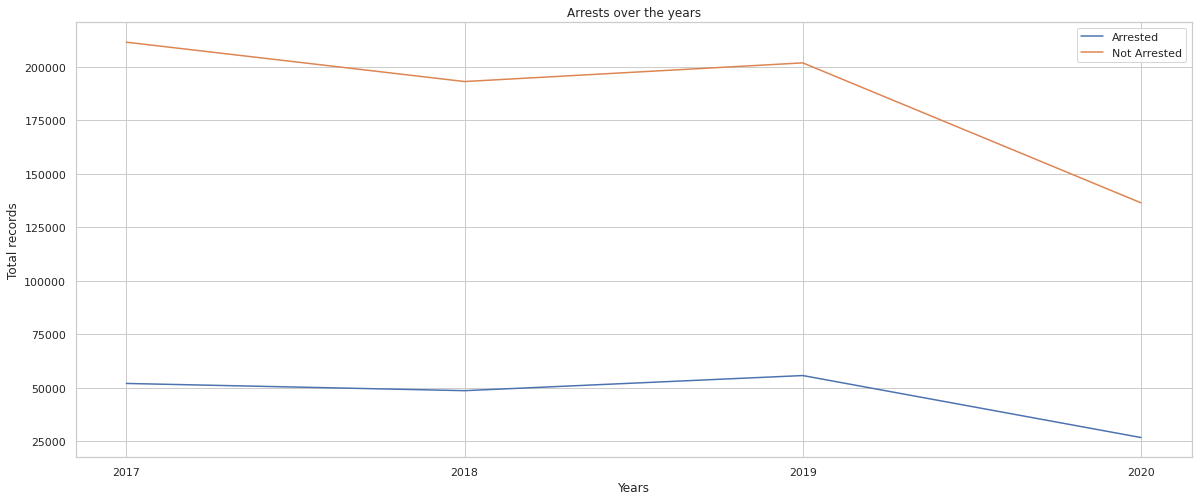
**1. Visualize top ten crimes in Chicago**

****

From the above fig, the visualization depicts the top ten primary type of crime that occur in the city of Chicago and it’s count.

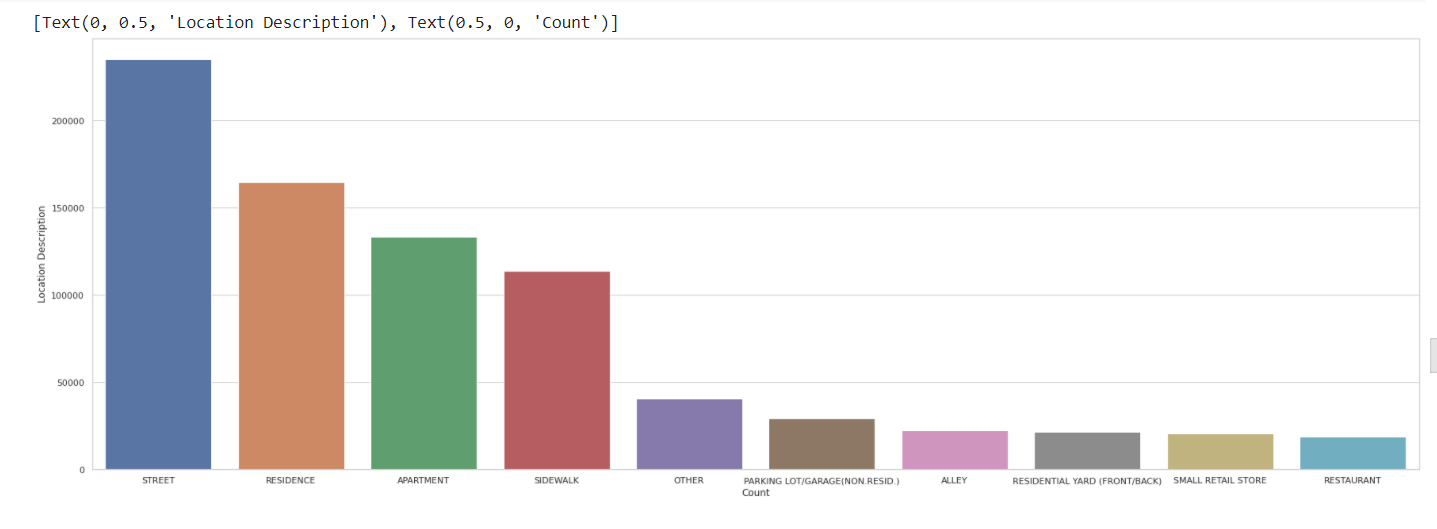
* Theft being the most committed primary crime, with a count above 200000
* Followed by Battery, Criminal damage, Narcotics, assault, other offenses, Burglary, Deceptive Practice and Motor Vehicle Theft
* The least committed primary is shown to be Robbery

**2. Visualize arrests over the years**

****

Blue indicates Arrested and orange indicates Not Arrested, hence we can see the arrest pattern over the years from 2017- 2020

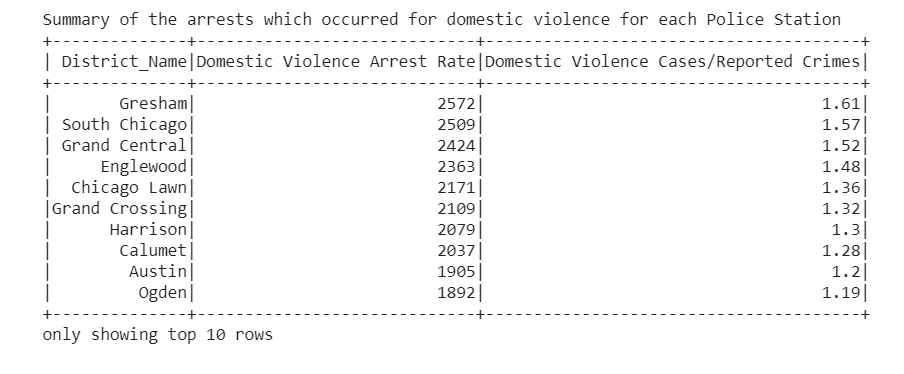
**3. Visualize number of crimes in each Location type (top 10)**

****

The figure depicts the count of crime in every location type

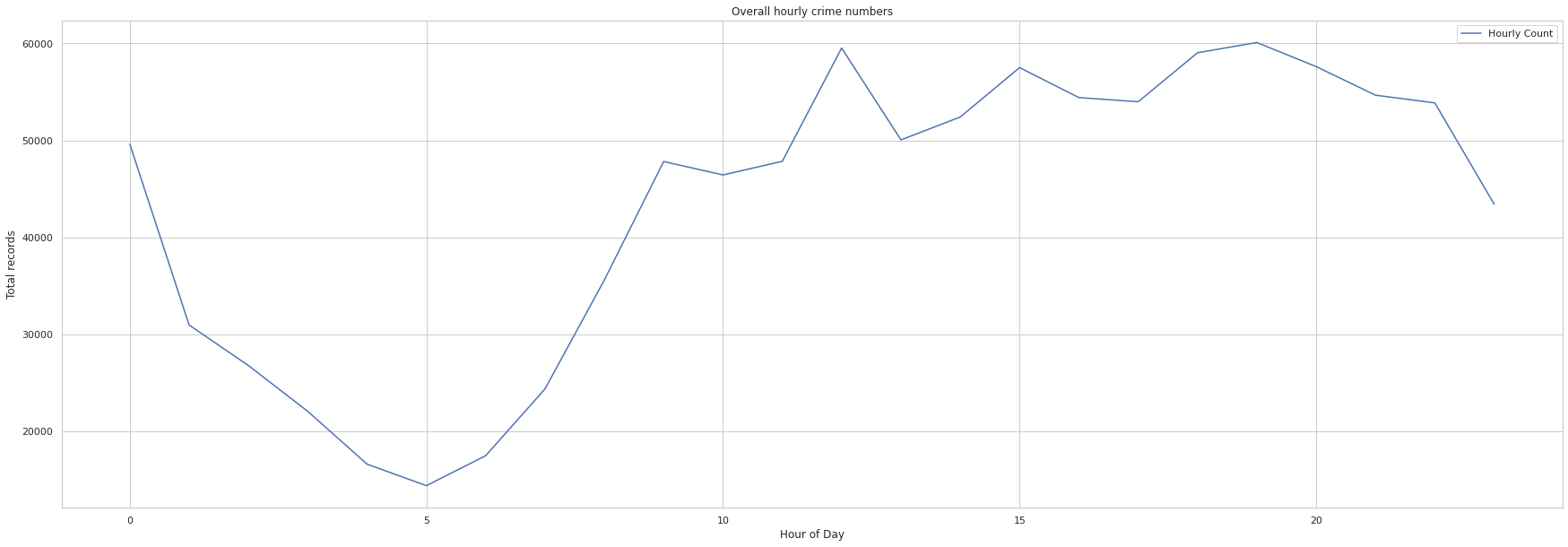
* Street is shown to have the highest occurrence rate of any crime type
* Followed by Residence, Apartment, Sidewalk, Others, Alley, Small Rental Store
* Restaurant is shown to have least amount of crime count

**4.****Top 10 police stations (with District names) having highest-lowest domestic violence arrest rate**

****

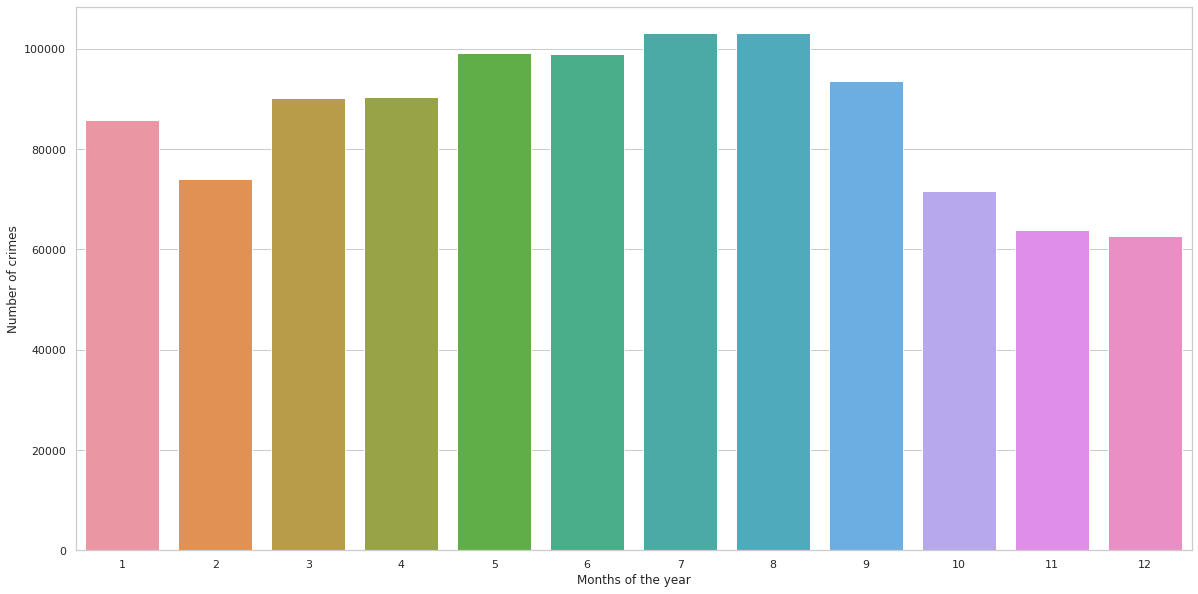
The table above shows the arrest made on domestic violence charges, arranged in descending order. Gresham Police Stations makes the highest arrest whereas Ogden makes the least number of arrests related to domestic crime.

**5. Let’s visualize which is the** **peak hour of the day for crimes in Chicago**

****

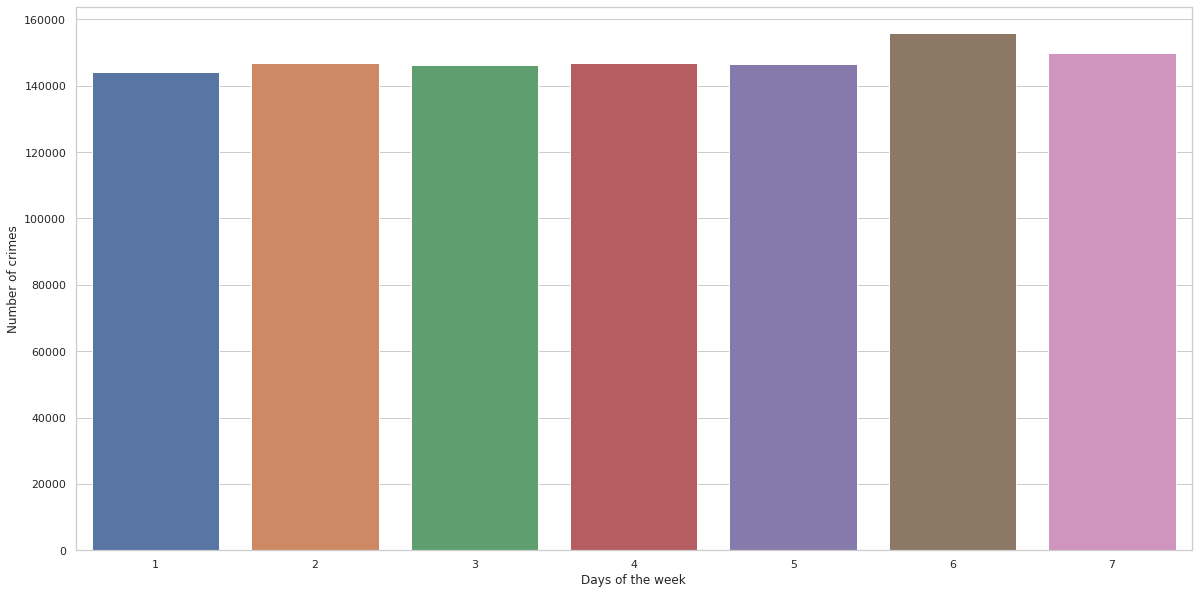
The peak hour of crime occurrence in the city of Chicago is normally around 12pm and 7pm. The least chances of a crime to occur is around 5AM.

**6. Visualize which month has the highest crime rate in Chicago**

****

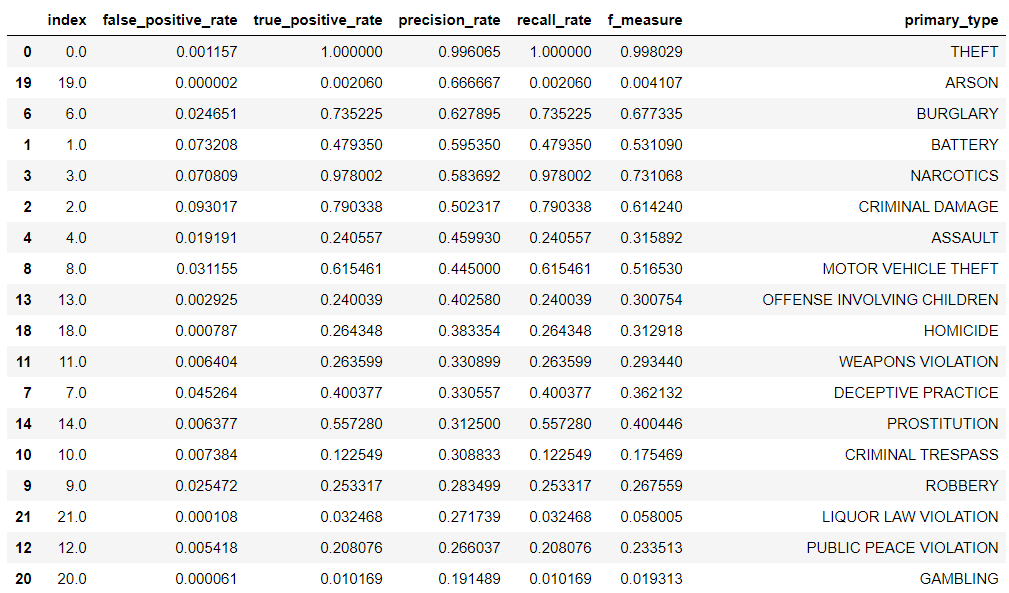
As we can see July and August have crime rate equally distributed and December is the month where the crime rate is the lowest.

**7. Visualize which day of the week has the highest crime rate in Chicago**

****

The above visualization displays Saturday to be highly susceptible for crime and Monday to be the least, and the rest of the week days are equally distributed in range.

**PREDICTIVE MODEL RESULT:**

****

Result contains all primary type of crimes, ordered by precision rate

Our training model predicts arson the best type of crime, but our test model predicts theft the best type of crime

Accuracy of the model was found to be 61.223%