PHILLY CITY CRIME DATASET ANALYSIS

BY

NIMMANI VYSHNAVI

DATASET

- Philly city crime data analysis for the last ten years (2010-2021)
  The dataset was collected from the Philadelphia official website which has many public datasets.
- For this project I have used two datasets ,the first dataset is the actual crime dataset which contains information about the crime that is time, location, category and primary type. The second dataset is weather dataset. Weather data consists information about Datetime, MaxTemperatue, MinTemp, Snow, CloudCover and Precipitation.
- https://metadata.phila.gov/index.html#home/ is official page for getting both datasets.

The dataset consists of 14 columns and 1.9 Million rows.

Cleaning Dataset

```
crime_df_dropnull = (crime_df_new.dropna()
                     .where(col("DispatchDate")<=to_date(lit("2021-06-30"),"yyyy-MM-dd"))
                    )
```

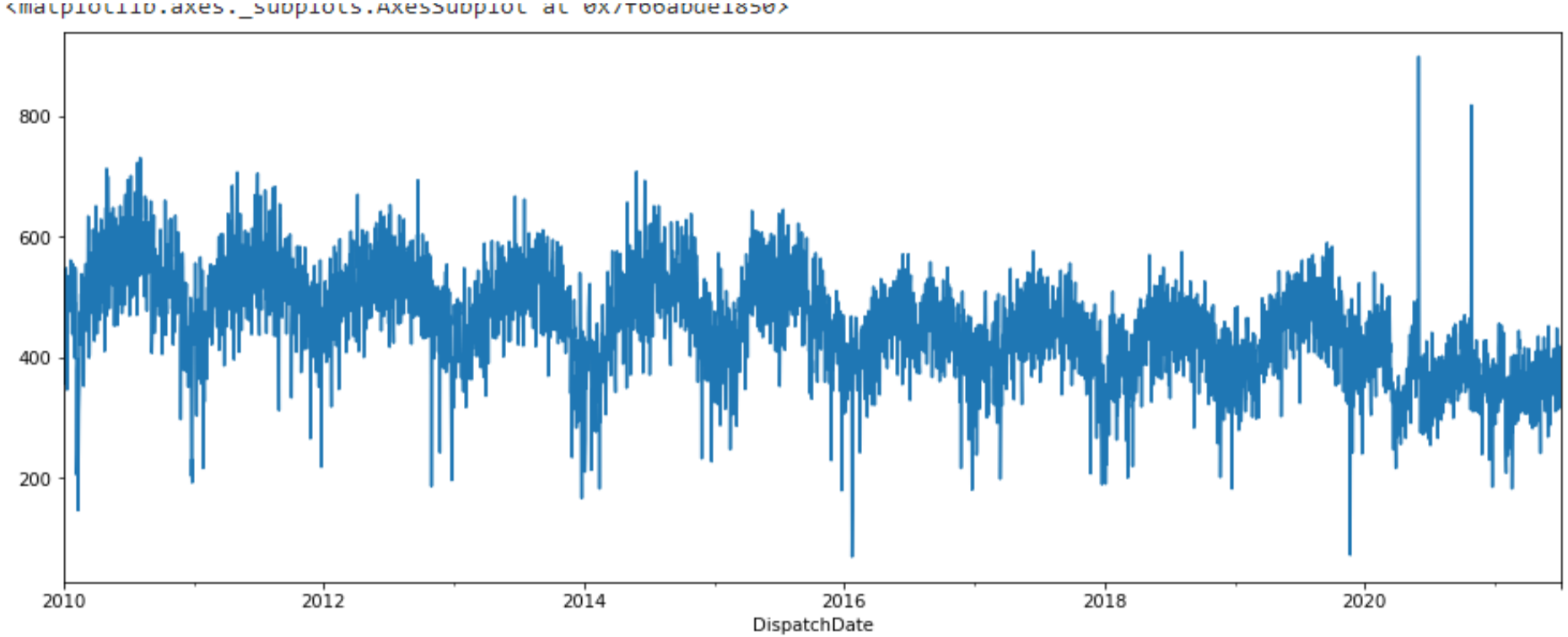
```
[44] crime_df2 = crime_df_dropnull.dropDuplicates()
```


```
44] crime_df2 = crime_df_dropnull.dropDuplicates()
```


```
45] weather_df2 = (weather_df
                   .withColumn("YearPart", split(col("DateTime"),'/')[2])
                   .withColumn("TimePart", split(col("YearPart"),' ')[1])
                   .withColumn("Year", split(col("YearPart"),' ')[0])
                   .withColumn("Day", split(col("DateTime"),'/')[1])
                   .withColumn("Month", split(col("DateTime"),'/')[0])
                   .withColumn("MonthVal", when(length("Month")==1, concat(lit("0"),col("Month")))
                               .otherwise(col("Month")))
                   .withColumn("DayVal", when(length("Day")==1, concat(lit("0"),col("Day")))
                               .otherwise(col("Day")))
                   .withColumn("DatePart", concat_ws('-',col("Year"),col("MonthVal"),col("DayVal")))
                   .withColumn("Date", to_date(col("DatePart"),"yyyy-MM-dd"))
                   .withColumn("Hour", split(col("TimePart"),':')[0])
                   .drop("YearPart")
                   .drop("TimePart")
                   .drop("Year")
                   .drop("Day")
```

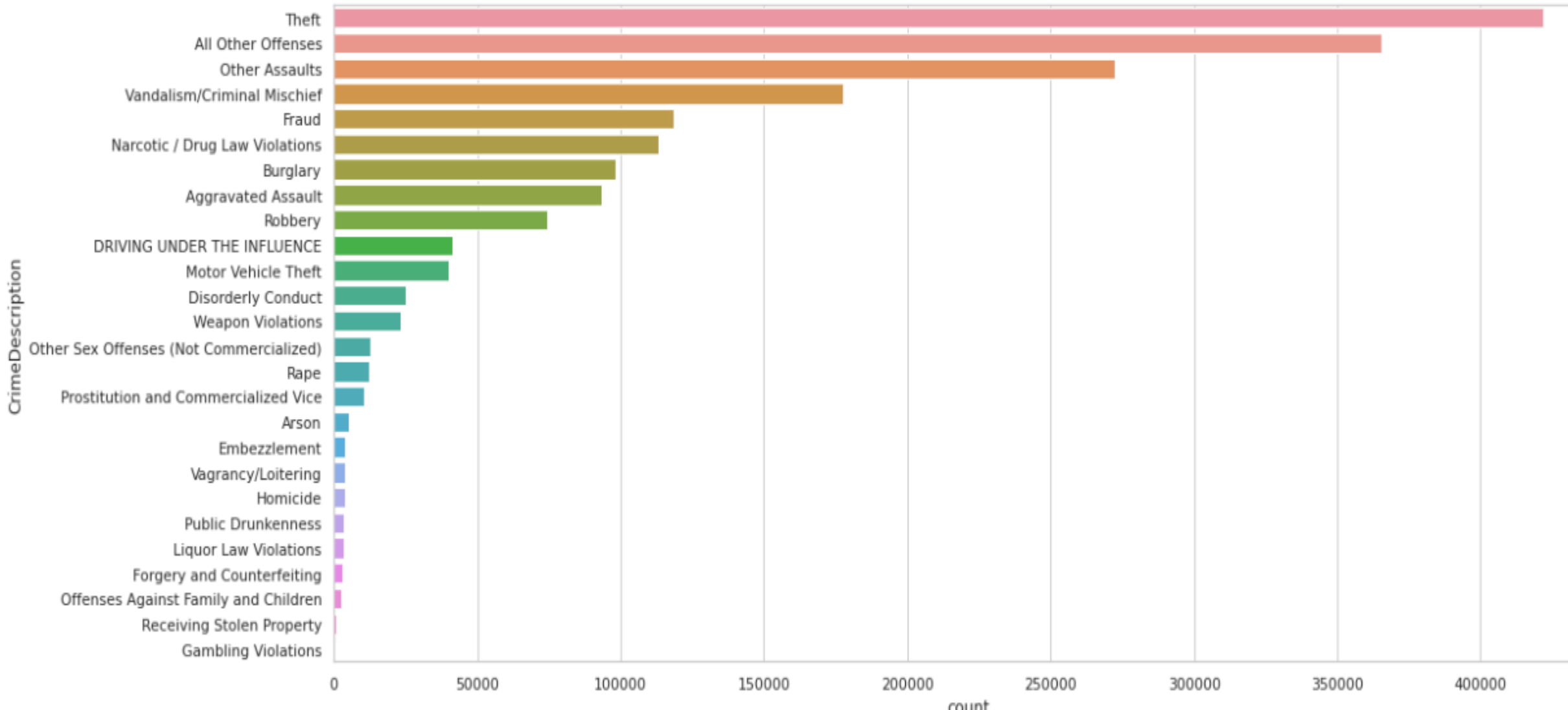
```
[84] # Join Crime and weather dataset based on date and hour, The weather data that we are using is hourly dataset
     crime_df_join = crime_df_new.join(weather_df2, ((crime_df3['DispatchDate'] == weather_df2['Date']) & (crime_df3['DispatchHour'] == weather_df2['Hour']))).drop("[

     # Partition the dataset based on Year
     crime_df_join.repartition("DispatchYear")
```

Which year and date had highest number of crime

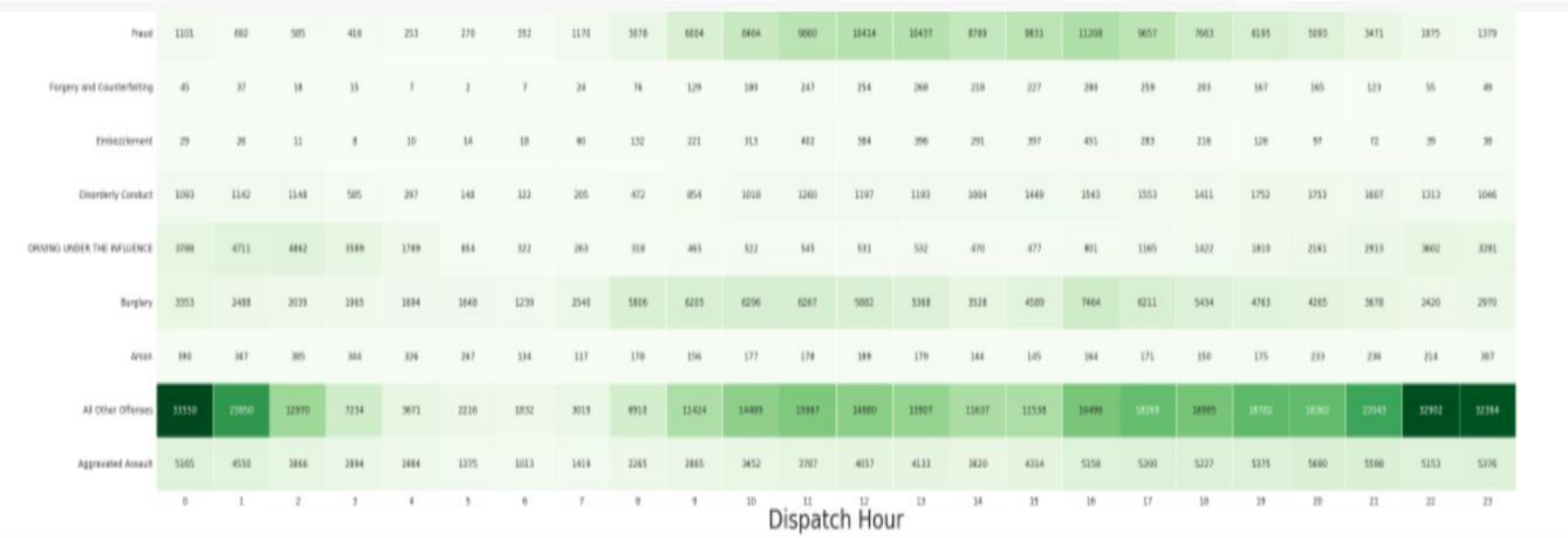


Crime Count

[ ]


Theft
All Other Offenses
Other Assaults
Vandalism/Criminal Mischief
Fraud
Narcotic / Drug Law Violations
Burglary
Aggravated Assault
Robbery
DRIVING UNDER THE INFLUENCE
Motor Vehicle Theft
Disorderly Conduct
Weapon Violations
Other Sex Offenses (Not Commercialized)
Rape
Prostitution and Commercialized Vice
Arson
Embezzlement
Vagrancy/Loitering
Homicide
Public Drunkenness
Liquor Law Violations
Forgery and Counterfeiting
Offenses Against Family and Children
Receiving Stolen Property
Gambling Violations
CrimeDescription
0
50000
100000
150000
200000
250000
300000
350000
400000
count

Crime types analysis on hour basis using heatmap



Which year had highest number of crimes

```
(crime_df_join
 .select("DispatchYear")
 .groupBy("DispatchYear")
 .count()
 .orderBy("count", ascending=Fals
 .show(n=10, truncate=False))
```

```
+------------+------+
|DispatchYear|count |
+------------+------+
|2010        |191247|
|2011        |187552|
|2012        |186990|
|2013        |176790|
|2014        |176442|
|2015        |174723|
|2019        |161087|
|2016        |159606|
|2017        |157314|
|2018        |152700|
+------------+------+
only showing top 10 rows
```

Comparison with weather conditions and crime data

```
display(crime_df_data
        .select("CrimeDescription")
        .where((col("MinTemperature") < 0) & (col("Snow") != 0) & (col("CloudCover") > 0) & (col("
        .groupBy("CrimeDescription")
        .count()
        .orderBy("count", ascending=False))
```

CrimeDescription	count
Theft	1660
All Other Offenses	1058
Other Assaults	960
Vandalism/Criminal Mischief	691
Fraud	467
Narcotic / Drug Law Violations	436
Burglary	416
Aggravated Assault	337

Conclusions

- Crime rates are low in extreme cold and extreme hot temperatures
- Most crimes happened on No-Snow days
- There are less crimes on cloudy days
- There are very less crimes reported on rainy days

THANK YOU