Chicago Crime Dataset: A Data Analysis Project

In this project I'll be focusing in the Primary Type variable

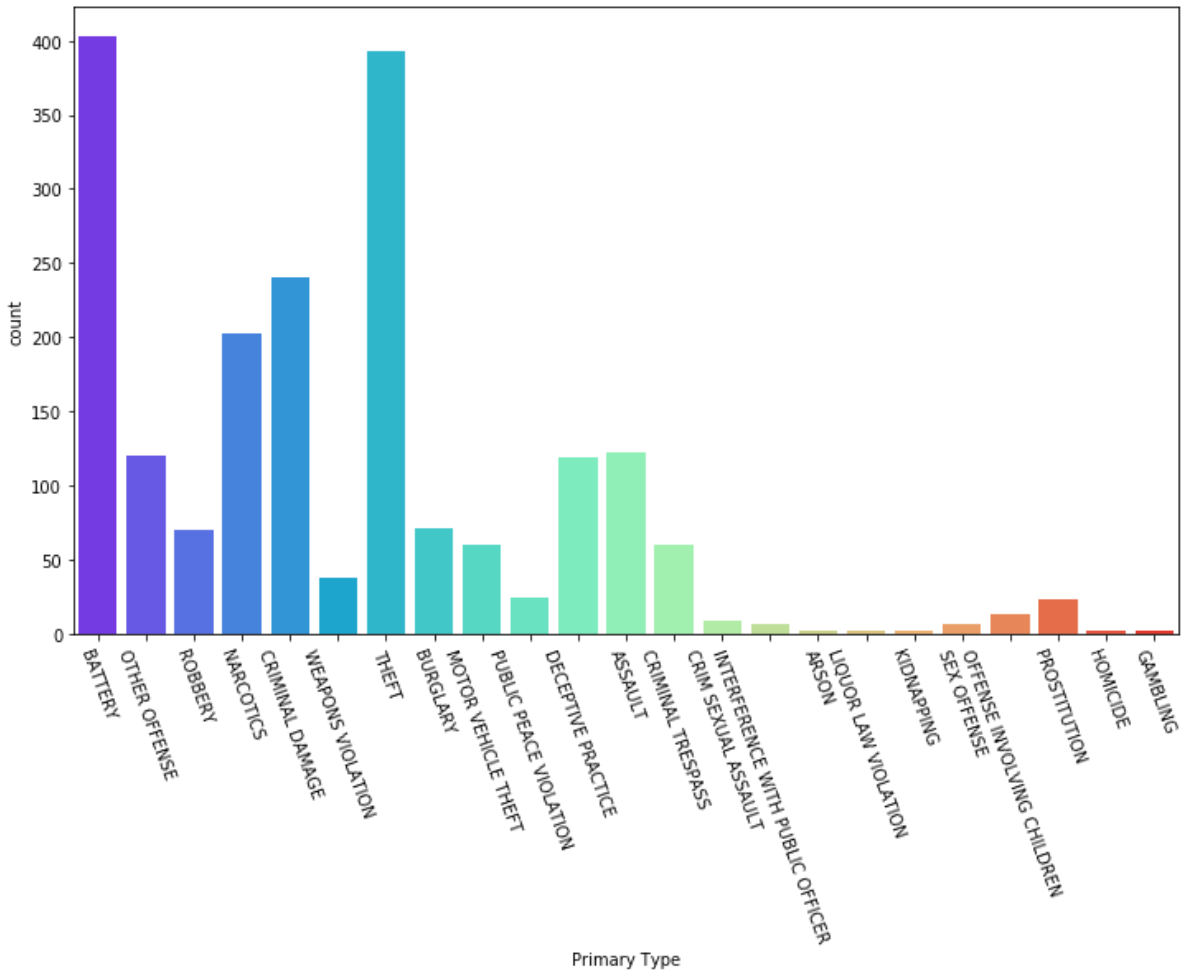
In [17]:
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
from pandas import ExcelWriter
from pandas import ExcelFile
from datetime import date
import calendar

df = pd.read_excel('crm.xlsx', sheet_name='crimes')
```

Let's take a better picture of the 'Primary Type' variable

In [2]:
```
plt.figure(figsize=(12,7))
plt.xticks(rotation=-70)
sns.countplot(x='Primary Type',data=df,palette='rainbow')
```

Out[2]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc9b6a42510>



It seems that there are a lot of different crime types and too much information for a countplot, maybe a different approch will help. Let's look on the date

In [3]:
```
df['Date'] = pd.to_datetime(df['Date'])
```

In [4]:
```
df['Hour'] = df['Date'].apply(lambda time: time.hour)
df['Month'] = df['Date'].apply(lambda time: calendar.month_name[time.month])
df['Day of Week'] = df['Date'].apply(lambda time: calendar.day_name[time.weekd
ay()])
```

let's create Primary Type by hour/day/month clustermaps to see if we can find more information. for Primary Type that didn't occur in a specific time we will replace the missing value with 0. the reason for that is because 0 crimes of a specific type occurred in that time unit.

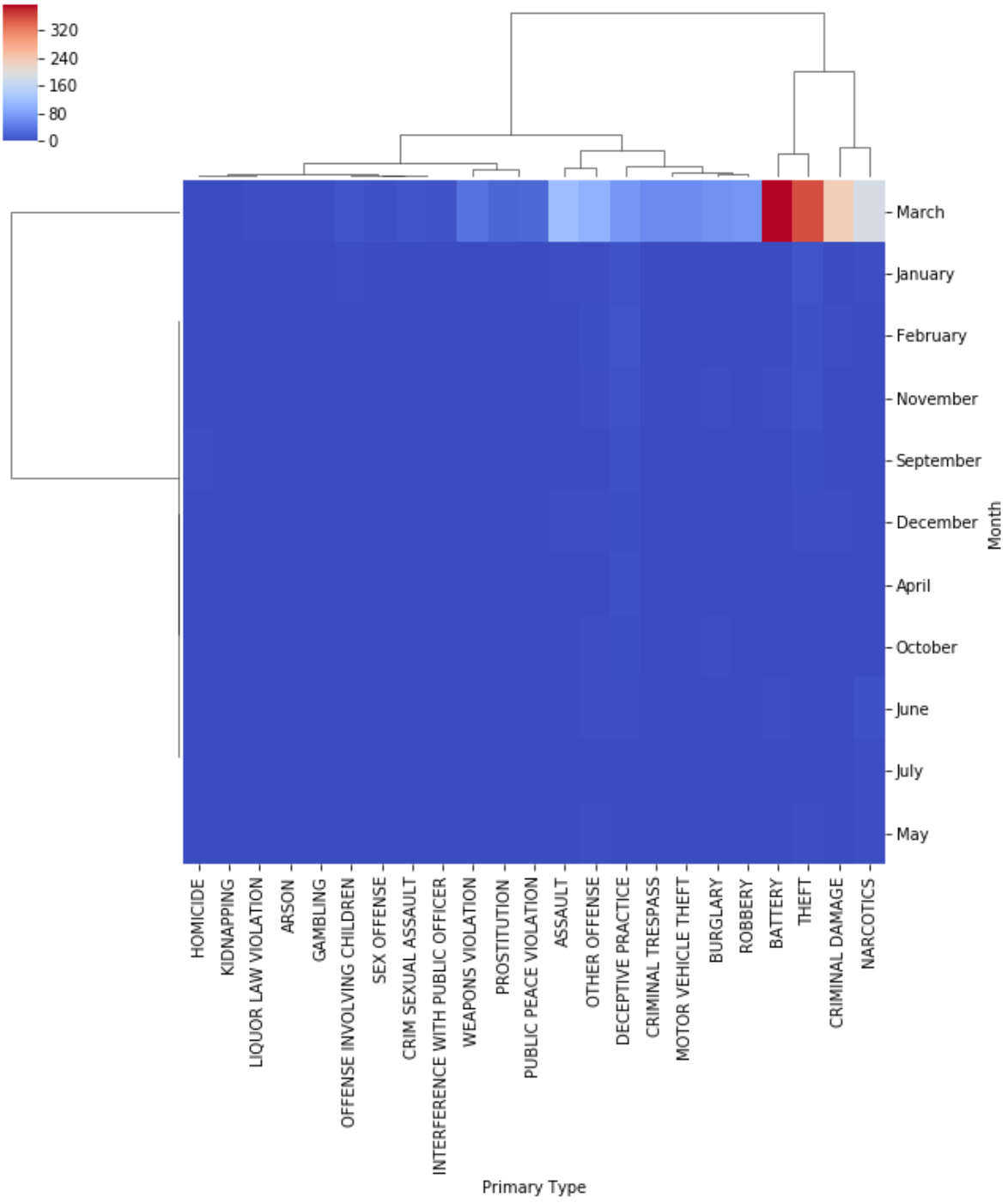
In [5]:
```
byDay = df.groupby(by=['Day of Week','Primary Type']).count()['ID'].unstack()
byDay = byDay.fillna(value=0)
```

In [6]:
```
byMonth = df.groupby(by=['Month','Primary Type']).count()['ID'].unstack()
byMonth = byMonth.fillna(value=0)
```

In [7]:
```
byHour = df.groupby(by=['Hour','Primary Type']).count()['Case Number'].unstack
()
byHour = byHour.fillna(value=0)
```

In [8]:
```
sns.clustermap(byMonth,cmap='coolwarm')
```

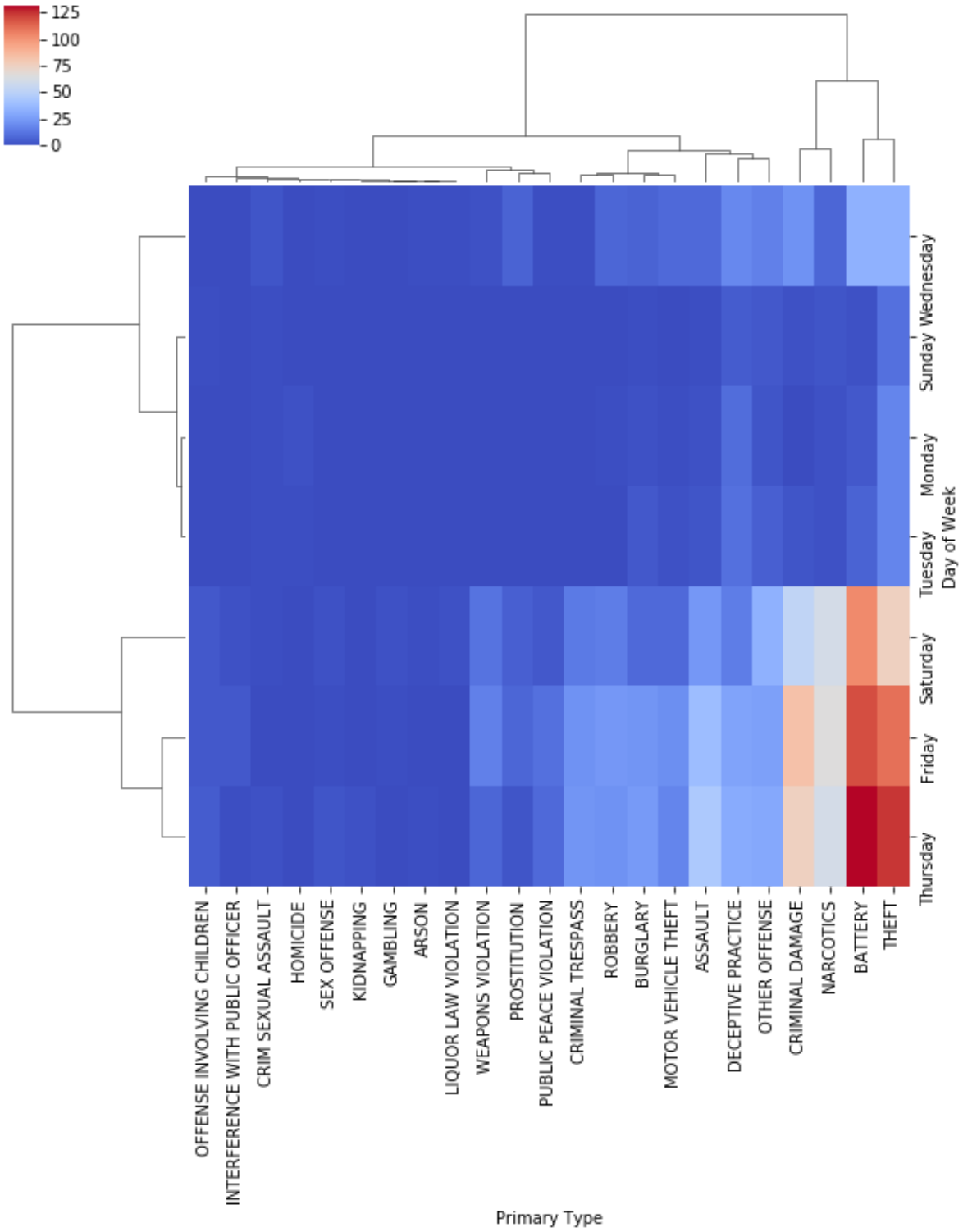
Out[8]: <seaborn.matrix.ClusterGrid at 0x7fc9e5fd3350>



We can see that most of the crimes happened in March

In [9]:
```
sns.clustermap(byDay,cmap='coolwarm')
```

Out[9]: <seaborn.matrix.ClusterGrid at 0x7fc9b484b750>

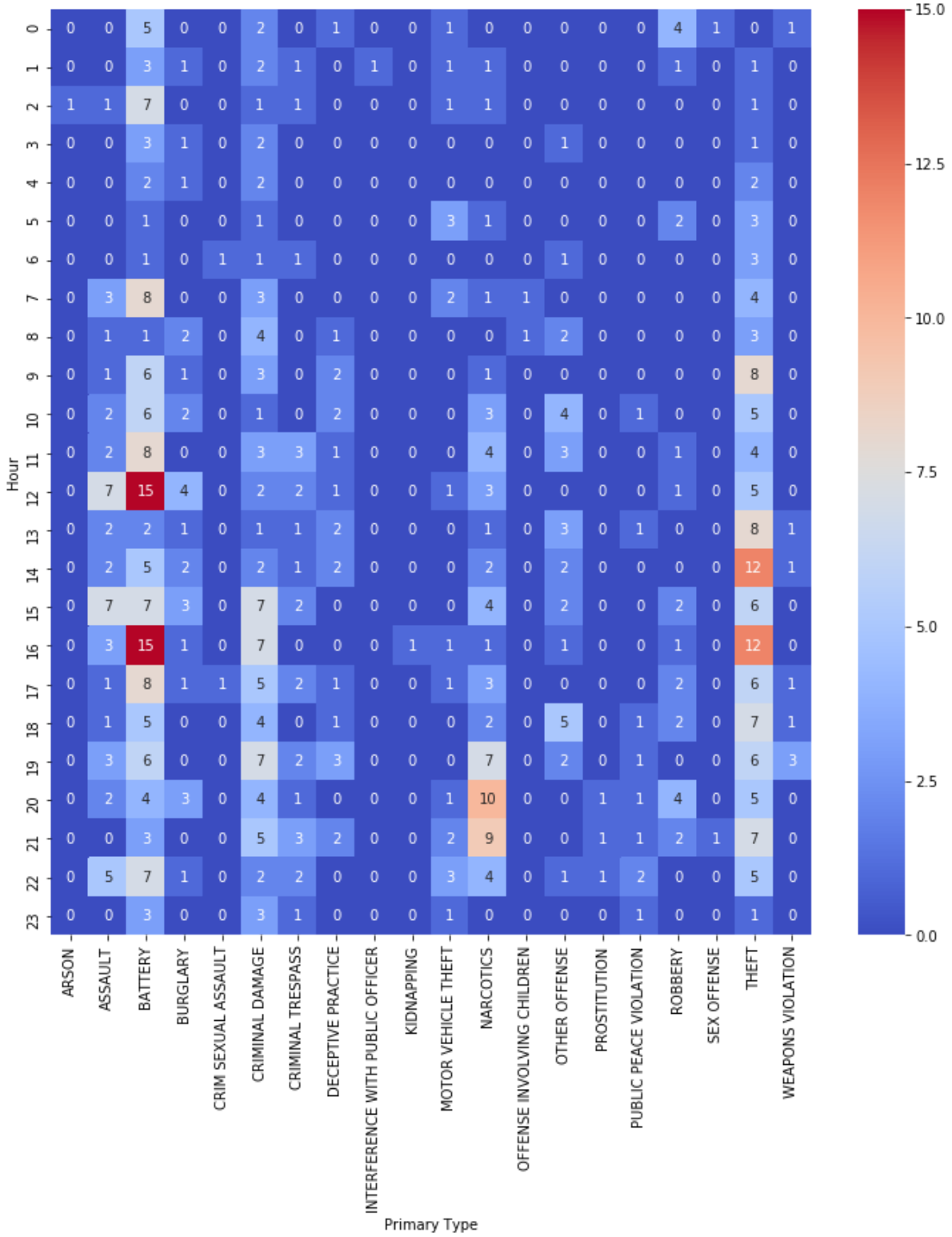


We can see that most of the crimes occur between Thursday to Saturday

Let's say that an officer in a specific month and day wants to know what kind of crime types will occur during the present day. We will create a function for an hour by Primary Type heatmap for a specific month and day

In [10]:
```
def md(month,day):
    df2=df[(df['Month']==month) & (df['Day of Week']==day)]
    byHour2 = df2.groupby(by=['Hour','Primary Type']).count()['Case Number'].u
nstack()
    byHour2 = byHour2.fillna(value=0)
    byHour2
    plt.figure(figsize=(12,12))
    sns.heatmap(byHour2,cmap='coolwarm',annot=True)
```

In [11]:
```
md('March','Thursday')
```



For a typical Thursday in March we can see that most of the crime will probably happen after 12 PM, perhaps a BATTERY crime type. later on, a THEFT crime type in 2 PM and a NARCOTICS crime type in 8 PM.

Still, even in the busiest month only a few crime types a likely to occur. let's focus on the top 5 most common crime types

In [12]:
```
df['Primary Type'].value_counts().head(5)
```

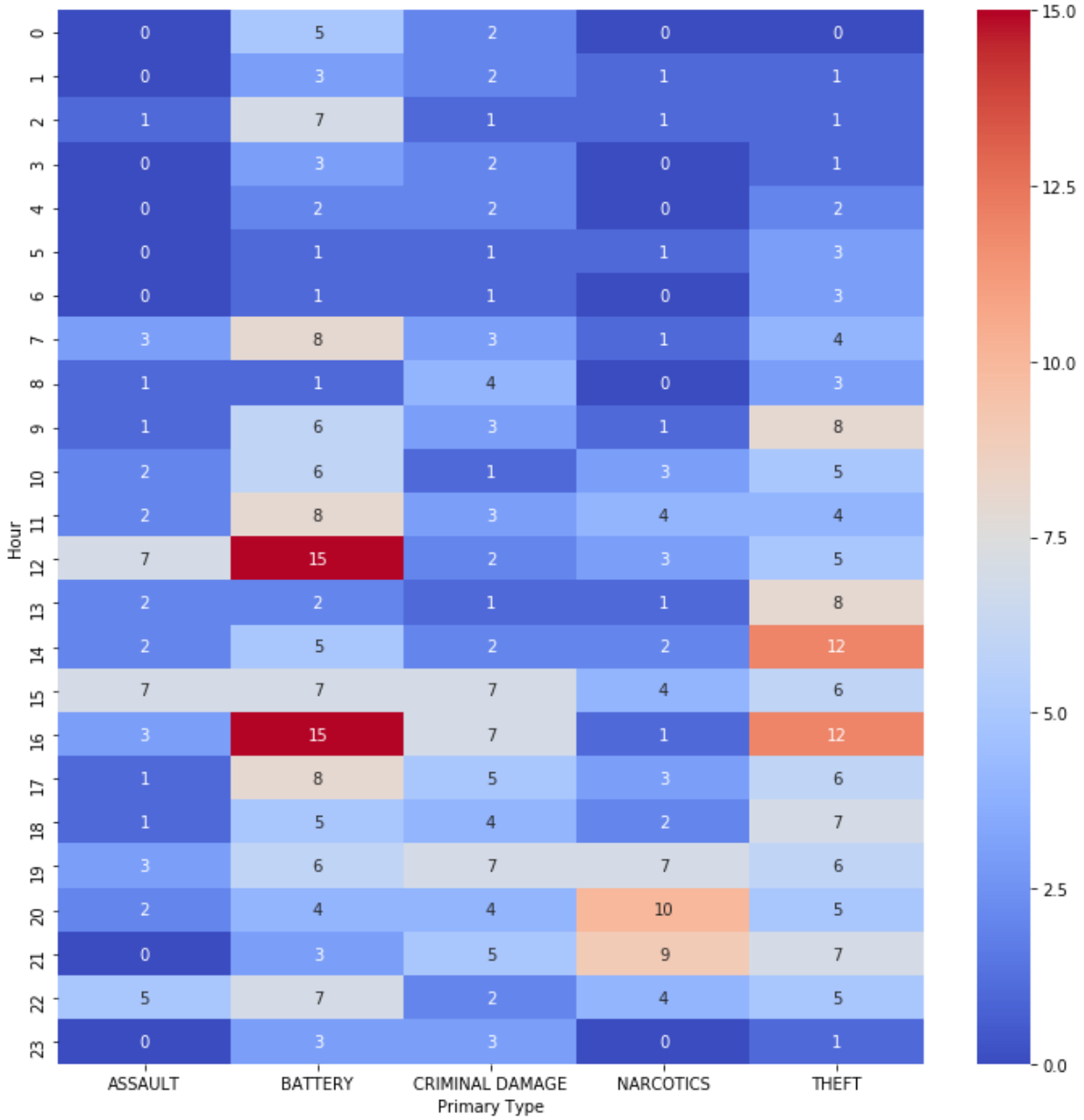
Out[12]:
```
BATTERY            403
THEFT              393
CRIMINAL DAMAGE    240
NARCOTICS          203
ASSAULT            123
Name: Primary Type, dtype: int64
```

In [13]:
```
topcrm = pd.DataFrame({'Primary Type': df['Primary Type'].value_counts().index
})[['Primary Type']].iloc[:5]
```

In [14]:
```
df = df.merge(topcrm,on='Primary Type',how='inner')
```

In [15]:
```
def md(month,day):
    df2=df[(df['Month']==month) & (df['Day of Week']==day)]
    byHour2 = df2.groupby(by=['Hour','Primary Type']).count()['Case Number'].u
nstack()
    byHour2 = byHour2.fillna(value=0)
    plt.figure(figsize=(12,12))
    sns.heatmap(byHour2,cmap='coolwarm',annot=True)
```

In [16]:
```
md('March','Thursday')
```



We have here the same results, but now it easier to focus in the most likely crime types

In this project I used clustermaps to highlight the month and day each type of crime is most likely to occur. I found that the weekends and the month of March have the highest probability for crimes to be committed. Seeing that the top 5 crimes contribute the lion share of the crimes committed, I found ample reason to focus solely on the aforementioned crimes