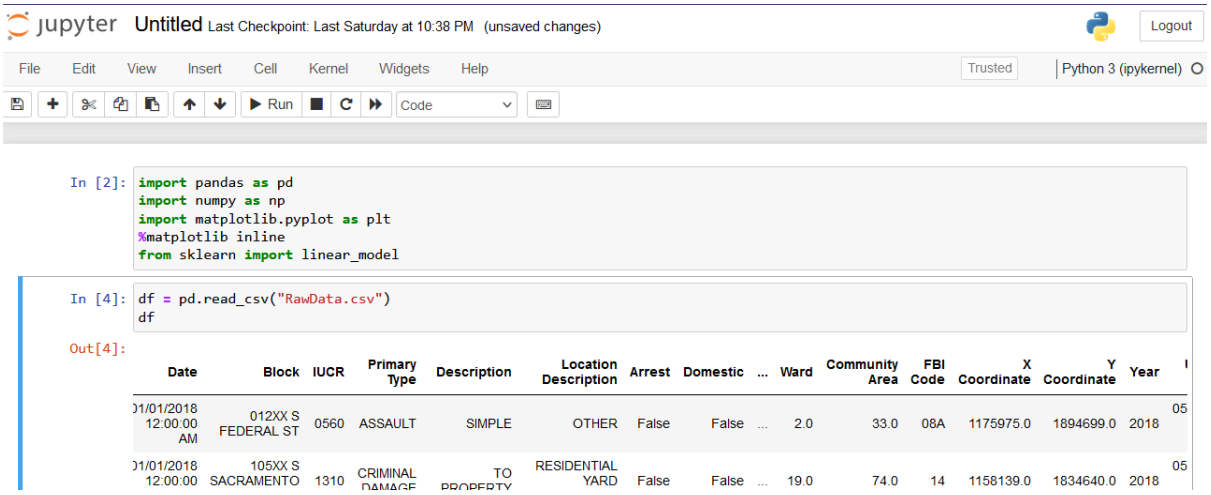
Data Pre-processing

Step1: Load libraries and the csv file on to Jupyter Notebook



jupyter Untitled Last Checkpoint: Last Saturday at 10:38 PM (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

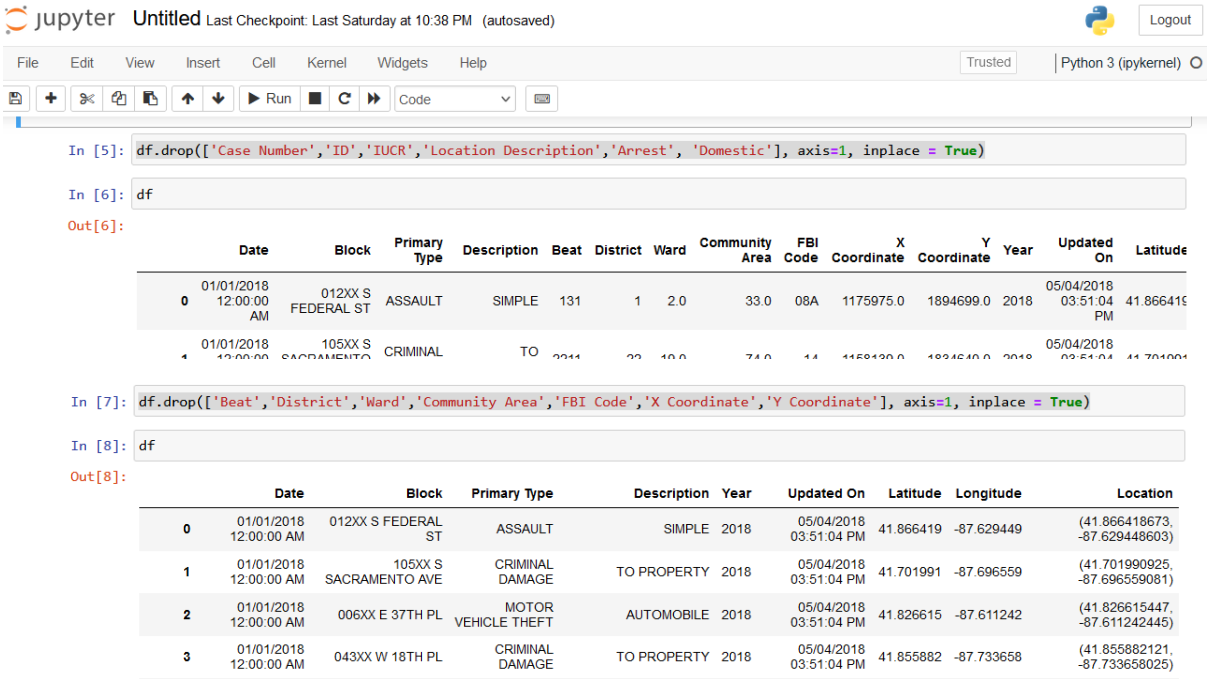
In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import linear_model
```

In [4]:
```
df = pd.read_csv("RawData.csv")
df
```

Out[4]:

Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	
01/01/2018 12:00:00 AM	012XX S FEDERAL ST	0560	ASSAULT	SIMPLE	OTHER	False	False	...	2.0	33.0	08A	1175975.0	1894699.0	2018	05
01/01/2018 12:00:00	105XX S SACRAMENTO	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENTIAL YARD	False	False	...	19.0	74.0	14	1158139.0	1834640.0	2018	05

Step 2: Drop the Irrelevant attributes in the dataset



jupyter Untitled Last Checkpoint: Last Saturday at 10:38 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [5]:
```
df.drop(['Case Number','ID','IUCR','Location Description','Arrest', 'Domestic'], axis=1, inplace = True)
```

In [6]:
```
df
```

Out[6]:

	Date	Block	Primary Type	Description	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude
0	01/01/2018 12:00:00 AM	012XX S FEDERAL ST	ASSAULT	SIMPLE	131	1	2.0	33.0	08A	1175975.0	1894699.0	2018	05/04/2018 03:51:04 PM	41.866419
1	01/01/2018	105XX S	CRIMINAL	TO	2211	22	19.0	74.0	14	1158139.0	1834640.0	2018	05/04/2018	41.701991

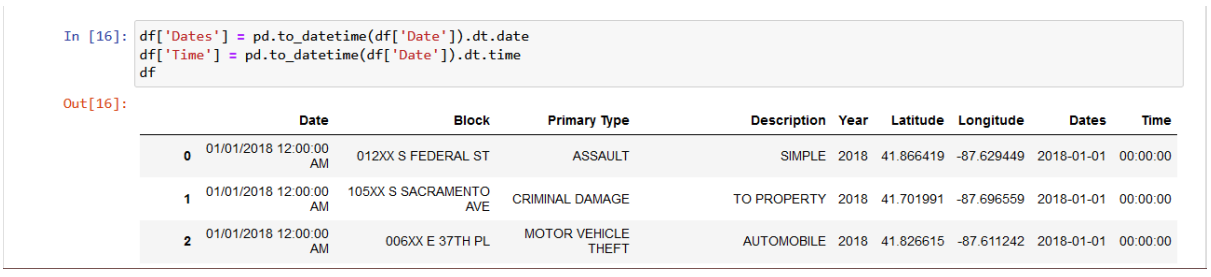
In [7]:
```
df.drop(['Beat','District','Ward','Community Area','FBI Code','X Coordinate','Y Coordinate'], axis=1, inplace = True)
```

In [8]:
```
df
```

Out[8]:

	Date	Block	Primary Type	Description	Year	Updated On	Latitude	Longitude	Location
0	01/01/2018 12:00:00 AM	012XX S FEDERAL ST	ASSAULT	SIMPLE	2018	05/04/2018 03:51:04 PM	41.866419	-87.629449	(41.866418673, -87.629448603)
1	01/01/2018 12:00:00 AM	105XX S SACRAMENTO AVE	CRIMINAL DAMAGE	TO PROPERTY	2018	05/04/2018 03:51:04 PM	41.701991	-87.696559	(41.701990925, -87.696559081)
2	01/01/2018 12:00:00 AM	006XX E 37TH PL	MOTOR VEHICLE THEFT	AUTOMOBILE	2018	05/04/2018 03:51:04 PM	41.826615	-87.611242	(41.826615447, -87.611242445)
3	01/01/2018 12:00:00 AM	043XX W 18TH PL	CRIMINAL DAMAGE	TO PROPERTY	2018	05/04/2018 03:51:04 PM	41.855882	-87.733658	(41.855882121, -87.733658025)

Step 3: Segregate Date and Time column into 2 separate columns:



In [16]:
```
df['Dates'] = pd.to_datetime(df['Date']).dt.date
df['Time'] = pd.to_datetime(df['Date']).dt.time
df
```

Out[16]:

	Date	Block	Primary Type	Description	Year	Latitude	Longitude	Dates	Time
0	01/01/2018 12:00:00 AM	012XX S FEDERAL ST	ASSAULT	SIMPLE	2018	41.866419	-87.629449	2018-01-01	00:00:00
1	01/01/2018 12:00:00 AM	105XX S SACRAMENTO AVE	CRIMINAL DAMAGE	TO PROPERTY	2018	41.701991	-87.696559	2018-01-01	00:00:00
2	01/01/2018 12:00:00 AM	006XX E 37TH PL	MOTOR VEHICLE THEFT	AUTOMOBILE	2018	41.826615	-87.611242	2018-01-01	00:00:00

Step 4: Reduce dataset size from 2018-Present to 2020-Present

```
In [23]: df.drop(df[(df['Year'] < 2020)].index, inplace=True)
```

```
In [24]: df
```

Out[24]:

	Date	Block	Primary Type	Description	Year	Latitude	Longitude	Dates	Time
182192	01/01/2020 12:00:00 AM	020XX N LOCKWOOD AVE	ASSAULT	SIMPLE	2020	41.917922	-87.758419	2020-01-01	00:00:00
182193	01/01/2020 12:00:00 AM	013XX S KILBOURN AVE	ASSAULT	AGGRAVATED PO: HANDGUN	2020	41.863700	-87.737326	2020-01-01	00:00:00
182194	01/01/2020 12:00:00 AM	082XX S MARYLAND AVE	BURGLARY	FORCIBLE ENTRY	2020	41.745354	-87.603799	2020-01-01	00:00:00
182195	01/01/2020 12:00:00 AM	040XX W 21ST ST	CRIMINAL DAMAGE	TO PROPERTY	2020	41.853481	-87.725910	2020-01-01	00:00:00
182196	01/01/2020 12:00:00 AM	057XX S PAULINA ST	BURGLARY	FORCIBLE ENTRY	2020	41.789476	-87.666898	2020-01-01	00:00:00
...	...	...	...	...	...	...	...	...	...

Step 5: Reformat the block attribute

```
In [26]: df['Area'] = df['Block'].str[7:]
```

```
In [32]: df
```

Out[32]:

	Date	Block	Primary Type	Description	Year	Latitude	Longitude	Dates	Time	Area
182192	01/01/2020 12:00:00 AM	020XX N LOCKWOOD AVE	ASSAULT	SIMPLE	2020	41.917922	-87.758419	2020-01-01	00:00:00	LOCKWOOD AVE
182193	01/01/2020 12:00:00 AM	013XX S KILBOURN AVE	ASSAULT	AGGRAVATED PO: HANDGUN	2020	41.863700	-87.737326	2020-01-01	00:00:00	KILBOURN AVE
182194	01/01/2020 12:00:00 AM	082XX S MARYLAND AVE	BURGLARY	FORCIBLE ENTRY	2020	41.745354	-87.603799	2020-01-01	00:00:00	MARYLAND AVE
182195	01/01/2020 12:00:00 AM	040XX W 21ST ST	CRIMINAL DAMAGE	TO PROPERTY	2020	41.853481	-87.725910	2020-01-01	00:00:00	21ST ST
182196	01/01/2020 12:00:00 AM	057XX S PAULINA ST	BURGLARY	FORCIBLE ENTRY	2020	41.789476	-87.666898	2020-01-01	00:00:00	PAULINA ST
...	...	...	...	...	...	...	...	...	...	...

Step 6: Drop the exisiting Block Column

```
In [50]: df.drop(['Block'], axis=1, inplace = True)
```

```
In [51]: df
```

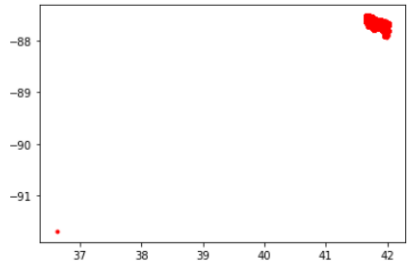
Out[51]:

	Date	Primary Type	Description	Year	Latitude	Longitude	Dates	Time	Area
182192	01/01/2020 12:00:00 AM	ASSAULT	SIMPLE	2020	41.917922	-87.758419	2020-01-01	00:00:00	LOCKWOOD AVE
182193	01/01/2020 12:00:00 AM	ASSAULT	AGGRAVATED PO: HANDGUN	2020	41.863700	-87.737326	2020-01-01	00:00:00	KILBOURN AVE
182194	01/01/2020 12:00:00 AM	BURGLARY	FORCIBLE ENTRY	2020	41.745354	-87.603799	2020-01-01	00:00:00	MARYLAND AVE

Step 7: Visualize the latitudes and the longitude:

```
In [42]: plt.scatter(df['Latitude'],df['Longitude'],color = 'red',marker = '.')
```

```
Out[42]: <matplotlib.collections.PathCollection at 0x1fc022f6c40>
```



Final Step : clean the data for better Visualization

In [43]:
```
#--------Because of that one corner value the mapping of points is messy----------------------------------------------------
```

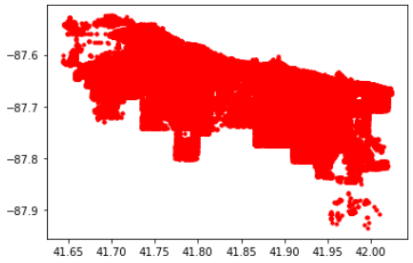
In [44]:
```
df['Latitude'].min()
```

Out[44]:
```
36.619446395
```

In [45]:
```
df.drop(df[(df['Latitude'] < 40)].index, inplace=True)
```

In [46]:
```
plt.scatter(df['Latitude'],df['Longitude'],color = 'red',marker = '.')
```

Out[46]:
```
<matplotlib.collections.PathCollection at 0x1fc046ca610>
```



In [47]:
```
#-----------------Now we get a clearer image of latitudes and longitudes-------------------------------------------------
```