Data Acquisition and Cleaning

Data Acquisition

The data acquired for this project is from Kaggle. data source of the project uses a Boston crime data that shows the crime per Street in Boston. The dataset contains the following columns:

INCIDENT_NUMBER
OFFENSE_CODE
OFFENSE_CODE_GROUP
OFFENSE_DESCRIPTION
DISTRICT
REPORTING_AREA
SHOOTING
OCCURRED_ON_DATE
YEAR
MONTH
DAY_OF_WEEK
HOUR
UCR_PART
STREET
Lat
Long
Location

Data Cleaning

The data preparation is done separately. From the Boston crime data, the crimes during the most recent year (2018) are only selected. And all the unnecessary fields are deleted.

```
df_boston = pd.read_csv('/Users/adityaghag/Desktop/crime.csv', encoding= 'unicode_escape')
df_boston.head()
```

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00

```
df_boston.drop(['OFFENSE_CODE', 'OFFENSE_CODE_GROUP', 'REPORTING_AREA', 'SHOOTING', 'DAY_OF_WEEK', 'UCR_PART', 'HOUR'],
```

```
df_boston.head()
```

	INCIDENT_NUMBER	OFFENSE_DESCRIPTION	DISTRICT	OCCURRED_ON_DATE	YEAR	MONTH	STREET	Lat	Long	Location
0	I182070945	LARCENY ALL OTHERS	D14	2018-09-02 13:00:00	2018	9	LINCOLN ST	42.357791	-71.139371	(42.35779134, -71.13937053)
1	I182070943	VANDALISM	C11	2018-08-21 00:00:00	2018	8	HECLA ST	42.306821	-71.060300	(42.30682138, -71.06030035)
2	I182070941	TOWED MOTOR VEHICLE	D4	2018-09-03 19:27:00	2018	9	CAZENOVE ST	42.346589	-71.072429	(42.34658879, -71.07242943)
3	I182070940	INVESTIGATE PROPERTY	D4	2018-09-03 21:16:00	2018	9	NEWCOMB ST	42.334182	-71.078664	(42.33418175, -71.07866441)
4	I182070938	INVESTIGATE PROPERTY	B3	2018-09-03 21:05:00	2018	9	DELHI ST	42.275365	-71.090361	(42.27536542, -71.09036101)

```
df_boston.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65685 entries, 0 to 65684
Data columns (total 10 columns):
INCIDENT_NUMBER         65685 non-null object
OFFENSE_DESCRIPTION     65685 non-null object
DISTRICT                65141 non-null object
OCCURRED_ON_DATE        65685 non-null object
YEAR                    65685 non-null int64
MONTH                   65685 non-null int64
STREET                  64542 non-null object
Lat                     61464 non-null float64
Long                    61464 non-null float64
Location                65685 non-null object
dtypes: float64(2), int64(2), object(6)
memory usage: 5.0+ MB
```