NYPD Crime data

Summary -

The data provides an insight into crime in new york city

The various columns are analysed as follows -

1)Complaint\_from\_date (represents exact date when the even occurred) column2

These are the top 10 values for the complaints specifying dates when they occured and the no of complaints . Overall we can see that the most complaints/crimes occur during the New year days

01/01/2010 2343

01/01/2008 2273

01/01/2007 2259

01/01/2011 2253

01/01/2014 2211

01/01/2006 2156

08/01/2007 2053

01/01/2009 2036

01/01/2013 2030

01/01/2012 2025

2)Complaint\_from\_time (represents exact time the crime occurred) column3

Following top 10 values indicate that crimes are more probable after 12 pm than before 12 pm

12:00:00 135632

15:00:00 116221

18:00:00 112102

20:00:00 106703

16:00:00 103994

17:00:00 103510

19:00:00 100290

14:00:00 92385

21:00:00 92016

22:00:00 90635

3)report\_date (date when report was reported to NYPD) column6

Top 10 values

5/29/2007 1903

11/01/2006 1889

10/23/2007 1854

06/21/2006 1841

06/05/2007 1829

06/20/2006 1820

07/05/2006 1803

07/29/2008 1791

05/30/2006 1785

07/31/2007 1783

4)offense\_code

offense code and number of offenses column7

341 822508

578 604141

344 521611

351 433368

109 429205

235 285801

361 283102

105 198774

107 191407

106 184124

5)offense\_description column8

description for offense codes

PETIT LARCENY 822498

HARRASSMENT 2 604070

ASSAULT 3 & RELATED OFFENSES 521538

CRIMINAL MISCHIEF & RELATED OF 505774

GRAND LARCENY 429196

DANGEROUS DRUGS 348469

OFF. AGNST PUB ORD SENSBLTY & 283065

ROBBERY 198772

BURGLARY 191406

FELONY ASSAULT 184069

as we can see 4 is not exactly equal to 5 in terms of the second column. This shows a slight error in data collection at source.

6) crime indicator (completed vs attempted) column11

COMPLETED 5013311

ATTEMPTED 87913

7)boro\_names The names of the borough in which the incident occurred column14

BROOKLYN 1526213

MANHATTAN 1216249

BRONX 1103514

QUEENS 1011002

STATEN ISLAND 243790

Time series analysis in python pandas -

On running time series analysis in python pandas we were unable to initially load the data as the csv file is 1.3 gb .

we got an initial error saying - Specify dtype option on import or set low\_memory=False

1)This meant that we need to remove the memory restrictions by providing low\_memory=False

but this method is not time efficient . What it does it sets the memory limit to false

2) Instead setting dtype to several values helps in saving time as pandas does not have to specifically scan the entire row to intially find the datatype.

Therefore we wwent with dtype={"HADEVELOPT": str } as an additional parameter in the read\_csv() function

On doing this we still got a warning of #DtypeWarning: Columns (17) have mixed types. Specify dtype option on import or set low\_memory=False.

Hence we added column 17 to the function ie- data = pd.read\_csv('/home/karan/crime.csv', dtype={"HADEVELOPT": str ,"PREM\_TYP\_DESC":str})

This enabled us to retreive the data but the time required to retreive it was long ( as the dataset is large)