Big Data Project Progress Report - Part I

Li Lin Qin (llq205) Yidi Zhang (yz3464) Yurui Mu (ym1495)

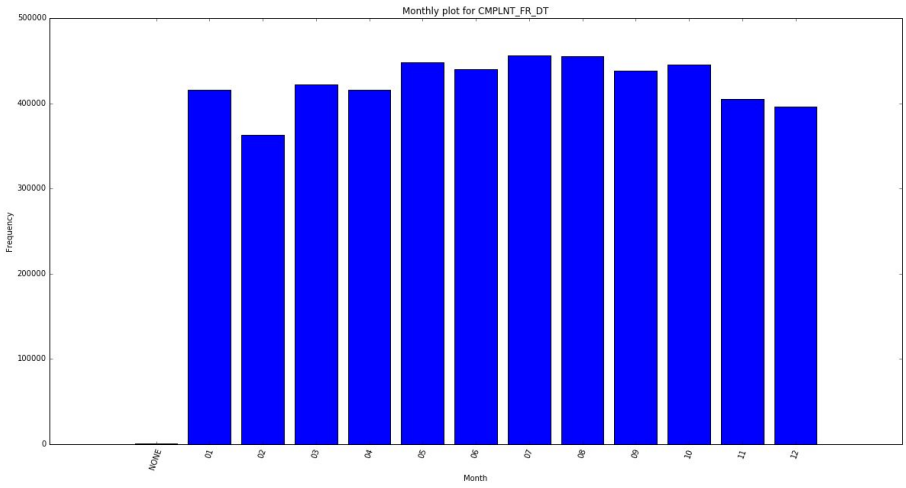
For this project, the dataset we will be using is NYPD Complaint Historical Data, which includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2015.

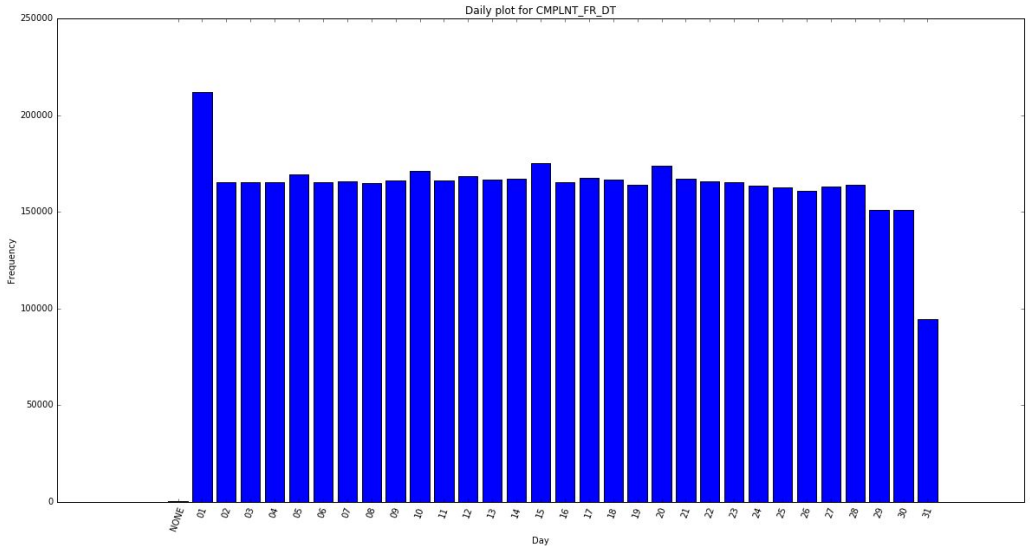
For the first part, we will analyze the data and generate a descriptive summary of their contents as well as a list of data quality issues. We start by providing comprehensive analysis for each of the column in the dataset. The code for generating the summary table can be found in our github site.

There are 24 columns in the NYPD_Complaint_Historic.csv.

1. Complaint Number ID is Randomly generated persistent ID for each complaint. There are 5101231 unique integers assigning to each cases, which is essentially the number of observations in our dataset. All of the entries are valid integers.

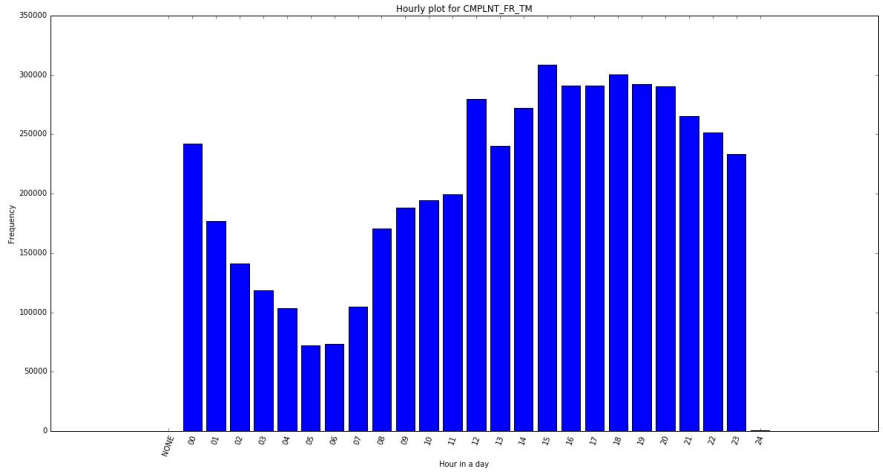
2. Complaint From Date is the date each event is filed. Using datetime.date() function, there are 5100576 valid DATETIME type dates and 655 null values. Among all, most complaints were filed in 2005-2015. There are also 7 records with dates in 1015, which might be typos, here we still consider them as valid, because they fit in the form of '%m/%d/%Y'. When performing analysis, we should change them into year 2015. In addition, there are also early records dated back to 1900s. Since those records do not seem to be mistakes, we counted them as valid records along with the others.





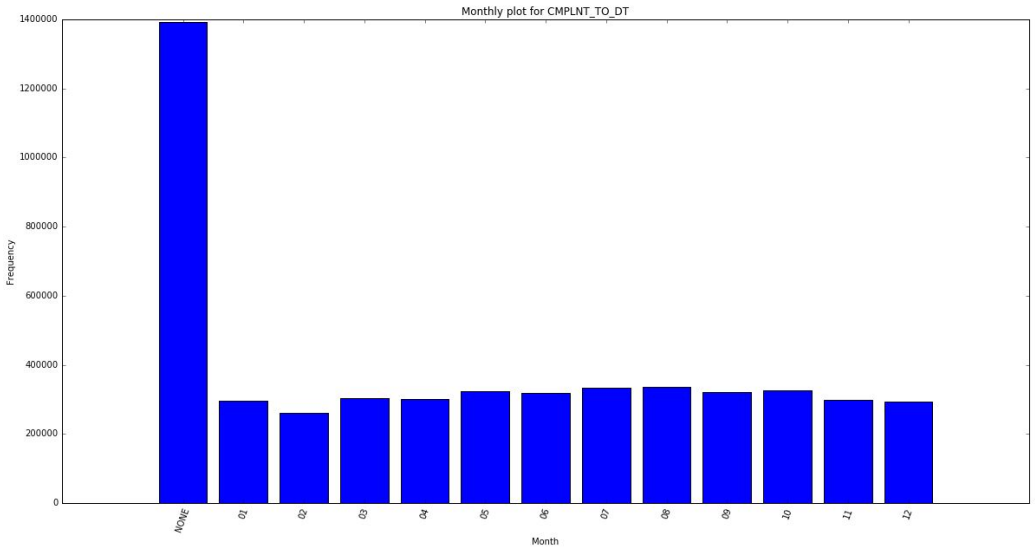
The above graphs indicate counts in a monthly or daily basis over years. The crimes spread rather evenly among all the months, with slight fewer ones in February. We can see that a lot more crimes happened during the first day of the month. Our guess is that it is probably a default number when the exact day of the month is unknown. Furthermore, the 31st of a month had least crimes, for the reason that some months do not have 31st as their last day.

3. Complaint From Time is the specific time of the day each event is filed. Using datetime.time() function, there are 5100280 valid DATETIME type time, 903 invalid time, and 48 null values. Invalid values come from '24:00:00', whereas midnight is defined as '00:00:00'. Thus we are not sure if it happens at midnight or did they simply don't know about the time.



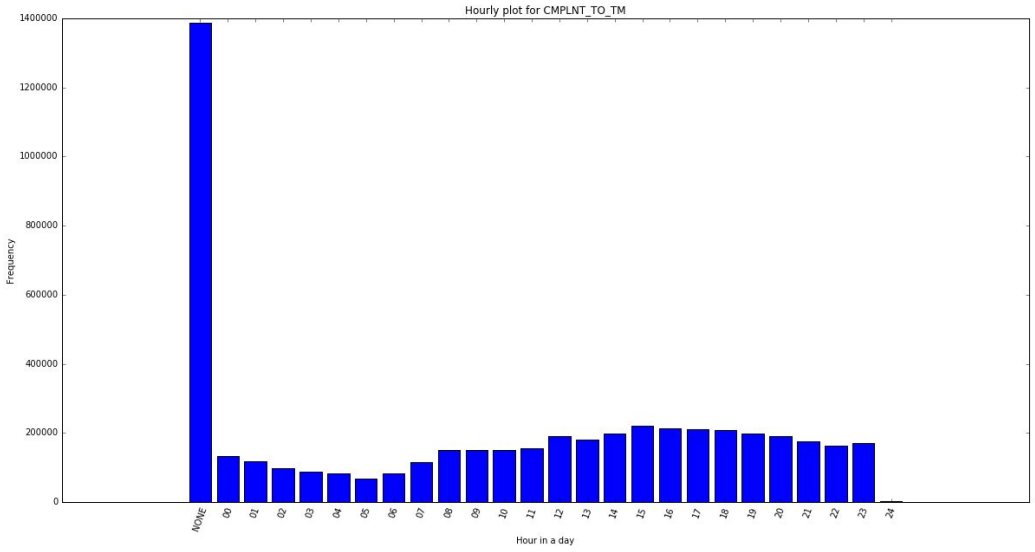
A majority of the crimes happened in the afternoon till later of the day. Fewer crimes took place during midnight or early morning.

4. Complaint To Date is the date when event ended. Using datetime.date() function, we found 3709213 valid DATETIME type dates, 540 invalid DATETIME type dates and 1391478 null values. Invalid dates include 1 record in '2090' and 539 invalid ending dates before starting dates.



Most of the entries are null values, as shown in the graph. The remaining ones are spread relatively evenly among all the months in a year.

5. Complaint To Time is specific time of the day when event ended. Using datetime.date() function, we found 3712070 valid DATETIME type time, 1376 invalid time(24:00:00), and 1387785 null values. From the below graph, it is clear that most of the values are null.



6. Complaint Report Date is the date police filed reports. In this datetime column, all of the 5101231 values are valid, ranging from year 2006 to 2015.

7. Key Code is the three digits code assigned to each offense category. Type Integer. There are no missing values. All codes are valid three digits codes, splitting into 74 categories. Then we pair up KY_CD with OFNS_DESC as key, we received 122 pairs of keys. Other than the reason that there are cases with valid key code, while missing offense descriptions. There are also issues due to multiple names of a single key code, like ‘kidnapping’ vs ‘kidnapping & related’ and ‘agriculture & mrkts law-unclassified’ vs ‘other state laws’.

8. Offense Description is the semantic description of offenses reported. It includes 70 unique valid string values and 19080 null values.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
OFNS_DESC	String	Offense Description	5082151	0	19080

9. PD Code is a three digit integer representing the internal classification code. This column includes 415 unique valid float values and 4574 null values.

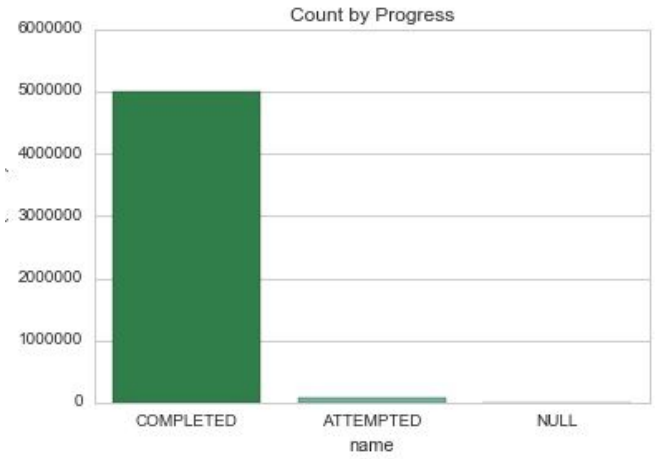
Column name	Basetype	Semantic Type	Valid	Invalid	Null
PD_CD	Float	Classification Code	5096657	0	4574

10. PD Description is a text description of internal classification corresponding with PD code. This column includes 403 unique valid string values and 4574 null values.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
PD_DESC	String	Classification Description	5096657	0	4574

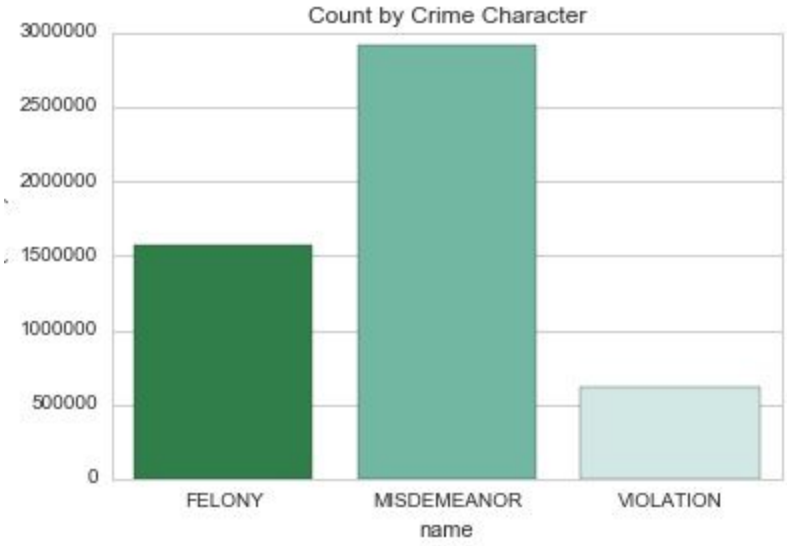
11. Crime Attempted Code is an integer indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely. ‘Completed’ and ‘attempted’ are two values contained in this column, and nearly 80% of the values are ‘completed’.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
CRM_ATPT_CPTD_CD	String	Progress	5101224	0	7



12. Law Category Code is an integer describing the level of offense: felony, misdemeanor, violation.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
LAW_CAT_CD	String	Law Category	5101231	0	0



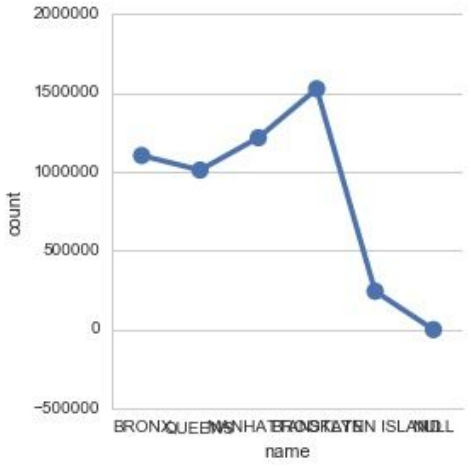
13. Jurisdiction Decision is text indicating the jurisdiction responsible for incident. It has 25 unique values without null type.

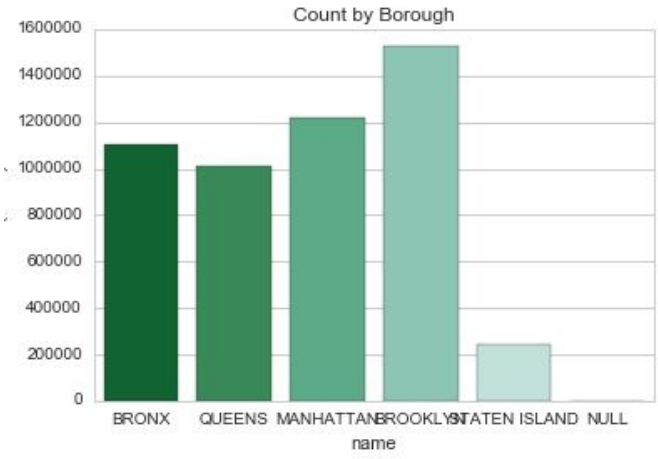
Column name	Basetype	Semantic Type	Valid	Invalid	Null
JURIS_DESC	String	Jurisdiction Decision	5101231	0	0

14. Borough Name is the name of the borough in which the incident occurred. It is a string type. This column contains 5 unique string values, including 'BRONX', 'QUEENS',

'MANHATTAN', 'BROOKLYN', 'STATEN ISLAND', which is five borough name in New York.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
BORO_NM	String	Classification Code	5100768	0	463

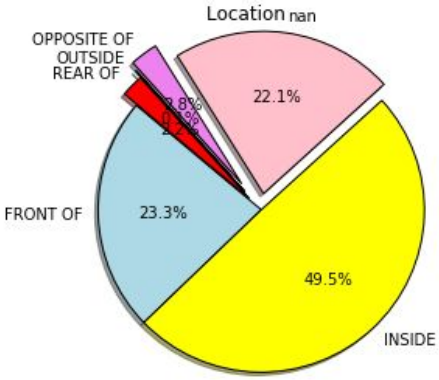




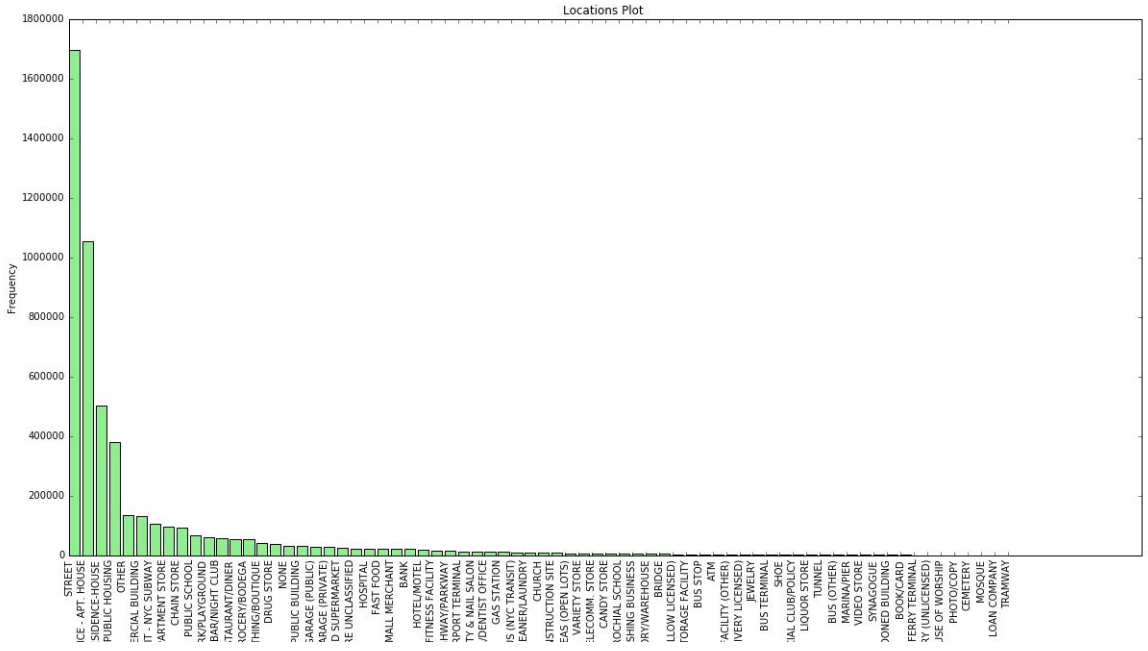
15. Address Precinct: The precinct in which the incident occurred.

Column name	Basetype	Semantic Type	Valid	Invalid	Null
ADDR_PCT_CD	Float	Address Precinct	5100841	0	390

16. Location of Occurrence is the relative location of occurred around places. Entries include “front of”, “inside”, “opposite of”, “outside”, and “rear of”, with respective counts of 1189787, 2527543, 140606, 2765,113189. There are also 1,127,128 null values in this column.

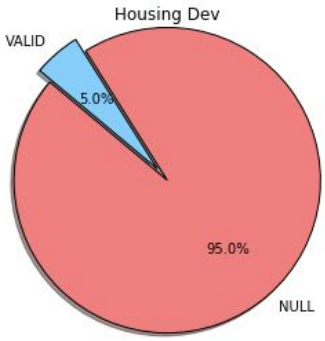


17. PREM_TYP_DESC: Specific description of premises; grocery store, residence, street, etc. There are 33279(0.65%) null values and 5067952(99.35%) valid values. There are 70 valid types in total.

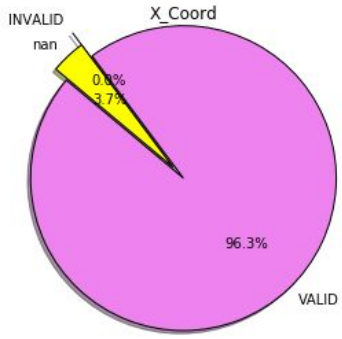


18. Park names includes text of the name of NYC park, playground or greenspace of occurrence, if applicable. There are 7597 valid entries (0.15%) and 5093634 null values (99.85%). It seems that a majority of crime are happening outside of NYC park area. Among all 863 parks, Central Parks has the most crimes reported: 543. Following by Flushing Meadows Corona Park with 301 crimes, and Riverside Park with 188 crimes reported.

19. Hadevelpt is the name of NYCHA housing development of occurrence, if applicable. There are a total of 4848026 null values, as well as 253205 valid ones. Among all valid values, there are 278 unique types.



20. X Coordinate is a float number representing x-coordinate of the place where the crime happened. There are 4913085 valid entries and 188146 null values. The valid entries are all within the New York area.



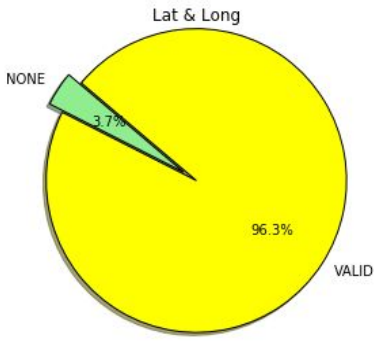
21. Y Coordinate is a float number representing the y-coordinate of the place where the crime happened. Same as the column of x-coordinate, there are 4913085 valid entries and 188146 null values. The valid entries are all within the New York area. The plot should look the same as the above graph for X Coordinate.

22. Latitude is a float number depicting the latitude where the crime took place. The valid entries are all within the New York area. There are 188146 empty entries and 4913085 valid counts.

23. Longitude is a float number depicting the longitude where the crime took place. The valid entries are all within the New York area. There are 188146 empty entries and 4913085 valid counts.

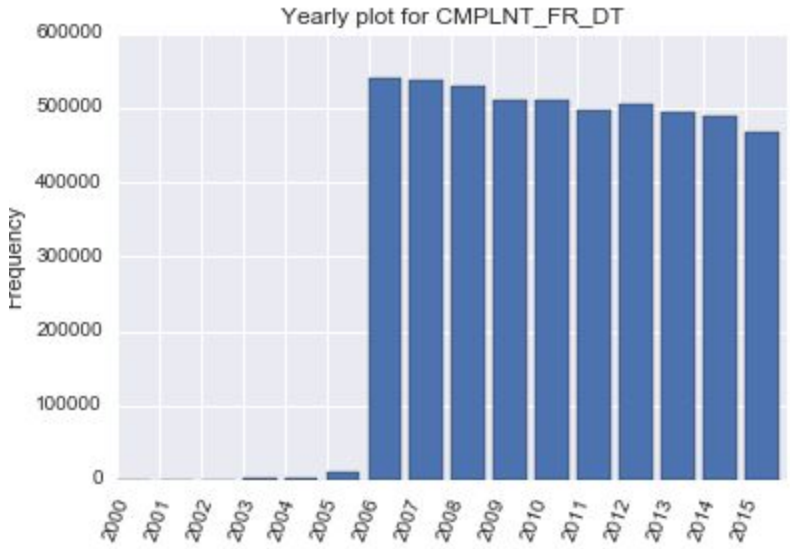
24. Lat_Lon is a column that combines the two previous column, latitude and longitude to provide a location. It has a form of “(float, float)”. So we checked for validity by

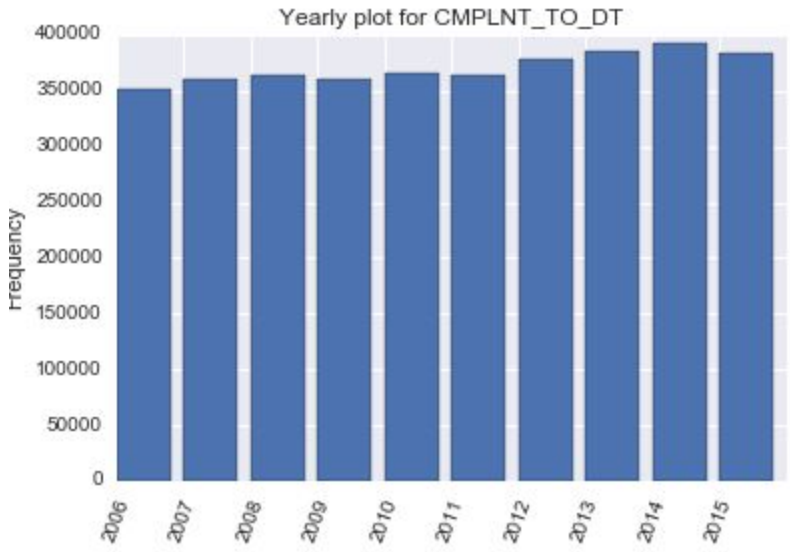
confirming that the two numbers in the bracket match the two previous numbers in the same row. Besides 188146 empty entries, the remaining 4913085 are all valid.



Data Sets issue:

1. PD_CD and PD_DESC should have same number of unique values, however not. We will explore and visualize this questions in the next step.
2. There are some invalid values in DATETIME columns, such as 1015 (change to 2015). And count before 2006 is far smaller compared to the count after 2006. After cleaning, the result are shown as follows:




Yearly plot for CMPLNT_TO_DT
Frequency
400000
350000
300000
250000
200000
150000
100000
50000
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015