Data Description

For this class, I have selected a data package called NYPD_Complaint_Data_Historic. This data consists of 5,580,035 observation and 24 variables collected between 2006 and 2016 by the New York Police Department (Fig.1, Fig 2). While there are 24 variables, many of the variables report the same variable at different levels. The variables KY_CD, OFNS_DESC, PD_CD, PD_DESC, and LAW_CAT_CD all give information on the type of crime committed. KY_CD and OFNS_DESC report the exact same information; KY_CD gives a numerical code for the type of crime, while OFNS_DESC gives the description matching the numerical code. PD_DESC then clumps the OFNS_DESC into more general categories (with an associated numerical code in PD_CD), and lastly LAW_CAT_CD classifies each crime as either a felony, misdemeanor, or violation. (E.g., 344 = Assault 3 & Related Offenses > 101 = Assault 3 > Misdemeanor). Likewise, the last five variables (X_COORD_CD through Lat_Lon) all concern exact location coordinates for each crime. Other data include neighborhood, venue type, time and date of the incident and the report, whether the crime was successful or not, the name of the park or housing development if applicable, jurisdiction, and so on.

Project Aim

My goal at the end of the project is to predict the type of crime occurring given the location and the time. For example, if a police car responds to a call from the Bronx between the hours of 2am and 3am on a weekend, is the crime more likely to be x, y, or z? As this data has no bearing on my thesis work, I will most likely not pursue it after the class, and thus my 'class aim' is the same as my overall aim.

While working towards this goal, I will also look at various potential trends one might expect to find in the data. For example, is there any association with crime type and frequency with holidays? Are there regional differences in the type of crime?

Basic Model

The response variable will be some level of crime classification. I expect my primary predictors to be location (at some level), and time, possibly with some interaction with day or date.

Next Steps

I am currently at the point of sorting, organizing, and cleaning my data.

Selections

- To simplify my data, I will select only for Completed crimes.

Deletions

- Because of the nature of many of the crimes, I will nix the following columns: CMPLNT_TO_DT, CMPLNT_TO_TM, and RPT_DT and focus solely on the date and time of the start of the crime. Likewise, CMPLNT_NUM is somewhat useless for my purposes.
- Because X_COORD_CD and Y_COORD_CD are useful only on the New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIP 3104) Map, I will nix those variables. Due to redundancy, I will delete Lat_Lon, and keep Latitude and Longitude in their separated columns. I will also delete ADDR_PCT_CD, or the address code for each precinct.

Compressions

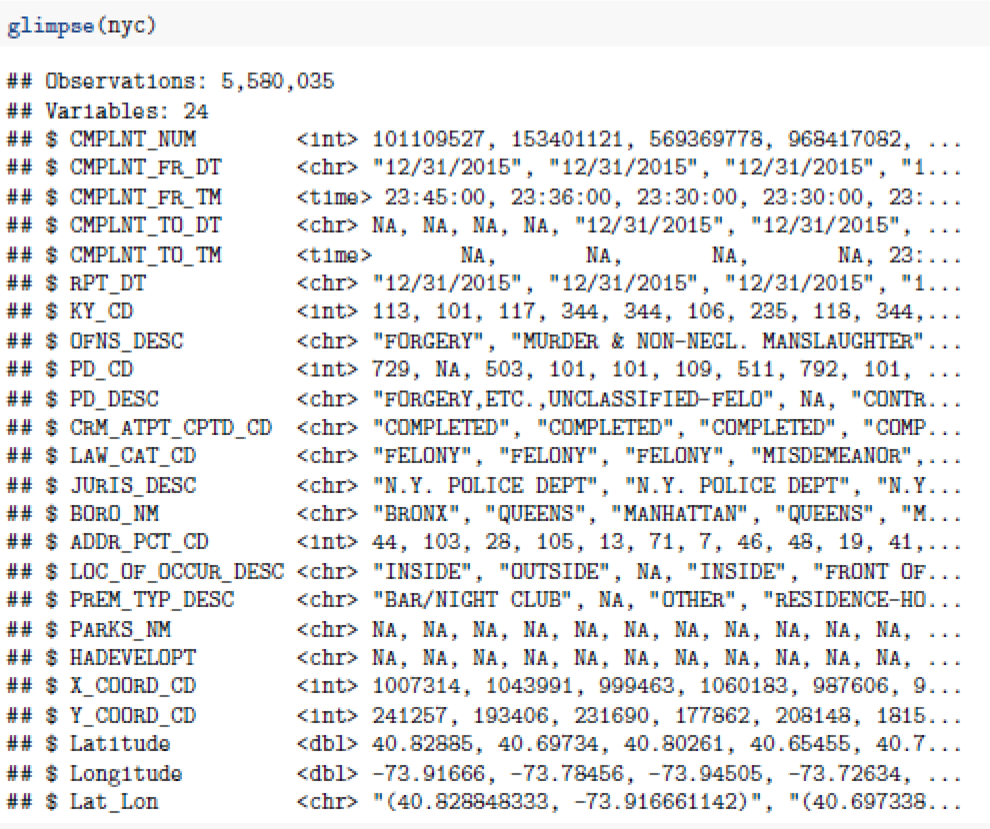
- LOC_OF_OCCUR_DESC will be compressed into Inside and Outside (possibly as a simple indicator variable), PARKS_NM will be reduced to an indicator variable (0 for NA, 1 for all others), and the same with HADEVELOPT (housing development).
- Based on feedback, I agree that binning the time into one-hour bins will greatly simplify my data.

Insertions

- I will splice CMPLNT_FR_DT into three separate columns: year, month, and day. I will also generate a new variable that classifies each date as Weekday or Weekend (perhaps using R's isWeekday/isWeekend functions).
- In order to analyze patterns regarding holidays, I will generate a categorical variable for each date: 0 for non-holiday, 1 for holiday (corresponding to New Year, Valentine's Day, St. Patrick's Day, Easter, Memorial Day, Labor Day, July 4th, Halloween, Thanksgiving, and Christmas).

Questions

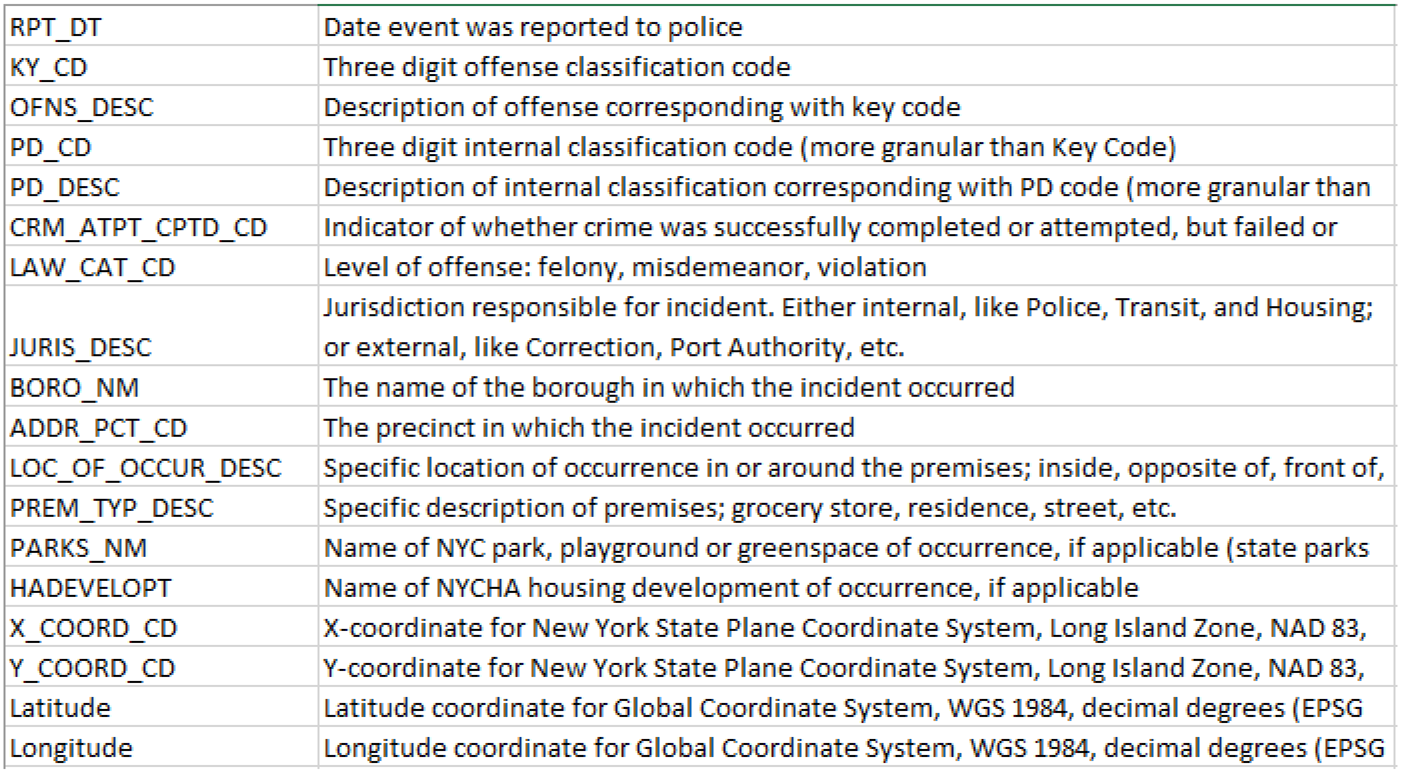
- For jurisdiction, I can either select only for crimes that fell under NYPD jurisdiction and delete the rest, compress the variable into NYPD and !NYPD as an indicator variable, or just straight up delete the column. I'm not entirely sure what patterns would emerge from this variable and am inclined to ignore it.
- Because Thanksgiving and Christmas are such large holidays –generally three to four days surrounding the actual day of the holiday (and Memorial Day is often associated with Memorial Weekend)–I'm not sure if I should just code a week for those or stick with the actual holiday day as 1 and all surrounding days as 0.
- I am still unsure how to compress the crime type levels. I am inclined to use the second level, PD_DESC, and LAW_CAT_CD, ignoring the most descriptive level in favor for the higher, more broad categorical levels.



```
glimpse(nyc)
```

```
## Observations: 5,580,035
## Variables: 24
## $ CMPLNT_NUM        <int> 101109527, 153401121, 569369778, 968417082, ...
## $ CMPLNT_FR_DT      <chr> "12/31/2015", "12/31/2015", "12/31/2015", "1...
## $ CMPLNT_FR_TM      <time> 23:45:00, 23:36:00, 23:30:00, 23:30:00, 23:...
## $ CMPLNT_TO_DT      <chr> NA, NA, NA, NA, "12/31/2015", "12/31/2015", ...
## $ CMPLNT_TO_TM      <time>       NA,       NA,       NA,       NA, 23:...
## $ RPT_DT            <chr> "12/31/2015", "12/31/2015", "12/31/2015", "1...
## $ KY_CD             <int> 113, 101, 117, 344, 344, 106, 235, 118, 344,...
## $ OFNS_DESC         <chr> "FORGERY", "MURDER & NON-NEGL. MANSLAUGHTER"...
## $ PD_CD             <int> 729, NA, 503, 101, 101, 109, 511, 792, 101, ...
## $ PD_DESC           <chr> "FORGERY,ETC.,UNCLASSIFIED-FELO", NA, "CONTR...
## $ CRM_ATPT_CPTD_CD  <chr> "COMPLETED", "COMPLETED", "COMPLETED", "COMP...
## $ LAW_CAT_CD        <chr> "FELONY", "FELONY", "FELONY", "MISDEMEANOR",...
## $ JURIS_DESC        <chr> "N.Y. POLICE DEPT", "N.Y. POLICE DEPT", "N.Y...
## $ BORO_NM           <chr> "BRONX", "QUEENS", "MANHATTAN", "QUEENS", "M...
## $ ADDR_PCT_CD       <int> 44, 103, 28, 105, 13, 71, 7, 46, 48, 19, 41,...
## $ LOC_OF_OCCUR_DESC <chr> "INSIDE", "OUTSIDE", NA, "INSIDE", "FRONT OF...
## $ PREM_TYP_DESC     <chr> "BAR/NIGHT CLUB", NA, "OTHER", "RESIDENCE-HO...
## $ PARKS_NM          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ HADEVELOPT        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ X_COORD_CD        <int> 1007314, 1043991, 999463, 1060183, 987606, 9...
## $ Y_COORD_CD        <int> 241257, 193406, 231690, 177862, 208148, 1815...
## $ Latitude          <dbl> 40.82885, 40.69734, 40.80261, 40.65455, 40.7...
## $ Longitude         <dbl> -73.91666, -73.78456, -73.94505, -73.72634, ...
## $ Lat_Lon           <chr> "(40.828848333, -73.916661142)", "(40.697338...
```

Figure 1: glimpse(nyc) displays the relevant information about the dataset.



RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of,
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83,
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83,
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG

Figure 2: Table of Variable names and descriptions.

Feedback

I: 5, L: 4, F: 5