NYC Crime Project

Zane Wolf

September 27, 2017

I have chosen a data set called NYPD_Complaint_Data_Historic, available at https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.

> *Can I produce a model that can predict what type of crime has likely occurred based on location and date.*

Overview of the Data

```
glimpse(nyc)
```

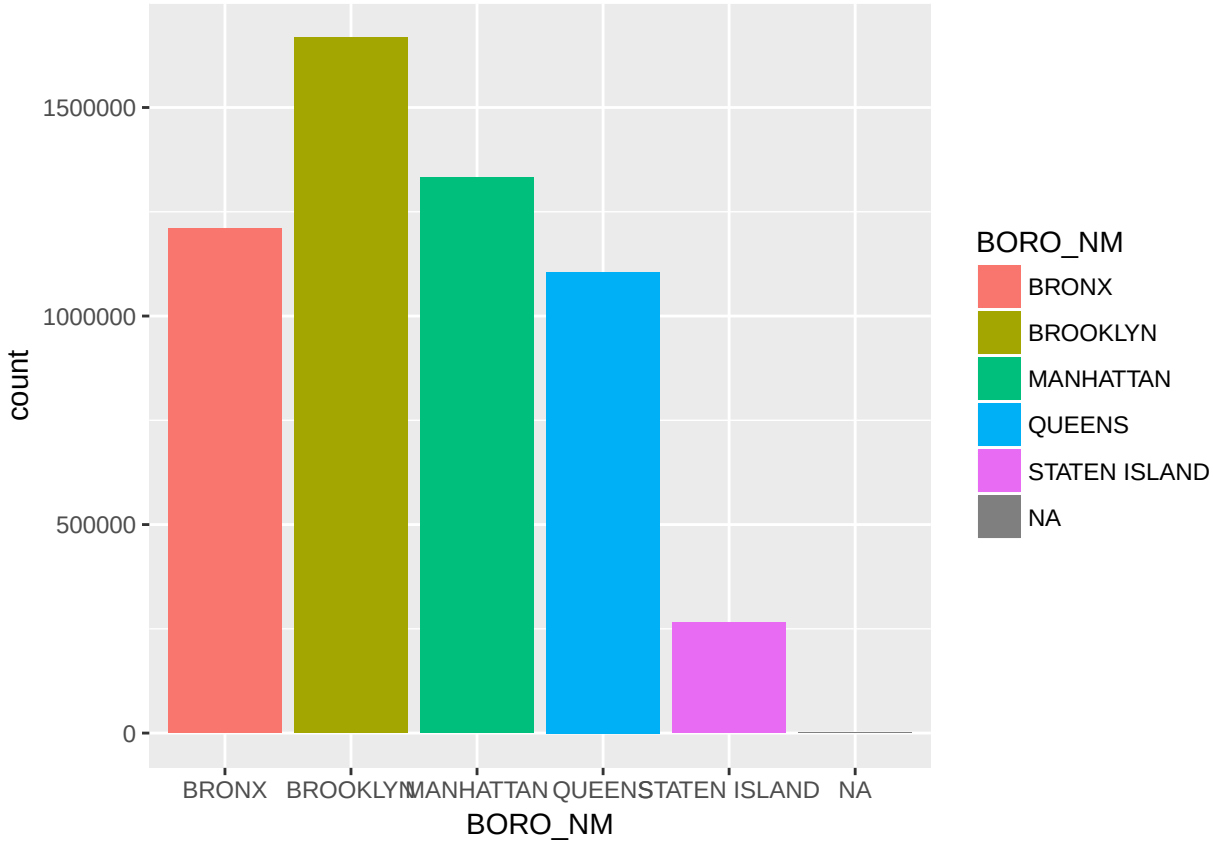
```
## Observations: 5,580,035
## Variables: 24
## $ CMPLNT_NUM        <int> 101109527, 153401121, 569369778, 968417082, ...
## $ CMPLNT_FR_DT      <chr> "12/31/2015", "12/31/2015", "12/31/2015", "1...
## $ CMPLNT_FR_TM      <time> 23:45:00, 23:36:00, 23:30:00, 23:30:00, 23:...
## $ CMPLNT_TO_DT      <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...
```

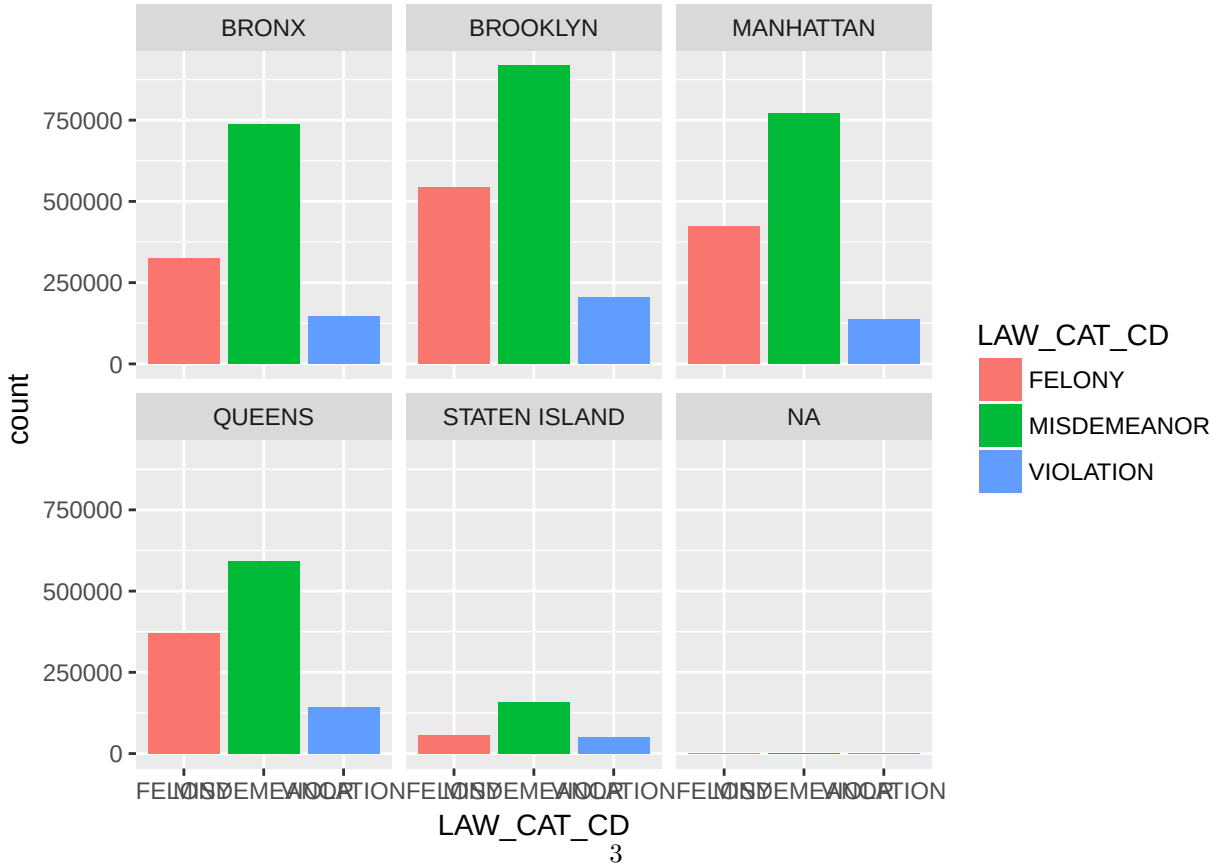
```
summary(nyc)
```

```
##    CMPLNT_NUM        CMPLNT_FR_DT       CMPLNT_FR_TM
##  Min.   :100000228   Length:5580035     Length:5580035
##  1st Qu.:324985515   Class :character   Class1:hms
##  Median :549944466   Mode  :character   Class2:difftime
##  Mean   :549961482                      Mode  :numeric
##  3rd Qu.:774926532
##  Max.   :999999904
##
```

```
##  CMPLNT_TO_DT       CMPLNT_TO_TM         RPT_DT              KY_CD
##  Length:5580035     Length:5580035    Length:5580035     Min.   :101.0
##  Class :character   Class1:hms        Class :character   1st Qu.:117.0
##  Mode  :character   Class2:difftime   Mode  :character   Median :341.0
##                     Mode  :numeric                       Mean   :293.7
##                                                          3rd Qu.:351.0
##                                                          Max.   :881.0
##
##   OFNS_DESC             PD_CD         PD_DESC          CRM_ATPT_CPTD_CD
##  Length:5580035     Min.   :101.0   Length:5580035     Length:5580035
##  Class :character   1st Qu.:254.0   Class :character   Class :character
##  Mode  :character   Median :393.0   Mode  :character   Mode  :character
##                     Mean   :417.5
##                     3rd Qu.:637.0
##                     Max.   :975.0
##                     NA's   :4909
##   LAW_CAT_CD         JURIS_DESC          BORO_NM           ADDR_PCT_CD
##  Length:5580035     Length:5580035     Length:5580035     Min.   :  1.00
##  Class :character   Class :character   Class :character   1st Qu.: 40.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 63.00
##                                                           Mean   : 63.43
##                                                           3rd Qu.: 94.00
##                                                           Max.   :123.00
##                                                           NA's   :390
##  LOC_OF_OCCUR_DESC  PREM_TYP_DESC        PARKS_NM
##  Length:5580035     Length:5580035     Length:5580035
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HADEVELOPT           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:5580035     Min.   : 913319   Min.   :120829   Min.   :40.50
##  Class :character   1st Qu.: 991704   1st Qu.:184162   1st Qu.:40.67
##  Mode  :character   Median :1004292   Median :205586   Median :40.73
##                     Mean   :1004648   Mean   :206913   Mean   :40.73
##                     3rd Qu.:1016282   3rd Qu.:235177   3rd Qu.:40.81
##                     Max.   :1067298   Max.   :271820   Max.   :40.91
##                     NA's   :195868    NA's   :195868   NA's   :195868
##    Longitude         Lat_Lon
##  Min.   :-74.26   Length:5580035
##  1st Qu.:-73.97   Class :character
##  Median :-73.93   Mode  :character
##  Mean   :-73.93
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :195868
```

Some Preliminary Plots





Points to think about

Variable Exclusion

- Due to the lack of consistent data in these fields or the seeming irrelevance to my goal, I want to nix:
  - CMPLNT_TO_DT
  - CMPLNT_TO_TM
  - CRM_ATPT_CPTD_CD
  - JURIS_DESC
  - ADDR_PCT_CD
  - PARKS_NM
  - HADEVELOPT

Variable Simplification

The following variables all give some idea of the type of crime committed:

```
head(nyc)[names(nyc)[c(7,8,10,12)]]
```

```
## # A tibble: 6 x 4
##   KY_CD                        OFNS_DESC                          PD_DESC
##   <int>                            <chr>                            <chr>
## 1   113                          FORGERY FORGERY,ETC.,UNCLASSIFIED-FELO
## 2   101 MURDER & NON-NEGL. MANSLAUGHTER                              <NA>
## 3   117                 DANGEROUS DRUGS CONTROLLED SUBSTANCE,INTENT TO
## 4   344    ASSAULT 3 & RELATED OFFENSES                       ASSAULT 3
## 5   344    ASSAULT 3 & RELATED OFFENSES                       ASSAULT 3
## 6   106                  FELONY ASSAULT       ASSAULT 2,1,UNCLASSIFIED
## # ... with 1 more variables: LAW_CAT_CD <chr>
```

What is the best way of handling multiple variables that basically present the same information but in differing levels of classification?

Likewsie, the variables X_COORD_CD, Y_COORD_CD, Latitude, Longitude, and Lat_Lon all give me numerical values of location. I want to just use Latitude and Longitude, and maybe create a heat map of NYC based on crime type and frequency.

Priors

- For example, I would expect more crimes downtown to be white-collar crimes, while crimes in the surrounding neighborhoods might be more violent.
- How do I make this into a prior?
- Should I find 2000-2005 and use that data as an 'existing body of literature'?