Project: Capstone Project 1: Data Wrangling

Daniel Loew

This report summarizes what was done in the Data Cleaning Jupyter notebooks, that can be found at these two links:

Data Cleaning for Cannabis Crimes notebook:

<https://github.com/danloew/Springboard/blob/master/1st_Capstone_Data_Cleaning_cannabis_Daniel_Loew_Final.ipynb>

Data Cleaning for Non-Cannabis Crimes notebook:

<https://github.com/danloew/Springboard/blob/master/1stCapstone_Data_Cleaning_non-cannabis_Daniel_Loew_Final.ipynb>

The following data cleaning and data wrangling steps were performed on the NYPD data set “NYPD Complaint Data Historic” that records data on all crimes committed in New York City between 2006 and the first part of 2018:

1. Dropped two columns that had multiple data types and were either of unlikely relevance to the predictive model that is central to this project, were unlabeled and therefore meaningless, or had too many values for predictive analysis. These two variables were ‘PARKS\_NM’ and ‘HOUSING\_PSA’, which showed which city park the crime reportedly occurred in, and a list of unlabeled numerical codes having to do with housing projects, respectively. ‘PARKS\_NM’ had 1,130 values, far too many for the predictive model. There are also other geographic variables to use, and only 9.5% of cannabis crimes were recorded as occurring in city parks. The data dictionary had no information on the meaning of ‘HOUSING\_PSA’.
2. A new dataset was subsetted from the larger dataset to only include cannabis crimes. This only includes crimes with penal codes (PD\_CD) 566-570. These codes are 566) Marijuana Possession, 567) Marijuana Possession 4 & 5, 568) Marijuana Possession 1, 2, & 3, 569) Marijuana Sale 4 & 5, and 570) Marijuana Sale 1, 2, & 3.
3. Binary features (i.e., columns) were created for cannabis possession (PD\_CDs 566-568), cannabis sales (PD\_CDs 569 and 570), misdemeanors (LAW\_CAT\_CD = misdemeanor), violations (LAW\_CAT\_CD = violation), and felonies (LAW\_CAT\_CD = felony).
4. The features in step 3 were then used to create features for the following crime levels: misdemeanor cannabis possession, violation cannabis possession, felony cannabis possession, misdemeanor cannabis sales, violation cannabis sales, and felony cannabis sales. There were no violation sales cases, but the other five features will serve as the target variables for the machine learning classification of cannabis crime levels that will be carried out later in the capstone project pipeline.
5. The chained .isnull().sum() and .isna().sum() methods were then used to show how many missing values there were in each of the features in the feature set. A set of features had their missing values filled in with an ‘unknown’ value or some other similar feature-specific value (like “not transit-related” for the TRANSIT\_DISTRICT feature). This set included the following features (missing value n reported in parentheses): HADEVELOPT (193,639), CMPLNT\_TO\_DT (67,454), CMPLNT\_TO\_TM (67,381), TRANSIT\_DISTRICT (217,273), STATION\_NAME (217,273), BORO\_NM (185), LOC\_OF\_OCCUR\_DESC (92,077), PREM\_TYP\_DESC (1,731), SUSP\_AGE\_GROUP (186,008), SUSP\_RACE (185,550), SUSP\_SEX (185,587), PATROL\_BORO (1), VIC\_AGE\_GROUP (188,373), VIC\_RACE (54), and VIC\_SEX (54). The date variable had missing values coded as ‘00/00/0000’, and the time variable had missing values coded as ‘00:00:00’.
6. Another set of features had missing values dropped, as imputation of these values would be specious and biasing to the results, and the number of missing values were rather small overall. This set included the following features (missing value n reported in parentheses): CMPLNT\_FR\_DT (2), CMPLNT\_FR\_TM (1), X\_COORD\_CD (472), Y\_COORD\_CD (472), Latitude (472), Longitude (472), and Lat\_Lon (472).
7. Datetime crime start, crime end, and crime duration variables were created from the crime start date, crime start time, crime end date, and crime end time variables that came with the dataset.
8. The raw crime start time variable was used to create a set of time-window features that may be predictive of cannabis crimes. The start time variable 'CMPLNT\_FR\_TM' was used, as all crimes have a start time in the data set but not necessarily an end time. It also makes simple logical sense to timestamp a crime at the time that it starts. The derived time window features include daytime, night time, early morning, morning rush hour, the traditional work day, the lunch hour, evening rush hour, dinner hour, evening, and late night.
9. The distance of each cannabis crime from prominent NYC landmarks was encoded into continuous data features. All latitudes and longitudes were found from Google searches. Both driving/walking/biking distances ("\_taxi" variables) and shortest distances ("\_crow" variables) were computed. The landmarks included the World Trade Center, the New York Stock Exchange, Brooklyn Bridge, New York City Hall, Manhattan Bridge, Williamsburg Bridge, Washington Square Park, Union Square, Penn Station, Times Square, Rockefeller Center, Empire State Building, Lincoln Center, Central Park, Apollo Theatre, Yankee Stadium, Mets Stadium, the center of Queens Borough, the center of Prospect Park, the center of downtown Brooklyn, Staten Island Ferry Terminal, Port Authority Bus Station, New York Police Department headquarters, Manhattan Detention Center, Rikers Island, and the New York Supreme Court. With this step, each cannabis crimes’ distance from key geographic landmarks in New York City is known. This information will be used in the predictive analysis to follow.
10. Isolated year, month, date, hour and minute features were extracted from the crime start datetime variable. Along with being useful data to have on their own, the extracted date variables were used to define holidays that fall on the same day every year. These included New Year’s Eve, New Year’s Day, Christmas Eve, Christmas Day, July 4th, Valentine’s Day, Halloween, and St. Patrick’s Day. Time variables were kept for machine learning classification purposes.
11. Boolean masks and variable assignment were used to create new binary features for major holidays that do not fall on the same day every year. The first step created separate boolean masks for the date of the holiday in question for each year, then a boolean mask was created for all years based on the boolean masks for each individual year, and then the holiday's feature was assigned. To keep the dataset tidy, the variables for each holiday’s year were then deleted, as they were no longer needed. These holidays include Martin Luther King Day, President’s Day, Easter, Diwali, the Puerto Rican Day Parade, Yom Kippur, Rosh Hashanah, Eid al-Fitr, Eid al-Adha, Hannukkah, Memorial Day, Labor Day, and Thanksgiving.
12. Cases outside of the stated year range of the dataset were dropped, that is, cases earlier than 2006.
13. Unclear values were recoded to ‘unknown’ for the suspect and victim age group, race, and sex variables. First, the value counts were called for these features for reference, a cleaned version of the feature was created by mapping unusual values to ‘unknown’, and then value counts were called on the cleaned version of the feature to ensure all unclear values had been mapped to ‘unknown’. Unclean versions of these features were dropped from the dataset. ‘JURISDICTION\_CODE’ was also dropped, as it had numerical codes that were meaningless as they weren't labeled, and there is a descriptive variable for information on which jurisdiction the crime fell under.
14. A separate DataFrame was subsetted from the overall DataFrame of cannabis crimes, with just those cannabis crimes that had suspect race reported. This is for later machine learning classification of cannabis crime types for just the cases whose suspect race was reported.
15. A .csv file was then exported for both the overall DataFrame of cannabis crimes and the smaller DataFrame of cannabis crimes that had suspect race reported. These DataFrames keep their categorical features intact, to be used for visual and statistical EDA in the next report. This was done before the .get\_dummies() method described in the next step that makes the DataFrames machine learning friendly.
16. The remaining categorical variables were transformed into individual binary features using .get\_dummies(), so that each categorical value of each of these features ends up having its own feature. This was done for both the larger DataFrame of all cannabis crimes, and for the smaller subsetted DataFrame with cannabis crimes whose suspect race was reported. This is done for machine learning classification purposes later in the capstone project pipeline. The features that were transformed included: ADDR\_PCT\_CD, SUSP\_AGE\_GROUP\_cleaned, SUSP\_RACE\_cleaned, SUSP\_SEX\_cleaned, VIC\_AGE\_GROUP\_cleaned, VIC\_RACE\_cleaned, VIC\_SEX\_cleaned, BORO\_NM, HADEVELOPT, JURIS\_DESC, LOC\_OF\_OCCUR\_DESC, PATROL\_BORO, PREM\_TYP\_DESC, TRANSIT\_DISTRICT, and STATION\_NAME.
17. A list of variables was then dropped from both DataFrames as they were superfluous or were in a data type which would not be readable by machine learning methods (like ‘string’ and ‘datetime’). Some of the variables also duplicate the information stored in the target variables and will distort the fit and predict classifier functions. The variables that were dropped were: possession, sales, misdemeanor, violation, felony, viol\_sales, KY\_CD, X\_COORD\_CD, Y\_COORD\_CD, Lat\_Lon, LAW\_CAT\_CD, PD\_DESC, OFNS\_DESC, CMPLNT\_FR\_DT, CMPLNT\_TO\_DT, RPT\_DT, CMPLNT\_FR\_TM, CMPLNT\_TO\_TM, date\_time\_start, date\_time\_end, duration, CRM\_ATPT\_CPTD\_CD, and CMPLNT\_NUM. PD\_CD was kept to serve as the target variable for the classification of different crime types amongst the cannabis crimes dataset.
18. The cleaned DataFrames were then exported to a .csv file for easy loading for the upcoming machine learning classification.
19. A separate DataFrame of cannabis crimes was created for a planned second round of machine learning classification of cannabis crimes vs non-cannabis crimes. The non-cannabis crime DataFrame was created in a separate Jupyter notebook that followed the exact same cleaning protocol as the cannabis crime DataFrame, but was randomly sampled. PD code and law category codes were compared between the overall population of non-cannabis crimes and the sample, to ensure that there was not an oversampling of any specific crime type, so as not to bias the results later.
20. Because this second round of classification won't need to differentiate between types of cannabis crime, and for easy merging with the cleaned non-cannabis crime DataFrame, the cannabis crime type target variables were dropped from the cannabis crime DataFrame. A new label variable was added to the cannabis crime and non-cannabis crime DataFrames for classification and supervised learning purposes.
21. Versions of cannabis crime and non-cannabis crime DataFrames were also created for just those cases where the suspect’s race was reported, as that will be an important part of the analysis.
22. Csv exports were done for these last four DataFrames.

There were outliers in the years column. I omitted them completely as they were obviously errors, as the stated year range for the dataset on NYC Open Data was 2006-2018. After a careful review of the value counts of all other features, no other outliers were found.

I transformed a 35 feature DataFrame into much higher-dimensional DataFrames that will be capable of robust machine learning that will hopefully provide a highly predictive model that can identify the most salient predictive factors for cannabis arrests in New York City according to the data the NYPD has collected on the subject. I’m hoping that this work will help flesh out research done on long-standing racial bias in cannabis arrests in the city to help illustrate a larger context for the reasons for these arrests.