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/SpringboardFirstCapstone/blob/master/Data_Cleaning_cann_Final.ipynb

Data Cleaning for Non-Cannabis Crimes notebook:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/DataCleaning_ncann_Final.ipynb

The following data cleaning and wrangling steps were performed on the "NYPD Complaint Data Historic" dataset:

1. The dataset was loaded in full into a Pandas DataFrame.
2. The 'PARKS_NM' and 'HOUSING_PSA' variables were found to be of mixed data type. The 'HOUSING_PSA' variable was dropped from the DataFrame because it contained numeric codes for public housing developments whose character names were located elsewhere in the 'HADEVELOPMENT' variable, and there was no key in the data dictionary which provided correspondence between the numeric codes and character names of the housing developments. The 'JURISDICTION_CODE' feature was also dropped, as there was already a character data type feature called 'JURIS_DESC'. The 'PARKS_NM' variable which contained the names of NYC parks where crimes were committed was kept, but was coerced to the 'string' data type.
3. To ensure that there were no duplicate rows, which can be a source of machine learning model leakage, the .drop_duplicates() method was used. No duplicate rows were found.
4. A new Pandas DataFrame was subsetted from the larger DataFrame to only include cannabis crimes. This only included crimes with penal codes (PD_CD) 566-570. These codes are listed in the NYPD's data dictionary as:
    a. 566 - Marijuana, Possession (violation level)
    b. 567 - Marijuana, Possession, 4th & 5th degree (misdemeanor level)
    c. 568 - Marijuana, Possession, 1st, 2nd, and 3rd degree (felony level)
    d. 569 - Marijuana, Sales 4th & 5th degree (misdemeanor level)
    e. 570 - Marijuana, Sales, 1st, 2nd, and 3rd degree (felony level)
5. Binary features 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).
6. The binary features created in step 5 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 cannabis sales cases as that is not a valid legal category of crime, but the other five features are the target features for the machine learning classification of the five cannabis crime levels.

7. 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 another similar feature-specific value (like "not transit-related" for the TRANSIT_DISTRICT feature). The date variable had missing values coded as '00/00/0000', and the time variable had missing values coded as '00:00:00'. This set included the following features (missing value n reported in parentheses):

    a. HADEVELOPT (193,639),
    b. CMPLNT_TO_DT (67,454),
    c. CMPLNT_TO_TM (67,381),
    d. TRANSIT_DISTRICT (217,273),
    e. STATION_NAME (217,273),
    f. BORO_NM (185),
    g. PARKS_NM (218,332),
    h. LOC_OF_OCCUR_DESC (92,077),
    i. PREM_TYP_DESC (1,731),
    j. SUSP_AGE_GROUP (186,008),
    k. SUSP_RACE (185,550),
    l. SUSP_SEX (185,587),
    m. PATROL_BORO (1),
    n. VIC_AGE_GROUP (188,373),
    o. VIC_RACE (54),
    p. and VIC_SEX (54)
8. 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. These dropped cases amounted to 475 out of 220,817 rows (or 0.2%), which was an acceptable loss for a uniform DataFrame with no missing values across all features. This set included the following features (missing value n reported in parentheses):
    a. CMPLNT_FR_DT (2),
    b. CMPLNT_FR_TM (1),
    c. X_COORD_CD (472),
    d. Y_COORD_CD (472),
    e. Latitude (472),
    f. Longitude (472), and
    g. Lat_Lon (472)
9. Datetime crime start, crime end, and crime duration features were created from the crime start date ('CMPLNT_FR_DT'), crime start time ('CMPLNT_FR_TM'), crime end date ('CMPLNT_TO_DT'), and crime end time ('CMPLNT_TO_TM') features native to the NYPD's dataset. These new features were named 'date_time_start', 'date_time_end', and 'duration'. 'Duration' was created to measure the amount of time elapsed between the start of the crime complaints and their end.
10. The newly created 'duration' feature was checked for null values, which it had 67,212 of as this number of cases only had a 'date_time_start' value. Null values were edited to the specific value of zero days, hours, minutes, and seconds.
11. The new 'duration' feature was of TimeDelta data type, which is unable to be used by the Logistic Regression classifiers. Therefore, it was converted into the integer data type feature

'duration_days', in order to identify just how many days a case transpired over. The original TimeDelta ‘duration’ feature was then dropped from the DataFrame.

12. Discrete year, month, date, hour, minute and second features were extracted from the ‘date_time_start’ feature as integer data type features for use in the classification models. Unusual values were checked for in these features, and none were found except for several cases that did not occur within the prescribed year range of 2006 to 2018. Along with being useful features to have on their own, the extracted date and time features were used later in the data cleaning pipeline to create other features.
13. The cases with year values outside of the range of 2006 to 2018 were identified and dropped from the DataFrame.
14. Because the 'duration_days' feature uses 'date_time_end' in its valuation of the number of days a crime takes to occur, extracting datetime features from 'date_time_end' is not needed. In other words, the 'duration_days' feature stores the information of the date and time that the crime ends. Therefore, the 'date_time_end' feature was not used in the feature set for classification models, and was dropped from the DataFrame. 'CMPLNT_TO_DT', 'CMPLNT_TO_TM', and 'end_year' were also no longer needed, and were dropped.
15. The 'RPT_DT', or date that the crime was reported by the NYPD, differed from the complaint date stored in the 'CMPLNT_FR_DT' feature 7.2% of the time. This difference was postulated as a possible predictive feature. Before creating a predictive feature based on the difference between the two dates, ‘RPT_DT’ was first converted into a feature of DateTime data type and then year, month, and date features were extracted from it so that any outliers outside of the prescribed date range of 2006 to 2018 could be identified and dropped. There were none such cases. The DateTime feature and extracted year, month, and date features were no longer needed and were dropped. A binary feature was then created to store the information of whether the complaint date matches the police-reported date. Because the original 'RPT_DT' feature contains the same information as 'CMPLNT_FR_DT' 92.8% of the time, the 'RPT_DT' feature itself will be dropped later in the data cleaning pipeline so as to avoid duplication of information that could theoretically adversely affect the classification models.
16. The native crime start time variable ‘CMPLNT_FR_TM’ was used to create a set of time-window features that may be predictive of cannabis crimes. The derived time window features included daytime, night time, early morning, morning rush hour, the traditional work day, the lunch hour, evening rush hour, dinner hour, evening, and late night.
17. Boolean masks and feature assignment were used to create new binary features for major holidays that always fall on the same day of the year. These holidays could be predictors of cannabis crimes because different holidays have special meaning to demographic groups that may be differentially targeted for cannabis arrests, and more generally holidays are drivers of behavior in the United States. These major holidays are New Year's Day, Valentine's Day, St. Patrick's Day, July 4th, Halloween, Christmas Eve, Christmas, and New Year's Eve. April 20th is included as it is an emerging day of cannabis celebrations and could be relevant. Intriguingly, this day has more cannabis arrests than any holiday. The creation of these binary fixed holiday features was done by first creating a DataFrame with the month and day of each fixed holiday. Then, the contents of the DataFrame were used by 'for' loops to create individual features storing whether a crime occurred on that particular fixed holiday or not. Fixed holiday features were then cast into integer type.

18. Pd.DataFrame and 'for' loops were used to create new binary features for crimes that occur on major holidays that do not fall on the same day every year, i.e., "floating holidays". This is done by first creating a DataFrame with the month, day, and year of each floating holiday, for each of the years between 2006 and 2018. Then, the contents of the DataFrame were used with a combination of 'for' loops, boolean assignment across the years, and integer data type casting to create individual features storing whether a crime occurred on that particular floating holiday (1) or not (0). These floating holidays included Martin Luther King Day, President's Day, Easter, Diwali, the Puerto Rican Day Parade, Yom Kippur, Rosh Hashanah, Eid al-Fitr, Eid al-Adha, Hanukkah, Memorial Day, Labor Day, and Thanksgiving.
19. Outlier and erroneous values of several features native to the NYPD's data set were looked for and dropped if necessary. All values of these features were verified with online sources specified in the Jupyter notebooks. These features were the:
    a. police precinct ('ADDR_PCT_CD'),
    b. NYC borough ('BORO_NM'),
    c. location of crime occurrence in relation to the premises ('LOC_OF_OCCUR_DESC'),
    d. premises type description of crime occurrence ('PREM_TYP_DESC'),
    e. jurisdiction of crimes ('JURIS_DESC'),
    f. NYC parks crimes occurred in ('PARKS_NM'),
    g. housing developments crimes occurred in ('HADEVELOPT')
    h. transit districts crimes occurred in ('TRANSIT_DISTRICT'),
    i. latitude and longitude crimes occurred in ('LATITUDE', 'LONGITUDE'),
    j. NYPD patrol borough crimes occurred in ('PATROL_BORO'), and
    k. MTA transit stations ('STATION_NAME')
20. The only features that had outliers were latitude and longitude. There was one case that occurred outside of New York City, which was discovered through a scatterplot of the two features. This case was dropped from the DataFrame.
21. Subway station entrances can be places where people sell, purchase, consume, and get cannabis delivered in New York City, and the Open NY project has latitudes and longitudes of all NYC subway stations and their entrances (available at https://data.ny.gov/widgets/i9wp-a4ja) Features were created that determined both the L1 (Euclidean or "taxi") and L2 (straight-line or "as the crow flies") distances in units of latitude/longitude to the closest subway station for each cannabis crime.
22. The distance of each cannabis crime from prominent NYC landmarks was encoded into continuous data features. All latitudes and longitudes for these landmarks were found via Google search. Both L1 (Euclidean or "taxi") and L2 (straight-line or "as the crow flies") distances in units of latitude/longitude 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, the Staten Island Ferry Terminal, the Port Authority Bus Station, the New York Police Department headquarters, Metropolitan Detention Center in Manhattan, Riker's Island, and the New York Supreme Court. With this step, each cannabis crimes' distance from key geographic landmarks in New York City is known.

23. Unclear values were recoded to 'unknown' for the suspect and victim age group, race, and sex features. First, the value counts were run for each feature. Then, a cleaned version of the features were created by mapping unusual values to the 'unknown' value. Approximately 84% of the cases had an ‘unknown’ value for these six demographic features.
24. For ease of use in doing visual and statistical EDA, two versions of the DataFrame were then exported before the binarization of categorical data features (which is implemented later in the data cleaning pipeline for use in the machine learning models). The first version of the DataFrame contained all cannabis crime cases, while the second version only contained cannabis crime cases where the suspect’s race was recorded by the arresting officer (approximately 16% of the overall cases).
25. The hypothesis testing done in the Statistical Data Analysis notebook required a sample of all NYC crimes. After the data cleaning process described above, a 10% sample was taken and exported to a .csv file. The same process was done in the non-cannabis crime data cleaning notebook and the two sample DataFrames were concatenated in the Statistical Data Analysis notebook (see below).
26. The remaining categorical features were transformed into individual binary features via the Pandas .get_dummies() method in a separate machine learning version of the DataFrame, so that each value of categorical features has its own binary feature. This is done for later machine learning classification of different cannabis crime types within the universe of NYC cannabis crimes between 2006 and 2018. Several features were identified as being superfluous or as causing leakage in the machine learning pipeline, as they contain the same information as the target feature being predicted. Meanwhile, a set of valuable features were binarized and kept in the machine learning version of the DataFrame. A breakdown of which features were dropped and which were binarized follows:

    a. Because the five target features for the classification model of the five types of cannabis crime were already instantiated, some of the features that were used to create the target features were no longer needed. These included 'misdemeanor', 'violation', and 'felony'. 'Possession' was kept as the target feature for classification models differentiating cannabis possession crimes from cannabis sales crimes, which will be explored with a second round of classification. However, 'sales' was not needed, so it was dropped as well. So the 'misdemeanor', 'violation', 'felony', and 'sales' features were dropped, as NOT doing so would introduce leakage to the classification models.
    b. Although violation sales ('viol_sales') started as a target feature, no cases were designated as violation sales as there is no legal category of violation cannabis sales in New York City. So it was also dropped from the feature set.
    c. 'CMPLNT_NUM' was kept at this juncture as it was needed to label all cases as cannabis crimes for the classification model of cannabis vs. non-cannabis crimes (see below). Also, dropping it before concatenating the cannabis crimes with non-cannabis crimes for this model was shown to cause false positive duplicates in earlier versions of the notebook.
    d. The 'PD_DESC', 'PD_CD', and 'LAW_CAT_CD' features were superfluous as they contained the same information as the target features for each model, namely what kind of crime each case is. So to avoid leakage, 'PD_DESC', 'PD_CD', and 'LAW_CAT_CD' were dropped.

e. 'OFNS_DESC' was also deemed to be superfluous, as it is another way of describing type of crime. Using 'OFNS_DESC' in the classification models will introduce leakage, as it contains the information that will be predicted by them. Also, all cases were either labeled as 'DANGEROUS DRUGS' or 'MISCELLANEOUS PENAL LAW'.
f. 'KY_CD' was also dropped, as it duplicates the information stored in the target features and will introduce leakage to the classification models. 'KY_CD' is the numeric version of the string feature 'LAW_CAT_CD', as described in the NYPD's data dictionary.
g. 'Lat_Lon' was dropped, as separate features for Latitude and Longitude were already in the DataFrame.
h. 'X_COORD_CD' and 'Y_COORD_CD' were superfluous as they contained the same geo-coordinate information as latitude and longitude under a different data convention, so they were dropped as well.
i. The 'CMPLNT_FR_DT', 'CMPLNT_FR_TM', and 'date_time_start' features were all dropped as they contained the same date and time information contained in the 'start_year', 'start_month', 'start_day', 'start_hour', 'start_minute', and 'start_seconds' features. In the case of 'date_time_start', it was a datetime formatted feature unable to be processed by the machine learning classifiers.
j. The 'rpt_cmplnt_dt_match' feature was kept, but its parent feature 'RPT_DT' was dropped so as not to duplicate 92.8% of the information found in the 'start_year', 'start_month', and 'start_day' features.
k. 'CRM_ATPT_CPTD_CD', the feature which stores information on whether a crime was completed or attempted, essentially contains the same information that exists in the target feature, i.e. a crime was completed. So it was dropped.
l. Victim info would not necessarily be available at the time of trying to predict whether a new crime is a cannabis crime or a non-cannabis crime, because not every cannabis crime has a victim and therefore victim age, race, and sex. Therefore, 'VIC_AGE_GROUP_cleaned', 'VIC_RACE_cleaned', and 'VIC_SEX_cleaned' was dropped.

27. A separate DataFrame (‘nyc_cann_ml_alt’) was created for machine learning classification of cannabis crimes and non-cannabis crimes through making a copy of the 'nyc_cann_ml' DataFrame. Because this round of classification won't need to differentiate between types of cannabis crime, and for easy merging with the cleaned non-cannabis crime DataFrame created in a separate notebook (see below), the cannabis crime type target features, along with 'possession', were dropped.
28. For the machine learning classification model differentiating cannabis crimes from non-cannabis crimes, a label feature was needed that pre-labels cannabis crimes and non-cannabis crimes. This feature called 'cannabis_crime' was created by assigning all rows in the ‘nyc_cann_ml_alt’ DataFrame with a ‘cannabis_crime’ value of 1. As the 'nyc_cann_ml_alt' DataFrame only contains cannabis crimes, and all cases have a positive integer complaint number ('CMPLNT_NUM' variable), assignment of '1' for the feature 'cannabis_crime' was implemented with a simple condition of the ‘CMPLTN_NUM’ being positive.
29. The cleaned machine learning DataFrame of just the universe of cannabis crimes (‘nyc_cann_ml’), and the cleaned machine learning DataFrame of the universe of cannabis crimes with the ‘cannabis_crime’ flag feature (‘nyc_cann_ml_alt’), were exported to separate

.csv files. Both of these files had 220,304 cases. The ‘nyc_cann_ml’ DataFrame was used for the classification models differentiating cannabis possession crimes from cannabis sales crimes and the five models differentiating each of the five cannabis crime types from each other. The ‘nyc_cann_ml_alt’ DataFrame was used for the classification model differentiating cannabis crime from all other crimes, after concatenation with a DataFrame of non-cannabis crimes.

The non-cannabis crime DataFrame was created in a separate Jupyter notebook. This Jupyter notebook followed the exact same cleaning protocol as the cannabis crime DataFrame detailed above, except for the following differences:

1. A DataFrame was subsetted from the original NYPD dataset which only included non-cannabis crimes, i.e. crimes that did not have a police crime code (‘PD_CD’) of 566-570.
2. The ‘date_time_end’ feature derived from the ‘CMPLNT_TO_DT’ and ‘CMPLNT_TO_TM’ made clear that there were three cases that needed to be dropped because they had erroneous values later than the circumscribed year range of 2006-2018. These three cases were dropped.
3. For non-cannabis crimes, there were 129 geographic outlier cases that occurred outside of the latitude/longitude range of the five boroughs of New York City. The frame of this study is crimes that occurred inside New York City between 2006-2018, so these outliers had to be dealt with. For situations where a significant proportion of cases are geographic outliers, their latitude/longitude values would be trimmed to the maximum or minimum point of the range, whichever is nearest. However, because of the vanishingly small proportion of outliers (129 out of 6,238,620 cases), removing these cases would not bias the results and they were dropped from the DataFrame.
4. A random sample of cases of equal size to the cleaned cannabis crime universe (n=220,304) was taken from the non-cannabis DataFrame. PD code and law category codes were compared between the cleaned universe of non-cannabis crimes and its random sample, to ensure that there was not an oversampling of any specific crime type that could bias the results later. There was no such oversampling.
5. The distance to the closest subway entrance feature was derived after the random sample was taken because of a shortage of computational resources. In other words, the analyst’s computer was not up to the task of computing the values of this feature for approximately 6.2 million rows.
6. An EDA version of the sample DataFrame with the categorical features intact was exported to a csv file for use in the Data Story & Exploratory Data Analysis section below.
7. A value of ‘0’ was assigned to the derived ‘cannabis_crime’ feature, so that all non-cannabis crimes in the sample were flagged as being non-cannabis crimes.
8. A .csv file named ‘nyc_non_cann_for_ML’ was exported and then concatenated with the ‘nyc_cann_ml’ DataFrame using the pandas.concat() method. The presence of duplicate rows was checked for with the pandas.drop_duplicates() method, and there were none.
9. The presence of null values was checked with the chained Pandas methods of .isnull().sum() and .isna().sum(). The features with null values were put into a list called 'fill_w_zero', and then any null values for these features were filled with zeroes. This was needed as there were many housing developments, transit stations, and city parks that exist in the cannabis crime DataFrame but not the non-cannabis DataFrame, or vice versa.

10. The 'CMPLNT_NUM' feature was dropped from the concatenated DataFrame of cannabis and non-cannabis crimes because it was not needed any longer, and would likely introduce leakage to the classification model. The leakage would occur as each record had a unique value for 'CMPLNT_NUM', which would create absolute overfitting if 'CMPLNT_NUM' was used as a predictive feature.
11. The concatenated DataFrame was finally exported to a .csv file for use in the machine learning classification model of differentiating cannabis crimes from non-cannabis crimes.