Przetwarzanie i analiza dużych zbiorów danych2021/2022

Prowadzący: mgr inż. Rafał Woźniak Czwartek, 15:45

Szymon Gruda	239661	239661@edu.p.lodz.pl
Jan Karwowski	239671	239671@edu.p.lodz.pl
Michał Kidawa	239673	239673@edu.p.lodz.pl
Kamil Kowalewski	239676	239676@edu.p.lodz.pl

Checkpoint 1

Spis treści

1. Wprowadzenie

Głównym celem projektu jest przeprowadzenie kompleksowej analizy zbioru *NYPD Complaint Data Historic* [1]. Jego dokładny opis został umieszczony w sekcji 3 natomiast jego cele oraz pytania na które będziemy chcieli znaleźć odpowiedź w trakcie wykonywania analizy zostały przedstawione w sekcji 4.

2. Podział obowiązków w zespole

— Szymon Gruda - Przygotowanie statystycznego opisu zbioru danych dla danych określonych jako lokalizacja zdarzenia, cechy podejrzanego i ofiary oraz zależności między podejrzanym a ofiarą. Dodatkowo udział w grupowych rozważaniach.
— Jan Karwowski - sformułowanie w przystępny sposób opisu celów przygotowanych przez zespół. Dodatkowo udział w grupowych rozważaniach.
— Michał Kidawa - Przygotowanie opisu kolumn oraz zebranie w formie tekstowej informacji o wstępnym przetwarzaniu danych. Dodatkowo udział w grupowych rozważaniach.
— Kamil Kowalewski - Przygotowanie ogólnego opisu danych zbioru oraz statystycznego opisu zbioru danych dla danych określonych jako data i czas zgłoszenia, typ i opis wykroczenia, status przestępstwa (czy się udało) oraz otoczenie zdarzenia. Dodatkowo udział w grupowych rozważaniach.

3. Statystyczny opis zbioru danych

3.1. Opis kolumn

Zbiór danych zawiera kolumny których nazwy oraz opisy w pogrupowany sposób zostały przedstawione poniżej:

1. Identyfikator
   — CMPLNT_NUM - Losowo generowany trwały identyfikator dla każdego zgłoszenia
2. Data i czas zdarzenia
   — CMPLNT_FR_DT - Dokładna data wystąpienia zgłoszonego zdarzenia (lub data początkowa wystąpienia, jeżeli CMPLNT_TO_DT istnieje)
   — CMPLNT_FR_TM - Dokładny czas wystąpienia zgłoszonego zdarzenia (lub czas rozpoczęcia wystąpienia, jeżeli CMPLNT_TO_TM istnieje)
   — CMPLNT_TO_DT - Data końcowa wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany
   — CMPLNT_TO_TM - Końcowy czas wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany
3. Data i czas zgłoszenia

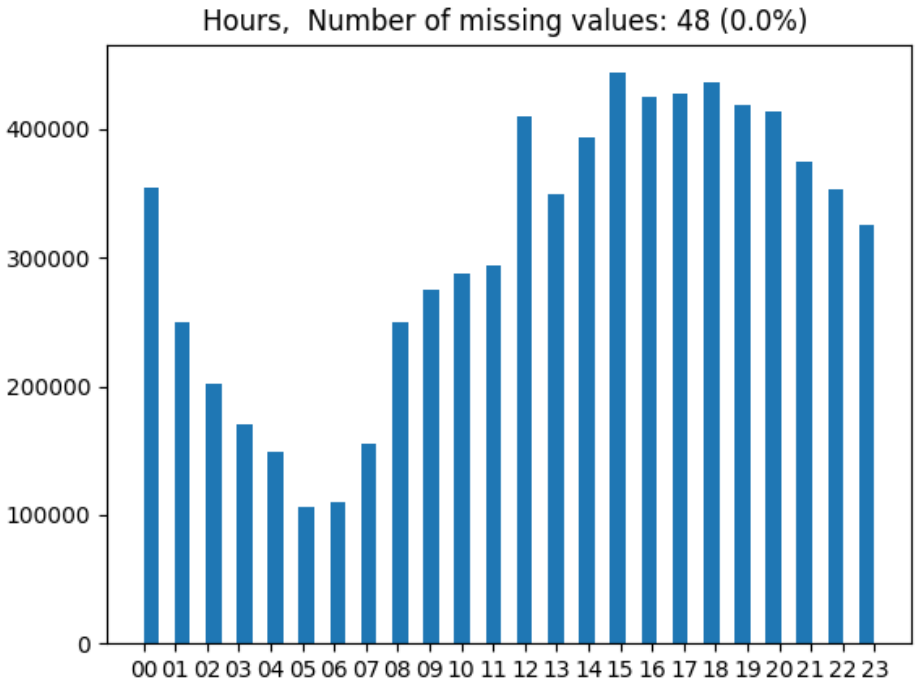
   — RPT_DT - Data zgłoszenia zdarzenia na policję
4. Typ i opis wykroczenia/przestępstwa
   — KY_CD - Trzycyfrowy kod klasyfikacji wykroczeń
   — OFNS_DESC - Opis wykroczenia odpowiadający kodowi klucza
   — PD_CD - Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code)
   — PD_DESC - Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa)
   — LAW_CAT_CD - Poziom wykroczenia: przestępstwo, wykroczenie, naruszenie
5. Czy się udało
   — CRM_ATPT_CPTD_CD - Status, czy przestępstwo zostało dokonane czy była próba jego popełnienia lub czy zostało ono przerwane
6. Otoczenie zdarzenia
   — PREM_TYP_DESC - Rodzaj otoczenia; sklep spożywczy, domek jednorodzinny, ulica itp.
   — LOC_OF_OCCUR_DESC - Lokalizacja w stosunku do otoczenia; wewnątrz, naprzeciw, z przodu, z tyłu
7. Lokalizacja zdarzenia
   — ADDR_PCT_CD - Posterunek
   — BORO_NM - Dzielnica
   — JURIS_DESC - Opis kodu jurysdykcji
   — JURISDICTION_CODE - Kod jurysdykcji na której miało miejsce to zdarzenie
   — PARKS_NM - Nazwa parku, placu zabaw lub terenów zielonych w Nowym Jorku, jeśli dotyczy (parki stanowe nie są uwzględnione)
   — HADEVELOPT - Nazwa osiedla NYCHA miejsca zdarzenia, jeśli dotyczy
   — HOUSING_PSA - Kod poziomu rozwoju
   — X_COORD_CD - Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
   — Y_COORD_CD - Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
   — TRANSIT_DISTRICT - Okręg tranzytowy, w którym doszło do wykroczenia.
   — Latitude - Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
   — Longitude - Współrzędna długości bloku środkowego dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
   — Lat_Lon - Punkt lokalizacji geoprzestrzennej (łącznie szerokość i długość geograficzna)
   — PATROL_BORO - Nazwa dzielnicy patrolowej, w której doszło do incydentu
   — STATION_NAME - Nazwa stacji tranzytowej
8. Cechy podejrzanego

— SUSP_AGE_GROUP - Grupa wiekowa podejrzanego
— SUSP_RACE - Rasa podejrzanego
— SUSP_SEX - Płeć podejrzanego

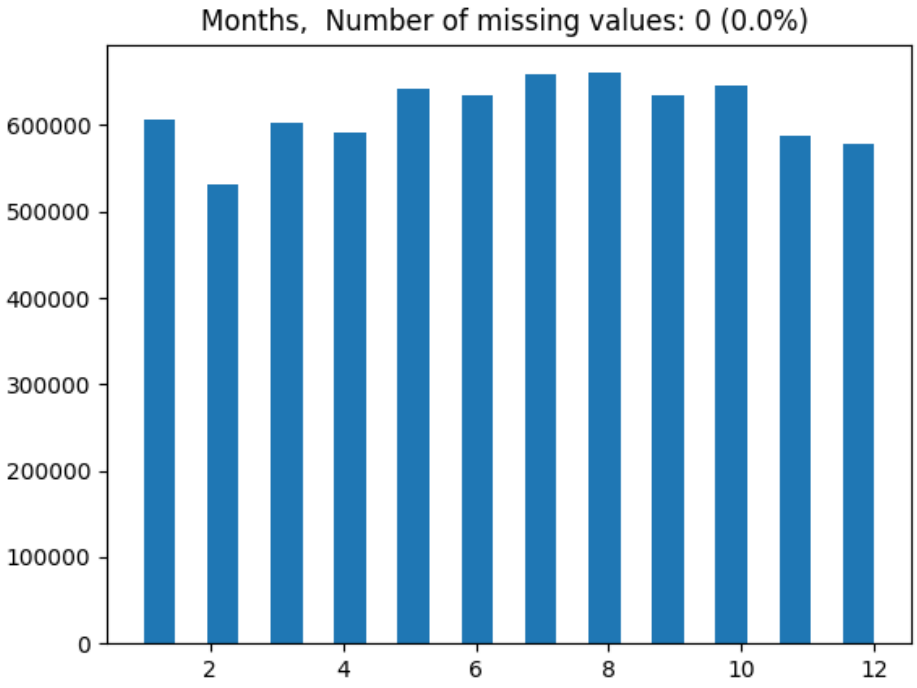
9. Cechy ofiary
— VIC_AGE_GROUP - Grupa wiekowa ofiary
— VIC_RACE - Rasa ofiary
— VIC_SEX - Płeć ofiary

3.2. Wykresy

3.2.1. Data i czas zgłoszenia



Rysunek 1. Histogram liczby przestępstw w zależności od godziny



Rysunek 2. Histogram liczby przestępstw w zależności od miesiąca

Na podstawie rysunku 1 można zaobserwować, że największa liczba przestępstw/wykroczeń miała miejsce w godzinach popołudniowych, natomiast najmniejsza w godzinach wczesnoporonnych tzn godzina 5 i 6 rano.

Na podstawie rysunku 2 można zaobserwować, że wyniki liczby przestępstw/wykroczeń są do siebie zbliżone natomiast większa liczba była w czasie miesięcy letnich.

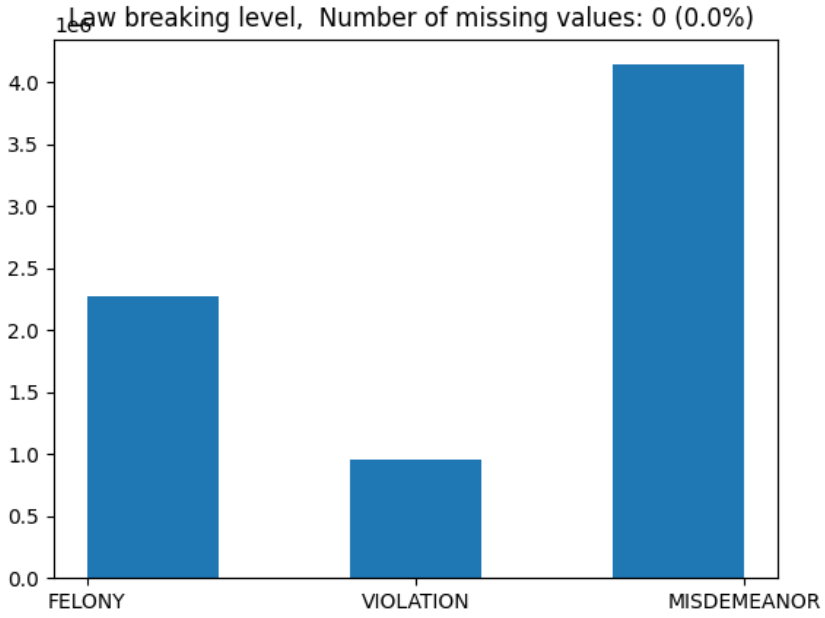
Dla rekordów w, których występuję wartość *CMPLNT_TO_DT* oraz *CMPLNT_TO_TM* została obliczony średni czas między początkiem a końcem zdarzenia i są to następujące wartości

— średnia: 8 days 21:03:19.644759679
— odchylenie standardowe: 190 days 21:44:14.282341806

Dla rekordów została policzony średni czas oraz odchylenie standardowe w dniach między wystąpieniem zdarzenia a jego zgłoszeniem na policje i są to następujące wartości:

— średnia: 14 days 19:01:46.072720350
— odchylenie standardowe: 233 days 01:34:24.475636108

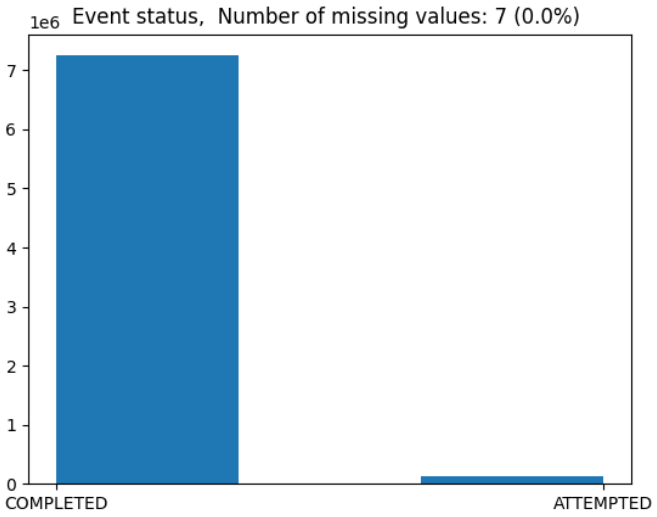
3.2.2. Typ i opis wykroczenia/przestępstwa



Rysunek 3. Histogram liczby przestępstw/wykroczeń zależnie od jego typu

Na podstawie rysunku 3 można zaobserwować, że największy odsetek stanowiły wykroczenia, na drugim miejscu uplasowały się przestępstwa natomiast najmniej było występków(jest to pośredni czy miedzy wykroczeniem a przestępstwem).

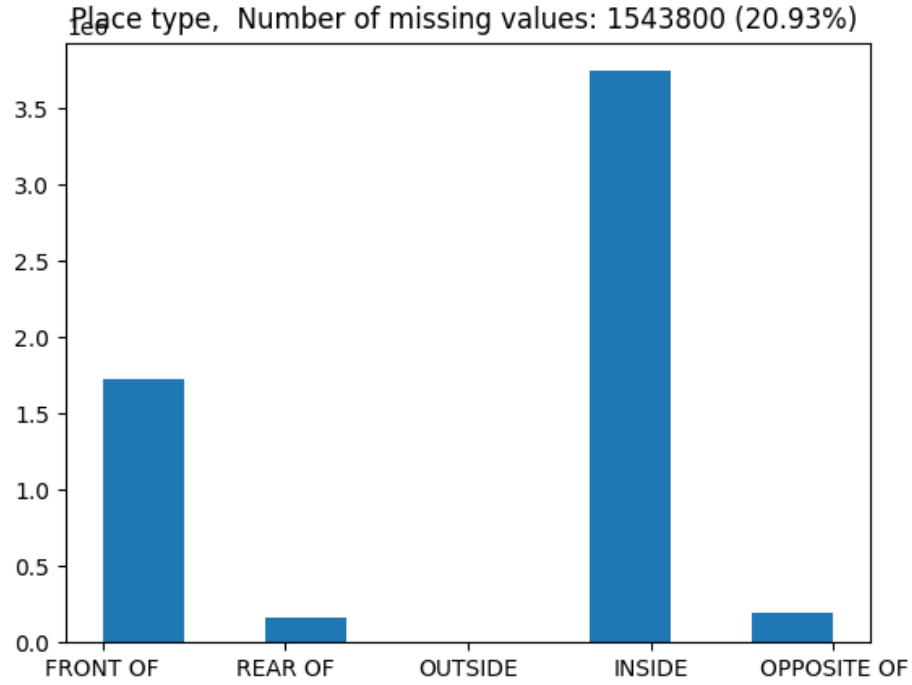
3.2.3. Czy się udało



Rysunek 4. Histogram liczby przestępstw/wykroczeń w zależności od tego czy doszło ono do skutku czy zostało zatrzymane

Na podstawie rysunku 4 można zaobserwować, że praktycznie wszystkie przestępstwa/wykroczenia zostały dokonane a naprawdę niewielki odsetek został udaremniony.

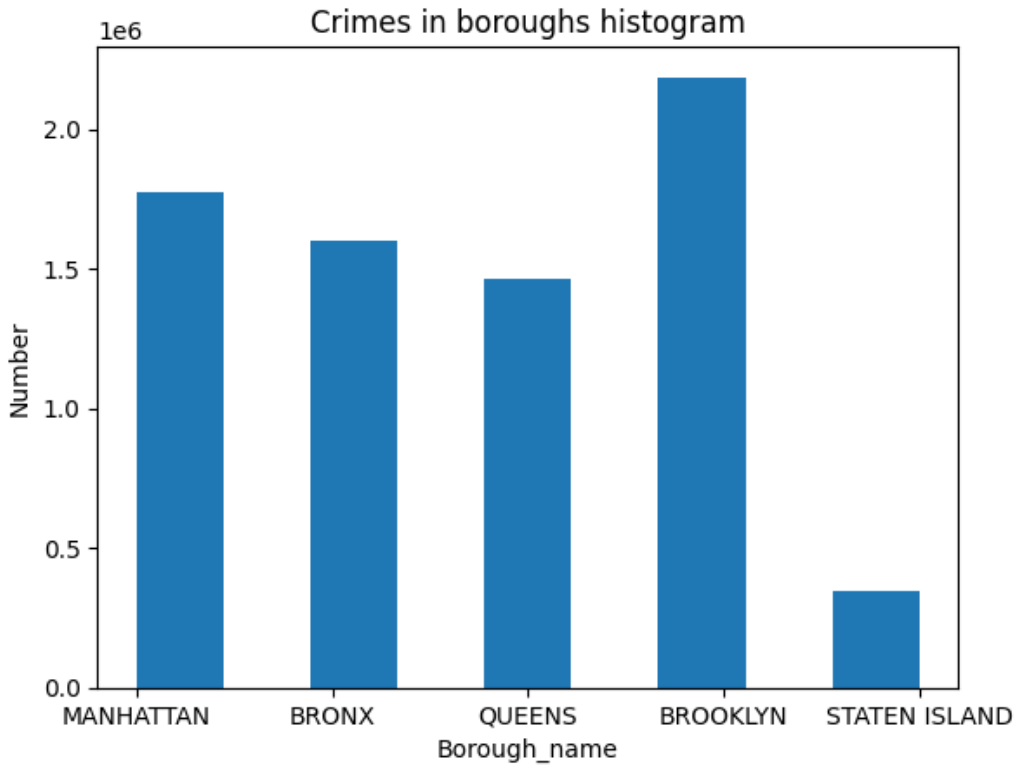
3.2.4. Otoczenie zdarzenia



Rysunek 5. Histogram liczby przestępstw/wykroczeń w zależności od typu miejsca w jakim się zdarzyło

Na podstawie rysunku 5 można zaobserwować, że wg raportów policji znakomita większość miała miejsce w środku danego miejsca, drugim najczęściej wskazanym otoczeniem było miejsce przed np. danym budynkiem.

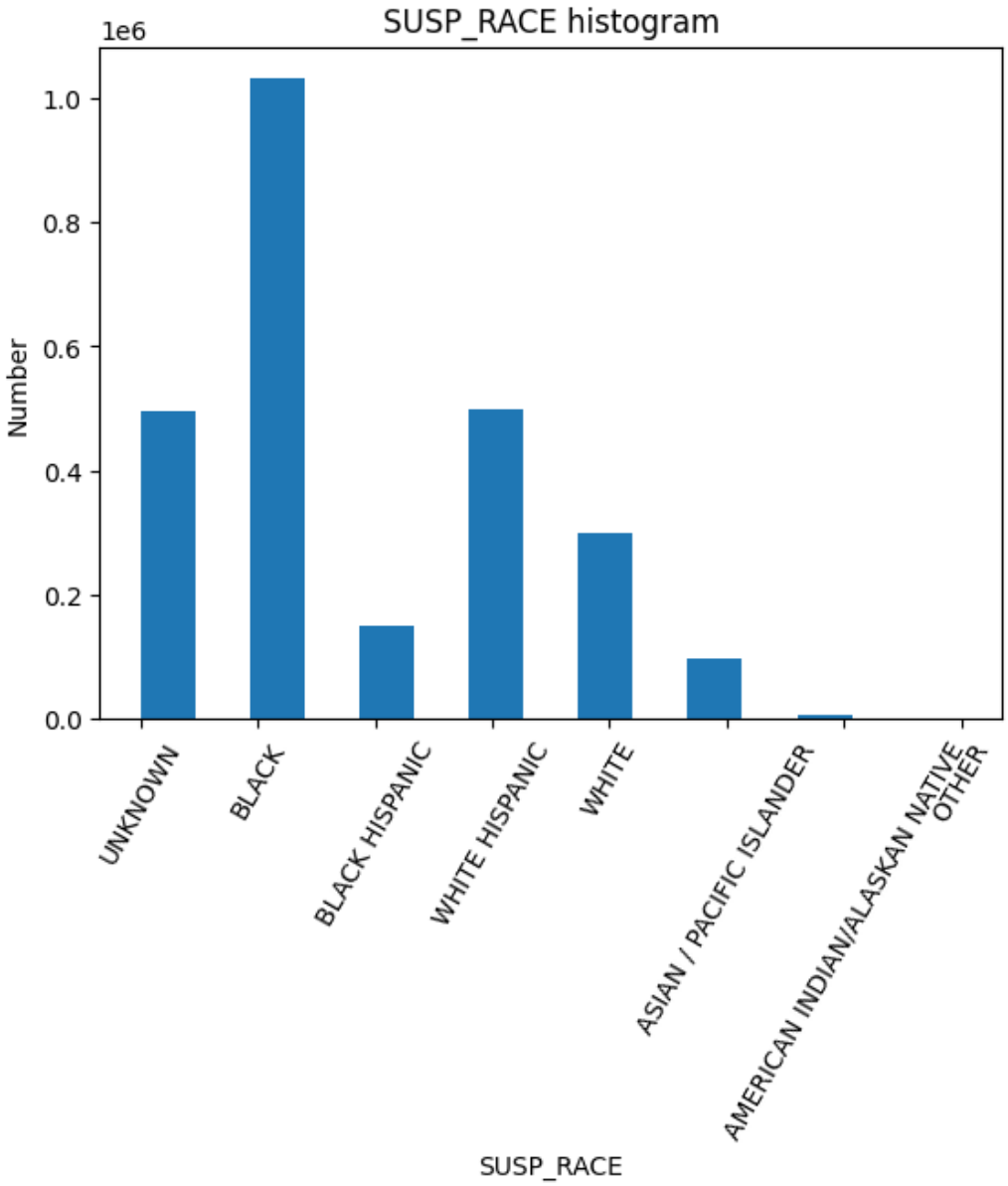
3.2.5. Lokalizacja zdarzenia



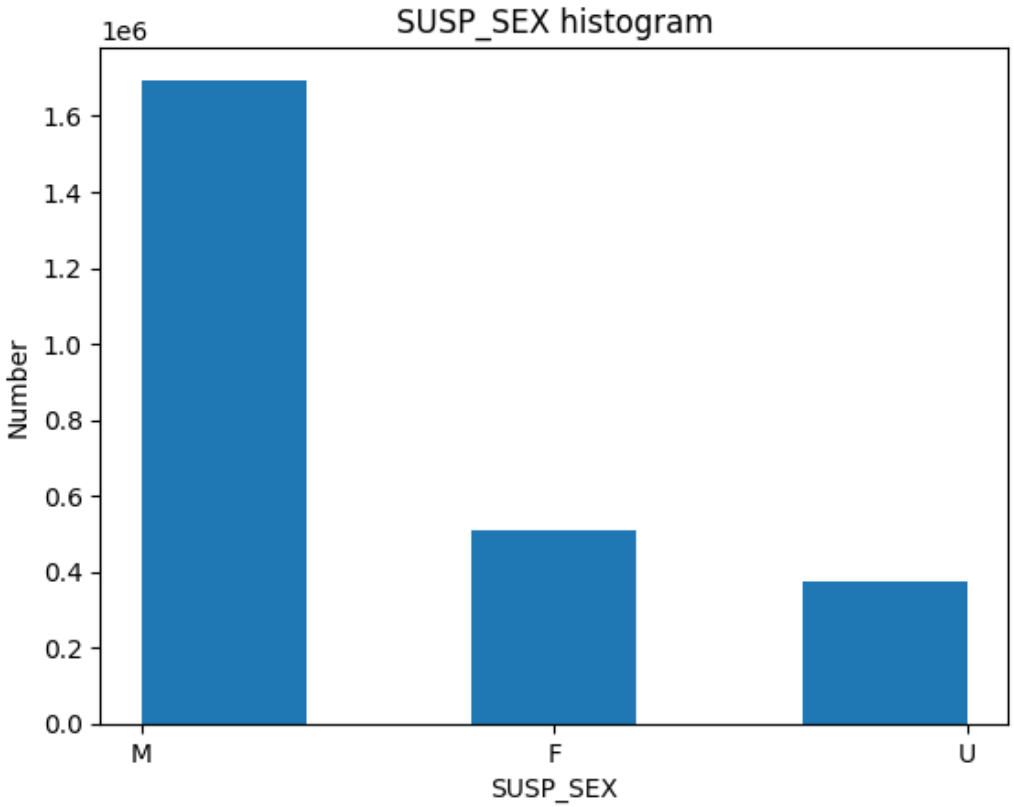
Rysunek 6. Histogram liczby przestępstw w zależności od dzielnicy

Na podstawie rysunku 6. można zobaczyć, że w dzielnicy Brooklyn dochodzi do największej liczby przestępstw, w Staten Island najmniej, natomiast pozostałe dzielnice posiadają podobną liczbę popełnionych przestępstw. Przyczyną odbiegających od nich wyników dla Brooklyn'u i Staten Island może być, np. ich powierzchnia.

3.2.6. Cechy podejrzanego



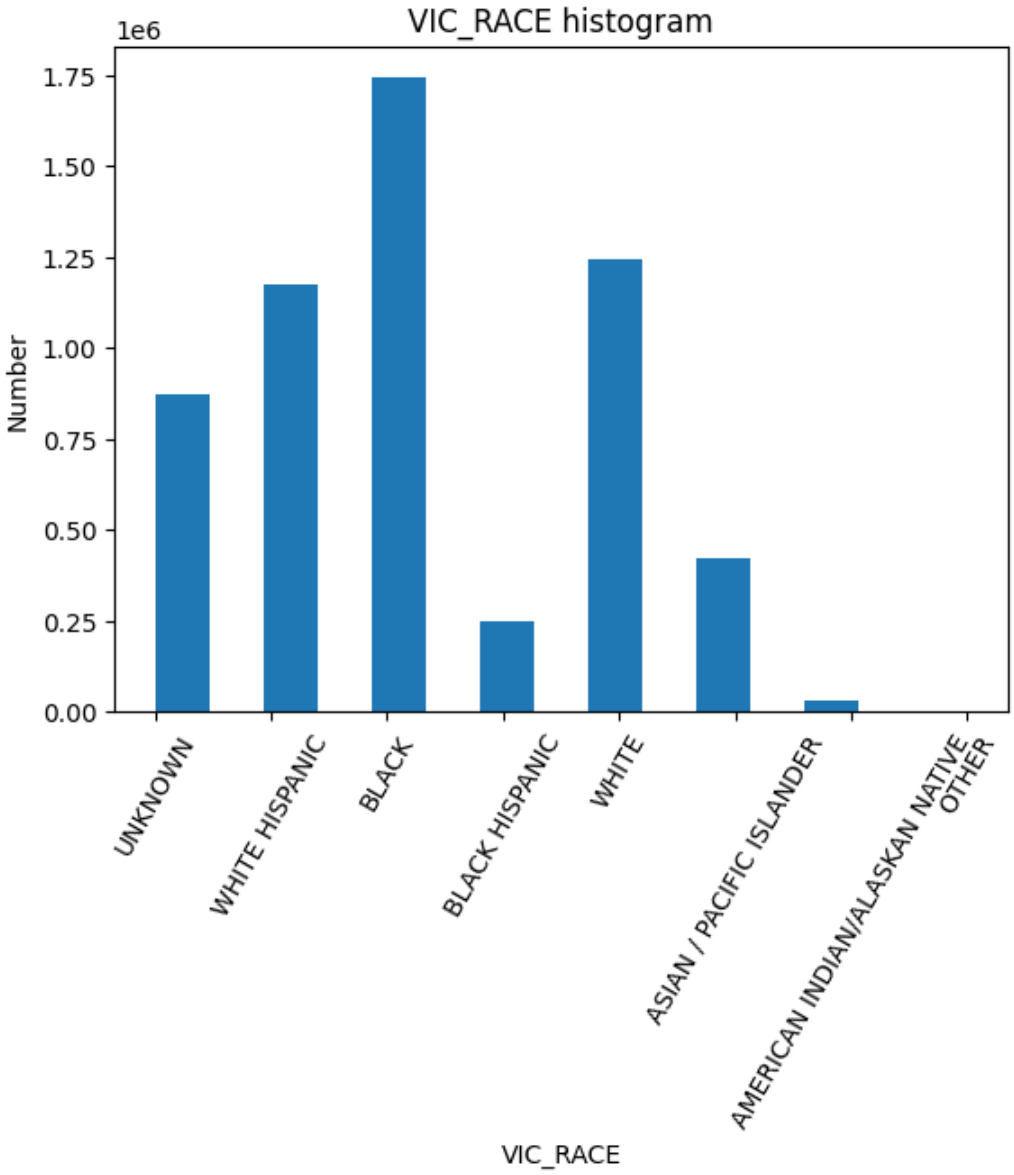
Rysunek 7. Histogram liczby przestępstw popełnianych przez określoną rasę.



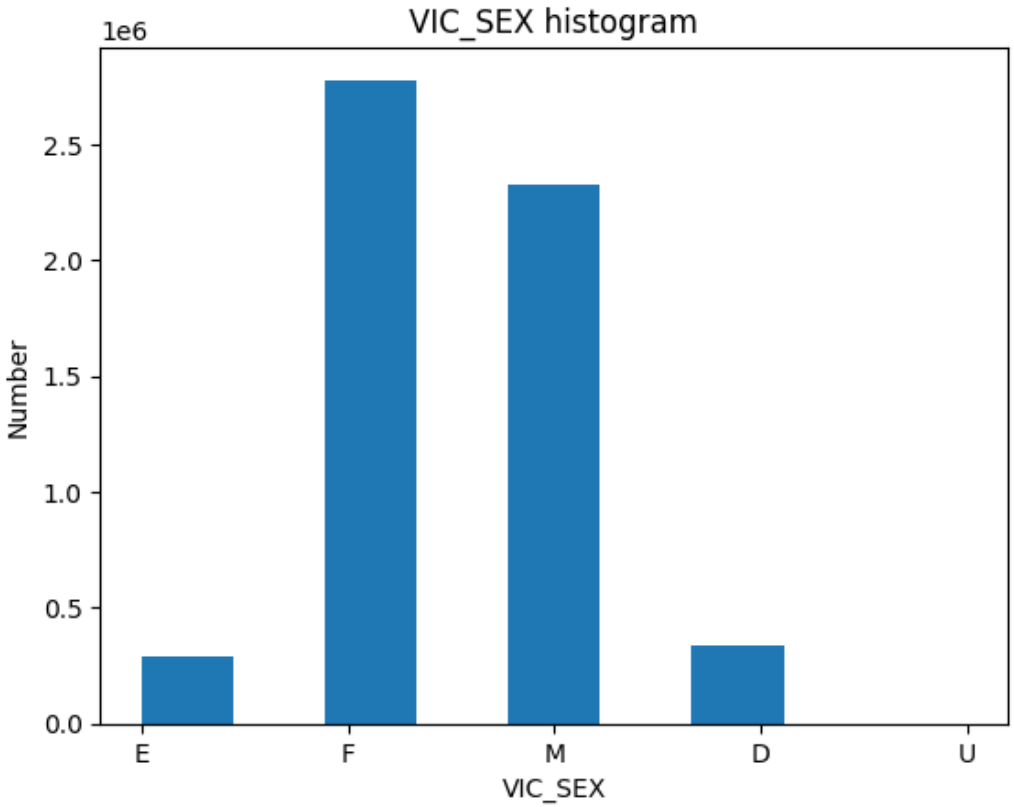
Rysunek 8. Histogram liczby przestępstw popełnianych przez określoną płeć.

Na podstawie rysunków 7. i 8 można zauważyć, że najwięcej podejrzanych o przestępstwo cechuje się czarnym kolorem skóry, lub płcią męską. Dysproporcje pomiędzy rasą, mogą być spowodowane np. nierównym procentowym udziałem tych ras w społeczeństwie.

3.2.7. Cechy ofiary



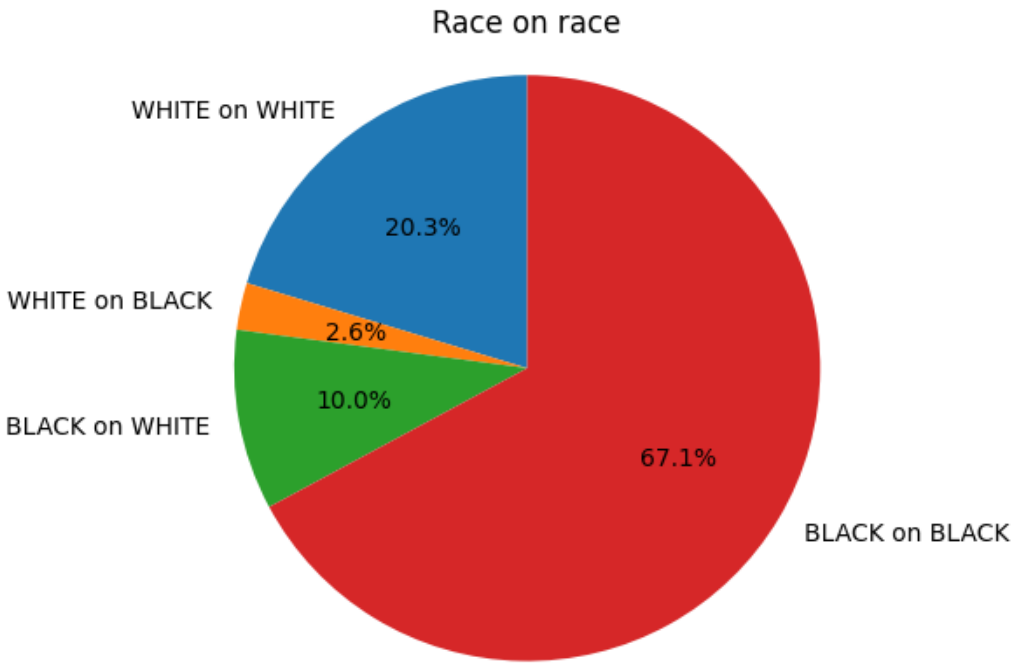
Rysunek 9. Histogram liczby przestępstw, których ofiarą jest określona rasa.



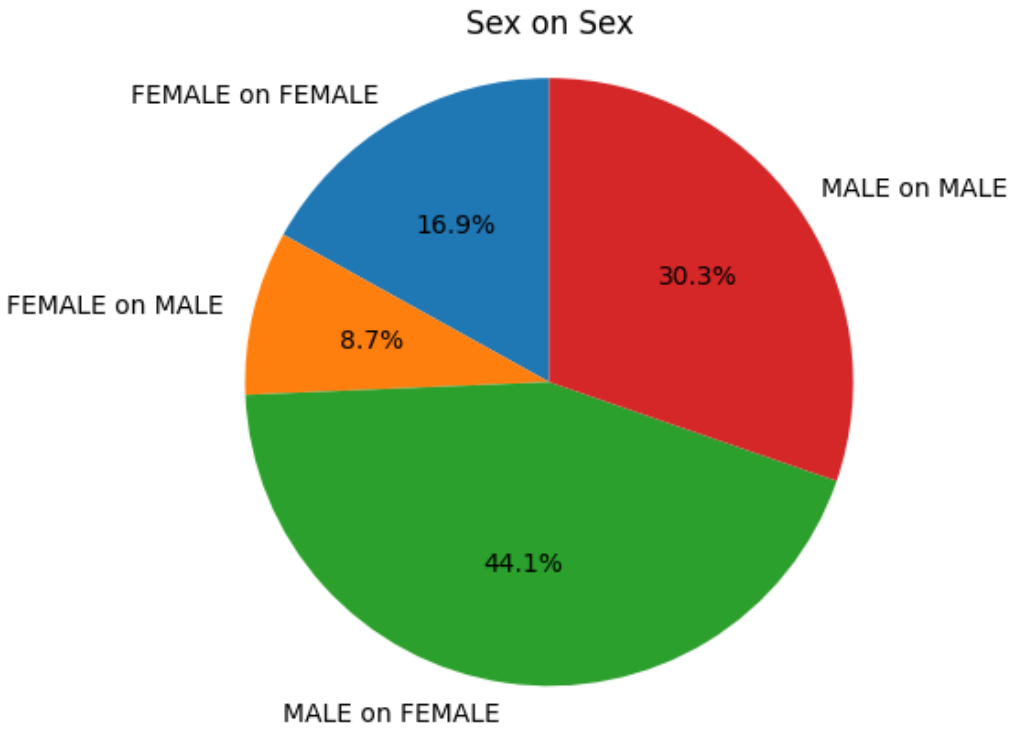
Rysunek 10. Histogram liczby przestępstw, których ofiarą jest określona płeć.

Na podstawie rysunków 9. i 10 można zauważyć, że najwięcej ofiar przestępstw cechuje się czarnym kolorem skóry, lub płcią żeńską (chociaż nie ma dużej różnicy pomiędzy nią, a płcią męską). Dysproporcje pomiędzy rasą, mogą być spowodowane np. nierównym procentowym udziałem tych ras w społeczeństwie. Albo cechami społeczności, niektóre osoby z pewnego powodu, mogą nie czuć obowiązku zgłaszania przestępstwa.

3.2.8. Zależności między podejrzanym a ofiarą



Rysunek 11. Wykres przedstawiający zależność pomiędzy podejrzanym a ofiarą, na podstawie ich rasy (uwzględnione zostały tylko dwie).



Rysunek 12. Wykres przedstawiający zależność pomiędzy podejrzanym a ofiarą, na podstawie ich płci (uwzględnione zostały tylko dwie).

Do wykresów 11 i 12 zostały wybrane dwie rasy i płcie, które największą liczbę razy uczestniczyły w przestępstwie jako ofiara lub podejrzany. W ten sposób się zaobserwować, że największy odsetek przestępstw dotyczył tej samej rasy, można zatem stwierdzić, że odsetek przestępstw między rasowych jest niższy niż można było się spodziewać. Natomiast sprawdzenie rozkładu płci, nie zaskutkowało niespodziewanymi wnioskami. Mężczyźni ogólnie częściej są podejrzanymi o przestępstwa i częściej popełniają takie, w których ofiarami są kobiety niż mężczyźni.

4. Cele projektu

W ramach projektu sformułowane zostały trzy następujące cele.

4.1. Klasyfikacja rodzaju lub poziomu przestępstwa

W ramach tego etapu przeprowadzony zostanie szereg eksperymentów mających na celu stworzenie klasyfikatora typu przestępstwa (KY_CD). Jeżeli okaże się to niemożliwe podjęta zostanie próba klasyfikacji poziomu wykroczenia, stanowiącego bardziej ogólną informacje. Klasyfikacja ta będzie odbywała się na podstawie następujących informacji:
— godzina zdarzenia
— dzień tygodnia zdarzenia
— odstęp między zgłoszeniem a zdarzeniem
— czy doszło do skutku (CRM_ATPT_CPTD_CD)
— otoczenie zdarzenia
— cechy podejrzanego
— cechy ofiary
— poziom lub typ przestępstwa w zależności od tego co klasyfikujemy

Wybrane cechy mogą ulec drobnym modyfikacjom w trakcie trwania eksperymentów. Opcjonalnie podjęta będzie próba wykorzystania informacji o czasie trwania przestępstwa (dla tych zdarzeń, dla których jest dostępna).

Informacja o dokładnej lokalizacji zdarzenia nie jest wykorzystana ze względu na chęć stworzenia uniwersalnego narzędzia.

Aby zrealizować zaproponowany cel wykorzystane zostaną następujące metody: metody imputacji brakujących danych, prosta ekstrakcja cech (zwłaszcza z daty), naiwny klasyfikator Bayesa i lasy losowe.

4.2. Regresja czasu trwania przestępstwa

Drugim celem będzie regresja czasu trwania przestępstwa dla tych przestępstw, dla których jest on znany. Tak więc wybrany zostanie podzbiór głównego zbioru danych i na podstawie czasu rozpoczęcia i zakończenia przestępstwa wyekstrahowany zostanie czas jego trwania. Na podstawie pozostałych kolumn (poza tych z datą i czasem) podjęta będzie próba wyuczenia modelu regresyjnego estymowania tego czasu.

Wykorzystane zostaną modele przystosowane do zadań regresji, tak więc przede wszystkim lasy losowe, ewentualnie klasyfikator liniowy, maszyny wektorów nośnych czy wreszcie sieć neuronowa. Wykorzystane metody będą

dobierane odpowiednio w zależności od wyników otrzymywanych w trakcie trwania eksperymentów i od napotkanych problemów. Dodatkowo planowane jest wykorzystanie biblioteki *XGBoost*.

4.3. Grupowanie przestępstw na podstawie podzbiorów cech

W ramach tego celu przeprowadzone zostanie automatyczne grupowanie zdarzeń. Spośród cech podejrzanego i ofiary kilkakrotnie wybrany zostanie podzbiorów stanowiących przedmiot eksperymentu - etykietę. Dla każdej wybranej etykiety (która może być złożeniem kilku atrybutów) przeprowadzona zostanie seria eksperymentów, mająca na celu automatyczne pogrupowanie zdarzeń zgodnie z tą etykietą, wykorzystując wszystkie pozostałe atrybuty (poza tworzącymi etykietę). Po przeprowadzeniu grupowania z wykorzystaniem kilku różnych algorytmów (DBSCAN, k-means, algorytm aglomeracyjny), zmierzona zostanie jakość grupowania za pomocą metryk zewnętrznych (accuracy) względem wybranej etykiety. Dodatkowo porównana zostanie jakość grupowania między seriami eksperymentów (dla różnych etykiet) za pomocą metryk wewnętrznych.

5. Wstępne przetwarzanie danych

Do wstępnego przetwarzania danych został wykorzystany język Python oraz biblioteka Pandas. Za ich pomocą zostało wygenerowane podsumowanie zbioru poprzez wywołanie funkcji *describe()*, pomogła ona w znalezieniu zależności oraz wybraniu ciekawych danych do prezentacji na wykresach i histogramach. W przypadku przegotowywania danych do histogramów gdy występowała pusta wartość była ona pomijana.

Literatura

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i