Hurtownie danych – Projekt HD

PWr. Wydział Informatyki i Telekomunikacji Data: 13.06.2022

Student	----------	Ocena
Indeks	252463	
Imię	Daniil	
Nazwisko	Hardzetski	

1. Tytuł projektu

Działania policji w Nowym Yorku w latach 2006-2020.

2. Charakterystyka dziedziny problemowej

2.1 Opis obszaru analizy (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

Do analizy zostały wzięte dane o skargach, aresztowaniach i strzelaninach w Nowym Yorku.

2.2 Problemy

- P01 –Większa częstotliwość zatrzymań osób innej rasy niż biała.
- P02 –Zagrożenie bycia ofiarą przestępstwa dla obywateli Nowego Yorku.
- P03 – Nie wiadomo gdzie potrzebne dodatkowe patroli.
- P04 – Brak wiedzy o tendencjach przestępstw i tym samym wiedzy dla wykrycia przyczyn.

2.3 Cel przedsięwzięcia

2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji (pytania badawcze)

Będziemy szukać pytania na następne pytania

- Jakie są najczęstsze naruszenia, ogólnie i szczegółowo?
- Jakie są tendencje wystąpienia zdarzeń w czasie
- Czy dochodzi do większej liczby zatrzymań osób czarnoskórych lub ogólnie innej rasy niż biała i z jakich powodów?
- Jakie okresy czasowe i lokacje sprzyjają największej liczby ofiar?
- Czy są jakieś korelacje pomiędzy typem ofiary i przestępcy
- W jakich miejscach są największe liczby zabitych ofiar?
- Jakie okresy czasowe sprzyjają największej liczbie zabójstw?
- Kto najczęściej jest inicjatorem i ofiarą strzelanin

2.3.2 Zakres analizy – badane aspekty

Analiza dostarczając odpowiedzi na te wszystkie pytania powinna być przydatna do odpowiedniego zarządzania oddziałami policji w Nowym Yorku. Pozwoli ona na prawidłowe rozdzielenie patroli na terenie miasta. Ponadto może ona dostarczyć informacji o tym, czy policjanci częściej nie aresztują osób o innym kolorze skóry niż biała, co może świadczyć o przejawach rasizmu, który wciąż zdaje się być problemem w Stanach Zjednoczonych.

Zdobyte odpowiedzi na pytania mogą być również bardzo przydatne dla osób mieszkających lub wybierających się do Nowego Yorku. Uzyskane informacje mogą pomóc zidentyfikować miejsca oraz pory dnia, w których dana jednostka jest bardziej narażona na zostanie ofiarą przestępstwa.

Z uzyskanej wiedzy będą mogli skorzystać aktywiści socjalne, które na podstawie danych o przestępstwach mogą wywnioskować przyczyny i podjąć działania socjalne na dla ich likwidacji.

Wszystkie uzyskane wiadomości, wyciągnięte wnioski oraz podjęte na ich podstawie działania mogą przyczynić się do zwiększenia bezpieczeństwa na terenie Nowego Yorku dla osób przebywających na jego terenie.

2.3.3 Potencjalni użytkownicy projektowanej hurtowni danych

Baza analityczna będzie wspierać procesy decyzyjne policji oraz udostępniać informacje dziennikarzom zainteresowanym tematyką zatrzymań przez policję w Nowym Yorku, na przykład spółki społeczne i naukowcy społeczeństwa.

3. Dane źródłowe

3.1. Źródła danych

Tabela 1. Źródło danych dla tematu „***Działania policji w Nowym Yorku w latach 2006-2020***”

Lp.	Plik, bazy danych	Typ	Rok	Liczba rek.	Opis
1.	NYPD_Arrest_Data__Historic_	csv	2006-2020	~ 5 099 376	Plik zawiera dane z aresztowań na terenie Nowego Yorku w danych podanych latach
2.	NYPD_Complaint_Data__Historic_	csv	2006-2020	5 153 369	Plik zawiera dane ze skarg złożonych na policję na terenie Nowego Yorku w danych latach

3.	NYPD_Shooting_Incident_Data__Historic_	csv	2006-2020	23 580	Plik zawiera listę wszystkich strzelanin, które miały miejsce w Nowym Yorku w danych latach

3.2. Lokalizacja, dostępność danych źródłowych

Dane znajdują się na stronie miasta Nowy York (https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Police+Department+%28NYPD%29&page=1). Dostępne są w postaci plików csv(csv for Excel).

Data Provided by
Police Department (NYPD)
Dataset Owner
NYC OpenData

3.3. Słownik danych – interpretacja

Tabela 2. Słownik danych w pliku *„NYPD_Complaint_Data__Historic_.csv”*

Lp.	Kolumna	Typ	Zakres dop. Wart. (puste oznacza brak ograniczeń)	Opis
1.	CMPLNT_NUM	Numer	100000000 - 999999999	Id incydentu
2.	ADDR_PCT_CD	Liczba	1-123	Dzielnica, w której doszło do incydentu
3.	BORO_NM	Tekst		Nazwa okręgu, w której doszło do incydentu
4.	CMPLNT_FR_DT	Data	01/01/2006 – 31/12/2020	Dokładna data wystąpienia zgłoszonego zdarzenia (lub data początkowa wystąpienia, jeżeli CMPLNT_TO_DT istnieje)
5.	CMPLNT_FR_TM	Data	00:00:00 – 23:59:59	Dokładny czas wystąpienia zgłoszonego zdarzenia (lub czas rozpoczęcia wystąpienia, jeżeli CMPLNT_TO_TM istnieje)
6.	CMPLNT_TO_DT	Data	01/01/2006 – 31/12/2020	Data końcowa wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany
7.	CMPLNT_TO_TM	Data	00:00:00 – 23:59:59	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
8.	CRM_ATPT_CPTD_CD	Tekst	Długość: 9	Wskaźnik, czy przestępstwo zostało pomyślnie zakończone lub usiłowane, ale nie powiodło się lub zostało przerwane przedwcześnie
9.	HADEVELOPT	Tekst		Nazwa osiedla NYCHA miejsca zdarzenia, jeśli dotyczy
10.	HOUSING_PSA	Liczba		Kod poziomu rozwoju
11.	JURISDICTION_CODE	Liczba		Jurysdykcja odpowiedzialna za incydent. Albo wewnętrzne, jak Policja (0), Tranzyt (1) i Mieszkania (2); lub zewnętrzne(3), takie jak Correction, Port Authority itp.
12.	JURIS_DESC	Tekst		Opis kodu jurysdykcji
	KY_CD	Liczba	Długość:	Trzycyfrowy kod klasyfikacji wykroczeń

13.			3	
14.	LAW_CAT_CD	Tekst	Długość: 6-11	Poziom wykroczenia: felony, misdemeanor, violation
15.	LOC_OF_OCCUR_DESC	Tekst	Długość: 6-11	Konkretna lokalizacja zdarzenia na terenie lub wokół obiektu; inside, opposite of, front of, rear of
16.	OFNS_DESC	Tekst		Opis wykroczenia odpowiadający kodowi klucza
17.	PARKS_NM	Tekst		Nazwa parku, placu zabaw lub terenów zielonych w Nowym Jorku, jeśli dotyczy (parki stanowe nie są uwzględnione)
18.	PATROL_BORO	Tekst		Nazwa dzielnicy patrolowej, w której doszło do incydentu
19.	PD_CD	Liczba	Długość: 3	Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code)
20.	PD_DESC	Tekst		Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa)
21.	PREM_TYP_DESC	Tekst		Szczegółowy opis pomieszczeń; grocery store, residence, street itp.
22.	RPT_DT	Data	01/01/2006 – 31/12/2020	Data zgłoszenia zdarzenia na policję
23.	STATION_NAME	Tekst		Nazwa stacji tranzytowej
24.	SUSP_AGE_GROUP	Tekst	Długość: 2-7	Grupa wiekowa podejrzanego
25.	SUSP_RACE	Tekst		Opis rasy podejrzanego
26.	SUSP_SEX	Tekst	F, M, U	Opis płci podejrzanego
27.	TRANSIT_DISTRICT	Liczba		Okręg tranzytowy, w którym doszło do wykroczenia.
28.	VIC_AGE_GROUP	Tekst	Długość: 2-7	Grupa wiekowa ofiary

29.	VIC_RACE	Tekst		Opis rasy ofiary
30.	VIC_SEX	Tekst	F, M, U	Opis płci ofiary
31.	X_COORD_CD	Liczba	111-1067298	Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
32.	Y_COORD_CD	Liczba	111-7250292	Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
33.	Latitude	Liczba	-90 - 90	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
34.	Longitude	Liczba	-180 - 180	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
35.	Lon_Lat	Punkt		Współrzędne długości i szerokości geograficznej do mapowania

Tabela 3. Słownik danych w pliku *„**NYPD_Arrest_Data__Historic_.csv**”*

Lp.	**Kolumna**	**Typ**	**Zakres**	**Opis**
1.	ARREST_KEY	Tekst		Trwały identyfikator generowany losowo dla każdego aresztowania
2.	ARREST_DATE	Data	01/01/2006 – 31/12/2020	Dokładna data aresztowania za zgłoszone zdarzenie
3.	PD_CD	Liczba	Długość: 3	Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code)
4.	PD_DESC	Tekst		Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa)
5.	KY_CD	Liczba	Długość: 3	Trzycyfrowy kod klasyfikacji wewnętrznej (kategoria bardziej ogólna niż kod PD)
6.	OFNS_DESC	Tekst		Opis klasyfikacji wewnętrznej odpowiadającej kodowi KY (kategoria bardziej ogólna niż opis PD)

7.	LAW_CODE	Tekst		Opłaty kodeksowe odpowiadające ustawie karnej stanu Nowy Jork, VTL i innym różnym lokalnym przepisom
8.	LAW_CAT_CD	Tekst	Długość: 6-11	Poziom wykroczenia: felony, misdemeanor, violation
9.	ARREST_BORO	Tekst	Długość: 1	Okręg w której doszło do aresztowania. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens)
10.	ARREST_PRECINCT	Liczba		Dzielnica, w której doszło do aresztowania
11.	JURISDICTION_CODE	Liczba		Jurysdykcja odpowiedzialna za areszt. Kody jurysdykcji 0 (Patrol), 1 (tranzyt) i 2 (mieszkanie) reprezentują NYPD, podczas gdy kody 3 i więcej reprezentują jurysdykcje inne niż NYPD
12.	AGE_GROUP	Tekst	Długość: 3-7	Wiek sprawcy w ramach kategorii
13.	PERP_SEX	Tekst	F, M, U	Opis płci sprawcy
14.	PERP_RACE	Tekst		Opis rasy sprawcy
15.	X_COORD_CD	Liczba		Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
16.	Y_COORD_CD	Liczba		Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
17.	Latitude	Liczba	-90 - 90	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
18.	Longitude	Liczba	-180 - 180	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)

Tabela 4. Słownik danych w pliku ***„NYPD_Shooting_Incident_Data__Historic_csv"***

Lp.	**Kolumna**	**Typ**	**Zakres**	**Opis**
1.	INCIDENT_KEY	Tekst	9953245-230611229	Trwały identyfikator generowany losowo dla każdego aresztowania

2.	OCCUR_DATE	Data	01/01/2006 – 31/12/2020	Dokładna data zdarzenia strzeleckiego
3.	OCCUR_TIME	Tekst	00:00:00 – 23:59:59	Dokładny czas zdarzenia strzeleckiego
4.	BORO	Tekst		Okręg, w której doszło do strzelaniny
5.	PRECINCT	Liczba		Dzielnica, w której doszło do strzelaniny
6.	JURISDICTION_CODE	Liczba		Jurysdykcja, w której doszło do strzelaniny. Kody jurysdykcji 0 (Patrol), 1 (tranzyt) i 2 (mieszkanie) reprezentują NYPD, podczas gdy kody 3 i więcej reprezentują jurysdykcje inne niż NYPD
7.	LOCATION_DESC	Tekst		Miejsce zdarzenia strzeleckiego
8.	STATISTICAL_MURDER_FLAG	Bool	1	Postrzelenie zakończyło się śmiercią ofiary, która zostałaby zaliczona jako morderstwo
9.	PERP_AGE_GROUP	Tekst	Długość: 3-7	Wiek sprawcy w ramach kategorii
10.	PERP_SEX	Tekst	F, M, U	Opis płci sprawcy
11.	PERP_RACE	Tekst		Opis rasy sprawcy
12.	VIC_AGE_GROUP	Tekst	Długość: 3-7	Grupa wiekowa ofiary
13.	VIC_SEX	Tekst	F, M, U	Opis rasy ofiary
14.	VIC_RACE	Tekst		Opis płci ofiary
15.	X_COORD_CD	Liczba		Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)

16.	Y_COORD_CD	Liczba		Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
17.	Latitude	Liczba	-90 - 90	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
18.	Longitude	Liczba	-180 - 180	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)

3.4. Ocena jakościowa danych

Tabela 5. Ocena jakości danych w pliku *„NYPD_Complaint_Data__Historic_.csv”*

Lp.	Kolumna	Typ	Zakres	Ocena jakości
1.	CMPLNT_NUM	Numer	100000065 - 999999904	Id incydentu 99.96% unikalne, usunąć powtarzające się ID
2.	ADDR_PCT_CD	Liczba	1-123	Dzielnica, w której doszło do incydentu
3.	BORO_NM	Tekst	Długość: 5-13	Nazwa okręgu, w której doszło do incydentu 1% null, usuwamy
4.	CMPLNT_FR_DT	Data	14/05/1010 – 31/12/2020	Dokładna data wystąpienia zgłoszonego zdarzenia (lub data początkowa wystąpienia, jeżeli CMPLNT_TO_DT istnieje) Pozbyć się błędnych danych
5.	CMPLNT_FR_TM	Data	00:00:00 – 23:59:00	Dokładny czas wystąpienia zgłoszonego zdarzenia (lub czas rozpoczęcia wystąpienia, jeżeli CMPLNT_TO_TM istnieje)
6.	CMPLNT_TO_DT	Data	15/10/1010-06/04/2090	Data końcowa wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany Pozbyć się błędnych danych
7.	CMPLNT_TO_TM	Data	00:00:00 – 23:59:00	Ending time of occurrence for the reported event, if exact time of occurrence is unknown 1% null, usuwamy
8.	CRM_ATPT_CPTD_CD	Tekst	Długość: 9	Wskaźnik, czy przestępstwo zostało pomyślnie zakończone lub usiłowane, ale nie powiodło się lub zostało przerwane przedwcześnie
9.	HADEVELOPT	Tekst	Długość: 4-43	Nazwa osiedla NYCHA miejsca zdarzenia, jeśli dotyczy 95% null
10.	HOUSING_PSA	Liczba	Długość: 1-7	Kod poziomu rozwoju

				93% null
11.	JURISDICTION_CODE	Liczba	0-97	Jurysdykcja odpowiedzialna za incydent. Albo wewnętrzne, jak Policja (0), Tranzyt (1) i Mieszkania (2); lub zewnętrzne(3), takie jak Correction, Port Authority itp.
12.	JURIS_DESC	Tekst	Długość: 5/35	Opis kodu jurysdykcji
13.	KY_CD	Liczba	101-881	Trzycyfrowy kod klasyfikacji wykroczeń
14.	LAW_CAT_CD	Tekst	Długość: 6-11	Poziom wykroczenia: felony, misdemeanor, violation
15.	LOC_OF_OCCUR_DESC	Tekst	Długość: 6-11	Konkretna lokalizacja zdarzenia na terenie lub wokół obiektu; inside, opposite of, front of, rear of 17% null, robimy UNKNOWN
16.	OFNS_DESC	Tekst	Długość: 4-36	Opis wykroczenia odpowiadający kodowi klucza 1% null, usuwamy
17.	PARKS_NM	Tekst	Długość: 2-83	Nazwa parku, placu zabaw lub terenów zielonych w Nowym Jorku, jeśli dotyczy (parki stanowe nie są uwzględnione) 99% null, robimy NONE
18.	PATROL_BORO	Tekst	Długość: 17-25	Nazwa dzielnicy patrolowej, w której doszło do incydentu
19.	PD_CD	Liczba	100-975	Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code)
20.	PD_DESC	Tekst	Długość: 6-71	Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa)
21.	PREM_TYP_DESC	Tekst	Długość: 3-28	Szczegółowy opis pomieszczeń; grocery store, residence, street itp. 1% null, usuwamy
22.	RPT_DT	Data	01/01/2006 – 31/12/2020	Data zgłoszenia zdarzenia na policję
23.	STATION_NAME	Tekst	Długość: 6-30	Nazwa stacji tranzytowej 97% null

24.	SUSP_AGE_GROUP	Tekst	Długość: 2-7	Grupa wiekowa podejrzanego 65% null, zamienic na UNKNOWN Usuwamy dane które nie są okresem lat
25.	SUSP_RACE	Tekst	Długość: 5-30	Opis rasy podejrzanego 46% null, zamienic na UNKNOWN
26.	SUSP_SEX	Tekst	Długość: 1	Opis płci podejrzanego 48% null, zamienic na U
27.	TRANSIT_DISTRICT	Liczba	1-34	Okręg tranzytowy, w którym doszło do wykroczenia. 97% null
28.	VIC_AGE_GROUP	Tekst	Długość: 2-7	Grupa wiekowa ofiary 1% null, zamienic na UNKNOWN Usuwamy dane które nie są okresem lat
29.	VIC_RACE	Tekst	Długość: 5-30	Opis rasy ofiary 1% null, zamienic na UNKNOWN
30.	VIC_SEX	Tekst	Długość: 1	Opis płci ofiary 1% null, zamienic na UNKNOWN Zamienic niepasujące dane na U
31.	X_COORD_CD	Liczba	111-1067298	Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104) 1% null, usuwamy
32.	Y_COORD_CD	Liczba	111-7250292	Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104) 1% null, usuwamy
33.	Latitude	Liczba	40,112709974 – 59,657273946	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
34.	Longitude	Liczba	-77,519206334 - -73.684788384	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
35.	Lon_Lat	Punkt		Współrzędne długości i szerokości geograficznej do mapowania

Tabela 6. Ocena jakości danych w pliku *„NYPD_Arrest_Data__Historic_.csv”*

Lp.	Kolumna	Typ	Zakres	Ocena jakości
1.	ARREST_KEY	Tekst	9926901 - 222500606	Trwały identyfikator generowany losowo dla każdego aresztowania
2.	ARREST_DATE	Data	01/01/2006 – 31/12/2020	Dokładna data aresztowania za zgłoszone zdarzenie
3.	PD_CD	Liczba	0-997	Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code) 1% null, usuwamy
4.	PD_DESC	Tekst	Długość: 6-54	Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa) 1% null, usuwamy
5.	KY_CD	Liczba	101-995	Trzycyfrowy kod klasyfikacji wewnętrznej (kategoria bardziej ogólna niż kod PD) 1% null, usuwamy
6.	OFNS_DESC	Tekst	Długość: 4-43	Opis klasyfikacji wewnętrznej odpowiadającej kodowi KY (kategoria bardziej ogólna niż opis PD) 1% null, usuwamy
7.	LAW_CODE	Tekst	Długość: 2-10	Opłaty kodeksowe odpowiadające ustawie karnej stanu Nowy Jork, VTL i innym różnym lokalnym przepisom
8.	LAW_CAT_CD	Tekst	Długość: 1	Poziom wykroczenia 1% null, usuwamy Usuwamy niepasujące dane Wpisujemy cala nazwę
9.	ARREST_BORO	Tekst	Długość: 1	Okręg, w której doszło do aresztowania. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens) Wpisujemy cala nazwę
10.	ARREST_PRECINCT	Liczba	1-123	Dzielnica, w której doszło do aresztowania
11.	JURISDICTION_CODE	Liczba	0-97	Jurysdykcja odpowiedzialna za areszt. Kody jurysdykcji 0 (Patrol), 1 (tranzyt) i 2 (mieszkanie) reprezentują NYPD, podczas gdy kody 3 i więcej reprezentują jurysdykcje inne niż NYPD
12.	AGE_GROUP	Tekst	Długość:	Wiek sprawcy w ramach kategorii

			3-7	Usuwamy dane które nie są okresem lat
13.	PERP_SEX	Tekst	Długość: 1	Opis płci sprawcy
14.	PERP_RACE	Tekst	Długość: 5-30	Opis rasy sprawcy
15.	X_COORD_CD	Liczba	913357-1067302	Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104) 1% null, usuwamy
16.	Y_COORD_CD	Liczba	121131-8202360	Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104) 1% null, usuwamy
17.	Latitude	Liczba	40.49890536 – 62.08307497	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326) 1% null, usuwamy
18.	Longitude	Liczba	-74.254938736 - -73.681780268	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326) 1% null, usuwamy

Tabela 7. Ocena jakości danych w pliku ***„NYPD_Shooting_Incident_Data__Historic_csv”***

Lp.	**Kolumna**	**Typ**	**Zakres**	**Ocena jakości**
1.	INCIDENT_KEY	Tekst	9953245-230611229	Trwały identyfikator generowany losowo dla każdego aresztowania
2.	OCCUR_DATE	Data	01/01/2006 – 31/12/2020	Dokładna data zdarzenia strzeleckiego
3.	OCCUR_TIME	Tekst	00:00:00 – 23:59:00	Dokładny czas zdarzenia strzeleckiego
	BORO	Tekst	Długość:	Okręg, w której doszło do strzelaniny

4.			5-13	
5.	PRECINCT	Liczba	1-123	Dzielnica, w której doszło do strzelaniny
6.	JURISDICTION_CODE	Liczba	0-2	Jurysdykcja, w której doszło do strzelaniny. Kody jurysdykcji 0 (Patrol), 1 (tranzyt) i 2 (mieszkanie) reprezentują NYPD, podczas gdy kody 3 i więcej reprezentują jurysdykcje inne niż NYPD 1% null, usuwamy
7.	LOCATION_DESC	Tekst	Długość: 3-25	Miejsce zdarzenia strzeleckiego 58% null, zamienić na UNKNOWN
8.	STATISTICAL_MURDER_FLAG	Bool	1	Postrzelenie zakończyło się śmiercią ofiary, która zostałaby zaliczona jako morderstwo
9.	PERP_AGE_GROUP	Tekst	Długość: 3-7	Wiek sprawcy w ramach kategorii 35% null, zamienić na UNKNOWN Usuwamy dane które nie są okresem lat
10.	PERP_SEX	Tekst	Długość: 1	Opis płci sprawcy 35% null, zamienić na U
11.	PERP_RACE	Tekst	Długość: 5-30	Opis rasy sprawcy 35% null, zamienić na UNKNOWN
12.	VIC_AGE_GROUP	Tekst	Długość: 3-7	Grupa wiekowa ofiary
13.	VIC_SEX	Tekst	Długość: 1	Opis rasy ofiary
14.	VIC_RACE	Tekst	Długość: 5-30	Opis płci ofiary
15.	X_COORD_CD	Liczba	914928- 1066815	Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
16.	Y_COORD_CD	Liczba	125756- 271127	Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)

17.	Latitude	Liczba	40.51158556 – 40.91081808	Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
18.	Longitude	Liczba	-74.25930349 - -73.70204616	Środkowa współrzędna długości geograficznej dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)

4. Analityczne modele wielowymiarowe

4.1. Fakty podlegające analizie oraz ich miary

Tabela 8. Fakty oraz ich miary opracowywanych modeli analitycznych

Lp.	Fakt	Miary
1.	Zdarzenie	- Liczba ofiar -Liczba przestępców - Liczba zginiętych ofiary

4.2. Kontekst analizy faktów

Tabela 9. Zidentyfikowane wymiary wraz z ich własnościami (charakterystykami) opracowywanych modeli analitycznych

Lp.	Fakt	Miary	Efekt
1.	Data zdarzenia	- Dzień - Miesiąc - Rok	Pozwala na analizę zdarzeń w czasie i wyjaśnienie najbardziej aktywnych okresów czasowych
2.	Czas zdarzenia	- Godzina - Przedział godzinowy - Minuta - Przedział minutowy	Pozwala na analizę zdarzeń w czasie i wyjaśnienie najbardziej aktywnych okresów czasowych

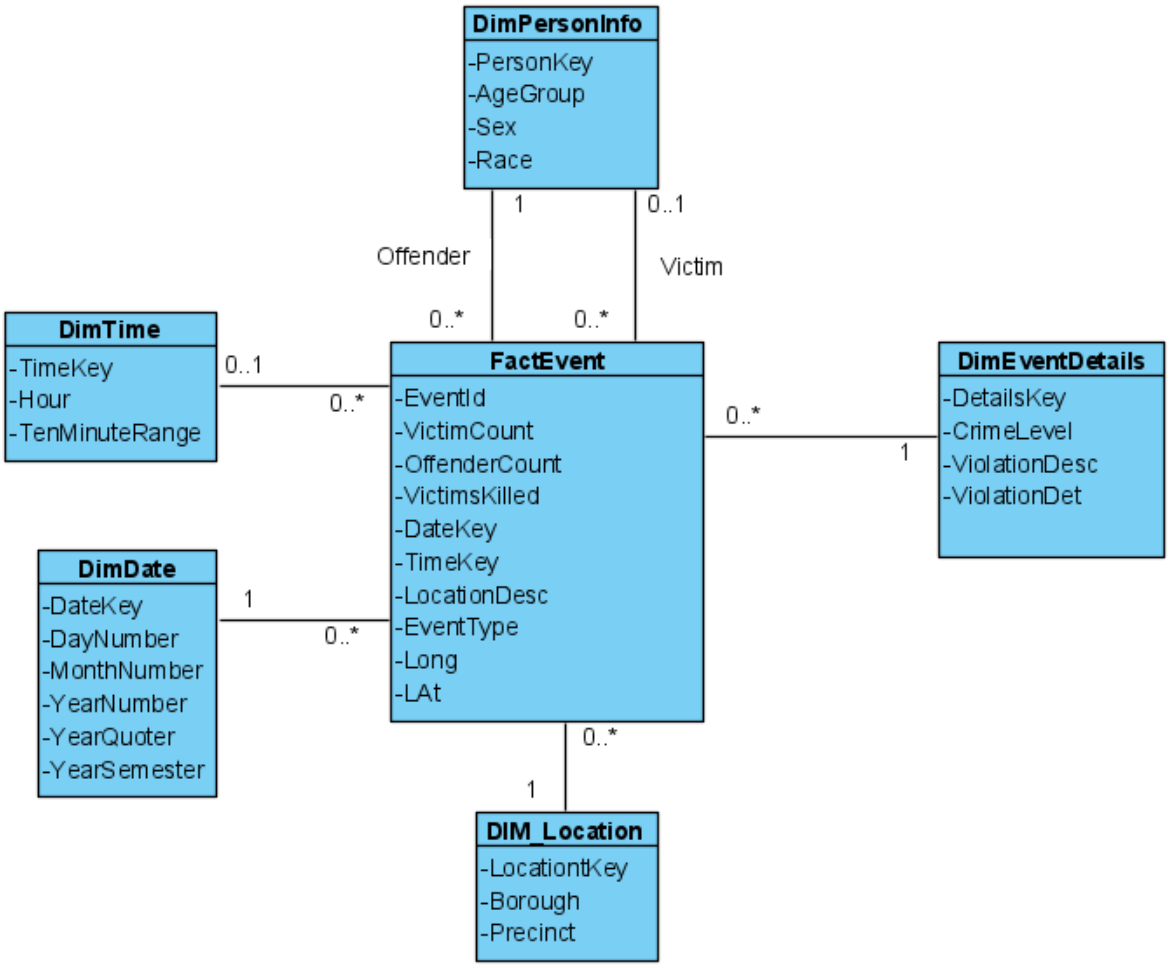
3.	Osoba biorąca udział w zdarzeniu	- Wiek - Płeć - Rasa	Pozwala na analizę według charakterystyk ofiary i przestępcy, kto ma największa szansa zostać kim i kto częściej na kogo napada
4.	Miejsce zdarzenia	- Region - Okolica - Opis lokacji	Pozwala na analizę miejsc zdążenia, jakie regiony są kryminalne i w jakich miejscach najczęściej dochodzi do przestępstwa
5.	Charakterystyka zdarzenia	- Rodzaj zdarzenia - Poziom przestępstwa - Opis przestępstwa - Szczególny opis przestępstwa	Pozwala na analizę według charakterystyk przestępstw, jakie są najczęstsze i w jakich warunkach

4.3. Modele wielowymiarowe (UML)

Do analizy dane o aresztowaniach, skargach i strzelanin będą złączone do jednego ogólnego rodzaju zdarzenia - zdarzenie. W końcu otrzymujemy 1 tabele faktów i 5 tabel kontekstów.

Niektóre dane, które są bardziej szczegółowe dla zdarzeń, są przechowywane w tabeli faktów jak wymiary zdegenerowane, bo ich włączenie do tabel wymiarów zbyt dużo powieli dane, co zniszczy sens rozdzielenia danych na tabele.

Podczas złączenia typów zdarzeń w jedna kategorie był problem, ze nie wszystkie atrybuty są współdzielone przez wszystkie zdarzenia. Gdzie było to możliwe zostały wykorzystane odpowiednie wartości. Jednak taka transformacja dla na końcu dwa powiazania wiele do zero lub jeden. Areszty, które stanowią prawie polowe wszystkich rekordów, nie maja ofiary i czasu. Powiazanie 0..1 z ofiara było zastawione, bo dane przestępców i ofiar są przechowywane w jednej tabeli. Powiazanie 0..1 z czasem było zostawione, aby było wygodnie definiować dodatkowe miary dla czasu i nie dodawać zbyt dużo atrybutów do tabeli faktów dla wygodności pracy.



Rysunek 1 Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

5. Projekt procesu ETL

5.1. Schemat bazy danych HD (SQL)

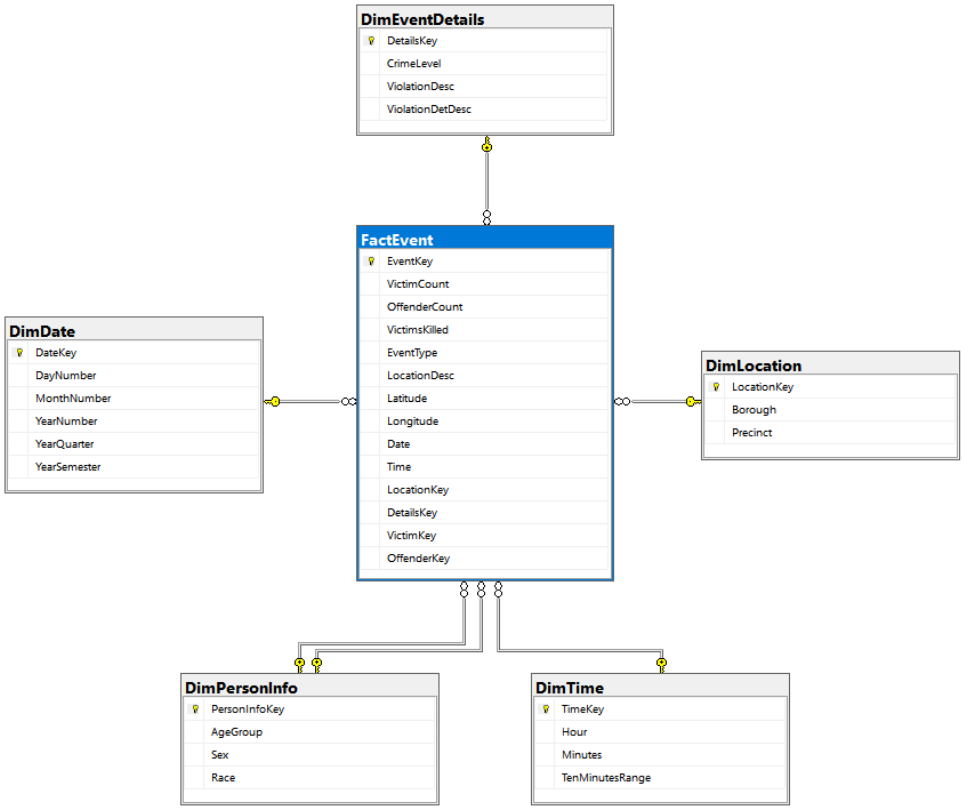
Baza danych została stworzona według tych zapytań SQL

Tabela 10. Skrypty tworzące tabele bazy danych

```
CREATE TABLE DimDate (
DateKey DATE NOT NULL,
DayNumber tinyint NOT NULL,
MonthNumber tinyint NOT NULL,
YearNumber smallint NOT NULL,
YearQuarter tinyint NOT NULL,
YearSemester tinyint NOT NULL
CONSTRAINT PK_DimDate_DateKey PRIMARY KEY (DateKey)
);

CREATE TABLE DimTime (
TimeKey TIME NOT NULL,
Hour int NOT NULL,
Minutes int NOT NULL,
TenMinutesRange varchar(5) NOT NULL,
CONSTRAINT PK_DimTime_TimeKey PRIMARY KEY (TimeKey)
);



CREATE TABLE DimLocation  (
    [LocationKey] bigint NOT NULL,
    [Borough] nvarchar(13) NOT NULL,
    [Precinct] bigint NOT NULL,
CONSTRAINT PK_DimLocation_LocationKey PRIMARY KEY (LocationKey)
)

CREATE TABLE DimEventDetails  (
    [DetailsKey] bigint NOT NULL AUTO_INCREMENT,
    [CrimeLevel] nvarchar(11) NOT NULL,
    [ViolationDesc] nvarchar(50) NOT NULL,
    [ViolationDetDesc] nvarchar(100) NOT NULL,
CONSTRAINT PK_DimEventDetails_DetailsKey PRIMARY KEY (DetailsKey)
)

CREATE TABLE DimPersonInfo (
    [PersonInfoKey] bigint NOT NULL AUTO_INCREMENT,
    [AgeGroup] nvarchar(7) NOT NULL,
    [Sex] nvarchar(1) NOT NULL,
    [Race] nvarchar(30) NOT NULL,
CONSTRAINT PK_DimPerson_PersonInfoKey PRIMARY KEY (PersonInfoKey)
)

CREATE TABLE FactEvent(
    [EventKey] bigint NOT NULL,
    [VictimCount] int NOT NULL,
    [OffenderCount] int NOT NULL,
```

```
    [VictimsKilled] int NOT NULL,
    [EventType] nvarchar(10) NOT NULL,
    [LocationDesc] nvarchar(100) NOT NULL,
    [Latitude] float NOT NULL,
    [Longitude] float NOT NULL,
    [Date] DATE NOT NULL,
    [Time] TIME,
    [LocationKey] bigint NOT NULL,
    [DetailsKey] bigint NOT NULL,
    [VictimKey] bigint,
    [OffenderKey] bigint,
CONSTRAINT PK_FactEvent_EventKey PRIMARY KEY (EventKey),
CONSTRAINT FK_FactEvent_DimDate FOREIGN KEY (Date) REFERENCES DimDate (DateKey),
CONSTRAINT FK_FactEvent_DimTime FOREIGN KEY (Time) REFERENCES DimTime (TimeKey),
CONSTRAINT FK_FactEvent_DimLocation FOREIGN KEY (LocationKey) REFERENCES DimLocation
(LocationKey),
CONSTRAINT FK_FactEvent_DimEventDetails FOREIGN KEY (DetailsKey) REFERENCES
DimEventDetails (DetailsKey),
CONSTRAINT FK_FactEvent_DimPerson_Victim FOREIGN KEY (VictimKey) REFERENCES
DimPersonInfo(PersonInfoKey),
CONSTRAINT FK_FactEvent_DimPerson_Offender FOREIGN KEY (OffenderKey) REFERENCES
DimPersonInfo (PersonInfoKey)
)
```

Wygenerowany przez Microsoft SQL Server Management Studio schemat bazy został przedstawiony na rys.



Rysunek 2 Schemat bazy danych utworzonej przy pomocą skryptu (tabela 10)

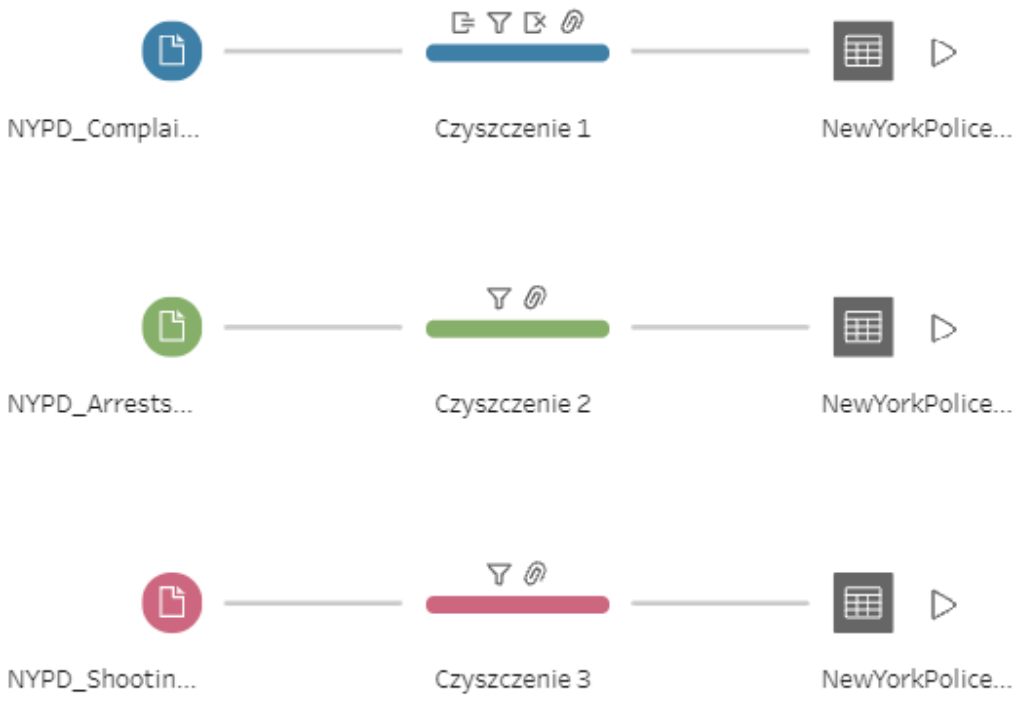
5.2. Specyfikacja procesów ETL (Control Flow + Data Flow)

Proces ETL składa się z obróbki danych przy pomocy Tableu prep i procesu SSIS.

Tableu prep

W Tableu prep została wykonana usuwanie wartości null, rozpisane skrócone dane i usunięte literówki. Narzędzie zostało użyte z powodu, ze Fuzzy Grouping nie jest zbyt precyzyjne i nie może naprawić wartości z kilka powodów. Po pierwsze rekord może nie mieć poprawnego odpowiednika, po drugie są pewne skróty w zapisie, które można rozpisać dla większej czytelności, po trzecie, co jest najbardziej istotnym powodem, literówki, które występują, mogą być różnicą o jeden symbol lub o cale słowo, dlatego potrzebna jest odnośnie niski próg podobieństwa, ale w bazie występują takie wartości jak „ABORTION", „ABORTION 1", „ASSAULT 2,3" lub „ARSON 1" i „ARSON 2,3,4", które są podobne, ale różne, i Fuzzy Grouping o niskim progu podobieństwa może zmienić te nazwy i dane nie będą odpowiadały rzeczywistości. Dlatego zdecydowano na naprawę wartości ręcznie przy pomocy Tableu Prep.

Procesy obróbki danych są zrobione dla wszystkich plików źródłowej i ich zawartości zostały załadowane do bazy danych – poczekalni.



Rysunek 3 Data Flow w Tableu Prep do obróbki danych z plików źródłowych

Przykład problematycznych danych

Skróty

CUSTODIAL INTERFERENCE 2
DIS. CON.,AGGRAVATED
DISORDERLY CONDUCT
DRUG PARAPHERNALIA, PO

Rysunek 4 Przykład skróconej wartości atrybutu

Brak kilku słów/znacznej ilości znaków

LARCENY, PETIT OF MOPED
LARCENY,GRAND BY ACQUIRING LOS
LARCENY,GRAND BY ACQUIRING LOST CREDIT CARD

Rysunek 5 Przykład braku dużej ilości znaków w wartosci atrybutu

SSIS

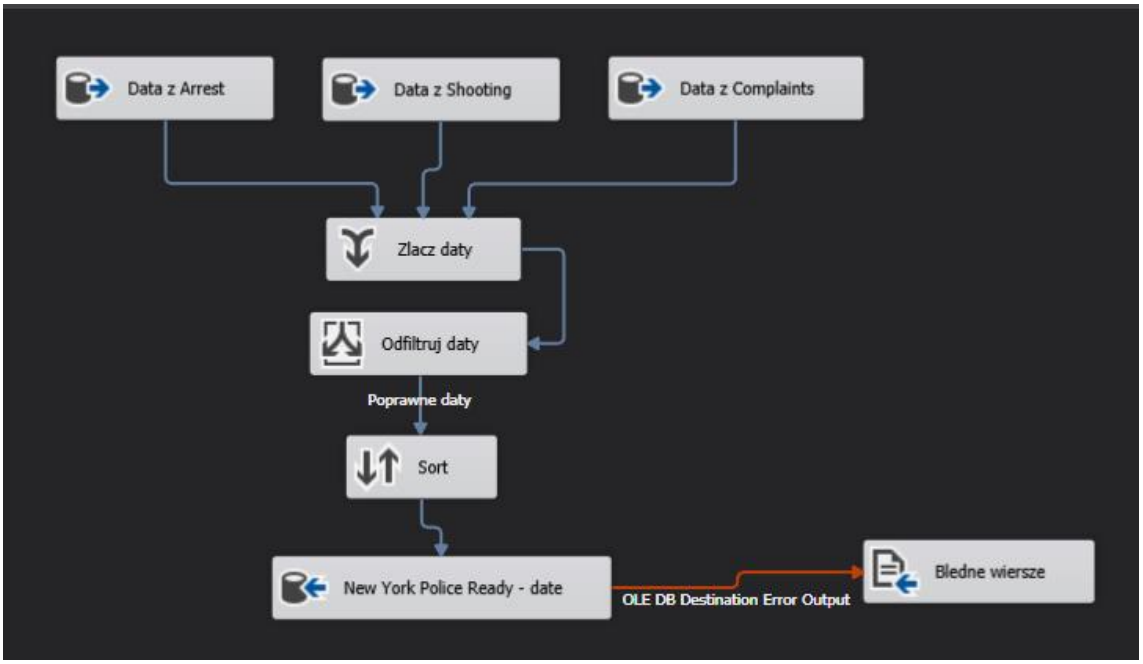
Został zdefiniowany jeden pakiet ETL, który polega na lądowaniu, sprawdzeniu i dodawaniu danych do bazy danych HD.

Control Flow:

Rysunek 6 Schemat Control Flow procesu w Visual Studio

W co bloczku bierzemy potrzebne wartości z co tabeli, kształtujemy do potrzebnej postaci i ładujemy do odpowiednich tabel, sprawdzając duplikaty. Dane wstępnie są opracowane w Tableu Prep i załadowany do bazy danych - poczekalni, proces ETL tylko kształtuje dany do modelu analitycznego, konwertuje dane do potrzebnego typu, sprawdza poprawność pewnych atrybutów, wartości których są znane, na przykład nazwy okręgów i zakres czasowy badanych danych, występując filtrem.

Dla daty i czasu od razu są brane dane ze wszystkich tabel. Podczas ładowania sprawdzamy zakres przy pomocy Conditional Split. Używa się klucz naturalny dla daty, czasu i faktów, dlatego sama baza danych kontroluje duplikaty.



Rysunek 7 Schemat Data Flow dla wymiaru dat

Dane dat i czasu są brane w następujący sposób

SQL command text:

```
SELECT DISTINCT
    ARREST_DATE AS "DateKey",
            DATEPART([DAY], ARREST_DATE) AS "DayNumber",
            DATEPART([MONTH], ARREST_DATE) AS
"MonthNumber",
            DATEPART([YEAR], ARREST_DATE) AS "YearNumber",
            DATEPART(QUARTER, ARREST_DATE) AS
"YearQuarter",
            CASE WHEN DATEPART(quarter, [ARREST_DATE]) >=
3 THEN 2 ELSE 1 END AS "YearSemester"
FROM dbo.ArrestData;
```

Rysunek 8 Skrypt do pobierania danych daty z bazy-poczekalni

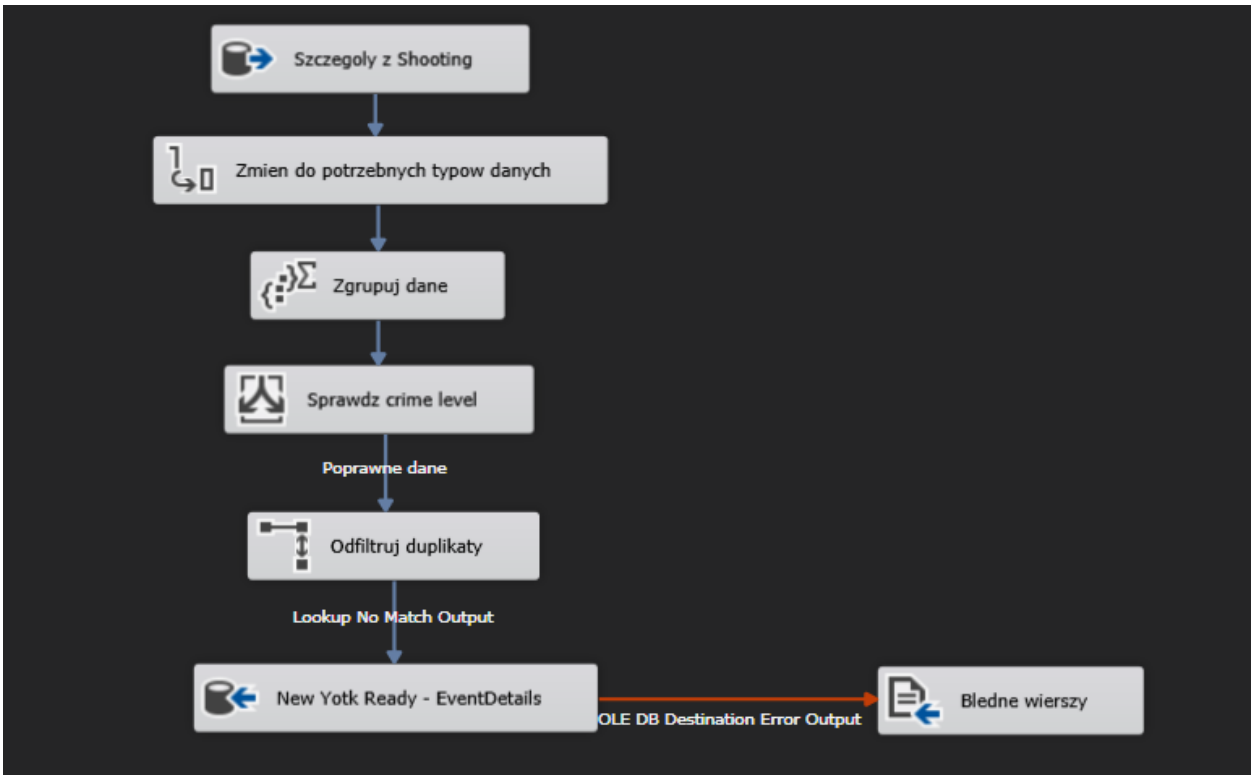
SQL command text:

```
SELECT DISTINCT
    OCCUR_TIME AS "FullTime",
            DATEPART([HOUR], OCCUR_TIME ) AS "Hour",
            DATEPART([MINUTE], OCCUR_TIME ) AS "Minutes",
           CASE
  WHEN DATEPART([MINUTE], OCCUR_TIME) <=10 THEN '0-9'
  WHEN DATEPART([MINUTE], OCCUR_TIME) <=20 THEN '10-19'
  WHEN DATEPART([MINUTE], OCCUR_TIME) <=30 THEN '20-29'
  WHEN DATEPART([MINUTE], OCCUR_TIME) <=40 THEN '30-39'
  WHEN DATEPART([MINUTE], OCCUR_TIME) <=50 THEN '40-49'
  ELSE '50-59'
END AS "TenMinutesRange"
FROM dbo.ShootingData
WHERE not( OCCUR_TIME is NULL);
```

Rysunek 9 Skrypt do pobierania danych czasu z bazy-poczekalni

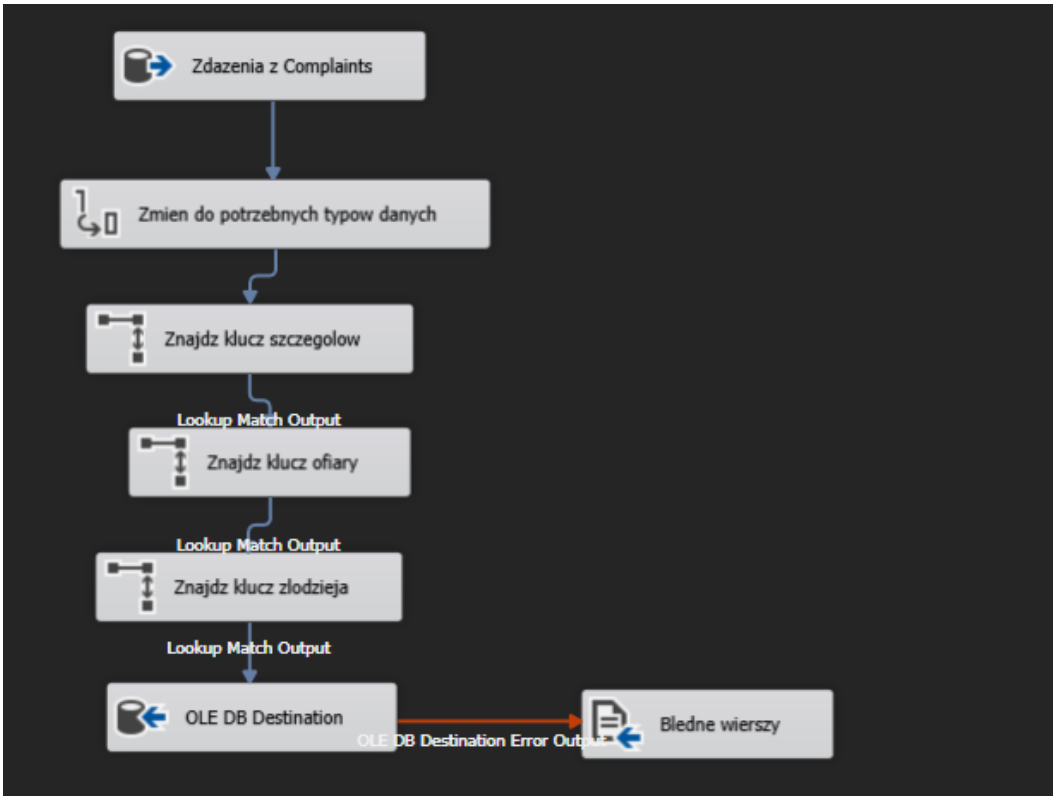
Dla kolejnych wymiarów bierzemy dane, grupujemy wartości przy pomocy Aggregate i sprawdzamy duplikaty przy pomocy lookup, szukamy potrzebne wartości w bazie i rekordy, dla których nie zostało znaleziono parę, są przekazywane do bazy. Gdzie możliwie, sprawdzamy wartości atrybutów.

(Tak jak dane brane z bazy danych, można wykorzystać DISTINCT, ale zdecydowano na zostawienie tego elementu, aby w razie zmiany źródła zapomnienie o używaniu DISTINCT nie psuło ładowania danych)



Rysunek 10 Data Flow do ładowania danych z danych o strzelaninach do tablicy wymiaru szczegółów zdarzenia

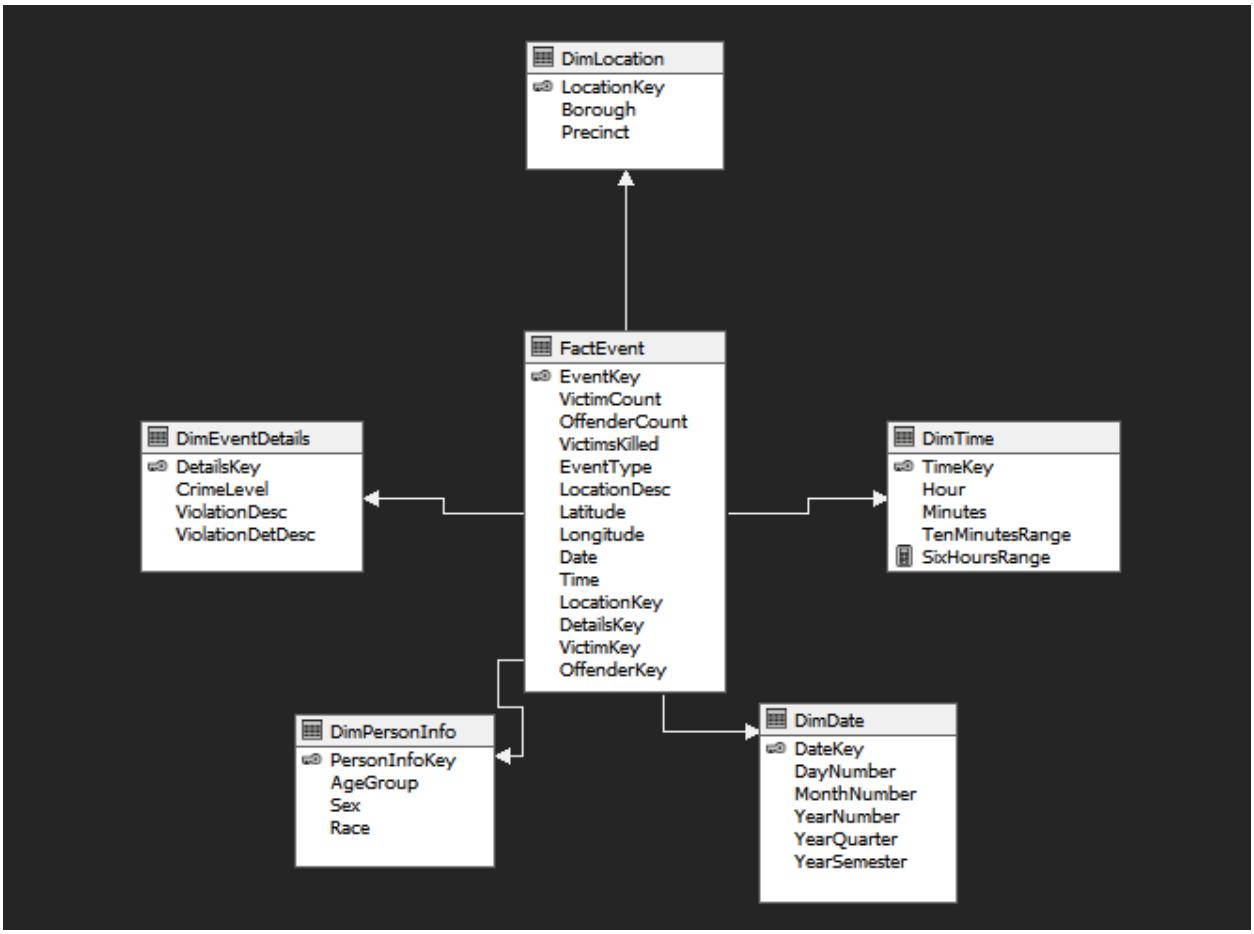
Przy lądowaniu tabeli faktów przy pomocy lookup znajdujemy klucze dla poszczególnych wymiarów i dodajemy rekordy do bazy. Używany jest klucz naturalny, dlatego duplikaty są kontrolowane przez bazę.



Rysunek 11 Data Flow do ładowania danych z danych o skargach do tablicy faktu zdarzenia

6. Implementacja modeli wielowymiarowych

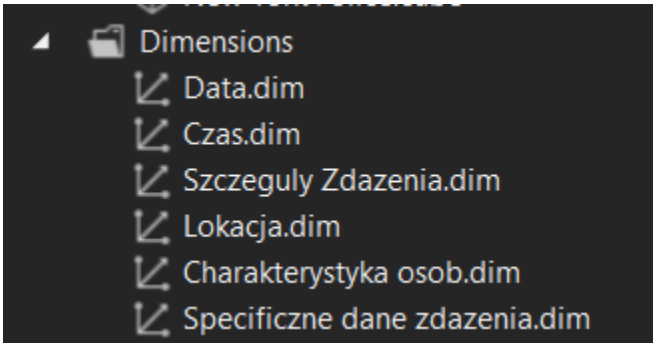
6.1. Widok danych



Rysunek 12 Schemat widoku danych w projekcie definiowania modeli wielowymiarowej

6.2. Wymiary

Zostały stworzone następne wymiary

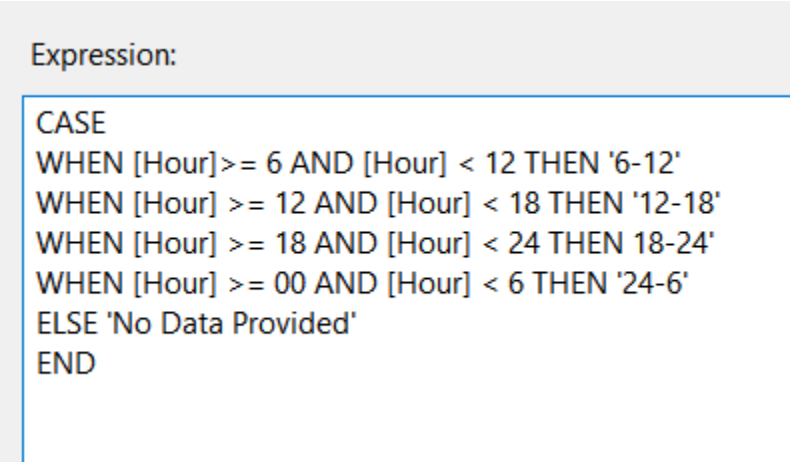


Rysunek 13 Spis zdefiniowanych wymiarów

Specyficzne dane zdążenia jest wymiarem, stworzonym na podstawie wymiarów zdegenerowanych.

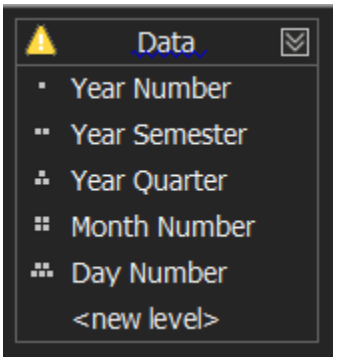
Dodatkowo zostały zdefiniowane następne atrybuty:

DimTime: SixHoursRange

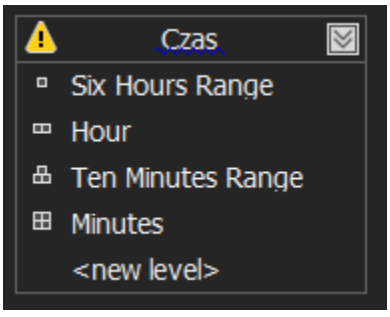


Rysunek 14 Wyrażenie SQL do definiowania New Named Calculation

Zostały zdefiniowane następne hierarchie



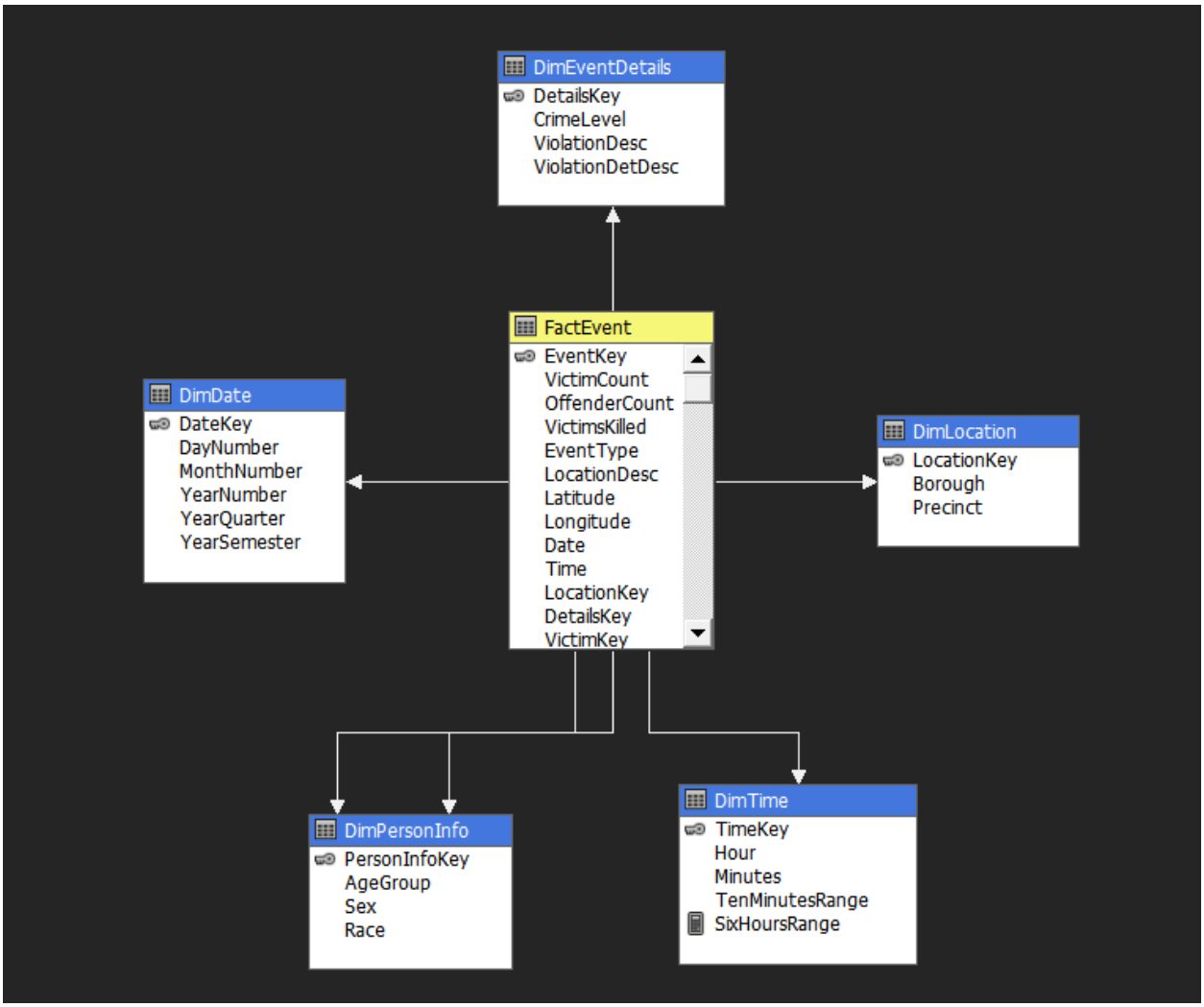
Rysunek 15 Hierarchia dat



Rysunek 16 Hierarchia czasu

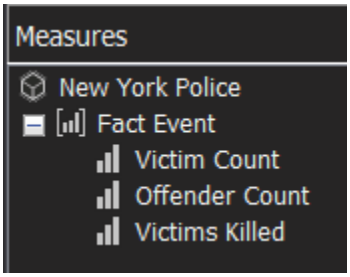
Rysunek 17 Hierarchie dotyczące informacji o osobie

6.3. Modele wielowymiarowe - kostki



Rysunek 18 Schemat kostki

Miary zostały zdefiniowany na etapie tworzenia kostki



Rysunek 19 Lista miar kostki

7. Analiza danych

Procesy analityczne wykonałam za pomocą Tableau Desktop i zrealizowałam według następującego scenariusza, które jest podzielone według pytań badawczych:

- Jakie są najczęstsze naruszenia, ogólnie i szczegółowo?

Liczba przestępców (de facto liczba zdarzeń) z zależności od:

Proste: opis przestępstwa, poziom przestępstwa

Złożone: opis przestępstwa (top 2) + szczegóły przestępstwa(top 5)

- Jakie są tendencje wystąpienia zdarzeń w czasie

Liczba przestępców (de facto liczba zdarzeń) z zależności od:

Proste: rok

Złożone: rok + kwartał, poziom przestępstwa + rok

- Czy dochodzi do większej liczby zatrzymań osób czarnoskórych lub ogólnie innej rasy niż biała i z jakich powodów?

Liczba przestępców (de facto liczba zdarzeń) z zależności od:

Złożone: typ zdarzenia(Arrest) + rasa przestępcy, typ zdarzenia(Arrest) + rasa (top 1) + opis przestępstwa

- Jakie okresy czasowe i lokacje sprzyjają największej liczby ofiar?

Liczba ofiar (de facto zdarzenia bez aresztowań) z zależności od:

Proste: okręg, opis lokacji, przedział godzinowy

Złożone: okręg + przedział godzinowy, okręg + opis lokacji

- Czy są jakieś korelacje pomiędzy typem ofiary i przestępcy

Liczba ofiar (de facto zdarzenia bez aresztowań) z zależności od:

Złożone: rasa ofiary + rasa przestępcy, kategoria wiekowa przestępcy + kategoria wiekowa ofiary, pleć ofiary + pleć przestępcy,

- W jakich miejscach są największe liczby zabitych ofiar?

Liczba zabitych ofiar (de facto liczba zabójstw przy strzelaninach) w zależności od:

Proste: okręg, opis lokacji

Złożone: okręg + opis lokacji (top 5)

- Jakie okresy czasowe sprzyjają największej liczbie zabójstw?

Liczba zabitych ofiar (de facto liczba zabójstw przy strzelaninach) w zależności od:

Proste: przedział godzinowy

Złożone: przedział godzinowy + przedział minutowy

- Kto najczęściej jest inicjatorem i ofiara strzelanin

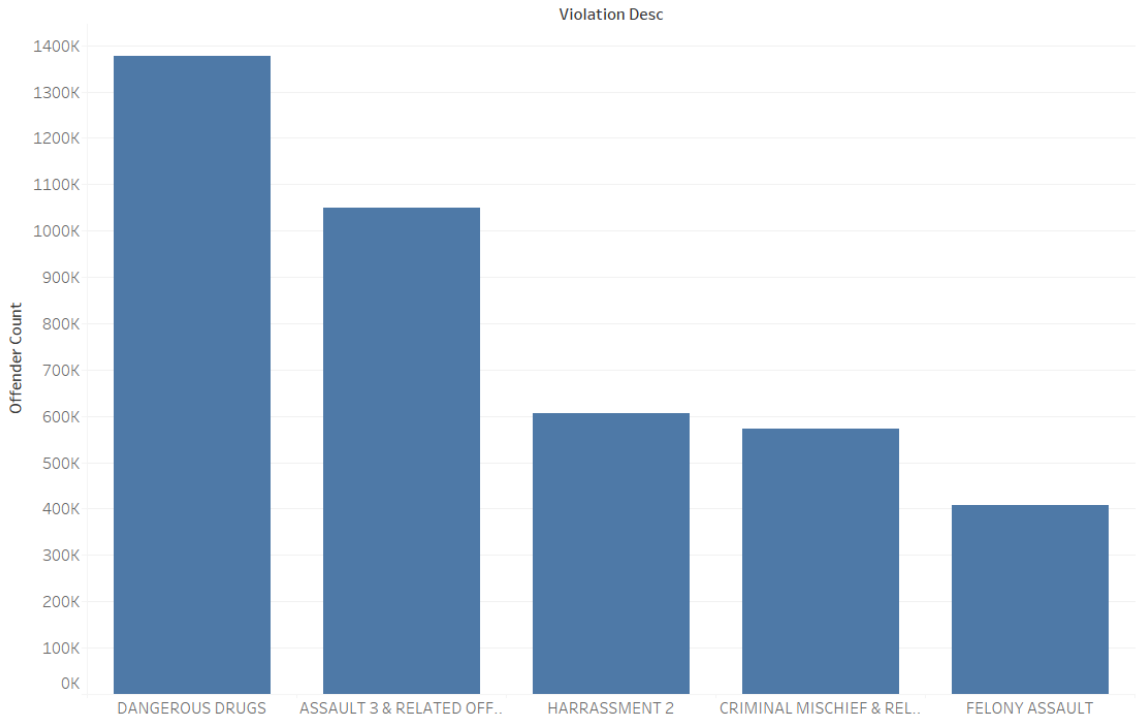
Liczba zabitych ofiar (de facto liczba zabójstw przy strzelaninach) w zależności od:

Proste: rasa ofiary, rasa przestępcy, kategoria wiekowa przestępcy, kategoria wiekowa ofiary, pleć ofiary, pleć przestępcy

Złożone: rasa ofiary + kategoria wiekowa ofiary, pleć ofiary + kategoria wiekowa ofiary, rasa przestępcy + kategoria wiekowa przestępcy, pleć przestępcy + kategoria wiekowa przestępcy

7.1. Realizacje procesów analitycznych

Liczba przestępców w zależności od opisu przestępstwa (top 5)



Rysunek 20 Wykres liczby przestępców od opisu przestępstwa

Możemy zobaczyć, ze najczęstszym naruszeniem są narkotyki.

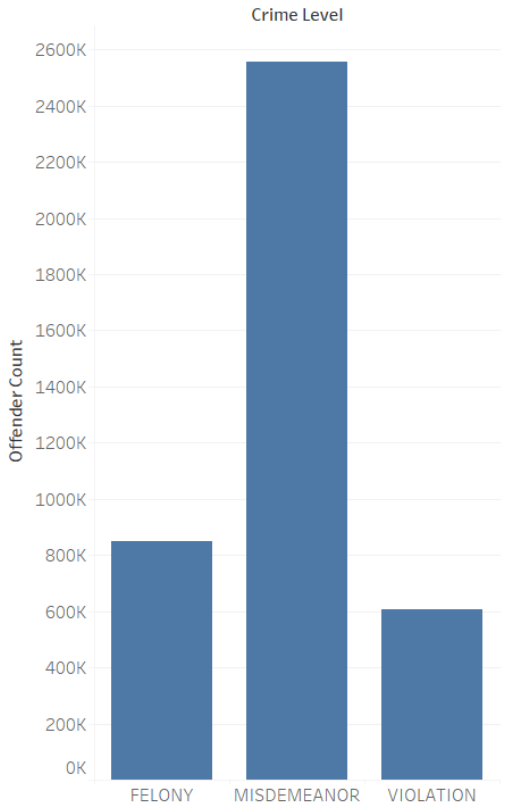
Drugim jest atak trzeciego stopnia, czyli celowy atak i zranienie osoby, i związane przestępstwa.

Na trzecim miejscu molestowanie 2 stopnia, co włącza podążanie za osobą lub denerwowanie ją.

Czwarte miejsce zajmuje niszczenie mienia, a piąte zajmuje znów napad, jednak bardziej ogólna kategoria.

NYC jest miastem o ludności 8 milionów, dlatego mieszkańcy są bardziej skłonny do stresu. Dodatkowo społeczność Nowego Yorku jest składana z różnych grup społecznych. Te czynniki mogą wyjaśnić wymienione przestępstwa.

Liczba przestępców w zależności od poziomu przestępstwa



Rysunek 21 Wykres liczby przestępców w zależności od poziomu przestępstwa

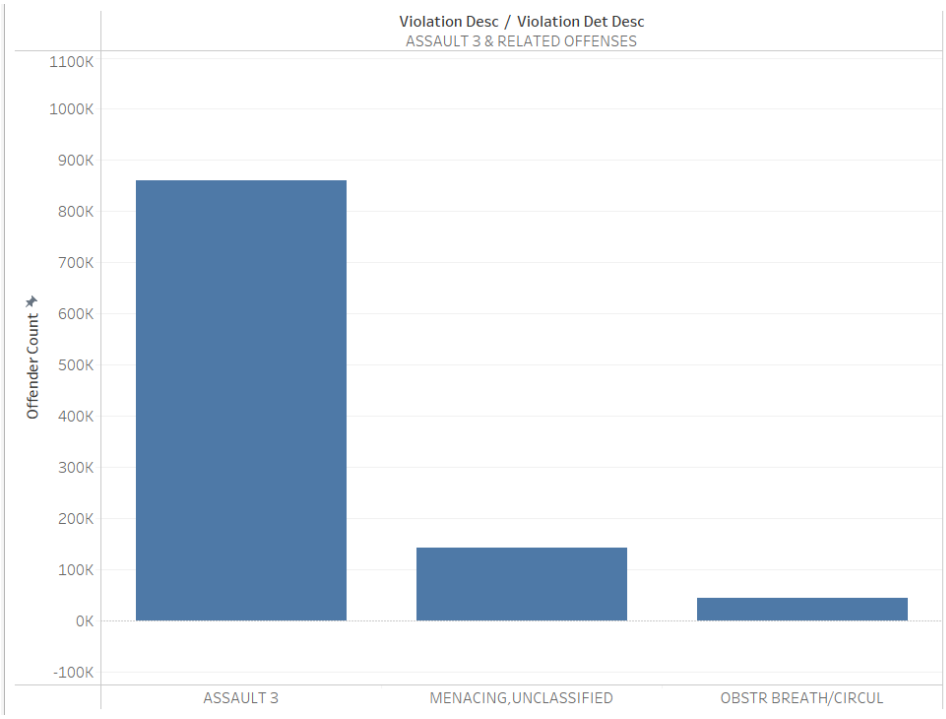
Możemy zobaczyć, ze najczęstszym poziomem przestępstw są średnie, które włączają w sobie posiadanie marihuany i atak i włącza pozbawienie wolności do roku.0

Drugie miejsce to niski poziom, są karane mandatem.

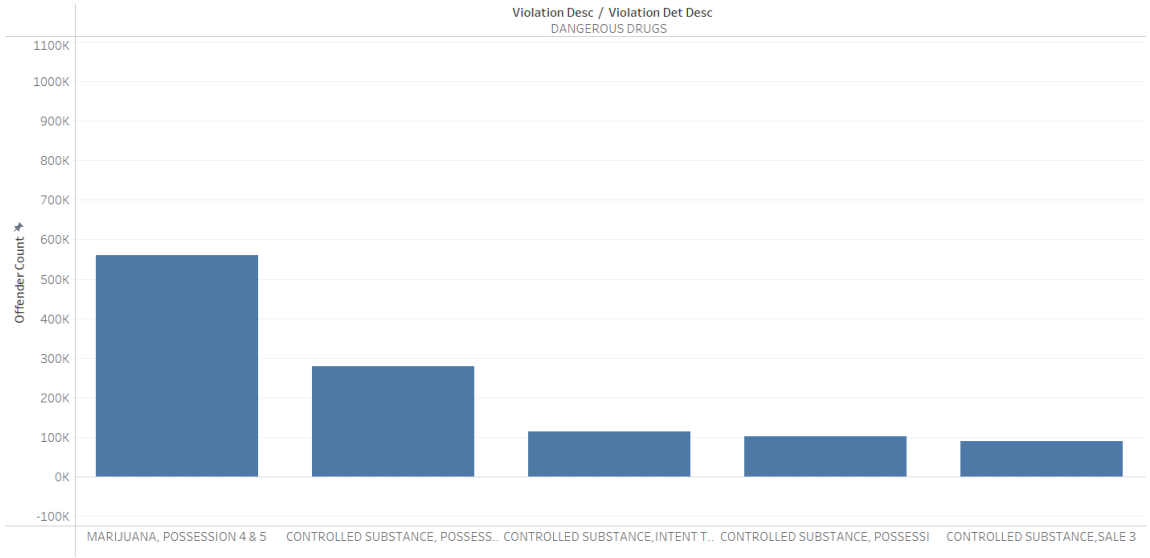
Na ostatnim miejscu są ciężki poziom, jak zabójstwo.

Duża liczba średniego poziomu jest związana z tym, ze włącza w sobie karę za narkotyki, z czym Nowy York ma duży problem od wielu lat [5].

Liczba przestępców w zależności od opisu i szczegółów przestępstwa



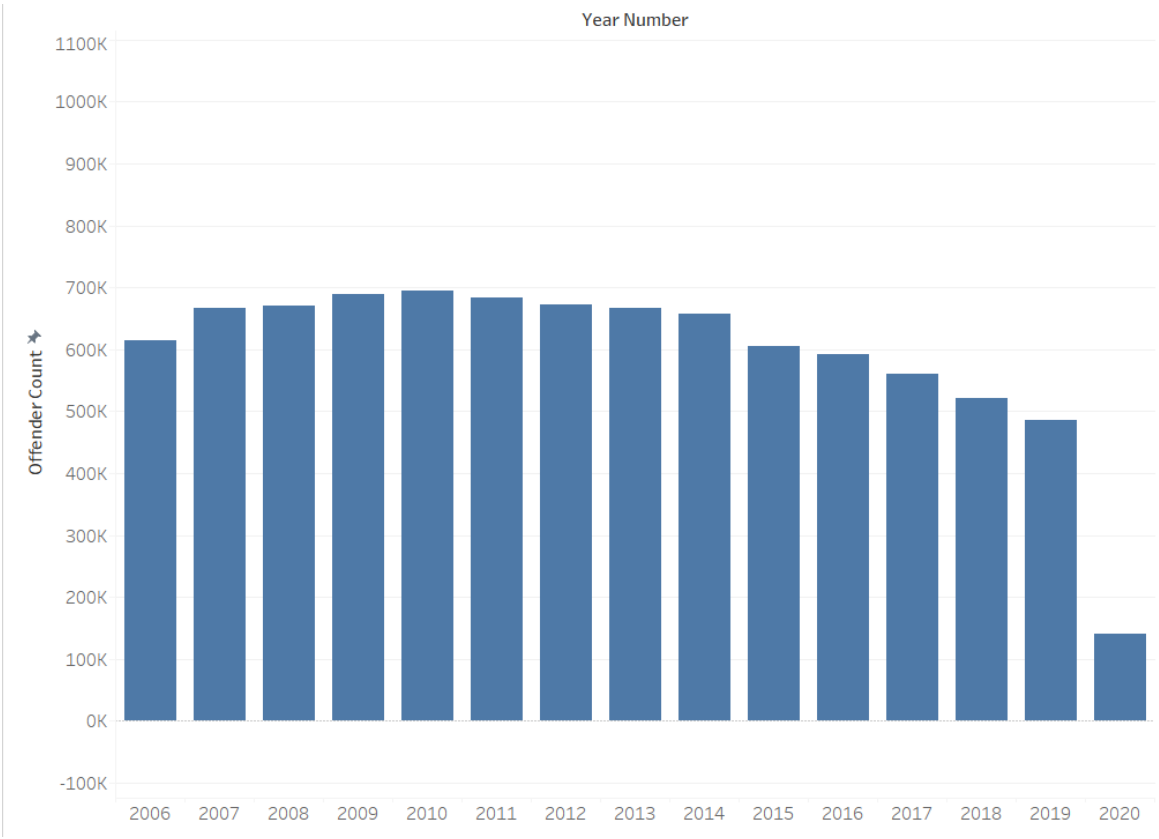
Rysunek 22 Wykres liczby przestępstw w zależności od opisu przestępstwa



Rysunek 23 Wykres liczby przestępstw w zależności od szczegółów przestępstwa

Ciekawe dane możemy uzyskać z szczegółów kategorii naruszeń, związanych z narkotykami. Największa kategoria to posiadanie marihuany. To jest związane z duża popularnością tego narkotyku z powodu odnośnej bezpieczności i, jak wspomniano wcześniej, stresem dużego miasta, wszystkie inne kategorie to posiadanie lub sprzedaż substancji o różnym stopniu ciężkości.

Liczba przestępców w zależności od roku

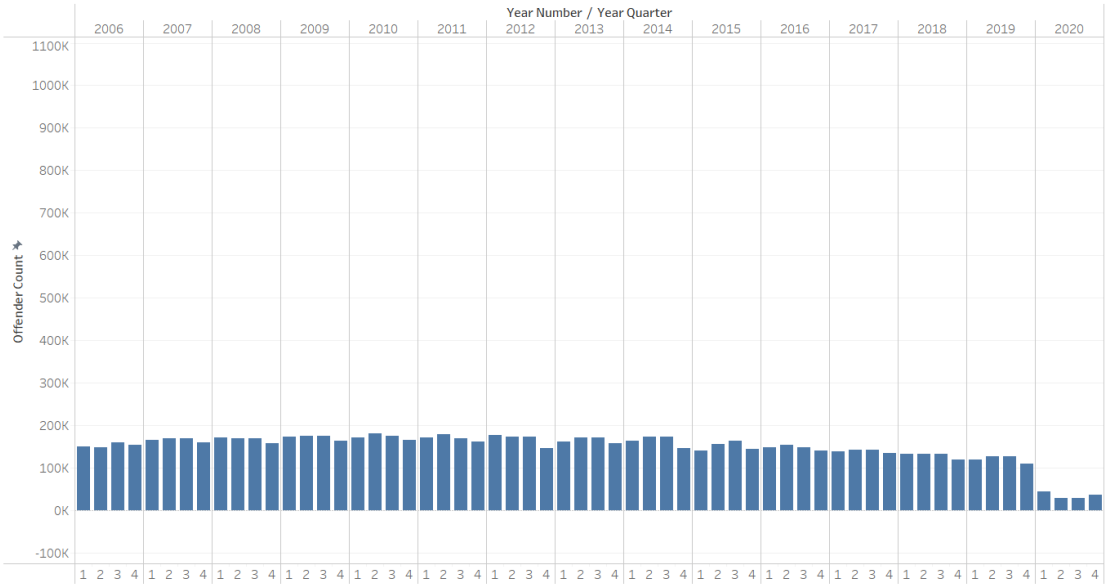


Rysunek 24 Wykres liczby przestępców w zależności od roku

W okresie 2006-2014 liczba złodziejów, tym samym naruszeń, była prawie jednego poziomu, ale od 2010 możemy zobaczyć tendencje do zmniejszenia tej miary.

W 2020 liczba przestępców znacznie spadla, co jest związane z pandemią koronawirusa.

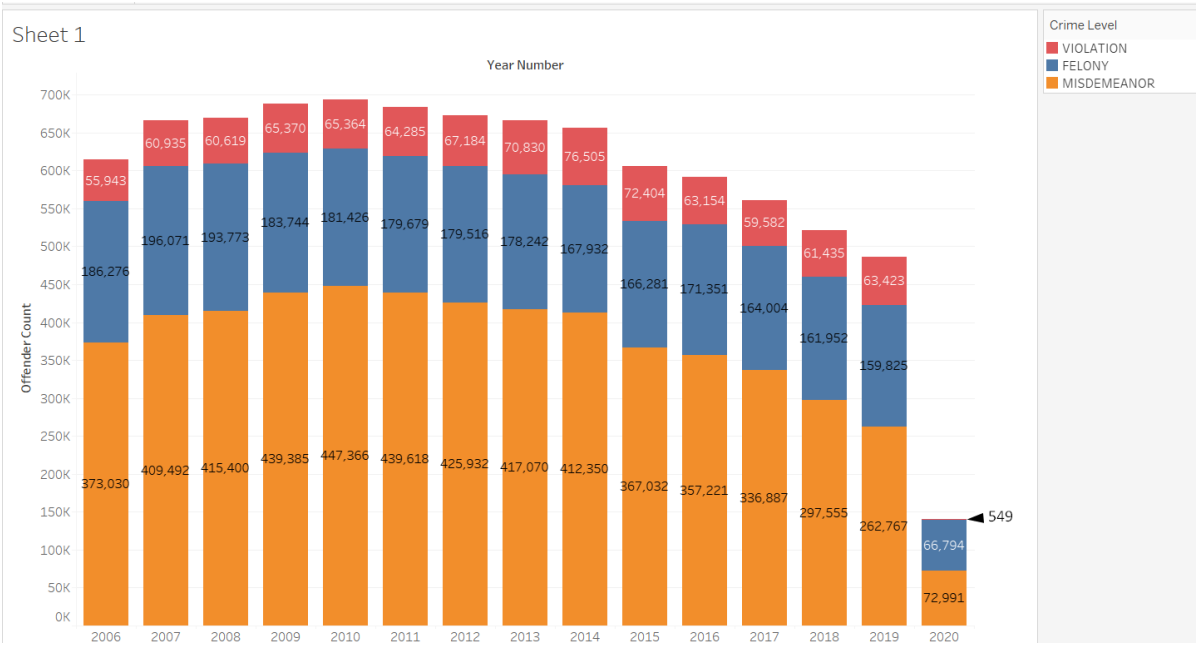
Liczba przestępców w zależności od roku i kwartału



Rysunek 25 Wykres liczby przestępców w zależności od roku i kwartału

Tutaj możemy wywnioskować, ze nie ma znaczących zależności pomiędzy liczbą przestępców i kwartałem roku: liczba przestępców jest równomiernie rozdzielona w ciągu roku.

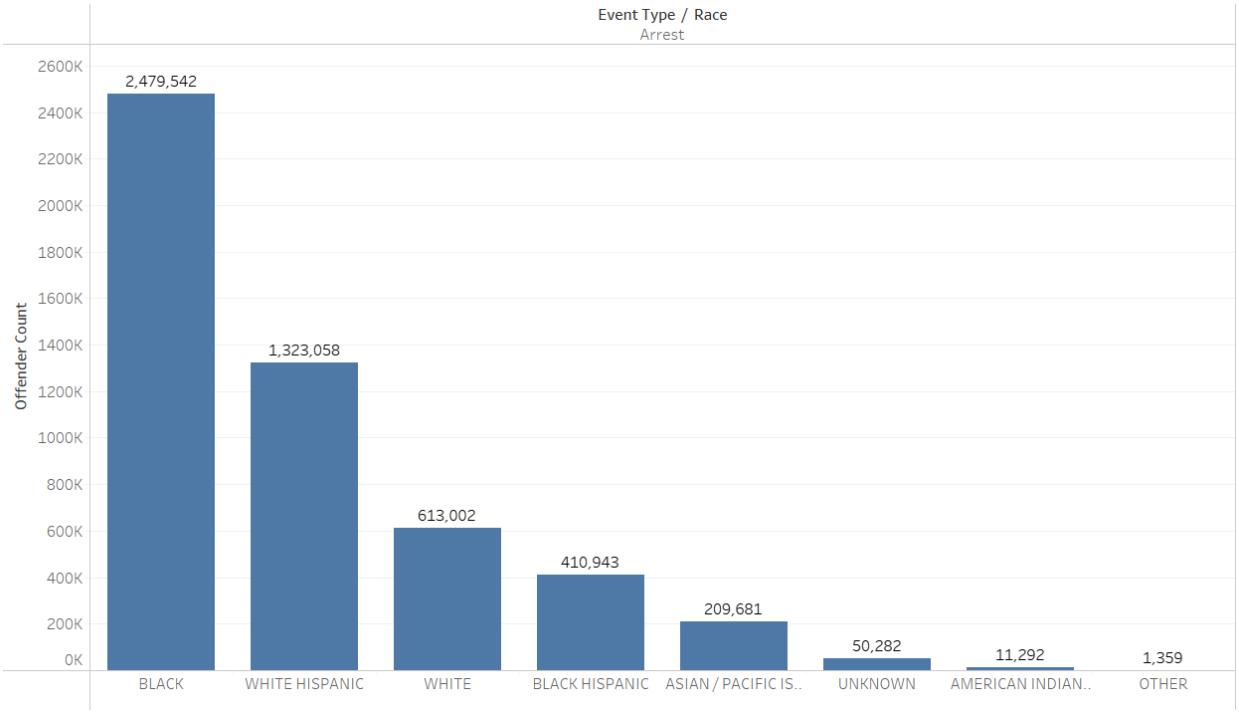
Liczba przestępców w zależności od roku i poziomu przestępstwa



Rysunek 26 Wykres liczby przestępców w zależności od roku i poziomu przestępstwa

Nie widzimy, że istnieje tendencja zmiany proporcji zdarzeń o rożnym poziomu ciężkości w zależności od roku. Zmiana liczby zdarzeń o poszczególnych poziomach jest związana ze zmianami ogólnej liczby naruszeń.

Liczba przestępców według aresztu i rasy przestępcy



Rysunek 27 Wykres liczby przestępców według aresztu i rasy przestępcy

Najczęściej dochodzi do zatrzymań ludzi o rasie czarnej, potem rasy białej i białej latynos, koleinie czarnych latynos, indianów i innych.

Duża liczba przestępstw poprzez ludzi o rasie czarnej jest związana z nierównością dochodów, 63% czarnych są w dolnej połowie poziomy dochodów [1]

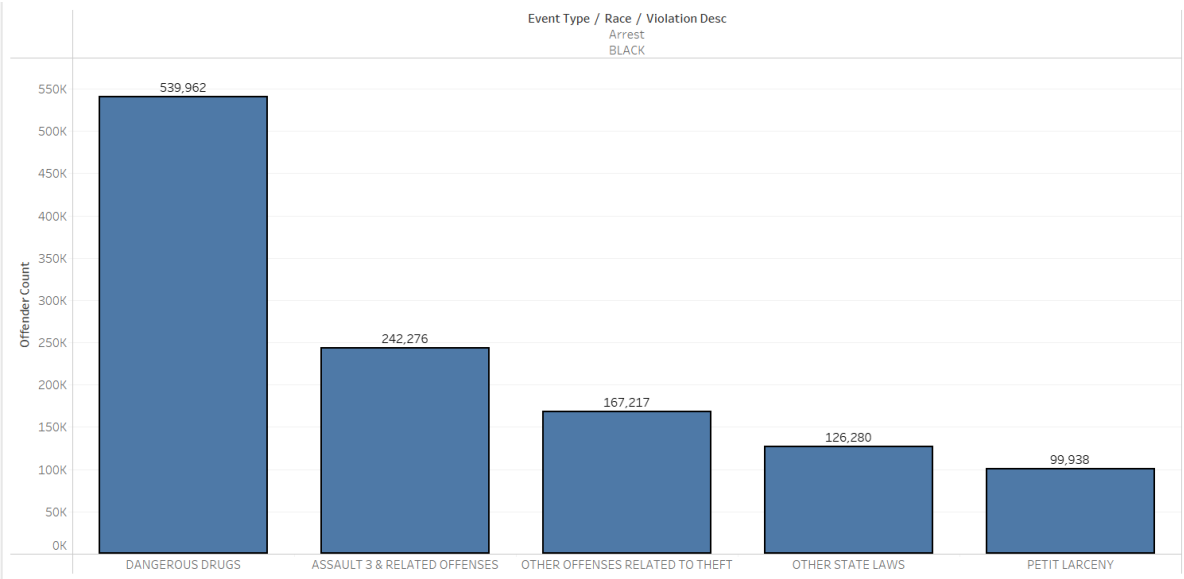
Jednak dodatkowo popatrzymy na demografie Nowego Jorku według rasy [4]:

Tabela 11. Demografia nowego Yorku według rasy

Rasa	% mieszkańców
Biała	41.33
Czarna	23.82
Azjacka	14.29
Inne	20.56

Możemy zobaczyć, ze dane o chętliwości przestępstw i o ilości mieszkańców z co grupy nie korelują pomiędzy sobą. Nie możemy wyjaśnić dużej ilości przestępców o rasie czarnej dużym procentem w liczbie mieszkańców.

Liczba przestępców według aresztu, rasy przestępcy i opisu przestępstwa

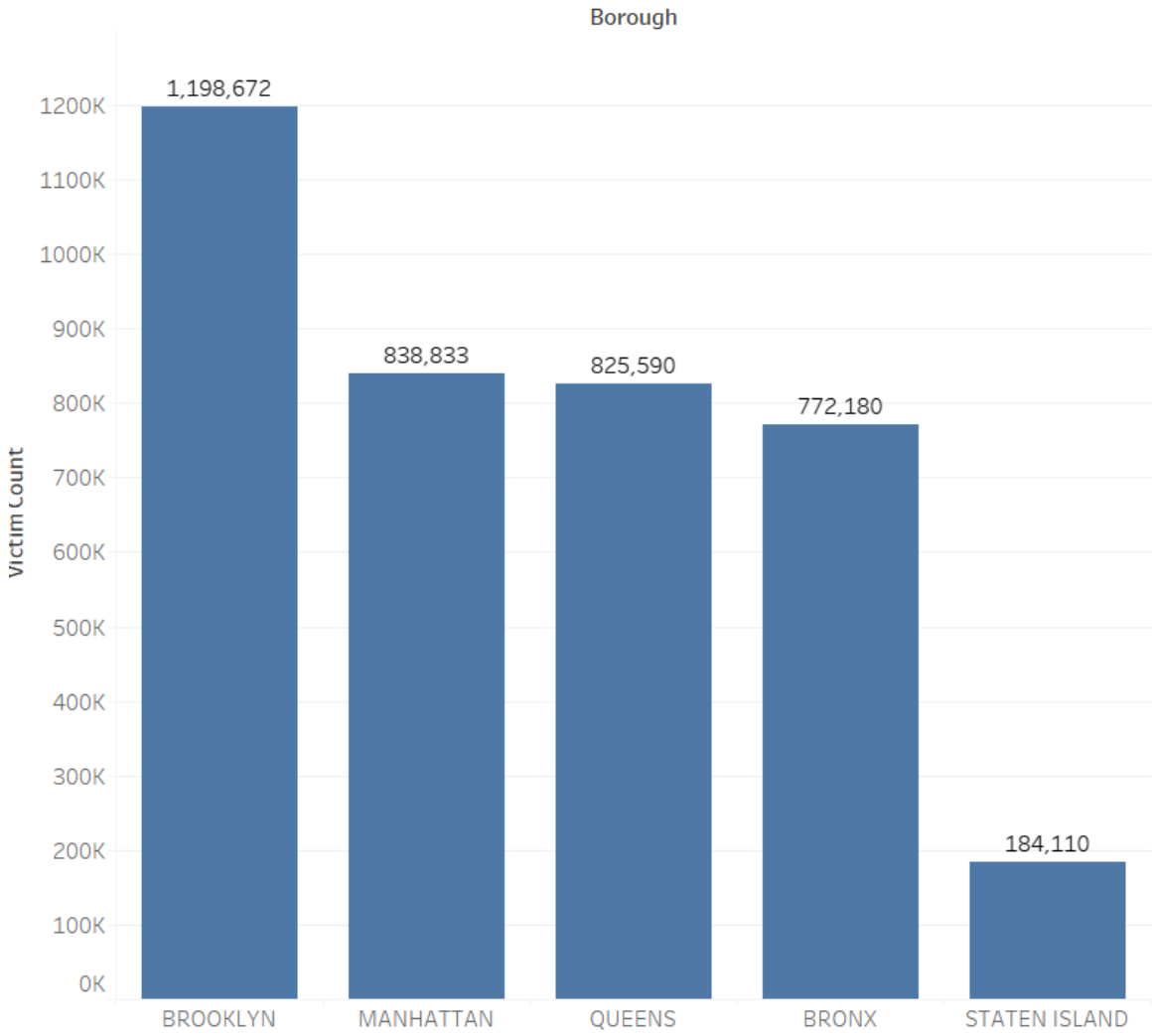


Rysunek 28 Wykres liczby przestępców według aresztu, rasy przestępcy i opisu przestępstwa

Najczęstszym powodem aresztu ludzi o rasie czarnej to narkotyki, potem idą przestępstwa związane z kradzieżą i naruszeni innych praw NYC.

Te przestępstwa można bezpośrednie związać z nierównością dochodów [1].

Liczba ofiar w zależności od okręgu

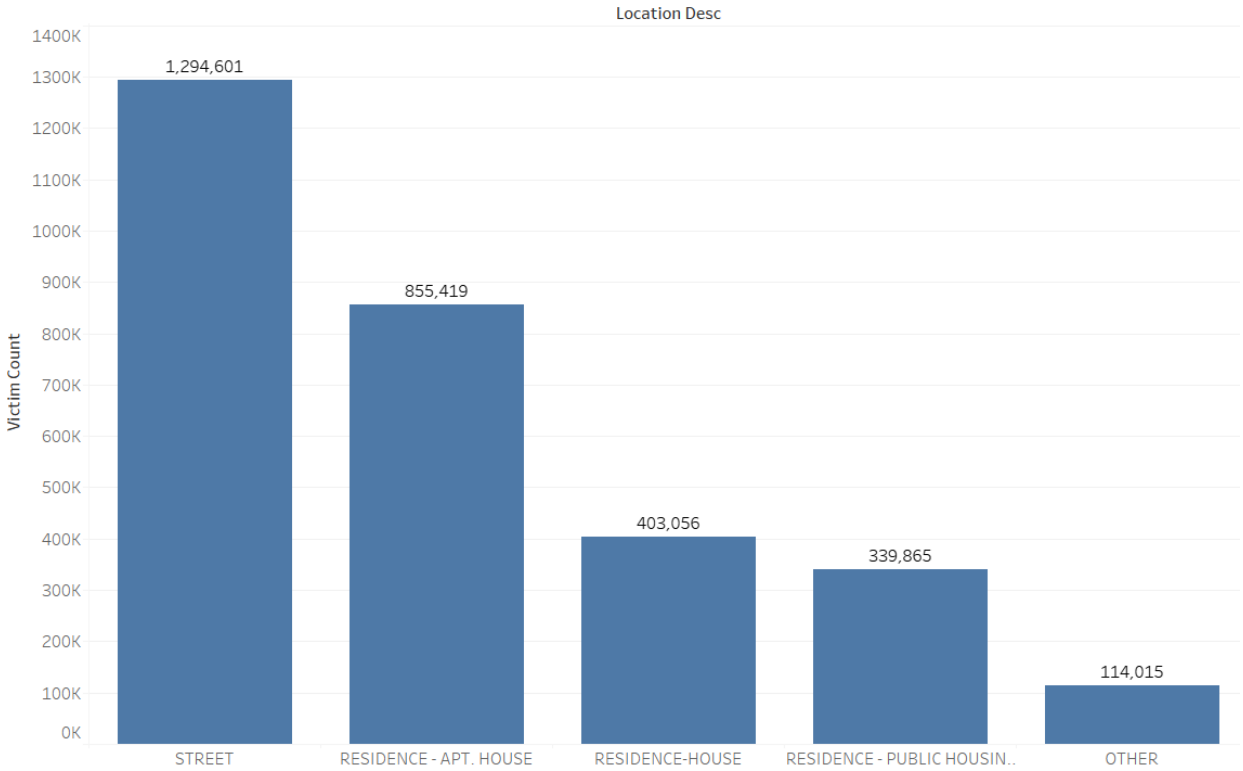


Rysunek 29 Wykres liczby ofiar w zależności od okręgu

Największa liczba ofiar jest w Brooklyn'ie, a najmniej s Staten Island.

Brooklyn ma najwięcej mieszkańców, kiedy Staten Island ma najmniej mieszkańców wśród wszystkich okręgów Nowego Yorku, dlatego możemy zrobić sugestie ze jest to związane z tym powodem [2].

Liczba ofiar w zależności od opisu lokacji

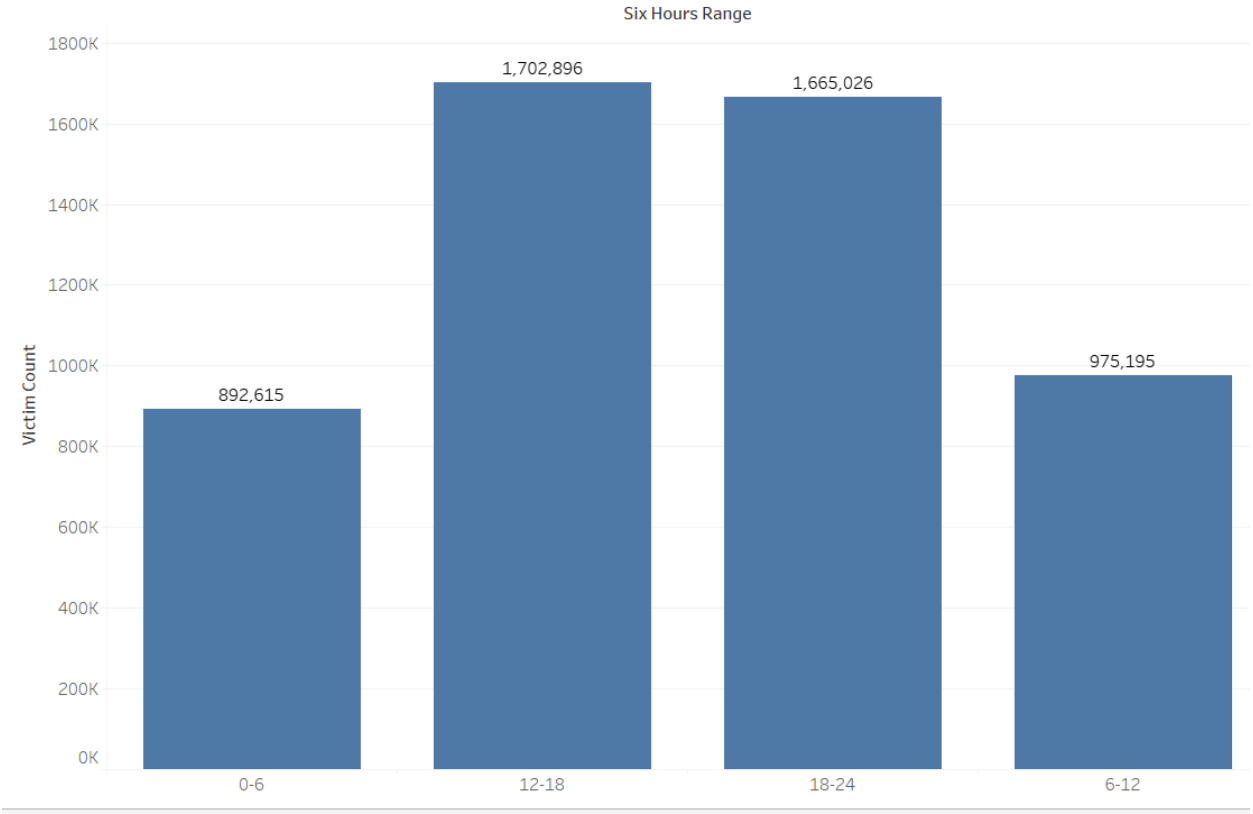


Rysunek 30 Wykres liczby ofiar w zależności od opisu lokacji

Możemy zobaczyć, ze najwięcej ofiar jest na ulicy, na drugim miejscu w apartamentach, potem w domu, mieszkaniach socjalnych i na piątym miejscu inne miejsca.

Duża liczba ofiar na ulicy i apartamentach możemy związać z tym, że na ulicach i domach z apartamentami są dużo ludzi, dlatego w tych miejscach najprawdopodobniej popaść na oczy przestępcy lub pokłócić się z innym człowiekiem.

Liczba ofiar w zależności od przedziału czasowego



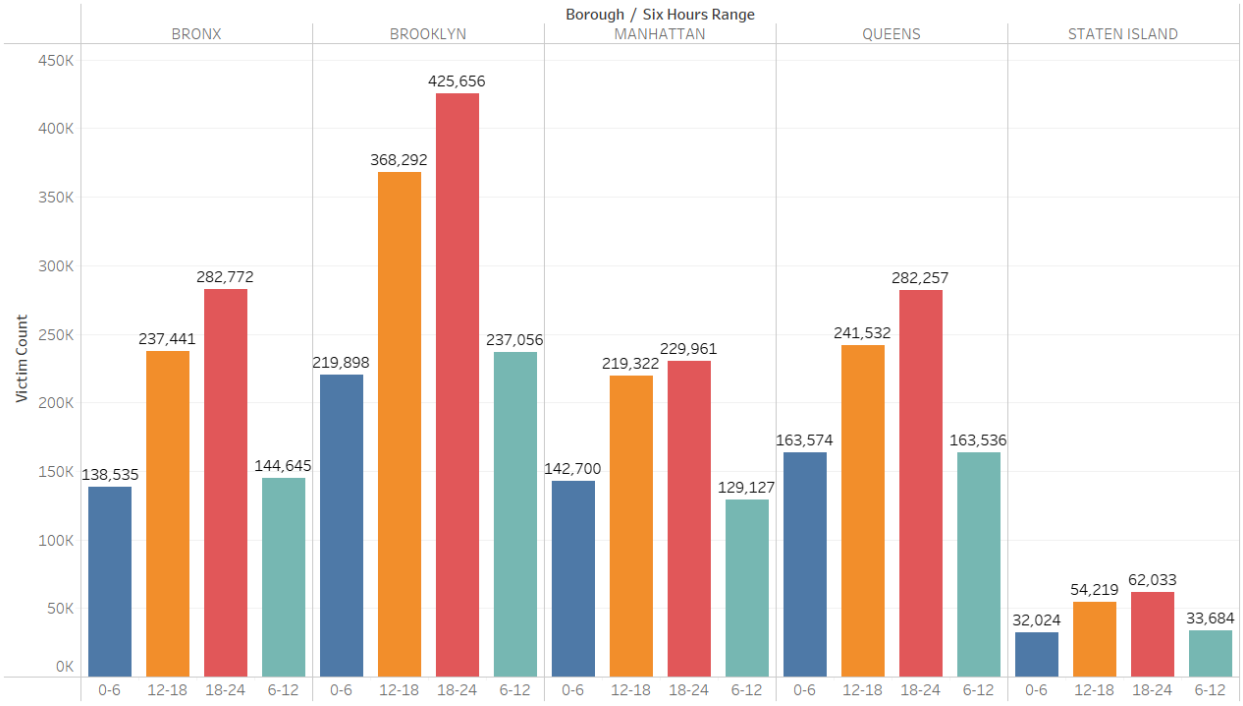
Rysunek 31 Wykres liczby ofiar w zależności od przedziału czasowego

Najwięcej uszkodzonych ofiar jest w przedziałach 12-18 i 18-24.

Godziny 12-18 to czas największej aktywności człowieka i, tym samym, na ulicach.

W godzinach 18-24 ludzie idą z pracy (w USA zwykle pracują od dziewiątej rano do piątej wieczorem) i tez jest dużo ludzi na ulicach.

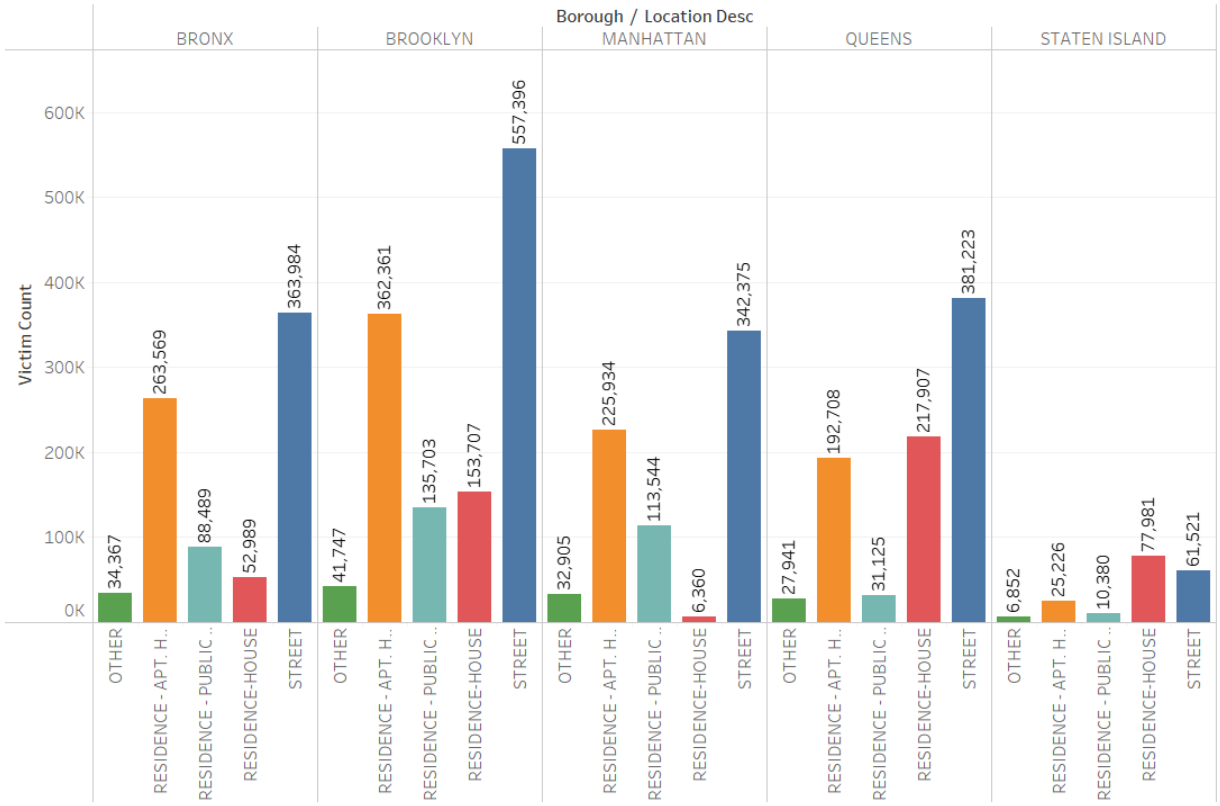
Liczba ofiar od okręgu i przedziału godzinowego



Rysunek 32 Wykres liczby ofiar od okręgu i przedziału godzinowego

Patrząc na wykres możemy stwierdzić, ze odnośnie co okręgu ranking przedziałów czasowych według liczby ofiar jest taka sama. Co okręg ma takie same najniebezpieczniejsze godziny.

Liczba ofiar od okręgu i opisu lokacji



Rysunek 33 Wykres liczby ofiar od okręgu i opisu lokacji

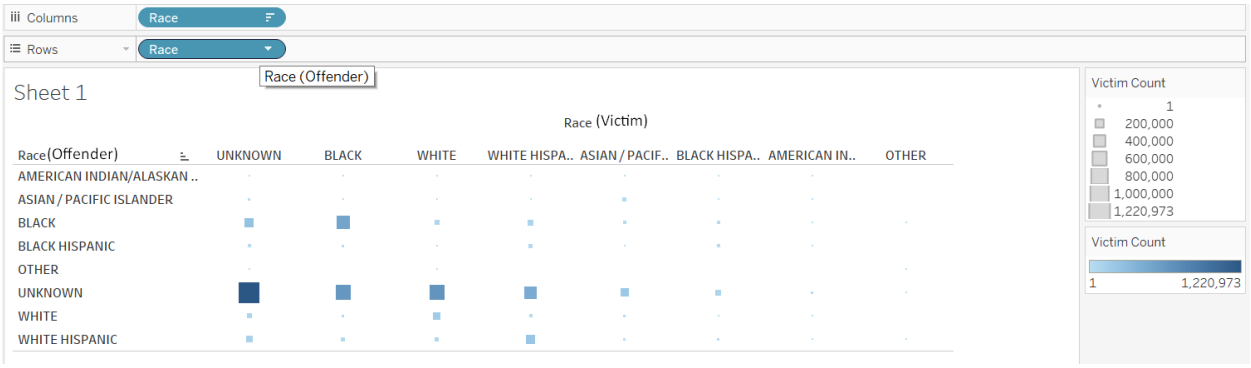
Patrząc na wykres możemy stwierdzić, ze nie ma jednej tendencji miejsc przestępstw w co okręgu.

Dla wszystkich okręgów na pierwszym miejscu jest ulica, oprócz Staten Island, gdzie większość przestępstw są w domach, kiedy ulica jest na drugim miejscu. Może być to związane z tym, ze ten okręg jest bardziej spokojny i nie ma dużo miejsc do kadi da dużo ludzi, co jest pośrednio potwierdzone tym, ze Staten Island jest najmniej głośnym okręgiem w Nowym Jorku [6]

W pozostałych na drugim miejscu są apartamenty, oprócz Queens, gdzie na tym miejscu jest dom, i Staten Island z ulicą.

Pozostałe tendencje można wywnioskować w wykresu.

Liczba ofiar w zależności od rasy ofiary i rasy przestępcy

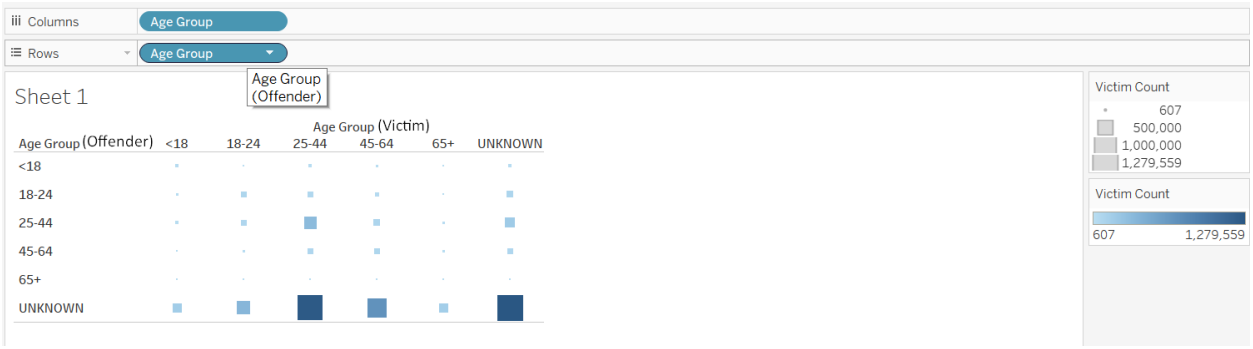


Rysunek 34 Wykres liczby ofiar w zależności od rasy ofiary i rasy przestępcy

Największa liczba przestępstw są pomiędzy niewiadomymi ofiarami i przestępcami. Dla każdej rasy ofiary możemy zobaczyć, że najczęściej przestępca jest niewiadomy, a na drugim miejscu znajduje się ta sama rasa.

Dlatego nie można powiedzieć ze jakaś wymieniona na tym obrazie grupa społeczna celowo szkodzi innej.

Liczba ofiar w zależności od grupy wiekowej ofiary i grupy wiekowej przestępcy



Rysunek 35 Wykres liczby ofiar w zależności od grupy wiekowej ofiary i grupy wiekowej przestępcy

Na tym wykresie możemy zobaczyć, ze dla wszystkich grup wiekowych najczęściej przestępca jest o wiadomym wieku są ludzie o 25-44 lat, dla ofiar o 25-44 lat to widać najwyraźniej.

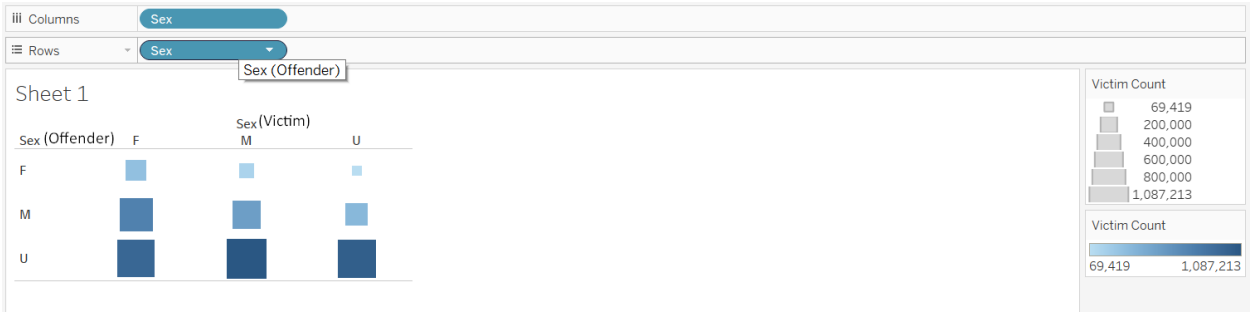
Patrząc na demografie Nowego Jorku[3] możemy ocenić, ile ludzi należą do co kategorii

Tabela 12. Demografia nowego Yorku według kategorii wiekowej

Grupa wiekowa	% mieszkańców
0-18	27.4%
18-24	6.6%
25-44	30.7%
45-65	22.4%
65	10.8%

Grupa wiekowa 25-44 jest większością mieszkańców, co wyjaśnia dlaczego oni najczęściej są ofiarami i przestępcami. Chociaż 0-18 i 45-65 maja bliski procent mieszkańców, w tych kategoriach znaczna cześć ludzi, które z powodu charakterystyk lub problem fizycznych nie mogą robić przestępstwa.

Liczba ofiar w zależności od płci ofiary i płci przestępcy



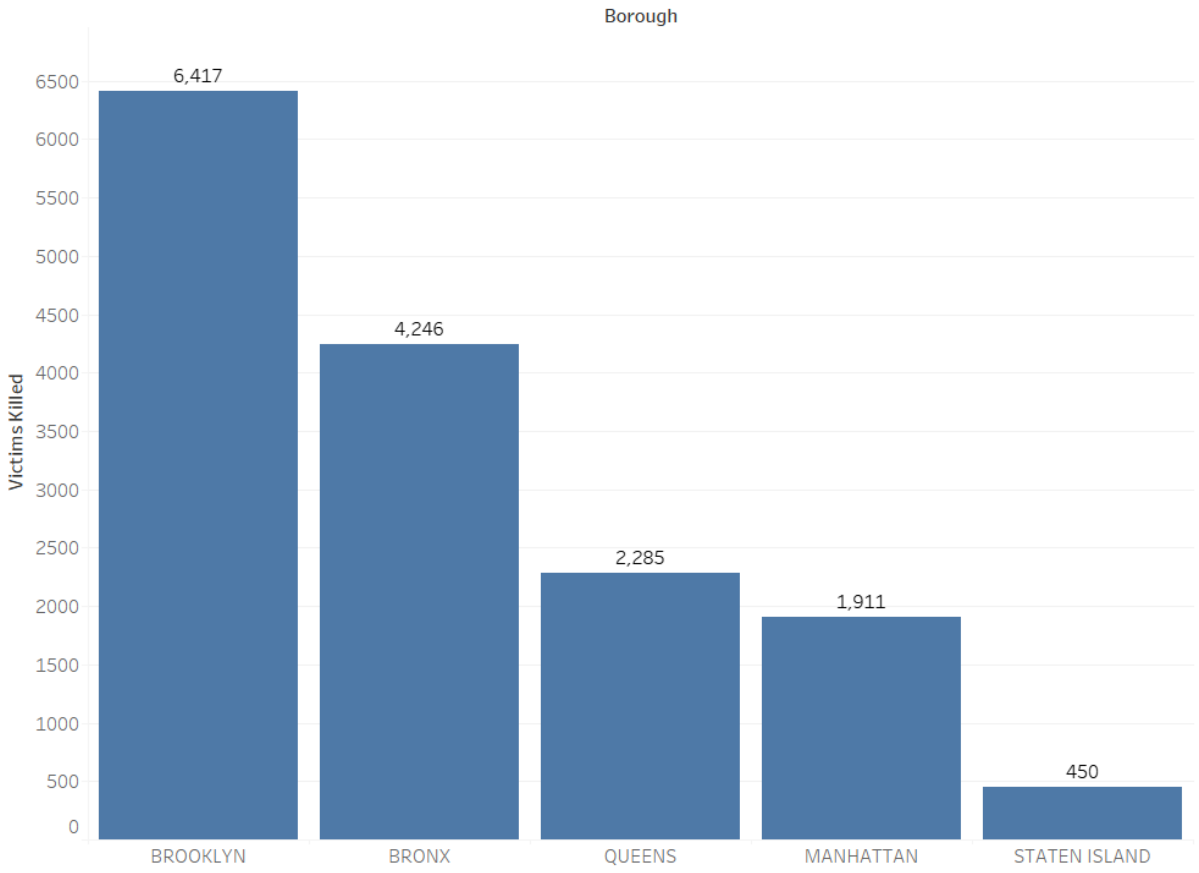
Rysunek 36 Wykres liczby ofiar w zależności od płci ofiary i płci przestępcy

W większości przypadkach przestępca nie jest znany, ale na drugim miejscu są mężczyźni. Możemy to związać z ogólna przewaga fizycznej płci męskiej i oczekiwaniami socjalnymi od mężczyzn.

Przestępca mężczyzna najczęściej robi ofiarę z kobiet, możemy to tez powiązać z ogólną przewagą fizyczne płci męskiej i problemami z seksizmem.

Przestępca kobieta i przestępca mężczyzna najczęściej napadają na kobiet..

Liczba zabitych ofiar w zależności od okręgu



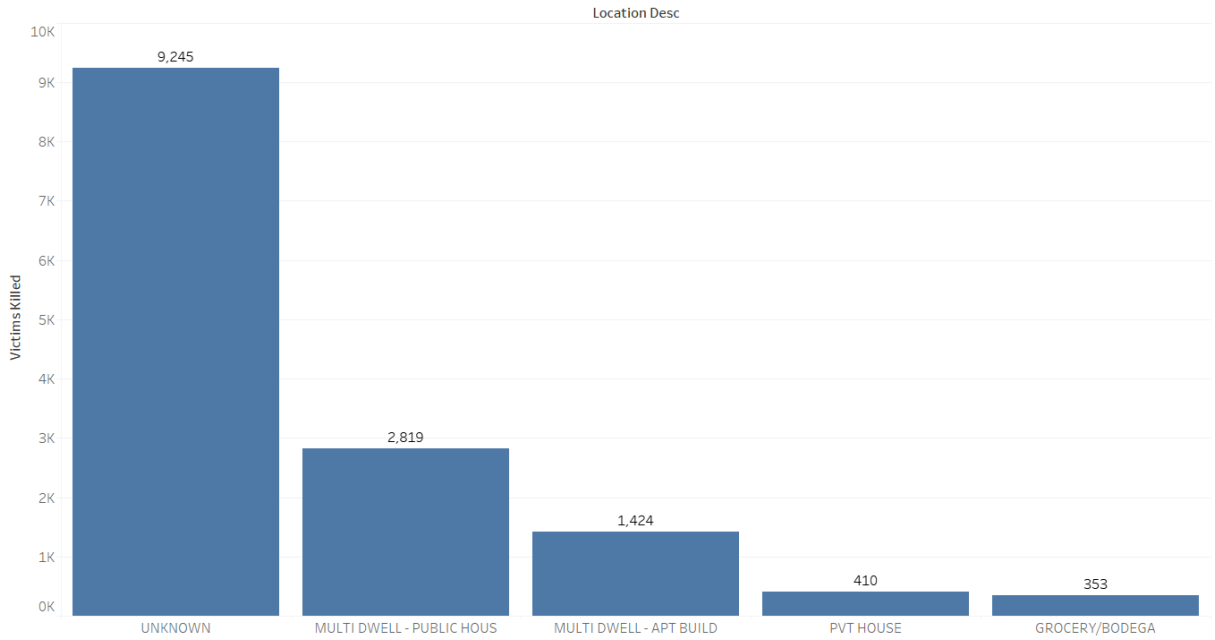
Rysunek 37 Wykres liczby zabitych ofiar w zależności od okręgu

Możemy zobaczyć, ze najwięcej zabójstw jest w Brooklinie i Bronksie, najmniej w Staten Island.

Porównywając z analizą przestępców w zależności od okręgu zobaczymy, że te dwa wykresy nie korelują pomiędzy sobą. Brooklyn ma najwięcej przestępstw i zabójstw, ale Manhattan, który jest drugi z góry według liczby przestępców, jest drugi z dołu według liczby zabitych ofiar. Analogiczna sytuacja jest i z Bronx'em, ale na odwrót.

Z tego wnioskujemy ze liczba przestępców i liczba zabójstw nie są powiązane pomiędzy sobą.

Liczba zabitych ofiar w zależności od opisu lokacji

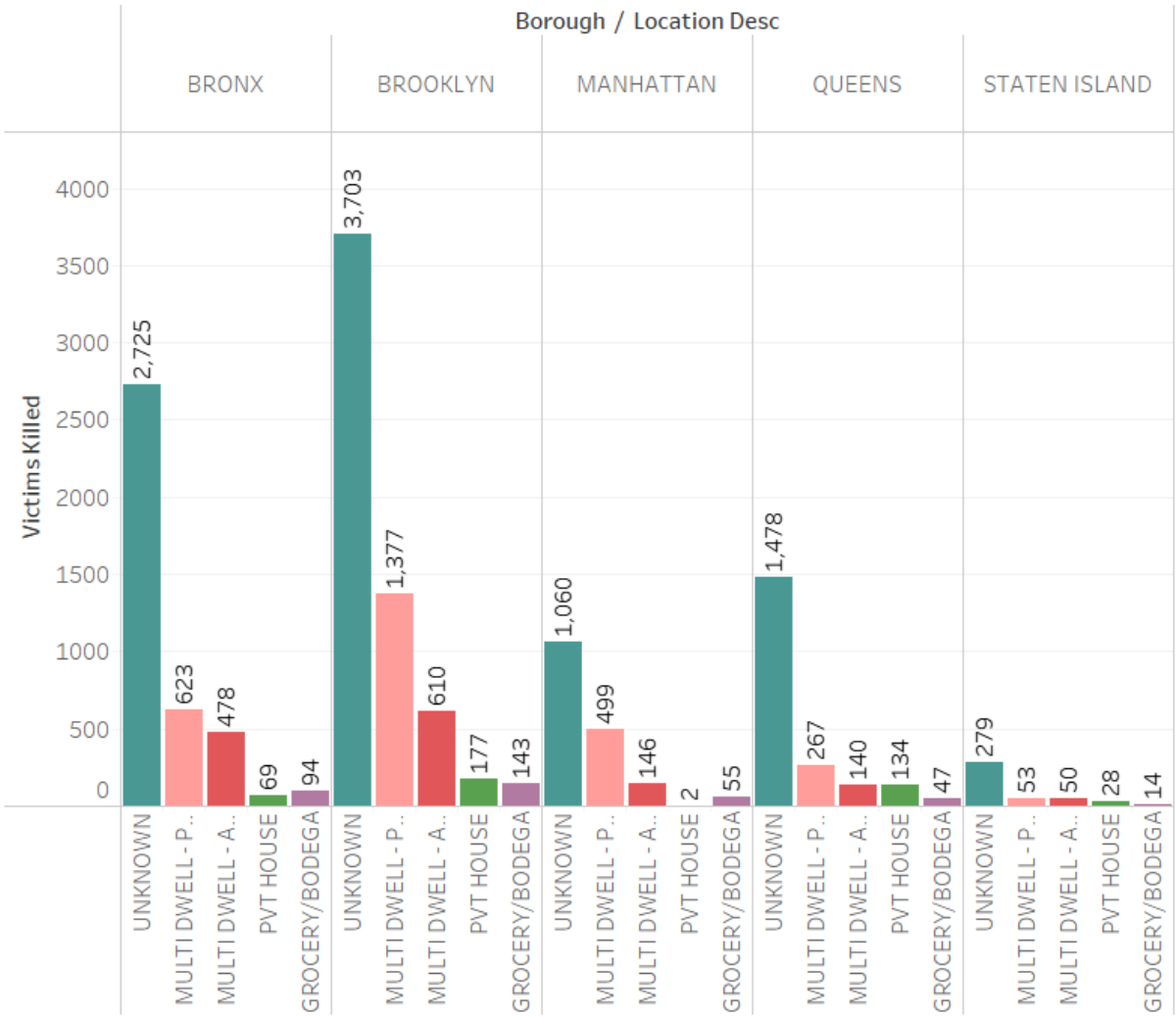


Rysunek 38 Wykres liczby zabitych ofiar w zależności od opisu lokacji

W większości przypadkach lokacja nie jest wyznaczona, na drugim miejscu są socjalne domy dla kilka rodzin, na trzecim domy z apartamentami, na czwartym miejscu są zwykle domy , a na ostatnim są sklepy spożywcze.

Chętliwość śmierci w domach może być związana ze włamaniami do nich lub z powodu konfliktów pomiędzy mieszkańcami.

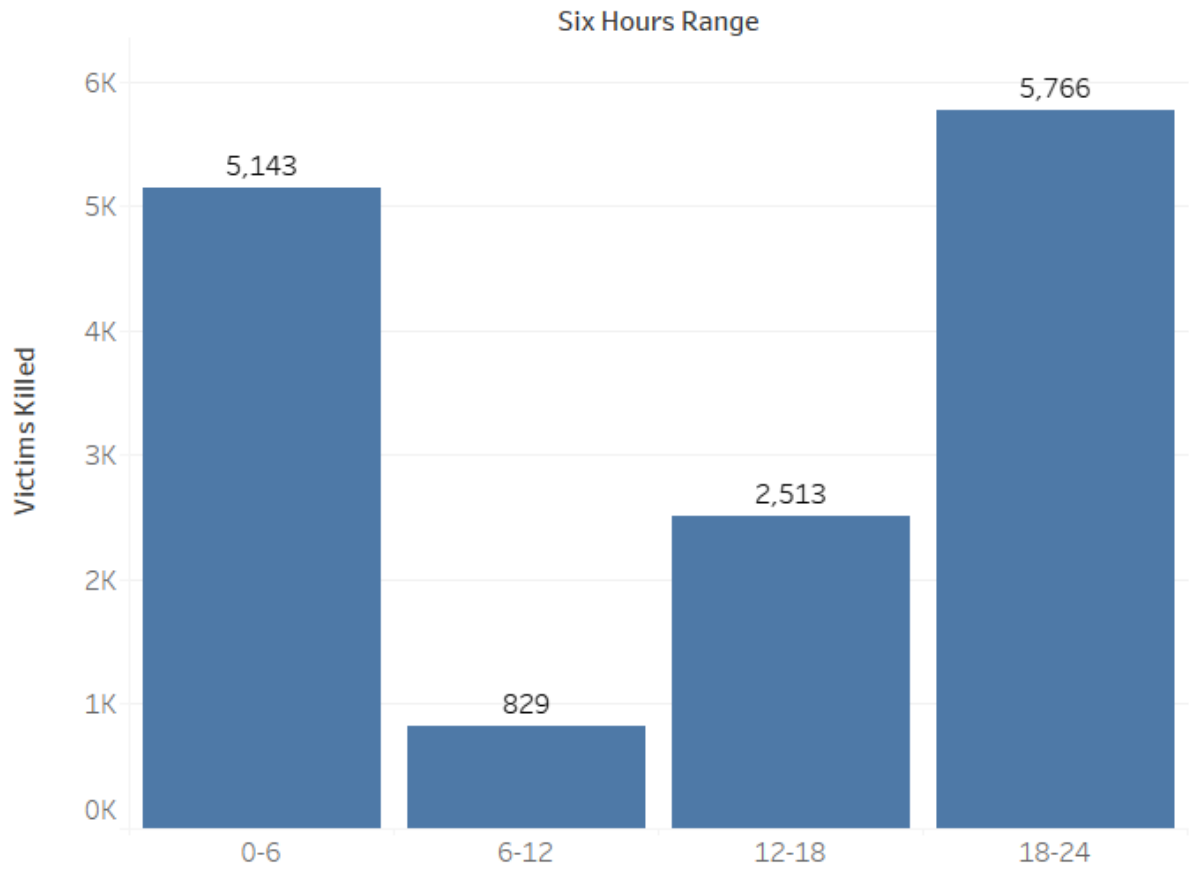
Liczba zabitych ofiar w zależności od okręgu i opisu lokacji



Rysunek 39 Wykres liczby zabitych ofiar w zależności od okręgu i opisu lokacji

Możemy zobaczyć, ze tendencje miejsc, gdzie doszło do zabójstwa w co okręgu jest prawie taka sama.

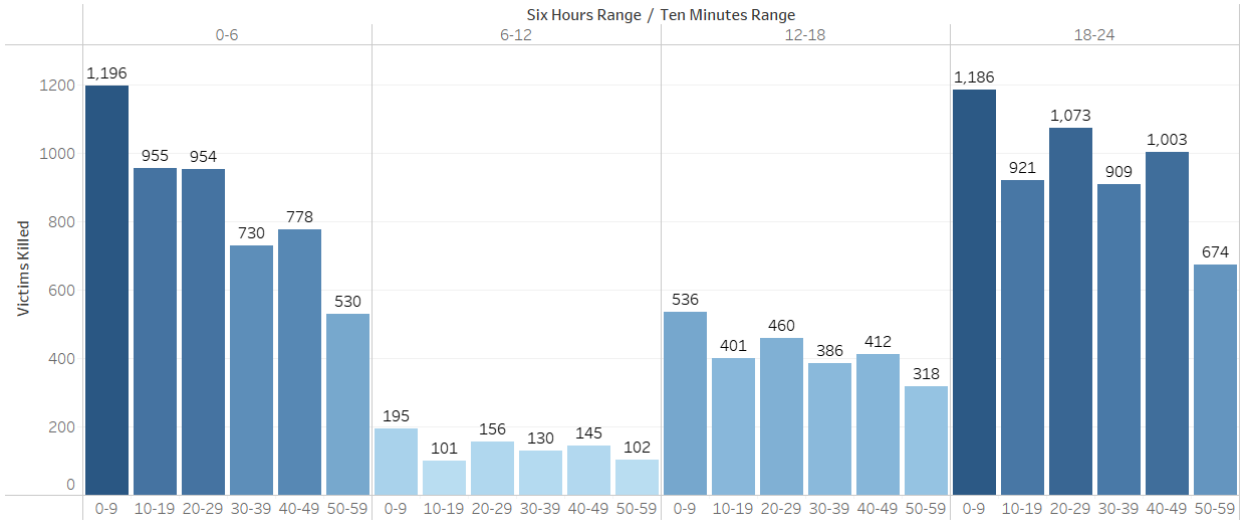
Liczba zabitych ofiar w zależności od przedziału godzinowego



Rysunek 40 Wykres liczby zabitych ofiar w zależności od przedziału godzinowego

Najczęstszej zabójstwa bieżą miejsce po pracy lub w głębokiej nocy, co możemy spróbować wyjaśnić tym, ze te godziny są bardziej wygodne dla kradzieży, włamania do domu lub chowaniu broni, bo w tych godzinach liczba ludzi na ulicach się zmniejsza.

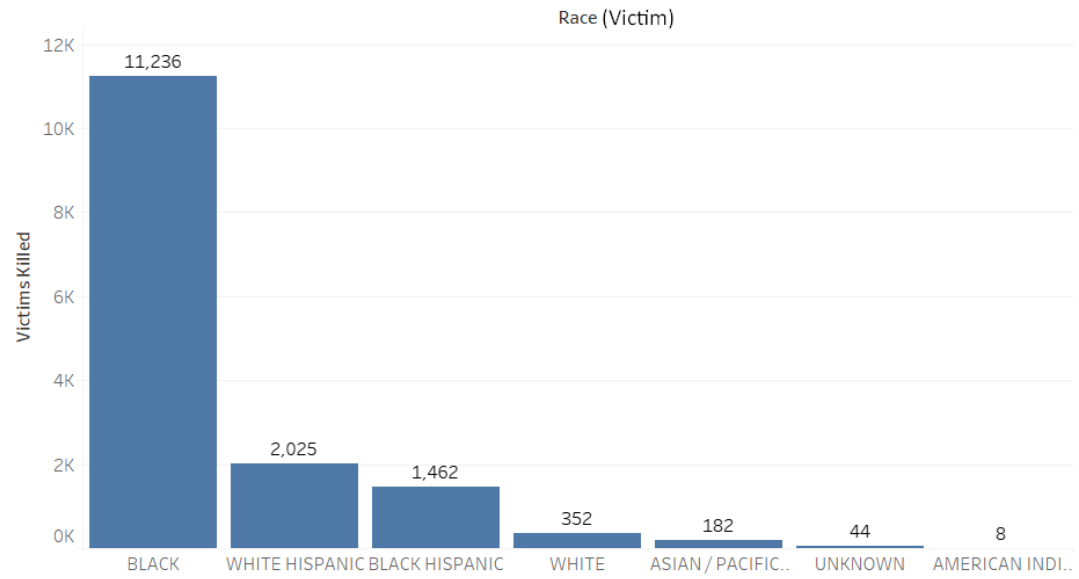
Liczba zabitych ofiar w zależności od przedziału godzinowego i minutowego



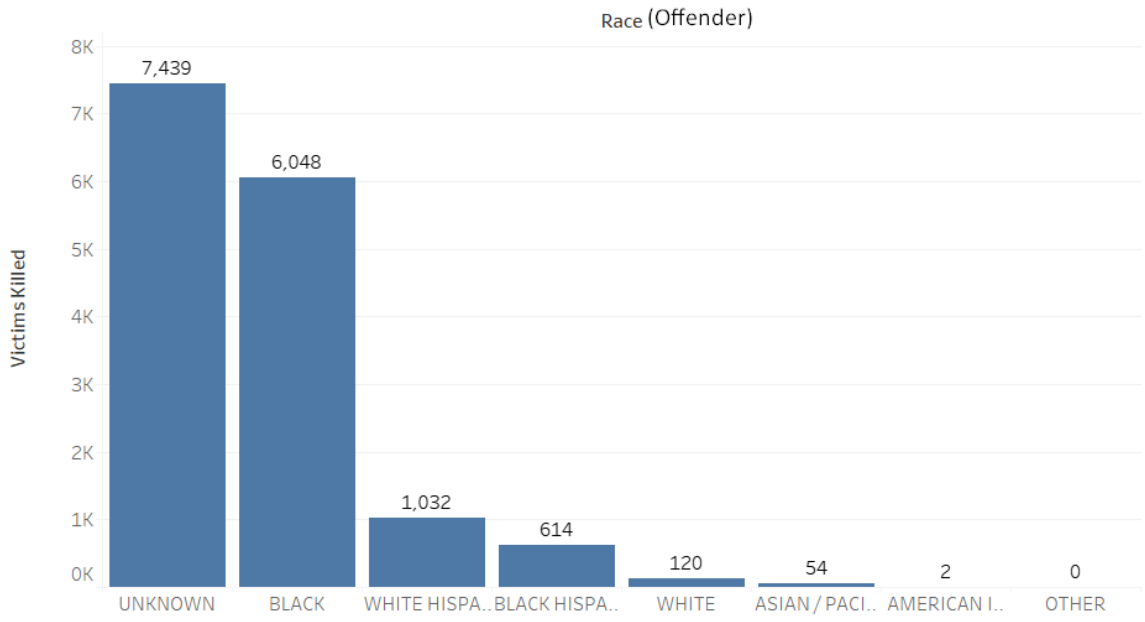
Rysunek 41 Wykres liczby zabitych ofiar w zależności od przedziału godzinowego i minutowego

Możemy zobaczyć, ze najczęściej zabójstwa zdarzają się w początku godziny, szczególnie w okresie „0-6”, co może wskazać na planowanie akcji.

Liczba zabitych ofiar w zależności od rasy ofiary i rasy przestępcy (osobnie)



Rysunek 42 Wykres liczby zabitych ofiar w zależności od rasy ofiary



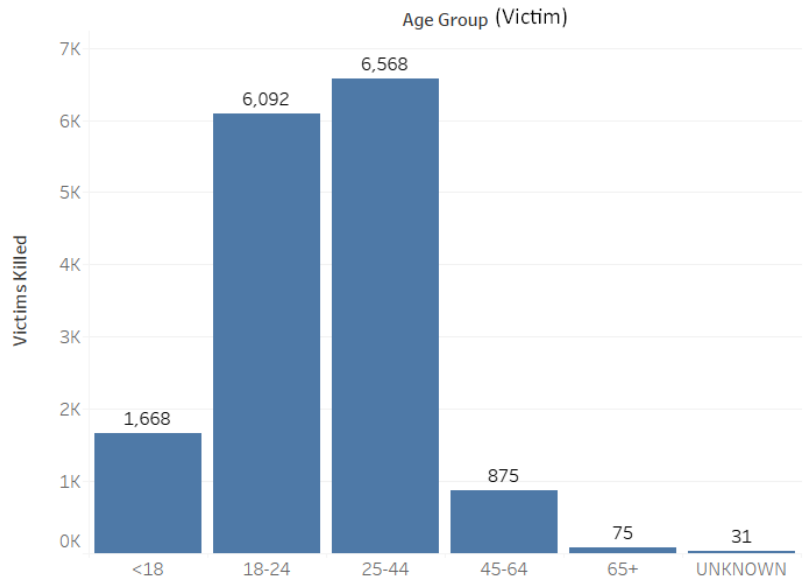
Rysunek 43 Wykres liczby zabitych ofiar w zależności od rasy przestępcy

Największa kategoria ofiar to rasa czarna.

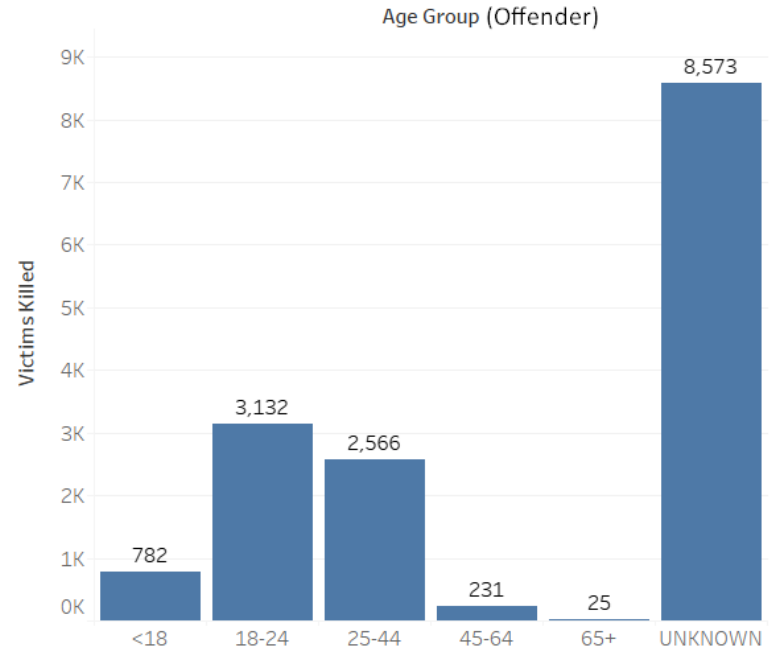
Największa kategoria przestępców są „nie wiadome”, na drugim miejscu rasa czarna.

Te dwa wykresy, bez uwzględnienia niewiadomych przestępców, korelują pomiędzy sobą.

Liczba zabitych ofiar w zależności od kategorii wiekowej ofiary i kategorii wiekowej przestępcy (osobnie)



Rysunek 44 Wykres liczby zabitych ofiar w zależności od kategorii wiekowej ofiary

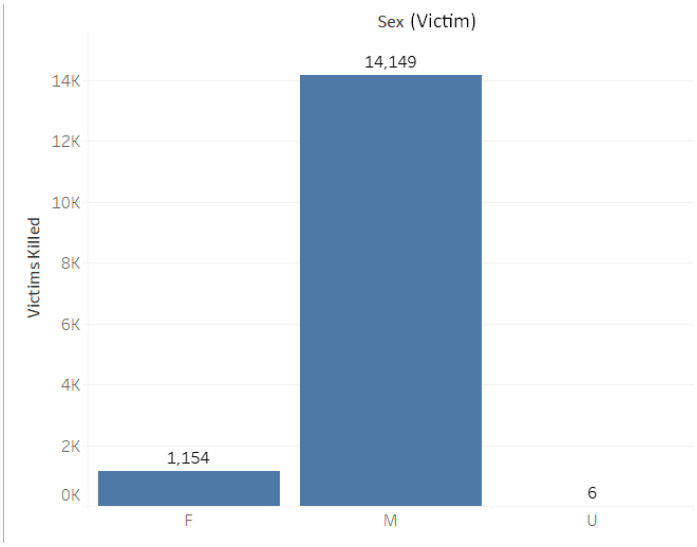


Rysunek 45 Wykres liczby zabitych ofiar w zależności od kategorii wiekowej przestępcy

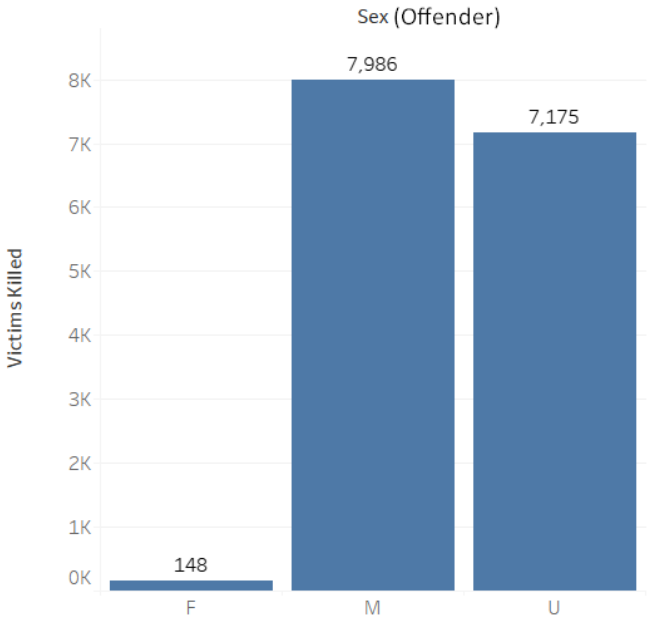
Najczęściej ofiarami zabójstw to ludzie o wieku 18-24 i 25-44.

Napastnik najczęściej nie jest znany, ale na drugim miejscu są ludzie o wieku 18-24

Liczba zabitych ofiar w zależności od płci ofiary i płci przestępcy (osobnie)



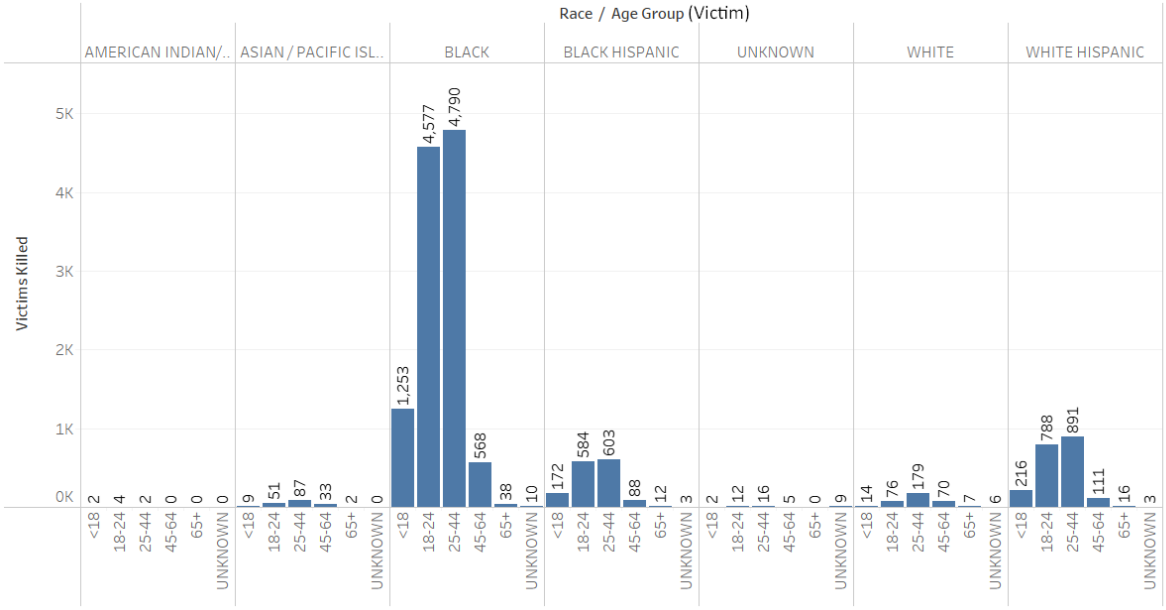
Rysunek 46 Wykres liczby zabitych ofiar w zależności od płci ofiary



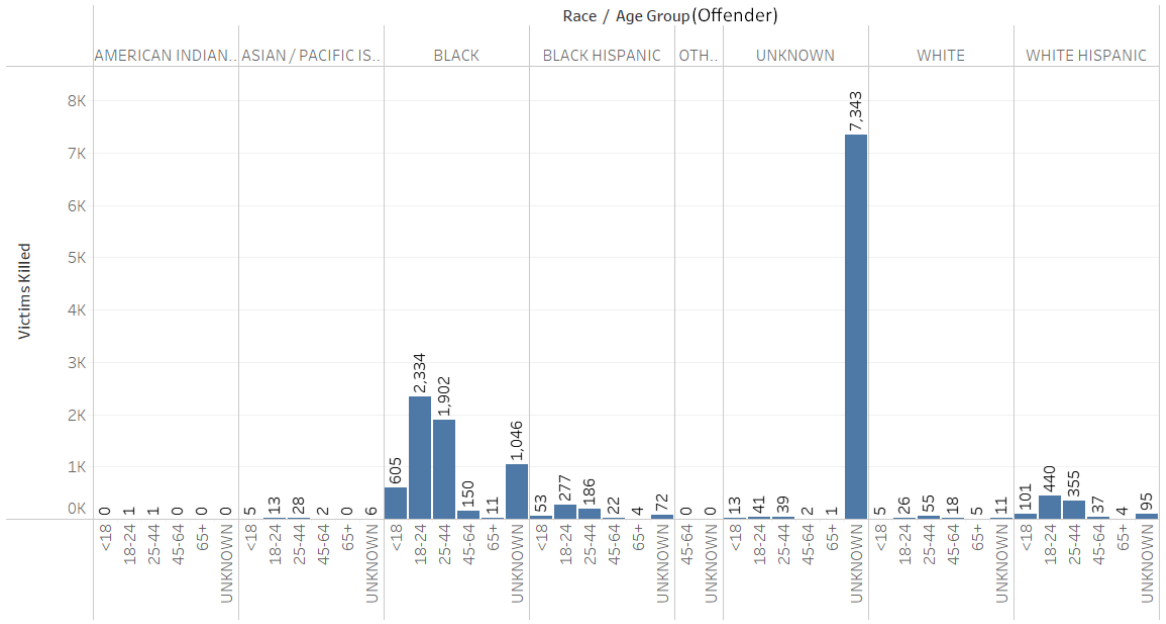
Rysunek 47 Wykres liczby zabitych ofiar w zależności od płci przestępcy

Najczęstszymi ofiarami i napastnikami są mężczyźni. Jednak duża cześć napastników jest niewiadoma. Te dane nie korelują z danymi analizy korelacji płci dla liczby ofiar.

Liczba zabitych ofiar w zależności od rasy i kategorii wiekowej ofiary i rasy i kategorii wiekowej przestępcy (osobnie)



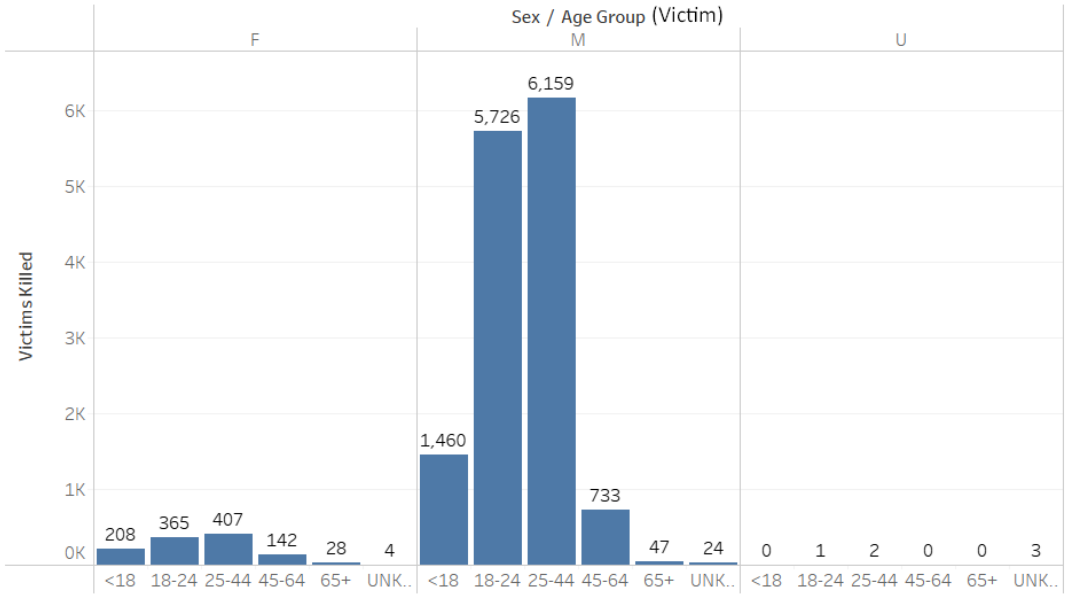
Rysunek 48 Wykres liczby zabitych ofiar w zależności od rasy i kategorii wiekowej ofiary



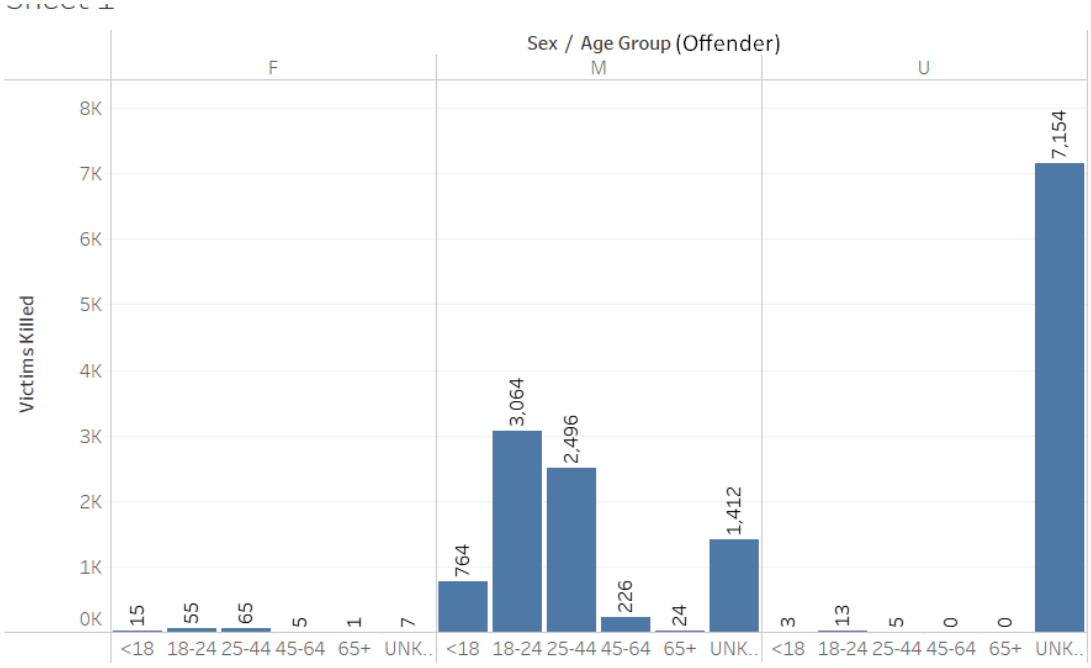
Rysunek 49 Wykres liczby zabitych ofiar w zależności od rasy i kategorii wiekowej przestępcy

Ogólnie tendencje ofiar i napastników według kategorii wiekowych są prawdziwe dla co rasy. Możemy tylko powiedzieć, ze grupa wiekowa jest najczęściej nieznana kiedy i rasa jest nieznana.

Liczba zabitych ofiar w zależności od płci i kategorii wiekowej ofiary i płci i kategorii wiekowej przestępcy (osobnie)



Rysunek 50 Wykres liczby zabitych ofiar w zależności od płci i kategorii wiekowej ofiary



Rysunek 51 Wykres liczby zabitych ofiar w zależności od płci i kategorii wiekowej przestępcy

Ogólnie tendencje ofiar i napastników według kategorii wiekowych są prawdziwe dla każdej płci. Możemy tylko powiedzieć, ze grupa wiekowa jest najczęściej nieznana kiedy i pleć jest nieznana.

7.2. Podsumowania, wnioski z analizy

Analiza danych miała pomóc w znalezieniu odpowiedzi na następujące pytania:

- Jakie są najczęstsze naruszenia, ogólnie i szczegółowo?

  Najczęstszymi typami naruszeń są: narkotyki, napad (głownie 3 stopnia), molestowanie i niszczenia mienia.

  Wewnątrz kategorii narusz, związanych z narkotykami, najpopularniejszym naruszeniem jest posiadanie marihuany.

  Najczęściej występują naruszenia średniego poziom.

- Jakie są tendencje wystąpienia zdarzeń w czasie

Z punktu widzenia czasu liczba naruszeń się zmniejszała, zaczynając od 2010 doku. W 2020 jest znaczny spad liczby naruszeń z powodu koronawirusa. W ciągu roku przestępstwa występują prawie równomiernie.

Nie odnaleziono tendencji do zmiany proporcji zdarzeń według poziomu przestępstw w czasie.

- Czy dochodzi do większej liczby zatrzymań osób czarnoskórych lub ogólnie innej rasy niż biała i z jakich powodów?

Tak, rasa czarna jest najbardziej aresztowana kategoria ludzi. Najczęściej one są aresztowani z powodów naruszeń związanych z narkotykami, napadami kradzieżami i z naruszeniami innych, bardziej szczególnych praw.

- Jakie okresy czasowe i lokacje sprzyjają największej liczby ofiar?

Największa liczba ofiar występuje w okresach godzinowych 12-18 i 18-24. Taka tendencja jest prawdą dla każdego okręgu.

Najwięcej ofiar jest w okręgach Brooklyn, Manhattan i Queens.

Najwięcej ludzie padają ofiarami na ulicach i w swoich domach, najczęściej w apartamentach. Jednak nie jest to prawda dla każdego okręgu, i w Staten Island największa liczba ofiar znajdowała się w swoim domu.

- Czy są jakieś korelacje pomiędzy typem ofiary i przestępcy

Z punktu widzenia rasy najczęściej ofiara i złodziej są o takiej samej rasie. Nie ma wskaźników ze jedna grupa społeczna celowo krzywdzi inną grupę.

Z punktu widzenia grupy wiekowej każda grupa najczęściej staje się ofiara przestępcy o grupie wiekowej 25-44. Złodziej z grupy wiekowej 25-44 najczęściej napada na ofiarę o tej samej grupie wiekowej.

Z punktu widzenia płci każda pleć najczęściej staje się ofiara płci męskiej. Złodziej kobieta i złodziej mężczyzna najczęściej napadają na kobietę.

- W jakich miejscach są największe liczby zabitych ofiar?

Najwięcej zabójstw jest w Brooklyn'ie, Bronx'ie i Queens. Co nie koreluje z okręgami z najczęstszymi występowaniami naruszeń.

Bardziej szczegółowo, najczęściej zabójstw jest w domach i sklepach spożywczych. W różnym stopniu, ale taka tendencja jest prawdziwa dla każdego okręgu.

- Jakie okresy czasowe sprzyjają największej liczbie zabójstw?

Największa liczba zabójstw występuje w nocy: w godzinach 0-6 i 18-24. Najwięcej zabójstw jest w początku godziny.

- Kto najczęściej jest inicjatorem i ofiarą strzelanin

Z punktu widzenia rasy najczęściej napastnikiem i ofiara jest człowiek o rasie białej (nie licząc napastników o niewiadomej rasie)

Z punktu widzenia płci najczęściej napastnikiem i ofiara jest człowiek o płci męskiej.

Z punktu widzenia grupy wiekowej najczęściej napastnikiem jest człowiek o wieku 18-24 (nie licząc napastników o niewiadomym wieku), ofiarą jest człowiek o wieku 25-44. Taka tendencja zachowuje się dla każdej badanej grupy społecznej (rasy i płci)

8. Problemy

8.1. Problemy

Jednym z problemem było wymyślenie analitycznego modelu danych, z którym po jakimś czasie się poradziłem.

Głównym problemem, z którym się spotkałem i nie zdążyłem do końca rozwiązać to naprawa danych tekstowych. Jak było opisane w opisie procesu ETL, nie można było zgrupować dane przy pomocy Fuzzy Grouping, bo może uszkodzić dane. Dlatego zdecydowane było na wykorzystanie Tableu Prep do takiej naprawy danych. Jednak nie zostało zrealizowane to idealnie z powodu braku wcześnie ustandaryzowanych poprawnych wartości i specyfika pracy z Tableu Prep, która nie jest najlepsza w

planie rekcji na zdarzenia. Nie wpłynęło to na wykonywaną analize, ale w końcu nie wiem jaki jest najlepszy i najefektywniejszy sposób rozwiązania tego problemu.

8.2. Pozyskana wiedza i doświadczenie

Podczas realizacji projektu bardziej zrozumiałem dziedzinę analizy danych. Zrozumiałem ważność pasywnego sprawdzania danych po tym, jak załadowałem do wymiaru niepoprawne dane kategorii wiekowych. Jednak dodatkowo zobaczyłem, ze automatyczne naprawianie danych ma swoje wady, głownie z nieprecyzyjnością.

Zrozumiałem jak musi wyglądać proces ETL i jak zrobić go w bardziej efektywny sposób. Ten projekt nie jest stworzony w najbardziej optymalny sposób, bo materiały na zajęciach były przedstawiane częściowo i pewne etapy były tworzone na podstawie posiadanej wiedzy w ten moment czasu, nie wiedzac o bardziej wygodnych metodach rozwiązań problemów . Od tych których zależały kolejne etapy, co zrobiło przerabianie pewnych rzeczy w bardziej efektywny sposób złożonym i czasochłonnym. Robiąc kolejny projekt, z nowa wiedza, zdążę zrobić nadziej wygodny proces.

Bardzo podobała mi się praca z Tableu dla reprezentacji graficznej danych. Z początku nie rozumiałem go możliwości, ale z czasem i doświadczeniem zrozumiałem, jak go można wykorzystać do pobierania potrzebnej informacji.

9. Źródła informacji zużyte podczas analizy danych

[1] http://fiscalpolicy.org/wp-content/uploads/2017/03/Racial-Dimension-of-Income-Inequality.pdf

[2] https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

[3] https://www.infoplease.com/us/census/new-york/demographic-statistics

[4] https://worldpopulationreview.com/us-cities/new-york-city-ny-population

[5] https://rockinst.org/issue-area/growing-drug-epidemic-new-york/

[6] https://www.soundproofcow.com/quietest-neighborhoods-new-york-city/

Uwaga:

- Niekompletny projekt nie będzie sprawdzany i tym samym ocena będzie negatywna!
- Kompletna dokumentacja musi być przesłana do sprawdzenia w formie pliku pdf nie później niż trzy dni przed terminem odbioru i prezentacji opracowanej hurtowni danych!