Zespół: Zuzanna Sieńko, Katarzyna Rogalsk

Opis celu projektu oraz planowane korzyści z perspektywy odbiorcy rozwiązania

Cel biznesowy:

Analiza Wypadków Drogowych i Interwencji Pogotowia Ratunkowego w Nowym Jorku w Zależności od Czynników Zewnętrznych

Celem projektu jest zrozumienie wpływu czynników zewnętrznych takich jak pora dnia, pora roku, limity prędkości, i innych czynników demograficznych takich jak populacja, czy zamożność w danym regionie na częstotliwość wypadków oraz efektywność interwencji pogotowia ratunkowego w Nowym Jorku w roku 2020.

Cel główny:

Opracowanie modelu analitycznego, który pozwoli na zidentyfikowanie kluczowych czynników wpływających na liczbę wypadków drogowych oraz interwencji pogotowia ratunkowego w Nowym Jorku.

Cele szczegółowe:

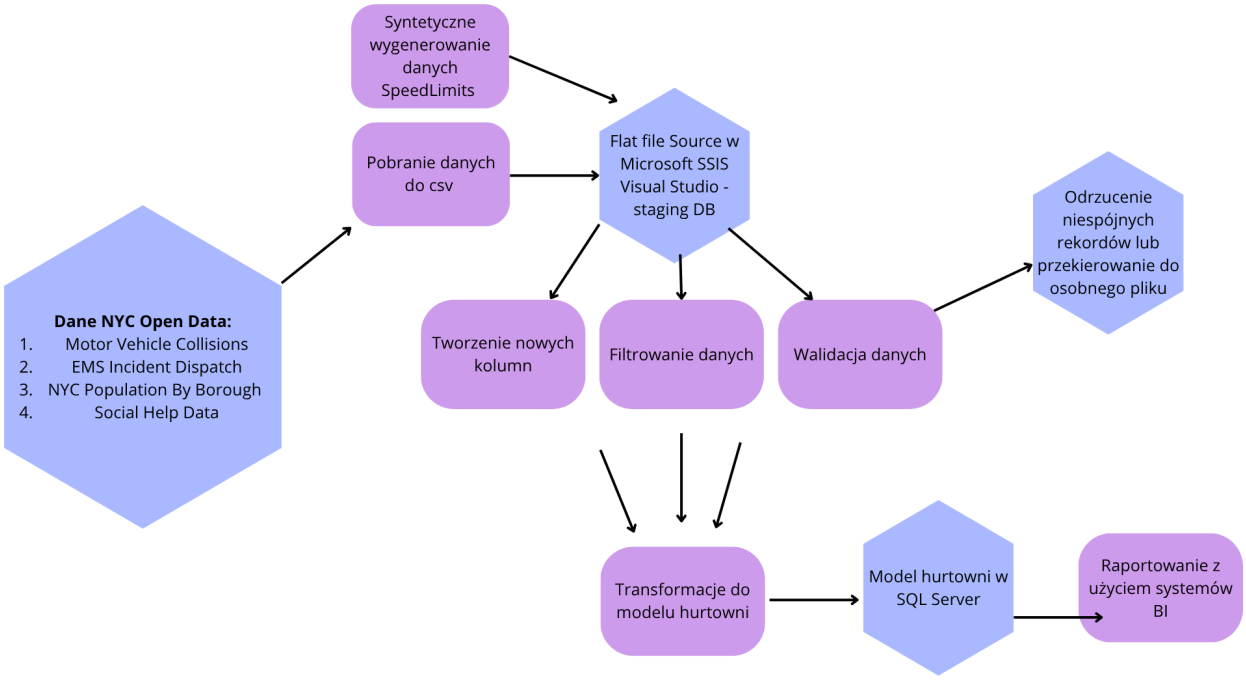
- Zidentyfikowanie zależności między limitami prędkości i oznaczeniami drogowymi, a liczbą wypadków.
- Rekomendacje dotyczące działań poprawiających bezpieczeństwo drogowe i efektywność interwencji pogotowia.

Benefity

- **Poprawa bezpieczeństwa publicznego** - identyfikacja miejsc o podwyższonym ryzyku wypadków drogowych. Możliwość wdrożenia prewencyjnych środków (np. zmiany w organizacji ruchu, infrastruktura ochronna, lepsze oznakowanie).

- **Zwiększenie efektywności służb ratunkowych** - optymalizacja czasów dojazdu do miejsca interwencji na podstawie danych historycznych, zwiększająca szanse na powodzenie akcji ratunkowych.
- **Lepsze zarządzanie budżetem** państwowym w sektorze medycznym, optymalizacja kosztów interwencji
- **Możliwość prognozowania** gdzie i kiedy wystąpi zwiększone ryzyko niebezpiecznych incydentów

Finalna architektura rozwiązania



Projekt zakładał pobranie danych w formatach csv ze strony https://opendata.cityofnewyork.us/ udostępniającej dane tabelaryczne z różnych dziedzin. Finalna architektura różni się względem planowanej ilością pobranych tabel. Z powodu zerowego dopasowania kolumn z danych dotyczących natężenia ruchu drogowego zdecydowano się zrezygnować z tabeli Traffic wspomnianej w pierwszym raporcie. Kolejną wprowadzoną zmianą jest syntetyczne generowanie danych

dotyczących ograniczeń prędkości, również z powodu małego dopasowania kolumn z pobranej tabeli SpeedLimits. Zaimplementowano także mechanizm walidacji danych, który zakładał brak załadowania niespójnych rekordów do tabel w bazie danych oraz przekierowanie niektórych z nich do osobnych plików w celu ręcznej weryfikacji powodu ich odrzucenia. Użyte zostały planowane narzędzia jakimi są Visual Studio 2022 z pakietem SSIS do procesu ETL, SQL Server do przechowywania baz danych i implementacji testów oraz Power BI wykorzystany do raportowania.

Opis wykorzystywanych zbiorów danych

Finalnie wykorzystane dane różnią się względem przedstawionych we wstępnej dokumentacji zrezygnowaniem z dwóch tabel - Traffic oraz SpeedLimits. Tabela Traffic została zaniechana w całym modelu hurtowni, natomiast dane SpeedLimits zostały syntetycznie wygenerowane z użyciem języka Python, co zostanie również opisane w tej części.

- **Motor Vehicle Collisions-Crashes (Facts)**

  https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data_preview

  Tabela z opisem i detalami wypadków samochodowych (uwzględnia również zdarzenie, w których uczestniczyły rowery i motocykle). Dane pochodzą z raportów policji w Nowym Jorku, raport dotyczący kolizji jest wymagany gdy osoba jest ranna, martwa lub straty w wyniku wypadku przekraczają 1000$.

  Zawartość tabeli:

- Crash Date w formacie MM/DD/YYYY
- Crash Time w formacie HH:MM
- Borough - dzielnica Nowego Jorku (Manhattan, Bronx, Queens, Brooklyn, Staten Island), niektóre rekordy mają puste nazwy, ale mają latitude i longitude albo nazwę ulicy
- ZipCode
- Latitude

- Longitude
- Location w formacie (latitude, longitude)
- On Street Name
- Cross Street Name
- Off Street Name
- Number of persons injured
- Number of persons killed
- Number of pedestrians injured
- Number of pedestrians killed
- Number of cyclists injured
- Number of cyclists killed
- Number of motorists injured
- Number of motorists killed
- Contributing Factor Vehicle (1-5) - opisana przyczyna wypadku
- CollisionID - unikalny numer identyfikujący wypadek
- Vehicle Type (1-5) - typ pojazdu biorący udział w zdarzeniu

- **EMS Incident Dispatch Data (Facts)**

  https://data.cityofnewyork.us/Public-Safety/EMS-Incident-Dispatch-Data/76xm-jjuj/about_data

  Dane wygenerowane przez EMS Computer Aided Dispatch System. Zawiera informacje o incydentach od momentów pojawienia się w systemie do czasu ich zakończenia. Dostarczone przez Fire Department of New York City (FDNY).

  Zawiera:

  - CAD_INCIDENT_ID - unikalny numer zgłoszenia w systemie, tworzony na podstawie daty incydentu i stringu 4 liczbowego resetującego się każdego dnia
  - INCIDENT_DATETIME - data i godzina, w którym zgłoszenie zostało stworzone w systemie dispatch

- INITIAL_CALL_TYPE - przypisany rodzaj zgłoszenia przy tworzeniu zgłoszenia, nie musi dokładnie opisywać stanu pacjenta, jest przypisywany na podstawie informacji pochodzących od zgłaszającego
- INITIAL_SEVERITY_LEVEL_CODE - numer priorytetu przypisywany przy tworzeniu zgłoszenia, zakres 0-9
- FINAL_CALL_TYPE - rodzaj zgłoszenia w momencie zamknięcia zgłoszenia
- FINAL_SEVERITY_LEVEL_CODE - numer priorytetu przypisywany przy zamykaniu zgłoszenia, zakres 0-9
- FIRST_ASSIGNMENT_DATETIME - data i czas kiedy pierwsza jednostka jest przypisana do zdarzenia
- VALID_DISPATCH_RSPNS_TIME_INDC - czy elementy do wyliczenia DISPATCH_RESPONSE_SECONDS_QY są poprawne
- DISPATCH_RESPONSE_SECONDS_QY - czas w sekundach, który upłynął od incident_datetime do first_assignment_datetime
- FIRST_ACTIVATION_DATETIME - data i czas, w którym pierwsza jednostka deklaruje, że jest w drodze do miejsca zdarzenia
- FIRST_ON_SCENE_DATETIME - data i czas, w którym pierwsza jednostka deklaruje, że jest na miejscu zdarzenia
- VALID_INCIDENT_RSPNS_TIME_INDC - czy elementy do wyliczenia INCIDENT_RESPONSE_SECONDS_QY są poprawne
- INCIDENT_RESPONSE_SECONDS_QY - czas w sekundach między incident_datetime a first_on_scene_datetime
- INCIDENT_TRAVEL_TM_SECONDS_QY - czas w sekundach między first_assignment_datetime a first_on_scene_datetime.
- FIRST_TO_HOSP_DATETIME - data i czas, w którym pierwsza jednostka deklaruje, że jest w drodze do szpitala
- FIRST_HOSP_ARRIVAL_DATETIME - data i czas, w którym pierwsza jednostka deklaruje, że dojechała do szpitala
- INCIDENT_CLOSE_DATETIME - data i czas, kiedy zgłoszenie jest zamykane w systemie
- HELD_INDICATOR - indykator, że z jakiś powodów jednostka nie mogła zostać przypisana od razu do zgłoszenia
- INCIDENT_DISPOSITION_CODE - kod opisujący wynik interwencji (patrz excel z informacjami z about data sekcja Incident Dispositions)

- BOROUGH - dzielnica Nowego Jorku (BRONX, BROOKLYN, RICHMOND/STATEN ISLAND, QUEENS, UNKNOWN, MANHATTAN)
- INCIDENT_DISPATCH_AREA - najmniejsza jednostka dzielnicy, w której jest zlokalizowany incydent
- ZIPCODE
- POLICEPRECINCT - kod okręgu policyjnego incydentu
- To samo z CITYCOUNCILDISTRICT, COMMUNITYDISTRICT, COMMUNITYSCHOOLDISTRICT, CONGRESSIONALDISTRICT
- REOPEN_INDICATOR - czy zgłoszenie po zamknięciu zostało ponownie otwarte
- SPECIAL_EVENT_INDICATOR - czy incydent miał miejsce podczas specjalnego wydarzenia, np NYC Manhattan
- STANDBY_INDICATOR - czy jednostki zostały przypisane jako standby w razie potrzebnej pomocy
- TRANSFER_INDICATOR - indykator, czy zgłoszenie zostało stworzone jako transport pacjenta, np ze szpitala do domu opieki

- **Speed Limit (Dim)**

  Początkowo planowano skorzystać z danych udostępnionych na stronie https://data.cityofnewyork.us/Transportation/VZV_Speed-Limits/7n5j-865y, jednak w procesie ETL okazało się, że nazwy ulic dostępnych w tej tabeli pokrywają się w bardzo małym stopniu z ulicami pojawiającymi się w tabeli faktów. Stąd, zdecydowano się na syntetyczne wygenerowanie tabeli o następującej strukturze:

  - StreetName - kolumna odpowiadająca nazwie ulicy wygenerowana jako lista unikalnych ulic z załadowanej tabeli faktów
  - SpeedLimit - losowo wygenerowane ograniczenie prędkości 25,30,40 MPH
  - IsSigned - losowo wygenerowana flag YES/NO informująca czy dana ulica jest oznakowana, czy nie

- **NYC Population (Dim)**

https://data.cityofnewyork.us/City-Government/New-York-City-Population-by-Borough-1950-2040/xywu-7bv9/about_data

Dane zawierają dane o populacji z Nowego Jorku w podziale na dzielnice w latach 1950 - 2040 co 10 lat, dane >2020 przewidywane wartości.

Zawartość:

- Borough - nazwa dzielnicy
- 1950 - populacja dzielnicy na rok 1950
- 1950 - Boro share of NYC total - procent udziału mieszkańców dzielnicy w całości populacji Nowego Jorku w 1950
- Analogicznie dla pozostałych lat

- **Social Help (Dim)**

https://data.cityofnewyork.us/Social-Services/Borough-Community-District-Report/5awp-wfkt/data_preview

Zbiór zawiera informacje o liczbie osób i gospodarstw domowych otrzymujących świadczenia SNAP (Supplemental Nutrition Assistance Program), pomoc pieniężną (CA - Cash Assistance) lub ubezpieczenie Medicaid (MA).

Dane zawierają:

- Month - pierwszy dzień raportowanego miesiąca w formacie MM/01/YYYY
- Borough - nazwa dzielnicy w Nowym Jorku
- Community District (CD) - kod okręgu społecznego w danej dzielnicy
- Borough Consultation Total SNAP Recipients - liczba osób otrzymujących pomoc żywieniową
- Borough Consultation Total SNAP Households - liczba gospodarstw domowych otrzymujących pomoc żywieniową
- Borough Consultation Total Cash Assistance Recipients - liczba osób, które otrzymują pomoc pieniężną, w tym osoby z aktywnym świadczeniem, jednorazową pomocą lub objęte sankcjami

- Borough Consultation Total Cash Assistance Cases - liczba spraw, czyli gospodarstw lub jednostek objętych pomocą pieniężną – podobnie jak wyżej, obejmuje aktywne, jednorazowe lub sankcjonowane przypadki.
- Borough Consultation Total Medicaid Only Enrollees - liczba osób zapisanych tylko do programu Medicaid – czyli takich, które nie otrzymują żadnych innych świadczeń (np. SNAP lub CA).
- Borough Consultation Total Medicaid Enrollees - całkowita liczba osób zapisanych do programu Medicaid, niezależnie od tego, czy otrzymują inne świadczenia.

Raport tworzony jest dla roku 2020, dlatego wszystkie dane dotyczące wypadków i zgłoszeń w tabelach faktów traktowane są jako dane historyczne. Dopuszczamy jedynie odświeżanie i modyfikację rekordów dotyczących sytuacji demograficznej dzielnic oraz zmian atrybutów takich jak limity prędkości czy oznakowania na danej ulicy. Możliwość takich sytuacji została zaadresowana w projekcie wykorzystaniem mechanizmu SCD2 w dwóch tabelach wymiarów.

Opis kluczowych miar i atrybutów w modelu

Spośród wszystkich kolumn dostępnych w znalezionych ramkach danych wybrano atrybuty i miary kluczowe dla celu biznesowego. Poniżej przedstawiony jest ich szczegółowy opis. Metody pozyskania tych atrybutów w procesie ETL zostały opisane w kolejnej sekcji.

Tabela Faktów - Kolizje drogowe (Sources: Motor Vehicle Collisions-Crashes, Traffic) :

- CrashID - unikalne ID wypadku (PK)
- CrashDateID - data, w której doszło do wypadku (FK)
- CrashTimeID - czas, w którym doszło do wypadku (FK)
- BoroughKey - klucz obcy do informacji o dzielnicy, w której doszło do wypadku (FK)
- StreetKey - klucz obcy do informacji o ulicy, na której doszło do wypadku (FK)
- Latitude - współrzędne geograficzne miejsca kolizji
- Longitude - współrzędne geograficzne miejsca kolizji
- PeopleKilled - liczba osób, które zginęły w danym wypadku
- PeopleInujred - liczba osób, które zostały ranne w danym wypadku
- PedestriansKilled - liczba przechodniów, którzy zginęli w danym wypadku

- PedestriansInjured - liczba przechodniów, którzy zostali ranni w danym wypadku
- CyclistsKilled - liczba rowerzystów, którzy zginęli w danym wypadku
- CyclistsInjured - liczba rowerzystów, którzy zostali ranni w danym wypadku
- MotoristsKilled - liczba kierowców, którzy zginęli w danym wypadku
- MotoristsInjured - liczba kierowców, którzy zostali ranni w danym wypadku
- VehiclesAmt - wyrażona liczbowo ilość pojazdów biorących udział w wypadku

Tabela faktów kolizji drogowych połączona jest za pomocą kluczy obcych z tabelami wymiarów (Sources: Motor Vehicle Collisions-Crashes,Populacja Nowego Jorku, Speed Limit) :

1. **VehicleTypes_Dim** - wymiar połączony z tabelą faktów mechanizmem bridge zawierający informacje o typach pojazdów biorących udział w danym wypadku
2. **ContributingFactors_Dim** - wymiar połączony z tabelą faktów mechanizmem bridge zawierający informacje o przyczynach danego wypadku
3. **SpeedLimits Dim** - wymiar przechowujący informację o limitach prędkości ustalonych w miejscu gdzie doszło do wypadku
4. **DistrictDetails Dim** - wymiar przechowujący informacje dotyczące dzielnicy, w której doszło do wypadku

Kluczowe atrybuty każdej z tabeli wymiarów to:

1. SpeedLimits
    - StreetID - sztuczny klucz główny (PK)
    - StreetName - nazwa ulicy
    - SpeedLimit - limit prędkości obowiązujący na danej ulicy wyrażony w mph
    - IsSigned - flaga ‘Yes’/’No’ zawierająca informację o tym, czy dany odcinek drogowy jest oznakowany czy nie
2. VehicleTypes
    - VehicleType ID - sztuczny klucz główny (PK)
    - VehicleTypeName - pole tekstowe zawierające typ pojazdu biorącego udział w wypadku np. ‘Sedan’, ‘Motorbike’
3. ContributingFactors
    - ContributingFactor ID - sztuczny klucz główny (PK)

- ContributingFactorName - pole tekstowe zawierające opis przyczyny wypadku np. ‘Speeding’, ‘Rain’

4. DistrictDetails
    - DistrictID - sztuczny klucz główny (PK)
    - DistrictName - pełna nazwa dystryktu w Nowym Jorku
    - AvgSNAPBeneficientsAmt - średnia ilość osób korzystających z programu pomocy żywieniowej SNAP na przestrzeni lat w danej dzielnicy (Low, Medium, High)
    - AvgMedicalSupportAmt - średnia ilość osób korzystających z socialnej pomocy w opłaceniu usług zdrowotnych na przestrzeni lat w danej dzielnicy
    - AvgPopulation - średnie zaludnienie danej dzielnicy na przestrzeni lat
    - ShareOfTotalPopulation - odsetek populacji dystryktu w porównaniu z całym NYC

Tabela Faktów 2 - wezwania pogotowia ratunkowego (Source : EMS Incident Dispatch Data)

- EMS_ID - unikalne ID zgłoszenia (PK)
- IncidentDateKey - klucz obcy do wymiaru daty (FK)
- IncidentTimeKey - klucz obcy do wymiaru czasu dzielącego dobę na pory dnia (FK)
- InitialCallTypeKey - klucz obcy do wymiaru przechowującego informację o typie zgłoszenia - możliwych typów jest ok. 300 (FK)
- FinalCallTypeKey - klucz obcy do wymiaru przechowującego informację o finalnym typie zgłoszenia - możliwych typów jest ok. 300 (FK)
- InitialSeverityLevelKey - klucz obcy do wymiaru przechowującego informację o priorytecie zgłoszenia - możliwych typów jest 10
- FinalSeverityLevelKey - klucz obcy do wymiaru przechowującego informację o finalnym priorytecie zgłoszenia - możliwych typów jest 10
- FistActivationDateKey - data aktywacji pierwszego zespołu ratunkowego w związku z danym zdarzeniem (FK)
- FirstActivityTimeKey - czas aktywacji pierwszego zespołu ratunkowego w związku z danym zdarzeniem (FK)
- FirstOnSceneDateKey - data przybycia na miejsce zdarzenia pierwszego zespołu ratunkowego (FK)
- FirstOnSceneTimeKey - godzina przybycia na miejsce zdarzenia pierwszego zespołu ratunkowego (FK)

- FistHospitalArrivalDateKey - data przybycia do szpitala (FK)
- FistHospitalArrivalTimeKey - godzina przybycia do szpitala (FK)
- ActivationDuration - miara obliczona jako różnica w datach aktywowania pierwszego zespołu ratunkowego a wykonaniem telefonu przedstawiona w minutach
- OnSceneArrivalDuration - miara obliczana jako różnica w datach przybycia na miejsce zdarzenia a aktywacją zespołu ratunkowego przedstawiona w minutach
- HospitalArrivalDuration - miara obliczana jako różnica w datach przybycia do szpitala a przybycia na miejsce zdarzenia przedstawiona w minutach
- InterventionDuration - całościowa długość interwencji obliczona jako różnica w datach zamknięcia zgłoszenia w systemie a wykonania połączenia alarmowego przedstawiona w minutach

Z drugą tabelą faktów połączone są następujące tabele wymiarów:

1. **CallTypes Dimension** - wymiar role playing dim przechowujący informację o typach zgłoszeń, dwukrotnie połączony z tabelą faktów kolumnami InitialCallTypeID i FinalCallTypeID
2. **SeverityLevel Dimension** - wymiar role playing dim przechowujący informację o priorytecie zgłoszenia, dwukrotnie połączony z tabelą faktów kolumnami InitialSeverityLevelID, FinalSeverityLevelID
3. **DistrictDetails Dimension** - wymiar wspólny dla tabeli faktów wypadków drogowych i interwencji pogotowia, przechowuje szczegóły demograficzne opisujące dzielnicę zdarzenia

Atrybuty wymiaru **CallTypes**:

- CallTypeID - sztuczny klucz główny od 0 do ok 300 (PK)
- CallTypeShort - kolumna tekstowa zawierająca skrót typu zgłoszenia
- CallType - kolumna tekstowa zawierająca pełne rozwinięcie typu zgłoszenia

Atrybuty wymiaru **SeverityLevels**:

- SeverityLevelID - naturalnie występujący w danych klucz główny od -1 do 9 (PK)
- SeverityLevelName - kategoryczna nazwa priorytetu zgłoszenia np. ‘Critical Condition’

Stworzono także wymiary daty i czasu o następujących kolumnach:

Wymiar daty:

- DateID - sztuczny klucz główny (PK)
- FullDate - data w formie date
- Day - dzień miesiąca
- WeekdayNumber - dzień tygodnia
- WeekdayName - nazwa dnia tygodnia
- Month - miesiąc
- MonthName - nazwa miesiąca
- Year - rok
- WeekendFlag - flaga ‘Yes’/’No’ informująca czy dany dzień to sobota lub niedziela
- Season - nazwa pory roku

Wymiar czasu:

- TimeID - sztuczny klucz główny
- TimeValue - czas w formacie time zaokrąglony do pełnych godzin
- DayPart - pora dnia jako jedna z : Early Morning, Morning, Noon, Early Afternoon, Afternoon, Evening, Night, Late Night

Hierarchie występujące w danych znajdują się jedynie w wymiarze daty i czasu gdzie mamy

Rok → Miesiąc → Dzień miesiąca → Dzień tygodnia

Pora dnia → Pełna godzin

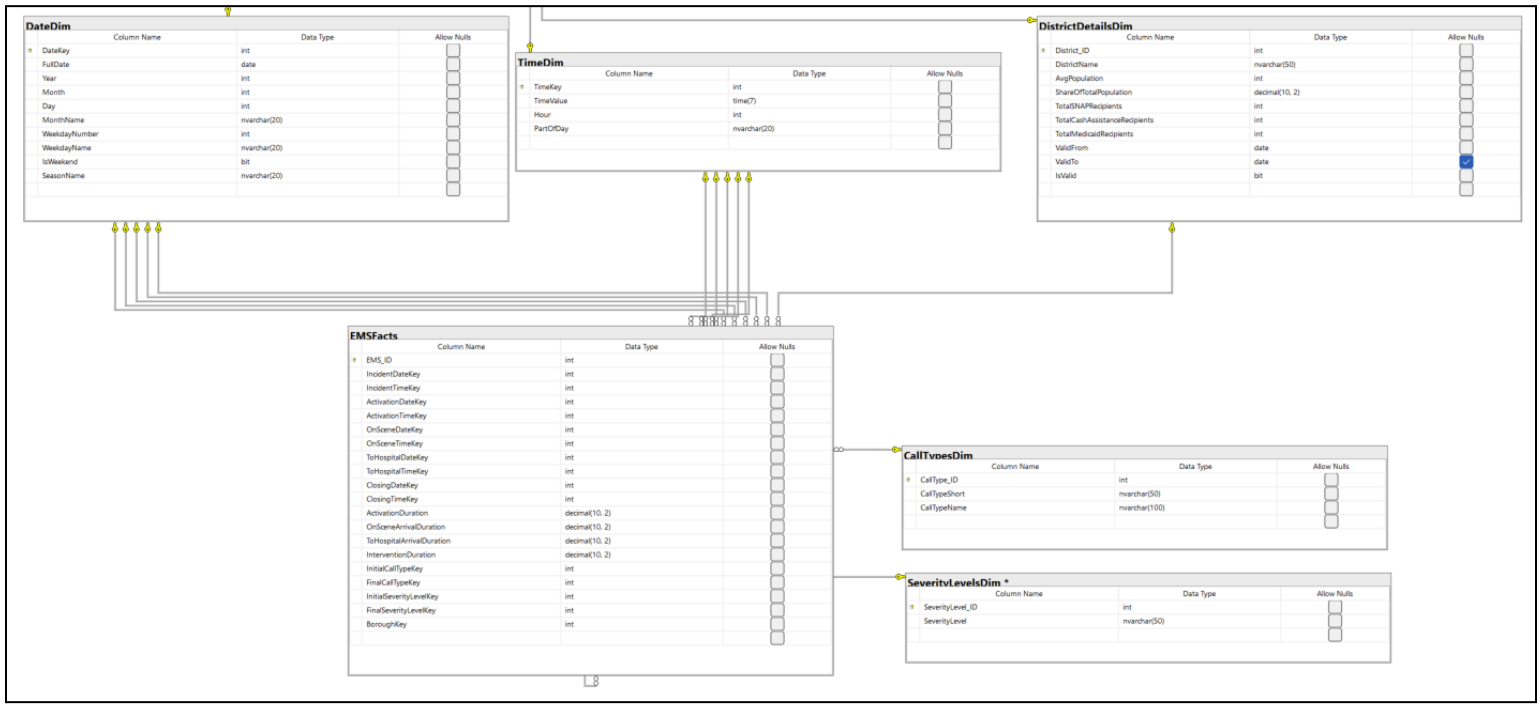
Model fizyczny hurtowni

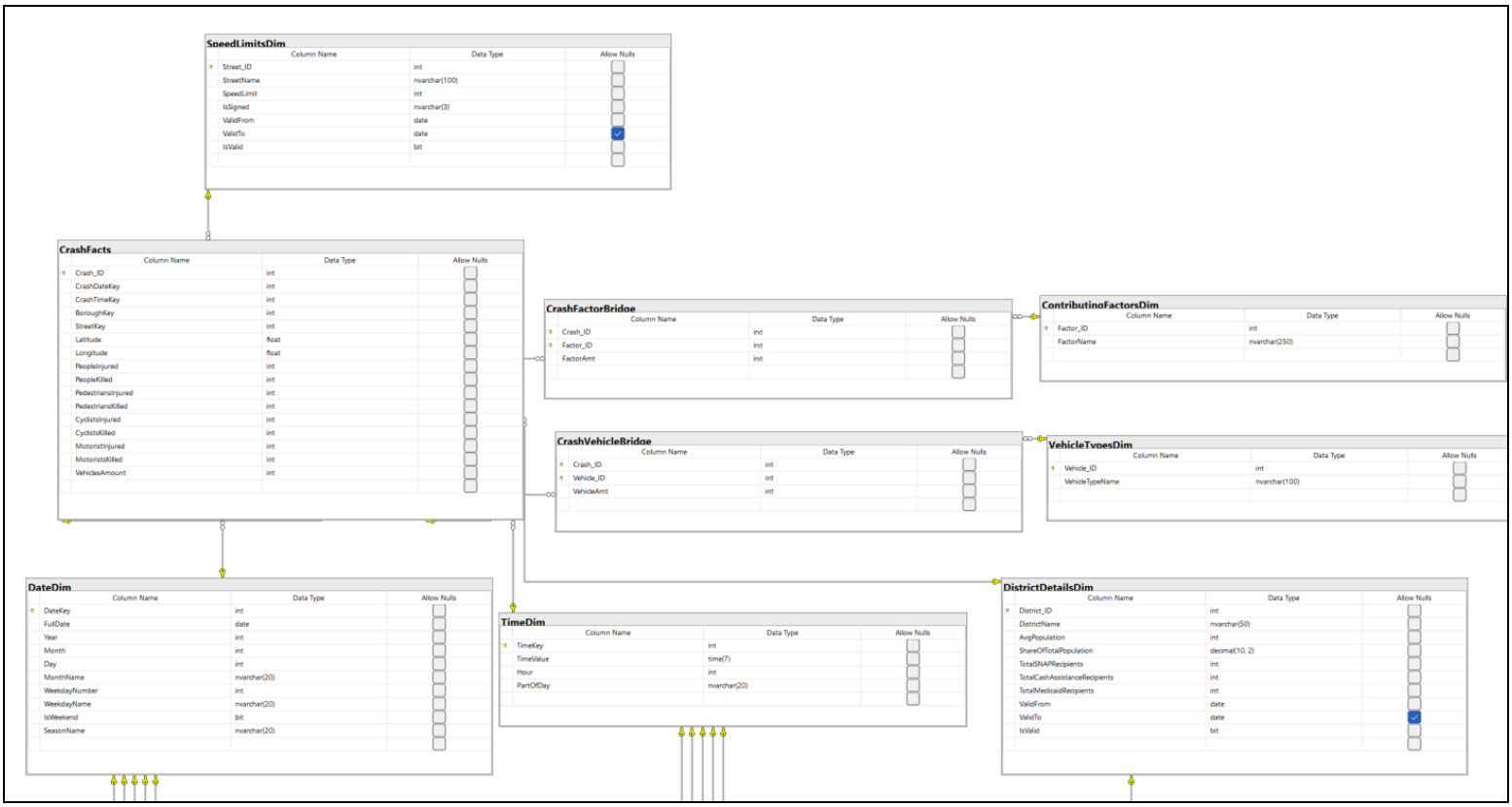
Hurtownia została zaimplementowana w **schemacie galaktyki**. Występują dwie tabele faktów CrashFacts oraz EMSFacts mające połączenia z wymiarami w następujący sposób:

1. Wymiary dotyczące tylko tabeli EMS: CallTypes Dim, SeverityLevels Dim

2. Wymiary dotyczące tylko tabeli Crashes : SpeedLimits Dim, VehicleTypes Dim, ContributingFactors Dim
3. Wymiary wspólne dla obu faktów: Date Dim, Time Dim, DistrictDetails Dim

W modelu występują tzw. **Role Playing Dimensions** mające kilka połączeń z tabelami faktów. Są to Date Dim, Time Dim, CallTypes Dim oraz SeverityLevels Dim. Dodany został **mechanizm Bridge** do rozwiązania sytuacji relacji wiele do wielu tabeli CrashFacts i ContributingFactors Dim oraz VehicleTypes Dim. Dwie tabele wymiarów DistrictDetails oraz SpeedLimits mają także zaimplementowany tryb śledzenia zmian **SCD2**.





Proces ETL

Proces ETL został przeprowadzony z wykorzystaniem narzędzia SSIS w połączeniu z SQL Server. Do niektórych komponentów w SSIS dodano mechanizm przekierowania błędów do pliku w celu zaprezentowania tego mechanizmu, jednak nie było to wykorzystane w całym projekcie, ponieważ błędy powinny wstrzymywać działanie procesu na tym etapie rozwoju w celu efektywnego poprawiania ich na bieżąco. Proces ETL został podzielony na kilka kluczowych etapów. Pierwszym z nich było załadowanie pobranych ze strony NYC Data plików csv do pośredniej bazy danych Staging DB wraz z ich wstępnym czyszczeniem i transformacją. Następnie wczytane dane były używane do załadowania tabel wymiarów i faktów w docelowej hurtowni. Na koniec sprawdzano poprawność działania procesu oraz jego kolejnych iteracji za pomocą licznych testów znajdujących się w dołączonym pliku Testy.pdf.

Załadowanie bazy Staging DB

W pierwszym etapie procesu ETL do bazy pośredniej zostały załadowane następujące pliki csv: EMS, Social Help, Speed Limits, NYC Population oraz Crashes i wykonano na nich poniższe transformacje.

Tabela EMS

Tabela EMS to główne źródło danych do jednej z tabeli faktów, stąd proces czyszczenia danych rozpoczęto już na etapie ładowania do bazy pośredniej.

Kolumny, dla których wiersze z brakami danych zostały pominięte:

- Wszystkie kolumny typu '[...]DateTime' np. IncidentDateTime, OnSceneDateTime itd - będą one zamieniane na mapowanie do wymiaru daty i czasu i są jednymi z głównych danych potrzebnych w analizie raportowej, więc nie powinny być brakujące.

Kolumny, dla których zastosowano imputację braków danych:

- Initial oraz Final CallType - brak danych zamieniony na 'Unknown'.
- Initial oraz Final SeverityLevel - brak danych zamieniony na wartość -1, która w wymiarze odpowiada opisowi 'Unknown'.
- Borough - brak danych zamieniony na 'Unknown', jednak w danych źródłowych nie odnotowano przypadku, gdzie nazwa dzielnicy byłaby brakująca.

Wiersze, które zostały usunięte ze względu na niespójności przekazywanych informacji:

- Podczas ładowania danych zauważono, że zdarzają się rekordy, w których chronologia dat jest nieprawidłowa. To znaczy, przykładowo z wartości w kolumnach wynikałoby, że dany pacjent wcześniej dojechał do szpitala, niż powiadomił służby ratunkowe o zdarzeniu. Z oczywistych przyczyn takie rekordy, w których zaobserwowano niespójności dla którejkolwiek z dat, zostały pomijane, aby zapewnić wysoką jakość przekazywanych informacji raportowych.
- Usunięto także rekordy zawierające tzw. Outliery, gdzie którakolwiek z wyliczonych miar czasu wynosiła powyżej 24 godzin. W większości przypadków oznaczało to błędne wartości w kolumnach dat, co zostało usunięte, aby nie takie obserwacje nie zaburzały całego przekazu.

- Ponieważ domyślnie w danych kolumna EMS_ID powinna być unikalnym identyfikatorem zdarzenia, usunięto wiersze, gdzie wystąpił duplikat klucza głównego, aby wyeliminować zjawisko powtarzających się informacji.

Nowe kolumny, które zostały dodane do tabeli EMS:

- Activation duration - obliczona jako różnica w minutach między datą i czasem zgłoszenia, a datą i czasem aktywacji pierwszej jednostki.
- On scene arrival duration - obliczona jako różnica w minutach między datą i czasem aktywacji pierwszej jednostki, a datą i czasem jej przybycia na miejsce zdarzenia.
- To hospital arrival duration - obliczona jako różnica w minutach między datą i czasem przybycia służb na miejsce zdarzenia, a datą i czasem dotarcia do szpitala.
- Intervention duration - obliczona jako różnica w minutach między datą i czasem zgłoszenia, a datą i czasem zamknięcia zgłoszenia w systemie.

Wczytane do tabeli kolumny zostały zmapowane na następujące typy danych:

- IncidentDateTime, FirstActivationDateTime, FirstOnSceneDateTime, FirstToHospitalDateTime, ClosingDateTime → datetime
- InitialCallType, FinalCallType, Borough → nvarchar
- InitialSeverityLevel, FinalSeverityLevel → int
- Wszystkie kolumny obliczone jako duration w minutach → decimal

Podczas ładowania danych z pliku csv do Staging DB zaimplementowano mechanizm, który przed wstawieniem nowego rekordu sprawdza, czy taka obserwacja nie znajduje się już w bazie danych w celu uniknięcia dodawania do bazy duplikatów w kolejnych iteracjach. Kolumną identyfikującą wiersze jest EMS_ID będący kluczem pobranym z oryginalnych danych. Założono, że dane o zgłoszeniach są traktowane jako dane historyczne i nie mogą się zmieniać, ewentualnie jedynie napływać nowe.

Po odfiltrowaniu danych zgodnie z powyższymi kryteriami do tabeli EMS załadowano **820 880** wierszy.

Tabele Social Help i NYC Population

Tabele Social Help i NYC Population są tabelami zasilającymi wymiar DistrictDetails w finalnej hurtowni. We wstępnym etapie wczytywania do bazy pośredniej zostały one tylko załadowane wraz z mapowaniem typów danych w następujący sposób:

Tabela Social Help zawiera kolumny:

- Borough → nvarchar
- TotalSNAPRecipients, TotalCashAssistanceRecipients, TotalMEDICAIDRecipients → int

Na tym etapie nie zaimplementowano jeszcze agregacji tabeli po dzielnicach. Występuje więc sytuacja, w której dla jednej dzielnicy widnieje w tabeli wiele rekordów, które będą następnie uśrednione w obrębie dzielnicy w kolejnym kroku.

Tabela NYC Population zawiera kolumny:

- Borough → nvarchar
- AvgPopulation → int
- ShareOfTotalPopulation → decimal

Oryginalne dane zawierają po jednym rekordzie dla każdej dzielnicy, dlatego nie wystąpiła konieczność transformacji tej tabeli.

Tabela Crashes

Tabela Crashes to tabela zasilająca drugą tabelę faktów w finalnej hurtowni. Podobnie jak w przypadku EMS proces czyszczenia danych rozpoczęto już na etapie ładowania do Staging DB.

Kolumny, dla których wiersze z brakami danych zostały pominięte:

- CrashDate i CrashTime - ponownie, są to kolumny mapowane w kolejnych etapach na klucze obce do tabeli wymiarów oraz wartości kluczowe w analizie biznesowej, stąd braki danych nie zostały dopuszczone.
- Borough, StreetName, Latitude, Longitude - analogicznie jak w przypadku powyżej, kolumny te są kluczowe i braki danych zostały pominięte.
- Metryki takie jak PeopleKilled, PedestriansInjured itd - tabela wypadków nie zawiera zbyt wiele metryk numerycznych, dlatego zdecydowano się skupić tylko na rekordach, w których istnieje możliwość dostępu do pełnych informacji o wypadku. Uznano, że sztuczna imputacja braków np. Wartościami ujemnymi nie okaże się korzystna raportowo.

Przeprowadzona imputacja braków danych:

- Możliwe braki danych występują w kolumnach typu ContributingFactor oraz VehicleType. Podczas analizy danych przyjęto następującą logikę:
    a) Jeśli obie kolumny dla danego pojazdu są brakujące, to nie brał on udziału w zdarzeniu. Np. Factor3 = NULL i Vehicle3 = NULL, to pojazd nr. 3 nie uczestniczył w wypadku.
    b) Jeśli tylko jedna z kolumn dla danego pojazdu jest brakująca, to uzupełniona zostaje wartością ‘Unknown’. Np. Factor3 = ‘Rainy weather’ i Vehicle3 = NULL, to Vehicle3 zamieniony zostaje na wartość ‘Unknown’

Nowe dodane kolumny:

- Dodaną metryką na poziomie wypadku jest VehicleAmount informujący o ilości pojazdów biorących udział w zdarzeniu. Metryka obliczana jest zgodnie z logiką z poprzedniego podpunktu, to znaczy, tylko jeśli obie kolumny Factor i Vehicle są brakujące, można uznać, że pojazd nie brał udziału w zdarzeniu.

Konwersja typów danych nastąpiła zgodnie z:

- CrashDate → Date
- CrashTime → Time
- Borough, StreetName, Factor1,...Factor5, Vehicle1,...Vehicle5 → nvarchar
- Latitude, Longitude → float
- PeopleInjured, …, MotoristsKilled → int

Klucz główny tabeli Crashes również został pobrany z oryginalnych danych. Nie zidentyfikowano przypadku, gdzie łamałby on zasadę unikalności.

Po odfiltrowaniu danych zgodnie z powyższymi kryteriami otrzymano **43 537** wierszy w tabeli Crashes.

Tabela Speed Limits

Z powodu problemów zaistniałych z dopasowaniem znalezionych danych do rekordów z tabeli Crashes zdecydowano się na syntetyczne wygenerowanie tabeli SpeedLimits zawierającej losowe wartości SpeedLimit i IsSigned dla każdej ulicy zidentyfikowanej w

tabeli Crashes. W procesie ładowania danych do Staging DB zmieniono jedynie typy danych zgodnie z:

- StreetName → nvarchar
- SpeedLimit → int
- IsSigned → char(3)

Wszystkie wyżej wymienione tabele mają zaimplementowany mechanizm sprawdzenia duplikatów przed dodaniem nowego rekordu do Staging DB.

Wypełnianie wymiarów

Kolejnym etapem procesu ETL było stworzenie tabel wymiarów i wypełnienie ich danymi pobieranymi z przejściowej bazy danych lub z dodatkowych plików csv. Tabele wymiarów związane z EMS to : CallTypesDim oraz SeverityLevelsDim. Z tabelą Crashes związane są SpeedLimitsDim, VehicleTypesDim, ContributingFactorsDim, a także wspólne wymiary dla obu tabel faktowych takie jak DateDim, TimeDim i DistrictDetailsDim.

CallTypes Dimension

Wymiar słownikowy powstały bezpośrednio przez załadowanie pliku call_types_map.csv, który został dołączony jako dodatkowy materiał do danych EMS. Zawiera kolumny Call Type i Call Type Description. Tabela służy do mapowania skrótów związanych z typem zgłoszenia na ich pełne nazwy. Finalna struktura tej tabeli to:

- CallType_ID - sztuczny inkrementowany klucz główny
- CallTypeShort, CallTypeName - nvarchar

SeverityLevels Dimension

Wymiar słownikowy służący do mapowania kodów priorytetu zgłoszenia na pełne nazwy takie jak ‘Critical Condition’ itd. Z powodu braku jednoznacznie udostępnionego pliku zawierającego mapowania kodów, tabela wymiarów została wypełniona ręcznie bazując na materiałach znalezionych online. Zawiera ona dwie kolumny SeverityLevel_ID (będący kluczem głównym) oraz SeverityLevel wypełnione za pomocą skryptu SQL w następujący sposób:

```
INSERT INTO SeverityLevelsDim (SeverityLevel_ID, SeverityLevel)
VALUES
    (1, 'Immediate Life Threat'),
    (2, 'Critical Condition'),
    (3, 'Severe Condition'),
    (4, 'Moderate Condition'),
    (5, 'Minor Condition'),
    (6, 'Non Urgent'),
    (7, 'Informational Call'),
    (8, 'Informational Call'),
    (9, 'Misuse'),
    (-1,'Unknown');
```

DistrictDetails Dimension

Wymiar wspólny dla obu tabel faktowych dostarczający informacji o danych dotyczących konkretnych dzielnic Nowego Jorku. Wymiar powstał przez połączenie informacji zawartych w tabelach NYC Population i Social Help. W tym celu wykonano następujące kroki:

Tabela NYC Population

1. Pobranie wyczyszczonych danych z Staging DB
2. Konwersja nazw dzielnic na wielkie litery
3. Mapowanie wartości ‘Staten Island’ na ‘Richmond / Staten Island’
4. Sortowanie danych po wartościach kolumny Borough

Tabela Social Help

1. Pobranie wyczyszczonych danych z Staging DB
2. Grupowanie danych po kolumnie Borough oraz uśrednienie pozostałych kolumn numerycznych
3. Konwersja nazw dzielnic na wielkie litery
4. Mapowanie wartości ‘Staten_Island’ na ‘Richmond / Staten Island’
5. Sortowanie danych po wartościach kolumny Borough

Następnie wykonano Inner Join obu tabel używając kolumny Borough jako klucz. Potem przekonwertowano uśrednione wartości numeryczne na typ int i załadowano dane do tabeli StagingDistrictTable w hurtowni danych. Tabela staging służy jako bufor danych w implementacji mechanizmu SCD2 dla tego wymiaru. Dane najpierw trafiają do tabeli Staging, w kolejnym kroku sprawdzane jest, czy rekord dla danej dzielnicy istnieje już w tabeli wymiarów:

a) Identyczny rekord już istnieje i ma flagę IsValid = 1 → nie ładuje się ponownie.

b) Istnieje rekord dla danej dzielnicy, ale zawiera inne wartości w pozostałych kolumnach → wykonywany jest update mechanizmem SCD2. Rekord zostaje dodany do tabeli wymiaru z kolumnami ValidFrom = data dodania, ValidTo = NULL, IsValid = 1, a poprzedni istniejący rekord zostaje zmodyfikowany i jego kolumna ValidTo ustawiona jest na datę dodania nowego rekordu oraz flaga IsValid zostaje zmieniona na 0.

c) Nie istnieje rekord z daną dzielnicą → jest on dodawany jako zupełnie nowy rekord do tabeli wymiaru.

SpeedLimits Dimension

Wymiar wypełniony przy pomocy danych ze Staging DB. Nie wymaga żadnych dodatkowych transformacji w porównaniu z załadowaniem do bazy pośredniej. Dane są najpierw ładowane do Staging table w docelowej hurtowni, która również służy jako bufor dla mechanizmu SCD2 działającego analogicznie jak w przypadku DistrictDetails Dim.

ContributingFactors Dimension

Wymiar łączony z tabelą faktów za pomocą mechanizmu bridge zawierający wszystkie unikalne wartości kolumn Factor1,... Factor5 z tabeli Crashes. W celu załadowania wymiaru wykonano następujące kroki:

1. Stworzono 5 osobnych OLE DB Source, w każdym wybrano jedną kolumnę spośród Factor1,... Factor5 i zmieniono jej nazwę na Factor
2. Dołączono do siebie wszystkie kolumny w pionie tworząc jedną długą kolumnę Factor
3. Usunięto braki danych z powstałej kolumny
4. Usunięto duplikaty oraz zamieniono wszystkie wartości na wielkie litery

5. Załadowano wszystkie znalezione unikalne wartości do tabeli wymiarów wraz ze sztucznie inkrementowanym kluczem głównym

VehicleTypes Dimension

Wymiar łączony z tabelą faktów za pomocą mechanizmu bridge zawierający wszystkie unikalne wartości kolumn VehicleType1,...VehicleType5 z tabeli Crashes. W celu załadowania wymiaru wykonano następujące kroki:

1. Stworzono 5 osobnych OLE DB Source, w każdym wybrano jedną kolumnę spośród VehicleType1,...,VehicleType5 i zmieniono jej nazwę na VehicleType
2. Dołączono do siebie wszystkie kolumny w pionie tworząc jedną długą kolumnę VehicleType
3. Usunięto braki danych z powstałej kolumny
4. Usunięto duplikaty oraz zmieniono wszystkie wartości na wielkie litery i dodano kilka ręcznych poprawek w nazwach
5. Załadowano wszystkie znalezione unikalne wartości do tabeli wymiarów wraz ze sztucznie inkrementowanym kluczem głównym

Date oraz Time Dimensions

Wymiary dat i czasu zostały zainicjalizowane za pomocą skryptu SQL. Wymiar dat zawiera kolumny takie jak:

- DateKey - klucz główny w formacie YYYYMMDD
- FullDate, Year, Month, Day - numeryczne wartości pól daty
- MonthName, WeekDayName - tekstowe pola nazw miesięcy i dni
- IsWeekend - flaga 0/1 informująca czy dany dzień przypada w weekend
- SeasonName - nazwa pory roku

Wymiar daty został wypełniony datami od 1 stycznia 2020 do 31 grudnia 2020, co skutkowało finalną ilością 366 rekordów (rok przestępny).

Wymiar czasu został zaimplementowany z granulacją co godzinę, czyli składał się z 24 wierszy o kolumnach:

- TimeKey - klucz główny w formacie HH0000
- TimeValue - pełna godzina o typie time
- Hour - numeryczna wartość godziny, gdzie 0 to północ, 1 to pierwsza w nocy itd.

- PartOfDay - tekstowe pole pory dnia. Godziny 00-04 klasyfikowane są jako noc, 05-08 jako wczesny poranek, 09-11 późny poranek, 12-16 jako popołudnie, 17-20 wieczór oraz 21-23 późny wieczór.

Wypełnienie faktów

Kluczowym etapem procesu ETL było wypełnienie tabel faktów oraz poprawne zmapowanie kolumn w nich zawartych na klucze obce do tabel wymiarów.

EMS Facts

Podczas procesu ETL korzystając z załadowanej tabeli w Staging DB oraz wypełnionych już tabel wymiarów załadowano tabelę faktów wykonując następujące kroki:

1. Pobrano dane z tabeli EMS w Staging DB
2. Dodano nowe kolumny mapujące pola typu DateTime na kolumny w formatach zgodnych z kluczami głównymi wymiarów dat i czasu zgodnie z przykładowymi formułami:

   Dla wymiaru dat:
   ```
   YEAR(IncidentDateTime) * 10000 + MONTH(IncidentDateTime) * 100
   + DAY(IncidentDateTime)
   ```

   oraz dla wymiaru czasu:
   ```
   DATEPART("hour",FirstOnSceneDateTime) * 10000)
   ```

   Jak można zauważyć podczas mapowania danych godzinowych na klucze obce nie uwzględniono minut, tylko pełne godziny zegarowe. Nie zaimplementowano także mechanizmów zaokrąglania, jedynie wybrano część godzinową z pola time.

3. Za pomocą komponentu Lookup dodano mapowania nowo stworzonych kolumn do kluczy głównych wymiarów z ustawieniem opcji Fail Component, jeśli jakiś wiersz nie znalazłby dopasowania.

4. W kolejnych komponentach Lookup dodano mapowanie kolumn Initial i Final CallType do wymiaru CallTypes Dim. Jako klucz dopasowania użyto kolumny CallType w tabeli EMS oraz CallType w tabeli wymiarów zwracając jako wynik dopasowania odpowiadający mu klucz główny z wymiaru. Ponieważ ten krok bazuje na dostarczonej przez twórców danych tabeli słownikowej, dopuszczono możliwość pewnych niespójności, takich jak brak uwzględnienia pojawiających się w tabeli EMS typów zgłoszeń w słowniku. Stąd, dodana została opcja ‘Redirect Rows with No match output’, jako przekierowanie rekordów, które nie znalazły odpowiednika do osobnego pliku. Następnie ręcznie analizowano rekordy przekierowane do tego pliku w porównaniu z dostarczonym słownikiem csv i rzeczywiście zauważono braki pewnych typów zgłoszeń w słowniku. Są to skróty takie jak np. CARDBR, EDPC itd.

5. Kolejną transformacją było dodanie mapowania kolumn Initial oraz Final SeverityLevel z użyciem kolumny SeverityLevelCode jako klucza. W tym przypadku nie dopuszczano możliwości nieznalezienia dopasowania, więc ponownie ustawiono ‘Fail component’ gdyby takie zdarzenie miało miejsce.

6. Ostatnie mapowanie dotyczyło wymiaru DistrictDetailsDim, gdzie dopasowywano nazwy dzielnic i również zwracano klucz główny występujący w tabeli wymiaru nie dopuszczając możliwości braku dopasowania.

W taki sposób powstała pierwsza tabela faktów zawierająca kolumny:

- EMD_ID - klucz główny załadowany z EMS Staging DB
- IndicentDateKey, ActivtionDateKey, OnSceneDateKey, ToHospitalDateKey, ClosingDateKey - klucze obce do wymiaru dat
- IncidentTimeKey, ActivationTimeKey, OnSceneTimeKey, ToHospitalTimeKey, ClosingTimeKey - klucze obce do wymiaru czasu
- InitialCallTypeKey, FinalCallTypeKey - klucze obce do wymiaru CallTypesDim
- InitialSeverityLevelKey, FinalSeverityLevelKey - klucze obce do wymiaru SeverityLevelsDim
- BoroughKey - klucz obcy do wymiaru DistrictDetailsDim
- ActivationDuration, OnSceneArrivalDuration, ToHospitalArivalDuration, InterventionDuration - obliczone wcześniej miary pobrane z tabeli EMS Staging DB

W procesie ETL zaimplementowano także mechanizm zapobiegający dodawaniu duplikatów do tabeli faktów w kolejnych iteracjach. Dane o zgłoszeniach są uznawane jako dane historyczne, więc atrybuty zgłoszenia o danym ID nie powinny być zmieniane, jedynie dodawane jako nowe rekordy. Mechanizm ten zapewnia także uaktualnienie kluczy obcych do wymiarów, które podlegają mechanizmowi SCD2 tak, aby fakt był zawsze połączony z najbardziej aktualną wersją wymiaru.

Po kolejnych odfiltrowaniach oraz przekierowaniach rekordów bez dopasowań w tabeli faktów pozostało **726 033 wierszy.**

Crash Facts

Drugą tabelą faktów była tabela dotycząca wypadków samochodowych załadowana za pomocą tabeli Crashes Staging DB oraz połączeń z wymiarami. W tym celu wykonano następujące kroki:

1. Pobrano dane wcześniej wczytane do Crashes Staging DB
2. Dodano nowego kolumny mapujące CrashDate oraz CrashTime na klucze obce w takim samym formacie jak EMS, to znaczy klucze dat w formacie YYYYMMDD oraz czasu z uwzględnieniem tylko pełnej godziny w formacie HH0000.
3. Kolejno zmapowano kolumny CrashDate i CrashTime na klucze obce za pomocą komponentu Lookup nie dopuszczającego sytuacji braku dopasowania.
4. Dodano zmianę wartości w kolumnie Borough z ‘Staten Island’ na ‘Richmond / Staten Island’ oraz wykonano mapowanie do wymiaru DistrictDetails używając nazw dzielnic jako klucz. Nie dopuszczano sytuacji, w której rekord nie znajdzie dopasowania.
5. Analogiczne mapowanie wykonano do wymiaru SpeedLimits korzystając z kolumny StreetName jako klucza. Również nie dopuszczano sytuacji, w której rekord nie znajdzie dopasowania, ponieważ dane SpeedLimits zostały wygenerowane syntetycznie tak, aby zawierały wszystkie ulice występujące w danych.

Zaimplementowano także mechanizm zapobiegający wstawianiu do tabeli faktów zduplikowanych danych. Założono, że są to dane historyczne, dlatego atrybuty wypadku o konkretnym ID nie powinny się zmieniać. Mechanizm jedynie uaktualnia klucze obce odnoszące się do tabel wymiarów z SCD2 tak, aby fakt zawsze odnosił się do aktualnego rekordu w wymiarze. W tak zaimplementowanej tabeli faktów znalazło się **43 537 rekordów**.

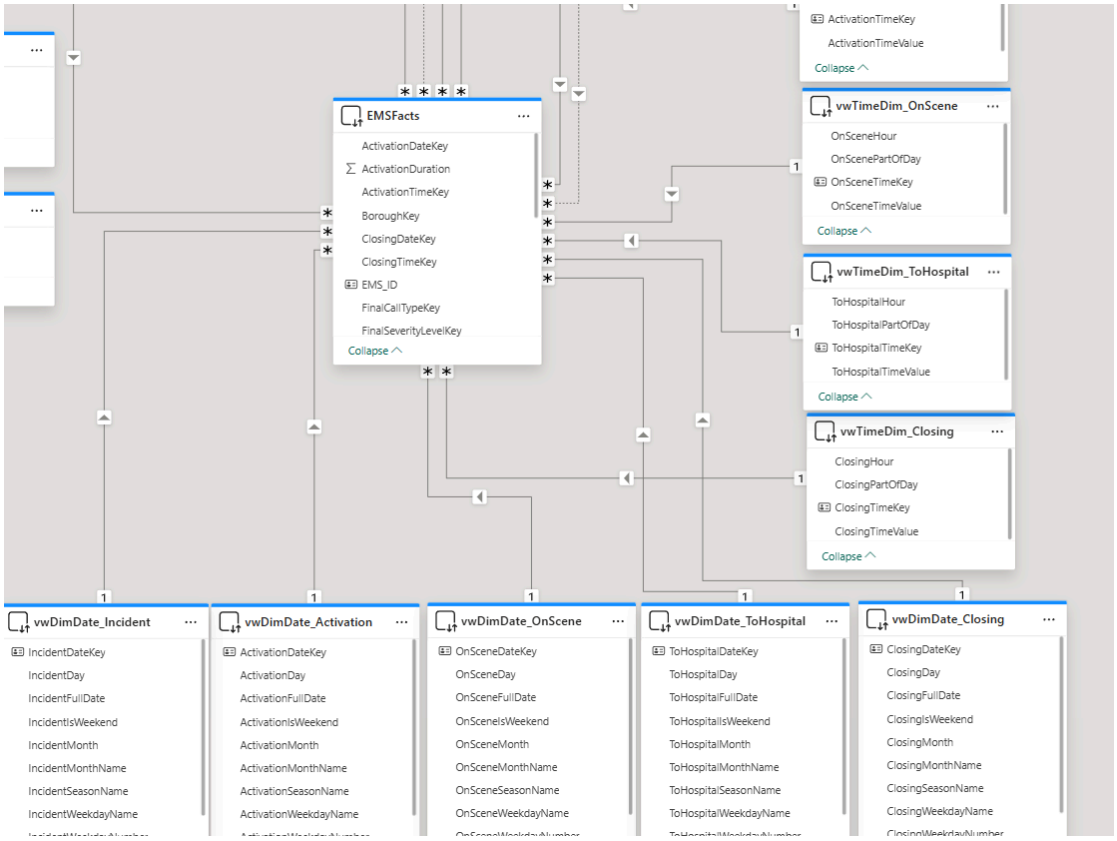
Deployment

Po zweryfikowaniu poprawności działania procesu ETL oraz załadowaniu hurtowni danych wykonano ostatni krok, czyli deployment pakietu do SQL Server, którego poszczególne kroki wraz ze zrzutami ekranu zostały zaprezentowane w pliku Testy.pdf.

Opis warstwy raportowej

Warstwa raportowa została zaimplementowana z wykorzystaniem Microsoft Power BI. Dane zostały wczytane do modelu raportowego za pomocą połączenia DirectQuery do bazy danych SQL Server, zawierającej przetworzone dane – po zakończeniu procesów ETL.

Wymiary czasowe DateDim i TimeDim zostały zaimplementowane jako widoki (views), aby umożliwić elastyczne i czytelne odwzorowanie różnych kontekstów czasowych w analizie danych. Dla każdego kontekstu czasowego został przygotowany dedykowany widok, który odwołuje się do wspólnej tabeli źródłowej (DateDim lub TimeDim), zawiera kolumny z przekształconymi nazwami, które jasno odzwierciedlają dany kontekst analityczny (np. CrashMonth, CrashSeasonName, IncidentHour, ActivationHour). Relacje tabela faktowa - wymiar (view) zostały dodane ręcznie. Poniżej przykład implementacji widoków dla tabeli faktowej EMSFacts, analogicznie dla CrashFacts.



Wszystkie klucze główne zostały ręcznie zdefiniowane w modelu Power BI, ponieważ mechanizm importu przy połączeniu DirectQuery nie przenosi definicji kluczy głównych z bazy danych SQL Server.
Domyślnie, w Power BI wszystkie klucze obce prowadzące do tabel wymiarów oraz odpowiadające im klucze główne zostały automatycznie zidentyfikowane jako metryki liczbowe, co mogło skutkować możliwością ich nieprawidłowej agregacji w raportach. Z tego powodu wyłączono możliwość agregacji dla tych pól w modelu danych (Summarize by: None).

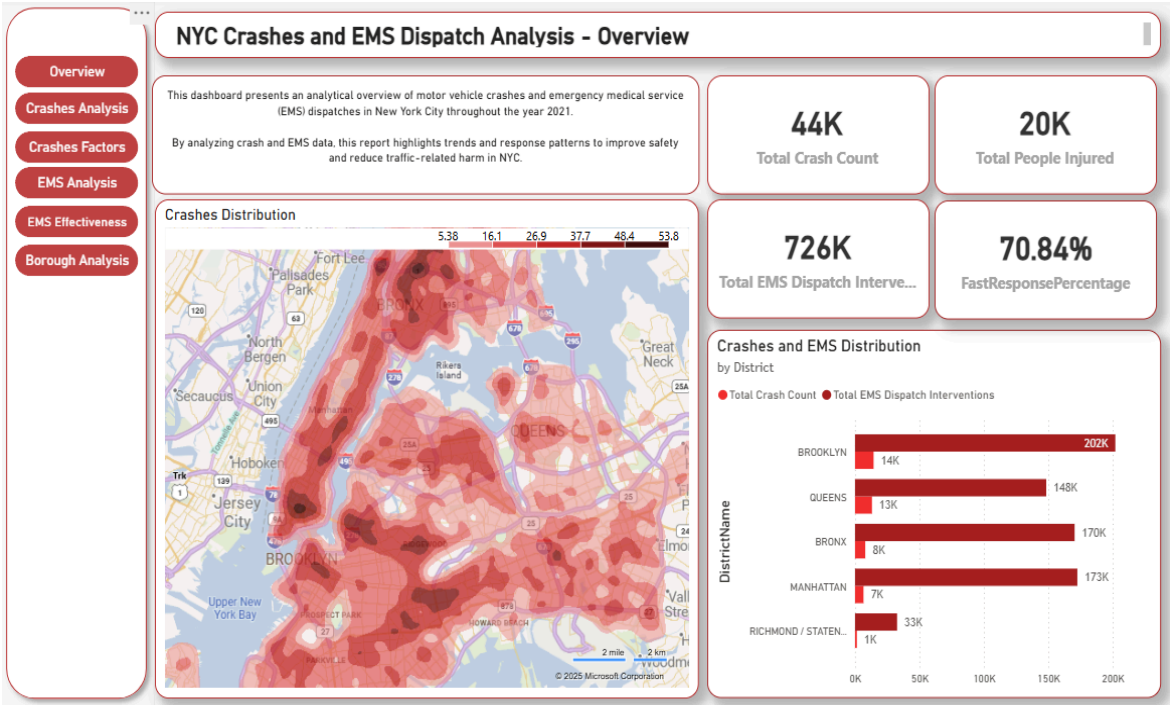
Dokonano identyfikacji kluczowych miar biznesowych, które będą wykorzystywane w analizach. Przeprowadzono również weryfikację i dostosowanie domyślnych typów agregacji tych miar, w celu zapewnienia ich zgodności z kontekstem analitycznym.

- Domyślnie wszystkie miary liczbowe zostały zidentyfikowane jako agregowane za pomocą funkcji `SUM`. W wielu przypadkach było to zgodne z oczekiwanym sposobem prezentacji danych, jednak w niektórych obszarach konieczne było zastosowanie innej logiki agregacji.

- W tabeli faktów EMSFacts, dla miar opisujących czasowe aspekty zdarzenia (takie jak: *czas aktywacji zdarzenia*, *czas dotarcia zespołu ratunkowego na miejsce*, *czas trwania interwencji*, *czas transportu do szpitala*) zastosowano agregację typu AVERAGE (średnia), aby uzyskać reprezentatywny obraz wydajności działań ratunkowych w przekroju czasu i lokalizacji.
- W tabeli CrashFacts, dla miar opisujących liczbę osób poszkodowanych lub ofiar śmiertelnych, zastosowano agregację SUM, co umożliwia analizę łącznej skali zdarzeń w wybranych przekrojach (np. czasowym, geograficznym).
- Zmieniono kategorie dla danych geograficznych w tabeli faktowej CrashFacts - Latitude i Longitude przekonwertowane zostały na typy geograficzne.

Na każdej stronie raportu zastosowano ogólny filtr do tabeli wymiarowej opisującej dzielnice zawierającą SCD2, IsValid - True

Strona raportowa 1 - Overview



Elementy:

- **Krótki opis problemu i tematu raportu**
- **Zestaw KPI cards**

  Zawiera najważniejsze dane liczbowe w skondensowanej formie: całkowita liczba wypadków, liczba osób poszkodowanych, liczba interwencji służb ratunkowych (EMS), procent interwencji poniżej 8 minut traktowanych jako standard medyczny.
- **Heatmapa wypadków – rozkład przestrzenny zdarzeń drogowych**

  Mapa prezentuje geograficzny rozkład wypadków drogowych. Kolorystyka wskazuje intensywność występowania zdarzeń – im ciemniejszy odcień, tym większe ich natężenie (Crash Count).
- **Wykres skumulowany – liczba wypadków i interwencji służb ratunkowych w podziale na dzielnice**

  Clustered bar chart przedstawia porównanie liczby wypadków oraz liczby wyjazdów służb ratunkowych (EMS) dla każdej dzielnicy.

  Zaimplementowane miary:
- Total Crash Count

```
Total Crash Count = DISTINCTCOUNT(CrashFacts[Crash_ID])
```

- Total People Injured

```
Total People Injured = SUM(CrashFacts[PeopleInjured])
```

- Total EMS Dispatch Intervations

```
Total EMS Dispatch Interventions = DISTINCTCOUNT(EMSFacts[EMS_ID])
```

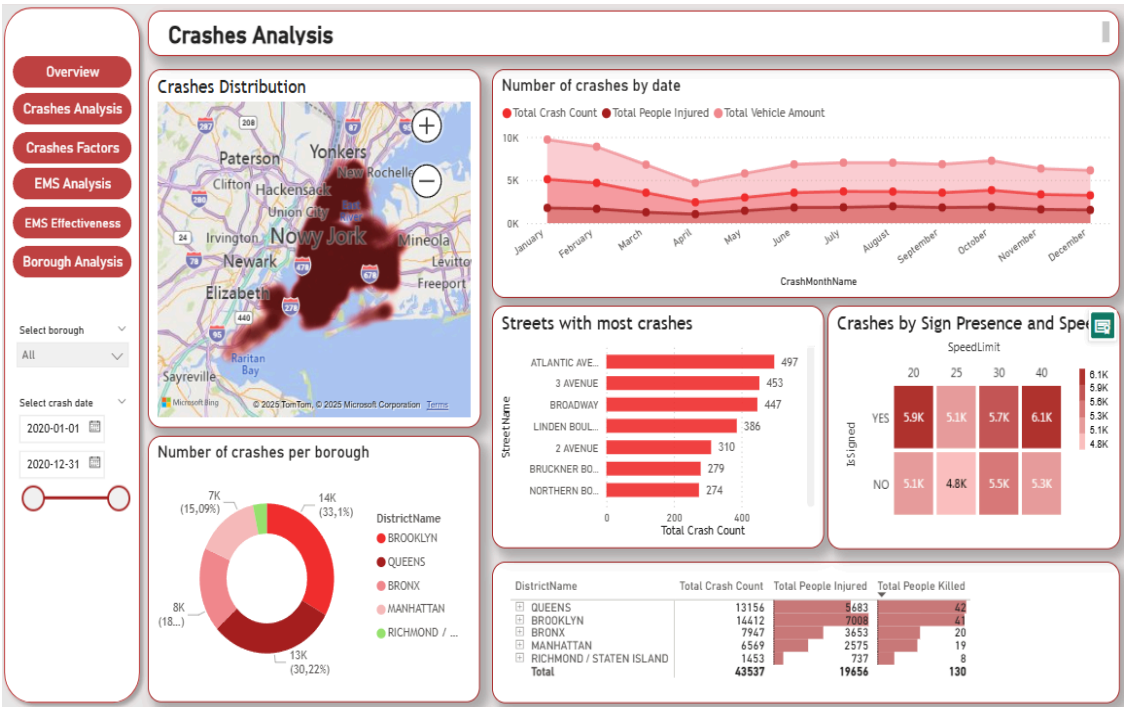
- Fast Response Percentage

```
Fast Response Percentage =
DIVIDE(
    COUNTROWS(
        FILTER(
            EMSFacts,
            EMSFacts[OnSceneArrivalDuration] < 8
        )
    ),
    COUNTROWS(EMSFacts),
    0
)
```

- Total EMS Dispatch Interventions

```
Total EMS Dispatch Interventions = COUNT(EMSFacts[EMS_ID])
```

Strona raportowa 2 - Crashes Analysis



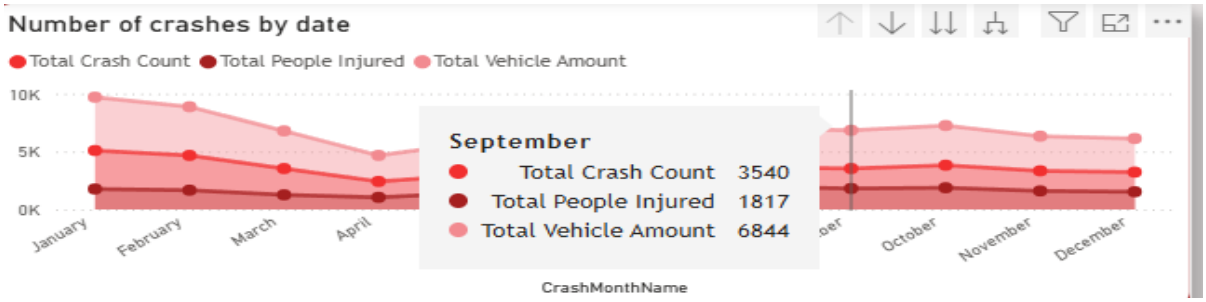
Elementy:

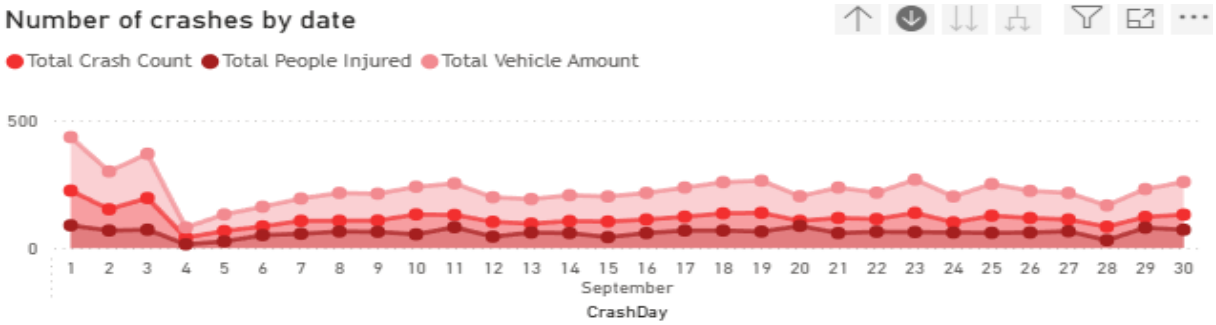
- **Interaktywna heatmapa intensywności wypadków drogowych (Crashes Distribution)**

  Mapa wizualizuje przestrzenny rozkład wypadków samochodowych. Użytkownik ma możliwość filtrowania danych według zakresu dat oraz wybranej dzielnicy. Dodatkowo, po wskazaniu konkretnej ulicy w innych wizualizacjach, mapa automatycznie prezentuje lokalizację oraz natężenie wypadków na wybranym odcinku drogi.

- **Wykres liniowy – liczba wypadków, poszkodowanych i ofiar śmiertelnych w ujęciu miesięcznym (Number of crashes by date)**

  Wykres przedstawia zmiany liczby wypadków drogowych, osób poszkodowanych oraz zabitych w podziale na miesiące. Zaimplementowana hierarchia pozwala użytkownikowi na zagłębienie się w szczegóły danego miesiąca – po rozwinięciu prezentowane są dane dzienne.

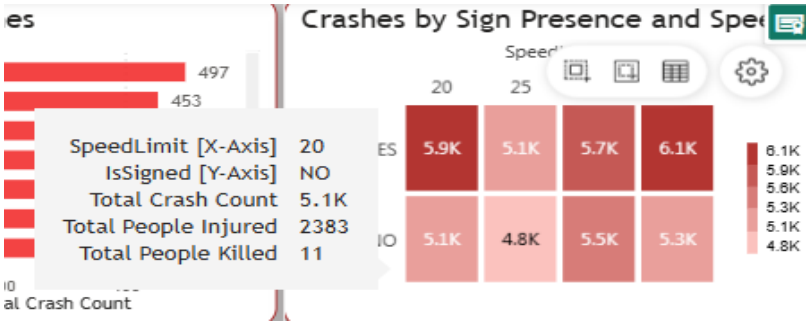




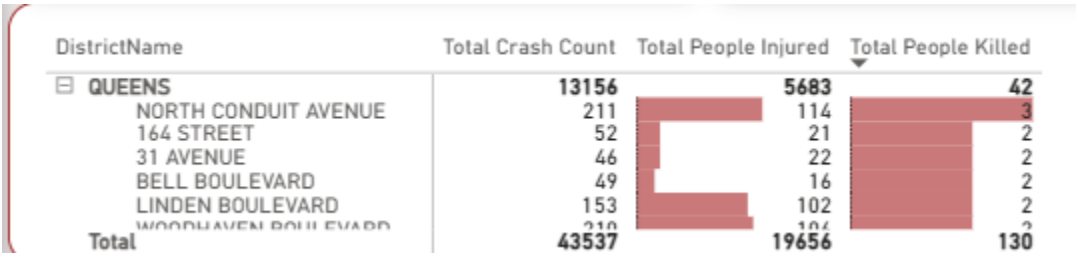
- **Donut Chart – procentowy udział dzielnic w liczbie wypadków (Number of crashes per borough)**
  Wykres typu donut prezentuje udział procentowy każdej z dzielnic w całkowitej liczbie wypadków drogowych.

- **Wykres słupkowy – ulice o największej liczbie wypadków (Streets with Most Crashes)**
  Przedstawia ranking ulic według liczby wypadków, posortowany malejąco. Możliwość przewijania pozwala na zapoznanie się z pełną listą. Dodatkowo, interaktywne podpowiedzi (tooltips) zawierają szczegółowe dane o liczbie poszkodowanych oraz ofiar śmiertelnych dla każdej ulicy.

- **Mapa korelacji – zależność liczby wypadków od obecności oznakowania drogowego i prędkości (Crashes by Sign Presence and Speed)**
  Wizualizacja pokazuje wpływ obecności znaków drogowych oraz ograniczeń prędkości na liczbę wypadków. Po najechaniu kursorem na dany punkt, wyświetlane są szczegółowe informacje: liczba wypadków, liczba osób poszkodowanych oraz zabitych.



- **Tabela podsumowująca – statystyki wypadków w podziale na dzielnice**
  Interaktywna tabela przedstawia kluczowe dane dla każdej dzielnicy: łączną liczbę wypadków liczbę osób poszkodowanych oraz liczbę ofiar śmiertelnych. Użytkownik ma możliwość rozwinięcia każdego wiersza, aby uzyskać szczegółowe statystyki dla poszczególnych ulic znajdujących się w danej dzielnicy.



DistrictName	Total Crash Count	Total People Injured	Total People Killed
QUEENS	13156	5683	42
NORTH CONDUIT AVENUE	211	114	3
164 STREET	52	21	2
31 AVENUE	46	22	2
BELL BOULEVARD	49	16	2
LINDEN BOULEVARD	153	102	2
Total	43537	19656	130

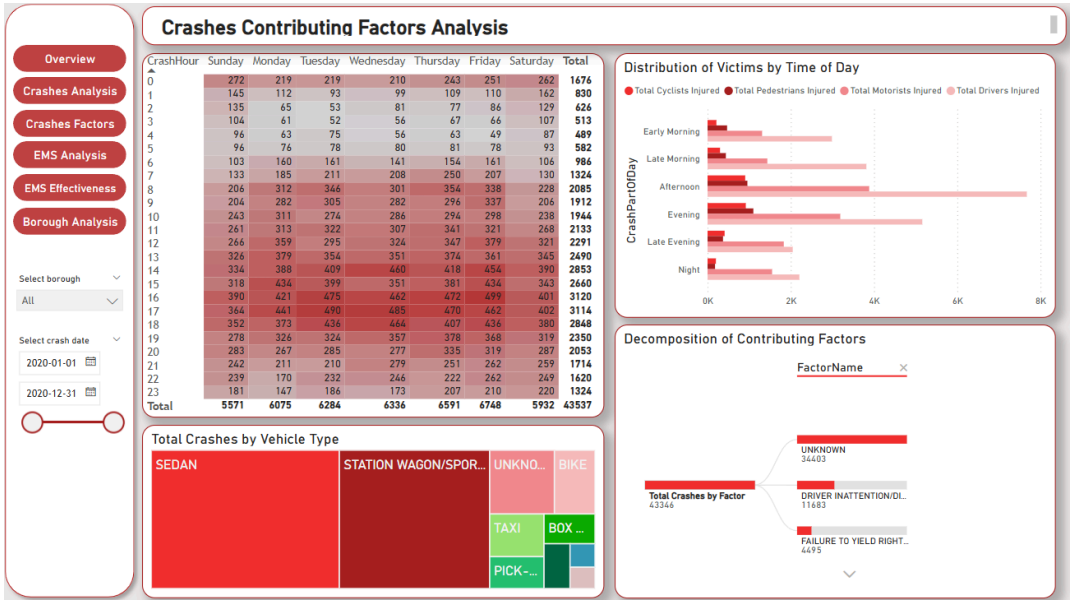
  Filtry daty oraz wyboru dzielnicy są globalnie zastosowane we wszystkich komponentach wizualizacji, z wyjątkiem wykresu typu Donut Chart, który prezentuje ogólny udział procentowy dzielnic i nie podlega filtrowaniu według konkretnej dzielnicy.
  Zaimplementowane miary (nowe, nie używane do tej pory):
- Total Vehicle Amount
  ```
  Total Vehicle Amount = SUM(CrashFacts[VehiclesAmount])
  ```
- Total People Killed
  ```
  Total People Killed= SUM(CrashFacts[PeopleKilled])
  ```

Zaimplementowana hierarchia:

- CrashDate Hierarchy (vwDimDate_Crash): CrashMonthName -> CrashDate

Kolumny CrashMonthName i CrashWeekdayName w vwDimDate_Crash zostały posortowane odpowiednio przez CrashMonth i CrashWeekdayNumber.

Strona raportowa 3 - Crashes Contributing Factors

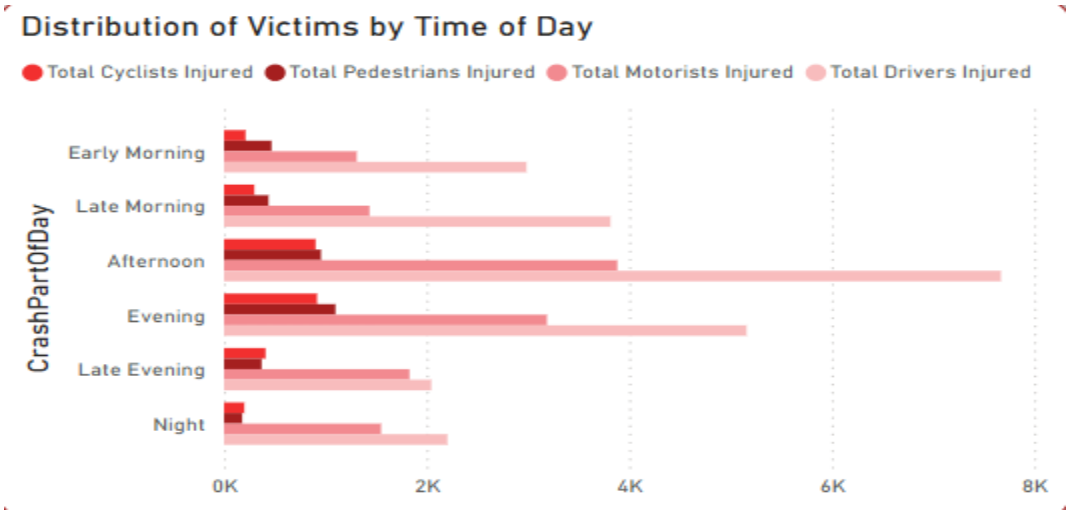


Elementy:

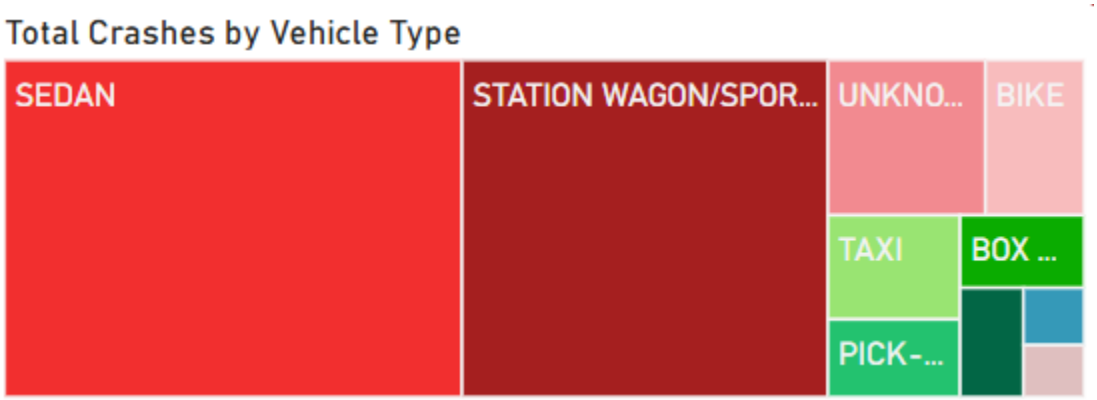
- **Macierz zależności – dzień tygodnia vs godzina zdarzenia**
  Dwuwymiarowa macierz przedstawia liczbę wypadków w zależności od dnia tygodnia i godziny. Komórki są kolorystycznie cieniowane w zależności od natężenia zdarzeń – im intensywniejszy kolor, tym większa liczba wypadków. Dodatkowe kolumny i wiersze sumaryczne prezentują łączną liczbę zdarzeń dla każdego dnia tygodnia oraz dla poszczególnych godzin.

CrashHour	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Total
0	272	219	219	210	243	251	262	1676
1	145	112	93	99	109	110	162	830
2	135	65	53	81	77	86	129	626
3	104	61	52	56	67	66	107	513
4	96	63	75	56	63	49	87	489
5	96	76	78	80	81	78	93	582
6	103	160	161	141	154	161	106	986
7	133	185	211	208	250	207	130	1324
8	206	312	346	301	354	338	228	2085
9	204	282	305	282	296	337	206	1912
10	243	311	274	286	294	298	238	1944
11	261	313	322	307	341	321	268	2133
12	266	359	295	324	347	379	321	2291
13	326	379	354	351	374	361	345	2490
14	334	388	409	460	418	454	390	2853
15	318	434	399	351	381	434	343	2660
16	390	421	475	462	472	499	401	3120
17	364	441	490	485	470	462	402	3114
18	352	373	436	464	407	436	380	2848
19	278	326	324	357	378	368	319	2350
20	283	267	285	277	335	319	287	2053
21	242	211	210	279	251	262	259	1714
22	239	170	232	246	222	262	249	1620
23	181	147	186	173	207	210	220	1324
Total	5571	6075	6284	6336	6591	6748	5932	43537

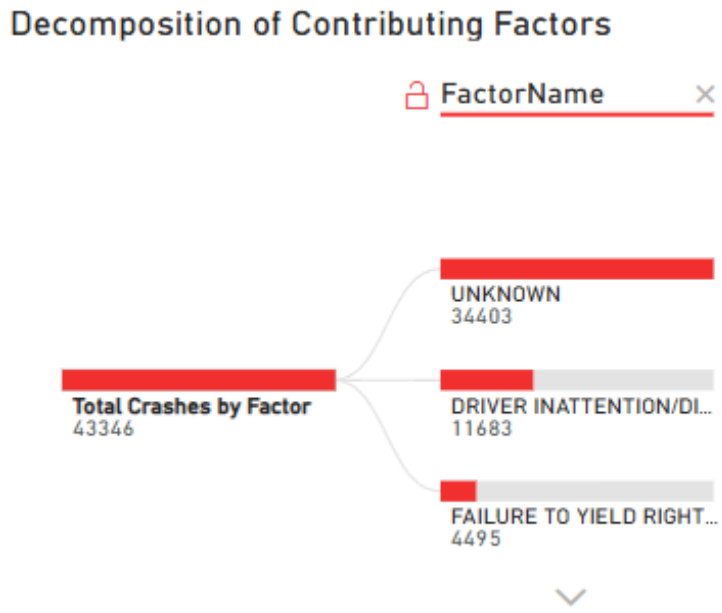
- **Wykres skumulowany z podziałem – liczba poszkodowanych wg typu uczestnika ruchu i pory dnia**
  Clustered stacked bar chart przedstawia liczbę poszkodowanych osób w rozbiciu na cztery kategorie: rowerzyści, piesi, motocykliści oraz kierowcy. Dane pogrupowane są według pory dnia, co pozwala zaobserwować różnice w strukturze uczestników zdarzeń w zależności od momentu doby



- **Tree map – 10 najczęściej uczestniczących w wypadkach typów pojazdów**
  Wizualizacja typu Tree Map przedstawia dziesięć typów pojazdów najczęściej biorących udział w zdarzeniach drogowych. Każdy prostokąt reprezentuje konkretny typ pojazdu, a jego wielkość odzwierciedla liczbę wypadków z jego udziałem.



- **Drzewo dekompozycji – analiza przyczyn wypadków (Contributing Factors)**
  Decomposition Tree umożliwia interaktywną eksplorację przyczyn wypadków drogowych.



Zaimplementowane miary:

- Total Crashes by Vehicle

```
Total Crashes by Vehicle =
COUNTROWS(
    SUMMARIZE(
        CrashVehicleBridge,
        CrashVehicleBridge[Crash_ID]
    )
)
```

- Total Crashes by Factor

```
Total Crashes by Factor =
COUNTROWS(
    SUMMARIZE(
        CrashFactorBridge,
        CrashFactorBridge[Crash_ID]
    )
)
```

- Total Cyclists Injured

```
Total Cyclists Injured = SUM(CrashFacts[CyclistsInjured])
```

- Total Pedestrians Injured

```
Total Pedestrians Injured = SUM(CrashFacts[PedestriansInjured])
```

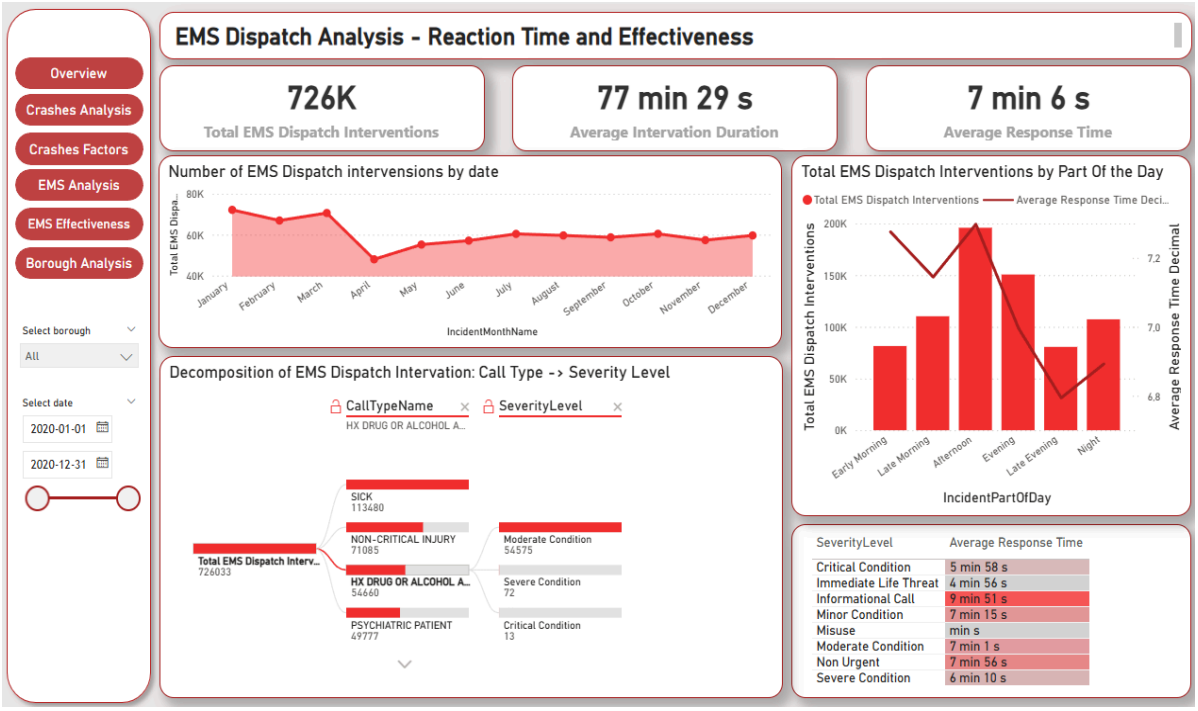
- Total Motorists Injured

```
Total Motorists Injured = SUM(CrashFacts[MotoristInjured])
```

- Total Drivers Injured

```
Total Drivers Injured = CrashFacts[TotalCrashCount] - CrashFacts[Total Cyclists Injured] - CrashFacts[Total Motorists Injured] - CrashFacts[Total Pedestrians Injured]
```

Strona raportowa 4 - EMS Dispatch Analysis



Elementy:

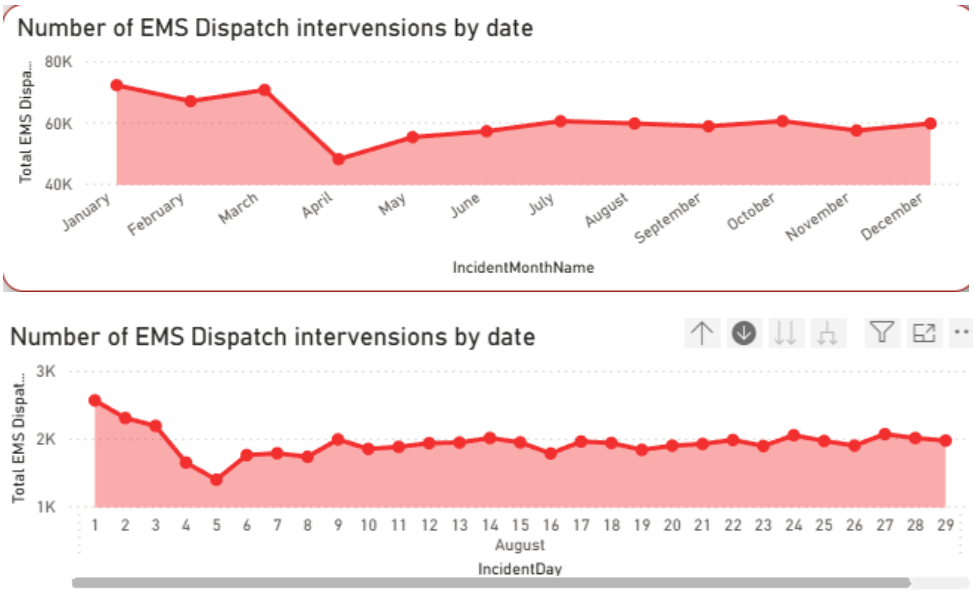
- **KPI Cards:**

  Total EMS Dispatch Interventions - liczba interwencji służb ratunkowych w analizowanym okresie czasu

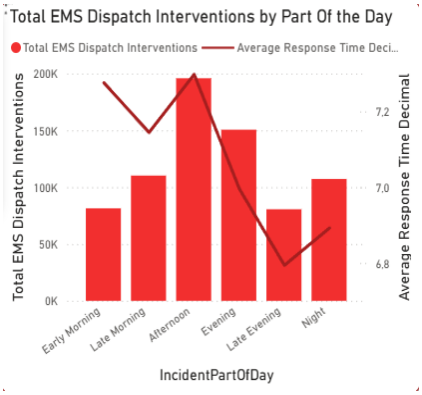
  Average Intervention Duration - średni czas trwania interwencji, czas liczony od aktywacji zgłoszenia do jego zamknięcia w systemie, liczony w minutach.

  Average Response Time - średni czas przybycia służb ratunkowych na miejsce zdarzenia, liczony w minutach.

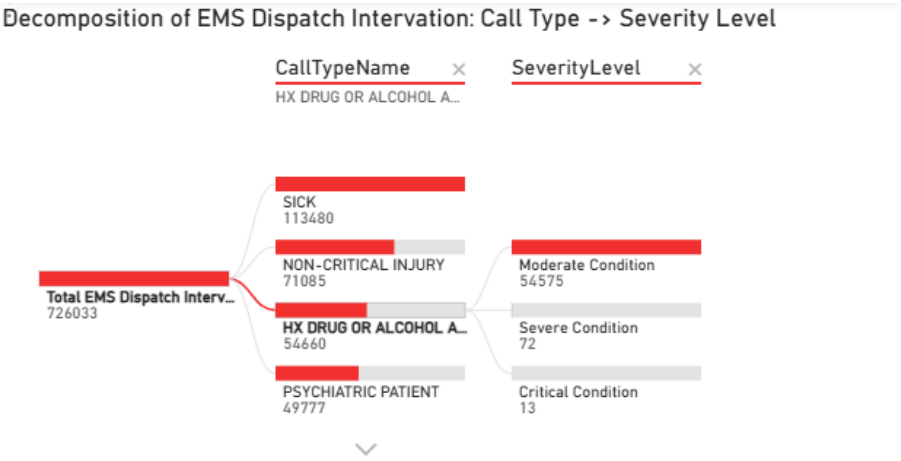
- **Wykres liniowy - liczba interwencji w zależności od miesiąca.**
  Dzięki zastosowaniu hierarchii dat możliwa jest interaktywna analiza w trybie drill down do poziomu konkretnych dni.



- **Line and clustered column chart ilości interwencji i średniego czas reakcji**
  Wizualizacja typu Line and Clustered Column Chart przedstawia liczbę interwencji w zależności od pory dnia, zestawioną z średnim czasem przybycia służb ratunkowych na miejsce zdarzenia. Słupki reprezentują liczbę zdarzeń w poszczególnych przedziałach dnia, natomiast linia ilustruje średni czas reakcji dla każdego z tych przedziałów. Możliwość drill down do poziomu konkretnego zakresu godzin.



- **Decomposition Tree – interwencje służb ratunkowych według typu zdarzenia i poziomu ciężkości**
  Przedstawia strukturę interwencji służb ratunkowych z podziałem na typy zgłoszeń (Call Type Name), a następnie na przypisane im poziomy ciężkości (Severity Level). Każda gałąź drzewa reprezentuje kolejny poziom szczegółowości – od ogólnego typu zdarzenia po jego klasyfikację pod względem powagi.



- **Tabela – Poziom ciężkości zdarzenia a średni czas reakcji**
  Wizualizacja w formie tabeli przedstawia zestawienie poziomów ciężkości zdarzeń (Severity Level) z odpowiadającym im średnim czasem reakcji służb ratunkowych (Average Response Time).

SeverityLevel	Average Response Time
Critical Condition	5 min 58 s
Immediate Life Threat	4 min 56 s
Informational Call	9 min 51 s
Minor Condition	7 min 15 s
Moderate Condition	7 min 1 s
Non Urgent	7 min 56 s
Severe Condition	6 min 10 s

Zaimplementowane miary:

- Average Intervention Duration Decimal

```
Average Intervention Duration Decimal = AVERAGE(EMSFacts[InterventionDuration])
```

- Average Intervention Duration

```
Average Intervention Duration =
VAR TotalSeconds = [Average Intervention Duration Decimal] * 60
VAR Minutes = INT(TotalSeconds / 60)
VAR Seconds = ROUND(MOD(TotalSeconds, 60), 0)
RETURN
Minutes & " min " & Seconds & " s"
```

- Average Response Time Decimal

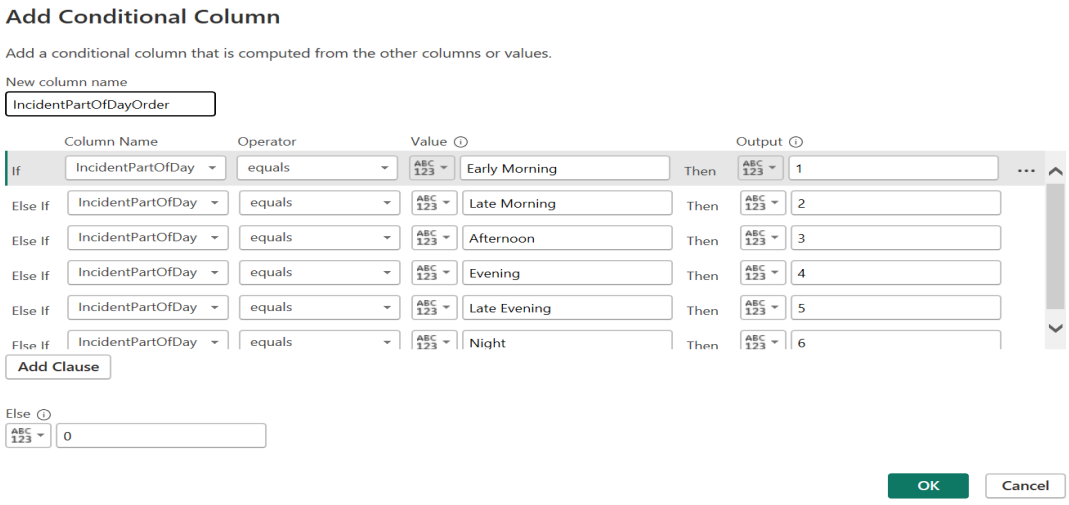
```
Average Response Time Decimal = AVERAGE(EMSFacts[OnSceneArrivalDuration])
```

- Average Response Time

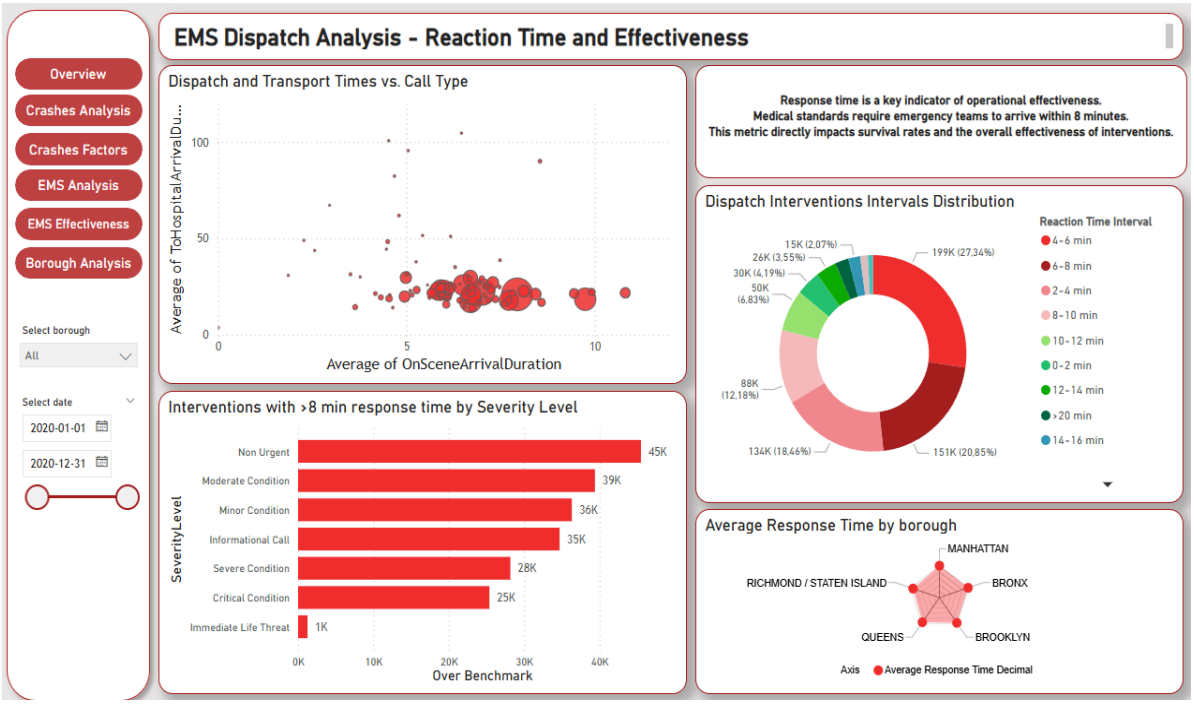
```
Average Response Time =
VAR TotalSeconds = [Average Response Time Decimal] * 60
VAR Minutes = INT(TotalSeconds / 60)
VAR Seconds = ROUND(MOD(TotalSeconds, 60), 0)
RETURN
Minutes & " min " & Seconds & " s"
```

Hierarchie:

- IncidentDate Hierarchy: IncidentMonthName (posortowana za pomocą IncidentMonth - kolumna numeryczna), IncidentDay
- IncidentPartOfDay Hierarchy: IncidentPartOfDay (posortowana za pomocą stworzonej kolumny IncidentPartOfDayOrder), IncidentHour

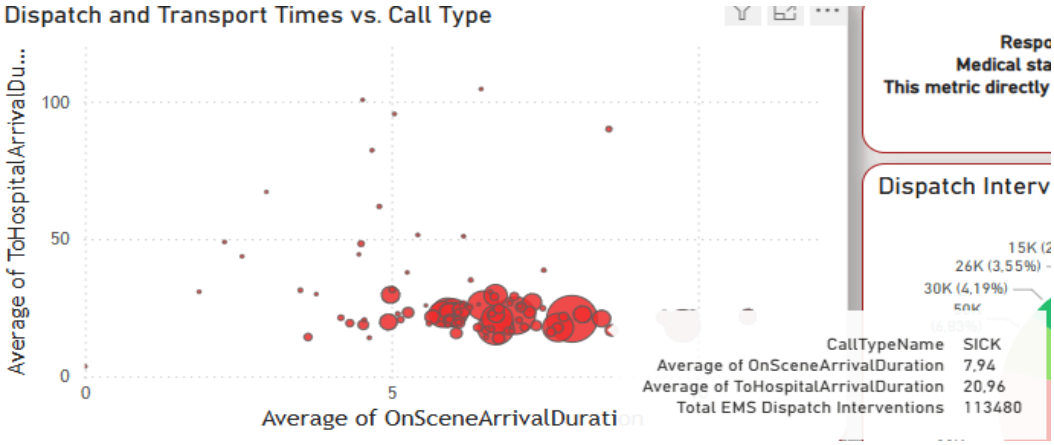


Strona raportowa 5 - EMS Dispatch Effectiveness



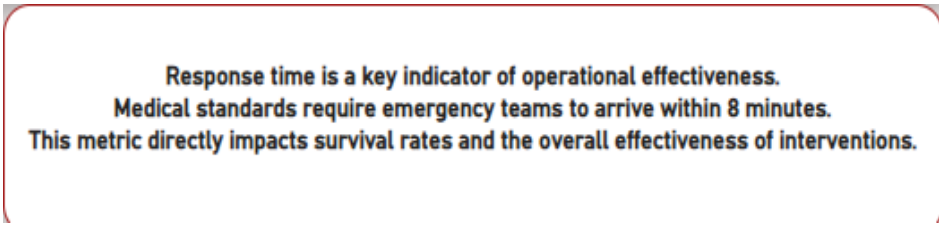
Elementy:

- **Scatter Plot – rozkład średnich czasów reakcji według typu zgłoszenia**
  Przedstawia zależność pomiędzy czasem przyjazdu służb na miejsce zdarzenia a czasem transportu do szpitala, w podziale na typy zgłoszeń (Call Type). Wielkość punktów odpowiada liczbie interwencji dla danego typu zdarzenia. Po najechaniu na punkt użytkownik uzyskuje szczegółowe informacje w tooltipie, takie jak typ zgłoszenia, dokładne wartości czasów i liczba przypadków.

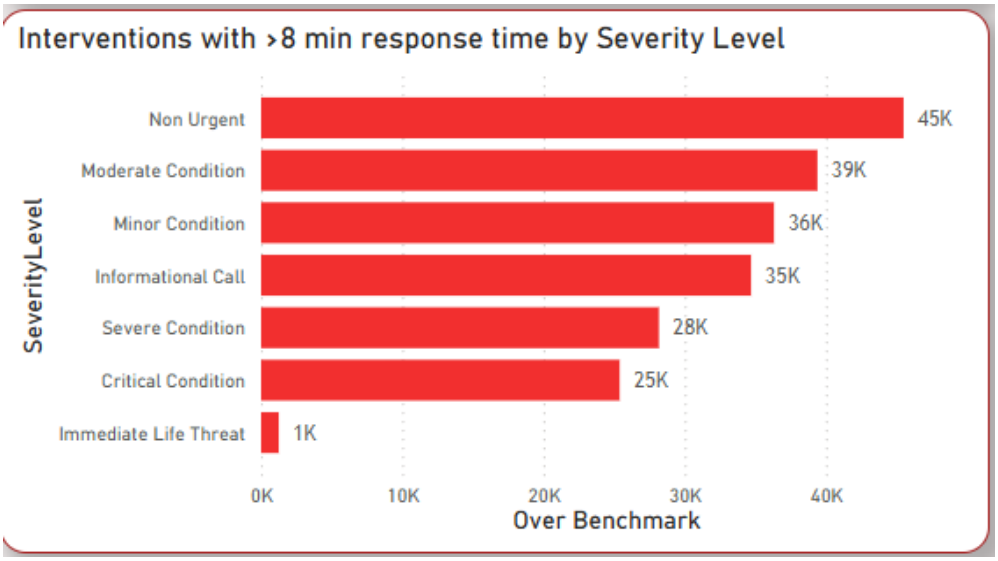


- **Text Box – standardowy czas reakcji EMS**

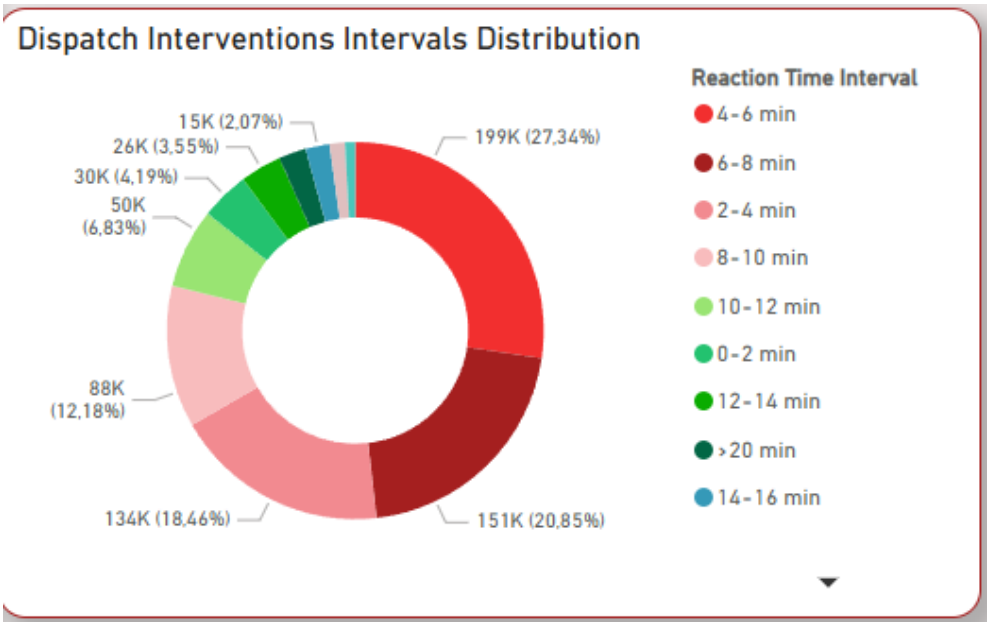
Zawiera opis kluczowego wskaźnika efektywności działań ratowniczych, jakim jest czas dotarcia na miejsce zdarzenia.



- **Clustered Bar Chart – przekroczenie standardu 8 minut według poziomu ciężkości**
  Przedstawia liczbę interwencji, w których czas reakcji przekroczył standard medyczny 8 minut, w podziale na poziomy ciężkości zdarzenia (Severity Level).

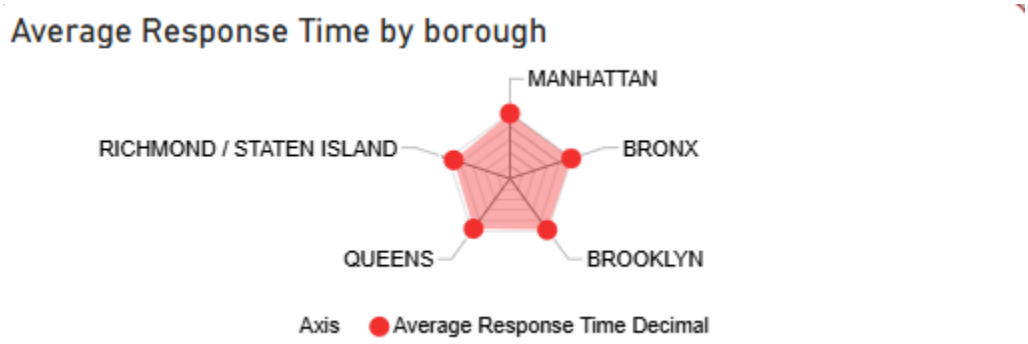


- **Donut Chart – rozkład czasów dotarcia na miejsce zdarzenia**
  Prezentuje rozkład interwencji ratunkowych w zdefiniowanych przedziałach czasowych (co 2 minuty: 0-2 min, 2-4min, …., 18-20min, >20min) przyjazdu na miejsce zdarzenia.



- **Radar Chart – średni czas reakcji wg dzielnicy**

  Przedstawia średni czas reakcji służb ratunkowych w podziale na dzielnice miasta.Każde ramię wykresu reprezentuje jedną dzielnicę, a odległość od środka odpowiada wartości średniego czasu.



Zaimplementowane miary:

- Over Benchmark

```
Over Benchmark =
SUMX(
    EMSFacts,
    IF(EMSFacts[Average Response Time Decimal] > 8, 1, 0)
)
```

Utworzona została również kolumna definiująca interwały czasowe.

New column name

Reaction Time Interval

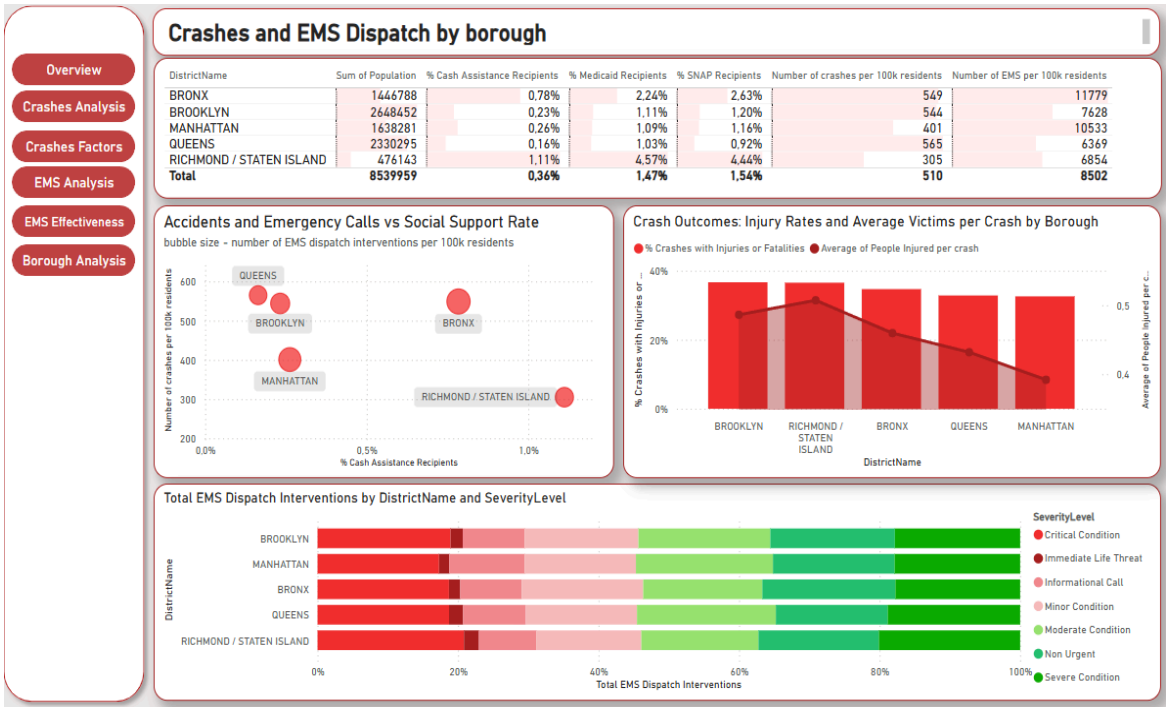
	Column Name	Operator	Value		Output
If	OnSceneArrivalDu...	is less than	2	Then	0-2 min
Else If	OnSceneArrivalDu...	is less than	4	Then	2-4 min
Else If	OnSceneArrivalDu...	is less than	6	Then	4-6 min
Else If	OnSceneArrivalDu...	is less than	8	Then	6-8 min
Else If	OnSceneArrivalDu...	is less than	10	Then	8-10 min
Else If	OnSceneArrivalDu...	is less than	12	Then	10-12 min

Add Clause

Else

>20 min

Strona raportowa 6 - Crashes and EMS Dispatch by Borough



Elementy:

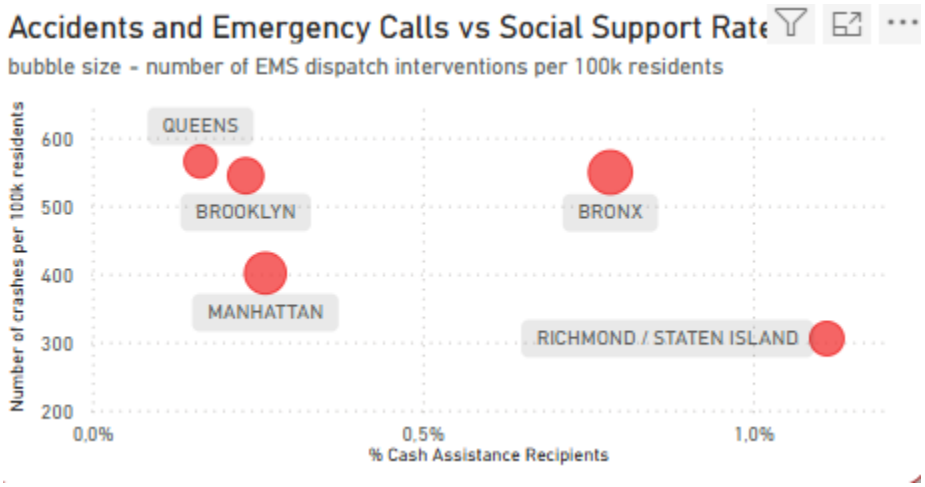
- **Tabela podsumowująca charakterystyki dzielnic**

  Tabela prezentuje kluczowe wskaźniki społeczno-demograficzne oraz dane dotyczące zdarzeń ratunkowych dla poszczególnych dzielnic. Zawiera informacje o liczbie mieszkańców, odseteku osób korzystających z pomocy finansowej, pomocy medycznej, programu SNAP, liczbie wypadków drogowych oraz interwencji EMS w przeliczeniu na 100 000 mieszkańców.

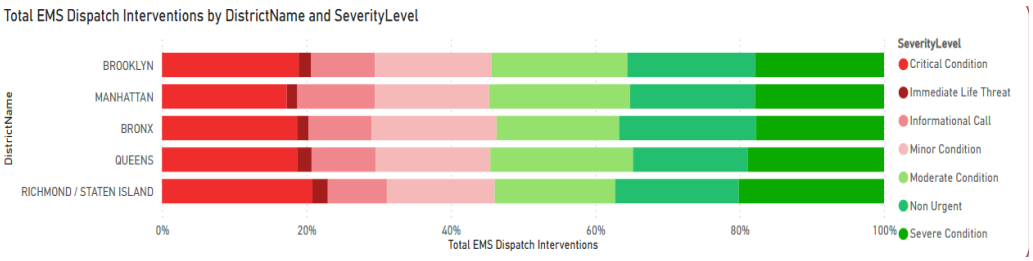
DistrictName	Sum of Population	% Cash Assistance Recipients	% Medicaid Recipients	% SNAP Recipients	Number of crashes per 100k residents	Number of EMS per 100k residents
BRONX	1446788	0,78%	2,24%	2,63%	549	11779
BROOKLYN	2648452	0,23%	1,11%	1,20%	544	7628
MANHATTAN	1638281	0,26%	1,09%	1,16%	401	10533
QUEENS	2330295	0,16%	1,03%	0,92%	565	6369
RICHMOND / STATEN ISLAND	476143	1,11%	4,57%	4,44%	305	6854
Total	**8539959**	**0,36%**	**1,47%**	**1,54%**	**510**	**8502**

- **Scatter Plot – zależność między pomocą społeczną a liczbą zdarzeń ratunkowych i wypadków**

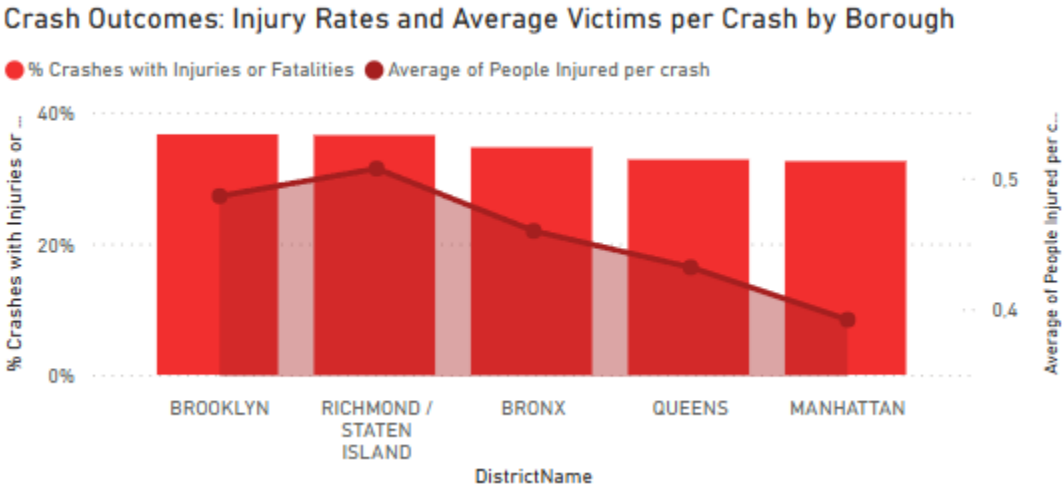
  Przedstawia zależność pomiędzy wskaźnikiem wsparcia społecznego (% Cash Assistance Recipients) a liczbą wypadków drogowych i interwencji EMS, w podziale na dzielnice. Każdy punkt reprezentuje jedną dzielnicę, a wielkość bąbelka odzwierciedla liczbę interwencji EMS.



- **100% Stacked Bar Chart – struktura poziomów ciężkości interwencji według dzielnicy**
  Przedstawia udział poszczególnych poziomów ciężkości zdarzeń (Severity Level) w ogólnej liczbie interwencji dla każdej dzielnicy. Dzięki normalizacji do 100% umożliwia porównanie struktury typów interwencji niezależnie od ich łącznej liczby w danej dzielnicy.



- **Line and clustered column plot - udział wypadków z ofiarami oraz średnia liczba poszkodowanych na wypadek per dzielnica**
  Przedstawia odsetek wypadków drogowych skutkujących obrażeniami lub ofiarami śmiertelnymi (słupki) oraz średnią liczbę osób poszkodowanych na jeden wypadek (linia) per dzielnica.



Zaimplementowane miary:

- % Cash Assistance Recipients

```
% Cash Assistance Recipients =
DIVIDE(
    SUM('DistrictDetailsDim'[TotalCashAssistanceRecipients]),
    SUM('DistrictDetailsDim'[Population]),
    0
)
```

- % Medicaid Recipients

```
% Medicaid Recipients =
DIVIDE(
    SUM('DistrictDetailsDim'[TotalMedicaidRecipients]),
    SUM('DistrictDetailsDim'[Population]),
    0
)
```

- % SNAP Recipients

```
% SNAP Recipients =
DIVIDE(
    SUM('DistrictDetailsDim'[TotalSNAPRecipients]),
    SUM('DistrictDetailsDim'[Population]),
    0
)
```

- Number of crashes per 100k residents

```
Number of crashes per 100k residents =
ROUND(
    DIVIDE(
        COUNTROWS(CrashFacts),
        SUM('DistrictDetailsDim'[Population]) / 100000,
        0
    ),
    0
)
```

- Number of EMS per 100k residents

```
Number of EMS per 100k residents =
ROUND(
    DIVIDE(
        COUNTROWS(EMSFacts),
        SUM('DistrictDetailsDim'[Population]) / 100000,
        0
    ),
    0
```

```
)
```

- % Crashes with Injuries or Fatalities

```
% Crashes with Injuries or Fatalities =
DIVIDE(
    CALCULATE(
        COUNTROWS(CrashFacts),
        CrashFacts[PeopleInjured] > 0 || CrashFacts[PeopleKilled] > 0
    ),
    COUNTROWS(CrashFacts),
    0
)
```

- Average People Injured per crash

```
Average of People Injured per crash = AVERAGE(CrashFacts[PeopleInjured])
```

Podsumowanie prac

Podczas projektu udało się przeprowadzić kompleksowy proces ETL uwzględniający integrację, transformację i czyszczenie danych źródłowych oraz wykorzystanie ich w wypełnieniu hurtowni wykorzystanej następnie do stworzenia raportu. Mimo napotkanych podczas procesu przeszkód, takich jak niespójności, braki czy brak dopasowań w danych, udało się znaleźć rozwiązania wszystkich tych problemów zapewniając końcowo wysoki standard danych i raportu. Wszystkie zaplanowane we wstępnej dokumentacji postulaty, z wyjątkiem analizy natężenia ruchu drogowego, zostały spełnione i zaprezentowane w niniejszym sprawozdaniu oraz podczas testów. Dla celów biznesowych stworzony został także szczegółowy raport umożliwiający analizę interesujących obszarów i wzorców związanych z wypadkami oraz interwencjami pogotowia ratunkowego pozwalający między innymi : zidentyfikować okresy czasowe, w których warto wzmożyć gotowość służb ratunkowych oraz dzielnice, w których pomoc może okazać się szczególnie potrzebna. Dokument umożliwia analizę zdarzeń z różnych perspektyw, co zapewnia szerokie pole korzyści biznesowych.

Podział prac w zespole

1. Wstępna dokumentacja wraz z planem modelu hurtowni i znalezieniem danych - Kasia i Zuzia
2. Stworzenie struktury hurtowni danych oraz bazy pośredniej w narzędziu SQL - Kasia
3. Przeprowadzenie procesu ETL skutkującego w wypełnieniu zaprojektowanej hurtowni danymi oraz deployment pakietu - Kasia
4. Przeprowadzenie testów za pomocą skryptów SQL mających na celu zbadanie spójność, poprawność i struktury danych wczytanych do hurtowni - Kasia
5. End to End test - Kasia i Zuzia
6. Raport i testy warstwy raportowej - Zuzia
7. Przygotowanie finalnego dokumentu opisującego poszczególne etapy projektu oraz prezentacja końcowa - Kasia i Zuzia