

Projekt Hurtowni Danych: Analiza Wypadków Drogowych

Mikołaj Mróz, Tymoteusz Urban



Wprowadzenie

- Stworzenie hurtowni danych, która umożliwi analizę wypadków drogowych z uwzględnieniem danych pogodowych oraz technicznych pojazdów
- Hurtownia będzie przydatna dla instytucji takich jak:
 - rząd
 - policja
 - firmy ubezpieczeniowe
 - firmy transportowe



Koncepcja i cel biznesowy

- Rząd i Policja: "Identyfikacja niebezpiecznych miejsc, prowadzenie częstszych kontroli modeli pojazdów powodujących wypadki."
- Firmy Ubezpieczeniowe: "Ocena ryzyka wypadków w określonych warunkach pogodowych, lepsze ustalanie stawek ubezpieczeniowych."
- Planowanie Infrastruktury: "Lepsze planowanie dróg, sygnalizacji świetlnej i infrastruktury drogowej."
- Firmy Transportowe: "Optymalizacja tras z uwzględnieniem warunków pogodowych, minimalizacja ryzyka opóźnień i wypadków."

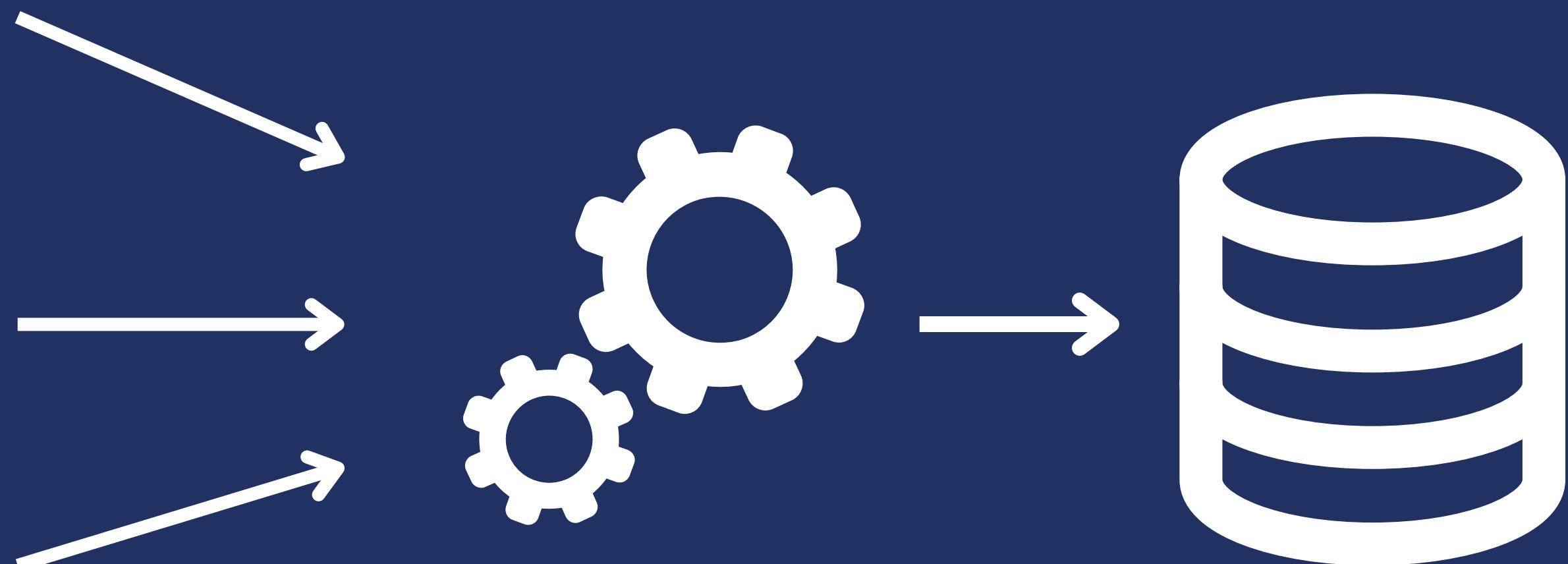
Koncepcja i cel biznesowy

Dane pogodowe

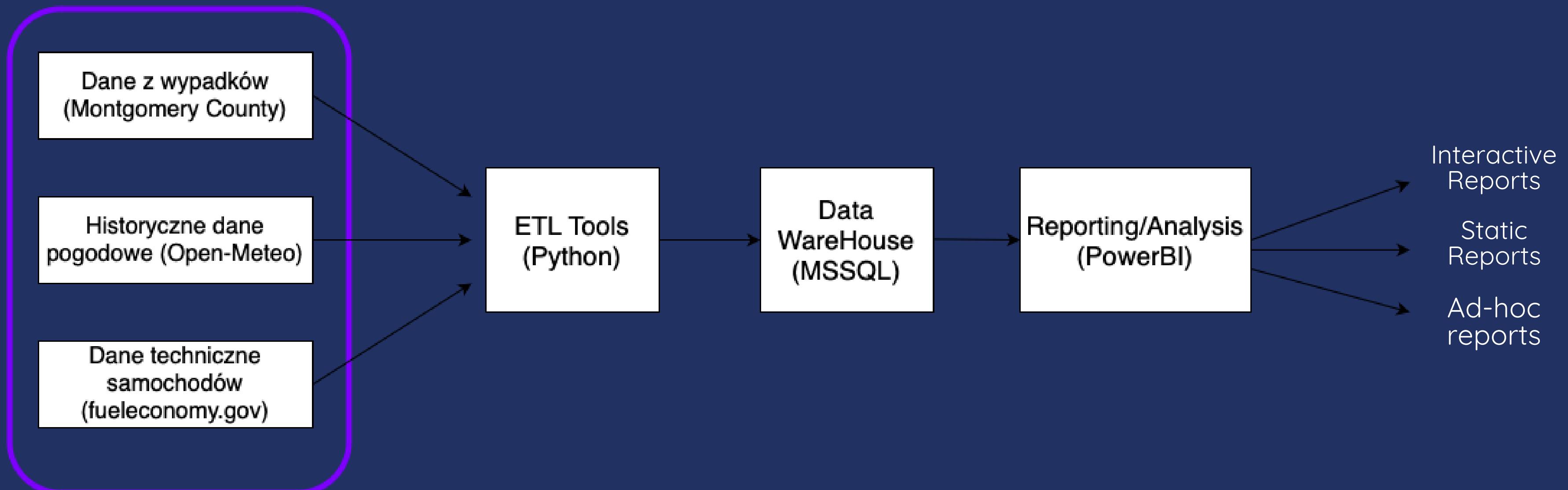
Dane dotyczące
wypadków
drogowych w
Hrabstwie
Montgomery

Dane techniczne
pojazdów

Dane
geograficzne



Architektura systemu

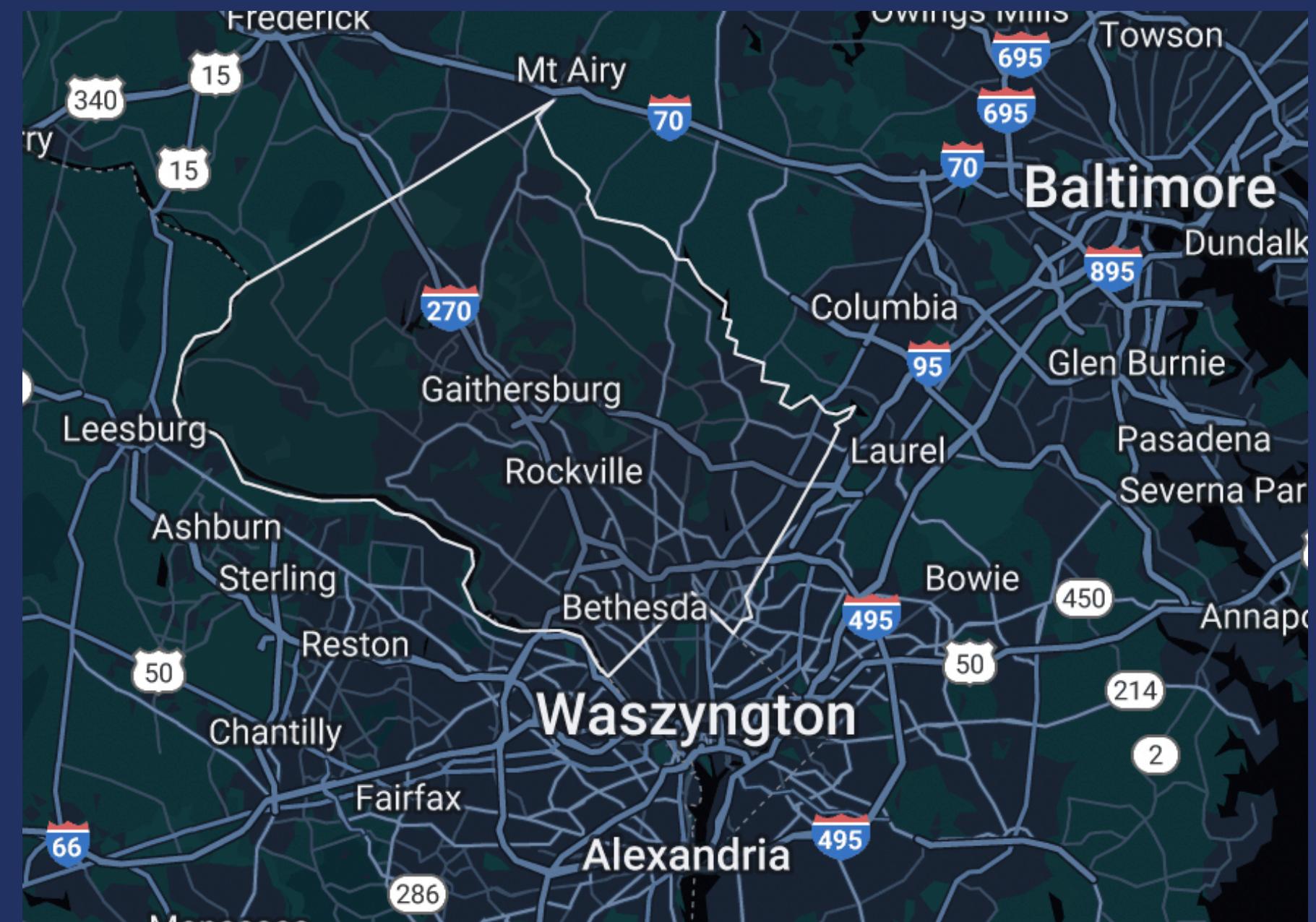


Źródła danych

Dane dotyczące wypadków w hrabstwie Montgomery

Format: CSV

Częstotliwość odświeżania:
Co miesiąc



Źródła danych

Dane dotyczące pogody z darmowego API Open-Meteo

Format: JSON

Częstotliwość odświeżania:
Co godzinę



Źródła danych

Dane techniczne
samochodów

Format: CSV

Częstotliwość odświeżania:
Nieregularnie, co kilka dni



Źródła danych

Dane geograficzne

Format: CSV

Częstotliwość odświeżania:
Nigdy



Struktura Hurtowni Danych

Tabele Faktów:

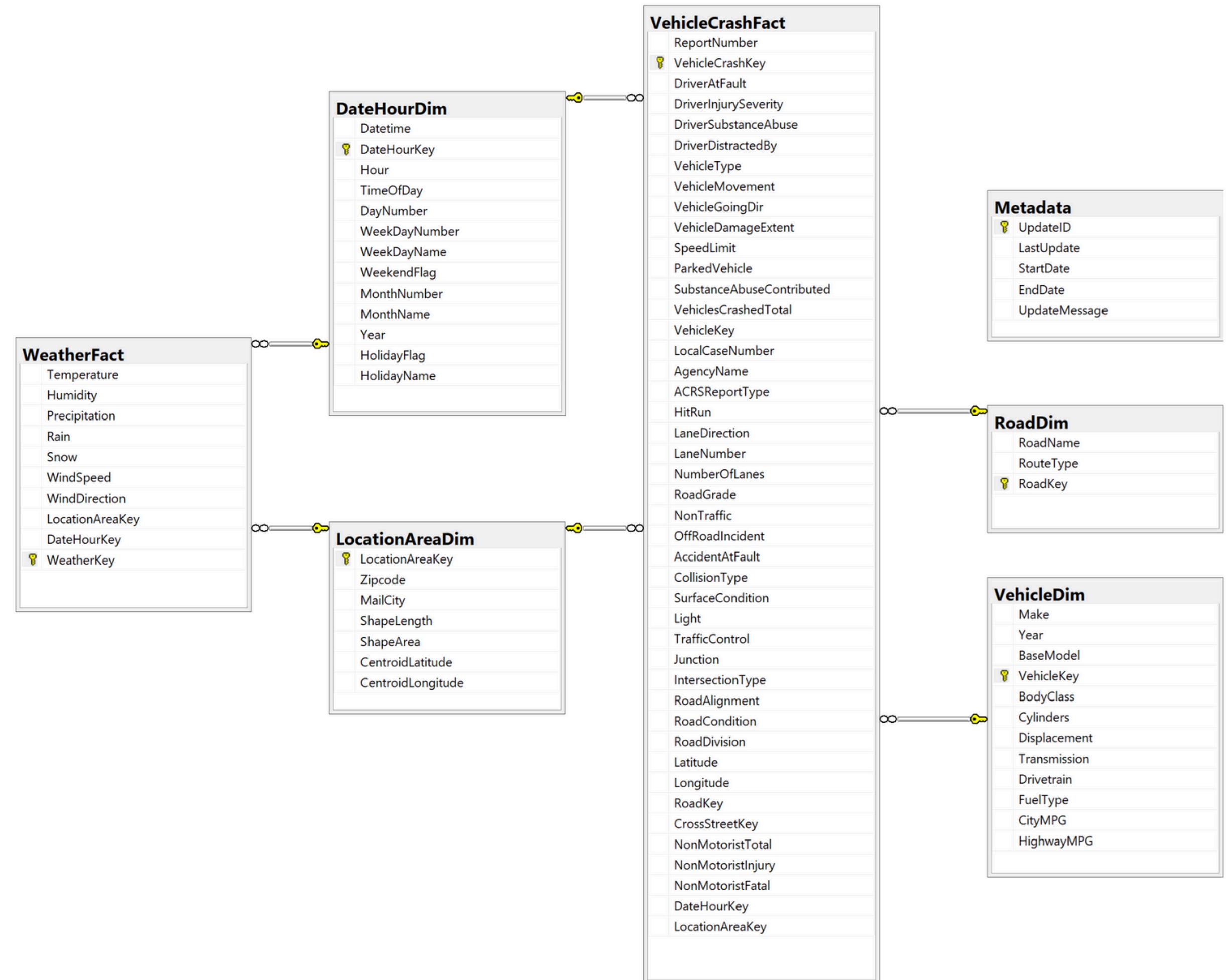
- VehicleCrashFact: Udział pojedynczego pojazdu w wypadku.
- WeatherFact: Dane pogodowe dla określonych obszarów i czasów.

Wymiary:

- VehicleDim: Dane techniczne pojazdu.
- RoadDim: Nazwa i typ drogi.
- DateHourDim: Data, godzina, atrybuty przydatne do analizy.
- LocationAreaDim: Obszar powiązany z długością i szerokością miejsca wypadku.



Diagram hurtowni danych



Model ETL

Struktura aplikacji

Dla każdej tabeli oddzielny moduł a ponadto:

etl.py
config.py
insertion.py
utils.py
tests.py

/static
.env

ETL

crash_data
drivers_data
nonmotorists_data
vehicles_data
road_data
weather_data
datehour_data
location_data
merged_data

extract_data()
transform_data()
join_data()
merge_data()
load_data()

Pipeline

load_static_files()
check_last_update()
calculate_update_range()
ETL()
ETL.extract_data(range)
ETL.transform_data()
ETL.join_data()
ETL.load_data()
update_metadata()

Model ETL

config.py

Config

DWH_INITIALIZATION
DEBUG
N_RETRIES
...

insertion.py

```
query = f"""
BEGIN TRY
    {insert_statement}
END TRY
BEGIN CATCH
    IF ERROR_NUMBER() = 2601 OR
        ERROR_NUMBER() = 2627
    BEGIN
        PRINT 'Duplicate key, skipping row'
    END
    ELSE
    BEGIN
        THROW;
    END
END CATCH
"""
```

utils.py

Static

BRANDS_DICT
MODELS_DICT
AREA_MAPPER
ZIPCODES

Model ETL

DateHourDim	
Column Name	Data Type
Datetime	datetime2(7)
 DateHourKey	bigint
Hour	tinyint
TimeOfDay	nvarchar(50)
DayNumber	int
WeekDayNu...	tinyint
WeekDayNa...	nvarchar(50)
WeekendFlag	bit
MonthNumb...	tinyint
MonthName	nvarchar(50)
Year	int
HolidayFlag	bit
HolidayName	nvarchar(50)

EKSTRAKCJA

- generowanie wierszy - każdy wiersz to kolejna godzina

TRANSFORMACJE

- wyciągnięcie atrybutów daty
- utworzenie flag
- utworzenie klucza głównego

Model ETL

LocationAreaDim	
Column Name	Data Type
LocationAreaKey	bigint
Zipcode	smallint
MailCity	nvarchar(50)
ShapeLength	float
ShapeArea	float
CentroidLatitude	float
CentroidLongitude	float

EKSTRAKCJA

- jednorazowe załadowanie z pliku

TRANSFORMACJE

- wyliczenie punktów centralnych
- dodanie pustej lokalizacji
- utworzenie klucza głównego

Model ETL

WeatherFact	
Column Name	Data Type
Temperature	float
Humidity	float
Precipitation	float
Rain	float
Snow	float
WindSpeed	float
WindDirection	float
LocationAreaKey	bigint
DateHourKey	bigint
 WeatherKey	bigint

EKSTRAKCJA

- zapytanie do API dla każdej godziny dla każdej lokalizacji

TRANSFORMACJE

- przesunięcie czasu
- utworzenie klucza głównego i kluczy obycz

Model ETL

VehicleDim	
Column Name	Data Type
Make	nvarchar(50)
Year	smallint
BaseModel	nvarchar(50)
 VehicleKey	bigint
BodyClass	nvarchar(50)
Cylinders	int
Displacement	float
Transmission	nvarchar(50)
Drivetrain	nvarchar(50)
FuelType	nvarchar(50)
CityMPG	tinyint
HighwayMPG	tinyint

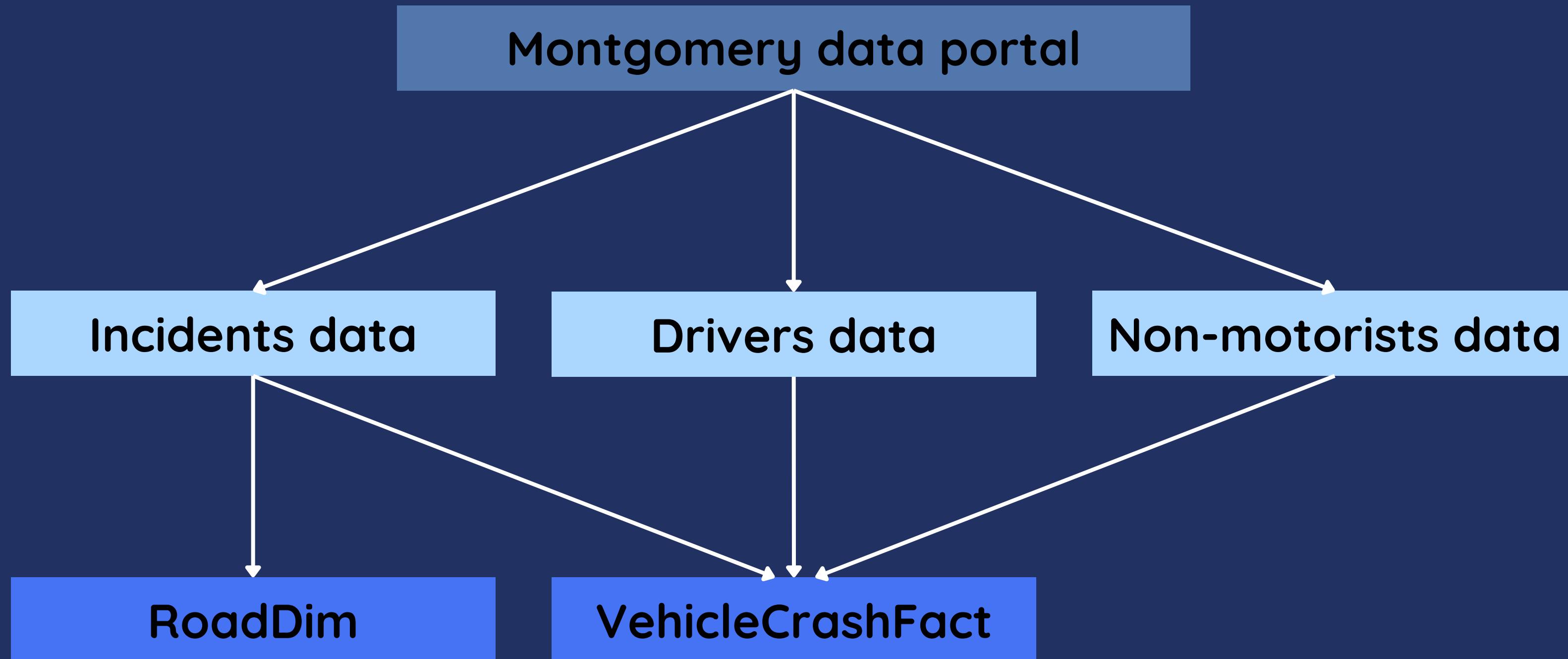
EKSTRAKCJA

- odczytanie CSV z URL

TRANSFORMACJE

- wypełnienie nulli wartością ‘Unknown’
- stworzenie własnych klas dla ‘Transmission’ i ‘Drivetrain’
- wygenerowanie pustych modeli dla każdej marki
- wygenerowanie kluczy głównych

Model ETL



Model ETL

RoadDim	
Column Name	Data Type
RoadName	nvarchar(50)
RouteType	nvarchar(50)
 RoadKey	bigint

TRANSFORMACJE

- utworzenie wierszy na podstawie wpisów z wypadków
- zastąpienie nulli wartością ‘Unknown’
- usunięcie duplikatów
- utworzenie klucza głównego

Model ETL

Drivers data

TRANSFORMACJE

- zastąpienie nulli wartością ‘Unknown’ i zerami
- wyczyszczenie wartości w ‘DriverSubstanceAbuse’
- własne kategorie w ‘VehicleType’
- dodanie flag w ‘SubstanceAbuseContributed’, ‘DriverAtFault’, ‘ParkedVehicle’
- dodanie zagregowanej wartości ‘CrashedTotal’
- mapowanie marek, modeli i daty produkcji oraz wygenerowanie klucza obcego dla VehicleDim

VehicleCrashFact
ReportNumber
VehicleCrashKey
DriverAtFault
DriverInjurySeverity
DriverSubstanceAbuse
DriverDistractedBy
VehicleType
VehicleMovement
VehicleGoingDir
VehicleDamageExtent
SpeedLimit
ParkedVehicle
SubstanceAbuseContributed
VehiclesCrashedTotal
VehicleKey
LocalCaseNumber
AgencyName
ACRSReportType
HitRun
LaneDirection
LaneNumber
NumberOfLanes
RoadGrade

NonTraffic
OffRoadIncident
AccidentAtFault
CollisionType
SurfaceCondition
Light
TrafficControl
Junction
IntersectionType
RoadAlignment
RoadCondition
RoadDivision
Latitude
Longitude
RoadKey
CrossStreetKey
NonMotoristTotal
NonMotoristInjury
NonMotoristFatal
DateHourKey
LocationAreaKey

Model ETL

Incidents data

TRANSFORMACJE

- zastąpienie nulli wartością ‘Unknown’
- skrócenie wartości w ‘ACRSReportType’
- przekonwertowanie daty
- dodanie flag w ‘HitRun’, ‘NonTraffic’ i ‘OffRoadIncident’
- mapowanie lokalizacji na regiony
- wygenerowanie kluczy obcych dla RoadDim, VehicleDim, DateHourDim
- scalenie z ‘Drivers data’

VehicleCrashFact	
ReportNumber	
VehicleCrashKey	NonTraffic
DriverAtFault	OffRoadIncident
DriverInjurySeverity	AccidentAtFault
DriverSubstanceAbuse	CollisionType
DriverDistractedBy	SurfaceCondition
VehicleType	Light
VehicleMovement	TrafficControl
VehicleGoingDir	Junction
VehicleDamageExtent	IntersectionType
SpeedLimit	RoadAlignment
ParkedVehicle	RoadCondition
SubstanceAbuseContributed	RoadDivision
VehiclesCrashedTotal	Latitude
VehicleKey	Longitude
LocalCaseNumber	RoadKey
AgencyName	CrossStreetKey
ACRSReportType	NonMotoristTotal
HitRun	NonMotoristInjury
LaneDirection	NonMotoristFatal
LaneNumber	DateHourKey
NumberOfLanes	LocationAreaKey
RoadGrade	

Model ETL

Non-motorists data

TRANSFORMACJE

- własne kategorie dla ‘InjurySeverity’
- stworzenie zagregowanych miarek
‘NonMotoristTotal’, ‘NonMotoristInjury’,
‘NonMotoristFatal’
- scalenie z ‘Incidents data’ i wypełnienie nulli zerami

VehicleCrashFact
ReportNumber
VehicleCrashKey
DriverAtFault
DriverInjurySeverity
DriverSubstanceAbuse
DriverDistractedBy
VehicleType
VehicleMovement
VehicleGoingDir
VehicleDamageExtent
SpeedLimit
ParkedVehicle
SubstanceAbuseContributed
VehiclesCrashedTotal
VehicleKey
LocalCaseNumber
AgencyName
ACRSReportType
HitRun
LaneDirection
LaneNumber
NumberOfLanes
RoadGrade

NonTraffic
OffRoadIncident
AccidentAtFault
CollisionType
SurfaceCondition
Light
TrafficControl
Junction
IntersectionType
RoadAlignment
RoadCondition
RoadDivision
Latitude
Longitude
RoadKey
CrossStreetKey
NonMotoristTotal
NonMotoristInjury
NonMotoristFatal
DateHourKey
LocationAreaKey

Narzędzia Raportowania

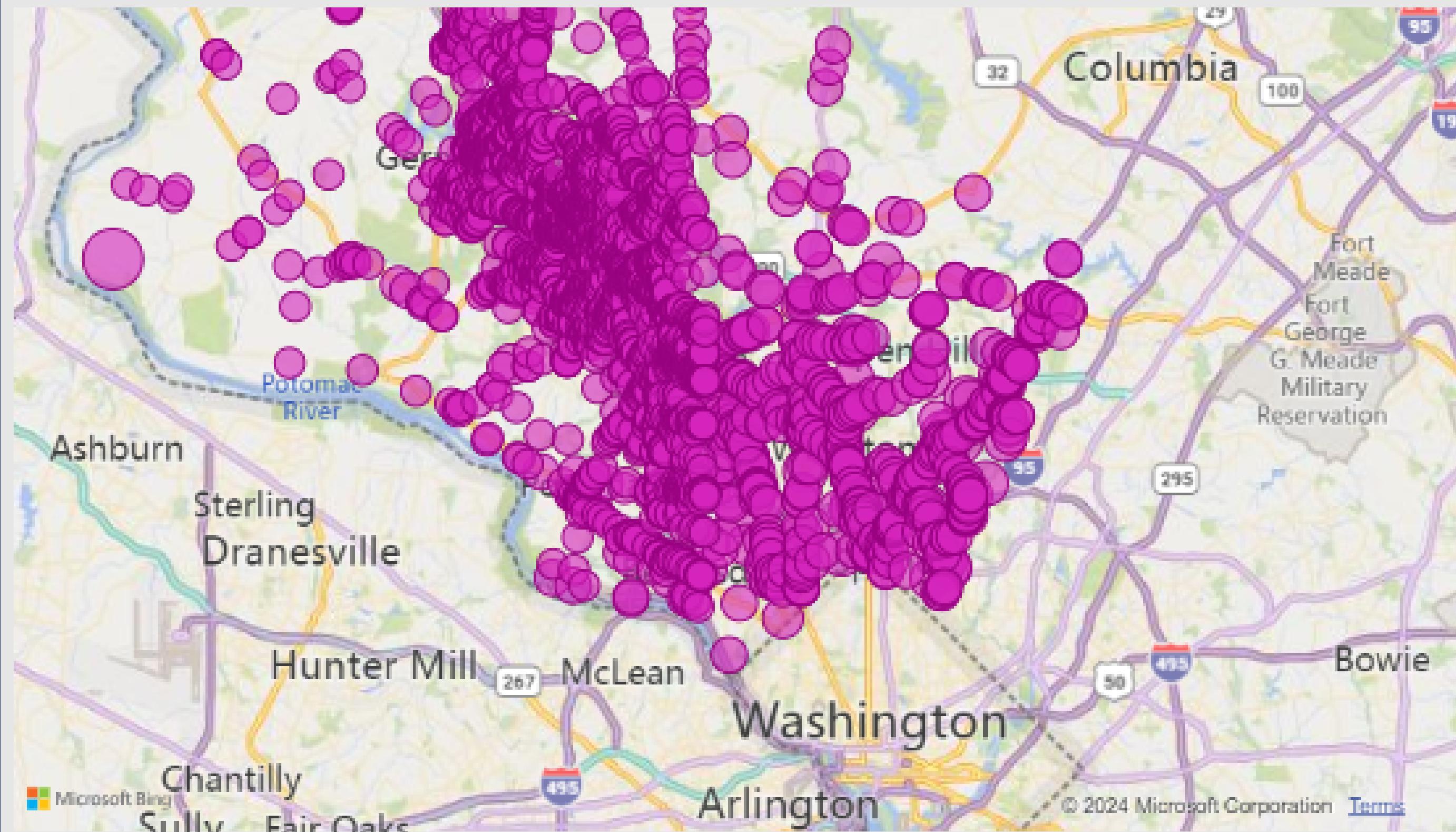
Wykorzystywane do tworzenia interaktywnych raportów, raportów ad-hoc i raportów statycznych.

Pozwalają na dokładniejszą analizę trendów i ciekawych zależności w danych

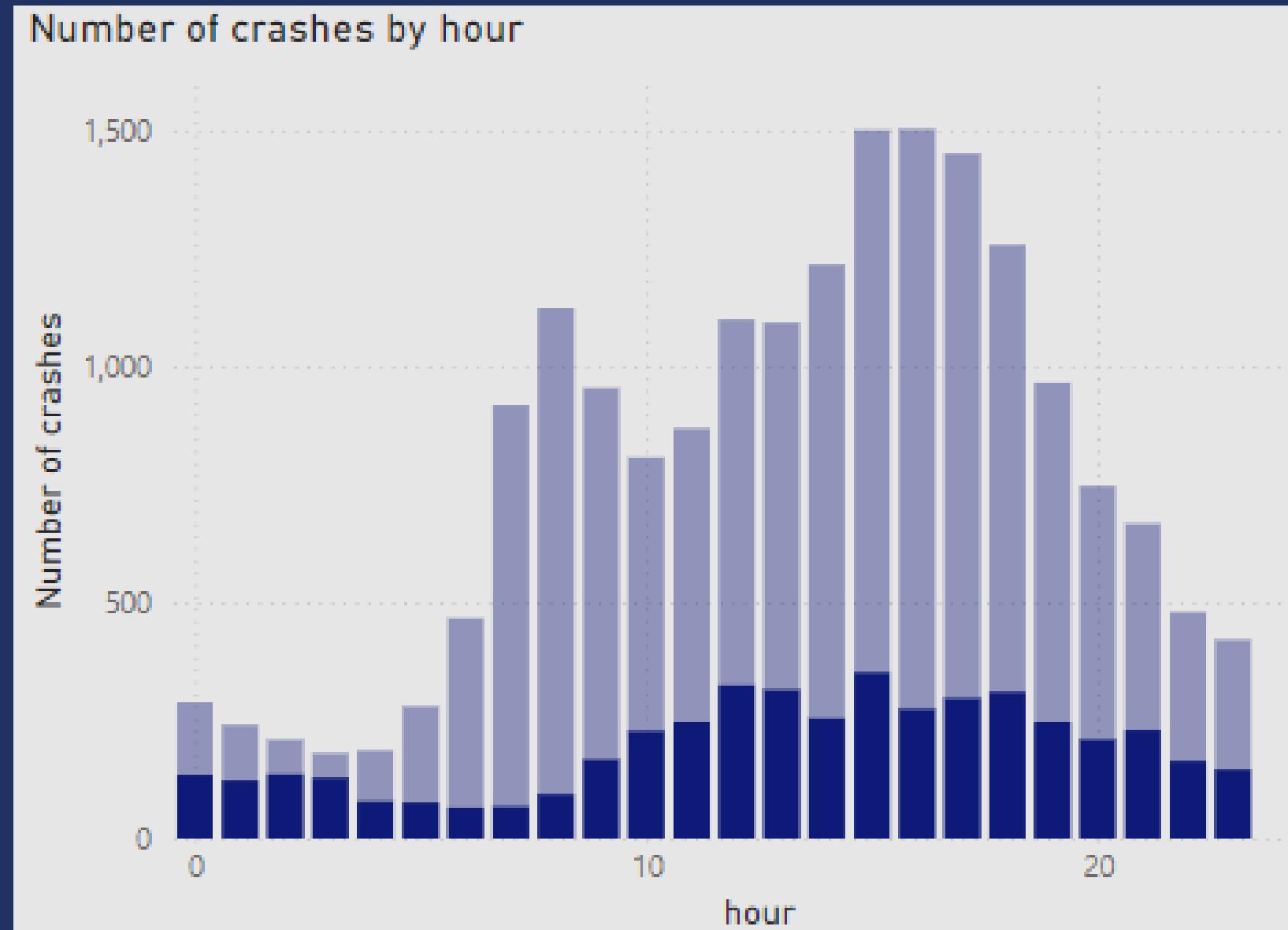


Power BI

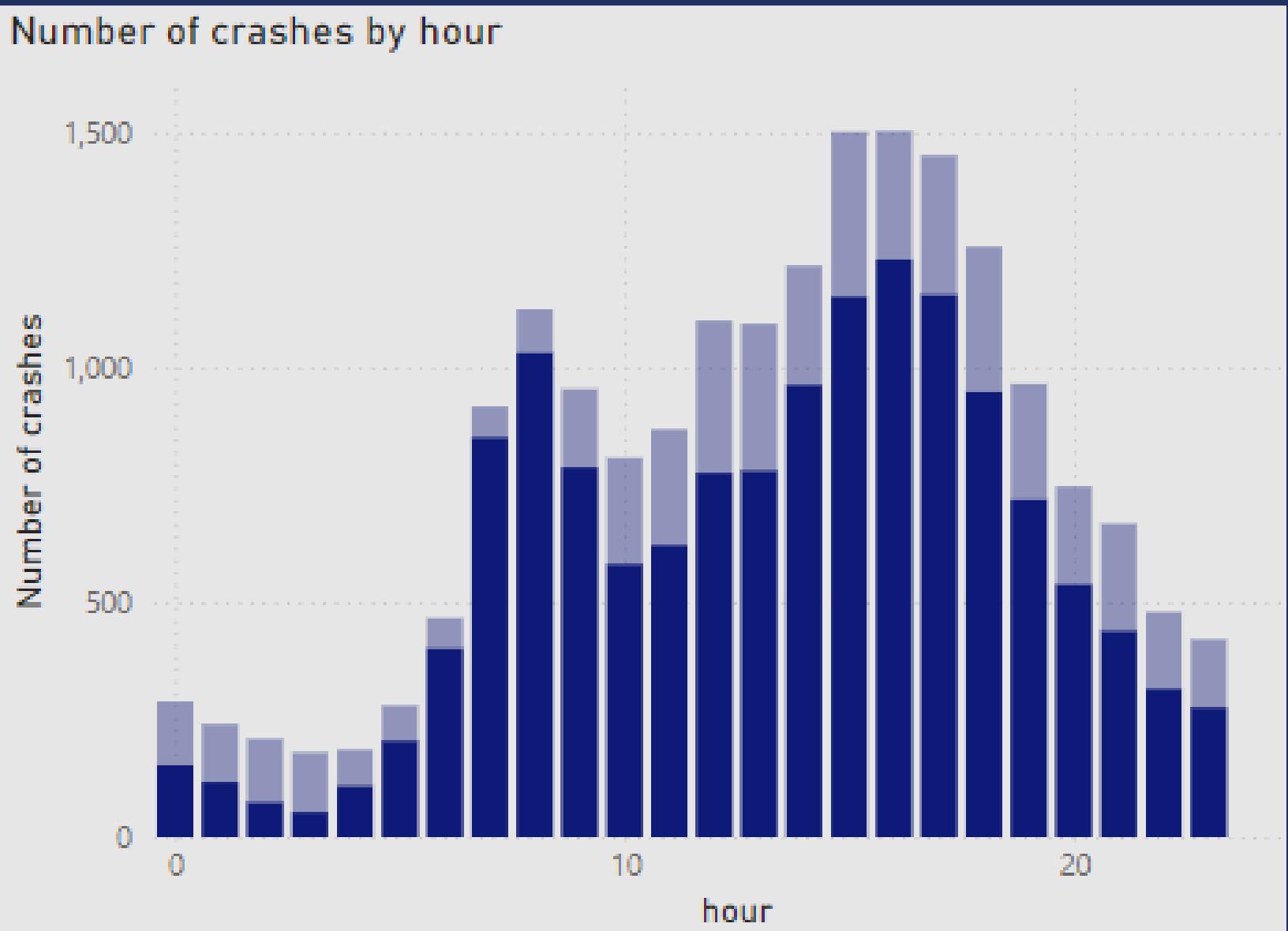
Map of crashes in Montgomery County by Latitude and Longitude



Weekend

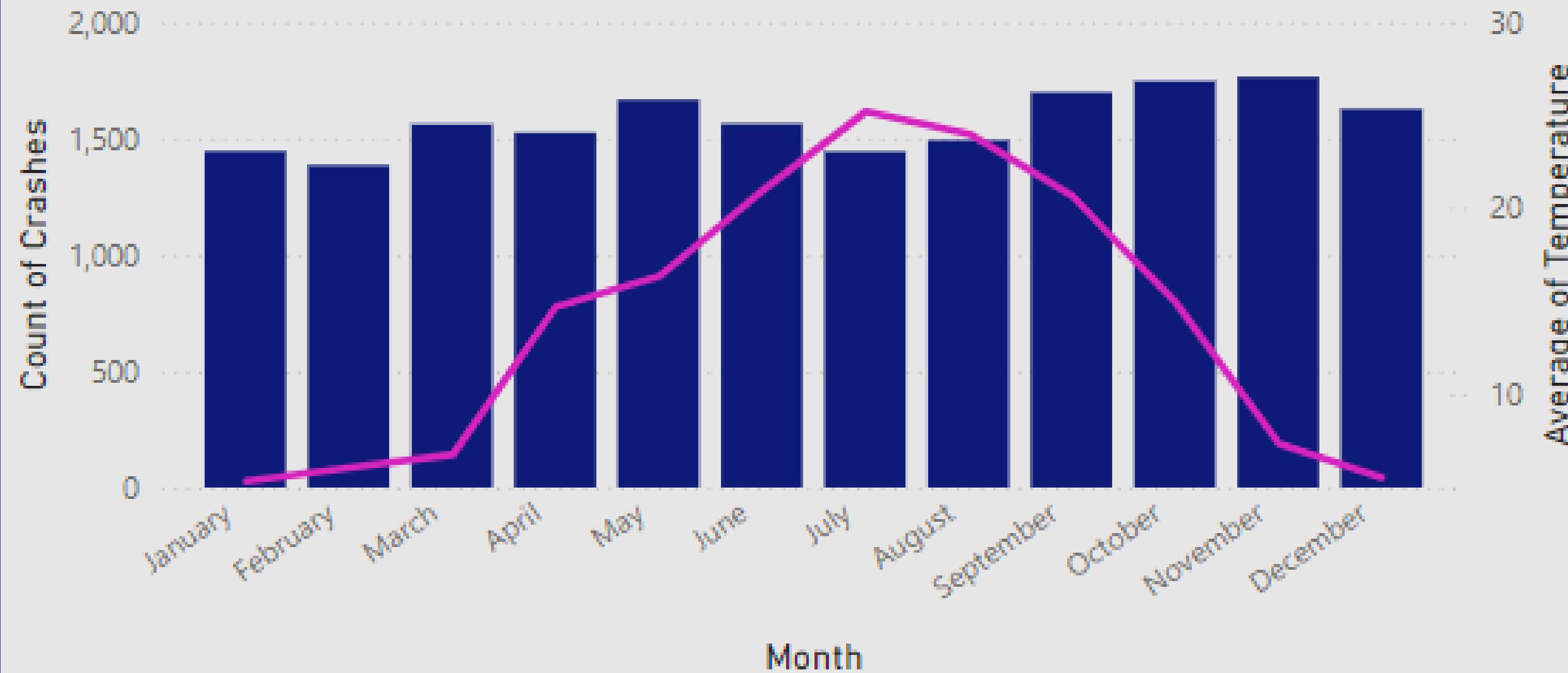


Dzień roboczy

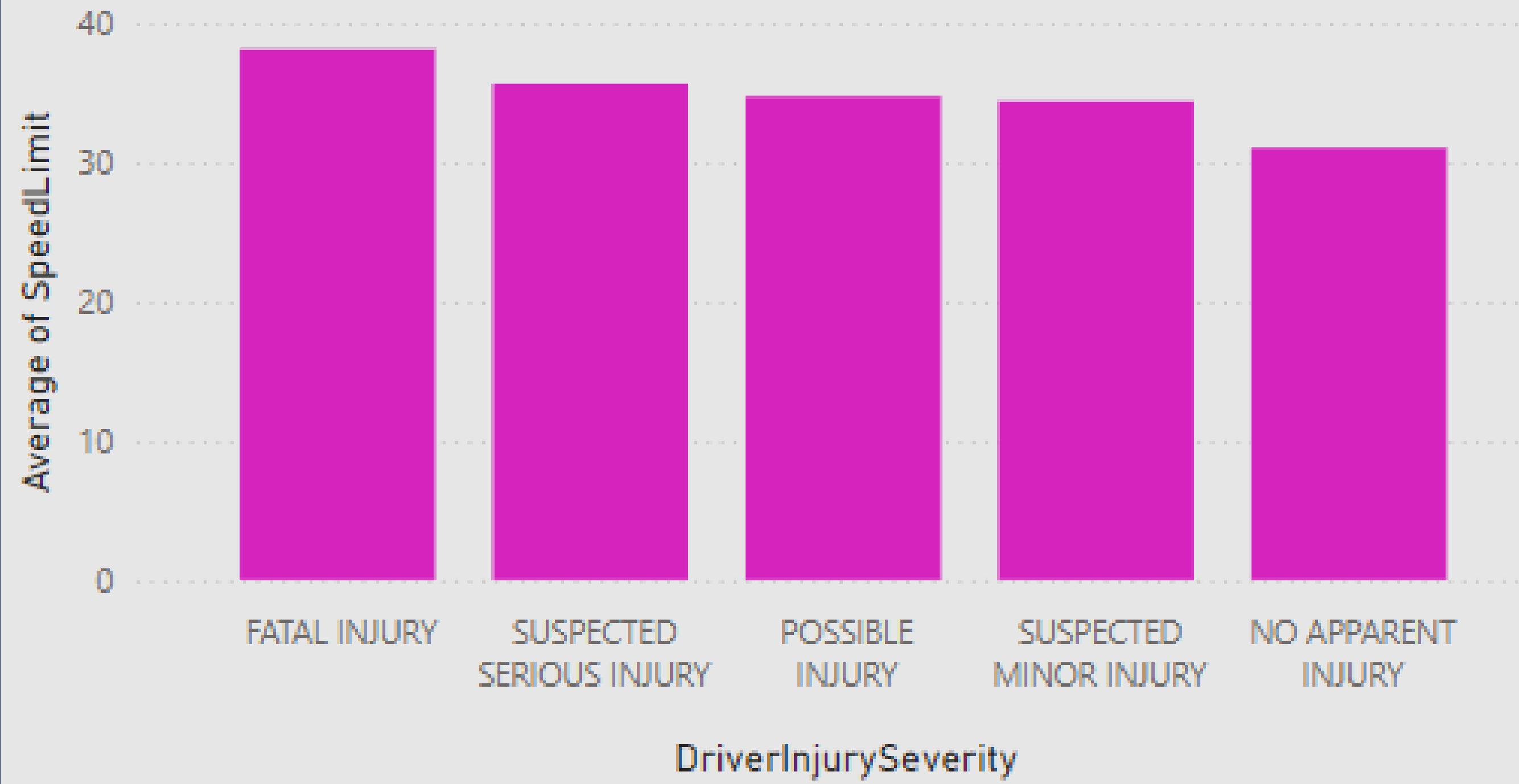


Count of Crashes and Average of Temperature by Month

● Count of Crashes ● Average of Temperature



Average of SpeedLimit by DriverInjurySeverity



Prezentacja działania hurtowni i symulacja dodawania nowych danych