Sveučilište Jurja Dobrile

Fakultet informatike u Puli

Matej Višnjić

Kriminal u Bostonu

Seminarski rad

Pula, 6.6.2021

Sveučilište Jurja Dobrile

Fakultet informatike u Puli

Matej Višnjić

Kriminal u Bostonu

Seminarski rad

JMBAG: 0303092423, redoviti student

Smjer: Informatika

Kolegij: Sustavi poslovne inteligencije

Mentor: doc. dr. sc. Goran Oreški

Pula, 6.6.2021

Sadržaj

1. UVOD

U današnjem dobu sve više se susrećemo sa raznoraznim podacima, istraživanjima, i analitikama podataka. Poslovna inteligencija je širok pojam koji obuhvaća rudarenje podataka, analizu procesa, usporedbu, i opisnu analitiku.

Poslovna inteligencija postupak je za analiziranje podataka i pružanje korisnih informacija menadžerima, radnicima ili korisnicima u donošenju poslovnih odluka. Vrlo je važno da su podaci koji su analizirani točni u suprotnom menadžer može donijeti loše odluke što bi moglo rezultirati gubitke. Krajnji cilj poslovne inteligencije je donošenje boljih poslovnih odluka, povećanje prihoda, poboljšanje operativne učikovitosti i slično.

Podaci poslovne inteligencije obično se pohranjuju u nekakvo skladište podatka. Podaci mogu sadržavati povijesne informacije i podatke u stvarnom vremenu prikupljenih iz izvornih sustava. Prije nego što se podaci koriste u alatima, podaci se prvo moraju integrirati, objedniti i očistiti. Ovime se osigurava da podaci budu što točniji.

2. Poslovna inteligencija

Poslovna inteligencija je tehnološki postupak za analizu podataka i pružanje korisnih informacija koje pomažu rukovoditeljima poduzeća, menadžerima i radnicima u donošenju poslovnih odluka.

Proces se temelji na:

- transformaciji podataka u informacije
- odluci
- akciji

Također, imamo različite primjene poslovne inteligencije:

1. deskriptivna analitika – Što se dogodilo?
    a. poslovno izvještavanje
    b. skladišta podataka
    c. kontrolne ploče

    Rezultat – dobro definiran poslovni problem i mogućnost

2. prediktivna analitika – Što će se dogoditi?
    a. rudarenje podataka
    b. rudarenje tekstualnih podataka
    c. sustavi za predikciju

    Rezultat – precizne projekcije budućih događaja i rezultata

3. preskriptivna – Što bi se trebalo poduzeti?
    a. optimizacija
    b. simulacija
    c. ekspertni sustavi

    Rezultat – najbolje poslovne odluke i akcije

Osnovne komponente poslovnih sustava:

- skladište podataka
- alati za obradu i prezentaciju podataka
- korisničko sučelje

Ključni dijelovi poslovnih procesa:

1. Podaci iz glavnog sustava u itegrirani i učitani u skladište podataka ili neki sličan analitički repozitorij

2. Podaci su organizirani u analitičkom modelu ili OLAP, da ih pripreme za analitiku.

3. Analitičari ispituju podatke.

4. Rezultati podataka se prikazuju pomoću grafova, nadzornih ploča ili izvještaja.

5. Rukovoditelji i radnici koriste informacije za donošenje odluka.

2.1. Podaci

Podaci predstavljaju skup činjenica koje se najčešće prikupljaju kroz; eksperimente, opservacije i transkacije. Na najvišoj razini podaci se dijele na; strukturirane, nestrukturirane i polu-strukturirane.

Karakteristike koje određuju spremnost podataka za izradu poslovnog modela:

- pouzdanost izvora
- točnost sadržaja
- dostupnost
- sigurnost i privatnost
- potpunost
- konzistentnost
- vremenska dimenzija
- granularnost
- ispravnost
- relevantnost

2.2 Opis i analiza projekta

Kao temu za ovaj projekt odabrao sam kriminal u Bostonu, koji je praćen u periodu od 2015. godine sve do 2018. godine. Skup podataka pronađen je na internetu: link.

1. Postoji preko 300 000 podataka
2. Različiti tipovi podataka
3. Postoji vremenska dimenzija
4. Postoji lokacijska dimenzija

Za ispitivanje podataka koristio sam python, biblioteka „pandas“. Biblioteka „pandas“ je vrlo dobra i korisna biblioteka za rad sa skupovima podataka. Pomoću pythona vidimo koliko je unikatnih vrijednosti, koliko podataka nije upisano, da li nešto fali i slično.

Ovdje se radi o deskriptivnoj analitici jer si postavljamo pitanje „što se dogodilo?“

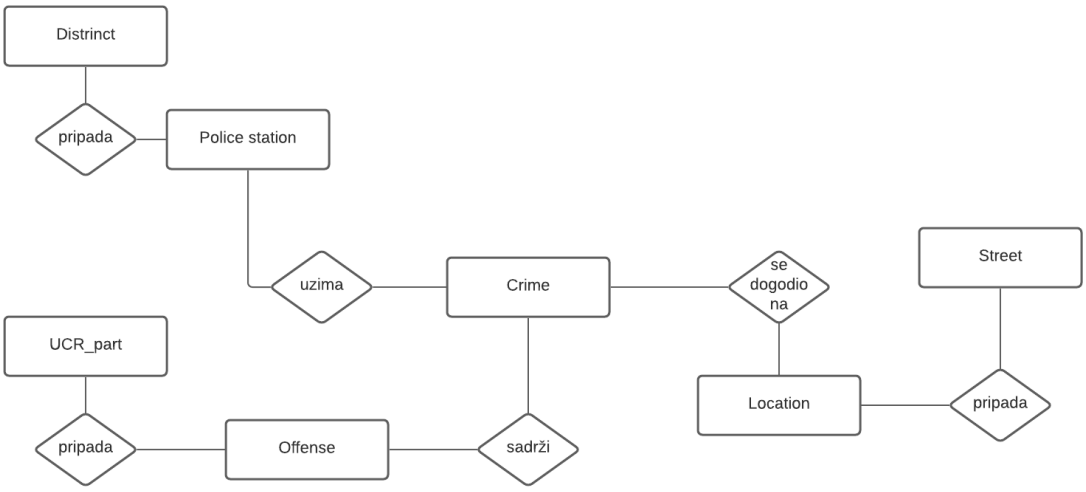
Pratiti ćemo koliko se dogodilo slučajeva, kakve vrsta slučajeva, da li je pucano, u koje vrijeme se dogodio kriminal, kad ima više slučajeva kada manje, u kojoj ulici se dogodio slučaj...

Cilj projekta je da pronađemo 5 dimenzije preko kojih ćemo vršiti razne analize podataka. Podatke iz CSV datoteke prebacimo u relacijsku bazu podataka, iz baze, koristeći pentaho, uspoređujemo podatke sa istima da bi vidjeli da li se podaci podudaraju, te na konačnici ubacimo podatke u dimenziju bazu podataka i preko tih dimenzija stvaramo grafove i pratimo analitiku kad se što dogodilo.

2.3 Relacijski model podataka

Nakon što smo pronašli podatke te ih analizirali, moramo kreirati bazu podataka i ubaciti podatke iz CSV datoteke.

2.3.1 EER Diagram jednostavan

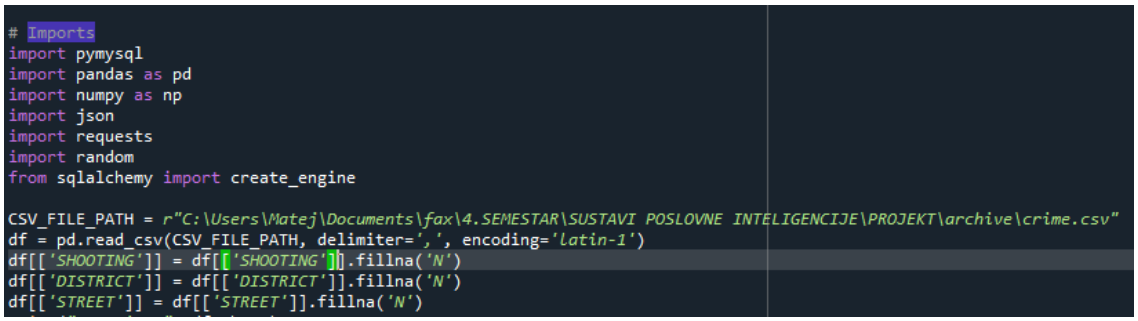


Slika 1. – EER Diagram – jednostavni

2.3.2 Ubacivanje u relacijsku bazu podataka

Za pravljenje baze i ubacivanje u relacijski model koristio sam program „Spyder“ te bazu podataka „MySQL“.

Najprvo sam vrijednosti koje nedostaju(Nan vrijednosti), sam dodao slovo „N“ kako bi znali da za te podatke ne postoji vrijednost.



```
# Imports
import pymysql
import pandas as pd
import numpy as np
import json
import requests
import random
from sqlalchemy import create_engine

CSV_FILE_PATH = r"C:\Users\Matej\Documents\fax\4.SEMESTAR\SUSTAVI POSLOVNE INTELIGENCIJE\PROJEKT\archive\crime.csv"
df = pd.read_csv(CSV_FILE_PATH, delimiter=',', encoding='latin-1')
df[['SHOOTING']] = df[['SHOOTING']].fillna('N')
df[['DISTRICT']] = df[['DISTRICT']].fillna('N')
df[['STREET']] = df[['STREET']].fillna('N')
```

Slika 2. – promjena NaN vrijednosti

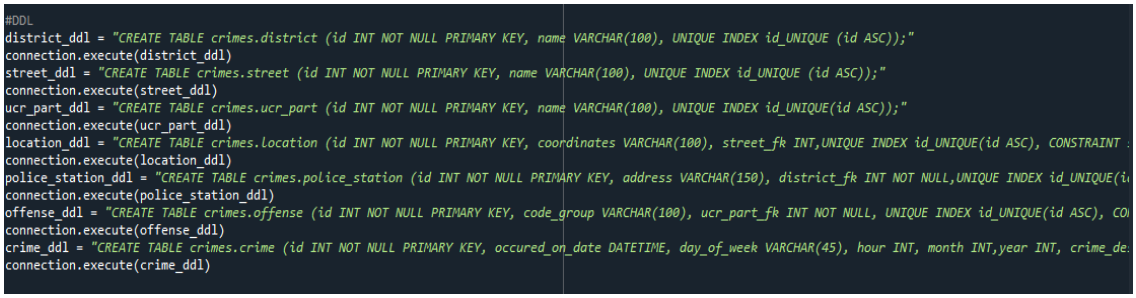
Nakon što su podaci spremni za ubacivanje u bazu, moramo napraviti konekciju na bazu te napraviti skriptu za tablice.

```
#Konekcija na bazu
user = 'root'
pasw = '1234'
host = 'localhost'
port = 3306
database = 'crimes'

mydb = create_engine('mysql+pymysql://' + user + ':' + pasw + '@' +host+ ':' +str(port)+ '/' +database, echo = False)
print(mydb)
connection = mydb.connect()
```

Slika 3. – konekcija na bazu

Kada smo uspješno spojeni sa bazom, onda možemo stvoriti tablice za punjenje podataka.



```
#DDL
district_ddl = "CREATE TABLE crimes.district (id INT NOT NULL PRIMARY KEY, name VARCHAR(100), UNIQUE INDEX id_UNIQUE (id ASC));"
connection.execute(district_ddl)
street_ddl = "CREATE TABLE crimes.street (id INT NOT NULL PRIMARY KEY, name VARCHAR(100), UNIQUE INDEX id_UNIQUE (id ASC));"
connection.execute(street_ddl)
ucr_part_ddl = "CREATE TABLE crimes.ucr_part (id INT NOT NULL PRIMARY KEY, name VARCHAR(100), UNIQUE INDEX id_UNIQUE(id ASC));"
connection.execute(ucr_part_ddl)
location_ddl = "CREATE TABLE crimes.location (id INT NOT NULL PRIMARY KEY, coordinates VARCHAR(100), street_fk INT,UNIQUE INDEX id_UNIQUE(id ASC), CONSTRAINT
connection.execute(location_ddl)
police_station_ddl = "CREATE TABLE crimes.police_station (id INT NOT NULL PRIMARY KEY, address VARCHAR(150), district_fk INT NOT NULL,UNIQUE INDEX id_UNIQUE(i
connection.execute(police_station_ddl)
offense_ddl = "CREATE TABLE crimes.offense (id INT NOT NULL PRIMARY KEY, code_group VARCHAR(100), ucr_part_fk INT NOT NULL, UNIQUE INDEX id_UNIQUE(id ASC), CO
connection.execute(offense_ddl)
crime_ddl = "CREATE TABLE crimes.crime (id INT NOT NULL PRIMARY KEY, occured_on_date DATETIME, day_of_week VARCHAR(45), hour INT, month INT,year INT, crime_de
connection.execute(crime_ddl)
```

Slika 4. – stvaranje tablica

Sada je baza spremna za punjenje, moramo uzeti podatke iz CSV datoteke i ubaciti u bazu.

```
#POLICE_STATION
# id,address,district_fk
district_id=[]
for i,row in df.iterrows():
    district_id.append(int(district_data['id'].iloc[i]))
policestation_data = pd.DataFrame({'id':list(range(1,len(district_id)+1)), 'district_fk': district_id})
policestation_data.to_sql(con=mydb, name='police_station', if_exists='append', index=False)
```

Slika 5. – primjer ubacivanja u relacijsku bazu

Trenutno nemamo adrese policijskih stanica koje su poredane po okruzima za koje su zadužene tako da smo preko web stranice bpdnews.com pronašli kojim okruzima pripadaju.

```
#Adrese za svaku stanicu posebno
policestation_A1 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '40 New Sudbury Street Boston, MA 02114' WHERE d.n
connection.execute(policestation_A1)
policestation_A15 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '40 New Sudbury Street Boston, MA 02114' WHERE d.
connection.execute(policestation_A15)
policestation_A7 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '69 Paris Street East Boston, MA 02128' WHERE d.na
connection.execute(policestation_A7)
policestation_B2 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '2400 Washington Street Roxbury, MA 02119' WHERE d
connection.execute(policestation_B2)
policestation_B3 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '1165 Blue Hill Avenue Mattapan, MA 02124' WHERE d
connection.execute(policestation_B3)
policestation_C11 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '40 Gibson Street Dorchester, MA 02122' WHERE d.n
connection.execute(policestation_C11)
policestation_C6 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '101 West Broadway South Boston, MA 02127' WHERE d
connection.execute(policestation_C6)
policestation_D14 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '301 Washington Street Brighton, MA 02135' WHERE
connection.execute(policestation_D14)
policestation_D4 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '650 Harrison Avenue Boston, MA 02116' WHERE d.nam
connection.execute(policestation_D4)
policestation_E13 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '3347 Washington Street Jamaica Plain, MA 02130'
connection.execute(policestation_E13)
policestation_E18 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '1249 Hyde Park Avenue Hyde Park, MA 02136' WHERE
connection.execute(policestation_E18)
policestation_E5 = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '1708 Centre Street West Roxbury, MA 02132' WHERE
connection.execute(policestation_E5)
policestation_N = "UPDATE police_station ps INNER JOIN district d on d.id = ps.district_fk SET ps.address= '' WHERE d.name = 'N';" # ne znamo koja je policijs
connection.execute(policestation_N)
```

Slika 6. – dodavanje adrese za svaku

id	occured_on_date	day_of_week	hour	month	year	crime_description	shooting	offense_fk	location_fk	police_station_fk
1	2018-09-02 13:00:00	Sunday	13	9	2018	LARCENY ALL OTHERS	N	1	1	1
2	2018-08-21 00:00:00	Tuesday	0	8	2018	VANDALISM	N	2	2	2
3	2018-09-03 19:27:00	Monday	19	9	2018	TOWED MOTOR VEHICLE	N	3	3	3
4	2018-09-03 21:16:00	Monday	21	9	2018	INVESTIGATE PROPERTY	N	4	4	4
5	2018-09-03 21:05:00	Monday	21	9	2018	INVESTIGATE PROPERTY	N	5	5	5
6	2018-09-03 21:09:00	Monday	21	9	2018	M/V ACCIDENT INVOLVING PEDESTRIAN - INJURY	N	6	6	6
7	2018-09-03 21:25:00	Monday	21	9	2018	AUTO THEFT	N	7	7	7
8	2018-09-03 20:39:37	Monday	20	9	2018	VERBAL DISPUTE	N	8	8	8
9	2018-09-03 20:48:00	Monday	20	9	2018	ROBBERY - STREET	N	9	9	9
10	2018-09-03 20:38:00	Monday	20	9	2018	VERBAL DISPUTE	N	10	10	10
11	2018-09-03 19:55:00	Monday	19	9	2018	VERBAL DISPUTE	N	11	11	11
12	2018-09-03 20:19:00	Monday	20	9	2018	INVESTIGATE PROPERTY	N	12	12	12
13	2018-09-03 19:58:00	Monday	19	9	2018	FIRE REPORT - HOUSE, BUILDING, ETC.	N	13	13	13
14	2018-09-03 20:39:00	Monday	20	9	2018	THREATS TO DO BODILY HARM	N	14	14	14
15	2018-09-02 14:00:00	Sunday	14	9	2018	PROPERTY - LOST	N	15	15	15

Slika 7. – tablica crimes

id	address	district_fk
1	301 Washington Street Brighton, MA 02135	1
2	40 Gibson Street Dorchester, MA 02122	2
3	650 Harrison Avenue Boston, MA 02116	3
4	650 Harrison Avenue Boston, MA 02116	4
5	1165 Blue Hill Avenue Mattapan, MA 02124	5
6	40 Gibson Street Dorchester, MA 02122	6
7	2400 Washington Street Roxbury, MA 02119	7
8	2400 Washington Street Roxbury, MA 02119	8
9	101 West Broadway South Boston, MA 02127	9
10	40 Gibson Street Dorchester, MA 02122	10
11	101 West Broadway South Boston, MA 02127	11
12	101 West Broadway South Boston, MA 02127	12
13	650 Harrison Avenue Boston, MA 02116	13
14	1165 Blue Hill Avenue Mattapan, MA 02124	14
15	1165 Blue Hill Avenue Mattapan, MA 02124	15

Slika 8. – tablica police station

id	name
1	D14
2	C11
3	D4
4	D4
5	B3
6	C11
7	B2
8	B2
9	C6
10	C11
11	C6
12	C6
13	D4
14	B3
15	B3

Slika 9. – tablica district

id	name
1	Part One
2	Part Two
3	Part Three
4	Part Three
5	Part Three
6	Part Three
7	Part One
8	Part Three
9	Part One
10	Part Three
11	Part Three
12	Part Three
13	Part Three
14	Part Two
15	Part Three

Slika 10. – tablica ucr_part

id	coordinates	street_fk
1	(42.35779134, -71.13937053)	1
2	(42.30682138, -71.06030035)	2
3	(42.34658879, -71.07242943)	3
4	(42.33418175, -71.07866441)	4
5	(42.27536542, -71.09036101)	5
6	(42.29019621, -71.07159012)	6
7	(42.30607218, -71.08273260)	7
8	(42.32701648, -71.10555088)	8
9	(42.33152148, -71.07085307)	9
10	(42.29514664, -71.05860832)	10
11	(42.31957856, -71.04032766)	11
12	(42.34011469, -71.05339029)	12
13	(42.35038760, -71.08785290)	13
14	(42.28647012, -71.08714661)	14
15	(42.27924052, -71.09667382)	15

Slika 11. – tablica street

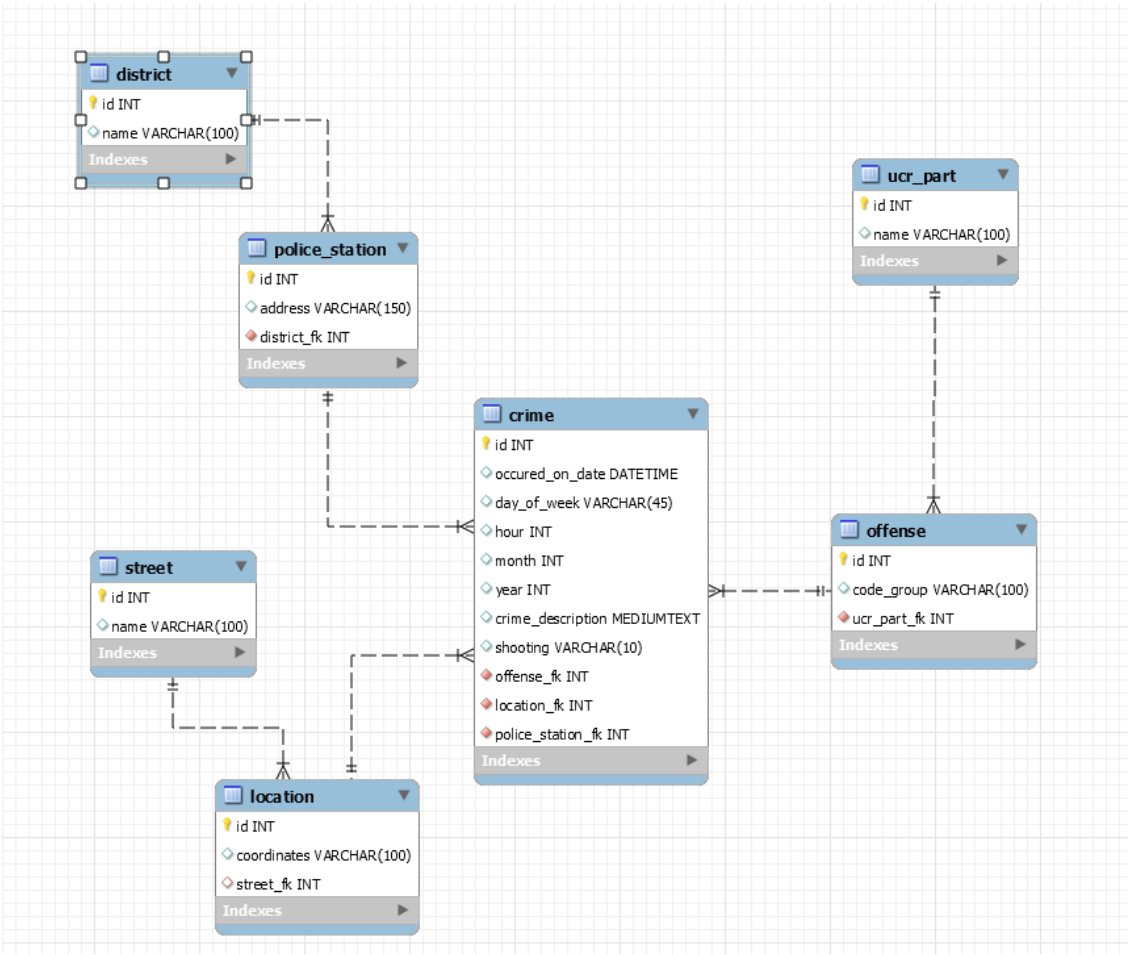
id	code_group	ucr_part_fk
1	Larceny	1
2	Vandalism	2
3	Towed	3
4	Investigate Property	4
5	Investigate Property	5
6	Motor Vehicle Accident Response	6
7	Auto Theft	7
8	Verbal Disputes	8
9	Robbery	9
10	Verbal Disputes	10
11	Verbal Disputes	11
12	Investigate Property	12
13	Fire Related Reports	13
14	Other	14
15	Property Lost	15

Slika 12. – tablica offense

id	coordinates	street_fk
1	(42.35779134, -71.13937053)	1
2	(42.30682138, -71.06030035)	2
3	(42.34658879, -71.07242943)	3
4	(42.33418175, -71.07866441)	4
5	(42.27536542, -71.09036101)	5
6	(42.29019621, -71.07159012)	6
7	(42.30607218, -71.08273260)	7
8	(42.32701648, -71.10555088)	8
9	(42.33152148, -71.07085307)	9
10	(42.29514664, -71.05860832)	10
11	(42.31957856, -71.04032766)	11
12	(42.34011469, -71.05339029)	12
13	(42.35038760, -71.08785290)	13
14	(42.28647012, -71.08714661)	14
15	(42.27924052, -71.09667382)	15

Slika 13. – tablica location

2.3.3 EER Model



Slika 14. – EER Model

3. Dimenzijski model podataka

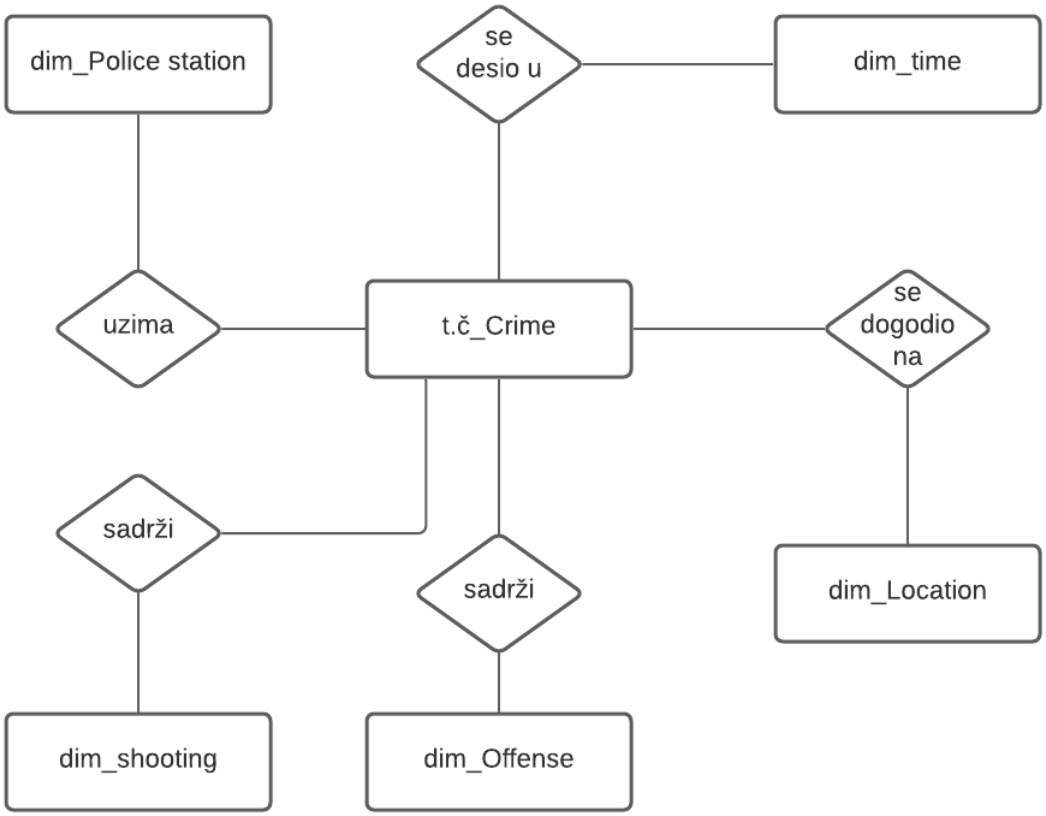
Nakon što smo napunili relacijsku bazu podataka, sada prelazimo na dimenzijski model podataka. Prednost dimenzijskog modela:

- standardizacija
- pohranjivanje povijesnih podataka
- modularnost
- dohvat velike količine podataka
- performanse upita
- razumljivost

predmet modeliranja su činjenice i dimenzije

U mom projektu koristimo sporo degenenirane dimenzije te tablicu činjenica bez činjenica. Shema je star shema.

3.1 Dimenzijski EER diagram – jednostavan



Slika 15. – jednostavan EER

Odabrali smo dimenzije i tablicu činjenica, te kreirali jednostavan EER diagram koji nam služi kao skica. Koristeći softver „Pentaho“ napuniti ćemo našu dimenzijsku bazu podataka.

3.2 Ubacivanje podataka - Pentaho

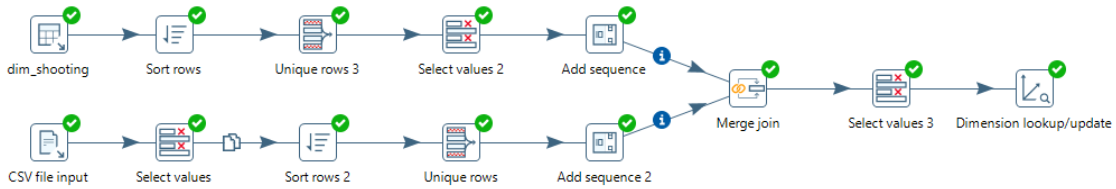
U ovom dijelu projekta prelazimo na punjenje dimenzijskih tablica koristeći program „Pentaho“ verzije 9.1. Prvo trebamo skinuti mysql konektor koji se nalazi na MySQL web stranici. Link.

Kada uspješno instaliramo mysql konektora spajamo na našu relacijsku bazu podataka koju smo napunili. U pentahu uspoređujemo podatake iz CSV datoteke i naše relacijske baze podataka, tako što mičemo sve vrijednosti koji su višak, to jest unikatne vrijednosti ostavljamo. U konačnici cilj nam je dobiti sve podatke u tablici činjenica, a dimenzije sadržavaju unikatne vrijednosti. Ovim načinom smanjiti ćemo vrijednosti koje se ponavljaju.

3.2.1 Dimenzija dim_shooting

Počeo sam sa dimenzijom dim_shooting i u njoj ću pohraniti dvije vrijednosti, Y i N.

„Y“ će govoriti da li je pucano a slovo „N“ da nije pucano.



Slika 16. – ubacivanje u dim_shooting

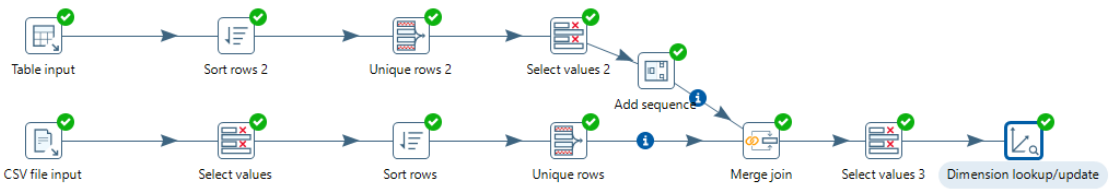
Najprije učitavamo iz relacijske tablice crime, gdje uzimamo podatak shooting. Onda podatke sortiramo i uzimamo vrijednosti koje su unikatne, također istu stvar radimo sa CSV datotekom. Kasnije ih spajamo u jedno, te u Dimensions lookup/update odaberemo vrijednosti koje će biti u tablici dim_shooting. Rezultat ove pentaho skripte je prikazan na slici broj 17.

shooting_sk	id	shooting	version	date_from	date_to
0	NULL	NULL	1	NULL	NULL
1	1	N	1	1900-01-01 00:00:00	2200-01-01 00:00:00
2	2	Y	1	1900-01-01 00:00:00	2200-01-01 00:00:00
NULL	NULL	NULL	NULL	NULL	NULL

Slika 17. – dim_shooting

3.2.2 Dimenzija dim_police_station

Uzimamo podatke iz tablice police_station i CSV datoteke. Uzimamo samo unikatne vrijednosti, kasnije ih spajamo u jedno i u konačnici biramo podatke koje ćemo staviti u dimenzijsku tablicu dim_police_station. Rezultat ove pentaho skripte biti će 12 policijskih stanica, svaka će sadržavati svoj okrug i 13. policijska stanica koja ima okrug N, jer za neke podatke nisu upisan podatak za okrug. Također rezultat možete pogledati na slici 19.



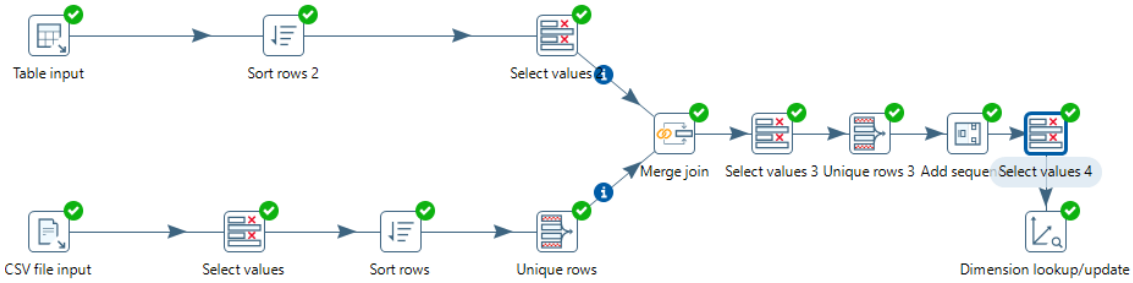
Slika 18. – ubacivanje u dim_policestation

police_station_sk	id	address	district	version	date_from	date_to
0	NULL	NULL	NULL	1	NULL	NULL
1	1	40 New Sudbury Street Boston, MA 02114	A1	1	1900-01-01 00:00:00	2200-01-01 00:00:00
2	2	40 New Sudbury Street Boston, MA 02114	A15	1	1900-01-01 00:00:00	2200-01-01 00:00:00
3	3	69 Paris Street East Boston, MA 02128	A7	1	1900-01-01 00:00:00	2200-01-01 00:00:00
4	4	2400 Washington Street Roxbury, MA 02119	B2	1	1900-01-01 00:00:00	2200-01-01 00:00:00
5	5	1165 Blue Hill Avenue Mattapan, MA 02124	B3	1	1900-01-01 00:00:00	2200-01-01 00:00:00
6	6	40 Gibson Street Dorchester, MA 02122	C11	1	1900-01-01 00:00:00	2200-01-01 00:00:00
7	7	101 West Broadway South Boston, MA 02127	C6	1	1900-01-01 00:00:00	2200-01-01 00:00:00
8	8	301 Washington Street Brighton, MA 02135	D14	1	1900-01-01 00:00:00	2200-01-01 00:00:00
9	9	650 Harrison Avenue Boston, MA 02116	D4	1	1900-01-01 00:00:00	2200-01-01 00:00:00
10	10	3347 Washington Street Jamaica Plain, MA 0...	E13	1	1900-01-01 00:00:00	2200-01-01 00:00:00
11	11	1249 Hyde Park Avenue Hyde Park, MA 02136	E18	1	1900-01-01 00:00:00	2200-01-01 00:00:00
12	12	1708 Centre Street West Roxbury, MA 02132	E5	1	1900-01-01 00:00:00	2200-01-01 00:00:00
13	13	N	N	1	1900-01-01 00:00:00	2200-01-01 00:00:00
NULL	NULL	NULL	NULL	NULL	NULL	NULL

Slika 19. – dim_policestation

3.2.3 Dimenzija dim_offense

Sličnu stvar napravio sam kod dimenzije dim_offense. Uzeti su podaci iz CSV datoteke, i iz relacijske tablice offense koja je spojena sa tablicom ucr_part kako bi dobili točne podatke. Nakon što smo uzeli samo unikatne vrijednosti, i sortirali po abecedi, onda smo ju stavili u dimenzijsku tablicu dim_offense. Postoji 67 unikatnih grupa zločina u koje mogu spadati zločini iz naših podataka. Rezultat je prikazan na slici 20.



Slika 18. – ubacivanje u dim_offense

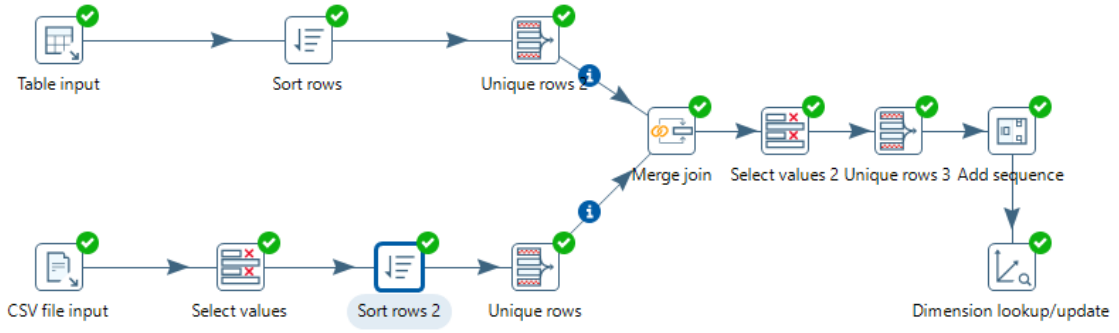
offense_sk	id	ucr_part	code_group	version	date_from	date_to
0	NULL	NULL	NULL	1	NULL	NULL
1	1	Part One	Aggravated Assault	1	1900-01-01 00:00:00	2200-01-01 00:00:00
2	2	Part Three	Aircraft	1	1900-01-01 00:00:00	2200-01-01 00:00:00
3	3	Other	Arson	1	1900-01-01 00:00:00	2200-01-01 00:00:00
4	4	Part Three	Assembly or Gathering Violations	1	1900-01-01 00:00:00	2200-01-01 00:00:00
5	5	Part One	Auto Theft	1	1900-01-01 00:00:00	2200-01-01 00:00:00
6	6	Other	Auto Theft Recovery	1	1900-01-01 00:00:00	2200-01-01 00:00:00
7	7	Part Two	Ballistics	1	1900-01-01 00:00:00	2200-01-01 00:00:00
8	8	Part Two	Biological Threat	1	1900-01-01 00:00:00	2200-01-01 00:00:00
9	9	Part Two	Bomb Hoax	1	1900-01-01 00:00:00	2200-01-01 00:00:00
10	10	Other	Burglary - No Property Taken	1	1900-01-01 00:00:00	2200-01-01 00:00:00
11	11	Part One	Commercial Burglary	1	1900-01-01 00:00:00	2200-01-01 00:00:00
12	12	Part Two	Confidence Games	1	1900-01-01 00:00:00	2200-01-01 00:00:00
13	13	Part Two	Counterfeiting	1	1900-01-01 00:00:00	2200-01-01 00:00:00
14	14	Part Two	Criminal Harassment	1	1900-01-01 00:00:00	2200-01-01 00:00:00

Slika 20. – dim_offense

tablica nije prikazana u cijelosti

3.2.4 Dimenzija dim_location

Dimenzija dim_location napravljena je koristeći relacijsku tablicu location i street te CSV datoteku gdje smo uzeli samo unikatne vrijednosti i sortirana je po imenu. Na kraju smo došli do rezultata da ima 4658 različitih imena ulica. Rezultat možete pogledati na slici 22.



Slika 21. – ubacivanje u dim_location

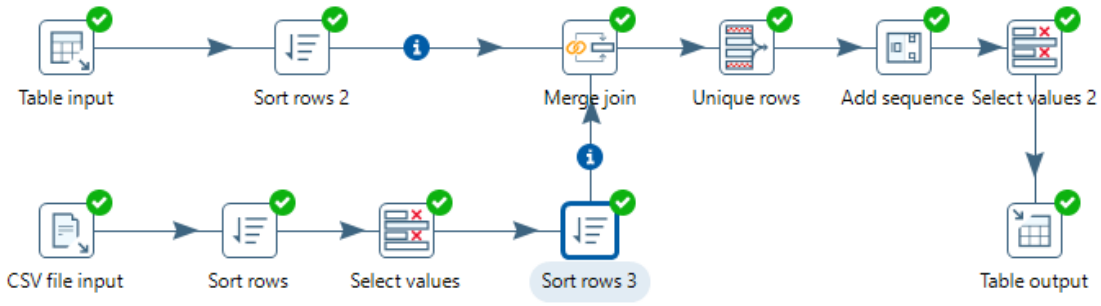
location_sk	id	coordinates	street_name	version	date_from	date_to
0	NULL	NULL	NULL	1	NULL	NULL
1	1	(42.33361000, -71.07337000)	ALBANY ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00
2	2	(42.31730370, -71.07799648)	BLUE HILL AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
3	3	(42.34060371, -71.08174185)	COLUMBUS AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
4	4	(42.35053956, -71.13106722)	COMMONWEALTH AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
5	5	(42.33253100, -71.07213000)	MASSACHUSETTS AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
6	6	(0.00000000, 0.00000000)	0 BURRELL	1	1900-01-01 00:00:00	2200-01-01 00:00:00
7	7	(0.00000000, 0.00000000)	0 MASS AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
8	8	(0.00000000, 0.00000000)	00 CENTRE ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00
9	9	(0.00000000, 0.00000000)	00 MASS AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
10	10	(0.00000000, 0.00000000)	00 OLD COLONY AVE	1	1900-01-01 00:00:00	2200-01-01 00:00:00
11	11	(0.00000000, 0.00000000)	12 BREED ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00
12	12	(0.00000000, 0.00000000)	12 BROOKVIEW ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00
13	13	(0.00000000, 0.00000000)	12 WOODWARD PARK ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00
14	14	(0.00000000, 0.00000000)	13 DIXWELL ST	1	1900-01-01 00:00:00	2200-01-01 00:00:00

Slika 22. – dim_location

tablica nije prikazana u cijelosti

3.2.5 Dimenzija dim_time

Dimenzija dim_time na isti način je implementirana kao i ostale dimenzije do sada. Podaci su uzeti iz relacijske baze podataka i CSV datoteke, zatim su sortirani i nakon spajanja su dodani u bazu. U konačnici u dimenziji time imati ćemo 233,229 podataka, koji su unikatni. Ti podaci biti će povezani sa tablicom činjenica. Rezulatat možete pogledati na slici 24.



Slika 23. – ubacivanje u dim_time

time_sk	id	day_of_week	hour	month	YEAR	occured_on_date
1	1	Friday	0	1	2016	2016-01-01 00:00:00
2	2	Friday	0	1	2016	2016-01-01 00:01:00
3	3	Friday	0	1	2016	2016-01-01 00:05:00
4	4	Friday	0	1	2016	2016-01-01 00:09:00
5	5	Friday	0	1	2016	2016-01-01 00:13:00
6	6	Friday	0	1	2016	2016-01-01 00:14:00
7	7	Friday	0	1	2016	2016-01-01 00:15:00
8	8	Friday	0	1	2016	2016-01-01 00:16:00
9	9	Friday	0	1	2016	2016-01-01 00:28:00
10	10	Friday	0	1	2016	2016-01-01 00:30:00
11	11	Friday	0	1	2016	2016-01-01 00:35:00
12	12	Friday	0	1	2016	2016-01-01 00:37:00
13	13	Friday	0	1	2016	2016-01-01 00:38:00
14	14	Friday	0	1	2016	2016-01-01 00:39:00
15	15	Friday	0	1	2016	2016-01-01 00:43:00

Slika 24. – dim_time

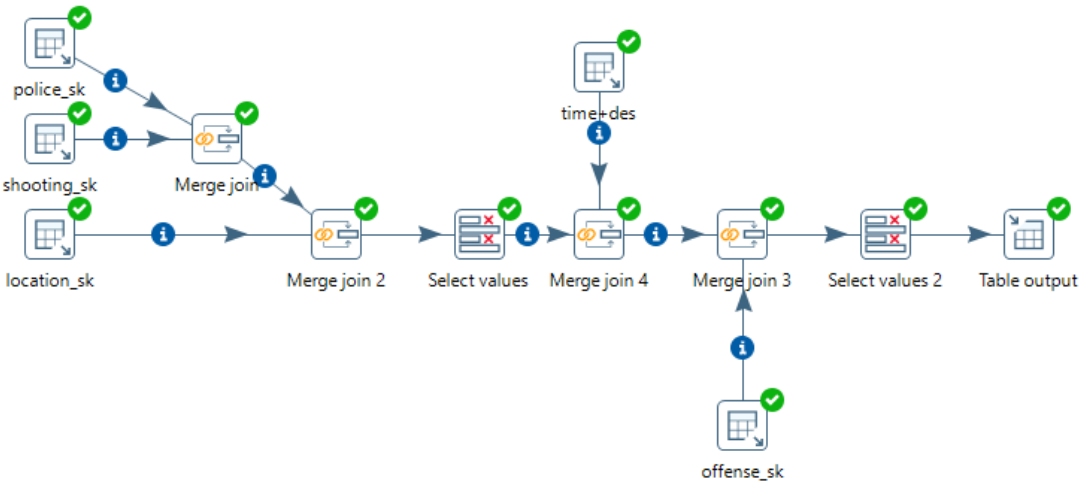
tablica nije prikazana u cijelosti

3.2.6 Tablica činjenica dim_crime

Tablica činjenica je bez činjenica i sadržavati će samo opis slučaja koji se dogodio. Za tablicu činjenica nije korišten CSV datoteka kao u dimenzijama, a razlog tomu je što se podaci nisu mogli lijepo učitati iz CSV datoteke. Zbog nemogućnosti korištenja CPython script executora nisam mogao hvatati ključeve sa pythonom, nego sam ih hvatao sam sql select upitom. Nakon što sam podatke uhvatio, moramo paziti na kako su podaci sortirani, meni su sortirani po zadanoj vrijednosti kao u CSV datoteci. Nakon što smo ih sve pospajali sa merge joinom, spremni smo ju ubaciti u tablicu činjenica, dim_crime. Primjer hvatanja ključa, slika 25. Rezultat možete pogledati na slici 27.

```
SELECT dp.police_station_sk, c.id
FROM crimes.crime c,
     crimes.offense o,
     crimes.location l,
     crimes.police_station p,
     crimes.district d
INNER JOIN crimes_dimenzije.dim_police_station dp
ON dp.district = d.name
WHERE c.offense_fk = o.id
AND   c.location_fk = l.id
AND   c.police_station_fk = p.id
AND   d.id = p.district_fk;
```

Slika 25. – primjer hvatanje ključa SQL upitom



Slika 26. – ubacivanje u dim_crime

	crime_sk	id	crime_description	dim_location_sk	dim_offense_sk	dim_time_sk	dim_police_station_sk	dim_shooting_sk
▶	1	1	LARCENY ALL OTHERS	2536	35	110302	8	1
	2	2	VANDALISM	2074	64	5917	6	1
	3	3	TOWED MOTOR VEHICLE	785	63	190628	9	1
	4	4	INVESTIGATE PROPERTY	3067	33	213269	9	1
	5	5	INVESTIGATE PROPERTY	1241	33	213267	5	1
	6	6	M/V ACCIDENT INVOLVING PEDESTRIAN - INJURY	4075	44	213268	6	1
	7	7	AUTO THEFT	3100	5	213270	4	1
	8	8	VERBAL DISPUTE	2460	65	202386	4	1
	9	9	ROBBERY - STREET	2741	59	202387	7	1
	10	10	VERBAL DISPUTE	2504	65	202384	6	1
	11	11	VERBAL DISPUTE	3160	65	190631	7	1
	12	12	INVESTIGATE PROPERTY	1174	33	202383	7	1
	13	13	FIRE REPORT - HOUSE, BUILDING, ETC.	2709	20	190632	9	1
	14	15	PROPERTY - LOST	2968	53	122881	5	1
	15	16	SICK/INJURED/MEDICAL - PERSON	2987	41	190629	13	1

Slika 27. – dim_crime

tablica nije prikazana u cijelosti

4. Vizualizacija podataka

Završni dio ovoga projekta je vizualizacija projekta te prikaz razno raznih podataka koristeći OLAP (Online Analytical Processing), tehnologija koja stoji iza mnogih aplikacija poslovne inteligencije. OLAP je moćna tehnologija za otkrivanje podataka, uključujući mogućnost za pregledavanje izvješća, složene analitične izračune te planiranje scenarija „što ako“ planiranje.

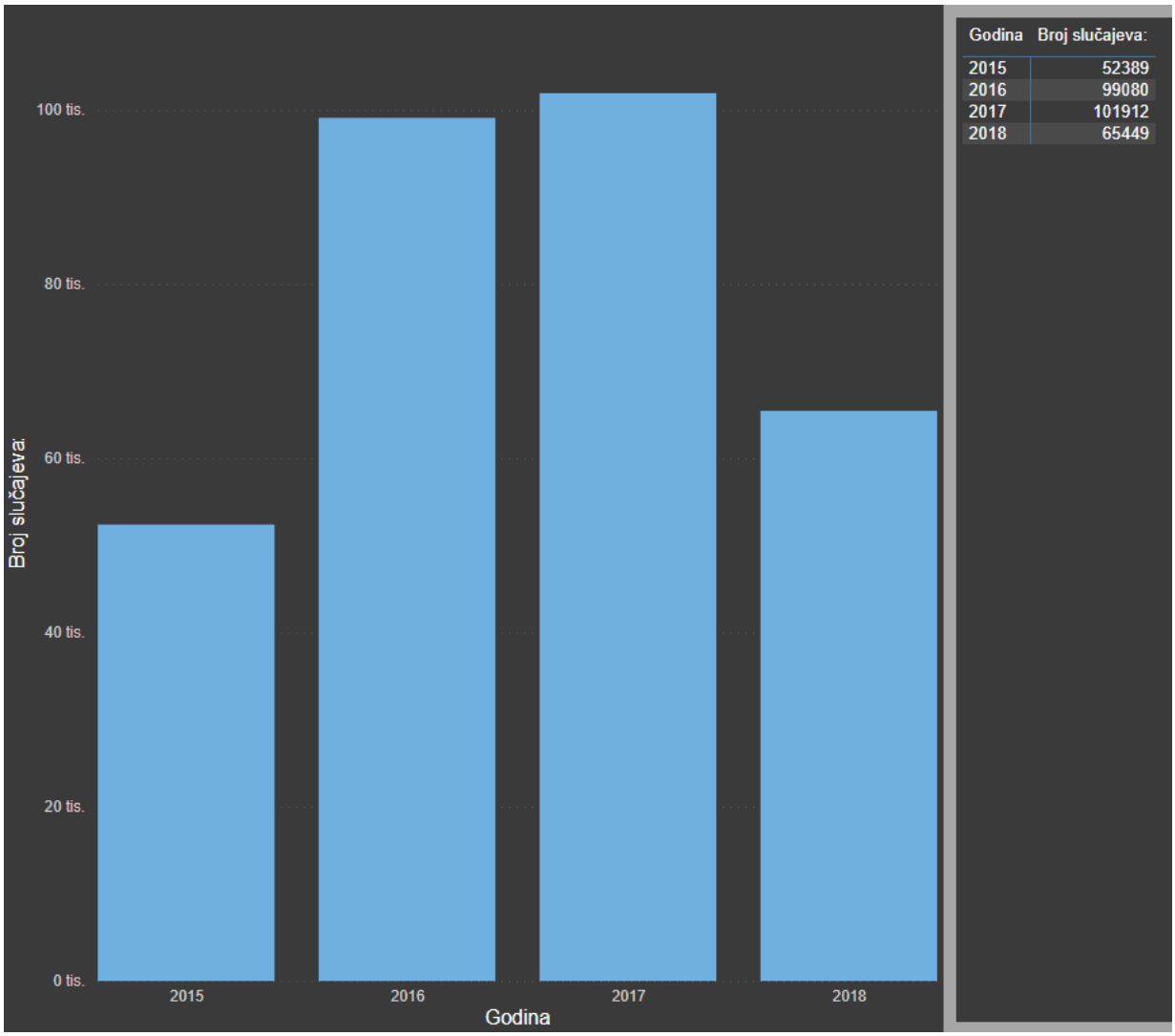
Za razliku od relacijskih baza podataka, OLAP alati ne pohranjuju pojedinačne zapise transakcija u dvodimenzionalom formatu reda po stupcu, već umjesto toga koriste višedimenzionalne strukture baze podatka, poznate pod nazivom OLAP kocke. Podaci i formule pohranjuju se u optimiziranu višedimenzionalnu bazu podataka, u mom slučaju dimenzijska baza podataka koju sam implementirao gore, dok se prikazi podataka izrađuju na zahtjev.

Implementacija OLAP tehnologije ne ovisi samo o vrsti softvera, već i o temeljnim izvorima podataka i planiranim poslovnim ciljevima.

Ja, osobno koristio sam Microsoft Power BI. Microsoft Power BI usluga poslovne analitike tvrtke Microsoft. Softver je napravljen da pruži korisnicima interaktivne vizualizacije i mogućnost poslovne inteligencije s jednostavnim sučeljem da krajnji korisnici mogu stvoriti vlastita izvješća i nadzorne ploče. Power BI omogućuje lako povezivanje sa izvorima podataka i vizualizaciju i otkrivanje onoga što je važno za daljnje poslovanje.

Power BI softver nije kompliciran za korištenje, na početnom zaslonu odaberemo spajanje sa MySQL bazom podataka. Kada se uspješno spojimo onda uzimamo dimenzijske tablice te ih spajamo sa činjeničnom tablicom. Kada to napravimo, u prozoru nam se prikaže graf koji smo odabrali. Softver nam nudi preko 10-ak raznih grafova i mogućnosti vizualizacija podataka.

4.1 Prikaz broj slučaja prema godinama

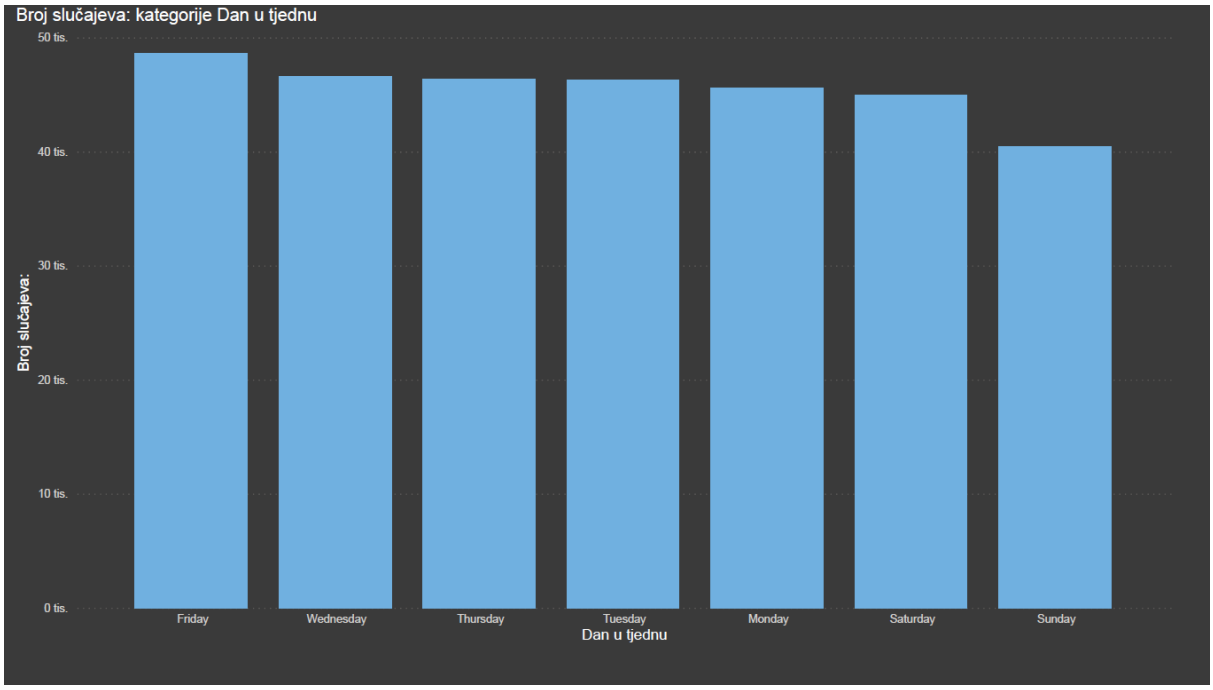


Slika 28. – graf- broj slučaja prema godinama

Za implementiranje ovog grafa, uzeo sam YEAR koja se nalazi u dim_time, te ju spojio sa crimes_sk ključem koji se nalazi u tablici činjenica.

U ovome grafu, prikazani su svi slučajevi koji su se dogodili u periodu od 2015. do 2018. godine. U desnom gornjem kutu vidimo podatke za svaku godinu. Iz ovog grafa vidimo da se najviše slučajeva dogodilo 2017. godine.

4.2 Prikaz broja slučajeva po danima



Slika 29. – graf- broj slučaja prema danima u tjednu

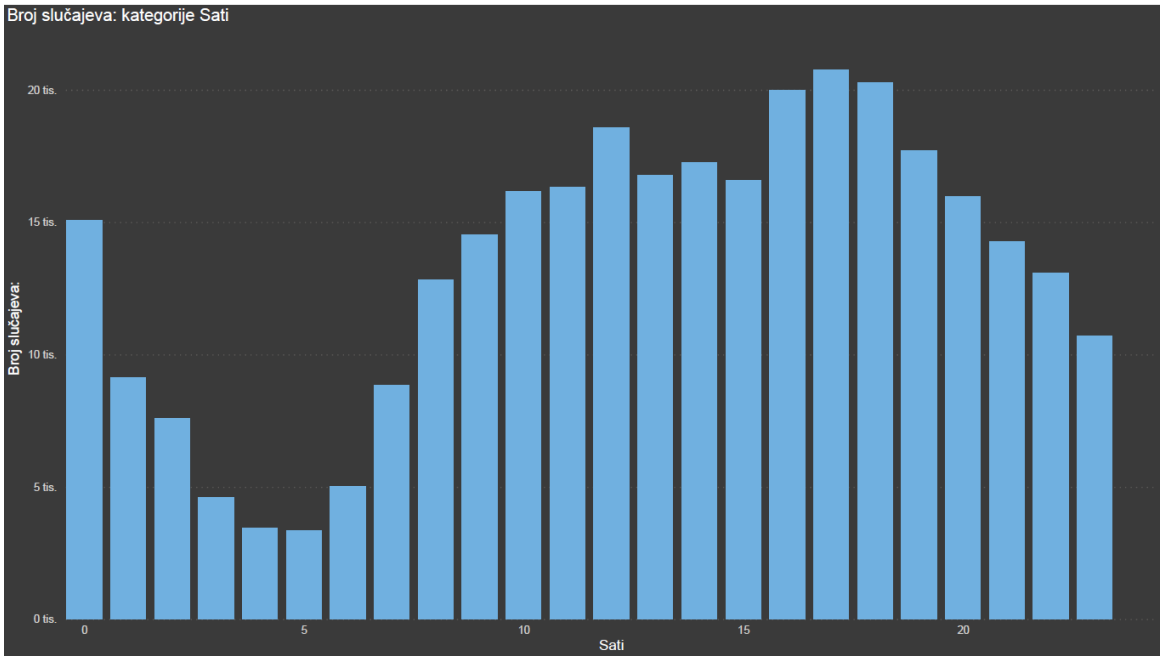
Za implementaciju ovog grafa koristio sam day_of_week iz dimenzije dim_time, te crime_sk iz tablice činjenica.

Graf sadrži podatke koliko se slučajeva desilo po danima u tjednu.

4.3 Prikaz broja slučajeva prema satima

Implementacija je slična ko i u prijašnjim grafovima, uzet je podatak hour iz dim_time, te crime_sk iz činjenične tablice.

Graf sadrži podatke koliko se slučajeva desilo prema satima. Graf se nalazi na slici 30.

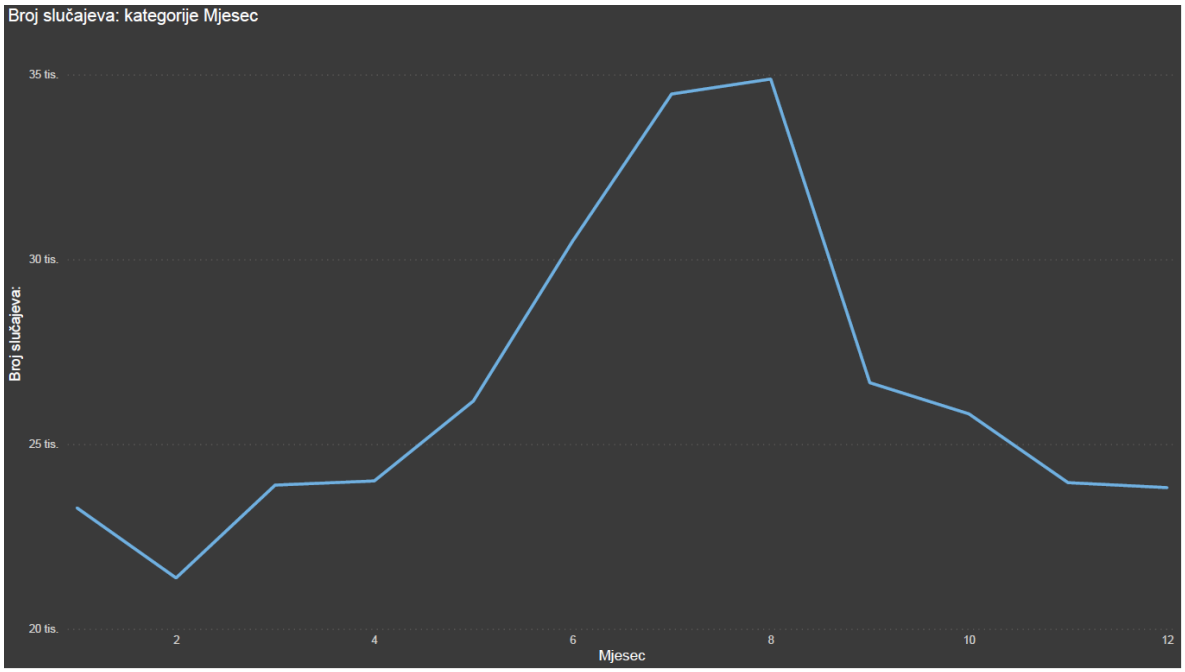


Slika 30. – graf- broj slučaja prema danima u tjednu

4.4 Prikaz broja slučajeva po mjesecima

Ovaj graf se razlikuje od ostalih jer je linijski, a uzeti su podaci month i dim_tim, te povezan sa crime_sk iz tablice činjenica.

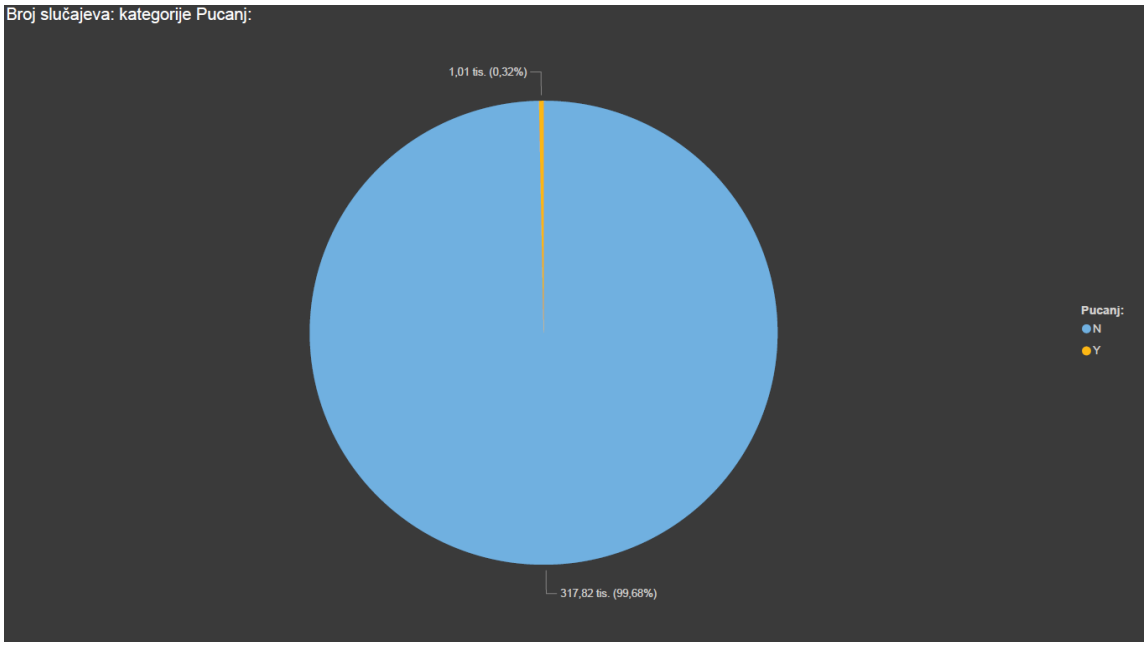
Graf prikazuje broj slučajeva poredanih po mjesecima.



Slika 31. – graf- broj slučaja prema mjesecima

4.5 Prikaz broja slučaja u kojemu je pucano iz vatrenog oružja

Graf je tortni te prikazuje podatak da li je pucano ili ne. Implementacija je ista ko i do sada, samo što je uzet podatak shooting, iz tablice dim_shooting te je povezan sa tablicom činjenica po ključu crime_sk.



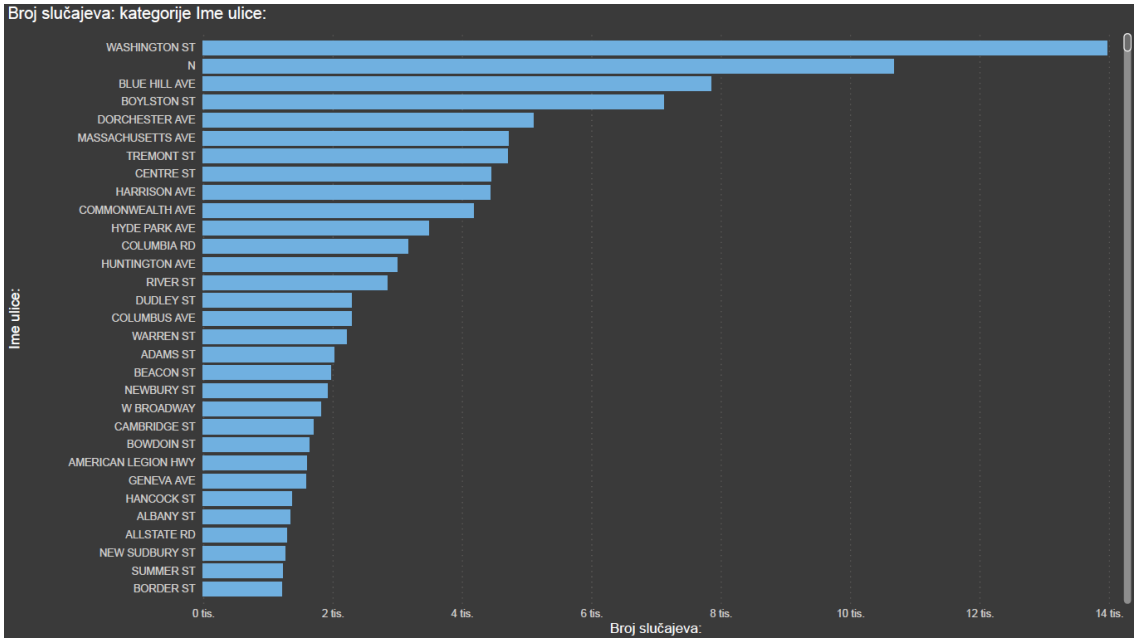
Slika 32. – graf- broj slučaja u kojemu je pucano iz oružja

4.6 Broj slučaja po imenima ulica

Graf je grupirani trokutast graf i sadrži podatke broj slučaja prema ulicama, poredanih od najvećeg prema najmanjem.

Implementacija je slična, uzet je atribut street_name iz tablice dim_location i povezan sa ključem crimes_sk iz tablice činjenica.

Graf je na slici 33.

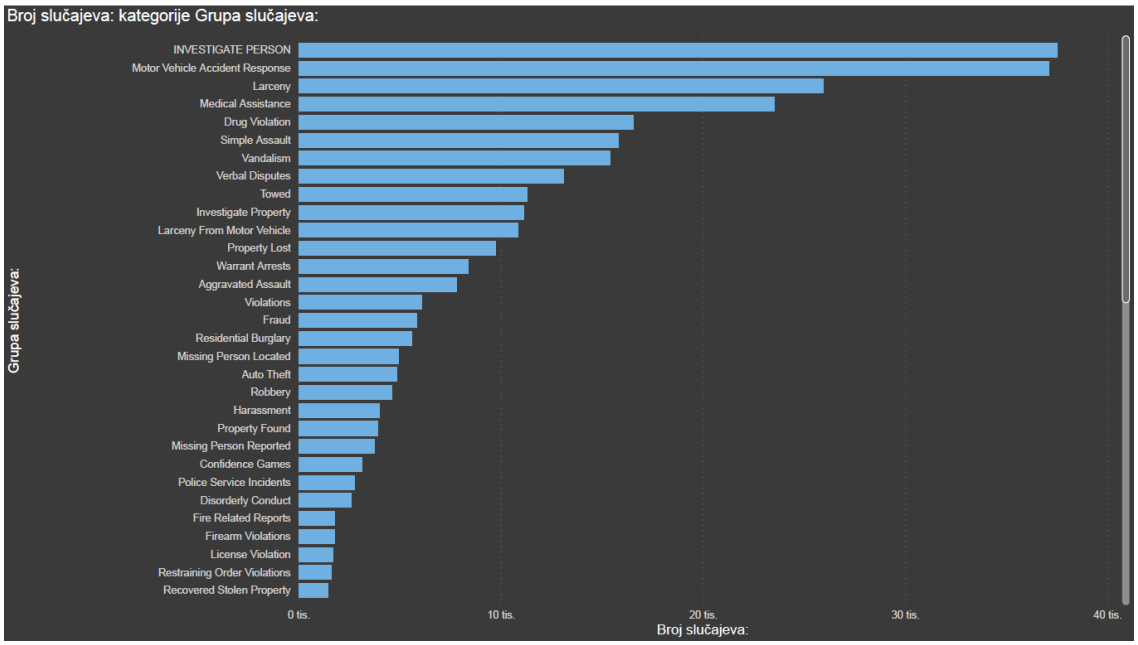


Slika 33. – graf- broj slučaja prema imenima ulice

4.7 Broj slučajeva prema vrsti prekršaja

Graf je također grupirani trokutast i prikazuje broj slučajeva prema vrsti prekršaja koji se dogodio.

Implementacija je slična samo je korišten code_group iz tablice dim_offense te je povezan sam ključem crime_sk iz tablice činjenica.

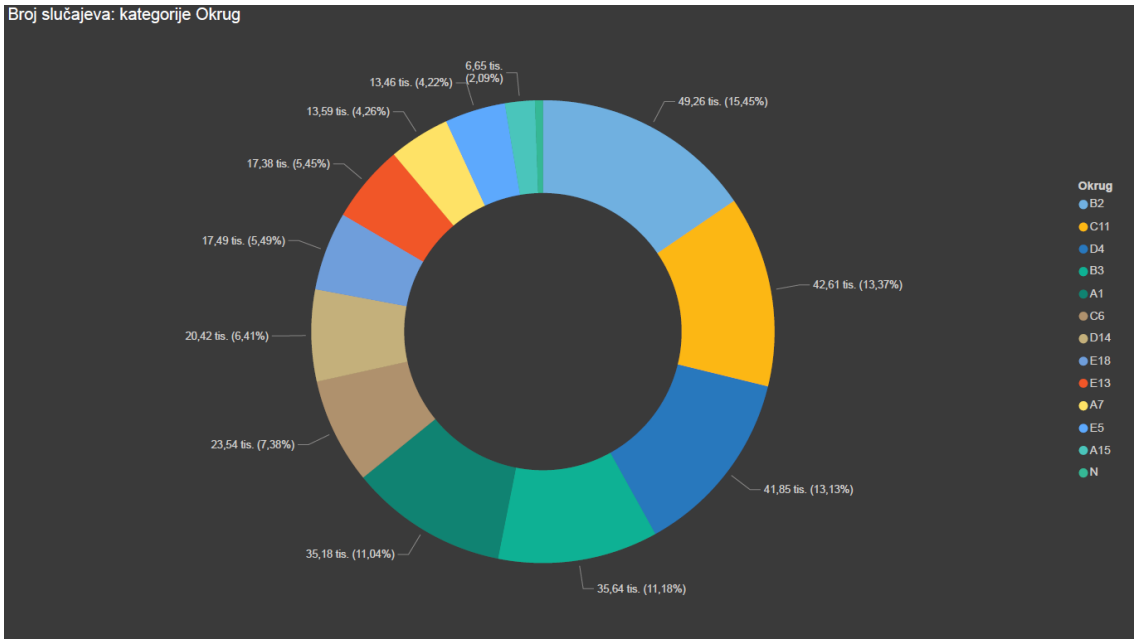


Slika 34. – graf- broj slučaja prema vrsti prekršaja

4.8 Broj slučaja grupiranim po okruzima

Prstenasti grafikon koji prikazuje broj slučaja koji su se desili prema okruzima. Ima 12 okruga, to jest 13 jer jedan je „N“ za vrijednosti koje nisu upisane.

Implementacija grafa je slična., korišten je atribut district iz dimenzije dim_district te je povezan sa ključem crimes_sk iz tablice činjenica.



Slika 35. – graf- broj slučaja prema okruzima

5. Zaključak

Ovim projektom možemo zaključiti koliko je ustvari skladištenje podataka vrlo široka pojam, ima puno načina za pohranu podataka. Isto tako poslovna inteligencija je jako bitna stvar danas, pogotovo kod velikih organizacija koje imaju velike prihode i rashode.

Bitno je pratiti sve te podatke, da znamo „sutra“ kakvu ćemo odluku donositi. Sustavi poslovne inteligencije mogu nam odgovoriti na pitanja; „što se dogodilo?“, „što će se dogoditi?“ i „što treba poduzeti?“ a ta tri pitanja su ključna za donošenje prave odluke. Isto tako treba paziti da su podaci točni jer u suprotnom može doći do donošenja krive odluke, što rezultira gubitcima.

Projektom „Kriminal u Bostonu“ najviše smo se fokusirali na deskriptivnu analitiku a to je „što se dogodilo?“. Što se dogodilo od 2015.-2018. godine?

Kada smo pretvorili podatke CSV datoteke u relacijsku bazu podataka, pa iz nje pretvarali u dimenzijske, u konačnici dobili smo ustvari grafove i dijagrame iz kojih možemo malo bolje, i točnije popratiti kad,šta i gdje se dogodilo. Ovime bi mogli smanjiti postotak kriminala tako što bi pojačali ophodnje na lokacijama gdje se događa najviše slučajeva.

6. Literatura

1. Business Intelligence – BI

Link : https://www.investopedia.com/terms/b/business-intelligence-bi.asp

2. Business Intelligence

Link : https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI

3. OLAP

Link: https://olap.com/olap-definition/

4. Kaggle – izvor podataka

Link : https://www.kaggle.com/AnalyzeBoston/crimes-in-boston?select=crime.csv

5. Microsoft Power BI

Link 1: https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Link 2: https://powerbi.microsoft.com/en-us/