

Analiza saobraćajnih nezgoda na teritoriji Republike Srbije

Seminarski rad u okviru kursa Poslovna inteligencija
Univerzitet u Novom Sadu
Prirodno-matematički fakultet

Nikola Pujaž - 21m/19
nikola.pujaž@gmail.com

29. septembar 2024.

Sadržaj

1	Uvod	2
2	Skup podataka	2
3	Motivacija istraživanja	2
4	Ciljevi rada	3
5	Učitavanje podataka i preprocesiranje	3
6	Analiza skupa podataka	4
6.1	Analiza vremena događanja saobraćajnih nezgoda	4
6.2	Analiza tipova nezgoda, broja i statusa učesnika	7
6.3	Analiza nezgoda na osnovu teritorija opština	9
7	Statistički testovi	11
8	Obrada velikih podataka	13
9	Zaključak	14
	Literatura	14

1 Uvod

U ovom izveštaju su predstavljeni rezultati istraživanja saobraćajnih nezgoda na teritoriji Republike Srbije. Istraživanje je sprovedeno sa intuitivnim ciljem pronalaženja odgovora na pitanja koji se odnose na koorelaciju tipova saobraćajnih nezgoda sa njihovim geografskim položajem, kao i uvidom u aktuelno stanje bezbednosti u saobraćaju na teritoriji Republike Srbije.

Za uspešnu realizaciju ovih ciljeva korišćeno je *R studio* okruženje, zajedno sa programskim jezikom *R*, njegovim konstrukcijama i bibliotekama. Implementacija rešenja se nalazi u prilogu ovog izveštaja. Konstante i promenljive su na Engleskom jeziku radi uniformnosti. U narednoj sekciji biće opisan skup podataka koji je korišćen i elementi infrastrukture koji su bili neophodni za realizaciju ovog rada.

2 Skup podataka

Podaci za ovo istraživanje su preuzeti sa portala otvorenih podataka vlade Republike Srbije na [2] i portala *GeoSrbija* na [1].

Skup podataka [2] sadrži podatke o saobraćajnim nezgodama i podeljen je po datumu ažuriranja na *CSV* datoteke. Konkretno u ovom radu korišćeni su podaci u periodu od 2015. do 2023. godine. Podaci u ovom skupu su podeljeni na 9 kolona od kojih prvu kolonu predstavlja jedinstveni broj saobraćajnog slučaja, zatim kolonu sa podacima o nadležnoj policijskoj upravi, opštini, datumu registrovane prijave nesreće, dve kolone namenjene geolociranim koordinatama nesreće, kategoriju procene štete, kategoriju broja učesnika, kao i klasifikaciju saobraćajne nezgode.

Pridruženi skup podataka sadrži prostorne informacije o teritoriji gradova na teritoriji Republike Srbije u *CSV* formatu.

Da bi bilo moguće doneti adekvatne zaključke u vezi sa koorelacijama između pojedinačnih podataka saobraćajnih nezgoda, bilo je potrebno dodatno formatirati, proširiti i ponovo definisati delove skupova podataka. Učitavanje i preprocesiranje ovih podataka je objašnjeno u okviru sekcije 5.

3 Motivacija istraživanja

Suštinska motivacija ovog istraživanja se može objasniti iz nekoliko perspektiva. Prva prespektiva podrazumeva poboljšanje saobraćajne bezbednosti identifikacijom faktora koji doprinose dešavanju saobraćajnih nezgoda. Ovi faktori mogu poslužiti kao osnova za donošenje adekvatnih mera sa ciljem smanjenja rizika od dešavanja nesreća.

Druga perspektiva podrazumeva potencijalno poboljšanje saobraćajne infrastrukture. Geolociranje kritičnih tačaka u saobraćajnoj infrastrukturnoj mreži može biti bitan indikator za njenu poboljšanje, optimizacijom mreže puteva, unapređivanjem signalizacije ili rekonstrukcijom kritičnih deonica.

Treća perspektiva podrazumeva smanjenje ekonomskog štete. Visoki troškovi sanacije saobraćajnih nezgoda mogu potencijalno biti umanjeni primenom preventivnih mera na osnovu identifikovanih faktora i obrazaca.

Četvrta perspektiva podrazumeva korišćenje računarskih tehnologija u oblasti formiranja infrastrukture, predloga, propisa i zakona, kao i saobraćaja generalno. U narednoj sekciji biće predstavljeni ciljevi ovog rada.

4 Ciljevi rada

Rad na ovim skupovima podataka omogućava konkretnu deskriptivnu analizu i identifikaciju teritorijalnih zona u kojima se dešavaju saobraćajne nezgode određenog tipa i određenog ishoda, stoga ovo predstavlja jedan od primarnih ciljeva ovog rada.

Na rezultatima prethodno opisanog cilja formiran je naredni cilj rada koji podrazumeva identifikaciju koorelacije između određenih tipova saobraćajnih nezgoda, njihovih ishoda i broja učesnika, što je omogućeno adekvatnom konstrukcijom nominalnih atributa skupa podataka.

5 Učitavanje podataka i preprocesiranje

Za adekvatnu analizu podataka bilo je prvenstveno potrebno učitati podatke koristeći se tipom radnog okvira (*eng. dataframe*). Zbog velikog broja podataka podeljenih u datoteke, podaci su učitavani u maniru delova (*eng. batches*).

Prilikom učitavanja skupa podataka, formirane su funkcije *classify_accident_type*, *classify_participants_nominal_count*, *classify_participants_status*, koje jednostavno enkodiraju vrednosti ovih kolona u celobrojne numeričke vrednosti, pri čemu su vrednosti kolone *accident_type*, koja predstavlja tip saobraćajne nezgode date na osnovu deskripcije date u koloni *description*.

Sa radnim okvirom saobraćajnih nezgoda ujedinjena i informacija o gradu na kome je nezgoda prijavljena, ukoliko se nezgoda dogodila na teritoriji grada. Ovo je postignuto korišćenjem *PostgreSQL* baze podataka i njenog proširenja *PostGIS*. Usled nemogućnosti funkcija iz datoteke *sf* da adekvatno prepoznaju geometrijski tip, *PostGIS* baza podataka je korišćena kao medijum za izvršavanje funkcije *ST_Cover* koja proverava da li se geolokaci-

ja saobraćajne nezgode, izražena kao tačka definisana geografskom dužinom i širinom, nalazi u okviru geometriskoj poligona koji opisuje teritoriju grada, a zatim je naziv tog grada dodeljen koloni *city_name* koja će kasnije biti korišćena u istraživanju.

Filtriranjem radnog okvira, uzete su u obzir samo kolone od značaja za ovo istraživanje, a to su: *case_number*, *police_department*, *reported_on*, *longitude*, *latitude*, *accident_type*, *participants_status*, *participants_nominal_count*, *municipality_name*, *city_name*.

Nakon što su izlistane osnovne informacije o radnom okviru korišćenjem funkcija *summary* i *head* kojom se dobija prvih nekoliko redova radnog okvira, izlistane su nedostajuće vrednosti po kolonama. Izveden je zaključak da skup sadrži saobraćajne nezgode sa različitim geolokacijama na teritoriji Republike Srbije te instance skupa podataka koje imaju nedostajuće vrednosti u koloni *city_name* ne treba uklanjati jer se jednostavno nisu dogodile na teritoriji grada. Sa druge strane, zabeleženo je 5 instanci koje nemaju vrednost u koloni *reported_on*, koja predstavlja datum i vreme kada je nezgoda prijavljena. Ove instance su uklonjene iz skupa podataka, a sa ovim korakom skup podataka je pripremljen za dalju analizu.

Bitno je napomenuti da duplikati nisu uklanjeni iz skupa podataka, jer su manuelnim pregledom skupa uočeni različiti tipovi saobraćajnih nezgoda zavedeni pod istim jedinstvenim brojem slučaja, pa je na taj način izbegnut gubitak dragocenih podataka.

6 Analiza skupa podataka

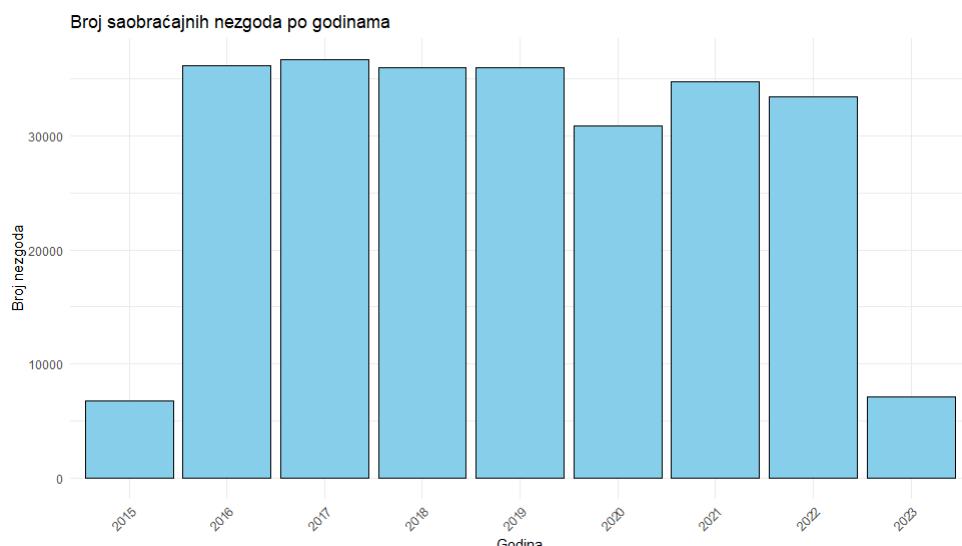
Prvi korak u analizi skupa podataka je bio utvrđivanje broja instanci skupa nakon procesa pretpreparacije. Broj instanci skupa podataka, tj. saobraćajnih nezgoda uzetih u razmatranje, iznosio je 257690.

6.1 Analiza vremena događanja saobraćajnih nezgoda

Konvertovanjem kolone *reported_on* u datumski tip omogućeno je utvrđivanje najranijeg i poslednjeg zabeleženog saobraćajnog incidenta. U ovom skupu prva nezgoda je prijavljena 1. januara 2015, a poslednja 31. marta 2023. godine.

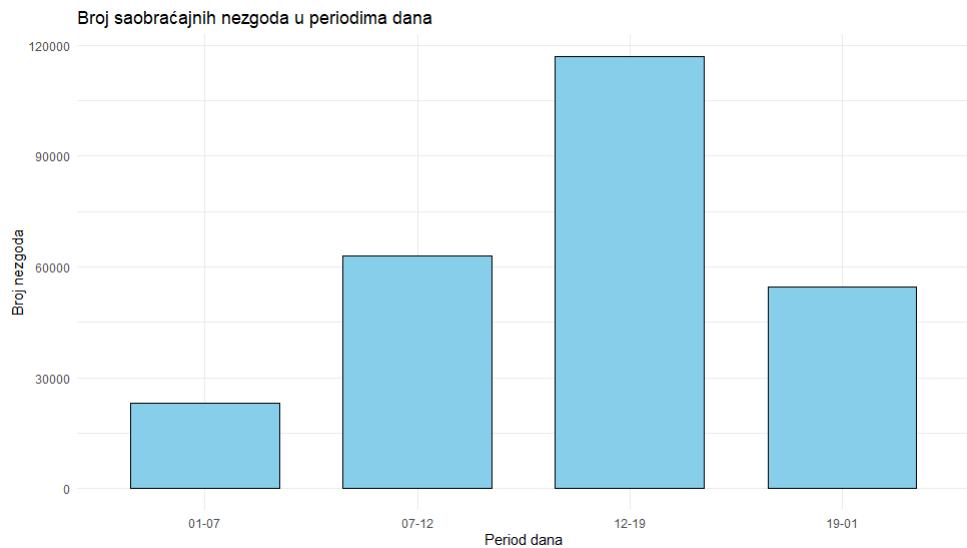
Na slici 1 prikazana je distribucija prijavljenih saobraćajnih nezgoda prema godinama iz čega se može zaključiti da je za godine između 2016 i 2022, osim za 2020. godinu, prijavljeno minimum 30000 saobraćajnih nezgoda, pri čemu je blago odstupanje primećeno samo za pomenutu 2020. godinu. Interesantan podatak je da je iste godine saobraćaj bio ograničen vanrednim sta-

njem tokom pandemije korona virusa, međutim, ova koorelacija se ne može jednostavno utvrditi. Najveći broj saobraćajnih nezgoda u ovom skupu je prijavljen tokom 2017. godine. Manji broj zabeleženih saobraćajnih nezgoda tokom 2015. godine verovatno ukazuje na uspostavljanje mehanizma prikupljanja i formiranja skupa podataka, koji se stabilizuje narednih godina. Sa druge strane, broj nezgoda za 2023. godine je znatno manji jer pokazuje podatke samo za prvi kvartal pomenute godine.



Slika 1: Broj saobraćajnih nezgoda po godinama

Daljim ispitivanjem vrednosti kolone koja predstavlja datum i vreme prijavljene nezgode odnosi se na period dana u kome su nezgode prijavljene. Izabrana su četiri perioda dana - jutarnji od (7 do 12 časova), popodnevni (od 12 do 19 časova), večernji (od 19 do 1 čas posle ponoći) i noćni (od 1 čas posle ponoći do 7 časova). Na osnovu informacija iz medija postoji pretpostavka da se saobraćajne nezgode najviše događaju tokom noći, pri čemu se napominje i konzumacija alkohola kao bitan faktor uticaja, međutim, iz podataka prikazanih na slici 2 se jasno vidi da to nije slučaj, već da se najviše saobraćajnih nezgoda dešava u periodu izmeđi podneva i večernjih časova. Intuitivno se može zaključiti da je razlog ovakve pojave poprilično direkstan, zbog toga što najveća cirkulacija saobraćaja je upravo u ovom periodu dana.



Slika 2: Broj saobraćajnih nezgoda u periodima dana

Iz slike 3, sa predstavljenog *pie* dijagrama se može videti procenat saobraćajnih nezgoda u odnosu na izabrani period dana, pri čemu 45.4% saobraćajnih nezgoda se dogodi između 12 i 19 časova, 24.4% između 7 i 12 časova, 21.2% između 19 i 1 čas posle ponoći, a 9% tokom noći između 1 i 7 časova.



Slika 3: Procenat saobraćajnih nezgoda u periodima dana

6.2 Analiza tipova nezgoda, broja i statusa učesnika

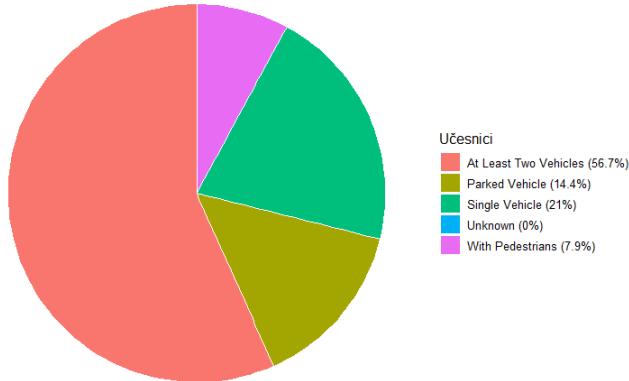
Istraživanje se u ovoj sekciji fokusira na tipove saobraćajnih nezgoda, broja učesnika, kao i statusa učesnika nakon nesreće. Uvidom u procentualni prikaz saobraćajnih nezgoda prema statusu učesnika na slici 4, dolazi se do zaključka da je najveći broj saobraćajnih nezgoda okarakterisan kao nezgoda sa materijalnom štetom (59.8%). Nezgode u kojima su ishodi okarakterisani kao nezgode sa povređenima iznose 38.8% od ukupnog broja nezgoda, dok nezgode sa nastrandalim licima iznose 1.5%.



Slika 4: Procenat različitih ishoda saobraćajnih nezgoda

Naredna figura 5 daje uvid u procentualni pregled broja učesnika u saobraćaju. Zabeženo je 56.7% saobraćajnih nezgoda u kojoj su učestvovala bar dva vozila, dok nezgode sa jednim vozilom iznose 21%. Nezgode sa parkiranim vozilima iznose 14.4%, dok procenat nezgoda u kojima su učestvovali pešaci iznosi 7.9%

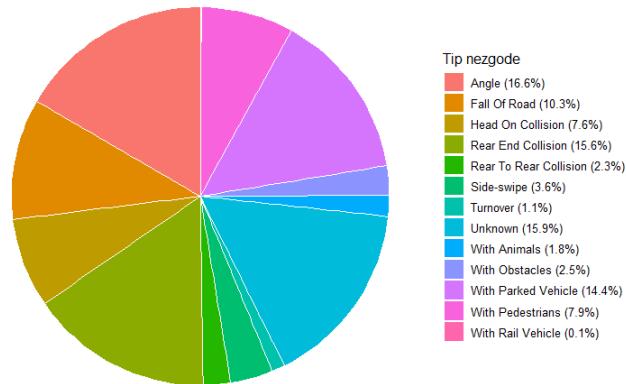
Procenat broja učesnika u saobraćajnim nezgoda



Slika 5: Procenat različitih ishoda saobraćajnih nezgoda

Finalna *pie* figura 6 predstavlja procentualni prikaz tipova saobraćajnih nezgoda, pri čemu su bočni sudari pod uglom (*eng. Angle*), sudari pri sustizanju (*Rear End Collision*), kao i sudari sa parkiranim vozilima (*eng. With Parked Vehicle*) zabeleženi kao tri najčešća tipa saobraćajnih nezgoda i iznose 16.6%, 15.6% i 14.4% redom. Bitno je napomenuti da sličan deo zauzimaju i neklasifikovane nezgode - 15.9%.

Procenat tipova saobraćajnih nezgoda

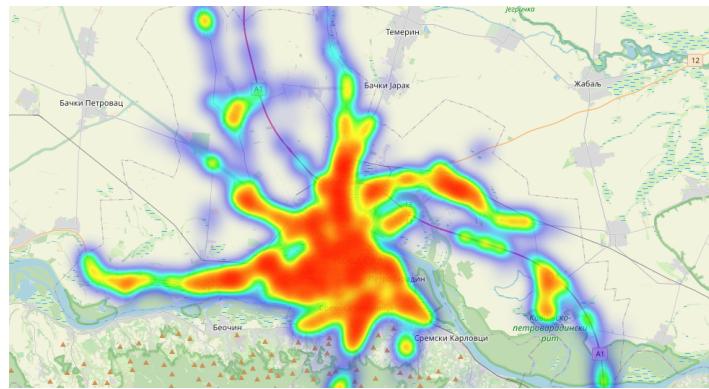


Slika 6: Procenat različitih ishoda saobraćajnih nezgoda

6.3 Analiza nezgoda na osnovu teritorija opština

U ovoj podsekciji razmatrane su saobraćajne nezgode na osnovu teritorije opština na kojoj su se dogodile sa ciljem odgovora na pitanje na kojim opštinama se događa najviše saobraćajnih nezgoda sa određenim ishodom i kog su tipa te nezgode.

Koristeći se konverzijom, kategorizacijom i agregacijom podataka uočeno je da se najviše saobraćajnih nezgoda u kojima su učesnici pretrpeli povredu dogodio na teritoriji opštine **Novi sad**, pri čemu je zabeleženo 7623 nezgode od kojih 1617 pripada tipu nezgoda koje su se dogodile sudsatom pod uglom. Slikoviti prikaz pomenutih izdvojenih nezgoda dobijen je korišćenjem *heatmap* isrtavanja korišćenjem funkcija biblioteke *leaflet* i prikazan je na slikama 7 i 8.

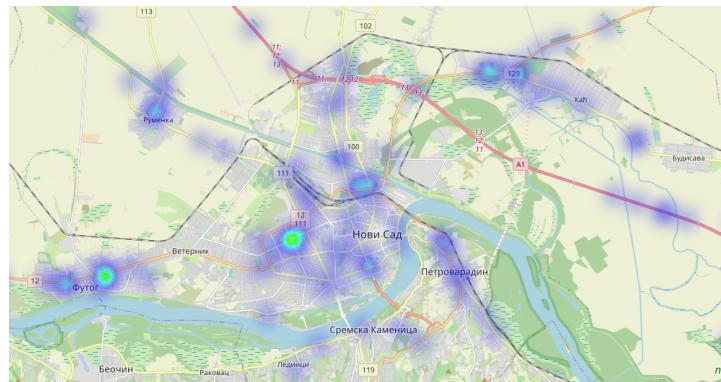


Slika 7: *Heatmap* nezgoda sa povredenima na teritoriji opštine Novi Sad

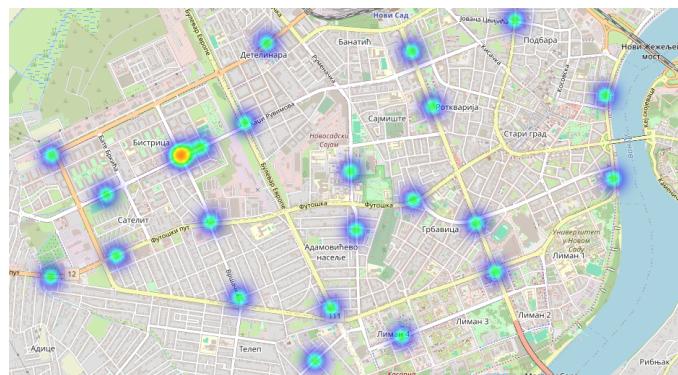


Slika 8: *Heatmap* nezgoda sa povredenima na teritoriji opštine Novi Sad

Interesantno je da je pomenutim tehnikama opština Novi Sad ponovo izdvojena kao opština sa najvećim brojem saobraćajnih nezgoda u kojima je bilo žrtava (130 nezgoda), a u kojima su najviše učestvovali pešaci (43. nezgode). Slike 9 i 10 sadrže *heatmap* prikaz geolokacije izdvojenih nezgoda.



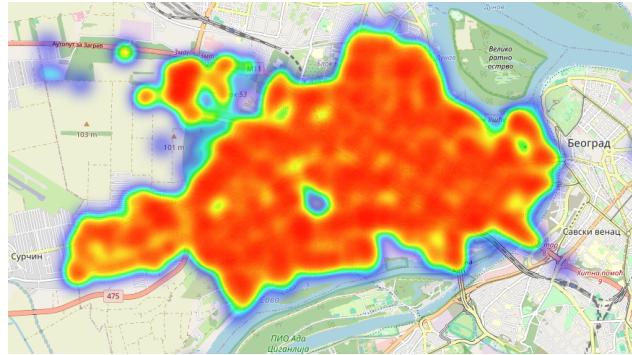
Slika 9: *Heatmap* nezgoda sa nastrandalima na teritoriji opštine Novi Sad



Slika 10: *Heatmap* nezgoda sa nastrandalima na teritoriji opštine Novi Sad

Ovaj podataka ukazuje na potencijalno neophodnu evaluaciju saobraćajne infrastrukture namenjene pešacima na teritoriji opštine Novi Sad.

Opština **Novi Beograd** je izdvojena kao opština sa najviše saobraćajnih nezgoda sa ishodom materijalne štete (13164 nezgode), u kojima je sudar pri sustizanju zabeležen kao najčešći tip nezgode (3266 nezgoda). I u ovom slučaju je grafički prikazana geolokacijska učestalost ovih nesreća korišćenjem *heatmap* prikaza na slikama 11 i 12.



Slika 11: *Heatmap* nezgoda sa nastradalima na teritoriji opštine Novi Sad



Slika 12: *Heatmap* nezgoda sa nastradalima na teritoriji opštine Novi Sad

7 Statistički testovi

Statistički testovi pružaju mogućnost donošenja adekvatnih zaključaka o međusobnoj zavisnosti između atributa koji nisu uočljivi iz rezultata dobijenih primenom tehnika i metoda deskriptivne statistike.

Kako je najveći broj atributa u skupu podataka koji su predmet ovog istraživanja nominalnog (kategoričkog) tipa, pogodan test za utvrđivanje međusobne zavisnosti između određenih atributa predstavlja Hi-kvadrat test (*eng. Chi-squared test*) primjenjen na osnovu definisanih hipoteza u nastavku.

Izdvojena su dva istraživačka pitanja sa ciljem utvrđivanja koja potvrđuju jednu od definisanih prepostavki. Na nivou opština postavljeno je pitanje međusobne zavisnosti između opštine na kojoj je zabeležena nezgoda i tipa nezgode:

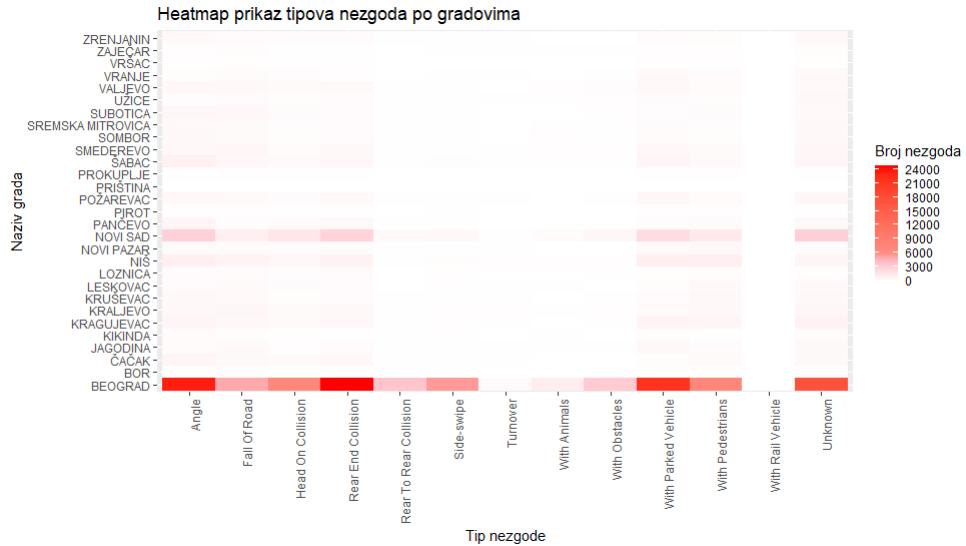
- H_0 : Informacija o opštini na kojoj je zabeležena saobraćajna nezgoda je nezavisan atribut u odnosu na tip saobraćajne nezgode. (*nulta hipoteza*)
- H_1 : Tip saobraćajne nezgode je zavisan atribut u odnosu na opštinu na kojoj je zabeležena saobraćajna nezgoda. (*alternativna hipoteza*)

Dodatno, definisane su dodatne hipoteze u slučaju gradova:

- H_0 : Za saobraćajne nezgode koje su zabeležene na teritoriji grada u Republici Srbiji, informacija o gradu na kome je zabeležena nezgoda je nezavisan atribut u odnosu na tip saobraćajne nezgode. (*nulta hipoteza*)
- H_1 : Saobraćajne nezgode određenog tipa se dešavaju češće na određenim teritorijama gradova (tip saobraćajne nezgode zavisi od grada na kome je zabeležena nezgoda). (*alternativna hipoteza*)

Razmatranjem rezultata Hi-kvadrat testa, u oba slučaja je zabeležena visoka razlika između očekivanih i posmatranih ferkvencija u unakrsnoj tabeli (*eng. contingency table*), što implicira visokoj zavisnosti između atributa. Kako je vrednost p manja od standardnog praga od 0.05 ($p = 0.0002$) u oba slučaja, može se pristupiti odbacivanju nulte hipoteze i u slučaju kada su to u pitanju opštine, i u slučaju kada su to u pitanju gradovi. Kako alternativna hipoteza stoji, može se zaključiti da postoji zavisnosti između teritorijalnog obeležja i tipa saobraćajne nezgode.

Dodatno, *heatmap* grafikom 13 je prikazana unakrsna tabela gradova i tipova nezgoda, pri čemu se vidi da se Beograd najviše ističe kao grad sa najvećem brojem saobraćajnih nezgoda, pogotovo kada je reč o tipovima sudara pod uglom, sudara sa parkiranim vozilima, kao i sudarima pri sustizanju. U više gradova je prisutan značajan broj nezgoda kategorisanih sa nepoznatim tipom nezgode. Ovo može biti pokazatelj da bi trebalo dodatno analizirati ili poboljšati sistem prijavljivanja saobraćajnih nezgoda.



Slika 13: Heatmap unakrsne tabele

8 Obrada velikih podataka

Tehnike obrade velikih podataka omogućavaju još dublju spoznaju o podacima, pružajući mogućnost donošenja zaključaka na osnovu nelinearnih modela koji izdvajaju nelinearne obrazce iz skupova podataka.

Korišćenjem *apriori* pravila - pravila asocijacije formiran je model pravila koji utvrđuje da li se na osnovu saobraćajne nezgode, broja učesnika u saobraćaju i ishoda saobraćajne nezgode mogu izvesti pravila koja ukazuju na to da vrednosti jednog impliciraju u drugo. Dodatno, definisane se naredne vrednosti metrika: *support = 0.01, confidence = 0.8*.

Ovom tehnikom je izdvojeno narednih 6 pravila sa visokim stepenom pouzdanosti i *lift* vrednosti, odnosno stepena zavisnosti. Pravila na osnovu vrednosti kolone *participants_nominal_count*:

1. With Pedestrians (Sa pešacima) => Injured (Sa povređenima)
2. Parked Vehicle (Sa parkiranim vozilima) => Material Damage (Sa materijalnom štetom)

Pravila na osnovu vrednosti kolone *participants_nominal_count* i kolone *municipality_name*:

1. NOVI BEOGRAD, Parked Vehicle (Sa parkiranim vozilima) => Material Damage (Sa materijalnom štetom)

2. VRAČAR, At Least Two Vehicles (Sa dva vozila) => Material Damage (Sa materijalnom štetom)
3. STARI GRAD, At Least Two Vehicles (Sa dva vozila) => Material Damage (Sa materijalnom štetom)

Pravila samo na osnovu vrednosti kolone *municipality_name*:

1. STARI GRAD => Material Damage (Sa materijalnom štetom)

9 Zaključak

U okviru ovog istraživanja urađena je analiza skupa podataka saobraćajnih nezgoda na teritoriji Republike Srbije. Zahvaljujući tehnikama deskriptivne i inferencijske statistike, kao i tehnikama obrade velikih podataka, pružen je detaljniji uvid u stanje saobraćaja u državi Srbiji, pri čemu su dodatno ustanovljene bitne koorelacije između određenih podataka u skupu. Ovaj proces je započet utvrđivanjem međusobne zavisnosti između tipa saobraćajne nezgode i teritorije na kojoj je ona zabeležena, kao i izdvajanjem *apriori* pravila koja ukazuju na direktnu implikaciju između broja učesnika i opštine na kojoj je zabeležena nezgoda sa ishodom saobraćajne nezgode. Ove koorelacije mogu služiti kao osnova za dalju i dublju analizu, a samim tim i precizniju identifikaciju činioca koji mogu poslužiti prevenciji događanja saobraćajnih nezgoda.

Literatura

- [1] GeoSrbija Nacionalna Infrastruktura Geoprostornih Podataka. <https://opendata.geosrbija.rs/>. [Accessed 29-09-2024].
- [2] Ministarstvo unutrašnjih poslova Republike Srbije. Podaci o saobraćajnim nezgodama po policijskim upravama i opštinama - otvoreni podaci — data.gov.rs. <https://data.gov.rs/>. [Accessed 29-09-2024].

Prilog

Uz ovaj rad su priložene dve *R* datoteke sa nazivima:

- *preprocessing.R*
- *data-analysis.R*