

Seminarska naloga 2

Računalniška podpora statistike

Alen Kahteran

22. 11. 2020

Uvod

Pred kratkim sem si hotel odgovoriti na vprašanje ali obstaja kakšna razlika med moškimi in ženskami, ki povzročijo nesrečo z osebnim avtomobilom v različnih okoliščinah.

Podatke o vseh (od leta 1995 dalje) prometnih nesrečah v Sloveniji, lahko dobimo na spletni strani policije¹. Odločil sem se, da raziščemo podatke za leto 2020, saj je možno, da je kje opazen tudi vpliv COVID-19. Seveda za ta namen bi bila potrebna podrobnejša analiza, ter primerjava s prejšnjimi leti, vendar je v vsakem primeru najprej potrebno analizirati le letošnje podatke. Trenutni podatki vsebujejo prometne nesreče le do konca avgusta 2020.

Naše podatke bomo primarno delili na moške in ženske, za tem pa še na posamezne skupine, glede na določeno spremenljivko. Okoliščine so lahko različnih oblik. Večina je povezanih s samo prometno nesrečo (gostota prometa, lokacija, vremenske razmere, itd.), nekaj pa jih je z voznikom (starost, vozniški staž, itd.).

Najprej je bilo podatke korektno pripraviti za obdelavo.

Čiščenje podatkov

Preden se kakorkoli dotaknemo podatkov, je potrebno vedeti kakšne podatke sploh imamo. Tu nam je policija poleg podatkov, pripravila tudi opise (in formate) vseh spremenljivk. Žal niso točno povedali kateri opis pripada kateri spremenljivki, vendar je že samo ime spremenljivke povedalo večino informacij. Tako da v tabeli 1 vidimo imena spremenljivk, njihove opise ter ali so bile uporabljene za analizo ali ne.

Za večino spremenljivk je bil podan format. Ponekod, kjer je bil podatek ločen na dva stolpca, kot npr. `DatumPN` in `UraPN`, smo ustrezno pretvorili v datum z uro.

Spremenljivke ki smo jih obdržali so bile `ZaporednaStevilkaPN`, `DatumPN`, `VNaselju`, `VremenskeOkoliscine`, `Starost`, `Spol` in `VrednostAlkotesta`. `Povzrocitelj`, `VrstaUdelezenca` in `Drzavljanstvo` sem uporabil le toliko, da smo dobili pravilen vzorec naših začetnih podatkov (povzročitelje nesreč, ki so vozili osebni avtomobil in imajo slovensko državljanstvo).

Ostale spremenljivke so ali preveč razdrobljene, ali neuporabne saj se osredotočamo na slovensko populacijo, ali pa vezane na posledice nesreče.

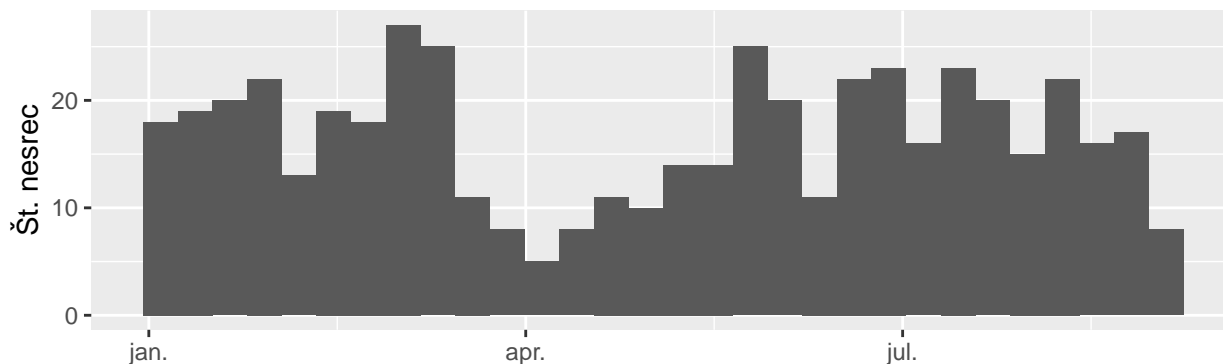
¹<https://www.policija.si/o-slovenski-policiji/statistika/prometna-varnost>

Tabela 1: Opisi spremenljivk

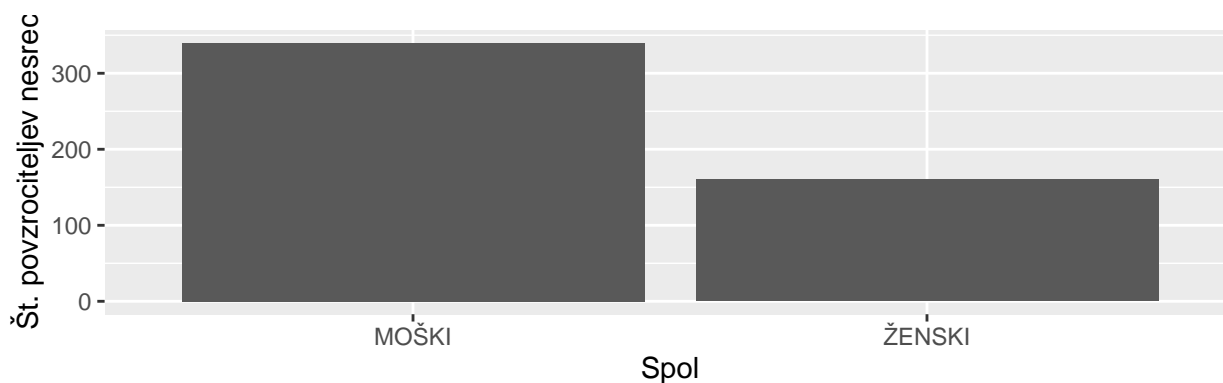
Ime spremenljivke	Opis spremenljivke (in morebiten format)	Uporaba
ZaporednaStevilkaPN	številka za štetje in ločevanje posamezne prometne nesreče	da
KlasifikacijaNesrece	klasifikacija nesreče glede na posledice (Izračuna se avtomatično glede na najhujšo posledico pri udeležencih v prometni nesreči)	ne
UpravnaEnotaStoritve	upravna enota, na območju katere se je zgodila prometna nesreča	ne
DatumPN	datum nesreče (format: dd.mm.llll)	da
UraPN	ura nesreče (format: hh)	ne
VNaselju	indikator ali se je nesreča zgodila v naselju (D) ali izven (N)	da
Lokacija	lokacija nesreče	ne
VrstaCesteNaselja	vrsta ceste ali naselja na kateri je prišlo do nesreče	ne
SifraCesteNaselja	oznaka ceste ali šifra naselja kjer je prišlo do nesreče	ne
TekstCesteNaselja	tekst ceste ali naselja, kjer je prišlo do nesreče	ne
SifraOdsekaUlice	oznaka odseka ceste ali šifra ulice, kjer je prišlo do nesreče	ne
TekstOdsekaUlice	tekst odseka ali ulice, kjer je prišlo do nesreče	ne
StacionazaDogodka	točna stacionaža ali hišna številka, kjer je prišlo do nesreče	ne
OpisKraja	opis prizorišča nesreče	ne
VzrokNesrece	glavni vzrok nesreče	ne
TipNesrece	tip nesreče	ne
VremenskeOkoliscine	vremenske okoliščine v času nesreče	da
StanjePrometa	stanje prometa v času nesreče	ne
StanjeVozisca	stanje vozišča v času nesreče	ne
VrstaVozisca	stanje površine vozišča v času nesreče	ne
GeoKoordinataX	Geo Koordinata X (Gauß-Krüger-jev koordinatni sistem)	ne
GeoKoordinataY	Geo Koordinata Y (Gauß-Krüger-jev koordinatni sistem)	ne
ZaporednaStevilkaOsebeVPN	številka za štetje in ločevanje oseb, udeleženih v prometnih nesrečah	ne
Povzročitelj	kot kaj nastopa oseba v prometni nesreči	da
Starost	starost osebe (LL)	da
Spol	spol	da
UEStalnegaPrebivalisca	upravna enota stalnega prebivališča	ne
Drzavljanstvo	državljanstvo osebe	da
PoskodbaUdelezenca	poškodba osebe	ne
VrstaUdelezenca	vrsta udeleženca v prometu	da
UporabaVarnostnegaPasu	ali je oseba uporabljala varnostni pas ali čelado (polje se interpretira v odvisnosti od vrste udeleženca) (Da/Ne)	ne
VozniskiStazVLetih	vozniški staž osebe za kategorijo, ki jo potrebuje glede na vrsto udeleženca v prometu (LL)	ne
VozniskiStazVMesecih	vozniški staž osebe za kategorijo, ki jo potrebuje glede na vrsto udeleženca v prometu (MM)	ne
VrednostAlkotesta	vrednost alkotesta za osebo, če je bil opravljen (n.nn)	da
VrednostStrokovnegaPregleda	vrednost strokovnega pregleda za osebo, če je bil odrejen in so rezultati že znani (n.nn)	ne

Pregled podatkov

Najprej si pogledjmo na sliki 1 porazdeljenost nesreč v času. V normalnih okoliščinah bi pričakovali, da so nesreče čez celo leto enakomerno porazdeljene (dogodki so časovno neodvisni). V našem časovnem obdobju (od 1. 1. 2020 do 31. 8. 2020) vidimo da temu ni tako, saj se je v sredini marca 2020 začelo ustavljanje javnega življenja zaradi COVID-19. Ker je takrat večina ljudi bila doma, in ne v avtu, je v našem časovnem obdobju pričakovano, da bo v obdobju, ko je bilo ustavljeno javno življenje, manj nesreč. Kljub temu to ne bi smelo vplivati na rezultate, saj se ne osredotočamo na to kdaj so se zgodili, temveč kdo jih je storil.



Slika 1: Porazdelitev nesreč v času.



Slika 2: Porazdelitev spola povzročiteljev nesreč.

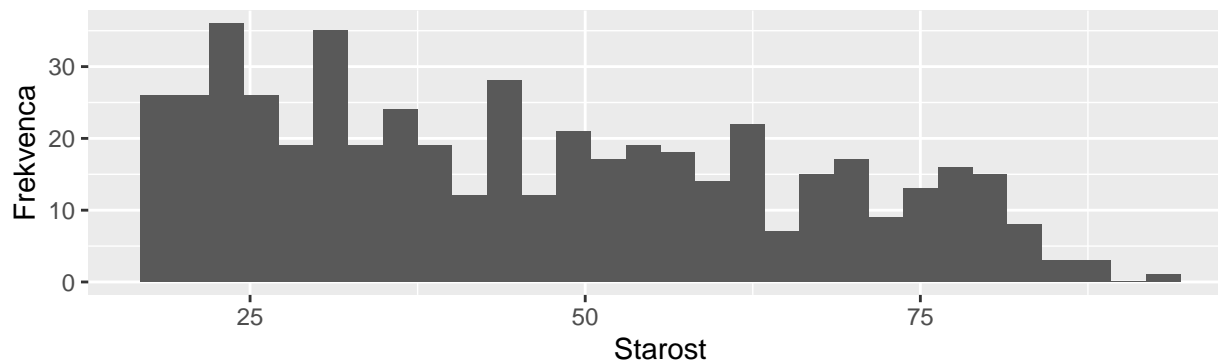
Zato si na sliki 2 najprej pogledjmo porazdelitev našega vzorca po spolu. Videti je da je moških približno dvakrat toliko kot žensk (moških - 340, žensk - 160). Pogledjmo si naš vzorec še za ostale spremenljivke.

Starost

Porazdelitev starosti lahko vidimo na sliki 3. Videti je da št. povzročiteljev nesreč pada s starostjo. Če primerjamo našo porazdelitev s starostno porazdelitvijo Slovenije², je videti da ta ne pada s starostjo. To padanje starosti bi mogoče lahko pripisali dvem stvarim, ali temu da so mlajši vozniki osebnih avtomobilov bolj pogosti, da povzročijo nesrečo (zaradi manj izkušenj), ali pa temu, da starejši ljudje manj vozijo avte.

Poglejmo si še opisne statistike starosti za oba spola, za moške in za ženske v tabeli 2 in še grafični prikaz opisnih statistik starosti za moške in ženske s škatlo z ročaji na sliki 4. Iz grafa okvirjev z ročaji, je videti da sta si porazdelitvi precej podobni. Videti je da so porazdelitve asimetrične (Zgornji del je daljši), in je zato mediana smiselna mera središčnosti, medtem ko je interkvartilni razmik smiselna mera razpršenosti. Žal

²<https://pxweb.stat.si/SiStatData/pxweb/sl/Data/-/05C5002S.px/>



Slika 3: Porazdelitev starosti.

iz teh slik ne moremo nič sklepati. Poleg tega je iz tabele razvidno, da naš vzorec vsebuje celoten spekter starosti, od mladih, ki so komaj opravili izpit, do starejših (75+).

Tabela 2: Opisne statistike starosti

	Skupno	Moški	Ženske
Min.	17.0	17.0	18.0
1st Qu.	28.0	28.0	30.0
Median	44.0	45.0	41.0
Mean	45.6	46.1	44.5
3rd Qu.	61.0	62.2	59.2
Max.	92.0	92.0	83.0
IQR	33.0	34.2	29.2

V naselju

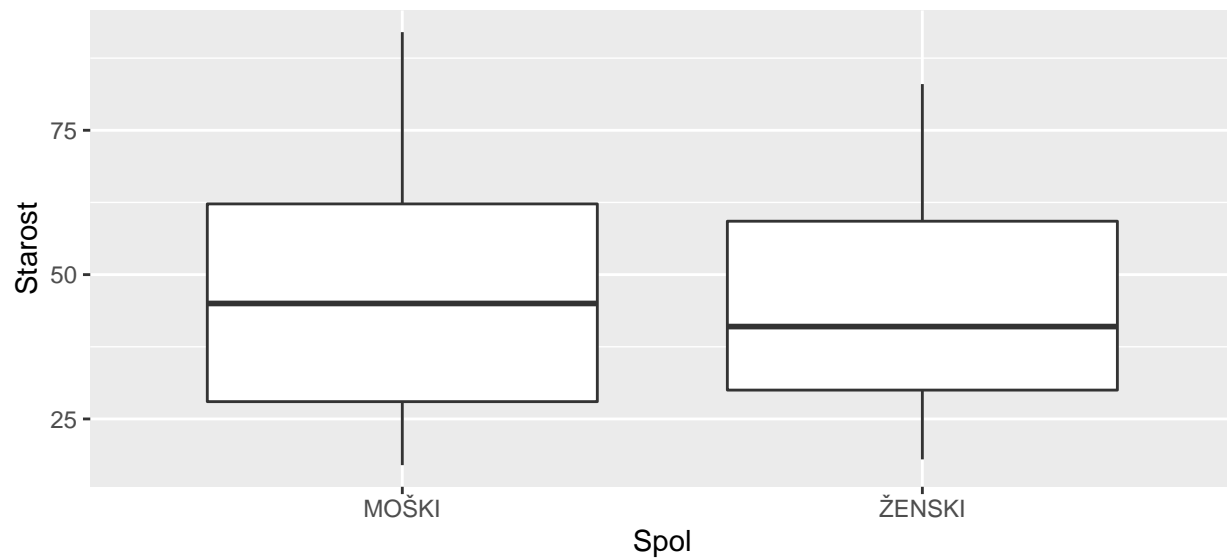
Enostaven način kako ločiti podatke kje se nesreča zgodi, je ali se je ta zgodila v naselju ali ne. Posledično si lahko pogledamo kje je več nesreč. Vidimo da se v našem vzorcu zgodi več nesreč v naselju (V naselju - 336, Izven naselja - 164). To je bilo pričakovano, saj se najverjetneje največ vožnje z avtomobilom zgodi ravno v naselju. Podoben razmislek je tudi da zato, ker je tam največ ljudi in avtov.

V tabeli 3 si oglejmo kontingenčno tabelo opazovanih frekvenc za spol povzročiteljev nesreč, ter ali je bila nesreča v naselju, ali ne. Po občutku bi rekli, da večjih odstopanj od pričakovanj (če gledamo vsote) ni. To bomo v naslednjem poglavju preverili s testom χ^2 .

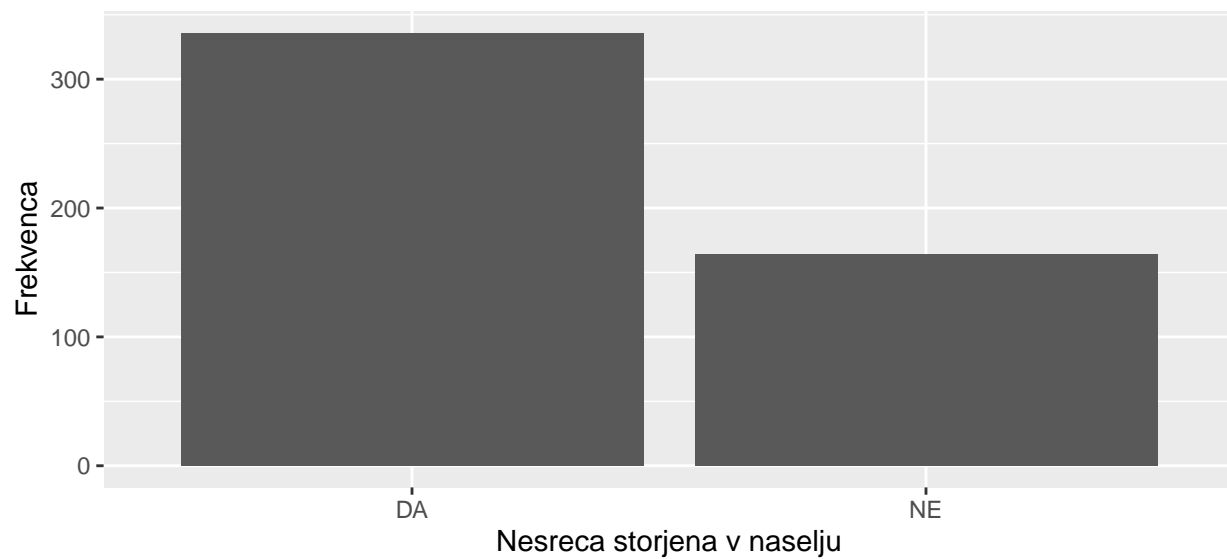
Tabela 3: Kontingenčna tabela za spol in ali je bila nesreča storjena v naselju

	MOŠKI	ŽENSKI	VSOTA
DA	223	113	336
NE	117	47	164
VSOTA	340	160	500

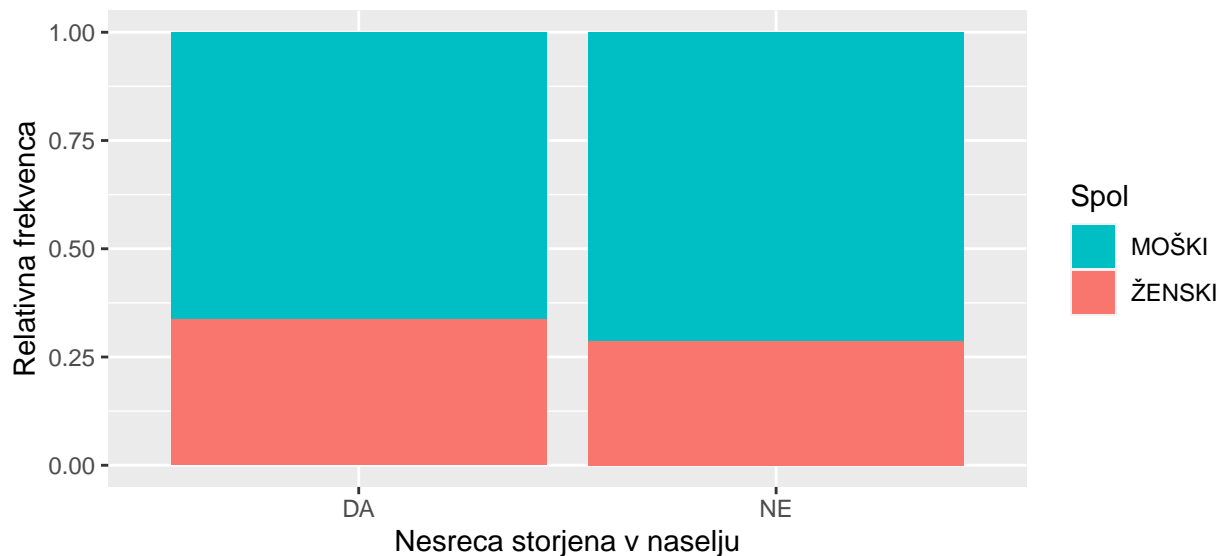
Podobno lahko sklepamo iz slike 6, saj vidimo da nekih velikih razlik v relativnih frekvencah ni.



Slika 4: Škatle z brki za starost po spolu.



Slika 5: Frekvenca ali je bila nesreča storjena v naselju ali ne.

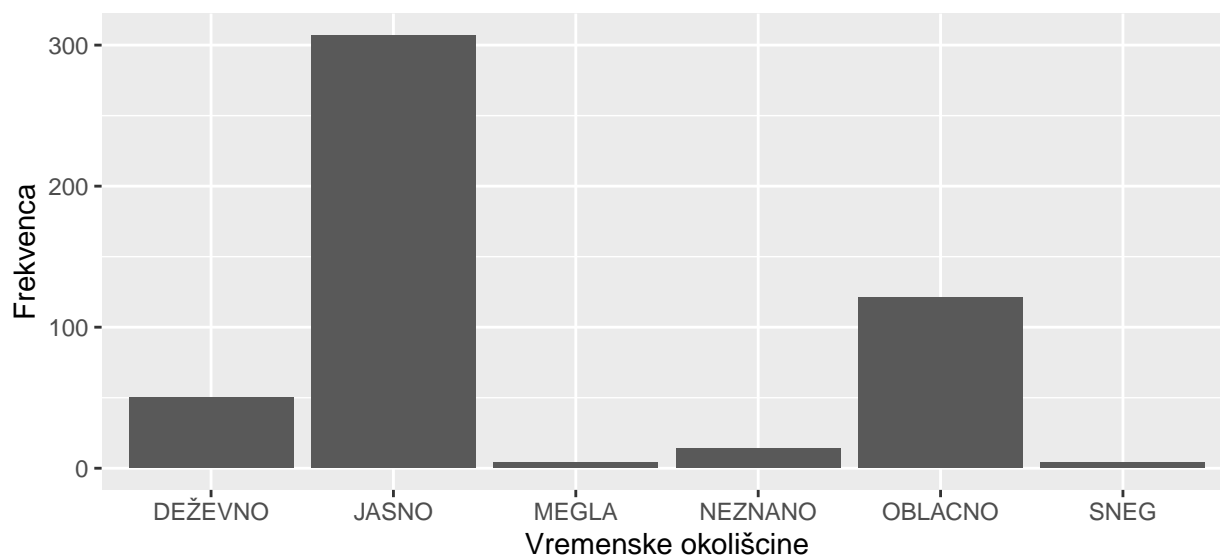


Slika 6: Relativen prikaz, ali je bila nesreča storjena v naselju ali ne, po spolu.

VremenskeOkoliscine

Pri vremenskih okoliščinah nas zanima, če obstajajo take vremenske okoliščine, ki bi pomenile da določen spol povzroči več nesreč v določenem vremenu.

Na sliki 7 so prikazane absolutne frekvence nesreč v različnih vremenskih okoliščinah. Vidimo da se največ nesreč zgodi pri vremenu JASNO.



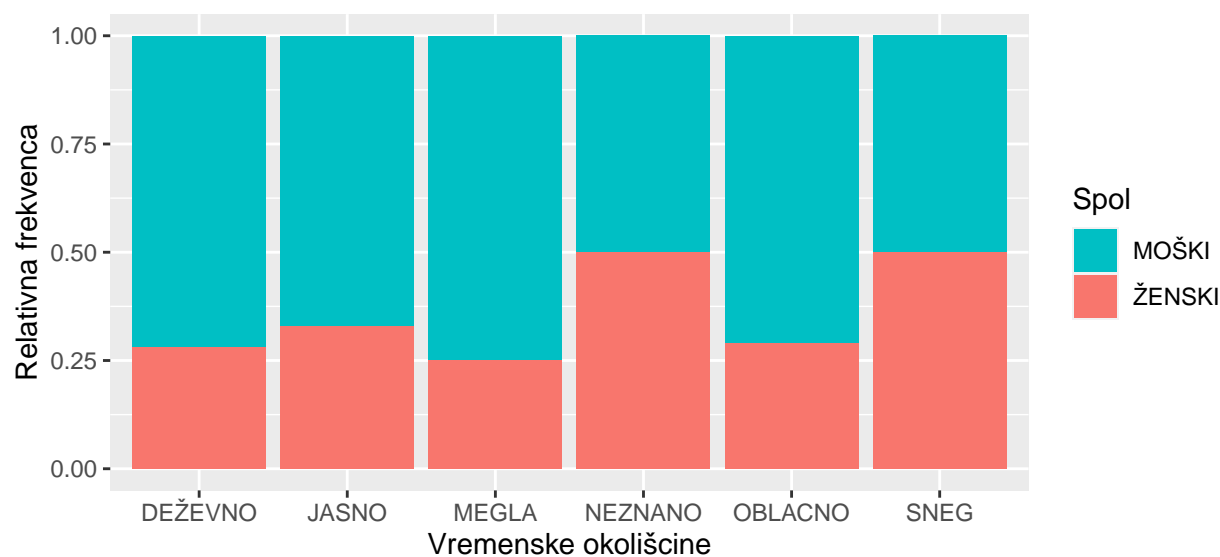
Slika 7: Frekvence vremenskih okoliščin.

Če si pogledamo še dejanske številke v tabeli 4, vidimo da so nekatere izmed skupin manjše od zahtevanega za χ^2 test (vsaj 80% skupin mora imeti vsaj 5 opazovanj). Tu lahko storimo dve stvari, ali odstranimo skupine, ki so manjše od 10 (nastavljeno na 10, da bo manj skupin) opazovanj, ali pa jih združimo v novo skupino - OSTALO. Sam sem se odločil da jih združim, saj bomo tako še vedno imeli 500 opazovanj. Še prej pa lahko

vidimo na sliki 8 relativne frekvence. Takoj vidimo, da **SNEG** in **NEZNANO** izstopata, kar je razložljivo s tem, da so tu absolutne frekvence zelo nizke (tabela 4), in ne moremo nič sklepati iz te slike.

Tabela 4: Kontingenčna tabela za spol in vremenske okoliščine

	MOŠKI	ŽENSKI	VSOTA
DEŽEVNO	36	14	50
JASNO	206	101	307
MEGLA	3	1	4
NEZNANO	7	7	14
OBLAČNO	86	35	121
SNEG	2	2	4
VSOTA	340	160	500

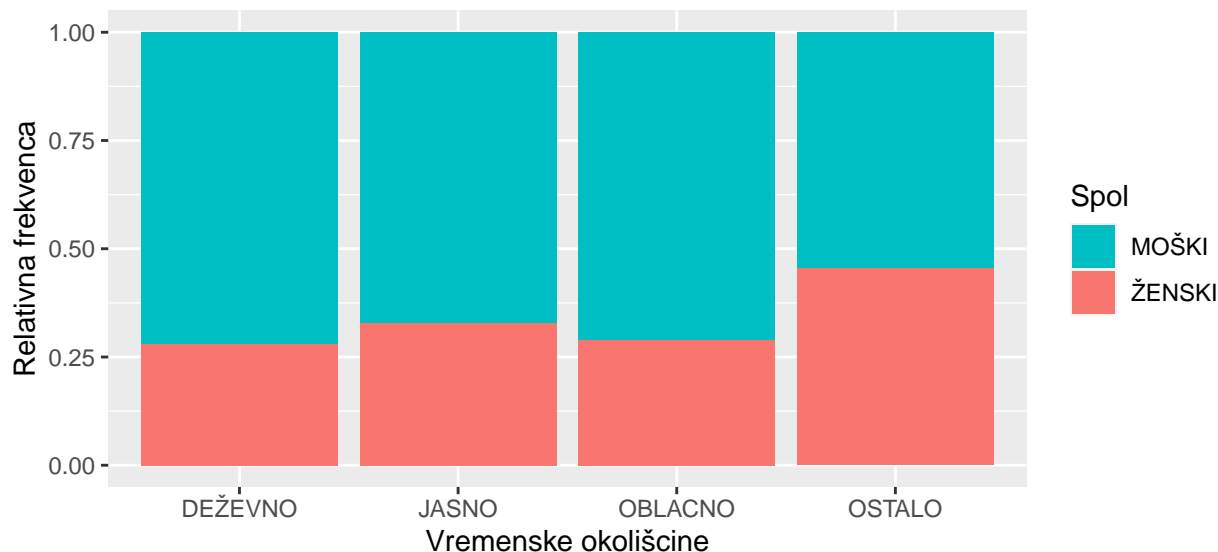


Slika 8: Relativen prikaz vremenskih razmer po spolu.

Zato si pogledjmo še spremenjeno tabelo (tabela 5), oziroma sliko (slika 9), kjer smo majhne skupine združili v skupino **OSTALO**. Vidimo da relativna frekvenca žensk v skupini **OSTALO** nekoliko odstopa, v primerjavi z ostalimi skupinami. Kot je razvidno iz tabele 5, je ta skupina še vedno nekoliko majhna, in je lahko to razlog za odstopanje. Je pa nova tabela primerna za χ^2 test, kar bomo preverili naknadno.

Tabela 5: Kontingenčna tabela za spol in vremenske okoliščine (združene majhne skupine)

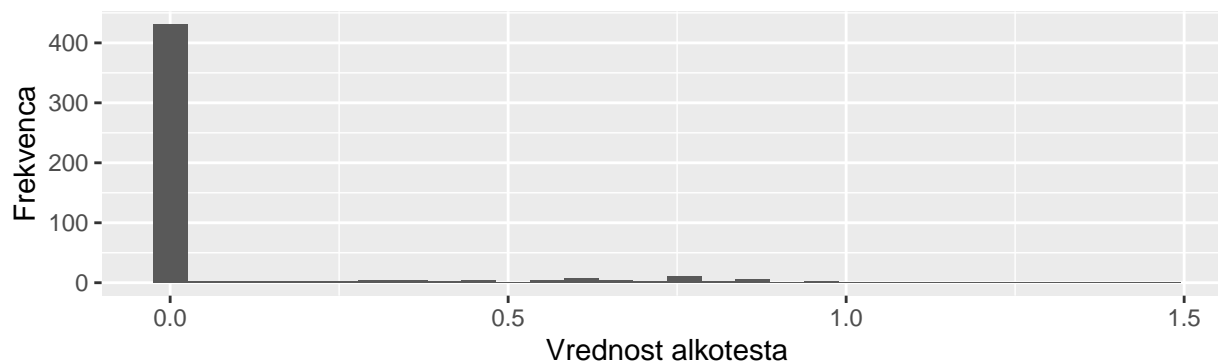
	MOŠKI	ŽENSKI	VSOTA
DEŽEVNO	36	14	50
JASNO	206	101	307
OBLAČNO	86	35	121
OSTALO	12	10	22
VSOTA	340	160	500



Slika 9: Relativen prikaz vremenskih razmer (združene majhne skupine) po spolu.

Vrednost Alkotesta

Vrednost alkotesta, če je ta bil izveden (če ni, je vrednost 0), bi lahko bil dober pokazatelj za razliko med moškimi in ženskami. Najprej si pogledjmo porazdelitev na sliki 10.



Slika 10: Porazdelitev vrednosti alkotesta.

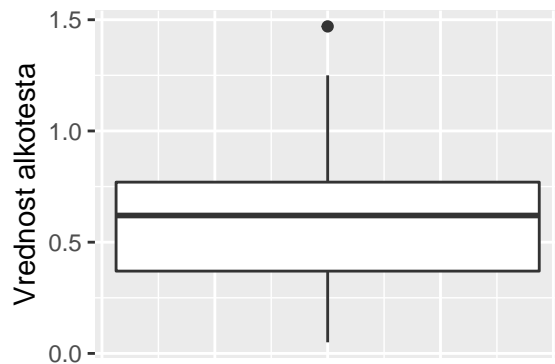
Vidimo da je večina voznikov ki so povzročili nesrečo imelo vrednost alkotesta enako 0. Pogledjmo si sliko 11, kjer je prikazan okvir z ročajo za tiste, ki so imeli vrednost alkotesta večjo od 0. Vidimo da je porazdelitev nekoliko asimetrična, vendar iz te slike ne moremo nič sklepati.

Zato je smiselno pogledati še opisne statistike v tabeli 6. Te so deljene tudi po spolu, saj nas zanima če so kakšne razlike med spoloma.

Iz teh opisnih statistik težko razberemo kakšno informacijo, zato si pogledjmo še opisne statistike za tiste povzročitelje nesreč ki so imeli vrednost alkotesta večjo od 0.

Večjih razlik v spolu ni videti, zato si pogledjmo še kontingenčno tabelo 8 za spol in ali je bil alkotest pozitiven ali ne.

Tu je videti, da se razmerje med moškimi in ženskami podre (drugod $\sim 2 : 1$). Pričakujem, da bomo s χ^2 testom tu zavrnili ničelno hipotezo. To bomo preverili v naslednjem poglavju. Najprej si pogledjmo če obstaja



Slika 11: Porazdelitev vrednosti alkotesta za tiste ki so imeli vrednost večjo od 0.

Tabela 6: Opisne statistike vrednosti alkotesta

	Skupno	Moški	Ženske
Min.	0.00	0.00	0.00
1st Qu.	0.00	0.00	0.00
Median	0.00	0.00	0.00
Mean	0.08	0.11	0.03
3rd Qu.	0.00	0.00	0.00
Max.	1.47	1.47	1.04
IQR	0.00	0.00	0.00

Tabela 7: Opisne statistike vrednosti alkotesta, za tiste ki so imeli alkotest večji od 0

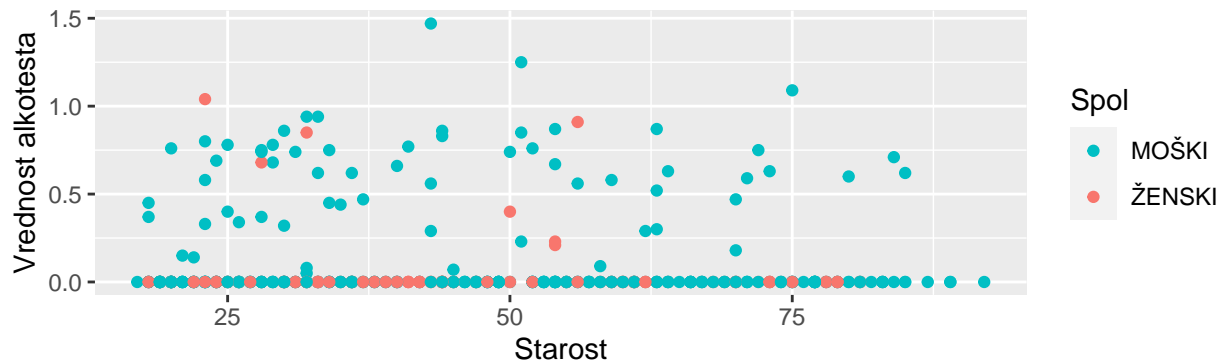
	Skupno	Moški	Ženske
Min.	0.0	0.0	0.2
1st Qu.	0.4	0.4	0.3
Median	0.6	0.6	0.7
Mean	0.6	0.6	0.6
3rd Qu.	0.8	0.8	0.9
Max.	1.5	1.5	1.0
IQR	0.4	0.4	0.6

Tabela 8: Kontingenčna tabela za spol in ali je nekdo imel alkotest večji od 0

	MOŠKI	ŽENSKI	VSOTA
< 0.0	278	153	431
> 0.0	62	7	69
VSOTA	340	160	500

še kaka odvisnost od starosti. Vidimo da večje odvisnosti ni. Podatkov za ženske, ki so povzročile nesrečo pod vplivom alkohola je le 7. Iz toliko zapisov težko kaj sklepamo.

Korelacijski koeficient le za tiste, ki so imeli vrednost alkotesta večjo od nič, je 0.048. Kar praktično pomeni da ni linearne povezanosti med starostjo in vrednostjo alkohola v krvi.



Slika 12: Razsevni diagram za vrednost alkotesta v odvisnosti od starosti.

Rezultati

Torej naš zastavljen problem je bil, da najdemo morebitne razlike med moškimi in ženskami ki so z osebnim avtomobilom povzročili prometno nesrečo. Analizirali smo spremenljivke **VNaselju**, **VremenskeOkoliscine**, **Starost**, in **VrednostAlkotesta**. Sklepati karkoli iz samih tabel in grafov bi bilo neodgovorno, zato bomo uporabili 2 statistična testa za primerjavo med moškimi in ženskami.

1.) Predpostavke testov

Pri primerjavi številskih spremenljivk bomo uporabili t -test za neodvisna vzorca, saj primerjamo dve različni skupini. En vzorec bo predstavljal moške, drugi ženske. Predpostavka t -testa je, da ali je vzorec normalno porazdeljen, ali pa da je vzorec dovolj velik ($n \geq 30$), saj zahteva da so povprečja vzorcev normalno porazdeljena (kar drži zaradi centralno limitnega izreka). Pri starosti ta predpostavka drži, saj je vzorec moških velikosti 340 in žensk 160. t -test primerja povprečja skupin, in za ničelno hipotezo trdi, da sta povprečji enaki.

Če bomo pa primerjali opisne spremenljivke, bomo uporabili χ^2 -test. Predpostavka testa je, da mora biti velikost (vsaj 80% polj) v kontingenčni tabeli, večja od 5, da je testna statistika porazdeljena po χ^2 porazdelitvi. Preverja če obstaja dovolj velika razlika med opazovanimi in pričakovanimi frekvencami v eni ali več skupinah. Ničelna hipoteza pravi, da ni dovolj velikih odstopanj, tj. da so frekvence opazovanih skupin enake frekvencam pričakovanih skupin.

2.) Primerjava starosti

Testna statistika t -testa za ta primer je bila 0.869, medtem ko je bila p vrednost enaka 0.385 in posledično ničelne hipoteze pri $\alpha = 0.05$, ne zavrnamo. Torej ne moremo trditi da se v povprečju starosti povzročiteljev nesreč v osebnih avtomobilih med moškimi in ženskami razlikujejo na našem vzorcu.

3.) Primerjava VNaselju

Pri preverjanju χ^2 -testa za razlike kje je bila storjena nesreča (v naselju ali izven), smo izračunali p vrednost, ki je bila 0.309 in posledično ničelne hipoteze pri $\alpha = 0.05$, ne zavrnamo. Se pravi ne moremo trditi da obstajajo dovolj velike razlike med opazovanimi in pričakovanimi frekvencami. Lahko bi tudi rekli, da ne moremo trditi da bodisi moški bodisi ženske povzročijo več ali manj nesreč v naselju, oz. izven naselja.

4.) Primerjava VremenskeOkoliscine

Kot smo videli pri preverjanju podatkov, za nekatere skupine v spremenljivki **VremenskeOkoliscine** imamo premalo opazovanj, te smo nato združili v skupino **OSTALO** in s tem pridobili dovolj velike skupine, da lahko preverimo če obstajajo razlike med opazovanimi in pričakovanimi frekvencami. Za to smo zopet uporabili χ^2 -test.

p vrednost χ^2 -testa je bila 0.417. Na podlagi p vrednosti pri stopnji tveganja $\alpha = 0.05$ ničelno hipotezo ne zavrnamo. Torej ne moremo trditi da obstajajo razlike med moškimi in ženskami v različnih vremenskih

okolščinah.

5.) Primerjava `VrednostAlkotesta`

Pri pregledovanju naših podatkov, smo videli, da pri preverjanju odvisnosti vrednosti alkotesta od starosti zaradi majhnega števila podatkov žensk, ne moremo uporabiti linearne regresije. Če pa vrednosti spremenimo v binarno spremenljivko, kjer so bile vrednosti ali enake 0 ali pa večje od 0 lahko uporabimo χ^2 -test, saj so še vedno izpolnjene predpostavke testa.

V našem primeru je bila vrednost p enaka 0.0001 in zato ničelno hipotezo zavrnamo pri stopnji tveganja $\alpha = 0.05$. Kot je bilo razvidno iz tabele so moški tisti, ki so večkrat povzročili nesrečo pod vplivom alkohola kot ženske.

Zaključek

Na podlagi našega vzorca kjer smo imeli 500 opazovanj, lahko trdimo da obstaja razlika med moškimi in ženskami, ki povzročijo nesrečo z osebnim avtomobilom. Ta razlika je v našem primeru bila le pri številu tistih, ki so nesrečo povzročili pod vplivom alkohola. Sumim, da bi se lahko tudi to spremenilo, če bi uporabili ali drug vzorec, ali večji vzorec. V ostalih pogledih ne moremo trditi da razlike obstajajo.