

SVEUČILIŠTE U ZAGREBU  
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5511

**Eksploratorna analiza  
telekomunikacijskih podataka u  
svrhu detekcije anomalija**

Tin Ivan Križ

Zagreb, lipanj 2018.

Zagreb, 14. ožujka 2018.

## **ZAVRŠNI ZADATAK br. 5511**

**Pristupnik:** Tin Ivan Križ (0036490962)  
**Studij:** Računarstvo  
**Modul:** Računarska znanost

**Zadatak:** Eksploratorna analiza telekomunikacijskih podataka u svrhu detekcije anomalija

### **Opis zadatka:**

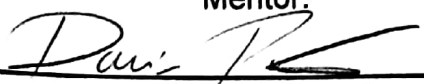
Pojam velikih podataka označava količinu podataka koja zbog svojeg obujma nije primjerena upravljanju uz pomoć tradicionalnih sustava kao što su relacijske baze podataka. Telekomunikacijske tvrtke primjer su tvrtki koje se tijekom svoj poslovanja susreću sa problemom velikih podataka: veliki broj korisnika na dnevnoj bazi koristi veliki broj telekomunikacijskih usluga o čemu se stvaraju i pohranjuju zapisi. Tipičan primjer ovakvih zapisa su popisi detalja o pozivima, čija detaljna analiza može otkriti niz korisnih informacija vezanih uz navike i ponašanje korisnika, a koje mogu koristiti za daljnje unaprjeđenje usluga, povećanje zadovoljstva korisnika i ostvarivanje većih profita. Dodatno, analizom ovakvih zapisa mogu se otkriti razne anomalije koje upućuje na moguće nepravilnosti i zlouporabe sustava.

Vaš zadatak je provesti eksploratornu analizu stvarnih telekomunikacijskih podataka s posebnim osvrtom na uočene anomalije. Analiza mora uključivati tehnologije vezne uz velike podatke, a može sadržavati i rezultate primjene deskriptivnih metoda strojnog učenja. Konačno rješenje mora biti oblikovano u programski skriptu za generiranje izvještaja koji će na adekvatan način prezentirati rezultate izvršene analize.

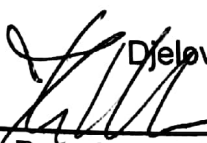
**Zadatak uručen pristupniku:** 16. ožujka 2018.

**Rok za predaju rada:** 15. lipnja 2018.

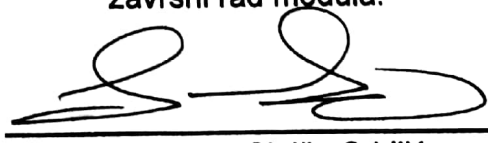
**Mentor:**

  
Doc. dr. sc. Damir Pintar

**Djelovođa:**

  
Doc. dr. sc. Tomislav Hrkać

**Predsjednik odbora za  
završni rad modula:**

  
Prof. dr. sc. Siniša Srbljić

*Zahvaljujem se mentoru doc. dr. sc. Damiru Pintaru i suradnicima iz firme Multicom d.o.o. na strpljenju i pomoći u napretku ovoga rada.*

# SADRŽAJ

<b>1. Uvod</b>	<b>1</b>
<b>2. Općenito o eksploratornoj analizi</b>	<b>2</b>
2.1. Prilagodba podataka . . . . .	2
2.2. Čišćenje podataka . . . . .	3
2.2.1. Kardinalnost varijable . . . . .	3
2.2.2. Uklanjanje nepostojećih vrijednosti . . . . .	4
2.2.3. Detekcija anomalija . . . . .	4
2.3. Istraživanje podataka . . . . .	5
2.4. Izvještavanje . . . . .	6
<b>3. Analiza stvarnih podataka telekomunikacijskog sustava</b>	<b>7</b>
3.1. Učitavanje i prilagodba podataka . . . . .	7
3.2. Čišćenje podataka . . . . .	9
3.2.1. Duljina poziva . . . . .	9
3.2.2. Cijena poziva . . . . .	10
3.2.3. Vrijeme uspostave poziva . . . . .	10
3.2.4. Odredište poziva . . . . .	11
3.3. Istraživanje podataka . . . . .	12
3.3.1. Broj poziva u ovisnosti o vremenu uspostave . . . . .	12
3.3.2. Duljina razgovora u ovisnosti o vremenu uspostave poziva . .	18
3.3.3. Zarada u ovisnosti o danu u mjesecu . . . . .	22
3.3.4. Osvrt na pozive prema inozemstvu . . . . .	24
3.4. Zaključak analize . . . . .	26
<b>4. Zaključak</b>	<b>28</b>
<b>Literatura</b>	<b>29</b>

<b>A. Isječci koda u jeziku R</b>	<b>30</b>
A.1. Učitavanje podataka . . . . .	30
A.2. Pretvaranje tipova . . . . .	30
A.3. Identifikacija nepostojećih vrijednosti . . . . .	30
A.4. Popunjavanje nepostojećeg izvorišnog operatera pretpostavljenim (VAS)	31
A.5. Provjera da su svi zapisi vezani uz pozive . . . . .	31
A.6. Provjera da su svi pozivi uspostavljeni u siječnju 2017. godine . . . .	31
A.7. Uklanjanje vrste usluge . . . . .	31
A.8. Univarijantna analiza duljine poziva . . . . .	32
A.9. Univarijantna analiza cijene poziva . . . . .	32
A.10. Univarijantna analiza vremena uspostave poziva . . . . .	32
A.11. Ostvarenje prikaza ovisnosti prometa o satu u danu . . . . .	32
A.12. Ostvarenje prikaza ovisnosti prometa o satu u danu i radnosti . . . .	32
A.13. Ostvarenje prikaza ovisnosti prometa o satu u danu kroz tjedan . . . .	33
A.14. Ostvarenje prikaza ovisnosti prometa o danu u mjesecu . . . . .	33
A.15. Ostvarenje prikaza ovisnosti prometa o danu u mjesecu i radnosti . . .	34
A.16. Ostvarenje prikaza ovisnosti prometa o danu u tjednu . . . . .	34
A.17. Ostvarenje prikaza ovisnosti prometa o danu u tjednu i operateru . . .	35
A.18. Ostvarenje prikaza ovisnosti duljine razgovora o satu u danu . . . . .	36
A.19. Ostvarenje prikaza ovisnosti duljine razgovora o satu u danu i danu u tjednu . . . . .	36
A.20. Ostvarenje prikaza ovisnosti duljine razgovora o danu u mjesecu . . .	37
A.21. Ostvarenje prikaza ovisnosti duljine razgovora o dana u mjesecu i rad- nosti dana . . . . .	37
A.22. Ostvarenje prikaza ovisnosti zarade i dana u mjesecu . . . . .	38
A.23. Istraživanje razdioba poziva prema inozemstvu kroz mjesec . . . . .	38
A.24. Uljepšan prikaz ovisnosti broja poziva o danu u tjednu . . . . .	39
A.25. Uljepšan prikaz ovisnosti duljine poziva o danu u tjednu . . . . .	39

# 1. Uvod

Znanost o podacima (engl. *data science*) je, uz strojno učenje (engl. *machine learning*), jedna od najbrže rastućih grana računarstva današnjice. Iako se područje počelo razvijati već sredinom dvadesetog stoljeća, njena je važnost došla do izražaja tek u moderno doba. Razlog leži u činjenici da se računalna snaga dovoljno razvila da može obrađivati ogromne količine podataka u relativno kratkom roku, a podataka za obrađivanje ima nezamislivo puno. Ti neobrađeni podaci skrivaju vrijedno znanje koje se može iskoristiti za razne optimizacije ili predikcije.

Prvi korak u analizi podataka se naziva eksploratorna analiza i upravo je ona tema ovog rada. U radu prolazimo kroz osnovnu teoretsku pozadinu eksploratorne analize, a onda ćemo je primijeniti na stvarnim telekomunikacijskim podacima u nadi da ćemo u naizgled neiskoristivim podacima pronaći zanimljivo i korisno znanje. Konkretna analiza je poduprijeta i isječcima koda u programskog jeziku R čime se potiče čitatelja da se i sam okuša u eksploratornoj analizi.

## 2. Općenito o eksploratornoj analizi

Eksploratorna analiza podataka (engl. *exploratory data analysis*) je prvi korak u analizi podataka kojemu je cilj upoznati se s dobivenim podacima, te iz njih izvući korisno i zanimljivo znanje. Općenito, podaci koje obrađujemo su ništa drugo nego velika tablica vrijednosti. Svaki red u podatkovnom skupu nazivamo opažanjem (engl. *observation*), a svaki stupac značajkom ili varijablom (engl. *variable*). Dobiveno znanje najčešće predstavljaju uočene relacije među značajkama opažanja u podacima koje su uglavnom podržane grafičkim prikazima.

Kada bismo za uočene relacije dodatno razvili matematičke modele koji ih najbolje opisuju, onda bismo pričali o prediktivnoj analizi (engl. *predictive data analysis*) koja se više usredotočila na predviđanje vrijednosti novih podataka iz postojećih.

Bitno je napomenuti da eksploratorna analiza nije strogo definirani koncept. Drugim riječima, ne postoji algoritam za eksploratornu analizu. Svaka analiza je drugačija i najviše ovisi o dobivenim podacima i željenim rezultatima. Unatoč tome, svaka eksploratorna analiza prolazi kroz sljedeće tri faze:

- prilagodba podataka (engl. *data wrangling*)
- čišćenje podataka (engl. *data cleaning*)
- istraživanje podataka (engl. *data exploration*)

Ove faze se najčešće isprepliću i iterativno ponavljaju kako se naše postojeće znanje i apetit za novim znanjem o podacima povećava.

### 2.1. Prilagodba podataka

Podaci koji dolaze analitičaru, skupljeni iz stvarnog svijeta kroz razna istraživanja ili mjerenja, se nazivaju sirovi podaci (engl. *raw data*). Prilagodba podataka je proces obrađivanja sirovih podataka s ciljem pripreme za daljnju analizu.

Jednom kada podatke učitamo u odgovarajuću računalnu strukturu<sup>1</sup>, potrebno je

---

<sup>1</sup>Npr. okvir podataka (engl. *data frame*) u programskom jeziku R

uvjeriti se da je svaka varijabla odgovarajućeg tipa i da se prazna polja u cijelom podatkovnom skupu označavaju na jednak način. Ovdje je također moguće stvarati nove stupce iz postojećih ili sažeti više stupaca u jedan zbog lakše analize (npr. rastaviti stupac “ime i prezime” na dva stupca: “ime” i “prezime” ili obrnuto), te mijenjati vrijednosti pojedinih stupaca kako bi bile razumljivije ili pogodnije za analizu (npr. pretvoriti stupac dužine iz inča u centimetre).

## 2.2. Čišćenje podataka

Jednom kada smo stvorili semantički ispravne podatke, želimo se posvetiti kvaliteti dobivenih podataka i njenom poboljšanju. Samo iz kvalitetnih podataka možemo izvući kvalitetne zaključke. Ispitivanje kvalitete podataka se svodi na univarijantnu analizu (engl. *univariate analysis*) zasebnih varijabli. Drugim riječima, za svaku varijablu se izolirano promatra njena razdioba, iznosi kvartila<sup>2</sup> i srednje vrijednosti. Ovdje je moguće uočiti i ukloniti više različitih problema:

- kardinalnost varijable (engl. *cardinality*)
- uklanjanje nepostojećih vrijednosti (engl. *handling missing values*)
- detekcija anomalija (engl. *anomaly detection*)

Rezultat čišćenja podataka je kvalitetan i istražen skup izoliranih varijabli, spreman za proučavanje njihovih međusobnih relacija i interakcija.

### 2.2.1. Kardinalnost varijable

Kardinalnost varijable je općenito broj različitih vrijednosti varijable koji se pojavljuje u podacima. Problem nastaje kad kardinalnost pojedine varijable nije u skladu s našim očekivanjima i znanju o domeni podataka.

Prvi mogući problem je ako je kardinalnost jednaka 1. To bi značilo da varijabla poprima istu vrijednost za svako opažanje. Kao takva nam ne donosi nikakvu novu informaciju i možemo je ukloniti iz podatkovnog skupa.

Drugi problem se tiče kontinuiranih varijabli<sup>3</sup> Očekivano je da će kardinalnost kontinuirane varijable biti približno jednaka broju opažanja u podacima. Ako se radi o očigledno manjoj kardinalnosti, onda možemo razmišljati o kategorizaciji varijable. Npr.

---

<sup>2</sup>N-ti kvartil varijable predstavlja broj ispod kojeg se nalazi N četvrtina vrijednosti varijable

<sup>3</sup>Varijable koje mogu poprimiti neograničen broj različitih vrijednosti zovemo kontinuiranim varijablama. Nasuprot njima, varijable koje mogu poprimiti konačan broj različitih vrijednosti zovemo diskretnim ili kategorijskim varijablama.



ako nam varijabla označava broj vrata na autu i svi auti imaju ili 3 ili 5 vrata, onda broj vrata ne možemo promatrati kao kontinuiranu varijablu, već bi je bilo pametnije promatrati kao kategorijsku.

Zadnji mogući problem nastaje kada je kardinalnost puno veća od očekivane. Npr. kada za spol osobe dobijemo kardinalnost veću od 2. Takav problem može nastati kada se za istu vrijednost varijable koriste različite oznake ("M", "muško", "Muški" ..).

### **2.2.2. Uklanjanje nepostojećih vrijednosti**

Dok smo u prilagodbi podataka više bili usredotočeni na jednoznačno prepoznavanje nepostojećih vrijednosti, u ovoj fazi čišćenja podataka pokušavamo utvrditi koliko nam količina uočenih nepostojećih vrijednosti utječe na kvalitetu podataka. Utjecaj nepostojećih vrijednosti najbolje možemo razumjeti ako znamo koji je razlog njihovog nastanka.

Ako su nepostojeće vrijednosti nastale kao pogreške u mjerenju ili istraživanju, onda moramo ozbiljno razmisliti o uklanjanju varijable s velikim postotkom nepostojećih vrijednosti. Iako je nemoguće odrediti čvrstu granicu za visoki udio, možemo reći da ako varijabla ima udio nepostojećih vrijednosti veći od 60% onda je količina informacije dobivena od nje toliko malena da ju je najbolje ukloniti iz podatkovnog skupa [2].

S druge strane, nepostojeće vrijednosti mogu biti namjerno integrirane u podatke. One mogu predstavljati pretpostavljane vrijednosti (npr. 0) ili mogu biti izostavljene zbog raznih ograničenja domene kao što su zaštita osobnih podataka. Ovdje nam nepostojeća vrijednost donosi informaciju pa varijable koje imaju visoki udio ovakvih nepostojećih vrijednosti najčešće nemamo potrebu ukloniti iz podatkovnog skupa. Njih možemo prilagoditi unosom pretpostavljenih vrijednosti varijabli ili kategorizirati na to postoji li upisana vrijednost ili ne.

### **2.2.3. Detekcija anomalija**

Općenito govoreći, anomalije su vrijednosti koje očigledno odstupaju od očekivane razdiobe pojedine varijable. Iako nam anomalije mogu biti vrlo informativne za razumijevanje podataka, one nam najčešće predstavljaju problem pri izradi modela i testiranju relacija. Postavlja se pitanje: kako prepoznati anomalije?

Razlikujemo opravdane i neopravdane anomalije. Neopravdane anomalije su pogrešno upisani podaci koji ne predstavljaju opažanja iz stvarnog svijeta. One mogu nastati kroz kvarove mjernih instrumenata ili pogrešnim prepisivanjem podataka. Npr.

ako meteorološka stanica očita vrijednost temperature 2000 °C umjesto 20 °C ili ako se visina osobe u registar slučajno upiše kao 1.75 cm umjesto 175 cm.

S druge strane, opravdane anomalije su stvarna opažanja koja su nastala u ekstremnim ili neočekivanim uvjetima. Npr. ako promatramo duljinu razgovora na određeni dan, ne želimo uzeti u obzir razgovor koji je trajao 12 sati jer to nije realno očekivano trajanje razgovora i vjerojatno je nastalo kao slučajan splet okolnosti potpuno neovisno o promatranoj relaciji.

Jako je teško fiksirati iznos nakon kojeg neko opažanje možemo promatrati kao anomaliju. Je li razgovor od 5 sati anomalija? A razgovor od 2 sata? Ako kažemo da je razgovor od 2 sata anomalija, znači li to da razgovor od 1 sat i 59 minuta nije anomalija? Postoji iscrpno područje koje se bavi isključivo detekcijom anomalija i njenim istraživanjem.

U sklopu statistike, anomalijom se uglavnom smatraju vrijednosti koje se nalaze na više od 1.5 interkvartilnog raspona (engl. *interquartile range*)<sup>4</sup> ispod prvog kvartila i iznad trećeg kvartila. U praksi se ipak odlučujemo na malo slobodniji odabir granica prihvatljivosti iznosa varijabli na temelju općeg znanja o domeni podataka.

## 2.3. Istraživanje podataka

Kada smo napokon došli do kvalitetnih podataka koje smo dobro upoznali kroz univarijantnu analizu, sva mašta i znatiželja analitičara dolaze do izražaja u sljedećoj fazi eksploratorne analize – istraživanju podataka.

Cilj istraživanja je pronaći zanimljive relacije i kvalitetno ih vizualizirati. Da bismo došli do cilja, potrebno je dugo kopati i isprobavati različite i sve finije relacije<sup>5</sup> među varijablama. Kako istraživane relacije postaju finije, tako često imamo potrebu vraćati se na prijašnje faze prilagodbe i čišćenja podataka.

Spomenute relacije općenito možemo zamisliti kao odgovore na neka pitanja. Npr. Ovisi li iznos varijable A o iznosu varijable B? Može li povećanje varijable A utjecati na promjenu varijable B? Odgovornost je analitičara da iskoristi svoju znatiželju kako bi iznova smišljao pitanja i odgovarao na njih.

Odgovore na pitanja najčešće pronalazimo kroz razne vizualizacije. Jednom kada se skupi zadovoljavajući broj argumenata u prilog neke zanimljive ili korisne relacije, rezultate je potrebno interpretirati i pripremiti podatke za matematičko modeliranje relacije (prediktivnu analizu). Pri interpretaciji je bitno imati na umu pristranost nas

---

<sup>4</sup>Razlika vrijednosti trećeg i prvog kvartila

<sup>5</sup>Preciznije ili konkretnije relacije

kao analitičara i kako je način na koji su početni podaci nastali mogao utjecati na rezultate.

## 2.4. Izvještavanje

U analizi podataka nam najčešće nije bitno samo doći do korisnih i zanimljivih znanja o podacima, već nam je to isto znanje bitno na što bolji i intuitivniji način prenijeti drugim ljudima. Tada pričamo o izvještavanju (engl. *reporting*).

U eksploratornoj analizi smo tražili vrijedno znanje o podacima kroz odgovore na pitanja. Nađeno znanje je samo vrh sante leda naspram silnih postavljenih pitanja koja smo se morali pitati na putu do nađenog znanja. Potrebno je izdvojiti što je bitno čitateljima rezultata analize, a što je bitno analitičaru. Jednom kada se odabere što se želi prenijeti čitateljima, potrebno je pripremiti grafove, te slijedno i intuitivno opisati rezultate analize kako bi ih čitatelji što lakše razumjeli.

Iako ćemo imati na umu da rad bude što čitljiviji, u nastavku rada se nećemo toliko baviti izvještavanjem rezultata eksploratorne analize koliko ćemo se baviti samim postupkom.

## 3. Analiza stvarnih podataka telekomunikacijskog sustava

Analizu podataka moguće je provesti raznim programskim jezicima kao što su Python, R, SAS, Julia, itd. U ovom ćemo poglavlju demonstrirati i komentirati proces eksploratorne analize podataka kroz programski jezik R. Podatke je ustupila tvrtka Multicom d.o.o. Ustupljeni podaci sadrže točno milijun fiksnih poziva hrvatskih klijenata iz perspektive teleoperatera VAS raspoređenih kroz siječanj 2017. godine.

U skladu s aktualnim zakonom o osobnim podacima koji je 25. svibnja 2018. stupio na snagu u svim zemljama Europske unije, sve varijable u dobivenim podacima koje bi se mogle povezati sa stvarnim osobama su maskirane (identifikatori, brojevi, paketi, tarife, itd.).

Daljnja analiza bit će načelno podijeljena u spomenute tri faze (prilagodba, čišćenje i istraživanje podataka). Cilj opisane analize nije samo otkriti zanimljive relacije u podacima, već i približiti općenito proces eksploratorne analize stvarnih podataka čitatelju. Za većinu opisanih postupaka će biti priložen programski kod u jeziku R u dodatku A kako bi se čitatelj i sam mogao okušati u eksploratornoj analizi.

### 3.1. Učitavanje i prilagodba podataka

Dobiveni sirovi podaci bili su spremljeni u obliku 168 MB<sup>1</sup> velike DSV<sup>2</sup> datoteke. Srećom, radilo se o vrijednostima odvojenim znakom ‘;’ (točka i zarez) kojeg programski jezik R zna izravno učitati (dodatak A.1). Opis dobivenog podatkovnog skupa možemo vidjeti u tablici 3.1.

Učitani podaci su puni praznih polja i krivo postavljenih tipova varijabli pa je potrebno svaki stupac pretvoriti u odgovarajući tip (dodatak A.2) i osigurati da se nepos-

---

<sup>1</sup>Megabajt, 1024 kilobajta

<sup>2</sup>Delimiter Separated Values

**Tablica 3.1:** Opis dobivenog podatkovnog skupa

Varijabla	Opis
N	Redni broj poziva
TRANS_ID	Identifikator transakcije
BUYER_ID_MASK <sup>a</sup>	Maskirani identifikator kupca
DEBTOR_ID_MASK	Maskirani identifikator osobe koja plaća
TRANS_DATE	Vrijeme poziva
OFFER_CODE_MASK	Maskirani paket
SUBSCRIPTION_CODE_MASK	Maskirani identifikator paketa
CHARGE_CODE_MASK	Maskirani identifikator tarife
AMOUNT	Cijena poziva
ITEM_CALL_DIRECTION	Odlazni(1) ili dolazni(0) poziv?
ITEM_CALLER_NUMBER_MASK	Maskirani broj pozivatelja
ITEM_CALLEE_NUMBER_MASK	Maskirani pozvani broj
ITEM_CALL_DURATION	Duljina razgovora
ITEM_ORIGINATING_CARRIER_ID	Izvorišni operater
ITEM_TERMINATING_CARRIER_ID	Ciljni operater
COMPUTED_ZONE_MASK	Maskirana zona prema kojoj je upućen poziv (lokalna, fiksna, mobilna..)
COMPUTED_DESTINATION	Geografska ili mobilna mreža odredišta
COMPUTED_PRICE_PER_MINUTE	Cijena po minuti
COMPUTED_ON_PEAK_PRICE_PER_M	Cijena unutar veće tarife (ako je promet pojačan)
COMPUTED_OFF_PEAK_PRICE_PER_M	Cijena unutar niže tarife (ako je promet slab)
COMPUTED_ESTABLISHMENT_FEE	Cijena uspostave poziva
PARAMETER_PACKET_MASK	Maskirani naziv paketa
PARAMETER_SERVICE_GROUP	Vrsta usluge

<sup>a</sup>Maskirani stupci ne predstavljaju stvarne podatke zbog zaštite osobnih podataka klijenata

tojeće vrijednosti označavaju na jednak način u cijelom podatkovnom skupu<sup>3</sup> (dodatak A.3). Također, možemo primijetiti uzorak u podacima pri kojem se na svim odlaznim pozivima ne zapisuje izvorišni operater, te se samo za odlazne pozive zna računati cijena uspostave i odredište poziva (geografska ili mobilna mreža). U razgovoru s tvrtkom koja nam je ustupila podatke, došli smo do zaključka da se na svim odlaznim pozivima pretpostavlja teleoperater VAS kao izvorišni teleoperater, te da se samo za odlazišne pozive može računati odredišna mreža (dodatak A.4).

Rezultat ove faze je tablica podataka koja jednoznačno opisuje nepostojeće vrijednosti i čijim varijablama možemo lako manipulirati.

## 3.2. Čišćenje podataka

Cilj sljedeće faze je provjeriti kvalitetu semantički ispravnih podataka. Najprije smo se uvjerali da se zaista radi o zapisima koji su vezani uz pozive i to u siječnju 2017. godine. Pošto su svi zapisi vezani uz pozive, kardinalnost varijable “Parameter service group” je jednaka 1 i zbog toga je možemo ukloniti iz podataka. Programski kod za navedeni postupak je dostupan u dodacima A.5 - A.7.

Iako je sada sljedeći logičan korak provoditi univarijantnu analizu pojedinih varijabli, možemo uočiti da je nema potrebe provoditi za baš svaku varijablu. Naime, početni podatkovni skup sadrži 22 varijable (stupca) koje opisuju svaki poziv (tablica 3.1). U narednoj analizi ćemo se usredotočiti na samo one varijable koje nisu prethodno maskirane, ali me svejedno mogu odvesti do zanimljivih zaključaka. Takve varijable nazivamo bitnim varijablama (engl. *important variables*). Konkretno, usredotočit ćemo se na sljedeće bitne varijable:

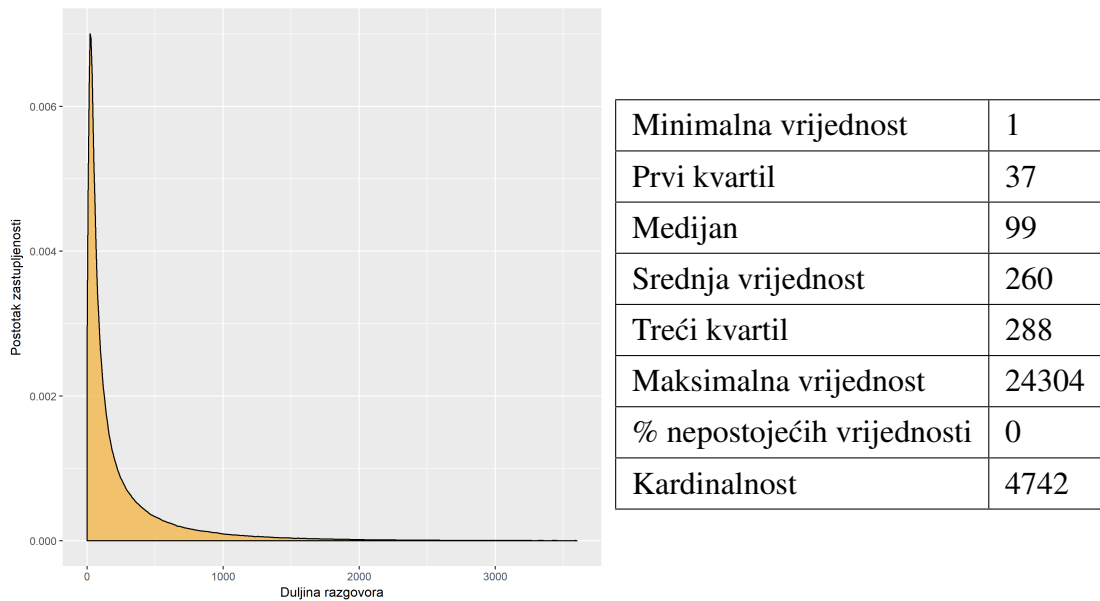
- duljina poziva
- cijena poziva
- vrijeme uspostave poziva
- odredište poziva

### 3.2.1. Duljina poziva

Varijabla duljine poziva ne sadrži nepostojeće vrijednosti i velike je kardinalnosti što znači da je dobar kandidat za istraživanje. Prije nego krenemo istraživati interakciju

---

<sup>3</sup>Svaka nepostojeća vrijednost zamijenjena je s vrijednosti NA koja u programskom jeziku R upravo za to smišljena (engl. *Not Available*)



**Slika 3.1:** Podaci o razdiobi duljine razgovora izražene u sekundama (dodatak A.8)

duljine razgovora s ostalim značajkama poziva, želimo ukloniti potencijalne anomalije u podacima.

Kada bismo se odlučili za statističku granicu anomalija<sup>4</sup>, dobili bismo da su anomalije svi pozivi dulji od 11 minuta. Ova granica nam je preniska pošto želimo u analizi uključiti i pozive ljudi koji se dugo nisu čuli ili vidjeli. Odokativno se odlučujemo za granicu od 1 sata. Time smo iz podataka isključili 1811 poziva.

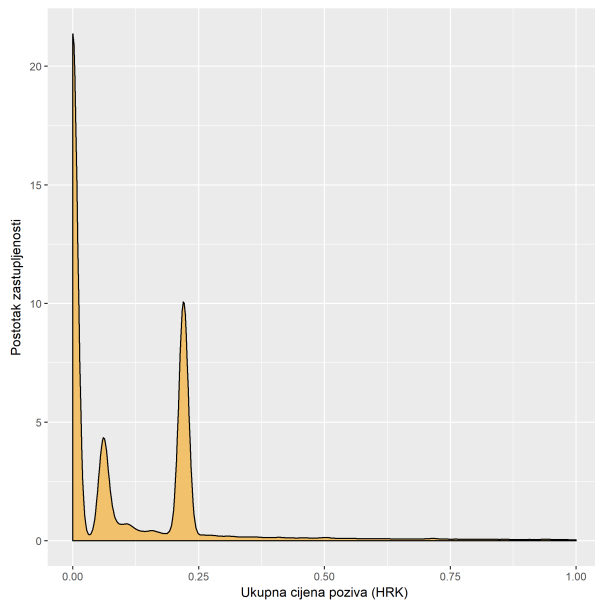
### 3.2.2. Cijena poziva

Varijabla cijene poziva, slično kao duljina poziva, nema nepostojeće vrijednosti i velike je kardinalnosti. Ovdje nemamo potrebu izbacivati anomalije s obzirom na to da je najskuplji poziv naplaćen 550 kuna. Takva cijena poziva ne zvuči pretjerano jer troškovi mogu nastati iz raznih razloga pa odlučujemo da i takve skuplje pozive želimo uključiti u analizu.

### 3.2.3. Vrijeme uspostave poziva

Vrijeme uspostave je također definirano za cijeli podatkovni skup i ima visoku kardinalnost. Kvartilne vrijednosti se približno podudaraju s tjednima u mjesecu tako da pretpostavljamo da su podaci raspoređeni po cijelom siječnju.

<sup>4</sup>1.5 \* vrijednost interkvartilnog raspona



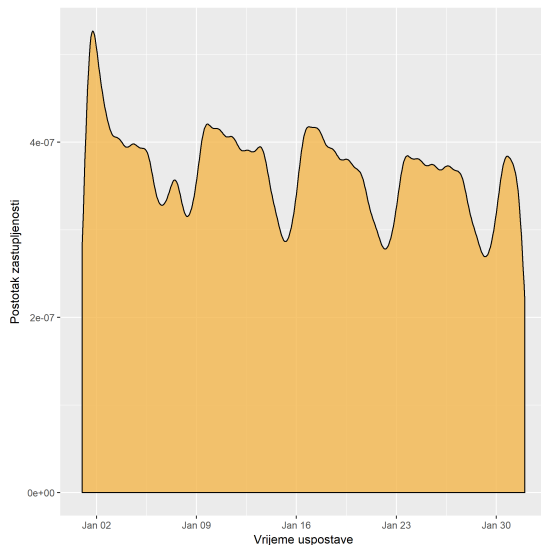
Minimalna vrijednost	0.00
Prvi kvartil	0.00
Medijan	0.06
Srednja vrijednost	0.3759
Treći kvartil	0.22
Maksimalna vrijednost	558.61
% nepostojećih vrijednosti	0
Kardinalnost	19533

**Slika 3.2:** Podaci o razdiobi cijene razgovora izražene u kunama (dodatak A.9)

### 3.2.4. Odredište poziva

Odredište se računa samo za odlazne pozive i zato ima visoki postotak nepostojećih vrijednosti (41%). Ovu značajku poziva ćemo pokušati izbjeći koristiti ili ćemo je koristiti uz veliku količinu sumnje u rezultate zato što trenutno ne možemo sa sigurnošću reći da odlaznost poziva ne utječe na odredišnu mrežu poziva.





Minimalna vrijednost	2017-01-01 00:00:04
Prvi kvartil	2017-01-07 19:42:25
Medijan	2017-01-15 18:47:13
Srednja vrijednost	2017-01-15 21:50:08
Treći kvartil	2017-01-23 17:26:29
Maksimalna vrijednost	2017-01-31 23:58:53
% nepostojećih vrijednosti	0
Kardinalnost	721557

**Slika 3.3:** Podaci o razdiobi vremena uspostave poziva oblika: "godina-mjesec-dan sat:minuta:sekunda" (dodatak A.10)

### 3.3. Istraživanje podataka

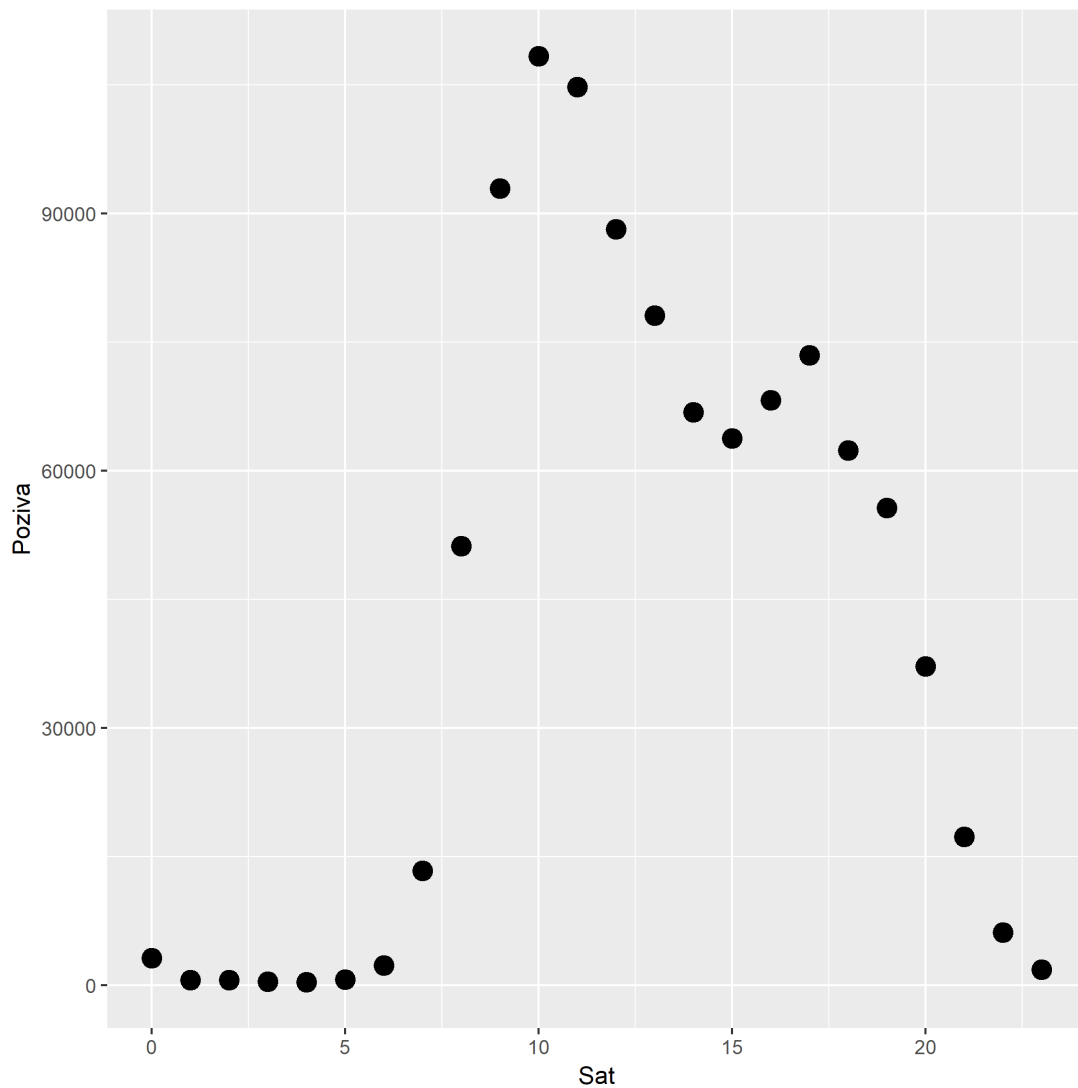
Svjesni kvalitete podataka koji su nam na raspolaganju, spremni smo pustiti mašti na volju i istraživati sva pitanja koja nam padnu na pamet. Ova iscrpna cjelina istraživanja podataka podijeljena je na poglavlja s obzirom na promatrane varijable.

#### 3.3.1. Broj poziva u ovisnosti o vremenu uspostave

Vrijeme uspostave poziva u svojem punom obliku nije nam praktično za analizu pa odlučujemo provjeriti kako ovisi prosječan broj poziva o karakteristikama vremena uspostave kao što su dan u mjesecu, dan u tjednu, sat u danu, je li dan radan ili neradan, radi li se možda o blagdanu ili prazniku...

##### Broj poziva u ovisnosti o satu u danu

Prvi od niza pitanja glasi: ovisi li broj poziva o satu u danu? Rezultantni graf (slika 3.4) prikazuje krivulju s dva lokalna maksimuma – jednim u 10 sati i jednim u 17 sati gdje je onaj u 10 sati ujedno i globalni maksimum. Ova dva broja dosta koreliraju s radnim navikama ljudi pa nas to potiče na sljedeće pitanje: razlikuje li se prosječan broj poziva kroz dan za radne i neradne dane?

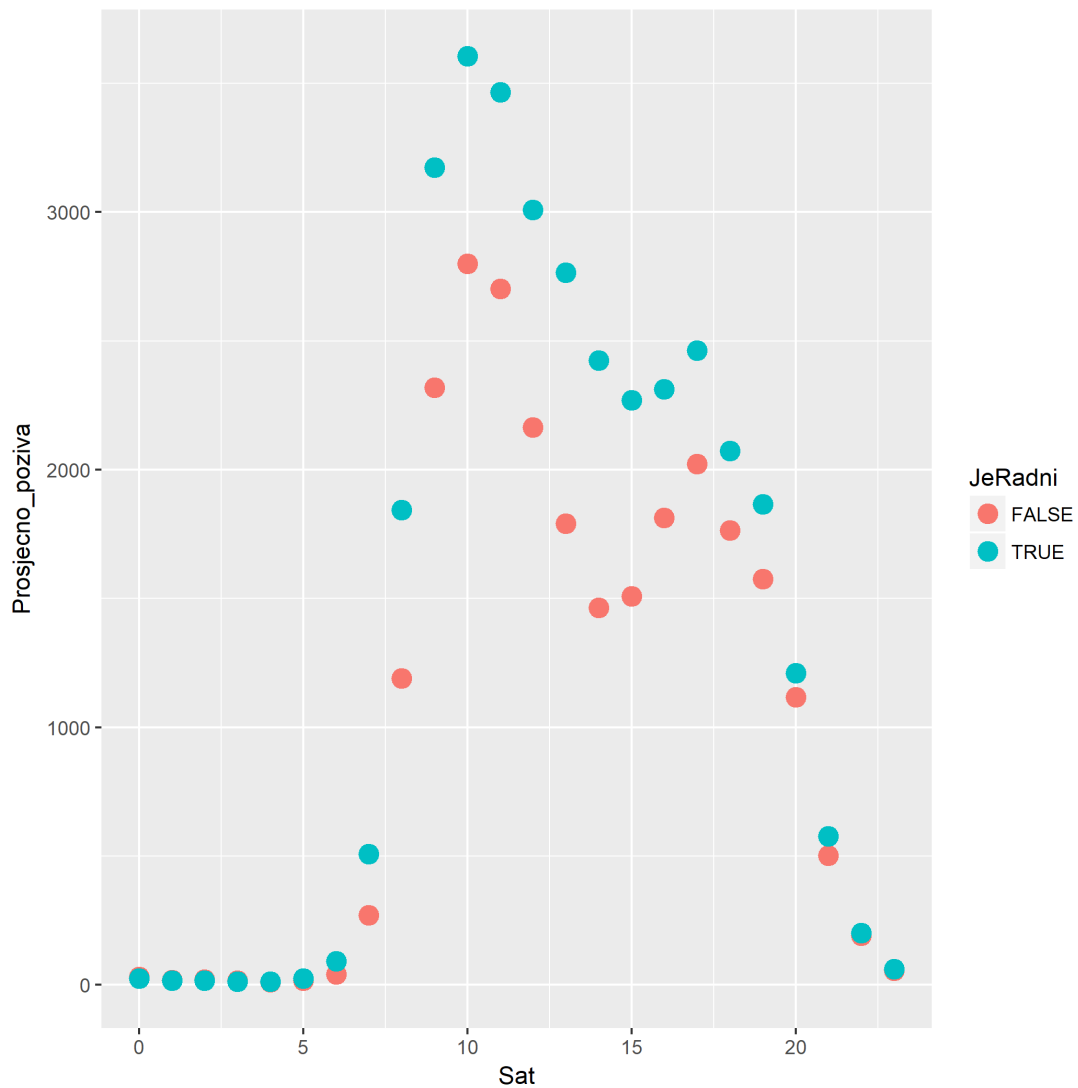


**Slika 3.4:** Prikaz ovisnosti broja poziva o satu u danu (dodatak A.11)

### Broj poziva u ovisnosti o satu u danu i radnosti

Kada bismo za radne i neradne dane dobili različite oblike krivulja, mogli bismo zaključiti da radnost dana utječe na razdiobu poziva kroz dan.

Međutim, resultantni graf (slika 3.5) upućuje na to da su razdiobe jednake za radne i neradne dane. Ovdje naslućujemo da su radni dani općenito prometniji od neradnih, ali im je krivulja kroz sate u danu jednaka. Kako bismo se dodatno uvjerali, provjeravamo relaciju i na razini dana u tjednu.



**Slika 3.5:** Prikaz ovisnosti broja poziva o satu u danu i radnosti dana (dodatak A.12)

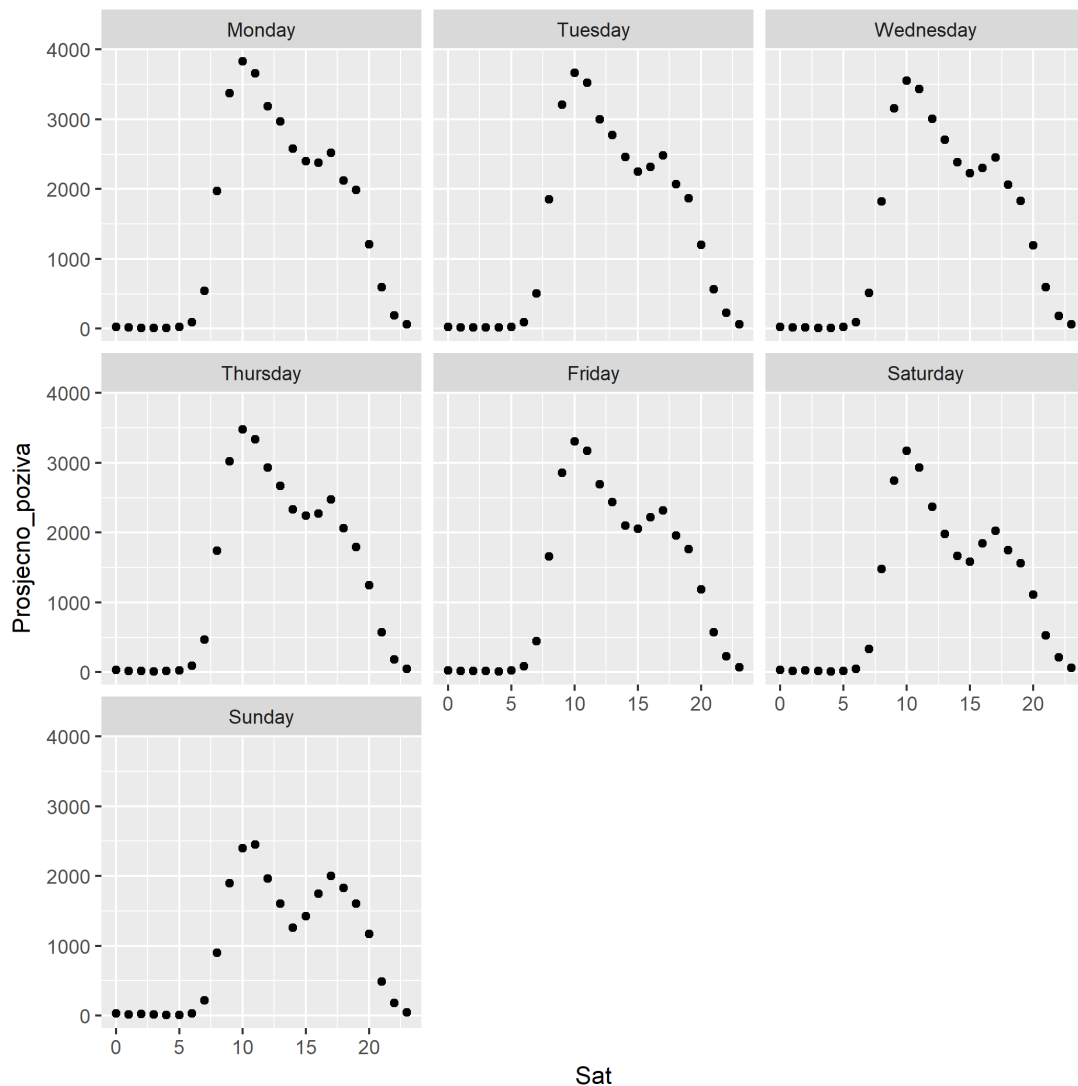
### Broj poziva u ovisnosti o satu u danu i danu u tjednu

Dobivši sličnu razdiobu poziva kroz sate u danu za svaki dan u tjednu (slika 3.6), možemo zaključiti da činjenica da ljudi upućuju najviše poziva u danu u 10 sati ujutro ovisi o nekoj drugoj ljudskoj sklonosti koju iz podataka ne možemo očitati.

### Broj poziva u ovisnosti o danu u mjesecu

Iako nismo uspjeli dokazati da radnost dana utječe na razdiobu poziva kroz dan, graf na slici 3.5 nam daje naslutiti da su radni dani općenito prometniji od neradnih. Za početak se odlučujemo na malo općenitije pitanje: ovisi li broj poziva o danu u mjesecu?

Ako uzmemo u obzir da se radi samo o siječnju, ne možemo napraviti generaliza-

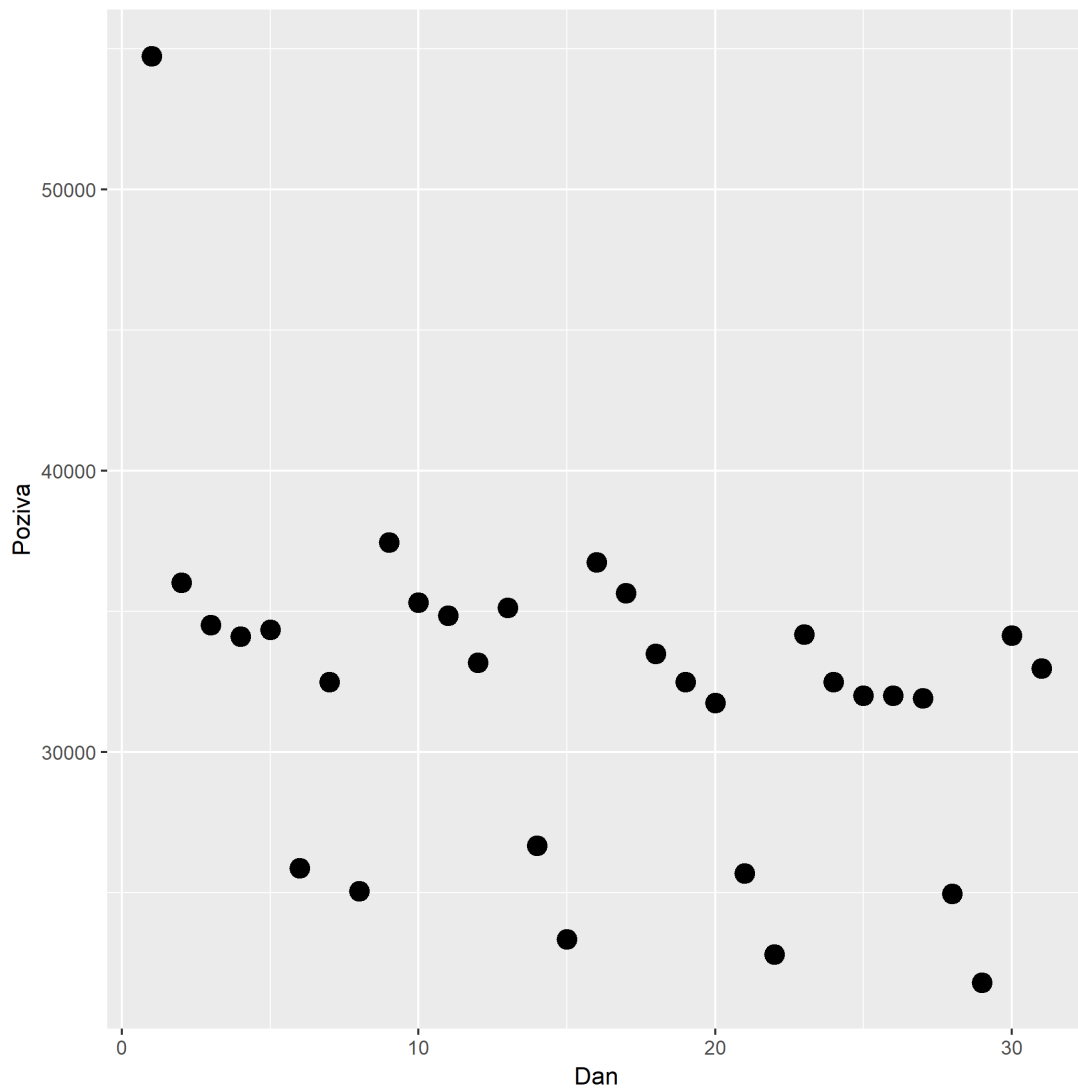


**Slika 3.6:** Usporedba razdiobe prosječnog broja poziva kroz dan za svaki pojedini dan u tjednu (dodatak A.13)

ciju na sve mjesece u godini, ali nam ovakav početak rješavanja pitanja otvara nove probleme (slika 3.7).

Za početak, uočavamo da praznik Nova godina (1. siječnja) ima daleko najviše poziva u mjesecu. To je i očekivano ponašanje, ali nam kao takvo samo smeta za ispitivanje daljnjih relacija i odlučujemo je ukloniti. Slično Novoj godini, blagdani Sveta tri kralja (6. siječnja) i pravoslavni Božić (7. siječnja) se također uklanjaju iz ispitivanja relacija jer njihove karakteristike očigledno odudaraju od svakodnevnih.

Također, možemo uočiti uzorak koji se čini da bi mogao predstavljati dane u tjednu i to nas tjera na postavljanje sljedećeg pitanja: ovisi li broj poziva kroz mjesec o radnosti dana?

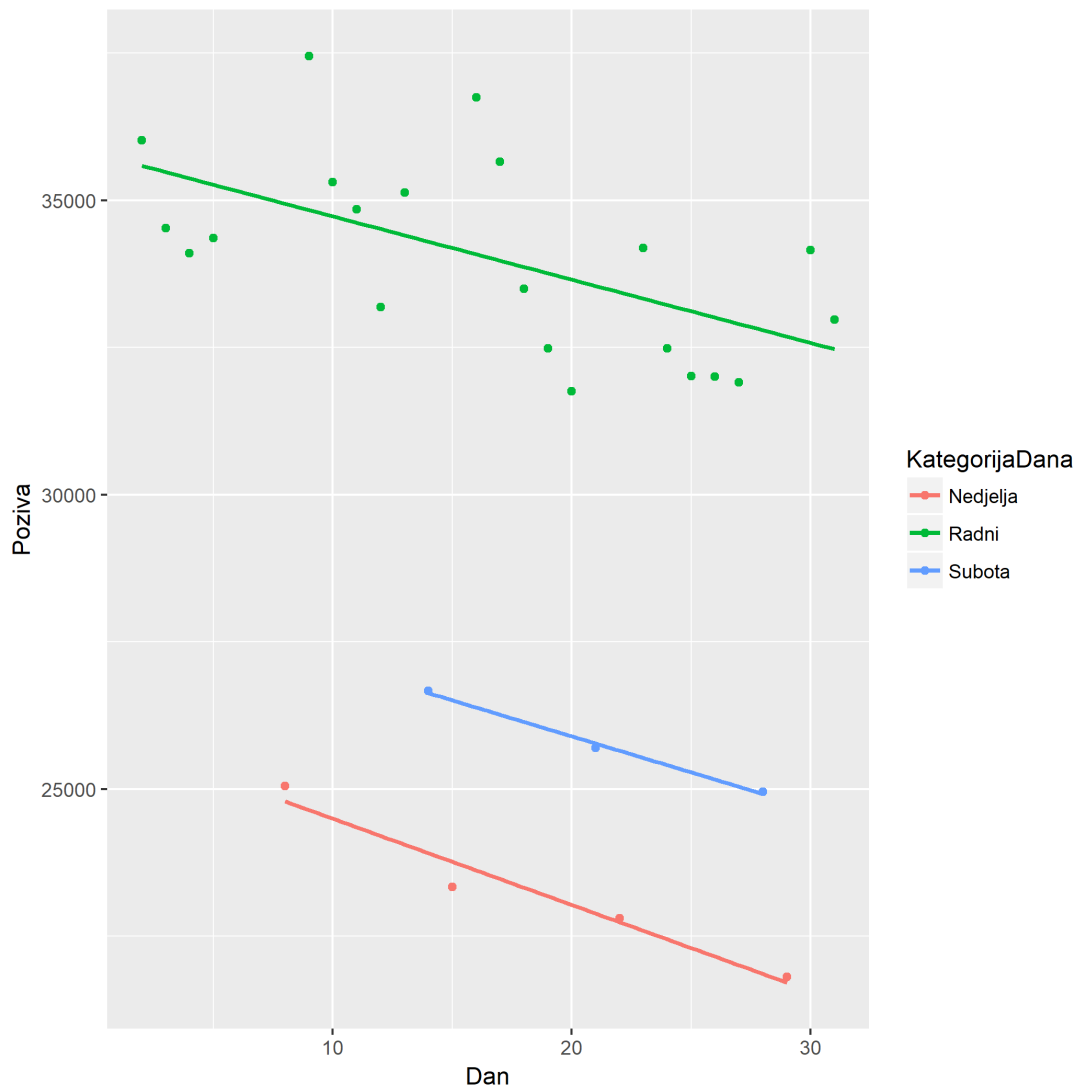


**Slika 3.7:** Prikaz ovisnosti broja poziva o danu u mjesecu(dodatak A.14)

### Broj poziva u ovisnosti o radnosti

Subota je u puno većoj mjeri radni dan nego li je to nedjelja i zato se odlučujemo podijeliti tjedan na radne dane, subotu i nedjelju.

Dobiveni graf (slika 3.8) zaista daje naslutiti da broj poziva u danu ovisi o njegovoj radnosti. Također, možemo primijetiti generalno smanjenje broja poziva kroz mjesec što nam na razini jednog mjeseca ne znači puno. Sada smo spremni dodatno potvrditi svoje slutnje promatrajući razdiobu poziva kroz tjedan.



**Slika 3.8:** Prikaz ovisnosti broja poziva o danu u mjesecu i njegovoj radnosti (dodatak A.15)

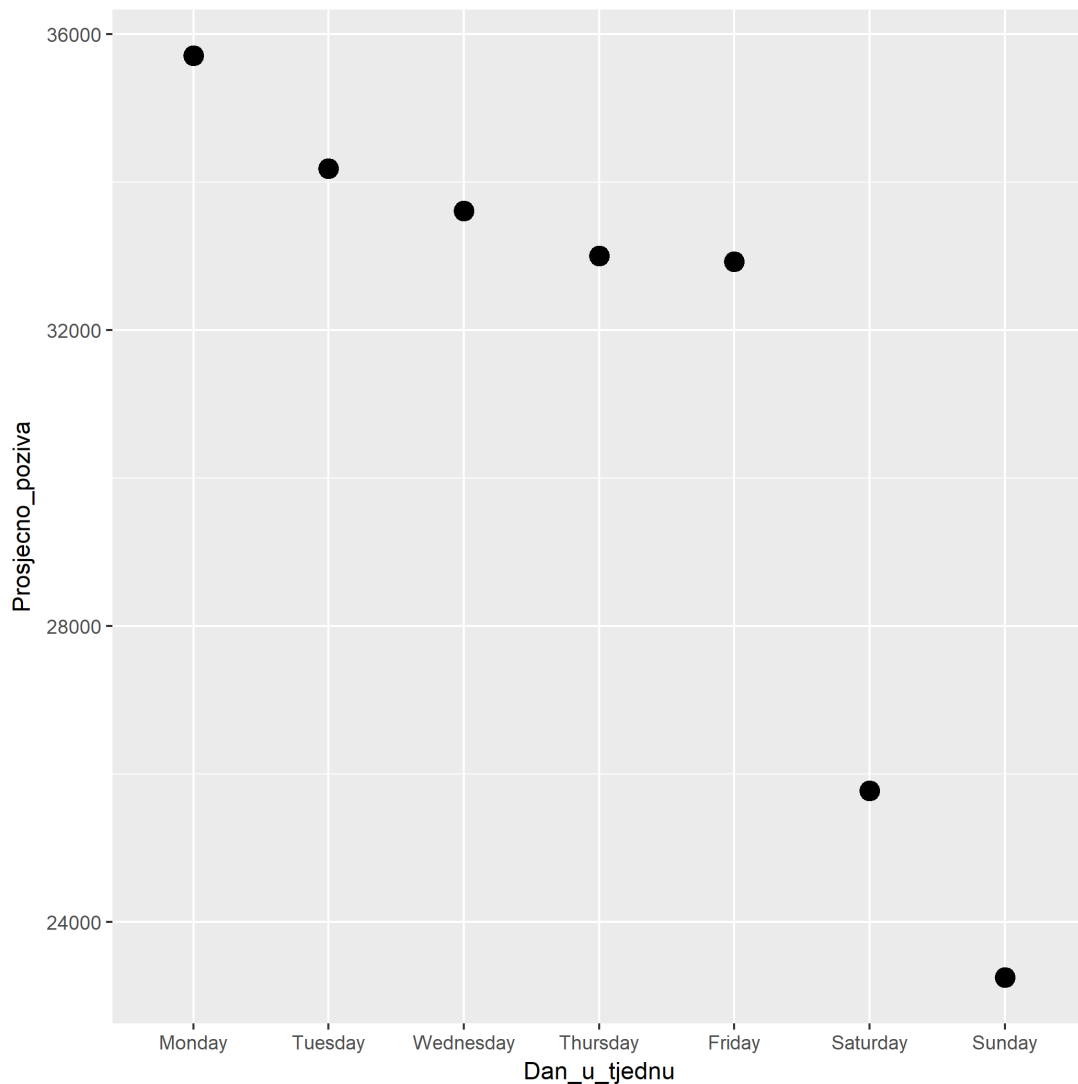
### Broj poziva u ovisnosti o danu u tjednu

Rezultantni graf (slika 3.9) daje zaključiti da broj poziva pada kako tjedan odmiče i vikend se bliži. U problematiku možemo ući još dublje pitajući se: je li razdioba poziva kroz tjedan univerzalna za svakog od operatera?

### Broj poziva u ovisnosti o danu u tjednu i operateru

Ako uzmemo izvorišnog operatera, dobivamo zadovoljavajuće rezultate (slika 3.10). Sada smo već toliko učvrstili svoju sigurnost u ispitanu relaciju da je ona spremna za razvijanje matematičkog modela.

U idealnom slučaju, rezultat prediktivne analize nad navedenom relacijom bi nam

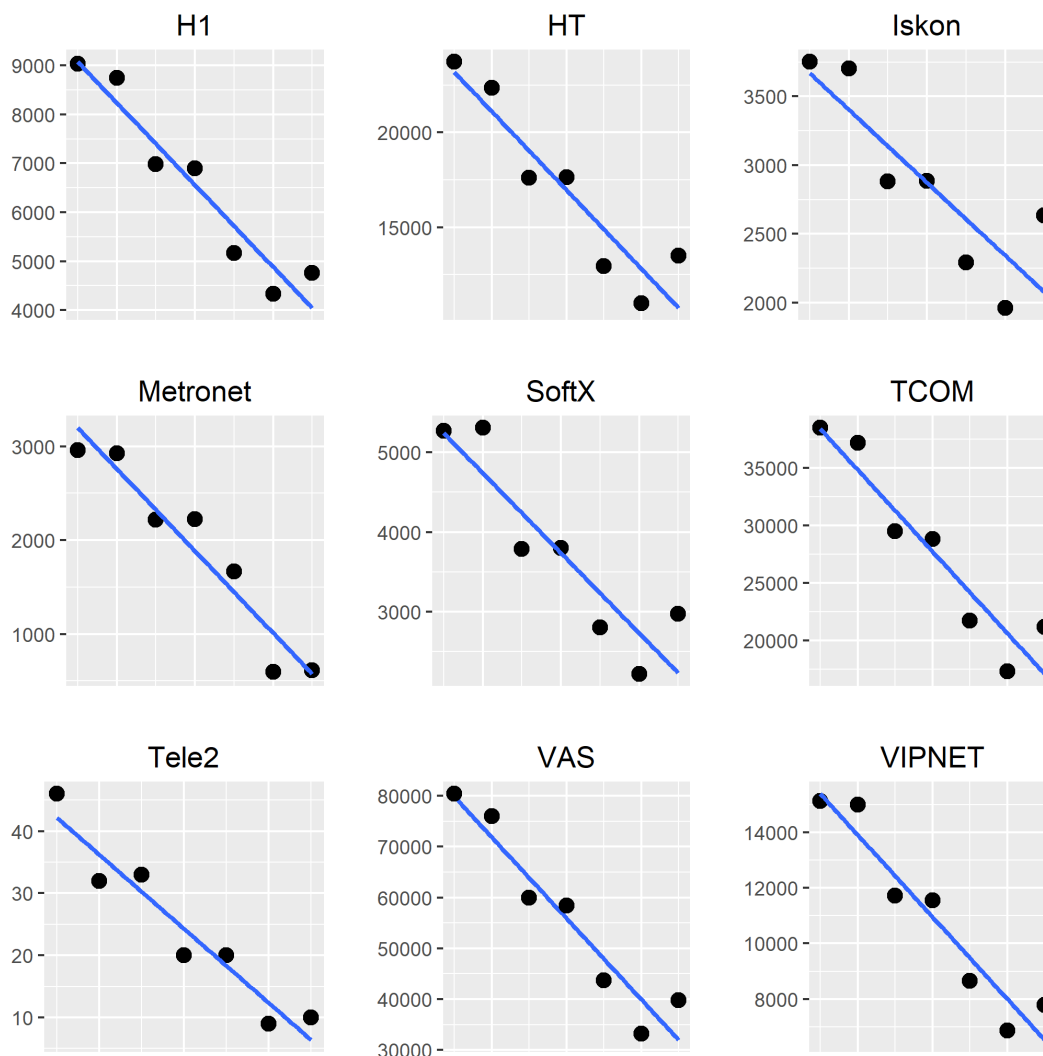


**Slika 3.9:** Prikaz ovisnosti prosječnog broja poziva o danu u tjednu (dodatak A.16)

dao model koji može predvidjeti broj poziva koji će biti upućen na zadani dan u tjednu od svakog pojedinog operatera, ako uzmemo u obzir da dan nije blagdan ili praznik (anomalija). Da bi takve idealne rezultate dobili i da bismo u njih bili sigurni, bili bi nam potrebni podaci za više mjeseci jer ovako imamo samo 4 opažanja za svaki od dana u tjednu (jer mjesec ima 4 tjedna).

### 3.3.2. Duljina razgovora u ovisnosti o vremenu uspostave poziva

Zadovoljni dobivenim rezultatima za prosječan broj poziva u ovisnosti o vremenu, zanima nas možemo li na isti način doći do sličnih zaključaka i za prosječnu duljinu razgovora. Daljnja ispitivanja su slična onima koje smo proveli za ovisnost broja poziva o vremenu uspostave poziva. Tijek analize je gotovo identičan, tako da ćemo se u



**Slika 3.10:** Usporedba razdioba broja poziva kroz tjedan za svakog pojedinog izvorišnog operatera (dodatak A.17)

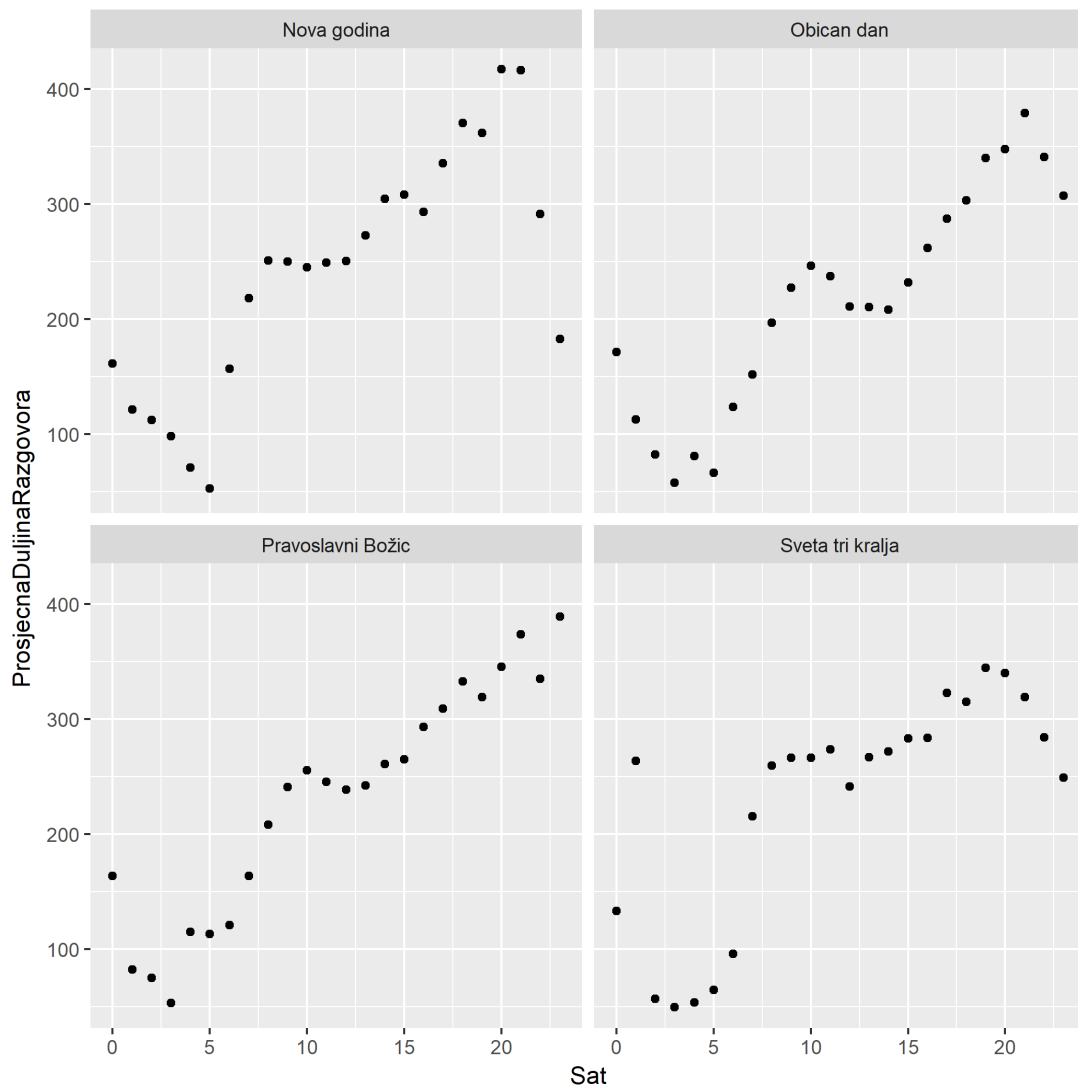
nastavku ispitivanja ove relacije malo manje usredotočiti na proces analize, a više na njene rezultate.

### **Duljina razgovora u ovisnosti o satu u danu**

Za početak nas zanima u koje doba dana ljudi najdulje razgovaraju. Na temelju dobivenih podataka, čini se da ljudi najdulje razgovaraju oko 9 sati navečer (slika 3.11). Razdioba poziva kroz dan za svaki od blagdana je dovoljno slična običnom danu da ih nemamo potrebu isključiti iz daljnje analize. Primjećujemo nagli skok u 1 sat ujutro na blagdan Sveta tri kralja.

Dubljom analizom možemo utvrditi da se radi o dva poziva od 20 minuta kategorije



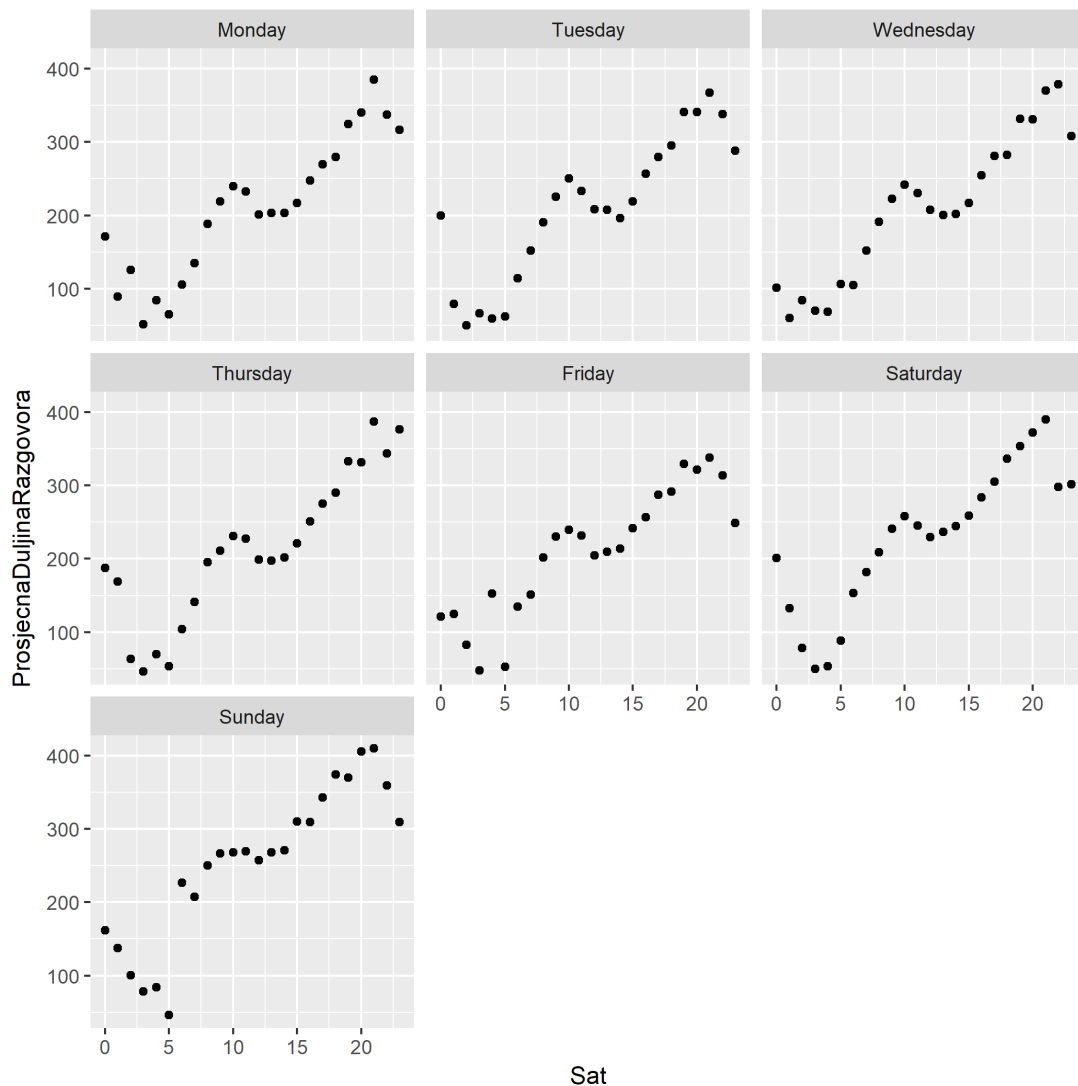


**Slika 3.11:** Usporedni prikaz prosječne duljine razgovora kroz dan za blagdane i normalne dane (dodatak A.18)

“sadržaja neprimjerenog za djecu” koja su uspjela utjecati na prosjek malog broja poziva nastalih u 1 sat ujutro na Sveta tri kralja. Slobodno procjenjujemo da takvi pozivi nisu vezani uz činjenicu da se radi o blagdanu pa ih nemamo potrebu dalje analizirati ili isključiti blagdane iz skupa podataka.

### Duljina razgovora u ovisnosti o satu u danu i danu u tjednu

Napokon možemo provjeriti ovisi li razdioba prosječne duljine poziva kroz dan u ovisnosti o radnosti dana. Rezultati ovog ispitivanja (slika 3.12) pokazuju da razdioba poprima jednaku krivulju za svaki od dana u tjednu, ali različitih intenziteta. Za svaki dan vrijedi da su razgovori sve dulji i dulji kako dan odmiče s maksimumom u 9 sati



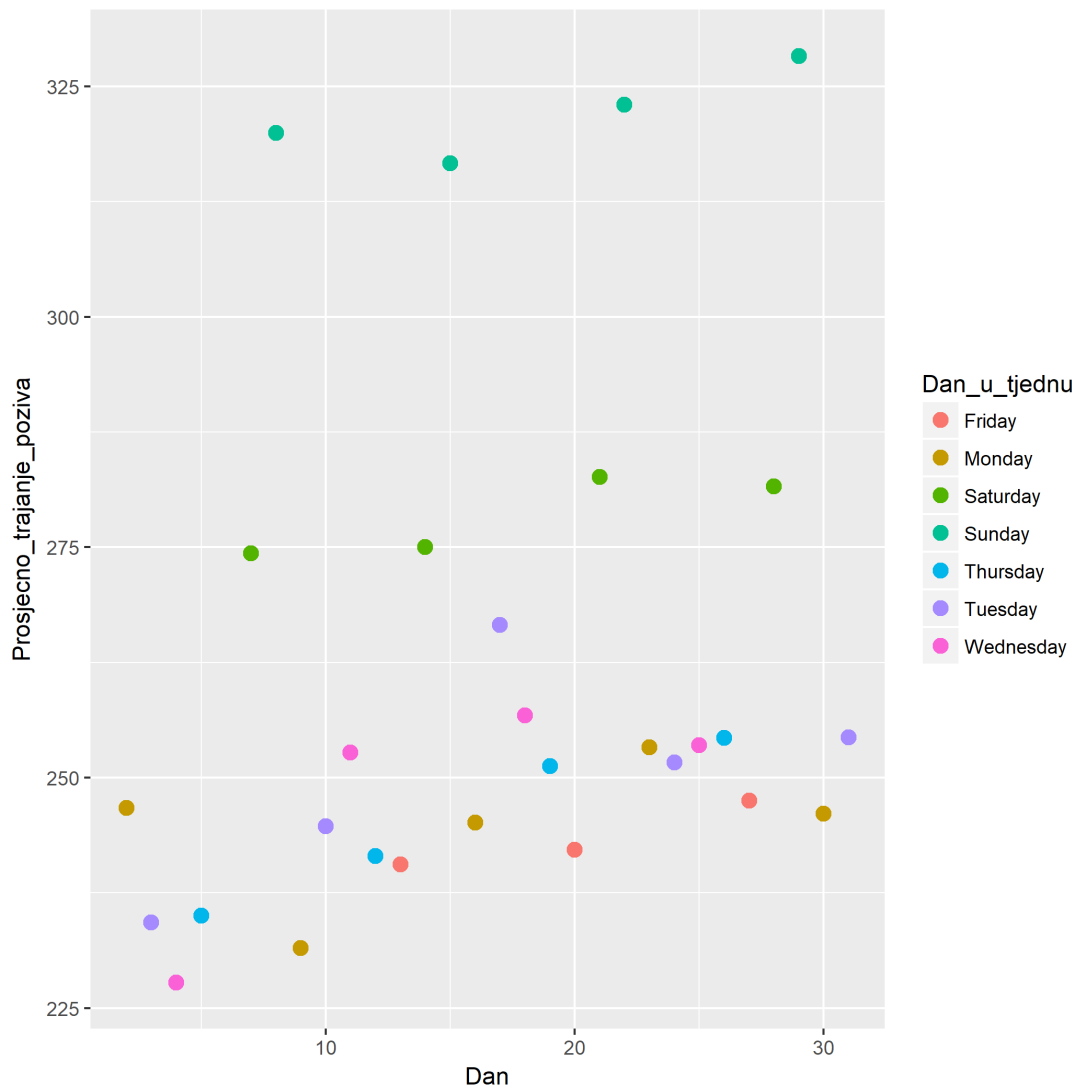
**Slika 3.12:** Usporedni prikaz razdiobe prosječne duljine poziva kroz dan posebno za svaki dan u tjednu (dodatak A.19)

navečer. Stoga, slično kao za prosječan broj poziva, ni za duljinu razgovora ne možemo utvrditi da radnost dana utječe na razdiobu kroz dan.

### Duljina razgovora u ovisnosti o danu u mjesecu i radnosti

Ako promatramo razdiobu duljine razgovora kroz mjesec (slika 3.13), uočavamo da opet imamo potrebu izostaviti blagdane i praznike iz daljnje analize. Nova godina ima kraće pozive od ostalih nedjelja, dok su pozivi na Sveta tri kralja primjetno duži od poziva na ostale petke.

Kada ih uklonimo, dobijemo jasan grafički prikaz (slika 3.14) iz kojeg možemo naslutiti da ljudi na radnije dane kraće razgovaraju. Razdioba prosječne duljine poziva



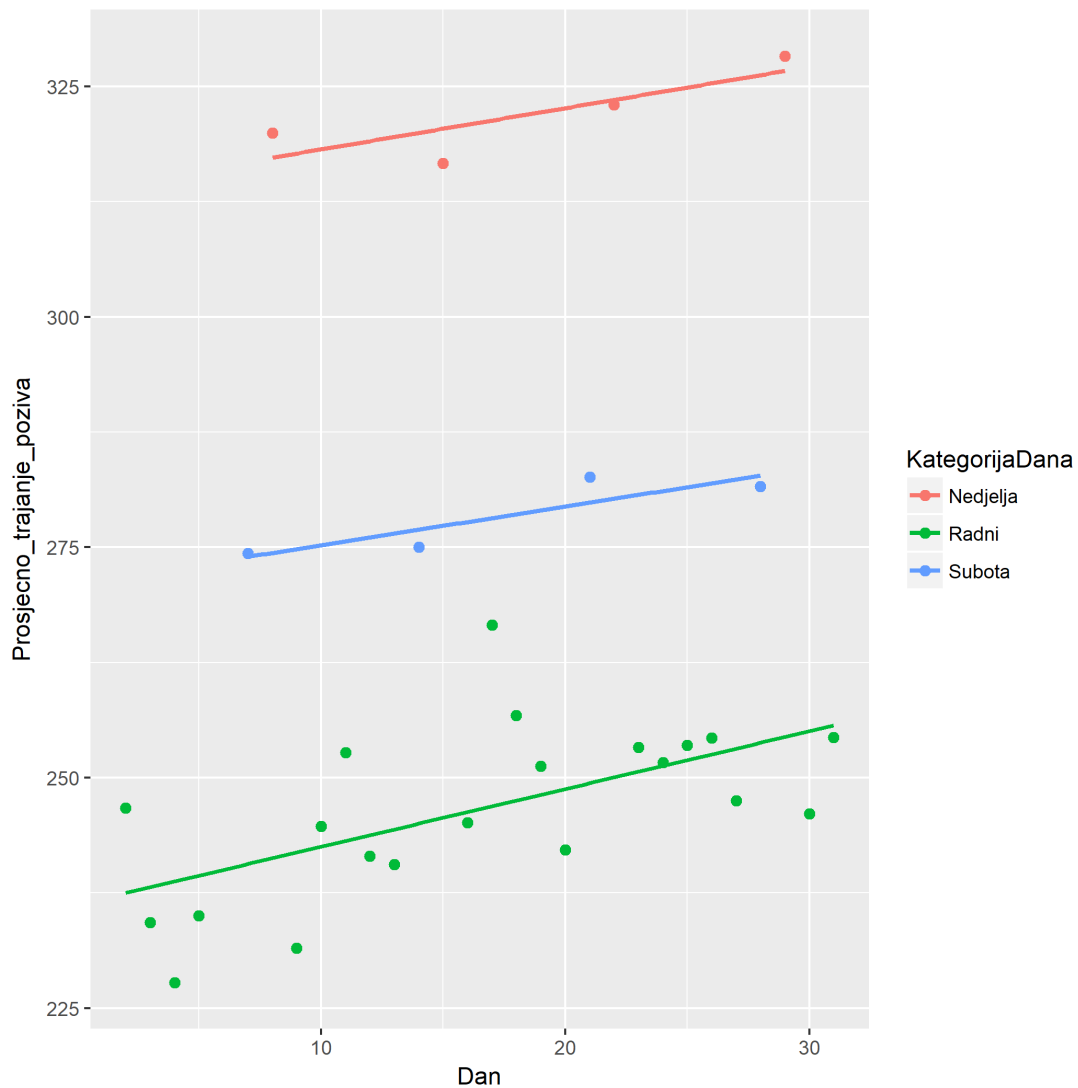
**Slika 3.13:** Prikaz prosječne duljine razgovora kroz mjesec obojano po danima u tjednu (do-datak A.20)

kroz radni tjedan je gotovo uniformna. Rezultati nisu toliko pogodni za razvoj modela kao što je to bio slučaj kod prosječnog broja poziva, ali su zasigurno zanimljivi.

Također, možemo primijetiti blagi porast u duljini razgovora kroz mjesec što nam samo može potaknuti znatiželju o duljini razgovora na razini godine.

### 3.3.3. Zarada u ovisnosti o danu u mjesecu

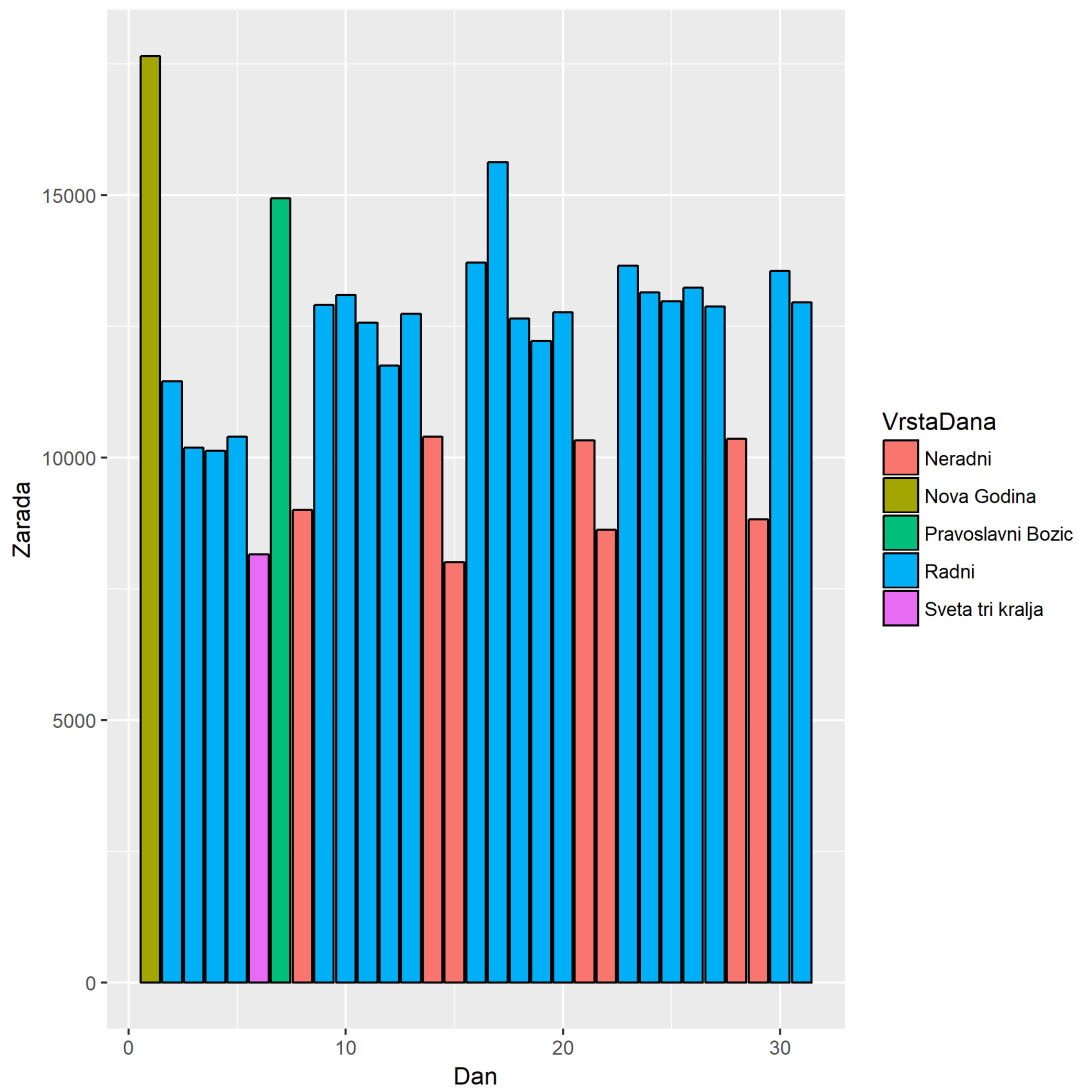
Za telekom je dobro znati kada treba očekivati veliki broj poziva i kada se očekuje da će oni dulje trajati kako bi njihovi korisnici bili zadovoljni uslugom. Ipak, ono što je vjerojatno telekomu zanimljivije čuti kao rezultat analize je kada i kako se uprili najviše novaca.



**Slika 3.14:** Prikaz prosječne duljine razgovora kroz mjesec grupirano po radnosti dana (dodatak A.21)

Pošto se ukupna cijena poziva sastoji od cijene uspostave i cijene po minuti, očekujemo da će prihodi korelirati s većim prosječnim brojem poziva i većom prosječnom duljinom poziva. Prethodna je analiza dala zaključiti da su na dane kada je poziva više, oni najčešće i kraći. Ako pogledamo graf na slici 3.15, vidimo da telekom više uprihodi na radne nego neradne dane, dok najviše uprihodi na blagdane Nove godine i pravoslavni Božić. Pravoslavni Božić nije blagdan u Hrvatskoj pa povlači pitanje: je li broj poziva upućenih prema Srbiji i BiH zamjetno veći na pravoslavni Božić?

Treba spomenuti i anomaliju 17. siječnja kao dan na koji je ostvaren očigledno veći prihod nego na ostale radne dane u mjesecu. Iako nismo dužni istražiti uzrok anomalija kroz eksploratornu analizu, uvijek je pametno na njih obratiti pažnju ljudima koji su



**Slika 3.15:** Prikaz zarade po danima kroz siječanj 2017. (dodatak A.22)

nam podatke ustupili kako bi se one mogle potencijalno dalje analizirati.

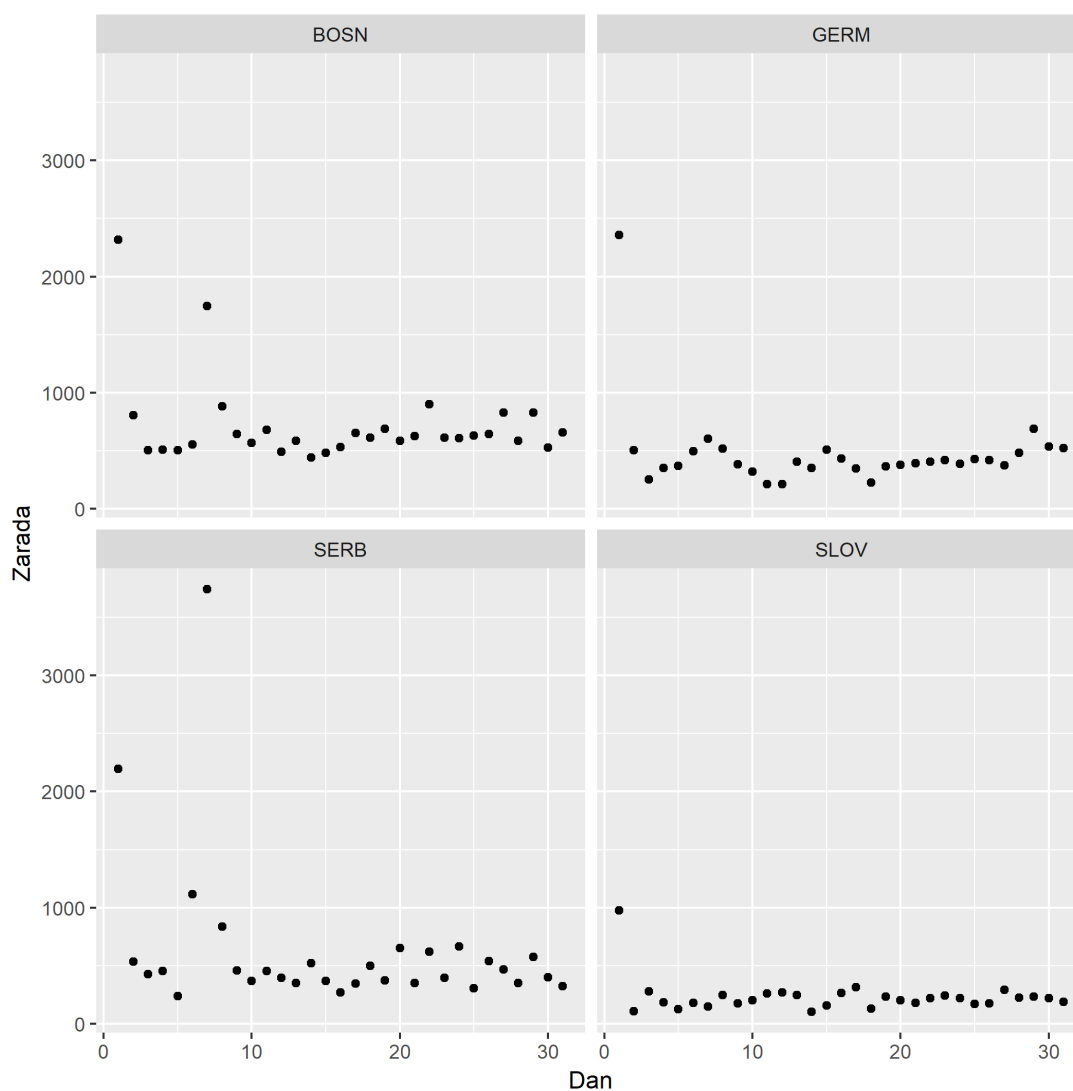
### 3.3.4. Osvrt na pozive prema inozemstvu

Dobiveni podaci sadrže izračunatu ciljnu geografsku ili mobilnu mrežu poziva za koju smo u fazi čišćenja podataka utvrdili da je dostupna samo na odlazne pozive.

Analizom vrijednosti ciljne mreže, utvrdili smo da su pozivi hrvatskih klijenata za koje imamo podatke najčešće upućeni prema Bosni i Hercegovini, te Srbiji (tablica 3.2). Ako promatram razdiobu poziva prema najpopularnijim državama (slika 3.16), možemo uočiti da je pravoslavni Božić prema Srbiji i BiH zaista puno prometniji od ostalih dana u mjesecu pa možemo zaključiti da je to razlog povećanog prometa 7. siječnja.

**Tablica 3.2:** Države prema kojima je upućeno najviše odlaznih poziva

Zemlja	Broj poziva
BiH	5245
Srbija	5109
Njemačka	4697
Slovenija	1586
Italija	987
Austrija	937



**Slika 3.16:** Usporedni prikaz razdioba odlaznih broja poziva za četiri najčešće zvane države (dodatak A.23)

### 3.4. Zaključak analize

Dobili smo sirove podatke vezane za fiksne pozive u siječnju 2017. godine. Podaci su sadržavali puno različitih značajki, ali je većina bila maskirana zbog zaštite podataka i kao takva je bila slabo upotrebljiva. Uočili smo bitne varijable (vrijeme uspostave, cijena, duljina i odredišna mreža poziva) i iz njih odlučili pronaći zanimljive informacije. Osim malog udjela vrlo dugačkih poziva, nije bilo potrebe uklanjati pozive iz podatkovnog skupa.

Dobiveni podaci nam daju naslutiti da su pozivi na radnije dane sve češći i sve kraći (slika 3.17). To i ima smisla s obzirom na to da na radne dane ljudi češće imaju konkretnu poruku koju žele prenijeti, dok se na neradne dane češće zove samo da bi se razgovaralo. Ova relacija je dobro istraжена i ima potencijala da se za nju izgradi matematički model na većem podatkovnom skupu raširenom kroz više mjeseci.

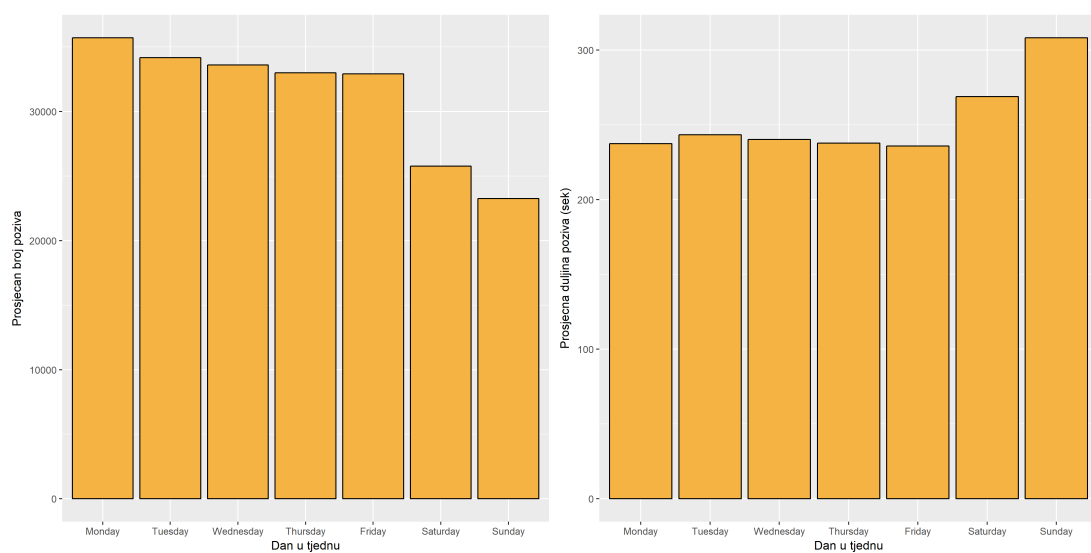
Iznimke predstavljaju blagdani. Ako se radi o blagdanima koji se čestitaju (Nova godina i pravoslavni Božić) onda dan ima više obilježja radnog dana - poziva je sve više i kraći su. S druge strane, ako se blagdan ne čestita (Sveta tri kralja) onda dan ima više obilježja neradnog dana - pozivi su rjeđi i duži.

Istražujući prosječan broj i duljinu poziva kroz dan, zaključili smo da ljudi najviše poziva upućuju oko 10 sati ujutro, a najdulje razgovaraju oko 9 sati navečer. Razlog navedenog ponašanja smo pokušali potražiti u radnim navikama ljudi, ali nismo uspjeli dokazati interakciju s obzirom na to da se isti uzorak ponavlja i za neradne dane i blagdane.

Zarada operatera uglavnom prati frekventnost poziva, ali se opet pojavljuje pravoslavni Božić kao anomalija. Ako promatramo samo ovaj skup od 6775 hrvatskih korisnika i njihovih fiksnih poziva kroz prvi mjesec, mogli bismo zaključiti da Hrvati zamjetno najviše poziva upućuju prema BiH, Srbiji i Njemačkoj. Pogotovo je veliki broj poziva upućen prema BiH i Srbiji na pravoslavni Božić što zbog veće naplate uspostave poziva rezultira većim prihodima operatera.

Iako smo na puno pitanja dali odgovor, postoje još mnoga druga koja preostaju za neku daljnju analizu od kojih mogu izdvojiti:

- što izaziva povećanu zaradu 17.siječnja 2017.?
- zašto ljudi najviše zovu u 10 sati ujutro, a najduže u 9 sati navečer?
- kako se navedene relacije ponašaju na razini godine?



**Slika 3.17:** Usporedba prosječnog broja poziva i duljine poziva kroz tjedan (dodaci A.24 i A.25)



## 4. Zaključak

Ovom je radu bio cilj približiti općeniti proces eksploratorne analize podataka i predstaviti konkretne rezultate analize stvarnih telekomunikacijskih podataka. Cilj smo pokušali ostvariti prolaskom kroz teoretsku pozadinu općenite eksploratorne analize i demonstrirajući tijek analize na stvarnim podacima. Na kraju smo došli do raznih zanimljivih i korisnih zaključaka o ljudskim pozivnim navikama, a neke smo čak znali i objasniti dubljim pogledom u podatke.

Sljedeći logični korak bi bila prediktivna analiza na većem podatkovnom skupu kojoj je cilj stvoriti matematički model iz kvalitetno istraženih relacija. Također je moguće dublje istražiti nastanak anomalija u podacima i njihova svojstva.

# LITERATURA

- [1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.
- [2] John D. Kelleher, Brian MacNamee, Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2015.
- [3] Cole Nussbaumer Knaflic. Exploratory vs explanatory analysis, 2014. URL <http://www.storytellingwithdata.com/blog/2014/04/exploratory-vs-explanatory-analysis>. 8.6.2018.
- [4] Thomas Maydon. The 4 Types Of Data Analytics, 2017. URL <https://insights.principa.co.za/4-types-of-data-analytics-descriptive-diagnostic-predictive-prescriptive>. 3.6.2018.
- [5] Damir Pintar. Programiranje u R-u, 2016. URL [https://ratnip.github.io/FER\\_OPJR\\_2016/](https://ratnip.github.io/FER_OPJR_2016/). 31.5.2018.
- [6] John Spacey. 14 Types of Data Analysis, 2015. URL <https://simplicable.com/new/data-analysis>. 2.6.2018.
- [7] Hadley Wickham. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition*. O'Reilly Media, 2016.

## A. Isječci koda u jeziku R

### A.1. Učitavanje podataka

```
teldata <- read.csv2("1mj_1M_1MilUTF.dsv", encoding = "
  UTF-8" , stringsAsFactors = F, header=TRUE)
teldata <- tbl_df(teldata)
```

### A.2. Pretvaranje tipova

```
teldata <- mutate(teldata , TRANS_DATE = dmy_hms(TRANS_
  DATE))
teldata <- mutate(teldata , AMOUNT = as.double(AMOUNT))
teldata <- mutate(teldata , COMPUTED_PRICE_PER_MINUTE = as
  .double(COMPUTED_PRICE_PER_MINUTE))
teldata <- mutate(teldata , COMPUTED_OFF_PEAK_PRICE_PER_M
  = as.double(COMPUTED_OFF_PEAK_PRICE_PER_M))
teldata <- mutate(teldata , COMPUTED_ON_PEAK_PRICE_PER_M =
  as.double(COMPUTED_ON_PEAK_PRICE_PER_M))
teldata <- mutate(teldata , COMPUTED_ESTABLISHMENT_FEE =
  as.double(COMPUTED_ESTABLISHMENT_FEE))
teldata <- mutate(teldata , COMPUTED_ROUNDED_DURATION = as
  .double(COMPUTED_ROUNDED_DURATION))
```

### A.3. Identifikacija nepostojećih vrijednosti

```
teldata <- mutate(teldata , COMPUTED_PRICE_PER_MINUTE =
  ifelse(is.na(COMPUTED_PRICE_PER_MINUTE) , 0 , COMPUTED_
  PRICE_PER_MINUTE))
teldata <- mutate(teldata , COMPUTED_ESTABLISHMENT_FEE =
```

```

    ifelse(is.na(COMPUTED_ESTABLISHMENT_FEE), 0, COMPUTED_
    ESTABLISHMENT_FEE))
teldata <- mutate(teldata, COMPUTED_ON_PEAK_PRICE_PER_M =
    ifelse(is.na(COMPUTED_ON_PEAK_PRICE_PER_M), 0,
    COMPUTED_ON_PEAK_PRICE_PER_M))
teldata <- mutate(teldata, COMPUTED_OFF_PEAK_PRICE_PER_M
    = ifelse(is.na(COMPUTED_OFF_PEAK_PRICE_PER_M), 0,
    COMPUTED_OFF_PEAK_PRICE_PER_M))
# teldata <- mutate(teldata, COMPUTED_ROUNDED_DURATION =
    ifelse(COMPUTED_ROUNDED_DURATION==0, NA, COMPUTED_
    ROUNDED_DURATION))
teldata <- mutate(teldata, COMPUTED_DESTINATION = ifelse(
    COMPUTED_DESTINATION=="", NA, COMPUTED_DESTINATION))

```

#### **A.4. Popunjavanje nepostojećeg izvorišnog operatera pretpostavljenim (VAS)**

```

teldata <- mutate(teldata, ITEM_ORIGINATING_CARRIER_ID =
    ifelse(ITEM_ORIGINATING_CARRIER_ID=="", "385-VAS",
    ITEM_ORIGINATING_CARRIER_ID))

```

#### **A.5. Provjera da su svi zapisi vezani uz pozive**

```

mean(teldata$PARAMETER_SERVICE_GROUP == "voice")
# 1

```

#### **A.6. Provjera da su svi pozivi uspostavljeni u siječnju 2017. godine**

```

min(teldata$TRANS_DATE)
# "2017-01-01 00:00:04 UTC"
max(teldata$TRANS_DATE)
# "2017-01-31 23:58:53 UTC"

```

#### **A.7. Uklanjanje vrste usluge**

```
teldata$PARAMETER_SERVICE_GROUP <- NULL
```

## **A.8. Univarijantna analiza duljine poziva**

```
ggplot(teldata , aes (ITEM_CALL_DURATION)) + geom_density (
  fill="#f4b342" , alpha = 3/4) + xlim(0,3600) + xlab ("
  Duljina_razgovora") + ylab ("Postotak_zastupljenosti")
```

## **A.9. Univarijantna analiza cijene poziva**

```
ggplot(teldata , aes (AMOUNT)) + geom_density ( fill="#
  f4b342" , alpha = 3/4) + xlim(0,1) + xlab ("Ukupna_
  cijena_poziva_(HRK)") + ylab ("Postotak_zastupljenosti "
  )
```

## **A.10. Univarijantna analiza vremena uspostave poziva**

```
ggplot(teldata , aes (TRANS_DATE)) + geom_density ( fill="#
  f4b342" , alpha=3/4) + ylab ("Udio_poziva") + xlab ("
  Vrijeme_uspostave") + ylab ("Postotak_zastupljenosti")
```

## **A.11. Ostvarenje prikaza ovisnosti prometa o satu u danu**

*# Izdvajanje*

```
teldataPrometniSati <- group_by(teldata , Sat = hour(TRANS
  _DATE)) %>% summarise(Poziva = n())
```

*# Vizualizacija*

```
ggplot(teldataPrometniSati , aes(x=Sat , y = Poziva)) +
  geom_point ( size=4)
```

## **A.12. Ostvarenje prikaza ovisnosti prometa o satu u danu i radnosti**

*# Opcenitija relacija*

```
dataSatDanRadni <- group_by(teldata , Dan = day(TRANS_DATE
), Dan_u_tjednu = weekdays(TRANS_DATE), Sat = hour(
TRANS_DATE)) %>% summarise(Poziva = n()) %>% mutate(
JeNeradniDan = Dan_u_tjednu %in% c("Sunday", "Saturday
") || Dan == 6 , JeRadni = !JeNeradniDan) %>% filter(
Dan != 1)
```

*# Izdvajanje*

```
dataSatRadni <- dataSatDanRadni %>% group_by(Sat, JeRadni
) %>% summarise(Prosjecno_poziva = mean(Poziva))
```

*# Vizualizacija*

```
ggplot(dataSatRadni , aes(x=Sat , y=Prosjecno_poziva , color
= JeRadni)) + geom_point(size=4) #+ stat_smooth(
method = "loess" , se = FALSE)
```

### **A.13. Ostvarenje prikaza ovisnosti prometa o satu u danu kroz tjedan**

```
ggplot(dataSatDan , aes(Sat , Prosjecno_poziva)) +
geom_point() +
facet_wrap(~ Dan_u_tjednu)
```

### **A.14. Ostvarenje prikaza ovisnosti prometa o danu u mjesecu**

*# Izdvajanje*

```
teldataPrometniDani <- group_by(teldata , Dan = day(TRANS_
DATE)) %>% summarise(Poziva = n())
```

*# head(teldataPrometniDani)*

*# Vizualizacija*

```
ggplot(teldataPrometniDani , aes(x=Dan, y = Poziva)) +
geom_point(size=4)
```

## A.15. Ostvarenje prikaza ovisnosti prometa o danu u mjesecu i radnosti

*# Izdvajanje*

```
teldataPrometniDaniUTjednu <- group_by(teldata , Dan = day
  (TRANS_DATE) , Dan_u_tjednu = weekdays(TRANS_DATE)) %>%
  summarise(Poziva = n()) %>% mutate(JeNeradniDan = Dan
  _u_tjednu %in% c("Sunday", "Saturday") || Dan == 6 )
%>% filter (Dan!=1 & Dan!=6 & Dan != 7)
```

*# Faktorizacija – omogućuje da redoslijed prati stvarne dane u tjednu*

```
teldataPrometniDaniUTjednu <- teldataPrometniDaniUTjednu
%>% mutate(KategorijaDana = if_else(JeNeradniDan , if_
  else(Dan_u_tjednu == "Sunday", "Nedjelja", "Subota") ,
  "Radni"))
```

*# Vizualizacija*

```
ggplot(teldataPrometniDaniUTjednu , aes(x=Dan , y=Poziva ,
  color = KategorijaDana)) + geom_point()+ stat_smooth(
  method="lm" , se = FALSE)
```

## A.16. Ostvarenje prikaza ovisnosti prometa o danu u tjednu

*# Izdvajanje*

```
teldataPrometniDaniUTjednu <- group_by(
  teldataPrometniDaniUTjednu , Dan_u_tjednu) %>% mutate(
  Prosjecno_poziva = mean(Poziva))
```

*# Faktorizacija*

```
teldataPrometniDaniUTjednu$Dan_u_tjednu <- factor(
  teldataPrometniDaniUTjednu$Dan_u_tjednu , levels = c("
  Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
  "Saturday", "Sunday"))
```

```
# Vizualizacija
```

```
ggplot(teldataPrometniDaniUTjednu , aes(Dan_u_tjednu , y =  
  Prosjecno_poziva)) + geom_point(size=4)
```

## **A.17. Ostvarenje prikaza ovisnosti prometa o danu u tjednu i operateru**

```
# Iako je općenito pametnije koristiti funkciju facet_  
wrap za uspoređivanje grafova, ona izjednaci  
vrijednosti osi i na taj način onemogućuje analizu malih  
operatera
```

```
# Izdvajanje
```

```
teldataPrometniDaniUTjednuOperateri <- teldata %>% filter  
  (!day(TRANS_DATE)%in% c(1,6,7)) %>% mutate(Dan_u_  
  tjednu = weekdays(TRANS_DATE)) %>% group_by(Dan_u_  
  tjednu , ITEM_ORIGINATING_CARRIER_ID) %>% summarise(  
  Prosjecno_poziva = mean(n()))
```

```
# Faktorizacija
```

```
teldataPrometniDaniUTjednuOperateri$Dan_u_tjednu <-  
  factor(teldataPrometniDaniUTjednuOperateri$Dan_u_  
  tjednu , levels = c("Monday", "Tuesday", "Wednesday", "  
  Thursday", "Friday", "Saturday", "Sunday"))
```

```
# Za svakog od 9 operatera odvajamo posebne podatke
```

```
danUTjednuTCOM <- teldataPrometniDaniUTjednuOperateri %>%  
  filter(ITEM_ORIGINATING_CARRIER_ID == "385-TCOM")
```

```
# Onda za svakog stvaramo graficki prikaz
```

```
gTCOM <- ggplot(danUTjednuTCOM , aes(x=as.numeric(Dan_u_  
  tjednu), y = Prosjecno_poziva)) + geom_point(size=3) +  
  stat_smooth(method="lm", se=FALSE) + xlab("") + ylab(  
  "")+ ggtitle("TCOM") + theme(plot.title = element_text
```



```
(hjust = 0.5), axis.text.x=element_blank(), axis.ticks
.x=element_blank())

# I na kraju sve graficke prikaze spajamo u jednu mrezu
grid.arrange(gH1, gHT, gIskon, gMetronet, gSoftX, gTCOM,
gTele2, gVAS,gVIPNET, ncol = 3, nrow = 3)
```

## A.18. Ostvarenje prikaza ovisnosti duljine razgovora o satu u danu

```
# Izdvajanje
DuljinaRazgovoraSatObican <- teldata %>% mutate(Sat =
  hour(TRANS_DATE), Dan_u_tjednu = weekdays(TRANS_DATE),
  Dan=day(TRANS_DATE), Kategorija = if_else(Dan ==1 , "
Nova_godina", if_else(Dan == 6, "Sveta_tri_kralja", if
_else(Dan == 7, "Pravoslavni_Bozic", "Obican_dan"))))
%>% group_by(Sat, Kategorija) %>% summarise(
  ProsjecnaDuljinaRazgovora = mean(ITEM_CALL_DURATION))

# Vizualizacija
ggplot(DuljinaRazgovoraSatObican, aes(Sat,
  ProsjecnaDuljinaRazgovora)) +
  geom_point() +
  facet_wrap(~ Kategorija)
```

## A.19. Ostvarenje prikaza ovisnosti duljine razgovora o satu u danu i danu u tjednu

```
# Izdvajanje
DuljinaRazgovoraSat <- teldata %>% mutate(Sat = hour(
  TRANS_DATE), Dan_u_tjednu = weekdays(TRANS_DATE), Dan=
  day(TRANS_DATE)) %>% group_by(Sat, Dan_u_tjednu) %>%
  summarise(ProsjecnaDuljinaRazgovora = mean(ITEM_CALL_
  DURATION))
```

```
# Faktorizacija
```

```
DuljinaRazgovoraSat$Dan_u_tjednu <- factor(  
  DuljinaRazgovoraSat$Dan_u_tjednu , levels = c("Monday",  
    "Tuesday", "Wednesday", "Thursday", "Friday", "  
    Saturday", "Sunday"))
```

```
# Vizualizacija
```

```
ggplot(DuljinaRazgovoraSat , aes(Sat ,  
  ProsjecnaDuljinaRazgovora)) +  
  geom_point() +  
  facet_wrap(~ Dan_u_tjednu)
```

## **A.20. Ostvarenje prikaza ovisnosti duljine razgovora o danu u mjesecu**

```
# Izdvajanje
```

```
duljinaPozivaVrijeme <- teldata %>% mutate(Dan=day(TRANS_  
  DATE) , Dan_u_tjednu = weekdays(TRANS_DATE) , JeRadni =  
  !(Dan_u_tjednu %in% c("Sunday", "Saturday") | Dan ==  
  6)) %>% group_by(Dan , Dan_u_tjednu , JeRadni) %>%  
  summarise(Prosjecno_trajanje_poziva = mean(ITEM_CALL_  
  DURATION))
```

```
# Vizualizacija
```

```
ggplot(duljinaPozivaVrijeme , aes(x=Dan, y=Prosjecno_  
  trajanje_poziva , color = Dan_u_tjednu)) + geom_point(  
  size=3) #+ stat_smooth(method="lm", se = FALSE)
```

## **A.21. Ostvarenje prikaza ovisnosti duljine razgovora o dana u mjesecu i radnosti dana**

```
# Izdvajanje
```

```
duljinaPozivaVrijemeRadnost <- duljinaPozivaVrijeme %>%  
  mutate(KategorijaDana = if_else(JeRadni , "Radni" , if_  
    else(Dan_u_tjednu == "Sunday", "Nedjelja", "Subota")))
```

```
ggplot(duljinaPozivaVrijemeRadnost, aes(x=Dan, y=
  Prosjecno_trajanje_poziva, color = KategorijaDana)) +
  geom_point(size=2)+ stat_smooth(method="lm", se =
  FALSE)
```

## A.22. Ostvarenje prikaza ovisnosti zarade i dana u mjesecu

```
# Izdvajanje
zaradaDan <- teldata %>% mutate(Dan = day(TRANS_DATE),
  Dan_u_tjednu = weekdays(TRANS_DATE), JeRadni = !(Dan_u_
  _tjednu %in% c("Sunday", "Saturday")), VrstaDana =
  ifelse(Dan==1, "Nova_Godina", if_else(Dan==6, "Sveta_
  tri_kralja",
  if_else(Dan == 7, "Pravoslavni_Bozic", if_else(JeRadni, "
  Radni", "Neradni"))))) %>% group_by(Dan, JeRadni,
  VrstaDana) %>% summarise(Zarada = sum(AMOUNT))

ggplot(zaradaDan, aes(x=Dan, y=Zarada, fill=VrstaDana)) +
  stat_summary(geom="bar", color="Black")
```

## A.23. Istraživanje razdioba poziva prema inozemstvu kroz mjesec

```
# Izdvajanje za tablicu
zemljaPoziva <- teldataDestEstCost %>% filter(COMPUTED_
  DESTINATION == toupper(COMPUTED_DESTINATION)) %>%
  group_by(Zemlja = substr(COMPUTED_DESTINATION,0,4))
  %>% summarise(Poziva = sum(N)) %>% arrange(desc(Poziva
  ))

head(zemljaPoziva) # prikaz prvih 6 zemalja

# Izdvajanje za graf
```

```
zaradaInozemstvo <- teldata %>% mutate(Zemlja = substr(
  COMPUTED_DESTINATION,0,4)) %>% filter(Zemlja %in% c("
  BOSN","SERB","GERM","SLOV")) %>% mutate(Dan = day(
  TRANS_DATE), Dan_u_tjednu = weekdays(TRANS_DATE),
  JeRadni = !(Dan_u_tjednu %in% c("Sunday", "Saturday")
  | Dan == 6 | Dan == 1)) %>% group_by(Dan, JeRadni,
  Zemlja) %>% summarise(Zarada = sum(AMOUNT))
```

```
# Vizualizacija
```

```
ggplot(zaradaInozemstvo, aes(Dan, Zarada)) +
  geom_point() +
  facet_wrap(~ Zemlja)
```

## A.24. Uljepšan prikaz ovisnosti broja poziva o danu u tjednu

```
ggplot(teldataPrometniDaniUTjednu %>% select(Dan_u_tjednu
, Prosjecno_poziva) %>% unique, aes(x=Dan_u_tjednu, y
= Prosjecno_poziva)) + geom_bar(stat="identity",
color="Black", fill="#f4b342") + xlab("Dan_u_tjednu")
+ ylab("Prosjecan_broj_poziva") # + ggtitle("Prosjecan
broj poziva kroz tjedan") + theme(plot.title =
element_text(hjust = 0.5))
```

## A.25. Uljepšan prikaz ovisnosti duljine poziva o danu u tjednu

```
ggplot(duljinaPozivaKrozTjedan, aes(x=Dan_u_tjednu, y =
Prosjecno_trajanje_poziva)) + geom_bar(stat="identity"
, color="Black", fill="#f4b342") + xlab("Dan_u_tjednu"
) + ylab("Prosjecna_duljina_poziva_(sek)")# + ggtitle
("Prosjecna duljina poziva kroz tjedan") + theme(plot
.title = element_text(hjust = 0.5))
```

# **Eksploratorna analiza telekomunikacijskih podataka u svrhu detekcije anomalija**

## **Sažetak**

Ovaj rad prolazi kroz osnovne teoretske principe eksploratorne analize i njenu primjenu na stvarnim podacima iz telekomunikacijske domene. U središtu rada je tijek eksploratorne analize, ali se dotiče i problema detekcije i analize anomalija, prediktivne analize i izvještavanja. Sustavni pristup analizi podržan je i isječcima koda u programskog jeziku R kojima se čitatelja poziva na aktivno sudjelovanje.

**Ključne riječi:** Univarijantna analiza, vizualizacija, kvaliteta podataka, prilagodba podataka, izvještavanje, bitne varijable

# **Exploratory analysis of telecommunication data for the purpose of anomaly detection**

## **Abstract**

This paper goes through the underlying theoretical principles of exploratory analysis and its application on real-time data from the telecommunications domain. The focus of the paper is the course of exploratory analysis, but it also concerns the problems of anomaly detection and analysis, predictive analysis and reporting. The systematic approach to the analysis is supported by code snippets in the R programming language that invite readers to active participation.

**Keywords:** Univariate analysis, visualisation, data quality, data wrangling, reporting, important variables