



Univerza v Mariboru

Fakulteta za organizacijske vede

Vuk Stojković

Analiza podatkov pridobljenih z aplikacijo KAMbi

Diplomsko delo

Kranj, Julij 2021



Univerza v Mariboru

Fakulteta za organizacijske vede

Vuk Stojković

Analysis of data collected by KAMbi application

Bachelor thesis

Kranj, July 2021

Analiza podatkov pridobljenih z aplikacijo KAMbi

Diplomsko delo

Študent: Vuk Stojković

Študijski program: Univerzitetni študij

Smer: Organizacija in management informacijskih sistemov

Mentor: red. prof. dr. Robert Leskovar

ZAHVALA

Zahvaljujem se profesorju dr. Robertu Leskovarju za profesionalen pristop, strokovne nasvete ter navodila v procesu raziskovanja ter seveda družini in prijateljem, ki so me svojim znanjem in izkušnjami nenehno spodbujali v času študija.

Analiza podatkov pridobljenih z aplikacijo KAMbi

Ključne besede: analiza podatkov, KAMbi, R, programiranje, znanost o podatkih

Povzetek

Diplomska naloga obravnava podatke, ki so bili pridobljeni s spletno aplikacijo KAMbi. Ta je namenjena dijakom zaključnih letnikov srednjih šol kot pomoč pri izbiri inženirskega študija. Aplikacija uporabnika vodi preko 78 vprašanj o želeni naravi dela in samooceni lastnih kompetenc. Obdelano je bilo 1750 anket. Predstavljena so programska orodja za izvedbo ankete in obdelavo: a) aplikacija KAMbi, ki je napisana z orodjem za malo-kodno programiranje Oracle Application Express in b) jezik R z nekaj najpomembnejših knjižnic, ki so bile uporabljene pri analizi podatkov. Razvit je skript v jeziku R, katerega glavne komponente so: a) branje vhodnih podatkov iz izvoženih tabel baze ali pa direktno branje iz baze podatkov, b) priprava delovnih spremenljivk kot so matrika odgovorov in matrika časov, c) funkcije za izračun dosežka anketiranca po področjih in primerjavo podobnosti odgovorov, d) opisna statistika s histogrami odgovorov ter področij, e) korelacijska analiza ter f) analiza gruč in čiščenje podatkov. Iz opisne statistike je razvidno, da odgovori pogosteje izražajo pozitivne lastnosti anketiranca. Časi reševanja posameznih odgovorov so od 1 do 30 sekund. Pri čiščenju podatkov smo predpostavili, da anketiranec s poprečnimi kognitivnimi sposobnostmi ne more pošteno odgovoriti na vsa vprašanja v manj kot 380 sekundah. Korelacije med odgovori in področji so statistično šibke in nesignifikantne. Analiza gruč je odgovore klasificirala v dve skupini tako v primeru vseh anket kot v primeru očiščenih podatkov.

Analysis of data collected by KAMbi application

Keywords: data analysis, KAMbi, R, programming, data science

Abstract

The research addresses the analysis of the data obtained with the KAMbi web application. The application is intended as support to choose higher education study programme in engineering. The application guides the user over 78 questions about the desired nature of work and self-assessment of their own competencies. In total 1750 surveys were processed. The software tools for conducting the survey and processing are presented: a) the KAMbi application, which is written in Oracle Application Express, and the R language with most important libraries that were used in the data analysis. A script in R language has been developed. The main components of this script: a) reading input data from exported database tables or reading directly from the database, b) preparation of working variables such as answer matrix and time matrix, c) functions for calculating the respondent's achievement in areas and comparison of similarity of responses, d) descriptive statistics with histograms of responses and areas, e) correlation analysis, and f) cluster analysis and data cleaning. Descriptive statistics shows that the answers more often express the positive characteristics of the respondent. The times for answering individual answers vary from 1 to 30 seconds. In data cleaning step we assumed that a respondent with average cognitive abilities could not honestly answer all the questions in less than 380 seconds. Correlations between responses and areas are statistically weak and insignificant. The cluster analysis classified the responses into two groups both in the case of all surveys and in the case of cleaned data.

Številka: 423-41/2021/2/210-SB

Datum in kraj: Kranj, 11.03.2021

Na osnovi 330. člena Statuta Univerze v Mariboru (Uradni list RS, št. 29/2017-UPB12, 32/2019 in 14/2020) izdajam:

SKLEP O ZAKLJUČNEM DELU

Stojković Vuk, študentu-ki UN študijskega programa prve stopnje ORGANIZACIJA IN MANAGEMENT INFORMACIJSKIH SISTEMOV se dovoljuje izdelati zaključno delo.

Tema zaključnega dela je pretežno s področja Katedre za informatiko.

Mentor/-ica: red. prof. dr. ROBERT LESKOVAR, univ. dipl. org.

Somentor/-ica:

Naslov zaključnega dela: ANALIZA PODATKOV PRIDOBLENIH Z APLIKACIJO KAMBI

Naslov zaključnega dela v angleškem jeziku: ANALYSIS OF DATA COLLECTED BY KAMBI APPLICATION

Rok za izdelavo in oddajo zaključnega dela je **30.09.2021**. Zaključno delo je potrebno izdelati skladno z »Navodili o postopku priprave in zagovora zaključnega dela na študijskih programih prve in druge stopnje FOV« in ga v treh izvodih oddati v referatu za študentske zadeve FOV. Hkrati se odda tudi izjava mentor-ja/-ice (in morebitne/-ga somentorj-ja/-ice) o ustreznosti zaključnega dela.

Pravni pouk: Zoper ta sklep je možna pritožba na Senat FOV v roku 10 delovnih dni od prejema sklepa.

V. d. prodekanice za izobraževalno dejavnost:
izr. prof. dr. Anja Žnidaršič

Obvestiti:

- Stojković Vuk
- red. prof. dr. ROBERT LESKOVAR, univ. dipl. org.
- mapa 41059046



KAZALO VSEBINE

KAZALO SLIK	VII
1 Uvod	1
2 Metodologija	2
2.1 Definicija problema	2
2.2 Definicija ciljev	2
2.3 Uporabljene metode in orodja	2
3 Predstavitev orodij	3
3.1 Aplikacija KAMbi	3
3.2 Baza podatkov za aplikacijo KAMbi	6
3.3 Programski jezik R in RStudio	6
3.4 Najpomembnejši paketi za konkretno nalogu	7
4 Razvoj skripta za obdelavo podatkov	9
5 Opisna statistika in korelacijska analiza	22
5.1 Opisna statistika za odgovore	22
5.2 Opisna statistika za čase odgovorov	34
5.3 Opisna statistika - področja	45
5.4 Korelacijska analiza odgovorov	49
5.5 Analiza podobnosti odgovorov med kandidati	53
5.6 Analiza gruč in čiščenje podatkov	54
6 Diskusija	64
7 Zaključki	65
8 Literatura	66

KAZALO SLIK

Slika 3.1: Prva skupina vprašanj aplikacije KAMbi	3
Slika 3.2: Druga skupina vprašanj aplikacije KAMbi	3
Slika 3.3: KAMbi - ujemanje preferenc o naravi dela kandidata z izbranim poklicem	4
Slika 3.4: KAMbi - ujemanje kompetenc kandidata z izbranim poklicem	4
Slika 3.5: Relacijski model podatkovne baze KAMbi aplikacije	6
Slika 3.6: Dodatek 1: SQL skripta za kreiranje tabel, sekvenc in sprožilcev KAMbi baze	6
Slika 3.7: IDE - R Studio	7
 Slika 5.1: Histogram: vprašanja 1-9	26
Slika 5.2: Histogram: vprašanja 10-18	27
Slika 5.3: Histogram: vprašanja 19-27	28
Slika 5.4: Histogram: vprašanja 28-36	29
Slika 5.5: Histogram: vprašanja 37-39	29
Slika 5.6: Histogram: vprašanja 40-48	30
Slika 5.7: Histogram: vprašanja 49-57	31
Slika 5.8: Histogram: vprašanja 58-66	32
Slika 5.9: Histogram: vprašanja 67-75	33
Slika 5.10: Histogram: vprašanja 76-78	33
Slika 5.11: Histogram časov: vprašanja 1-9	37
Slika 5.12: Histogram časov: vprašanja 10-18	38
Slika 5.13: Histogram časov: vprašanja 19-27	39
Slika 5.14: Histogram časov: vprašanja 28-36	40
Slika 5.15: Histogram časov: vprašanja 37-39	40
Slika 5.16: Histogram časov: vprašanja 40-48	41
Slika 5.17: Histogram časov: vprašanja 49-57	42
Slika 5.18: Histogram časov: vprašanja 58-66	43
Slika 5.19: Histogram časov: vprašanja 67-75	44
Slika 5.20: Histogram časov: vprašanja 76-78	44
Slika 5.21: Histogram področij 1-9	46
Slika 5.22: Histogram področij 10-18	47
Slika 5.23: Histogram področij 19-26	48
Slika 5.24: Prikaz korelacije med odgovori	49

Slika 5.25: Signifikantne korelacijske matrice med odgovori	50
Slika 5.26: Prikaz korelacijske matrice med področji	51
Slika 5.27: Signifikantne korelacijske matrice med področji	52
Slika 5.28: Toplotni zemljevid podobnosti odgovorov v anketah	53
Slika 5.29: Toplotni zemljevid podobnosti prvih 100 anket	54
Slika 5.30: Optimalno število gruč - odgovori	55
Slika 5.31: Gruče odgovorov	56
Slika 5.32: Gruče področij	57
Slika 5.33: Gruče področij	58
Slika 5.34: Gruče področij	59
Slika 5.35: Število gruč odgovorov po čiščenju podatkov	61
Slika 5.36: Gruče odgovorov očiščenih podatkov	62
Slika 5.37: Gruče področij	63

1 Uvod

Podatkovna analitika je izjemno agilno področje, ki obsega množico metod in pristopov. Kot osnovo sicer uporablja statistične metode. Statistične metode in raziskave pomagajo, da smiselno razvrstimo in pravilno interpretiramo masovne podatke v poslovnih primerih uporabe. Podatkovno analitiko v tem smislu opazujemo in definiramo z dveh vidikov: s tehničnega in s poslovnega. Tehnični vidik zajema načine in instrumente, s katerimi podatke zberemo in obdelamo. Za ta del so pomembna znanja s področja informacijskih sistemov in programskega inženirstva. Poslovni vidik pa predstavlja interpretacijo rezultatov in realizacijo aktivnosti za doseganje poslovnih ciljev. V raziskavi bom obravnaval primer aplikacije KAMbi, ki je sicer nastala kot pomoč pri izbiri inženirskega poklica dijakom v zaključnih letnikih srednjih šol. Vendar gre tu tudi za poslovni vidik visokošolskih inštitucij, ki si na razpisih za vpis konkurirajo za te iste dijake.

Podatkovna analitika v podjetjih lahko vpliva na poslovne procese in aktivnosti zaradi izboljšanje procesa odločanja ali reševanja kompleksnih poslovnih problemov. S podatkovno analitiko podjetja lahko spremenijo poslovno strategijo, poslovne modele in poslovne cilje. Primer aktivnosti v podjetju, ki je ugotovilo, da ima pre malo kupcev v starostni skupini do 25 let, je sprememba dizajna izdelka ali kreiranje novega modela, ki bolj privlači mlajše kupce. Uporaba podatkovne analitike je proces, ki lahko vključuje različne oddelke in strokovnjake. Pojav znanosti o podatkih (ang. data science) in poklica *podatkovni znanstvenik* (ang. data scientist) ni le posledica masovnih podatkov (ang. big data) in s tem povezanimi tehnologijami. Predpostavimo lahko, da so k temu pripomogle zahteve poslovnih uporabnikov, ki so želeli podatke pretvoriti v poslovno vrednost in so opazili, da ima tudi statistična analiza še precej potenciala. Ključna razlika med kompetencami, ki jih potrebujeta analitik in podatkovni znanstvenik, je v tem, da se znanstveniki ukvarjajo z večimi podatkovnimi zbirkami, masovnimi podatki in zato potrebujejo širša znanja s področja računalništva, informatike, matematike, statistike in poslovanja.

V tej diplomski nalogi bomo zato demonstrirali interakcijo navedenih disciplin na primeru spletnne aplikacije, razvite z orodjem za malo-kodno programiranje Oracle Application Express ter analize zajetih podatkov z jezikom R.

2 Metodologija

2.1 Definicija problema

S spletno aplikacijo KAMbi, ki je namenjena dijakom zaključnih letnikov srednjih šol kot pomoč pri odločanju za inženirske poklice, je bilo zajeto 1750 anket, pri čemer vsaka anketa vsebovala 78 vprašanj. Čeprav je osnovni namen aplikacije posamezniku ponuditi rangiran seznam najprimernejših visokošolskih in višješolskih programov glede na samooceno lastnosti, doslej ni bila izvedena analiza vseh zajetih podatkov. Niso znane osnovne statistične mere posameznih odgovorov, korelacije med odgovori, morebitne skupine podobnih odgovorov niti obstoj slabih podatkov. Poseben problem predstavlja pretvorba zapisov iz baze podatkov v obliko, ki je primerna za statistično obdelavo.

2.2 Definicija ciljev

Glavni cilji naloge so:

- preštudirati literaturo o jeziku R in knjižnicah za obdelavo podatkov
- proučiti spletno aplikacijo KAMbi, razvojno orodje Oracle Application Express in strukturo baze podatkov omenjene aplikacije
- izvesti analizo podatkov, ki bo obsegala opisno statistiko s histogrami odgovorov in področij, korelacijsko analizo, analizo gruč in čiščenje podatkov

2.3 Uporabljene metode in orodja

Glavne metode: programiranje v jeziku R, opisna statistika, korelacijska analiza, analiza gruč

Orodja: R, RStudio, Oracle Application Express, L^AT_EX

3 Predstavitev orodij

3.1 Aplikacija KAMbi

KAMbi je spletna aplikacija, namenjena podpori odločanju za inženirski poklic oz. izbiro visokošolskega ali višješolskega programa, ki daje ustrezno izobrazbo. Vprašalnik, ki ga kandidat rešuje, vsebuje dva dela: prvi del (39 vprašanj tipa da/ne) se nanaša na naravo delo in drugi del (39 vprašanj, 5-stopenjska lestvica pogostosti obnašanja) se nanaša na samooceno kompetenc.

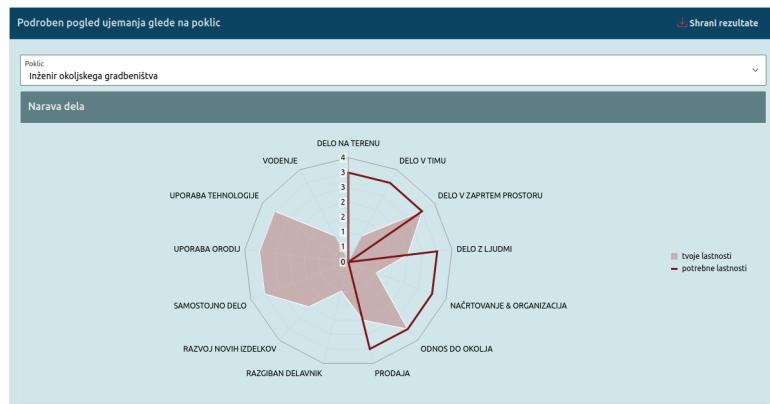


Slika 3.1: Prva skupina vprašanj aplikacije KAMbi

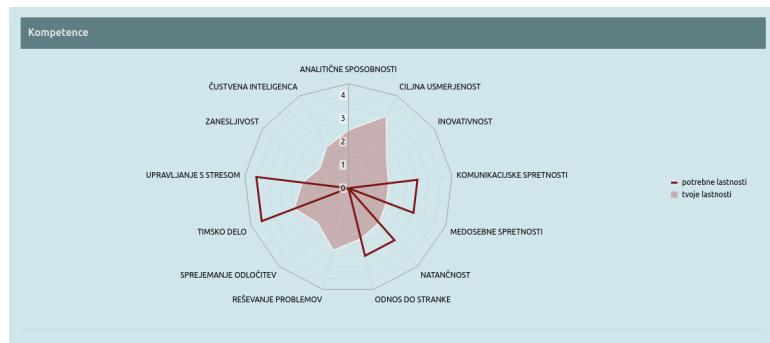


Slika 3.2: Druga skupina vprašanj aplikacije KAMbi

Na podlagi odgovorov se izračuna stopnja skladnosti kandidatovih lastnosti z zahtevanimi lastnostmi 23 različnih inženirskih poklicev. Poklici so rangirani po stopnji skladnosti, vendar kandidat lahko pogleda radarska grafa ujemanja preferenc narave dela in kompetenc za kateregakoli izmed 23 poklicev v bazi podatkov.



Slika 3.3: KAMbi - ujemanje preferenc o naravi dela kandidata z izbranim poklicem



Slika 3.4: KAMbi - ujemanje kompetenc kandidata z izbranim poklicem

Slika 3.3 prikazuje ujemanje kandidatovih želja in izbranega poklica glede na naravo dela, na Sliki 3.4 pa so prikazane kandidatove samoocene kompetenc in zahtevane kompetence izbranega poklica. Poleg grafičnih prikazov kandidat lahko generira poročilo o stopnji ujemanja samo-ocenjenih lastnosti z zahtevanimi lastnostmi poklicev, ki so shranjeni v bazi podatkov. Poročilo vsebuje tudi unikaten žeton za ponoven ogled rezultatov.



BOMO!

Rezultat vprašalnika

Najbolj ustrezena izobrazba

Fizik	85 %
-------	------

V čem bi bil še dober

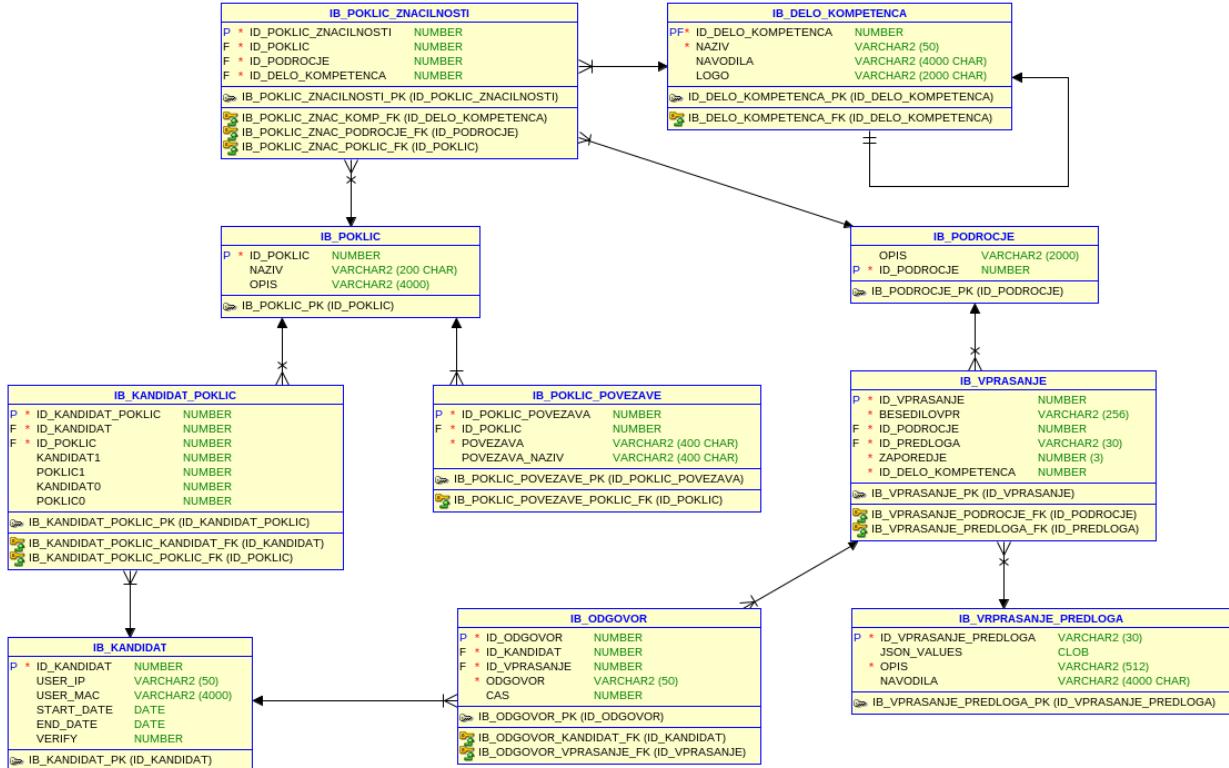
Inženir laboratorijske biomedicine	79 %
Inženir kemijskega inženirstva	78 %
Organizator informatik	75 %
Inženir tehniskega varstva okolja	74 %
Inženir računalništva in informacijskih tehnologij	72 %
Inženir metalurgije	71 %
Inženir informatike in tehnologij komuniciranja	71 %
Inženir računalništva in matematike	71 %
Inženir računalništva in informatike	70 %
Inženir strojništva	69 %
Inženir radiološke tehnologije	69 %
Sanitarni inženir	68 %
Inženir tehniške varnosti	66 %
Inženir tehnologije prometa	64 %
Gospodarski inženir	63 %
Inženir elektrotehnikе	63 %
Inženir mehatronike	62 %
Inženir telekomunikacij	62 %
Inženir multimedije	62 %
Organizator poslovnih sistemov	61 %
Inženir okoljskega gradbeništva	53 %
Inženir gradbeništva	52 %

Za ponovni ogled rezultatov lahko uporabite spodnjo povezavo :

https://apex.oracle.com//pls/apex/kpo2019/r/inzenirji-bomo/rezultat?p2_token=256822874493115438027897865366898259006

3.2 Baza podatkov za aplikacijo KAMbi

Baza podatkov obsega 10 tabel, relacijski model pa je prikazan na sliki 3.5.



Slika 3.5: Relacijski model podatkovne baze KAMbi aplikacije

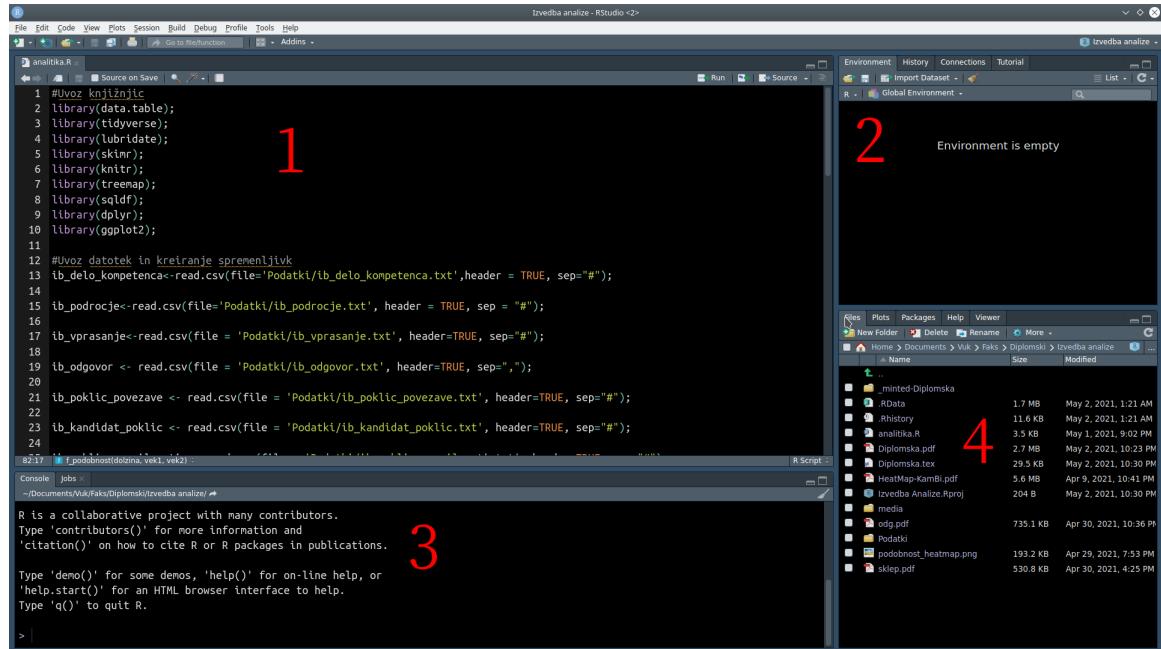


Dodatek 1: SQL skripta za kreiranje tabel, sekvenc in sprožilcev KAMbi baze

3.3 Programski jezik R in RStudio

Jezik R, ki je odprtokodna programska oprema, je trenutno eno najpomembnejših orodij za podatkovno analizo. RStudio je integrirano razvojno okolje za pisanje skript v jeziku R, izvajanje analiz in predstavitev rezultatov. Glavni repozitorij knjižnic oz. paketov je CRAN (The Comprehensive R Archive Network). Ta poleg statističnih metod obsega še veliko specifičnih rešitev od grafike, generiranja poročil, nevronskeih mrež, strojnega učenja, sistemsko dinamike in podobno.

Integrirano razvojno okolje RStudio si uporabnik lahko prilagodi. Na sliki 3.7 je primer prikaza štirih glavnih področij: izvorna koda, konzola z rezultati, delovne spremenljivke in grafika.



Slika 3.7: IDE - R Studio

Vsako področje ima še več zavihkov v svoji kategoriji.

3.4 Najpomembnejši paketi za konkretno nalogu

Paket **tidyverse** je eden najpogosteje uporabljenih paketov za podatkovno analizo. Za delovanje potrebuje več odvisnih paketov za risanje grafov, obdelavo podatkov, oblikovanje izpisov, branje vhodnih virov in podobno: **ggplot2**, **dplyr**, **tidyr**, **readr**, **purrr**, **tibble**, **stringr**, **forcats**.

V primeru, da je nameščanje paketa **tidyverse** neuspešno, je najbolj verjeten vzrok nenameščena sistemski knjižnica. Na mojem računalniku z operacijskim sistemom Ubuntu sem dodatno namestil tri knjižnice z ukazom:

```
1 $ sudo apt install libcurl4 openssl devdev
```

Nato sem v RStudiu lahko uspešno namestil paket **tidyverse**:

```
1 install.packages("tidyverse");
```

Paket **sqldf** omogoča programerju da izvaja poizvedbe z jezikom SQL nad podatkovnimi okviri v R.

Paket **data.table** je namenjen obdelavi tabel: združevanje tabel, dodajanje, brisanje

stolpcev, odstranjevanje podvojenih vrstic, grupiranje podatkov in podobno. Paket je združljiv s podatkovnimi okvirom, ki se v terminologiji R imenuje *data.frame* (Matt Dowle, stran 4).

Paket **skimr** pomaga pri hitrem in učinkovitem kreiranju povzetka analize. Omogoča tudi uporabo funkcije *pipe()*.

Paket **dplyr** omogoča manipuliranje s podatki.

Paket **Hmisc** omogoča korelacijsko analizo. Za prikaz korelacijskih diagramov se uporablja paket **corrplot**, ki ima za oblikovanje zelo veliko število parametrov.

Paket **psych** omogoča izdelavo opisne statistike . V nalogi sta uporabljeni funkciji *stat.desc()* in *describe()*. Paket **cluster** se uporablja za analizo dokler paket **factoextra** omogoča vizualizacijo gruč.

V nalogi sem uporabljal tako konzolo programa R kot tudi RStudio.

4 Razvoj skripta za obdelavo podatkov

Pri razvoju skripta sem poskušal posamezne funkcionalnosti zapisati čim bolj pregleдno: nalaganje knjižnic, branje podatkov iz izvoženih tabel baze ali pa direktno branje iz baze podatkov, priprava delovnih spremenljivk kot so matrika odgovorov in matrika časov, funkcije za izračun dosežka anketiranca po področjih in primerjava podobnosti odgovorov, opisna statistika s histogrami odgovorov ter področij, korelacijska analiza ter analiza gruč in čiščenje podatkov. Nalaganje knjižnic:

```
1 library(tidyverse)
2 library(data.table)
3 library(skimr)
4 library(sqldf)
5 library(dplyr)
6 library(corrplot)
7 library(pastecs)
8 library(psych)
9 library(cluster)
10 library(RColorBrewer)
11 library(ggpubr)
12 library(factoextra)
13 library(fpc)
```

Branje podatkov iz datotek ali pa direktno branje iz baze podatkov je prikazano v naslednjem segmentu skripta. Podatki so bili iz baze izvoženi v datoteke, ki imajo zaglavje z imeni stolpcev, nato pa sledijo podatki z ločilom #. Če so podatki v bazi Oracle, lahko do njih dostopamo tudi direktno. Branje podatkov iz datotek:

```
1 delo_kompetenca <- read.csv(file='NoviPodatki/
  ib_delo_kompetenca.txt', header = TRUE, sep="#")
2 podrocje <- read.csv(file='NoviPodatki/ib_podrocje.txt',
  header = TRUE, sep = "#")
3 vprasanja <- read.csv(file = 'NoviPodatki/ib_vprasanje.txt',
  header=TRUE, sep="#")
4 odgovori <- read.csv(file = 'NoviPodatki/ib_odgovor.txt',
  header=TRUE, sep="#")
5 poklic_povezave <- read.csv(file = 'NoviPodatki/
  ib_poklic_povezave.txt', header=TRUE, sep="#")
```

```

6 kandidat_poklic <- read.csv(file = 'NoviPodatki/
7   ib_kandidat_poklic.txt', header=TRUE, sep="#")
8 poklic_znacilnosti <- read.csv(file = 'NoviPodatki/
9   ib_poklic_znacilnosti.txt', header=TRUE, sep="#")
10 kandidat <- read.csv(file = 'NoviPodatki/ib_kandidat.txt',
11   header=TRUE, sep="#")
12 poklic <- read.csv(file='NoviPodatki/ib_poklic.txt', header
13   = TRUE, sep = '#')

```

V okolju R zgoraj navedenimi ukazi kreiramo podatkovne okvire (npr. področje). Drugi način za pripravo podatkovnih okvirjev je branje iz baze podatkov Oracle. V tem primeru je potreno naložiti dodatne knjižnice.

Branje podatkov iz Oracle baze:

```

1 # knjiznice
2 library(rJava)
3 library(RJDBC)
4 library(sqldf)
5
6 # klic gonilnika JDBC za Oracle XE
7 driver <- JDBC("oracle.jdbc.driver.OracleDriver",
8                 classPath="/opt/rii/konektorji/ojdbc8.jar", "")
9 # vzpostavitev povezave
10 con <- dbConnect(driver, "jdbc:oracle:thin:@
11   //164.8.76.27:1521/XEPDB1", "un","pw")
12 # branje v podatkovne okvire
13 vprasanje_predloga <- dbReadTable(con, "
14   ib_vprasanje_predloga")
15 vprasanje <- dbReadTable(con, "ib_vprasanje")
16 poklic_znacilnosti <- dbReadTable(con, "
17   ib_poklic_znacilnosti")
18 poklic_povezave <- dbReadTable(con, "ib_poklic_povezave")
19 poklic <- dbReadTable(con, "ib_poklic")
20 podrocje <- dbReadTable(con, "ib_podrocje")
21 odgovor <- dbReadTable(con, "ib_odgovor")
22 odgovor_sort <- odgovor[order(odgovor$ID_KANDIDAT,
23   odgovor$ID_VPRASANJE),]
24 kandidat <- dbReadTable(con, "ib_kandidat")
25 kandidat_poklic <- dbReadTable(con, "ib_kandidat_poklic")

```

Priprava delovnih spremenljivk in podatkovnih struktur kot so polja, matrike in podatkovni okviri:

```
1 # stevilo vrstic (anketirancev), stolpcov (vprasanj) in
  podrocijs
2 nVrstic <- dim(kandidat)[1]
3 nStolpc <- dim(vprasanja)[1]
4 nVprasanj <- dim(vprasanja)[1]
5 nPodrocij <- dim(podrocje)[1]
```

Začetni in končni indeksi področij:

```
1 zac_ind_pod <- c
  (1,4,7,10,13,16,19,22,25,28,31,34,37,40,44,47,50,53,56,
2 58,62,65,68,71,74,77)
3 kon_ind_pod <- c
  (3,6,9,12,15,18,21,24,27,30,33,36,39,43,46,49,52,55,57,
4 61,64,67,70,73,76,78)
```

Poizvedba za pretvorbo tekstovnih odgovorov v numerične vrednosti iz podatkovnega okvira *odgovori* in kreiranje vektorja odgovorov:

```
1 vektor_odgovorov <- as.numeric(unlist(sqldf("select CASE
2 WHEN ODGOVOR = 'NE' THEN 0.0
3 WHEN ODGOVOR = 'NIKOLI' THEN 0.0
4 WHEN ODGOVOR = 'DA' THEN 1.0
5 WHEN ODGOVOR = 'VEDNO' THEN 1.0
6 WHEN ODGOVOR = 'NAJPOGOSTEJE' THEN 0.75
7 WHEN ODGOVOR = 'VCASIH DA/VCASIH NE' THEN 0.50
8 WHEN ODGOVOR = 'REDKO' THEN 0.25
9 END AS tocke
10 from odgovori order by ID_KANDIDAT, ID_VPRAŠANJE"))))
```

Pretvorba vektorja odgovorov v matriko odgovorov:

```
1 names(vektor_odgovorov) <- NULL
2 odg_matrika <- t(matrix(vektor_odgovorov,
3                           nrow = nStolpc,
4                           ncol = nVrstic))
```

Poleg transponiranja je pomembno, da se matrika odgovorov polni po stolpcih in ne vrsticah. Rezultat je matrika z odgovori, ki ima 1750 vrstic in 78 stolpcov. Skupno je bilo shranjenih 136500 odgovorov. Preverimo ali kak podatek manjka:

```

1 # Ali obstajajo neznane vrednosti?
2 any(is.na(odg_matrika))
> [1] FALSE

```

Kreiranje matrike časov za odgovore anketirancev:

```

1 vektor_hitrosti <- as.numeric(unlist(sqlldf("select CAS from
      odgovori order by ID_KANDIDAT, ID_VPRASANJE")))
2 names(vektor_hitrosti) <- NULL
3
4 # Vpisovanje v matriko
5 cas_matrika <- t(matrix(vektor_hitrosti,
6                           nrow = nStolpc,
7                           ncol = nVrstic))

```

Kreiranje matrike področij obsega izdelavo funkcije *f_podrocja* in polnjenje z dvojni zanki *for*:

```

1 f_podrocja <- function(odgovori, zac_pod, kon_pod) {
2   rezultat <- replicate(nPodrocij, 0)
3   for(j in 1:nPodrocij) {
4     rezultat[j] <- sum(odgovori[zac_pod[j]:kon_pod[j]])
5   }
6   return (rezultat)
7 }

8
9 # Kreiranje prazne matrike
10 podrocja_matrika <- matrix (0L, nrow = nVrstic, ncol =
11   nPodrocij)
12 colnames(podrocja_matrika) <- seq(1, nPodrocij)
13 narava_dela <- podrocje$OPIS[1:13]
14 kompetenca <- podrocje$OPIS[14:26]
15 # Polnjenje matrike področij
16 for(j in 1:nVrstic) {
17   podrocja_matrika[j,] <- f_podrocja(
18     odg_matrika[j,1:nVprasanj],
19     zac_ind_pod, kon_ind_pod)
20 }

```

Rezultat je matrika področij, ki ima 1750 vrstic in 26 stolpcev.
Opisna statistika odgovorov in histogrami pogostosti odgovorov:

```
1 # funkcija za prikaz opisne statistike
2 stat.desc(odg_matrika)
3
4 # stevilo vrstic in stolpcev v mrezi grafov
5 par(mfrow = c(3,3))
6
7 for(i in 1:9) {
8     hist(odg_matrika[,i], col="orange",
9           main=vprasanja$BESEDILOVPR[i],
10          xlab = "NE / DA", breaks=2,
11          labels = TRUE, freq=TRUE)
12 }
13
14 for(i in 10:18) {
15     hist(odg_matrika[,i], col="orange",
16           main=vprasanja$BESEDILOVPR[i],
17          xlab = "NE / DA", breaks=2,
18          labels = TRUE, freq=TRUE)
19 }
20
21 for(i in 19:27) {
22     hist(odg_matrika[,i], col="orange",
23           main=vprasanja$BESEDILOVPR[i],
24          xlab = "NE / DA", breaks=2,
25          labels = TRUE, freq=TRUE)
26 }
27
28 for(i in 28:36) {
29     hist(odg_matrika[,i], col="orange",
30           main=vprasanja$BESEDILOVPR[i],
31          xlab = "NE / DA", breaks=2,
32          labels = TRUE, freq=TRUE)
33 }
34
35 for(i in 37:39) {
```

```

36 hist(odg_matrika[,i], col="orange",
37       main=vprasanja$BESEDILOVPR[i],
38       xlab = "NE / DA", breaks=2,
39       labels = TRUE, freq=TRUE)
40 }
41
42
43 for(i in 40:48) {
44   hist(odg_matrika[,i], col="orange",
45       main=vprasanja$BESEDILOVPR[i],
46       xlab = "Nikoli / Redko / čVasihda-čVasih Ne /
47           Najpogosteje / Vedno", breaks=5,
48       labels = TRUE, freq=TRUE)
49 }
50
51 for(i in 49:57) {
52   hist(odg_matrika[,i], col="orange",
53       main=vprasanja$BESEDILOVPR[i],
54       xlab = "Nikoli / Redko / čVasihda-čVasih Ne /
55           Najpogosteje / Vedno", breaks=5,
56       labels = TRUE, freq=TRUE)
57 }
58 for(i in 58:66) {
59   hist(odg_matrika[,i], col="orange",
60       main=vprasanja$BESEDILOVPR[i],
61       xlab = "Nikoli / Redko / čVasihda-čVasih Ne /
62           Najpogosteje / Vedno", breaks=5,
63       labels = TRUE, freq=TRUE)
64 }
65 for(i in 67:75) {
66   hist(odg_matrika[,i], col="orange",
67       main=vprasanja$BESEDILOVPR[i],
68       xlab = "Nikoli / Redko / čVasihda-čVasih Ne /
           Najpogosteje / Vedno", breaks=5,
       labels = TRUE, freq=TRUE)

```

```

69 }
70
71 for(i in 76:78) {
72   hist(odg_matrika[,i], col="orange",
73       main=vprasanja$BESEDILOVPR[i],
74       xlab = "Nikoli / Redko / čVasihda-čVasih Ne /
75           Najpogosteje / Vedno", breaks=5,
76       labels = TRUE, freq=TRUE)
77 }
```

Opisna statistika časov za odgovore in histogrami pogostosti:

Zanimajo me statistični podatki za drugo spremenljivko, oziroma za čas odgovorov. Poklical bom drugo funkcijo za izračun vrednosti in isto za izdelavo histogramov.

```

1 # Opisna statisike
2 describe(cas_matrika)
3 # histogrami
4 par(mfrow = c(3,3))
5 for(i in 1:9) {
6   hist(cas_matrika[,i], col="blue",
7       main=vprasanja$BESEDILOVPR[i],
8       xlim = c(0,30),
9       breaks=max(cas_matrika[,i]),
10      labels = TRUE, freq=TRUE)
11 }
12 for(i in 10:18) {
13   hist(cas_matrika[,i], col="blue",
14       main=vprasanja$BESEDILOVPR[i],
15       xlim = c(0,30),
16       breaks=max(cas_matrika[,i]),
17       labels = TRUE, freq=TRUE)
18 }
19 for(i in 19:27) {
20   hist(cas_matrika[,i], col="blue",
21       main=vprasanja$BESEDILOVPR[i],
22       xlim = c(0,30),
23       breaks=max(cas_matrika[,i]),
```

```

24         labels = TRUE, freq=TRUE)
25     }
26     for(i in 28:36) {
27       hist(cas_matrika[,i], col="blue",
28             main=vprasanja$BESEDILOVPR[i],
29             xlim = c(0,30),
30             breaks=max(cas_matrika[,i]),
31             labels = TRUE, freq=TRUE)
32   }
33   for(i in 37:39) {
34     hist(cas_matrika[,i], col="blue",
35           main=vprasanja$BESEDILOVPR[i],
36           xlim = c(0,30),
37           breaks=max(cas_matrika[,i]),
38           labels = TRUE, freq=TRUE)
39   }
40   for(i in 40:48) {
41     hist(cas_matrika[,i], col="blue",
42           main=vprasanja$BESEDILOVPR[i],
43           xlim = c(0,30),
44           breaks=max(cas_matrika[,i]),
45           labels = TRUE, freq=TRUE)
46   }
47   for(i in 49:57) {
48     hist(cas_matrika[,i], col="blue",
49           main=vprasanja$BESEDILOVPR[i],
50           xlim = c(0,30),
51           breaks=max(cas_matrika[,i]),
52           labels = TRUE, freq=TRUE)
53   }
54
55   for(i in 58:66) {
56     hist(cas_matrika[,i], col="blue",
57           main=vprasanja$BESEDILOVPR[i],
58           xlim = c(0,30),
59           breaks=max(cas_matrika[,i]),
60           labels = TRUE, freq=TRUE)

```

```

61 }
62 for(i in 67:75) {
63   hist(cas_matrika[,i], col="blue",
64       main=vprasanja$BESEDILOVPR[i],
65       xlim = c(0,30),
66       breaks=max(cas_matrika[,i]),
67       labels = TRUE, freq=TRUE)
68 }
69 for(i in 76:78) {
70   hist(cas_matrika[,i], col="blue",
71       main=vprasanja$BESEDILOVPR[i],
72       xlim = c(0,30),
73       breaks=max(cas_matrika[,i]),
74       labels = TRUE, freq=TRUE)
75 }
```

Opisna statistika in histogrami po področjih:

```

1 # opisna statistika
2 describe(podrocja_matrika)
3 # histogrami
4 par(mfrow = c(3,3))
5 for(i in 1:9) {
6   hist(
7     podrocja_matrika[,i], col = "blue",
8     main=podrocje$OPIS[i],
9     xlim = c(0,4),
10    breaks=8,
11    labels = TRUE, freq=TRUE)
12 }
13 for(i in 10:18) {
14   hist(
15     podrocja_matrika[,i], col = "blue",
16     main=podrocje$OPIS[i],
17     xlim = c(0,4),
18     breaks=8,
19     labels = TRUE, freq=TRUE)
```

```

20 }
21 for(i in 19:26) {
22   hist(
23     podrocja_matrika[,i], col = "blue",
24     main=podrocje$OPIS[i],
25     xlim = c(0,4),
26     breaks=8,
27     labels = TRUE, freq=TRUE)
28 }
```

Funkcija podobnosti in matrika podobnosti odgovorov:

```

1 # funkcija podobnosti
2
3 f_podobnost <- function(dolz, vek1, vek2) {
4   rezultat <- 0
5   for(j in 1:dolz) {
6     if (vek1[j] == vek2[j] ) {
7       rezultat <- rezultat+1}
8   }
9   return (rezultat)
10 }
11
12 # matrika podobnosti
13 matrika_podobnosti <- matrix(0L, nrow = nVrstic, ncol =
14   nVrstic)
15 for(i in 1:nVrstic) {
16   for(j in 1:nVrstic){
17     matrika_podobnosti[i,j] = f_podobnost(nVprasanj,
18       as.vector(odg_matrika[i,1:nVprasanj]),
19       as.vector(odg_matrika[j,1:nVprasanj]))/nVprasanj
20   }}
```

Toplotni zemljevid matrike podobnosti:

```

1 # tolpotni zemljevid
2 heatmap (matrika_podobnosti ,
3           scale="column",
4           xlab="Kandidat",
5           ylab="",
6           main="Matrika podobnosti",
7           col = colorRampPalette(brewer.pal(8, "Reds"))(25)
8
9 # legenda
10 legend(x="bottomright", legend=c("min", "avg", "strong", "max"),
11         fill=colorRampPalette(brewer.pal(8, "Reds"))(4))

```

Ker je matrika podobnosti velika (1750 x 1750), je težko natančno pregledovati toplotni zemljevid. Posamezne segmente izberemo tako, da določimo spodnji in zgornji meji po vrsticah in stolpcih. Primer prikaza toplotnega zemljevida za prvih 100 anket:

```

1 # toplotni zemljevid dimenzije 100 x 100
2 heatmap(matrika_podobnosti[1:100,1:100] ,
3           scale="column",
4           xlab="Kandidat",
5           ylab="",
6           main="Matrika podobnosti 100x100",
7           col = colorRampPalette(brewer.pal(8, "Reds"))(25))
# legenda
9 legend(x="bottomright", legend=c("min", "avg", "strong", "max"),
10         fill=colorRampPalette(brewer.pal(8, "Reds"))(4))

```

Korelacijska analiza (Pearsonov koeficient) in prikazi signifikantnih povezav (odgovori in področja):

```

1 # korelacie
2 korelacija_odgovorov <- rcorr(odg_matrika)
3 korelacija_podrocij <- rcorr(podrocja_matrika)

```

```

4
5 # prikaz korelacijs in signifikantnih povezav (odgovori)
6 corrplot(korelacija_odgovorov$r, type = "upper", order = "
7   hclust",
8     tl.col = "black", tl.srt = 45,
9       main = "Korelacija med odgovori")
10 corrplot(korelacija_odgovorov$P, type = "upper", order = "
11   hclust",
12     p.mat = korelacija_odgovorov$P, sig.level = 0.01,
13       insig = "blank",
14         main = "Pomembnost povezanosti med odgovori")
15
16
17 # prikaz korelacijs in signifikantnih povezav (področja)
18 corrplot(korelacija_podrocij$r, type = "upper", order = "
19   hclust",
20     tl.col = "black", tl.srt = 45,
21       main = "Korelacija med podrocji")
22 # Prikaz pomembnosti zvezne med področji
23 corrplot(korelacija_podrocij$P, type = "upper", order = "
24   hclust",
25     p.mat = korelacija_podrocij$P, sig.level = 0.01,
26       insig = "blank",
27         main = "Pomembnost korelacijs med podrocji")

```

Analiza gruč (odgovori in področja):

```

1 # statistika vrzeli
2 gap_odgovori <- clusGap(odg_matrika, FUN = kmeans, K.max =
3   10, B=100)
4 print)(gap_odgovori, method = "firstmax")
5 # prikaz optimalnega stevila gruc
6 fviz_gap_stat(gap_odgovori)
7 # optimalno stevilo gruc je 2
8 cluster_odgovori <- kmeans(odg_matrika, 2)
9 # prikaz gruc
10 fviz_cluster(cluster_odgovori, data = odg_matrika,
11   palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
12

```

```

11         geom = "point",
12         ellipse.type = "convex",
13         ggtheme = theme_bw())
14
15 gap_podrocja <- clusGap(podrocja_matrika, FUN = kmeans, K.
16   max = 10, B = 100)
17 # Prikazem rezultate operacije
18 print(gap_podrocja, method = "firstmax")
19 fviz_gap_stat(gap_podrocja)
20 cluster_podrocja <- kmeans(podrocja_matrika, 2)
21 fviz_cluster(cluster_podrocja, data = podrocja_matrika,
22   main = "Gruce podrocij")

```

Čiščenje podatkov - odstranitev anket, ki so bile rešene izven intervala [380-624] sekund:

```

1 spodnja_meja <- 380
2 zgornja_meja <- 624
3
4 veljavne_ankete <- subset(odg_matrika , rowSums(cas_matrika
5   [,1:78]) > spodnja_meja & rowSums(cas_matrika[,1:78])<
6   zgornja_meja)
#length(veljavne_ankete)
7 gap_odgovori2 <- clusGap(veljavne_ankete, FUN = kmeans, K.
8   max = 10, B=100)
print(gap_odgovori2, method = "firstmax")
fviz_gap_stat(gap_odgovori2)
9
10 cluster_odgovori2 <- kmeans(veljavne_ankete , 2)
11 print(cluster_odgovori2)
12 fviz_cluster(cluster_odgovori2, data = veljavne_ankete ,
13   palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
14   geom = "point",
15   ellipse.type = "convex",
16   ggtheme = theme_bw())

```

5 Opisna statistika in koreacijska analiza

Opisna statistika obsega statistične mere kot so srednja vrednost, standardni odklon, porezana srednja vrednost, poprečni absolutni odklon, minimum, maksimum, razpon, asimetričnost, sploščenost in standardna napaka. Pri koreacijski analizi je bil uporabljen Pearsonov koeficient korelacije. Tako v primeru opisne statistike kot tudi pri koreacijski analizi imamo zelo veliko število spremenljivk, kar zelo zmanjšuje preglednost.

5.1 Opisna statistika za odgovore

V prvi skupini odgovorov (vprašanja 1 do 39) je bilo največ odgovorov NE pri vprašanjih 39, 28 in 3: **Nekoč bom imel svoj laboratorij & Sem 'geek' & Zanimajo me raziskave s področja varovanja okolja.** Največ odgovorov DA pa pri vprašanjih 16, 2, 23: **Ohranjam dobre odnose z ljudmi, ki so mi pomembni & Verjamem, da lahko vsak od nas naredi nekaj za boljši jutri & Z drugimi delim svoje znanje.**

V drugi skupini vprašanj je bilo največ odgovorov NIKOLI pri vprašanjih 49, 65, 77: **Rad si naredim to-do listo & Nimam težav s spoznavanjem novih ljudi & Ostajam miren, tudi, ko se stvari ne odvijajo, kot si želim.** Največ odgovorov REDKO se je pojavilo pri vprašanjih 65, 49, 50: **Nimam težav s spoznavanjem novih ljudi & Rad si naredim to-do listo & Pogosto najdem rešitve, ki jih drugi spregledajo.** Največje število odgovorov VČASIH DA/VČASIH NE je pri vprašanjih 50, 52, 40: **Pogosto najdem rešitve, ki jih drugi spregledajo & Moji predlogi vedno ustrezajo konkretnemu cilju & Pri delu si določim cilj in sestavim načrt, kako ga bom dosegel.** Največ odgovorov NAJPOGOSTEJE je bilo pri vprašanjih 70, 67, 51: **Ko rešujem težave, poiščem čim boljšo rešitev & Znam predstaviti prednosti ali slabosti določene stvari & Sposoben sem prilagoditi svojo idejo, da je sprejemljiva za vse ter največ odgovorov VEDNO** pri vprašanjih 62, 66, 63: **Spodbujam pozitivno vzdušje in dobre odnose & Z ljudmi, ki so mi pomembni, se pogosto družim & Pri pogovoru se trudim, da se sogovornik počuti upoštevanega.**

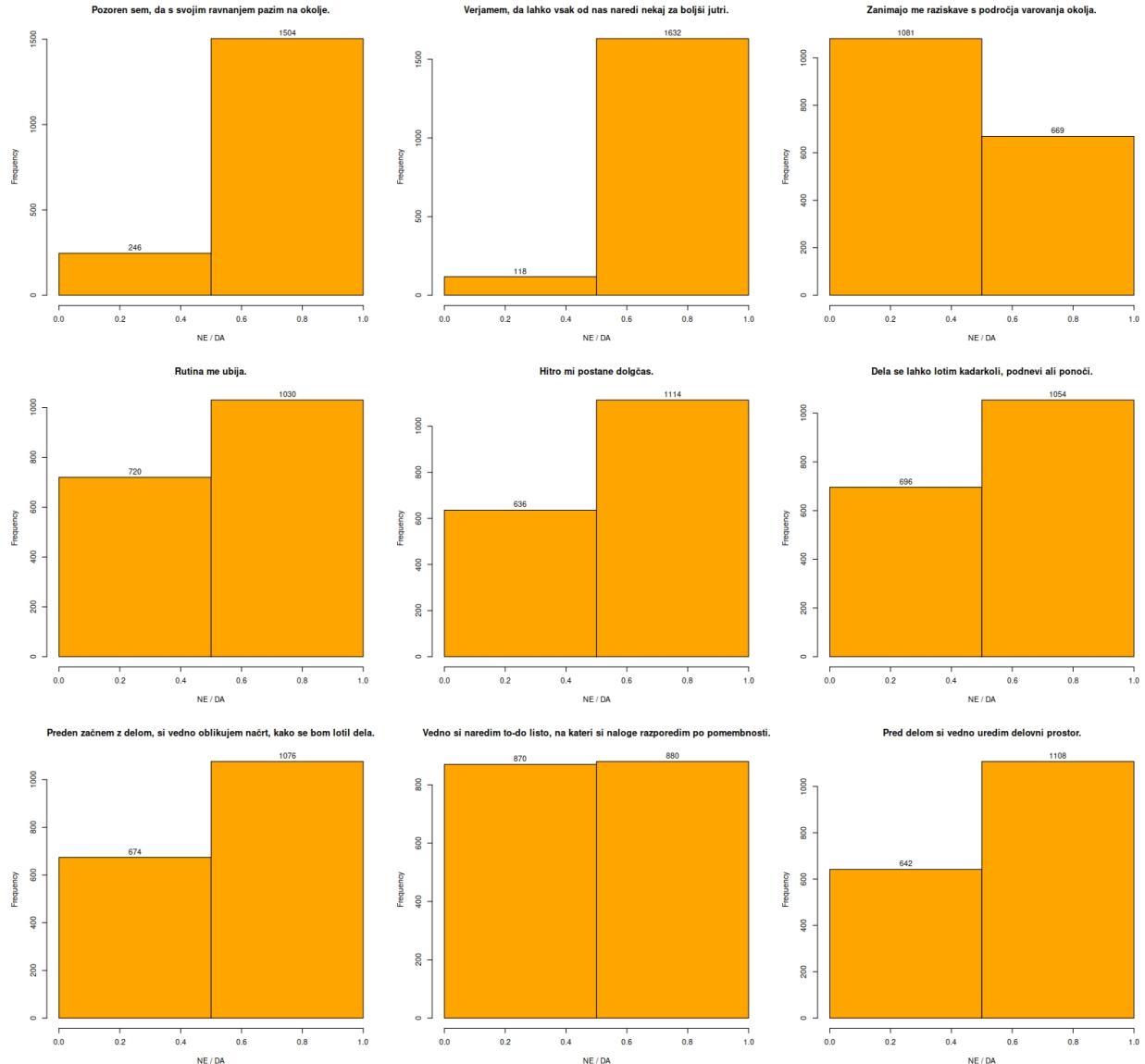
V nadaljevanju sledi prikaz frekvenc odgovorov v tabelarni oblikih in histogramih.

	vars	n	mean	sd	median	trimmed	mad	min	max	range
1	X1	1	1750	0.86	0.35	1.00	0.95	0.00	0	1
2	X2	2	1750	0.93	0.25	1.00	1.00	0.00	0	1
3	X3	3	1750	0.38	0.49	0.00	0.35	0.00	0	1
4	X4	4	1750	0.59	0.49	1.00	0.61	0.00	0	1
5	X5	5	1750	0.64	0.48	1.00	0.67	0.00	0	1
6	X6	6	1750	0.60	0.49	1.00	0.63	0.00	0	1
7	X7	7	1750	0.61	0.49	1.00	0.64	0.00	0	1
8	X8	8	1750	0.50	0.50	1.00	0.50	0.00	0	1
9	X9	9	1750	0.63	0.48	1.00	0.67	0.00	0	1
10	X10	10	1750	0.89	0.31	1.00	0.99	0.00	0	1
11	X11	11	1750	0.59	0.49	1.00	0.62	0.00	0	1
12	X12	12	1750	0.83	0.37	1.00	0.91	0.00	0	1
13	X13	13	1750	0.84	0.37	1.00	0.92	0.00	0	1
14	X14	14	1750	0.74	0.44	1.00	0.80	0.00	0	1
15	X15	15	1750	0.85	0.36	1.00	0.94	0.00	0	1
16	X16	16	1750	0.96	0.20	1.00	1.00	0.00	0	1
17	X17	17	1750	0.81	0.39	1.00	0.88	0.00	0	1
18	X18	18	1750	0.89	0.31	1.00	0.99	0.00	0	1
19	X19	19	1750	0.60	0.49	1.00	0.62	0.00	0	1
20	X20	20	1750	0.55	0.50	1.00	0.56	0.00	0	1
21	X21	21	1750	0.76	0.43	1.00	0.83	0.00	0	1
22	X22	22	1750	0.53	0.50	1.00	0.54	0.00	0	1
23	X23	23	1750	0.90	0.30	1.00	1.00	0.00	0	1
24	X24	24	1750	0.80	0.40	1.00	0.88	0.00	0	1
25	X25	25	1750	0.70	0.46	1.00	0.75	0.00	0	1
26	X26	26	1750	0.72	0.45	1.00	0.77	0.00	0	1
27	X27	27	1750	0.62	0.49	1.00	0.65	0.00	0	1
28	X28	28	1750	0.34	0.48	0.00	0.30	0.00	0	1
29	X29	29	1750	0.50	0.50	1.00	0.50	0.00	0	1
30	X30	30	1750	0.72	0.45	1.00	0.77	0.00	0	1
31	X31	31	1750	0.80	0.40	1.00	0.88	0.00	0	1
32	X32	32	1750	0.42	0.49	0.00	0.40	0.00	0	1
33	X33	33	1750	0.56	0.50	1.00	0.57	0.00	0	1
34	X34	34	1750	0.69	0.46	1.00	0.74	0.00	0	1
35	X35	35	1750	0.73	0.44	1.00	0.79	0.00	0	1
36	X36	36	1750	0.68	0.47	1.00	0.72	0.00	0	1

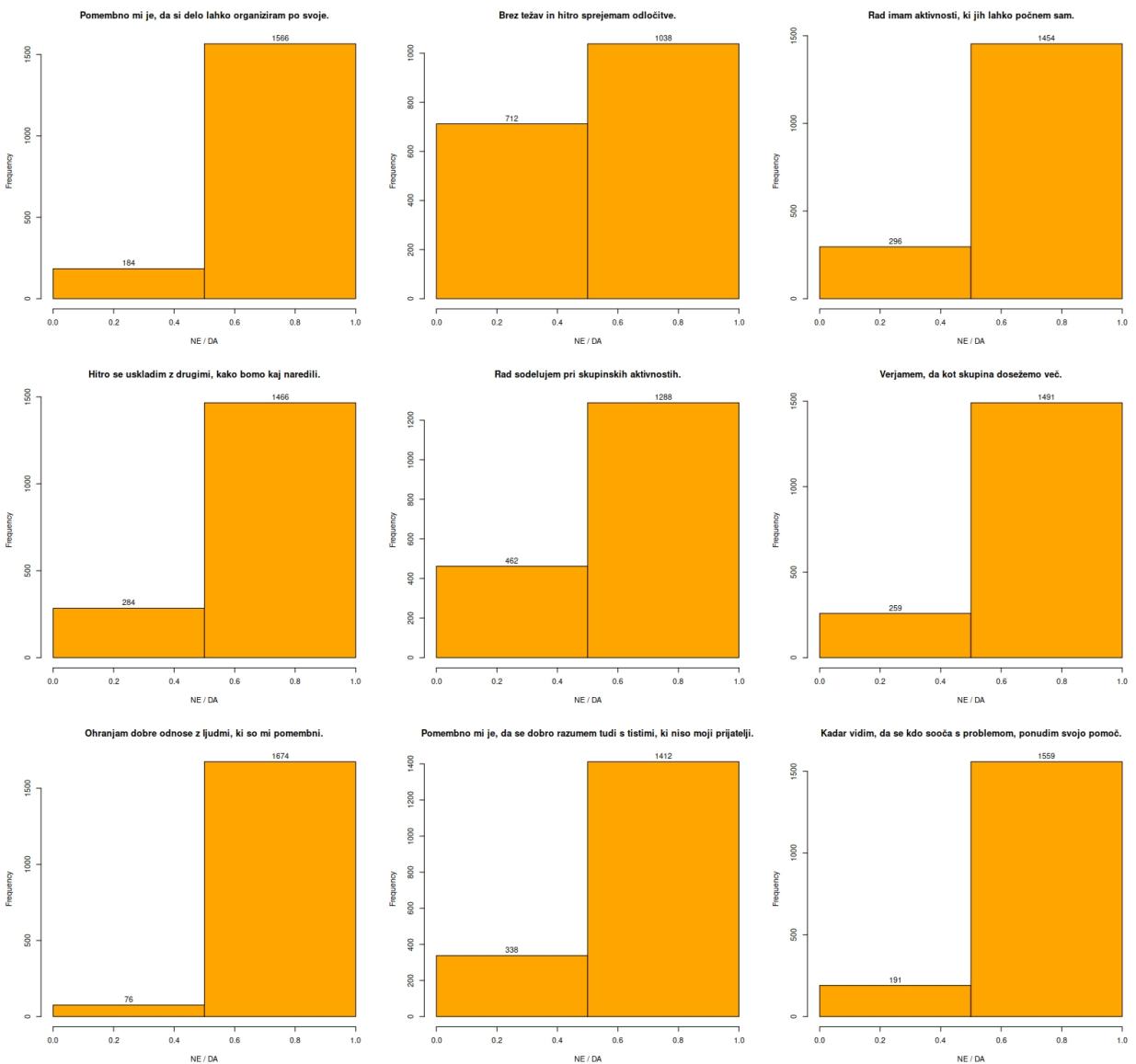
38	X37	37	1750	0.59	0.49	1.00	0.62	0.00	0	1	1
39	X38	38	1750	0.42	0.49	0.00	0.40	0.00	0	1	1
40	X39	39	1750	0.25	0.43	0.00	0.19	0.00	0	1	1
41	X40	40	1750	0.66	0.24	0.75	0.68	0.37	0	1	1
42	X41	41	1750	0.71	0.25	0.75	0.74	0.37	0	1	1
43	X42	42	1750	0.71	0.26	0.75	0.73	0.37	0	1	1
44	X43	43	1750	0.81	0.21	0.75	0.83	0.37	0	1	1
45	X44	44	1750	0.78	0.23	0.75	0.81	0.37	0	1	1
46	X45	45	1750	0.78	0.22	0.75	0.80	0.37	0	1	1
47	X46	46	1750	0.72	0.24	0.75	0.74	0.37	0	1	1
48	X47	47	1750	0.78	0.22	0.75	0.81	0.37	0	1	1
49	X48	48	1750	0.70	0.24	0.75	0.72	0.37	0	1	1
50	X49	49	1750	0.62	0.32	0.75	0.64	0.37	0	1	1
51	X50	50	1750	0.64	0.24	0.75	0.65	0.37	0	1	1
52	X51	51	1750	0.72	0.22	0.75	0.73	0.37	0	1	1
53	X52	52	1750	0.67	0.22	0.75	0.68	0.37	0	1	1
54	X53	53	1750	0.72	0.24	0.75	0.74	0.37	0	1	1
55	X54	54	1750	0.72	0.25	0.75	0.75	0.37	0	1	1
56	X55	55	1750	0.73	0.24	0.75	0.76	0.37	0	1	1
57	X56	56	1750	0.78	0.22	0.75	0.81	0.37	0	1	1
58	X57	57	1750	0.80	0.22	0.75	0.83	0.37	0	1	1
59	X58	58	1750	0.75	0.23	0.75	0.77	0.37	0	1	1
60	X59	59	1750	0.75	0.24	0.75	0.79	0.37	0	1	1
61	X60	60	1750	0.79	0.23	0.75	0.82	0.37	0	1	1
62	X61	61	1750	0.79	0.23	0.75	0.82	0.37	0	1	1
63	X62	62	1750	0.84	0.21	1.00	0.88	0.00	0	1	1
64	X63	63	1750	0.82	0.21	0.75	0.85	0.37	0	1	1
65	X64	64	1750	0.78	0.23	0.75	0.81	0.37	0	1	1
66	X65	65	1750	0.66	0.32	0.75	0.69	0.37	0	1	1
67	X66	66	1750	0.81	0.23	0.75	0.85	0.37	0	1	1
68	X67	67	1750	0.77	0.21	0.75	0.79	0.37	0	1	1
69	X68	68	1750	0.74	0.21	0.75	0.76	0.37	0	1	1
70	X69	69	1750	0.68	0.25	0.75	0.70	0.37	0	1	1
71	X70	70	1750	0.79	0.20	0.75	0.82	0.37	0	1	1
72	X71	71	1750	0.71	0.26	0.75	0.74	0.37	0	1	1
73	X72	72	1750	0.77	0.23	0.75	0.79	0.37	0	1	1
74	X73	73	1750	0.72	0.23	0.75	0.74	0.37	0	1	1

75	X74	74	1750	0.76	0.25	0.75	0.79	0.37	0	1	1
76	X75	75	1750	0.74	0.25	0.75	0.78	0.37	0	1	1
77	X76	76	1750	0.77	0.23	0.75	0.80	0.37	0	1	1
78	X77	77	1750	0.67	0.26	0.75	0.69	0.37	0	1	1
79	X78	78	1750	0.75	0.27	0.75	0.79	0.37	0	1	1

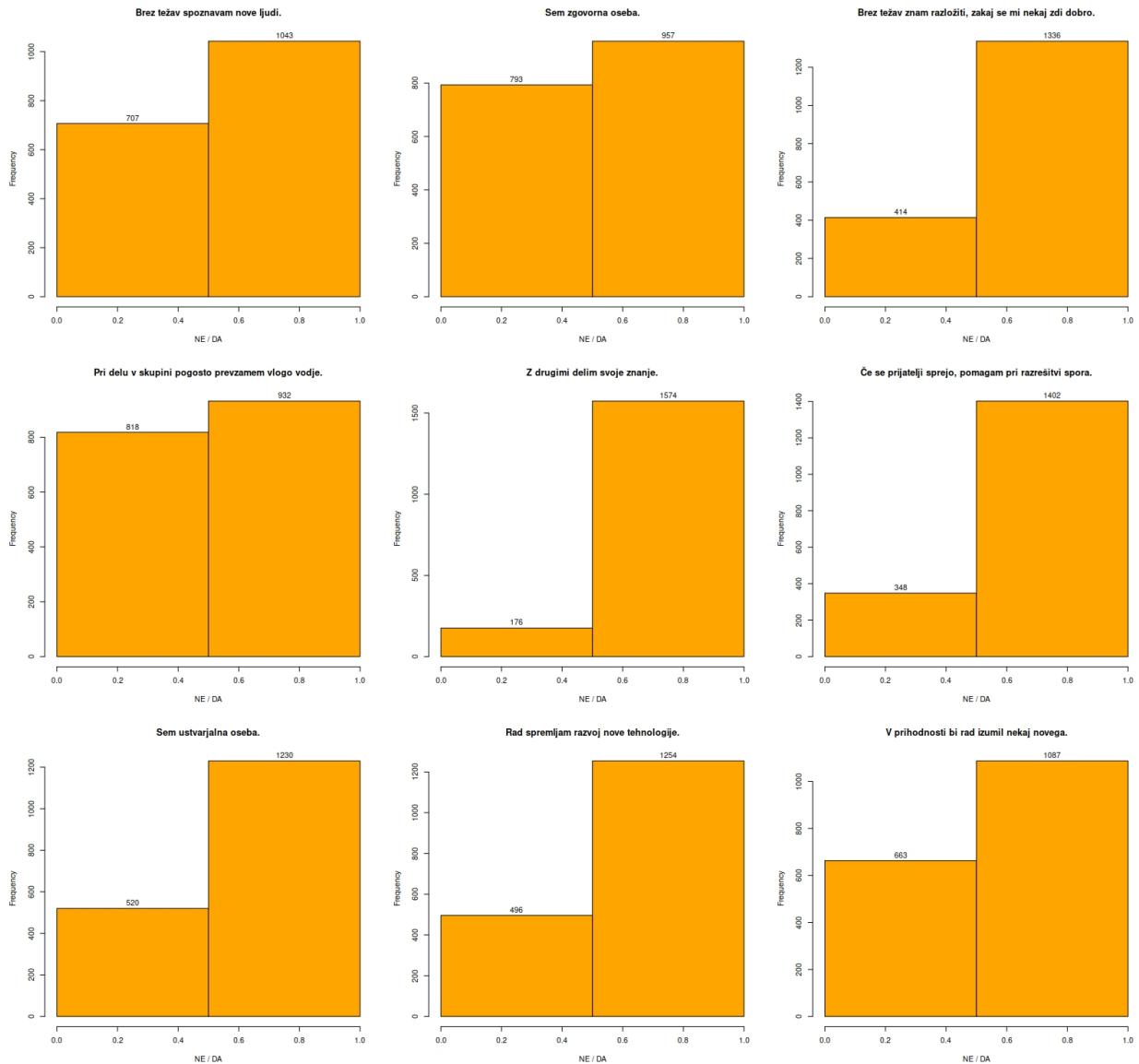
Histogrami za celoten nabor vprašanj so prikazani na naslednjih slikah.



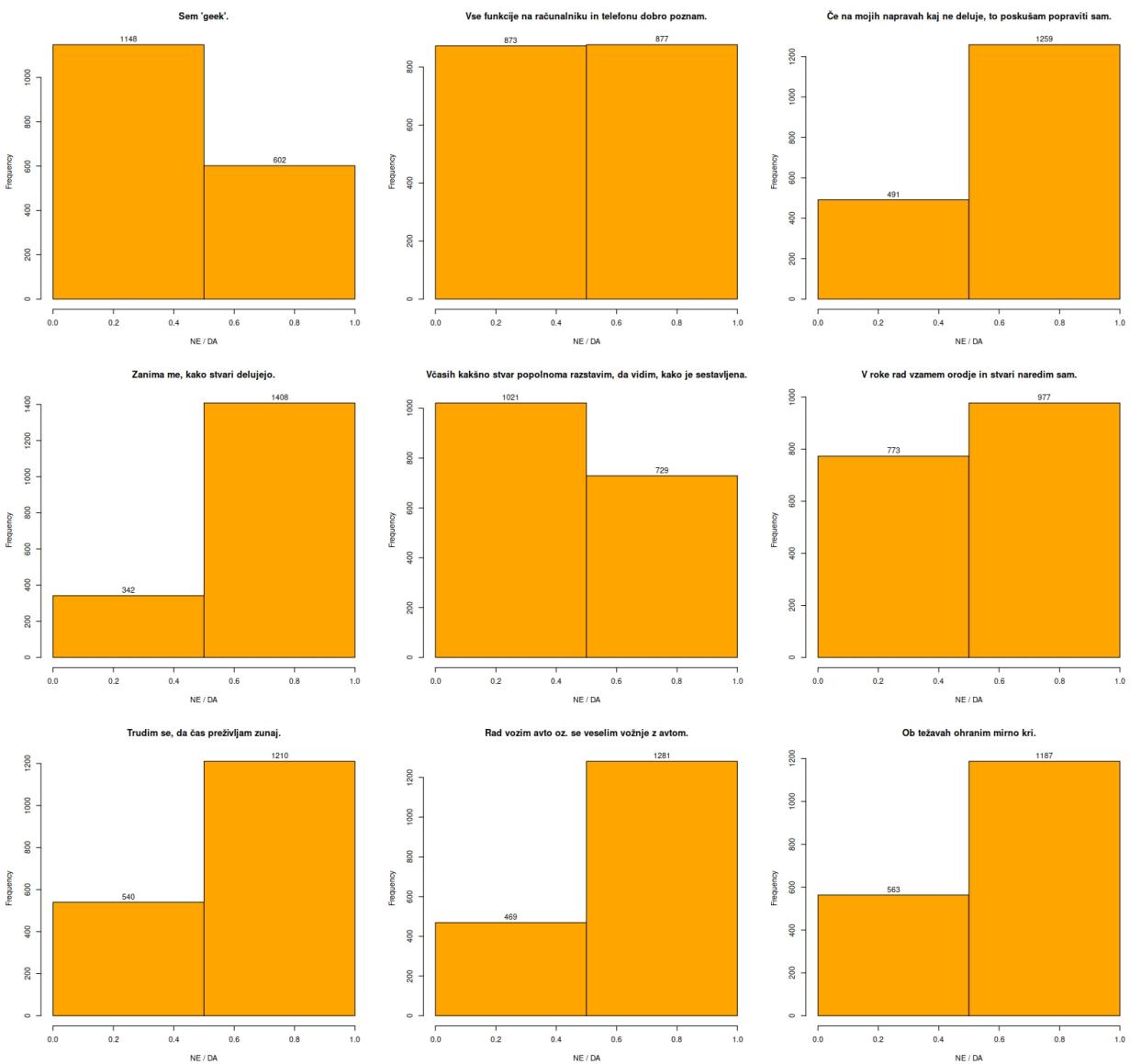
Slika 5.1: Histogram: vprašanja 1-9



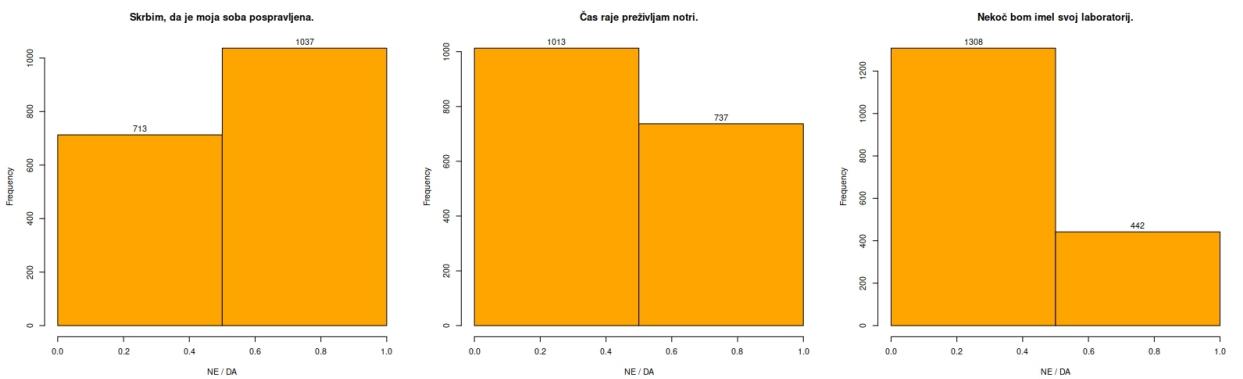
Slika 5.2: Histogram: vprašanja 10-18



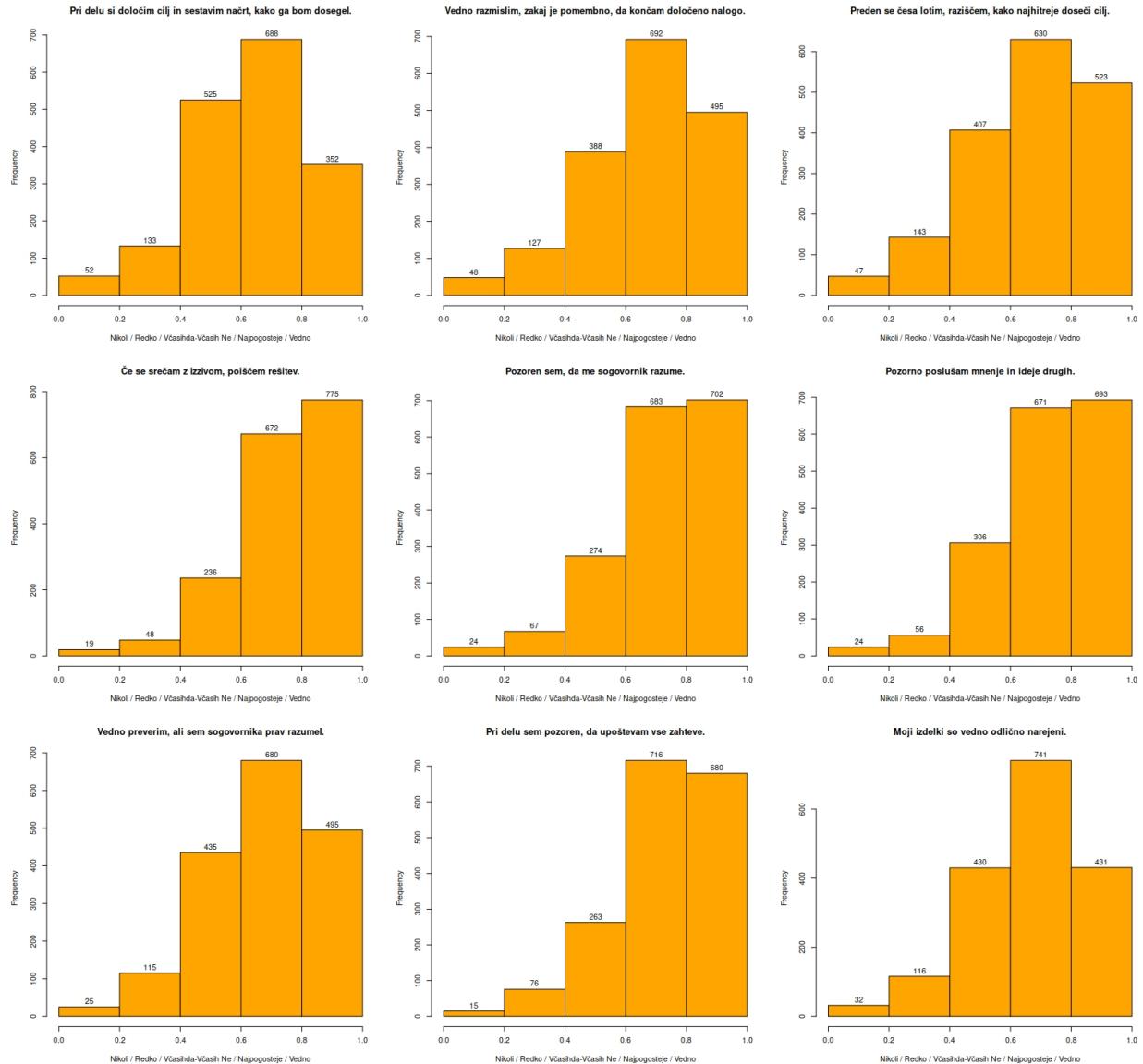
Slika 5.3: Histogram: vprašanja 19-27



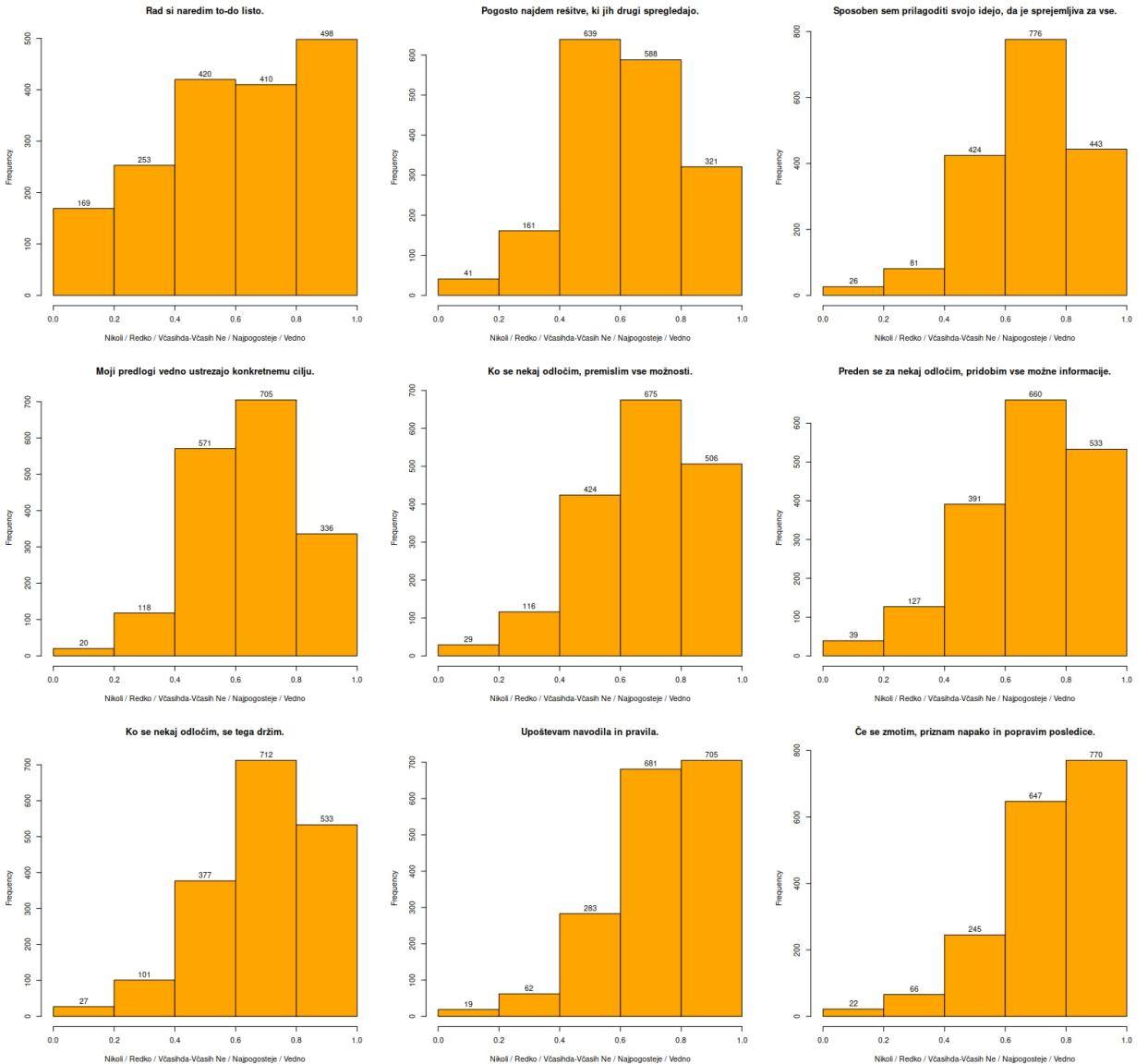
Slika 5.4: Histogram: vprašanja 28-36



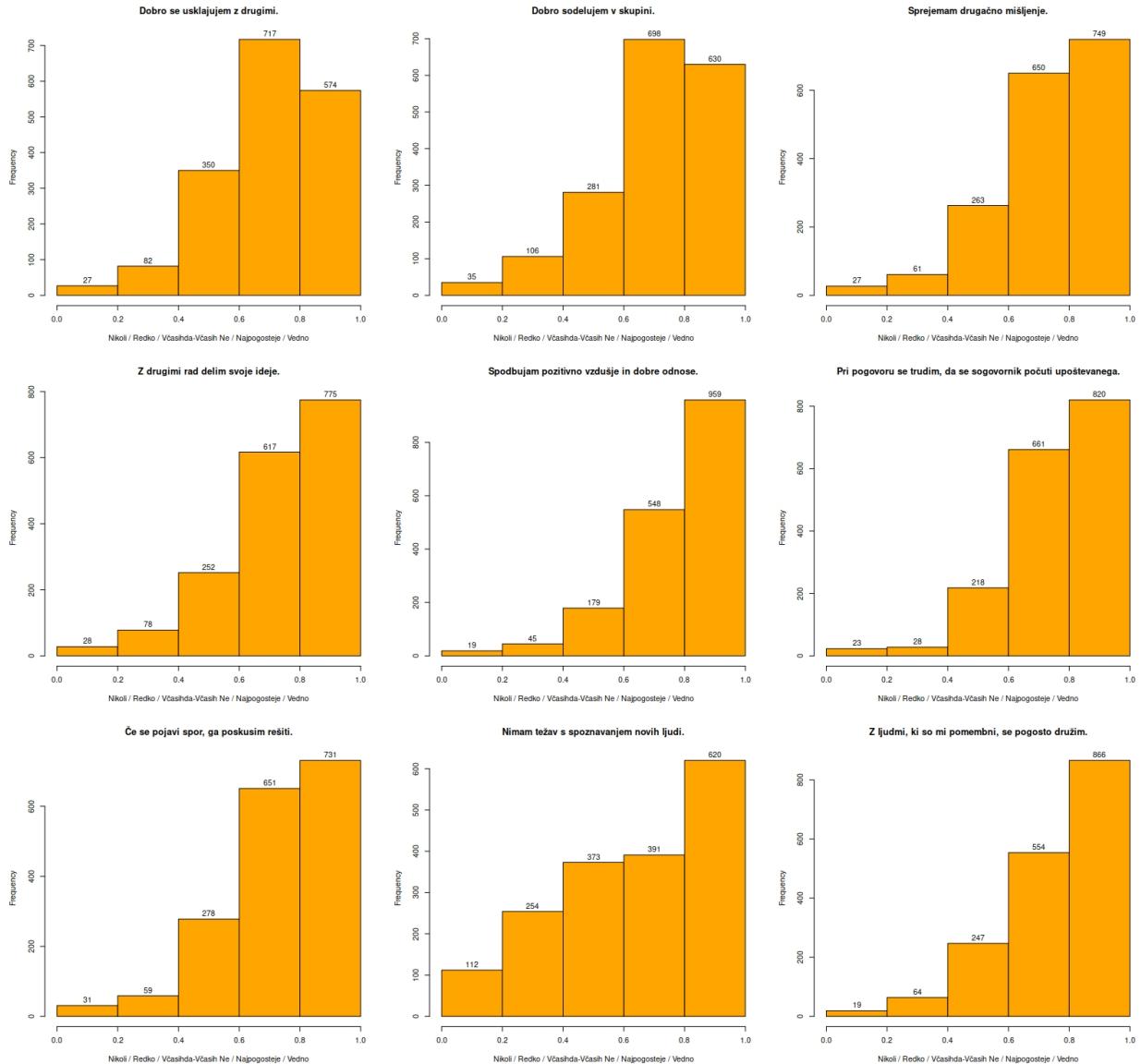
Slika 5.5: Histogram: vprašanja 37-39



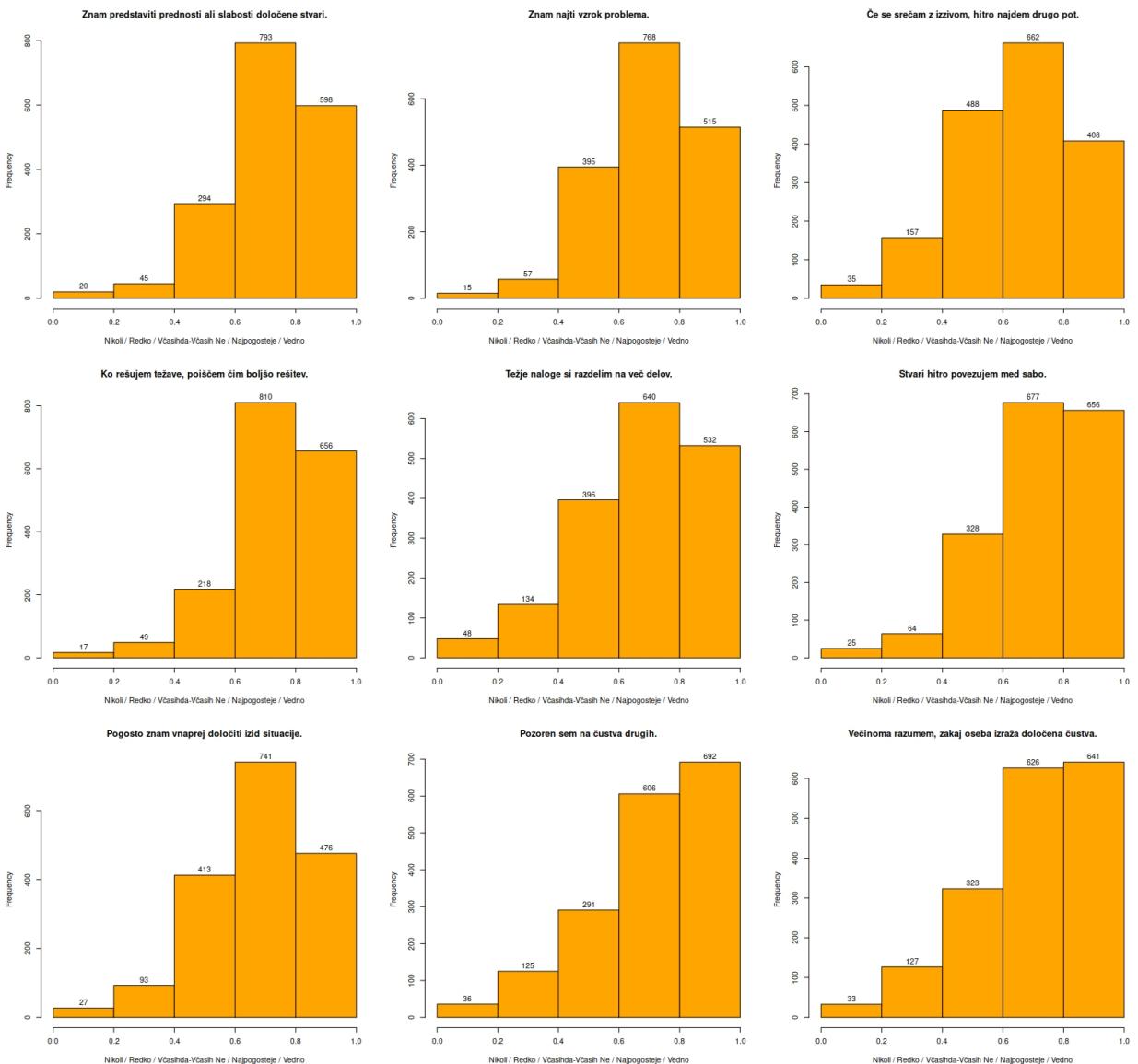
Slika 5.6: Histogram: vprašanja 40-48



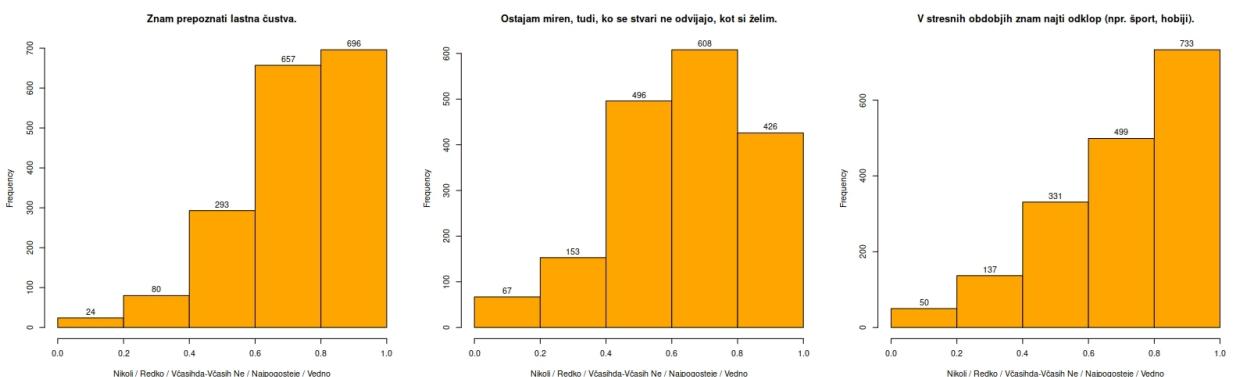
Slika 5.7: Histogram: vprašanja 49-57



Slika 5.8: Histogram: vprašanja 58-66



Slika 5.9: Histogram: vprašanja 67-75



Slika 5.10: Histogram: vprašanja 76-78

5.2 Opisna statistika za čase odgovorov

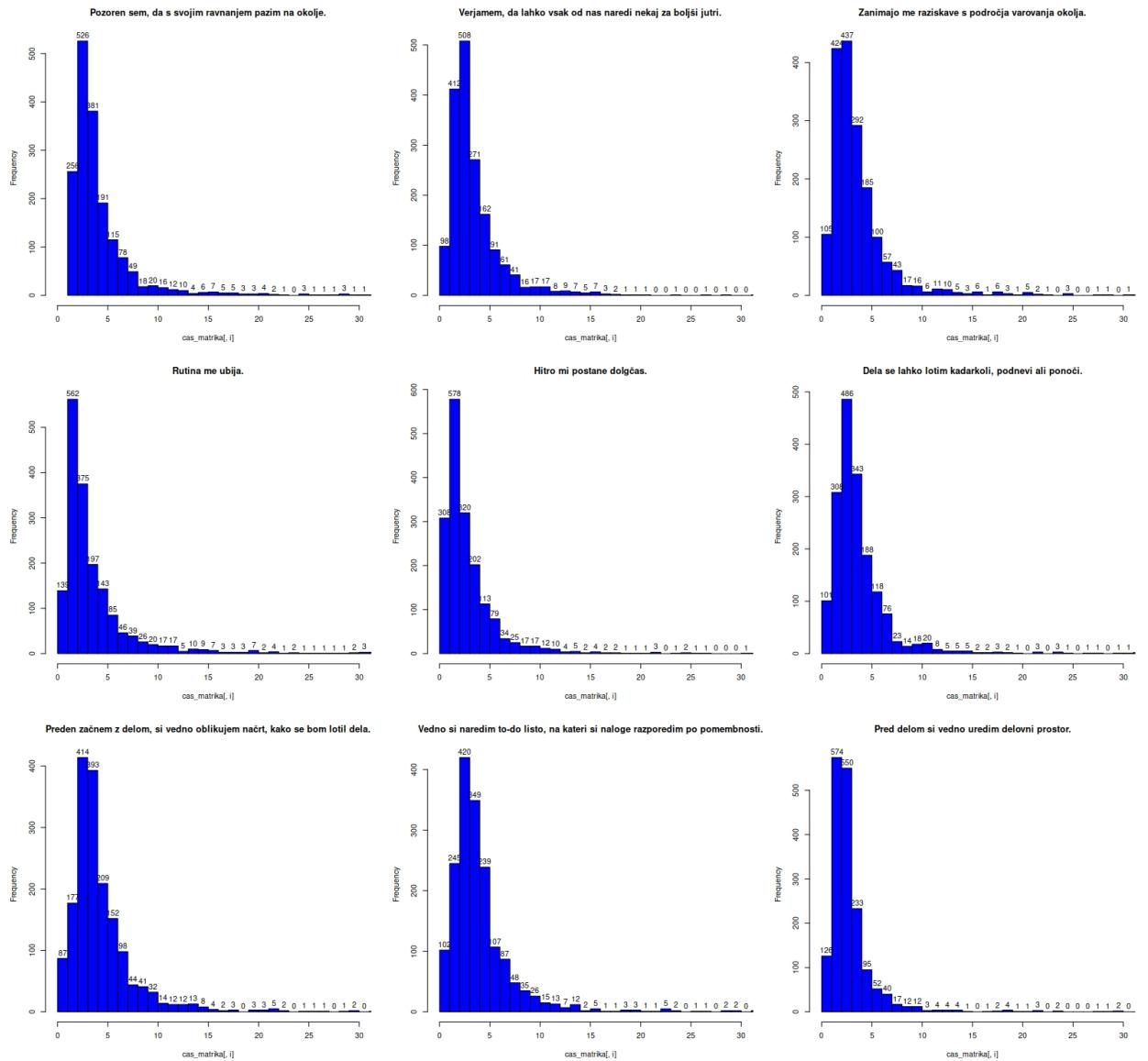
Opisna statistika časov je prikazana v spodnji tabeli.

	vars	n	mean	sd	median	trimmed	mad	min	max
1	X1	1	1750	7.89	58.39	4	4.06	1.48	1 2253
2	X2	2	1750	5.24	28.80	3	3.48	1.48	0 826
3	X3	3	1750	4.47	8.07	3	3.53	1.48	0 189
4	X4	4	1750	4.60	7.34	3	3.38	1.48	0 177
5	X5	5	1750	3.83	18.89	2	2.82	1.48	0 781
6	X6	6	1750	4.34	4.25	3	3.69	1.48	0 56
7	X7	7	1750	6.33	39.69	4	4.28	1.48	0 1597
8	X8	8	1750	10.14	171.91	4	4.04	1.48	0 6872
9	X9	9	1750	3.93	17.71	3	2.92	1.48	0 706
10	X10	10	1750	4.10	7.26	3	3.24	1.48	0 223
11	X11	11	1750	3.90	5.23	3	3.04	1.48	0 83
12	X12	12	1750	4.22	8.21	3	3.19	1.48	0 261
13	X13	13	1750	5.31	63.40	3	3.17	1.48	0 2649
14	X14	14	1750	3.14	6.35	2	2.33	1.48	0 131
15	X15	15	1750	4.15	47.20	2	2.45	1.48	0 1965
16	X16	16	1750	3.19	4.34	2	2.65	1.48	0 105
17	X17	17	1750	8.05	123.61	4	4.05	1.48	0 5161
18	X18	18	1750	4.38	7.71	3	3.50	1.48	-18 195
19	X19	19	1750	3.02	9.11	2	2.32	1.48	0 340
20	X20	20	1750	2.21	2.46	2	1.77	1.48	0 28
21	X21	21	1750	5.35	15.50	4	4.05	1.48	0 537
22	X22	22	1750	3.76	4.61	3	3.02	1.48	0 77
23	X23	23	1750	3.06	10.10	2	2.33	1.48	0 398
24	X24	24	1750	5.11	33.79	3	3.38	1.48	0 1272
25	X25	25	1750	3.78	40.90	2	2.12	1.48	0 1695
26	X26	26	1750	3.31	13.66	2	2.56	1.48	0 560
27	X27	27	1750	2.94	4.36	2	2.29	1.48	0 73
28	X28	28	1750	5.06	19.43	2	2.30	1.48	0 580
29	X29	29	1750	3.48	3.59	3	2.95	1.48	0 55
30	X30	30	1750	3.50	3.36	3	3.08	1.48	0 83
31	X31	31	1750	4.01	56.51	2	2.12	1.48	0 2357
32	X32	32	1750	4.52	12.43	3	3.37	1.48	0 356
33	X33	33	1750	4.32	31.04	3	2.82	1.48	0 1276

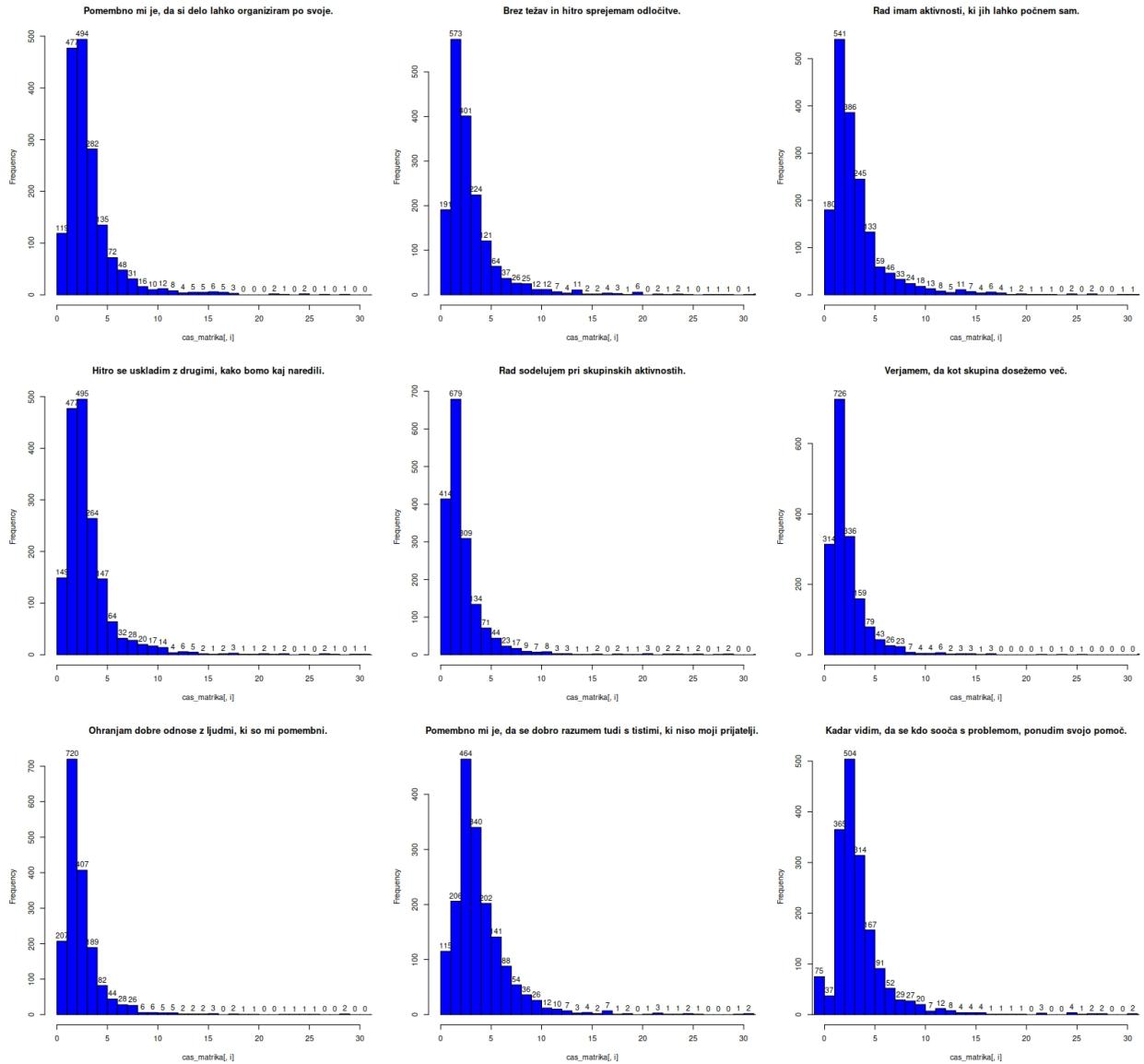
35	X34	34	1750	2.97	2.97	2	2.57	1.48	0	49
36	X35	35	1750	2.87	3.41	2	2.45	1.48	0	67
37	X36	36	1750	3.66	7.99	3	2.78	1.48	0	198
38	X37	37	1750	3.04	6.36	2	2.38	1.48	0	208
39	X38	38	1750	3.61	18.17	2	2.65	1.48	0	745
40	X39	39	1750	9.40	199.11	2	2.22	1.48	0	6750
41	X40	40	1750	11.32	88.67	6	6.30	2.97	0	3393
42	X41	41	1750	7.43	55.82	5	4.90	2.97	0	2292
43	X42	42	1750	6.00	22.82	4	4.38	1.48	0	887
44	X43	43	1750	3.99	4.91	3	3.36	1.48	0	93
45	X44	44	1750	4.64	39.85	3	3.25	1.48	0	1663
46	X45	45	1750	4.82	62.73	3	2.77	1.48	0	2622
47	X46	46	1750	3.96	4.93	3	3.40	1.48	0	87
48	X47	47	1750	4.78	6.53	4	3.84	1.48	0	170
49	X48	48	1750	3.32	6.13	3	2.80	1.48	0	204
50	X49	49	1750	3.08	4.98	2	2.52	1.48	0	116
51	X50	50	1750	5.85	38.35	4	4.01	1.48	0	1550
52	X51	51	1750	5.87	18.78	4	4.29	1.48	0	531
53	X52	52	1750	8.88	83.18	3	3.81	1.48	0	2315
54	X53	53	1750	5.73	32.38	4	3.74	1.48	0	1049
55	X54	54	1750	4.62	10.31	3	3.62	1.48	0	358
56	X55	55	1750	3.66	6.81	3	3.06	1.48	0	242
57	X56	56	1750	3.06	6.44	2	2.40	1.48	0	180
58	X57	57	1750	4.05	19.19	3	2.97	1.48	0	772
59	X58	58	1750	2.97	3.90	2	2.42	1.48	0	73
60	X59	59	1750	2.06	4.00	2	1.65	1.48	0	115
61	X60	60	1750	2.64	3.03	2	2.22	1.48	0	53
62	X61	61	1750	2.69	2.54	2	2.34	1.48	0	45
63	X62	62	1750	3.81	11.35	2	2.74	1.48	0	379
64	X63	63	1750	6.36	55.67	3	3.47	1.48	0	1957
65	X64	64	1750	3.22	13.32	2	2.35	1.48	0	535
66	X65	65	1750	3.51	6.12	2	2.63	1.48	0	168
67	X66	66	1750	3.94	5.80	3	3.31	1.48	0	155
68	X67	67	1750	5.32	31.34	3	3.71	1.48	0	1295
69	X68	68	1750	5.99	121.40	2	2.54	1.48	0	5079
70	X69	69	1750	4.68	7.09	4	3.80	2.97	0	213
71	X70	70	1750	8.45	133.61	3	3.34	1.48	0	5059

72	X71	71	1750	3.84	4.93	3	3.22	1.48	0	74
73	X72	72	1750	3.86	24.39	2	2.62	1.48	0	1005
74	X73	73	1750	4.45	13.52	3	3.26	1.48	0	366
75	X74	74	1750	2.60	2.65	2	2.27	1.48	0	49
76	X75	75	1750	3.94	4.88	3	3.33	1.48	0	84
77	X76	76	1750	3.57	36.89	2	2.35	1.48	0	1541
78	X77	77	1750	4.72	23.90	3	3.41	1.48	0	951
79	X78	78	1750	4.67	14.67	3	3.67	1.48	0	559

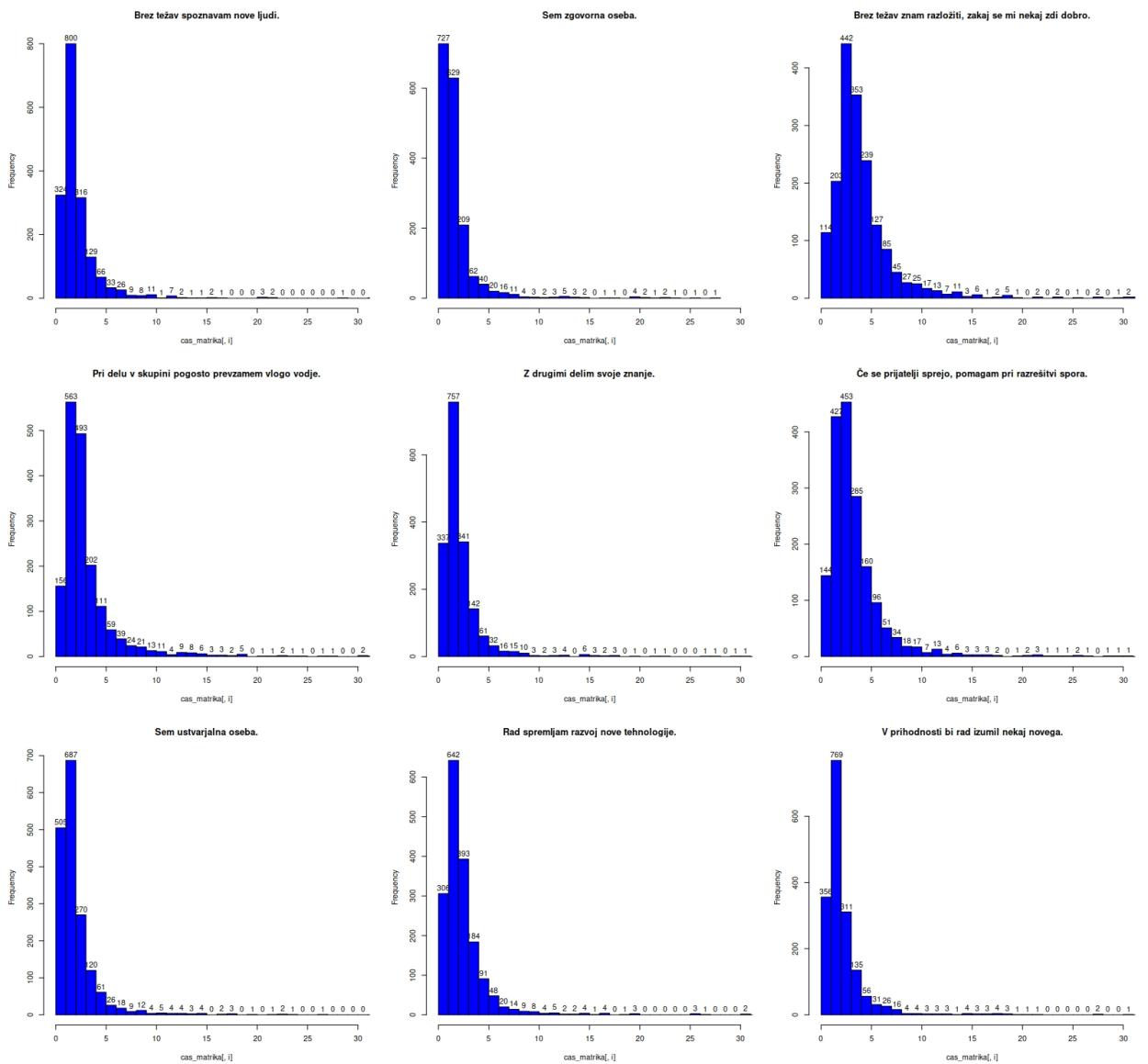
V nadaljevanju sledi prikaz frekvenc časov na histogramih.



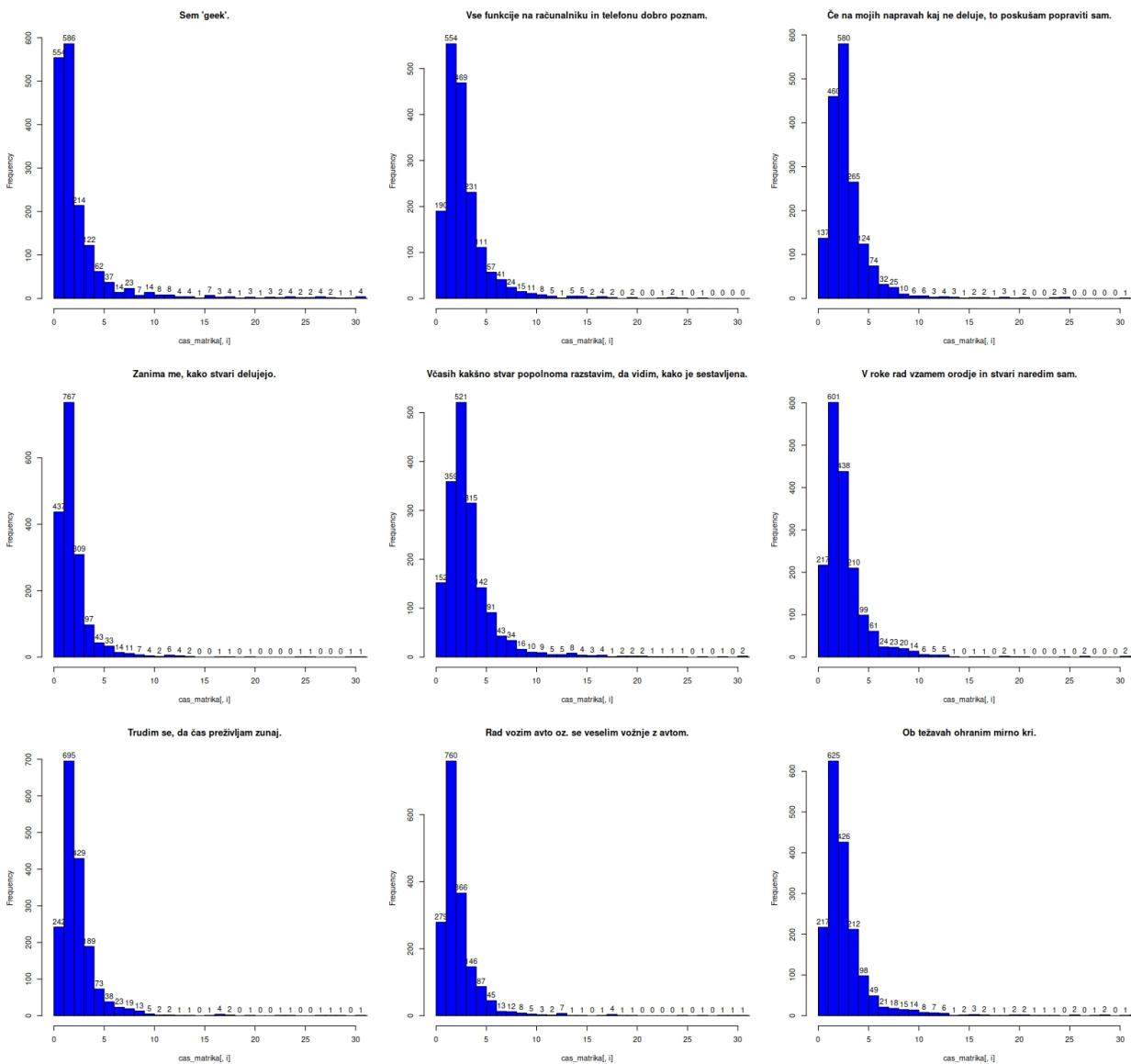
Slika 5.11: Histogram časov: vprašanja 1-9



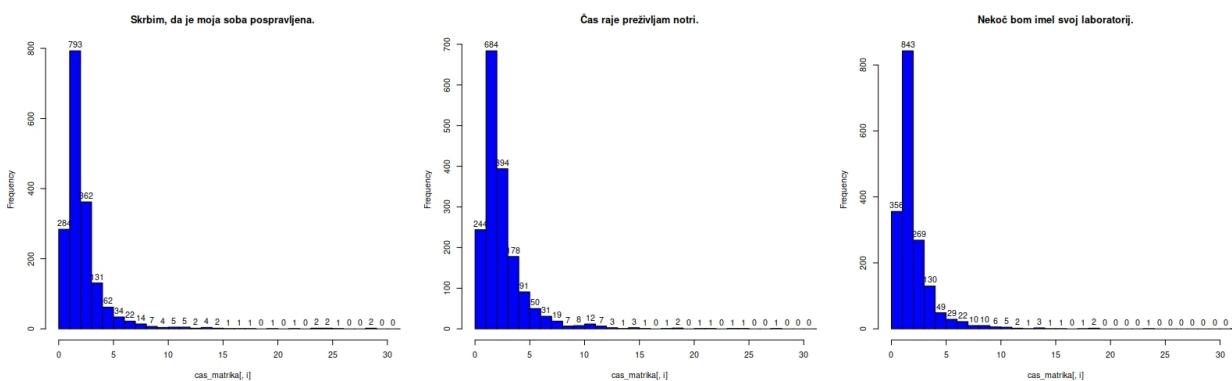
Slika 5.12: Histogram časov: vprašanja 10-18



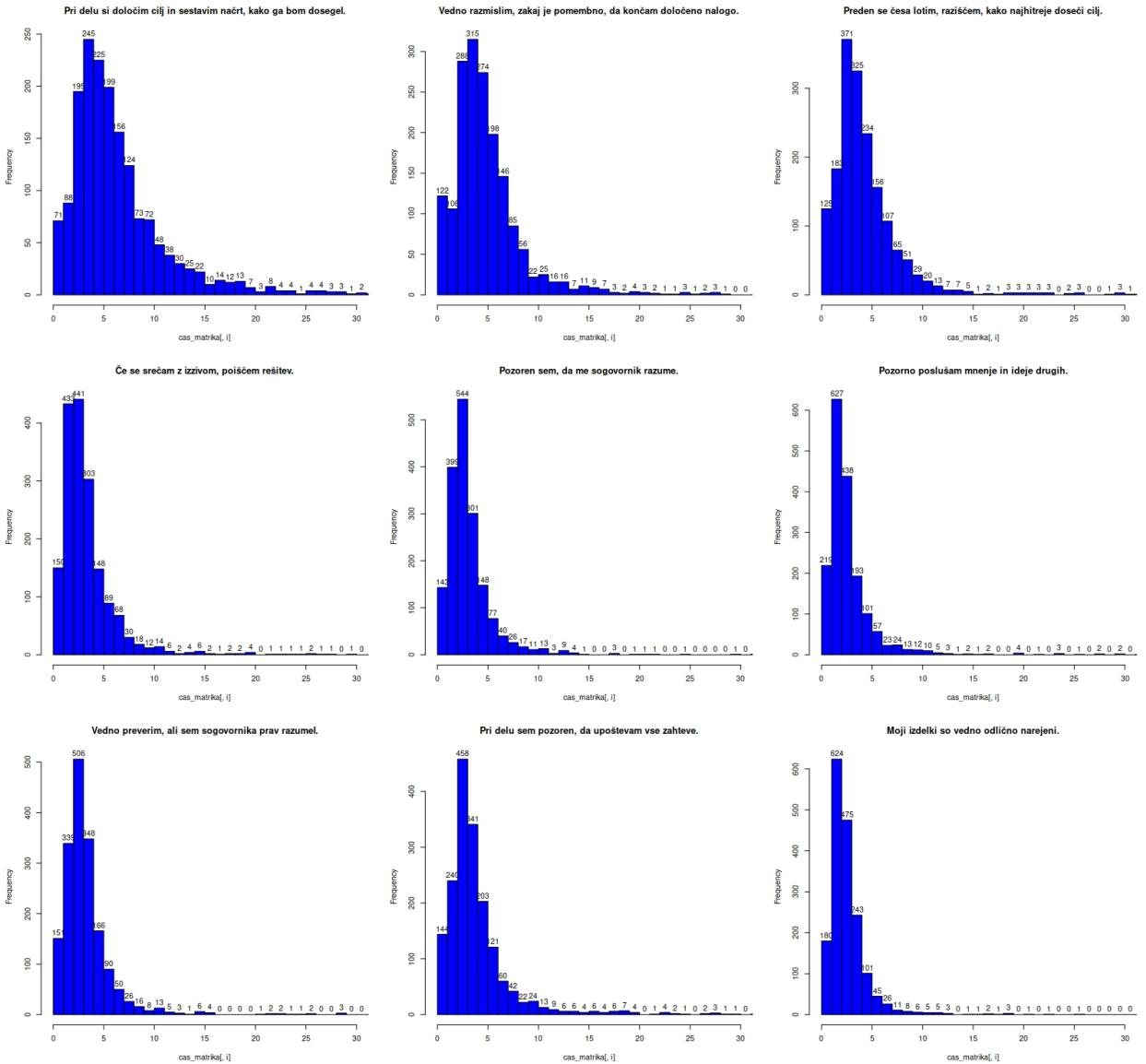
Slika 5.13: Histogram časov: vprašanja 19-27



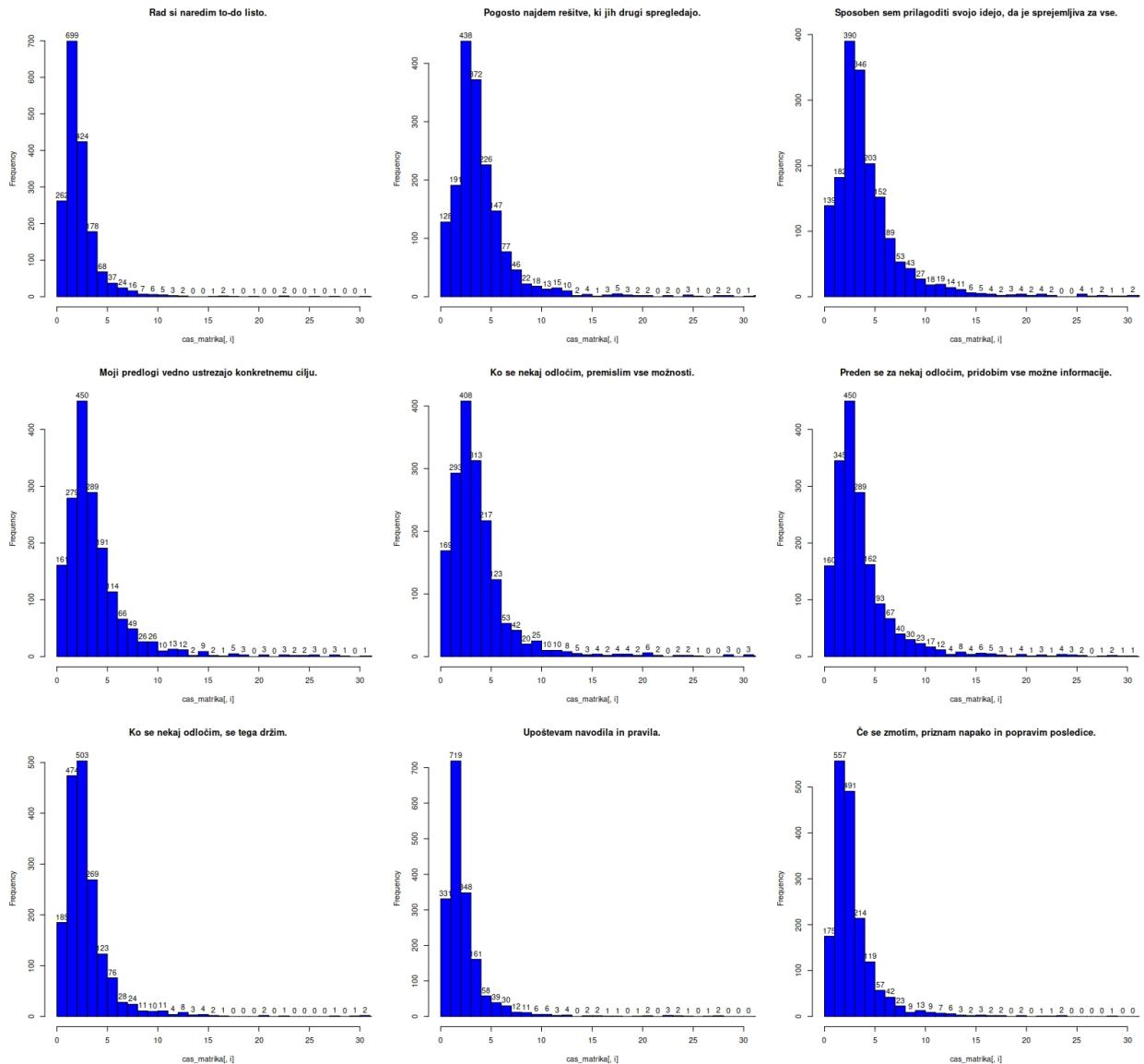
Slika 5.14: Histogram časov: vprašanja 28-36



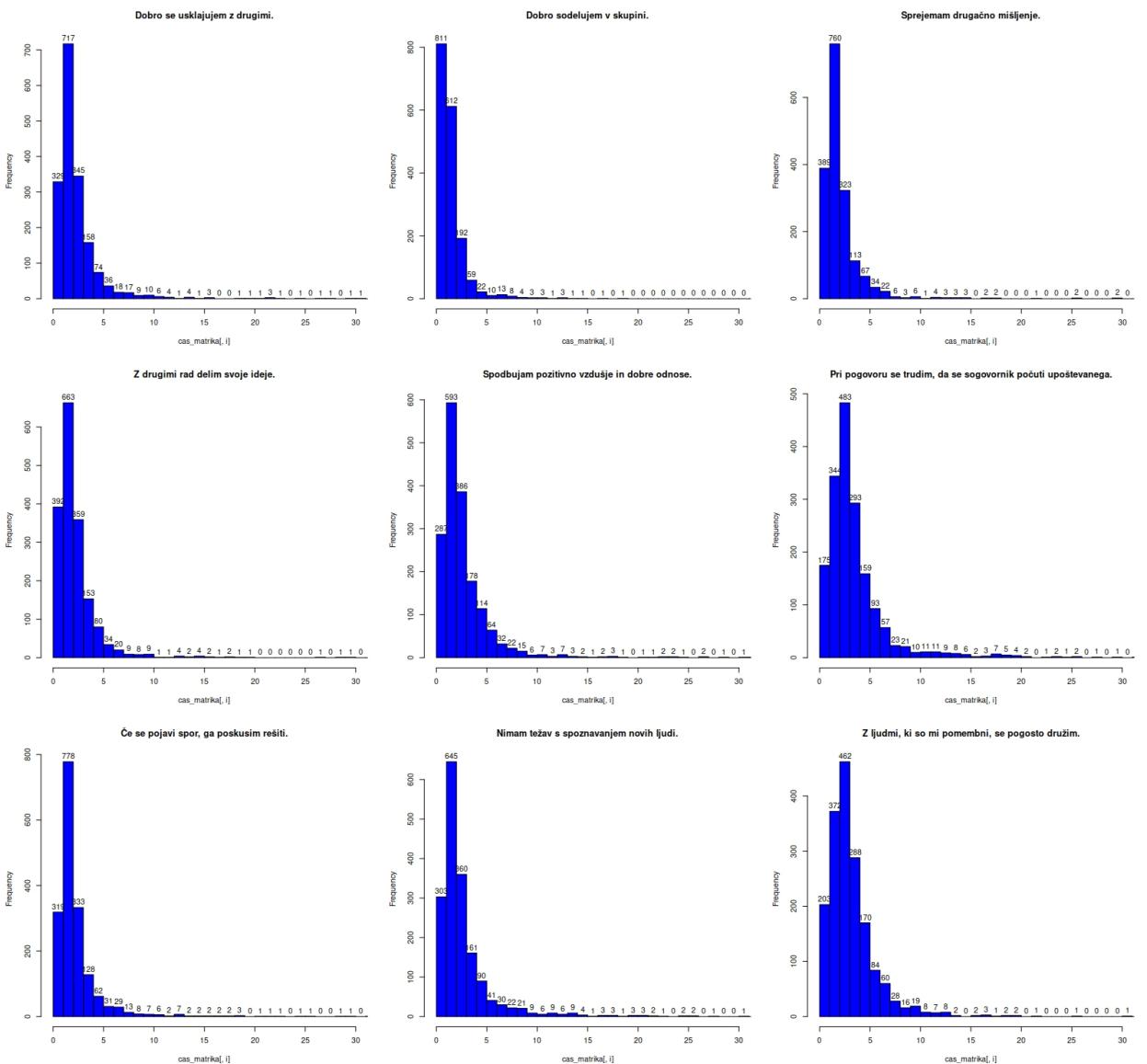
Slika 5.15: Histogram časov: vprašanja 37-39



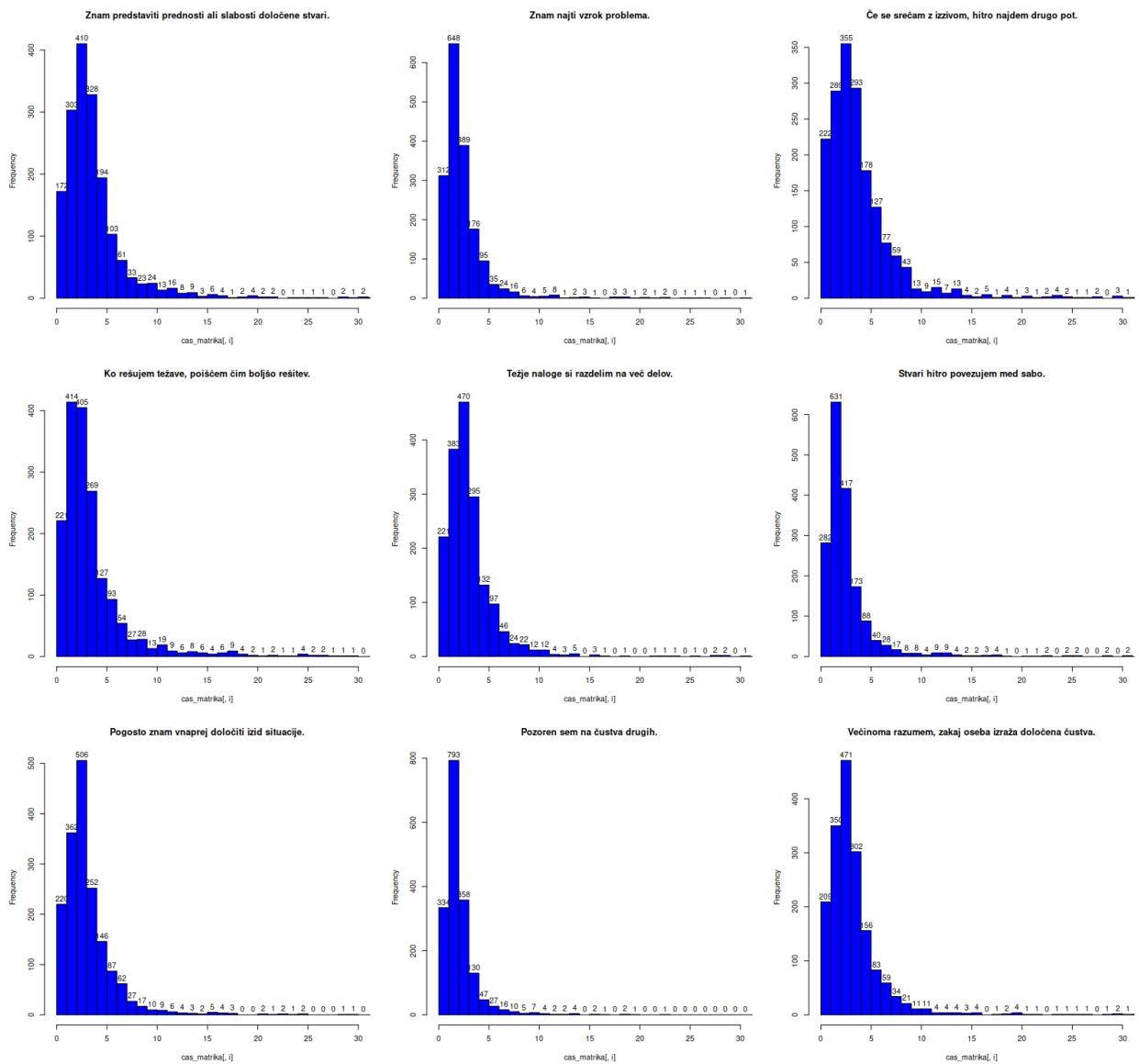
Slika 5.16: Histogram časov: vprašanja 40-48



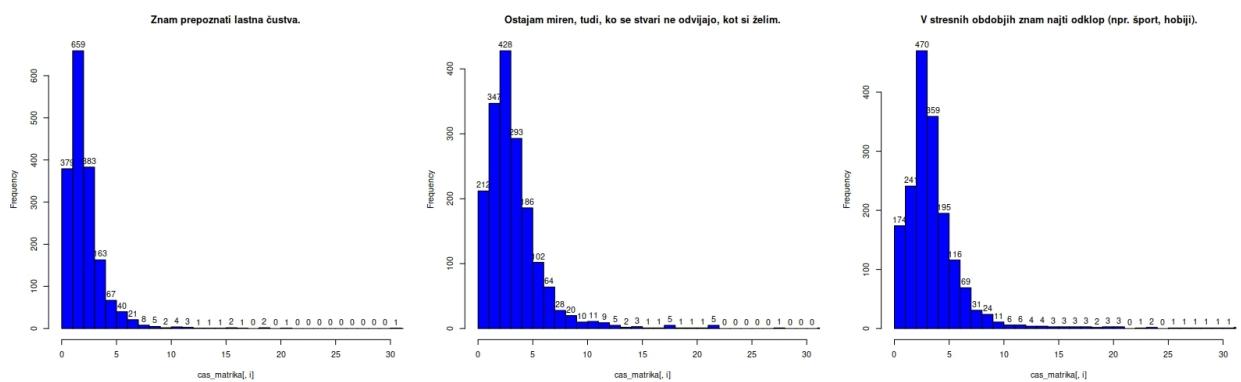
Slika 5.17: Histogram časov: vprašanja 49-57



Slika 5.18: Histogram časov: vprašanja 58-66



Slika 5.19: Histogram časov: vprašanja 67-75



Slika 5.20: Histogram časov: vprašanja 76-78

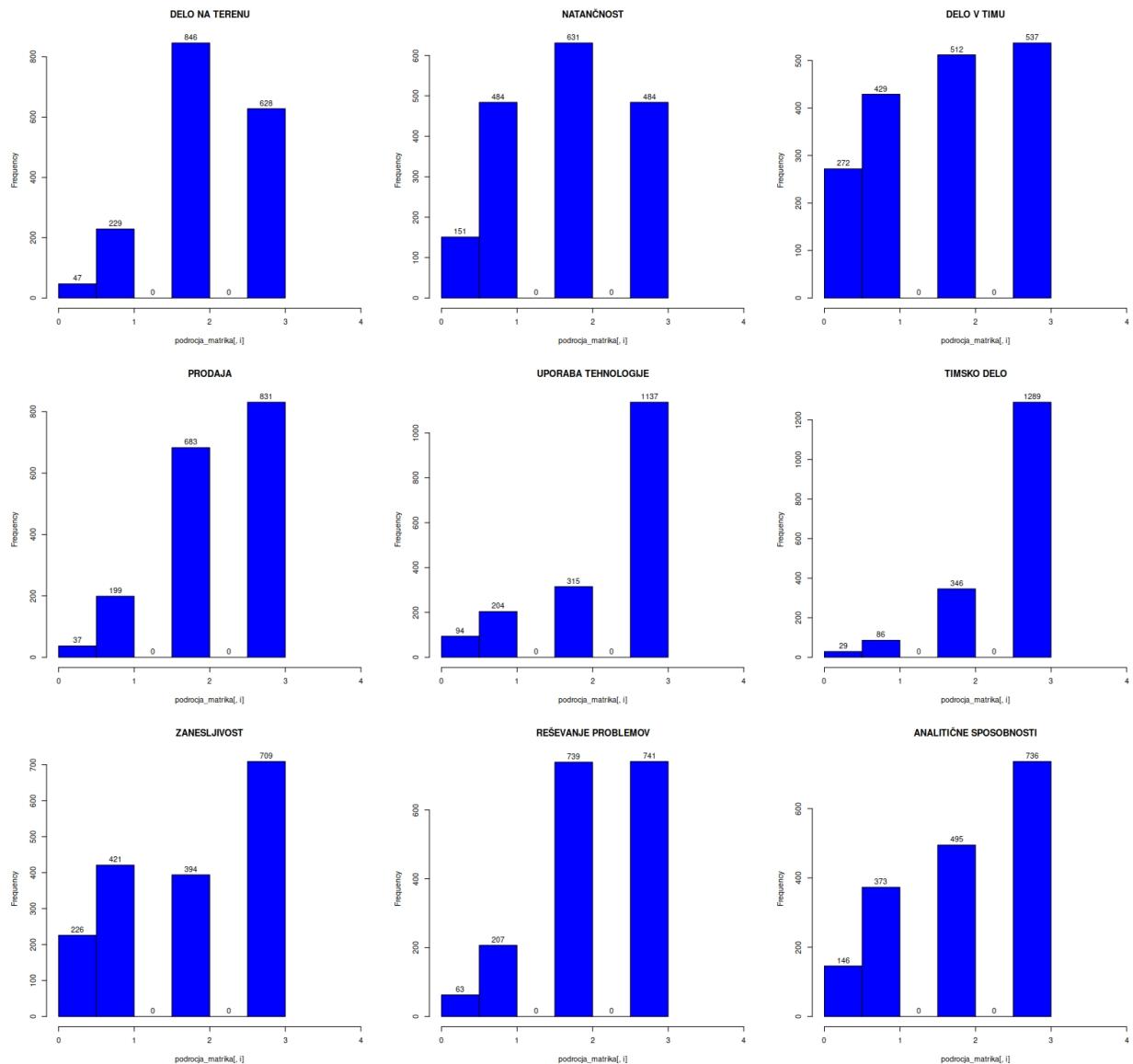
5.3 Opisna statistika - področja

Detajljna frekvenčna porazdelitev je prikazana v spodnji tabeli.

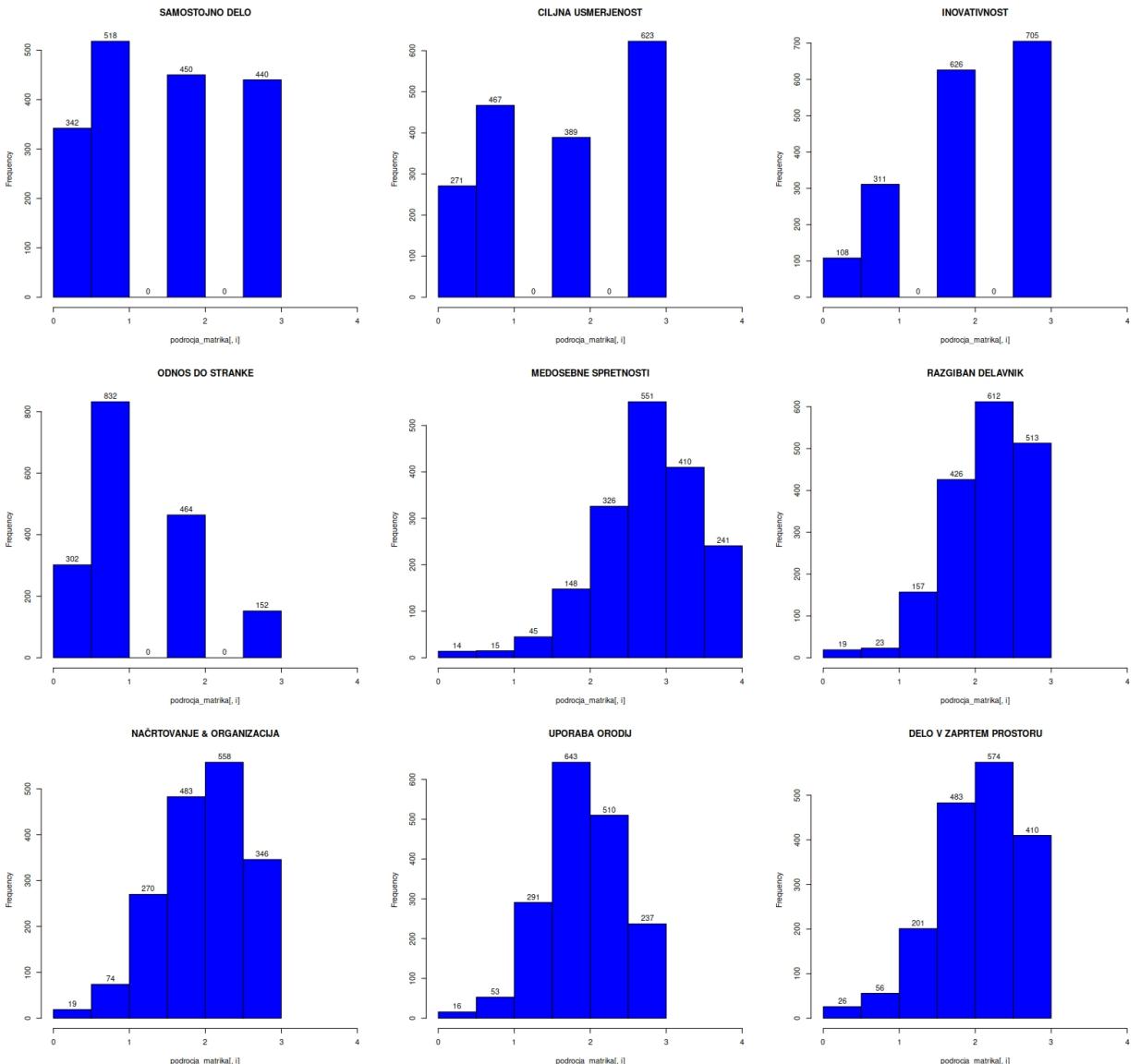
	vars	n	mean	sd	median	trimmed	mad	min	max	
1	1	1	1750	2.17	0.75	2.00	2.25	1.48	0	3
2	2	2	1750	1.83	0.93	2.00	1.89	1.48	0	3
3	3	3	1750	1.75	1.05	2.00	1.81	1.48	0	3
4	4	4	1750	2.32	0.76	2.00	2.42	1.48	0	3
5	5	5	1750	2.43	0.89	3.00	2.60	0.00	0	3
6	6	6	1750	2.65	0.65	3.00	2.80	0.00	0	3
7	7	7	1750	1.91	1.07	2.00	2.01	1.48	0	3
8	8	8	1750	2.23	0.79	2.00	2.34	1.48	0	3
9	9	9	1750	2.04	0.98	2.00	2.15	1.48	0	3
10	10	10	1750	1.56	1.07	2.00	1.58	1.48	0	3
11	11	11	1750	1.78	1.09	2.00	1.85	1.48	0	3
12	12	12	1750	2.10	0.90	2.00	2.20	1.48	0	3
13	13	13	1750	1.27	0.85	1.00	1.22	1.48	0	3
14	14	14	1750	2.88	0.69	3.00	2.92	0.74	0	4
15	15	15	1750	2.28	0.54	2.25	2.32	0.37	0	3
16	16	16	1750	2.10	0.57	2.25	2.13	0.74	0	3
17	17	17	1750	2.03	0.53	2.00	2.04	0.37	0	3
18	18	18	1750	2.17	0.58	2.25	2.20	0.74	0	3
19	19	19	1750	1.58	0.37	1.50	1.63	0.37	0	2
20	20	20	1750	3.08	0.76	3.00	3.16	0.74	0	4
21	21	21	1750	2.44	0.56	2.50	2.52	0.74	0	3
22	22	22	1750	2.25	0.59	2.25	2.30	0.74	0	3
23	23	23	1750	2.21	0.55	2.25	2.24	0.74	0	3
24	24	24	1750	2.20	0.56	2.25	2.23	0.37	0	3
25	25	25	1750	2.28	0.64	2.25	2.35	0.74	0	3
26	26	26	1750	1.41	0.46	1.50	1.46	0.37	0	2

Histogrami za področja se so prikazani na spodnjih slikah.

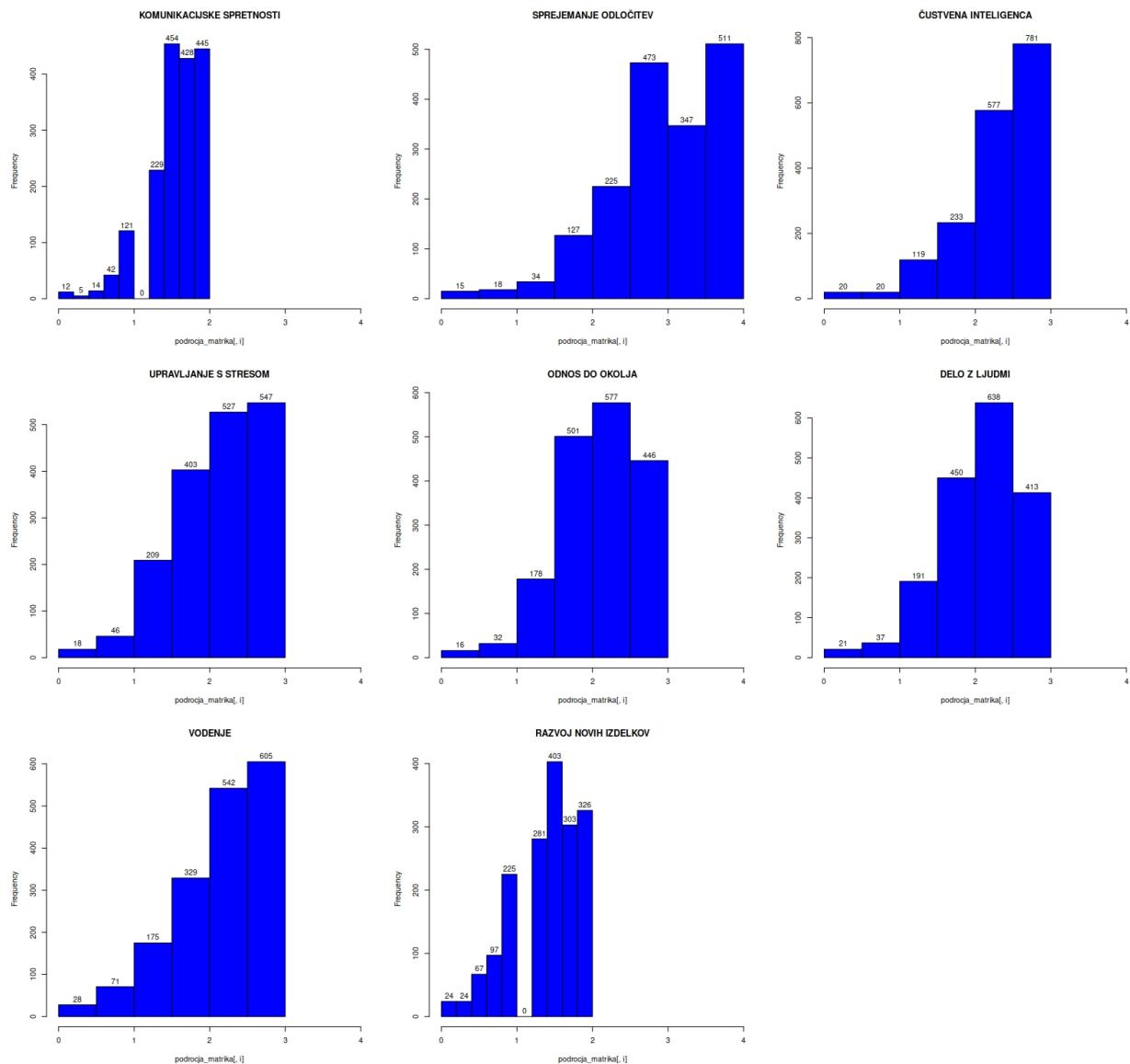
Slika 5.21: Histogram področij 1-9



Slika 5.22: Histogram področij 10-18



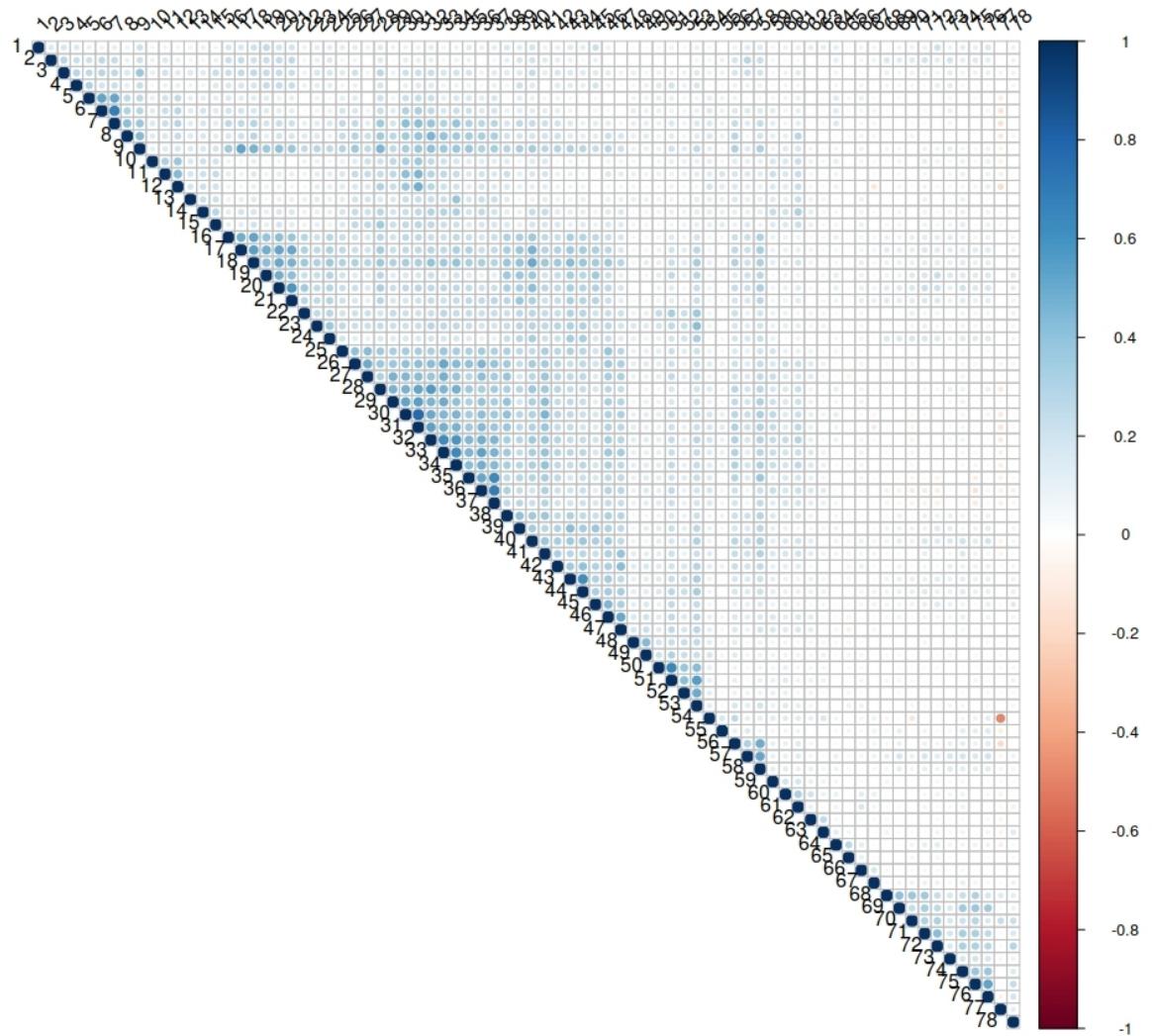
Slika 5.23: Histogram področij 19-26



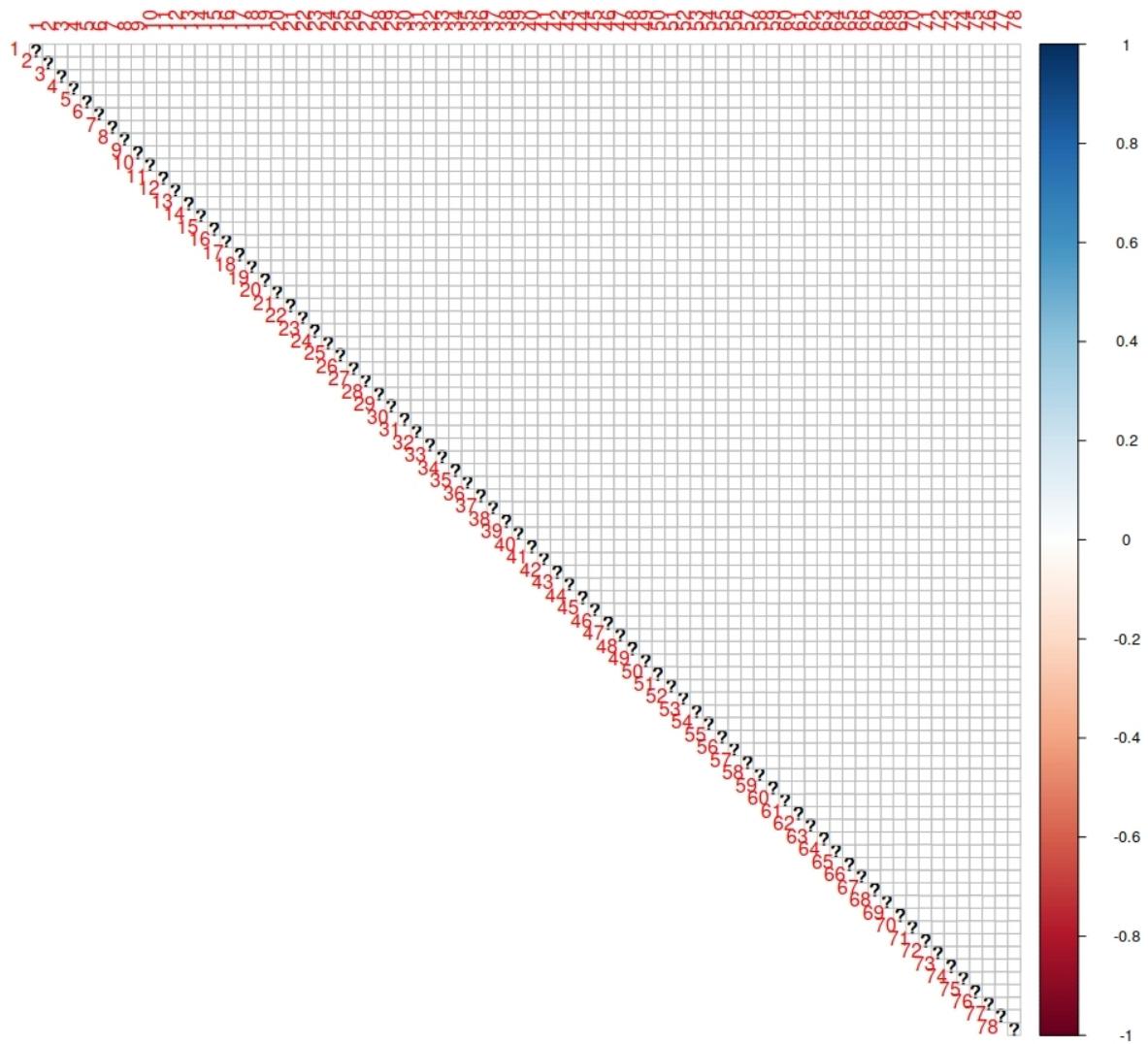
5.4 Korelacijska analiza odgovorov

Moč povezave med dvema spremenljivkama merimo s korelacijskim koeficientom v intervalu $[-1, +1]$. Poznamo več vrst korelacijskih koeficientov kot so Pearsonov, Spearmanov, Kendalov. V nalogi smo uporabili privzetega (Pearsonovega), čeprav bi bila glede na vrsto spremenljivk in njihovo domeno primernejša Spearmanov in Kendalov. Najprej so obravnavani odgovori, nato pa področja. Pri obeh je najprej grafično predstavljena korelacijska matrika nad glavno diagonalo, nato pa še signifikantne korelacije nad glavno diagonalo. Pokazalo se je, da so povezave šibke in ne-signifikantne.

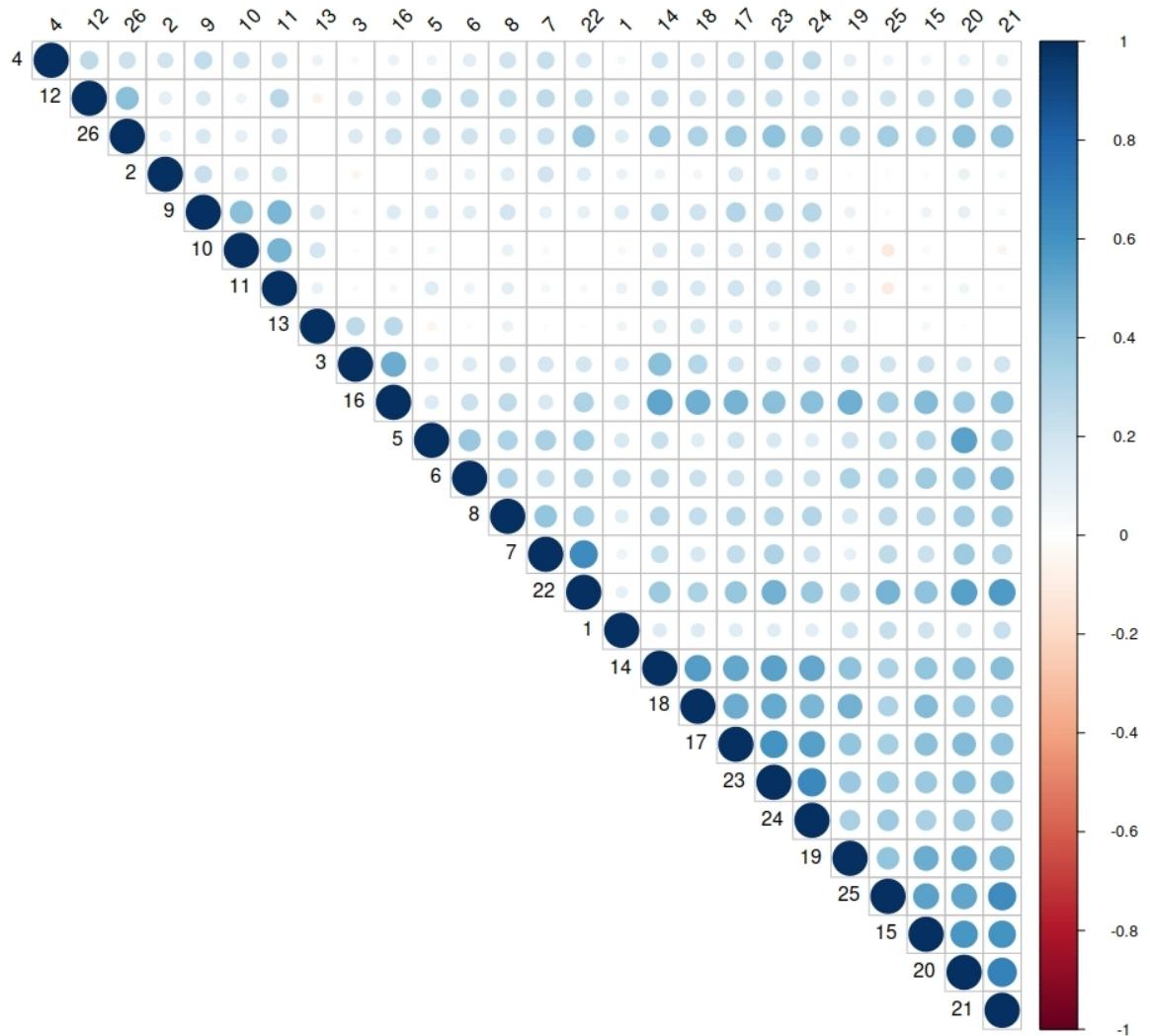
Slika 5.24: Prikaz korelacije med odgovori



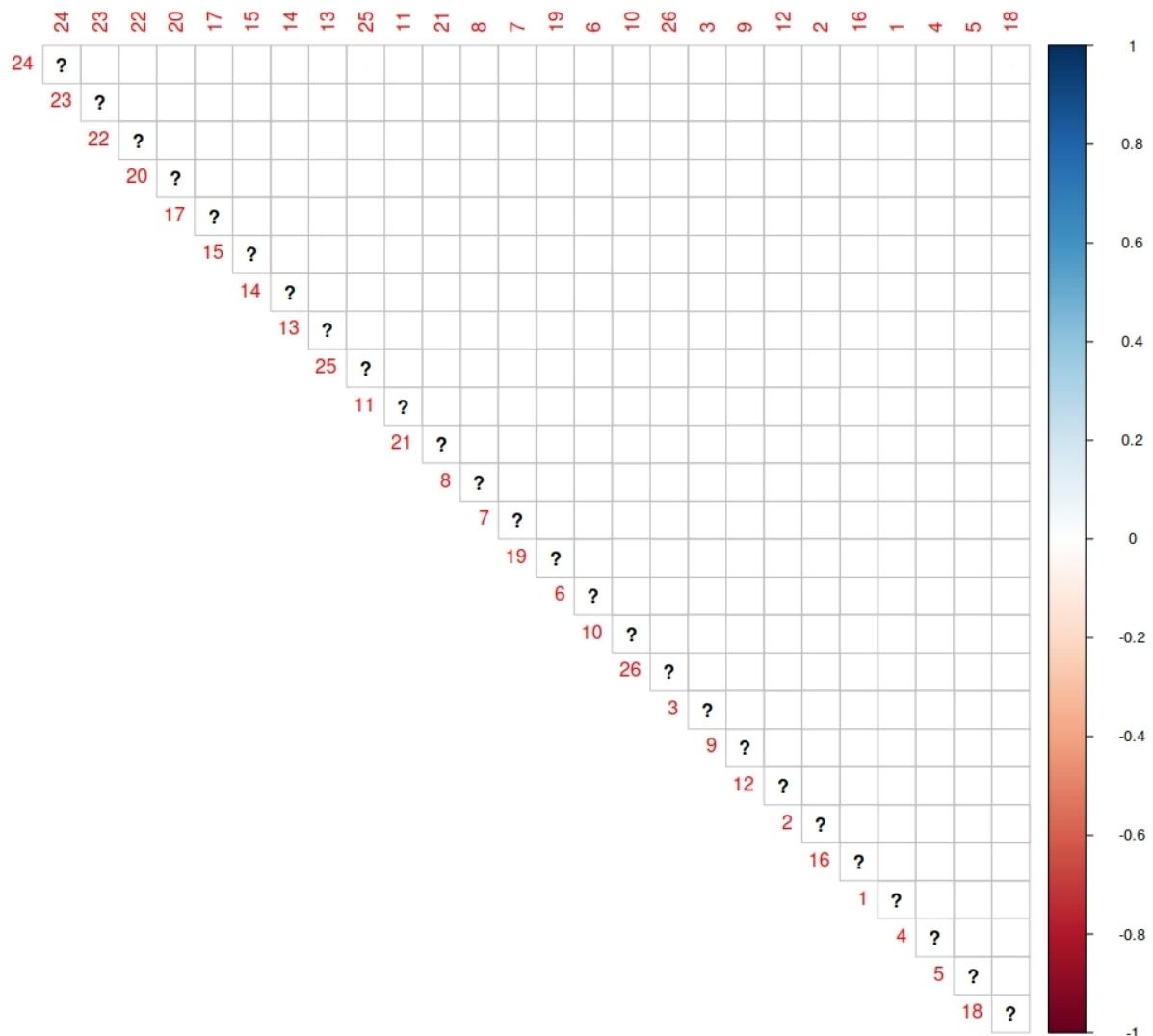
Slika 5.25: Signifikantne korelacijske odgovore



Slika 5.26: Prikaz korelacije med področji

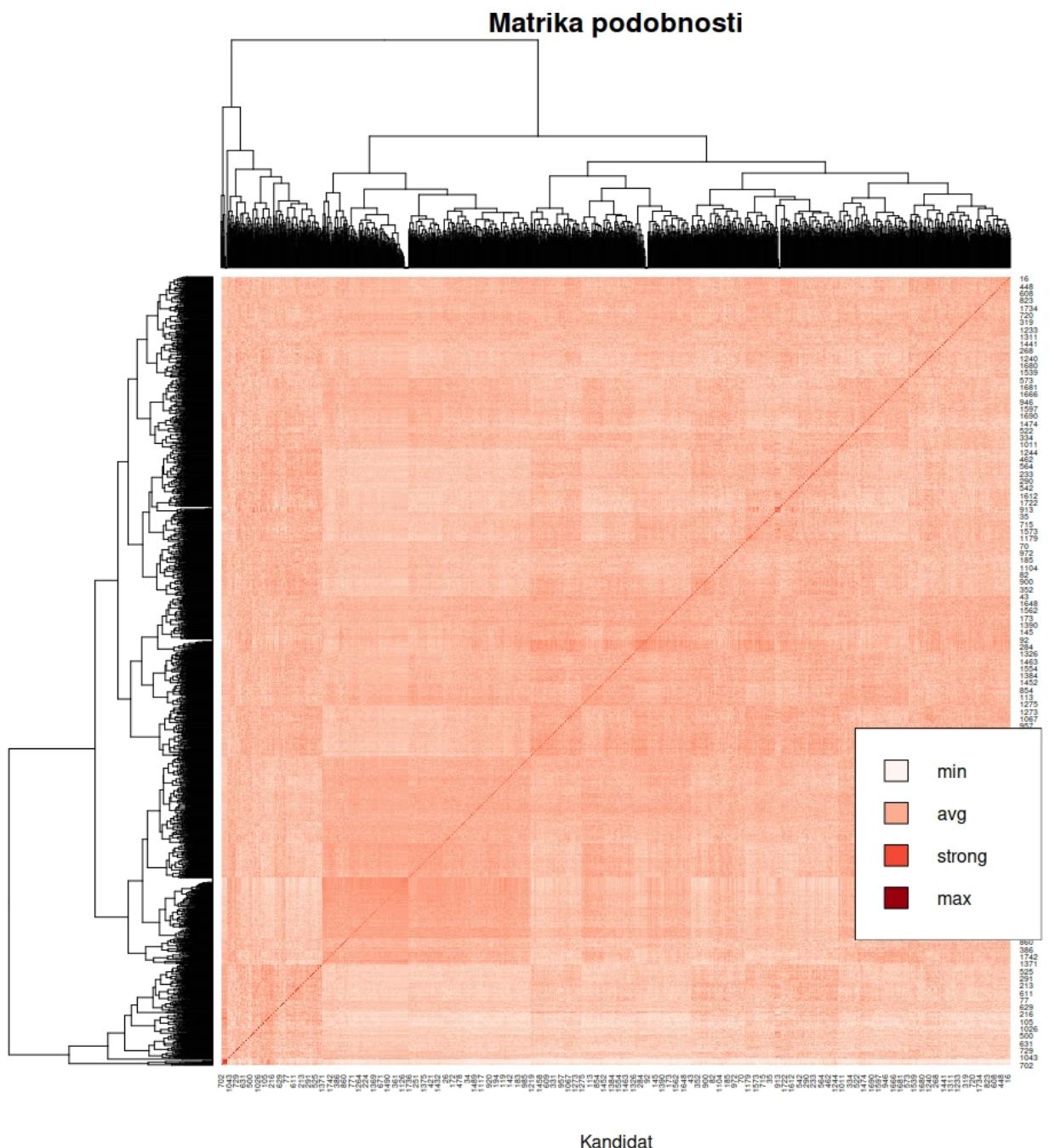


Slika 5.27: Signifikantne korelacijske področje

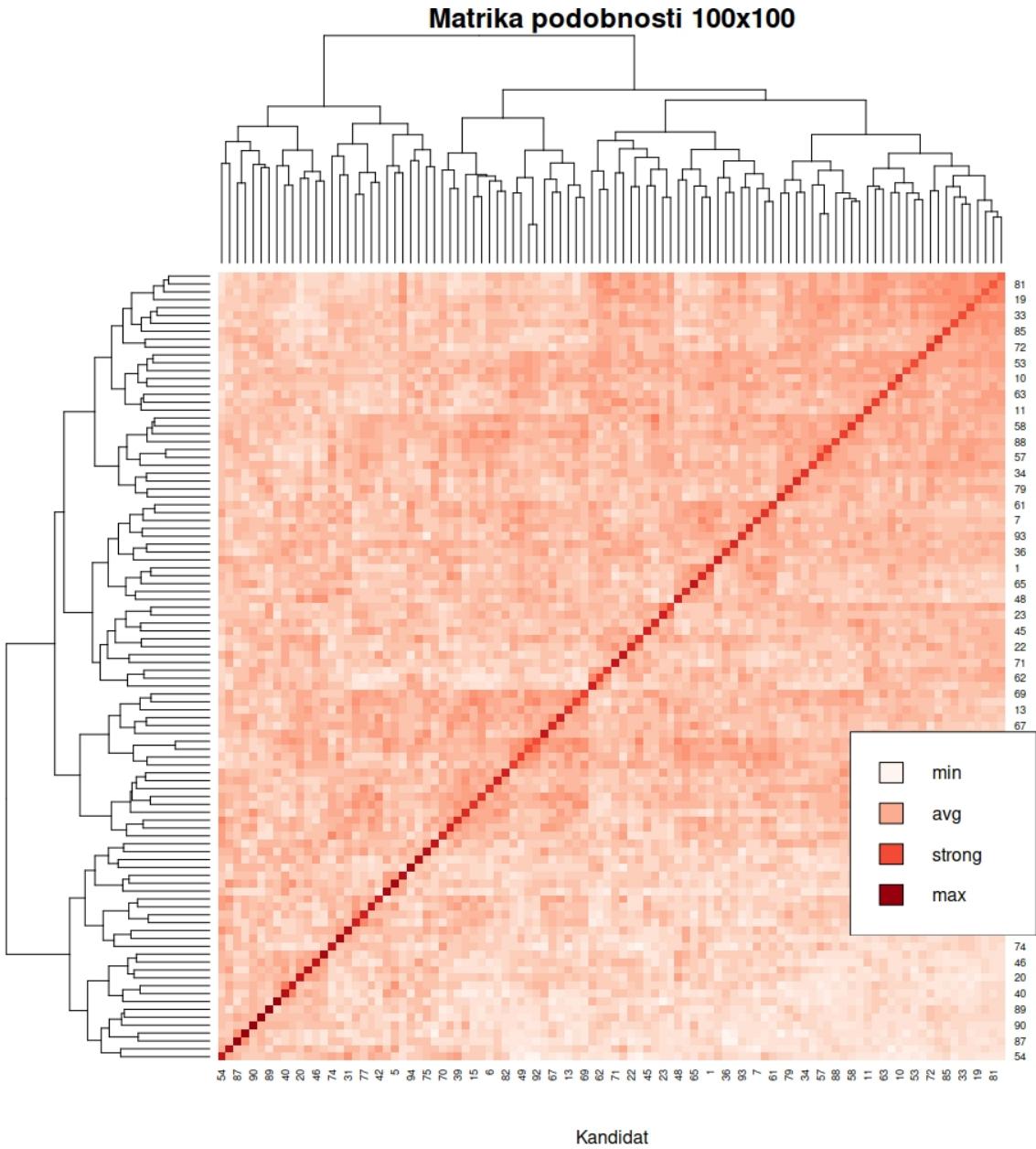


5.5 Analiza podobnosti odgovorov med kandidati

Podobnost odgovorov je prikazana na dveh toplotnih zemljevidih, prvi zajema vseh 1750 anket, na drugi pa je prikazan le izsek prvih stop. Poprečna podobnost je 0.4663187, minimalna 0 in maksimalna 1. Porazdelitev podobnosti je zvonaste oblike (normalna porazdelitev), nekoliko asimetrična in srednje sploščena.



Slika 5.28: Toplotni zemljevid podobnosti odgovorov v anketah

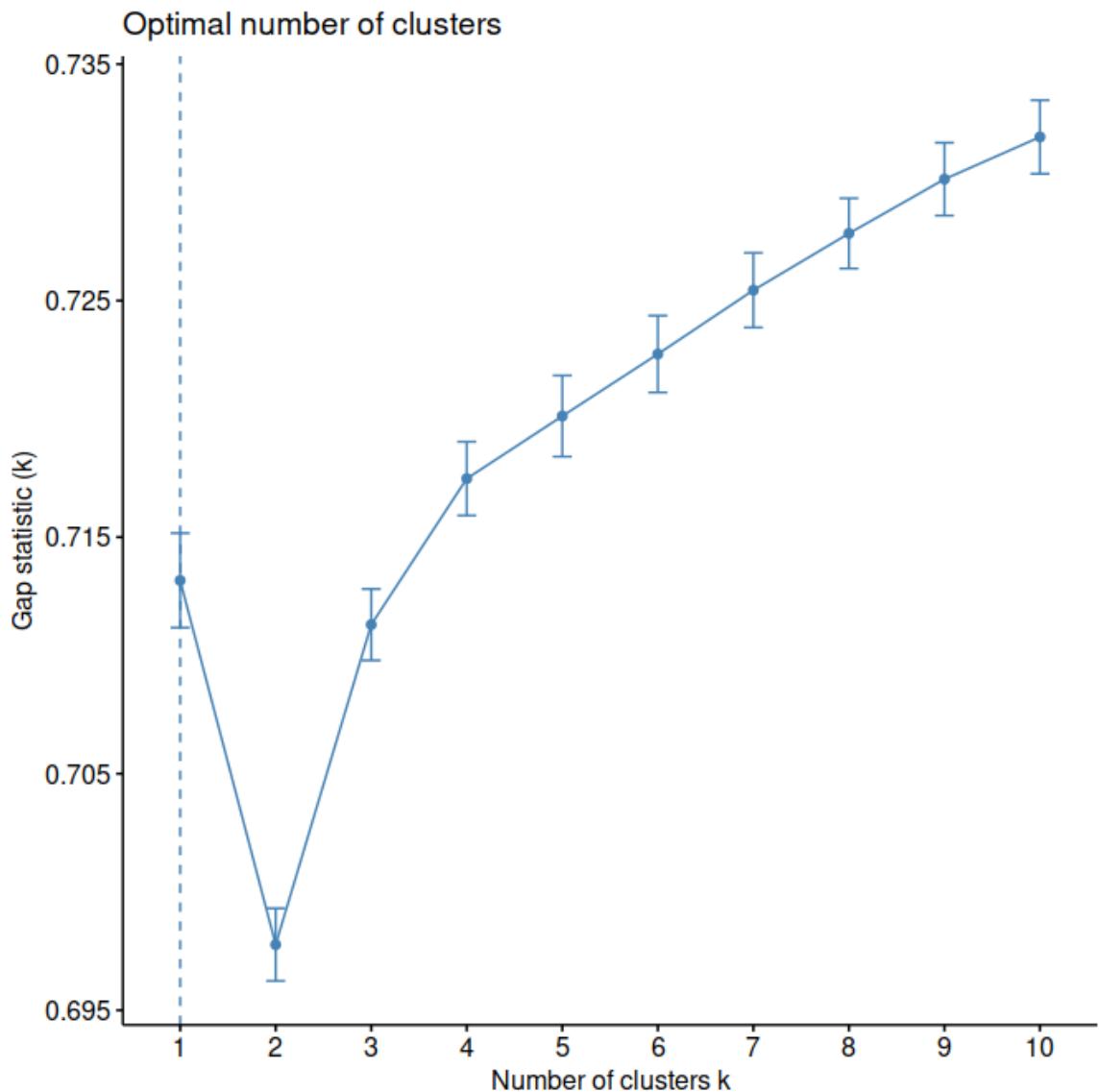


Slika 5.29: Toplotni zemljevid podobnosti prvih 100 anket

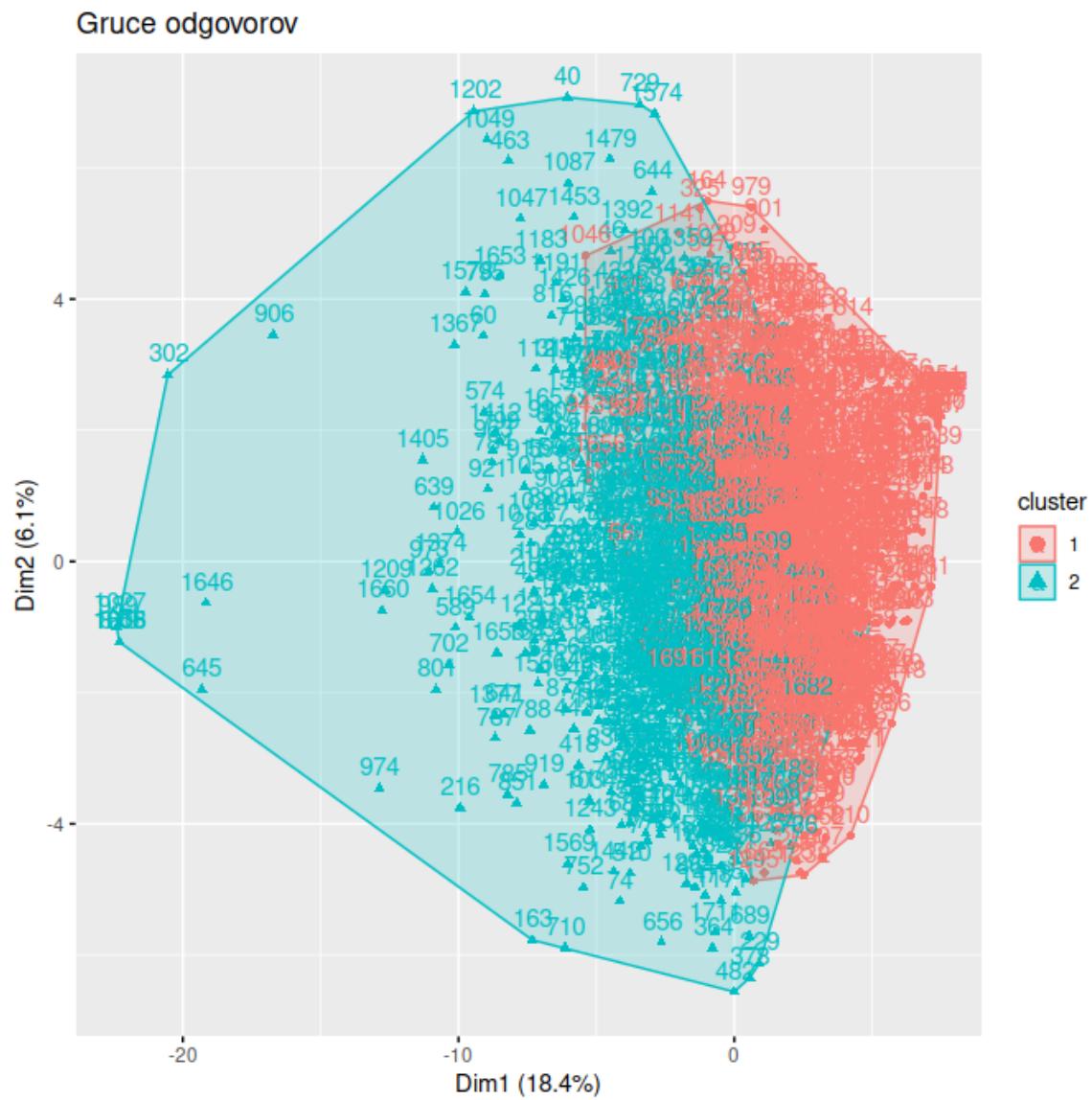
5.6 Analiza gruč in čiščenje podatkov

Pri analizi gruč je eno ključnih vprašanj število gruč. V nalogi smo uporabili statistiko vrzeli in v vseh primerih izračunali, da je optimalno število gruč 2.

Slika 5.30: Optimalno število gruč - odgovori



Slika 5.31: Gruče odgovorov

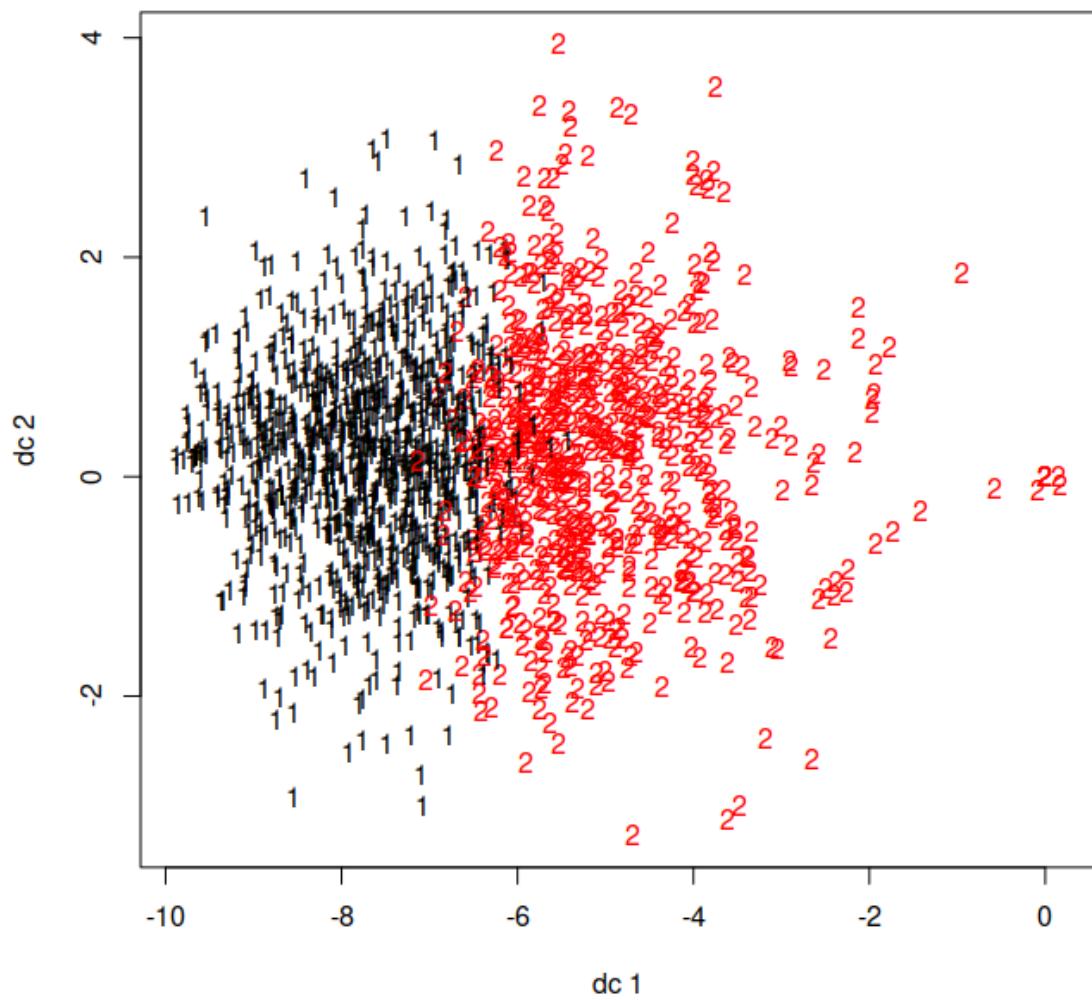


Slika 5.32: Gruče področij

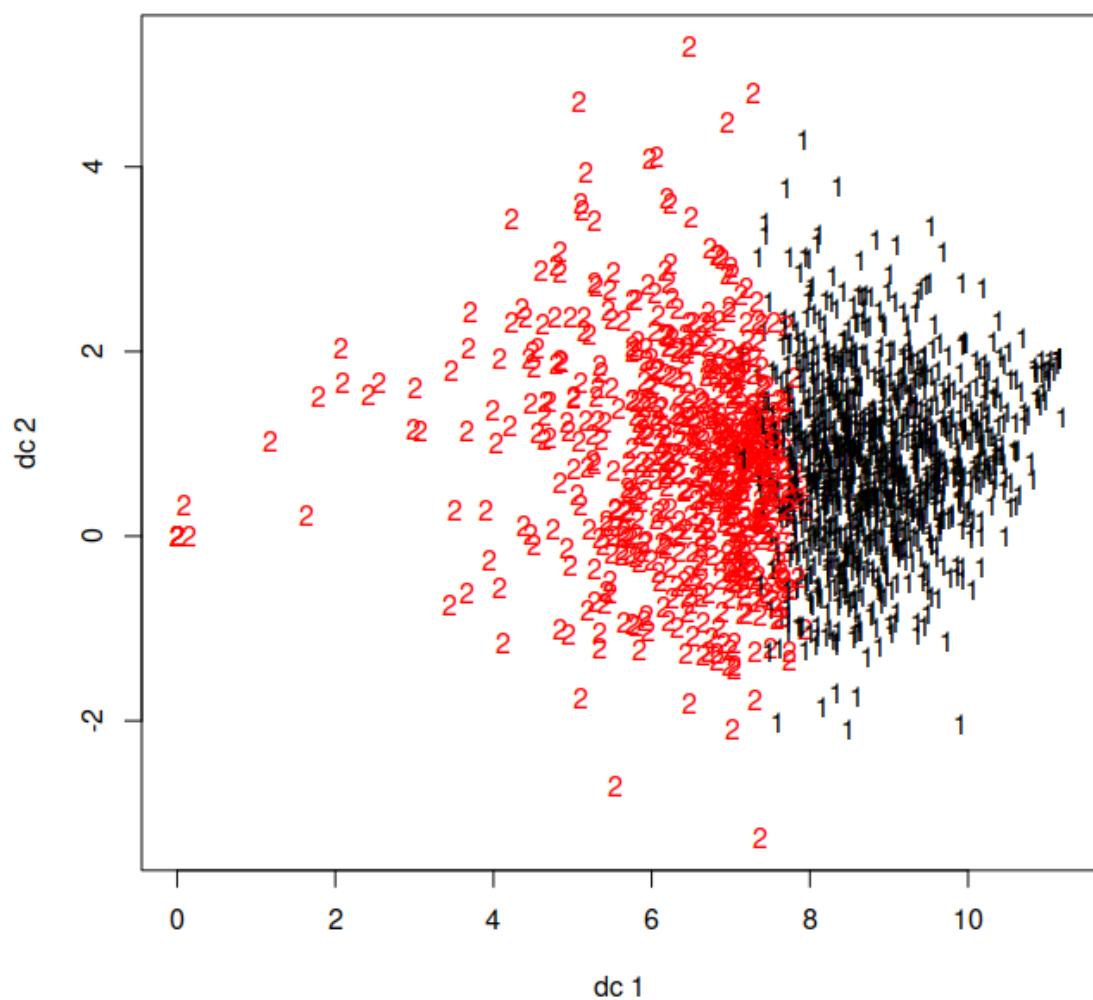


Drugi način prikaza analize gruč:

Slika 5.33: Gruče področij



Slika 5.34: Gruče področij



Čiščenje podatkov

Pri obdelavi velikega števila podatkov je pomembno, da odstranimo slabe, nepopolne ali napačne podatke. V konkretnem primeru je bilo vseh 1750 anket izpolnjeno pri vseh vprašanjih. Pri analizi časov reševanja anket pa je razvidno, da so nekatere ankete naklikano veliko prehitro, da bi jih lahko upoštevali kot relevantne. Po lastnih meritvah se mi je pokazalo, da v manj kot 380 sekundah nisem mogel pošteno odgovoriti na vseh 78 vprašanj. Zgornjo mejo sem po občutku določil na 624 sekund.

```
1 # Kreiram meje
2 spodnja_meja <- 380
3 zgornja_meja <- 624
4
5 # matriko relevantnih oziroma veljavnih anket
6 veljavne_ankete <- subset(odg_matrika , rowSums(cas_matrika
  [,1:78]) > spodnja_meja & rowSums(cas_matrika[,1:78]) <
  zgornja_meja)
```

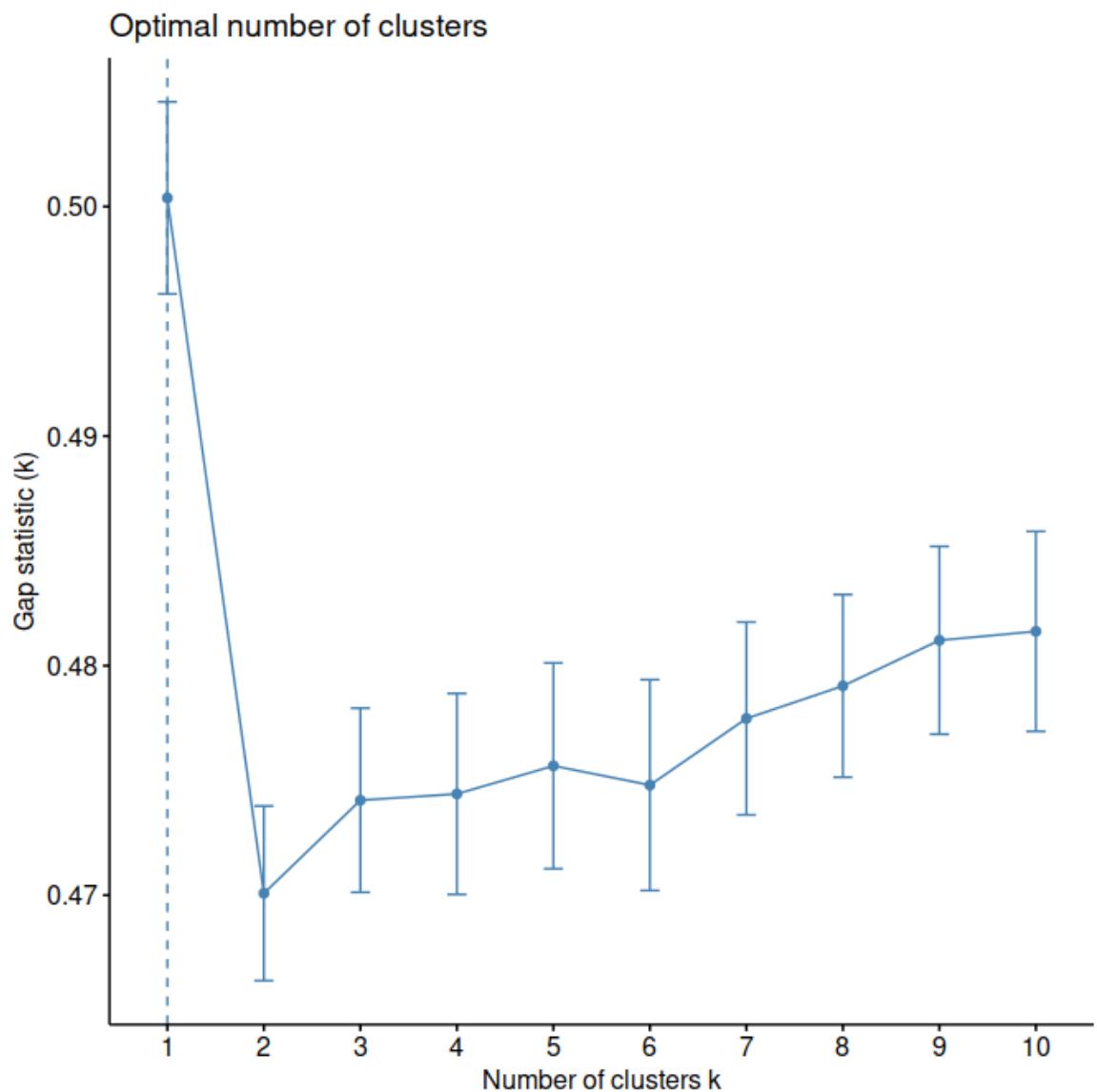
Reducirana matrika odgovorov obsega 269 anket.

```
1 length(veljavne_ankete)/78
2 > [1] 269
```

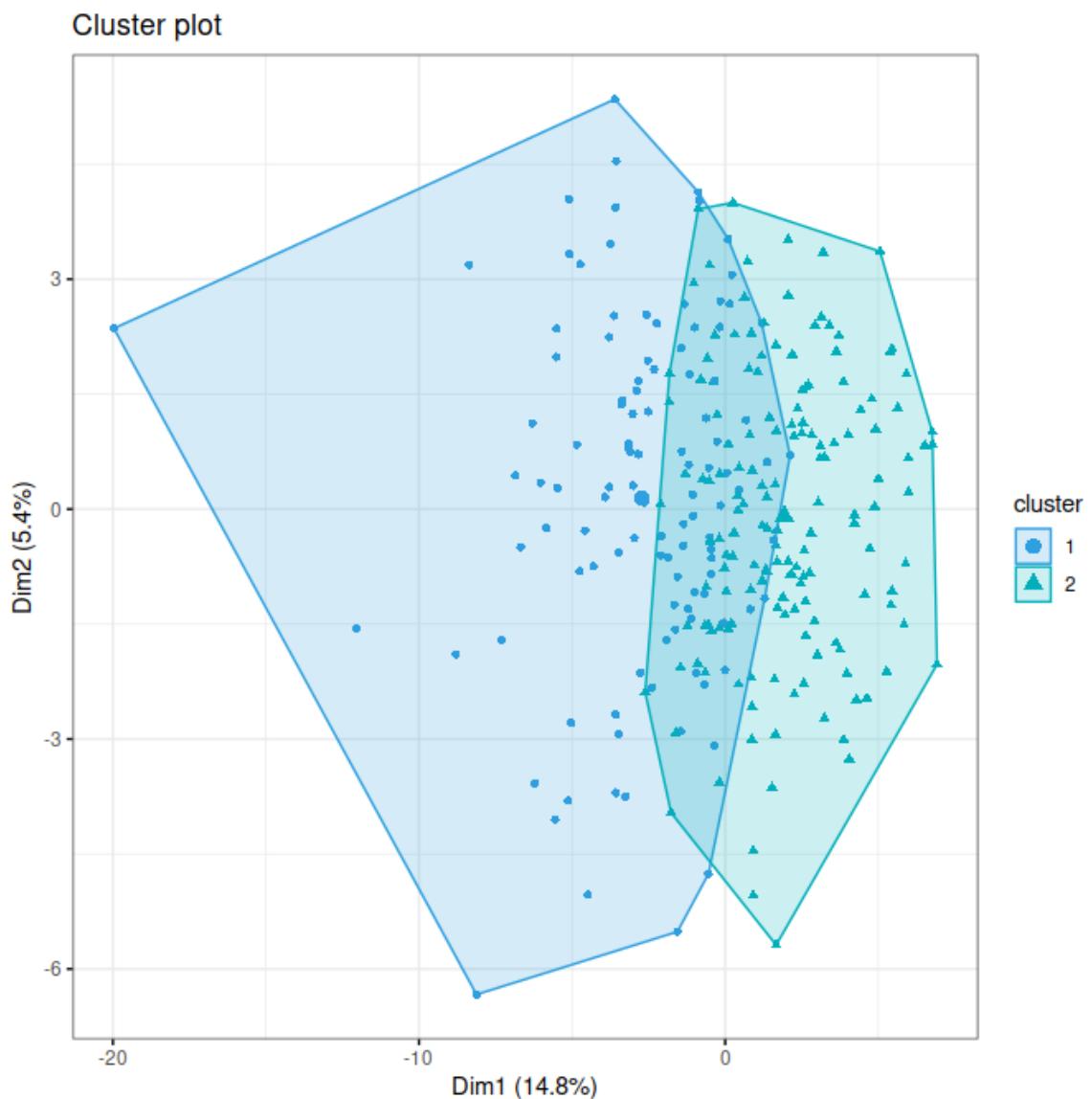
Ponoven izračun optimalnega števila gruč in klasificiranje v dve gruči sta pokazala, da je med gručama še vedno veliko prekrivanje.

Alternativni način prikaza gruč po čiščenju odgovorov:

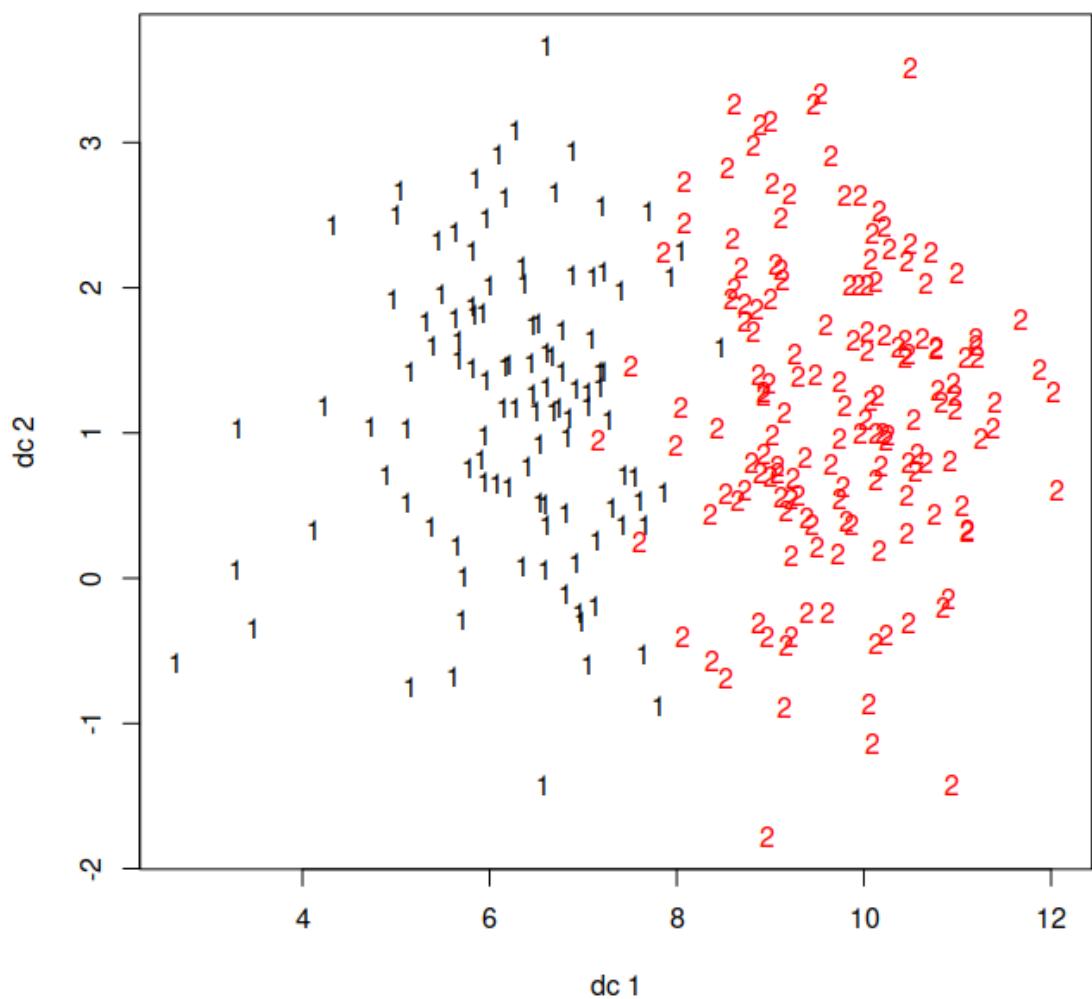
Slika 5.35: Število gruč odgovorov po čiščenju podatkov



Slika 5.36: Gruče odgovorov očiščenih podatkov



Slika 5.37: Gruče področij



6 Diskusija

Na podlagi izračunov opisne statistike lahko ugotovimo nekaj zakonitosti o podatkih. Povprečna vrednost za prvo skupino odgovorov nam pove, da je več kandidatov odgovorilo DA na prvih 39 vprašanj. V drugi skupini vprašanje je najpogostejši odgovor NAJPOGOSTEJE. Najmanjša vrednost standardnega odklona je 0.20 pri vprašanju 16 (*Ohranjam dobre odnose z ljudmi, ki so mi pomembni.*). Če predpostavimo, da so odgovori pošteni, pomeni, da anketirancem odnos do ljudi veliko pomeni. Majhen standardni odklon (vrednosti med 0.20 in 0.32) zasledimo pri vprašanju 70 (*Ko rešujem težave, poiščem čim boljšo rešitev.*).

Povprečen čas reševanja za eno vprašanje je 4.58 sekunde. Prvi in drugi del ankete se ne razlikujeta bistveno (4.47 s in 4.6 s). Najdalj (poprečje 11.32 s) so anketiranci reševali vprašanje 40 (*Pri delu si določim cilj in sestavim načrt, kako ga bom dosegel.*). Predvidevamo lahko, da se je to zgodilo zaradi spremembe tipa vprašanja (binarni tip, Likertov tip). Standardni odklon tega časa je izjemno velik. Zanimiv in nepojasnjen je velik odklon (199.11) na vprašanje 39 (*Nekoč bom imel svoj laboratorij.*). Najmanjši standardni odklon čas je pri vprašanju 74 (*Pozoren sem na čustva drugih.*), kar spet govori o tem, da so anketiranci pozorni do drugih.

Pri področjih so poprečne vrednosti med 1.27 do 3.08. Standardni odklon je največji (1.09) za področje *Ciljna usmerjenost*, najmanjši (0.37) pa za področje *Komunikacijske spretnosti*. Po korelacijski analizi vidimo, da je najmočnejša negativna korelacija med vprašanjema 55 in 77 (*Ko se nekaj odločim, se tega držim. Če Ostajam miren, tudi, ko se stvari ne odvijajo, kot si želim.*). Sicer so vse korelacije statistično nesignifikantne na nivoju tveganja 5 odstotkov.

Izid analize gruč odgovorov je po čiščenju podatkov zelo podoben predhodnemu. Z metodo *Gap Stat* (statistika vrzeli) sem ugotovil, da je optimalno število gruč spet 2.

7 Zaključki

Celotna raziskava je zajemala več aktivnosti. Najprej sem preštudiral precej literature o jeziku R in iskal primerne knjižnice za statistično analizo. Spoznal sem nekaj zgodovine, namen in razvoj tega programskega jezika. Ugotovil sem, da je jezik R pomemben za področje informatike in znanosti o podatkih. Pri analizi aplikacije KAMbi sem poglobil poznavanje baz podatkov in povezanost z disciplino podatkovne analitike. Zelo veliko razvijalcev nenehno razvija nove in dopolnjuje obstoječe pakete za jezik R, pri čemer so tisti, ki so objavljeni v repozitoriju CRAN odprto-kodni. Zasledil sem številne pakete za področje strojnega učenja in umetne inteligence, ki pa jih v nalogi nisem uporabiljal. Diplomska naloga je bila zelo dobra priložnost, da spoznam ustrezno literaturo in preskusim jezik R na konkretnem primeru analize podatkov.

Aplikacija KAMbi mi je omogočila, da povečam izpopolnim poznavanje Oracle APEX in baze Oracle. Kljub temu, da je Oracle APEX malo-kodna rešitev za izgradnjo aplikacij, je celovita in močna platforma za izdelavo polno funkcionalnih aplikacij. Spoznal sem možnosti PL/SQL-a in temeljito analiziral arhitekturo aplikacije. Oraclov sistem je konsistenten, pragmatičen in vključuje integrirane metodami, ki razvijalcu zelo olajšajo ustvarjanje aplikacije. Menim, da je aplikacija KAMbi pomagala določenem delu anketirancev pri izbiri inženirskega poklica in izobraževalne inštitucije.

V raziskovalnem delu sem preštudiral in uporabil različne metode statističnih obdelav. Pri delu z velikim številom podatkov sem zaznal kompleksnost podatkovne analitike.

Menim, da je bila aplikacija KAMbi odlična priložnost za spoznavanje in konkretno analizo podatkov. Večkrat sem analizo tudi ponavljal, da bi se prepričal o rezultatih in pravilnosti mojega pristopa.

Cilji naloge so bili v celoti doseženi, vendar se s to analizo odpirajo številna nova vprašanja, ki jih še nisem uspel odgovoriti.

8 Literatura

1. Wiley, M., Wiley. J. F. (2020) *Beginning R* 4. Apress, Berkeley, CA. Victoria College, Victoria USA
2. Stojanović, D. (2018) *Osnove R-a (inovativnost, zanat, jezik)* na <https://rzanat.rs/predgovor.html>
3. Wickham, H. (2015) *R Packages*, O'Reilly books, USA.
4. Oracle Academy - *PL/SQL*
5. *Zapiski predavanj:* dr. Robert Leskovar, Računalništvo in informatika & Sistemska analiza, 2018/19.
6. Jonge, D., Loo, D.M. (2013) *An introduction to data cleaning with R*. Statistics Netherlands, The Hague - Herlen
7. James G, Witten D, Hastie T, Tibshirani R (2017) *An Introduction to Statistical Learning with Applications in R*. Springer New York
8. Wickham H, Grolemund G (2016) *R for Data Science*. O'Reilly Media