Eksplorativna analiza podataka iz sustava za isporuku oglasa

Miljević, Marinko

Master's thesis / Diplomski rad

2016

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek / Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:200:413940

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Repository / Repozitorij:

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek

zir.nsk.hr

Image not found or type unknown

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni diplomski studij računarstva

EKSPLORATIVNA ANALIZA PODATAKA IZ SUSTAVA ZA ISPORUKU OGLASA

Diplomski rad

Marinko Miljević

Osijek, 2016

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Obrazac D1: Obrazac za imenovanje Povjerenstva za obranu diplomskog rada

Osijek, 04.10.2016.

Odboru za završne i diplomske ispite

Imenovanje Povjerenstva za obranu diplomskog rada

Ime i prezime studenta:	Marinko Miljević
Studij, smjer:	Diplomski sveučilišni studij Računarstvo, smjer Procesno računarstvo
Mat. br. studenta, godina upisa:	D 726 R, 14.10.2014.
OIB studenta:	22470227024
Mentor:	Doc.dr.sc. Zdravko Krpić
Sumentor:	
Predsjednik Povjerenstva:	Doc.dr.sc. Mirko Köhler
Član Povjerenstva:	Doc.dr.sc. Josip Balen
Naslov diplomskog rada:	Eksplorativna analiza podataka iz sustava za isporuku oglasa
Znanstvena grana rada:	**Obradba informacija (zn. polje računarstvo)**
Zadatak diplomskog rada:	Zadatak ovog diplomskog rada je proučiti utjecaj različitih atributa na CTR (engl. click to response). Neki atributi imaju pozitivan, negativan i neutralan utjecaj na CTR i svi oni utječu na rad sustava, odnosno obradu i čuvanje podataka. Potrebno je provesti eksplorativnu analizu podataka (engl. Exploratory Data Analysis - EDA) te pronaći određene uzorke (engl. pattern) koji se pojavljuju među atributima, veze između ciljnog atributa i ostalih atributa i otkriti moguće anomalije (engl. outlier) u podacima koje negativno utječu na performanse sustava. Analizu je potrebno provesti na više nivoa jer različita granulacija može dati različite rezultate. Potrebno je otkriti koliko se može utjecati na stvarne sustave za isporuku oglasa odnosno kako ih se može poboljšati. Također bi bilo dobro otkriti neke nove, izvedene atribute koji bi mogli imati pozitivan utjecaj na ciljni atribut, odnosno CTR. U radu je potrebno
Prijedlog ocjene pismenog dijela ispita (diplomskog rada):	Izvrstan (5)
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: 3 Postignuti rezultati u odnosu na složenost zadatka: 3 Jasnoća pismenog izražavanja: 3 Razina samostalnosti: 3
Datum prijedloga ocjene mentora:	04.10.2016.

Potpis mentora za predaju konačne verzije rada u Studentsku službu pri završetku studija:	Potpis:
	Datum:

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA
I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O ORIGINALNOSTI RADA

Osijek, 12.10.2016.

Ime i prezime studenta:	Marinko Miljević
Studij:	Diplomski sveučilišni studij Računarstvo, smjer Procesno računarstvo
Mat. br. studenta, godina upisa:	D 726 R, 14.10.2014.
Ephorus podudaranje [%]:	1%

Ovom izjavom izjavljujem da je rad pod nazivom**: Eksplorativna analiza podataka iz sustava za isporuku oglasa**

izrađen pod vodstvom mentora Doc.dr.sc. Zdravko Krpić

i sumentora

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

Sadržaj

1. UVOD 1
2. OPIS ZADATKA 2
2.1. Alati za obradu podataka 2
2.2. Izvorni skup podataka 3
3. EKSPLORATIVNA ANALIZA PODATAKA 5
3.1. Opis atributa 7
3.2. Problem nedostajućih vrijednosti 10
3.3. Korelirani atributi 10
3.4. Istraživanje odnosa ciljnog atributa i ostalih atributa 18
3.5. Istraživanje nominalnih atributa 19
3.6. Istraživanje numeričkih atributa 27
3.7. Analiza veza između više atributa i ciljnog atributa 35
4. ANALIZA REZULTATA 52
5. ZAKLJUČAK 55
LITERATURA 57
SAŽETAK 59
ABSTRACT 60
ŽIVOTOPIS 61
PRILOZI 62
1. Tablice 62
2. Slike 64

1. UVOD

Današnja svakodnevica konstantno je izložena raznim oglasima. Oglasi su u časopisima, na televiziji, radiju, pa i na internetu. Prednost interneta u odnosu na televiziju, radio, časopise ili novine jest da njegovi korisnici nisu u potpunosti nepoznati, anonimni. Upravo nedostatak anonimnosti može pomoći ciljanom prikazivanju oglasa. Ako se određenom korisniku ciljano prikaže oglas koji bi mogao potaknuti njegov interes, veća je vjerojatnost da će korisnik izvršiti akciju, primjerice kliknuti na oglas ili kupiti proizvod. Personalizirani način oglašavanja u velikom je interesu tvrtkama, prodavačima i oglašivačima koji oglasima pokušavaju doći do što većeg broja korisnika ili kupaca. Ciljano oglašavanje je učinkovitije i isplativije.

U počecima internet oglašavanja su prema [1] oglašivači plaćali oglašavanje u obliku malog *bannera* na internet stranici na određeno vrijeme. Tek sredinom devedesetih dolazi do razvoja prvih programa koji su mogli pratiti koliko korisnika vidi određeni oglas i koliko ih je kliknulo na njega. Način naplaćivanja oglašavanja se zbog toga izmijenio jer se moglo plaćati ovisno o tome koliko je oglašivač puta želio da se njegov oglas prikaže korisnicima.

Prema [1] postoji više vrsta oglašavanja, a koji će se odabrati ovisi o vrsti internet stranice, potrebama oglašivača i o publici, tj. korisnicima kojima je oglas namijenjen. Sponzorstvo (engl. *Sponsorship*) je vrsta tradicionalnog oblika oglašavanja gdje oglašivač zakupi cijeli prostor na internet stranici samo za njegov oglas. Oglašavanje banerom (engl. *Banner Run*) je vjerojatno najosnovniji oblik tradicionalnog oglašavanja gdje oglašivač plaća najčešće za svakih tisuću prikaza oglasa korisniku u obliku *bannera* na internet stranici kroz određeni period vremena. Partnersko oglašavanje (engl. *Affiliate*) je vrsta oglašavanja u kojoj se koristi poseban kod za praćenje koji bilježi tko je izvršio kupnju, preko koje stranice je došao (na kojoj internet stranici je korisnik kliknuo oglas) i slično. Oglašivač na kraju plaća uslugu oglašavanja ovisno o tome koliko je ona bila uspješna (koliko je klikova ostvarila ili koliko je puta ostvarena kupnja i slično). Plaćanje po kliku (engl. *Pay-per-click*) vrsta oglašavanja je povezana s internet tražilicama i kontekstualnim oglašavanjem, a oglašivač obično plaća za svaki klik na oglas.

2. OPIS ZADATKA

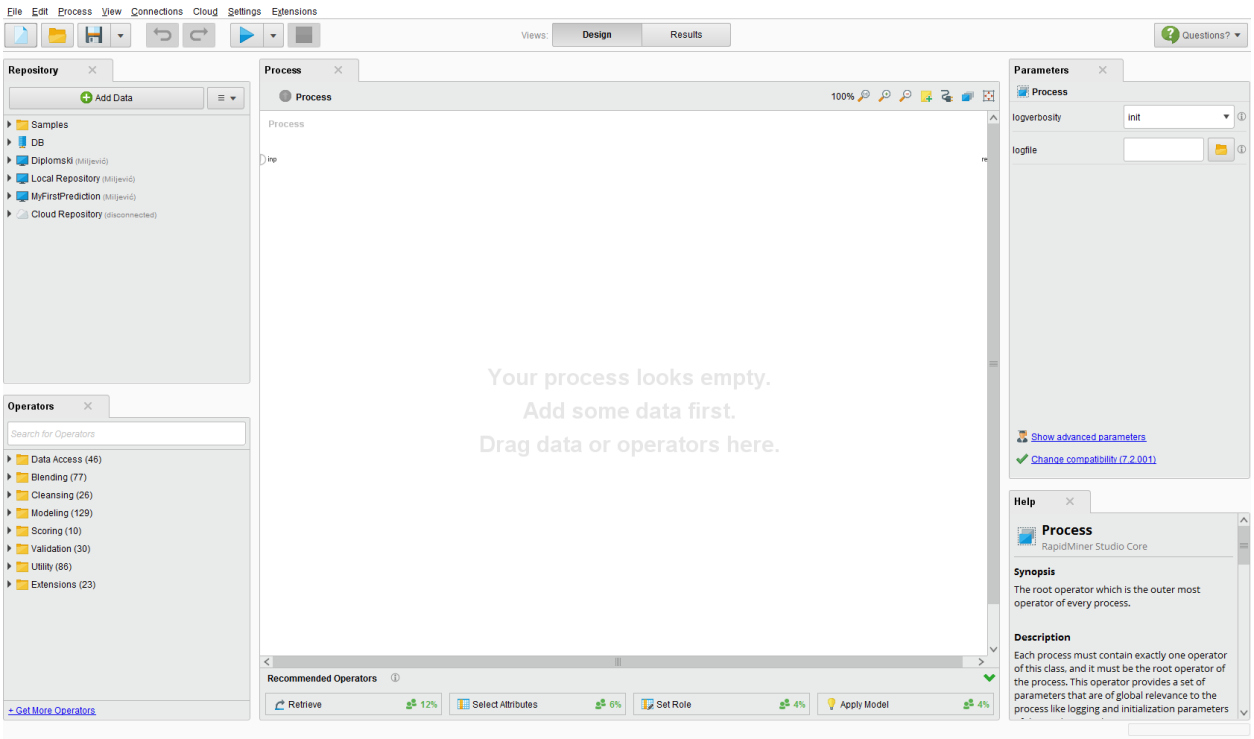
Zadatak ovog diplomskog rada je proučiti utjecaj različitih atributa na stopu klikanja odnosno klikovni postotak, skraćeno CTR (engl. *Click-through rate*). Neki atributi imaju pozitivan, negativan i neutralan utjecaj na CTR i svi oni utječu na rad sustava, odnosno obradu i čuvanje podataka. Potrebno je provesti istraživačku ili eksplorativnu analizu podataka (engl. *Exploratory Data Analysis* – EDA) te pronaći određene uzorke (engl. *pattern*) koji se pojavljuju među atributima, veze između ciljnog atributa i ostalih atributa i otkriti moguće anomalije (engl. *outlier*) u podacima koje negativno utječu na performanse sustava. Analizu je potrebno provesti na više razina jer različita granulacija može dati različite rezultate. Potrebno je otkriti koliko se može utjecati na stvarne sustave za isporuku oglasa odnosno kako ih se može poboljšati. Dodatno se mogu otkriti neki novi, izvedeni atributi koji bi mogli imati pozitivan utjecaj na ciljni atribut, odnosno CTR.

U radu je potrebno primijeniti skup različitih grafičkih tehnika kako bi se dobio bolji uvid u skup podataka. Alat koji je korišteni u svrhu izvršenja zadatka ovog diplomskog rada je *RapidMiner*.

2.1. Alati za obradu podataka

U ovom radu je za proučavanje, operacije i vizualizaciju rezultata korišten program *RapidMiner*. *RapidMiner* je prema [2] platforma koja omogućuje okruženje za strojno učenje, rudarenje podataka, prediktivnu analitiku i slično. Odlikuje se jednostavnim grafičkim sučeljem za povezivanje operatora kao što je navedeno u [3]. Ti operatori realiziraju različite funkcije transformacije podataka i indukcije modela. Ovaj program se prema [2] koristi u poslovne, istraživačke i obrazovne svrhe. Podržava pripremu podataka, vizualizaciju podataka, validaciju i optimizaciju te sadrži mnoštvo naprednih analitičkih rješenja kroz razne predloške zbog čega olakšava rad i smanjuje mogućnost pogreške te minimizira potrebe za pisanjem koda. Također se može proširiti korištenjem **R** i *Python* skripti te raznih *plugin*-ova.

Izgled sučelja *RapidMiner*-a je prikazan na slici Sl. 2.1.



Sl. 2.1. Izgled sučelja programskog alata *RapidMiner*.

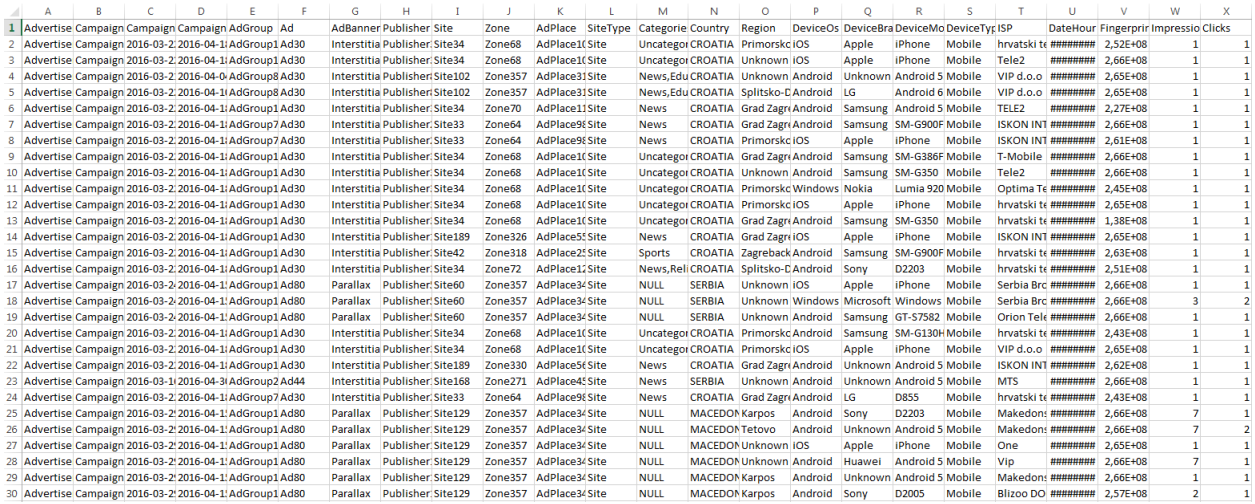
Osim *RapidMiner*-a za izvršenje zadatka se također moglo koristiti programsko okruženje **R**. Moguće ga je koristiti kao dodatak u *RapidMiner*-u pod nazivom *R Scripting*. Koristi se ukoliko postoji potreba za pisanjem vlastitog koda.

2.2. Izvorni skup podataka

Kod integriranja i transformiranja podataka česta je upotreba alata *Microsoft Excel*. Da bi se podaci mogli analizirati trebaju biti prikazani u obliku tablice. Redovi tablice prikazuju primjere (engl. *example*, *instance*), a svaki stupac prikazuje neki od atributa (engl. *attribute*) što znači da tablicu čine primjeri opisani atributima. Kako bi rezultati analize bili kvalitetni potrebno je prema [3] imati dovoljnu količinu primjera i dovoljnu količinu korisnih atributa.

Podaci korišteni u diplomskom radu su dobiveni iz stvarnog sustava za isporuku oglasa i dani su u CSV formatu. Sustav za isporuku oglasa prema [4] funkcionira tako da na osnovu različitih prikupljenih podataka o posjetitelju internet stranice ili korisniku mobilne aplikacije, kao što su lokacija korisnika, vrsta uređaja i operacijskog sustava koji korisnik koristi i drugih dostupnih podataka te parametara oglasa kao što su tip banera, starost oglasa, kategorija oglasnog mjesta i slično, vrši prediktivnu analitiku kako bi odabrao najbolji oglas, odnosno oglas koji ima najveću vjerojatnost da će biti kliknut.

Ukupno ima 253703 zapisa (redci tablice) i svaki zapis ili podatak je opisan s 24 atributa (stupci tablice). Na slici Sl. 2.2. nalazi se prikaz podataka u programu *Microsoft Excel*.



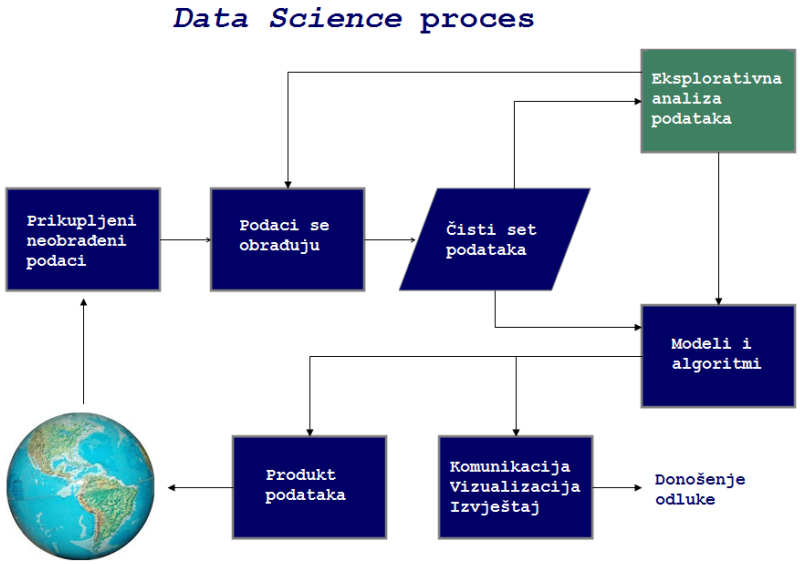
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Advertise	Campaign	Campaign	Campaign	AdGroup	Ad	AdBanner	Publisher	Site	Zone	AdPlace	SiteType	Categorie	Country	Region	DeviceOs	DeviceBra	DeviceMo	DeviceTyp	ISP	DateHour	Fingerprir	Impressio	Clicks
2	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Primorskc	iOS	Apple	iPhone	Mobile	hrvatski te	########	2,52E+08	1	1
3	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Unknown	iOS	Apple	iPhone	Mobile	Tele2	########	2,66E+08	1	1
4	Advertise	Campaign	2016-03-2	2016-04-0	AdGroup8	Ad30	Interstitia	Publisher	Site102	Zone357	AdPlace31	Site	News,Edu	CROATIA	Unknown	Android	Unknown	Android 5	Mobile	VIP d.o.o	########	2,65E+08	1	1
5	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup8	Ad30	Interstitia	Publisher	Site102	Zone357	AdPlace31	Site	News,Edu	CROATIA	Splitsko-D	Android	LG	Android 6	Mobile	VIP d.o.o	########	2,65E+08	1	1
6	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone70	AdPlace11	Site	News	CROATIA	Grad Zagr	Android	Samsung	Android 5	Mobile	TELE2	########	2,27E+08	1	1
7	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup7	Ad30	Interstitia	Publisher	Site33	Zone64	AdPlace9(	Site	News	CROATIA	Grad Zagr	Android	Samsung	SM-G900F	Mobile	ISKON INT	########	2,66E+08	1	1
8	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup7	Ad30	Interstitia	Publisher	Site33	Zone64	AdPlace9(	Site	News	CROATIA	Primorskc	iOS	Apple	iPhone	Mobile	ISKON INT	########	2,61E+08	1	1
9	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Grad Zagr	Android	Samsung	SM-G386F	Mobile	T-Mobile	########	2,66E+08	1	1
10	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Unknown	Android	Samsung	SM-G350	Mobile	Tele2	########	2,66E+08	1	1
11	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Primorskc	Windows	Nokia	Lumia 920	Mobile	Optima Te	########	2,45E+08	1	1
12	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Primorskc	iOS	Apple	iPhone	Mobile	hrvatski te	########	2,65E+08	1	1
13	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Grad Zagr	Android	Samsung	SM-G350	Mobile	hrvatski te	########	1,38E+08	1	1
14	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site189	Zone326	AdPlace5!	Site	News	CROATIA	Grad Zagr	iOS	Apple	iPhone	Mobile	ISKON INT	########	2,65E+08	1	1
15	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site42	Zone318	AdPlace2!	Site	Sports	CROATIA	Zagreback	Android	Samsung	SM-G900F	Mobile	hrvatski te	########	2,63E+08	1	1
16	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone72	AdPlace1:	Site	News,Reli	CROATIA	Splitsko-D	Android	Sony	D2203	Mobile	hrvatski te	########	2,51E+08	1	1
17	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site60	Zone357	AdPlace3(	Site	NULL	SERBIA	Unknown	iOS	Apple	iPhone	Mobile	Serbia Brc	########	2,66E+08	1	1
18	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site60	Zone357	AdPlace3(	Site	NULL	SERBIA	Unknown	Windows	Microsoft	Windows	Mobile	Serbia Brc	########	2,66E+08	3	2
19	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site60	Zone357	AdPlace3(	Site	NULL	SERBIA	Unknown	Android	Samsung	GT-S7582	Mobile	Orion Tele	########	2,66E+08	1	1
20	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Primorskc	Android	Samsung	SM-G130H	Mobile	hrvatski te	########	2,43E+08	1	1
21	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site34	Zone68	AdPlace1(	Site	Uncategor	CROATIA	Primorskc	iOS	Apple	iPhone	Mobile	VIP d.o.o	########	2,65E+08	1	1
22	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad30	Interstitia	Publisher	Site189	Zone330	AdPlace5(	Site	News	CROATIA	Grad Zagr	Android	Unknown	Android 5	Mobile	ISKON INT	########	2,62E+08	1	1
23	Advertise	Campaign	2016-03-1	2016-04-3	AdGroup2	Ad44	Interstitia	Publisher	Site168	Zone271	AdPlace4!	Site	News	SERBIA	Unknown	Android	Unknown	Android 5	Mobile	MTS	########	2,66E+08	1	1
24	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup7	Ad30	Interstitia	Publisher	Site33	Zone64	AdPlace9(	Site	News	CROATIA	Grad Zagr	Android	LG	D855	Mobile	hrvatski te	########	2,43E+08	1	1
25	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Karpos	Android	Sony	D2203	Mobile	Makedons	########	2,66E+08	7	1
26	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Tetovo	Android	Unknown	Android 5	Mobile	Makedons	########	2,66E+08	7	2
27	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Unknown	iOS	Apple	iPhone	Mobile	One	########	2,65E+08	1	1
28	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Unknown	Android	Huawei	Android 5	Mobile	Vip	########	2,66E+08	7	1
29	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Karpos	Android	Unknown	Android 5	Mobile	Makedons	########	2,66E+08	1	1
30	Advertise	Campaign	2016-03-2	2016-04-1	AdGroup1	Ad80	Parallax	Publisher	Site129	Zone357	AdPlace3(	Site	NULL	MACEDON	Karpos	Android	Sony	D2005	Mobile	Blizoo DO	########	2,57E+08	2	1

Sl. 2.2. Prikaz podataka i atributa u programu *Microsoft Excel*.

Atributi koji opisuju podatke su *Advertiser*, *Campaign*, *CampaignTimeStart*, *CampaignTimeEnd*, *AdGroup*, *Ad*, *AdBannerType*, *Publisher*, *Site*, *Zone*, *AdPlace*, *SiteType*, *Categories*, *Country*, *Region*, *DeviceOs*, *DeviceBrand*, *DeviceModel*, *DeviceType*, ISP, *DateHour*, *Fingerprint*, *Impressions* i *Clicks*. Kombinacijom određenih izvornih podataka, u radu su kreirani dodatni atributi kako bi se dobio bolji uvid u podatke, otkrili neki skriveni uzorci te što bolje provela analiza, a o njima će biti više riječi u sljedećim poglavljima.

3. EKSPLORATIVNA ANALIZA PODATAKA

Eksplorativna analiza podataka je, prema [5], pristup koji se koristi za sumiranje glavnih karakteristika podataka iz skupova podataka. Mnoge njene tehnike se koriste u analizama velikih skupova podataka (engl. *Big Data*) i u dubinskoj analizi podataka (engl. *Data Mining*). Grana znanosti koja koristi eksplorativnu analizu podataka je znanost o podacima (engl. *Data Science*). Tu se kombiniraju znanja iz statistike, obrade podataka, programiranja i vizualizacije podataka.



Sl. 3.1. Eksplorativna analiza podataka kao dio znanosti o podacima.

Veliki doprinos nastanku i razvoju eksplorativne analize podataka je dao američki matematičar John Wilder Tukey. Upravo zahvaljujući njemu je narasla popularnost primjene eksplorativne analize podataka što je s vremenom dovelo i do razvoja statističkih računalnih paketa među koje se ubraja i **S** koji je nastao u *Bell Labs*, a koji je poslužio kasnijem razvoju **R** programskog paketa.

Osnovni ciljevi eksplorativne analize podataka prema [6] su:

1. dobiti bolji uvid u skup podataka,
2. ispitati odnose među atributima,
3. identificirati zanimljive podskupove podataka,
4. pronaći veze između ciljnog i ostalih atributa.

Prema [7], tehnike koje se koriste za eksplorativnu analizu podataka mogu se podijeliti na grafičke tehnike i kvantitativne tehnike. Najviše se koriste grafičke tehnike kao što su iscrtavanje neobrađenih podataka (*histogram*, *bihistogram*, *block plot*, *run sequence plot*, *scatter plot*), iscrtavanje jednostavne statistike (*box plot*, *mean plot*, *standard deviation plot*) i druge te će neke od njih kao što su *histogram*, i *scatter plot* biti primijenjene u ovom radu.

Te tehnike koje čine eksplorativnu analizu omogućavaju prema [7]:

1. povećanje uvida u skup podataka,
2. otkrivanje temeljnih struktura,
3. izvlačenje važnih varijabli,
4. otkrivanje odudaranja i anomalija,
5. testiranje temeljnih pretpostavki,
6. razvijanje osnovnih (škrtih) modela,
7. određivanje optimalnih postavki faktora.

Grafičke metode omogućuju analitičarima da bolje otkriju strukturalne tajne u podacima i omogućuju nove i ponekad neočekivane uvide u podatke.

Priprema podataka prije analize je prema [3] vrlo važan korak jer kvaliteta rezultata analize u velikoj mjeri ovisi o tome. Priprema podataka iziskuje dosta vremena i zahtijeva poznavanje ciljeva analize te značenje podataka. Što je veći broj sakupljenih podataka, bolje se može izvršiti analiza u velikoj većini slučajeva. Ali osim kvantitete bitna je i kvaliteta, što bi značilo da je poželjno da podaci imaju poznate sve vrijednosti atributa i da primjeri dobro reprezentiraju cijelu populaciju koja je predmet istraživanja. U prikupljenim podacima se dio podataka koristi za analizu, a manji dio za nezavisnu verifikaciju rezultata.

Osim količine podataka, prema [3] dobro je i da je količina atributa što veća, ali isto treba voditi računa o kvaliteti atributa. Atributi su kvalitetniji što su bolje povezani s predmetom istraživanja. Atributi koji nisu povezani s predmetom istraživanja samo dodatno odužuju vrijeme potrebno za prikupljanje te mogu usporavati izvođenje procesa analize.

3.1. Opis atributa

Razlikuju se dvije osnovne grupe atributa, a to su atribut koji se analizira (ciljni atribut) i ostali atributi koji su u nekom odnosu s ciljnim atributom i pomoću kojih se objašnjava ciljni atribut. Osim te podjele, atributi se još mogu podijeliti i na numeričke i nominalne (lat. *nome* = ime) atribute. Numerički atributi su cjelobrojne ili realne vrijednosti i mogu se podijeliti još na redne, intervalne i razmjerne atribute, a nominalne atribute čine riječi, odnosno unaprijed definiran skup vrijednosti.

Kao što je prije navedeno, podaci korišteni za ovaj rad su opisani s 24 atributa. Osim tih atributa još neki atributi su dodani među kojima je jedan jako bitan za ovaj rad, a to je CTR koji je i ciljni atribut. Od 24 izvorna atributa, 18 atributa je nominalno, a 6 atributa su numerički. Od 6 numeričkih atributa, 3 atributa pripadaju intervalnim (datum i vrijeme), a ostali pripadaju razmjernim atributima.

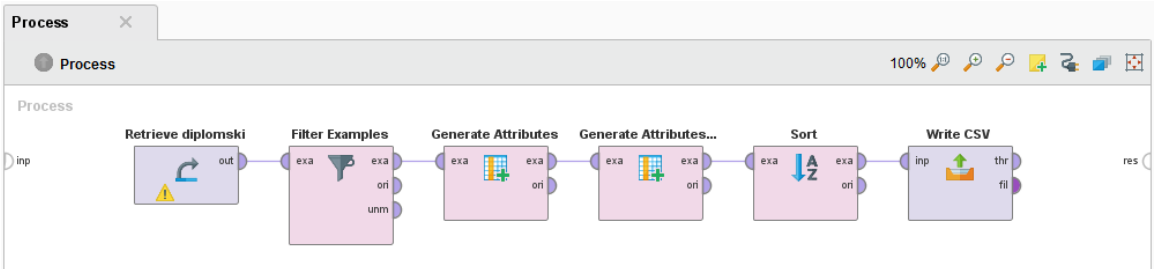
Većina nominalnih atributa može poprimiti velik broj različitih vrijednosti pa tako na primjer atribut *Region* može imati vrijednost naziva regije koja se može nalaziti u bilo kojoj državi koja je obuhvaćena u podacima ili *Fingerprint* koji je za svakog korisnika drukčiji. Tu su još i *Advertiser*, *Campaign*, *AdGroup*, *Ad*, *Publisher*, *Site* i drugi. Manji raspon vrijednosti mogu poprimiti atributi kao što su *Country*, *DeviceOs*, *DeviceBrand*, *SiteType*, *AdBannerType* i *DeviceType*.

Naziv i značenje svih izvornih atributa koji opisuju podatke su dani u tablici 3.1.

Tablica 3.1. Izvorni atributi i njihovo značenje.

Redni broj	Naziv atributa	Značenje atributa
1.	Advertiser	Oglašivač koji oglašava oglas
2.	Campaign	Naziv marketing kampanje
3.	CampaignTimeStart	Datum i vrijeme početka marketing kampanje
4.	CampaignTimeEnd	Datum i vrijeme završetka marketing kampanje
5.	AdGroup	Grupa oglasa (hijerarhijski viši nivo od Ad)
6.	Ad	Oglas
7.	AdBannerType	Vrsta banera u kojem se prikazuje oglas
8.	Publisher	Izdavač internet stranice gdje se oglašava oglas
9.	Site	Internet stranica
10.	Zone	Zona (hijerarhijski viši nivo od Site)
11.	AdPlace	Pozicija na kojoj je prikazan oglas
12.	SiteType	Opisuje da li se oglas prikazuje na internet stranici ili u aplikaciji
13.	Categories	Kategorija kojoj pripada pozicija na internet stranici ili u aplikaciji na kojoj se prikazuje oglas
14.	Country	Država u kojoj je oglas prikazan
15.	Region	Regija u državi
16.	DeviceOs	Operacijski sustav uređaja na kojem je prikazan oglas
17.	DeviceBrand	Marka uređaja na kojem je prikazan oglas
18.	DeviceModel	Model određene marke na kojem je prikazan oglas
19.	DeviceType	Vrsta uređaja na kojem je prikazan oglas
20.	ISP	Pružatelj internet usluga (engl. *Internet Service Provider*)
21.	DateHour	Datum i vrijeme kada je prikazan i kliknut oglas
22.	Fingerprint	Digitalni, virtualni otisak prsta
23.	Impressions	Ukupan broj prikaza oglasa
24.	Clicks	Broj klikova na oglas

Osim postojećih atributa uvedena su još 3 nova atributa, a to su CTR, *Duration* i *Hour*. Kreirani su pomoću *Generate Attributes* bloka u *RapidMiner*-u kao što je prikazano na slici Sl. 3.2.



Sl. 3.2. Kreiranje novih atributa.

CTR ili klikovni postotak je prema [8] omjer korisnika koji su kliknuli određeni oglas (reklamu) i ukupnog broja korisnika koji su posjetili određenu internet stranicu, reklamu ili e-mail, odnosno ukupnog broja korisnika kojima je oglas prikazan. Izračunava se prema formuli (3-1). Ova metoda se često koristi u internet oglašavanju kako bi se ocijenilo koliko je neka marketing kampanja uspješna. Postoje i druge metode za mjerenje uspješnosti nekog oglasa, ali ova metoda je najjednostavnija i najraširenija. Vrijednost CTR-a se izražava u postotku.

$$CTR = \frac{broj\ klikova}{ukupan\ broj\ koliko\ je\ puta\ oglas\ prikazan} * 100 \qquad (3\text{-}1)$$

U formuli (3-1) za izračun CTR-a atribut *Clicks* iz podataka predstavlja broj klikova, a ukupan broj koliko je puta oglas prikazan predstavlja atribut *Impressions*.

Budući da prosječne vrijednosti CTR-a za internet oglase u pravilu nisu jako visoke, nastoji se procijeniti interese korisnika kako bi im se dostavili oglasi koji bi ih mogli zanimati te bi time postojala veća vjerojatnost da će korisnik kliknuti na oglas.

Atribut *Duration* predstavlja trajanje marketing kampanje. Računa se u *RapidMiner*-u kao razlika između kraja i početka marketing kampanje te se izražava u sekundama. Izraz kojim se računa ovaj atribut u *RapidMiner* bloku *Generate Attributes* je prikazan pod (3-2).

```
abs(date_diff(CampaignTimeEnd,CampaignTimeStart))/1000
```
(3-2)

Atribut *Hour* je izveden iz atributa DateHour koji predstavlja datum i vrijeme kada je oglas prikazan. Ovaj atribut predstavlja točno vrijeme prikazivanja oglasa, a izražen je u satima, minutama i sekundama. Izraz kojim je izveden ovaj atribut u *RapidMiner* bloku *Generate Attributes* je prikazan pod (3-3).

```
date_str(DateHour,DATE_FULL, DATE_SHOW_TIME_ONLY)
```
(3-3)

Sva tri nova izvedena atributa spadaju u numeričke atribute, a zajednički s opisima su prikazani u tablici 3.2.

Tablica 3.2. Izvedeni atributi i njihovo značenje.

Redni broj	Naziv atributa	Značenje atributa
1.	CTR	Broj koji predstavlja postotak koliko je puta oglas kliknut s obzirom na to koliko je puta bio prikazan
2.	Duration	Trajanje marketing kampanje
3.	Hour	Vrijeme u satima kada je oglas prikazan

3.2. Problem nedostajućih vrijednosti

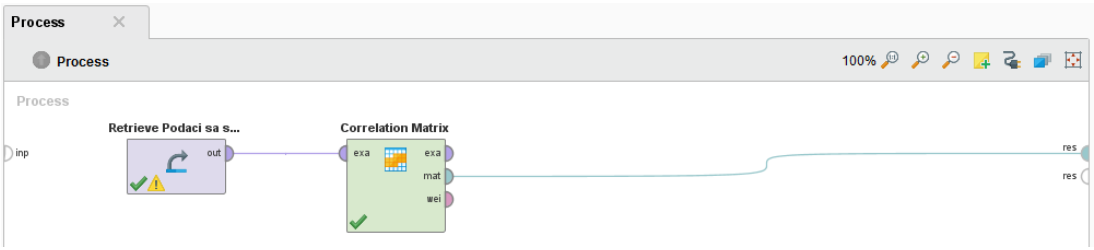
Ponekad se može dogoditi da u nekom podatku postoji atribut koji nema vrijednost, odnosno da mu je vrijednost nepoznata. Razlog tome prema [3] može biti da atribut nije izmjeren, nije poznat ili nije upisan. Količinu atributa s nedostajućim vrijednostima bi bilo dobro svesti na minimum kako bi sustav što bolje funkcionirao.

U skupu podataka koji se koristi u ovom diplomskom radu ima nekoliko atributa koji poprimaju nepoznatu vrijednost. Tu spadaju atributi *Categories*, *Region* i *DeviceBrand*. Uobičajena praksa je da se za vrijednost nepoznatog numeričkog atributa postavi srednja vrijednost tog atributa izvedena iz ostalih podataka koji imaju vrijednost tog atributa, a za vrijednost nepoznatog nominalnog atributa unese određeni string i to za sve nepoznate vrijednosti istog atributa isti string. U slučaju podataka obrađenima u ovom radu svi atributi koji imaju nepoznate vrijednosti su nominalni atributi i već imaju zamijenjene vrijednosti pa se za nepoznatu vrijednost atributa *Categories* koristi string NULL, za nepoznatu vrijednost atributa *Region* koristi string *Unknown*, a za nepoznatu vrijednost atributa *DeviceBrand* također string *Unknown*.

3.3. Korelirani atributi

Ponekad se može dogoditi na između nekih atributa postoji jaka veza što uzrokuje to da se promjenom vrijednosti jednog atributa mijenja vrijednosti drugog atributa (povećava ili smanjuje). Prema [9], takvi atributi se nazivaju korelirani atributi i pomoću njih se može otkriti koji atributi utječu na ciljni atribut. Korelirani atributi mogu predstavljati problem jer tijekom procesa istraživanja dolazi do značajnog utjecaja pojedinog atributa što dovodi do formiranja nepouzdanih i nestabilnih modela.

Pomoću programskog alata *RapidMiner* moguće je otkriti postojanje koreliranih atributa u skupu podataka i to iscrtavanjem dijagrama rasipanja (engl. *Scatter plot*). Također je odnose između atributa prema [9] moguće ispitati i matricom korelacije atributa (engl. *correlation matrix*) koja se dobije upotrebom *RapidMiner* bloka *Correlation Matrix* kako je prikazano na slici Sl. 3.3.



Sl. 3.3. *Correlation Matrix* blok u *RapidMiner*-u.

Iz matrice korelacija koja se nalazi na slici Sl. 1. u prilogu P.2. je vidljivo da postoje određene veze među atributima. Prema [10], za iznos korelacije od 0 do ±0,25 smatra se da nema povezanosti. Za vrijednosti od ±0,26 do ±0,50 smatra se da postoji slaba povezanost. Kod vrijednosti od ±0,51 do ±0,75 postoji umjerena do dobra povezanost, a za vrijednosti iznad ±0,76 povezanost je vrlo dobra do izvrsna.

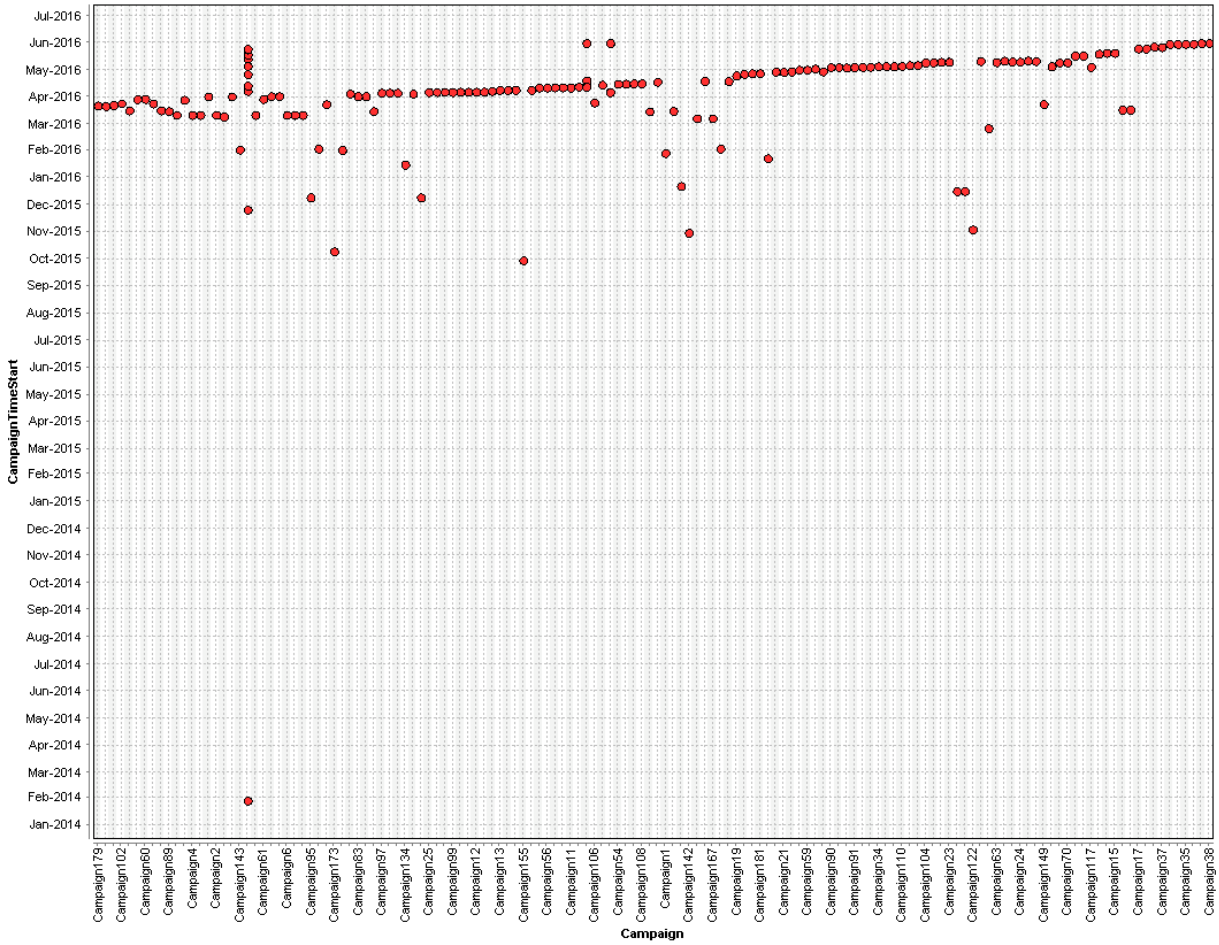
Najjače veze su među atributima koji predstavljaju kampanju, početak i kraj kampanje, grupu oglasa i *DateHour* atributa. Međusobna povezanost atributa koji predstavljaju datum i vrijeme (*CampaignTimeStart*, *CampaignTimeEnd* i *DateTime*) je logična. U ovom razmatranju će se gledati samo atributi između kojih postoji umjerena i izvrsna povezanost, odnosno gdje je iznos korelacije iznad 0,5.

Također, osim matrice korelacija, za vizualizaciju odnosa između atributa će se koristiti dijagram rasipanja. Dijagram rasipanja se prema [11] koristi za otkrivanje odnosa, odnosno povezanosti dviju varijabli koja može biti linearna, nelinearna, kvadratna i slično. Takav odnos se očituje u svakom iscrtavanju strukture koja nije slučajna.

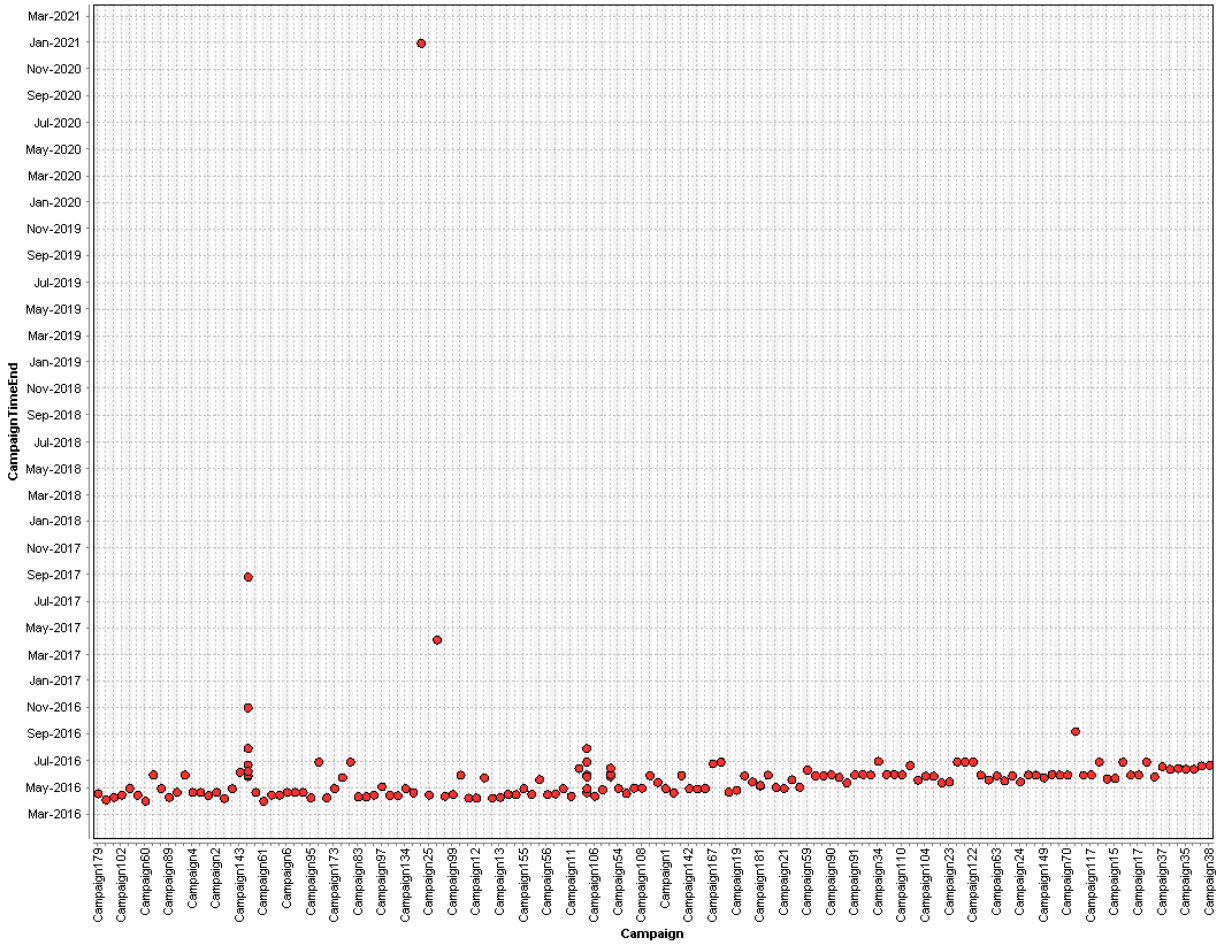
Na dijagramu se prikazuje vrijednost X u odnosu na vrijednost Y. Na x-osi se nalazi varijabla X koja je najčešće neka varijabla za koju sumnjamo da bi mogla biti povezana s odzivom koji se nalazi na y-osi kako je navedeno u [11].

Dijagram rasipanja prema [11] daje odgovor na pitanje jesu li varijable X i Y povezane i kako te ima li iznimaka. Koristan je za otkrivanje povezanosti među varijablama, ali ako povezanost postoji, on najčešće ne može dokazati uzrok i posljedicu.

Veza između atributa *Campaign* i *CampaignTimeStart*, odnosno između atributa *Campaign* i *CampaignTimeEnd* je prikazana pomoću dijagrama rasipanja na slikama Sl. 3.4. i Sl. 3.5. Ovi atributi su povezani zbog toga što u svim podacima s istom vrijednosti atributa *Campaign* kampanja počinje u isto vrijeme i završava u isto vrijeme jer se radi o istoj kampanji, a i većina kampanja je počela i završila u intervalu između dva mjeseca (ožujak i lipanj), jedna za drugom.

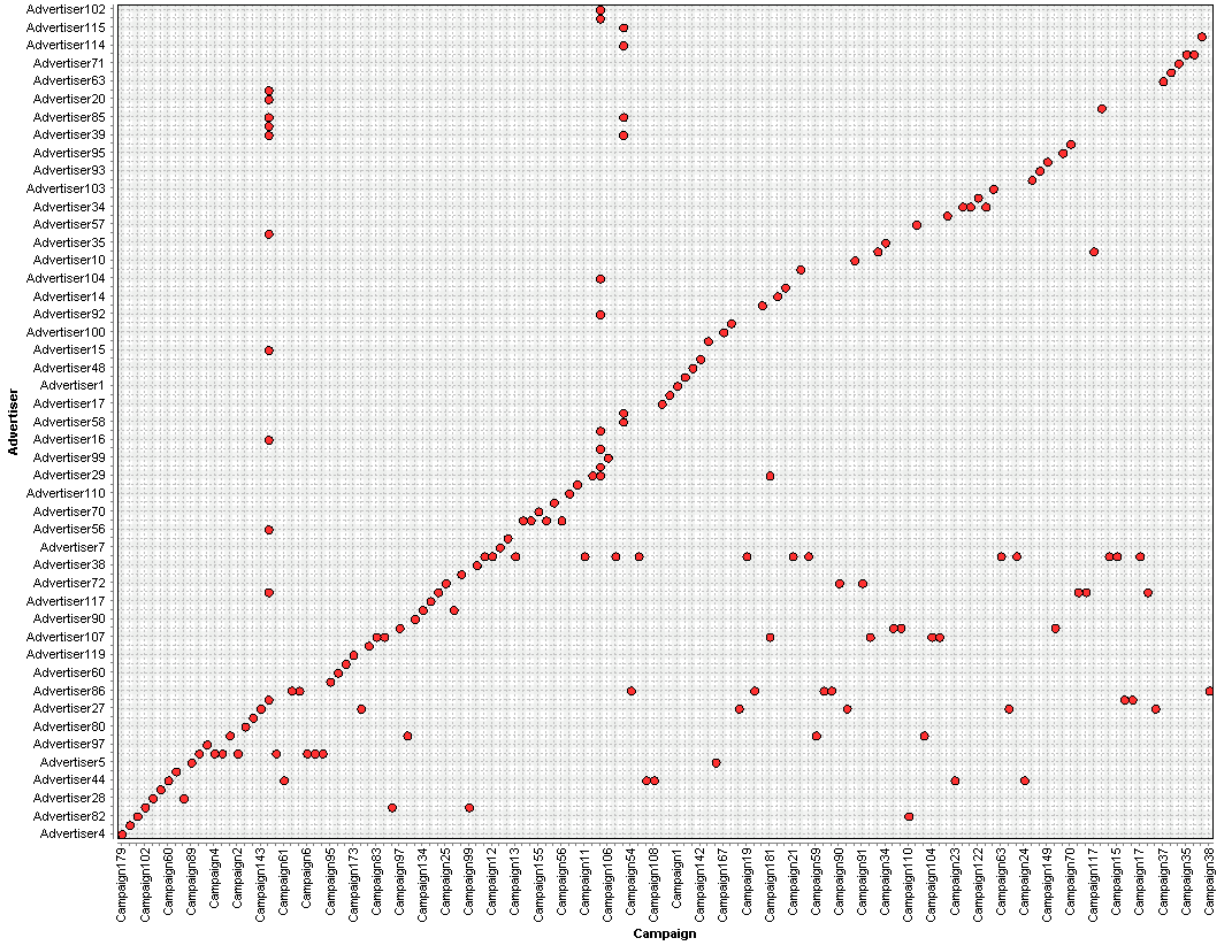


Sl. 3.4. Dijagram rasipanja za atribute *Campaign* i *CampaignTimeStart*.



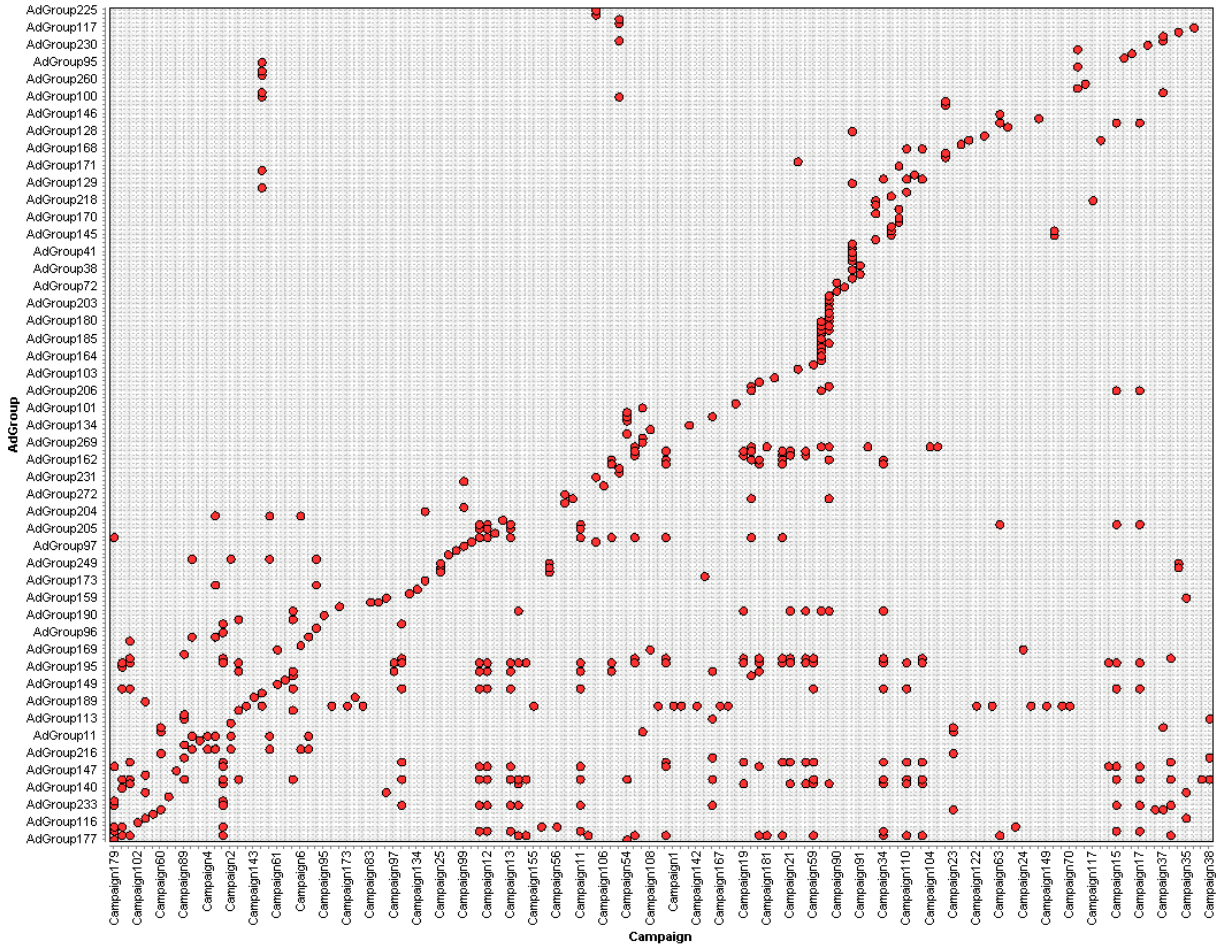
Sl. 3.5. Dijagram rasipanja za atribute *Campaign* i *CampaignTimeEnd*.

Iz matrice korelacija je također vidljiva dosta jaka veza između atributa *Advertiser* i *Campaign*, a to se vidi i na dijagramu na slici Sl. 3.6. Razlog tome je također taj što je određeni oglašivač u većini slučajeva povezan za određenu kampanju u kojoj se oglas oglašava, iako ima kampanja u kojima oglašava više oglašivača kao i oglašivača koji oglašavaju u više različitih kampanja.



Sl. 3.6. Dijagram rasipanja za atribute *Advertiser* i *Campaign*.

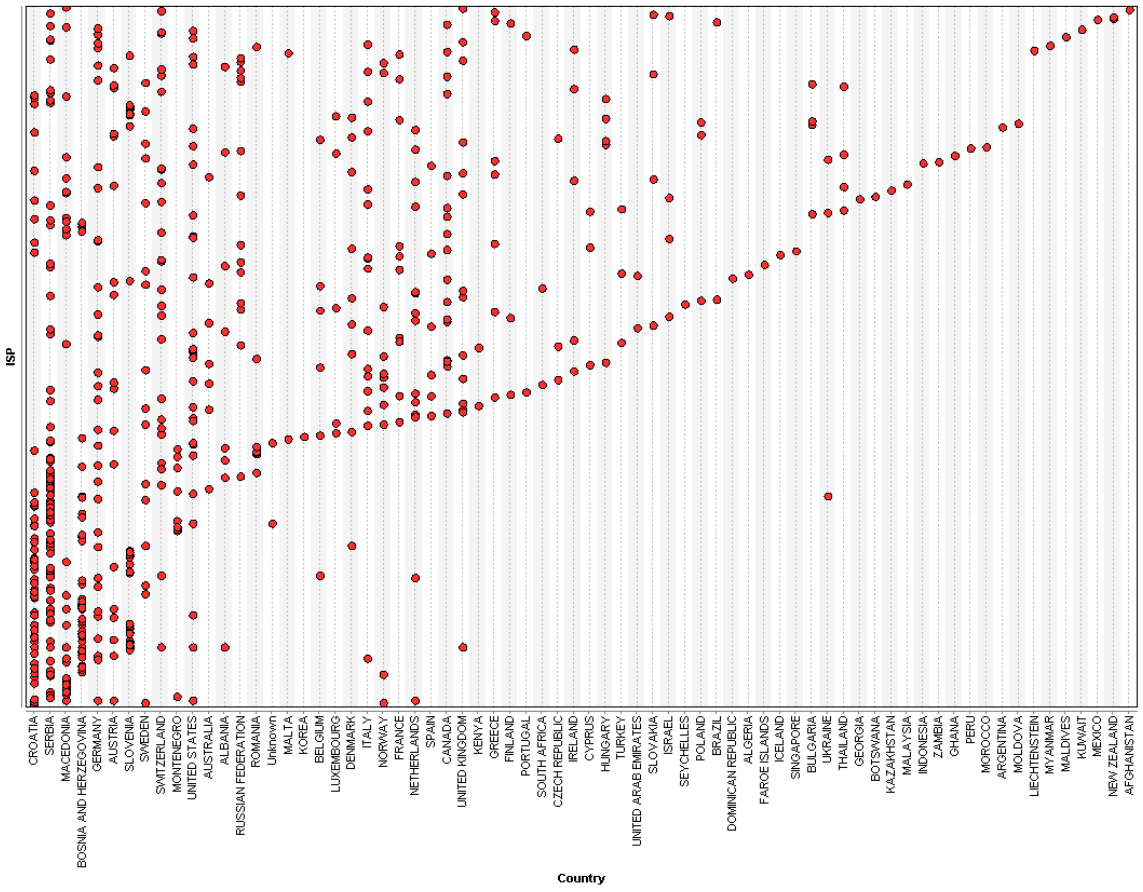
Malo slabija veza nego kod prethodnih atributa pojavljuje se i između atributa *AdGroup* i *Campaign*, a prikazana je na slici Sl. 3.7. Razlog povezanosti je sličan kao i u prethodnom primjeru.



Sl. 3.7. Dijagram rasipanja za atribute *AdGroup* i *Campaign*.

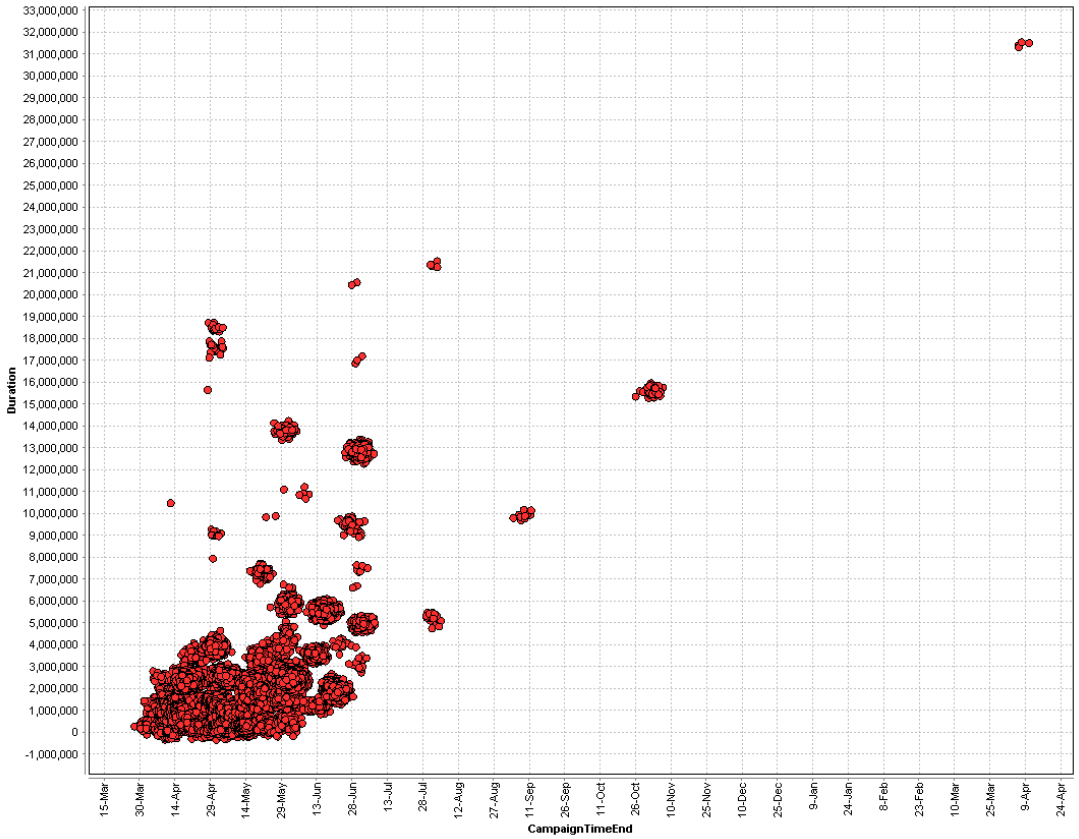
Slične su povezanosti i između atributa *Site* i *AdPlace* te S*iteType* i *Country*.

Još jedna od logičnih veza među atributima je veza između atributa *Country* i ISP zbog toga što je određeni pružatelj internet usluga povezan sa određenom državom u kojoj pruža usluge uz par iznimki gdje isti pružatelj pruža usluge u više država. Dijagram je prikazan na slici Sl. 3.8.



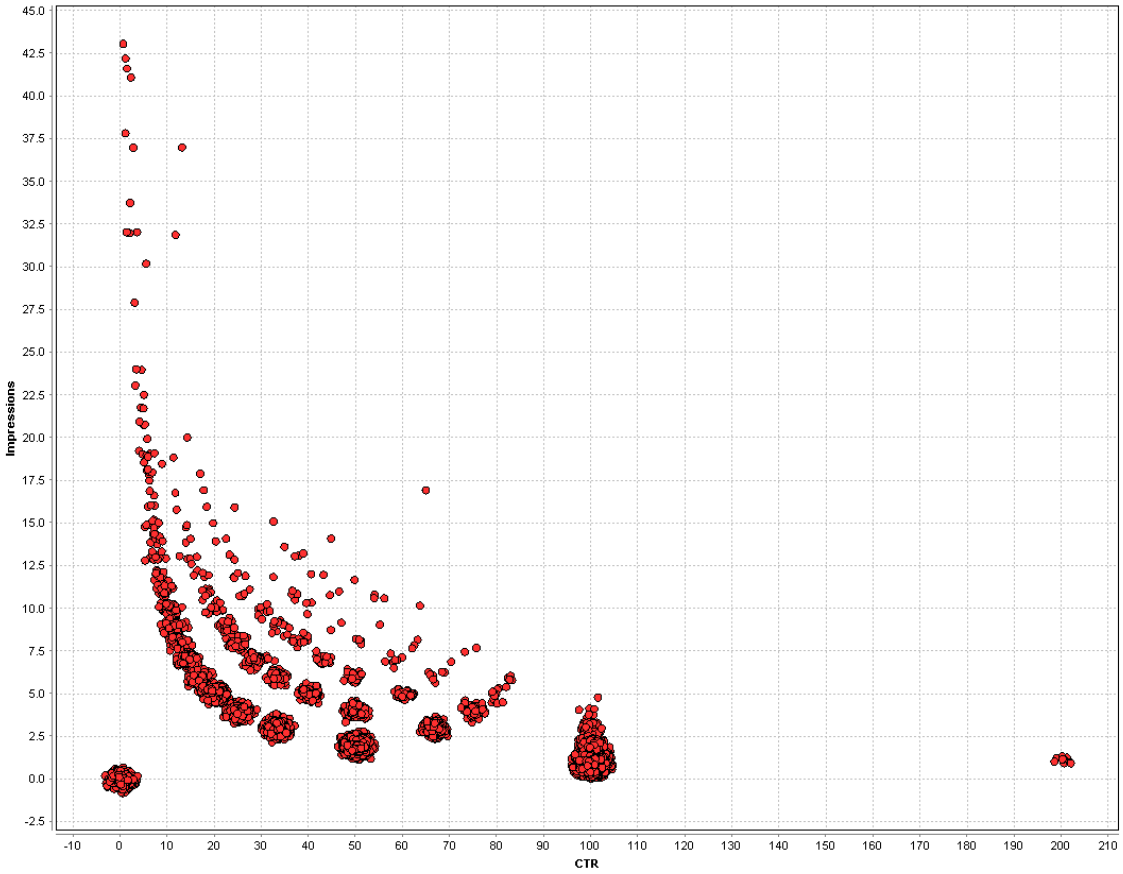
Sl. 3.8. Dijagram rasipanja za atribute *Country* i ISP.

Atribut *Duration* ima dosta jaku vezu sa atributom *CampaignTimeEnd*, a razlog tome je taj da što je kraj kampanje dalji, to je vrijednost atributa *Duration* veća jer kampanja traje duže. Odnos ova dva atributa je prikazan na slici Sl. 3.9. Također razlog njihove povezanosti je i taj što je atribut *Duration* izveden iz atributa *CampaignTimeStart* i *CampaignTimeEnd*.



Sl. 3.9. Dijagram rasipanja za atribute *Duration* i *CampaignTimeEnd*.

Posljednji atributi između kojih je možda i najzanimljivija veza su atributi *Impressions* i CTR. Ova dva atributa su izrazito negativno korelirana. Odnos između njih je prikazan dijagramom na slici Sl. 3.10. Iz dijagrama je vidljivo da što je veća vrijednost atributa *Impressions* za neki podatak, to je vrijednost atributa CTR niža. Naravno, njihova povezanost proizlazi i iz toga što je atribut CTR nastao iz atributa *Impressions* i *Clicks*. Osim toga vidljivo je i da je najveći broj podataka koji imaju CTR 100 prikazan samo jednom.



Sl. 3.10. Dijagram rasipanja za atribute *Impressions* i CTR.

3.4. Istraživanje odnosa ciljnog atributa i ostalih atributa

Kod istraživanja odnosa ciljnog i ostalih atributa, prema [9], česta je upotreba koeficijenta korelacije. Upotreba koeficijenta korelacije je bitna kako bi se bolje razumjeli podaci. U tu svrhu se izrađuju matrice korelacije za sve parove atributa. U ovom radu postoji samo jedan ciljni atribut, a to je CTR koji je izveden. Veći CTR znači i bolji rezultat. Odnosi između ciljnog atributa i ostalih atributa prikazani su u tablici Tablica 3.3.

Iz tablice se vidi da atribut CTR ima najveću vrijednost korelacije s atributom *Impressions* dok sa svim ostalim atributima vrijednost korelacije se kreće između 0,08 i -0,18 što znači da nema povezanosti među tim atributima. S većinom atributa atribut CTR ima negativnu korelaciju što bi značilo da se povećanjem vrijednosti tih atributa vrijednost atributa CTR smanjuje.

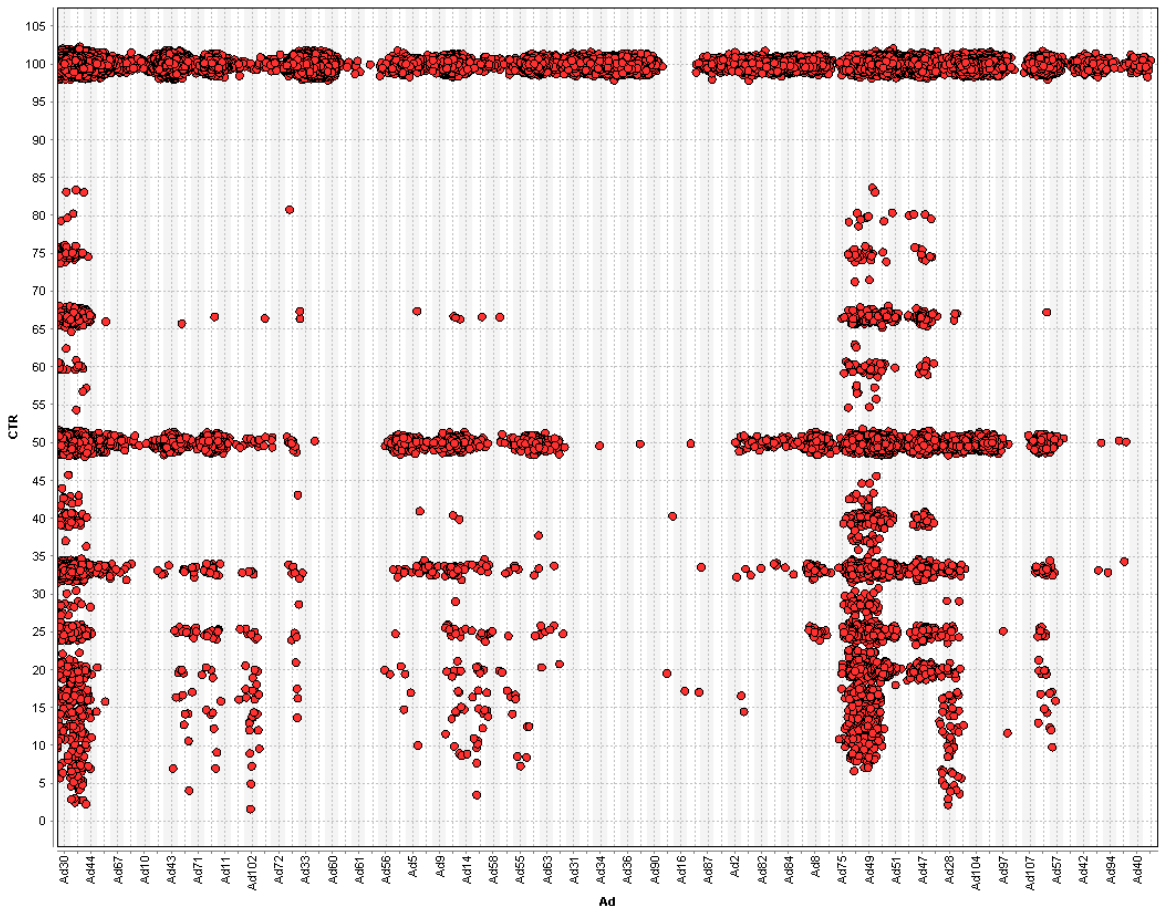
Tablica 3.3. Iznosi korelacije atributa CTR s ostalim atributima.

Matrica korelacije za atribut CTR	
Advertiser	-0,104
Campaign	-0,094
CampaignTimeStart	-0,059
CampaignTimeEnd	-0,010
AdGroup	0,037
Ad	-0,179
AdBannerType	-0,113
Publisher	-0,091
Site	-0,075
Zone	0,076
AdPlace	0,044
SiteType	0,014
Categories	-0,090
Country	-0,015
Region	0,005
DeviceOs	-0,020
DeviceBrand	0,021
DeviceModel	0,008
DeviceType	-0,058
ISP	-0,005
DateHour	-0,056
Fingerprint	-0,023
Impressions	-0,748
Clicks	-0,147
Duration	0,039
Hour	-0,012

3.5. Istraživanje nominalnih atributa

Većina nominalnih atributa može poprimiti jako velik broj vrijednosti pa je zbog toga jako teško promatrati njihove odnose sa ciljnim atributom. Kao primjer prikazan je dijagram rasipanja

za atribute oglas i CTR na slici Sl. 3.11. Ostali atributi koji poprimaju manji broj vrijednosti bit će prikazani grafički.



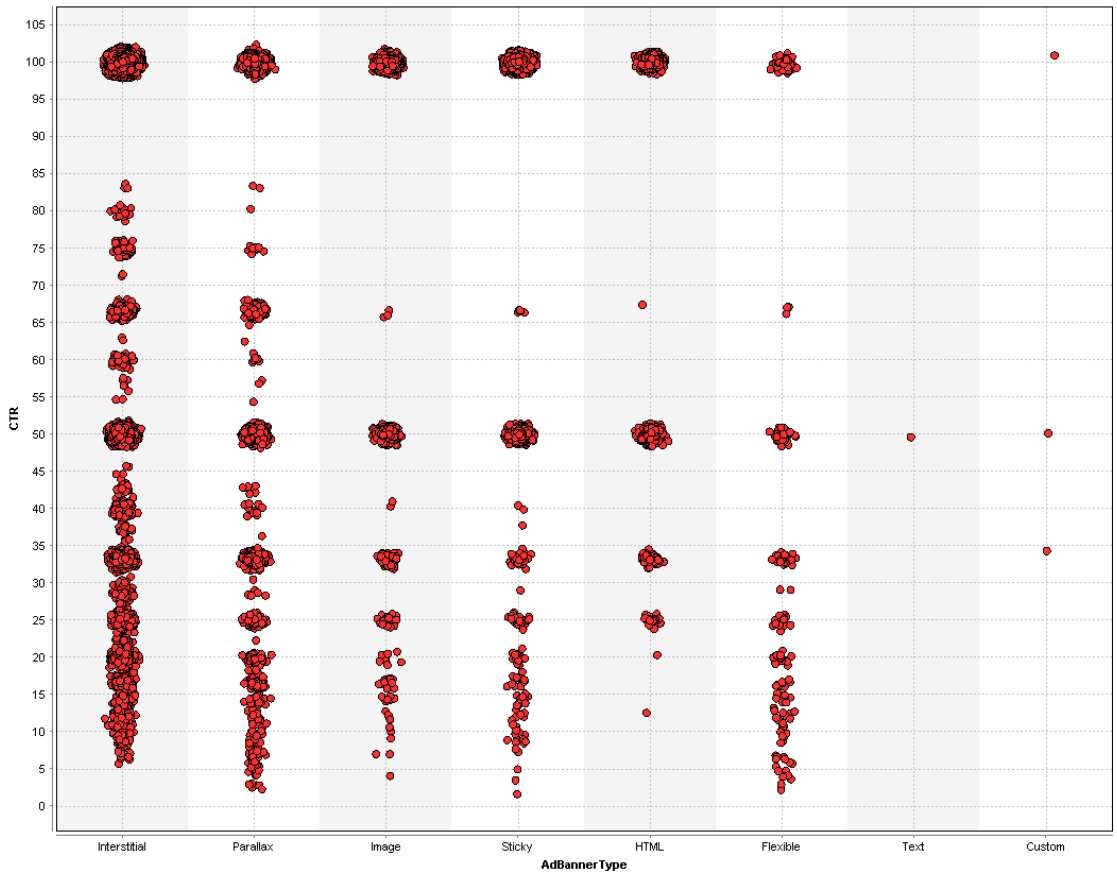
Sl. 3.11. Dijagram rasipanja za atribute CTR i *Ad*.

Iz dijagrama sa slike Sl. 3.11. se može vidjeti da se određeni oglasi pojavljuju u više podataka te da neki oglasi vrlo često poprimaju vrijednosti za CTR ispod 50. Zbog velike količine različitih vrijednosti koje može poprimiti atribut oglas teško se može iščitati koji su to oglasi, a i kada se iščita, opet se ne zna ništa više o njima osim njihovog naziva u podacima.

Prvi od atributa s manjim rasponom vrijednosti je atribut vrsta banera koji se uspoređuje s ciljnim atributom CTR. U podacima se pojavljuje 8 tipova banera, a to su *Interstitial*, *Parallax*, *Image*, *Sticky*, HTML, *Flexible*, *Text* i *Custom*. *Interstitial* je prema [12] statički tip banera koji prekriva cijeli sadržaj na ekranu. Kada se korisniku pojavi ovakav tip banera on može kliknuti na njega ili ga zatvoriti i nastaviti pregledavati internet stranicu ili koristiti aplikaciju. *Parallax* je prema [13] tip banera koji se ne miješa sa sadržajem koji korisnik gleda ili koristi već se pojavljuje kao dio veće pozadine čiji se dijelovi otkrivaju kako korisnik pomiče stranicu gore, dolje. *Text* i

Image su prema [13] stari tipovi statičkih banera koji prikazuju tekst ili sliku u malom okviru na internet stranici ili u aplikaciji. *Sticky* su prema [14] statički baneri fiksnih dimenzija koji se nalaze negdje na ekranu (gore, dolje ili sa strane) i ne mijenjaju poziciju pomicanjem stranice gore, dolje. *Flexible* je dinamički, rastezljivi tip banera, a *Custom* je tip banera izrađen po narudžbi, odnosno prema željama oglašivača.

Budući da je korišten dijagram rasipanja moguće je dobiti i okvirni uvid u to koliko podataka spada u neki atribut. Iz dijagrama sa slike Sl. 3.12. se vidi da najviše oglasa spada u *Interstitial* i *Parallax* tip banera. Najrjeđi tipovi su *Text* i *Custom* pa je za njih jako teško i donijeti bilo kakve zaključke. Dva najbrojnija tipa ostvaruju i najbolji CTR u intervalu vrijednosti za CTR od 55 do 95. Također iz dijagrama se može zaključiti da HTML tip ima veći najniži CTR spram ostalih, odnosno vrijednosti kreću od vrijednosti iznad 10. Također i kod najbrojnijeg tipa banera, *Interstatial*-a vrijednosti za CTR kreću od oko 5%. Najviše je oglasa koji imaju CTR vrijednost 100, a zatim onih koji imaju vrijednost 50.

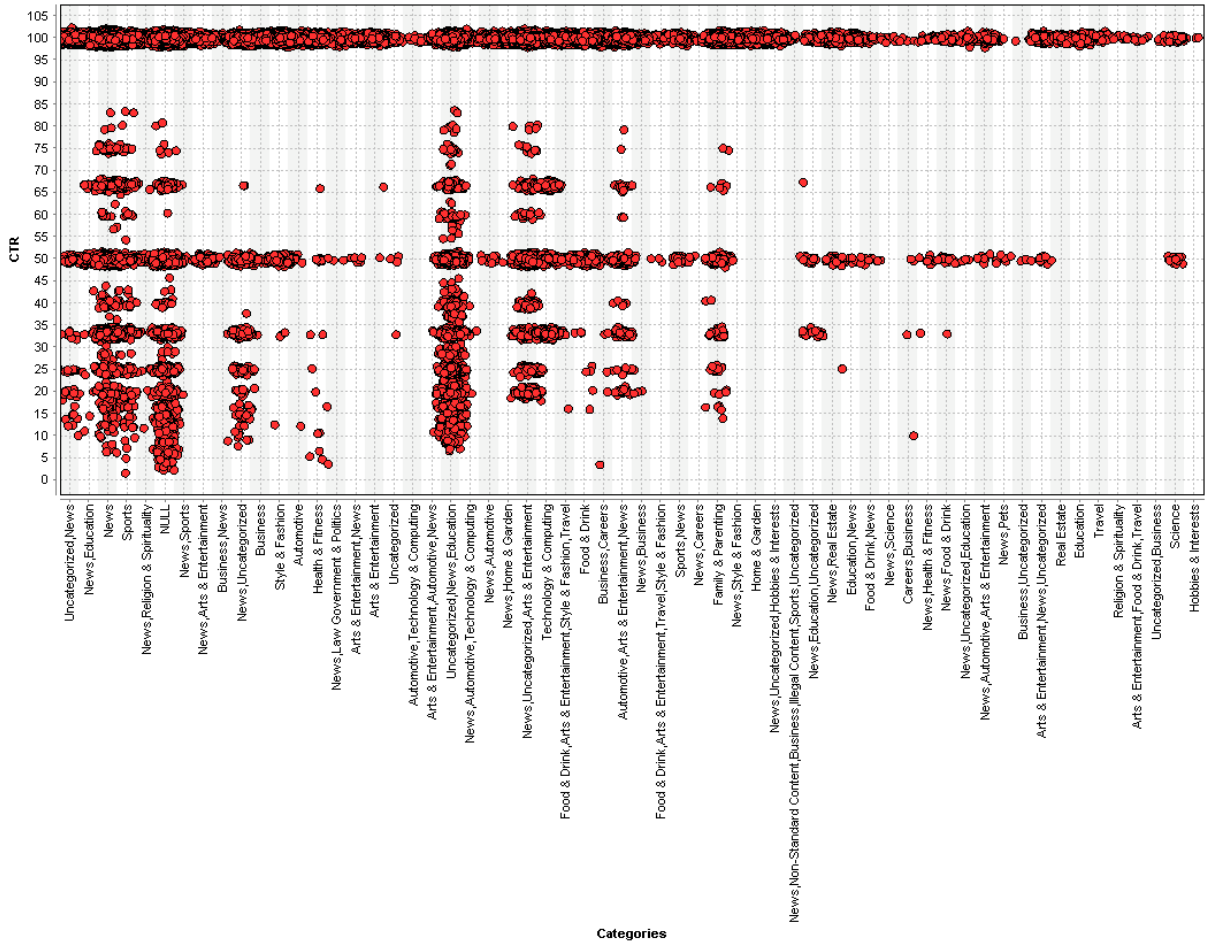


Sl. 3.12. Dijagram rasipanja za atribute CTR i *AdBannerType*.

Drugi atribut je atribut kategorije kojoj pripada pozicija na internet stranici ili u aplikaciji na kojoj se prikazuje oglas. Iako ovaj atribut može poprimiti vrlo širok spektar vrijednosti, ipak je odlučeno staviti ga u razmatranje zbog toga što su vrijednosti koje poprima razumljive pa se iz toga može možda izvući neka zanimljiva pretpostavka. Na dijagramu je vidljivo da se kod nekih kategorija češće pojavljuju vrijednosti CTR-a ispod 50 dok su kod nekih vrijednosti uglavnom 50 i 100. Pregledom dijagrama ustanovljeno je da se veće vrijednosti CTR-a pojavljuju najčešće kod kategorija kao što su "*Travel*", "*Real Estate*", "*Religion & Spirituality*", "*Arts & Entertainment, Food & Drink, Travel*", "*Education*", "*News, Science*", "*Home & Garden*" i dr. Lošije vrijednosti CTR-a su češće kod kategorija kao što su "NULL", "*News*", "*Sport*", "*Uncategorized, News, Education*", "*Business, Careers", Family & Parenting*" i dr. Naravno da lošijem CTR-u svakako kod tih kategorija doprinosi i to što su brojniji podaci pod tim kategorijama, a i većina je kategorija mješavina više pojedinačnih kategorija.

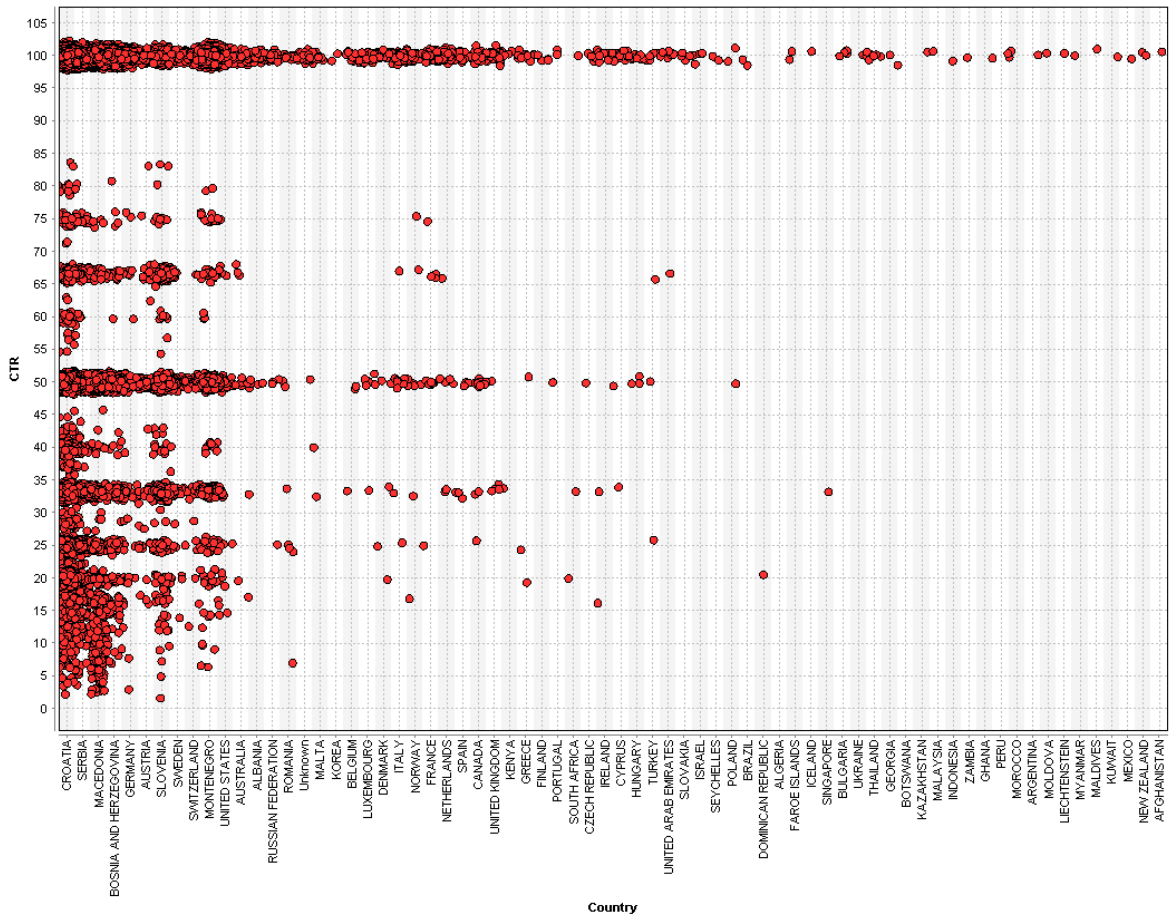
Do ovakvih rezultata je najvjerojatnije došlo zbog različitih tipova korisnika koji koriste aplikacije, odnosno posjećuju internet stranice određenog sadržaja. Korisnici najviše koriste internet radi zabave, čitanja vijesti, obrazovanja i traženja pomoći, a među njima je najviše mladih. Upravo zbog toga bi kategorije koje obuhvaćaju zabavu, vijesti, sport, obrazovanje, hobije, tehnologiju i slično trebale privući najveći broj korisnika.

Kako je već na prethodnim dijagramima bilo prikazano da je vrijednost CTR-a najčešće 100 kada je broj prikaza 1 tako je teško zaključiti da su ovdje izdvojene kategorije stvarno uspješnije. Upravo zbog toga ovaj atribut će se malo dublje istražiti u sljedećim potpoglavljima.



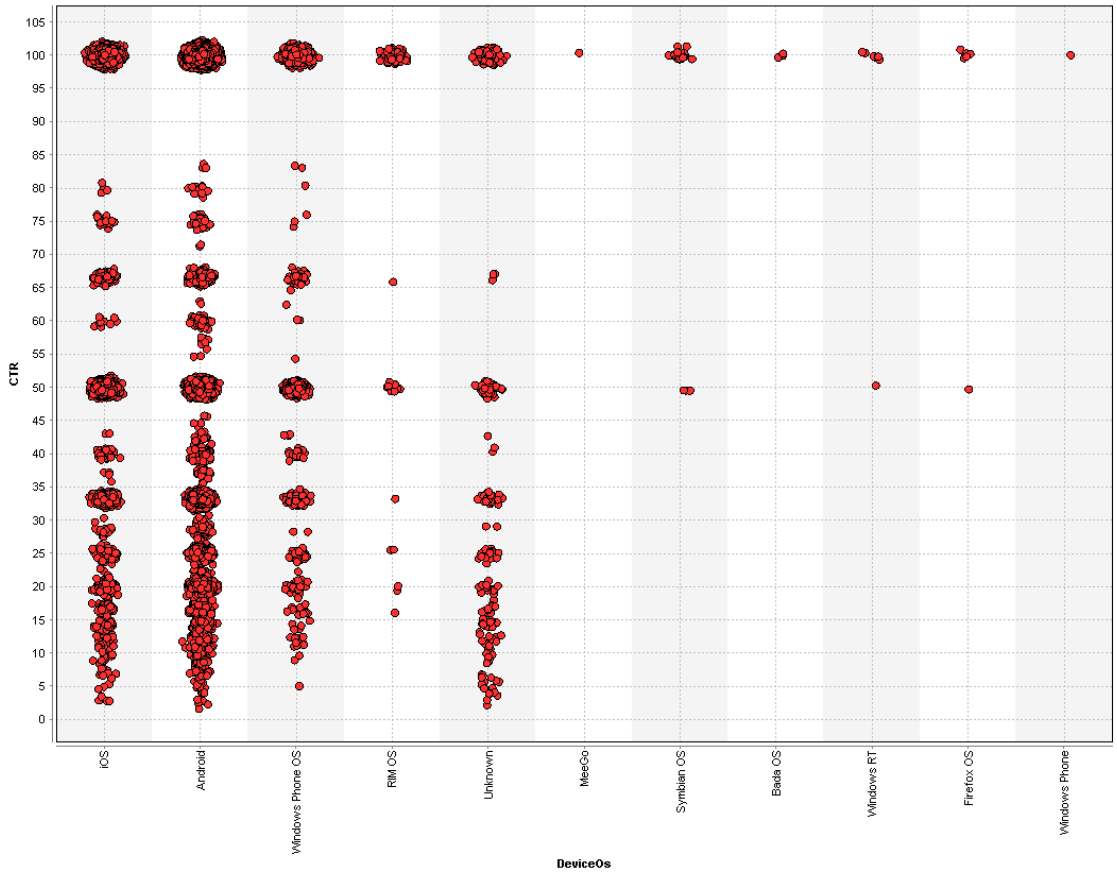
Sl. 3.13. Dijagram rasipanja za atribute CTR i *Categories*.

Sljedeći nominalni atribut koji se uspoređuje s atributom CTR je atribut koji predstavlja državu u kojoj je oglas prikazan. Atribut države u kojoj je oglas prikazan slično kao i atribut kategorije može poprimiti poprilično velik broj vrijednosti ali su vrijednosti lako razumljive. Također se može primijetiti kako se velika većina podataka odnosi na države jugoistočne Europe (Hrvatska, Bosna i Hercegovina, Srbija, Slovenija, Crna Gora i Makedonija). Zbog toga će se pri detaljnijem proučavanju odnosa između atributa CTR i atributa države u kojoj je oglas prikazan u potpoglavlju 3.7. posvetiti više pažnje tim državama. Prikaz će tada biti puno jasniji.



Sl. 3.14. Dijagram rasipanja za atribute CTR i *Country*.

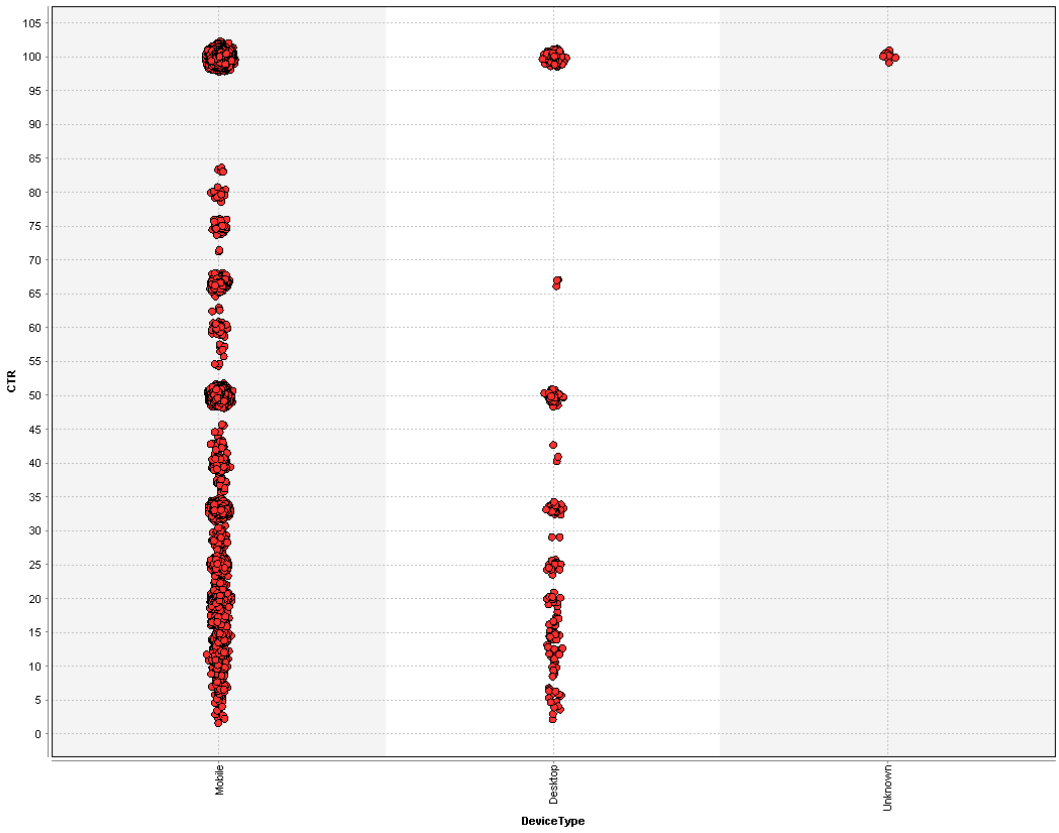
Sljedeći atribut je operacijski sustav uređaja na kojem je prikazan oglas. Iz dijagrama sa slike Sl. 3.15. se vidi da većinom 4 operacijska sustava dominiraju u podacima. To su iOS, *Android*, *Windows Phone* OS i *Unknown* koji označava nepoznati operacijski sustav. Iz ovakvog prikaza se ne može primijetiti utjecaj određenog operacijskog sustava na visinu CTR-a. Poprilično je ujednačeno sve raspoređeno. Zbog toga je ovaj atribut detaljnije proučen i kombiniran s drugim atributima u potpoglavlju 3.7.



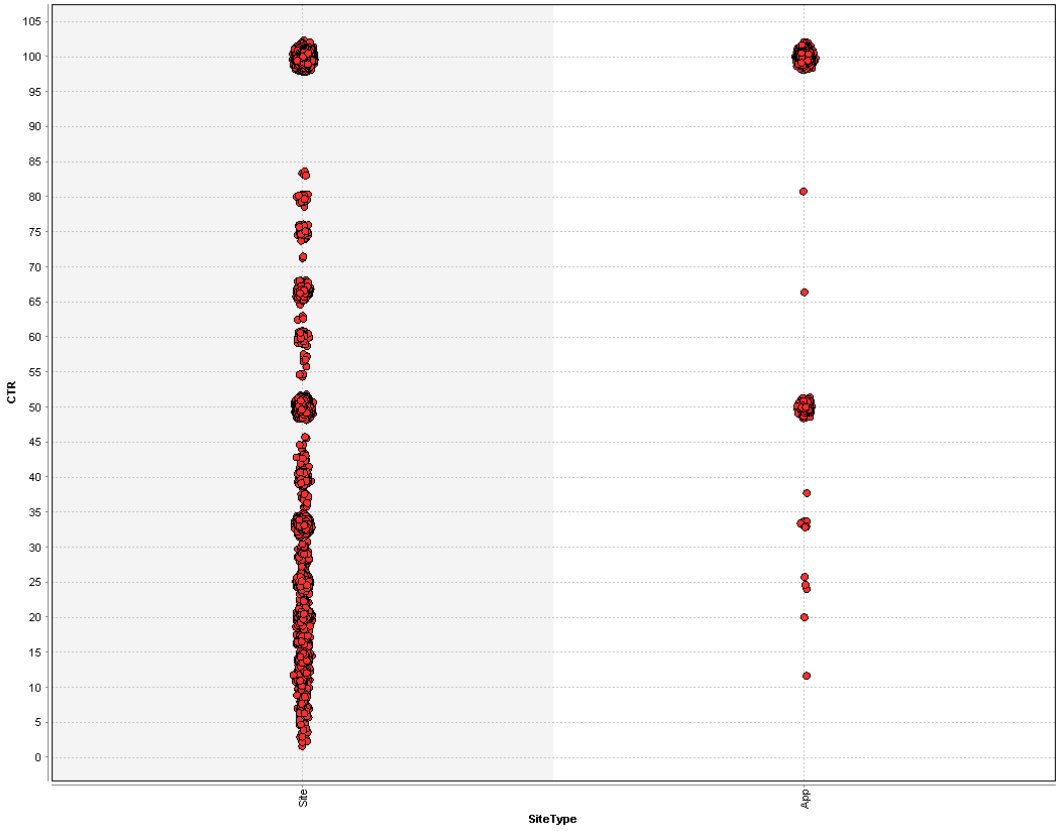
Sl. 3.15. Dijagram rasipanja za atribute CTR i *DeviceOs*.

Atribut vrsta uređaja poprima 3 vrijednosti, od kojih je jedna *Unknown*, odnosno nepoznata, a druge dvije su *Mobile* i *Desktop*. Većina podataka je prikupljena sa mobilnih uređaja i može se primijetiti da češće poprima veće vrijednosti za CTR od *Desktop* uređaja iako je razlog tome najviše razlika u količini podataka prikupljenih za jedne i za druge. Dijagram je prikazan na slici Sl. 3.16.

Atribut koji opisuje da li se oglas prikazuje na internet stranici ili u aplikaciji može poprimiti dvije vrijednosti, a to su *Site* i *App*. Puno češće su oglasi prikazani u aplikacijama pa je zbog toga teže zaključiti postoji li neki uzorak, odnosno gdje je češće veći CTR.



Sl. 3.16. Dijagram rasipanja za atribute CTR i *DeviceType*.



Sl. 3.17. Dijagram rasipanja za atribute CTR i *SiteType*.

3.6. Istraživanje numeričkih atributa

Za ovo istraživanje se koriste histogram i histogram u boji. Osnovna svrha histograma prema [15] je grafički prikazati distribuciju univarijatnog skupa podataka. Najčešće se dobiva cijepanjem niza podataka i njihovim smještanjem u klase jednakih veličina. Na x-osi dijagrama se nalazi varijabla odziva dok se na y-osi nalazi frekvencija ili učestalost pojavljivanja neke vrijednosti. Vrijednosti koje se koriste pri izradi histograma mogu biti u izvornom obliku ili mogu biti normalizirane.

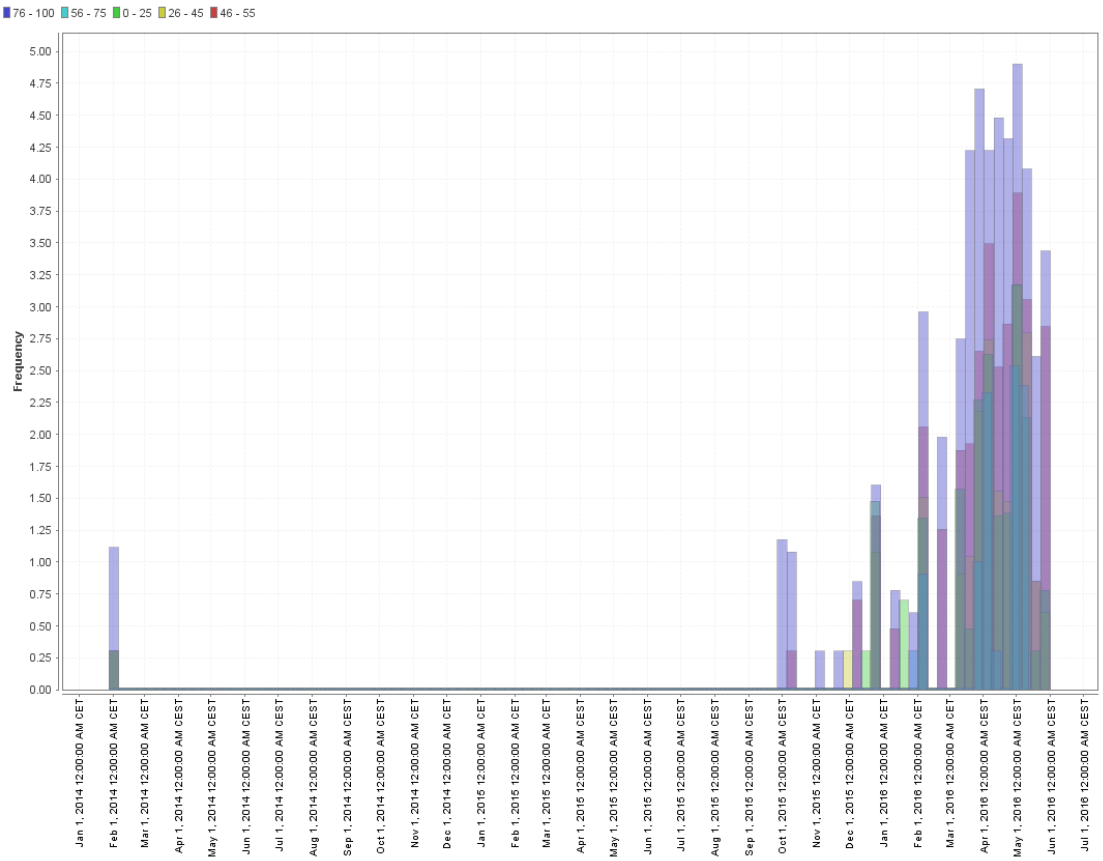
Upotreba histograma je jako raširena i popularna tako da se osim u statističkim programskim alatima često koriste i u programima opće namjene za iscrtavanje dijagrama iz tablica te u poslovnim grafičkim programima kako je navedeno u [15]. Za ovo istraživanje se uz histogram koristi i dijagram histogram u boji koji je njegova izvedenica. Razlika između njega i običnog histograma je da je u histogramu u boji moguće promatrati dva atributa gdje se drugim atributom dijeli prvi u grupe pomoću boja.

Zbog toga što atribut CTR poprima veliki raspon vrijednosti (u ovom radu od 2,326 do 200), za ovo istraživanje izveden je novi atribut koji je nazvan grupirani CTR (engl. *Group* CTR). On je izveden kako bi se mogli uspoređivati numerički atributi s vrijednostima atributa CTR pomoću histograma u boji (svaka grupa vrijednosti CTR-a je prikazana drugom bojom), a i kako bi razlike u vrijednosti CTR-a bile jače izražene. U atributu *Group* CTR vrijednosti CTR-a su grupirane u 5 grupa. U prvu grupu spadaju podaci čiji je CTR od 0 do 25, u drugu podaci s vrijednosti CTR-a od 26 do 45, u treću oni sa vrijednosti od 46 do 55, zatim od 56 do 75 i na kraju oni od 76 do 100. Umjesto 4 grupe koje bi obuhvatile jednake raspone vrijednosti, u ovom radu je dodana odvojena grupa sa vrijednostima CTR-a od oko 50 zbog toga što velik broj podataka ima upravo vrijednost CTR-a 50. Formula za generiranje atributa *Group* CTR u *RapidMiner* bloku *Generate Attribute* je prikazana pod (3-4).

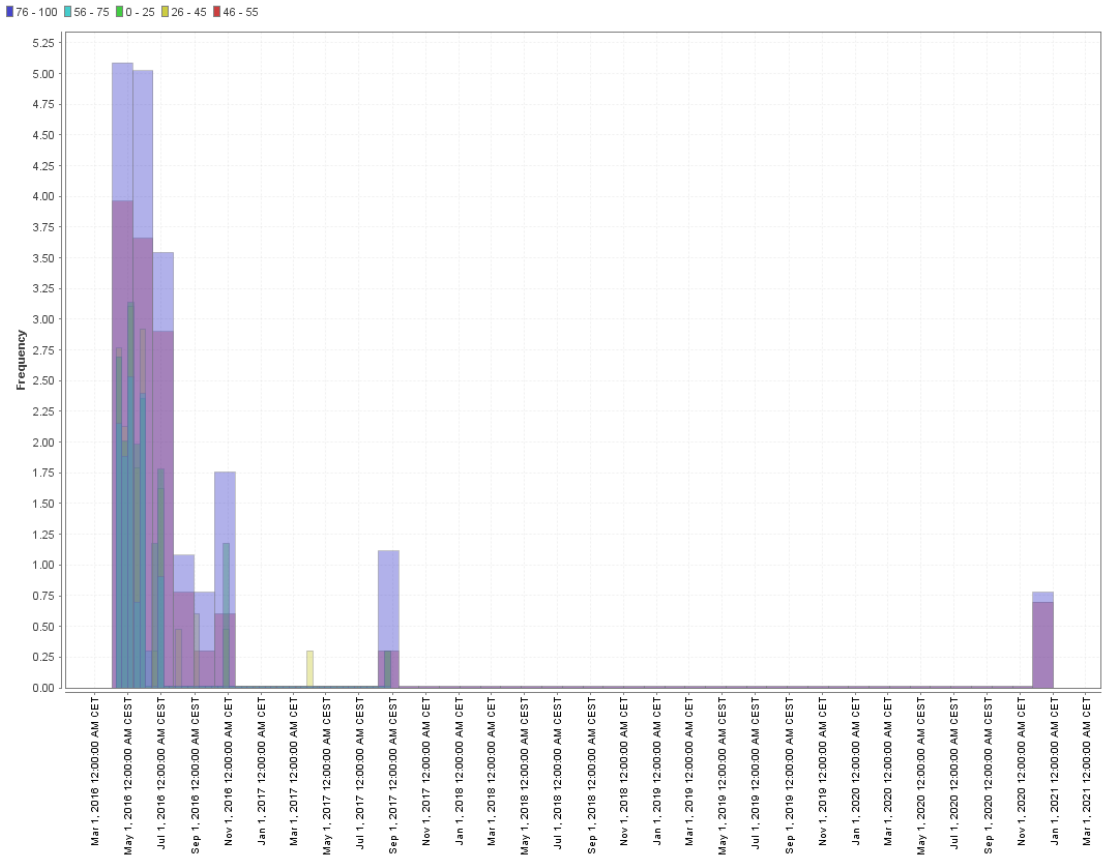
```
if(CTR < 26, "0 - 25", if(CTR < 45, "26 - 45", if(CTR < 56, "46 - 55", if(CTR < 76, "56 - 75", "76 - 100"))))
```
(3-4)

U podacima je izražena dominacija određenih vrijednosti kod pojedinih atributa zbog čega ostale vrijednosti nije moguće ili je jako teško očitati. Zbog toga je u većini prikazanih dijagrama u ovom poglavlju uključena opcija logaritamske skale (engl. *Log Scale*). Ova opcija se koristi kod histograma kako bi raspon vrijednosti frekvencija bio niži, a odnosi između različitih vrijednosti atributa jasnije vidljivi.

Na dijagramima na slikama Sl. 3.18. i 3.19. su prikazani atribut koji prikazuje datum i vrijeme početka marketing kampanje i grupirani CTR, odnosno atribut koji prikazuje datum i vrijeme kraja marketing kampanje i grupirani CTR. Uključena je *Log Scale* opcija. Podacima je obuhvaćen uzak period vremena pa nije moguće izvesti kvalitetan zaključak, a i podaci su poprilično ujednačeno raspoređeni. Zbog toga se ovi atributi neće detaljnije proučavati. Veći značaj za ovo istraživanje ima atribut datuma i vremena kada je oglas prikazan i kliknut koji će biti prikazan kasnije u ovom poglavlju.

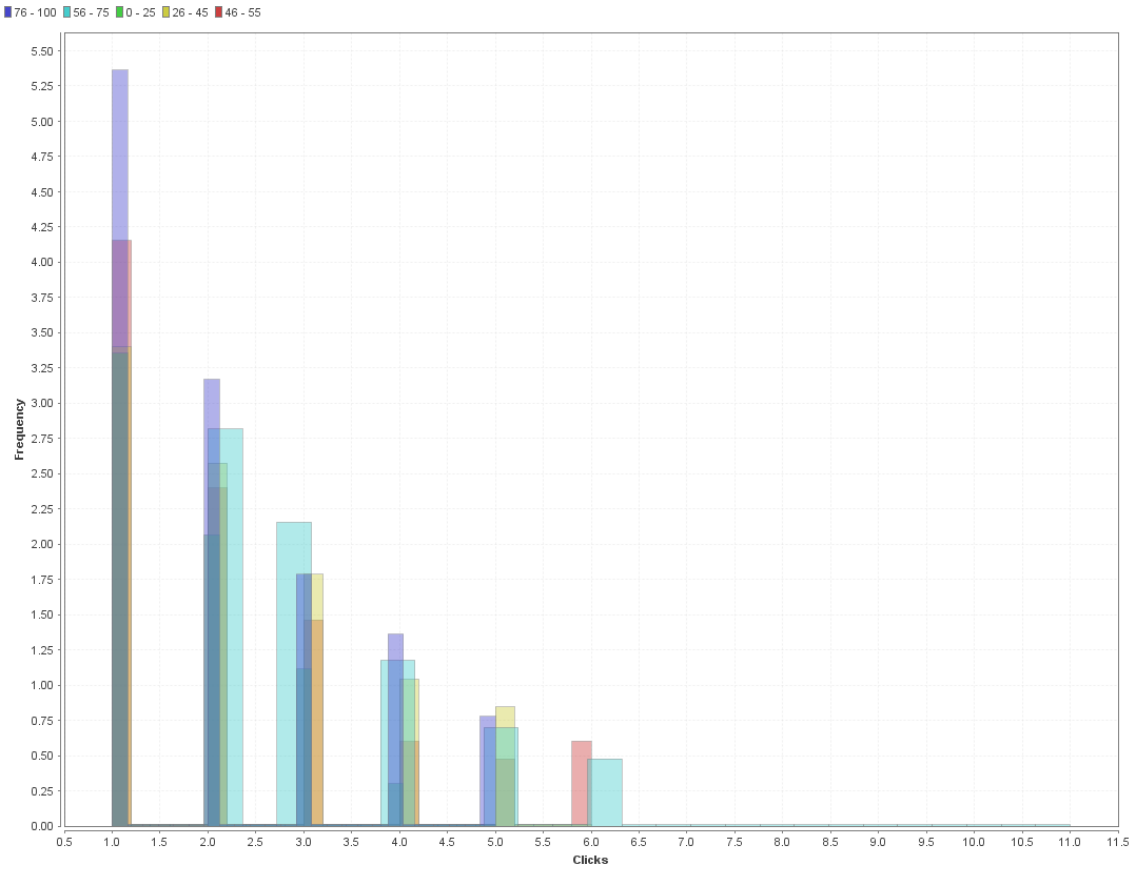


Sl. 3.18. Histogram u boji za atribute *CampaignTimeStart* i *Group* CTR.



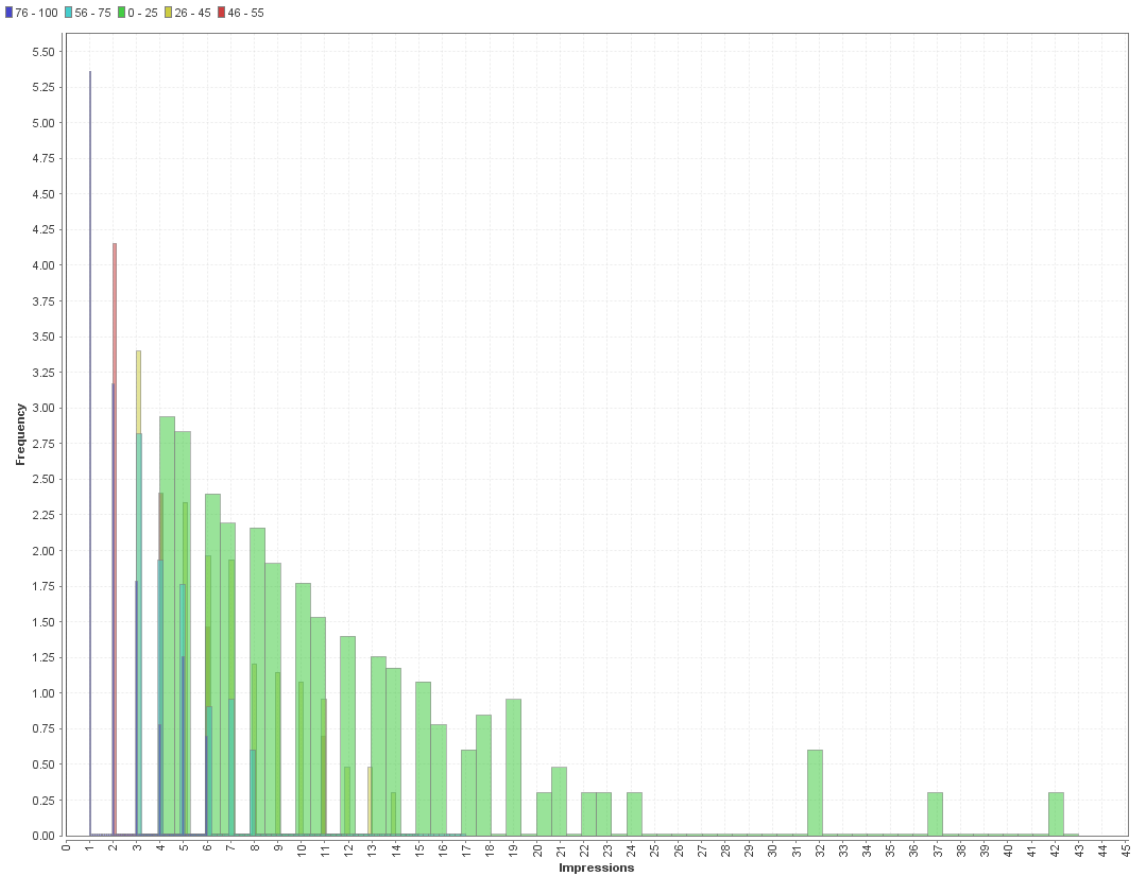
Sl. 3.19. Histogram u boji za atribute *CampaignTimeEnd* i *Group* CTR.

Na slici Sl. 3.20. nalazi se dijagram atributa klikovi na oglas i grupirani CTR. Uključena je *Log Scale* opcija. Iz dijagrama se vidi da ja najviše oglasa koji su kliknuti samo jednom, a među njima je najviše onih koji imaju najveći CTR, što znači da su samo jednom bili i prikazani. U grupu s najvećim CTR-om spadaju i oglasi koji su kliknuti 2 i 4 puta. Samo u oglasima koji su prikazani 5 puta vodstvo ima grupa s vrijednostima za CTR od 26 do 45.



Sl. 3.20. Histogram u boji za atribute *Clicks* i *Group* CTR.

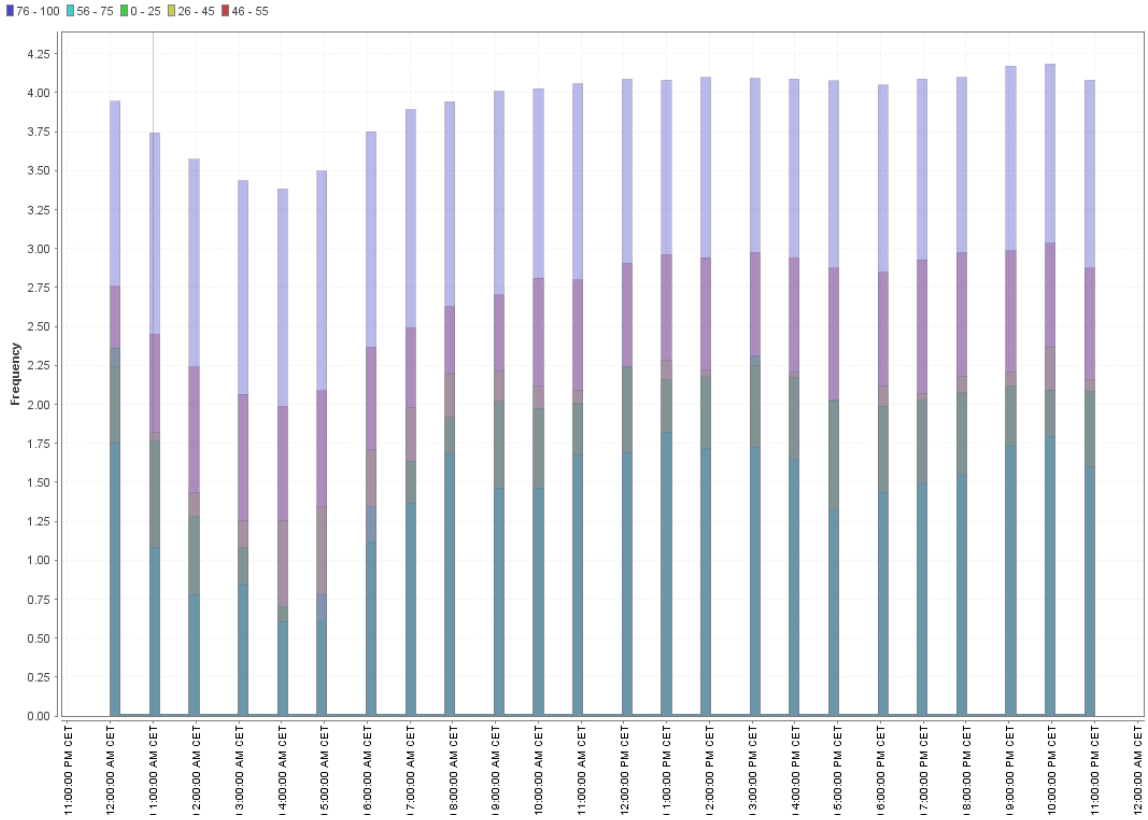
Već je prije utvrđena značajna veza između atributa prikaza oglasa i CTR. Na dijagramu sa slike Sl. 3.21. vidljivo je da je CTR najveći kada je broj prikaza oglasa 1. Uključena je opcija *Log Scale*. Kada je broj prikaza 2, tada je najdominantniji CTR u rasponu od 45 do 55 što znači da se oglasi koji se prikažu dva puta najčešće samo jednom kliknu. Ali ne treba zanemariti niti to da su oglasi s vrijednostima za CTR od 76 do 100 ovdje još uvijek izrazito česti. Daljnjim povećanjem broja prikaza oglasa CTR u rasponu od 0 do 25 postaje dominantan, a na kraju i jedini, što jasno daje do znanja da veći broj prikaza oglasa nikako ne garantira veći CTR, nego upravo suprotno.



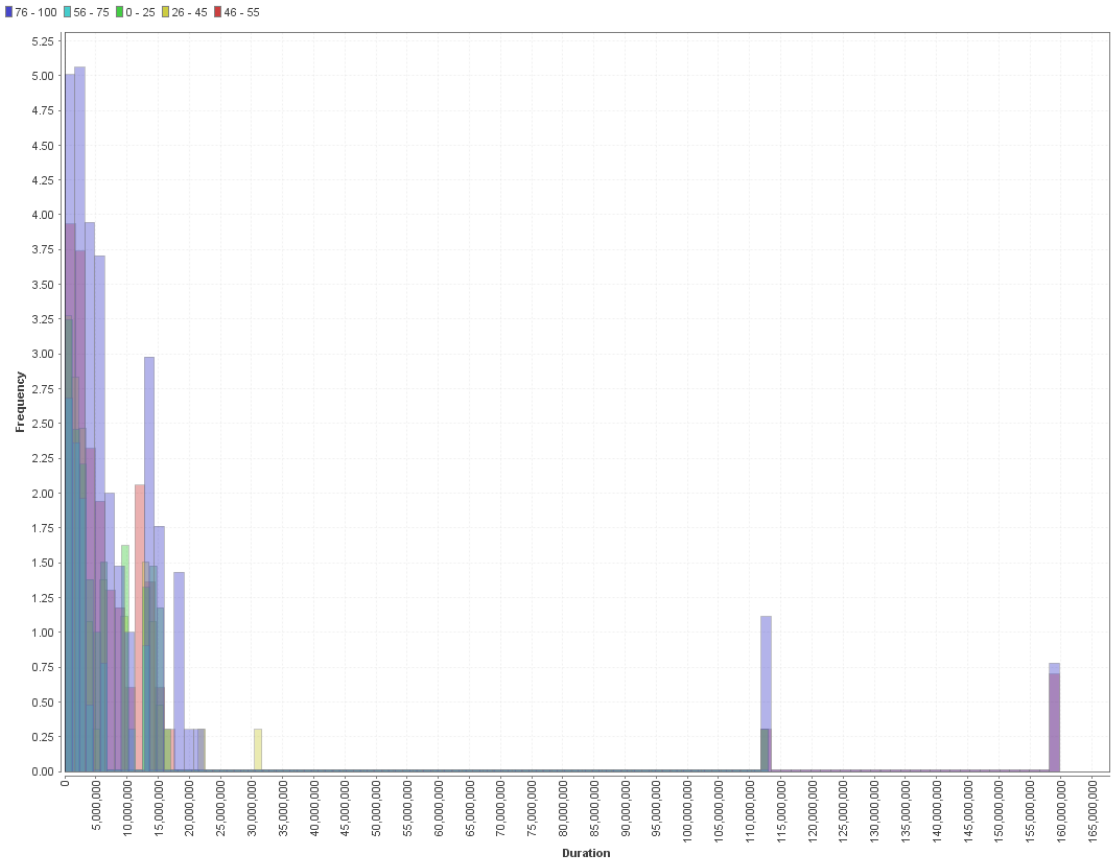
Sl. 3.21. Histogram u boji za atribute *Impressions* i *Group* CTR.

Iz dijagrama sa slike Sl. 3.22. vidi se usporedba atributa sata kad je prikazan i kliknut oglas i grupiranog CTR-a. Uključena je *Log Scale* opcija. Iz dijagrama se može vidjeti kako su grupe CTR-a od 76 do 100, od 56 do 75 i od 46 do 55 u poprilično jednakim odnosima u svim satima. Također se može primijetiti da je vrijednost grupe CTR-a od 26 do 45 dominantnija u jutarnjim satima, odnosno od 2:00 do 10:00 te od 18:00 do 23:00 sata od grupe CTR-a od 0 do 25. Također iz dijagrama je jasno vidljivo da se oglasi puno rjeđe prikazuju od ponoći do 9:00 sati, a razlog je taj što je tada slabija posjećenost internet stranica gdje se ti oglasi prikazuju.

Iz dijagrama sa slike Sl. 3.23. vidi se odnos između dužine trajanja kampanje i atributa grupirani CTR. Uključena je opcija *Log Scale*. Iz dijagrama se jasno vidi da su kraće kampanje najčešće, ali i najuspješnije. Također su jako uspješne kampanje koje traju oko 170 dana te kampanje koje traju oko 230 dana. Najlošije kampanje su one koje traju oko 110 dana. Naravno da je teško reći da je ovo nekakav jasan znak zbog toga što se ne zna koliko su puta pojedini oglasi bili prikazani i o kakvim se oglasima radi.

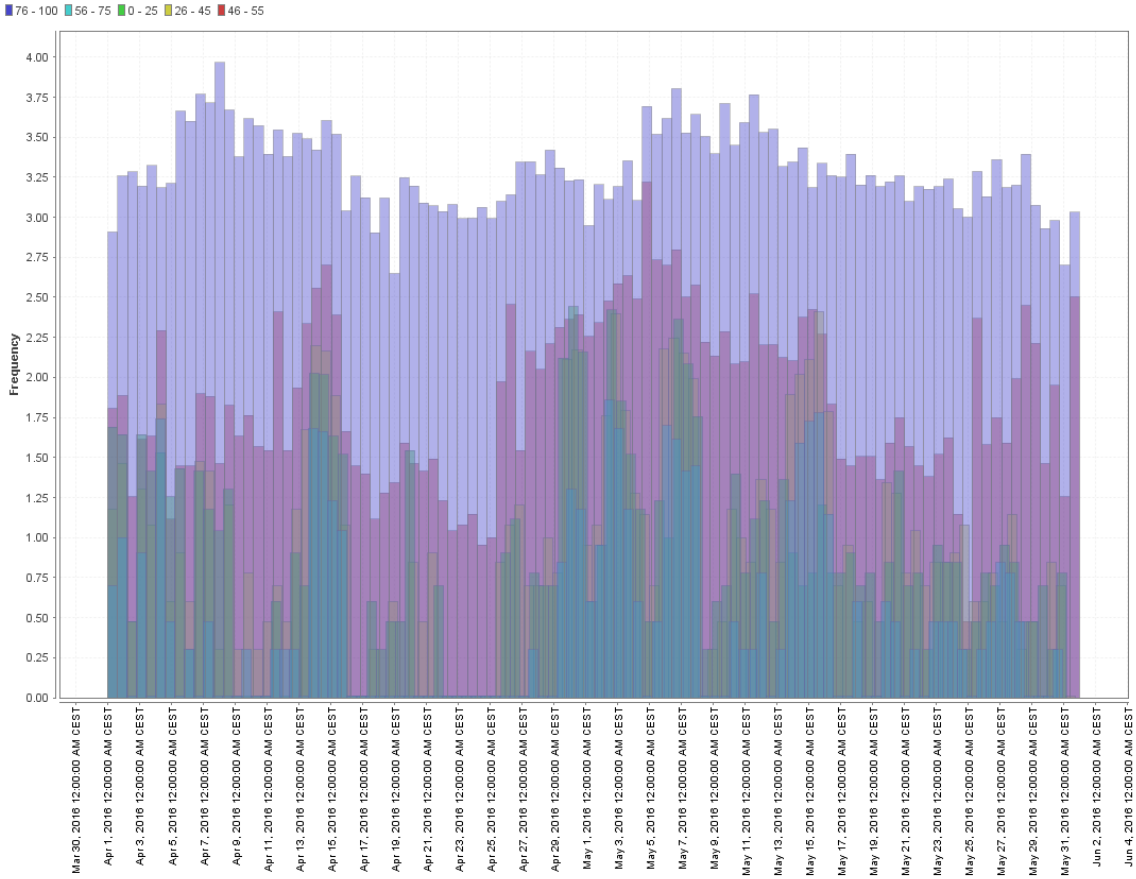


Sl. 3.22. Histogram u boji za atribute *Hour* i *Group* CTR.



Sl. 3.23. Histogram u boji za atribute *Duration* i *Group* CTR.

Odnos između atributa datuma i vremena kada je oglas prikazan i kliknut i atributa grupirani CTR je prikazan na slici Sl. 3.24. Uključena je opcija *Log Scale*. Iz dijagrama se vidi da je udio CTR-a od 0 do 45 poprilično visok početkom travnja dok je sredinom travnja dosta izražena grupa koja obuhvaća CTR od 56 do 75. Ova grupa je bitna jer se sigurno radi o oglasima koji su prikazani više od 2 puta, a budući da je CTR iznad 50 znači da su i češće kliknuti. Slično bi se moglo zaključiti i za početak mjeseca svibnja koji je zanimljiv termin zbog toga što u taj termin spadaju praznici kao što su Praznik rada i u nekim državama pravoslavni Uskrs. Lošiji rezultati, gdje je puno češći CTR od 0 do 45, su početkom travnja i sredinom svibnja. Ostali dani spadaju u neki prosjek, a još se mogu istaknuti dani između 21. i 26. travnja gdje je CTR bio ili jako visok ili srednji.



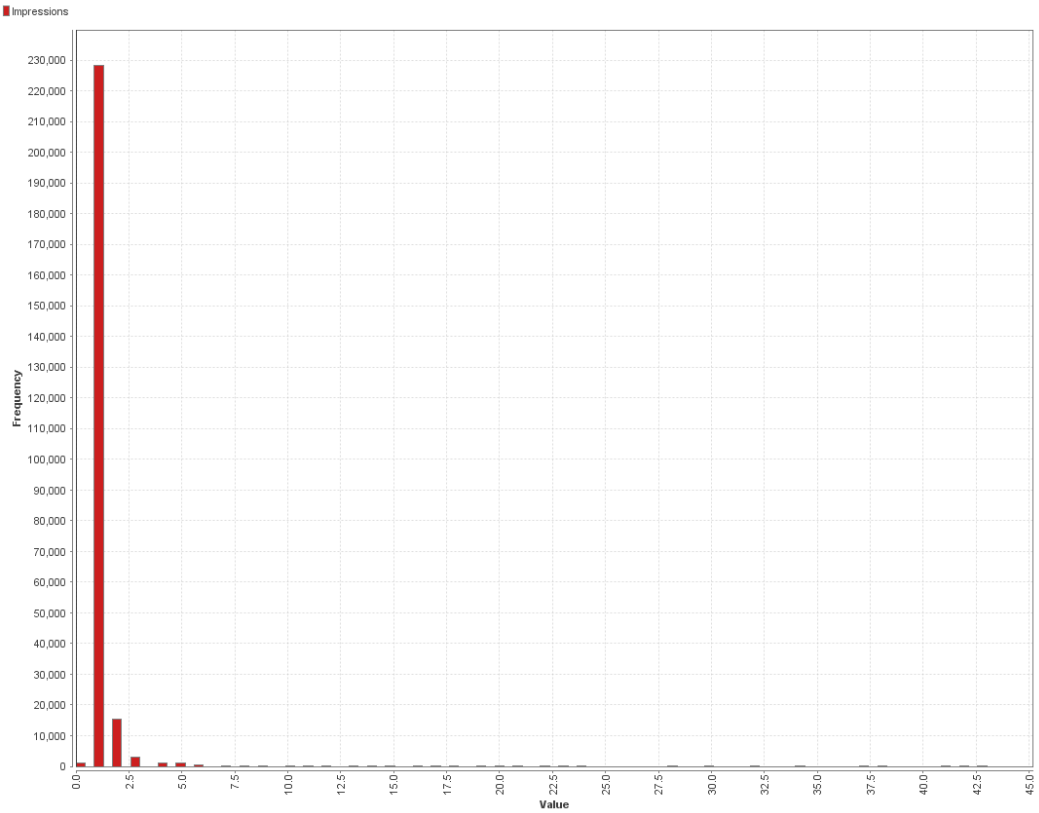
Sl. 3.24. Histogram u boji za atribute *DateHour* i *Group* CTR.

Za atribut virtualnog otiska prsta neće biti prikazan dijagram jer on samo predstavlja vrstu identifikacijskog broja iz kojeg se ne mogu iščitati informacije korisne za analizu u ovom radu. Iako ga se neće detaljno analizirati, ovaj atribut može biti koristan sustavima za isporuku oglasa

jer je moguće izvući neke zaključke o korisniku te mu na taj način isporučivati oglase koje on više preferira.

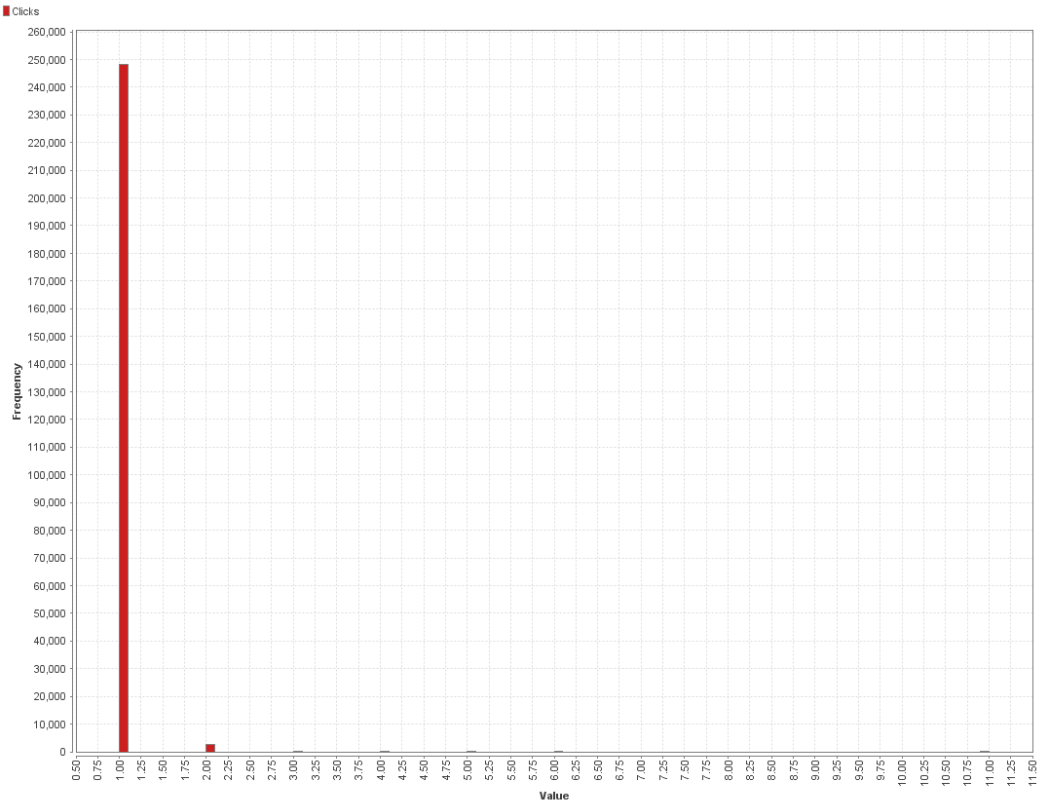
Zaključak istraživanja nominalnih atributa je da se promatranjem samo jednog atributa sa ciljnim atributom ne može izvući puno korisnih informacija. Jedini zaključak koji je potpuno jasan iz prikazanih dijagrama jest da CTR znatno opada s povećanjem broja prikaza oglasa. Također je jako teško donijeti bilo kakav realan zaključak vezan uz CTR jer je najviše oglasa koji su prikazani samo jednom, a daleko poslije njih slijede oglasi koji su prikazani dva puta. U ovom istraživanju uočena je i anomalija, a to su oglasi koji su prikazani 0 puta, a bili su kliknuti. U tim slučajevima se najvjerojatnije radi o oglasima za čiju isporuku nije bio odgovoran sustav za isporuku oglasa obrađen u ovom radu.

Kada se promotri histogram za atribut prikaza oglasa na slici Sl. 3.25, broj oglasa prikazan više od dva puta je zanemariv. Takva izrazito neravnomjerna raspodjela atributa predstavlja problem u ovom istraživanju zbog toga što u rezultatima dominira uglavnom samo jedna ili dvije vrijednosti i na osnovu toga je teško donijeti zaključke.



Sl. 3.25. Histogram za atribut *Impressions*.

Ista situacija je i kada se pogleda broj klikova. Znatno najveći broj oglasa je kliknut samo jednom, a daleko ispod je broj oglasa koji su kliknuti dva puta. Ostali su u zanemarivom broju.



Sl. 3.26. Histogram za atribut *Clicks*.

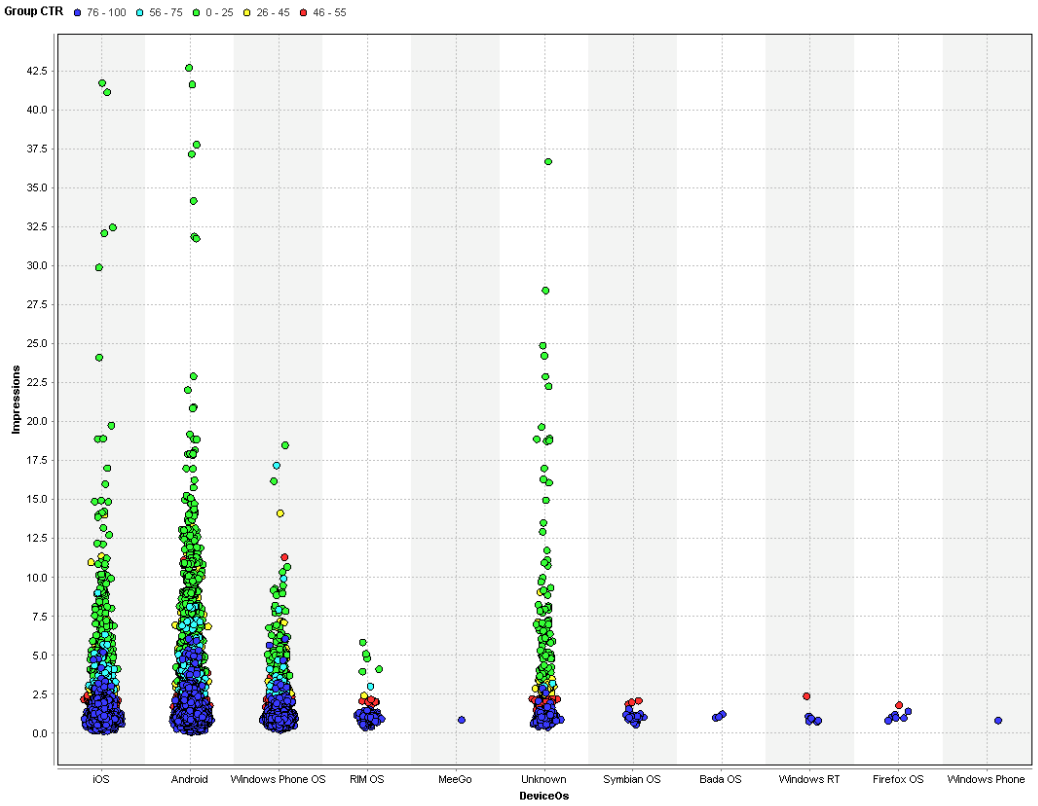
Upravo zbog navedenih razloga CTR je u najvećem broju slučajeva 100 ali je teško zaključiti da li ta brojka doista znači da je oglas uspješan, pogotovo ako se uzme u obzir da velika većina oglasa prikazanih više puta ima puno lošiji CTR, najčešće ispod 50.

3.7. Analiza veza između više atributa i ciljnog atributa

Ponekad može postojati veza između ciljnog atributa i kombinacije dva ili više atributa čak i u slučaju kada između pojedinačnih atributa ne postoji direktna veza s ciljnim atributom. Zbog toga je prema [9] potrebno koristiti tehnike prikaza podataka koje mogu na istom dijagramu prikazati odnose više atributa s ciljnim atributom.

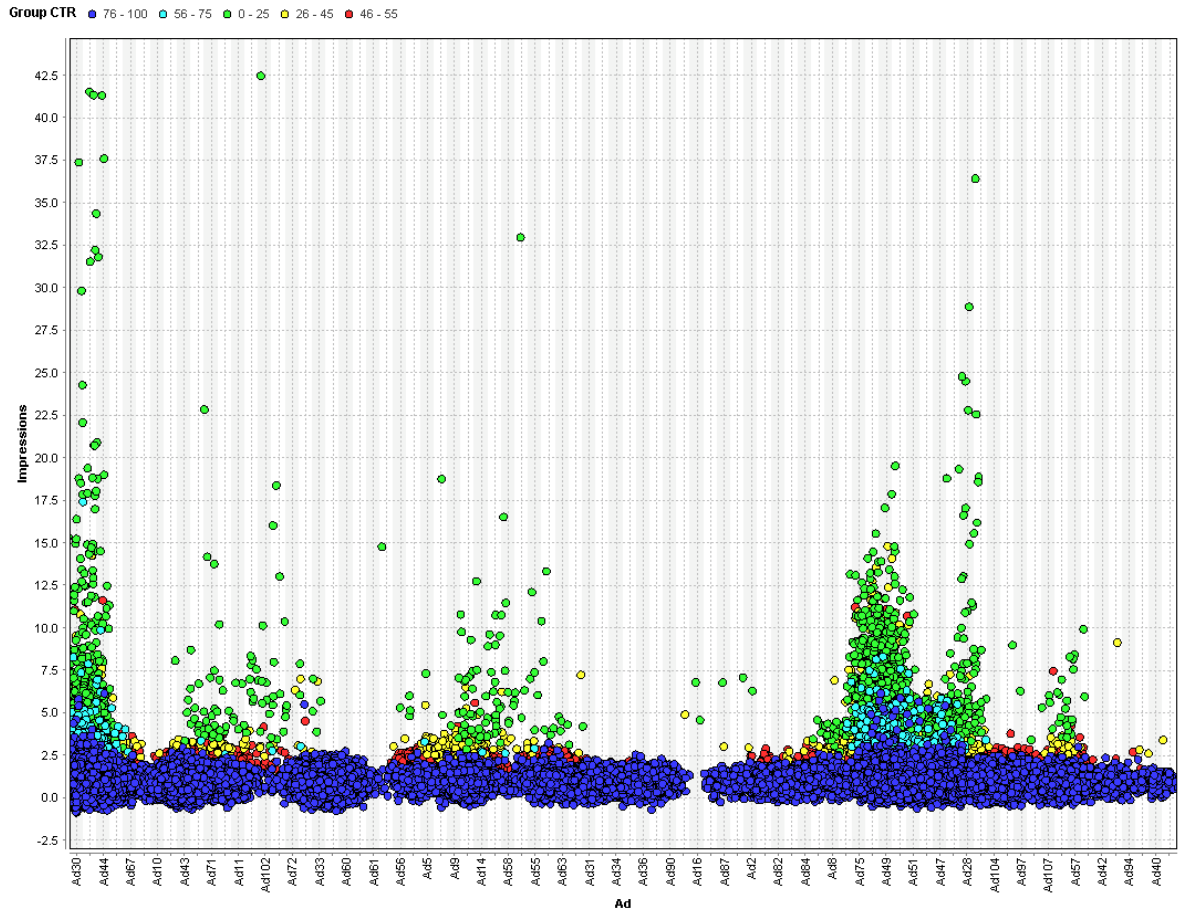
Na slici Sl. 3.27. prikazan je dijagram koji kombinira atribute *DeviceOs*, *Impressions* i *Group* CTR. Na x-osi je atribut *DeviceOs*, na y-osi *Impressions* i bojama je prikazan atribut *Group* CTR. Odabran je *Group* CTR kako bi granice u bojama bile jasnije izražene. Iz dijagrama se može primijetiti da korisnici koji koriste *Android* operacijski sustav ostvaruju veću vrijednost za CTR

kada je broj prikaza 4 i 8. Također se može primijetiti da korisnici *Windows Phone* operacijskog sustava ostvaruju veći CTR od ostalih za vrijednosti atributa *Impressions* od 14 do 18, ali budući da je broj podataka za *Windows Phone* OS u tom rasponu znatno manji od *Androida* i iOS-a, velika je vjerojatnost da se prije radi o iznimci nego o pravilu.



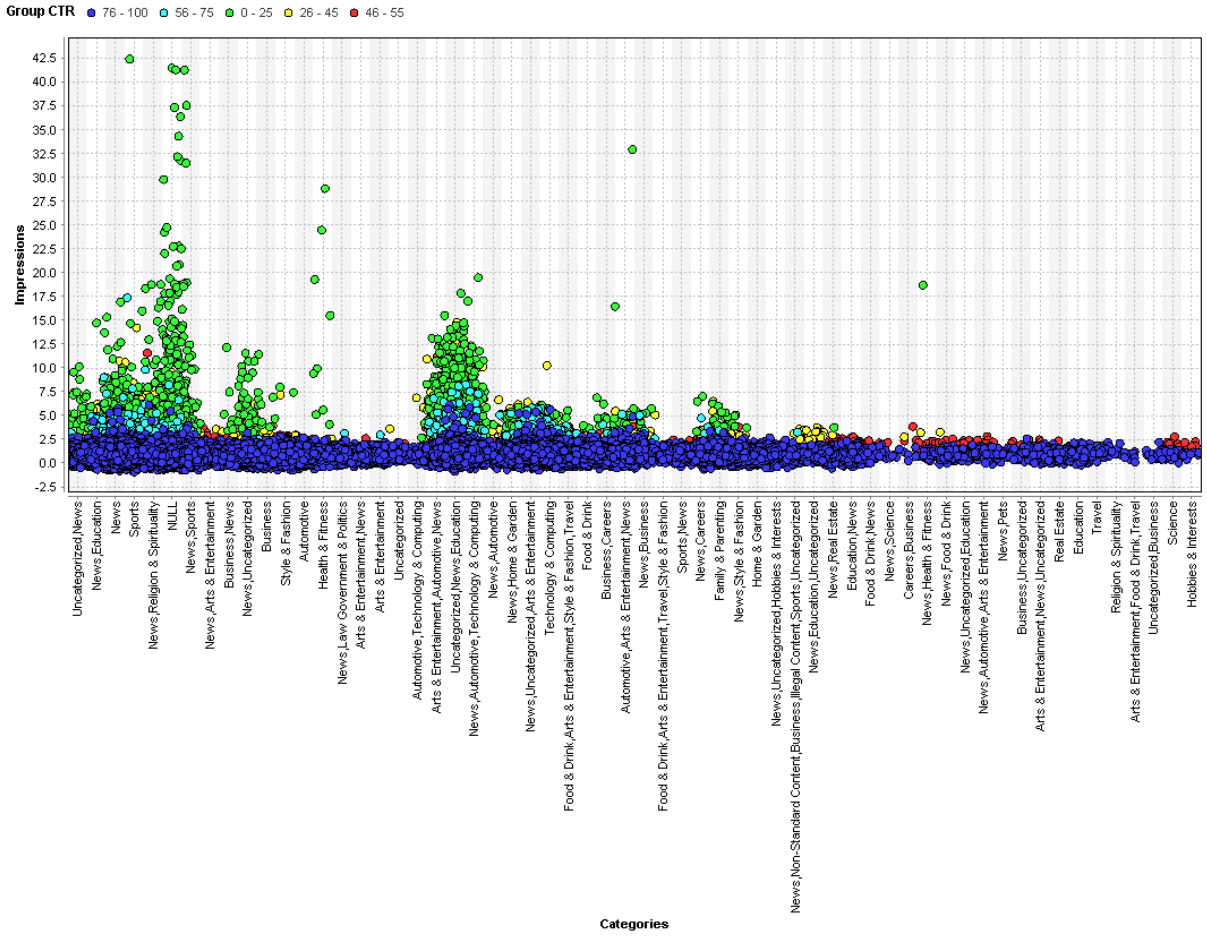
Sl. 3.27. Dijagram rasipanja za atribute *DeviceOs*, *Impressions* i *Group* CTR.

Budući da je u potpoglavlju 3.5., gdje su uspoređivani nominalni atributi s ciljnim atributom, spomenuto kako bi neki oglasi mogli biti uspješniji od ostalih, napravljen je jedan dijagram kako bi se navedeno usporedilo. U dijagramu se promatraju odnosi između atributa *Ad*, *Impressions* i *Group* CTR, a prikazan je na slici Sl. 3.28. Na dijagramu se može vidjeti da određeni oglasi (na dijagramu oglasi oko *Ad30* i *Ad44* te oglasi između *Ad75* i *Ad28*) poprimaju veće vrijednosti za CTR kada je broj prikaza oglasa iznad 3, ali i niske vrijednosti za veći broj prikaza oglasa. Neki oglasi poprilično su neuspješni kada se gleda CTR (npr. oglasi kod *Ad102* i *Ad16*).



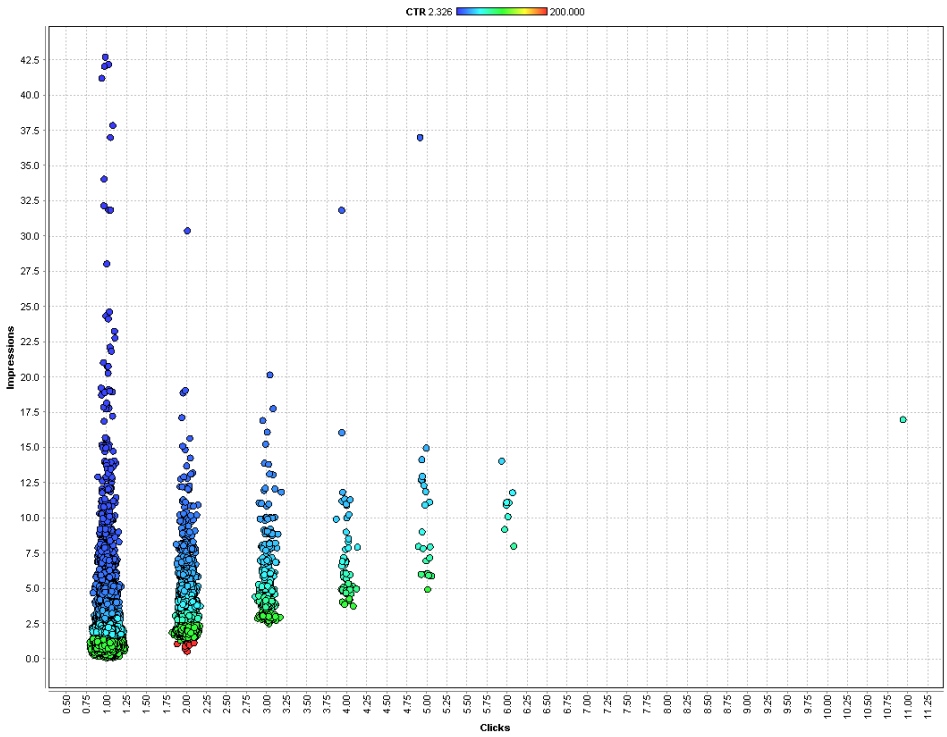
Sl. 3.28. Dijagram rasipanja za atribute *Ad*, *Impressions* i *Group* CTR.

Također u potpoglavlju 3.5. je rečeno kako bi atribut *Categories* mogao biti zanimljiv pa se navedenom atributu posvećuje pažnja u ovom potpoglavlju. Na slici Sl. 3.29. je dijagram s atributima *Categories*, *Impressions* i *Group* CTR. Kao što je i pretpostavljeno, pojedini oglasi pokazuju veći CTR za više od 3 prikaza, ali također se može primijetiti da većina tih kategorija ima najveći broj prikaza općenito i da oni imaju najlošiji CTR.



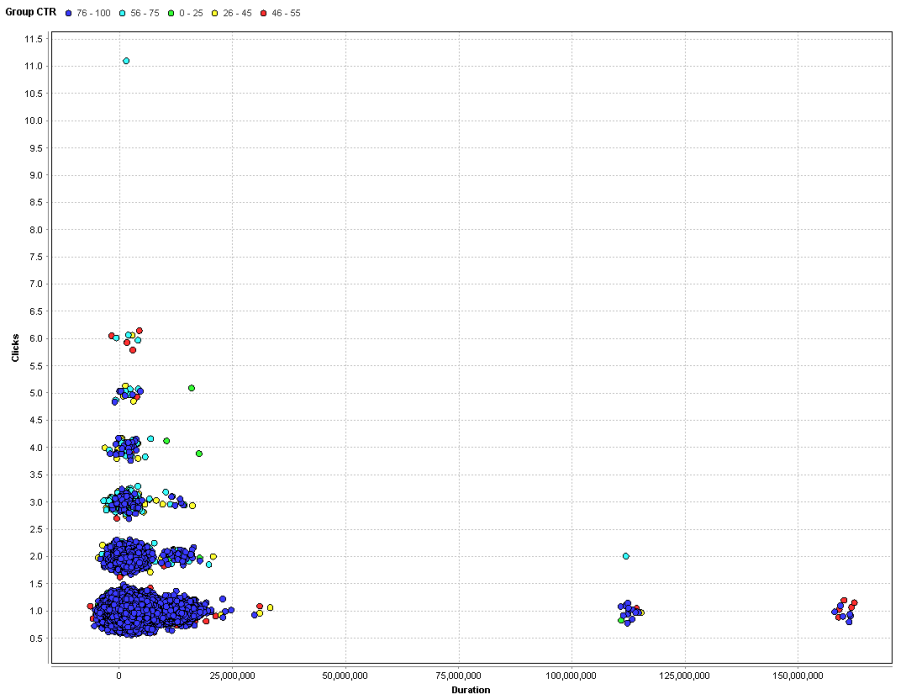
Sl. 3.29. Dijagram rasipanja za atribute *Categories*, *Impressions* i *Group* CTR.

Na dijagramu sa slike Sl. 3.30. prikazan je odnos između atributa *Impressions* i *Clicks*, a bojama je naznačen CTR. Na slici je vidljivo da se anomalije pojavljuju samo u slučaju kada je broj prikaza 1, a broj klikova 2. Također se vidi da je u slučajevima najlošijeg CTR-a broj klikova bio najčešće 1. Iz dijagrama se ponovno vidi velika grupiranost podataka koji su jednom prikazani i jednom kliknuti što daje jako visok CTR koji nije realan. Treba napomenuti i da su iz rezultata izbačeni podaci kod kojih je broj prikaza oglasa (*Impressions*) bio 0.



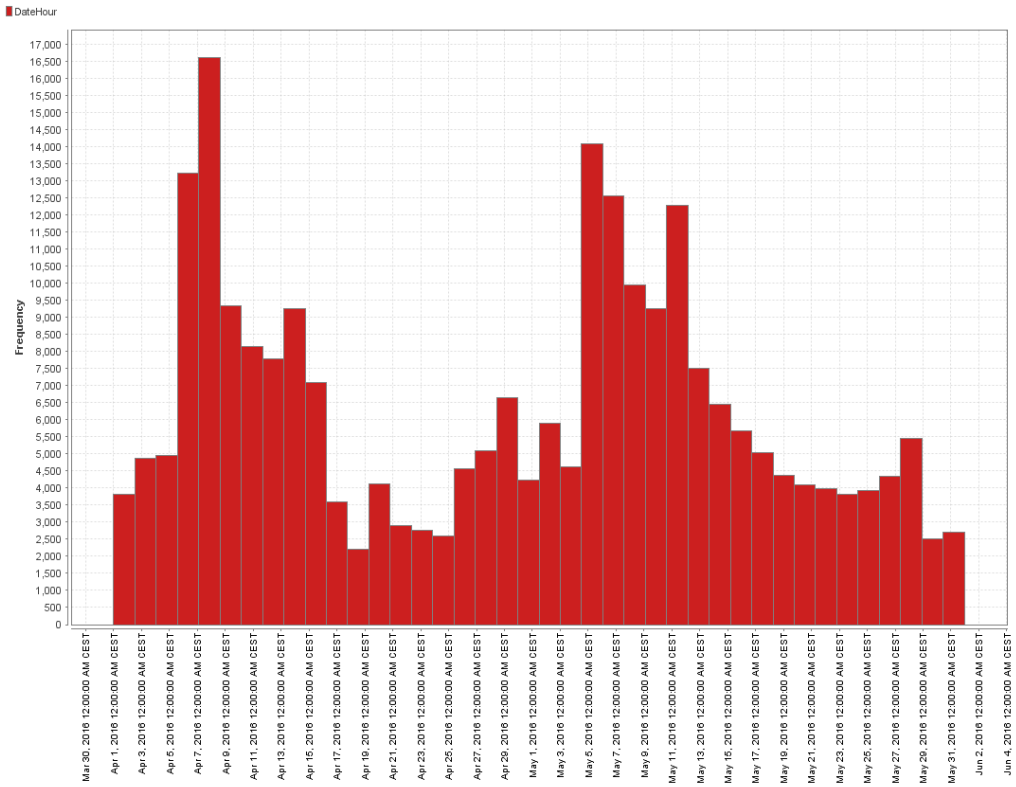
Sl. 3.30. Dijagram rasipanja za atribute *Clicks*, *Impressions* i CTR.

Iz dijagrama sa slike Sl. 3.31. vidi se da najviše klikova ostvaruju oglasi čije kampanje kraće traju, ali također su one i najviše puta prikazane. Ako se zanemare oglasi prikazani jednom i dva puta, najuspješniji su oglasi koji se prikažu 3, 4 i 5 puta.



Sl. 3.31. Dijagram rasipanja za atribute *Duration*, *Clicks* i *Group* CTR.

Iz dijagrama sa slike Sl. 3.32. vidi se da je većina podataka prikupljena početkom mjeseca travnja i svibnja. Zbog toga je odlučeno malo više pažnje posvetiti danima u tom razdoblju kako bi se malo prorijedili podaci te kako bi se vidjelo postoji li neka poveznica između CTR-a i dana u tjednu.



Sl. 3.32. Histogram za atribut *DateHour*.

Atributi koji su najviše promatrani su *Hour*, *Categories* i *DateHour*, a prikazivani su zajedno s atributima *Impressions* i *Group* CTR. Zaključci iz promatranih intervala uneseni su u tablicu Tablica 1. koja se nalazi u prilogu P.1.

Po tablici se da zaključiti da se bolji CTR u podacima koji se češće prikazuju ostvaruje oko 11:00 i 14:00 sati te dosta često i oko 21:00 i 22:00 sata. Bitno je napomenuti da su se uglavnom gledali podaci s brojem prikaza većim od 2 te da su se zajedno s podacima visokog CTR-a najčešće u puno većem broju nalazili i podaci niskog CTR-a.

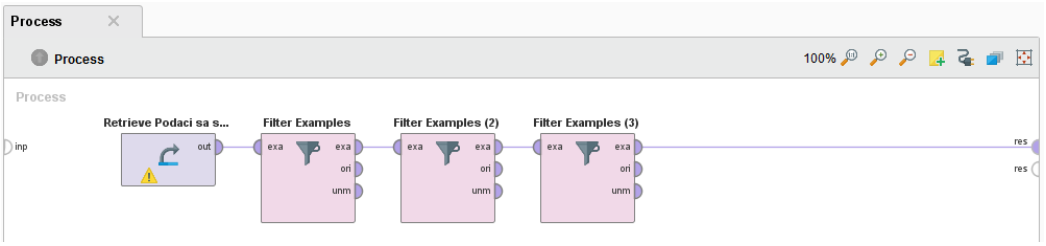
Što se kategorija tiče, treba svakako napomenuti da bi neke kategorije možda bile uspješnije da su više puta prikazane, ali kod većine je broj prikaza oko 1 i 2 puta. Uglavnom su lošije rezultate davale kategorije "*Health & Fitness*", "NULL" te u nekim situacijama i kombinacijama i kategorija "*News*". "NULL" kategorija predstavlja nepoznatu vrijednost atributa,

kada sustavu nije bio dostupan opis pozicije na kojoj se prikazuje oglas. Bolje rezultate daju kategorije kao što su "*Arts & Entertainment*", "*Food & Drink*" te "*News, Education*". Atribut *Categories* bi svakako trebalo još malo detaljnije analizirati.

Od dana u tjednu, vrlo često se ponedjeljak i srijeda javljaju među danima s više puta prikazanim oglasima i visokim CTR-om.

Konačan zaključak detaljnijeg proučavanja dana u navedena 2 mjeseca jest da se ne može sa prevelikom sigurnošću reći da su neki dani ili sati u danu uspješniji ili manje uspješni za prikazivanje oglasa. Previše je varijacija, a navedeni zaključci o danima i satima su temeljeni na nijansama. Razlog tome je što se kod oglasa s visokim CTR-om često pojavljuju i oglasi s jako niskim CTR-om. Što se kategorija oglasa tiče, one će biti dodatno proučene u ovom potpoglavlju u dijagramu na slici Sl. 3.39.

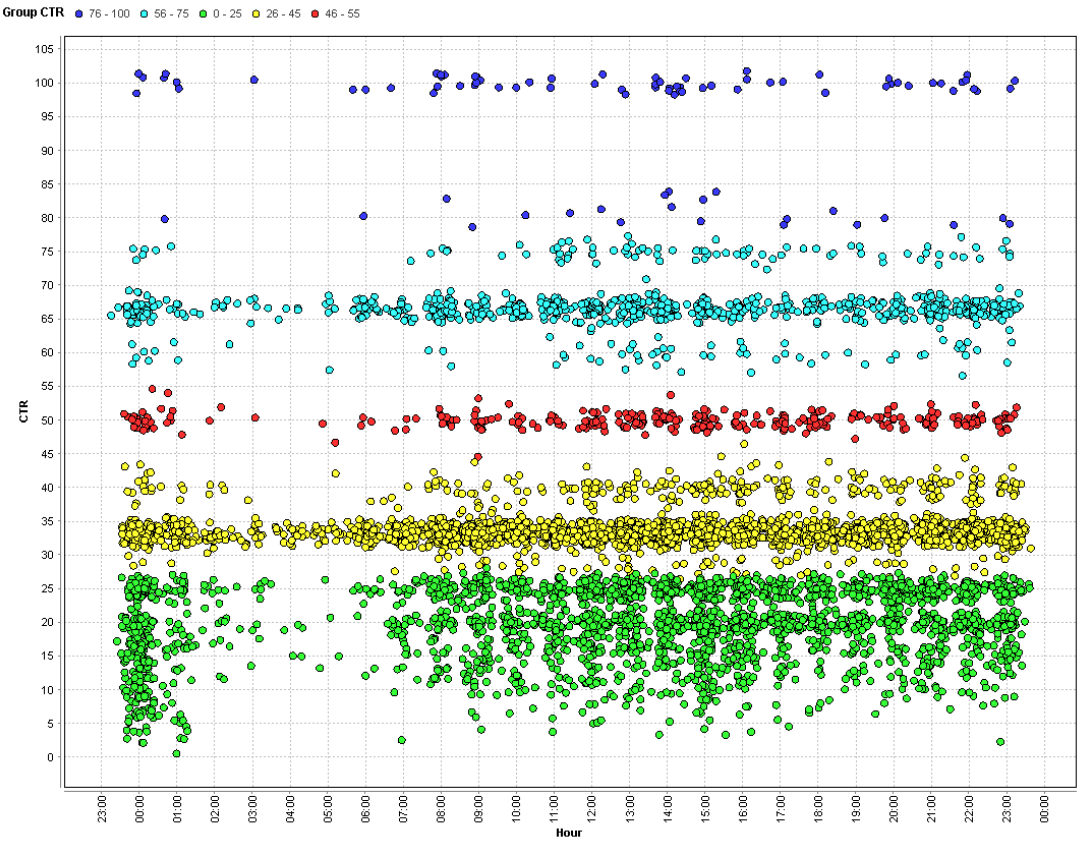
Sljedeći korak u analizi je smanjiti broj podataka na način da se uklone podaci s određenom vrijednosti atributa koji se pojavljuju u jako velikom broju slučajeva. Ovdje se radi o vrijednostima atributa *Clicks* i *Impressions*. Uklonjeni su oni podaci koji i za atribut *Clicks* i za atribut *Impressions* imaju vrijednost 1 i podaci koji imaju za atribut *Clicks* vrijednost 1, a za atribut *Impressions* vrijednost 2. Osim navedenih podataka, u određenim analizama su još uklonjeni podaci koji i za atribut *Clicks* i za atribut *Impressions* imaju vrijednost 2. Cilj je bio dobiti uvid u odnose među atributima u slučaju većeg broja prikaza oglasa, jer se tu najviše vidi razlika između uspješnijih i manje uspješnih oglasa. Većina podataka ima ciljni atribut CTR u vrijednosti 100, ali isto tako većina tih podataka je prikazana samo jednom, pa se teško može reći da li je neki oglas bio uspješniji od drugog. Na većem broju prikaza oglasa bi se razlike trebale jasnije vidjeti. Izgled blokova za filtriranje podataka u *RapidMiner*-u prikazan je na slici Sl. 3.33.



Sl. 3.33. Filtriranje podataka u *RapidMiner-u*.

Za početak, na slici Sl. 3.34. je prikazan dijagram s atributima *Hour* i CTR, a atribut CTR je još dodatno naglašen bojama atributa *Group* CTR. Iz dijagrama se može vidjeti da se dosta oglasa s visokim CTR-om prikazivalo u 14:00 sati, a dobri termini su i 08:00 i 09:00 sati te 15:00

i 16:00. I u prethodnom analiziranju su također oglasi u 14:00 sati ostvarivali najbolje rezultate za CTR. Najlošije vrijeme, gdje je puno podataka sa izrazito slabim CTR-om je 00:00, a slijedi ga 15:00 sati. Termin od 15:00 sati se pojavljuje i u slučaju većeg broja oglasa s visokom vrijednosti CTR-a i u slučaju većeg broja oglasa s niskom vrijednosti CTR-a.



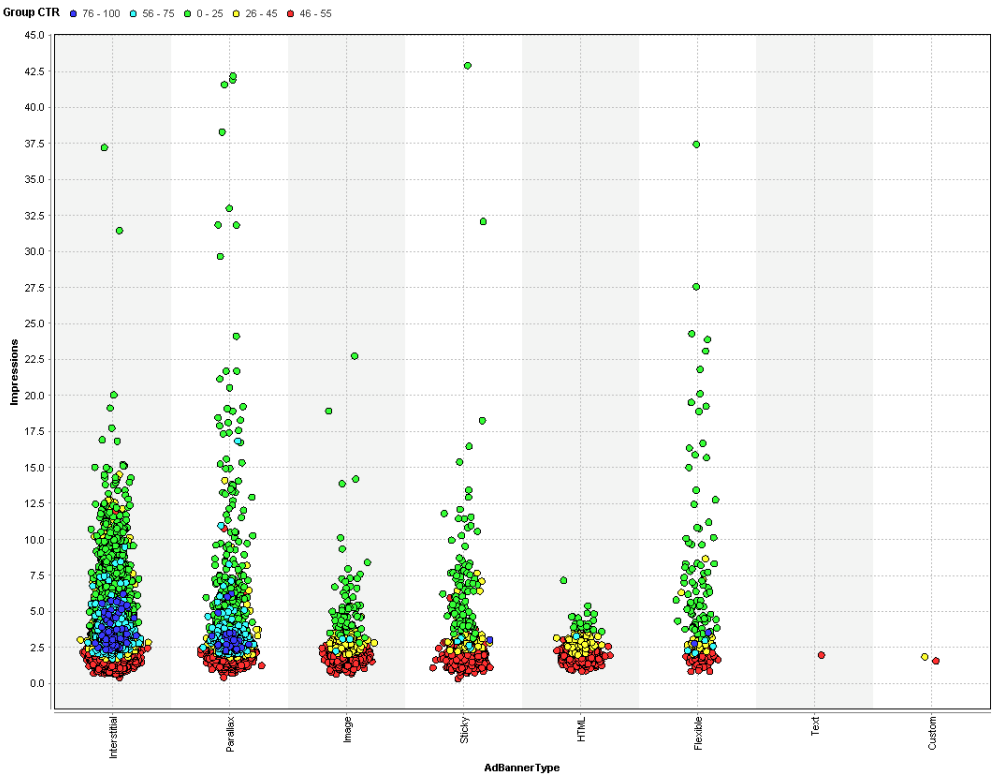
Sl. 3.34. Dijagram rasipanja za atribute *Hour*, CTR i *Group* CTR.

Na slici Sl. 3.35. nalazi se dijagram sa atributima *Hour*, *AdBannerType* i *Group* CTR. Iz dijagrama se vidi da dva tipa banera s najveći brojem prikaza, *Interstitial* i *Parallax*, ostvaruju najveći CTR u 08:00, 09:00 i 14:00 sati. Također se jasno vidi da je *Sticky* tip banera najčešće prikazivan u 00:00 ali je ostvario jako slab CTR.

Kada se pogledaju vrste banera prema broju prikaza oglasa na slici Sl. 3.36. vidi se da je *Interstitial* tip banera najuspješniji po broju oglasa s jako visokim CTR-om, ali je isto tako i najbrojniji.

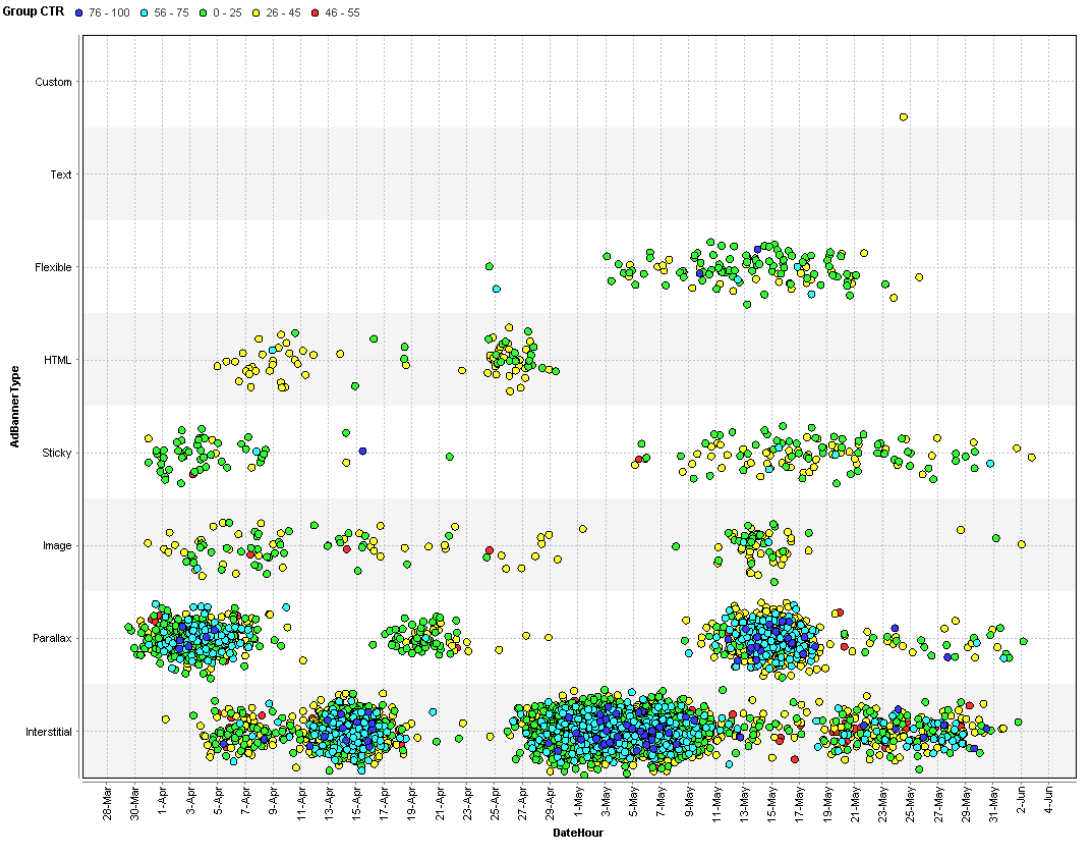


Sl. 3.35. Dijagram rasipanja za atribute *Hour*, *AdBannerType* i *Group* CTR.



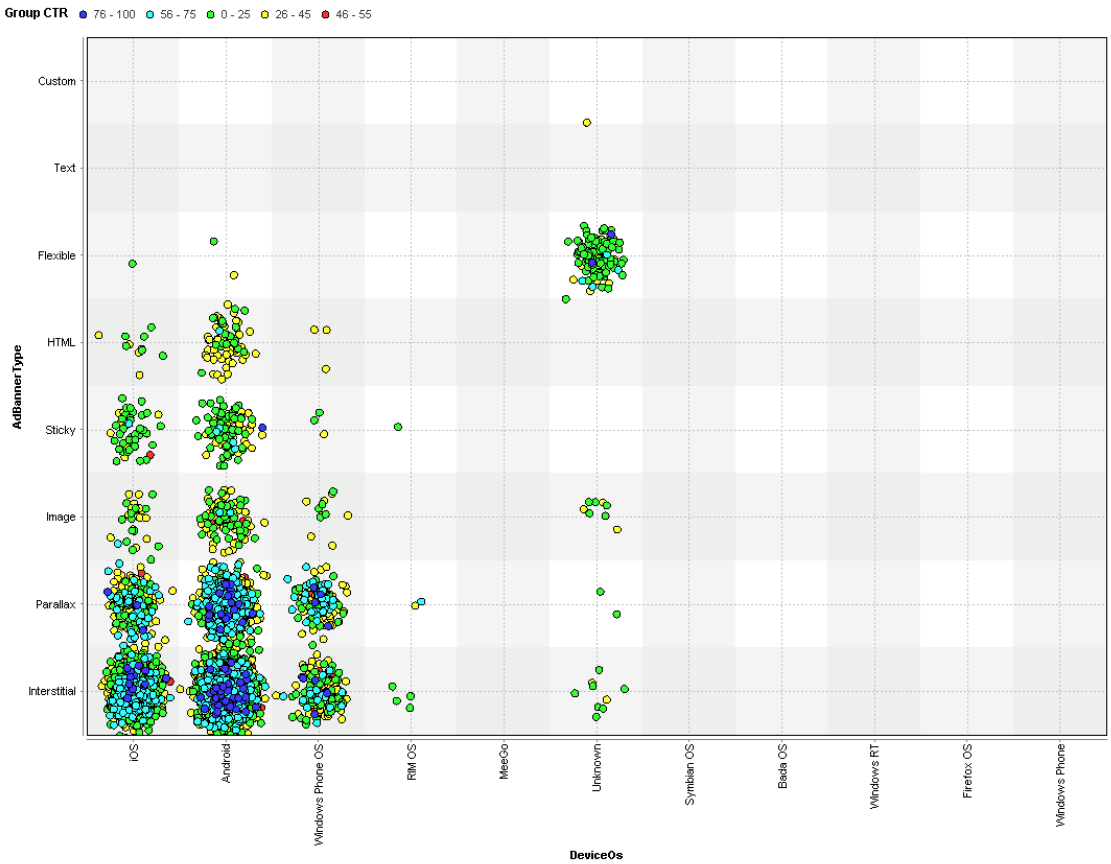
Sl. 3.36. Dijagram rasipanja za atribute *AdBannerType*, *Impressions* i *Group* CTR.

Kada se pogleda atribut *AdBannerType* zajedno s atributom *DateHour* na slici Sl. 3.37. vidi se da nisu u isto vrijeme određeni tipovi banera imali najviše prikaza. U razdoblju od 3. do 11. travnja kada su i *Interstitial* i *Parallax* imali dosta visok broj prikaza, *Parallax* je bio uspješniji. U ostalim danima obuhvaćenim ovim podacima izmjenjuju se navedena dva tipa banera tako da više nema dana u kojima su oba imala visok broj prikaza oglasa, pa se ne mogu niti uspoređivati. Tu se opet javlja problem istraživanja budući da nisu podjednako zastupljeni svi tipovi banera u svim danima, pa uspješnost određenog tipa banera leži u nečem drugom.



Sl. 3.37. Dijagram rasipanja za atribute *DateHour*, *AdBannerType* i *Group* CTR.

Kada se usporede atributi *DeviceOs* i *AdBannerType*, vidi se da su oglasi na operacijskom sustavu *Android* bili najbrojniji ali i da su imali najveći broj oglasa s visokim CTR-om. Na slici Sl. 3.38. se također vidi da je *AdBannerType* tipa *Flexible* najviše bio prikazivan na uređajima kod kojih sustav za isporuku oglasa nije mogao prepoznati tip operacijskog sustava, a i da je njegov CTR za oglase koji su više puta prikazani jako slab. Osim *Parallax* i *Interstitial* tipa banera skoro niti jedan ne ostvaruje visok CTR kada je broj prikaza oglasa veći od 2.



Sl. 3.38. Dijagram rasipanja za atribute *DeviceOs*, *AdBannerType* i *Group* CTR.

S obzirom da je ranije primijećeno da pojedini oglasi ostvaruju bolje rezultate CTR-a, navedena se pojava detaljnije istražuje u ovom potpoglavlju i pokušava se odrediti što neke oglase čini uspješnijima. Proučavanjem dijagrama na slici Sl. 3.28. izdvojeni su oglasi *Ad30*, *Ad80*, *Ad44*, *Ad33*, *Ad75*, *Ad46*, *Ad49*, *Ad50*, *Ad51*, *Ad48*, *Ad47* i *Ad96*.

Rezultati proučavanja su da se navedeni oglasi pojavljuju samo u 3 tipa banera, a to su *Parallax* (samo *Ad80*), *Flexible* (samo *Ad75*) i *Interstitial* (svi ostali). Primijećeno je također da se u oglasima *Ad30*, *Ad50* i *Ad47* pojavljuju već prije navedene anomalije (vrijednost CTR-a je 200). Što se kategorija tiče, izdvojeni oglasi su prisutni gotovo u svim kategorijama koje se nalaze u podacima. U tablici su navedene kategorije u kojima su ostvarene najbolje vrijednosti CTR-a te su odvojene ovisno o tome kolika je česta pojava i niskog CTR-a u tim kategorijama.

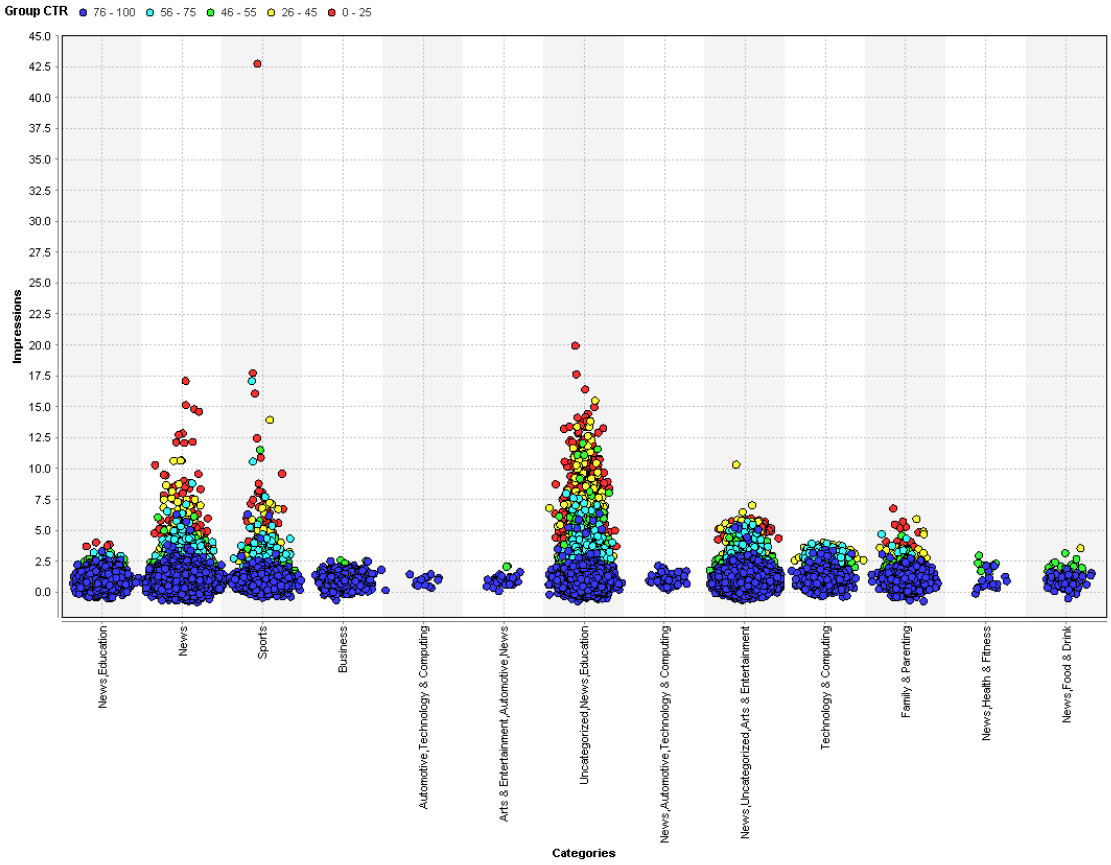
Tablica 3.4. Kategorije s većim brojem podataka visoke vrijednosti CTR-a.

Kategorije bez najlošije grupe CTR-a	Kategorije s manjim brojem najlošije grupe CTR-a	Kategorije s većim brojem najlošije grupe CTR-a
News, Health & Fitness	News, Education	Automotive, Technology & Computing
Technology & Computing	News	Uncategorized, News, Education
News, Food & Drink	Sports	News, Uncategorized, Arts & Entertainment

Osim pojedinih oglasa, možda je zanimljivije provesti istraživanje među odvojenim kategorijama. Kategorije koje su uzete u razmatranje su neke od kategorija koje su razmatrane prethodno u ovom potpoglavlju, kod proučavanja dana u mjesecu travnju i svibnju, a to su: "*Sports*", "*News, Education* ", "*News*", "*News, Health & Fintess*", "*News, Food & Drink*", "*Uncategorized, News, Education*", "*News, Uncategorized, Arts & Entertainment*", "*Automotive, Technology & Computing*", "*Family & Parenting*" i "*Technology & Computing*" (navedene u tablici Tablica 1. u prilogu P.1.). Osim tih kategorija, još su dodane kategorije: "*Bussines*", "*Arts & Entertainment, Automotive, News*" i "*News, Automotive, Technology & Computing*".

Nakon provedenog istraživanja zaključeno je da su prisutni gotovo svi tipovi banera, a *Interstitial* i *Parallax* ostvaruju najbolje rezultate za CTR i njih je najviše. *Interstitial* i *Sticky* također imaju najviše oglasa s lošim CTR-om, ali za *Interstitial* je to i očekivano s obzirom na učestalost tog tipa banera. Promotreni su rezultati u slučaju kada su izbačeni oglasi sa vrijednosti 100 za atribut CTR, uz broj prikaza 1 ili 2 te sa vrijednosti 50 za atribut CTR kada je broj prikaza 2. Kao i u prijašnjim istraživanjima i ovdje je utvrđeno da je u 14:00 sati najviše podataka s visokim CTR-om za veći broj prikaza oglasa.

Kategorije koje imaju najviše oglasa s visokim CTR-om su "*News, Uncategorized, Arts & Entertainment*" i "*Technology & Computing*", ali nema puno podataka prikazanih više od 5 puta. Također i kod kategorija "*News*", "*Sports*", "*Uncategorized, News, Education*" i "*Family & Parenting*" ima puno podataka s dobrim rezultatima za CTR, ali i puno podataka s niskim CTR-om za visoke vrijednosti atributa *Impressions*. I ostale kategorije su mogle biti uspješne ali za njih nema puno podataka s visokim vrijednostima za atribut *Impressions*.

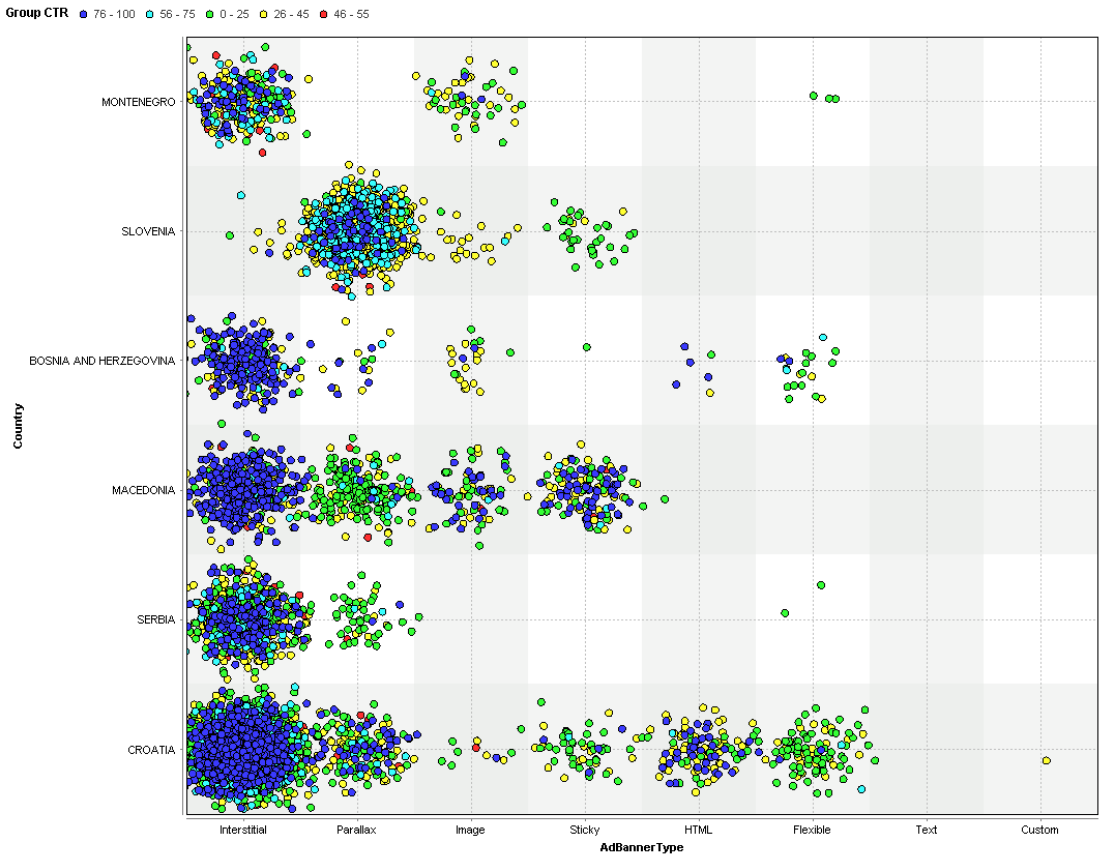


Sl. 3.39. Dijagram rasipanja za atribute *Categories*, *Impressions* i *Group* CTR.

Posljednja provedena analiza obuhvaća filtriranje država kako bi se provjerilo postoje li razlike između država što se tiče tipa oglasa, kategorije i slično. Odabrane su već prije spomenute države jugoistočne Europe za koje ima najviše podataka, a to su Hrvatska, Bosna i Hercegovina, Srbija, Slovenija, Crna Gora i Makedonija. Osim filtriranja država, izvršeno je i dodatno filtriranje podataka gdje su izbačeni podaci s CTR-om jednakim 100 za 1 prikaz te 50 za 2 prikaza kako bi se odmah dobili rezultati koje je lakše proučavati.

Na slici Sl. 3.40. je prikazan dijagram s atributima *AdBannerType* i *Country* uz boje za atribut *Group* CTR. Iz dijagrama se vidi da Slovenija od svih država ima najbolje rezultate za *Parallax* tip banera dok podataka za *Interstitial* gotovo i nema. Kod ostalih država je *Interstitial* tip banera najčešći i najuspješniji. U Sloveniji su također rezultati za *Sticky* najlošiji.

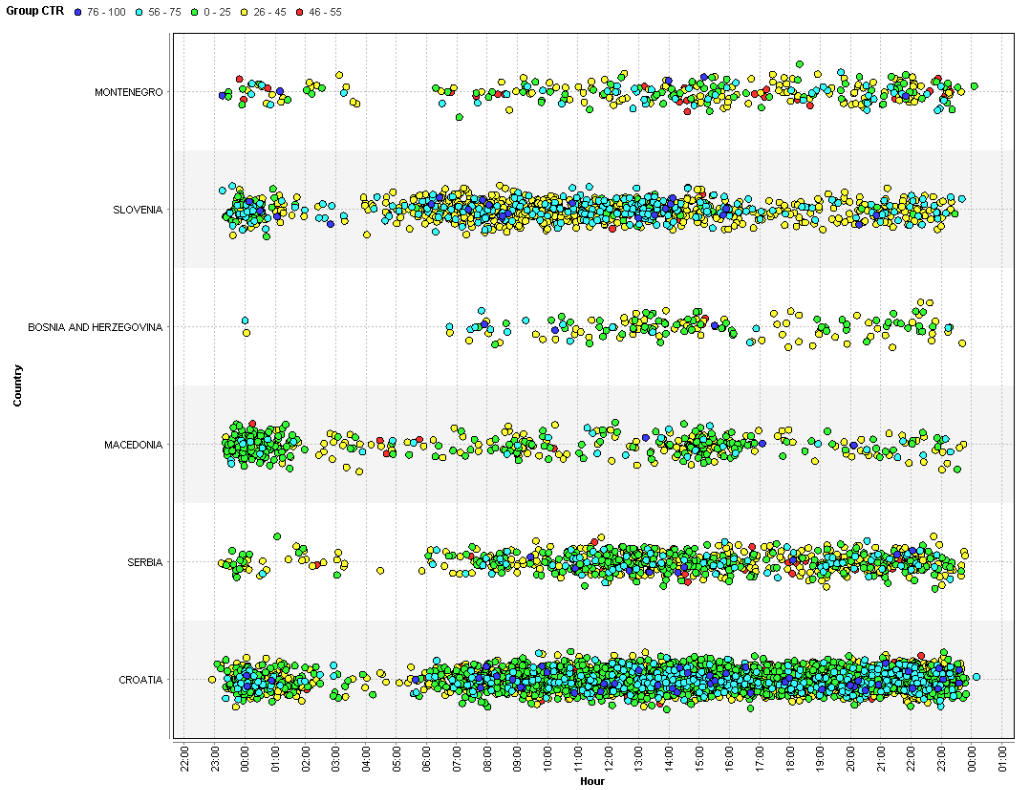
Baneri *Image* i *Sticky* su uspješniji u Makedoniji, a i HTML je uspješan u Hrvatskoj i Bosni i Hercegovini. Treba uzeti u obzir da su takvi rezultati samo kada su u razmatranje uključeni podaci koji imaju vrijednost atributa *Impressions* jednaku 2 i atributa CTR jednaku 100. Bez toga su rezultati ipak lošiji za *Image*, *Sticky* i HTML.



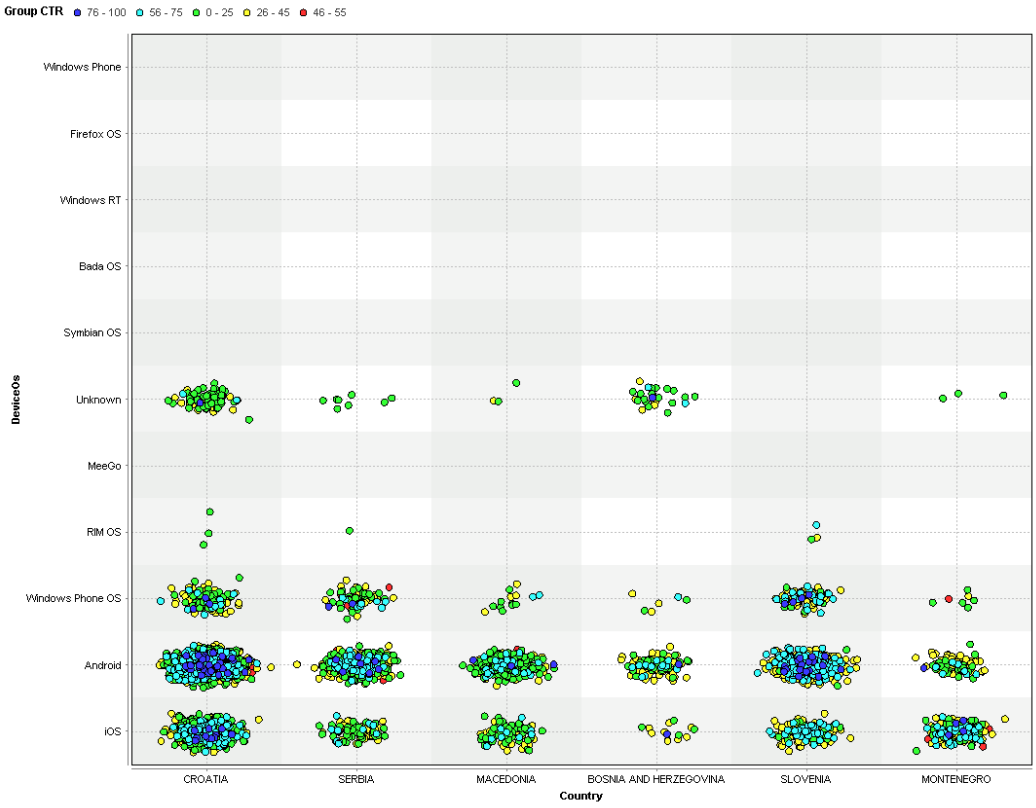
Sl. 3.40. Dijagram rasipanja za atribute *AdBannerType*, *Country* i *Group* CTR.

Što se atributa *Hour* tiče, u Sloveniji su najbolji rezultati ostvareni u vremenu od 06:00 do 09:00 te od 10:00 do 15:00 sati. Kod Slovenije se to najjasnije primjećuje kao što se može vidjeti na slici Sl. 3.41. U Hrvatskoj su najbolji rezultati ostvareni od 07:00 do 11:00 te od 12:00 do 17:00 i od 20:00 do 23:00 sata. Za Srbiju se već slabije može primijetiti isticanje određenih vrijednosti, ali malo bolji rezultati se ostvaruju od 10:00 do 15:00 sati. Za prikaz dijagrama su još uklonjeni i podaci koji su za vrijednost atributa *Impressions* imali 2, a za vrijednost atributa CTR 100.

Na dijagramu sa slike Sl. 3.42. se vidi da su u Sloveniji i Srbiji *Android* i *Windows Phone* OS bolji od iOS-a kada se gleda brojnost podataka s visokim CTR-om i velikim brojem prikaza. U Hrvatskoj su brojniji i bolji *Android* i iOS. U Crnoj Gori je iOS najbrojniji, a u Makedoniji *Android* pa zbog toga oni daju i malo bolje rezultate.

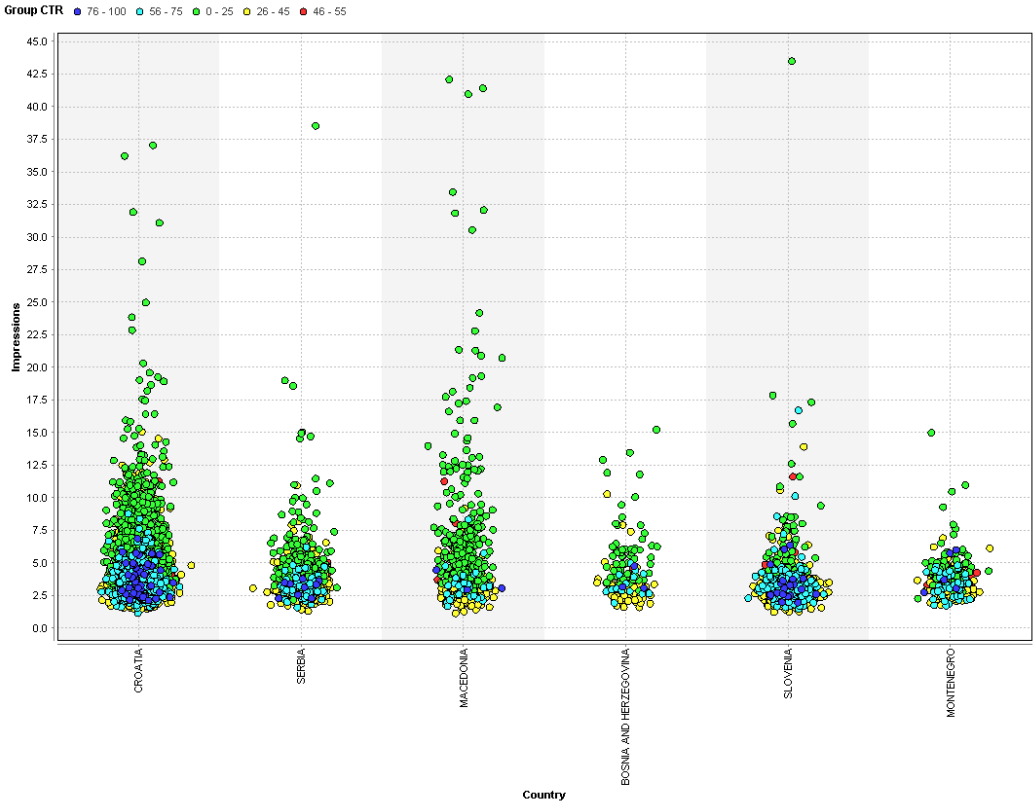


Sl. 3.41. Dijagram rasipanja za atribute *Hour*, *Country* i *Group* CTR.



Sl. 3.42. Dijagram rasipanja za atribute *Country*, *DeviceOs* i *Group* CTR.

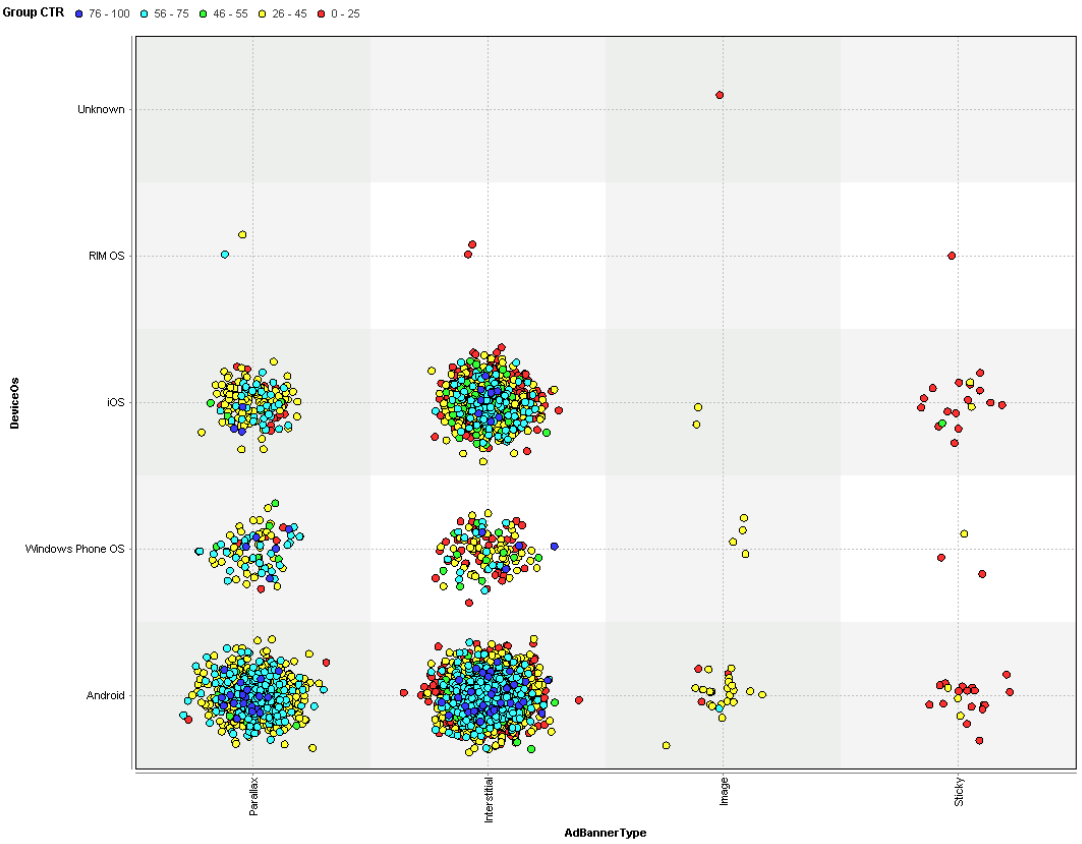
Kada se općenito gleda broj prikaza oglasa i države, na dijagramu sa slike Sl. 3.43. se vidi da se najveći CTR s obzirom na brojnost podataka ostvaruje u Sloveniji i Crnoj Gori, a nakon njih slijede Hrvatska i Srbija.



Sl. 3.43. Dijagram rasipanja za atribute *Country*, *Impressions* i *Group* CTR.

Kod kategorija se ne može previše zaključiti jer nisu sve najuspješnije kategorije bile prisutne u svim državama. Samo je kategorija "*News*" bila u svim državama i svugdje je podjednako uspješna kada se u obzir uzme brojnost podataka.

Na samom kraju napravljen je još jedan dijagram. Za njega su podaci iz prethodnog razmatranja još dodatno filtrirani tako što su odabrane samo kategorije s većim CTR-om što je također prije bilo korišteno. Dijagram je prikazan na slici Sl. 3.44. Na njemu se vidi da *Windows Phone* OS ostvaruje veći CTR u *Parallax* tipu banera nego u *Interstitial*. Kod ostalih *Interstitial* ostvaruje najveći CTR.



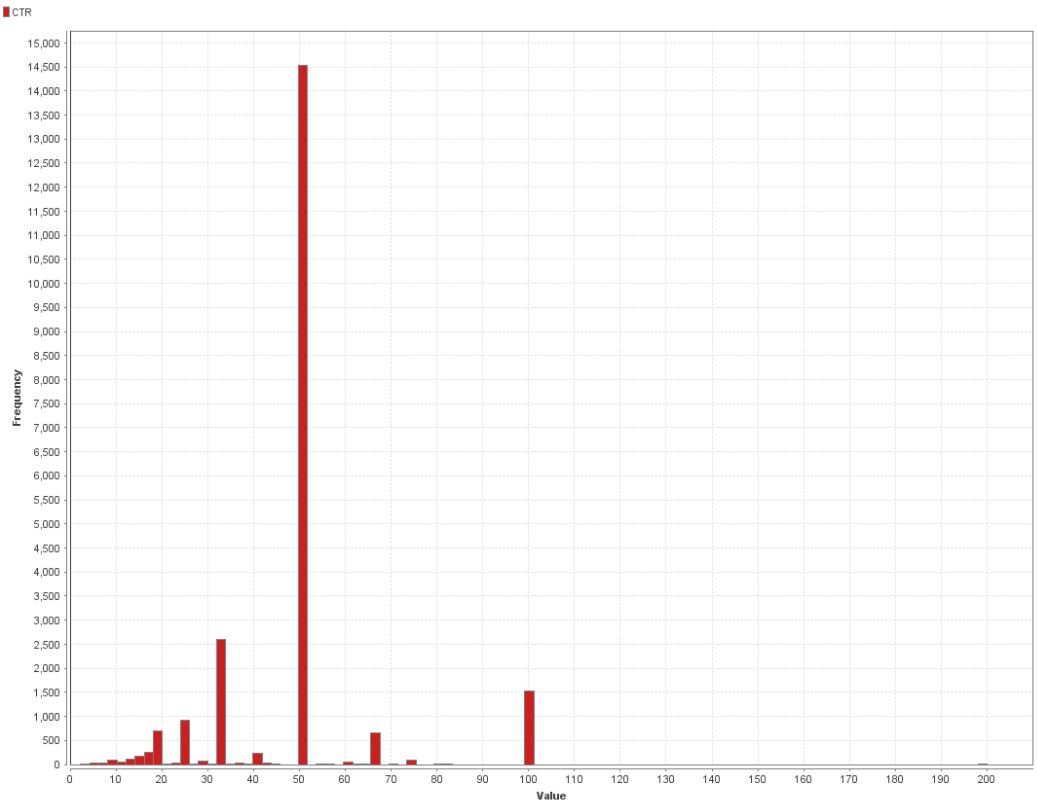
Sl. 3.44. Dijagram rasipanja za atribute *AdBannerType*, *DeviceOs* i *Group* CTR.

4. ANALIZA REZULTATA

Podaci su dosta nejednako raspoređeni po atributima, ali to je i očekivano budući da se radi o stvarnim podacima. Ta nejednakost stvara probleme jer se ne može zaključivati o nečemu za što nema dovoljno podataka. Jedan od takvih primjera je atribut *Categories*. Kod njega postoji velik broj kategorija za koje je broj prikaza bio samo 1 ili 2, dok se kod nekih kategorija pojavljuje velik broj podataka s velikim brojem prikaza.

Najjasniji zaključak je da CTR opada s povećanjem broja prikaza, skoro u svim slučajevima, i to drastično. U razgovoru s domenskim stručnjacima zaključeno je da je to normalna pojava. Osim toga, naglašeno je da oglasi u najboljim slučajevima dostižu vrijednosti za CTR najviše oko 10% u stvarnim sustavima. Ako se gleda CTR u ovoj analizi, može se zaključiti da je sustav za isporuku oglasa jako uspješan, odnosno idealan s obzirom na dominantan broj oglasa s atributom CTR u iznosu od 100%. Ali ako se malo bolje prouče podaci, onda se vidi da nema oglasa koji nisu uopće kliknuti (možda takvi oglasi nisu prikupljani), dok se istovremeno u podacima pojavljuju oglasi koji nisu nikada bili prikazani. U razgovoru sa stručnjacima, kao razlog zašto se pojavljuju oglasi koju nisu prikazani, a bili su kliknuti, rečeno je da u tim slučajevima njihov sustav nije bio odgovoran za isporuku tog oglasa pa se nisu mogli pratiti njegovi prikazi, ali su zabilježeni klikovi na taj oglas.

Također velika većina podataka je prikazana samo jednom i jednom je bila kliknuta. Kada se takvi podaci izbace iz analize, onda je situacija malo drugačija, a može se vidjeti na dijagramu na slici Sl. 4.1. Prosječni CTR je u tom slučaju 48.783, ali ni taj iznos nije realan.



Sl. 4.1. Histogram za atribut CTR.

Tek kada se izbace svi oglasi koji su prikazani manje od 10 puta, za prosječnu vrijednost CTR-a se dobije 14.807 što bi prema razgovoru s domenskim stručnjacima bilo bliže nekim stvarnim vrijednostima za takve sustave. Budući da količina takvih oglasa u ovim podacima nije velika, teško se nad njima mogu vršiti kvalitetnije i detaljnije analize. Za točniji izračun vrijednosti CTR-a broj prikaza oglasa bi trebao biti što veći, najbolje iznad 100. Ako se pogledaju podaci u ovoj analizi, takvih oglasa nema, ali čak i u ovakvim podacima se mogu uočiti razlike između uspješnijih i manje uspješnih oglasa. Također veliki je broj nominalnih atributa spram numeričkih.

Jedan od atributa koji bi bilo dobro imati je atribut koji bi opisivao kojoj kategoriji pripadaju oglasi. Bilo bi zanimljivo analizirati uspješnost pojedine kategorije oglasa te promatrati uspješnost oglasa s obzirom na kategoriju pozicije na internet stranici ili u aplikaciji na kojoj je oglas prikazan.

Za bolji uvid o broju podataka s kojima se radilo i kakav je odnos u broju podataka među različitim vrijednostima atributa napravljene su tablice Tablica 4.1. i Tablica 4.2.

Tablica 4.1. Raspodjela podataka prema *Group* CTR atributu.

Group CTR	Svi podaci osim onih koji imaju Impressions = 0	Odabrane države jugoistočne Europe i bez Impressions = 0	Svi podaci osim onih koji imaju Impressions = 0 Impressions = 1 + Clicks = 1 Impressions = 2 + Clicks = 1	Svi podaci osim onih koji imaju Impressions = 0 Impressions = 1 + Clicks = 1 Impressions = 2 + Clicks = 1 Impressions = 2 + Clicks = 2
76 – 100	231 828	229 999	1 565	99
56 – 75	823	794	823	823
46 – 55	14 530	14 302	291	291
26 – 45	2 954	2 884	2 954	2 954
0 - 25	2 388	2 331	2 388	2 388

Tablica 4.2. Brojnost podataka po različitim kriterijima.

Opis	Broj podataka (oglasa)
Ukupno podataka	253 704
Odabrane države jugoistočne Europe	251 480
Oglasi koji imaju 100% CTR	231 797
Oglasi koji imaju 50% CTR	14 524
Oglasi koji imaju 100% CTR, a prikazani su više od jednom	1 534
Oglasi prikazani samo jednom	230 272
Oglasi prikazani više od jednom	22 251
Oglasi prikazani više od 10 puta	253
Oglasi prikazani 0 puta	1 180
Oglasi prikazani jednom, a kliknuti 2 puta	9
Svi oglasi osim Impressions = 1 + CTR = 100 Impressions = 2 + CTR = 50	8 021
Svi oglasi osim Impressions = 1 + CTR = 100 Impressions = 2 + CTR = 50 Impressions = 2 + CTR = 100	6 555

5. ZAKLJUČAK

U ovom diplomskom radu nastojalo se analizom podataka iz sustava za isporuku oglasa doći do zaključaka i novih saznanja u vezi funkcioniranja sustava za isporuku oglasa, kako bi se moglo poboljšati njegovo djelovanje, odnosno kako bi oglasi ostvarivali veći CTR. U tu svrhu je korištena eksplorativna ili istraživačka analiza podataka.

U radu se prvo proučavalo da li postoji korelacija između atributa koja bi mogla loše djelovati na sustav. Sve uočene veće korelacije su bile i očekivane te nisu predstavljale problem.

Nakon toga se proučavao odnos jednog nominalnog ili numeričkog atributa s ciljnim atributom. Tijekom tog analiziranja su uočeni neki atributi kao što su *Categories, Ad, AdBannerType, DeviceOs, Country, DateHour, Hour* te *Impressions* i *Clicks* koji bi mogli imati veći značaj. Također je zaključeno da se povećanjem vrijednosti atributa *Impressions* smanjuje vrijednost atributa CTR.

Nakon toga proučene su kombinacije dva atributa s ciljnim atributom. U ovom istraživanju se više pažnje posvetilo atributima za koje se prethodno pretpostavilo da imaju veći značaj. Kao rezultat su izdvojene kategorije i vremena u kojima se ostvarivao veći CTR, ali i vrste banera te operacijskog sustava na kojem je prikazan oglas. Kako bi se dobili još jasniji rezultati provedeno je dodatno filtriranje i dijeljenje podataka. Najprije su se filtriranjem izbacili podaci kojih je najviše, a koji sigurno ne daju realnu sliku sustava. Tu spadaju oglasi prikazani samo jednom ili dva puta. Nakon toga su odvojeni oglasi i kategorije kod kojih se pretpostavilo u prethodnim analiziranjima da ostvaruju veći CTR te je donesen zaključak o uspješnosti pojedinih kategorija te o njihovoj zastupljenosti u pojedinim tipovima banera i među operacijskim sustavima uređaja.

Uspješnije kategorije su bile one koje se odnose na vijesti, obrazovanje, tehnologiju, piće, hranu, zdravlje, računarstvo, zabavu i sport, a najčešći i najuspješniji tipovi banera su *Parallax* i *Interstitial*. Također je zaključeno da se najveći CTR najčešće ostvaruje od 08:00 do 10:00 sati te od 14:00 do 17:00 sati. Na kraju su još proučene izdvojene države te su uočene razlike među njima u visini CTR-a ovisno o vrsti operacijskog sustava uređaja, tipu banera i vremenu kada je oglas prikazan.

Tijekom analize su uočeni određeni problemi. Neravnomjerna raspodjela podataka po atributima je onemogućila da se svi atributi promatraju po istim mjerilima, ali to je bilo i očekivano budući da se radi o stvarnim podacima. Također su i vrijednosti koje su dobivene za CTR uglavnom bile 100 što nije realno, a uzrok tome je niska vrijednost atributa *Impressions*. Za oglase

je uobičajeno da se prikazuju više stotina pa i tisuća puta iz čega onda proizlazi puno niži CTR. U ovim podacima je većina oglasa prikazana samo jednom pa do najviše 45 puta. Razlog tome može biti kratki vremenski period iz kojeg su uzeti podaci ili ne bilježenje oglasa ako nisu ostvarili niti jedan klik za određenu kombinaciju atributa.

Atributi iz kojih se izvuklo najviše zaključaka u ovoj analizi su atributi *AdBannerType*, *Categories*, *Country*, *DeviceOs*, *Hour* i *Impressions*. Uzimajući u obzir njihove vrijednosti za koje se zaključilo da ostvaruju veći CTR može se poboljšati rad sustava za isporuku oglasa tako da se oglasi češće oglašavaju na uspješnijim mjestima na internet stranici te pomoću uspješnijih tipova banera. Također se može voditi računa i o tome kada i na kojem mjestu je bolje prikazivati oglas u kojoj državi. Do još boljih i točnijih rezultata bi se došlo kada bi se prikupio puno veći broj podataka koji su prikazani više puta i kada bi se obuhvatio veći vremenski period. Također uvođenje nekih novih atributa, kao što je npr. atribut koji bi opisivao kojoj kategoriji pripadaju oglasi, moglo bi imati pozitivan utjecaj na poboljšanje djelovanja sustava za isporuku oglasa.

LITERATURA

[1] J. Taylor (2009) Guide to Online Advertising [online]. AdJuggler. Dostupno na: http://adjuggler.com/docs/AdJuggler_guidetoonlineadv.pdf [27. lipnja 2016.]

[2] M. Hofmann i R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, Boca Raton, 2013.

[3] D. Gamberger (2011) Otkrivanje znanja dubinskom analizom podataka - Priručnik za istraživače i studente. Zagreb: Institut R. Bošković [online], Verzija 1.46. Dostupno na: http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf [26. lipnja 2016.]

[4] M. Biberović (2015) Domaći AdCumulus bira najefikasnije banere oglašivačima – štedeći im živce (i budžet) [online]. Netokracija. Dostupno na: http://www.netokracija.com/adcumulus-103643 [26. rujna 2016.]

[5] N. R. Barth (2013) Exploratory data analysis [online]. Wikipedia. Dostupno na https://en.wikipedia.org/wiki/Exploratory_data_analysis [22. lipnja 2016.]

[6] A. Kovačević: Materijali sa predmeta SIAP. Fakultet tehničkih nauka, Novi Sad, 2012.

[7] J. J. Filliben (2013) Engineering Statistic Handbook: Exploratory Data Analysis. e-Handbook of Statistical Methods [online]. NIST/SEMATECH. Dostupno na: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm [21. lipnja 2016.]

[8] X. S. Hua, T. Mei, A. Hanjalic, Online Multimedia Advertising: Techniques and Technologies, Information Science Reference, Hershey, 2011.

[9] J. Lazarević: Primena istraživanja i analize podataka u cilju formiranja namenske polise osiguranja za novog klijenta osiguravajuće kuće. Fakultet tehničkih nauka, Novi Sad, 2015.

[10] L. Bilić-Zulle (2008) Materijali sa predmeta Statistika [online]. Farmaceutsko-biokemijski fakultet, Zagreb. Dostupno na: http://mi.medri.hr/assets/P6_Korelacija.pdf [15. rujna 2016.]

[11] J. J. Filliben (2013) Engineering Statistic Handbook: Exploratory Data Analysis. e-Handbook of Statistical Methods [online]. NIST/SEMATECH. Dostupno na: http://www.itl.nist.gov/div898/handbook/eda/section3/scatterp.htm [23. lipnja 2016.]

[12] D. Hafner (2015) Mobile marketing glossary – 30 terms ready to use [online]. Bern: Adcumulus. Dostupno na: http://adcumulus.com/blog-posts/mobile-marketing-glossary-30-terms-ready-to-use/ [26. rujna 2016.]

[13] V. Zezelj (2016) Types of banner ads that will make your brand stand out [online]. Bern: Adcumulus. Dostupno na:

http://adcumulus.com/blog-posts/types-of-banner-ads-that-will-make-your-brand-stand-out/ [26. rujna 2016.]

[14] T. Maier (2014) Sticky Ads Demo [online]. Advanced Ads. Dostupno na: https://wpadvancedads.com/sticky-ads/demo/ [26. rujna 2016.]

[15] J. J. Filliben (2013) Engineering Statistic Handbook: Exploratory Data Analysis. e-Handbook of Statistical Methods [online]. NIST/SEMATECH. Dostupno na: http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm [23. lipnja 2016.]

SAŽETAK

U ovom radu je izvršena eksplorativna analiza podataka iz sustava za isporuku oglasa. Analizom se pokušalo otkriti utjecaj pojedinih atributa na ciljni atribut CTR. Provedenom analizom došlo se do zaključaka da atributi koji opisuju datum i vrijeme prikaza oglasa (*Hour* i *DateHour*), atributi koji opisuju gdje i kako je oglas prikazan (*AdBannerType*, *Categories*, *Country* i *DeviceOS*) i atribut *Impressions* više utječu na ciljni atribut. Tim atributima se u radu posvetila veća pažnja u analizi tako da su dodatno filtrirani te kombinirani međusobno kako bi se osim utjecaja svakog pojedinačnog atributa na ciljni atribut provela i analiza utjecaja različitih kombinacija dva atributa na ciljni atribut. Kao rezultat toga zaključeno je da su uspješnije kategorije one koje se odnose na vijesti, obrazovanje, tehnologiju, piće, hranu, zdravlje, računarstvo, zabavu i sport, da su najčešći i najuspješniji tipovi banera *Parallax* i *Interstitial* i da se najveći CTR najčešće ostvaruje od 08:00 do 10:00 sati te od 14:00 do 17:00 sati.

Na rezultate analize su utjecali nejednaka raspodjela vrijednosti atributa i niski broj prikaza oglasa. Rezultati dobiveni analizom mogu poslužiti za usporedbe s rezultatima budućih sličnih analiza, a također mogu poslužiti osobama koje bolje poznaju način funkcioniranja sustava za isporuku oglasa u unaprjeđenju tog sustava.

KLJUČNE RIJEČI:

Eksplorativna analiza podataka, veliki skup podataka, znanost o podacima, RapidMiner, klikovni postotak.

ABSTRACT

The task of this thesis was to analyze the data from the system for ad delivering using Exploratory data analysis. The analysis attempts to discover the impact of certain attributes on the target attribute CTR. The analysis led to the conclusion that certain attributes such as attributes that describe date and time when the ad was displayed (*Hour* and *DateHour*), attributes that describe where and how the ad was displayed (*AdBannerType*, *Categories*, *Country* and *DeviceOS*) and attribute *Impressions* have higher influence on the target attribute. The thesis devoted more attention to the analysis of these attributes in a way that they were further filtered and combined with each other to see the effect of different combinations of two attributes on the target attribute. As a result it has been concluded that categories related to news, education, technology, drink, food, health, computing, entertainment and sport are the most successful, that most common and most successful types of banners are Interstitial and Parallax and that the highest CTR is usually achieved from 08:00 a.m. to 10:00 a.m. and from 02:00 p.m. to 05:00 p.m.

Unequal distribution of attribute values and low number of ad impressions influenced the results of the analysis. Results obtained from the analysis can be used for comparisons with the results of similar analysis in future, but also can be used by people with better knowledge of how the ad delivery system works to improve that system.

EXPLORATORY ANALYSIS OF DATA FROM AD DELIVERY SYSTEM

KEYWORDS:

Exploratory data analysis, Big Dana, Data Science, RapidMiner, Click-through rate.

ŽIVOTOPIS

Marinko Miljević rođen je 4.2.1992. godine u Starim Mikanovcima. Osnovnu školu Stjepana Antolovića u Privlaci je pohađao u razdoblju od 1999. do 2007. godine. Od 2007. do 2011. je pohađao Gimnaziju Matije Antuna Reljkovića u Vinkovcima, opći smjer. Godine 2011. upisuje preddiplomski studij računarstva na Elektrotehničkom fakultetu sveučilišta Josipa Jurja Strossmayera u Osijeku kojeg i završava 2014. godine te na istom sveučilištu iste godine upisuje diplomski studij računarstva, smjer procesno računarstvo.

PRILOZI

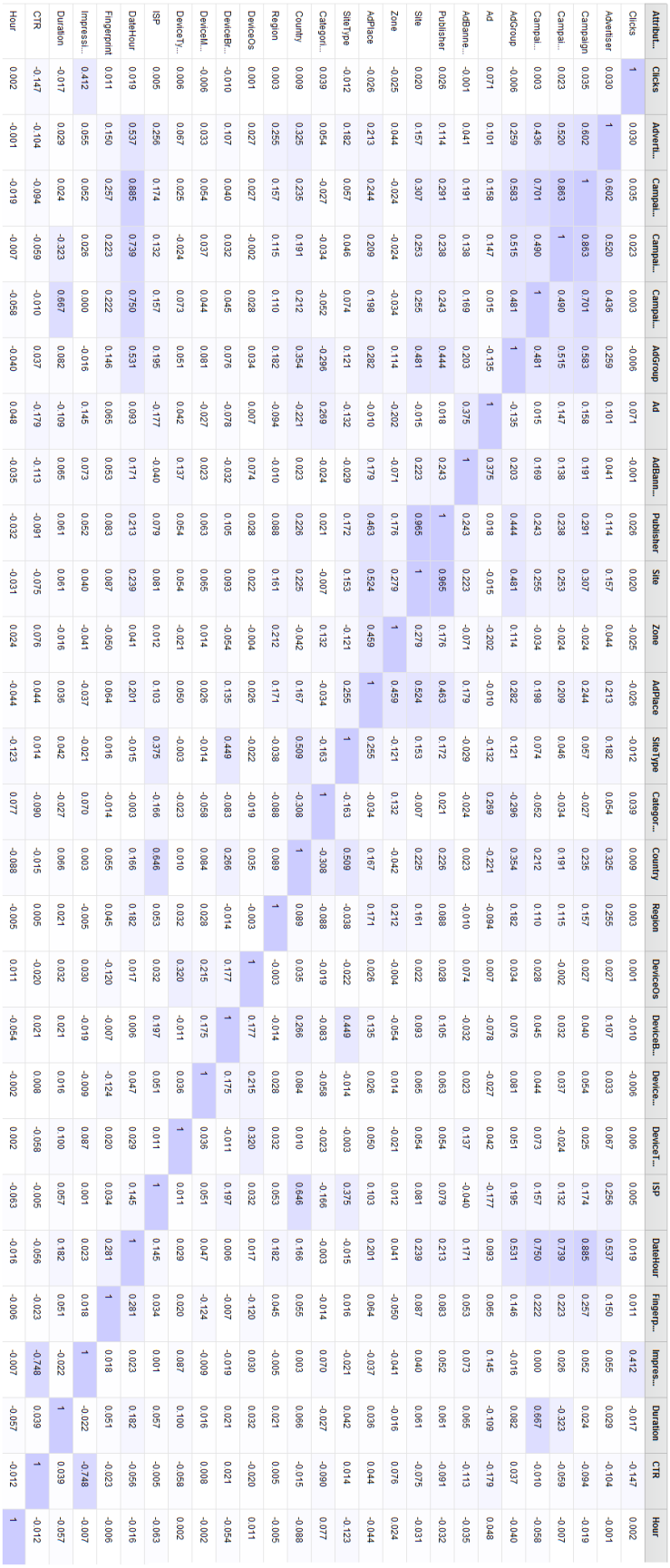
1. Tablice

Tablica 1. Rezultati analize određenih intervala dana u mjesecima travnju i svibnju.

Vrijeme	Hour	Categories	DateHour
1. – 10. travanj	Najbolji rezultati za CTR ostvareni su u terminu od 11:00 do 14:00 sati s tim da se najviše ističe 14:00 sati. Dobri su rezultati i od 00:00 do 01:00 ali tada je također jako velik broj rezultata sa jako slabim CTR-om.	Kod kategorija, najbolji rezultati su ostvareni u kategoriji "Sports", a zatim "News, Education, Uncategorized" te "Careers, Business". Najviše loših rezultata CTR-a je pod kategorijom "NULL".	Najbolji su rezultati ostvareni u ponedjeljak i srijedu.
10. – 20. travanj	Ovdje su rezultati slabije pregledni ali moglo bi se izdvojiti razdoblje od 13:00 do 16:00 sati. Također i 8:00 sati i 22:00 sata.	Od kategorija se ističu "News", "Health & Fintess", "Food & Drink". Također "News" i "NULL" spadaju i u kategoriju s najlošijim CTR-om s velikim brojem visokog broja prikaza.	Najbolji su srijeda, četvrtak i petak ali svakako treba napomenuti da je u tim danima najveći broj prikaza pa je i bilo moguće ostvariti veći CTR na većem broju prikaza. Također kad se gleda i najlošiji dan tu je opet srijeda ali ona 20. travnja.

1. – 10. svibanj	Može se izdvojiti 11:00 sati, razdoblje od 14:00 do 17:00 sati, 21:00 i 22:00 sata te 01:00 sat.	I najbolje i najlošije rezultate daju "Uncategorized, News, Education" i "NULL". Izrazito dobre rezultate daju "News, Uncategorized, Arts & Entertainment" i "Automotive, Arts & Entertainment, News". Jako loše rezultate daje "Health & Fitness".	Nedjelja i ponedjeljak te petak i subota daju najbolje rezultate, ali u tim danima ima i više prikaza oglasa te je mnogo njih s lošim CTR-om.
10. – 20. svibanj	5:00 sati, 10:00 sati i 21:00 - 22:00 sata pokazuju malo bolje rezultate.	Najuspješnije su "News", "Food & Drink, Arts & Entertainment, Style & Fashion, Travel", "Family & Parenting" i "Technology & Computing". Sigurno najbolja među njima je "Arts & Entertainment". Među najlošije se ubrajaju "NULL", "News, Uncategorized" i "Health & Fitness".	Srijeda, četvrtak, ponedjeljak i petak. Od toga srijeda je u oba tjedna uspješna. Jedna srijeda, četvrtak i petak imaju također i dosta oglasa s jako velikim brojem prikaza pa i lošim CTR-om.

2. Slike



Sl. 1. Matrica korelacija (engl. *Correlation Matrix*).