UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Martina Kotar

Vpliv manjkajočih podatkov na analize raziskav v Podjetju X

Diplomsko delo

Ljubljana, 2014

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Martina Kotar

Mentor: red. prof. dr. Vasja Vehovar

Vpliv manjkajočih podatkov na analize raziskav v Podjetju X

Diplomsko delo

Ljubljana, 2014

Univerza
v Ljubljani
Fakulteta
za družbene vede

Kardeljeva ploščad 5
1000 Ljubljana, Slovenija

I Z J A V A O A V T O R S T V U

diplomskega dela

Spodaj podpisani/-a Martina Kotar, z vpisno številko 21070552, sem avtor/-ica diplomskega dela z naslovom: Vpliv manjkajočih podatkov na analize raziskav v Podjetju X.

S svojim podpisom zagotavljam, da:

- je predloženo diplomsko delo izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbel/-a, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena oz. citirana v skladu s fakultetnimi navodili;
- sem poskrbel/-a, da so vsa dela in mnenja drugih avtorjev oz. avtoric navedena v seznamu virov, ki je sestavni element predloženega dela in je zapisan v skladu s fakultetnimi navodili;
- sem pridobil/-a vsa dovoljenja za uporabo avtorskih del, ki so v celoti prenesena v predloženo delo in sem to tudi jasno zapisal/-a v predloženem delu;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata bodisi v obliki skoraj dobesednega parafraziranja bodisi v grafični obliki, s katerim so tuje misli oz. ideje predstavljene kot moje lastne – kaznivo po zakonu (Zakon o avtorski in sorodnih pravicah (UL RS, št. 16/07-UPB3, 68/08, 85/10 Skl.US: U-I-191/09-7, Up-916/09-16)), prekršek pa podleže tudi ukrepom Fakultete za družbene vede v skladu z njenimi pravili;
- se zavedam posledic, ki jih dokazano plagiatorstvo lahko predstavlja za predloženo delo in za moj status na Fakulteti za družbene vede;
- je elektronska oblika identična s tiskano obliko diplomskega dela ter soglašam z objavo diplomskega dela v zbirki »Dela FDV«;
- je diplomsko delo lektorirano in urejeno skladno s fakultetnim Pravilnikom o diplomskem delu.

V Ljubljani, dne 23. 9. 2014 Podpis avtorja/-ice: _____

ZAHVALA

Zahvaljujem se mentorju red. prof. dr. Vasji Vehovarju

za vse strokovne nasvete, pomoč in usmerjanje pri delu.

Posebna zahvala pa gre mojemu Johnu

za zaupanje in podporo.

Vpliv manjkajočih podatkov na analize raziskav v Podjetju X

Kadar v anketni raziskavi manjka del podatkov, ki smo jih sicer pričakovali, imenujemo to manjkajoči podatki. Poznamo različne kategorije manjkajočih podatkov, do katerih pride zaradi različnih razlogov. Anketiranec lahko zavrača sodelovanje, lahko se zgodi, da vprašanje preprosto spregleda, morda noben odgovor zanj ne drži … Obstaja veliko pristopov za obravnavo manjkajočih podatkov. Lahko jih zanemarimo, spregledamo, lahko pa uporabimo posebne pristope. Včasih so manjkajoči podatki nepomembni, nerelevantni in njihovo neupoštevanje ne spremeni statistične analize, obstaja pa možnost, da zaradi neupoštevanja manjkajočih podatkov pride do velikih napak v analizah in interpretacijah. V diplomskem delu sem se poglobila v konkretno raziskavo Podjetja X, kjer me je zanimalo, kakšen vpliv imajo manjkajoči podatki na rezultate. Ugotovila sem, da neupoštevanje dejstva, da nekateri podatki manjkajo, lahko spremeni rezultate, zato je primerno, da se uporabijo postopki za obravnavo manjkajočih podatkov.

Ključne besede: manjkajoči podatki, kategorije manjkajočih podatkov, razlogi in pristopi za obravnavo.

The impact of missing data on research in Company X

When part of data is missing in a survey researh we call that missing data. There are diffrent categories of missing data, which occur for various different reasons. The respodent may be uncooperative, the question may be overlooked, maybe no answer is the right one for the respondent… A variety of approaches have been developed for handling missing data. They can be ignored, overlooked, or we can use specific approaches for them. Sometimes missing data are in fact irrelevant, insignificant and disregarding them does not change statistical analysis. But, there is a possibility that major and significant errors occur in statistical analysis and intepretations, when we fail to consider missing data. In this paper I looked into specific research of Company X, where I was wondering if missing data have impact on the analysis results. I found out, that ignoring missing data can change the results and for that reason, it is appropriate to use specific approaches for handling missing data.

Key words: missing data, category of missing data, reasons and approaches for handling missing data.

Kazalo

Kazalo tabel

Kazalo grafov

1 Uvod

V praksi se problem manjkajočih podatkov pojavlja zelo pogosto. Kljub statističnemu in metodološkemu napredku v preteklih treh desetletjih manjkajoči podatki še vedno predstavljajo probleme v kvantitativnih raziskavah, saj statistiki in raziskovalci niso zagotovili zadostnih informacij o fenomenu teh podatkov (William R. Sterner 2009). Kadar v praksi pridemo do manjkajočih podatkov, jih pogosto preprosto ignoriramo. Toda vprašanje je, kdaj je neupoštevanje primerna rešitev?

Diplomsko delo smo naredili v sodelovanju z uspešnim slovenskih podjetjem, ki kljub krizi, v kateri je Slovenija, pomeni koncept uspešnega poslovanja. Za uspešno vodenje omenjenega podjetja, ki ga bomo v nadaljnje imenovali Podjetje X, je poznavanje potrošnikov ključnega pomena za njihov obstoj in uspešnost. Za posamezno podjetje namreč ni dovolj, da nudi določene izdelke ali storitve. Pomembno je, da zna svoje izdelke ali storitve tudi predstaviti potencialnim strankam. To lahko stori na več načinov. Če svojih obstoječih in potencialnih strank ne pozna, lahko to naredi z neciljnim *(non-target)* oglaševanjem. Če pa svoje stranke pozna, lahko točno ve, kje in kako mora predstaviti nove izdelke, da bodo slišani in videni v čim večjem številu. Podjetje X vlaga del svojega denarja v raziskave na tržišču. Zanima jih, kdo so njihove stranke, kaj najpogosteje kupujejo, kje najraje kupujejo, zakaj se vedno znova odločajo za njihove izdelke. Lahko rečemo, da svoje stranke poznajo.

V nalogi nas bo zanimala specifična raziskava omenjenega podjetja in odgovarjajoča kakovost podatkov. Na sedežu podjetja smo namreč pridobili anketno raziskavo o profiliranju potrošnikov. V tej raziskavi pa, kot v vsaki, manjka del podatkov, ki bi morali biti, a jih ni.

V diplomskem delu bomo s pregledom literature najprej poskusili razumeti, zakaj pride v raziskavah do manjkajočih podatkov, kakšne so tehnike spopadanja z njimi in zakaj so ti podatki sploh pomembni. Sledi podrobna opredelitev manjkajočih podatkov in predstavitev problema, ki je, ugotoviti, kakšno vlogo imajo manjkajoči podatki in kako vplivajo na analize raziskav v Podjetju X.

Najprej bomo pogledali, kakšni tipi manjkajočih podatkov obstajajo in katere izmed njih lahko najdemo v raziskavi Podjetja X. Pričakujemo, da bo manjkajočih podatkov v raziskavi z 22.171 enotami kar veliko, zato menim, da njihov vpliv na končne analize ne bo zanemarljiv.

Sledi empirični del, kjer bomo preskočili načrtovanje raziskave in zbiranje podatkov, saj že imamo raziskavo, ki smo jo dobili na sedežu Podjetja X. Tako se bomo v tem delu naloge takoj poglobili v analizo pridobljenih podatkov. Ta del bo sestavljen iz treh delov:

- obdelava in raziskava pridobljenih podatkov na klasičen način z neupoštevanjem manjkajočih podatkov,
- obdelava in raziskava pridobljenih podatkov z ustrezno metodo obravnave manjkajočih podatkov,
- primerjava obdelav in raziskav pridobljenih podatkov iz zgornjih točk.

Na koncu bomo zaokrožili izsledke naloge. Zanimivo bo videti, kakšne kakovosti so pridobljeni podatki. Glede na to, da gre za pomembne podatke, domnevamo, da so analize, ki so jih izvajali ob predpostavki, da je mogoče manjkajoče podatke ne upoštevati, korektne. Lahko pa se zgodi, da ustrezne tehnike obravnave manjkajočih podatkov pomembno spremenijo rezultate analize raziskav. Naša osnovna hipoteza je, da lahko manjkajoče podatke v tem primeru ne upoštevamo, in to brez posledic za vsebinske analize.

2 Teoretski uvod

2.1 Terminološki uvod

Podatki, s katerimi smo delali analize, so bili pridobljeni na različne načine, zato najprej poglejmo osnovne definicije.

Anketa (*angl. survey)* pomeni v najširšem smislu kakršnokoli statistično opazovanje oziroma zbiranje podatkov, vključno s popisom, na osnovi vnaprej pripravljenega standardiziranega vprašalnika (Vehovar 2007).

Anketno zbiranje podatkov preko interneta, ali natančneje preko svetovnega spleta, pomeni računalniško podprto samoanketiranje, pri katerem anketiranci sami odgovarjajo na anketni vprašalnik, torej brez pomoči anketarja. Pri tem je vprašalnik postavljen na spletno mesto, anketiranci ga izpolnjujejo s pomočjo spletnega pregledovalnika in njihovi odgovori se samodejno preko interneta prenesejo na strežnik raziskovalne organizacije (Lozar Manfreda in drugi 2002).

Telemarketing je načrtna uporaba telefona kot prodajnega medija, v povezavi s tradicionalnimi marketinškimi tehnikami, z namenom povečanja prodaje, zmanjšanja stroškov in povečanja dobička (»Telemarketing« [Klicni center podjetja Lisac & Lisac, d. o. o.], b.d.).

2.2 Manjkajoči podatki

V najširšem smislu delimo manjkajoče podatke na načrtovane in nenačrtovane. Med načrtovane podatke spadajo manjkajoči podatki enot, ki smo jih izločili iz vzorca, in manjkajoči podatki, ki nastanejo zaradi preskokov v vprašalniku. Nenačrtovani podatki pa niso odvisni in pod nadzorom zbiralca podatkov. Sem spadajo manjkajoče vrednosti, ki nastanejo kot posledica neodgovora na posamezno vprašanje ter zavrnitev ali prekinitev sodelovanja pred koncem raziskave (van Buuren 2012).
Manjkajoči podatki so podatki, s katerimi se srečamo v skoraj vsaki anketni raziskavi.

Kadar pri izbrani enoti manjka vrednost določene spremenljivke, govorimo o manjkajoči vrednosti (*missing value)* oziroma manjkajočem podatku (*missing data)*. Dejanska vrednost spremenljivke obstaja tudi v tem primeru, vendar je ne poznamo.

Vprašanja brez odgovora so v tržnih raziskavah povzročena zaradi nezmožnosti ali nepripravljenosti subjekta, da odgovori na specifično vprašanje. Taka vprašanja so velik problem v marketinških raziskavah, saj segajo široko čez razpon različnih tipov marketinških raziskovalnih podatkov (Wagner and Wedel in Factor analysis and missing data, 2000). Te manjkajoče vrednosti pa so pri tržnih raziskavah prej pravilo kot pričakovanje (Wagner A. Kamura and Michel Wedel v Factor analysis and missing data 2000).
Vsak program ima svoj način prikazovanja manjkajočih podatkov. Za SPSS je običajno, da manjkajoče podatke označimo z piko, program R označuje manjkajoče vrednosti s simbolom NA. Kadar beremo podatke iz tekstualnih (ASCII) ali excel datotek, se prazna polja pretvorijo v NA.

2.2.1 Kategorije manjkajočih podatkov

V anketnih raziskavah ločimo tri tipe manjkajočih podatkov, in sicer: nepokritje, neodgovor in ostale vzroke manjkajočih podatkov (Vehovar 2007, 11).

- **Nepokritje** *(non-coverage)* vsebuje manjkajoče podatke, nastale v fazi načrtovanja raziskave. Določene enote iz ciljne populacije v vzorčni okvir sploh niso bile vključene, čeprav vanje sodijo. Problem nepokritja je poseben problem načrtovanja vzorcev in ga le deloma obravnavamo kot problem nepopolnih podatkov. Dejstvo, da bodo določene enote manjkale, smo poznali že ob koncipiranju raziskave, zato tudi nismo pričakovali, da bomo razpolagali z njihovimi vrednostmi.
- **Neodgovor** *(non-response)* nastane v fazi zbiranja podatkov *(field-work stage)* pri enoti, ki je za anketo sicer ustrezna. Ključno za neodgovor je dejstvo, da je anketiranec ustrezen za raziskavo, torej sodi v ciljno populacijo, torej dejanska vrednost spremenljivke obstaja, a je nismo uspeli pridobiti. Pri določeni ustrezni enoti torej nekaterih podatkov nismo uspeli zbrati.

Temu pravimo tudi neodzivnost *(non-responsiveness)* enote. Izraz neodgovor nedvoumno predpostavlja, da je bil anketiranec pred tem izpostavljen nekemu vprašanju, na katerega preprosto ni odgovoril. Razlogi za neodgovore so različni, od nedosegljivosti anketiranca do zavrnitve sodelovanja.

- **Ostali vzroki manjkajočih podatkov** se nanašajo na manjkajoče podatke, nastale v drugih fazah statistične analize: urejanje podatkov, proces kontrole in obdelave podatkov.

Neodgovori se lahko nanašajo na enote ali na spremenljivke.

- **Neodgovor enote** *(unit non-response)* pomeni manjkajoče vrednosti v celotni vrstici podatkovne matrike. V takem primeru nismo uspeli pri določenemu anketirancu zabeležiti odgovora na nobeno vprašanje, zato manjkajo vrednosti za vse njegove spremenljivke. Nekateri temu pravijo tudi popolni neodgovor (*total non-response)*.
- **Neodgovor spremenljivke** *(item non-response)* pomeni, da pri spremenljivki manjkajo vrednosti le pri nekaterih enotah. V anketnih raziskavah pride do neodgovora spremenljivke takrat, ko anketiranci ne želijo odgovoriti na določena vprašanja, se ne morejo odločiti za nobeno izmed ponujenih možnosti, ali takrat ko gre za preskok vprašanja.

Nadaljnja razmejitev neodgovorov (Vehovar 2007):

- Neodgovor in nepokritje; neodgovor enote je, ko je bila le-ta ustrezna in vključena v vzorec, ampak zanjo nismo dobili nobenega odgovora. Nepokritje pa imamo, takrat ko določene ustrezne enote niso bile vključene v vzorec, ker so izpadle iz vzorčnega okvira.
- Neodgovor in neustrezne enote; neustrezne enote so tiste enote, ki so vključene v vzorec, a ne sodijo v ciljno populacijo. Ločimo tuje in prazne enote.
- Neodgovor in odgovor »ne vem«; v primeru, da med odgovori v anketnem vprašalniku ni odgovora *»ne vem«,* lahko neodgovor spremenljivke sporoča, da anketiranec v resnici ne zna odgovoriti na vprašanje. Neodgovor je zato navidezen, saj anketiranec ima odgovor *»ne vem«* in ga je tudi pripravljen izraziti, vendar mu zaprti tip vprašalnika tega ne omogoča.

Glede na proces, ki generira podatke, je Rubin (1976) kategoriziral tri tipe manjkajočih podatkov, ki so postali standard za vse razprave, ki obravnavajo tematiko manjkajočih podatkov. To so: povsem naključno manjkajoči podatki (MCAR), naključno manjkajoči podatki (MAR), nenaključno manjkajoči podatki (NMAR).

Zdaj pa poglejmo, kdaj mehanizma odzivnosti *R* ni treba upoštevati.

- Manjkajoče vrednosti manjkajo popolnoma naključno (MCAR – *missing completely at random*). Verjetnost za manjkajočo vrednost je neodvisna od njene vrednosti, kot tudi od drugih opazovanih vrednosti. Ti podatki so definirani kot vrednosti, ki manjkajo povsem naključno. Podatki spadajo v to skupino, če ni jasnega vzorca manjkajočih vrednosti med katerimikoli poljubnimi spremenljivkami v študiji (Croninger & Douglas, 2005). Manjkajoči podatki so naključno porazdeljeni po naboru podatkov, kjer je enaka možnost, da manjkajo podatki iz katerekoli študije. Verjetnost za manjkajočo vrednost je torej neodvisna tako od njene vrednosti kot tudi od ostalih opazovanih vrednosti. Primer MCAR manjkajočega podatka: anketiranec poroča o svojem osebnem dohodku neodvisno od spola, zakonskega stanu in dohodka. Podatek manjka povsem naključno.
- Manjkajoče vrednosti manjkajo naključno (MAR – *missing at random*). Manjkajoče vrednosti manjkajo naključno, kadar je verjetnost za to pri danih opazovanih vrednostih neodvisna od manjkajočih vrednosti. Mehanizem odzivnosti ima torej lastnost MAR, če imamo dovolj opazovanih podatkov, da lahko z njimi pojasnimo variiranje stopnje odgovorov. Ista realizirana shema mehanizma odzivnosti *R* pa lahko v primeru pomanjkanja teh podatkov pomeni, da imamo lastnosti NMAR. Pogrešanost podatkov ni funkcija manjkajoče vrednosti, ampak bolj funkcija karakteristike udeležencev v raziskavi (Sinharay 2001). Te vrednosti niso naključno porazdeljene po celotni raziskavi, ampak imajo naključne vzorce znotraj specifičnih podskupin. Manjkajoči podatki tipa MAR se torej pojavljajo pogojno na vrednosti opazovanih podatkov opazovanih spremenljivk. Primer MAR manjkajočega podatka: poročeni respondenti v anketnih vprašalnikih raje podajo svoj osebni dohodek kot neporočeni in samski; verjetnost je torej odvisna od njihovega zakonskega statusa in ne od velikosti dohodka.

    - V primeru, da podatki ne manjkajo naključno (MAR), je prisoten ekspliciten mehanizem manjkajočih vrednosti (MDM – *missing data mechanism)*. V takem primeru mehanizem odzivnosti označujemo NMAR *(not missing at random).* Do NMAR podatka pride, če je verjetnost za shemo manjkajočih vrednosti *R* odvisna od vrednosti manjkajočih spremenljivk. Primer NMAR manjkajočega podatka: osebe z višjim osebnim dohodkom v anketnih vprašalnikih raje podajo svoj osebni dohodek kot osebe z nižjih dohodkom; verjetnost je torej odvisna od velikosti dohodka.
- Opazovani podatki so opazovani naključno (OAR – *observed at random)*. V takem primeru je verjetnost za pojav določene sheme neodgovorov *R* neodvisna od opazovanih vrednosti.

Čeprav poznamo ta klasifikacijski sistem, pa je včasih vseeno težko opredeliti, v katero skupino spadajo posamezni manjkajoči podatki, saj se velikokrat zgodi, da je v eni raziskavi kombinacija različnih tipov manjkajočih podatkov (Croninger & Douglas 2005).

Manjkajoče podatke lahko delimo tudi glede na to, ali jih lahko ignoriramo ali ne. V nadaljevanju sta predstavljena dva mehanizma odzivnosti (Vehovar 2007).

- **Zanemarljiv** *(ignorable)* mehanizem odzivnosti. V takšnem primeru mehanizma odzivnosti sploh ni treba upoštevati. Običajno tega mehanizma ne poznamo, če pa ga poznamo in ga vključimo v statistično sklepanje, se postopek zelo zaplete. Za problem manjkajočih podatkov je zato pomembno opredeljevanje pogojev, ki so potrebni in zadostni za zanemarljivost mehanizma. Kadar imamo opravka s takimi manjkajočimi podatki, lahko s posebnimi pristopi izvajamo analize, ki zaradi manjkajočih podatkov niso prizadete. Rubin (1976) je pokazal, da za zanemarljivost tega mehanizma zadošča lastnost MAR.
- Mehanizem **ni zanemarljiv**. Sem spadajo NMAR podatki, ki so pogosto problematični zato, ker je njihova odsotnost povezana z vrednostmi, ki jih raziskovalci niso zmožni opaziti (Sinharay 2001). Kadar imamo opravka s takimi podatki, je analiza bolj zapletena kot pri MCAR in MAR podatkih.

2.2.2 Razlogi za manjkajoče podatke

Razlogov za manjkajoče podatke je veliko. Lahko gre za nedosegljivost anketiranca v času anketiranja, zavrnitev sodelovanja anketiranca, napake, nastale v fazi urejanja podatkov, procesa kontrole in obdelave podatkov. Posebni primeri neodgovora enote nastopajo v panelnih anketah, kjer anketiramo enote po večkrat: **neodgovor v valu** *(wave non-response)* nastane, kadar sicer kooperativne enote ne odgovorijo zgolj v določeni fazi panelne ankete; **osip** *(attrition)* nastopi, kadar enote iz panelne raziskave po določenem času sodelovanja oziroma odgovarjanja dokončno izpadejo iz sodelovanja. Neodgovori spremenljivke pogosto nastanejo zaradi občutljivosti teme določenega vprašanja, o katerem anketiranci ne govorijo radi (Vehovar 2007).

2.2.3 Vzroki za neodgovor enote

Neanketirane enote *(non-surveyed units)* so enote, ki so za raziskavo ustrezne, čeprav anketiranje ni bilo izvedeno iz različnih razlogov.

- Kljub poskusom z določenimi enotami niso vzpostavili stika – **nekontaktirane enote** *(non-contact unit)*, zaradi krajše, daljše ali trajne odsotnosti anketiranca.
- Izrecna **zavrnitev sodelovanja** *(refusal)* nekaterih enot v raziskavi.
- Zaradi jezikovne, psihološke, zdravstvene ali katerekoli druge ovire nekateri anketiranci ne morejo sodelovati v anketi, čemur pravimo **nezmožnost odgovora.**
- Drugi razlogi neanketiranja, zaradi katerih anketarji niso mogli opraviti anketiranja.

Zavržene enote *(deferred units)* so enote, za katere je bilo anketiranje opravljeno, ampak zaradi različnih razlogov v fazi zbiranja podatkov niso bile sprejete v nadaljnjo obdelavo.

2.2.4 Vzroki za neodgovor spremenljivke

Zavrnitev odgovora: predvsem takrat, kadar gre za občutljivo tematiko vprašanja.

Nezmožnost odgovora: kadar je vprašanje anketirancu nerazumljivo ali pa med odgovori ni nakazane možnosti, da bi izrazil svoje mnenje.

Neveljaven odgovor: kadar gre za nejasen, izpuščen odgovor ali grobo tehnično napako pri vnašanju.

2.3 Pristopi za obravnavo manjkajočih podatkov

V preteklem času so različni statistiki razvili različne tehnike spopadanja z manjkajočimi podatki. Pristopi k reševanju problema manjkajočih podatkov so zelo različni (Vehovar 2007).

- Lahko jih delimo glede na nastanek. Tu ločimo metode, ki so nastale v statistični praksi, in sodobne, formalno izdelane metode, ki so jih statistiki razvili posebej za takšne probleme.
- Ločimo tri strategije obravnave manjkajočih podatkov glede na način obravnave nepopolnih podatkov:
    - **metode uteževanja** (z določenimi utežmi poskušamo popraviti napako v ocenah zaradi nepopolnih podatkov),
    - **metode neposredne analize** (analiziramo obstoječo podatkovno matriko, pri čemer moramo upoštevati, da za določena polja nimamo podatkov),
    - **metode vstavljanja** (vstavimo nadomestno vrednosti tam, kjer imamo manjkajoče vrednosti), ki jih delimo na:
        - preprosto vstavljanje povprečne vrednosti opazovanih enot,
        - vstavljanje večjega števila nadomestnih vrednosti na osnovi Bayesovega pristopa.
- Pristopi glede na način statističnega sklepanja:
    - metode, ki temeljijo na klasičnem ali naključnem pristopu, in
    - metode, ki temeljijo na modelskem pristopu.
- Pristopi, ki se nanašajo na predpostavko o fiksnih populacijskih vrednostih:
    - **frekvenlistični** *(frequenlist)* pristop (predpostavlja, da so populacijske vrednosti in parametri fiksni, vzorčne ocene pa okoli njih variirajo) in
    - **Bayesov** pristop (predpostavlja, da so populacijske vrednosti naključne spremenljivke).

V nadaljevanju se bomo ukvarjali izključno z metodami za obravnavo manjkajočih podatkov, vezanih na neodgovor spremenljivke. V tem okviru je – poleg neupoštevanja manjkajočih vrednosti – najbolj običajen pristop vstavljanja. Schafer in Graham (2002) sta definirala imputacijo (vstavljanje) kot proces nadomeščanja manjkajočih podatkov. Obstaja več različnih metod imputiranja. Nedavni napredki v statistični analizi in novejša programska oprema so pokazali rezultate v novih tehnikah imputiranja. Z uporabo te metode imajo raziskovalci dostop do celotnega seta podatkov, kar zmanjša problem moči in posplošljivosti zaradi majhnega vzorca na minimum. Če so podatki v raziskavi zadostni, da lahko napovemo manjkajoče vrednosti, je uporaba te metode v tej raziskavi zanesljivejša in manj pristranska. Uporaba metode imputacije omogoča raziskovalcem razviti popoln set podatkov, s katerim lahko naredijo celostnejše analize, kar omogoča drugim dosleden dostop in analize teh podatkov.

V marketinških raziskavah so najpogosteje uporabljene tehnike spopadanja z manjkajočimi podatki: case/variable deletion (sem spadata listeise in pairwise deletion) in zamenjava/vstavljanje povprečij spremenljivk. Najpogostejši tehniki spopadanja z manjkajočimi podatki sta listwise in pairwise deletion (Peugh & Enders 2004). Pri teh dveh pristopih pa je pomembno, da ju lahko uporabljamo samo takrat, kadar imamo opravka s podatki, ki manjkajo popolnoma naključno (MCAR).

- Case/variable deletion. To klasifikacijo je opredelil William R. Sterner, ki pravi, da morajo raziskovalci najprej razumeti vzorec manjkajočih podatkov, šele nato lahko določijo, kateri pristop je najprimernejši. Pri tem pristopu enote z manjkajočimi vrednostim preprosto odstranimo in analiziramo popolne podatke, ki ostanejo. Brisanje primerov ali spremenljivk je običajen pristop pri soočanju z manjkajočimi vrednostmi. Ta pristop se najpogosteje uporablja v študijah, čeprav je velikokrat nedopusten. Predvsem se ga ne bi smelo uporabljati, kadar imamo opravka z majhnim vzorcem, kadar so manjkajoči podatki naključno porazdeljeni po vzorcu in kadar se ti podatki nanašajo samo na malo število spremenljivk ali primerov (Hair 2006). Prednost tega pristopa je, da se raziskovalci izognejo umetno ustvarjenim podatkom. Sem spadata metodi listwise deletion in pairwise deletion.

- *Pairwise deletion* oziroma analiza podatkov, ki so na voljo. Ta pristop minimalizira izgubo podatkov, ki se zgodi z listwise deletion. To najlažje razumemo s korelacijsko matriko. Korelacije merijo moč povezanosti med spremenljivkami. Za vsak par spremenljivk, za katere imamo na voljo podatke, bo korelacijski koeficient upošteval podatke. Tako ta pristop maksimizira vse razpoložljive podatke na osnovi analize. Čeprav se tej tehniki običajno daje prednost pred listwise deletion, pa prav tako kot omenjena tehnika predvideva, da so podatki MCAR. Pristop pa ima tudi slabe strani, in sicer standardne napake, ki jih izračunava večina programskih paketov, uporabljajo povprečne velikosti vzorca za analize. To pa lahko ustvari standardne napake, ki so podcenjene ali precenjene. Raziskovalci so to tehniko spopadanja z manjkajočimi podatki povezali tudi z nepozitivno definiranimi matrikami v multivariatni in sodobni statistični analizi kot strukturno enačbeno modeliranje.
- *Listwise deletion* oziroma popolna analiza podatkov. Tehnika spopadanja z manjkajočimi podatki občutno zmanjša velikost vzorca, saj odstrani vse enote, ki vsebujejo eno ali več manjkajočih vrednosti (Wagner in Wedel in Factor analysis and missing data 2000), kar je za multivariatne analize nujno, saj pri njih potrebujemo podatke za vse spremenljivke in enote. V večini primerov je ta tehnika spopadanja z manjkajočimi podatki neugodna za uporabo, saj je izguba podatkov prevelika.

- *Mean replacement/substitution* oziroma zamenjava/vstavljanje povprečja, ki sistematično podcenjuje kovariance. Pomanjkljivost tega pristopa je ta, da ne nudi novih informacij. Skupno povprečje se v tem primeru ne spremeni. Toda standardna napaka se z uporabo tega pristopa občutno in umetno zmanjša, saj vzorcu ne dodajamo novih informacij, ampak le povečamo njegovo velikost; neustrezno pa se spremeni tudi porazdelitev. Bistvo povečanja velikosti vzorca je v tem, da povečamo skupni imenovalec za izračun standardne napake, s čimer zmanjšamo standardno napako (Cohen et al. 2003). Postopek pripisovanje sredine spada k nadomeščanju manjkajočih podatkov in računa povprečne vrednosti na podlagi podatkov spremenljivk, ki vsebujejo manjkajoče podatke (Little & Rubin 2002).

- Ko je izračunana povprečna vrednost za vse primere spremenljivke, ki vsebuje manjkajoče podatke, je povprečna vrednost nadomeščena za vse primere, povezane s tisto spremenljivko, ki vsebuje manjkajoče podatke (Vriens & Melton 2002). Raziskovalci uporabljajo to metodo predvsem takrat, kadar imajo opravka z majhnim deležem manjkajočih podatkov v raziskavi in kadar nočejo izbrisati celotnega seta ali posameznih podatkov. Ker se pri tej metodi računa pripisovanje sredine, raziskovalcem ni treba ugibati manjkajoče vrednosti. Dva najpogostejša pristopa pripisovanja sredine sta: računanje sredine iz vseh podatkov, ki so na voljo. To je previdnejši pristop ocenjevanja manjkajočih vrednosti, saj se celotno povprečje udeležencev v raziskavi ne spremeni (Tabachnick & Fidell 2007). Ta metoda je priljubljena med raziskovalci, kadar želijo nadomestiti vrednost, ki je odsev celotne študije in ne posameznega vidika študije. Vstavljanje povprečne vrednosti, posebej za moške, posebej za ženske, ki riše črte med specifičnimi skupinami udeležencev v raziskavi, lahko koristi takoj, ko se manjkajoča vrednost nagiba proti določeni skupini udeležencev (Tabachnick & Fidell 2007). Pri tej metodi je skrb vzbujajoča težnja po rezultatu, ki je blizu povprečja, kar minimalizira variabilnost podatkov in zmanjša korelacije med spremenljivkami (Hair 2006). To je tudi razlog, da narašča konsenz, da ta metoda morda le ni najboljša metoda zamenjave, ampak je med najslabšimi.

Zgornje metode so sicer v praksi zelo pogoste, s statističnih vidikov pa so neprimerne. V primeru MAR podatkov se pogosto uporabljajo naslednji pristopi:

a) *Regression imputation* oziroma nadomeščanje manjkajočih podatkov z regresijo. To je eden od pristopov vstavljanja vrednosti. Ta pristop zagotavlja boljšo oceno manjkajočih podatkov kot vstavljanje povprečij. Ključni koncept te metode je ta, da uporablja napovedane vrednosti manjkajočih podatkov na osnovi razpoložljivih opazovanih spremenljivk (Little & Rubin 2002). Ta metoda običajno uporablja za določanje napovedanih vrednosti postopek najmanjših kvadratov in ima nekaj podobnih značilnosti kot metoda vstavljanja povprečnih vrednosti. Od nje pa se razlikuje v tem, da je posamezna vstavljena vrednost pridobljena iz različnih spremenljivk iz nabora podatkov (McKnight 2007).

Če se raziskovalec odloči za uporabo te metode, mora upoštevati uporabo stohastične metode nadomeščanja regresije in vključiti tudi naključne reziduale. Vključno z dodajanjem naključne napake ta pristop zmanjša nevarnosti za premajhno variabilnost (Little & Rubin 2002).

b) EM ALGORITEM je splošen algoritem za iskanje ML (metoda največjega verjetja pri manjkajočih podatkih) rešitev v primeru manjkajočih vrednosti. Gre torej za metodo ocenjevanja parametrov, ki pa se na koncu lahko pogojno uporabi tudi za vstavljanje manjkajočih vrednosti. Potrebni pogoj je zahteva, da je funkcija verjetja linearna funkcija opazovanih podatkov ali pa odgovarjajočih zadostnih statistik. Prednosti EM algoritma so v preprostosti, saj za oceno ni treba izračunavati odvodov informacijske matrike (Vehovar 2007). Čeprav se EM algoritem zaradi preprostosti in stabilnosti pogosto uporablja, pa ne smemo pozabiti na njegove slabosti. To so: razmeroma počasna konvergenca v primeru velikega deleža manjkajočih vrednosti, nevarnost lokalnih ekstremov in problemi pri zagotavljanju ocene variančno kovariančne matrike, ki pa se v našem primeru niso pojavile oziroma nas niso ovirale. Uporaba EM algoritma za vstavljanje vrednosti pa ima lahko podobne slabosti kot vstavljanje povprečij, predvsem podcenjevanje variance.

c) Z razvojem računalniških metod je vse bolj pogost tudi pristop večkratnega vstavljanja, ki na osnovi določenega modela izvede vstavljanje večkrat, tako da dobimo večkrat izpopolnjeno podatkovno matriko.

Poleg navedenega obstajajo še številne druge metode vstavljanja podatkov. Od *»hot-deck«* do *»nearest neighbour«*, kar v pričujočem pogledu podrobneje ne bomo obravnavali, saj se bomo osredotočili predvsem na primerjavo najpogosteje uporabljenih pristopov neupoštevanj in EM logaritma.

Kadar uporabljamo katerokoli izmed metod vstavljanja vrednosti, moramo upoštevati različne dejavnike (McKnight 2007). Najprej mora biti jasna utemeljitev, da so parametri, ki jih uporabimo za določitev regresijskih koeficientov, empirično ali teoretično podprti.

Poleg tega morajo biti spremenljivke, ki jih uporabimo za napovedovanje manjkajočih vrednosti, različne od spremenljivk, uporabljenih v statističnem modelu. Nato moramo biti pozorni, da za vsako spremenljivko z manjkajočimi vrednostmi najprej opravimo ločen postopek regresije. Zadnji pogoj za uporabo metode vstavljanja vrednosti pa je, da imajo manjkajoči podatki lastnost MCAR ali MAR.

3 Empirični del

V raziskovalnem delu smo analizirali podatke, ki smo jih dobili na sedežu Podjetja X. Gre za zasebno podatkovno bazo podjetja. Podatki so bili pridobljeni na različne načine, na koncu pa so vsi podatki združeni v eni bazi, saj ima podjetje le tako popoln pregled nad svojimi strankami. Del podatkov, zajetih v anketi, je bil pridobljen s telefonskim anketiranjem, del podatkov s spletnim anketiranjem, del pa s klasično anketo, ki jo je izpolnil kupec v trgovini, ko je opravil določen nakup. Iz baze ni razvidno, na kak način je bil anketiran posameznik, zato ni bilo mogoče opraviti analiz ločeno, glede na način pridobivanja podatkov. Navedeno je sicer samo zase že zelo resen problem, ki pa se ga v diplomskem delu ne lotevamo. Osredotočamo se na dejstvo, da v bazi manjka del podatkov, ki smo jih sicer pričakovali, kar bomo obravnavali z različnimi pristopi za delo z manjkajočimi podatki.

V empiričnem delu bomo poskušala ugotoviti, ali lahko pri tej raziskavi zanemarimo manjkajoče podatke ali ne. Naša osnovna hipoteza je, da lahko dejstvo, da podatki manjkajo, zanemarimo, kar tudi pomeni, da podatki manjkajo povsem naključno (MCAR).

H_0 = Podatki manjkajo povsem naključno (MCAR) in analize lahko izvedemo z neupoštevanjem manjkajočih podatkov.

V spodnji tabeli so predstavljene vse spremenljivke, ki imajo vsaj 1,3 % manjkajočih vrednosti. Ostale spremenljivke imajo zanemarljivo majhen delež manjkajočih podatkov (manj kot 0,3 %), zato v tabeli niso prikazane. Povprečje spremenljivk je določeno glede na vse enote. Spremenljivke pa so razporejene glede na število manjkajočih enot, od tiste, ki ima največ manjkajočih enot, na vrhu, do tiste, ki ima najmanj manjkajočih enot, na dnu tabele 3.1. Krepko so poudarjene spremenljivke, ki jih bomo obravnavali v analizi manjkajočih vrednosti.

Tabela 3.1: Frekvence spremenljivk z vsaj 1,3 % manjkajočih vrednosti

Spremenljivke	Manjkajoči podatki		Nemanjkajoči podatki	Povprečje
	Število	Odstotek		
Vstop potrošnikov v klub	17854	80,5	4317	2008,75
Število otrok	14236	64,2	7935	9,04
Trajanje kluba	9870	44,5	12301	16,31
Tip hiše	9800	44,2	12371	2,18
Vstop v klub	9573	43,2	12598	2004,98
Nivo kluba	9568	43,2	12603	1,82
Ocena partnerja	8897	40,1	13274	4,66
Neoglaševana naročila preko katalogov (€)	8570	38,7	13601	42,818
Vrednost naročil preko katalogov (€)	**7401**	**33,4**	**14770**	**0,72**
Izobrazba	**4987**	**22,5**	**17184**	**2,39**
Leto rojstva	**3552**	**16,0**	**18619**	**1957,66**
Odhodni klic (min)	2767	12,5	19404	4125,64
Pošta_obveščanje	2545	11,5	19626	1,30
Epošta_obveščanje	2545	11,5	19626	1,30
Sms_obveščanje	2543	11,5	19628	1,31
Telefon_obveščanje	2543	11,5	19628	1,31
Splošno_obveščanje	2022	9,1	20149	2,61
Status partnerja	1529	6,9	20642	1,56
Prihodni klic (min)	1419	6,4	20752	1875,26
Vrednost vseh nakupov	**1049**	**4,7**	**21122**	**322,179**
Regija	**348**	**1,6**	**21823**	
Spol	**280**	**1,3**	**21891**	

Podatke, ki jih obravnavamo v diplomskem delu, smo dobili kot *.xlsx* datoteko, ki smo jo lahko odprli le v excelu. Naša prva naloga v raziskovalnem delu je torej bila uvoz podatkov v SPSS. Že tu je prišlo do manjkajočih vrednosti, saj ima vsak program svoj način branja podatkov, ki ne sovpada z drugimi. Zato smo morali podatke najprej v *.xlsx* datoteki pretvoriti v tako obliko, ki jo SPSS lahko prebere, šele nato smo lahko začeli z analizami v programu SPSS.

Manjkajoči podatki so bili ločeni:

- Za neodgovor spremenljivke (*item non-response)* je bila v SPSS ».«. To so podatki, ki jih v bazi ni. Sem štejemo:
    - podatke, na katere je odgovor neznan *(»unknown«),*
    - podatke, na katere anketiranci niso želeli odgovoriti,
    - podatke, na katere nimamo odgovora iz drugih razlogov.

  To vrsto manjkajočih podatkov smo v SPSS označili s skupno vrednostjo (−1), kar je lahko v določenih vidikih poenostavitev, saj se lahko manjkajoči podatki, glede na vzroke, obnašajo različno. Podrobnejša obravnava presega obseg diplomskega dela, poleg tega pa v tem pogledu podatki niso ustrezno označeni v sami datoteki.
- Manjkajoči podatki zaradi preskokov – *IF* stavek. To so manjkajoči podatki, ki so pričakovani. V SPSS so bili sicer označeni s ».«, ampak se jih lahko rekonstruira. To vrsto manjkajočih podatkov smo označili z −2.
- *Respondent breakoff* – ko anketiranec preneha odgovarjati na vprašanja. Te vrste manjkajočih podatkov v bazi skoraj ni bilo, kar pa jih je bilo, se pa ni ločilo.
- Za *Null* manjkajoči podatki. V datoteki smo te manjkajoče podatke označili z 0. Gre za podatke, ki niso bili enaki nič (npr. vrednost nakupa je lahko enaka nič), ampak so manjkali.

V raziskovalnem delu diplomskega dela se bomo ukvarjali s podatki, označenimi z −1 (neodgovor spremenljivke).

3.1 Deskriptivne statistike

Preden začnemo z deskriptivno analizo naših spremenljivk, poglejmo celotno bazo podatkov, ki nam bo omogočila celovit vpogled v podatke. Dobljeni vzorec sestavlja 22.171 enot. V bazi imamo 79 različnih spremenljivk. Od vseh spremenljivk jih kar 37 vsebuje manjkajoče vrednosti. Kar 98,6 % celotnega vzorca manjka vsaj ena vrednost, samo za 303 enote imamo vse podatke. Manjka 7,1 % vseh podatkov (tu so šteti vsi manjkajoči podatki v celotni bazi). V diplomskem delu bomo izmed vseh spremenljivk obravnavali 8 izbranih. Med njimi jih 6 vsebuje manjkajoče vrednosti, ki so označene z −1.

Ciljna skupina v diplomskem delu je bil delež slovenskih potrošnikov, pri katerih smo preučevali nakupovalne navade.

Pri kontrolnih spremenljivkah najdemo manjkajoče vrednosti pri starosti, spolu, izobrazbi, tipu stanovanja, številu otrok, oceni partnerja in regiji. Ker je teh spremenljivk veliko, jih bomo v analizi obravnavali le nekaj. To so: spol, leto rojstva, izobrazba in regija. Prav tako bomo med ciljnimi spremenljivkami v analizo vključili le vrednost vseh nakupov (v €), vrednost prvega nakupa, število nakupov preko katalogov, vrednost nakupov preko katalogov.

V spodnji tabeli so prikazane veljavne in manjkajoče vrednosti za izbrane kontrolne in ciljne spremenljivke. Vse izbrane kontrolne spremenljivke vsebujejo manjkajoče vrednosti. Pri ciljni spremenljivki *Skupno število naročil preko katalogov* opazimo, da nimamo manjkajočih vrednosti, medtem ko jih pri *Skupni vrednosti naročil preko katalogov* imamo. Ker je delež manjkajočih vrednosti pri omenjeni spremenljivki zelo velik, lahko domnevamo, da te vrednosti ne manjkajo povsem naključno.

Tabela 3.2: Kontrolne in ciljne spremenljivke

Kontrolne spremenljivke			Ciljne spremenljivke		
Spol	Veljavne	21891	Vrednost prvega nakupa	Veljavne	22171
	Manjkajoče (−1)	280		Manjkajoče	0
Leto rojstva	Veljavne	18619	Vrednost vseh nakupov (−1)	Veljavne	21122
	Manjkajoče (−1)	3552		Manjkajoče	1049
Izobrazba	Veljavne	16732	Število naročil preko katalogov	Veljavne	22171
	Manjkajoče (−1)	5439		Manjkajoče	0
Regija	Veljavne	22138	Vrednost naročil preko katalogov (−1)	Veljavne	14770
	Manjkajoče (−1)	33		Manjkajoče	7401

Zdaj pa poglejmo mersko lestvico za spremenljivke, ki nas bodo zanimale v nadaljevanju.

- Spol: 1 »moški«, 2 »ženski«, 3 »neznan« *(manjkajoča vrednost)*. Spol je nominalna spremenljivka.
- Izobrazba: 1 »osnovna šola«, 2 »srednja šola«, 3 »srednja strokovna šola«, 4 »univerzitetna izobrazba«, 5 »magisterij ali doktorat«, 6 »ne želim odgovoriti« *(manjkajoča vrednost)*, 7 »neznana« *(manjkajoča vrednost)*. Izobrazba je ordinalna spremenljivka.

- Regija: 1 »ljubljanska regija«, 2 »mariborska regija«, 3 »celjska regija«, 4 »kranjska regija«, 5 »novogoriška regija«, 6 »koprska regija«, 7 »novomeška regija«, 8 »murskosoboška regija«, 9 »neznana« *(manjkajoča vrednost)*. Regija je nominalna spremenljivka.
- Vrednost prvega nakupa. Pri tej spremenljivki so respondenti že v vprašalniku poročali v kategorijah in ne v vrednostih. Kategorije so bile naslednje: 1 »ni nakupa«, 2 »od 1 do 20 €«, 3 »od 20 do 50 €«, 4 »od 50 do 100 €«, 5 »od 100 do 200 €«, 6 »od 200 do 400 €«, 7 »od 400 do 600 €«, 8 »od 600 do 800 €«, 9 »od 800 do 1000 €«, 10 »več kot 1000 €«. Vrednost prvega nakupa je ordinalna spremenljivka.
- Leto rojstva je razmernostna spremenljivka, ki sem jo v nadaljnji analizi rekodirala v starostne skupine: 1 »od 0 do 20 let«, 2 »od 21 do 30 let«, 3 »od 31 do 40 let«, 4 »od 41 do 50 let«, 5 »od 51 do 60 let«, 6 »od 61 do 70 let«, 7 »od 71 do 80 let«, 8 »več kot 80 let«.

KONTROLNE SPREMENLJIVKE

SPOL

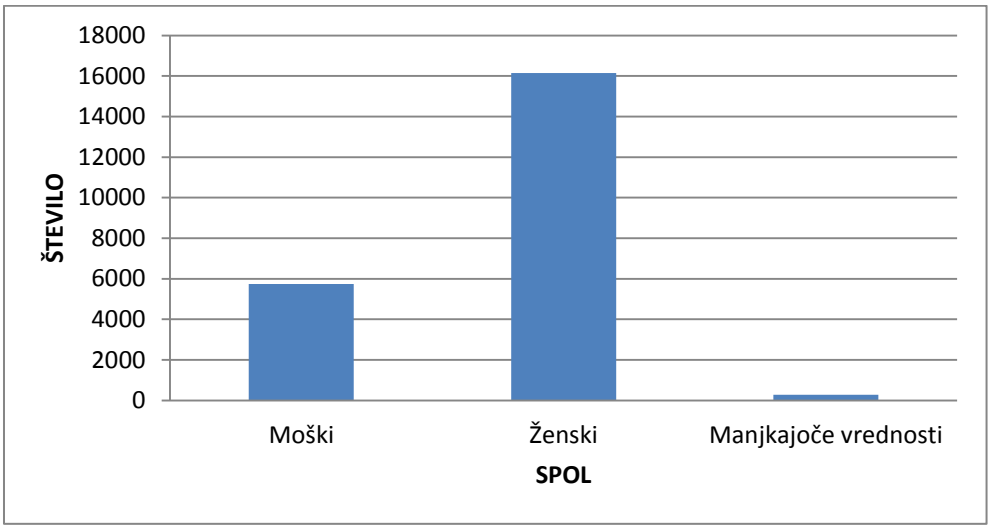
Dobljeni vzorec sestavlja 22.171 enot, med katerimi je 5.740 (25,9 %) moških in 16.151 (72,8 %) žensk. Za 280 (1,3 %) oseb ne vemo, katerega spola so, kot je prikazano v spodnji tabeli.

Tabela 3.3: Frekvenčna porazdelitev za spol

		Število	N % po vrsticah
Spol	Moški	5740	25,9
	Ženski	16151	72,8
	Manjkajoče vrednosti	280	1,3
	Skupaj	22171	100,0

Na grafu 3.1 je spremenljivka *Spol* prikazana še grafično, na histogramu.

Graf 3.1: Histogram za spol



LETO ROJSTVA

V tabeli 3.4 so prikazane opisne statistike za leto rojstva anketirancev. Standardni odklon spremenljivke *leto rojstva* je 14.525. Leto rojstva se od povprečja, ki je 1957.66, standardno odklanja za približno 14,5 let. Najpogostejša vrednost v populaciji (modus) je 1957. To pomeni, da je največ oseb rojenih leta 1957. Koeficient asimetrije je malo večji od 0 (0,051), kar pomeni, da gre pri tej spremenljivki za rahlo asimetrijo v desno. Koeficient mere sploščenosti pa je −0.439, kar pomeni, da gre za sploščeno porazdelitev.

Tabela 3.4: Opisne statistike za leto rojstva

Število enot	Veljavne	18619
	Manjkajoče	3552
Poprečje		1957,66
Modus		1957
Standardni odklon		14,525
Koeficient asimetrije		0,051
Koeficient sploščenosti		−0,439

V tabeli 3.5 je prikazana frekvenčna porazdelitev starosti anketirancev. Arbitrarno smo se odločili, da iz analize odstranimo enote, ki močno odstopajo od povprečja. To so osebe, ki so starejše od 102 let in mlajše od 14 let. V tabeli 3.5 imamo prikazano starost anketirancev od 14 od 102 let. Spremenljivko *Leto rojstva* smo rekodirali v starostne razrede.

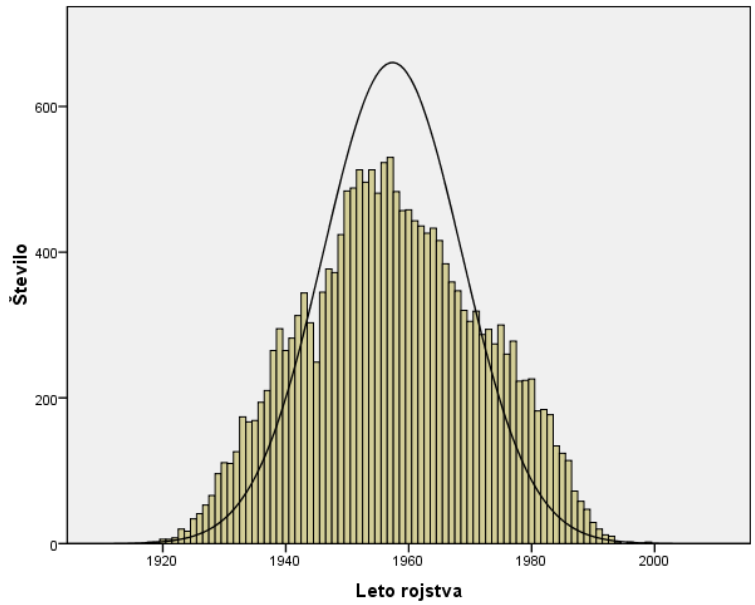
Vidimo, da je največ oseb starih od 51 do 60 let, kar 21,4 % vseh sodelujočih v raziskavi. Najmanj oseb, ki so opravljali nakupe pri Podjetju X, pa je starih do 20 let. To so otroci in mladostniki, ki verjetno nimajo lastnih prihodkov, ki bi jim omogočali samostojne nakupe. Pri spremenljivki *Starost* imamo 3552 manjkajočih vrednosti, kar pomeni, da za 16 % vseh anketirancev ne vemo, katerega leta so bili rojeni.

Tabela 3.5: Frekvenčna porazdelitev za starost

		Število	N % po vrsticah
Starost	14 to 20	6	0,0
	21 to 30	620	2,8
	31 to 40	2328	10,5
	41 to 50	3464	15,6
	51 to 60	4750	21,4
	61 to 70	4051	18,3
	71 to 80	2337	10,6
	80 do 96	1039	4,7
	Manjkajoče vrednosti	3552	16,0
	Skupaj	22147	100,0

V spodnjem grafu 3.2 je prikazana frekvenčna porazdelitev za *Leto rojstva* anketirancev, po odstranjenih enotah, ki so bile rojene pred letom 1918 in po letu 1999. Vidimo, da gre pri tej spremenljivki za porazdelitev, ki se približuje normalni, z rahlo asimetrijo v desno.

Graf 3.2: Histogram za leto rojstva



IZOBRAZBA

V tabeli 3.6 imamo podane opisne statistike za doseženo izobrazbo anketirancev. Izobrazba je ordinalna spremenljivka, ki pa se nekoliko približuje intervalni porazdelitvi, zato lahko ob tej predpostavki, z nekaj poenostavitve, pogojno ilustrativno izračunamo nekatere deskriptivne statistike. Pri tej spremenljivki imamo 5426 manjkajočih vrednosti, kar pomeni, da za 24,5 % vseh sodelujočih v raziskavi ne vemo, kakšno izobrazbo so dosegli. Anketiranci so najpogosteje dosegali vrednost 2, kar pomeni, da najpogosteje dosegali srednješolsko izobrazbo. Koeficient asimetrije je enak 0,625 (> 0), kar pomeni, da je tudi izobrazba asimetrična v desno. Negativni koeficient sploščenosti (−0,451) pa nam pove, da gre pri izobrazbi za sploščeno porazdelitev.

Tabela 3.6: Opisne statistike za izobrazbo

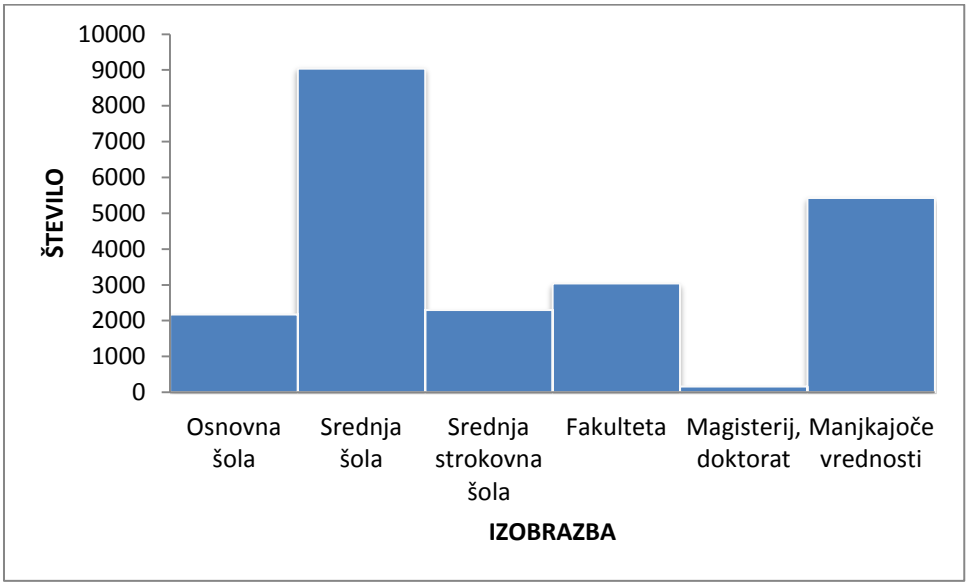
Število enot	Veljavne	16721
	Manjkajoče	5426
Modus		2,00
Standardni odklon		0,96099
Koeficient asimetrije		0,625
Koeficient sploščenosti		−0,451

V tabeli 3.7 je prikazana frekvenčna porazdelitev za izobrazbo potrošnikov Podjetja X. Vidimo, da je največji delež vprašanih dosegel srednješolsko izobrazbo (40,8 %), visoko izobraženih potrošnikov pa je zelo malo, le 14,4 % vseh ima končano univerzitetno ali visokošolsko izobrazbo in magisterij ali doktorat. Izobrazba ima manjkajoče vrednosti tudi v kategoriji *»Ne želim odgovoriti«.*

Tabela 3.7: Frekvenčna porazdelitev za izobrazbo

		Število	N % po vrsticah
Izobrazba	Osnovna šola	2176	9,8
	Srednja šola	9043	40,8
	Srednja strokovna šola	2295	10,4
	Fakulteta	3042	13,7
	Magisterij, doktorat	165	0,7
	Ne želim odgovoriti	451	2,0
	Manjkajoče vrednosti	4975	22,5
	Skupaj	22147	100,0

Graf 3.3: Histogram za izobrazbo



REGIJA

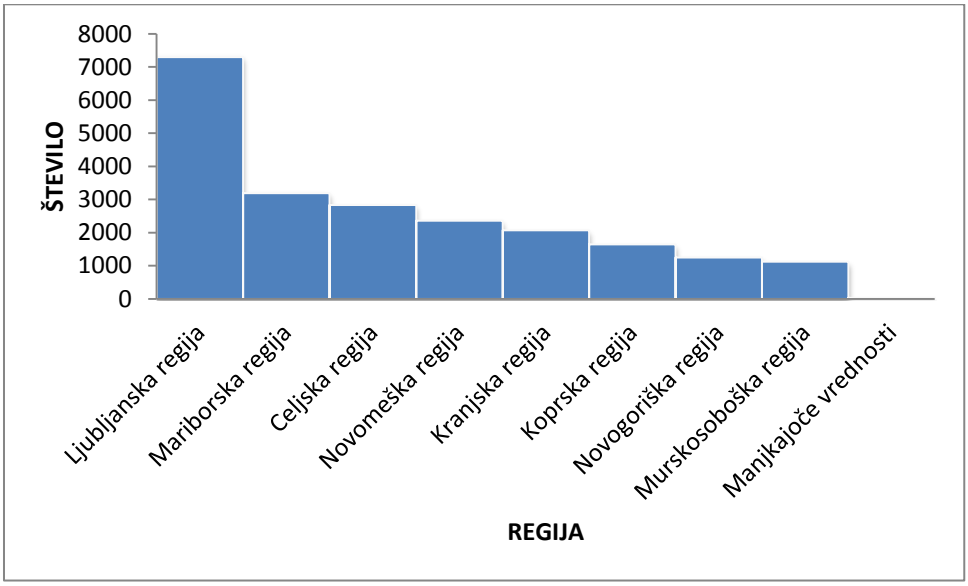
Pomembna spremenljivka je tudi regija, ki nam pove, iz katerega dela Slovenije je največ potrošnikov. V tabeli 3.8 in grafu 3.4 je jasno prikazano, da je to ljubljanska regija. Verjetno zato, ker je to največja slovenska regija in se v njej nahajajo največje trgovine. Najmanjši delež potrošnikov pa je iz murskosoboške regije, le 5,1 %. Nekaj slovenskih potrošnikov je tudi tujih državljanov (315). Pri tej spremenljivki imamo tudi nekaj manjkajočih vrednosti, in sicer 33 (0,1 %). Ker nas zanima samo slovenski delež potrošnikov, smo pri tej spremenljivki arbitrarno odstranili enote, ki niso iz Slovenije.

Tabela 3.8: Frekvenčna porazdelitev za regijo

		Število	N % po vrsticah
Regija	Ljubljanska regija	7301	33,0
	Mariborska regija	3197	14,4
	Celjska regija	2837	12,8
	Novomeška regija	2365	10,7
	Kranjska regija	2074	9,4
	Koprska regija	1645	7,4
	Novogoriška regija	1255	5,7
	Murskosoboška regija	1125	5,1
	Tujci	315	1,4
	Manjkajoče vrednosti	33	0,1
	Skupaj	22147	100,0

V spodnjem grafu je prikazana frekvenčna porazdelitev potrošnikov po odstranitvi enot, ki ne spadajo v našo analizo (tujci). Regija ima samo 33 manjkajočih vrednosti, ki pa se na grafu ne vidijo.

Graf 3.4: Histogram za regijo



CILJNE SPREMENLJIVKE

V nadaljevanju so predstavljene še frekvenčne porazdelitve in opisne statistike za ciljne spremenljivke. To so: vrednost prvega nakupa, skupno število nakupov, število naročil preko katalogov in vrednost naročil preko katalogov. Za te spremenljivke so najprej podane frekvenčne tabele, opisne statistike in histogrami porazdelitve. Število vseh enot posameznih spremenljivk se manjša, ker sproti odstranjujemo enote, ki močno odstopajo od povprečja ali pa iz drugih razlogov ne spadajo v analizo.

VREDNOST VSEH NAKUPOV

Najprej je predstavljena analiza za spremenljivko *Vrednost vseh nakupov* v €. Iz opisnih statistik v tabeli 3.9 vidimo, da ima spremenljivka *Vrednost vseh nakupov* 1048 (5,0 %) manjkajočih vrednosti. *Vrednost vseh nakupov* se od povprečja, ki je 322,179 €, standardno odklanja za 277,0 €. 7 € pa je vrednost, okoli katere so vrednosti v populacije najgostejše. Tudi pri tej spremenljivki nam koeficienta asimetrije in sploščenosti kažeta na rahlo asimetrijo v desno in rahlo koničasto porazdelitev spremenljivke.

Tabela 3.9:Opisne statistike za vrednost vseh nakupov

Število enot	Veljavne	19938
	Manjkajoče	1048
Povprečje		322,179
Modus		7,0
Standardni odklon		277,02
Koeficient asimetrije		1,038
Koeficient sploščenosti		0,339

V tabeli 3.10 so prikazane frekvence za *Vrednost vseh nakupov.* Največ potrošnikov Podjetja X je za nakupe zapravilo med 100 in 200 €. Zelo majhen pa je delež tistih oseb, ki so za nakupe pri Podjetju X zapravili velike količine denarja. Le 3,1 % celotnega vzorca je za nakupe zapravilo več kot 1000 €.

Tabela 3.10: Frekvenčna porazdelitev za vrednost vseh nakupov

		Število	Odstotek
Vrednost vseh nakupov	1 do 10	755	3,6
	11 do 25	1299	6,2
	26 do 50	1296	6,2
	51 do 75	828	3,9
	76 do 100	756	3,6
	101 do 200	4292	20,5
	201 do 300	2367	11,3
	301 do 400	1989	9,5
	401 do 500	1729	8,2
	501 do 750	2536	12,1
	751 do 1000	1437	6,8
	1001 do 1500	654	3,1
	Manjkajoče vrednosti	1048	5,0
	Skupaj	20986	100,0

Tudi za to spremenljivko bomo pogledali, če vsebuje enote, ki preveč odstopajo od povprečja in ga tako kvarijo. Če pogledamo v tabelo ekstremov 3.11, vidimo, da za spodnji ekstrem take vrednosti ne obstajajo, za zgornjega pa. To bomo preverili z *outlier labeling rule,* kjer smo upoštevali *g* faktor v vrednosti 1,5, kar je utemeljil Tukey, J. W. leta 1977 v Exploratory Data Analysis. Izračunan zgornji ekstrem ima vrednost 1187, kar pomeni, da vse prikazane enote za zgornji ekstrem presegajo to vrednost, zato bi jih morali odstraniti in ponoviti *outlier labeling rule.*

Naveden kriterij pa je zelo strog, zato smo ponovili izračune in se arbitrarno odločili, da za zgornjo mejo vzamemo vrednost 47383 in iz nadaljnje analize izločimo samo 3 enote z ekstremnimi vrednostmi.

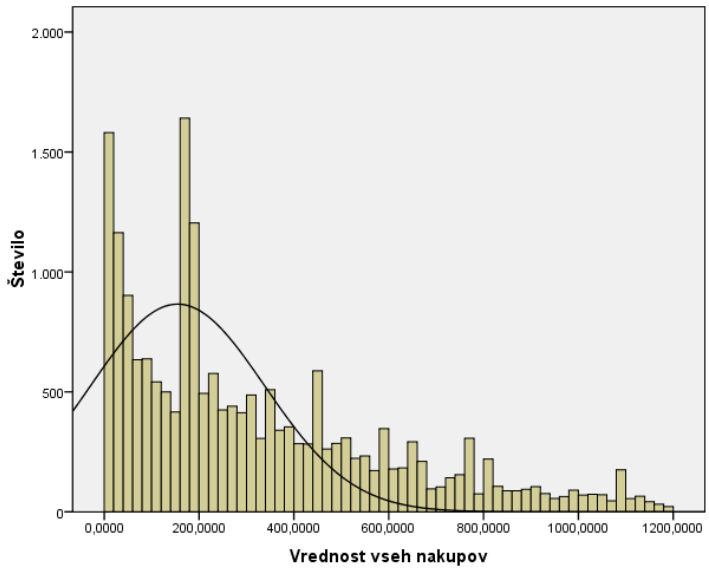
Tabela 3.11: Ekstremne vrednosti za vrednost vseh nakupov

			Številka enote	Vrednost
Vrednost vseh nakupov	Zgornji ekstrem	1	4	144805,0000
		2	18	126562,0000
		3	70	119702,0000
		4	132	47383,0000
		5	141	44988,0000
	Spodnji ekstrem	1	16474	2,0000
		2	16456	2,0000
		3	16407	2,0000
		4	16301	3,0000
		5	16265	3,0000[a]

a. V tabeli z spodnjimi ekstremi je prikazan samo delni seznam enot z vrednostjo 3.

Sledi še histogram za *Vrednost vseh nakupov* po odstranjenih outlierjih. Na grafu so prikazane le enote, ki so za nakupe porabile do 1200 €. Kot smo omenili zgoraj, gre pri omenjeni spremenljivki za močno asimetrijo v desno.

Graf 3.5: Histogram za vrednost vseh nakupov



VREDNOST PRVEGA NAKUPA

V nadaljevanju si poglejmo opisne statistike spremenljivke V*rednost prvega nakupa,* za katero, kot rečeno, respondenti niso navajali točne vrednosti, ampak so se izražali v kategorijah. Vrednost modusa pri vrednosti prvega nakupa je enaka 4, kar pomeni, da so vrednosti populacije najbolj goste okrog razreda »*od 50 do 100 €*«. Vrednosti za omenjeno spremenljivko so bile razvrščene v razrede, kot je prikazano v tabeli 3.13. Največ oseb je za prvi nakup zapravilo med 50 in 100 €, najmanj pa več kot 600 € (23 oseb). Koeficienta asimetrije in sploščenosti nam povesta, da gre v primeru te spremenljivke za koničasto porazdelitev in rahlo asimetrijo v desno, kar je razvidno tudi iz grafa 3.6.

Tabela 3.12: Opisne statistike za vrednost prvega nakupa

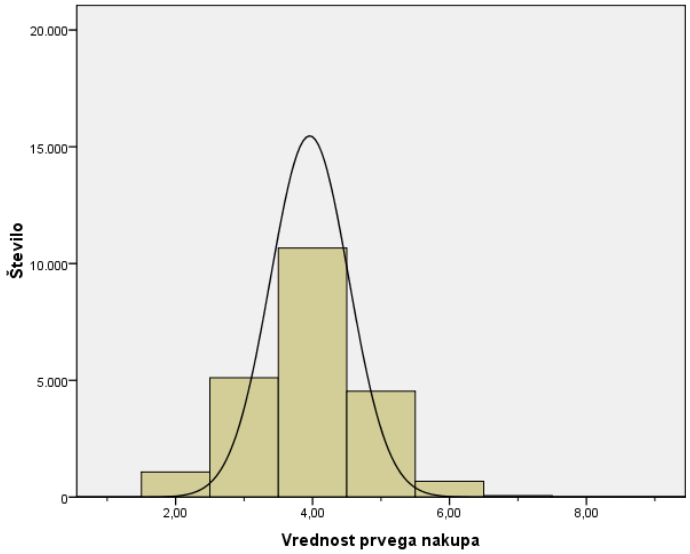
Število enot	Veljavne	22147
	Manjkajoče	0
Modus		4,00
Standardni odklon		0,89423
Koeficient asimetrije		0,189
Koeficient sploščenosti		0,941

Tabela 3.13: Frekvenčna porazdelitev za vrednost prvega nakupa

		Število	N % po vrsticah
Vrednost prvega nakupa	Ni nakupa	3	0,0
	Od 1 do 20 €	1075	4,8
	Od 20 do 50 €	5110	23,1
	Od 50 do 100 €	10663	48,1
	Od 100 do 200 €	4534	20,5
	Od 200 do 400 €	677	3,1
	Od 400 do 600 €	65	0,3
	Od 600 do 800 €	11	0,0
	Od 800 do 1000 €	9	0,0
	Več kot 1000 €	3	0,0
	Skupaj	22147	100,0

Če pogledamo graf 3.6, vidimo, da pri tej spremenljivki ne obstajajo enote, ki bi izstopale iz povprečja *(outlier)*. To smo tudi preverili z *outlier labeling rule,* kjer smo upoštevali *g* faktor v vrednosti 1,5, kar je utemeljil Tukey, J. W. leta 1977 v Exploratory Data Analysis. Iz grafa je tudi razvidno, da gre pri *Vrednosti prvega nakupa* za koničasto porazdelitev z rahlo asimetrijo v desno.

Graf 3.6: Histogram za vrednost prvega nakupa



ŠTEVILO NAROČIL PREKO KATALOGOV

V diplomskem delu pa nas bo zanimala predvsem kataloška prodaja Podjetja X. V tabeli 3.14 so prikazane opisne statistike za spremenljivko *Število naročil preko katalogov.* Povprečje števila naročil je zelo nizko, 0,84. To pa pomeni, da je potrošnik v povprečju opravil manj kot 1 nakup. Tu pride do zelo nizke povprečne vrednosti *Števila naročil preko katalogov* zato, ker več kot polovica (66,0 %) potrošnikov ni opravila nakupa preko kataloga. Najpogostejša vrednost nakupa je tako 0. Keoficienta asimetrije in sploščenosti nam povesta, da gre za spremenljivko, ki je asimetrična v desno in zelo koničasta.

Tabela 3.14: Opisne statistike za število naročil preko katalogov brez izločanja

Število enot	Veljavne	22147
	Manjkajoče	0
Povprečje		0,84
Modus		0
Standardni odklon		1,835
Koeficient asimetrije		4,372
Koeficient sploščenosti		31,134

V tabeli 3.15 je prikazano, koliko nakupov je bilo opravljenih preko katalogov. Vidimo, da imamo za vse osebe prikazano število nakupov, torej pri tej spremenljivki nimamo manjkajočih vrednosti. Pri tem zanemarjamo možnost, da so lahko osebe z manjkajočimi podatki pomotoma izenačene z osebami, ki niso opravile nakupa. V nadaljevanju to možnost izključujemo, ker presega problematiko, ki jo obravnavam v tem diplomskem delu. Ločeno smo sicer opravili tudi analizo brez izločanja, ki jo v nadaljevanju tudi navajamo v opombi.

OPOMBA: ANALIZA ŠTEVILA NAROČIL PREKO KATALOGOV BREZ IZLOČANJA

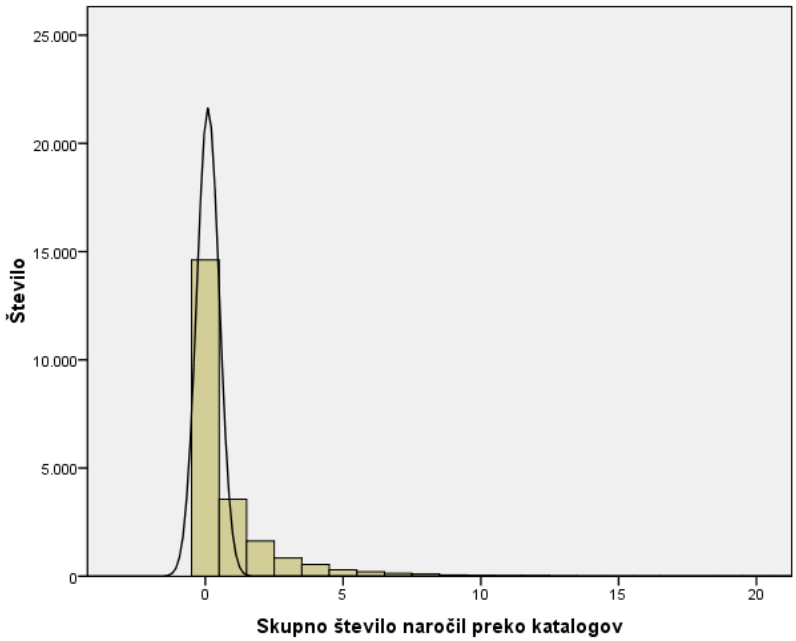
V analizi brez izločanja enot pri spremenljivki *Število naročil preko katalogov* je prikazano, da je bilo največ potrošnikov takih, ki nakupa niso opravili (kar 66,0 %) in veliko (16,1 %) takih, ki so opravili samo 1 nakup. Le 121 oseb je opravilo preko katalogov več kot 10 nakupov.

Tabela 3.15: Frekvenčna porazdelitev za število naročil preko katalogov brez izločanja

		Število	N % po vrsticah
Število naročil preko katalogov	0	14612	66,0
	1	3560	16,1
	2	1625	7,3
	3	842	3,8
	4	544	2,5
	5	297	1,3
	6 do 10	546	2,5
	11 do 20	110	0,5
	21 do 32	11	0,0
	Skupaj	22171	100,0

Sledi še graf porazdelitve nakupov pri analizi brez izločanja. Kot je bilo že prikazano v tabeli 3.14, gre v tem primeru za spremenljivko, ki je zelo koničasta in asimetrična v desno.

Graf 3.7: Histogram za število naročil preko katalogov brez izločanja



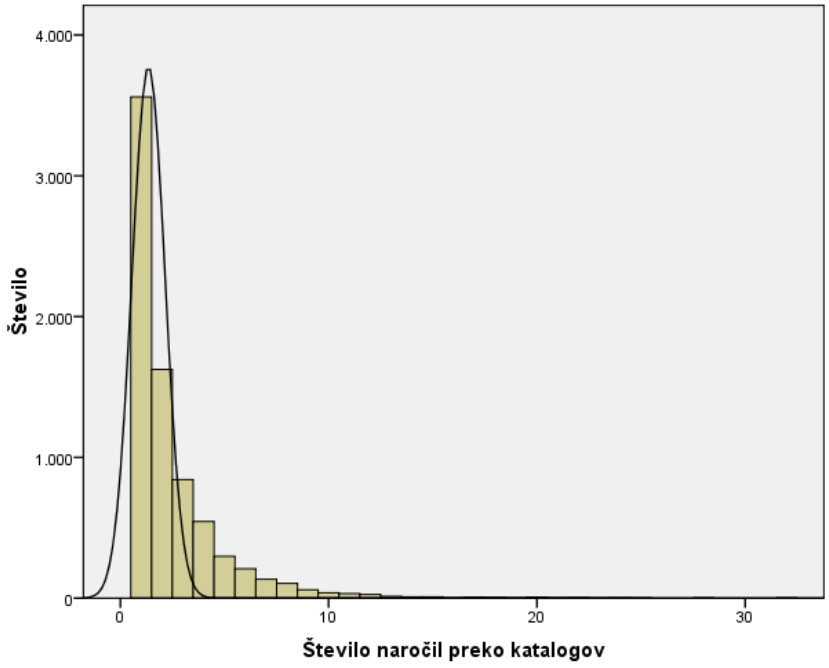
Zdaj pa poglejmo še analizo spremenljivke *Število naročil preko katalogov,* ob predpostavki, da izločimo vse tiste enote, ki niso opravile nakupa (tiste enote, ki so imele pri tej spremenljivki vrednost enako 0). Iz opisnih statistik v tabeli 3.16 vidimo, da se je povprečje spremenljivke povečalo na 2,48. To pomeni, da so potrošniki v povprečju opravili skoraj 2,5 nakupa in ne manj kot enega, kot je bilo prikazano v analizi brez izločanja. Spremenila se je tudi vrednost modusa. Prej je bila ta vrednost enaka 0, sedaj pa je enaka 1. To pomeni, da je največ oseb opravilo 1 nakup. Koeficienta asimetrije in sploščenosti še vedno kažeta na koničasto porazdelitev z asimetrijo v desno, vendar vrednosti koeficientov niso več tako visoke kot v analizi brez izločanja.

Table 3.16: Opisne statistike za število naročil preko katalogov

Število enot	Veljavne	7535
	Manjkajoče	0
Povprečje		2,48
Modus		1
Standardni odklon		2,417
Koeficient asimetrije		3,369
Koeficient sploščenosti		18,266

Poglejmo še spremembe na grafu te spremenljivke. Kot smo omenili že zgoraj, gre tudi po odstranitvi enot z vrednostjo 0 za koničasto porazdelitev z asimetrijo v desno.

Figure 3.8: Histogram za število naročil preko katalogov



VREDNOST NAROČIL PREKO KATALOGOV

V nadaljevanju pa poglejmo, koliko denarja so osebe, ki so opravile nakup preko kataloga, porabile in kakšne so opisne statistike za spremenljivko *Vrednost naročil preko katalogov.*

Pri tej spremenljivki imamo veliko manjkajočih enot, za kar 7397 potrošnikov ne vemo, koliko denarja so porabili za nakupe.

Enote, ki niso naročale preko katalogov, smo že v prejšnjem koraku analize odstranili, tako da imamo tu opravka samo z enotami, ki so opravile nakup preko kataloga. Potrošniki so torej v povprečju porabili 76,91 € za nakup. Če bi upoštevali vse enote (tudi tiste, ki niso opravile nakupa preko kataloga), bi bilo povprečje vrednosti nakupa bistveno manjše. Koeficienta asimetrije in sploščenosti sta pri tej spremenljivki pozitivna in kažeta na asimetrijo v desno in zelo koničasto porazdelitev.

Table 3.17: Opisne statistike za vrednost naročil preko katalogov

Število enot	Veljavne	138
	Manjkajoče	7397
Povprečje		76,91
Modus		33
Standardni odklon		105,885
Koeficient asimetrije		6,669
Koeficient sploščenosti		58,887

Kot smo že omenili, imamo pri *Vrednosti naročil preko katalogov* veliko manjkajočih vrednosti, kar 33,4 % celotnega vzorca. 6 oseb je porabilo več kot 200 € za nakupe preko kataloga, 56 oseb (0,3 %) pa je zapravilo za nakupe med 26 in 50 €.

Tabela 3.18: Frekvenčna porazdelitev za vrednost naročil preko katalogov

		Število	N % po vrsticah
Skupna vrednost naročil preko katalogov	1 do 10	6	0,0
	11 do 25	10	0,1
	26 do 50	56	0,7
	51 do 75	22	0,3
	76 do 100	19	0,3
	101 do 200	19	0,3
	Več kot 200	6	0,1
	Manjkajoči podatki	7397	98,2
	Skupaj	7535	100,0

V spodnji tabeli ekstremnih vrednosti, tabela 3.19, je prikazano, koliko enot izstopa iz povprečja pri spremenljivki *Vrednost naročil preko kataloga*. Vidimo, da ima spremenljivka eno takšno vrednost na zgornjem ekstremu. Ta enota ima vrednost 1078 in kvari povprečje spremenljivke. Ker je enota outlier, ji spremenimo vrednost v 339, tako da enota spada v skupino, ki je za nakupe preko kataloga porabila največ denarja.

Ker smo v prejšnjem koraku analize odstranili enote, ki niso opravile nakupa preko kataloga, zdaj na spodnjem ekstremu nimamo ničelnih vrednosti, ampak imajo enote, ki so porabile najmanj denarja za nakupe preko kataloga, vrednost enako 10.

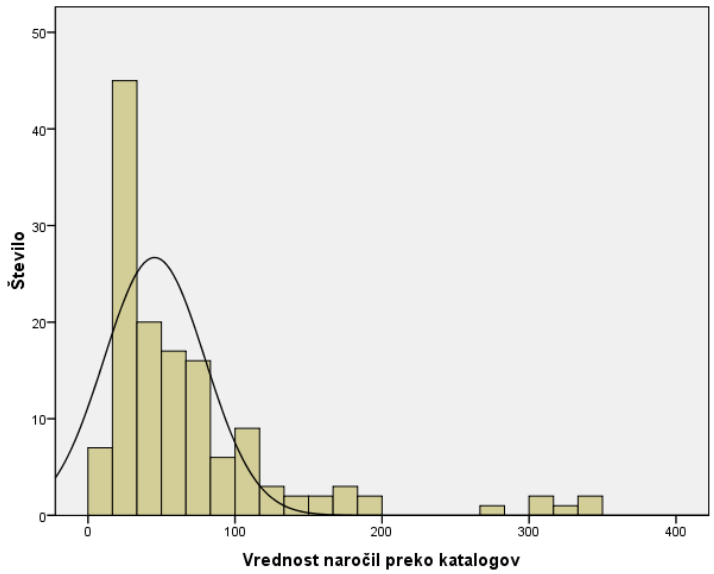
Tabela 3.19: Ekstremne vrednosti za skupno vrednost naročil preko kataloga

			Številka enote	Vrednosti
Skupna vrednost nakupov preko kataloga	Zgornji ekstremi	1	1	1078
		2	2	339
		3	3	333
		4	4	305
		5	5	300
	Spodnji ekstremi	1	138	10
		2	137	10
		3	136	10
		4	135	10
		5	134	10[a]

a. V tabeli z spodnjimi ekstremi je prikazan samo delni seznam enot z vrednostjo 0.

V grafu 3.9 je prikazan histogram za *Vrednost naročil preko katalogov*, brez enot, ki odstopajo od povprečja – *outlierjev*. Iz histograma je razvidno, da je samo en potrošnik porabil za nakupe med 200 in 300 €.

Graf 3.9: Histogram za vrednost naročil preko katalogov



3.2 Analiza glede na spol

Ilustrativno si oglejmo še izbrane bivariatne analize, kjer bomo dobili vtis o povezanosti spremenljivk in morebitnem specifičnem obnašanju manjkajočih vrednosti. Analiza za *Vrednost vseh nakupov* in *Vrednost prvega nakupa* glede na spol je bila narejena na podlagi vseh razpoložljivih podatkov (brez *outlierjev)*, medtem ko je bila analiza za *Število* in *Vrednost naročil preko katalogov* narejena na podlagi enot, ki so opravile vsaj en nakup.

VREDNOST VSEH NAKUPOV GLEDE NA SPOL

Najprej si poglejmo, ali je vrednost vseh nakupov odvisna od spola. V spodnji preglednici so prikazani rezultati. Za celotne nakupe so največ denarja zapravile ženske. Vrednost vseh nakupov je bila tako med moškimi in ženskami največkrat med 101 in 200 €. Kar 21,6 % vseh moških in 19,7 % vseh žensk za vse nakupe pri Podjetju X porabilo omenjeno vsoto. Tam, kjer nimamo podatka o spolu, je bilo prav tako največ vrednosti (35,7 %) med 101 in 200 €. Manjkajoče vrednosti pri nakupih se porazdeljujejo podobno kot pri ostalih znanih vrednostih. Približno tretjino teh nakupov so opravili moški, dve tretjine ženske, dober odstotek pa osebe, čigar spola ne poznamo.

Table 3.20: Frekvenčna porazdelitev vrednosti vseh nakupov glede na spol

		Moški		Ženski		Manjkajoče vrednosti		Skupaj	
		Število	N % vrsta	Število	N % vrsta	Število	N % vrsta	Število	N % vrsta
Vrednost vseh nakupov	1 do 10	205	3,8	543	3,6	7	2,7	755	3,6
	11 do 25	346	6,4	942	6,2	11	4,2	1299	6,2
	26 do 50	328	6,0	960	6,3	8	3,1	1296	6,2
	51 do 75	216	4,0	612	4,0	0	0,0	828	3,9
	76 do 100	195	3,6	560	3,7	1	0,4	756	3,6
	101 do 200	1175	21,6	3017	19,7	100	38,6	4292	20,5
	201 do 300	599	11,0	1747	11,4	21	8,1	2367	11,3
	301 do 400	502	9,2	1473	9,6	14	5,4	1989	9,5
	401 do 500	449	8,2	1261	8,3	19	7,3	1729	8,2
	501 do 750	594	10,9	1910	12,5	32	12,4	2536	12,1
	751 do 1000	346	6,4	1068	7,0	23	8,9	1437	6,8
	1001 do 1500	163	3,0	476	3,1	10	3,9	649	3,1
	Manjkajoče vrednosti	328	6,0	707	4,6	13	5,0	1048	5,0
	Skupaj	5446	100,0	15276	100,0	259	100,0	20981	100,0

VREDNOST PRVEGA NAKUPA GLEDE NA SPOL

Najprej poglejmo primer kontrolne spremenljivke, ki nima večjega deleža manjkajočih vrednosti (le 1,3 %) oziroma s ciljnimi spremenljivkami ni povezana.

Tabela 3.21: Frekvenčna porazdelitev vrednosti prvega nakupa glede na spol

		Spol							
		Moški		Ženski		Manjkajoče vrednosti		Skupaj	
		Število	N % po vrsticah	Število	N % po vrsticah	Število	N % po vrsticah	Število	N % po vrsticah
Vrednost prvega nakupa	Ni nakupa	0	0,00	2	0,01	1	0,36	3	0,01
	Od 1 do 20 €	282	4,91	739	4,58	54	19,29	1075	4,85
	Od 20 do 50 €	1197	20,85	3818	23,64	103	36,79	5118	23,08
	Od 50 do 100 €	2605	45,38	7992	49,48	74	26,43	10671	48,13
	Od 100 do 200 €	1374	23,94	3133	19,4	31	11,07	4538	20,47
	Od 200 do 400 €	243	4,23	422	2,61	13	4,64	678	3,06
	Od 400 do 600 €	30	0,52	33	0,2	2	0,71	65	0,29
	Od 600 do 800 €	4	0,07	7	0,04	0	0	11	0,05
	Od 800 do 1000 €	4	0,07	4	0,02	1	0,36	9	0,04
	Več kot 1000 €	1	0,02	1	0,01	1	0,36	3	0,01
	Skupaj	5740	100	16151	100	280	100	22171	100

Moški so se večinoma (45,3 %) odločali za prvi nakup v vrednosti od 50 do 100 €. Tudi večina žensk (49,4 %) je za prvi nakup zapravila od 50 do 100 €. Za prvi nakup je več kot 1000 € zapravil 1 moški, ena ženska in 3 osebe, čigar spola ne vemo. Tam, kje spol manjka, je posebej veliko vrednosti v kategoriji *»od 20 do 50 €«*, zelo malo vrednosti pa v kategorijah nad 400 €. Za prvi nakup so osebe, za katere nimamo podatka o spolu, porabile manj denarja kot drugi potrošniki.

ŠTEVILO NAROČIL PREKO KATALOGOV GLEDE NA SPOL

Pri tej spremenljivki delamo analize z zelo majhnim vzorcem, saj operiramo le s tistimi enotami, ki so opravile vsaj 1 nakup. V tabeli 3.22 vidimo, da je največ nakupov opravil ženski del populacije. Skoraj polovica moških in žensk je opravila samo 1 nakup. Med osebami, čigar spol nam je neznan, se jih je 7 odločilo za en nakup, ena oseba pa je opravila 4 nakupe.

Tabela 3.22: Frekvenčna porazdelitev števila naročil preko katalogov glede na spol

		Spol							
		Moški		Ženski		Manjkajoče vrednosti		Skupaj	
		Število	N % po stolpcu	Število	N % po stolpcu	Število	N % po stolpcu	Število	N % po stolpcu
Število naročil preko katalogov	1	914	48,6	2639	46,7	7	87,5	3560	47,2
	2	390	20,7	1235	21,9	0	0,0	1625	21,6
	3	196	10,4	646	11,4	0	0,0	842	11,2
	4	140	7,4	403	7,1	1	12,5	544	7,2
	5	60	3,2	237	4,2	0	0,0	297	3,9
	6 do 10	145	7,7	401	7,1	0	0,0	546	7,2
	11 do 20	33	1,8	77	1,4	0	0,0	110	1,5
	21 do 32	3	0,2	8	0,1	0	0,0	11	0,1
	Skupaj	1881	100,0	5646	100,0	8	100,0	7535	100,0

VREDNOST NAROČIL PREKO KATALOGOV GLEDE NA SPOL

Zdaj pa poglejmo še, koliko denarja so porabili potrošniki za nakupe preko katalogov. Vidimo, da je največ moških, ki so opravili nakup, zanj/zanje zapravilo od 26 do 50 €, enako kot ženske. Tam, kjer imamo manjkajoče vrednosti za spol, nimamo nobenih vrednosti na spremenljivki *Vrednost naročil preko katalogov.* To pomeni, da osebe, ki niso navedle spola, tudi niso povedale, koliko denarja so zapravile za nakupe preko katalogov. Večina oseb ima manjkajoče vrednosti na spremenljivki *Vrednost naročil preko katalogov.* Kar 98,4 % moških, 98,1 % žensk in vse osebe, čigar spola ne vemo.

Tabela 3.23: Frekvenčna porazdelitev vrednosti naročil preko katalogov glede na spol

		Spol							
		Moški		Ženski		Manjkajoče vrednosti		Skupaj	
		Število	N % vrsta	Število	N % vrsta	Število	N % vrsta	Število	N % vrsta
Vrednost naročil preko katalogov	1 do 10	0	0,0	6	0,1	0	0,0	6	0,1
	11 do 25	1	0,1	8	0,1	0	0,0	9	0,1
	26 do 50	15	0,8	39	0,7	0	0,0	54	0,8
	51 do 75	5	0,3	17	0,3	0	0,0	22	0,3
	76 do 100	3	0,2	15	0,3	0	0,0	18	0,3
	101 do 200	1	0,1	17	0,3	0	0,0	18	0,3
	Več kot 201	4	0,2	1	0,0	0	0,0	5	0,1
	Manjkajoče vrednosti	1747	98,4	5231	98,1	8	100,0	6986	98,1
	Skupaj	1776	100,0	5334	100,0	8	100,0	7118	100,0

3.3 Analiza glede na starost

Ker nas v diplomskem delu zanima predvsem kataloška prodaja, sem v nadaljevanju predstavila analizo glede na starost in izobrazbo le za spremenljivki *Število* in *Vrednost naročil preko katalogov*.

ŠTEVILO NAROČIL PREKO KATALOGOV

V tabeli 3.24 je prikazano, da so največ naročil opravili posamezniki v starostni skupini *»od 51 do 60 let«*. Mlajši so opravili manj nakupov kot starejši potrošniki. Razlog je morda v tem, da znajo mladi bolj uporabljati internet in več kupujejo preko spletnih trgovin, medtem ko starejši verjetno redkeje nakupujejo preko spleta. Samo 117 mladih (do 30. leta) je opravilo nakup preko kataloga.

Tabela 3.24: Frekvenčna porazdelitev števila naročil preko katalogov glede na starost

	14 do 30		31 do 40		41 do 50		51 do 60		61 do 70		71 do 96		Manjkajoče vrednosti		Skupaj	
	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp
1	60	51,3	322	54,8	519	49,2	859	46,2	749	43,3	595	42,2	193	76,9	3297	47,0
2	28	23,9	131	22,3	227	21,5	403	21,7	378	21,8	305	21,6	38	15,1	1510	21,5
3	13	11,1	55	9,4	117	11,1	232	12,5	204	11,8	158	11,2	10	4,0	789	11,3
4	5	4,3	30	5,1	78	7,4	129	6,9	150	8,7	117	8,3	5	2,0	514	7,3
5	1	0,9	18	3,1	32	3,0	67	3,6	93	5,4	64	4,5	2	0,8	277	4,0
6 do 10	9	7,7	28	4,8	69	6,5	137	7,4	128	7,4	138	9,8	2	0,8	511	7,3
11 do 20	1	0,9	3	0,5	10	0,9	29	1,6	25	1,4	32	2,3	1	0,4	101	1,4
21 do 32	0	0	1	0,2	3	0,3	3	0,2	4	0,2	0	0	0	0	11	0,2
Skupaj	117	100	588	100	1055	100	1859	100	1731	100	1409	100	251	100	7010	100

VREDNOST NAROČIL PREKO KATALOGOV

Glede na to, da so ljudje med 51 in 60 letom opravili največ nakupov, pričakujem, da bodo porabili tudi največ denarja. To je tudi potrjeno v tabeli 3.25. Najmanj pa so za nakupe preko kataloga porabili mladi do 30. leta starosti. Ta spremenljivka ima veliko manjkajočih vrednosti, zato npr. za 116 mladih (do 30. leta) ne vemo, koliko so porabili, čeprav vemo, da so nakup preko kataloga opravili.

Tabela 3.25: Frekvenčna porazdelitev za vrednost naročil preko katalogov glede na starost

	14 do 30		31 do 40		41 do 50		51 do 60		61 do 70		71 do 96		Manjkajoče vrednosti		Skupaj	
	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp
1 do 10	0	0	0	0	0	0	3	0,2	1	0,1	2	0,1	0	0	6	0,1
11 do 25	0	0	0	0	3	0,3	2	0,1	3	0,2	0	0	1	0,4	9	0,1
26 do 50	0	0	7	1,2	5	0,5	16	0,9	7	0,4	16	1,1	3	1,2	54	0,8
51 do 75	1	0,9	0	0	0	0	12	0,6	5	0,3	2	0,1	2	0,8	22	0,3
76 do 100	0	0	1	0,2	4	0,4	4	0,2	4	0,2	5	0,4	0	0	18	0,3
101 do 200	0	0	3	0,5	5	0,5	7	0,4	1	0,1	1	0,1	1	0,4	18	0,3
Več kot 201	0	0	0	0	2	0,2	0	0	0	0	1	0,1	2	0,8	5	0,1
Manjkajoče vrednosti	116	99,1	577	98,1	1036	98,2	1815	97,6	1710	98,8	1382	98,1	242	96,4	6878	98,1
Skupaj	117	100	588	100	1055	100	1859	100	1731	100	1409	100	251	100	7010	100

3.4 Analiza glede na izobrazbo

Tudi pri analizi glede na izobrazbo prikazujemo frekvenčne porazdelitve samo za dve spremenljivki, in sicer za *Število naročil preko katalogov* in *Vrednost naročil preko katalogov*.

ŠTEVILO NAROČIL PREKO KATALOGOV

V tabeli 3.26 so prikazane frekvence za *Število naročil preko katalogov* glede na izobrazbo. Vidimo, da so največ naročil opravile osebe s srednješolsko izobrazbo, najmanj pa osebe z visoko stopnjo izobrazbe. Pri manjkajočih vrednostih izobrazbe je veliko oseb opravilo 1 nakup, nihče pa ni opravil več kot 21 nakupov.

Table 3.26: Frekvenčna porazdelitev števila naročil preko katalogov glede na izobrazbo

	Osnovna šola		Srednja šola		Srednja strokovna šola		Univerza		Magisterij, doktorat		Manjkajoče vrednosti		Skupaj	
	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp	Število	N % stolp
1	411	42,4	1536	45,6	404	43,1	449	46,2	27	47,4	470	66,8	3297	47,0
2	193	19,9	738	21,9	221	23,6	221	22,8	9	15,8	128	18,2	1510	21,5
3	126	13,0	377	11,2	104	11,1	121	12,5	7	12,3	54	7,7	789	11,3
4	81	8,4	274	8,1	70	7,5	64	6,6	3	5,3	22	3,1	514	7,3
5	55	5,7	146	4,3	29	3,1	34	3,5	3	5,3	10	1,4	277	4,0
6 do 10	81	8,4	246	7,3	88	9,4	71	7,3	8	14,0	17	2,4	511	7,3
11 do 20	19	2,0	49	1,5	20	2,1	10	1,0	0	0	3	0,4	101	1,4
21 do 32	4	0,4	5	0,1	1	0,1	1	0,1	0	0	0	0	11	0,2
Skupaj	970	100	3371	100	937	100	971	100	57	100	704	100	7010	100

VREDNOST NAROČIL PREKO KATALOGOV

V tabeli 3.27 vidimo, koliko denarja so porabili potrošniki za nakupe preko kataloga glede na stopnjo izobrazbe. Kot že rečeno, imamo pri tej spremenljivki veliko manjkajočih vrednosti. Za kar 6878 oseb ne vemo, koliko denarja so porabile za nakup preko kataloga. Največ manjkajočih vrednosti imamo pri srednješolski izobrazbi. Med osebami, za katere vemo, koliko denarja so porabile za nakupe, je največ takih s srednješolsko izobrazbo. Med tistimi z visoko stopnjo izobrazbe pa nihče ni povedal, koliko denarja je porabil za nakupe.

Table 3.27: Frekvenčna porazdelitev vrednosti naročil preko katalogov glede na izobrazbo

	Osnovna šola		Srednja šola		Srednja strokovna šola		Univerza		Magisterij, doktorat		Manjkajoče vrednosti		Skupaj	
	N	N % stolp	N	N % stolp	N	N % stolp	N	N % stolp	N	N % stolp	N	N % stolp	N	N % stolp
1 do 10	0	0	4	0,1	1	0,1	0	0	0	0	1	0,1	6	0,1
11 do 25	0	0	4	0,1	0	0	1	0,1	0	0	4	0,6	9	0,1
26 do 50	14	1,4	17	0,5	8	0,9	10	1,0	0	0	5	0,7	54	0,8
51 do 75	1	0,1	16	0,5	1	0,1	2	0,2	0	0	2	0,3	22	0,3
76 do 100	2	0,2	7	0,2	1	0,1	4	0,4	0	0	4	0,6	18	0,3
101 do 200	5	0,5	6	0,2	1	0,1	4	0,4	0	0	2	0,3	18	0,3
Več kot 201	1	0,1	2	0,1	0	0	0	0	0	0	2	0,3	5	0,1
Manjkajoče vrednosti	947	97,6	3315	98,3	925	98,7	950	97,8	57	100	684	97,2	6878	98,1
Skupaj	970	100	3371	100	937	100	971	100	57	100	704	100	7010	100

4 Analiza manjkajočih vrednosti

Kadar imamo manjkajoče podatke, najpogosteje uporabimo tehniko vstavljanja manjkajočih podatkov**.**

Pri analizi manjkajočih podatkov moramo slediti trem osnovnim korakom:

1. opis vzorcev manjkajočih podatkov (kje imamo manjkajoče vrednosti, kakšen je obseg manjkajočih vrednosti, ali imajo pari spremenljivk manjkajoče vrednosti v več primerih, ali so vrednosti podatkov ekstremne, ali podatki manjkajo naključno),
2. ocene povprečij, standardnih odklonov in morebitnih kovarianc oziroma korelacij,
3. vstavljanje manjkajočih vrednosti na osnovi izbrane metode.

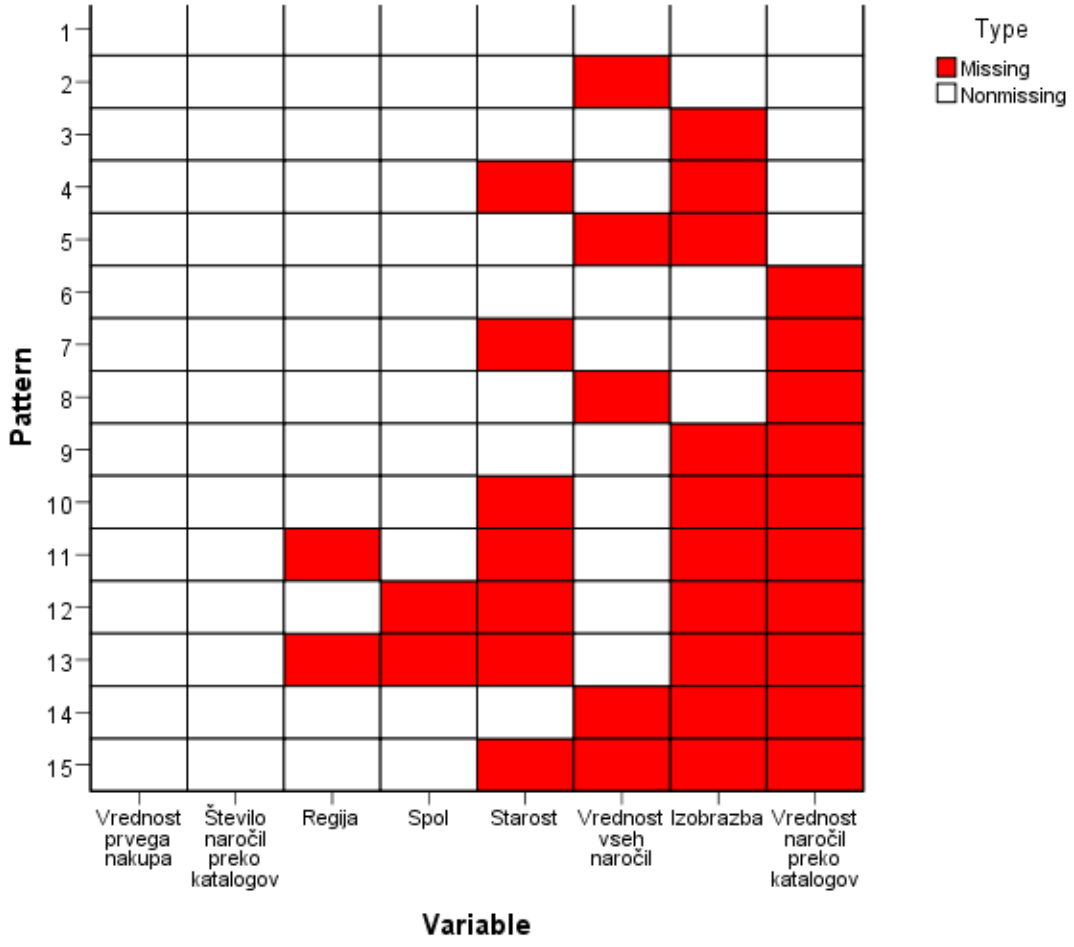
Najprej bomo pregledali odsotnost podatkov v bazi. Naredili smo analizo manjkajočih enot, a le za spremenljivke, ki so nas zanimale v diplomskem delu. V analizo manjkajočih vrednosti smo vključili tudi dve spremenljivki, ki sta sicer ordinalni (*Vrednost prvega nakupa* in *Izobrazba),* a se nekoliko približujeta intervalni porazdelitvi, zato lahko ob tej predpostavki, z nekaj poenostavitve, pogojno izvedemo tudi analizo manjkajočih vrednosti in EM algoritem.

Prikaz vzorca manjkajočih vrednosti in frekvenčni graf vzorca smo dobili v SPSS z ukazom:
»In the main Missing Value Analysis dialog box, select the variable(s) for which you want to display missing value patterns. Click Patterns. Select the pattern table(s) that you want to display«.

V spodnjem grafu je prikazan vzorec manjkajočih vrednosti. Vsaka vrstica v grafu prikazuje skupino enot z enakim vzorcem manjkajočih vrednosti. Vzorci skupin enot so prikazani glede na to, kje ima enota manjkajočo vrednost na posamezni spremenljivki. Prva vrsta v vzorcu kaže spremenljivke, ki ne vsebujejo manjkajočih vrednosti. Druga vrsta prikazuje enote, ki imajo manjkajoče vrednosti pri *Vrednosti vseh naročil*. 10. vrsta prikazuje enote, ki imajo manjkajoče vrednosti na *Starosti, Izobrazbi* in *Vrednosti naročil preko katalogov*. Spremenljivke na x osi so prikazane glede na število manjkajočih vrednosti.

Od leve proti desni sta najprej dve spremenljivki, ki nimata manjkajočih vrednosti (*Vrednost prvega nakupa* in *Število naročil preko katalogov*), sledijo spremenljivke, ki imajo manj manjkajočih vrednosti (*Regija, Spol, Starost* in *Vrednost vseh naročil*), do tistih, ki imajo največ manjkajočih enot (*Izobrazba, Vrednost naročil preko katalogov*). Vidimo, da manjkajoče vrednosti po vzorcu naraščajo. Ker ima naš vzorec manjkajočih vrednosti posamezne »otoke« manjkajočih in nemanjkajočih celic, lahko sklepamo, da odsotnost podatkov ne prikaže monotonosti.

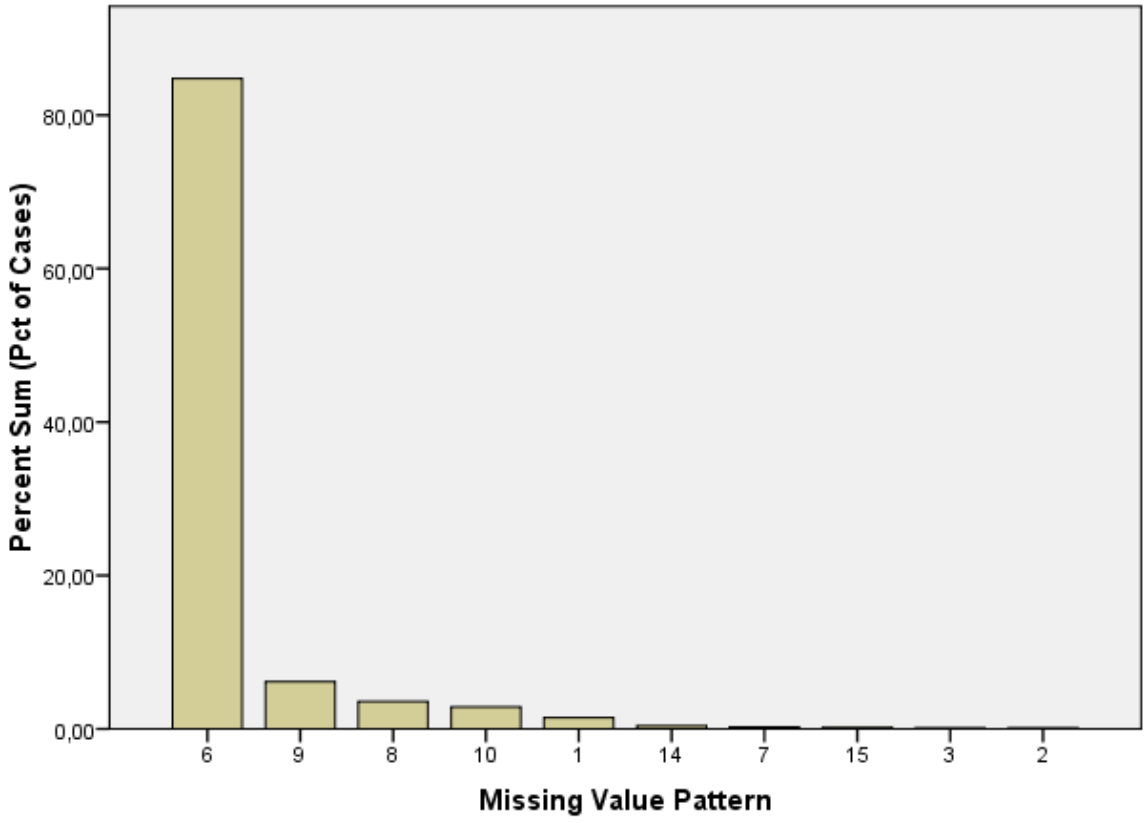
Graf 4.1: Vzorec manjkajočih vrednosti



Poglejmo še frekvenčni graf vzorca. Histogram prikazuje, da je šesti vzorec (tisti, ki ima manjkajoče vrednosti na spremenljivki *Vrednost naročil preko katalogov*) najbolj razširjen, ker se v vzorcu pojavlja najpogosteje.

Drugi najpogostejši je 9. vzorec (manjkajoče vrednosti pri *Izobrazbi* in *Vrednosti naročil preko katalogov*), sledi mu 8. (manjkajoče vrednosti pri *Vrednosti vseh naročil* in *Vrednosti naročil preko katalogov*) in 10. vzorec (manjkajoče vrednosti pri *Starosti, Izobrazbi* in *Vrednosti naročil preko katalogov*). Drugi vzorci se pojavljajo redko, ampak približno enako pogosto. Malo izstopa prvi vzorec, ki se pojavlja malo pogosteje kot 14., 7., 15., 3. in 2. vzorec.

Graf 4.2: Histogram vzorca primernosti



Prikazanih je 10 najpogostejših vzorcev.

Univariatne statistike nam omogočajo vpogled na razširjenost manjkajočih podatkov za vsako spremenljivko posebej. V tabeli 4.1 vidimo, da ima največ manjkajočih enot spremenljivka V*rednost naročil preko katalogov* (6878), sledi ji *Izobrazba* (704), *Vrednost vseh nakupov* (303), *Leto rojstva* (251), *Spol* (8) in *Regija* (3). Povprečja teh spremenljivk so se od povprečij, prikazanih v tabeli 3.1, spremenila. Do razlik pride zato, ker smo v poglavju *Deskriptivne statistike* delali analizo na osnovi vseh razpoložljivih enot, tukaj pa smo analizirali samo določene enote (spremenljivkam smo odstranili outlierje in enote, ki niso opravile nakupa preko kataloga).

Ob odstranitvi enot, ki niso opravile nakupa preko kataloga, se je število potrošnikov občutno zmanjšalo. Analiza v tabeli 3.1 je bila narejena na podlagi 22171 enot, univariatne statistike v tabeli 4.1 pa na podlagi 7010 enot.

Tabela 4.1: Univariatne statistike

	Število manjkajočih enot	Odstotek manjkajočih enot	Število nemanjkajočih enot	Povprečje	Standardni odklon
Vrednost vseh nakupov	303	4,3	6707	322,515	277,0213
Vrednost prvega nakupa	0	0,0	7010		
Število naročil preko katalogov	0	0,0	7010	2,49	2,424
Vrednost naročil preko katalogov	6878	98,1	132	70,15	62,969
Spol	8	0,1	7002		
Leto rojstva	251	3,6	6759	1955,12	13,417
Izobrazba	704	10,0	6306		
Regija	3	0,0	7007		

4.1 Analiza na osnovi popolnih enot *(»Listwise deletion«)*

Za to metodo pridejo v poštev samo popolne enote. To so tiste enote, ki nimajo manjkajočih vrednosti na spremenljivkah. V našem primeru analiziramo 6 spremenljivk in 7010 enot. Za pristop popolnih enot pridejo v poštev samo 103 enote. Samo toliko enot ima vrednosti na spremenljivki *Vrednost naročil preko katalogov.* Vse ostale enote pri tej spremenljivki nimajo vrednosti, zato ne pridejo v poštev za analizo na osnovi popolnih enot. Ta pristop je zelo zmanjšal velikost vzorca (za 6872 enot) in ni uporabil vseh informacij.

4.2 Analiza podatkov, ki so na voljo (»Pairwise deletion«)

Drugi pristop analize manjkajočih vrednosti je analiza podatkov, ki so na voljo. Ta metoda pregleda pare spremenljivk in uporabi posamezno enoto samo, če ima nemanjkajoče vrednosti na obeh parih spremenljivk. Ker so druge manjkajoče vrednosti v posameznem primeru ignorirane, korelacijska in kovariančna matrika za dva para spremenljivk nista odvisni od manjkajočih vrednosti na katerikoli drugi spremenljivki.

V nadaljevanju poglejmo, koliko enot pogleda metoda za posamezne pare spremenljivk. Pri parih spremenljivk, kjer nimamo manjkajočih vrednosti, imamo vzorec 7010 enot.

Najmanj enot imamo pri spremenljivki *Vrednost naročil preko katalogov* (112 enot) v paru z *Izobrazbo.* Sledi ji *Leto rojstva* (123 enot) v paru z *Vrednostjo naročil preko katalogov*.

Tabela 4.2: Pairwise frekvence

	Izobrazba	Vrednost prvega nakupa	Vrednost vseh nakupov	Število naročil preko katalogov	Vrednost naročil preko katalogov	Leto rojstva
Izobrazba	6306					
Vrednost prvega nakupa	6306	7010				
Vrednost vseh nakupov	6047	6707	6707			
Število naročil preko katalogov	6306	7010	6707	7010		
Vrednost naročil preko katalogov	112	132	122	132	132	
Leto rojstva	6288	6759	6470	6759	123	6759

4.3 EM algoritem in primerjava podatkov

Preden začnemo z EM algoritmom, se moramo prepričati, da so naši podatki manjkajoči popolnoma po naključju. To naredimo z *Little's missing completly at random* testom. Ker je vrednost $\pi <$ 0,5 (0,000), ne moremo sprejeti ničelne hipoteze in skleniti, da podatki manjkajo povsem naključno (MCAR). Obstaja torej sistematični vzorec, po katerem podatki manjkajo, pristranskost v naših podatkih. To pomeni, da manjkajočih podatkov ne moremo ignorirati, ker obstaja razlog, zakaj podatkov ni (ni naključje). Lahko, da so respondenti občutljivi na določene teme in zato ne morejo odgovoriti na naša vprašanja in tako prinesejo določeno pristranskost v vzorec. V nadaljevanju bomo zato predpostavili, da podatki manjkajo MAR, in jih obravnavali na osnovi različnih metod, ki so primerne za tak tip manjkajočih podatkov.

Dodati velja, da brez nekih zunanjih preverjanj ne moremo zares potrditi, če velja MAR; ali pa imamo ne neznan mehanizem manjkajočih vrednosti (MDM). Predpostavka MAR je sicer običajna v statistični analizi.

Za analizo v SPSS smo uporabili privzeti EM algoritem, ki vstavlja vrednosti le za razmernostne spremenljivke. Za te spremenljivke uporablja normalno porazdelitev.

V SPSS izvedemo analizo na osnovi EM algoritma z ukazom: *»From the menus choose: Analyze > Missing Value Analysis. Select at least one quantitative (scale) variable for estimating statistics and optionally imputing missing values«.*

EM algoritem je nato na osnovi zgornjih algoritmov vstavil ML vrednost za vse manjkajoče podatke, tako da manjkajočih vrednosti v popolnjeni matriki ni več. Zato lahko analiziramo vseh 7010 enot. Pri tem pa ne smemo pozabiti, da niso povsod pravi podatki.

4.3.1 Analiza povprečij

Rezultati analize povprečij so prikazani spodaj. Najprej so podana povprečja za analizo na osnovi popolnih enot (listwise), sledijo povprečja spremenljivk za analizo na osnovi podatkov, ki so na voljo (pairwise) in povprečja po metodi EM algoritma. Na koncu poglavja je še primerjava povprečij.

V tabeli 4.3 so prikazana povprečja spremenljivk za 138 enot, ki niso imele manjkajočih vrednosti. Pri tej metodi se spremenijo tudi povprečja spremenljivk, ki nimajo manjkajočih vrednosti, saj se iz analize izključi tudi tiste enote, ki imajo na primer vrednost za *Leto rojstva*, nimajo pa vrednosti za S*pol*. Zelo nizko vrednost ima še vedno spremenljivka *Število naročil preko katalogov* (1,29), medtem ko se je povprečje za *Vrednost naročil preko katalogov* že povečalo (65,45). Čeprav sta *Izobrazba* in *Vrednost prvega nakupa* ordinalni spremenljivki, se približujeta nominalni, zato smo tudi zanju izračunali povprečje, ki pa nam ne da jasne interpretacije. Pove nam samo to, da se povprečna izobrazba nahaja nekje med 2. (srednja šola) in 3. stopnjo izobrazbe (srednja strokovna šola). Povprečna vrednost pri spremenljivki *Vrednost prvega nakupa* je 4,126, kar pomeni, da je 138 potrošnikov, ki so bili zajeti v listwise deletion metodo, v povprečju porabilo za prvi nakup nekje med 50 in 100 € (kategorija 4 pri omenjeni spremenljivki) in od 100 do 200 € (kategorija 5 pri V*rednosti prvega nakupa*).

Tabela 4.3: Povprečja Listwise

Število enot	103
Vrednost vseh nakupov	313,834
Vrednost prvega nakupa	4,126
Število naročil preko katalogov	1,29
Vrednost naročil preko katalogov	65,45
Leto rojstva	1954,26
Izobrazba	2,25

V spodnji tabeli 4.4 so prikazana povprečja po parih spremenljivk. Povprečje *Vrednosti vseh nakupov* ob prisotnosti *Vrednosti prvega nakupa* je 322,515. Povprečje *Leta rojstva* je ob prisotnosti katerekoli spremenljivke podobno. Razlikuje se za manj kot eno leto. Vidimo, da so povprečja katerekoli spremenljivke ob prisotnosti ostalih spremenljivk podobna. Malo izstopa par *Število naročil preko katalogov* ob prisotnosti *Vrednost naročil preko katalogov.* Tu je povprečna vrednost 1,27 nakupa, kar je manj kot pa v odnosu z drugimi spremenljivkami, kjer je vrednost 2,49 in več.

Tabela 4.4: Povprečja Pairwise

	Izobrazba	Vrednost prvega nakupa	Vrednost vseh nakupov	Število naročil preko katalogov	Vrednost naročil preko katalogov	Leto rojstva
Izobrazba	2,3298	4,0592	320,8710	2,58	67,84	1955,02
Vrednost prvega nakupa	2,3298	4,0521	322,5151	2,49	70,15	1955,12
Vrednost vseh nakupov	2,3233	4,0510	322,5151	2,50	68,14	1955,00
Število naročil preko katalogov	2,3298	4,0521	322,5151	2,49	70,15	1955,12
Vrednost naročil preko katalogov	2,2768	4,0909	314,0901	1,27	70,15	1955,38
Leto rojstva	2,3286	4,0580	321,6904	2,53	67,09	1955,12

Tabela 4.5 prikazuje povprečja analize EM algoritma. Povprečja so pri tej metodi drugačna kot pri drugih dveh. Do največje razlike pride pri spremenljivki *Vrednost naročil preko katalogov*, kjer povprečje naraste na 72,3 € (pri metodi Listwise je bilo povprečje enako 65,45) in pri spremenljivki *Število naročil preko katalogov,* ki naraste iz 1,29 (pri metodi Listwise) na 2,49. Spremenljivka *Vrednost naročil preko katalogov* je imela 7401 manjkajočih vrednosti od skupno 22171 (izločili smo sistemske missinge, ki tega vprašanja niso mogli dobiti, saj niso opravili nakupa preko kataloga oziroma so imeli vrednost enako 0). Povprečje naraste tudi pri spremenljivki *Vrednost vseh nakupov* za skoraj 10 €. Pri nobeni spremenljivki pa ne govorimo o zmanjšanju povprečne vrednosti. Povprečje pri *Letu rojstva* je bilo na podlagi vseh razpoložljivih vrednosti 1957,66 (tabela 3.1), zdaj pa manjkajočih vrednosti ni več in ima *Leto rojstva* povprečje 1955,14. Vidimo pa, da se kljub velikemu deležu manjkajočih vrednosti po EM algoritmu povprečje ni veliko spremenilo, kar pomeni, da podatki niso manjkali na način, kjer bi vzorec manjkajočih podatkov vključeval informacijo, ki bi lahko vplivala na postopek vstavljanja. V takih primerih vidimo, zakaj je potreben poseben pristop obravnavanja manjkajočih podatkov. Potrošniki so v resnici več zapravili za nakupe, kar nam navadna analiza z neupoštevanjem manjkajočih podatkov ni pokazala. Šele zdaj, z uporabljenim EM algoritmom, vidimo, da je pravi pristop pomemben.

Tabela 4.5: Povprečja EM

Izobrazba	2,3315
Vrednost prvega nakupa	4,0521
Vrednost vseh nakupov	322,3797
Število naročil preko katalogov	2,49
Vrednost naročil preko katalogov	72,30
Leto rojstva	1955,14

4.3.2 Analiza korelacij

V tem poglavju so prikazane korelacije med posameznimi pari spremenljivk za vse tri obravnavane metode. Najprej si poglejmo korelacije za osnovi analize popolnih enot.

V korelacijski matriki, tabeli 4.6, smo pogledali, kolikšno moč povezave imajo posamezni pari spremenljivk. Vidimo, da popolna korelacija med katerimakoli spremenljivkama ne obstaja.

Najbližje popolni korelaciji je povezanost med *Vrednostjo* in *Številom naročil preko katalogov*. Ta par ima vrednost 0,774, kar pomeni, da ti dve spremenljivki naraščata ali padata skupaj. Torej, kadar se veča število naročil preko katalogov, se veča tudi vrednost naročil preko katalogov. Spremenljivka *Izobrazba* negativno korelira z vsemi ostalimi spremenljivkami. To pomeni, da se z večanjem katerekoli spremenljivke manjša stopnja izobrazbe potrošnikov.

Tabela 4.6: Korelacijska matrika Listwise

	Izo-brazba	Vrednost prvega nakupa	Vrednost vseh nakupov	Število naročil preko katalogov	Vrednost naročil preko katalogov	Leto rojstva
Izobrazba	1					
Vrednost prvega nakupa	−0,158	1				
Vrednost vseh nakupov	−0,171	0,026	1			
Število naročil preko katalogov	−0,008	0,005	−0,018	1		
Vrednost naročil preko katalogov	−0,072	0,198	−0,007	0,744	1	
Leto rojstva	−0,019	0,145	0,018	0,073	0,089	1

Sledi korelacijska matrika na osnovi analize podatkov, ki so na voljo (pairwise).

Vsak korelacijski koeficient pri tej analizi je produkt enot s popolnimi podatki za vsak par spremenljivk, ki korelira. Vrednost korelacijskega koeficienta je pri paru *Vrednost naročil preko katalogov* in *Število naročil preko katalogov* enaka 0,554, kar pomeni, da obstaja pozitivna povezanost med omenjenima spremenljivkama. Potrošniki, ki opravijo več naročil preko katalogov, zanje porabijo tudi več denarja. Močna pozitivna korelacija obstaja tudi med *Vrednostjo naročil preko katalogov* in *Vrednostjo prvega nakupa.* Korelirata tudi spremenljivki *Vrednost naročil preko katalogov* in *Leto rojstva.* Vrednost korelacijskega koeficienta je tu enaka 0,162, kar pomeni, da se z večanjem vrednosti prve spremenljivke veča tudi druga. Starejši ljudje porabijo več denarja za nakupe preko kataloga kot mlajši.

Tabela 4.7: Korelacijska matrika Pairwise

	Izobrazba	Vrednost prvega nakupa	Vrednost vseh nakupov	Število naročil preko katalogov	Vrednost naročil preko katalogov	Leto rojstva
Izobrazba	1					
Vrednost prvega nakupa	0,042	1				
Vrednost vseh nakupov	−0,027	−0,018	1			
Število naročil preko katalogov	−0,027	−0,047	0,015	1		
Vrednost naročil preko katalogov	−0,055	0,412	−0,074	0,554	1	
Leto rojstva	0,037	0,178	−0.,006	−0,080	0,162	1

Zdaj si poglejmo še korelacijsko matriko za metodo vstavljanja najverjetnejših vrednosti, EM algoritem.

V tabeli 4.8 vidimo, da pozitivna korelacija še vedno obstaja med *Vrednostjo in Številom naročil preko katalogov* ter med *Vrednostjo naročil preko katalogov* in *Vrednostjo prvega nakupa.* Z naraščanjem *Števila naročil preko katalogov* narašča tudi *Vrednost naročil preko katalogov* in *Vrednost prvega nakupa*. Negativna korelacija pa obstaja med *Izobrazbo* in *Vrednostjo vseh nakupov.* Z večanjem stopnje izobrazbe se manjša vrednost nakupovanja. Bolj izobraženi potrošniki opravljajo manj nakupov. Tudi tu smo predpostavili, da je izobrazba razmernostna spremenljivka, čeprav je v resnici ordinalna.

Tabela 4.8: Korelacijska matrika EM

	Izobrazba	Vrednost prvega nakupa	Vrednost vseh nakupov	Število naročil preko katalogov	Vrednost naročil preko katalogov	Leto rojstva
Izobrazba	1					
Vrednost prvega nakupa	0,043	1				
Vrednost vseh nakupov	−0,026	−0,018	1			
Število naročil preko katalogov	−0,026	−0,047	0,015	1		
Vrednost naročil preko katalogov	−0,053	0,397	−0,068	0,188	1	
Leto rojstva	0,037	0,183	−0,006	−0,079	0,151	1

PRIMERJAVA POVPREČIJ IN KORELACIJ

V tem poglavju je najprej podana tabela z opisnimi statistikami za spremenljivke glede na različne pristope, v nadaljevanju pa sledi interpretacija vrednosti v tabeli.

Tabela 4.9: Opisne statistike za spremenljivke glede na različne pristope

		VSE ENOTE	ENOTE NA VOLJO (AVALIABLE)	LISTWISE	PAIRWISE	EM
Izobrazba	Veljavne	17184	6306	103	6306	7010
	Manjkajoče	4987	704	6907	704	0
Povprečje		2,39	2,3298	2,25	2,32	2,33
Standardni odklon		0,9609	0,9439	0,9773	0,9439	0,9439
Minimum		1	1	1	1	1
Maksimum		6	6	6	6	6
Leto rojstva	Veljavne	18619	6759	103	6759	7010
	Manjkajoče	3552	251	6907	251	0
Povprečje		1957,66	1955,12	1954,26	1955,12	1955,14
Standardni odklon		14,405	13,417	13,748	13,417	13,416
Minimum		1900	1918	1918	1918	1918
Maksimum		2013	1999	1999	1999	1999
Vrednost vseh nakupov	Veljavne	21122	6707	103	6707	7010
	Manjkajoče	1049	303	6907	303	0
Povprečje		322,179	322,5151	313,834	322,5151	322,3797
Standardni odklon		1,6726	277,0213	265,1024	277,0213	277,0222
Minimum		2	2	2	2	2
Maksimum		144805	1188	1188	1188	1188
Vrednost prvega nakupa	Veljavne	22171	7010	103	7010	7010
	Manjkajoče	0	0	6907	0	0
Povprečje		3,5087	4,0521	4,126	4,0521	4,0521
Standardni odklon		2,1194	0,7009	0,5887	0,7009	0,7009
Minimum		1	1	1	1	1
Maksimum		10	10	10	10	10
Število naročil preko kataloga	Veljavne	22171	7010	103	7010	7010
	Manjkajoče	0	0	6907	0	0
Povprečje		0,84	2,49	1,29	2,49	2,49
Standardni odklon		1,834	2,424	0,681	2,424	2,424
Minimum		0	1	1	1	1
Maksimum		32	32	32	32	32
Vrednost naročil preko kataloga	Veljavne	14770	132	103	132	7010
	Manjkajoče	7401	6878	6907	6878	0
Povprečje		0,72	70,15	65,45	70,15	72,30
Standardni odklon		12,599	62,969	54,902	62,969	63,1713
Minimum		0	0	0	0	0
Maksimum		1078	339	339	339	339

V prvem stolpcu, z naslovom »VSE ENOTE«, so podane opisne statistike za izbrane spremenljivke na podlagi vseh 22171 enot, brez izločanja in odstranjevanja *outlierjev*. Vrednosti v tem stolpcu se ujemajo z vrednostmi v tabeli 3.1. V vseh ostalih stolpcih pa so prikazane opisne statistike za izbrane spremenljivke po odstranitvi enot, ki niso opravile nakupa preko kataloga.

Vidimo, da se vrednosti spremenljivk, glede na različne pristope, spreminjajo. Spremenljivki *Vrednost prvega nakupa* in *Število naročil preko katalogov* nimata manjkajočih vrednosti, zato deskriptivne statistike ostajajo enake, ne glede na metodo obravnave. Do razlik pride le v prvem stolpcu, saj smo tam računali povprečje in standardni odklon na podlagi vseh enot. Vrednosti se razlikujejo tudi v »LISTWISE« stolpcu, saj smo tam računali povprečje in standardni odklon na podlagi popolnih enot.

Drugače pa je pri tistih spremenljivkah, ki imajo manjkajoče vrednosti. Pri *Vrednosti vseh nakupov* je bilo povprečje na podlagi vseh enot 322,179, pri razpoložljivih podatkih (po odstranitvi enot, ki niso opravile nakupa preko kataloga) 322,515, z metodo Listwise se je povprečje zmanjšalo na 313,834 €, po vstavljenem EM algoritmu pa je povprečje spet naraslo na 322,379. Pri tej spremenljivki smo imeli 303 manjkajoče vrednosti, ki pa niso statistično značilno vplivale na interpretacijo analize. Tudi če bi delali analizo ob neupoštevanju manjkajočih vrednosti, le-ta ne bi bila popolnoma napačna.

Do sprememb pride tudi pri spremenljivki *Vrednost naročil preko katalogov.* Ta ima samo 132 veljavnih vrednosti, 6878 je manjkajočih vrednosti, ostalih 15151 pa smo že na začetku izločili iz analize, saj toliko potrošnikov ni opravilo nakupa preko kataloga, in posledično zanj ni porabilo denarja. Pri analizi na osnovi 132 razpoložljivih vrednosti ima ta spremenljivka povprečje enako 70,15, pri listwise metodi pa 65,45. Po uporabljenem EM algoritmu se povprečje zopet dvigne, na 72,30 €. Uporaba EM algoritma pri tej spremenljivki ni prinesla velikih sprememb, razen v primerjavi s prvim stolpcom, kjer vse manjkajoče vrednosti upoštevamo kot *Null* manjkajoče vrednosti, kar pa je napačna predpostavka.

Tudi pri kontrolnih spremenljivkah opazimo razlike v povprečjih. Čeprav je izobrazba ordinalna spremenljivka, ki pa se nekoliko približuje razmernostni, smo ob tej predpostavki tudi zanjo izračunali opisne statistike, ki pa nam ne dajo jasne interpretacije. Pri *Izobrazbi,* ki je imela 704 manjkajoče vrednosti, se povprečje z metodami za obravnavo manjkajočih podatkov spremeni, a le malo. Prav tako ne opazimo velikih razlik pri *Letu rojstva* potrošnikov.

Pri Listwise metodi je povprečno leto rojstva 1954,26, pri metodi Pairwise se dvigne za 1 leto, po EM algoritmu ostaja 1955,14. Tu se povprečja zelo malo spremenijo, starostne skupine potrošnikov ostajajo iste, tako da manjkajoči podatki pri *Letu rojstva* ne vplivajo na analize raziskav v Podjetju X. Povprečje na podlagi vseh enot je namreč enako 1957,66, kar se sicer razlikuje od ostalih povprečij za več kot eno leto, a starostne skupine potrošnikov še vedno ostajajo iste.

Toda opozoriti moramo na dejstvo, da smo delali analizo manjkajočih vrednosti samo za enote, ki niso imele manjkajočih vrednosti pri spremenljivki *Število naročil preko katalogov.* Ta spremenljivka je v naši raziskavi kritična, saj je od nje odvisna celotna analiza in interpretacija rezultatov.

4.3.3 Analiza porazdelitve

Zdaj pa poglejmo še primerjave porazdelitve posameznih spremenljivk na podlagi razpoložljivih vrednosti po odstranitvi enot, ki niso opravile nakupa preko kataloga (v grafu so te enote označene z *»avaliable«*), in po izvedenih korakih za analizo z manjkajočimi vrednostmi (v grafu so te enote označene z *»EM«)*. Pri vsaki spremenljivki imamo 7010 enot, saj smo odstranili tiste potrošnike, ki niso opravili nakupa preko kataloške prodaje in tako zmanjšali število enot za 15161 (22171 na 7010).

Seveda je EM algoritem pri izračunih simulantno upošteval vse spremenljivke, vendar za ilustrativen pregled navajamo pregled učinka na porazdelitev posameznih spremenljivk.

Porazdelitev za spremenljivke, ki niso imele manjkajočih vrednosti *(Vrednost prvega nakupa, Število naročil preko kataloga),* ostaja nespremenjena, medtem ko je pri spremenljivkah z manjkajočimi vrednostmi prišlo do sprememb.

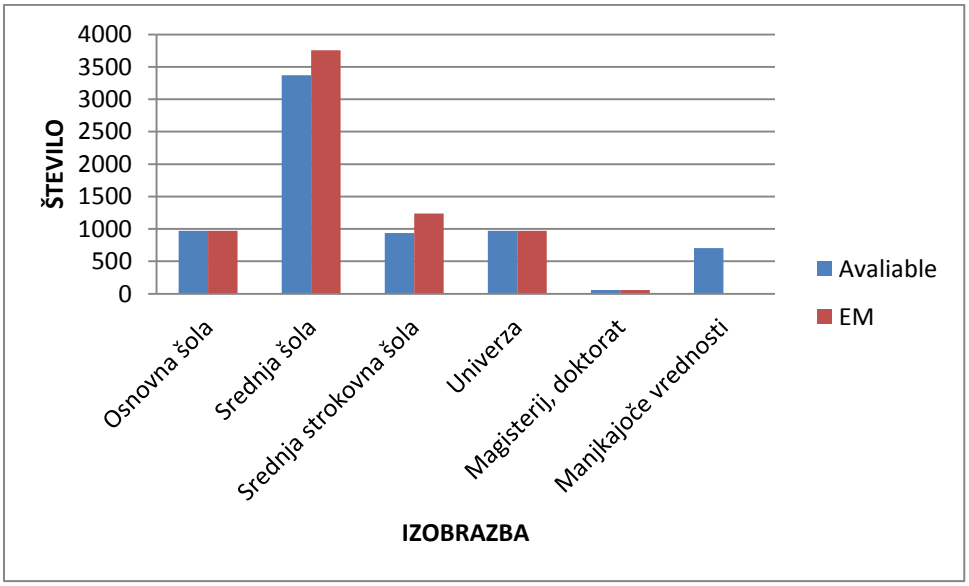
Najprej poglejmo manjkajoče vrednosti za posamezne spremenljivke, ob upoštevanju odstranitve enot, ki niso opravile nakupa preko kataloga. Iz tabele 4.9 je razvidno, da sta imeli spremenljivki *Spol* in *Regija* zelo malo manjkajočih vrednosti, zato teh dveh spremenljivk ne bomo prikazali s frekvenčnimi porazdelitvami. Za omenjeni spremenljivki (spol – nominalna in regija – ordinalna) EM algoritem tudi ne zna računati manjkajočih vrednosti.

Za nominalne spremenljivke je EM algoritem uporabil multiplo logistično regresijo, za ordinalne prilagojeno regresijo, za razmernostne spremenljivke pa normalno porazdelitev. Čeprav je *Izobrazba* ordinalna spremenljivka, smo tudi zanjo naredili analizo vstavljanja najverjetnejših vrednosti (EM algoritem), saj se le-ta približuje intervalni porazdelitvi.

IZOBRAZBA

Prva kontrolna spremenljivka, ki jo prikazujemo v analizi porazdelitve, je *Izobrazba*. Pri njej je porazdelitev drugačna, kot je bila na začetku. Razlika je v tem, da je EM algoritem vsem manjkajočim vrednostim dodelil neke številke, ki smo jih razdelili v skupine. Če je dodelil EM algoritem neki enoti, ki je imela manjkajočo vrednost pri izobrazbi vrednost 2,37, smo 37 % vrednosti razvrstili v 3. stopnjo izobrazbe (srednja strokovna šola), 63 % vrednosti pa v 2. stopnjo izobrazbe (srednja šola). Glede na spodnji graf 4.3 lahko rečemo, da manjkajoče vrednosti pri izobrazbi niso odločilno vplivale na rezultate analize v Podjetju X, saj ostaja v kategoriji *srednja šola* še vedno največ potrošnikov.

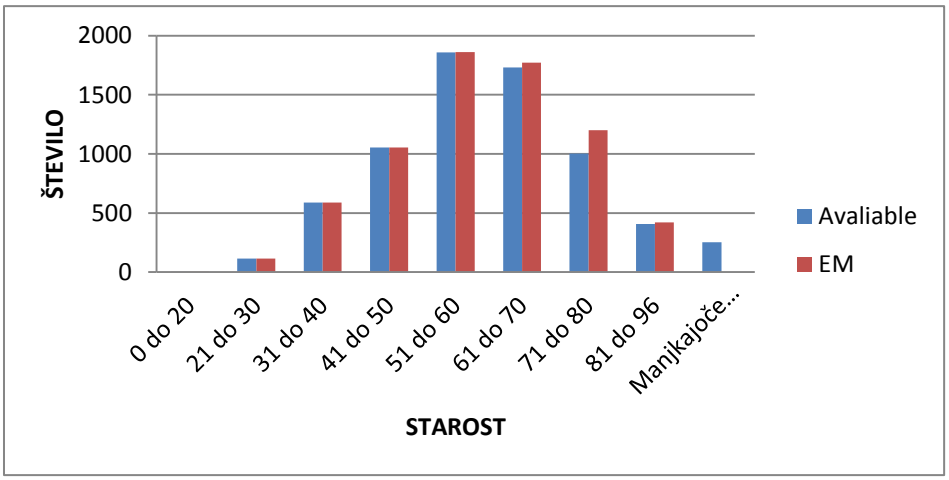
Graf 4.3: Primerjava_histogram izobrazba za razpoložljive podatke in za celotno popolnjeno matriko z EM algoritmom



STAROST

Druga kontrolna spremenljivka, katere porazdelitev prikazujemo, je starost. Pri S*tarosti* smo rekodirali spremenljivko *Leto rojstva* tako, da smo dobili starostne razrede. Iz grafa 4.4 je razvidno, da je EM algoritem manjkajoče vrednosti pri starosti dal v starostno skupino »*od 61 do 70 let*« in od »*71 do 80 let*«. Za analizo to sicer ni ključnega pomena, saj govorimo o starostnih skupinah, ki zapravijo največ denarja. Toda pomemben podatek je, da so tisti potrošniki, ki niso izdali svoje starosti, v starostni skupini od *61 do 80 let*. Očitno starejši potrošniki prikrivajo informacije o svojih letih pogosteje kot mladi.

Graf 4.4: Primerjava_histogram starost za razpoložljive podatke in za celotno popolnjeno matriko z EM algoritmom

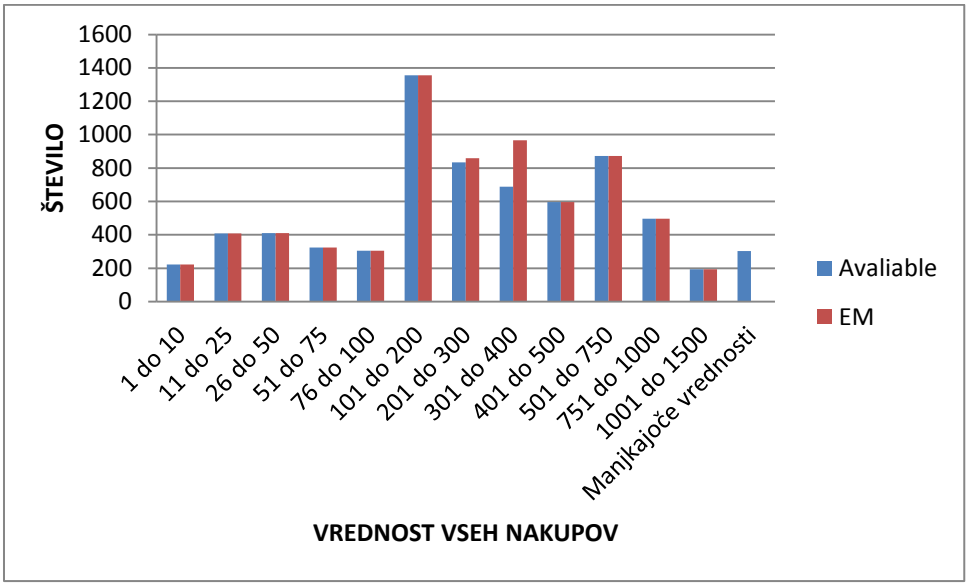


Zdaj pa poglejmo še porazdelitve za ciljne spremenljivke. Tu prikazujemo samo tiste spremenljivke, ki so imele manjkajoče vrednosti, saj pri drugih ni prišlo do sprememb.

VREDNOST VSEH NAKUPOV

Prva spremenljivka, ki jo prikazujemo, je *Vrednost vseh nakupov.* Do vrednosti 200 € za vse nakupe se porazdelitveni graf ujema v obeh primerih (avaliable in EM), enote se prekrivajo. Iz grafa 4.5 je razvidno, da je EM algoritem manjkajoče vrednosti na tej spremenljivki dal v razrede od 201 do 400 €. Pri razredih od 201 do 400 € se povprečje spremenljivke, po uporabljeni EM metodi, dvigne. Enote z manjkajočimi vrednostmi so trošile nadpovprečno, največ v kategoriji od 301 do 400 €.

Graf 4.5: Primerjava_histogram vrednost vseh nakupov za razpoložljive podatke in za celotno popolnjeno matriko z EM algoritmom



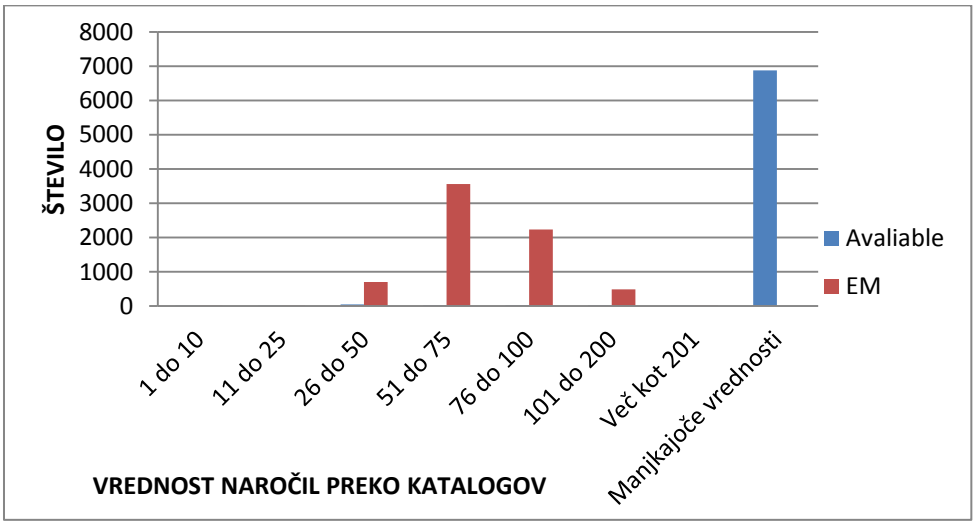
VREDNOST NAROČIL PREKO KATALOGOV

Ta spremenljivka je imela veliko manjkajočih vrednosti, kar 6878. V tabeli 4.10 je zato prikazana frekvenčna porazdelitev za *Vrednost naročil preko katalogov,* iz katere je razvidno, kako je EM algoritem vstavljal vrednosti. Vidimo, da je največji delež manjkajočih vrednosti, po izvedenem EM algoritmu, v kategoriji *»od 51 do 75 €«*. Zelo se je povečala tudi kategorija »od 76 do 100 €« (iz 0,3 % na 31,9 vseh vrednosti). Manjše povečanje pa je bilo v kategoriji »od 100 do 200 €«. Tu gre za povečanje iz 18 na 490 enot. Ostale kategorije spremenljivke *Vrednost naročil preko katalogov* pa so imele tako majhno povečanje enot po izvedenem EM algoritmu, da na grafu 4.6 ni opazno.

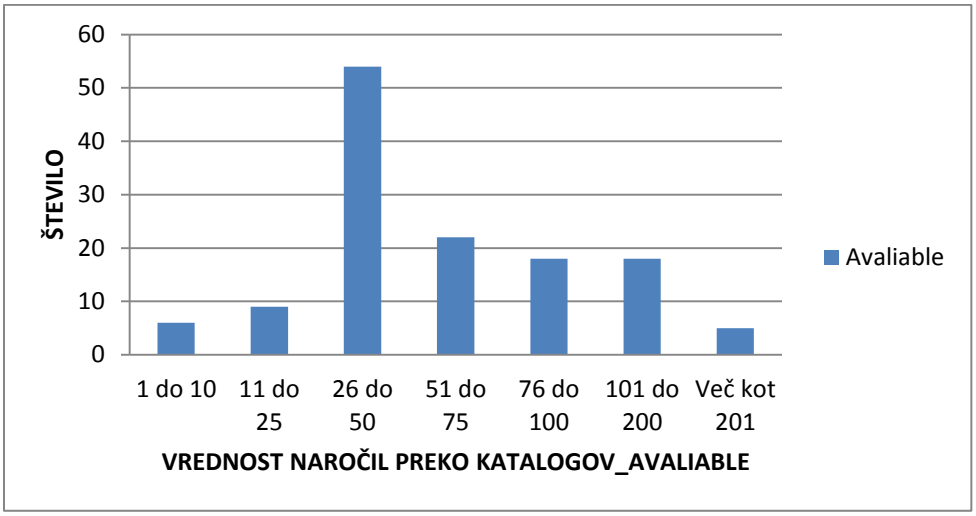
Na grafu 4.7 smo ločeno prikazali porazdelitev za to spremenljivko, za *Avaliable* enote. Iz grafa je razvidno, da je najpogostejša vrednost nakupa kategorija *»26 do 50€«,* kar smo pokazali že v tabeli 3.17, kjer je bil modus 33, kar spada v omenjeno kategorijo. Potrošniki, ki so za naročila preko katalogov zapravili 33€, so imeli večinoma osnovnošolsko ali univerzitetno izobrazbo, kar je relativno nizka oziroma relativno visoka stopnja izobrazbe. Po uporabi EM algoritma, pa se modus spremeni. EM algoritem je namreč velikemu deležu manjkajočih vrednosti dodelil vrednosti od 51 do 75€.

Potrošniki, katerim je EM algoritem vstavil takšne vrednosti, so najpogosteje dosegali srednješolsko ali srednješolsko strokovno izobrazbo. Em algoritem je torej pri vstavljanju manjkajočih vrednosti upošteval stopnjo izobrazbe potrošnikov. Kot je bilo že prikazano v analizi glede na izobrazbo, pa so tisti potrošniki z srednješolsko ali srednješolsko strokovno izobrazbo, v povprečju za naročila preko katalogov porabili več denarja, kot tisti z doseženo osnovnošolsko ali univerzitetno izobrazbo. EM algoritem je pri vstavljanju manjkajočih vrednosti upošteval tudi starost potrošnikov. Starejšim osebam je pripisal večje vrednosti, kot mlajšim. Vsi potrošniki, ki so po uporabi EM algoritma za naročila preko katalogov porabili *»51 do 75€«* so namreč v starostnih skupinah nad 51 let. To pa so potrošniki, ki porabijo za nakupe največ denarja, kar je prikazano tudi v tabeli 3.25. Modus spremenljivke *Vrednost naročil preko katalogov* se torej, po uporabi EM algoritma, spremeni zato, ker je EM agoritem pri vstavljanju vrednosti upošteval izobrazbo in starost potrošnikov. Potrošnikom, ki so dosegali povprečno izobrazbo (srednješolska ali srednješolska strokovna) in višjo starost je pripisal večje vrednosti. Zato se spremeni tudi modus spremenljivke. Tu moramo še opozoriti na dejstvo, da na tej točki analize (EM algoritma) nismo imeli manjkajočih vrednosti na nobeni spremenljivki.

Graf 4.6: Primerjava_histogram Vrednost naročil preko katalogov za razpoložljive podatke in za celotno popolnjeno matriko z EM algoritmom



Graf 4.7: Histogram za Vrednost naročil preko katalogov za razpoložljive podatke brez manjkajočih vrednosti



V tabeli 4.10 je porazdelitev frekvenc med *Avaliable* in *EM* precej drugačna, zato se, glede na različne pristope, spremeni tudi povprečje spremenljivke.

Tabela 4.10: Primerjava frekvenčne porazdelite za vrednost naročil preko katalogov

		Available	N %	EM	N %
Vrednost naročil preko katalogov	1 do 10	6	0,1	6	0,1
	11 do 25	9	0,1	10	0,1
	26 do 50	54	0,8	705	10,1
	51 do 75	22	0,3	3557	50,7
	76 do 100	18	0,3	2235	31,9
	101 do 200	18	0,3	490	7,0
	Več kot 201	5	0,1	7	0,1
	Manjkajoče vrednosti	6878	98,1	0	0,0
	Skupaj	7010	100,0	7010	100,0

5 SKLEP

Manjkajoči podatki so problem, s katerim se pogosto srečamo v družboslovnem raziskovanju. Če je bilo raziskovalno vprašanje na začetku naloge, ali *podatki manjkajo povsem naključno in analize lahko izvedemo z neupoštevanjem manjkajočih podatkov,* lahko zdaj trdim, da je za korektne analize vsekakor treba preveriti, kakšnega izvora so manjkajoči podatki, in v skladu s tem tudi pristopati do njih. Čeprav je analiza manjkajočih vrednosti kompleksna in zapletena, je vseeno nujna, če želimo predstaviti pravilne rezultate. Morda se bo po izvršeni analizi manjkajočih vrednosti pokazalo, da le-ta sploh ni bila potrebna, morda pa je ključnega pomena za doseganje pravilnih rezultatov.

V raziskavi smo videli veliko manjkajočih vrednosti, ki so vplivale na potek raziskovanja. Tu velja še enkrat poudariti, da smo delali analizo za 8 izbranih spremenljivk. Zato se sklep nanaša na teh 8 obravnavanih spremenljivk in ne na celotno bazo podatkov, ki je obsegala več kot 70 spremenljivk. Toda spremenljivke, ki smo jih izbrali za obdelavo v diplomskem delu, se porazdeljujejo podobno kot ostale, zato lahko trdim, da sklep na podlagi obravnavanih spremenljivk velja za celotno bazo.

Pri kontrolni spremenljivki *Leto rojstva* se povprečje ni bistveno spremenilo. Tehnike obravnavanja manjkajočih podatkov so zvišale ali pa znižale povprečno starost, a ne za več kot eno leto. To pa ne spremeni niti starostne skupine potrošnikov. Za to spremenljivko lahko rečemo, da manjkajoče vrednosti ne vplivajo na končne rezultate. Drugače pa je bilo pri tistih spremenljivkah, ki so imele veliko manjkajočih vrednosti. Tehnik za obravnavno manjkajočih podatkov je veliko in pomembno je, da najdemo ustrezno. Po Little`s MCAR testu se je v našem primeru pokazalo, da je ustrezna metoda uporaba EM algoritma. Ker pa je bila to raziskovalna naloga, smo vseeno prikazali tudi analizo podatkov, ki so na voljo, in celotno analizo podatkov *(pairwise in listwise deletion).* Pri raziskovalni nalogi je namreč pomembno, da predstavimo različne vidike in pristope reševanja problema.

Kot najpomembnejšo ugotovitev diplomskega dela moramo vsekakor izpostaviti dejstvo, da manjkajoči podatki niso imeli večjega učinka na analize raziskave. Neupoštevanje manjkajočih vrednosti v našem primeru vsekakor ni pravi pristop k analizi podatkov, skoraj izključno zaradi ene spremenljivke (*Vrednost naročil preko katalogov).*

Pri tej spremenljivki pa je bilo kjučnega pomena, kako smo se lotili analize. Če bi delali analizo ob napačni predpostavki, da bi vse manjkajoče vrednosti omenjene spremenljivke, zamenjali za *Null* manjkajoče vrednosti, bi bila razlika v povprečjih zelo velika in statistična analiza napačna. Ker pa smo predhodno odstranili sistemske manjkajoče vrednosti, so nam dali rezultati analize podobne interpretacije statističnih obdelav, kot uporaba EM algoritma. Razlika v povprečjih je vseeno opazna, povprečje spremenljivke *Vrednost naročil preko katalogov,* se namreč razlikuje za 2€ (med *Avaliable enotami* – brez enot, ki niso opravile nakupa preko kataloga je povprečje enako 70,15€, ob uporabi EM algoritma pa 72,3€). Pri tej spremenljivki je torej pomemben pravilen pristop in razumevanje ter razlikovanje različnih tipov manjkajočih podatkov. Poudariti velja, da je imela ta spremenljivka 7401 manjkajočih vrednosti. Od 14770 veljavnih vrednosti pa je bilo 14637 vrednosti 0, saj odgovarjajoče enote niso opravile nakupa, kar pomeni, da so to sistemske manjkajoče vrednosti in zanje ne moremo vstavljati nadomestnih vrednosti. Posledično imamo pri tej spremenljivki le 133 veljavnih, neničelnih vrednosti.

Ker manjkajoči podatki niso manjkali naključno, njihovo neupoštevanje vsekakor ni pravi pristop k analizi. Manjkajoči podatki torej vplivajo na analizo raziskave v Podjetju X v večji ali manjši meri.

Ugotovili smo, da kljub znatnemu delu manjkajočih podatkov uporaba EM algoritma, ki predpostavlja MAR, ni prinesla večjih sprememb. To pa zgolj zato, ker smo pred analizo manjkajočih vrednosti pregledali vse tipe manjkajočih podatkov in izločili sistemske manjkajoče vrednsoti. Če tega ne bi naredili, bi bili rezultati popolnoma drugačni. V celoti gledano je torej neupoštevanje manjkajočih vrednosti ustrezno, ampak izključno ob pravilnih predpostavkah manjkajočih vrednosti.

Seveda pa se pojavlja tudi problem izračuna variance, česar se v nalogi nismo lotevali. Običajni izračuni na osnovi popolnjene podatkovne matrike podcenjujejo varianco, saj predpostavljajo, da imamo večji vzorec kot v resnici. Navedeno pa presega zastavljeni cilj diplomskega dela, ki bi zahteval bodisi obravnavo izračuna kovarianc EM algoritma bodisi uporabo večkratnega imputiranja.

Vsekakor smo pokazali, da prinaša neupoštevanje manjkajočih vrednosti velike nevarnosti in napačne interpretacije v statističnih analizah.

Seveda se v delu tudi nismo posvetili veljavnosti predpostavke MAR, ampak smo predvidevali, da podatki manjkajo na tak način. Problematizacija predpostavke MAR pa prinaša nadaljnjo kompleksno analizo v obravnavi manjkajočih podatkov.

6 LITERATURA

1. Blejec, Andrej. 2008. *Multivariatne metode in manjkajoči podatki.* Dostopno prek: http://ablejec.nib.si/R/mva07.pdf (7. marec 2001).
2. van Buuren, Stef. 2012. *Flexible imputation on missing data.* Boca Raton: Champan & Hall/CRC.
3. Drukker, David M. 2011. *Missing Data Methods*. *Advances in Econometrics.* Bingley, U. K.: Emerald.
4. Hair, J. F. 2006. *Multivariate data analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.
5. IBM Corporation. 1989. *IBM SPSS Missing Values 21*. Dostopno prek: http://www.sussex.ac.uk/its/pdfs/SPSS_Missing_Values_21.pdf (1. avgust 2014).
6. McKnight, P. E., K. M., McKnight, S., Sidani, A.J., Figuerdo. 2007. *Missing Data: A Gentle Introduction.* Guildford Press, New York.
7. Lisac in Lisac. 2014. *Telemarketing.* Dostopno prek: http://www.lisac-lisac.si/sl/Klicni+center/Telemarketing (25. avgust 2014).
8. Little, J. A. Roderick in Donald B. Rubin. 2002. *Statistical Analysis with Missing Data.* New York: Wiley.
9. Lozar Manfreda, Katja, Zenel Batagelj in Vasja Vehovar. 2002. Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer Mediated Communication* 7(3). Dostopno prek http://jcmc.indiana.edu/vol7/ issue3/vehovar.html (10. januar 2008).
10. Peugh, James in Craig K. Enders. 2004. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research.* 74 (4): 525–556.
11. Rubin, Donald B. 1976: Inference and Missing Data. *Biometrika* 63 (3): 581–592.
12. Sawilowskiy, Shlomo S. 2007. Real Data Analysis. *Quantitative Methods in Education and the Behavioral Sciences.* Charlotte, N.C.: Information Age Pub.
13. Schafer L. Joseph in John W. Graham. 2002. Missing data: Our View of the State of The Art. *Psychological Methods* 7 (2): 147–177.

14. Sinharay, S., H.S. Stern, David Russel. 2001. The use of multiple imputation for the analysis of missing data. *Psychological Methods* 6. Dostopno prek: http://eds.b.ebscohost.com.nukweb.nuk.uni-lj.si/eds/pdfviewer/pdfviewer?sid=f6779f9f-333e-4c54-9ad8-e9b1d7657218%40sessionmgr115&vid=5&hid=115 (3. september 2014).
15. Starkweather, Jon. 2014. *Multiple imputation.* Dostopno prek: http://www.unt.edu/rss/class/Jon/SPSS_SC/Module6/SPSS_M6_2.htm (4. september 2014).
16. William R. Sterner: What Is Missing in Counceling Research? Reporting Missing Data. 2011. *Journal of Counseling & Developmen* 89. Dostopno prek: http://content2.learntoday.info/ben/NTR697_2014/media/Week%203/Assignments/Readings/article%20-%20handling%20reporting%20missing%20data.pdf (1. september 2014).
17. Tabachnick, B. G. in L. S. Fidell. 2007. *Using multivariate statistics.* Boston: Pearson/Allyn & Bacon.
18. Vehovar, Vasja. 2007. *Nepopolni podatki v anketah.* Ljubljana: interno gradivo.
19. Wagner A., Kamuakura, Michel Wedel. 2000. Factor Analysis and Missing Data. *Journal of Marketing Research:* November 2000, 37 (4): 490–498.

PRILOGE

Priloga A: Sintaksa analize manjkajočih vrednosti v SPSS

```
DATASET ACTIVATE DataSet2.
DATASET DECLARE  EMLOGARITEM.
MVA VARIABLES=total_nof_orders_notaxeur_w_rec nof_valid_order_on_PRINT
    total_value_of_valid_orders_on_PRINT starost education_recR avg_BV_recR gender_recR
    region_recR
  /MAXCAT=25
  /CATEGORICAL=gender_recR region_recR
  /LISTWISE
  /PAIRWISE
  /EM(TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25 OUTFILE=EMLOGARITEM).
```