Univerza *v Ljubljani*

# Poročilo

o obdelavi podatkov iz spletne trgovine v jeziku R, za predmet Programiranje 2

Avtor: Gašper Šenk
Šenčur, 21.8.2012

Mentor: prof. Vladimir Batagelj

# 1 Uvod

Za zaključni projekt pri predmetu Programiranje 2 sem zbral zajetno količino podatkov z izbrane spletne trgovine. Po zbiranju podatkov je sledil naslednji korak, obdelava v jeziku R. V poročilu opisujem postopek obdelave  in tudi grafično predstavitev zbranih podatkov.

# 2 Pridobitev podatkov

Za vir informacij sem si izbral spletno trgovino www.ToysRUs.com, kjer sem zbral

- Spletni naslov artikla
- Ime artikla
- Redno ceno
- Spletno ceno(cena s popustom)
- Popust v številki
- Popust v procentih
- Oceno
- Število komentarjev

o približno 20 000 različnih izdelkih s programom napisanem v jeziku Python. Te podatke sem shranil v datoteko Podatki.csv, kjer je bil v vsaki vrstici en izdelek, podatki o njem so pa ločeni s podpičji.

# 3 Obdelava podatkov

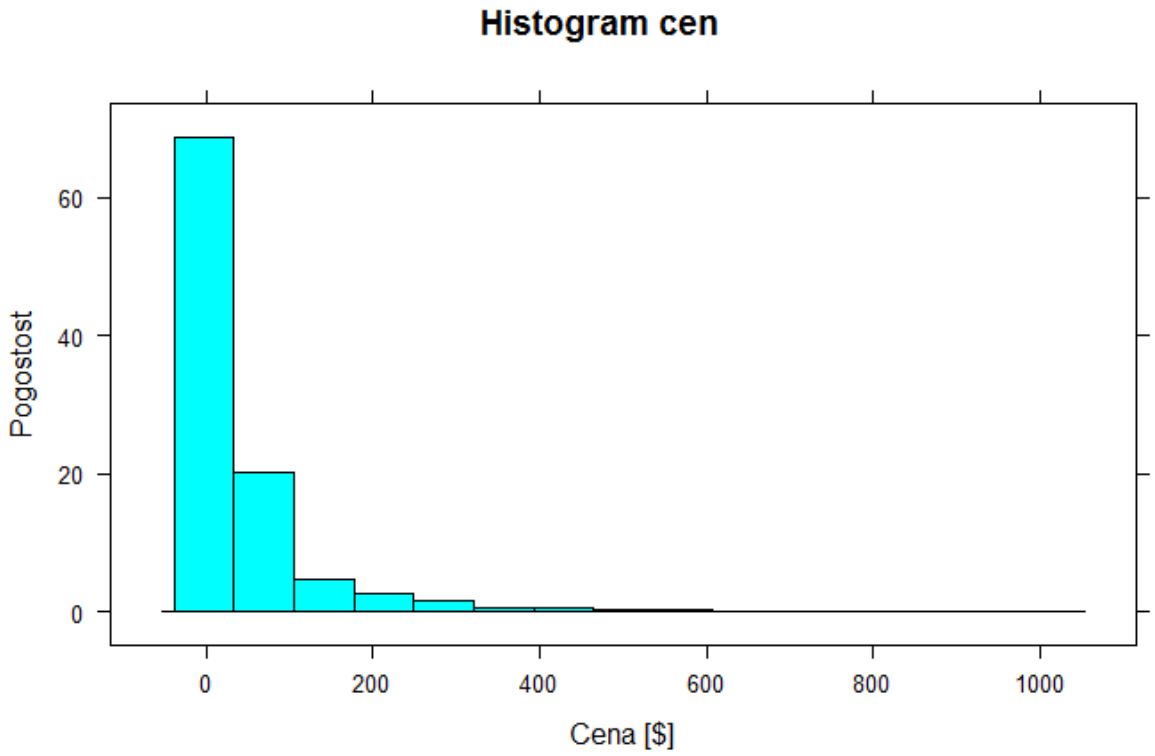
Podatke sem uvozil v spremenljivko *dat*. Za nadaljno obdelavo sem posamezne stolpce shranil v svoje spremenljivke, kjer  je bilo potrebno odstranit nekatere artikle, ki na spletni strani niso imeli na enak način določene *cene* in je moj program ni izluščil. Pri statistiki *ocen* in *komentarjev*, sem uporabil samo artikle pri katerih sta bili količini definirani.

Ker želim pri obdelavi s histrogrami *y-os* v procentih najprej vpeljem:

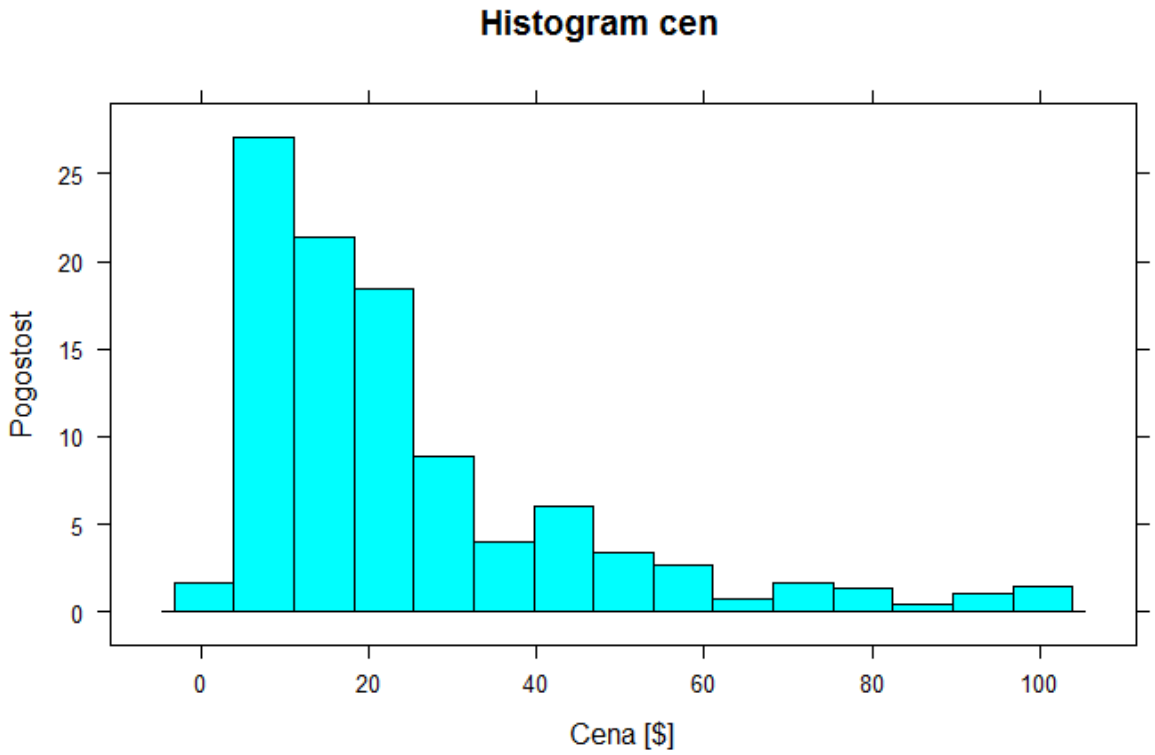
- *library(lattice)* in uporavljam funkcijo *histogram*, namesto *hist*.

## 3.1 Za začetek pogledmo porazdelitev cen artiklov

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.79 | 10.99 | 19.99 | 51.41 | 51.41 | 5800.00 |



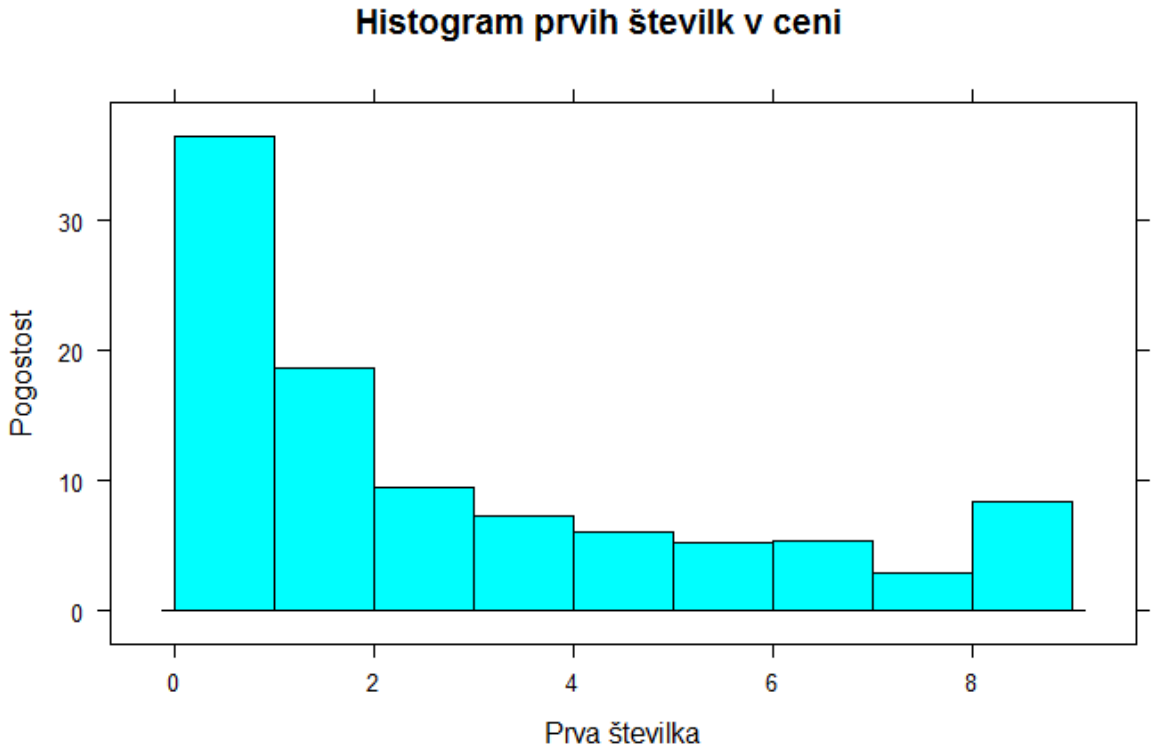

kot vidimo je velika večina cen v predelu nizkih cen, zato bom dodal tudi graf, kjer je najvišja cena 100$.



## 3.2 Zanima nas pokoravanje cen Benfordovemu zakonu

Benfordov zakon je zanimiv s finančnega stališča, saj z njim odkrivajo finančne malverzacije, poleg tega pa lahko določimo kako se nek set podatkov prilagaja naravnim številom v naravi. Porazdelitev bazira na enakomerni porazdelitvi kjer je x os v logaritemski skali. Saj s tem pokažemo relativna povečanja vrednosti. Recimo podvojitev z 1 na 2, 4, 8.

| Min. | 1st Qu. | Median | Mean | Mean 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1.000 | 1.000 | 2.000 | 3.304 | 5.000 | 9.000 |

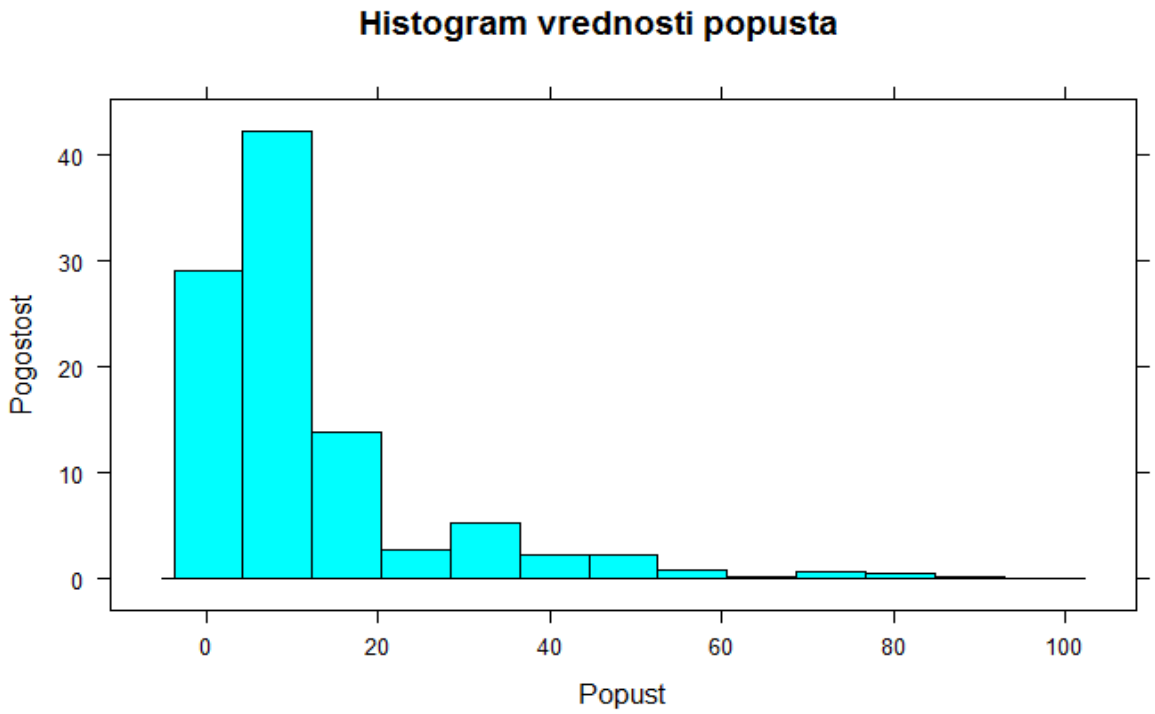



Kot vidimo se naš graf kar lepo prilega teoretičnim vrednostim, razen številke 9, ki je verjetno ustvarjena umetko, saj so se tako trgovci znebili več mestnim številkam in navidezno ustvarili manjšo vrednost, čeprav je razlika majhna.

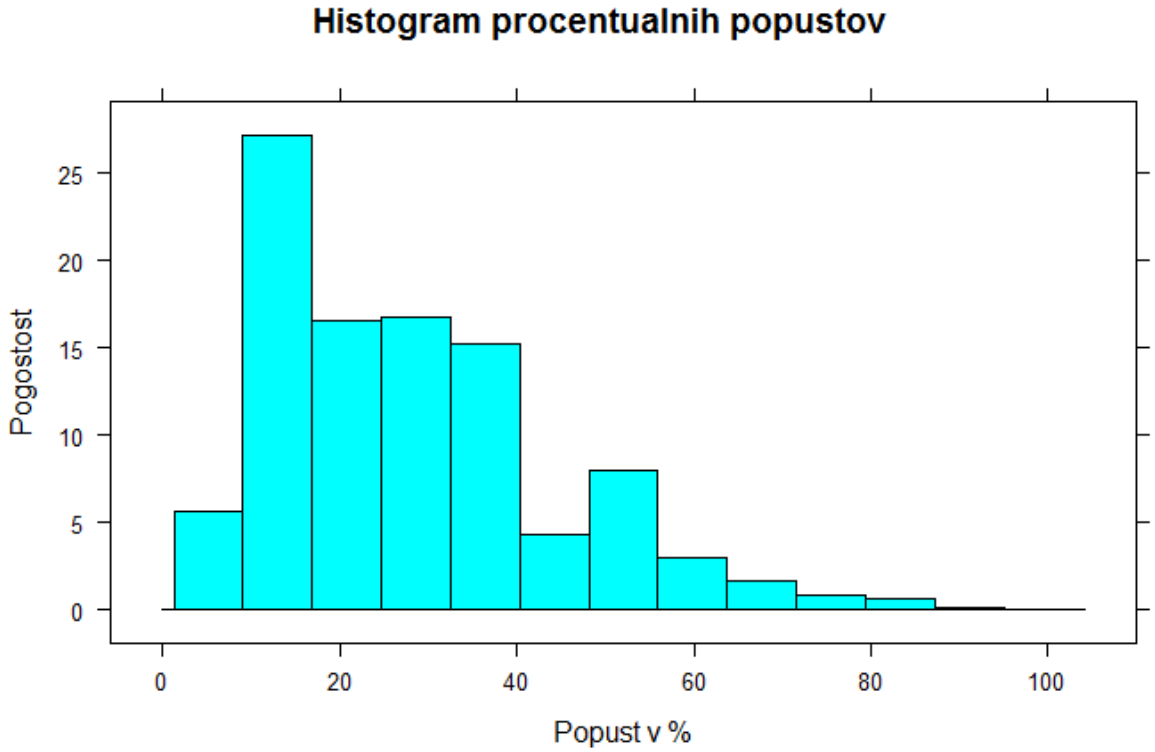
## 3.3 Popusti

| Min. | 1st Qu. | Median | Mean | Mean 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.100 | 3.992 | 8.000 | 15.800 | 15.010 | 3160.000 |

Predmet z največjim popustom je znižan z 3190 na 29.99, kar je verjetno anomalija, saj so pozabili decimalno piko v originalni ceni. Drugo največje znižanje je z 3799.99 na 2999.99, ker znese 800$

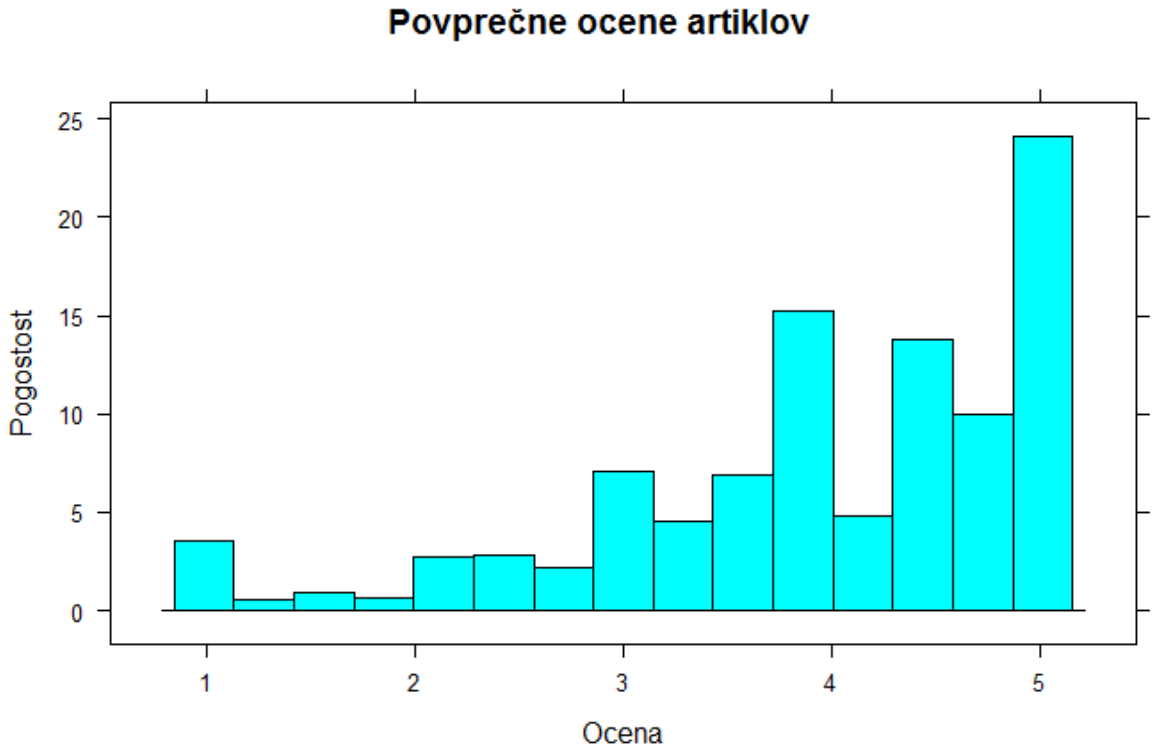



Ker s tem nimamo predstave o dejanskem popustu, saj so cene artiklov zelo različne dodajam



## 3.4 Porazdelitev ocen artiklov

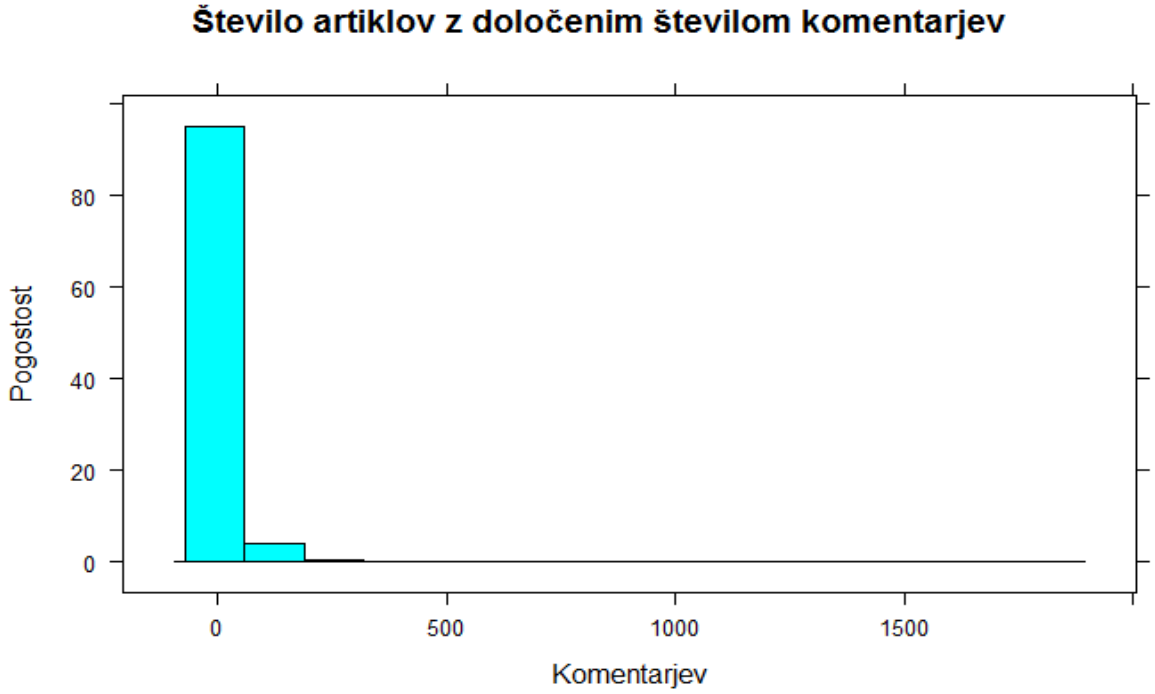
| Min. | 1st Qu. | Median | Mean | Mean 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1.000 | 3.400 | 4.200 | 3.952 | 4.800 | 5.000 |



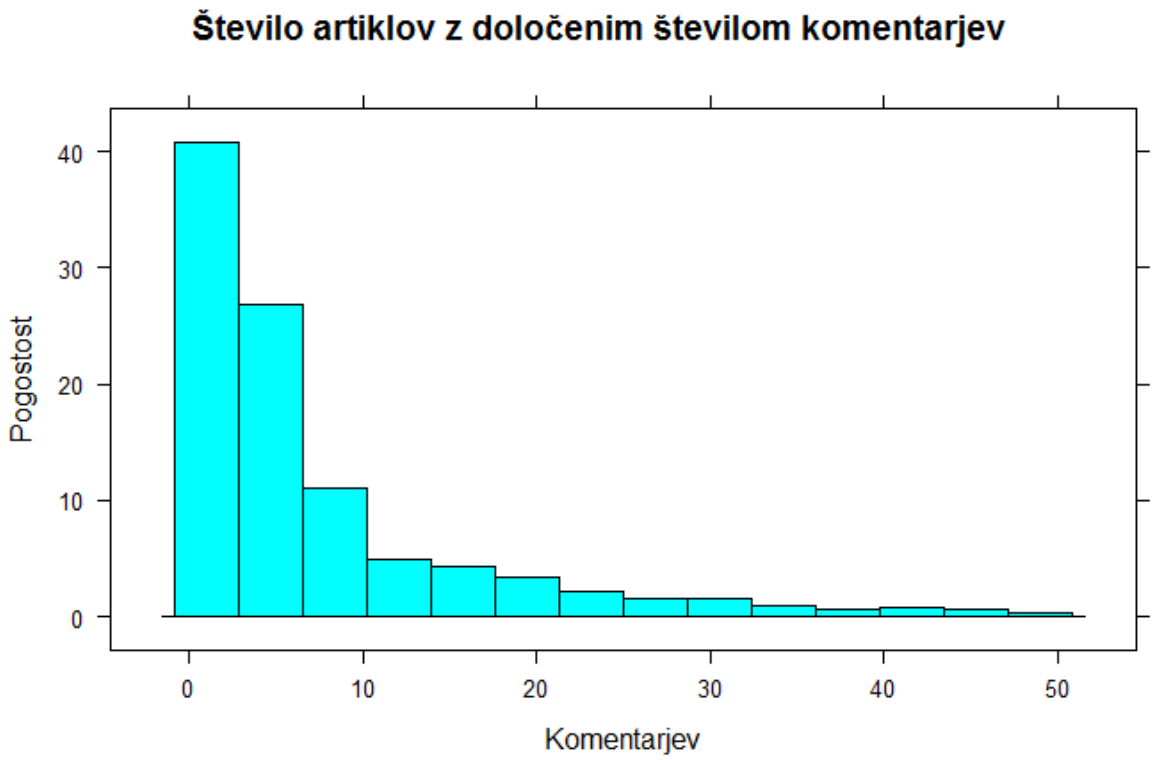

## 3.5 Število komentarjev posameznega artikla

Večina artiklov je brez komentarjev, zato le te odstranim pri sledeči analizi.

| Min. | 1st Qu. | Median | Mean | Mean 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1.000 | 2.000 | 4.000 | 14.74 | 11.00 | 1799.00 |




Ker je spet večina podatkov  v nižjih vrstah, se osredotočimo na predmete z manj kot 50 komentarji.




## 3.6 Zanimiva je še ocena izdelka v odvisnosti od cene