SEMINARSKI RAD IZ KOLEGIJA

**OSNOVE VJEROJATNOSTI I STATISTIKE**

**Izradili:**

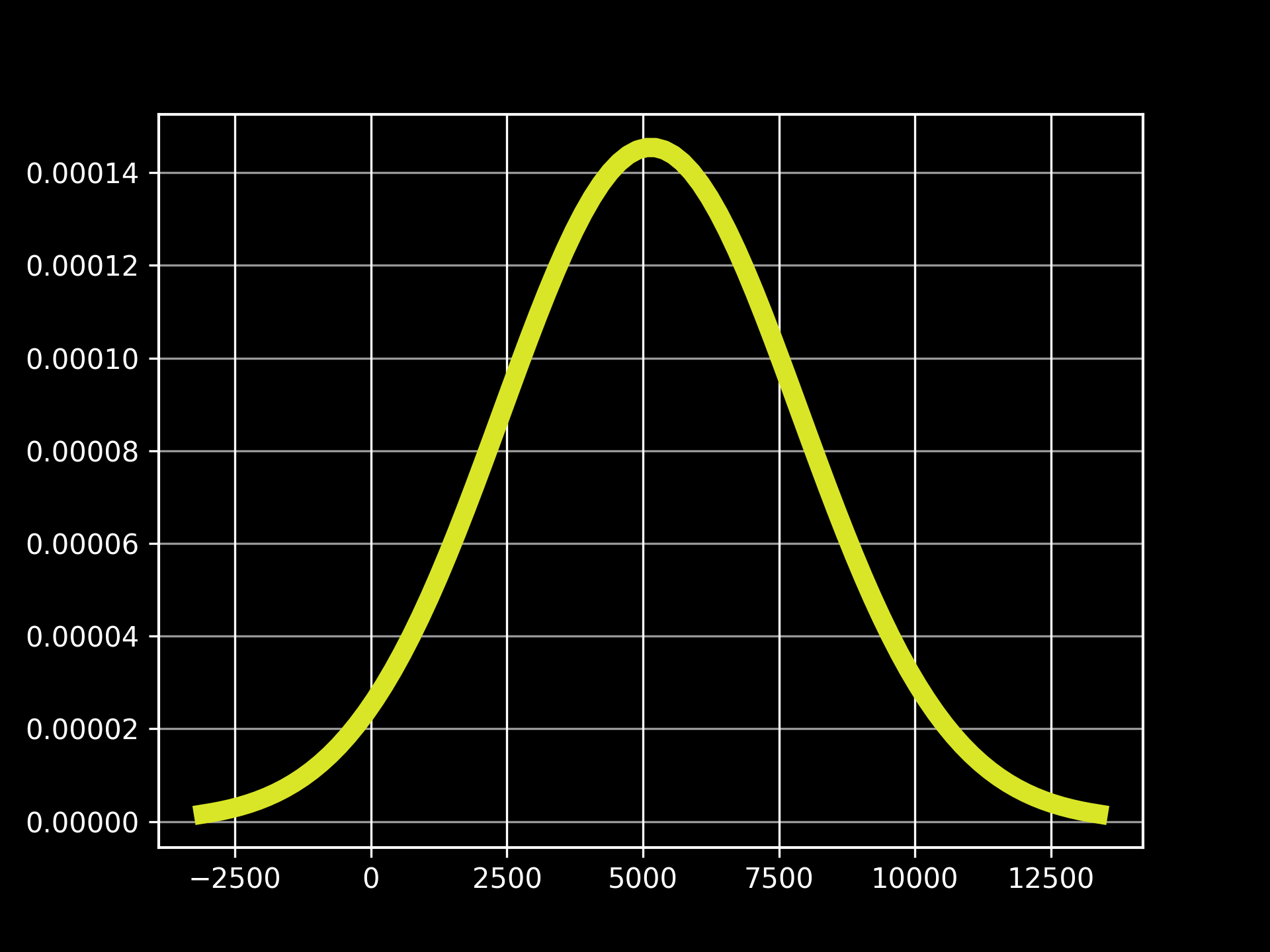
Marino Linić,

Dorian Manjarić,

Fran Poje,

Luka Krulčić,

Ivan Matejčić



*Normalna distribucija nasumično generiranih podataka iz uzorka u radu*

*Rijeka, 19. lipnja 2021.*

Tablica sadržaja

[Korišteni Python moduli i biblioteke 3](#_Toc75032499)

[Uvod 4](#_Toc75032500)

[Deskriptivna statistika 5](#_Toc75032501)

[Tablica frekvencija 5](#_Toc75032502)

[Histogram frekvencija i relativnih frekvencija 6](#_Toc75032503)

[Kumulativne frekvencije 8](#_Toc75032504)

[Mjere centralne tendencije 9](#_Toc75032505)

[Položajne mjere centralne tendencije 11](#_Toc75032506)

[Raspon varijacije 12](#_Toc75032507)

[Kvantili 13](#_Toc75032508)

[Percentili 14](#_Toc75032509)

[Suma apsolutnih vrijednosti odstupanja od srednje vrijednosti 15](#_Toc75032510)

[Prosječno apsolutno odstupanje od aritmetičke sredine 15](#_Toc75032511)

[Varijanca 15](#_Toc75032512)

[Standarda devijacija 16](#_Toc75032513)

[Korigirana varijanca 16](#_Toc75032514)

[Korigirana standardna devijacija 16](#_Toc75032515)

[Kutijasti (Box-plot) dijagram 17](#_Toc75032516)

[Koeficijent asimetrije 18](#_Toc75032517)

[Pearsonova mjera asimetrije 18](#_Toc75032518)

[Bowleyjeva mjera asimetrije 19](#_Toc75032519)

[Mjera zaobljenosti 20](#_Toc75032520)

[Linijski graf za distribuciju podataka 21](#_Toc75032521)

[Histogrami frekvencija sa svim mjerama uzorka 22](#_Toc75032522)

[Intervali pouzdanosti 23](#_Toc75032523)

[Zaključak 24](#_Toc75032524)

[Cijeli ispis u konzoli 25](#_Toc75032525)

[Kôd 25](#_Toc75032526)

[Tablica opisa slika 26](#_Toc75032527)

# Korišteni Python moduli i biblioteke

Za izradu seminarskog rada, kao i za rad sa uzorkom, korištena je nekolicina python-ovih biblioteka i modula. Ispod su navedene knjižice, popraćene objašnjenjem.

**Python Statistics** – biblioteka koja omogućuje rad sa standardnim statističkim funkcionalnostima poput: aritmetičke sredine, geometrijske sredine, harmonijske sredine, itd.

**Python Scypy** – biblioteka koja za znanstvene i tehničke izračune poput: linearne algebre, integrale, signalne izračune, statističke izračune i sl.

**Python Numpy** – biblioteka koja koji omogućuje rad sa višedimenzionalnim listama, vektorima, matricama i dr.

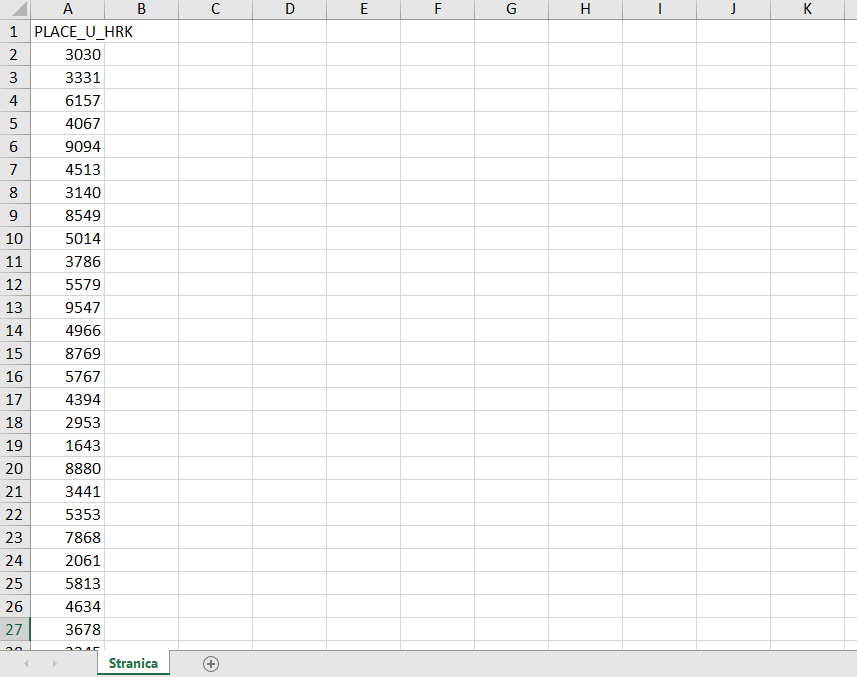
**Python Pandas** – biblioteka koja omogućuje funkcionalnosti za rad s podacima.

**Python Matplotlib** – biblioteka pomoću koje možemo crtati grafove i generirati slike grafova.

**Python Tabulate** – biblioteka koja nam omogućuje formatiranje generiranih podataka u uređenu tablicu.

# Uvod

Za ovaj rad uzeli smo izmišljeni i nasumično generirani uzorak od 200 podataka u iznosima plaća sa web stranice <https://randat.com> i spremili smo ga u datoteku ***place.xml***. U ovom istraživanju promatramo numeričke varijable, odnosno kvantitativne.



Slika 1. XML dokument "place.xml" s našim uzorkom od 200 podataka.

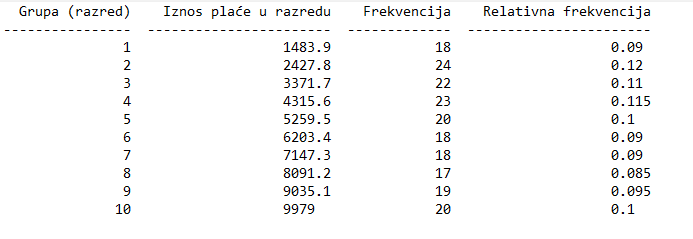
Kôd u programskom jeziku Python pisali smo zajedno u online okruženju <https://replit.com>. Ujedno i preporučamo pokretanje kôda u istom okruženju. Sve slike i tablice, ukoliko nisu dovoljne veličine, mogu se povećati korištenjem raznih preglednih alata, ali nalaze se i u zapakiranoj RAR datoteci.

# Deskriptivna statistika

## Tablica frekvencija

Tablicu frekvencija smo ispisali u konzolu i u datoteku „**tablica.txt**“.

1. *„Grupa (razred)“* stupac – označava 10 razreda s jednakom udaljenosti ((maksimalna vrijednost – minimalna vrijednost) / br. razreda) po iznosu plaća, s obzirom na to da želimo sažeti informacije o plaćama i intuitivno ih analizirati (bez grupacije bi tablica bila samo granično korisnija od sirovih podataka).
2. *„Iznos plaće u razredu“* stupac – označava prosječnu plaću u razredu (pojam prosjeka ćemo kasnije navesti pod podnaslovom „Mjere centralnih tendencija“ i nadalje *aritmetičkom sredinom*)
3. *„Frekvencija“* stupac – označava broj podataka u razredu, odnosno frekvenciju pojavljivanja razreda unutar uzorka. Konkretnije, ako je vrijednost podatka iz uzorka unutar ranga razreda, onda se pridodaje frekventnosti toga razreda.
4. *„Relativna frekvencija“* stupac – označava frekvencije razreda unutar skupa razreda, odnosno da je totalna frekvencija = 1.

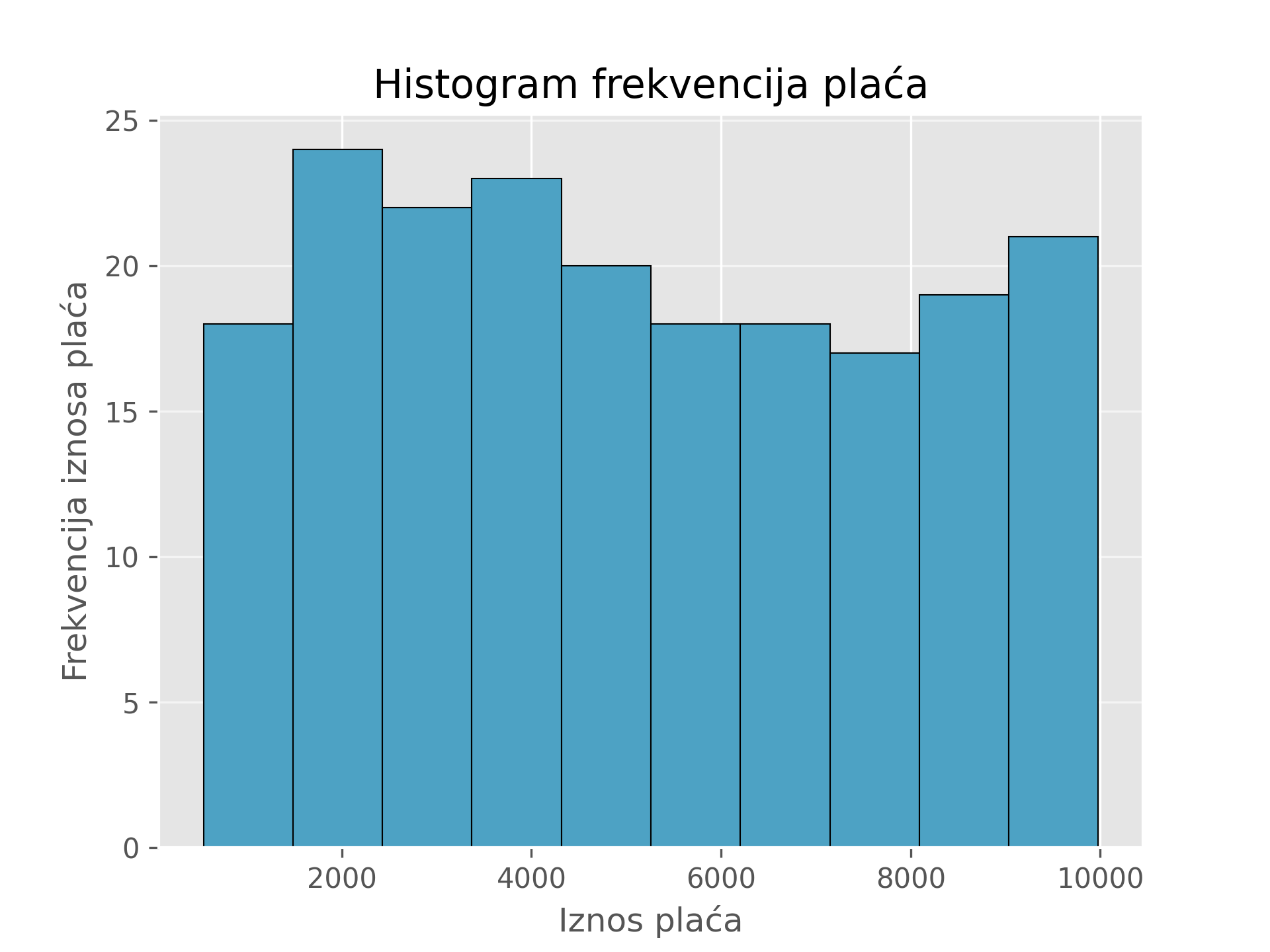


Slika 2. Tablica frekvencija

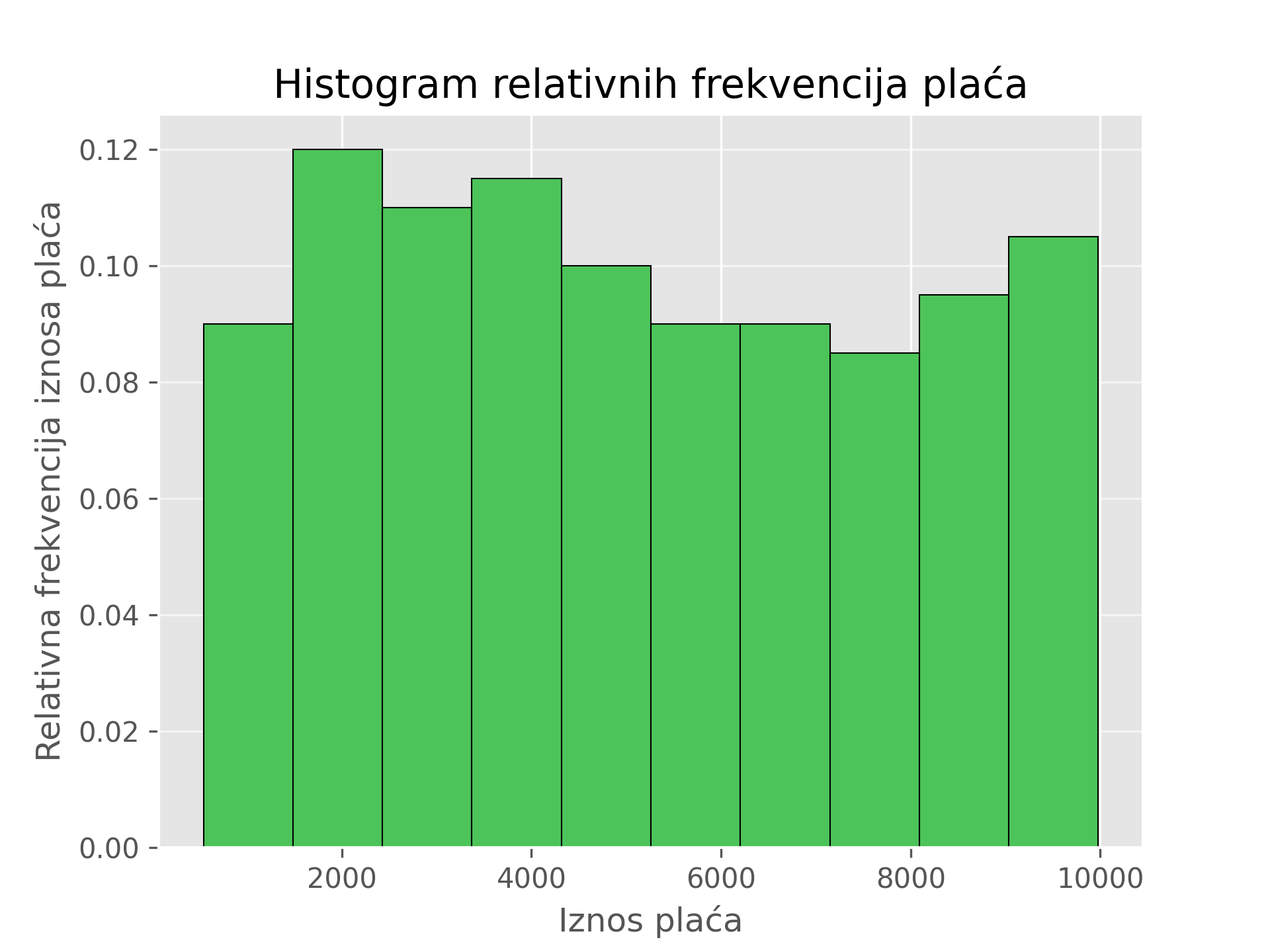


## Histogram frekvencija i relativnih frekvencija

Histogram je grafički prikaz odnosa podataka i njihovih frekvencija. Izgleda poput stupčastog grafikona. U našem slučaju, podaci su grupirani u razrede od 10 na X osi, te frekvencije do 25 na Y osi. Sve grafičke prikaze smo definirali koristeći Python biblioteku „Matplotlib“, stoga je nakon ovoga nećemo više navoditi. U kôdu smo naveli veličinu razreda na x osi pomoću jedne od funkcija biblioteke.



Slika 3. Histogram frekvencija plaća

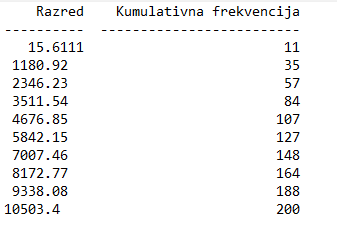


Slika 4. Histogram relativnih frekvencija plaća

*Kraj 1. zadatka.*

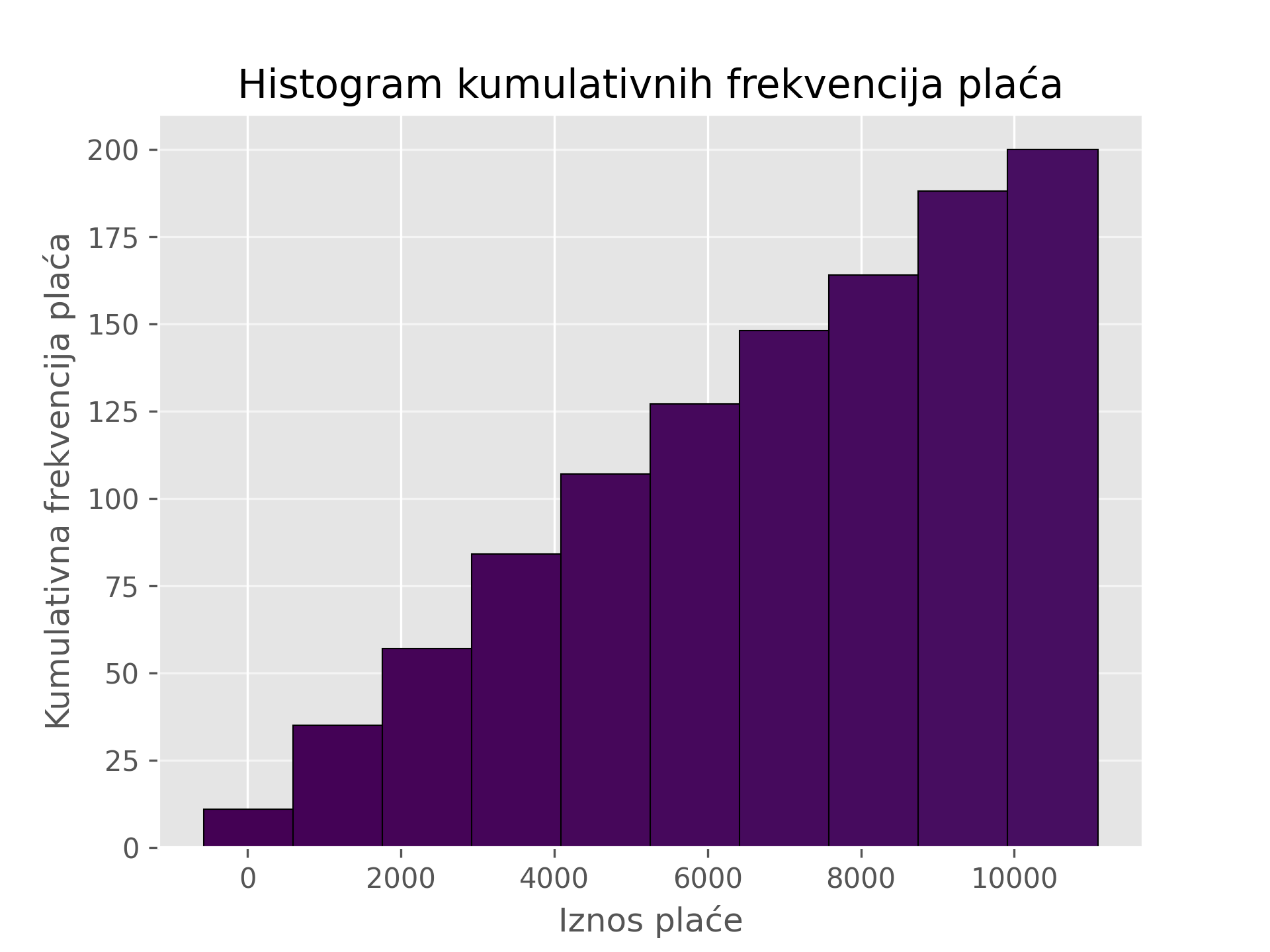
## Kumulativne frekvencije

Kumulativna frekvencija distribucija je po kojoj se analizira i zbraja frekventnost ispod i unutar razreda (odnosno raspona razreda). Frekventnost se ne može smanjivati kako raspon vrijednosti raste.



Slika 5. Tablica kumulativnih frekvencija



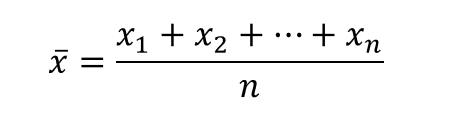


Slika 6. Graf kumulativnih frekvencija

*Kraj 2. zadatka.*

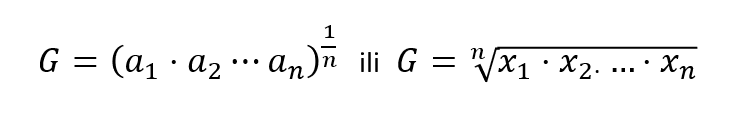
## Mjere centralne tendencije

Aritmetička sredina je jedna od središnjih vrijednosti koje se koriste u statistici i računa se za neki kvocijent sume članova skupa i broja članova tog skupa. Aritmetička sredina je najčešće korištena mjera centralne tendencije i računa se po formuli:



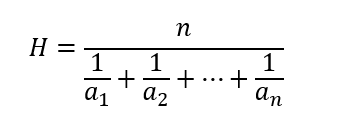
Aritmetička sredina za uzorak iz zadatka iznosi: **5145.635**

Geometrijska sredina je također jedna od mjere središnje tendencije, a pretežno se koristi kao mjera prosječne brzine nekih promjena. Računa se po formuli:

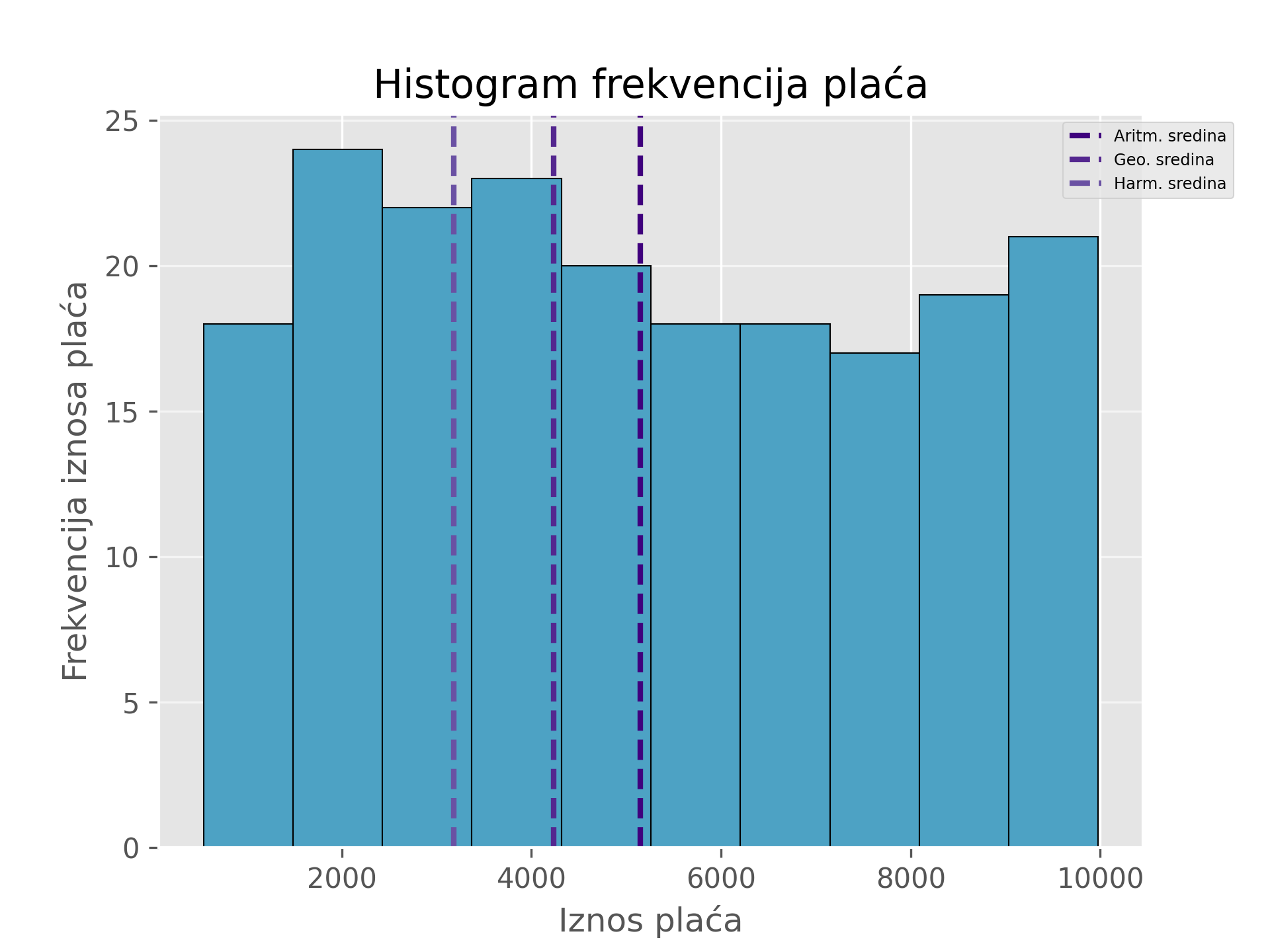


Geometrijska sredina za uzorak iz zadatka iznosi: **4231.869**

Harmonijska sredina je srednja vrijednost koja je recipročna vrijednost aritmetičke sredine recipročnih vrijednosti zadanih vrijednosti. Računa se po formuli:



Harmonijska sredina za uzorak iz zadatka iznosi: **3181.014**



Slika 7. Histogram frekvencija plaća s mjerama centralne tendencije

Tu možemo i vizualizirati mjere centralne tendencije u našem uzorku.

*Kraj 3. zadatka.*

## Položajne mjere centralne tendencije

Modus ili skraćeno mod je još jedna mjera centralne tendencije i predstavlja najčešću vrijednost u nekom nizu tako zvanu sredinu distribucije.

Nadalje, ako je distribucija numeričke varijable grupirana u prave razrede, onda se razred s najvećom korigiranom frekvencijom *b* naziva se modali razred. Ako je *L1*donja granica tog razreda, a *l* njegova veličina te *a* frekvencija razreda koji prethodi modalnom, a *c* frekvencija razreda koji slijedi iza modalnog, onda se Mod aproksimira formulom

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

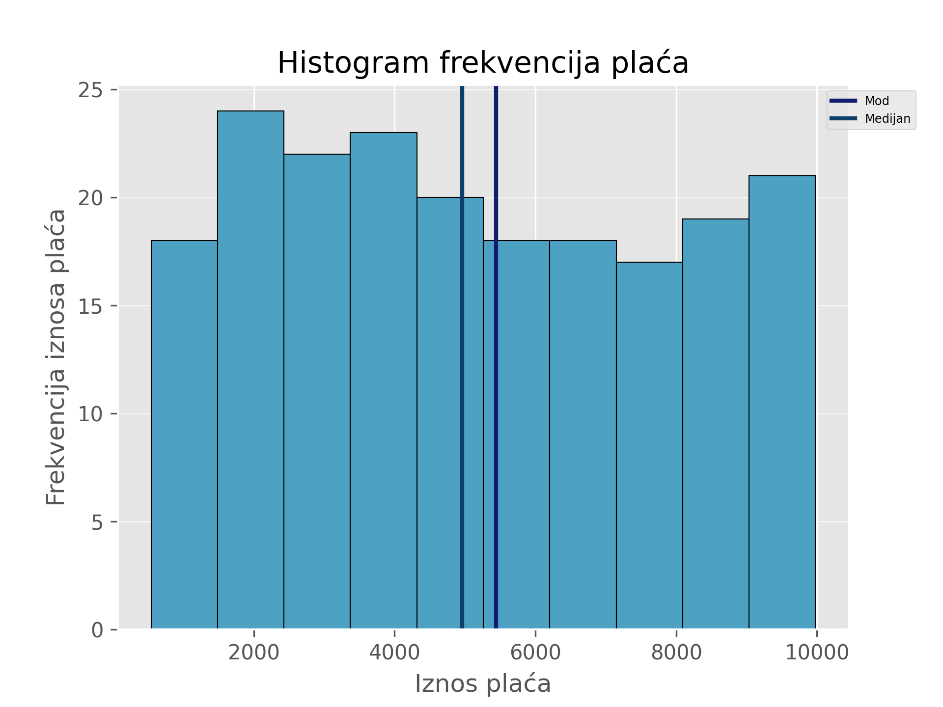
Mod za uzorak iz zadatka iznosi: **5440**

Medijan je vrijednost nekog skupa za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti i 50% podataka je veće ili jednako njoj. Ako je broj promatranih podataka paran onda medijan dobijemo s polu zbrojem središnjih podataka.

U slučaju da je N (članovi populacije) paran, Medijan se računa kao poluzbroj vrijednosti središnjih članova po formuli

*R = xmax – xmin.*

Medijan za uzorak iz zadatka iznosi: **4961.5**



Slika 8. Histogram frekvencije plaća s položajnim mjerama centralne tendencije

*Kraj 4. zadatka.*

## Raspon varijacije

Raspon podataka je mjera koja prikazuje koliko su podaci raspršeni. Računa se kao razlika najveće i najmanje vrijednosti u zadanom uzorku.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Raspon za uzorak u zadatku iznosi: **9439**

Najveći element skupa je **9979.**

Najmanji element skupa je **540.**

*Kraj 5. zadatka.*

## Kvantili

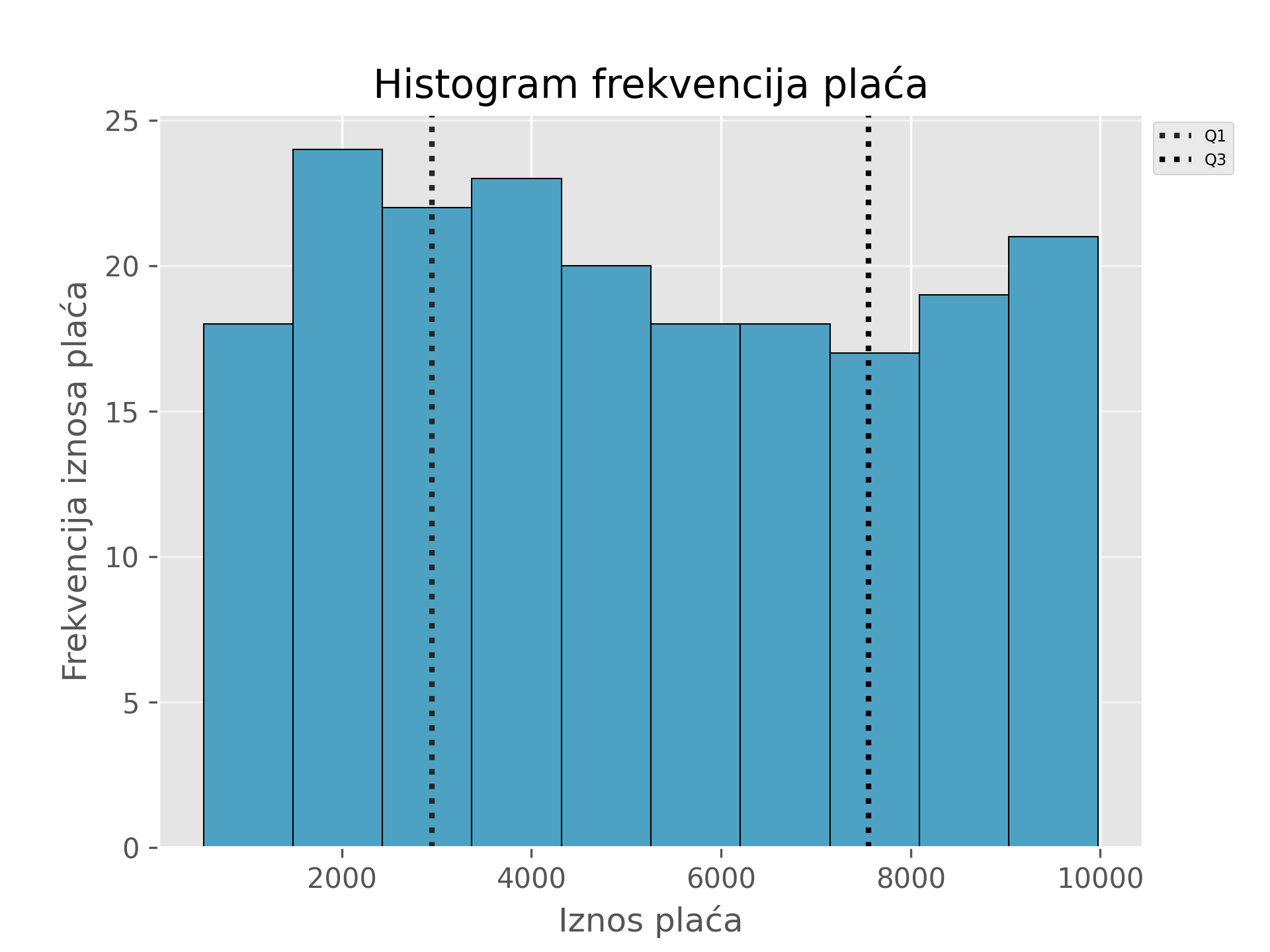
Kvantili su položajne vrijednosti obilježja koje uređene statističke nizove dijele na jednake dijelove. U praksi se najčešće koriste neki specifični kvantili: četvrtog reda (kvartili), desetog reda (decili) i stotog reda (centili/percentili). Kvartili su položajne vrijednosti koje uređene statističke nizove dijele na četiri jednaka dijela.

Prvi ili donji kvartil je broj od kojeg je 25% podataka manje ili jednako tom broju. Računa se tako da nađemo koji broj se nalazi na 25% cijelog promatranog uzorka.

Prvi kvartil za uzorak iz zadatka iznosi: **2947.25**

Treći ili gornji kvartil je broj od kojega je 75% manje ili jednako tom broju. Računa se tako da nađemo koji broj se nalazi na 75% cijelog promatranog uzorka.

Treći kvartil za uzorak iz zadatka iznosi: **7554.25**

****

Slika 9. Histogram frekvencija plaća s prvim i trećim kvartilom

*Kraj 6. zadatka.*

## Percentili

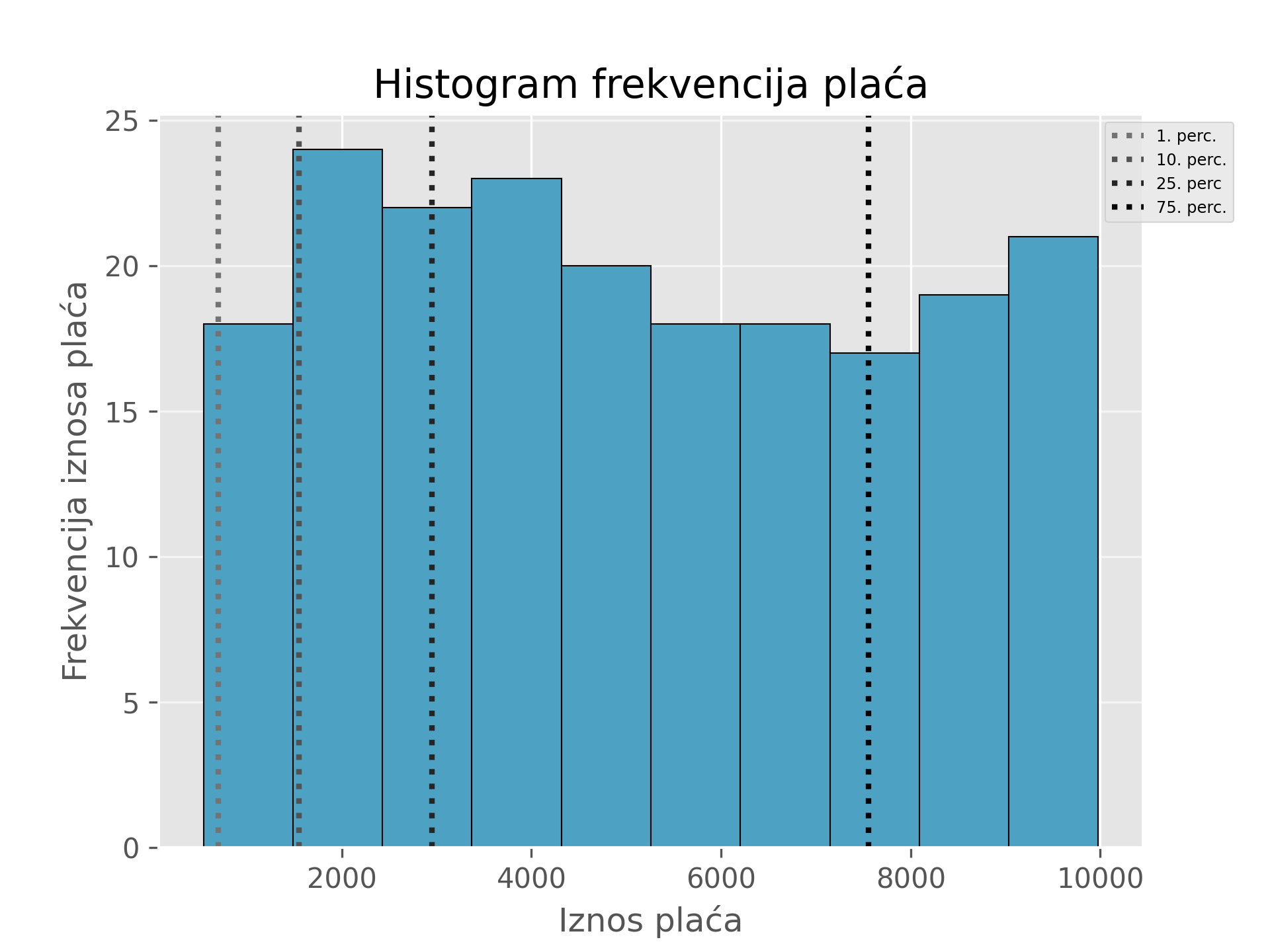
Percentil je statistička mjera koja spada u mjere lokacije. Percentil je broj koji predstavlja koliki je postotak brojeva manji ili jednak njemu.

1. percentil: **699.24**

10. percentil: **1548.4**

25. percentil je ujedno i 1. kvartil te iznosi: **2947.25**

75. percentil je ujedno i 3. kvartil te iznosi: **7554.25**



Slika 10. Histogram frekvencija plaća s percentilima

*Kraj 7. zadatka.*

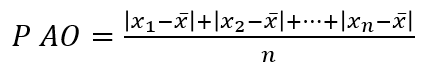
## Suma apsolutnih vrijednosti odstupanja od srednje vrijednosti

Suma apsolutnih vrijednosti odstupanja služi da vidimo rasipanje podataka oko srednje vrijednosti. Računa se po formuli :

Suma apsolutnih vrijednosti odstupanja za uzorak iz zadatka iznosi: **475335.350**

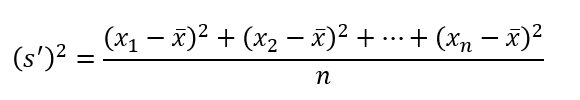
## Prosječno apsolutno odstupanje od aritmetičke sredine

Prosječno apsolutno odstupanje od aritmetičke sredine računamo tako da sumu apsolutnih vrijednosti odstupanja podijelimo s brojem uzoraka:



Prosječno apsolutno odstupanje od aritmetičke sredine za uzorak iz zadatka iznosi: **2376.677**

## Varijanca

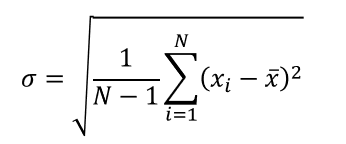
Varijanca uzorka definira se kao prosječno kvadratno odstupanje od prosjeka i računa se po formuli:

Varijanca za uzorak iz zadanog zadatka iznosi: **7519289.372**

*Kraj 8. zadatka.*

## Standarda devijacija

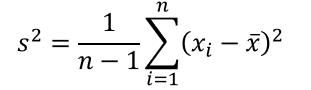
Standardna devijacija je mjera koja se koristi da kvantificira iznos varijacije ili disperzije vrijednosti podataka. Računa se po formuli:



Standardna devijacija za uzorak iz zadanog zadatka iznosi: **2742.132**

## Korigirana varijanca

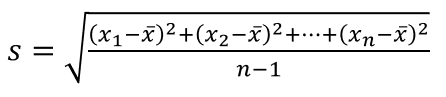
Korigirana varijanca jednaka je zbroju svih kvadrata razlike elemenata uzorka i aritmetičke sredine podijeljenih s duljinom uzorka umanjenom za jedan. Računa se po formuli:



Korigirana varijanca za uzorak u zadanom zadatku iznosi: **7557074.746**

## Korigirana standardna devijacija

Korigirana standardna devijacija uzorka koristi se za procjenu standardne devijacije populacije. Računa se po formuli:



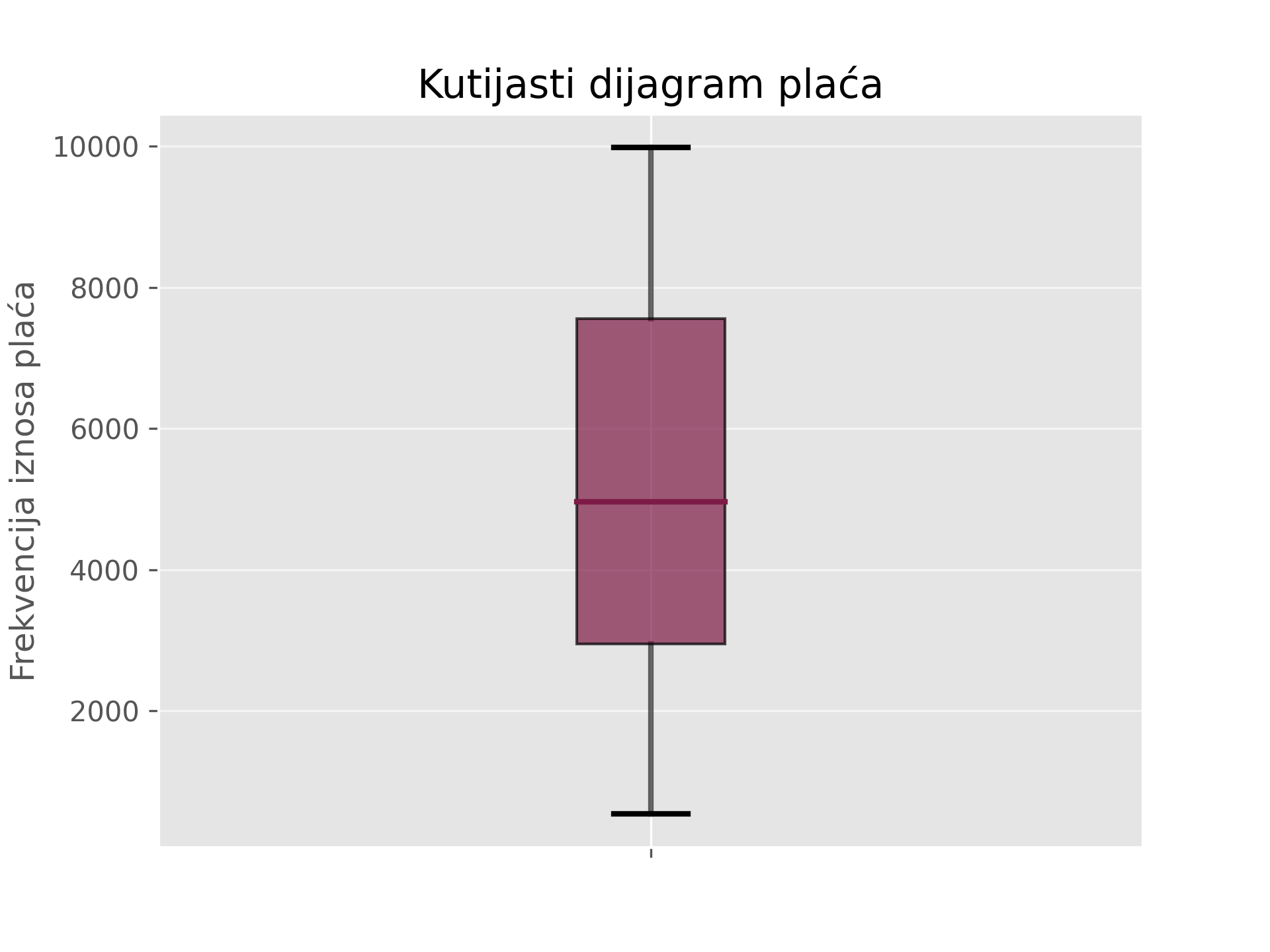
Korigirana standardna devijacija uzorka iz zadatka iznosi: **2749.013**

*Kraj 9. zadatka.*

## Kutijasti (Box-plot) dijagram

Box-plot ili kutijasti dijagram sastoji se od box-a odnosno pravokutnika koji predstavlja podatke od 1. odnosno donjeg kvartila do gornjeg odnosno 4. kvartila.

Sve točke koje se nalaze izvan područja koje predstavlja podatke od 1. do 4. kvartila nazivaju se „stršeće vrijednosti” koje predstavljaju vrijednosti sa malom frekvencijom pojavljivanja i velikim otklonom od srednje vrijednosti, medijana i moda.

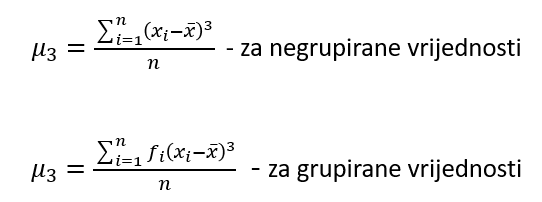


Slika 11. Kutijasti dijagram plaća

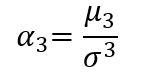
*Kraj 10. zadatka.*

## Koeficijent asimetrije

Koeficijent asimetrije koristi sva odstupanja vrijednosti numeričke varijable od aritmetičke sredine i po tome je potpuna mjera asimetrije. Za mjerenje asimetrije polazna je veličina aritmetička sredina na treću potenciju odnosno treći moment oko sredine koja se računa po formuli:



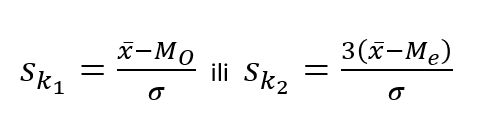
A koeficijent asimetrije α\_3 računa se po formuli:



Koeficijent asimetrije za uzorak iz zadatka iznosi: **0.113** i zbog toga je distribucija podataka pozitivno asimetrična.

## Pearsonova mjera asimetrije

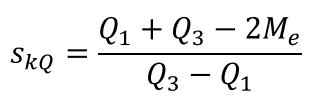
Pearsonova mjera asimetrije je standardizirano odstupanje medijana ili moda od aritmetičke sredine. Pearsonova mjera je nepotpuna mjera asimetrije i manje je informativna od koeficijenata asimetrije, ali je jednostavnija i brže se izračuna. U pravilu vrijednosti su joj u intervalu [-3, 3]. Računa se po formuli:



Pearsonova mjera asimetrije uzorka iz zadatka iznosi: **-0.108** i **0.201**

## Bowleyjeva mjera asimetrije

Bowleyjeva mjera asimetrije je mjera asimetrije koja se temelji na odnosima kvartila i medijana. U pravilu zauzima vrijednosti u intervalu [-1, 1]. Kao i Pearsonova mjera, Bowleyjeva mjera je također nepotpuna mjera asimetrije. Računa se po formuli:



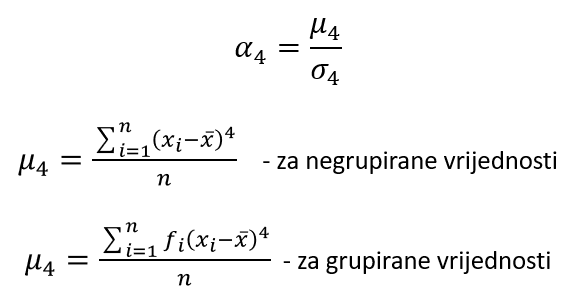
Bowleyjeva mjera asimetrije za uzorak iz zadatka iznosi: **10499.346**

Distribucija uzorka iz zadatka je pozitivno asimetrična distribucija zato što je razlika 4. kvartila i medijana veća od razlike medijana i 1. kvartila.

*Kraj 11. zadatka.*

## Mjera zaobljenosti

Mjera zaobljenosti mjeri zaobljenost modalnog vrha distribucije. Izračunava se kao omjer četvrtog momenta oko sredine i standardne devijacije na četvrtu potenciju. Za različite rezultate koji se uspoređuju s brojem 3 postoje različite distribucije podataka. Računa se po formuli:



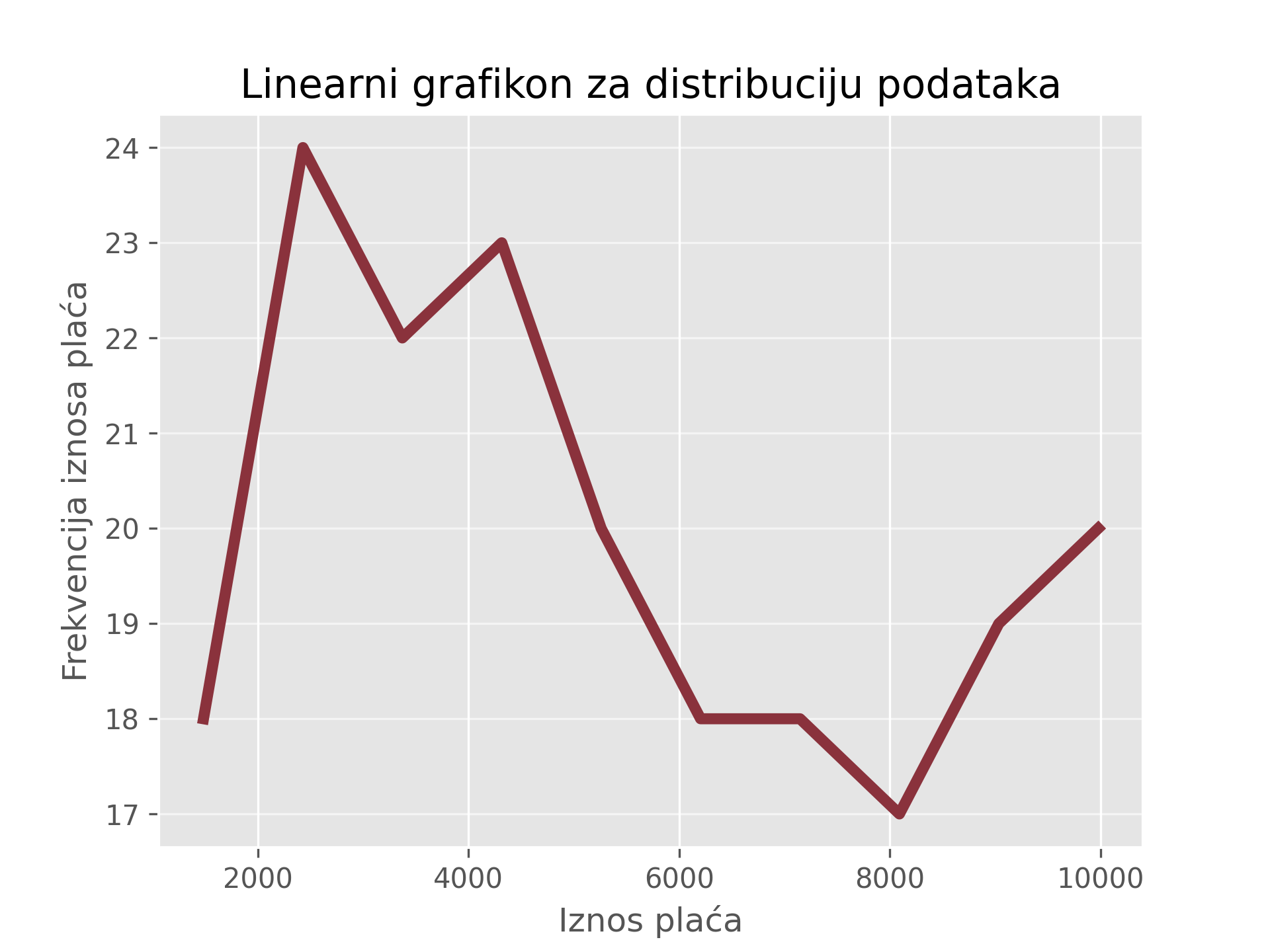
Mjera zaobljenosti za uzorak iz zadatka iznosi: **-1.198**

*Kraj 12. zadatka.*

## Linijski graf za distribuciju podataka

Na ovom grafu prikazana je distribucija plaća po njihovim frekvencijama i iznosima.

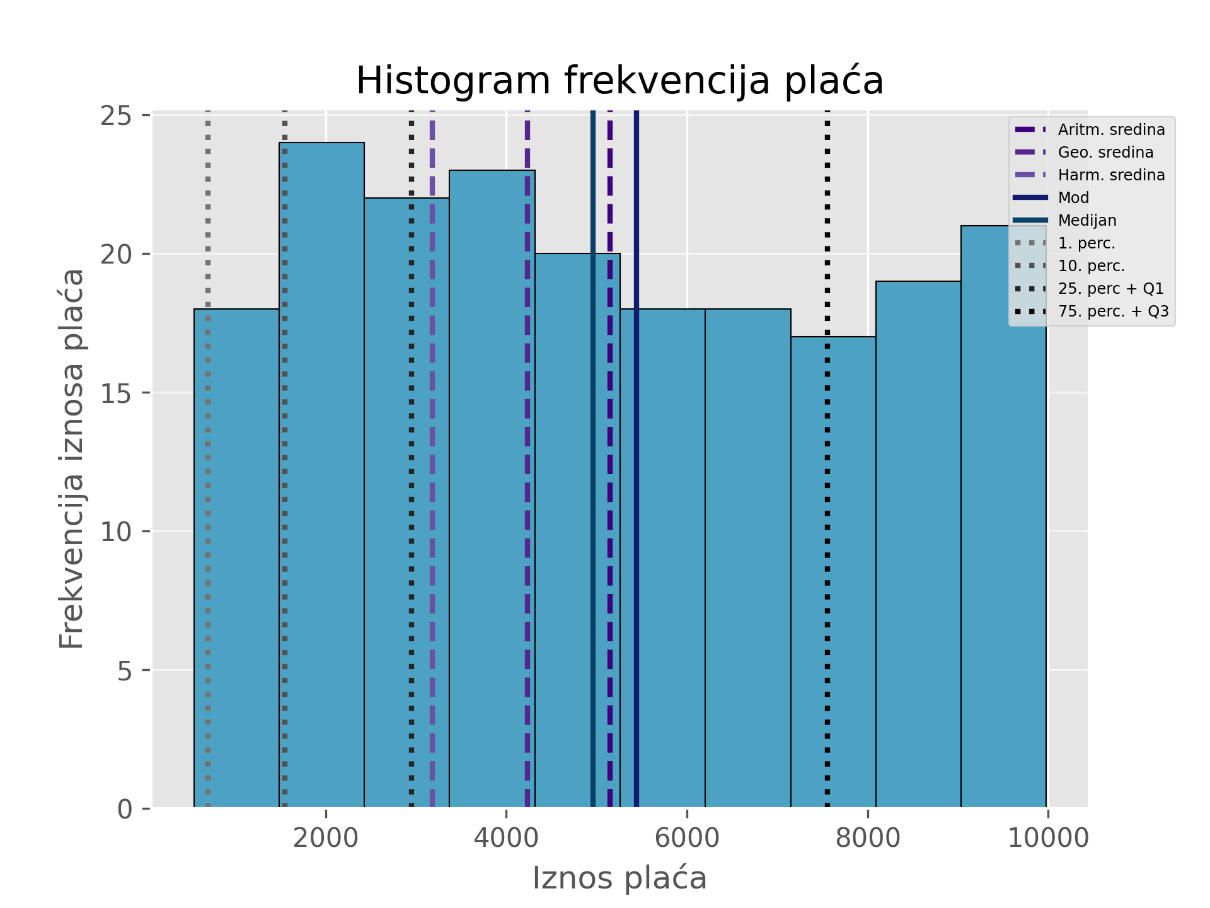
Frekvencije iznosa plaća su predstavljeni pomoću Y osi, a iznosi placa X osi.



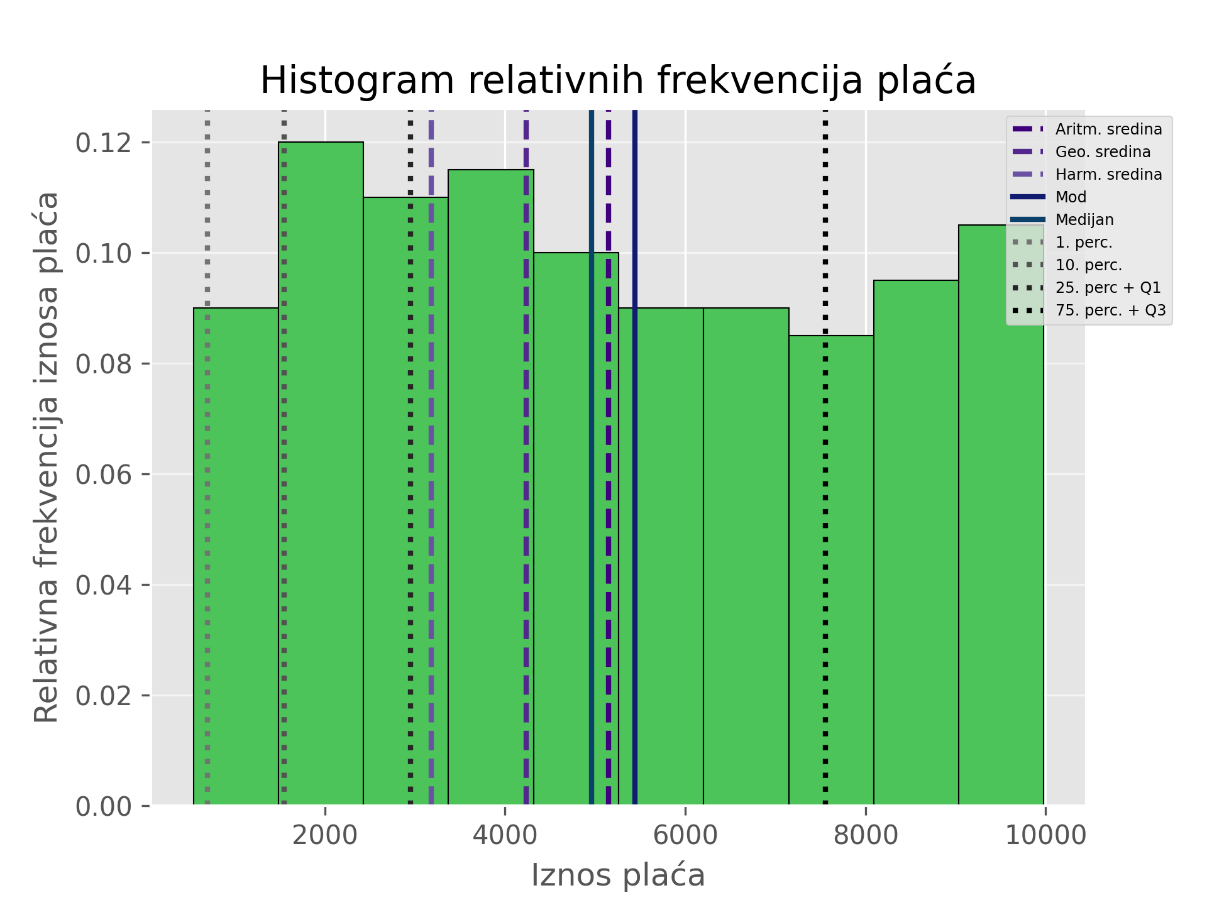
Slika 12. Linearni graf distribucije podataka

*Kraj 13. zadatka.*

## Histogrami frekvencija sa svim mjerama uzorka



Slika 13. Histogram frekvencija plaća sa svim mjerama uzorka



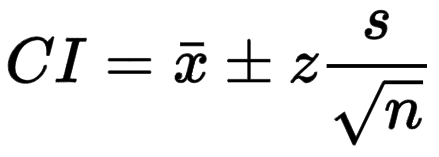
Slika 14. Histogram relativnih frekvencija plaća sa svim mjerama uzorka

# Intervali pouzdanosti

Interval pouzdanosti je interval u kojem se s predodređenoj pouzdanosti nalaze vrijednosti određenog parametra s odgovarajućom vjerojatnosti. To ne znači da se u, primjerice, intervalu pouzdanosti od 95% nalazi 95% podataka, nego da se u tom intervalu nalazi traženi parametar s 95% pouzdanosti. U našem slučaju, taj parametar je *očekivanje danog uzorka*.

Interval pouzdanosti koristimo da bi uzeli u obzir činjenicu da ne analiziramo cijelu populaciju uzorka te da imamo pogrešku uzorkovanja. Pouzdanost će rasti s povećanjem intervala, zato što obuhvaćanje više podataka znači i manji prostor za pogrešku, ali inversno onda imamo i manju korist od našeg intervala.

Formula za interval pouzdanosti slijedi:



Gdje je:  
CI = interval pouzdanosti  
x̄ = aritmetička sredina uzorka (očekivanje)

z = vrijednost pouzdanosti

s = standardna devijacija

n = broj podataka u uzorku

Interval pouzdanosti za očekivanje danog uzorka uz pouzdanost od 95%:   
**4762.31 ≤ x ≤ 5528.95**,gdje x označava interval pouzdanosti.

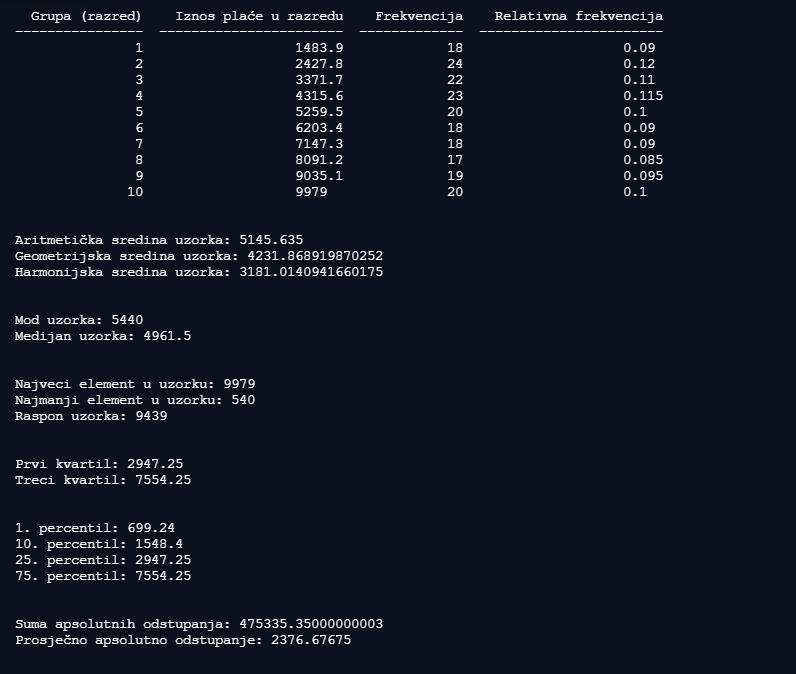
Interval pouzdanosti za očekivanje danog uzorka uz pouzdanost od 85%: **4864.73 ≤ x ≤ 5426.54**,gdje x označava interval pouzdanosti.

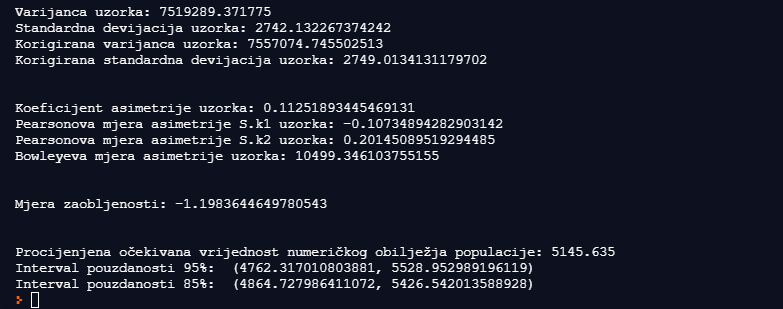
*Kraj 1. i 2. zadatka intervala pouzdanosti.*

# Zaključak

Nimalo začuđujuće s obzirom na to da su podaci generirani specifično u okviru plaća na web stranici s koje su preuzeti, naši podaci općenito indiciraju standardno variranje i ostale parametre analize našeg uzorka. Unatoč tome, plaće se svejedno mogu razlikovati skoro desetorostruko unutar samo 200 podataka, i relativno su jednako raspoređene po razredima.

## Cijeli ispis u konzoli





## Kôd



# Tablica opisa slika

[Slika 1. XML dokument "place.xml" s našim uzorkom od 200 podataka. 4](#_Toc75032452)

[Slika 2. Tablica frekvencija 5](#_Toc75032453)

[Slika 3. Histogram frekvencija plaća 6](#_Toc75032454)

[Slika 4. Histogram relativnih frekvencija plaća 7](#_Toc75032455)

[Slika 5. Tablica kumulativnih frekvencija 8](#_Toc75032456)

[Slika 6. Graf kumulativnih frekvencija 8](#_Toc75032457)

[Slika 7. Histogram frekvencija plaća s mjerama centralne tendencije 10](#_Toc75032458)

[Slika 9. Histogram frekvencije plaća s položajnim mjerama centralne tendencije 11](#_Toc75032459)

[Slika 10. Histogram frekvencija plaća s prvim i trećim kvartilom 13](#_Toc75032460)

[Slika 11. Histogram frekvencija plaća s percentilima 14](#_Toc75032461)

[Slika 12. Kutijasti dijagram plaća 17](#_Toc75032462)

[Slika 13. Linearni graf distribucije podataka 21](#_Toc75032463)

[Slika 14. Histogram frekvencija plaća sa svim mjerama uzorka 22](#_Toc75032464)

[Slika 15. Histogram relativnih frekvencija plaća sa svim mjerama uzorka 22](#_Toc75032465)