

SQL

za analizu podataka

Upom Malik
Matt Goldwasser
Benjamin Johnston

Izvršite brzu i efikasnu analizu podataka uz pomoć moćnog SQL-a



SQL

za analizu podataka

**Upom Malik
Matt Goldwasser
Benjamin Johnston**



Packt

Izdavač:



Obalskih radnika 4a, Beograd

Tel: 011/2520272

e-mail: kombib@gmail.com

internet: www.kombib.rs

Urednik: Mihailo J. Šolajić

Za izdavača, direktor:

Mihailo J. Šolajić

Autori: Upom Malik
Matt Goldwasser
Benjamin Johnston

Prevod: Biljana Tešić

Lektura: Miloš Jevtović

Slog: Zvonko Aleksić

Znak Kompjuter biblioteke:

Miloš Milosavljević

Štampa: „Pekograf“, Zemun

Tiraž: 500

Godina izdanja: 2019.

Broj knjige: 520

Izdanje: Prvo

ISBN: 978-86-7310-543-7

SQL for Data Analytics

Upom Malik
Matt Goldwasser
Benjamin Johnston

ISBN 978-1-78980-735-6
Copyright © 2019 Packt Publishing

All right reserved. No part of this book may be reproduced or transmitted in any form or by means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher. Autorizovani prevod sa engleskog jezika edicije u izdanju „Packt Publishing“, Copyright © 2019.

Sva prava zadržana. Nije dozvoljeno da nijedan deo ove knjige bude reprodukovan ili snimljen na bilo koji način ili bilo kojim sredstvom, elektronskim ili mehaničkim, uključujući fotokopiranje, snimanje ili drugi sistem presnimavanja informacija, bez dozvole izdavača.

Zaštitni znaci

Kompjuter Biblioteka i „Packt Publishing“ su pokušali da u ovoj knjizi razgraniče sve zaštitne oznake od opisnih termina, prateći stil isticanja oznaka velikim slovima.

Autor i izdavač su učinili velike napore u pripremi ove knjige, čiji je sadržaj zasnovan na poslednjem (dostupnom) izdanju softvera. Delovi rukopisa su možda zasnovani na predizdanju softvera dobijenog od strane proizvođača. Autor i izdavač ne daju nikakve garancije u pogledu kompletnosti ili tačnosti navoda iz ove knjige, niti prihvataju ikakvu odgovornost za performanse ili gubitke, odnosno oštećenja nastala kao direktna ili indirektna posledica korišćenja informacija iz ove knjige.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд,
се добија на захтев

The top half of the page features a background of marbled paper with intricate, swirling patterns in shades of grey and white. The word "UVOD" is printed in a large, bold, black, sans-serif font on the right side of this section.

UVOD

O ovom odeljku

U ovom odeljku ukratko su predstavljeni autori, teme koje su obuhvaćene u knjizi, neophodne tehničke veštine za početak rada i hardverski i softverski zahtevi koji su potrebni za završetak svih navedenih radnji i vežbi.

O KNJIZI

Razumevanje i pronalaženje obrazaca u podacima predstavljaju najbolji način za poboljšanje poslovnih odluka. Ako poznajete osnove SQL-a, ali ne znate kako da koristite SQL da biste iz podataka stekli uvid u posao, knjiga „SQL za analizu podataka“ je za vas.

Ova knjiga obuhvata sve što vam je potrebno za napredak - od jednostavnog poznavanja osnovnog SQL-a, do pričanja priča i identifikacije trendova u podacima. Moći ćete da počnete da istražujete podatke prepoznavanjem obrazaca i „otključavanjem“ detaljnijih informacija. Takođe ćete steći iskustvo u analiziranju korišćenjem različitih tipova podataka u SQL-u, uključujući vremenske serije, geoprostorne podatke i tekstualne podatke. Na kraju, shvat ćete kako da postanete produktivni, koristeći SQL pomoću profilisanja i automatizacije da biste brže dobili informacije.

Kada pročitate u celosti ovu knjigu, moći ćete efikasno da koristite SQL u svakodnevnom poslovnim situacijama i da pregledate podatke „kritičkim očima“ profesionalnog analitičara.

O autorima

Upom Malik je analitičar podataka, koji je zaposlen u tehnološkoj industriji više od 6 godina. Magistrirao je hemijsko inženjerstvo na Univerzitetu „Cornell“ i diplomirao biohemiju na Univerzitetu „Duke“. Koristi SQL i druge alatke da bi rešio zanimljive izazove u finansijama, energetici i potrošačkoj elektronici. Dok je rešavao analitičke probleme, bio je stalno na putu, kao „digitalni nomad“. U slobodno vreme voli da čita, da pešači stazama na severoistoku Sjedinjenih Američkih Država i da uživa u činijama za ramen supu iz celog sveta.

Matt Goldwasser je vodeći analitičar podataka u kompaniji „T. Rowe Price“. Uživa u demistifikaciji nauke o podacima zainteresovanim stranama iz poslovnog sektora i u primeni proizvodnih rešenja za mašinsko učenje. Koristio je SQL za analizu podataka u finansijskoj industriji tokom poslednjih osam godina. Diplomirao je mašinsko i vazduhoplovno inženjerstvo na Univerzitetu „Cornell“. U slobodno vreme uživa da svoje dete podučava u analizi podataka.

Benjamin Johnston je viši analitičar podataka jedne od vodećih svetskih medicinskih kompanija, koja se oslanja na podatke, i učestvuje u razvoju inovativnih digitalnih rešenja tokom čitavog puta razvoja proizvoda – od definicije problema, do istraživanja i razvoja rešenja, pa sve do konačne primene. Trenutno završava doktorat iz mašinskog učenja, specijalizirajući obradu slike i duboke konvolucione neuronske mreže. Ima više od 10 godina iskustva u dizajniranju i razvoju medicinskih proizvoda (učestvovao je u raznim tehničkim „ulogama“), a ima i diplomu sa počastima iz inženjerskih i medicinskih nauka Univerziteta u Sidneju, u Australiji.

Ciljevi učenja

Kada pročitate u celosti ovu knjigu, moći ćete:

- **da** izvršite napredne statističke proračune pomoću funkcije WINDOWS
- **da** koristite SQL upite i podupite za pripremu podataka za analizu
- **da** uvezete i izvezete podatke, koristeći tekstualnu datoteku i psql
- **da** primenite posebne SQL klauzule i funkcije za generisanje opisne statistike
- **da** analizirate specijalne tipove podataka u SQL-u, uključujući geoprostorne podatke i podatke o vremenu
- **da** optimizujete upite da biste poboljšali njihove performanse, radi dobijanja bržih rezultata
- **da** debugujete upite koji se ne mogu izvršiti
- **da** koristite SQL da biste sumirali i identifikovali obrasce u podacima.

Publika

Ako ste administrator baze podataka koji želi da pređe na analitiku ili backend inženjer koji želi da bolje razume proizvodne podatke, smatraćete ovu knjigu korisnom. Ova knjiga je idealna i za analitičare podataka ili poslovne analitičare koji žele da poboljšaju svoje veštine iz oblasti analize podataka korišćenjem SQL-a. Poznavanje osnovnih koncepata SQL-a i baze podataka će pomoći u razumevanju koncepata koji su obuhvaćeni u ovoj knjizi.

Pristup

U knjizi „*SQL za analizu podataka*“ savršeno su usklađene teorija i praktične vežbe i obezbeđen je praktičan pristup analizi podataka. Fokus je na obezbeđivanju praktičnih uputstava za SQL i statističku analizu da biste bolje razumeli svoje podatke. Knjiga uklanja „mrvice“ i usredsređena je na praktičnost. Sadrži više aktivnosti u kojima se koriste poslovni scenariji iz stvarnog života za vežbanje i primenu novih veština u izuzetno relevantnom kontekstu.

Hardverski zahtevi

Za optimalni doživljaj preporučujemo sledeću konfiguraciju hardvera:

- **Procesor:** Intel Core i5 ili ekvivalent
 - **Memorija:** 4 GB RAM memorije
 - **Skladištenje:** 5 GB slobodnog prostora
-

Softverski zahtevi

Takođe preporučujemo da unapred instalirate sledeće softvere:

- **Operativni sistem:** 64-bitni Windows 7 SP1, 64-bitni Windows 8.1, 64-bitni Windows 10, Linux (Ubuntu 16.04 ili noviji, Debian, Red Hat ili Suse) ili najnovija verzija macOS-a
- **PostgreSQL 10.9** (<https://www.postgresql.org/download/>)
- **Anaconda Python 3.7** (<https://www.anaconda.com/distribution/#downloadsection>)
- **Git 2** ili noviji

Konvencije

Reči koda u tekstu, nazivi tabele baze podataka, nazivi direktorijuma, nazivi datoteka, ekstenzije datoteka, nazivi putanja, skraćeni URL-ovi, korisnički unos i Twitter postovi su prikazani na sledeći način:

„Ovde treba napomenuti da formatiranje za komandu `\copy` može izgledati malo neuredno, jer ne dozvoljava komande sa novim linijama. Jednostavan način za omogućavanje komande sa novim linijama je kreiranje prikaza koji sadrži vaše podatke pre komande `\copy`, a zatim se dodaje prikaz nakon što je završena komanda `\copy`.“

Blok koda je prikazan na sledeći način:

```
CREATE TEMP VIEW customers_sample AS ( SELECT *
    FROM customers LIMIT 5
);
\copy customers_sample TO 'my_file.csv' WITH CSV
HEADER DROP VIEW customers_sample;
```

Instalacija i podešavanje

Svako veliko putovanje započinje skromnim korakom, a naša predstojeća „avantura“ u svetu obrade podataka nije izuzetak. Da bismo mogli dobro da iskoristimo podatke, moramo da pripremimo najproduktivnije okruženje. U ovom kratkom odeljku videćete kako se priprema to okruženje.

Instaliranje softvera PostgreSQL 10.9

Instaliranje na Windowsu:

Preuzmite program za instalaciju softvera PostgreSQL verzija 10 pomoću linka <https://www.postgresql.org/download/windows/> i pratite uputstva.

Instaliranje na Linuxu:

PostgreSQL možete instalirati na Ubuntu ili Debian Linuxu pomoću komandne linije, koristeći komandu:

```
sudo apt-get install postgresql-11
```

Instaliranje na macOS-u:

Preuzmite program za instalaciju PostgreSQL verzije 10 pomoću linka <https://www.postgresql.org/download/macosx/> i pratite uputstva.

Instaliranje Pythona

Instaliranje Pythona na Windowsu:

1. Pronađite željenu verziju Pythona na zvaničnoj stranici za instalaciju na adresi <https://www.anaconda.com/distribution/#windows>.
2. Izaberite Python 3.7 sa stranice za preuzimanje.
3. Pobrinite se da instalirate odgovarajuću arhitekturu za vaš računarski sistem (32-bitnu ili 64-bitnu arhitekturu). Ove informacije možete pronaći u prozoru **System Properties** u vašem operativnom sistemu.
4. Nakon što preuzmete program za instalaciju, samo dva puta kliknite na datoteku i sledite uputstva na ekranu koja su prilagođena korisnicima.

Instaliranje Pythona na Linuxu:

Za instaliranje Pythona na Linuxu imate nekoliko dobrih opcija:

1. Otvorite komandni odzivnik (Command Prompt) i proverite da li je **p\Python 3** već instaliran, tako što ćete pokrenuti komandu **python3 --version**.
2. Da biste instalirali Python 3, pokrenite komandu:

```
sudo apt-get update  
sudo apt-get install python3.7
```

3. Ako naiđete na probleme, u njihovom rešavanju mogu vam pomoći brojni izvori na mreži.
4. Instalirajte Anaconda Linux, tako što ćete preuzeti program za instalaciju sa linka <https://www.anaconda.com/distribution/#linux> i pratiti uputstva.

Instaliranje Pythona na macOS-u:

1. Slično kao i za Linux, imate nekoliko metoda za instaliranje Pythona na Macu. Da biste instalirali Python na operativnom sistemu macOS X, uradite sledeće:
2. Otvorite Terminal za Mac pritiskom na *CMD + taster za razmak*, ukucajte **terminal** u otvorenom polju za pretraživanje i pritisnite *Enter*.
3. Instalirajte Xcode pomoću komandne linije pokretanjem komande **xcode-select --install**. Najlakši način za instaliranje Pythona 3 je korišćenje Homebrea koji se instalira pomoću komandne linije pokretanjem komande **ruby -e „\$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)“**.
4. Dodajte Homebrew u promenljivu okruženja **\$PATH**. Otvorite svoj profil u komandnoj liniji pokretanjem komande **sudo nano ~/.profile** i na dnu umetnite **export PATH=\"/usr/local/opt/python/libexec/bin:\$PATH“**.
5. Završni korak je instalacija Pythona. U komandnoj liniji pokrenite komandu **brew install python**.
6. Ponavljamo: možete instalirati Python pomoću programa za instalaciju Anaconda, koji je dostupan na adresi <https://www.anaconda.com/distribution/#macos>.

Instaliranje Gita

Instaliranje Gita na operativnom sistemu Windows ili macOS X:

Git za Windows/Mac može se preuzeti i instalirati pomoću linka <https://git-scm.com/>. Međutim, za poboljšani korisnički doživljaj preporučuje se instaliranje Gita pomoću naprednog klijenta, kao što je GitKraken (<https://www.gitkraken.com/>).

Instaliranje Gita na Linuxu

Git se može jednostavno instalirati pomoću komandne linije:

```
sudo apt-get install git
```

Ako više volite grafički korisnički interfejs, imajte na umu da je GitKraken (<https://www.gitkraken.com/>) dostupan i za Linux.

Učitavanje primera baze podataka

U većini vežbi u ovoj knjizi koristi se primer baze podataka `sqllda`, koja sadrži izmišljene podatke za izmišljenu kompaniju električnih vozila „ZoomZoom“. Da biste instalirali bazu podataka na PostgreSQL-u, kopirajte datoteku `data.dump` iz fascikle `Datasets` u GitHub spremište knjige (<https://github.com/Training-ByPackt/SQL-for-Data-Analytics/tree/master/Datasets>). Zatim, učitajte datoteku `data.dump` iz komandne linije pomoću komande:

```
psql < data.dump
```

Ovde je `psql` PostgreSQL klijent.

Pokretanje SQL datoteka

Komande i iskazi mogu se izvršiti pomoću datoteke `*.sql` iz komandne linije korišćenjem komande:

```
psql <commands.sql
```

Alternativno, oni se mogu izvršiti pomoću SQL interpretera:

```
database=#
```

Dodatni resursi

Komplet kodova za knjigu možete da preuzmete sa našeg sajta:

<http://bit.ly/2osAppu>

Komplet kodova za ovu knjigu takođe se nalazi na GitHubu na adresi:

<http://bit.ly/2nEzjHc>

Grafički paket za knjigu možete preuzeti sa adrese:

<http://bit.ly/2nRn4XN>



Postanite član Kompjuter biblioteke

Kupovinom jedne naše knjige stekli ste pravo da postanete član Kompjuter biblioteke. Kao član možete da kupujete knjige u pretplati sa 40% popusta i učestvujete u akcijama kada ostvarujete popuste na sva naša izdanja.

Potrebno je samo da se prijavite preko formulara na našem sajtu.

Link za prijavu: <http://bit.ly/2TxeK5a>

Skenirajte QR kod
registrujte knjigu
i osvojite nagradu





1

Razumevanje i opisivanje podataka

Ciljevi učenja

Kada pročitate u celosti ovu knjigu, moći ćete:

- da objasnite podatke i tipove podataka
- da klasifikujete podatke na osnovu njihovih karakteristika
- da izračunate osnovnu univarijantnu statistiku podataka
- da identifikujete neuobičajene vrednosti (outliers)
- da koristite bivarijantnu analizu da biste razumeli vezu između dve promenljive

U ovom poglavlju će biti predstavljene osnove analize podataka i statistike. Takođe ćete naučiti kako da prepoznate neuobičajene vrednosti i shvatite veze između promenljivih.

UVOD

Podaci su, u osnovi, transformisali 21. vek. Zahvaljujući jednostavnom pristupu računarima, kompanije i organizacije uspele su da promene način na koji koriste veće i složenije skupove podataka. Sada se upotrebom podataka, i to pomoću samo nekoliko linija računarskog koda, mogu pronaći informacije koje je bilo praktično nemoguće saznati pre 50 godina. U ovom poglavlju ćemo razmotriti šta su podaci i kako se mogu koristiti za „otključavanje“ informacija i prepoznavanje obrazaca.

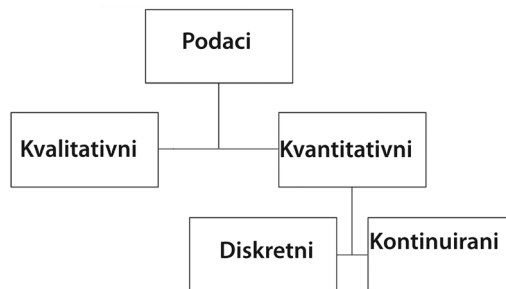
„SVET“ PODATAKA

Krenimo od prvog pitanja - šta su podaci. Podaci se mogu smatrati zapisanim merenjima nečega u stvarnom svetu. Na primer, lista visina su podaci - to jest, visina je merilo rastojanja između glave i nogu neke osobe. Nešto što podaci opisuju se obično zove **jedinica posmatranja**. U slučaju ovih visina osoba je jedinica posmatranja.

Kao što možete zamisliti, postoji mnogo podataka koje možemo prikupiti da bismo opisali neku osobu - uključujući njenu starosnu dob, težinu, da li je pušač i još mnogo štošta. Jedno ili više merenja koja se koriste za opisivanje jedne određene jedinice posmatranja zove se tačka podataka, a svako merenje u tački podataka zove se promenljiva (ona se često naziva i funkcija). Kada imate nekoliko tačaka podataka zajedno, imate skup podataka.

Tipovi podataka

Podaci se mogu podeliti i na dve osnovne kategorije: **kvantitativni** i **kvalitativni**.



Slika 1.1 Klasifikacija tipova podataka

Kvantitativni podaci su rezultati merenja koje se može opisati brojem, a kvalitativni podaci su opisani vrednostima koje nisu numeričke, kao što je tekst. Vaša visina je podatak koji bi se mogao opisati kao kvantitativni. Međutim, opisi neke osobe kao „pušača“ ili „nepušača“ smatraju se kvalitativnim podacima.

Kvantitativni podaci mogu se dalje klasifikovati u dve potkategorije: **diskretni** i **kontinuirani**. Diskretne kvantitativne vrednosti mogu da prihvate fiksni nivo preciznosti – obično, cele brojeve. Na primer, broj operacija koje ste imali u životu je diskretna vrednost - možete imati 0, 1 ili više operacija, ali ne možete imati 1,5 operacija. Kontinuirana promenljiva je vrednost koja bi se teoretski mogla podeliti proizvoljnom količinom preciznosti. Na primer, vaša telesna masa može se opisati proizvoljnom preciznošću – 55, 55,3, 55,32 i tako dalje. U praksi, naravno, merni instrumenti ograničavaju našu preciznost. Međutim, ako vrednost može da se opiše većom preciznošću, onda se ona, u suštini, smatra kontinuiranom.



Kvalitativni podaci se, uglavnom, mogu pretvoriti u kvantitativne, a kvantitativni se takođe mogu pretvoriti u kvalitativne. To je objašnjeno kasnije u ovom poglavlju pomoću primera.

Razmislite o ovome, tako što ćete upotrebiti primer „pušača“ nasuprot primera „nepušača“. Iako možete da opišete da spadate u kategoriju „pušača“ ili „nepušača“, ove kategorije možete da zamislite i kao odgovore na izjavu da „pušite redovno“, a zatim da upotrebite Bulove vrednosti 0 i 1 da biste predstavili „tačan“ ili „netačan“ odgovor (tim redom).

Slično tome, u suprotnom smeru, kvantitativni podaci, poput visine, mogu se pretvoriti u kvalitativne. Na primer, umesto da o visini odrasle osobe razmišljate kao o broju u inčima ili santimetrima (cm), možete da je klasifikujete u grupama ljudi koji su viši od 72 inča (tj. 183 cm) u kategoriji „visok“, ljudi koji su visoki između 63 i 72 inča (tj. između 160 i 183 cm) kao „srednji“ i ljudi niži od 63 inča (tj. 152 cm) kao „niski“.

Analiza i statistika podataka

„Sirovi“ podaci su, sami po sebi, jednostavno grupa vrednosti. Međutim, u ovom obliku nisu previše zanimljivi. Tek kada počnemo da pronalazimo obrasce u podacima i da ih interpretiramo, možemo početi da radimo nešto zanimljivo – na primer, predviđanje budućnosti i identifikovanje neočekivanih promena. Ovi obrasci u podacima se zovu **informacije**. Na kraju, velika organizovana kolekcija postojanih i opsežnih informacija i iskustva koja se može koristiti za opisivanje i predviđanje pojava u stvarnom svetu naziva se **znanje**. **Analiza podataka** je proces pomoću kojeg pretvaramo podatke u informacije, a, nakon toga, u znanje. Kada se analiza podataka kombinuje sa predviđanjima, onda je to **analiza podataka**.

Dostupno je mnogo alatki koje omogućavaju razumevanje podataka. Jedna od najmoćnijih alatki u okviru alatki za analizu je primena matematike na skupovima podataka. Jedna od tih matematičkih alatki je **statistika**.

Tipovi statistike

Statistika se može dodatno podeliti na dve potkategorije: **opisnu statistiku** i **inferencijalnu statistiku**.

Opisna statistika se koristi za opisivanje podataka. Opisna statistika koja se primenjuje na jednu promenljivu u skupu podataka naziva se univarijantna analiza, dok se opisna statistika koja se istovremeno primenjuje na dve ili više promenljivih naziva **multivarijantna analiza**.

Suprotno tome, inferencijalna statistika skupove podataka smatra **uzorkom** ili malim delom merenja iz veće grupe koja se zove **populacija**. Na primer, istraživanje o učešću 10.000 birača na parlamentarnim izborima je uzorak celokupne populacije birača u nekoj zemlji. Inferencijalna statistika se koristi za pokušaj zaključivanja o svojstvima stanovništva na osnovu svojstava uzorka.



U ovoj knjizi ćemo se prvenstveno fokusirati na opisnu statistiku. Više informacija o inferencijalnoj statistici potražite u udžbeniku o statistici „Statistics“, koji su napisali David Freedman, Robert Pisani i Roger Purves.

Primer

Zamislite da ste zdravstveni analitičar i da ste dobili skup podataka sa podacima o pacijentima kao na sledećoj slici.

Year of Birth	Country of Birth	Height (cm)	Eye Color	Number of Doctor Visits in the Year 2018
1977	Egypt	182	Blue	1
1988	China	196	Hazel	2
1986	USA	180	Brown	2
1990	USA	166	Brown	1
1975	India	181	Green	3
1951	Germany	184	Brown	1
2000	Australia	174	Gray	5
1995	India	183	Brown	1
1992	China	187	Brown	2
1987	USA	169	Blue	2

Slika 1.2 Zdravstveni podaci

Kada se dobije skup podataka, često je korisno klasifikovati osnovne podatke. U ovom primeru jedinica posmatranja za skup podataka je pojedinačni pacijent, jer svaki red predstavlja pojedinačno posmatranje, tj. jedinstvenog pacijenta. Postoji 10 tačaka podataka, od kojih svaka ima pet promenljivih. Kolone **Year of Birth**, **Height** i **Number of Doctor Visits** su kvantitativne, zato što su predstavljene brojevima, a **Eye Color** i **Country of Birth** su kvalitativne.

Aktivnost 1: Klasifikacija novog skupa podataka

U ovoj aktivnosti ćemo podeliti podatke na skup podataka. Pretpostavimo da ćete uskoro započeti posao u novom gradu. Uzbudjeni ste što treba da započnete svoj novi posao, ali pre toga ste odlučili da prodate sve svoje stvari, među kojima je i vaš automobil - niste sigurni po kojoj ceni ćete ga otuđiti, pa ste odlučili da prikupite neke podatke. Pitali ste neke prijatelje koji su nedavno prodali svoje automobile po kojim cenama su to učinili. Na osnovu tih informacija, sada imate skup podataka.

Date	Make	Sales Amount (Thousands of \$)
2/1/18	Ford	12
2/2/18	Honda	15
2/2/18	Mazda	19
2/3/18	Ford	20
2/4/18	Toyota	10
2/4/18	Toyota	10
2/4/18	Mercedes	30
2/5/18	Ford	11
2/6/18	Chevy	12.5
2/6/18	Chevy	19

Slika 1.3 Podaci o prodaji polovnih automobila

Koraci koje treba da pratite:

1. Odredite jedinicu posmatranja.
2. Klasifikujte tri kolone kao kvantitativne ili kvalitativne.
3. Pretvorite kolonu **Make** u kvantitativnu kolonu podataka.



Rešenje za ovu aktivnost možete pronaći na strani 314.

METODI OPISNE STATISTIKE

Kao što je ranije pomenuto, opisna statistika jedan je od načina na koji možemo analizirati podatke da bismo ih razumeli. Univarijantna i multivarijantna analiza mogu nam dati uvid u ono što se može desiti sa određenom pojavom. U ovom odeljku ćemo detaljnije pogledati osnovne matematičke tehnike koje možemo bolje iskoristiti da bismo razumeli i opisali skup podataka.

Univarijantna analiza

Jedna od glavnih grana statistike je univarijantna analiza. Ona se koristi za razumevanje jedne promenljive u skupu podataka. U ovom odeljku ćemo pogledati neke od tehnika univarijantne analize koje se najčešće koriste.

Distribucija frekvencije podataka

Distribucija podataka je jednostavno izračunavanje broja vrednosti koje se nalaze u skupu podataka. Na primer, pretpostavimo da imamo skup podataka od 1.000 medicinskih kartona, a jedna od promenljivih u skupu podataka je boja očiju. Ako pogledamo skup podataka i ustanovimo da 700 ljudi ima smeđe, 200 ljudi zelene, a 100 ljudi plave oči, onda smo upravo opisali distribuciju skupa podataka. Konkretno, opisali smo **apsolutnu distribuciju frekvencije**. Ako brojeve ne opisujemo stvarnim brojem pojava u skupu podataka, već proporcijom ukupnog broja podataka, onda opisujemo **relativnu distribuciju frekvencije**. U prethodnom primeru boje očiju relativna frekvencija je 70 odsto smeđih, 20 odsto zelenih i 10 odsto plavih očiju.

Lako je izračunati distribuciju kada promenljiva može prihvatiti mali broj fiksnih vrednosti, kao što je boja očiju. Međutim, šta je sa kvantitativnom promenljivom koja može prihvatiti različite vrednosti, kao što je visina? Opšti način za izračunavanje distribucija za ove tipove promenljivih je kreiranje grupa intervala kojima se ove vrednosti mogu dodeliti, a zatim izračunavanje distribucije pomoću tih grupa. Na primer, visina se može raščlaniti na grupe intervala od 5 cm da bi se postigla apsolutna distribucija koja je prikazana u nastavku (pogledajte *sliku 1.6*). Zatim, možemo podeliti svaki red u tabeli ukupnim brojem tačaka podataka (tj. brojem 10.000) i dobiti relativnu distribuciju.

Još jedan koristan način na koji možete iskoristiti distribucije je da ih grafički predstavite. Sada ćemo pomoću grupa intervala kreirati **histogram**, koji je grafički prikaz kontinuirane distribucije.

Vežba 1: Kreiranje histograma

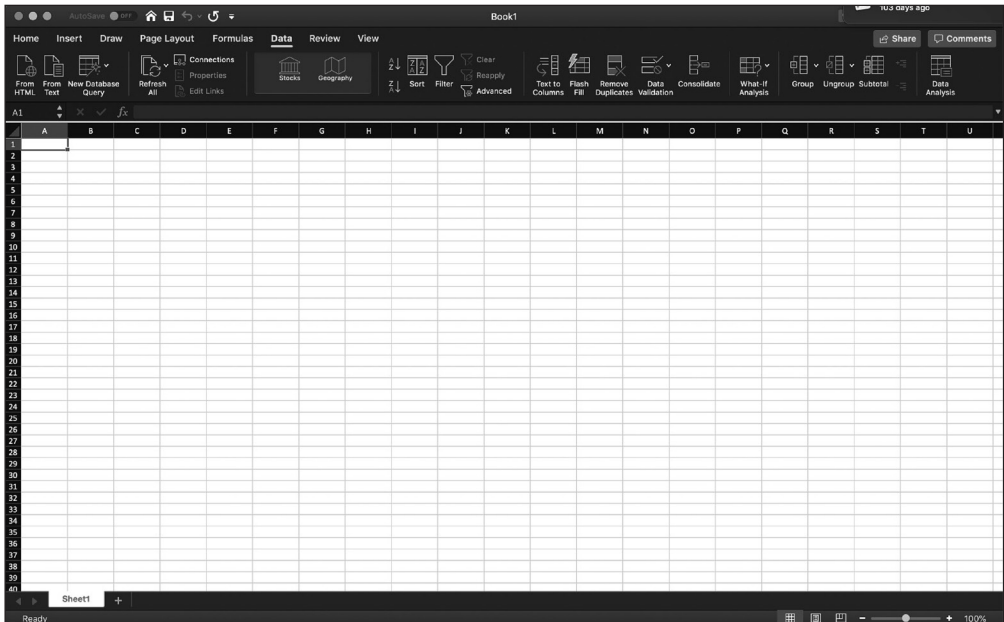
U ovoj vežbi ćemo koristiti Microsoft Excel da bismo kreirali histogram. Zamislite da kao zdravstveni analitičar želite da vidite distribuciju visine da biste primetili obrasce. Da biste izvršili ovaj zadatak, potrebno je da kreirate histogram.



Za izradu histograma možemo da koristimo softver za tabelarno izračunavanje, kao što su Excel, Python ili R. Radi praktičnosti, korišćićemo Excel. Sve skupove podataka koji se koriste u ovom poglavlju možete pronaći na GitHubu <https://github.com/TrainingByPackt/SQL-for-Data-Analytics/tree/master/Datasets>.

Izvršite sledeće korake:

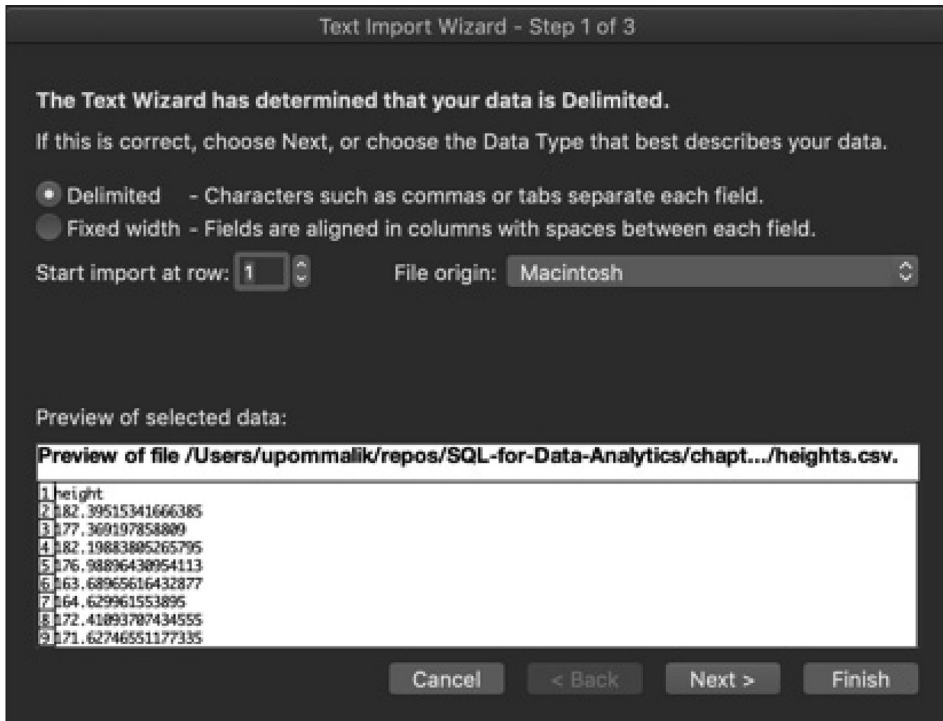
1. Otvorite Microsoft Excel praznu radnu svesku.



Slika 1.4 Prazna Excel radna sveska

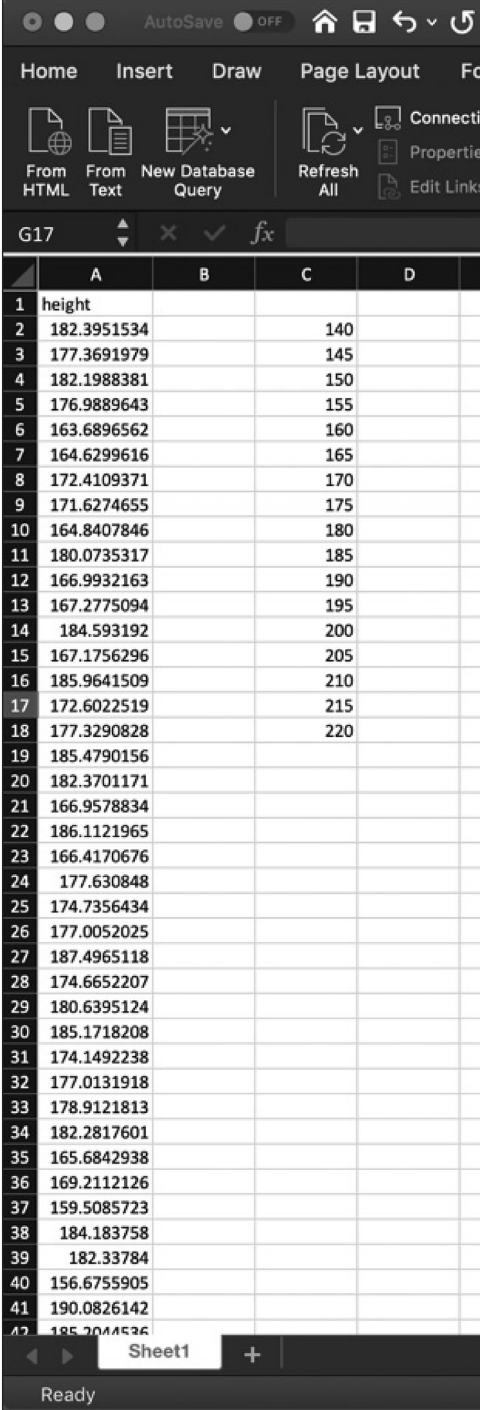
2. Otvorite karticu **Data** i kliknite na **From Text**.
3. Datoteku skupa podataka **heights.csv** možete pronaći u fascikli GitHub spremišta **Datasets**. Nakon što pristupite ovoj datoteci, kliknite na **OK**.

- Izaberite opciju **Delimited** u okviru za dijalog **Text Import Wizard** i pobrinite se da započnete uvoz u redu 1. Sada kliknite na **Next**.



Slika 1.5 Izbor opcije Delimited

- Izaberite graničnik za vašu datoteku. Pošto je ova datoteka samo jedna kolona, nema graničnika, iako CSV-ovi tradicionalno koriste zareze kao graničnike (ubuduće koristite ono što je prikladno za vaš skup podataka). Sada kliknite na **Next**.
- Izaberite **General** za **Column Data Format**, pa kliknite na **Finish**.
- U okviru za dijalog sa pitanjem **Where you want to put the data?** izaberite stavku **Existing Sheet** i ne menjajte ono što se nalazi u polju za tekst pored ove stavke. Sada kliknite na **OK**.
- U kolonu **C** upišite brojeve **140**, **145** i **150**, sa povećanjem od 5 do **220** u ćelijama **C2** do **C18**, kao što je prikazano na slici 1.6.

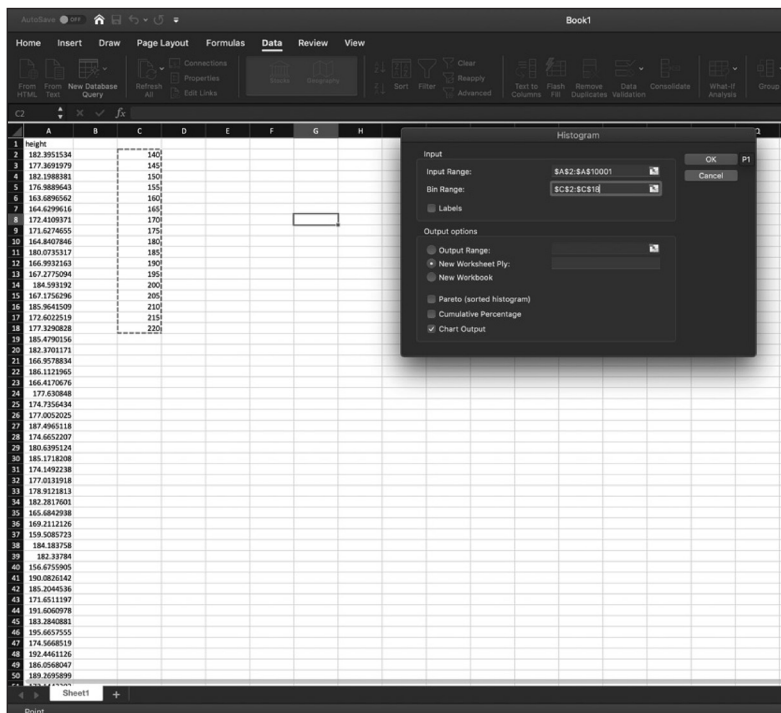


The screenshot displays the Microsoft Excel application window. The ribbon at the top shows the 'Home' tab selected, with options for 'From HTML', 'From Text', 'New Database Query', 'Refresh All', 'Connections', 'Properties', and 'Edit Links'. The formula bar shows 'G17'. The main area contains a table with the following data:

	A	B	C	D
1	height			
2	182.3951534		140	
3	177.3691979		145	
4	182.1988381		150	
5	176.9889643		155	
6	163.6896562		160	
7	164.6299616		165	
8	172.4109371		170	
9	171.6274655		175	
10	164.8407846		180	
11	180.0735317		185	
12	166.9932163		190	
13	167.2775094		195	
14	184.593192		200	
15	167.1756296		205	
16	185.9641509		210	
17	172.6022519		215	
18	177.3290828		220	
19	185.4790156			
20	182.3701171			
21	166.9578834			
22	186.1121965			
23	166.4170676			
24	177.630848			
25	174.7356434			
26	177.0052025			
27	187.4965118			
28	174.6652207			
29	180.6395124			
30	185.1718208			
31	174.1492238			
32	177.0131918			
33	178.9121813			
34	182.2817601			
35	165.6842938			
36	169.2112126			
37	159.5085723			
38	184.183758			
39	182.33784			
40	156.6755905			
41	190.0826142			
42	185.2044536			

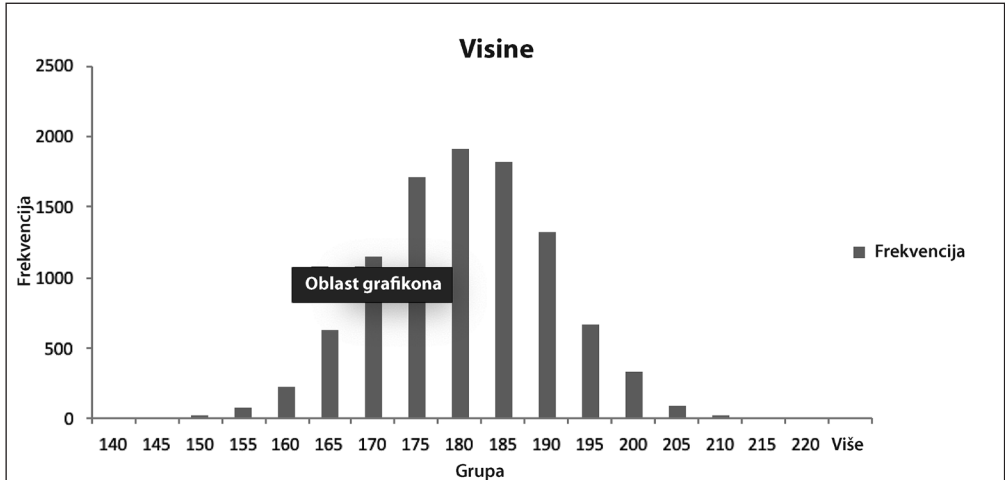
Slika 1.6 Unos podataka u Excel list

9. U kartici **Data** kliknite na **Data Analysis**. Ako ne vidite karticu **Data Analysis**, pratite ova uputstva da biste je instalirali: <https://support.office.com/en-us/article/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4>.
10. U okviru za izbor koji se prikazuje izaberite **Histogram**, pa kliknite na **OK**.
11. Za **Input Range** kliknite na dugme za izbor na krajnjoj desnoj strani polja za tekst. Trebalo bi da budete vraćeni na **Sheet1** radnu svesku, zajedno sa praznim okvirom sa dugmetom koje na sebi ima crvenu strelicu. Prevcucite i označite sve podatke u radnoj svesci **Sheet1** od **A2** do **A10001**. Sada kliknite na crveno dugme sa strelicom.
12. Za **Big Range** kliknite na dugme za izbor na krajnjoj desnoj strani polja za tekst. Trebalo bi da budete vraćeni na **Sheet1** radnu svesku, zajedno sa praznim okvirom sa dugmetom koje na sebi ima crvenu strelicu. Prevcucite i označite sve podatke u radnoj svesci **Sheet1** od **C2** do **C18**. Sada kliknite na crveno dugme sa strelicom.
13. U opciji **Output Options** izaberite **New Worksheet Ply** i pobrinite se da bude izabrana stavka **Chart Output**, kao što je prikazano na slici 1.7. Sada kliknite na **OK**.



Slika 1.7 Izbor stavke New Worksheet Ply

Kliknite na **Sheet2**. Pronađite grafikon i dva puta kliknite na naslov **Histogram**. Otkucajte reč **Heights**. Treba da kreirate grafikon koji je sličan onome na sledećem dijagramu.



Slika 1.8 Distribucija visine za muškarce

Ako pogledate oblik distribucije, možete da pronađete zanimljive obrasce. Ovde ćete primetiti simetrični zvonasti oblik ove distribucije, koja se često može naći u mnogim skupovima podataka i poznata je kao *normalna distribucija*. U ovoj knjizi neće biti previše detalja o ovoj distribuciji, ali obratite pažnju na nju u vašoj analizi podataka, jer se pojavljuje prilično često.

Kvantili

Jedan od načina za kvantitativnu distribuciju podataka je upotreba kvantila. N -kvantili su skup $n-1$ tačaka koje se koriste za podelu promenljive na n grupe. Ove tačke se često nazivaju **tačke preseka**. Na primer, kvantil četvrtog reda (koji se naziva i kvartil) je grupa od tri tačke koja deli promenljivu na četiri približno jednake grupe brojeva. Postoji nekoliko uobičajenih naziva za kvantile koji se koriste naizmenično:

N	UOBIČAJENI NAZIV
3	tercili
4	kvantili
5	kvantili
10	decili
20	ventili
100	percentili

Slika 1.9 Uobičajeni nazivi za n -kvantile

Postupak za izračunavanje kvantila se razlikuje od mesta do mesta. Sledeći postupak ćemo koristiti da bismo izračunali n -kvantile za d tačke podataka jedne promenljive:

1. Rasporedite tačke podataka od najnižih do najviših.
2. Odredite n broj n -kvantila koje želite da izračunate i broj tačaka preseka $n-1$.
3. Odredite koju k tačku preseka želite da izračunate, tj. broj od 1 do $n-1$. Kada započinjete izračunavanje, k treba da bude jednako 1.
4. Pronađite indeks i za k tačku preseka, koristeći sledeću jednačinu:

$$i = \left\lceil \frac{k}{n} (d - 1) \right\rceil + 1$$

Slika 1.10 Indeks

5. Ako je i izračunato u broju 3 ceo broj, samo izaberite tu numerisanu stavku iz raspoređenih tačaka podataka. Ako k tačka preseka nije ceo broj, pronađite numerisanu stavku koja je manja od i i onu iza nje. Pomnožite razliku između numerisane stavke i one posle nje, a zatim pomnožite decimalnim delom indeksa. Dodajte ovaj broj stavci sa najmanjim brojem.
6. Ponovite *korake od 2 do 5*, tako što ćete upotrebiti različite k vrednosti, dok ne izračunate sve tačke preseka.

Ovi koraci su malo komplikovani za razumevanje, pa ćemo sada preći na vežbu. Zahvaljujući savremenim alatcima, uključujući i SQL, računari mogu brzo izračunati kvantile pomoću ugrađenih funkcija.

Vežba 2: Izračunavanje kvartila za prodaju dodataka

Pre nego što započnete svoj novi posao, novi šef želi da pogledate neke podatke da biste imali bolji uvid u jedan od problema koji ćete rešavati, tj. sve veću prodaju dodataka i nadgradnje za kupovinu automobila. Vaš šef šalje listu 11 kupljenih automobila i koliko je potrošeno na dodatke i nadgradnju osnovnog modela novog automobila ZoomZoom Model Chi. U ovoj vežbi ćemo klasifikovati podatke i izračunati kvartile za kupovinu automobila pomoću Excela. Ovo su vrednosti za **Add-on Sales (\$)**: 5.000, 1.700, 8.200, 1.500, 3.300, 9.000, 2.000, 0, 2.300 i 4.700.



Sve skupove podataka koji su korišćeni u ovom poglavlju možete pronaći na GitHubu: <https://github.com/TrainingByPackt/SQL-for-Data-Analytics/tree/master/Datasets>.

Da biste dovršili vežbu, izvršite sledeće korake:

1. Otvorite Microsoft Excel praznu radnu svesku.
2. Otvorite karticu **Data** i kliknite na **From Text**.
3. Datoteku skupa podataka **auto_upgrades.csv** možete pronaći u fascikli GitHub spremišta **Datasets**. Pristupite ovoj datoteci i kliknite na **OK**.
4. Izaberite opciju **Delimited** u okviru za dijalog **Text Import Wizard** i pobrinite se da započnete uvoz u redu 1. Sada kliknite na **Next**.

5. Izaberite graničnik za vašu datoteku. Pošto je ova datoteka samo jedna kolona, nema graničnika, iako CSV-ovi tradicionalno koriste zarez kao graničnike (ubuduće koristite ono što je prikladno za vaš skup podataka). Sada kliknite na **Next**.
6. Izaberite **General** za **Column Data Format**, pa kliknite na **Finish**.
7. U okviru za dijalog sa pitanjem **Where you want to put the data?** izaberite stavku **Existing Sheet** i ne menjajte ono što se nalazi u polju za tekst pored nje. Sada kliknite na **OK**.
8. Kliknite na ćeliju **A1**. Zatim, kliknite na karticu **Data**, pa na **Sort** iz kartice.
9. Otvoriće se sortirani okvir za dijalog. Sada kliknite na **OK**. Vrednosti će biti sortirane od najnižih do najviših. U listi na *slici 1.11* prikazane su sortirane vrednosti.

	A
1	Add-on Sales (\$)
2	0
3	0
4	1500
5	1700
6	2000
7	2300
8	3300
9	4700
10	5000
11	8200
12	9000
13	

Slika 1.11 Sortirani podaci o prodaji dodataka

10. Sada odredite broj n -kvantila i tačaka preseka koje treba da izračunate. Kvantili su ekvivalentni broju 4, kao što možete videti na *slici 1.9*. Pošto je

broj tačaka preseka samo za 1 manji od broja n -kvantila, znamo da postoje tri tačke preseka.

11. Izračunajte indeks za prvu tačku preseka. U ovom primeru je $k=1$, d , broj tačaka podataka, jednako je 10, a n , broj n -kvantila, jednako je 4. Kada ovo uključimo u jednačinu sa slike 1.12, dobićemo 3.5.
12. Pošto indeks 3.5 nije ceo broj, prvo ćete pronaći treću i četvrtu stavku, tj. 1.500 i 1.700 (tim redom). Pronađite razliku između njih, koja je 200, a zatim to pomnožite sa decimalnim delom 0,5, pri čemu ćete dobiti 100. Broj 100 dodajte trećoj numerisanoj stavci 1.500 i dobijate 1.600.
13. Ponovite korake od 2 do 5 za $k = 2$ i $k = 4$ da biste izračunali drugi i treći kvartil. Trebalo bi da dobijete 2.300 i 4.850 (tim redom).

$$\begin{aligned}i &= \left[\frac{k}{n} (d - 1) \right] + 1 \\i &= \left[\frac{1}{4} (11 - 1) \right] + 1 \\i &= \frac{10}{4} + 1 \\i &= \frac{10}{4} + 1 \\i &= 2.5 + 1 = 3.5\end{aligned}$$

Slika 1.12 Izračunavanje indeksa za prvu tačku preseka

U ovoj vežbi ste naučili kako se klasifikuju podaci i izračunavaju kvartili pomoću Excela.

Centralna tendencija

Jedno od uobičajenih pitanja koja se postavljaju o promenljivoj u skupu podataka odnosi se na tipičnu vrednost za ovu promenljivu. Ova vrednost se često opisuje kao centralna tendencija promenljive. Postoji mnogo izračunatih brojeva iz skupa podataka koji se često koriste za opisivanje njegove centralne tendencije, a svaki od njih ima svoje prednosti i mane. Neki od načina za merenje centralne tendencije uključuju sledeće:

- **modus** – Modus je vrednost koja se najčešće pojavljuje u distribuciji promenljive. U primeru boje očiju na slici 1.2 modus bi bio „smeđe oči“, jer se najčešće pojavljuje u skupu podataka. Ako je za najčešću promenljivu vezano više vrednosti, ona se naziva **multimodalna** i prijavljuju se sve najviše vrednosti. Ako se ni jedna vrednost ne ponavlja, ne postoji modus za te skupove vrednosti. Modus je koristan kada promenljiva može prihvatiti mali fiksni broj vrednosti. Međutim, problematično je izvršiti proračun kada je promenljiva kontinuirana kvantitativna promenljiva, kao u našem problemu koji se odnosi na visinu. Zahvaljujući ovim promenljivim, drugi proračuni su pogodniji za utvrđivanje centralne tendencije.
- **prosek/sredina** – Prosek promenljive (koji se naziva i **sredina**) je vrednost koja je izračunata kada se zbir svih vrednosti promenljive podeli sa brojem podataka. Na primer, recimo da ste imali mali skup podataka starosne dobi: 26, 25, 31, 35 i 29. Prosek ovih godina starosti bi bio 29,2, jer to je broj koji dobijete kada saberete pet brojeva, a zatim ih podelite brojem 5, tj. brojem podataka. Sredinu je lako izračunati, a ona generalno dobro opisuje „tipične“ vrednosti za promenljivu, pa nije čudo što je jedna od najčešće objavljivanih opisnih statistika u literaturi. Međutim, prosek, kao centralna tendencija, ima jednu veliku manu – osetljiv je na **neuobičajene vrednosti**. Neuobičajene vrednosti su podaci koji se značajno razlikuju u odnosu na ostale podatke i pojavljuju se veoma retko. Ove vrednosti se često mogu identifikovati pomoću grafičkih tehnika (kao što su dijagrami rasipanja i kutijasti dijagrami) i identifikovanjem tačaka podataka koje su veoma udaljene od ostalih podataka. Kada skup podataka ima neuobičajene vrednosti, on se naziva „**iskrivljeni**“ skup podataka. Neki uobičajeni razlozi zbog kojih se pojavljuju neuobičajene vrednosti obuhvataju „nečiste“ podatke, izuzetno retke događaje i probleme u vezi sa mernim instrumentima. Neuobičajene vrednosti često „iskrive“ prosek do te mere da više ne mogu da predstavljaju tipičnu vrednost u podacima.

- **medijana** – Medijana (koja se naziva i drugi kvartil i pedeseti percentil) je vrsta čudne mere centralne tendencije, ali ima neke ozbiljne prednosti u odnosu na prosek. Da biste je izračunali, upotrebite brojeve za promenljivu i sortirajte ih od najnižeg do najvišeg, a zatim odredite srednji broj. Za neparan broj podataka srednji broj je jednostavno srednja vrednost raspoređenih podataka. Ako postoji parni broj tačaka podataka, upotrebite prosek dva srednja broja.

Iako je medijanu malo teže izračunati, na nju manje utiču neuobičajene vrednosti nego na sredinu. Da bismo ilustrovali ovu činjenicu, izračunaćemo medijanu „iskrivljenog“ skupa podataka starosne dobi od 26, 25, 31, 35, 29 i 82. Ovoga puta, kada izračunamo sredinu skupa podataka, dobićemo vrednost 30, koja je mnogo bliža tipičnoj vrednosti skupa podataka nego prosek 38. Ova otpornost na neuobičajene vrednosti je jedan od glavnih razloga zbog kojeg se izračunava medijana.

Kao opšte pravilo, dobra je ideja izračunati sredinu i medijanu promenljive. Ako postoji značajna razlika u vrednosti proseka i medijane, skup podataka može imati neuobičajene vrednosti.

Vežba 3 Izračunavanje centralne tendencije prodaje dodataka

U ovoj vežbi ćemo izračunati centralnu tendenciju konkretnih podataka. Da biste bolje razumeli podatke **Add-on Sales**, treba da razumete šta je tipična vrednost za ovu promenljivu. Sada ćemo izračunati modus, sredinu i medijanu podataka **Add-on Sales**. Ovo su podaci za 11 kupljenih automobila: 5.000, 1.700, 8.200, 1.500, 3.300, 9.000, 2.000, 0, 0, 2.300 i 4.700.

Da biste implementirali vežbu, izvršite sledeće korake:

1. Da biste izračunali modus, pronađite vrednost koja se najčešće pojavljuje. S obzirom da je 0 vrednost koja se najčešće pojavljuje u skupu podataka, modus je 0.
2. Da biste izračunali sredinu, saberite brojeve u podacima **Add-on Sales**, a rezultat bi trebalo da bude 37.700. Zatim, podelite zbir sa brojem vrednosti 11 i dobićete srednju vrednost od 3.427,27.

3. Na kraju, izračunajte medijanu, tako što ćete sortirati podatke, kao što je prikazano na slici 1.13.

	A
1	Add-on Sales (\$)
2	0
3	0
4	1500
5	1700
6	2000
7	2300
8	3300
9	4700
10	5000
11	8200
12	9000
13	

Slika 1.13 Sortirani podaci o prodaji dodataka

Utvrđite srednju vrednost. Pošto postoji 11 vrednosti, srednja vrednost će biti šesta na listi. Sada uzimamo šesti elemenat u raspoređenim podacima i dobijamo medijanu od 2.300.



Kada uporedimo sredinu i medijanu, videćemo da postoji značajna razlika između te dve vrednosti. Kao što je ranije pomenuto, to je znak da u našem skupu podataka imamo neuobičajene vrednosti. U narednim odeljcima ćemo razmatrati kako se utvrđuje koje vrednosti su neuobičajene.

Disperzija

Još jedno svojstvo koje je značajno za skup podataka je otkrivanje koliko su tačke podataka u promenljivoj blizu jedne druge. Na primer, skupovi brojeva [100, 100, 100] i [50, 100, 150] imaju srednju vrednost 100, ali brojevi u drugoj grupi su više rasuti od brojeva u prvoj. Ovo svojstvo koje opisuje kako se podaci šire zove se **disperzija**.

Postoji mnogo načina za merenje disperzije promenljive. Ovo su neki od najčešćih načina za procenu disperzije:

- **opseg** – Opseg je jednostavno razlika između najviše i najniže vrednosti promenljive. On se veoma jednostavno izračunava, ali je vrlo osetljiv na neuobičajene vrednosti. Osim toga, ne obezbeđuje mnogo informacija o širenju vrednosti u sredini skupa podataka.
- **standardna devijacija/varijansa** – Standardna devijacija je jednostavno kvadratni koren proseka kvadrirane razlike između svake tačke podataka i sredine. Vrednost standardne devijacije je u opsegu od 0 do pozitivne beskonačnosti. Što je standardna devijacija bliža 0, brojevi u skupu podataka se manje razlikuju. Ako je standardna devijacija 0, to znači da su sve vrednosti promenljive skupa podataka iste.

Suptilna razlika koju treba posebno napomenuti je da postoje dve različite formule za standardnu devijaciju koje su prikazane na *slici 1.14*. Kada skup podataka predstavlja čitavu populaciju, trebalo bi da izračunate standardnu devijaciju populacije, koristeći formulu A sa *slike 1.14*. Ako vaš uzorak predstavlja deo opservacija, trebalo bi da koristite formulu B za standardnu devijaciju uzorka, kao što je prikazano na *slici 1.14*. Kada ste u nedoumici, koristite varijansu uzorka, jer se ona smatra konzervativnijom. Razlika između dve formule u praksi je vrlo mala kada postoji mnogo tačaka podataka.

Standardna devijacija je, obično, kvantitet koji se najčešće koristi za opisivanje disperzije. Međutim, baš kao i na opseg, na nju mogu uticati neuobičajene vrednosti, iako ne toliko ekstremno kao na opseg. Standardna devijacija se, baš kao i opseg, može uključiti u proračun. Međutim, savremene alatke obično olakšavaju izračunavanje standardne devijacije.

Treba napomenuti i da ćete povremeno možda videti navedenu povezanu vrednost, tj. varijansu. Ovaj kvantitet je jednostavno kvadrat standardne devijacije.

$$A) \sqrt{\frac{\sum_{i=1}^n (x_i - u_x)^2}{n}} \quad B) \sqrt{\frac{\sum_{i=1}^n (x_i - u_x)^2}{n-1}}$$

Slika 1.14 Formule standardne devijacije za A) populaciju i B) uzorak

- **Interkvartilni opseg (IQR, Interquartile Range)** – Interkvartilni opseg je razlika između prvog kvartila Q1 (koji se naziva i donji kvartil) i trećeg kvartila Q3 (koji se naziva i gornji kvartil).



Više informacija o izračunavanju kvantila i kvartila potražite u odeljku „Distribucija podataka“ u ovom poglavlju.

Za razliku od opsega i standardne devijacije, IQR je otporniji na neuobičajene vrednosti, ali, mada je jedna od najsloženijih funkcija za izračunavanje, obezbeđuje robusniji način za merenje širenja skupa podataka. U stvari, IQR se često koristi za definisanje neuobičajenih vrednosti. Ako je vrednost u skupu podataka manja od $Q1 - 1,5 \times IQR$ ili veća od $Q3 + 1,5 \times IQR$, vrednost se smatra neobičajenom.

Vežba 4: Disperzija prodaje dodataka

Da biste bolje razumeli prodaju dodataka i nadgradnji, potrebno je da bolje pogledate raširenost podataka. U ovoj vežbi ćemo izračunati opseg, standardnu devijaciju, IQR i neuobičajene vrednosti prodaje dodataka (**Add-on Sales**). Ovo su podaci za 11 kupljenih automobila: 5.000, 1.700, 8.200, 1.500, 3.300, 9.000, 2.000, 0, 0, 2.300 i 4.700.

Da biste uradili vežbu, pratite sledeće korake:

1. Da biste izračunali opseg, treba da pronađete najmanju vrednost podataka 0 i da je oduzmete od maksimalne vrednosti podataka 9.000, pri čemu ćete dobiti 9.000.
2. Za izračunavanje standardne devijacije treba da uradite sledeće: da utvrdite da li želite da izračunate standardnu devijaciju uzorka ili standardnu devijaciju populacije. Pošto tih 11 tačaka podataka predstavljaju samo mali deo svih kupovina, treba da izračunate standardnu devijaciju uzorka.
3. Zatim, pronađite srednju vrednost skupa podataka (izračunatu u „Vežbi 2: Izračunavanje kvartila za prodaju dodataka“), koja treba da bude 3.427,27.

4. Sada oduzmite svaku tačku podataka od srednje vrednosti i kvadrirajte rezultat. Rezultati su sumirani na sledećem dijagramu.

Add-on Sales (\$)	Difference with Mean	Difference with Mean Squared
5000	1572.727273	2473471.074
1700	-1727.272727	2983471.074
8200	4772.727273	22778925.62
1500	-1927.272727	3714380.165
3300	-127.2727273	16198.34711
9000	5572.727273	31055289.26
2000	-1427.272727	2037107.438
0	-3427.272727	11746198.35
0	-3427.272727	11746198.35
2300	-1127.272727	1270743.802
4700	1272.727273	1619834.711

Slika 1.15 Zbir izračunavanja kvadrata

5. Saberite razlike vrednosti **Differences with Mean Squared**, pri čemu ćete dobiti rezultat 91.441.818,6.
6. Podelite zbir brojem tačaka podataka, koji je u ovom primeru 10, i oduzmite 1, a zatim upotrebite njegov kvadratni koren. Rezultat ovog proračuna treba da je 3.023,93 kao standardna devijacija uzorka.
7. Da biste izračunali IQR, pronađite prvi i treći kvartil. Izračunavanje kvartila možete videti u „*Vežbi 2: Izračunavanje kvartila za prodaju dodataka*“, a rezultati su 1.600 i 4.850. Zatim, oduzmite dva da biste dobili vrednost 3.250.