FAKULTET TEHNIČKIH NAUKA

UNIVERZITET U NOVOM SADU

Dokumentacija projekta

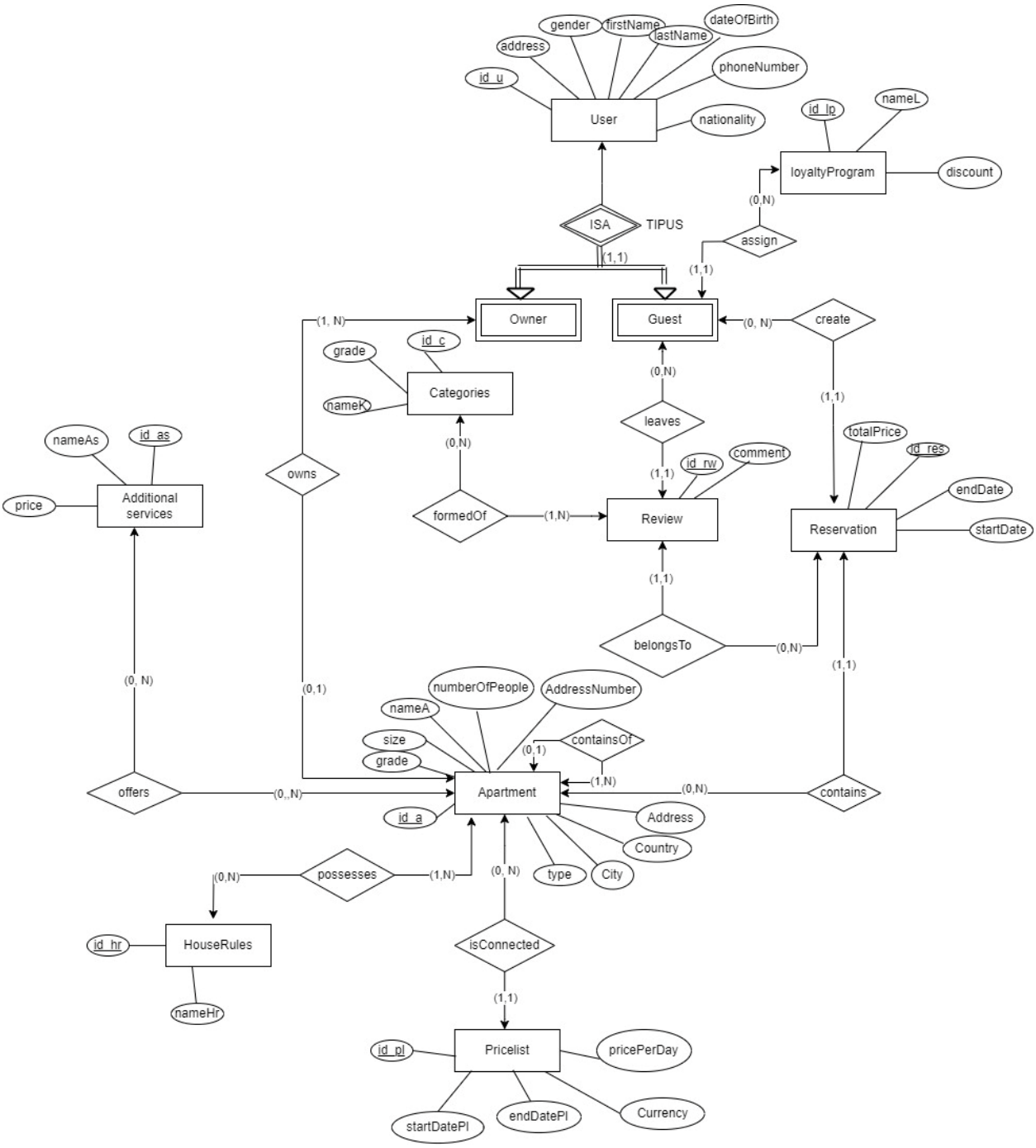
Student: Bogdan Blagojević Indeks: E2 71/2023

Predmet: Sistemi skladišta podataka

Februar 2024.

Sadržaj

1.Opis ER modela



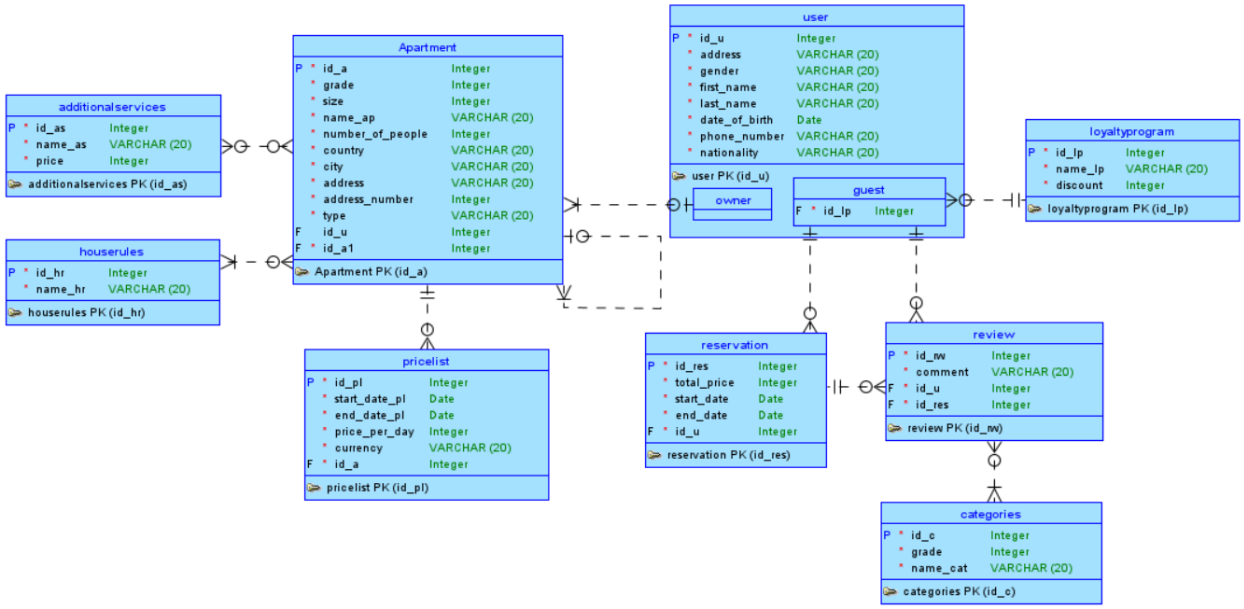
Slika 1. ER Model

- Korisnik može biti ili vlasnik smeštaja ili gost
- Vlasnik smeštaja poseduje bar jedan apartman
- Apartman pripada najviše jednom vlasniku
- Gostu smeštaja je dodeljen tačno jedan program lojalnosti
- Program lojalnosti može biti dodeljen više gostiju, a ne mora ni jednom
- Gost može da kreira više rezervacija, a ne mora ni jednu
- Rezervaciju je kreirao tačno jedan gost
- Gost može da ostavi više komentara, a ne mora da ostavi ni jedan
- Komentar je ostavljen od strane tačno jednog gosta smeštaja
- Komentar se sastoji od bar jedne kategorije
- Neka kategorija može da sačinjava više komentara, a ne mora ni jedan
- Komentar pripada tačno jednoj rezervaciji
- Rezervaciji može da pripada više komentara, a ne mora ni jedan
- Rezervacija sadrži tačno jedan apartman
- Apartman može biti deo više rezervacija, a ne mora ni jedne
- Apartman se sastoji od bar jedne smeštajne jedinice
- Smeštajna jedinica pripada najviše jednom apartmanu, a ne mora ni jednom
- Apartman može biti povezan sa više cenovnika, a ne mora ni sa jednim
- Cenovnik je povezan za tačno jedan apartman
- Apartman poseduje bar jedno pravilo kućnog reda
- Pravilo kućnog reda može da pripada više apartmana, a ne mora ni jednom
- Apartman može da nudi više dodatnih usluga, a ne mora ni jednu
- Dodatna usluga može biti nuđena u vise apartmana, a ne mora ni u jednom

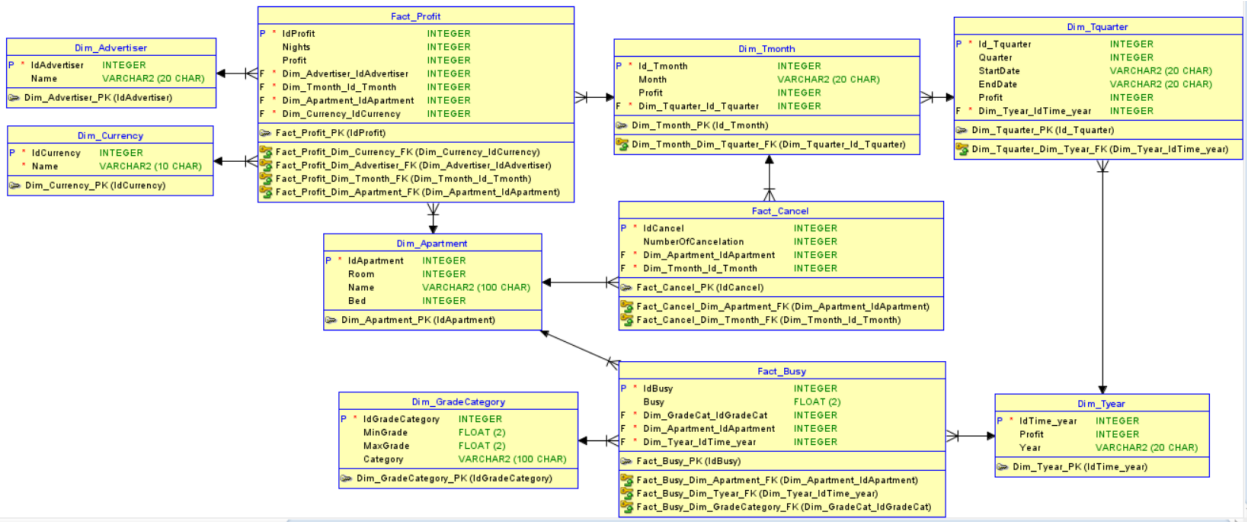
Skup obečežja

Mnemonik	Pun opis
Id_u	Identifikaciona oznaka korisnika
Address	Adresa korisnika
Gender	Pol korisnika
FirstName	Ime korisnika
LastName	Prezime korisnika
dateOfBirth	Datum rođenja korisnika
phoneNumber	Broj telefona korisnika
Nationality	Državljansto korisnika
TIPUS	Vrsta korisnika
Id_lp	Identifikaciona oznaka programa lojalnosti
NameL	Naziv programa lojalnosti
Discount	Popust koji donosi određeni program loyalnosti
Id_rw	Identifikaciona oznaka komentara
Comment	Tekstualni dodatak komentaru
Id_c	Identifikaciona oznaka categorije
Grade	Ocena kategorije
NameK	Naziv kategorije
Id_res	Identifikaciona oznaka rezervacije
totalPrice	Ukupna cena rezervacije
startDate	Datum pocetka rezervacije
endDate	Datum završetka rezervacije
Id_a	Identifikaciona oznaka apartmana
Grade	Ocena apartmana
Size	Veličina apartmana
nameA	Naziv apartmana
numberOfPeople	Maksimalan broj osoba u apartmanu
Country	Drzava u kojoj se nalati apartman
City	Grad u kojem se nalazi apartman
Address	Adresa apartmana
addressNumber	Kućni broj apartmana
Type	Tip apartmana
Id_pl	Identifikaciona oznaka cenovnika
pricePerDay	Cena noćenja
startDatePl	Datum pocetka važenja cenovnika
endDatePl	Datum završetka važenja cenovnika
Id_hr	Identifikaciona oznaka pravila kućnog reda
nameHr	Naziv pravila kućnog reda
Id_as	Identifikaciona oznaka dodatne usluge
nameAs	Naziv dodatne usluge
Price	Cena dodatne usluge

2.Opis OLTP šeme baze podataka



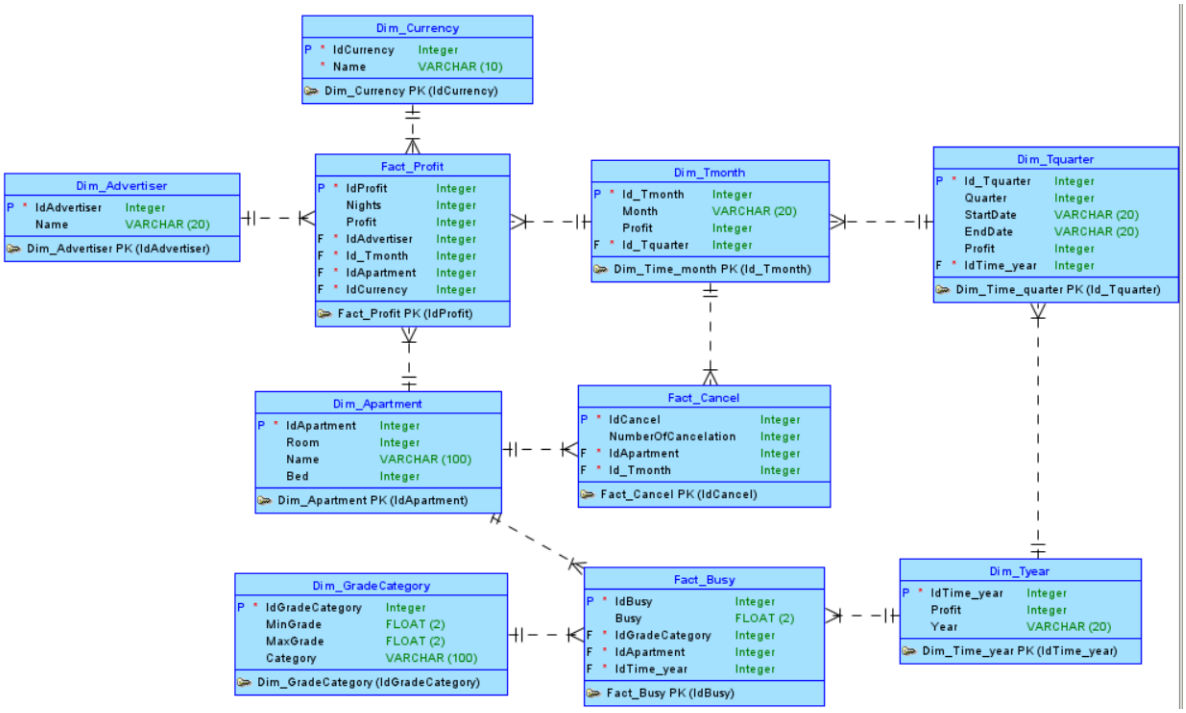
Slika 2. Logička šema OLTP baze podataka



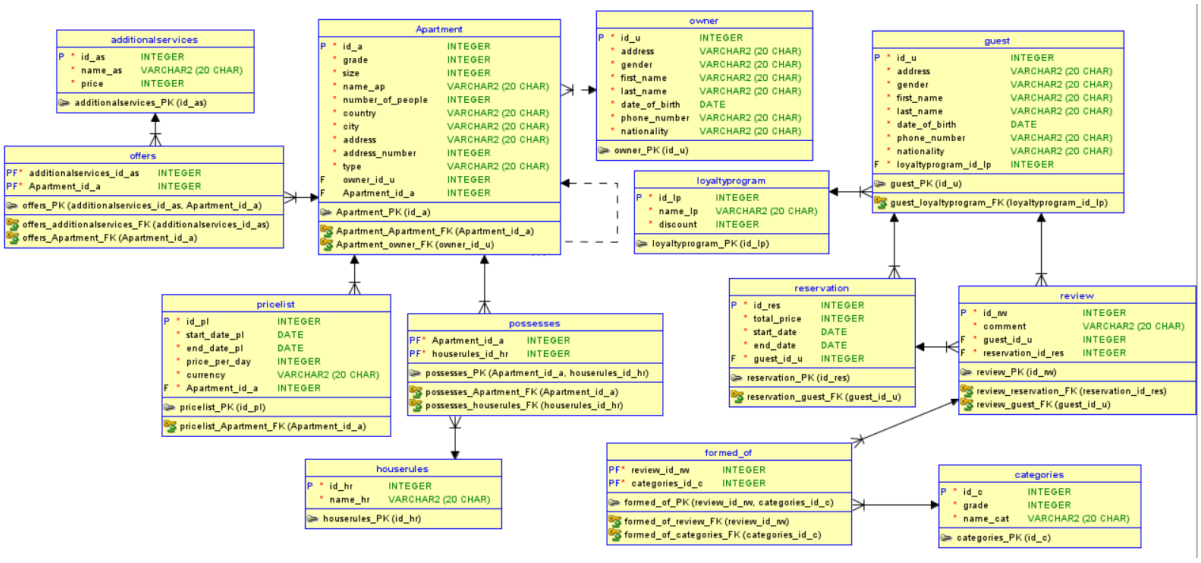
Slika 3. Relaciona šema OLTP baze podataka

3.Opis data warehouse šeme baze podataka

Data warehouse šema baze podataka sačinjena je od tri činjenične tabele i 6 dimenzionih tabela. Dimenzija vreme predstavlja normalizovanu hijerarhijsku strukturu dok je dimencija ocena zapravo *Bracketing* dimenziju.



Slika 4. Logički šema data warehouse baze podatka



Slika 5. Relaciona šema data warehouse baze podataka

3.1 Teme

1. Praćenje priliva novca i broja noćenja za svaki apartman u odnosnu na način oglašivanja
2. Analiza zauzetosti apartmana kao i kategoriji ocene kojoj pripada na godišnjem novou
3. Praćenje broja otkazanih rezervacija za svaki apartman na mesečnom nivou

3.2 Bus Matrix

	Fact_Profit	Fact_Cancel	Fact_Busy
Dim_Advertiser	X		
Dim_Currency	X		
Dim_Apartment	X	X	X
Dim_Tmonth	X	X	
Dim_Tquarter			
Dim_Tyear			X
Dim_GradeCategory			X

Tabela 1. Prikaz Bus Matrix

3.3 Dimenzione tabele

- **Dim_Advertiser:** Sadrži naziv oglašivača koji može biti: *Booking*, *Airbnb* ili *OLTP* (odnosno applikacija za izdavanje smeštaja). Jedinstveno se identifikuje preko svog primarnog ključa IdAdvertiser.
- **Dim_Currency:** Predstavlja valutu koja može biti RSD, DOLLAR ili EUR, dok se jedinstveno identifikuje preko primarnog ključa IdCurrency.
- **Dim_Apartment:** Sadrži naziv apartmana kao i broj soba i broj kreveta svakog pojedinačnog apartmana koji se jedinstevno identifikuju preko primarnog ključa IdApartment.
- **Dim_Tmonth:** Predstavlja vremensku dimenziju koja sadrži agregirani podatak *Profit.* Takođe porede identifikacionog broja sadrži i naziv meseca kao i strain ključ ka kvartalu kom pripada.
- **Dim_Tquarter:** Vremenska dimenzija koja sadrži agregirani atribut *Profit* kao i strani ključ ka godini kojoj pripada. Dodatno, sadrži I mesec početka i završetka svakog kvartala kao I identifikacioni broj.
- **Dim_Tyear:** Vremenska dimenzija koja sadrži agregirani atribut *Profit,* naziv meseca kao i identifikacioni broj.
- **Dim_GradeCategory:** *Bracketing* dimenzija koja pored identifikacionog broja sadži gornju I donju granicu za svaku kategoriju ocene koja može biti: *Very Poor, Poor, Passable, Good, Superb* ili *NotGraded*.

3.4 Činjenične tabele

- **Fact_Profit:** Sadrži profit i broj noćenja po mesecu, apartmanu, oglašivaču i valuti kojom oglašivač posluje. Sadrži i polje IdProfit po kom se jedinstveno identifikuje.
- **Fact_Cancel:** Pruža uvid u broj otkazanih rezervacija za svaki apartman u bilo kom mesecu, te sadrži i polje IdCancel koje predsavlja identifikacioni broj.
- **Fact_Busy:** Procentualno prikazana zauzetost svakog apartmana godišnje kao i kategoričkoj oceni kojoj pripada. Polje IdBusy služi kao jedinstveni identifikatior.

4.Opis ECTL procesa

U folderu *data* se nalaze dva foldera: *processed* i *row*. U *row* folderu se nalaze neobrađeni podaci strukturiratni po godinama, dok se u *processed* folderu nalaze sledeći csv fajlovi: *ProcessedAirbnb* ,*ProcessedBooking* i *ProcessedData*. U python fajlovima *ProcessingAirbnb*, *ProcessingBooking*, *ProcessingData* i *TLData* se nalazi logika obrađivanja podataka. Pre svega vršeno je učitavanje excel fajlova i nakon njihove obrade vršeno je upisivanje u csv fajlove.

Procesiranje podataka generisanih od strane *Booking* oglašivača vršeno je na sledeći način. Pre svega dodata je kolona *Advertiser* kako bi se znalo koja je vrsta oglašivača. Zatim su izbačeni redovi koji imaju istu vrednost za *Book Number.* Nedostajuće vrednosti u kolonama *Guest Name(s)* i *Booked by* su zamenje na praznim prostorom te su modifikovane vrednosti tako što svaka reč počinje velikim početnim slovom. Format datuma je izmenjen u nam pregledniji (DD.MM.YYYY) i ostavljeni su samo validni datumi te je vršeno njihovo preimenovanje, kao i preimenovanje vrednosti za kolonu Status. Dodata je kolona *Adults* koja predstavlja razliku kolone *People* i *Children.* Iz kolona *Commission Amount* i *Price* uklonjene su vrednosti EUR. Vršeno je i preimenovanje kolone *Accommodation* te su odbačene sve nevalidne vrednosti. Vršeno je i preimenovanje kolona *Guest Name(s)* u *Guest Name*, *People* u *NumberOfPeople*, *Adults* u *NumberOfAdults*, *Children* u *NumberOfChildren*, *Children's age(s)* u *Children's age*, *Booked by* u *Booked By*, *Booked on* u *Booked On*. Nakon uočenih pravilnosti među godinama napravljena je funkcija koja ponavlja isti posao. Dodatno za 2021 i 2022 godinu, kolona *Occypancy* poslužila je kako bi se kreirala kolona *NumberOfChildren* i *Children's age* te je nakon toga odbačena. Nakon obrađenih svih excel fajlova vezanih za *Booking* vršeno je njihovo spajanje u jedan csv fajl pod nazivom *ProcessedBooking* te je dodatno vršeno I preimenovanje kolona kao I odbacivanje redova koji za kolonu *Price* imaju vrednost manju od nule.

Procesiranje podataka generisanih od strane *Airbnb* oglašivača vršeno je na sledeći način. Na samom početku dodata je kolona *Advertiser* kako bi se znalo koja je vrsta oglašivača. Vršena je izmena formata datuma u DD.MM.YYYY kao i odbacivanje nevalidnih datuma. Takođe odbačeni su i redovi čija se vrednost za kolonu *Confirmation code ponavlja.* Kolona *Guest Name* obrađena je kao i kod *Booking* oglašivača. Kolona *# of infants* te pridodata na kolonu *# of children* te je odbačena. Nedostajuće vrednosti u koloni *Grade* popunjene su sa nula isto kao i kod *Booking* oglašivača. Vršeno je i kolona: *Confirmation code* u *Book Number*, *Guest name* u *Guest Name*, *# of adults* u *NumberOfAdults*, *# of children* u *NumberOfChildren*, *Start date u Start Date*, *End date u End Date*, *# of nights* u *NumberOfNights*, *Booked* u *Booked On*, *Listing* u *Accommodation*, *Listing name* u *Accommodation Type*, *# of people* u *NumberOfPeople*, *Earnings* u *Price_DOLLAR*. Nakon toga napravljena je i funkcija koja radi posao ispravljanje istih nepravilnosti među godinama. Dodatno za 2021 godinu vršena je izmena Statusa: *Arriving in 49 days, Arring in 82 days, Arriving tomorrow i*

Arriving in 13 days u *Arriving* . Nakon obrađenih svih godina vršeno je spajanje njihovo spajanje i odbacivanje redova koji poseduju vrednosti za *Price_DOLLAR* manje od nule. Na samom kraju vršeno je upisivanje u csv fajl pod nazivom *ProcessedAirbnb*.

Python fajl *ProccessingData* namenjen je da grupiše podatke o *Airbnb* i *Booking* oglašivačima u jedan fajl. Na samom početku vrši učivatanje iz csv fajlova i preimenovanje kolona. Nakon toga odbacuju se redovi koji imaju istu vrednost za kolonu *Book Number* i dodaje se kolona *Bed* koja predstavlja koliko svaki apartman ima kreveta. Kao poslednji korak vrši se upisivanje u csv fajl pod nazivom *ProcessedData*.

Transformacija i učitavanje podataka u *data warehouse* vršeno je u *TLData* python fajlu. Kao prvi korak vršeno je učitavanje *ProcessedData* fajla i kreirana je konekcija sa *Oracle* bazom podataka kroz dve funkcije za čitanje i pisanje. Podaci o nazivu oglašivača prikupljeni su iz kolone *Advertiser* i dodat je oglašivač OLTP koji predstaša aplikaciju za izdavanje smeštaja. Podaci o dimenziji apartman dobijeni su iz kolona *Room, Accommodation i Bed* iz procesiranog skupa podataka. Naziv unutar dimezije valuta generisan je uz pomoć naziva oglašivača kao i aplikacije za izdavanje smeštaja. *Bracketing* dimenzija kategorije ocene kreiran je tako što su donju granicu uzete ocene redom 1,3,5,7,9 dok za gornju 2,4,6,8,10 te im je dodeljen i naziv. Vremenska dimenzija godina tako što je posmatrano koje se jedinstvene godine pojavljuju u OLTP bazi kao i procesiranom fajlu. Dodatno je izračunat profit na nivou svake godine ponaosob. Dimenzija kvartal je kreirana tako što je za su za početak uzete vrednosti januar,april. jul, oktobar, dok za kraj mart, jun, septembar i decembar. Izračunar je profit na nivou svakog kvartala i dodeljen je godini kojoj pripada. Na sličan način kao i kod dimenzije kvartal i godina računata je i dimencija mesec, odnosno posmatrajući podatke iz OLTP kao i iz procesiranog fajla. Tabela činjenica *Busy* kreirana je uz pomoć dimenzija apartman, kategorije ocena kao i godine, dok su procesirani podaci i OLTP baza podataka služili kako bi se računao procenat zauzetnosti apartmana. U konačan procenat nisu ulazile rezervacije koje nisu bile ocenjene. Strani ključevi u tabeli činjecina *Cancel* prikupljeni su iz dimenzije apartman i vremenske dimenzije mesec te je broj otkaza računat za svaki mesec i svaki apartkam ponaosob. U činjeničnoj tabeli *Profit* strani ključ valute određen je uz pomoć oglašivača. Dodatno sadrži strane ključeve prema mesecu i apartmanu. Profit je računat u odnosu na valutu u kojoj ogređeni ograšivač posluje, dok je broj noćenja računat kao razlika *Start Date i End Date.*

5.Testiranje

5.1 Dimenzije

	IDADVERTISER	NAME
1	1	Booking
2	2	Airbnb
3	3	OLTP

Slika 6. Dimenzija oglašivač

	IDCURRENCY	NAME
1	1	RSD
2	2	EUR
3	3	DOLLAR

Slika 7. Dimenzija valuta

	IDAPARTMENT	ROOM	NAME	BED
1	1	4	One-Bedroom Apartment with Balcony and Sea View	4
2	2	2	Studio with Patio and Sea View	2
3	3	4	One-Bedroom Apartment with Patio and Sea View	4
4	4	2	Studio with Patio	2

Slika 7. Dimenzija apartman

	IDGRADECATEGORY	MINGRADE	MAXGRADE	CATEGORY
1	1	1	2	Very poor
2	2	3	4	Poor
3	3	5	6	Passable
4	4	7	8	Good
5	5	9	10	Superb
6	6	0	0	NotGraded

Slika 8. Dimenzija kategorije ocene

	IDTIME_YEAR	YEAR	PROFIT
1	1	2019	3531291
2	2	2020	595859
3	3	2021	1255456
4	4	2022	2245116
5	5	2023	4460667

Slika 9. Dimenzija godina

	ID_TMONTH	MONTH	DIM_TQUARTER_ID_TQUARTER	PROFIT
1	1	April	1	240802
2	2	May	1	392109
3	3	June	1	569326
4	4	July	2	726492
5	5	August	2	736090
6	6	September	2	618979
7	7	October	3	247494
8	8	June	4	13772
9	9	July	5	114532
10	10	August	5	340627
11	11	September	5	126928
12	12	February	6	12000
13	13	June	7	103484
14	14	July	8	567118
15	15	August	8	490184
16	16	September	8	82671
17	17	January	9	3000
18	18	March	9	7000
19	19	May	10	99236
20	20	June	10	515696
21	21	July	11	663456
22	22	August	11	406736
23	23	September	11	469523
24	24	October	12	80471
25	25	January	13	5000
26	26	January	13	9000
27	27	April	14	250896
28	28	May	14	451333
29	29	June	14	689530
30	30	July	15	971796
31	31	August	15	1072929
32	32	September	15	839590
33	33	October	16	161592
34	34	December	16	9000

Slika 10. Dimenzija mesec

	ID_TQUARTER	QUARTER	DIM_TYEAR_IDTIME_YEAR	STARTDATE	ENDDATE	PROFIT
1	1	2	1	April	June	1202237
2	2	3	1	July	September	2081561
3	3	4	1	October	December	247494
4	4	2	2	April	June	13772
5	5	3	2	July	September	582087
6	6	1	3	January	March	12000
7	7	2	3	April	June	103484
8	8	3	3	July	September	1139973
9	9	1	4	January	March	10000
10	10	2	4	April	June	614931
11	11	3	4	July	September	1539714
12	12	4	4	October	December	80471
13	13	1	5	January	March	14000
14	14	2	5	April	June	1391760
15	15	3	5	July	September	2884315
16	16	4	5	October	December	170592

Slika 11. Dimenzija kvartal

5.2 Ispitivanje tačnosti rezultata

Check-in	Check-out	Booked on	Status	Rooms	People	Adults	Children	Children's age(s)	Price	Commission %	Commission Amount
2019-04-03	2019-04-06	2019-03-30 19:53:46	ok	1	2				117 EUR	18	21.06 EUR
2019-04-06	2019-04-11	2019-03-01 09:25:42	ok	1	2				200 EUR	18	36 EUR
2019-04-07	2019-04-08	2019-03-31 07:50:45	ok	1	2				39 EUR	18	7.02 EUR
2019-04-10	2019-04-11	2019-04-02 13:30:27	ok	1	2				39 EUR	18	7.02 EUR
2019-04-11	2019-04-12	2019-04-02 15:03:10	ok	1	2				40 EUR	18	7.2 EUR
2019-04-12	2019-04-13	2019-03-13 21:32:44	ok	1	2				50 EUR	18	9 EUR
2019-04-12	2019-04-14	2019-03-28 13:31:43	ok	1	2				100 EUR	18	18 EUR
2019-04-15	2019-04-20	2019-04-01 22:00:38	ok	1	2				250 EUR	18	45 EUR
2019-04-16	2019-04-19	2019-03-20 20:33:15	ok	1	4	2	2	5, 7	180 EUR	18	32.4 EUR
2019-04-16	2019-04-18	2019-04-08 21:17:14	ok	1	2				80 EUR	18	14.4 EUR

Slika 12. Excel fajl Booking oglašivača iz 2019. Godine

Na slici 12. prikazani su podaci iz excel fajla iz 2019. godine kreirane od strane *Booking* oglašivača. Možemo primetiti da razlika između *Check-out* i *Check-in* kolone u stvari predstavlja kolonu *nights* iz činjenične tabele profit (Slika 13.). Takođe, kolona *Profit* (Slika 13.) predstavlja razliku kolona *Price* i *Commission Amount* iz excel fajla. Dodatno, može se primetiti da *EUR* unutar kolona *Commision Amount* i *Price* u stvari predstavlja valutu koja je preko svog identifikacionog broja iz dimenzione tabele valuta (Slika 7.) povezana sa kolonom dim_currency_idcurrency (Slika 13.). Sa slike 6. može se primetiti da je identifikaciona oznaka *Booking* oglašivača jedan baš kao i što piše u koloni dim_advertiser_idadvertiser u činjeničnoj tabeli *Profit* (Slika 13.)

	IDPROFIT	NIGHTS	PROFIT	DIM_ADVERTISER_IDADVERTISER	DIM_TMONTH_ID_TM...	DIM_APARTMENT_IDAPARTMENT	DIM_CURRENCY_IDCURRENCY
1	1	3	96	1	1	1	2
2	2	5	164	1	1	1	2
3	3	1	32	1	1	2	2
4	4	1	32	1	1	3	2
5	5	1	33	1	1	2	2
6	6	1	41	1	1	2	2
7	7	2	82	1	1	1	2
8	8	5	205	1	1	2	2
9	9	3	148	1	1	1	2
0	10	2	66	1	1	3	2

Slika 13. Prikaz dela činjenicke tabele Profit

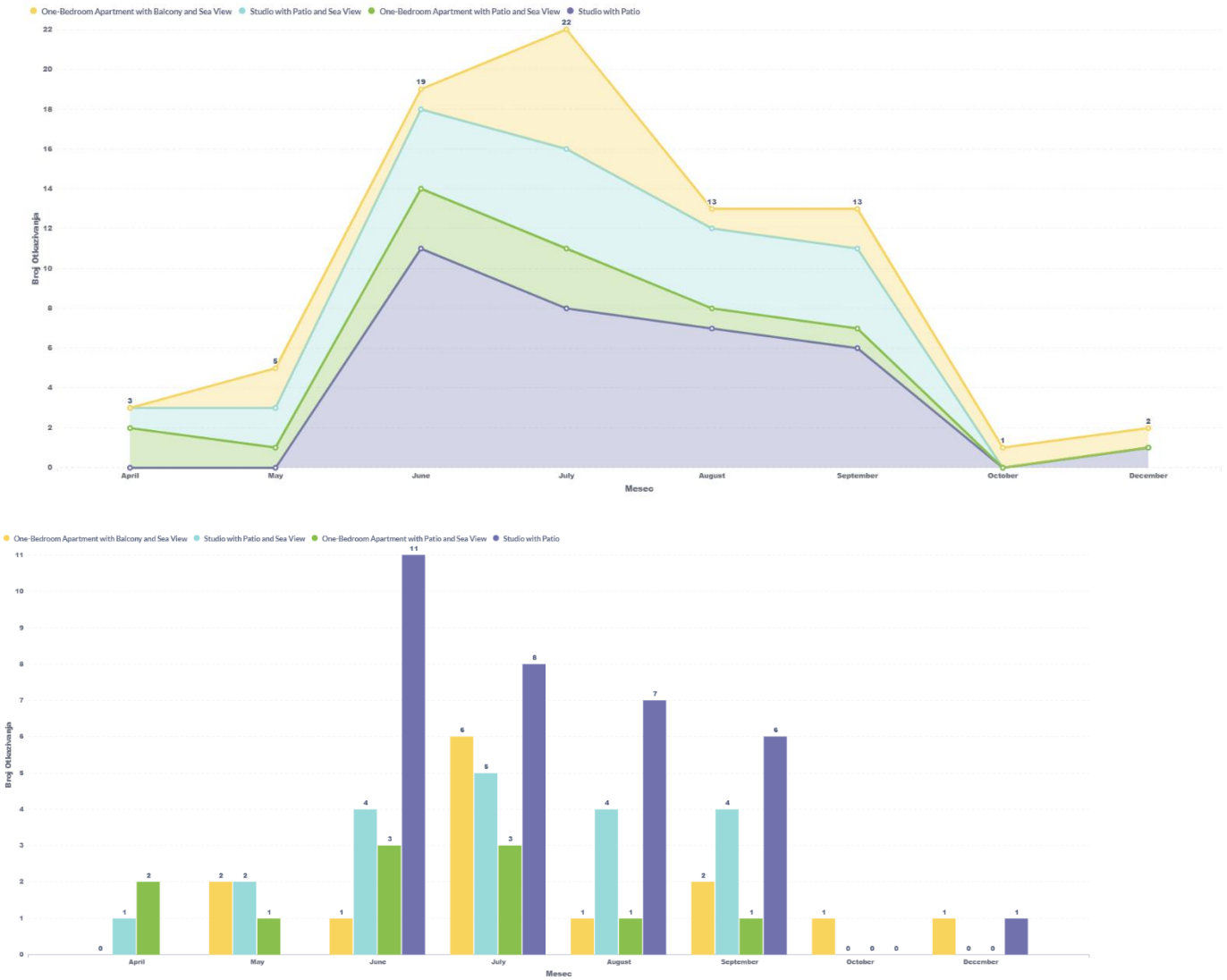
6.Materijalizovani pogledi

6.1 Prikazati broj otkazanih rezervacija svakog apartmana mesečno

```
create materialized view otkazi_po_mesecima_i_apartmanima
nologging
build immediate
refresh force
on demand
enable query rewrite
as  select SUM(numberofcancelation) AS broj_otkazivanja, dtm.month mesec, da.name apartman
    from fact_cancel fc, dim_tmonth dtm, dim_apartment da
    where fc.dim_tmonth_id_tmonth = dtm.id_tmonth and fc.dim_apartment_idapartment = da.idapartment
    group by dtm.month, da.name;
```

Slika 14. Kreiranje prvog materijalizovanog pogleda

6.1.1 Rezultati



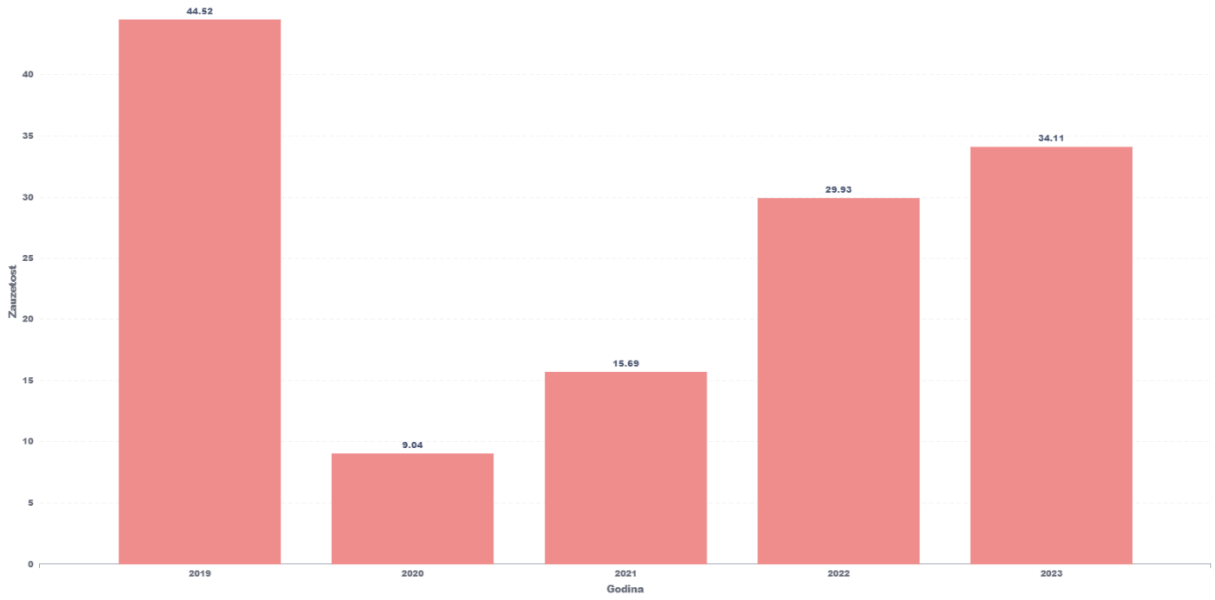
Slika 14. Prikaz rezultata prvog dashboard-a

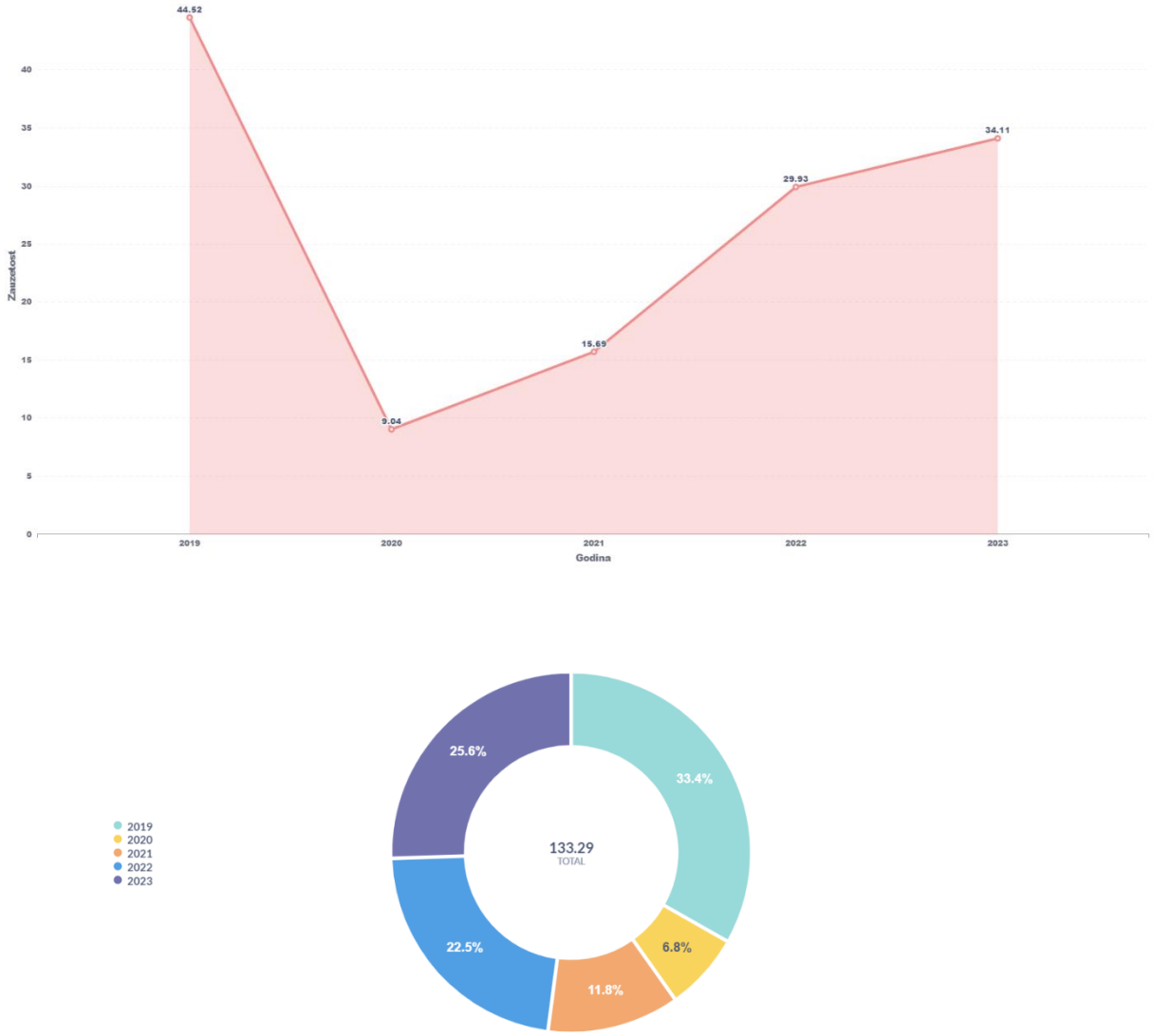
6.2. Prikazati godišnju zauzetnost apartmana izraženu u procentima

```
create materialized view zauzetost_po_godini
nologging
build immediate
refresh force
on demand
enable query rewrite
as  select round(avg(fb.busy),2) zauzetost ,dtm.year godina
    from fact_busy fb, dim_tyear dtm
    where fb.dim_tyear_idtime_year = dtm.idtime_year
    group by dtm.year;
```

Slika 15. Kreiranje drugog materijalizovanog pogleda

6.2.1 Rezultati





Slika 16. Prikaz rezultata drugog dashboard-a

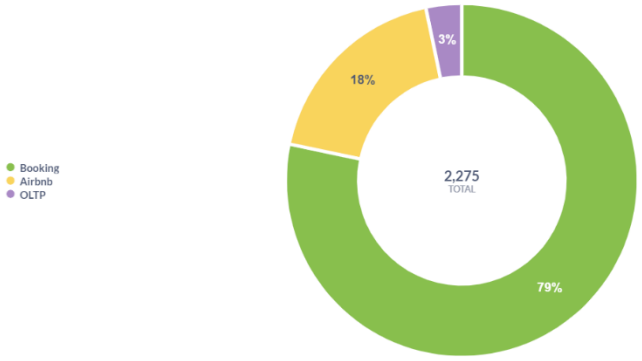
6.3 Prikazati ostvareni profit, broj noćenja kao i prosečnu cenu izdavanja apartmana u valuti u kojoj oglašivač posluje

```
create materialized view profit_nocenje_po_oglasivacu
nologging
build immediate
refresh force
on demand
enable query rewrite
as  select sum(fp.profit) profit, sum(fp.nights) broj_nocenja, da.name oglasivac, dc.name valuta, round(sum(fp.profit)/sum(fp.nights),2) prosek
    from fact_profit fp, dim_advertiser da, dim_currency dc
    where fp.dim_advertiser_idadvertiser = da.idadvertiser and fp.dim_currency_idcurrency = dc.idcurrency
    group by da.name, dc.name;
```

Slika 17. Kreiranje trećeg materijalizovaog pogleda

6.3.1 Rezultati

Profit	Oglasivac	Valuta	Prosek
85,862	Booking	EUR	47.97
18,356	Airbnb	DOLLAR	44.13
138,400	OLTP	RSD	2,005.8



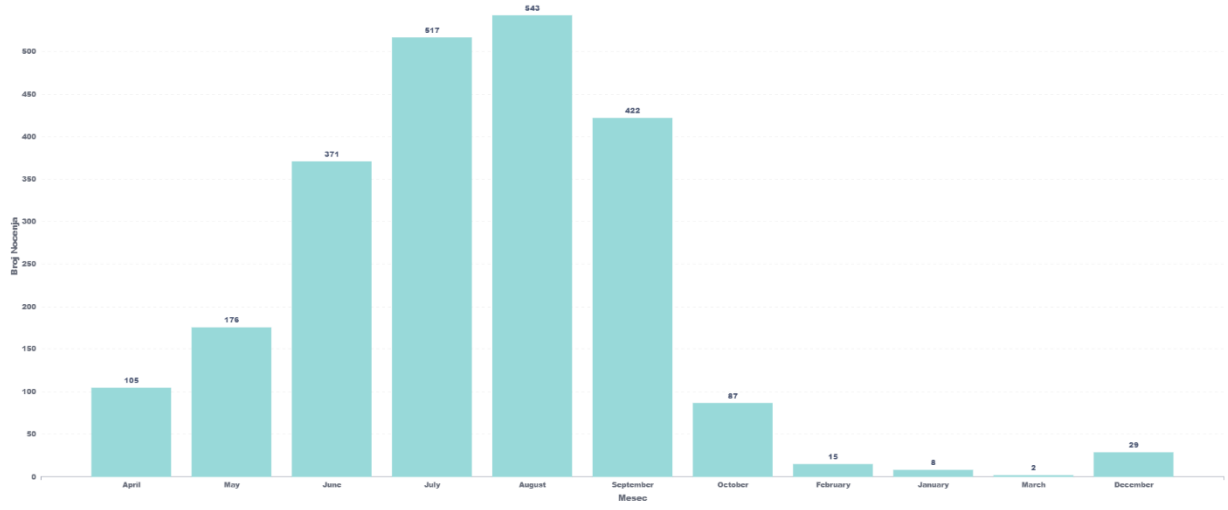
Slika 17. Prikaz rezultata trećeg dashboard-a

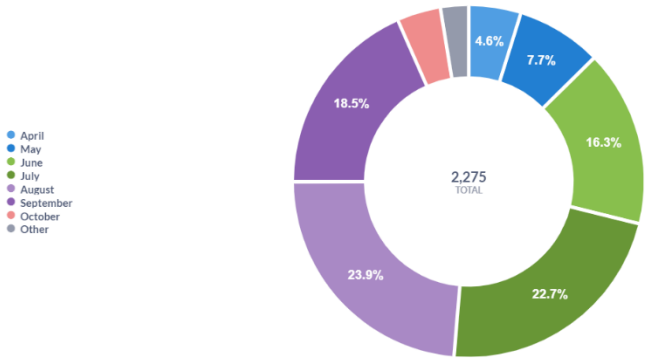
6.4 Prikazati ostvareni broj noćenja za svaki mesec u godini

```
create materialized view broj_nocenja_po_mesecu
nologging
build immediate
refresh force
on demand
enable query rewrite
as  select sum(fp.nights) broj_nocenja, dtm.month mesec
    from fact_profit fp, dim_tmonth dtm
    where fp.dim_tmonth_id_tmonth = dtm.id_tmonth
    group by dtm.month;
```

Slika 18. Kreiranje četvrtog materijalizovanog pogleda

6.4.1 Rezultati





Slika 19. Prikaz rezultata četvrtog dashboard-a