

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 2](#_Toc159683687)

[2.Opis OLTP šeme baze podataka 5](#_Toc159683688)

[3.Opis data warehouse šeme baze podataka 6](#_Toc159683689)

[3.1 Teme 7](#_Toc159683690)

[3.2 Bus Matrix 7](#_Toc159683691)

[3.3 Dimenzione tabele 7](#_Toc159683692)

[3.4 Činjenične tabele 8](#_Toc159683693)

[4.Opis ECTL procesa 9](#_Toc159683694)

[5.Testiranje 11](#_Toc159683695)

[5.1 Dimenzije 11](#_Toc159683696)

[5.2 Ispitivanje tačnosti rezultata 13](#_Toc159683697)

[6.Materijalizovani pogledi 15](#_Toc159683698)

[6.1 Prikazati broj otkazanih rezervacija svakog apartmana mesečno 15](#_Toc159683699)

[6.1.1 Rezultati 15](#_Toc159683700)

[6.2. Prikazati godišnju zauzetnost apartmana izraženu u procentima 16](#_Toc159683701)

[6.2.1 Rezultati 16](#_Toc159683702)

[6.3 Prikazati ostvareni profit, broj noćenja kao i prosečnu cenu izdavanja apartmana u valuti u kojoj oglašivač posluje 18](#_Toc159683703)

[18](#_Toc159683704)

[6.3.1 Rezultati 18](#_Toc159683705)

[6.4 Prikazati ostvareni broj noćenja za svaki mesec u godini 19](#_Toc159683706)

[6.4.1 Rezultati 19](#_Toc159683707)

# 1.Opis ER modela

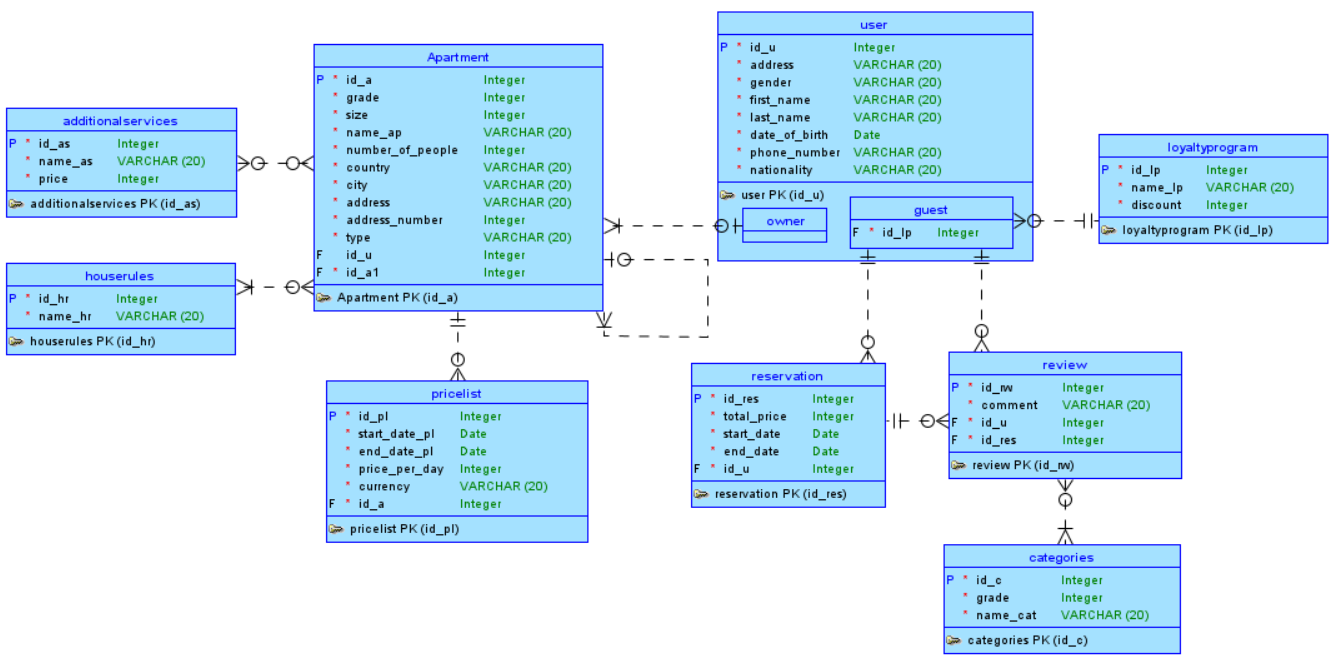
Slika 1. ER Model

* 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

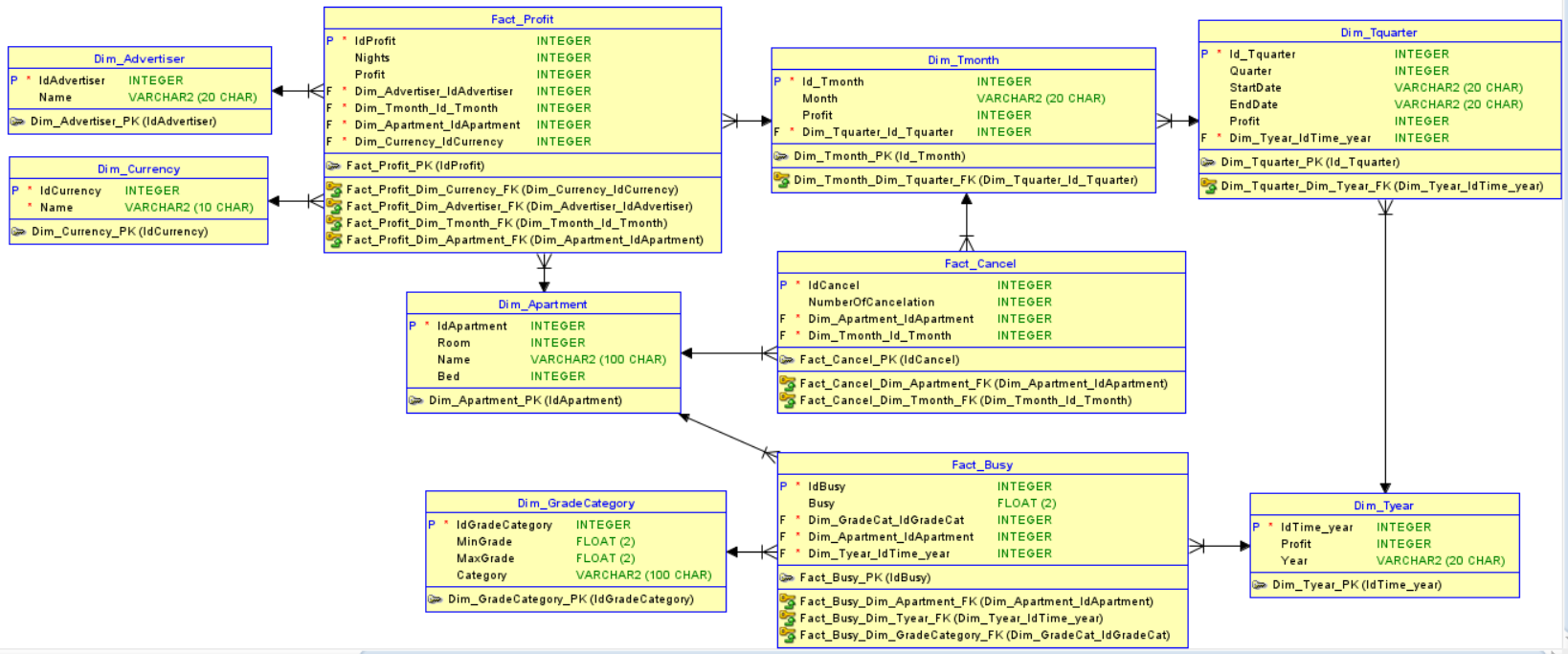
Skup obečežja

|  |  |
| --- | --- |
| 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

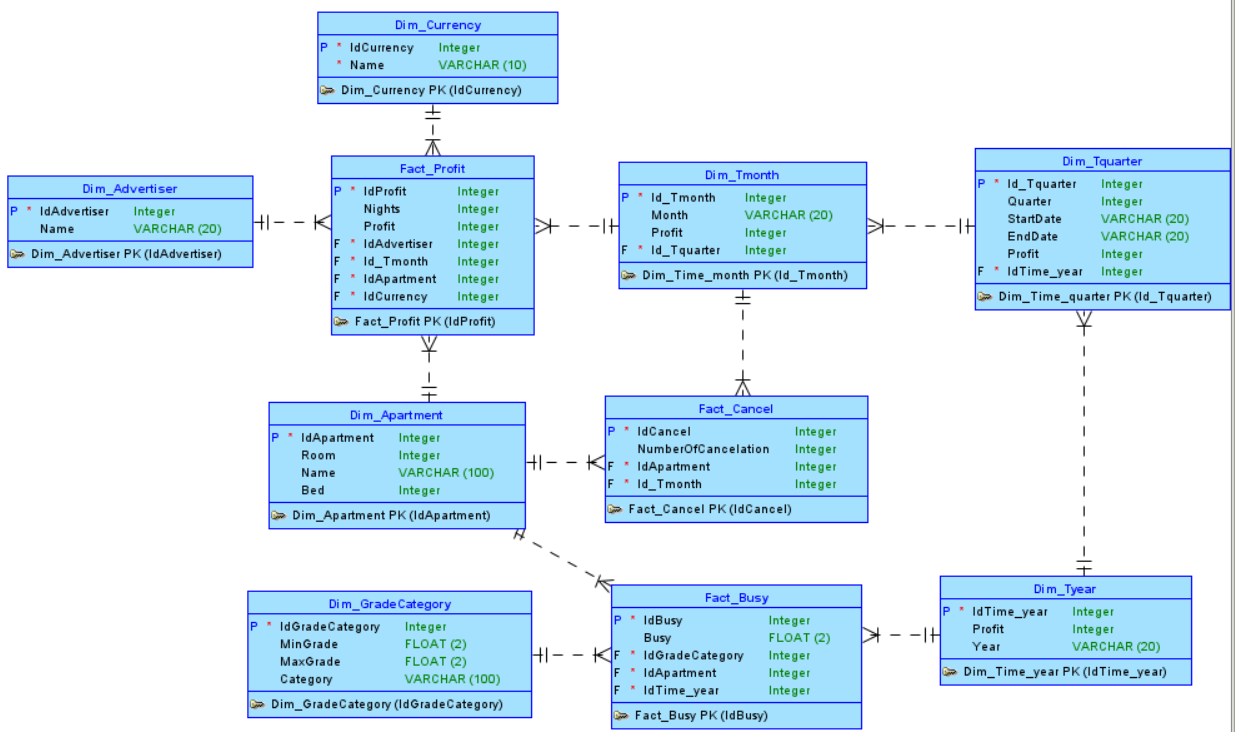


Slika 2. Logička šema OLTP baze podataka

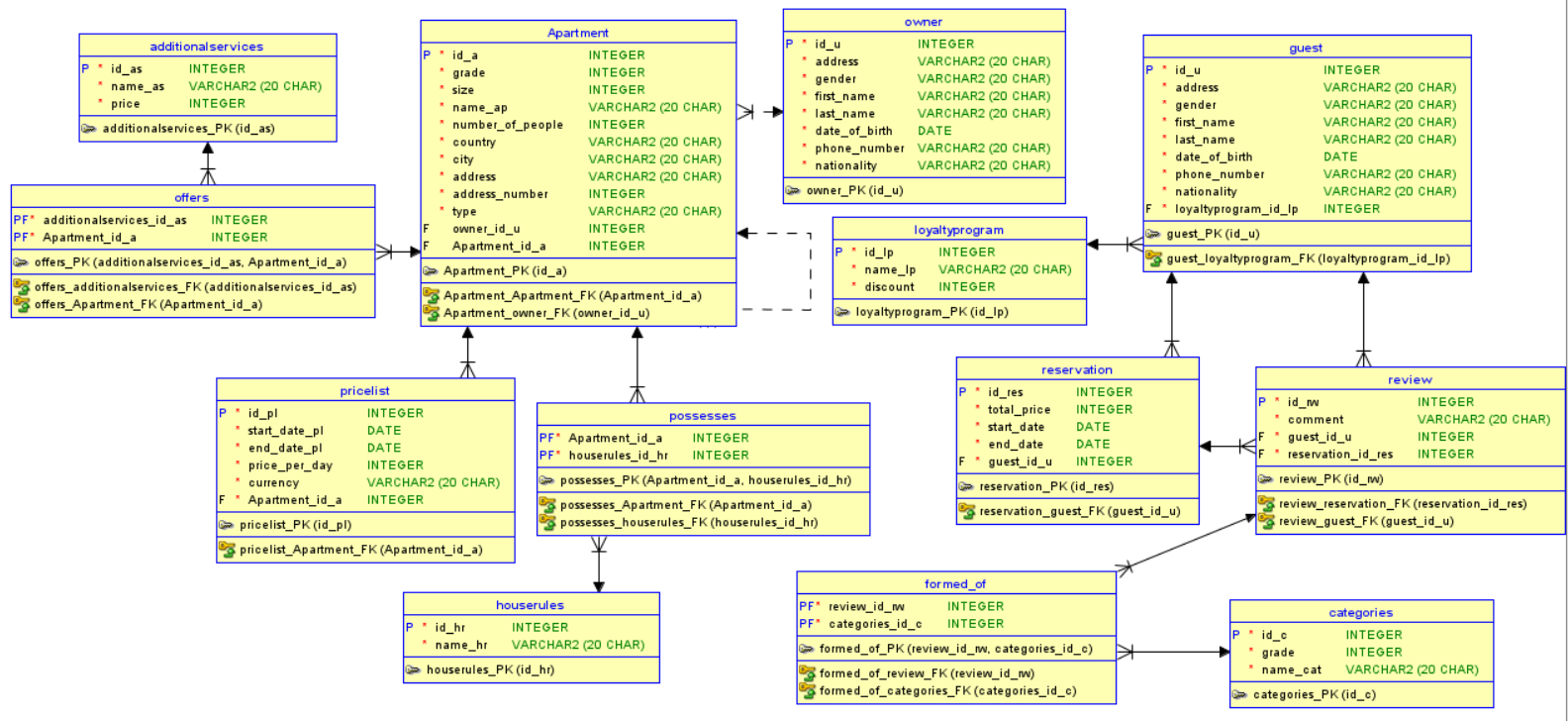


Slika 3. Relaciona šema OLTP baze podataka

# 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 mesecakao 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.

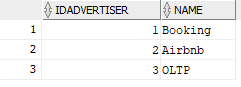
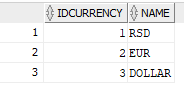
Procesiranje podataka generisanih od strane *Airbnb* oglašivača vršeno je na sledeći način. Na samom početku dodata je kolona *Advertiser* kako bi se znalo koja je vrsta oglašivača. Vršena je izmena formata datuma u DD.MM.YYYY kao i odbacivanje nevalidnih datuma. Takođe odbačeni su i redovi čija se vrednost 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*.

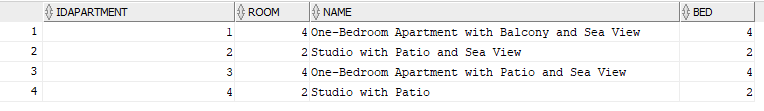
Transformacija i učitavanje podataka u *data warehouse* vršeno je u *TLData* python fajlu. Kao prvi korak vršeno je učitavanje *ProcessedData* fajla i kreirana je konekcija sa *Oracle* bazom podataka kroz dve funkcije za čitanje i pisanje. Podaci o nazivu oglašivača prikupljeni su iz kolone *Advertiser* 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 kljućeve 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

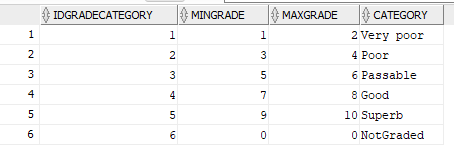
## 5.1 Dimenzije



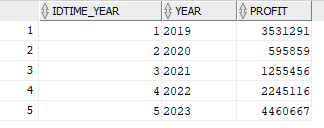
Slika 6. Dimenzija oglašivač Slika 7. Dimenzija valuta



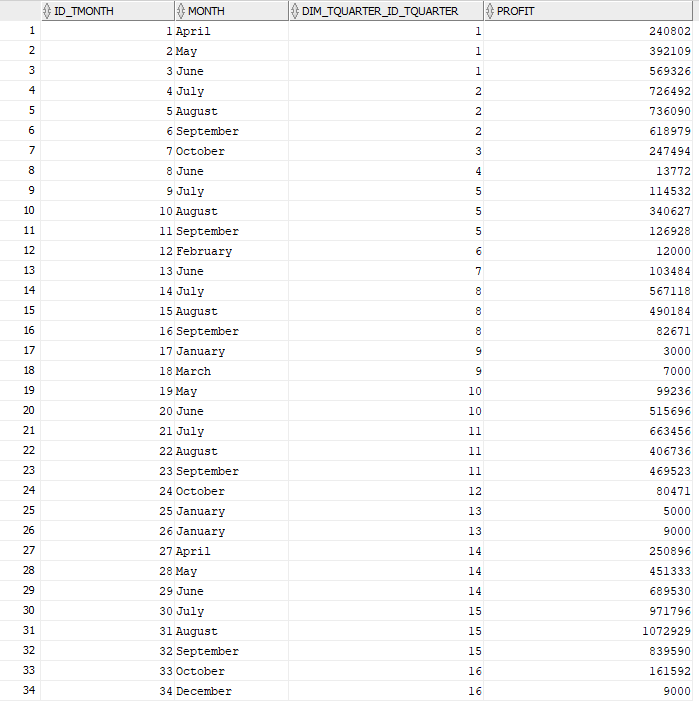
Slika 7. Dimenzija apartman



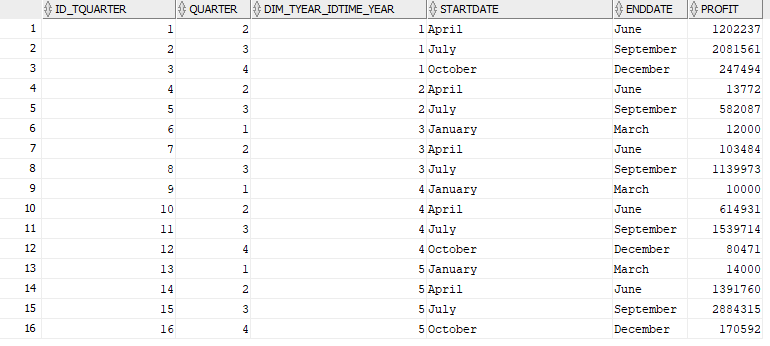
Slika 8. Dimenzija kategorije ocene



Slika 9. Dimenzija godina

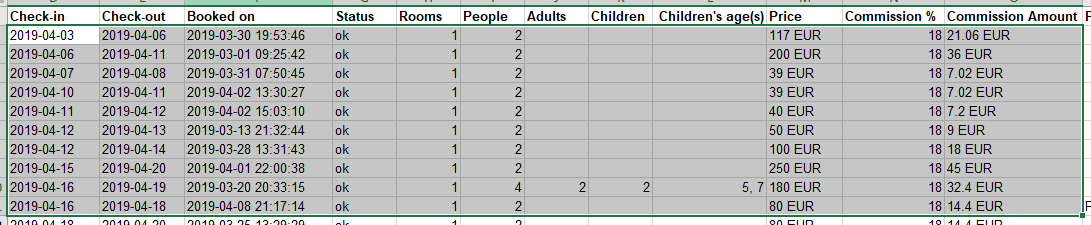


Slika 10. Dimenzija mesec



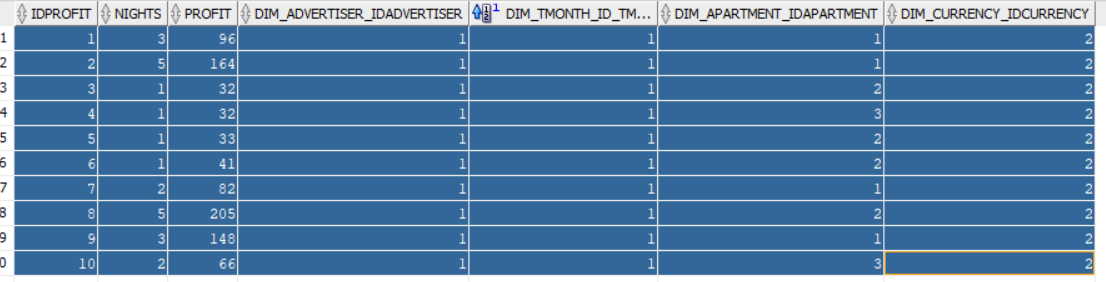
Slika 11. Dimenzija kvartal

## 5.2 Ispitivanje tačnosti rezultata



Slika 12. Excel fajl Booking oglašivača iz 2019. Godine

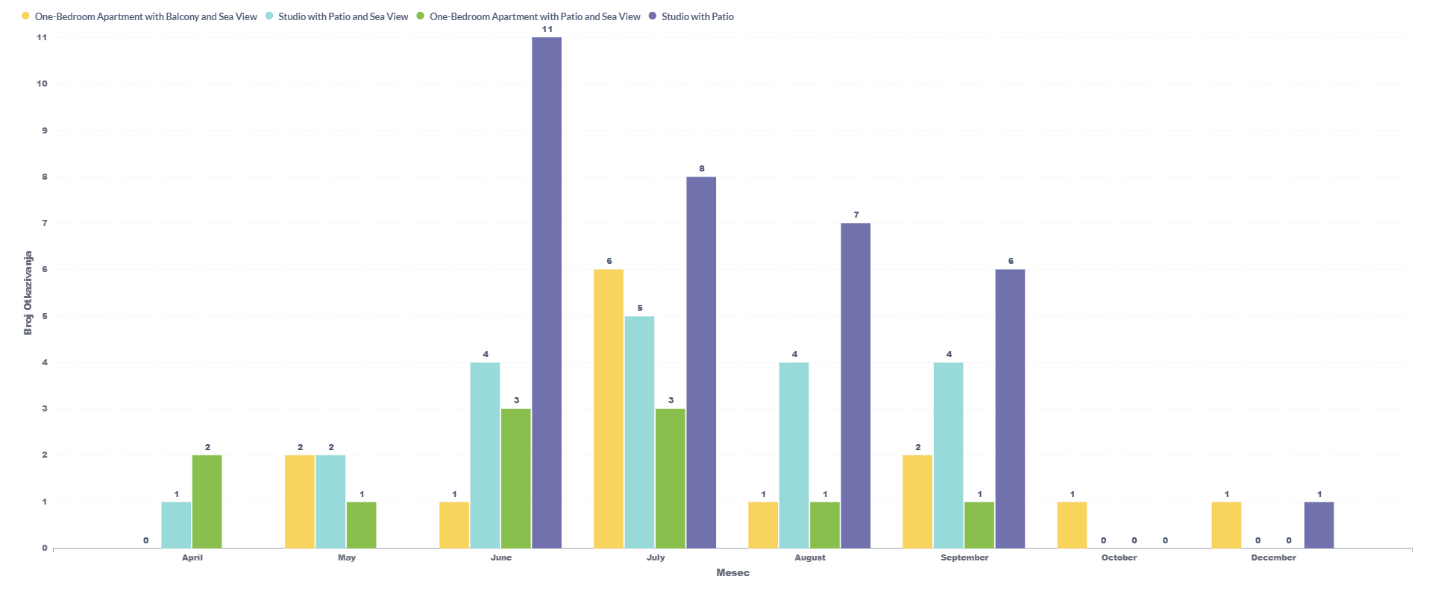
Na slici 12. prikazani su podaci iz excel fajla iz 2019. godine kreirane od strane *Booking* oglašivača. Možemo primetiti da razlika između *Check-out* i *Check-in* kolone u stvari predstavlja kolonu *nights* iz činjenične tabele profit (Slika 13.). Takođe, kolona *Profit* (Slika 13.) predstavlja razliku 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.)



Slika 13. Prikaz dela činjenicke tabele Profit

# 6.Materijalizovani pogledi

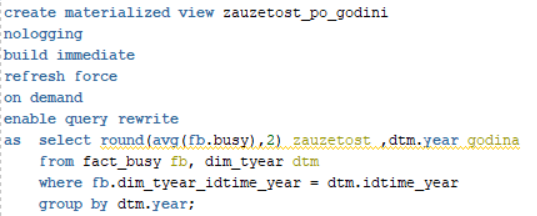
## 6.1 Prikazati broj otkazanih rezervacija svakog apartmana mesečno

Slika 14. Kreiranje prvog materijalizovanog pogleda

### 6.1.1 Rezultati

Slika 14. Prikaz rezultata prvog dashboard-a

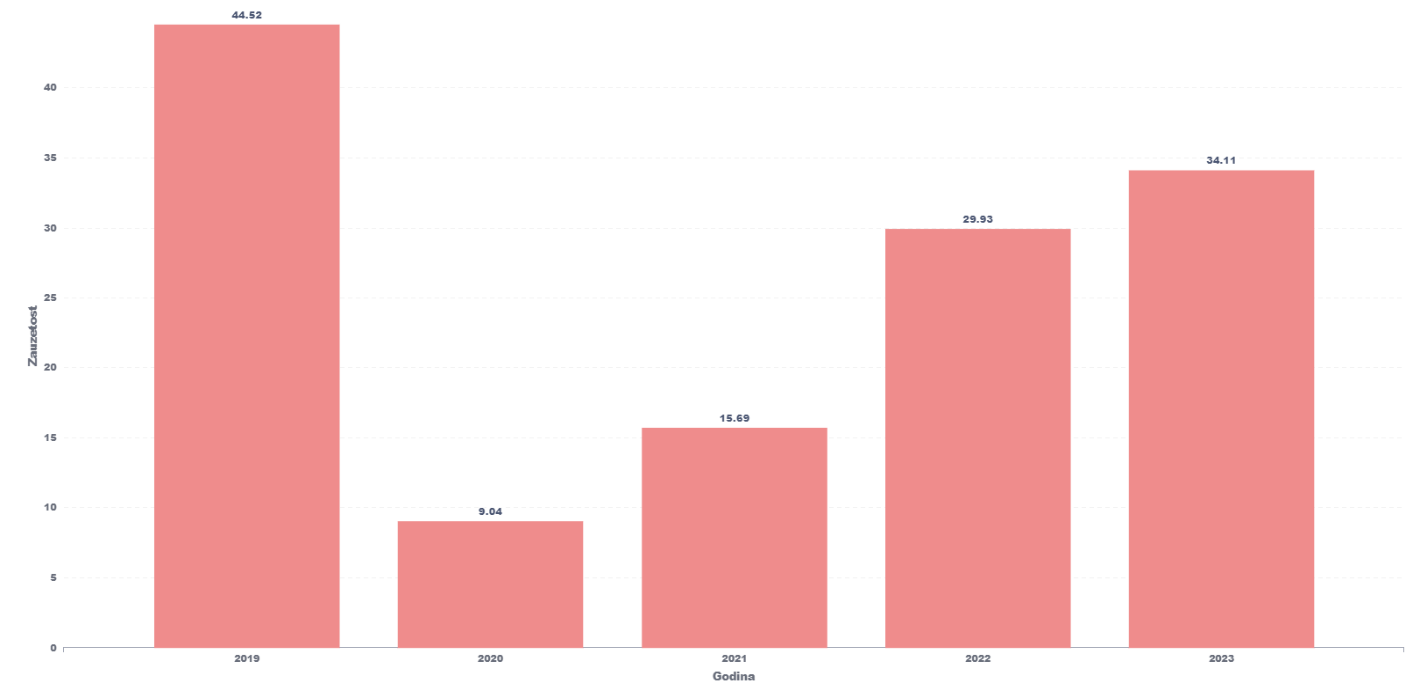
## 6.2. Prikazati godišnju zauzetnost apartmana izraženu u procentima

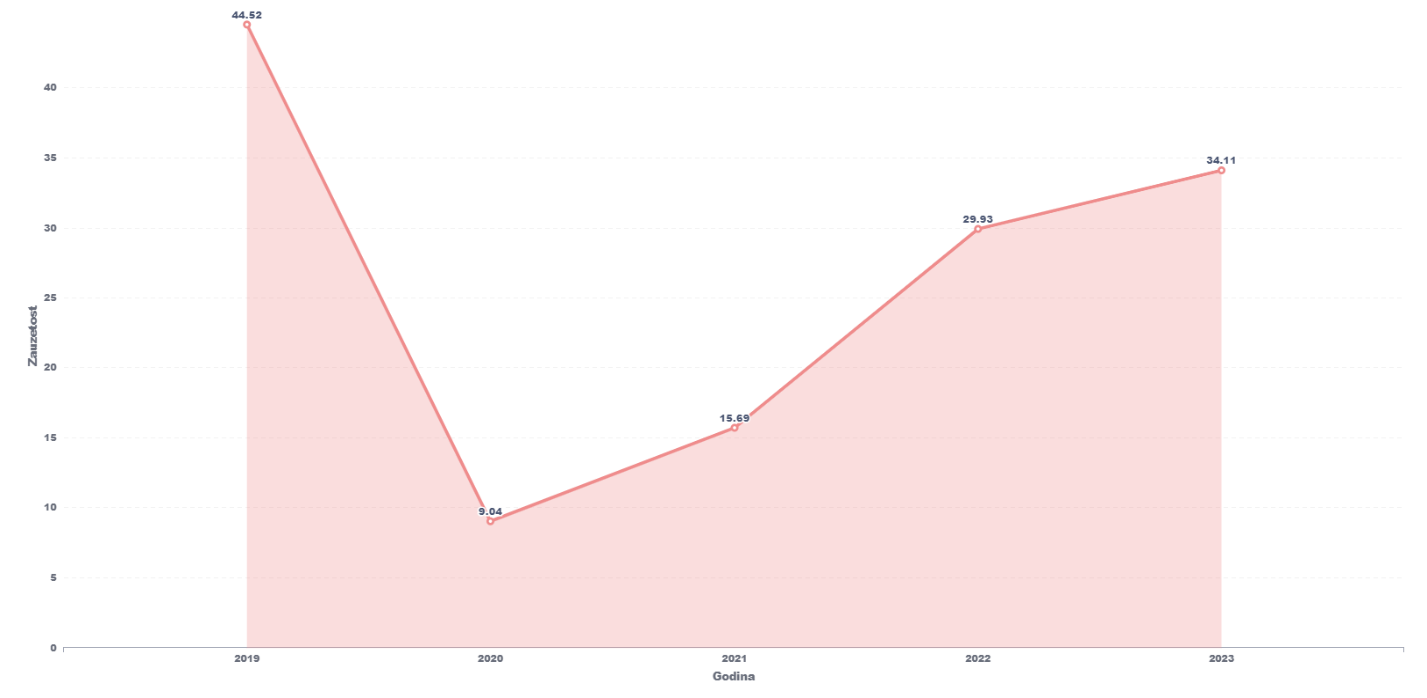
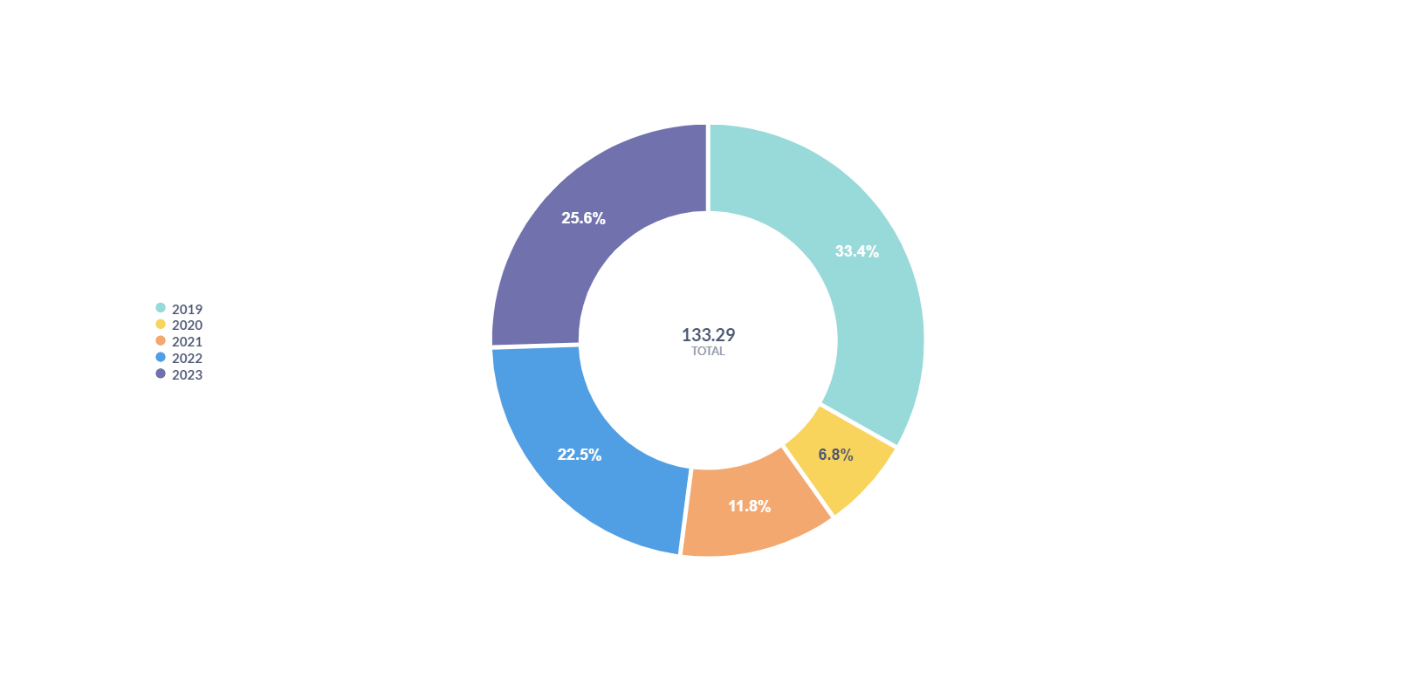


Slika 15. Kreiranje drugog materijalizovanog pogleda

### 

### 6.2.1 Rezultati



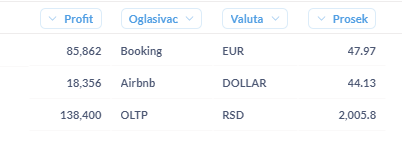
Slika 16. Prikaz rezultata drugog dashboard-a

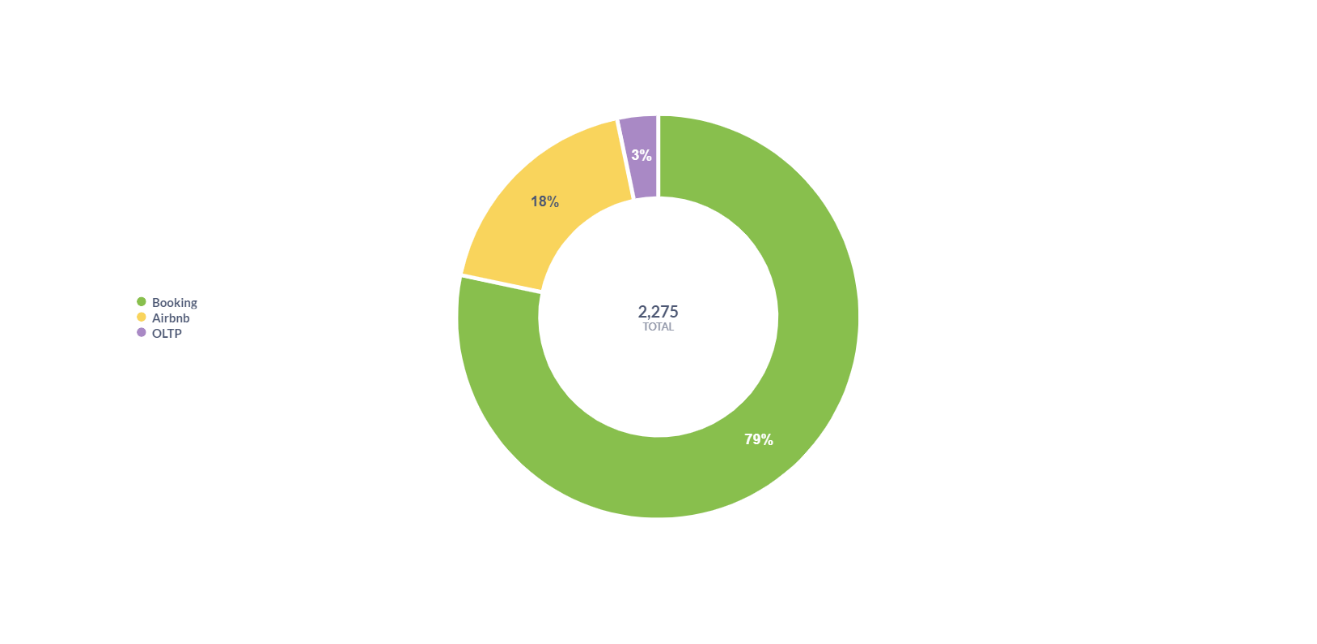
## 6.3 Prikazati ostvareni profit, broj noćenja kao i prosečnu cenu izdavanja apartmana u valuti u kojoj oglašivač posluje

## 

Slika 17. Kreiranje trećeg materijalizovaog pogleda

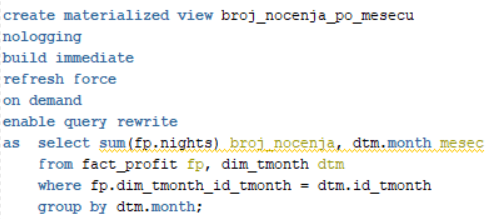
### 6.3.1 Rezultati





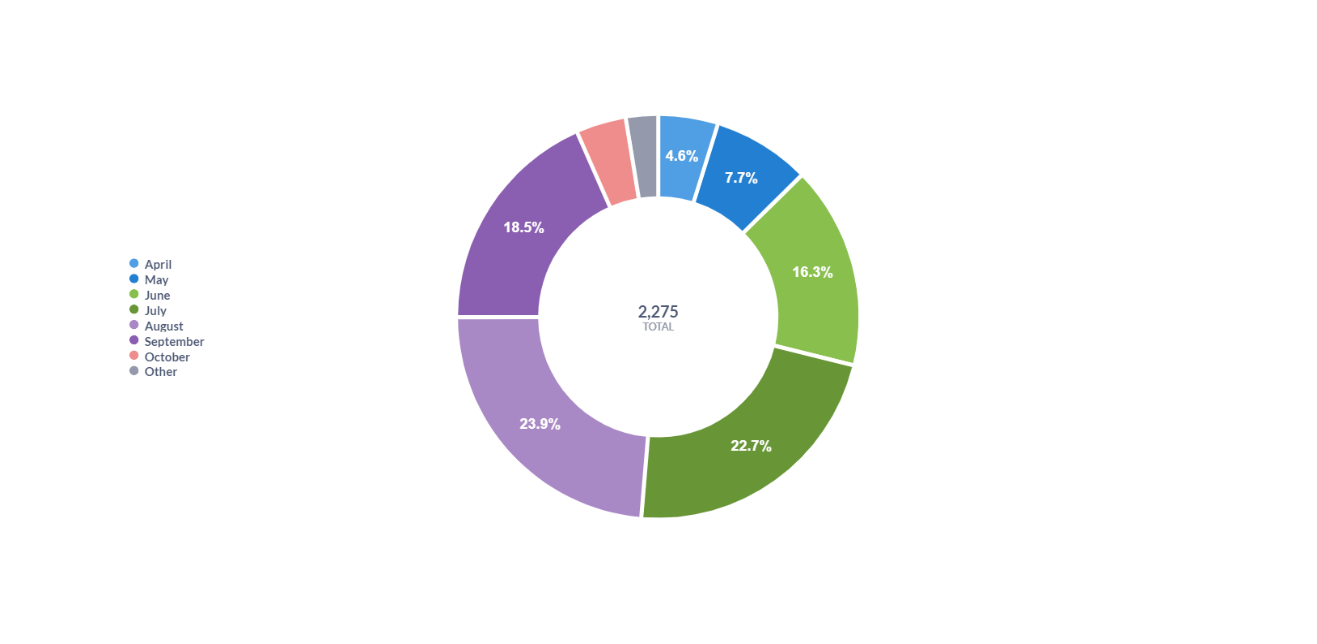
Slika 17. Prikaz rezultata trećeg dashboard-a

## 6.4 Prikazati ostvareni broj noćenja za svaki mesec u godini



Slika 18. Kreiranje četvrtog materijalizovanog pogleda

### 6.4.1 Rezultati



Slika 19. Prikaz rezultata četvrtog dashboard-a