Sveučilište Jurja Dobrile u Puli

Fakultet Informatike u Puli

**TOMISLAV BIĆANIĆ**

**Prodaja bicikla u Europi u periodu od 2013-2016.godine**

Seminarski rad

Sveučilište Jurja Dobrile u Puli

Fakultet Informatike u Puli

**TOMISLAV BIĆANIĆ**

**Prodaja bicikla u Europi u periodu od 2013-2016.godine**

Seminarski rad

**JMBAG :0303069482 , redovni student**

**Studijski smjer: Informatika**

**Kolegij: Sustavi poslovne inteligencije**

**Profesor: doc.dr.sc Goran Oreški**

Sadržaj

[1. Uvod 3](#_Toc82746957)

[2. Skupa podataka 4](#_Toc82746958)

[2.1 Analiza 4](#_Toc82746959)

[3. Relacijski model podataka 5](#_Toc82746960)

[3.1 Entity/Relationship model 5](#_Toc82746961)

[3.2 Popunjavanje baze podataka podacima 6](#_Toc82746962)

[4. Dimenzijski model podataka 9](#_Toc82746963)

[5. Proces dohvata, transformacije i učitavanja 11](#_Toc82746964)

[5.1. Stvaranje i punjenje dimenzije „customer“ 11](#_Toc82746965)

[5.2. Stvaranje i punjenje dimenzije „location“ 13](#_Toc82746966)

[5.3. Stvaranje i punjenje dimenzije „production\_cost“ 14](#_Toc82746967)

[5.4. Stvaranje i punjenje dimenzije „time\_of\_order“ 16](#_Toc82746968)

[5.5. Stvaranje i punjenje dimenzije „product“ 17](#_Toc82746969)

[5.6. Punjenje tablice činjenica 18](#_Toc82746970)

[6. Vizualizacija podataka 19](#_Toc82746971)

[7. Zaključak 21](#_Toc82746972)

[8. Literatura 22](#_Toc82746973)

# Uvod

Podatci u današnjem vremenu predstavljaju jednu od najvažnijih komponenata u informatičkom svijetu. Od jednostavnog skupa podataka pa do najsloženijih baza, potrebno je znati kako možemo manipulirati sa njima i što nama sve to omogućuje. Poslovna inteligencija je jedna od grana informatičkog svijeta koja se bavi upravo tim. Koriste se analitički alati kako bi se donesle poslovne odluke. Upravo nam ti alati služe za pretvorbu podataka i na temelju dobivene informacije doprinesti kvaliteti i razvoju sustava ili poslovnog okruženja.

U ovom projektu ćemo prikazati kako se obrađuje veliki skup podataka na relativno brz i efikasan način. Nakon obrade podataka imat ćemo uvid u trenutno stanje određene situacije.

Nakon što odaberemo naš skup podataka prvo ga trebamo analizirati da li ima i smisla ga uopće obrađivati, ako bude zadovoljavao naše kriterije onda ćemo definirati njegove elemente da bismo mogli stvoriti potrebne modele.

Podatke ćemo prvo spremiti u relacijsku bazu podataka pomoću MySql-a gdje ćemo definirati sve veze koje imamo te onda krećemo na dimenzijski model kojeg ćemo puniti putem transakcijskog sustava Pentaho. Na kraju ćemo vizualizirati naše podatke pomoću Microsoft Power Bi.

Cilj nam je napraviti što jednostavnije skladište podataka koji je lako uporabljiv i kompletan.

# Odabir skupa podataka

Prvobitno je da se odabere kvalitetan skup podataka koji zadovoljava određene zahtjeve. Skup podataka pronađen je preko Interneta koji već sadrži gotove skupove podataka. Podaci su preuzeti u obliku CSV datoteke na linku**:** https://www.kaggle.com/sadiqshah/bike-sales-in-europe

Podatci predstavljaju prodaju bicikala u periodu između 2013.do 2016 godine na europskom tržištu

## Analiza skupa podataka

Najvažniji elementi pri odabiru skupa podataka su podatci te ti isti podatci trebaju biti dovoljno raznoliki , različitih tipova i dovoljno opširni kako bi se mogla provesti analiza tih istih podataka. Najbitnije karakteristike koje određuju spremnost podataka za izradu modela su: pouzdanost izvora, točnost sadržaja, potpunost, konzistentnost, itd. Kod skupa podataka bitno je da sadržava dovoljno dimenzija te jednu vremensku dimenziju. Skup podataka ne smije imati previše nepostojećih vrijednosti te također ne smije imati malo jedinstvenih vrijednosti.

Jupyter smo koristili da bismo analizirali podatke. Utvrdili smo da podaci imaju dovoljan broj podataka 113037 i sastoje se od 18 stupca. Podaci su različitih tipova, sadrže brojeve i stringove. Različiti su i imaju više dimenzija od kojih nam je najbitnija vremenska dimenzija kojom ćemo se kasnije više pozabaviti.

# Relacijski model podataka

Nakon što se odabere i analizira skupove podataka potrebno je napraviti relacijski model podataka koji je tako reći transakcijski model koji podržava dnevno poslovanje organizacije i odgovoran je da podatcima popuni bazu podataka. On sadrži trenutne, detaljne, promjenjive podatke.

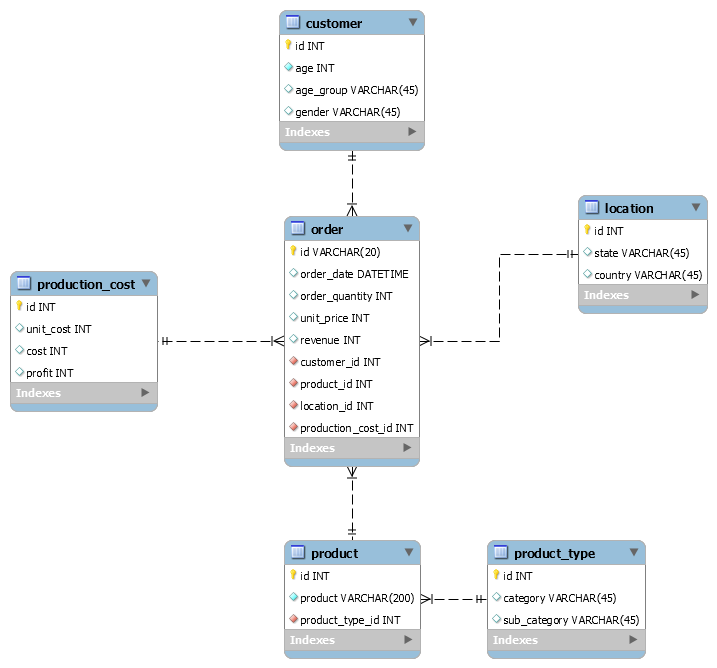
Pri samom početku prije korištenja baze podataka potrebno ju je dizajnirati da učinkovito podržava poslovanje organizacije, identificira tablice, atribute, ograničenja, itd. Rezultat dizajna je apstraktan prikaz stvarnosti, isključujući mnogo detalja iz stvarnog svijeta, poput smanjenja složenosti i fokusiranja samo na osnovne detalje

## Entity/Relationship model

Model entiteta/odnosa koristi se za idejno projektiranje. Omogućuje konceptualni pregled baze podataka. Potrebno je odrediti entitet koji predstavlja objekt od interesa, koji je prikazan kao pravokutnik sa zaobljenim vrhovima u ER dijagramu, koji predstavlja činjenice i atribute entiteta, te prikazan kao elipsa, koja definira odnos između dva ili više Entitet i kardinalnost veze prikazane u dijamantu.

Skup podataka podijeljen je u 6 odnosa: narudžba, kupac, proizvod, vrsta proizvoda, lokacija i tržnica. Središnji odnos predstavljat će tablicu činjenica u dimenzionalnom modelu dim\_order i sadržavat će podatke o svim narudžbama, dobiti i poštarini. Odnos s klijentima sadrži podatke o kupcu, podatke o lokaciji o mjestu narudžbe, tržnicu o mjestu prodajnog mjesta, proizvod sadrži podatke o proizvodu, njihov naziv, cijenu, količinu narudžbe i podatke o vrsti proizvoda o kategoriji i potkategorija proizvoda.

.



*Slika 1. Prikaz ER modela*

## Popunjavanje baze podataka podacima

Popunjavanje baze podataka prema stvorenom ER dijagramu je sljedeći korak. Koristit ću MySQL bazu podataka koji je besplatan i otvoreni sustav za upravljanje bazom podataka. Kako bi se ispunila baza podataka potrebno je napisati skriptu za popunjavanje bazu podataka podatcima. Koristim programski jezik Python za popunjavanje podataka u CSV datoteci u bazu podataka.

Morate koristiti funkciju pandas za uspostavu veze sa MySQL bazom podataka putem Pythona i učitavanje podataka iz CSV-a.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

*Slika 2. Konekcija na MySQL bazu i učitavanje CSV-a*

Nakon toga morate koristiti naredbu „CREATE TABLE“ za stvaranje tablice. Napravio sam 6 tablica u koje će spremati učitane podatke iz CSV-a.

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

*Slika 3. Kreiranje tablica*

Prikaz popunjenih tablica u MySQL-u:

Slika na kojoj se prikazuje stol

Opis je automatski generiran Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 4. Prikaz tablice customer Slika 5. Prikaz tablice location*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 6. Prikaz tablice product*

Slika na kojoj se prikazuje stol

Opis je automatski generiran Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 7. Prikaz tablice product\_type Slika 8. Prikaz tablice production\_cost*

*Slika na kojoj se prikazuje stol

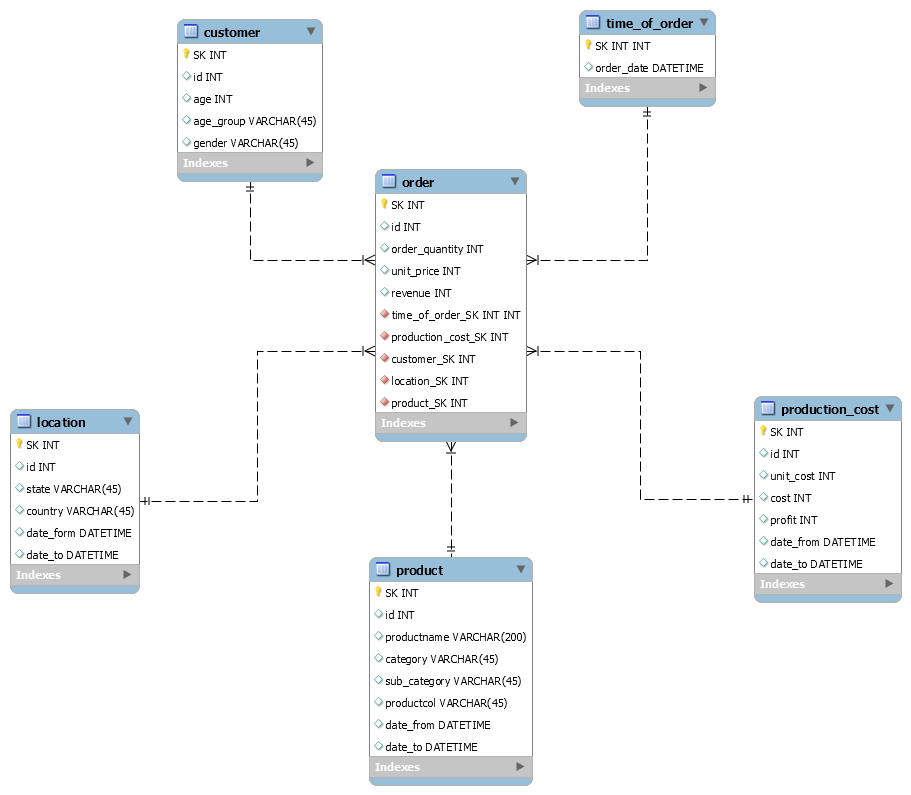
Opis je automatski generiran*

*Slika 9. Prikaz tablice order*

# Dimenzijski model podataka

Nakon stvaranja i popunjavanja relacijskog modela morate stvorit i popuniti dimenzionalni model. Modeliranje dimenzija omogućuje poslovnim korisnicima razumljivost podataka i brzu izvedbu upita.. Sastoji se od tablice činjenica i dimenzija.

Ona se nalazi u središtu dimenzijskog modela i sliži nam za modeliranje ostalih grana. Sastoji se od velikog broja zapisa i ima cilj da sadrži najnižu razinu granularnosti. Sadrži surogat ključ koji predstavlja primarni ključ tablice.



*Slika 10. Prikaz dimenzijskog modela*

# ETL proces

ETL nam predstavlja temeljni princip za analiziranje podataka i za machine learing

Nakon izrade dimenzijskog modela podataka slijedi punjenje skladišta podataka sa podacima iz transakcijskog sustava.

Za ovaj proces koristit ću alat Pentaho , softver za poslovnu inteligenciju koja pruža integraciju podataka, OLAP usluge, izvještavanje, mogućnost rudarenja podataka, izdvajanje, transformaciju i učitavanje.

## Stvaranje i punjenje dimenzije „customer“

Slika na kojoj se prikazuje tekst

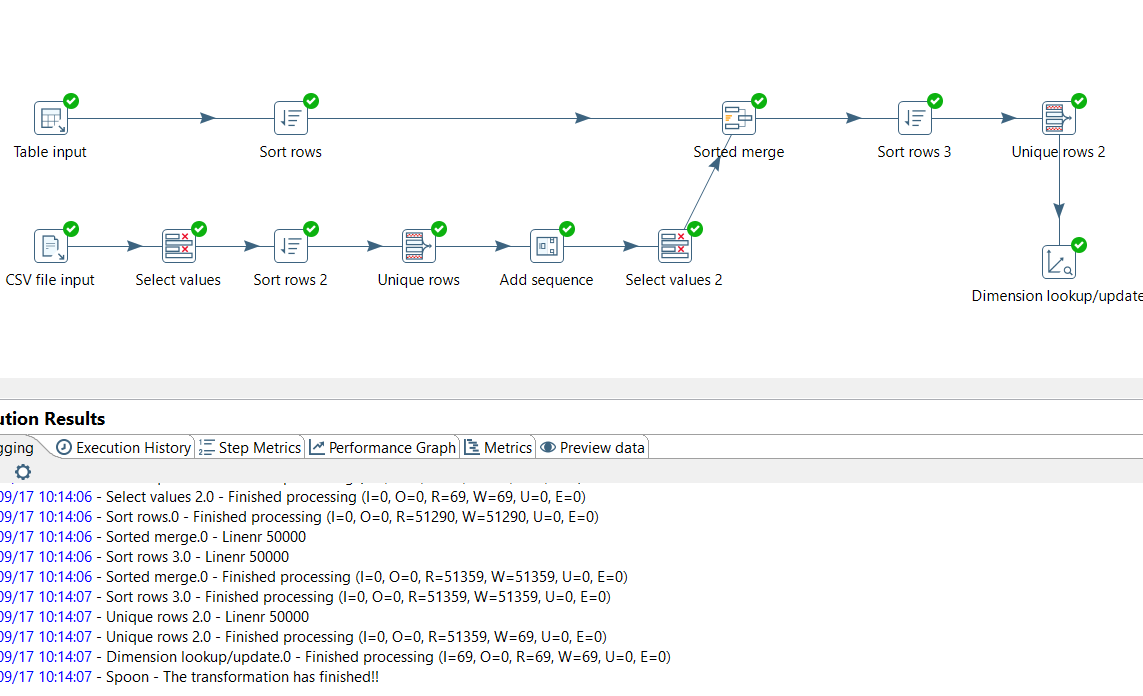
Opis je automatski generiran

*Slika 11. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 12. Prikaz dimenzijske tablice customer*

**

*Slika14. Proces punjenja dimenzijske tablice customer*

## Stvaranje i punjenje dimenzije „location“

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

*Slika 14. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 15. Prikaz dimenzijske tablice location*

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

*Slika 16. Proces punjenja dimenzijske tablice location*

## Stvaranje i punjenje dimenzije „production\_cost“

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

*Slika 17. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 18. Prikaz dimenzijske tablice production\_cost*

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

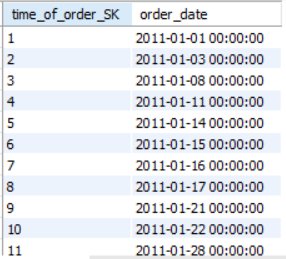
*Slika 19. Proces punjenja dimenzijske tablice production\_cost*

## Stvaranje i punjenje dimenzije „time\_of\_order“

Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 20. Prikaz dimenzijske tablice time\_of\_order*

**

*Slika 21. Proces punjenja dimenzijske tablice time\_of\_order*

## Stvaranje i punjenje dimenzije „product“

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

*Slika 22. Prikaz dimension lookup/update*

**

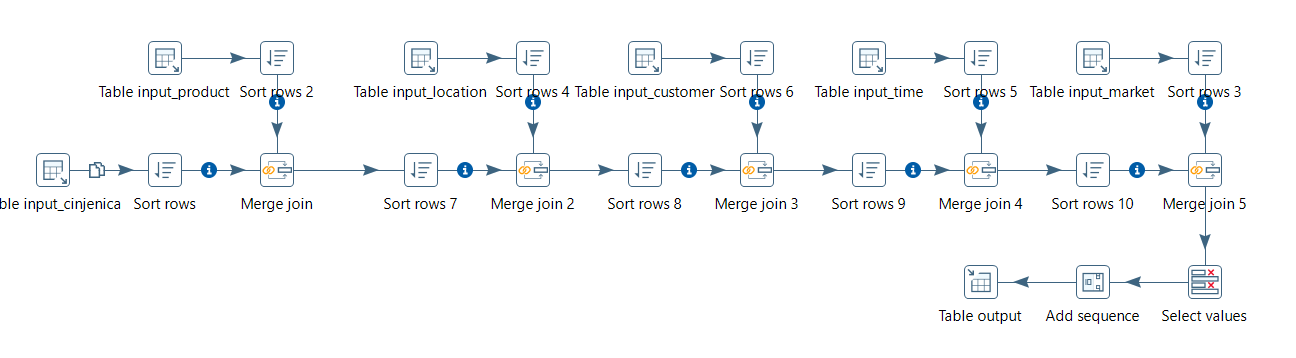
*Slika 23. Prikaz dimenzijske tablice product*

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

*Slika 24. Proces punjenja dimenzijske tablice product*

## Punjenje tablice činjenica



*Slika 25. Proces punjenja tablice činjenica order*

*Slika na kojoj se prikazuje stol

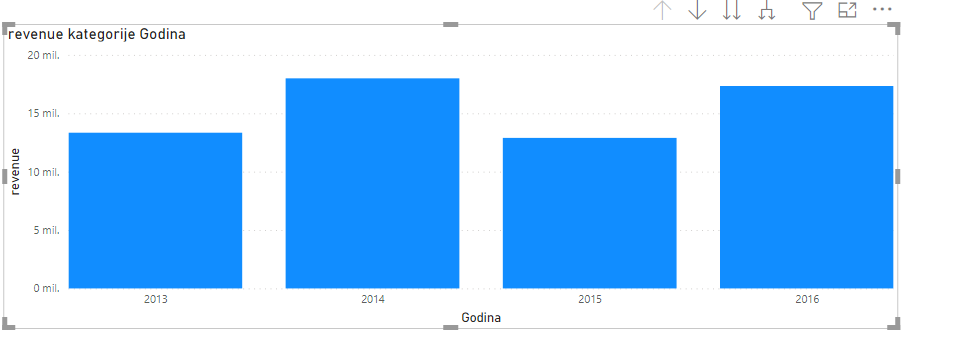
Opis je automatski generiran*

*Slika 26. Prikaz tablice činjenica order*

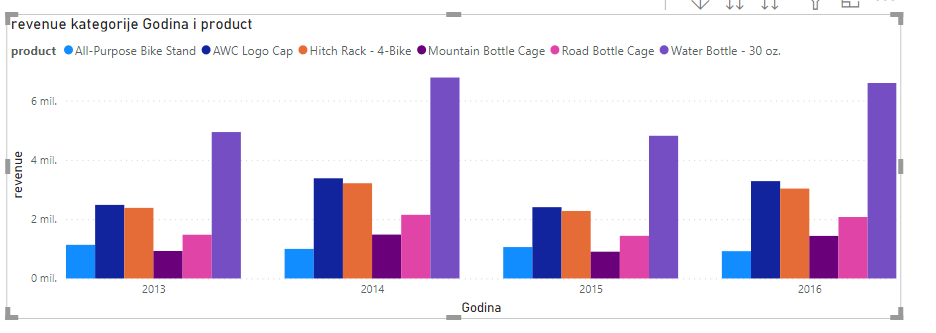
# Vizualizacija podataka

OLAP (OnLine Analytic Processing) je metoda koja podržava i pruža korisniku jednostavnu analizu multi-dimenzijskih podataka. Rezultati su dani u grafičkom obliku. Možemo promatrati dimenzije iz različitih uglova-

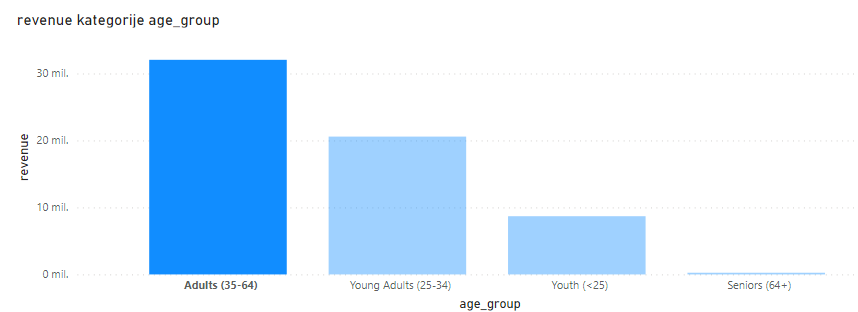
Nakon instalacije programa potrebno je uspostaviti konekciju MySQL-a i MicrosoftPower BI te odabrati sve tablice dimenzijskog modela.



*Slika 27. Prikaz ukupno ostvarenog profita za svaku godinu od 2013 do 2016.*



*Slika 28. Prikaz ukupno ostvarenog profita po pod kategorijama proizvoda za svaku godinu*



*Slika 39. Prikaz ukupnog prihoda po dobnim skupinama potrošača*

# Zaključak

Ovaj projekt temelji se na stvaranju skladišta podataka koji je jedinstven, kompletan repozitorij podataka prikupljen iz raznih izvora i predstavljen krajnjem korisniku na razumljiv način.

Proces stvaranja je prilično zahtjevan. Započeo je s odabirom skupa podataka s interneta koji su detaljno analizirani. Od analiziranih podatka stvoren je relacijski model podataka. Pomoću programskog jezika Python podaci su filtrirani kako bi popunili relacijsku bazu. Relacijski model pretvoren je u dimenzijski model koristeći ETL proces, izdvajanjem, transformiranjem i učitavanjem podataka. Alatom Pentaho Data Integration podaci su premljeni u dimenzijski model koji se sastoji od dimenzija i tablice činjenica. Na kraju, alatom Microsoft Power BI napravljena je vizualizacija podataka za potrebne analize na temelju kojih se donose važne poslovne odluke.

Rezultat mora biti jednostavan za korištenje poslodavcu te pružiti brz, lak pristup uvidima u trenutno stanje organizacije na temelju dostupnih podataka te lakše donošenje poslovnih odluka.

# Literatura

Pentaho 9.0, Online dokumentacija:

<https://help.pentaho.com/Documentation/9.0>