SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Iva Melnjak

Analiza trendova YouTube videozapisa

PROJEKT IZ KOLEGIJA SKLADIŠTA PODATAKA I POSLOVNA INTELIGENCIJA

Varaždin, 2021.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Iva Melnjak

Matični broj: 0016121322

Studij: Organizacija poslovnih sustava

Analiza trendova YouTube videozapisa

PROJEKT IZ KOLEGIJA SKLADIŠTA PODATAKA I POSLOVNA INTELIGENCIJA

Mentor/Mentorica:

Prof. dr. sc. Kornelije Rabuzin

Varaždin, lipanj 2021.

Sadržaj

1. Uvod

U ovom projektu je potrebno izraditi skladište podataka i prikupljene podatke vizualno prikazati. Uvjet koji prvotno treba zadovoljiti je dobar dataset, odnosno pronalazak podataka iz kojih se može izraditi model zvijezde kojeg će uz činjeničnu tablicu činiti minimalno tri dimenzijske tablice. Prilikom preuzimanja podataka, obrade i prikaza istih, korišteni su različiti alati. Dataset je preuzet u csv formatu te je zatim otvoren i prilagođen u Microsoft Excelu i ponovno spremljen kao csv radi lakšeg uvođenja u Microsoft SQL Server Management Studio.

Po završetku procesa izrade i uređenja podataka prelazi se u vizualizaciju istih. Vizualizacijom se prikazuje svrha podataka, a ona se u ovom radu prikazuje pomoću alata Power BI, kreiranjem izvještaja. Sve korištene tehnologije i provedeni koraci bit će objašnjeni u nastavku rada.

2. Opis domene prikazane skladištem podataka

YouTube je mrežna usluga za razmjenu videozapisa na kojoj korisnici mogu postavljati, pregledavati i ocjenjivati videozapise. Za pregledavanje videozapisa nije potrebna registracija, ali za postavljanje se potrebno registrirati čime se kreira takozvani kanal. Na YouTube-u se svakodnevno mogu vidjeti trendovi, odnosno najpopularniji videozapisi tog dana.

U ovom radu će biti prikazani podaci o videozapisima koji su na platformi YouTube imali najviše pregleda u 2021. godini, točnije u prva tri mjeseca 2021. godine. Skup podataka svakodnevno bilježi najpopularnije YouTube videozapise. Podaci su uzeti za 4 različite države, a osim naziva videozapisa, sadrže i naziv kanala preko kojeg se isti objavljuje te ostale informacije kao što su oznake i opis videozapisa. Osim broja pregleda, ostali faktori koji se uzimaju kako bi se utvrdilo koji su videozapisi bili najpopularniji su broj komentara i broj pozitivnih reakcija.

Nakon što se kreira skladište podataka, pomoću alata za vizualizaciju će se svi podaci uspoređivati, a naglasak će biti na usporedbi broja pregleda, komentara i broja pozitivnih i negativnih reakcija na najpopularnije videozapise.

3. Opis korištenih tehnologija

Za izgraditi skladište podataka i provesti analizu nad podacima potrebno je provesti tri koraka, dohvaćanje, obradu i učitavanje podataka (engl. Extract, Transform, Load). Nakon izrade skladišta podatke je potrebno prikazati pomoću izvještaja. Za provođenje ovih koraka i izradu ovog projekta korištene su redom sljedeće tehnologije:

- Microsoft Excel
- Microsoft SQL Server Management Studio
- Microsoft Power BI

Sam dataset se sastojao od četiri csv datoteke koje su preuzete s Kaggle-a. Nakon preuzimanja datoteke su otvorene u Microsoft Excelu radi uređivanja i dodavanja potrebnih atributa. Nakon Excela ponovno su spremljene u csv format kako bi se mogle uvesti u Microsoft SQL Server Management Studio (SSMS). SSMS je integrirano okruženje za upravljanje bilo kojom SQL infrastrukturom.

SSMS nudi alate za konfiguriranje, nadgledanje i administriranje instanci SQL Servera i baza podataka. Koristi se za postavljanje upita, dizajn i upravljanje bazama podataka i skladištima podataka na lokalnom računalu ili u oblaku [1]. Za ovaj projekt kreirana je baza podataka na lokalnom poslužitelju. Prije samog kreiranja potrebnih tablica, uvezene su četiri csv datoteke i naredbom SELECT i INSERT INTO su spojene u jednu tablicu kako bi kasnije bilo lakše popunjavati ostale tablice. Nakon toga su kreirane četiri dimenzijske tablice i jedna činjenična te su one spojene vanjskim ključevima generiranjem ERA modela. Izrađene tablice su zatim popunjene podacima kreiranjem novih upita iz već spomenute tablice nastale spajanjem csv datoteka.

Nakon što je imamo sve potrebno u SSMS-u želimo analizirati podatke iz tablica i prikazati njihovu povezanost. Za to je korišten alat Power BI koji pruža interaktivne vizualizacije i mogućnosti poslovne inteligencije. Ima jednostavno sučelje i omogućuje dohvaćanje podataka iz različitih izvora za stvaranje vlastitih izvješća [2]. Podaci iz izrađenog skladišta podataka su učitani u Power BI te su zatim kreirani izvještaji koje je moguće urediti po svojoj želji preko raznih mogućnosti koje ovaj alatu nudi.

4. Izrada skladišta podataka

Kao početak izrade skladišta podataka kreira se model podataka gdje su definirani svi podaci koji će kasnije biti korišteni za generiranje izvještaja. Nakon kreiranja modela podataka provest će se ETL proces, odnosno dohvatit će se podaci, pripremiti za uvoz u skladište i na kraju učitati u skladište podataka.

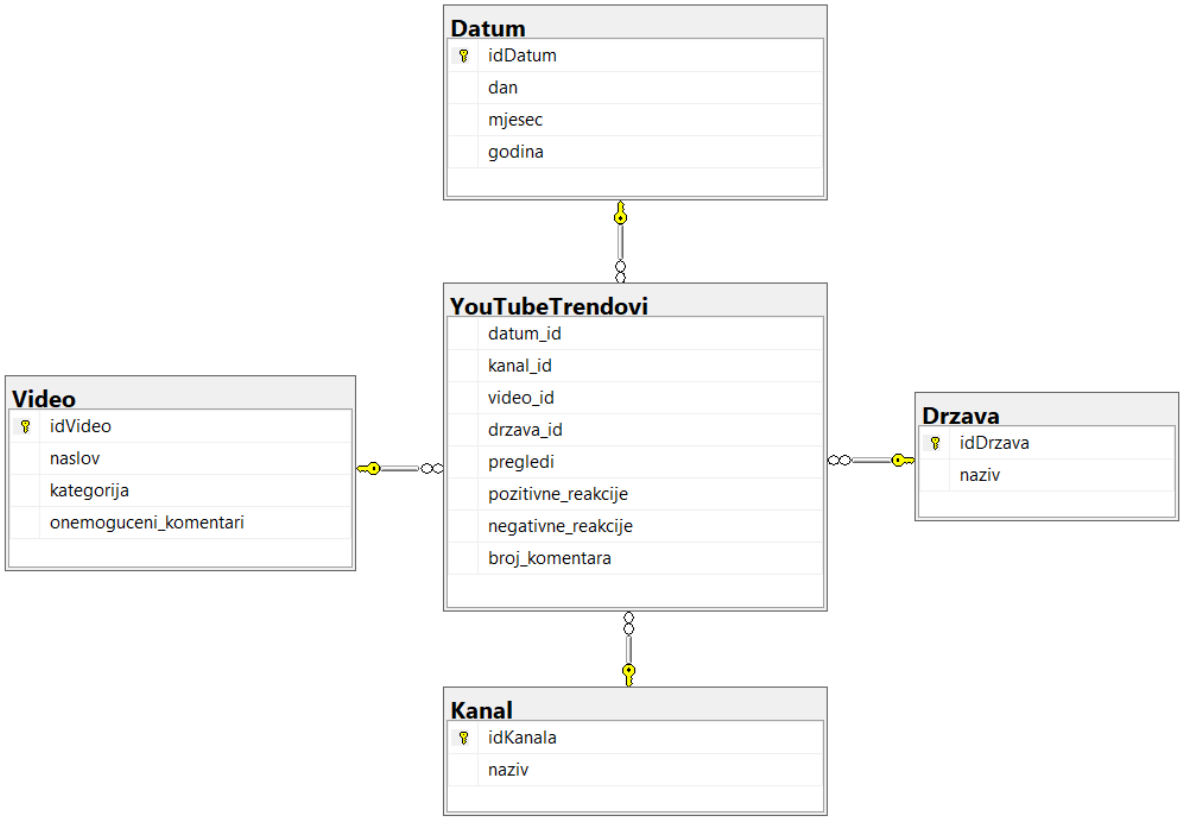
4.1. Opis korištenog skupa podataka

Za izradu ovog projekta korišten je dataset o najpopularnijim YouTube videozapisima iz 2021. godine preuzet sa stranice Kaggle. Kaggle je službena stranica na kojoj su godinama objavljivani podaci i možemo ju smatrati pouzdanim izvorom. Sam skup podataka ukupno sadrži 11 csv datoteka za 11 različitih država. Svaka datoteka ima jednake atribute i svaka ima 6000 redova. Za ovaj projekt uzeti su podaci iz 4 datoteke kako bi se razumio sadržaj istih jer su podaci u tim datotekama većinom na engleskom jeziku za razliku od drugih datoteka. Države za koje su uzeti podaci su Kanada, Meksiko, Ujedinjeno Kraljevstvo i Sjedinjenje Američke Države.

Osim csv datoteka za kreiranje skladišta podataka preuzeti su i podaci iz drugog dataseta u kojem su u .json obliku navedene kategorije videozapisa za svaku pojedinu državu [3]. Dataset sadrži ukupno 16 stupaca koji opisuju zanimljive značajke najpopularnijih YouTube videozapisa kao što su naslov videozapisa, datum objavljivanja istog, opis videozapisa, ime kanala s kojeg je videozapis objavljen i slično. Također dataset ima i numeričke podatke koji zapravo i pokazuju popularnost nekog videa a to su broj pregleda, broj pozitivnih i broj negativni reakcija i broj komentara na nekom videozapisu. Kako su podaci objavljeni na stranici Kaggle prije 2 mjeseca sadrže podatke za prva tri mjeseca 2021. godine, odnosno za siječanj, veljaču i ožujak. Podaci o najpopularnijim videozapisima prikupljaju se svakodnevno pa je u tih 6000 prikupljenih redova moguće da je neki videozapis više puta naveden jer može biti najpopularniji više puta, odnosno više dana za redom. Isto tako osoba, odnosno jedan YouTube kanal može objaviti više videozapisa koji će postati najpopularniji, pa se tako u datasetu neki kanal može pojaviti više puta.

4.2. Opis modela izgrađenog skladišta podataka

Prvi korak u izradi skladišta je izrada logičkog modela podataka u ovom konkretnom slučaju ERA modela.



Slika 1: ERA model skladišta podataka

Slika 1. prikazuje ERA model skladišta podataka koji je generiran nakon što su napravljene sve tablice s potrebnim atributima. Skladište je oblikovano prema modelu zvijezde u kojem se u sredini nalazi činjenična tablica, a oko nje se nalaze dimenzijske tablice, kojih je u ovom slučaju četiri. U činjeničnoj tablici *YouTubeTrendovi* spojeni su svi podaci i ona daje odgovor na pitanje što mjerimo. U ovom slučaju mjerimo, odnosno uspoređujemo broj pregleda, broj pozitivnih i negativnih reakcija na videozapise i broj komentara ostavljenih ispod videozapisa. Dimenzijske tablice podijeljene su na kanal, video, datum i državu. Dimenzija *Kanal* prikazuje nazive kanala s kojih su objavljeni najpopularniji videozapisi. Dimenzija *Video* prikazuje sve najpopularnije videozapise, odnosno njihovo ime, prikazuje kategorije u koje ti videozapisi spadaju i da li su komentari ispod videozapisa trenutno onemogućeni ili ne. Dimenzija *Datum* navodi dan, mjesec i godinu kada je određeni videozapis bio u trendu, a dimenzija *Drzava* prikazuje ime država, u ovom slučaju četiri države čije podatke promatramo.

4.3. Opis provedenog ETL procesa

Kako je već ranije spomenuto skup podataka je preuzet kao četiri csv datoteke koje su zatim otvorene u Excel-u i tamo uređene. Bilo je potrebno dodati u svaku tablicu jedan stupac koji će sadržavati ime države kako bi kasnije kod spajanja mogli razlikovati iz koje države su koji podaci i također je bilo potrebno zamijeniti ID kategorija s pravim nazivima kategorija koji su preuzeti iz .json datoteka. Nakon što su tablice bile spremne za korištenje, spremljene su ponovno u csv formatu kako bi ih mogli uvesti u SSMS. U SSMS-u su podaci uvezeni preko Import Flat File opcije, odnosno preko Export Wizard-a. Uvezeni skup podataka se tako sastojao od 4 tablice od kojih je svaka imala 14 stupaca i 6000 redaka. Radi lakšeg kasnijeg popunjavanja dimenzijskih i činjeničnih tablica ove 4 tablice s podacima su spojene u jednu novu tablicu YouTube koristeći naredbe INSERT INTO i SELECT te sada ta tablica ima 14 stupaca i 24000 redaka.

Kako su u tablici postojali duplicirani podaci poput imena videozapisa, imena kanala, država, datuma i kategorija prilikom popunjavanja dimenzijskih tablica je bilo potrebno osim naredbe INSERT INTO koristiti i naredbu SELECT DISTINCT, koja će spriječiti duple vrijednosti u tablicama. Sve dimenzijske tablice popunjene su na jednak način koji je prikazan na slici 2.

```
INSERT INTO Datum SELECT DISTINCT day, month, year FROM YouTube;
INSERT INTO Kanal SELECT DISTINCT channelTitle FROM YouTube;
INSERT INTO Drzava SELECT DISTINCT drzava FROM YouTube;
INSERT INTO Video SELECT DISTINCT title, categoryId, comments_disabled FROM YouTube;
```

Slika 2: Upiti za popunjavanje dimenzijskih tablica

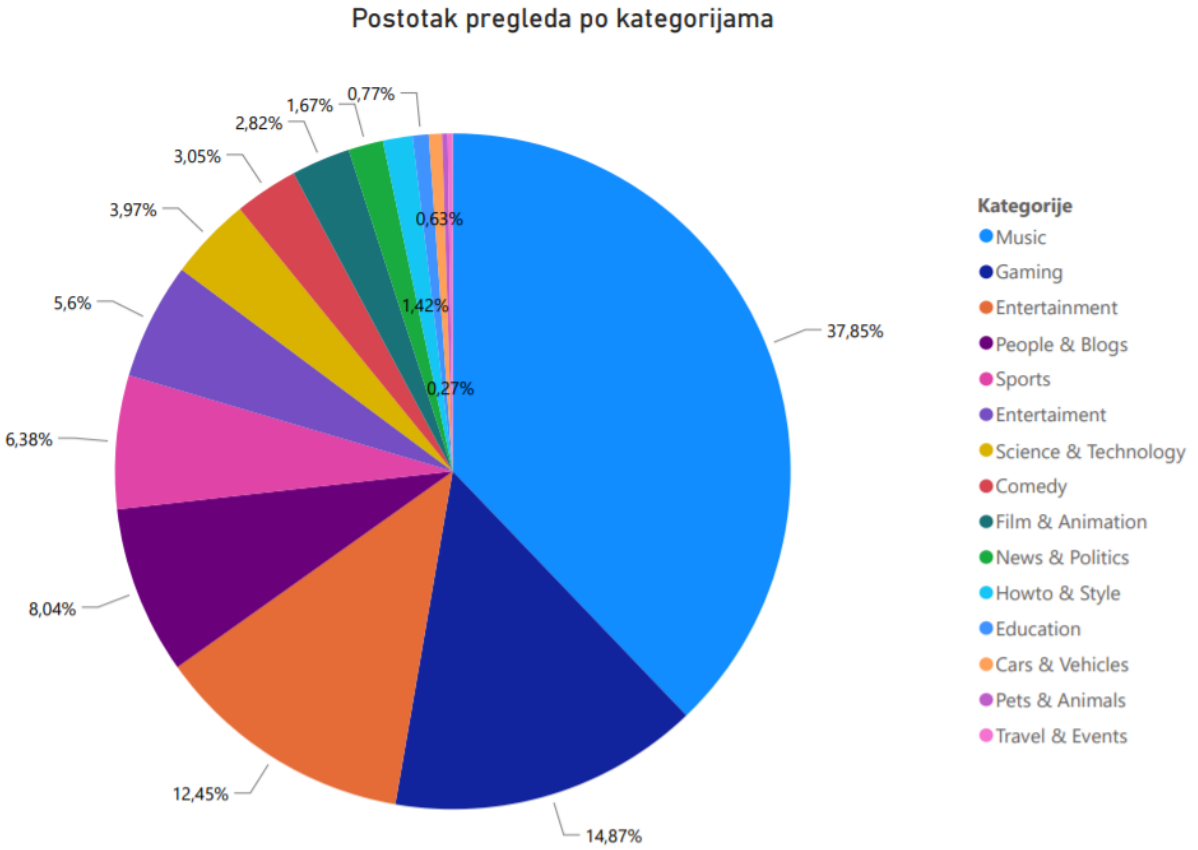
Nakon što su popunjene sve dimenzijske tablice možemo popuniti činjeničnu tablicu s podacima iz dimenzijskih tablica i iz izvorne tablice YouTube. Ovaj upit je malo složeniji i zahtijeva korištenje naredba INSERT INTO, SELECT i JOIN a prikazan je na slici 3.

```
INSERT INTO YouTubeTrendovi(datum_id,kanal_id,video_id,drzava_id,pregledi,pozitivne_reakcije,negativne_reakcije,broj_komentara)
SELECT Datum.idDatum, Kanal.idKanala, Video.idVideo, Drzava.idDrzava, YouTube.view_count, YouTube.likes, YouTube.dislikes, YouTube.comment_count
FROM YouTube
JOIN Datum ON (YouTube.day=Datum.dan AND YouTube.month=Datum.mjesec AND YouTube.year=Datum.godina)
JOIN Kanal ON (YouTube.channelTitle=Kanal.naziv)
JOIN Video ON (YouTube.title=Video.naslov AND Youtube.categoryId=Video.kategorija AND YouTube.comments_disabled=Video.onemoguceni_komentari)
JOIN Drzava ON (YouTube.drzava=Drzava.naziv);
```

Slika 3: Upit za popunjavanje činjenične tablice

5. Provedba analize podataka

Za analizu podataka kroz izradu izvještaja korišten je prethodno opisan alat Microsoft Power BI. Alat se može koristiti i online ali bilo ga je potrebno instalirati kao desktop verziju. Nakon instalacije potrebno je bilo uvesti sve podatke i tablice iz SSMA-a kako bi na temelju njih mogli izraditi izvještaje. Kako bi se dohvatili podaci potrebno je u Power BI-u odabrati opciju SQL Server i u novootvoreni prozor upisati poslužitelja i po potrebi ime baze podataka. Kada smo unijeli podatke, s desne strane alata nam se otvaraju sve tablice i njihovi atributi koje odabrana baza podataka sadrži. U nastavku će biti prikazani i opisani kreirani izvještaji.

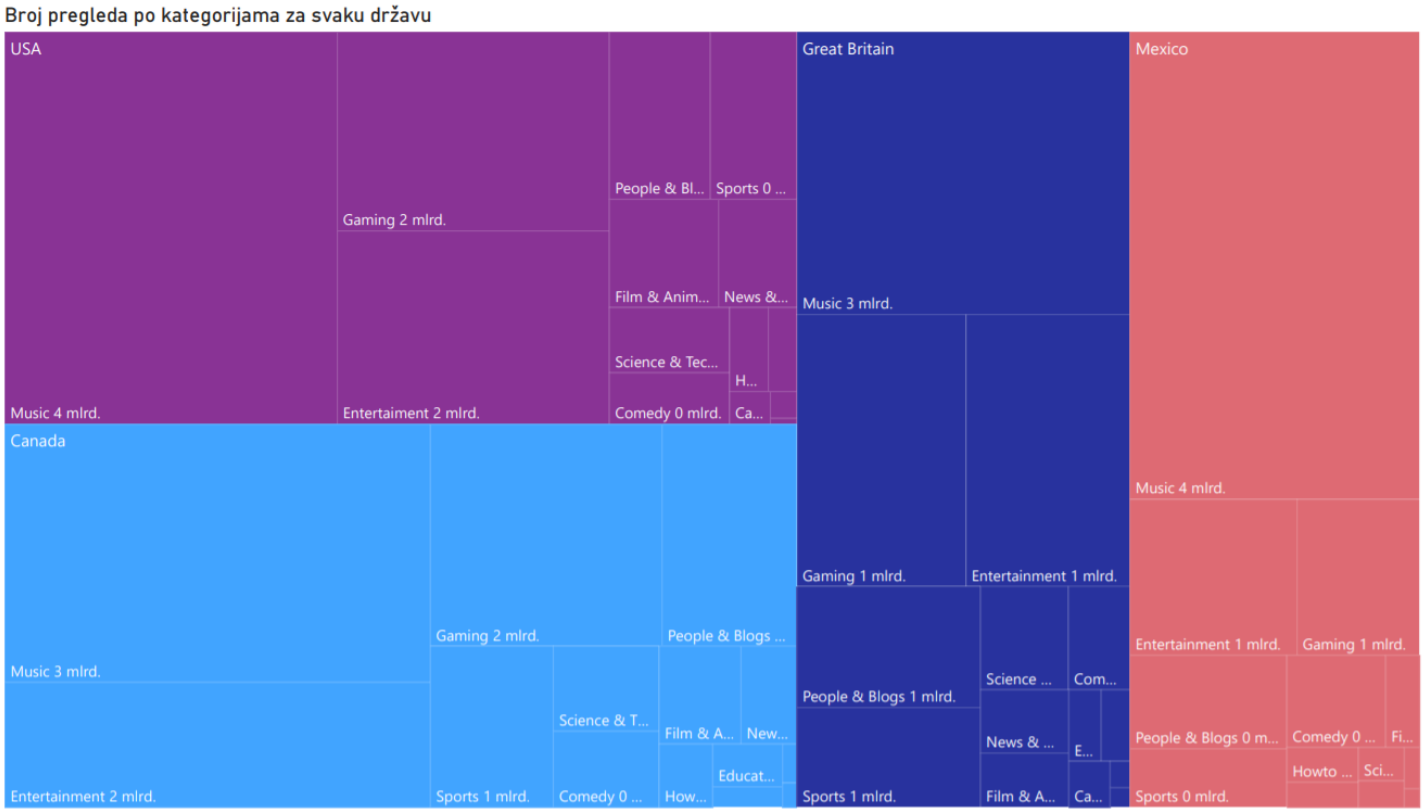


Slika 4: Izvještaj o broju pregleda po kategorijama (izraženo u postotku)

Prvi izvještaj na slici 4. prikazuje koje kategorije videozapisa imaju najviše pregleda, odnosno koje kategorije su najpopularnije. U ovaj izvještaj su uključeni podaci iz svih država te on prikazuje ukupan broj pregleda neke kategorije ali izražen u postocima radi jasnijeg prikaza odnosa kategorija. Kao što se vidi na prikazu, najviše pregleda imaju videozapisi koji se svrstavaju pod kategoriju glazbe i to skoro 40%. Na

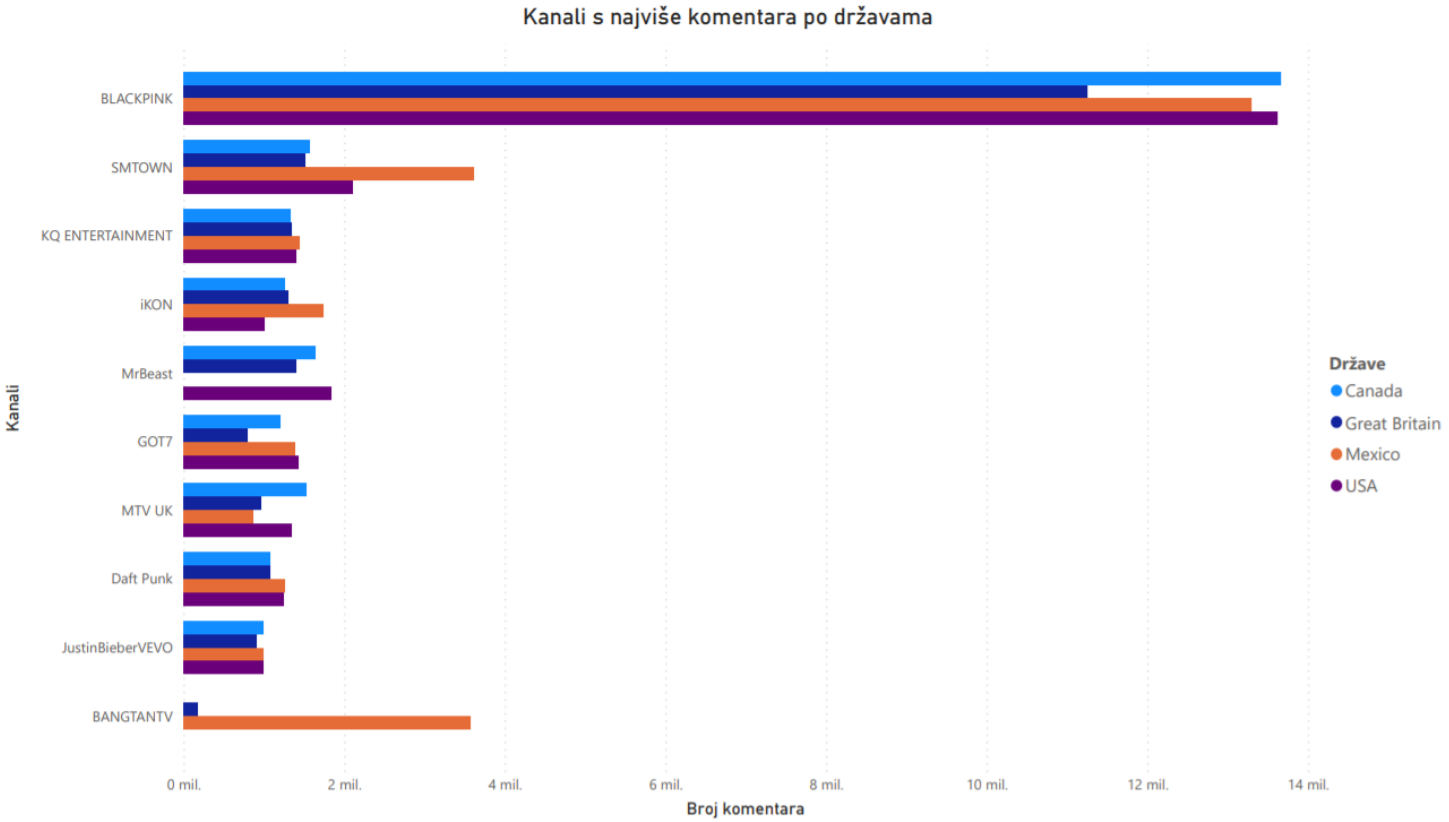
drugom mjestu s otprilike 15% su videozapisi iz kategorije videoigara a zatim s nešto manjim postotkom slijedi kategorija zabave. Najmanje pregleda imaju videozapisi iz kategorija automobila i vozila, ljubimci i životinje i putovanje i događaji.

Kako bi vidjeli da li se redoslijed mijenja ako promatramo svaku državu zasebno kreiran je još jedan izvještaj koji prikazuje preglede po kategorijama ali za svaku državu posebno. Taj izvještaj prikazan je na slici 5.



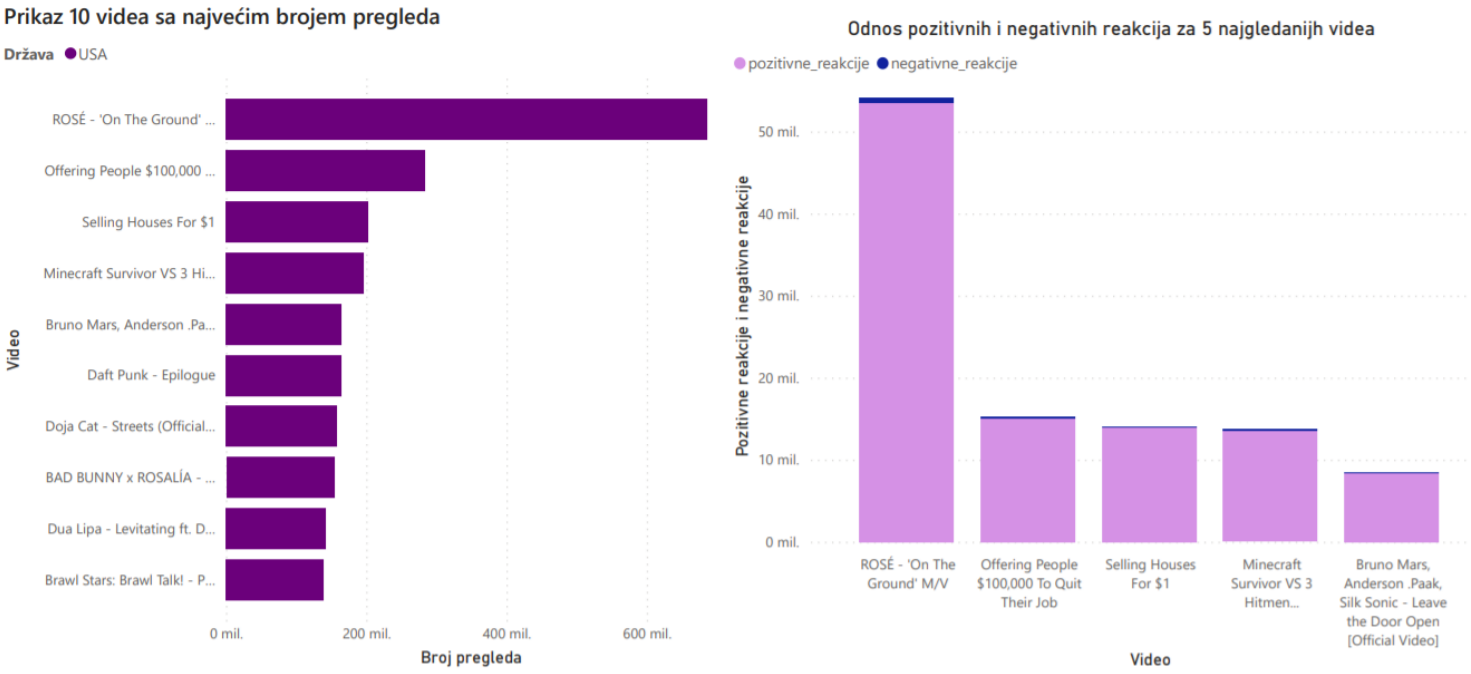
Slika 5: Izvještaj o broju pregleda po kategorijama za svaku državu zasebno

Izvještaj prikazuje 4 države koje se razlikuju po bojama, kategorije i broj pregleda po kategorijama koji su naznačeni u donjem dijelu svakog kvadrata. Možemo vidjeti kako se redoslijed nije značajno promijenio iz čega možemo zaključiti da je odnos broja pregleda po kategorijama u svim državama gotov jednak. U svim promatranim državama kategorija Glazba ima najviše pregleda, odnosno videozapisi iz te kategorije su najpopularniji.



Slika 6: Izvještaj o kanalima s najvećim brojem komentara po državama

Sljedeći izvještaj, prikazan na slici 4. prikazuje 10 kanala koji imaju najveći broj komentara. Komentari su podijeljeni po državama, odnosno izvještaj različitim bojama prikazuje broj komentara iz različitih država. Kao što možemo vidjeti najveći broj komentara ima kanal BLACKPINK koji je zapravo popularan u svim državama. U Kanadi i Sjedinjenim Američkim Državama ima najviše komentara ali isto tako u Meksiku i u Ujedinjenom Kraljevstvu ima uvelike više komentara nego li drugoplasirani kanal. Do velike razlike između prva dva kanala zapravo dolazi zbog toga jer je kanal BLACKPINK izbacio više videozapisa koji su dosegli veliku popularnost odnosno koji se nalaze u datasetu, a kanal SMTOWN nema toliki broj popularnih videozapisa ali ipak videozapisi koji jesu u podacima imaju jako velik broj komentara. Možemo vidjeti i kako kanal koji je posljednji na listi nema komentara iz SAD-a i Kanade ali broj komentara iz Meksika je dovoljno velik da se taj kanal nalazi u top 10.



Slika 7: Izvještaj o videozapisima s najvećim brojem pregleda i njihovim pozitivnim i negativnim reakcijama za 3. mjesec

Posljednji izvještaj, a možda i najbitniji prikazuje najpopularnije videozapise prema broju pregleda iz Sjedinjenih Američkih Država u 3. mjesecu 2021. godine. Broj pregleda i pozitivnih i negativnih reakcija nije trenutni broj koji taj videozapis ima, nego je to zbroj od različitih dana jer su videozapisi bili u trendu nekoliko dana, pa tako izvještaj pokazuje podatke iz svih dana, ne samo jedan dan. Sastoji se od dva dijela. Lijevi dio izvještaja prikazuje 10 videozapisa koji imaju najveći broj pregleda u SAD-u. Na prvom mjestu je ponovno videozapis iz kategorije Glazba i ima puno veći broj pregleda od drugoplasiranog videozapisa. Kako bi vidjeli kakav doživljaj ostavljaju najpopularniji videozapisi napravljen je desni dio izvještaja koji pokazuje odnos pozitivnih i negativnih reakcija na videozapise za prvi 5 najgledanijih videozapisa. Izvještaj pokazuje kako je puno više pozitivnih reakcija na svim videozapisima i u većini slučajeva je broj negativnih reakcija zanemariv.

6. Zaključak

Za ovaj projekt najvažnije je bilo pronaći pravi skup podataka koji bi se kasnije mogao urediti, prikazati kao model zvijezde i koristiti za analizu. Nakon pronalaska pravog skupa podataka isti je otvoren i uređen u Microsoft Excelu kako bi se mogao uvesti u SSMS. U SSMS-u se kreiralo skladište podataka čije su se tablice punile tim skupom podataka. Kreirane su tablice koje su četiri dimenzijske tablice i jedna činjenična tablica koje zajedno tvore model zvijezde. Nakon kreiranja i popunjavanja tablica preko različitih upita, podaci su spremni za analizu. Analiza se provodila u alatu Power Bi kreiranjem izvještaja. Alat Power Bi je zbog jednostavnog sučelja i raznih mogućnosti koje nudi bilo odlično koristiti za prikaz podataka i odnosa među podacima.

Analizom podataka prikazano je bilo koji su videozapisi najpopularniji, koji su kanali imali ukupno najviše pregleda, kakav je odnos pozitivnih i negativnih reakcija na videozapisima i koji kanali imaju najveći broj komentara. Pomoću izvještaja je zaključeno kako najviše pregleda imaju videozapisi iz kategorije Glazba i to u svim promatranim državama. Isto tako je prema pregledima utvrđeno da je kategorija Glazba i sveukupno najpopularnija odnosno videozapisi iz te kategorije su najgledaniji.

Popis literature

[1] Microsoft, SQL Server Management Studio (SSMS), 2021. godina, dostupno: https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15, preuzeto: 30.5.2021.

[2] Microsoft, Power BI, dostupno: https://powerbi.microsoft.com/en-us/desktop/, preuzeto: 30.5.2021.

[3] Kaggle, Trending YouTube Video Statistics https://www.kaggle.com/datasnaek/youtube-new?select=CA_category_id.json, preuzeto: 28.5.2021.

Popis slika