OSNOVE STATISTIČKOG PROGRAMIRANJA

Obrada skupa podataka: UFO Sightings

Projektni zadatak

2023/2024

Autori:

Luka Blagaić

Valentina Ivanić

Dominik Ožvald

Nikola Vlahović

Uvodna obrada i prilagodba podataka:

Podaci se nalaze u datoteci nazvanoj complete.csv, pri čemu su podaci odvojeni zarezima. Za učitavanje ovih podataka koristimo funkciju read\_csv(). Nakon inicijalnog učitavanja i pregleda dokumentacije povezane s ovim skupom podataka, primijetili smo da se u stupcu "comments" nalaze nizovi znakova poput "&#44", koji označavaju točke, zareze i druge simbole koji mogu izazvati probleme pri spremanju u formatu csv datoteke. Zamijenili smo ove nizove znakova razmakom koristeći regularne izraze.

Mnogi reci podataka nisu imali navedenu državu, stoga smo poduzeli sljedeće korake kako bismo odredili pripadnost državi za svaki podatak:

* Kreirali smo popis država u SAD-u i svaki redak koji ima vrijednost u stupcu "state" koja se nalazi unutar ovog popisa smo dodijelili SAD-u.
* Svaki redak koji u atributu "city" ima navedenu državu u zagradi smo pridružili odgovarajućoj državi (na primjer, "chester (uk/england)" smo dodijelili Velikoj Britaniji).
* Za sve preostale podatke na temelju atributa "city", odredili smo pripadajući grad i pridružili ga državi u kojoj se taj grad nalazi.
* Podatke koji nisu imali niti jedan od atributa "city", "state" ili "country" smo označili kao pripadajuće državi "Nepoznato".

Početno, atribut "country" je sadržavao kratice država ("us", "ca", "gb"…). Ove kratica smo zamijenili punim imenima država na hrvatskom jeziku. Sve navedene promjene smo pohranili u posebnu csv datoteku koju smo koristili za daljnju analizu podataka.

Daljnja obrada podataka, uključujući izlučivanje specifičnih podataka i izradu grafova, zahtijevala je dodatne korake, koji će biti opisani uz svaki pojedini graf.

Analiza podataka:

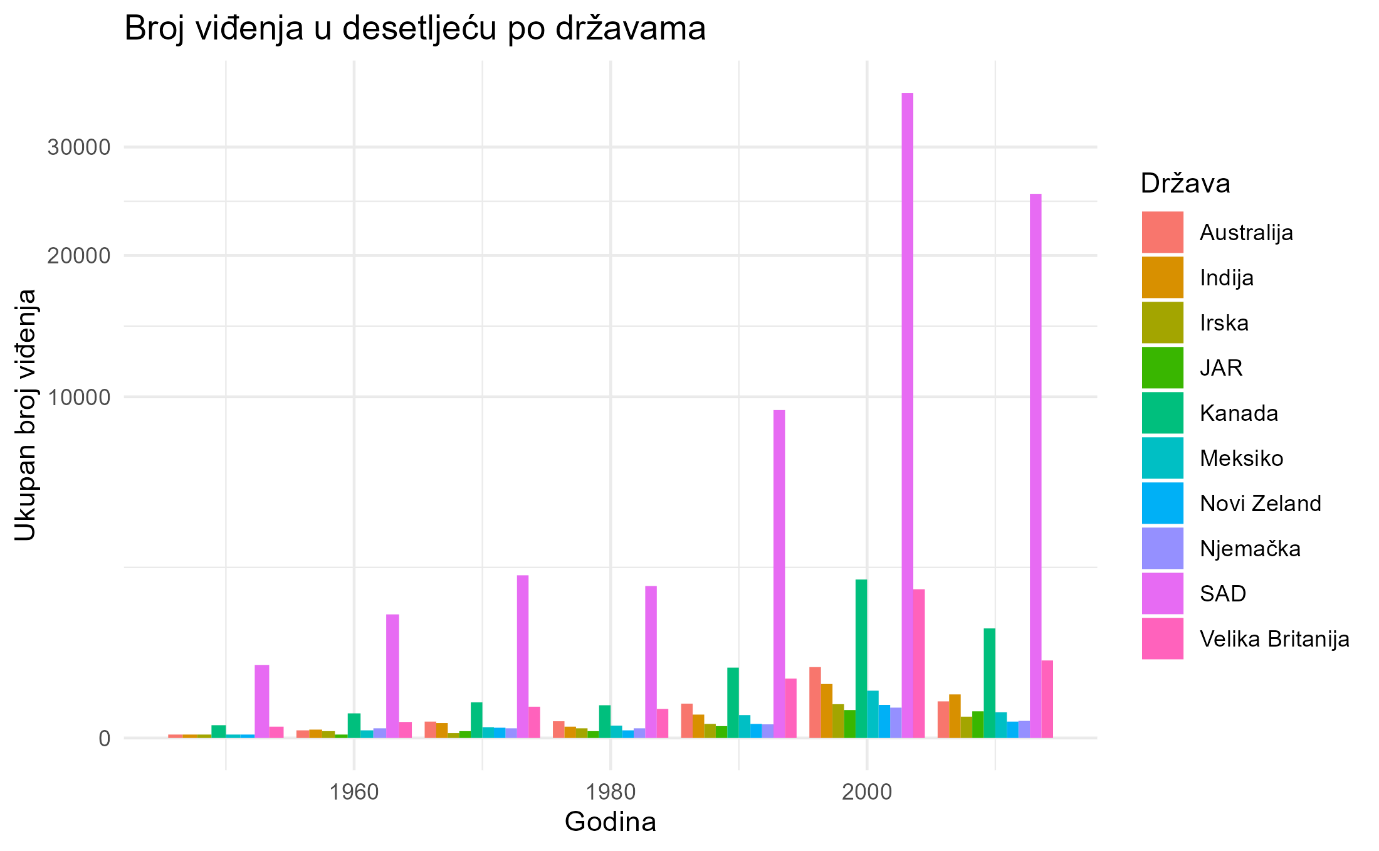
Prve pretpostavke koje smo željeli provjerit bile su da će broj prijavljenih UFO viđenja naglo rasti nakon što su dronovi postali lako dvostupačni i da je država s najviše prijavljenih viđenja SAD. Dronovi se u lošim uvjetima lako mogu zamijeniti za nepoznatu letjelicu. Da bi smo to provjerili stvorili smo graf(slika 1) koji prikazuje broj UFO viđenja svake godine za nekoliko različitih država. Na to me grafu očekujemo nagli rast oko 2000-te godine. A graph of a graph of a stock market

Description automatically generated with medium confidence

Slika Broj viđenja u godini po državama

Prikazano na graf(slika 1) možemo vidjeti da nije došlo do naglog rasta oko 2000-te godine nego je taj rast krenuo već u devedesetim godinama to nam je opovrgnulo prvu pretpostavku. Naša druga pretpostavka je točna SAD ima najviše viđenja.

Željeli smo provjerit kolika je rizika u broju viđenja u prošlih nekoliko desetljeća. Možemo vidjeti na sljedećem grafu(slika 2) kako od šezdesetih do osamdesetih godina broj viđanja ostaje malo promijenjen dok se u devedesetima naglo udvostručio, te nakon toga u dvije tisućitima utrostručio.



Slika Broj viđenja u desetljeću

Pretpostavili smo da broj viđanja 1969. kada je čovjek prvi put sletio na mjesec. Jer su ljudi u to vrijeme, pogotovo u SAD-u, bili jako zainteresirani za svemirsku tehnologiju. To bi moglo povećati vjerojatnost da se neki obični događaj interpretira kao nešto nepoznato iz svemira. Na grafu (slika 3) smo prikazali razdoblje od 1960-1980 kako bi provjerili ovu pretpostavku.

A graph with lines and lines

Description automatically generated

Slika Broj viđanja od 1960-1980

Začuđujuće broj viđenja u SAD-u pada dok u ostalim država dolazi do malih promjena. Lokani minimum je oko 1972, ali možemo vidjeti da je broj slučajeva počeo padati već prije 1969 što je suprotno od naše pretpostavke.