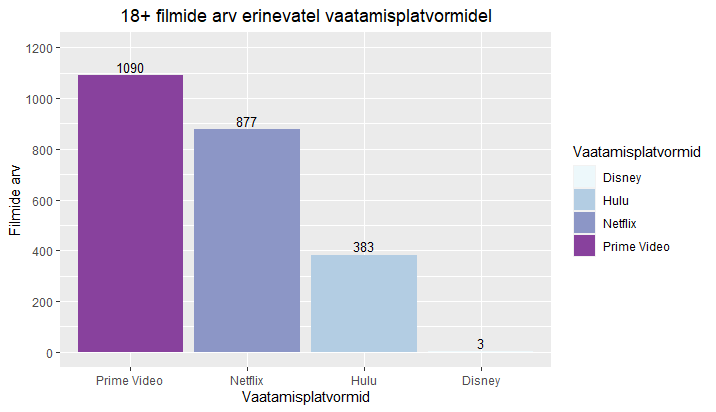
**ANDMESTIKU KIRJELDUS**

**Uuritud objektid**

Antud projektis on uuritavaks objektiks võetud andmestik, mis sisaldab andmeid filmivaatamis platvormide kohta. Andmebaas sisaldab 10 erinevat tulpa: indeks andmebaasis, filmi ID, filmi pealkiri, filmi tootmisaasta, vanusegrupp, IMDb hinnang, *Rotten Tomatoes* hinnang, Kas film on Netflix-is, Kas film on Hulu-s, Kas film on Prime Video-s, Kas film on Disney+-is. Enne hüpoteeside tõestamist, korrastame ka andmeid. Esialgu viime IMDb ning Rotten Tomates reitingud samale tasemele. Kuna IMDb reiting on 10-skaalas ja Rotten Tomatoes on protsentuaalne hinnang, on vaja saada reitingud samale vormile, et neid saaks uurida koos. Selleks korrutame IMDb reitingud 10-ga, et saada mõlemad 100-skaalale. Mõlemalt hinnangult eemaldame ka skaala osa, et numbritega saaks arvutada (x**/10** ; x**/100**). Eemaldame kõik tühjad väljad andmestikust ning teeme andmestiku kirjeldused. Andmestikus kirjeldused on järgnev info: Mis on kõige suurem keskmine reiting, Mis on keskmine filmide reiting, Mis on kõige madalam keskmine reiting, Mis on kõige pikem film, Mis on keskmine filmi kestvus ning Mis on kõige lühem film. Leidsime et kõige kõrgem keskmine reiting filmil oli 92,5%. Kõikide filmide keskmine reiting oli 60,3%. Kõige madalam keskmine reiting oli 23%. Pikim film oli 359 minutit ehk 5 tundi ja 59 minutit. Keskmine filmipikkus on 99 minutit ehk 1 tund ja 39 minutit. Ning lõpuks kõige lühem film oli 2 minutit pikk.

**Hüpotees 1:**

Esimeseks hüpoteesiks oli püstitatud väide, et Disney+ platvormil on kõige vähem täiskasvanutele mõeldud sisuga filme. Hüpoteesi kontrollimiseks filtreerisime andmebaasi, et alles oleks jäetud ainult read, kus filmi vanusehinnang on „18+“. Seejärel loetakse kokku, mitu filmi on igal vaatamisplatvormil alles jäänud. Saadud andmetest tehakse andmestik*,* mis tehakse ggplot2 abil tulpdiagrammiks.



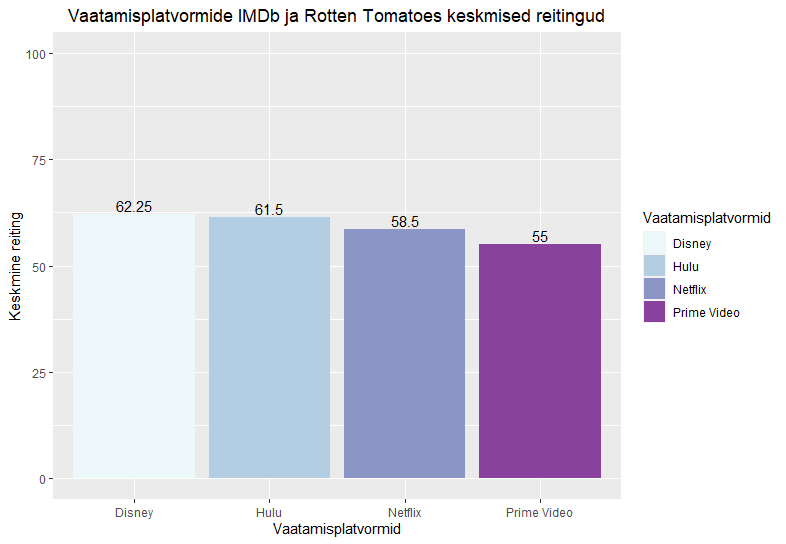
**Joonis 1. 18+ hinnanguga filmide arv erinevatel platvormidel**

Nagu me ka arvasime, pidas hüpotees paika nagu on näha joonisel 1. Arvame, et hüpotees pidas paika, sest Disney+ on tuntud kui kogupereplatvorm, mistõttu ei ole 18+ filmid sobilikud. Kuna Disney+ sisaldab vähem filme ning sarju kui teised platvormid, sest nad näitavad ainult enda toodangut, on väga tähtis, et vähest valikut vaadatakse võimalikult palju.

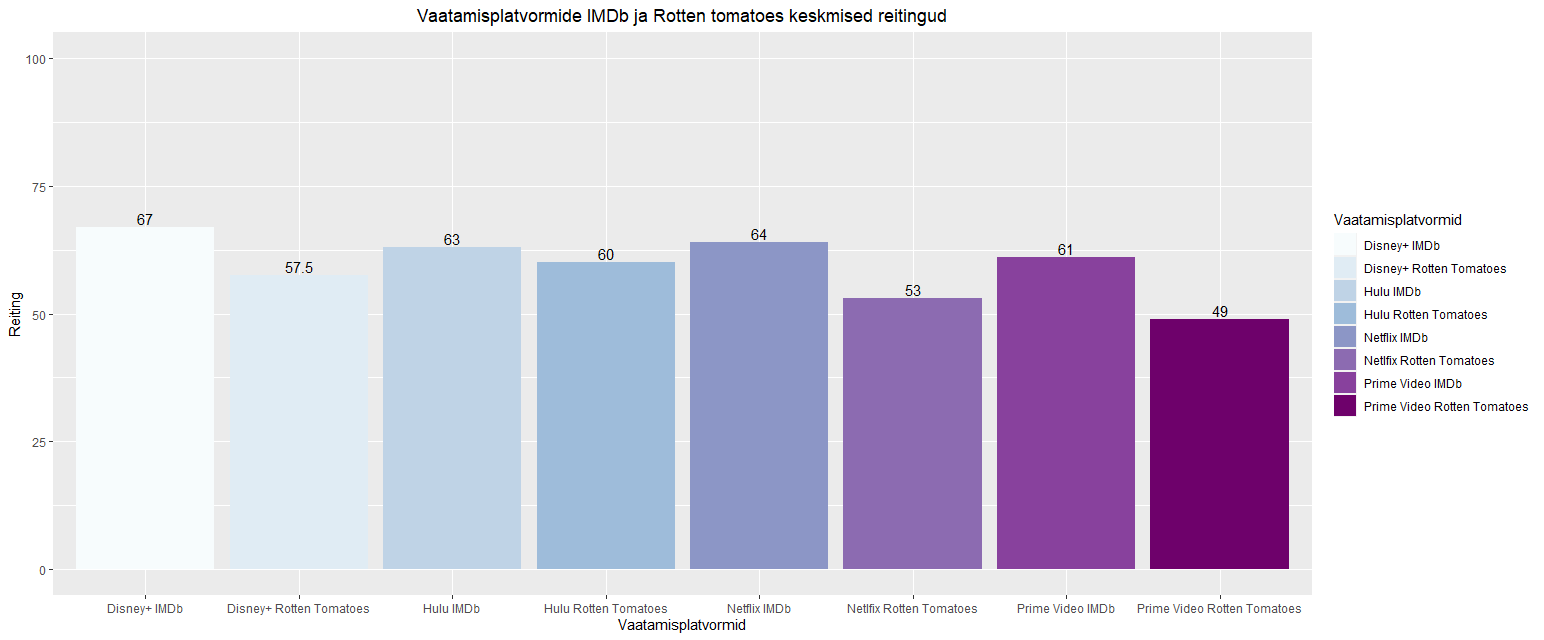
**Hüpotees 2:**

Teiseks hüpoteesiks on „Netflix-i filmivaliku keskmine *rating* on suurem kui muudel vaatamisplatvormidel.“ Hüpoteesi testimiseks valisime välja andmestikust järgmised tulbad: IMDb (filmi IMDb hinnang), Rotten Tomatoes(filmi Rotten Tomatoes hinnang), Netflix, Hulu, Prime Video, Disney+(Viimased neli tulpa ütlevad, kas antud film on sellel platvormil leitav või mitte). Seejärel tegime funktsiooni, mis leiab filmi keskmise hinnangu IMDb ning Rotten Tomatoes vahel. Funktsioon arvutab IMDb keskmise reitingu sisseantud platvormil, Rotten Tomatoes reitingu sisseantud platvormil ning mõlema allika keskmise reitingi sisseantud platvormil. Kasutame hinnangute keskmise leidmisfunktsiooni igal vaatamisplatvormi tulbal, et leida iga platvormi keskmise hinnangu. Seejärel väljastame graafi.

**Joonis 2. Vaatamisplatvormide keskmised reitingud (IMDb ja Rotten Tomatoes kokku)**



**Joonis 3. Vaatamisplatvormide keskmised reitingud (IMDb ning Rotten Tomatoes eraldi)**

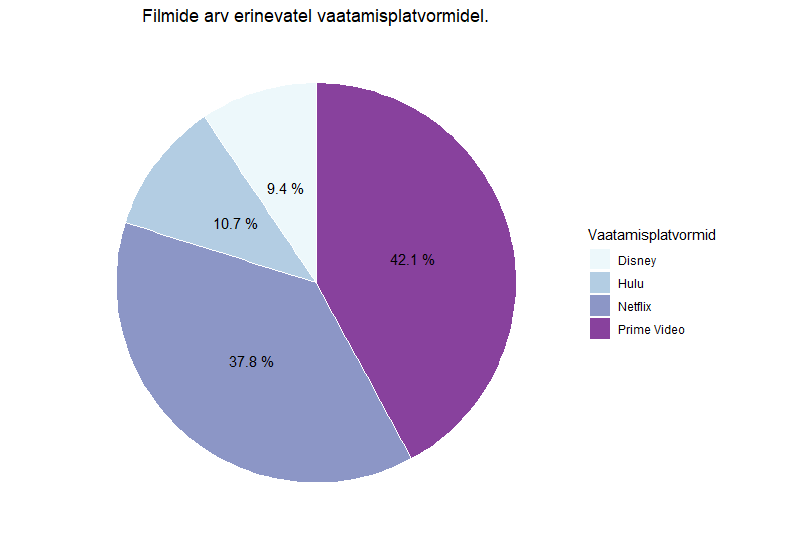


Seekord ei pidanud meie püstitatud hüpotees paika. Meie hüpotees oli, et Netflix on kõige kõrgema reitinguga, kuid tegelikult ei ole nii. Kui kokku panna mõlema lehe reitingud, mille tulemust on näha joonisel 2, on Netflix lausa kolmandal kohal. Kuid kuna IMDb kogub kasutajaandmeid reitinguteks ja Rotten Tomatoes on kriitikute hinnangud, otsustasime uurida reitinguid eraldi. Kuid isegi eraldi uurides, ei ole kummalgi alal Netflix esimesel kohal nagu on näha joonisel 3.

**Hüpotees 3:**

Viimaseks hüpoteesiks püstitasime väite, et Netflix-i filmivalik on kõige suurem neljast vaatamisplatvormist. Kontrollimiseks valisime andmestikust ainult tulbad, mis sisaldavad andmeid filmi platvormi asukohast. Lisaks eemaldasime tühjad väljad (na). Puhastatud andmed teisendasime laiast formaadist pikka formaati, kasutades melt() käsku. Kuna filmide arvu platvormil leidmiseks peame tegema sama asja mitme tulbaga, tegime selleks ka vastava funktsiooni. Funktsioon võtab sisse andmestiku kuhu andmed lisame ning tulbanime, kust tahame lugeda kokku, mitu filmi on antud tulbas. Antud andmestikku kasutasime uue andmestiku tegemiseks, et joonise tegemine oleks lihtsam. Viimase andmestikuga tegime sektordiagrammi andmete illustreerimiseks.

**Joonis 4. Filmide arv erinevatel platvormidel**



Andmestikus on kõikide platvormide peale kokku 9777 filmi, millest 42,1% filmidest on Amazon Prime Video-s saadaval, 37,8% Netflixi-s, 10,7% Hulu-s ning 9,4% Disney+-is. Arvuliselt, pärast andmetepuhastust on Prime Video-s 4113, Netflix-is 3695, Hulu-s 1047, Disney+-is 922 filmi. Protsendilised väärtused on arvutatud kokku filmikogusest. Jällegi, ei olnud hüpotees tõene, kuna suurim filmivalik on Amazon Prime Video-s. Arvestades ka eelnevat hüpoteesi, võib järeldada, et Amazon Prime Video puhul on tegu *quantity over quality* ehk kvantiteet oli tähtsam kui kvaliteet. Kuigi Prime Video-s on kõige rohkem filme, on aga igal alal Prime Video valiku keskmine reiting kõige madalam.

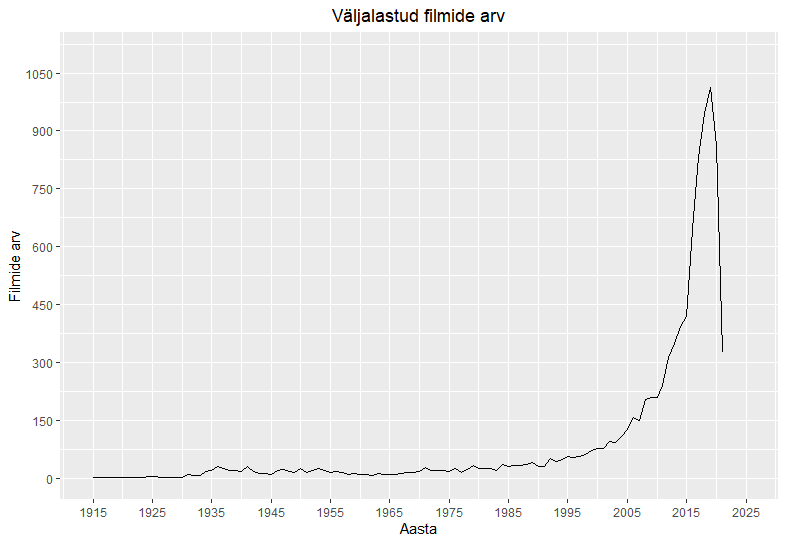
**Lisauuringud**

Projekti käigus tekkisid meil ka muud huvitavad küsimused, millele tahtsime vastuseid saada.

**Filmide väljalaskmine läbi aastate**

Kuna andmestikus olevad filmide seas on ka filme, mis on nüüdseks üle 100 aasta vanad, tahtsime uurida, kuidas on filminduse trend muutunud läbi aastate. Selleks valisime välja algandmestikust aastaarvud ning sorteerisime kasvujärjekorras. Lisaks eemaldasime kõik tühjad väljad. Kasutades neid andmeid koostasime joondiagrammi

**Joonis 5. Väljastatud filmide arv läbi aastate**

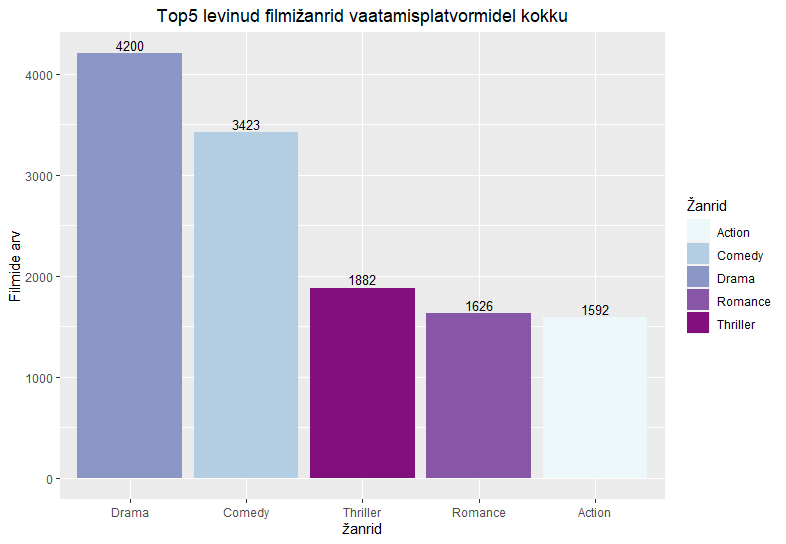


Diagrammilt on näha, kuidas filmide väljatulek on pidevalt tõusutrendis olnud. Selline kasv on ilmselge, kuna filmindus on väga suur meelelahutusosa. 21. sajandi alguses toimub väga järsk tõus. Tõus toimub selle tõttu, et väga vanad filmid ei kutsu väga suurt vaatajaskonda. Lisaks on selleks ajaks tehnoloogia areng nii kaugel, et filme üleüldiselt toodetakse rohkem. Samuti on diagrammil näha ka praegu väga aktuaalset teemat, COVID-19 pandeemiat. Pandeemia tõttu on filmindus pausile pandud, mistõttu viimastel aastatel ei ole nii palju filme tehtud. Lisaks ei ole nii palju uusi filme andmestikus, kuna uued filmid on kas alles kinos või vaatamisplatvormid pole veel saanud endale uute filmide litsentse, et neid näidata.

**Levinuimad žanrid vaatamisplatvormidel**

Lisaks filmide väljastamisaastale, huvitas meid ka filmižanri valik vaatamisplatvormidel. Tahtsime teada saada, millised filmižanrid on kõige levinumad vaatamisplatvormidel. Sellele küsimusele vastuse leidmiseks oli vaja andmestiku andmeid töödelda. Kuna enamusel filmidest on mitu žanri, eraldasime komakoha pealt žanrid ning lisasime need listi. Saadud listiga koostasime ka andmestiku, kus lugesime kokku iga žanri esinemiskordi. Nende andmetega tegime ka diagrammi, et visuaalselt näha ka 5 kõige levinumat žanri.

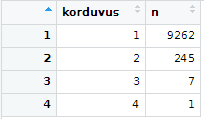
**Joonis 6. Top 5 levinumat filmižanri vaatamisplatvormidel**



Tulemusena leidsime, et kõige levinum žanr on draama. Tulemus on väga mõistetav, sest enamus filme on osaliselt draamad ning seda on näha ka filmikirjeldustel (Comedy, Drama ; Action Drama ; Romance Drama jne). Mis meid isiklikult hämmastas on see, et trillerfilme on rohkem kui romantikat.

**Filmide korduvus erinevatel platvormidel**

Andmeid uurides, tekkis ka küsimus, kui palju filme kattuvad erinevatel platvormidel. Selleks tegime tabeli, kus vaatasime üle kõik filmid ning mitmel platvormil nad on.



**Joonis 7. Filmide kattuvus**

Nagu näha jooniselt 7, valdav enamus filme on ainult ühel platvormil. Väike kogus filme on ka kahel platvormil korraga ning ainult üks film on leitav kõikidel platvormidel. Selleks filmiks on 2014 aasta “Zapped”, mis on koguperefilm, mille IMDb reiting on 5.2 ning Rotten Tomatoes on 59%.