

Datavisualisatie projectverslag

Floris Kornelis van Dijken & Ruben
Vandamme - groep 9

1e Master Informatica

19 mei 2025



Introductie:

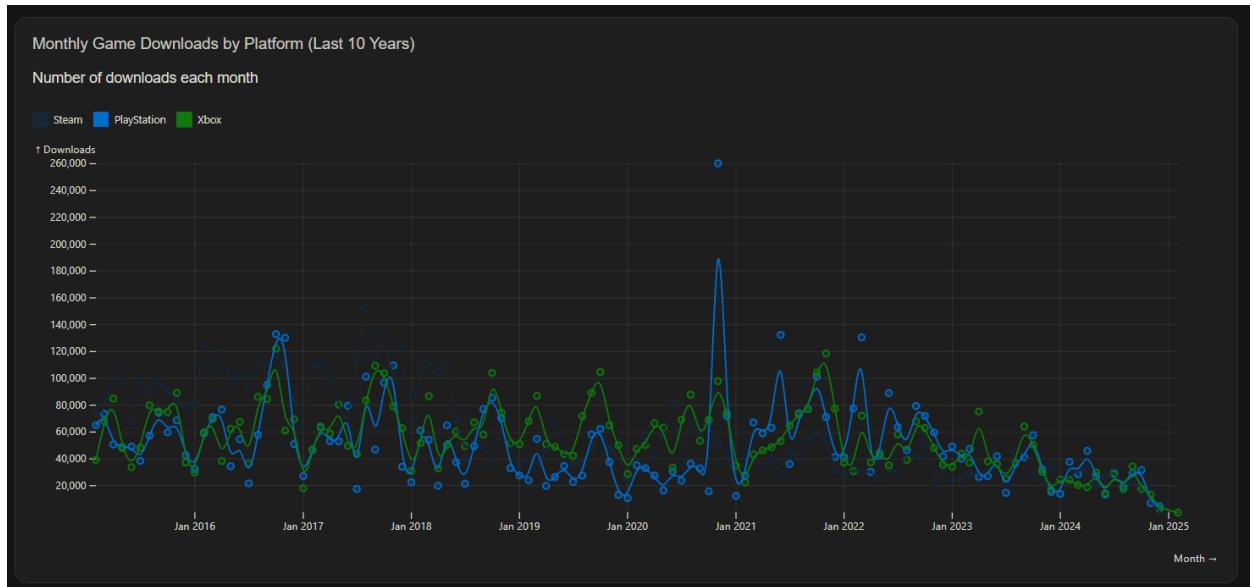
Voor onze datavisualisaties hebben we gekozen voor de Gaming Profiles dataset. Deze dataset was onze favoriet van de drie voorgestelde, omdat:

- Het real-world data is.
- Er erg veel data is (~60GB).
- Er veel mogelijk te onderzoeken dimensies zijn (user activity, rankings, price changes, historical data, ...).

Aanvankelijk wilden we onze visualisaties vooral richten op game developers, zodat zij op basis van de inzichten slimme beslissingen over hun games konden nemen (zie appendix). Tijdens het verkennend werk merkten we echter al snel dat er weinig opvallende conclusies te trekken vielen en dat de dataset bepaalde verbanden niet kon leggen die we aanvankelijk wel verwachtten, zoals een logische koppeling van landen op basis van vrienden. Door deze beperkingen is onze webpagina uiteindelijk meer algemeen informatief van aard geworden, met een nadruk op het gedrag van gamers.

Toen bleek dat de beoogde meerdere invalshoeken geen extra meerwaarde boden, ontstonden er nog andere problemen. Na cross-validatie met externe bronnen kwam aan het licht dat de dataset nauwelijks representatief is voor de werkelijkheid: veel van de genoemde 'spellen' blijken geen echte games te zijn en het lijkt alsof de Verenigde Staten de grootste gebruikersbasis vormen, terwijl China in de praktijk veel groter zou moeten zijn. Daarnaast werkte ons uitgangspunt dat een grote dataset automatisch tot nauwkeurigere resultaten leidde niet in ons voordeel. Met een omvang van ruim 60 GB laden sommige visualisaties, zoals de scatterplot, onvermijdelijk traag. Hoewel we waar mogelijk hebben geoptimaliseerd, bleek dit niet altijd toereikend om de performance op peil te houden.

What do gamers play on?



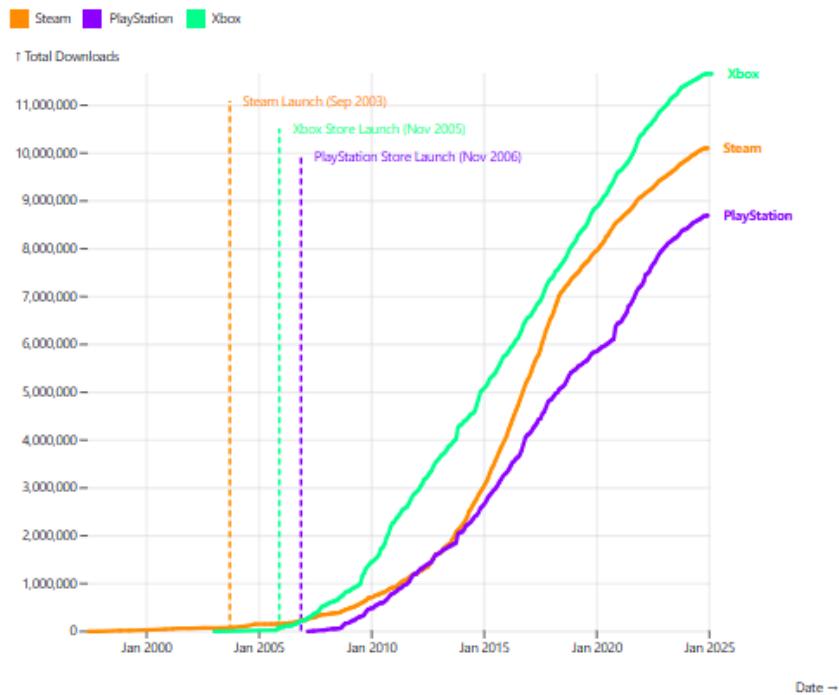
Oorspronkelijk bevatte deze pagina een grafiek die het aantal downloads per maand weergaf, zoals ook zichtbaar is in de bovenstaande afbeelding. We hebben er echter voor gekozen deze te vervangen door een cumulatieve visualisatie, omdat de oorspronkelijke grafiek een vertekend beeld gaf van de werkelijkheid. De dataset bevat namelijk geen informatie over het exacte moment waarop een game wordt gedownload; enkel de releasedatum en het totale aantal downloads per game zijn beschikbaar.

Aanvankelijk werd dit opgelost door alle downloads toe te wijzen aan de releasedatum van de betreffende game. Deze aanpak bleek echter misleidend, aangezien het impliceert dat alle downloads plaatsvinden op het moment van release, wat in de praktijk zelden het geval is. Om dit probleem te mitigeren, is gekozen voor een cumulatieve benadering waarbij het totale aantal downloads in de tijd wordt opgebouwd. Daarnaast is de tijdsinterval van deze grafiek verlengd om het verloop realistischer te kunnen weergeven.

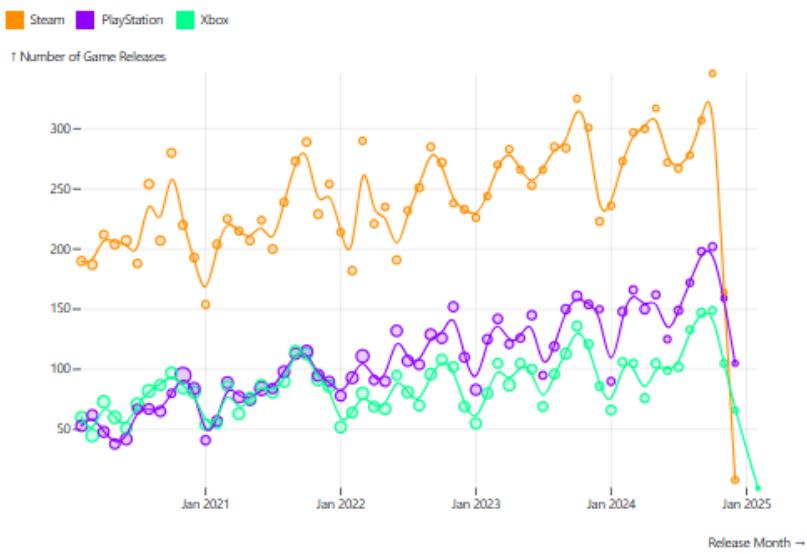
Een tweede grafiek op deze pagina toont het aantal game-releases per maand; een eigenschap dat wel betrouwbaar in de dataset aanwezig is. De grootte van de bolletjes in de grafiek represeneert het gemiddelde aantal downloads van games die in die maand zijn uitgebracht, waarvoor eveneens voldoende gegevens beschikbaar zijn. Tot slot is ook de releasedatum van de verschillende platformen toegevoegd. Dit leek ons een relevant en nuttig datapunt om uit te lichten.

De finale versie ziet er als volgt uit:

Cumulative game downloads



Monthly game releases



Where are gamers?

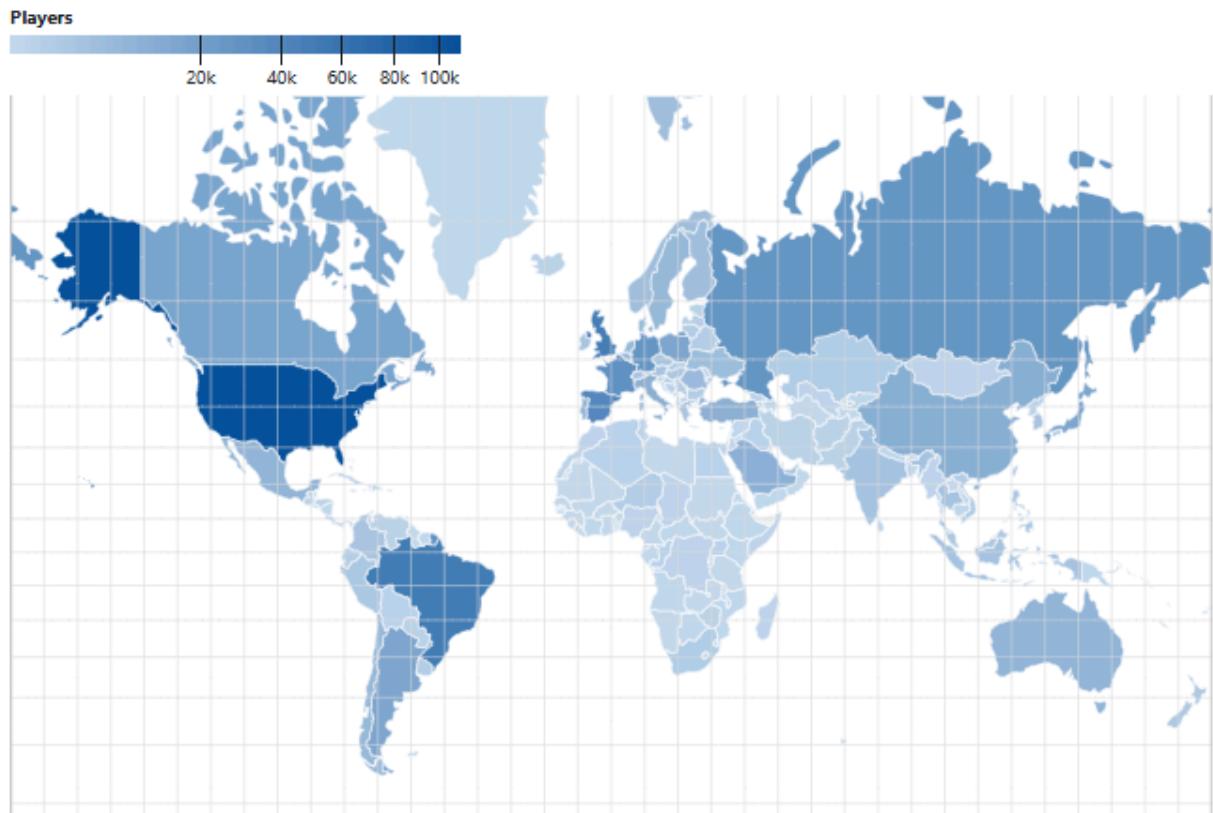
Voor deze visualisatie was al snel duidelijk dat we een choropleth map wilden gebruiken om de geografische spreiding van de data inzichtelijk te maken. Na enkele technische beperkingen met Vega-Lite besloten we over te stappen op een geo plot van Observable. Daarbij liepen we echter tegen het probleem aan dat de landnamen in onze dataset niet overeenkwamen met de benamingen die Observable hanteert, zoals zichtbaar is in de onderstaande afbeelding. Merk op dat de Verenigde Staten grijs weergegeven wordt.



Om dit op te lossen hebben we een mapping opgesteld die de landnamen uit onze dataset vertaalt naar de namen die door de geo plot worden herkend. Ter aanvulling hebben we dezelfde gegevens ook weergegeven in de vorm van een bar chart, zodat exacte waarden eenvoudiger af te lezen zijn.

Tijdens het analyseren van deze gegevens ontdekten we bovendien een scheefheid in de dataset, die we ook op onze website kort toelichten.

De finale versie ziet er als volgt uit:



What do gamers think?

Voor de pagina rond reviewanalyse moesten eerst enkele praktische obstakels worden overwonnen. Een eerste uitdaging was dat de dataset reviews bevatte in meerdere talen. Om de analyse consistent te maken, hebben we ervoor gekozen om uitsluitend Engelstalige reviews te behouden.

Na deze filtering bleek het meest voorkomende woord in de dataset “the” te zijn; een veelvoorkomend, maar inhoudelijk betekenisloos woord in deze context. Daarom voegden we een tweede filter toe die enkel game-gerelateerde termen behield. Dit resulteerde echter in “game” als meest frequente woord, wat evenmin relevante inzichten opleverde.

De uiteindelijke doorbraak kwam toen we besloten ons te richten op combinaties van bijvoeglijk naamwoord en zelfstandig naamwoord. Deze aanpak leverde betekenisvollere patronen op, die we visualiseerden in de vorm van een word cloud.

Ter aanvulling hebben we, net als bij de choropleth map, ook een bar chart toegevoegd waarin de exacte frequenties van de gebruikte woordcombinaties terug te vinden zijn.

De finale versie ziet er als volgt uit:



How do gamers spend? - Game yields:

Een eerste versie zag er als volgt uit:



Deze plot werd oorspronkelijk ingevoerd als onderdeel van de inleiding, om game developers een duidelijker beeld te geven van twee cruciale metrics: de prijs en het aantal downloads. Later hebben we, om een sterkere verhaallijn te creëren en te focussen op gamer behaviour, de titels van de assen aangepast naar specifieke vragen. Deze plot is sindsdien onderdeel geworden van de vraag: "*How do gamers spend?*"

Eerste indruk en top 5-labels

Op het eerste gezicht vertelt deze grafiek weinig; de gametitels zijn pas zichtbaar bij hoveren. Daarom heb ik de vijf populairste titels permanent op de grafiek geplaatst. Omdat de namen veel overlapten, heb ik de labels afwisselend rechts en links van de datapunten geplaatst. Dit werkt vooral goed nadat de top 5 is gerangschikt, zodat dicht opeenvolgende punten niet met elkaar botsen.

Van "Profit" naar "Yield"

Na feedback van de assistent bleek het woord "*Profit*" niet helemaal accuraat: de kosten waren immers niet meegenomen in de berekening. Daarom is de term gewijzigd naar "*Yield*" (opbrengst), wat beter overeenkomt met het onderliggende concept.

De rode lijn: gemiddelde opbrengst

Een bijkomende vraag die voortkwam uit de feedback was: "Wat stelt die rode lijn voor?" Hoewel dit in de begeleidende tekst werd toegelicht, hoort het visueel direct duidelijk te zijn. Daarom heb ik een label toegevoegd met de tekst "*Average yield*", inclusief de numerieke

waarde. Dit onderstreept dat het om een constante referentielijn gaat én benadrukt het feit dat dit cijfer op zichzelf ook een interessant inzicht biedt.

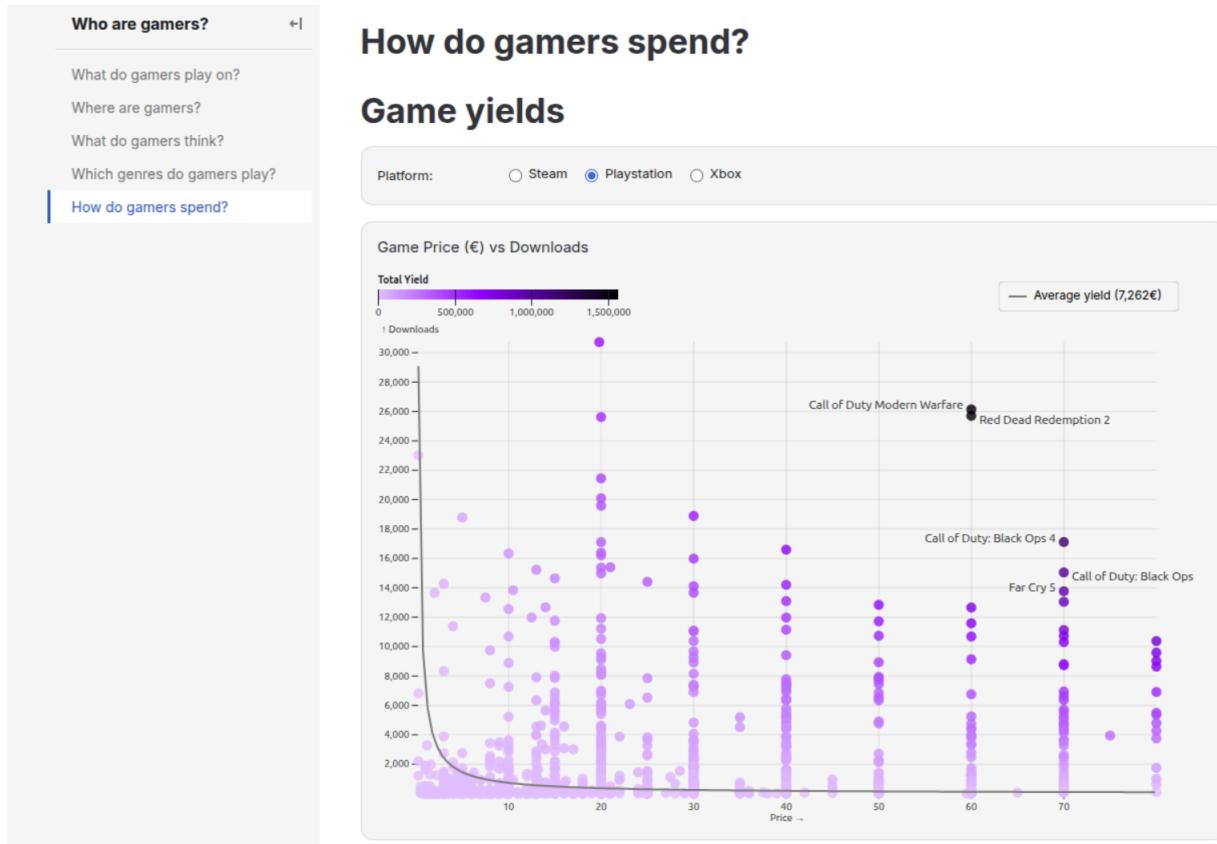
Platform-filter

Aanvankelijk gebruikten we alle games gezamenlijk in één grafiek. Om echter verschillen tussen platforms zichtbaar te maken, hebben we een radiobutton-selector toegevoegd. Hiermee kunnen gebruikers de plot filteren op platform. Deze interactiviteit leidde tot verrassende inzichten in platform-specifiek gedrag zoals de eerder besproken gemiddelde opbrengst.

Datacleaning en performance

- **Clustering van kleine punten:** Er zaten zeer veel datapunten met een lage opbrengst, waardoor de visualisatie geclutterd raakte en langzaam laadde. Met een kansberekening (kleine opbrengst grotere kans dan grote opbrengst) hebben we willekeurig een groot deel van deze punten verwijderd, zodat alleen de statistisch representatieve steekproef overblijft. De rode gemiddelde-lijn is echter nog altijd berekend over de volledige dataset, om vertekening te voorkomen.
- **Clipping van de assen:** Er waren enkele outliers met kostprijs >100€, maar met vrijwel geen downloads. Om zulke extreme waarden de leesbaarheid van de rest niet te laten beïnvloeden, hebben we de assen afgekapt waar nodig.

De finale versie ziet er als volgt uit:



Which genres do gamers play? - Correlation between genres:

Een eerste versie zag er als volgt uit:



Voor deze grafiek werd per genre een dictionary bijgehouden met als keys opnieuw alle genres en als values een teller voor co-occurrence. Concreet:

- Voor elke game is er een lijst van bijbehorende genres.
- Voor elk genre in die lijst worden voor alle andere genres (*) uit dezelfde lijst de corresponderende teller in de dictionary verhoogt.

Diagonaal en correlatie

Een probleem was dat de diagonaal geen maximale correlatie toonde, terwijl elk genre per definitie voor zichzelf 100% co-occurrence zou moeten hebben. Oorzaak: we negeerden bij (*) de eigen genre-combinaties. Door simpelweg alle genres, inclusief het huidige, mee te nemen in de iteratie, komt de diagonaal uit op de maximale waarde. Echter, aangezien deze maximale waarden voor verschillende genres kunnen verschillen, heeft voorlopig nog niet elke cel dezelfde kleur.

Absoluut vs. relatief

De oorspronkelijke tellers waren absolute waarden, maar we willen relatieve frequenties. Daarom delen we voor elke cel de co-occurrence door het totaal aantal games van de corresponderende genre (y-as). Dit maakt de matrix asymmetrisch en onthult extra verbanden. De diagonaal kleurt nu wel uniform als 100%; geheel in lijn met de co-occurrence van een genre met zichzelf.

Interactieve tooltips

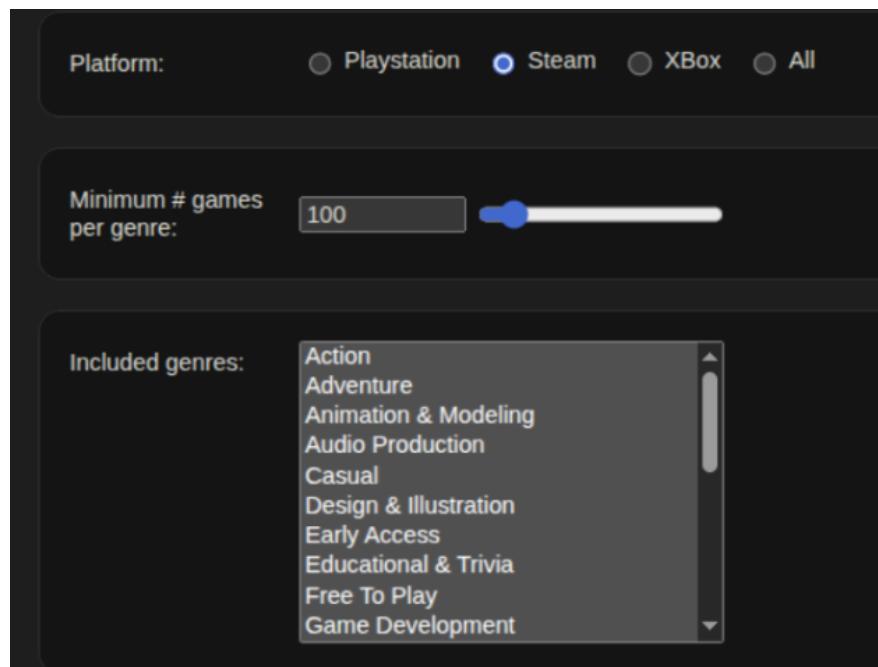
Een co-occurrence matrix is complex voor wie er niet in thuis is. Daarom hebben we bij hoveren een tooltip toegevoegd met een heldere uitleg, bijvoorbeeld:

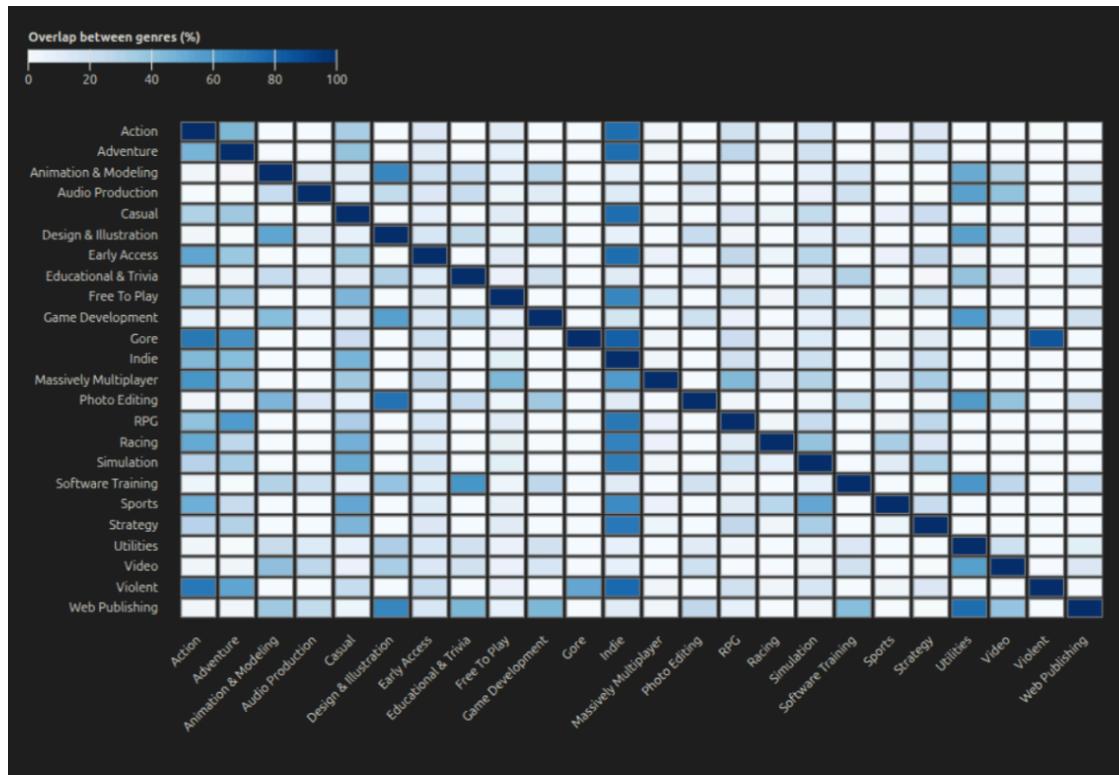
-> *Racing games die ook Simulation zijn: 32,1%*

Overzicht bij veel genres

Er zijn te veel genres om ze allemaal overzichtelijk weer te geven. Daarom:

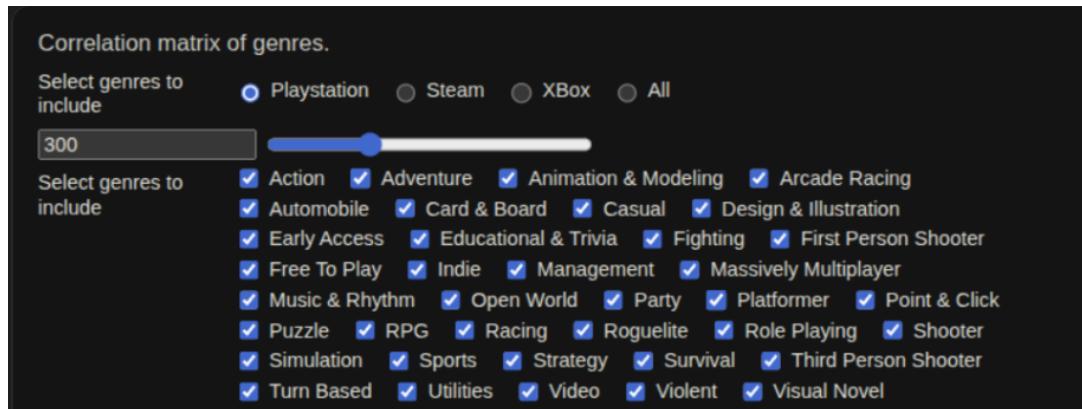
- **Schuine x-labels:** De genre-namen lopen niet meer in elkaar over, waardoor er meer labels passen.
- **Minimum-drempel slider:** Gebruikers bepalen hoeveel games een genre minimaal moet bevatten om te worden weergegeven. Zo verbergen we zeldzame genres tenzij explicet gewenst.
- **Genre-selector:** Voor gericht onderzoek kunnen gebruikers één of meerdere genres selecteren om specifieke co-occurrence verbanden te analyseren zonder cluttering.
- **Sorteerfunctionaliteit:** Genres zijn geordend op naam, zodat men snel het gewenste genre vindt.





UX-verbeteringen

- Checkbox-grid in plaats van dropdown:** Checkboxes tonen in één oogopslag welke genres actief zijn, zonder scrollen. Door ze in een grid te plaatsen blijft het overzichtelijk, zelfs bij veel opties.
- Invoerveld voor minimum & verduidelijking ervan:** Een invoerveld i.p.v. invoerveld + slider maakt de vele opties compacter. Het daadwerkelijke percentage games dat na filtering wordt meegenomen, maakt de impact van de gekozen drempel onmiddellijk inzichtelijk en wordt onder de grafiek weergegeven.

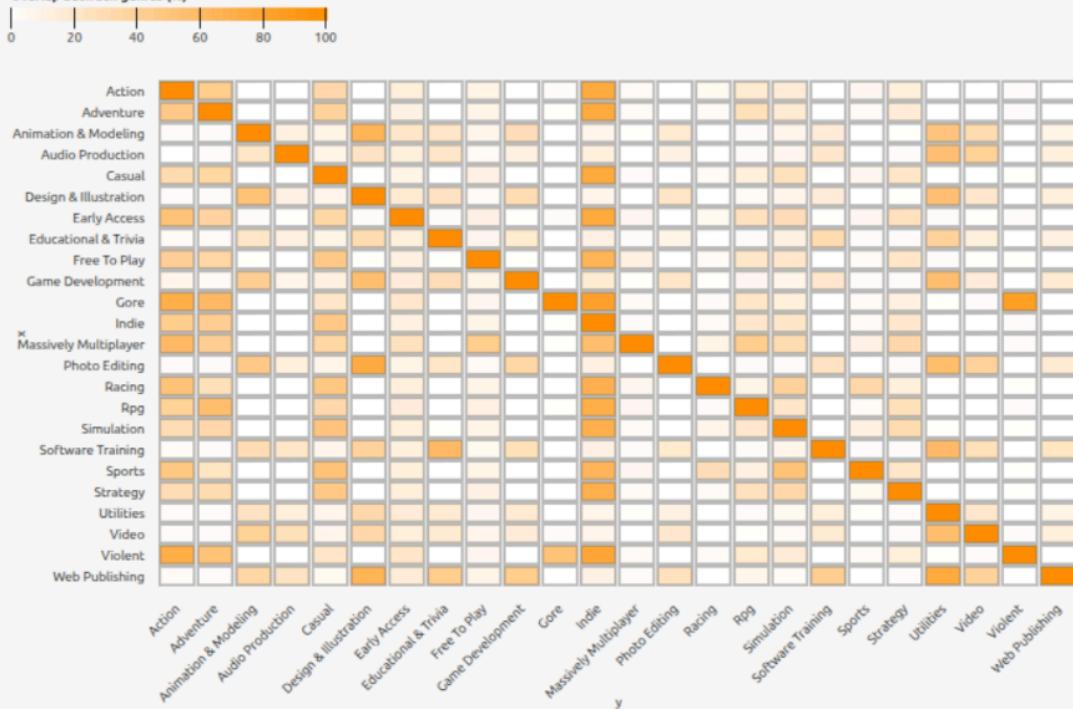


Steam Playstation Xbox All

Minimum amount of games:

- | | | |
|---|--|---|
| <input checked="" type="checkbox"/> Action | <input checked="" type="checkbox"/> Adventure | <input checked="" type="checkbox"/> Animation & Modeling |
| <input checked="" type="checkbox"/> Audio Production | <input checked="" type="checkbox"/> Casual | <input checked="" type="checkbox"/> Design & Illustration |
| <input checked="" type="checkbox"/> Early Access | <input checked="" type="checkbox"/> Educational & Trivia | <input checked="" type="checkbox"/> Free To Play |
| <input checked="" type="checkbox"/> Game Development | <input checked="" type="checkbox"/> Gore | <input checked="" type="checkbox"/> Indie |
| <input checked="" type="checkbox"/> Massively Multiplayer | <input checked="" type="checkbox"/> Photo Editing | <input checked="" type="checkbox"/> Racing |
| <input checked="" type="checkbox"/> Rpg | <input checked="" type="checkbox"/> Simulation | <input checked="" type="checkbox"/> Software Training |
| <input checked="" type="checkbox"/> Sports | <input checked="" type="checkbox"/> Strategy | <input checked="" type="checkbox"/> Utilities |
| <input checked="" type="checkbox"/> Video | <input checked="" type="checkbox"/> Violent | <input checked="" type="checkbox"/> Web Publishing |

Overlap between genres (%)

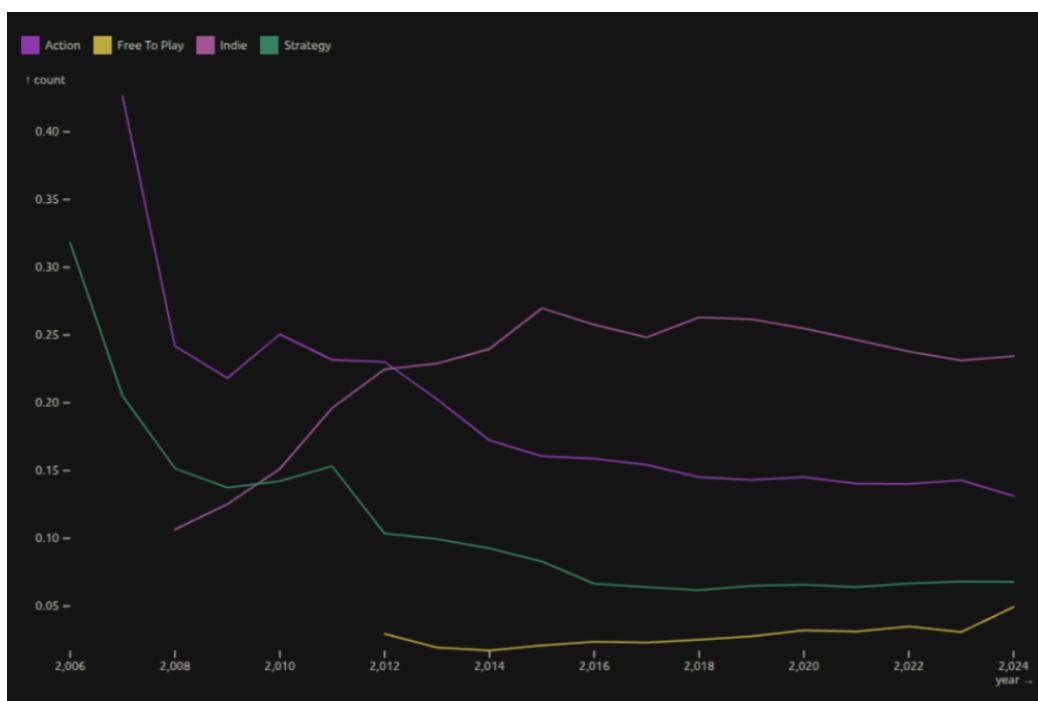
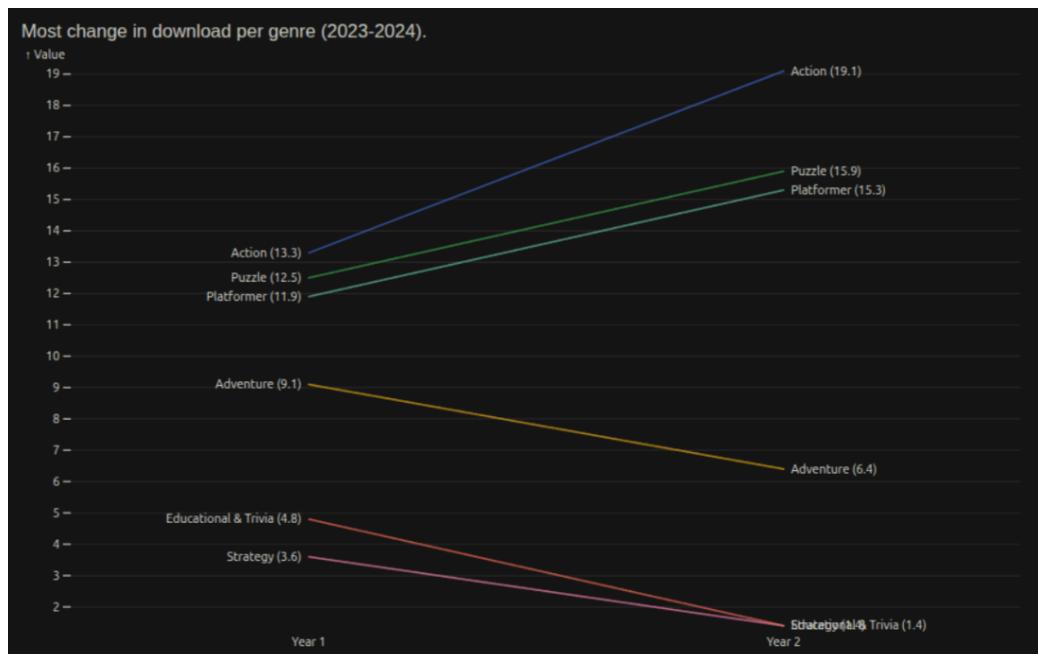


Included games:

99.92%

Which genres do gamers play? - Game release trend:

Het oorspronkelijke idee was om een slope graph te gebruiken, gecombineerd met twee selectoren om de te vergelijken jaren te kiezen. Al snel bleek echter dat deze aanpak te omslachtig was: het vereiste veel interactie van de gebruiker en leverde geen duidelijke inzichten op zonder actief te zoeken. Daarom hebben we ervoor gekozen om alle jaren tegelijk te tonen via een line graph, wat trends in één oogopslag veel inzichtelijker maakt.



Probleem van te veel genres

Een grote uitdaging bij deze grafiek was opnieuw het hoge aantal genres. Om dit te mitigeren, hebben we dezelfde technieken toegepast als bij de correlatiematrix.

Kleurcodering van lijnen

We wilden de verschillen tussen de lijnen zo duidelijk mogelijk maken. In eerste instantie probeerden we de `iwanthue`-package, die ontworpen is om n visueel onderscheidende kleuren te genereren. Hoewel dit enige verbetering bracht, voldeed het niet aan de verwachtingen. Uiteindelijk kozen we voor een eenvoudige regenboogkleurenreeks, die in de praktijk visueel het meest bruikbaar bleek.

Markeren van interessante trends

Omdat er relatief weinig opvallende trends zichtbaar waren, hebben we bij de initiële weergave de meest interessante ontwikkelingen aangeduid.

Relatieve vs. absolute waarden

Wanneer we met absolute aantallen werken, lijkt het alsof het aantal games in elk genre stijgt. Dit is deels een vertekening: elk jaar worden er simpelweg meer games uitgebracht. Daarom bieden we ook de mogelijkheid om de gegevens relatief te bekijken, genormaliseerd ten opzichte van het totaal aantal games dat jaar.

Probleem bij relatieve weergave: kleine aantallen

Bij relatieve data ontstaat echter een probleem bij genres met weinig games. Een stijging van 1 naar 2 games geeft al een toename van 100%, wat tot verkeerde interpretaties kan leiden. Daarom laten we het invoerveld voor het minimum aantal games niet lager dan 100 gaan, om de kans op misleidende inzichten te minimaliseren.

Uitsluiting van het jaar 2025

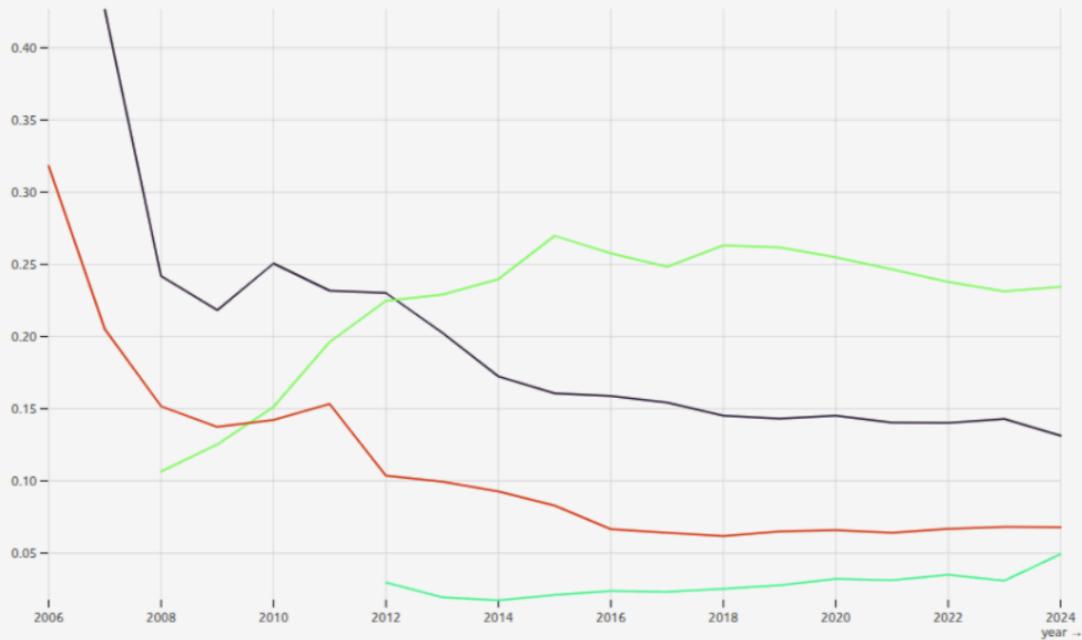
Het jaar 2025 is in de dataset nog niet compleet. Wanneer we dit jaar tonen, lijkt het alsof het aantal games plots sterk daalt. Om verwarring te voorkomen, wordt 2025 voorlopig niet weergegeven in de grafiek.

relative absolute

(The relative option will normalize the a genre's game count to the total amount of games released that year.)

Action Free To Play Indie Strategy

t count



Included games:

99.75%

Algemene aanpassingen:

In dit project hebben we verschillende algemene verbeteringen doorgevoerd:

1. Kleurenschema per platform

We kozen voor drie vaste kleuren per platform, gebaseerd op een triadisch schema op de kleurencirkel. Dit zorgt voor voldoende contrast en visuele harmonie.

2. Inhoudsopgave

Om gebruikers een snel overzicht te geven van de verschillende secties, hebben we een interactieve content table toegevoegd.

3. Verhaallijn via vragen

In plaats van louter beschrijvende titels, structureren we de presentatie aan de hand van kernvragen. Dit versterkt de narratieve flow en stimuleert de nieuwsgierigheid van de gebruiker.

4. Uitgebreide storytelling

Tijdens eerdere presentaties bleek dat er meer kleine verhaaltjes en contextuele anekdotes bij de grafieken verwacht werden, in plaats van uitsluitend technische toelichting.

5. Kleurwijziging thema

Voor betere zichtbaarheid werd besloten een wit thema in plaats van een zwart thema te gebruiken.

Taakverdeling:

Ruben	Floris
How do gamers spend? - Game yields	Opzetten Observable Framework
Which genres do gamers play? - Correlation between genres	What do gamers play on? - Cumulative downloads line graph
Which genres do gamers play? - Game release trend	What do gamers play on? - Monthly releases line graph
Automatische upload naar github pages	Where are gamers? - Gamer distribution chloropleth map
	Where are gamers? - Gamer distribution bar chart
	What do gamers think? - Word cloud
	What do gamers think? - Word frequency bar chart

(Wat niet vermeld staat, werd samen gedaan.)

Appendix:

Gaming Profiles 2025 (Steam, PlayStation, Xbox)

Links:

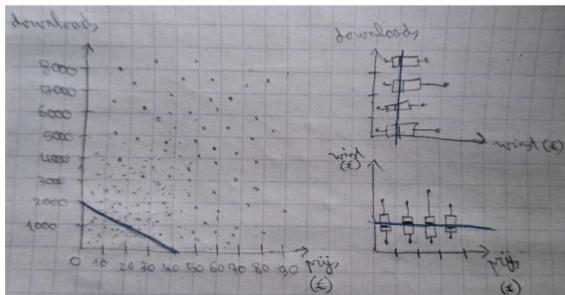
<https://www.kaggle.com/datasets/artyomkruglov/gaming-profiles-2025-steam-playstation-xbox>
<https://github.com/Smipe-a/gamestatshub>

Beschrijving:

Deze dataset geeft meer informatie over spellen gespeeld op steam, playstation of xbox. Ook krijgen we informatie wie deze spellen gespeeld heeft. We kunnen deze combinatie gebruiken om linken tussen de twee te gaan onderzoeken en zo verbanden te zoeken. Mogelijke vragen die we ons kunnen stellen zijn: "Spelen zowel vrouwen als mannen dezelfde spellen?", "Welke leeftijdsgroep brengt voor de producent het meeste geld op?", "Zijn er bepaalde talen die steeds populairder worden in spellen?", ... Het doelpubliek zou spelontwikkelaars zijn die zich willen vormen naar de huidige trends, maar ook in het algemeen geïnteresseerde spelliefhebbers. Moest er extra tijd over zijn, dan zou het interessant zijn om extra verwerking van de data te doen om verborgen verbanden te ontdekken (machine learning, text processing, sentiment analysis,

INLEIDING (Ruben)

- Link tussen prijs en downloads
- Prijs t.o.v. gemiddelde winst
- Downloads t.o.v. gemiddelde winst
- **Scatter plot** (per game een punt op xy-plot. Bereken de gemiddelde winst en plot als lijn)
+ 2 box-and-whisker plots



=> Hoeveel ga ik vragen voor mijn spel? Er kan inschatting gemaakt worden voor een afweging tussen verspreiding van het spel (downloads) en winst. Dit is nuttig voor specifieke doelen of een bepaalde focus van een team.

- Populairste games (downloads, winst)
- Populairste developers (totale downloads, totale winst)
- Populairste publishers (totale downloads, totale winst)
- **Bar Chart** (over de tijd heen)

LAND (Floris, basis done)

- Downloads per land
- **Choropleth Map**
- Link tussen landen op basis van vrienden. (geen hovering is grootste)
- **Map with arrows on hovering** (size indicates amount)
- Belangrijkste genre(s) per land (geen hovering is grootste)
- **Map with list of genres on hovering**

=> Welke landen hebben grote invloed? Als ik zie dat mijn spel het goed doet in een land? In welke andere landen zou ik aan marketing kunnen doen?

PLATFORM (Floris)

- Trend (downloads, prijs, winst) over tijd van platform van de games. Kan zowel globaal als per land.
- **Line Graph** (can all be on one graph as the actual values don't have to be displayed. You are making a comparison, not the actual values are important. You are making the game anyways. If necessary, make a selector, or make tooltips)

=> Voor welk platform ga ik mijn spel maken?

GENRE (Ruben)???

- Trend (downloads, prijs, winst) over tijd van genre van de games. Kan zowel globaal als per land.
- **Slope Graph**
- Correlatie tussen genres (op basis van downloads, prijs, winst). Welke genres staan vaak samen.
- **Heatmap**
- Correlatie tussen developers/publisher. Een developer/publisher die een bepaald genre heeft gemaakt, maakt ook vaak een ander genre?
- **Correlatiematrix**

=> In welke genres gaan we de game maken?

TAAL

- (downloads, prijs, winst) voor een bepaalde taal, van hoog naar laag.
- **Bar Chart**
- Per land, welke ondersteunde talen het populairst zijn
- **Pie Chart**, land selector by map

=> Welke talen gaan we ondersteunen? We zien dat ons spel populair wordt in een bepaald land. Welke extra talen gaan we toevoegen?

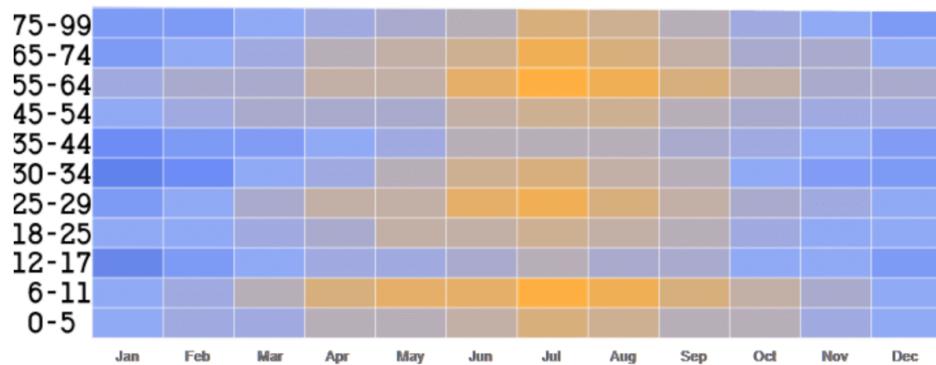
ACHIEVEMENT (nee)

- Achievement rarity distributie
- **Stacked Bar Chart**
- Aantal achievements (per genre?) (downloads, prijs, winst)
- **Box plot**
- Correlatie tussen achievement rarity en populariteit van de game?

=> Welke moeilijkheidsgraad moet ik mijn achievements geven? Heeft het nut om achievements toe te voegen? Hoeveel/hoe moeilijk zijn de achievements binnen een bepaald genre?

TIJD (nee)

- User engagement over tijd, met focus op specifieke demographics
- Heatmap, tijd op x axis, specifieke demographic op y axis, en een dropdown om specifieke demographic criteria te selecteren (leeftijd, regio, multiplayer gamer, singleplayer gamer)
- Kan evt. aparte plots voor elk criterium
- Voorbeeld met leeftijd:



=> Gegeven een specifieke doelgroep, welke periode is het beste om mijn spel in te releasen?

REVIEWS (Floris, done)

- Word cloud/bar chart van meest gebruikte woorden in reviews
- Preprocessing zodat woorden als "de", "een", niet het meest frequent, maar eerder woorden als "moeilijk", "slechte graphics" het meest frequent zijn

Deployen via github pages