|  |
| --- |
| DataScience |
| Specialisatie Opdracht |
| Kunst maken met games |
| Roy Hendriks, Rick Schoeman en Nick Schouten [Klas: SIE-V2A] Team-Bob Ross |

|  |  |  |
| --- | --- | --- |
| **DataScience** | | |
| **Document datum** | 12-04-2017 | |
| **Document versie** | V1.0 | |
| **Versiebeheer** | V1.0 | Geen Verandering |
|  |  |  |

Inhoudsopgave

[Requirements 2](#_Toc479780025)

[Data munging 2](#_Toc479780026)

[Statistische analyse 2](#_Toc479780027)

[Machine learning 2](#_Toc479780028)

[Visualisatie 2](#_Toc479780029)

[Beschrijving extra databronnen 3](#_Toc479780030)

# Requirements

* gamesales.csv
* metacritic.csv
* 2015.csv

# Data munging

We lezen 3 datasets in; een voor gamesales, een voor gamereviews en een voor de worldhappiness. De datasets voor gamesales en voor reviews voegen we samen tot een dataset. In die nieuw gemaakte dataset verwijderen we de kolom rank omdat die overbodig is door de index. Ook verwijderen we alle rijen waar een NaN veld instaat omdat we die data niet kunnen gebruiken.

# Statistische analyse

We maken aan aantal diagrammen aan. Een diagram laat zien wat de totale sales per platform zijn. Een diagram laat zien wat de sales per genre zijn. En de andere diagrammen zijn de critic en user reviews per genre en platform. Ook maken we aken we twee heatmaps aan die laten zien in welke regio de hoogste worldhappiness is en in welke regio de meeste games worden verkocht. Zo kun je zien dat in de gebieden waar meer games worden verkocht meer happiness is.

# Machine learning

Voor supervised learning maken we een formule aan voor de prediction van hoeveel sales er gaan komen in een jaar. Deze formule gebruiken we voor het plotten van twee graffieken. Een met de huidige jaartallen en een met jaartallen in de toekomst. Ook maken we een vergelijkbare graffiek maar dan zonder de gemaakte formule, maar dan aan de hand van de functie lmplot van seaborn.  
Voor unsupervised maken we een veld aan met vlakken op basis van de som van de gamesales op basis van het jaar. Doormiddel van deze vlakken kun je een hoeveelheid sales invullen en dan laat het systeem weten in welk jaar die hoeveelheid sales het meest valt.

# Visualisatie

We maken twee datasets aan met de gemiddelde reviewscore van critics en users aan op basis van het jaartal. Met deze datasets maken we een grafiek aan met de gemiddelde userscore en de gemiddelde criticscore per jaar(criticscore in het rood en userscore in het blauw). Ook hebben we twee sliders om aan te geven wat het minimale jaartal en het maximale jaartal moet zijn.

# Beschrijving extra databronnen

Voor deze opdracht gaan we pandas, bokeh, seaborn, sklearn en matplotlib gebruiken. Pandas gebruiken we voor de datasets zodat we goed de data kunnen beheren. Bokeh gebruiken we voor de visualisatie. Seaborn gebruiken we om heatmaps te maken. Sklearn gebruiken we voor het maken van de prediction. Matplotlib gebruiken we voor de diagrammen voor de prediction