Modèles statistiques

Projet proglit

Quentin Blondel, Andréas Dedieu-Meille

24 mars 2017

Contents

Introduction

Context / Dataset description

Ce projet a été réalisé dans le cadre du cours de Modèles statistiques de la L3 MIAGE de l'Université Grenoble Alpes. Pour ce projet nous devions choisir des data sur lesquelles travailler parmi les sources suivantes :

- FR Data.gouv: https://www.data.gouv.fr/fr/
- Insee: https://www.insee.fr/fr/accueil
- Kaggle: https://www.kaggle.com/datasets
- US Data.gov: https://catalog.data.gov/dataset

Nous avons choisi de travailler sur les dataset de la célébre plateforme de téléchargement de jeux vidéo : Steam. Les data ont été récupéré sur Kaggle Kaggle dans un fichier csv comportant 200 000 lignes. L'auteur de la page Kaggle indique que ces données proviennent des données publiques de Steam, mais aucune source n'est citée.

Le fichier comporte 5 colonnes :

- l'ID du joueur
- le nom du jeu
- le statut qui est soit "purchase", soit "play"
- le temps de jeu en heures pour "play" ou bien 1.0 pour "purchase"
- la dernière colonne n'a pas d'utilité et ne comporte que des zéro (elle sera filtrée par la suite)

Ainsi pour chaque jeu pour un joueur donné, on a une ligne indiquant qu'il a acheté le jeu et une seconde ligne indiquant le temps de jeu (s'il a joué à ce jeu).

Description of the question

Notre objectif par l'étude de ces données est de mettre en évidence s'il y a ou non une corrélation entre le nombre d'achats et le temps total joué ? On peut par exemple supposé que plus un jeu est acheté, plus il est joué. Est-ce le cas ?

Methodology

Data clean-up procedures

Chargement des librairies :

```
library(ggplot2);
library(magrittr);
library(dplyr);
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr);
```

Comme le fichier brut ne comportait pas d'entête de colonne, la première ligne était considérée comme l'entête. Cela posait problème car lorsqu'on redéfinissait les noms des colonnes on perdait la première ligne, ce qui fausse les données. Afin de remédier à ce problème nous avons modifier directement le fichier brut afin de rajouter des entêtes de colonnes, ce qui nous permet de conserver la première ligne. On commence tout d'abord par charger le fichier dans une varible (ici "df"), on en profite également pour fixer un nom à nom colonne et éliminer la colonne 5 qui ne comporte aucune données.

```
#Lecture du fichier et nommage des colonne
df <- read.csv("steam-200k.csv", header=TRUE, col.names=c("ID", "Game", "Statut",
                                                          "Temps joué", "ras"))
#Suppression de la dernière colonne qui n'a pas d'utilité
df$ras <- NULL
#Affichage des 6 premières lignes du fichier
head(df);
```

```
##          ID                         Game   Statut Temps.joué
## 1 151603712 The Elder Scrolls V Skyrim purchase        1.0
## 2 151603712 The Elder Scrolls V Skyrim     play      273.0
## 3 151603712                   Fallout 4 purchase        1.0
## 4 151603712                   Fallout 4     play       87.0
## 5 151603712                       Spore purchase        1.0
## 6 151603712                       Spore     play       14.9
```

Scientific workflow

Tout d'abord, nous avons voulu traiter tous les cas : nous avons créé une relation "jouéTous" qui pour chaque jeu relie le cumul du temps passé dessus par tous les joueurs.

```
df %>% filter(Statut=="play") %>%
  group_by(Game) %>%
  summarize(TempsJoué=sum(Temps.joué)) -> joueTous;
head(joueTous);
```

```
## # A tibble: 6 × 2
##                                                      Game TempsJoué
##                                                    <fctr>     <dbl>
## 1                                             007 Legends       0.7
## 2                                               ORBITALIS       1.2
## 3 1... 2... 3... KICK IT! (Drop That Beat Like an Ugly Baby)      20.0
## 4                                           10 Second Ninja       5.9
## 5                                                10,000,000       3.6
## 6                                          100% Orange Juice      78.3
```

Puis, nous avons récupéré le nombre total d'achats pour chaque jeu.

```
df %>%
  filter(Statut == "purchase") %>%
  group_by(Game) %>%
  summarize(NbAchat = n()) -> purchaseTous;
head(purchaseTous);
```

```
## # A tibble: 6 × 2
##                                                      Game NbAchat
##                                                    <fctr>   <int>
## 1                                             007 Legends       1
## 2                                               ORBITALIS       3
## 3 1... 2... 3... KICK IT! (Drop That Beat Like an Ugly Baby)       7
## 4                                           10 Second Ninja       6
## 5                                                10,000,000       1
## 6                                          100% Orange Juice      10
```

Enfin, on relie les 2 tables avec un merge.

```
merge(joueTous, purchaseTous, by="Game") -> result1;
head(result1);
```

```
##                                                      Game TempsJoué
## 1                                             007 Legends       0.7
## 2                                               ORBITALIS       1.2
## 3 1... 2... 3... KICK IT! (Drop That Beat Like an Ugly Baby)      20.0
## 4                                           10 Second Ninja       5.9
## 5                                                10,000,000       3.6
## 6                                          100% Orange Juice      78.3
##   NbAchat
## 1       1
## 2       3
## 3       7
## 4       6
## 5       1
## 6      10
```

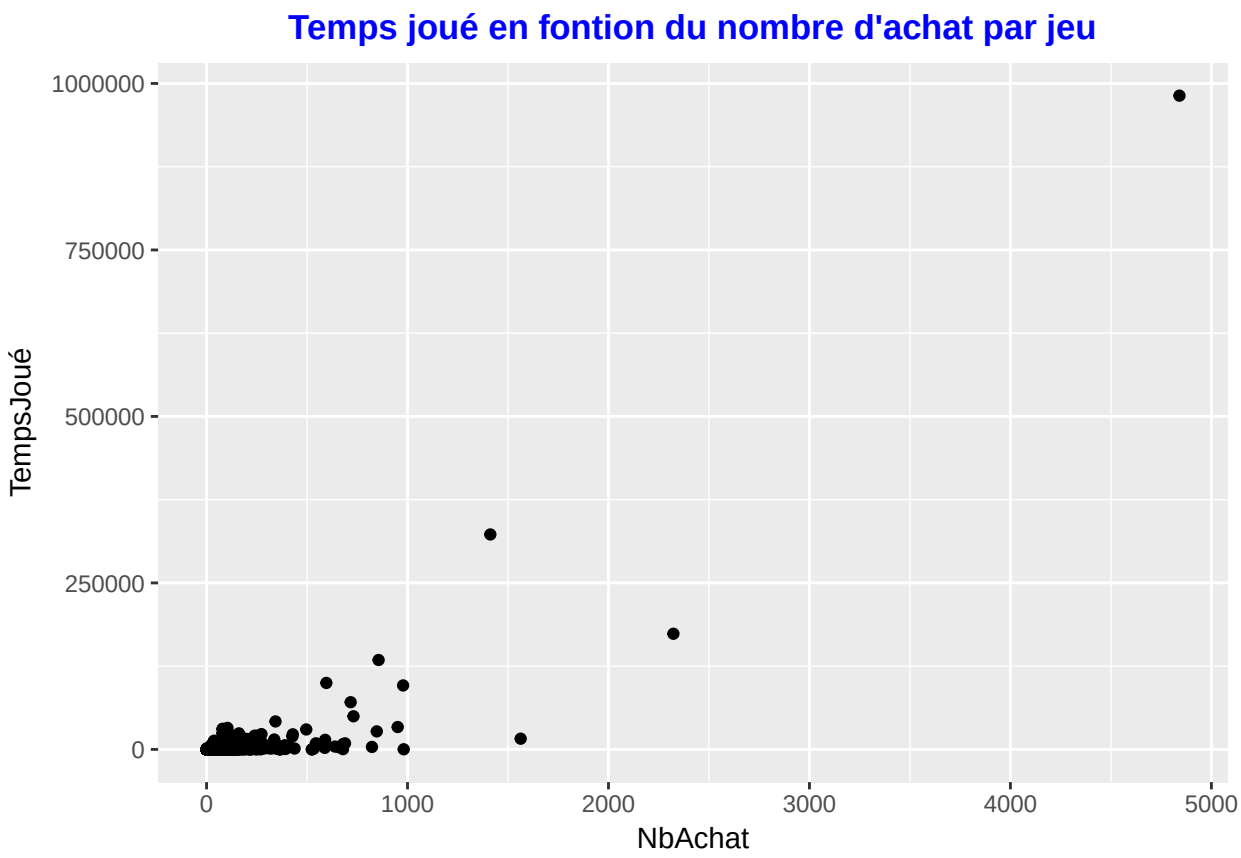
Data representation choices

Comme représentation graphique, nous avons choisi de mettre en abscisse le nombre d'achats et en ordonnée le temps total de jeu. Chaque jeu sera représenté par un point afin de visualiser pour chacun des jeux le rapport entre le nombre d'achats et le temps de jeu.

Analysis in Literate Programming

Production du graphique :

```
result1 %>%
  ggplot(aes(x = NbAchat, y = TempsJoué)) +
  ggtitle("Temps joué en fontion du nombre d'achat par jeu") +
  theme(plot.title = element_text(color = "blue", face="bold", hjust = 0.5)) +
  geom_point()
```



Comme on peut le voir, un grand nombre de jeux sont concentrés proches de l'origine, avec quelques points qui sortent de la masse. De plus, on ne peut pas différencier tous les jeux (l'idée était de colorer chaque point qui représenterait un jeu). Il fallait restreindre notre liste. C'est pourquoi nous nous sommes dits que l'on pourrait prendre les jeux qui ont été achetés plus de X fois, et les jeux qui ont été joués plus de Y heures.

La méthode effectuée est la suivante :

- Création d'une fonction "joués" qui créé une relation entre les jeux présents dans la base et la somme du temps de jeu de chaque joueur pour ce jeu. "joués" prend un paramètre qui est le palier en heures de jeu à partir duquel on prend en compte les jeux concernés. Exemple : joués(limite = 10000) -> renvoie un tableau qui contient les jeux joués plus de 10000 heures au total.
- Création d'une fonction "jouésSansDota" qui est la même que "joués" mais on enlève le jeu "Dota 2" de la liste, car nous verrons que ce jeu "casse" les donnés avec des valeurs extêmes.
- Création d'une fonction "purchase" qui renvoie la quantité totale d'exemplaires achetée pour un jeu. La fonction comporte elle aussi un paramètre, une limite où on ne prend en compte que la nombre d'achats supérieurs à cette limite.

Puis, on relie la fonction "joués" (ou "jouésSansDota") avec "purchase" dans une table. Avec cette table, nous faisons un plot : en abscisses de Nombre d'achat total des jeux, en ordonnée le temps passé total, les

points représentent chacun un jeu.

Dans notre exemple, les jeux ont doivent être joués pendant au moins 10000h au total, et doivent être achetés au moins 500 fois. Ces paramètres peuevent êtres changés dans les paramètres des fonctions.

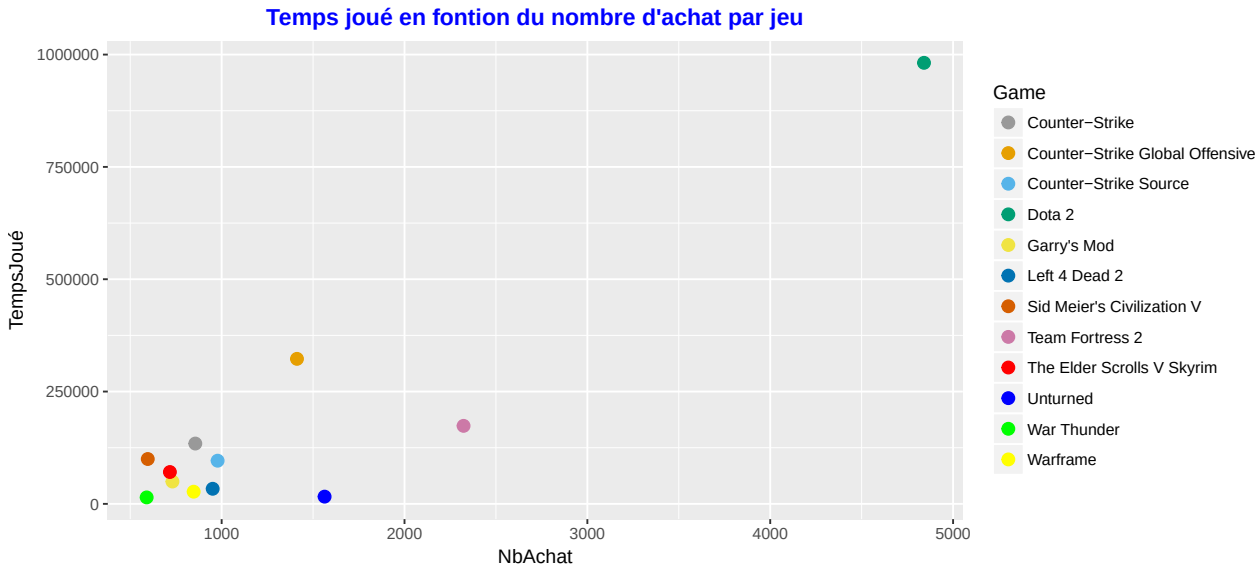
Comme cité plus haut, le jeu Dota 2 a des valeurs extêmes, c'est pourquoi nous avons choisi de faire un dernier graphique sans celui-ci.

```
# Fonction joués(limite) renvoie un tableau où les jeux indiqués ont été joués plus de
# temps que la limite entrée en paramètre
joués <- function(limite = 10000)
{
df %>%
    filter(Statut=="play") %>%
    group_by(Game) %>%
    summarize(TempsJoué=sum(Temps.joué)) %>%
    filter(TempsJoué > limite);
}
joués(limite = 10000)-> jeu;


# Fonction jouésSansDota(limite) renvoie un tableau où les jeux indiqués ont été joués
# plus de temps que la limite entrée en paramètre, mais le jeu "Dota 2" est exclu
# comme réultat
jouésSansDota <- function(limite = 10000)
{
 df %>%
    filter(Statut=="play") %>%
    filter(Game != "Dota 2") %>%
    group_by(Game) %>%
    summarize(TempsJoué=sum(Temps.joué)) %>%
    filter(TempsJoué > limite);
}
jouésSansDota(limite = 10000) -> jeuSD;


# Fonction purchase(limite) renvoie où les jeux indiqués ont été achetés plus que la
# limite entrée en paramètres
purchase <- function(limite = 500)
{
  df %>%
    filter(Statut == "purchase") %>%
    group_by(Game) %>%
    summarize(NbAchat = n()) %>%
    filter(NbAchat > limite);
}
purchase(limite = 500) -> achat;

# Liaison de la table de "joués" avec "purchase"
merge(jeu, achat, by="Game") -> test;
test;
```

```
##                             Game TempsJoué NbAchat
## 1                 Counter-Strike  134261.1     856
## 2  Counter-Strike Global Offensive  322771.6    1412
## 3            Counter-Strike Source   96075.5     978
```

```
## 4                          Dota 2  981684.6    4841
## 5                    Garry's Mod   49725.3     731
## 6                   Left 4 Dead 2   33596.7     951
## 7       Sid Meier's Civilization V   99821.3     596
## 8                  Team Fortress 2  173673.3    2323
## 9       The Elder Scrolls V Skyrim   70889.3     717
## 10                        Unturned   16096.4    1563
## 11                   War Thunder   14381.6     590
## 12                       Warframe   27074.6     847
```

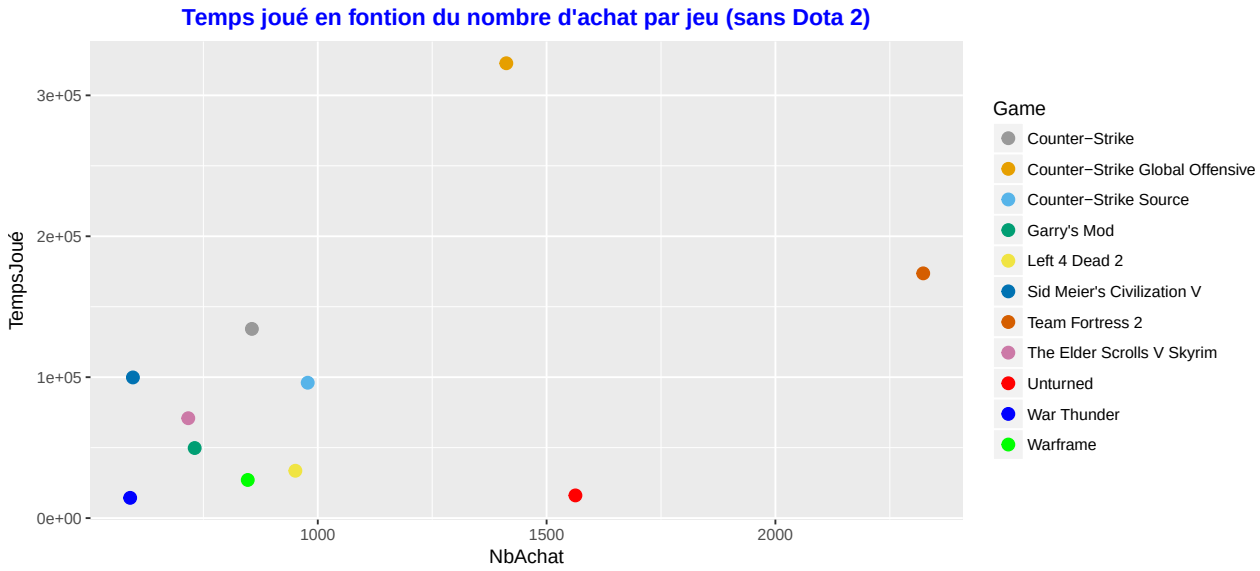
```
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442",
               "#0072B2", "#D55E00", "#CC79A7", "Red", "Blue", "Green", "Yellow");

test %>%
  ggplot(aes(x = NbAchat, y = TempsJoué, color = Game)) +
  ggtitle("Temps joué en fontion du nombre d'achat par jeu") +
  theme(plot.title = element_text(color = "blue", face="bold", hjust = 0.5))  +
  geom_point(size = 3) +
  scale_colour_manual(values=cbPalette);
```




```
# Liaison de la table "jouésSansDota" avec "purchase"
# En soit, c'est un zoom du plot précédent
merge(jeuSD, achat, by="Game") -> test2;
test2;
```

```
##                              Game TempsJoué NbAchat
## 1                  Counter-Strike  134261.1     856
## 2  Counter-Strike Global Offensive  322771.6    1412
## 3            Counter-Strike Source   96075.5     978
## 4                    Garry's Mod   49725.3     731
## 5                   Left 4 Dead 2   33596.7     951
## 6       Sid Meier's Civilization V   99821.3     596
## 7                  Team Fortress 2  173673.3    2323
## 8       The Elder Scrolls V Skyrim   70889.3     717
## 9                         Unturned   16096.4    1563
## 10                   War Thunder   14381.6     590
## 11                       Warframe   27074.6     847
```

```
test2 %>%
  ggplot(aes(x = NbAchat, y = TempsJoué, color = Game)) +
  ggtitle("Temps joué en fontion du nombre d'achat par jeu (sans Dota 2)") +
  theme(plot.title = element_text(color = "blue", face="bold", hjust = 0.5))  +
  geom_point(size = 3)  +
  scale_colour_manual(values=cbPalette);
```



On remarque que le jeu le plus joué et le plus acheté est Dota 2, sans le considérer c'est Team Fortress 2 qui est le plus acheté et Counter-Strike : Global Offensive le plus joué.

La corrélation est difficile entre le nombre d'achats d'un jeu et la quantité de temps passé dessus. La logique voudrait que plus un jeu est acheté, plus les joueurs passent du temps dessus, donc une relation proportionnelle entre les deux. Dans les faits, ce n'est pas le cas.

On note quand même les extrêmes : le jeu "Unturned" est celui le moins joué malgré un nombre d'achats important : on peut supposer qu'il a bénéficié d'une publicité positive malgré des désavantages comme peut être la qualité médiocre du jeu. Après vérification, ce jeu est actuellement gratuit au téléchargement mais intègre des achats en jeu, ce qui expliquerait un nombre important d'"achats" (de téléchargements). Les joueurs n'ont toutefois pas accroché, l'entreprise pourrait utiliser cette donnée pour déterminer ce qu'il déplaît et adapter son produit.

A l'inverse, "Counter-Strinke : Global Offensive" compte beaucoup de temps de jeu et relativement peu d'achats. Il arrive à fidéliser ses clients, puisqu'ils passent plus de temps sur ce jeu plutôt que de partir sur un autre.

"Dota 2" est vraiment un cas à part, dans le sens où il est vraiment excentré des autres valeurs. C'est apparemment le jeu le plus joué de la plateforme, et aussi le plus acheté, même si pour ce dernier les nombres sont faciles à monter vu que le jeu est gratuit à l'achat...

Il n'y a donc pas de relation directe que l'on peut voir de visu entre la quantité d'achats pour un jeu et la quantité de temps joué dessus.

Nous avons trouvé les statistiques de jeu du jour actuel, les premiers (au moment où l'on rédige) correspondent bien à Dota 2 et Counter-Strike, ce qui est conforme à notre base de données. En revanche, d'autres jeux sont apparemment actuellement très joués, qui n'apparaissent pas dans la base : on peut supposer que la base n'est pas récente.

Conclusion

Comme nous l'avons vu, on ne peut pas trouver de corrélation entre le nombre d'achats d'un jeu et son total d'heures de jeu. Cela semble à priori propre à chaque jeu.

On peut également regretter le manque de données de notre base. Il aurait en effet été intéressant d'avoir des données en plus comme par exemple l'année de lancement d'un jeu. Cela nous aurait permis de pouvoir comparer le temps de jeu par rapport à l'année de lancement du jeu.

Il aurait également été intéréssant d'avoir un échantillon plus important, car Steam comporte plusieurs millions de joueur, alors qu'ici nous avons qu'un échantillon de quelques milliers d'utilisateurs.

References

- Data sur Kaggle : https://www.kaggle.com/tamber/steam-video-games
- Cours de Modèles statistiques : https://github.com/schnorr/proglit (par Jean-Marc Vincent et Lucas Mello Schnorr)
- Github de Quentin Blondel : https://github.com/blondelq
- Github d'Andréas Dedieu-Meille : https://github.com/dedieuma