

Game Recommendations on Steam

Using Play/Purchase data to generate a
predictive game recommendation model

What data have you gathered, and how did you gather it?

I used an open sourced project on Kaggle to find data. With a Kaggle account, it is easy to access data- data science projects and datasets

<https://www.kaggle.com/tamber/steam-video-games>

Steam Video Games

Recommend video games from 200k steam user interactions.

Now Available

by **Tamber** · last updated a month ago

-15%
\$9.99
\$8.49

tamber

Overview

Kernels

Discussion

Activity

Download (2 MB)

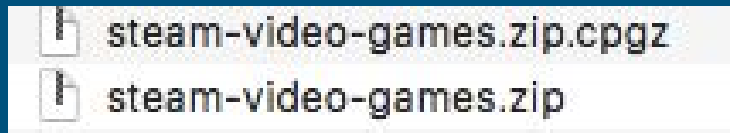
New Notebook

New Script

14

Downloading the Data was a little weird...

- Downloaded as a zip file
- When attempted to unzip, file would become .zip.cpgz, then zip again



- Trick to this is using command line “unzip + file extension”

Which areas of the data have been cleaned, and which areas still need cleaning?

The data is a “list of user behaviors”

Columns: user-id, game-title, behavior-name, value. The behaviors included in this data are 'purchase' and 'play' where play is set to the time played, and purchase is always set to 1 to mean yes, it has been purchased (or downloaded in the case for free games)

As is, the columns aren't particularly helpful and they need to be moved around a bit. I wanted play and purchase in the same row.

	A	B	C	D	E
1	151603712	The Elder Sc	purchase	1	0
2	151603712	The Elder Sc	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	0
7	151603712	Fallout New	purchase	1	0
8	151603712	Fallout New	play	12.1	0
9	151603712	Left 4 Dead 2	purchase	1	0
10	151603712	Left 4 Dead 2	play	8.9	0
11	151603712	HuniePop	purchase	1	0
12	151603712	HuniePop	play	8.5	0
13	151603712	Path of Exile	purchase	1	0
14	151603712	Path of Exile	play	8.1	0
15	151603712	Poly Bridge	purchase	1	0
16	151603712	Poly Bridge	play	7.5	0
17	151603712	Left 4 Dead	purchase	1	0
18	151603712	Left 4 Dead	play	3.3	0
19	151603712	Team Fortre	purchase	1	0
20	151603712	Team Fortre	play	2.8	0

Opening the df in ipython/jupyter

```
In [102]: steam.head(5)
```

```
Out[102]:
```

	player	game	behavior	hours	numeric
0	151603712	The Elder Scrolls V Skyrim	purchase	1.0	0
1	151603712	The Elder Scrolls V Skyrim	play	273.0	0
2	151603712	Fallout 4	purchase	1.0	0
3	151603712	Fallout 4	play	87.0	0
4	151603712	Spore	purchase	1.0	0

What steps have been taken to explore the data?

What insights have been gained from the exploration?

- Shape- 200000, 5
- Numeric column always = 0 (not useful) I need to remove these columns
- Having a play column and purchase column doubles the rows needed. I decided I wanted them both in a single row
- Need to make a 'play' df and 'purchase' df to create a new df
- I will probably use KNN and/or clustering in data model to find and predict trends in gaming. (will be based on play hours) I think I can use this data to answer the initial question of trends and predictions in gaming

Play/Purchase

- Separating out from one df to two- one with play data, and one with purchase data
- View of both df.head()

```
In [66]: df = pd.DataFrame(steam)
```

```
In [67]: df_play = df[df.behavior == 'play']  
df_purchase = df[df.behavior == 'purchase']
```

```
In [68]: df_purchase.head()
```

Out[68]:

	player	game	behavior	hours	numeric
0	151603712	The Elder Scrolls V Skyrim	purchase	1.0	0
2	151603712	Fallout 4	purchase	1.0	0
4	151603712	Spore	purchase	1.0	0
6	151603712	Fallout New Vegas	purchase	1.0	0
8	151603712	Left 4 Dead 2	purchase	1.0	0

```
In [69]: df_play.head()
```

Out[69]:

	player	game	behavior	hours	numeric
1	151603712	The Elder Scrolls V Skyrim	play	273.0	0
3	151603712	Fallout 4	play	87.0	0
5	151603712	Spore	play	14.9	0
7	151603712	Fallout New Vegas	play	12.1	0
9	151603712	Left 4 Dead 2	play	8.9	0

Making the new df

Next, creating one df_new with df_play and df_purchase

```
In [86]: df_new = df_play.merge(df_purchase, left_on=['player', 'game'], right_on=['player', 'game'] )  
df_new
```

Out[86]:

	player	game	behavior_x	hours_x	numeric_x	behavior_y	hours_y	numeric_y
0	151603712	The Elder Scrolls V Skyrim	play	273.0	0	purchase	1.0	0
1	151603712	Fallout 4	play	87.0	0	purchase	1.0	0
2	151603712	Spore	play	14.9	0	purchase	1.0	0
3	151603712	Fallout New Vegas	play	12.1	0	purchase	1.0	0
4	151603712	Left 4 Dead 2	play	8.9	0	purchase	1.0	0
5	151603712	HuniePop	play	8.5	0	purchase	1.0	0
6	151603712	Path of Exile	play	8.1	0	purchase	1.0	0

Dropping non-useful 'numeric' values

```
In [100]: del df_new['numeric_x']  
del df_new['numeric_y']  
df_new.head()
```

Out[100]:

	player	game	behavior_x	hours_x	behavior_y	hours_y
0	151603712	The Elder Scrolls V Skyrim	play	273.0	purchase	1.0
1	151603712	Fallout 4	play	87.0	purchase	1.0
2	151603712	Spore	play	14.9	purchase	1.0
3	151603712	Fallout New Vegas	play	12.1	purchase	1.0
4	151603712	Left 4 Dead 2	play	8.9	purchase	1.0

```
In [101]: df_new.shape
```

Out[101]: (70785, 6)