We are started gather all Kaggle repos we could find that had game related data. We ended up with 9 csv’s having various amounts of data in varying degrees of completeness. We eliminated one after the other until we settled on one with three files found in the Kaggle repo:

<https://www.kaggle.com/sidtwr/videogames-sales-dataset>

Three files containing video game data.

* PS4\_GamesSales.csv
* Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv
* XboxOne\_GameSales.csv

The columns/data broke down as follows

RangeIndex: 16719 entries, 0 to 16718

Data columns (total 16 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Name 16717 non-null object

1 Platform 16719 non-null object

2 Year\_of\_Release 16450 non-null float64

3 Genre 16717 non-null object

4 Publisher 16665 non-null object

5 NA\_Sales 16719 non-null float64

6 EU\_Sales 16719 non-null float64

7 JP\_Sales 16719 non-null float64

8 Other\_Sales 16719 non-null float64

9 Global\_Sales 16719 non-null float64

10 Critic\_Score 8137 non-null float64

11 Critic\_Count 8137 non-null float64

12 User\_Score 7590 non-null float64

13 User\_Count 7590 non-null float64

14 Developer 10096 non-null object

15 Rating 9950 non-null object

dtypes: float64(10), object(6)

We are going to drop columns 10-15 from video game sales since the other files do not contain matching columns.

We are going to create self-referencing columns for platform in both the Xbox and PS4.

This will leave us with 10 columns in each dataframe.

We then merged the 3 into one.

We drop duplicates.

We dropped NaN.

That left us with XXXXX rows of data.

We then started breaking the data down into the various normal forms