Data Science - Homework 5 (EDA)

- Introduction of datasets

1. Video Game Sales with Ratings (from Kaggle)

This is a video game sales dataset including game sales of North America, European, Japan, and other area, together they make the global sale. The data also gives information about the critic score, user score, and the counts of critics or users who gave these two kind of scores.

This dataset tells us quite a bit of information.

There are:

① 16719 rows

② 16 columns

③ Not many NA values. Only about 5 columns that have a significant amount of NA values

④ 9 of the variables are float types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16719 entries, 0 to 16718
Data columns (total 16 columns):
Name               16717 non-null object
Platform           16719 non-null object
Year_of_Release    16450 non-null float64
Genre              16717 non-null object
Publisher          16665 non-null object
NA_Sales           16719 non-null float64
EU_Sales           16719 non-null float64
JP_Sales           16719 non-null float64
Other_Sales        16719 non-null float64
Global_Sales       16719 non-null float64
Critic_Score       8137 non-null float64
Critic_Count       8137 non-null float64
User_Score         10015 non-null object
User_Count         7590 non-null float64
Developer          10096 non-null object
Rating             9950 non-null object
dtypes: float64(9), object(7)
memory usage: 2.0+ MB
```

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Sc
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	N
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	N

Taking the top 5 rows of the data frame, we see that Nintendo takes all 5 positions. This would make sense because they have been successful since they have entered the Video Game market. In addition, the games listed are known by the masses.

2. US Police Shootings (from Kaggle)

In the recent killings of US, a popular topic came into being, "Racism". So, I chose this dataset to take out some insights and analyze the story around racism in America. However, this dataset looks cleaned with no null values in each of the columns, I decided to manually delete some record to implement this assignment.

This dataset tells us quite a bit of information.

There are:

① 4895 rows

② 15 columns

③ Not any NA values.

④ Many different data types of the variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4895 entries, 0 to 4894
Data columns (total 15 columns):
id                         4895 non-null int64
name                       4895 non-null object
date                       4895 non-null object
manner_of_death            4895 non-null object
armed                      4895 non-null object
age                        4895 non-null float64
gender                     4895 non-null object
race                       4895 non-null object
city                       4895 non-null object
state                      4895 non-null object
signs_of_mental_illness    4895 non-null bool
threat_level               4895 non-null object
flee                       4895 non-null object
body_camera                4895 non-null bool
arms_category              4895 non-null object
dtypes: bool(2), float64(1), int64(1), object(11)
memory usage: 506.8+ KB
```

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_camera	arms_ca
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	M	Asian	Shelton	WA	True	attack	Not fleeing	False	
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	M	White	Aloha	OR	False	attack	Not fleeing	False	
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	M	Hispanic	Wichita	KS	False	other	Not fleeing	False	U
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	M	White	San Francisco	CA	True	attack	Not fleeing	False	Other
4	9	Michael Rodriguez	2015-01-04	shot	nail gun	39.0	M	Hispanic	Evans	CO	False	attack	Not fleeing	False	Piercing

- Use some data sets with missing data to conduct Exploratory Data Analysis (EDA) and Explain what you find and why you choose these EDA methods

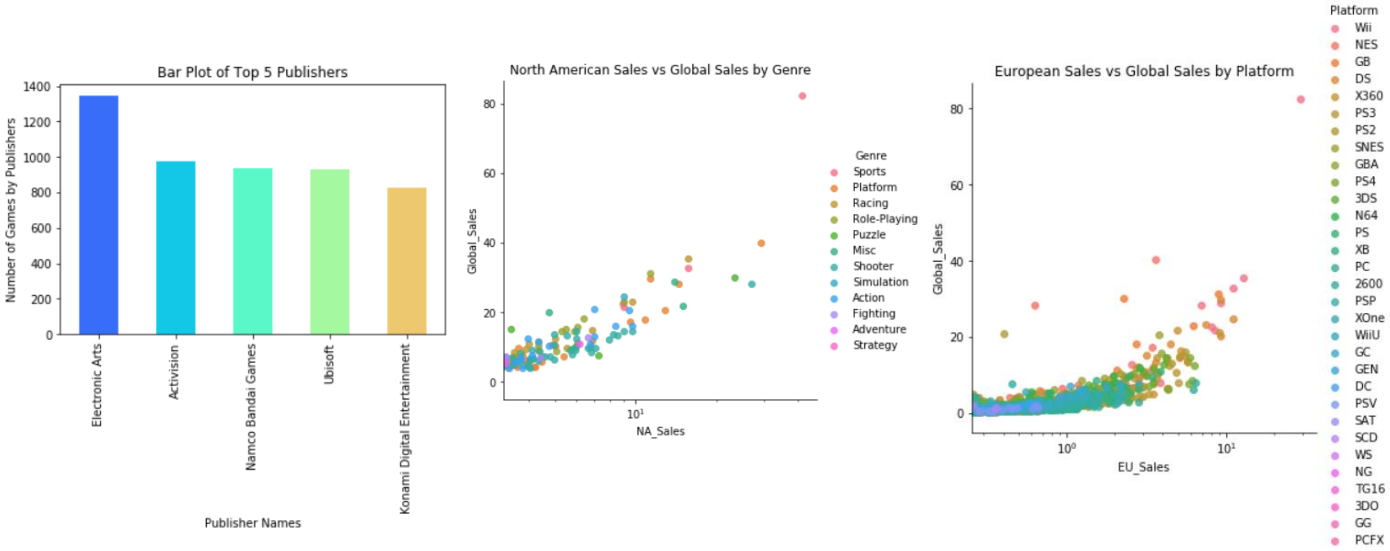
Dataset 1 - Video Game Sales with Ratings

It has been a ongoing question on what makes a best selling platform? In this assignment I will go through the processes of attaining, cleaning, preprocessing, and analyzing data to explore this question.

From the right statistics, we notice that User_Count has the most N/A values (about 9,100) showing that it will be difficult to work with since it will have very little significant values. It is a bit interesting how both Critic_Count and Critic_Score have the same amount of N/A values. However, it is more understandable for Rating to have many N/A values as many games are difficult to find a rating for.

```
User_Count         9129
Critic_Count       8582
Critic_Score       8582
Rating             6769
User_Score         6704
Developer          6623
Year_of_Release     269
Publisher            54
Genre                 2
Name                  2
Global_Sales          0
Other_Sales           0
JP_Sales              0
EU_Sales              0
NA_Sales              0
Platform              0
dtype: int64
```

When examining this data set, there wasn't any bad data to clean, such as misspelling. The only data that I will pursue to clean is 'Year_of_Release', 'Publisher', and 'Genre'. These columns are of interest since genre and publisher are a part of my goal. In addition, the amount of NA's present and will be dropped in these columns shouldn't be significant to effect my model later.

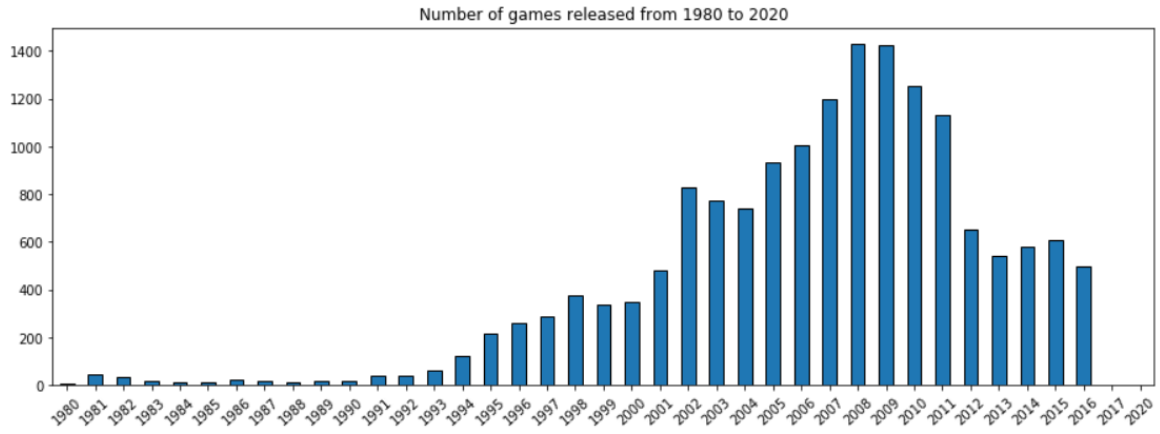


The top publisher for games is Electronic Arts with a significant lead over the middle three publishers. Next, I think that the lmplot is a helpful method to analyze the genre and platforms how to affects the markets and compare them, because the regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.

In the middle result of the genre column, by comparing North American sales and Global sales, one outlier is the top right pink circle representing sports. If you can recall from my *df.head()* the game "Wii Sports" is the top game of the data set. This would make sense because the game falls under 'Sports'.

In the right result of the platforms column, I went on to compare European sales and Global sales, again I observed that the same outlier at the top right, the pink circle representing Wii platform.

The other goal I was very curious about is the number of games released in the recent years. To simplify perception, I present a histogram grouped by name.



I found that the gaming industry has been actively developing since 1994. It should be noted that the ESRP rating has been used since 1993. The peaks are in 2006 - 2011, then we see a decline and since 2012 some leveling off. This may be due to the development of games on mobile devices for mobile phones, which are not in the list of platforms.

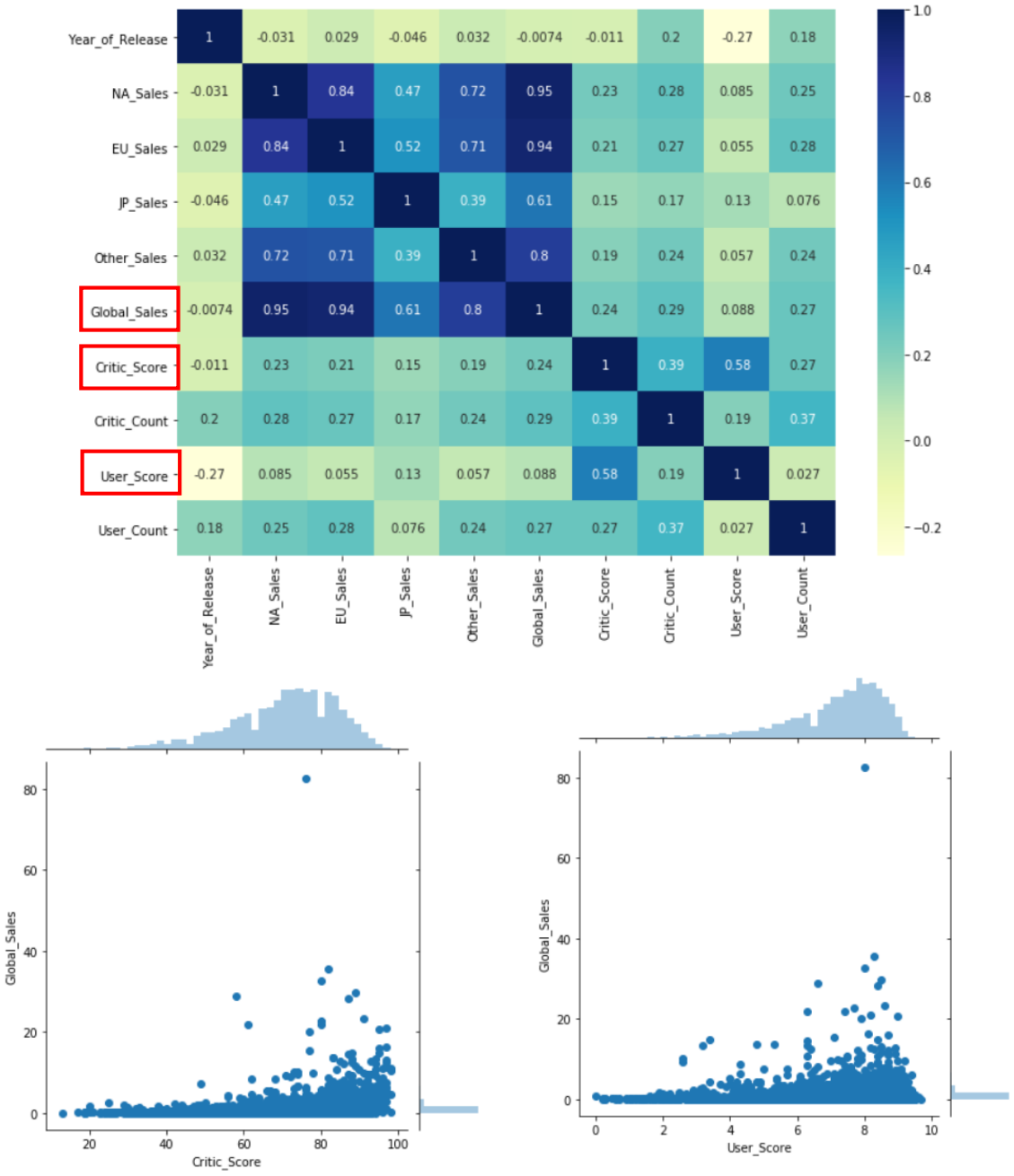
Let's take the period from 2012 to 2016 - from the moment of a sharp drop in sales and the trend of a decrease in the number of games produced. Therefore, I especially grabbed the sales of all released games from 2012 to 2016 to observe.

I see a decrease in the number of games created on the old generation consoles and an increase in production for new consoles: PS4 and XOne. Among the portable consoles, the 3DS can be distinguished. But a decline can be seen on the PSP.

Year_of_Release	2012	2013	2014	2015	2016	2017	2020	total
Platform								
total	653	544	581	606	502	3	1	2890
PS3	148	126	108	73	38	0	0	493
PSV	53	63	100	110	85	2	0	413
3DS	93	91	80	86	46	0	0	396
PS4	0	16	75	137	164	1	0	393
X360	106	75	63	35	13	0	0	292
PC	61	38	47	50	54	0	0	250
XOne	0	19	61	80	87	0	0	247
PSP	106	54	10	3	0	0	0	173
WiiU	32	42	31	28	14	0	0	147
Wii	31	12	6	4	1	0	0	54
DS	23	8	0	0	0	0	1	32

In recent years, new game platforms such as ps5 and switch have been launched one after another. Because the statistics of this data set are as of 2016, it may be a different trend now.

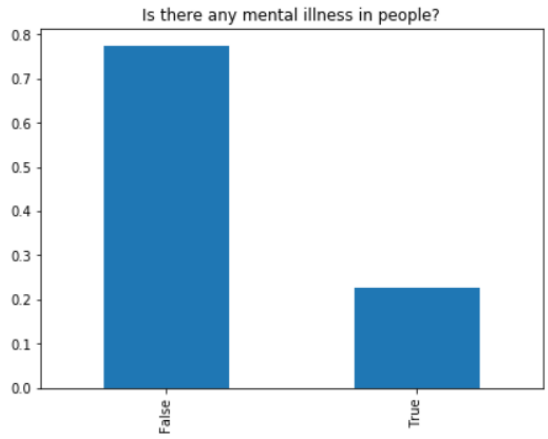
Another interesting thing about this dataset is critics' score and user scores. To inspect the correlations between all the numeric features and see how one feeds into the other, I shall plot a heatmap that seaborn provided. I wanted to focus on correlation between critic scores and sales and user scores and sales.

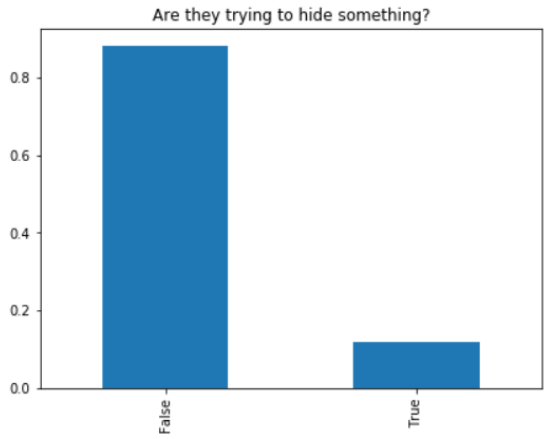


From the above, I observe a positive correlation between critics' score and sales and user scores and sales. It is worth noting that there is an average correlation between the ratings of users and critics, which may be due to some influence of the rating of the critic on the opinion of some ordinary voting people. Distributions are asymmetric.

Dataset 2 - US Police Shootings

First thing I must claim is why I selected this dataset, because as long as there is a violent or murder incident, the topic that is often brought up for discussion is whether the defendant suffers from a mental illness. This issue also plagues judges' decisions and needs to strike a balance between the victim and human rights. As a result, I want to know if the suspects or offenders usually has a mental illness? The other reason is the racial issues in America.



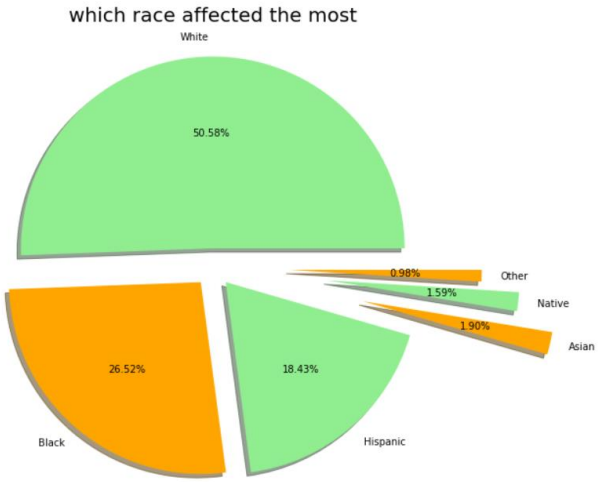


① About 78% people won't have any mental illness.

② About 90% time police does not have their body_camera ON to record the event which rises many question against them.

As we can see, almost more than half the number of shootings is of whites while about 26.5 % shootings is of black offenders.

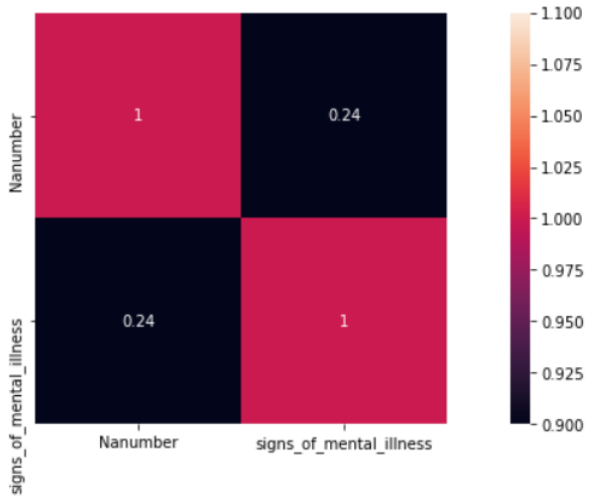
This is definitely contrary to popular belief which shows that the police had to shoot down more number of whites than blacks.



The last part is that I randomly delete the record myself and deal with it through some feature engineering techniques. I want to try a special method that assistant teach in the course, that is

“We can add a new feature “Nanumber” into our df, strong positive correlation may improve our model performance”

The results are not ideal, because the data is randomly deleted by myself, it looks normal, I just want to implement this feature extraction method, which may be useful in my future assignments.



After the above step, I used the frequent values to impute the missing values of the categorical features, the result is the closest compared to the original dataframe. In the numerical features, I used the MICE strategy to impute the missing values, it seemed to work fine.