How effective are video game categories for their sales?

Beyzanur Mollaoğlu
Computer Engineering 3rd Grade
University of Galatasaray
Istanbul, Turkey

Abstract- This document examines how categories affect video game sales, based on a dataset grouping various games.

Index Terms- data, values, analysis, statistics

I. Introduction

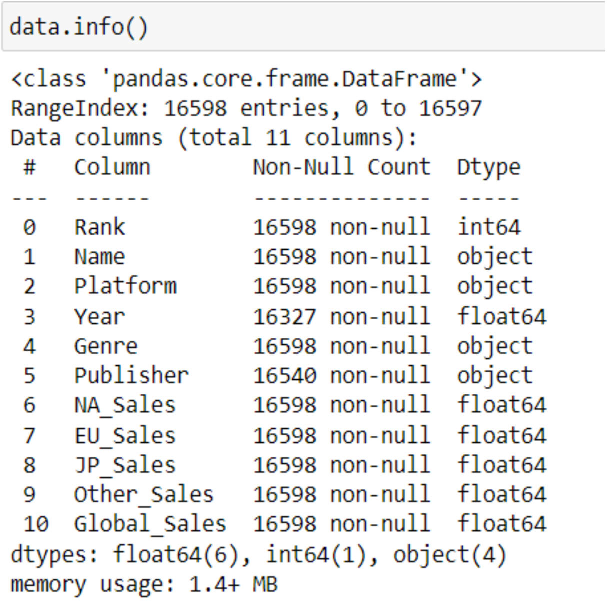
Video games entered our lives with the cathode ray tube entertainment device made in 1947. Later, this developed into console games and arcade machines. Since the 90s, computer games have taken over these. Over the years, game enthusiasts have increased and gaming has become an industry.

In this paper, we will look into a dataset is related a list of video games with sales greater than 100,000 copies. First, we will introduce the data. Next, we will ask a research question to conclude the topic and technical questions to help answer it. Finally, we will give a description of the methods we intend to use to answer the questions we have previously asked.

II. Dataset

First of all, what we need for analysis is a detailed introduction of data. It is a data set with a sample size of approximately 11600 and has eleven main columns (Fig. 1.).

- Rank – Ranking of overall sales
- Name – The games name
- Platform – Platform of the games release (i.e. PC, PS4, etc.)
- Year – Year of the game's release
- Genre – Genre of the game
- Publisher – Publisher of the game
- NA-Sales – Sales in North America (in millions)
- EU-Sales – Sales in Europe (in millions)
- JP-Sales – Sales in Japan (in millions)
- Other-Sales – Sales in the rest of the world (in millions)
- Global-Sales – Total worldwide sales.

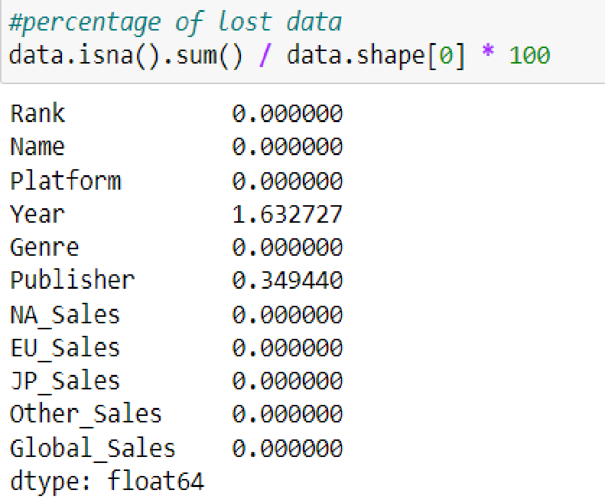


```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Rank          16598 non-null  int64  
 1   Name          16598 non-null  object 
 2   Platform      16598 non-null  object 
 3   Year          16327 non-null  float64
 4   Genre         16598 non-null  object 
 5   Publisher     16540 non-null  object 
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB
```

Fig.1 The output of the input "data.info()"

Before examining each value in the data, I will examine the null values in the data. When I looked at the percentages of null values (Fig 2.), I saw that the maximum was 0.16. Therefore, I decided to delete the null values from the data.

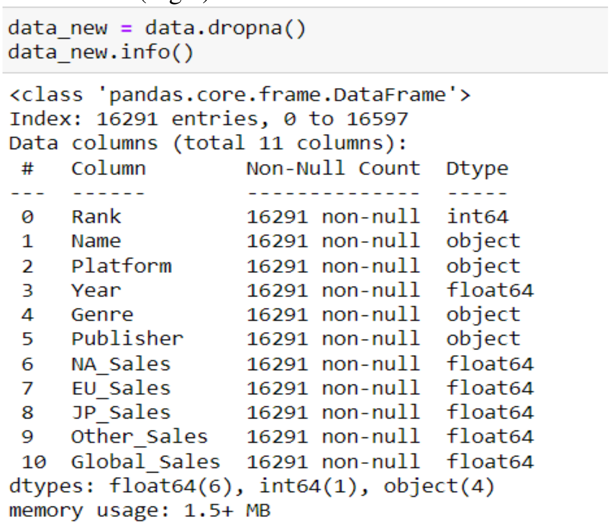


```
#percentage of lost data
data.isna().sum() / data.shape[0] * 100
```

```
Rank            0.000000
Name            0.000000
Platform        0.000000
Year            1.632727
Genre           0.000000
Publisher       0.349440
NA_Sales        0.000000
EU_Sales        0.000000
JP_Sales        0.000000
Other_Sales     0.000000
Global_Sales    0.000000
dtype: float64
```

Fig.2 The percentages of null values.

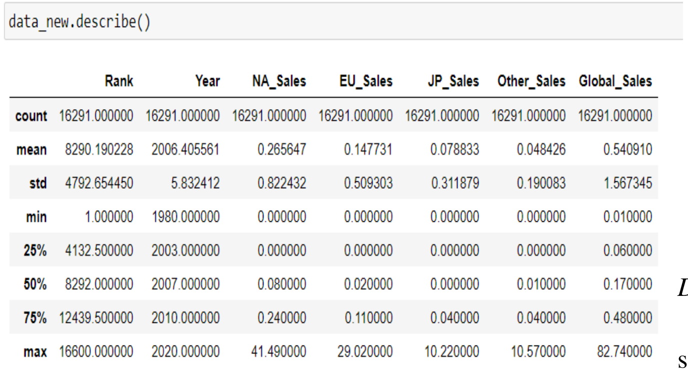
After dearing the null value in the dataset, we obtained a dataset with sample size of 16300 and consisting of eleven main columns (Fig.3).



```
data_new = data.dropna()
data_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 16291 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16291 non-null  int64
 1   Name          16291 non-null  object
 2   Platform      16291 non-null  object
 3   Year          16291 non-null  float64
 4   Genre         16291 non-null  object
 5   Publisher     16291 non-null  object
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.5+ MB
```

Fig.3 The info of the new data



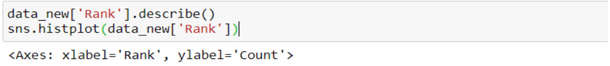
```
data_new.describe()
```

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Fig.4 The describe of the new data

A. *Rank*

Rank is a value that ensures that the data is in a certain order. It just lists from one to the end of the table.



```
data_new['Rank'].describe()
sns.histplot(data_new['Rank'])
```

```
<Axes: xlabel='Rank', ylabel='Count'>
```

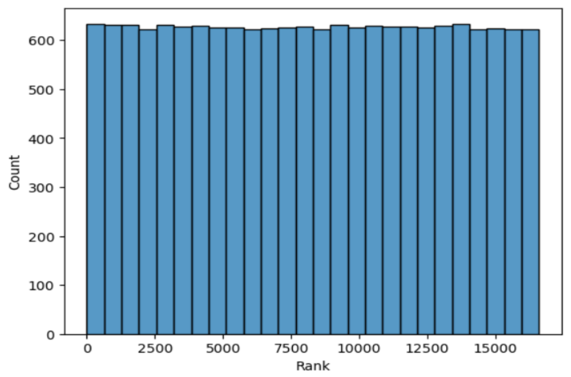
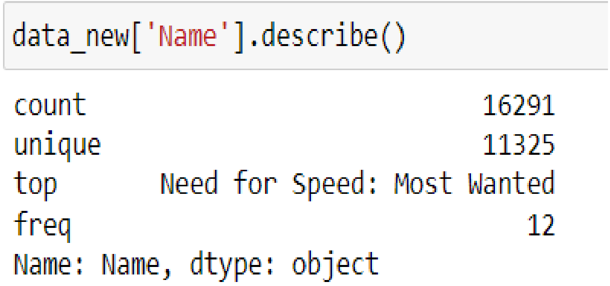


Fig.5 Rank histogram

B. *Name*

This column is the column where the names of the video games in the dataset are written.



```
data_new['Name'].describe()
```

```
count                             16291
unique                            11325
top       Need for Speed: Most Wanted
freq                                 12
Name: Name, dtype: object
```

Fig.6 Simple explanation of name values

C. *Platform*

Nowadays, games are produced specifically for many different platforms. This column shows on which platforms the video games in the dataset are sold. It seems that mostly games are listed on 13 percent DS and 13 percent PS2 platforms.

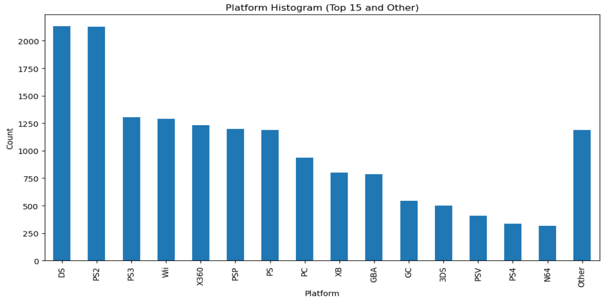
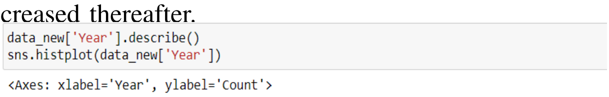


Fig.7 Platform histogram

D. *Year*

This column lists the years in which the games in the data set were published, from 1980 to 2020. As can be seen in Fig 8, most of the games were released in 2010. Although it increased continuously until 2010, this increase gradually decreased thereafter.



```
data_new['Year'].describe()
sns.histplot(data_new['Year'])
```

```
<Axes: xlabel='Year', ylabel='Count'>
```

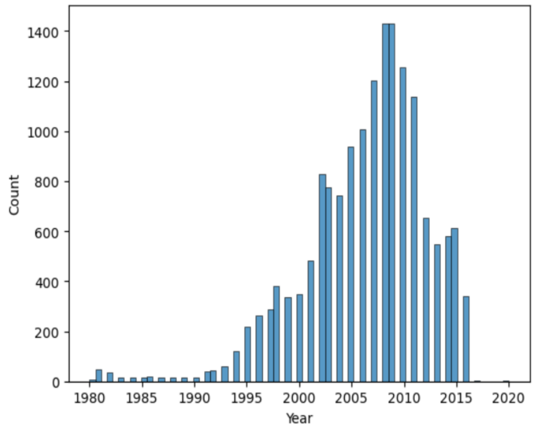
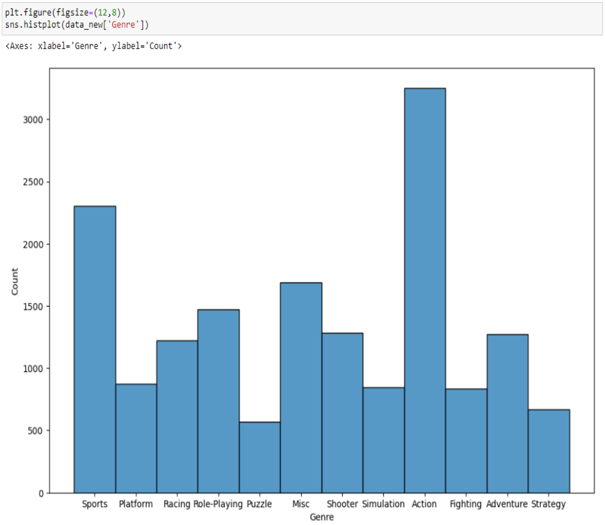


Fig.8 Year histogram

E. *Genre*

Since the first game was made, many different categories of games have emerged. In this column these game categories are listed according to the video games in the dataset.



```
plt.figure(figsize=(12,8))
sns.histplot(data_new['Genre'])
```

```
<Axes: xlabel='Genre', ylabel='Count'>
```

Fig.9 Genre histogram

F. Publisher

As games and players increased, gaming became an industry. Therefore, there are many different game producers and publishers. This column lists which publisher produced the video games in this dataset. When we examine the graph in fig 10, we observe that this graph is not a normally distributed graph and that some video game publishers produce many more games.

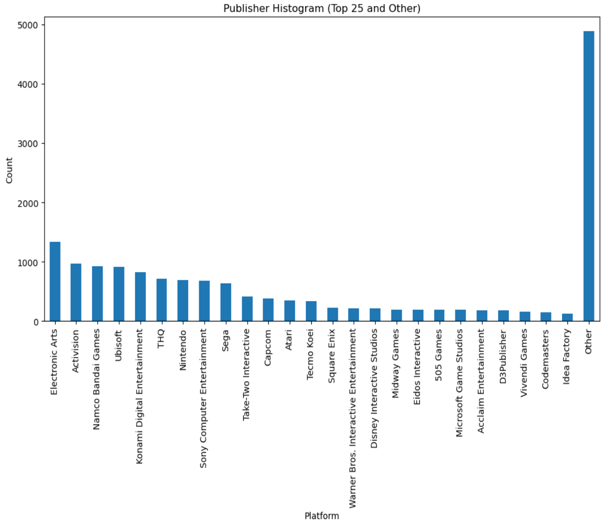
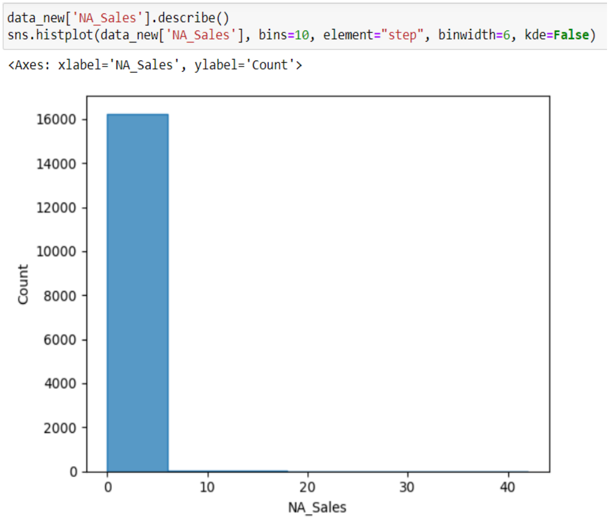


Fig.10 Publisher histogram

G. NA-Sales

This column lists the sales of all games released from 1980 to 2020 in North America. The data shows how many million games have been sold. Since the origin of video games is generally linked to North America, it can be said that they sell more here. Although the game that sold the most is 40 million, it is generally observed that games sell between 0 and 10 million.



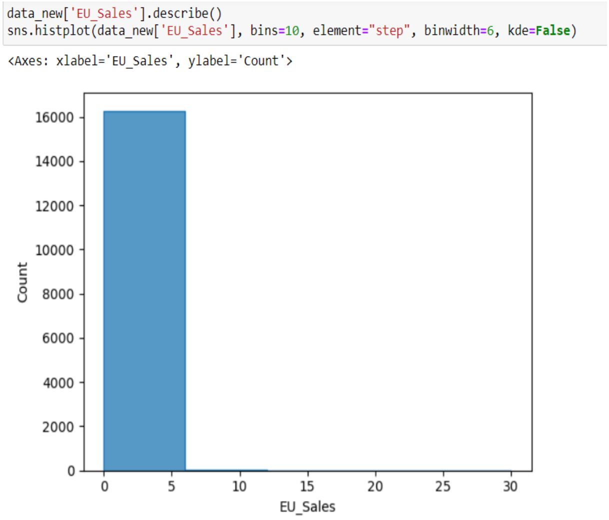
```
data_new['NA_Sales'].describe()
sns.histplot(data_new['NA_Sales'], bins=10, element="step", binwidth=6, kde=False)
```

```
<Axes: xlabel='NA_Sales', ylabel='Count'>
```

Fig.11 North America Sales histogram

H. EU-Sales

This column lists the sales of all games released from 1980 to 2020 in Europe. The data shows how many million games have been sold in Europe. Although the game that sold the most is 30 million, it is generally observed that games sell between 0 and 10 million. Europe sales is smaller than North America sales ,as can be seen it Fig.12.



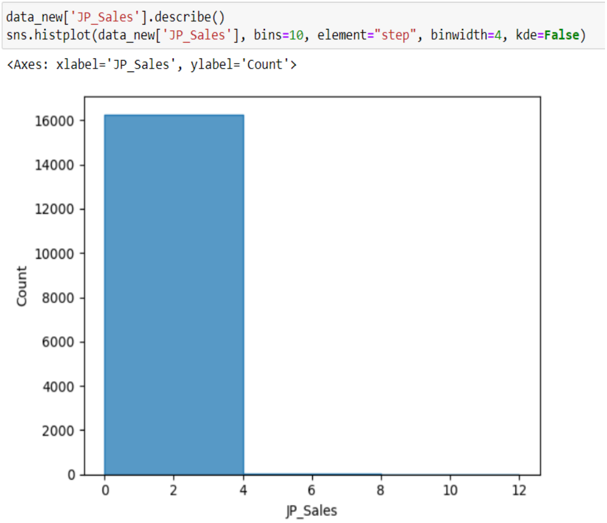
```
data_new['EU_Sales'].describe()
sns.histplot(data_new['EU_Sales'], bins=10, element="step", binwidth=6, kde=False)
```

```
<Axes: xlabel='EU_Sales', ylabel='Count'>
```

Fig.12 Europe Sales histogram

I. JP-Sales

This column lists the sales of all games released from 1980 to 2020 in Japan. The data shows how many million games have been sold in Japan. Although video game sales rates in Japan seem lower than the other two regions, Japan is a country that has had a big place in this industry from the very beginning, but the best-selling game is 12 million and in general, games seem to be sold between 0 and 4 million.



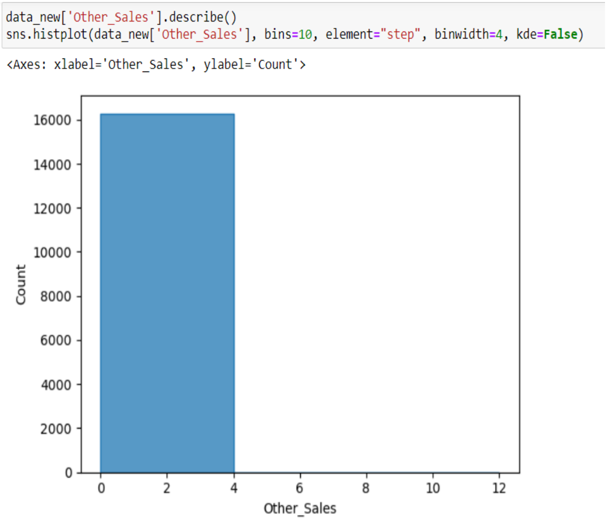
```
data_new['JP_Sales'].describe()
sns.histplot(data_new['JP_Sales'], bins=10, element="step", binwidth=4, kde=False)
```

```
<Axes: xlabel='JP_Sales', ylabel='Count'>
```

Fig.13 Japan sales histogram

J. Other-Sales

Although games are sold in large numbers in 3 different regions, there are also many people consuming game content in the rest of the world. This column shows the video game sales rates in the remaining countries and when we look at the chart of this column, we can understand that most of these sales are in North America, Europe and Japan.



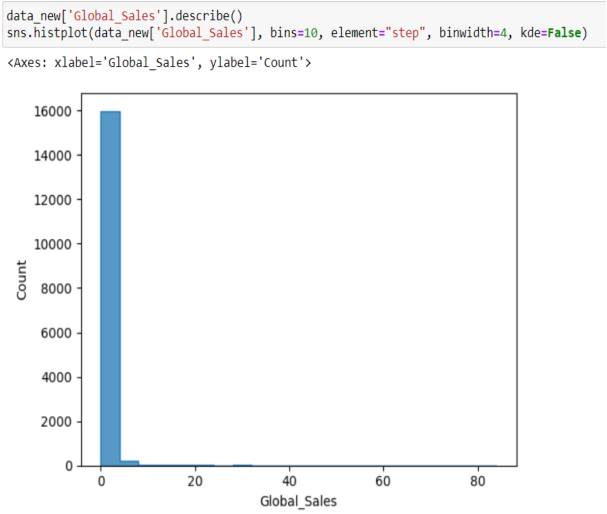
```
data_new['Other_Sales'].describe()
sns.histplot(data_new['Other_Sales'], bins=10, element="step", binwidth=4, kde=False)
```

```
<Axes: xlabel='Other_Sales', ylabel='Count'>
```

Fig.14 Other Sales histogram

K. Global-Sales

This column generally shows how much all games in the dataset have sold worldwide. So it is the sum of the other 4 columns.



```
data_new['Global_Sales'].describe()
sns.histplot(data_new['Global_Sales'], bins=10, element="step", binwidth=4, kde=False)
```

```
<Axes: xlabel='Global_Sales', ylabel='Count'>
```

Fig.15 Global Sales histogram

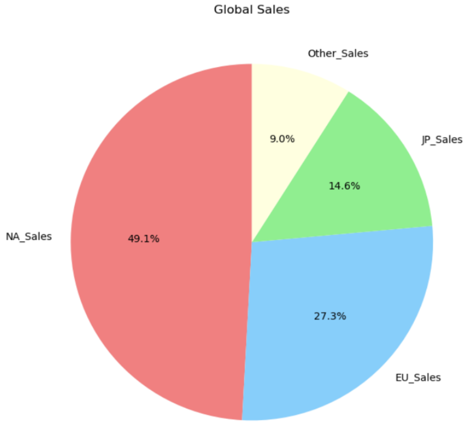


Fig.16 Global Sales Pie Chart

And finally, when we examine figure 16, we can see that almost half of the game sales are in North America. Additionally, as we have examined in other graphs, North America, Japan and Europe regions account for the majority of game sales.

III. Research Questions

A. The High Level Research Question

We have a data set containing the categories and sales rates of approximately 11,600 games. For this reason, I thought that the subject of the research should be to examine the relationship between these two more closely. This situation brought to my mind the following question that can be asked about this idea: How effective are video game categories in sales?

So what I want to investigate here is how video games are offered for sale in many different categories and how much the categories of these games affect the buyer's purchase of the game.

B. The Low Level Research Questions

Of course, to answer this high level question, I need to do some research and support it with some low level questions.

These low level questions should be indirectly related to the topic in the high level question and lead us to the conclusion.

- When we examine the test parameters, what are the status and connections of sales rates by region?
- Does the category distribution of games change depending on the years they were published? Could the interaction of these two affect the sales of the game?
- What are the interdependent or independent parameters that affect sales? How do they affect each other?

IV. Answers to low level questions

We have determined some low level questions to answer our main question. So, let's determine how well these low level questions help us achieve our main goal.

A. Status and Connections of Sale Rates by Region

With the pie chart we created in Figure 16, we examined that sales differ by region. Looking again, we can say that sales are highest in North America, followed by Europe, Japan and the rest of the world. So what is the connection between these sales? Let's examine how much the games sold in all regions.

We need to do some analysis and tests to understand the relationship between the regional distribution of sales with the help of the pie chart we obtained. The most logical one for this is the Anova test.

B. The Interaction of The Years and The Genres

What we want to achieve in this question is to find out whether certain video games are popular in certain years or whether years do not affect the categories.If we can find the answer to this, we will be one step closer to answering the question of how much video game categories affect game sales.First, let's create a graph to find the relationship between these two.

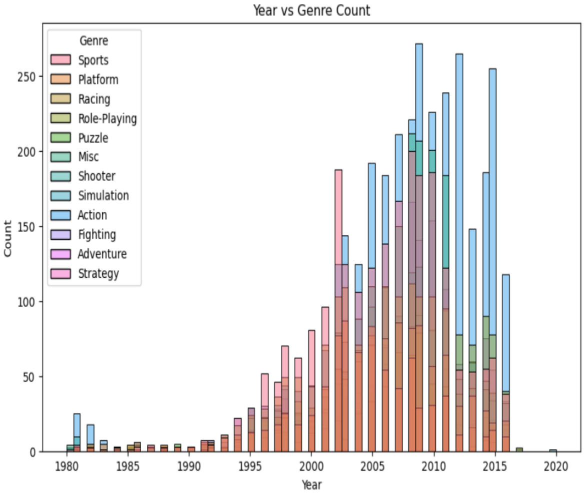


Fig.17 Year vs Genre Count

As can be seen in Fig.17, different categories of video games have different release rates in different years. But our high level question is not related to this issue. It seems to be a question that does not provide enough information to reach the answer to the high level question.

C. Parameters that Affect Sales

This question is actually a question aimed at finding a solution to the main problem. We can examine the factors affecting game sales under many different headings, but the part we want to examine in this analysis is how much do video game categories affect video game sales? So in this case, the parameters we want to examine are genre and global sales rates. Instead of examining the regions separately, let's examine the situation of total sales by categories.

The most logical way to compare the sales rates of 12 different game categories and reach a statistical conclusion would be an Anova test.

V. Constructing Hypothesis tests and methods

We formulate some hypotheses to apply the necessary tests to all these questions.

1) Null Hypothesis (H0): Regional differences do not affect sales
   Alternative Hypothesis (H1): Regional differences affect sales
2) Null Hypothesis (H0): There is no statistically significant difference in sales averages between game categories.
   Alternative Hypothesis (H1): The sales average in at least one of the game categories is statistically different from the others.

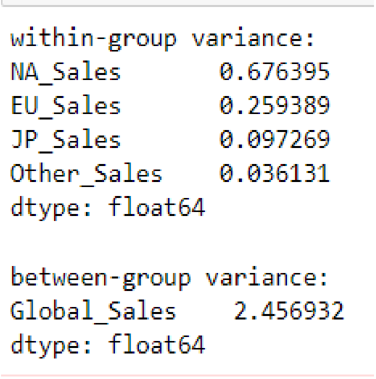
VI. Application of methods

A. First Hypothesis

- Null Hypothesis (H0): Regional differences do not affect sales
- Alternative Hypothesis (H1): Regional differences affect sales

First of all, we have 4 groups under the headings of North America, Europe, Japan and other regions. We will examine the relationships between these groups and within each other.

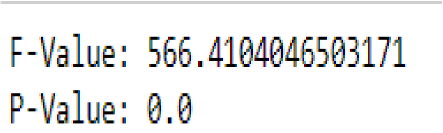
For this, we first need to make some calculations and find the within-group variance and between-group variance.



```
within-group variance:
NA_Sales        0.676395
EU_Sales        0.259389
JP_Sales        0.097269
Other_Sales     0.036131
dtype: float64

between-group variance:
Global_Sales    2.456932
dtype: float64
```

Fig.18 Variances for regions

After finding the variances, we can find the F value and p value after finding our actual values by finding the average of the variances within the group.



```
F-Value: 566.4104046503171
P-Value: 0.0
```

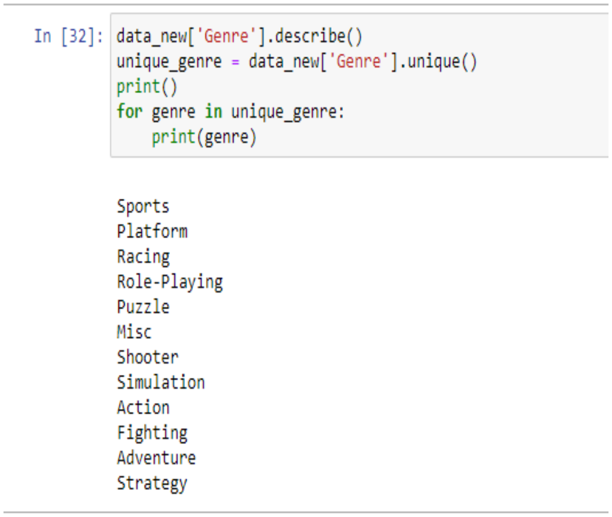
Fig.19 F value and p value for regions

As can be seen in Fig.19, the p value being 0.0 shows us that there is a statistically significant difference between the regions. Because 0.0 actually indicates that the p value is not exactly zero, but very close to 0. If we accept the significance level, i.e. alpha, as 0.05, if p is less than ten, this proves that the alternative hypothesis is correct. As a result, we can say that regional differences affect video game sales.

B. *Second Hypothesis*

- Null Hypothesis (H0): There is no statistically significant difference in sales averages between game categories.
- Alternative Hypothesis (H1): The sales average in at least one of the game categories is statistically different from the others.

After the first hypothesis, in which I evaluated regional sales, we created a new hypothesis to evaluate how sales interact according to categories. In this hypothesis, instead of separating sales regionally, we will make our transactions according to the total sales amount globally.



```
In [32]: data_new['Genre'].describe()
         unique_genre = data_new['Genre'].unique()
         print()
         for genre in unique_genre:
             print(genre)


         Sports
         Platform
         Racing
         Role-Playing
         Puzzle
         Misc
         Shooter
         Simulation
         Action
         Fighting
         Adventure
         Strategy
```

Fig.20 Video game genres in the dataset

As seen in Fig.20, there are a total of 12 video game categories in the dataset. Based on this information, what we need to do is to group the global sales rates according to these video game categories.

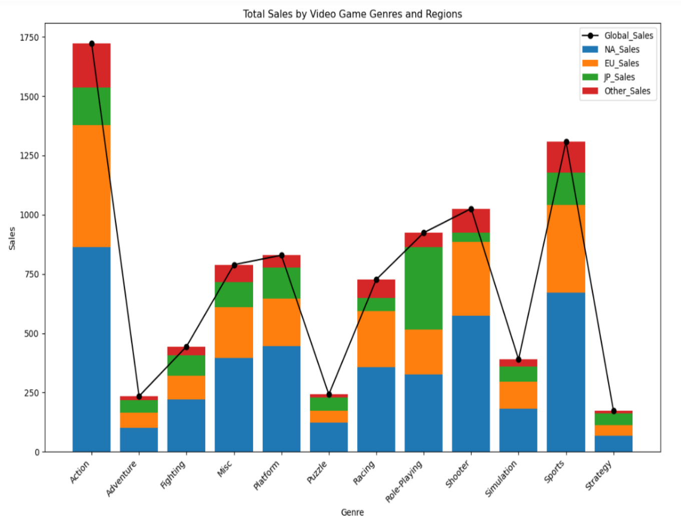
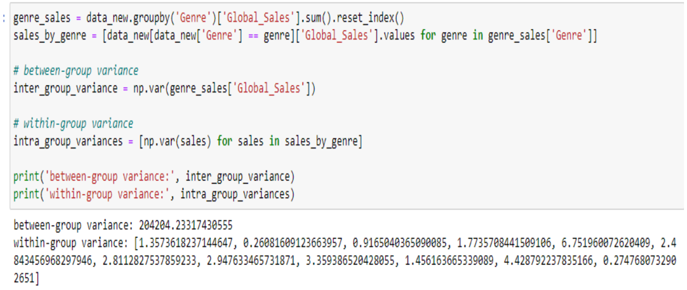


Fig.21 Total Sales by Video Game Genre and Regions

This grouping process is an important step in continuing the Anova test. After this step, what we need to do is find the variance between and within these groups and make the necessary calculations with the formula below.

$$F = \frac{MS_{between}}{MS_{within}}$$

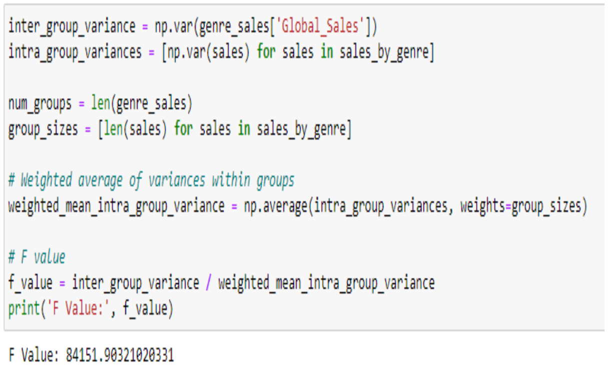
We calculated within-group and between-group variances.



```
: genre_sales = data_new.groupby('Genre')['Global_Sales'].sum().reset_index()
  sales_by_genre = [data_new[data_new['Genre'] == genre]['Global_Sales'].values for genre in genre_sales['Genre']]

  # between-group variance
  inter_group_variance = np.var(genre_sales['Global_Sales'])

  # within-group variance
  intra_group_variances = [np.var(sales) for sales in sales_by_genre]

  print('between-group variance:', inter_group_variance)
  print('within-group variance:', intra_group_variances)

  between-group variance: 204204.23317430555
  within-group variance: [1.3573618237144647, 0.26081609123663957, 0.9165040365090085, 1.7735708441509106, 6.751960072620409, 2.4
  843456968297946, 2.8112827537859233, 2.947633465731871, 3.359386520428055, 1.456163665339089, 4.428792237835166, 0.274768073290
  2651]
```

Fig.22 Variances

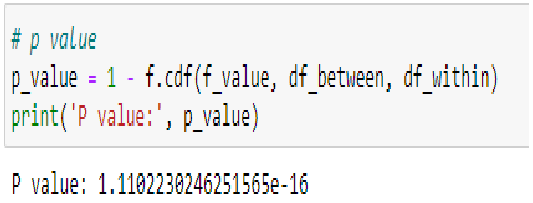
After this variance calculation, we need to find the F value so we can move on to the next step for the Anova test. First of all, we must find the weight average of the group-within variance that occurs more than once and calculate the ratio between the group-within variance and group-between variance.



```
inter_group_variance = np.var(genre_sales['Global_Sales'])
intra_group_variances = [np.var(sales) for sales in sales_by_genre]

num_groups = len(genre_sales)
group_sizes = [len(sales) for sales in sales_by_genre]

# Weighted average of variances within groups
weighted_mean_intra_group_variance = np.average(intra_group_variances, weights=group_sizes)

# F value
f_value = inter_group_variance / weighted_mean_intra_group_variance
print('F Value:', f_value)
```

```
F Value: 84151.90321020331
```

Now that we have found the F value, we can also find the p value and make a general evaluation.



```
# p value
p_value = 1 - f.cdf(f_value, df_between, df_within)
print('P value:', p_value)
```

```
P value: 1.1102230246251565e-16
```

Thanks to our calculations, we reached both the f value and the p value. As the last stage, we need to look at the relationship of these values according to the significance level we determined.

0.05 is generally used as the significance level. In this case, when we examine the p value, we can say that the value is very small and very close to zero. and finally;

If p $<$ alpha ,

our alternative hypothesis is confirmed and we can say that the sales average in at least one of the game categories is statistically different from the others.

VII. Conclusion

And as a result, with the help of the tests we conducted and the graphs we produced, we can say that video game categories have a global impact on video game sales. Although there are many different factors affecting video game sales, video game categories are definitely among these factors.