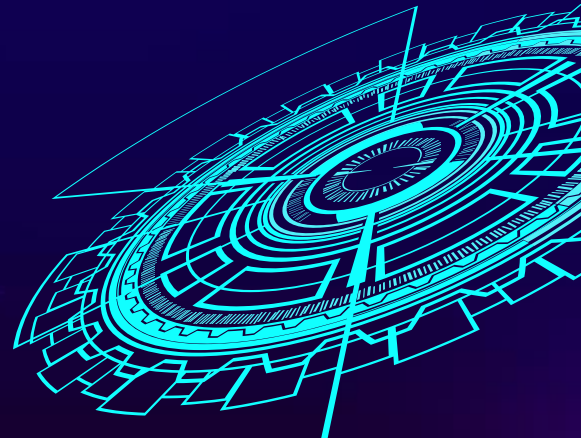


# Data Science Project



# Timeline



# The research question

- As gamers, we both have a personal connection to the subject matter and own multiple gaming consoles. Our shared passion for video games not only brings us enjoyment but also provides us with a financial incentive to make informed decisions about which consoles or genres to invest in for the future. This lead us to our research question:
- Can we predict the success of a particular game in different areas according to its sales ratio?**



# Data source

<https://www.vgchartz.com>


We chose this site thanks to its user friendly user interface, in addition it stores massive amounts of data regarding game sales throughout many years.

Scraping resulted in 65,768 games.

## Data Scraped

Each game consists of the following 11 properties, resulting in 723,448 items of data.

- Video game name
- Developer
- Platform
- Genre
- Release date
- Europe sales
- North America sales
- Japan sales
- “Other” sales
- Total sales
- Total shipments

Scraper 

```
def extract_items(genre, html_object):  
    """ The function below is responsible for extracting the content of the webpage obtained from the preceding function.  
    Once all the games on the page have been scraped, the data is sent to a saving function for storage.  
    The function takes the following inputs:  
    Genre: Data to be included in the JSON file.  
    HTML object: The HTML content obtained from the aforementioned function.  
    The function returns an error if the number of results on a page is less than 5.  
    This threshold accounts for the header and a few irrelevant div elements on the page.  
    If there are fewer than 5 items on a page, it indicates that the page contains no relevant  
    data for the given genre, signifying the completion of processing for that genre. """  
  
    try:  
        amount = 0  
        games_from_page = []  
        soup = bs(html_object, "html.parser")  
        table_of_games = soup.find('div', id="generalBody").find_all('tr')  
        if len(table_of_games) > 5:  
            del table_of_games[:3]  
            del table_of_games[-1]  
            for game in table_of_games:  
                row = game.find_all('td')  
                game_record = {'name': row[2].text,  
                               'developer': row[4].text,  
                               'platform': row[3].find('img')['alt'],  
                               'genre': genre,  
                               'total_shipments': row[5].text,  
                               'total_sales': row[6].text,  
                               'na_sales': row[7].text,  
                               'pal_sales': row[8].text,  
                               'japan_sales': row[9].text,  
                               'other_sales': row[10].text,  
                               'release': row[11].text  
                               }  
                amount = save_data(game_record, "database.json")  
            print(f'Collected: {amount}')  
            return True  
        else:  
            return False  
    except Exception as e:  
        return False
```

# Work Flow

- ◀ We used python scraping libraries such as Requests and BeautifulSoup 4.
- ◀ We used “Requests” library to send requests to the website <https://www.vgchartz.com>  
The number of games shown per page is custom, ranging from 10 to 200 records.
- ◀ A list of genres was chosen to iterate over all pages, which allowed us to scrape all of the data in one continuous run.
- ◀ We scraped all of the data for each genre, when we got to the last page, it loaded, but showed no content.  
That was how we knew to continue to the next genre.
- ◀ After completing the scraping, the data was saved in a Json file, which was later transferred to a CSV file for further work on the data.
- ◀ We used the `time.sleep(n)` function when we got 503 status code, which allowed us to run the requests repeatedly and retry attempts until we did establish a connection to the webpage.





# Cleaning the Data

# Cleaning the Data

Before

	name	developer	platform	genre	total_sales	na_sales	pal_sales	japan_sales	other_sales	release
0	God of War	SIE Santa Monica Studio	Series	Action	NaN	NaN	NaN	NaN	NaN	22nd Mar 05
1	Warriors	Omega Force	Series	Action	NaN	NaN	NaN	NaN	NaN	30th Jun 97
2	Devil May Cry	Capcom	Series	Action	NaN	NaN	NaN	NaN	NaN	16th Oct 01
3	God of War (2018)	SIE Santa Monica Studio	All	Action	NaN	NaN	NaN	NaN	NaN	20th Apr 18
4	Dynasty Warriors	Omega Force	Series	Action	NaN	NaN	NaN	NaN	NaN	30th Jun 97
5	Grand Theft Auto V Read the review	Rockstar North	PS3	Action	20.32m	6.37m	9.85m	0.99m	3.12m	17th Sep 13
6	Frogger	Konami	Series	Action	NaN	NaN	NaN	NaN	NaN	23rd Oct 81
7	God of War (2018) Read the review	SIE Santa Monica Studio	PS4	Action	NaN	NaN	NaN	NaN	NaN	20th Apr 18
8	Grand Theft Auto V	Rockstar North	PS4	Action	19.39m	6.06m	9.71m	0.60m	3.02m	18th Nov 14
9	Grand Theft Auto: San Andreas	Rockstar North	PS2	Action	NaN	NaN	NaN	NaN	NaN	26th Oct 04
10	Uncharted 4: A Thief's End	Naughty Dog	PS4	Action	NaN	NaN	NaN	NaN	NaN	10th May 16
11	Grand Theft Auto: Vice City	Rockstar North	PS2	Action	16.15m	8.41m	5.49m	0.47m	1.78m	28th Oct 02
12	Grand Theft Auto V	Rockstar North	X360	Action	15.86m	9.06m	5.33m	0.06m	1.42m	17th Sep 13
13	Grand Theft Auto III	DMA Design	PS2	Action	13.10m	6.99m	4.51m	0.30m	1.30m	23rd Oct 01
14	Grand Theft Auto V	Rockstar North	PC	Action	NaN	NaN	NaN	NaN	NaN	14th Apr 15

65768 rows × 10 columns

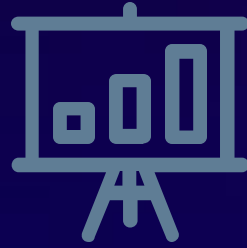
## Work Flow

- ◀ Changed N/A items to NaN in the “release” column.
- ◀ Deleted the ‘m’ (as in “million”) in our sales columns.
- ◀ Changed pal\_sales (Europe, New Zealand, Australia, Middle East, India, and South Africa) to eu\_sales for convenience.
- ◀ Changed date format in “release” column to year only.
- ◀ A row with the same game name and platform as another row is considered a duplicate row, which we deleted.

After

	name	developer	platform	genre	total_sales	na_sales	eu_sales	japan_sales	other_sales	release_year
0	Grand Theft Auto V Read the review	Rockstar North	PS3	Action	20.32	6.37	9.85	0.99	3.12	2013
1	Grand Theft Auto V	Rockstar North	PS4	Action	19.39	6.06	9.71	0.60	3.02	2014
2	Grand Theft Auto: Vice City	Rockstar North	PS2	Action	16.15	8.41	5.49	0.47	1.78	2002
3	Grand Theft Auto V	Rockstar North	X360	Action	15.86	9.06	5.33	0.06	1.42	2013
4	Grand Theft Auto III	DMA Design	PS2	Action	13.10	6.99	4.51	0.30	1.30	2001
...	...	...	...	...	...	...	...	...	...	...
18914	Nora, Princess, and Stray Cat	Harukaze	NS	Visual Novel	0.00	0.01	0.01	0.00	0.01	2018
18915	Memories Off: Innocent File	5pb. Games	NS	Visual Novel	0.00	0.01	0.01	0.00	0.01	2018
18916	Enkan no Memoria: Kakeru Tomoshi	A'sRing	PSV	Visual Novel	0.00	0.01	0.01	0.00	0.01	2018
18917	Disorder 6	5pb. Games	X360	Visual Novel	0.00	0.01	0.01	0.00	0.01	2013
18918	Zero Escape: Virtue's Last Reward Read the...	ChunSoft	PSV	Visual Novel	0.00	0.01	0.00	0.01	0.00	2012

18919 rows × 10 columns



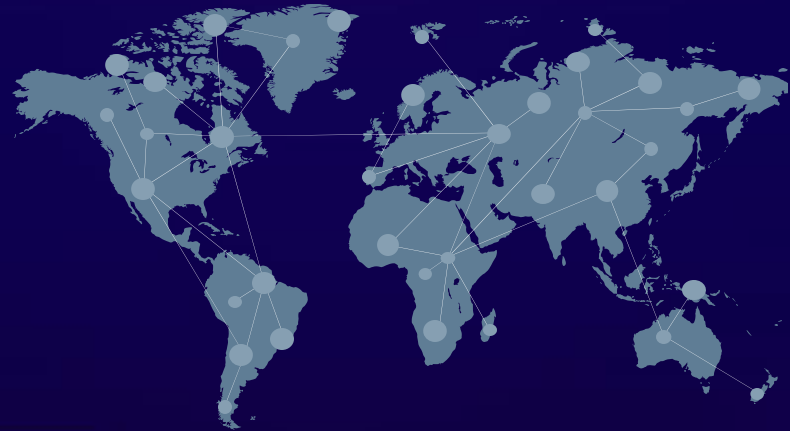
# EDA – Exploratory Data Analysis



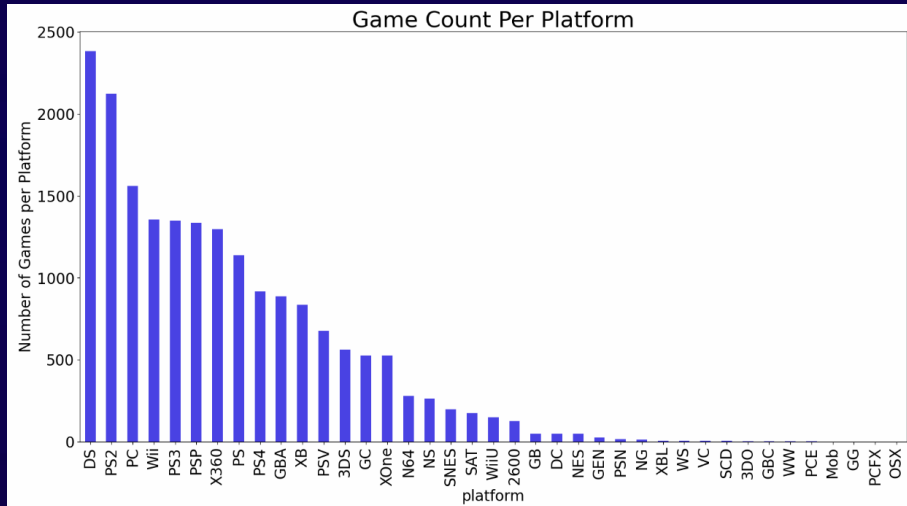
# The Dataset Categories

	name	developer	platform	genre	total_sales	na_sales	eu_sales	japan_sales	other_sales	release_year
0	Grand Theft Auto V Read the review	Rockstar North	PS3	Action	20.32	6.37	9.85	0.99	3.12	2013

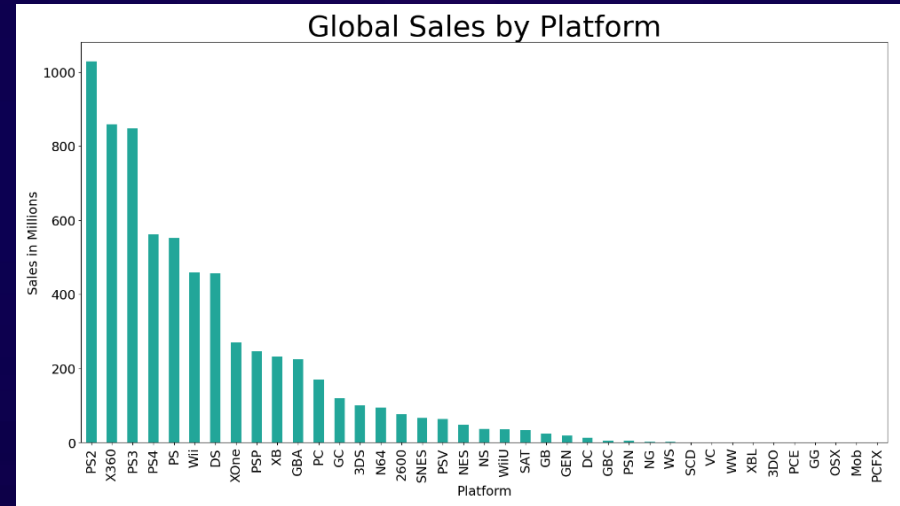
- ◀ name: Game name.
- ◀ platform: Name of the gaming device. For example, DS.
- ◀ genre: Genre of the game. For example “GTA V” – Action.
- ◀ total\_sales: Game Sales for all markets.
- ◀ na\_sales: Game Sales in North American market.
- ◀ eu\_sales: Game Sales in European market.
- ◀ japan\_sales: Game Sales in Japanese market.
- ◀ other\_sales: Game Sales in the rest of the regions’ markets.
- ◀ release\_year: The year in which the game was released.



# Exploratory Data Analysis

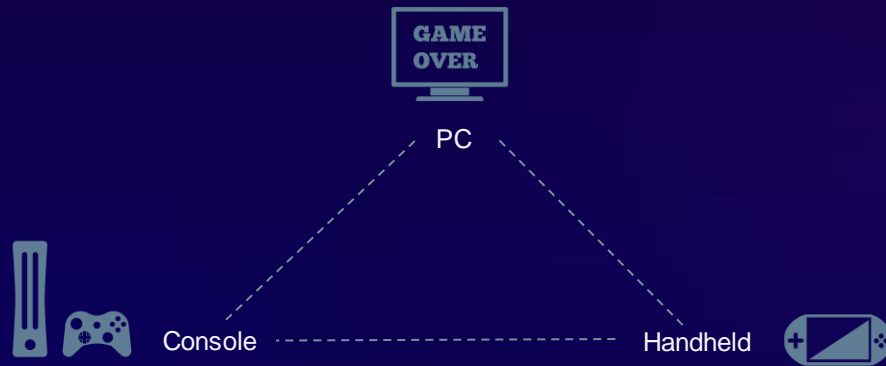


The plot above shows how many games per platform our data contains.  
All platforms will be generalized to 3 platform types: handhelds, consoles, and PC.



PS2 has sold the most games even though DS holds the highest game count.

# Exploratory Data Analysis



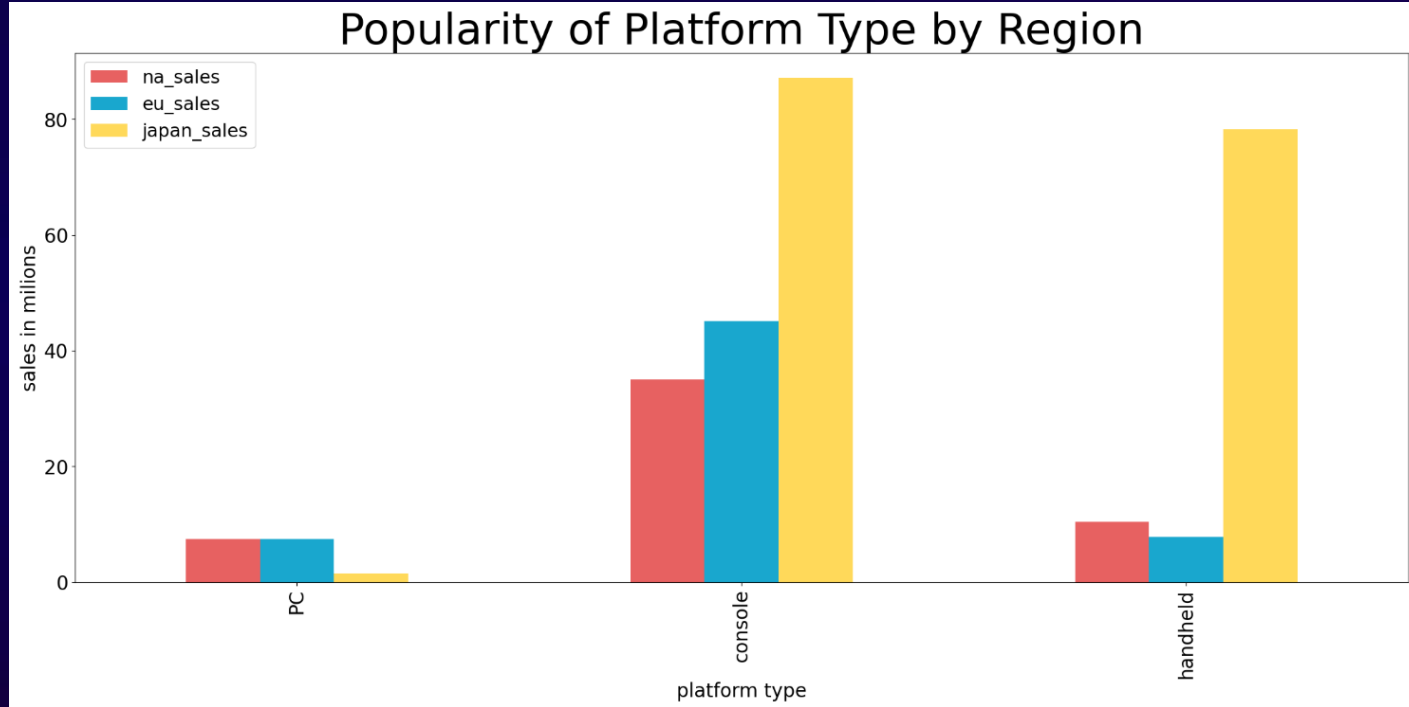
A directory was needed in order to map consoles to console type:

```
platform_type = {'2600': 'console', '3DO': 'console', '3DS': 'handheld', 'DC': 'console', 'DS': 'handheld',  
                'GB': 'handheld', 'GBA': 'handheld', 'GBC': 'handheld', 'GC': 'console', 'GEN': 'console',  
                'GG': 'console', 'Mob': 'handheld', 'N64': 'console', 'NES': 'console', 'NG': 'console',  
                'NS': 'handheld', 'OSX': 'PC', 'PC': 'PC', 'PCE': 'console', 'PCFX': 'console', 'PS': 'handheld',  
                'PS2': 'console', 'PS3': 'console', 'PS4': 'console', 'PSN': 'PC', 'PSP': 'handheld',  
                'PSV': 'handheld', 'SAT': 'console', 'SCD': 'handheld', 'SNES': 'console', 'VC': 'console',  
                'Wii': 'console', 'WiiU': 'handheld', 'WS': 'handheld', 'WW': 'console', 'X360': 'console',  
                'XB': 'console', 'XBL': 'console', 'XOne': 'console'}
```

```
df["platform_type"] = df["platform"].map(lambda x: platform_type[x])
```

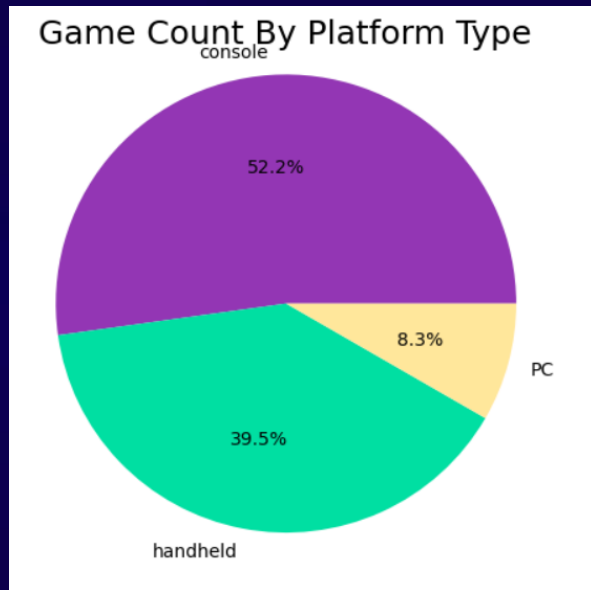


# Exploratory Data Analysis

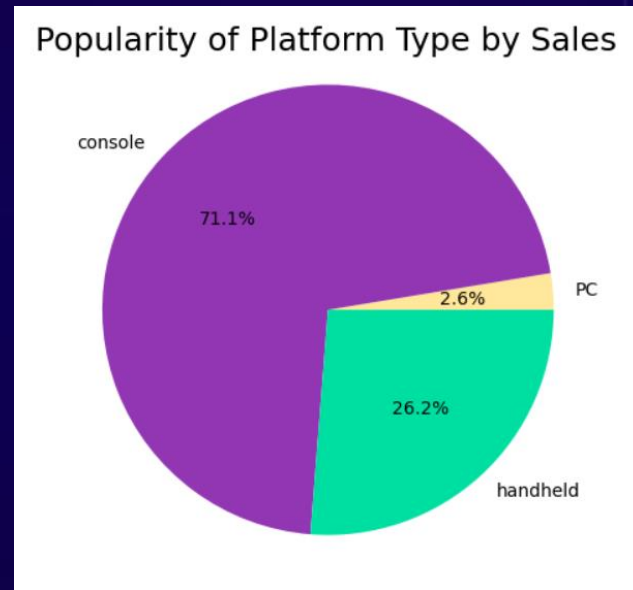


Japan holds the vast majority of global sales of console and handheld platforms.

# Exploratory Data Analysis



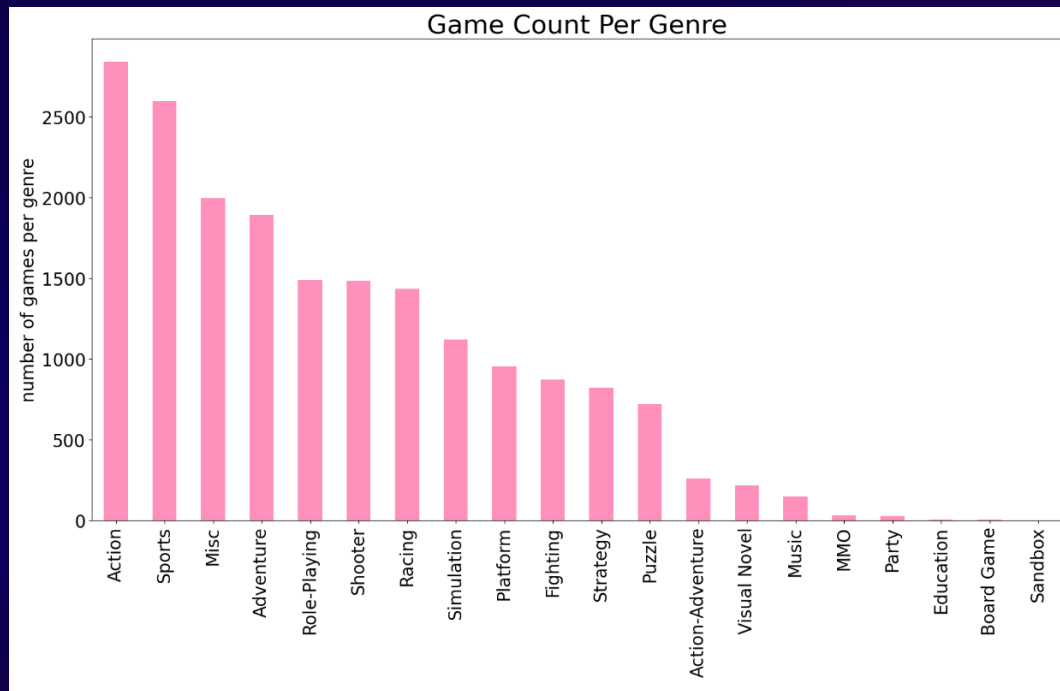
Percentage of games of a certain platform (PC, Handheld and Console) from the df.



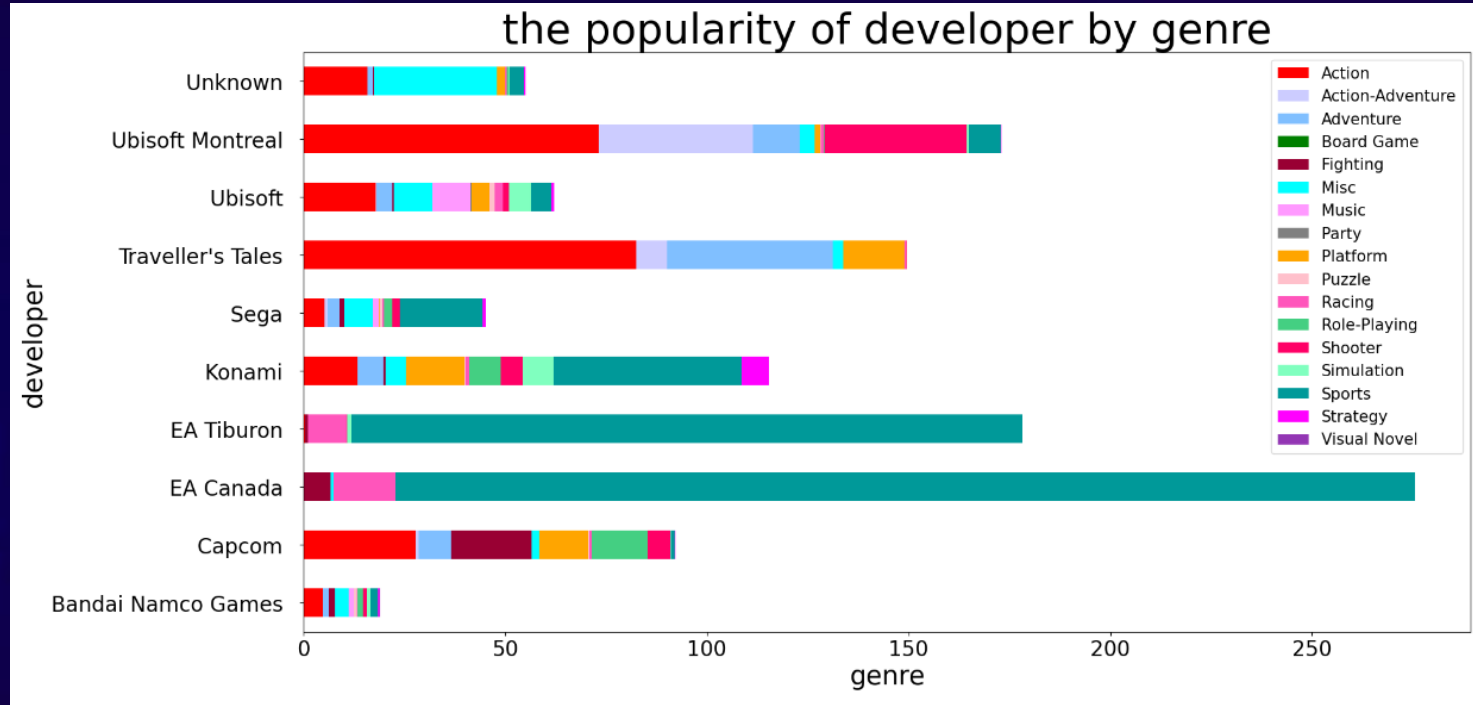
Percentage of game sales for each platform.

More than half of the games sold are console games. The next pie chart will show the correlating sales.

# Exploratory Data Analysis



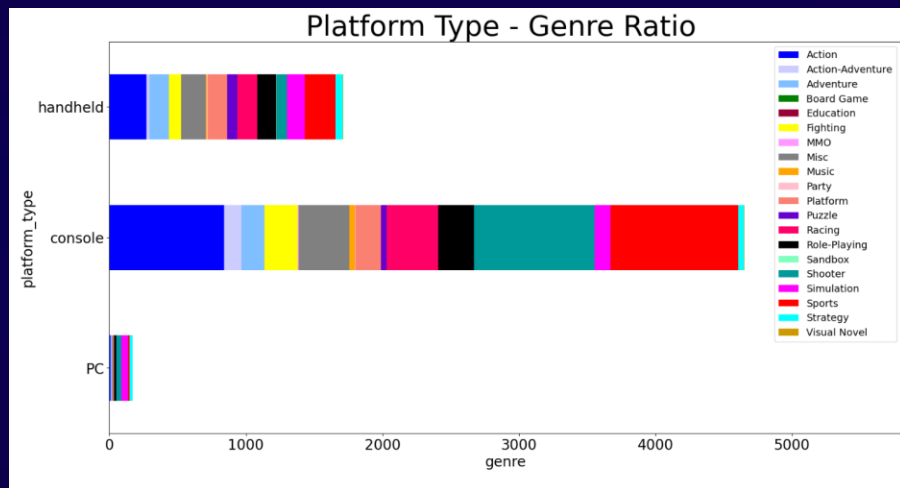
The action – genre games are developed the most.



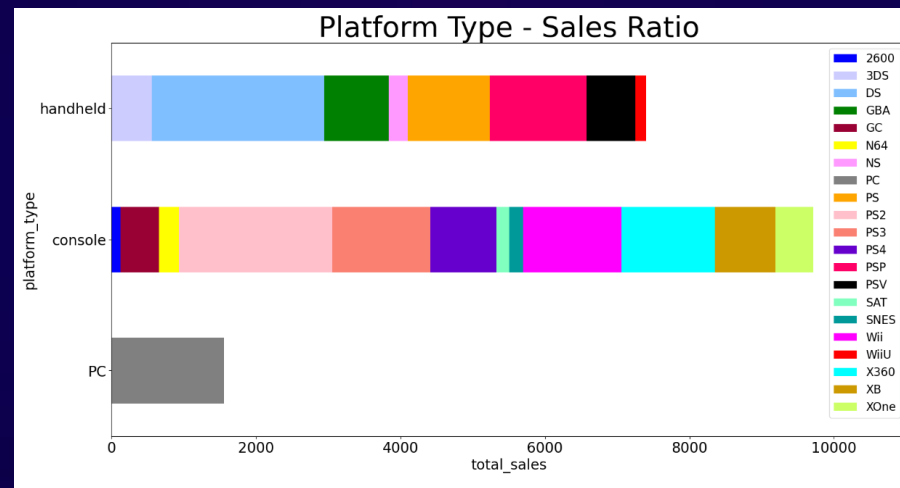
This graph depicts the genre distribution of developers.

We can notice a dominant genre in most development companies, for example: EA specializes in the sports genre and sticks to it.

# Exploratory Data Analysis



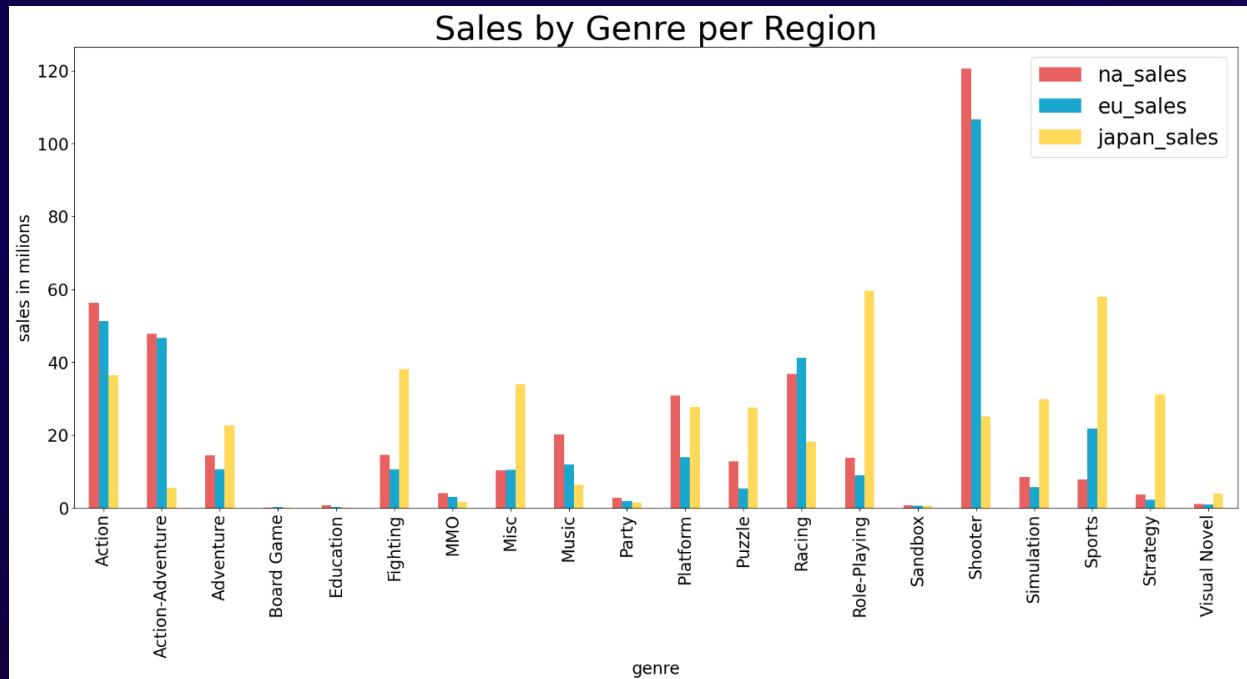
Above is the visualization of most developed consoles by platform type (which console has the most games, by platform type).



Above is the visualization of the best selling consoles by platform type (which console has sold the most games, by platform type).



# Exploratory Data Analysis



We took the 100 best-selling games in each genre, and examined their sales in each region

Shooter games are much more successful in NA and Europe than in Japan

Most of Japan's best selling genres are North America's and Europe's least selling genres, which shows the contrast between said cultures.

While the Japanese market does not have a clear leading genre, the European and North American markets have a clear leading genre: Shooter.

# Machine Learning

# Machine Learning

Our research question:

**“Can we predict the success of a particular game in different areas according to its sales ratio?”**

is categorized as a classification problem.

```
column_names_to_normalize = ['other_sales', 'na_sales', 'eu_sales', 'japan_sales']
min_max_scaler = preprocessing.MinMaxScaler()
x = df[column_names_to_normalize].values
x_scaled = min_max_scaler.fit_transform(x)
df_temp = pd.DataFrame(x_scaled, columns=column_names_to_normalize, index = df.index)
df[column_names_to_normalize] = df_temp
```

Here we use the “fit\_transform” method of “min\_max\_scaler” to normalize the data stored in x.

This method scales the values in x to a specific range, typically between 0 and 1, based on the minimum and maximum values of the data.

	name	developer	platform	genre	na_sales	eu_sales	japan_sales	other_sales	release_year	platform_type
0	Grand Theft Auto V Read the review	Rockstar North	PS3	Action	0.652664	1.000000	0.456221	1.000000	2013	console
1	Grand Theft Auto V	Rockstar North	PS4	Action	0.620902	0.985787	0.276498	0.967949	2014	console
2	Grand Theft Auto: Vice City	Rockstar North	PS2	Action	0.861680	0.557360	0.216590	0.570513	2002	console
3	Grand Theft Auto V	Rockstar North	X360	Action	0.928279	0.541117	0.027650	0.455128	2013	console
4	Grand Theft Auto III	DMA Design	PS2	Action	0.716189	0.457868	0.138249	0.416667	2001	console
...	...	...	...	...	...	...	...	...	...	...
18914	Nora, Princess, and Stray Cat	Harukaze	NS	Visual Novel	0.001025	0.001015	0.000000	0.003205	2018	handheld
18915	Memories Off: Innocent File	5pb. Games	NS	Visual Novel	0.001025	0.001015	0.000000	0.003205	2018	handheld
18916	Enkan no Memoria: Kakera Tomoshi	A'sRing	PSV	Visual Novel	0.001025	0.001015	0.000000	0.003205	2018	handheld
18917	Disorder 6	5pb. Games	X360	Visual Novel	0.001025	0.001015	0.000000	0.003205	2013	console
18918	Zero Escape: Virtue's Last Reward Read the...	ChunSoft	PSV	Visual Novel	0.001025	0.000000	0.004608	0.000000	2012	handheld

18919 rows × 10 columns

# Machine Learning

- Afterwards we checked which region in every game had the most sales in relation to the other regions. The resulting boolean values are then converted to integers (0 = False , 1 = True) using the astype(int) function.

```
df['na_succses'] = (df['max_region'].str.lower() == 'na_sales').astype(int)
df['eu_succses'] = (df['max_region'].str.lower() == 'eu_sales').astype(int)
df['japan_succses'] = (df['max_region'].str.lower() == 'japan_sales').astype(int)
df['other_succses'] = (df['max_region'].str.lower() == 'other_sales').astype(int)
df
```

- The Dummy columns we added to represent the best selling region of each game:

max_region	na_succses	eu_succses	japan_succses	other_succses
eu_sales	0	1	0	0
eu_sales	0	1	0	0
na_sales	1	0	0	0
na_sales	1	0	0	0
na_sales	1	0	0	0

# Classification models

- We used 3 classification models:  
logistic regression, KNN and decision tree.
- We transferred the dataset from categorial to numeric for compatibility, then ran the models and tried to predict the na\_succses

	developer	genre	na_sales	eu_sales	japan_sales	other_sales	release_year	platform_type	na_succses
0	2067	0	0.652664	1.000000	0.456221	1.000000	2013	1	0
1	2067	0	0.620902	0.985787	0.276498	0.967949	2014	1	0
2	2067	0	0.861680	0.557360	0.216590	0.570513	2002	1	1
3	2067	0	0.928279	0.541117	0.027650	0.455128	2013	1	1
4	621	0	0.716189	0.457868	0.138249	0.416667	2001	1	1
...	...	...	...	...	...	...	...	...	...
18914	1099	19	0.001025	0.001015	0.000000	0.003205	2018	2	0
18915	44	19	0.001025	0.001015	0.000000	0.003205	2018	2	0
18916	57	19	0.001025	0.001015	0.000000	0.003205	2018	2	0
18917	44	19	0.001025	0.001015	0.000000	0.003205	2013	1	0
18918	490	19	0.001025	0.000000	0.004608	0.000000	2012	2	0

18919 rows × 9 columns

## Logistic regression:

LR y prediction TRAIN

	Predicted Negative	Predicted Positive
Actual Negative	8907	190
Actual Positive	675	5363

## NA Model

### KNN:

KNN y prediction TRAIN

	Predicted Negative	Predicted Positive
Actual Negative	8868	229
Actual Positive	209	5829

### Decision tree:

DT y prediction TRAIN

	Predicted Negative	Predicted Positive
Actual Negative	9025	72
Actual Positive	22	6016

LR y prediction TEST

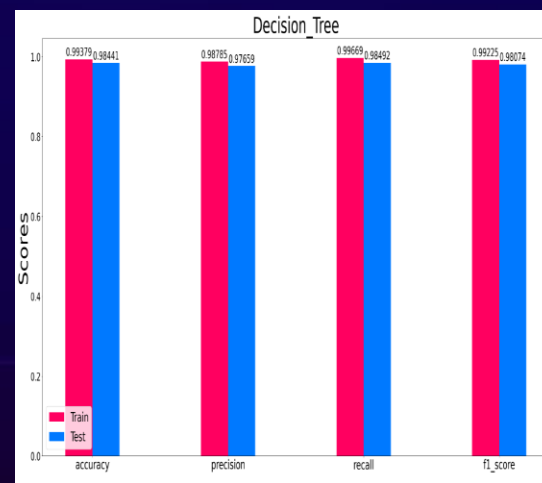
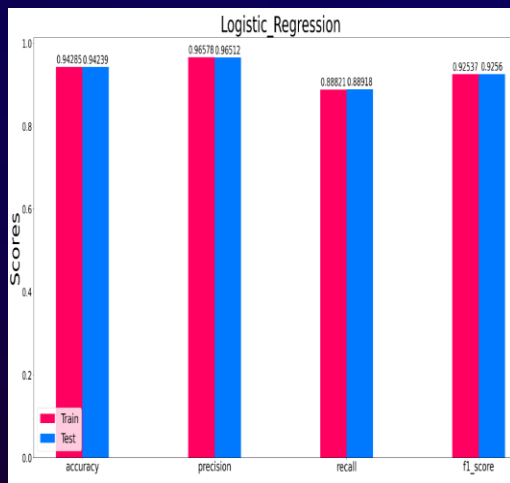
	Predicted Negative	Predicted Positive
Actual Negative	2210	49
Actual Positive	169	1356

KNN y prediction TEST

	Predicted Negative	Predicted Positive
Actual Negative	2148	111
Actual Positive	126	1399

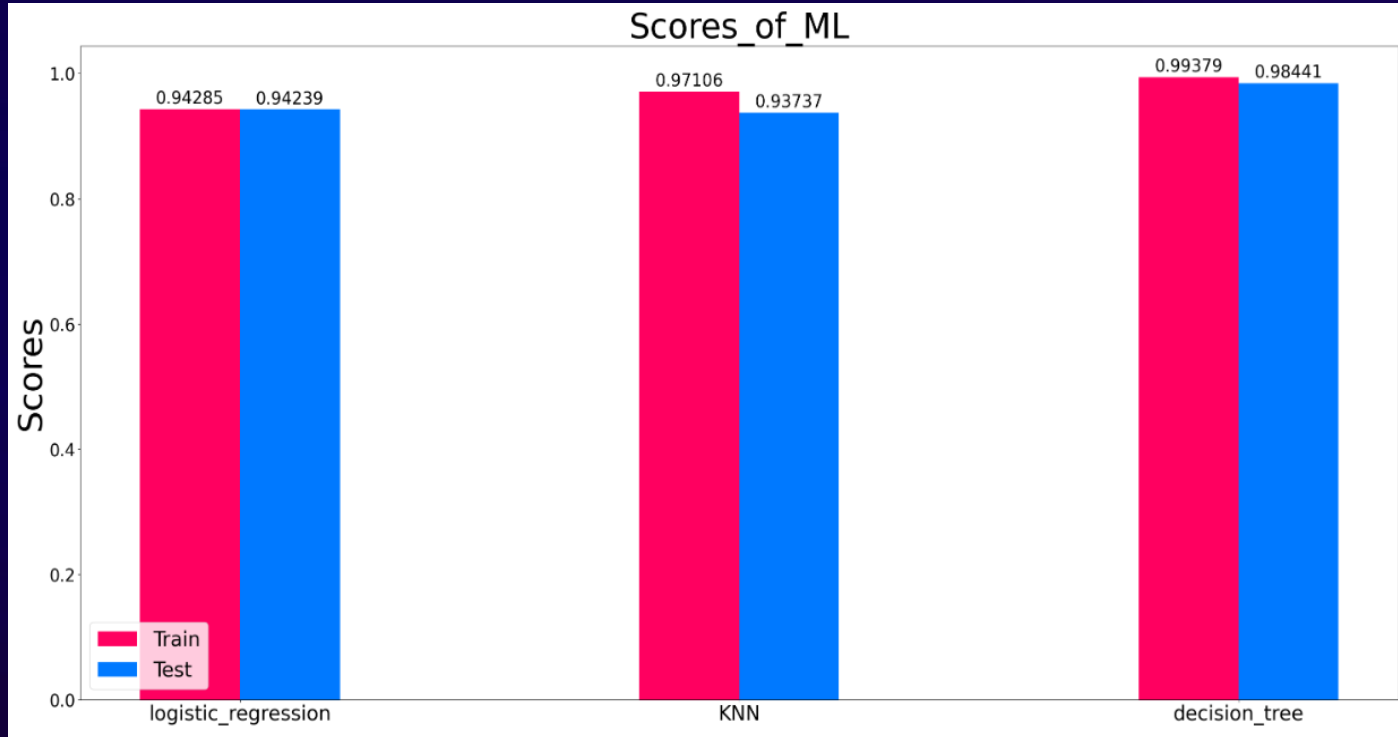
DT y prediction TEST

	Predicted Negative	Predicted Positive
Actual Negative	2223	36
Actual Positive	21	1504

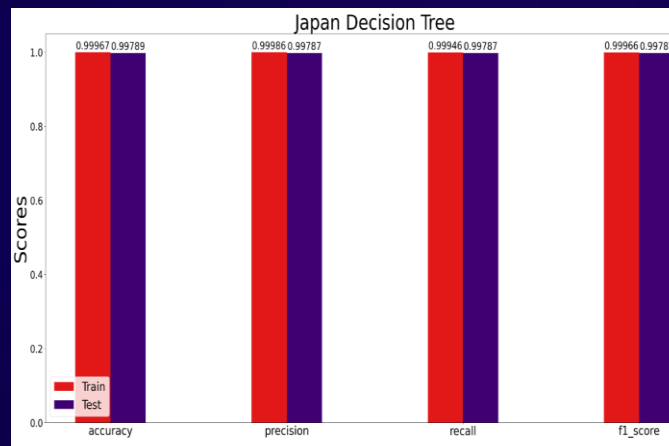
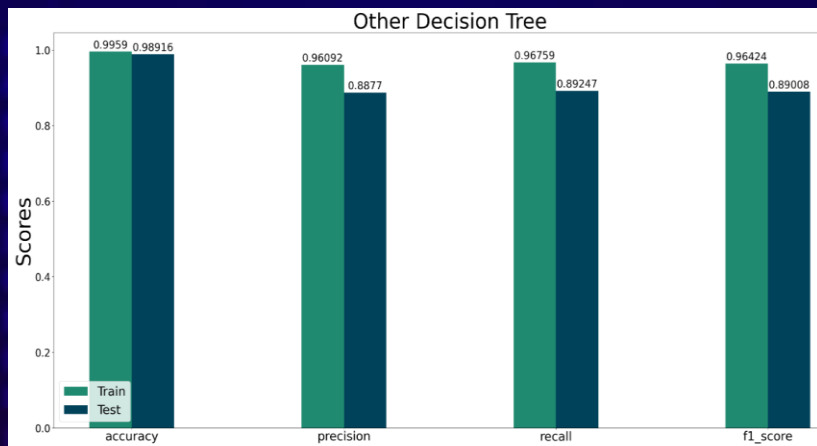
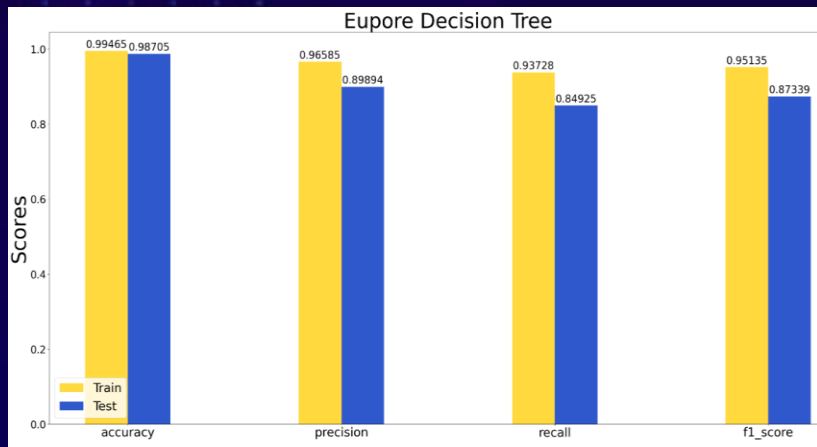


# Classification models - accuracy

- We infer that the best model to use is decision tree, so we will continue with it to predict the rest of the regions' success.



## Decision trees of rest of regions





# Conclusions

Globally, console is the best-selling platform type

There isn't always a linear correlation between a platform's games count to its' total sales

The Japanese market differs greatly from the American and European markets, both in terms of platform popularity and video game genre popularity

Shooter, sports and action genre were most successful for videogames.  
Shooter being the most successful.

We can notice a dominant genre in most development companies, for example:  
EA specializes in the sports genre and sticks to it.

Sources used: Jupyter Notebook, Stack Overflow, Campus IL, OpenAI, Youtube.