



**Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)**

**Факультет «Информатика и системы управления»
Кафедра ИУ5 «Системы обработки информации и управления»**

**Отчёт по лабораторной работе
по дисциплине «Методы Машинного Обучения»**

Выполнил:
студент группы № ИУ5-21М
Кучеренко Михаил Александрович
_____, _____

Проверил:
к.т.н., доц., Ю.Е. Гапанюк
_____, _____

2021 г.

Лабораторная работа №1

Создание "истории о данных" (Data Storytelling).

Рассмотрим исторические данные по выходу и продажам видеоигр из [на 2019 год](#)

Задание:

Оригинал

Выбрать набор данных (датасет). Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

- История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
- На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
- Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
- Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
- История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib
4 import matplotlib.pyplot as plt
5 import seaborn as sns
6 %matplotlib inline
```

```
1 from joypy import joyplot
```

```
1 sns.set(style="darkgrid")
```

```
1 df = pd.read_csv('vgsales-12-4-2019-short.csv')
2 print(f'Total loaded {len(df.index)} video games')
```

```
1 Total loaded 55792 video games
```

Немного предварительной обработки кривого датасета

```
1 # count NaNs by each column
2 df.isna().sum()
```

```
1 Rank          0
2 Name          0
3 Genre         0
4 ESRB_Rating   32169
5 Platform      0
6 Publisher     0
7 Developer     17
8 Critic_Score  49256
9 User_Score    55457
10 Total_Shipped 53965
11 Global_Sales 36377
12 NA_Sales     42828
13 PAL_Sales    42603
14 JP_Sales     48749
15 Other_Sales  40270
16 Year         979
17 dtype: int64
```

```

1 # Объединим продажи в одну колонку
2 df['Sales'] = df[['Total_Shipped', 'Global_Sales']].max(axis=1)

```

```

1 # Определим колонки, пустые более чем на 65%
2 is_part_empty_threshold = 0.65

```

```

1 part_empty_cols = []
2 for column, count in df.isna().sum().items():
3     if (count / len(df.index)) > is_part_empty_threshold:
4         part_empty_cols.append(column)
5
6 print(part_empty_cols)

```

```

1 ['Critic_Score', 'User_Score', 'Total_Shipped', 'Global_Sales', 'NA_Sales', 'PAL_Sales', 'JP_Sales',
  'Other_Sales']

```

```

1 # Уберем плохо заполненные и не особо важные (т.к. ввели Sales) колонки
2 df = df.drop(columns = ['Total_Shipped', 'Global_Sales', 'NA_Sales', 'PAL_Sales', 'JP_Sales',
  'Other_Sales'])

```

```

1 df.head()

```

```

1 .dataframe tbody tr th {
2     vertical-align: top;
3 }
4
5 .dataframe thead th {
6     text-align: right;
7 }

```

	Rank	Name	Genre	ESRB_Rating	Platform	Publisher	Developer	Critic_Score	User_Score	Year
0	1	Wii Sports	Sports	E	Wii	Nintendo	Nintendo EAD	7.7	NaN	2006.
1	2	Super Mario Bros.	Platform	NaN	NES	Nintendo	Nintendo EAD	10.0	NaN	1985.
2	3	Mario Kart Wii	Racing	E	Wii	Nintendo	Nintendo EAD	8.2	9.1	2008.
3	4	PlayerUnknown's Battlegrounds	Shooter	NaN	PC	PUBG Corporation	PUBG Corporation	NaN	NaN	2017.
4	5	Wii Sports Resort	Sports	E	Wii	Nintendo	Nintendo EAD	8.0	8.8	2009.

Шаг 1 - Количество игр по году выхода

```

1 df.loc[df['Year'].isna()].head()

```

```

1 .dataframe tbody tr th {
2     vertical-align: top;
3 }
4
5 .dataframe thead th {
6     text-align: right;
7 }

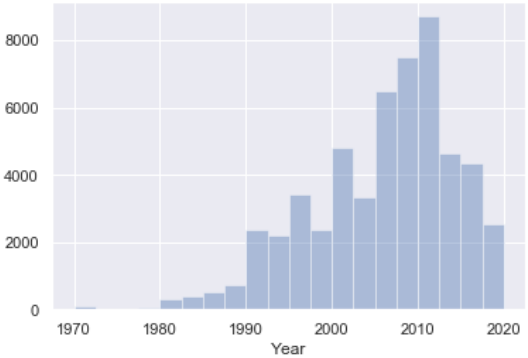
```

	Rank	Name	Genre	ESRB_Rating	Platform	Publisher	Developer	Critic_Score	User_Score	Year	Sales
15224	15225	Tour de France 2011	Sports	E	X360	Unknown	Cyanide Studio	NaN	NaN	NaN	0.05
15653	15654	The History Channel: Great Battles - Medieval	Strategy	NaN	PS3	Unknown	Slitherine Software	NaN	NaN	NaN	0.04
15791	15792	B.L.U.E.: Legend of Water	Adventure	NaN	PS	Unknown	Unknown	NaN	NaN	NaN	0.04
20065	20066	Wii de Asobu: Metroid Prime 2: Dark Echoes	Shooter	NaN	Wii	Unknown	Retro Studios	NaN	NaN	NaN	0.00
20086	20087	Iron Master: The Legendary Blacksmith	Action	RP	DS	Unknown	Barunson Creative	NaN	NaN	NaN	0.00

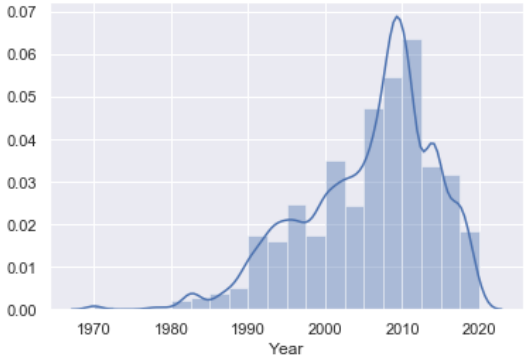
```
1 # Очевидно это не очень успешные игры от неизвестных разработчиков, в топку
2 df = df.dropna(subset=['Year', 'Developer'])
```

Т.к. нам еще не известна структура данных то для визуализации хорошо подойдет простое распределение плотности количества игр на года.

```
1 sns.distplot(df["Year"], kde=False, norm_hist=False, bins=20)
2 plt.show()
```



```
1 sns.distplot(df["Year"], hist=True, bins=20)
2 plt.show()
```



Видим, что пик выхода видео игр приходится на 2010-2012 годы.
Рискну предположить, что индустрия наполнялась проектами, после чего стало выгодным издавать только наиболее успешные и крупные.

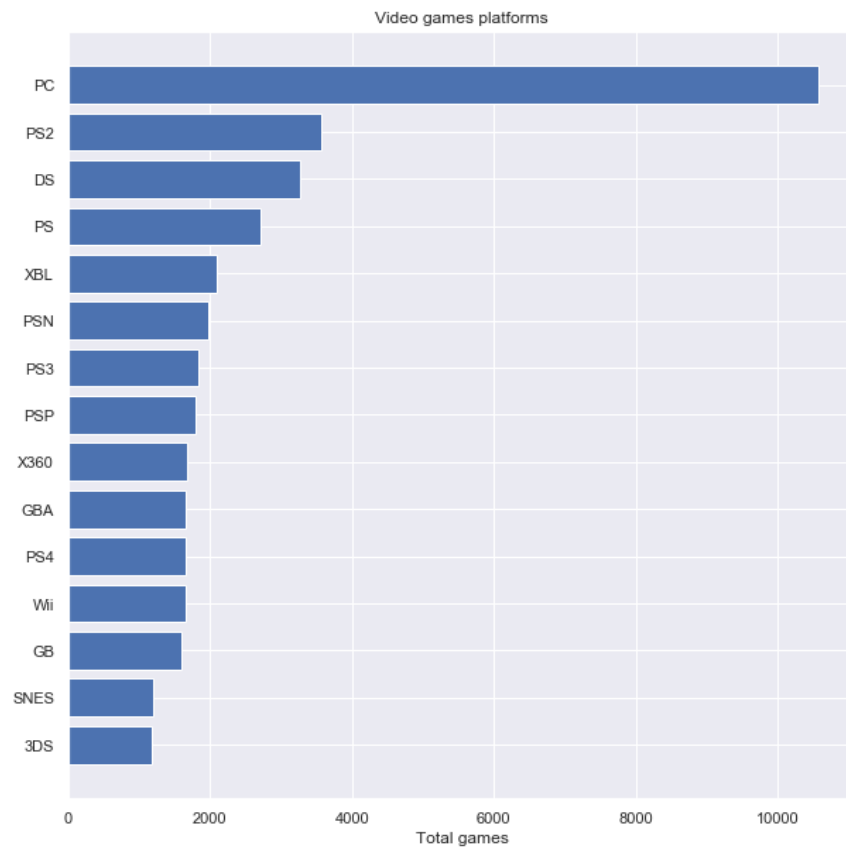
Шаг 2 - Определение популярности платформ

Построим вспомогательный график платформ, на которых выходило наибольшее количество игр за все время.

```

1 fig, ax = plt.subplots(figsize=(10, 10))
2
3 counts = df['Platform'].value_counts().sort_values(ascending=False).head(15)
4 ax.barh(counts.index, counts)
5 ax.invert_yaxis() # labels read top-to-bottom
6 ax.set_xlabel('Total games')
7 ax.set_title('Video games platforms')
8 plt.show()

```



Очевиден большой отрыв для ПК, там нет характерных для консолей "поколений", а значит все игры за все время объединены в один столбец.

Шаг 3 - Определение количества игр на платформах

```

1 df_top_platforms = df.loc[df['Platform'].isin(counts.index)]
2 df_top_platforms.head()

```

```

1 .dataframe tbody tr th {
2     vertical-align: top;
3 }
4
5 .dataframe thead th {
6     text-align: right;
7 }

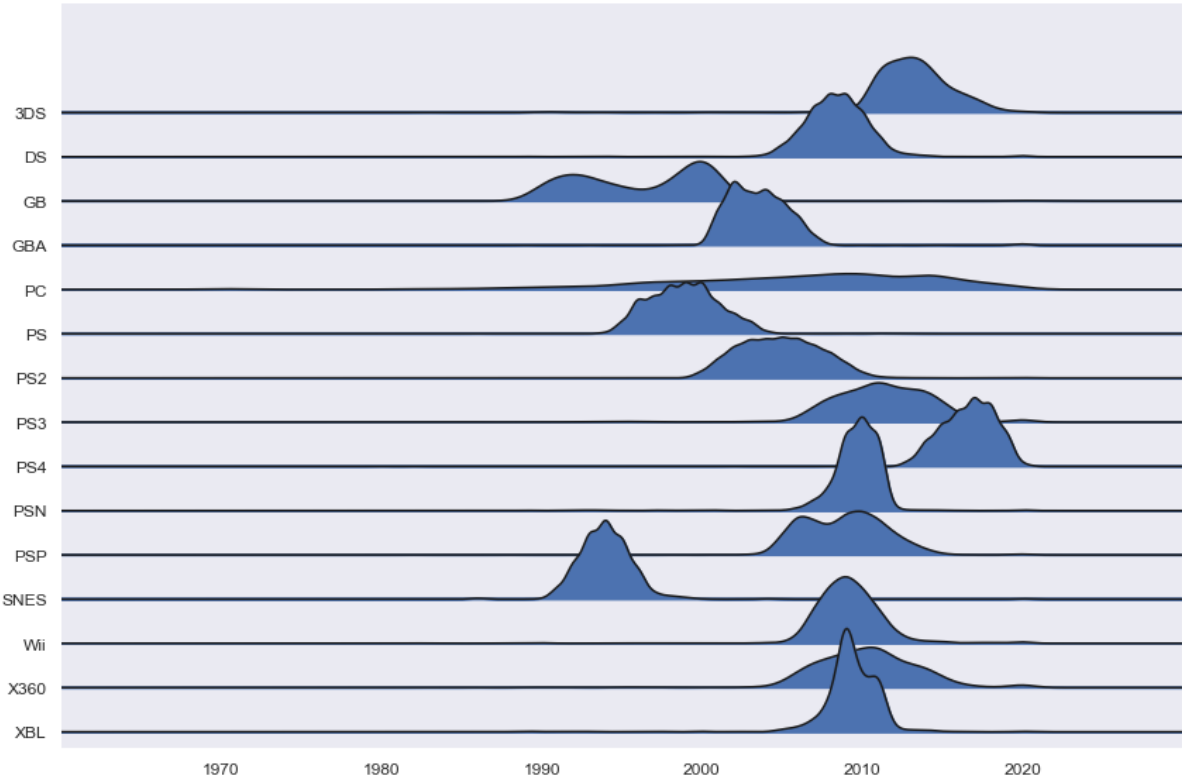
```

	Rank	Name	Genre	ESRB_Rating	Platform	Publisher	Developer	Critic_Score	User_Score	Year	Sales
0	1	Wii Sports	Sports	E	Wii	Nintendo	Nintendo EAD	7.7	NaN	2006.0	82.86
2	3	Mario Kart Wii	Racing	E	Wii	Nintendo	Nintendo EAD	8.2	9.1	2008.0	37.14
3	4	PlayerUnknown's Battlegrounds	Shooter	NaN	PC	PUBG Corporation	PUBG Corporation	NaN	NaN	2017.0	36.60
4	5	Wii Sports Resort	Sports	E	Wii	Nintendo	Nintendo EAD	8.0	8.8	2009.0	33.09
5	6	Pokemon Red / Green / Blue Version	Role-Playing	E	GB	Nintendo	Game Freak	9.4	NaN	1998.0	31.38

```
1 # Немного почи́м выбросы
2 df_top_platforms = df_top_platforms.loc[(df_top_platforms['Year'] > 1970) | (df_top_platforms['Platform'] == 'PC')]
```

Очевидно, что должны наблюдаться последовательные пики количества игр на каждой платформе, т.к. среди платформ характерна концепция "смены поколений". Для визуализации выберем Ridgeline.

```
1 joyplot(
2     data=df_top_platforms[['Year', 'Platform']],
3     by='Platform',
4     figsize=(12, 8)
5 )
6 plt.show()
```



Видно, что для PC игры выходили на протяжении всего рассматриваемого периода, а количество выпускаемых игр на консоли сменяется поколениями консолей (отчетливо видно последовательность PS, PS2, PS3, PS4).

Для консолей, имеющих несколько ревизий отчетливо видны "пики". Так, для GameBoy можно выделить два пика - в 1991 (через год после выхода первой ревизии) и в 1999 (через год после выхода GameBoy Light). Для PSP аналогичная структура - пик через некоторое время после выхода первой ревизии, а затем еще один после выхода PSP-Go в 2009.

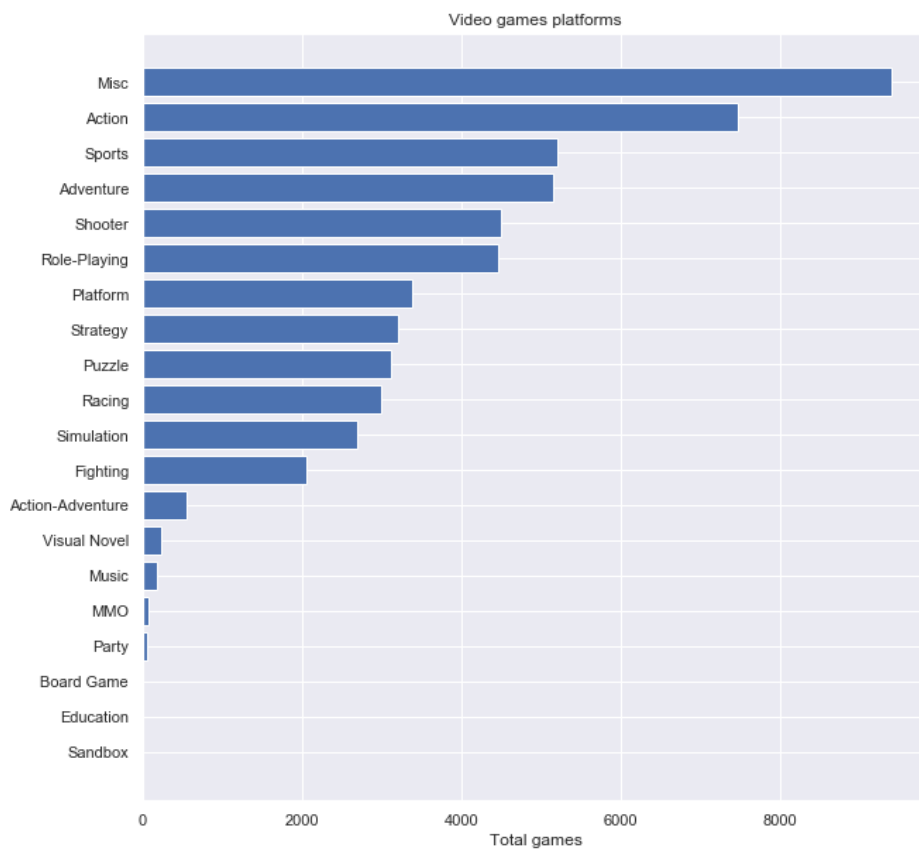
Шаг 4 - Определение популярности жанров

Построим предварительный график, аналогичный шагу 2, но для жанров, чтобы отбросить самые редкие

```

1 fig, ax = plt.subplots(figsize=(10, 10))
2
3 counts_genre = df['Genre'].value_counts().sort_values(ascending=False).head(20)
4 ax.barh(counts_genre.index, counts_genre)
5 ax.invert_yaxis() # labels read top-to-bottom
6 ax.set_xlabel('Total games')
7 ax.set_title('Video games platforms')
8 plt.show()

```



```

1 df_top_genre = df.loc[df['Genre'].isin(counts_genre.index)]
2 df_top_genre.head()

```

```

1 .dataframe tbody tr th {
2     vertical-align: top;
3 }
4
5 .dataframe thead th {
6     text-align: right;
7 }

```

	Rank	Name	Genre	ESRB_Rating	Platform	Publisher	Developer	Critic_Score	User_Score	Year	Sales
0	1	Wii Sports	Sports	E	Wii	Nintendo	Nintendo EAD	7.7	NaN	2006.0	82.86
1	2	Super Mario Bros.	Platform	NaN	NES	Nintendo	Nintendo EAD	10.0	NaN	1985.0	40.24
2	3	Mario Kart Wii	Racing	E	Wii	Nintendo	Nintendo EAD	8.2	9.1	2008.0	37.14
3	4	PlayerUnknown's Battlegrounds	Shooter	NaN	PC	PUBG Corporation	PUBG Corporation	NaN	NaN	2017.0	36.60
4	5	Wii Sports Resort	Sports	E	Wii	Nintendo	Nintendo EAD	8.0	8.8	2009.0	33.09

```

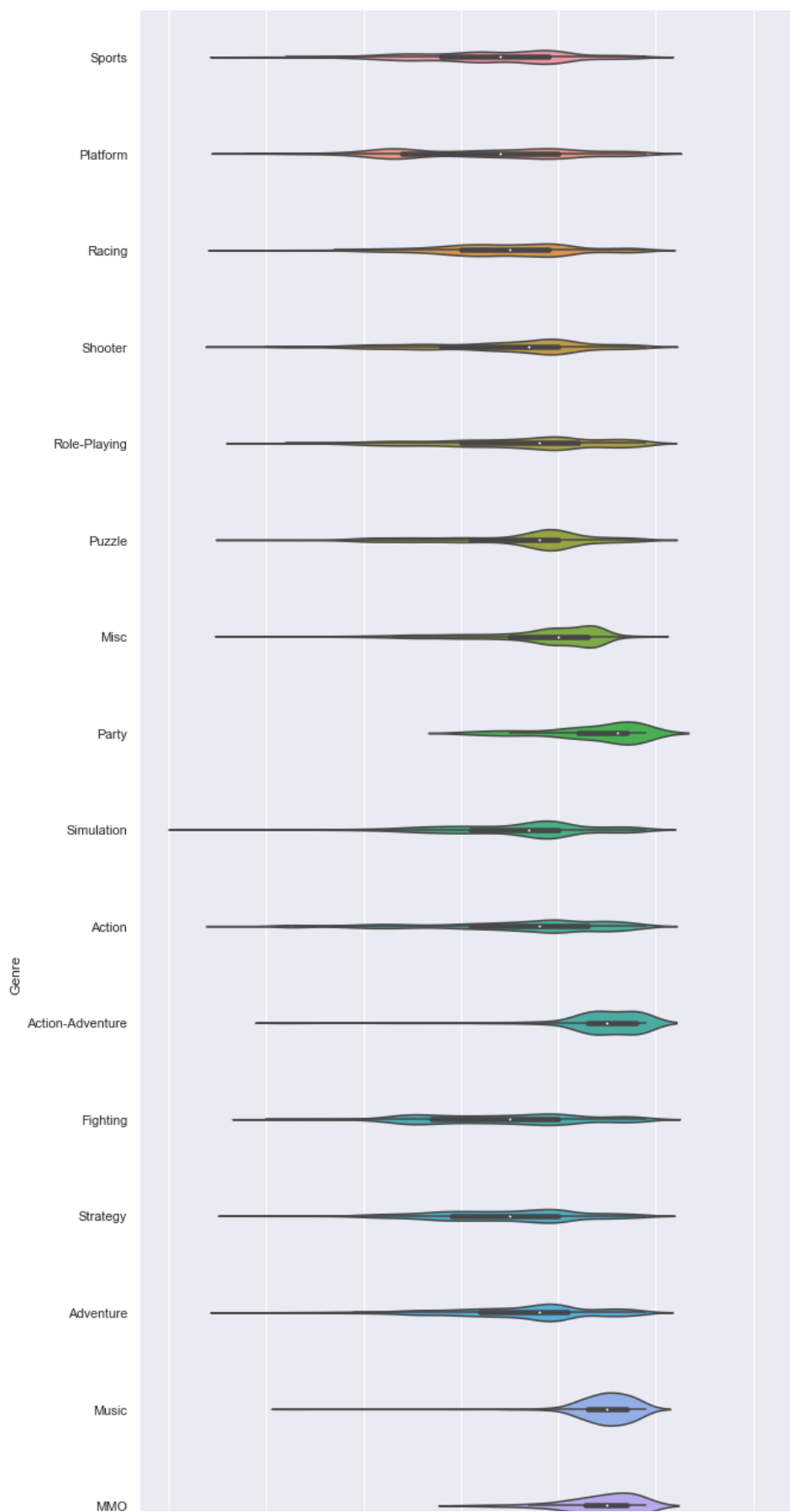
1 # Немного почистим выбросы
2 df_top_genre = df_top_genre.loc[(df_top_genre['Year'] > 1970) & (df_top_genre['Year'] <= 2019)]

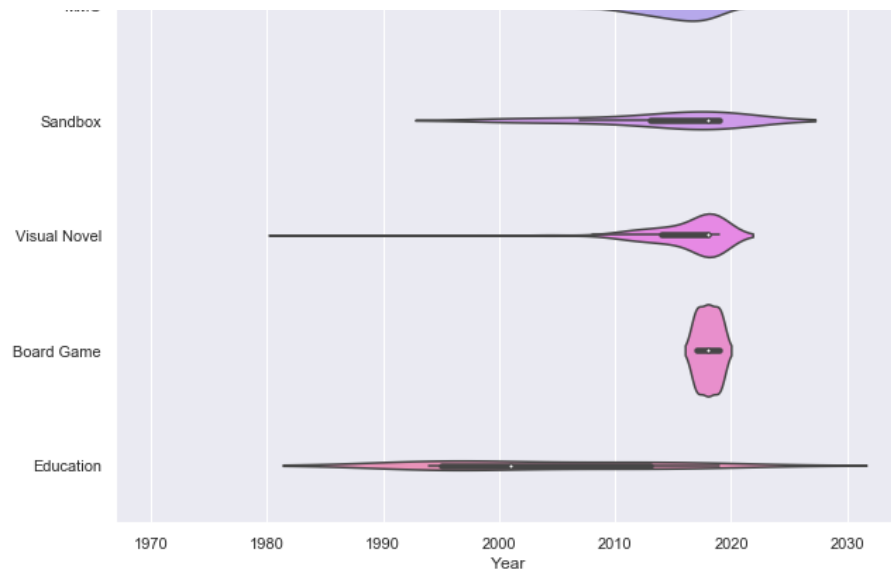
```

```

1 fig, ax = plt.subplots(figsize=(10, 30))
2 sns.violinplot(ax=ax, x=df_top_genre['Year'], y=df_top_genre['Genre'])
3 plt.show()

```





Наиболее старым популярным жанром является платформер, первый пик его популярности приходится на начало 1990-ых (например игры Super Mario Bros, Metroid, итд).

До и после 2010 года с выходом консолей 7-го поколения (PS3, Xbox360, Wii) наблюдается смена популярности жанров на появившиеся музыкальные игры, визуальные новеллы, и экшн-адвенчуры.

Занятно также, что жанр настольных видеоигр появился совсем недавно и пока не получил развития.

Шаг 5 - Определение частоты наиболее популярных жанров на платформах

```
1 counts_genre.index
```

```
1 Index(['Misc', 'Action', 'Sports', 'Adventure', 'Shooter', 'Role-Playing',
2       'Platform', 'Strategy', 'Puzzle', 'Racing', 'Simulation', 'Fighting',
3       'Action-Adventure', 'Visual Novel', 'Music', 'MMO', 'Party',
4       'Board Game', 'Education', 'Sandbox'],
5       dtype='object')
```

```
1 df.groupby('Platform')['Genre'].value_counts()
```

```
1 Platform  Genre
2 2600      Action      299
3           Shooter      68
4           Sports      39
5           Misc       21
6           Puzzle      19
7           ...
8 iQue      Shooter       2
9           Adventure     1
10          Fighting     1
11          Platform     1
12          Simulation     1
13 Name: Genre, Length: 758, dtype: int64
```

```
1 platform_genre_df = pd.DataFrame()
2
3 for genre in counts_genre.index:
4     selection = (df['Genre'] == genre) & (df['Platform'].isin(counts.index))
5     platform_genre_df[genre] = df.loc[selection]['Platform'].value_counts()
```

```
1 # NaN = не было игр такого типа, заменим нулями
2 platform_genre_df = platform_genre_df.fillna(0)
```

```

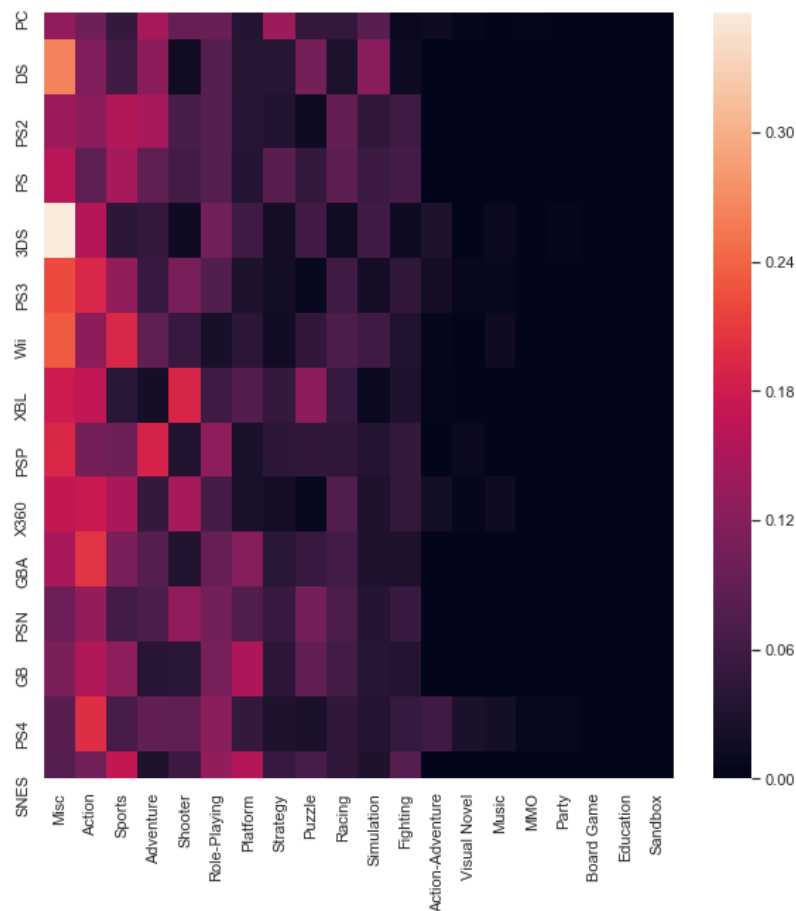
1 # Нормализуем по каждой строке, чтобы PC не вызывал перекоса в сравнении с консолями
2 platform_genre_df = platform_genre_df.div(platform_genre_df.sum(axis=1), axis=0)

```

```

1 # Построим распределение жанров по платформам
2 fig, ax = plt.subplots(figsize=(10, 10))
3 sns.heatmap(platform_genre_df, ax=ax)
4 plt.show()

```



На всех платформах популярным оказался жанр Экшн и Другое. Однако у некоторых консолей наблюдается явная специализация в определенный жанр:

- PC: Стратегии (согласуется с высокой сложностью управления и наличием клавиатуры)
- SNES / GameBoy / GameBoy Advance: Платформеры (самые технологически простые игры)
- Wii: Спортивные (виртуальный фитнес Wii Sports самая продаваемая игра в списке)
- Xbox: Шутеры (лучшая производительность железа среди конкурентов)
- Playstation Portable: Приключения (продвинутые игры для портативной консоли)
- Nintendo DS/3DS: Другое (согласуется с уникальной особенностью консолей - два экрана)

Выводы

Если учесть данные с шагов 3, 4 и 5 - можно сделать вывод, что игровая индустрия развивалась последовательно и годы выхода консолей совпадают с началом популярности определенных жанров.

Для более старых консолей NES, SNES, GameBoy - это платформеры. В 2010ых стали популярны спортивные игры из-за чего Wii стала самой продаваемой консолью и WiiSports самой продаваемой игрой. А для новых поколений консолей это приключения и шутеры. Что полностью совпадает с пиками популярности жанров и специализацией платформ.