



**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»**

**Лабораторная работа №1
«Создание “истории о данных”»
по дисциплине «Методы машинного обучения»**

Выполнил:
студент группы ИУ5-25М
Тураев Г.В.
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.
Подпись и дата:

Цель: изучение различных методов визуализации данных и создание истории на основе данных.

Импортируем нужные нам библиотеки для выполнения лабораторной работы №1:

```
[5] import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

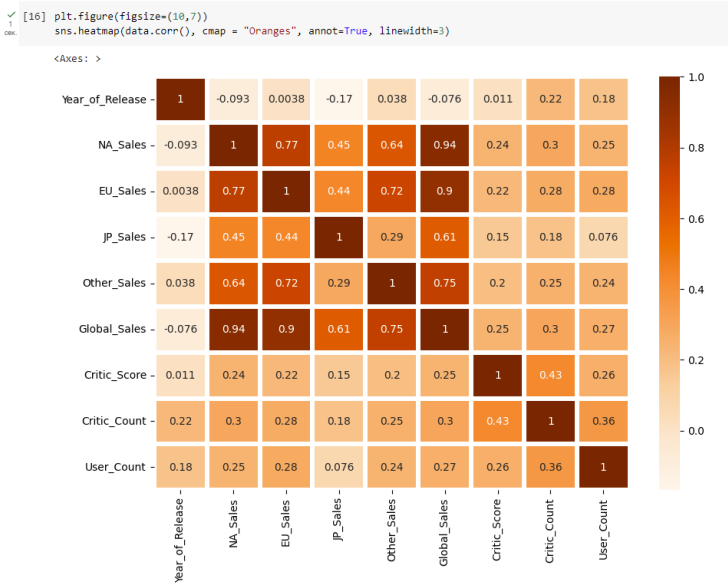
Загрузим датасет и проанализируем его, выведя первые 5 строк, запросив информацию о нем:

```
[10] data = pd.read_csv('video_games_sales.csv')
[11] data.head()
[12] data.shape
[15] data.info()
```

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN

```
<class 'pandas.core.frame.DataFrame'>
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           18015 non-null  object
13  User_Count           7598 non-null   float64
14  Developer            10096 non-null  object
15  Rating               9950 non-null   object
dtypes: float64(9), object(7)
memory usage: 2.0+ MB
```

График №1. Корреляция.

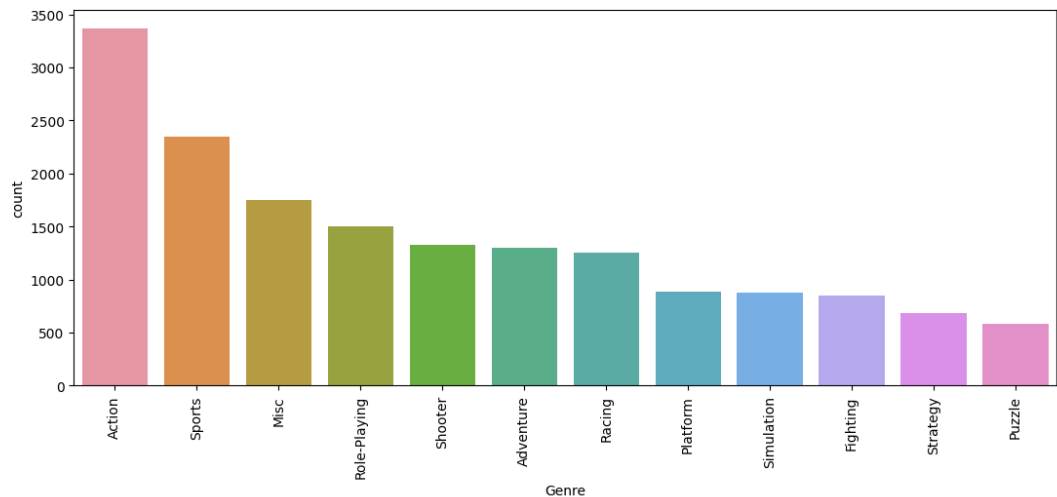


Вывод по графику №1: видно, что наиболее сильно коррелируют показатели продаж Северной Америки и Европы.

График №2. Гистограмма.

```
[45] plt.figure(figsize=(13, 5))
sns.countplot(x="Genre", data=data, order = data['Genre'].value_counts().index)
plt.xticks(rotation=90)
```

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'Action'),
  Text(1, 0, 'Sports'),
  Text(2, 0, 'Misc'),
  Text(3, 0, 'Role-Playing'),
  Text(4, 0, 'Shooter'),
  Text(5, 0, 'Adventure'),
  Text(6, 0, 'Racing'),
  Text(7, 0, 'Platform'),
  Text(8, 0, 'Simulation'),
  Text(9, 0, 'Fighting'),
  Text(10, 0, 'Strategy'),
  Text(11, 0, 'Puzzle')])
```



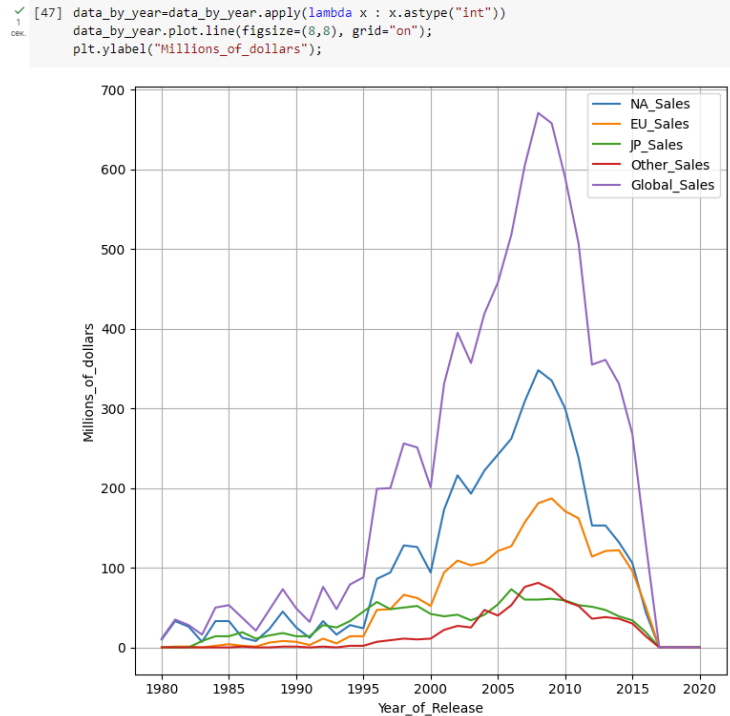
Вывод по графику №2: больше всего игр в жанре «Action», меньше всего – «Puzzle».

Сгруппируем по годам сумму показателей по каждому виду продаж:

```
[22] data_by_year = data.groupby(by = 'Year_of_Release').sum()
data_by_year.drop(columns=["Critic_Count", "User_Count", "Critic_Score"],inplace=True)
data_by_year
```

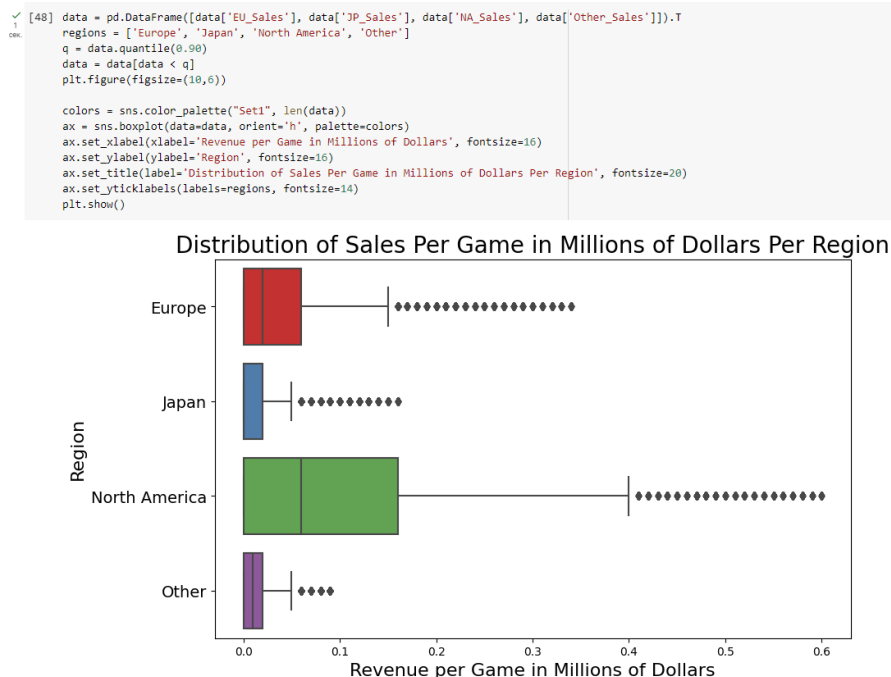
	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Year_of_Release					
1980.0	10.59	0.67	0.00	0.12	11.38
1981.0	33.40	1.96	0.00	0.32	35.77
1982.0	26.92	1.65	0.00	0.31	28.86
1983.0	7.76	0.80	8.10	0.14	16.79
1984.0	33.28	2.10	14.27	0.70	50.36
1985.0	33.73	4.74	14.56	0.92	53.94
1986.0	12.50	2.84	19.81	1.93	37.07
.....					
2012.0	153.26	114.59	51.80	36.19	355.84
2013.0	153.65	121.55	47.69	38.35	361.24
2014.0	132.27	122.74	39.69	36.83	331.51
2015.0	106.86	96.72	34.09	30.31	268.05
2016.0	44.93	51.22	19.31	14.48	130.10
2017.0	0.00	0.00	0.06	0.00	0.06
2020.0	0.27	0.00	0.00	0.02	0.29

График №3. Линейный.



Построим линейный график, где представим игры по продажам в разные года по разным регионам. Проанализировав, видно, что на период 2005-2010 годов пришел бум продаж игр по всему миру. Больше всего игр было продано в NA (Северная Америка), меньше всего в JP (Японии).

График №4. «Ящик с усами».



Вывод: заметим, что Северная Америка лидирует по продажам игр как в размахе, так и по медианному значению.

```

[49] top_sale_reg = data[['NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales']]
top_sale_reg = top_sale_reg.sum().reset_index()
top_sale_reg = top_sale_reg.rename(columns={"index": "Region", 0: "Sales"})
top_sale_reg

```

	Region	Sales
0	NA_Sales	1674.46
1	EU_Sales	744.53
2	JP_Sales	242.07
3	Other_Sales	227.81

```

[50] labels = top_sale_reg['Region']
      sizes = top_sale_reg['Sales']

```

График №5. Кольцевая диаграмма.

```

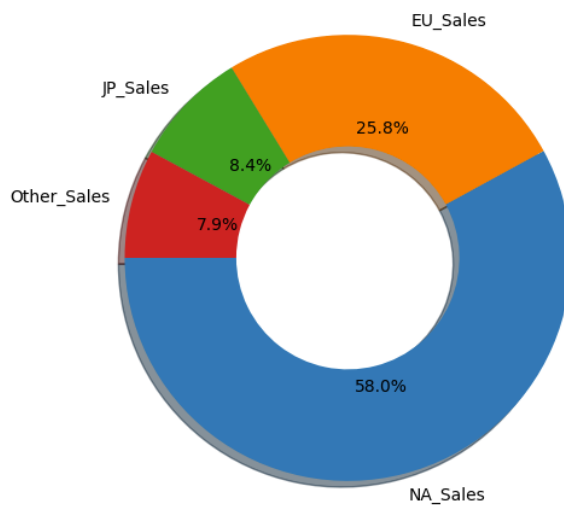
[51] plt.figure(figsize=(8, 6))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', wedgeprops=dict(width=0.5), shadow=True, startangle=180)

```

```

([<matplotlib.patches.Wedge at 0x7fd01c96b160>,
 <matplotlib.patches.Wedge at 0x7fd01c96b040>,
 <matplotlib.patches.Wedge at 0x7fd01c969310>,
 <matplotlib.patches.Wedge at 0x7fd01c969be0>],
 [Text(0.2723019312452782, -1.0657634156979174, 'NA_Sales'),
 Text(0.2836793891660941, 1.062791609000726, 'EU_Sales'),
 Text(-0.7982850337767683, 0.7567965412500403, 'JP_Sales'),
 Text(-1.0664161445551974, 0.2697343260173396, 'Other_Sales')],
 [Text(0.14852832613378808, -0.5813254994715913, '58.0%'),
 Text(0.15473421227241493, 0.5797045140003959, '25.8%'),
 Text(-0.4354282002418736, 0.4127981134091128, '8.4%'),
 Text(-0.581681533393744, 0.14712781419127613, '7.9%')])

```



Вывод: видно, что NA (Северная Америка) имеет наибольшую долю продаж во всем мире.

Общий вывод по лабораторной работе №1:

- «Action» – самый популярный жанр игр во всем мире;
- Северная Америка имеет самую большую долю продаж в мире;
- В периоде 2005-2010 годов произошел бум продаж игр по всему миру, где Япония оказалась «аутсайдером» по показателю.