Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и системы управления»

Кафедра ИУ5. Курс «Методы машинного обучения»

Отчет по лабораторной работе №1: «Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5-21М
Курганова Александра

Подпись и дата:

Проверил:
Гапанюк Ю.Е.

Подпись и дата:

Москва, 2021 г.

Задание

1. Выбрать набор данных (датасет). Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.
2. Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:
3. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
4. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
5. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
6. 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.
7. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.
8. Сформировать отчет и разместить его в своем репозитории на github.

Выполнение

```
from google.colab import drive
drive.mount('/content/drive', force_remount=True)
path = "/content/drive/My Drive/магистратура/2 семестр/ммо"

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

```
Mounted at /content/drive
```

[280]
```
data = pd.read_csv(path+'/Video_Games_Sales.csv')
```

[281]
```
data.head()
```

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	U
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	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	
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	

```
data.info()
```

```
<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       10015 non-null  object
 13  User_Count       7590 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
```

```
data.isnull().all()
```

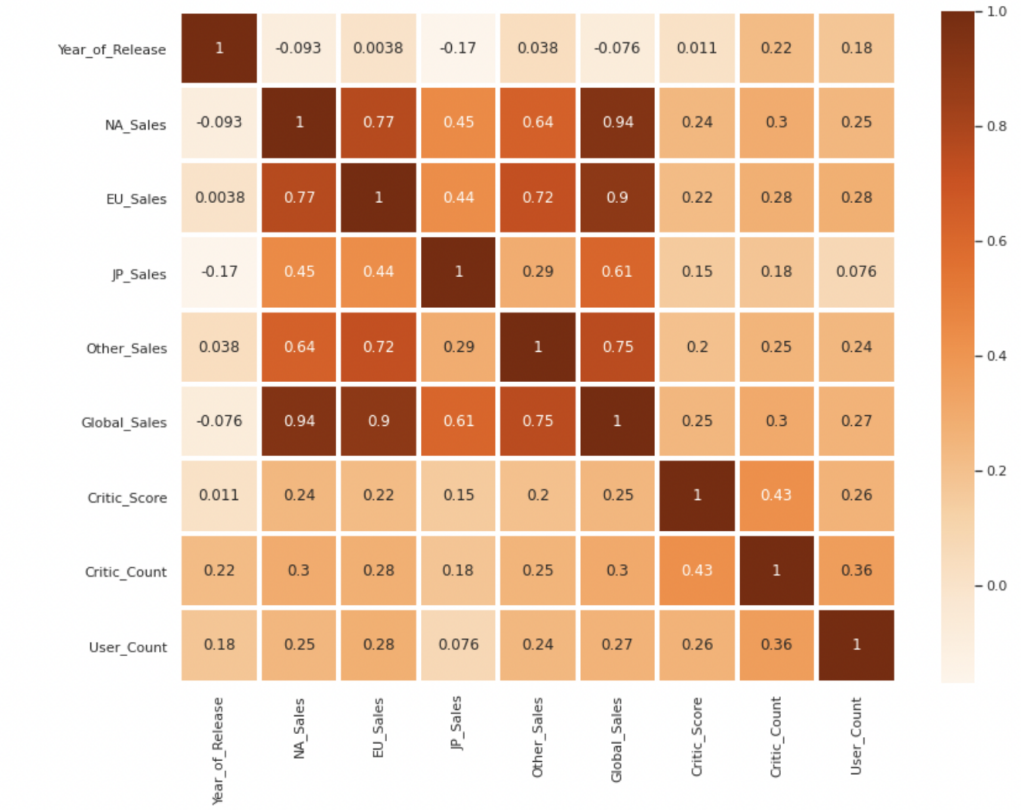
```
Name               False
Platform           False
Year_of_Release    False
Genre              False
Publisher          False
NA_Sales           False
EU_Sales           False
JP_Sales           False
Other_Sales        False
Global_Sales       False
Critic_Score       False
Critic_Count       False
User_Score         False
User_Count         False
Developer          False
Rating             False
dtype: bool
```

[284]
```
data["EU_Sales"].value_counts()
```

```
0.00    5874
0.01    1494
0.02    1308
0.03     926
0.04     709
        ...
3.59       1
4.02       1
2.24       1
2.27       1
3.75       1
Name: EU_Sales, Length: 307, dtype: int64
```

```
plt.figure(figsize=(13,10))
sns.heatmap(data.corr(), cmap="Oranges", annot=True, linewidths=3)
```

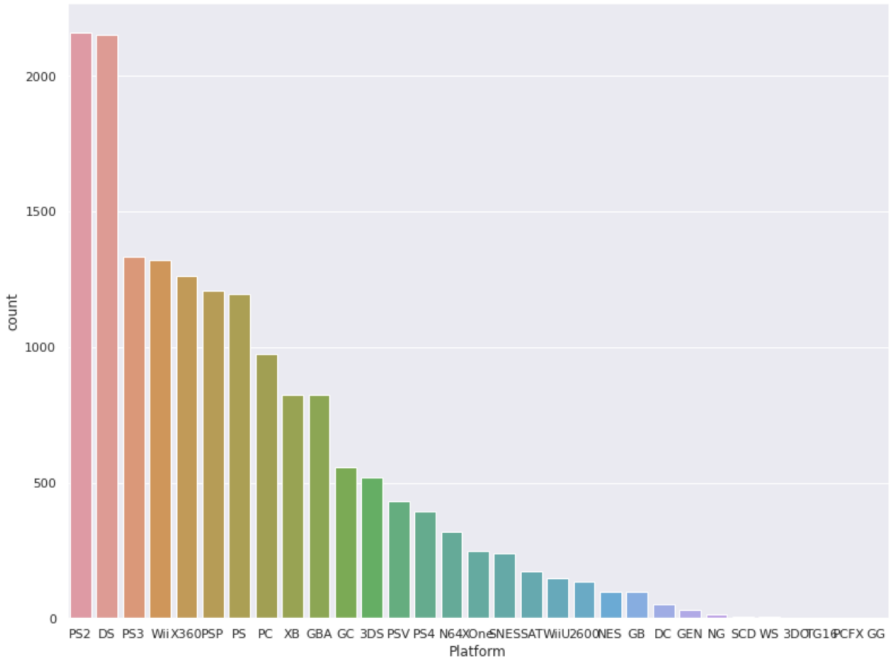
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f62f0ae6350>
```



Из матрицы корреляции видно, что наиболее сильно коррелируют показатели продаж Северной Америки и Европы

```
plt.figure(figsize=(13,10))
sns.countplot(x="Platform", data=data, order=data["Platform"].value_counts().index)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f62f0c68b50>
```



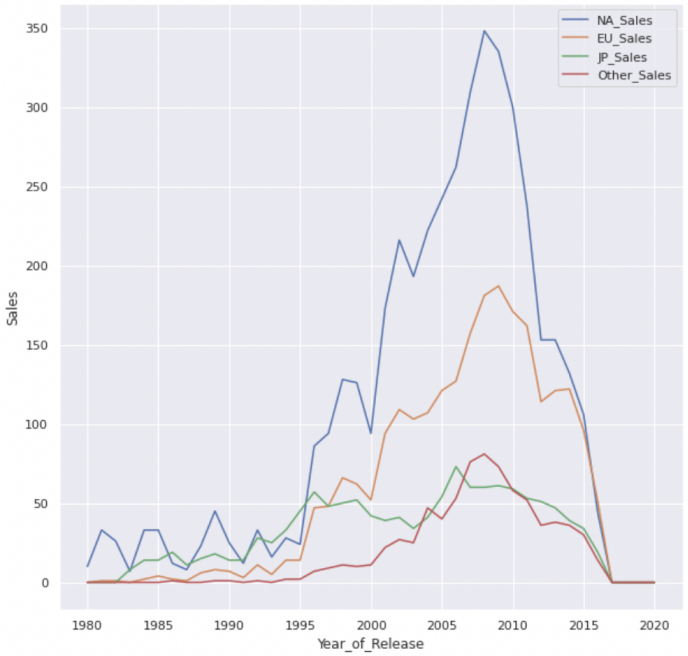
Из графика видно, что количество игр на платфоме "PS2" наибольшее, дальше идёт жанр "DS", а меньше всего игр на платформе "GG".

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

Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales
1980.0	10.59	0.67	0.00	0.12
1981.0	33.40	1.96	0.00	0.32
1982.0	26.92	1.65	0.00	0.31
1983.0	7.76	0.80	8.10	0.14
1984.0	33.28	2.10	14.27	0.70
1985.0	33.73	4.74	14.56	0.92
1986.0	12.50	2.84	19.81	1.93
1987.0	8.46	1.41	11.63	0.20
1988.0	23.87	6.59	15.76	0.99
1989.0	45.15	8.44	18.36	1.50
1990.0	25.46	7.63	14.88	1.40
1991.0	12.76	3.95	14.78	0.74
1992.0	33.89	11.71	28.91	1.65
1993.0	16.90	5.18	25.36	0.97
1994.0	28.16	14.88	33.99	2.20
1995.0	24.83	14.90	45.75	2.64
1996.0	86.76	47.26	57.44	7.69
1997.0	94.75	48.32	48.87	9.13
1998.0	128.36	66.90	50.04	11.01
1999.0	126.06	62.67	52.34	10.04
2000.0	94.50	52.77	42.77	11.62
2001.0	173.98	94.89	39.86	22.73

```
sales_data_year = sales_data_year.apply(lambda x: x.astype("int"))
sales_data_year.plot.line(figsize=(10,10), grid="on");
plt.ylabel("Sales")
```

```
Text(0, 0.5, 'Sales')
```



После разбиения продажи игр по года получается, что в 2011 году больше всех заработала Северная Америка. Затем Европа. Далее идут другие страны. Япония получила наименьший доход.

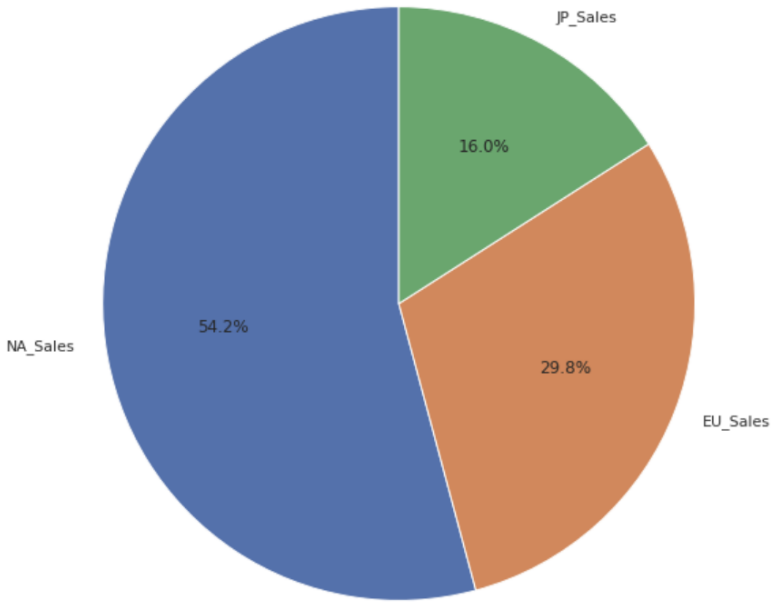
```
sales_region = data[["NA_Sales", "EU_Sales", "JP_Sales"]]
sales_region = sales_region.sum().reset_index()
sales_region = sales_region.rename(columns={"index": "region", 0: "sale"})
sales_region
```

	region	sale
0	NA_Sales	4402.62
1	EU_Sales	2424.67
2	JP_Sales	1297.43

```
[291] values = sales_region["region"]
      sizes = sales_region["sale"]
```

```
plt.figure(figsize=(10,10))
plt.pie(sizes, labels=values, autopct="%1.1f%%", startangle=90)
```

```
([<matplotlib.patches.Wedge at 0x7f62f1e64c90>,
  <matplotlib.patches.Wedge at 0x7f62f1e5f410>,
  <matplotlib.patches.Wedge at 0x7f62f1e5fc90>],
 [Text(-1.0904930590581252, -0.14430830934514013, 'NA_Sales'),
  Text(1.0255163791147273, -0.397889628122447, 'EU_Sales'),
  Text(0.5289876929807663, 0.964454260540585, 'JP_Sales')],
 [Text(-0.5948143958498864, -0.07871362327916734, '54.2%'),
  Text(0.5593725704262148, -0.21703070624860743, '29.8%'),
  Text(0.2885387416258725, 0.5260659602948645, '16.0%')])
```



Северная Америка имеет большую долю в продаже игр. На втором месте Европа, а на последнем Япония.

Вывод

В ходе выполнения лабораторной работы был проведён анализ данных о продажах видео-игр. После проведения анализа было получено, что наибольшую прибыль от продажи видеоигр получают в Северной Америке и Европе. Однако в 2011 году Северная Америка заняла лидерские позиции, в отличие от Европы и Японии. Самой популярной площадкой для игр является "PS2".