Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Создание "истории о данных"»

Выполнил:
студент группы ИУ5И-22М
Лю Чжинань

Москва — 2023 г.

1. Цель лабораторной работы

Изучение различных методов визуализация данных и создание истории на основе данных.

2. Задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

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

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. 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.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

```
import pandas as pd
import random
import matplotlib.pyplot as plt
import math as math
%matplotlib inline
import re
```

```
url="/content/Movies.csv"
dataset = pd.read_csv(url)
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<ipython-input-2-34e00ff2610b> in <module>
      1 url="/content/Movies.csv"
----> 2 dataset = pd.read_csv(url)

7 frames
/usr/local/lib/python3.8/dist-packages/pandas/io/common.py in get_handle(path_or_buf, mode, encoding, compression, memory_map, is_text, errors, storage_options)
    700         if ioargs.encoding and "b" not in ioargs.mode:
    701             # Encoding
--> 702             handle = open(
    703                 handle,
    704                 ioargs.mode,

FileNotFoundError: [Errno 2] No such file or directory: '/content/Movies.csv'
```

SEARCH STACK OVERFLOW

```
dataset
```

	index	Title	Release Date	Year	Description	URL	Rating	Runtime	Genres	Votes	Directors	Series	Order
0	0	101 Dalmatians	18-11-1996	1996.0	NaN	https://www.imdb.com/title/tt0115433/	5.7	103.0	Adventure, Comedy, Crime, Family	98439.0	Stephen Herek	101 Dalmatians	1
1	1	102 Dalmatians	22-11-2000	2000.0	NaN	https://www.imdb.com/title/tt0211181/	4.9	100.0	Adventure, Comedy, Family	33823.0	Kevin Lima	101 Dalmatians	2
2	2	12 Rounds	19-03-2009	2009.0	NaN	https://www.imdb.com/title/tt1160368/	5.6	108.0	Action, Crime, Thriller	26828.0	Renny Harlin	12 Rounds	1
3	3	12 Rounds 2: Reloaded	04-06-2013	2013.0	NaN	https://www.imdb.com/title/tt2317524/	5.3	95.0	Action, Adventure, Thriller	5141.0	Roel Reiné	12 Rounds	2
4	4	21 Jump Street	12-03-2012	2012.0	NaN	https://www.imdb.com/title/tt1232829/	7.2	109.0	Action, Comedy, Crime	498876.0	Christopher Miller, Phil Lord	21 Jump Street	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
861	861	[Rec]²	02-09-2009	2009.0	NaN	https://www.imdb.com/title/tt1245112/	6.5	85.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	67100.0	Jaume Balagueró, Paco Plaza	[Rec]	2
862	862	[Rec]³: Génesis	09-03-2012	2012.0	NaN	https://www.imdb.com/title/tt1649444/	5.0	80.0	Action, Comedy, Horror, Romance, Sci-Fi, Thriller	32388.0	Paco Plaza	[Rec]	3
863	863	[REC] 4: Apocalipsis	09-09-2014	2014.0	NaN	https://www.imdb.com/title/tt1649443/	5.3	95.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	15599.0	Jaume Balagueró	[Rec]	4
864	864	xXx	09-08-2002	2002.0	NaN	https://www.imdb.com/title/tt0295701/	5.9	124.0	Action, Adventure, Thriller	170874.0	Rob Cohen	xXx	1
865	865	xXx: State of the Union	27-04-2005	2005.0	NaN	https://www.imdb.com/title/tt0329774/	4.4	101.0	Action, Adventure, Crime, Sci-Fi, Thriller	66166.0	Lee Tamahori	xXx	2

866 rows × 13 columns

```
ds1 = dataset
ds1 = ds1[ds1["Year"]>2000]
ds1 = ds1[ds1["Order"]<10]
ds1
```

	index	Title	Release Date	Year	Description	URL	Rating	Runtime	Genres	Votes	Directors	Series	Order
2	2	12 Rounds	19-03-2009	2009.0	NaN	https://www.imdb.com/title/tt1160368/	5.6	108.0	Action, Crime, Thriller	26828.0	Renny Harlin	12 Rounds	1
3	3	12 Rounds 2: Reloaded	04-06-2013	2013.0	NaN	https://www.imdb.com/title/tt2317524/	5.3	95.0	Action, Adventure, Thriller	5141.0	Roel Reiné	12 Rounds	2
4	4	21 Jump Street	12-03-2012	2012.0	NaN	https://www.imdb.com/title/tt1232829/	7.2	109.0	Action, Comedy, Crime	498876.0	Christopher Miller, Phil Lord	21 Jump Street	1
5	5	22 Jump Street	04-06-2014	2014.0	NaN	https://www.imdb.com/title/tt2294449/	7.0	112.0	Action, Comedy, Crime	336672.0	Christopher Miller, Phil Lord	21 Jump Street	2
6	6	28 Days Later...	01-11-2002	2002.0	NaN	https://www.imdb.com/title/tt0289043/	7.6	113.0	Drama, Horror, Sci-Fi, Thriller	366735.0	Danny Boyle	28 Days Later...	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
861	861	[Rec]²	02-09-2009	2009.0	NaN	https://www.imdb.com/title/tt1245112/	6.5	85.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	67100.0	Jaume Balagueró, Paco Plaza	[Rec]	2
862	862	[Rec]³: Génesis	09-03-2012	2012.0	NaN	https://www.imdb.com/title/tt1649444/	5.0	80.0	Action, Comedy, Horror, Romance, Sci-Fi, Thriller	32388.0	Paco Plaza	[Rec]	3
863	863	[REC] 4: Apocalipsis	09-09-2014	2014.0	NaN	https://www.imdb.com/title/tt1649443/	5.3	95.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	15599.0	Jaume Balagueró	[Rec]	4
864	864	xXx	09-08-2002	2002.0	NaN	https://www.imdb.com/title/tt0295701/	5.9	124.0	Action, Adventure, Thriller	170874.0	Rob Cohen	xXx	1
865	865	xXx: State of the Union	27-04-2005	2005.0	NaN	https://www.imdb.com/title/tt0329774/	4.4	101.0	Action, Adventure, Crime, Sci-Fi, Thriller	66166.0	Lee Tamahori	xXx	2

532 rows × 13 columns

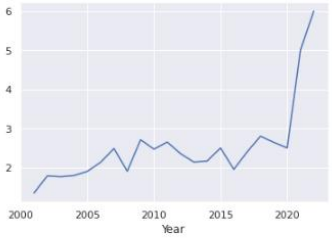
```
#uniq_Year = ds1.drop_duplicates(subset=["Year"])
#ds1["Year"]
ds1["Year"]=ds1["Year"].astype('int')
#uniq_Year["Year"]
ds1
```

	index	Title	Release Date	Year	Description	URL	Rating	Runtime	Genres	Votes	Directors	Series	Order
2	2	12 Rounds	19-03-2009	2009	NaN	https://www.imdb.com/title/tt1160368/	5.6	108.0	Action, Crime, Thriller	26828.0	Renny Harlin	12 Rounds	1
3	3	12 Rounds 2: Reloaded	04-06-2013	2013	NaN	https://www.imdb.com/title/tt2317524/	5.3	95.0	Action, Adventure, Thriller	5141.0	Roel Reiné	12 Rounds	2
4	4	21 Jump Street	12-03-2012	2012	NaN	https://www.imdb.com/title/tt1232829/	7.2	109.0	Action, Comedy, Crime	498876.0	Christopher Miller, Phil Lord	21 Jump Street	1
5	5	22 Jump Street	04-06-2014	2014	NaN	https://www.imdb.com/title/tt2294449/	7.0	112.0	Action, Comedy, Crime	336672.0	Christopher Miller, Phil Lord	21 Jump Street	2
6	6	28 Days Later...	01-11-2002	2002	NaN	https://www.imdb.com/title/tt0289043/	7.6	113.0	Drama, Horror, Sci-Fi, Thriller	366735.0	Danny Boyle	28 Days Later...	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
861	861	[Rec]²	02-09-2009	2009	NaN	https://www.imdb.com/title/tt1245112/	6.5	85.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	67100.0	Jaume Balagueró, Paco Plaza	[Rec]	2
862	862	[Rec]³: Génesis	09-03-2012	2012	NaN	https://www.imdb.com/title/tt1649444/	5.0	80.0	Action, Comedy, Horror, Romance, Sci-Fi, Thriller	32388.0	Paco Plaza	[Rec]	3
863	863	[REC] 4: Apocalipsis	09-09-2014	2014	NaN	https://www.imdb.com/title/tt1649443/	5.3	95.0	Action, Adventure, Fantasy, Horror, Sci-Fi, Th...	15599.0	Jaume Balagueró	[Rec]	4
864	864	xXx	09-08-2002	2002	NaN	https://www.imdb.com/title/tt0295701/	5.9	124.0	Action, Adventure, Thriller	170874.0	Rob Cohen	xXx	1
865	865	xXx: State of the Union	27-04-2005	2005	NaN	https://www.imdb.com/title/tt0329774/	4.4	101.0	Action, Adventure, Crime, Sci-Fi, Thriller	66166.0	Lee Tamahori	xXx	2

532 rows × 13 columns

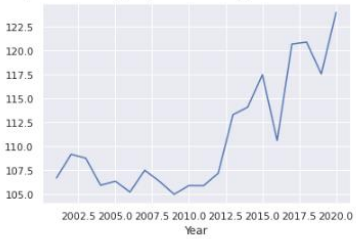
```
ds1.groupby("Year").Order.mean().plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9ae5dde100>
```

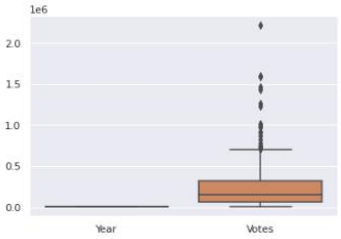



```
[ ] ds1.groupby("Year").Runtime.mean().plot()
```

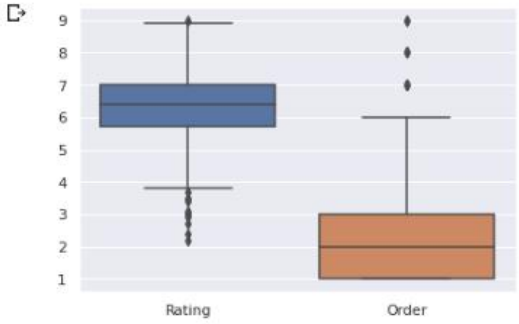
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9ae4415640>
```




```
[ ] # libraries & dataset
    import seaborn as sns
    import matplotlib.pyplot as plt
    # set a grey background (use sns.set_theme() if seaborn version 0.11.0 or above)
    sns.set(style="darkgrid")
    sns.boxplot(data=ds1.loc[:, ['Year', 'Votes']])
    plt.show()
```

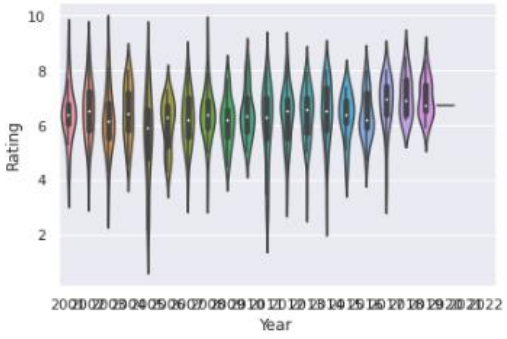


```
sns.boxplot(data=ds1.loc[:, ['Rating', 'Order']])
plt.show()
```



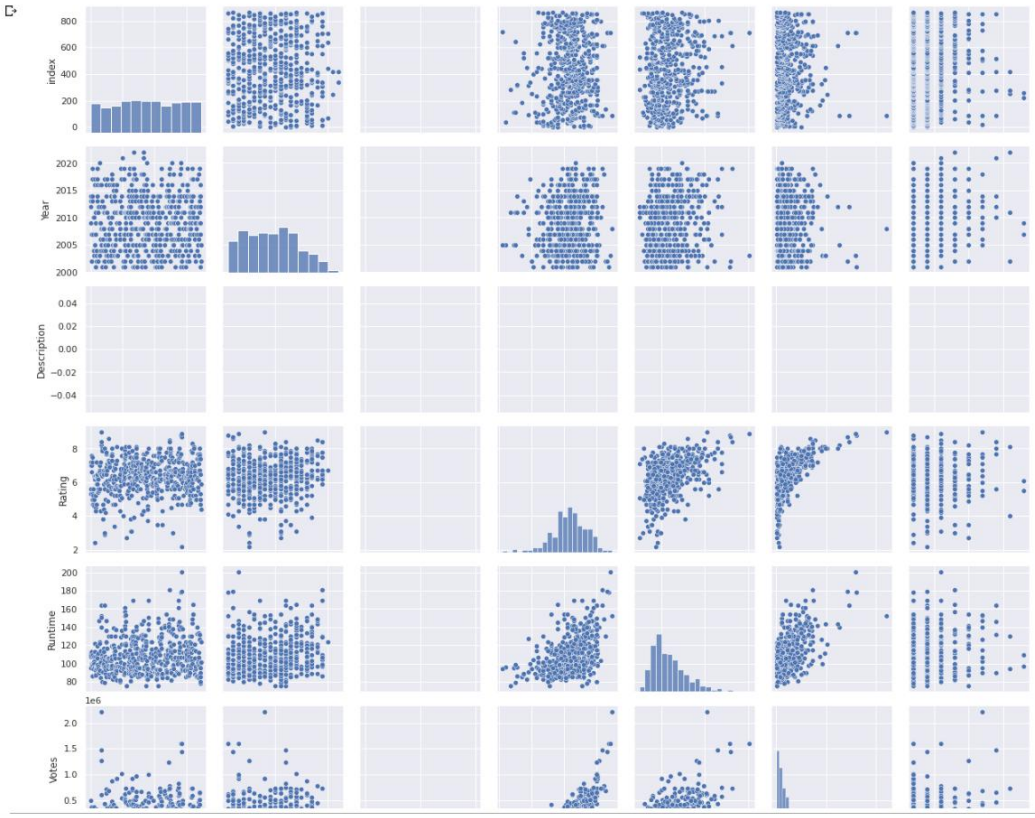

```
[ ] # libraries & dataset
    import seaborn as sns
    import matplotlib.pyplot as plt
    # set a grey background (use sns.set_theme() if seaborn version 0.11.0 or above)
    sns.set(style="darkgrid")

    # plot
    sns.violinplot( x=ds1["Year"], y=ds1["Rating"] )
    plt.show()
```



```
# libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Basic correlogram
sns.pairplot(ds1)
plt.show()
```



+ 代码 + 文本

```
[ ] ds2=pd.DataFrame(ds1,  columns=['Runtime','Order','Rating'])
    ds2
```

	Runtime	Order	Rating
2	108.0	1	5.6
3	95.0	2	5.3
4	109.0	1	7.2
5	112.0	2	7.0
6	113.0	1	7.6
...	...	...	...
861	85.0	2	6.5
862	80.0	3	5.0
863	95.0	4	5.3
864	124.0	1	5.9
865	101.0	2	4.4

532 rows × 3 columns

```
ds3  =  ds2[ds2["Runtime"]<90]
ds3
```

	Runtime	Order	Rating
15	86.0	1	5.0
16	86.0	2	4.7
39	88.0	2	4.5
40	87.0	3	4.3
76	84.0	1	7.2
...	...	...	...
846	84.0	1	6.1
856	88.0	1	7.6
860	78.0	1	7.4
861	85.0	2	6.5
862	80.0	3	5.0

67 rows × 3 columns

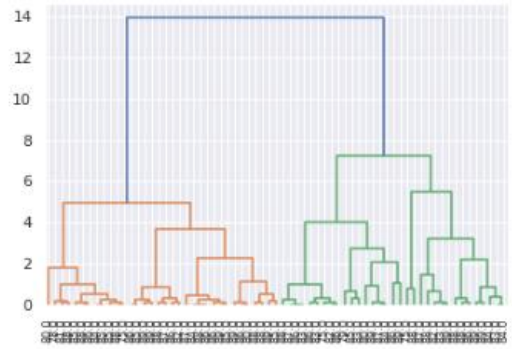
```
import pandas as pd
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
import numpy as np

# Data set
df = ds3.set_index('Runtime')

# Calculate the distance between each sample
Z = linkage(df,'ward')

# Plot with Custom leaves
dendrogram(Z, leaf_rotation=90, leaf_font_size=8, labels=df.index)

# Show the graph
plt.show()
```




```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

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

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets