Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет
Рубежный контроль №1

По курсу «Технологии машинного обучения»

Вариант 10

ИСПОЛНИТЕЛЬ:

Сафин Рустам
Группа ИУ5-64

____

"__"___________2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк. Ю.Е.

____

"__"___________2020 г.

Москва 2020

1. Условие

Задача №2:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Набор данных №2:

https://www.kaggle.com/fivethirtyeight/fivethirtyeight-comic-characters-dataset (файл dc-wikia-data.csv)

Дополнительное требование:

Для произвольной колонки данных построить график
«Скрипичная диаграмма» (violin plot).

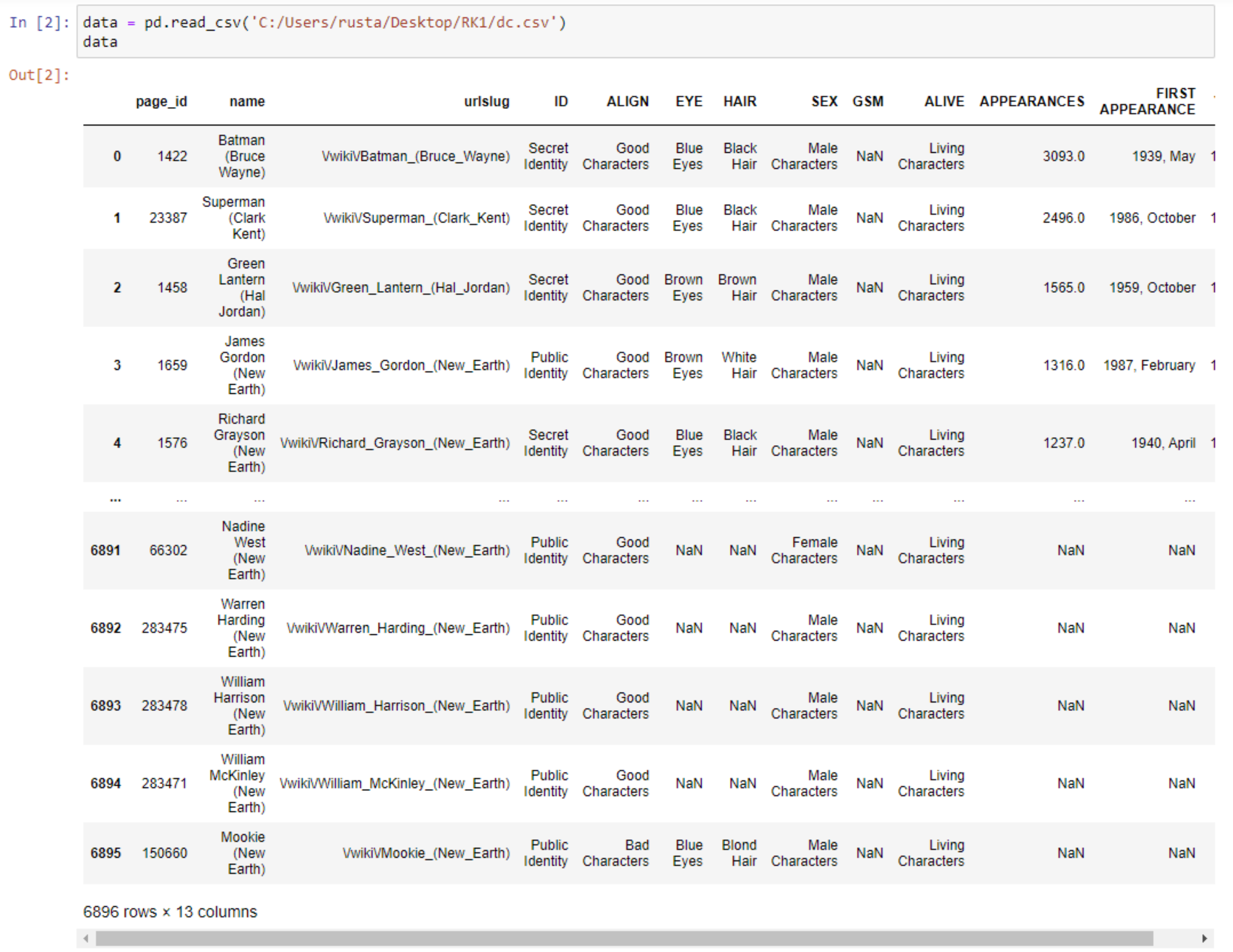
2. Выполнение

См. на следующей странице

In [1]:

```
import pandas as pd
import numpy as np
```

Извлечение dataset



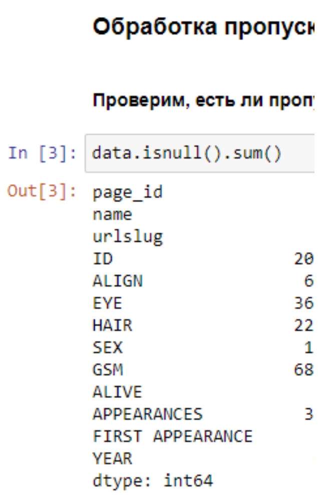
In [2]:

```
data = pd.read_csv('C:/Users/rusta/Desktop/RK1/dc.csv')
data
```

Out[2]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES	FIRST APPEARANCE	
0	1422	Batman (Bruce Wayne)	VwikiVBatman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	3093.0	1939, May	1
1	23387	Superman (Clark Kent)	VwikiVSuperman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	2496.0	1986, October	1
2	1458	Green Lantern (Hal Jordan)	VwikiVGreen_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	1565.0	1959, October	1
3	1659	James Gordon (New Earth)	VwikiVJames_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	NaN	Living Characters	1316.0	1987, February	1
4	1576	Richard Grayson (New Earth)	VwikiVRichard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	1237.0	1940, April	1
...	...	...	...	...	...	...	...	...	...	...	...	...	
6891	66302	Nadine West (New Earth)	VwikiVNadine_West_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Female Characters	NaN	Living Characters	NaN	NaN	
6892	283475	Warren Harding (New Earth)	VwikiVWarren_Harding_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN	NaN	
6893	283478	William Harrison (New Earth)	VwikiVWilliam_Harrison_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN	NaN	
6894	283471	William McKinley (New Earth)	VwikiVWilliam_McKinley_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN	NaN	
6895	150660	Mookie (New Earth)	VwikiVMookie_(New_Earth)	Public Identity	Bad Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	NaN	NaN	

6896 rows × 13 columns



Обработка пропуск

Проверим, есть ли проп

In [3]:

```
data.isnull().sum()
```

Out[3]:

```
page_id
name
urlslug
ID                  20
ALIGN                6
EYE                 36
HAIR                22
SEX                  1
GSM                 68
ALIVE
APPEARANCES          3
FIRST APPEARANCE
YEAR
dtype: int64
```

1. Замена пустых значений на среднее

Выполним замену для количественного признака APPEARANCES.

1. Количество нулевых значений:

```
In [4]: data['APPEARANCES'].isna().sum()
```

```
Out[4]: 355
```

2. Получим среднее:

```
In [5]: mean = data['APPEARANCES'].mean()
        mean
```

```
Out[5]: 23.62513377159456
```

3. Выполним замену и проверим количество пустых значений:

```
In [6]: data['APPEARANCES'].fillna(mean, inplace=True)
        data['APPEARANCES'].isna().sum()
```

```
Out[6]: 0
```

2. Удаление пустых значений

Выполним удаление для категориального признака ALIGN.

1. Количество нулевых значений:

```
In [7]: data['ALIGN'].isna().sum()
```

```
Out[7]: 601
```

```
In [8]: data = data[~data['ALIGN'].isna()]
        data
```

Out[8]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES	FIRST APPEARANCE	
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	3093.000000	1939, May	1
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	2496.000000	1986, October	1
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	1565.000000	1959, October	1
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	NaN	Living Characters	1316.000000	1987, February	1
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	1237.000000	1940, April	1
...	...	...	...	...	...	...	...	...	...	...	...	...	
6891	66302	Nadine West (New Earth)	\/wiki\/Nadine_West_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Female Characters	NaN	Living Characters	23.625134	NaN	
6892	283475	Warren Harding (New Earth)	\/wiki\/Warren_Harding_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	23.625134	NaN	
6893	283478	William Harrison (New Earth)	\/wiki\/William_Harrison_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	23.625134	NaN	
6894	283471	William McKinley (New Earth)	\/wiki\/William_McKinley_(New_Earth)	Public Identity	Good Characters	NaN	NaN	Male Characters	NaN	Living Characters	23.625134	NaN	
6895	150660	Mookie (New Earth)	\/wiki\/Mookie_(New_Earth)	Public Identity	Bad Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	23.625134	NaN	

6295 rows × 13 columns

Как можно видеть, количество строк датасета уменьшилось.

3. Проверим количество пустых значений поля ALIGN:

```
In [9]: data['ALIGN'].isna().sum()
```

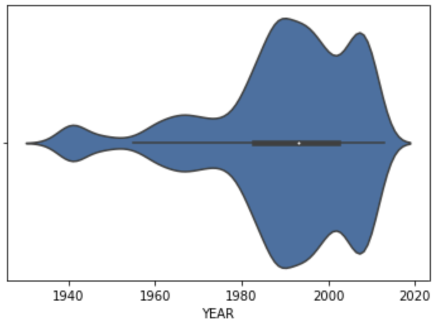
```
Out[9]: 0
```

Дополнительное задание

Построим график "Скрипичная диаграмма" (Violin plot) для поля YEAR

```
In [12]: import seaborn as sns
         sns.violinplot(x=data['YEAR'])
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x24a8c670>
```


YEAR