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)

Рубежный контроль №1

по курсу «Технологии машинного обучения»

Вариант 19

Выполнил

студент группы ИУ5-64Б

Шпак И.Д.

Москва, 2021

1 Исходное задание

Задача №3.

Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

2 Исходный код

[1]:
```
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
pd.set_option("display.max_rows", None, "display.max_columns", None)
```

[2]:
```
data = pd.read_csv("marvel-wikia-data.csv")
sc = MinMaxScaler()
tmpdata = sc.fit_transform(data[["APPEARANCES"]])
data["APPEARANCES"] = pd.DataFrame(tmpdata,columns=["APPEARANCES"])
```

[3]:
```
le = LabelEncoder()
encoded = le.fit_transform(pd.DataFrame(data["SEX"]))
data["SEX"] = encoded
data.head()
```

```
/home/igor/.local/lib/python3.9/site-packages/sklearn/utils/validation.
 ↪py:63:
DataConversionWarning: A column-vector y was passed when a 1d array was
expected. Please change the shape of y to (n_samples, ), for example␣
 ↪using
ravel().
  return f(*args, **kwargs)
```

[3]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES	FIRST APPEARANCE	Year
0	1678	Spider-Man (Peter Parker)	\/Spider-Man_(Peter_Parker)	Secret Identity	Good Characters	Hazel Eyes	Brown Hair	3	NaN	Living Characters	1.000000	Aug-62	1962.0
1	7139	Captain America (Steven Rogers)	\/Captain_America_(Steven_Rogers)	Public Identity	Good Characters	Blue Eyes	White Hair	3	NaN	Living Characters	0.831024	Mar-41	1941.0
2	64786	Wolverine (James \"Logan\" Howlett)	\/Wolverine_(James_%22Logan%22_Howlett)	Public Identity	Neutral Characters	Blue Eyes	Black Hair	3	NaN	Living Characters	0.757051	Oct-74	1974.0
3	1868	Iron Man (Anthony \"Tony\" Stark)	\/Iron_Man_(Anthony_%22Tony%22_Stark)	Public Identity	Good Characters	Blue Eyes	Black Hair	3	NaN	Living Characters	0.732311	Mar-63	1963.0
4	2460	Thor (Thor Odinson)	\/Thor_(Thor_Odinson)	No Dual Identity	Good Characters	Blue Eyes	Blond Hair	3	NaN	Living Characters	0.558387	Nov-50	1950.0

[4]:
```
data = pd.get_dummies(data, columns=["ALIGN"], prefix = ["align"])
data.head()
```

[4]:

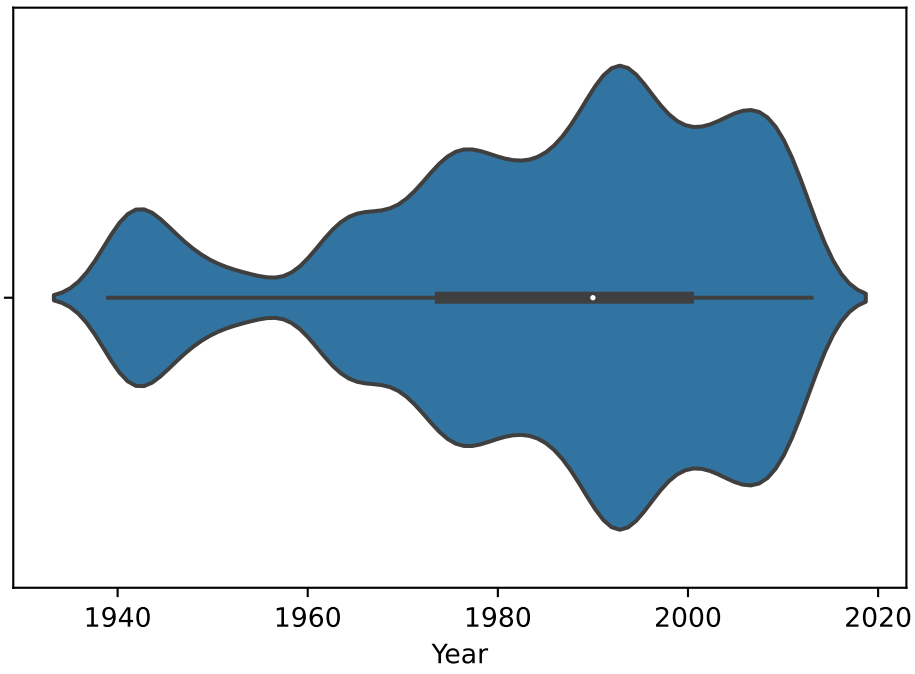
```
   page_id                                name  \
0     1678            Spider-Man (Peter Parker)
1     7139      Captain America (Steven Rogers)
2    64786  Wolverine (James \"Logan\" Howlett)
3     1868    Iron Man (Anthony \"Tony\" Stark)
4     2460                  Thor (Thor Odinson)

                                  urlslug                ID        ␣
↪EYE  \
0             \/Spider-Man_(Peter_Parker)   Secret Identity  Hazel Eyes
1       \/Captain_America_(Steven_Rogers)   Public Identity   Blue Eyes
2  \/Wolverine_(James_%22Logan%22_Howlett)   Public Identity   Blue Eyes
3    \/Iron_Man_(Anthony_%22Tony%22_Stark)   Public Identity   Blue Eyes
4                   \/Thor_(Thor_Odinson)  No Dual Identity   Blue Eyes

         HAIR  SEX  GSM              ALIVE  APPEARANCES FIRST␣
↪APPEARANCE  \
0  Brown Hair    3  NaN  Living Characters     1.000000           Aug-62
1  White Hair    3  NaN  Living Characters     0.831024           Mar-41
2  Black Hair    3  NaN  Living Characters     0.757051           Oct-74
3  Black Hair    3  NaN  Living Characters     0.732311           Mar-63
4  Blond Hair    3  NaN  Living Characters     0.558387           Nov-50

     Year  align_Bad Characters  align_Good Characters  \
0  1962.0                     0                      1
1  1941.0                     0                      1
2  1974.0                     0                      0
3  1963.0                     0                      1
4  1950.0                     0                      1
```

	align_Neutral	Characters
0		0
1		0
2		1
3		0
4		0

[9]:
```
data = pd.read_csv("marvel-wikia-data.csv")
sns.violinplot(data["Year"])
```

```
/home/igor/.local/lib/python3.9/site-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From␣
 ↪version
0.12, the only valid positional argument will be `data`, and passing␣
 ↪other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

[9]: <AxesSubplot:xlabel='Year'>



3 Описание используемых методов

Для решения задачи нормализации данных используется класс MinMaxScaler библиотеки sklearn.preprocessing, который преобразует данные в диапазон от 0 до 1.

Преобразование категориальных признаков в числовые осуществлялось при помощи класса label encoder, библиотеки sklearn, и метода get_dummies библиотеки pandas. Использование этих методов диктуется заданием, и минимальным количеством необходимых библиотек.