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

Отчёт

“Методы машинного обучения”

Лабораторная работа № 3

“Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных”

ИСПОЛНИТЕЛЬ:

Студент группы ИУ5-21М

Коростелёв В. М. _____________

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю. Е. _____________

Москва – 2019

Задание

Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи: обработку пропусков в данных; кодирование категориальных признаков; масштабирование данных.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [2]:
```
data = pd.read_csv('dc-wikia-data.csv', sep=",")
data.head()
```

Out[2]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	G
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	I
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	I
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	I
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	I
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	I

In [3]:
```
data.isnull().sum()
```

Out[3]:
```
page_id                0
name                   0
urlslug                0
ID                  2013
ALIGN                601
EYE                 3628
HAIR                2274
SEX                  125
GSM                 6832
ALIVE                  3
APPEARANCES          355
FIRST APPEARANCE      69
YEAR                  69
dtype: int64
```

```
In [4]: data.dtypes
```

```
Out[4]: page_id                int64
        name                  object
        urlslug               object
        ID                    object
        ALIGN                 object
        EYE                   object
        HAIR                  object
        SEX                   object
        GSM                   object
        ALIVE                 object
        APPEARANCES          float64
        FIRST APPEARANCE      object
        YEAR                 float64
        dtype: object
```

Обработка пропусков в данных 1.1. Простые стратегии - удаление или заполнение нулями

```
In [5]: data_new_1 = data.dropna(axis=1, how='any')
        (data.shape, data_new_1.shape)
```

```
Out[5]: ((6896, 13), (6896, 3))
```

```
In [6]: data_new_2 = data.dropna(axis=0, how='any')
        (data.shape, data_new_2.shape)
```

```
Out[6]: ((6896, 13), (38, 13))
```

```
In [7]: data_new_3 = data.fillna(0)
        data_new_3.head()
```

Out[7]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	G
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	

1.2. "Внедрение значений" - импьютация (imputation) 1.2.1. Обработка пропусков в числовых данных

In [8]:

```
# Выберем числовые колонки с пропущенными значениями
# Цикл по колонкам датасета
num_cols = []
for col in data.columns:
    # Количество пустых значений
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if temp_null_count>0 and (dt=='float64' or dt=='int64'):
        num_cols.append(col)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}.'.format(col,
dt, temp_null_count))
```

```
Колонка APPEARANCES. Тип данных float64. Количество пустых значений 355.
Колонка YEAR. Тип данных float64. Количество пустых значений 69.
```

In [9]:

```
# Фильтр по пустым значениям поля MasVnrArea
data[data['YEAR'].isnull()]
# Сохраняем индексы
flt_index = data[data['YEAR'].isnull()].index
flt_index
```

Out[9]:

```
Int64Index([ 386, 1400, 1401, 1832, 1937, 1938, 2065, 2066, 2067, 2230, 2231,
            2232, 2413, 2414, 2841, 2842, 3104, 3105, 3431, 3432, 3433, 3434,
            3435, 3819, 3820, 3821, 3822, 3823, 3824, 4320, 4321, 4322, 4323,
            4826, 4827, 4828, 4829, 5525, 5526, 5527, 5528, 5529, 5530, 5531,
            5532, 5533, 5534, 5535, 5536, 5537, 5538, 6532, 6533, 6534, 6535,
            6536, 6537, 6538, 6539, 6540, 6887, 6888, 6889, 6890, 6891, 6892,
            6893, 6894, 6895],
           dtype='int64')
```

In [10]:

```
import warnings
warnings.filterwarnings('ignore')
for rows in flt_index:
  data.YEAR[rows]=data.YEAR.median()
```

In [11]:

```
# Фильтр по пустым значениям поля MasVnrArea
data[data['YEAR'].isnull()]
# Сохраняем индексы
flt_index = data[data['YEAR'].isnull()].index
flt_index
```

Out[11]:

```
Int64Index([], dtype='int64')
```

In [12]:

```
data[data['APPEARANCES'].isnull()]
# Сохраняем индексы
flt_index = data[data['APPEARANCES'].isnull()].index
flt_index
```

Out[12]:

```
Int64Index([6541, 6542, 6543, 6544, 6545, 6546, 6547, 6548, 6549, 6550,
            ...
            6886, 6887, 6888, 6889, 6890, 6891, 6892, 6893, 6894, 6895],
           dtype='int64', length=355)
```

In [13]:

```
data.APPEARANCES = data.APPEARANCES.mean()
```

1.2.2. Обработка пропусков в категориальных данных

In [14]:

```
# Выберем категориальные колонки с пропущенными значениями
# Цикл по колонкам датасета
cat_cols = []
for col in data.columns:
    # Количество пустых значений
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if temp_null_count>0 and (dt=='object'):
        cat_cols.append(col)
        temp_perc = round((temp_null_count / data.shape[0]) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format
(col, dt, temp_null_count,temp_perc))
```

```
Колонка ID. Тип данных object. Количество пустых значений 2013, 29.19%.
Колонка ALIGN. Тип данных object. Количество пустых значений 601, 8.72%.
Колонка EYE. Тип данных object. Количество пустых значений 3628, 52.61%.
Колонка HAIR. Тип данных object. Количество пустых значений 2274, 32.98%.
Колонка SEX. Тип данных object. Количество пустых значений 125, 1.81%.
Колонка GSM. Тип данных object. Количество пустых значений 6832, 99.07%.
Колонка ALIVE. Тип данных object. Количество пустых значений 3, 0.04%.
Колонка FIRST APPEARANCE. Тип данных object. Количество пустых значений 69, 1.0%.
```

In [15]:

```
MaxPassEmbarked = data.groupby('ALIVE').count()['page_id']
data.ALIVE[data.ALIVE.isnull()] = MaxPassEmbarked[MaxPassEmbarked == MaxPassEmbarked.
max()].index[0]

data[data[col].isnull()].shape[0]
```

Out[15]:

```
0
```

Преобразование категориальных признаков в числовые

In [16]:

```
data.ALIGN.replace({'Good Characters':'1','Bad Characters':'0'},inplace=True)
data.head()
```

Out[16]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	1	Blue Eyes	Black Hair	Male Characters	NaN
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	1	Blue Eyes	Black Hair	Male Characters	NaN
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	1	Brown Eyes	Brown Hair	Male Characters	NaN
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	1	Brown Eyes	White Hair	Male Characters	NaN
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	1	Blue Eyes	Black Hair	Male Characters	NaN

In [17]:

```
from sklearn.preprocessing import LabelEncoder
label = LabelEncoder()
dicts = {}

data.ALIGN = label.fit_transform(data.ALIGN.astype(str))
label.fit(data.ALIGN.drop_duplicates()) #задаем список значений для кодирования

dicts['ALIGN'] = list(label.classes_)
data.ALIGN = label.transform(data.ALIGN) #заменяем значения из списка кодами закодированных элементов
flt_index = data['ALIGN'].unique()
flt_index
```

Out[17]:

```
array([1, 0, 2, 4, 3], dtype=int64)
```

In [18]:

```
import pandas
cat_columns = ['ID']
data_processed = pandas.get_dummies(data, prefix_sep="__",
                                    columns=cat_columns)
data_processed
```

Out[18]:

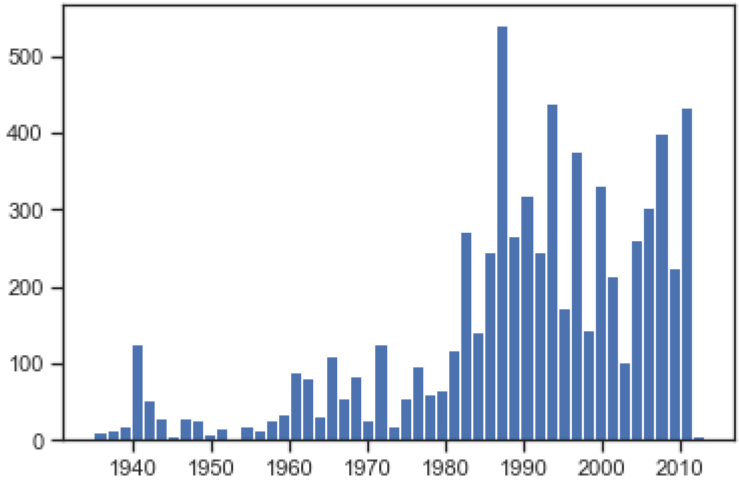
	page_id	name	urlslug	ALIGN	EYE	HAIR	SEX	G
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	1	Blue Eyes	Black Hair	Male Characters	M
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	1	Blue Eyes	Black Hair	Male Characters	M
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	1	Brown Eyes	Brown Hair	Male Characters	M
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	1	Brown Eyes	White Hair	Male Characters	M
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	1	Blue Eyes	Black Hair	Male Characters	M
5	1448	Wonder Woman (Diana Prince)	\/wiki\/Wonder_Woman_(Diana_Prince)	1	Blue Eyes	Black Hair	Female Characters	M
6	1486	Aquaman (Arthur Curry)	\/wiki\/Aquaman_(Arthur_Curry)	1	Blue Eyes	Blond Hair	Male Characters	M
7	1451	Timothy Drake (New Earth)	\/wiki\/Timothy_Drake_(New_Earth)	1	Blue Eyes	Black Hair	Male Characters	M
8	71760	Dinah Laurel Lance (New Earth)	\/wiki\/Dinah_Laurel_Lance_(New_Earth)	1	Blue Eyes	Blond Hair	Female Characters	M
9	1380	Flash (Barry Allen)	\/wiki\/Flash_(Barry_Allen)	1	Blue Eyes	Blond Hair	Male Characters	M
10	403631	GenderTest	\/wiki\/GenderTest	1	Blue Eyes	Blond Hair	Female Characters	M
11	1459	Alan Scott (New Earth)	\/wiki\/Alan_Scott_(New_Earth)	1	Blue Eyes	Blond Hair	Male Characters	M
12	1905	Barbara Gordon (New Earth)	\/wiki\/Barbara_Gordon_(New_Earth)	1	Blue Eyes	Red Hair	Female Characters	M
13	1386	Jason Garrick (New Earth)	\/wiki\/Jason_Garrick_(New_Earth)	1	Blue Eyes	Brown Hair	Male Characters	M
14	23383	Lois Lane (New Earth)	\/wiki\/Lois_Lane_(New_Earth)	1	Blue Eyes	Black Hair	Female Characters	M
15	1456	Alfred Pennyworth (New Earth)	\/wiki\/Alfred_Pennyworth_(New_Earth)	1	Blue Eyes	Black Hair	Male Characters	M
16	1849	Carter Hall (New Earth)	\/wiki\/Carter_Hall_(New_Earth)	1	Blue Eyes	Brown Hair	Male Characters	M
17	4320	Kyle Rayner (New Earth)	\/wiki\/Kyle_Rayner_(New_Earth)	1	Green Eyes	Black Hair	Male Characters	M
18	1706	Raymond Palmer (New Earth)	\/wiki\/Raymond_Palmer_(New_Earth)	1	Brown Eyes	NaN	Male Characters	M
19	1480	Alexander Luthor (New Earth)	\/wiki\/Alexander_Luthor_(New_Earth)	0	Green Eyes	NaN	Male Characters	M
20	1556	Roy Harper (New Earth)	\/wiki\/Roy_Harper_(New_Earth)	2	Green Eyes	Red Hair	Male Characters	M

	page_id	name	urlslug	ALIGN	EYE	HAIR	SEX	G
21	1580	Kara Zor-L (Earth-Two)	\/wiki\/Kara_Zor-L_(Earth-Two)	1	Blue Eyes	Blond Hair	Female Characters	N
22	4849	Ted Grant (New Earth)	\/wiki\/Ted_Grant_(New_Earth)	4	Blue Eyes	Black Hair	Male Characters	N
23	1611	Garfield Logan (New Earth)	\/wiki\/Garfield_Logan_(New_Earth)	1	Green Eyes	Green Hair	Male Characters	N
24	1479	Guy Gardner (New Earth)	\/wiki\/Guy_Gardner_(New_Earth)	1	Blue Eyes	Red Hair	Male Characters	N
25	1582	Victor Stone (New Earth)	\/wiki\/Victor_Stone_(New_Earth)	1	Brown Eyes	Black Hair	Male Characters	N
26	14006	Kon-El (New Earth)	\/wiki\/Kon-El_(New_Earth)	1	Blue Eyes	Black Hair	Male Characters	N
27	1484	Ralph Dibny (New Earth)	\/wiki\/Ralph_Dibny_(New_Earth)	4	Blue Eyes	Red Hair	Male Characters	N
28	23391	James Olsen (New Earth)	\/wiki\/James_Olsen_(New_Earth)	1	Green Eyes	Red Hair	Male Characters	N
29	1478	John Stewart (New Earth)	\/wiki\/John_Stewart_(New_Earth)	1	Brown Eyes	Black Hair	Male Characters	N
...	...	...	...	...	...	...	...	
6866	162822	Baron Tyrano (New Earth)	\/wiki\/Baron_Tyrano_(New_Earth)	0	Blue Eyes	NaN	Male Characters	N
6867	10025	Brains (New Earth)	\/wiki\/Brains_(New_Earth)	1	NaN	NaN	Male Characters	N
6868	10030	Cracker (New Earth)	\/wiki\/Cracker_(New_Earth)	1	NaN	NaN	Male Characters	N
6869	10031	Hard Head (New Earth)	\/wiki\/Hard_Head_(New_Earth)	1	NaN	Black Hair	Male Characters	N
6870	10032	Zig-Zag (New Earth)	\/wiki\/Zig-Zag_(New_Earth)	1	NaN	NaN	Male Characters	N
6871	228659	Dragonfly (New Earth)	\/wiki\/Dragonfly_(New_Earth)	0	NaN	Black Hair	Female Characters	N
6872	129755	Carl Bradford (New Earth)	\/wiki\/Carl_Bradford_(New_Earth)	0	NaN	NaN	Male Characters	N
6873	1449	Donna Troy (New Earth)	\/wiki\/Donna_Troy_(New_Earth)	1	Blue Eyes	Black Hair	Female Characters	N
6874	128098	Bartholomew Magan (New Earth)	\/wiki\/Bartholomew_Magan_(New_Earth)	0	NaN	NaN	Male Characters	N
6875	22325	James Moon (New Earth)	\/wiki\/James_Moon_(New_Earth)	4	NaN	NaN	Male Characters	N
6876	1383	Flash (Wally West)	\/wiki\/Flash_(Wally_West)	1	Green Eyes	Red Hair	Male Characters	N
6877	1485	J'onn J'onzz (New Earth)	\/wiki\/J%27onn_J%27onzz_(New_Earth)	1	Red Eyes	NaN	Male Characters	N
6878	34617	Dorothea Tane (New Earth)	\/wiki\/Dorothea_Tane_(New_Earth)	4	NaN	Blond Hair	Female Characters	N
6879	238641	Dmane (Earth-Two)	\/wiki\/Dmane_(Earth-Two)	0	Blue Eyes	NaN	Male Characters	N
6880	258830	Maximillian O'Leary (New Earth)	\/wiki\/Maximillian_O%27Leary_(New_Earth)	1	NaN	Black Hair	Male Characters	N

	page_id	name	urlslug	ALIGN	EYE	HAIR	SEX	G
6881	1624	Doris Zuel (New Earth)	\/wiki\/Doris_Zuel_(New_Earth)	0	Green Eyes	Red Hair	Female Characters	N
6882	22701	Doris Lee (New Earth)	\/wiki\/Doris_Lee_(New_Earth)	4	Brown Eyes	Brown Hair	Female Characters	N
6883	1581	Patrick O'Brian (New Earth)	\/wiki\/Patrick_O%27Brian_(New_Earth)	1	Blue Eyes	Black Hair	Male Characters	N
6884	1473	Basil Karlo (New Earth)	\/wiki\/Basil_Karlo_(New_Earth)	0	Black Eyes	Black Hair	Male Characters	N
6885	1460	Catwoman (Selina Kyle)	\/wiki\/Catwoman_(Selina_Kyle)	2	Green Eyes	Black Hair	Female Characters	N
6886	289378	Bedivere (New Earth)	\/wiki\/Bedivere_(New_Earth)	4	NaN	NaN	Male Characters	N
6887	283661	Herbert Hoover (New Earth)	\/wiki\/Herbert_Hoover_(New_Earth)	1	NaN	NaN	Male Characters	N
6888	283657	William Howard Taft (New Earth)	\/wiki\/William_Howard_Taft_(New_Earth)	1	NaN	NaN	Male Characters	N
6889	21655	Frank Fitzsimmons (New Earth)	\/wiki\/Frank_Fitzsimmons_(New_Earth)	1	NaN	Grey Hair	Male Characters	N
6890	283482	James Garfield (New Earth)	\/wiki\/James_Garfield_(New_Earth)	1	NaN	NaN	Male Characters	N
6891	66302	Nadine West (New Earth)	\/wiki\/Nadine_West_(New_Earth)	1	NaN	NaN	Female Characters	N
6892	283475	Warren Harding (New Earth)	\/wiki\/Warren_Harding_(New_Earth)	1	NaN	NaN	Male Characters	N
6893	283478	William Harrison (New Earth)	\/wiki\/William_Harrison_(New_Earth)	1	NaN	NaN	Male Characters	N
6894	283471	William McKinley (New Earth)	\/wiki\/William_McKinley_(New_Earth)	1	NaN	NaN	Male Characters	N
6895	150660	Mookie (New Earth)	\/wiki\/Mookie_(New_Earth)	0	Blue Eyes	Blond Hair	Male Characters	N

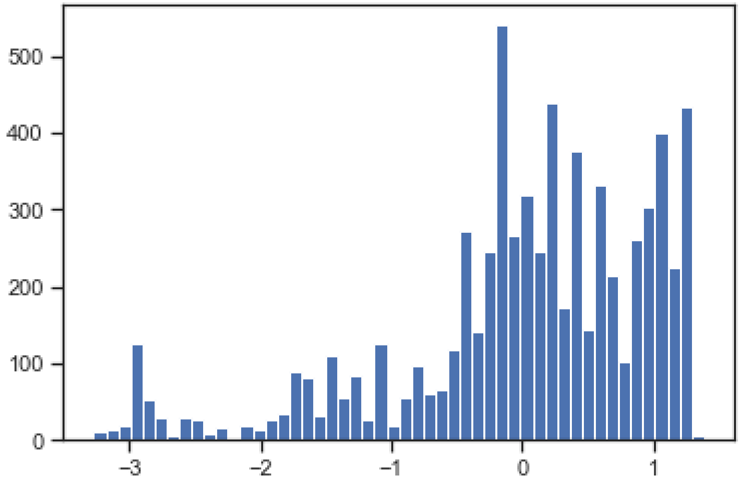
6896 rows × 15 columns

```
In [19]: from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer

         sc1 = MinMaxScaler()
         sc1_data = sc1.fit_transform(data[['YEAR']])
         plt.hist(data['YEAR'], 50)
         plt.show()
```



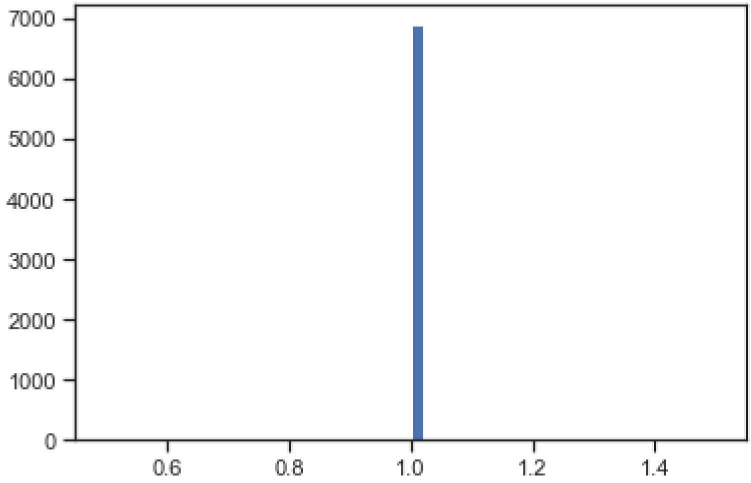
Масштабирование данных на основе Z-оценки

```
In [20]: sc2 = StandardScaler()
         sc2_data = sc2.fit_transform(data[['YEAR']])

         plt.hist(sc2_data, 50)
         plt.show()
```



In [21]:

```
sc3 = Normalizer()
sc3_data = sc3.fit_transform(data[['YEAR']])
flt_index = data['YEAR'].unique()
flt_index
plt.hist(sc3_data, 50)
plt.show()
```


7000
6000
5000
4000
3000
2000
1000
0
0.6
0.8
1.0
1.2
1.4