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

“Методы машинного обучения”

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

Лабораторная работа № 3

ИСПОЛНИТЕЛЬ:

Студент группы ИУ5-21М

Гузилов А.В. _____

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е._____

Москва 2019

Цель лабораторной работы: изучение способов предварительной обработки данных для дальнейшего формирования моделей.

Задание:

1. 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.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:

- обработку пропусков в данных;
- кодирование категориальных признаков;
- масштабирование данных.

In [3]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [6]:
```
data = pd.read_csv('marvel-wikia-data.csv', sep=",")
data.head(17000)
```

Out[6]:

	page_id	name	urlslug	ID	ALIG
0	1678	Spider-Man (Peter Parker)	\/Spider-Man_(Peter_Parker)	Secret Identity	Goo Character
1	7139	Captain America (Steven Rogers)	\/Captain_America_(Steven_Rogers)	Public Identity	Goo Character
2	64786	Wolverine (James \"Logan\" Howlett)	\/Wolverine_(James_%22Logan%22_Howlett)	Public Identity	Neutra Character
3	1868	Iron Man (Anthony \"Tony\" Stark)	\/Iron_Man_(Anthony_%22Tony%22_Stark)	Public Identity	Goo Character
4	2460	Thor (Thor Odinson)	\/Thor_(Thor_Odinson)	No Dual Identity	Goo Character
5	2458	Benjamin Grimm (Earth-616)	\/Benjamin_Grimm_(Earth-616)	Public Identity	Goo Character
6	2166	Reed Richards (Earth-616)	\/Reed_Richards_(Earth-616)	Public Identity	Goo Character
7	1833	Hulk (Robert Bruce Banner)	\/Hulk_(Robert_Bruce_Banner)	Public Identity	Goo Character
8	29481	Scott Summers (Earth-616)	\/Scott_Summers_(Earth-616)	Public Identity	Neutra Character
9	1837	Jonathan Storm (Earth-616)	\/Jonathan_Storm_(Earth-616)	Public Identity	Goo Character
10	15725	Henry McCoy (Earth-616)	\/Henry_McCoy_(Earth-616)	Public Identity	Goo Character
11	1863	Susan Storm (Earth-616)	\/Susan_Storm_(Earth-616)	Public Identity	Goo Character
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	Neutra Character
13	2614	Ororo Munroe (Earth-616)	\/Ororo_Munroe_(Earth-616)	Public Identity	Goo Character

14	1803	Clinton Barton (Earth-616)	\/Clinton_Barton_(Earth-616)	Public Identity	Goo Character
15	1396	Matthew Murdock (Earth-616)	\/Matthew_Murdock_(Earth-616)	Public Identity	Goo Character
16	55534	Stephen Strange (Earth-616)	\/Stephen_Strange_(Earth-616)	Public Identity	Goo Character
17	1978	Mary Jane Watson (Earth-616)	\/Mary_Jane_Watson_(Earth-616)	No Dual Identity	Goo Character
18	1872	John Jonah Jameson (Earth-616)	\/John_Jonah_Jameson_(Earth-616)	No Dual Identity	Neutra Character
19	35350	Robert Drake (Earth-616)	\/Robert_Drake_(Earth-616)	Secret Identity	Goo Character
20	1557	Henry Pym (Earth-616)	\/Henry_Pym_(Earth-616)	Public Identity	Goo Character
21	65255	Charles Xavier (Earth-616)	\/Charles_Xavier_(Earth-616)	Public Identity	Goo Character
22	1073	Warren Worthington III (Earth-616)	\/Warren_Worthington_III_(Earth-616)	Public Identity	Goo Character
23	1346	Piotr Rasputin (Earth-616)	\/Piotr_Rasputin_(Earth-616)	Secret Identity	Goo Character
24	2512	Wanda Maximoff (Earth-616)	\/Wanda_Maximoff_(Earth-616)	Public Identity	Goo Character
25	1671	Nicholas Fury (Earth-616)	\/Nicholas_Fury_(Earth-616)	No Dual Identity	Neutra Character
26	1976	Janet van Dyne (Earth-616)	\/Janet_van_Dyne_(Earth-616)	Public Identity	Goo Character
27	65219	Jean Grey (Earth-616)	\/Jean_Grey_(Earth-616)	Public Identity	Goo Character
28	6545	Natalia Romanova (Earth-616)	\/Natalia_Romanova_(Earth-616)	Public Identity	Goo Character
29	2223	Kurt Wagner (Earth-616)	\/Kurt_Wagner_(Earth-616)	Secret Identity	Goo Character
...	...	...	...	...	.
16346	42106	Thiazzi (Earth-616)	\/Thiazzi_(Earth-616)	NaN	Na
16347	462671	TOR\/test	\/User:TOR\/test	NaN	Na
16348	725312	Toxin (Luminals) (Earth-616)	\/Toxin_(Luminals)_(Earth-616)	Secret Identity	Na
16349	507239	Urizen Ul'var (Earth-616)	\/Urizen_Ul%27var_(Earth-616)	Public Identity	Goo Character
16350	40061	Valka (Earth-616)	\/Valka_(Earth-616)	NaN	Na

16351	505759	Vayor (Earth-616)	\/Vayor_(Earth-616)	No Dual Identity	Ba Character
16352	16569	Viridian (Earth-616)	\/Viridian_(Earth-616)	NaN	Na
16353	16296	William Burke (Earth-616)	\/William_Burke_(Earth-616)	NaN	Na
16354	120833	William Falsworth (Earth-616)	\/William_Falsworth_(Earth-616)	NaN	Goo Character
16355	693077	William Shakespeare (Earth-616)	\/William_Shakespeare_(Earth-616)	Public Identity	Neutra Character
16356	473710	Yamata no Orichi (Earth-616)	\/Yamata_no_Orichi_(Earth-616)	Public Identity	Ba Character
16357	670512	Zero G. Priestly (Legion Personality) (Earth-616)	\/Zero_G._Priestly_(Legion_Personality)_(Earth...	Secret Identity	Neutra Character
16358	12132	Zora Loftus (Earth-616)	\/Zora_Loftus_(Earth-616)	Secret Identity	Ba Character
16359	727035	Agar (Earth-616)	\/Agar_(Earth-616)	No Dual Identity	Ba Character
16360	703546	Ana (Natasha Romanoff's neighbor) (Earth-616)	\/Ana_(Natasha_Romanoff%27s_neighbor)_(Earth-616)	No Dual Identity	Goo Character
16361	713903	Dante's mother (Earth-616)	\/Dante%27s_mother_(Earth-616)	No Dual Identity	Neutra Character
16362	508693	Farbauti (Earth-616)	\/Farbauti_(Earth-616)	Public Identity	Ba Character
16363	694577	Finch (Kate Bishop's neighbor) (Earth-616)	\/Finch_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16364	655292	Jack O'Lantern (Impostor) (Earth-616)	\/Jack_O%27Lantern_(Impostor)_(Earth-616)	Secret Identity	Ba Character
16365	684262	K'thol (Earth-616)	\/K%27thol_(Earth-616)	NaN	Goo Character
16366	643435	Karen (Hijack's girlfriend) (Earth-616)	\/Karen_(Hijack%27s_girlfriend)_(Earth-616)	Public Identity	Neutra Character
16367	694583	Marcus (Kate Bishop's neighbor) (Earth-616)	\/Marcus_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16368	703892	Marcy (Offer's employee) (Earth-616)	\/Marcy_(Offer%27s_employee)_(Earth-616)	Public Identity	Neutra Character
		Melanie			

16369	660799	Melanie Kapoor (Earth-616)	\/Melanie_Kapoor_(Earth-616)	Public Identity	Goo Character
16370	674414	Phoenix's Shadow (Earth-616)	\/Phoenix%27s_Shadow_(Earth-616)	NaN	Neutra Character
16371	657508	Ru'ach (Earth-616)	\/Ru%27ach_(Earth-616)	No Dual Identity	Ba Character
16372	665474	Thane (Thanos' son) (Earth-616)	\/Thane_(Thanos%27_son)_(Earth-616)	No Dual Identity	Goo Character
16373	695217	Tinkerer (Skrull) (Earth-616)	\/Tinkerer_(Skrull)_(Earth-616)	Secret Identity	Ba Character
16374	708811	TK421 (Spiderling) (Earth-616)	\/TK421_(Spiderling)_(Earth-616)	Secret Identity	Neutra Character
16375	673702	Yologarch (Earth-616)	\/Yologarch_(Earth-616)	NaN	Ba Character

16376 rows × 13 columns

In [130]:
```
data.isnull().sum()
```

Out[130]:
```
page_id                0
name                   0
urlslug                0
ID                  3770
ALIGN               2812
EYE                 9767
HAIR                4264
SEX                  854
GSM                16286
ALIVE                  3
APPEARANCES         1096
FIRST APPEARANCE     815
Year                 815
dtype: int64
```

In [7]:
```
data.dtypes
```

Out[7]:
```
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 [8]:
```
data_new_1 = data.dropna(axis=1, how='any')
(data.shape, data_new_1.shape)
```

Out[8]:
```
((16376, 13), (16376, 3))
```

In [9]:
```
data_new_2 = data.dropna(axis=0, how='any')
(data.shape, data_new_2.shape)
```

Out[9]:
```
((16376, 13), (58, 13))
```

In [21]:
```
data_new_3 = data.fillna(0)
data_new_3.head(50)
```

Out[21]:

	page_id	name	urlslug	ID	ALIGN	EYE
0	1678	Spider-Man (Peter Parker)	\/Spider-Man_(Peter_Parker)	Secret Identity	Good Characters	Hazel Eyes
1	7139	Captain America (Steven Rogers)	\/Captain_America_(Steven_Rogers)	Public Identity	Good Characters	Blue Eyes
2	64786	Wolverine (James \"Logan\" Howlett)	\/Wolverine_(James_%22Logan%22_Howlett)	Public Identity	Neutral Characters	Blue Eyes
3	1868	Iron Man (Anthony \"Tony\" Stark)	\/Iron_Man_(Anthony_%22Tony%22_Stark)	Public Identity	Good Characters	Blue Eyes
4	2460	Thor (Thor Odinson)	\/Thor_(Thor_Odinson)	No Dual Identity	Good Characters	Blue Eyes
5	2458	Benjamin Grimm (Earth-616)	\/Benjamin_Grimm_(Earth-616)	Public Identity	Good Characters	Blue Eyes
6	2166	Reed Richards (Earth-616)	\/Reed_Richards_(Earth-616)	Public Identity	Good Characters	Brown Eyes
7	1833	Hulk (Robert Bruce Banner)	\/Hulk_(Robert_Bruce_Banner)	Public Identity	Good Characters	Brown Eyes
8	29481	Scott Summers (Earth-616)	\/Scott_Summers_(Earth-616)	Public Identity	Neutral Characters	Brown Eyes
9	1837	Jonathan Storm (Earth-616)	\/Jonathan_Storm_(Earth-616)	Public Identity	Good Characters	Blue Eyes
10	15725	Henry McCoy (Earth-616)	\/Henry_McCoy_(Earth-616)	Public Identity	Good Characters	Blue Eyes
11	1863	Susan Storm (Earth-616)	\/Susan_Storm_(Earth-616)	Public Identity	Good Characters	Blue Eyes
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	Neutral Characters	Green Eyes

13	2614	Ororo Munroe (Earth-616)	\/Ororo_Munroe_(Earth-616)	Public Identity	Good Characters	Blue Eyes
14	1803	Clinton Barton (Earth-616)	\/Clinton_Barton_(Earth-616)	Public Identity	Good Characters	Blue Eyes
15	1396	Matthew Murdock (Earth-616)	\/Matthew_Murdock_(Earth-616)	Public Identity	Good Characters	Blue Eyes
16	55534	Stephen Strange (Earth-616)	\/Stephen_Strange_(Earth-616)	Public Identity	Good Characters	Grey Eyes
17	1978	Mary Jane Watson (Earth-616)	\/Mary_Jane_Watson_(Earth-616)	No Dual Identity	Good Characters	Green Eyes
18	1872	John Jonah Jameson (Earth-616)	\/John_Jonah_Jameson_(Earth-616)	No Dual Identity	Neutral Characters	Blue Eyes
19	35350	Robert Drake (Earth-616)	\/Robert_Drake_(Earth-616)	Secret Identity	Good Characters	Brown Eyes
20	1557	Henry Pym (Earth-616)	\/Henry_Pym_(Earth-616)	Public Identity	Good Characters	Blue Eyes
21	65255	Charles Xavier (Earth-616)	\/Charles_Xavier_(Earth-616)	Public Identity	Good Characters	Blue Eyes
22	1073	Warren Worthington III (Earth-616)	\/Warren_Worthington_III_(Earth-616)	Public Identity	Good Characters	Blue Eyes
23	1346	Piotr Rasputin (Earth-616)	\/Piotr_Rasputin_(Earth-616)	Secret Identity	Good Characters	Blue Eyes
24	2512	Wanda Maximoff (Earth-616)	\/Wanda_Maximoff_(Earth-616)	Public Identity	Good Characters	Green Eyes
25	1671	Nicholas Fury (Earth-616)	\/Nicholas_Fury_(Earth-616)	No Dual Identity	Neutral Characters	Brown Eyes
26	1976	Janet van Dyne (Earth-616)	\/Janet_van_Dyne_(Earth-616)	Public Identity	Good Characters	Blue Eyes
27	65219	Jean Grey (Earth-616)	\/Jean_Grey_(Earth-616)	Public Identity	Good Characters	Green Eyes
28	6545	Natalia Romanova (Earth-616)	\/Natalia_Romanova_(Earth-616)	Public Identity	Good Characters	Green Eyes
29	2223	Kurt Wagner (Earth-616)	\/Kurt_Wagner_(Earth-616)	Secret Identity	Good Characters	Yellow Eyes
30	2414	Vision (Earth-616)	\/Vision_(Earth-616)	Secret Identity	Good Characters	Gold Eyes
31	8650	May Reilly (Earth-616)	\/May_Reilly_(Earth-616)	No Dual Identity	Good Characters	Blue Eyes
32	2527	Katherine Pryde (Earth-616)	\/Katherine_Pryde_(Earth-616)	Secret Identity	Good Characters	Hazel Eyes
33	1970	Carol	\/Carol_Danvers_(Earth-616)	Public	Good	Blue

		Danvers (Earth-616)		Identity	Characters	Eyes
34	37690	Jennifer Walters (Earth-616)	\/Jennifer_Walters_(Earth-616)	Public Identity	Good Characters	Green Eyes
35	3765	Emma Frost (Earth-616)	\/Emma_Frost_(Earth-616)	Public Identity	Neutral Characters	Blue Eyes
36	2312	Frank Castle (Earth-616)	\/Frank_Castle_(Earth-616)	Public Identity	Neutral Characters	Blue Eyes
37	1265	Luke Cage (Earth-616)	\/Luke_Cage_(Earth-616)	No Dual Identity	Good Characters	Brown Eyes
38	1677	Rogue (Anna Marie) (Earth-616)	\/Rogue_(Anna_Marie)_(Earth-616)	Secret Identity	Good Characters	Green Eyes
39	25222	Conan (Earth-616)	\/Conan_(Earth-616)	No Dual Identity	Neutral Characters	Blue Eyes
40	6537	Joseph Robertson (Earth-616)	\/Joseph_Robertson_(Earth-616)	No Dual Identity	Good Characters	Brown Eyes
41	1592	Pietro Maximoff (Earth-616)	\/Pietro_Maximoff_(Earth-616)	Public Identity	Good Characters	Blue Eyes
42	1818	Hercules (Earth-616)	\/Hercules_(Earth-616)	No Dual Identity	Good Characters	Blue Eyes
43	1448	Victor von Doom (Earth-616)	\/Victor_von_Doom_(Earth-616)	Public Identity	Bad Characters	Brown Eyes
44	2081	Max Eisenhardt (Earth-616)	\/Max_Eisenhardt_(Earth-616)	Public Identity	Neutral Characters	Grey Eyes
45	2307	Elizabeth Braddock (Earth-616)	\/Elizabeth_Braddock_(Earth-616)	Secret Identity	Neutral Characters	Blue Eyes
46	2556	Norrin Radd (Earth-616)	\/Norrin_Radd_(Earth-616)	Public Identity	Good Characters	Blue Eyes
47	1740	Norman Osborn (Earth-616)	\/Norman_Osborn_(Earth-616)	Public Identity	Bad Characters	Blue Eyes
48	8651	Eugene Thompson (Earth-616)	\/Eugene_Thompson_(Earth-616)	Secret Identity	Good Characters	Blue Eyes
49	2009	Simon Williams (Earth-616)	\/Simon_Williams_(Earth-616)	Public Identity	Good Characters	Red Eyes

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

In [131]:
```
# 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
1096.
Колонка Year. Тип данных float64. Количество пустых значений 815.
```

In [132]:
```
data_num = data[num_cols]
data_num
```

Out[132]:

	APPEARANCES	Year
0	4043.0	1962.0
1	3360.0	1941.0
2	3061.0	1974.0
3	2961.0	1963.0
4	2258.0	1950.0
5	2255.0	1961.0
6	2072.0	1961.0
7	2017.0	1962.0
8	1955.0	1963.0
9	1934.0	1961.0
10	1825.0	1963.0
11	1713.0	1961.0
12	1528.0	NaN
13	1512.0	1975.0
14	1394.0	1964.0
15	1338.0	1964.0
16	1307.0	1963.0
17	1304.0	1965.0
18	1266.0	1963.0
19	1265.0	1963.0
20	1237.0	1962.0
21	1233.0	1963.0
22	1230.0	1963.0
23	1162.0	1975.0
24	1161.0	1964.0
25	1137.0	1963.0
26	1120.0	1963.0
27	1107.0	1963.0
28	1050.0	1964.0

29	1047.0	1975.0
...	...	...
16346	NaN	NaN
16347	NaN	NaN
16348	NaN	NaN
16349	NaN	NaN
16350	NaN	NaN
16351	NaN	NaN
16352	NaN	NaN
16353	NaN	NaN
16354	NaN	NaN
16355	NaN	NaN
16356	NaN	NaN
16357	NaN	NaN
16358	NaN	NaN
16359	NaN	NaN
16360	NaN	NaN
16361	NaN	NaN
16362	NaN	NaN
16363	NaN	NaN
16364	NaN	NaN
16365	NaN	NaN
16366	NaN	NaN
16367	NaN	NaN
16368	NaN	NaN
16369	NaN	NaN
16370	NaN	NaN
16371	NaN	NaN
16372	NaN	NaN
16373	NaN	NaN
16374	NaN	NaN
16375	NaN	NaN

16376 rows × 2 columns

In [68]:
```
# Фильтр по пустым значениям поля MasVnrArea
data[data['Year'].isnull()]
```

Out[68]:

	page_id	name	urlslug	ID	ALIG
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	Neutra Character
38	1677	Rogue (Anna Marie)	\/Rogue_(Anna_Marie)_(Earth-616)	Secret Identity	Goo Character

		(Earth-616)			
80	67048	Blaine Colt (Earth-616)	\/Blaine_Colt_(Earth-616)	Public Identity	Na
114	37751	Monica Rambeau (Earth-616)	\/Monica_Rambeau_(Earth-616)	Secret Identity	Goo Character
259	25255	James Bradley (Earth-616)	\/James_Bradley_(Earth-616)	Secret Identity	Goo Character
310	535259	Steel (Earth-616)	\/Steel_(Earth-616)	Public Identity	Goo Character
413	626857	Howard Hanover (Earth-616)	\/Howard_Hanover_(Earth-616)	Public Identity	Na
683	13066	Fen (Earth-616)	\/Fen_(Earth-616)	No Dual Identity	Goo Character
789	103136	Thakorr (Earth-616)	\/Thakorr_(Earth-616)	No Dual Identity	Ba Character
854	30883	Brightwind (Earth-616)	\/Brightwind_(Earth-616)	NaN	Na
997	18445	Ahura Boltagon (Earth-616)	\/Ahura_Boltagon_(Earth-616)	No Dual Identity	Na
1118	17489	Francesca Grace (Earth-616)	\/Francesca_Grace_(Earth-616)	Secret Identity	Na
1158	1771	Alistaire Smythe (Earth-616)	\/Alistaire_Smythe_(Earth-616)	Public Identity	Ba Character
1316	2126	Martinique Wyngarde (Earth-616)	\/Martinique_Wyngarde_(Earth-616)	Secret Identity	Ba Character
1454	93233	Leonard McKenzie (Earth-616)	\/Leonard_McKenzie_(Earth-616)	No Dual Identity	Goo Character
1564	40064	Hrimhari (Earth-616)	\/Hrimhari_(Earth-616)	Secret Identity	Goo Character
1565	638526	David Bond (Earth-616)	\/David_Bond_(Earth-616)	Secret Identity	Goo Character
1845	631202	Aldrif Odinsdottir (Earth-616)	\/Aldrif_Odinsdottir_(Earth-616)	Secret Identity	Neutra Character
1937	655072	Shogo Lee (Earth-616)	\/Shogo_Lee_(Earth-616)	No Dual Identity	Neutra Character
2033	6632	Leila Davis (Earth-616)	\/Leila_Davis_(Earth-616)	Public Identity	Neutra Character
2147	665003	Theodore Marley Brooks (Earth-616)	\/Theodore_Marley_Brooks_(Earth-616)	No Dual Identity	Goo Character
2148	670564	Arno Stark (Earth-616)	\/Arno_Stark_(Earth-616)	No Dual Identity	Goo Character
2398	694425	John Renwick	\/John_Renwick_(Earth-616)	No Dual	Goo Character

		(Earth-616)		Identity	
2399	640754	Proxima Midnight (Earth-616)	\/Proxima_Midnight_(Earth-616)	No Dual Identity	Ba Character
2549	21721	Perrikus (Earth-616)	\/Perrikus_(Earth-616)	NaN	Ba Character
2727	635523	Doombot (Avenger) (Earth-616)	\/Doombot_(Avenger)_(Earth-616)	Public Identity	Goo Character
2728	678469	Selah Burke (Earth-616)	\/Selah_Burke_(Earth-616)	Secret Identity	Goo Character
2909	17957	Teresa Martin (Earth-616)	\/Teresa_Martin_(Earth-616)	Public Identity	Goo Character
2910	635275	Alexis (Earth-616)	\/Alexis_(Earth-616)	Public Identity	Goo Character
2911	674895	Kamala Khan (Earth-616)	\/Kamala_Khan_(Earth-616)	Secret Identity	Goo Character
...	...	...	...	...	.
16346	42106	Thiazzi (Earth-616)	\/Thiazzi_(Earth-616)	NaN	Na
16347	462671	TOR\/test	\/User:TOR\/test	NaN	Na
16348	725312	Toxin (Luminals) (Earth-616)	\/Toxin_(Luminals)_(Earth-616)	Secret Identity	Na
16349	507239	Urizen Ul'var (Earth-616)	\/Urizen_Ul%27var_(Earth-616)	Public Identity	Goo Character
16350	40061	Valka (Earth-616)	\/Valka_(Earth-616)	NaN	Na
16351	505759	Vayor (Earth-616)	\/Vayor_(Earth-616)	No Dual Identity	Ba Character
16352	16569	Viridian (Earth-616)	\/Viridian_(Earth-616)	NaN	Na
16353	16296	William Burke (Earth-616)	\/William_Burke_(Earth-616)	NaN	Na
16354	120833	William Falsworth (Earth-616)	\/William_Falsworth_(Earth-616)	NaN	Goo Character
16355	693077	William Shakespeare (Earth-616)	\/William_Shakespeare_(Earth-616)	Public Identity	Neutra Character
16356	473710	Yamata no Orichi (Earth-616)	\/Yamata_no_Orichi_(Earth-616)	Public Identity	Ba Character
16357	670512	Zero G. Priestly (Legion Personality) (Earth-616)	\/Zero_G._Priestly_(Legion_Personality)_(Earth...	Secret Identity	Neutra Character
16358	12132	Zora Loftus (Earth-616)	\/Zora_Loftus_(Earth-616)	Secret Identity	Ba Character
16359	727035	Agar (Earth-616)	\/Agar_(Earth-616)	No Dual Identity	Ba Character

16360	703546	Ana (Natasha Romanoff's neighbor) (Earth-616)	\/Ana_(Natasha_Romanoff%27s_neighbor)_(Earth-616)	No Dual Identity	Goo Character
16361	713903	Dante's mother (Earth-616)	\/Dante%27s_mother_(Earth-616)	No Dual Identity	Neutra Character
16362	508693	Farbauti (Earth-616)	\/Farbauti_(Earth-616)	Public Identity	Ba Character
16363	694577	Finch (Kate Bishop's neighbor) (Earth-616)	\/Finch_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16364	655292	Jack O'Lantern (Impostor) (Earth-616)	\/Jack_O%27Lantern_(Impostor)_(Earth-616)	Secret Identity	Ba Character
16365	684262	K'thol (Earth-616)	\/K%27thol_(Earth-616)	NaN	Goo Character
16366	643435	Karen (Hijack's girlfriend) (Earth-616)	\/Karen_(Hijack%27s_girlfriend)_(Earth-616)	Public Identity	Neutra Character
16367	694583	Marcus (Kate Bishop's neighbor) (Earth-616)	\/Marcus_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16368	703892	Marcy (Offer's employee) (Earth-616)	\/Marcy_(Offer%27s_employee)_(Earth-616)	Public Identity	Neutra Character
16369	660799	Melanie Kapoor (Earth-616)	\/Melanie_Kapoor_(Earth-616)	Public Identity	Goo Character
16370	674414	Phoenix's Shadow (Earth-616)	\/Phoenix%27s_Shadow_(Earth-616)	NaN	Neutra Character
16371	657508	Ru'ach (Earth-616)	\/Ru%27ach_(Earth-616)	No Dual Identity	Ba Character
16372	665474	Thane (Thanos' son) (Earth-616)	\/Thane_(Thanos%27_son)_(Earth-616)	No Dual Identity	Goo Character
16373	695217	Tinkerer (Skrull) (Earth-616)	\/Tinkerer_(Skrull)_(Earth-616)	Secret Identity	Ba Character
16374	708811	TK421 (Spiderling) (Earth-616)	\/TK421_(Spiderling)_(Earth-616)	Secret Identity	Neutra Character
16375	673702	Yologarch (Earth-616)	\/Yologarch_(Earth-616)	NaN	Ba Character

815 rows × 13 columns

In [75]:
```
# Запоминаем индексы строк с пустыми значениями
flt_index = data[data['Year'].isnull()].index
flt_index
```

Out[75]:
```
Int64Index([   12,    38,    80,   114,   259,   310,   413,   683,
789,
              854,
             ...
             16366, 16367, 16368, 16369, 16370, 16371, 16372, 16373,
16374,
             16375],
            dtype='int64', length=815)
```

In [76]:
```
# Проверяем что выводятся нужные строки
data[data.index.isin(flt_index)]
```

Out[76]:

	page_id	name	urlslug	ID	ALIG
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	Neutr Character
38	1677	Rogue (Anna Marie) (Earth-616)	\/Rogue_(Anna_Marie)_(Earth-616)	Secret Identity	Goo Character
80	67048	Blaine Colt (Earth-616)	\/Blaine_Colt_(Earth-616)	Public Identity	Na
114	37751	Monica Rambeau (Earth-616)	\/Monica_Rambeau_(Earth-616)	Secret Identity	Goo Character
259	25255	James Bradley (Earth-616)	\/James_Bradley_(Earth-616)	Secret Identity	Goo Character
310	535259	Steel (Earth-616)	\/Steel_(Earth-616)	Public Identity	Goo Character
413	626857	Howard Hanover (Earth-616)	\/Howard_Hanover_(Earth-616)	Public Identity	Na
683	13066	Fen (Earth-616)	\/Fen_(Earth-616)	No Dual Identity	Goo Character
789	103136	Thakorr (Earth-616)	\/Thakorr_(Earth-616)	No Dual Identity	Ba Character
854	30883	Brightwind (Earth-616)	\/Brightwind_(Earth-616)	NaN	Na
997	18445	Ahura Boltagon (Earth-616)	\/Ahura_Boltagon_(Earth-616)	No Dual Identity	Na
1118	17489	Francesca Grace (Earth-616)	\/Francesca_Grace_(Earth-616)	Secret Identity	Na
1158	1771	Alistaire Smythe (Earth-616)	\/Alistaire_Smythe_(Earth-616)	Public Identity	Ba Character
1316	2126	Martinique Wyngarde (Earth-616)	\/Martinique_Wyngarde_(Earth-616)	Secret Identity	Ba Character
1454	93233	Leonard McKenzie (Earth-616)	\/Leonard_McKenzie_(Earth-616)	No Dual Identity	Goo Character
1564	40064	Hrimhari	\/Hrimhari_(Earth-616)	Secret	Goo

		(Earth-616)		Identity	Character
1565	638526	David Bond (Earth-616)	\/David_Bond_(Earth-616)	Secret Identity	Good Character
1845	631202	Aldrif Odinsdottir (Earth-616)	\/Aldrif_Odinsdottir_(Earth-616)	Secret Identity	Neutral Character
1937	655072	Shogo Lee (Earth-616)	\/Shogo_Lee_(Earth-616)	No Dual Identity	Neutral Character
2033	6632	Leila Davis (Earth-616)	\/Leila_Davis_(Earth-616)	Public Identity	Neutral Character
2147	665003	Theodore Marley Brooks (Earth-616)	\/Theodore_Marley_Brooks_(Earth-616)	No Dual Identity	Good Character
2148	670564	Arno Stark (Earth-616)	\/Arno_Stark_(Earth-616)	No Dual Identity	Good Character
2398	694425	John Renwick (Earth-616)	\/John_Renwick_(Earth-616)	No Dual Identity	Good Character
2399	640754	Proxima Midnight (Earth-616)	\/Proxima_Midnight_(Earth-616)	No Dual Identity	Bad Character
2549	21721	Perrikus (Earth-616)	\/Perrikus_(Earth-616)	NaN	Bad Character
2727	635523	Doombot (Avenger) (Earth-616)	\/Doombot_(Avenger)_(Earth-616)	Public Identity	Good Character
2728	678469	Selah Burke (Earth-616)	\/Selah_Burke_(Earth-616)	Secret Identity	Good Character
2909	17957	Teresa Martin (Earth-616)	\/Teresa_Martin_(Earth-616)	Public Identity	Good Character
2910	635275	Alexis (Earth-616)	\/Alexis_(Earth-616)	Public Identity	Good Character
2911	674895	Kamala Khan (Earth-616)	\/Kamala_Khan_(Earth-616)	Secret Identity	Good Character
...	...	...	...	...	.
16346	42106	Thiazzi (Earth-616)	\/Thiazzi_(Earth-616)	NaN	Na
16347	462671	TOR\/test	\/User:TOR\/test	NaN	Na
16348	725312	Toxin (Luminals) (Earth-616)	\/Toxin_(Luminals)_(Earth-616)	Secret Identity	Na
16349	507239	Urizen Ul'var (Earth-616)	\/Urizen_Ul%27var_(Earth-616)	Public Identity	Good Character
16350	40061	Valka (Earth-616)	\/Valka_(Earth-616)	NaN	Na
16351	505759	Vayor (Earth-616)	\/Vayor_(Earth-616)	No Dual Identity	Bad Character
16352	16569	Viridian (Earth-616)	\/Viridian_(Earth-616)	NaN	Na
16353	16296	William	\/William_Burke_(Earth-616)	NaN	Na

		Burke (Earth-616)			
16354	120833	William Falsworth (Earth-616)	\/William_Falsworth_(Earth-616)	NaN	Goo Character
16355	693077	William Shakespeare (Earth-616)	\/William_Shakespeare_(Earth-616)	Public Identity	Neutra Character
16356	473710	Yamata no Orichi (Earth-616)	\/Yamata_no_Orichi_(Earth-616)	Public Identity	Ba Character
16357	670512	Zero G. Priestly (Legion Personality) (Earth-616)	\/Zero_G._Priestly_(Legion_Personality)_(Earth...	Secret Identity	Neutra Character
16358	12132	Zora Loftus (Earth-616)	\/Zora_Loftus_(Earth-616)	Secret Identity	Ba Character
16359	727035	Agar (Earth-616)	\/Agar_(Earth-616)	No Dual Identity	Ba Character
16360	703546	Ana (Natasha Romanoff's neighbor) (Earth-616)	\/Ana_(Natasha_Romanoff%27s_neighbor)_(Earth-616)	No Dual Identity	Goo Character
16361	713903	Dante's mother (Earth-616)	\/Dante%27s_mother_(Earth-616)	No Dual Identity	Neutra Character
16362	508693	Farbauti (Earth-616)	\/Farbauti_(Earth-616)	Public Identity	Ba Character
16363	694577	Finch (Kate Bishop's neighbor) (Earth-616)	\/Finch_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16364	655292	Jack O'Lantern (Impostor) (Earth-616)	\/Jack_O%27Lantern_(Impostor)_(Earth-616)	Secret Identity	Ba Character
16365	684262	K'thol (Earth-616)	\/K%27thol_(Earth-616)	NaN	Goo Character
16366	643435	Karen (Hijack's girlfriend) (Earth-616)	\/Karen_(Hijack%27s_girlfriend)_(Earth-616)	Public Identity	Neutra Character
16367	694583	Marcus (Kate Bishop's neighbor) (Earth-616)	\/Marcus_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16368	703892	Marcy (Offer's employee) (Earth-616)	\/Marcy_(Offer%27s_employee)_(Earth-616)	Public Identity	Neutra Character
16369	660799	Melanie Kapoor (Earth-616)	\/Melanie_Kapoor_(Earth-616)	Public Identity	Goo Character
16370	674414	Phoenix's Shadow (Earth-616)	\/Phoenix%27s_Shadow_(Earth-616)	NaN	Neutra Character
16371	657508	Ru'ach	\/Ru%27ach_(Earth-616)	No	Ba

		(Earth-616)		Dual Identity	Character
16372	665474	Thane (Thanos' son) (Earth-616)	\/Thane_(Thanos%27_son)_(Earth-616)	No Dual Identity	Good Character
16373	695217	Tinkerer (Skrull) (Earth-616)	\/Tinkerer_(Skrull)_(Earth-616)	Secret Identity	Bad Character
16374	708811	TK421 (Spiderling) (Earth-616)	\/TK421_(Spiderling)_(Earth-616)	Secret Identity	Neutral Character
16375	673702	Yologarch (Earth-616)	\/Yologarch_(Earth-616)	NaN	Bad Character

815 rows × 13 columns

In [82]:
```
data[data['Year'].isnull()]
# Сохраняем индексы
flt_index = data[data['Year'].isnull()].index
flt_index
```

Out[82]:
```
Int64Index([   12,    38,    80,   114,   259,   310,   413,   683,
789,
              854,
             ...
            16366, 16367, 16368, 16369, 16370, 16371, 16372, 16373,
16374,
            16375],
           dtype='int64', length=815)
```

In [83]:
```
# фильтр по колонке
data_num[data_num.index.isin(flt_index)]['Year']
```

Out[83]:
```
12      NaN
38      NaN
80      NaN
114     NaN
259     NaN
310     NaN
413     NaN
683     NaN
789     NaN
854     NaN
997     NaN
1118    NaN
1158    NaN
1316    NaN
1454    NaN
1564    NaN
1565    NaN
1845    NaN
1937    NaN
2033    NaN
2147    NaN
2148    NaN
2398    NaN
2399    NaN
2549    NaN

2727    NaN
```

```
2728     NaN
2909     NaN
2910     NaN
2911     NaN
          ..
16346    NaN
16347    NaN
16348    NaN
16349    NaN
16350    NaN
16351    NaN
16352    NaN
16353    NaN
16354    NaN
16355    NaN
16356    NaN
16357    NaN
16358    NaN
16359    NaN
16360    NaN
16361    NaN
16362    NaN
16363    NaN
16364    NaN
16365    NaN
16366    NaN
16367    NaN
16368    NaN
16369    NaN
16370    NaN
16371    NaN
16372    NaN
16373    NaN
16374    NaN
16375    NaN
Name: Year, Length: 815, dtype: float64
```

In [86]:
```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
```

In [96]:
```
data[['Year']].describe()
```

Out[96]:

	Year
count	15561.000000
mean	1984.951803
std	19.663571
min	1939.000000
25%	1974.000000
50%	1990.000000
75%	2000.000000
max	2013.000000

In [105]:
```
# 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
1096.
Колонка Year. Тип данных float64. Количество пустых значений 815.
```

In [107]:
```
# Фильтр по пустым значениям поля Year
data[data['Year'].isnull()]
# Сохраняем индексы
flt_index = data[data['Year'].isnull()].index
flt_index
```

Out[107]:
```
Int64Index([   12,    38,    80,   114,   259,   310,   413,   683,
789,
              854,
            ...
            16366, 16367, 16368, 16369, 16370, 16371, 16372, 16373,
16374,
            16375],
           dtype='int64', length=815)
```

In [116]:
```
for rows in flt_index:
  data.Year[rows]=data.Year.median()
```

```
c:\program files\python36\lib\site-packages\ipykernel_launcher.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
```

1.2.2. Обработка пропусков в категориальных данных

In [208]:
```
# 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 3770, 23.0
2%.
Колонка ALIGN. Тип данных object. Количество пустых значений 2812, 1
7.17%.

Колонка EYE. Тип данных object. Количество пустых значений 9767, 59.
64%.
Колонка HAIR. Тип данных object. Количество пустых значений 4264, 2
```

```
6.04%.
Колонка SEX. Тип данных object. Количество пустых значений 854, 5.2
1%.
Колонка GSM. Тип данных object. Количество пустых значений 16286, 9
9.45%.
Колонка ALIVE. Тип данных object. Количество пустых значений 3, 0.0
2%.
Колонка FIRST APPEARANCE. Тип данных object. Количество пустых значе
ний 815, 4.98%.
```

In [209]:
```
data_num = data[cat_cols]
data_num
```

Out[209]:

	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	FIRST APPEARANCE
0	Secret Identity	Good Characters	Hazel Eyes	Brown Hair	Male Characters	NaN	Living Characters	Aug-62
1	Public Identity	Good Characters	Blue Eyes	White Hair	Male Characters	NaN	Living Characters	Mar-41
2	Public Identity	Neutral Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	Oct-74
3	Public Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	Mar-63
4	No Dual Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	Nov-50
5	Public Identity	Good Characters	Blue Eyes	No Hair	Male Characters	NaN	Living Characters	Nov-61
6	Public Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	Nov-61
7	Public Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	May-62
8	Public Identity	Neutral Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	Sep-63
9	Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	Nov-61
10	Public Identity	Good Characters	Blue Eyes	Blue Hair	Male Characters	NaN	Living Characters	Sep-63
11	Public Identity	Good Characters	Blue Eyes	Blond Hair	Female Characters	NaN	Living Characters	Nov-61
12	No Dual Identity	Neutral Characters	Green Eyes	Black Hair	Male Characters	NaN	Living Characters	NaN
13	Public Identity	Good Characters	Blue Eyes	White Hair	Female Characters	NaN	Living Characters	May-75
14	Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	Sep-64
15	Public Identity	Good Characters	Blue Eyes	Red Hair	Male Characters	NaN	Living Characters	Apr-64
16	Public Identity	Good Characters	Grey Eyes	Black Hair	Male Characters	NaN	Living Characters	Jul-63
17	No Dual Identity	Good Characters	Green Eyes	Red Hair	Female Characters	NaN	Living Characters	Jun-65
18	No	Neutral	Blue	Black	Male	NaN	Living	Mar-63

	Dual Identity	Characters	Eyes	Hair	Characters		Characters	
19	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	Sep-63
20	Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	Jan-62
21	Public Identity	Good Characters	Blue Eyes	Bald	Male Characters	NaN	Deceased Characters	Sep-63
22	Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters	NaN	Living Characters	Sep-63
23	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	May-75
24	Public Identity	Good Characters	Green Eyes	Brown Hair	Female Characters	NaN	Living Characters	Mar-64
25	No Dual Identity	Neutral Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	May-63
26	Public Identity	Good Characters	Blue Eyes	Auburn Hair	Female Characters	NaN	Living Characters	Jun-63
27	Public Identity	Good Characters	Green Eyes	Red Hair	Female Characters	NaN	Deceased Characters	Sep-63
28	Public Identity	Good Characters	Green Eyes	Red Hair	Female Characters	Bisexual Characters	Living Characters	Apr-64
29	Secret Identity	Good Characters	Yellow Eyes	Blue Hair	Male Characters	NaN	Living Characters	May-75
...	...	...	...	...	...	...	...	...
16346	NaN	NaN	NaN	NaN	Male Characters	NaN	Deceased Characters	NaN
16347	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
16348	Secret Identity	NaN	NaN	No Hair	NaN	NaN	Living Characters	NaN
16349	Public Identity	Good Characters	White Eyes	NaN	Male Characters	NaN	Living Characters	NaN
16350	NaN	NaN	NaN	NaN	NaN	NaN	Living Characters	NaN
16351	No Dual Identity	Bad Characters	NaN	Black Hair	Male Characters	NaN	Living Characters	NaN
16352	NaN	NaN	NaN	NaN	Male Characters	NaN	Living Characters	NaN
16353	NaN	NaN	NaN	NaN	Male Characters	NaN	Living Characters	NaN
16354	NaN	Good Characters	NaN	NaN	Male Characters	NaN	Deceased Characters	NaN
16355	Public Identity	Neutral Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN
16356	Public Identity	Bad Characters	NaN	No Hair	Male Characters	NaN	Living Characters	NaN
16357	Secret Identity	Neutral Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN
16358	Secret Identity	Bad Characters	NaN	Blond Hair	Female Characters	NaN	Living Characters	NaN
16359	No Dual	Bad Characters	Black Eyes	No Hair	Male Characters	NaN	Living Characters	NaN

	Identity							
16360	No Dual Identity	Good Characters	Black Eyes	Grey Hair	Female Characters	NaN	Living Characters	NaN
16361	No Dual Identity	Neutral Characters	NaN	NaN	Female Characters	NaN	Deceased Characters	NaN
16362	Public Identity	Bad Characters	Red Eyes	NaN	Female Characters	NaN	Living Characters	NaN
16363	Public Identity	Good Characters	Black Eyes	Bald	Male Characters	Homosexual Characters	Living Characters	NaN
16364	Secret Identity	Bad Characters	Hazel Eyes	Bald	Male Characters	NaN	Living Characters	NaN
16365	NaN	Good Characters	NaN	NaN	Male Characters	NaN	Deceased Characters	NaN
16366	Public Identity	Neutral Characters	Brown Eyes	Black Hair	Female Characters	NaN	Living Characters	NaN
16367	Public Identity	Good Characters	Hazel Eyes	Bald	Male Characters	Homosexual Characters	Living Characters	NaN
16368	Public Identity	Neutral Characters	NaN	Brown Hair	Female Characters	NaN	Living Characters	NaN
16369	Public Identity	Good Characters	Blue Eyes	Black Hair	Female Characters	NaN	Living Characters	NaN
16370	NaN	Neutral Characters	NaN	NaN	NaN	NaN	Living Characters	NaN
16371	No Dual Identity	Bad Characters	Green Eyes	No Hair	Male Characters	NaN	Living Characters	NaN
16372	No Dual Identity	Good Characters	Blue Eyes	Bald	Male Characters	NaN	Living Characters	NaN
16373	Secret Identity	Bad Characters	Black Eyes	Bald	Male Characters	NaN	Living Characters	NaN
16374	Secret Identity	Neutral Characters	NaN	NaN	Male Characters	NaN	Living Characters	NaN
16375	NaN	Bad Characters	NaN	NaN	NaN	NaN	Living Characters	NaN

16376 rows × 8 columns

In [198]:
```
# Фильтр по пустым значениям поля Year
data[data['HAIR'].isnull()]
# Сохраняем индексы
flt_index = data[data['HAIR'].isnull()].index
flt_index
```

Out[198]:
```
Int64Index([], dtype='int64')
```

In [212]:
```
data[['HAIR']].describe()
```

Out[212]:

	HAIR
count	16376
unique	25

top	Purple Hair
freq	4311

In [211]:
```
MaxPassEmbarked = data.groupby('HAIR').count()['name']
data.HAIR[data.HAIR.isnull()] = MaxPassEmbarked[MaxPassEmbarked == Ma
xPassEmbarked.median()].index[0]

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

```
c:\program files\python36\lib\site-packages\ipykernel_launcher.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
```

Out[211]: 815

Преобразование категориальных признаков в числовые

In [182]:
```
data.ALIGN.replace({'Good Characters':'1','Bad Characters':'0','Neutr
al Characters':'2'},inplace=True)
data.head(50)
```

Out[182]:

	page_id	name	urlslug	ID	ALIGN	EYE	
0	1678	Spider-Man (Peter Parker)	\/Spider-Man_(Peter_Parker)	Secret Identity	1	Hazel Eyes	
1	7139	Captain America (Steven Rogers)	\/Captain_America_(Steven_Rogers)	Public Identity	1	Blue Eyes	\
2	64786	Wolverine (James \"Logan\" Howlett)	\/Wolverine_(James_%22Logan%22_Howlett)	Public Identity	2	Blue Eyes	E
3	1868	Iron Man (Anthony \"Tony\" Stark)	\/Iron_Man_(Anthony_%22Tony%22_Stark)	Public Identity	1	Blue Eyes	E
4	2460	Thor (Thor Odinson)	\/Thor_(Thor_Odinson)	No Dual Identity	1	Blue Eyes	E
5	2458	Benjamin Grimm (Earth-616)	\/Benjamin_Grimm_(Earth-616)	Public Identity	1	Blue Eyes	
6	2166	Reed Richards (Earth-616)	\/Reed_Richards_(Earth-616)	Public Identity	1	Brown Eyes	
7	1833	Hulk (Robert Bruce Banner)	\/Hulk_(Robert_Bruce_Banner)	Public Identity	1	Brown Eyes	
8	29481	Scott Summers (Earth-616)	\/Scott_Summers_(Earth-616)	Public Identity	2	Brown Eyes	
9	1837	Jonathan	\/Jonathan_Storm_(Earth-616)	Public	1	Blue	E

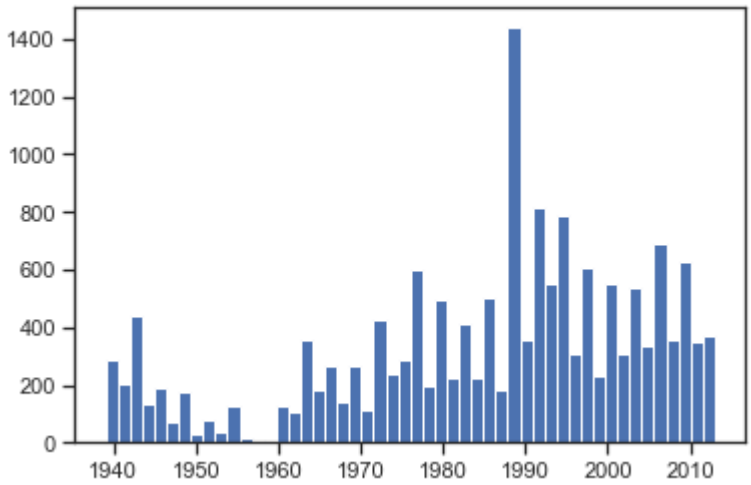
		Storm (Earth-616)		Identity		Eyes	
10	15725	Henry McCoy (Earth-616)	\/Henry_McCoy_(Earth-616)	Public Identity	1	Blue Eyes	
11	1863	Susan Storm (Earth-616)	\/Susan_Storm_(Earth-616)	Public Identity	1	Blue Eyes	B
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	2	Green Eyes	B
13	2614	Ororo Munroe (Earth-616)	\/Ororo_Munroe_(Earth-616)	Public Identity	1	Blue Eyes	W
14	1803	Clinton Barton (Earth-616)	\/Clinton_Barton_(Earth-616)	Public Identity	1	Blue Eyes	B
15	1396	Matthew Murdock (Earth-616)	\/Matthew_Murdock_(Earth-616)	Public Identity	1	Blue Eyes	
16	55534	Stephen Strange (Earth-616)	\/Stephen_Strange_(Earth-616)	Public Identity	1	Grey Eyes	B
17	1978	Mary Jane Watson (Earth-616)	\/Mary_Jane_Watson_(Earth-616)	No Dual Identity	1	Green Eyes	
18	1872	John Jonah Jameson (Earth-616)	\/John_Jonah_Jameson_(Earth-616)	No Dual Identity	2	Blue Eyes	B
19	35350	Robert Drake (Earth-616)	\/Robert_Drake_(Earth-616)	Secret Identity	1	Brown Eyes	
20	1557	Henry Pym (Earth-616)	\/Henry_Pym_(Earth-616)	Public Identity	1	Blue Eyes	B
21	65255	Charles Xavier (Earth-616)	\/Charles_Xavier_(Earth-616)	Public Identity	1	Blue Eyes	
22	1073	Warren Worthington III (Earth-616)	\/Warren_Worthington_III_(Earth-616)	Public Identity	1	Blue Eyes	B
23	1346	Piotr Rasputin (Earth-616)	\/Piotr_Rasputin_(Earth-616)	Secret Identity	1	Blue Eyes	B
24	2512	Wanda Maximoff (Earth-616)	\/Wanda_Maximoff_(Earth-616)	Public Identity	1	Green Eyes	
25	1671	Nicholas Fury (Earth-616)	\/Nicholas_Fury_(Earth-616)	No Dual Identity	2	Brown Eyes	
26	1976	Janet van Dyne (Earth-616)	\/Janet_van_Dyne_(Earth-616)	Public Identity	1	Blue Eyes	
27	65219	Jean Grey (Earth-616)	\/Jean_Grey_(Earth-616)	Public Identity	1	Green Eyes	
28	6545	Natalia Romanova (Earth-616)	\/Natalia_Romanova_(Earth-616)	Public Identity	1	Green Eyes	
29	2223	Kurt	\/Kurt_Wagner_(Earth-616)	Secret	1	Yellow	

		Wagner (Earth-616)		Identity		Eyes	
30	2414	Vision (Earth-616)	\/Vision_(Earth-616)	Secret Identity	1	Gold Eyes	
31	8650	May Reilly (Earth-616)	\/May_Reilly_(Earth-616)	No Dual Identity	1	Blue Eyes	
32	2527	Katherine Pryde (Earth-616)	\/Katherine_Pryde_(Earth-616)	Secret Identity	1	Hazel Eyes	
33	1970	Carol Danvers (Earth-616)	\/Carol_Danvers_(Earth-616)	Public Identity	1	Blue Eyes	E
34	37690	Jennifer Walters (Earth-616)	\/Jennifer_Walters_(Earth-616)	Public Identity	1	Green Eyes	
35	3765	Emma Frost (Earth-616)	\/Emma_Frost_(Earth-616)	Public Identity	2	Blue Eyes	
36	2312	Frank Castle (Earth-616)	\/Frank_Castle_(Earth-616)	Public Identity	2	Blue Eyes	E
37	1265	Luke Cage (Earth-616)	\/Luke_Cage_(Earth-616)	No Dual Identity	1	Brown Eyes	E
38	1677	Rogue (Anna Marie) (Earth-616)	\/Rogue_(Anna_Marie)_(Earth-616)	Secret Identity	1	Green Eyes	
39	25222	Conan (Earth-616)	\/Conan_(Earth-616)	No Dual Identity	2	Blue Eyes	E
40	6537	Joseph Robertson (Earth-616)	\/Joseph_Robertson_(Earth-616)	No Dual Identity	1	Brown Eyes	V
41	1592	Pietro Maximoff (Earth-616)	\/Pietro_Maximoff_(Earth-616)	Public Identity	1	Blue Eyes	S
42	1818	Hercules (Earth-616)	\/Hercules_(Earth-616)	No Dual Identity	1	Blue Eyes	
43	1448	Victor von Doom (Earth-616)	\/Victor_von_Doom_(Earth-616)	Public Identity	0	Brown Eyes	
44	2081	Max Eisenhardt (Earth-616)	\/Max_Eisenhardt_(Earth-616)	Public Identity	2	Grey Eyes	
45	2307	Elizabeth Braddock (Earth-616)	\/Elizabeth_Braddock_(Earth-616)	Secret Identity	2	Blue Eyes	
46	2556	Norrin Radd (Earth-616)	\/Norrin_Radd_(Earth-616)	Public Identity	1	Blue Eyes	
47	1740	Norman Osborn (Earth-616)	\/Norman_Osborn_(Earth-616)	Public Identity	0	Blue Eyes	
48	8651	Eugene Thompson (Earth-616)	\/Eugene_Thompson_(Earth-616)	Secret Identity	1	Blue Eyes	S E
49	2009	Simon [illegible]	\/Simon_Williams_(Earth-616)	Public [illegible]	1	Red [illegible]	

	Williams (Earth-616)	Identity	Eyes

In [58]:
```
from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normal
izer
```

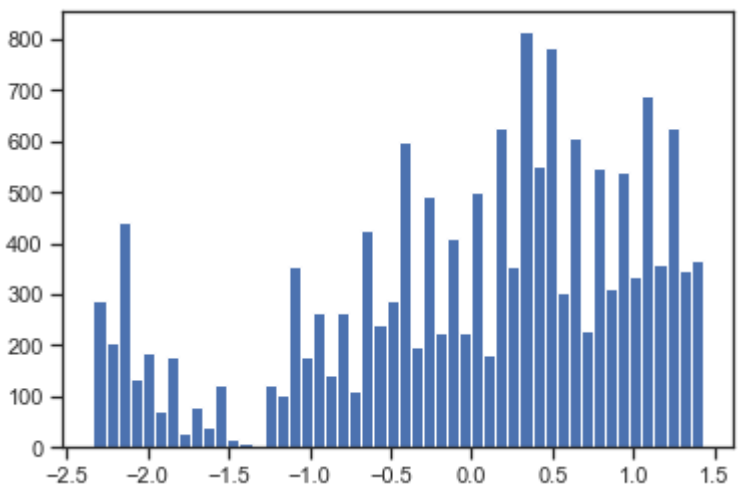
In [124]:
```
sc1 = MinMaxScaler()
sc1_data = sc1.fit_transform(data[['Year']])
plt.hist(data['Year'], 50)
plt.show()
```



Масштабирование данных на основе Z-оценки

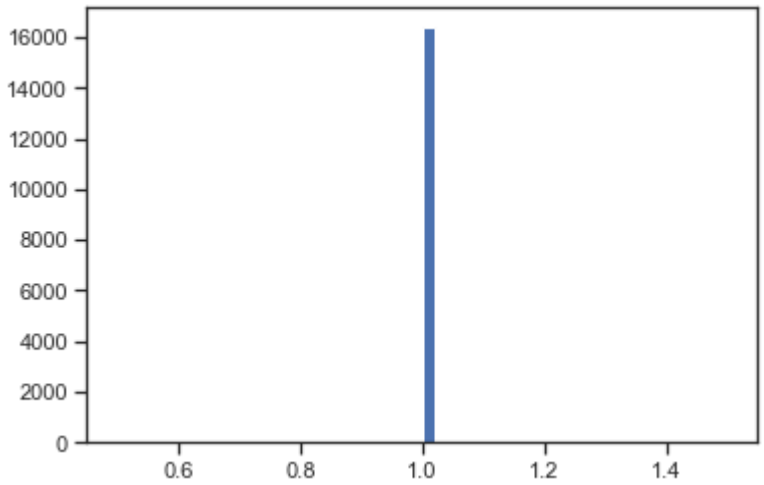
In [53]:
```
sc2 = StandardScaler()
sc2_data = sc2.fit_transform(data[['Year']])
plt.hist(sc2_data, 50)
plt.show()
```

```
c:\program files\python36\lib\site-packages\numpy\lib\histograms.py:
824: RuntimeWarning: invalid value encountered in greater_equal
  keep = (tmp_a >= first_edge)
c:\program files\python36\lib\site-packages\numpy\lib\histograms.py:
825: RuntimeWarning: invalid value encountered in less_equal
  keep &= (tmp_a <= last_edge)
```



In [120]:
```
sc3 = Normalizer()
sc3_data = sc3.fit_transform(data[['Year']])
```

In [121]:
```
plt.hist(sc3_data, 50)
plt.show()
```



In [213]:
```
data.head(17000)
```

Out[213]:

	page_id	name	urlslug	ID	ALIG
0	1678	Spider-Man (Peter Parker)	\/Spider-Man_(Peter_Parker)	Secret Identity	Goo Character
1	7139	Captain America (Steven Rogers)	\/Captain_America_(Steven_Rogers)	Public Identity	Goo Character
2	64786	Wolverine (James \"Logan\" Howlett)	\/Wolverine_(James_%22Logan%22_Howlett)	Public Identity	Neutr Character
3	1868	Iron Man (Anthony \"Tony\" Stark)	\/Iron_Man_(Anthony_%22Tony%22_Stark)	Public Identity	Goo Character
4	2460	Thor (Thor Odinson)	\/Thor_(Thor_Odinson)	No Dual Identity	Goo Character
5	2458	Benjamin Grimm (Earth-616)	\/Benjamin_Grimm_(Earth-616)	Public Identity	Goo Character
6	2166	Reed Richards (Earth-616)	\/Reed_Richards_(Earth-616)	Public Identity	Goo Character
7	1833	Hulk (Robert Bruce Banner)	\/Hulk_(Robert_Bruce_Banner)	Public Identity	Goo Character
8	29481	Scott Summers (Earth-616)	\/Scott_Summers_(Earth-616)	Public Identity	Neutr Character
9	1837	Jonathan Storm (Earth-616)	\/Jonathan_Storm_(Earth-616)	Public Identity	Goo Character
10	15725	Henry McCoy (Earth-616)	\/Henry_McCoy_(Earth-616)	Public Identity	Goo Character
11	1863	Susan Storm	\/Susan_Storm_(Earth-616)	Public	Goo

		(Earth-616)		Identity	Character
12	7823	Namor McKenzie (Earth-616)	\/Namor_McKenzie_(Earth-616)	No Dual Identity	Neutra Character
13	2614	Ororo Munroe (Earth-616)	\/Ororo_Munroe_(Earth-616)	Public Identity	Goo Character
14	1803	Clinton Barton (Earth-616)	\/Clinton_Barton_(Earth-616)	Public Identity	Goo Character
15	1396	Matthew Murdock (Earth-616)	\/Matthew_Murdock_(Earth-616)	Public Identity	Goo Character
16	55534	Stephen Strange (Earth-616)	\/Stephen_Strange_(Earth-616)	Public Identity	Goo Character
17	1978	Mary Jane Watson (Earth-616)	\/Mary_Jane_Watson_(Earth-616)	No Dual Identity	Goo Character
18	1872	John Jonah Jameson (Earth-616)	\/John_Jonah_Jameson_(Earth-616)	No Dual Identity	Neutra Character
19	35350	Robert Drake (Earth-616)	\/Robert_Drake_(Earth-616)	Secret Identity	Goo Character
20	1557	Henry Pym (Earth-616)	\/Henry_Pym_(Earth-616)	Public Identity	Goo Character
21	65255	Charles Xavier (Earth-616)	\/Charles_Xavier_(Earth-616)	Public Identity	Goo Character
22	1073	Warren Worthington III (Earth-616)	\/Warren_Worthington_III_(Earth-616)	Public Identity	Goo Character
23	1346	Piotr Rasputin (Earth-616)	\/Piotr_Rasputin_(Earth-616)	Secret Identity	Goo Character
24	2512	Wanda Maximoff (Earth-616)	\/Wanda_Maximoff_(Earth-616)	Public Identity	Goo Character
25	1671	Nicholas Fury (Earth-616)	\/Nicholas_Fury_(Earth-616)	No Dual Identity	Neutra Character
26	1976	Janet van Dyne (Earth-616)	\/Janet_van_Dyne_(Earth-616)	Public Identity	Goo Character
27	65219	Jean Grey (Earth-616)	\/Jean_Grey_(Earth-616)	Public Identity	Goo Character
28	6545	Natalia Romanova (Earth-616)	\/Natalia_Romanova_(Earth-616)	Public Identity	Goo Character
29	2223	Kurt Wagner (Earth-616)	\/Kurt_Wagner_(Earth-616)	Secret Identity	Goo Character
...	...	...	...	...	.
16346	42106	Thiazzi (Earth-616)	\/Thiazzi_(Earth-616)	NaN	Na
16347	462671	TOR\/test	\/User:TOR\/test	NaN	Na

16348	725312	Toxin (Luminals) (Earth-616)	\/Toxin_(Luminals)_(Earth-616)	Secret Identity	Na
16349	507239	Urizen Ul'var (Earth-616)	\/Urizen_Ul%27var_(Earth-616)	Public Identity	Goo Character
16350	40061	Valka (Earth-616)	\/Valka_(Earth-616)	NaN	Na
16351	505759	Vayor (Earth-616)	\/Vayor_(Earth-616)	No Dual Identity	Ba Character
16352	16569	Viridian (Earth-616)	\/Viridian_(Earth-616)	NaN	Na
16353	16296	William Burke (Earth-616)	\/William_Burke_(Earth-616)	NaN	Na
16354	120833	William Falsworth (Earth-616)	\/William_Falsworth_(Earth-616)	NaN	Goo Character
16355	693077	William Shakespeare (Earth-616)	\/William_Shakespeare_(Earth-616)	Public Identity	Neutra Character
16356	473710	Yamata no Orichi (Earth-616)	\/Yamata_no_Orichi_(Earth-616)	Public Identity	Ba Character
16357	670512	Zero G. Priestly (Legion Personality) (Earth-616)	\/Zero_G._Priestly_(Legion_Personality)_(Earth...	Secret Identity	Neutra Character
16358	12132	Zora Loftus (Earth-616)	\/Zora_Loftus_(Earth-616)	Secret Identity	Ba Character
16359	727035	Agar (Earth-616)	\/Agar_(Earth-616)	No Dual Identity	Ba Character
16360	703546	Ana (Natasha Romanoff's neighbor) (Earth-616)	\/Ana_(Natasha_Romanoff%27s_neighbor)_(Earth-616)	No Dual Identity	Goo Character
16361	713903	Dante's mother (Earth-616)	\/Dante%27s_mother_(Earth-616)	No Dual Identity	Neutra Character
16362	508693	Farbauti (Earth-616)	\/Farbauti_(Earth-616)	Public Identity	Ba Character
16363	694577	Finch (Kate Bishop's neighbor) (Earth-616)	\/Finch_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character
16364	655292	Jack O'Lantern (Impostor) (Earth-616)	\/Jack_O%27Lantern_(Impostor)_(Earth-616)	Secret Identity	Ba Character
16365	684262	K'thol (Earth-616)	\/K%27thol_(Earth-616)	NaN	Goo Character
16366	643435	Karen (Hijack's girlfriend) (Earth-616)	\/Karen_(Hijack%27s_girlfriend)_(Earth-616)	Public Identity	Neutra Character
16367	694583	Marcus (Kate	\/Marcus_(Kate_Bishop%27s_neighbor)_(Earth-616)	Public Identity	Goo Character

		Bishop's neighbor) (Earth-616)			
16368	703892	Marcy (Offer's employee) (Earth-616)	\/Marcy_(Offer%27s_employee)_(Earth-616)	Public Identity	Neutra Character
16369	660799	Melanie Kapoor (Earth-616)	\/Melanie_Kapoor_(Earth-616)	Public Identity	Goo Character
16370	674414	Phoenix's Shadow (Earth-616)	\/Phoenix%27s_Shadow_(Earth-616)	NaN	Neutra Character
16371	657508	Ru'ach (Earth-616)	\/Ru%27ach_(Earth-616)	No Dual Identity	Ba Character
16372	665474	Thane (Thanos' son) (Earth-616)	\/Thane_(Thanos%27_son)_(Earth-616)	No Dual Identity	Goo Character
16373	695217	Tinkerer (Skrull) (Earth-616)	\/Tinkerer_(Skrull)_(Earth-616)	Secret Identity	Ba Character
16374	708811	TK421 (Spiderling) (Earth-616)	\/TK421_(Spiderling)_(Earth-616)	Secret Identity	Neutra Character
16375	673702	Yologarch (Earth-616)	\/Yologarch_(Earth-616)	NaN	Ba Character

16376 rows × 13 columns

In [ ]: