

Лабораторная работа №3  
по дисциплине  
«Методы машинного обучения»  
на тему  
«Обработка пропусков, кодирование категориальных  
признаков, масштабирование данных»

Выполнил:  
студент группы РТ5-61Б  
Корякин Д.

---

# 1. Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.

Мы научимся обрабатывать пропуски в данных для количественных (числовых) и категориальных признаков и масштабировать данные. Также мы научимся преобразовывать категориальные признаки в числовые.

```
[1]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
pd.set_option("display.width", 70)
```

## 1.1. Загрузка и первичный анализ данных

Используем данные из соревнования House Prices: Advanced Regression Techniques

```
[2]: # Будем использовать только обучающую выборку
data = pd.read_csv('data/gun_violence.csv', sep=",")
```

```
[3]: # размер набора данных
data.shape
```

```
[3]: (162867, 29)
```

```
[4]: # ТИПЫ КОЛОНОК
data.dtypes
```

```
[4]: incident_id      int64
date                object
state               object
city_or_county      object
address             object
n_killed            int64
n_injured           int64
incident_url        object
source_url          object
incident_url_fields_missing  bool
congressional_district  float64
gun_stolen          object
gun_type            object
incident_characteristics  object
latitude            float64
location_description  object
longitude           float64
n_guns_involved     float64
notes               object
participant_age      object
participant_age_group  object
```

```

participant_gender      object
participant_name        object
participant_relationship object
participant_status      object
participant_type        object
sources                object
state_house_district    float64
state_senate_district   float64
dtype: object

```

```
[5]: # проверим есть ли пропущенные значения
data.isnull().sum()
```

```

[5]: incident_id      0
date                 0
state                0
city_or_county       0
address              12303
n_killed             0
n_injured            0
incident_url         0
source_url           276
incident_url_fields_missing 0
congressional_district 4865
gun_stolen           99311
gun_type             99299
incident_characteristics 242
latitude             4715
location_description 140476
longitude            4715
n_guns_involved      99299
notes                56008
participant_age       63464
participant_age_group 27678
participant_gender    23832
participant_name      84207
participant_relationship 152618
participant_status    18510
participant_type      16327
sources               516
state_house_district  24163
state_senate_district 20659
dtype: int64

```

```
[6]: # Первые 5 строк датасета
data.head()
```

```

[6]:   incident_id      date      state city_or_county \
0      461105  2013-01-01  Pennsylvania  Mckeesport
1      460726  2013-01-01   California  Hawthorne

```

2	478855	2013-01-01	Ohio	Lorain
3	478925	2013-01-05	Colorado	Aurora
4	478959	2013-01-07	North Carolina	Greensboro

	address	n_killed	n_injured
0	1506 Versailles Avenue and Coursin Street	0	4
1	13500 block of Cerise Avenue	1	3
2	1776 East 28th Street	1	3
3	16000 block of East Ithaca Place	4	0
4	307 Mourning Dove Terrace	2	2

	incident_url
0	http://www.gunviolencearchive.org/incident/461105
1	http://www.gunviolencearchive.org/incident/460726
2	http://www.gunviolencearchive.org/incident/478855
3	http://www.gunviolencearchive.org/incident/478925
4	http://www.gunviolencearchive.org/incident/478959

	source_url
0	http://www.post-gazette.com/local/south/2013/0...
1	http://www.dailybulletin.com/article/zz/201301...
2	http://chronicle.northcoastnow.com/2013/02/14/...
3	http://www.dailydemocrat.com/20130106/aurora-s...
4	http://www.journalnow.com/news/local/article_d...

	incident_url_fields_missing
0	False
1	False
2	False
3	False
4	False

	participant_age
0	0::20
1	0::20
2	0::25 1::31 2::33 3::34 4::33
3	0::29 1::33 2::56 3::33
4	0::18 1::46 2::14 3::47

	participant_age_group
0	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
1	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
2	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
3	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
4	0::Adult 18+ 1::Adult 18+ 2::Teen 12-17 3::...

	participant_gender
0	0::Male 1::Male 3::Male 4::Female

1		0::Male
2	0::Male  1::Male  2::Male  3::Male  4::Male	
3	0::Female  1::Male  2::Male  3::Male	
4	0::Female  1::Male  2::Male  3::Female	

	participant_name \
0	0::Julian Sims
1	0::Bernard Gillis
2	0::Damien Bell  1::Desmen Noble  2::Herman Sea...
3	0::Stacie Philbrook  1::Christopher Ratliffe  ...
4	0::Danielle Imani Jameison  1::Maurice Eugene ...

	participant_relationship \
0	NaN
1	NaN
2	NaN
3	NaN
4	3::Family

	participant_status \
0	0::Arrested  1::Injured  2::Injured  3::Injure...
1	0::Killed  1::Injured  2::Injured  3::Injured
2	0::Injured, Unharmed, Arrested  1::Unharmed, A...
3	0::Killed  1::Killed  2::Killed  3::Killed
4	0::Injured  1::Injured  2::Killed  3::Killed

	participant_type \
0	0::Victim  1::Victim  2::Victim  3::Victim  4::...
1	0::Victim  1::Victim  2::Victim  3::Victim  4::...
2	0::Subject-Suspect  1::Subject-Suspect  2::Vic...
3	0::Victim  1::Victim  2::Victim  3::Subject-Su...
4	0::Victim  1::Victim  2::Victim  3::Subject-Su...

	sources \
0	<a href="http://pittsburgh.cbslocal.com/2013/01/01/4-pe...">http://pittsburgh.cbslocal.com/2013/01/01/4-pe...</a>
1	<a href="http://losangeles.cbslocal.com/2013/01/01/man-...">http://losangeles.cbslocal.com/2013/01/01/man-...</a>
2	<a href="http://www.morningjournal.com/general-news/201...">http://www.morningjournal.com/general-news/201...</a>
3	<a href="http://denver.cbslocal.com/2013/01/06/officer-...">http://denver.cbslocal.com/2013/01/06/officer-...</a>
4	<a href="http://myfox8.com/2013/01/08/update-mother-sho...">http://myfox8.com/2013/01/08/update-mother-sho...</a>

	state_house_district	state_senate_district
0	NaN	NaN
1	62.0	35.0
2	56.0	13.0
3	40.0	28.0
4	62.0	27.0

[5 rows x 29 columns]

```
[7]: total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

Всего строк: 162867

## 2. 1. Обработка пропусков в данных

### 2.1. 1.1. Простые стратегии - удаление или заполнение нулями

```
[8]: # Удаление колонок, содержащих пустые значения
data_new_1 = data.dropna(axis=1, how='any')
(data.shape, data_new_1.shape)
```

```
[8]: ((162867, 29), (162867, 8))
```

```
[9]: # Удаление строк, содержащих пустые значения
data_new_2 = data.dropna(axis=0, how='any')
(data.shape, data_new_2.shape)
```

```
[9]: ((162867, 29), (450, 29))
```

```
[10]: data.head()
```

```
[10]:
```

	incident_id	date	state	city_or_county \			
0	461105	2013-01-01	Pennsylvania	McKeesport			
1	460726	2013-01-01	California	Hawthorne			
2	478855	2013-01-01	Ohio	Lorain			
3	478925	2013-01-05	Colorado	Aurora			
4	478959	2013-01-07	North Carolina	Greensboro			

		address	n_killed	n_injured
→ \	0	1506 Versailles Avenue and Coursin Street	0	4
	1	13500 block of Cerise Avenue	1	3
	2	1776 East 28th Street	1	3
	3	16000 block of East Ithaca Place	4	0
	4	307 Mourning Dove Terrace	2	2

	incident_url \
0	http://www.gunviolencearchive.org/incident/461105
1	http://www.gunviolencearchive.org/incident/460726
2	http://www.gunviolencearchive.org/incident/478855
3	http://www.gunviolencearchive.org/incident/478925
4	http://www.gunviolencearchive.org/incident/478959

	source_url \
0	http://www.post-gazette.com/local/south/2013/0...
1	http://www.dailybulletin.com/article/zz/201301...
2	http://chronicle.northcoastnow.com/2013/02/14/...
3	http://www.dailydemocrat.com/20130106/aurora-s...

```

4 http://www.journalnow.com/news/local/article_d...

incident_url_fields_missing ... \
0 False ...
1 False ...
2 False ...
3 False ...
4 False ...

participant_age \
0 0::20
1 0::20
2 0::25||1::31||2::33||3::34||4::33
3 0::29||1::33||2::56||3::33
4 0::18||1::46||2::14||3::47

participant_age_group \
0 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
1 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
2 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
3 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
4 0::Adult 18+||1::Adult 18+||2::Teen 12-17||3::...

participant_gender \
0 0::Male||1::Male||3::Male||4::Female
1 0::Male
2 0::Male||1::Male||2::Male||3::Male||4::Male
3 0::Female||1::Male||2::Male||3::Male
4 0::Female||1::Male||2::Male||3::Female

participant_name \
0 0::Julian Sims
1 0::Bernard Gillis
2 0::Damien Bell||1::Desmen Noble||2::Herman Sea...
3 0::Stacie Philbrook||1::Christopher Ratliffe||...
4 0::Danielle Imani Jameison||1::Maurice Eugene ...

participant_relationship \
0 NaN
1 NaN
2 NaN
3 NaN
4 3::Family

participant_status \
0 0::Arrested||1::Injured||2::Injured||3::Injure...
1 0::Killed||1::Injured||2::Injured||3::Injured
2 0::Injured, Unharmd, Arrested||1::Unharmd, A...
3 0::Killed||1::Killed||2::Killed||3::Killed
4 0::Injured||1::Injured||2::Killed||3::Killed

```

```

                                participant_type \
0  0::Victim|1::Victim|2::Victim|3::Victim|4:...
1  0::Victim|1::Victim|2::Victim|3::Victim|4:...
2  0::Subject-Suspect|1::Subject-Suspect|2::Vic...
3  0::Victim|1::Victim|2::Victim|3::Subject-Su...
4  0::Victim|1::Victim|2::Victim|3::Subject-Su...

```

```

                                sources \
0  http://pittsburgh.cbslocal.com/2013/01/01/4-pe...
1  http://losangeles.cbslocal.com/2013/01/01/man-...
2  http://www.morningjournal.com/general-news/201...
3  http://denver.cbslocal.com/2013/01/06/officer-...
4  http://myfox8.com/2013/01/08/update-mother-sho...

```

```

state_house_district state_senate_district
0                    NaN                    NaN
1                    62.0                    35.0
2                    56.0                    13.0
3                    40.0                    28.0
4                    62.0                    27.0

```

[5 rows x 29 columns]

```

[11]: # Заполнение всех пропущенных значений нулями
      # В данном случае это некорректно, так как нулями заполняются в
      → том числе категориальные колонки
      data_new_3 = data.fillna(0)
      data_new_3.head()

```

```

[11]: incident_id      date      state city_or_county \
0      461105  2013-01-01  Pennsylvania  Mckeesport
1      460726  2013-01-01  California    Hawthorne
2      478855  2013-01-01  Ohio          Lorain
3      478925  2013-01-05  Colorado      Aurora
4      478959  2013-01-07  North Carolina  Greensboro

                                address  n_killed  n_injured
→ \
0  1506 Versailles Avenue and Coursin Street      0      4
1              13500 block of Cerise Avenue      1      3
2              1776 East 28th Street      1      3
3      16000 block of East Ithaca Place      4      0
4              307 Mourning Dove Terrace      2      2

                                incident_url \
0  http://www.gunviolencearchive.org/incident/461105
1  http://www.gunviolencearchive.org/incident/460726
2  http://www.gunviolencearchive.org/incident/478855
3  http://www.gunviolencearchive.org/incident/478925

```



```

4 http://www.gunviolencearchive.org/incident/478959

                                source_url \
0 http://www.post-gazette.com/local/south/2013/0...
1 http://www.dailybulletin.com/article/zz/201301...
2 http://chronicle.northcoastnow.com/2013/02/14/...
3 http://www.dailydemocrat.com/20130106/aurora-s...
4 http://www.journalnow.com/news/local/article_d...

incident_url_fields_missing ... \
0 False ...
1 False ...
2 False ...
3 False ...
4 False ...

                                participant_age \
0 0::20
1 0::20
2 0::25||1::31||2::33||3::34||4::33
3 0::29||1::33||2::56||3::33
4 0::18||1::46||2::14||3::47

                                participant_age_group \
0 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
1 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
2 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
3 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...
4 0::Adult 18+||1::Adult 18+||2::Teen 12-17||3::...

                                participant_gender \
0 0::Male||1::Male||3::Male||4::Female
1 0::Male
2 0::Male||1::Male||2::Male||3::Male||4::Male
3 0::Female||1::Male||2::Male||3::Male
4 0::Female||1::Male||2::Male||3::Female

                                participant_name \
0 0::Julian Sims
1 0::Bernard Gillis
2 0::Damien Bell||1::Desmen Noble||2::Herman Sea...
3 0::Stacie Philbrook||1::Christopher Ratliffe||...
4 0::Danielle Imani Jameison||1::Maurice Eugene ...

participant_relationship \
0 0
1 0
2 0
3 0
4 3::Family

```

```

                                participant_status \
0  0::Arrested||1::Injured||2::Injured||3::Injure...
1      0::Killed||1::Injured||2::Injured||3::Injured
2  0::Injured, Unharmed, Arrested||1::Unharmed, A...
3      0::Killed||1::Killed||2::Killed||3::Killed
4      0::Injured||1::Injured||2::Killed||3::Killed

                                participant_type \
0  0::Victim||1::Victim||2::Victim||3::Victim||4:...
1  0::Victim||1::Victim||2::Victim||3::Victim||4:...
2  0::Subject-Suspect||1::Subject-Suspect||2::Vic...
3  0::Victim||1::Victim||2::Victim||3::Subject-Su...
4  0::Victim||1::Victim||2::Victim||3::Subject-Su...

                                sources \
0  http://pittsburgh.cbslocal.com/2013/01/01/4-pe...
1  http://losangeles.cbslocal.com/2013/01/01/man-...
2  http://www.morningjournal.com/general-news/201...
3  http://denver.cbslocal.com/2013/01/06/officer-...
4  http://myfox8.com/2013/01/08/update-mother-sho...

state_house_district state_senate_district
0                    0.0                    0.0
1                    62.0                    35.0
2                    56.0                    13.0
3                    40.0                    28.0
4                    62.0                    27.0

[5 rows x 29 columns]

```

## 2.2. 1.2. “Внедрение значений” - импьютация (imputation)

### 2.2.1. 1.2.1. Обработка пропусков в числовых данных

```

[12]: # 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)
              temp_perc = round((temp_null_count / total_count) * 100.
→0, 2)
              print('Колонка {}. Тип данных {}. Количество пустых
→значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))

```

Колонка congressional\_district. Тип данных float64. Количество  
 ↳пустых значений  
 4865, 2.99%.  
 Колонка latitude. Тип данных float64. Количество пустых значений  
 ↳4715, 2.9%.  
 Колонка longitude. Тип данных float64. Количество пустых значений  
 ↳4715, 2.9%.  
 Колонка n\_guns\_involved. Тип данных float64. Количество пустых  
 ↳значений 99299,  
 60.97%.  
 Колонка state\_house\_district. Тип данных float64. Количество  
 ↳пустых значений  
 24163, 14.84%.  
 Колонка state\_senate\_district. Тип данных float64. Количество  
 ↳пустых значений  
 20659, 12.68%.

```
[13]: # Фильтр по колонкам с пропущенными значениями
data_num = data[num_cols]
data_num
```

```
[13]:
```

	congressional_district	latitude	longitude	\
0	14.0	40.3467	-79.8559	
1	43.0	33.9090	-118.3330	
2	9.0	41.4455	-82.1377	
3	6.0	39.6518	-104.8020	
4	6.0	36.1140	-79.9569	
...	...	...	...	
162862	13.0	33.7938	-84.5894	
162863	13.0	37.7338	-122.1790	
162864	NaN	NaN	NaN	
162865	1.0	34.2190	-88.7378	
162866	8.0	35.0708	-89.6713	

	n_guns_involved	state_house_district	state_senate_district
0	NaN	NaN	
↳NaN			
1	NaN	62.0	
↳35.0			
2	2.0	56.0	
↳13.0			
3	NaN	40.0	
↳28.0			
4	2.0	62.0	
↳27.0			
...	...	...	...
162862	1.0	39.0	
↳38.0			

162863	1.0	18.0	□
→ 9.0			
162864	1.0	NaN	□
→ NaN			
162865	1.0	16.0	□
→ 7.0			
162866	1.0	95.0	□
→ 32.0			

[162867 rows x 6 columns]

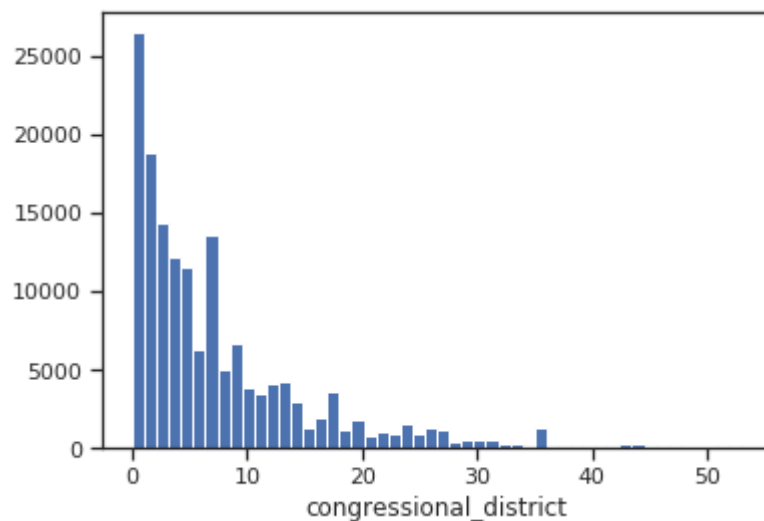
```
[14]: # Гистограмма по признакам
for col in data_num:
    plt.hist(data[col], 50)
    plt.xlabel(col)
    plt.show()
```

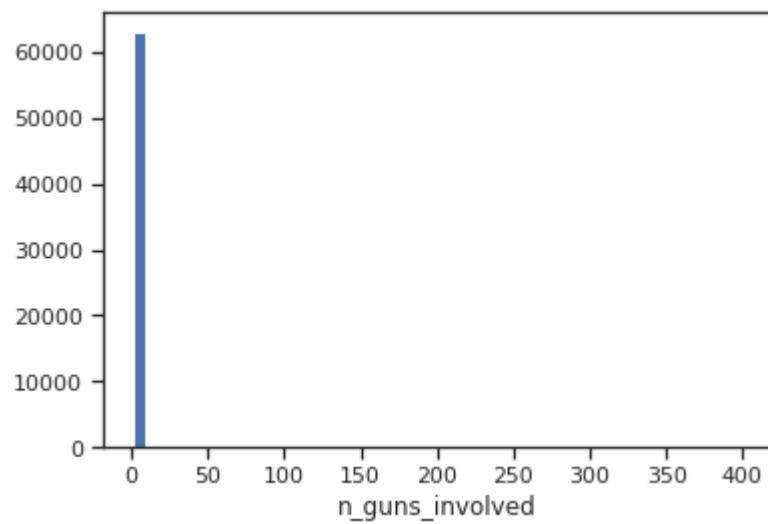
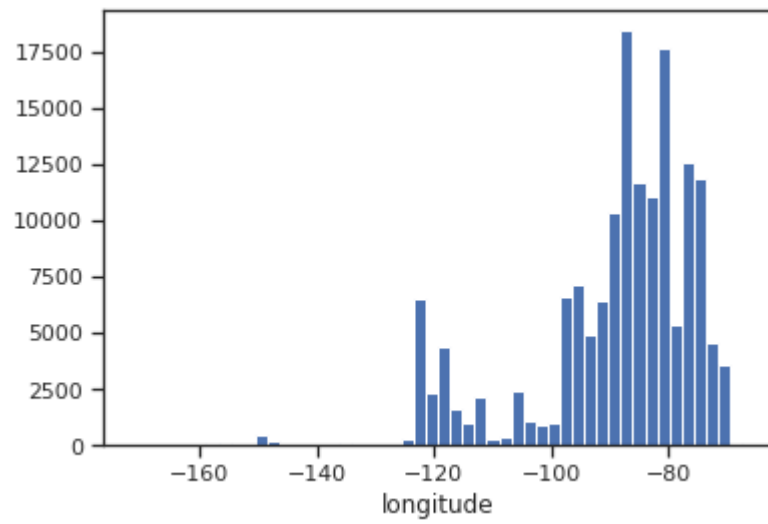
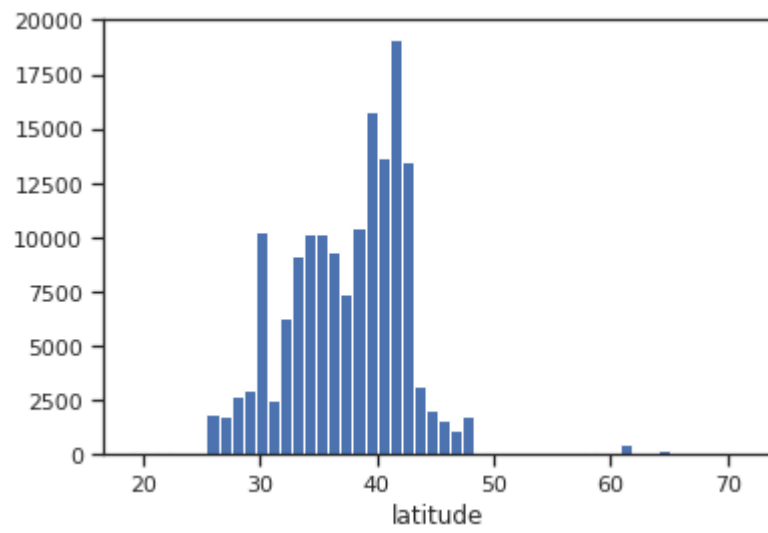
/home/dan/anaconda3/lib/python3.7/site-packages/numpy/lib/  
→ histograms.py:839:

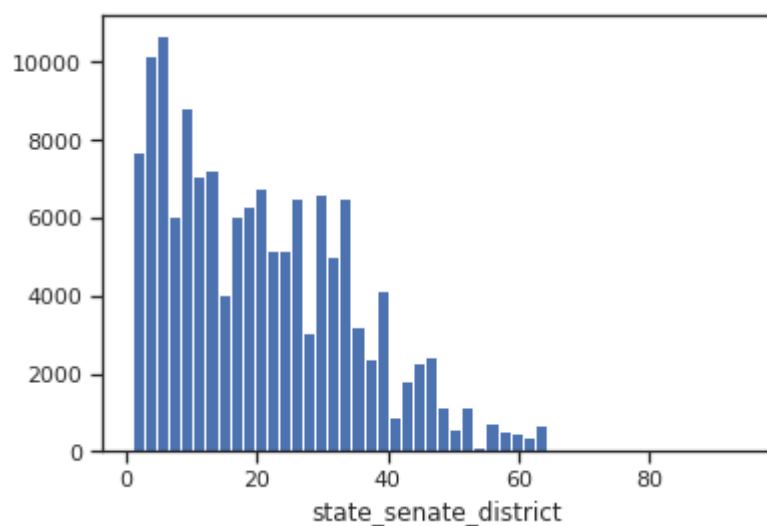
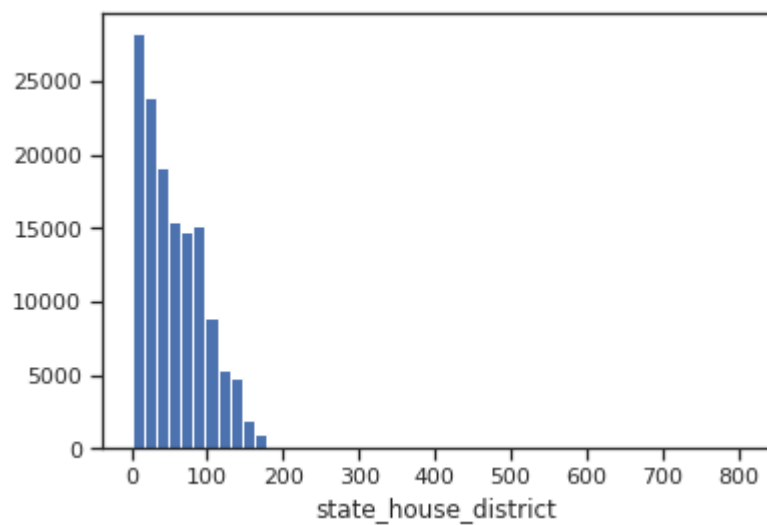
RuntimeWarning: invalid value encountered in greater\_equal  
keep = (tmp\_a >= first\_edge)

/home/dan/anaconda3/lib/python3.7/site-packages/numpy/lib/  
→ histograms.py:840:

RuntimeWarning: invalid value encountered in less\_equal  
keep &= (tmp\_a <= last\_edge)







```
[15]: # Фильтр по пустым значениям поля n_guns_involved
data[data['n_guns_involved'].isnull()]
```

```
[15]:
```

	incident_id	date	state	city_or_county \
0	461105	2013-01-01	Pennsylvania	McKeesport
1	460726	2013-01-01	California	Hawthorne
3	478925	2013-01-05	Colorado	Aurora
5	478948	2013-01-07	Oklahoma	Tulsa
7	479374	2013-01-21	Louisiana	New Orleans
...	...	...	...	...
161927	729430	2016-12-16	Wisconsin	Madison
162166	730843	2016-12-18	Florida	Naples
162373	729453	2016-12-19	California	Brawley

162740	730974	2016-12-21	Arkansas	Fayetteville
162801	732054	2016-12-22	Florida	Palm Harbor

address □

↪n_killed \	
0	1506 Versailles Avenue and Coursin Street □
↪0	
1	13500 block of Cerise Avenue □
↪1	
3	16000 block of East Ithaca Place □
↪4	
5	6000 block of South Owasso □
↪4	
7	LaSalle Street and Martin Luther King Jr. Boul... 0
...	...
161927	Fourth Street □
↪0	
162166	Pine Ridge Rd and Airport Pulling Road □
↪0	
162373	500 block of North Imperial □
↪1	
162740	800 South School Avenue □
↪1	
162801	252 Whisper Lake Road □
↪1	

n_injured \	
0	4
1	3
3	0
5	0
7	5
...	...
161927	0
162166	0
162373	0
162740	0
162801	0

incident_url \	
0	<a href="http://www.gunviolencearchive.org/incident/461105">http://www.gunviolencearchive.org/incident/461105</a>
1	<a href="http://www.gunviolencearchive.org/incident/460726">http://www.gunviolencearchive.org/incident/460726</a>
3	<a href="http://www.gunviolencearchive.org/incident/478925">http://www.gunviolencearchive.org/incident/478925</a>
5	<a href="http://www.gunviolencearchive.org/incident/478948">http://www.gunviolencearchive.org/incident/478948</a>
7	<a href="http://www.gunviolencearchive.org/incident/479374">http://www.gunviolencearchive.org/incident/479374</a>
...	...
161927	<a href="http://www.gunviolencearchive.org/incident/729430">http://www.gunviolencearchive.org/incident/729430</a>
162166	<a href="http://www.gunviolencearchive.org/incident/730843">http://www.gunviolencearchive.org/incident/730843</a>
162373	<a href="http://www.gunviolencearchive.org/incident/729453">http://www.gunviolencearchive.org/incident/729453</a>

162740 <http://www.gunviolencearchive.org/incident/730974>  
 162801 <http://www.gunviolencearchive.org/incident/732054>

source\_url \

0	<a href="http://www.post-gazette.com/local/south/2013/0...">http://www.post-gazette.com/local/south/2013/0...</a>
1	<a href="http://www.dailybulletin.com/article/zz/201301...">http://www.dailybulletin.com/article/zz/201301...</a>
3	<a href="http://www.dailydemocrat.com/20130106/aurora-s...">http://www.dailydemocrat.com/20130106/aurora-s...</a>
5	<a href="http://usnews.nbcnews.com/_news/2013/01/07/163...">http://usnews.nbcnews.com/_news/2013/01/07/163...</a>
7	<a href="http://www.nola.com/crime/index.ssf/2013/01/no...">http://www.nola.com/crime/index.ssf/2013/01/no...</a>

...

161927	<a href="http://www.nbc15.com/content/news/2-teens-arre...">http://www.nbc15.com/content/news/2-teens-arre...</a>
162166	<a href="http://www.naplesnews.com/story/news/crime/201...">http://www.naplesnews.com/story/news/crime/201...</a>
162373	<a href="http://www.kyma.com/news/fatal-officer-involve...">http://www.kyma.com/news/fatal-officer-involve...</a>
162740	<a href="http://www.4029tv.com/article/officer-involved...">http://www.4029tv.com/article/officer-involved...</a>
162801	<a href="http://web.tampabay.com/news/publicsafety/crim...">http://web.tampabay.com/news/publicsafety/crim...</a>

incident\_url\_fields\_missing ... \

0	False	...
1	False	...
3	False	...
5	False	...
7	False	...

...

161927	False	...
162166	False	...
162373	False	...
162740	False	...
162801	False	...

participant\_age \

0	0::20
1	0::20
3	0::29 1::33 2::56 3::33
5	0::23 1::23 2::33 3::55
7	NaN

...

161927	0::18 1::18
162166	0::24
162373	NaN
162740	0::25
162801	0::55

participant\_age\_group \

0	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
1	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
3	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
5	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...
7	NaN

...

161927	0::Adult 18+ 1::Adult 18+
--------	---------------------------



162166	0::Adult 18+
162373	0::Adult 18+
162740	0::Adult 18+
162801	0::Adult 18+

	participant_gender \
0	0::Male 1::Male 3::Male 4::Female
1	0::Male
3	0::Female 1::Male 2::Male 3::Male
5	0::Female 1::Female 2::Female 3::Female 4::...
7	0::Male 1::Male 2::Male 3::Male 4::Male
...	...
161927	0::Male 1::Male
162166	0::Male
162373	0::Male
162740	0::Male
162801	0::Male

	participant_name \
0	0::Julian Sims
1	0::Bernard Gillis
3	0::Stacie Philbrook 1::Christopher Ratliffe ...
5	0::Rebeika Powell 1::Kayetie Melchor 2::Mist...
7	NaN
...	...
161927	0::Taylor Loving 1::Theron Walker
162166	0::Sean Blackwell
162373	NaN
162740	0::Benjamin Ortiz
162801	0::Stanley Eversole

	participant_relationship \
0	NaN
1	NaN
3	NaN
5	NaN
7	NaN
...	...
161927	NaN
162166	NaN
162373	NaN
162740	NaN
162801	NaN

	participant_status \
0	0::Arrested 1::Injured 2::Injured 3::Injure...
1	0::Killed 1::Injured 2::Injured 3::Injured
3	0::Killed 1::Killed 2::Killed 3::Killed
5	0::Killed 1::Killed 2::Killed 3::Killed 4::...
7	0::Injured 1::Injured 2::Injured 3::Injured...

```

...
161927      0::Unharmed, Arrested||1::Unharmed, Arrested
162166      0::Unharmed, Arrested
162373      0::Killed
162740      0::Killed
162801      0::Killed

...
participant_type \
0      0::Victim|1::Victim|2::Victim|3::Victim|4:...
1      0::Victim|1::Victim|2::Victim|3::Victim|4:...
3      0::Victim|1::Victim|2::Victim|3::Subject-Su...
5      0::Victim|1::Victim|2::Victim|3::Victim|4:...
7      0::Victim|1::Victim|2::Victim|3::Victim|4:...

...
161927      0::Subject-Suspect||1::Subject-Suspect
162166      0::Subject-Suspect
162373      0::Subject-Suspect
162740      0::Subject-Suspect
162801      0::Subject-Suspect

...
sources \
0      http://pittsburgh.cbslocal.com/2013/01/01/4-pe...
1      http://losangeles.cbslocal.com/2013/01/01/man-...
3      http://denver.cbslocal.com/2013/01/06/officer-...
5      http://www.kjrh.com/news/local-news/4-found-sh...
7      http://www.huffingtonpost.com/2013/01/21/new-o...

...
161927      http://www.nbc15.com/content/news/2-teens-arre...
162166      http://www.naplesnews.com/story/news/crime/201...
162373      http://www.kyma.com/news/fatal-officer-involve...
162740      http://www.4029tv.com/article/officer-involved...
162801      http://www.nbcmiami.com/news/local/Deputies-Sh...

state_house_district state_senate_district
0      NaN      NaN
1      62.0      35.0
3      40.0      28.0
5      72.0      11.0
7      93.0      5.0

...
161927      76.0      26.0
162166      106.0      23.0
162373      56.0      40.0
162740      85.0      4.0
162801      NaN      NaN

```

[99299 rows x 29 columns]

```

[16]: # Запоминаем индексы строк с пустыми значениями
      flt_index = data[data['n_guns_involved'].isnull()].index

```

```
flt_index
```

```
[16]: Int64Index([    0,    1,    3,    5,    7,    8,
    ↪9,
           14,   17,   19,
           ...
           160630, 160803, 160878, 161236, 161836, 161927,
    ↪162166,
           162373, 162740, 162801],
      dtype='int64', length=99299)
```

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

```
[17]:
```

	incident_id	date	state	city_or_county	\
0	461105	2013-01-01	Pennsylvania	McKeesport	
1	460726	2013-01-01	California	Hawthorne	
3	478925	2013-01-05	Colorado	Aurora	
5	478948	2013-01-07	Oklahoma	Tulsa	
7	479374	2013-01-21	Louisiana	New Orleans	
...	...	...	...	...	
161927	729430	2016-12-16	Wisconsin	Madison	
162166	730843	2016-12-18	Florida	Naples	
162373	729453	2016-12-19	California	Brawley	
162740	730974	2016-12-21	Arkansas	Fayetteville	
162801	732054	2016-12-22	Florida	Palm Harbor	

	n_killed	\	address	
0		1506 Versailles Avenue and Coursin Street		
↪0				
1		13500 block of Cerise Avenue		
↪1				
3		16000 block of East Ithaca Place		
↪4				
5		6000 block of South Owasso		
↪4				
7		LaSalle Street and Martin Luther King Jr. Boul...	0	
...				
161927		Fourth Street		
↪0				
162166		Pine Ridge Rd and Airport Pulling Road		
↪0				
162373		500 block of North Imperial		
↪1				
162740		800 South School Avenue		
↪1				
162801		252 Whisper Lake Road		
↪1				

	n_injured	\
0	4	
1	3	
3	0	
5	0	
7	5	
...	...	
161927	0	
162166	0	
162373	0	
162740	0	
162801	0	

	incident_url	\
0	<a href="http://www.gunviolencearchive.org/incident/461105">http://www.gunviolencearchive.org/incident/461105</a>	
1	<a href="http://www.gunviolencearchive.org/incident/460726">http://www.gunviolencearchive.org/incident/460726</a>	
3	<a href="http://www.gunviolencearchive.org/incident/478925">http://www.gunviolencearchive.org/incident/478925</a>	
5	<a href="http://www.gunviolencearchive.org/incident/478948">http://www.gunviolencearchive.org/incident/478948</a>	
7	<a href="http://www.gunviolencearchive.org/incident/479374">http://www.gunviolencearchive.org/incident/479374</a>	
...	...	
161927	<a href="http://www.gunviolencearchive.org/incident/729430">http://www.gunviolencearchive.org/incident/729430</a>	
162166	<a href="http://www.gunviolencearchive.org/incident/730843">http://www.gunviolencearchive.org/incident/730843</a>	
162373	<a href="http://www.gunviolencearchive.org/incident/729453">http://www.gunviolencearchive.org/incident/729453</a>	
162740	<a href="http://www.gunviolencearchive.org/incident/730974">http://www.gunviolencearchive.org/incident/730974</a>	
162801	<a href="http://www.gunviolencearchive.org/incident/732054">http://www.gunviolencearchive.org/incident/732054</a>	

	source_url	\
0	<a href="http://www.post-gazette.com/local/south/2013/0...">http://www.post-gazette.com/local/south/2013/0...</a>	
1	<a href="http://www.dailybulletin.com/article/zz/201301...">http://www.dailybulletin.com/article/zz/201301...</a>	
3	<a href="http://www.dailydemocrat.com/20130106/aurora-s...">http://www.dailydemocrat.com/20130106/aurora-s...</a>	
5	<a href="http://usnews.nbcnews.com/_news/2013/01/07/163...">http://usnews.nbcnews.com/_news/2013/01/07/163...</a>	
7	<a href="http://www.nola.com/crime/index.ssf/2013/01/no...">http://www.nola.com/crime/index.ssf/2013/01/no...</a>	
...	...	
161927	<a href="http://www.nbc15.com/content/news/2-teens-arre...">http://www.nbc15.com/content/news/2-teens-arre...</a>	
162166	<a href="http://www.naplesnews.com/story/news/crime/201...">http://www.naplesnews.com/story/news/crime/201...</a>	
162373	<a href="http://www.kyma.com/news/fatal-officer-involve...">http://www.kyma.com/news/fatal-officer-involve...</a>	
162740	<a href="http://www.4029tv.com/article/officer-involved...">http://www.4029tv.com/article/officer-involved...</a>	
162801	<a href="http://web.tampabay.com/news/publicsafety/crim...">http://web.tampabay.com/news/publicsafety/crim...</a>	

	incident_url_fields_missing	...	\
0	False	...	
1	False	...	
3	False	...	
5	False	...	
7	False	...	
...	...	...	
161927	False	...	
162166	False	...	
162373	False	...	

162740	False	...
162801	False	...

	participant_age	\
0	0::20	
1	0::20	
3	0::29 1::33 2::56 3::33	
5	0::23 1::23 2::33 3::55	
7	NaN	
...	...	
161927	0::18 1::18	
162166	0::24	
162373	NaN	
162740	0::25	
162801	0::55	

	participant_age_group	\
0	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...	
1	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...	
3	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...	
5	0::Adult 18+ 1::Adult 18+ 2::Adult 18+ 3::A...	
7	NaN	
...	...	
161927	0::Adult 18+ 1::Adult 18+	
162166	0::Adult 18+	
162373	0::Adult 18+	
162740	0::Adult 18+	
162801	0::Adult 18+	

	participant_gender	\
0	0::Male 1::Male 3::Male 4::Female	
1	0::Male	
3	0::Female 1::Male 2::Male 3::Male	
5	0::Female 1::Female 2::Female 3::Female 4::...	
7	0::Male 1::Male 2::Male 3::Male 4::Male	
...	...	
161927	0::Male 1::Male	
162166	0::Male	
162373	0::Male	
162740	0::Male	
162801	0::Male	

	participant_name	\
0	0::Julian Sims	
1	0::Bernard Gillis	
3	0::Stacie Philbrook 1::Christopher Ratliffe ...	
5	0::Rebeika Powell 1::Kayetie Melchor 2::Mist...	
7	NaN	
...	...	
161927	0::Taylor Loving 1::Theron Walker	

162166	0::Sean Blackwell
162373	NaN
162740	0::Benjamin Ortiz
162801	0::Stanley Eversole

	participant_relationship \
0	NaN
1	NaN
3	NaN
5	NaN
7	NaN
...	...
161927	NaN
162166	NaN
162373	NaN
162740	NaN
162801	NaN

	participant_status \
0	0::Arrested 1::Injured 2::Injured 3::Injure...
1	0::Killed 1::Injured 2::Injured 3::Injured
3	0::Killed 1::Killed 2::Killed 3::Killed
5	0::Killed 1::Killed 2::Killed 3::Killed 4:...
7	0::Injured 1::Injured 2::Injured 3::Injured...
...	...
161927	0::Unharmed, Arrested 1::Unharmed, Arrested
162166	0::Unharmed, Arrested
162373	0::Killed
162740	0::Killed
162801	0::Killed

	participant_type \
0	0::Victim 1::Victim 2::Victim 3::Victim 4:...
1	0::Victim 1::Victim 2::Victim 3::Victim 4:...
3	0::Victim 1::Victim 2::Victim 3::Subject-Su...
5	0::Victim 1::Victim 2::Victim 3::Victim 4:...
7	0::Victim 1::Victim 2::Victim 3::Victim 4:...
...	...
161927	0::Subject-Suspect 1::Subject-Suspect
162166	0::Subject-Suspect
162373	0::Subject-Suspect
162740	0::Subject-Suspect
162801	0::Subject-Suspect

	sources \
0	<a href="http://pittsburgh.cbslocal.com/2013/01/01/4-pe...">http://pittsburgh.cbslocal.com/2013/01/01/4-pe...</a>
1	<a href="http://losangeles.cbslocal.com/2013/01/01/man-...">http://losangeles.cbslocal.com/2013/01/01/man-...</a>
3	<a href="http://denver.cbslocal.com/2013/01/06/officer-...">http://denver.cbslocal.com/2013/01/06/officer-...</a>
5	<a href="http://www.kjrh.com/news/local-news/4-found-sh...">http://www.kjrh.com/news/local-news/4-found-sh...</a>
7	<a href="http://www.huffingtonpost.com/2013/01/21/new-o...">http://www.huffingtonpost.com/2013/01/21/new-o...</a>

```
...
161927 http://www.nbc15.com/content/news/2-teens-arre...
162166 http://www.naplesnews.com/story/news/crime/201...
162373 http://www.kyma.com/news/fatal-officer-involve...
162740 http://www.4029tv.com/article/officer-involved...
162801 http://www.nbcmiami.com/news/local/Deputies-Sh...
```

	state_house_district	state_senate_district
0	NaN	NaN
1	62.0	35.0
3	40.0	28.0
5	72.0	11.0
7	93.0	5.0
...	...	...
161927	76.0	26.0
162166	106.0	23.0
162373	56.0	40.0
162740	85.0	4.0
162801	NaN	NaN

[99299 rows x 29 columns]

```
[18]: # фильтр по колонке
data_num[data_num.index.isin(flt_index)]['n_guns_involved']
```

```
[18]: 0      NaN
      1      NaN
      3      NaN
      5      NaN
      7      NaN
      ..
161927 NaN
162166 NaN
162373 NaN
162740 NaN
162801 NaN
Name: n_guns_involved, Length: 99299, dtype: float64
```

Будем использовать встроенные средства импутации библиотеки scikit-learn - <https://scikit-learn.org/stable/modules/impute.html#impute>

```
[116]: data_num_guns = data_num[['n_guns_involved']]
data_num_guns.head()
```

```
[116]: n_guns_involved
0      NaN
1      NaN
2      2.0
3      NaN
4      2.0
```

```
[117]: from sklearn.impute import SimpleImputer
       from sklearn.impute import MissingIndicator

[118]: # Фильтр для проверки заполнения пустых значений
       indicator = MissingIndicator()
       mask_missing_values_only = indicator.fit_transform(data_num_guns)
       mask_missing_values_only
```

```
[118]: array([[ True],
              [ True],
              [False],
              ...,
              [False],
              [False],
              [False]])
```

С помощью класса SimpleImputer можно проводить импьютацию различными показателями центра распределения

```
[119]: strategies=['mean', 'median', 'most_frequent']
```

```
[120]: def test_num_impute(strategy_param):
       imp_num = SimpleImputer(strategy=strategy_param)
       data_num_imp = imp_num.fit_transform(data_num_guns)
       return data_num_imp[mask_missing_values_only]
```

```
[121]: strategies[0], test_num_impute(strategies[0])
```

```
[121]: ('mean',
       array([1.5237069, 1.5237069, 1.5237069, ..., 1.5237069, 1.5237069,
              1.5237069]))
```

```
[122]: strategies[1], test_num_impute(strategies[1])
```

```
[122]: ('median', array([1., 1., 1., ..., 1., 1., 1.]))
```

```
[123]: strategies[2], test_num_impute(strategies[2])
```

```
[123]: ('most_frequent', array([1., 1., 1., ..., 1., 1., 1.]))
```

```
[124]: # Более сложная функция, которая позволяет задавать колонку и вид
       ↪ импьютации
       def test_num_impute_col(dataset, column, strategy_param):
           temp_data = dataset[[column]]

           indicator = MissingIndicator()
           mask_missing_values_only = indicator.fit_transform(temp_data)

           imp_num = SimpleImputer(strategy=strategy_param)
           data_num_imp = imp_num.fit_transform(temp_data)

           filled_data = data_num_imp[mask_missing_values_only]
```



```

    return column, strategy_param, filled_data.size,
    ↪filled_data[0], filled_data[filled_data.size-1]

```

```
[125]: test_num_impute_col(data, 'n_guns_involved', strategies[0])
```

```
[125]: ('n_guns_involved', 'mean', 99299, 1.5237068965517242, 1.
    ↪5237068965517242)
```

```
[126]: test_num_impute_col(data, 'n_guns_involved', strategies[1])
```

```
[126]: ('n_guns_involved', 'median', 99299, 1.0, 1.0)
```

```
[127]: test_num_impute_col(data, 'n_guns_involved', strategies[2])
```

```
[127]: ('n_guns_involved', 'most_frequent', 99299, 1.0, 1.0)
```

### 2.2.2. 1.2.2. Обработка пропусков в категориальных данных

```

[128]: # 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 / total_count) * 100.
    ↪0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых
    ↪значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))

```

Колонка address. Тип данных object. Количество пустых значений  
 ↪12303, 7.55%.

Колонка source\_url. Тип данных object. Количество пустых значений  
 ↪276, 0.17%.

Колонка gun\_stolen. Тип данных object. Количество пустых значений  
 ↪99311, 60.98%.

Колонка gun\_type. Тип данных object. Количество пустых значений  
 ↪99299, 60.97%.

Колонка incident\_characteristics. Тип данных object. Количество  
 ↪пустых значений  
 242, 0.15%.

Колонка location\_description. Тип данных object. Количество пустых  
 ↪значений  
 140476, 86.25%.

Колонка notes. Тип данных object. Количество пустых значений  
 ↪56008, 34.39%.

Колонка participant\_age. Тип данных object. Количество пустых  
 ↳ значений 63464,  
 38.97%.  
 Колонка participant\_age\_group. Тип данных object. Количество  
 ↳ пустых значений  
 27678, 16.99%.  
 Колонка participant\_gender. Тип данных object. Количество пустых  
 ↳ значений 23832,  
 14.63%.  
 Колонка participant\_name. Тип данных object. Количество пустых  
 ↳ значений 84207,  
 51.7%.  
 Колонка participant\_relationship. Тип данных object. Количество  
 ↳ пустых значений  
 152618, 93.71%.  
 Колонка participant\_status. Тип данных object. Количество пустых  
 ↳ значений 18510,  
 11.37%.  
 Колонка participant\_type. Тип данных object. Количество пустых  
 ↳ значений 16327,  
 10.02%.  
 Колонка sources. Тип данных object. Количество пустых значений  
 ↳ 516, 0.32%.

Класс SimpleImputer можно использовать для категориальных признаков со стратегиями “most\_frequent” или “constant”.

```
[129]: cat_temp_data = data[['gun_stolen']]
cat_temp_data.head()
```

```
[129]:          gun_stolen
0                NaN
1                NaN
2  0::Unknown|1::Unknown
3                NaN
4  0::Unknown|1::Unknown
```

```
[130]: cat_temp_data['gun_stolen'].unique()[0:10]
```

```
[130]: array([nan, '0::Unknown|1::Unknown', '0::Unknown',
              '0::Unknown|1::Unknown|2::Unknown|3::Unknown',
              '0::Not-stolen|1::Unknown', '0::Unknown|1::Unknown|2::
↳ Unknown',
              '0::Stolen|1::Stolen', '0::Not-stolen', '0::Stolen',
              '0::Stolen|1::Stolen|2::Unknown|3::Unknown'],
↳ dtype=object)
```

```
[131]: cat_temp_data[cat_temp_data['gun_stolen'].isnull()].shape
```

```
[131]: (99311, 1)
```

```
[132]: # Импьютация наиболее частыми значениями
imp2 = SimpleImputer(missing_values=np.nan,
    ↪strategy='most_frequent')
data_imp2 = imp2.fit_transform(cat_temp_data)
data_imp2
```

```
[132]: array([[ '0::Unknown'],
             [ '0::Unknown'],
             [ '0::Unknown|1::Unknown'],
             ...,
             [ '0::Unknown'],
             [ '0::Unknown'],
             [ '0::Unknown']], dtype=object)
```

```
[133]: # Пустые значения отсутствуют
np.unique(data_imp2)[0:5]
```

```
[133]: array([ '0::Not-stolen', '0::Not-stolen|1::Not-stolen',
             '0::Not-stolen|1::Not-stolen|2::Not-stolen',
             '0::Not-stolen|1::Not-stolen|2::Not-stolen|3::
    ↪Not-stolen',
             '0::Not-stolen|1::Not-stolen|2::Not-stolen|3::
    ↪Not-stolen|4::Not-
stolen'],
             dtype=object)
```

```
[134]: # Импьютация константой
imp3 = SimpleImputer(missing_values=np.nan, strategy='constant',
    ↪fill_value='!!!')
data_imp3 = imp3.fit_transform(cat_temp_data)
data_imp3
```

```
[134]: array([[ '!!!'],
             [ '!!!'],
             [ '0::Unknown|1::Unknown'],
             ...,
             [ '0::Unknown'],
             [ '0::Unknown'],
             [ '0::Unknown']], dtype=object)
```

```
[135]: np.unique(data_imp3)[0:5]
```

```
[135]: array([ '!!!', '0::Not-stolen', '0::Not-stolen|1::Not-stolen',
             '0::Not-stolen|1::Not-stolen|2::Not-stolen',
             '0::Not-stolen|1::Not-stolen|2::Not-stolen|3::
    ↪Not-stolen'],
             dtype=object)
```

```
[136]: data_imp3[data_imp3=='!!!'].size
```

```
[136]: 99311
```

## 3. 2. Преобразование категориальных признаков в числовые

```
[137]: cat_enc = pd.DataFrame({'c1':data_imp2.T[0]})
cat_enc
```

```
[137]:
```

	c1
0	0::Unknown
1	0::Unknown
2	0::Unknown 1::Unknown
3	0::Unknown
4	0::Unknown 1::Unknown
...	...
162862	0::Unknown
162863	0::Unknown
162864	0::Unknown
162865	0::Unknown
162866	0::Unknown

[162867 rows x 1 columns]

### 3.1. 2.1. Кодирование категорий целочисленными значениями - label encoding

```
[138]: from sklearn.preprocessing import LabelEncoder, OneHotEncoder
```

```
[139]: le = LabelEncoder()
cat_enc_le = le.fit_transform(cat_enc['c1'])
```

```
[140]: cat_enc['c1'].unique()[0:5]
```

```
[140]: array(['0::Unknown', '0::Unknown|1::Unknown',
          '0::Unknown|1::Unknown|2::Unknown|3::Unknown',
          '0::Not-stolen|1::Unknown', '0::Unknown|1::Unknown|2::
          ↪Unknown'],
          dtype=object)
```

```
[141]: np.unique(cat_enc_le)[0:10]
```

```
[141]: array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```

```
[142]: le.inverse_transform([0, 1, 2, 3])
```

```
[142]: array(['0::Not-stolen', '0::Not-stolen|1::Not-stolen',
          '0::Not-stolen|1::Not-stolen|2::Not-stolen',
          '0::Not-stolen|1::Not-stolen|2::Not-stolen|3::
          ↪Not-stolen'],
          dtype=object)
```

### 3.2. 2.2. Кодирование категорий наборами бинарных значений - one-hot encoding

```
[143]: ohe = OneHotEncoder()  
cat_enc_ohe = ohe.fit_transform(cat_enc[['c1']])
```

```
[144]: cat_enc.shape
```

```
[144]: (162867, 1)
```

```
[145]: cat_enc_ohe.shape
```

```
[145]: (162867, 277)
```

```
[146]: cat_enc_ohe
```

```
[146]: <162867x277 sparse matrix of type '<class 'numpy.float64'>'  
      with 162867 stored elements in Compressed Sparse Row  
      ↪format>
```

```
[147]: cat_enc_ohe.todense()[0:10]
```

```
[147]: matrix([[0., 0., 0., ..., 0., 0., 0.],  
            [0., 0., 0., ..., 0., 0., 0.],  
            [0., 0., 0., ..., 0., 0., 0.],  
            ...,  
            [0., 0., 0., ..., 0., 0., 0.],  
            [0., 0., 0., ..., 0., 0., 0.],  
            [0., 0., 0., ..., 0., 0., 0.]])
```

```
[148]: cat_enc.head(10)
```

```
[148]:
```

	c1
0	0::Unknown
1	0::Unknown
2	0::Unknown   1::Unknown
3	0::Unknown
4	0::Unknown   1::Unknown
5	0::Unknown
6	0::Unknown   1::Unknown
7	0::Unknown
8	0::Unknown
9	0::Unknown

## 4. 3. Масштабирование данных

Термины “масштабирование” и “нормализация” часто используются как синонимы. Масштабирование предполагает изменение диапазона измерения величины, а нормализация - изменение распределения этой величины.

Если признаки лежат в различных диапазонах, то необходимо их нормализовать. Как правило, применяют два подхода: - MinMax масштабирование:

$$x = \frac{x - \min(X)}{\max(X) - \min(X)}$$

В этом случае значения лежат в диапазоне от 0 до 1. - Масштабирование данных на основе Z-оценки:

$$x = \frac{x - \text{AVG}(X)}{\sigma(X)}$$

В этом случае большинство значений попадает в диапазон от -3 до 3.

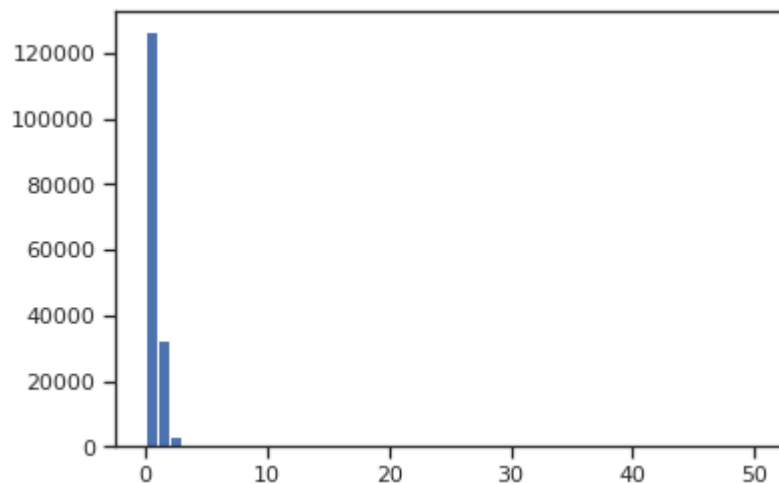
где  $X$  - матрица объект-признак,  $\text{AVG}(X)$  - среднее значение,  $\sigma$  - среднеквадратичное отклонение.

```
[149]: from sklearn.preprocessing import MinMaxScaler, StandardScaler, \
      ↪ Normalizer
```

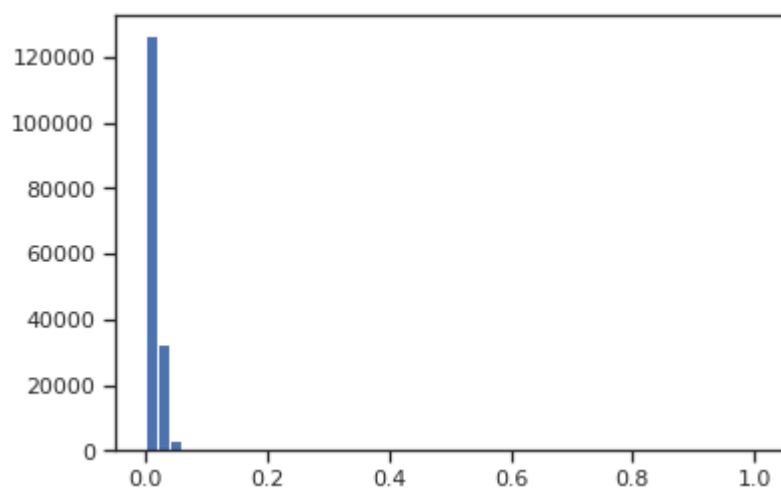
### 4.1. 3.1. MinMax масштабирование

```
[150]: sc1 = MinMaxScaler()
      sc1_data = sc1.fit_transform(data[['n_killed']])
```

```
[151]: plt.hist(data['n_killed'], 50)
      plt.show()
```



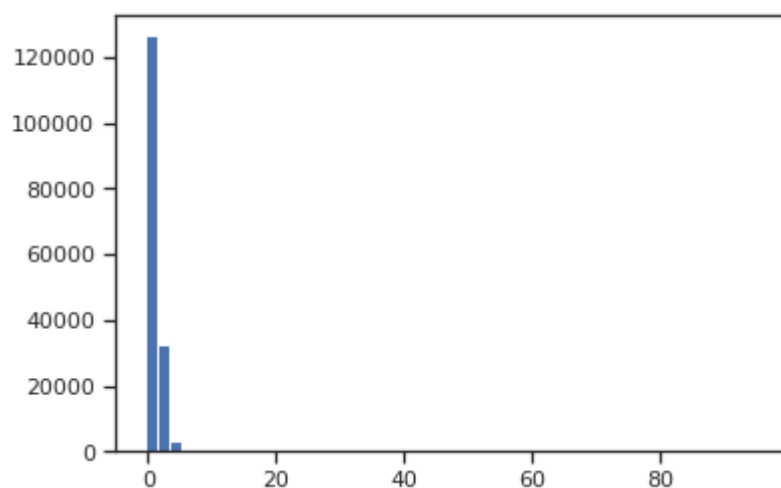
```
[152]: plt.hist(sc1_data, 50)
      plt.show()
```



#### 4.2. 3.2. Масштабирование данных на основе Z-оценки - StandardScaler

```
[153]: sc2 = StandardScaler()
       sc2_data = sc2.fit_transform(data[['n_killed']])
```

```
[154]: plt.hist(sc2_data, 50)
       plt.show()
```



#### 4.3. 3.3. Нормализация данных

```
[155]: sc3 = Normalizer()
       sc3_data = sc3.fit_transform(data[['n_killed']])
```

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

