Лабораторная работа №2 по курсу "Технологии машинного обучения"

Выполнила Попова Дарья, студентка группы РТ5-61Б

Обработка пропусков в данных. Замена нулевых значений на медиану в колонке "возраст"

In [1]:

```
import numpy as np
import pandas as pd
homicide = pd.read_csv('C:\\Users\\Дасупс\\Downloads\\homicide.csv')
```

```
C:\ProgramData\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3146: DtypeWarning: Columns (16) h
ave mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [2]:

```
homicide.shape
```

Out[2]:

```
(638454, 24)
```

In [3]:

```
homicide.head()
```

Out[3]:

	Record ID	Agency Code	Agency Name	Agency Type	City	State	Year	Month	Incident	Crime Type	...	Victim Ethnicity	Perpetrator Sex	Perpetrator Age	
0	1	AK00101	Anchorage	Municipal Police	Anchorage	Alaska	1980	January	1	Murder or Manslaughter	...	Unknown	Male	15	Ame
1	2	AK00101	Anchorage	Municipal Police	Anchorage	Alaska	1980	March	1	Murder or Manslaughter	...	Unknown	Male	42	
2	3	AK00101	Anchorage	Municipal Police	Anchorage	Alaska	1980	March	2	Murder or Manslaughter	...	Unknown	Unknown	0	
3	4	AK00101	Anchorage	Municipal Police	Anchorage	Alaska	1980	April	1	Murder or Manslaughter	...	Unknown	Male	42	
4	5	AK00101	Anchorage	Municipal Police	Anchorage	Alaska	1980	April	2	Murder or Manslaughter	...	Unknown	Unknown	0	

5 rows × 24 columns

In [4]:

```
homicide = homicide.rename(columns={'Perpetrator Sex':'Perpetrator_Sex','Perpetrator Age':'Perpetrator_Age','I
```

Посчитаем число нулей в колонке Perpetrator_Age

In [5]:

```
homicide.query('Perpetrator_Age==0').agg({'Record_ID':'count'})
```

Out[5]:

```
Record_ID    211079
dtype: int64
```

In [18]:

```
homicide['Perpetrator_Age'].isnull().any()
```

Out[18]:

```
False
```

При моей первой попытке построить гистограмму оказалось, что это не получается сделать из-за невозможности представить значения колонки Perpetrator_Age в числовом формате - они были object'ами.

In [6]:

```
homicide.dtypes
```

Out[6]:

```
Record_ID                int64
Agency Code             object
Agency Name             object
Agency Type             object
City                    object
State                   object
Year                     int64
Month                   object
Incident                 int64
Crime Type              object
Crime Solved            object
Victim Sex              object
Victim Age               int64
Victim Race             object
Victim Ethnicity        object
Perpetrator_Sex         object
Perpetrator_Age         object
Perpetrator Race        object
Perpetrator Ethnicity   object
Relationship            object
Weapon                  object
Victim Count             int64
Perpetrator Count        int64
Record Source           object
dtype: object
```

Далее я предприняла попытку воспользоваться функцией to_numeric из библиотеки pandas, но тут этого не удалось сделать из-за того, что в строке с индексом 634666 в колонке Perpetrator_Age содержится пробел " ", который не может быть преобразован в числовое значение.

In [7]:

```
homicide.loc[[634666]]
```

Out[7]:

	Record_ID	Agency Code	Agency Name	Agency Type	City	State	Year	Month	Incident	Crime Type	...	Victim Ethnicity	Perpetrator_Sex	Perpetrat
634666	634667	OK07205	Tulsa	Municipal Police	Tulsa	Oklahoma	2014	June	104	Murder or Manslaughter	...	Not Hispanic	Unknown	

1 rows × 24 columns

In [8]:

```
homicide = homicide.drop([634666])
```

In [9]:

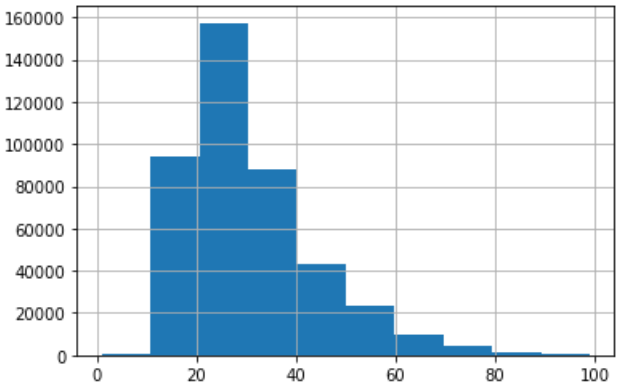
```
homicide['Perpetrator_Age'] = pd.to_numeric(arg=homicide.Perpetrator_Age)
```

In [10]:

```
homicide.query('Perpetrator_Age > 0').Perpetrator_Age.hist()
```

Out[10]:

```
<AxesSubplot:>
```



In [11]:

```
homicide.query('Perpetrator_Age > 0').median()
```

Out[11]:

```
Record_ID            310118.5
Year                   1994.0
Incident                  2.0
Victim Age               30.0
Perpetrator_Age          27.0
Victim Count              0.0
Perpetrator Count         0.0
dtype: float64
```

In [12]:

```
homicide.loc[homicide.Perpetrator_Age == 0] = 27
```

In [13]:

```
homicide.query('Perpetrator_Age==0').agg({'Record_ID':'count'})
```

Out[13]:

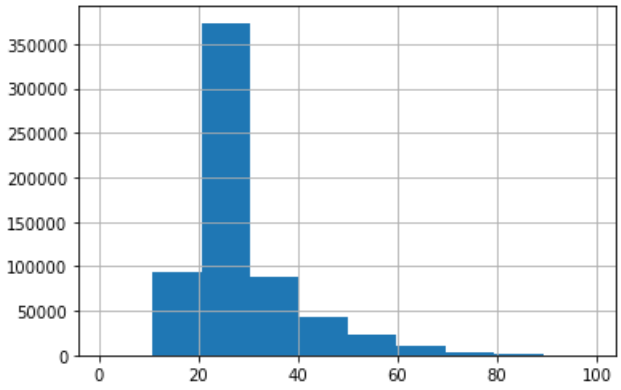
```
Record_ID    0
dtype: int64
```

In [14]:

```
homicide.Perpetrator_Age.hist()
```

Out[14]:

```
<AxesSubplot:>
```



In [ ]: