Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения» Отчет по лабораторной работе №2

«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.»

|  |  |
| --- | --- |
| Выполнил: | Проверил: |
| студент группы ИУ5-63Б | преподаватель каф. ИУ5 |
| Кащеев Максим | Гапанюк Ю.Е. |

# Москва, 2022 г.

**Описание задания:**

# 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.)

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

# обработку пропусков в данных;

* + кодирование категориальных признаков;

# масштабирование данных.

Лаборатораня работа №2: Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.

## Обработка пропусков в данных

In [1]:

**import** pandas **as** pd

**import** numpy **as** np

**from** matplotlib **import** pyplot **as** plt

**import** seaborn **as** sns

**import** math

**from** sklearn.impute **import** SimpleImputer

**from** sklearn.preprocessing **import** MinMaxScaler, StandardScaler, Normalizer

In [2]:

data**.**head()

data **=** pd**.**read\_csv('crimes.csv')

In [3]:

Out[3]:

### Location

**Description Inside/Outside Weapon Post District Neighborhood Location 1 Total**

### Incidents

300 SAINT PAUL

|  |  |  |
| --- | --- | --- |
| **CrimeDate** | **CrimeTime** | **CrimeCode** |
| **0** 11/12/2016 | 02:35:00 | 3B |
| **1** 11/12/2016 | 02:56:00 | 3CF |
| **2** 11/12/2016 | 03:00:00 | 6D |
| **3** 11/12/2016 | 03:00:00 | 6D |
| **4** 11/12/2016 | 03:00:00 | 6E |

PL

ROBBERY - STREET O NaN 111.0 CENTRAL Downtown (39.2924100000, 1

-76.6140800000)

800 S BROADWAY

ROBBERY - COMMERCIAL

I FIREARM 213.0 SOUTHEASTERN Fells Point (39.2824200000, 1

-76.5928800000)

1500

PENTWOOD RD

LARCENY FROM

AUTO

O NaN 413.0 NORTHEASTERN Stonewood-

Pentwood-Winston

(39.3480500000,

-76.5883400000) 1

6600 MILTON LN LARCENY FROM

AUTO

O NaN 424.0 NORTHEASTERN Westfield (39.3626300000, 1

-76.5516100000)

300 W BALTIMORE ST

LARCENY O NaN 111.0 CENTRAL Downtown (39.2893800000, 1

-76.6197100000)

In [4]:

data**.**info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 285807 entries, 0 to 285806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

* 1. CrimeDate 285807 non-null object
  2. CrimeTime 285807 non-null object
  3. CrimeCode 285807 non-null object
  4. Location 284184 non-null object
  5. Description 285807 non-null object
  6. Inside/Outside 281611 non-null object
  7. Weapon 97396 non-null object
  8. Post 285616 non-null float64
  9. District 285749 non-null object
  10. Neighborhood 284106 non-null object
  11. Location 1 284188 non-null object
  12. Total Incidents 285807 non-null int64 dtypes: float64(1), int64(1), object(10)

memory usage: 26.2+ MB

In [5]:

bad\_cols **=** []

**for** col **in** data**.**columns:

**if** data[col]**.**isnull()**.**sum(): bad\_cols**.**append(col)

print(f'{col}: {data[col]**.**isnull()**.**sum()} ({round(data[col]**.**isnull()**.**sum() **/** data**.**shape[0] **\*** 100, 3)}%)') print('\nbad columns:', bad\_cols)

CrimeDate: 0 (0.0%)

CrimeTime: 0 (0.0%)

CrimeCode: 0 (0.0%)

Location: 1623 (0.568%)

Description: 0 (0.0%)

Inside/Outside: 4196 (1.468%)

Weapon: 188411 (65.922%)

Post: 191 (0.067%)

District: 58 (0.02%)

Neighborhood: 1701 (0.595%)

Location 1: 1619 (0.566%)

Total Incidents: 0 (0.0%)

bad columns: ['Location', 'Inside/Outside', 'Weapon', 'Post', 'District', 'Neighborhood', 'Location 1']

In [6]:

print("Количество уникальных значений\n")

**for** col **in** data**.**columns:

print(f'{col}: {data[col]**.**unique()**.**size}')

Количество уникальных значений

CrimeDate: 2143

CrimeTime: 4236

CrimeCode: 81

Location: 25950

Description: 15

Inside/Outside: 5

Weapon: 5

Post: 190

District: 14

Neighborhood: 281

Location 1: 97952

Total Incidents: 1

In [7]:

data**.**drop(['Total Incidents', 'Weapon'], axis**=**1, inplace**=True**) *# малоинформативные столбцы*

data**.**head()

Out[7]:

### CrimeDate CrimeTime CrimeCode Location Description Inside/Outside Post District Neighborhood Location 1

**0** 11/12/2016 02:35:00 3B 300 SAINT PAUL PL ROBBERY - STREET O 111.0 CENTRAL Downtown (39.2924100000,

-76.6140800000)

**1** 11/12/2016 02:56:00 3CF 800 S BROADWAY ROBBERY -

COMMERCIAL

I 213.0 SOUTHEASTERN Fells Point (39.2824200000,

-76.5928800000)

### CrimeDate CrimeTime CrimeCode Location Description Inside/Outside Post District Neighborhood Location 1

**2** 11/12/2016 03:00:00 6D 1500 PENTWOOD

RD

LARCENY FROM AUTO O 413.0 NORTHEASTERN Stonewood-Pentwood-

Winston

(39.3480500000,

-76.5883400000)

**3** 11/12/2016 03:00:00 6D 6600 MILTON LN LARCENY FROM AUTO O 424.0 NORTHEASTERN Westfield (39.3626300000,

-76.5516100000)

**4** 11/12/2016 03:00:00 6E 300 W BALTIMORE

ST

LARCENY O 111.0 CENTRAL Downtown (39.2893800000,

-76.6197100000)

In [8]:

data['Inside/Outside']**.**unique()

Out[8]: array(['O', 'I', 'Outside', 'Inside', nan], dtype=object)

In [9]:

data['Inside/Outside']**.**replace('I', 'Inside', inplace**=True**) data['Inside/Outside']**.**replace('O', 'Outside', inplace**=True**) data['Inside/Outside']**.**unique()

Out[9]: array(['Outside', 'Inside', nan], dtype=object)

In [10]:

data['Inside/Outside']**.**value\_counts()

Out[10]: Inside 142531

Outside 139080

Name: Inside/Outside, dtype: int64

In [11]:

data['Post']**.**value\_counts()

Out[11]: 111.0 9776

212.0 4982

922.0 4655

211.0 4513

913.0 4470

...

2.1 1

925.0 1

0.7 1

0.6 1

1.0 1

Name: Post, Length: 189, dtype: int64

In [12]:

data['District']**.**value\_counts()

Out[12]: NORTHEASTERN 44832

SOUTHEASTERN 39245

CENTRAL 33782

SOUTHERN 33031

NORTHERN 32005

NORTHWESTERN 28690

SOUTHWESTERN 26242

EASTERN 24168

WESTERN 23266

NORTHESTERN 280

SOUTHESTERN 205

Central 2

Gay Street 1

Name: District, dtype: int64

In [13]:

data['Neighborhood']**.**value\_counts()

Out[13]: Downtown 9666

Frankford 6791

Belair-Edison 6133

Brooklyn 4528

Cherry Hill 4273

...

Mt Pleasant Park 12

Blythewood 5

EASTERN 1

Dundalk Marine Terminal 1

NORTHEASTERN 1

Name: Neighborhood, Length: 280, dtype: int64

In [14]:

data **=** data**.**dropna(subset**=**['Location 1'], axis**=**0)

x **=** [float(elem[1:14]) **for** elem **in** data['Location 1']] y **=** [float(elem[17:30]) **for** elem **in** data['Location 1']]

data['X'] **=** pd**.**Series(x)**.**reindex(data**.**index, method**=**'ffill') data['Y'] **=** pd**.**Series(y)**.**reindex(data**.**index, method**=**'ffill')

data**.**drop(['Location 1'], axis**=**1, inplace**=True**)

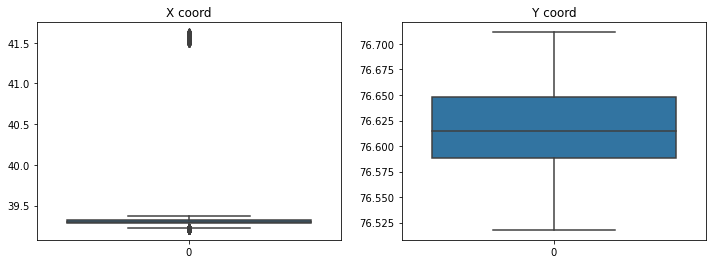
In [15]:

**def** show\_coords\_dist():

fig, (ax1, ax2) **=** plt**.**subplots(1, 2, figsize**=**(12, 4)) sns**.**boxplot(data**=**data['X'], ax**=**ax1);

sns**.**boxplot(data**=**data['Y'], ax**=**ax2);

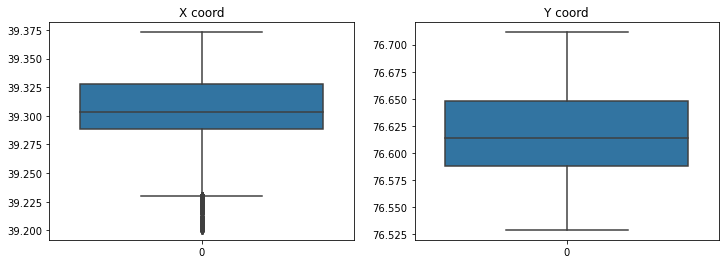
ax1**.**set\_title('X coord') ax2**.**set\_title('Y coord') plt**.**show()

In [16]:

show\_coords\_dist()

In [17]:

data**.**drop(data[data['X'] **>** 40]**.**index, inplace**=True**)

In [18]:

show\_coords\_dist()

In [19]:

SimpleImputer(strategy**=**'constant', fill\_value**=**'-', copy**=False**)**.**fit\_transform(data['Location']**.**values**.**reshape(**-**1, 1))

SimpleImputer(strategy**=**'constant', fill\_value**=**'-', copy**=False**)**.**fit\_transform(data['Neighborhood']**.**values**.**reshape(**-**1, 1))

SimpleImputer(strategy**=**'constant', fill\_value**=**'Inside', copy**=False**)**.**fit\_transform(data['Inside/Outside']**.**values**.**reshape(**-**1, 1))

data **=** data**.**dropna(subset**=**['Post'], axis**=**0)

data **=** data**.**dropna(subset**=**['District'], axis**=**0) data **=** data**.**dropna(subset**=**['X'], axis**=**0)

data **=** data**.**dropna(subset**=**['Y'], axis**=**0)

In [20]:

data**.**isnull()**.**sum()

Out[20]: CrimeDate 0

CrimeTime 0

CrimeCode 0

Location 0

Description 0

Inside/Outside 0

Post 0

District 0

Neighborhood 0

X 0

Y 0

dtype: int64

## Кодирование категориальных признаков

In [21]:

data**.**head()

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Out[21]: | **CrimeDate** | **CrimeTime** | **CrimeCode** | **Location** | **Description** | **Inside/Outside** | **Post** | **District** | **Neighborhood** | **X** | **Y** |
|  | **0** 11/12/2016 | 02:35:00 | 3B | 300 SAINT PAUL PL | ROBBERY - STREET | Outside | 111.0 | CENTRAL | Downtown | 39.29241 | 76.61408 |
|  | **1** 11/12/2016 | 02:56:00 | 3CF | 800 S BROADWAY | ROBBERY - COMMERCIAL | Inside | 213.0 | SOUTHEASTERN | Fells Point | 39.28242 | 76.59288 |
|  | **2** 11/12/2016 | 03:00:00 | 6D | 1500 PENTWOOD RD | LARCENY FROM AUTO | Outside | 413.0 | NORTHEASTERN | Stonewood-Pentwood-Winston | 39.34805 | 76.58834 |
|  | **3** 11/12/2016 | 03:00:00 | 6D | 6600 MILTON LN | LARCENY FROM AUTO | Outside | 424.0 | NORTHEASTERN | Westfield | 39.36263 | 76.55161 |
|  | **4** 11/12/2016 | 03:00:00 | 6E | 300 W BALTIMORE ST | LARCENY | Outside | 111.0 | CENTRAL | Downtown | 39.28938 | 76.61971 |

In [22]:

data['Hours'] **=** [int(elem[:2]) **for** elem **in** data['CrimeTime']]

data['Minutes'] **=** [int(elem[3:5]) **for** elem **in** data['CrimeTime']] data**.**head()

Out[22]:

### CrimeDate CrimeTime CrimeCode Location Description Inside/Outside Post District Neighborhood X Y Hours Minutes

300 SAINT PAUL

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **0** | 11/12/2016 | 02:35:00 | 3B |
| **1** | 11/12/2016 | 02:56:00 | 3CF |
| **2** | 11/12/2016 | 03:00:00 | 6D |
| **3** | 11/12/2016 | 03:00:00 | 6D |
|  | **4** | 11/12/2016 | 03:00:00 | 6E |

PL

ROBBERY - STREET Outside 111.0 CENTRAL Downtown 39.29241 76.61408 2 35

800 S BROADWAY

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | Winston |  | | | |
| Outside | 424.0 | NORTHEASTERN | Westfield | 39.36263 | 76.55161 | 3 | 0 |
| Outside | 111.0 | CENTRAL | Downtown | 39.28938 | 76.61971 | 3 | 0 |

ROBBERY - COMMERCIAL

Inside 213.0 SOUTHEASTERN Fells Point 39.28242 76.59288 2 56

1500

PENTWOOD RD

LARCENY FROM

AUTO

Outside 413.0 NORTHEASTERN Stonewood-Pentwood-

39.34805 76.58834 3 0

6600 MILTON LN LARCENY FROM

AUTO

300 W BALTIMORE ST

LARCENY

In [23]:

category\_cols **=** ['CrimeCode', 'Location', 'Description', 'Inside/Outside', 'District', 'Neighborhood']

In [24]:

print("Количество уникальных значений\n")

**for** col **in** category\_cols:

print(f'{col}: {data[col]**.**unique()**.**size}')

Количество уникальных значений CrimeCode: 81

Location: 25937

Description: 15

Inside/Outside: 2

District: 12

Neighborhood: 281

In [25]:

category\_cols**.**remove('Location')

category\_cols**.**remove('Neighborhood')

In [26]:

**for** col **in** category\_cols:

data **=** pd**.**concat([data, pd**.**get\_dummies(data[col])], axis**=**1)

In [27]:

data**.**head()

Out[27]:

### CrimeDate CrimeTime CrimeCode Location Description Inside/Outside Post District Neighborhood X ... Gay

**Street**

### NORTHEASTERN NORTHERN NORT

**Gay Street**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CrimeDate** | **CrimeTime** | **CrimeCode** | **Location** | **Description** | **Inside/Outside** | **Post** | **District** | **Neighborhood** | **X** | **...** |

### NORTHEASTERN NORTHERN NORT

**0** 11/12/2016 02:35:00 3B 300 SAINT

PAUL PL

**1** 11/12/2016 02:56:00 3CF 800 S

BROADWAY

1500

**2** 11/12/2016 03:00:00 6D PENTWOOD

RD

ROBBERY - STREET

ROBBERY - COMMERCIAL

LARCENY FROM AUTO

0 0 0

0 0 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outside | 111.0 | CENTRAL | Downtown | 39.29241 | ... |
| Inside | 213.0 | SOUTHEASTERN | Fells Point | 39.28242 | ... |
| Outside | 413.0 | NORTHEASTERN | Stonewood- Pentwood- | 39.34805 | ... |
|  |  |  | Winston |  |  |

0 1 0

**3** 11/12/2016 03:00:00 6D 6600

MILTON LN

300 W

LARCENY FROM AUTO

Outside 424.0 NORTHEASTERN Westfield 39.36263 ... 0 1 0

**4** 11/12/2016 03:00:00 6E

BALTIMORE

ST

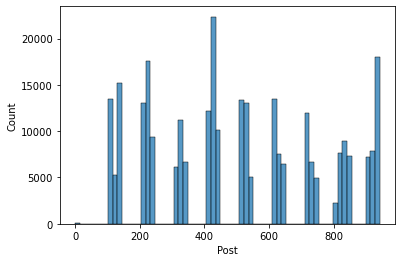
LARCENY Outside 111.0 CENTRAL Downtown 39.28938 ... 0 0 0

5 rows × 123 columns

## Масштабирование данных

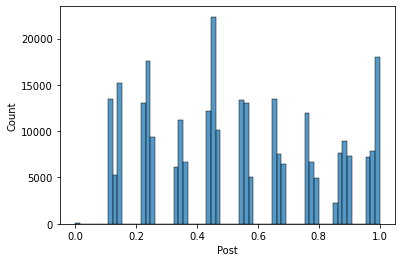
In [28]:

sns**.**histplot(data['Post'])

Out[28]: <AxesSubplot:xlabel='Post', ylabel='Count'>

In [29]:

data['Post'] **=** MinMaxScaler()**.**fit\_transform(data['Post']**.**values**.**reshape(**-**1, 1)) sns**.**histplot(data['Post']);



In [30]:

data['Hours'] **=** MinMaxScaler()**.**fit\_transform(data['Hours']**.**values**.**reshape(**-**1, 1))

data['Minutes'] **=** MinMaxScaler()**.**fit\_transform(data['Minutes']**.**values**.**reshape(**-**1, 1)) data['X'] **=** MinMaxScaler()**.**fit\_transform(data['X']**.**values**.**reshape(**-**1, 1))

data['Y'] **=** MinMaxScaler()**.**fit\_transform(data['Y']**.**values**.**reshape(**-**1, 1))

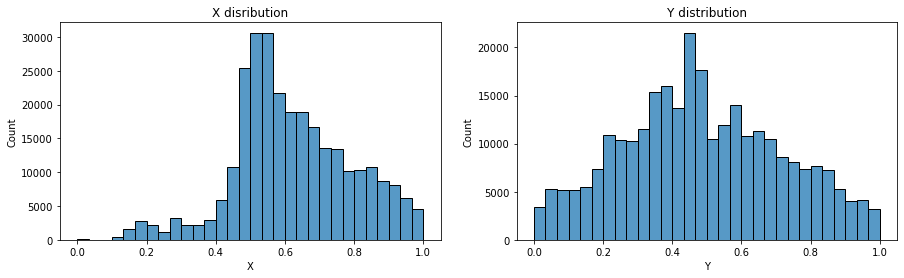
In [31]:

fig, (ax1, ax2) **=** plt**.**subplots(1, 2, figsize**=**(15, 4)) sns**.**histplot(data['X'], ax**=**ax1, bins**=**30)

sns**.**histplot(data['Y'], ax**=**ax2, bins**=**30) ax1**.**set\_title('X disribution')

ax2**.**set\_title('Y distribution')

plt**.**show()



In [32]:

fig, (ax1, ax2) **=** plt**.**subplots(1, 2, figsize**=**(15, 4)) sns**.**histplot(data['Hours'], ax**=**ax1)

sns**.**histplot(data['Minutes'], ax**=**ax2) ax1**.**set\_title('Hours disribution')

ax2**.**set\_title('Minutes distribution')

plt**.**show()

