**Московский государственный технический**

**университет им. Н.Э. Баумана.**

Факультет «Информатика и управление»

Кафедра ИУ5.

Курс «Технологии машинного обучения»

Отчет по лабораторной работе №3

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

|  |  |  |
| --- | --- | --- |
| Выполнил: |  | Проверил: |
| студент группы ИУ5-62 |  | преподаватель каф. ИУ5 |
| Король Константин |  | Гапанюк Ю.Е. |
| Подпись и дата: |  | Подпись и дата: |

Москва, 2019 г.

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](https://github.com/ugapanyuk/ml_course/blob/master/common/notebooks/missing/handling_missing_norm.ipynb)решить следующие задачи:

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

**Код программы**

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

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

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

**import** **matplotlib.pyplot** **as** **plt**

%matplotlib inline

<https://www.kaggle.com/gsutters/the-human-freedom-index>

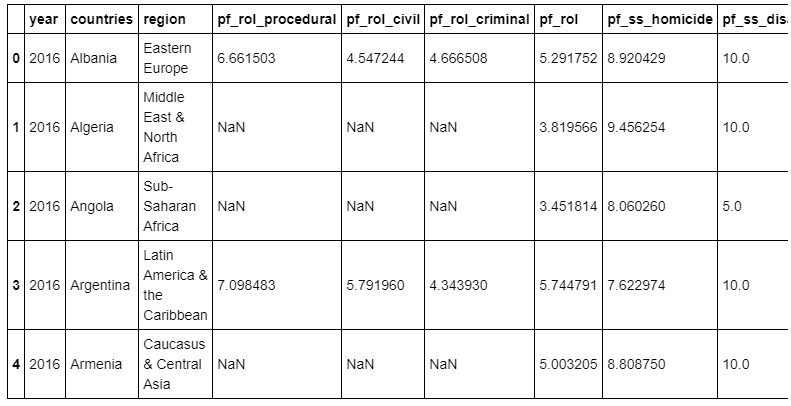
data = pd.read\_csv('data/hfi.csv', sep=',')

data.shape

OUT: (1458, 123)

data.drop(['ISO\_code'], axis='columns', inplace=**True**)

data.head()



data.isnull().sum()

OUT: year 0

countries 0

region 0

pf\_rol\_procedural 578

pf\_rol\_civil 578

pf\_rol\_criminal 578

pf\_rol 80

pf\_ss\_homicide 80

pf\_ss\_disappearances\_disap 89

pf\_ss\_disappearances\_violent 80

pf\_ss\_disappearances\_organized 179

pf\_ss\_disappearances\_fatalities 80

pf\_ss\_disappearances\_injuries 80

pf\_ss\_disappearances 80

pf\_ss\_women\_fgm 172

pf\_ss\_women\_missing 120

pf\_ss\_women\_inheritance\_widows 541

pf\_ss\_women\_inheritance\_daughters 541

pf\_ss\_women\_inheritance 119

pf\_ss\_women 100

pf\_ss 80

pf\_movement\_domestic 98

pf\_movement\_foreign 98

pf\_movement\_women 141

pf\_movement 80

pf\_religion\_estop\_establish 731

pf\_religion\_estop\_operate 731

pf\_religion\_estop 329

pf\_religion\_harassment 94

pf\_religion\_restrictions 94

...

ef\_trade\_black 87

ef\_trade\_movement\_foreign 164

ef\_trade\_movement\_capital 89

ef\_trade\_movement\_visit 85

ef\_trade\_movement 80

ef\_trade 81

ef\_regulation\_credit\_ownership 172

ef\_regulation\_credit\_private 72

ef\_regulation\_credit\_interest 100

ef\_regulation\_credit 80

ef\_regulation\_labor\_minwage 91

ef\_regulation\_labor\_firing 171

ef\_regulation\_labor\_bargain 170

ef\_regulation\_labor\_hours 88

ef\_regulation\_labor\_dismissal 110

ef\_regulation\_labor\_conscription 81

ef\_regulation\_labor 84

ef\_regulation\_business\_adm 169

ef\_regulation\_business\_bureaucracy 102

ef\_regulation\_business\_start 90

ef\_regulation\_business\_bribes 175

ef\_regulation\_business\_licensing 101

ef\_regulation\_business\_compliance 90

ef\_regulation\_business 84

ef\_regulation 80

ef\_score 80

ef\_rank 80

hf\_score 80

hf\_rank 80

hf\_quartile 80

Length: 122, dtype: int64

data.dtypes

OUT: year int64

countries object

region object

pf\_rol\_procedural float64

pf\_rol\_civil float64

pf\_rol\_criminal float64

pf\_rol float64

pf\_ss\_homicide float64

pf\_ss\_disappearances\_disap float64

pf\_ss\_disappearances\_violent float64

pf\_ss\_disappearances\_organized float64

pf\_ss\_disappearances\_fatalities float64

pf\_ss\_disappearances\_injuries float64

pf\_ss\_disappearances float64

pf\_ss\_women\_fgm float64

pf\_ss\_women\_missing float64

pf\_ss\_women\_inheritance\_widows float64

pf\_ss\_women\_inheritance\_daughters float64

pf\_ss\_women\_inheritance float64

pf\_ss\_women float64

pf\_ss float64

pf\_movement\_domestic float64

pf\_movement\_foreign float64

pf\_movement\_women float64

pf\_movement float64

pf\_religion\_estop\_establish float64

pf\_religion\_estop\_operate float64

pf\_religion\_estop float64

pf\_religion\_harassment float64

pf\_religion\_restrictions float64

...

ef\_trade\_black float64

ef\_trade\_movement\_foreign float64

ef\_trade\_movement\_capital float64

ef\_trade\_movement\_visit float64

ef\_trade\_movement float64

ef\_trade float64

ef\_regulation\_credit\_ownership float64

ef\_regulation\_credit\_private float64

ef\_regulation\_credit\_interest float64

ef\_regulation\_credit float64

ef\_regulation\_labor\_minwage float64

ef\_regulation\_labor\_firing float64

ef\_regulation\_labor\_bargain float64

ef\_regulation\_labor\_hours float64

ef\_regulation\_labor\_dismissal float64

ef\_regulation\_labor\_conscription float64

ef\_regulation\_labor float64

ef\_regulation\_business\_adm float64

ef\_regulation\_business\_bureaucracy float64

ef\_regulation\_business\_start float64

ef\_regulation\_business\_bribes float64

ef\_regulation\_business\_licensing float64

ef\_regulation\_business\_compliance float64

ef\_regulation\_business float64

ef\_regulation float64

ef\_score float64

ef\_rank float64

hf\_score float64

hf\_rank float64

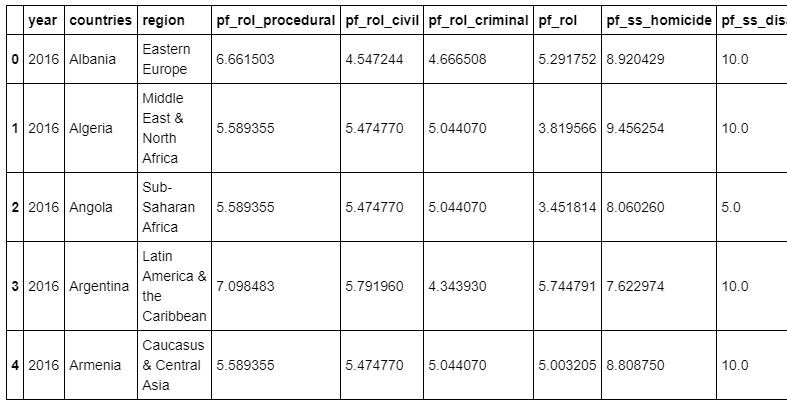
hf\_quartile float64

Length: 122, dtype: object

Замена пустых числовых значений на среднее их столбца

data=data.fillna(data.mean())

data.head()



Label Encoding

**from** **sklearn.preprocessing** **import** LabelEncoder, OneHotEncoder

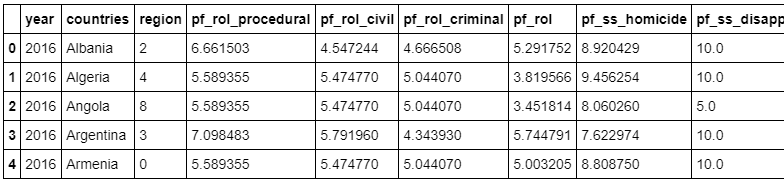
le = LabelEncoder()

le.fit(data.region)

data['region'] = le.transform(data.region)

data.head()

len(data['region'].unique())



OUT: 10

le.inverse\_transform([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

OUT: array(['Caucasus & Central Asia', 'East Asia', 'Eastern Europe',

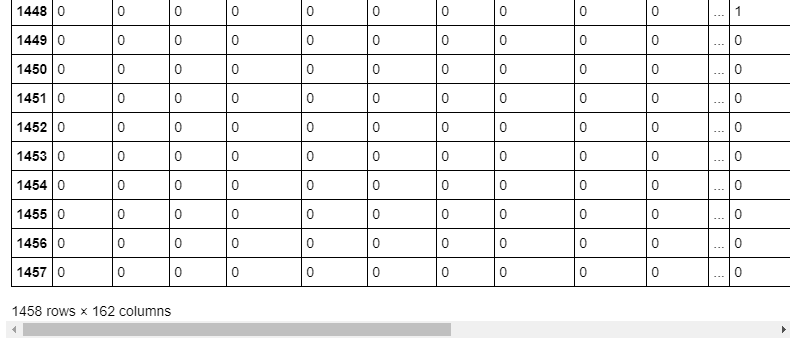
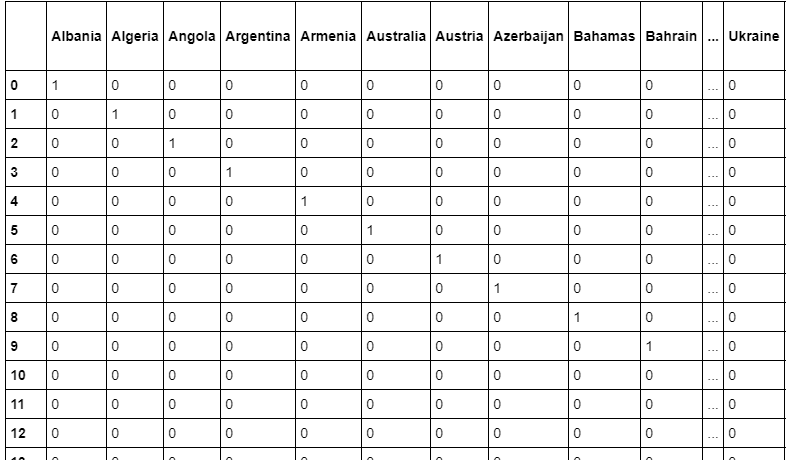
'Latin America & the Caribbean', 'Middle East & North Africa',

'North America', 'Oceania', 'South Asia', 'Sub-Saharan Africa',

'Western Europe'], dtype=object)

One Hot Encoding

pd.get\_dummies(data.countries)



MinMax масштабирование

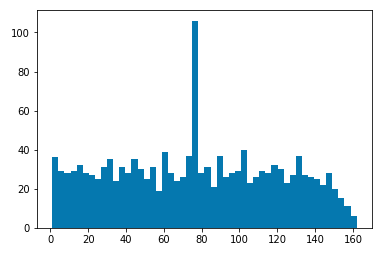
**from** **sklearn.preprocessing** **import** MinMaxScaler, StandardScaler, Normalizer

sc1 = MinMaxScaler()

sc1\_data = sc1.fit\_transform(data[['ef\_rank']])

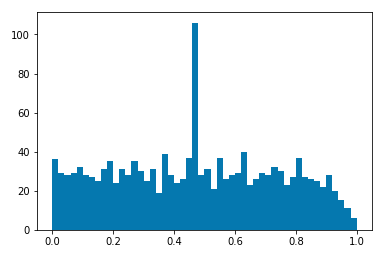
plt.hist(data['ef\_rank'], 50)

plt.show()



plt.hist(sc1\_data, 50)

plt.show()



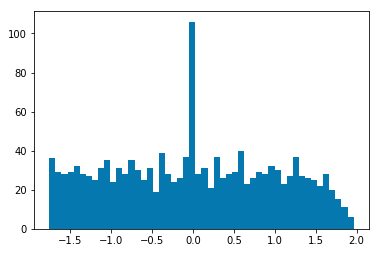
Масштабирование данных на основе Z-оценки

sc2 = StandardScaler()

sc2\_data = sc2.fit\_transform(data[['ef\_rank']])

plt.hist(sc2\_data, 50)

plt.show()



Нормализация данных

sc3 = Normalizer()

sc3\_data = sc3.fit\_transform(data[['ef\_rank']])

plt.hist(sc3\_data, 50)

plt.show()

