Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Лабораторная работа №3
по дисциплине «Технология машинного обучения» на тему:

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

Выполнил:
студент группы № ИУ5-62
Чернышев Павел
подпись, дата

Проверил:
Ю.Е. Гапанюк
подпись, дата

2020 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 решить следующие задачи:

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

In [2]:

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

Загрузка и первичный анализ данных

In [3]:

```
data = pd.read_csv('winemag-data-130k-v2.csv')
data.head()
```

Out[3]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	vari
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	Wh Ble
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portugu F
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot G
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesl
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot N

In [4]:

```
data.shape
```

Out[4]:

(129971, 14)

In [5]:

```
data.dtypes
```

Out[5]:

```
Unnamed: 0             int64
country               object
description           object
designation           object
points                 int64
price                float64
province              object
```

```
region_1                object
region_2                object
taster_name             object
taster_twitter_handle   object
title                   object
variety                 object
winery                  object
dtype: object
```

In [6]:

```
data.isnull().sum()
```

Out[6]:

```
Unnamed: 0                 0
country                   63
description                0
designation            37465
points                     0
price                   8996
province                  63
region_1                21247
region_2                79460
taster_name             26244
taster_twitter_handle   31213
title                       0
variety                     1
winery                      0
dtype: int64
```

In [7]:

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

```
Всего строк: 129971
```

1. 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]:

```
((129971, 14), (129971, 5))
```

In [9]:

```
data_new_2 = data.dropna(axis=0, how='any')
(data.shape, data_new_2.shape)
```

Out[9]:

```
((129971, 14), (22387, 14))
```

1.2. "Внедрение значений" - импьютация (imputation)

1.2.1. Обработка пропусков в числовых данных

In [10]:

```
num_cols = []
for col in data.columns:
    # Количество пустых значений
```

```
    # 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 price. Тип данных float64. Количество пустых значений 8996, 6.92%.

In [11]:

```
data_num = data[num_cols]
data_num
```

Out[11]:

	price
0	NaN
1	15.0
2	14.0
3	13.0
4	65.0
5	15.0
6	16.0
7	24.0
8	12.0
9	27.0
10	19.0
11	30.0
12	34.0
13	NaN
14	12.0
15	24.0
16	30.0
17	13.0
18	28.0
19	32.0
20	23.0
21	20.0
22	19.0
23	22.0
24	35.0
25	69.0
26	13.0
27	10.0
28	17.0
29	16.0
...	...
129941	20.0
129942	35.0
129943	29.0
129944	25.0
129945	20.0
129946	17.0

	price
129947	20.0
129948	43.0
129949	35.0
129950	35.0
129951	30.0
129952	22.0
129953	25.0
129954	15.0
129955	40.0
129956	19.0
129957	17.0
129958	35.0
129959	57.0
129960	48.0
129961	30.0
129962	40.0
129963	20.0
129964	NaN
129965	28.0
129966	28.0
129967	75.0
129968	30.0
129969	32.0
129970	21.0

129971 rows × 1 columns

In [12]:

```
data[data['price'].isnull()]
```

Out[12]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nico Vulk
13	13	Italy	This is dominated by oak and oak-driven aromas...	Rosso	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	N S 201
30	30	France	Red cherry fruit comes laced with light tannin...	Nouveau	86	NaN	Beaujolais	Beaujolais-Villages	NaN	Roger Voss	@vossroger	Doma Madc I (Bea
31	31	Italy	Merlot and Nero d'Avola form the base for this...	Calanìca Nero d'Avola-Merlot	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Sa 2010 (Nero
32	32	Italy	Part of the extended Calanìca series, this Gri...	Calanìca Grillo-Viognier	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Sa 2011 (Grillo
50	50	Italy	This blend of Nero d'Avola and Syrah opens wit...	Scialo	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	V A Canic Sci

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	Co
54	54	Italy	A blend of Nero d'Avola and Nerello Mascalese,...	Rosso	85	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Ro
79	79	Portugal	Grown on the sandy soil of Tejo, the wine is t...	Bridão	86	NaN	Tejo	NaN	NaN	Roger Voss	@vossroger	Coc do 201
137	137	South Africa	This is great Chenin Blanc, wood fermented but...	Hope Marguerite	90	NaN	Walker Bay	NaN	NaN	Roger Voss	@vossroger	B 20 M Cher
159	159	Italy	Intense aromas of ripe red berry, menthol, esp...	Filo di Seta	91	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	F 201 (
163	163	France	Produced from vineyards donated to a charitabl...	Hospices Civils de Romanèche Thurins	91	NaN	Beaujolais	Moulin-à-Vent	NaN	Roger Voss	@vossroger	E H F
182	182	Italy	Loaded with bold, ripe fruit, exotic spice and...	M. Vigna	88	NaN	Tuscany	Brunello di Montalcino	NaN	NaN	NaN	Arder (Br
194	194	Italy	Here's a traditional Chianti Classico with pre...	NaN	87	NaN	Tuscany	Chianti Classico	NaN	NaN	NaN	Camp 200
200	200	Italy	Aromas of mature black-skinned berry, tobacco ...	Riserva	90	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	T Se (Br
222	222	Italy	Chopped herb, forest floor, leather, espresso ...	Trentennale	90	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	Tale Tre (Br M
223	223	Italy	Bright and creamy, this savory white offers ar...	NaN	90	NaN	Northeastern Italy	Alto Adige	NaN	Kerin O'Keefe	@kerinokeefe	Ter Pinc (Alt
285	285	Austria	This is a very aromatic wine that's rich, spic...	Steiner Kögl Erste Lage	92	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	S Undl Stei Er
288	288	Austria	While it feels rich and round, this is also a ...	Kremser Gebling Erste Lage	92	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	Jose 2011 Gebli L
290	290	France	This is a wine that has great potential—you ca...	NaN	92	NaN	Bordeaux	Saint-Estèphe	NaN	Roger Voss	@vossroger	Lafor 20
291	291	Italy	This powerful Sagrantino opens with aromas of ...	NaN	92	NaN	Central Italy	Sagrantino di Montefalco	NaN	Kerin O'Keefe	@kerinokeefe	Dior Sagr Mo
299	299	Italy	Pinay is a tight and defined red blend from no...	Pinay	87	NaN	Piedmont	Langhe	NaN	NaN	NaN	Attilio 20 Red
302	302	Italy	This fresh Nebbiolo-based Roero from Cantina	NaN	87	NaN	Piedmont	Roero	NaN	NaN	NaN	Ca Nebbi

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
304	304	Italy	This aged expression from the Nizza subzone of...	Vignassa	87	NaN	Piedmont	Barbera d'Asti Superiore Nizza	NaN	NaN	NaN	Ca Ghe V
313	313	Italy	Roero Rabino is accented by sweet touches of c...	NaN	87	NaN	Piedmont	Roero	NaN	NaN	NaN	Rab
316	316	France	An intriguing blend of Sauvignon Blanc and Mus...	Les Amants Mont-Pérat	86	NaN	Bordeaux	Bordeaux Blanc	NaN	Roger Voss	@vossroger	Mo 2 Amar
317	317	France	Soft, ripe, dominated by peaches, pears and a ...	NaN	86	NaN	Bordeaux	Bordeaux Blanc	NaN	Roger Voss	@vossroger	Châ de Mi E
377	377	Greece	A fresh, tangy, vibrant wine that has ripe gre...	Vigneto Massoni	88	NaN	Naoussa	NaN	NaN	Roger Voss	@vossroger	Zen (N
395	395	Italy	This is the first vintage of this blend of Cha...	Rylint	87	NaN	Northeastern Italy	Collio	NaN	Roger Voss	@vossroger	Fo 20 Whit
417	417	Italy	This is a sweet style of Prosecco with a lumin...	NaN	85	NaN	Veneto	Prosecco del Veneto	NaN	NaN	NaN	Pizz Pros
418	418	Austria	Using many biodynamic practices, Werner Michli...	Biokult Zweigelt Pinot Noir	85	NaN	Burgenland	NaN	NaN	Roger Voss	@vossroger	Mich Zweig No
...	...	...	...	...	...	...	...	...	...	...	...	
129371	129371	Austria	A perfumed wine from the Krems Vineyards, with...	Donau Riesling	87	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	Fo 201
129383	129383	France	The entry-level wine from Château d'Esclans, t...	Whispering Angel	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	d Wh Ang
129384	129384	France	This is a textured wine that's rich and round....	Pétale de Rose	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Ch l'Evêc F
129387	129387	Italy	This is a broad, round Prosecco that doesn't o...	Extra Dry	85	NaN	Veneto	Prosecco del Veneto	NaN	NaN	NaN	Far E (Pros
129388	129388	Argentina	Slightly tart and like rhubarb at first, but i...	Aduentus Classic	85	NaN	Mendoza Province	Mendoza	NaN	Michael Schachner	@wineschach	Anti A Cla (M
129395	129395	Italy	This Prosecco Extra Dry has full notes of almo...	Extra Dry	85	NaN	Veneto	Prosecco di Valdobbiadene	NaN	NaN	NaN	La Frati Dry (F
129432	129432	Italy	Earthy aromas of wild berry, underbrush,	Sorano	89	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Asch

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
			leath...									
129469	129469	France	Alain Thiénot's non vintage is ripe, rich and ...	Brut	88	NaN	Champagne	Champagne	NaN	Roger Voss	@vossroger	Th (Cha
129488	129488	Portugal	A spicy, ripe wine that is part of the second ...	DouRosa	88	NaN	Douro	NaN	NaN	Roger Voss	@vossroger	Qui Ro DouR
129512	129512	Italy	This opens with aroma suggesting raw oak and t...	Gibin Riserva	88	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Gem Gibi
129519	129519	Portugal	The 12-months of oak aging has given this wine...	Adega de Pegões	88	NaN	Península de Setúbal	NaN	NaN	Roger Voss	@vossroger	Coc Ag Sar de F
129525	129525	France	This is a powerful, big-hearted Malbec, full o...	Pont du Diable	89	NaN	France Other	Vin de France	NaN	Roger Voss	@vossroger	Lion & (Diable
129552	129552	Italy	Fragrant blue flower, rose, new leather, cake ...	del Comune di Serralunga d'Alba	93	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Rive del Co Se
129610	129610	Italy	La Fagiana is a bright, ruby-colored wine with...	La Fagiana	84	NaN	Tuscany	Toscana	NaN	NaN	NaN	F San Fagi
129638	129638	France	Very rich and packed with yellow and tropical ...	Moulin des Dames	90	NaN	Southwest France	Bergerac Sec	NaN	Roger Voss	@vossroger	Châ des 201 des
129684	129684	France	This concentrated, full-bodied wine has fine, ...	Les Baudes Premier Cru	93	NaN	Burgundy	Chambolle-Musigny	NaN	Roger Voss	@vossroger	2 Prem
129696	129696	France	Soft, fragrant wine, layering raspberry and re...	Côte de Rouffach	87	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	Re 2011 P
129721	129721	Italy	Made with 60% Carricante and 40% Catarratto, t...	Bianco	90	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Tenu Gor Bianc
129758	129758	France	This is a barely pink rosé, perfect to reflect...	Les Voiles de Saint-Tropez	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Les Vign la P d
129759	129759	France	Soft and ripe, it is gently fruity with light ...	Cuvée G	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Les V G 2013 Ros
129764	129764	France	Only lightly spicy, it has a smooth texture an...	NaN	87	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	Pie 20 Gris
129787	129787	France	Soft, round and perfumed, this has rich red fr...	NaN	89	NaN	Burgundy	Volnay	NaN	Roger Voss	@vossroger	Poull et
129794	129794	Spain	Attractive and fruity on the front end, this h...	Secreto Reserva	89	NaN	Northern Spain	Ribera del Duero	NaN	Michael Schachner	@wineschach	Viñ 2006 (R

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
129807	129807	Italy	Here is yet another Roero Riserva from the suc...	Roche d'Ampsej Riserva	89	NaN	Piedmont	Roero	NaN	NaN	NaN	C 200 (R
129813	129813	Italy	Here's a more informal and approachable expres...	Chirlet	89	NaN	Piedmont	Barolo	NaN	NaN	NaN	Scale
129844	129844	Italy	Doga delle Clavule is a neutral, mineral-drive...	Doga delle Clavule	86	NaN	Tuscany	Morellino di Scansano	NaN	NaN	NaN	Capa Do (More
129860	129860	Portugal	This rich wine has a firm structure as well as...	Pacheca Superior	90	NaN	Douro	NaN	NaN	Roger Voss	@vossroger	C Pache Supe
129863	129863	Portugal	This mature wine that has 50% Touriga Nacional...	Reserva	90	NaN	Dão	NaN	NaN	Roger Voss	@vossroger	Se 2011 R
129893	129893	Italy	Aromas of passion fruit, hay and a vegetal not...	Corte Menini	91	NaN	Veneto	Soave Classico	NaN	Kerin O'Keefe	@kerinokeefe	Le M 20 Menin (
129964	129964	France	Initially quite muted, this wine slowly develo...	Domaine Saint-Rémy Herrenweg	90	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	Ehrh Sai

8996 rows × 14 columns

In [13]:

```
flt_index = data[data['price'].isnull()].index
flt_index
```

Out[13]:

```
Int64Index([    0,    13,    30,    31,    32,    50,    54,    79,
              137,   159,
            ...
            129764, 129787, 129794, 129807, 129813, 129844, 129860, 129863,
            129893, 129964],
           dtype='int64', length=8996)
```

In [14]:

```
data[data.index.isin(flt_index)]
```

Out[14]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nico Vulk
13	13	Italy	This is dominated by oak and oak-driven aromas...	Rosso	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	S 201
30	30	France	Red cherry fruit comes laced with light tannin...	Nouveau	86	NaN	Beaujolais	Beaujolais-Villages	NaN	Roger Voss	@vossroger	Doma Madc (Bea
			Merlot and	Calanìca			Sicily &					Sa

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
31	31	Italy	form the base for this...	Nero d'Avola-Merlot	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	2010 Nero
32	32	Italy	Part of the extended Calanìca series, this Gri...	Calanìca Grillo-Viognier	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Sa 2011 Grillo
50	50	Italy	This blend of Nero d'Avola and Syrah opens wit...	Scialo	86	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Canic Sci
54	54	Italy	A blend of Nero d'Avola and Nerello Mascalese,...	Rosso	85	NaN	Sicily & Sardinia	Sicilia	NaN	NaN	NaN	Co Rc
79	79	Portugal	Grown on the sandy soil of Tejo, the wine is t...	Bridão	86	NaN	Tejo	NaN	NaN	Roger Voss	@vossroger	Coc do 201
137	137	South Africa	This is great Chenin Blanc, wood fermented but...	Hope Marguerite	90	NaN	Walker Bay	NaN	NaN	Roger Voss	@vossroger	B 20 M Cher
159	159	Italy	Intense aromas of ripe red berry, menthol, esp...	Filo di Seta	91	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	F 20 (
163	163	France	Produced from vineyards donated to a charitabl...	Hospices Civils de Romanèche Thurins	91	NaN	Beaujolais	Moulin-à-Vent	NaN	Roger Voss	@vossroger	E H F
182	182	Italy	Loaded with bold, ripe fruit, exotic spice and...	M. Vigna	88	NaN	Tuscany	Brunello di Montalcino	NaN	NaN	NaN	Arder (Br
194	194	Italy	Here's a traditional Chianti Classico with pre...	NaN	87	NaN	Tuscany	Chianti Classico	NaN	NaN	NaN	Camp 200
200	200	Italy	Aromas of mature black-skinned berry, tobacco ...	Riserva	90	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	T Se (Br
222	222	Italy	Chopped herb, forest floor, leather, espresso ...	Trentennale	90	NaN	Tuscany	Brunello di Montalcino	NaN	Kerin O'Keefe	@kerinokeefe	Tale Tre (Br M
223	223	Italy	Bright and creamy, this savory white offers ar...	NaN	90	NaN	Northeastern Italy	Alto Adige	NaN	Kerin O'Keefe	@kerinokeefe	Ter Pinc (Alt
285	285	Austria	This is a very aromatic wine that's rich, spic...	Steiner Kögl Erste Lage	92	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	S Undl Stei Er
288	288	Austria	While it feels rich and round, this is also a ...	Kremser Gebling Erste Lage	92	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	Jose 2011 Gebli L
290	290	France	This is a wine that has great potential—you ca...	NaN	92	NaN	Bordeaux	Saint-Estèphe	NaN	Roger Voss	@vossroger	Lafor 20
			This powerful									

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
291	291	Italy	Sagrantino opens with aromas of ...	NaN	92	NaN	Central Italy	Sagrantino di Montefalco	NaN	Kerin O'Keefe	@kerinokeefe	Dior Sagr Mo
299	299	Italy	Pinay is a tight and defined red blend from no...	Pinay	87	NaN	Piedmont	Langhe	NaN	NaN	NaN	Attilio 20 Red
302	302	Italy	This fresh Nebbiolo-based Roero from Cantina d...	NaN	87	NaN	Piedmont	Roero	NaN	NaN	NaN	Ca Nebbi
304	304	Italy	This aged expression from the Nizza subzone of...	Vignassa	87	NaN	Piedmont	Barbera d'Asti Superiore Nizza	NaN	NaN	NaN	Ca Ghe
313	313	Italy	Roero Rabino is accented by sweet touches of c...	NaN	87	NaN	Piedmont	Roero	NaN	NaN	NaN	Rab
316	316	France	An intriguing blend of Sauvignon Blanc and Mus...	Les Amants Mont-Pérat	86	NaN	Bordeaux	Bordeaux Blanc	NaN	Roger Voss	@vossroger	Mo 2 Amar
317	317	France	Soft, ripe, dominated by peaches, pears and a ...	NaN	86	NaN	Bordeaux	Bordeaux Blanc	NaN	Roger Voss	@vossroger	Châte de Mi E
377	377	Greece	A fresh, tangy, vibrant wine that has ripe gre...	Vigneto Massoni	88	NaN	Naoussa	NaN	NaN	Roger Voss	@vossroger	Zen (N
395	395	Italy	This is the first vintage of this blend of Cha...	Rylint	87	NaN	Northeastern Italy	Collio	NaN	Roger Voss	@vossroger	Fo 20 White
417	417	Italy	This is a sweet style of Prosecco with a lumin...	NaN	85	NaN	Veneto	Prosecco del Veneto	NaN	NaN	NaN	Pizz Pros
418	418	Austria	Using many biodynamic practices, Werner Michli...	Biokult Zweigelt Pinot Noir	85	NaN	Burgenland	NaN	NaN	Roger Voss	@vossroger	Mich Zweig No
...	...	...	...	...	...	...	...	...	...	...	...	
129371	129371	Austria	A perfumed wine from the Krems Vineyards, with...	Donau Riesling	87	NaN	Kremstal	NaN	NaN	Roger Voss	@vossroger	Fo 201
129383	129383	France	The entry-level wine from Château d'Esclans, t...	Whispering Angel	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	d Wh Ang
129384	129384	France	This is a textured wine that's rich and round....	Pétale de Rose	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Ch l'Evêc F
129387	129387	Italy	This is a broad, round Prosecco that doesn't o...	Extra Dry	85	NaN	Veneto	Prosecco del Veneto	NaN	NaN	NaN	Far E (Pros
			Slightly tart									Anti

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
129388	129388	Argentina	and like rhubarb at first, but i...	Adventus Classic	85	NaN	Mendoza Province	Mendoza	NaN	Michael Schachner	@wineschach	A Cla (M
129395	129395	Italy	This Prosecco Extra Dry has full notes of almo...	Extra Dry	85	NaN	Veneto	Prosecco di Valdobbiadene	NaN	NaN	NaN	La Frati Dry (F
129432	129432	Italy	Earthy aromas of wild berry, underbrush, leath...	Sorano	89	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Asch
129469	129469	France	Alain Thiénot's non vintage is ripe, rich and ...	Brut	88	NaN	Champagne	Champagne	NaN	Roger Voss	@vossroger	Th (Cha
129488	129488	Portugal	A spicy, ripe wine that is part of the second ...	DouRosa	88	NaN	Douro	NaN	NaN	Roger Voss	@vossroger	Qui Ro DouR
129512	129512	Italy	This opens with aroma suggesting raw oak and t...	Gibin Riserva	88	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Gem Gibi
129519	129519	Portugal	The 12-months of oak aging has given this wine...	Adega de Pegões	88	NaN	Península de Setúbal	NaN	NaN	Roger Voss	@vossroger	Coc Ag San de F
129525	129525	France	This is a powerful, big-hearted Malbec, full o...	Pont du Diable	89	NaN	France Other	Vin de France	NaN	Roger Voss	@vossroger	Lion & Diabl
129552	129552	Italy	Fragrant blue flower, rose, new leather, cake ...	del Comune di Serralunga d'Alba	93	NaN	Piedmont	Barolo	NaN	Kerin O'Keefe	@kerinokeefe	Rive del Co Se
129610	129610	Italy	La Fagiana is a bright, ruby-colored wine with...	La Fagiana	84	NaN	Tuscany	Toscana	NaN	NaN	NaN	F San Fagi
129638	129638	France	Very rich and packed with yellow and tropical ...	Moulin des Dames	90	NaN	Southwest France	Bergerac Sec	NaN	Roger Voss	@vossroger	Châ des 201 des
129684	129684	France	This concentrated, full-bodied wine has fine, ...	Les Baudes Premier Cru	93	NaN	Burgundy	Chambolle-Musigny	NaN	Roger Voss	@vossroger	2 Prem
129696	129696	France	Soft, fragrant wine, layering raspberry and re...	Côte de Rouffach	87	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	Re 2011 P
129721	129721	Italy	Made with 60% Carricante and 40% Catarratto, t...	Bianco	90	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Tenu Gor Bianc
129758	129758	France	This is a barely pink rosé, perfect to reflect...	Les Voiles de Saint-Tropez	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Les Vign la P d
129759	129759	France	Soft and ripe, it is gently fruity with light ...	Cuvée G	87	NaN	Provence	Côtes de Provence	NaN	Roger Voss	@vossroger	Les V G 2013 Ros
129764	129764	France	Only lightly spicy, it has a	NaN	87	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	Pie 20

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	
129787	129787	France	Soft, round and perfumed, this has rich red fr...	NaN	89	NaN	Burgundy	Volnay	NaN	Roger Voss	@vossroger	I Poulle et I
129794	129794	Spain	Attractive and fruity on the front end, this h...	Secreto Reserva	89	NaN	Northern Spain	Ribera del Duero	NaN	Michael Schachner	@wineschach	Viñ 2006 (R
129807	129807	Italy	Here is yet another Roero Riserva from the suc...	Roche d'Ampsej Riserva	89	NaN	Piedmont	Roero	NaN	NaN	NaN	C 200 R
129813	129813	Italy	Here's a more informal and approachable expres...	Chirlet	89	NaN	Piedmont	Barolo	NaN	NaN	NaN	Scale
129844	129844	Italy	Doga delle Clavule is a neutral, mineral-drive...	Doga delle Clavule	86	NaN	Tuscany	Morellino di Scansano	NaN	NaN	NaN	Capa D (More
129860	129860	Portugal	This rich wine has a firm structure as well as...	Pacheca Superior	90	NaN	Douro	NaN	NaN	Roger Voss	@vossroger	C Pache Supe
129863	129863	Portugal	This mature wine that has 50% Touriga Nacional...	Reserva	90	NaN	Dão	NaN	NaN	Roger Voss	@vossroger	Se 2011 R
129893	129893	Italy	Aromas of passion fruit, hay and a vegetal not...	Corte Menini	91	NaN	Veneto	Soave Classico	NaN	Kerin O'Keefe	@kerinokeefe	Le M 20 Menin
129964	129964	France	Initially quite muted, this wine slowly develo...	Domaine Saint-Rémy Herrenweg	90	NaN	Alsace	Alsace	NaN	Roger Voss	@vossroger	I Ehrh I Sai

8996 rows × 14 columns

In [15]:

```
data_num[data_num.index.isin(flt_index)]['price']
```

Out[15]:

```
0      NaN
13     NaN
30     NaN
31     NaN
32     NaN
50     NaN
54     NaN
79     NaN
137    NaN
159    NaN
163    NaN
182    NaN
194    NaN
200    NaN
222    NaN
223    NaN
285    NaN
288    NaN
290    NaN
291    NaN
299    NaN
302    NaN
304    NaN
```

313 NaN
316 NaN
317 NaN
377 NaN
395 NaN
417 NaN
418 NaN
..
129371 NaN
129383 NaN
129384 NaN
129387 NaN
129388 NaN
129395 NaN
129432 NaN
129469 NaN
129488 NaN
129512 NaN
129519 NaN
129525 NaN
129552 NaN
129610 NaN
129638 NaN
129684 NaN
129696 NaN
129721 NaN
129758 NaN
129759 NaN
129764 NaN
129787 NaN
129794 NaN
129807 NaN
129813 NaN
129844 NaN
129860 NaN
129863 NaN
129893 NaN
129964 NaN
Name: price, Length: 8996, dtype: float64

In [16]:

```
data_num_price = data_num[['price']]
data_num_price.head()
```

Out[16]:

	price
0	NaN
1	15.0
2	14.0
3	13.0
4	65.0

In [17]:

```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
```

In [18]:

```
# Фильтр для проверки заполнения пустых значений
indicator = MissingIndicator()
mask_missing_values_only = indicator.fit_transform(data_num_price)
mask_missing_values_only
```

Out[18]:

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

In [19]:

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

In [20]:

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

In [21]:

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

Out[21]:

```
('mean', array([35.36338913, 35.36338913, 35.36338913, ..., 35.36338913,
        35.36338913, 35.36338913]))
```

In [22]:

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

Out[22]:

```
('median', array([25., 25., 25., ..., 25., 25., 25.]))
```

In [23]:

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

Out[23]:

```
('most_frequent', array([20., 20., 20., ..., 20., 20., 20.]))
```

In [24]:

```
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]
```

In [25]:

```
test_num_impute_col(data, 'price', strategies[1])
```

Out[25]:

```
('price', 'median', 8996, 25.0, 25.0)
```

1.2.2. Обработка пропусков в категориальных данных

In [26]:

```
# 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, 4)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

```
Колонка country. Тип данных object. Количество пустых значений 63, 0.0485%.
Колонка designation. Тип данных object. Количество пустых значений 37465, 28.8257%.
Колонка province. Тип данных object. Количество пустых значений 63, 0.0485%.
Колонка region_1. Тип данных object. Количество пустых значений 21247, 16.3475%.
Колонка region_2. Тип данных object. Количество пустых значений 79460, 61.1367%.
Колонка taster_name. Тип данных object. Количество пустых значений 26244, 20.1922%.
Колонка taster_twitter_handle. Тип данных object. Количество пустых значений 31213, 24.0154%.
Колонка variety. Тип данных object. Количество пустых значений 1, 0.0008%.
```

In [27]:

```
cat_temp_data = data[['country']]
cat_temp_data.head()
```

Out[27]:

	country
0	Italy
1	Portugal
2	US
3	US
4	US

In [28]:

```
cat_temp_data['country'].unique()
```

Out[28]:

```
array(['Italy', 'Portugal', 'US', 'Spain', 'France', 'Germany',
       'Argentina', 'Chile', 'Australia', 'Austria', 'South Africa',
       'New Zealand', 'Israel', 'Hungary', 'Greece', 'Romania', 'Mexico',
       'Canada', nan, 'Turkey', 'Czech Republic', 'Slovenia',
       'Luxembourg', 'Croatia', 'Georgia', 'Uruguay', 'England',
       'Lebanon', 'Serbia', 'Brazil', 'Moldova', 'Morocco', 'Peru',
       'India', 'Bulgaria', 'Cyprus', 'Armenia', 'Switzerland',
       'Bosnia and Herzegovina', 'Ukraine', 'Slovakia', 'Macedonia',
       'China', 'Egypt'], dtype=object)
```

In [29]:

```
cat_temp_data[cat_temp_data['country'].isnull()].shape
```

Out[29]:

```
(63, 1)
```

In [30]:

```
imp2 = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
data_imp2 = imp2.fit_transform(cat_temp_data)
data_imp2
```

Out[30]:

```
array([['Italy'],
       ['Portugal'],
       ['US'],
       ...,
       ['France'],
       ['France'],
       ['France']], dtype=object)
```

In [31]:

```
np.unique(data_imp2)
```

Out[31]:

```
array(['Argentina', 'Armenia', 'Australia', 'Austria',
       'Bosnia and Herzegovina', 'Brazil', 'Bulgaria', 'Canada', 'Chile',
       'China', 'Croatia', 'Cyprus', 'Czech Republic', 'Egypt', 'England',
       'France', 'Georgia', 'Germany', 'Greece', 'Hungary', 'India',
       'Israel', 'Italy', 'Lebanon', 'Luxembourg', 'Macedonia', 'Mexico',
       'Moldova', 'Morocco', 'New Zealand', 'Peru', 'Portugal', 'Romania',
       'Serbia', 'Slovakia', 'Slovenia', 'South Africa', 'Spain',
       'Switzerland', 'Turkey', 'US', 'Ukraine', 'Uruguay'], dtype=object)
```

In [32]:

```
imp3 = SimpleImputer(missing_values=np.nan, strategy='constant', fill_value='!!!')
data_imp3 = imp3.fit_transform(cat_temp_data)
data_imp3
```

Out[32]:

```
array([['Italy'],
       ['Portugal'],
       ['US'],
       ...,
       ['France'],
       ['France'],
       ['France']], dtype=object)
```

In [33]:

```
np.unique(data_imp3)
```

Out[33]:

```
array(['!!!', 'Argentina', 'Armenia', 'Australia', 'Austria',
       'Bosnia and Herzegovina', 'Brazil', 'Bulgaria', 'Canada', 'Chile',
       'China', 'Croatia', 'Cyprus', 'Czech Republic', 'Egypt', 'England',
       'France', 'Georgia', 'Germany', 'Greece', 'Hungary', 'India',
       'Israel', 'Italy', 'Lebanon', 'Luxembourg', 'Macedonia', 'Mexico',
       'Moldova', 'Morocco', 'New Zealand', 'Peru', 'Portugal', 'Romania',
       'Serbia', 'Slovakia', 'Slovenia', 'South Africa', 'Spain',
       'Switzerland', 'Turkey', 'US', 'Ukraine', 'Uruguay'], dtype=object)
```

In [34]:

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

Out[34]:

63

2. Преобразование категориальных признаков в числовые

In [35]:

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

Out[35]:

	c1
0	Italy
1	Portugal
2	US
3	US
4	US
5	Spain
6	Italy
7	France

	c1
8	Germany
9	France
10	US
11	France
12	US
13	Italy
14	US
15	Germany
16	Argentina
17	Argentina
18	Spain
19	US
20	US
21	US
22	Italy
23	US
24	Italy
25	US
26	Italy
27	Italy
28	Italy
29	US
...	...
129941	US
129942	US
129943	Italy
129944	Israel
129945	US
129946	Germany
129947	Italy
129948	Argentina
129949	US
129950	US
129951	France
129952	US
129953	New Zealand
129954	New Zealand
129955	New Zealand
129956	New Zealand
129957	Spain
129958	New Zealand
129959	France
129960	Portugal
129961	Italy
129962	Italy
129963	Israel
129964	France
129965	France

	c1
129966	Germany
129967	US
129968	France
129969	France
129970	France

129971 rows × 1 columns

2.1. Кодирование категорий целочисленными значениями - label encoding

In [36]:

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

In [37]:

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

In [38]:

```
cat_enc['c1'].unique()
```

Out[38]:

```
array(['Italy', 'Portugal', 'US', 'Spain', 'France', 'Germany',
       'Argentina', 'Chile', 'Australia', 'Austria', 'South Africa',
       'New Zealand', 'Israel', 'Hungary', 'Greece', 'Romania', 'Mexico',
       'Canada', 'Turkey', 'Czech Republic', 'Slovenia', 'Luxembourg',
       'Croatia', 'Georgia', 'Uruguay', 'England', 'Lebanon', 'Serbia',
       'Brazil', 'Moldova', 'Morocco', 'Peru', 'India', 'Bulgaria',
       'Cyprus', 'Armenia', 'Switzerland', 'Bosnia and Herzegovina',
       'Ukraine', 'Slovakia', 'Macedonia', 'China', 'Egypt'], dtype=object)
```

In [39]:

```
np.unique(cat_enc_le)
```

Out[39]:

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
       17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
       34, 35, 36, 37, 38, 39, 40, 41, 42])
```

In [40]:

```
le.inverse_transform([0, 1, 2, 3, 4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
       17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
       34, 35, 36, 37, 38, 39, 40, 41, 42])
```

Out[40]:

```
array(['Argentina', 'Armenia', 'Australia', 'Austria',
       'Bosnia and Herzegovina', 'Brazil', 'Bulgaria', 'Canada', 'Chile',
       'China', 'Croatia', 'Cyprus', 'Czech Republic', 'Egypt', 'England',
       'France', 'Georgia', 'Germany', 'Greece', 'Hungary', 'India',
       'Israel', 'Italy', 'Lebanon', 'Luxembourg', 'Macedonia', 'Mexico',
       'Moldova', 'Morocco', 'New Zealand', 'Peru', 'Portugal', 'Romania',
       'Serbia', 'Slovakia', 'Slovenia', 'South Africa', 'Spain',
       'Switzerland', 'Turkey', 'US', 'Ukraine', 'Uruguay'], dtype=object)
```

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

In [41]:

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

In [42]:

```
cat_enc.shape
```

Out[42]:

```
(129971, 1)
```

In [43]:

```
cat_enc_ohe.shape
```

Out[43]:

```
(129971, 43)
```

In [44]:

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

Out[44]:

```
matrix([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
         0., 0., 0., 0., 0., 0., 1., 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.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.,
         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., 0., 0., 0., 0., 0., 0., 0.,
         0., 0., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 1., 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., 1., 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., 1.,
         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., 0., 0., 0., 0., 0., 0., 0.,
         0., 1., 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., 0., 0., 0., 0., 1.,
         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.]])
```

In [45]:

```
cat_enc.head(10)
```

Out[45]:

	c1
0	Italy
1	Portugal
2	US
3	US
4	US
5	Spain
6	Italy
7	France

8	Germany
9	France

2.3. Pandas get_dummies - быстрый вариант one-hot кодирования

In [46]:

```
pd.get_dummies(cat_enc).head()
```

Out[46]:

	c1_Argentina	c1_Armenia	c1_Australia	c1_Austria	c1_Bosnia and Herzegovina	c1_Brazil	c1_Bulgaria	c1_Canada	c1_Chile	c1_China	...	c1_Serbia	c1_Slovak
0	0	0	0	0	0	0	0	0	0	0	...	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	

5 rows × 43 columns

In [47]:

```
pd.get_dummies(cat_temp_data, dummy_na=True).head()
```

Out[47]:

	country_Argentina	country_Armenia	country_Australia	country_Austria	country_Bosnia and Herzegovina	country_Brazil	country_Bulgaria	country_Canada	country
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	

5 rows × 44 columns

3. Масштабирование данных

In [48]:

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

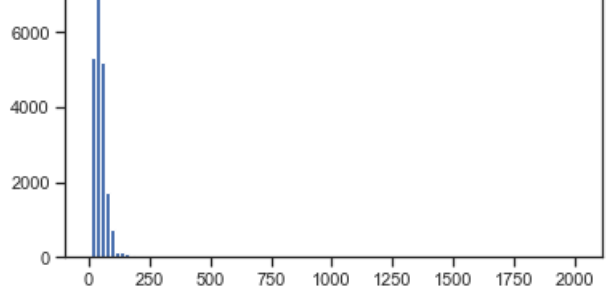
3.1. MinMax масштабирование

In [49]:

```
sc1 = MinMaxScaler()
sc1_data = sc1.fit_transform(data_new_2[['price']])
```

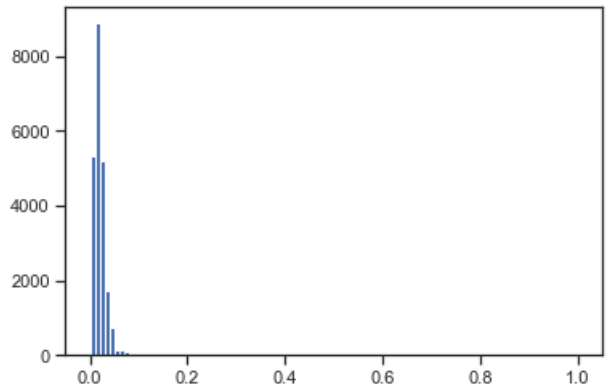
In [50]:

```
plt.hist(data_new_2['price'], 100)
plt.show()
```



In [51]:

```
plt.hist(sc1_data, 100)
plt.show()
```



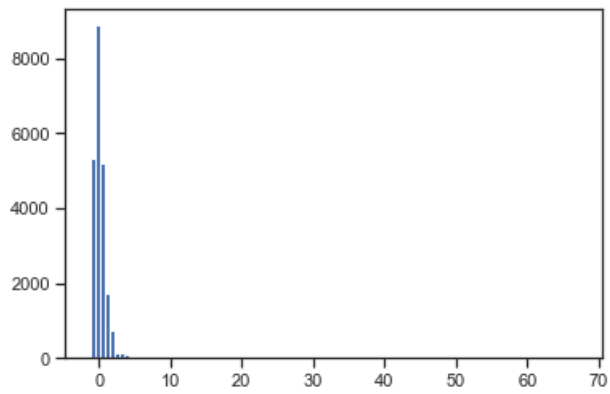
3.2. Масштабирование данных на основе Z-оценки - StandardScaler

In [52]:

```
sc2 = StandardScaler()
sc2_data = sc2.fit_transform(data_new_2[['price']])
```

In [53]:

```
plt.hist(sc2_data, 100)
plt.show()
```



3.3. Нормализация данных

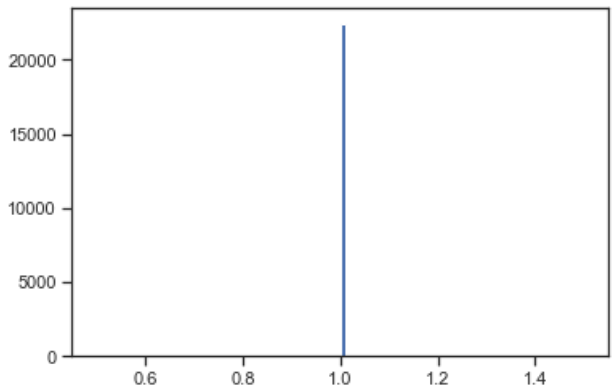
In [54]:

```
sc3 = Normalizer()
sc3_data = sc3.fit_transform(data_new_2[['price']])
```

In [55]:

```
plt.hist(sc3_data, 100)
```

```
plt.show()
```



In [56]:

```
data.isnull().sum()
```

Out[56]:

```
Unnamed: 0                 0
country                   63
description                0
designation            37465
points                     0
price                   8996
province                  63
region_1               21247
region_2               79460
taster_name            26244
taster_twitter_handle  31213
title                      0
variety                    1
winery                     0
dtype: int64
```

In [ ]:

```

```