Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Рубежный контроль №1
по курсу «Методы машинного обучения»
Вариант №6

ИСПОЛНИТЕЛЬ:

Ерохин И.А.
Группа ИУ5-24М

_____

"__"_____2022 г.

Москва 2022

Вариант:

Номер варианта	Номер задачи 1	Номер задачи 2
6	6	26

Дополнительные требования по группам:

Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Выполнение:

Для задачи 1

In [219…
```
# задача 6
# Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с
# использованием метода заполнения средним значением.

data2 = pd.read_csv('houses_to_rent.csv', sep = ',')
```

In [220…
```
data2.head()
```

Out[220…

	Unnamed: 0	city	area	rooms	bathroom	parking spaces	floor	animal	furniture	hoa	rent amount	property tax	fire insurance	total
0	0	1	240	3	3	4	-	acept	furnished	R$0	R$8,000	R$1,000	R$121	R$9,121
1	1	0	64	2	1	1	10	acept	not furnished	R$540	R$820	R$122	R$11	R$1,493
2	2	1	443	5	5	4	3	acept	furnished	R$4,172	R$7,000	R$1,417	R$89	R$12,680
3	3	1	73	2	2	1	12	acept	not furnished	R$700	R$1,250	R$150	R$16	R$2,116
4	4	1	19	1	1	0	-	not acept	not furnished	R$0	R$1,200	R$41	R$16	R$1,257

Для задачи 2

In [230…
```
data = pd.read_csv('WineQT.csv', sep = ',')
```

In [231…
```
data.head()
```

Out[231…

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

Дополнительные требования

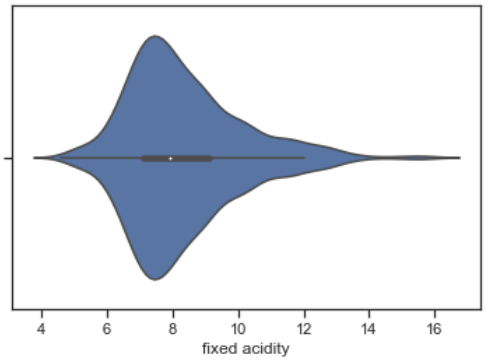
In [234…

```
# Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график
#"Скрипичная диаграмма (violin plot)

sns.violinplot(x = data['fixed acidity'])
```

Out[234…

```
<AxesSubplot:xlabel='fixed acidity'>
```



Задача 1

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения средним значением.

```
# задача 6
# Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с
# использованием метода заполнения средним значением.

data2 = pd.read_csv('houses_to_rent.csv', sep = ',')
```

```
data2.head()
```

	Unnamed: 0	city	area	rooms	bathroom	parking spaces	floor	animal	furniture	hoa	rent amount	property tax	fire insurance	total
0	0	1	240	3	3	4	-	acept	furnished	R$0	R$8,000	R$1,000	R$121	R$9,121
1	1	0	64	2	1	1	10	acept	not furnished	R$540	R$820	R$122	R$11	R$1,493
2	2	1	443	5	5	4	3	acept	furnished	R$4,172	R$7,000	R$1,417	R$89	R$12,680
3	3	1	73	2	2	1	12	acept	not furnished	R$700	R$1,250	R$150	R$16	R$2,116
4	4	1	19	1	1	0	-	not acept	not furnished	R$0	R$1,200	R$41	R$16	R$1,257

```
data2.dtypes
```

```
Unnamed: 0         int64
city               int64
area               int64
rooms              int64
bathroom           int64
parking spaces     int64
floor             object
animal            object
furniture         object
hoa               object
rent amount       object
property tax      object
fire insurance    object
total             object
dtype: object
```

```
# data2['floor'] = data2['floor'].astype('float')
data2['floor'] = pd.to_numeric(data2['floor'], errors = 'coerce')
```

```
def get_loss(some_data):
    for col in some_data.columns:
        #some_data[col] = np.where(some_data[col] == '-', np.nan, some_data[col])
        null_counter = some_data[some_data[col].isnull()].shape[0]
        print("{} : {}".format(col,null_counter))
```

```
get_loss(data2)
```

```
Unnamed: 0 : 0
city : 0
area : 0
rooms : 0
bathroom : 0
parking spaces : 0
floor : 1555
animal : 0
furniture : 0
hoa : 0
rent amount : 0
property tax : 0
fire insurance : 0
total : 0
```

```
data3 = data2.copy()
data3['floor'].fillna(data2['floor'].mean(), inplace = True)
```

```
get_loss(data3)
```

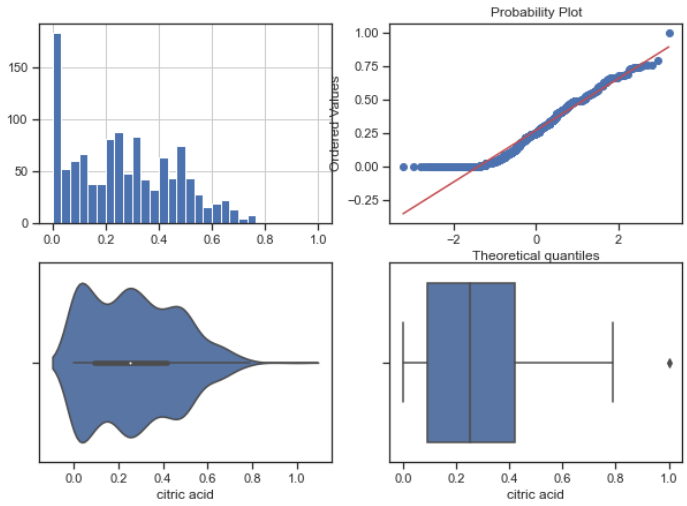
```
Unnamed: 0 : 0
city : 0
area : 0
rooms : 0
bathroom : 0
parking spaces : 0
floor : 0
animal : 0
furniture : 0
hoa : 0
rent amount : 0
property tax : 0
fire insurance : 0
total : 0
```

Задача 2

Для набора данных для одного (произвольного) числового признака проведите обнаружение и замену (найденными верхними и нижними границами) выбросов на основе правила трех сигм.

In [236…

```
# Задача 2б
# Для набора данных для одного (произвольного) числового признака проведите
# обнаружение и замену (найденными верхними и нижними границами) выбросов на основе правила трех сигм.
col_name = 'citric acid'
diagnostic_plots(data, col_name)
```



In [238…

```
# среднее арифм. +\- среднеквадратичное отклонение * 3
k = 3
low = data[col_name].mean() - (k*data[col_name].std())
up = data[col_name].mean() + (k*data[col_name].std())

# Изменение данных
data[col_name] = np.where(data[col_name] > up, up, np.where(data[col_name] < low, low, data[col_name]))
diagnostic_plots(data, col_name)
```