

**Московский государственный технический  
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 13

Выполнил:  
Пермяков С.С.  
группа ИУ5-65Б

Проверил:  
Гапанюк Ю.Е.

Дата: 11.04.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

## **Задание:**

Номер варианта: **1**

Номер задачи: **2**

Номер набора данных, указанного в задаче: **5**

(<https://www.kaggle.com/mohansacharya/graduate-admissions> (файл Admission\_Predict.csv))

### **Задача №2.**

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

## Ход выполнения:

- 1) Загрузил набор данных, просмотрел начало, проверил пропуски и выяснил, что пропуски отсутствуют.

```
import pandas as pd
Execute Cell (Ctrl+Alt+Enter) as np

[1] ✓ 2.7s

df = pd.read_csv('Admission_Predict.csv')
df.columns = df.columns.str.strip()
df.head()

[3] ✓ 0.0s

...
  Serial No.  GRE Score  TOEFL Score  University Rating  SOP  LOR  CGPA  Research  Chance of Admit
0          1         337          118                4    4.5  4.5  9.65          1          0.92
1          2         324          107                4    4.0  4.5  8.87          1          0.76
2          3         316          104                3    3.0  3.5  8.00          1          0.72
3          4         322          110                3    3.5  2.5  8.67          1          0.80
4          5         314          103                2    2.0  3.0  8.21          0          0.65

df.info()

[22] ✓ 0.0s

...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Serial No.            400 non-null   int64
1   GRE Score             400 non-null   float64
2   TOEFL Score           400 non-null   int64
3   University Rating      400 non-null   object
4   SOP                   400 non-null   float64
5   LOR                   400 non-null   float64
6   CGPA                  400 non-null   float64
7   Research               400 non-null   int64
8   Chance of Admit        400 non-null   float64
dtypes: float64(5), int64(3), object(1)
memory usage: 28.3+ KB

print(df.isnull().sum()) # пропусков нет

[18] ✓ 0.0s

...
Serial No.      0
GRE Score       0
TOEFL Score     0
University Rating 0
SOP             0
LOR             0
CGPA            0
Research        0
Chance of Admit 0
dtype: int64
```

- 2) Создал 5% пропусков в колонках 'GRE Score' и 'University Rating' искусственно, предварительно сделав 'University Rating' категориальным.

Так как мы выяснили, что пропусков нет, то создадим 5% пропусков искусственно. Также сделаем University Rating категориальным

```
rating_map = {
    1: 'Very Bad',
    2: 'Bad',
    3: 'Normal',
    4: 'Good',
    5: 'Excellent'
}
df['University Rating'] = df['University Rating'].map(rating_map)
for col in ['GRE Score', 'University Rating']:
    missing_indices = df.sample(frac=0.05, random_state=52).index
    df.loc[missing_indices, col] = np.nan
```

```
df.info() #проверяем создание пропусков
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Serial No.      400 non-null   int64
1   GRE Score       380 non-null   float64
2   TOEFL Score     400 non-null   int64
3   University Rating 380 non-null   object
4   SOP             400 non-null   float64
5   LOR             400 non-null   float64
6   CGPA            400 non-null   float64
7   Research        400 non-null   int64
8   Chance of Admit 400 non-null   float64
dtypes: float64(5), int64(3), object(1)
memory usage: 28.3+ KB
```

- 3) Заменял пропуски в 'GRE Score' и 'University Rating' отсеченным средним и модой соответственно.

```
# Заменяем пропуски в GRE Score усеченным средним
q_05 = df['GRE Score'].dropna().quantile(0.05)
q_95 = df['GRE Score'].dropna().quantile(0.95)
print(f"5-й перцентиль (q_05): {q_05}")
print(f"95-й перцентиль (q_95): {q_95}")
filtr_data = df[(df['GRE Score'] > q_05) & (df['GRE Score'] < q_95)][['GRE Score']]
print(f"Количество значений между 5% и 95% квантилями: {len(filtr_data)}")
filtr_data_mean = filtr_data.mean()
print(f"Отсеченное среднее для GRE Score: {filtr_data_mean}")
df['GRE Score'] = df['GRE Score'].fillna(filtr_data_mean)

# Заменяем пропуски в University Rating модой
mode_rating = df['University Rating'].mode()[0]
print(f"Мода для University Rating: {mode_rating}")
df['University Rating'] = df['University Rating'].fillna(mode_rating)
```

[47]

✓ 0.0s

...

```
5-й перцентиль (q_05): 298.0
95-й перцентиль (q_95): 335.05
Количество значений между 5% и 95% квантилями: 336
Отсеченное среднее для GRE Score: 317.17857142857144
Мода для University Rating: Normal
```

...

```
df.info()
```

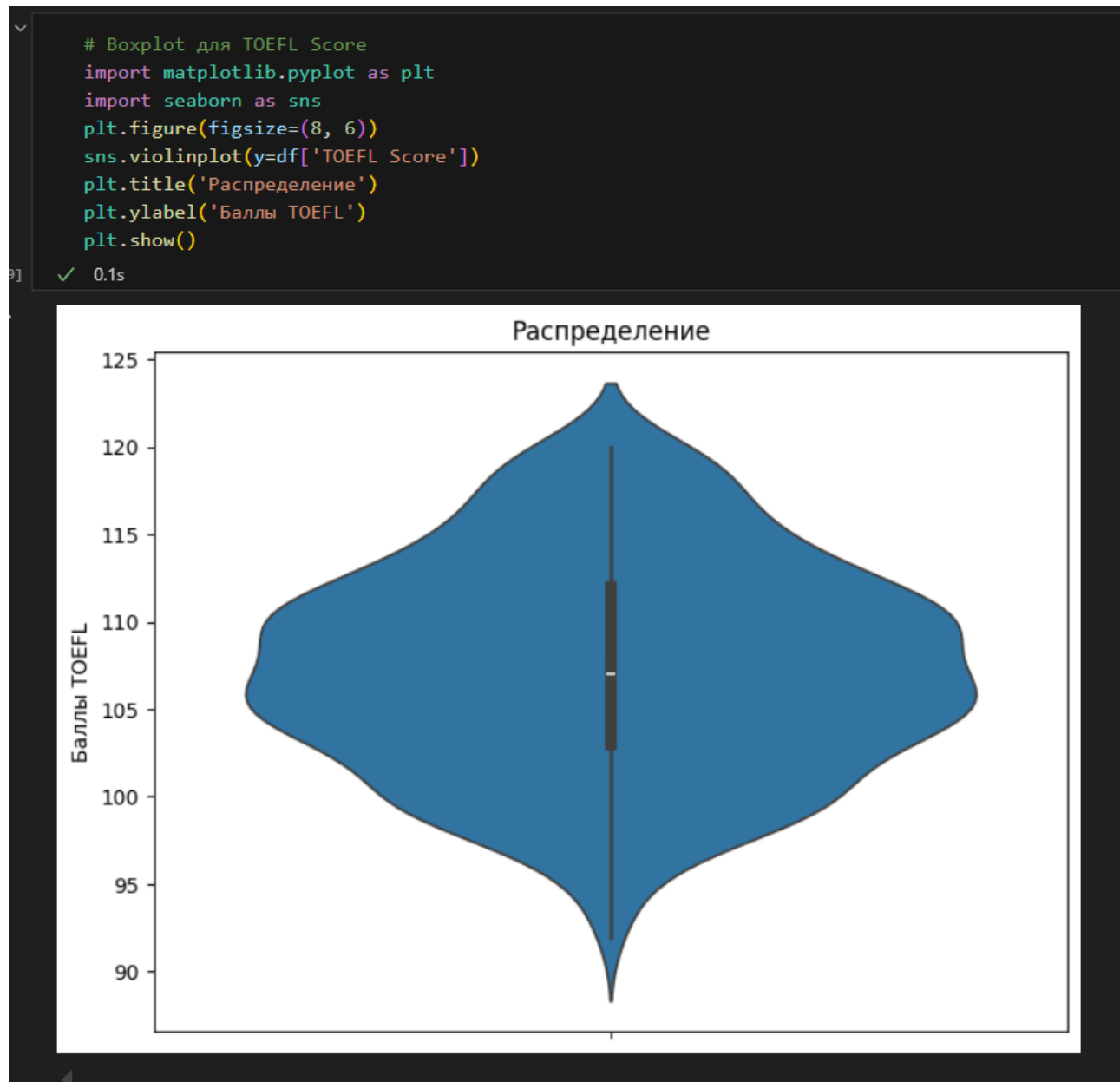
[48]

✓ 0.0s

...

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   Serial No.            400 non-null   int64  
 1   GRE Score              400 non-null   float64
 2   TOEFL Score            400 non-null   int64  
 3   University Rating      400 non-null   object  
 4   SOP                    400 non-null   float64
 5   LOR                    400 non-null   float64
 6   CGPA                   400 non-null   float64
 7   Research               400 non-null   int64  
 8   Chance of Admit        400 non-null   float64
dtypes: float64(5), int64(3), object(1)
memory usage: 28.3+ KB
```

#### 4) Построил violin plot для TOEFL Score



Далее для построения моделей машинного обучения я буду использовать все признаки(после преобразования University Rating обратно в числовой). В столбцах признаков нет пропусков, они являются числовыми и согласно моему представлению о поступлении в магистратуру все эти признаки так или иначе имеют влияние на шанс поступления.