Низовцев Р. А. ИУ5-64 РК1. Вариант 14.

Варианты заданий.

Номер варианта	Номер задачи	Номер набора данных, указанного в задаче
14	2	6

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?

Импорт библиотек и выгрузка датасета из внутреннего хранилища Google Colab:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt %matplotlib inline
sns.set(style="ticks")
data = pd.read_csv('/content/drive/MyDrive/Colab
Notebooks/Admission_Predict.csv', sep=",")
```

Датасет содержит информацию об абитуриентах магистратуры университета.

Шапка датасета:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Admission_Predict.csv', sep=",")
data.head()
```

	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

Описание столбцов

1) Serial No. - серийный номер студента
2) GRE Score - количество баллов за экзамен GRE, от 0 до 340 баллов (целые числа)
3) TOEFL Score - количество баллов за экзамен на знание Английского языка TOEFL, от 0 до 120 баллов (целые числа)
4) University rating - рейтинг университета, от 1 до 5 (целые числа)
5) Statement of Purpose and Letter of Recommendation Strength - заявление о целях студентах и сила его рекоммендаций, от 0 до 5 (целые числа)
6) Undergraduate GPA - средний балл студента, от 0 до 10 (целые числа)
7) Research - есть ли опыт в научных исследованиях, 0 или 1
8) Chance of admit - шанс принятия студента в университет, 0 до 1 (действительные числа)

Типы данных

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Admission_Predict.csv', sep=",")
#data.head()
data.dtypes
```

```
Serial No.             int64
GRE Score              int64
TOEFL Score            int64
University Rating      int64
SOP                  float64
LOR                  float64
CGPA                 float64
Research               int64
Chance of Admit      float64
dtype: object
```

Проверка на пустые ячейки

```
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

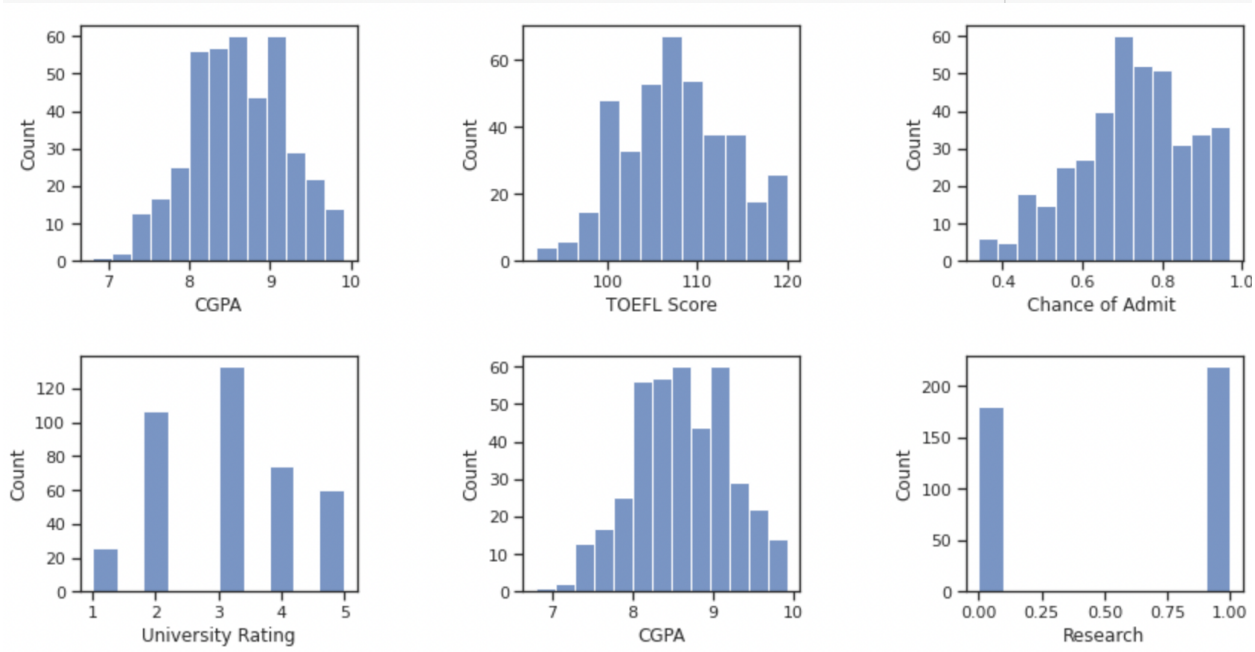
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Admission_Predict.csv', sep=",")
#data.head()
#data.dtypes
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

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

Пустых ячеек в датасете нет.

Стобчатые гистограммы по каждой характеристике:

```
_, axes = plt.subplots(2, 3, figsize= (14, 7))
sns.histplot(data['CGPA'], ax=axes[0][0])
sns.histplot(data['TOEFL Score'], ax=axes[0][1])
sns.histplot(data['Chance of Admit '], ax=axes[0][2])
sns.histplot(data['University Rating'], ax=axes[1][0])
sns.histplot(data['CGPA'], ax=axes[1][1])
sns.histplot(data['Research'], ax=axes[1][2])
plt.subplots_adjust(hspace=0.4, wspace=0.6)
plt.show()
```



В выводе можно отметить, что абитуриентов, участвоваших в исследованиях больше, чем не участвовавших. Больше всего выбирают вузы с рейтингом 3. Экзамен TOEFL сдают в основном от 100 до 120 баллов. Студентов больше со средним баллом от 8 до 9.

Основные статистические характеристки датасета

```
data.describe()
```

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	200.500000	316.807500	107.410000	3.087500	3.400000	3.452500	8.598925	0.547500	0.724350
std	115.614301	11.473646	6.069514	1.143728	1.006869	0.898478	0.596317	0.498362	0.142609
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.000000	6.800000	0.000000	0.340000
25%	100.750000	308.000000	103.000000	2.000000	2.500000	3.000000	8.170000	0.000000	0.640000
50%	200.500000	317.000000	107.000000	3.000000	3.500000	3.500000	8.610000	1.000000	0.730000
75%	300.250000	325.000000	112.000000	4.000000	4.000000	4.000000	9.062500	1.000000	0.830000
max	400.000000	340.000000	120.000000	5.000000	5.000000	5.000000	9.920000	1.000000	0.970000

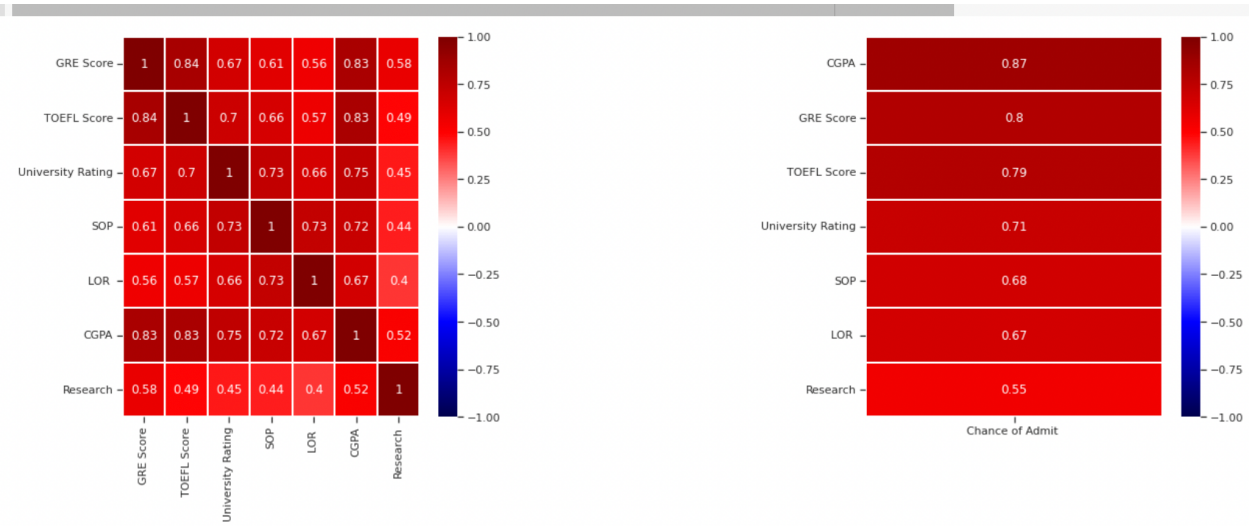
Корелляционный анализ

```
data.corr()
```

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.097526	-0.147932	-0.169948	-0.166932	-0.088221	-0.045608	-0.063138	0.042336
GRE Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.580391	0.802610
TOEFL Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.489858	0.791594
University Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.447783	0.711250
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.444029	0.675732
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.396859	0.669889
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.521654	0.873289
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.000000	0.553202
Chance of Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.553202	1.000000

Построим корреляционную матрицу, убрав характеристику Serial No.

```
_, axes = plt.subplots(1, 2, figsize=(20, 7))
sns.heatmap(data.drop(['Chance of Admit ', 'Serial No.'], axis=1).corr(),
annot=True, vmin = -1, vmax = 1, cmap='seismic', linewidth=1, ax =
axes[0])
sns.heatmap(pd.DataFrame(data.drop('Serial No.', axis=1).corr()['Chance of
Admit '].sort_values(ascending=False)[1:]),
            annot=True, vmin=-1, vmax=1, cmap='seismic', linewidth=1,
ax=axes[1])
plt.subplots_adjust(wspace=1)
plt.show()
```



Наиболее сильной кореляцией обладают:
Gre Score и TOEFL Score - 0.84
Gre Score и CGPA - 0.83
CGPA и TOEFL Score - 0.83

Шанс поступления в большей степени зависит от CGPA, Gre Score, TOEFL Score.

Выводы и ответы на вопросы задания:

Способы для обработки пропусков никакие не использовались, так как их в датасете не оказалось.

Одновременное использование следующих пар признаков: Gre Score и TOEFL Score, Gre Score и CGPA, CGPA и TOEFL Score в моделях машинного обучения приведет к мультиколлинеарности. Следовательно, необходимо выбрать один признак.
Из второй матрицы видно, что наибольшей корреляции с прогнозируемой величиной обладает признак CGPA. Оставляем его и все остальные признаки, убирая признаки TOEFL Score, Gre Score.
Подводя итог, оставляем для обучения модели машинного обучения следующие признаки: CGPA, University Rating, SOP, LOR, Research.

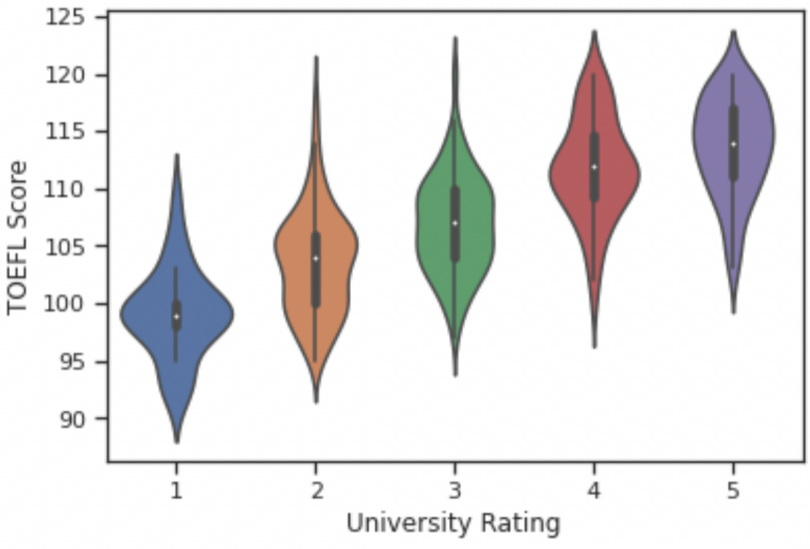
Дополнительное задание для группы:

- Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

```
sns.violinplot(x=data["University Rating"], y=data["TOEFL Score"])
```

```
sns.violinplot(x=data["University Rating"], y=data["TOEFL Score"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe88d4ab350>
```



На скрипичной диаграмме показана зависимость рейтинга университета от количества баллов за экзамен на знание английского языка TOEFL. Мы видим, что чем выше рейтинг вуза, тем более высокие баллы набирают абитуриенты туда поступающие.