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 Системы обработки информации и управления

Рубежный контроль №1
«Технологии разведочного анализа и обработки данных»
по курсу «Технологии машинного обучения»

Вариант №5

Выполнил:
Студент группы ИУ5Ц-81Б
Тураев Глеб

Проверил:
Преподаватель кафедры ИУ5
Гапанюк Ю.Е.

Москва 2020

Данные варианта:

Номер варианта	Номер задачи	Номер набора данных
30	4	6

Задача:

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

Дополнительное задание:

Для пары произвольных колонок данных построить график «Диаграмма рассеяния».

Выполнение рубежного контроля:

1) **Текстовое описание набора данных**

В качестве набора данных используем набор данных о прогнозировании поступления выпускников.

https://www.kaggle.com/mohansacharya/graduate-admissions

Анализ подобного набора данных содержит несколько параметров, которые считаются важными при подаче заявки на магистерские программы, а также позволяющие поступить выпускникам в те или иные ВУЗы.

Датасет состоит из одного файла: Admission_Predict.csv.

Файл содержит следующие колонки:

- Serial No – порядковый номер строки;
- GRE Scores – количество баллов GRE из всех возможных 340;
- TOEFL Scores – количество баллов TOEFL из всех возможных 120;
- University Rating – рейтинг университета, оцениваемый от 1 до 5;
- Statement of Purpose – формулировка цели поступления;
- Letter of Recommendation Strength – сила рекомендательного письма;
- Undergraduate GPA – средний академический балл: от 1 до 10;
- Research Experience – опыт исследования: либо 0, либо 1;
- Chance of Admit – вероятность признания в диапазоне от 0 до 1.

2) Импорт библиотек

Осуществим импорт библиотек с помощью команды **import:**

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

```
/usr/local/lib/python3.6/dist-packages/statsmo
  import pandas.util.testing as tm
```

3) Загрузка данных

Загрузим файлы датасета с помощью библиотеки **Pandas**:

```
[7]  data = pd.read_csv('Admission_Predict.csv', sep=",")
```

4) Проверка на наличие пропусков в данных

```
[8]  data.isnull().sum()
```

```
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
```

5) Основные характеристики набора данных

Выведем первые «5» строк нашего датасета:

```
[9] 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

Узнаем размер датасета:

```
[10] data.shape
```

```
(400, 9)
```

Выведем основные статистические характеристики набора данных:

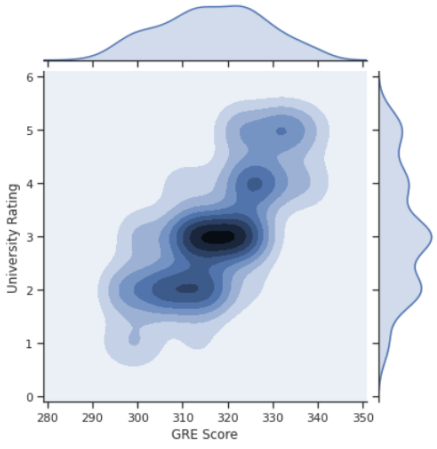
```
[17] data.describe()
```

	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

6) Построим основные графики, входящие в этап разведочного анализа данных:

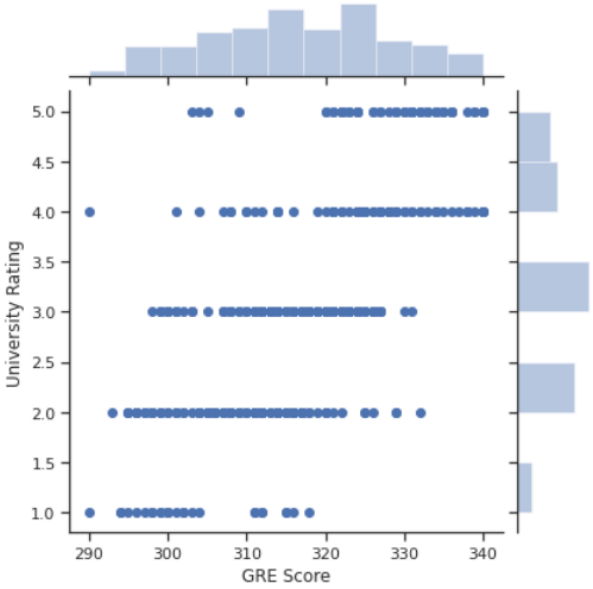
```
[12] sns.jointplot(x='GRE Score', y='University Rating', data=data, kind="kde")
```

```
<seaborn.axisgrid.JointGrid at 0x7f337fd369e8>
```



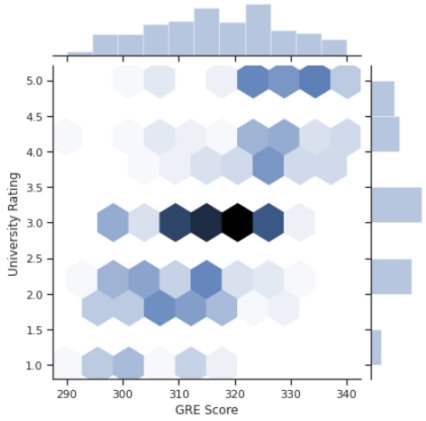
```
[13] sns.jointplot(x='GRE Score', y='University Rating', data=data)
```

```
<seaborn.axisgrid.JointGrid at 0x7f337ccf34e0>
```



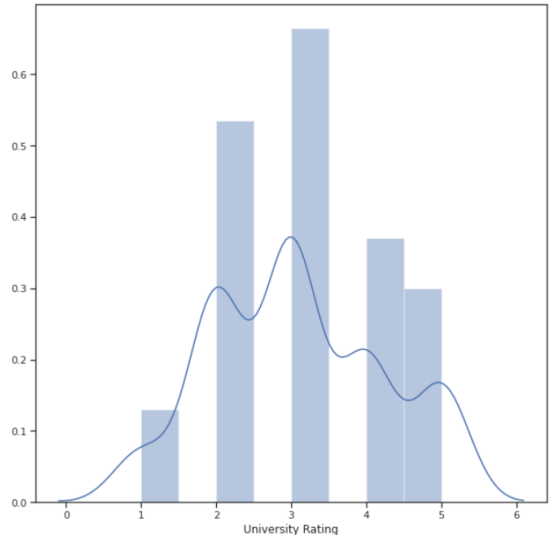

```
[16] sns.jointplot(x='GRE Score', y='University Rating', data=data, kind="hex")
```

```
<seaborn.axisgrid.JointGrid at 0x7f337ca28e80>
```




```
[14] fig, ax = plt.subplots(figsize=(10,10))
     sns.distplot(data['University Rating'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f337cbde358>
```



Вывод: данные графики нам отображают зависимость между двумя важными компонентами данного датасета: **GRE Scores** *(количество баллов GRE из всех возможных 340)* и **University Rating** *(рейтинг университета, оцениваемый от 1 до 5)*. С помощью графиков выпускники могут сделать вывод о том, как влияет количество баллов на рейтинг того или иного университета и определиться с его выбором.

7) **Выполним дополнительное задание: для пары произвольных колонок данных построим график «Диаграмма рассеяния», используя колонки GRE Score и University Rating**

```
[11] fig, ax = plt.subplots(figsize=(8,8))
     sns.scatterplot(ax=ax, x='GRE Score', y='University Rating', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f337fd36400>
```