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 №7

Выполнил:
Студент группы РТ5-61
Калин Владимир

_____

Проверил:
Преподаватель кафедры ИУ5
Гапанюк Ю.Е.

_____

Москва 2020

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_Ver1.1.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**:

```
[ ]  data = pd.read_csv('/Admission_Predict_Ver1.1.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:

```
data.shape
```

```
(500, 9)
```

Выведем основные статистические характеристики набора данных:

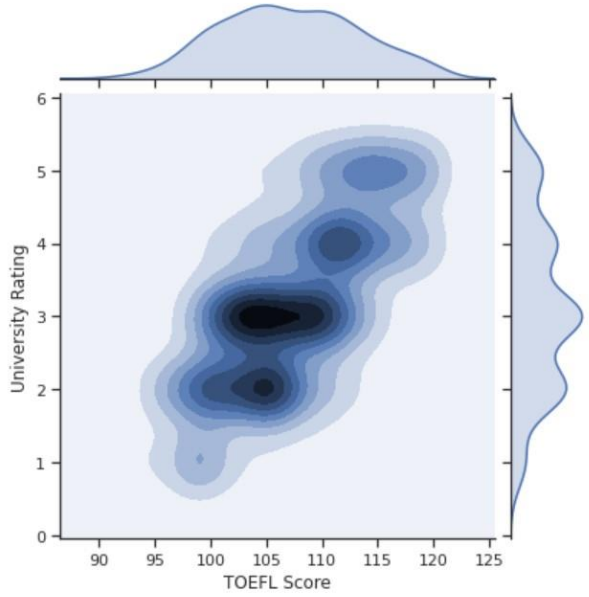
```
[18] data.describe()
```

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

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

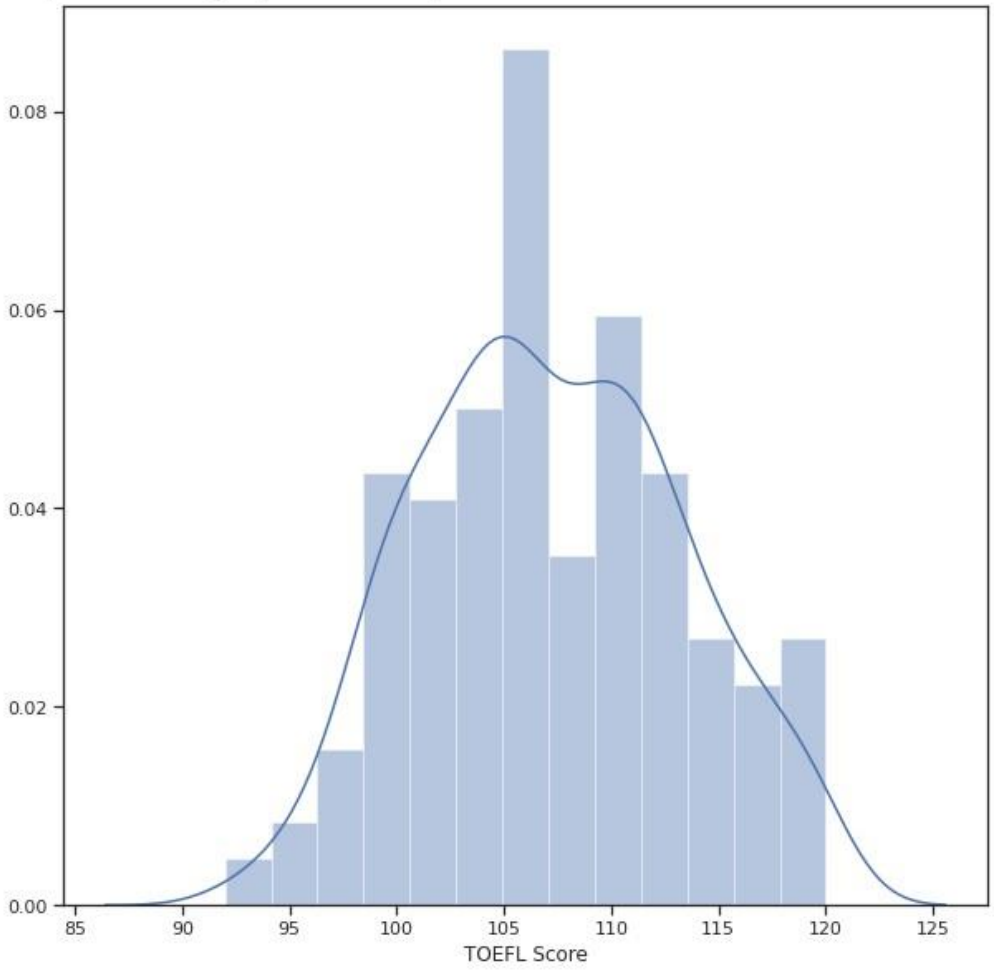
```
sns.jointplot(x='TOEFL Score', y='University Rating', data=data, kind="kde"
```

```
<seaborn.axisgrid.JointGrid at 0x7f3174cb15f8>
```



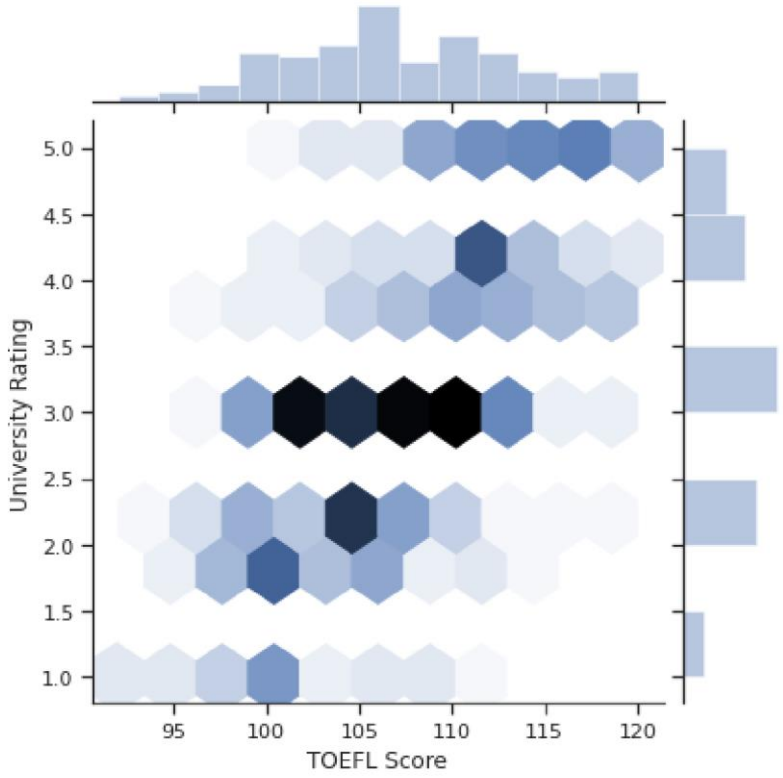

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['TOEFL Score'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f31746fd0b8>
```



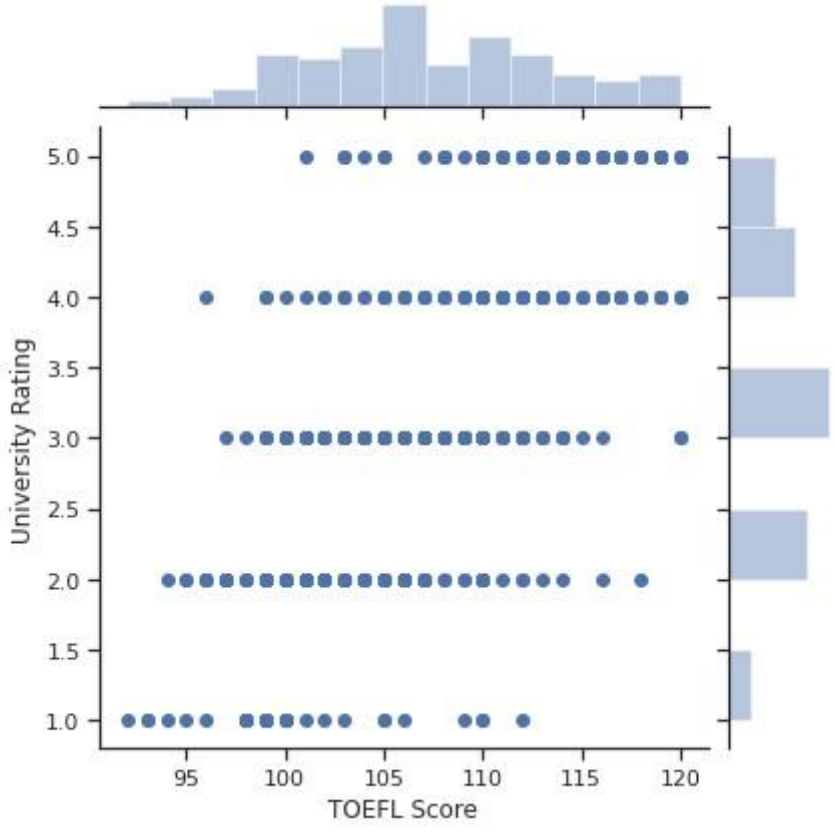
```
] sns.jointplot(x='TOEFL Score', y='University Rating', data=data, kind="hex")
```

```
<seaborn.axisgrid.JointGrid at 0x7f3174df0128>
```




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

```
<seaborn.axisgrid.JointGrid at 0x7f3174e0e9e8>
```



Вывод: данные графики нам отображают зависимость между двумя важными компонентами данного датасета: **TOEFL Scores** *(количество баллов GRE из всех возможных 120)* 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 TOEFL Score и University Rating**

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f31749540b8>
```