МГТУ имени Баумана
Факультет «Информатика и Системы управления»
Кафедра «Системы обработки информации и управления»
Дисциплина «Технологии машинного обучения»

Отчет по РК №1
Вариант 7

# «Технологии разведочного анализа и обработки данных.»

Выполнила:
Студентка группы ИУ5-61Б
Громова О.А.

Преподаватель:
Гапанюк Ю.Е.

Москва, 2021г.

Задание:

## Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Набор данных:

https://www.kaggle.com/lava18/google-play-store-apps

## Листинг программы:

In [1]:
```
#Громова Ольга ИУ5-51
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

In [2]:
```
data = pd.read_csv('Admission_Predict.csv')
```

B [3]:
```
data.head()
```

Out[3]:

| | 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 |

B [4]:
```
data.dtypes
```

Out[4]:
```
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
```

B [5]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[5]:
```
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
```

В [6]:
```
data.info()
```

```
<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    int64  
 2   TOEFL Score        400 non-null    int64  
 3   University Rating  400 non-null    int64  
 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(4), int64(5)
memory usage: 28.2 KB
```

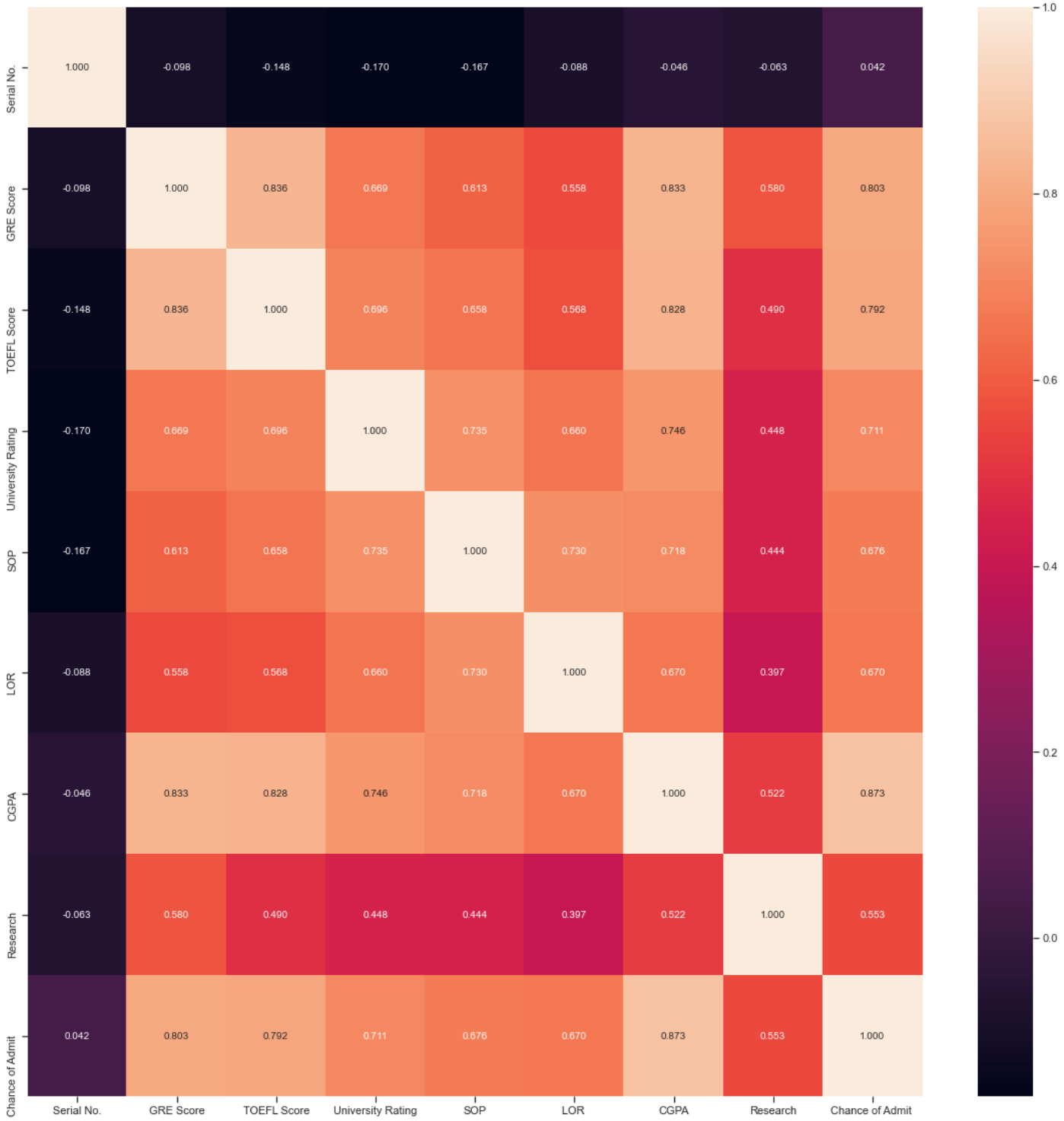
В [7]:
```
## Корр. анализ
corr_matrix = data.corr()
```

В [8]:
```
corr_matrix['GRE Score']
```

Out[8]:
```
Serial No.          -0.097526
GRE Score            1.000000
TOEFL Score          0.835977
University Rating    0.668976
SOP                  0.612831
LOR                  0.557555
CGPA                 0.833060
Research             0.580391
Chance of Admit      0.802610
Name: GRE Score, dtype: float64
```

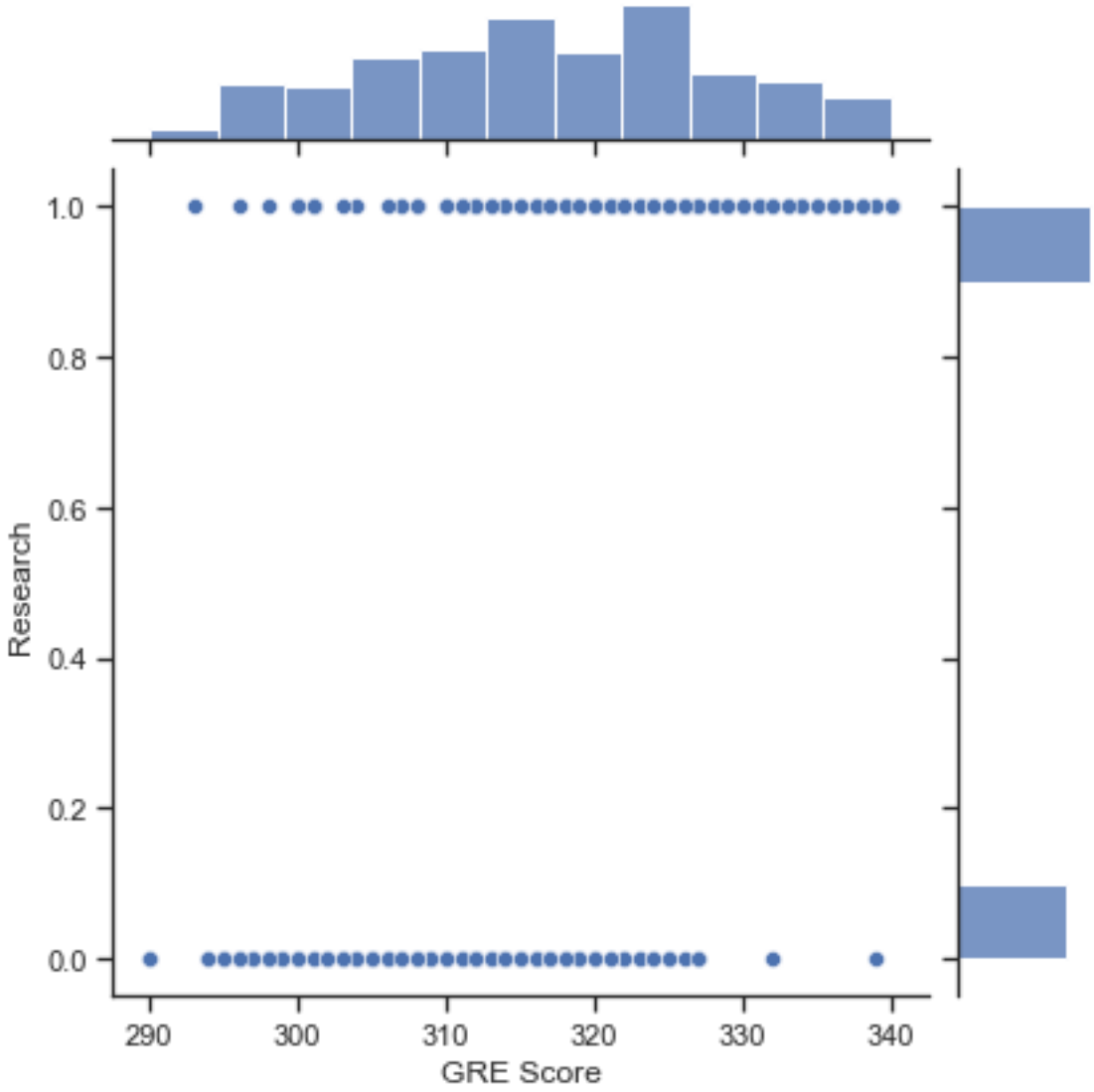
In [9]:
```
plt.figure(figsize=(20,20))
sns.heatmap(corr_matrix, annot=True, fmt='.3f')
```

Out[9]: <AxesSubplot:>




In [11]:
```
# Увеличенные диаграммы рассеяния
sns.jointplot(x = "GRE Score", y = "Research", kind="scatter", data = data)
```

Out[11]: <seaborn.axisgrid.JointGrid at 0x24c681134f0>


Research
1.0
0.8
0.6
0.4
0.2
0.0
290
300
310
320
330
340
GRE Score