Рубежный контроль №1

Тема: Технологии разведочного анализа и обработки данных

Зубарева А. М. ИУ5-65Б Вариант 7

Загрузка необходимых библиотек:

In [33]:
```
import pandas as pd
import seaborn as sns
from sklearn import preprocessing
```

In [34]:
```
data = pd.read_csv('/Users/toffee/Downloads/archive/Admission_Predict.csv')
```

In [35]:
```
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
```

In [36]:
```
data.shape
```

Out[36]:
```
(400, 9)
```

(строк, колонок)

In [39]:
```
data.head()
```

Out[39]:

	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

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
4	5	314	103	2	2.0	3.0	8.21	0	0.65

In [40]:
```
data.dtypes
```

Out[40]:
```
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
```

Все значения числовые

Проверим наличие пропусков

In [42]:
```
data.isnull().sum()
```

Out[42]:
```
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
```

Здесь видно, что пропусков в данных нет ни в одном столбце

In [43]:
```
data.value_counts()
```

Out[43]:
```
Serial No.  GRE Score  TOEFL Score  University Rating  SOP  LOR   CGPA  Resear
ch  Chance of Admit
1           337        118          4                 4.5  4.5   9.65  1
0.92                1
264         324        111          3                 2.5  1.5   8.79  1
0.70                1
274         312        99           1                 1.0  1.5   8.01  1
0.52                1
273         294        95           1                 1.5  1.5   7.64  0
0.49                1
272         299        96           2                 1.5  2.0   7.86  0
0.54                1

..
131         339        114          5                 4.0  4.5   9.76  1
0.96                1
130         333        118          5                 5.0  5.0   9.35  1
0.92                1
129         326        112          3                 3.5  3.0   9.10  1
0.84                1
128         319        112          3                 2.5  2.0   8.71  1
0.78                1
400         333        117          4                 5.0  4.0   9.66  1
```

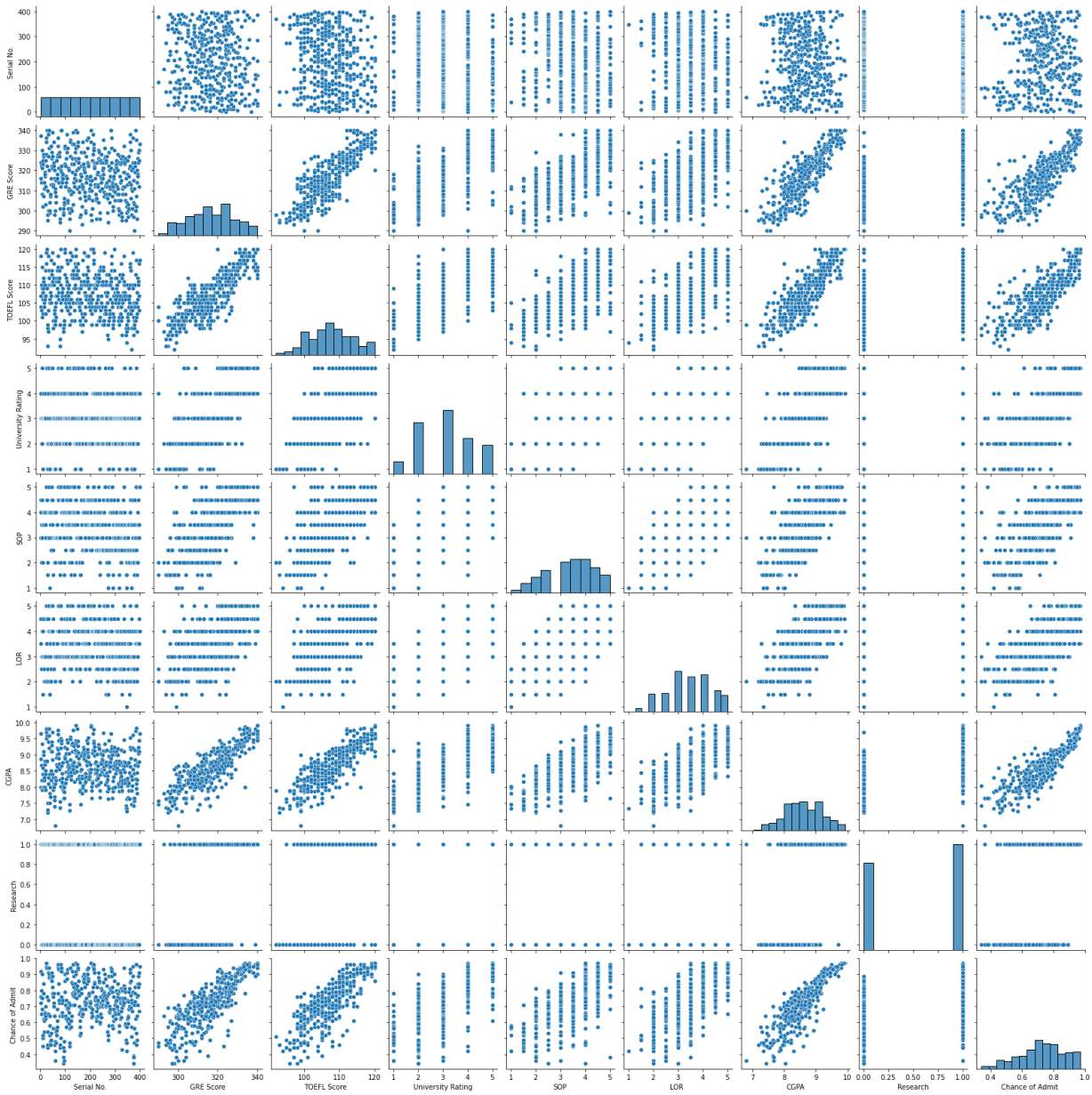
```
0.95                  1
Length: 400, dtype: int64
```

In [44]:

```
sns.pairplot(data)
```

Out[44]:

```
<seaborn.axisgrid.PairGrid at 0x7fad58883df0>
```



Построили pairplot, здесь уже можно увидеть корреляцию полей

In [45]:

```
data.corr()
```

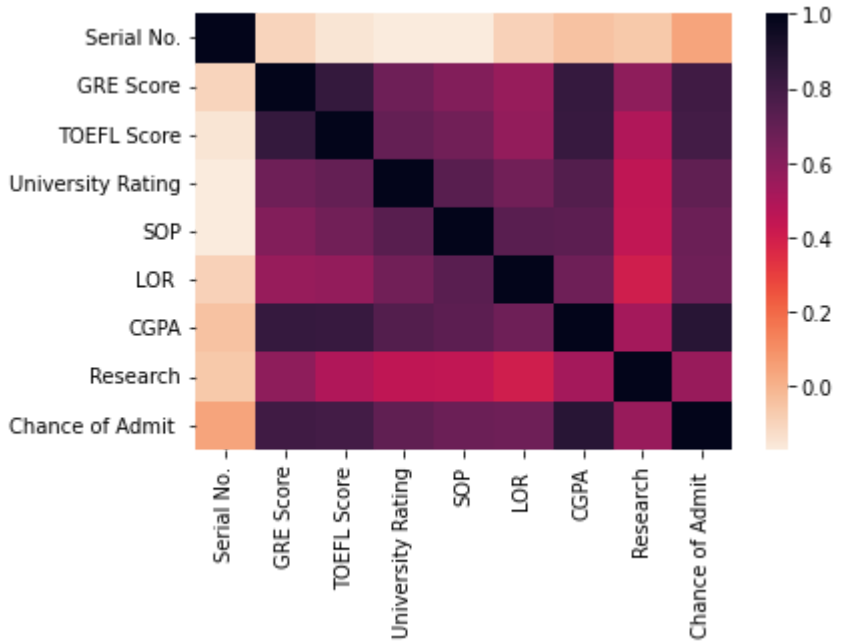
Out[45]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Res
Serial No.	1.000000	-0.097526	-0.147932	-0.169948	-0.166932	-0.088221	-0.045608	-0.0
GRE Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.5
TOEFL Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.4
University Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.4

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Res
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.44
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.39
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.5
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.0
Chance of Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.5

Построим heatmap для лучшего визуального представления всех корреляций

In [46]:
```
cmap = sns.cm.rocket_r
ax = sns.heatmap(data.corr(), cmap=cmap)
```



Наиболее интересно для построения модели как коррелируют все поля с Chance of admit. \ Видим, что у нас наиболее влиятельные - поля CGPA, GRE Score и TOEFL Score. Соответственно, они должны вносить наибольший вклад в итоговую модель

До построения модели необходимо нормализировать поля, так как все они числовые, и находятся порой в разных диапазонах (GRE score имеет значения порядка 300, а CGPA - порядка 10)

In [52]:
```
normalized_data = preprocessing.normalize(data)
```