Вардумян А.Т. ИУ5-61Б

1 Оглавление

2 Задание (к оглавлению)

> 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.

> Для пары произвольных колонок данных построить график "Диаграмма рассеяния".

> Необходимо подготовить отчет по рубежному контролю и разместить его в Вашем репозитории. Вы можете использовать титульный лист, или в начале ноутбука в текстовой ячейке указать Ваши Ф.И.О. и группу.

3 Описание датасета (к оглавлению)

Датасет `Graduate Admission 2` создан для прогнозирования поступления в аспирантуру. Он состоит из двух таблиц:

- `Admission_Predict.csv`
- `Admission_Predict_Ver1.1.csv`

В текущей работе используем второй.

Таблица `Admission_Predict_Ver1.1.csv` состоит из следующих столбцов:

- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose and Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)

- Research Experience (either 0 or 1)
- Chance (ranging from 0 to 1)

4 Импорт библиотек (к оглавлению)

Ввод [1]:

```
import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

5 Загрузка и первичный анализ данных (к оглавлению)

Ввод [2]:

```
data = pd.read_csv("Admission_Predict_Ver1.1.csv", sep=",")
data
```

Out[2]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance
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
...	...	...	...	...	...	...	...	...	...
495	496	332	108	5	4.5	4.0	9.02	1	0.87
496	497	337	117	5	5.0	5.0	9.87	1	0.96
497	498	330	120	5	4.5	5.0	9.56	1	0.93
498	499	312	103	4	4.0	5.0	8.43	0	0.73
499	500	327	113	4	4.5	4.5	9.04	0	0.84

500 rows × 9 columns

Ввод [3]:

```
# Переименуем столбцы, чтобы избавиться от пробелов в именах
data = data.rename(columns={
    "Serial No.": "ID",
    "GRE Score": "GRE",
    "TOEFL Score": "TOEFL",
    "University Rating": "Rating",
    "CGPA": "GPA",
})
```

Ввод [4]:

```
data.describe()
```

Out[4]:

	ID	GRE	TOEFL	Rating	SOP	LOR	GPA	Re
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.(
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.5
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.4
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.(
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.(
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.(
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.(
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.(

Ввод [5]:

```
data.shape
```

Out[5]:

```
(500, 9)
```

Ввод [6]:

```
data.dtypes
```

Out[6]:

```
ID             int64
GRE            int64
TOEFL          int64
Rating         int64
SOP          float64
LOR          float64
GPA          float64
Research       int64
Chance       float64
dtype: object
```

Ввод [7]:

```
# Количество пустых значений
total_count = data.shape[0]
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    temp_perc = round((temp_null_count / total_count) * 100.0, 2)
    print('Колонка {} — {}, {}%'.format(col, temp_null_count, temp_perc))
```

```
Колонка ID — 0, 0.0%
Колонка GRE — 0, 0.0%
Колонка TOEFL — 0, 0.0%
Колонка Rating — 0, 0.0%
Колонка SOP — 0, 0.0%
Колонка LOR — 0, 0.0%
Колонка GPA — 0, 0.0%
Колонка Research — 0, 0.0%
Колонка Chance — 0, 0.0%
```

6 Визуализация (к оглавлению)

Ввод [8]:

```
# Удалим столбец ID, он неинформативен
df = data.drop('ID', axis=1)
df
```

Out[8]:

	GRE	TOEFL	Rating	SOP	LOR	GPA	Research	Chance
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...
495	332	108	5	4.5	4.0	9.02	1	0.87
496	337	117	5	5.0	5.0	9.87	1	0.96
497	330	120	5	4.5	5.0	9.56	1	0.93
498	312	103	4	4.0	5.0	8.43	0	0.73
499	327	113	4	4.5	4.5	9.04	0	0.84

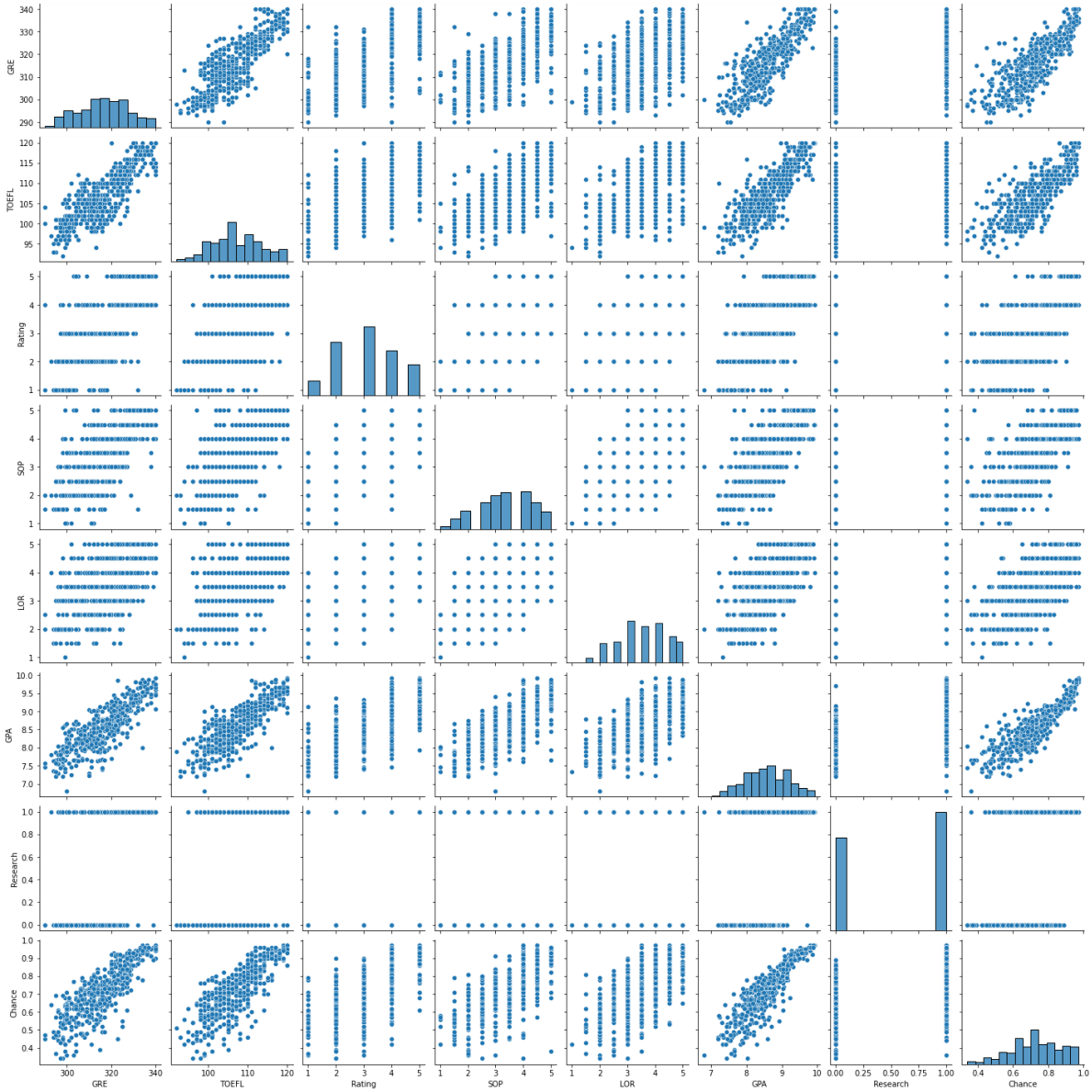
500 rows × 8 columns

Ввод [9]:

```
sns.pairplot(df)
```

Out[9]:

```
<seaborn.axisgrid.PairGrid at 0x7fb397ea3f40>
```

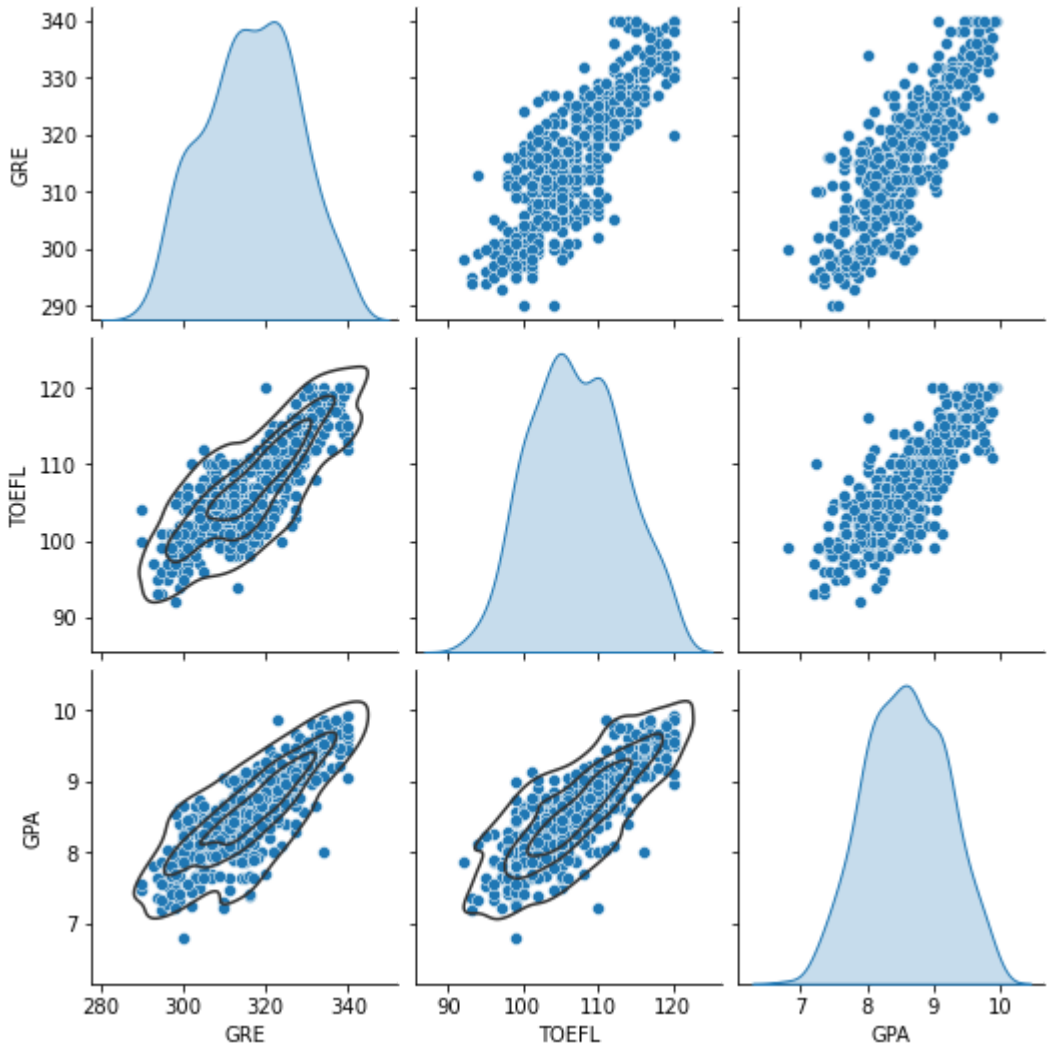


Ввод [10]:

```
g = sns.pairplot(
    df,
    vars=['GRE', 'TOEFL', 'GPA'],
    diag_kind='kde'
)
g.map_lower(sns.kdeplot, levels=4, color=".2")
```

Out[10]:

```
<seaborn.axisgrid.PairGrid at 0x7fb397f32700>
```

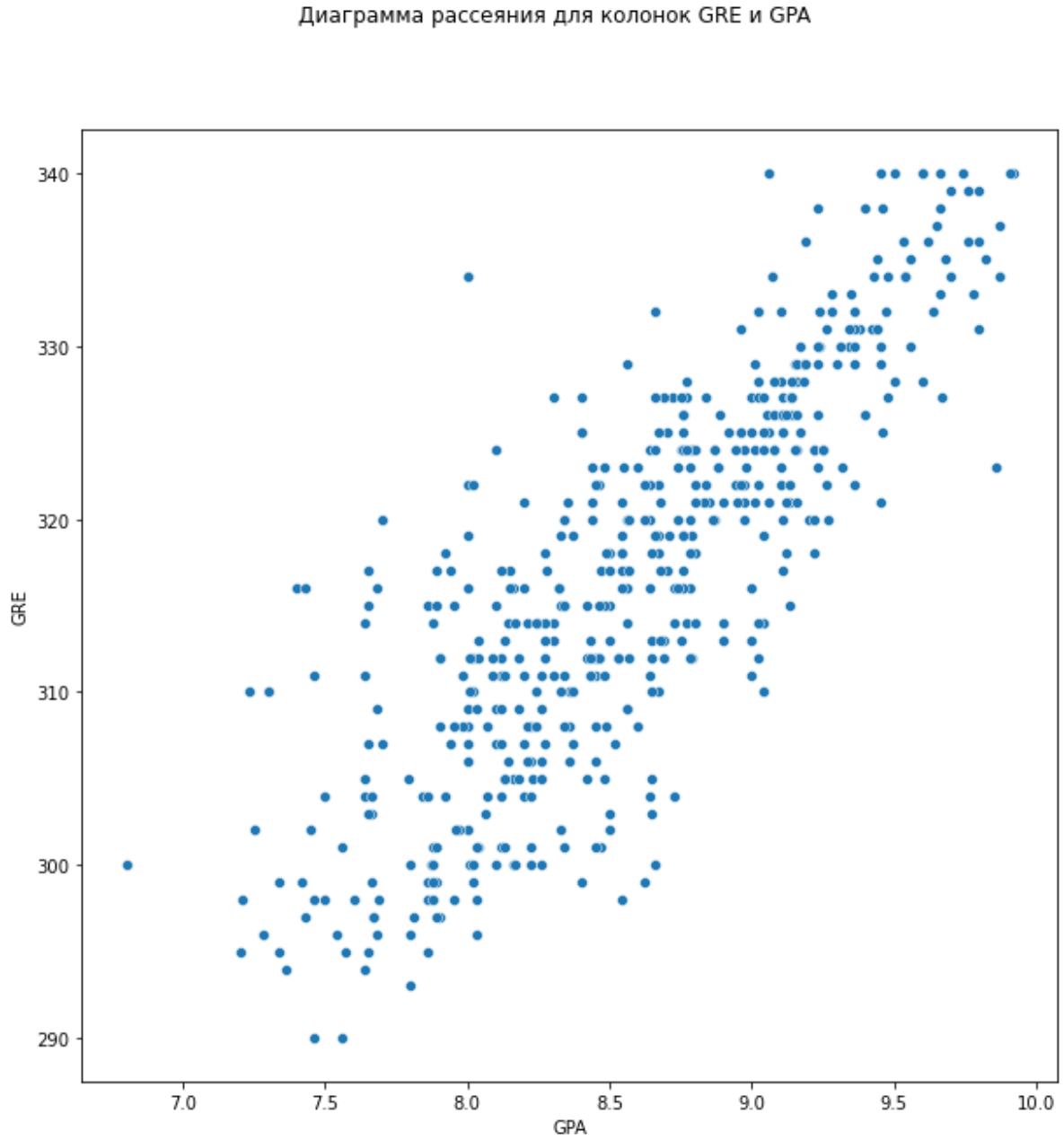


Ввод [11]:

```
fig, ax = plt.subplots(figsize=(10,10))
fig.suptitle("Диаграмма рассеяния для колонок GRE и GPA")
sns.scatterplot(ax=ax, x='GPA', y='GRE', data=data)
```

Out[11]:

```
<AxesSubplot:xlabel='GPA', ylabel='GRE'>
```



7 Корреляционный анализ (к оглавлению)

Ввод [12]:

```
df.corr()
```

Out[12]:

	GRE	TOEFL	Rating	SOP	LOR	GPA	Research	Chance
GRE	1.000000	0.827200	0.635376	0.613498	0.524679	0.825878	0.563398	0.810351
TOEFL	0.827200	1.000000	0.649799	0.644410	0.541563	0.810574	0.467012	0.792228
Rating	0.635376	0.649799	1.000000	0.728024	0.608651	0.705254	0.427047	0.690132
SOP	0.613498	0.644410	0.728024	1.000000	0.663707	0.712154	0.408116	0.684137
LOR	0.524679	0.541563	0.608651	0.663707	1.000000	0.637469	0.372526	0.645365
GPA	0.825878	0.810574	0.705254	0.712154	0.637469	1.000000	0.501311	0.882413
Research	0.563398	0.467012	0.427047	0.408116	0.372526	0.501311	1.000000	0.545871
Chance	0.810351	0.792228	0.690132	0.684137	0.645365	0.882413	0.545871	1.000000

Ввод [13]:

```
df.corr()['Chance']
```

Out[13]:

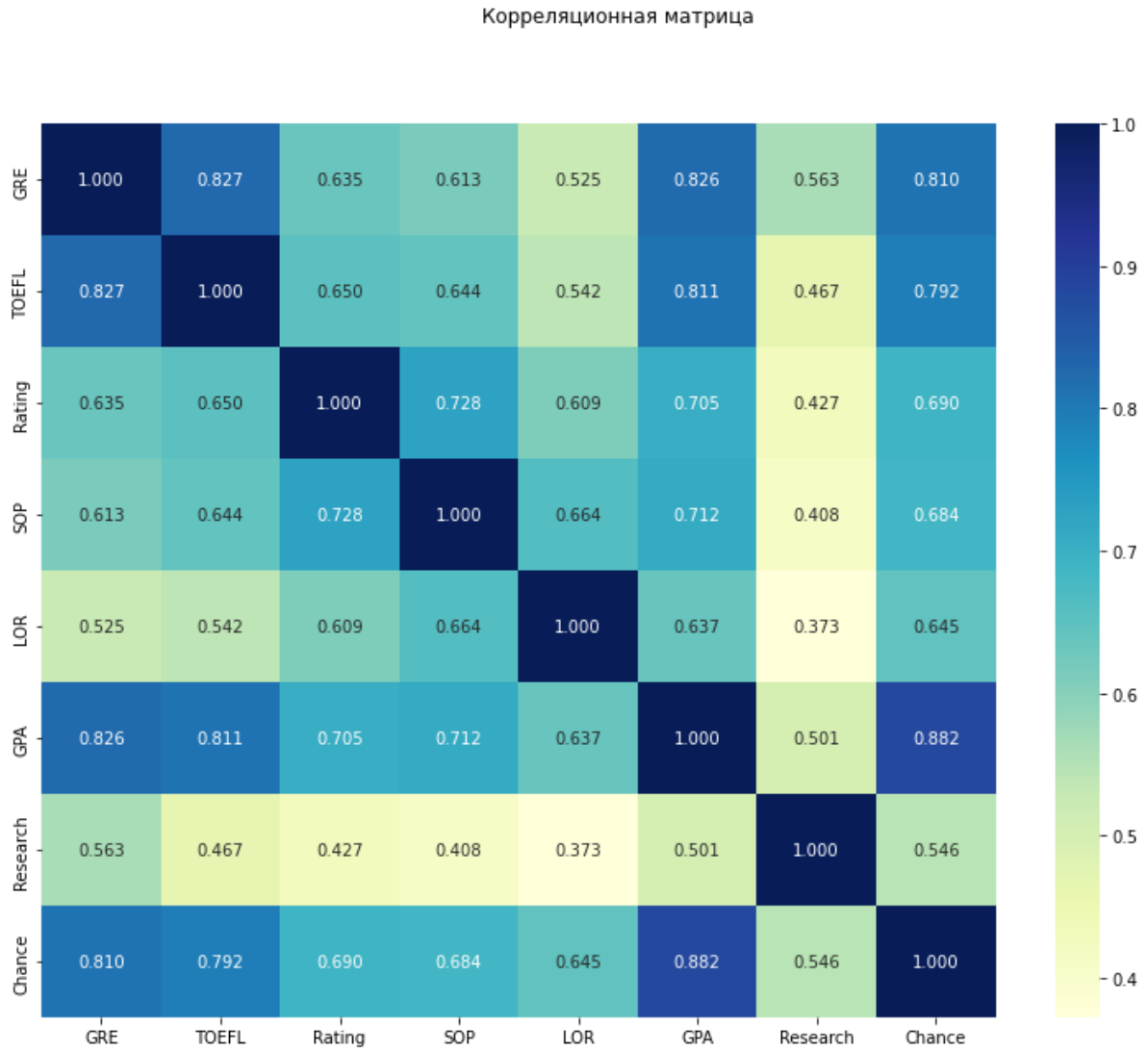
```
GRE         0.810351
TOEFL       0.792228
Rating      0.690132
SOP         0.684137
LOR         0.645365
GPA         0.882413
Research    0.545871
Chance      1.000000
Name: Chance, dtype: float64
```

Ввод [14]:

```
fig, ax = plt.subplots(1, 1, sharex='col', sharey='row', figsize=(13,10))
fig.suptitle('Корреляционная матрица')
sns.heatmap(df.corr(), ax=ax, annot=True, fmt='.3f', cmap='YlGnBu')
```

Out[14]:

```
<AxesSubplot:>
```



На основе корреляционной матрицы можно сделать следующие выводы.

Все признаки достаточно хорошо коррелируют с целевым признаокм `Chance` :

Признак	Корреляция
GRE	0.810351
TOEFL	0.79222
Rating	0.690132
SOP	0.684137
LOR	0.645365
GPA	0.882413
Research	0.545871

Признаки `GRE` , `TOEFL` , `GPA` сильно коррелируют между собой, следовательно, для того, чтобы повысить качество модели, стоит выбрать из них только один, который наиболее сильно коррелирует с целевым признаком, т.е. `GPA` .

Таким образом, для построения модели использовались бы следующие признаки: `Rating` , `SOP` , `LOR` , `GPA` , `Research` .