Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по рубежному контролю

«Технологии разведочного анализа и обработки данных»

Вариант 6

In [10]:
```
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [3]:
```
df = pd.read_csv('Admission_Predict.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	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

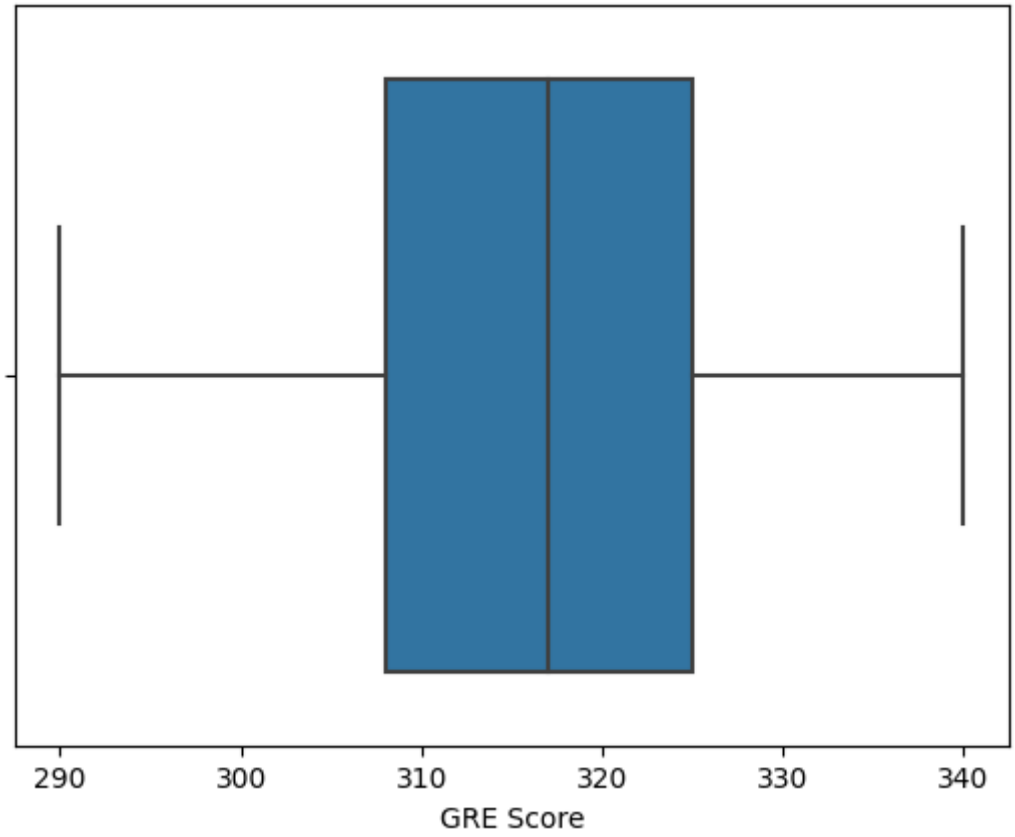
In [5]:
```
df.info()
```

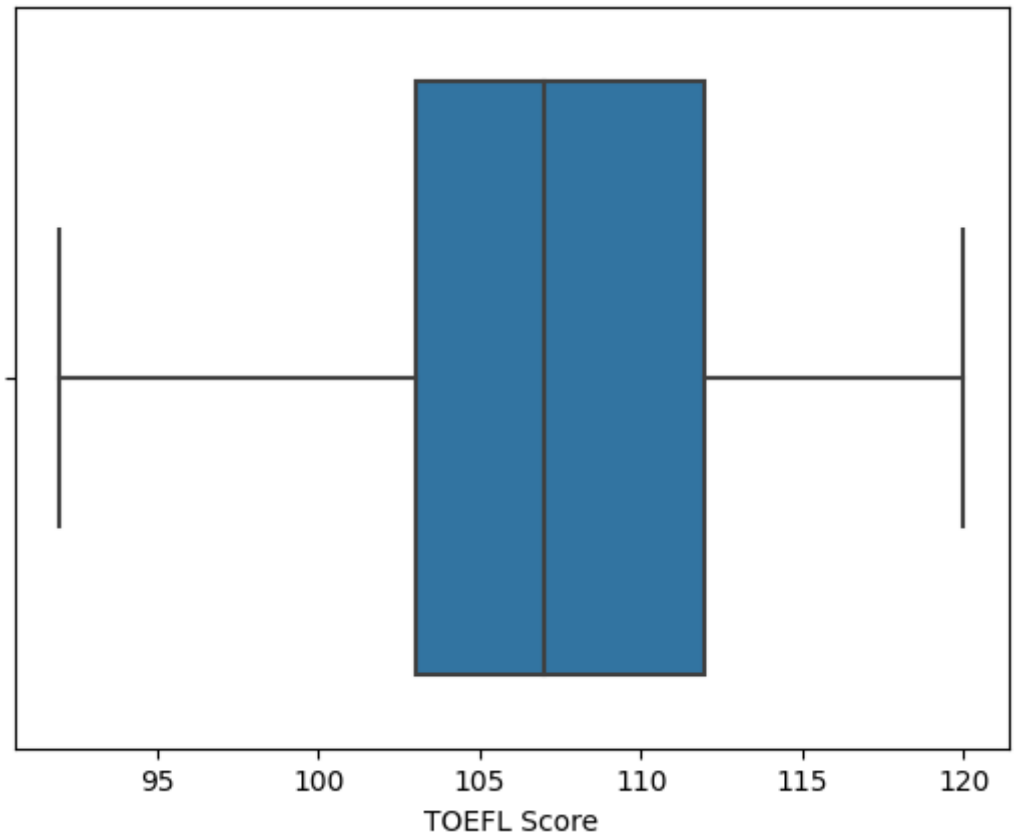
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Serial No.         400 non-null    int64  
 1   GRE Score          400 non-null    int64  
 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
```

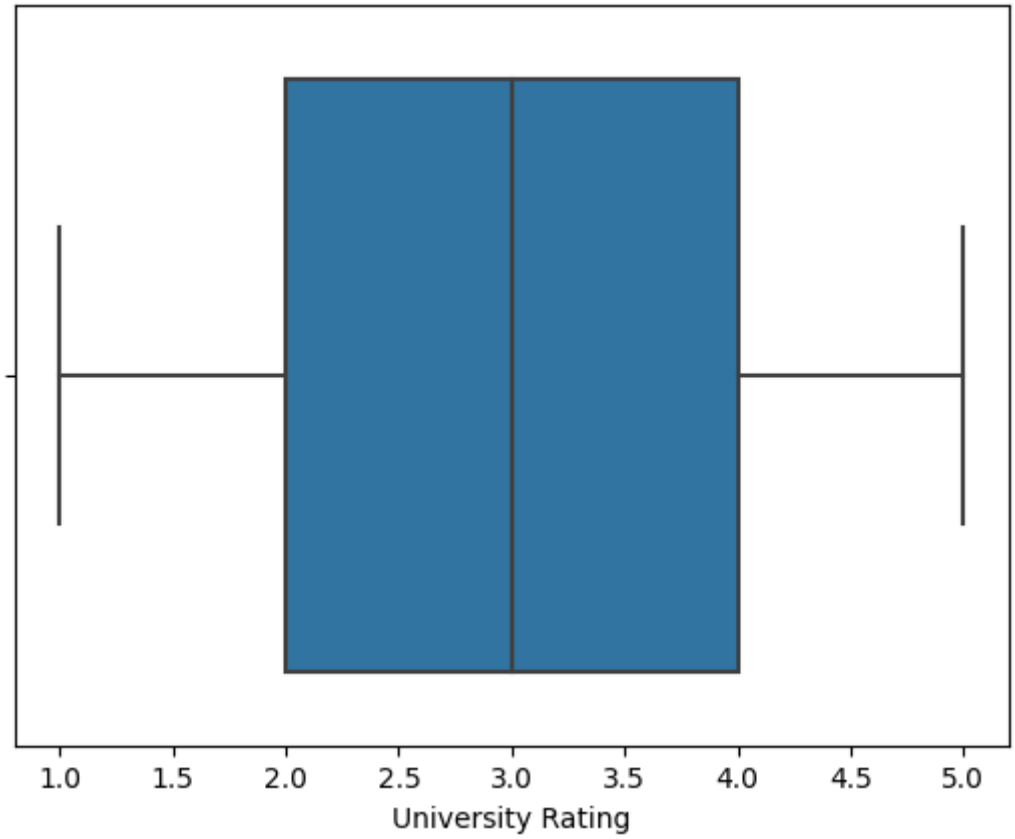
In [6]:
```
df = df.drop(['Serial No.'], axis=1)
df.isnull().sum()
```

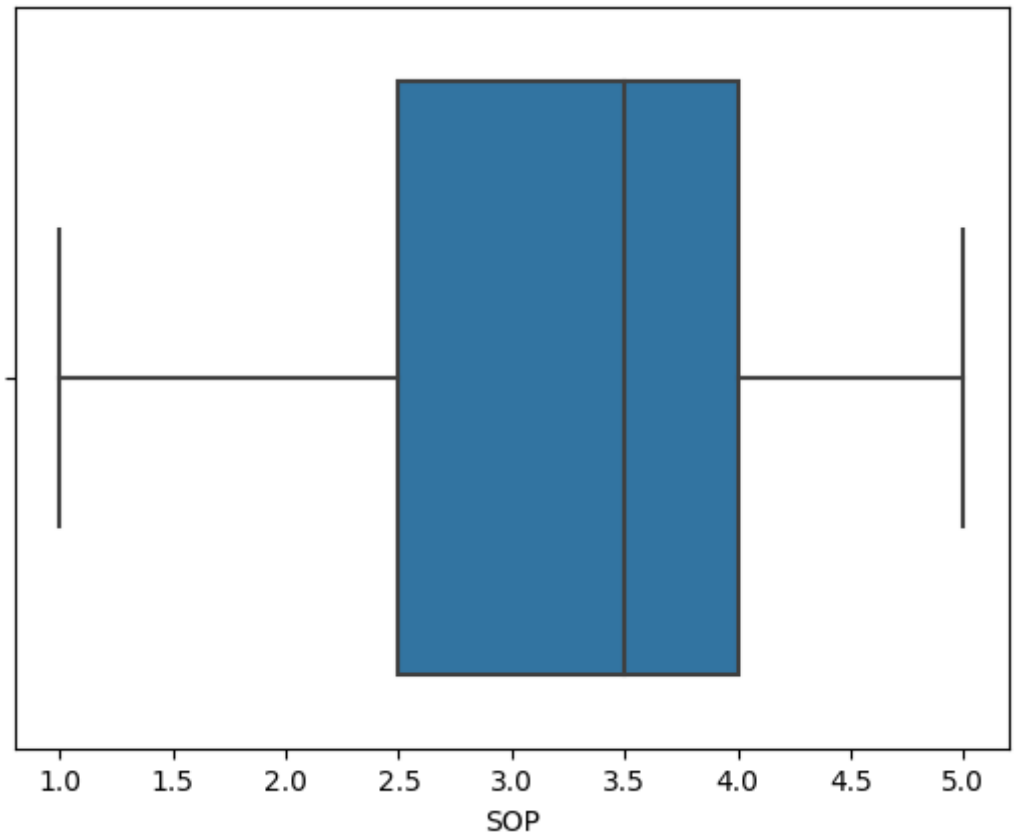
Out[6]:
```
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

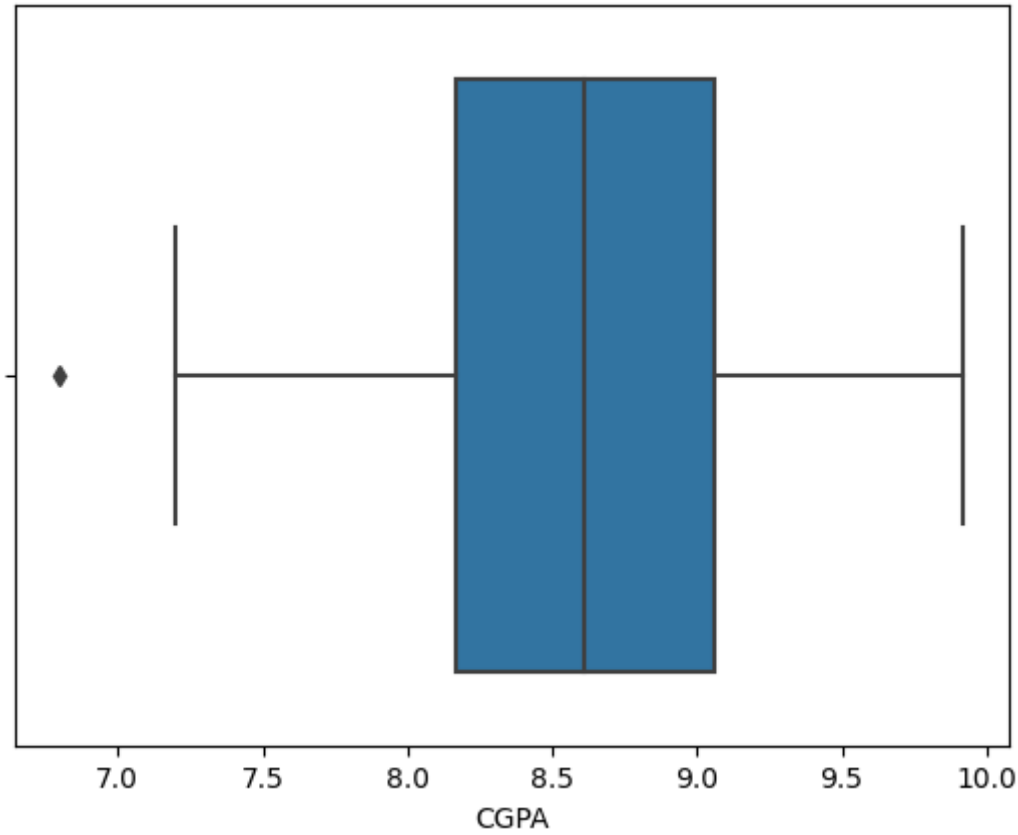
In [19]:
```
fig = sns.boxplot(df['GRE Score'])
plt.show()

fig = sns.boxplot(df['TOEFL Score'])
plt.show()

fig = sns.boxplot(df['University Rating'])
plt.show()

fig = sns.boxplot(df['SOP'])
plt.show()

fig = sns.boxplot(df['CGPA'])
plt.show()

plt.show()
```


290
300
310
320
330
340
GRE Score

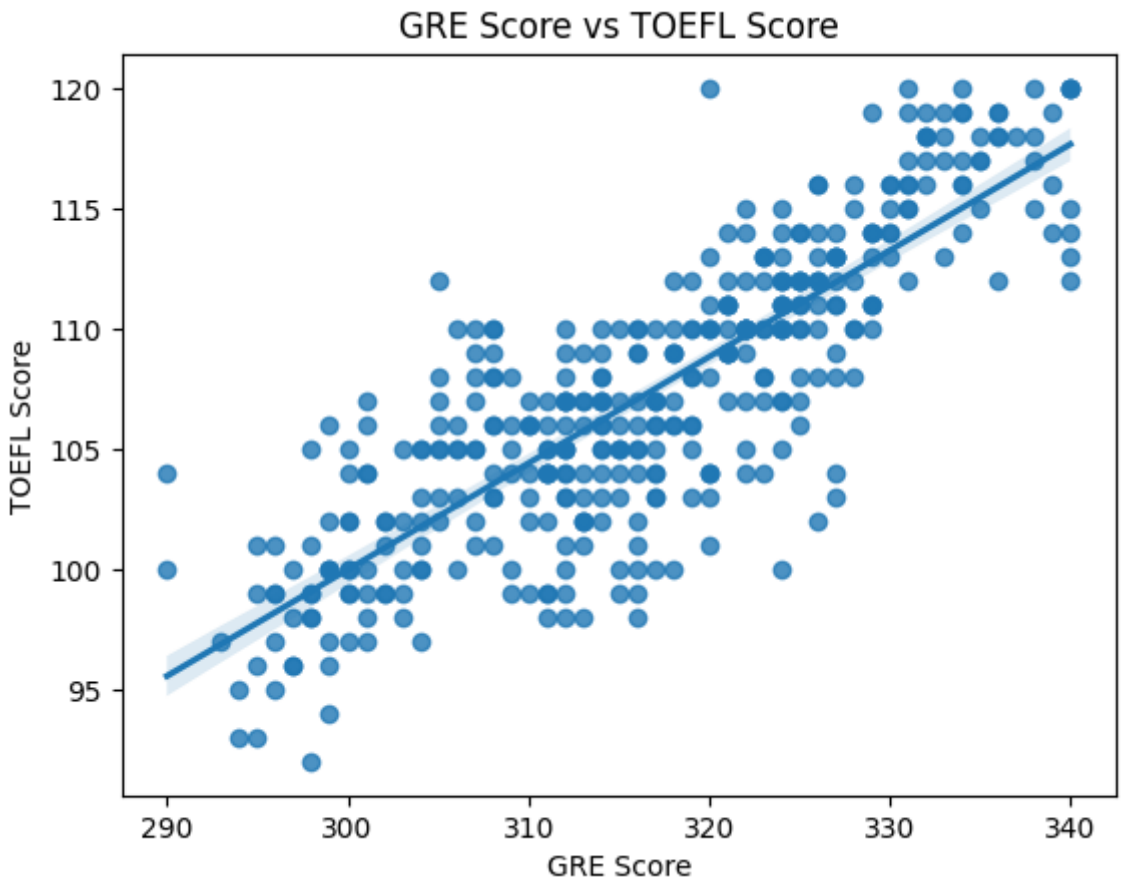

95
100
105
110
115
120
TOEFL Score


1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
University Rating


1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SOP



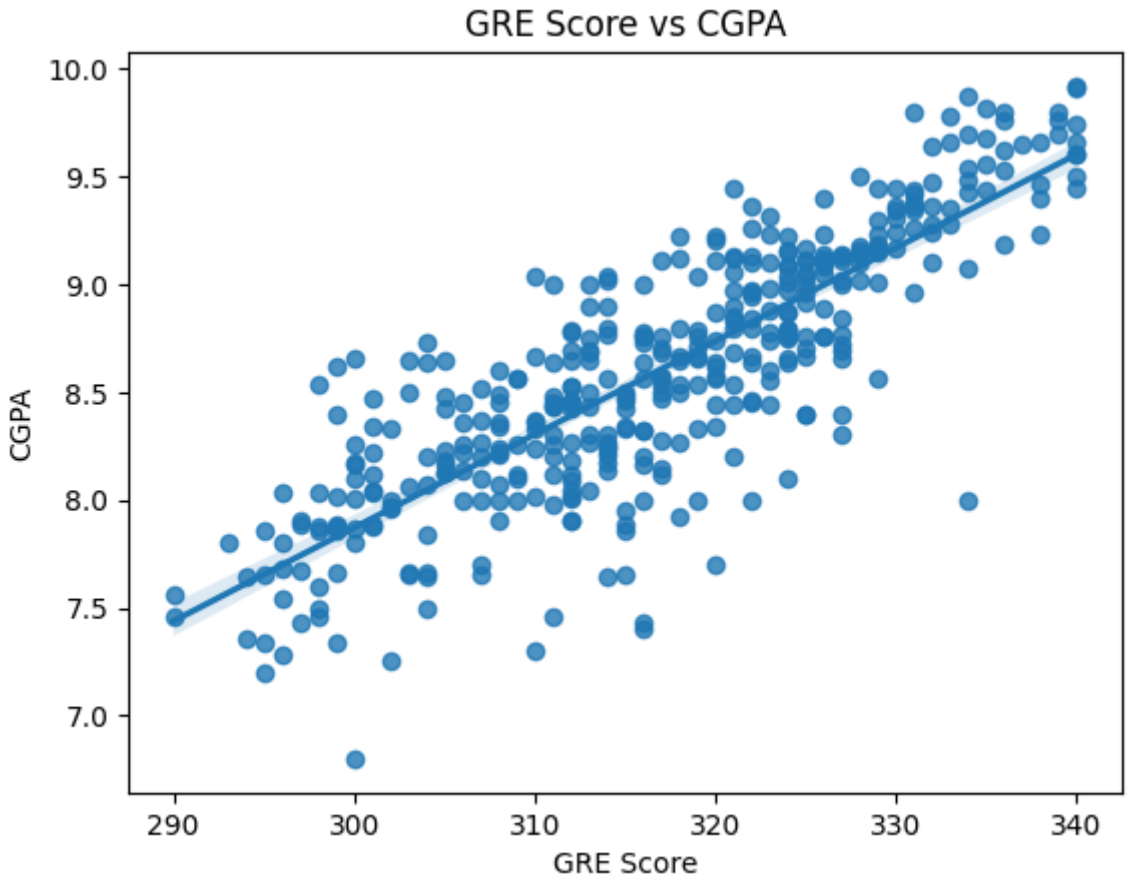
In [7]:
```
fig = sns.regplot(x="GRE Score", y="TOEFL Score", data=df)
plt.title("GRE Score vs TOEFL Score")
plt.show()
```



Люди с более высокими баллами GRE также имеют более высокие баллы TOEFL, что оправдано, тк они связаны между собой (хоть и косвенно)

- GRE — тест, который необходимо сдавать для поступления в аспирантуру, магистратуру или иной последипломный курс в вузе США и ряда других стран.
- TOEFL — стандартизованный тест на знание английского языка, результаты которого могут использоваться для подтверждения уровня владения английским языком абитуриентами из неанглоязычных стран при поступлении в вузы США, а также Европы и Азии.

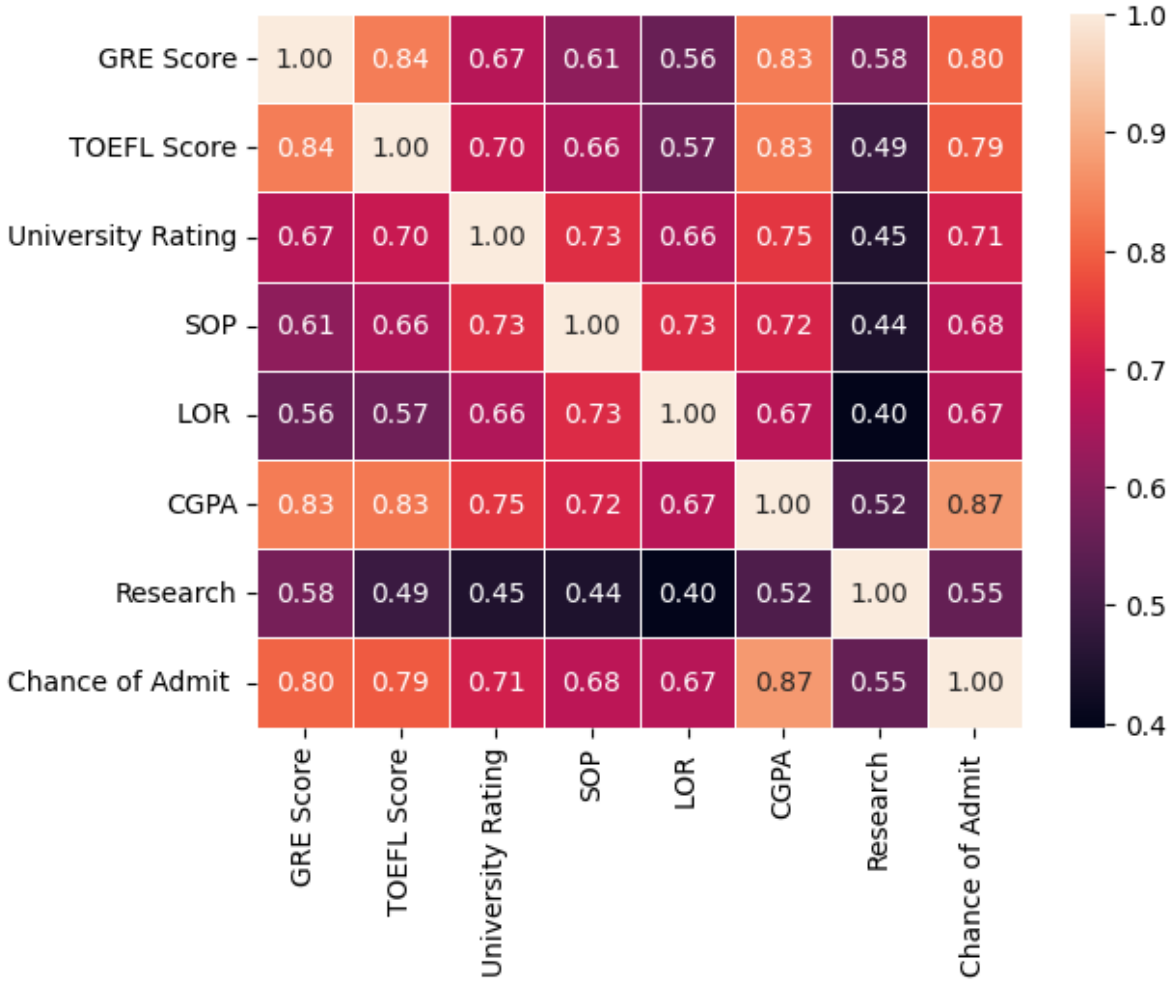
In [8]:
```
fig = sns.regplot(x="GRE Score", y="CGPA", data=df)
plt.title("GRE Score vs CGPA")
plt.show()
```



Люди с более высоким CGPA обычно имеют более высокие баллы GRE

- Систе́ма оце́нивания зна́ний — система оценивания качества освоения образовательных программ учащимся, важнейший элемент образовательного процесса.

In [15]:
```
corr = df.corr()
sns.heatmap(corr, linewidths=.5, annot=True, fmt=".2f")
plt.show()
```



1) С целевым признаком "Chance of Admit" наиболее коррелируют признаки "CGPA", "GRE Score" , "TOEFL Score". При построении модели машинного обучения перечисленные признаки будут наиболее информативными.

2) Стоит отметить корреляцию признаков "SOP" и "University Rating".

3) Можно построить модель машинного обучения на основе признаков "CGPA", "GRE Score", "TOEFL Score", "LOR", "Research". Первые 3 признака наиболее сильно повлияют на результат ввиду их высокой корреляции. Обученные модели позволят бакалаврам оценить свои возможности для поступления на магистратуру.