Московский государственный технический
университет им. Н. Э. Баумана
Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Рубежный контроль №1

Группа: РТ5-61

Студент: Коржов С.Ю.

Преподаватель: Гапанюк Ю.Е.

Москва, 2020 г.

Тема: Технологии разведочного анализа и обработки данных.

Задание:

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 произвольной колонки данных построить гистограмму.

Набор данных: https://www.kaggle.com/mohansacharya/graduate-admissions (файл Admission_Predict_Ver1.1.csv)

Дятленко Елена Александровна Группа ИУ5-62Б

In [0]:
```
import pandas as pd
import numpy as np
```

In [0]:
```
data = pd.read_csv('Admission_Predict_Ver1.1.csv')
```

In [5]:
```
data.head()
```

Out[5]:

	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 [6]:
```
data.columns
```

Out[6]:
```
Index(['Serial No.', 'GRE Score', 'TOEFL Score', 'University Rating', 'SOP',
       'LOR ', 'CGPA', 'Research', 'Chance of Admit '],
      dtype='object')
```

In [7]:
```
data.shape
```

Out[7]:
```
(500, 9)
```

In [0]:
```
data = data.dropna()
```

In [12]:
```
data.shape
```

Out[12]:
```
(500, 9)
```

In [13]:
```
data.isnull().sum()
```

Out[13]:
```
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
```

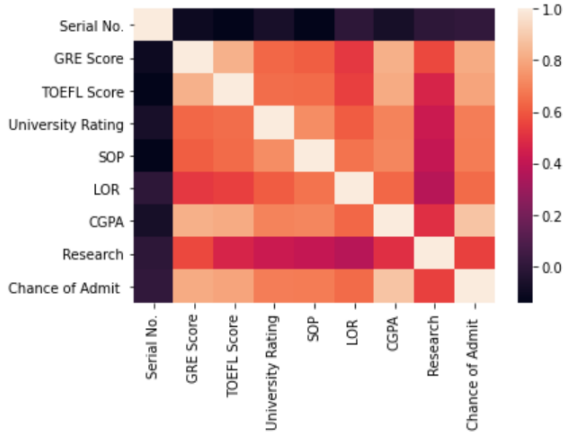
In [14]:
```
data.corr()
```

Out[14]:

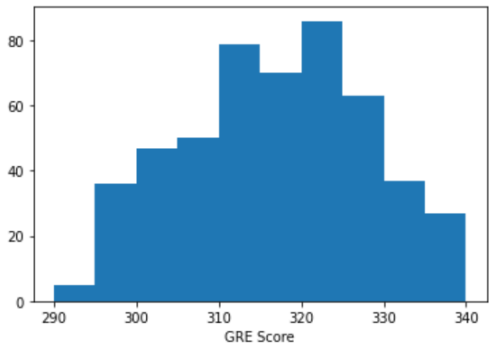
	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.103839	-0.141696	-0.067641	-0.137352	-0.003694	-0.074289	-0.005332	0.008505
GRE Score	-0.103839	1.000000	0.827200	0.635376	0.613498	0.524679	0.825878	0.563398	0.810351
TOEFL Score	-0.141696	0.827200	1.000000	0.649799	0.644410	0.541563	0.810574	0.467012	0.792228
University Rating	-0.067641	0.635376	0.649799	1.000000	0.728024	0.608651	0.705254	0.427047	0.690132
SOP	-0.137352	0.613498	0.644410	0.728024	1.000000	0.663707	0.712154	0.408116	0.684137
LOR	-0.003694	0.524679	0.541563	0.608651	0.663707	1.000000	0.637469	0.372526	0.645365
CGPA	-0.074289	0.825878	0.810574	0.705254	0.712154	0.637469	1.000000	0.501311	0.882413
Research	-0.005332	0.563398	0.467012	0.427047	0.408116	0.372526	0.501311	1.000000	0.545871
Chance of Admit	0.008505	0.810351	0.792228	0.690132	0.684137	0.645365	0.882413	0.545871	1.000000

In [21]:
```
import seaborn as sns
corr = data.corr()
sns.heatmap(corr,
xticklabels=corr.columns.values,
yticklabels=corr.columns.values)
```

Out[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3c6836fa20>
```



In [32]:
```
plt.hist(data['GRE Score'], )
plt.xlabel('GRE Score')
plt.show()
```



Корреляция данных высокая, следовательно, можно построить модель машинного обучения. Все признаки полезны, кроме "Serial No.", т.к. корреляция у данного столбца отсутвует (это номер строки).