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 №1
«Технологии разведочного анализа и обработки данных»
Вариант №5

Выполнил:
студент группы ИУ5-62Б
Долинский Александр
Александрович

Подпись:_____

Дата:_____

Проверил:
преподаватель каф. ИУ5
Гапанюк Юрий
Евгеньевич

Подпись:_____

Дата:_____

Москва, 2023 г.

Задача

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.

Дополнительное требование

Для студентов групп ИУ5-62Б, ИУ5Ц-82Б - для произвольной колонки данных построить гистограмму.

Выполнение работы

Для выполнения задачи проведения корреляционного анализа данных был использован набор данных Admission_Predict.

In [5]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [6]:
```
data = pd.read_csv('Admission_Predict.csv', sep=',');
```

In [7]:
```
data.head()
```

Out[7]:

	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

Типы данных всех полей являются числовыми.

In [9]:
```
data.dtypes
```

Out[9]:
```
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
```

В наборе данных отсутствуют пропуски и дубликаты.

```
In [8]: for col in data.columns:
            temp_null_count = data[data[col].isnull()].shape[0]
            print('{} - {}'.format(col, temp_null_count))
```

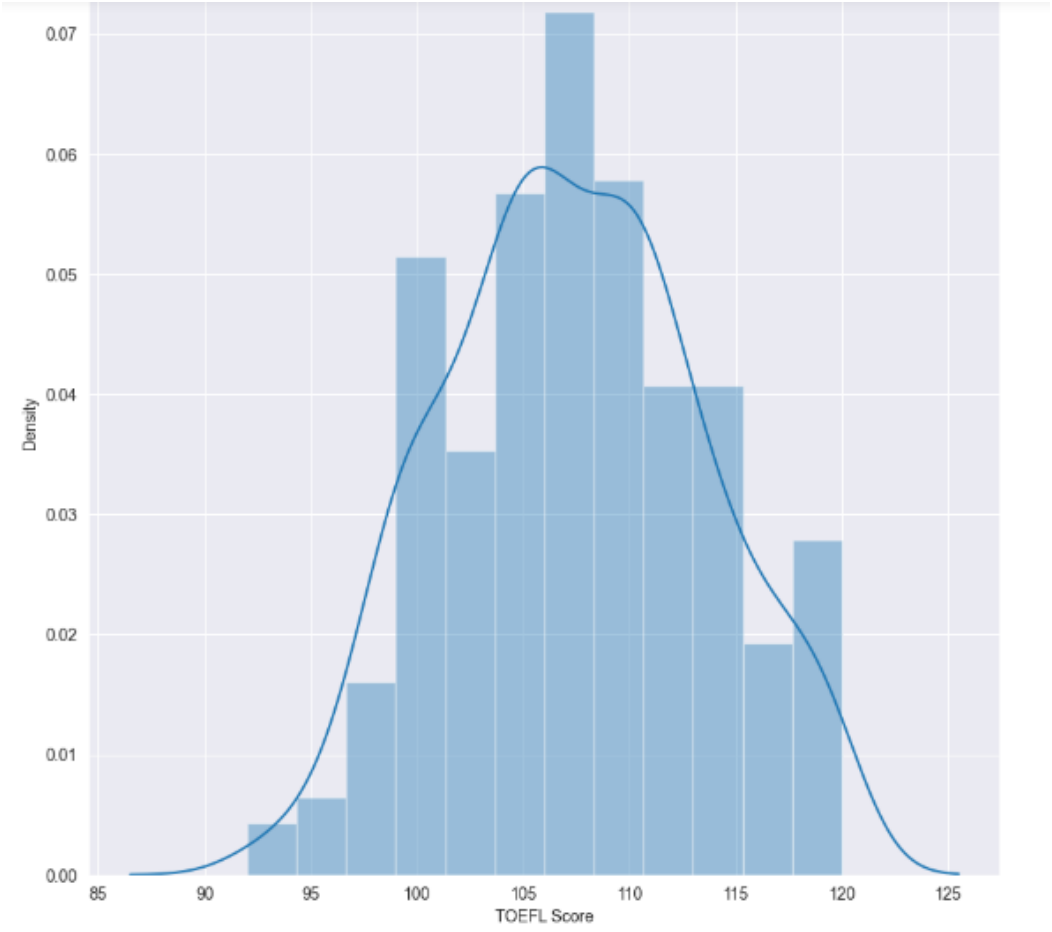
```
Out[8]: Serial No.           0
        GRE Score            0
        TOEFL Score          0
        University Rating    0
        SOP                  0
        LOR                  0
        CGPA                 0
        Research             0
        Chance of Admit      0
        dtype: int64
```

```
In [16]: data.duplicated().sum()
```

```
Out[16]: 0
```

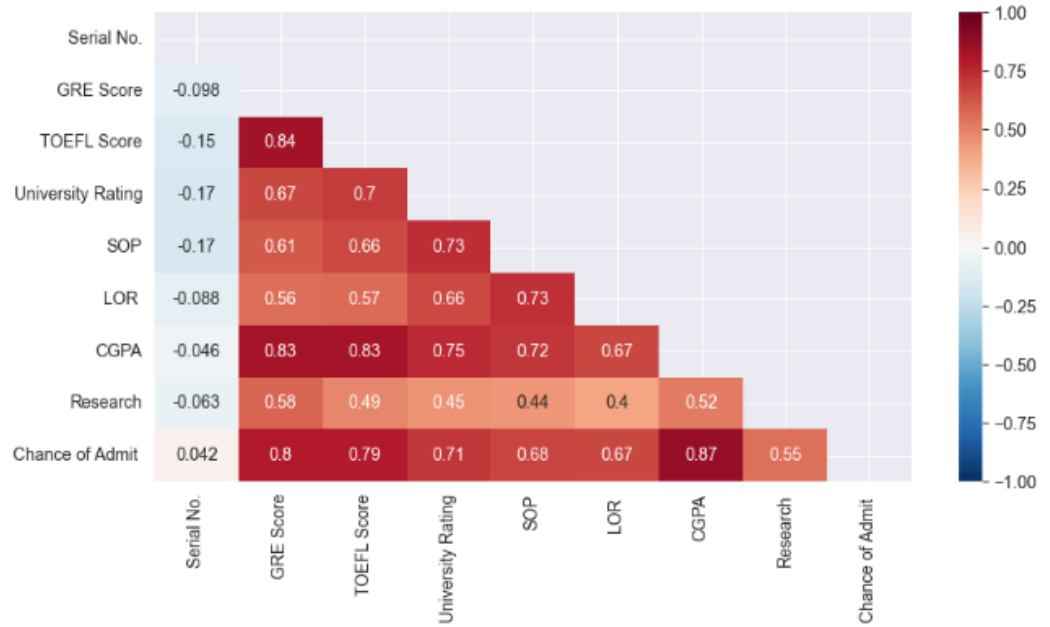
Построю гистограмму для колонки “TOEFL Score”.

```
In [25]: fig, ax = plt.subplots(figsize=(10,10))
         sns.distplot(data['TOEFL Score'])
```



Для визуализации корреляционной матрицы была использована “тепловая карта”.

```
In [23]: plt.figure(figsize=(10, 5))
         mask=np.triu(np.ones_like(data.corr(), dtype=bool))
         sns.heatmap(data.corr(), mask=mask, annot=True, vmin=-1.0, vmax=1, center=0, cmap='RdBu_r')
Out[23]: <Axes: >
```



С целевым признаком “Chance of Admit” наиболее коррелируют признаки “CGPA” (0,87), “GRE Score” (0,8), “TOEFL Score” (0,79). При построениимодели машинного обучения перечисленные признаки будут наиболее информативными.

Целевой признак “Chance of Admit” коррелирует с признаками “University Rating ” (0,71), SOP (0,68), LOR (0,67) и Research (0,55) которые также можно применять в процессе обучения модели.

Признак “Serial No.” не коррелирует не только с целевым признаком (0,042), но и со всеми остальными ввиду того, что предназначен для нумерации записей в наборе данных. Такой признак не привнесёт пользы в обучение моделей, и его следует изъять.

Стоит отметить корреляцию признаков “SOP” и “University Rating” (0,73). Ввиду того, что оценка рекомендательного письма зависит от статуса университета, можно не учитывать “SOP” при обучении модели, заменив этот признак более весомым “University Rating”.

Наконец, 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.