Рубежный контроль №1

Коваленко Артём, ИУ5-64, Вариант №7, Набор данных №7

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

Дополнительное задание

Для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)"

Решение

Подключим необходимые библиотеки и загрузим набор данных

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

%matplotlib inline

# Устанавливаем тип графиков
sns.set(style="ticks")

# Для лучшего качествоа графиков
from IPython.display import set_matplotlib_formats
set_matplotlib_formats("retina")

# Устанавливаем ширину экрана для отчета
pd.set_option("display.width", 70)

# Загружаем данные
data = pd.read_csv('Admission_Predict_Ver1.1.csv')
data.head()
```

Out[1]:

	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 [2]:

```
data.dtypes
```

Out[2]:

```
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
```

In [3]:

```
data.shape
```

Out[3]:

```
(500, 9)
```

In [4]:

```
data.isnull().sum()
```

Out[4]:

```
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
```

Как видим, в наборе данных отсутствуют пропуски

Проведем корреляционный анализ

In [5]:

```
data.corr()
```

Out[5]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Resea
Serial No.	1.000000	-0.103839	-0.141696	-0.067641	-0.137352	-0.003694	-0.074289	-0.005
GRE Score	-0.103839	1.000000	0.827200	0.635376	0.613498	0.524679	0.825878	0.563
TOEFL Score	-0.141696	0.827200	1.000000	0.649799	0.644410	0.541563	0.810574	0.467
University Rating	-0.067641	0.635376	0.649799	1.000000	0.728024	0.608651	0.705254	0.427
SOP	-0.137352	0.613498	0.644410	0.728024	1.000000	0.663707	0.712154	0.408
LOR	-0.003694	0.524679	0.541563	0.608651	0.663707	1.000000	0.637469	0.372
CGPA	-0.074289	0.825878	0.810574	0.705254	0.712154	0.637469	1.000000	0.501
Research	-0.005332	0.563398	0.467012	0.427047	0.408116	0.372526	0.501311	1.000
Chance of Admit	0.008505	0.810351	0.792228	0.690132	0.684137	0.645365	0.882413	0.545

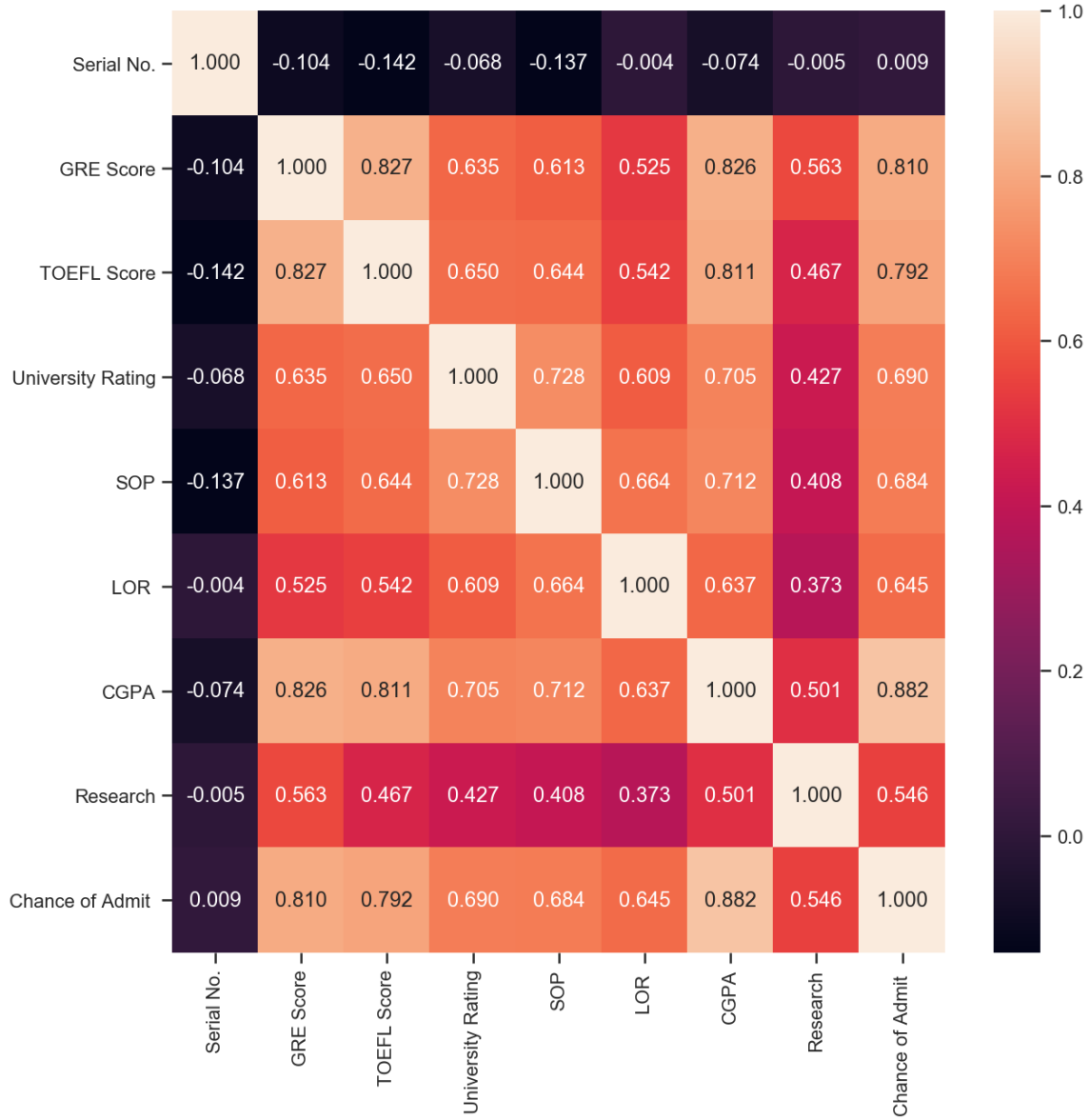
Построим тепловую карту

In [6]:

```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a5899bfb80>
```



На основе корреляционной матрицы можно сделать следующие выводы:

- Признак Serial No. можно исключить из модели, так как он слабо коррелирует с целевым признаком;
- Целевой признак наиболее сильно коррелирует с признаком CGPA. Его обязательно нужно оставить;
- Признак CGPA сильно зависим с признаками GRE Score, TOEFL Score. Их можно будет попробовать исключить из модели;
- Также можно попробовать исключить признак Research, так как он слабо коррелирует с целевым признаком, но он слабо зависим от других признаков, поэтому его можно оставить.

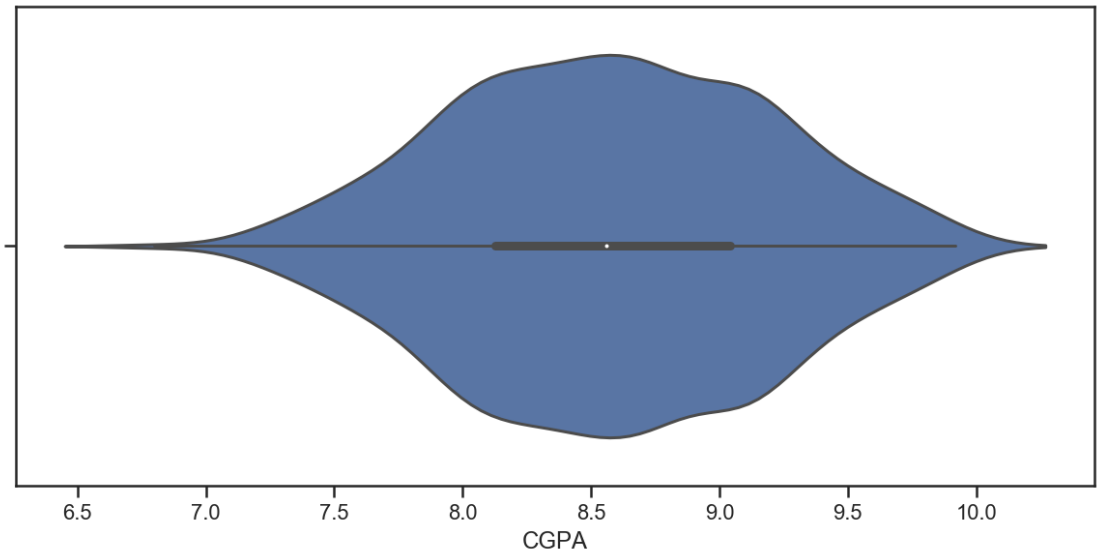
Построим violin plot для колонки CGPA

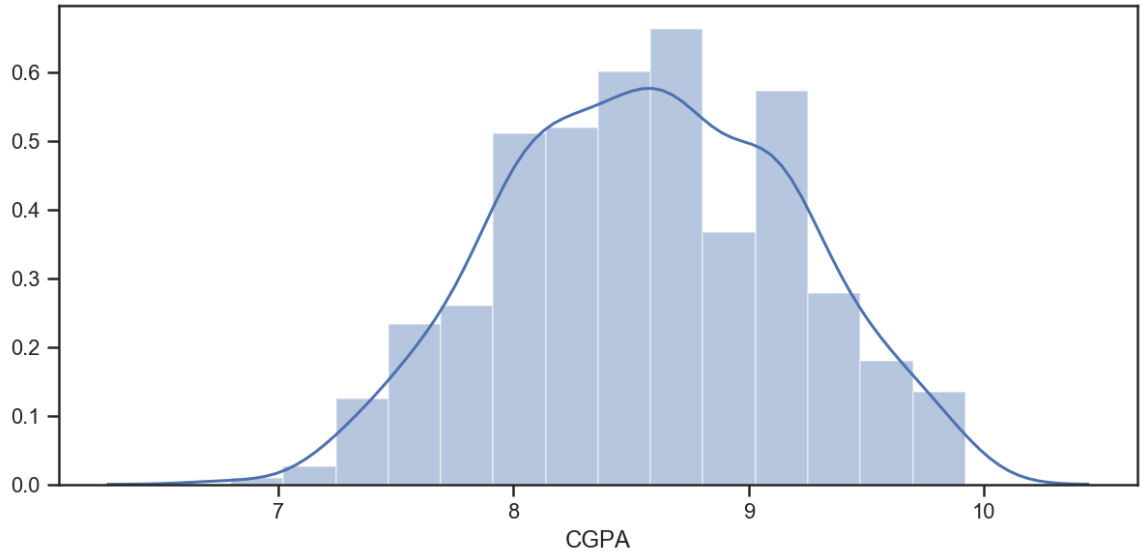
In [7]:

```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['CGPA'])
sns.distplot(data['CGPA'], ax=ax[1])
```

Out[7]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a58d42eee0>
```





In [ ]:

In [ ]: