МГТУ им. Н. Э. Баумана

кафедра ИУ5

курс «Технологии машинного обучения»

Рубежный контроль №1

«Технологии разведочного анализа и обработки данных»

Вариант 6

ВЫПОЛНИЛ:

Климанов А.С.

Группа

ИУ5ц-81Б

ПРОВЕРИЛ:

Гапанюк Ю. Е.

Москва, 2020 г.

Номер варианта	Номер задачи	Номер набора данных, указанного в задаче
6	1	6

Задача №1

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.

Дополнительные требования по группам:

Для студентов групп ИУ5-61Б - для пары произвольных колонок данных построить график «Диаграмма рассеяния».

Наборы данных:

6. https://www.kaggle.com/mohansacharya/graduate-admissions
файл `Admission_Predict.csv`

Выполнение работы

1. Загрузка данных

```
In [47]: import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         %matplotlib inline
         sns.set(style="ticks")
```

```
In [48]: data = pd.read_csv('data/Admission_Predict.csv', sep=",")
```

2. Первичный анализ и удаление пропусков

```
In [49]: data.shape
```

```
Out[49]: (400, 9)
```

```
In [50]: data.dtypes
```

```
Out[50]: 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 [51]:
```
data.isnull().sum()
```

Out[51]:
```
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 [52]:
```
data.head()
```

Out[52]:

	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

Как видно из результата функции data.isnull().sum(), пропусков нет.

3. Корреляционный анализ:

In [57]:
```
# Можно анализировать исходные данные, т.к. пустых нет
data.corr(method='pearson')
```

Out[57]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.097526	-0.147932	-0.169948	-0.166932	-0.088221	-0.045608	-0.063138	0.042336
GRE Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.580391	0.802610
TOEFL Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.489858	0.791594
University Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.447783	0.711250
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.444029	0.675732
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.396859	0.669889
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.521654	0.873289
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.000000	0.553202
Chance of Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.553202	1.000000

In [58]:
```
data.corr(method='kendall')
```

Out[58]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.064791	-0.101776	-0.119730	-0.121769	-0.057908	-0.033369	-0.051616	0.020912
GRE Score	-0.064791	1.000000	0.667509	0.544494	0.467137	0.414973	0.659007	0.492727	0.639933
TOEFL Score	-0.101776	0.667509	1.000000	0.567294	0.514420	0.424169	0.653665	0.421512	0.625485
University Rating	-0.119730	0.544494	0.567294	1.000000	0.638358	0.547389	0.611896	0.411914	0.599076
SOP	-0.121769	0.467137	0.514420	0.638358	1.000000	0.600787	0.565253	0.385807	0.545497
LOR	-0.057908	0.414973	0.424169	0.547389	0.600787	1.000000	0.510758	0.350789	0.518433
CGPA	-0.033369	0.659007	0.653665	0.611896	0.565253	0.510758	1.000000	0.434767	0.720655
Research	-0.051616	0.492727	0.421512	0.411914	0.385807	0.350789	0.434767	1.000000	0.480270
Chance of Admit	0.020912	0.639933	0.625485	0.599076	0.545497	0.518433	0.720655	0.480270	1.000000

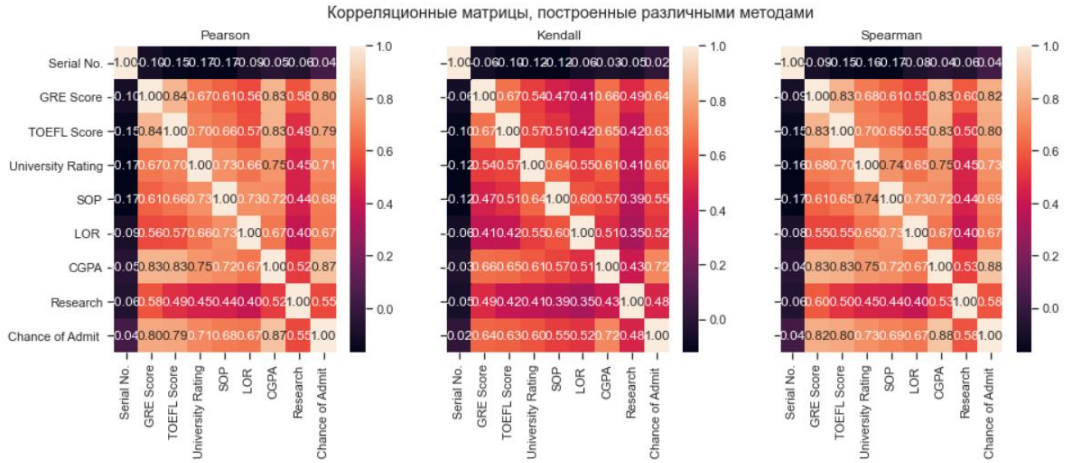
In [59]:
```
data.corr(method='spearman')
```

Out[59]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.093525	-0.146489	-0.161542	-0.170409	-0.081427	-0.042829	-0.063138	0.038328
GRE Score	-0.093525	1.000000	0.831860	0.676265	0.613743	0.547786	0.831848	0.595911	0.815352
TOEFL Score	-0.146489	0.831860	1.000000	0.696868	0.652922	0.549405	0.825720	0.504322	0.795573
University Rating	-0.161542	0.676265	0.696868	1.000000	0.740387	0.653256	0.750562	0.454131	0.731977
SOP	-0.170409	0.613743	0.652922	0.740387	1.000000	0.727178	0.724348	0.443648	0.694715
LOR	-0.081427	0.547786	0.549405	0.653256	0.727178	1.000000	0.666012	0.400385	0.670562
CGPA	-0.042829	0.831848	0.825720	0.750562	0.724348	0.666012	1.000000	0.530265	0.878403
Research	-0.063138	0.595911	0.504322	0.454131	0.443648	0.400385	0.530265	1.000000	0.581742
Chance of Admit	0.038328	0.815352	0.795573	0.731977	0.694715	0.670562	0.878403	0.581742	1.000000

Более наглядное представление в виде тепловой карты:

In [60]:
```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```



Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков.

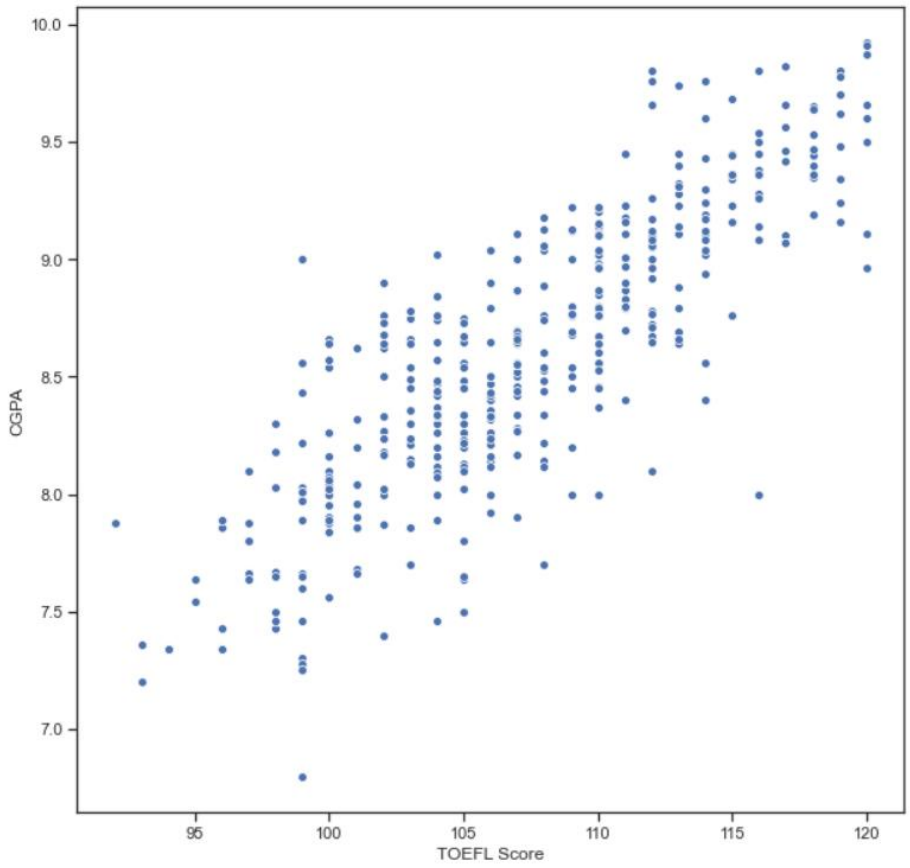
На основе сформированной корреляционной матрицы можно сделать следующие выводы:

1. Целевой признак `Chance of Admit` очень сильно коррелирует со всеми признаками, кроме `Serial No.` и `Research`. Эти признаки обязательно нужно оставить в модели.
2. Целевой признак отчасти коррелирует с `Research`. Этот признак стоит также оставить в модели.
3. Целевой признак слабо коррелирует с `Serial No.` (0,02-0,03). Его нужно убрать из модели, т.к. он, возможно, ухудшает ее качество.
4. Нет сильной взаимной корреляции между другими признаками. Нет необходимости убирать какой-либо из них.

4. Диаграмма рассеяния:

In [64]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='TOEFL Score', y='CGPA', data=data)
```

Out[64]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xd4226f0>
```



Таким образом, согласно диаграмме рассеяния, выбранные значения (`TOEFL Score` и `CGPA`) нормально коррелируют и имеют положительный тип корреляции (оба значения увеличиваются).