Желтова Александра Алексеевна, ИУ5-63Б

Рубежный контроль №1

Вариант 7

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.

Доп задание: для произвольной колонки данных построить график "Ящик с усами (boxplot)".

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

[2]:
```
data = pd.read_csv('AdmPre1.1.csv', sep=",")
```

[3]:
```
data.shape
```

[3]:
```
(500, 9)
```

[4]:
```
data.dtypes
```

[4]:
```
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
```

[5]:
```
data.isnull().sum()
```

[5]:
```
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
```

[6]:
```
data.head()
```

[6]:

	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

[7]:
```
data.describe()
```

[7]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

[8]:
```
#уникальные значения для целевого признака
data['Research'].unique()
```

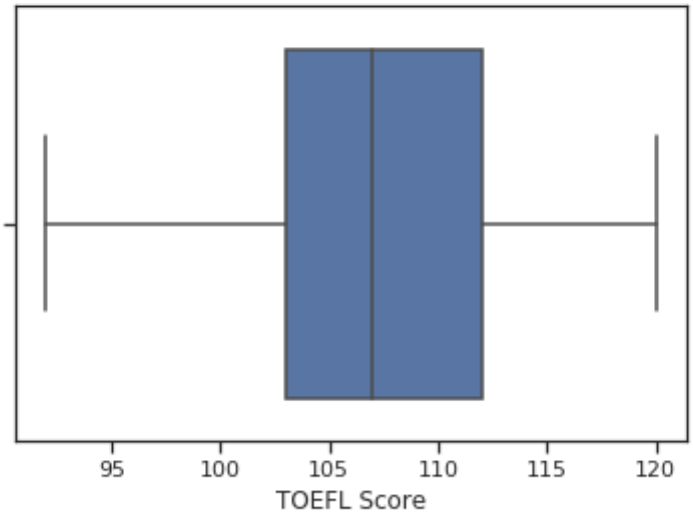
[8]:
```
array([1, 0])
```

Целевой признак бинарный, т.к. содержит только значения 0 и 1

Ящик с усами

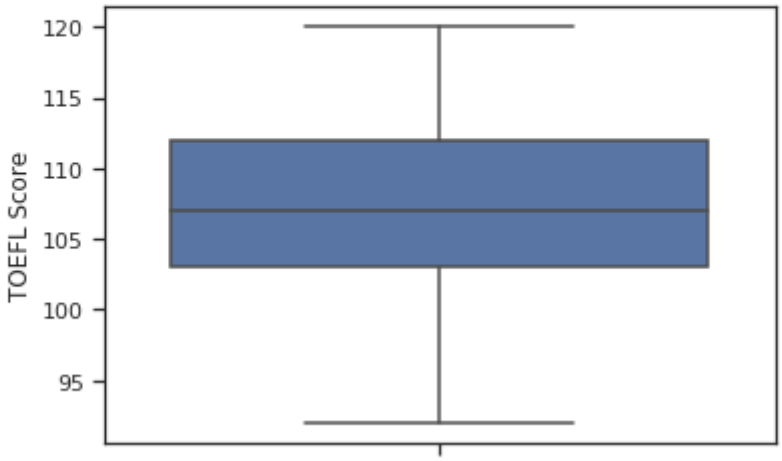
```
[10]: #одномерное распределение вероятности
      sns.boxplot(x=data['TOEFL Score'])
```

```
[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb034120210>
```




```
[11]: sns.boxplot(y=data['TOEFL Score'])
```

```
[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb034094890>
```



Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере это колонка "Occupancy"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно

отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.

2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

```
[12]: data.corr()
```

[12]:

	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

Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков.

Корреляционная матрица симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой).

На основе корреляционной матрицы можно сделать следующие выводы:

- Целевой признак наиболее сильно коррелирует с освещенностью (0.9) и концентрацией углекислого газа (0.71). Эти признаки обязательно следует оставить в модели.
- Целевой признак отчасти коррелирует с температурой (0.54). Этот признак стоит также оставить в модели.
- Целевой признак слабо коррелирует с влажностью (0.13) и HumidityRatio (0.3). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат качество модели.
- Влажность и HumidityRatio очень сильно коррелируют между собой (0.96). Это неудивительно, ведь HumidityRatio величина производная от влажности. Поэтому из этих признаков в модели можно оставлять только один.
- Также можно сделать вывод, что выбирая из признаков влажность и HumidityRatio лучше выбрать HumidityRatio, потому что он сильнее коррелирован с целевым признаком. Если линейно зависимые

признаки сильно коррелированы с целевым, то оставляют именно тот признак, который коррелирован с целевым сильнее. Но для этой пары признаков этот вывод нельзя считать надежным, потому что и 0.13 и 0.3 являются довольно малыми величинами.

По умолчанию при построении матрицы используется коэффициент корреляции Пирсона. Возможно также построить корреляционную матрицу на основе коэффициентов корреляции Кендалла и Спирмена. На практике три метода редко дают значимые различия.

[13]:
```
data.corr(method='pearson')
```

[13]:

	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

[14]:
```
data.corr(method='kendall')
```

[14]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.068496	-0.098656	-0.040534	-0.101583	0.002344	-0.053469	-0.004358	-0.005993
GRE Score	-0.068496	1.000000	0.655920	0.514842	0.475974	0.386159	0.651313	0.478379	0.647169
TOEFL Score	-0.098656	0.655920	1.000000	0.520345	0.504574	0.403507	0.635410	0.396523	0.622481
University Rating	-0.040534	0.514842	0.520345	1.000000	0.624569	0.497402	0.565745	0.394370	0.570844
SOP	-0.101583	0.475974	0.504574	0.624569	1.000000	0.535641	0.558255	0.355953	0.552719
LOR	0.002344	0.386159	0.403507	0.497402	0.535641	1.000000	0.485466	0.328867	0.494280
CGPA	-0.053469	0.651313	0.635410	0.565745	0.558255	0.485466	1.000000	0.417418	0.731828
Research	-0.004358	0.478379	0.396523	0.394370	0.355953	0.328867	0.417418	1.000000	0.467002
Chance of Admit	-0.005993	0.647169	0.622481	0.570844	0.552719	0.494280	0.731828	0.467002	1.000000

[15]:
```
data.corr(method='spearman')
```

[15]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.099592	-0.142607	-0.055424	-0.144249	0.004220	-0.075126	-0.005332	-0.001733
GRE Score	-0.099592	1.000000	0.823853	0.643423	0.620688	0.514352	0.829251	0.578487	0.822201
TOEFL Score	-0.142607	0.823853	1.000000	0.645533	0.644715	0.523434	0.809485	0.474540	0.793634
University Rating	-0.055424	0.643423	0.645533	1.000000	0.729399	0.602319	0.703333	0.435351	0.703742
SOP	-0.144249	0.620688	0.644715	0.729399	1.000000	0.662653	0.717384	0.409088	0.702799
LOR	0.004220	0.514352	0.523434	0.602319	0.662653	1.000000	0.639563	0.376166	0.643627
CGPA	-0.075126	0.829251	0.809485	0.703333	0.717384	0.639563	1.000000	0.509264	0.888786
Research	-0.005332	0.578487	0.474540	0.435351	0.409088	0.376166	0.509264	1.000000	0.565715
Chance of Admit	-0.001733	0.822201	0.793634	0.703742	0.702799	0.643627	0.888786	0.565715	1.000000

В случае большого количества признаков анализ числовой корреляционной матрицы становится неудобен.

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap которая показывает степень корреляции различными цветами.

```
[16]: # Вывод значений в ячейках
      sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

```
[16]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb03400fa90>
```