第一题

- a:分类器，推断。根据利润、员工人数等特征训练无监督分类模型，将美国500强进行分类，然后推断每一类在特征上的不同点。
- b:分类器，预测。将产品的价格成本等变量作为自变量，成功或失败作为因变量，训练二分类模型，然后根据产品数据预测产品是否会成功。
- c:回归模型，推断。将每周数据中美元百分比变化率作为因变量，其他全球股市周变动率作为自变量，训练回归模型，推断美元百分比变化率随全球股市周变动率而变动的规律。

第二题

a

In [6]:

```
import pandas as pd
college = pd.read_csv('../data/College.csv',index_col=[0])
college.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 777 entries, Abilene Christian University to York College of Pe
nnsylvania
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Private      777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
memory usage: 115.3+ KB
```

以上输出可以看到数据集college的基本信息及缺失值情况。

b

In [7]:

```
college.head()
```

Out[7]:

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12
Adrian College	Yes	1428	1097	336	22	50	1036	99	11
Agnes Scott College	Yes	417	349	137	60	89	510	63	12
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7

观察数据的前五行

c

1-5

In [11]:

```
college.describe()
```

Out[11]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Under
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.29
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.43
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.00
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.00
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.00
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.00
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.00

In [12]:

```
college['Private'].value_counts()
```

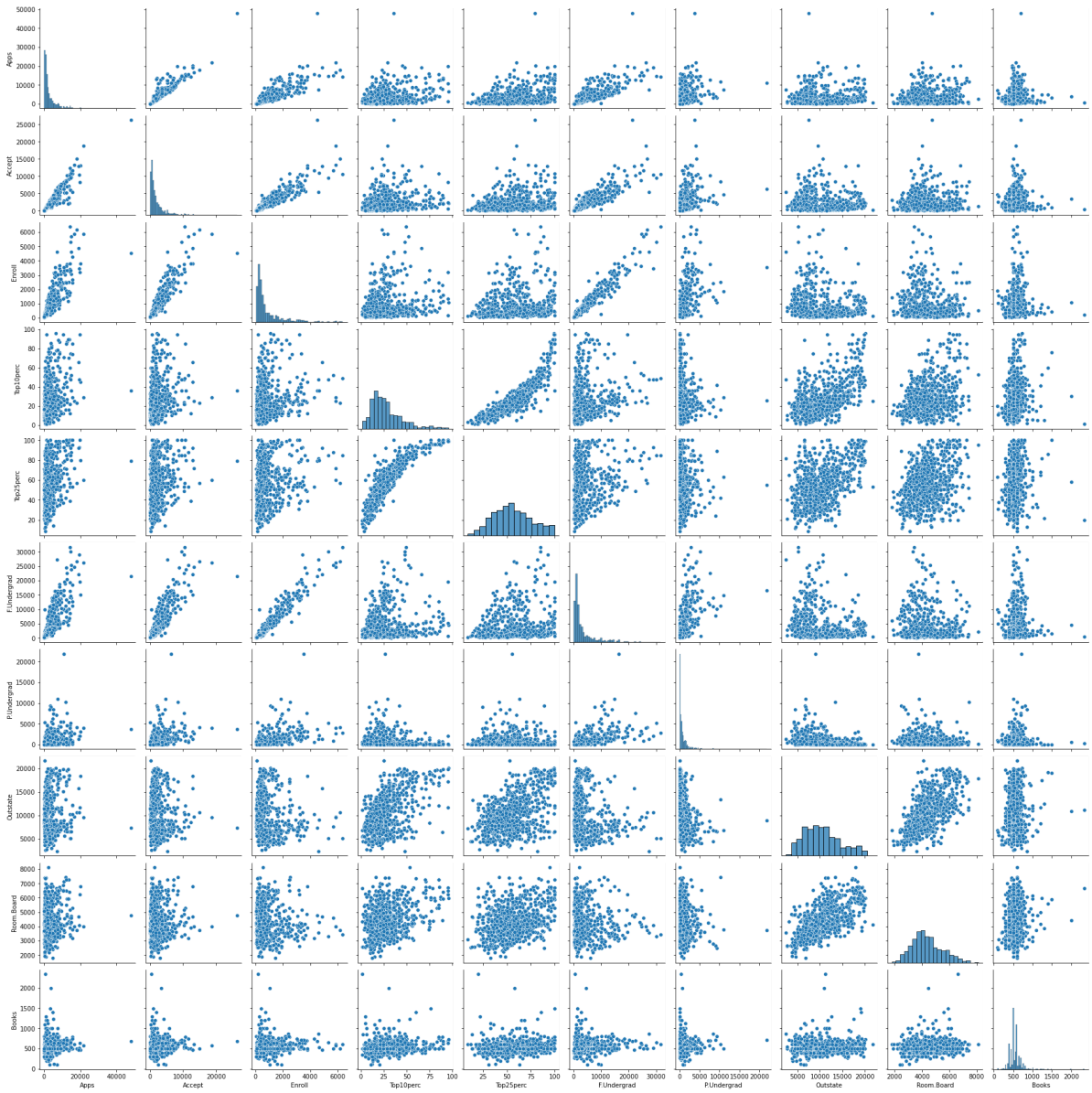
Out[12]:

```
Yes    565
No     212
Name: Private, dtype: int64
```

以上输出了每个变量的汇总信息，由于“Private”是分类变量，这里单独对其进行统计。

In [18]:

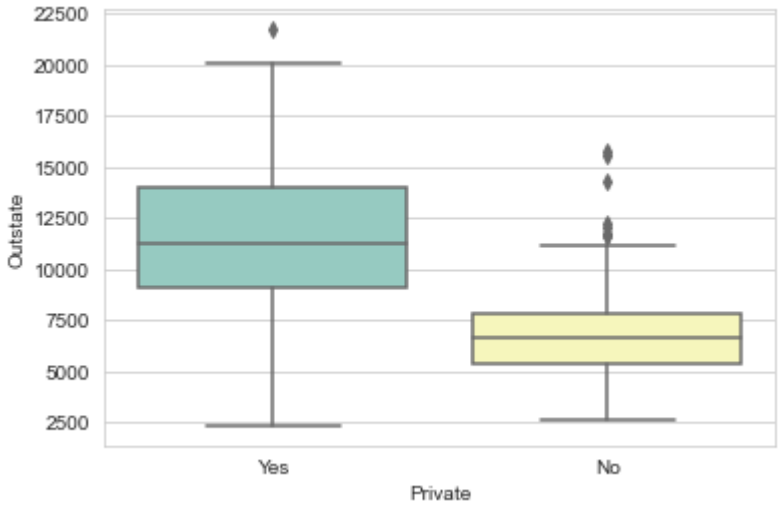
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.pairplot(college.iloc[:,[i for i in range(0,11)]])
plt.show()
```



以上是十个连续型变量绘制散点图矩阵。

In [20]:

```
import seaborn as sns
sns.set_style("whitegrid")
ax = sns.boxplot(x="Private", y="Outstate",
data=college, palette="Set3")
```



以上按照“Private”变量分组对“Outstate”绘制箱线图，可见私立学校非本州学生学费更高，但非私立学校非本州学生学费分布更集中。

In [24]:

```
college['Elite'] = 'No'
college.loc[college['Top10perc']>50,'Elite']='Yes'
college.head()
```

Out[24]:

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12
Adrian College	Yes	1428	1097	336	22	50	1036	99	11
Agnes Scott College	Yes	417	349	137	60	89	510	63	12
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7

成功增加最后一列，用来指示是否有百分之50以上的学生来自前百分之10的顶尖高中。

In [26]:

```
college['Elite'].value_counts()
```

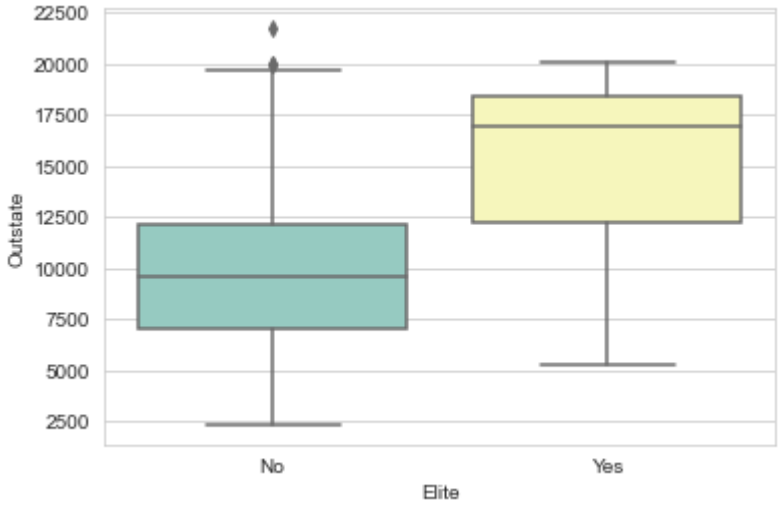
Out[26]:

```
No     699
Yes     78
Name: Elite, dtype: int64
```

有78所精英大学。

In [27]:

```
import seaborn as sns
sns.set_style("whitegrid")
ax = sns.boxplot(x="Elite", y="Outstate",
data=college, palette="Set3")
```



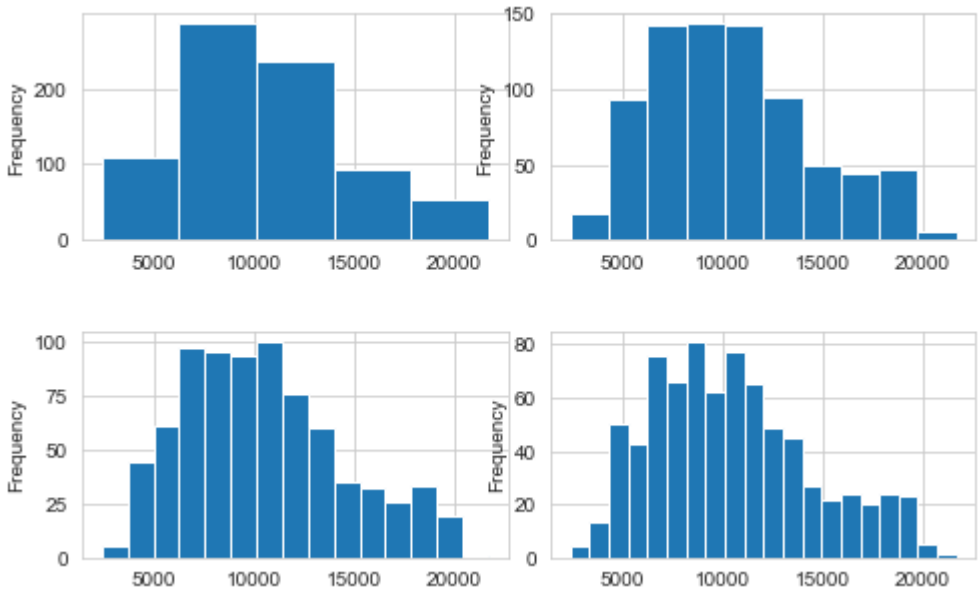
以上按照“Elite”变量分组对“Outstate”绘制箱线图，可见精英大学非本州学生学费更高，且学费数额的前百分之50非常集中。

In [35]:

```
fig = plt.figure(figsize = (8,5))
plt.subplots_adjust(wspace = 0.1) #设置两个图之间的横向间隔
plt.subplots_adjust(hspace = 0.4)#纵向间隔
ax1 = fig.add_subplot(2,2,1)
# plt.hist(college['Outstate'],ax = ax1)
college['Outstate'].plot(kind='hist',bins=5,grid=True,ax=ax1)
ax2 = fig.add_subplot(2,2,2)
college['Outstate'].plot(kind='hist',bins=10,grid=True,ax=ax2)
ax3 = fig.add_subplot(2,2,3)
college['Outstate'].plot(kind='hist',bins=15,grid=True,ax=ax3)
ax4 = fig.add_subplot(2,2,4)
college['Outstate'].plot(kind='hist',bins=20,grid=True,ax=ax4)
```

Out[35]:

```
<AxesSubplot:ylabel='Frequency'>
```



分别分5，10，15，20组绘制定量变量“Outstate”的直方图如上述所示。

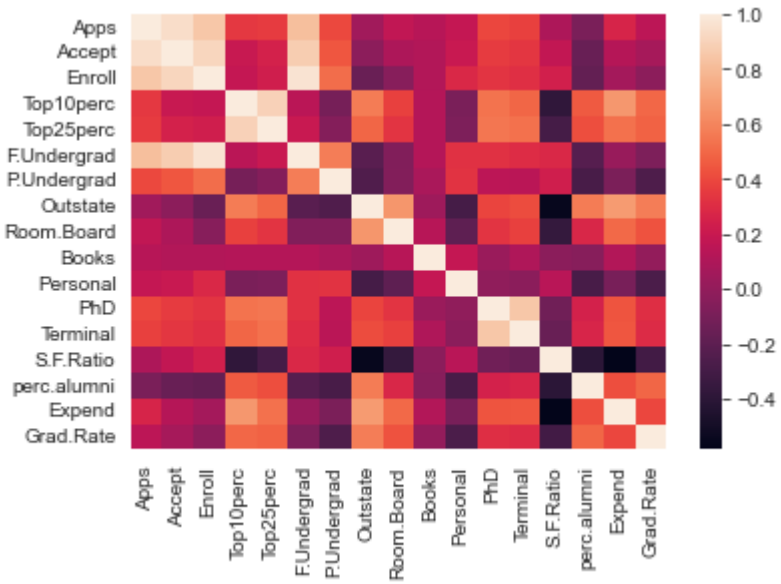
6:数据探索性分析

In [37]:

```
sns.heatmap(college.corr())
```

Out[37]:

<AxesSubplot:>



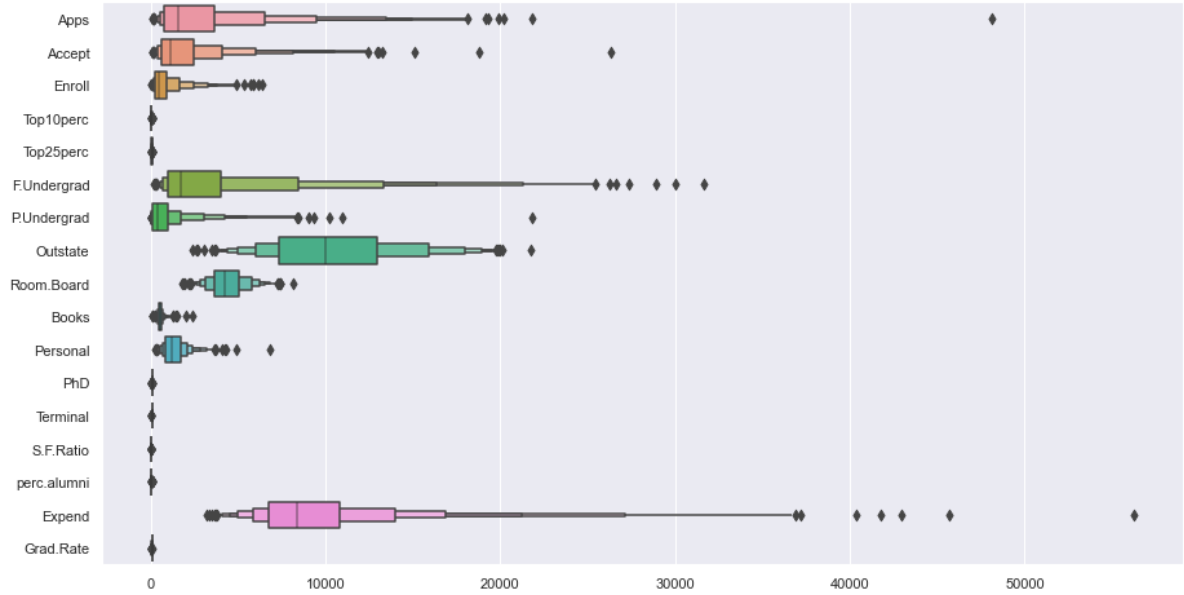
查看定量数据的皮尔逊相关性热力图。

In [42]:

```
sns.set(rc = {'figure.figsize':(15,8)})
sns.boxenplot(data = college, orient = "h")
```

Out[42]:

```
<AxesSubplot:>
```



为所有定量数据绘制增强的箱线图，可以看出各个变量的分布情况和取值对比。

第三题

- 学习小结：

  学习数据科学，知行合一尤为重要，理论学习和工具学习让我们掌握了许多“武器”，而最终用它们创造价值则是归宿。大数据时代，数据量大，速度快，种类多，作为个人学习，需要有一个从掌握通用技能到深挖某一领域（如视觉，图数据，自然语言，结构化数据等）的能力。

  所谓数据科学，既要掌握科学，又要了解数据，而数据的质量和使用容易被忽略，各大机构和研究所提出参数越来越大的模型，已经超越了优质公开数据集的发展速度，个人进行数据挖掘联系常常受制于数据质量；不过，也应当具体问题具体分析。

  二分类问题中，混淆矩阵、ROC曲线、AUC值的关系，准确率、精确率、召回率的概念和计算，其他评估指标，如F1_Score的计算需要掌握。

$$F_\beta = (1+\beta^2) \times \frac{precision \times recall}{(1+\beta^2) \times precision + recall}$$

  线性回归的“线性”概念指的是参数和变量之间是线性的（βX）。

  数据挖掘，可解释性是很重要的部分，可视化是一种比较生动的解释方法。数据预处理（清洗、整合、变换、特征选择、抽样）是数据挖掘中最有挑战性的部分。

  云计算可以促进计算资源高效应用。GPU/MIC并行计算能力较强。

  数据，算法，算力，三者缺一不可，数据挖掘是三方面的集合。

“没有免费的午餐”，即没有通用的算法，应用什么算法应该取决于业务。没有必要一味追求复杂算法，够用即可。