# Exercise 1 - Python Fundamentals and Data Handling

Name: Sona V
Reg. No: 2021239022
Course: M.Sc. Integrated Computer Science

**(a) Use the pd.read_csv() function to read the data into Python. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.**

In [2]:
```
import pandas as pd
```

In [57]:
```
college = pd.read_csv("/content/College.csv")
college.head(5)
```

Out[57]:

| | Unnamed: 0 | Private | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad |
|---|---|---|---|---|---|---|---|---|
| **0** | Abilene Christian University | Yes | 1660 | 1232 | 721 | 23 | 52 | 2885 |
| **1** | Adelphi University | Yes | 2186 | 1924 | 512 | 16 | 29 | 2683 |
| **2** | Adrian College | Yes | 1428 | 1097 | 336 | 22 | 50 | 1036 |
| **3** | Agnes Scott College | Yes | 417 | 349 | 137 | 60 | 89 | 510 |
| **4** | Alaska Pacific University | Yes | 193 | 146 | 55 | 16 | 44 | 249 |

**(b) Look at the data used in the notebook by creating and running a new cell with just the code college in it. You should notice that the first column is just the name of each university in a column named something like Unnamed: 0. We don't really want pandas to treat this as data. However, it may be handy to have these names for later. Try the following commands and similarly look at the resulting data frames:**

```
college2 = pd. read_csv ('College .csv ', index_col =0)
college3 = college.rename ({'Unnamed : 0': 'College '}, axis
```

```
=1)
college3 = college3 . set_index ('College ')
```

**This has used the first column in the file as an index for the data frame. This means that pandas has given each row a name corresponding to the appropriate university. Now you should see that the first data column is Private. Note that the names of the colleges appear on the left of the table. We also introduced a new python object above: a dictionary, which is specified by dictionary (key, value) pairs. Keep your modified version of the data with the following:**

```
college = college3
```

In [58]:
```
college = college.rename(columns={'Unnamed: 0': 'College'})
college = college.set_index('College')
college.head(5)
```

Out[58]:

| College | Private | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | F |
|---|---|---|---|---|---|---|---|---|
| **Abilene Christian University** | Yes | 1660 | 1232 | 721 | 23 | 52 | 2885 | |
| **Adelphi University** | Yes | 2186 | 1924 | 512 | 16 | 29 | 2683 | |
| **Adrian College** | Yes | 1428 | 1097 | 336 | 22 | 50 | 1036 | |
| **Agnes Scott College** | Yes | 417 | 349 | 137 | 60 | 89 | 510 | |
| **Alaska Pacific University** | Yes | 193 | 146 | 55 | 16 | 44 | 249 | |

**(c) Use the describe() method of to produce a numerical summary of the variables in the data set.**

In [59]:
```
college.describe()
```

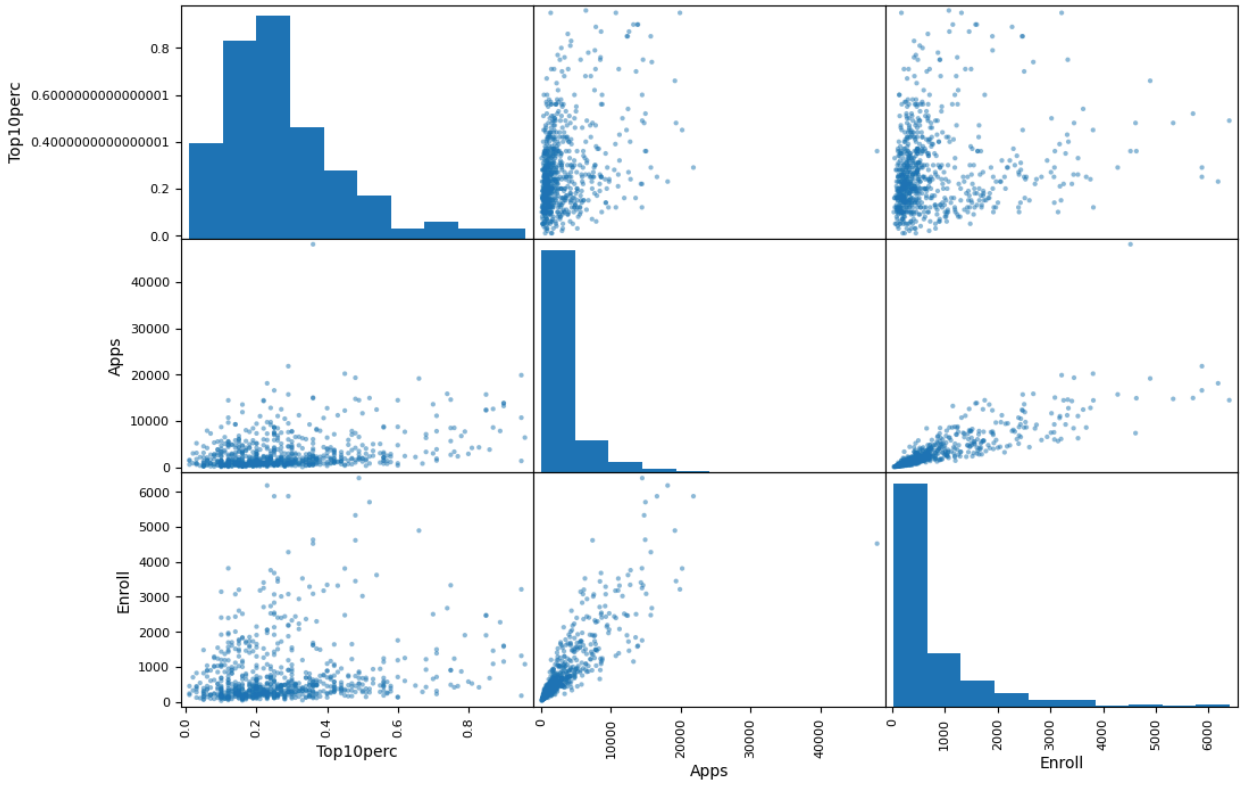
Out[59]:

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Unde |
|---|---|---|---|---|---|---|
| **count** | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.0 |
| **mean** | 3001.638353 | 2018.804376 | 779.972973 | 27.558559 | 55.796654 | 3699.9 |
| **std** | 3870.201484 | 2451.113971 | 929.176190 | 17.640364 | 19.804778 | 4850.4 |
| **min** | 81.000000 | 72.000000 | 35.000000 | 1.000000 | 9.000000 | 139.0 |
| **25%** | 776.000000 | 604.000000 | 242.000000 | 15.000000 | 41.000000 | 992.0 |
| **50%** | 1558.000000 | 1110.000000 | 434.000000 | 23.000000 | 54.000000 | 1707.0 |
| **75%** | 3624.000000 | 2424.000000 | 902.000000 | 35.000000 | 69.000000 | 4005.0 |
| **max** | 48094.000000 | 26330.000000 | 6392.000000 | 96.000000 | 100.000000 | 31643.0 |

**(d) Use the pd.plotting.scatter_matrix() function to produce a scatterplot matrix of the first columns [Top10perc, Apps, Enroll]. Recall that you can reference a list C of columns of a data frame A using A[C]**

In [60]:
```
import matplotlib.pyplot as plt
```

In [72]:
```
pd. plotting . scatter_matrix(college[['Top10perc', 'Apps', 'Enroll']], figsiz
```

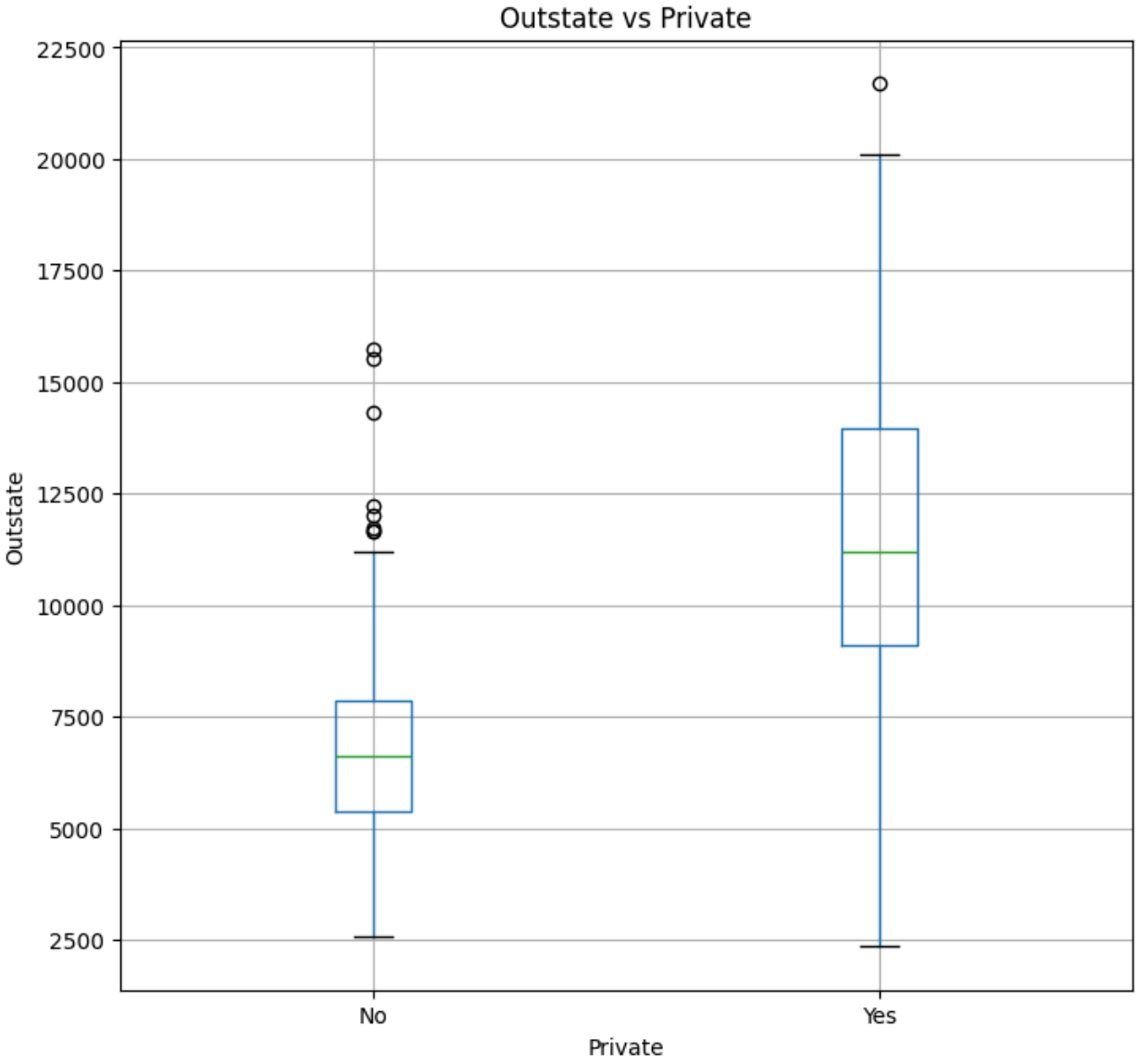



**(e) Use the boxplot() method of college to produce side-by-side boxplots of Outstate versus Private.**

In [64]:
```
import matplotlib.pyplot as plt
```

In [65]:
```
fig, ax = plt.subplots(figsize=(8, 8))
college.boxplot('Outstate', by='Private', ax=ax);
plt.suptitle('')
plt.title('Outstate vs Private')
plt.ylabel('Outstate')
```

Out[65]: Text(0, 0.5, 'Outstate')




**(f) Create a new qualitative variable, called Elite, by binning the Top10perc variable into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.**

```
college['Elite '] = pd.cut(college['Top10perc '], [0 ,0.5 ,1] ,
```

```
    labels =['No', 'Yes '])
```

**Use the value_counts() method of college['Elite'] to see how many elite universities there are. Finally, use the boxplot() method again to produce side-by-side boxplots of Outstate versus Elite**

In [66]:
```
college['Top10perc'] = college['Top10perc']/100
college['Elite'] = pd.cut(college['Top10perc'], [0, 0.5, 1], labels=['No', 'Ye
college['Elite'].value_counts()
```

Out[66]:

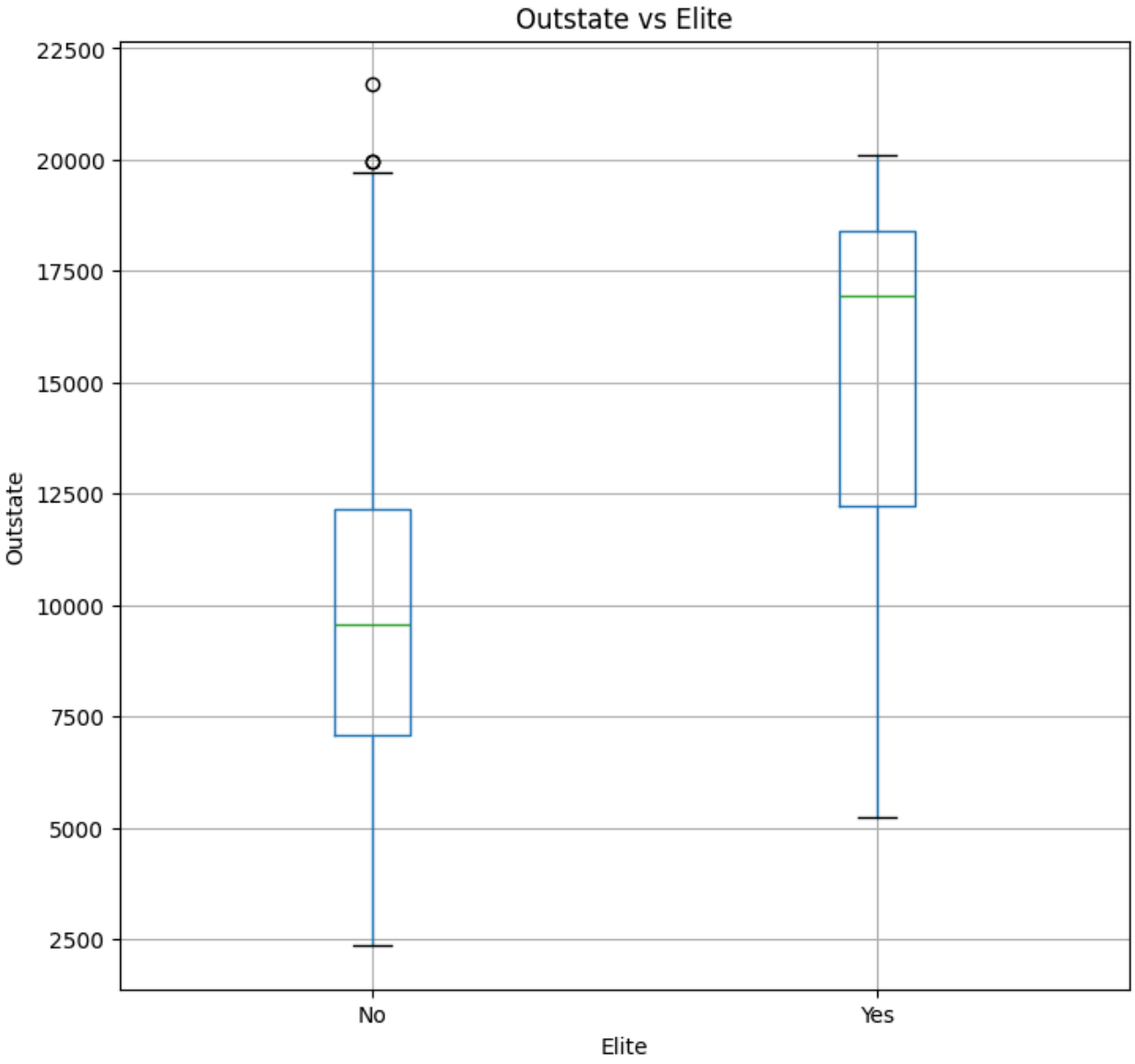
| Elite | count |
|---|---|
| **No** | 699 |
| **Yes** | 78 |

**dtype:** int64

In [73]:
```
import matplotlib.pyplot as plt
```

In [67]:
```
fig, ax = plt.subplots(figsize=(8, 8))
college.boxplot('Outstate', by='Elite', ax=ax);
plt.suptitle('')
plt.title('Outstate vs Elite')
plt.ylabel('Outstate')
```

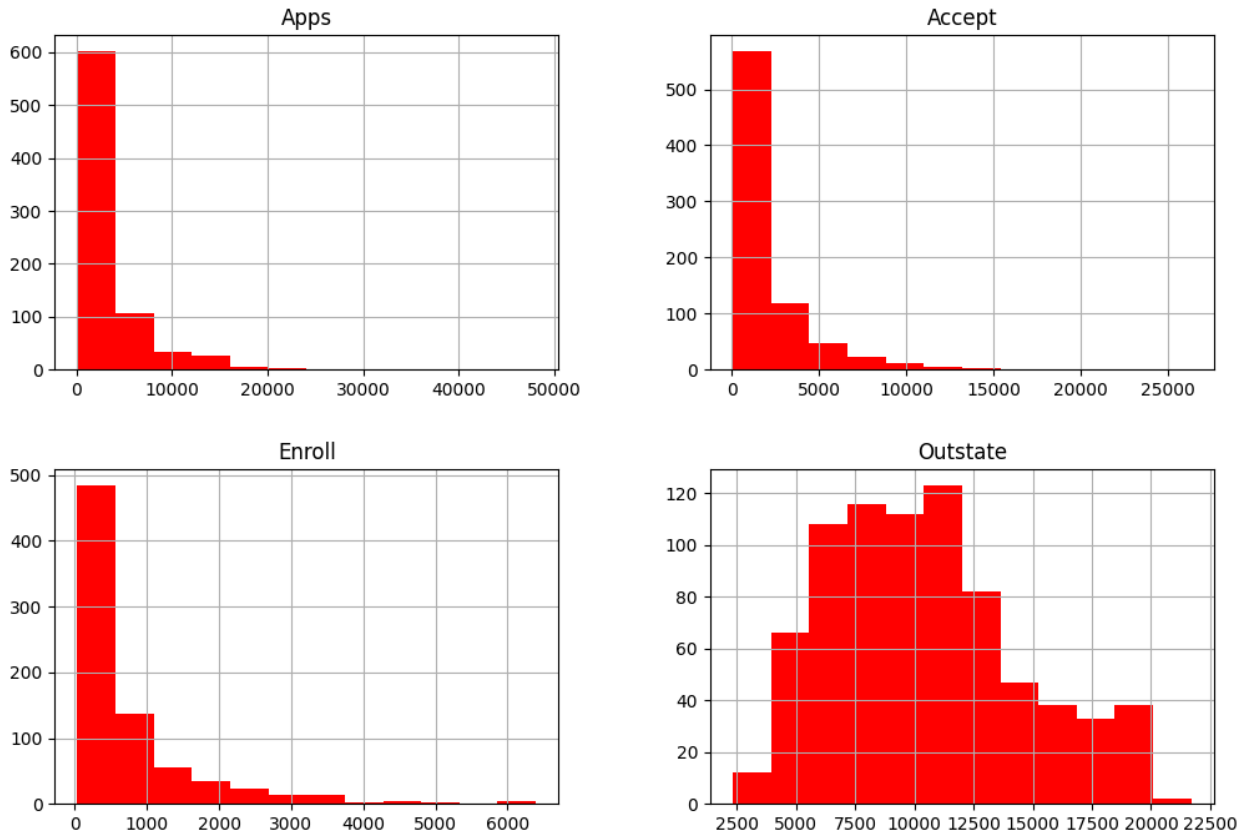
Out[67]:
```
Text(0, 0.5, 'Outstate')
```




**(g) Use the plot.hist() method of college to produce some histograms with differing numbers of bins for a few of the quantitative variables. The command plt.subplots(2, 2) may be useful: it will divide the plot window into four regions so that four plots can be made simultaneously. By changing the arguments you can divide the screen up in other combinations.**

In [70]:
```
quantitative = ['Apps', 'Accept', 'Enroll', 'Outstate']
fig, ax = plt.subplots(2,2, figsize = (12,8))
college[quantitative].hist(ax = ax, bins = 12, color = "red")
```

Out[70]:
```
array([[<Axes: title={'center': 'Apps'}>,
        <Axes: title={'center': 'Accept'}>],
       [<Axes: title={'center': 'Enroll'}>,
        <Axes: title={'center': 'Outstate'}>]], dtype=object)
```




**(h) Continue exploring the data, and provide a brief summary of what you discover.**

- Private and elite colleges have higher out of state tuition.
- Most colleges get fewer applications but a few get more.