```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

```
dataset = pd.read_csv('vehicle.csv')
```

```
dataset.head()
```

	compactness	circularity	distance_circularity	radius_ratio	pr.axis_aspect_ratio
0	95	48.0	83.0	178.0	72.0
1	91	41.0	84.0	141.0	57.0
2	104	50.0	106.0	209.0	66.0
3	93	41.0	82.0	159.0	63.0
4	85	44.0	70.0	205.0	103.0

```
dataset.shape
```

```
(846, 19)
```

```
dataset.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	
compactness	846.0	93.678487	8.234474	73.0	87.00	93.0	100.0	
circularity	841.0	44.828775	6.152172	33.0	40.00	44.0	49.0	
distance_circularity	842.0	82.110451	15.778292	40.0	70.00	80.0	98.0	
radius_ratio	840.0	168.888095	33.520198	104.0	141.00	167.0	195.0	
pr.axis_aspect_ratio	844.0	61.678910	7.891463	47.0	57.00	61.0	65.0	
max.length_aspect_ratio	846.0	8.567376	4.601217	2.0	7.00	8.0	10.0	
scatter_ratio	845.0	168.901775	33.214848	112.0	147.00	157.0	198.0	
elongatedness	845.0	40.933728	7.816186	26.0	33.00	43.0	46.0	
pr.axis_rectangularity	843.0	20.582444	2.592933	17.0	19.00	20.0	23.0	
max.length_rectangularity	846.0	147.998818	14.515652	118.0	137.00	146.0	159.0	
scaled_variance	843.0	188.631079	31.411004	130.0	167.00	179.0	217.0	
scaled_variance.1	844.0	439.494076	176.666903	184.0	318.00	363.5	587.0	1
scaled_radius_of_gyration	844.0	174.709716	32.584808	109.0	149.00	173.5	198.0	
scaled_radius_of_gyration.1	842.0	72.447743	7.486190	59.0	67.00	71.5	75.0	
skewness_about	840.0	6.364286	4.920649	0.0	2.00	6.0	9.0	
skewness_about.1	845.0	12.602367	8.936081	0.0	5.00	11.0	19.0	

```
dataset.dtypes
```

```
compactness                     int64
circularity                   float64
distance_circularity          float64
radius_ratio                  float64
pr.axis_aspect_ratio          float64
max.length_aspect_ratio         int64
scatter_ratio                 float64
elongatedness                 float64
pr.axis_rectangularity        float64
max.length_rectangularity       int64
scaled_variance               float64
scaled_variance.1             float64
scaled_radius_of_gyration     float64
scaled_radius_of_gyration.1   float64
skewness_about                float64
skewness_about.1              float64
skewness_about.2              float64
hollows_ratio                   int64
class                          object
dtype: object
```

```
dataset['class'].value_counts()
```

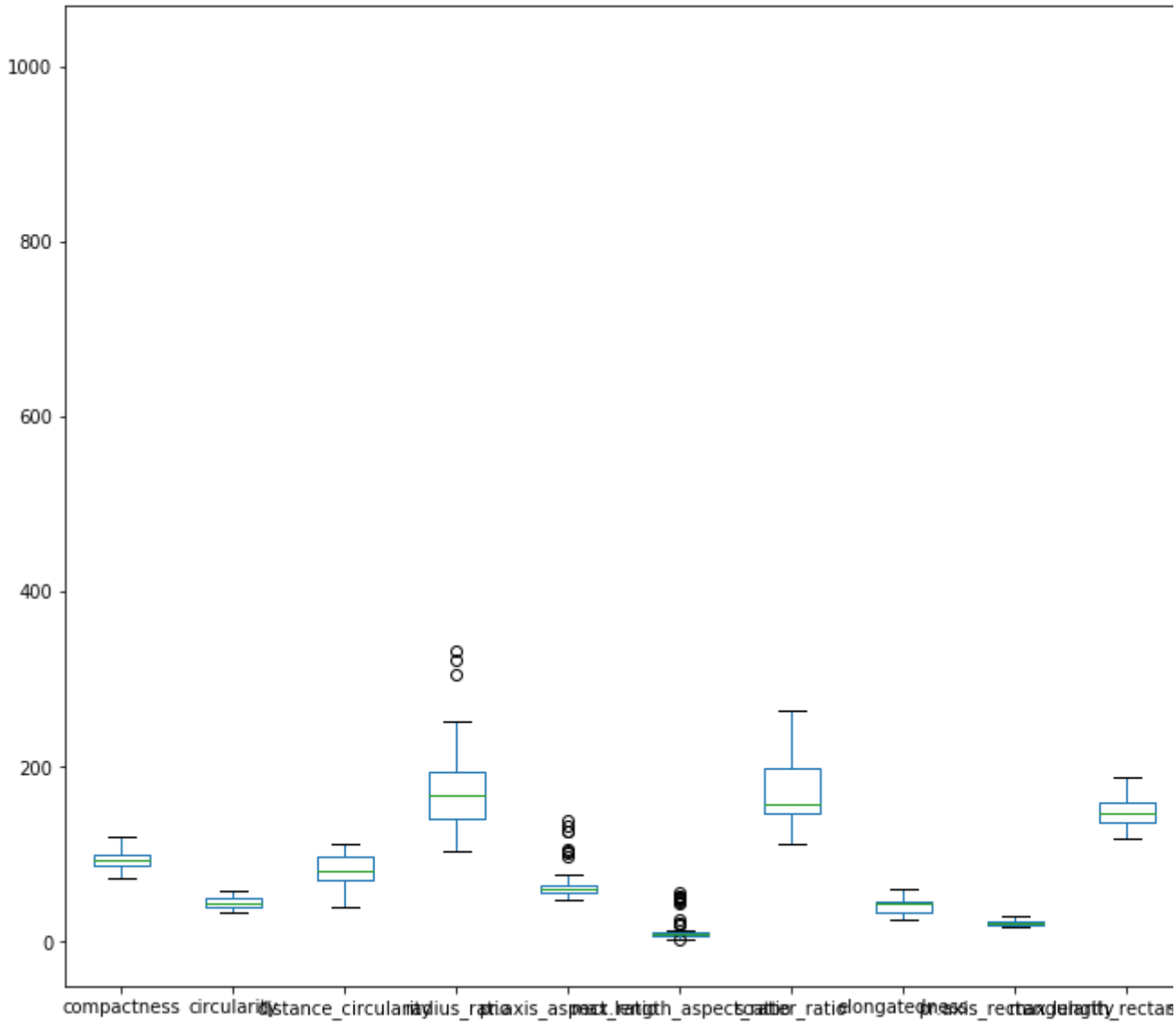
```
car    429
bus    218
```

```
dataset.groupby('class').size()
```

```
class
bus    218
car    429
van    199
dtype: int64
```

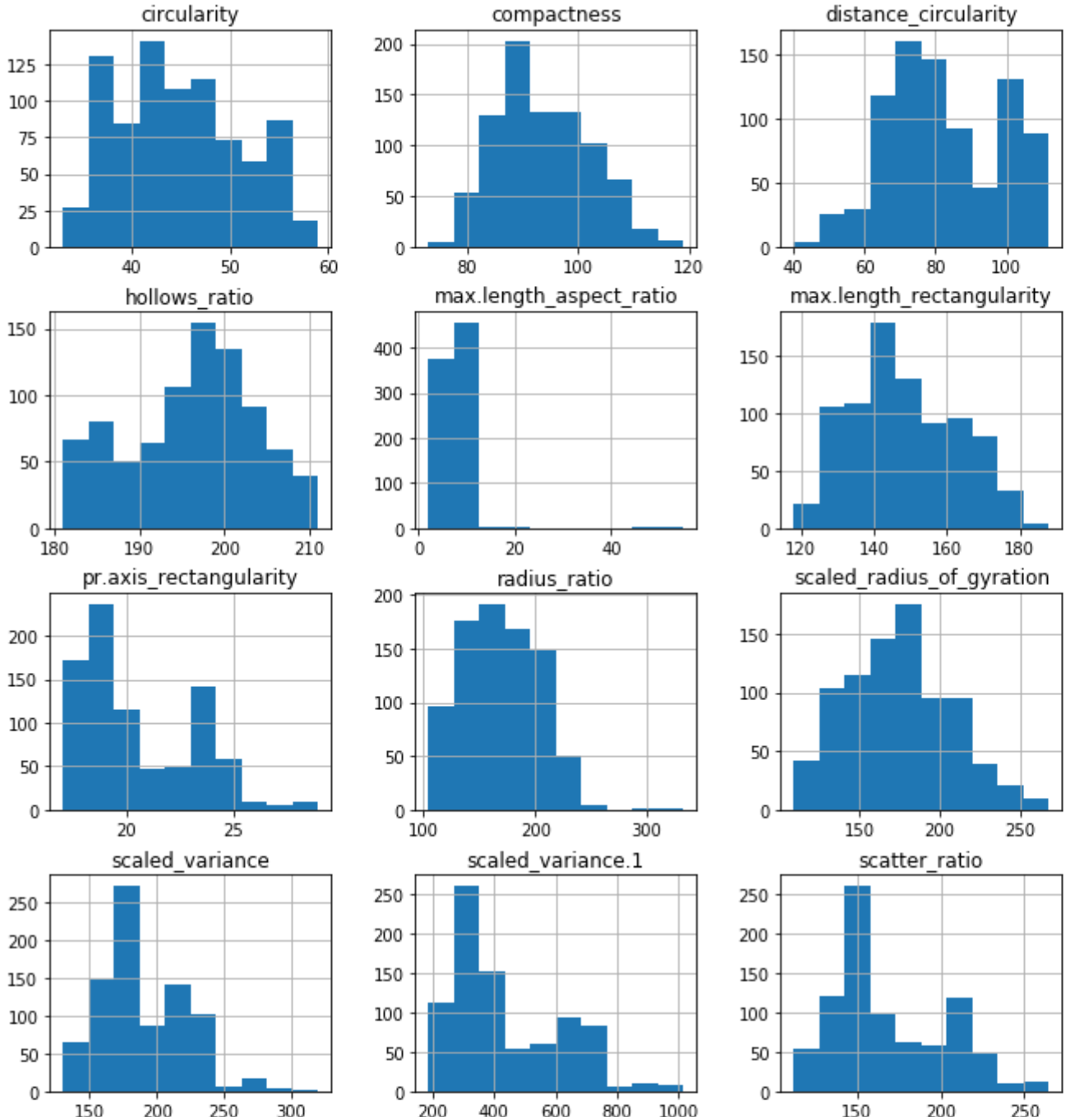
```
dataset.plot(kind='box', figsize=(20,10))
```

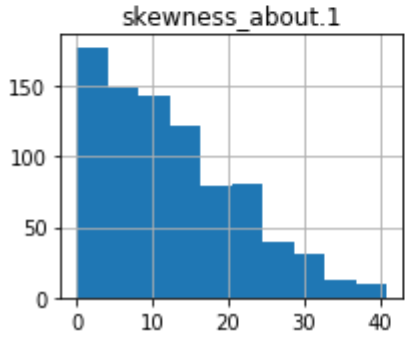
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7868b09310>
```

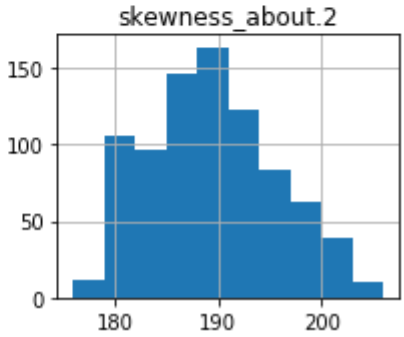



```
dataset.hist(figsize=(15,15))
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f7820f4bb50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820e665d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820ee6350>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820df8890>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7820f0cf10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820ec3490>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820d83e10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820cd0850>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7820cf9750>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820cc8d90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820c17d50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820c427d0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f78208ec050>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820bfbf10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820864990>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820832350>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f78207f0d50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f78207bf810>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7868cc6050>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7820717a90>]],
      dtype=object)
```








```
dataset.isnull().sum()
```

```
compactness                    0
circularity                    5
distance_circularity           4
radius_ratio                   6
pr.axis_aspect_ratio           2
max.length_aspect_ratio        0
scatter_ratio                  1
elongatedness                  1
pr.axis_rectangularity         3
max.length_rectangularity      0
scaled_variance                3
scaled_variance.1              2
scaled_radius_of_gyration      2
scaled_radius_of_gyration.1    4
skewness_about                 6
skewness_about.1               1
skewness_about.2               1
hollows_ratio                  0
class                          0
dtype: int64
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 846 entries, 0 to 845
```

```
for i in dataset.columns[:-1]:
    median_value = dataset[i].median()
    dataset[i] = dataset[i].fillna(median_value)
```

```
distance_circularity           842 non-null float64
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 846 entries, 0 to 845
Data columns (total 19 columns):
compactness                    846 non-null int64
circularity                    846 non-null float64
distance_circularity           846 non-null float64
radius_ratio                   846 non-null float64
pr.axis_aspect_ratio           846 non-null float64
max.length_aspect_ratio        846 non-null int64
scatter_ratio                  846 non-null float64
elongatedness                  846 non-null float64
pr.axis_rectangularity         846 non-null float64
max.length_rectangularity      846 non-null int64
scaled_variance                846 non-null float64
scaled_variance.1              846 non-null float64
scaled_radius_of_gyration      846 non-null float64
scaled_radius_of_gyration.1    846 non-null float64
skewness_about                 846 non-null float64
skewness_about.1               846 non-null float64
skewness_about.2               846 non-null float64
hollows_ratio                  846 non-null int64
class                          846 non-null object
dtypes: float64(14), int64(4), object(1)
memory usage: 125.6+ KB
```

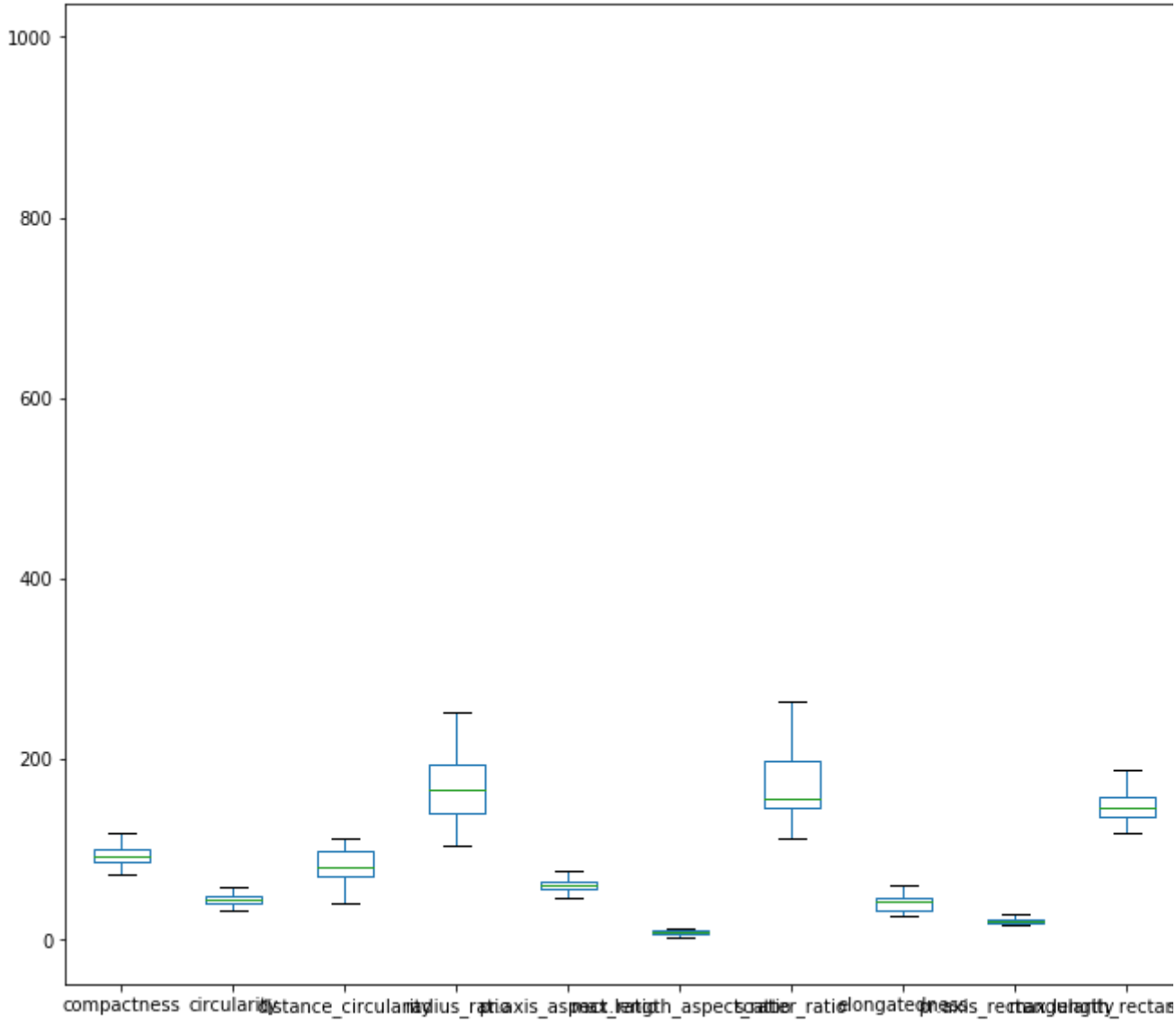
```
for col_name in dataset.columns[:-1]:
    q1 = dataset[col_name].quantile(0.25)
    q3 = dataset[col_name].quantile(0.75)
    iqr = q3 - q1

    low = q1-1.5*iqr
    high = q3+1.5*iqr

    dataset.loc[ (dataset[col_name] < low) | (dataset[col_name] > high), col_name] = dataset[col_
```

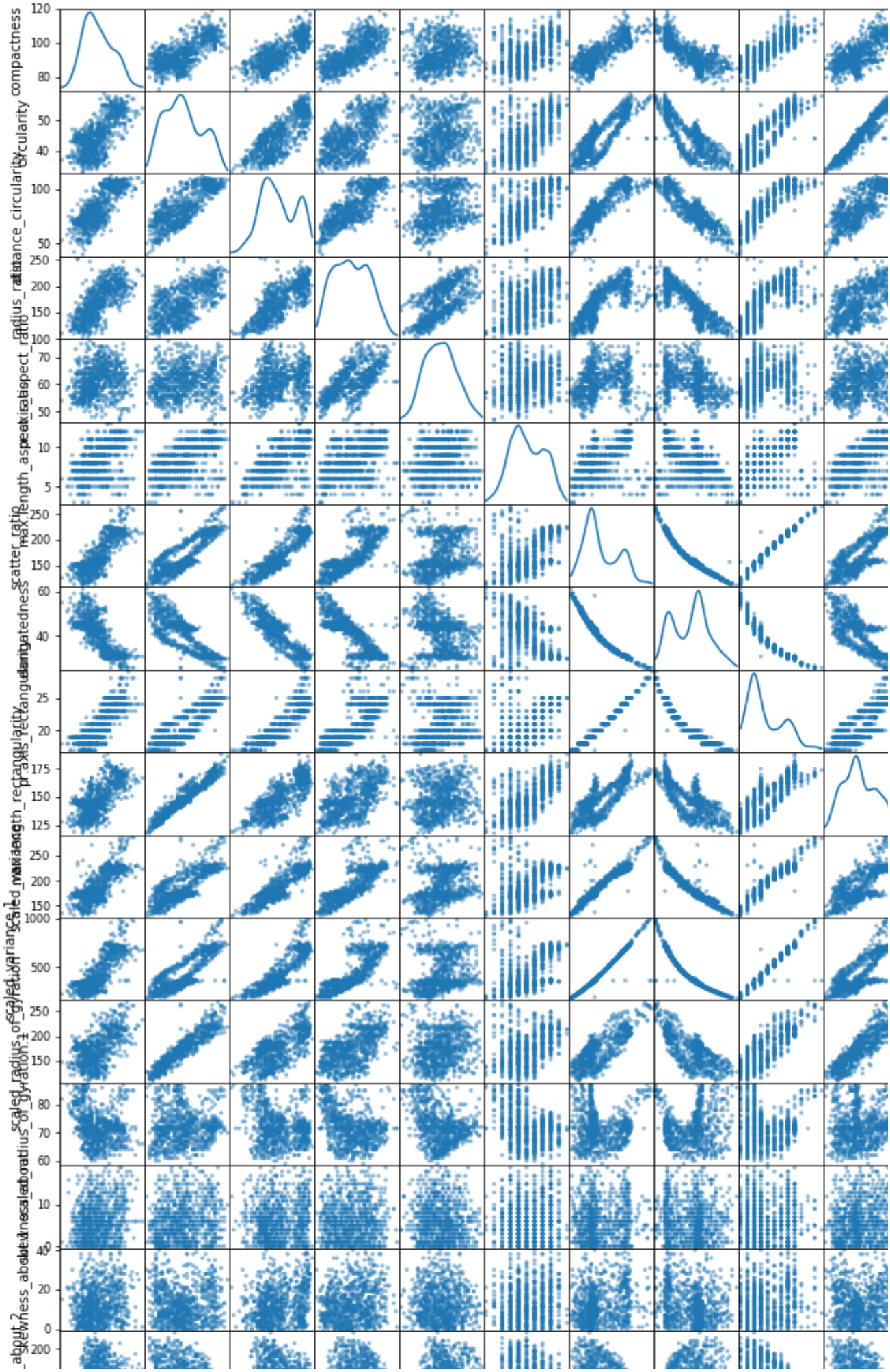
```
dataset.plot(kind='box', figsize=(20,10))
```

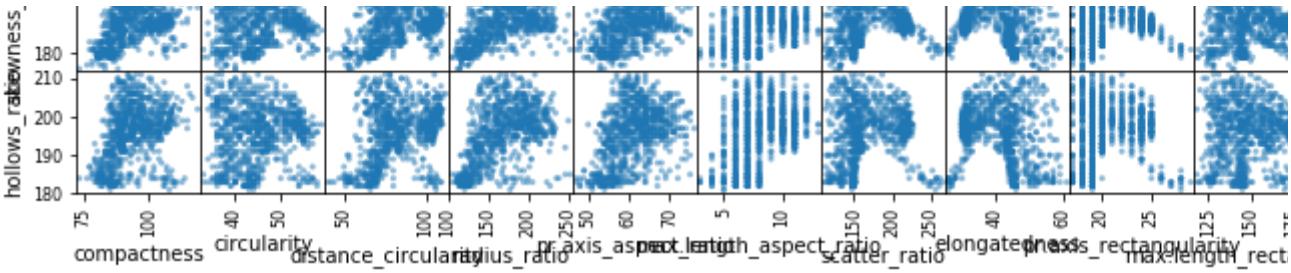
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f781e9dc990>
```




```
spd = pd.scatter_matrix(dataset, figsize=(20,20), diagonal='kde')
```

```
/usr/local/anaconda/python2/lib/python2.7/site-packages/ipykernel_launcher.py:1: Futu
  """Entry point for launching an IPython kernel.
```



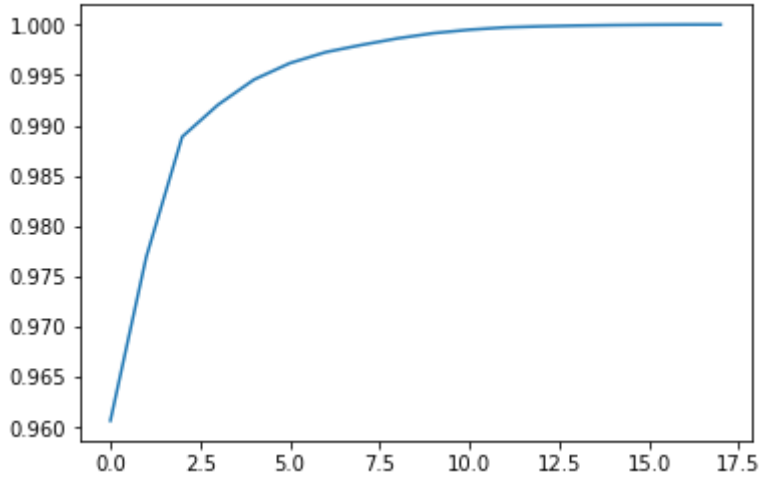



```
dataset.corr()
```

	compactness	circularity	distance_circularity	radius_rat
compactness	1.000000	0.684887	0.789928	0.7219
circularity	0.684887	1.000000	0.792320	0.6382
distance_circularity	0.789928	0.792320	1.000000	0.7942
radius_ratio	0.721925	0.638280	0.794222	1.0000
pr.axis_aspect_ratio	0.192864	0.203253	0.244332	0.6505
max.length_aspect_ratio	0.499928	0.560470	0.666809	0.4639
scatter_ratio	0.812620	0.847938	0.905076	0.7699
elongatedness	-0.788750	-0.821472	-0.911307	-0.8253
pr.axis_rectangularity	0.813694	0.843400	0.893025	0.7441
max.length_rectangularity	0.676143	0.961318	0.774527	0.5794
scaled_variance	0.769871	0.802768	0.869584	0.7861
scaled_variance.1	0.806170	0.827462	0.883943	0.7602
scaled_radius_of_gyration	0.585243	0.925816	0.705771	0.5507
scaled_radius_of_gyration.1	-0.246681	0.068745	-0.229353	-0.3904
skewness_about	0.197308	0.136351	0.099107	0.0357
skewness_about.1	0.156348	-0.009666	0.262345	0.1796
skewness_about.2	0.298537	-0.104426	0.146098	0.4058
hollows_ratio	0.365552	0.046351	0.332732	0.4917

```
X = dataset.iloc[:,0:18]
from sklearn.decomposition import PCA
pca = PCA().fit(X)
plt.plot(np.cumsum(pca.explained_variance_ratio_))
```

```
[<matplotlib.lines.Line2D at 0x7f7808ea9cd0>]
```




```
pca = PCA(n_components=10)
pca.fit(X)

X = pca.transform(X),
```

```
from sklearn import cross_validation

Y = dataset.iloc[:,18]
test_size=0.2
X_train, X_test, Y_train, Y_test = cross_validation.train_test_split(X, Y, test_size=0.2, random_
```

```
from sklearn.naive_bayes import GaussianNB

from sklearn.svm import SVC
```

```
from sklearn.model_selection import GridSearchCV

model = SVC()

params = {'C': [0.01, 0.1, 0.5, 1], 'kernel': ['linear', 'rbf']}

model1 = GridSearchCV(model, param_grid=params, verbose=5)

model1.fit(X_train, Y_train)

print("Best Hyper Parameters:\n", model1.best_params_)
```

```
Fitting 3 folds for each of 8 candidates, totalling 24 fits
[CV] kernel=linear, C=0.01 ........................................
[CV] ...... kernel=linear, C=0.01, score=0.845132743363, total=   0.0s
[CV] kernel=linear, C=0.01 ........................................
[CV] ...... kernel=linear, C=0.01, score=0.835555555556, total=   0.0s
[CV] kernel=linear, C=0.01 ........................................
[CV] ...... kernel=linear, C=0.01, score=0.897777777778, total=   0.1s
[CV] kernel=rbf, C=0.01 ...........................................
[CV] ......... kernel=rbf, C=0.01, score=0.504424778761, total=   0.0s
[CV] kernel=rbf, C=0.01 ...........................................
[CV] ......... kernel=rbf, C=0.01, score=0.502222222222, total=   0.0s
[CV] kernel=rbf, C=0.01 ...........................................
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:    0.1s remaining:    0.0s
[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:    0.1s remaining:    0.0s
[Parallel(n_jobs=1)]: Done   4 out of   4 | elapsed:    0.2s remaining:    0.0s
[CV] ......... kernel=rbf, C=0.01, score=0.502222222222, total=   0.0s
[CV] kernel=linear, C=0.1 .........................................
[CV] ....... kernel=linear, C=0.1, score=0.849557522124, total=   0.1s
[CV] kernel=linear, C=0.1 .........................................
[CV] ....... kernel=linear, C=0.1, score=0.831111111111, total=   0.2s
[CV] kernel=linear, C=0.1 .........................................
[CV] ....... kernel=linear, C=0.1, score=0.893333333333, total=   0.2s
[CV] kernel=rbf, C=0.1 ............................................
[CV] .......... kernel=rbf, C=0.1, score=0.504424778761, total=   0.0s
[CV] kernel=rbf, C=0.1 ............................................
[CV] .......... kernel=rbf, C=0.1, score=0.502222222222, total=   0.0s
[CV] kernel=rbf, C=0.1 ............................................
[CV] .......... kernel=rbf, C=0.1, score=0.502222222222, total=   0.0s
[CV] kernel=linear, C=0.5 .........................................
[CV] ....... kernel=linear, C=0.5, score=0.836283185841, total=   0.7s
[CV] kernel=linear, C=0.5 .........................................
[CV] ....... kernel=linear, C=0.5, score=0.826666666667, total=   1.2s
[CV] kernel=linear, C=0.5 .........................................
[CV] ....... kernel=linear, C=0.5, score=0.888888888889, total=   1.0s
[CV] kernel=rbf, C=0.5 ............................................
[CV] .......... kernel=rbf, C=0.5, score=0.504424778761, total=   0.0s
[CV] kernel=rbf, C=0.5 ............................................
[CV] .......... kernel=rbf, C=0.5, score=0.502222222222, total=   0.0s
[CV] kernel=rbf, C=0.5 ............................................
[CV] .......... kernel=rbf, C=0.5, score=0.502222222222, total=   0.0s
[CV] kernel=linear, C=1 ...........................................
[CV] ......... kernel=linear, C=1, score=0.849557522124, total=   1.8s
[CV] kernel=linear, C=1 ...........................................
[CV] ......... kernel=linear, C=1, score=0.826666666667, total=   2.4s
[CV] kernel=linear, C=1 ...........................................
[CV] ......... kernel=linear, C=1, score=0.888888888889, total=   1.8s
[CV] kernel=rbf, C=1 ..............................................
[CV] ............ kernel=rbf, C=1, score=0.508849557522, total=   0.0s
[CV] kernel=rbf, C=1 ..............................................
[CV] ............ kernel=rbf, C=1, score=0.502222222222, total=   0.0s
[CV] kernel=rbf, C=1 ..............................................
[CV] ............ kernel=rbf, C=1, score=0.506666666667, total=   0.0s
('Best Hyper Parameters:\n', {'kernel': 'linear', 'C': 0.01})
[Parallel(n_jobs=1)]: Done  24 out of  24 | elapsed:    9.8s finished
```

```
from sklearn.cross_validation import cross_val_score

model = SVC(C=0.01, kernel="linear")

scores = cross_val_score(model, X, Y, cv=10)

print(scores)
```

```
[0.82352941 0.84705882 0.84705882 0.84705882 0.82352941 0.83529412
 0.89411765 0.82352941 0.8452381  0.86585366]
```

```
model = GaussianNB()

scores = cross_val_score(model, X, Y, cv=10)

print(scores)
```

```
[0.74117647 0.8        0.76470588 0.77647059 0.8        0.74117647
 0.81176471 0.67058824 0.80952381 0.81707317]
```