lab1

September 9, 2020

[1]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from sklearn.datasets import load_breast_cancer
from sklearn.metrics import confusion_matrix
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report

import seaborn as sns
sns.set()
breast_cancer = load_breast_cancer()

X = pd.DataFrame(breast_cancer.data, columns=breast_cancer.feature_names)
```

[2]:
```
print (X)
```

```
     mean radius  mean texture  mean perimeter  mean area  mean smoothness  \
0          17.99         10.38          122.80     1001.0          0.11840
1          20.57         17.77          132.90     1326.0          0.08474
2          19.69         21.25          130.00     1203.0          0.10960
3          11.42         20.38           77.58      386.1          0.14250
4          20.29         14.34          135.10     1297.0          0.10030
..           …             …               …          …               …
564        21.56         22.39          142.00     1479.0          0.11100
565        20.13         28.25          131.20     1261.0          0.09780
566        16.60         28.08          108.30      858.1          0.08455
567        20.60         29.33          140.10     1265.0          0.11780
568         7.76         24.54           47.92      181.0          0.05263

     mean compactness  mean concavity  mean concave points  mean symmetry  \
0             0.27760         0.30010              0.14710         0.2419
1             0.07864         0.08690              0.07017         0.1812
2             0.15990         0.19740              0.12790         0.2069
3             0.28390         0.24140              0.10520         0.2597
4             0.13280         0.19800              0.10430         0.1809
..                …               …                    …              …
```

```
564          0.11590         0.24390              0.13890         0.1726
565          0.10340         0.14400              0.09791         0.1752
566          0.10230         0.09251              0.05302         0.1590
567          0.27700         0.35140              0.15200         0.2397
568          0.04362         0.00000              0.00000         0.1587

     mean fractal dimension  …  worst radius  worst texture  \
0                   0.07871  …        25.380          17.33
1                   0.05667  …        24.990          23.41
2                   0.05999  …        23.570          25.53
3                   0.09744  …        14.910          26.50
4                   0.05883  …        22.540          16.67
..                      …  …           …             …
564                 0.05623  …        25.450          26.40
565                 0.05533  …        23.690          38.25
566                 0.05648  …        18.980          34.12
567                 0.07016  …        25.740          39.42
568                 0.05884  …         9.456          30.37

     worst perimeter  worst area  worst smoothness  worst compactness  \
0             184.60      2019.0           0.16220            0.66560
1             158.80      1956.0           0.12380            0.18660
2             152.50      1709.0           0.14440            0.42450
3              98.87       567.7           0.20980            0.86630
4             152.20      1575.0           0.13740            0.20500
..               …           …                …                 …
564           166.10      2027.0           0.14100            0.21130
565           155.00      1731.0           0.11660            0.19220
566           126.70      1124.0           0.11390            0.30940
567           184.60      1821.0           0.16500            0.86810
568            59.16       268.6           0.08996            0.06444

     worst concavity  worst concave points  worst symmetry  \
0             0.7119                0.2654          0.4601
1             0.2416                0.1860          0.2750
2             0.4504                0.2430          0.3613
3             0.6869                0.2575          0.6638
4             0.4000                0.1625          0.2364
..               …                   …               …
564           0.4107                0.2216          0.2060
565           0.3215                0.1628          0.2572
566           0.3403                0.1418          0.2218
567           0.9387                0.2650          0.4087
568           0.0000                0.0000          0.2871

     worst fractal dimension
0                    0.11890
1                    0.08902
```

```
2                    0.08758
3                    0.17300
4                    0.07678
..                       …
564                  0.07115
565                  0.06637
566                  0.07820
567                  0.12400
568                  0.07039

[569 rows x 30 columns]
```

[3]:
```
dir(breast_cancer)
```

[3]:
```
['DESCR',
 'data',
 'feature_names',
 'filename',
 'frame',
 'target',
 'target_names']
```

[4]:
```
y=pd.Categorical.from_codes(breast_cancer.target,breast_cancer.target_names)
print(y)
```

```
[malignant, malignant, malignant, malignant, malignant, …, malignant,
malignant, malignant, malignant, benign]
Length: 569
Categories (2, object): [malignant, benign]
```

[5]:
```
X.describe()
```

[5]:

	mean radius	mean texture	mean perimeter	mean area	\
count	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	
std	3.524049	4.301036	24.298981	351.914129	
min	6.981000	9.710000	43.790000	143.500000	
25%	11.700000	16.170000	75.170000	420.300000	
50%	13.370000	18.840000	86.240000	551.100000	
75%	15.780000	21.800000	104.100000	782.700000	
max	28.110000	39.280000	188.500000	2501.000000	

	mean smoothness	mean compactness	mean concavity	mean concave points	\
count	569.000000	569.000000	569.000000	569.000000	
mean	0.096360	0.104341	0.088799	0.048919	
std	0.014064	0.052813	0.079720	0.038803	
min	0.052630	0.019380	0.000000	0.000000	
25%	0.086370	0.064920	0.029560	0.020310	

```
50%          0.095870          0.092630        0.061540             0.033500
75%          0.105300          0.130400        0.130700             0.074000
max          0.163400          0.345400        0.426800             0.201200

       mean symmetry  mean fractal dimension  …  worst radius  \
count     569.000000              569.000000  …    569.000000
mean        0.181162                0.062798  …     16.269190
std         0.027414                0.007060  …      4.833242
min         0.106000                0.049960  …      7.930000
25%         0.161900                0.057700  …     13.010000
50%         0.179200                0.061540  …     14.970000
75%         0.195700                0.066120  …     18.790000
max         0.304000                0.097440  …     36.040000

       worst texture  worst perimeter   worst area  worst smoothness  \
count     569.000000       569.000000   569.000000        569.000000
mean       25.677223       107.261213   880.583128          0.132369
std         6.146258        33.602542   569.356993          0.022832
min        12.020000        50.410000   185.200000          0.071170
25%        21.080000        84.110000   515.300000          0.116600
50%        25.410000        97.660000   686.500000          0.131300
75%        29.720000       125.400000  1084.000000          0.146000
max        49.540000       251.200000  4254.000000          0.222600

       worst compactness  worst concavity  worst concave points  \
count         569.000000       569.000000            569.000000
mean            0.254265         0.272188              0.114606
std             0.157336         0.208624              0.065732
min             0.027290         0.000000              0.000000
25%             0.147200         0.114500              0.064930
50%             0.211900         0.226700              0.099930
75%             0.339100         0.382900              0.161400
max             1.058000         1.252000              0.291000

       worst symmetry  worst fractal dimension
count      569.000000               569.000000
mean         0.290076                 0.083946
std          0.061867                 0.018061
min          0.156500                 0.055040
25%          0.250400                 0.071460
50%          0.282200                 0.080040
75%          0.317900                 0.092080
max          0.663800                 0.207500

[8 rows x 30 columns]
```

[6]:
```
#We will do this using SciKit-Learn library in Python using the␣
 ↪train_test_split method.
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size = 0.25,␣
 ↪random_state = 0)
#Feature Scaling to bring attribute to one range (say 0-100 or 0-1)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

[7]:
```
print(X_train)
print(X_test)
```

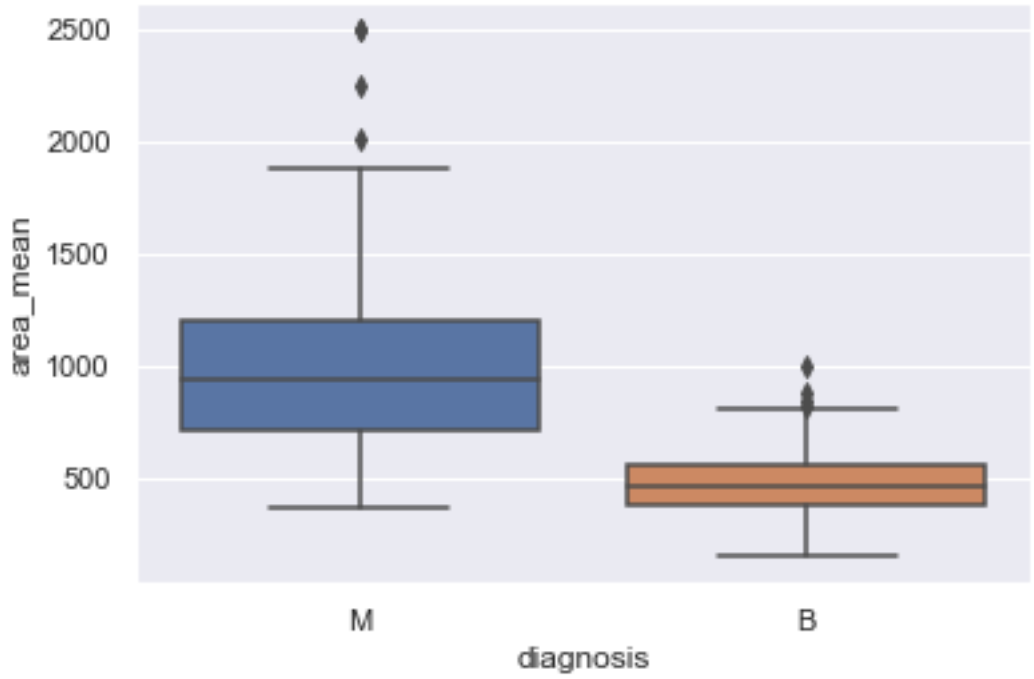
```
[[-0.65079907 -0.43057322 -0.68024847 … -0.36433881  0.32349851
  -0.7578486 ]
 [-0.82835341  0.15226547 -0.82773762 … -1.45036679  0.62563098
  -1.03071387]
 [ 1.68277234  2.18977235  1.60009756 …  0.72504581 -0.51329768
  -0.96601386]
 …
 [-1.33114223 -0.22172269 -1.3242844  … -0.98806491 -0.69995543
  -0.12266325]
 [-1.25110186 -0.24600763 -1.28700242 … -1.75887319 -1.56206114
  -1.00989735]
 [-0.74662205  1.14066273 -0.72203706 … -0.2860679  -1.24094654
   0.2126516 ]]
[[-0.21395901  0.3125461  -0.14355187 …  1.37043754  1.08911166
   1.53928319]
 [-0.26750714  1.461224   -0.32955207 … -0.84266106 -0.71577388
  -0.88105993]
 [-0.03922298 -0.86770223 -0.10463112 … -0.505318   -1.20298225
  -0.92494342]
 …
 [-0.51270124 -1.69096186 -0.54095317 … -0.12632201  0.33773512
  -0.42872244]
 [-0.17732081 -2.01395163 -0.17345939 … -0.62875108 -0.29500302
  -0.65432858]
 [ 1.5305829  -0.26300709  1.57961296 …  1.6694843   1.18085869
   0.48889253]]
```

[16]:
```
print("Cancer data set dimensions : {}".format(X.shape,y.shape))
```

```
Cancer data set dimensions : (569, 30)
```

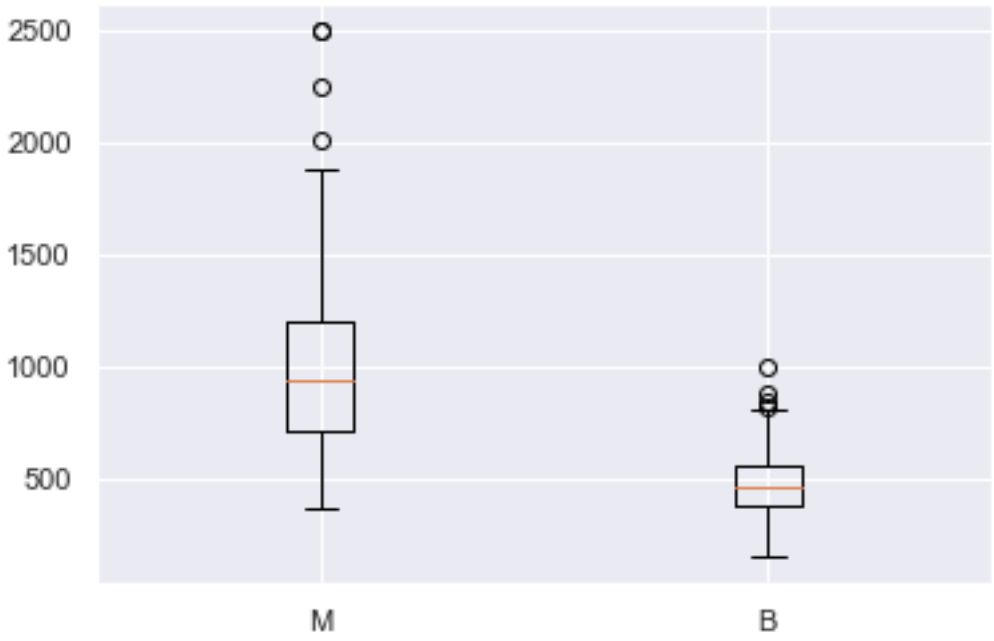
[17]:
```
df=pd.read_csv(r"D:\msc3\machine learning\lab1\data.csv")
sns.boxplot(x='diagnosis', y='area_mean', data=df)
```

[17]: <matplotlib.axes._subplots.AxesSubplot at 0x23766372388>

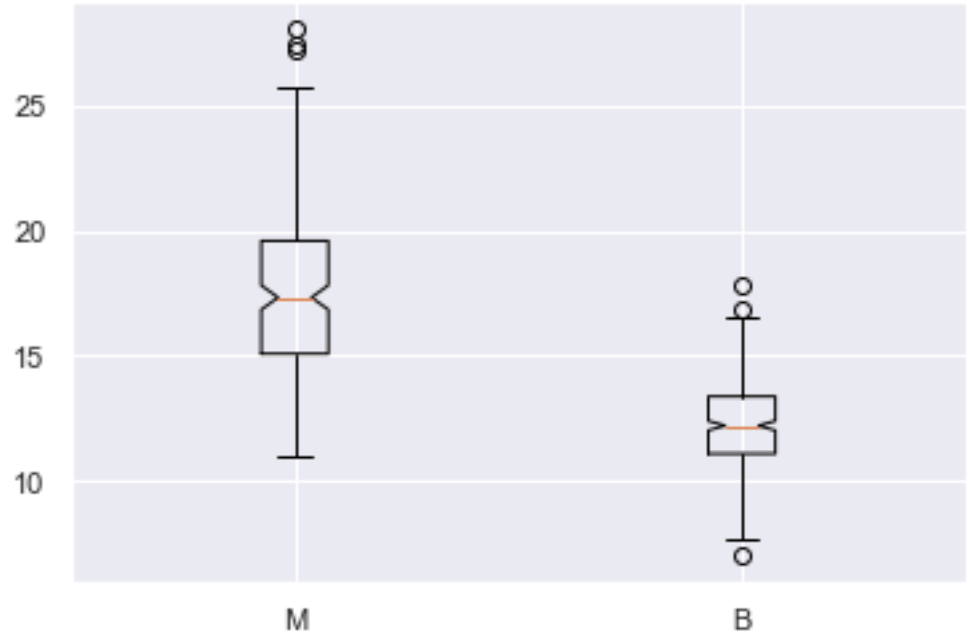


[18]:
```
malignant = df[df['diagnosis']=='M']['area_mean']
benign = df[df['diagnosis']=='B']['area_mean']
fig = plt.figure()
ax = fig.add_subplot(111)
ax.boxplot([malignant,benign], labels=['M', 'B'])
```

[18]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x23771870288>,
  <matplotlib.lines.Line2D at 0x23771884a08>,
  <matplotlib.lines.Line2D at 0x237718918c8>,
  <matplotlib.lines.Line2D at 0x2377188a848>],
 'caps': [<matplotlib.lines.Line2D at 0x23771884ec8>,
  <matplotlib.lines.Line2D at 0x23771884bc8>,
  <matplotlib.lines.Line2D at 0x23771891f48>,
  <matplotlib.lines.Line2D at 0x2377189a908>],
 'boxes': [<matplotlib.lines.Line2D at 0x23771879fc8>,
  <matplotlib.lines.Line2D at 0x2377188aec8>],
 'medians': [<matplotlib.lines.Line2D at 0x2377188a8c8>,
  <matplotlib.lines.Line2D at 0x2377189ae88>],
 'fliers': [<matplotlib.lines.Line2D at 0x2377188ad88>,
  <matplotlib.lines.Line2D at 0x2377189ab88>],
 'means': []}
```

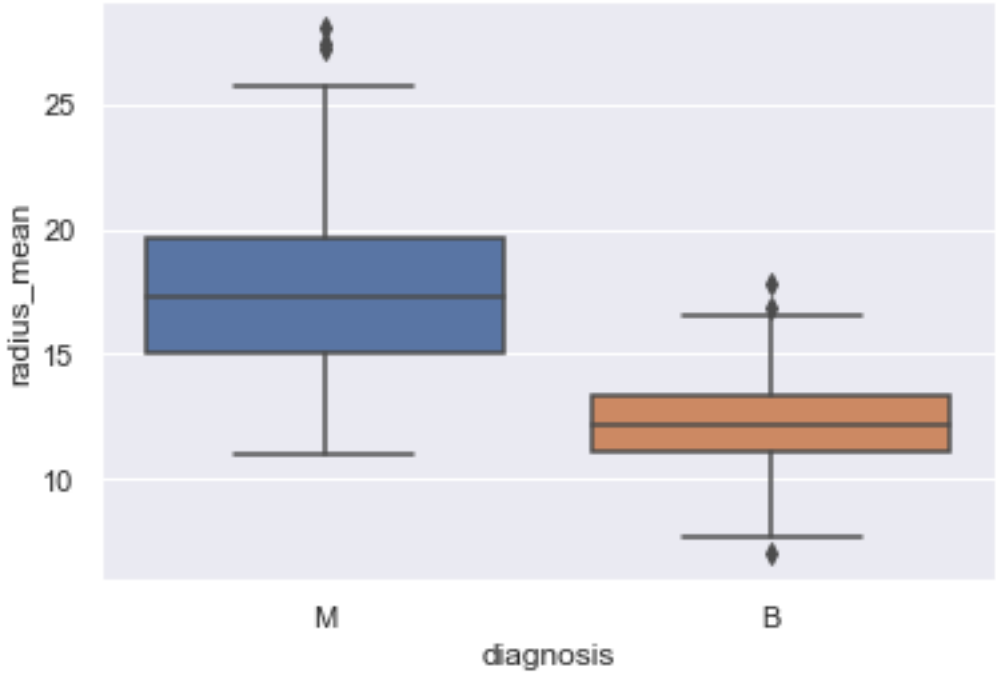


[19]:
```
malignant = df[df['diagnosis']=='M']['radius_mean']
benign = df[df['diagnosis']=='B']['radius_mean']
fig = plt.figure()
ax = fig.add_subplot(111)
ax.boxplot([malignant,benign], notch = True, labels=['M', 'B']);
```



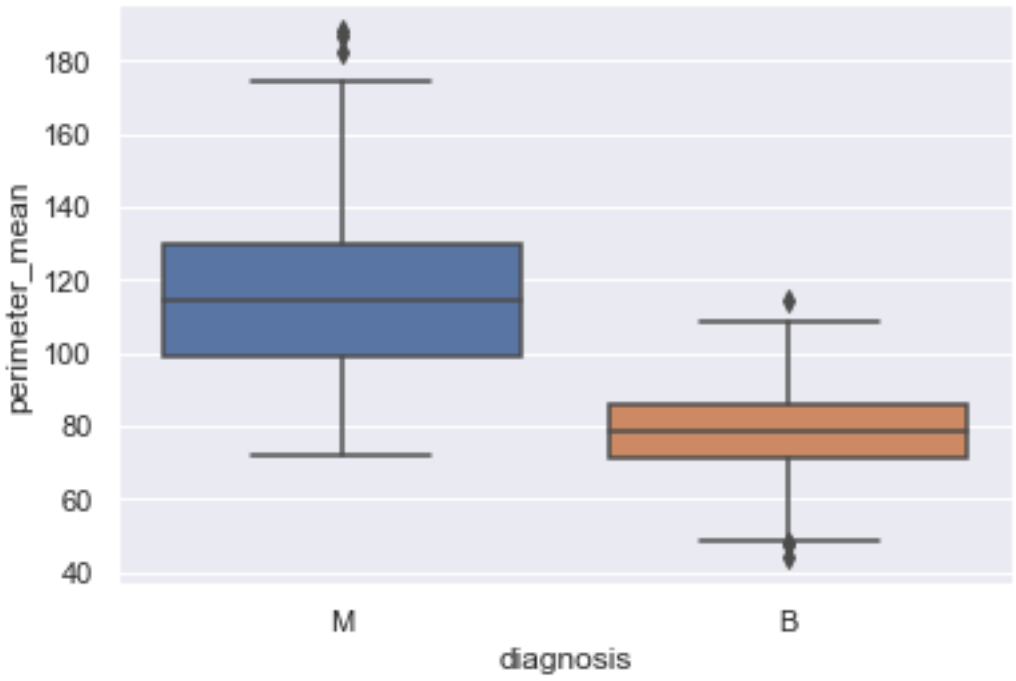
[20]:
```
sns.boxplot(x='diagnosis', y='radius_mean', data=df)
```

[20]: <matplotlib.axes._subplots.AxesSubplot at 0x23771926048>

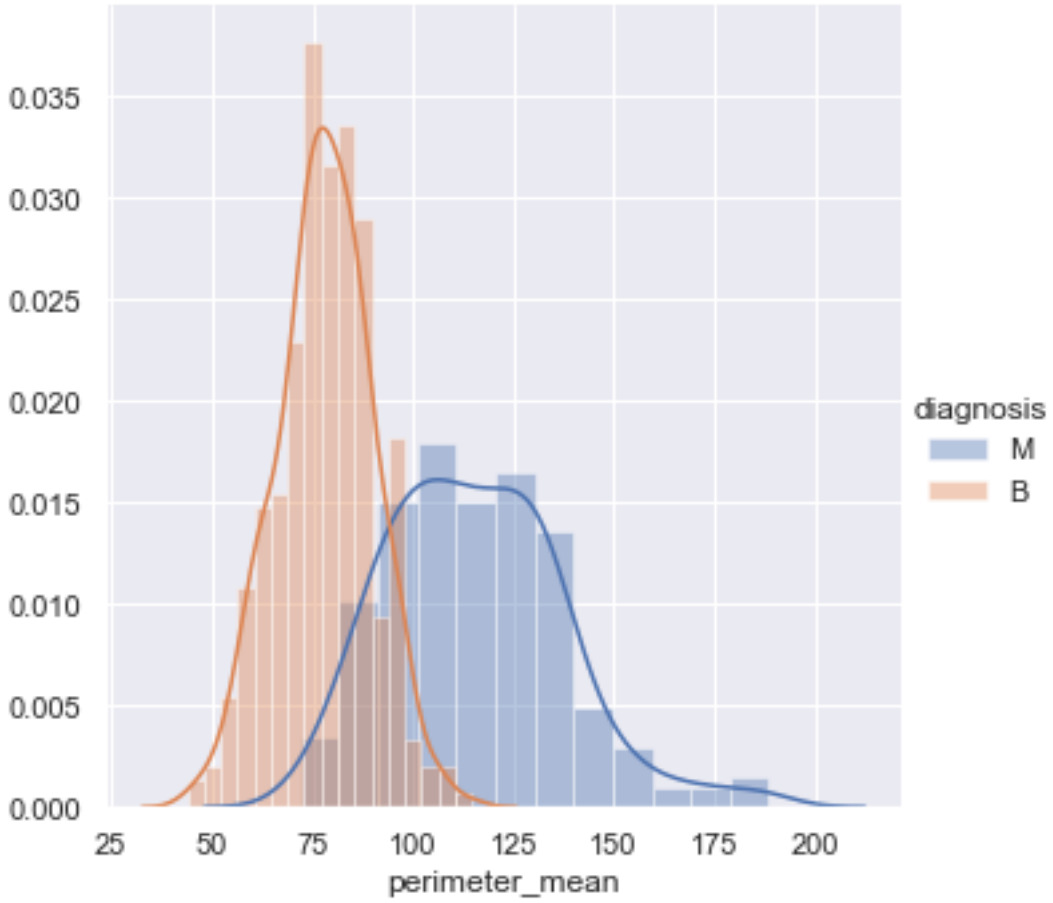


[21]:
```
sns.boxplot(x='diagnosis', y='perimeter_mean', data=df)
```

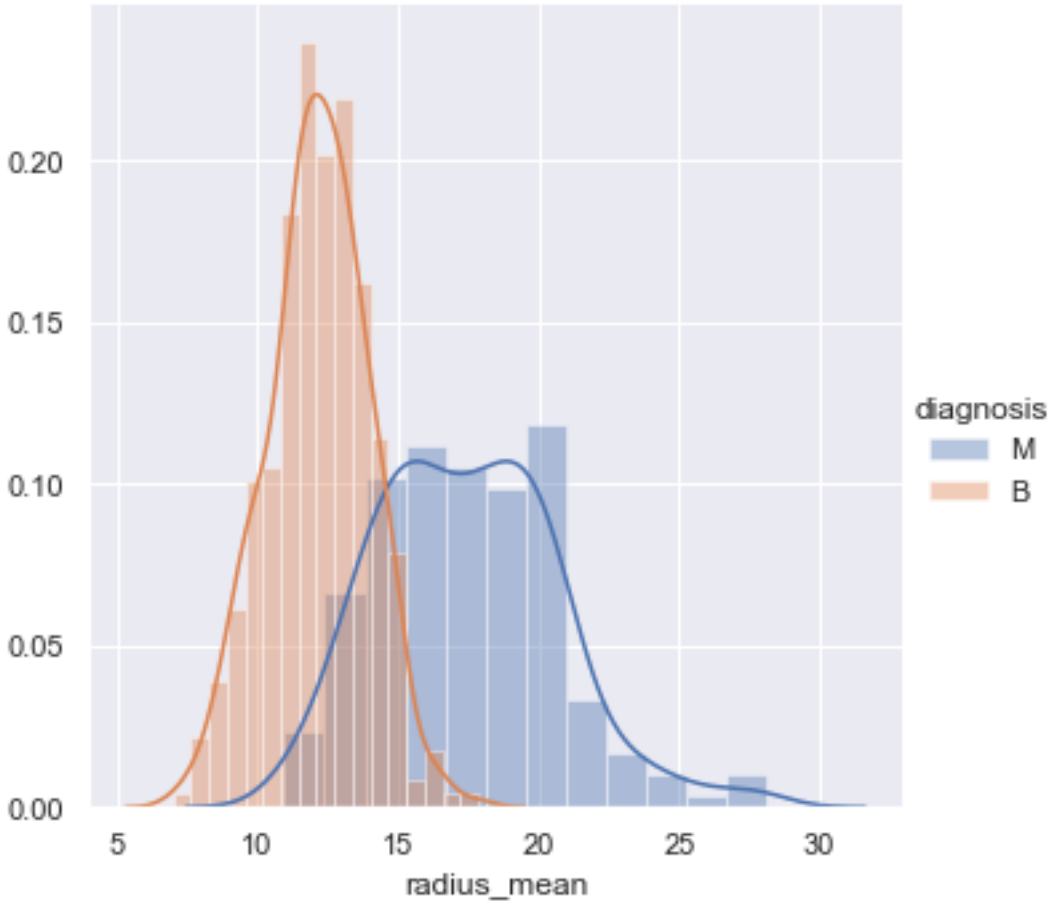
[21]: <matplotlib.axes._subplots.AxesSubplot at 0x23771997c88>



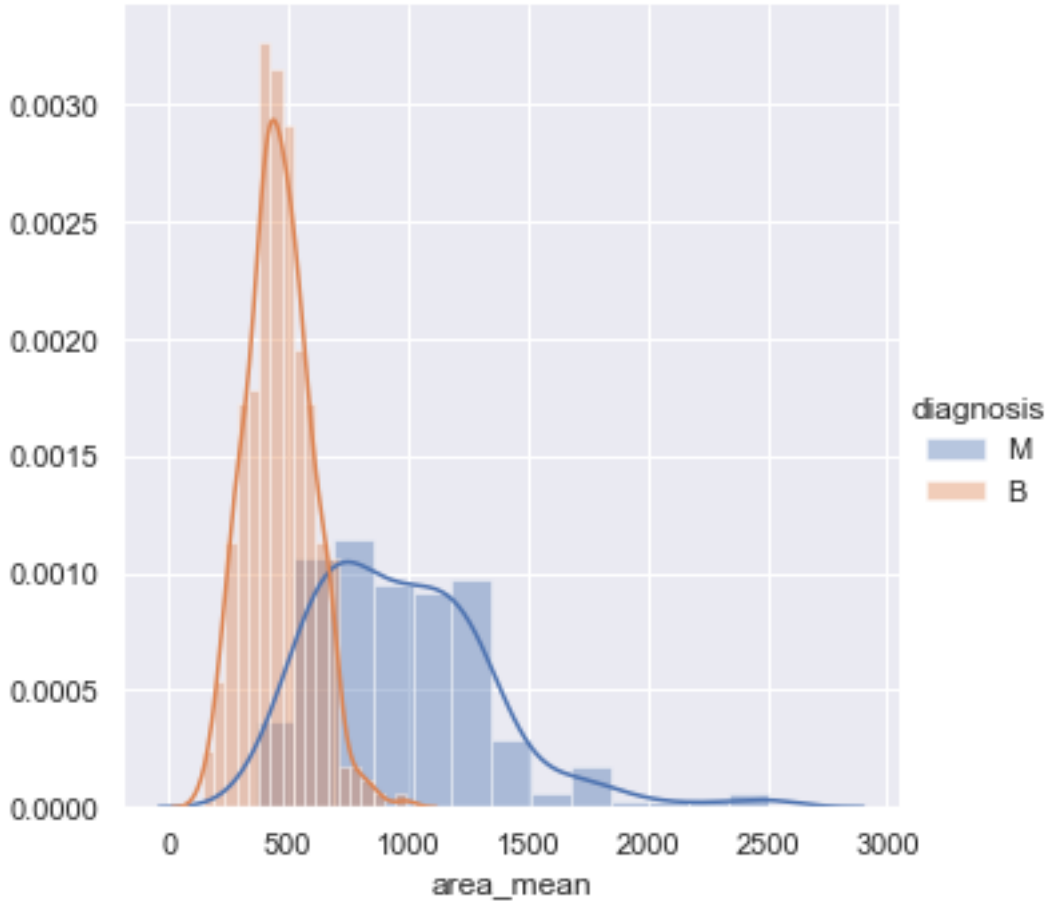
[22]:
```
sns.FacetGrid(df, hue='diagnosis', height=5) \
  .map(sns.distplot, 'perimeter_mean') \
  .add_legend()
plt.show()
```



[23]:
```
sns.FacetGrid(df, hue='diagnosis', height=5) \
  .map(sns.distplot, 'radius_mean') \
  .add_legend()
plt.show()
```



[24]:
```
sns.FacetGrid(df, hue='diagnosis', height=5) \
  .map(sns.distplot, 'area_mean') \
  .add_legend()
plt.show()
```



[25]:
```
df.isnull().sum()
df.isna().sum()
```

[25]:
```
id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
```

```
area_se                  0
smoothness_se            0
compactness_se           0
concavity_se             0
concave points_se        0
symmetry_se              0
fractal_dimension_se     0
radius_worst             0
texture_worst            0
perimeter_worst          0
area_worst               0
smoothness_worst         0
compactness_worst        0
concavity_worst          0
concave points_worst     0
symmetry_worst           0
fractal_dimension_worst  0
dtype: int64
```

[ ]:

[ ]: