ECE - 657A Homework 1

Tongdan Su 20754736

Q1:

```
Mean:
 feature_1    4.442167
feature_2    3.150805
feature_3    3.215227
feature_4    2.830161
feature_5    3.234261
feature_6    3.544656
feature_7    3.445095
feature_8    2.869693
feature_9    1.603221
dtype: float64
Mode:
    feature_1  feature_2  feature_3  feature_4  feature_5  feature_6  \
0          1          1          1          1          2          1

   feature_7  feature_8  feature_9
0          3          1          1
Skew:
 feature_1    0.587654
feature_2    1.226404
feature_3    1.157890
feature_4    1.509181
feature_5    1.703716
feature_6    0.990016
feature_7    1.095270
feature_8    1.420431
feature_9    3.511476
dtype: float64
 Standard Deviation:
  feature_1    2.820761
 feature_2    3.065145
 feature_3    2.988581
 feature_4    2.864562
 feature_5    2.223085
 feature_6    3.643857
 feature_7    2.449697
 feature_8    3.052666
 feature_9    1.732674
 dtype: float64
 Variance Values:
  feature_1     7.956694
 feature_2     9.395113
 feature_3     8.931615
 feature_4     8.205717
 feature_5     4.942109
 feature_6    13.277695
 feature_7     6.001013
 feature_8     9.318772
 feature_9     3.002160
 dtype: float64
```

Q2:

Computing results for all the PCC numbers of each pair of features are shown as follows:

	PCC	p-value
feature_1__feature_2	0.642481	8.964173e-81
feature_1__feature_3	0.653470	2.064616e-84
feature_1__feature_4	0.487829	4.027956e-42
feature_1__feature_5	0.523596	2.411759e-49
feature_1__feature_6	0.593091	4.050902e-66
feature_1__feature_7	0.553742	3.880813e-56
feature_1__feature_8	0.534066	1.260114e-51
feature_1__feature_9	0.350957	3.148289e-21
feature_2__feature_3	0.907228	2.567742e-258
feature_2__feature_4	0.706977	1.544338e-104
feature_2__feature_5	0.753544	3.535863e-126
feature_2__feature_6	0.691709	2.402961e-98
feature_2__feature_7	0.755559	3.184894e-127
feature_2__feature_8	0.719346	7.433029e-110
feature_2__feature_9	0.460755	3.398097e-37
feature_3__feature_4	0.685948	4.146228e-96
feature_3__feature_5	0.722462	3.061101e-111
feature_3__feature_6	0.713878	1.807287e-107
feature_3__feature_7	0.735344	3.567289e-117
feature_3__feature_8	0.717963	3.018221e-109
feature_3__feature_9	0.441258	6.531371e-34
feature_4__feature_5	0.594548	1.627087e-66
feature_4__feature_6	0.670648	2.058229e-90
feature_4__feature_7	0.668567	1.153511e-89
feature_4__feature_8	0.603121	6.883291e-69
feature_4__feature_9	0.418898	2.125287e-30
feature_5__feature_6	0.585716	3.828707e-64
feature_5__feature_7	0.618128	3.199163e-73
feature_5__feature_8	0.628926	1.734754e-76
feature_5__feature_9	0.480583	9.266128e-41
feature_6__feature_7	0.680615	4.391890e-94
feature_6__feature_8	0.584280	9.158074e-64
feature_6__feature_9	0.339210	7.473326e-20
feature_7__feature_8	0.665602	1.312645e-88
feature_7__feature_9	0.346011	1.214400e-20
feature_8__feature_9	0.433757	1.053441e-32

PCC results would be in range from -1 to 1, the greater PCC result is, the stronger positive correlation there is between 2 features. And the smaller PCC result is, the stronger negative correlation there would be. Independence could imply when PCC result is 0.

Specifical results are shown as follow:

For those features that PCC > 0.5, they have strong positive correlation:

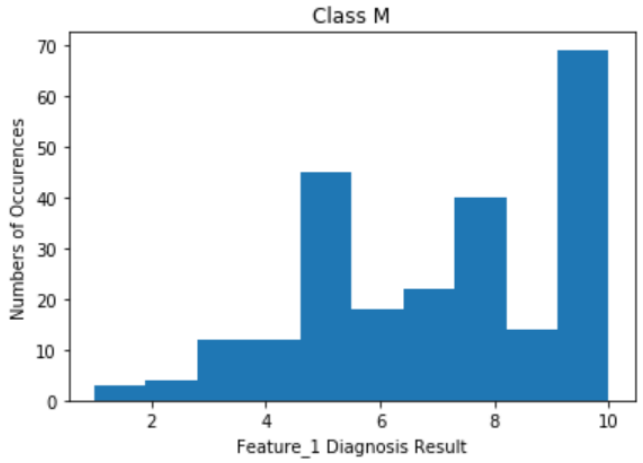
	PCC	p-value
feature_1__feature_2	0.642481	8.964173e-81
feature_1__feature_3	0.653470	2.064616e-84
feature_1__feature_5	0.523596	2.411759e-49
feature_1__feature_6	0.593091	4.050902e-66
feature_1__feature_7	0.553742	3.880813e-56
feature_1__feature_8	0.534066	1.260114e-51
feature_2__feature_3	0.907228	2.567742e-258
feature_2__feature_4	0.706977	1.544338e-104
feature_2__feature_5	0.753544	3.535863e-126
feature_2__feature_6	0.691709	2.402961e-98
feature_2__feature_7	0.755559	3.184894e-127
feature_2__feature_8	0.719346	7.433029e-110
feature_3__feature_4	0.685948	4.146228e-96
feature_3__feature_5	0.722462	3.061101e-111
feature_3__feature_6	0.713878	1.807287e-107
feature_3__feature_7	0.735344	3.567289e-117
feature_3__feature_8	0.717963	3.018221e-109
feature_4__feature_5	0.594548	1.627087e-66
feature_4__feature_6	0.670648	2.058229e-90
feature_4__feature_7	0.668567	1.153511e-89
feature_4__feature_8	0.603121	6.883291e-69
feature_5__feature_6	0.585716	3.828707e-64
feature_5__feature_7	0.618128	3.199163e-73
feature_5__feature_8	0.628926	1.734754e-76
feature_6__feature_7	0.680615	4.391890e-94
feature_6__feature_8	0.584280	9.158074e-64
feature 7 feature 8	0.665602	1.312645e-88

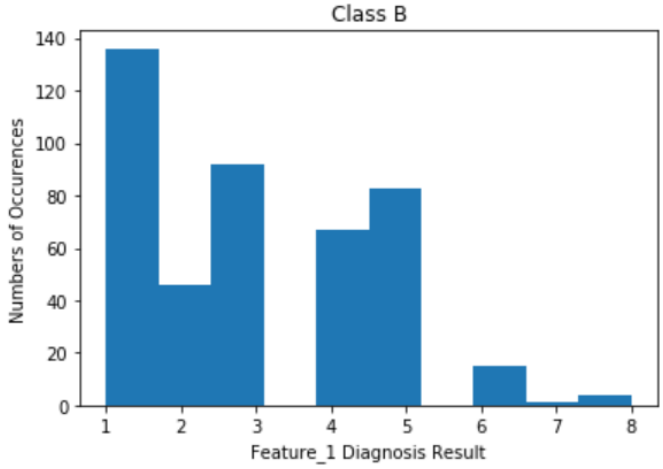
For those features that 0< PCC < 0.5, they have weak positive correlation:

	PCC	p-value
feature_1__feature_4	0.487829	4.027956e-42
feature_1__feature_9	0.350957	3.148289e-21
feature_2__feature_9	0.460755	3.398097e-37
feature_3__feature_9	0.441258	6.531371e-34
feature_4__feature_9	0.418898	2.125287e-30
feature_5__feature_9	0.480583	9.266128e-41
feature_6__feature_9	0.339210	7.473326e-20
feature_7__feature_9	0.346011	1.214400e-20
feature_8__feature_9	0.433757	1.053441e-32

Q3

I take the firt feature to show the diagnosis results in graph for both the Class M and Class B.





Codes are attached as below:

In [23]:

```
import numpy as np
import pandas as pd
from pandas import Series,DataFrame
import itertools
import matplotlib.pyplot as plt
from scipy.stats.stats import pearsonr
data=pd.read_csv('Desktop/ece657/breast-cancer-wisconsin.csv',header=None,names=['id','feature_1',"feature_2",'feature_3
data1=data.astype(str)
delet=data1[~data1['feature_6'].str.contains("\?")]
data2=delet.astype(int)
data2.drop_duplicates()
data3=data2.drop(['id'],axis=1)
data3=data3.drop(['class'],axis=1)
def describ(x):
    print ("Mean:\n",x.mean())
    print("Mode:\n",x.mode())
    print("Skew:\n",x.skew())
    print("Standard Deviation:\n",x.std())
    print("Variance Values:\n",x.var())
df=describ(data3)
def PCC(x):
    correlations = {}
    columns = x.columns.tolist()

    for col_a, col_b in itertools.combinations(columns, 2):
        correlations[col_a + '__' + col_b] = pearsonr(x.loc[:, col_a], x.loc[:, col_b])

    result = DataFrame.from_dict(correlations, orient='index')
    result.columns = ['PCC', 'p-value']
    print(result.sort_index())
    return result
df1=PCC(data3)
df_s=df1[df1['PCC']>0.5]
print("\n For those features that PCC > 0.5, they have strong positive correlation: \n")
print (df_s)
df_w=df1[df1['PCC']<0.5]
print("\n For those features that 0< PCC < 0.5, they have weak positive correlation: \n")
print (df_w)
data_m=data2[~data2['class'].isin([2])]
data_b=data2[~data2['class'].isin([4])]
def plot(x):
    fig = plt.figure()
    ax = fig.add_subplot(111)
    ax.hist(x['feature_1'])
    plt.title('Class M')
    plt.xlabel('Feature_1 Diagnosis Result')
    plt.ylabel('Numbers of Occurences')
    plt.show()
plot(data_m)
def plot(x):
    fig = plt.figure()
    ax = fig.add_subplot(111)
    ax.hist(x['feature_1'])
    plt.title('Class B')
    plt.xlabel('Feature_1 Diagnosis Result')
    plt.ylabel('Numbers of Occurences')
    plt.show()
plot(data_b)
```