North South University

Department of Mathematics & Physics

Course Code : MAT 361

Course Title: Probability And Statistics

Course Instructor: Dr. M.A. Masud

Date of Submission: 16/12/2018

Section: 09

Submitted By	Score
Student Name : Townim Faisal ID: 1721327042	

Dataset: https://www.kaggle.com/saurabh00007/diabetescsv

Code: (Attached the code with the file)

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

get_ipython().run_line_magic('matplotlib', 'inline')

diabetes_data = pd.read_csv('diabetes.csv')
print(diabetes_data.shape)
diabetes_data.head(10)

grp1 = 0
grp2 = 0
grp3 = 0
grp4 = 0
grp5 = 0

for i in range(diabetes_data.shape[0]):
    if diabetes_data['Age'][i]<=15 and diabetes_data['Outcome'][i] == 1:
        grp1+=1
    if diabetes_data['Age'][i]<=30 and diabetes_data['Age'][i]>15 and diabetes_data['Outcome'][i] == 1:
        grp1+=1
    if diabetes_data['Age'][i]<=45 and diabetes_data['Age'][i]>30 and diabetes_data['Outcome'][i] == 1:
        grp1+=1
    elif diabetes_data['Age'][i]<=60 and diabetes_data['Age'][i]>45 and diabetes_data['Outcome'][i] ==
1:
        grp4 += 1
    else:
        grp5 += 1


diabetes_data['Age'][0]

age_data = {'Patient Count': pd.Series(data=[grp1,grp2,grp3,grp4,grp5],
                                       index=['0-15', '15-30', '31-45', '46-60','60-100'])}
age_table = pd.DataFrame(age_data)
age_table

fig = plt.figure()

plt.title("Distribution of Diabetes Patient of different ages")
plt.xlabel('Age')
plt.ylabel('Count')
plt.bar(['0-15', '15-30', '31-45', '46-60','60-100'],[grp1,grp2,grp3,grp4,grp5])

fig.savefig('Age.png')
plt.show()
```

```
fig = plt.figure()

plt.title("Histogram of Body Mass Index (BMI) of patients")
plt.xlabel('BMI Distribution')
plt.ylabel('Count')

bin_edges = np.arange(0,diabetes_data['BMI'].max()+5, 5)
plt.hist(data=diabetes_data,x='BMI', bins=bin_edges)

fig.savefig('BMI.png')

bmi_updated = []

for d in diabetes_data.BMI:
   if d >10 and d<60:
      bmi_updated.append(d)


from scipy.stats import gamma

bmi_updated = np.array(bmi_updated)

sb.distplot(bmi_updated)
plt.xlabel('BMI Distribution')
plt.ylabel('Count')

shape, scale = bmi_updated.mean(), bmi_updated.std()
s = np.random.gamma(shape, scale)


import scipy.special as sps
count, bins, ignored = plt.hist(bmi_updated, 50, density=True)
y = bins**(shape-1)*(np.exp(-bins/scale)/(sps.gamma(shape)*scale**shape))
plt.plot(bins, y, linewidth=2, color='r')
plt.show

count, bins, ignored = plt.hist(s, 50, density=True)
y = bins**(shape-1)*(np.exp(-bins/scale) /
                (sps.gamma(shape)*scale**shape))
plt.plot(bins, y, linewidth=2, color='r')
plt.show()

age_table['Patient Count'].value_counts()

color = [(0.8392156862745098, 0.15294117647058825, 0.1568627450980392),
       (0.17254901960784313, 0.6274509803921569, 0.17254901960784313),
       (0.09019607843137255, 0.7450980392156863, 0.8117647058823529),
       (0.12156862745098039, 0.4666666666666667, 0.7058823529411765),
       (0.09019607843137255, 0.7450980392156863, 0.8117647058823529)]

fig = plt.figure()
plt.title("Pie chart of Diabetes Patient of different ages ")
```

```
plt.pie([grp1,grp2,grp3,grp4,grp5],labels=['0-15', '15-30', '31-45', '46-60','60-100'],colors=color)

fig.savefig('Pie.png')
plt.show()

sb.boxplot(x=bmi_updated)
plt.boxplot(diabetes_data['BMI'])

plt.xlabel('BMI')
plt.ylabel('BMI index')
plt.show()

diabetes_data.BMI.std()

from scipy import stats

mean, sd = stats.norm.fit(bmi_updated)
print(mean,sd)

n, bins, patches = plt.hist(bmi_updated, 10, density=1)
y = stats.norm.pdf(bins, mean, sd)
plt.plot(bins, y, color='r')
plt.show
```

Graph:

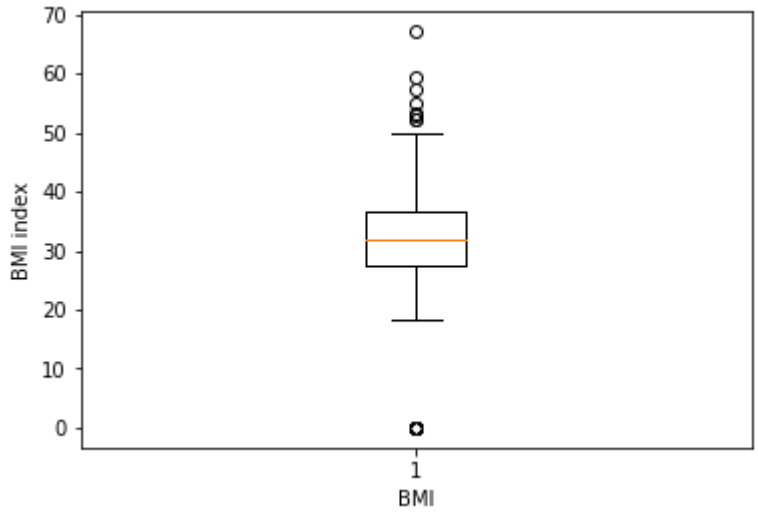
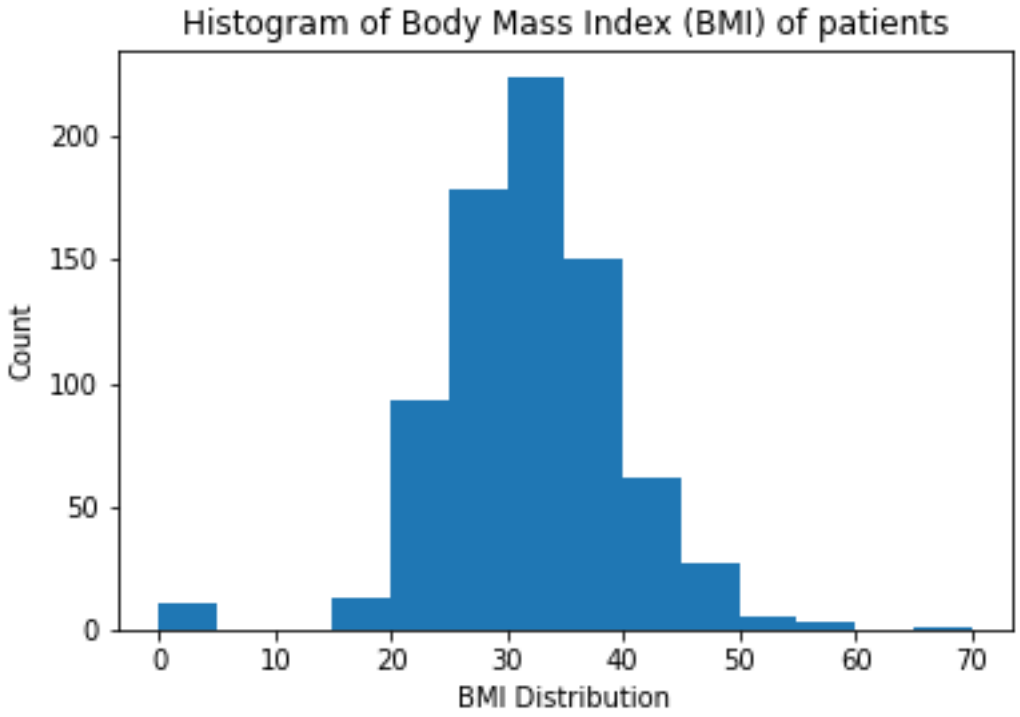
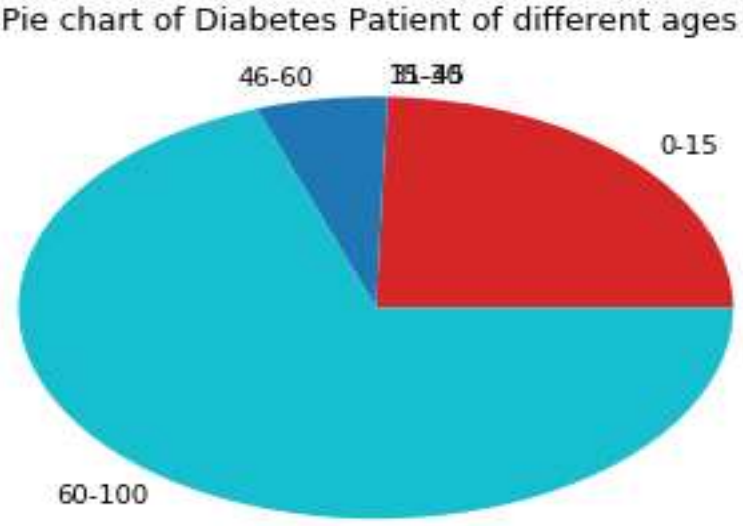


Figure: Boxplot of BMI


Histogram of Body Mass Index (BMI) of patients
Count
200
150
100
50
0
0
10
20
30
40
50
60
70
BMI Distribution


Pie chart of Diabetes Patient of different ages
46-60
0-15
60-100

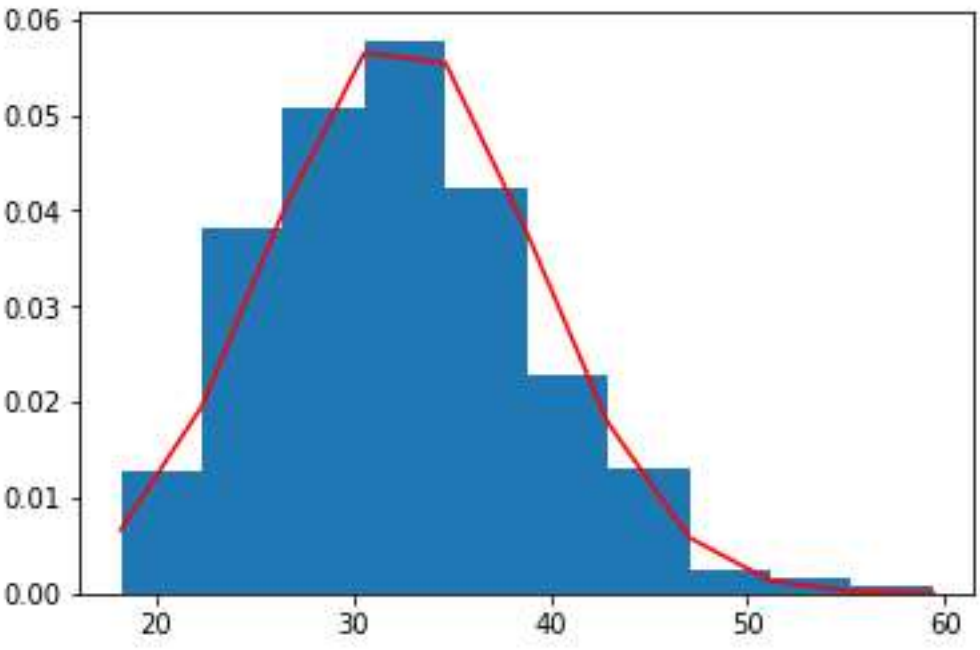


Figure: Normal distribution of BMI

Standard deviation and Mean:

Mean = 32.41164021164021

Standard deviance = 6.809252727468057

Explanation: Here in boxplot we found outlier. So we do not choose that part in our distribution. We have plotting the next histogram by following normal distribution. And following that distribution, we get our standard deviance and mean value.