Assignment 2 Plotting Data Scientists

Introduction

Assume that we have a data-science club where data scientists meet and discuss data analysis and visualization. The members in the club are either paid accounts or unpaid accounts. You are provided a list of tuples. Each tuple contains three elements:

- tenure, which is the number of years as a data scientist,
- salary, which is how much the data scientist ears,
- account, which is a number that is either 1 for a paid account or 0 for an unpaid account.

In [3]:

```
import random
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

In [4]:

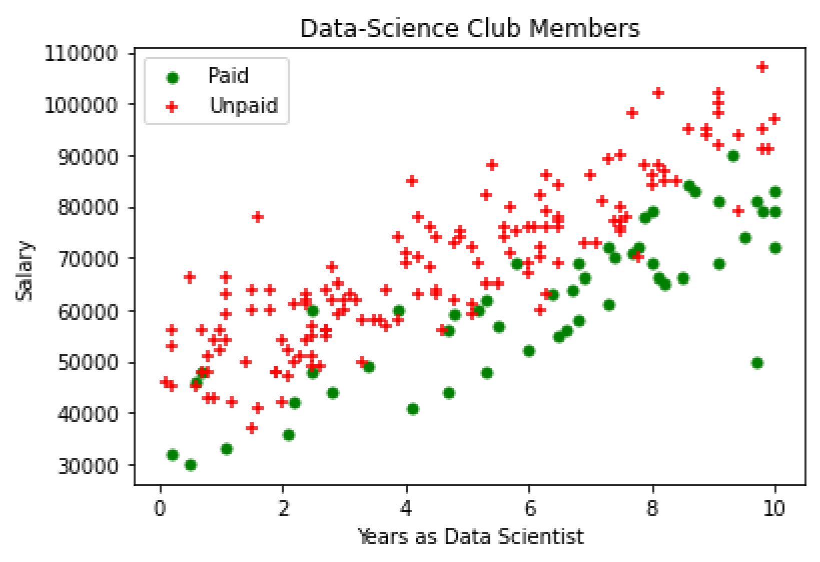
```
#do not change the below statement
data = [(0.7,48000,1),(1.9,48000,0),(2.5,60000,1),(4.2,63000,0),(6,76000,0),(6.5,69000,0),(7.5,76000,0),(8.1,8
```

Problem 1: Plotting the Club Members

For this problem, you need to plot the data scientists in the data-science club so that we can conveniently visualize their tenured years, salaries, and paid accounts or not. Your plot should be similar to the figure that can be displayed below by running the code cell.

In [82]:

```
from IPython.core.display import Image
Image(filename='p1.png')

paidy=[]
paidx=[]
unpaidy=[]
unpaidx=[]

for i in data:
    if(i[2]==1):
        paidy.append(i[0])
        paidx.append(i[1])
    else:
        unpaidy.append(i[0])
        unpaidx.append(i[1])

plt.scatter(paidy, paidx, label="Paid", color="green", marker=".", s=100)
plt.scatter(unpaidy, unpaidx, label="Unpaid", color="red", marker="+", s=30)
plt.xlabel('Years as Data Scientist')
plt.ylabel('Salary')
plt.title('Data-Science Club Members')
plt.legend()

plt.show()
```

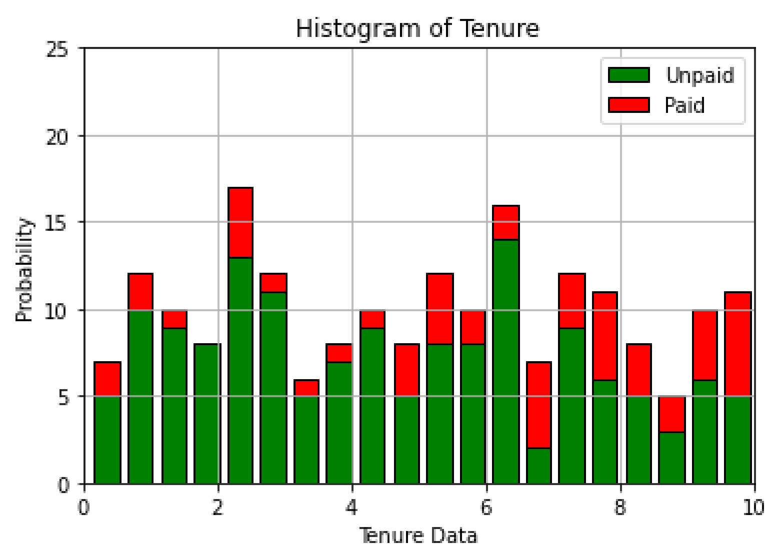


Problem 2

For this problem, you need to create a histogram plot. The histogram plot allows us discover the underlying frequency distribution of tenure data associated with the data scientists including paid and unpaid members in the club. Your plot should be similar to the figure that can be displayed below by running the code cell.

In [80]:

```
from IPython.core.display import Image
Image(filename='p2.png')

unpaid = []
paid = []

for i in data:
    if i[2] == 1:
        unpaid.append(i[0])
    else:
        paid.append(i[0])

plt.hist([paid, unpaid], stacked = True, color = ['green','red'], bins = 20, edgecolor = 'black', linewidth =
plt.xlabel('Tenure Data')
plt.ylabel('Probability')
plt.title('Histogram of Tenure')
plt.legend(['Unpaid', 'Paid'], loc = 0)
plt.axis([0, 10, 0, 25])
plt.grid(True)
plt.show()
```



Problem 3

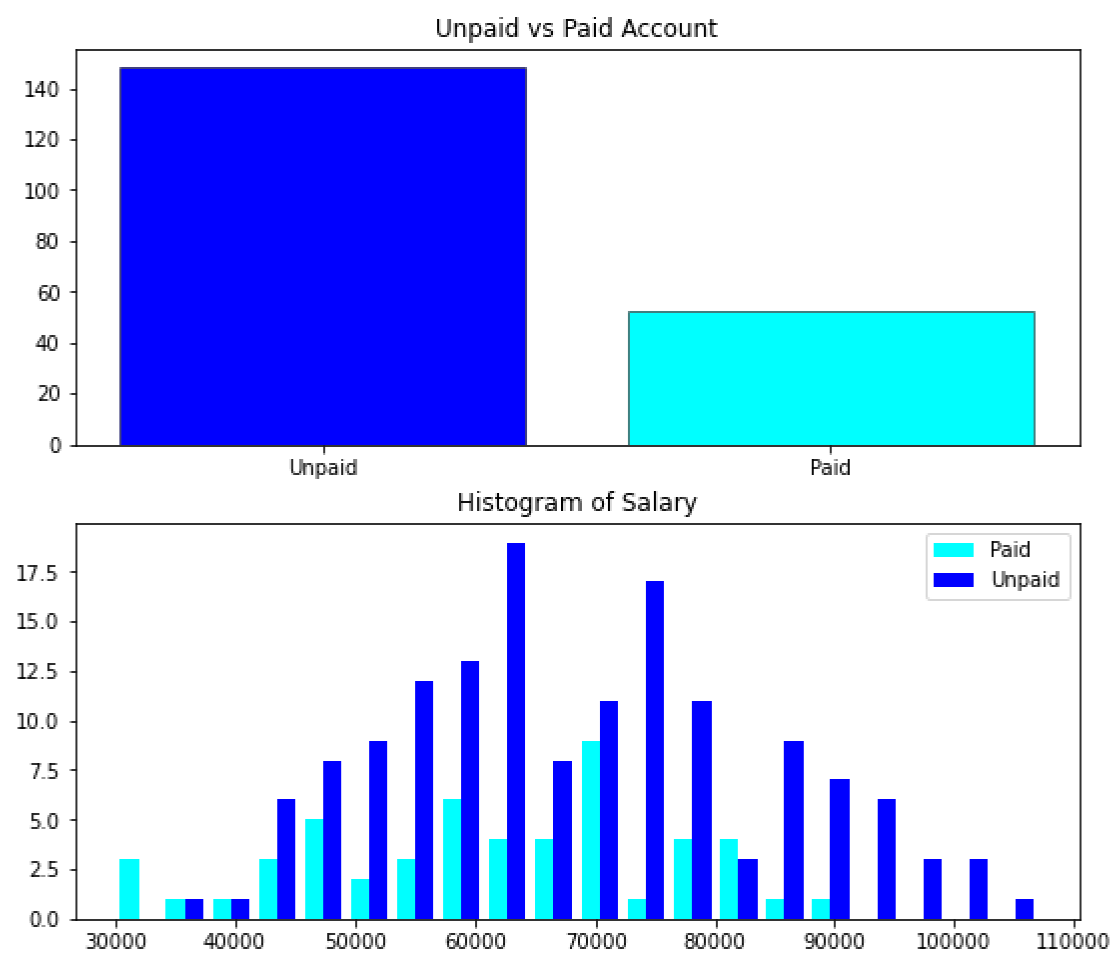
For this problem, you need to make a figure with two subplots. One is a bar plot, and the other is a histogram plot. The bar plot can conveniently support users to visualize how many accounts are paid and how many are unpaid. The histogram plot allows us discover the underlying frequency distribution of salary data associated with the two types of data scientists (paid and unpaid). Your plot should be similar to the figure that can be displayed below by running the code cell.

In [79]:

```
from IPython.core.display import Image
Image(filename='p3.png')

paid=0

for i in data:
    if (i[2]==1):
        paid+=1

unpaid=len(data)-paid
bar1=[unpaid,paid]
bar2=["Unpaid","Paid"]

paid=[]
unpaid=[]
for i in data:
    if(i[2]==0):
        unpaid.append(i[1])
    else:
        paid.append(i[1])

fig, (ax1, ax2) = plt.subplots(nrows=2,ncols=1, figsize = (9,8))

ax1.set_title("Unpaid vs Paid Account")
ax2.set_title("Histogram of Salary")

ax1.bar(bar2, bar1, color = ['blue', 'cyan'], linewidth = 0.5, edgecolor = "black")
labels = ["Paid","Unpaid"]
ax2.hist([paid,unpaid], bins = 20, color = ["cyan","blue"], label = labels)
ax2.legend(['Paid', 'Unpaid'], loc=0)
```

Out[79]:

```
<matplotlib.legend.Legend at 0x214e76aa7c0>
```

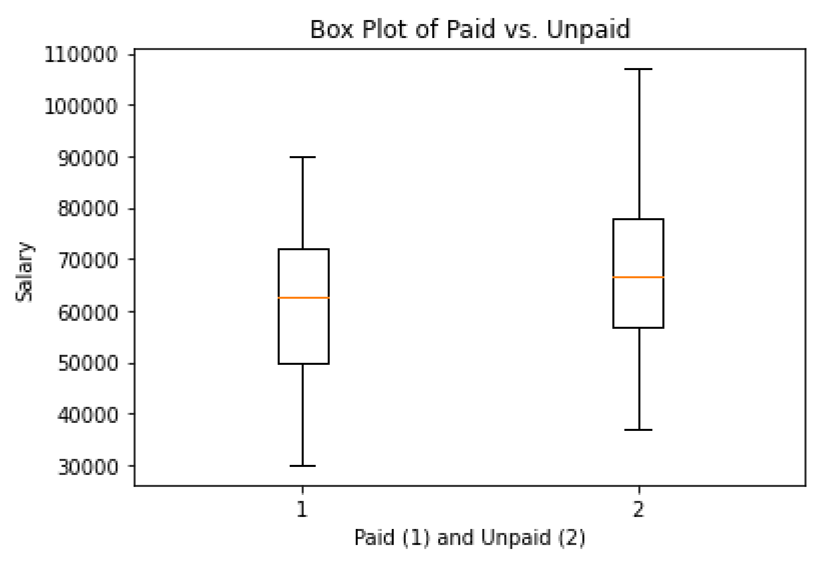


Problem 4 One More Thing

Create another plot that is different from the previous ones you created in this assignment. Provide your code that can be applied to create the plot and briefly describe what the generated plot is following your code cell.

In [78]:

```
from IPython.core.display import Image
Image(filename='p1.png')

paid=[]
unpaid=[]
for i in data:
    if(i[2]==0):
        unpaid.append(i[1])
    else:
        paid.append(i[1])

fig, ax1 = plt.subplots()

ax1.boxplot([paid, unpaid])
ax1.set_title('Box Plot of Paid vs. Unpaid')
ax1.set_ylabel('Salary')
ax1.set_xlabel('Paid (1) and Unpaid (2)')

plt.show()


#-------------------------READ BELOW---------------------------------------------
#The plot generated for Question 4 is a box plot, box plot 1 representing paid data scientists and box plot 2
```



Turn-in

Turn in your notebook including your Python code and answers to the questions to D2L Assignments folder **Assignment 2**. In addition to the notebook document, you need to provide a pdf exported from your executed notebook. (That is, if a code cell's execution has generated output, the pdf file should include the output presentation.)