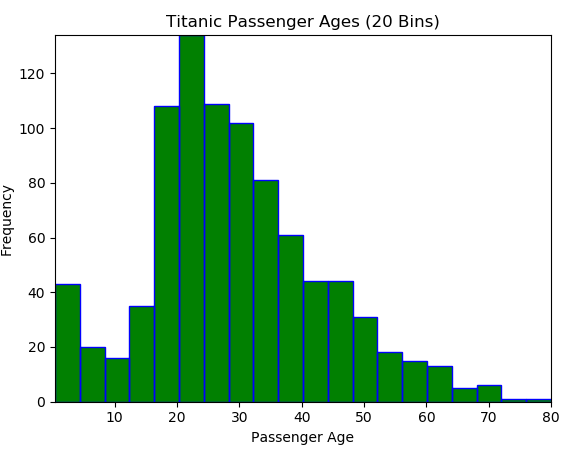
Fundamentals of Data Science Homework 1

1. Many things can be concluded from the age histogram. Firstly, the most common age group for titanic passengers was young adults, from late teenagers to early 30s. After that, the number of older passengers decreases steadily; there were only 2 passengers older than 71. Also, compared to adults, there were not many children on board, especially aged 9-13.

Another insight learned is that the bin size of histograms has a significant effect on how the data is presented. The fewer bins, the easier the graph is to read but important detail could be lost.

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.patches as patches

import matplotlib.path as path

import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

ages = []

for column in reader:

ages.append(column[3])

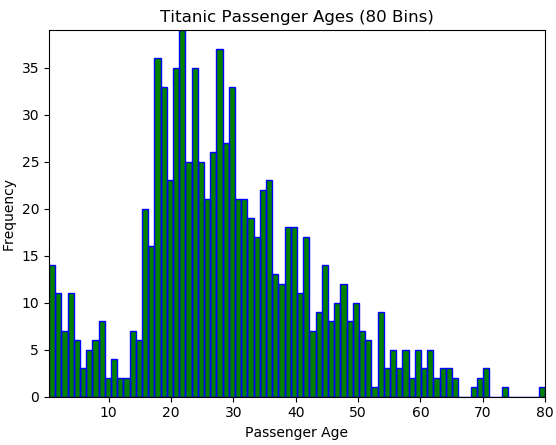
ages.pop(0)

ages = list(map(float, ages))

data = ages

n, bins = np.histogram(data, 20)

left = np.array(bins[:-1])

right = np.array(bins[1:])

bottom = np.zeros(len(left))

top = bottom + n

nrects = len(left)

nverts = nrects \* (1 + 3 + 1)

verts = np.zeros((nverts, 2))

codes = np.ones(nverts, int) \* path.Path.LINETO

codes[0::5] = path.Path.MOVETO

codes[4::5] = path.Path.CLOSEPOLY

verts[0::5, 0] = left

verts[0::5, 1] = bottom

verts[1::5, 0] = left

verts[1::5, 1] = top

verts[2::5, 0] = right

verts[2::5, 1] = top

verts[3::5, 0] = right

verts[3::5, 1] = bottom

fig, ax = plt.subplots()

barpath = path.Path(verts, codes)

patch = patches.PathPatch(barpath, facecolor='green', edgecolor='blue', alpha=2)

ax.add\_patch(patch)

ax.set\_xlim(left[0], right[-1])

ax.set\_ylim(bottom.min(), top.max())

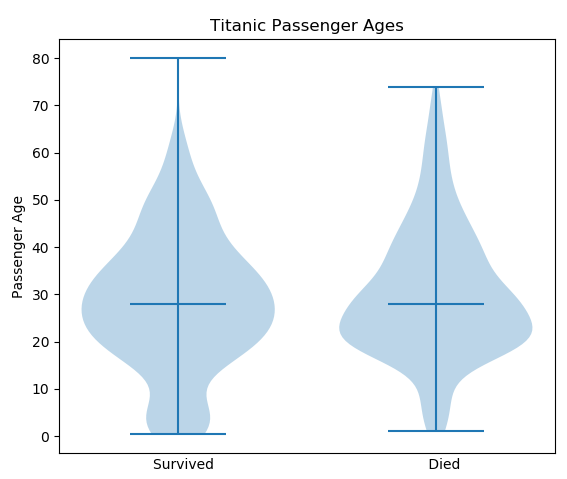
plt.xlabel('Passenger Age')

plt.ylabel('Frequency')

plt.show()

1. Conclusions drawn from the violin graph include children being a larger proportion of survivors than of casualties. There were many more elders that died than survived as well. Interestingly, the median age of the survivors and fatalities were very close.

Violin graphs are somewhat limited in the detail they show, they are effective at showing proportionality but bad for estimating specific frequencies without a scale.

import numpy as np

import matplotlib.pyplot as plt

import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

agesLived = []

agesDied = []

for column in reader:

if (column[0] == '1'):

agesLived.append(column[3])

else:

agesDied.append(column[3])

agesDied.pop(0)

agesLived = list(map(float, agesLived))

agesDied = list(map(float, agesDied))

pos = [1.5, 2.5]

data = [agesLived, agesDied]

plt.violinplot(data, pos, points=100, widths=0.75, showextrema=True, showmedians=True)

plt.title('Titanic Passenger Ages')

plt.xlabel('Survived Died')

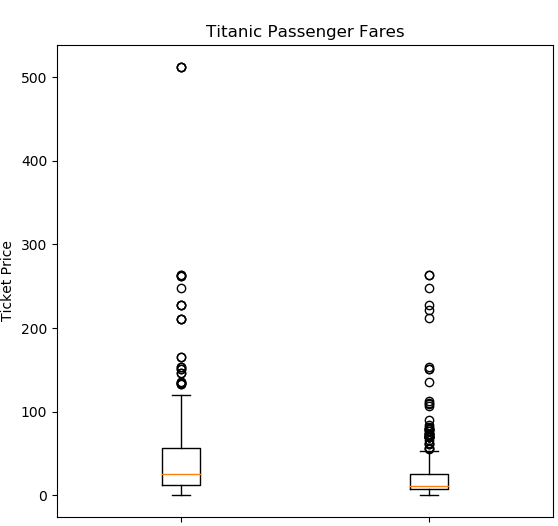
plt.ylabel('Passenger Age')

frame1 = plt.gca()

frame1.axes.get\_xaxis().set\_ticks([])

plt.show()

1. The 2 box and whisker plots show a correlation between surviving and economic class. Comparing the box of the survivors’ plot on the left and the box of the casualties’ box on the right shows that the lower class made up a larger proportion of the casualties than the survivors. The prices of the survivors’ tickets were much more spread out.



import numpy as np

import matplotlib.pyplot as plt

import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

pricesLived = []

pricesDied = []

for column in reader:

if (column[0] == '1'):

pricesLived.append(column[6])

else:

pricesDied.append(column[6])

pricesDied.pop(0)

pricesLived = list(map(float, agesLived))

pricesDied = list(map(float, agesDied))

data = [agesLived, agesDied]

plt.title('Titanic Passenger Fares')

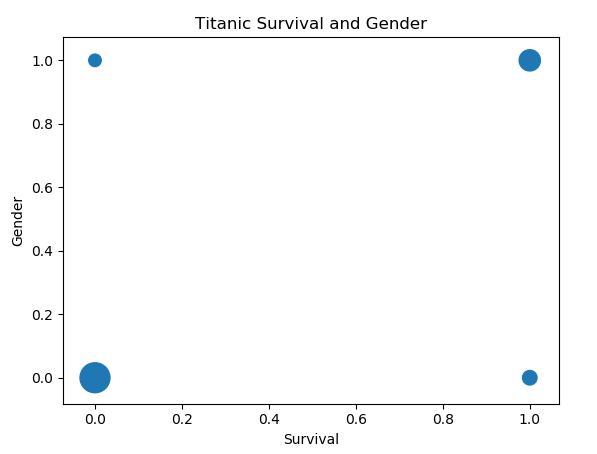
plt.ylabel('Ticket Price')

fig, ax = plt.subplots()

ax.boxplot(data)

plt.show()

1. Few conclusions can be made from the gender and survival scatter plot. There are 2 binary axes and thus only 4 possible combinations. Adding a size variable gives more detail, but still does not allow precise estimations or comparisons. For 2 binary axes, a scatter plot is a horrible visualization option. Yet it can still be determined that there were more male deaths and more female survivors.

import numpy as np

import matplotlib.pyplot as plt

import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

survivedAndMale = 0

survivedAndFemale = 0

diedAndMale = 0

diedAndFemale = 0

for column in reader:

if (column[0] == '1' and column[2] == '0'):

survivedAndMale += 1

if (column[0] == '1' and column[2] == '1'):

survivedAndFemale += 1

if (column[0] == '0' and column[2] == '0'):

diedAndMale += 1

if (column[0] == '0' and column[2] == '1'):

diedAndFemale += 1

plt.scatter([0, 0, 1, 1], [0, 1, 0, 1], s = [diedAndMale, diedAndFemale, survivedAndMale, survivedAndFemale])

plt.title('Titanic Survival and Gender')

plt.ylabel('Gender')

plt.xlabel('Survival')

plt.show()

1. Histograms are a much better graph for displaying the frequency of 4 groups. From this survival and gender histogram, it can be concluded that the number of women that survived was double the number of men that survived. Also, the number of men that died was nearly 5 times the number of women that died. The histogram shows that there were far more casualties than survivors in total from the titanic accident.

import numpy as np

import matplotlib.pyplot as plt

from matplotlib.ticker import MaxNLocator

from collections import namedtuple

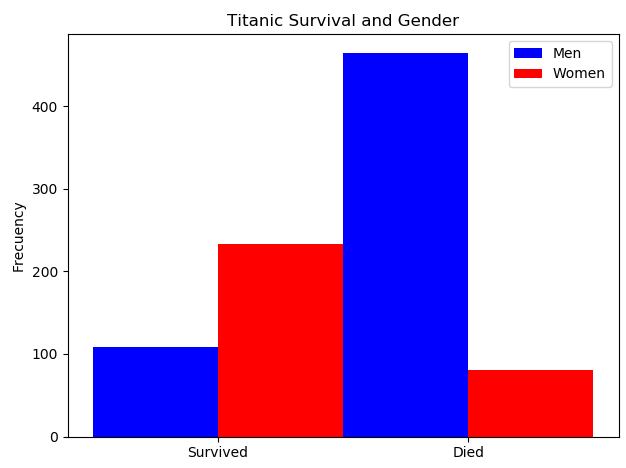
import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

survivedAndMale = 0

survivedAndFemale = 0

 diedAndMale = 0

diedAndFemale = 0

for column in reader:

if (column[0] == '1' and column[2] == '0'):

survivedAndMale += 1

if (column[0] == '1' and column[2] == '1'):

survivedAndFemale += 1

if (column[0] == '0' and column[2] == '0'):

diedAndMale += 1

if (column[0] == '0' and column[2] == '1'):

diedAndFemale += 1

n\_groups = 2

means\_men = (survivedAndMale, diedAndMale)

means\_women = (survivedAndFemale, diedAndFemale)

fig, ax = plt.subplots()

index = np.arange(n\_groups)

bar\_width = 0.5

rects1 = ax.bar(index, means\_men, bar\_width, color='b', label='Men')

rects2 = ax.bar(index + bar\_width, means\_women, bar\_width, color='r', label='Women')

ax.set\_ylabel('Frecuency')

ax.set\_title('Titanic Survival and Gender')

ax.set\_xticks(index + bar\_width / 2)

ax.set\_xticklabels(('Survived', 'Died'))

ax.legend()

fig.tight\_layout()

plt.show()

1. It can be concluded that the number of women that survived was double the number of men that survived. Also, the number of men that died was nearly 5 times the number of women that died. The histogram shows that there were far more casualties than survivors in total from the titanic accident.

import numpy as np

import matplotlib.pyplot as plt

from matplotlib.ticker import MaxNLocator

from collections import namedtuple

import csv

with open('D:/Coding/Python/FundDataSci/Titanic\_data.csv') as file:

reader = csv.reader(file)

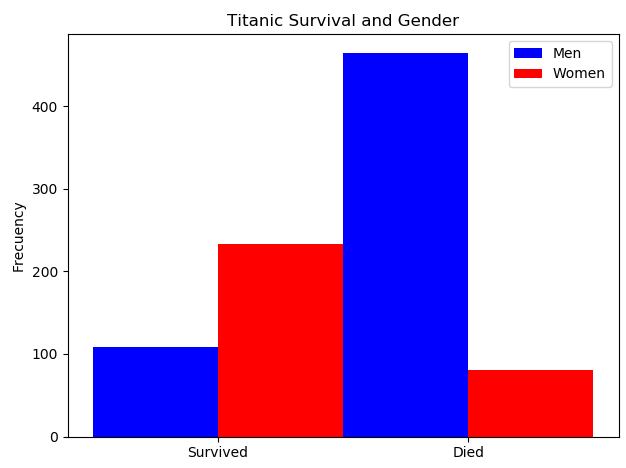
survivedAndMale = 0

survivedAndFemale = 0

diedAndMale = 0

diedAndFemale = 0

for column in reader:

 if (column[0] == '1' and column[2] == '0'):

survivedAndMale += 1

if (column[0] == '1' and column[2] == '1'):

survivedAndFemale += 1

if (column[0] == '0' and column[2] == '0'):

diedAndMale += 1

if (column[0] == '0' and column[2] == '1'):

diedAndFemale += 1

n\_groups = 2

means\_men = (survivedAndMale, diedAndMale)

means\_women = (survivedAndFemale, diedAndFemale)

fig, ax = plt.subplots()

index = np.arange(n\_groups)

bar\_width = 0.5

rects1 = ax.bar(index, means\_men, bar\_width, color='b', label='Men')

rects2 = ax.bar(index + bar\_width, means\_women, bar\_width, color='r', label='Women')

ax.set\_ylabel('Frecuency')

ax.set\_title('Titanic Survival and Gender')

ax.set\_xticks(index + bar\_width / 2)

ax.set\_xticklabels(('Survived', 'Died'))

ax.legend()

fig.tight\_layout()

plt.show()