NAMES: Tyler Hudson tkh51

Data Mining and Visualization Homework 1

1) Download the titanic.txt file to a location whose filepath you know. For instance in my laptop this file location is [C:\Users\rm84\Desktop\Teaching\3339](file:///C:\C:\Users\rm84\Desktop\Teaching\3339\titanic.txt) In your laptop it is:

C:\Users\tyler\OneDrive\Documents\school2k20\Fall2020\CIS3339\Hw1

2) Write down the script you would need to have in order to import the flat text file and equate it . For instance if I were doing this homework I would answer this question as:

file1=pd.read\_csv("[C:/Users/rm84/Desktop/Teaching/3339/titanic.txt](file:///C:\C:\Users\rm84\Desktop\Teaching\3339\titanic.txt)",sep=",")

file1=pd.read\_csv("C:/Users/tyler/OneDrive/Documents/school2k20/Fall2020/CIS3339/Hw1/titanic.txt",sep=",")

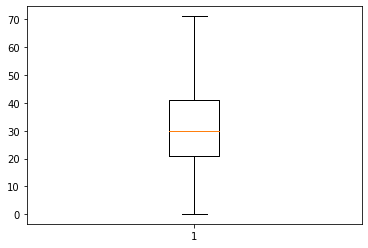
Note that if you are a MAC user your operating System is NOT going to have a [C:\](file:///C:\C:\)

3) Find the dimensions of the titanic text file. Write the dimension and the script.

In[19]: file1.shape

Out[19]: (1313, 11)

4) Do a boxplot of all individuals’ ages. You need to use np.nanpercentile instead of np.percentile function. Copy/Paste Script and the boxplot. Np is the abbreviation for numpy package. This is already done when you run import numpy as np .

plt.boxplot(file1[file1['age'].notnull()]['age'])

5) Find the IQR for age. You need to use np.nanpercentile function.

Write the script and its numerical evalution.

IQR=np.nanpercentile(file1['age'],75)-np.nanpercentile(file1['age'],25)

IQR= 20

6) Use the 1.5 multiplier with the IQR to come up with the Upper and Lower Bound. Copy/Paste the script to do this.

UpperBound=np.nanpercentile(file1['age'],75)+1.5\*IQR

LowerBound=np.nanpercentile(file1['age'],25)-1.5\*IQR

7) How many “age” values are above the upper bound and how many are below the lower bound. Write the script to do this and the numerical result.

howmanyoutliers=sum(file1['age']>UpperBound) 🡺 0

howmanyoutliers=sum(file1['age']<LowerBound) 🡺 0

8) Set a Crita boolean object that will allow you to subset titanic to only male passengers via the sex variable. COPY/PASTE the script to do this.

df.sex[df.sex==”male”]

9) The and operator in Python and a lot of other programming languages is &. The OR operator in Python is (either google this or find it on the lecture1introduction.py file. I use this operator in the python file to find the number of individuals with missing age or ticket numbers.

The OR operator in Python is |.