CS215 Chapter7A Lab – Data Analysis with DataFrames Name:_____

Background on the Data Set: A study was undertaken to examine physicians working for an emergency department at a major hospital. The goal was to determine which of a number of factors are related to the number of complaints received during the preceding year. For each of the 44 phyrisicans, the following was recorded.

- num_visits: number of patient visits attended to
- num_complaints: number of complaints received
- residency: whether or not the physician is training
- gender: male/female
- revenue: dollars per hour
- hrs: workload at the emergency department

***Create a Python file called Lab7B_MoreOnDFs.py.

1. Import the attached .csv file into a data frame called df.

2. Print the column headers, dimensions, and shape of the data.

3. Let's say physician A has 5 complaints and physician B has 1 complaint. It seems like physician A is doing poorly. But what if physician A got 5 complaints after working 5000 hours and physician B got 1 complaint after working 1000 hours? Well, in that case, they have the same rate of complaints. So let's calculate the number of complaints per hours worked for each physician.

   In the first record/row, we see that a physician worked 1287.25 hours and got 2 complaints, meaning the physician received 0.00155 complaints for each hour worked. That's hard to wrap our heads around, so multiplying by 1000, we get 1.55, telling us that the physician got 1.55 complaints per 1000 hours worked.

   Create a column in df called complaints_1000hrs. The first few rows/records give these values.

   1.553700
   5.149478
   1.118725
   1.082690
   3.319502

4. Call the appropriate method that you learned in the last lab to find the descriptive stats for the new column (mean/std deviation/etc). What is the max and min of the new column? 6.5 and 0

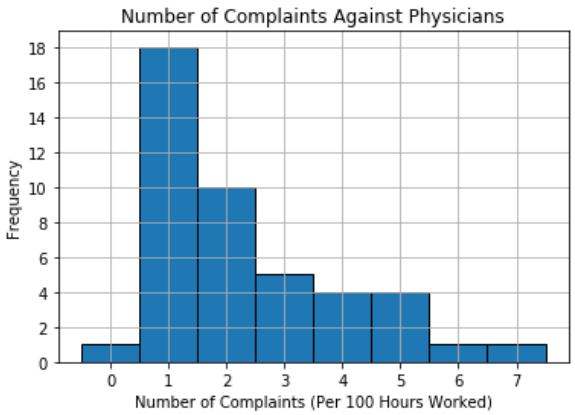
5. Our goal is to create the histogram for the complaints_1000hrs shown below.

- Use the procedure from last lab to create a histogram. It won't look quite as nice as the one at the right yet.

  FYI: You don't want to use xBins from 0-100 because our data ranges between a different man and min, so update xBins appropriately.

  I wanted 0, 1, 2, etc to be the center of all of my bars. So I chose a very special xBins: I started a little below 0 and I went up by 1 till a little above the max. So what could you put in the blank below?

```
xBins = np.arange(-0.5,8_____,1)
```



- I did not show you how to create a title for the x and y axis last time. You use the label functions like below. Every graph should have an overall title and 2 axes tiles.

  plt.xlabel('Number of Complaints (Per 100 Hours Worked)')
  plt.ylabel('Frequency')

- Run your code. The histogram is still not so pretty.

  To add the black lines: Pass in a parameter of edgecolor='black' to the hist function.

  To control where the tick marks appear on the x and y axis, add the following code BEFORE plt.show()

```
plt.xticks(range(0, 8)) #Get a tick mark at every spot 1-7 on the x axis
plt.yticks(range(0, 20, 2))  #Get a tick mark at 0, 2, 4, ... on the y axis
```

6. You have a nice histogram. Now move the describe method below the histogram so that right below the histogram is all of the interesting information.

7. What do we learn from this information? The following would be a well-written answer worthy of full credit on a test.

   The average number of complaints (per 1000 hours) is 2.25, but the data can range from 0 to 7. However, most people have received 1-2 complaints (per 1000 hours). I also notice that the shows that only 1 person has received no complaints so that seems rare. Also, the data is skewed right in a descending fashion (so fewer and fewer people receive more and more complaints).

8. Now let's see if the number of complaints (per 1000 hours) seems similar for men and women. This means we have to get the Female Rows ('F') and Male Rows ('M') out of the table using the gender column. Here's how to do this.

```
criteria = (df["gender"] == 'F')   #This creates a criteria comparing
                                      #gender in each row to 'F'
females = df[criteria] #This selects all rows from our data frame that meet
                         #our criteria
```

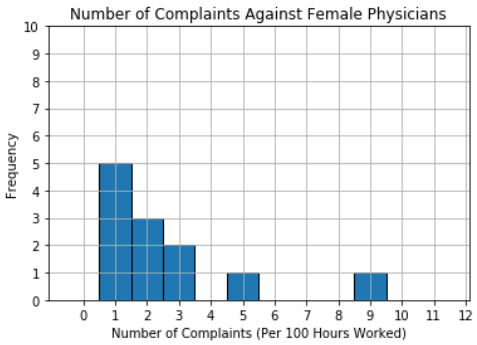
9. Create a males data frame in a similar fashion.

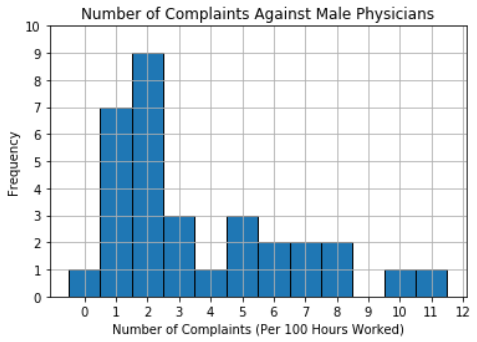
10. Get the descriptive stats for the new female's data frame and create the histograms shown at the right for the data. Then do the same for the males. Notice that I have made the x/y axis scales the same in both histograms because this allows for easy comparison. You should always do this.

    You may wonder why there is less "blue" area for the females versus the males. Well, how many rows/records are in each set?
    Use the descriptive stats information to decide.

    #records in male data frame: 32

    #records in female data frame: 12





11. What do we learn from this information? The following would be a well-written answer worthy of full credit on a test.

Women: The average number of complaints (per 1000 hours) is 1.82 which is below average (compared to 2.25, the mean of the whole data set), but the data can range from 1 to 9. However, most females receive 1-2 complaints (per 1000 hours). More than 3 complaints is seems rare. Also, the data is skewed right in a descending fashion (so fewer and fewer females receive more and more complaints).

Men: The average number of complaints (per 1000 hours) for men is 2.42 which is higher than women and above average (compared to 2.25, the mean of the whole data set), but the data can range from 0 to 11. However, most males receive 1-2 complaints (per 1000 hours). More than 3 complaints is seems rare. Also, the data is skewed right in a somewhat descending fashion (so fewer and fewer males receive more and more complaints). However, the number of males who received 3-8 complaints is pretty consistent. The 2 who received 10-11 complaints could be outliers.

Overall: This data set may suggest that men seem to get more complaints than women on average. If I had to speculate to explain why this might be the case, I might say that perhaps people often are used to being taken care of traditionally by their mothers and so the male persona in such situations may seem jarring to patients, even if the male nursing staff are no different than the female staff.

One thing to note is that there is only 12 females versus 32 males so this is a small data set. I would like to look at a larger set to see if I can trust the results more.

12. To get full credit for this lab, first pick one of the questions below.

    - How do the distribution of complaints compare for residents vs. non-residents?
    - How do the distribution of complaints compare for those who receive above average or below average revenue?

    Then do the following.

    - Calculate the descriptive stats and the histograms data for number of complaints (per 1000 hours) for the categories related to your problem.
    - Write a summary indicating what you learn from these results. Then speculate and try to explain why this might be the case. (You'll put this in the text submission area when you submit.)

SUBMISSION INFO

TO GET CREDIT FOR THIS LAB, UPLOAD THE FOLLOWING TO THE SUBMISSION AREA.

- Your pdf doc with any blanks filled in.
- Your code
- Your summary placed in the text submission area.