Lab 05: Pandas Data Analysis Library

In this lab/discussion, we will start to learn about the Pandas library for data analysis. Pandas is part of the Anaconda distribution, and there are decent tutorials on most aspects of Pandas; I would recommend the following:


There are three problems, the first an extended tutorial, and the last two actual problems. You should read through problem 0 carefully, perhaps with a partner, and try all the examples. Problems 1 and 2 are actual activities based on the material in problem 0, you should do these yourself to hand in with the homework.

In [63]:

    # Here are some imports which will be used in code that we write for CS 237
    # Jupyter notebook specific
    from IPython.display import Image
    from IPython.core.display import HTML
    from IPython.display import display_html
    from IPython.display import display
    from IPython.display import Math
    from IPython.display import Latex
    # Imports potentially used for this lab
    import numpy as np  # arrays and functions which operate on array
    from numpy import linspace, arange
    import matplotlib.pyplot as plt  # normal plotting
    import seaborn as sns  # Fancy plotting
    import pandas as pd  # Data input and manipulation
    from numpy.random import random, randint, uniform, choice, shuffle
    from collections import Counter
    %matplotlib inline

Problem One

Download the following file of student data: studentdata.csv and write a function Prob01() which does the following:

- Print out the mean GPA for men;
- Print out the mean GPA for seniors (U4);
- Print out the 10 individuals with the largest number of credits earned, sorted in descending order by GPA;
- Display a histogram of the GPA of all individuals, with bins for each letter grade, i.e., 0.0, 1.0, 2.0, 2.33, 2.67, 3.0, 3.33, 3.67, and 4.0.
Problem One

(1) GPA of Men is 3.1148

(2) GPA for Seniors is 3.196

(3)

<table>
<thead>
<tr>
<th>Gender</th>
<th>ClassYear</th>
<th>GPA</th>
<th>CreditsEarned</th>
<th>TransferCredits</th>
<th>APCredits</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>U4</td>
<td>3.65</td>
<td>159.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>3.63</td>
<td>155.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>3.37</td>
<td>173.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F</td>
<td>U4</td>
<td>3.32</td>
<td>154.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>3.12</td>
<td>167.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>3.10</td>
<td>152.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F</td>
<td>U4</td>
<td>2.89</td>
<td>152.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>2.87</td>
<td>152.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M</td>
<td>U4</td>
<td>2.64</td>
<td>162.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F</td>
<td>U4</td>
<td>2.06</td>
<td>156.0</td>
<td>4.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

(4)

Problem Two

Download the following file of heights and weights of 25,000 individuals: biometricdata.csv and write a function Prob02() which does the following:

- Print out the maximum, minimum, mean, and (unbiased) standard deviation (all functions listed above) for the heights of all individuals;
- Print out the mean height for all individuals weighing more than 130 pounds;
- Print out how many individuals have a height >= 65 inches and <= 70 inches;
- Display a histogram of the heights of all individuals from the minimum to the maximum, where each bin represents 1 inch.

When doing your histograms, change the figure size so that they are large enough to see:

```python
import matplotlib.pyplot as plt
.....
plt.figure(figsize=(10,5)) # this will change size of figure when you call hist(....)
Students.hist(.....)
```

Play around with the parameters 10 and 5 until you like the result.

Also, I found that when I called hist(...) several times, Pandas would add the new histogram to the old one (in weird colors). To keep them separate, you need to show a histogram and then close it; after a class to hist(...), add these two lines:
Problem 2

(1) Max of Heights: 75.1528
    Min of Heights: 60.2784
    Mean of Heights: 67.9931
    Std of Heights: 1.9017

(2) Mean height with weight > 130: 75.1528

(3) How many with 65 <= height <= 70: 19846

# Your solution

```python
Bio = pd.read_csv('biometricdata.csv')
def Prob2():
    pass
Prob2()
```

In [89]:

```python
Students['GPA'].hist(.....)
plt.show()
plt.close()
```

In [ ]: