Math221_ExcelFiles_ForLoops

March 6, 2018

Math 221 - Applied Statistics (Data Analysis)
Dr. Kamal Dingle (Spring 2018)
Reading Excel Files, for loops, and removing mistaken values
Reading text files with numpy is easy, but often data sets are given as Excel sheets. To read Excel files we will use **pandas**, a powerful set of data analysis tools

```
In [1]: import numpy as np
        import matplotlib.pyplot as plt
        import pandas as pd
```

Reading in the data file with pandas is easy

```
In [2]: data_file = pd.ExcelFile('Sample_Excel_Data.xls')
        data_file = data_file.parse()
```

Let's print the first 3 rows of the data to see what it looks like

```
In [3]: print ('First 3 rows are\n',data_file[:3])
```

```
First 3 rows are
     Age  Weight (kg)  Income (KWD) Smoker
0  20.0           69           681      Y
1  19.5           70           731      Y
2  18.5           70           936      N
```

So we see that the data file contains information about people: age, weight, salary, and whether or not they smoke. The 0,1,2,.. on the left is just the row number, it is not part of the data.
We could study these data using the tools of pandas, but to keep things simple, we will convert the data to a numpy matrix

```
In [4]: X = data_file.values # a numpy matrix. the column headings are automatically discarded

In [5]: print ('First 3 rows of X are\n',X[:3]) # print the first three rows of X
```

```
First 3 rows of X are
 [[20.0 69 681 'Y']
 [19.5 70 731 'Y']
 [18.5 70 936 'N']]
```

Before we study these data, we want to look for any obvious mistaken values, which don't make sense. We will use a for loop for this to find the largest and smallest values.

```
In [6]: for col in range(4):# there are 4 columns in X
            print ('\nColumn number=',col)
            print (np.min(X[:,col]))# smallest value in the column
            print (np.max(X[:,col])) # largest value in the column
```

```
Column number= 0
10.0
67.0

Column number= 1
42
112

Column number= 2
472
1072

Column number= 3
N
Y
```

It looks like there are no obvious mistaken values. If there were, we would have to remove the samples with mistakes.

Now let's how many people smoke, and how many don't smoke. We use a for loop again.

```
In [7]: Smoker_counter = 0
        Nonsmoker_counter = 0
        for j in range(len(X)):
            # Y/N for smoking is in column 3
            if X[j,3]=='Y': # they smoke
                Smoker_counter = Smoker_counter +1
            elif X[j,3]=='N':
                Nonsmoker_counter = Nonsmoker_counter +1

        print ('The number of smokers is',Smoker_counter)
        print ('The number of nonsmokers is',Nonsmoker_counter)
```

```
The number of smokers is 10
The number of nonsmokers is 17
```