To get familiar with Pandas, first, I looked through the “10 Minutes to pandas” to have a general idea about it. And to learn how to manipulate with Pandas, I take classes on Lynda.com. I think it is kind of like SPSS that I use in my undergraduate. They both process data frame, but Pandas has less clicking and more typing. I will talk about what I have learnt about Pandas in the sections below.

1. Data Input and Validation
   First of all, we should import pandas. It is usually renamed as pd as a naming convention. And just like what we do with text files, we read csv files by pd.read_csv(). In the parentheses, we put the file rout so python can find it. Next, we should have a general idea about the file we work on. Shape method can help us to get a tuple in which the first element is the number of row, and the second is the number of column. To have a intuitive idea about what our csv file looks like, we use head() or tail() to see the first n rows in the csv with the headers. If we don’t set up the number n, it will be five by default. And to have comprehensive information, we use info method. This method will tell us the number of entries, the data type and the number of non-null entries for each series in the data frame.

2. Indexing
   What we will get from set_index is a data frame with changed label which is used to be the series. By doing this, I think I can solve the second question that I raised in task 2. And there is one thing that worth mentioning. If you want to change the label of the original file, you should set the inplace parameter as ture. Otherwise the original file won’t change. And resset_index is to return a data frame to its default. In addition, sort_index helps to sort data by the index. This also needs to assign inplace as ture. Further, if we want data shown as descending order, we can set ascending as false.
   To explore data of a certain value of a series, set_index and loc will be helpful. First, set the series that you want to find value in as index. And use loc to find the data of certain value. Moreover, with iloc, we can do traditional Python slicing. This one is integer based.

3. Some Basic Analysis
   By using value_counts, we can find the total number of unique values of a series. It is shown by descending order of the count number by default. For example, if there is a gender series, we can find out how many men and how many women are there by using value_counts. While sort_values works much like the method sort_index. It sorts the value of series and order them by ascending order by default. However, the result will not be saved into the original file. So, if we want to capture it. We should assign it into a new variable. And of course, we can sort multiple series together. We use by = [] to enter the series we want into a list after the “by”.

Next, I would like to talk about the code that I wrote.

For task one, the hardest thing is to think about how to let user choose whether to continue to next step. So, I use nested while loops, the out one is to control whether to select data from another file, the inside one is to control whether to select another attribute from the file. And I also set a gate keeper—a while loop to check if user if user input the correct rout of file.

Task two related to two kinds of data—one is baseball data and one is softball data. After manipulate data from each part, I compare two kinds of data and select the specific data I want.

Task three needs more Pandas knowledge than the former two. I am confused at the beginning about choosing right function. After getting Madam's help, I managed to finish the code. I learn two more function: set_value and df.to_csv.