**###PYTHON TRAINING – working with Pandas, DataFrame(s), and baseball examples**

'''

**PART 0**

**Download all relevant files into a single folder. REVIEW THE FILES BEFORE YOU BEGIN THE FOLLOWING EXERCISES**

**Set the working directory in the IDE of your choice.**

# For details on how to set the working directory, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/

'''

'''

**PART 1 – Dallas Keuchel**

The 2015 pitching logs for Dallas Keuchel have been mishandled! The logs are now split into 2 separate files and must be recombined.

There is missing data in both files and some incorrect data:

The only complete and correct column in both files for Dallas Keuchel is the column "Gtm".

For the first file (incomplete1), the "Opp" column has some incorrect entries, specifically rows index 3 to 6 (Excel rows 5-8). There is no way Dallas Keuchel would have faced the Montreal Expos (MTL) in 2015! (The Expos no longer exist, among other reasons)

All other data in the first file (incomplete1) is correct.

The second file (incomplete2) tried to correct this mistake, but is missing other entries in the "Opp" column.

**Read the relevant .csv files into Python as dataframes.**

**Merge the two files into one complete and correct dataframe, with sensible/original column names and no unnecessary/duplicate columns, using the information provided above.**

# For details on how to read files into Python, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/

# For details on how to merge dataframes and which method to use, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/Merging+Data

# For details on how to rename columns and delete columns in dataframes, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/

'''

'''

**PART 2 – Scott Kazmir**

John and Jane were the ones tracking pitching for Scott Kazmir in 2015. At the end of the season, when they compared their files, they found that their files had different data! They agree that only data that appear in both of their files is correct.

**Read the relevant .csv files into Python as dataframes.**

**Create a dataframe of 2015 Scott Kazmir pitching data that is correct by merging John and Jane's files.**

# For details on how to merge dataframes and which method to use, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/Merging+Data

'''

'''

**PART 3 – The other Astros starters and a readable master file**

Import the data for other Astros starting pitchers: Feldman, Hernandez, McCullers, McHugh. All of these files contain complete and correct data.

**Read the relevant .csv files into Python as dataframes.**

For readability, the organization wants a master file that stacks all of the correct pitcher data for their starting pitchers.

**Create a dataframe that stacks all of the correct pitcher data for their starting pitchers.**

**Reset the index of the stacked dataframe so there are no duplicates in the index column**

**Sort the stacked dataframe by "Gtm"**

Currently, the readability of this dataframe is still poor. The Astros analysts and scouts want the following changes made:

**For the 'Home/Away' column, fill in blank cells with "Home" and change cells with "@" to "Away"**

**For the 'Dec' column, fill in blank cells with "No Decision"**

**Reorder the columns of the stacked dataframe so that the order of the columns matches that of “2015 McHugh, Collin.csv”**

# For details on how to stack dataframes and which method to use, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/Stacking+Data

'''

'''

**PART 4 – Filtering through the stacked dataframe**

At least one of the Houston starting pitchers in 2015 were acquired via trade mid-season.

**Team scouts and data analysts want an Excel file of pitching logs of just the games these starting pitchers played for the Houston Astros.**

**They want the name of this file to be "2015 Houston Starts.xlsx"**

**As for the games played by the starting pitcher(s) for other team(s), the analysts want it in a .csv file named "2015 Non Houston Starts.csv" without an index column.**

Both of these files should have columns in the same order as the stacked dataframe

# For more details, refer to the IPG MB wiki site here: https://wiki.mbww.com/display/INA/

'''

'''

**PART 5 – Slightly more advanced stuff**

The scouts and analysts want to create (slightly) more advanced pitching stats from existing data so they can better evaluate the players’ performances.

**Create a new dataframe with calculations of : SO/BB, SO/9, P/Inning, PAB, WHIP for each pitcher from data in the stacked dataframe (not the filtered dataframes). For those not familiar with these baseball stat terms, here is a quick breakdown:**

**SO/BB: strikeouts per walk ratio**

**SO/9: strikeouts per 9 innings pitched**

**P/Inning: pitches per inning**

**PAB: pitches per at-bat ratio**

**WHIP: (BB + H) per inning pitched**

**The .groupby() and .sum() functions could be of great use here!!!**

**It may be advantageous to review the columns of the stacked/master dataframe created in PART 3**

Caution/Hint: The way baseball innings are recorded is quite terrible modular arithmetic. For example, “4.1” innings is actually 4 and 1/3 innings. “4.2” innings is actually 4 and 2/3 innings. Therefore, “4.1” + “4.2” innings is not “8.3” innings, but actually 9 innings. There might need to be a new column created in order to add innings correctly…

**This new data frame should have columns in this order and no other columns:**

**['Pitcher', 'IP', 'H', 'SO', 'BB', 'WHIP', 'SO/BB', 'SO/9', 'H/9', 'P/Inning', 'PAB']**

**Export this new dataframe as a .csv file named "2015 Houston Astros extra stats.csv" without an index column**

'''