Heads up poker analysis
Part 1: Data Wrangling

I bought 500,000 hands that were played on 888poker.com between 2018 and 2020. The games are low stakes at $0.01/0.02 per big blind. The data came in a very messy format: it consists of 28,972 text files that are located in 587 folders and each file has between 1 and 40 hands in it.

Here is an example of a hand:

#Game No : 1180858642
***** 888poker Hand History for Game 1180858642 *****
$0.01/$0.02 Blinds No Limit Holdem - *** 03 06 2018 18:51:45
Table Berkeley 9 Max (Real Money)
Seat 4 is the button
Total number of players : 2
Seat 4: turron3618 ($2.47)
Seat 5: trebs15 ($0.72)
turron3618 posts small blind [$0.01]
trebs15 posts big blind [$0.02]
** Dealing down cards **
turron3618 calls [$0.01]
trebs15 checks
** Dealing flop ** [As, Jh, 6h]
trebs15 checks
turron3618 bets [$0.04]
trebs15 folds
** Summary **
turron3618 collected [$0.04]

To analyze this data, I need to have it in an accessible format and I decided to transfer it into a Pandas dataframe. The difficulty here is, of course, reading the file correctly. Here is how I solved the problem:
I looped over a list of files names and directories and for each of them, I read the contents into an array. Within this array, I searched for lines that indicate a new hand and created an array of hands. Each hand is an array of lines where is line is an array of characters. I looped over the entire structure and used sliced to get the information I needed like bet sizes, flop cards, and hole cards. For each street (preflop, flop, turn, river) I recorded all actions (bet, raise, check, fold, etc) as an array of tuple where the tuple has 2 entries: the action and the corresponding dollar amount.

I ended up with a dataframe like this:

Game ID	Player SB	Player BB	Preflop actions	Flop actions	Turn actions	River actions	Flop	Turn	River	SB stack	BB stack	SB cards	BB cards
502874582	Oracool1	sirstadiljus	[(f, 0)]	[]	[]	[]	[]	[]	[]	1.22	2.01	[]	[]
603604223	ArcticWin	Ceaban	[(call, 0.01), (check, 0)]	[(check, 0), (check, 0)]	[(check, 0), (check, 0)]	[(bets, 0.04), (call, 0.04)]	[5c, 9c, 5s]	7h	2c	2	1	[Qh, Tc]	[6h, 2h]
603604231	Ceaban	ArcticWin	[(f, 0)]	[]	[]	[]	[]	[]	[]	1.94	1.06	[]	[]
603604236	ArcticWin	Ceaban	[(f, 0)]	[]	[]	[]	[]	[]	[]	1.95	1.05	[]	[]
603604241	Ceaban	ArcticWin	[(f, 0)]	[]	[]	[]	[]	[]	[]	1.94	1.06	[]	[]

Here Game ID is a unique game identifier and stack column show players' stack sizes which are important for analyzing gameplay.