Zihao Li, Ye Zhang, Brody Vogel, and Jinghao Yan

Project Assignment #1

**Write-up.docx** - contains project write-up answering all questions for Project part 1.

**FinanceScraper.py** is the script for getting the stocks data from Yahoo! Finance. DO NOT RUN. This script has a runtime of ~40 minutes and produces the file STOCKS\_RAW.csv.

**STOCKS\_RAW.csv** is the raw finance data as scraped from Yahoo! Finance by FinanceScraper.py

**Sports\_Scraper.py** is the script for getting sports results from various reference sites. DO NOT RUN. This script has a runtime of ~2 hours, 45 minutes and produces the file Sports\_Consolidated\_RAW.csv.

**Sports\_Consolidated\_RAW.csv** is the raw sports results data as scraped from baseballreference, basketballreference, and pro-footballreference.

**Group4Cleaner.py** reads in STOCKS\_RAW.csv and Sports\_Consolidated\_RAW.csv, cleans them, prints the cleaning metrics, and then produces our two final datasets: SPORTS\_FINAL.csv and STOCKS\_FINAL.csv. This script should run in a few minutes, at the most.

###NOTE###

SPORTS\_FINAL.csv, as turned in, contains the 4 created variables. We’ve commented out the part of Group4Cleaner.py that creates those variables, though, because it adds a few minutes to the runtime. If you choose to un-comment that section, the script should run in 5-7 minutes.