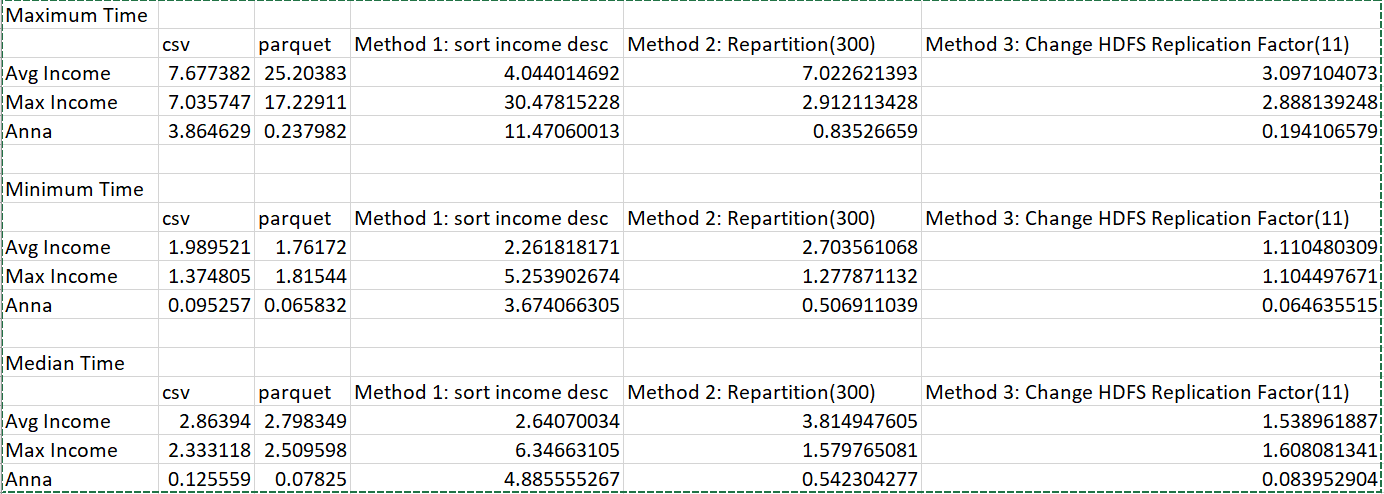
DS-1004 Big Data Lab4 Report

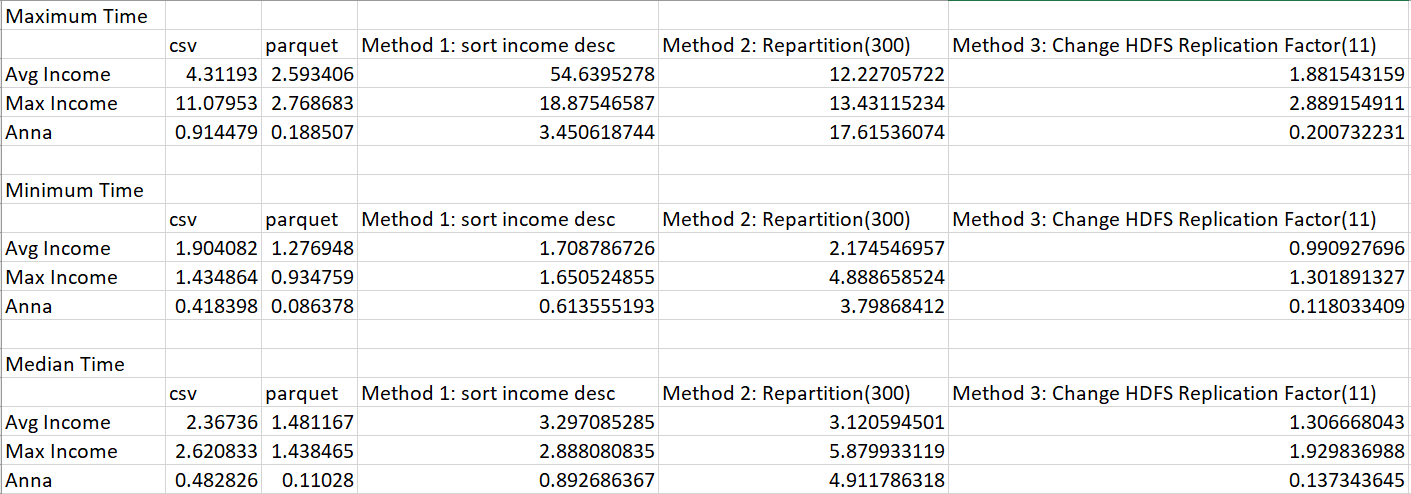
Frank Jiang

Results for part 1, 2 and 3

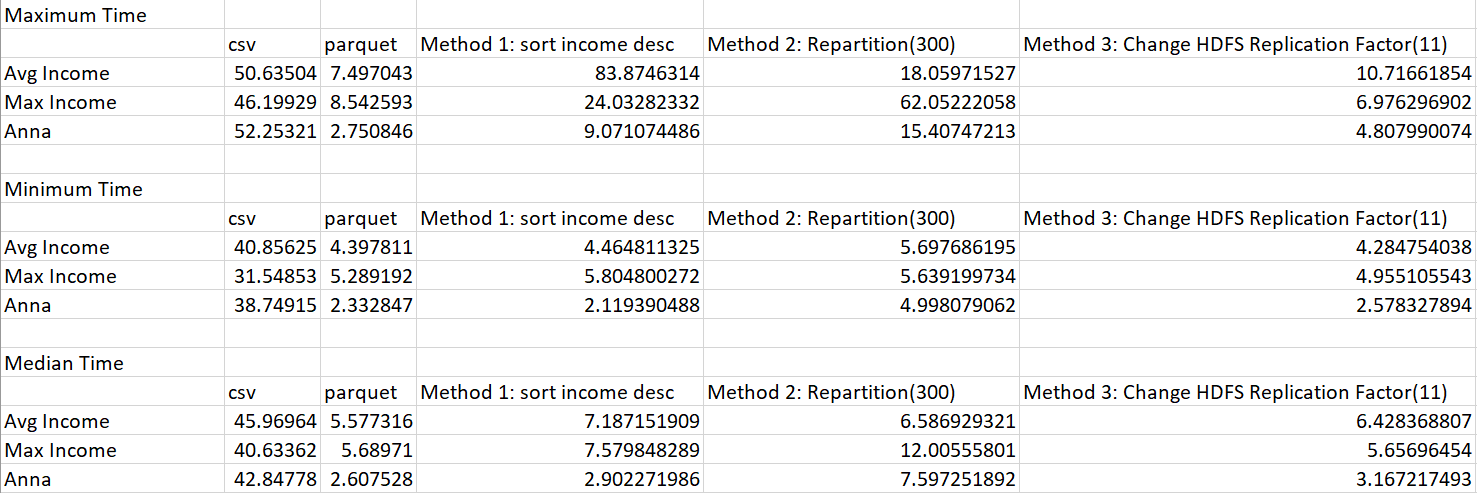
For small size people dataset with 5000 records



For medium size people dataset with 500,000 records



For Large size people dataset with 50,000,000 records



From the results table, we can find out for all three size of people dataset, save it as a parquet file improve the computation speed significantly. The bigger size the datafile is, the more efficient it is to use parquet file instead of csv file. For part 3, different method for optimizing the parquet file seems to have a different effect on different queries and different size of datafile.

For part 3, I use to 3 different methods try to optimize parquet. Method 1 is to sort the income column in descending order and then save it as a parquet file. Method 2 is to change the partition size of the data frame to 300. Method 3 is to change the HDFS replication factor of the entire hj1325 folder, which I stored all the parquet file from a default of 3 to 11.

Since, we run the queries 25 times on each parquet file, the median time seems to be more representative of how much time a single run will take.

For small size people datafile, method 1 and 3 seems to shorten the processing time on the average income query. Method 2 and 3 have a great effect on the maximum income query. None of the 3 methods have a great impact on the anna query.

For medium size people datafile, method 3 have some impact on the average income query. Similarly, none of the configurations works on the anna query.

For large size people datafile, only method 3 shorten the processing time on the maximum income query.

Overall, I believe increasing the HDFS replication factor might be the better solution to optimize parquet and shorten processing time.