**Programming for Big Data**

**Assignment 3**

**D10126532 John Warde DT230B**

**NOTE: Must be exported/saved to a PDF before submission!!!**

# 1. Tax Underpayment

Blah

## Part A. Suitable Analyses

Blah

## Part B. Data Storage

Blah

# 2. Stock Performance

For the non-parallel solution, I coded the functionality into self-contained re-usable functions as much as possible.

* Chose to read in the numeric data and apply the stock names using factors – allows more flexibility if stocks are added/removed.

For parallel solution,

* In the getAveragesPerStock() function, I filtered the stockData on the current stock code within the parallelised foreach/%dopar% loop because the cluster is on the same machine. If a cluster of physical machine was available to me I would consider taking the following line out of the   
    
  dfForStock <- dfStock[dfStock$stock==stockNamesAsLevels[i],]  
    
  to possibly allow the transfer of a small amount of data to the other machines, however I would benchmark the function before and after to get a clear picture of performance.

# 3. R and Hadoop

Blah