1. Return information is in “Returns”, Ret\_Var.xlsx, which includes monthly excess returns for 18,719 stocks ever existed from 196307 to 201912 (let’s drop 202001). Note many returns are missing because the stock does not exist at a particular point in time. The data format is a 18,719\*678 matrix.

Also in “Returns” we have other variables including market capitalization (ME\_Var), industry classification (SIC\_Var), etc., that we may use later. For example, we may calculate value weighted returns based on ME\_Var. These variables are not predictive signals. They are just additional stock characteristics that help create portfolios.

2. “HighTunrover”, “LowTurnover”, “MidTurnover”, “ChenCRSP”, and “GHZ” include five sets of predictive signals, organized based on their economic interpretations. Each variable is arranged in the same 18719\*678 format. Note each variable has a range of 0 to 100 at each point in time, corresponding to the ranking of the variable. We can probably standardize each variable to -1 to 1 (at each point in time) to facilitate comparison with the existing literature.

In terms of variable timing, return for a given month, e.g., 196307, is realized during this month. Whereas for a predictive signal for the same month, e.g., BtM for 196307, its value is the beginning of the month value (or, equivalently, the end of the previous month value). As such, the predictive signal for a given month can be used to forecast returns realized over the same month. There is no look-ahead bias.

Same as returns, there are many missing variables. If a stock has non-missing data on Size (in “LowTurnover”), BtM (in “LowTurnover”), GPA (in “LowTurnover, set the missing values for other variables to be the within-industry average, where industry classification is based on the two-digit SIC provided in SIC\_Var (in “Returns”).