**Predictive Analytics Project 2**

***Venkatesh Subramaniam, Ravi Teja Ayyagari, Abdulaziz Alsayegh***

**Objective**

The objective of the project is to predict a time series model using 15 dates and predicting the coefficients for the dates from 16 – 19. With the help of the coefficients that have been predicted from above phenomenon, we are supposed to calculate expected log returns by dividing the data into five buckets that have been sorted based on the expected log returns and check if the mean actual returns and mean of expected log returns are similar or vary drastically. This methodology is to be performed for both the entire universe of stocks as well as a specific chosen industry.

**Results**

For the entire universe, the mean expected log returns and the actual log returns as distributed as follow:

## Flag Mean\_Return Mean\_Target  
## 1 A -0.05214809 2.044093  
## 2 B -0.06560267 3.506867  
## 3 C -0.06822487 4.053031  
## 4 D -0.06623533 4.489411  
## 5 E -0.05464837 4.846144

So, if we have a look at the observations we can find that there is a deviation between the actual and the expected log returns since the observations are in logarithmic scale and not on the original value of the data.

For the health care industry, we can find the observations for the actual and the expected log returns as below:

## Flag Mean\_Return Mean\_Target  
## 1 A 0.005224148 -0.2243830  
## 2 B 0.024854527 0.8100563  
## 3 C 0.004697563 1.1372756  
## 4 D -0.015074969 1.4232331  
## 5 E 0.008391145 1.7689001

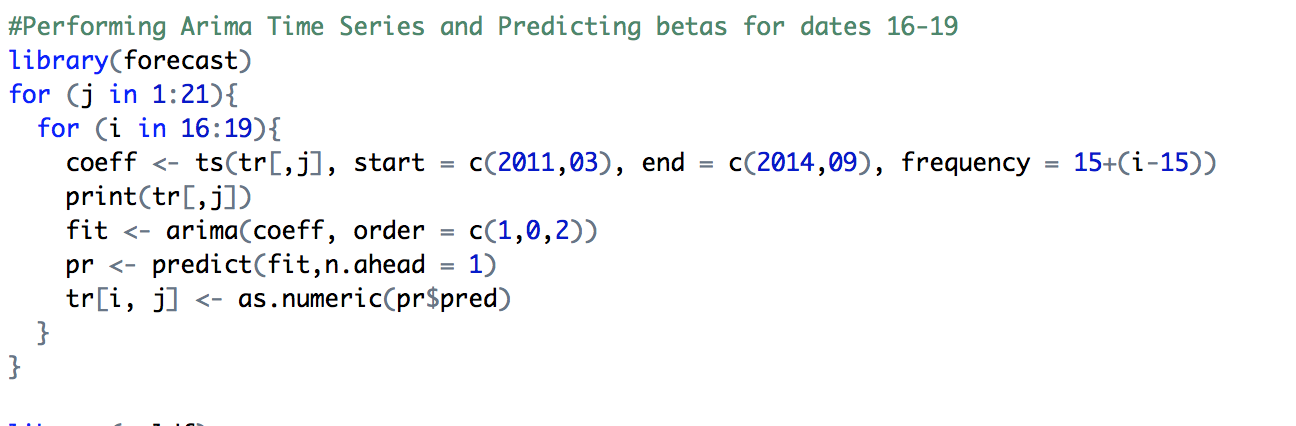
These observations show not much of a deviation as compared to the entire universe of the stocks, since the data is specific to the industry and hence the result.

**Methodology**

The basic process is almost the same for both the cases. Since we have already discussed about the regression, we shall discuss from the time series.

*Time Series*

Since we are interested in the dates from 16-19, the loop also consists of the same interval. The arima concept is used in the time series model and the predicted values for the betas for the 4 dates are calculated. We are ideally calculating 21 betas starting from b0 through b20 for the 4 dates

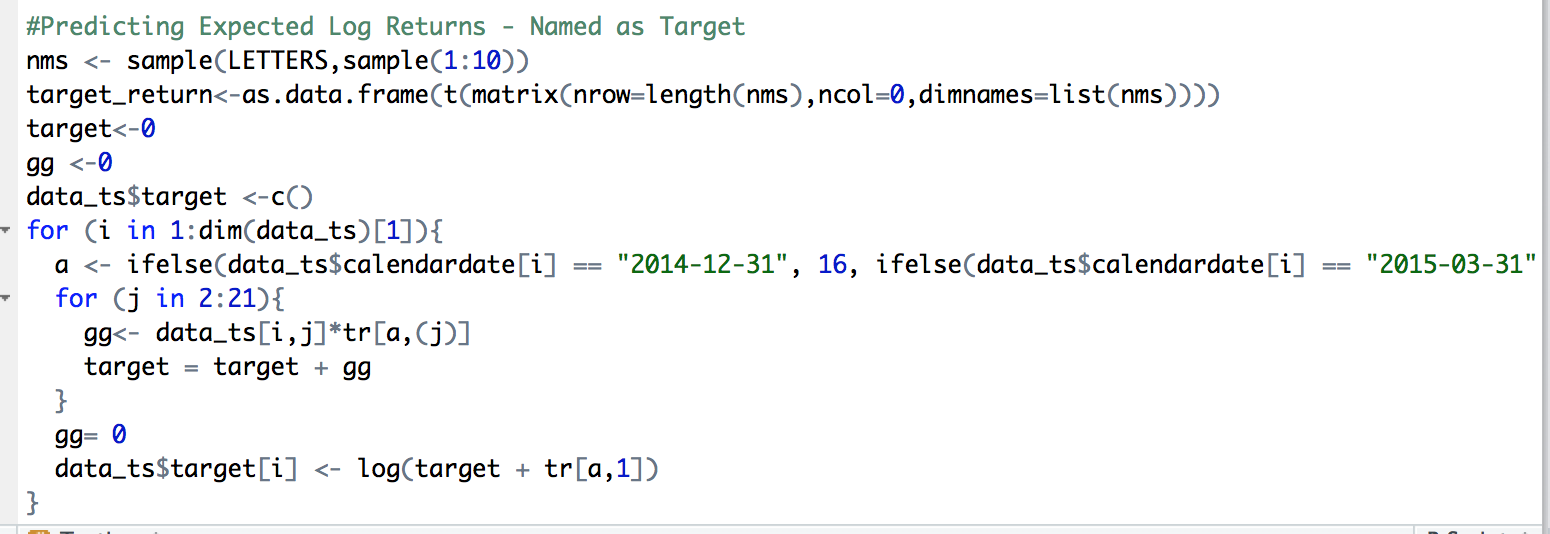


*Predicting the Expected Log Returns*

The prediction of log returns can be carried out for the 4 dates for which the betas have been calculated that were from the time series prediction. Ideally, we are completing the linear regression

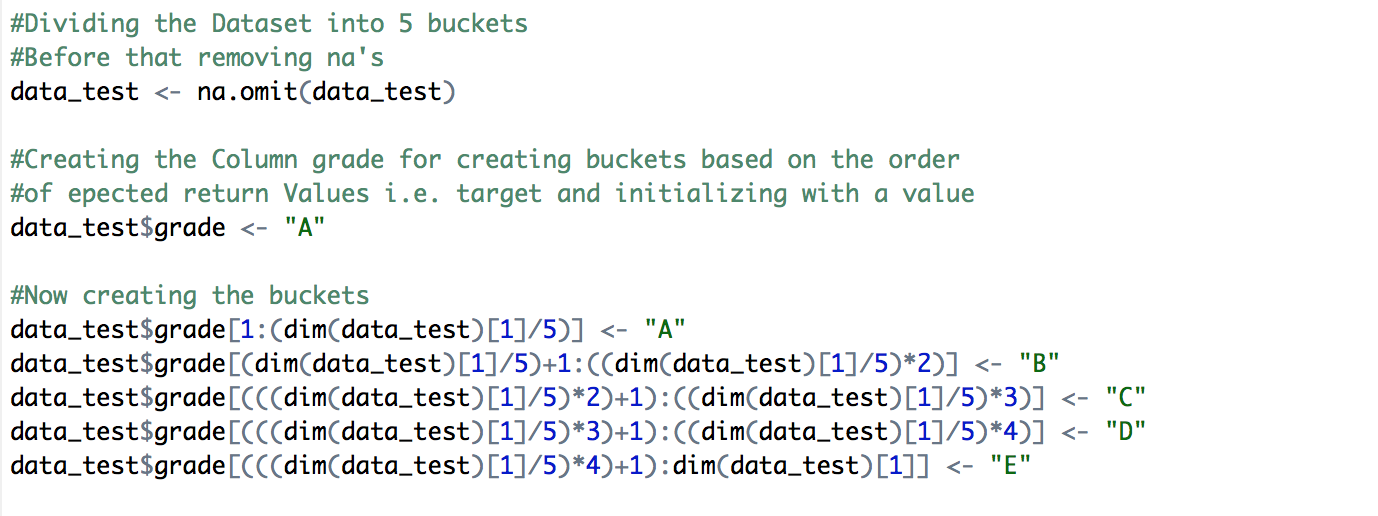
log(y) = b0 + b1\*ind1 +…..+ b20\*ind20

This process put in the for-loop will give the predicted log returns for the entire universe of stocks as well as the specific industry



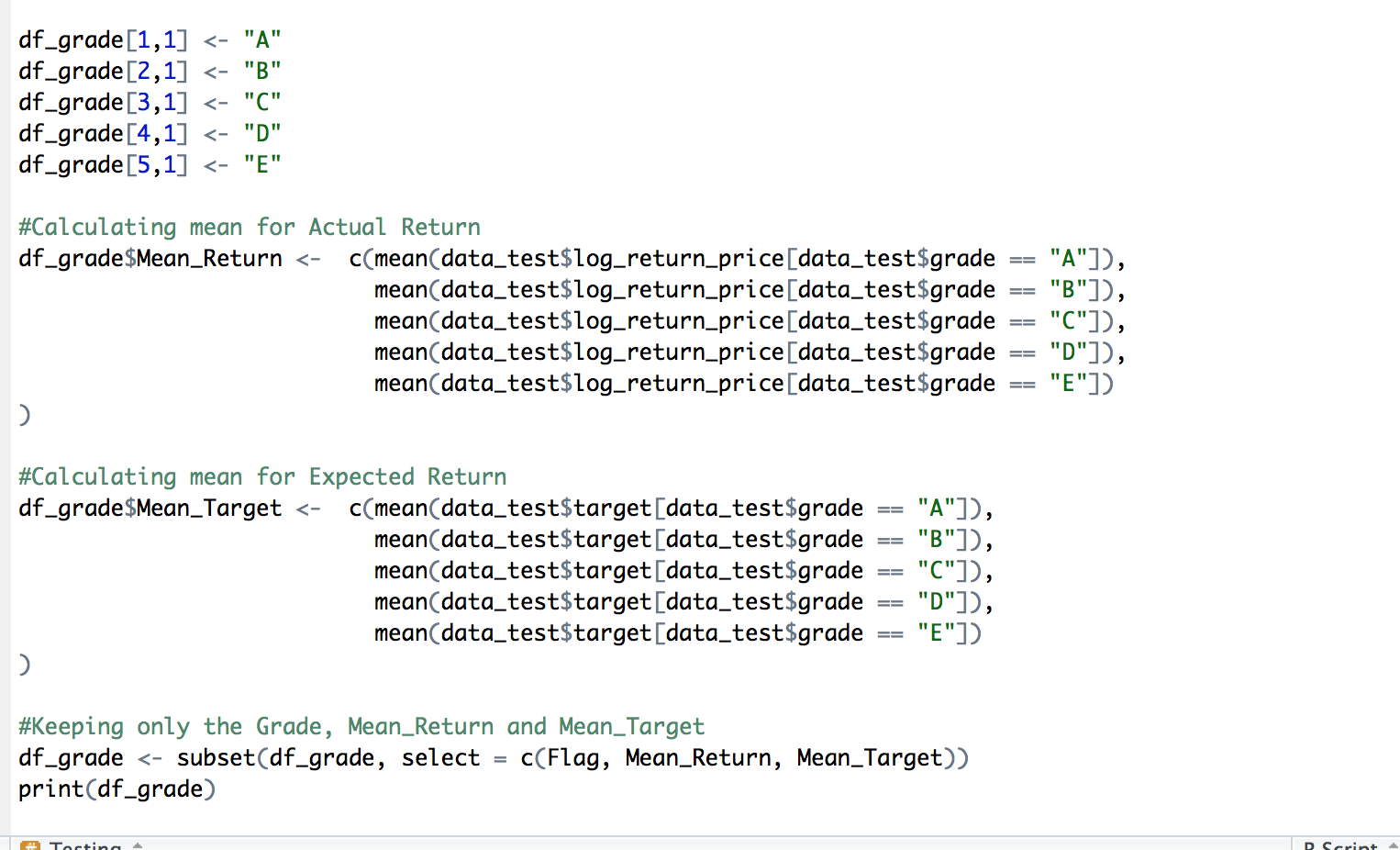
*Dividing the data into 5 buckets*

The process of dividing the data into buckets is as follows. The data is pre sorted according to the expected log returns – in this case target



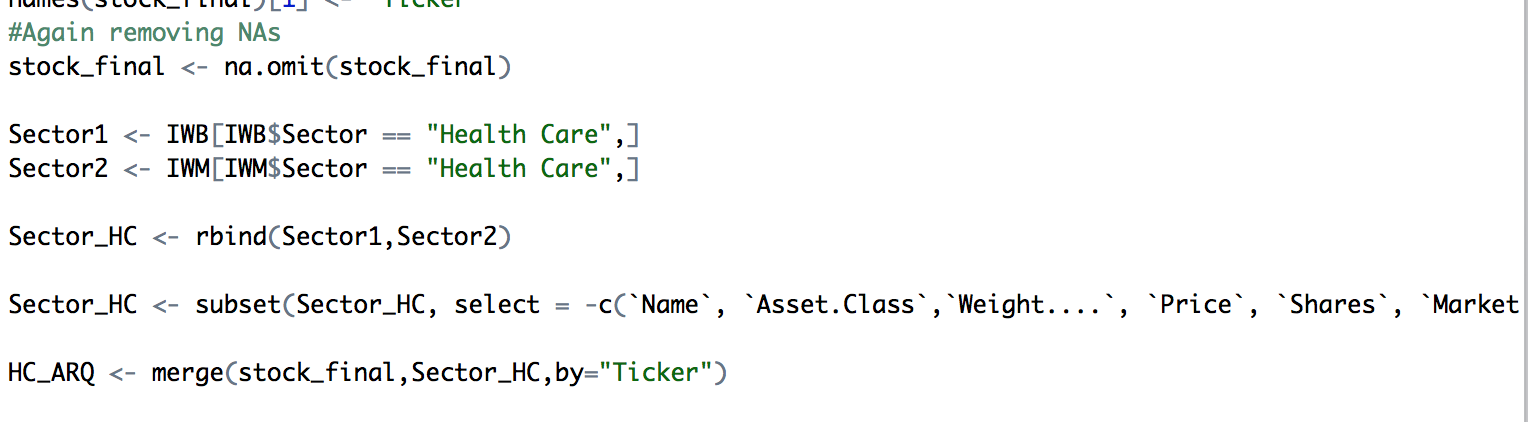
*Finding the mean of expected and actual log returns*

This is the final step of the data, where we calculate the 5 means for 5 grades of the dataset that were bucketed into previously.



*Specific Industry*

There is a slight modification in the process for the industry process, since the data needs to be merged and then the analysis needs to be performed.



**Conclusion**

There is a difference in the actual log returns and the expected log returns for both the universe as well as the industry cases. In case of the entire industry, there is a good deviation while the industry case does not show much of the deviation.