1. Data Processing

1.1 Join all datasets with Security Reference Data

Firstly, after went through the definition of different datasets, I decided to join all the data to the security reference data by date and security id and got a full dataset. Afterwards, I looked through dataset and focused on the missing value.

Column Name	Missing Rows	Missing Ratio
security_id	0	0
close_price	0	0
volume	0	0
group_id	0	0
in_trading_universe	0	0
ret1d	992	0.000140727
rf1	0	0
rf2	0	0
rf3	0	0
rf4	0	0
rf5	0	0
rf6	0	0
df1	6945605	0.985317565
df2	6950287	0.985981763
df3	6955558	0.986729517
df4	6945494	0.985301818
df5	6950210	0.985970839
df6	6955493	0.986720296
df7	6955271	0.986688803
df8	157035	0.022277303
df9	309019	0.04383806
df10	2924785	0.414915912
df11	2924785	0.414915912

From the table above, we can find that for dataset1 to dataset7, these datasets have over 98% of data is missing, and for dataset10 and dataset11, about 40% of the data is missing. Since I didn't have the meaning of the dataset1 – dataset11, it is not reasonable to add back the data by any common methods. In this way, I decided to drop these columns. After that, I only considered the six risk factors.

As for the column of 1-day return, I grouped by security id and checked the percentage change of the close price. I found that the 1-day return is matched with the calculation result of close price, then I added the missing value by calculating the close price change.

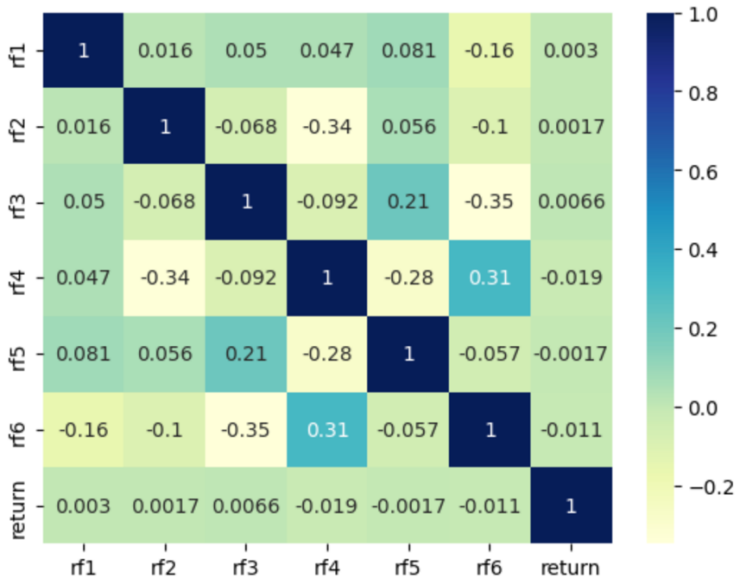
Moreover, considering that the portfolio only considers the stocks with "in_trading_universe = Y", I selected the stock with this condition to do further analysis.

Finally, the dataset is like below.

data_date	security_id	close_price	volume	group_id	in_trading_universe	rf1	rf2	rf3	rf4	rf5	rf6	return
2010-01-05	78001	24.23	496257.0	20101010	Y	0.532	-0.207	-0.187	0.808	0.640	0.303	0.019352
2010-01-05	78401	19.77	2881636.0	20104020	Y	0.768	-0.441	-0.504	-0.387	-0.050	0.070	0.006619
2010-01-05	85301	6.66	3458084.0	45201020	Y	1.143	-1.492	0.441	0.235	-1.038	-1.666	0.012158
2010-01-05	97801	48.95	3965232.0	40301020	Y	2.076	0.276	-0.459	-0.242	0.064	0.779	0.029010
2010-01-05	100701	36.27	546563.0	55102010	Y	-0.951	0.060	-1.274	0.435	-0.836	0.413	-0.019730

1.2 Data Analysis

I analyzed that relationship between the risk factors and return and calculated the correlation. The result of Pearson correlation is as below.



From the correlation result, I found that risk factor 4 and risk factor 6 have relatively high correlation with the return which means these two factors have greater prediction power.

Furthermore, I ran regression of return on each factor to check whether the coefficient is significant or not. Using OLS method, the percentage change of return is the dependent variable and the value of risk factor is the independent variable. Then, I ran six linear regressions on these six factors respectively.

2. Model building and Portfolio Construction

2.1 Stock Selection

After checking that all the risk factors remained are effective, I want to aggregate the effect of these factors. Therefore, I first normalize each risk factor and then added whole the risk factor together.

To be more specific, I used mean-max normalization for the data standardization to transform the data to one scale and then summed up all the factors. After that, I gained a new risk factor which includes all the information of risk exposure.

Based on the value of this new risk factor ("rf_all"), I ranked the stock according to the value of rf_all, and selected stocks according to the rank. To illustrate more, I chose the top 20 stocks to long and bottom 20 stocks to short. Plus, I decided to monthly rebalance my portfolio and I used the last month's data to select stocks. Hence, I calculated the mean value of the rf_all factor, and then picked the top and bottom stocks.

2.2 Portfolio Construction

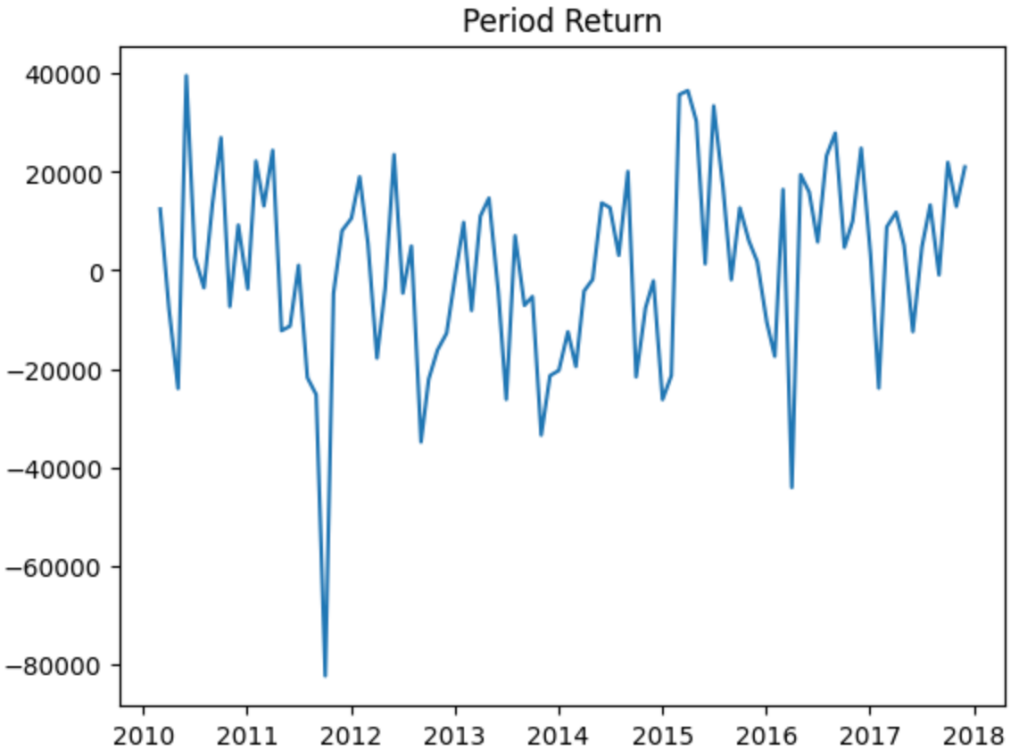
I decided to rebalance the portfolio monthly, so I chose new stocks at the first day each month based on the last month's data.

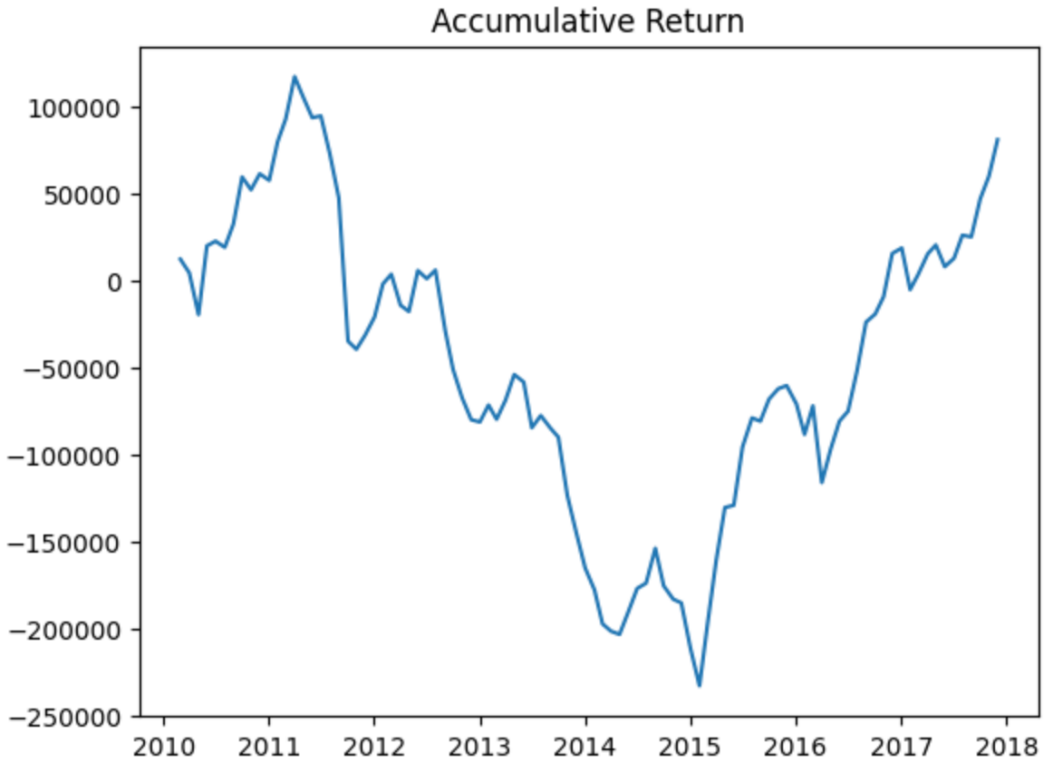
Due to the time limitation, I chose to equally weight each stock in my portfolio. Since the portfolio should be market neutral, I assumed each stocks allocated 10000$ (long position is 10000$, while short position is -10000$)

3. Result and Plot

From the result below, we can conclude that the portfolio average return is positive, and more than half of our period return is positive. But the volatility is relatively high, and the maximum drawdown ratio is high. This suggests that although the portfolio can earn a lot money, the risk is also high, and this problem need to be further deal with.

Volatility	Average Return	Shape Ratio	Maximum Draw down ratio (*100%)	% of positive return
56685.068	863.71	0.13	1.50	0.55




Accumulative Return
100000
50000
0
−50000
−100000
−150000
−200000
−250000
2010
2011
2012
2013
2014
2015
2016
2017
2018

4. Future Potential Improvement

4.1 Data Analysis Part

For the data part, I think dataset 8 and dataset 9 still have a lot of information which can be quite useful. In this way, I will try to analyze these two columns and find whether there is any correlation with other factors or return, and then I can deal with the missing value problem of these data.

In addition, I believe that the group_id is also useful. This is because sometimes, we can allocate the portfolio in different groups (stocks that are not in the similar industry), in order to enable the portfolio to be more diverse, which means that it can reduce some unsystematic risk. Besides, maybe during some specific period, the equities in one or two industries perform really great. Under this situation, by analyzing the relationship between group id and return may help a lot.

4.2 Portfolio Construction Part

For the portfolio construction part, I believe that by applying some other methods, like mean-variance, can help portfolio to maximize the return, as well as reduce risk. Plus, when I an selecting the stocks, I can add weight to different factors, which enable the more predictive factor to influence more during the process of choosing the stocks.