Algorithm Trading HW1

Q1(c)

Steps to make adjustment

Step1: Load S&P 500.xlsx file
Step2: Generate a Data Frame book judging if the price and size data needs adjustment for each ticker by comparing the average and initial values.
Step3: Rewrite data if needs adjustment or copy the original data directly to save time

Tickers needing adjustment

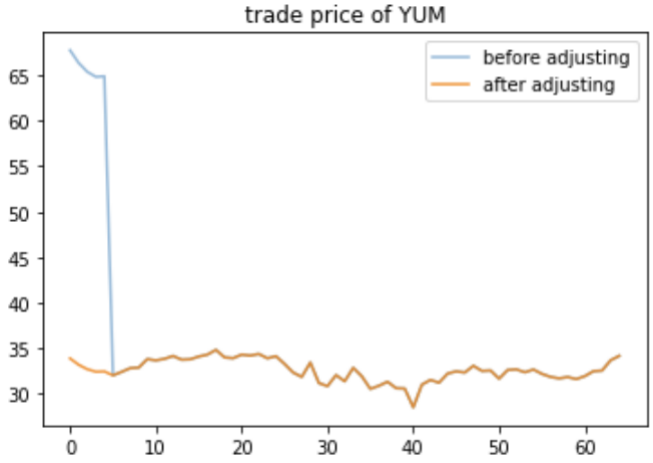
	ticker_id	law
100	TXT	False
146	OMC	False
194	TYC	False
219	MTW	False
230	LEN	False
298	MS	False
327	AGN	False
346	GILD	False
356	ESRX	False
388	ABC	False
419	KHD	False
425	YUM	False
444	NVDA	False
490	NE	False

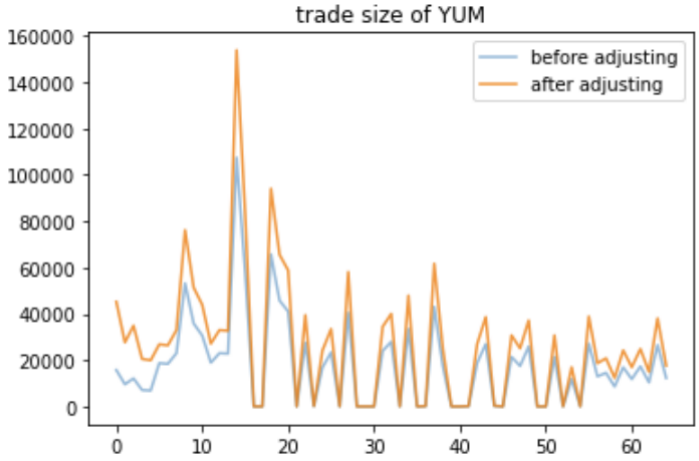
Testing Adjustment Result - YUM

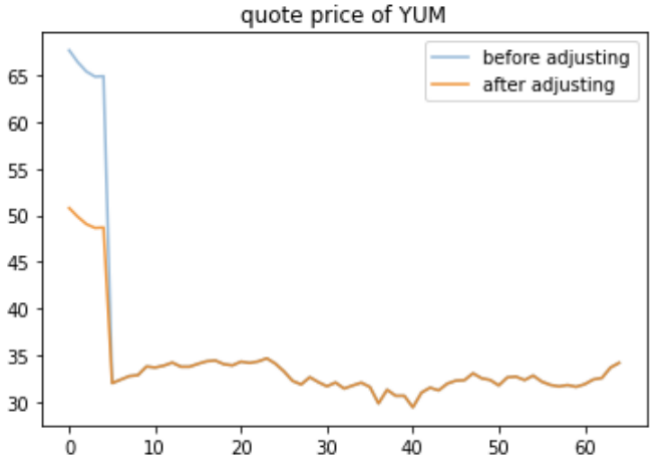
We can see that the adjustment factor does change in the 6th date.

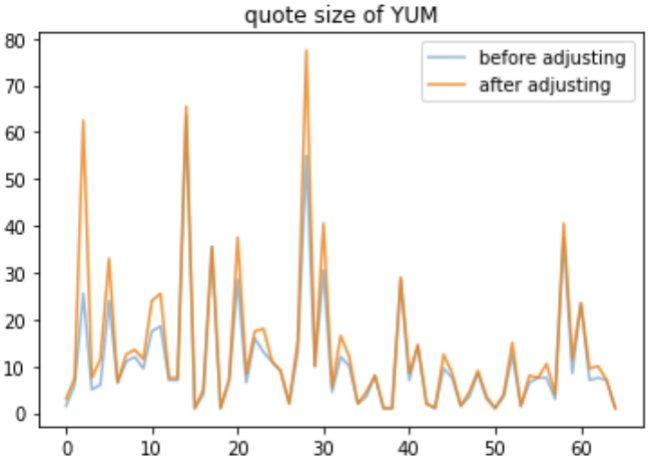
Cumulative Factor to Adjust Prices	Cumulative Factor to Adjust Shares/Vol
1	1
2.863075	2
2.863075	2
2.863075	2
2.863075	2
2.863075	2
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1
1.431537	1

For simplicity, we only load the data at the beginning of all dates to check if we implement successfully.









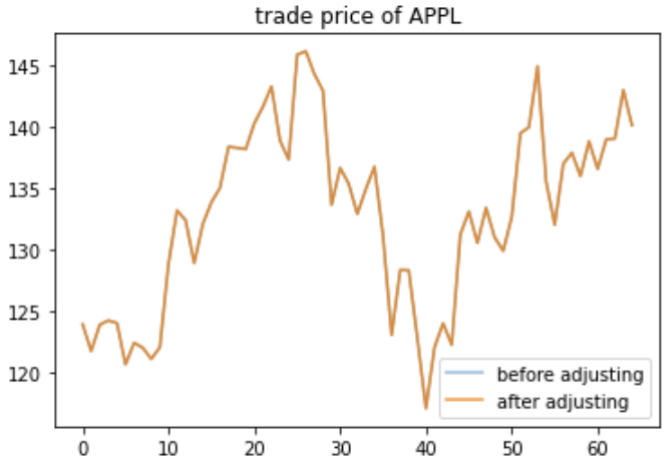
As we can see , the price and size successfully made adjustment after the 6th date as expected.

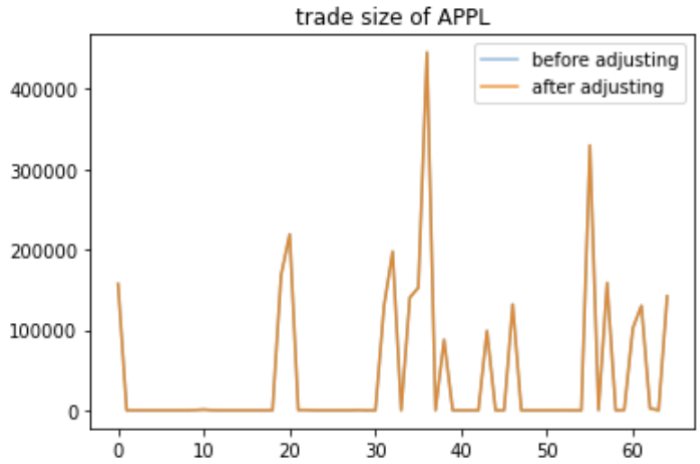
Testing Adjustment Result - AAPL

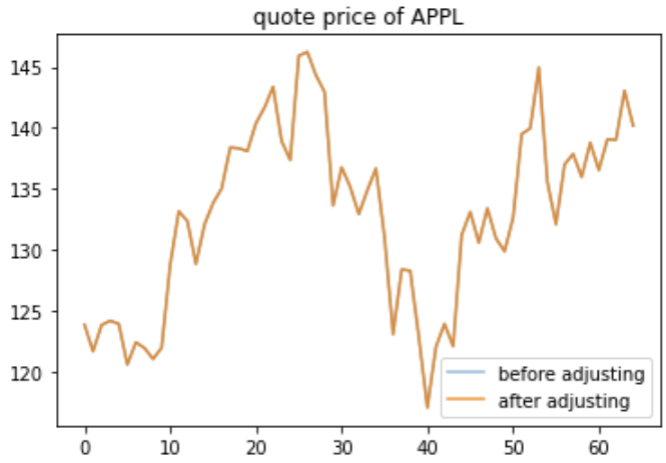
The adjustment factor of AAPL remain the same. So the expected adjusted data should be the same as original since we simply copy the data to save time.

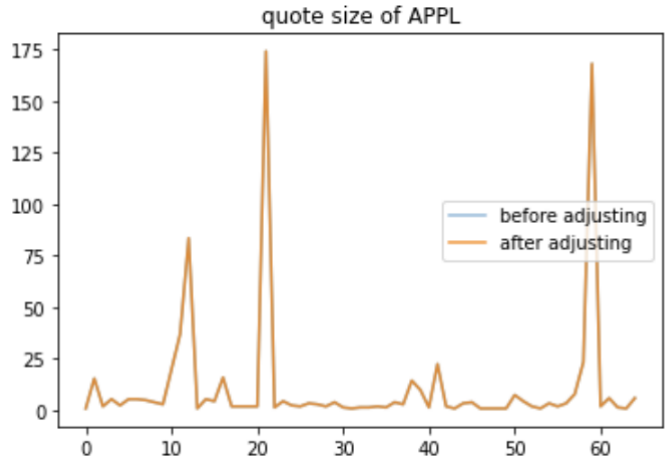
Cumulative Factor to Adjust Prices	Cumulative Factor to Adjust Shares/Vol
1	1
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7
7	7

The graph of data at the beginning of dates:









We can see that the result remains the same for AAPL as expected.

Conclusion

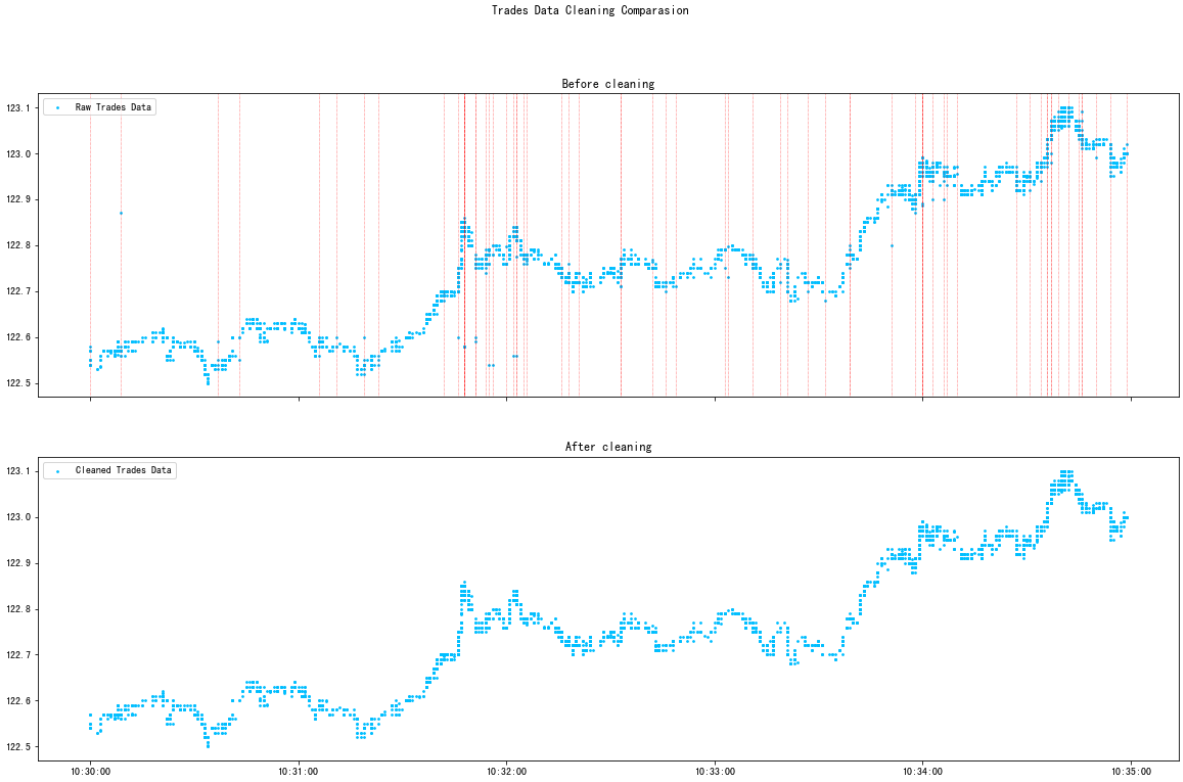
1. Two reasons we can omit the adjustment if it remain the same: The first is that we the scale of time series will not affect most of our analysis like trend of data (e.g. auto correlation, derivative of the graph) or relative comparison of statistics estimator between different time series. (e.g. rank of correlation, covariance). The second is that under most circumstance, we care about return instead of price which makes scale of price meaningless.
2. Among over 500 stocks, only 14 of them need adjustment over this time period. This means we can save much time only with a simple but looks negligible insight of adjustment. We should always look into whether we can save time by details before we implement it. (Actually tried without judgment about whether we need adjustment. we would spend more than 10 hours approximately if we did not make this judgement.)
3. The adjustment makes the trading price data comparable while there still seems some problem with quote price data.

Q1(d)

Cleaning result for AAPL Trades & Quotes

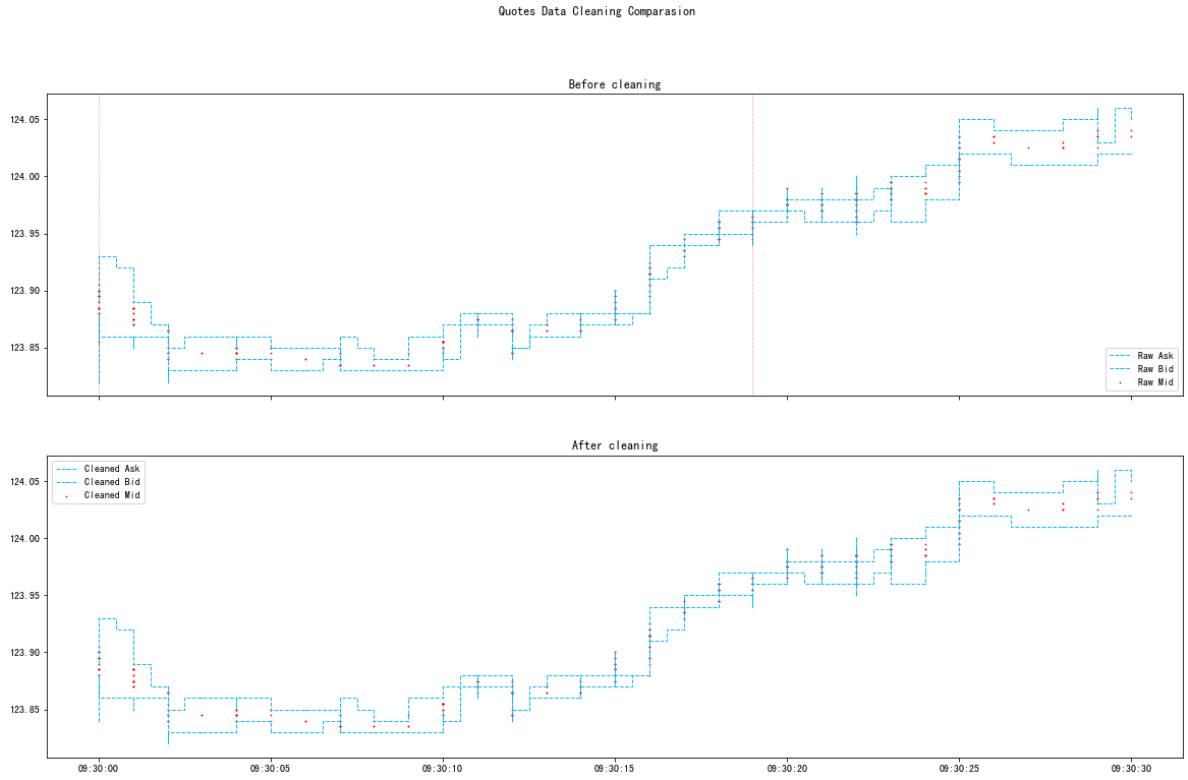
After cleaning the data using the suggested method, we test several parameters for filtering. We finally decide to clean the data with parameters $K = 21, \gamma = 0.00005$.

The cleaning result for Trades data of AAPL on 20070620 10:30:00 - 10:35:00 is:



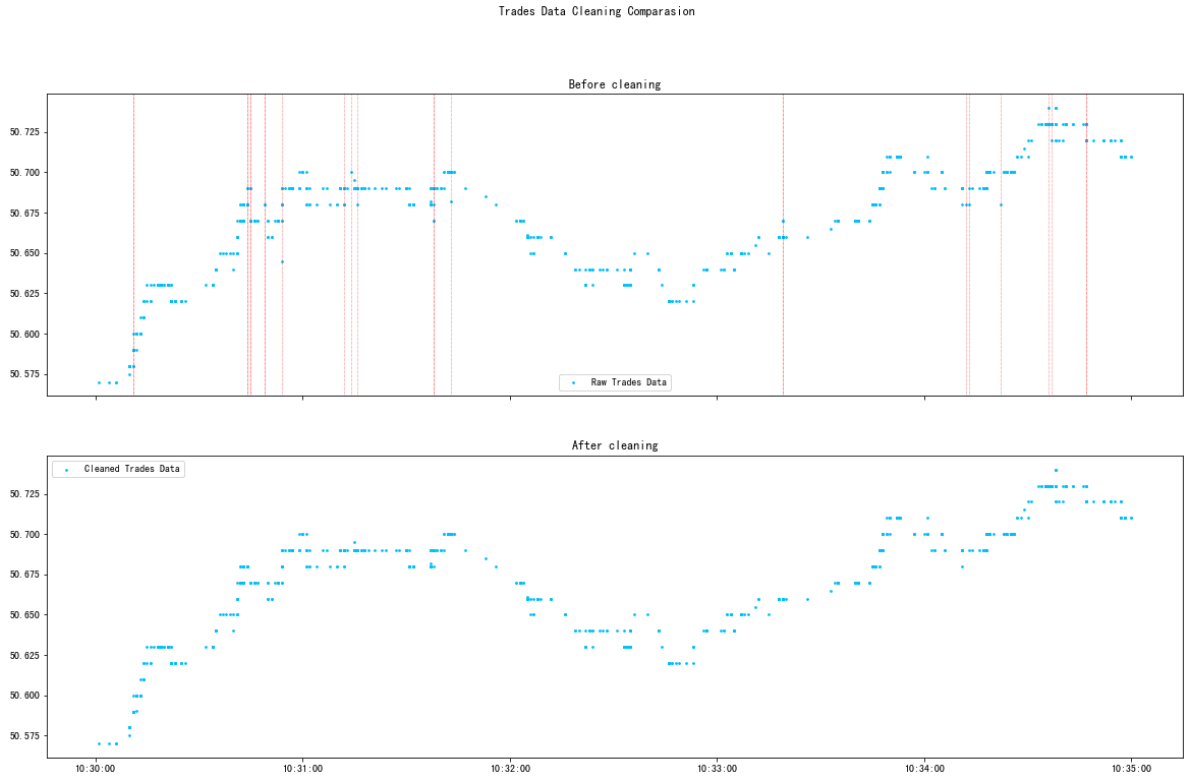
we can see that the trading on AAPL stock is very active, so that there are more outliers.

The cleaning result for Quotes data of AAPL on 20070620 9:30:00 - 9:30:30 is:

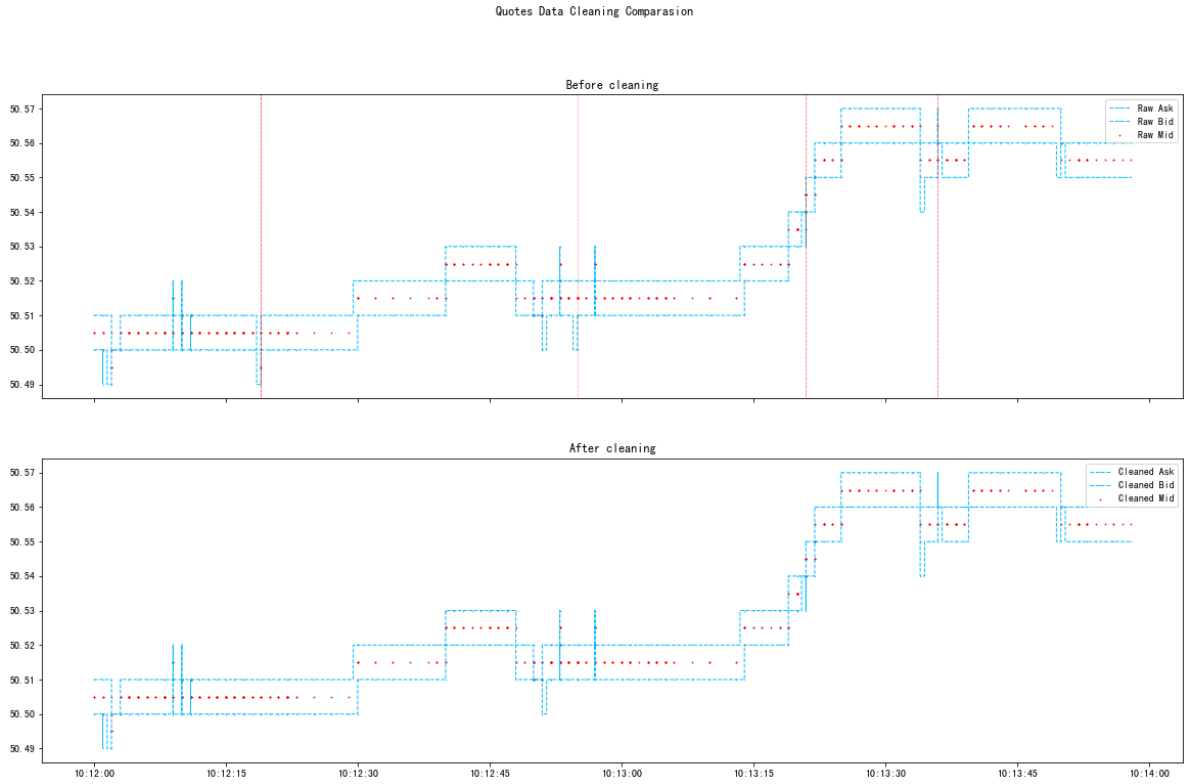


Cleaning result for JPM Trades & Quotes

The cleaning result for Trades data of JPM on 20070620 10:30:00 - 10:35:00 is:



The cleaning result for Quotes data of JPM on 20070620 10:12:00 - 10:14:00 is:



We can see that our cleaning procedure is effective.

Q2(c)

Combining the calculation for i, ii, iii together, the result for AAPL stock in our sample preiod is:

==========AAPL, Trade, frequency=10==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total trades	trades/quotes
Dirty		−0.570834	0	0.367713	181.697	−0.224994	112.699	0.0317498	65	11409653	0.457302
Clean	(5, 0.0005)	−0.573989	0	0.359039	181.182	−0.150375	9.42715	0.0124354	65	11408124	0.457241
Clean	(5, 0.0001)	−0.577384	0	0.357939	180.726	−0.161073	9.49811	0.0124354	65	11388411	0.456452
Clean	(5, 5e−05)	−0.575465	0	0.35781	180.649	−0.15679	9.50219	0.0124354	65	11345448	0.454745
Clean	(11, 0.0005)	−0.574269	0	0.358257	180.909	−0.155807	9.44365	0.0124354	65	11403503	0.457055
Clean	(11, 0.0001)	−0.575091	0	0.357378	180.514	−0.155697	9.4675	0.0124354	65	11341365	0.454578
Clean	(11, 5e−05)	−0.580007	0	0.357102	180.288	−0.151522	9.48014	0.0124354	65	11240846	0.450662
Clean	(21, 0.0005)	−0.574474	0	0.357685	180.717	−0.164419	9.36497	0.0124354	65	11400320	0.456928
Clean	(21, 0.0001)	−0.579428	0	0.357009	180.409	−0.163961	9.35003	0.0124354	65	11329837	0.454124
Clean	(21, 5e−05)	−0.578552	0	0.356458	179.9	−0.158929	9.3907	0.0124354	65	11223741	0.450068

==========AAPL, Quote, frequency=10==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total quotes	trades/quotes
Dirty		−0.604167	0	0.35095	175.078	−0.153494	9.78981	0.0131344	65	24949935	0.457302
Clean	(5, 0.0005)	−0.604167	0	0.35095	175.078	−0.153494	9.78981	0.0131344	65	24949935	0.457241
Clean	(5, 0.0001)	−0.604165	0	0.350956	175.082	−0.153507	9.78912	0.0131344	65	24949855	0.456452
Clean	(5, 5e−05)	−0.603017	0	0.350955	175.076	−0.153762	9.79127	0.0131344	65	24949025	0.454745
Clean	(11, 0.0005)	−0.604167	0	0.35095	175.078	−0.153494	9.78981	0.0131344	65	24949934	0.457055
Clean	(11, 0.0001)	−0.604168	0	0.350947	175.073	−0.153532	9.79028	0.0131344	65	24949203	0.454578
Clean	(11, 5e−05)	−0.604295	0	0.350935	175.062	−0.153768	9.78929	0.0131344	65	24942973	0.450662
Clean	(21, 0.0005)	−0.604167	0	0.35095	175.078	−0.153494	9.78981	0.0131344	65	24949932	0.456928
Clean	(21, 0.0001)	−0.607591	0	0.35098	175.072	−0.151365	9.81847	0.0131344	65	24948765	0.454124
Clean	(21, 5e−05)	−0.607156	0	0.35095	175.057	−0.151451	9.81703	0.0131344	65	24937876	0.450068

The sample frequency is $X = 10s$, we can see that comparing the cleaning result, different K does not change the statistics much, but as the γ decreases, the standard deviation, mean absolute deviation and skewness is going smaller, while the kurtosis is going larger. This is because a smaller γ put a more strict restriction on filtering, so that more data is classified to be outliers. Also, the cleaned data has a great improve on the skewness, kurtosis, and maximum drawdown.

The result is similar for JPM stocks:

==========JPM, Trade, frequency=10==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total trades	trades/quotes
Dirty		−0.263279	0	0.331027	156.366	−0.456166	40.0561	0.0239459	65	4968937	0.520853
Clean	(5, 0.0005)	−0.23536	0	0.325818	155.704	−0.460487	36.6329	0.0239459	65	4968495	0.520807
Clean	(5, 0.0001)	−0.212252	0	0.324549	155.34	−0.44187	36.57	0.0239459	65	4963875	0.520331
Clean	(5, 5e−05)	−0.200322	0	0.324111	155.181	−0.414248	36.4233	0.0239459	65	4955958	0.519523
Clean	(11, 0.0005)	−0.206453	0	0.324854	155.474	−0.495768	36.2176	0.0228136	65	4967473	0.520701
Clean	(11, 0.0001)	−0.230598	0	0.322356	153.981	−0.491598	36.8624	0.0228136	65	4922336	0.51601
Clean	(11, 5e−05)	−0.229514	0	0.318973	150.256	−0.501836	38.4571	0.0228136	65	4833743	0.50757
Clean	(21, 0.0005)	−0.175164	0	0.323967	155.311	−0.465474	35.985	0.0228136	65	4966605	0.520611
Clean	(21, 0.0001)	−0.164241	0	0.320516	152.747	−0.448722	37.1209	0.0228136	65	4910310	0.51484
Clean	(21, 5e−05)	−0.162436	0	0.3163	148.053	−0.463288	38.9342	0.0228136	65	4827178	0.507443

==========JPM, Quote, frequency=10==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total quotes	trades/quotes
Dirty		0.00937336	0	0.307429	140.729	−0.197061	24.8243	0.0210245	65	9539994	0.520853
Clean	(5, 0.0005)	0.00937336	0	0.307429	140.729	−0.197061	24.8243	0.0210245	65	9539994	0.520807
Clean	(5, 0.0001)	0.0125386	0	0.307329	140.703	−0.195051	24.8308	0.0210245	65	9539837	0.520331
Clean	(5, 5e−05)	0.0122708	0	0.307306	140.693	−0.195266	24.8365	0.0210245	65	9539438	0.519523
Clean	(11, 0.0005)	0.00936737	0	0.30741	140.721	−0.197314	24.8299	0.0210245	65	9539980	0.520701
Clean	(11, 0.0001)	0.0030796	0	0.307346	140.697	−0.195442	24.7941	0.0210245	65	9539220	0.51601
Clean	(11, 5e−05)	0.00435952	0	0.307308	140.662	−0.19577	24.7894	0.0210245	65	9523306	0.50757
Clean	(21, 0.0005)	0.00559786	0	0.3074	140.716	−0.197507	24.8326	0.0210245	65	9539951	0.520611
Clean	(21, 0.0001)	0.00217154	0	0.307214	140.663	−0.196286	24.8474	0.0210245	65	9537546	0.51484
Clean	(21, 5e−05)	0.00852482	0	0.307095	140.556	−0.197967	24.8666	0.0210245	65	9512745	0.507443

We can tune our sample frequency parameter X, and the result statistics is like:

$X = 30s$

==========AAPL, Trade, frequency=30==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total trades	trades/quotes
Dirty		−0.613361	0	0.356481	105.025	−0.296701	10.7328	0.0217453	65	11409653	0.457302
Clean	(5, 0.0005)	−0.61364	0	0.355699	104.903	−0.302776	10.7292	0.0217453	65	11408124	0.457241
Clean	(5, 0.0001)	−0.606921	0	0.35538	104.824	−0.306684	10.7378	0.0217453	65	11388411	0.456452
Clean	(5, 5e−05)	−0.60494	0	0.355418	104.834	−0.308462	10.7381	0.0217453	65	11345448	0.454745
Clean	(11, 0.0005)	−0.606856	0	0.355564	104.873	−0.305306	10.7312	0.0217453	65	11403503	0.457055
Clean	(11, 0.0001)	−0.608921	0	0.355225	104.792	−0.308238	10.7506	0.0217453	65	11341365	0.454578
Clean	(11, 5e−05)	−0.609804	0	0.355247	104.765	−0.309528	10.7686	0.0217453	65	11240846	0.450662
Clean	(21, 0.0005)	−0.607032	0	0.355069	104.802	−0.302189	10.6596	0.0217453	65	11400320	0.456928
Clean	(21, 0.0001)	−0.607609	0	0.354786	104.759	−0.309097	10.6559	0.0217453	65	11329837	0.454124
Clean	(21, 5e−05)	−0.608305	0	0.354721	104.697	−0.304135	10.6696	0.0217453	65	11223741	0.450068

==========AAPL, Quote, frequency=30==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total quotes	trades/quotes
Dirty		−0.616433	0	0.351604	103.088	−0.336483	11.1289	0.0214123	65	24949935	0.457302
Clean	(5, 0.0005)	−0.616433	0	0.351604	103.088	−0.336483	11.1289	0.0214123	65	24949935	0.457241
Clean	(5, 0.0001)	−0.616433	0	0.351604	103.088	−0.336483	11.1289	0.0214123	65	24949855	0.456452
Clean	(5, 5e−05)	−0.616435	0	0.3516	103.082	−0.336495	11.1296	0.0214123	65	24949025	0.454745
Clean	(11, 0.0005)	−0.616433	0	0.351604	103.088	−0.336483	11.1289	0.0214123	65	24949934	0.457055
Clean	(11, 0.0001)	−0.616437	0	0.351594	103.08	−0.336505	11.1304	0.0214123	65	24949203	0.454578
Clean	(11, 5e−05)	−0.616562	0	0.351587	103.077	−0.336587	11.1304	0.0214123	65	24942973	0.450662
Clean	(21, 0.0005)	−0.616433	0	0.351604	103.088	−0.336483	11.1289	0.0214123	65	24949932	0.456928
Clean	(21, 0.0001)	−0.619871	0	0.351601	103.083	−0.336356	11.1284	0.0214123	65	24948765	0.454124
Clean	(21, 5e−05)	−0.620424	0	0.351598	103.08	−0.336581	11.129	0.0214123	65	24937876	0.450068

$X = 60s$

==========AAPL, Trade, frequency=60==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total trades	trades/quotes
Dirty		−0.614788	0	0.357348	75.1421	−0.274998	8.39944	0.021975	65	11409653	0.457302
Clean	(5, 0.0005)	−0.614987	0	0.35679	75.0631	−0.278217	8.41902	0.021975	65	11408124	0.457241
Clean	(5, 0.0001)	−0.618613	0	0.356736	75.0552	−0.277195	8.41862	0.021975	65	11388411	0.456452
Clean	(5, 5e−05)	−0.616638	0	0.356749	75.0401	−0.279123	8.42224	0.021975	65	11345448	0.454745
Clean	(11, 0.0005)	−0.615009	0	0.35673	75.0582	−0.275027	8.41759	0.021975	65	11403503	0.457055
Clean	(11, 0.0001)	−0.621796	0	0.356739	75.0327	−0.277905	8.43105	0.021975	65	11341365	0.454578
Clean	(11, 5e−05)	−0.621229	0	0.356702	75.0182	−0.280676	8.43613	0.021975	65	11240846	0.450662
Clean	(21, 0.0005)	−0.615098	0	0.35648	75.0279	−0.277484	8.42599	0.021975	65	11400320	0.456928
Clean	(21, 0.0001)	−0.619839	0	0.356497	75.0207	−0.279974	8.42437	0.021975	65	11329837	0.454124
Clean	(21, 5e−05)	−0.620005	0	0.356356	74.9816	−0.281572	8.44048	0.021975	65	11223741	0.450068

==========AAPL, Quote, frequency=60==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total quotes	trades/quotes
Dirty		−0.634699	0	0.35431	74.1141	−0.303036	8.91667	0.0225328	65	24949935	0.457302
Clean	(5, 0.0005)	−0.634699	0	0.35431	74.1141	−0.303036	8.91667	0.0225328	65	24949935	0.457241
Clean	(5, 0.0001)	−0.634699	0	0.35431	74.1141	−0.303036	8.91667	0.0225328	65	24949855	0.456452
Clean	(5, 5e−05)	−0.6347	0	0.354305	74.1074	−0.303051	8.9173	0.0225328	65	24949025	0.454745
Clean	(11, 0.0005)	−0.634699	0	0.35431	74.1141	−0.303036	8.91667	0.0225328	65	24949934	0.457055
Clean	(11, 0.0001)	−0.634701	0	0.354304	74.1094	−0.303032	8.91744	0.0225328	65	24949203	0.454578
Clean	(11, 5e−05)	−0.63483	0	0.354288	74.1056	−0.302182	8.91564	0.0225328	65	24942973	0.450662
Clean	(21, 0.0005)	−0.634699	0	0.35431	74.1141	−0.303036	8.91667	0.0225328	65	24949932	0.456928
Clean	(21, 0.0001)	−0.638138	0	0.354317	74.1124	−0.302941	8.9152	0.0225328	65	24948765	0.454124
Clean	(21, 5e−05)	−0.638698	0	0.354295	74.109	−0.302715	8.91455	0.0225328	65	24937876	0.450068

$X = 300s$

==========AAPL, Trade, frequency=300==========

Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total trades	trades/quotes
Dirty		−0.64343	0	0.345231	32.4224	−0.456146	6.60853	0.0308368	65	11409653	0.457302
Clean	(5, 0.0005)	−0.64343	0	0.345231	32.4224	−0.456146	6.60853	0.0308368	65	11408124	0.457241
Clean	(5, 0.0001)	−0.645449	0	0.345206	32.4233	−0.457213	6.6089	0.0308368	65	11388411	0.456452
Clean	(5, 5e−05)	−0.64347	0	0.345176	32.4159	−0.452336	6.6085	0.0308368	65	11345448	0.454745
Clean	(11, 0.0005)	−0.64343	0	0.345231	32.4224	−0.456146	6.60853	0.0308368	65	11403503	0.457055
Clean	(11, 0.0001)	−0.654667	0	0.345036	32.4101	−0.465128	6.60254	0.0308368	65	11341365	0.454578
Clean	(11, 5e−05)	−0.654092	0	0.344985	32.4061	−0.46794	6.6372	0.0310617	65	11240846	0.450662
Clean	(21, 0.0005)	−0.649429	0	0.345066	32.4165	−0.461948	6.59819	0.0308368	65	11400320	0.456928
Clean	(21, 0.0001)	−0.653181	0	0.34504	32.4049	−0.468382	6.63808	0.0310617	65	11329837	0.454124
Clean	(21, 5e−05)	−0.652422	0	0.34488	32.4039	−0.463937	6.61854	0.0310617	65	11223741	0.450068

==========AAPL, Quote, frequency=300==========

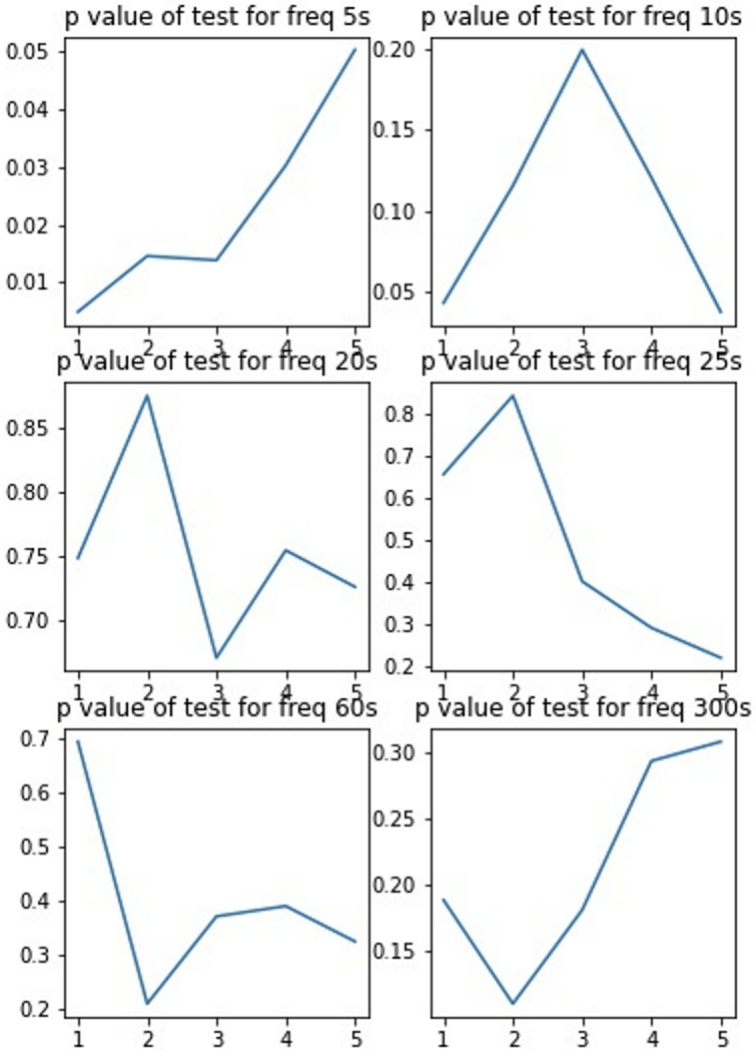
Type	Parameter	mean	median	std	mad	skewness	kurtosis	maximum drawdown	sample length	total quotes	trades/quotes
Dirty		−0.66871	0	0.343322	32.1558	−0.475865	6.75538	0.0312431	65	24949935	0.457302
Clean	(5, 0.0005)	−0.66871	0	0.343322	32.1558	−0.475865	6.75538	0.0312431	65	24949935	0.457241
Clean	(5, 0.0001)	−0.66871	0	0.343322	32.1558	−0.475865	6.75538	0.0312431	65	24949855	0.456452
Clean	(5, 5e−05)	−0.668709	0	0.343323	32.1562	−0.475863	6.75533	0.0312431	65	24949025	0.454745
Clean	(11, 0.0005)	−0.66871	0	0.343322	32.1558	−0.475865	6.75538	0.0312431	65	24949934	0.457055
Clean	(11, 0.0001)	−0.668711	0	0.343318	32.1518	−0.475882	6.75588	0.0312431	65	24949203	0.454578
Clean	(11, 5e−05)	−0.668834	0	0.343324	32.1519	−0.476137	6.75669	0.0312431	65	24942973	0.450662
Clean	(21, 0.0005)	−0.66871	0	0.343322	32.1558	−0.475865	6.75538	0.0312431	65	24949932	0.456928
Clean	(21, 0.0001)	−0.672189	0	0.343315	32.1553	−0.475441	6.75534	0.0312431	65	24948765	0.454124
Clean	(21, 5e−05)	−0.672764	0	0.343265	32.1495	−0.476927	6.75096	0.0312431	65	24937876	0.450068

As we can see when the sample frequency is smaller (as X goes larger), the kurtosis goes smaller, and the returns distribution is more like normal distribution, this is consistent with our intuition.

Q3(a)

We test the resampled return with frequency {5s, 10s, 20s, 25s, 60s, 300s} with lag {1, 2, 3, 4, 5}.

The graphs of p value:



If we set 0.1 as the threshold of whether there is a significant autocorrelation. We can see that only frequence: 20s and 60s passed ljungbox test for lag $\belong {1,2,3,4,5}$.

Since we also want to maximize the amount of data we use, it is better to make the resampling frequency as samll as possible. The best frequency we should choose is 20s.

Q3(b)

Results of Dicky Fuller test for cleaned data and drop nan data directly:

```
Dickey Fuller Testing result of trade data
   Test Statistic  p-value  #Lags Used
0      -34.427876      0.0           0
Dickey Fuller Testing result of quote data
   Test Statistic  p-value  #Lags Used
0      -31.685962      0.0           0
```

Results of Dicky Fuller test for cleaned data and filling nan data by zero:

```
Dickey Fuller Testing result of trade data
   Test Statistic  p-value  #Lags Used
0      -34.438101      0.0           0
Dickey Fuller Testing result of quote data
   Test Statistic  p-value  #Lags Used
0      -31.695417      0.0           0
```

We can see that whether we fill nan or drop nan, we cannot reject null hypothesis both for quote data and trade data so the 20s return should not be stationary.

Problem 4

(a)(b)

The optimizer calculating the optimal portfolio by constructing a quadratic programming system and solving it:

$$\begin{gathered}\text{Minimize } \frac{1}{2} x^{\intercal} \mu S x - x^{\intercal} \bar{p} \\ \text{Subject to } Gx \preceq h \\ \text{and } Ax = b\end{gathered}$$

S is the correlation matrix of the portfolio (in the case of the example, the portfolio has three assets with riska and a risk-free asset). μ ranges from 0.1 to 1, controlling the weight of risk-free asset in the portfolio and thereby, controlling the risk of the portfolio. $\bar{p}$ is the vector of each asset's return.

Condition $Gx \preceq h$ ensures no shorting, and condition $Ax = b$ ensures the weights sum to 1

By solving such a system, we can find our optimal portfolio and their expected return under different levels of risk.

Output:

The output xs is an array of vectors. Each vector represents a portfolio (weight of each asset in the portfolio) under a specific level of risk.

Figure 1 shows the expected returns under different levels of risk. Risk-free portfolio has standard deviation equals to 0 and expected return equals to 0.03. The portfolio with highest risk has standard deviation equals to 0.2 and expected return equals to 0.12.

Figure 2 shows the weights of assets under different levels of risk. If drawing a line perpendicular to the horizontal axis, the intersection points of the line with curves for the assets show the cumulative weights of these assets in an optimal portfolio under such standard deviation. Risk-free portfolio has only risk-free asset, and the portfolio with highest risk includes only $x1$.

Max iterations is 100, and the optimization will converge if the objective function value gap between 2 iterations is smaller than or equal to 10^{-7} (with other conditions).

(c)

For problem c, we calculated the weights for the market portfolio at 2007/06/20 and 2007/09/20 and wrote the result to two csv files jun_20_holdings.csv and sep_20_holdings.csv

For turnover rate, we used two method:

1. change the portfolio every day according to market capitalization, the result (annualized by multiplying the result by 4) was around 1.93;

2. only change the portfolio at 2007/09/20 according to market capitalization, the result (annualized by multiplying the result by 4) was 0.07.

The formula we used is

$$\text{turnover rate} = \frac{\min(\text{sum of buying in dollar amount}, \text{sum of selling in dollar amount})}{\text{average market portfolio value}}$$

We set the total portfolio value to $\$10^8$ and do the calculation.