UNIVERSITY OF MALAYA

DATA MINING (WQD7005)
Assignment 3

Instructor: Profesor Madya Dr. Teh Ying Wah

Group Members:

Name	**Matric No.**
Choy Siew Wearn	WQD180053
Teng Lung Yun	WQD180024
Wo Choy Yun	WQD180054
Xu Xiang	WQD180061
Zhu Ting	WQD180077

1.0 Introduction

Stock analysis is an essential step for investors to search for potential profitable stocks and analyse stocks. As such, analysts can perform Principal Component Analysis (PCA) and algorithms such as Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA). Both Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA) are time series representation techniques that can be used with optimization algorithms to help analysts to identify hidden and relevant patterns in financial time series data.

1.1 Principal Component Analysis And Covariance

Principal Component Analysis (PCA) was invented by Karl Pearson in 1901. PCA uses orthogonal transformation to transform a set of data with possible correlated variables into linearly uncorrelated variables which is known as the principal components. In most cases, PCA is used an a exploratory tool for data analysis to make prediction models. In this assignment, PCA is done through covariance matrix where the results will be shown later in section 2.1.

1.2 Symbolic Aggregate Approximation & Piecewise Aggregate Approximation

Both Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA) algorithms are invented by Eamonn Keogh and Jessica Lin. The main function of PAA is to reduce the dimensionality of the time-series by dividing the time-series into equal segments, then computing each of them by averaging the values in the segment. As for SAX, it is an extended based approach of PAA where it transforms the original time-series into PAA representation, then converts the PAA data into a string.

2.0 Data Preprocessing

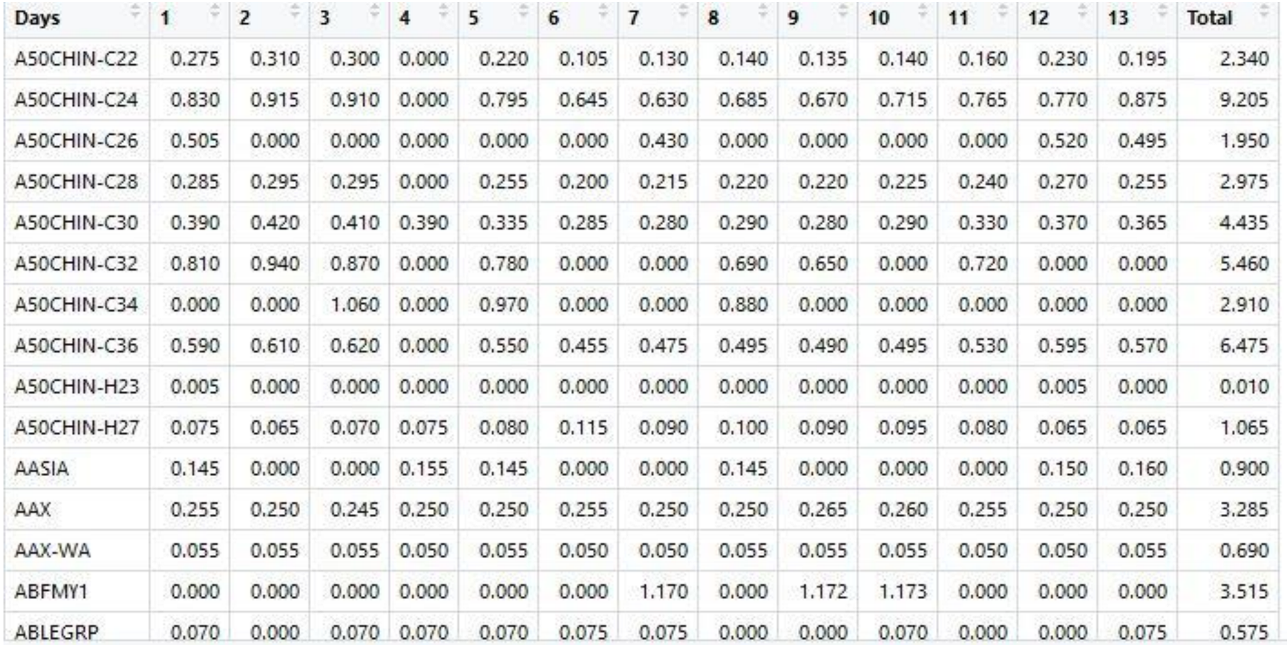
In this assignment, we are using two dataset. The first dataset is ss.csv from previous milestone and the second dataset are crawled data for 13 days.

First dataset are be used to calculate PCA and covariance between attributes. Second datasets are used to calculate SAX and PAA, because second dataset is time-series data.

For the second dataset, 13 days of data are compiled into one file using R. In order for us to perform SAX and PAA, data preprocessing is required such as cleaning data. For data preprocessing, we did:

a. Created a sum column at the right known as “Total” to sum all the stock values
b. Remove rows with sum value “0.000”
c. Replace NA with “0.000” in dataset
d. Change all columns attribute of those containing values to numeric

Figure 2.0.1 below shows the dataset that has already been cleaned or preprocessed. In the first column shows the list of companies, where from columns named “1” to “13” contain the stock values for each company from day 1 to day 13. The column named “Total” is the sum of all stock values from day 1 to day 13.

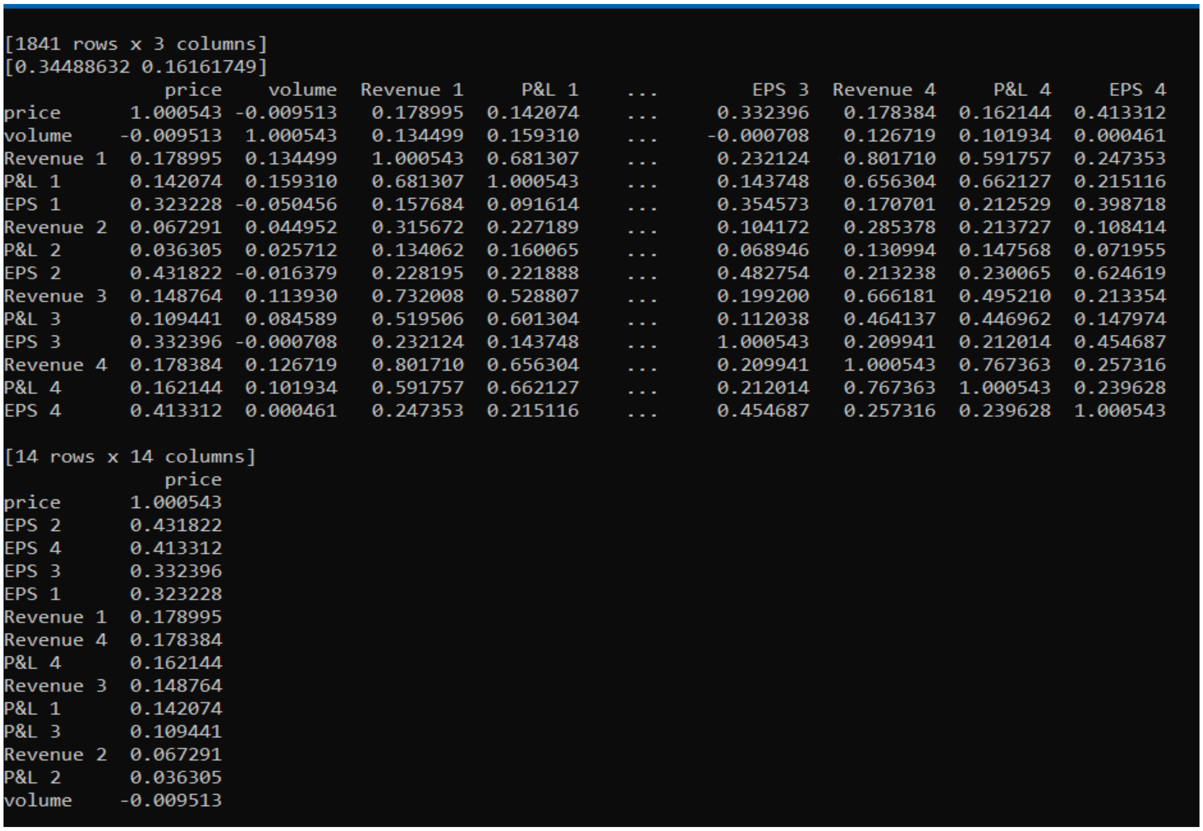


Days	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
A50CHIN-C22	0.275	0.310	0.300	0.000	0.220	0.105	0.130	0.140	0.135	0.140	0.160	0.230	0.195	2.340
A50CHIN-C24	0.830	0.915	0.910	0.000	0.795	0.645	0.630	0.685	0.670	0.715	0.765	0.770	0.875	9.205
A50CHIN-C26	0.505	0.000	0.000	0.000	0.000	0.000	0.430	0.000	0.000	0.000	0.000	0.520	0.495	1.950
A50CHIN-C28	0.285	0.295	0.295	0.000	0.255	0.200	0.215	0.220	0.220	0.225	0.240	0.270	0.255	2.975
A50CHIN-C30	0.390	0.420	0.410	0.390	0.335	0.285	0.280	0.290	0.280	0.290	0.330	0.370	0.365	4.435
A50CHIN-C32	0.810	0.940	0.870	0.000	0.780	0.000	0.000	0.690	0.650	0.000	0.720	0.000	0.000	5.460
A50CHIN-C34	0.000	0.000	1.060	0.000	0.970	0.000	0.000	0.880	0.000	0.000	0.000	0.000	0.000	2.910
A50CHIN-C36	0.590	0.610	0.620	0.000	0.550	0.455	0.475	0.495	0.490	0.495	0.530	0.595	0.570	6.475
A50CHIN-H23	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.010
A50CHIN-H27	0.075	0.065	0.070	0.075	0.080	0.115	0.090	0.100	0.090	0.095	0.080	0.065	0.065	1.065
AASIA	0.145	0.000	0.000	0.155	0.145	0.000	0.000	0.145	0.000	0.000	0.000	0.150	0.160	0.900
AAX	0.255	0.250	0.245	0.250	0.250	0.255	0.250	0.250	0.265	0.260	0.255	0.250	0.250	3.285
AAX-WA	0.055	0.055	0.055	0.050	0.055	0.050	0.050	0.055	0.055	0.055	0.050	0.050	0.055	0.690
ABFMY1	0.000	0.000	0.000	0.000	0.000	0.000	1.170	0.000	1.172	1.173	0.000	0.000	0.000	3.515
ABLEGRP	0.070	0.000	0.070	0.070	0.070	0.075	0.075	0.000	0.000	0.070	0.000	0.000	0.075	0.575

Figure 2.0.1: Second dataset Preprocessed Stock Data

2.1 Covariance

In this part of assignment, PCA and Covariance among attributes is done using Python. Using the first dataset, the results from the Python code determining covariance among attributes will be shown in Figure 2.1.1 below:



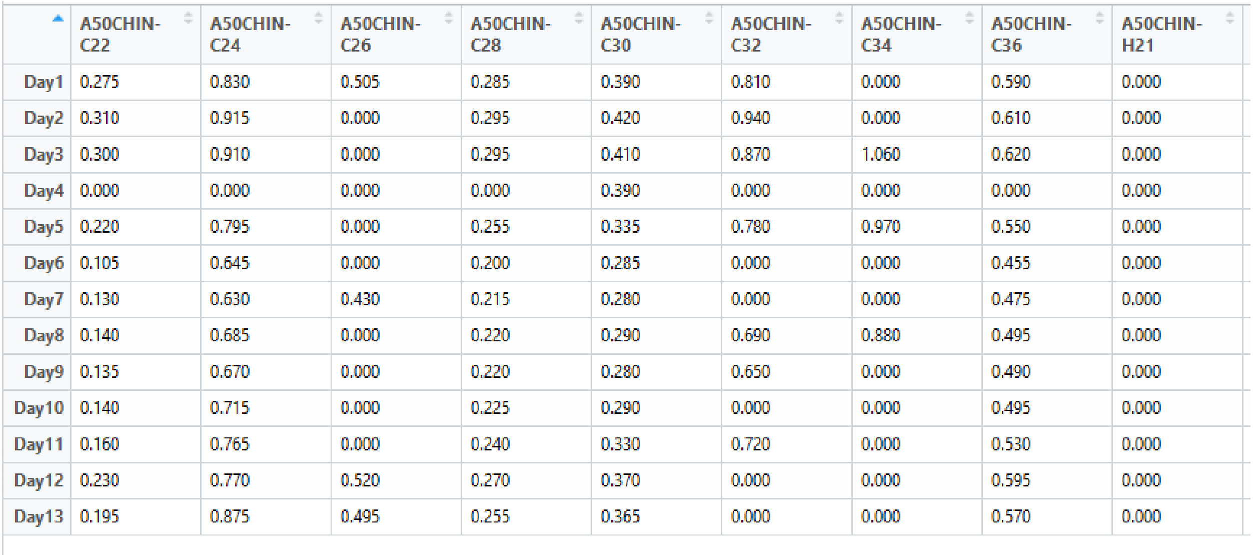
```
[1841 rows x 3 columns]
[0.34488632 0.16161749]
              price    volume  Revenue 1     P&L 1    ...       EPS 3  Revenue 4     P&L 4     EPS 4
price      1.000543 -0.009513   0.178995  0.142074    ...    0.332396   0.178384  0.162144  0.413312
volume    -0.009513  1.000543   0.134499  0.159310    ...   -0.000708   0.126719  0.101934  0.000461
Revenue 1  0.178995  0.134499   1.000543  0.681307    ...    0.232124   0.801710  0.591757  0.247353
P&L 1      0.142074  0.159310   0.681307  1.000543    ...    0.143748   0.656304  0.662127  0.215116
EPS 1      0.323228 -0.050456   0.157684  0.091614    ...    0.354573   0.170701  0.212529  0.398718
Revenue 2  0.067291  0.044952   0.315672  0.227189    ...    0.104172   0.285378  0.213727  0.108414
P&L 2      0.036305  0.025712   0.134062  0.160065    ...    0.068946   0.130994  0.147568  0.071955
EPS 2      0.431822 -0.016379   0.228195  0.221888    ...    0.482754   0.213238  0.230065  0.624619
Revenue 3  0.148764  0.113930   0.732008  0.528807    ...    0.199200   0.666181  0.495210  0.213354
P&L 3      0.109441  0.084589   0.519506  0.601304    ...    0.112038   0.464137  0.446962  0.147974
EPS 3      0.332396 -0.000708   0.232124  0.143748    ...    1.000543   0.209941  0.212014  0.454687
Revenue 4  0.178384  0.126719   0.801710  0.656304    ...    0.209941   1.000543  0.767363  0.257316
P&L 4      0.162144  0.101934   0.591757  0.662127    ...    0.212014   0.767363  1.000543  0.239628
EPS 4      0.413312  0.000461   0.247353  0.215116    ...    0.454687   0.257316  0.239628  1.000543

[14 rows x 14 columns]
              price
price      1.000543
EPS 2      0.431822
EPS 4      0.413312
EPS 3      0.332396
EPS 1      0.323228
Revenue 1  0.178995
Revenue 4  0.178384
P&L 4      0.162144
Revenue 3  0.148764
P&L 1      0.142074
P&L 3      0.109441
Revenue 2  0.067291
P&L 2      0.036305
volume    -0.009513
```

Figure 2.1.1: Covariance Results Among Attributes Using Python

Covariance calculations are used to find relationships between dimensions in high dimensional data sets where visualization is difficult. Figure 2.1.1 shows a covariance matrix between the target variable price and other variables and a reduced covariance matrix. The reduced covariance matrix shows that EPS 2, has the highest covariance in relative to price, followed by EPS 4, EPS 3 and EPS 1, etc. This shows that the price tends to be affected by EPS 2. Positive value of covariance indicates both dimensions increase or decrease together. A negative value indicates while one increases the other decreases, or vice-versa. The reduced covariance matrix shows all positive covariance values in relative to price except ‘volume’ variable.

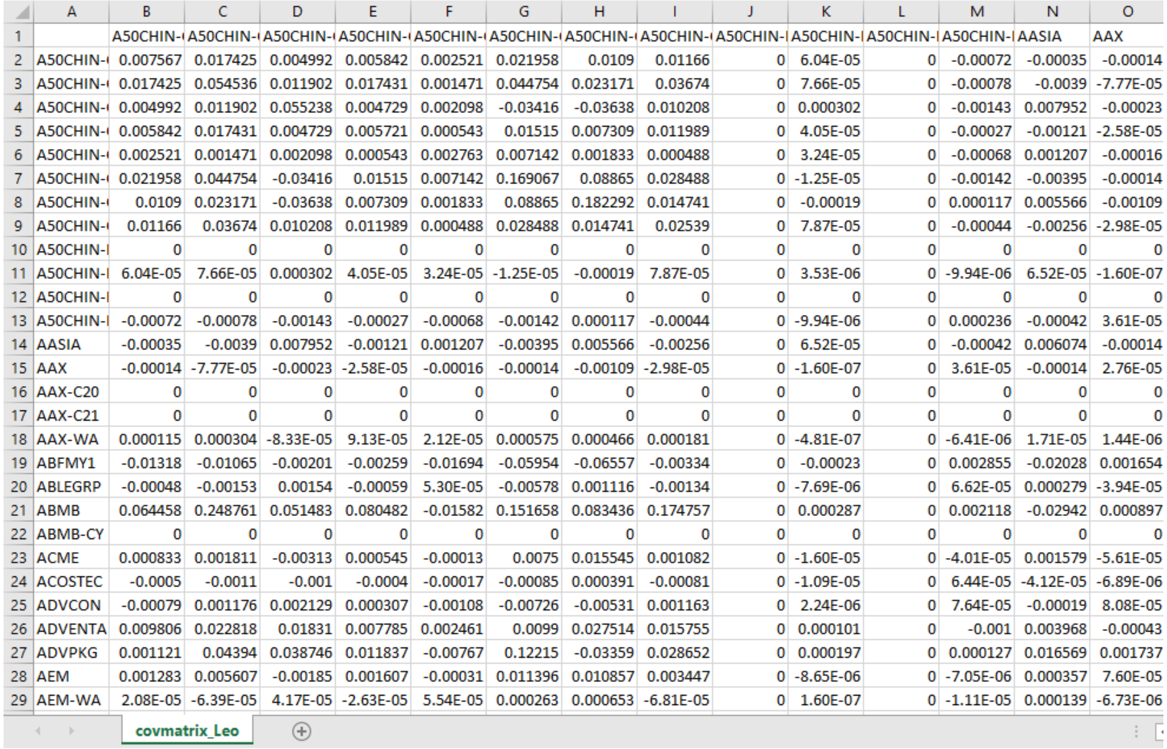
The second dataset preprocessed data is transposed to below figure:



	A50CHIN-C22	A50CHIN-C24	A50CHIN-C26	A50CHIN-C28	A50CHIN-C30	A50CHIN-C32	A50CHIN-C34	A50CHIN-C36	A50CHIN-H21
Day1	0.275	0.830	0.505	0.285	0.390	0.810	0.000	0.590	0.000
Day2	0.310	0.915	0.000	0.295	0.420	0.940	0.000	0.610	0.000
Day3	0.300	0.910	0.000	0.295	0.410	0.870	1.060	0.620	0.000
Day4	0.000	0.000	0.000	0.000	0.390	0.000	0.000	0.000	0.000
Day5	0.220	0.795	0.000	0.255	0.335	0.780	0.970	0.550	0.000
Day6	0.105	0.645	0.000	0.200	0.285	0.000	0.000	0.455	0.000
Day7	0.130	0.630	0.430	0.215	0.280	0.000	0.000	0.475	0.000
Day8	0.140	0.685	0.000	0.220	0.290	0.690	0.880	0.495	0.000
Day9	0.135	0.670	0.000	0.220	0.280	0.650	0.000	0.490	0.000
Day10	0.140	0.715	0.000	0.225	0.290	0.000	0.000	0.495	0.000
Day11	0.160	0.765	0.000	0.240	0.330	0.720	0.000	0.530	0.000
Day12	0.230	0.770	0.520	0.270	0.370	0.000	0.000	0.595	0.000
Day13	0.195	0.875	0.495	0.255	0.365	0.000	0.000	0.570	0.000

Figure 2.1.2: Stock with price for 13 days

Using data in Figure 2.1.2, covariance among stock price is determined using R. The results are shown in Figure 2.1.3 below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	A50CHIN-	AASIA	AAX
2	A50CHIN-	0.007567	0.017425	0.004992	0.005842	0.002521	0.021958	0.0109	0.01166	0	6.04E-05	0	-0.00072	-0.00035	-0.00014
3	A50CHIN-	0.017425	0.054536	0.011902	0.017431	0.001471	0.044754	0.023171	0.03674	0	7.66E-05	0	-0.00078	-0.0039	-7.77E-05
4	A50CHIN-	0.004992	0.011902	0.055238	0.004729	0.002098	-0.03416	-0.03638	0.010208	0	0.000302	0	-0.00143	0.007952	-0.00023
5	A50CHIN-	0.005842	0.017431	0.004729	0.005721	0.000543	0.01515	0.007309	0.011989	0	4.05E-05	0	-0.00027	-0.00121	-2.58E-05
6	A50CHIN-	0.002521	0.001471	0.002098	0.000543	0.002763	0.007142	0.001833	0.000488	0	3.24E-05	0	-0.00068	0.001207	-0.00016
7	A50CHIN-	0.021958	0.044754	-0.03416	0.01515	0.007142	0.169067	0.08865	0.028488	0	-1.25E-05	0	-0.00142	-0.00395	-0.00014
8	A50CHIN-	0.0109	0.023171	-0.03638	0.007309	0.001833	0.08865	0.182292	0.014741	0	-0.00019	0	0.000117	0.005566	-0.00109
9	A50CHIN-	0.01166	0.03674	0.010208	0.011989	0.000488	0.028488	0.014741	0.02539	0	7.87E-05	0	-0.00044	-0.00256	-2.98E-05
10	A50CHIN-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	A50CHIN-	6.04E-05	7.66E-05	0.000302	4.05E-05	3.24E-05	-1.25E-05	-0.00019	7.87E-05	0	3.53E-06	0	-9.94E-06	6.52E-05	-1.60E-07
12	A50CHIN-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	A50CHIN-	-0.00072	-0.00078	-0.00143	-0.00027	-0.00068	-0.00142	0.000117	-0.00044	0	-9.94E-06	0	0.000236	-0.00042	3.61E-05
14	AASIA	-0.00035	-0.0039	0.007952	-0.00121	0.001207	-0.00395	0.005566	-0.00256	0	6.52E-05	0	-0.00042	0.006074	-0.00014
15	AAX	-0.00014	-7.77E-05	-0.00023	-2.58E-05	-0.00016	-0.00014	-0.00109	-2.98E-05	0	-1.60E-07	0	3.61E-05	-0.00014	2.76E-05
16	AAX-C20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	AAX-C21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	AAX-WA	0.000115	0.000304	-8.33E-05	9.13E-05	2.12E-05	0.000575	0.000466	0.000181	0	-4.81E-07	0	-6.41E-06	1.71E-05	1.44E-06
19	ABFMY1	-0.01318	-0.01065	-0.00201	-0.00259	-0.01694	-0.05954	-0.06557	-0.00334	0	-0.00023	0	0.002855	-0.02028	0.001654
20	ABLEGRP	-0.00048	-0.00153	0.00154	-0.00059	5.30E-05	-0.00578	0.001116	-0.00134	0	-7.69E-06	0	6.62E-05	0.000279	-3.94E-05
21	ABMB	0.064458	0.248761	0.051483	0.080482	-0.01582	0.151658	0.083436	0.174757	0	0.000287	0	0.002118	-0.02942	0.000897
22	ABMB-CY	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	ACME	0.000833	0.001811	-0.00313	0.000545	-0.00013	0.0075	0.015545	0.001082	0	-1.60E-05	0	-4.01E-05	0.001579	-5.61E-05
24	ACOSTEC	-0.0005	-0.0011	-0.001	-0.0004	-0.00017	-0.00085	0.000391	-0.00081	0	-1.09E-05	0	6.44E-05	-4.12E-05	-6.89E-06
25	ADVCON	-0.00079	0.001176	0.002129	0.000307	-0.00108	-0.00726	-0.00531	0.001163	0	2.24E-06	0	7.64E-05	-0.00019	8.08E-05
26	ADVENTA	0.009806	0.022818	0.01831	0.007785	0.002461	0.0099	0.027514	0.015755	0	0.000101	0	-0.001	0.003968	-0.00043
27	ADVPKG	0.001121	0.04394	0.038746	0.011837	-0.00767	0.12215	-0.03359	0.028652	0	0.000197	0	0.000127	0.016569	0.001737
28	AEM	0.001283	0.005607	-0.00185	0.001607	-0.00031	0.011396	0.010857	0.003447	0	-8.65E-06	0	-7.05E-06	0.000357	7.60E-05
29	AEM-WA	2.08E-05	-6.39E-05	4.17E-05	-2.63E-05	5.54E-05	0.000263	0.000653	-6.81E-05	0	1.60E-07	0	-1.11E-05	0.000139	-6.73E-06

Figure 2.1.3: Stock Price Covariance

The readings above indicate the extent to which the volatility in prices move together. These readings are shown in covmatrix_Leo.csv.

A variance-covariance matrix is particularly useful when it comes to analysing the volatility between elements of a group of data. For instance, a variance-covariance matrix has particular applications when it comes to analysing portfolio returns.

If several assets with a high covariance are included in a portfolio, then this represents high risk. This indicates that several assets with high volatility move together, which is what investors would typically want to avoid.

On the other hand, selecting assets that show negative covariance allows for greater diversification of the portfolio, since the volatility of the assets do not tend to move together.

2.2 PCA

PCA is a dimensionality reduction algorithm where new features are created which represents the original feature dimensions in a lower dimension with a little loss of the total information. PCA is a linear transformation that chooses a new coordinate system for the dataset such that greatest variance by an y projection of the data set comes to lie on the first axis(then called the first principal component), the second greatest variance on the second axis, and so on.

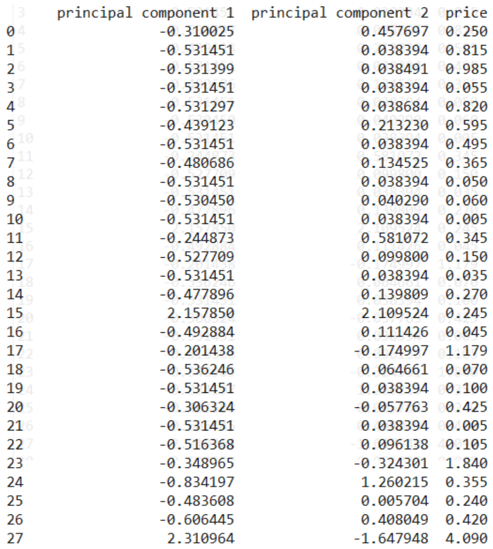
The dataset used for PCA is the first dataset – ss.csv. The Features for PCA calculations for this assignment are Volume, Revenue 1, P&L 1, EPS 1, Revenue 2, P&L 2, EPS 2, Revenue 3, P&L 3, EPS 3, Revenue 4, P&L 4, EPS 4.

Volume refers to buy and sales volume of the day.

Revenue, Profit and Loss(P&L) and Earning per Share(EPS) are the quarter results for each month for previous consecutive 4 months.

Target variable is the price of the stocks.

The features are reduced to 2 dimensions as below Figure 2.2.1



```
    principal component 1  principal component 2  price
0               -0.310025               0.457697  0.250
1               -0.531451               0.038394  0.815
2               -0.531399               0.038491  0.985
3               -0.531451               0.038394  0.055
4               -0.531297               0.038684  0.820
5               -0.439123               0.213230  0.595
6               -0.531451               0.038394  0.495
7               -0.480686               0.134525  0.365
8               -0.531451               0.038394  0.050
9               -0.530450               0.040290  0.060
10              -0.531451               0.038394  0.005
11              -0.244873               0.581072  0.345
12              -0.527709               0.099800  0.150
13              -0.531451               0.038394  0.035
14              -0.477896               0.139809  0.270
15               2.157850               2.109524  0.245
16              -0.492884               0.111426  0.045
17              -0.201438              -0.174997  1.179
18              -0.536246               0.064661  0.070
19              -0.531451               0.038394  0.100
20              -0.306324              -0.057763  0.425
21              -0.531451               0.038394  0.005
22              -0.516368               0.096138  0.105
23              -0.348965              -0.324301  1.840
24              -0.834197               1.260215  0.355
25              -0.483608               0.005704  0.240
26              -0.606445               0.408049  0.420
27               2.310964              -1.647948  4.090
```

Figure 2.2.1: Principal Component Values with the Target Variable – Closing Price

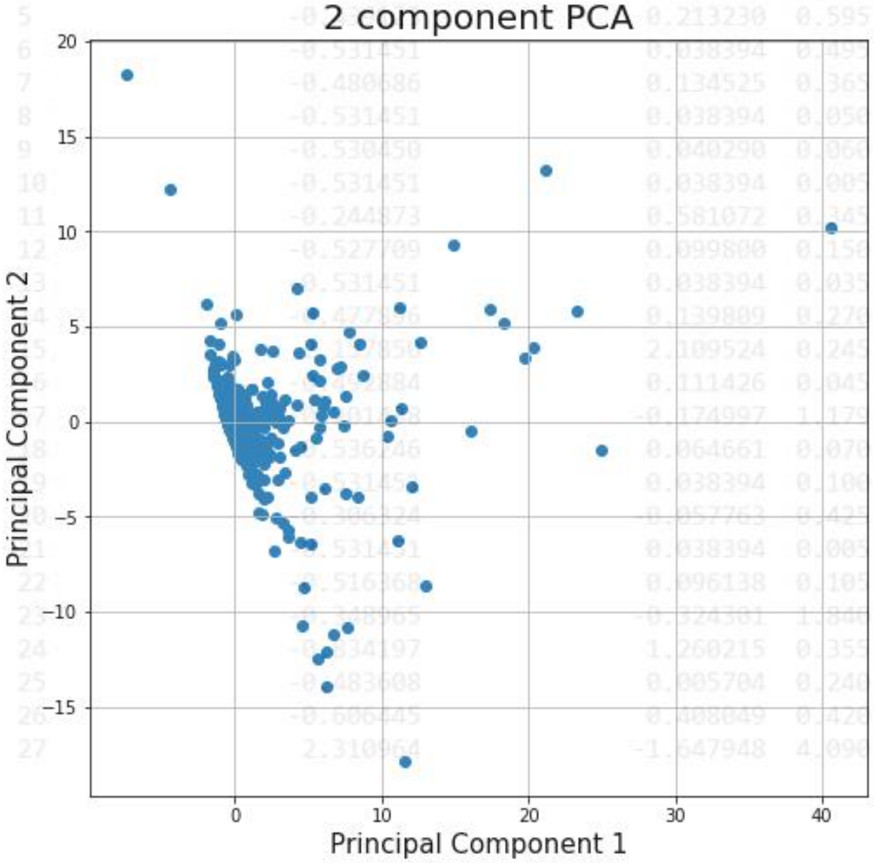
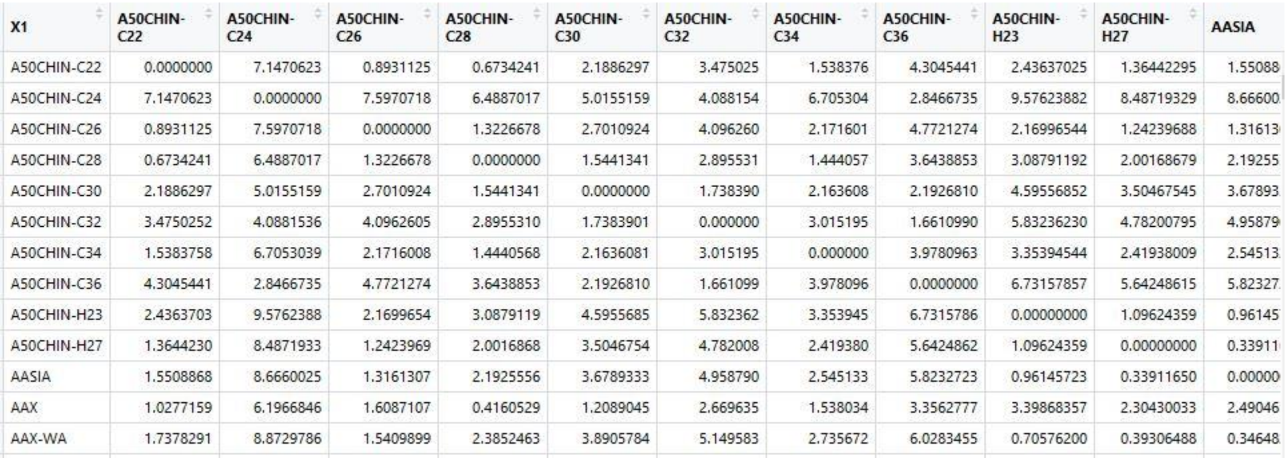


Figure 2.2.2: Graph of Principal Components

2.3 Work Flow On Performing SAX and PAA

In this assignment, SAX and PAA algorithms is performed using R. The steps and results will be shown below:

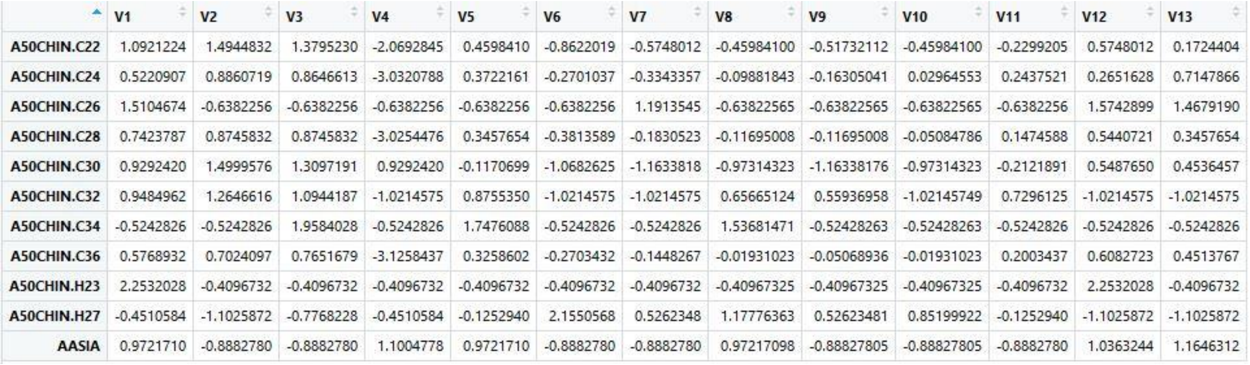
a. Perform Euclidean value to determine the distance relationship



X1	A50CHIN-C22	A50CHIN-C24	A50CHIN-C26	A50CHIN-C28	A50CHIN-C30	A50CHIN-C32	A50CHIN-C34	A50CHIN-C36	A50CHIN-H23	A50CHIN-H27	AASIA
A50CHIN-C22	0.0000000	7.1470623	0.8931125	0.6734241	2.1886297	3.475025	1.538376	4.3045441	2.43637025	1.36442295	1.55088
A50CHIN-C24	7.1470623	0.0000000	7.5970718	6.4887017	5.0155159	4.088154	6.705304	2.8466735	9.57623882	8.48719329	8.66600
A50CHIN-C26	0.8931125	7.5970718	0.0000000	1.3226678	2.7010924	4.096260	2.171601	4.7721274	2.16996544	1.24239688	1.31613
A50CHIN-C28	0.6734241	6.4887017	1.3226678	0.0000000	1.5441341	2.895531	1.444057	3.6438853	3.08791192	2.00168679	2.19255
A50CHIN-C30	2.1886297	5.0155159	2.7010924	1.5441341	0.0000000	1.738390	2.163608	2.1926810	4.59556852	3.50467545	3.67893
A50CHIN-C32	3.4750252	4.0881536	4.0962605	2.8955310	1.7383901	0.000000	3.015195	1.6610990	5.83236230	4.78200795	4.95879
A50CHIN-C34	1.5383758	6.7053039	2.1716008	1.4440568	2.1636081	3.015195	0.000000	3.9780963	3.35394544	2.41938009	2.54513
A50CHIN-C36	4.3045441	2.8466735	4.7721274	3.6438853	2.1926810	1.661099	3.978096	0.0000000	6.73157857	5.64248615	5.82327
A50CHIN-H23	2.4363703	9.5762388	2.1699654	3.0879119	4.5955685	5.832362	3.353945	6.7315786	0.00000000	1.09624359	0.96145
A50CHIN-H27	1.3644230	8.4871933	1.2423969	2.0016868	3.5046754	4.782008	2.419380	5.6424862	1.09624359	0.00000000	0.33911
AASIA	1.5508868	8.6660025	1.3161307	2.1925556	3.6789333	4.958790	2.545133	5.8232723	0.96145723	0.33911650	0.00000
AAX	1.0277159	6.1966846	1.6087107	0.4160529	1.2089045	2.669635	1.538034	3.3562777	3.39868357	2.30430033	2.49046
AAX-WA	1.7378291	8.8729786	1.5409899	2.3852463	3.8905784	5.149583	2.735672	6.0283455	0.70576200	0.39306488	0.34648

Figure 2.3.1: Stock Data with Euclidean Value Using R

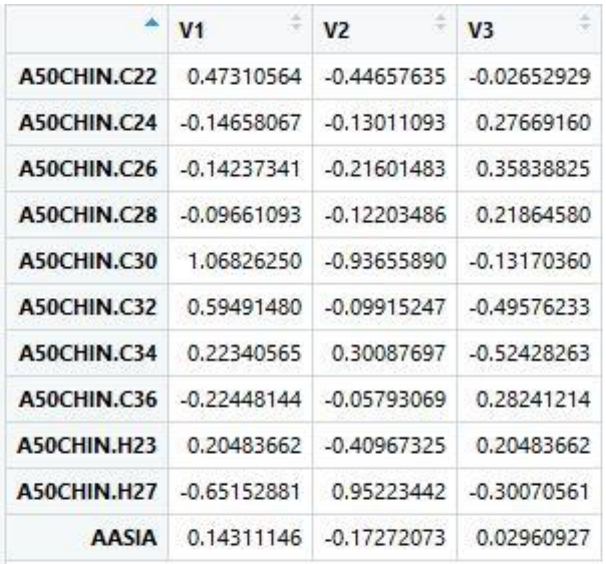
b. Execute Z-Normalization on the data to reduce the value gap between multiple lines



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
A50CHIN.C22	1.0921224	1.4944832	1.3795230	-2.0692845	0.4598410	-0.8622019	-0.5748012	-0.45984100	-0.51732112	-0.45984100	-0.2299205	0.5748012	0.1724404
A50CHIN.C24	0.5220907	0.8860719	0.8646613	-3.0320788	0.3722161	-0.2701037	-0.3343357	-0.09881843	-0.16305041	0.02964553	0.2437521	0.2651628	0.7147866
A50CHIN.C26	1.5104674	-0.6382256	-0.6382256	-0.6382256	-0.6382256	-0.6382256	1.1913545	-0.63822565	-0.63822565	-0.63822565	-0.6382256	1.5742899	1.4679190
A50CHIN.C28	0.7423787	0.8745832	0.8745832	-3.0254476	0.3457654	-0.3813589	-0.1830523	-0.11695008	-0.11695008	-0.05084786	0.1474588	0.5440721	0.3457654
A50CHIN.C30	0.9292420	1.4999576	1.3097191	0.9292420	-0.1170699	-1.0682625	-1.1633818	-0.97314323	-1.16338176	-0.97314323	-0.2121891	0.5487650	0.4536457
A50CHIN.C32	0.9484962	1.2646616	1.0944187	-1.0214575	0.8755350	-1.0214575	-1.0214575	0.65665124	0.55936958	-1.02145749	0.7296125	-1.0214575	-1.0214575
A50CHIN.C34	-0.5242826	-0.5242826	1.9584028	-0.5242826	1.7476088	-0.5242826	-0.5242826	1.53681471	-0.52428263	-0.52428263	-0.5242826	-0.5242826	-0.5242826
A50CHIN.C36	0.5768932	0.7024097	0.7651679	-3.1258437	0.3258602	-0.2703432	-0.1448267	-0.01931023	-0.05068936	-0.01931023	0.2003437	0.6082723	0.4513767
A50CHIN.H23	2.2532028	-0.4096732	-0.4096732	-0.4096732	-0.4096732	-0.4096732	-0.4096732	-0.40967325	-0.40967325	-0.40967325	-0.4096732	2.2532028	-0.4096732
A50CHIN.H27	-0.4510584	-1.1025872	-0.7768228	-0.4510584	-0.1252940	2.1550568	0.5262348	1.17776363	0.52623481	0.85199922	-0.1252940	-1.1025872	-1.1025872
AASIA	0.9721710	-0.8882780	-0.8882780	1.1004778	0.9721710	-0.8882780	-0.8882780	0.97217098	-0.88827805	-0.88827805	-0.8882780	1.0363244	1.1646312

Figure 2.3.2: Stock Data With Z-Normalization Using R

c. Carry out PAA algorithm on the dataset and we will get the result as follows:



	V1	V2	V3
A50CHIN.C22	0.47310564	-0.44657635	-0.02652929
A50CHIN.C24	-0.14658067	-0.13011093	0.27669160
A50CHIN.C26	-0.14237341	-0.21601483	0.35838825
A50CHIN.C28	-0.09661093	-0.12203486	0.21864580
A50CHIN.C30	1.06826250	-0.93655890	-0.13170360
A50CHIN.C32	0.59491480	-0.09915247	-0.49576233
A50CHIN.C34	0.22340565	0.30087697	-0.52428263
A50CHIN.C36	-0.22448144	-0.05793069	0.28241214
A50CHIN.H23	0.20483662	-0.40967325	0.20483662
A50CHIN.H27	-0.65152881	0.95223442	-0.30070561
AASIA	0.14311146	-0.17272073	0.02960927

Figure 2.3.3: Stock Data With PAA Algorithm Results Performed Using R

2.4 Visualisation

Results from PAA is visualised as follows:

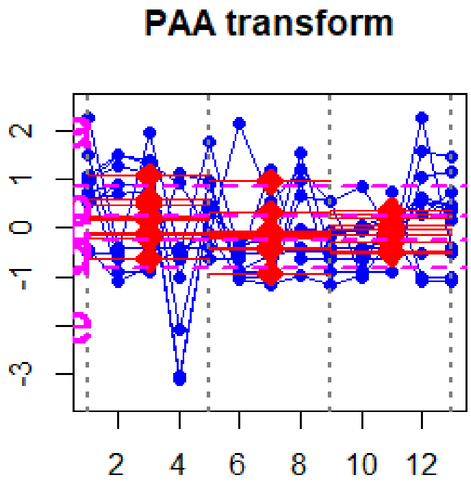
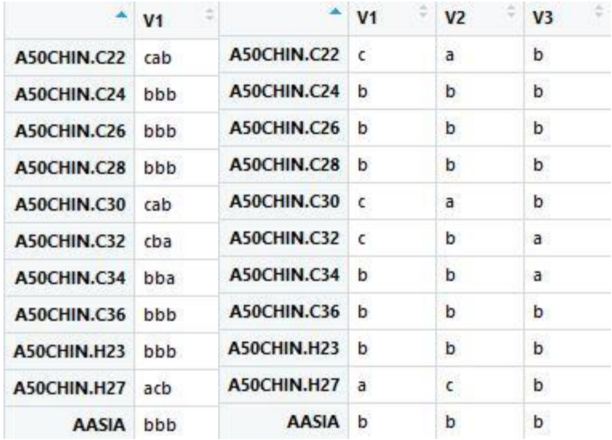


Figure 2.4.1: PAA Graph From R

Note that the PAA figure does not shows all data due to large amount of unique stock

3.0 Discussion

During the execution of SAX and PAA, we also did the PAA path output, which will be shown in Figure 3.0.1 below:



	V1
A50CHIN.C22	cab
A50CHIN.C24	bbb
A50CHIN.C26	bbb
A50CHIN.C28	bbb
A50CHIN.C30	cab
A50CHIN.C32	cba
A50CHIN.C34	bba
A50CHIN.C36	bbb
A50CHIN.H23	bbb
A50CHIN.H27	acb
AASIA	bbb

	V1	V2	V3
A50CHIN.C22	c	a	b
A50CHIN.C24	b	b	b
A50CHIN.C26	b	b	b
A50CHIN.C28	b	b	b
A50CHIN.C30	c	a	b
A50CHIN.C32	c	b	a
A50CHIN.C34	b	b	a
A50CHIN.C36	b	b	b
A50CHIN.H23	b	b	b
A50CHIN.H27	a	c	b
AASIA	b	b	b

Figure 3.0.1: PAA Path Output Using R

Then SAX is the difference value, where SAX will transform the discrete variables into strings. An example of how to determine the SAX distance in this assignment will be shown below.

Determine the SAX distance between AASIA and A50CHIN:

a. SAX transform strings through 3 PAA points for AASIA : b b b, same goes to A50CHIN
b. SAX distance is calculated by adding the the gap distance between each PAA points in the y-axis as shown in Figure 4.2, C30 = 0.5 + 0.3 +0.3 = 0.5 where C30 is the gap distance.

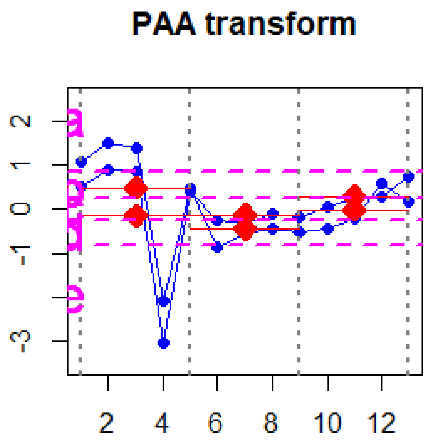


Figure 3.0.2: Visualised PAA Graph Using R

4.0 Contribution

Group Member	Contribution
Teng Lung Yun	- Crawl data for 13 days, do PCA code and covariance
Choy Siew Wearn	- Clean data, do PAA and SAX, provide brief report of PAA and SAX
Wo Choy Yun	- Compile, elaborate and completing full report, run through PAA, PCA and covariance code
Zhu Ting	- Assist in report completion
Xu Xiang	- Assist in report completion

5.0 Reference

1. Symbolic Aggregate Approximation. Retrieved from: https://jmotif.github.io/sax-vsm_site/morea/algorithm/SAX.html

2. Piecewise Aggregate Approximation of time series. Retrieved from: https://jmotif.github.io/sax-vsm_site/morea/algorithm/PAA.html

3. Symbolic Aggregate Approximation. Retrieved from: http://www.cs.ucr.edu/~eamonn/SAX.htm

4. A Beginner's Guide To Eigenvectors, Eigenvalues, Pca, Covariance and Entropy Retrieved from: https://skymind.ai/wiki/eigenvector

5. Principal Component Analysis Retrieved from:http://www.cse.psu.edu/~rtc12/CSE586/lectures/pcaLectureShort_6pp.pdf

6. Variance-Covariance Matrix: Stock Price Analysis in R (corpcor, covmat) Retrieved from: https://www.michael-grogan.com/variance-covariance-matrix-calculation-r/