RMIT UNIVERSITY MELBOURNE
MATH1309 - MULTIVARIATE ANALYSIS

Assignment 2

Udeshika Dissanayake (S3400652)
Udeshika.dissanayake@student.rmit.edu.au

Table of Contents

Question1

1.1. Mean and standard deviation for the 13 chemical concentrations

The MEANS Procedure

Variable	Mean	Std Dev
chem1	13.0006	0.8118
chem2	2.3363	1.1171
chem3	2.3665	0.2743
chem4	19.4949	3.3396
chem5	99.7416	14.2825
chem6	2.2951	0.6259
chem7	2.0293	0.9989
chem8	0.3619	0.1245
chem9	1.5909	0.5724
chem10	5.0581	2.3183
chem11	0.9574	0.2286
chem12	2.6117	0.7100
chem13	746.8933	314.9075

1.2. Correlation matrix and a scatterplot

The CORR Procedure

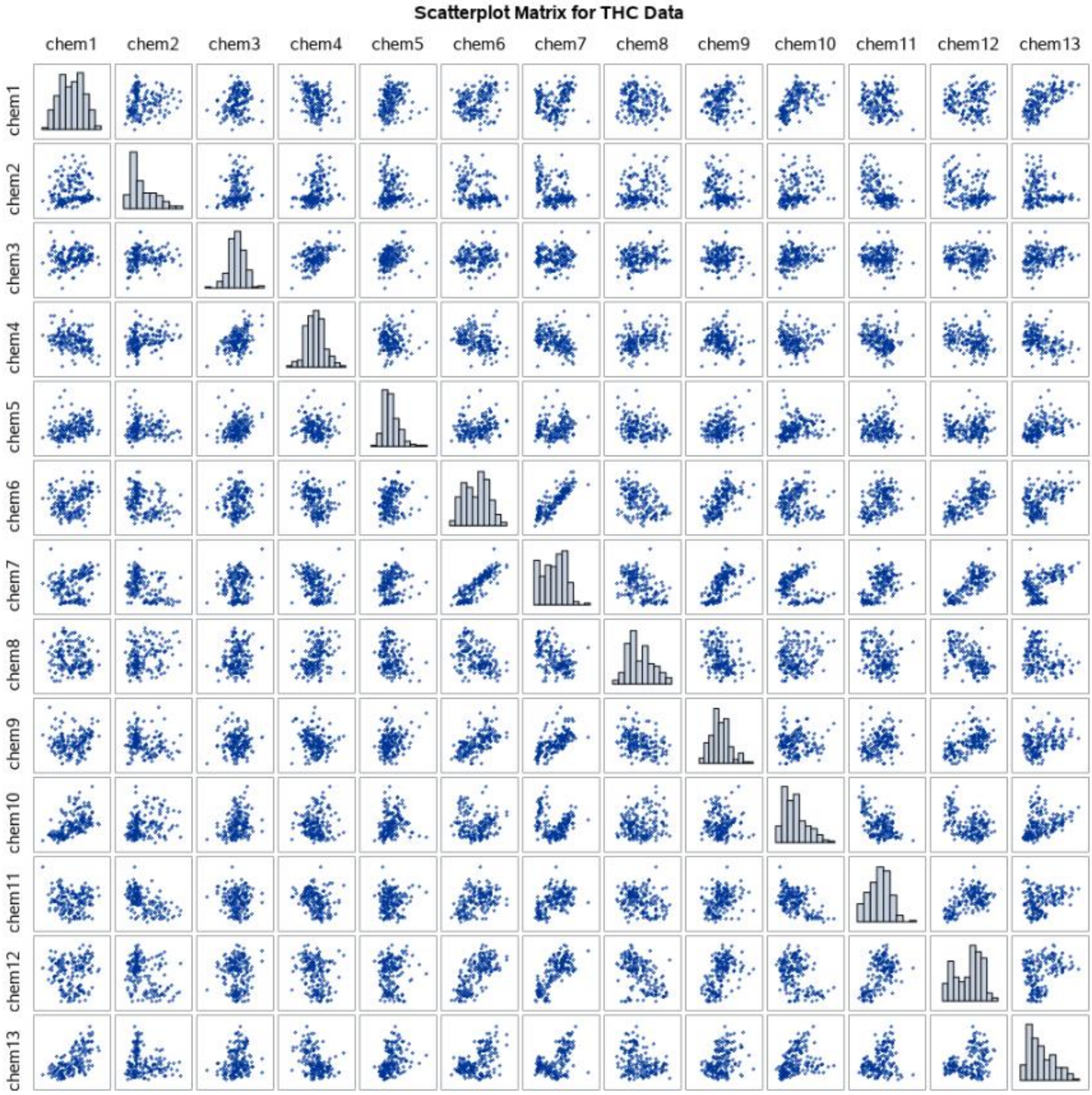
13 Variables:	chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
chem1	178	13.00062	0.81183	2314	11.03000	14.83000
chem2	178	2.33635	1.11715	415.87000	0.74000	5.80000
chem3	178	2.36652	0.27434	421.24000	1.36000	3.23000
chem4	178	19.49494	3.33956	3470	10.60000	30.00000
chem5	178	99.74157	14.28248	17754	70.00000	162.00000
chem6	178	2.29511	0.62585	408.53000	0.98000	3.88000
chem7	178	2.02927	0.99886	361.21000	0.34000	5.08000
chem8	178	0.36185	0.12445	64.41000	0.13000	0.66000
chem9	178	1.59090	0.57236	283.18000	0.41000	3.58000
chem10	178	5.05809	2.31829	900.34000	1.28000	13.00000
chem11	178	0.95745	0.22857	170.42600	0.48000	1.71000
chem12	178	2.61169	0.70999	464.88000	1.27000	4.00000
chem13	178	746.89326	314.90747	132947	278.00000	1680

Pearson Correlation Coefficients, N = 178

	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
chem1	1.00000	0.09440	0.21154	-0.31024	0.27080	0.28910	0.23681	-0.15593	0.13670	0.54636	-0.07175	0.07234	0.64372
chem2	0.09440	1.00000	0.16405	0.28850	-0.05458	-0.33517	-0.41101	0.29298	-0.22075	0.24899	-0.56130	-0.36871	-0.19201
chem3	0.21154	0.16405	1.00000	0.44337	0.28659	0.12898	0.11508	0.18623	0.00965	0.25889	-0.07467	0.00391	0.22363
chem4	-0.31024	0.28850	0.44337	1.00000	-0.08333	-0.32111	-0.35137	0.36192	-0.19733	0.01873	-0.27396	-0.27677	-0.44060
chem5	0.27080	-0.05458	0.28659	-0.08333	1.00000	0.21440	0.19578	-0.25629	0.23644	0.19995	0.05540	0.06600	0.39335
chem6	0.28910	-0.33517	0.12898	-0.32111	0.21440	1.00000	0.86456	-0.44994	0.61241	-0.05514	0.43368	0.69995	0.49811
chem7	0.23681	-0.41101	0.11508	-0.35137	0.19578	0.86456	1.00000	-0.53790	0.65269	-0.17238	0.54348	0.78719	0.49419
chem8	-0.15593	0.29298	0.18623	0.36192	-0.25629	-0.44994	-0.53790	1.00000	-0.36585	0.13906	-0.26264	-0.50327	-0.31139
chem9	0.13670	-0.22075	0.00965	-0.19733	0.23644	0.61241	0.65269	-0.36585	1.00000	-0.02525	0.29554	0.51907	0.33042
chem10	0.54636	0.24899	0.25889	0.01873	0.19995	-0.05514	-0.17238	0.13906	-0.02525	1.00000	-0.52181	-0.42881	0.31610
chem11	-0.07175	-0.56130	-0.07467	-0.27396	0.05540	0.43368	0.54348	-0.26264	0.29554	-0.52181	1.00000	0.56547	0.23618
chem12	0.07234	-0.36871	0.00391	-0.27677	0.06600	0.69995	0.78719	-0.50327	0.51907	-0.42881	0.56547	1.00000	0.31276
chem13	0.64372	-0.19201	0.22363	-0.44060	0.39335	0.49811	0.49419	-0.31139	0.33042	0.31610	0.23618	0.31276	1.00000



The correlation matrix is suitable for principal component analysis as there is a high correlation between some variables. E.g.

- chem1 and chem13→ 64 %
- chem2 and chem11→ -56 %
- chem6 and chem7→ 86 %
- chem6 and chem9→ 61 %
- chem6 and chem12→ 70 %
- chem7 and chem9→ 65 %
- chem7 and chem12→ 79 %

This means some of the variables can safely be ignored and entire data set can effectively be summarized by a fewer numbers of variables. i.e. principle components.

2.3. Principal component analysis on the raw data

The PRINCOMP Procedure

Observations	178
Variables	13

Simple Statistics

	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
Mean	13.00061798	2.336348315	2.366516854	19.49494382	99.74157303	2.295112360	2.029269663	0.3618539326	1.590898876	5.058089882	0.9574494382	2.611685393	746.8932584
StD	0.81182654	1.117146098	0.274344009	3.33956377	14.28248352	0.625851049	0.998858685	0.1244533403	0.572358863	2.318285872	0.2285715658	0.709990429	314.9074743

Covariance Matrix

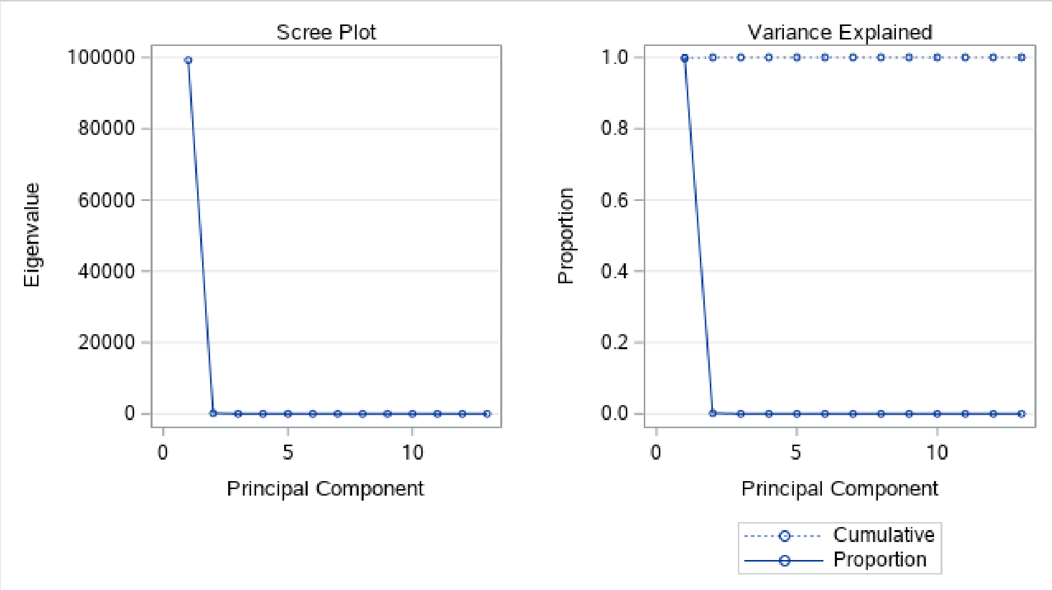
	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
chem1	0.65906	0.08561	0.04712	-0.84109	3.13988	0.14689	0.19203	-0.01575	0.06352	1.02828	-0.01331	0.04170	164.56718
chem2	0.08561	1.24802	0.05028	1.07633	-0.87078	-0.23434	-0.45863	0.04073	-0.14115	0.64484	-0.14333	-0.29245	-67.54887
chem3	0.04712	0.05028	0.07526	0.40621	1.12294	0.02215	0.03153	0.00636	0.00152	0.16465	-0.00468	0.00076	19.31974
chem4	-0.84109	1.07633	0.40621	11.15269	-3.97476	-0.67115	-1.17208	0.15042	-0.37718	0.14502	-0.20912	-0.65623	-463.35535
chem5	3.13988	-0.87078	1.12294	-3.97476	203.98934	1.91647	2.79309	-0.45556	1.93283	6.62052	0.18085	0.66931	1769.15870
chem6	0.14689	-0.23434	0.02215	-0.67115	1.91647	0.39169	0.54047	-0.03505	0.21937	-0.08000	0.06204	0.31102	98.17106
chem7	0.19203	-0.45863	0.03153	-1.17208	2.79309	0.54047	0.99772	-0.06687	0.37315	-0.39917	0.12408	0.55826	155.44749
chem8	-0.01575	0.04073	0.00636	0.15042	-0.45556	-0.03505	-0.06687	0.01549	-0.02606	0.04012	-0.00747	-0.04447	-12.20359
chem9	0.06352	-0.14115	0.00152	-0.37718	1.93283	0.21937	0.37315	-0.02606	0.32759	-0.03350	0.03866	0.21093	59.55433
chem10	1.02828	0.64484	0.16465	0.14502	6.62052	-0.08000	-0.39917	0.04012	-0.03350	5.37445	-0.27651	-0.70581	230.76748
chem11	-0.01331	-0.14333	-0.00468	-0.20912	0.18085	0.06204	0.12408	-0.00747	0.03866	-0.27651	0.05224	0.09177	17.00022
chem12	0.04170	-0.29245	0.00076	-0.65623	0.66931	0.31102	0.55826	-0.04447	0.21093	-0.70581	0.09177	0.50409	69.92753
chem13	164.56718	-67.54887	19.31974	-463.35535	1769.15870	98.17106	155.44749	-12.20359	59.55433	230.76748	17.00022	69.92753	99166.71736

Total Variance	99391.504992

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	99201.7895	99029.2543	0.9981	0.9981
2	172.5353	163.0972	0.0017	0.9998
3	9.4381	4.4469	0.0001	0.9999
4	4.9912	3.7623	0.0001	1.0000
5	1.2288	0.3878	0.0000	1.0000
6	0.8411	0.5621	0.0000	1.0000
7	0.2790	0.1276	0.0000	1.0000
8	0.1514	0.0393	0.0000	1.0000
9	0.1121	0.0404	0.0000	1.0000
10	0.0717	0.0341	0.0000	1.0000
11	0.0376	0.0165	0.0000	1.0000
12	0.0211	0.0129	0.0000	1.0000
13	0.0082		0.0000	1.0000

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
chem1	0.001659	0.001203	0.016874	0.141447	0.020337	0.194120	0.923280	0.284821	-.086601	-.002245	-.014972	-.015651	0.008029
chem2	-.000681	0.002155	0.122003	0.160390	-.612883	0.742473	-.150110	-.064674	-.015662	-.018509	-.023188	0.067296	-.011090
chem3	0.000195	0.004594	0.051987	-.009773	0.020176	0.041753	0.045010	-.149340	-.073650	-.086800	0.954011	-.132063	-.173686
chem4	-.004671	0.026450	0.938593	-.330965	0.064352	-.024065	0.031527	0.015154	-.002045	0.003554	-.052822	0.005394	0.001940
chem5	0.017868	0.999344	-.029780	-.005394	-.006149	-.001924	0.001797	-.003552	0.001964	-.000041	-.003025	0.000621	0.002285
chem6	0.000990	0.000878	-.040485	-.074585	0.315245	0.278717	-.020186	-.177238	-.255673	0.847195	0.008802	0.003883	-.026691
chem7	0.001567	-.000052	-.085443	-.169087	0.524761	0.433598	-.038869	-.248117	-.378307	-.520138	-.133205	-.037488	0.069599
chem8	-.000123	-.001354	0.013511	0.010806	-.029648	-.021953	-.004665	0.006498	-.036752	0.037713	0.199179	0.147552	0.966466
chem9	0.000601	0.005004	-.024659	-.050121	0.251183	0.241884	-.309799	0.870433	0.051520	0.009723	0.135621	-.013119	-.017604
chem10	0.002327	0.015100	0.291398	0.878894	0.331747	0.002740	-.112837	-.081287	0.099029	-.023147	-.009820	0.050356	-.004633
chem11	0.000171	-.000763	-.025978	-.060035	0.051524	-.023776	0.030820	-.002952	-.033065	-.038470	0.097511	0.975562	-.166551
chem12	0.000705	-.003495	-.070324	-.178200	0.260639	0.288913	0.101974	-.186715	0.873747	0.017017	0.028485	0.011630	0.044192
chem13	0.999823	-.017774	0.004529	-.003113	-.002299	-.001212	-.001076	0.000010	0.000073	0.000049	-.000240	-.000100	0.000036



1.3.a)

In total 99.99% of the total sample variation is accounted in first 3 PC's.

1.3.b)

First PC is having eigenvalue of 99201.7895 and it explains 99.81% of variation. There is high correlation between first PC and Chem13 ; which is 99.98%.

Second PC is having eigenvalue of 172.5353 and it explains 0.17% of variation. There is high correlation between second PC and Chem5 ; which is 99.93%.

Third PC is having eigenvalue of 9.4381 and it explains 0.01% of variation. There is high correlation between third PC and Chem4 ; which is 93.86%. and between third PC and Chem10 ; which is 29.13%.

1.3.c)

$$PC1 = Y_1 = 0.001659X_1 - 0.000681X_2 + 0.000195X_3 - 0.004671X_4 + 0.017868X_5 + 0.00099X_6 + 0.001567X_7 - 0.000123X_8 + 0.000601X_9 + 0.002327X_{10} + 0.000171X_{11} + 0.000705X_{12} + 0.999823X_{13}$$

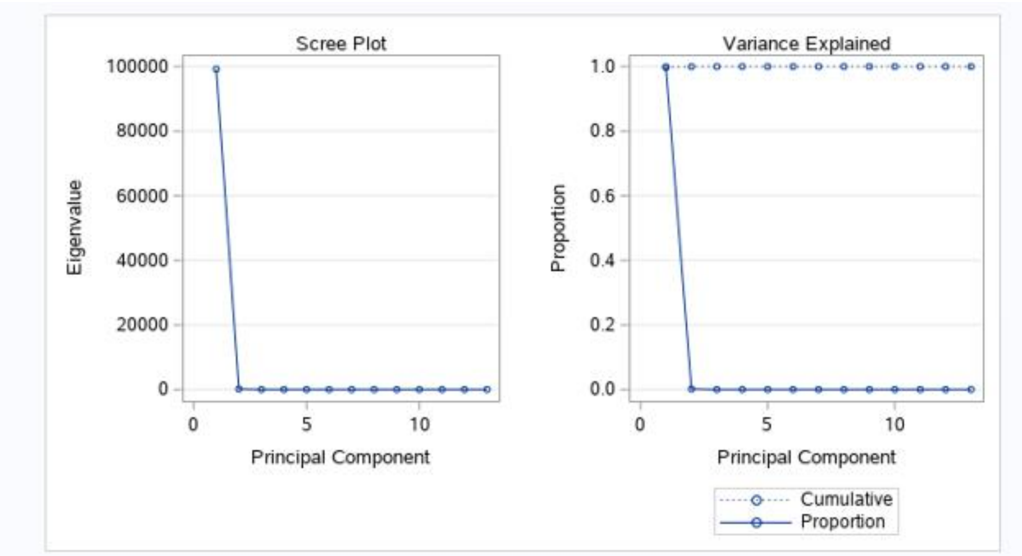
$$PC2 = Y_2 = 0.001203X_1 + 0.002155X_2 + 0.004594X_3 + 0.02645X_4 + 0.999344X_5 + 0.000878X_6 - 0.000052X_7 - 0.001354X_8 + 0.005004X_9 + 0.0151X_{10} - 0.000763X_{11} - 0.003495X_{12} - 0.017774X_{13}$$

$$PC3 = Y_3 = 0.016874X_1 + 0.122003X_2 + 0.051987X_3 + 0.938593X_4 - 0.02978X_5 - 0.040485X_6 - 0.085443X_7 + 0.013511X_8 - 0.024659X_9 + 0.291398X_{10} - 0.025978X_{11} - 0.070324X_{12} + 0.004529X_{13}$$

1.3.d)

Due to large variance, Chem13 completely dominates the first principle component in above covariance calculation. Moreover, the first principle component explains 99.81% of the total population variance. This means even though the dataset could be effectively summarized by a fewer variables through PC analysis, the covariance matrix approach is not suitable due to non-normalized behavior of variables, specially chem13.

1.3.e)



As per the scree plot, markers for component 2-13 are linear and first two PC's explain 99.99% of the variance. Hence two components will be retained.

2.4. Principal component analysis on the correlation matrix

The PRINCOMP Procedure

Observations	178
Variables	13

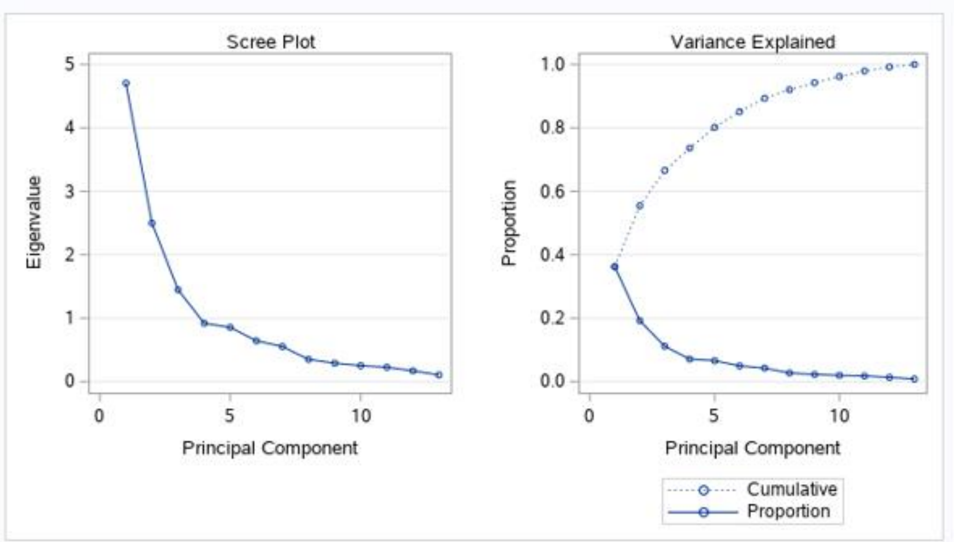
Simple Statistics

	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
Mean	13.00061798	2.336348315	2.366516854	19.49494382	99.74157303	2.295112360	2.029269663	0.3618539326	1.590898876	5.058089882	0.9574494382	2.611685393	746.8932584
StD	0.81182654	1.117146098	0.274344009	3.33956377	14.28248352	0.625851049	0.998858685	0.1244533403	0.572358863	2.318285872	0.2285715658	0.709990429	314.9074743

Correlation Matrix													
	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
chem1	1.0000	0.0944	0.2115	-.3102	0.2708	0.2891	0.2368	-.1559	0.1367	0.5464	-.0717	0.0723	0.6437
chem2	0.0944	1.0000	0.1640	0.2885	-.0546	-.3352	-.4110	0.2930	-.2207	0.2490	-.5613	-.3687	-.1920
chem3	0.2115	0.1640	1.0000	0.4434	0.2866	0.1290	0.1151	0.1862	0.0097	0.2589	-.0747	0.0039	0.2236
chem4	-.3102	0.2885	0.4434	1.0000	-.0833	-.3211	-.3514	0.3619	-.1973	0.0187	-.2740	-.2768	-.4406
chem5	0.2708	-.0546	0.2866	-.0833	1.0000	0.2144	0.1958	-.2563	0.2364	0.2000	0.0554	0.0660	0.3934
chem6	0.2891	-.3352	0.1290	-.3211	0.2144	1.0000	0.8646	-.4499	0.6124	-.0551	0.4337	0.6999	0.4981
chem7	0.2368	-.4110	0.1151	-.3514	0.1958	0.8646	1.0000	-.5379	0.6527	-.1724	0.5435	0.7872	0.4942
chem8	-.1559	0.2930	0.1862	0.3619	-.2563	-.4499	-.5379	1.0000	-.3658	0.1391	-.2626	-.5033	-.3114
chem9	0.1367	-.2207	0.0097	-.1973	0.2364	0.6124	0.6527	-.3658	1.0000	-.0252	0.2955	0.5191	0.3304
chem10	0.5464	0.2490	0.2589	0.0187	0.2000	-.0551	-.1724	0.1391	-.0252	1.0000	-.5218	-.4288	0.3161
chem11	-.0717	-.5613	-.0747	-.2740	0.0554	0.4337	0.5435	-.2626	0.2955	-.5218	1.0000	0.5655	0.2362
chem12	0.0723	-.3687	0.0039	-.2768	0.0660	0.6999	0.7872	-.5033	0.5191	-.4288	0.5655	1.0000	0.3128
chem13	0.6437	-.1920	0.2236	-.4406	0.3934	0.4981	0.4942	-.3114	0.3304	0.3161	0.2362	0.3128	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.70585025	2.20887652	0.3620	0.3620
2	2.49697373	1.05090176	0.1921	0.5541
3	1.44607197	0.52709805	0.1112	0.6653
4	0.91897392	0.06574575	0.0707	0.7360
5	0.85322818	0.21157115	0.0656	0.8016
6	0.64165703	0.09062872	0.0494	0.8510
7	0.55102831	0.20253095	0.0424	0.8934
8	0.34849736	0.05961742	0.0268	0.9202
9	0.28887994	0.03797746	0.0222	0.9424
10	0.25090248	0.02511384	0.0193	0.9617
11	0.22578864	0.05701840	0.0174	0.9791
12	0.16877023	0.06539230	0.0130	0.9920
13	0.10337794		0.0080	1.0000

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
chem1	0.144329	0.483652	-.207383	-.017856	0.265664	0.213539	-.056396	0.396139	0.508619	-.211605	-.225917	-.266286	0.014970
chem2	-.245188	0.224931	0.089013	0.536890	-.035214	0.536814	0.420524	0.065827	-.075283	0.309080	0.076486	0.121696	0.025964
chem3	-.002051	0.316069	0.626224	-.214176	0.143025	0.154475	-.149171	-.170260	-.307694	0.027125	-.498691	-.049622	-.141218
chem4	-.239320	-.010591	0.612080	0.060859	-.066103	-.100825	-.286969	0.427970	0.200449	-.052799	0.479314	-.055743	0.091683
chem5	0.141992	0.299634	0.130757	-.351797	-.727049	0.038144	0.322883	-.156361	0.271403	-.067870	0.071289	0.062220	0.056774
chem6	0.394661	0.065040	0.146179	0.198068	0.149318	-.084122	-.027925	-.405934	0.286035	0.320131	0.304341	-.303882	-.463908
chem7	0.422934	-.003360	0.150682	0.152295	0.109026	-.018920	-.060685	-.187245	0.049578	0.163151	-.025694	-.042899	0.832257
chem8	-.298533	0.028779	0.170368	-.203301	0.500703	-.258594	0.595447	-.233285	0.195501	-.215535	0.116896	0.042352	0.114040
chem9	0.313429	0.039302	0.149454	0.399057	-.136860	-.533795	0.372139	0.368227	-.209145	-.134184	-.237363	-.095553	-.116917
chem10	-.088617	0.529996	-.137306	0.065926	0.076437	-.418644	-.227712	-.033797	0.056218	0.290775	0.031839	0.604222	-.011993
chem11	0.296715	-.279235	0.085222	-.427771	0.173615	0.105983	0.232076	0.436624	0.085828	0.522399	-.048212	0.259214	-.089689
chem12	0.376167	-.164496	0.166005	0.184121	0.101161	0.265851	-.044764	-.078108	0.137227	-.523706	0.046423	0.600959	-.156718
chem13	0.286752	0.364903	-.126746	-.232071	0.157869	0.119726	0.076805	0.120023	-.575786	-.162116	0.539270	-.079402	0.014447



1.4.a)

In total 66.53% of the total sample variation is accounted for first 3 PC's.

1.4.b)

First PC is having eigenvalue of 4.70585 and it explains 36.2% of variation. There is high correlation between first PC and Chem7; which is 42.29%.

Second PC is having eigenvalue of 2.497 and it explains 19.21% of variation. There is high correlation between second PC and Chem10 ; which is 53%.

Third PC is having eigenvalue of 1.446 and it explains 11.1% of variation. There is high correlation between third PC and Chem3; which is 62.62%. and between third PC and Chem4; which is 61.4%

1.4.c)

$$PC1=Y_1 = 0.144329\left(\frac{X_1-\mu_1}{\sqrt{0.65906}}\right) - 0.245188\left(\frac{X_2-\mu_2}{\sqrt{1.24802}}\right) - 0.002051\left(\frac{X_3-\mu_3}{\sqrt{0.07526}}\right) - 0.23932\left(\frac{X_4-\mu_4}{\sqrt{11.15269}}\right) + 0.141992\left(\frac{X_5-\mu_5}{\sqrt{203.98}}\right) + 0.394661\left(\frac{X_6-\mu_6}{\sqrt{0.39169}}\right) + 0.422934\left(\frac{X_7-\mu_7}{\sqrt{0.99772}}\right) - 0.298533\left(\frac{X_8-\mu_8}{\sqrt{0.01549}}\right) + 0.313429\left(\frac{X_9-\mu_9}{\sqrt{0.32759}}\right) - 0.088617\left(\frac{X_{10}-\mu_{10}}{\sqrt{5.37445}}\right) + 0.296715\left(\frac{X_{11}-\mu_{11}}{\sqrt{0.05224}}\right) + 0.376167\left(\frac{X_{12}-\mu_{12}}{\sqrt{0.50409}}\right) + 0.286752\left(\frac{X_{13}-\mu_{13}}{\sqrt{99166.71}}\right)$$

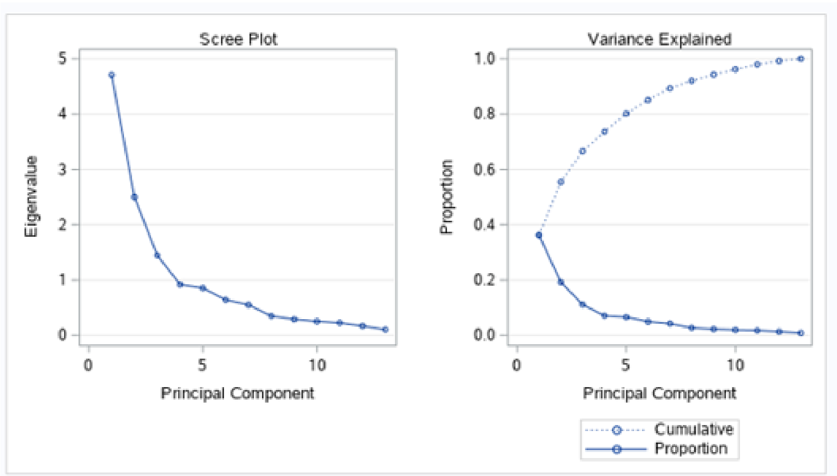
$$PC2 = Y_2 = 0.483652\left(\frac{X_1-\mu_1}{\sqrt{0.65906}}\right) + 0.224931\left(\frac{X_2-\mu_2}{\sqrt{1.24802}}\right) + 0.3160694\left(\frac{X_3-\mu_3}{\sqrt{0.07526}}\right) - 0.010591\left(\frac{X_4-\mu_4}{\sqrt{11.15269}}\right) + 0.299634\left(\frac{X_5-\mu_5}{\sqrt{203.98}}\right) + 0.06504\left(\frac{X_6-\mu_6}{\sqrt{0.39169}}\right) - 0.00336\left(\frac{X_7-\mu_7}{\sqrt{0.99772}}\right) + 0.028779\left(\frac{X_8-\mu_8}{\sqrt{0.01549}}\right) + 0.039302\left(\frac{X_9-\mu_9}{\sqrt{0.32759}}\right) + 0.529996\left(\frac{X_{10}-\mu_{10}}{\sqrt{5.37445}}\right) - 0.279235\left(\frac{X_{11}-\mu_{11}}{\sqrt{0.05224}}\right) - 0.164496\left(\frac{X_{12}-\mu_{12}}{\sqrt{0.50409}}\right) + 0.364903\left(\frac{X_{13}-\mu_{13}}{\sqrt{99166.71}}\right)$$

$$PC3 = Y_3 = -0.207383\left(\frac{X_1 - \mu_1}{\sqrt{0.65906}}\right) + 0.089013\left(\frac{X_2 - \mu_2}{\sqrt{1.24802}}\right) + 0.626224\left(\frac{X_3 - \mu_3}{\sqrt{0.07526}}\right)$$
$$+ 0.61208\left(\frac{X_4 - \mu_4}{\sqrt{11.15269}}\right) + 0.130757\left(\frac{X_5 - \mu_5}{\sqrt{203.98}}\right) + 0.146179\left(\frac{X_6 - \mu_6}{\sqrt{0.39169}}\right)$$
$$+ 0.150682\left(\frac{X_7 - \mu_7}{\sqrt{0.99772}}\right) + 0.170368\left(\frac{X_8 - \mu_8}{\sqrt{0.01549}}\right) + 0.149454\left(\frac{X_9 - \mu_9}{\sqrt{0.32759}}\right)$$
$$- 0.137306\left(\frac{X_{10} - \mu_{10}}{\sqrt{5.37445}}\right) + 0.085222\left(\frac{X_{11} - \mu_{11}}{\sqrt{0.05224}}\right) + 0.166005\left(\frac{X_{12} - \mu_{12}}{\sqrt{0.50409}}\right)$$
$$- 0.126746\left(\frac{X_{13} - \mu_{13}}{\sqrt{99166.71}}\right)$$

1.4.d)

The first four principle components account for 73% of the total population variance in the data set. This means, the first four principle components could replace the original 13 variables with significantly less loss of information.

1.4.e)



As per the scree plot, 4 components will be retained.

Question2

2.1 Dataset

Obs	Population	School	Employment	Services	HouseValue
1	5700	12.8	2500	270	25000
2	1000	10.9	600	10	10000
3	3400	8.8	1000	10	9000
4	3800	13.6	1700	140	25000
5	4000	12.8	1600	140	25000
6	8200	8.3	2600	60	12000
7	1200	11.4	400	10	16000
8	9100	11.5	3300	60	14000
9	9900	12.5	3400	180	18000
10	9600	13.7	3600	390	25000
11	9600	9.6	3300	80	12000
12	9400	11.4	4000	100	13000

2.2 Mean and Standard Deviation of the data

The MEANS Procedure

Variable	Mean	Std Dev
Population	6241.6667	3439.9943
School	11.4417	1.7865
Employment	2333.3333	1241.2115
Services	120.8333	114.9275
HouseValue	17000.0000	6367.5313

2.3 Factor Analysis on the raw data and the correlation matrix

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	12
Number of Records Used	12
N for Significance Tests	12

Means and Standard Deviations from 12 Observations		
Variable	Mean	Std Dev
Population	6241.667	3439.9943
School	11.442	1.7865
Employment	2333.333	1241.2115
Services	120.833	114.9275
HouseValue	17000.000	6367.5313

Correlations					
	Population	School	Employment	Services	HouseValue
Population	1.00000	0.00975	0.97245	0.43887	0.02241
School	0.00975	1.00000	0.15428	0.69141	0.86307
Employment	0.97245	0.15428	1.00000	0.51472	0.12193
Services	0.43887	0.69141	0.51472	1.00000	0.77765
HouseValue	0.02241	0.86307	0.12193	0.77765	1.00000

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

2 factors will be retained by the MINEIGEN criterion.

Factor Pattern		
	Factor1	Factor2
Population	0.58096	0.80642
School	0.76704	-0.54476
Employment	0.67243	0.72605
Services	0.93239	-0.10431
HouseValue	0.79116	-0.55818

Variance Explained by Each Factor	
Factor1	Factor2
2.8733136	1.7966601

Final Communality Estimates: Total = 4.669974				
Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

2.4

2.4.a)

First 2 factors explain 93.4% of the variance. Hence first two factors provide an adequate summary of the data.

2.4.b)

First 2 factors explain 93.4% of the variance.

2.4.c)

First 3 factors explain 97.7% of the variance

2.5. Scoring coefficient as eigenvectors

The PRINCOMP Procedure

Observations	12
Variables	5

Simple Statistics

	Population	School	Employment	Services	HouseValue
Mean	6241.666667	11.44166667	2333.333333	120.8333333	17000.00000
StD	3439.994274	1.78654483	1241.211529	114.9275134	6367.53128

Correlation Matrix

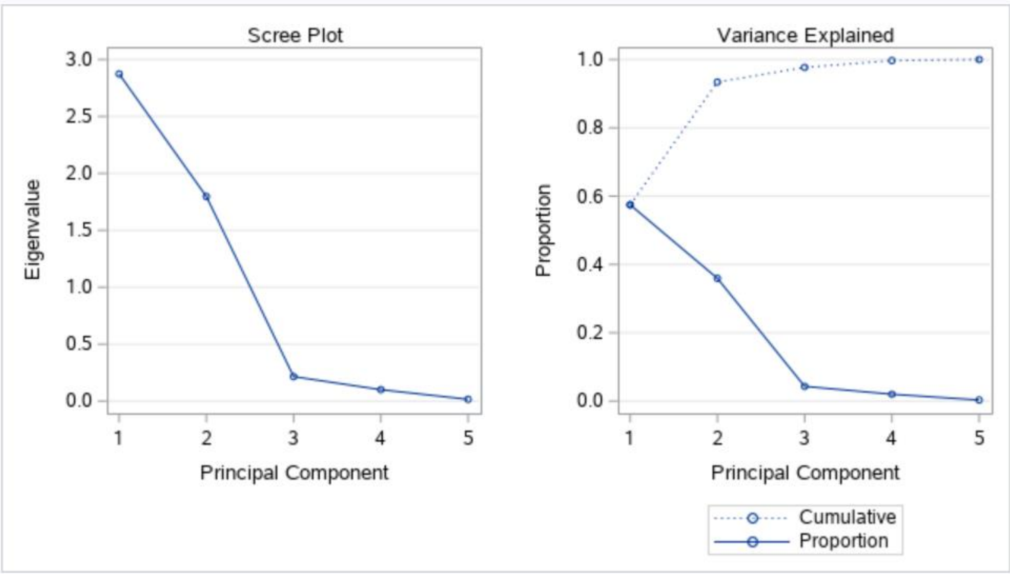
	Population	School	Employment	Services	HouseValue
Population	1.0000	0.0098	0.9724	0.4389	0.0224
School	0.0098	1.0000	0.1543	0.6914	0.8631
Employment	0.9724	0.1543	1.0000	0.5147	0.1219
Services	0.4389	0.6914	0.5147	1.0000	0.7777
HouseValue	0.0224	0.8631	0.1219	0.7777	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425



2.5.a) Eigen values and Eigen Vectors

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425

2.5.b)

The first and the second component account for 0.57 (57%) and 0.36 (36%) proportions of variance, respectively.

2.5.c)

First and second factors together account for the 93.4% of the standardized variance.

2.5.d)

Final communality estimates represent the proportion of each variable's variance that can be explained by the retained factors. As can be seen from the final communality estimate values (close to 1 for all variables) in this analysis for all the variables, it can be claimed that all the variables are well accounted by the retained factors.

Final Community Estimates: Total = 4.669974				
Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

2.6. Component scores as linear combination of the observed variables

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	12
Number of Records Used	12
N for Significance Tests	12

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

5 factors will be retained by the NFACTOR criterion.

Factor Pattern					
	Factor1	Factor2	Factor3	Factor4	Factor5
Population	0.58096	0.80642	0.02759	0.06450	0.08516
School	0.76704	-0.54476	0.31927	-0.11177	0.02160
Employment	0.67243	0.72605	0.11493	0.00725	-0.08621
Services	0.93239	-0.10431	-0.30780	-0.15818	-0.00002
HouseValue	0.79116	-0.55818	-0.06473	0.24126	-0.01018

Variance Explained by Each Factor				
Factor1	Factor2	Factor3	Factor4	Factor5
2.8733136	1.7966601	0.2148369	0.0999341	0.0152554

Final Communality Estimates: Total = 5.000000				
Population	School	Employment	Services	HouseValue
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000

The FACTOR Procedure
Initial Factor Method: Principal Components

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor				
Factor1	Factor2	Factor3	Factor4	Factor5
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000

Standardized Scoring Coefficients					
	Factor1	Factor2	Factor3	Factor4	Factor5
Population	0.20219	0.44884	0.12841	0.64542	5.58240
School	0.26695	-0.30320	1.48612	-1.11846	1.41574
Employment	0.23403	0.40411	0.53496	0.07256	-5.65135
Services	0.32450	-0.05806	-1.43273	-1.58288	-0.00100
HouseValue	0.27535	-0.31068	-0.30129	2.41419	-0.66734

2.6.a)

$FC_1 = 0.20219X_1 + 0.26695X_2 + 0.23403X_3 + 0.3245X_4 + 0.27535X_5$

2.6.b) FC using standardized scoring coefficients

$FC_2 = 0.44884X_1 - 0.3032X_2 + 0.40411X_3 - 0.05806X_4 - 0.31068X_5$

2.6.c) PC using Eigenvectors

$PC_1 = 0.34273X_1 + 0.452507X_2 + 0.396695X_3 + 0.550057X_4 + 0.466738X_5$

$PC_2 = 0.601629X_1 - .406414X_2 + 0.541665X_3 - .077817X_4 - 0.416429X_5$

Question3

3.1. Sample from Dataset

Obs	Class	Length	Left	Right	Bottom	Top	Diagonal
1	genuine	214.8	131.0	131.1	9.0	9.7	141.0
2	genuine	214.6	129.7	129.7	8.1	9.5	141.7
3	genuine	214.8	129.7	129.7	8.7	9.6	142.2
4	genuine	214.8	129.7	129.6	7.5	10.4	142.0
5	genuine	215.0	129.6	129.7	10.4	7.7	141.8
6	genuine	215.7	130.8	130.5	9.0	10.1	141.4
7	genuine	215.5	129.5	129.7	7.9	9.6	141.6

3.2 Mean and variance-covariance matrix for genuine notes

Class=genuine _TYPE_=COV

NAME	Length	Left	Right	Bottom	Top	Diagonal
Length	0.15024	0.05801	0.05729	0.05713	0.01445	0.00548
Left	0.05801	0.13258	0.08590	0.05665	0.04907	-0.04306
Right	0.05729	0.08590	0.12626	0.05818	0.03065	-0.02378
Bottom	0.05713	0.05665	0.05818	0.41321	-0.26347	-0.00019
Top	0.01445	0.04907	0.03065	-0.26347	0.42119	-0.07531
Diagonal	0.00548	-0.04306	-0.02378	-0.00019	-0.07531	0.19981

Class=genuine _TYPE_=MEAN

NAME	Length	Left	Right	Bottom	Top	Diagonal
	214.969	129.943	129.72	8.305	10.168	141.517

Class=genuine _TYPE_=N

NAME	Length	Left	Right	Bottom	Top	Diagonal
	100	100	100	100	100	100

3.3 Mean, STD and variance-covariance matrix for counterfeit notes

Class=counterf _TYPE_=COV

NAME	Length	Left	Right	Bottom	Top	Diagonal
Length	0.12401	0.031515	0.024001	-0.10060	0.01944	0.01157
Left	0.03152	0.065051	0.046768	-0.02404	-0.01192	-0.00505
Right	0.02400	0.046768	0.088940	-0.01858	0.00013	0.03419
Bottom	-0.10060	-0.024040	-0.018576	1.28131	-0.49019	0.23848
Top	0.01944	-0.011919	0.000132	-0.49019	0.40446	-0.02207
Diagonal	0.01157	-0.005051	0.034192	0.23848	-0.02207	0.31121

Class=counterf _TYPE_=MEAN

NAME	Length	Left	Right	Bottom	Top	Diagonal
	214.823	130.3	130.193	10.53	11.133	139.45

Class=counterf _TYPE_=STD

NAME	Length	Left	Right	Bottom	Top	Diagonal
	0.35215	0.25505	0.29823	1.13195	0.63597	0.55786

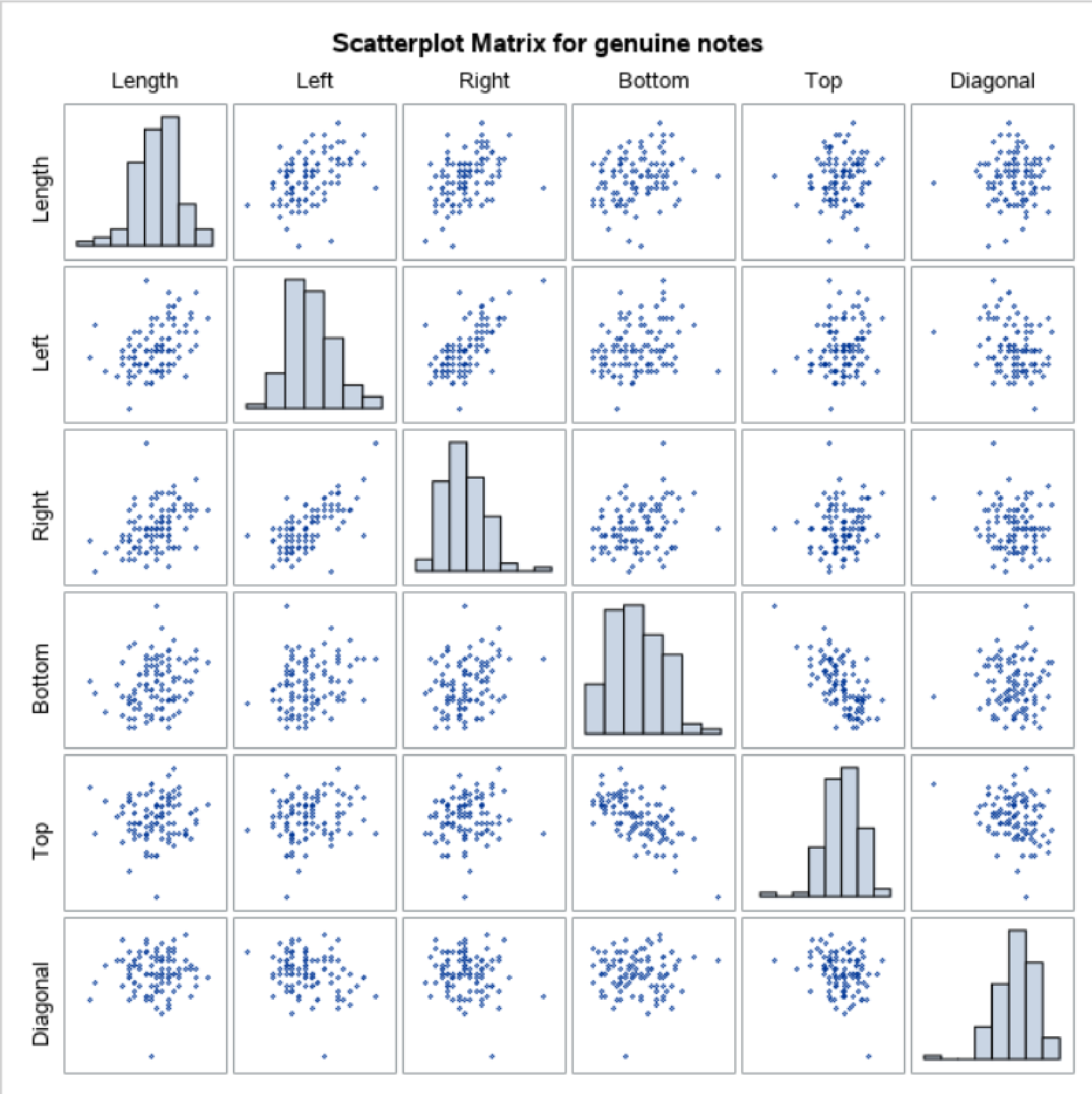
Class=counterf _TYPE_=N

NAME	Length	Left	Right	Bottom	Top	Diagonal
	100	100	100	100	100	100

3.4 Correlation matrix and the scatterplot matrix for genuine notes

Class=genuine _TYPE_=CORR

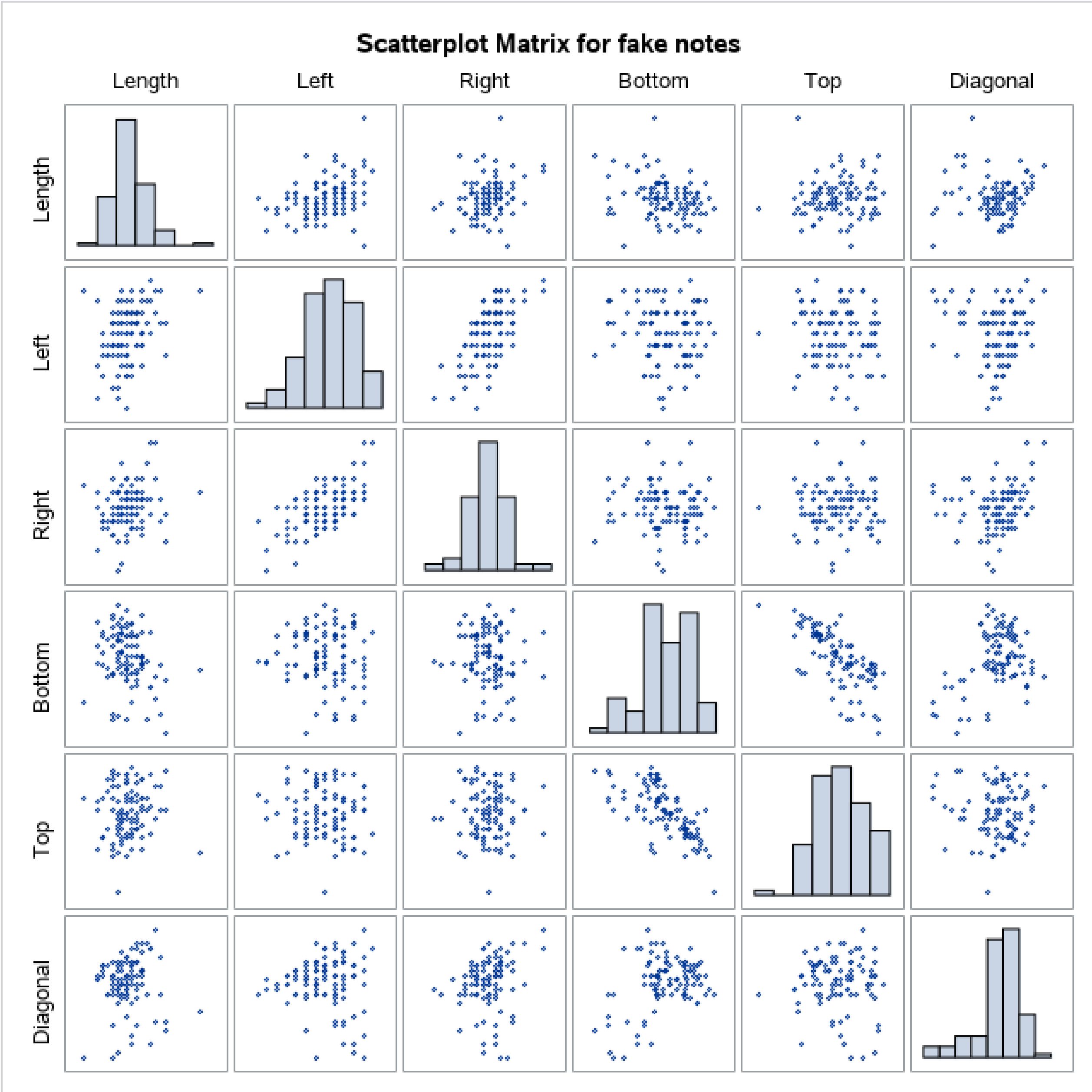
NAME	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00000	0.41105	0.41598	0.22928	0.05745	0.03164
Left	0.41105	1.00000	0.66392	0.24204	0.20764	-0.26458
Right	0.41598	0.66392	1.00000	0.25472	0.13289	-0.14970
Bottom	0.22928	0.24204	0.25472	1.00000	-0.63156	-0.00065
Top	0.05745	0.20764	0.13289	-0.63156	1.00000	-0.25960
Diagonal	0.03164	-0.26458	-0.14970	-0.00065	-0.25960	1.00000



3.5 Correlation matrix and the scatterplot matrix for counterfeit notes

Class=counterf _TYPE_=CORR

NAME	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00000	0.35088	0.22853	-0.25236	0.08678	0.05887
Left	0.35088	1.00000	0.61485	-0.08327	-0.07348	-0.03550
Right	0.22853	0.61485	1.00000	-0.05503	0.00070	0.20552
Bottom	-0.25236	-0.08327	-0.05503	1.00000	-0.68093	0.37766
Top	0.08678	-0.07348	0.00070	-0.68093	1.00000	-0.06221
Diagonal	0.05887	-0.03550	0.20552	0.37766	-0.06221	1.00000


Scatterplot Matrix for fake notes
Length
Left
Right
Bottom
Top
Diagonal

3.6 Discriminant Analysis

The DISCRIM Procedure

Total Sample Size	200	DF Total	199
Variables	6	DF Within Classes	198
Classes	2	DF Between Classes	1

Number of Observations Read	200
Number of Observations Used	200

Class Level Information					
Class	Variable Name	Frequency	Weight	Proportion	Prior Probability
counterf	counterf	100	100.0000	0.500000	0.010000
genuine	genuine	100	100.0000	0.500000	0.990000

Within Covariance Matrix Information		
Class	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
counterf	6	-10.79076
genuine	6	-11.21447
Pooled	6	-10.36654

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
121.899123	21	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

The DISCRIM Procedure

Generalized Squared Distance to Class		
From Class	counterf	genuine
counterf	-1.58042	43.66535
genuine	71.30651	-11.19437

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANKDATA
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Class			
From Class	counterf	genuine	Total
counterf	99 99.00	1 1.00	100 100.00
genuine	0 0.00	100 100.00	100 100.00
Total	99 49.50	101 50.50	200 100.00
Priors	0.01	0.99	

Error Count Estimates for Class			
	counterf	genuine	Total
Rate	0.0100	0.0000	0.0001
Priors	0.0100	0.9900	

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANKDATA
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Class			
From Class	counterf	genuine	Total
counterf	98 98.00	2 2.00	100 100.00
genuine	1 1.00	99 99.00	100 100.00
Total	99 49.50	101 50.50	200 100.00
Priors	0.01	0.99	

Error Count Estimates for Class			
	counterf	genuine	Total
Rate	0.0200	0.0100	0.0101
Priors	0.0100	0.9900	

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	0
Number of Observations Used	0

Number of Observations and Percent Classified into Class			
	counterf	genuine	Total
Total	0 0.00	0 0.00	0 0.00
Priors	0.01	0.99	

Obs	Length	Left	Right	Bottom	Top	Diagonal	counterf	genuine	_INTO_
1	214.9	130.1	129.9	9	10.6	140.5	.000002526	1.00000	genuine

3.6.a)

The test of homogeneity within the covariance matrices show significantly high Chi Sq value (121.9) compared to the full model. This suggests that $\Sigma_1 = \Sigma_2$.

3.6.b)

As the implemented SAS script, the classification of $X_0{}^T$ falls into "Genuine" category.

3.6.c) Confusion Matrix

		Predicted Membership		Number of
	Population	Genuine	Counterf	Observation
Actual Membership	Genuine	99	1	100
	Counterf	2	98	100

Appendix

SAS Codes for Question1

```
/* Read THC.csv data file*/

PROC IMPORT out= work.data
datafile='/home/u41080493/Udeshika/THC.csv'
      DBMS=CSV replace;
      GETNAMES=YES;
      DATAROW=2;
RUN;

%let variableList = chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9
chem10 chem11 chem12 chem13;

/* 1.1 Mean and standard deviation for the 13 chemical concentrations*/

proc means data=data maxdec=4 MEAN STD;
var &variableList;
run;

/* 1.2 Correlation matrix for the 13 chemical concentrations */
proc corr data=data noprob;
var &variableList;
run;


/* 1.2 Scatterplot for the 13 chemical concentrations */
proc sgscatter data=data;
  title "Scatterplot Matrix for THC Data";
  matrix chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11
chem12 chem13/DIAGONAL = (HISTOGRAM);


/* 1.3 Principal component analysis on the raw data*/
proc princomp data=data cov;
var &variableList;
run;


/* 1.4 Principal component analysis on the correlation matrix*/
proc princomp data=data ;
var &variableList;
run;
```

SAS Codes for Question2

```
/* 2.1 Prepare the dataset */
data SocioEconomics;
      input Population School Employment Services HouseValue;
      datalines;
    5700      12.8      2500      270       25000
    1000      10.9      600       10        10000
    3400      8.8       1000      10        9000
    3800      13.6      1700      140       25000
    4000      12.8      1600      140       25000
    8200      8.3       2600      60        12000
    1200      11.4      400       10        16000
    9100      11.5      3300      60        14000
    9900      12.5      3400      180       18000
    9600      13.7      3600      390       25000
    9600      9.6       3300      80        12000
    9400      11.4      4000      100       13000
run;

/* 2.1 Print the dataset */
proc print data=SocioEconomics;
run;

/* 2.2 Mean and standard deviation for the data */
proc means data=SocioEconomics maxdec=4 MEAN STD;
      /* 2.3 Factort analysis*/
proc factor data=SocioEconomics simple corr;
run;

/* 2.5 the scoring  coefficients as  eigenvalues*/
proc princomp data=SocioEconomics;
run;

/* 2.6 the component scores as linear combinations of the observed variable*/
proc factor data=SocioEconomics n=5 score;
```

SAS Codes for Question3

```
/* 1.1 Load dataset */
data bankData;
      infile "/home/u41080493/Udeshika/Swiss Bank data.csv" delimiter=','
missover
            firstobs=1;
      input Class $ Length Left Right Bottom Top Diagonal;
run;

/* 3.1 Print the dataset */
proc print data=bankData(obs=7);
run;

/* 3.2 mean and variance-covariance matrix for genuine notes */
proc corr data=bankData outp=CorrOut COV noprint;
      by Class notsorted;
      var Length Left Right Bottom Top Diagonal;
run;

proc print data=CorrOut(where=(_TYPE_ in ("N", "MEAN", "COV"))) noobs;
      where Class="genuine";

      /* just view information for one group */
      by Class _Type_ notsorted;
      var _NAME_ Length Left Right Bottom Top Diagonal;
run;

/* 3.3 mean, STD and variance-covariance matrix for counterfeit notes */
proc print data=CorrOut(where=(_TYPE_ in ("N", "MEAN", "STD", "COV"))) noobs;
      where Class="counterf";

      /* just view information for one group */
      by Class _Type_ notsorted;
      var _NAME_ Length Left Right Bottom Top Diagonal;
run;

/* 3.4 Correlation matrix for genuine notes */
proc corr data=bankData outp=CorrOut2 noprint;
      by Class notsorted;
      var Length Left Right Bottom Top Diagonal;
run;

proc print data=CorrOut2(where=(_TYPE_ in ("CORR"))) noobs;
      where Class="genuine";

      /* just view information for one group */
      by Class _Type_ notsorted;
      var _NAME_ Length Left Right Bottom Top Diagonal;
run;

proc sgscatter data=bankData;
      where Class="genuine";
```

```
      title "Scatterplot Matrix for genuine notes";
      matrix Length Left Right Bottom Top Diagonal/DIAGONAL=(HISTOGRAM);
run;

/* 3.5 Correlation matrix for counterfeit notes */
proc corr data=bankData outp=CorrOut2 noprint;
      by Class notsorted;
      var Length Left Right Bottom Top Diagonal;
run;

proc print data=CorrOut2(where=(_TYPE_ in ("CORR"))) noobs;
      where Class="counterf";

      /* just view information for one group */
      by Class _Type_ notsorted;
      var _NAME_ Length Left Right Bottom Top Diagonal;
run;

proc sgscatter data=bankData;
      where Class="counterf";
      title "Scatterplot Matrix for fake notes";
      matrix Length Left Right Bottom Top Diagonal/DIAGONAL=(HISTOGRAM);
run;

/* 3.6 Discriminant Analysis */
data test;
      input Length Left Right Bottom Top Diagonal;
      cards;
214.9 130.1 129.9 9 10.6 140.5;
run;

proc discrim data=bankData pool=test crossvalidate testdata=test testout=a;
      class Class;
      var Length Left Right Bottom Top Diagonal;
      priors "genuine"=0.99 "counterf"=0.01;
run;
```