Formative Assignment

Numerical and graphical summaries of the data

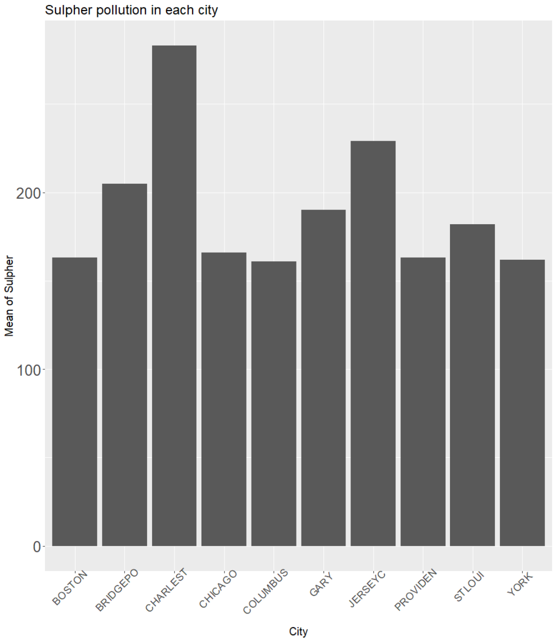
NUMERICAL SUMMARIES-AVERAGE OF P- VARIABLES

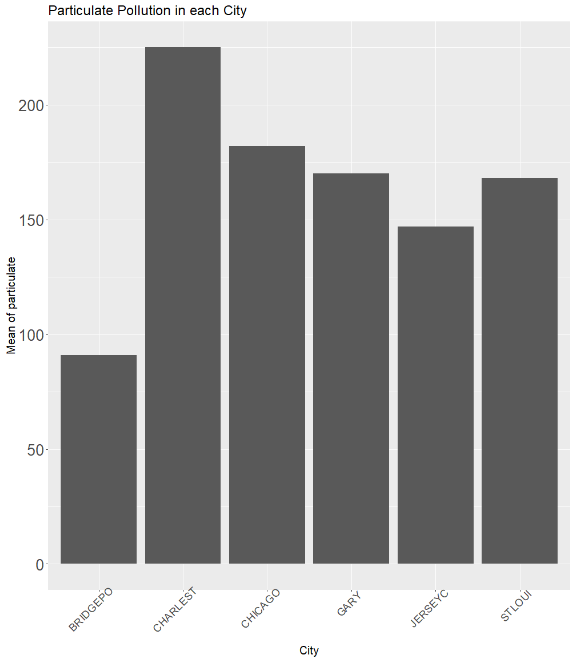
SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
47.1	99.65	219.875	44.5	116.725	275.5375	72.85875	87.2575	81.82875	85.875	56.55078

VARIATION OF P-VARIABLE

SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
913.1544	2542.939	14409.35	337.8481	1508.354	25312.5	23920.24	107.821	45.45271	465.4525	14.85679

GRAPHICAL SUMMARY





Inference- From the graph ‘charlest’ city has the highest Sulphur pollution as well as particulate pollution.

‘Chicago’ stands second in terms of particulate pollution.

‘JERSEYC’ city has the second most Sulphur pollution.

TOTAL VARIATION AND GENERALISED VARIANCE OF THE DATA

Total Variation = 69577.97

Generalized Variance = 8.72131e+29

Standardizing the data matrix and verifying that the sample mean vector of standardized data is composed of zeros and the sample covariance matrix is equal to the sample correlation matrix of the original data.

The result of this command gives : round(apply(scale(airpollution),2,mean),0)

SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
0	0	0	0	0	0	0	0	0	0	0

Standardized covariance matrix

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
SMIN	1	0.5740385	0.3024247	0.18044944	0.1554891	-0.001717436	0.47336207	0.12743151	0.194859	0.20157846	0.11627953
SMEAN	0.574038494	1	0.8319561	0.44808834	0.5535306	0.338628755	0.42093207	0.20826398	0.3315168	0.19154155	0.3767937
SMAX	0.302424735	0.8319561	1	0.34015646	0.5604334	0.473790925	0.19572114	0.2140634	0.2501953	0.0653095	0.25562284
PMIN	0.180449445	0.4480883	0.3401565	1	0.6950989	0.159553313	0.23960598	0.06025395	0.1563026	-0.05247788	0.32265315
PMEAN	0.155489111	0.5535306	0.5604334	0.69509894	1	0.656554363	0.16348532	0.17894753	0.2036115	-0.1136778	0.30352357
PMAX	-0.001717436	0.3386288	0.4737909	0.15955331	0.6565544	1	-0.01003661	0.09917366	0.1337081	-0.14713087	0.11841517
PM2	0.473362073	0.4209321	0.1957211	0.23960598	0.1634853	-0.010036612	1	0.05729714	0.2210385	0.11480895	0.26465782
PERWH	0.127431511	0.208264	0.2140634	0.06025395	0.1789475	0.09917366	0.05729714	1	0.6370699	0.52816946	0.06358122
NONPOOR	0.194859019	0.3315168	0.2501953	0.15630263	0.2036115	0.133708132	0.22103848	0.63706991	1	0.25565891	0.41713832
GE65	0.201578458	0.1915416	0.0653095	-0.05247788	-0.1136778	-0.147130871	0.11480895	0.52816946	0.2556589	1	0.09531892
LPOP	0.116279527	0.3767937	0.2556228	0.32265315	0.3035236	0.118415169	0.26465782	0.06358122	0.4171383	0.09531892	1

Standardized correlation matrix

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
SMIN	1	0.5740385	0.3024247	0.18044944	0.1554891	-0.001717436	0.47336207	0.12743151	0.194859	0.20157846	0.11627953
SMEAN	0.574038494	1	0.8319561	0.44808834	0.5535306	0.338628755	0.42093207	0.20826398	0.3315168	0.19154155	0.3767937
SMAX	0.302424735	0.8319561	1	0.34015646	0.5604334	0.473790925	0.19572114	0.2140634	0.2501953	0.0653095	0.25562284
PMIN	0.180449445	0.4480883	0.3401565	1	0.6950989	0.159553313	0.23960598	0.06025395	0.1563026	-0.05247788	0.32265315
PMEAN	0.155489111	0.5535306	0.5604334	0.69509894	1	0.656554363	0.16348532	0.17894753	0.2036115	-0.1136778	0.30352357
PMAX	-0.001717436	0.3386288	0.4737909	0.15955331	0.6565544	1	-0.01003661	0.09917366	0.1337081	-0.14713087	0.11841517
PM2	0.473362073	0.4209321	0.1957211	0.23960598	0.1634853	-0.010036612	1	0.05729714	0.2210385	0.11480895	0.26465782
PERWH	0.127431511	0.208264	0.2140634	0.06025395	0.1789475	0.09917366	0.05729714	1	0.6370699	0.52816946	0.06358122
NONPOOR	0.194859019	0.3315168	0.2501953	0.15630263	0.2036115	0.133708132	0.22103848	0.63706991	1	0.25565891	0.41713832
GE65	0.201578458	0.1915416	0.0653095	-0.05247788	-0.1136778	-0.147130871	0.11480895	0.52816946	0.2556589	1	0.09531892
LPOP	0.116279527	0.3767937	0.2556228	0.32265315	0.3035236	0.118415169	0.26465782	0.06358122	0.4171383	0.09531892	1

Conclusion: Both the covariance and correlation are same.

Reason for using standardized data than the raw data

It is better to use scale data than non-scaled data as PCA is not scale invariant. Principal components analysis based on the raw data may be dominated by a few of the original variables if they have much higher means and/or variances than the other variables.

Interpretation of the first two principal components

Rotation Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
SMIN	0.261362	0.1902472	0.48898065	0.3038969	-0.005036	0.16976617	-0.2335794	0.65047519	-0.10093751	0.10793791	0.1888865
SMEAN	0.450339	-0.013488	0.17831895	0.2221787	-0.077002	-0.25640015	-0.1668734	-0.14094009	0.10793985	-0.2413032	-0.7256666
SMAX	0.398857	-0.1344166	-0.05261114	0.3329914	-0.167514	-0.32783475	-0.2311736	-0.43584227	-0.07259588	0.21225036	0.529091
PMIN	0.312649	-0.2271652	0.07514298	-0.351073	0.6733466	0.02028976	-0.1170141	-0.0183562	0.21795695	0.45080432	-0.0564058
PMEAN	0.386827	-0.3402921	-0.19234925	-0.051451	0.2634797	0.14420155	0.14871	0.14558294	-0.18129015	-0.6852311	0.2428752
PMAX	0.252282	-0.3447943	-0.37450985	0.2514864	-0.308694	0.20626628	0.44615456	0.27668187	0.19043628	0.37810587	-0.1459211
PM2	0.240471	0.1463248	0.51477592	-0.117162	-0.114303	0.43974026	0.48679957	-0.43913254	-0.05078188	0.0204928	0.0528705
PERWH	0.207324	0.4594607	-0.43826348	0.0865834	0.1838354	0.25437827	-0.0752961	-0.09090255	-0.60313853	0.18709915	-0.1874091
NONPOOR	0.276427	0.3654429	-0.27541948	-0.297116	-0.271933	0.32090663	-0.341315	-0.01058218	0.54480664	-0.1376076	0.1289157
GE65	0.105928	0.5399086	-0.09313504	0.1705604	0.297455	-0.43689985	0.50321729	0.10479728	0.30194137	-0.0712142	0.1403633
LPOP	0.265188	0.0412993	0.04278853	-0.647814	-0.372273	-0.42692962	0.13009585	0.23828579	-0.31924187	0.09954071	0.0095443

First, two principle components are :

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
PC1	0.261362	0.190247	0.488981	0.3038969	-0.00503646	0.1697662	-0.233579	0.65047519	-0.1009375	0.10793791	0.188886532
PC2	0.450339	-0.01349	0.178319	0.2221787	-0.07700225	-0.2564	-0.166873	-0.14094009	0.10793985	-0.24130316	-0.72566658

Interpretation: In these results, the first principal component has large positive associations with SMEAN and PMAX, so this component primarily measures mean concentrations of sulphate and mean suspended particulate. The second component has a large positive association with demographic variables like GE65, Non-Poor etc, so this component primarily measures the demographic properties of the data. Additionally, the first pc1 constitutes 34.88 % of the total variation meanwhile the pc2 constitutes 17.26 % of the total variation.

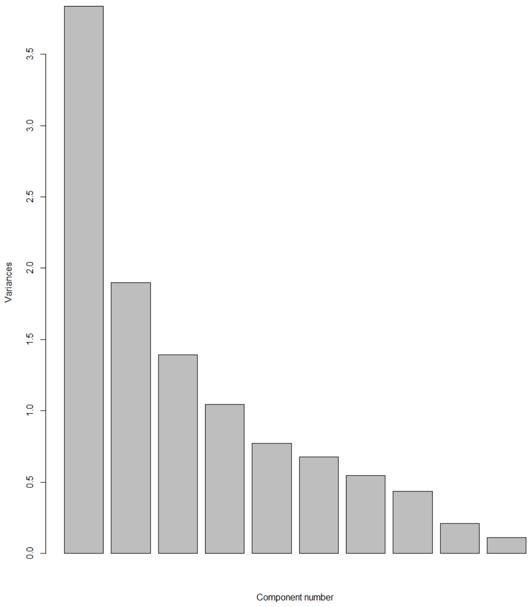
Number of Principle component used in the Analysis

Importance of components:

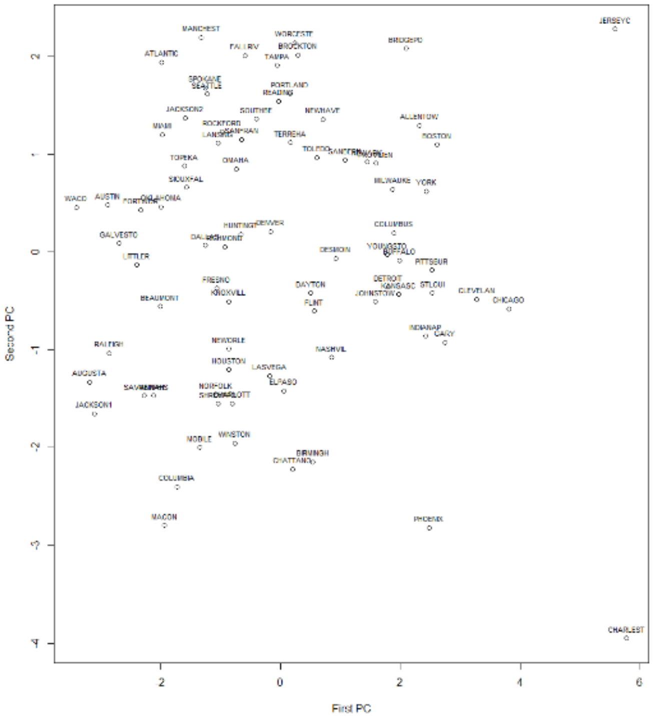
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Proportion	0.3488	0.1726	0.1264	0.09486	0.07025	0.06134	0.04949	0.03968	0.01901	0.00986	0.00762
Cumulative	0.3488	0.5214	0.6479	0.74274	0.81299	0.87433	0.92382	0.9635	0.98251	0.99238	1

From the cumulative variance percentage, we understand that the first four Principal component has to be selected since they account for around 74 % of the total variation.

Graphical Plot also shows the same idea



Scatter plot of the first principal component against the second component



Interpretation: As discussed earlier the pc1 axis measures mean concentrations of sulphate and mean suspended particulate which is the reason for the points corresponding to the cities ‘charlest’ and ‘Jerseyc’ located at the extreme position in the pc1 axis. Similarly, the pc2 axis represents the demographic properties so the cities located at the extreme position in pc2 imply that they have improved demographic qualities. A cluster can be seen located around the point corresponding to the city ‘Denver’.