

Data Mining HW2

Anil Varma B

September 24, 2019

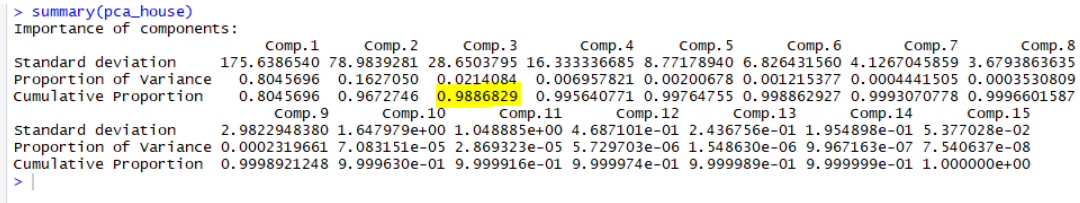
Use BostonHousing dataset to do the following.

1. Conduct a principal components analysis on the data and comment on the results. Select new variables that contain at least 98% of information and write down the new variables which are the linear combination of original variables.

```
#Read csv file
pca <- read.csv(file.choose(),header = TRUE)

#Finding the principal components
pca_house<-princomp(house)

#Summary to check the cumulative variance percentage
summary(pca_house)
```



	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	175.6386540	78.9839281	28.6503795	16.333336685	8.77178940	6.826431560	4.1267045859	3.6793863635
Proportion of Variance	0.8045696	0.1627050	0.0214084	0.006957821	0.00200678	0.001215377	0.0004441505	0.0003530809
Cumulative Proportion	0.8045696	0.9672746	0.9886829	0.995640771	0.99764755	0.998862927	0.9993070778	0.9996601587

	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
Standard deviation	2.9822948380	1.647979e+00	1.048885e+00	4.687101e-01	2.436756e-01	1.954898e-01	5.377028e-02
Proportion of Variance	0.0002319661	7.083151e-05	2.869323e-05	5.729703e-06	1.548630e-06	9.967163e-07	7.540637e-08
Cumulative Proportion	0.9998921248	0.999630e-01	0.999916e-01	0.999974e-01	0.999989e-01	0.999999e-01	1.000000e+00

Figure 0.1: Summary of principal component analysis.

```
> pca_houses$loadings
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
i..CRIM					0.272	0.930	0.157	0.151	0.106						
ZN			-0.632	0.763											
INDUS						-0.126	0.860		-0.465						
CHAS												0.990	-0.137		
NOX														0.999	
RM												-0.992	0.107		
AGE		0.752	0.641												
DIS						-0.110			0.112	0.984					
RAD					0.231	-0.360	-0.399	-0.797	-0.134						
TAX	0.949	-0.293													
PTRATIO								-0.153	0.973	-0.123					
B	-0.291	-0.956													
LSTAT					0.460	0.170	-0.813	0.276							
MEDV				-0.828	0.241	0.223	-0.377	0.176	0.133						
CAT..MEDV											-0.110	-0.135	-0.983		

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067
Cumulative Var	0.067	0.133	0.200	0.267	0.333	0.400	0.467	0.533	0.600	0.667	0.733	0.800	0.867	0.933	1.000

comp.15
SS loadings 1.000
Proportion Var 0.067
Cumulative Var 1.000

Figure 0.2: Loadings on principal component analysis.

New variables with linear combinations

Component1 = (0.949*TAX)+(-0.291*B)

Component2 = (-2.93*TAX)+(-0.956*B)

Component3 = (-0.632*ZN)+(0.752*AGE)

biplot(house_normalise_pca)

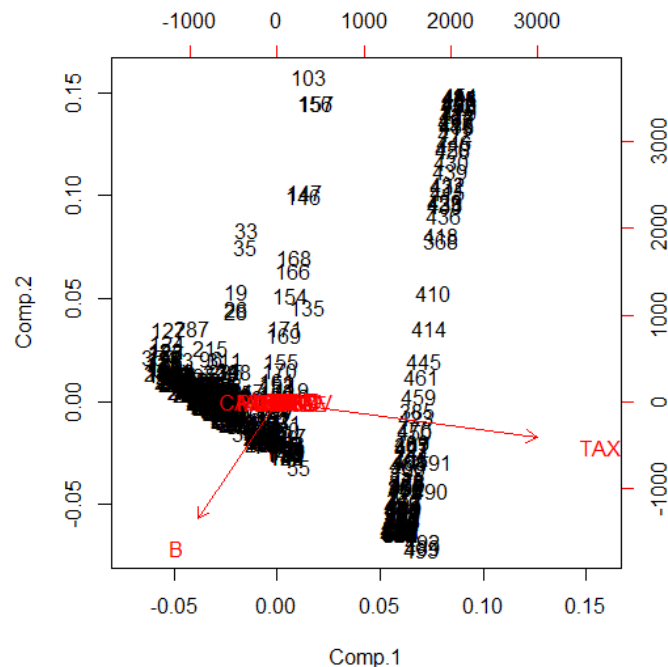


Figure 0.3: Bi plot of data.

2. Should the data be normalized? Discuss what characterizes the components you consider

key. Perform PCA on normalized data. Write down the new components and compare this result with a

```
#Min-Max normalisation function:
#=====
normalize <- function(x)
{
  return((x- min(x)) /(max(x)-min(x)))
}

#Apply Normalize function
house_normalise <- apply(house[1:15], 2, normalise)

#Finding the prinicpal components for normalized data set
house_normalise_pca <- princomp(house_normalise)

> summary(house_normalise_pca)
Importance of components:
      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation 0.6756580 0.3928210 0.3120698 0.2436001 0.21377908 0.18809036 0.16684899 0.14705170 0.13296725
Proportion of Variance 0.4768217 0.1611727 0.1017198 0.06198079 0.04773452 0.03695177 0.02907698 0.02258615 0.01846679
Cumulative Proportion 0.4768217 0.6379944 0.7397142 0.80169503 0.84942955 0.88638132 0.91545830 0.93804445 0.95651124
      Comp.10      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15
Standard deviation 0.10450676 0.088773800 0.083876952 0.080420698 0.074573792 0.061400540
Proportion of Variance 0.01140751 0.008231363 0.007348309 0.006755194 0.005808642 0.003937735
Cumulative Proportion 0.96791876 0.976150120 0.983498429 0.990253623 0.996062265 1.000000000
```

Figure 0.4: Summary of principal component analysis.

Plotting a scree plot to choosing the 'optimal' number of principal components.

Rplot_scree_house_norm_pca

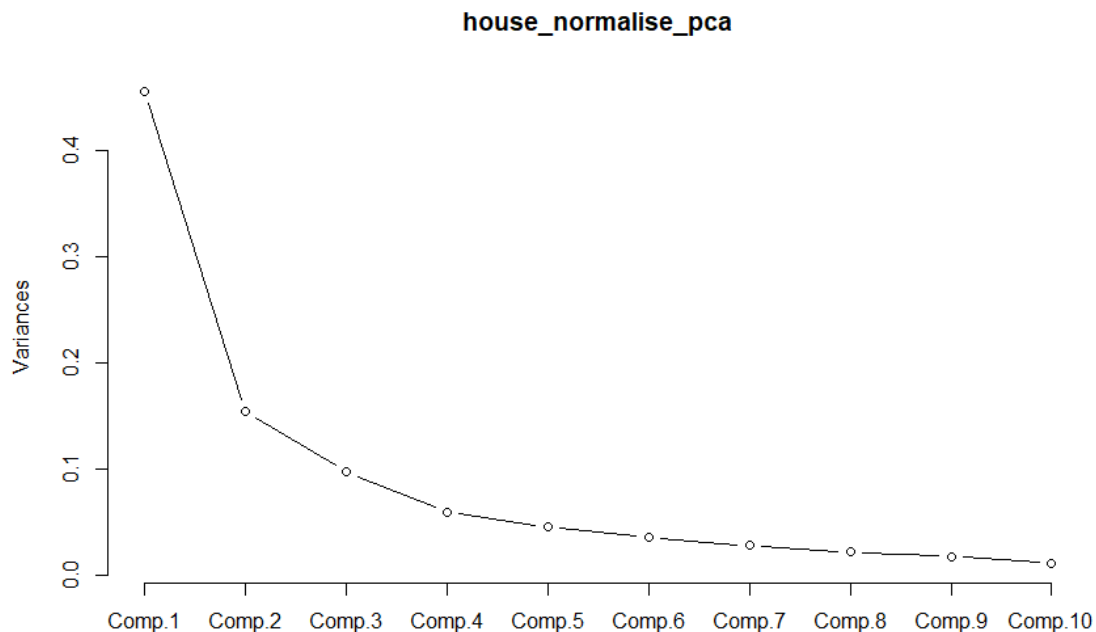


Figure 0.5: Scree plot of data.

```
> house_normalise_pca$loadings
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
1..CRIM									0.116		0.219		0.286	0.684	0.604
ZN	-0.204		0.360	-0.109	0.355	-0.404	0.411	-0.235	-0.364	-0.211	0.331				
INDUS	0.311		-0.175			-0.145	-0.254	-0.694	-0.124	0.291	0.122	-0.394		0.151	
CHAS		0.154	-0.323	-0.912		-0.114	0.130								
NOX	0.283	0.140	-0.240		0.178	-0.208	-0.175			-0.785	-0.202	-0.129			0.208
RM		0.183					-0.114	0.116	-0.361			-0.329	0.786	-0.245	
AGE	0.306	0.134	-0.482	0.220			0.466	0.302	-0.476	0.183			-0.149		
DIS	-0.199	-0.134	0.286	-0.114			0.179			0.120	-0.603	-0.546	-0.222		0.274
RAD	0.468	0.315	0.452	-0.157	-0.141			0.396	0.110		0.263	-0.334	-0.167		-0.225
TAX	0.419	0.205	0.310			-0.211		-0.130	-0.113	0.227	-0.442	0.524	0.151	-0.152	0.177
PTRATIO	0.192	-0.160	0.176		-0.552	0.452	0.369	-0.351	-0.145	-0.317					
B	-0.168		-0.129		-0.663	-0.688			0.120						
LSTAT	0.218				0.184		0.431	0.529	0.187	0.161	-0.159	0.183	-0.490	0.238	
MEDV	-0.207	0.303			-0.157		-0.259	-0.220	0.128	0.307		-0.357	-0.383	0.579	
CAT..MEDV	-0.285	0.778		0.214		0.111	0.228	-0.217	0.287	-0.176			0.122	-0.140	

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067
Cumulative var	0.067	0.133	0.200	0.267	0.333	0.400	0.467	0.533	0.600	0.667	0.733	0.800	0.867	0.933

	Comp.15
SS loadings	1.000
Proportion Var	0.067
Cumulative var	1.000

```
> |
```

Figure 0.6: Loadings on principal component analysis.

New variables with linear combinations

Component1 = $(-0.204 \cdot \text{ZN}) + (-0.311 \cdot \text{INDUS}) + (0.283 \cdot \text{NOX}) + (0.306 \cdot \text{AGE}) + (-0.199 \cdot \text{DIS}) + (0.468 \cdot \text{RAD}) + (0.419 \cdot \text{TAX}) + (0.192 \cdot \text{PTRATIO}) + (-0.168 \cdot \text{B}) + (0.218 \cdot \text{LSTAT}) + (-0.207 \cdot \text{MEDV}) + (-0.285 \cdot \text{CAT..MEDV})$
and so on till the 13th variable for achieving 98 percent.

We plot a biplot to show in case of a normalised data set we have the principal components are orthogonal to each other and also capture maximum data.

```
biplot(house_normalise_pca)
```

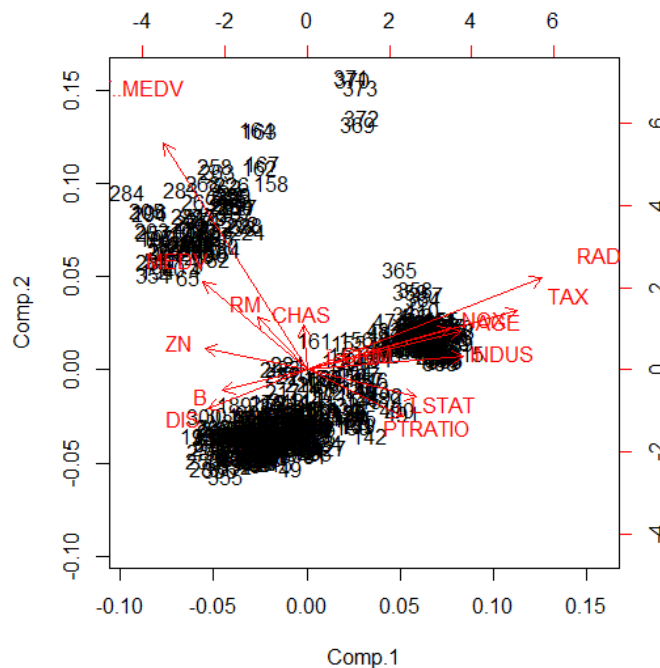


Figure 0.7: Bi plot of data.