Vandana Agrawal

Principal Component analysis: Data set-Food.csv

Principal Component is used to reduce dimensionality of a data set. The major goal of PCA is to identify the underlying structure in data so that we can reduce the data by identifying how the various variable correlated to each other to define another dimension. PCA reduces redundancy of data.

There are 24 cities and their food price of Bread Hamburger, butter Apples and tomato is provided in cent Looking at the data below points were noted:

	Highest Price	City	Lowest Price	City
Bread	70.9	Anchorage	28.9	Baltimore
Hamburger	135.6	Anchorage	84.5	San Diego
Butter	162.3	Kansas City	123.2	Milwaukee
Apple	65.1	Chicago	35.6	Buffalo
Tomato	104.5	Baltimore	75.9	Buffalo

From the above data it appears that in Baltimore Bread is cheap and Tomato is costly. In Anchorage Bread and Hamburger are Expensive. Also, we might say that Hamburgers are costly in Anchorage because Bread is expensive there and Hamburgers are made of bread. Apple and Tomato are cheap in Buffalo. Butter looks like most expensive food item among all food item with highest price in Kansas City 162.3 and its lowest price value is 123.2

Variance of Bread: 69.9, Variance of Hamburger: 135.3, Variance of Butter: 85.1, Variance of Apple: 69.8, Variance of Tomato: 54.7. Above data shows that Hamburger's price has highest variance and Tomato has the least. Price Variance for Bread and apple is quite same.

Below is the correlation matrix of the given data

```
              Bread Hamburger    Butter    Apples    Tomato
Bread     1.0000000 0.6490532 0.3301770 0.3187031 0.3620681
Hamburger 0.6490532 1.0000000 0.2447778 0.1908956 0.5557993
Butter    0.3301770 0.2447778 1.0000000 0.2351424 0.4361291
Apples    0.3187031 0.1908956 0.2351424 1.0000000 0.1333844
Tomato    0.3620681 0.5557993 0.4361291 0.1333844 1.0000000
```

From Above data we can deduce that Bread and Hamburger are positively and strongly co related meaning price of Hamburger will increase if the price of bread increases vice-versa. Next strongly co related food item with Hamburger is Tomato and then Butter. This might be because all these are ingredients to make Hamburger. Butter and tomato price are also strongly and positively co related. Higher the value in correlation matrix meaning data is more redundant.

Below is the percentage variance of Individual food item price

```
  Bread Hamburger    Butter    Apples    Tomato
0.1686394 0.3260335 0.2051058 0.1683354 0.1318860
```

Hamburger contributes 32% in total variance in total food price for all cities. Variance will signify which food item is most significant feature.

Principal Component is used for dimension reduction. We are going identify the principal component whose variance are as large as it can be. We have 5 food items hence the number of Principal components can be 5

```
eigenvalues
```

```
[1] 216.79440 79.12794  62.26846  34.67047  22.22005
```

Eigen values gives the variance of each principal component. Total variance of principal component=415.0813. First PCA(PCA1) has 52% of variance. PCA2 variance = 19% PCA 3 variance = 15%. So if we are going to use PCA1 instead of five original variables (Bread, Hamburger etc.) we are going to Account 52% of original data. Also if we consider PCA1, PCA2 we are going to account 71% of total data.

Total Variance of original variables (Bread, Hamburger etc.) is same as the total variance of new variables (Principal components). Meaning that there would be no data loss.

```
> eigenvectors
       [,1]       [,2]        [,3]       [,4]        [,5]
[1,] -0.4529089  0.05515147  0.21435116  0.6856702  0.52511130
[2,] -0.7146773  0.48679539 -0.02261341 -0.1338116 -0.48358009
[3,] -0.3391656 -0.75632931 -0.43256354  0.1976187 -0.29456459
[4,] -0.2203644 -0.42895099  0.81242257 -0.3231370 -0.05470465
[5,] -0.3471543 -0.06289354 -0.32619100 -0.6070257  0.63296724
```

Observation: In the First Principal Component Eigen value for Bread and Hamburger is quite predominate. So, Each unit increase in price of Hamburger will decrease in PCA1 value by .71 . Also all the variable has negatively coefficient w.r.t to PCA which indicates that increased price any of them will increase the value of PCA1.In PCA2 butter has highest magnitude and in PCA3 Apple is positive magnitude.

Same could be observed if we calculate the correlation between each principal component and the data

```
> cor(y1,data)
          Bread  Hamburger     Butter     Apples     Tomato
[1,] -0.7970559 -0.9045578 -0.5412281 -0.3881604 -0.6908451

Change in price of Bread And Hamburger will affects PCA1
> cor(y2,data)
          Bread Hamburger     Butter     Apples      Tomato
[1,] 0.05863756  0.372232 -0.7291558 -0.4564764 -0.07561449
```

Change in Price butter will affect PCA2

```
> cor(y3,data)
```

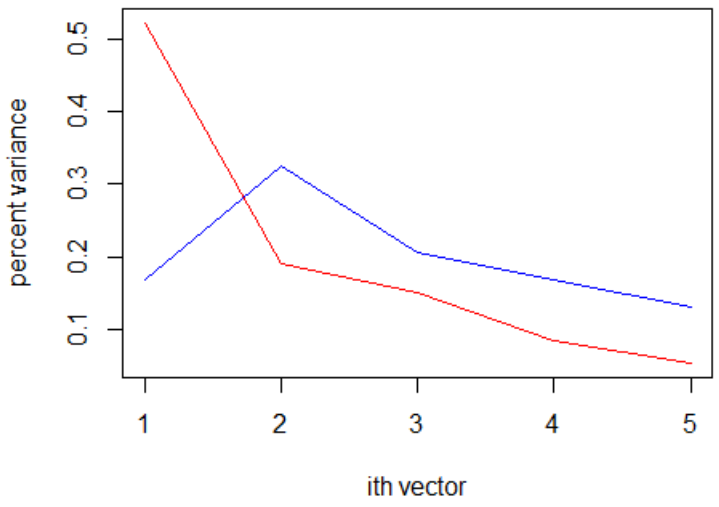
Vandana Agrawal

```
          Bread   Hamburger      Butter    Apples     Tomato
[1,] 0.2021687 -0.01533919 -0.3699376 0.7669406 -0.3478886
```

Principal Component 3 value is majorly depended on Apple price. Looks like this variable same as apple.

```
> eigenvectors[,1]*sqrt(eigenvalues[1])/sqrt(diag(vars))
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.7970559       -Inf       -Inf       -Inf       -Inf
[2,]       -Inf -0.9045578       -Inf       -Inf       -Inf
[3,]       -Inf       -Inf -0.5412281       -Inf       -Inf
[4,]       -Inf       -Inf       -Inf -0.3881604       -Inf
[5,]       -Inf       -Inf       -Inf       -Inf -0.6908451
```

```
eigenvectors[,2]*sqrt(eigenvalues[2])/sqrt(diag(vars))
           [,1]     [,2]       [,3]       [,4]        [,5]
[1,] 0.05863756      Inf        Inf        Inf         Inf
[2,]        Inf 0.372232        Inf        Inf         Inf
[3,]       -Inf     -Inf -0.7291558       -Inf        -Inf
[4,]       -Inf     -Inf       -Inf -0.4564764        -Inf
[5,]       -Inf     -Inf       -Inf       -Inf -0.07561449
```



Above Graph red line represents Variance of Principal component and Blue line represent original variable variance. Variance corresponding Principal component goes down to zero when we move from first principal component to fifth.

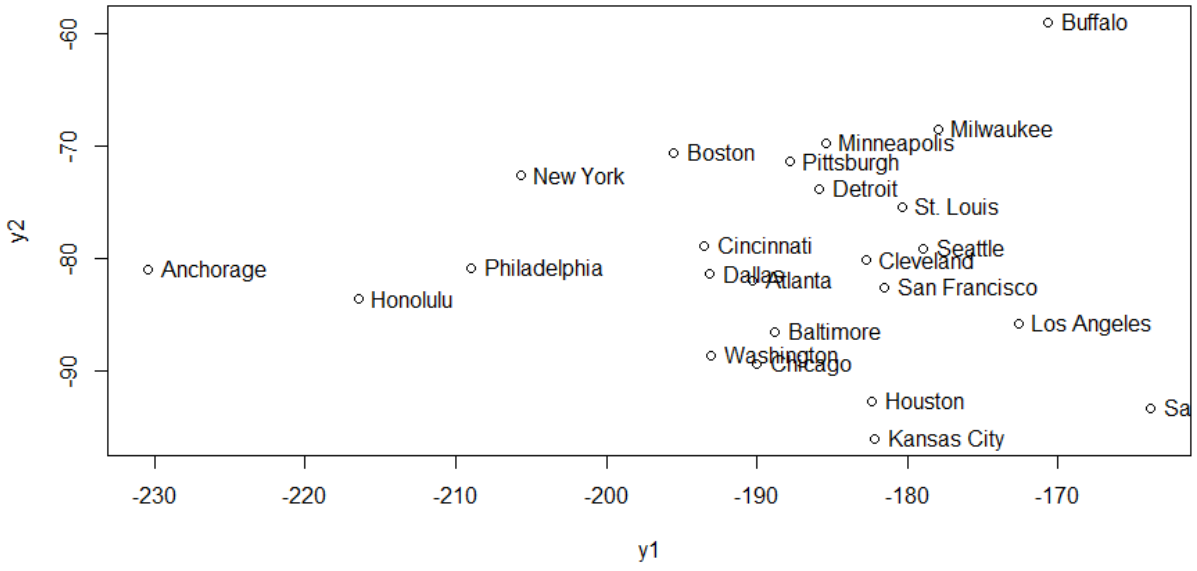
Vandana Agrawal

There is sharp drop in curve from First principal component to second and third. After Third principal component curve becomes steady. From this graph we can identify that what would be the minimum number of principal component require to create out model so that maximum variance in the data is accounted.

From the Correlation matrix we know that the original value of variable changes with change in other variable. We do not want our principal component to be correlated to avoid multicollinearity issue. Checking Correlation between Principal Components.

```
              [,1]          [,2]          [,3]          [,4]          [,5]
[1,]  1.000000e+00 -1.253822e-16  2.609254e-18  7.317803e-16 -1.606498e-16
[2,] -1.253822e-16  1.000000e+00  3.839127e-16 -2.777729e-18 -9.059390e-16
[3,]  2.609254e-18  3.839127e-16  1.000000e+00 -2.014107e-16  4.405486e-16
[4,]  7.317803e-16 -2.777729e-18 -2.014107e-16  1.000000e+00  1.369178e-16
[5,] -1.606498e-16 -9.059390e-16  4.405486e-16  1.369178e-16  1.000000e+00
```

Above values are almost tends to zero. This signifies that the principal components are not correlation with each other so change in one component doesn't affect the other. Hence redundancy is reduced here.



Plotting First and second principal Component. From the above plot we see that cities having low food price for bread and hamburger had high principal component values where cities Bread and hamburger price is more has low value of principal component 1 for example Kansas city has low value of bread and hamburger it falls high value of principal component.

Vandana Agrawal

Using regression model to test the model.

```
lm(formula = dv ~ as.matrix(data)) (Regression model with original variable)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7674  -5.3918   0.0916   6.1469  13.7811

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               15.4416    38.7806   0.398 0.695182
as.matrix(data)Bread       1.3439     0.3556   3.780 0.001372 **
as.matrix(data)Hamburger   0.5235     0.2736   1.913 0.071752 .
as.matrix(data)Butter      0.9464     0.2685   3.524 0.002424 **
as.matrix(data)Apples      0.6396     0.2728   2.345 0.030698 *
as.matrix(data)Tomato      1.5586     0.3760   4.145 0.000609 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 10.25 on 18 degrees of freedom
Multiple R-squared:  0.9188,   Adjusted R-squared:  0.8963
F-statistic: 40.74 on 5 and 18 DF,  p-value: 3.373e-09



lm(formula = dv ~ as.matrix(y)) (Regression model with new variables or using principal c
omponent.)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7674  -5.3918   0.0916   6.1469  13.7811

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.4416    38.7806   0.398  0.69518
as.matrix(y)1  -1.9858     0.1451 -13.684 5.93e-11 ***
as.matrix(y)2  -0.7592     0.2402  -3.161  0.00541 **
as.matrix(y)3  -0.1219     0.2708  -0.450  0.65792
as.matrix(y)4  -0.1143     0.3629  -0.315  0.75639
as.matrix(y)5   1.1253     0.4533   2.483  0.02313 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.25 on 18 degrees of freedom
Multiple R-squared:  0.9188,   Adjusted R-squared:  0.8963
F-statistic: 40.74 on 5 and 18 DF, p-value: 3.373e-09
```

When we are trying to predict the total food price for a city using regression. R squared value is same in both the cases. Also, we already know that total variance of original variable is equal to the total variance of principal component so. R squared value will be same.

```
> cor(dv,data)
         Bread Hamburger    Butter    Apples    Tomato
[1,] 0.7176994 0.7942445 0.5958123 0.4523603 0.7461481
```

From the above correlation it appears that the total food price of each city is more correlated to hamburger and Tomato these two seems to be the most important.

```
> summary(lm(dv~data$Hamburger+data$Tomato))

Call:
lm(formula = dv ~ data$Hamburger + data$Tomato)

Residuals:
    Min      1Q  Median      3Q     Max
-31.108  -8.015  -3.034  12.538  29.441

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     91.7538    43.6283   2.103 0.047691 *
data$Hamburger   1.5381     0.3563   4.317 0.000304 ***
data$Tomato      1.9416     0.5601   3.466 0.002309 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.52 on 21 degrees of freedom
Multiple R-squared:  0.7652,   Adjusted R-squared:  0.7428
F-statistic: 34.21 on 2 and 21 DF,  p-value: 2.471e-07
```

```
> summary(lm(dv~y1+y2))
```

```
Call:
lm(formula = dv ~ y1 + y2)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8517  -6.3435  -0.5877   5.9785  16.0048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2606    32.1418  -0.008   0.9936
y1           -2.0946     0.1389 -15.083 9.67e-13 ***
y2           -0.5345     0.2299  -2.325   0.0302 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.806 on 21 degrees of freedom
Multiple R-squared:  0.9173,   Adjusted R-squared:  0.9094
F-statistic: 116.4 on 2 and 21 DF, p-value: 4.309e-12
```

In above case Model with Principal component gives the high value of R squared and high F statistics value and this expla ins that 90% of the food price of given data is explained by two principal components.

Standardization of data: In above case with non-standardized data we saw that the food item having greater variance inf luence the value of principal components. (Hamburger has highest variance).It appears that Hamburger is only Compon ent which explains the variance in data.

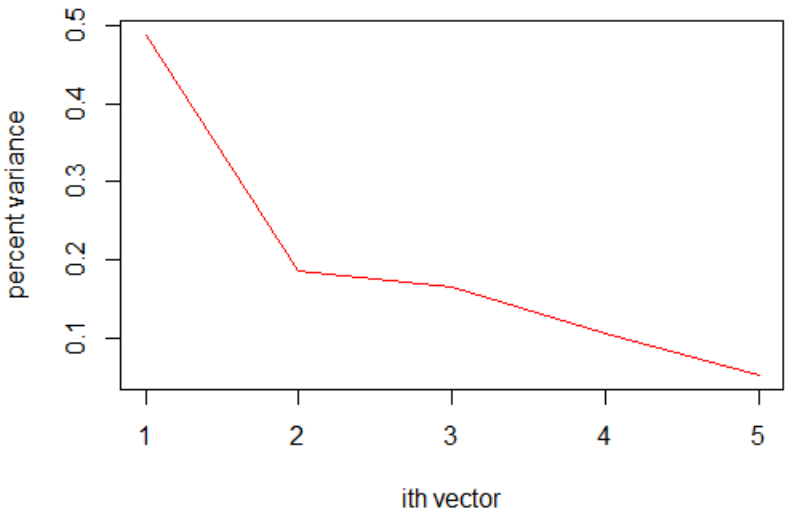
After standardization: variance of each variable is one meaning that unlike in the previous case we are giving importance to each data variable

```
> vars
    Bread Hamburger    Butter    Apples    Tomato
        1         1         1         1         1


> eigenvalues
[1] 2.4393555 0.9295911 0.8332378 0.5328728 0.2649429


> percentvars_pc
[1] 0.48787109 0.18591823 0.16664756 0.10657455 0.05298857
```

Vandana Agrawal



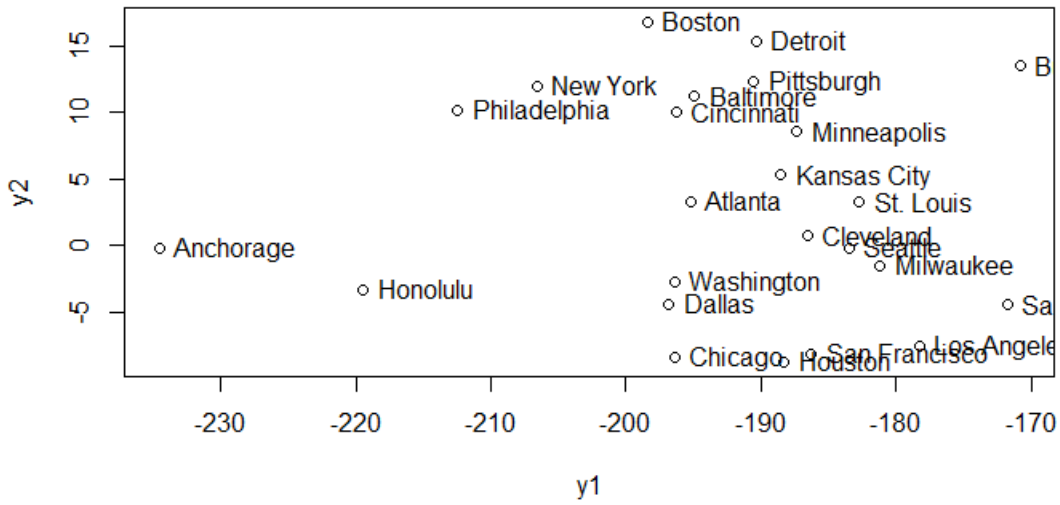
Three is sharp drop in value of variance after first principal component but it would not be sufficient to consider just one component. Variance drop is quite small from PCA2 to PCA3 , it would be good to consider PCA1 and pca2

```
> cor(y1,data)
          Bread  Hamburger     Butter     Apples     Tomato
[1,] -0.7983803 -0.8461256 -0.6039118 -0.4376169 -0.7217623
```

```
> cor(y2,data)
             Bread Hamburger      Butter     Apples    Tomato
[1,] -0.001041619 0.3462199 -0.09660557 -0.8165072 0.3683601
```

After standardization of data Principal component 1 value is affected by majorly three variables. Bread Hamburger and Tomato. Previously majorly Hamburger prices are affecting PCA1. Same applies for PCA2 Apple and tomatoes are influencing its value.

Vandana Agrawal



After standardization of the data we see quite changes in the position of cities in the graph. Kansas City moved up and appears in the middle location of the plot. And Chicago Dallas moved downwards in the graph. principal component analysis will tend to give more emphasis to those variables that have higher variances than to those variables that have lower variances. In effect the results of the analysis will depend on the units of measurement used to measure each variable. Standardization will give better result.

Verifying the using regression model.

```
lm(formula = dv ~ y1 + y2)

Residuals:
    Min      1Q  Median      3Q     Max
-18.096  -6.372  -1.147   7.974  16.237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.14846   26.65836   0.643    0.527
y1          -2.18408    0.13792 -15.836 3.76e-13 ***
y2          -0.06725    0.24088  -0.279    0.783
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.465 on 21 degrees of freedom
Multiple R-squared:  0.9229,   Adjusted R-squared:  0.9156
F-statistic: 125.8 on 2 and 21 DF,  p-value: 2.047e-12
```

R Squared value is increased after the standardization. Meaning after standardizing the data our model improved.