Using **Turtles** data set to complete Assignment 6.

Jolicoeur and Mosimann measured the length, width, and height of 48 painted turtle shells and conducted PCA on the data. Their work has been influential in the field of allometry, the study of the relative growth of a part of an organism in relation to an entire organism.

**(a)** Calculate the correlation matrix for the length, width and height variables. Do you see the highly correlation among the three variables?

**(b)** Conduct principal component analysis on the standardized length, width and height variables. Do this part in R, no need to show the result in the written part.

**(c)** Calculate the eigenvectors and eigenvalues of the correlation matrix.

**(d)** Calculate the principal components PC1, PC2, and PC3. No need to list all values for PC1, PC2, PC3.  In written part, show the calculation expression learned from the class, such as PC1 = 0.707\*Z1 + 0.707\*Z2, etc.

**(e)** Create a scatterplot of PC1 versus PC2, set the x-axis and y-axis with the same length. Which component explains more variability, PC1 or PC2? What is the correlation between PC1 and PC2?

**(f)** What percentage of the variability is explained by the first principal component? What percentage of the variability is explained by the first two principal components combined?  
   
**(g)** Create loadings plot of the first and second principal components. Write a brief interpretation of the plot.

**(h)** After the principal component analysis, can a majority of the variability in the data be explained with just one or two principal components? Use scree plot to confirm your thought.

**(a)** Calculate the correlation matrix for the length, width and height variables. Do you see the highly correlation among the three variables?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **length** | **width** | **height** |
| **length** | 1 | 0.978312 | 0.964695 |
| **width** | 0.9783116 | 1 | 0.960571 |
| **height** | 0.9646946 | 0.960571 | 1 |

**Length, width and height are highly corelated since the value of r is close to 1.**

**(b)** Conduct principal component analysis on the standardized length, width, and height variables. Do this part in R, no need to show the result in the written part.

R code

**(c)** Calculate the eigenvectors and eigenvalues of the correlation matrix.

|  |  |  |
| --- | --- | --- |
| **λ1** | **λ2** | **λ3** |
| 2.93573765 | 0.04284387 | 0.02141848 |

|  |  |  |
| --- | --- | --- |
| **V1** | **V2** | **V3** |
| -0.5787981 | -0.3250273 | 0.74789704 |
| -0.577984 | -0.4834699 | -0.65741263 |
| -0.5752628 | 0.8127817 | -0.09197088 |

**(d)** Calculate the principal components PC1, PC2, and PC3. No need to list all values for PC1, PC2, PC3.  In written part, show the calculation expression learned from the class, such as PC1 = 0.707\*Z1 + 0.707\*Z2, etc.

PC1 = -(0.5787981) \*Z1 + (-0.5779840) \*Z2 + -(0.5752628) \*Z3

PC2 = -(0.3250273) \*Z1 + -(0.48346990) \* Z2 + (0.8127817) \*Z3

PC3 = (0.74789704) \*Z1 + -(0.65741263) \*Z2 + -(0.09197088)\*Z3

**(e)** Create a scatterplot of PC1 versus PC2, set the x-axis and y-axis with the same length. Which component explains more variability, PC1 or PC2? What is the correlation between PC1 and PC2?

Most variability is explained by PC1.

There is 0 correlation between PC1 and PC2.

Chart, scatter chart

Description automatically generated

**(f)** What percentage of the variability is explained by the first principal component? What percentage of the variability is explained by the first two principal components combined?

PC1 explains 97 .85 % of variability.

PC1 and PC2 together explains 99.28 % of variability.

**(g)** Create loadings plot of the first and second principal components. Write a brief interpretation of the plot.

Chart, line chart

Description automatically generated

**Loading plot of PC1**

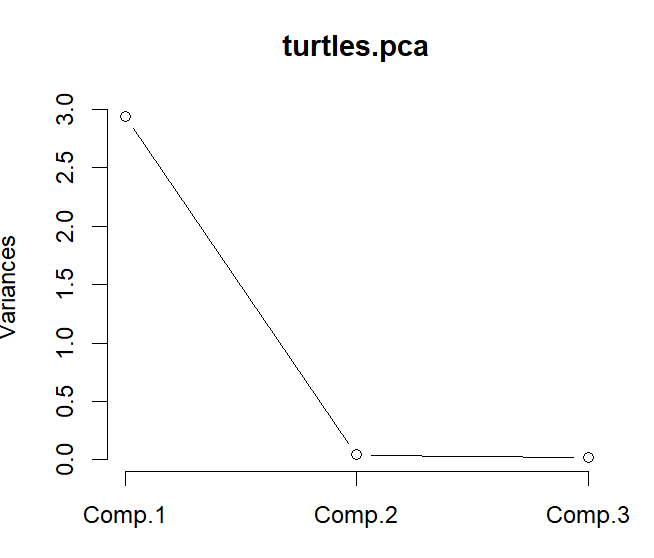
Loading plot of PC1 shows that all weights on PC1 are positive.

**Loading plot of PC2**

Loading plot of PC2 shows that both positive and negative weights are there on PC2.

PC2 has positive weight for length and width and negative weight for height

**(h)** After the principal component analysis, can a majority of the variability in the data be explained with just one or two principal components? Use scree plot to confirm your thought.



PC1 explains 97.85 % of variability, ie majority of the variability in the data can be explained with just one principal component.

Scree plot shows the change after comp 2 which means that we only need PC1 as our principal component, which confirms our thought.