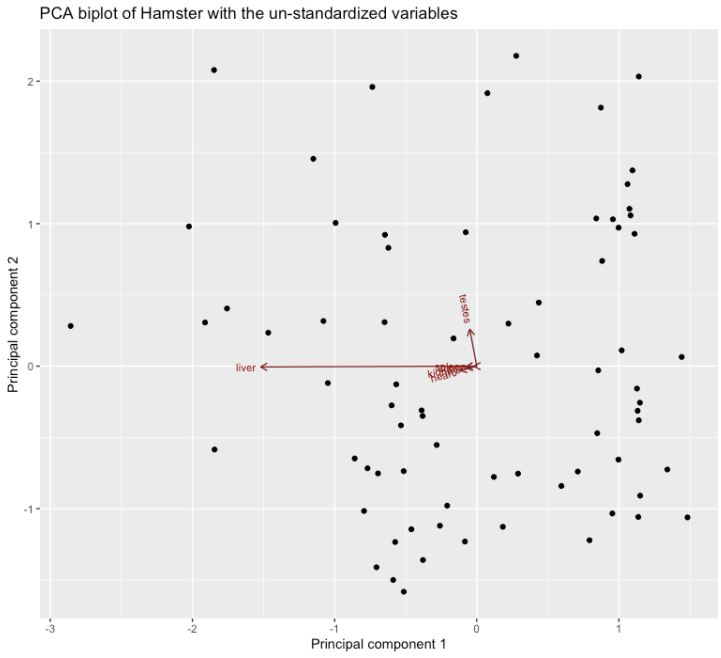
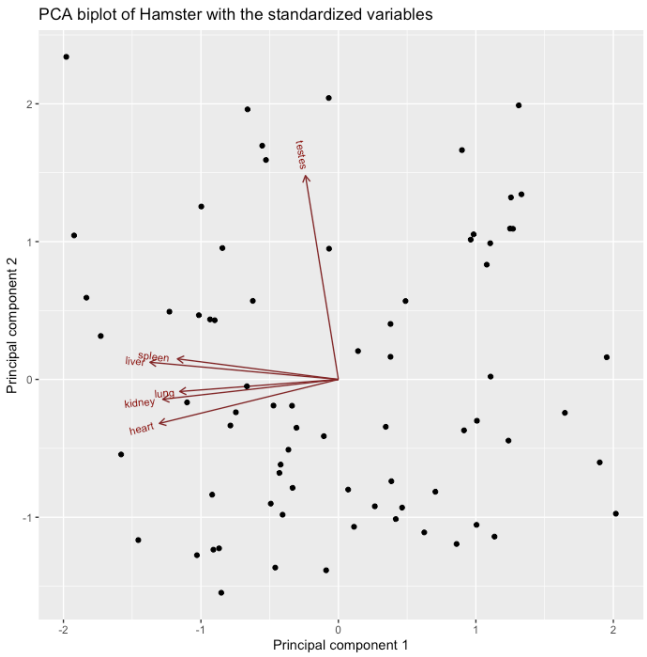
Ritwik Budhiraja rbudhira@iu.edu 2000916462

EDA ASSIGNMENT – 5

1. Do a PCA of the organ weights and create a PCA biplot. Do you think it's useful to standardize the variables (scale. = TRUE in the prcomp function)? Does the PCA with the standardized variables tell you something different than the PCA with variables on the un-standardized variables?





I would prefer standardizing the variables because:

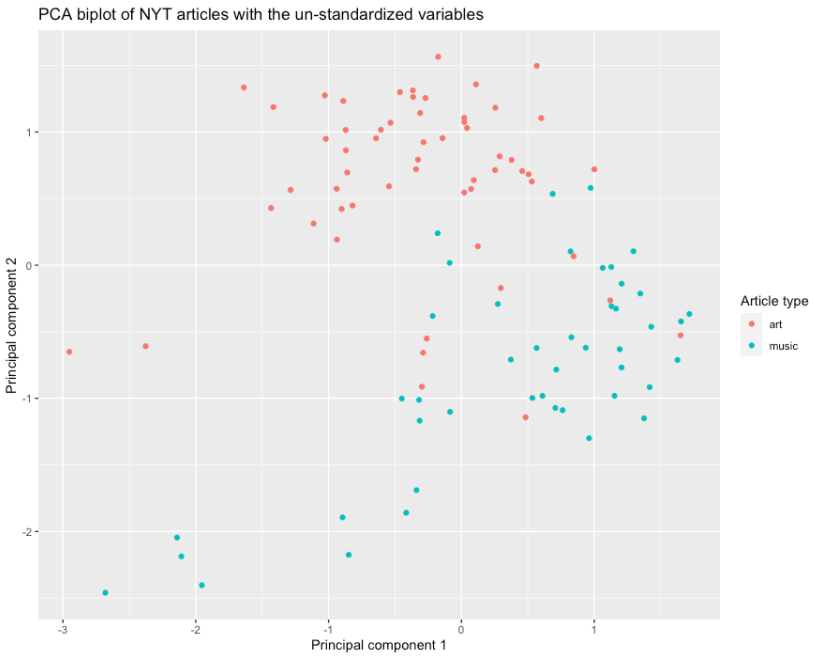
- The loadings (variables) appear cluttered in the unstandardized plot, and now in the standardized plot, it is easier to comment on the correlations between any two loadings (variables) by looking at the angles between them.
- In the unstandardized PCA, liver had the most influence on Principal component 1, but in the standardized plot, all the variables, except testes, have a somewhat similar influence on Principal component 1.

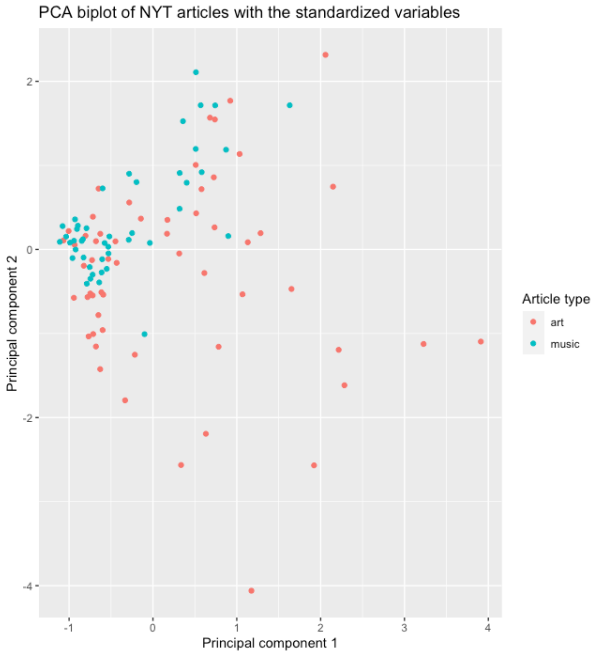
Some observations:

- The influence of testes on Principal component 2 increases in the standardized plot.
- All the pairs of loadings which have a really small angle, such as liver and spleen, are highly correlated. While, all the pairs of loadings which have an angle close to 90 degrees, such as kidneys and testes, have a negligible correlation.

2. Does the PCA with the standardized variables tell you something different than the PCA with the un-standardized variables? Which one is more useful for this data set? What does it tell you about the article?

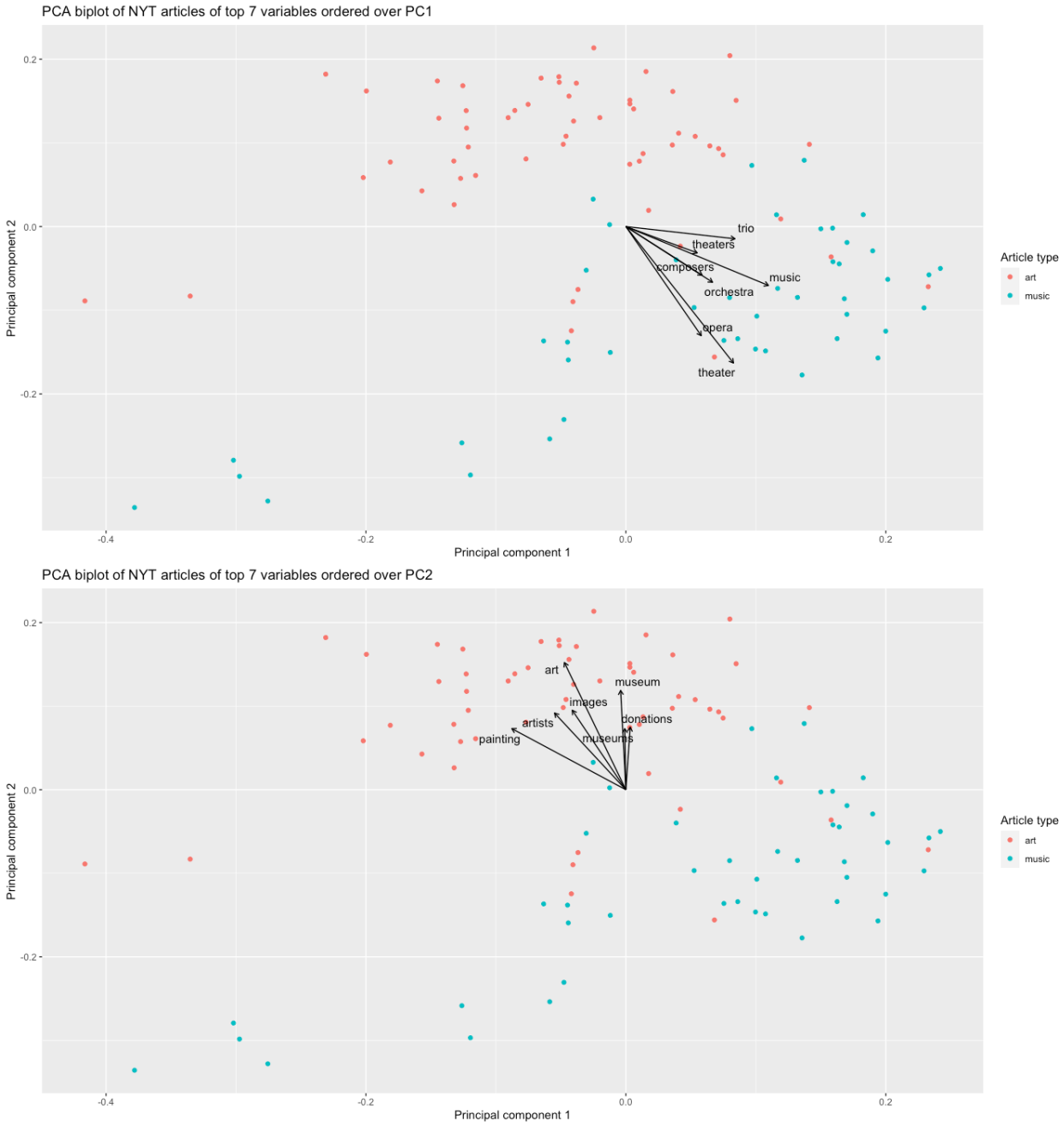
Ritwik Budhiraja rbudhira@iu.edu 2000916462





- In the case of the unstandardized plot, both the article types (other than some outliers) have a similar score on principal component one. Art article type has a larger score on principal component two as compared to the score of music article type.
- In the case of the standardized plot, both the articles have a similar score on both principal component 1 and principal component 2 (other than some outliers).
- I would prefer using the unstandardized plot more because it is easier to make distinguished comments about the data, and if loadings are also plotted, it would be easier to figure out the relationships and the contributions of each variable.

3. Plot the biplot axes corresponding to the variables with the largest loadings on the principal axes, and describe what the axes suggest about the differences betwen articles about art and articles about music.



- If we take the top variables ordered over Principal Component 1, then most of the contributions come from the words of article type music.

- If we take the top variables ordered over Principal Component 2, then most of the contributions come from the words of article type art.