# Lab 7 solutions

# Exercise 1

## Examine how the the ten ratings variables are correlated with the livability scores (the Score variable) given the cities. Write a short summary of your findings.

The correlations between livability score and community, health, safety, transportation are moderate and positive.

The correlation between livability score and house is moderate and negative.

The correlation between livability score and is weak and positive.

# Exercise 2

## All the scores are measured on a scale of 0‐100, so why is it still necessary to use the correlation matrix, or standardize the data, before doing PCA? (Hint: Compute some summary statistics or make some plots.)

Since the variables differ in their variances, we need to standardize the variables. Variables with larger variances will be given more weight in determination of the first few principal components unless the variables are standardized.

# Exercise 3

## Present a summary of the PCA including the table of eigenvectors, a list of eigenvalues (variance), and cumulative percent of total variance explained by the principal components. (Be sure to make your output readable, e.g. rounding digits appropriately.)

## 

## How many PCs would you need to use to explain 80% of the total variation?

6 PCA would be needed to explain at least 80% of the total variation

1. Explain how the Cumulative Proportion row of the summary of the PC was calculated.

The cumulative proportion row was calculated by adding the proportion of variance explained by the current PC to that of all the previous .

# Exercise 4

## Explain how the variables contribute to the first two principal components.

Cities with relatively low ratings for climate and urban, and relatively high ratings for diversions, economics, education, community, health, safety, and transportation have high scores for the first principal component. Cities with relatively high ratings for climate and urban, and relatively low ratings for diversions, economics, education, community, health, safety, and transportation have low scores for the first principal component.

Cities with relatively high ratings for climate, education, community, urban and safety, and relatively low ratings economic activity and health and housing have high scores for the second principal component. Transpiration and diversions don’t matter much. This is a comparison of climate, education, community and safety factors with economic and health factors.

## Using three pieces of information, where the elbow is in the scree plot, the proportion of total variation, and the interpretation of the PCs, make an argument for how many PCs would you recommend to summarize this data?

From the scree plot, the corner appears to be around 4. This indicates that 3 principal components should be used. Three components capture about 63% of the total variation. You might consider keeping the fourth principal component if it had a nice interpretation and could be interpreted. The component appears to contrast economic opportunities, education and health with sense of community. Cities that are healthy places to live with good economic and educational opportunities, but have low ratings on sense of community have high scores on this factor. This factor may contrast cities that are basically suburbs or bedroom communities for larger cities with more isolated small cities that have a greater sense of community but fewer educational and economic activities. It may be of sufficient interest to keep. After component four, the components seem to be capturing about equal amounts of random variation and account for less than 30% of the total variation.

**Exercise 5**

## Compare the scores for the first principal component with the Score variable in the data (this is the rating the article gives for each city). Which city would be rated first using the Score variable? Which city would be rated first using the scores for the first principal component? (You could make a plot of the Score variable against PC1, and compute the correlation between the two variables.) Do the two approaches give cities similar ratings? (You may need to multiple the scores for the first principal component by -1).

Mount Vernon, WA has the highest score variable. Bozeman, MT has the highest value for the first principal component. The correlation is moderate and positive (0.547) as shown in the following plot.

The two approaches might give cities different

scores because the magazine gives equal weight

to all

variables while the PC1 rating take considerations about the correlation among variables. We do not know how the magazine rated the cities. They could have used weighted sums of principal component scores.