**Chapter 8 Exercises**

**Concepts**

1. If a researcher finds that the data points in a factor analysis are surrounded by a

3-dimensional ellipsoidal space that is elongated along the x and y axes but concentrated within the z axes, what does this imply about the contribution of the variable mapped to the x, y, and z dimensions?

1. How do factor analysis and cluster analysis differ in the way they aggregate data?
2. What are the advantages and drawbacks associated with hierarchical versus nonhierarchical clustering methods?
3. What are the assumptions of factor analysis and how might spatial dependence within the data affect the results of a factor analysis?

**Exercises**

1. Explain and interpret the following rotated factor loading table:

|  |  |  |
| --- | --- | --- |
|  | **Factor 1** | **Factor 2** |
| % University degree | .13 | .88 |
| % Malay | .48 | .52 |
| % English speaking | .05 | .75 |
| % > 65 years old | .75 | .29 |
| % Homeowners | -.67 | -.24 |

1. Perform a hierarchical cluster analysis using the following data ranking prospective employees . Comment on your findings.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Test score** | **Time to complete test** | **Score from interview** |
| 1 | 41 | 21 | 52 |
| 2 | 32 | 37 | 44 |
| 3 | 31 | 22 | 35 |
| 4 | 33 | 29 | 37 |
| 5 | 23 | 43 | 12 |
| 6 | 23 | 28 | 22 |
| 7 | 19 | 18 | 25 |
| 8 | 36 | 25 | 48 |
| 9 | 31 | 30 | 50 |
| 10 | 31 | 30 | 49 |

1. How many significant factors are there in the following analysis? How many variables were there in the original analysis?

|  |  |  |
| --- | --- | --- |
| **Factor** | **Eigenvalue** | **Cumulative percent of variance explained** |
| 1 | 3.1 | 29 |
| 2 | 1.7 | 43 |
| 3 | 0.8 | 62 |
| 4 | 0.4 | 71 |
| 5 | 0.4 | 76 |
| 6 | 0.3 | 82 |
| 7 | 0.2 | 87 |
| 8 | 0.2 | 91 |
| 9 | 0.2 | 93 |
| 10 | 0.2 | 95 |
| 11 | 0.1 | 96 |
| 12 | 0.1 | 97 |
| 13 | 0.1 | 98 |
| 14 | 0.1 | 99 |
| 15 | 0.1 | 100 |

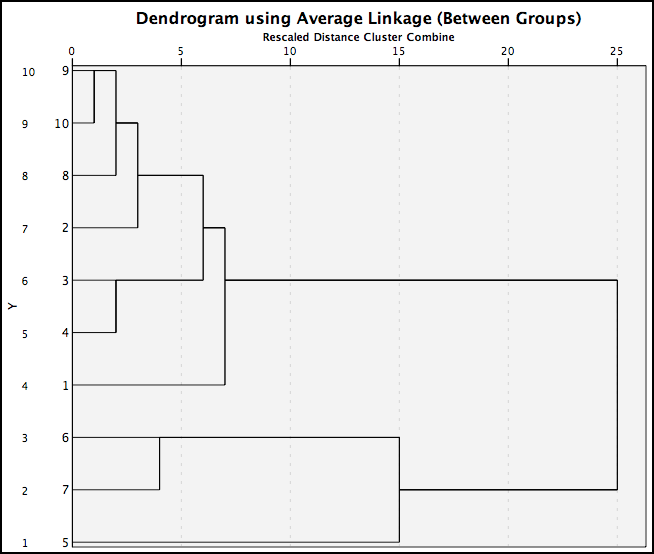
1. A marketing group collects the following data about the neighborhoods it serves. Use factor analysis to summarize it. How many factors are sufficient to describe the data? (that is, have eigenvalues greater than 1). What characterizes the factors?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **Average income (1000s)** | **>65** | **% English speakers** | **% Female** |
| 1 | 22 | 12 | 40 | 47 |
| 2 | 54 | 11 | 78 | 54 |
| 3 | 54 | 6 | 83 | 42 |
| 4 | 53 | 8 | 67 | 48 |
| 5 | 34 | 10 | 55 | 47 |
| 6 | 42 | 3 | 73 | 52 |
| 7 | 32 | 6 | 65 | 37 |
| 8 | 38 | 5 | 56 | 51 |
| 9 | 35 | 2 | 64 | 54 |
| 10 | 37 | 9 | 59 | 46 |
| 11 | 52 | 3 | 77 | 48 |
| 12 | 33 | 4 | 57 | 48 |

1. Using the data from Question 4, generate the factor scores and identify which neighborhoods the marketing firm might target when advertising an expensive English-language cable package?
2. Use the Milwaukee Sales dataset to carry out a k-means cluster analysis on the *z*-score standardized lot size, age, air conditioning, and number of bedrooms associated with a property.
3. Using the Singapore 2010 Census dataset, carry out factor analysis on the unemployment rate, illiteracy rate, percentage of the population with no school, percentage of the population who rent, the number of English speakers, and the percentage of the population with a university degree. Describe the rotated factor loadings, and identify the most important variables that comprise each factor.
4. Using the SPSS Housing Dataset, carry out a factor analysis on the property’s region, price, number of bedrooms, date of construction, and floor area. Describe the rotated factor loadings, and and identify the most important variables that comprise each factor.
5. Use the SPSS Housing Dataset, carry out a k-means cluster analysis on the z-score standardized house price, number of bedrooms, date of construction, and floor area.   
   Use k = 3.

**Chapter 8 Solutions**

1. The first factor is highly loaded with being older than 65, and home ownership. The second factor covers university degrees and English speakers, with less emphasis on Malay ethnicity.
2. Many of the applicants are quite similar, with the whole group tending to fall into two or arguably three groups. If three groups are selected, there are a few obvious outliers.



1. Arguably 3 to 4 significant factors. There were 8 variables in the original analysis.
2. Two factors are sufficient to describe the data. Factor 1 is characterized by the average income and the percentage of English speakers, while Factor 2 is characterized by the percentage of females.
3. Neighborhoods 3 and 11 score highly for Factor 1, which is correlated with income and English-speaking.
4. The first cluster is characterised by the presence of air conditioning, while the second focuses on high age score and slightly less high numbers of bedrooms. The third cluster is defined by large lot sizes.
5. The first component has all but the percent of renters loading highly, while the second component loads the percentage of renters. The illiteracy rate and percentage of the population with no school dominate the first component, so that it might be termed the education factor, while the second factor is exclusively the renter factor.
6. The rotated component matrix shows Factor 1 with strong loading of price, number of bedrooms, date of construction, while Factor 2 is loaded with the region and floor area. These two factors have eigenvalues above one and cumulatively explain 69.3% of the variance.
7. The three clusters are each quite large, with the smallest cluster containing 103 records. Each of the variables is significant, and the final clusters are characterized by the following: the first cluster tends to contain relatively older houses, which also tend to be cheaper. The second cluster contains relatively expensive, large houses with many bedrooms. The third cluster consists of newer houses, which are slightly smaller than the average.