On the left of ◘ Fig 9.26, we can now see a description of the three clusters, including their (relative) sizes. We find that the first cluster contains 40.2 % of the objects, the second cluster 33.5 % of the objects, and the third cluster contains 26.2 % of the objects. Further below, the output shows the distribution of the *gender* variable in each cluster. Moving the mouse over the boxes showing the clustering variable labels, SPSS shows their mean values as well as their relative importance in terms of predicting each object's membership per cluster. Darker shades (i.e., higher values in feature importance) denote the variable's greater importance for the clustering solution. Comparing the results, we learn that *gender* is by far the most important variable for each of the clusters, followed by *e5* ("Oddjob Airways provides you with a very pleasant travel experience"), *e21* ("Oddjob Airways makes traveling uncomplicated"), *e9* ("Oddjob Airways gives you a sense of safety"), *e22* ("Oddjob Airways provides you with interesting on-board entertainment, service, and information sources"), and *e1* ("with Oddjob Airways you will arrive on time").[6] Clicking on one of the boxes will show a graph with the frequency distribution of each cluster.

The auxiliary view on the right-hand side shows an overview of the variables' overall importance for predicting the clustering solution (i.e., across all clusters). The model viewer provides us with additional options for visualizing the results or comparing clustering solutions. It is worthwhile to simply play around with the different self-explanatory options. So go ahead and explore the model viewer's features yourself!



## 9.5 Oh, James! (Case Study)

**Case Study**

The James Bond movie series is one of the success stories of filmmaking. The movies are the longest continually running and the third-highest-grossing film series to date, which started in 1962 with Dr. No, starring Sean Connery as James Bond. As of 2018, there have been 24 movies with six actors having played James Bond. Interested in the factors that contributed to this running success, you decide to investigate the different James Bond movies' characteristics. Specifically, you want to find out whether the movies can be grouped into clusters, which differ in their box-office revenues. To do so, you draw on Internet Movie Database (www.imdb.com) and collect data on all 24 movies based on the following variables (variable names in parentheses):

- Title. (*title*)
- Actor playing James Bond. (*actor*)
- Year of publication. (*year*)
- Budget in USD, adjusted for inflation. (*budget*)
- Box-office revenues in the USA, adjusted for inflation. (*gross_usa*)
- Box-office revenues worldwide, adjusted for inflation. (*gross_worldwide*)
- Runtime in minutes. (*runtime*)
- Native country of the villain actor. (*villain_country*)
- Native country of the bondgirl. (*bondgirl_country*)

6 The strong emphasis of gender in determining the solution supports prior research, which found that two-step clustering puts greater emphasis on categorical variables in the results computation (Bacher et al. 2004).

- Haircolor of the bondgirl. (*bondgirl_hair*)

Use the dataset *James Bond.sav* (↓ Web Appendix → Downloads) to run a cluster analysis—despite potential objections regarding the sample size. Answer the following questions:

1. Which clustering variables would you choose in light of the study objective, their levels of measurement, and correlations?
2. Given the levels of measurement, which clustering method would you prefer? Carry out a cluster analysis using this procedure.
3. Interpret and profile the obtained clusters by examining cluster centroids. Compare the differences across clusters on the box-office revenue variables.
4. Use a different clustering method to test the stability of your results.

## 9.6 Review Questions

1. In your own words, explain the objective and basic concept of cluster analysis.
2. What are the differences between hierarchical and partitioning methods? When do we use hierarchical or partitioning methods?
3. Repeat the manual calculations of the hierarchical clustering procedure from the beginning of the chapter, but use complete linkage as the clustering method. Compare the results with those of the single linkage method.
4. Explain the different options to decide on the number of clusters to extract from the data. Should you rather on statistical measures or rather on practical reasoning?
5. Run the two-step clustering analysis on the Oddjob Airways data again (*Oddjob.sav*, ↓ Web Appendix → Downloads) but with a prespecified number of four and five clusters. Compare your results with the original three-cluster solution.
6. Which clustering variables could be used to segment:
   - The market for smartphones?
   - The market for chocolate?
   - The market for car insurances?

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). New York: Springer.

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods in marketing research* (pp. 160–189). Cambridge: Basil Blackwell & Mott, Ltd.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027–1035.

Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep Cluster – A first evaluation. Arbeits- und Diskussionspapiere/Universität Erlangen-Nürnberg, Sozialwissenschaftliches Institut, Lehrstuhl für Soziologie, 2004-2. http://www.ssoar.info/ssoar/handle/document/32715.

Becker, J.-M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. *Marketing Letters*, *26*(4), 643–659.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods*, *3*(1), 1–27.