1 Einführung

1.1 Explorative Datenanalyse

“Exploratory data analysis is detective work” (Tukey, 1977, S.1). So beschreibt John W. Tukey, Begründer der Explorativen Datenanalyse, sein Verständnis dieser Analysetechniken. Die Arbeit eines guten Ermittlers zeichnet sich dadurch aus, dass er weiß, wonach es sich an einem Tatort zu suchen lohnt und welche Hilfsmittel er dazu benötigt. (In der Regel ist es viel sinnvoller auf einer Türklinke nach Fingerabdrücken zu suchen als auf einer Glühlampe.) Auch explorative Datenanalyse lässt sich als Detektivarbeit auffassen, allerdings suchen Sie dabei nicht an einem Tatort, sondern in Daten. Wie bei einem guten Ermittler steht am Ende das Ziel **Muster oder Auffälligkeiten zu finden**, die neue Schlussfolgerungen ermöglichen bzw. etwas Unbekanntes erklären.

Die Explorative Datenanalyse (EDA) bezeichnet im engeren Sinne eine Reihe meist grafischer oder semigrafischer Verfahren, die bestimmte Informationen zu einzelnen Variablen oder zum Zusammenhang von zwei oder mehr Variablen sichtbar machen. Das können beispielsweise Kennwerte, Verteilungen, Unterschiede, Zusammenhänge oder auffällige Werte sein. Das Ziel der EDA ist es, Ihnen als „Datenanalytiker“ einen schnellen Überblick über einen bestimmten Sachverhalt (z. B. die Verteilung von Daten) zu verschaffen. Daraus können Sie dann Schlussfolgerungen ziehen oder Hinweise darauf bekommen, welche Aspekte der Daten Sie genauer untersuchen sollten.

Vor diesem Hintergrund geht es bei den grafischen Verfahren der EDA also in erster Linie um die Informationen, die *Sie* aus der Grafik ziehen. Im Gegensatz dazu haben Grafiken in wissenschaftlichen Publikationen das Ziel, dem Leser, meist eine Person, die mit den Daten nicht unmittelbar vertraut ist, eine bestimmte Information mitzuteilen. Manchmal gehen beide Ziele Hand in Hand und einige explorative Verfahren eignen sich hervorragend, um eine Aussage zu unterstreichen oder einen bestimmten Aspekt innerhalb Ihrer Daten zu kommunizieren. Sie sollten sich aber auch bewusst sein, dass das Verständnis und die Anwendung von EDA-Verfahren entsprechende Erfahrung oder zumindest eine Erklärung voraussetzt und nicht jedem unmittelbar die Interpretation der entsprechenden Ergebnisse zugänglich ist.

1.2 Inhalte des Buches

Im vorliegenden Buch beschreiben wir die wichtigsten Grundtechniken der explorativen Datenanalyse im Bereich der Sozial- und Humanwissenschaften. Wir haben versucht viele Beispiele und Abbildungen zu nutzen, um Ihnen den Einstieg in diese Thematik zu erleichtern. Daher bieten wir Ihnen auch keinen umfassenden Verfahrensüberblick (was auch gar nicht möglich ist, da sich die EDA an *Ihren* Daten orientieren muss.) Doch wenn Sie einmal mit der EDA vertraut sind, wird es Ihnen nicht schwerfallen Ihr Wissen zu ergänzen.

Für die EDA nutzen wir die frei erhältliche Programmiersprache *R*. Nach unserer Auffassung ist R das am besten geeignete Werkzeug für die EDA. Mit diesem Buch richten wir uns auch an Personen, die das erste Mal mit R arbeiten. Infolgedessen behandeln wir im ersten Teil (Kapitel 2 und 3) Grundlegendes zu R bzw. RStudio.

Wir haben uns dazu entschieden, RStudio als grafische Benutzeroberfläche für R zu verwenden. Nach unserem Dafürhalten erleichtert dies den Einstieg, da damit viele Aufgaben wie Öffnen und Speichern von Dateien auf konventionelle Art mit Schaltflächen durchgeführt werden können. Sie können sich dadurch besser auf die statistischen und datenbezogenen Aufgaben in R konzentrieren. Wenn Sie bereits Erfahrung im Umgang mit R haben, können Sie diesen Teil des Buches überspringen. Die Beispiele und Skripte sind so gestaltet, dass Sie auch direkt in R ausgeführt werden können. Sie sind also keineswegs an RStudio gebunden.

In diesem Buch verwenden wir „Explorative Datenanalyse“ in einem sehr weitgefassten Sinn, der auch die sogenannte beschreibende oder deskriptive Datenanalyse mit umfasst. Manche Autoren machen eine Trennung zwischen EDA und deskriptiver Datenanalyse. Jedoch ist die Trennlinie nicht genau definiert. Oft versteht man unter deskriptiver Analyse die Beschreibung von Kennwerten für einzelne Variablen (z. B. Mittelwerte und Streuungen) oder für den Zusammenhang von Variablen (z. B. Korrelationskoeffizienten). Aber auch das Erstellen von Häufigkeitsverteilungen und Streuungsdiagrammen wird manchmal als deskriptive Analyse bezeichnet. Wie dem auch sei: Die Kenntnis der wichtigsten statistischen Kennwerte ist auch notwendig, um die Ergebnisse von EDA-Analysen im engeren Sinn zu verstehen. Deswegen stellen wir sie im zweiten Teil dieses Buches vor (Kapitel 4 und 5). Wir gehen dabei auch exemplarisch auf Effektgrößen und Zusammenhangsmaße ein. Damit wollen wir den Ausgangspunkt für Ihre eigenen Analysen legen. Wenn Sie bereits mit diesen Themen vertraut sind, so können Sie auch diesen Bereich überspringen.

Der dritte Teil des Buches behandelt verschiedene Grafiken zur EDA. Wir gehen auf die Exploration von Verteilungen, Streudiagrammen sowie kategoriale und multivariate Daten ein (Kapitel 6, 7, 8, und 9). Abschließend geben wir einen Ausblick auf den Einsatz von EDA-Techniken zur Illustration inferenzstatistischer Resultate (Kapitel 10).

1.3 Skripte, Beispiele und Datensätze

In jedem Kapitel finden Sie mehrere Quellcodebeispiele, die zur Umsetzung der Inhalte dieses Buches dienen. Außerdem bieten wir Ihnen eine Webseite unter: http://www.r-stutorials.de/eda an, auf der Sie die Skripte und Datensätze herunterladen und zur Übung nutzen können. Aber gerade wenn Sie das erste Mal mit R arbeiten, ist es hilfreich, die eine oder andere Zeile Quellcode selbst zu schreiben.