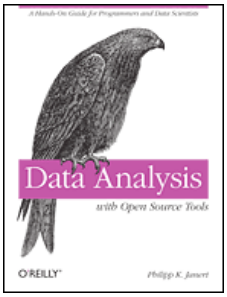

Data Analysis with Open Source Tools
By Phillip K. Janert
First Edition November 2010
ISBN 978-0-596-80235-6
530 Seiten
EUR38.00, SFR79,90

Buchbesprechung

Das Buch besteht aus mehreren Teilen, die sich mit den folgenden Themen befassen:

- Datenvisualisierung
- Modellierung
- Data Mining
- Anwendungen zur Datenverarbeitung

In jedem Kapitel wird zuerst ein Thema erörtert. Am Ende folgt dann ein Workshop, in dem ein OpenSource-Tool vorgestellt wird. Jedes Kapitel endet mit einer Liste weiterführender Literatur.

Im Vorwort stellt sich der Autor des Buches als Physiker vor, der auch in der Softwareentwicklung tätig war. Er erläutert, dass neben den Daten auch das Modell des Systems, welches der Ursprung der erhobenen Daten darstellt, von Bedeutung ist. Au8erdem stellt er im Vorwort einige Grundsätze auf, die man bei der Datenanalyse befolgen sollte. Ebenfalls stellt er klar, dass nicht alle Empfehlungen aus dem Buch, die er aus seiner eigenen Arbeit abgeleitet hat, allgemeingültig sein müssen.

Im Einleitungskapitel erklärt der Autor, dass sich das Buch an technisch orientierte Personen richtet, die in der Lage seien, Ihre Konzepte in einer Programmiersprache auszudrücken. Außerdem wurde noch auf die für das Verständnis des Buches notwendigen mathematischen Vorkenntnisse eingegangen. Danach wird in dem Kapitel der Aufbau des restlichen Buches skizziert. Am Ende wird noch dargestellt, dass sich der Autor auf die Analyse von Geschäftsdaten bezieht und welche Themen in dem Buch nicht abgehandelt werden.

Im ersten Teil des Buches wird die Visualisierung von Daten behandelt. Hierbei wird zuerst hervorgehoben, dass allein durch die Visualisierung der Daten wichtige Erkenntnisse gewonnen werden können. So kann z. B. festgestellt werden, ob sich sogenannte Cluster bilden, und ob es viele oder wenige Ausreißer gibt. Danach werden im ersten Kapitel verschiedene Diagrammtype für Datenmengen vorgestellt, die nur eine Variable besitzen. Hierbei wird auf richtige Parameterisierung bei der Erstellung der Diagramme eingegangen.
Außerdem erläutert der Autor, warum die klassischen Begriffe des Mittelwertes und der Standardabweichung aus der beschreibenden Statistik nicht für alle Datenmengen geeigent sind. Stattdessen wird ein Konzept entwickelt, welches auch für Datenmengen geeignet ist,

für die diese klassische Begriffe ungeeignet sind. In einem Workshop wird dann das Python-Paket NumPy vorgestellt.

Im zweiten Kapitel werden Datenbestände mit zwei Variablen behandelt. Hierbei wird u. a. Augenmerk auf die Frage: „Gibt es eine Beziehung zwischen den Variablen“ gelegt. Daraufhin werden Methoden vorgestellt, mit denen man das Rauschen der Daten unterdrücken kann, um eine glatte Kurve für die Visualisierung der Daten erstellen zu können.

Im dritten Kapitel wird das Thema Zeitreihenanalyse behandelt. Hierbei wird darauf eingegangen, dass sich die Daten einer Zeitreihe einen Bezug auf einem zeitlichen Verlauf besitzen und aus den Komponenten, Trent, saisonale Einflüsse und einem Rauschen bestehen. Nach er Vorstellung von Verfahren zur Zeitreihenanalyse wird das Pythonmodul scipy.signal im Workshop vorgestellt.

In dem darauffolgenden Kapitel werden dann verschiedene Visualisierungstechniken für Datenbestände mit mehr als zwei Variablen behandelt. Diese Techniken reichen von Falschfarbendarstellung bis zu computergestützten interaktiven Methoden, die es dem Benutzer erlauben, den Datenbestand aus unterschiedlichen „Blickwinkeln“ zu betrachten.

Der erste Teil des Buches wird durch ein Beispiel abgeschlossen, bei dem eine Datenanalyse mit Hilfe des Tools gnuplot demonstriert wird.

Der zweite Teil des Buches beginnt mit einem Kapitel, welches sich mit dem Schätzen von Werten und der Vorhersage der Schätzfehler beschäftigt. Im workshop wird hierbei die Gnu Scientific Library (GSL) vorgestellt.

Danach folgt ein Kapitel, welches sich mit der Modellierung von Grö0enrelationen beschäftigt. Hierbei wird auch darauf eingegangen, was ein gutes Modell auszeichnet und wo seine eventuellen Grenzen liegen. In diesem Kapitel wird auch das CAS-Tool Sage vorgestellt.

Iim darauffolgenden Kapitel werden dann Systeme beleuchte, deren Daten auf Zufallsereignissen beruhen. In diesem Kapitelwird auch das Statistiksystem R vorgestellt.

Der Abschluss des zweiten Teils bildet ein Kapitel, welches sich mit dem unüberlegten Gebrauch bekannter Begriffen der beschreibenden Statistik wie dem Mittelwert und der Standardabweichung beschäftigt.

Der dritte Teil des Buches beschäftigt sich mit Methoden, die allgemein als Data Mining bekannt sind.

Im ersten Kapitel des dritten Abschnittes wird der Einsatz von Simulationen zur Verfifkation und Verständnisbildung von Modellen erläutert. Weiterhin werden die Grenzen des Einsatzes von Simulationen aufgezeigt. Im Workshop dieses Kapitels wird das Python-Modul SimPy vorgestellt.

In dem darauffolgenden Kapitel wird die Untersuchung von Clustenn dargestellt. Als Bibliothek wird im Workshop PythonCluster vorgestellt.

Der vierten Teil beginnt mit einem Kapitel, der sich mit den Begriffen Datenwarehouse und Buisiness Inteligence (BI) beschäftigt. In diesem Kapitelgeht der Autor auch mit den

kommerziell erhältlichen Reporting Tools hart ins Gericht. Im Workshop werden die Datenbanken Berkeley DB und SQLite vorgestellt.

Da vielen Anforderungen bzgl. Datenanalyse aus einem geschäftlichen Umfeld erwachsen, folgt ein Kapitel mit einer Einführung in Finanzmathematik. Letztendlich geht es darum, alle ein- und ausgehende Geldströme abzuzinsen, um beurteilen zu können, ob sich ein Investment rentiert. Selbstverständlich stellt der Autor klar, dass auch andere Faktoren, die sich u. U. sich nit in einem Geldbetrag ausdrücken lassen, Einfluss auf die Entscheidungsfindung haben können.

Im Anhang wird zuerst eine Zusammenstellung über verbreitete Softwaresystem zum wissenschaftlichen Arbeiten zusammengetragen. Was mir hierbei auffiel ist die Erwähnung des kommerzeilen Tools Matlab in einem Buch, welches sich in seinem titel ausdrücklich auf OpenSource-Tools bezieht. Weiterhin enthält der Anhang noch ein Kapitel, in dem mathematische Grundlagen, die für das Verständnis des Buches hilfreich sind, wiederholt werden.

Fazit:

Das Buch streift ein breites Spektrum von Verfahren, welche für die Datenanalyse von Bedeutung sind, wobei aufgrund der Breite des Stoffs manche Themen nur stichpunktartig angerissen werden können. Bzgl. der mathematischen Voraussetzung enthält das Buch durchaus Kapitel, für die ein Verständnis von mathematischen Themen oberhalb des Ab iturniveaus (z. B. Lineare Algebra, Eigenwerte und –vektoren von Matrizen) vorteilhaft ist. Auch die Auswahl der in den Workshops der einzelnen Kapiteln vorgestellten Tools ist teilweise nicht nachvollziehbar.

Wer noch keine Erfahrung in diesem Bereich hat, sollte u. U. eine deutschsprachige Einführung bevorzugen. In der „Von Kopf bis Fuß“-Reihe des selben Verlages gibt es ein deutschsprachiges Buch zum Thema Datenanalyse.