

<<基于开源工具的数据分析>>

图书基本信息

书名：<<基于开源工具的数据分析>>

13位ISBN编号：9787564126742

10位ISBN编号：7564126744

出版时间：2011-5

出版时间：东南大学

作者：Philipp K. Janert

页数：509

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<基于开源工具的数据分析>>

### 内容概要

数据收集相对比较简单，而要把原始信息转化为有用的数据则需要知道如何精确地抽取你想要的内容。

通过这本书(作者Philipp

K. Janert)的深入讲解，那些对数据分析感兴趣的中等或者富有经验的程序员将可以学习到在商业环境中与数据打交道的技术。

你将了解到如何观察数据来找出它所包含的信息，如何在概念模型里捕捉到这些想法，然后把你的理解通过商业计划、度量标准的精确报告和其他方式反馈给你所在的机构。

你将会通过本书每章结束部分的动手实践来慢慢体验各种概念。

最重要的是，你将了解到如何思考你所希望获取的数据——而不是依赖于工具来替你思考。

## <<基于开源工具的数据分析>>

### 作者简介

Philipp

K. Janert目前提供数据分析和数学模型的咨询服务，他曾经是物理学家和软件工程师。

他是《Gnuplot in

Action : Understanding Data with Graphs》(Manning出版)的作者，他为O ' Reilly

Network . IBM

deVeloperWorks和IEEEsoftware写过文章。

他拥有Washington大学理论物理学的博士学位。

<<基于开源工具的数据分析>>

书籍目录

PREFACE

1 INTRODUCTION

Data Analysis

What's in This Book

What's with the Workshops?

What's with the Math?

What You'll Need

What's Missing

PART I Graphics: Looking at Data

2 A SINGLE VARIABLE: SHAPE AND DISTRIBUTION

Dot and Jitter Plots

Histograms and Kernel Density Estimates

The Cumulative Distribution Function

Rank-Order Plots and Lilliefors Tests

Only When Appropriate: Summary Statistics and Box Plots

Workshop: NumPy

Further Reading

3 TWO VARIABLES: ESTABLISHING RELATIONSHIPS

Scatter Plots

Conquering Noise: Smoothing

Logarithmic Plots

Banking

Linear Regression and All That

Showing What's Important

Graphical Analysis and Presentation Graphics

Workshop: matplotlib

Further Reading

TIME AS A VARIABLE: TIME-SERIES ANALYSIS

Examples

The Task

Smoothing

Don't Overlook the Obvious!

The Correlation Function

Optional: Filters and Convolutions

Workshop: scipy.signal

Further Reading

5 MORE THAN TWO VARIABLES: GRAPHICAL MULTIVARIATE ANALYSIS

False-Color Plots

A Lot at a Glance: Multiplots

Composition Problems

Novel Plot Types

Interactive Explorations

Workshop: Tools for Multivariate Graphics

Further Reading

6 INTERMEZZO: A DATA ANALYSIS SESSION

<<基于开源工具的数据分析>>

A Data Analysis Session

Workshop: gnuplot

Further Reading

PART II Analytic: Modeling Data

7 GUESSTIMATION AND THE BACK OF THE ENVELOPE

Principles of Guesstimation

How Good Are Those Numbers?

Optional: A Closer Look at Perturbation Theory and

Error Propagation

Workshop: The Gnu Scientific Library (GSL)

Further Reading

8 MODELS FROM SCALING ARGUMENTS

Models

Arguments from Scale

Mean-Field Approximations

Common Time-Evolution Scenarios

Case Study: How Many Servers Are Best?

Why Modeling?

Workshop: Sage

Further Reading

9 ARGUMENTS FROM PROBABILITY MODELS

The Binomial Distribution and Bernoulli Trials

The Gaussian Distribution and the Central Limit Theorem

Power-Law Distributions and Non-Normal Statistics

Other Distributions

Optional: Case Study--Unique Visitors over Time

Workshop: Power-Law Distributions

Further Reading

10 WHAT YOU REALLY NEED TO KNOW ABOUT CLASSICAL STATISTICS

Genesis

Statistics Defined

Statistics Explained

Controlled Experiments Versus Observational Studies

Optional: Bayesian Statistics--The Other Point of View

Workshop: R

Further Reading

11 INTERMEZZO: MYTHBUSTING--BIGFOOT, LEAST SQUARES, AND ALL THAT

How to Average Averages

The Standard Deviation

Least Squares

Further Reading

PART III Computation: Mining Data

12 SIMULATIONS

A Warm-Up Question

Monte Carlo Simulations

Resampling Methods

## <<基于开源工具的数据分析>>

Workshop: Discrete Event Simulations with Simpy

Further Reading

13 FINDING CLUSTERS

What Constitutes a Cluster?

Distance and Similarity Measures

Clustering Methods

Pre-and Postprocessing

Other Thoughts

A Special Case: Market Basket Analysis

A Word of Warning

Workshop: Python Cluster and the C Clustering Library

Further Reading

14 SEEING THE FOREST FOR THE TREES: FINDING

IMPORTANT ATTRIBUTES

Principal Component Analysis

Visual Techniques

Kohonen Maps

Workshop: PCA with R

Further Reading

15 INTERMEZZO: WHEN MORE IS DIFFERENT

A Horror Story

Some Suggestions

What About Map/Reduce?

Workshop: Generating Permutations

Further Reading

PART IV Applications: Using Data

16 REPORTING, BUSINESS INTELLIGENCE, AND DASHBOARDS

Business Intelligence

Corporate Metrics and Dashboards

Data Quality Issues

Workshop: Berkeley DB and SQLite

Further Reading

17 FINANCIAL CALCULATIONS AND MODELING

The Time Value of Money

Uncertainty in Planning and Opportunity Costs

Cost Concepts and Depreciation

Should You Care?

Is This All That Matters?

Workshop: The News Vendor Problem

Further Reading

18 PREDICTIVE ANALYTICS

Introduction

Some Classification Terminology

Algorithms for Classification

The Process

The Secret Sauce

The Nature of Statistical Learning

<<基于开源工具的数据分析>>

Workshop: Two Do-It-Yourself Classifiers

Further Reading

19 EPILOGUE: FACTS ARE NOT REALITY

A PROGRAMMING ENVIRONMENTS FOR SCIENTIFIC COMPUTATION  
AND DATA ANALYSIS

Software Tools

A Catalog of Scientific Software

Writing Your Own

Further Reading

B RESULTS FROM CALCULUS

Common Functions

Calculus

Useful Tricks

Notation and Basic Math

Where to Go from Here

Further Reading

WORKING WITH DATA

Sources for Data

Cleaning and Conditioning

Sampling

Data File Formats

The Care and Feeding of Your Data Zoo

Skills

Terminology

Further Reading

INDEX

## <<基于开源工具的数据分析>>

### 编辑推荐

《基于开源工具的数据分析(影印版)》(作者Philipp K . Janert)使用图形来描述带有一个、两个或者十多个变量的数据；使用粗略计算以及维度和概率参数来开发概念模型；使用诸如模拟和聚类的集约计算方法来挖掘数据；通过报告、信息板和其他度量程序来让你的结论更容易理解；理解财务计算，包括货币时间价值；利用降维技术或者预测分析来克服数据分析过程中面临的挑战；熟悉数据分析的不同开源编程环境。

<<基于开源工具的数据分析>>

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>