Wie Data Science und Python zusammenpassen

1

In diesem Kapitel

- Die Entdeckung der Vorzüge von Data Science
- Wie Data Science funktioniert
- Die Verbindung zwischen Python und Data Science
- Der Start mit Python

Data Science scheint eine Technologie zu sein, von der Sie glauben, sie niemals zu benötigen, aber da liegen Sie falsch. Ja, Data Science beinhaltet die Anwendung erweiterter Mathematikkenntnisse, wie Statistik oder Big Data. Data Science unterstützt Sie jedoch auch darin, intelligente Lösungen zu finden sowie Vorschläge auf Grundlage früherer Entscheidungen zu entwerfen, und hilft dabei, dass Roboter Objekte erkennen können. Tatsächlich nutzen die Leute Data Science auf unterschiedlichsten Wegen, die Sie buchstäblich nicht sehen können, oder sie tun etwas, ohne dass Sie die Auswirkungen dessen überhaupt mitbekommen. Kurz, Data Science ist die Person hinter dem Wunder der Technik. Ohne Data Science wäre vieles, was Sie als alltäglich empfinden, nicht möglich. Dies ist der Grund dafür, dass ein Data Scientist den geilsten Job des 21. Jahrhunderts hat.

Um Data Science für jeden nutzbar zu machen, der kein Mathematikgenie ist, werden Werkzeuge benötigt. Ihnen stehen zahlreiche solcher Werkzeuge zum Lösen von Data-Science-Aufgaben zur Verfügung, aber Python ist bestens geeignet, um die Arbeit mit solchen Daten zu vereinfachen. Zum einen stellt Python eine unglaubliche Anzahl von Mathematik-assoziierten Bibliotheken zur Verfügung, die Ihnen helfen werden, Aufgaben zu meistern, ohne genau wissen zu müssen, was passiert. Außerdem bietet Python Mehrfachkodierung und andere Dinge, die Ihnen Ihre Arbeit erleichtern. Selbstverständlich könnten Sie auch andere Sprachen nutzen, um Anwendungen für Data Science zu schreiben, aber Python reduziert Ihren Arbeitsaufwand, weshalb es das Mittel der Wahl für jene ist, die nicht wirklich schwer arbeiten wollen und es sich gerne einfach machen.

Dieses Kapitel ermöglicht Ihnen den Einstieg in Python. Obwohl dieses Buch kein komplettes Python-Tutorial enthält, wird die Erklärung einiger grundlegender Python-Ausgaben die Zeit für Ihren Einstieg verkürzen (wenn Sie ein gutes Tutorial für den Anfang benötigen, nutzen Sie bitte *Python programmieren lernen für Dummies*). Sie werden sehen, dass dieses Buch Verweise auf Tutorials und andere Hilfen zur Verfügung stellt, um Ihre Wissenslücken über Python zu schließen.

Die Wahl einer Data-Science-Sprache

Es gibt viele unterschiedliche Programmiersprachen auf der Welt – und die meisten davon wurden entwickelt, um Aufgaben auf eine bestimmte Weise zu lösen oder um die Lösbarkeit bestimmter Aufgabenbereiche zu vereinfachen. Wenn Sie die richtigen Werkzeuge benutzen, wird Ihr Leben leichter. Vergleichbar ist das mit einem Hammer, den Sie anstelle des Schraubenziehers zum Festdrehen einer Schraube benutzen. Sicher, der Hammer funktioniert, aber mit dem Schraubenzieher wird es viel einfacher und effektiver funktionieren. Data Scientists benutzen für gewöhnlich nur ein paar Sprachen, die die Arbeit mit ihren Daten vereinfachen. Daher werden hier die vier am meisten genutzten Sprachen für Data Science in der Reihenfolge ihrer bevorzugten Nutzung aufgeführt (von 91 Prozent der Wissenschaftler verwendet):

- ✔ **Python (allgemeine Verwendung):** Viele Wissenschaftler bevorzugen Python, da es viele Bibliotheken, wie NumPy, SciPy, MatPlotLib oder Scikit-learn zur Verfügung stellt, die die Arbeit mit Daten wesentlich vereinfachen. Python ist eine exakte Sprache, die es leicht macht, mit großen Datensätzen zu arbeiten, und zudem die Rechenzeit verkürzt. Die Data-Science-Community hat spezialisierte IDEs wie Anaconda entwickelt, die das IPython-Notebook-Konzept implementieren. Dadurch wird die Arbeit mit Kalkulationen signifikant vereinfacht (Kapitel 3 beschreibt die Verwendung von IPython, also keine Angst vor diesem Kapitel). Neben all diesen Vorzügen ist Python eine hervorragende Sprache, um Verbindungen mit Sprachen wie C++ oder Fortran herzustellen. Die aktuelle Python-Dokumentation beschreibt die Erstellung dieser Erweiterungen. Python findet Anwendung in diversen wissenschaftlichen Zusammenhängen.
- ✔ **R (speziell statistische Anwendung):** In vielerlei Hinsicht teilen sich Python und R die gleiche Art der Funktionalität, aber sie implementieren diese unterschiedlich. Abhängig von der Datengrundlage haben Python und R etwa die gleiche Anzahl an Befürwortern und manche Leute nutzen Python und R im Austausch (oder manchmal hintereinander). Anders als Python hat R eine eigene Entwicklungsumgebung, sodass keine dritte Plattform wie Anaconda benötigt wird. Allerdings stellt R nicht wie Python die Möglichkeit der Verbindung mit anderen Sprachen zur Verfügung.
- ✔ **SAS (statistische Businessanalysen):** Die Sprache des Statistischen Analyse Systems (SAS) ist beliebt, da Datenanalysen, Business Intelligence, Datenmanagement und vorhersagende Analysen sehr einfach sind. Das SAS Institute entwickelte SAS ursprünglich als Mittel zur Durchführung statistischer Analysen. Es ist daher eine Business-spezifische Sprache – man verwendet sie für Analysen, anstatt diese per Hand durchzuführen, oder um spezifische Muster zu erkennen.
- ✔ **SQL (Datenbankmanagement):** Das Wichtigste, das man über Structured Query Language (SQL) wissen sollte, ist, dass der Fokus auf den Daten liegt und nicht auf den zu erfüllenden Aufgaben. Unternehmen funktionieren nicht ohne ein gutes Datenmanagement – die Daten sind das Unternehmen. Große Organisationen nutzen eine Art relationale Datenbank, die normalerweise mit SQL zugänglich ist, um ihre Daten zu speichern. Die meisten Produkte von Datenbank-Management-Systemen (DBMS) ba-

sieren auf SQL als Hauptsprache, und DBMS bringen normalerweise eine große Anzahl an Datenanalyse- und Data-Science-Funktionen mit. Wenn Sie einen nativen Zugriff auf die Daten haben, wird die Geschwindigkeit oftmals erhöht, wenn Sie die Daten auf diese Weise auswerten. Datenbank-Administratoren (DBAs) nutzen SQL für gewöhnlich, um Daten zu organisieren oder zu verändern, und nicht unbedingt, um detaillierte Analysen durchzuführen. Der Data Scientist kann SQL aber genauso für unterschiedlichste Data-Science-Anwendungen nutzen und die resultierenden Skripte für die DBAs und ihren Bedarf zur Verfügung stellen.

Die Definition des geilsten Jobs des 21. Jahrhunderts

Allgemein werden Menschen, die mit Statistik arbeiten, als eine Art Buchhalter angesehen oder als verrückte Wissenschaftler. Viele finden Statistik und die Analyse von Daten langweilig. Allerdings ist Data Science eine jener Tätigkeiten, bei der Sie umso mehr wissen wollen, je mehr Sie lernen. Die Beantwortung einer Frage wirft oftmals mehr Fragen auf, die noch interessanter sind als jene, die Sie beantworten wollten. Die Sache, die Data Science so sexy macht, ist, dass Sie es überall antreffen und in einer schier unendlichen Anzahl von Möglichkeiten einsetzen können. Die folgenden Abschnitte beinhalten detaillierte Informationen darüber, warum Data Science ein so verblüffender Forschungsbereich ist.

Die Entstehung von Data Science

Data Science ist ein relativ neuer Begriff. William S. Cleaveland prägte den Begriff 2001 als Teil seiner Veröffentlichung »Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics«. Erst ein Jahr später erkannte das International Council for Science den Begriff Data Science an und rief ein Komitee dafür ins Leben. Die Columbia University begann im Jahr 2003 mit der Veröffentlichung des *Journal of Data Science*.

Gleichwohl ist die mathematische Basis hinter Data Science Jahrhunderte alt und im Wesentlichen ein Verfahren der Analyse von Statistiken und Wahrscheinlichkeiten. Der erste Gebrauch des Begriffes Statistik geht auf das Jahr 1749 zurück, der Begriff ist aber sicherlich viel älter. Die Menschen nutzen die Statistik zur Mustererkennung seit Tausenden von Jahren. Der Historiker Thucydides beschreibt (in der Geschichte des Peloponnesischen Krieges), wie die Athener die Höhe der Mauer von Platea im 5. Jahrhundert vor Christus berechneten, indem sie die Ziegelsteine in einem unverputzten Teil der Mauer zählten. Da die Zählung genau sein musste, berechneten sie den Durchschnitt aus mehreren Zählungen.

Der Prozess der Quantifizierung und des Verständnisses der Statistik ist relativ neu, allerdings ist die Wissenschaft selbst ziemlich alt. Ein früher Versuch, die Bedeutung der Statistik zu dokumentieren, war das *Manuscript on Deciphering Cryptographic Messages* von Al-Kindi. In dieser Veröffentlichung beschreibt Al-Kindi die Verwendung einer Kombination aus Statistik und Frequenzanalyse zur Entschlüsselung geheimer Botschaften. Schon damals wurde der

Nutzen von Statistik in der praktischen Anwendung auf Aufgaben gesehen, die nahezu unlösbar schienen. Data Science führt diesen Prozess weiter und für manche Menschen sieht es aktuell noch nach Magie aus.

Umriss der Kernkompetenzen eines Data Scientists

Wie heutzutage in den meisten Berufen muss ein Data Scientist ein breit gefächertes Wissen sowie bestimmte Fähigkeiten besitzen, um die geforderten Aufgaben erfüllen zu können. Es sind sogar so viele Anforderungen gefragt, dass Data Scientists oft in Teams arbeiten. Jemand, der sich gut mit der Sammlung von Daten auskennt, könnte ein Team mit einem Analysten und jemandem, der sich mit der Darstellung von Daten auskennt, bilden. Es ist schwierig, eine Person zu finden, die sämtliche geforderten Fähigkeiten besitzt. Die folgende Liste zeigt Bereiche, in denen sich ein Data Scientist auszeichnen kann (je mehr Kompetenzen, umso besser):

- ✔ **Datenerfassung:** Es ist unerheblich, welche Art von Mathekenntnissen Sie haben, wenn Sie keine Daten haben, die Sie analysieren können. Der Prozess der Datenerfassung beginnt mit der Verwaltung einer Datenquelle mit Datenbank-Management-Fähigkeiten. Rohdaten sind in vielen Fällen nicht sonderlich hilfreich – Sie müssen also die Herkunft der Daten kennen und beim Betrachten dieser Daten bereits Fragestellungen formulieren können. Schließlich benötigen Sie Fähigkeiten zur Modellierung von Daten, sodass Sie verstehen, wie die Daten miteinander in Verbindung stehen und wie sie strukturiert sind.
- ✔ **Analyse:** Nachdem Ihnen die zu verarbeiteten Daten vorliegen und Sie ihre Komplexität verstanden haben, können Sie mit der Analyse beginnen. Sie führen Analysen mithilfe von Tools sowie Ihrem Wissen über Statistik durch, das Sie während Ihrer Ausbildung erworben haben. Der Gebrauch spezieller mathematischer Operationen und Algorithmen kann eindeutige Muster in Ihren Daten hervorbringen oder Ihnen helfen, Schlussfolgerungen zu ziehen, die Sie mit bloßem Blick auf die Daten nicht hätten ziehen können.
- ✔ **Präsentation:** Viele Menschen können mit Zahlen nicht viel anfangen. Sie sehen einfach die Muster nicht, die ein Wissenschaftler sieht. Es ist daher wichtig, eine grafische Präsentation dieser Muster vorzubereiten, um zu visualisieren, was die Zahlen bedeuten und wie sie sinnvoll verwendet werden können. Viel wichtiger ist aber, dass die Präsentation so gestaltet ist, dass die Wirkung der Daten klar wird.

Die Verbindung von Data Science und Big Data

Die Kunst ist es, Daten so vorzubereiten, dass jeder sie für seine Analysen verwenden, das heißt extrahieren, transformieren und laden kann (ETL). Spezialisten nutzen Programmiersprachen wie Python, um die Daten aus unterschiedlichen Quellen zu extrahieren. Ein Unternehmen neigt dazu, die Daten an mehreren Speicherorten abzulegen, wodurch die Analysen zeitaufwendig sein können. Nachdem der ETL-Spezialist die Daten gefunden hat, transformiert eine Programmiersprache oder ein anderes Tool diese in ein gebräuchliches Format für weitere Analysen. Der Ladeprozess kann viele Formen annehmen, wobei sich dieses Buch ausschließlich auf Python beschränkt. Bei großen Projekten werden Sie möglicherweise auch Tools wie Informatica, MS SSIS oder Teradata für Ihre Zwecke verwenden wollen.

Das Verständnis der Rolle der Programmierung

Ein Data Scientist muss unterschiedliche Programmiersprachen kennen, um unterschiedliche Ziele zu erreichen. Zum Beispiel benötigen Sie SQL-Kenntnisse, um Daten aus relationalen Datenbanken zu extrahieren. Python hilft Ihnen beim Laden und Transformieren sowie bei Analysen. Trotzdem sollten Sie ein Tool wie MATLAB (das seine eigene Programmiersprache besitzt) oder PowerPoint (basierend auf VBA) wählen, um die Informationen präsentieren zu können. (Wenn es Sie interessiert, wie MATLAB im Vergleich zu Python funktioniert, sollten Sie das Buch *MATLAB für Dummies* heranziehen.) Die immense Menge an Daten, mit denen ein Data Scientist arbeitet, muss mehreren Stufen mit redundanten Analyseschritten unterzogen werden, um sie in verwertbare Daten umzuwandeln. Die manuelle Ausführung dieser Aufgaben ist zeitaufwendig und fehleranfällig, sodass Programmierung die beste Methode ist, um das Ziel des Erhalts von schlüssigen und verwendbaren Daten zu erreichen. Bei der Anzahl an Werkzeugen, die ein Data Scientist nutzt, ist es nicht möglich, sich nur auf eine Programmiersprache zu beschränken. Sicherlich, Python kann Daten laden, transformieren, analysieren und dem Nutzer präsentieren, aber das ist nur möglich, weil die Sprache die erforderlichen Funktionen bereitstellt. Sie werden andere Programmiersprachen nutzen müssen, um Ihren Werkzeugkasten zu vervollständigen. Welche Sprache Sie wählen sollten, ist von vielerlei Kriterien abhängig. Hier sind die Dinge, die es zu beachten gilt:

- ✔ Wie Sie Data Science für Ihren Code zu verwenden beabsichtigen (Sie müssen eine Reihe von Aufgaben wie Datenanalyse, Klassifizierung und Regression einbeziehen).
- ✔ Ihre Vertrautheit mit der Sprache.
- ✔ Die Notwendigkeit der Interaktion mit anderen Sprachen.
- ✔ Die Verfügbarkeit von Tools zur Vereinfachung der Entwicklungsumgebung.
- ✔ Die Verfügbarkeit von APIs und Bibliotheken, um Aufgaben leichter zu lösen.

Die Entwicklung einer Data-Science-Pipeline

Data Science ist teilweise Kunst und teilweise Ingenieursarbeit. Die Erkennung von Mustern innerhalb der Daten unter Einbeziehung der zu beantwortenden Frage und die Untersuchung, welcher Algorithmus am besten geeignet ist, sind Teil der Data-Science-Kunst. Um Data Science realisierbar zu machen, basiert der Teil der Ingenieursarbeit auf einem speziellen Prozess, um das Ziel zu erreichen. Diesen Prozess beschreibt die Data-Science-Pipeline. Es ist erforderlich, dass der Data Scientist besondere Schritte während der Vorbereitung, Analyse und Präsentation der Daten durchführt. Die folgenden Abschnitte beschreiben eine Data-Science-Pipeline genauer, damit Sie verstehen, wie in diesem Buch die Beispiele veranschaulicht werden.

Vorbereitung der Daten

Daten aus unterschiedlichen Quellen erhalten Sie nicht einfach so zusammengepackt, dass Sie direkt mit der Analyse beginnen können – ganz im Gegenteil. Die Rohdaten kommen in unterschiedlichen Formaten, und Sie müssen sie erst mal umwandeln, um sie für weitere

Analysen verfügbar zu machen. Die Transformation erfordert eine Änderung der Datentypen, die Festlegung der Reihenfolge, in der die Daten verarbeitet werden, sowie das Anlegen von Dateneinträgen, basierend auf den Informationen, die für bereits existierende Einträge zur Verfügung stehen.

Darstellung der beschreibenden Datenanalyse

Die Mathematik hinter der Datenanalyse basiert auf Prinzipien des Ingenieurwesens, die Ergebnisse müssen beweisbar und konsistent sein. Data Science ermöglicht den Zugang zu einer Fülle statistischer Methoden und Algorithmen, die Ihnen helfen werden, Muster in den Daten zu erkennen. Ein einziger Ansatz löst für gewöhnlich nicht das Problem. Sie werden typischerweise einen iterativen Prozess nutzen, um die Daten aus unterschiedlichen Sichtweisen zu bearbeiten. Der Gebrauch von Versuch und Irrtum ist ein Teil der Kunst von Data Science.

Von den Daten lernen

Wenn Sie verschiedene statistische Analysemethoden durchlaufen und Algorithmen anwenden, um Muster zu finden, werden Sie von den Daten lernen. Möglicherweise sagen Ihnen die Daten gar nicht das, was Sie erwartet haben, oder Sie erzählen Ihnen viele verschiedene Geschichten. Neues zu entdecken, ist ein Teil dessen, was einen Data Scientist ausmacht. Tatsächlich ist es sogar der lustigste Teil von Data Science, weil Sie im Voraus niemals genau wissen werden, was die Daten Ihnen offenbaren werden.

Sicherlich, der unsichere Ursprung der Daten und das Finden scheinbar zufälliger Muster darin erfordert einen klaren Kopf. Wenn Sie zu wissen glauben, was die Daten beinhalten, werden Sie nicht auf die Informationen stoßen, die sie wirklich beinhalten. Sie bringen sich um den Entdeckungsprozess und damit um Möglichkeiten, die wichtig sein können für Sie und auch für von Ihnen abhängige Personen.

Visualisierung

Visualisierung bedeutet, Muster in den Daten zu erkennen und darauf reagieren zu können. Sie bedeutet außerdem, zu erkennen, wenn ein Muster kein Teil der Daten ist. Sehen Sie sich als eine Art Bildhauer der Daten – das Entfernen der Daten, die sich nicht innerhalb eines Musters befinden (also der Ausreißer), ermöglicht anderen, das Kernstück der Informationen zu sehen. Ja, Sie können dieses Kernstück sehen, aber bis andere das auch tun, existiert es allein in Ihrem Kopf.

Erkenntnisse und Ergebnisse

Der Data Scientist scheint durch bloßes Betrachten der Daten auf Methoden zurückgreifen zu können. Der Prozess endet jedoch nicht, bis Sie ganz sicher verstehen, was die Daten aussagen. Die Erkenntnisse, die Sie aus der Manipulation und Analyse der Daten gewinnen, werden Ihnen helfen, realistische Aufgaben zu lösen. Sie können beispielsweise die Ergebnisse einer

Analyse nutzen, um eine geschäftliche Entscheidung zu treffen. Manchmal wird durch das Ergebnis einer Analyse eine automatische Antwort erzeugt. Wenn ein Roboter beispielsweise eine Ansammlung von Pixeln mithilfe einer Kamera erkennt, haben die Pixel, die ein Objekt erzeugen, eine spezielle Bedeutung. Die Programmierung des Roboters führt zu einer Interaktion mit diesem Objekt. Erst wenn der Data Scientist eine Anwendung programmiert hat, die laden, analysieren und Pixel visualisieren kann, sieht der Roboter überhaupt etwas.

Die Rolle von Python in Data Science

Mit der richtigen Datenquelle sowie den Analyse- und Präsentationsanforderungen können Sie Python an jeder Stelle einer Data-Science-Pipeline nutzen. In der Tat ist es genau das, was Sie mit diesem Buch tun. Jedes Beispiel nutzt Python, um Ihnen einen weiteren Teil der Data-Science-Gleichung zu erläutern. Von allen Programmiersprachen, die Sie für Data-Science-Aufgaben verwenden könnten, ist Python die flexibelste und leistungsfähigste, da sie über viele zusätzliche Bibliotheken verfügt. Die folgenden Abschnitte werden Ihnen zeigen, weshalb Python eine so gute Wahl für so viele (eigentlich die meisten) Data-Science-Anwendungen ist.

Das sich wandelnde Profil eines Data Scientists

Manche Leute sehen Data Scientists als unnahbare Nerds, die Wunder mit Daten und Mathematik vollführen. Der Data Scientist ist die Person hinter dem Vorhang bei einem Oz-ähnlichen Schauspiel. Diese Sichtweise beginnt sich zu verändern. In vielerlei Hinsicht sieht die Welt den Data Scientist jetzt entweder als Ergänzung zum Entwickler oder ganz neuen Typ von Entwickler. Der Vormarsch von lernenden Systemen ist der Kern dieser Veränderung. Damit eine Anwendung lernen kann, müssen große Datenmengen verändert und neue Muster darin gefunden werden. Zusätzlich muss die Anwendung in der Lage sein, neue Daten auf Grundlage der alten Daten zu erzeugen – als eine Art Vorhersage der Möglichkeiten. Diese neue Art der Anwendung beeinflusst die Menschen in einer Weise, die bis vor ein paar Jahren noch wie Science-Fiction gewirkt hätte. Sicherlich die bemerkenswertesten dieser Anwendungen definieren das Verhalten von Robotern, die in der Zukunft bestimmt viel enger als bisher mit Menschen zusammenarbeiten werden.

Aus der geschäftlichen Perspektive ist die notwendige Verschmelzung von Data Science und Anwendungsentwicklung offensichtlich: Im Geschäft muss eine Vielzahl von Analysen mit großen Datenmengen durchgeführt beziehungsweise gesammelt werden – um Sinn in die Informationen zu bringen und sie zukünftig für Vorhersagen zu nutzen. In Wahrheit aber liegen die weit größeren Auswirkungen der Verschmelzung der beiden Wissenschaften – Data Science und Anwendungsentwicklung – in der Schaffung völlig neuer Arten von Anwendungen, wovon einiges heute noch nicht einmal vorstellbar ist. Beispielsweise könnten solche neuen Anwendungen Lehrern durch die Analyse des Lernverhaltens der Schüler und der Schaffung neuer Lehrmethoden mit höherer Genauigkeit zeigen, welche Methoden für den einzelnen Schüler nützlich sind. Die Kombination der Wissenschaften könnte ebenso jede Menge Probleme in der Medizin lösen, die heute unmöglich lösbar scheinen – nicht nur beim In-Schach-Halten von Krankheiten, sondern auch bei der Lösung von Problemen, wie beispielsweise der Entwicklung von Prothesen, die aussehen und funktionieren wie natürliche Gliedmaßen.

Die Arbeit mit einer vielseitigen, einfachen und effizienten Sprache

Es gibt zahlreiche unterschiedliche Wege für die Lösung von Aufgaben in Data Science. Dieses Buch deckt nur eine unter vielen zur Verfügung stehenden Methoden ab. Allerdings bietet Python eine der wenigen Einzelkomplettlösungen, die Sie nutzen können, um komplexe Probleme aus dem Bereich Data Science zu lösen. Statt unterschiedlicher Tools können Sie dafür die einfache, leichte Sprache Python verwenden. Der Unterschied zu anderen Sprachen ist, dass Python eine Vielzahl wissenschaftlicher und mathematischer Bibliotheken mitbringt. Der Einsatz dieser Bibliotheken erweitert Python und ermöglicht die Lösung von Aufgaben, für die eine andere Sprache größeren Aufwand betreiben müsste.

Die Bibliotheken von Python sind das wichtigste Verkaufsargument, trotzdem bietet Python mehr als nur wiederverwendbaren Code. Das Wichtigste, das man bei der Arbeit mit Python beachten sollte, ist, dass vier unterschiedliche Typen von Code zur Verfügung stehen:

- ✔ **Funktional:** Jede Aussage wird als mathematische Gleichung behandelt und jede Form statischer oder veränderlicher Daten wird vermieden. Der große Vorteil dieses Ansatzes ist, dass keine Nebeneffekte zu beachten sind. Außerdem eignet sich dieser Stil des Programmierens besser als andere für parallele Verarbeitung, weil keine statischen Zustände zu beachten sind. Viele Entwickler bevorzugen diesen Programmierstil für Rekursion und das Lambda-Kalkül.
- ✔ **Imperativ:** Es werden Berechnungen als direkte Änderungen des Programmzustands durchgeführt. Diese Art ist besonders sinnvoll, wenn Datenstrukturen verändert werden müssen, aber dennoch eleganter, einfacher Code erstellt werden soll.
- ✔ **Objektorientiert:** Daten werden als Objekte behandelt und nur durch vorgeschriebene Methoden verändert. Python unterstützt diese Programmierform nicht komplett, da Merkmale wie das Data Hiding nicht implementiert werden können. Trotzdem ist es ein nützlicher Programmierstil für komplexe Anwendungen, da Verkapselungen und Polymorphie unterstützt werden. Diese Art der Programmierung begünstigt ebenfalls die Wiederverwendung von Code.
- ✔ **Verfahrensorientiert:** Aufgaben werden durch Iterationen Schritt für Schritt abgearbeitet, wobei häufige Operationen in Funktionen hinterlegt sind, die bei Bedarf aufgerufen werden. Dieser Programmierstil begünstigt Iteration, Sequenzierung, Auswahl und Modularisierung.

Der schnelle Einstieg in Python

Es ist an der Zeit, Python auszuprobieren, um eine Data-Science-Pipeline in Aktion zu erleben. Die folgenden Abschnitte stellen einen kurzen Überblick über den Prozess zur Verfügung, den Sie im Verlauf dieses Buches ausführlich betrachten werden. Sie müssen die Aufgaben der folgenden Abschnitte nicht tatsächlich lösen. Sie müssen Python bis Kapitel 3 nicht

einmal installieren, folgen Sie einfach dem Text. Seien Sie nicht beunruhigt, wenn Sie an dieser Stelle nicht gleich alles verstehen. Der Zweck dieser Abschnitte ist, Ihnen ein Verständnis dafür zu vermitteln, wie Python für Data Science angewendet wird. Viele Details werden Ihnen an dieser Stelle schwierig erscheinen, aber der Rest des Buches wird Ihnen helfen, alles zu verstehen.

Die Beispiele in diesem Buch beziehen sich auf eine webbasierte Anwendung, genannt IPython Notebook. Die Screenshots in diesem und den anderen Kapiteln zeigen, wie IPython Notebook in Firefox auf einem Windows-7-System aussieht. Die Daten, die Sie sehen werden, sind die gleichen, aber das eigentliche Interface ist abhängig von der Plattform (so, als würde man ein Notebook anstelle eines Desktopsystems nutzen), dem Betriebssystem und dem Browser. Machen Sie sich keine Gedanken, wenn Sie kleine Unterschiede zwischen Ihrer Anzeige und den Screenshots des Buches bemerken.

Sie müssen den Quellcode dieses Kapitels nicht per Hand eingeben. Leichter ist es, ihn sich herunterzuladen (schauen Sie in die Einleitung für Details zum Download des Quellcodes). Den Quellcode für dieses Kapitel finden Sie in der Datei `P4DS4D; 01; Quick Overview.ipynb`.

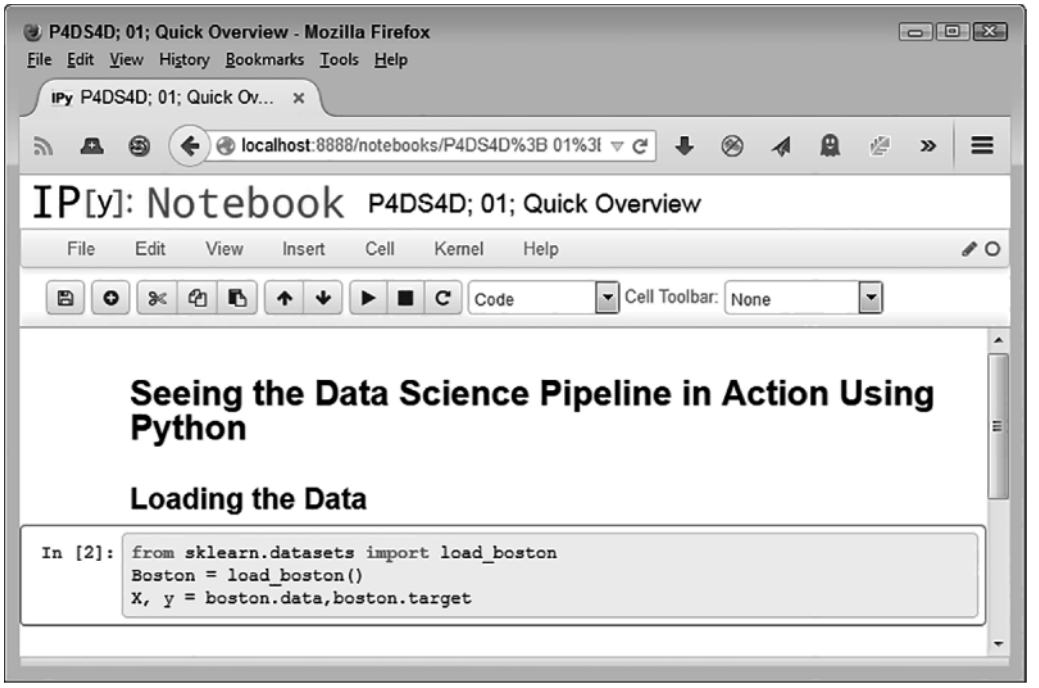

Abbildung 1.1: Sie laden Daten in Variablen, damit Sie sie bearbeiten können.

Daten laden

Bevor Sie irgendetwas tun können, müssen Sie einige Daten laden. Dieses Buch zeigt Ihnen alle Arten von Methoden für diese Aufgabe. In diesem Fall sehen Sie in Abbildung 1.1, wie Sie einen Datensatz laden. Dieser wurde *Boston* genannt und beinhaltet Immobilienpreise und andere Fakten über Gebäude in Boston. Der gesamte Datensatz wird in der Variablen `boston` hinterlegt und anschließend in den Variablen `x` und `y` abgelegt. Sie können sich Variablen wie Aufbewahrungsboxen vorstellen. Variablen sind wichtig, weil sie die Arbeit mit den Daten erst ermöglichen.

Ein Modell ableiten

Sobald die Daten vorliegen, können Sie damit arbeiten. In Python sind schon alle möglichen Algorithmen integriert. Abbildung 1.2 zeigt ein lineares Regressionsmodell. Machen Sie sich keine Gedanken, wenn Sie nicht wissen, was das ist. Ich gehe in späteren Kapiteln auf die lineare Regression ein. Worauf es hier ankommt, ist die Erkenntnis aus Abbildung 1.2, dass Python für die lineare Regression nur zwei Anweisungen benötigt und eine Ausgabevariable namens `hypothesis`.

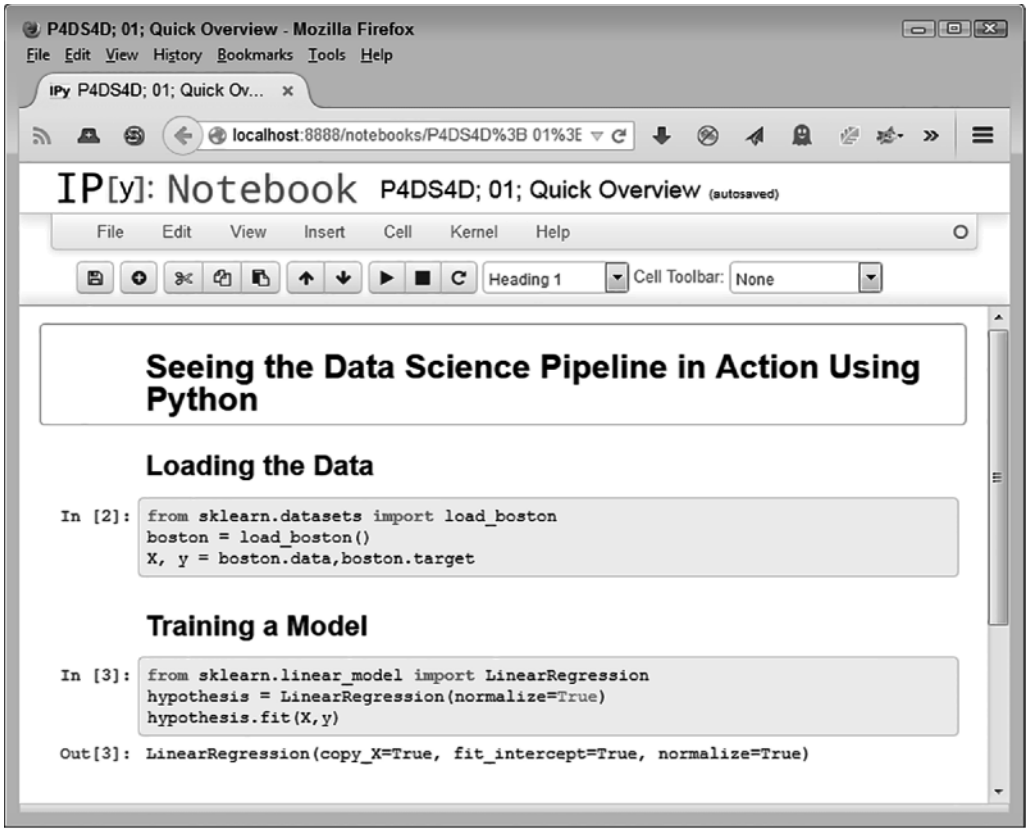

Abbildung 1.2: Verwendung des Variableninhalts für die Ableitung eines linearen Regressionsmodells

Anzeige eines Ergebnisses

Keine Art der Analyse lohnt sich, solange Sie keinen Nutzen in Form eines Ergebnisses erhalten. In diesem Buch werden alle Wege aufgezeigt, um sich eine Ausgabe anzeigen zu lassen, und in Abbildung beginnen wir mit etwas ganz Einfachem. In diesem Fall sehen Sie die Ausgabe der Koeffizienten einer linearen Regression.

Einer der Gründe dafür, dass in diesem Buch IPython Notebook verwendet wird, ist, dass dabei gut formatierte Ausgaben als Teil einer Anwendung entstehen. Schauen Sie auf Abbildung 1.3, und Sie sehen eine Darstellung, die Sie sofort ausdrucken und einem Arbeitskollegen zeigen könnten. Diese Ausgabe eignet sich nicht für jeden, aber diejenigen, die sich mit Python und Data Science auskennen, werden sie ziemlich nützlich und informativ finden.

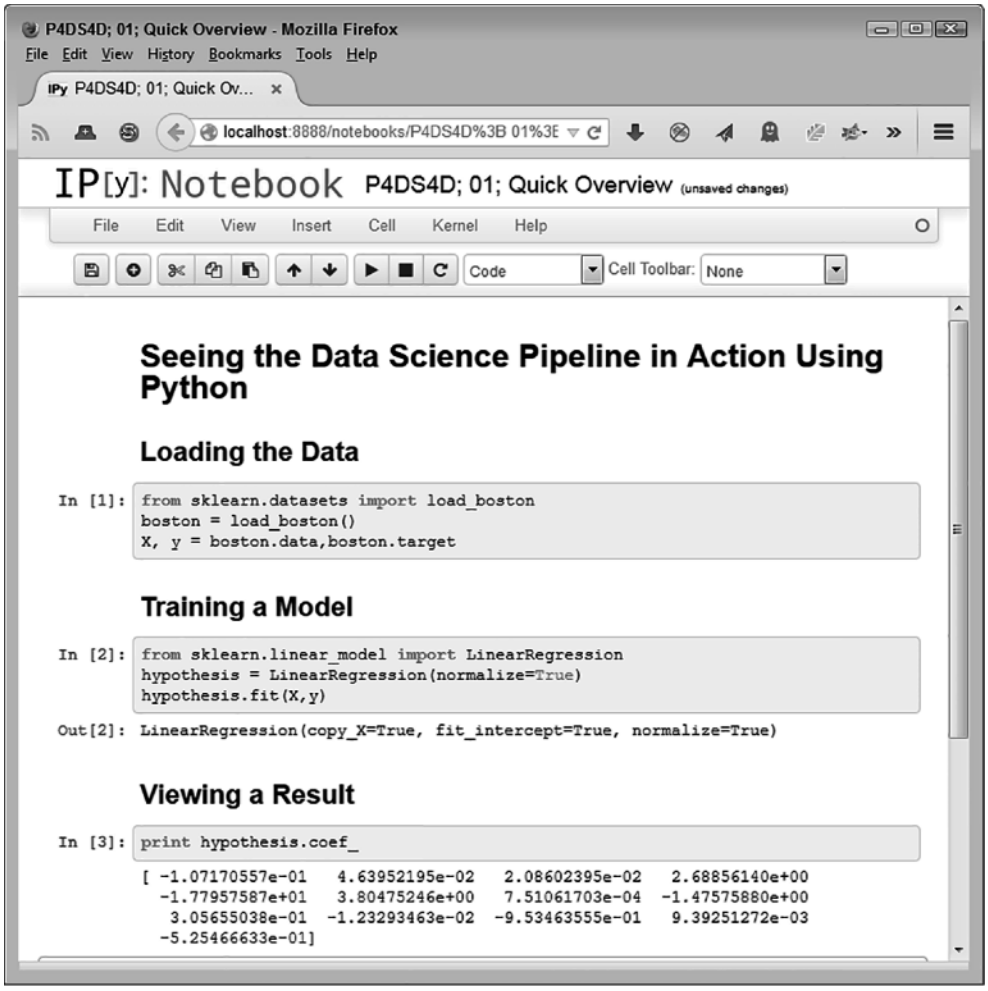

Abbildung 1.3: Ausgabe eines Ergebnisses als Antwort auf ein Modell