Python Business Intelligence Cookbook

The amount of data produced by businesses and devices is going nowhere but up. In this scenario, Python is an excellent tool for more specialized analysis tasks, and is powered with related libraries to process data streams, to visualize datasets, and to carry out scientific calculations.

Rather than spending day after day scouring Internet forums for "how-to" information, here you'll find more than 60 recipes that take you through the entire process of creating actionable intelligence from your raw data, no matter what shape or form it's in. Within the first 30 minutes of opening this book, you'll learn how to use the latest in Python and NoSQL databases to glean insights from data just waiting to be exploited.

What this book will do for you...

- Install Anaconda, MongoDB, and everything you need to get started with your data analysis
- Prepare data for analysis by querying, cleaning, and standardizing data
- Explore your data by creating a Pandas data frame from MongoDB
- Gain powerful insights, both statistical and predictive, to make informed business decisions
- Visualize your data by building dashboards and generating reports
- Create a complete data processing and business intelligence system

$ 39.99 US
£ 25.99 UK

Prices do not include local sales tax or VAT where applicable

Inside the Cookbook...
- A straightforward and easy-to-follow format
- A selection of the most important tasks and problems
- Carefully organized instructions to solve problems efficiently
- Clear explanations of what you did
- Solutions that can be applied to solve real-world problems

Leverage the computational power of Python with more than 60 recipes that arm you with the required skills to make informed business decisions

Free Sample
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 1 'Getting Set Up to Gain Business Intelligence'
- A synopsis of the book’s content
- More information on Python Business Intelligence Cookbook
About the Author

Robert Dempsey is a tested leader and technology professional who specializes in delivering solutions and products to solve tough business challenges. His experience of forming and leading agile teams, combined with more than 16 years of technology experience, enables him to solve complex problems while always keeping the bottom line in mind.

Robert has founded and built three start-ups in tech and marketing, developed and sold two online applications, consulted for Fortune 500 and Inc. 500 companies, and has spoken nationally and internationally on software development and agile project management.

He's the founder of Data Wranglers DC, a group that is dedicated to improving the craft of data engineering, as well as a board member of Data Community DC.

In addition to spending time with his growing family, Robert geeks out on Raspberry Pi, Arduinos, and automating more of his life through hardware and software.

Preface

Data! Everyone is surrounded by it, but few know how to truly exploit it. For those who do, glory awaits!

Okay, so that's a little dramatic; however, being able to turn raw data into actionable information is a goal that every organization is working to achieve. This book helps you achieve it.

Making sense of data isn't some esoteric art requiring multiple degrees—it's a matter of knowing the recipes to take your data through each stage of the process. It all starts with asking an interesting question.

My mission is that, by the end of this book, you will be equipped to apply Python to business intelligence tasks—preparing, exploring, analyzing, visualizing, and reporting—in order to make more informed business decisions using the data at hand.

Prepare for an awesome read, my friend!

A little context first. The code in this book is developed on Mac OS X 10.11.1, using Python 3.4.3, IPython 4.0.0, matplotlib 1.4.3, NumPy 1.9.1, scikit-learn 0.16.1, and Pandas 0.16.2—in other words, the latest or near-latest versions at the time of publishing.

What this book covers

Chapter 1, Getting Set Up to Gain Business Intelligence, covers a set of installation recipes and advice on how to install the required Python packages and libraries as well as MongoDB.

Chapter 2, Making Your Data All It Can Be, provides recipes to prepare data for analysis, including importing the data into MongoDB, cleaning the data, and standardizing it.

Chapter 3, Learning What Your Data Truly Holds, shows you how to explore your data by creating a Pandas DataFrame from a MongoDB query, creating a data quality report, generating summary statistics, and creating charts.
Chapter 4, *Performing Data Analysis for Non Data Analysts*, provides recipes to perform statistical and predictive analysis on your data.

Chapter 5, *Building a Business Intelligence Dashboard Quickly*, builds on everything that you've learned and shows you how to generate reports in Excel, and build web-based business intelligence dashboards.
Getting Set Up to Gain Business Intelligence

In this chapter, we will cover the following recipes:

- Installing Anaconda
- Installing, configuring, and running MongoDB
- Installing Rodeo
- Starting Rodeo
- Installing Robomongo
- Using Robomongo to query MongoDB
- Downloading the UK Road Safety Data dataset

Introduction

In this chapter, you'll get fully set up to perform business intelligence tasks with Python. We'll start by installing a distribution of Python called Anaconda. Next, we'll get MongoDB up and running for storing data. After that, we'll install additional Python libraries, install a GUI tool for MongoDB, and finally take a look at the dataset that we'll be using throughout this book.

Without further ado, let's get started!
Installing Anaconda

Throughout this book, we'll be using Python as the main tool for performing business intelligence tasks. This recipe shows you how to get a specific Python distribution—Anaconda, installed.

Getting ready

Regardless of which operating system you use, open a web browser and browse to the Anaconda download page at http://continuum.io/downloads.

The download page will automatically detect your operating system.

How to do it...

In this section, we have listed the steps to install Anaconda for all the major operating systems: Mac OS X, Windows, and Linux.

Mac OS X 10.10.4

1. Click on the I WANT PYTHON 3.4 link. We'll be using Python 3.4 throughout this book.
2. Next, click on the Mac OS X — 64-Bit Python 3.4 Graphical Installer button to download Anaconda.
3. Once the download completes, browse your computer to find the downloaded Anaconda, and double-click on the Anaconda installer file (a .pkg file) to begin the installation.
4. Walk through the installer steps to complete the installation. I recommend keeping the default settings.
5. To verify that Anaconda is installed correctly, open a terminal and type the following command:
   python
6. If the installer was successful, you should see something like this:

![Python 3.4 Command Prompt]

Windows 8.1

1. Click on the I WANT PYTHON 3.4 link. We'll be using Python 3.4 throughout this book.
2. Next, click on the Windows 64-Bit Python 3.4 Graphical Installer button to download Anaconda.
3. Once the download completes, browse your computer to find the downloaded Anaconda, and double-click on the Anaconda3-2.3.0-Windows-x86_64.exe file to begin the installation.
4. Walk through the installer steps to complete the installation. I recommend keeping the default settings.
5. To verify that Anaconda has installed correctly, open a terminal, or open a command prompt in Windows. Now type the following command:

   python
6. If the installation was successful, you should see something like this:

![Command Prompt](image)

**Linux Ubuntu server 14.04.2 LTS**

Linux servers have no graphical user interface (GUI), so you'll first need to log into your server and get a command prompt. With that complete, do the following:

1. On the Anaconda downloads page, select **Linux**.

2. Choose the **Python 3.4** link.

3. Right-click on the **Linux X 64-Bit** button, and copy the link.

4. At the command prompt on your server, use `curl` to download the file, pasting the following download link:
   
   ```
   curl -O <LINK TO DOWNLOAD>
   ```

5. I've created a special shortcut on my website that is a bit easier to type at the command line: [http://robertwdempsey.com/anaconda3-linux](http://robertwdempsey.com/anaconda3-linux).

6. Once Anaconda downloads, use the following command to start the installer:
   
   ```
   bash Anaconda3-2.3.0-Linux-x86_64.sh
   ```

7. Accept the license agreement to begin installation.

8. When asked if you would like Anaconda to prepend the Anaconda3 install location to the **PATH** variable, type yes.

   - To have the **PATH** update take effect immediately after the installation completes, type the following command in the command line:

     ```
     source ~/.bashrc
     ```
9. Once the installation is complete, verify the installation by typing `python` in the command line. If everything worked correctly, you should see something like this:

![](image)

**How it works...**

Anaconda holds many advantages over downloading Python from [http://www.python.org](http://www.python.org) or using the Python distribution included with your computer, some of which are as follows:

- Almost 90 percent of what you'll use on a day-to-day basis is already included. In fact, it contains over 330 of the most popular Python packages.
- Using Anaconda on both the computer you use for development and the server where your solutions will be deployed helps ensure that you are using the same version of the Python packages that your applications require.
- It's constantly updated; so, you will always be using the latest version of Python and the Python packages.
- It works on all the major operating systems—Linux, Mac, and Windows.
- It comes with tools to connect and integrate with Microsoft Excel.

At the time of writing this, the current version of Anaconda for Python 3 is 2.3.0.
Learn about the Python libraries we will be using

Seven Python libraries make up our Python business intelligence toolkit:

- **Pandas**: A set of high-performance, easy-to-use data structures and data analysis tools. Pandas are the backbone of all our business intelligence tasks.
- **Scikit-learn**: Gives us simple and efficient tools for data mining and data analysis including classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. This will be the workhorse library for our analysis.
- **Numpy**: An efficient multi-dimensional container of generic data that allows for arbitrary datatypes to be defined. We won't use numpy directly; however, Pandas relies on it.
- **Matplotlib**: A 2D plotting library. We'll use this to generate all our charts.
- **PyMongo**: Allows us to connect to and use MongoDB. We'll use this to insert and retrieve data from MongoDB.
- **XlsxWriter**: This allows us to access and create Microsoft Excel files. This library will be used to generate reports in the Excel format.
- **IPython Notebook (Jupyter)**: An interactive computational environment. We'll use this to write our code so that we can get feedback faster than running a script over and over again.

Installing, configuring, and running MongoDB

In this section, you'll see how to install, configure, and run MongoDB on all the major operating systems—Mac OS X, Windows, and Linux.

**Getting ready**

Open a web browser and visit: https://www.mongodb.org/downloads.

**How to do it...**

**Mac OS X**

The following steps explain how to install, configure, and run MongoDB on Mac OS X:

1. On the download page, click on the Mac OS X tab, and select the version you want.
2. Click on the **Download (TGZ)** button to download MongoDB.

3. Unpack the downloaded file and copy to any directory that you like. I typically create an **Applications** folder in my home directory where I install apps like this.

4. For our purpose, we’re going to set up a single instance of MongoDB. This means there is literally nothing to configure. To run MongoDB, open a command prompt and do the following:
   - At the root of your computer, make a data directory:
     ```bash
     sudo mkdir data
     ```
   - Make your user the owner of the directory using the `chown` command:
     ```bash
     chown your_user_name:proper_group data
     ```
   - Go to the directory where you have MongoDB.
   - Go to the MongoDB directory.
   - Type the following command:
     ```bash
     .mongod
     ```

5. You should see the following output from Mongo:

```
1. mongod
```

![MongoDB Output](image_url)

---

---
Getting Set Up to Gain Business Intelligence

Windows

The following steps explain how to install, configure, and run MongoDB on Windows:

1. Click on the Windows tab, and select the version you want.
2. Click on the Download (MSI) button to download MongoDB.
3. Once downloaded, browse to the folder where Mongo was downloaded, and double-click on the installer file.
   When asked which setup type you want, select Complete.
4. Follow the instructions to complete the installation.
5. Create a data folder at C:\data\db. MongoDB needs this directory in order to run. This is where, by default, Mongo is going to store all its database files.
6. Next, at the command prompt, navigate to the directory where Mongo was installed and run Mongo:
   
   ```
   cd C:\Program Files\MongoDB\Server\3.0\bin
   mongod.exe
   ```

   7. If you get any security warnings, give Mongo full access.
8. You should see an output like the following screenshot from Mongo, letting you know it's working:

   ![Command Prompt - mongod.exe](image)
Linux

The easiest way to install MongoDB in Linux is by using apt. At the time of writing, there are apt packages for 64-bit long-term support Ubuntu releases, specifically 12.04 LTS and 14.04 LTS. Since the URL for the public key can change, please visit the MongoDB Installation Tutorial to ensure that you have the most recent one: https://docs.mongodb.org/manual/tutorial/install-mongodb-on-ubuntu/.

Install Mongo as follows:

1. Log in to your Linux box
2. Import the public key:
   ```bash
   sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 7F0CEB10
   ```
3. Create a list file for MongoDB:
   ```bash
   echo "deb http://repo.mongodb.org/apt/ubuntu "$(lsb_release -sc)"/mongodb-org/3.0 multiverse" | sudo tee /etc/apt/sources.list.d/mongodb-org-3.0.list
   ```
4. Update apt:
   ```bash
   sudo apt-get update
   ```
5. Install the latest version of Mongo:
   ```bash
   sudo apt-get install -y mongodb-org
   ```
6. Run Mongo with the following command:
   ```bash
   sudo service mongod start
   ```
7. Verify that MongoDB is running by checking the contents of the log file at /var/log/mongodb/mongod.log for a line that looks like this: [initandlisten] waiting for connections on port 27017
8. You can stop MongoDB by using the following `mongod` command:
   ```bash
   sudo service mongod stop
   ```
9. Restart MongoDB with this command:
   ```bash
   sudo service mongod restart
   ```

MongoDB log file location

MongoDB stores its data files in /var/lib/mongodb and its log files in /var/log/mongodb.
How it works...

MongoDB's document data model makes it easy for you to store data of any structure and to dynamically modify the schema. In layman's terms, MongoDB provides a vast amount of flexibility when it comes to storing your data. This comes in very handy when we import our data. Unlike with an SQL database, we won't have to create a table, set up a scheme, or create indexes—all of that will happen automatically when we import the data.

Installing Rodeo

IPython Notebook, an interactive, browser-based tool for developing in Python, has become the de facto standard for creating and sharing code. We'll be using it throughout this book. The Python library that we're about to install—Rodeo—is an alternative you can use. The difference between IPython Notebook and Rodeo is that Rodeo has a built-in functionality to view data in a Pandas data frame, a functionality that can come in handy when you want to view, real-time, the changes that you are making to your data. Having said that, IPython Notebook is the current standard.

Getting ready

To use this recipe, you need a working installation of Python.

How to do it...

Regardless of the operating system, you install Rodeo with the following command:

```
pip install rodeo
```

That's all there is to it!

How it works...

The pitch for Rodeo is that it's a data centric IDE for Python. I use it as an alternative to IPython Notebook when I want to be able to view the contents of my Pandas data frames while working with my data. If you've ever used a tool like R Studio, Rodeo will feel very familiar.
Starting Rodeo

Using this recipe, you will get to learn how to start Rodeo.

Getting ready

To use this recipe, you need to have Rodeo installed.

How to do it...

To start an instance of Rodeo, change to the directory where you want to run it, and type the following command in your working directory:

```
rodeo .
```

Once Rodeo is up and running, open a browser and enter the following URL:

```
http://localhost:5000
```

Once there, you should see something like this:
Getting Set Up to Gain Business Intelligence

Installing Robomongo

Robomongo is a GUI tool for managing MongoDB that runs on Mac OS X, Windows, and Linux. It allows you to create new databases and collections and to run queries. It gives you the full power of the MongoDB shell in a GUI application, and has features including multiple shells, multiple results, and autocompletion. And to top it all, it’s free.

Getting ready

Open a web browser, and browse to http://robomongo.org/.

How to do it...

Mac OS X

The following steps explain how to install Robomongo on Mac OS X:

1. Click on the Download for Mac OS X button.
2. Click on the Mac OS X Installer (.dmg) link to download the file.
3. Once downloaded, double-click on the installer file.
4. Drag the Robomongo application to the Applications folder.
5. Open the Applications folder, and double-click on Robomongo to start it up.
6. In the MongoDB Connections window, create a new connection:
Chapter 1

7. Click on Save.
8. Highlight your new connection and click on Connect.
9. Assuming that you have MongoDB running, you should see the default system database.

Windows
The following steps explain how to install Robomongo on Windows:

1. Click on the Download for Windows button.
2. Click on the Windows Installer (.exe) link to download the file.
3. Once downloaded, double-click on the installer file, and follow the install instructions, accepting all the defaults.
4. Finally, run Robomongo.
5. In the MongoDB Connections window, create a new connection:

![MongoDB Connections Window](image)

6. Click on Save.
7. Highlight your new connection, and click on Connect.
8. In the View menu, select Explorer to start browsing the existing MongoDB databases. As this is a brand new instance, you will only have the system collection.
Using Robomongo to query MongoDB

Robomongo allows you to run any query against a MongoDB that would use the MongoDB command-line utility. This is a great way to test the queries that you’ll write and to view the results.

Getting ready

To use this recipe, you need to have a working installation of MongoDB and have Robomongo installed.

How to do it...

You can use Robomongo to run any query against MongoDB that you would run at the command line. Use the following command to retrieve a single record:

```javascript
db.getCollection('accidents').findOne()
```

You can view the results in multiple formats:

- Tree mode
- Table mode
- Text mode

By default, Robomongo will show you the results in tree mode as shown in the following screenshot:
Downloading the UK Road Safety Data dataset

In this section, we’re going to download and take a bird's eye view of the dataset we’ll be using throughout this book—the UK Road Safety Data. In total, this dataset provides more than 15 million rows across three CSV files.

How to do it...

2. Click on the red Download button on the right side of the page. I suggest creating a data directory to hold the data files.
3. Unpack the provided zip files in the directory you created.
4. You should see the following four files included in the expanded directory:
   - Accidents7904.csv
   - Casualty7904.csv
   - Road-Accident-Safety-Data-Guide-1979-2004.xls
   - Vehicles7904.csv

How it works...

The CSV files contain the data that we are going to use in the recipes throughout this book. The Excel file is pure magic, though. It contains a reference for all the data, including a list of the fields in each dataset as well as the coding used.

Coding data is a very important preprocessing step. Most analysis tools that you will use expect to see numbers rather than labels such as city or road type. The reason for this is that computers don't understand context like we humans do. Is Paris a city or a person? It depends. Computers can't make that judgment call. To get around this, we assign numbers to each text value. That's been done with this dataset.
Why we are using this dataset

It is said that up to 90 percent of the time spent on most data projects is for preparing the data for analysis. Anecdotal evidence from this author and those I speak with holds this to be true. While you will learn a number of techniques for cleaning and standardizing data, also known as preprocessing in the data world, the UK Road Safety Data dataset is an analysis-ready dataset. In addition, it provides a large amount of data—millions of rows—for us to work with.

This dataset contains detailed road safety data about the circumstances of personal injury road accidents in GB from 1979, the types (including Make and Model) of vehicles involved and the consequential casualties.
Where to buy this book

You can buy Python Business Intelligence Cookbook from the Packt Publishing website. Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers. Click here for ordering and shipping details.