Geospatial Data Science Quick Start Guide

Effective techniques for performing smarter geospatial analysis using location intelligence

Abdishakur Hassan
and Jayakrishnan Vijayaraghavan

Packt
www.packt.com

Geospatial Data Science Quick Start Guide

Effective techniques for performing smarter geospatial analysis using location intelligence

Abdishakur Hassan
Jayakrishnan Vijayaraghavan

Packt

BIRMINGHAM - MUMBAI

Geospatial Data Science Quick Start Guide

Commissioning Editor: Amey Varangaonkar
Acquisition Editor: Devika Battike
Content Development Editor: Roshan Kumar
Technical Editor: Snehal Dalmet
Copy Editor: Safis Editing
Project Coordinator: Namrata Swetta
Proofreader: Safis Editing
Indexer: Manju Arasan
Graphics: Alishon Mendonsa
Production Coordinator: Alishon Mendonsa

First published: May 2019

Production reference:1300519

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-78980-941-1

www.packtpub.com

To my wife, Sirad, for being my partner in life. To my children, Mohamed, Manal, and Almas, for always being a source of inspiration. To my mom, Mako, for raising us in difficult times, and finally, to my father, Awil, for being a guiding mentor and source of support.

– Abdishakur Hassan

To the three most beautiful women in my life: my mom, Vijayakumari, whose kindness is eternal and unfathomable; my wife, Athulya, who is amazing in every way imaginable; and my sister, Sujatha, who is always, always, an inspiration.

– JK

mapt.io

Mapt is an online digital library that gives you full access to over 5,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Improve your learning with Skill Plans built especially for you
- Get a free eBook or video every month
- Mapt is fully searchable
- Copy and paste, print, and bookmark content

Packt.com

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.packt.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at customercare@packtpub.com for more details.

At www.packt.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on Packt books and eBooks.

Contributors

About the authors

Abdishakur Hassan is a geographic information systems (GIS) expert and consultant with over 5 years of experience working with UN-Habitat. He holds an MSc in geoinformation science and earth observations. During his tenure as a GIS expert, Abdishakur has developed fully fledged GIS applications in the urban planning and land administration domains. He is interested in all things related to geospatial data science.

I would like to thank my wife, Sirad, who raises our three beautiful children, always prioritizing our family over her professional career, and succeeding in managing and balancing family life with her professional career. My children, Mohamed, Manal, and Almas, have also been an inspiration during writing of this book. I would also like to thank Roshan and Snehal and the editing team at Packt for always being constructive during this period.

Jayakrishnan Vijayaraghavan is a geospatial data scientist, innovator, and author of a book titled *ArcGIS for JavaScript developers*. He currently resides in the San Francisco Bay Area and has over 8 years of work experience. He has built patented technologies and products in the geospatial domain and has coached teams on leveraging mapping and spatial analytics tools for solving pertinent business problems. He is adept at computational geometry, especially in graph networks and in geospatial inferencing. He is a DAAD scholar and a winner of the UN-Habitat special jury award. He is keen on developing intelligent and ubiquitous mapping systems by integrating ML and DL techniques with GIS. He is also a novelist and a certified UAV pilot.

I would like to thank the Packt publishing team; the editors; my coauthor, Abdishakur; and the reviewers for helping me to bring this book, which I was so passionate about, to fruition. Also, a special shout-out to Roshan and Snehal, whose tireless support is highly appreciated.

About the reviewers

Jyoti Rohodia completed her bachelor of technology degree in geoinformatics at the University of Petroleum and Energy Studies. Her bachelor's curriculum included specializations in applied petroleum geology and geophysics and remote sensing. Currently, she is working as a senior GIS analyst at CyberTech Systems and Software Ltd, and provides GIS solutions to USA-based clients of ESRI Inc. She has extensive GIS proficiency working with the following ESRI products: ArcGIS Desktop, ArcGIS Pro, ArcGIS Online, and ArcGIS Enterprise. Additionally, she has sound knowledge of automating GIS workflows using the ArcPy module and ArcGIS API for Python.

Shital Dhakal is a seasoned GIS professional with over seven years' experience in the field of GIS and remote sensing. He has acquired industry and research experience in North America, Europe, and Asia. Currently, he works at a San Francisco Bay Area-based start-up and helps local government to implement enterprise GIS strategies. He is a certified GIS Professional (GISP) and has an MSc from Boise State University, Idaho. When he is not playing with spatial data, writing blogs, or making maps, he can be found hiking in the Sierra Nevada and, occasionally, in the Himalayas.

Packt is searching for authors like you

If you're interested in becoming an author for Packt, please visit `authors.packtpub.com` and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Table of Contents

Preface

Geospatial Data Science Quick Start Guide is all about providing a segue for data scientists into mapping technologies, and for GIS scientific researchers into data science and machine learning techniques. Geospatial data science is the core technology used at companies including Uber, Google Maps, Apple Maps, Foursquare, and by real estate search apps such as realtor.com. An integral reason for the enormous success of these companies can be traced to their early successes in effectively identifying and leveraging the power of location data in the context of data science. Location data is a pervasive component of data repositories of any company in Silicon Valley, and the actual benefits of this data are yet to tapped due to a lack of expertise in this area. *Geospatial Data Science Quick Start Guide* will try to bridge this gap by equipping the reader with important skills for handling location data and effectively synergizing location data with machine learning techniques.

Who this book is for

Data scientists who would like to leverage location-based data and want to use location-based intelligence in the models will benefit from this book considerably. This book is also for GIS Pythonistas who would like to sneak into the magical world of data science and make their analyses more powerful than ever before.

What this book covers

`Chapter 1`, *Introducing to Location Intelligence*, introduces the reader to location data and location data intelligence. It provides real-world examples of location data intelligence. This serves as a basic introduction, and has one section of code working on real-world examples of location data intelligence.

`Chapter 2`, *Consuming Location Data Like a Data Scientist*, covers machine learning models for predicting the trip time of taxi journeys based on location factors and other attributes. In this chapter, we will discuss how to leverage spatial data masquerading as tabular data, and apply machine learning techniques to it as any data scientist would.

`Chapter 3`, *Performing Spatial Operations Like a Pro*, lays the groundwork for dealing with geospatial data. In this chapter, we cover the basics of GeoDataFrame, coordinate systems and projections, as well as spatial operations such as buffer analysis and spatial joins. We look into foundational, as well as advanced, location data intelligence techniques using the Foursquare dataset.

`Chapter 4`, *Making Sense of Humongous Location Datasets*, explores ways to aggregate location data into meaningful chunks using machine learning clustering techniques, and deriving more value from it. This chapter further introduces the reader to topics such as spatial autocorrelation, and both global and local spatial autocorrelation are explored and discussed.

`Chapter 5`, *Nudging Check-Ins with Geofences*, introduces to the reader to geofencing, which is a popular tool that sees use in contexts from businesses to conservation work efforts. Geofencing refers to abstract fences that are created around a location, such that an alert or notification system can notify the relevant party should an event happens at or within the fence. The event can be something as simple as a customer entering the vicinity of a business location, or customers moving within range of a cell-phone tower – the applications are unlimited. This chapter moves onto methods for quickly building and deploying a geofencing system using Python.

`Chapter 6`, *Let's Build a Routing Engine*, teaches the reader about navigation and routing, which are indispensable features in today's apps. Be it navigational apps, food delivery apps, or a courier delivery app; routing is a key component to the effective delivery of their services. And most of these apps rely on a few key third-party APIs, such as Google Maps APIs, to provide such services. Do we really need to pay these vendors to build a simple routing app? This topic explains how to build our own routing and navigation solutions using open source data and Python libraries that implement graph algorithms.

`Chapter 7`, *Getting Location Recommender Systems*, is the final chapter of the book and discusses the process of building recommender systems that do not only apply to products on Amazon or movies on Netflix, but also locations. This chapter leverages popular recommender system techniques on offer, including collaborative filtering methods and location-based recommenders. We will use these recommender systems to build a restaurant venue recommendation system.

To get the most out of this book

This following is a short list of requirements to successfully explore this book:

- **A computer with a browser**: This book primarily uses Google Colaboratory Jupyter Notebooks. To be able to use the code effectively, you need a browser and internet access.
- **Basic Python programming skills**: The reader should have at least introductory knowledge of data types and functions in the Python language. GIS skills will be helpful, but this is not required.

Download the example code files

You can download the example code files for this book from your account at `www.packt.com`. If you purchased this book elsewhere, you can visit `www.packt.com/support` and register to have the files emailed directly to you.

You can download the code files by following these steps:

1. Log in or register at `www.packt.com`.
2. Select the **SUPPORT** tab.
3. Click on **Code Downloads & Errata**.
4. Enter the name of the book in the **Search** box and follow the onscreen instructions.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR/7-Zip for Windows
- Zipeg/iZip/UnRarX for Mac
- 7-Zip/PeaZip for Linux

The code bundle for the book is also hosted on GitHub at `https://github.com/PacktPublishing/Geospatial-Data-Science-Quick-Start-Guide`. In case there's an update to the code, it will be updated on the existing GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: `http://www.packtpub.com/sites/default/files/downloads/9781789809411_ColorImages.pdf`.

Conventions used

There are a number of text conventions used throughout this book.

`CodeInText`: Indicates code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles. Here is an example: "Mount the downloaded `WebStorm-10*.dmg` disk image file as another disk in your system."

A block of code is set as follows:

```
from shapely.geometry import Point
for lat, lon in zip(nyc['Latitude'][:5], nyc['Longtitude'][:5]):
   geometry = Point(lat, lon)
    print(geometry)
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
ax.set_xticklabels([])
ax.set_yticklabels([])
ax.set_title('Foursquare Points')
plt.show()
```

Any command-line input or output is written as follows:

```
!pip install networkx
```

Bold: Indicates a new term, an important word, or words that you see onscreen. For example, words in menus or dialog boxes appear in the text like this. Here is an example: "Select **System info** from the **Administration** panel."

Warnings or important notes appear like this.

Tips and tricks appear like this.

Get in touch

Feedback from our readers is always welcome.

General feedback: If you have questions about any aspect of this book, mention the book title in the subject of your message and email us at `customercare@packtpub.com`.

Errata: Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you have found a mistake in this book, we would be grateful if you would report this to us. Please visit `www.packt.com/submit-errata`, selecting your book, clicking on the Errata Submission Form link, and entering the details.

Piracy: If you come across any illegal copies of our works in any form on the Internet, we would be grateful if you would provide us with the location address or website name. Please contact us at `copyright@packt.com` with a link to the material.

If you are interested in becoming an author: If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, please visit `authors.packtpub.com`.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions, we at Packt can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about Packt, please visit `packt.com`.

1
Introducing Location Intelligence

"Everything that happens, happens somewhere."
- The first law of geography by Waldo Tobler

Location data is data with a geographic dimension. Location data is everywhere as all actions that occur in or near the Earth's surface happen to use geographic aspects. It is generally referred to as any data with coordinates (latitude, longitude, and sometimes altitude) but also encompasses different aggregated geographic units, including addresses, zip codes, landmarks, districts, cities, regions, and much more.

Location intelligence, on the other hand, is the process of turning geographic (spatial) data into insights and business outcomes. Any data with a geographical position, either implicitly or explicitly, requires location-aware preprocessing methods, visualization, as well as analytical methods to derive insights from it. Thus, location intelligence applications can reveal hidden patterns of spatial relationships that cannot be derived through other normal means. It leads to better decision making on spatial problems, where things happen, why they happen in some places, and the spatial trends in time-series analysis. Understanding the location dimension of today's challenges in, industrial, retail, agricultural, climate, and environment, can lead to a better understanding of why economic, social, and environmental activities tend to locate where they are.

In this chapter, we give an overview of location data and location data intelligence. Here, we briefly introduce different location data types and location data intelligence applications and examples. We cover how to identify location data from publicly available open datasets. We briefly discuss and highlight the difference between location data and other non-geographic data. At the end of this chapter, we explore how location data fits into data science and what opportunities and challenges bring location data into the interdisciplinarity of data science.

We will specifically focus on the following topics:

- Location data
- Location data intelligence
- Location data and data science
- A primer on Google Colab and Jupyter Notebooks

Location data

What is location data and why is it different than other data formats? It is quite common to see phrases such as *spatial data is special* or another more popular adage, *80% of data is geographic*. While these are not easily provable, we tend to witness an increased amount of location data. From geotagged images, text, and sensor data, location data is ubiquitous and the world is *datafied*. In this connected and data-driven driven era, we generate, keep track of, and store huge mounts of data every day. Think of the number of tweets, Instagram images, bank transactions, searches on the web, and routing requests from APIs. We collect more data than at any other period of time in the past, and thus the big data revolution. Many of the datasets collected have an inherent location dimension but are often hidden within the data and not utilized fully.

Understanding location data from various perspectives

We can examine location data from different perspectives: business, technical, and data perspectives.

From a business perspective

From a business perspective, the value of maps and location data is crucial in many business applications. A quick look at big companies such as Google, Apple, Microsoft, and Nokia shows that each of these companies has their own location and mapping services and products.

Think about how often you use Google Maps API's location service through your phone. This also highlights the importance of location data as all these companies would not go to such lengths to have their own in-house location data production if it was not necessary. Business applications in location data include not only individual uses of location data but also innovative applications spanning from individualized marketing, autonomous vehicles, logistics, and transportation to healthcare.

From a technical perspective

The technical perspective of location data indicates that it entails both opportunities as well as challenges. Location data, in contrast to other data, has a topology, which holds the relationships between geometry (points, lines, and polygons) and geographic features that they represent. In the case of conventional data, we store data into tables or a **Relational Database Management System** (**RDBMS**). However, spatial relations and topology require us to store the geometry of objects.

Due to the nature of location data, which is derived from **Tobler's first law** of geography, *Everything is related to everything else, but near things are more related than distant things*. The essence of this law entails also the presence of strong autocorrelation and interdependency in continuous near locations, which is not necessarily present in conventional data (non-spatial attributes).

From a data perspective

Having looked into the nature of location data from a technical perspective, let's also examine it from a data perspective. How is location data different than other data? In location data, we use geographic coordinates (2D) to represent the world (3D).

For example, **Digital Elevation Models** (**DEMs**) are used to represent heights and terrain surface. The first law of geography applies here as well. At a certain point of time, a particular terrain is very likely to have the same height with its relatively close surrounding, while we can expect a difference based on elevation in two areas distant from each other. As mentioned earlier, spatial autocorrelation in location data is assumed to be present in spatial data, while in other types of data, such as the statistical analysis of conventional data, we assume the independence of data points. That means location data can be categorized as **stochastic**, while other data is **probabilistic**.

Another complication in location data also arises from what we call **Modifiable Area Unit Problem** (**MAUP**), which arises from different aggregated units that produce different results. An example of this is poverty or crime estimates and aggregations. For example, areas of high poverty rates could be overestimated or underestimated depending on the boundaries of measured areas. By moving into different aggregations (that is, zip code, neighborhood, or district level), which can create different impressions and patterns created by the different scales and aggregations.

Types of location data

Geographic data types can be divided into two broad categories:

- **Vector data**: This is represented as points, lines, or polygons. The data is likely created by digitizing it and storing information in longitude and latitude. This type of data is useful for storing data that has discrete and distinct boundaries such as borders, land parcels, streets, and points of interest.
- **Raster data**: This stores information in cells and therefore is suitable for storing data that is continuous, such as satellite images, elevation models, and other aerial photographs.

Location data intelligence

Every industry uses **location intelligence**. It helps industries understand what their customers are doing, where their customers are based, what the geographic environment of their customers is, and what their interests are. Location intelligence is normally defined as using location data with other attributes to add context and derive useful information, services, and products that help organizations make effective and efficient decisions. The information derived through location intelligence can have a business and economic insights as well as environmental and social insights.

Application of location data intelligence

To illustrate how location intelligence is applied in a real-world application, we will take as an example **Foursquare** check-ins. Foursquare initially started in 2009 as a social platform to collect user check-ins and provide guides and search-results for its users to recommend places to visit near the user's current location. However, recently, Foursquare repositioned itself as a less social platform to a location intelligence company. The company describes itself as a "*technology company that uses location intelligence to build meaningful consumer experiences and business solutions*" and claims the following:

> *"If it tells you where, it's probably built on Foursquare."*

In its anonymized and aggregated trends of check-ins in physical brand locations, Foursquare provides insights and metrics that were not easily available before. Take, for example, the loyalty of customers, frequency of their visits, brand losses, and profits. This allows analysts and brands to understand their customers, reveal demographic insights and track patterns of customers, and look into and understand competition brands. To illustrate how powerful location intelligence is, let's explore a subset of Foursquare data in NYC. We will use this dataset later in `Chapter 3`, *Performing Spatial Operations Like a Pro*, but for now let's look into what it consists and how location intelligence is derived from it.

The NYC Foursquare check-in dataset has 10 months' worth of data spanning from April 12, 2012 to February 16, 2013.

Source: NYC Foursquare Check-in dataset first appeared in *Fine-Grained Preference-Aware Location Search Leveraging Crowdsourced Digital Footprints from LBSN, Dingqi Yang*, *Daqing Zhang*, *Zhiyong Yu*, and *Zhiwen Yu*, *p*roceedings of the 2013 *ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp* 2013), September 8 to 12, 2013, in Zurich, Switzerland.

The following table shows the first five rows of the data and consists of eight columns with a unique `UserID` and `VenueID`. Both of these features are anonymized for privacy issues; `VenueCategoryID` and `VenueCategoryName` indicate aggregated types of business. Here, we have more than 250 business types, including a medical center, arts store, burger joint, hardware store, and so on; `Latitude` and `Longitude` columns store the geographic coordinates of the venues.

The last two columns indicate the time of the check-in:

	UserID	VenueID	VenueCategoryID	VenueCategoryName	Latitude	Longtitude	Timezone	UTCtime
0	470	49bbd6c0f964a520f4531fe3	4bf58dd8d48988d127951735	Arts & Crafts Store	40.719810	-74.002581	-240	Tue Apr 03 18:00:09 +0000 2012
1	979	4a43c0aef964a520c6a61fe3	4bf58dd8d48988d1df941735	Bridge	40.606800	-74.044170	-240	Tue Apr 03 18:00:25 +0000 2012
2	69	4c5cc7b485a1e21e00d35711	4bf58dd8d48988d103941735	Home (private)	40.716162	-73.883070	-240	Tue Apr 03 18:02:24 +0000 2012
3	395	4bc7086715a7ef3bef9878da	4bf58dd8d48988d104941735	Medical Center	40.745164	-73.982519	-240	Tue Apr 03 18:02:41 +0000 2012
4	87	4cf2c5321d18a143951b5cec	4bf58dd8d48988d1cb941735	Food Truck	40.740104	-73.989658	-240	Tue Apr 03 18:03:00 +0000 2012

Foursquare data: first five rows

Here, we have the first five rows of the Foursquare data. In this chapter, we will only look at the data from a wider perspective. The code for this chapter is available, but you do not need to understand it right now. We will come to learn the details of reading and processing location data with Python in the next chapters.

So, what kind of location intelligence can be derived from this type of data? We will cover this from two broad perspectives: the user/customer perspective and the venue/business perspective.

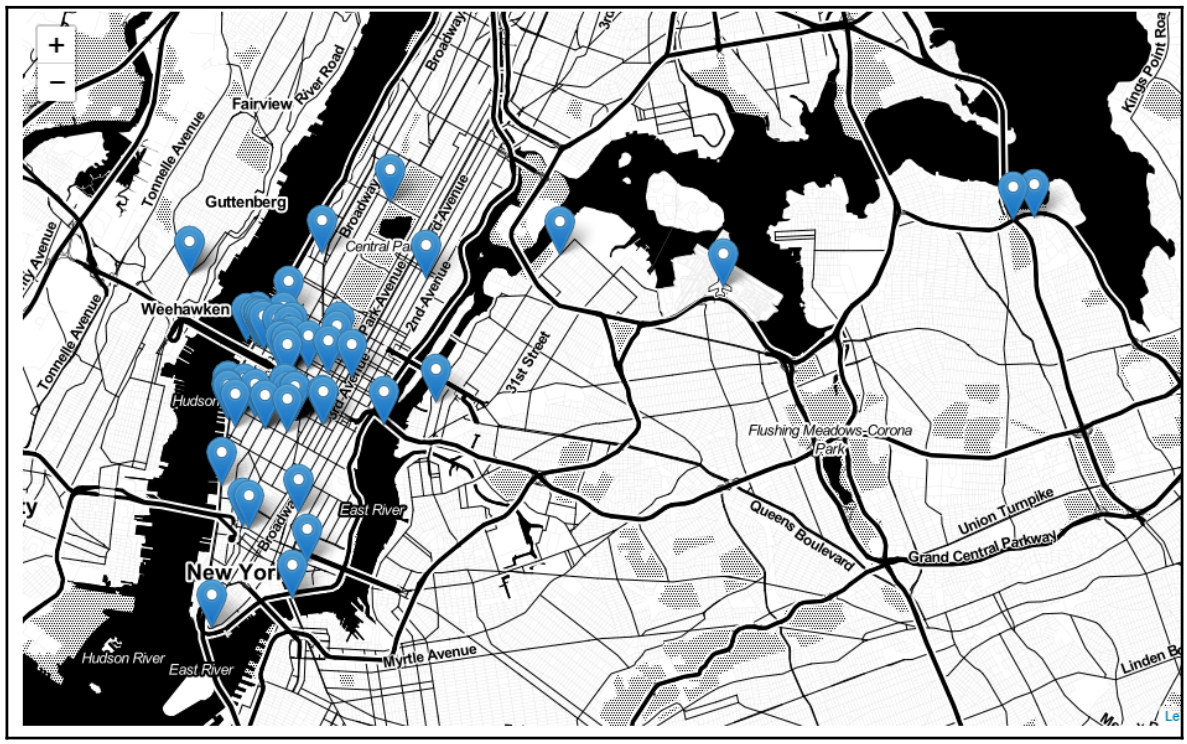
User or customer perspective

Here we will get a clear idea from a customer perspective. Often the following questions will come into picture:

Where does customer X spend his/her time? What does this place offer? How often does he/she visit these places? When does he/she visit these places?

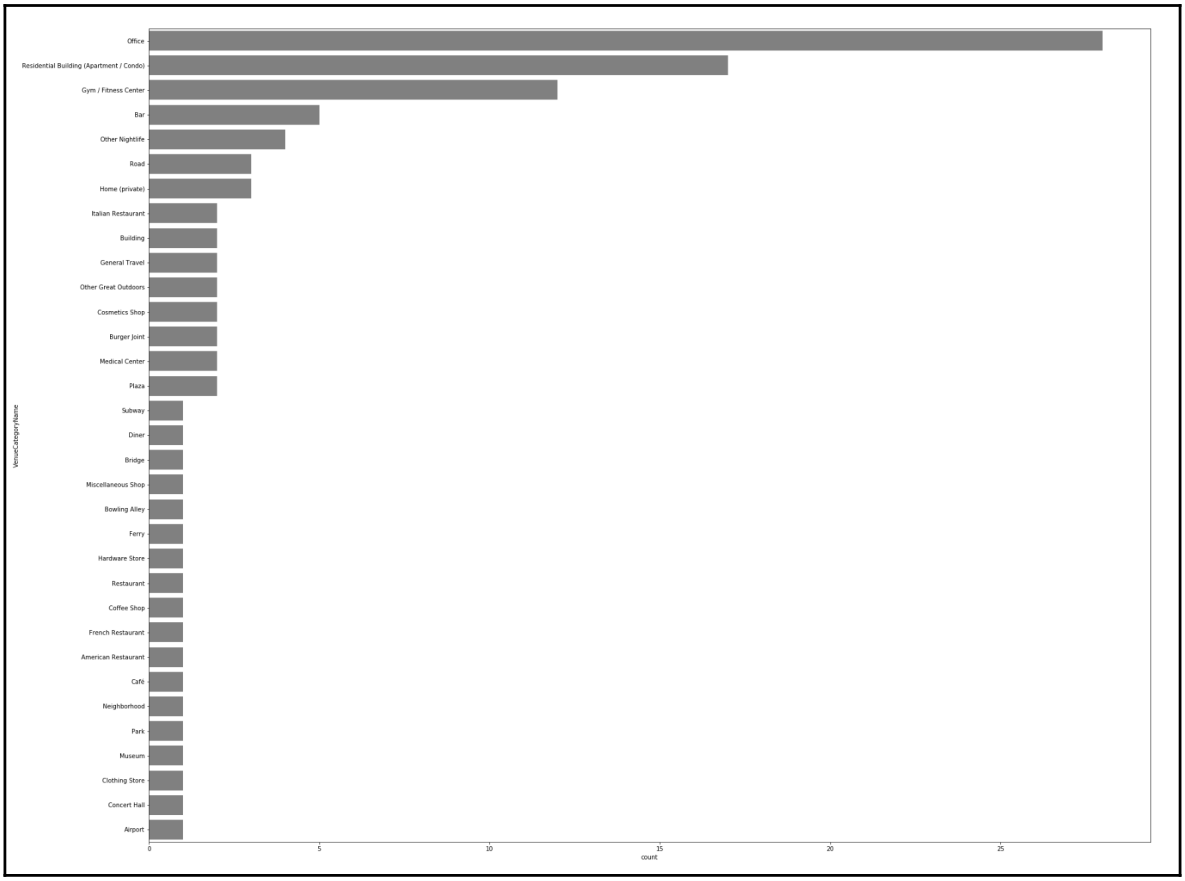
The code of this section is available in the accompanying Jupyter Notebook. You do not need to understand all the code right now, as it serves to give you the bigger picture of location data intelligence. Feel free to consult Jupyter Notebook of this chapter if you want to run the code and experiment.

Let's take an example for the `UserID` = `395` from the fourth row in the preceding table. This particular user has made `106` check-ins in total during this period of the dataset visiting `36` unique venues in NY (visualized as the map as follows):



User 395: Venues visited in NY

We can also look at what type of venues this particular user has visited. In this case, this user has visited frequently an office, a residential building, and a gym, in NYC. Other less-visited venues include an airport, outdoors, a medical center, and many others, as you can see from the following graph:



User 395: Check-ins plot

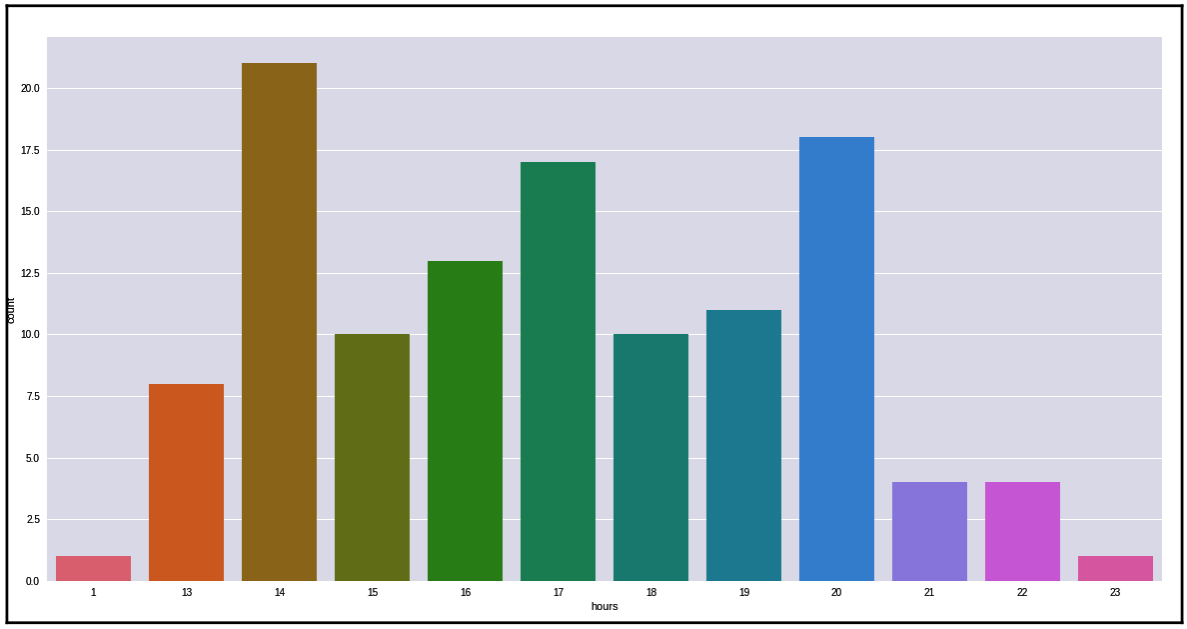
The user perspective can elicit many aspects related to the frequency of visits, preferences, and activities of the user that can guide location intelligence and decision making. Privacy issues in location data are very sensitive and require diligence. In this case, although it is anonymized data, it still reveals patterns and other useful information as we have shown. Now let's also look from the business perspective in the following section.

Venue or business perspective

Here we will get a clear idea from a business perspective. Often the following questions will come into picture:

How many customers does venue X Receive per day? What about per hour? What is the pattern? Can we estimate business value based on the check-ins?

We will use a gym venue as an example here, with `VenueID` = `4aca718ff964a520f6c120e3`. For this dataset, this gym has `118` check-ins. Although the data is small and cannot be generalizable in this particular `VenueID`, imagine it has enough data for a longer period of time. We can estimate the peak times of this gym as the following graph shows. There is a peak of check-ins at 14:00 and at 20:00:



Gym visit check-ins: per hour

This kind of business perspective analysis helps both decision makers and competitors to gain an insight into businesses. This is only an individual business example, but this can simply be extended to businesses in this dataset and look further into it. In fact, Foursquare predicted Chipotle's sales (link available in the information box), a Mexican grill, to drop 30% during the months of 2016 before the company announced its loss.

Foursquare predicted Chipotle's Sales Will Plummet 30%: `http://fortune.com/2016/04/15/chipotle-foursquare-swarm/`.

Let's now look at how location data science is different than data science in the next section.

Location data science versus data science

Now that we have learned that location data is beyond mapping, and specifically is manipulation and processing of geographic data and applying analytical methods, we will move into the interdisciplinarity of location data science. We have also studied location data intelligence and how insights are derived from location data by illustrating this with diagrams. But how is location data science (spatial data science) different than data science? How do they relate to each other? In this section, we will cover the commonalities as well as differences between location data science and data science as a discipline.

Data science

What is **data science**? Data science as a field consists of computer science, mathematics and statistics, and domain expertise and is generally referred to as the process of extracting insights and useful information from data. Mostly, it involves importing data and tidying it to make it ready for analysis. An iterative process of data science also implies transforming, visualizing, and modeling data to understand phenomena and hidden patterns within the data. The final process in data science which is often explored less, is to communicate the insights. Now you may realize from what we have covered so far that this is not far from location intelligence, and that is right. The location dimension is critical in many domains and applications with data science. Next, let's look at what spatial data science.

Location (spatial) data science

Adding location data and the underlying spatial science entails additional challenges and opportunities. It will form a combination of the interdisciplinary field consisting of computer science, mathematics and statistics, domain expertise, and spatial science. This does not only indicate the addition of spatial science but also whole new concepts, theories, and the application of spatial and location analysis, including spatial patterns, location clusters, hot spots, location optimization, and decision-making, as well as spatial autocorrelation and spatial exploratory data analysis. For example, in data science, histograms and scatter plots are used for data distributions analysis, but this won't help with location data analysis, as it requires specific methods, such as spatial autocorrelation and spatial distribution to get location insights.

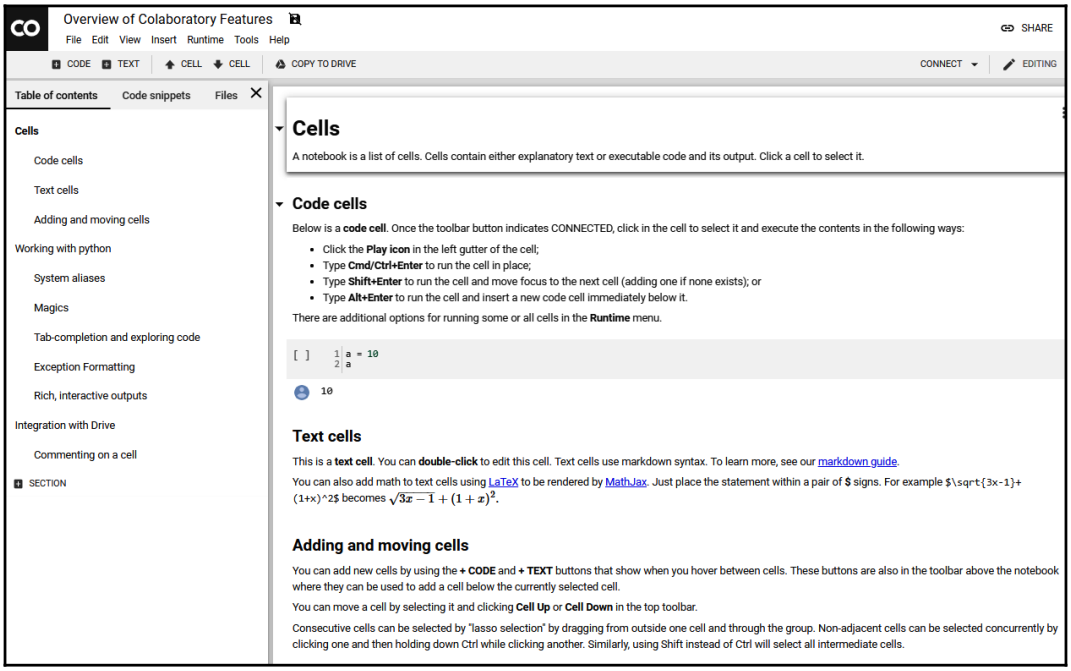
To get the reader up and running quickly and without burdening the local setup of Python environments, we will use Google Colab Jupyter Notebooks in this book. In the next section, we will cover a primer on how to use Google Colab and Jupyter Notebooks.

A primer on Google Colaboratory and Jupyter Notebooks

Jupyter Notebooks have become the favorite tool for data scientists, as they are flexible and combine code, computational output/multimedia, as well as comments. It is free, open source, and provides computational capabilities and interactive web-based notebooks. Anaconda distributions make the installation process easy if you want to install Jupyter Notebook on your local machine. The official Anaconda documents to install Jupyter Notebooks and Python is easy to follow and intuitive, so feel free to follow the instructions if you would like to work on your local machine.

However, we will use **Google Colab**, which is a free Jupyter Notebook environment that requires no installation or setup and runs entirely in the cloud, just like using Google Docs or Google Sheets. Google Colab enables you to write code, run the code, and share it. You just need to have a working Gmail to save and access Google Colab Jupyter Notebooks. In heavy computational tasks, such as machine learning or deep learning with big data, Google Colab allows you to use its **Graphics Processing Unit** (**GPU**) or **Tensor Processing Unit** (**TPU**) for free.

Google Colab interface is shown as follows. In the upper part, you have the main menu. The right part is where we can write our code and comments:



Google Colab

You can open Google Colab from this URL: `https://colab.research.google.com`. There are two main types of cells: code and text. With a code cell, you can write your code and execute it, while a text cell allows you to write down your text with a markdown. Here, you can have different text types, including several heading levels as well as a bulleted and a numbered list. To execute a cell, you can either use a *Ctrl* + *Enter* shortcut or press the **Run** button (small triangle) next to the cell.

We will learn Google by using it as our coding platform for this book. If you are new to Jupyter Notebooks or Google Colab, here is a useful guide to get started: `https://colab.research.google.com/notebooks/welcome.ipynb#scrollTo=GJBs_flRovLc`.