

Let's take a look at how data scientists use different datasets.

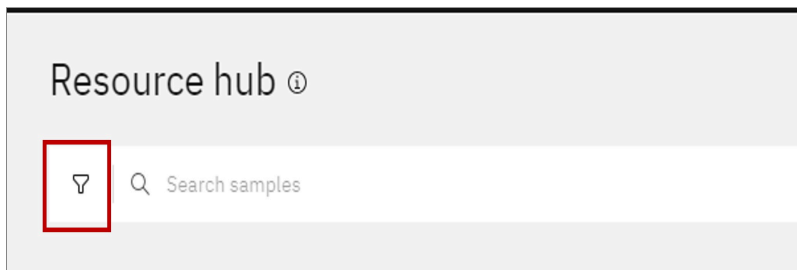
### Objectives :

You will learn to:

- Explore the IBM Cloud Resource hub
- Examine a numeric dataset
- Examine a dataset with non-numeric attributes
- Examine a Jupyter Notebook

### Exercise 1: Examine a numeric dataset

1. Click on the link: <https://dataplatfrom.cloud.ibm.com/gallery>.
2. Click the filter button in the top right of the window:



3. In the dropdown menu that appears, select the *Data* checkbox under *Sample type*. Then click on the *Tags* dropdown, and select the *Environment* checkbox.

- ☐ Data fabric
- ☐ Decision Optimization
- ☐ Economy & Business
- ☐ Energy & Utilities
- ☒ Environment
- ☐ Financial Markets
- ☐ Financial Services
- ☐ Geography
- ☐ Guided tutorial
- ☐ Health
- ☐ Industry Accelerator
- ☐ Insurance
- ☐ Knowledge Accelerators
- ☐ Law & Government

4. In the search results, click on *UCI: Forest Fires*.

Search results (35)

<div><div><div>01</div><div>00</div></div><div>Beijing PM 2.5</div><div>A data set on air pollutant level and other weather conditions in Beijing. Cleaned data based on Beijing PM 2.5 from UCI.</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>Scoring for Beijing PM 2.5</div><div>No description available</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>UCI: Forest fires</div><div>This is a difficult regression task, where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using...</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>UCI: Iris</div><div>The data set contains 150 samples, where each sample is a 4x4 matrix of floating-point values. One class is the set of samples that belong to the same species.</div><div>Data by IBM</div></div>
<div><div><div>01</div><div>00</div></div><div>Dry Bulb Temperature, by country, station and...</div><div>Part of World Meteorological Organization Standard Normals.</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>Worldwide Electricity Demand and Production...</div><div>Measured in Millions of Kilowatt-Hours</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>Renewable internal freshwater resources per...</div><div>The World Development Indicators (WDI) is the statistical benchmark that helps measure the progress of development. The WDI provides a...</div><div>Data by IBM</div></div>	<div><div><div>01</div><div>00</div></div><div>Greenhouse gas emissions, worldwide</div><div>The Greenhouse gas emissions, worldwide, contains the most recent information, covering the period from 1990 to 2019.</div><div>Data by IBM</div></div>
<div><div><div>01</div><div>00</div></div><div>Environment Statistics Database - Water</div></div>	<div><div><div>01</div><div>00</div></div><div>Environment Statistics Database - Waste</div></div>	<div><div><div>01</div><div>00</div></div><div>Energy use (kg of oil equivalent per capita) by...</div></div>	<div><div><div>01</div><div>00</div></div><div>Electric power consumption (kWh per capita) by...</div></div>

5. Preview the data using the *Preview* option.

Environment

Description					Preview
X	Y	month	day	FFMC	
7	5	mar	fri	86.2	
7	4	oct	tue	90.6	
7	4	oct	sat	90.6	
8	6	mar	fri	91.7	

6/20/24, 7:14 PM Predicted to forest fires where the aim is to predict the burned area of forest fires, about:blank east region of Portugal, by using meterological and other data.

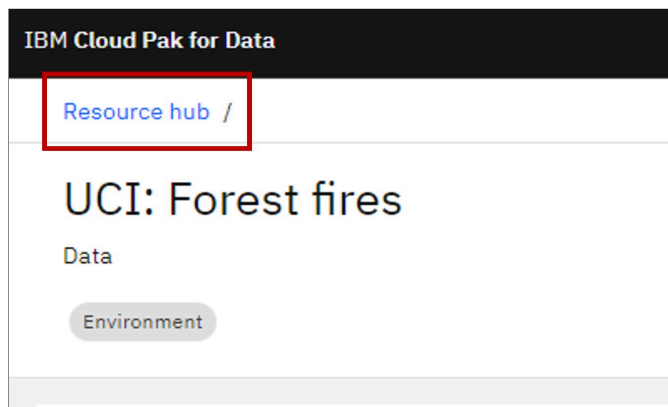
#### Attribute Information:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FPMC - FPMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84  
(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

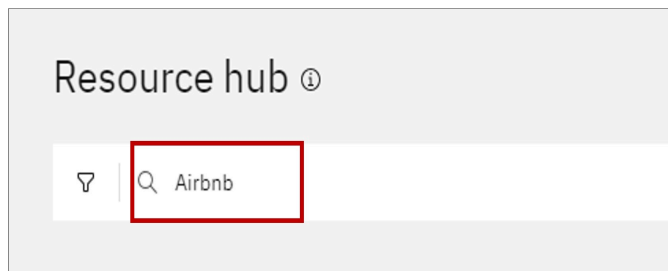
#### Exercise 2: Evaluate a non-numeric dataset

The data doesn't have to be only based on numbers. Data can be text, images and other types as well. Let's look at a dataset which has text values.

1. At the top of the page, select the *Resource hub* option.



2. Type *Airbnb* into the search bar.



3. Select the *Airbnb Data for Analytics: Trentino Reviews* option. You may need to scroll to find it.

<p>6/20/2024 4:17:14 PM</p> <p><b>Airbnb Data for Analytics: Venice Reviews</b></p> <p>Airbnb reviews for Venice, Veneto, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly available data...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Venice Listings</b></p> <p>Airbnb listings for Venice, Veneto, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly available data...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Trentino Listings</b></p> <p>Airbnb listings for Trentino, Trentino-Alto Adige/Südtirol, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Venice Calendar</b></p> <p>Airbnb calendar for Venice, Veneto, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly available data...</p> <p>Data by IBM</p>
<p><b>Airbnb Data for Analytics: Vancouver Listings</b></p> <p>Airbnb listings for Vancouver, British Columbia, Canada. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Vancouver Calendar</b></p> <p>Airbnb calendar for Vancouver, British Columbia, Canada. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Trentino Reviews</b></p> <p>Airbnb reviews for Trentino, Trentino-Alto Adige/Südtirol, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses...</p> <p>Data by IBM</p>	<p><b>Airbnb Data for Analytics: Trentino Calendar</b></p> <p>Airbnb calendar for Trentino, Trentino-Alto Adige/Südtirol, Italy. This dataset is sourced from Inside Airbnb which aggregates and cleanses publicly available data...</p> <p>Data by IBM</p>

4. Preview the data using the *Preview* option.

5064970	29436648	2015-04-07	11582326	Stephan	friendly. We enjoyed her apartment, that was very modern and clean with two rooms, a bathroom and the kitchen inside the living-room with a balcony that goes to the north. All in all a good flat to stay. Thanks!	apartment + Wi-Fi + parking!
5064970	33481368	2015-05-28	20223641	Annika	Marinas flat was a dream! Spotlessly clean, very cute decorated..... and the balcony was the biggest plus! Marina welcomed us in her flat and gave us many tips for hiking, mountainbiking and restaurants. You have to ask her for the best Gelateria in Riva. The best ice cream I 've ever eaten! We will definitely come back! Thank you Marina for the awesome time we could spend in your flat. Annika & Joachim	apartment + Wi-Fi + parking!

### Explore the data

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity. The minimum stay, price and number of reviews have been used to estimate the occupancy rate, the number of nights per year and the income per month for each listing.

You could use this data in multitude of ways - to analyze the star ratings of places, to analyze the location preferences of the customers, to analyze the tone and sentiment of customer reviews and many more. Airbnb uses location data to improve guest satisfaction.

💡 What else might you use this data for?

The dataset comprises of three main tables:

- listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing\_type (categorical), is\_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.
- reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing\_id (discrete), reviewer\_id (discrete) and comment (textual).
- calendar - Provides details about booking for the next year by listing. Four attributes in total including listing\_id (discrete), date(datetime), available (categorical) and price (continuous).

### Exercise 3: Evaluate Jupyter Notebook

Return to the Resource hub. Select *Notebook* from the *Sample type* menu that appears after clicking on the filter button. In the search bar type *Finding optimal locations* Select the card that says *Finding optimal locations of new stores using...*

Filter

×

Sample type

^

- ☐ Data
- ☐ Governance Content
- ☒ Notebook
- ☐ Project

Tags

▼

☰

IBM Cloud Pak for Data

All

▼

🔍

Search

← Back

Tags

Decision Optimization

# Finding optimal locations of new stores using Decision Optimization

This notebook shows you how Decision Optimization can help to prescribe decisions for a complex problem using CPLEX Modeling for Python to help determine the optimal location for a new store requires the Commercial Edition of CPLEX engines, which is included in the latest Python XS + DO Watson Studio.

## Finding Optimal Locations for New Stores

This notebook is an example of how **Decision Optimization** can help to prescribe decisions for a constrained problem.

When you finish this notebook, you'll have a foundational knowledge of *Prescriptive Analytics*.

This notebook requires the Commercial Edition of CPLEX engines, which is included Python 3.7 XS + DO in Watson Studio.

Table of contents:

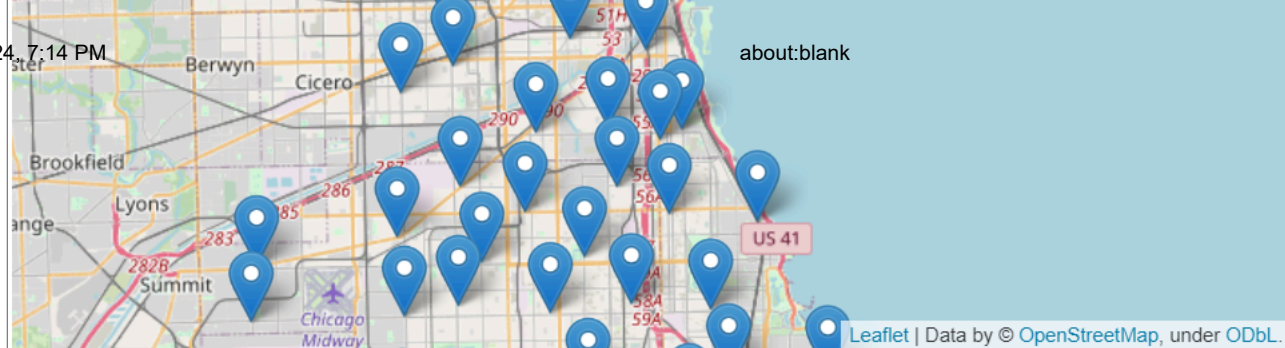
- [Describe the business problem](#)
- [How decision optimization \(prescriptive analytics\) can help](#)
- [Use decision optimization](#)

Part of the Python code in the notebook displays the locations of the libraries on a map.

about:blank

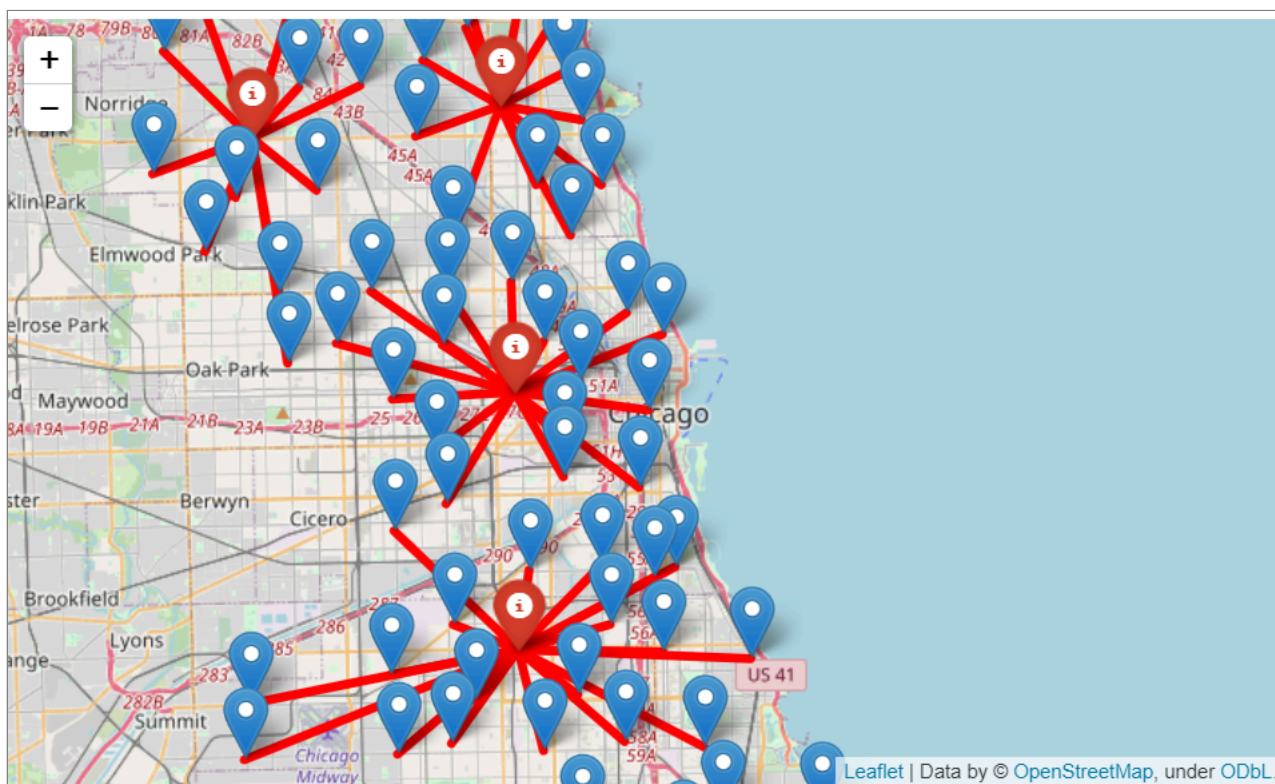
8/10





But with this data, you cannot determine the ideal location of the coffee shops by just looking at the map.

The code then solves this with an optimization model that will help determine possible locations for the coffee shops with the stipulation of minimizing the distance between the libraries and the shop.



## Summary

In this lab, you have learnt about to explore datasets and notebooks in IBM cloud Resource hub.

## Author(s)

[Malika Singla](#)

