Sample 3 contains a Jupyter Notebook, a tool which data scientists use to create models.

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

Objectives :

You will learn to:

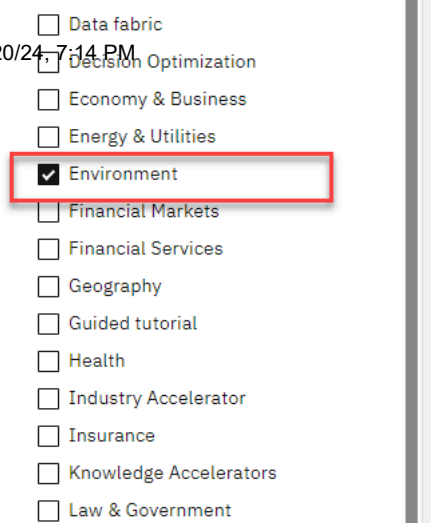
- Explore the IBM Cloud Resource hub
- Examine a numeric dataset
- Examine a dataset with non-numeric attributes
- Examine a Jupyter Notebook

Exercise 1: Examine a numeric dataset

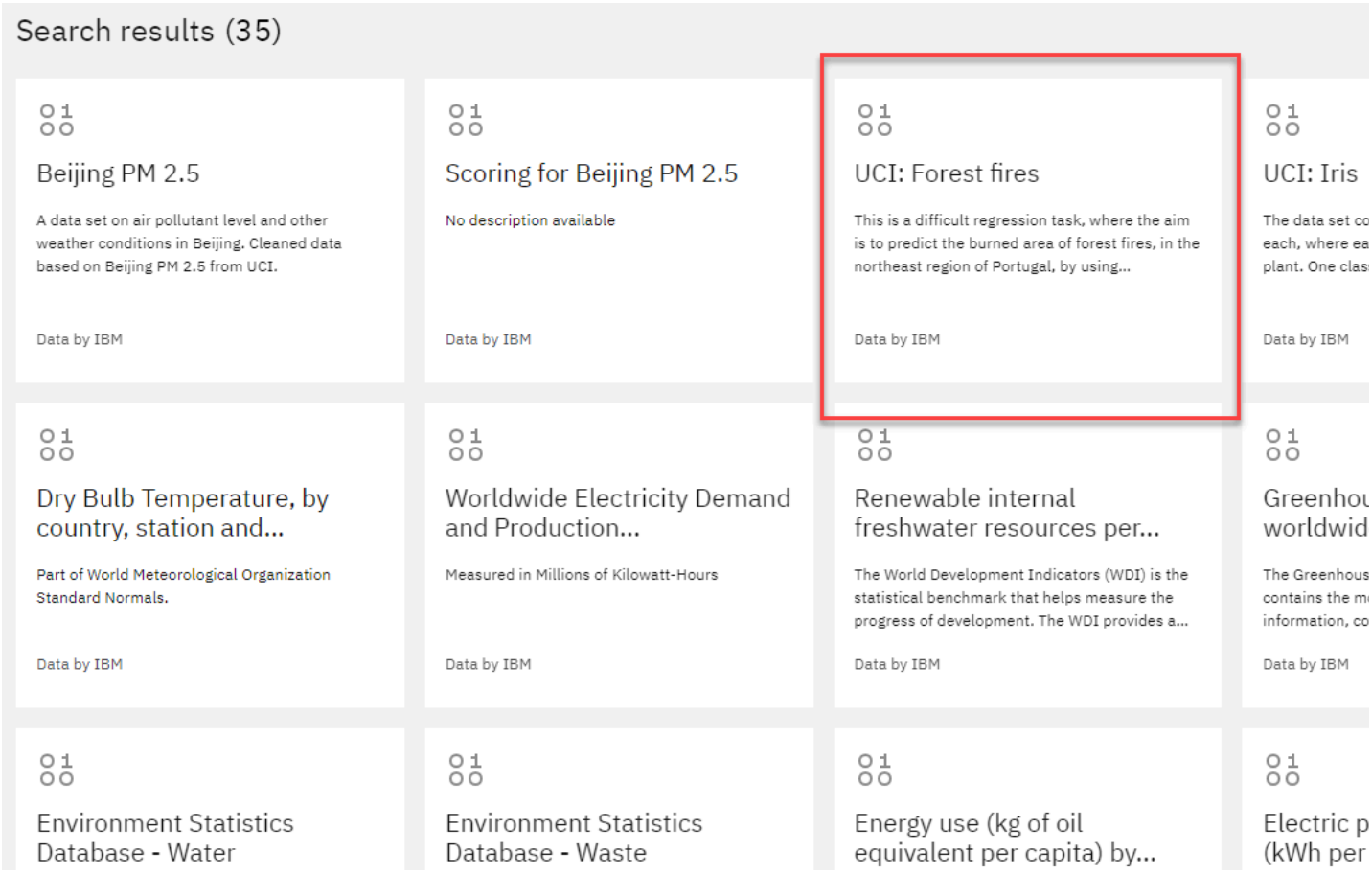
1. Click on the link: https://dataplatform.cloud.ibm.com/gallery

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.



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



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

Data

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

Explore the data

Data is related to forest fires where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using meterological and other data.

Attribute Information:

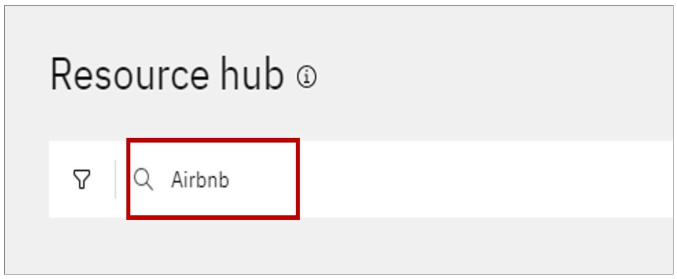
1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index 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

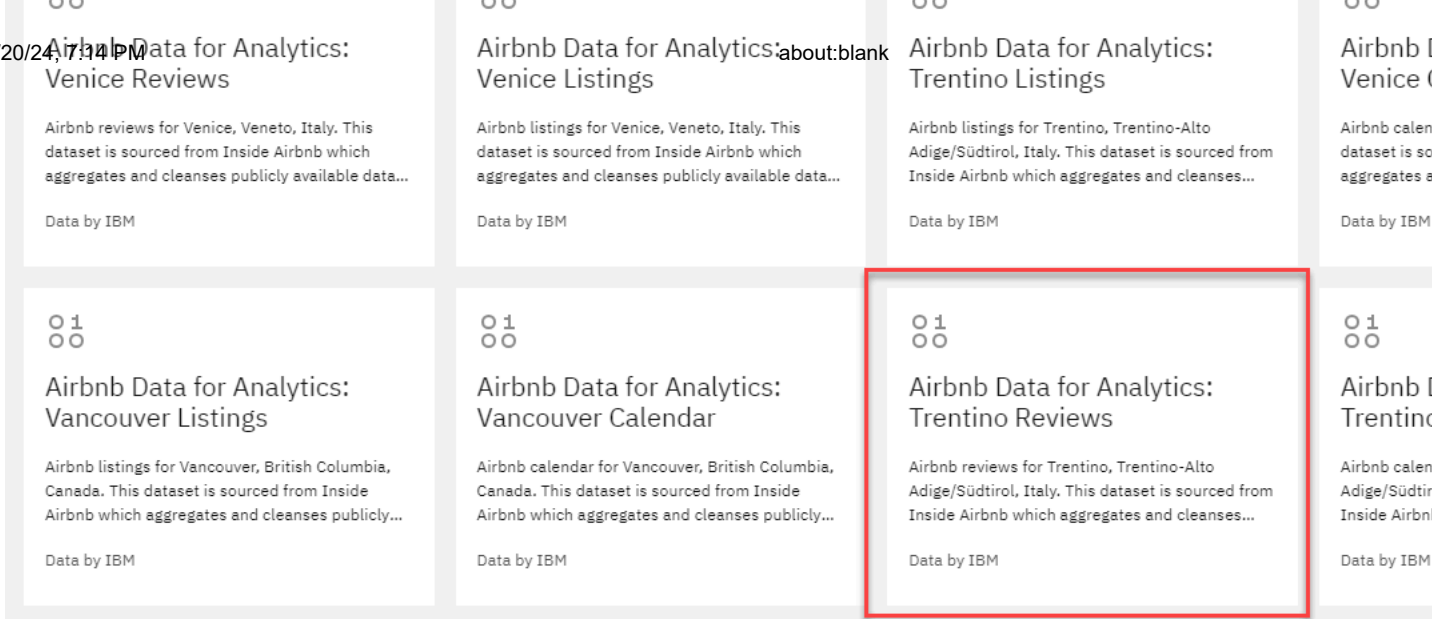
The data doesn't have to be only based on numbers. Data can be text, images and other types as well. Let's look at a dataset which has text values.

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.



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 definitly 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.

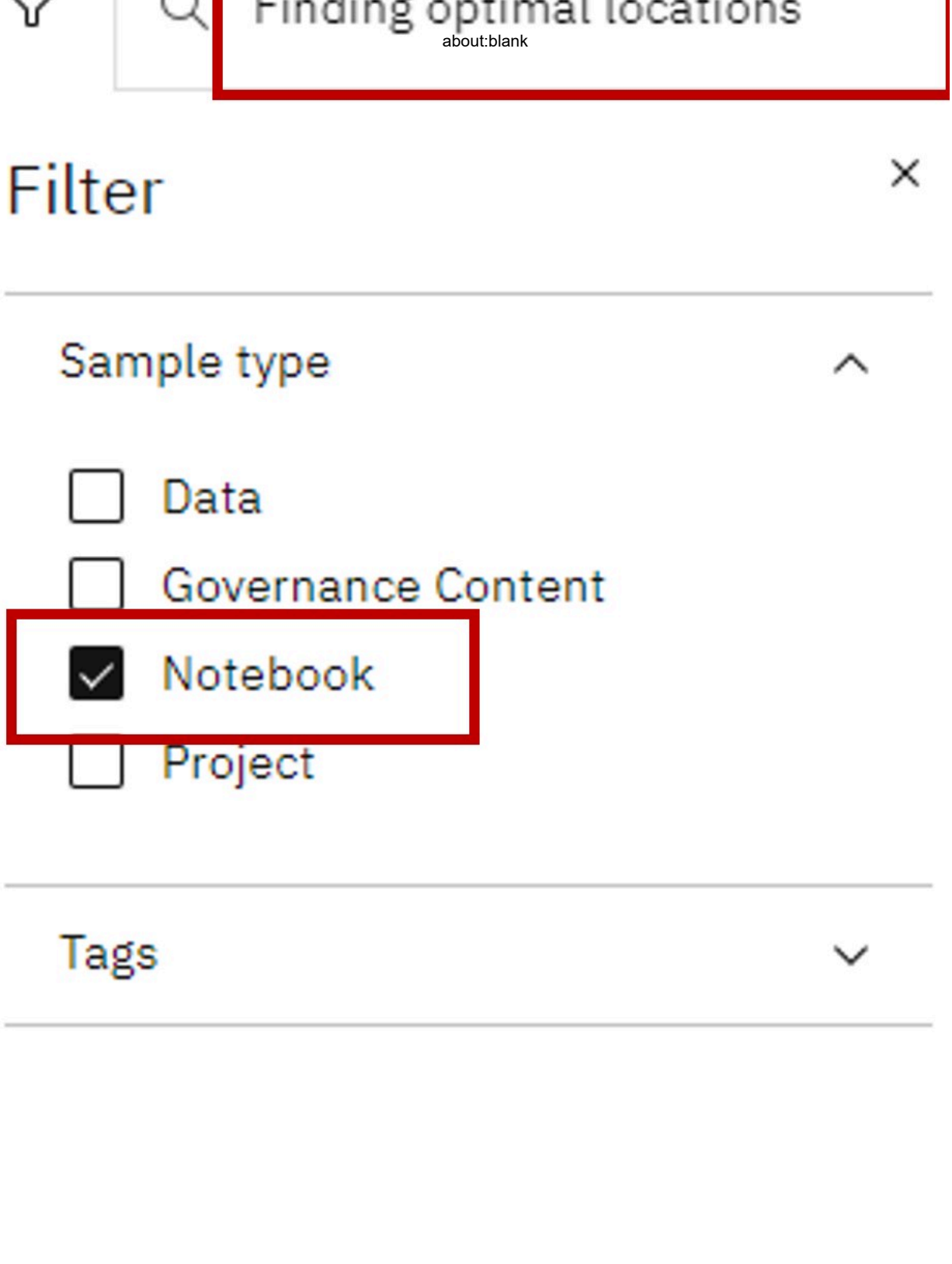
> 💡 What else might you use this data for?

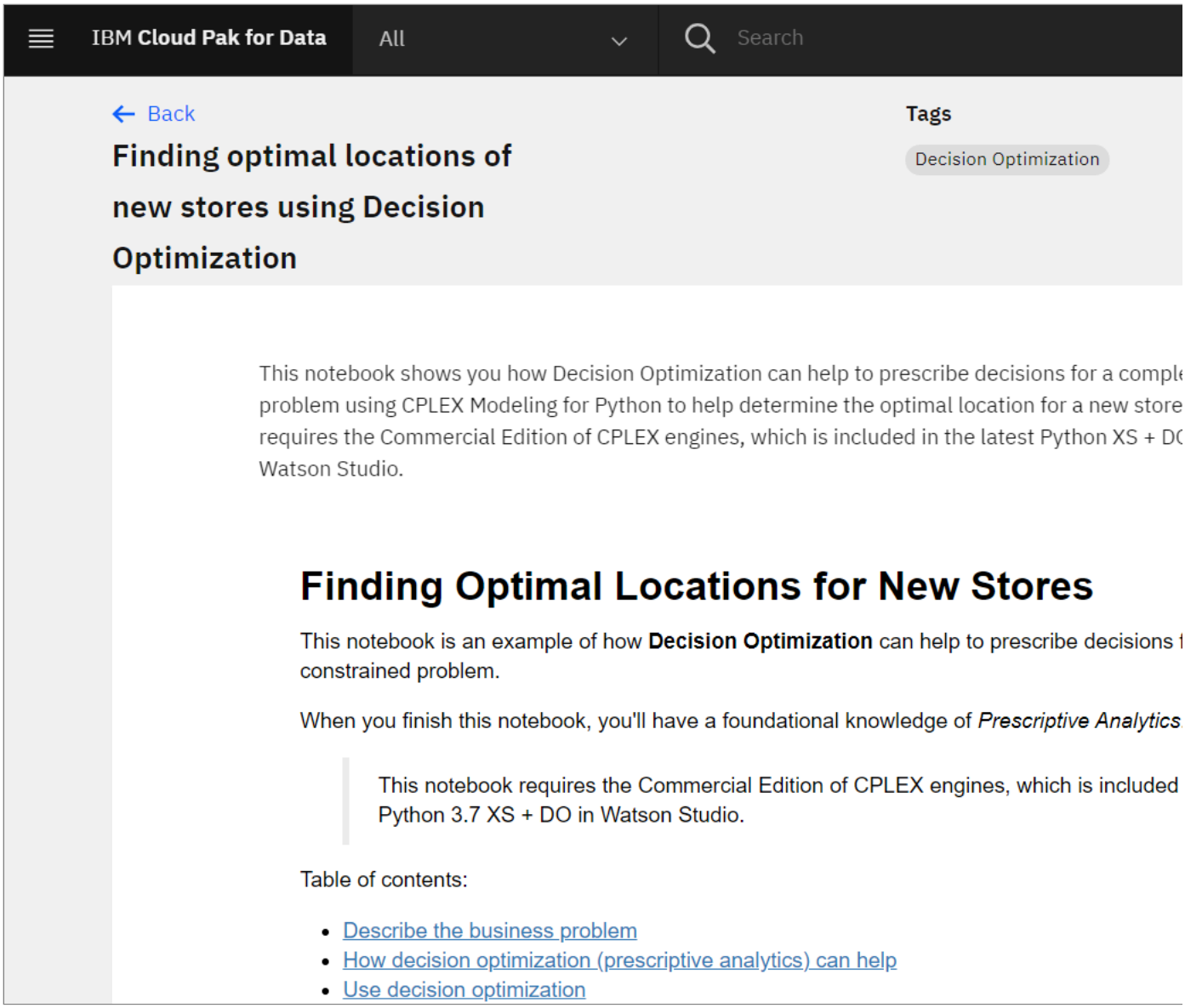
The dataset comprises of three main tables:

- 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).

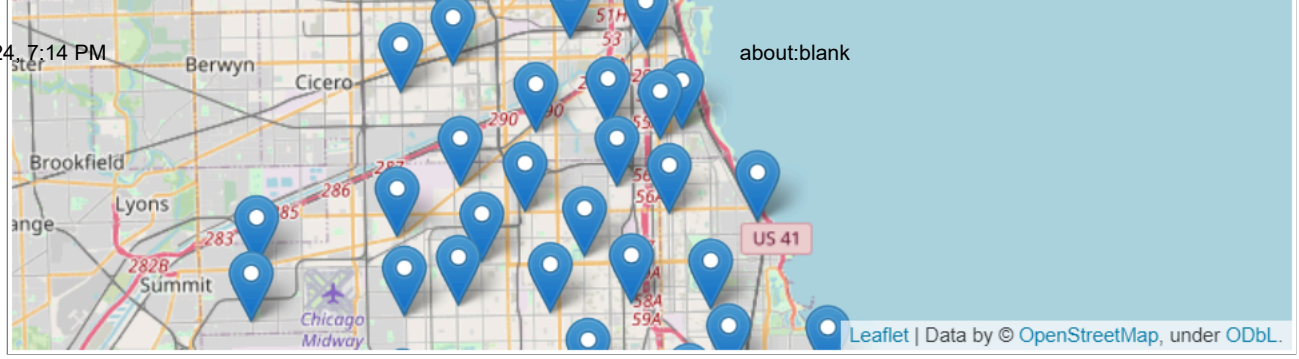
Exercise 3: Evaluate Jupyter Notebook

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…*


Finding optimal locations
Filter
Sample type
Data
Governance Content
Notebook
Project
Tags

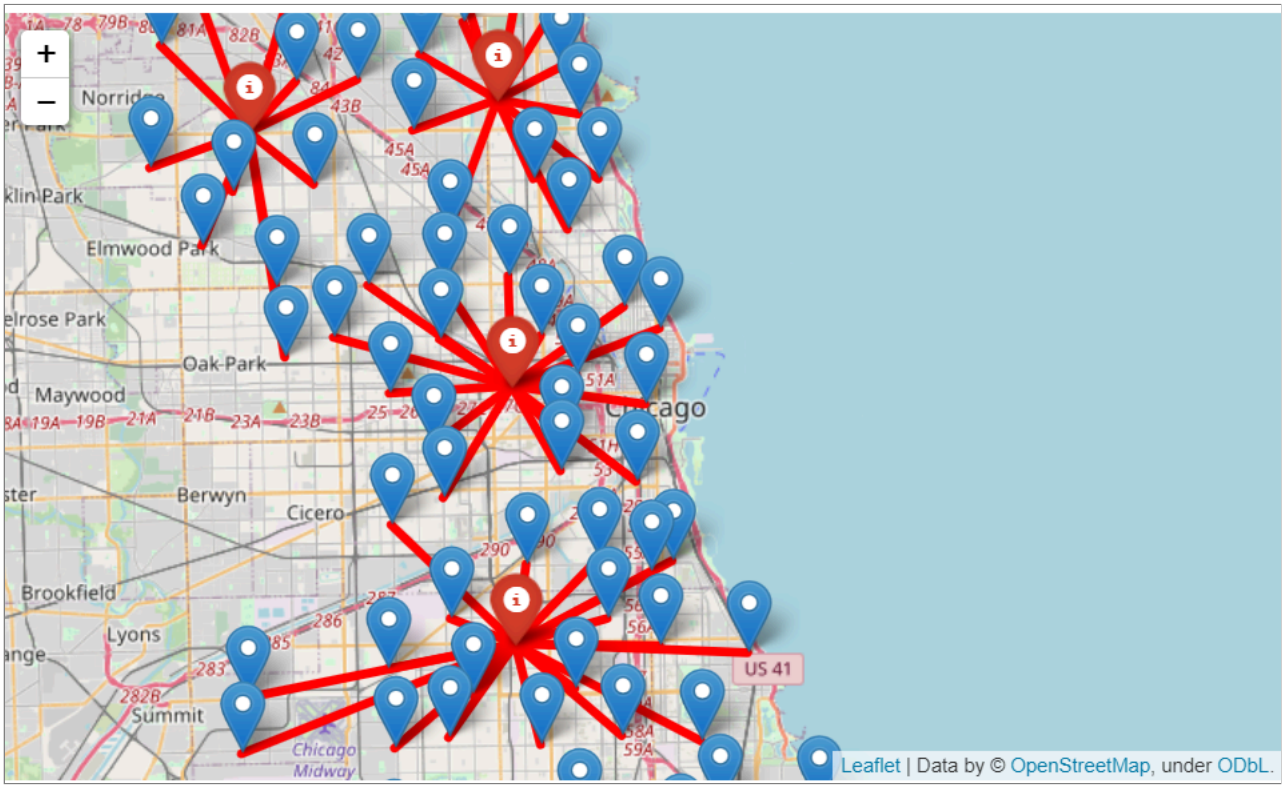


Part of the Python code in the notebook displays the locations of the libraries on a map.



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]