Establishing Data Mining Goals

The first step in data mining requires you to set up goals for the exercise. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the expected level of accuracy and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost-benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, the cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals.

Selecting Data

The output of a data-mining exercise largely depends upon the quality of data being used. At times, data are readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys. The type of data, its size, and frequency of collection have a direct bearing on the cost of data mining exercise. Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

Preprocessing Data

Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from further consideration. At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. Data should be subject to checks to ensure integrity. Lastly, you must develop a formal method of dealing with missing data and determine whether the data are missing randomly or systematically.

If the data were missing randomly, a simple set of solutions would suffice. However, when data are missing in a systematic way, you must determine the impact of missing data on the results. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. Therefore, you must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it.

Transforming Data

After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, variables may need to be transformed to help explain the phenomenon being studied. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

Often you need to transform variables from one type to another. It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual. This could help capture the non-linearities in the underlying behaviors.

Storing Data

The transformed data must be stored in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist. During data mining, new variables are created, which are written back to the original database, which is why the data storage scheme should facilitate efficiently reading from and writing to the database. It is also important to store data on servers or storage media that keeps the data secure and also prevents the data mining algorithm from unnecessarily searching for pieces of data scattered on different servers or storage media. Data safety and privacy should be a prime concern for storing data.

Mining Data

After data is appropriately processed, transformed, and stored, it is subject to data mining. This step covers data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms. A good starting point for data mining is data visualization. Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in developing a preliminary understanding of the trends hidden in the data set.

*Later sections in this chapter detail data mining algorithms and methods.*

Evaluating Mining Results

After results have been extracted from data mining, you do a formal evaluation of the results. Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an "in-sample forecast". In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

## Chapter 7. Why Tall Parents Don't Have Even Taller Children

You might have noticed that taller parents often have tall children who are not necessarily taller than their parents and that's a good thing. This is not to suggest that children born to tall parents are not necessarily taller than the rest. That may be the case, but they are not necessarily taller than their own "tall" parents. Why I think this to be a good thing requires a simple mental simulation. Imagine if every successive generation born to tall parents were taller than their parents, in a matter of a couple of millennia, human beings would become uncomfortably tall for their own good, requiring even bigger furniture, cars, and planes.

Sir Frances Galton in 1886 studied the same question and landed upon a statistical technique we today know as regression models. This chapter explores the workings of regression models, which have become the workhorse of statistical analysis. In almost all empirical pursuits of research, either in the academic or professional fields, the use of regression models, or their variants, is ubiquitous. In medical science, regression models are being used to develop more effective medicines, improve the methods for operations, and optimize resources for small and large hospitals. In the business world, regression models are at the forefront of analyzing consumer behavior, firm productivity, and competitiveness of public and private­ sector entities.

I would like to introduce regression models by narrating a story about my Master's thesis. I believe that this story can help explain the utility of regression models.

## The Department of Obvious Conclusions

In 1999, I finished my Masters' research on developing hedonic price models for residential real estate properties. It took me three years to complete the project involving 500,000 real estate transactions. As I was getting ready for the defense, my wife generously offered to drive me to the university. While we were on our way, she asked, "Tell me, what have you found in your research?". I was delighted to be finally asked to explain what I have been up to for the past three years. "Well, I have been studying the determinants of housing prices. I have found that larger homes sell for more than smaller homes," I told my wife with a triumphant look on my face as I held the draft of the thesis in my hands.

We were approaching the on-ramp for a highway. As soon as I finished the sentence, my wife suddenly turned the car to the shoulder and applied brakes. As the car stopped, she turned to me and said: "I can't believe that they are giving you a Master's degree for finding just that. I could have told you that larger homes sell for more than smaller homes."

At that very moment, I felt like a professor who taught at the department of obvious conclusions. How can I blame her for being shocked that what is commonly known about housing prices will earn me a Master's degree from a university of high repute?

I requested my wife to resume driving so that I could take the next ten minutes to explain to her the intricacies of my research. She gave me five minutes instead, thinking this may not require even that. I settled for five and spent the next minute collecting my thoughts. I explained to her that my research has not just found the correlation between housing prices and the size of housing units, but I have also discovered the magnitude of those relationships. For instance, I found that all else being equal, a term that I explain later in this chapter, an additional washroom adds more to the housing price than an additional bedroom. Stated otherwise, the marginal increase in the price of a house is higher for an additional washroom than for an additional bedroom. I found later that the real estate brokers in Toronto indeed appreciated this finding. I also explained to my wife that proximity to transport infrastructure, such as subways, resulted in higher housing prices. For instance, houses situated closer to subways sold for more than did those situated farther away. However, houses near freeways or highways sold for less than others did. Similarly, I also discovered that proximity to large shopping centers had a nonlinear impact on housing prices. Houses located very close (less than 2.5 km) to the shopping centers sold for less than the rest. However, houses located closer (less than 5 km, but more than 2.5 km) to the shopping center sold for more than did those located farther away. I also found that the housing values in Toronto declined with distance from downtown.

As I explained my contributions to the study of housing markets, I noticed that my wife was mildly impressed. The likely reason for her lukewarm reception was that my findings confirmed what we already knew from our everyday experience. However, the real value added by the research rested in quantifying the magnitude of those relationships.

## Why Regress?

A whole host of questions could be put to regression analysis. Some examples of questions that regression (hedonic) models could address include:

* How much more can a house sell for an additional bedroom?
* What is the impact of lot size on housing price?
* Do homes with brick exteriors sell for less than homes with stone exteriors?
* How much does a finished basement contribute to the price of a housing unit?
* Do houses located near high-voltage power lines sell for more or less than the rest?

## IBM Cloud Gallery

Estimated Time (45 min)

IBM Cloud Gallery is a growing collection of data sets, notebooks, and project templates. In this lab, you will use IBM cloud Gallery to explore different datasets. As we have learnt in the course, the data is not only about numbers, it can be anything such as numeric data, text data, images, videos, audios etc. You will work on three samples.

**Sample 1** in which you will learn about the dataset in which only numeric attributes are present.

**Sample 2** in which you will learn about the dataset in which numeric & text attributes are present.

**Sample 3** in which you will analyze how the Jupyter Notebooks look like. How a Data Scientist create the models?

Let's take a look that how different datasets are used by Data Scientist.

#### **Objectives :**

You will learn to:

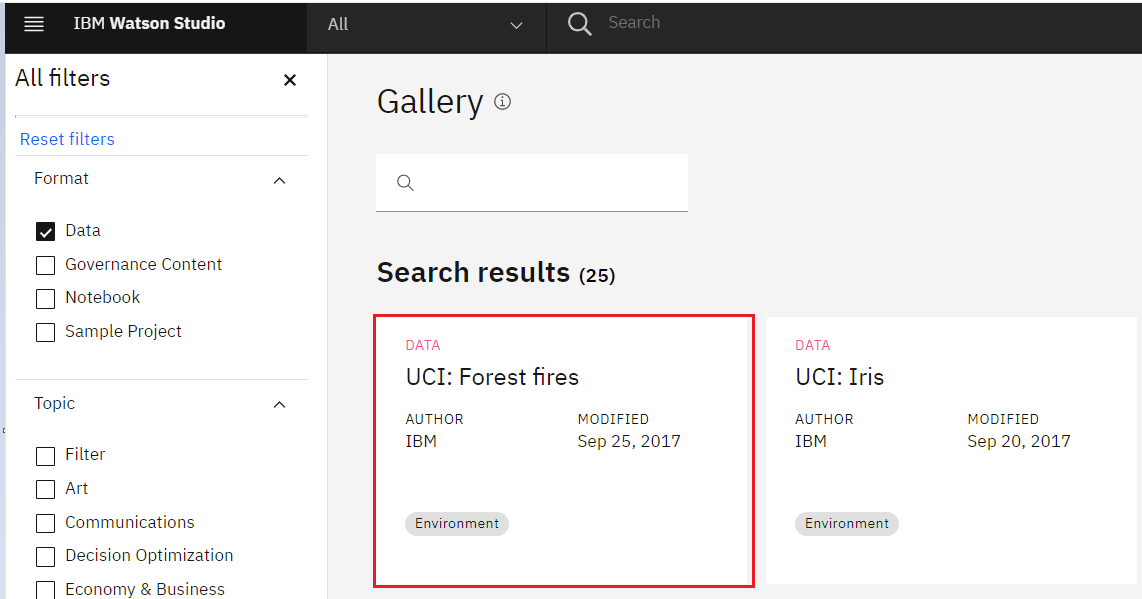
* Explore the IBM Cloud Gallery
* Evaluate Numeric dataset
* Evaluate dataset with Non-Numeric attributes
* Evaluate Jupyter Notebook

#### **Exercise 1: Evaluate Numeric dataset**

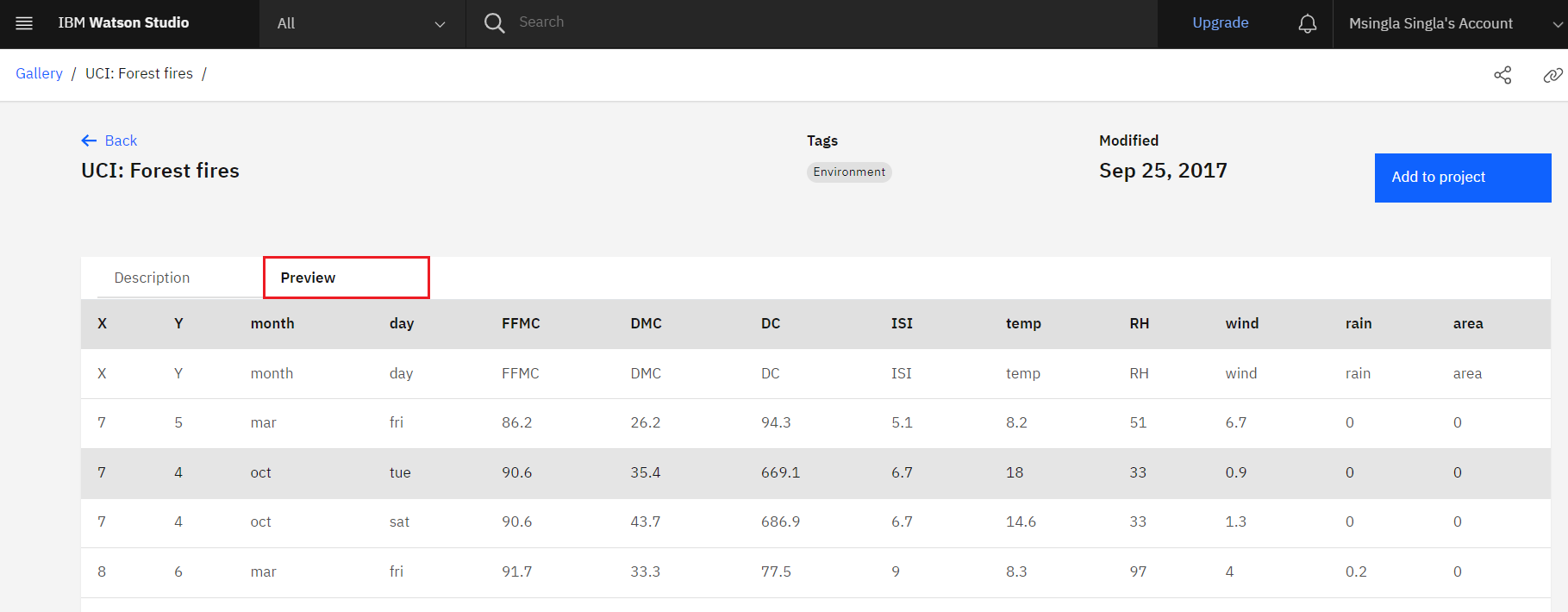
1. Click on the link: [https://dataplatform.cloud.ibm.com/gallery](https://dataplatform.cloud.ibm.com/gallery?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDS0101ENSkillsNetwork947-2022-01-01)
2. Select All Filters. From Format select Data and from Topic select Energy & Utilities, Enviornment and Industry Accelerator



1. Click on UCI: Forest Fires.



1. Preview the data using the Preview option.



##### **Explore the data**

The 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

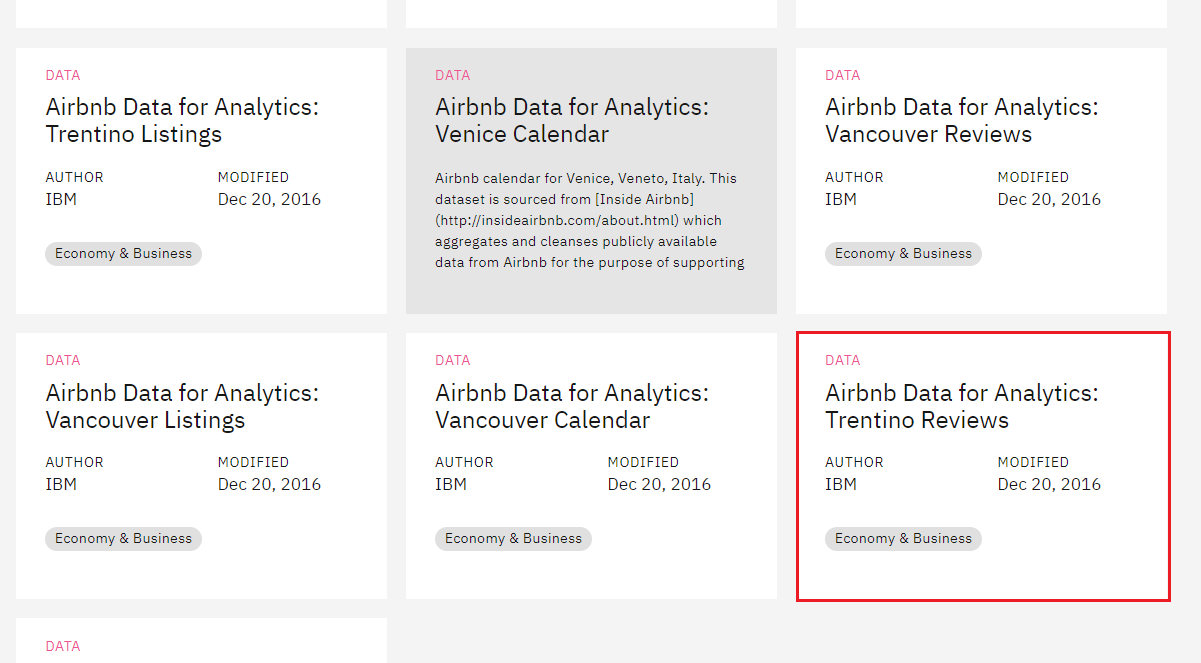
(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

### **Exercise 2: Evaluate Non-Numeric dataset**

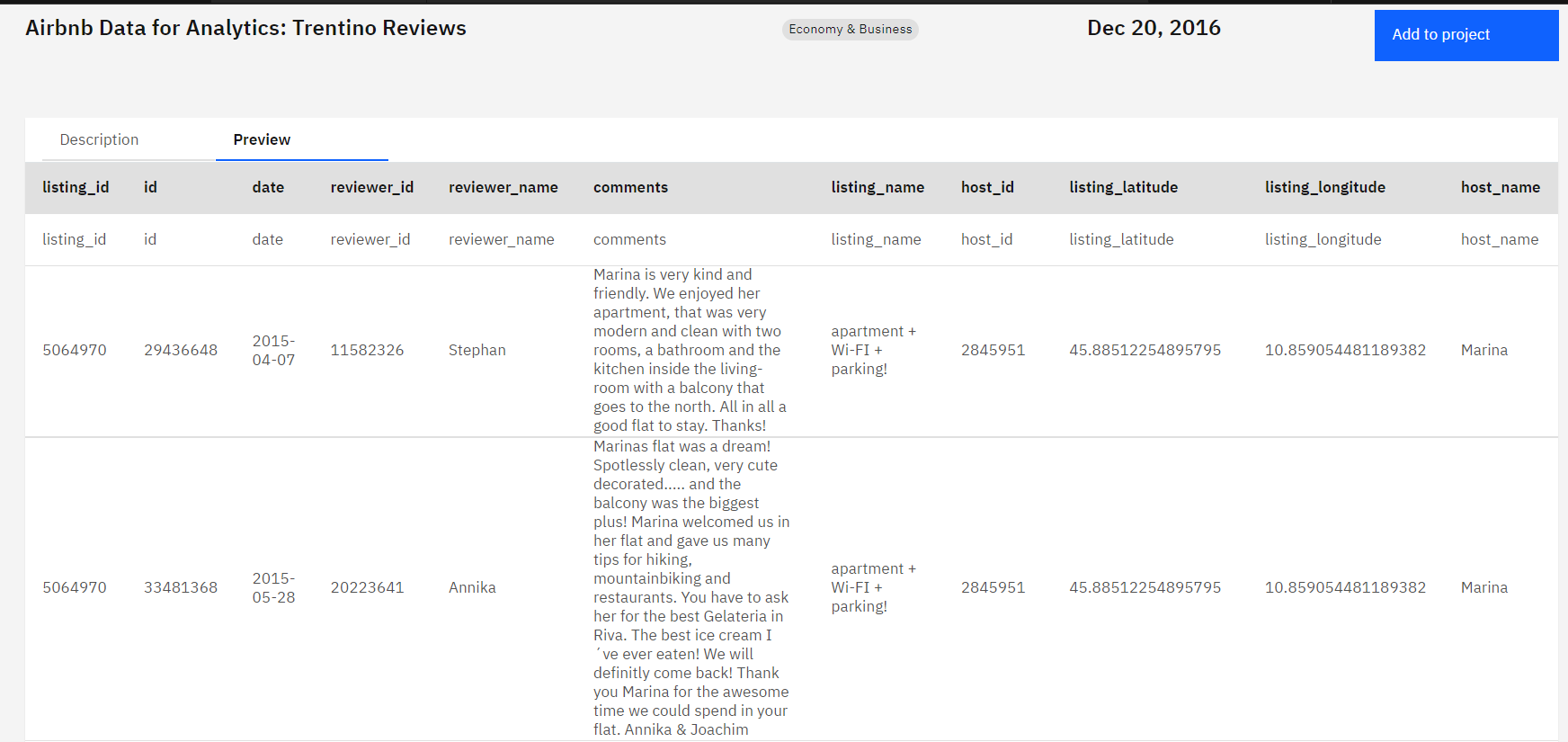
The data doesn't have to be only based on numbers. Data can be text, images and other types as well. Let's look into data having text values.

1. Use the All Filters. From Format select Data and from Topic select Economy and Business.

You will get mutiple datasets given. Scroll down and select Airbnb Data for Analytics: Trentino Reviews (If you will not get the data use the **Load More** option)



1. Preview the data using the Preview option.



##### **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.

This data can be used in various 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.

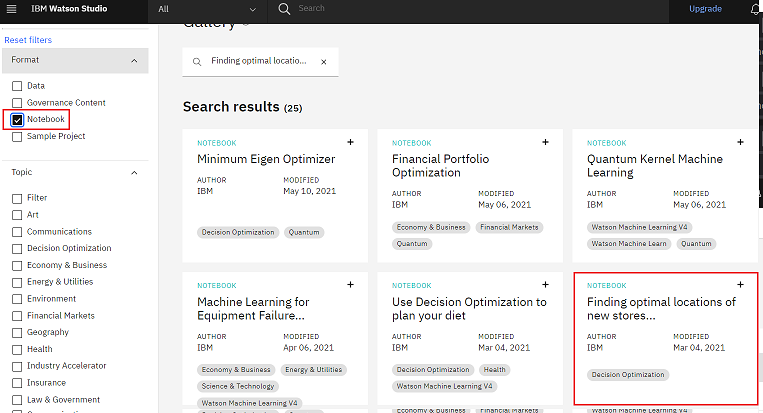
💡 Can you think of what you can use this data for?

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).

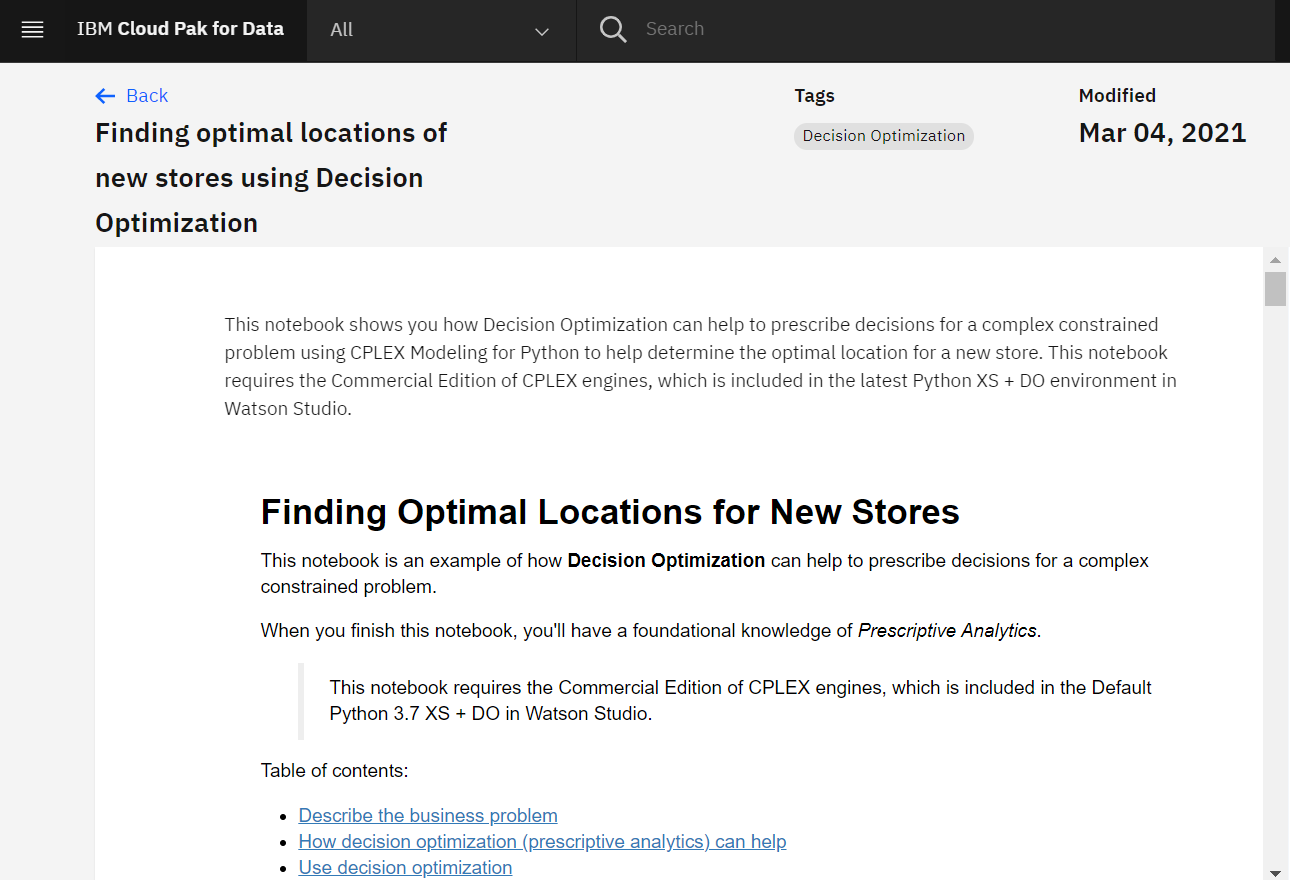
### **Exercise 3: Evaluate Jupyter Notebook**

Use the All Filters. From Format select Notebook and select Finding optimal locations of new stores using Decision Optimization (If you will not find the notebook use the **Load More** option to load the notebooks)

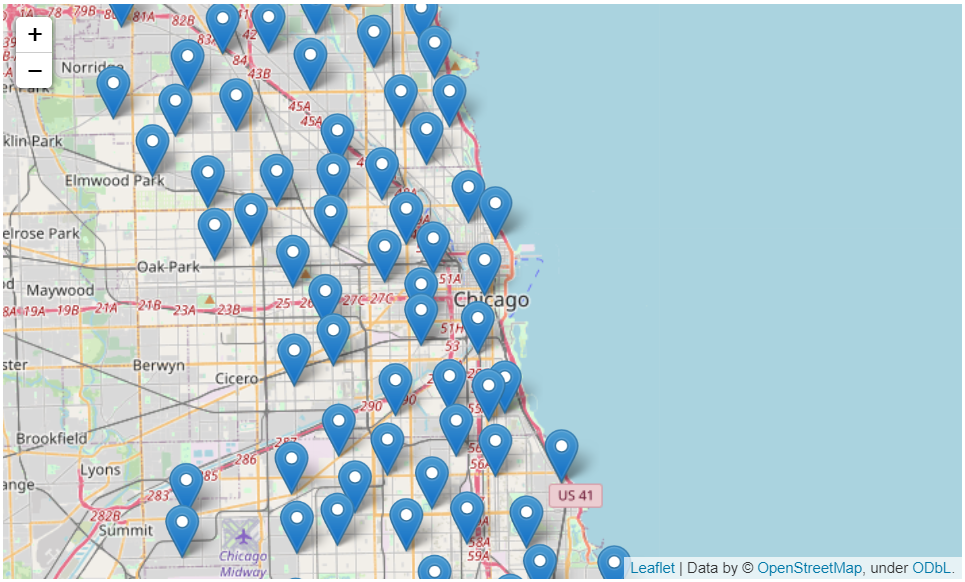


This notebook shows you how Decision Optimization can help to prescribe decisions for a complex constrained problem using Python to help determine the optimal location for a new store.

The objective is to minimize the total distance from libraries to coffee shops so that a book reader always gets to our coffee shop easily. It can be done by analyzing and displaying the location of the coffee shops on a map.

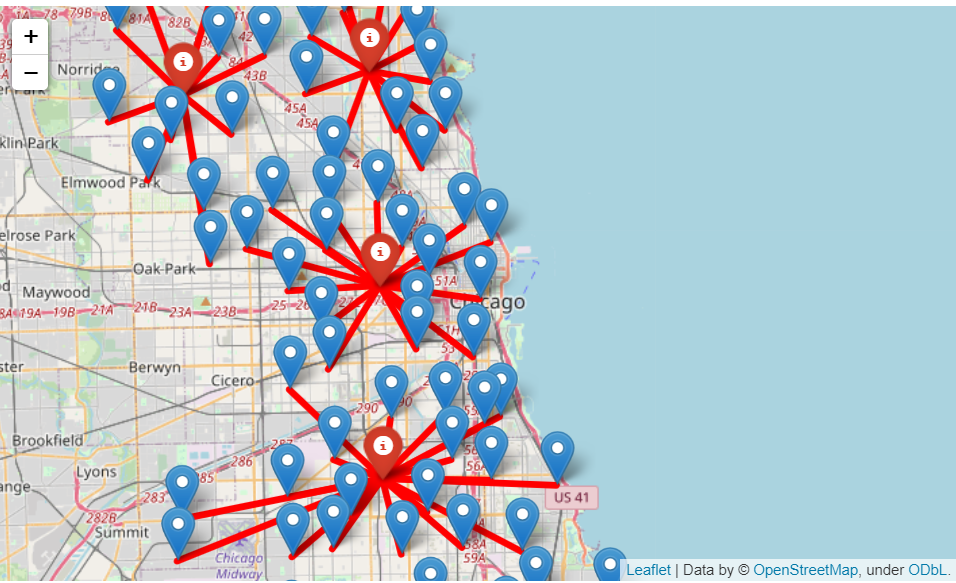


When we validate the dataset, the locations on map are seperated.



But it is impossible to determine where to ideally open the coffee shops by just looking at the map.

This is solved by an optimization model that will help us determine where to locate the coffee shops in an optimal way.



#### **Summary**

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