WHAT IS DATA MINING?

What is Data Mining?

Extracting knowledge from ***large*** amount of data

- **cleaning**: values that we get from user might not be meeting the standard set by the data owner.

For ex: age or job experience. It should be in years. What if you find 28.5 or 39.2 like values in the data? In Python, you would write some code like this: round(age)

Date formatting: it could be like 15-02-2023, or it could be like: 2023-06-09, or it could be like: Wednesday, June 7, 2023, numerous date formats are available, but machine needs a consistent format.

If the machine cannot understand the data, it cannot learn. Hence, data cleaning is required.

- **integration**: Let's say sender system is sending some string type of data. At the receiving end, you need integer, what do we do? Answer is: Type casting.

If the two systems are supposed to talk, then there should exist a protocol or a language that both systems understand. For example: in JavaScript we have 'true' and 'false'. While in Python, we have 'True' and 'False'. It is a difference of capitalization, but it is big difference from integration point of view if this data is to be passed around. If we need to zero down on a common standard, values like 0 and 1 could also help.

- **selection**: You have some data that's about the bio of a person. So maybe as a data scientist you want to explore relationship between height and weight. Or maybe between age and fat%.

In such a situation, you would have to select only the required columns from the tabular data.

- **transformation**:
type casting is again an example of transformation.

Another example could be transforming data to fit it to a scale of 0 to 1 (as in Min Max Scaling)

You want to plot the data in logarithmic scale (base could be 2, 10 or e=2.718) at that point you would be doing a transformation.
Logarithmic transformation required to convert exponential curve to linear curve.

- **mining/processing**: studying the data to get it's descriptive statistics, to know more about it's features and attributes.

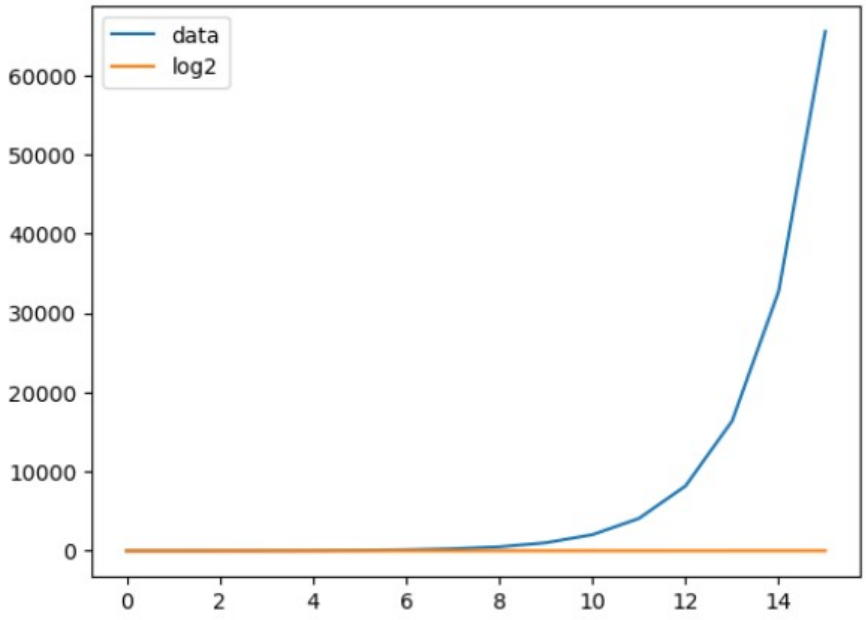
- **pattern evaluation:**
In stock market analysis: Have you heard of bearish market and bullish market?
Those are two kinds of patterns or trends in the market. There also exists a third type of trend which is 'stagnation'.
It is the job of market analysts to capture these trends.

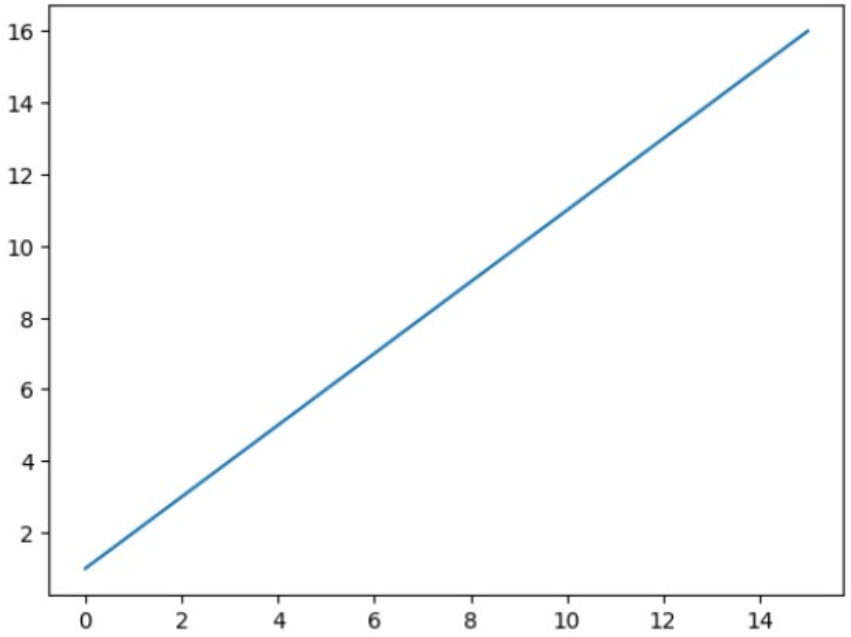
Another example could be of CPU usage. During normal processing, CPU usage would fluctuate around the mean. But when some large file is loaded or big data is loaded, CPU usage and memory usage might start to increase.

- **presentation**: drawing a scatter plot to see correlation, or drawing a line plot to see growth or decline, drawing a pie chart to show percentage composition of a composite category.

Transformation of data: from exponential scale to linear scale using logarithms

	data	log2
0	2	1.0
1	4	2.0
2	8	3.0
3	16	4.0
4	32	5.0
5	64	6.0
6	128	7.0
7	256	8.0
8	512	9.0
9	1024	10.0
10	2048	11.0
11	4096	12.0
12	8192	13.0
13	16384	14.0
14	32768	15.0
15	65536	16.0


data
log2
60000
50000
40000
30000
20000
10000
0
0
2
4
6
8
10
12
14


16
14
12
10
8
6
4
2
0
2
4
6
8
10
12
14

```
>>> age = float(input("Enter your age: "))
Enter your age: 28.5
>>> age
28.5
>>> age = round(age)
>>> age
28

>>> yob = input('Enter your yob: ')
Enter your yob: 2003
>>> type(yob)
<class 'str'>
>>>
>>> int(yob)
2003
>>>
```

What kind of patterns can be mined/found by data mining techniques?

Characterization and Discrimination:

Studying the properties of the given data. Laying down the lines for character identification of something from the data.

Frequent patterns, Associations, and Correlations

In a healthcare shop, you might notice that skin care soap goes along with skin care cream / night cream or sun screen cream.

Classification

As in image classification, cat versus dog. Or in medical science, differentiating between malignant tumor and benign tumor.

Prediction:

- monitoring of seismic waves to predict earthquake
- breakdown point of a circuit that is temperature sensitive

Cluster analysis: The task is to identify: (1) the number of clusters and

- (2) grouping the data points that make up those clusters.

Outlier analysis: what is happening (abnormally) in the extreme ends of the data?

- You have credit card transaction. You want to predict a fraudulent transaction. Every transaction has some features like: amount, place, mode of approval (touchless or via PIN)

Evolution analysis

Give examples of each of the following:

Characterization

- Characteristics of customers who buy a certain kind of product

  Maybe some person who only buys from FOSS (Free and Open Source Software) community.

  Maybe some person who only Apple products.

Discrimination

- Customers who buy product A vs customers who buy another product B

Give examples of each of the following:

- Frequent patterns, Associations, and Correlations
  - What kind of products do customers buy together?
  - If customer buys product A, what’s the chance that he/she will buy product B as well

Give examples of each of the following:

Classification

- You have get dataset containing images of animals. A task of classification would be to tell if it’s a cat image or not.
  This is a Cat and ‘Not Cat’ Image Classification problem.

Prediction

- Sales Forecasting: Predict the sales of the product
- House price prediction based on number of rooms that are there, area, number of balcony and bathroom, etc.

Give examples of each of the following:

- Cluster analysis
  - Divide the data into different groups of similar items
  - No. of cluster are not known apriori

  - Example: During Covid-19, we clustered the population into color coded regions. Red zone areas, orange zone and green zone etc.
  - Features:

    Number of infections

    Population density

    Severity of cases

    So and so forth

Give examples of each of the following:

- Outlier analysis
  - Deviant data from the expected
  - Centenary population: people who live past the age of 100. They are like outliers.

Give examples of each of the following:

- Evolution analysis
  - Time series analysis of data

Discuss whether or not each of the following activities is a data mining task.

Dividing the customers of a company according to their gender.
No. This is a simple database query.

Getting all male customers.

SQL> Select * from customers where gender = ‘M’;

Discuss whether or not each of the following activities is a data mining task.

Dividing the customers of a company according to their profitability.

No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

A customer's data could look like:
1. Operational expense data
2. Categories / Departments in which person makes purchases.

Threshold is set at: 5K INR in a month for grocery (viz. Department) to identify the customer as profitable in that department.
It could be 10K INR for home appliances and 2K INR for dairy products (just an example).

Setting this threshold is a problem of prediction (Data Mining).

Only requires writing a SQL query now.
Broad steps: aggregating and filtering if the threshold is given.

Discuss whether or not each of the following activities is a data mining task.

Predicting the outcomes of tossing a (fair) pair of dice.
No. Since the die is fair, this is a probability calculation.

Let’s say you want: 3 on the throw of a dice.

Size of sample space = {1, 2, 3, 4, 5, 6} = 6

P(event = 3) = 1 / 6

Discuss whether or not each of the following activities is a data mining task.

Predicting the future stock price of a company using historical records.

Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modelling.

Stock Market Analysis (or Prediction)

Discuss whether or not each of the following activities is a data mining task.

Monitoring seismic waves for earthquake activities.

Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.