Data Collection, Types of Data & Types of Variables

Data Collection

Data collection is the process of gathering and measuring information on targeted variables.

Data is generated through many ways –

- Data generated by humans e.g. surveys
- Data generated by machines/ systems e.g. ERP

Typically, as a Data Scientist one is not required to collect the data himself. He is provided with that data by the client/ organization. However, there could be:

- Multiple systems and sources involved *e.g. An ERP system, few .csv files, an RDBMS database*
- Multiple stakeholders involved *e.g. Marketing team, operations team, sales team*
- Multiple methods of integration involved *e.g. data transfer by mail, data transfer through RDBMS connection, data transfer through FTP etc.*

So a Data Scientist need to gather all the data at one place and one system so that he can start processing these datasets.

Types of Data

The primary ingredient for Data Science business problems is **DATA**. But data comes in all shapes and sizes. It can be structured, semi structured and unstructured.

STRUCTURED DATA

- Data is stored, processed, and manipulated in a tabular format (rows and columns)
- It can be easily mapped into pre-designed fields.

UNSTRUCTURED DATA

- Data that doesn't fit into structured format.
- E.g. images, scientific data, video, texts and emails, audio files etc

SEMI STRUCTURED

- Data doesn't fit into a structured database system in their raw form, but with some processes you can convert them into tabular format
- **Examples –** CSV, XML , JSON etc

Types of Data – Examples

UN-STRUCTURED

All of this data is only designed for humans to read and process. By looking at the page, you can easily differentiate between the product specifications and the images but can't expect the same of your machines

−<product>
 −<productname>
  Apple iPhone 5 A1428 Facto
 </productname>
 −<product_url>
  http://www.amazon.com/Apr
 </product_url>
 <product_id>B00ASURUYQ<
 <price>$369.99</price>
 <rating>3.8 out of 5 stars</r
 <color>black</color>
 <stock>2</stock>
</product>

SEMI-STRUCTURED

XML snippet. Each data point is clearly tagged, and it's easy to entitize each record encapsulating all these data points.

PRODUCT	PRODUCT NAME	PRODUCT URL	PRODUCT ID	PRICE
MOBILE	APPLE IPHONE	HTTP://WWW.	B00ASURUYQ	$369.

STRUCTURED

Data finally stored in the database. This is ready to be used for any analysis.

Types of data variables

- Data consists of a combination of "variables" which actually contain the values
- Variables at a high level are of two types depending on the kind of values they store:
  - Numerical
  - Categorical

Numerical variables

- Discrete
  - Arises from counting
  - Can take only a set of particular values including negative and fractional values
  - Examples: Credit score, number of credit cards owned by a person, number of states in a country, charge on electron etc.
- Continuous
  - Arises from measuring
  - Can take any value with in a specified range
  - Examples: Height, Amount of money, Age etc.

Categorical variables

- Binary (or Dichotomous)
  - Has only two categories
  - Examples: yes/no, male/female, pass/fail etc.
- Nominal
  - Has several unordered category
  - Examples: Type of bank account, type of insurance policy etc.
- Ordinal
  - Has several ordered category
  - Examples: questionnaire responses such as "strongly in favor / … / strongly against".

Types of Variables

	CONTINOUS	BINARY	NOMINAL	ORDINAL	DISCRETE	NOMINAL
FORM (TEXT/NUM)	NUM	TEXT	TEXT	TEXT	NUM	NUM
NO. OF DISTINCT VALUES (HIGH/LOW)	H	2	L	L	L	H
ORDER (YES / NO)	Y	N	N	Y	Y	N

Exercise: Give at least two examples of each of the variable type

Identify the variable type

Variable Name	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer Last Birthday	Gender of Customer	Marital Status of Customer	Annual Salary	Monthly Credit Card Usage
Value Stored	Name of the individual customer	Unique identifier	1, 2, 3…	18, 19, 20…	Male / Female	Married / Divorced / Never Married	Amount	Low(<25%) / Medium(<50%) / High(<75%) / Very High(>75%)
Variable Type								
Remarks								

Thanks!

Any questions ?

Next steps

- Attempt quiz –
- Share feedback –