Batch: A2	Roll No.: 16010122041
Experiment 01	
Grade: AA / AB / BB / BC / CC / CD /DD	

Title: Data Collection and finalizing dataset from problem domain

---

Objective:

1. To learn how to collect the dataset

2. To learn sources of dataset

3. To assess the dataset based on Metrics to Measure Data Quality

4. To finalize the features of dataset

Course Outcome:

CO1 : Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

Books/ Journals/ Websites referred:

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

https://grouplens.org/datasets/movielens/

https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Resources used:

1. Kaggle
2. IMDb
3. Cornell Edu
4. Full Movies Lens

Theory:

- **Problem Domain: Entertainment – Movies**

  There are thousands of movies existing in this world but when it comes to picking one, it becomes a hectic task. People are often confused which movies to watch on various occasions. We are building a **movie recommendation system** which will suggest movies based on the mood of the viewers, for better experience. That way, people can spend less time to choose the movies and more time in enjoying a great movie which is handpicked, to suit their moods. Each mood is linked with a particular genre so as per user's choice we will recommend them a movie relating to that genre.

- **Brain stormed features of Dataset:**

  The dataset should necessarily have attributes such as title, date of release and genre as they will be used in recommending movies. While searching for datasets, we came across three datasets:

  1. IMDB dataset: 85,000 movies, last updated 2020 and movies in a variety of languages with detailed information.
  2. Cornell dataset: 8000 movies, last updated 2009 and only English movies, with title, genre, plot, etc.
  3. Fullmovielens dataset: 2 lakh movies,last updated 2019, only English movies and information contained only title and genre.

     Some additional features which the dataset can have are date of release, duration, plot, director, actors, etc.

- **Motivation for the selected dataset**

  Out of the three datasets, we chose the dataset provided by IMDB, which was available on Kaggle.

  The dataset which we downloaded contains the name of the movies, the year they were released, the duration, country of release, language, director, actors, a short description, average votes, reviews from users and reviews from critics. According to us, this information is useful in suggesting a good movie for the users.

  There were 22 columns and 85,000 movies in our dataset. The motivation for choosing it is that it is very informative, updated, organized and easy to understand.

- **Source of dataset**

We found this dataset on Kaggle. This dataset is provided by IMDB. The link for downloading this dataset:
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

- **Sample of Finalized dataset and its source**

  This was the original dataset:

imdb_title_id	title	original_title	year	date_published	genre	duration	country	language
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None
tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70	Australia	None
tt0001892	Den sorte drøm	Den sorte drøm	1911	1911-08-19	Drama	53	Germany, Denmark	
tt0002101	Cleopatra	Cleopatra	1912	1912-11-13	Drama, History	100	USA	English
tt0002130	L'Inferno	L'Inferno	1911	1911-03-06	Adventure, Drama, Fantasy	68	Italy	Italian
tt0002199	From the Manger to the Cross; or, Jesus of Nazareth	From the Manger to the Cross; or, Jesus of Nazareth	1912	1913	Biography, Drama	60	USA	English
tt0002423	Madame DuBarry	Madame DuBarry	1919	1919-11-26	Biography, Drama, Romance	85	Germany	German
tt0002445	Quo Vadis?	Quo Vadis?	1913	1913-03-01	Drama, History	120	Italy	Italian
tt0002452	Independenta Romaniei	Independenta Romaniei	1912	1912-09-01	History, War	120	Romania	
tt0002461	Richard III	Richard III	1912	1912-10-15	Drama	55	France, USA	English
tt0002646	Atlantis	Atlantis	1913	1913-12-26	Drama	121	Denmark	Danish
tt0002844	Fantômas - À l'ombre de la guillotine	Fantômas - À l'ombre de la guillotine	1913	1913-05-12	Crime, Drama	54	France	French
tt0003014	Il calvario di una madre	Ingeborg Holm	1913	1915-10-18	Drama	96	Sweden	

We cleaned the dataset and excluded some irrelevant data according to our needs. This is a sample of the finalized dataset:

title	year	genre	duration	country	language	director	actors	description	avg_vote	reviews_fr	reviews_from_critics
Miss Jerry	1894	Romance	45	USA		Alexander	Blanche Ba	The adven	5.9	1	2
The Story (	1906	Biography,	70	Australia		Charles Ta	Elizabeth T	True story	6.1	7	7
Den sorte	1911	Drama	53	Germany,	Denmark	Urban Gad	Asta Nielse	Two men (	5.8	5	2
Cleopatra	1912	Drama, His	100	USA	English	Charles L.	Helen Garc	The fabled	5.2	25	3

- **Justification for choosing above dataset**

  1. We think 85,000 movies were sufficient for providing recommendations
  2. The dataset consists of old as well as new movies, the oldest movie being of 1894 and the latest movie being of 2020.
  3. The dataset has movies in a variety of languages like English, German, Italian, French, Hindi, etc.
  4. As we need genres for making recommendation, this dataset has detailed information of a wide variety of genre as well as information about the movie, plot, actors, etc.

Conclusion

Thus, we learnt how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

Post Lab Question:

1. **Explain Role of Data in the Application Design.**

   Application design proceeds in parallel with conceptual design in a DBMS-independent way. When a database system is being designed, the designers should be aware of the transactions/applications that will run on the database. An important part of database design is to specify the functional characteristics of these transactions early in the design process. This ensures that the database will include all the information required by these transactions.

   • Thus to fill these databases correctly the need for Data is of utmost importance as it serves as the backbone of any application

   • Only when the data is accurate and consistent can the rest of the application work in sync

   • Without adequate data there is possibility for errors and inaccuracy to creep into the code/software

   • All functionality rests on the database and designers find the need to work on correct data

2. **Write different types of Data with Example.**

   1. Quantitative data

   Quantitative data seems to be the easiest to explain. It answers key questions such as “how many, “how much” and “how often”.

   Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.

   Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

   Examples of quantitative data:

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

There are 2 general types of quantitative data: discrete data and continuous data. We will explain them later in this article.

2. Qualitative data

Qualitative data can't be expressed as a number and can't be measured. Qualitative data consist of words, pictures, and symbols, not numbers.

Qualitative data is also called categorical data because the information can be sorted by category, not by number.

Qualitative data can answer questions such as "how this has happened" or and "why this has happened".

Examples of qualitative data:

- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand and etc.
- Names as John, Patricia,…..
- Ethnicity such as American Indian, Asian, etc.

3. Nominal data

Nominal data is used just for labeling variables, without any type of quantitative value. The name 'nominal' comes from the Latin word "nomen" which means 'name'.

The nominal data just name a thing without applying it to order. Actually, the nominal data could just be called "labels."

Examples of Nominal Data:

- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)

As you see from the examples there is no intrinsic ordering to the variables.

Eye color is a nominal variable having a few categories (Blue, Green, Brown) and there is no way to order these categories from highest to lowest.

4. Ordinal data

Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data.

Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.

However, **you cannot do arithmetic with ordinal numbers** because they only show sequence.

We can also assign numbers to ordinal data to show their relative position. But we cannot do math with those numbers. For example: “first, second, third…etc.”

Examples of Ordinal Data:

- The first, second and third person in a competition.
- Letter grades: A, B, C, and etc.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- Economic status: low, medium and high.

5. Discrete data

Discrete data is a count that involves only integers. To put in other words, discrete data can take only certain values. The data variables cannot be divided into smaller parts.

It has a limited number of possible values e.g. days of the month.

Examples of discrete data:

- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

6. Continuous data

Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.

For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.

You can record continuous data at so many different measurements – width, temperature, time, etc. This is where the key difference from discrete types of data lies.

The continuous variables can take any value between two numbers. For example, between 50 and 72 inches, there are literally millions of possible heights: 52.04762 inches, 69.948376 inches and etc.

Examples of continuous data:

- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.