What is data and analysis?

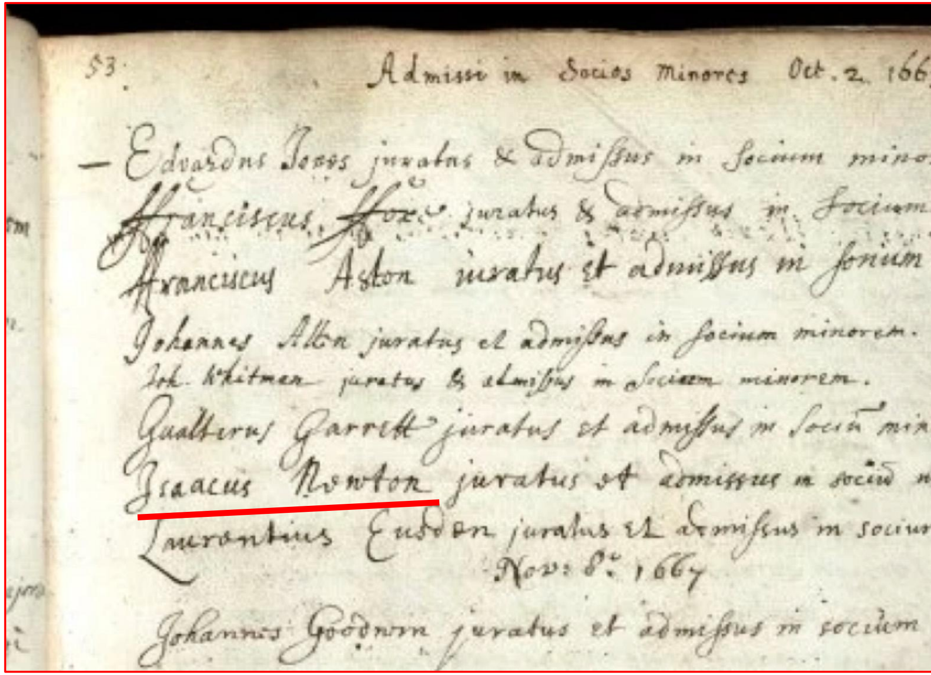
SF Free coding Bootcamp

What is data?

- Data is a recording of fact. It is generated as an outcome of *book-keeping.*
    - Sales records of a shop
    - Entry-exit records in a secure building
    - Transaction records in a bank
    - Census records for a state or a country
    - In drug trials, observations are recorded over time.
    - Exit Poll surveys
    - Automatically generated log every time a Google search is done.
- This definition is narrow, but for thinking about data analysis, this is accurate.

The medium of data collection

- Data has been collected for thousands of years - earliest known accounting was done on clay tablets!
- Today, most of the data is collected in a digital format. Older collections of data, in many cases, is already digitized.



53.

Admiſsi in Socios minores Oct. 2. 166

Edvardus Jones juratus & admiſſus in Socium mino

~~Franciscus Fox~~ juratus & admiſſus in Socium

Franciscus Aston juratus et admiſſus in Socium

Johannes Allen juratus et admiſſus in Socium minorem.

Joh. Whitman juratus & admiſſus in Socium minorem.

Gualterus Garrett juratus et admiſſus in Sociū min

Isaacus Newton juratus et admiſſus in sociū m

Laurentius Eusden juratus et admiſſus in socium

Nov: 8^{o}. 1667

Johannes Goodwin juratus et admiſſus in socium

Exercise: Creating a survey

- To understand the nature of data better, here is a small exercise.
- **Design a survey, with 5 questions, to determine what kind of music people prefer.**

- Now, imagine that you’ve surveyed 100 people using the survey questions you created. **Create a small snapshot of the resulting data.**

Discuss survey data and its representation

Structure of data

- Even though data can record many kinds of facts, there is a certain structure to all data.
- Data is represented in **tabular format**. Each record is listed as a separate row, and all the fields of data collected in a record are listed in columns.
- The list of columns are unique to the data. But, there is some structure to the kind of fields/columns present in different sources of data as well.

Structure of data: Kinds of fields

- Mandatory fields:
    - Date and time - data loses most of its value if we don’t know when it was collected.
    - ID - we need to be able to identify each row of data uniquely. In a large number of cases, it is just an incrementing record number.
- Types of fields:
    - Text - To store freeform data like name, address or any notes.
    - Number - for example, to record age, or air temperature
    - Fixed formats - for storing things like dates or phone numbers
    - A selected option among many choices - Like gender
- Required and Optional fields
    - In many surveys, some data fields may be optional and the data is still meaningful if those fields are not present in some rows.

Structure of data: Metadata

- In many complicated sets of data, the main data table is usually accompanied by multiple tables of **metadata.**
- **Metadata** is data which is not the main record of fact, but contains some additional information to complete your understanding of the main data.
- For example, in a survey the main data might record the zip code from where the data was located. A metadata table can help us identify the city to which the zip code belongs.

Structure of data: formats

- Since we can generalize the structure of data, most data programs (both for storage and analysis) are designed to assume this structure.
  - CSV (Comma Separated Values) is the most common file format in which data is shared.
  - Data is frequently stored in proprietary database formats (like MySQL DB).
  - Systems which generate very large amounts of data (like Google Search) have specially designed formats to minimize storage, and custom big data applications are typically used to analyze this logs.

What is data analysis?

- Let’s continue onwards from where we left off with the last exercise.
- You are starting your own music streaming service. **How can you utilize the data that was collected in the survey we created?**

Answering questions from data

- The aim of data analysis is to answer questions and gain insights.
- These insights are typically used to make informed decisions.
- It is important to keep in mind that any data is a snapshot over some period of time. Any insights derived from that data is thus truly valid only for that period of time.

The process of analysis

- Raw data typically has a lot of rows and columns, and it is very hard to simply peruse thousands of rows of data using by eye and conclude anything meaningful.
    - That said, perusing the data by eye is an essential first step in analysis, and informs subsequent steps.
- In order to gain insights using analysis, we need to **transform** our original data - usually multiple times.
- Most of these transformations are an attempt to **reduce the size of data**, either by filtering or summarizing.

The music survey

ID	Gender	Age group	How many hours of music listened per week	Favorite genre	Favorite artist
1.	M	30-40	10	Rock	Muse
2.	F	21-30	20	Pop	Beyonce
3.	F	21-30	25	Rap	Drake
4.	M	16-21	3	EDM	Avicii
5.	F	30-40	5	Jazz	Nina Simone

Your job is to highlight relevant artists on the homepage of your music streaming app.
How can you use this survey?

Top 5 genres

Steps

1. Create a frequency chart for each genre - adding one per Genre every time it’s mentioned in the data. (**Summarization -** group by and aggregate)
2. Select the genres with top 5 frequency scores (**Filtering**)

Top 5 artists among females.

Steps

1. Select only the data where gender is marked as Female. (**Filtering**)
2. The next two steps are the same as previous problem, except that we replace genre by artist (**Summarization**, **Filtering**).

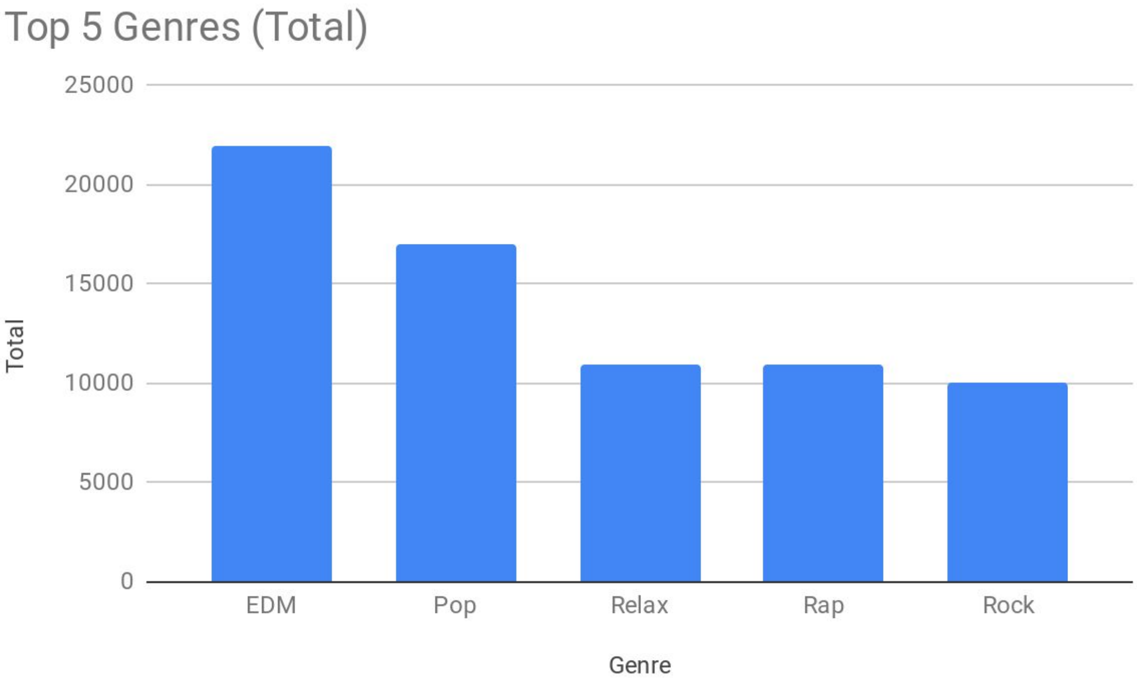
Outcome of analysis

- The outcome of any analysis is a new table containing the data which is easier to interpret.
- This outcome also conforms to the structure of data as we discussed earlier.
- Most commonly, this outcome is visualized in charts, to make it easier to understand and make inferences.

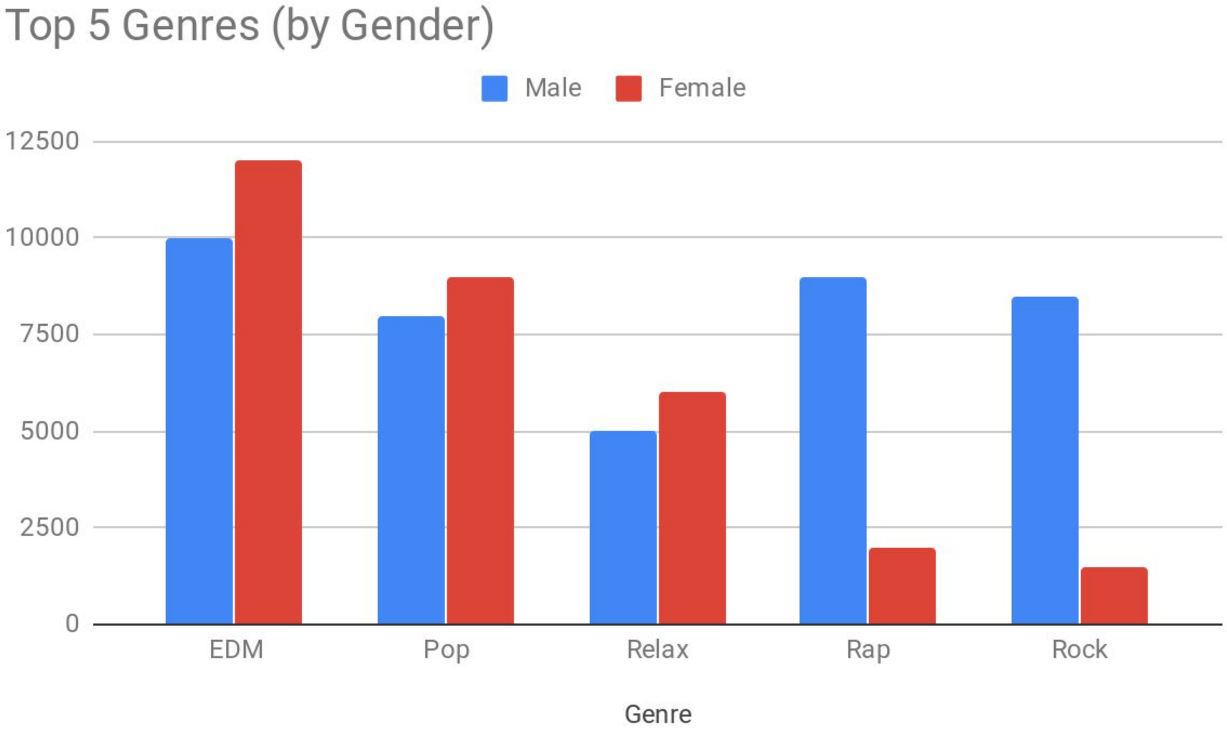
Outcome: Top 5 genres

Genre	Male	Female
EDM	10000	12000
Pop	8000	9000
Relax	5000	6000
Rap	9000	2000
Rock	8500	1500

Visualization: Top 5 genres



Visualization: Top 5 genres (another view)



Analyst’s job

- For an analyst, it is important to understand the domain of the data, and the purpose for which the analysis is being done.
- In real life, answering questions and gaining insights is not a one directional process. It is iterative.
- **For example**, let’s say you found the top 10 Artists from the survey. But none of the top 10 Artists produce music from “Relax” genre. What do you do?

Exercise

1. Define the list of fields/columns that you *log* whenever a Netflix user starts and finishes watching a show (one movie or episode).
2. Use a spreadsheet software to create a fake table (but realistic) with some rows of data.
3. You are working for the VP of content sourcing at Netflix. What kind of insights might she want from the data?
4. Define the outcome data table for those insights, and create fake data (but realistic) in that table.
5. Use the spreadsheet software to create charts (any type) for visualizing those insights.