THE LOVE FINDERS – ETL Pipeline Project

1. Project Description

You are single and are looking for love.
In XXI century, in the times of COVID-19 global pandemic and frequent lockdowns, chances are you would consider using a dating site / app. Which one will you go for? Which app has most users? Which has the best rating? What attributes people look for in their "second – half"? Does your age matter? How many people of your age are going to use the app? How big chances do you have to be successful?

The lockdown is off, and you can finally meet!
How to make a good impression and secure your second date? Should you even try to do that? Maybe being sincere is the best strategy? What traits of yours will bring you success / failure?

2. The choice of subject

Love is a natural human need. This is subject concerning us all. In times of COVID-19 a lot of people felt isolated. We are curious how people who are looking for love could succeed in current "love market".

3. Benefits of choosing this subject

By exploring the most popular dating apps and websites, we will know where to look for love to be most effective and efficient in the search. Analysing the most in demand traits will help us to understand how to flourish in our love life.
Using ETL pipeline will help us in preparation of the data for analysis and is useful as it enables moving and transformation of data internally. It is also a great foundation for the future if we wanted to use it as a base for data warehousing for business intelligence purposes (e.g., if we wanted to create our own product, conduct a meta-analysis of the subject, or analyse changes in the matter over time).

4. Datasets

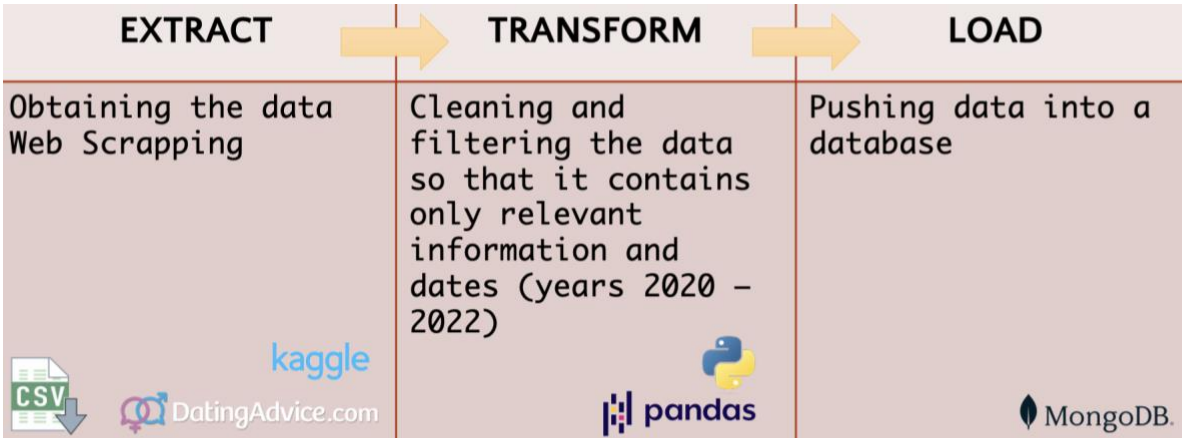
We used two datasets from Kaggle and scrapped one website:

- Speed Dating Analysis: https://www.kaggle.com/datasets/somesh24/speeddating
- Dating Apps Reviews 2017-2022: https://www.kaggle.com/datasets/sidharthkriplani/datingappreviews
- 24 Dating Sites With the Most Users: https://www.datingadvice.com/online-dating/dating-sites-with-the-most-users

5. Technologies

- Excel & CSV – as datasets come in this format
- Python – to clean scrapped data and connect to the database
- Pandas library – to clean data and create data frames, so our data can be uploaded to database
- Beautiful Soap library – to scrap the website
- Mongo Database - was chosen as the data is not related, so there is no reason to use relational database such as e.g., PostgreSQL
- Ms Word – to create and export this report
- Git & GitHub to collaborate with the team

6. **ETL Diagram**



EXTRACT	TRANSFORM	LOAD
Obtaining the data Web Scrapping	Cleaning and filtering the data so that it contains only relevant information and dates (years 2020 – 2022)	Pushing data into a database

7. **Data Transformation - Steps**
- Drop rows with empty cells
- Reduce columns and rename them
- Remove dates that were not relevant (everything before 2020)
- Scrap the website
- Storing the website source in a template for ease of use
- Searching through headline
- Removing tags

8. **Challenges**
- Very little amount of time to complete the project
- Technical problems – with Internet, different time zones

9. **Division of tasks by group members:**

Rita Starzyk:
- Writing Project Proposal
- Web Scrapping
- Uploading data to the Database
- Creating Read.me file in GitHub repository
- Combining all files together into one big project

Kouame Kwasi
- Cleaning and transforming "Speed Dating" Data
- Created connection with MongoDB using pymongo
- Uploading data to MongoDB

Daniela Shae-Bebeyi
- Cleaning and transforming "Reviews" Data
- Uploading data to the Database

Motasim Nasir
- Finding the data
- Conducting research in the subject matter and creating research summary
- Project extension and base for the portfolio