Data Analysis Lifecycle

Which aspects of the data analysis lifecycle are you primarily involved with on this project?

For this project I had to get involved with various stages of the data lifecycle. Firstly, I was given the objective - so what to do with the data. Secondly, I had to run queries to understand the data, such as finding out the column names and the size of the database, like amount of rows. Lastly, I had to do quite a bit of data cleaning, such as removing duplicates and changing the values in the queries to upper case.

What activities would you need to do before undertaking this project? Think about where the data came from.

My understanding of how I would collect the data needed for this project would be to use some sort of web scraping API. these is a lot of data in this dataset, so software would be more suited for this purpose. Once the data has been collected, I would use a reliable storage system, such as SQL, to hold all of this data reliably.

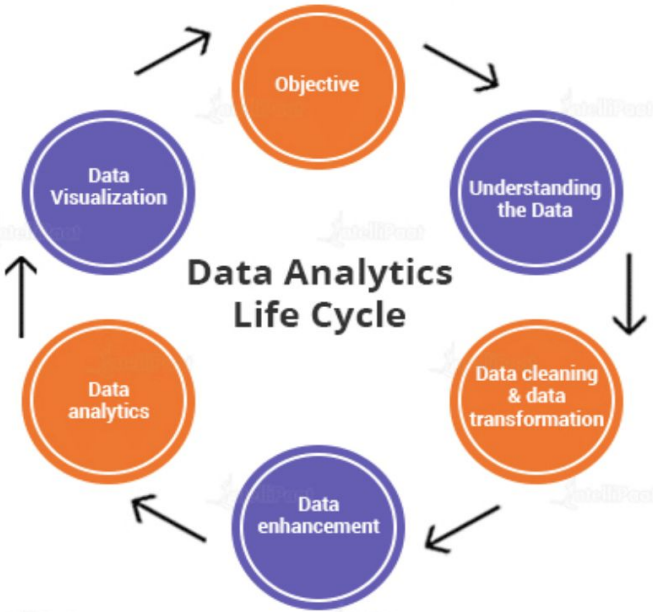
What would you need to do to allow this analysis to be repeated with updated data, and how would this solution be maintained?

Use a reliable storage system, set up an ETL (extract, transform, load) pipeline to keep on top of any new data to be added to the dataset, make sure this system is tested with duplicate data to make sure everything works. Lastly, find a way to reliably monitor this in a way that is almost semi-automated.

Data Analysis Lifecycle

How does this project fit into a broader data lifecycle?

This project is the glue between data extraction and data visualisation. The data extraction has probably been done with something like a web scraping API, which has then been stored once extracted. The data visualisation for this project would best be done in a tool such as Tableau, which allows for you to create interactive visualisations in order to drill down into any information required.



Requirements

What aspects of the requirements were not 100% clear to you?

Initially, some of the requirements weren't clear, but to complete this project I had to gain an understanding of all of them. The main example I can think of is the ERD, I struggle a bit to fully understand where and how to draw the links, but I believe that I have learned enough to create an accurate diagram for the purpose of this project.

Would it have been easier if you could talk directly to Oliver? If so, what sorts of questions would you want to ask him?

Definitely. If this was a real life scenario, I would ask for specific requirements, when points of the project are needed for, whether there would be time to arrange follow up meetings during the project, making sure everything I am doing is up to the standards required and so on…

How did you analyse the requirements?

I made sure to reference all of the queries to the data requirements sheet provided in the project brief. Once I was happy with the result of the query, I moved onto the next query.

Tools

Were the tools you used appropriate for the job?

Mostly, yes. I used Lucid charts for the Entity Relationship Diagram, which is a great visualisation tool for this purpose. To find out which types of cardinality to use, I found an image on Google, which shows the types of diagrams, I then chose which ones I felt were most appropriate for the purpose. For the coding, I used ChatGPT to do most of the heavy lifting. Once a code for the purpose I requested was generated, I made sure to read the code carefully to see if the results were fit for purpose, such as whether I felt the code could be shortened, check the spelling etc…I would also take time to try and read the code to further my understanding.

Why was SQLite used over PostgreSQL for this project?

SQLite is a serverless database designed for small scale operations in comparison to database sizes in general. A database that requires a server would incur extra costs and complications, which I feel is unnecessary for this project. Due to the lack of need for a complex queries in this database, it would be safe to assume that SQLite would be the most suitable tool for this project.

Tools

What are the benefits of adopting relational database technologies for an IT organisation?

There are many benefits of adopting relational database technologies to an IT organisation. One of the main benefits is security. Robust security features can be applied, such as encryption and authorisation. This is very useful considering some data held in these databases may come under GDPR issues if leaked. Relational databases are also scalable. Extra parts can be added to the data structure for future reference, and relational databases are able to store and function with lots of added data structures. Another very important part of relational databases are that they support backup and recovery mechanisms. This is crucial when working on projects as accidents can happen such as deleting or manipulating parts of the data that were meant to stay, so adopting backup and recovery mechanisms enable you to reduce overall risks, which I believe, is very important in any business.

Quality

What did you think of the organisation of the source data and how that mapped to the structure of your database? Was there a natural mapping from the CSV files to the database tables?

I would say so. Once the data tables were imported to the Jupyter Notebook and assigned a name, it was fairly easy to run the queries for each chosen table, as they were all assigned a unique, easy to remember name. What also helped was to have the tables open in Google Sheets. This helped me to find the column names in each table so that I could run the queries and also see what kind of results I would expect, such as integer or float.

Was the data consistent? In other words, were there any issues with the data that prevented you from producing good quality results?

For the purpose of the project, I would say there weren't really any issues. However, if I was to show this data to a stakeholder, I would want to use a data visualisation tool such as Tableau in order to make a real impact with the findings of the data.

Did you find the definitions of the data were detailed enough to assist you in the tasks?

Yes. The instructions in the Data Requirements sheet were very clear and easy to determine what was needed from all the instructions in each task.

Quality

What steps would you take to ensure that any concerns about data quality are dealt with appropriately? Who would you talk to about this?

I would set out the project in stages, similarly to how it has already been set out. Once each stage has been done, if possible, I would arrange a meeting with Melanie or Oliver discuss what I've done and whether the work I have conducted is up to the standards expected.

Security, Ethics and Legislation

Was the data sufficiently masked, or are there personal details present in the data?

There are a lot of personal details in this dataset. It would be wise to assume that all actions on a social media platform would constitute to personal details being gathered. Data points such as reputation and location are very strong for the purposes of targeted ads, cookies, etc…

Can you think of possible techniques (e.g. statistics) that could unmask the participants in this data?

I believe that statistics such as score and reputation, coupled with the profile image url, would be accurate enough to give a reliable representation of a user.

Are there any ethical considerations with our intended use of this data?

Of course. There is a lot of personal information held in these datasets, so we want to be careful with how this data is used. An important consideration would be to invest in a solid security system to minimise the chances of this information being leaked. Another important consideration would be to run some due diligence on all companies that ChatData are working with. In general, social media companies make their money from selling data, so we would not want that data being put into the wrong hands.

Security, Ethics and Legislation

Is the use of this data covered by any legislation, and if so, what is that legislation?

Yes, the use of this data comes under legislation such as GDPR, CCPA and the Data Protection Act, depending on where in the world a company is operating. In general, these legislations impose strict rules on how personal data is collected, processed, and protected, and they give individuals certain rights over their data. Each legislation has their own specific laws. A large social media company (such as Facebook) would typically operate internationally, that social media company would have to comply with all of these individual legislations in order to operate on an international scale.