**1. Introduction and motivations**

*A brief overview of our project and its purpose. We must mention the importance of understanding and comparing electronic product prices across different markets. Discuss the motivation behind the project, such as identifying pricing trends, market competitiveness, and potential cost savings for consumers.*

*Because we are planning on comparing sources from different countries, we must explain why we have chosen the ones we have chosen.*

**2. Research questions**

* How do electronic product prices vary across different geographical markets?
* What impact do promotions have on product prices in different regions?
* Can we identify patterns in consumer ratings and preferences based on geographical location?

**3. Sources**

* Electronic product lists (iPhone 14, 15, Samsung, etc.)
* Market reseller websites from three different countries / locations.
* Currency exchange rates

**3.1. Source 1 – reseller1.ch**

**3.2. Source 2 – reseller2.fr**

**3.3. Source 3 – reseller3.de**

**4. Context diagram**

3 Sources -> Extraction -> Transformation -> Storage

Source 1

reseller1.ch

*Jack Brown*

Source 2

reseller2.fr

*Ramon Burkhard*

Source 3

reseller3.de

*Daniel Herrera*

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Transformation**

Data cleaning  
Data enrichment

Python in PyCharm  
Packages: Pandas

**Merge of the data**

Based on model, storage, source, etc.

**Storage on MariaDB**

Merged data sets

Dataset 1

Dataset 2

Dataset 3

**5. Risks and Challenges**

We have outlined the proposed ETL (Extract, Transform, Load) workflow and undertaken a preliminary assessment of the websites chosen. It is foreseeable that challenges may arise during the scraping procedure. Since the main objective of our project is to collect data pertaining to the topics at hand, a solution in the event of unexpected issues would be to switch to a different source of data. A consequent additional challenge would be the identification of an adequate replacement website, which considering the broad topic of the project and the abundance of mobile phone resellers, should be straightforward.

Some additional challenges have been identified in the dynamic nature of these online resellers, with fluctuations of price data and ratings over time. Also, the consideration of foreign websites adds another layer of dynamic behaviour, with the variations in currency exchange data, and differences in pricing between countries. This is however a predictable circumstance, and partly the focus of our research questions.

As far as the characteristics of the data, naming conventions of variables and strings could differ within one dataset, as well as between sets from different portals and countries. With some thorough and systematic data cleaning, all three sources could be set to an equivalent standard of quality. This will also provide a significant aid for the next stage of the project, when all three sources are merged into one larger data set. An advisable approach would be to establish standardizations for variable naming and model-specific contents and rating codes, which will be elaborately described in a chapter of the final report. Only with a systematic review and standardization of the data can all three sources be acceptably combined and compared on equal terms.

**6. Reference list**

**7. List of tables and figures**