**Interim Project Report**

**Project Title:**

Real-Time Price Comparison Tool for Irish Supermarkets

**Project Definition:**

This project focuses on developing a web-based tool that allows consumers in Ireland to compare supermarket prices in real time. By using web scraping and data visualization, the tool will provide users with the latest price information from Tesco, Lidl, and Aldi, helping them make cost-effective shopping decisions.

The tool leverages technologies like Python for backend processing and Streamlit for the user interface. It will also explore how advancements in web scraping and dynamic data handling can improve digital shopping experiences.

**Area of Technology:**

This project applies web scraping, database management, and data visualization technologies:

* **Web Scraping:** Extracts price and product data directly from supermarket websites.
* **Data Processing:** Cleans and organizes the scraped data for display.
* **Data Visualization:** Creates a dashboard to help users quickly analyze price differences.

**Aims:**

1. Develop an easy-to-use platform to compare prices of groceries across Irish supermarkets.
2. Increase transparency for consumers by making it simple to identify price differences.
3. Showcase how web scraping and real-time data processing can address real-world challenges.

**Objectives:**

* Scrape essential data (product names, prices, categories, and unit sizes) from Tesco, Lidl, and Aldi websites.
* Store and organize the scraped data in a structured database.
* Build a user-friendly web interface to display and filter the data by product name, category, or store.
* Schedule regular updates to ensure the data remains current.
* Include basic visualizations, such as graphs to show price trends.

**Scope:**

This project is focused on groceries and household items, as they are frequently compared by Irish consumers. It excludes niche or luxury items and will only provide real-time price comparisons. Advanced features like historical data tracking or personalized shopping suggestions will not be included at this stage.

**Technologies Used:**

**1. Web Scraping:**

* **BeautifulSoup**: Extracts product data from static HTML pages.
* **Selenium**: Handles dynamic, JavaScript-rendered content.

**2. Data Storage:**

* **SQLite**: A lightweight database for storing structured data like product prices and categories.
* **MongoDB (optional)**: For unstructured data, if needed later.

**3. Frontend Development:**

* **Streamlit**: A Python framework for creating a visually appealing and interactive user interface.

**4. Backend Processing:**

* Python scripts will handle scraping, cleaning, and storing data. They will also refresh the data periodically.

**5. APIs (if available):**

* If supermarkets offer official APIs, they will be used for efficiency and accuracy.

**Research Topics:**

* Challenges in scraping dynamic websites with security measures like CAPTCHA.
* Ethical and legal considerations in web scraping, including compliance with supermarket policies.
* The impact of price transparency on consumer behavior in Ireland.

**Prototyping:**

**Initial Prototype:**

* Developed a basic Python script using BeautifulSoup to scrape product names and prices from Tesco’s website.

**Enhanced Prototype:**

* Incorporated Selenium to handle JavaScript-rendered pages.
* Stored the scraped data in SQLite and displayed it in a basic table using pandas.

**Key Testing Results:**

* Verified that prices and product details were accurately scraped.
* Tested scraping efficiency, ensuring it worked within reasonable time limits.

**Requirements:**

**Functional Requirements:**

1. Scrape key product details (name, price, unit size) from Tesco, Lidl, and Aldi.
2. Store the data in a structured format for easy querying.
3. Provide an interactive table where users can filter by product name, category, or store.
4. Allow users to search for specific products and view price differences instantly.

**Non-Functional Requirements:**

1. **Performance:** Data scraping and updates should complete within 15 minutes for all products.
2. **Compliance:** All scraping must adhere to legal and ethical standards (e.g., obeying robots.txt).
3. **Reliability:** The system should handle errors like layout changes or website downtime gracefully.

**Database Schema Design:**

The database design is based on three main tables:

1. **Products Table**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| product\_id | INTEGER (PK) | Unique product identifier. |
| product\_name | TEXT | Name of the product. |
| category | TEXT | Product category (e.g., Dairy, Snacks). |
| unit\_size | TEXT | Size or weight of the product (e.g., 500g, 1L). |
| store\_name | TEXT | Name of the supermarket (e.g., Tesco). |
| price | REAL | Current price of the product. |
| last\_updated | TIMESTAMP | Date and time of the last price update. |

1. **Prices Table**  
   | Column Name | Data Type | Description |  
   |-------------------|-------------|---------------------------------|  
   | price\_id | INTEGER (PK)| Unique price identifier. |  
   | product\_id | INTEGER (FK)| Links to the Products table. |  
   | store\_name | TEXT | Name of the supermarket. |  
   | price | REAL | Price of the product. |  
   | last\_updated | TIMESTAMP | Time of the last update. |
2. **Stores Table**  
   | Column Name | Data Type | Description |  
   |-------------------|-------------|---------------------------------|  
   | store\_id | INTEGER (PK)| Unique store identifier. |  
   | store\_name | TEXT | Name of the supermarket. |  
   | website\_url | TEXT | Supermarket website URL. |

**Visual Design:**

**Dashboard Layout:**

1. **Search Bar:**
   * Users can search for products (e.g., "Milk").
2. **Filter Panel:**
   * Dropdowns for category (e.g., Dairy, Vegetables), store name (Tesco, Lidl, Aldi), and unit size.
3. **Comparison Table:**
   * Displays product names, prices, unit sizes, and highlights the lowest price.
4. **Price Trends Visualization:**
   * A line graph showing price trends over time for selected products.

**Final Project Plan:**

**Phase 1: Data Collection (Dec 2024 – Jan 2025)**

* Develop Python scripts for scraping Tesco. Extend to Lidl and Aldi.

**Phase 2: Data Integration (Feb 2025)**

* Finalize the database schema and automate data cleaning/storage.

**Phase 3: Frontend Development (Mar 2025)**

* Use Streamlit to design a user-friendly interface.

**Phase 4: Testing & Deployment (Apr 2025)**

* Perform extensive testing and deploy the tool for public use.

**Conclusion:**

The project is progressing as planned, with initial prototypes successfully demonstrating core functionality. The next focus is on expanding the scraper to other supermarkets and refining the database integration. Once complete, this tool will make grocery shopping in Ireland more transparent and cost-effective for consumers.