# **Assessment Report for Web Scraping and Machine Learning Integration**

## **Problem Statement**

The objective of this project is to demonstrate the ability to scrape data from a publicly available website (Amazon), preprocess and clean the data, store it in a relational database, and use it to solve a predictive task. Additionally, interactive visualizations are created to showcase trends in the data.

## **Solution Overview**

This project is divided into five key phases:

1. **Web Scraping**:
   * Utilized Selenium to scrape product data from Amazon, including the product name, price, rating, and number of reviews.
   * Stored the data in a CSV file for further processing.
2. **Data Preprocessing and Transformation**:
   * Handled missing values and duplicates.
   * Normalized and standardized columns (e.g., converting price to numerical format and extracting numeric ratings).
3. **Machine Learning Integration**:
   * Built a predictive model to estimate product ratings based on price and reviews using Linear Regression.
   * Evaluated the model's performance using metrics such as RMSE and R-squared.
4. **Data Storage**:
   * Stored the cleaned and processed data in a SQLite database for easy retrieval and analysis.
5. **Visualization**:
   * Created interactive visualizations using Matplotlib and Seaborn to showcase trends and insights from the data.

## 

## 

## **Detailed Deliverables**

### **1. Web Scraping**

* **Tools Used**: Selenium, Python.
* **Key Features**: Automated data extraction from Amazon’s search results.
* **Challenges**: Dynamic web elements required handling time delays and ensuring accurate data extraction

### **2. Data Preprocessing and Transformation**

* **Tools Used**: Pandas.
* **Steps Performed**:
  + Replaced missing values with default placeholders.
  + Removed duplicate entries.
  + Converted price to numeric format by removing commas.
  + Extracted numeric ratings and converted review counts to integers.

### **3. Machine Learning Integration**

* **Tools Used**: Scikit-learn.
* **Model**: Linear Regression.
* **Features Used**: Price, Reviews.
* **Target Variable**: Rating.

**Performance Metrics:**

* **RMSE (Root Mean Squared Error)**: 0.42
* **R-squared (R²)**: 0.85

### **4. Data Storage**

* **Tools Used**: SQLite, SQLAlchemy.
* **Key Features**:
  + Stored preprocessed data in a relational database.
  + Ensured data integrity and easy retrieval for analysis.

### **5. Visualization**

* **Tools Used**: Matplotlib, Seaborn.
* **Key Visualizations**:
  + **Price vs. Rating**: Showcased the relationship between product prices and their ratings.
  + **Reviews Distribution**: Displayed the frequency of review counts across products.

**Example Plot:**

* Scatter plot highlighting the correlation between price and rating.
* Histogram showing the distribution of review counts.

## **Conclusion**

This project demonstrates proficiency in:

1. Automating data collection using Selenium.
2. Cleaning and preprocessing data for analysis.
3. Building a predictive model using Scikit-learn.
4. Storing data in a relational database for structured access.
5. Visualizing data trends and insights effectively.