**Project Documentation**

**( Salesforce AppExchange Data Scraping )**

## 

### **1. Project Overview**

**Objective**

The goal of this project is to scrape data from Salesforce AppExchange, focusing on multiple categories such as Business Needs, Products, Industries, Salesforce Labs, and Impact. Each category has various filters that need to be applied to ensure a comprehensive list of applications is extracted. And it has 3 sections as **Overview, Reviews** and **More Details.**

**Scope**

* **Phase 1:** Scrape data from various categories (Business Needs, Products, Industries, Salesforce Labs, Impact and Agentforce) using appropriate filters for each category.
* **Generate unique IDs** for each entry for easy identification and tracking so the initial template will contains Unique ID, App name and Link
* **Phase 2:** Navigate to each application’s link and extract the main profile content which contains App name, Company name, Average Ratings, Reviews, Price, Industries, Business Needs, Requires, Compatible with and App Description. At the same time it already shows Overview section details in which we need to extract details such as Company City, State, Country, Web Address, Social Impact, Email, Phone, Company description and Pricing Plan.
* **Phase 3:** After completion of extraction process of main profile and Overview section it should need to redirect to More Details section and in that section the extraction process will proceed for these details like Compatibility Requires, Compatible with, Salesforce Edition, Additional Requirements, App Version, First Release, Latest Release, Supported Features, Package Contents (Custom Objects, Custom Tabs, Custom Apps) , Lightning Components (Global, App Builder, Community Builder) and Languages.
* **Phase 4:** For Reviews data extraction we will used the same template which contains Unique ID, App name and Link so to directly navigate the Review section and extracting details such as Reviewer Name, Valence, Date, Review text, Likes and Badges Count such as MVP, Ranger and Top Review

### **2. Data Collection Process:**

**Step 1: Extracting Application List (Business needs)**

* Navigated to **Salesforce AppExchange Business Needs Category**.
* Selected all business filters to generate a **complete list of applications**.
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted **4,993** unique applications with their respective names and links.
* Stored this information in an Excel file (salesforce\_apps.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1tuuszd4S5M4Jr6WLOUY4RE5kkvrsTZm8/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 2: Extracting Application List (Products)**

* Navigated to **Salesforce AppExchange Products Category**.
* Selected radio button filters creates url and navigating that url of each filters
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted all available applications results of each radio button generated url with their respective names and links.
* Stored this information in an Excel file (salesforce\_all\_products.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1l_wOeEoMfz1wGIzbGV6Z6YERxguFNYAI/edit?usp=sharing&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 3: Extracting Application List (Industries)**

* Navigated to **Salesforce AppExchange Industries Category**.
* Selected all industries filters to generate a **complete list of applications**.
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted all available unique applications results with their respective names and links.
* Stored this information in an Excel file (salesforce\_industries\_apps\_data.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1dQB3ayVOxEqmTIbUmKnrUVB0JtJlvFHk/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 4: Extracting Application List (Salesforce Labs)**

* Navigated to **Salesforce AppExchange Salesforce Labs Category**.
* Selected all salesforce labs filters to generate a **complete list of applications**.
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted all available results.
* Stored this information in an Excel file (salesforce\_labs\_apps\_data.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1DTl8QnAaSFcLoLsUNK9EgcMeBKzO6yXm/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 5: Extracting Application List (Impact)**

* Navigated to **Salesforce AppExchange Impact Category**.
* Selected all impact filters to generate a **complete list of applications**.
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted all available unique applications results with their respective names and links.
* Stored this information in an Excel file (salesforce\_impact\_apps\_data.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/16NZfUrSyvKYccDHCH7rpd2Y1r_Fp_1AE/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 6: Extracting Application List (Agentforce)**

* Navigated to **Salesforce AppExchange Agentforce search term Category**.
* Searching “**Agentforce**” string and generating a **complete list of applications**.
* Handle properly the **Shadow DOM** and multiple slot elements.
* Successfully extracted all available unique applications results with their respective names and links.
* Stored this information in an Excel file (salesforce\_impact\_apps\_data.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1eQSRVMVcrPBpMnshMzUo3XEMlw42SWKL/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 7: Extracting Profile info, summary and overview section data**

* Navigated to **each link and properly extracted data by tags and element**
* Handle properly the **Shadow DOM** and multiple **slot** elements.
* Successfully extracted all available variables data.
* Stored this information in an Excel file (All Data.xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1VW5lYR-8v1ULLgsXacP8rvsIQxBt4OIO/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 8: Extracting More Details section data**

* Navigated to **each app link properly and again navigate to “More Details” section link and extracted data by tags and element**
* Handle properly the **Shadow DOM** and multiple **slot** elements.
* Successfully extracted all available variables data in the **More Details** section.
* Stored this information in an Excel file along with overview data (All Data (1).xlsx).
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1YBlC7tIkJE2k9OjllhjPfQwDBnmL63Bc/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

**Step 9: Extracting Reviews section data**

* Navigated to each apps list review section link directly by appending **“&tab=r”**
* Handle properly the **Shadow DOM** and multiple **slot** elements.
* Successfully extracted all available variables data in the **Reviews** section.
* Stored this information in an Excel file (Extracted Reviews.xlsx)
* Preview of the Data [Link](https://docs.google.com/spreadsheets/d/1sXrz0njfxopczwLq2Fx_qEfJYNm_UC6e/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

### **3. Implementation Details:**

**Technology Stack**

* **Python**: Core scripting language.
* **Selenium WebDriver**: For automating browser interactions.
* **Pandas**: For data handling and Excel file operations.
* **Chrome WebDriver**: Used to interact with the AppExchange pages.
* **Jupyter Notebook:** For interactive development and debugging.
* **Colab Notebook:** For running and testing scripts in a cloud-based environment.

### 4. Results & Achievements:

* Successfully extracted **Business Needs, Products, Industries, Salesforce Labs, Impact and Agentforce categories data** with their unique IDs, names, and links.
* Stored the results in in their respective destination xlsx file
* Successfully extracted Profile top section which contains **Company names, Average Ratings, Reviews, Pricing and App Description** variables
* Also extracted Profile Middle section which contains **Industries, Business Need, Requires and Compatible With**  Variables
* The **Overview** and **More Details** section have been successfully extracted and navigated to the appropriate variables
* Successfully navigated to review section and extracted data from appropriate variables
* Here is preview extract data of [Overview](https://docs.google.com/spreadsheets/d/1vHmKAZZHb2HbCeuRs3mM9PF5QPr_3DfO/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true), [More Details](https://docs.google.com/spreadsheets/d/1YBlC7tIkJE2k9OjllhjPfQwDBnmL63Bc/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true) and [Reviews](https://docs.google.com/spreadsheets/d/1sXrz0njfxopczwLq2Fx_qEfJYNm_UC6e/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

### 5. Implemented Source Codes:

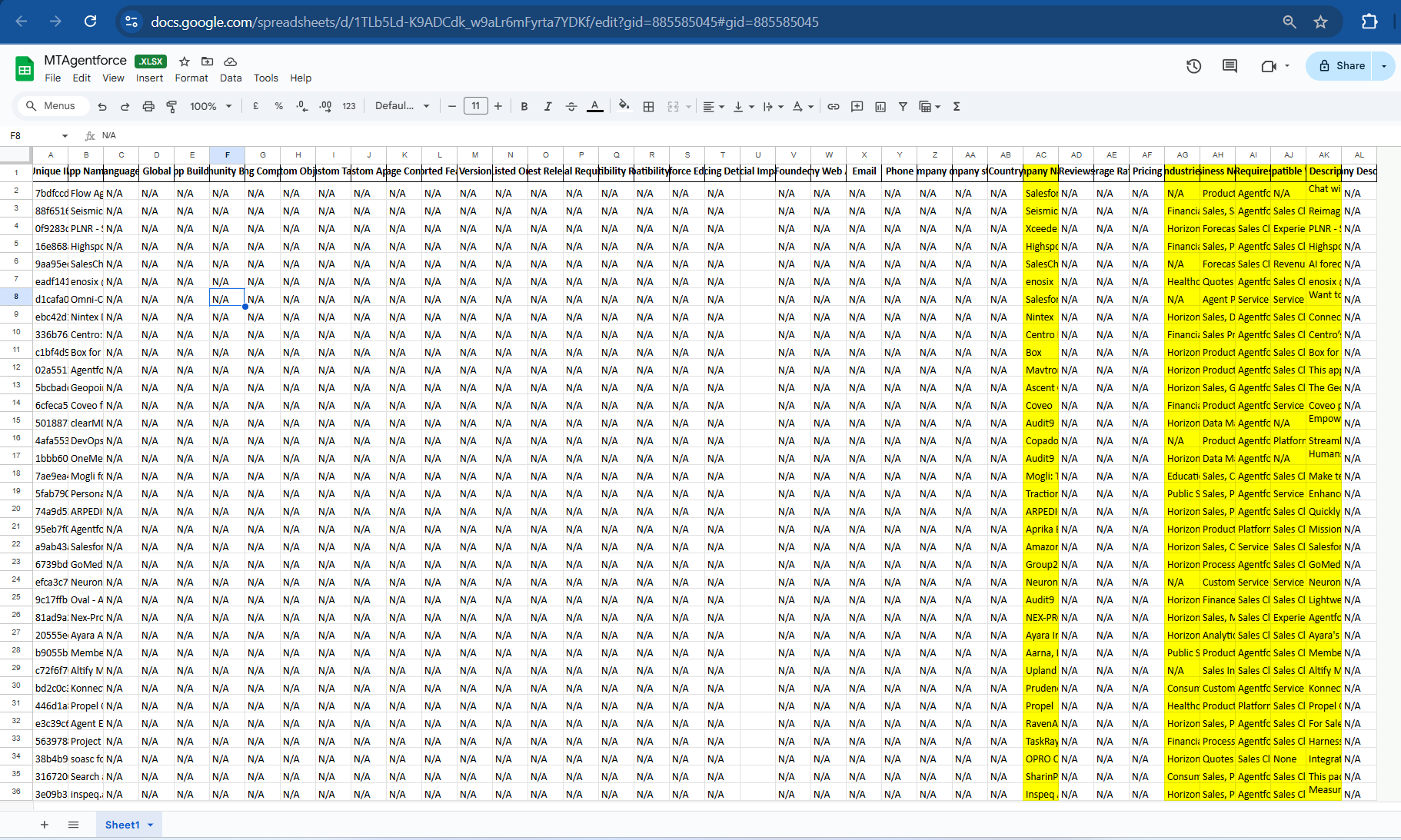
* **Business Needs -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Business%20Needs)
* **Products -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Products)
* **Industries -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Industries)
* **Salesforce Labs -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Salesforce%20Labs)
* **Impact -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Impact)
* **Agentforce -** [**Link**](https://github.com/AjayBidyarthy/Chris-Forman-Data-Management-Salesforce/tree/main/Agentforce)

### 6. Extracted Data and Review List:

* **Business Needs:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1w2gHLYLCcJ8SX-xLvuiITUiaC3ZHJ5d-/edit?usp=sharing&ouid=108758494584341020275&rtpof=true&sd=true), [Reviews](https://docs.google.com/spreadsheets/d/1_zl0ZbokKXlbkxMka9zSiIhK7PVW8JL7/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **Products:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1Kp28O5YczIt7hJV_qDE_3TCZSTjnu7Cj/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/1MM6wc6QP3NbbmsVLgyHWrX6BtlH272K5/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **Industries:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1nnF_e9pZqfgWcUD1JMcLCmc8qmxetY0X/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/12sYR4zKAqqBc-lR_UZcB8XWLlX1lecTA/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **Salesforce Labs:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1u9kvSXgNc0GBbBU0r9PGClPD3FoRChYo/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/1uVPYPvqzpz7UNGumBwRV9UaiF0eET6wT/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **Impact:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/18wHxHe0XYLcVBBLdqqEXZVo-Gyw4-ycE/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/1Uyzdt-QA-WlcUKzAJJxDVjzFbKfg7PRE/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **Agentforce:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1qnQxgc_vszJzS-jIsGPbEIAqqOQn_sLN/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/1Fq7e6y047jc1dTfbvaJmPsTXCOXdz80R/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)
* **All Combined:** [Extracted\_Data](https://docs.google.com/spreadsheets/d/1WZunrJAz7QGlno5Gm2n91xI8A7HTNuWA/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)**,** [Reviews](https://docs.google.com/spreadsheets/d/1izpsYKJUOzsv1hQy8OdDIom2msBjlhjG/edit?usp=drive_link&ouid=108758494584341020275&rtpof=true&sd=true)

### **7. Maintenance of the selenium script**

We need to perform some maintenance work in the website because there have been recent structural changes in the Salesforce AppExchange website—such as modifications in tag names, the addition of slot elements, and an extra shadow-root layer in the DOM. Due to these changes, our current script is unable to properly locate the required elements. For this we need to update and modify the script to accommodate these changes.

Here is the recent preview of extracted data which unable to locate it’s tag element before the recent changes in the webelement structure and hence most of the columns are missed and filled with **“N/A”**

To avoid this issue we need to update in selenium script code blocks frequently where it properly navigates the link and locates appropriate web elements and tag elements and extract reviews and metadata properly