🗂️ Project Requirements & Deliverables Document

Project Title:
Certified Real Estate Valuer Directory Scraping for Germany

🎯 Project Objective

To build a robust, reusable, and script-based web scraping solution to extract certified real estate valuer information from six public directories in Germany. The solution must provide clean, deduplicated, and structured data exports in `.xlsx` or `.csv` format, along with clear documentation for future use.

🌐 Target Data Sources (Total: 6 Websites)

1. **HypZert Gutachtersuche (http://www.hypzert.de/de/service/gutachtersuche)**

2. **DIACONSULTING Zertifizierungsverzeichnis (http://www.diaconsulting.de/de/16/Zertifizierungsverzeichnis)**

3. **Sprengnetter Sachverständigenverzeichnis (http://www.sprengnetter.de/gutachter-sachverstaendigenverzeichnis/)**

4. **IQ-Zert Personen Wertermittlung (http://www.iq-zert.de/verzeichnis)**

5. **BVS Sachverständige (Postal Code Based Search)**

   http://www.bvs-ev.de/sachverstaendigensuche

   - Search by PLZ with a **radius of 250 km**
   - **Postal Codes to use:**
     `10115, 20095, 50667, 70173, 80331, 01067, 90402, 04109, 28195, 65183, 39104, 99084, 24103, 66111, 45127, 79539, 02826`

6. **IHK Sachverständigenverzeichnis (Postal Code Based Search)**

http://svv.ihk.de

- Search by PLZ with a **radius of 100 km**
- **Postal Codes to use:**
  `01067, 04109, 06108, 10115, 14467, 19053, 20095, 23552, 24103, 26122, 28195, 30159, 33602, 34117, 37073, 39104, 44787, 47051, 50667, 53111, 56068, 60311, 66111, 68159, 70173, 72070, 74072, 80331, 86150, 89073, 90403, 93047, 94032, 96450, 99084`

🗒️ Required Fields to Extract per Entry

Each scraper will extract the following data (if publicly available):

- Full Name
- Address / City / Region
- Phone Number *(optional)*
- Email Address *(optional)*
- Website *(optional)*
- Certification Type / Details *(optional)*
- **Source Directory Name** *(tagged per record)*

💻 Technical Requirements

- **Reusable Python Scripts** using `requests`, `BeautifulSoup`, `Selenium`, etc.
- **Automatable**: Compatible with CRON or Windows Task Scheduler

- **Output Formats**: `.csv` and `.xlsx`
- **De-duplication**: Remove duplicate entries across all sources
- **Directory Tagging**: Field indicating source directory per record
- **Structured Output**: Tabular format with unified headers
- **Documentation**: `README.md` with clear usage instructions for non-developers

---

📦 Deliverables

✅ Six Python-based scraping scripts (1 per directory)
✅ Final combined and deduplicated dataset in `.xlsx` and `.csv` formats
✅ Tagging per record with the source site name
✅ Full source code with comments and folder structure
✅ `README.md` file for usage and maintenance
✅ *(Optional)*: Google-style basic search integration for public certs (pending further discussion)

🕑 Timeline

- **Completion**: Within 5 business days