**1. Types of Data to Collect**

We want to gather a wide range of datasets that can reflect the most common and sought-after data science skills. Here's a list of data sources and types:

**a. Job Postings Data**

* **Skills in demand:** Analyze the skills mentioned in data science job postings. This can help identify what companies are actively seeking.
* **Where to find it:**
  + **Job boards:** Websites like LinkedIn, Indeed, Glassdoor, and Kaggle.
  + **APIs:** You can use APIs from platforms like LinkedIn, Glassdoor, or Indeed to extract job listings.
  + **Datasets:** There are existing datasets like Indeed Job Postings on Kaggle that you can use.

**b. Resume Data**

* **Skills mentioned in resumes:** Identify the skills that data scientists tend to list on resumes, which can indicate demand in the job market.
* **Where to find it:**
  + **Online resume databases:** Sites like LinkedIn and Indeed might provide aggregated skills data.
  + **Data sharing platforms:** Kaggle has datasets like “Data Scientist Resume Dataset” which might have useful data.
  + **Scraping platforms:** If allowed, scraping platforms such as GitHub, LinkedIn, or even specialized resume sharing websites can help gather this data.

**c. Survey Data (Industry Reports)**

* **What data scientists are learning and using:** Gather data from surveys or industry reports that assess the popularity of various tools, techniques, and technologies.
* **Where to find it:**
  + **Kaggle surveys:** Kaggle conducts regular surveys on the state of data science (e.g., the Kaggle Data Science Survey).
  + **Stack Overflow Developer Surveys:** Stack Overflow surveys are valuable for understanding the tools and technologies that developers and data scientists are using.
  + **Annual reports:** Reports from organizations like O'Reilly, DataCamp, or Data Science Central may provide insight into industry trends and popular skills.

**d. College and University Course Catalogs**

* **Skills taught in formal education:** Understanding what skills are being taught in data science programs can help you determine the foundational skills that are being valued.
* **Where to find it:**
  + **University websites:** College and university course catalogs (e.g., MIT OpenCourseWare, Stanford, Harvard, etc.) often provide a list of courses.
  + **Online learning platforms:** Websites like Coursera, edX, Udacity, and DataCamp list popular data science courses and their contents. They also provide data on trends in skills learned.
  + **Course reviews and forums:** Sites like Reddit, Quora, and course reviews can provide feedback on which skills are most sought-after or taught in popular courses.

**e. Technical Blogs, Publications, and Research Papers**

* **Emerging skills in the industry:** Blogs and technical publications often highlight trending topics, new technologies, and methodologies.
* **Where to find it:**
  + **Blogs:** Sites like Towards Data Science, Medium, Analytics Vidhya, and Data Science Central.
  + **Academic publications:** Google Scholar, arXiv, or major conferences like NeurIPS and ICML can provide papers on new techniques and methodologies that are gaining popularity.

**f. Social Media (LinkedIn, Twitter, Reddit, etc.)**

* **Trending discussions about skills:** Social media can show what skills are being discussed and promoted in the data science community.
* **Where to find it:**
  + **LinkedIn:** Look at top skills in data science profiles, popular job postings, and trending articles.
  + **Reddit:** Subreddits like r/datascience, r/machinelearning, and r/statistics have discussions on skills, courses, and job trends.
  + **Twitter:** Follow leading data scientists, companies, and hashtags like #DataScience, #AI, #MachineLearning to see what’s trending.

**g. Freelance Platforms (Upwork, Fiverr, Toptal, etc.)**

* **Skills in demand for freelancers:** Freelance platforms often list the skills clients are looking for, which can help show demand for specific data science competencies.
* **Where to find it:**
  + **Freelance job postings:** You can use APIs or manually collect job posts from platforms like Upwork, Fiverr, and Freelancer.
  + **Skills & Tags:** These platforms often categorize jobs by specific skills (e.g., Python, TensorFlow, SQL), and you can analyze the frequency of these tags.

**h. Technical Skill Competitions and Kaggle**

* **Popular tools and techniques:** Kaggle competitions reveal which skills and techniques are being used to win data science challenges.
* **Where to find it:**
  + **Kaggle Datasets and Competitions:** Look at the most recent competitions and the techniques/data science tools mentioned.
  + **Kaggle Discussions and Kernels:** Read through Kaggle discussions to understand what skills and tools the community emphasizes.

**2. Data to Collect**

Once you have the data sources, here are some specific fields you might want to collect from each source:

* **For job postings**: Skill tags (e.g., Python, SQL, ML frameworks), job titles, experience level, location, and industry.
* **For resume data**: Listed skills (programming languages, data manipulation, visualization tools, ML models, etc.), job titles, and experience level.
* **For survey data**: Skill popularity rankings, tools used in the field, certification trends, and educational background.
* **For college/university courses**: Course names, topics covered, prerequisite skills, and course ratings.
* **For freelance platforms**: Job titles, required skills, project duration, and compensation rates.
* **For social media discussions**: Keywords or hashtags, mentions of specific tools, courses, or techniques.

**3. Methods for Data Collection**

* **APIs:** Many platforms (LinkedIn, Indeed, Kaggle, etc.) provide APIs to extract job postings, profile data, or other relevant information.
* **Scraping:** If no API is available, scraping websites like LinkedIn, Glassdoor, and Upwork could be useful (but make sure to follow each platform's terms of service).
* **Manual search and downloads:** Downloading pre-existing datasets from Kaggle or research papers might save time.

**4. Analysis Methods**

* **Data Cleaning and Preprocessing:** Remove duplicates, missing values, or irrelevant information.
* **Skill Frequency Analysis:** Count the frequency of certain skills appearing across datasets (e.g., Python, SQL, Machine Learning).
* **Trend Analysis:** Track how the demand for certain skills changes over time (e.g., in job postings, surveys, etc.).
* **Natural Language Processing (NLP):** Use NLP techniques to extract skills mentioned in job postings, resumes, or surveys automatically.

**5. Tools to Use**

* **Python:** Libraries like pandas for data manipulation, matplotlib/seaborn for visualizations, and BeautifulSoup/Scrapy for web scraping.
* **SQL Databases:** For storing large datasets like job postings, resume data, etc.
* **APIs and Web Scraping:** Requests and Scrapy for collecting data from websites.
* **Text Mining/NLP:** Use libraries like NLTK or spaCy for analyzing unstructured text data from job postings or resumes.

**6. Where to Start**

To get started, I suggest focusing on a few core datasets:

* Kaggle’s data science survey data.
* Job postings data from LinkedIn or Indeed.
* Resume datasets from Kaggle or LinkedIn.
* Kaggle competitions to spot in-demand techniques.