In [36]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import re
import warnings
from wordcloud import WordCloud
import nltk
from nltk.corpus import stopwords
from collections import Counter
```

In [37]:

```
# Global settings
warnings.filterwarnings('ignore', category=FutureWarning)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_rows', None)
```

=== 01. Data Pre-processing ===

Load the survey result downloaded from google form

In [38]:

```
file_path = 'data/data_03_22_2024.csv'
survey_res = pd.read_csv(file_path)
```

In [39]:

```
# Number of respondents
len(survey_res)
```

Out[39]:

```
98
```

Rename survey questions

In [40]:

```
survey_res.columns.tolist()
```

Out[40]:

```
['Timestamp',
 'Have you engaged in data-related work (such as data analysis, data engineeri
ng, or data science) as part of your job?',
 'When do you typically search for datasets? (select all that apply)',
 'Where do you typically look for datasets? (select all that apply)',
 'Could you specify and describe the tools you use for data searching?',
 'How do you usually find the correct dataset for your needs? (select all that
apply)',
 'Please briefly describe your approach when using the methods selected for fi
nding datasets. For instance, if you chose "Consultation with coworkers or exp
erts", what would you ask for?',
 'What content-related metadata do you find useful in locating relevant datase
ts? (select all that apply)',
 'What table-related metadata do you find useful in locating relevant dataset
s? (select all that apply)',
 'Imagine you had an ideal dataset search system, can you give an example quer
y (the query can be in natural language, doesn’t have to be SQL) that you woul
d like to find relevant datasets for?',
 'What challenges do you face in finding relevant datasets? (select all that a
pply)',
 "Could you provide a specific example about the challenges you've selected ab
ove?",
 'What is your current position?',
 'How many years of experience do you have working with data?',
 'Please specify your industry or the organization you work for.',
 'If you are interested in participating in further studies, please leave your
name and email. (Your personal information will be kept confidential and used
solely for research purposes.)']
```

In [41]:

```python
# Creating a dictionary for renaming columns
rename_dict = {
    'Have you engaged in data-related work (such as data analysis, data enginee
    'When do you typically search for datasets? (select all that apply)': 'sea
    'Where do you typically look for datasets? (select all that apply)': 'searc
    'Could you specify and describe the tools you use for data searching?': 'se
    'How do you usually find the correct dataset for your needs? (select all th
    'Please briefly describe your approach when using the methods selected for 
    'What content-related metadata do you find useful in locating relevant data
    'What table-related metadata do you find useful in locating relevant datase
    'Imagine you had an ideal dataset search system, can you give an example qu
    'What challenges do you face in finding relevant datasets? (select all that
    "Could you provide a specific example about the challenges you've selected 
    'What is your current position?': 'position',
    'How many years of experience do you have working with data?': 'year_of_exp
    'Please specify your industry or the organization you work for.': 'industry
    'If you are interested in participating in further studies, please leave yo
}

# Renaming the columns
survey_res.rename(columns=rename_dict, inplace=True)

survey_res.columns.tolist()
```

Out[41]:

```
['Timestamp',
 'is_data_worker',
 'search_purpose',
 'search_location',
 'search_tool',
 'data_discover_methods',
 'data_discover_methods_text',
 'content_metadata',
 'table_metadata',
 'ideal_query_example',
 'data_discover_challenges',
 'data_discover_challenges_text',
 'position',
 'year_of_experience',
 'industry',
 'participation_willingness']
```

Filter responses from those who indicated "Yes" to engaging in data-related work

In [42]:

```
survey_res = survey_res[survey_res['is_data_worker'] == 'Yes']

len(survey_res)
```

Out[42]:

```
95
```

Filter responses from researchers who are not our system's target user

In [43]:

```
survey_res_non_researcher = survey_res[(survey_res['position'] != 'Researcher'
                                        & (~survey_res['position'].str.contains
                                        & (~survey_res['position'].str.contains
                                       ]

len(survey_res_non_researcher)
```

Out[43]:

```
69
```

In [44]:

```
pd.DataFrame(survey_res_non_researcher['position'].unique())
```

Out[44]:

	0
0	Data analyst / data scientist
1	Machine learning engineer
2	Data Analyst Manager
3	Data Governance Lead and Data Catalog Product Owner
4	Finance and analytics head
5	Software developer
6	Data engineer
7	Software Engineer
8	Director
9	Management of data science group
10	Associate Director Career Services
11	Tech Lead (I wear many hats)
12	Applied DS&AI student, aspiring ML engineer
13	Manager
14	Analytics leader
15	Unemployed
16	Management Accountant (data geek)
17	Business analyst
18	BI developer
19	retired
20	Master Student

=== 02. Functions Handling Different Types of Questions Analysis ===

Helper functions

- ***Single choice barchart plot***

In [45]:
```
# Visualize the distribution of SINGLE-choice questions
def plot_single_choice_distribution(data, column_name, title, palette='cubehel
    # Get the distribution of the column
    distribution_data = data[column_name].value_counts()

    # Plot the distribution in barplot
    plt.figure(figsize=(10, 6))
    ax = sns.barplot(y=distribution_data.index, x=distribution_data.values, pa
    plt.title(title)

    # Correct percentage calculation
    total = distribution_data.sum()
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_width() / total)
        x = p.get_x() + p.get_width() + 0.02  # Shift the text to the right si
        y = p.get_y() + p.get_height() / 2
        ax.annotate(percentage, (x, y))

    plt.xlabel('Count')
    plt.ylabel(column_name)
    plt.show()
```

In [46]:

```
# Plot distributions for multiple single-choice questions in batch
def batch_plot_single_choice(data, column_names, titles=None, palette='cubehel
    if titles is None:
        titles = column_names

    # Ensure the titles list matches the length of the column_names list
    assert len(titles) == len(column_names)

    for column_name, title in zip(column_names, titles):
        plot_single_choice_distribution(data, column_name, title, palette)
```

- ***Multiple choice barchart plot***

In [48]:

```
# Plot the distribution of multi-choice question responses, handling predefine
def plot_multi_choice_distribution(data, column_name, predefined_options, titl
    # Initialize a counter for the choices
    choice_counts = Counter()

    # Iterate through each response, handling predefined options
    for response in data[column_name].dropna():
        # Split response into parts based on semicolon, handle each as a poten
        response_parts = [part.strip().lower() for part in response.split(';')

        # Track parts of the response to identify if it's a predefined option
        response_processed = []

        for part in response_parts:
            for option in predefined_options:
                option_lower = option.lower()
                # If the part contains the predefined option
                if option_lower in part:
                    choice_counts[option] += 1
                    response_processed.append(option_lower)
                    break
            else:
                # If the part is not one of the predefined options, count it a
                # Capitalize the first letter of each word for "other" respons
                other_label = 'Other: ' + ' '.join(word.capitalize() for word
                choice_counts[other_label] += 1

    # Convert counter to Series for plotting
    choice_series = pd.Series(choice_counts).sort_values(ascending=False)

    # Plotting
    plt.figure(figsize=(10, 6))
    ax = sns.barplot(x=choice_series.values, y=choice_series.index, palette=pa
    plt.title(title)
    plt.xlabel('Count')
    plt.ylabel(column_name)

    # Adding percentage annotations
    total = len(data[column_name].dropna())
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_width() / total)
        x = p.get_x() + p.get_width() + 0.02  # Shift the text to the right si
        y = p.get_y() + p.get_height() / 2
        ax.annotate(percentage, (x, y))

    plt.show()
```

In [70]:

```
# Adjusted multi-choice question distribution function to combine occurrences
# Mainly for the question "What content-related metadata do you find useful in
# The two options "Granularity of specific columns (e.g., geographic or tempora
# & "Level of detail in the dataset (e.g., item level vs. category level in pro
# were combined into one "Granularity of specific columns (e.g., product level,

def plot_combined_options_distribution(data, column_name, predefined_options,
    # Initialize a counter for the choices
    choice_counts = Counter()

    # Iterate through each response, handling predefined options and combining
    for response in data[column_name].dropna():
        # Split response into parts based on semicolon, handle each as a poten
        response_parts = [part.strip().lower() for part in response.split(';')

        for part in response_parts:
            combined = False
            # Check if the part matches any of the options to be combined
            for option in combined_options:
                if option.lower() in part:
                    # Increment the count for the updated combined option
                    choice_counts[combined_options[-1]] += 1
                    combined = True
                    break
            if not combined:
                # If it doesn't match the combined options, count it as is or
                for option in predefined_options:
                    if option.lower() in part:
                        choice_counts[option] += 1
                        break
                else:
                    # This part is considered an "other" response
                    other_label = 'Other: ' + ' '.join(word.capitalize() for w
                    choice_counts[other_label] += 1

    # Convert counter to Series for plotting
    choice_series = pd.Series(choice_counts).sort_values(ascending=False)

    # Plotting
    plt.figure(figsize=(10, 6))
    ax = sns.barplot(x=choice_series.values, y=choice_series.index, palette=pa
    plt.title(title)
    plt.xlabel('Count')
    plt.ylabel(column_name)

    # Adding percentage annotations
    total = len(data[column_name].dropna())
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_width() / total)
        x = p.get_x() + p.get_width() + 0.02  # Shift the text to the right sid
        y = p.get_y() + p.get_height() / 2
        ax.annotate(percentage, (x, y))

    plt.show()
```

In [49]:

```
# Plot distributions for multiple multi-choice questions in batch
def batch_plot_multi_choice(data, column_options_dict, titles=None, palette='cu
    if titles is None:
        titles = list(column_options_dict.keys())

    # Ensure the titles list matches the length of the column_options_dict keys
    assert len(titles) == len(column_options_dict)

    for column_name, predefined_options in column_options_dict.items():
        # Find the title for the current column, defaulting to the column name
        title = titles[list(column_options_dict.keys()).index(column_name)]
        # Call the plot_multi_choice_distribution function for each column
        plot_multi_choice_distribution(data, column_name, predefined_options, t
```

- ***Short answer questions wordcloud***

In [50]:

```
def plot_word_cloud(data, column_name):
    # Combine all responses into one large text string
    text = " ".join(response for response in data[column_name].dropna())

    # Generate a set of stopwords
    stop_words = set(stopwords.words('english'))

    # Create the word cloud object, setting the stopwords to the nltk stopwords
    wordcloud = WordCloud(stopwords=stop_words, background_color='white', widtl

    # Display the word cloud using matplotlib
    plt.figure(figsize=(15, 10))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')  # Hide the axes
    plt.show()
```

=== 03. Results Visualization ===

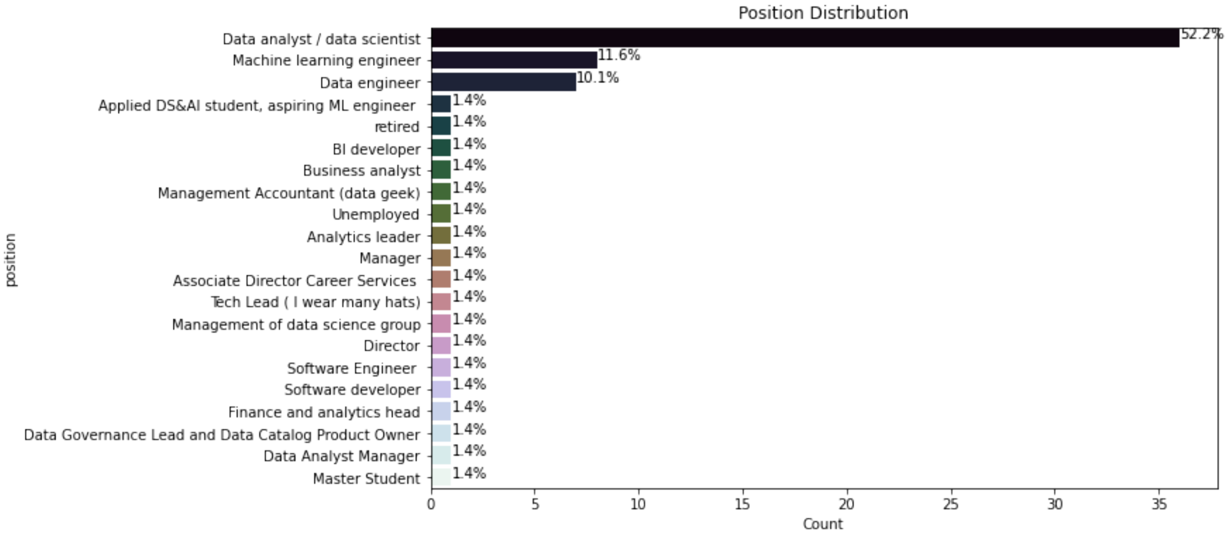
Respondents' demographics

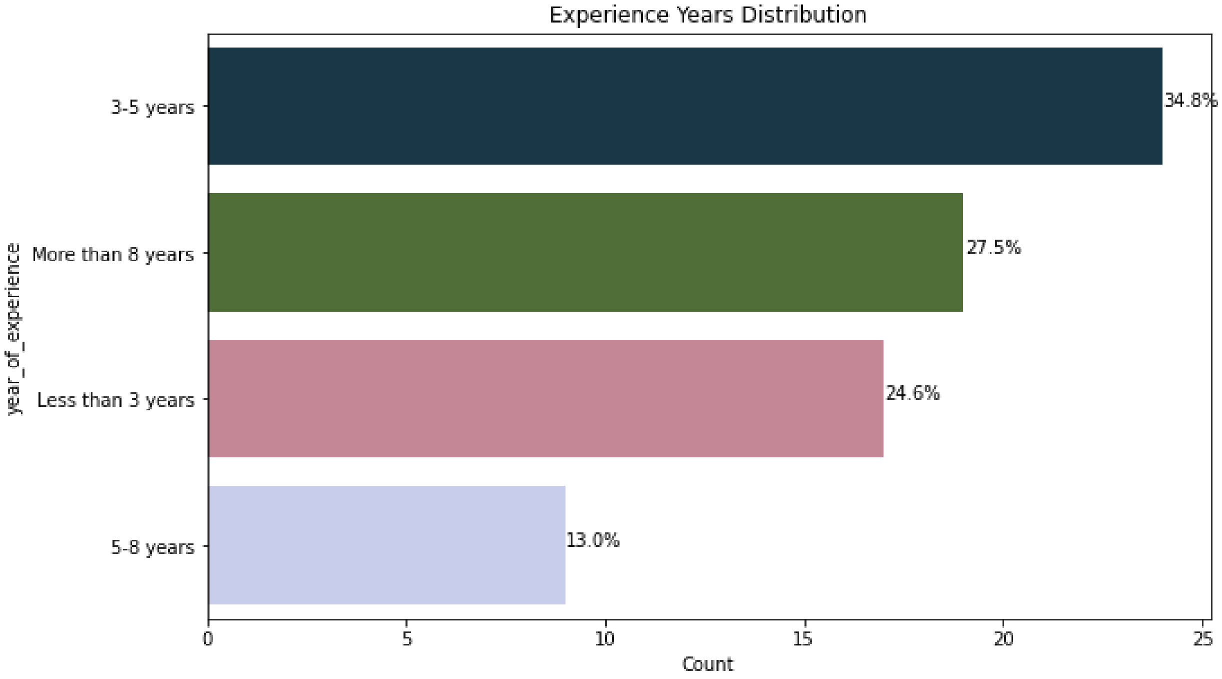
In [51]:

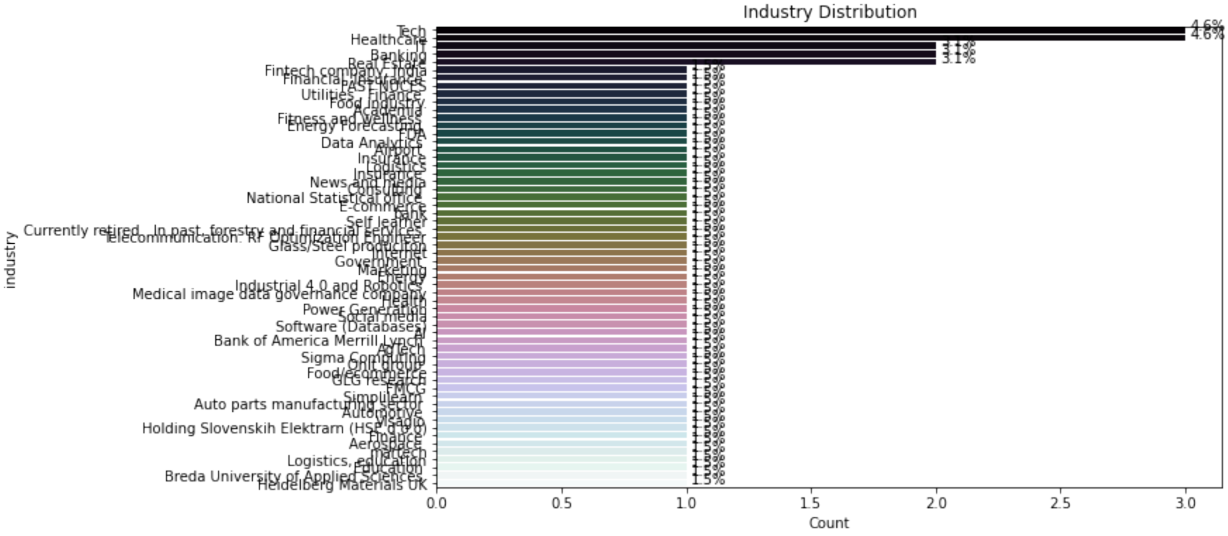
```
column_names_demo = ['position', 'year_of_experience', 'industry']
titles_demo = ['Position Distribution', 'Experience Years Distribution', 'Indus
```

In [52]:

```
batch_plot_single_choice(survey_res_non_researcher, column_names_demo, titles_c
```







In [53]:
```
# Show industry distribution in df for clarity
industry_counts = pd.DataFrame(survey_res_non_researcher['industry'].value_cou
industry_counts.columns = ['Counts']
```

```

# Calculate the percentage of each industry and add it as a new column
industry_counts['%'] = ((industry_counts['Counts'] / industry_counts['Counts']

# Resetting index to have "industry" as a column
industry_counts.reset_index(inplace=True)
industry_counts.rename(columns={'index': 'Industry'}, inplace=True)

industry_counts
```

Out[53]:

	industry	Counts	%
0	Tech	3	4.62
1	Healthcare	3	4.62
2	IT	2	3.08
3	Banking	2	3.08
4	Real Estate	2	3.08
5	Fintech company, India	1	1.54
6	Financial, Insurance	1	1.54
7	FAST NUCES	1	1.54
8	Utilities , Finance.	1	1.54
9	Food industry.	1	1.54
10	Academia	1	1.54
11	Fitness and wellness	1	1.54
12	Energy Forecasting	1	1.54
13	FDA	1	1.54
14	Data Analytics	1	1.54
15	Airport	1	1.54
16	Insurance	1	1.54
17	Logistics	1	1.54
18	Insurance	1	1.54
19	News and media	1	1.54
20	Consulting	1	1.54
21	National Statistical office	1	1.54
22	E-commerce	1	1.54
23	bank	1	1.54
24	Self learner	1	1.54
25	Currently retired. In past, forestry and financial services.	1	1.54
26	Telecommunication. RF Optimization Engineer	1	1.54
27	Glass/Steel produciton	1	1.54
28	Internet	1	1.54
29	Government	1	1.54
30	Marketing	1	1.54
31	Energy	1	1.54
32	Industrial 4.0 and Robotics	1	1.54
33	Medical image data governance company	1	1.54
34	Health	1	1.54

	industry	Counts	%
35	Power Generation	1	1.54
36	Social media	1	1.54
37	Software (Databases)	1	1.54
38	AI	1	1.54
39	Bank of America Merrill Lynch	1	1.54
40	AgTech	1	1.54
41	Sigma Computing	1	1.54
42	Onit group	1	1.54
43	Food/ecommerce	1	1.54
44	GLG research	1	1.54
45	FMCG	1	1.54
46	Simplilearn	1	1.54
47	Auto parts manufacturing sector	1	1.54
48	Automotive	1	1.54
49	Visagio	1	1.54
50	Holding Slovenskih Elektrarn (HSE.d.o.o)	1	1.54
51	Finance	1	1.54
52	Aerospace	1	1.54
53	martech	1	1.54
54	Logistics, education	1	1.54
55	Education	1	1.54
56	Breda University of Applied Sciences	1	1.54
57	Heidelberg Materials UK	1	1.54

Current dataset search practices & challenges

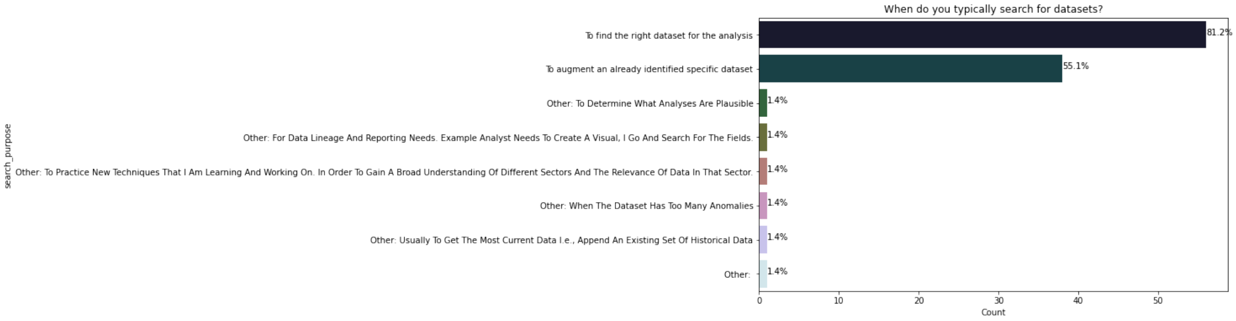
> **When** do you typically search for datasets?
>
> **Where** do you typically look for datasets?

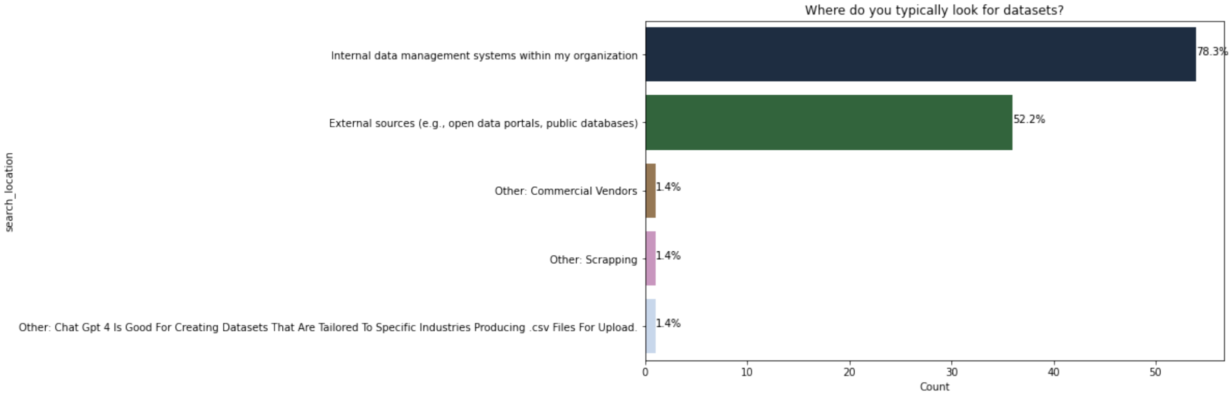
In [54]:

```
column_options_dict = {
    'search_purpose': [
        'To find the right dataset for the analysis',
        'To augment an already identified specific dataset'
    ],
    'search_location': [
        'Internal data management systems within my organization',
        'External sources (e.g., open data portals, public databases)'
    ]
}
```

```
titles = ['When do you typically search for datasets?', 'Where do you typically
```

```
In [55]: batch_plot_multi_choice(survey_res_non_researcher, column_options_dict, titles
```





> Could you specify and describe the **tools you use** for data searching?

```
In [56]: pd.DataFrame(survey_res_non_researcher['search_tool'])
```

Out[56]:

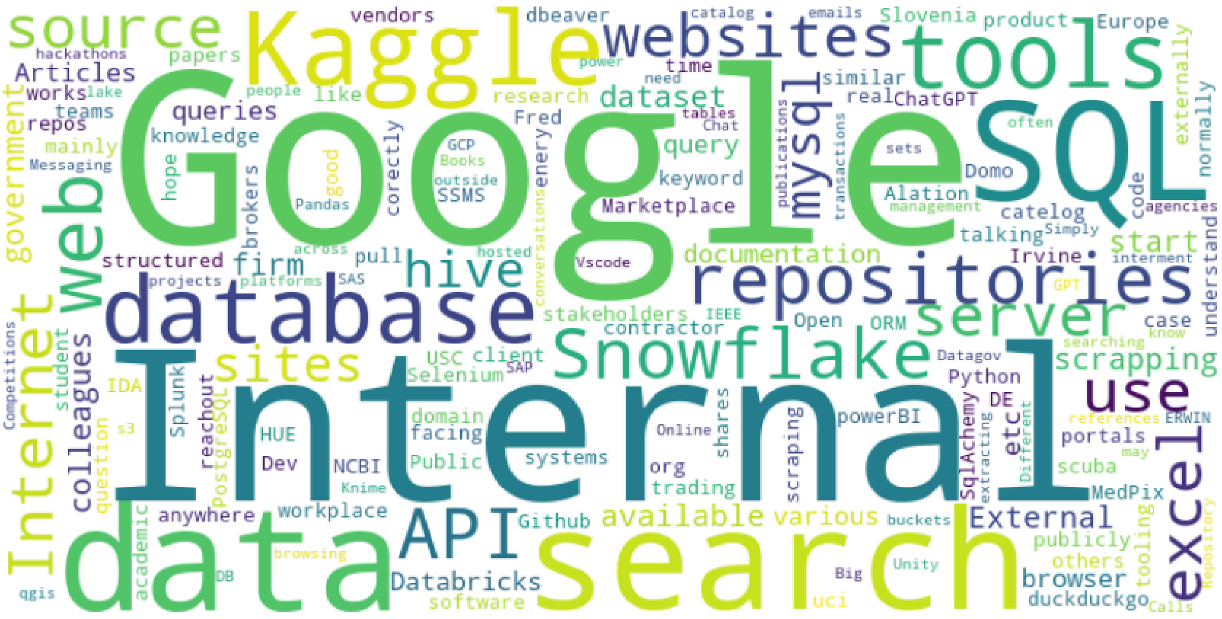
	search_tool
0	internal database repositories
1	hive database
2	mysql，powerBI，Google
3	talking with stakeholders and DE; data catelog
4	Google, NCBI, MedPix, IDA-USC
5	Sql, scuba, hive
6	Internal code repositories or contractor database sites
7	Websites like Kaggle, data scrapping with Selenium
9	internal workplace search
10	Alation
11	Internal shares
15	Google search
16	documentations/reachout to teams on product data for internal use case, externally - anywhere publicly available datasets, research papers are good start
17	Internal tools/portals that pull data from real time servers, similar to Splunk.
18	duckduckgo, domain knowledge, data brokers
20	Snowflake Marketplace, Github
21	mainly a sql client software (dbeaver)
23	Domo, snowflake, internal tools
25	Internet keyword search, vendors, no structured tools
29	Open source and Kaggle
33	Fred.org
34	Excel
35	Python
36	Internal search systems
37	Browsers and API for External Sources and internal tools as HUE, SQL Dev and others
39	As a student who works on enery trading firm in Europe (Slovenia), I normally search for data in firms database. (SSMS - MySQL), SqlAchemy - ORM (I hope I understand question corectly.
40	Google
42	Public facing APIs and other data repos
43	.
44	SQL/PostgreSQL
45	Google dataset, uci Irvine , government sites
46	Web scraping, academic databases, kaggle, Google it...
47	If it’s external, Google search, ChatGPT, Kaggle.
49	Internal tooling for data management (ERWIN)

	search_tool
50	SAS and SQL
51	Google
52	Messaging colleagues, searching documentation from other projects, browsing tables and s3 buckets
53	Online platforms kaggle data sets on various hackathons
54	Start with browser and may need web scrapping or extracting from API
56	Databricks Unity catalog, excel
57	google, kaggle
58	SQL over Snowflake
61	i google
62	Google, Chat GPT 4, Articles, Competitions. Books.
63	Datagov, kaggle, IEEE...
66	Simply Google
67	I use the database tools when available or I use SQL queries
68	Data repositories, Kaggle, etc
69	Web search
70	Search know websites
71	Internal data repositories
72	Pandas
74	Google, Databricks
75	I often search for DB from various interment sources such as government agencies and websites .
76	Big query GCP
77	Snowflake
78	Google search
79	Vscode
81	Internet
82	Google , Kaggle
84	sql
85	Repository on internet
86	google (or other web search tools); conversations & emails with people; references in publications & web hosted articles
87	Knime, sql, power query, excel and qgis
90	Internal tools, colleagues, etc...
91	Calls to outside or internal API's, queries from databases or across data lake
92	Google, SQL Server
93	NaN

	search_tool
97	Different transactions in SAP or SQL server

In [57]:

```
plot_word_cloud(survey_res_non_researcher, 'search_tool')
```



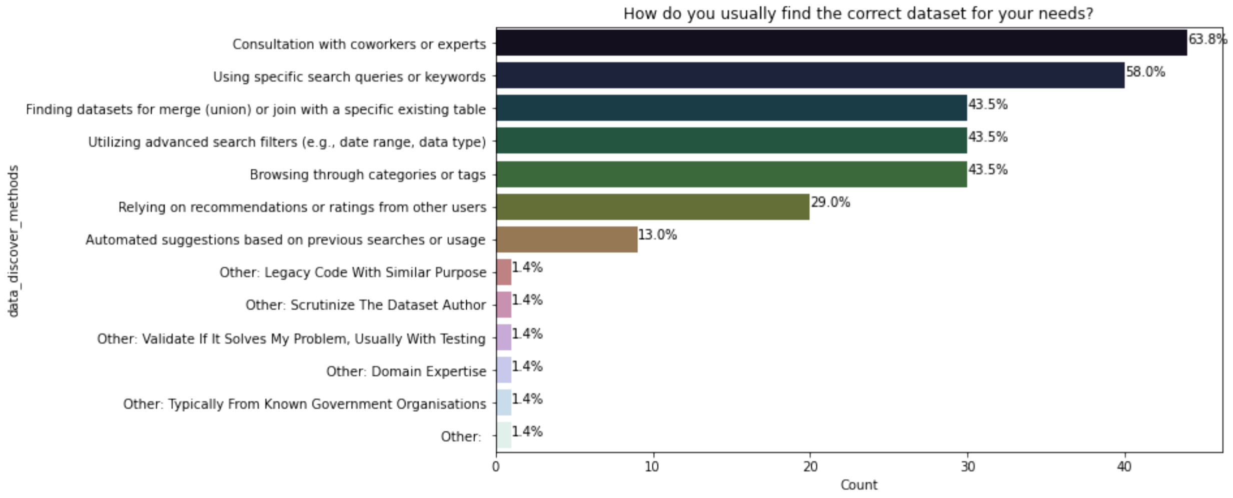
> **How** do you usually find the correct dataset for your needs?

In [58]:

```
column_options_dict = {
    'data_discover_methods': [
        'Using specific search queries or keywords',
        'Browsing through categories or tags',
        'Utilizing advanced search filters (e.g., date range, data type)',
        'Relying on recommendations or ratings from other users',
        'Consultation with coworkers or experts',
        'Automated suggestions based on previous searches or usage',
        'Finding datasets for merge (union) or join with a specific existing ta
    ]
}

titles = ['How do you usually find the correct dataset for your needs?']
```

In [59]:

```
batch_plot_multi_choice(survey_res_non_researcher, column_options_dict, titles
```



> Please briefly **describe your approach** when using the methods selected for finding datasets. For instance, if you chose "Consultation with coworkers or experts", what would you ask for?

```
In [60]: pd.DataFrame(survey_res_non_researcher['data_discover_methods_text'])
```

Out[60]:

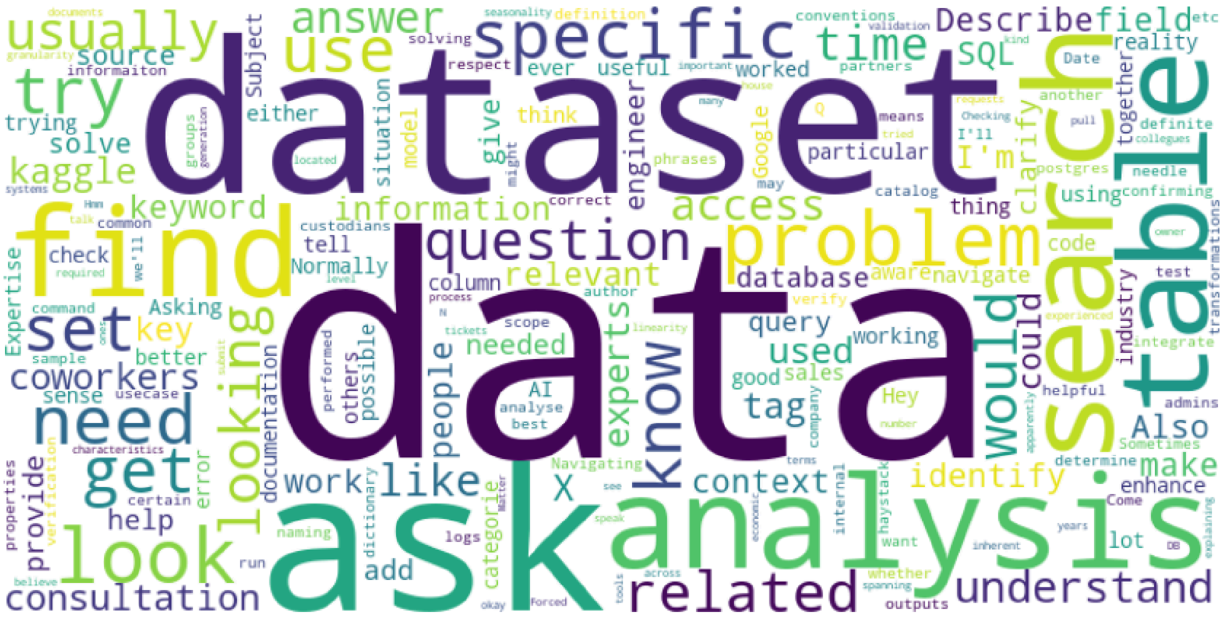
	data_discover_methods_text
0	Most of the time, I'll provide contexts to data engineer partners to help clarify where and how to find data needed for specific questions
1	Search with naming conventions and verify it with data engineers
2	1. to understand the data scope and definition;2. Get access to a sample set and run some analysis test; 3. If no problem need to find we to integrate data into analysis, if not find another data.
3	NaN
4	Use “like” command in sql
5	Search in company’s internal database and Q&A groups
6	I want to know which dataset will help me get to the correct answer or give me the best data to answer my question
7	Use the specific search query keyword and add dataset to it.
9	Ask about the key outputs i was looking for
10	Navigating the data catalog and confirming with data source admins/custodians for verification
11	Describe the problem that I try to solve and what data could be helpful
15	Come up with relevant search phrases.
16	if I know what I am looking for, I would ask whether you have particular data set that will tell me some definite properties etc, which postgres table it might be in, how to get access, who can tell more informaiton on what some of the field means, any data dictionary
17	- During consultation with coworkers, I try to determine the usecase we are solving and relevant data we'll need to analyse. - Sometimes we have to find the needle in the haystack so we search using common keywords in error logs -Date Time tags are most useful in that respect -
18	I check the sources used by the dataset author and which transformations they may have performed
20	I ask for datasets with certain characteristics, like sales data spanning N years with some number of sales categories, seasonality and some level of granularity.
21	Having so many tables, I ask more experienced collegues which ones are most inherent to the analysis I need to do. I then navigate through the categorie and tags to looks for others
23	I identify what is required for the analysis, then speak with the owner of the systems that I believe have that data and then either pull the data myself or submit tickets for data requests
25	Hmm. We talk about the problem. (Forced answer)
29	Subject Matter Expertise is more important.
33	Key economic terms
34	Checking the data linearity across
35	In house documents search and AI tools
36	A documentation explaining the data generation process
37	Did you ever used this kind of datasets? Do you think that it makes sense this datasets? I tried this thing and apparently things are okay, do you see other validation i should do?
39	Hey where is data for X located in DB? Or in majority of the time I use excel files where they already set SQL queries.
40	I look for the comments for other users in kaggle

	data_discover_methods_text
42	What data sets have you leveraged that will enhance my current analysis or which datasets do you think would enhance my analysis
43	.
44	If I’m searching in SQL i use to take a look to the dictionaries of the tables in the data base, if a find what I’m looking for I code the query. When consultation with coworkers or experts I ask them how they handle the situation or how how get the information if they had the same situation
45	Normally , data in my organization is a mess so I need to make a lot of question
46	I try to understand the reality of the context I am creating the analysis/AI model for, after that I can better understand what data more accurately represents this reality
47	Outside of obvious solutions like Google or Kaggle, I inquire about the type of data amongst people I trust or have a working relationship with. They either share it with me if the eg have access to it, suggest where I can look online for it, or reach out to people they know for industry data that can be anonymized.
49	Generally just ask something along the lines of "hey i have data X and would like to add information on Y, where could i find this?
50	Specific tables to look or useful column. Also how to join tables together
51	Check the manual for tags related to the subject
52	Asking if they know where data related to X resides, and how can grant access
53	asking for input on various realtiblity
54	Identify the peoblem and the data for the problem. Then based on the data needed to answer the problem used specific keyword or tag search. Also, identify people who have worked on similiar problems and try to contact them to understand data they used.
56	I'm looking to create this metric, do you know where this data lives ? I
57	I use search engines as other choices are not good enough
58	I ask for a table or view containing data I need.
61	i just google for data. i usually get related datasets from kaggle and github. a lot of the times from other random sites too.
62	I would typically ask if they were aware of a dataset that is good for practicing x , for instance or I would as if they were aware of a dataset that is tailored to a specific industry that was fun to explore.
63	I usually try to work on what others have worked with since I am new to the field.
66	Ask them about prior encounter with the dataset
67	Describe what I’m trying to model and get a database recommendation
68	most relevant dataset possible
69	“Does anyone know where to find XYZ healthcare reference code set or who publishes the authoritative table”
70	Usually simply update existing data from known providers
71	I normally communicate my needs to coworkers or the experts of a specific area. I could also navigate through the data warehouse by my own trying to make sense of the data.
72	Find the origin or the data.
74	Data governance
75	I usually use government websites , never do any consultations

	data_discover_methods_text
76	Objective of the work, decision making related questions.
77	Reviewing the contents of tables for details and keys
78	Give the context about the analysis or work Im doing and ask.
79	Kaggle
81	I do my own research
82	I search for datasets that try to solve the same ML problem
84	Ask for the tables that include the rows & columns i'm looking for
85	Filtering and applying to fit for optimum results
86	Describe question / analysis I am working on; why I'm doing it; why I thought of them and then ask for any experience / suggestions
87	My field was telecommunication. We should look through documentation and ask for experts. Usually we clarify the symptom and look at the counter related the problem
90	Quite often you dont have the expertise to decide if that particular dataset suits your needs or not. After talking with domain experts, I look at the data set, seeking for correlations with my features, targets
91	Consultation with coworkers: find out if better data available; using specific or advanced querying: improves the speed of putting together the dataset; unions: offer a bigger picture of the analysis with possible additional piece of information
92	Trail and Error
93	Have you ever resesrched on this? What challenges did you face?
97	Which transaction will provide the data for the task?

In [61]:

```
plot_word_cloud(survey_res_non_researcher, 'data_discover_methods_text')
```



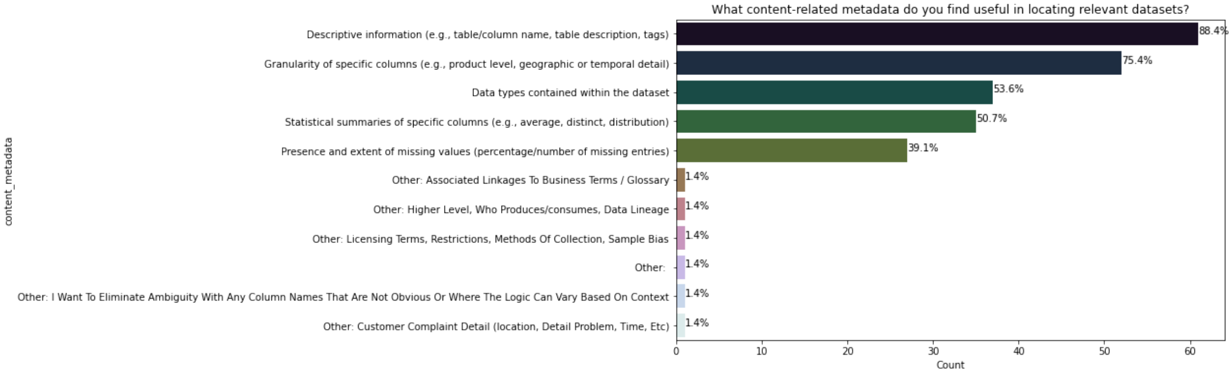
> What **content-related metadata** do you find useful in locating relevant datasets?

> What **table-related metadata** do you find useful in locating relevant datasets?

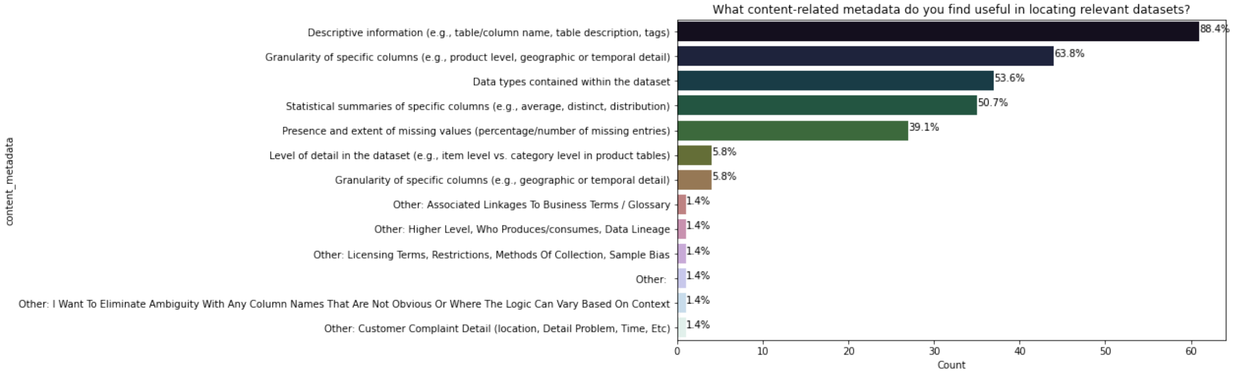
In [77]:
```
column_options_dict = {
    'content_metadata': [
        'Data types contained within the dataset',
        'Presence and extent of missing values (percentage/number of missing e
        'Descriptive information (e.g., table/column name, table description, 
        'Statistical summaries of specific columns (e.g., average, distinct, d
        'Granularity of specific columns (e.g., geographic or temporal detail)
        'Level of detail in the dataset (e.g., item level vs. category level i
        'Granularity of specific columns (e.g., product level, geographic or t
    ],
    'table_metadata': [
        'Previous SQL queries made by other users on the dataset',
        'Usage data (e.g., number of accesses, frequency of use)',
        'Update history (e.g., frequency and schedule of data updates)',
        'Availability of earliest data in the dataset',
        'Detailed schema of the dataset',
        'Dimension details (e.g., number of columns, size of dataset)',
        'Data lineage'
    ]
}

titles = ['What content-related metadata do you find useful in locating releva
```

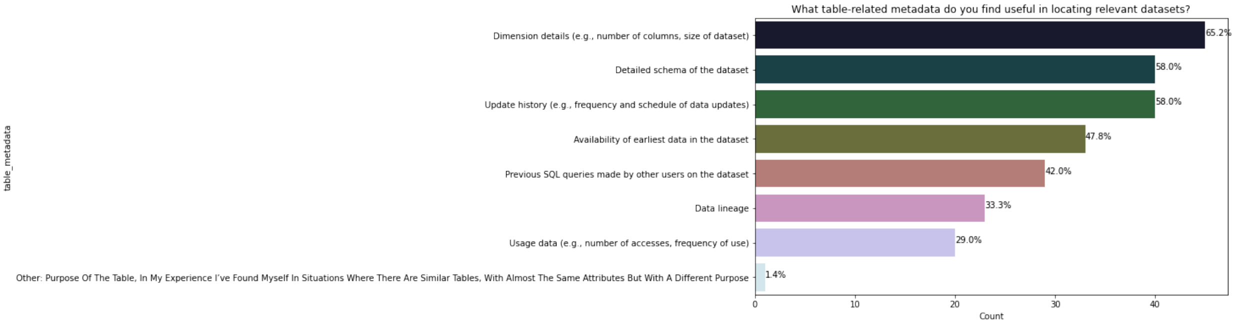
In [78]:
```
# Options to be combined
options_to_combine = [
    'Granularity of specific columns (e.g., geographic or temporal detail)',
    'Level of detail in the dataset (e.g., item level vs. category level in pr
    'Granularity of specific columns (e.g., product level, geographic or tempo
]
```

In [79]:
```
# Plotting the distribution for the content-related metadata question
plot_combined_options_distribution(
    data=survey_res_non_researcher,
    column_name="content_metadata",
    predefined_options=column_options_dict["content_metadata"],
    combined_options=options_to_combine,
    title="What content-related metadata do you find useful in locating releva
)
```



In [80]:
```
batch_plot_multi_choice(survey_res_non_researcher, column_options_dict, titles
```





> Imagine you had an ideal dataset search system, can you give an **example query** (the query can be in natural language, doesn't have to be SQL) that you would like to find relevant datasets for?

In [64]:

```
pd.DataFrame(survey_res_non_researcher['ideal_query_example'])
```

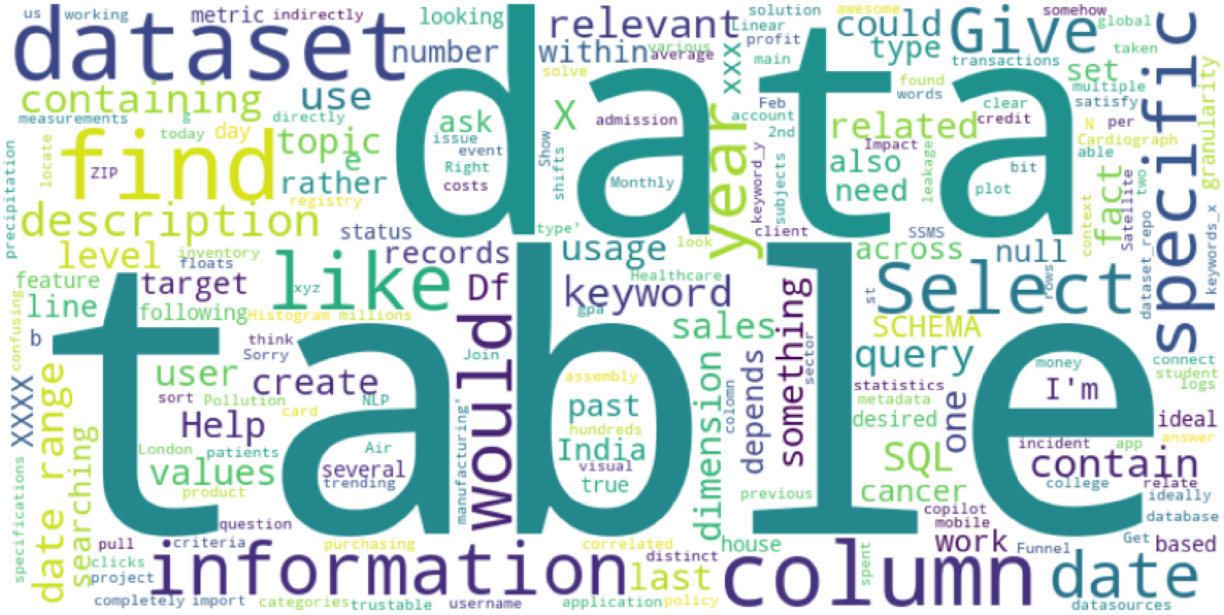
Out[64]:

	ideal_query_example
0	which table/ column can I use to query user level purchasing data?
1	Keywords of the subjects+ username of previous users
2	Help me locate the data relate to sales in database
3	help me find the data with XXXX information and relevant columns.
4	Find “target”;
5	N/a
6	Give me the inventory dataset with the most up to date categories within this date range
7	Cardiograph data set taken in the last 5 years.
9	Help me find the dataset for xxx information
10	CONTAINS LIKE "&SCHEMA.TABLE" or select all tables that contains words "%XXX%"
11	.
15	college admission statistics correlated with student gpa
16	give me the datasets related to directly or indirectly "Air Pollution in London"
17	"Find all the logs from 2nd Feb today with the keyword xyz"
18	this completely depends on the application I'm searching data for, doesn't it?
20	Show me product usage datasets where the main fact table is event-level usage data with hundreds of millions of records and there are dimension tables for user and account.
21	I think this answer depends on the project you are working on. Right now I would ask to find all tables containing data for a specific context (e.g. Healthcare) with a relevant number of distinct values per column. I would ask also to find registry data tables, rather than fact tables, rather than dimension tables.
23	ideally I would have something that could work across all of the various datasources and table and be able to use SQL (or a trustable NLP solution) and pull all the relevant data and metadata
25	Find incident cancer patients by ZIP-3 and cancer type for years 2018-2023 by year
29	Not clear
33	Select colomn from table where date > a.
34	Linear data
35	Find all datasets containing information on leakage measurements on assembly line st* across multiple shifts
36	Histogram or line plot of features
37	I need to find some datasets related to X information with Y specifications (Like date range, or containing specifics columns)
39	That would be awesome! If you could somehow connect SSMS with some sort of copilot.
40	Give me a dataset for the credit card transactions in us and India
42	SELECT * FROM dataset_repo WHERE type LIKE 'some type' AND sector = 'manufacturing'
43	.
44	Monthly average sales, costs and profit (as floats) of the last two years
45	Select * from x where column is not null

	ideal_query_example
46	From X import keywords_x + keyword_y... Sorry I found the question a bit confusing
47	Funnel data
49	Join usage over the past 5 years on the client numbers I have in this query
50	Give me all rows that satisfy the following criteria with the following information
51	Impact for money spent on policy
52	
53	not ideal
54	Dataset to <solve issue of...> with columns <1,2,3,...> on <granularity desired>
56	I'm looking to create a trending visual for this metric, what tables do I need to look at to create the metric.
57	find a global dataset with precipitation by date
58	Get dataset of all clicks from the mobile app.
61	"Satellite water bodies images", "UK Road Traffic accidents dataset"
62	I would like you to create a dataset for industry x , I would like three dimension tables , for a , b , c and a date table spanning 2 years with a date range e to b. The fact table should be at r granularity and contain 20,000 records. I would use something like this in Chat GPT 4 , for instance.
63	looking for a dataset containing information on temperature trends over the past century, broken down by region and updated annually. It's difficult to find in African countries mostly.
66	Select specific-columns from table where specific conditions
67	Workouts that have max performance power data.
68	Topic model search results, based on sentence similarity with the dataset description
69	"select * from tables where topic like keywords"
70	Not really relevant for my purposes
71	SELECT tables FROM master_database WHERE description ILIKE "%something%"
72	At first, filter then groupby followed by ranking
74	....
75	Df[Df[date]> a specific date and season \nDf[Colum]•unique() and searching for percentage of missing values in feature columns .
76	I use Big query, SELECT * FROM WHERE etc.
77	Sql query to find sales
78	Find topics based in country level within a specific date range
79	Please share one day, 15 days, 6 months kurtosis of all listed companies in india
81	Timeseries
82	Give all the datasets related to sequence recommendation systems
84	For such a scenario, syntax like SQL Wildcards is best. WHERE <desired col name> LIKE '%xy%'
85	Data set without null values
86	size, age, format. Ideal would be schema with descriptions of fields.

	ideal_query_example
87	Where and when the problem occurs also which site and area affected?
90	does featurexy corelate with my_feature/target
91	I work with filtering house properties. It is hard to find specific attributes in a house description: is it a townhome vs a standalone home, is the basement finished or not, is the heating central, if it has pool is it private or shared? So I had to correlate data over several columns to make a determination. Querying that allows for: if description in one column matches and specific keywords in another column exist, then it's a match. Or if flag is true in one of several columns (which either of them could contain information on a status), then that status is true.
92	Give me aggregated data for the last 3 years.
93	Can I have information on xxxx ?
97	SQL

In [65]:
```
plot_word_cloud(survey_res_non_researcher, 'ideal_query_example')
```

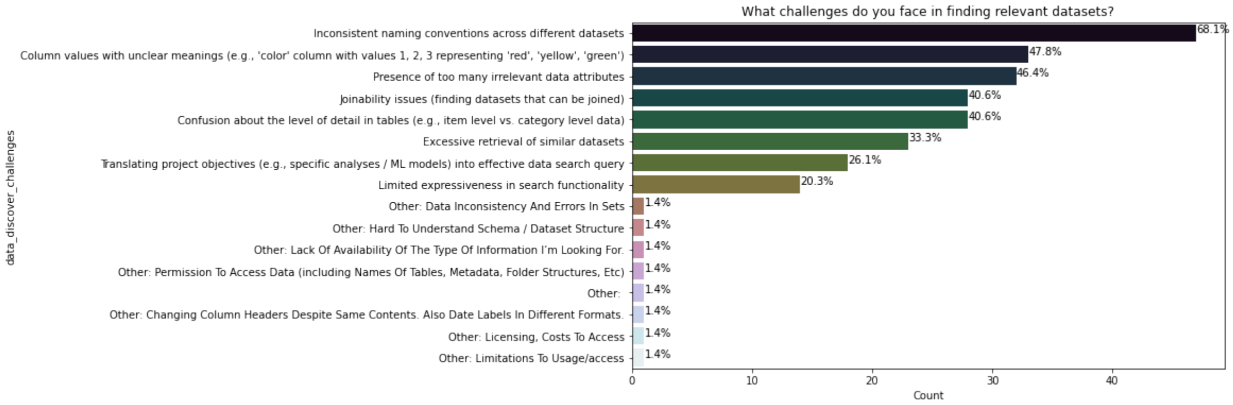


> What **challenges** do you face in finding relevant datasets?

In [66]:
```
column_options_dict = {
    'data_discover_challenges': [
        'Excessive retrieval of similar datasets',
        'Joinability issues (finding datasets that can be joined)',
        "Column values with unclear meanings (e.g., 'color' column with values
        'Presence of too many irrelevant data attributes',
        'Confusion about the level of detail in tables (e.g., item level vs. ca
        'Inconsistent naming conventions across different datasets',
        'Limited expressiveness in search functionality',
        'Translating project objectives (e.g., specific analyses / ML models) i
    ]
}

titles = ['What challenges do you face in finding relevant datasets?']
```

In [67]:

```
batch_plot_multi_choice(survey_res_non_researcher, column_options_dict, titles
```



> Could you provide **a specific example** about the **challenges** you've selected above?

In [68]:

```
pd.DataFrame(survey_res_non_researcher['data_discover_challenges_text'])
```

Out[68]:

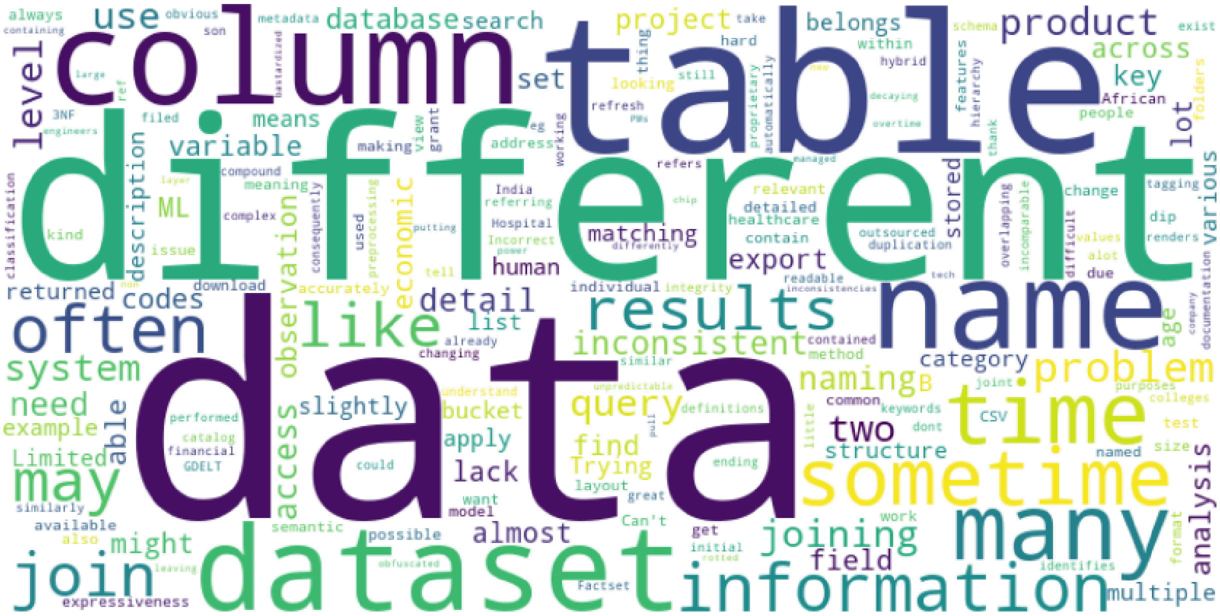
	data_discover_challenges_text
0	Most of the time there are too many table results after the initial search. Some of the filed definitions are slightly different (for example, when refers to age information, age bucket might be different).
1	No name description for data columns
2	In product table, we have different product hierarchy, so some product belongs to A category in table A, but belongs to category B in table B
3	same fields may have different names in two tables, and sometimes the same name may means different in two tables.
4	Hospital names across databases are different and ending up with “human” matching outsourced to India to join them.
5	Datasets don’t have joint key
6	Many data tables have very similar names or not human readable names
7	data duplication, not able to tell if the returned results were already returned
9	Same information can be stored in various datasets for different purposes
10	Incorrect classification and tagging, multiple datasets that are same structure but different refresh, dont exist in catalog, ref integrity in joining
11	,
15	Trying to find a list of colleges that my son could apply to.
16	if you take GDELT for eg, its difficult to understand what is metadata and what is data
17	Limited expressiveness: not many features to search/query keywords, alot of times changing query still renders same data results
18	2 different tables might have a similarly named column referring to almost the same thing, but each had slightly different preprocessing performed, making them incomparable
20	I used a Factset financial dataset for a project that had a complex schema (3NF+ with compound join keys) and column names that were proprietary / not obvious to me. There was no great documentation on the structure of the dataset or how I should be joining various tables.
21	I am working on a hybrid system (also thank to you) that automatically identifies the semantic meaning of the data contained in a column. Many of these columns often contain codes with little semantic meaning, often with values in overlapping with other columns and consequently unpredictable. Or often columns containing the same data are named differently between different tables. I often have many tables, but few columns with information power within them
23	within my large company the data has been managed like a 7-layer dip, with each new set of engineers, PMs, and non-tech people leaving inconsistencies and obfuscated data that has rotted overtime. Each data pull is like putting a chip into this decaying dip of bastardized data and hoping that someone doesn't get food poising from the results.
25	While several datasets do describe the same entities, there are no common keys or matching criteria at the granular level.
29	Categorical level of detailing is required, which is not possible this days.
33	.
34	I work on learner database which is filled by different individuals. Major problem is the inconsistency in Data representation
35	Data not clean
36	Data format and sizes can be different

	data_discover_challenges_text
37	I was looking for some dataset from population census and things were kind extensive
39	/
40	Multiple columns with same information
42	Definitely joining data sets on temporal attributes like between a given time window
43	.
44	Once I had to do an analysis but it was painful because almost every column had unrecognizable information (like encrypted) it took longer that I was expecting
45	Sometimes yo don't find a detailed description of the dataseyt
46	Limited expressiveness in functionality. For me this is the biggest challenge, find datasets that most accurately represent the reality of the problem I am trying to solve
47	I work in marketing analytics and most of the data that I’d use to do analysis isn’t typically publicly available making it hard to test out techniques.
49	I have faced many challenged with addresses, you want to join on address level but there are a million way to write down an address
50	Data to calculate target variable of a forecasting project was inconsistent over time as source systems had been changed over time so getting the data consistent was tedious
51	People track the same useless stuff with different names
52	A bunch of data resides in s3 buckets in a different account, so I can’t view folders/names, and it is a prolonged process to request access, and when requesting access I usually have to specify individual folders (because they don’t want to grant broad access, and need detailed business justification to grant any access)
53	Confusion about the level of detail in tables
54	Can't share due to NDA
56	When looking at two membership system, the member Id varies based on the state represented. Makes it very tough to join. The grain of the data is also different, for healthcare sometime the primary subscriber is hidden or can't see all of the dependents.
57	Datasets available, but not able to download due to size of the base data. Or API caps.
58	tables and views have a suffix of `_new` or `_v2` which I have no idea what it means.
61	i once had to use two separate datasets for analyzing road traffic accidents data. the datasets had common column which i used for joining them but it caused a lot of null rows and other redundant data
62	No real problems but sometimes tables can contain a lot of detail that is not always relevant for the analysis at hand. However, there is alsways scope to drop this columns or remove them and create a seperate dimension table , for instance.
63	The use of standardized economic indicators that may not accurately capture the unique socio-economic context of African nations. Metrics like Gross Domestic Product (GDP) per capita might not reflect the informal economy, subsistence agriculture, or other significant contributors to livelihoods in many African countries. This can lead to misleading comparisons or assessments of economic development.
66	Receiving data from meteorology sensors often are in vague and ambiguous naming.
67	Finding an open database that fulfills my data requirements for ML models.
68	NaN
69	Need to join on inconsistent identifiers or unclear if versions are the same. For example, in healthcare CPT codes are often retired or updated. It’s hard to get a comprehensive list that

	data_discover_challenges_text
	includes all possible codes.
70	Different column names over time. Say a dataset is a CSV that covers 2022 then in 2023 the relevant CSV has different headers.
71	The naming between transactional systems, reports and tables located in the warehouse are almost always completely different and denormalized.
72	Groupby
74	Too short space to answer
75	Doing a ML project where selecting features pivots from observation points. Sometimes these observations lack huge observations .
76	Wrong Columns name irritates me.
77	Relevance
78	Messy data
79	1. Different data nomenclature for same variables maintained by Different ecosystems, 2. lack of clear linkage path for text data
81	The tables do not have details about fields
82	Too many different formats for this kind of problem
84	"Inconsistent naming conventions across different datasets". This is quite a big issue as the same variable is stored with a different name across different teams.
85	Fitting the set on ML model for best results
86	Nothing comes to mind
87	How to check the root cause of problem using data, when sometime other data showing opposite trend
90	You have to do lot of steps in order to decide if the selected dataset quits your needs. You need to download it, look at statistics of columns of intereset, apply other tests, etc. This is very time consuming issue.
91	Oftentimes I am not able to drill more specifically (a more specific query) for lack of sufficient details to query on
92	Too may confusing tables
93	Information that is embedded in other cstegories
97	Inconsistent naming, typos in column names, which layout to use for export, which method to export, layout may change, field names sometimes change depending on export method

In [69]:

```python
plot_word_cloud(survey_res_non_researcher, 'data_discover_challenges_text')
```


data
different
table
column
name
dataset
information
many
sometime
time
join
often
results
like
may
problem
inconsistent
query
detail
joining
system
need
naming
two
access
observation
economic
slightly
variable
level
use
database
search
project
belongs
product
across
key
set
means
lot
various
export
human
matching
stored
description
ML
returned
codes
list
category
able
bucket
apply
find
lack
Trying
structure
field
might
almost
analysis
multiple
example
Limited
age
always
obvious
metadata
take
schema
exist
containing
son
hard
within
thing
hybrid
features
hierarchy
proprietary
automatically
still
refresh
looking
African
people
folders
3NF
filed
engineers
large
ref
bastardized
grant
view
making
address
eg
working
refers
managed
thank
decaying
overtime
compound
meaning
layer
complex
consequently
preprocessing
used
India
Hospital
referring
Incorrect
power
putting
differently
relevant
detailed
healthcare
contain
outsourced
duplication
overlapping
incomparable
difficult
change
tagging
dip
renders
values
alot
due
classification
kind
issue
download
accurately
tell
individual
integrity
readable
inconsistencies
tech
documentation
company
already
changing
contained
method
similar
understand
unpredictable
pull
little
definitions
common
keywords
joint
purposes
colleges
test
CSV
dont
named
size
performed
catalog
financial
GDELT
could
ending
layout
great
want
model
semantic
possible
Can't
get
initial
rotted
work
identifies
format
similarly
available
also
leaving
obfuscated
Factset
expressiveness
chip
PMs
B