Open Brewery Dataset

**Project Overview 📝**

At its core, your script is an **ETL (Extract, Transform, Load)** tool.

* **Extract:** It pulls data from an external source, the Open Brewery DB API.
* **Transform:** It processes this data by organizing it into a structured table.
* **Load:** It saves the data into two different file formats on your computer: a JSON file and a CSV file.

**Step-by-Step Explanation**

**1. Data Extraction (E)**

* **API endpoint: https://api.openbrewerydb.org/v1/breweries**
* **Pagination:**
  + **API returns a maximum of 200 results per page.**
  + **Loop until the API returns an empty list (no more data).**
* **Delay: 0.1 seconds between requests to avoid overwhelming the API.**
* **Error Handling:**
  + **Checks HTTP status code**
  + **Catches RequestException for network errors**
* **SSL: Disabled warnings for simplicity (verify=False).**

**Talking point:**

**“We ensure scalability by looping until no more data is found, which means if the API adds more breweries in the future, our code automatically fetches them.”**

**Data Transformation (T)**

* **Save raw JSON for full unprocessed data (all\_breweries\_data.json).**
* **Normalize JSON to DataFrame with pd.json\_normalize().**
* **Select relevant columns:**
  + **Brewery info (name, type, address, phone, coordinates, website).**
* **Null Value Summary: Check missing data to plan cleaning.**
* **Remove duplicates: Based on unique brewery id.**
* **Clean postal codes: Keep only the 5-digit ZIP.**
* **Format phone numbers: (XXX) XXX-XXXX.**
* **Geo dataset: Create df\_geo containing only breweries with latitude & longitude for mapping.**

**df = pd.json\_normalize(all\_breweries): This is a powerful Pandas function. The data from the API is a list of nested dictionaries. json\_normalize flattens this complex structure into a simple, 2D table (a DataFrame), which is perfect for a CSV.**

**Main Short Note**

This script extracts complete brewery data from the Open Brewery DB API by looping through all its pages. It handles errors and uses the requests library to fetch the information page-by-page. The raw data is first saved as a JSON file for a complete backup. Finally, it uses the pandas library to clean this data and save it as a user-friendly CSV file.

# why use json

APIs often send data in a nested format (e.g., a list of addresses within a single brewery record). JSON handles this complex structure perfectly.

Cleaning part

Loading and Initial Inspection ,removes any duplicates, and standardizes the formatting of columns like postal\_code and phone for consistency

Analysis

1. top 10 largest brewery chains

Approach

**Identify Top 10 Chains**:

First, df.groupby('name').size() counts the number of locations for each unique brewery name.

**Create a Pivot Table**: This is the key transformation step.

The code groups the data by both name and state and counts the locations.

used is a **Stacked Horizontal Bar Chart**. A Stacked Horizontal Bar Chart is a dual-purpose graph used to compare totals and see their internal breakdown at the same time.

* This contrast shows two paths to success: **wide geographic presence** vs. **deep regional dominance**.

Geographic distribution

The primary business insight from this visualization is that **there are two distinct and equally viable strategies for scaling a brewery chain to a national level.** The choice between these models fundamentally shapes a company's marketing, branding, and operations.

**The Widespread National Expansion Model 🌎**

This strategy focuses on achieving a broad presence across numerous states to capture a diverse national market.

**The Compact Regional Dominance Model 📍**

This strategy involves concentrating locations in a specific geographic area to saturate the market and build a loyal local following.

**2.Total brewery vs density breweries**

The best way to do this was to **normalize the brewery count by the state's land area**. This controls for the size variable and helps measure market concentration."

"

**. The Raw Count Chart: "Who has the MOST?"**

This first chart is simple. It answers the question, "Which state has the highest total number of breweries?"

* **California is #1 here simply because it is enormous.** It has a huge population and a massive land area, so it naturally has more of everything, including breweries. This chart is accurate, but it doesn't tell the whole story.

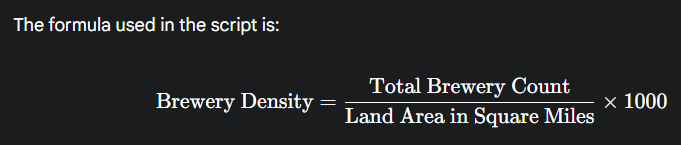
**2. The Density Chart: "Who is the MOST DEDICATED to breweries?"**

This second chart is the sophisticated one. It answers the better question, "Which state has the most breweries *for its size*?"

To do this accurately, I sourced a standard and reliable dataset for state land area, which I verified against figures from the **U.S. Census Bureau**.

* **California vanishes because its high brewery count is spread over an immense land area.** When you divide the number of breweries by its huge size, the "density" number becomes relatively small.

This new metric of **brewery density** completely reframed the results. It revealed that the true brewery hotspots—the states with the most intense concentration of breweries—are actually smaller states



This gives you the number of **breweries per 1,000 square miles**,

Graph used here Side-by-Side Horizontal Bar Charts.

Business Insights – Brewery Count vs. Brewery Density

1. Expansion Targeting:

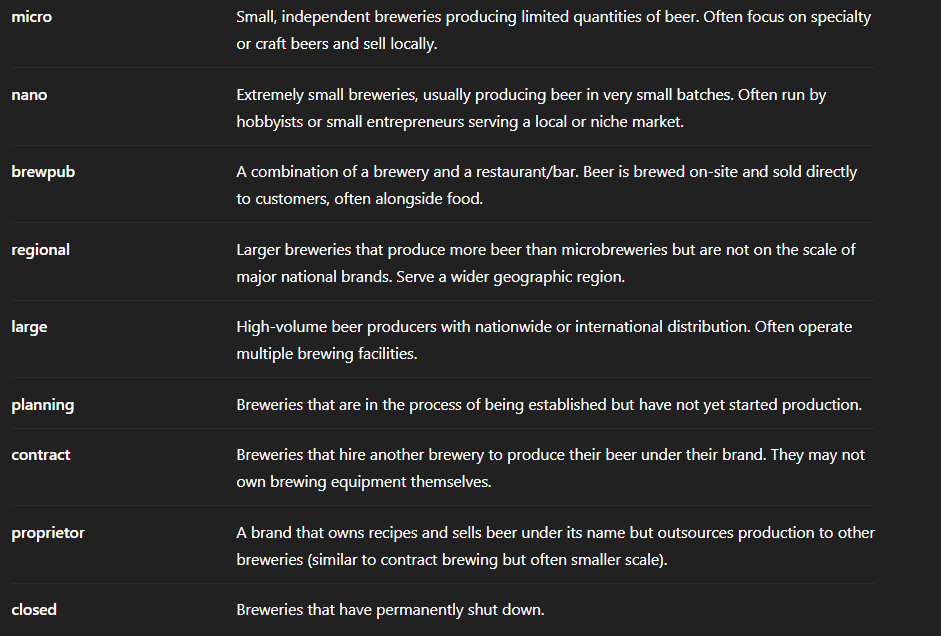
Large states with high total breweries but low density (e.g., California, Texas) may have **less-covered areas** where competition is lower, offering room for new entrants.

**2.Competitive Hotspots:**

* High-density states (e.g., Rhode Island, Massachusetts, Connecticut) signal **intense local competition**, requiring strong differentiation

**3.Distribution Efficiency:**

* Breweries in high-density areas can **reduce transportation costs** and reach customers faster, improving freshness and customer satisfaction.



Bids

The **Delaware website** is a Central hub where the state handles buying and supply-related work. It helps government offices get what they need—from services to products—by working with businesses.

### **What is a Bid Solicitation?** **suh·li·suht·ay·shn**

Imagine the **government wants to build a road** or **buy computers** for schools. But instead of just picking a company, they say:

“Hey, we need this work done. If you're a company that can do it, tell us how much you'll charge and how you'll do it.”

That public announcement is called a **bid solicitation**.

### **🔧 In Simple Terms:**

* It’s like a **job posting**, but for companies.
* The government says: “Here’s what we need. Who can do it best and at a good price?”
* Companies then **submit their offers** (called “bids”) explaining:
  + What they’ll do
  + How much it will cost
  + How long it will take

### **🏁 After That:**

* The government reviews all the bids.
* They choose the **best one** based on quality, price, and experience.
* That company gets the **contract** to do the job.

The **three categories** you see—**Open**, **Recently Closed**, and **Not Awarded**—refer to the **status of bid solicitations**.

### **Open**

These are **active solicitations** that are currently accepting bids or proposals from vendors. They include:

* **Request for Proposal (RFP)**
* **Invitation to Bid (ITB)**
* **Request for Information (RFI)**

Vendors can view details, ask questions, and submit their proposals before the deadline

### **1. Request for Proposal (RFP)**

* Used when the state needs a **solution to a problem**, not just a product.
* Vendors are expected to provide **detailed proposals** explaining how they will meet the state’s needs.
* Evaluation is based on **multiple factors**, not just price—like experience, approach, and innovation.

### **2. Invitation to Bid (ITB)**

* Used when the state knows **exactly what it wants** and is looking for the **lowest price**.
* Vendors submit bids based on **clear specifications**.
* The contract usually goes to the **lowest responsive and responsible bidder**.

### **Request for Information (RFI)**

* Not a request to buy something yet—it's used to **gather information** from vendors.
* Helps the state understand what’s available in the market before creating an RFP or ITB.
* No contract is awarded from an RFI directly.

### **Recently Closed**

These are solicitations that have **reached their submission deadline** and are no longer accepting bids. However:

* The evaluation process is still ongoing.
* No final decision or award has been made yet.

### **🔴 Not Awarded**

These are solicitations that were **closed without selecting a winning vendor**. This can happen due to:

* No suitable proposals received.
* Budget constraints.
* Project cancellation or scope change.

## **Key Insights You Can Extract from Bid Data**

### **1. Bid Volume & Frequency**

* How many bids are posted weekly/monthly?
* Which departments/agencies post the most bids?

### **2. Bid Categories & Types**

* What types of bids are most common? (RFP, ITB, RFI)
* Which categories (e.g., construction, IT, consulting) are most active?

### **3. Award Trends**

* Which companies are winning bids?
* What’s the average bid value or contract size?
* Are there repeat winners?

### **4. Bid Deadlines & Submission Patterns**

* What’s the average time window for submission?
* Are there seasonal trends (e.g., more bids in Q1 or Q3)?

### **5. Location-Based Opportunities**

* Are bids concentrated in certain regions or cities?
* Can your company target specific locations?

### **1. Agency Activity Analysis**

**Why?** To know which government departments are actively issuing bids.

* Identify agencies with the **highest bid volume**.
* Track **seasonal patterns**—some agencies may post more in certain quarters.
* Prioritize agencies that align with our company’s services.

### **2. Bid Submission Timelines**

**Why?** To plan resources and avoid last-minute submissions.

* Calculate the **average duration** between bid opening and deadline.
* Identify bids with **short windows** (urgent) vs. **long windows** (complex).
* Use this to build a **submission calendar** and allocate proposal teams accordingly.

### **3. Winning Vendor Patterns (from Closed Bids)**

**Why?** To understand the competition and pricing strategies.

* Analyze which vendors are **winning repeatedly**.
* Study their **bid types**, **pricing**, and **service categories**.
* Benchmark our offerings against theirs to improve competitiveness.

### **4. Quotation & Pricing Trends**

**Why?** To avoid overpricing or underpricing.

* Extract pricing data (if available) from awarded bids.
* Identify **average winning bid values** by category.
* Use this to **fine-tune our quotations** and stay within competitive range.

### **5. Bid Type Distribution**

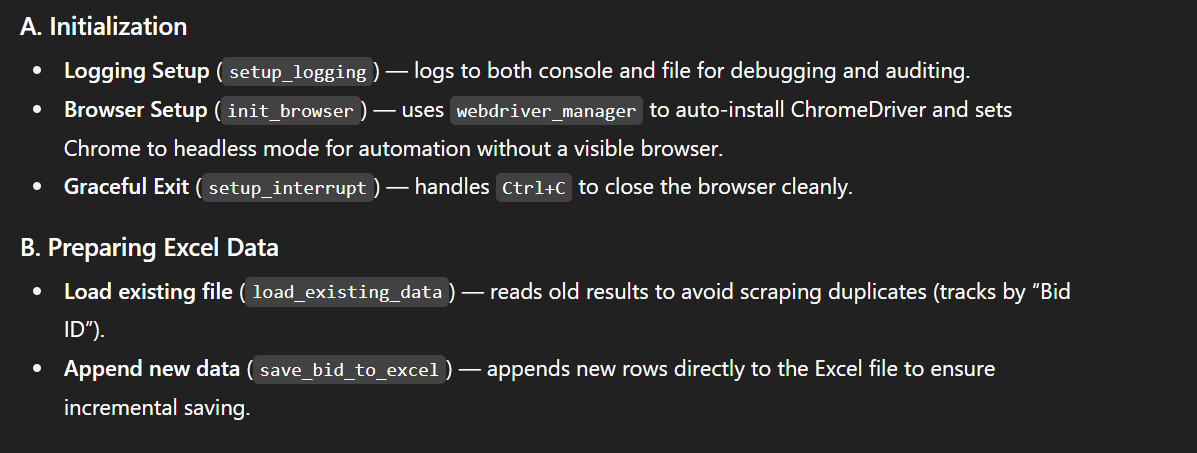
**Why?** To align our proposal strategy.

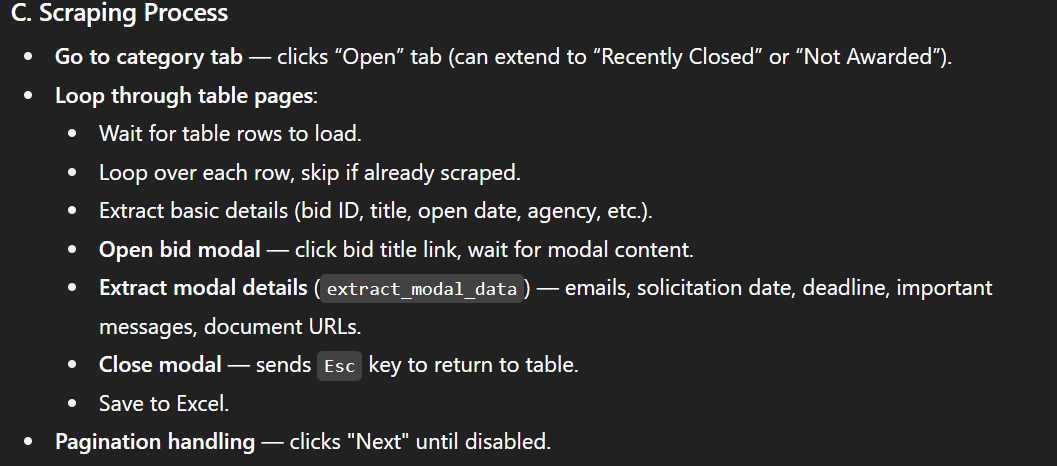
* Understand the mix of **RFPs, ITBs, RFIs**.
* RFPs require detailed solutions—ideal for consulting or tech firms.
* ITBs are price-driven—ideal for product suppliers.
* RFIs are exploratory—good for early engagement.

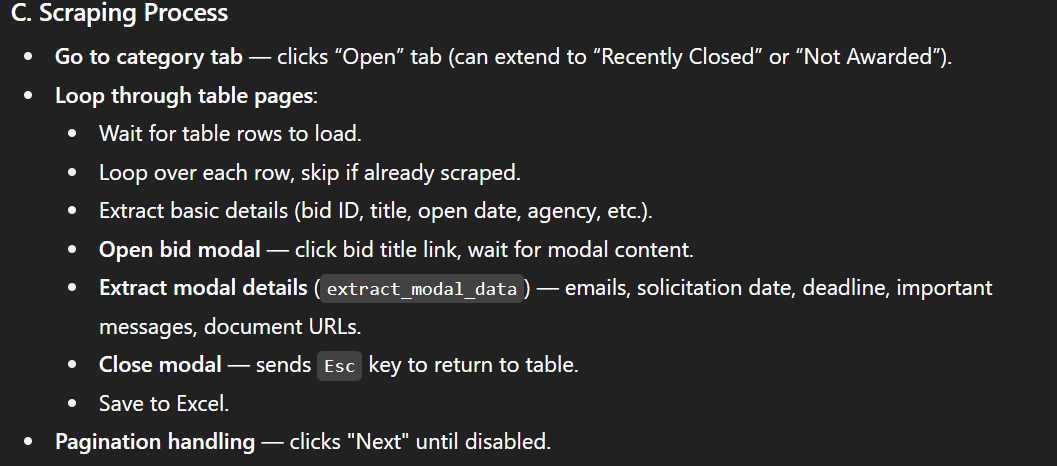
### **6. Reasons for Not Awarded Bids**

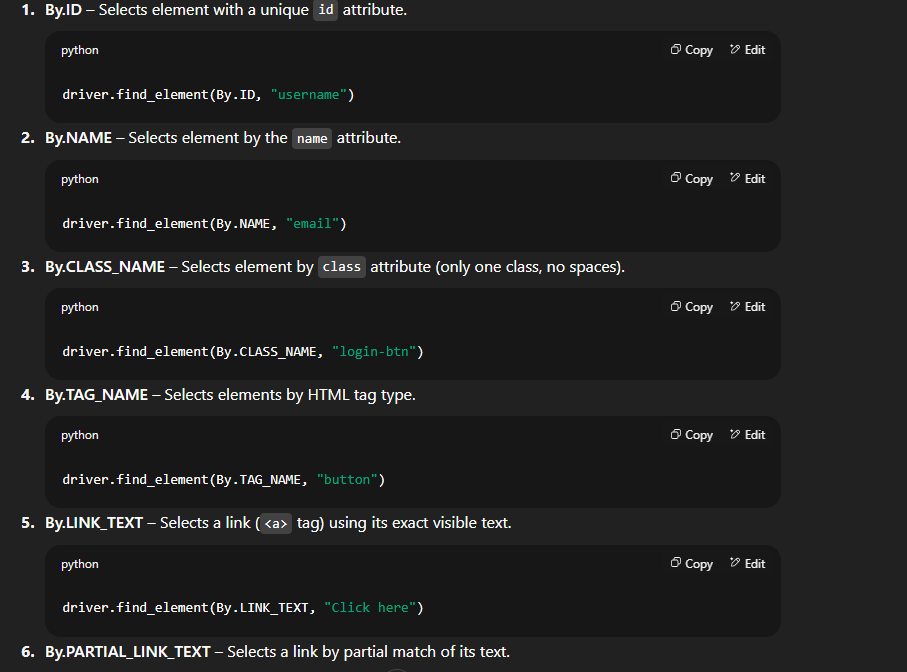
**Why?** To avoid common pitfalls.

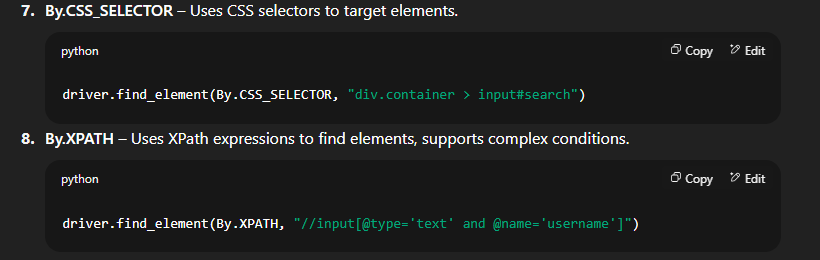
* Review bids that were **not awarded**.
* Look for patterns—e.g., lack of qualified vendors, unrealistic scope, budget issues.
* Use this to **refine our bid selection criteria**.

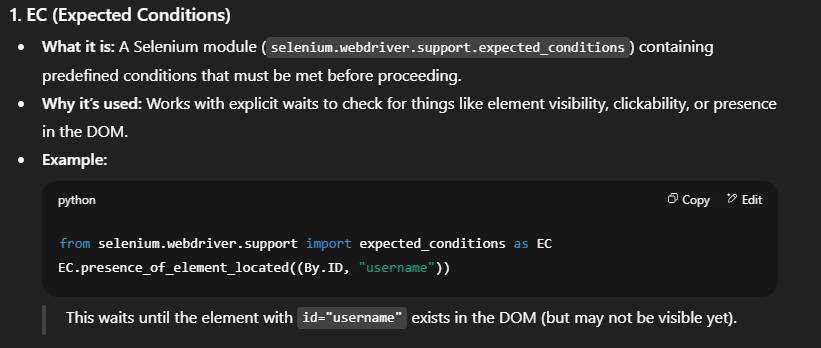


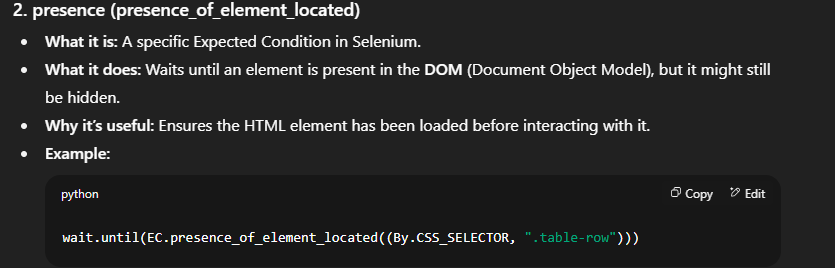


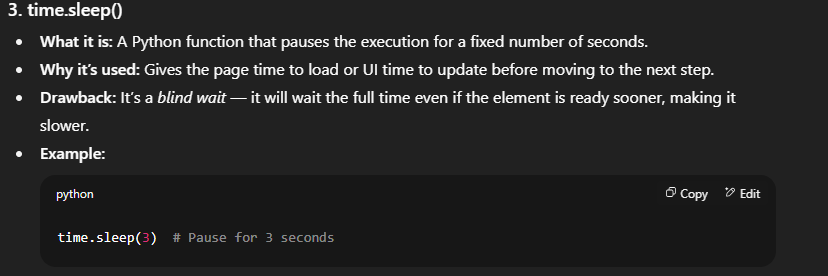










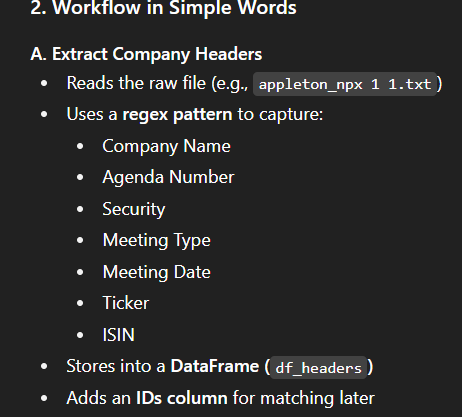


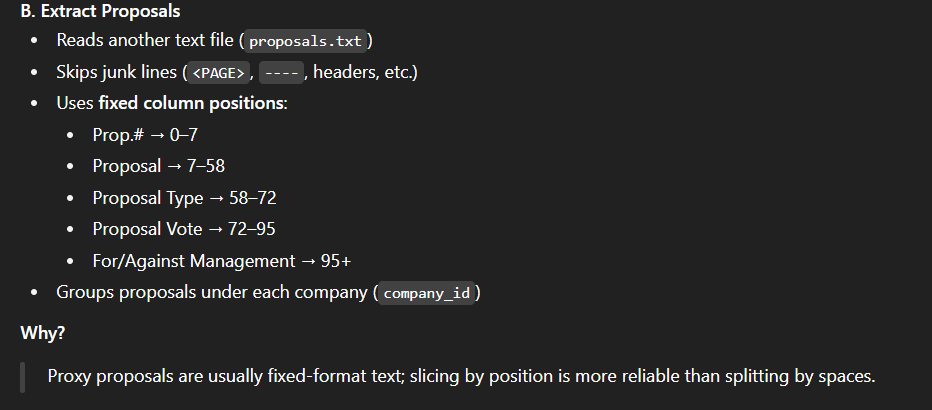
REGEX

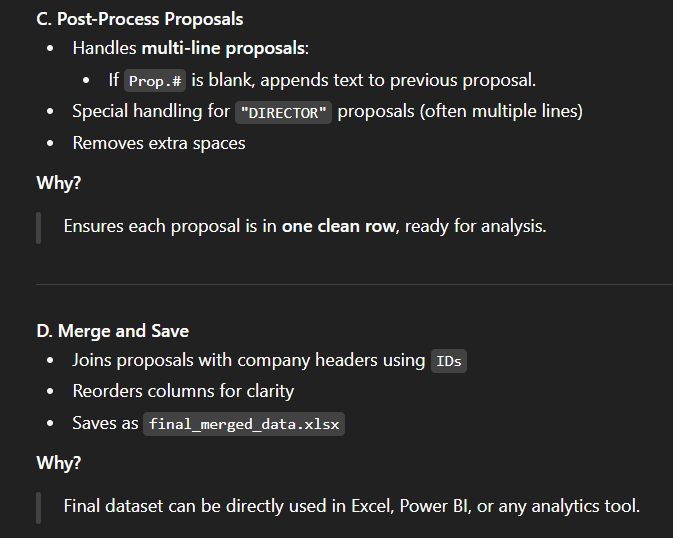
The dataset is derived from SEC NPX proxy voting text filings. Each record represents a proposal voted on during a company’s shareholder meeting.

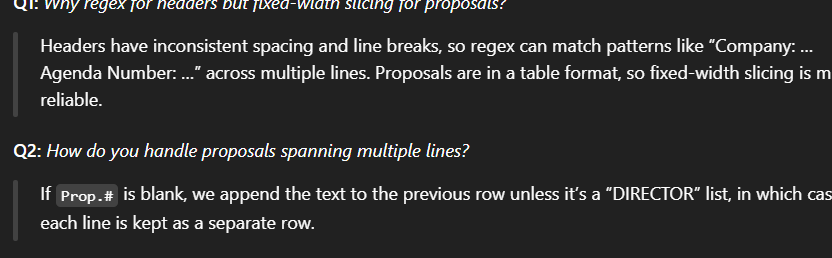
**It contains two levels of data:**  
**1. Company-level:** Company name, agenda number, meeting type & date, ticker symbol, ISIN, and security details.  
**2. Proposal-level:** Proposal number, description, type (e.g., Management, Shareholder), vote cast (For, Against, Abstain), and whether the vote aligned with management’s recommendation.

“The goal is to take raw proxy voting TXT files, extract both company header details and proposal-level voting data, clean and normalize them, then merge into a structured dataset for analysis.”









Insights

How does Appleton Funds typically vote on executive compensation proposals versus shareholder proposals?

Which companies have the most shareholder proposals?

Are there trends in voting patterns over time or across different industries?

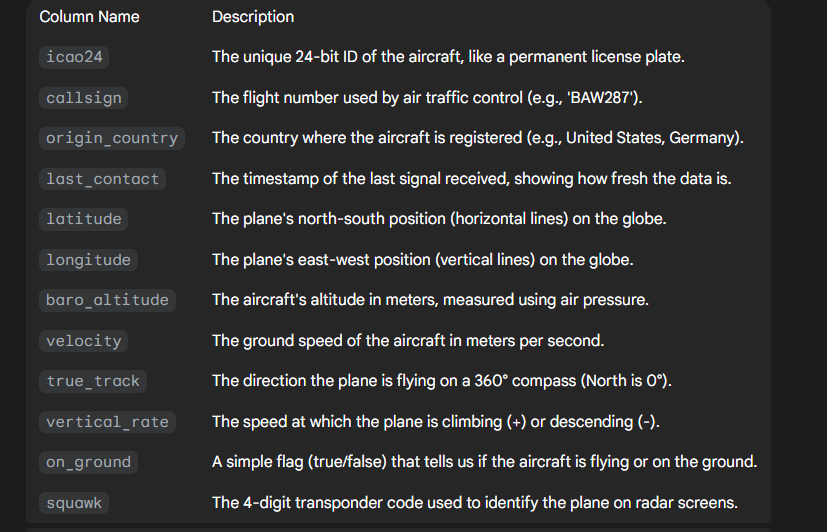
SQL

The OpenSky dataset captures comprehensive real-time and data for aircraft globally

**Group 1: Aircraft Identity: These first four columns tell us exactly who the plane is—its unique ID, flight number, and where it's registered."**

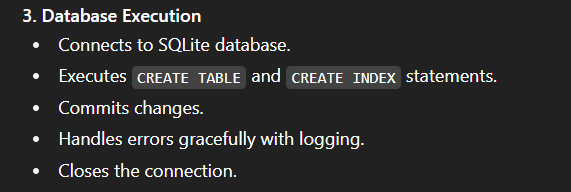
**Group 2: Position & Timing: This group tells us where the plane is—its precise location on the map, how high it is, and when we last heard from it."**

**Group 3: Flight Dynamics & Status:** **Finally, we have Flight Dynamics. This last group tells us what the plane is doing—how fast it's going, its direction, and whether it's in the air or on the ground."**

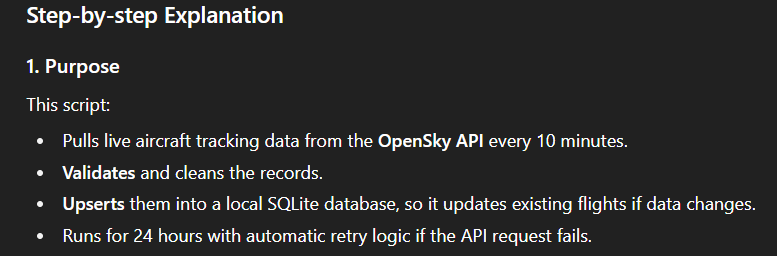
****

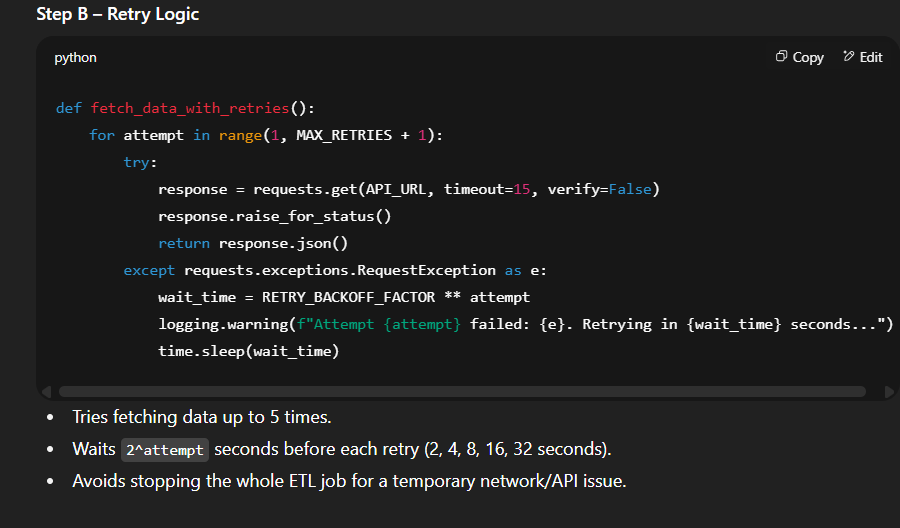
1 DB created





2nd file ETL





Insight 2

It ranks airlines based on how active they were during our data collection period — not by how many planes they own, but by how many times those planes were actually flying. The longer the bar, the busier the airline’s schedule.

It shows **who the big players are** in the market at the moment. If you’re competing in aviation, this tells you exactly who your biggest rivals are.

For companies that work with airlines — airports, fuel suppliers, maintenance services — this is a quick way to spot the *most active and most valuable customers*.

Insights 1

global air traffic

The line represents the total number of signals received from all aircraft during a specific hour,

It's measuring the total **volume** or **activity** of the traffic

The data shows a clear peak at **3:00 PM (15:00)** and a low point at **4:00 AM**. This allows airport and airline management to implement **demand-based scheduling**.

* **Insight:** You can schedule more air traffic controllers, ground crew, and gate agents for the afternoon rush and significantly reduce staff during the early morning lull.

Business Insight

the hourly trend is a clear proxy for passenger demand. Airlines can use this knowledge for **dynamic pricing strategies**, potentially increasing ticket prices for flights during peak demand hours and offering discounts during off-peak times to maximize revenue and fill seats.

Time:

04:00 UTC low → Europe still asleep, US in deep night, Asia just past peak.

09:00–15:00 UTC climb → Europe in full operation, US morning starting, Asia still active.

15:00–18:00 UTC peak → Europe & US overlap at full strength, transatlantic flights mid-route.

Insights 3

The key insight here is that we can now clearly see the world's most congested flight corridors. The data immediately draws our attention to two major hotspots: one over Western Europe and another over the Eastern United States.

Business

For an airline, this map is a direct guide to saving money. By planning routes that avoid these congested "red zones," they can **reduce fuel burn** from circling and **minimize costly delays**,