NOTES FOR PRODUCTION

* Open consolidated.csv
* Select All
* Format by column width

When cleansing the consolidated.csv file, you will need to

* Disregard the schema description on legacy.trade.gov
* Sort columns by **\_id** (Column A)
  + These are individual account numbers or identifiers. This will give you give you that intermingles the **ent\_num** from the **SDN AND OTHER SANCTIONS LISTS FROM OFAC AND TREASURY** with what appears to be hashed values that serve as individual account numbers for the rest of the entries in this set. Since this is messy, we will separate these values and create a **super\_id**.
  + Select All and Sort By **\_id** with Order Smallest to Largest.
* For fun, sort by **program** and get the count of **\_id**’s that are **SDN**. Then open the **SDN** file and verify the count against the Treasury’s data.
  + Make sure to Sort By **\_id** before going any further
* Insert Column into Column A

Before we go any further, let’s take a look at the **source** field of our data. Maybe later we can run a query to determine if some departments do not generate these values (some SDN hashed vs **ent\_num**(hypothetical)).

* Scroll down to where the hashed values begin and select the first row. *Control+Shift+Down* to highlight these values, and Sort By **source** with Order Smallest to Largest. Observe the source of these accounts.
  + While we’re down here, let’s take a look at the highest value of the ordered set. Take note of this for when we’re creating the **super­\_id**. (*later note:* it’s currently 7034489)
* Repeat selection for the above set. *Control+Shift+Up* to highlight these values, and Sort By **source** with Order Smallest to Largest and observe the data.

Now you are creating the **super\_id** which will serve as the new unique account identifier in this database.

* Since you are parsing the original **\_id** column and there are only two naming conventions, you will create two methods for generating the **super\_id**.
* Begin at the top. Select All and Sort By **\_id** with Smallest to Largest.
* In A1, type “=B1+4000000” (4 million). You are preserving the schema of the ent\_num to some extent so it might be easier to reference this account if more familiar with that identifier.
  + Double-Click the bottom-right square on the cell to Fill All to the bottom.
  + You will get the *#VALUES!* error on the hashed values. Delete these.
* To reserve room on the **super\_id** list for additional accounts imported, we will begin a different naming schema.
* In the first cell (aka. refcell1) with hashed values, type 8000000 (8 million).
* In the cell underneath, type = refcell1+1. Then autofill the rest of the fields by double-clicking the bottom-right square.

Beautiful, our **super\_id** has been created. For sanity’s sake let’s apply a Column Header.

* To retain this hashed data, let’s rename the **\_id** column to **minor­\_id**.

Now we can begin cleansing the rest of the data. Remember the source of your data, and the importance of what you are doing. This column represents the various SANCTIONS programs an individual or organization is associated with. Parsing this data is important now to stratify later.

* Bring attention to **programs**
  + In the SDN and other files, this column is named **program** – we would like to follow this convention, but will rename this Header to **program\_1**. We can make the same edit to **program** in the SDN list into another table later upon import.
  + \*\*NOTE that each account can belong to multiple programs and that they are separated by a semicolon.
* **GOING COWBOY**

Here’s where you use the tools at hand (I know it’s not ideal) to quickly separate these programs into separate columns for future analysis.

* We’re going to use Data > Text to Columns for these.
* Before starting, resize the right-adjacent column and the one to its right for good measure. It is large and makes inserting multiple columns tedious.
* Since we don’t know how many columns we need, we’re going to use Warnings as our guardrails. Insert 5 columns and test.
  + When run up against Warnings, add more columns.
  + If you are not going incrementally (if you are trying to get a sense of the range), you will need to Undo your work once you get a successful Text to Column and then proceed incrementally.
  + This will determine the current maximum count of programs that are assigned to an individual account. We don’t want to have extra columns in our schema unnecessarily, so for now let’s just determine the current max, and go incrementally until we don’t encounter the Warning.
* In Data > Text to Columns, Select *Delimited* and click Next. Then Select Semicolon as the Delimiter. Click Next and leave it as General and Finish. \*Make sure you confirm you have no additional columns\*
* NOTE: You will only need to add one additional column after you insert the original 5.

Now we need a naming convention for these headers – let’s start with **program\_1** and complete the rest, which will continue through **program\_8**.

* Sort by **program\_8** and observe the **name**. See if there is any information to be gained from observing the subject of 8 international sanctions.
  + You can also add additional Sort fields to see the rest of Top 10.

We want to anticipate that this may not be the upper limit for the number of SANCTIONS programs an account may be affiliated with. For good measure, let’s add two more columns with appropriate headers for a nice round **program\_10**.

Observe the **name** column and the format of its data. As we can see, it is not in first name/last name and there are also Aircraft, Entity, and Vessel fields. Let’s begin with Individuals. NOTE: Include an extra field for suffix when adding columns for name fields.

* For an easy way to do this (and really the only way), Select All and Sort By **type** in Order Z to A. We are looking for fields with the Individual value.
* Now we need to add columns to accommodate the new name fields. Add 2 columns right-adjacent to **name**.
* Select only the fields from the **name** field that are also attributed Individual – it may help to temporarily Hide a few columns so that you can see the beginning of the worksheet. \*Remember to Unhide.
  + With these, use Data > Text to Columns and Select Comma as the Delimiter. Finish.
* Now we can add the **last\_name**, **first\_name**, and **suffix\_name** to these headers. Sort by **suffix\_name** to explain your work.
* Starting with the Aircraft entries, Insert Column and name it **aircraft\_name**. From here we can carefully cut and paste our data between adjacent cells.

We are making some meaningful progress now and can already begin to see some of the data come together.

Inspect the **address** column, and observe that this data is not quite clean. For now, we’re going to do what we can to isolate the worst of it and parse the rest.

* Some of the data in here is separated by a semicolon. But on first glance there are many “C/O” references in here, especially between those delimiters. We want to exclude these from our **address** field because we are looking to isolate geographic data as best as possible.
  + We need to create a header for these outliers. Create a column at the end and call it **address\_extra**.
* Find and Select all field that contain “C/O” – Select All and apply a Cell Color.
* Escape the selection. Select All, then Sort On **address** with Order set to Cell Color.
* Select these fields within the **address** column

We will need to revisit the **address** field later, but for now let’s keep going.

I am not happy with **address** but for now it’s there. Let’s everything ready for MySQL.

I’m hoping that there will be enough clean data in the **address** field that I can still make some meaningful insights without parsing multiple international regions into new headers (example: *Room 201, 1296 Xuchang Road, Yangpu District, Shanghai, CN; China Communications Building, Block A, Desheng International, No. 85 Deshengmenwai Street, Xicheng District, Beijing, CN*).

We are also going to Data > Text to Column the **citizenships**, **dates\_of\_birth**, and **nationalities** columns.

* For **citizenships**,Sort By **citizenships** and determine the columns you will need (there are probably not many dual+ citizens). \*spoiler: 3 total columns.
* For **dates\_of\_birth**, Sort By **dates\_of\_birth** and determine the number of columns you will need. \*Spoiler: it’s 9.
* For **nationalities**, Sort By **nationalities** and determine the number of columns you will need. \*Spoiler: it’s 4.

Let’s clean some of the values we separated from the **dates\_of\_birth** Column by renaming the original column **dob\_1** and continuing through **dob\_9**. Select All and Sort By **dob\_1** and Order Oldest to Newest. Observe the data. Some are dates, other are ranges, and others are an estimate (“Circa”).

* To be honest, I’m not quite set on the best way of formatting these custom date ranges from text to DATE format. We’re going to take the entries that say “YEAR to YEAR” and for now store them as a string value in a new **dob\_range\_string** column.
* Select All and Sort By **dob\_1** in Order Smallest to Largest. Select and Cut the fields written in “YYYY to YYYY” format and paste into the **dob\_range\_string** column.
* Search and Replace “CIRCA ” with “” to isolate the integer
* Now Select All and Sort By dob\_1. Select All **dob\_#** columns and Format Cells as YYYY-MM-DD.

Now we need to name these new columns. Since we have already renamed some and will likely do so in some remaining columns, the below list should serve as a comprehensive summary of the changes we make to the Source Headers:

Header Columns that were renamed:

* **\_id** becomes **minor\_id**
* **entity\_num** becomes **ent\_num**.
* **source** becomes **account\_source.**
* **type** becomes **account\_type**.
* **programs** is split and becomes **program\_1** and continues through **program\_10**.
* **addresses** becomes **address\_1**. This may need to be parsed into more columns later.
* **name** is split and becomes **last\_name**, **first\_name**, **suffix\_name**, **aircraft\_name**, **entity\_name**, **vessel\_name**, and **unknown\_account\_type\_name**.
* **start\_date** and **end\_date** become **date\_start** and **date\_end**, respectively.
* **citizenships** is split and becomes **citizenship\_1** and continues through **citizenship\_3**.
* **dates\_of\_birth** is split and becomes **dob\_1** and continues through **dob\_9**.
  + To account for date ranges, we will also add **dob\_range**.
* **nationalities** is split and becomes **nationality\_1** and continues through **nationality\_3**.
* **ids** becomes **misc\_ids**.

Header Columns that were added:

* **super\_id**
* **address\_extra** (this is the field where we will place the “C/O” **address\_1** results)

Let’s begin writing our query to define our tables. I’ll be writing in VS Code and opening within MySQL. To note, some of the Header names in the Source file are reserved words in SQL. This is why some of the above changes were made to the dataset. NOTE: You will need to make sure that localhost privileges are set up if you are hosting your MySQL instance locally. (“*SET GLOBAL local\_infile=1;*”)

SCHEMA:

SET GLOBAL local\_infile=1;

CREATE TABLE blacklist(

super\_id INT NOT NULL,

minor\_id VARCHAR(500),

account\_source VARCHAR (500),

ent\_num INT,

account\_type VARCHAR(150),

program\_1 VARCHAR(100),

program\_2 VARCHAR(100),

program\_3 VARCHAR(100),

program\_4 VARCHAR(100),

program\_5 VARCHAR(100),

program\_6 VARCHAR(100),

program\_7 VARCHAR(100),

program\_8 VARCHAR(100),

program\_9 VARCHAR(100),

program\_10 VARCHAR(100),

aircraft\_name VARCHAR(100),

entity\_name VARCHAR(250),

vessel\_name VARCHAR(250),

unknown\_account\_type\_name VARCHAR(500),

last\_name VARCHAR(500),

first\_name VARCHAR(500),

suffix\_name VARCHAR(20),

title VARCHAR(500),

address\_1 VARCHAR(1000),

federal\_register\_notice VARCHAR(800),

date\_start DATE,

date\_end DATE,

standard\_order VARCHAR(10),

license\_requirement VARCHAR(500),

license\_policy VARCHAR(500),

call\_sign VARCHAR(25),

vessel\_type VARCHAR(100),

gross\_tonnage INT,

gross\_registered\_tonnage INT,

vessel\_flag VARCHAR(100),

vessel\_owner VARCHAR(100),

remarks VARCHAR(1001),

source\_list\_url VARCHAR(50),

alt\_names VARCHAR(500),

citizenship\_1 VARCHAR(500),

citizenship\_2 VARCHAR(10),

citizenship\_3 VARCHAR(10),

dob\_1 DATE,

dob\_2 DATE,

dob\_3 DATE,

dob\_4 DATE,

dob\_5 DATE,

dob\_6 DATE,

dob\_7 DATE,

dob\_8 DATE,

dob\_9 DATE,

dob\_range\_string VARCHAR(255),

nationality\_1 VARCHAR(500),

nationality\_2 VARCHAR(500),

nationality\_3 VARCHAR(500),

nationality\_4 VARCHAR(500),

places\_of\_birth VARCHAR(500),

source\_information\_url VARCHAR (150),

misc\_ids VARCHAR(1000),

address\_extra VARCHAR(1000),

PRIMARY KEY (super\_id)

);