Data Refinery Lab

Introduction

This lab will introduce the Data Refinery. Data Refinery is a self-service data preparation tool for data scientists, data engineers, and business analysts. Data Refinery provides profiling, visualization, and a robust set of transforms to prepare data for analytics purposes. You will use the 3 Female Human Trafficking data sets in this lab to demonstrate data profiling, data visualization, and data preparation capabilities of the Data Refinery tool.

End-to-End Data Science

The general flow of the End to End Data Science PoT will be guided by the activities shown in Figure 1- End to End Flow. This lab will focus on the Prepare Data activity.

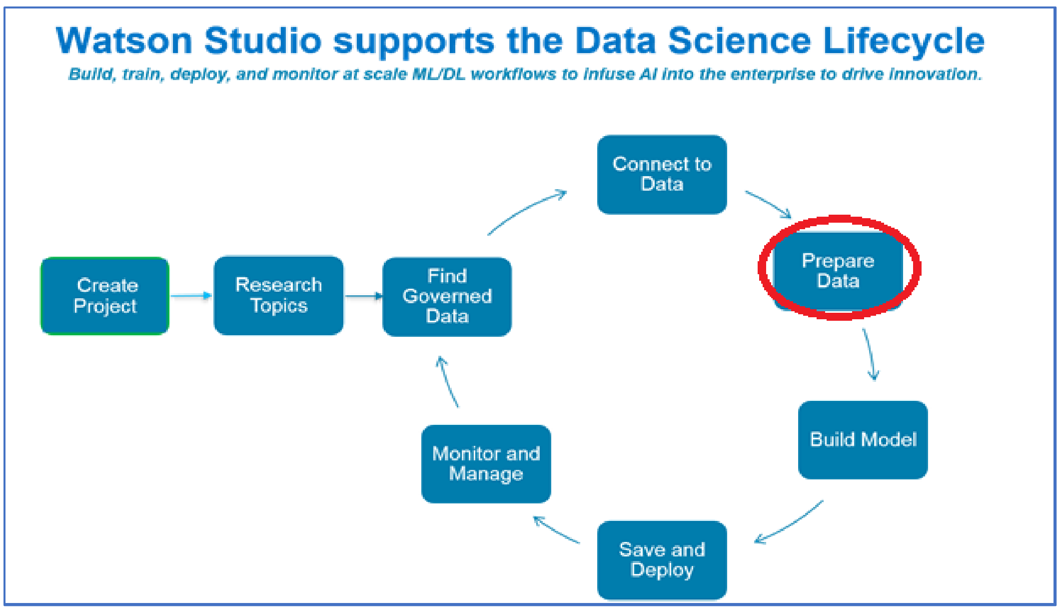


Figure 1- End to End Flow

Objectives

The goal of the lab is for the users to gain familiarity with the features of the Data Refinery. We will perform the following Data Refinery tasks:

- Create a new Data Flow
- Profile the data
- Visualize the data to gain a better understanding
- Prepare the data for modeling
- Run the sequence of data preparation operations on the entire data set.

The Create a new Data Flow task will be completed first, and the Run the sequence task will be completed last. The Profile, Visualize, and Prepare tasks will be intermixed.

Female Human Trafficking Data

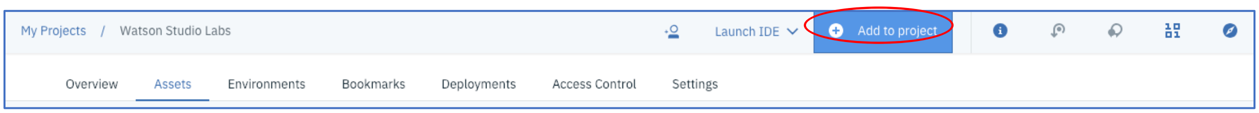
The data sets used for this lab consist of simulated travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING _LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL of low, medium, or high risk. Others have not yet been vetted.

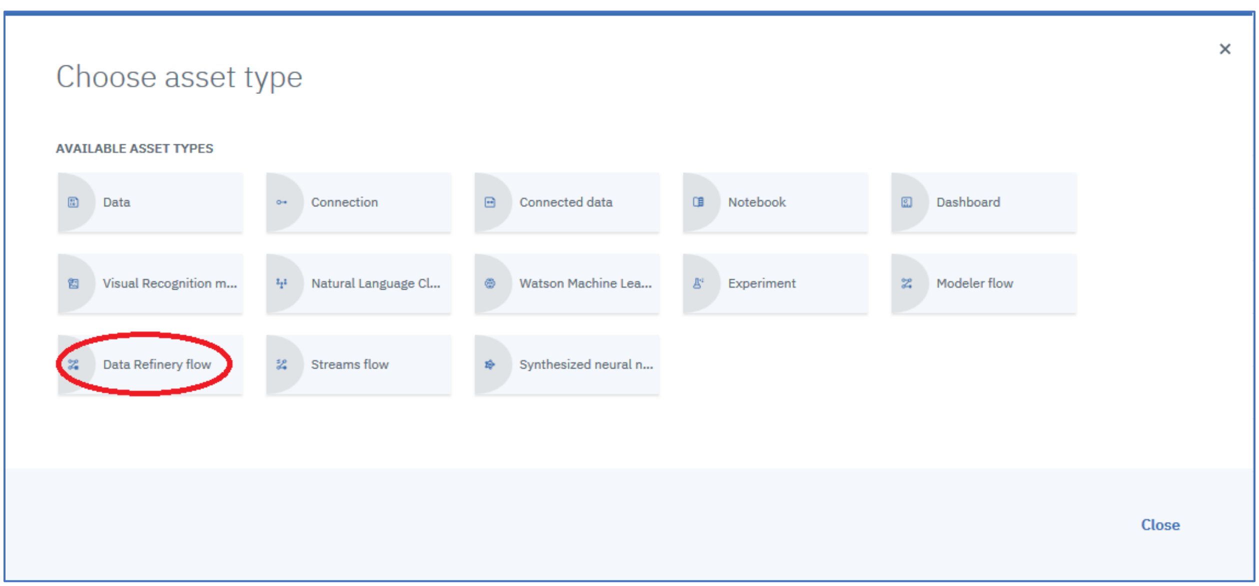
The OCCUPATION data included in the travel data is very granular. For modeling purposes, it was decided to categorize the OCCUPATION data. Two additional datasets are used for this purpose. The occupation.csv dataset maps the granular occupation data to a category code. The categories dataset maps a category code to a category description. These datasets will be joined to the main dataset to prepare the data for modeling.

Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

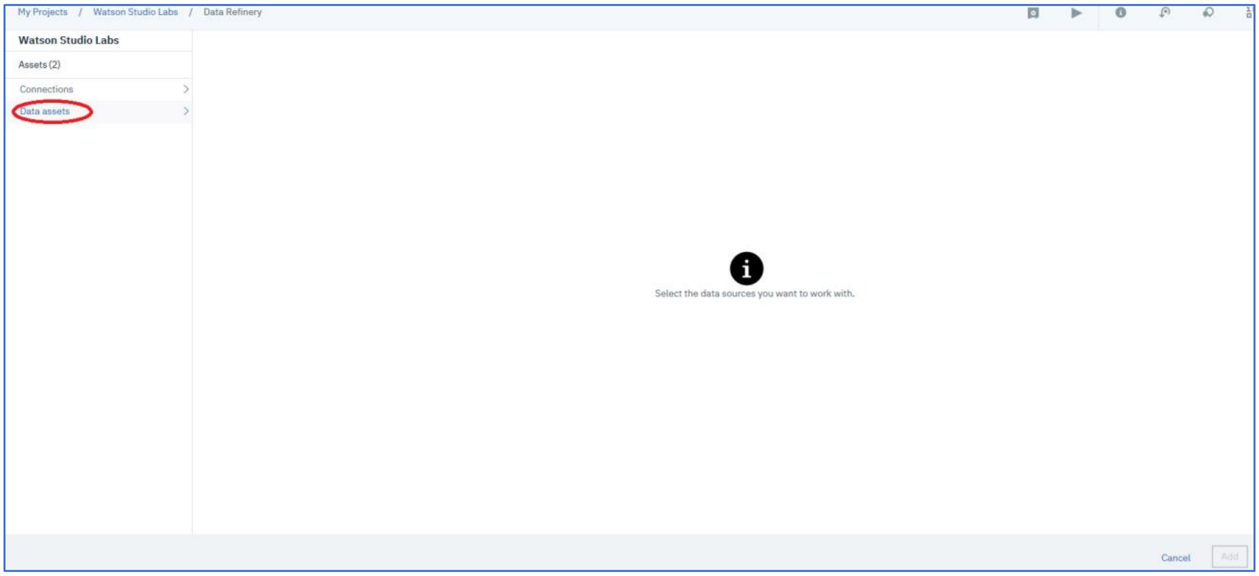
Create a new Data Flow

1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow.**

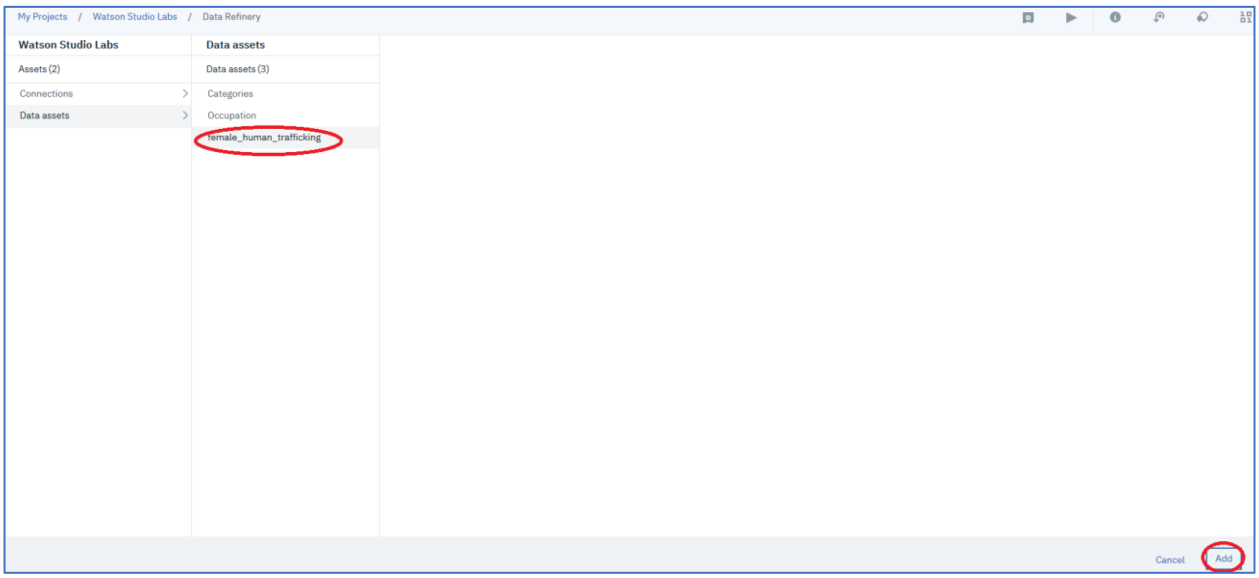




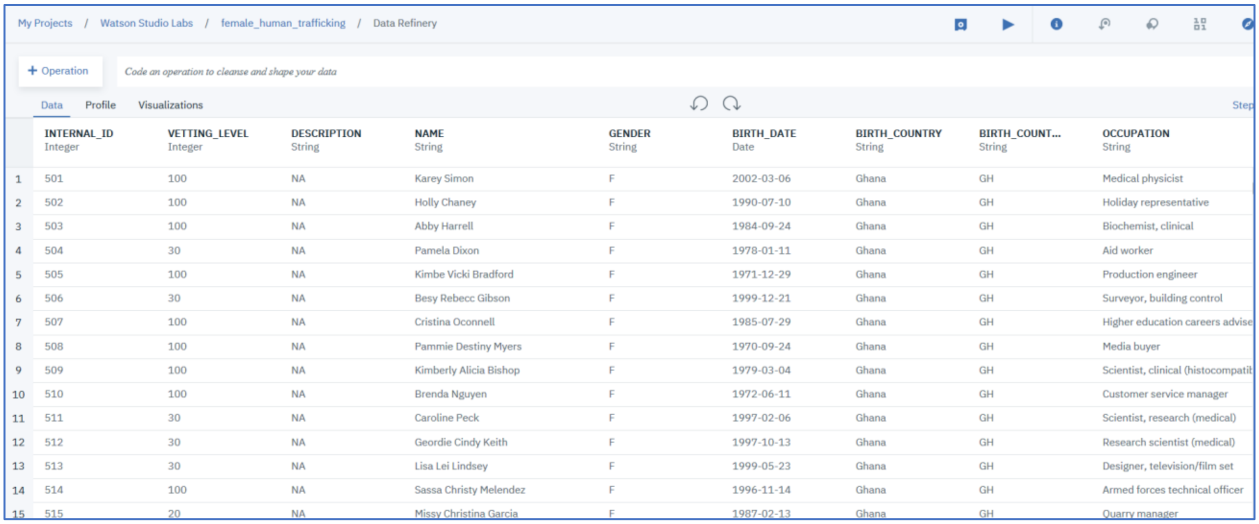
2. Click on **Data Assets**.



3. Click on **female_human_trafficking**, and then click on **Add**.



4. A sample of the data set (1000 rows) will be displayed.

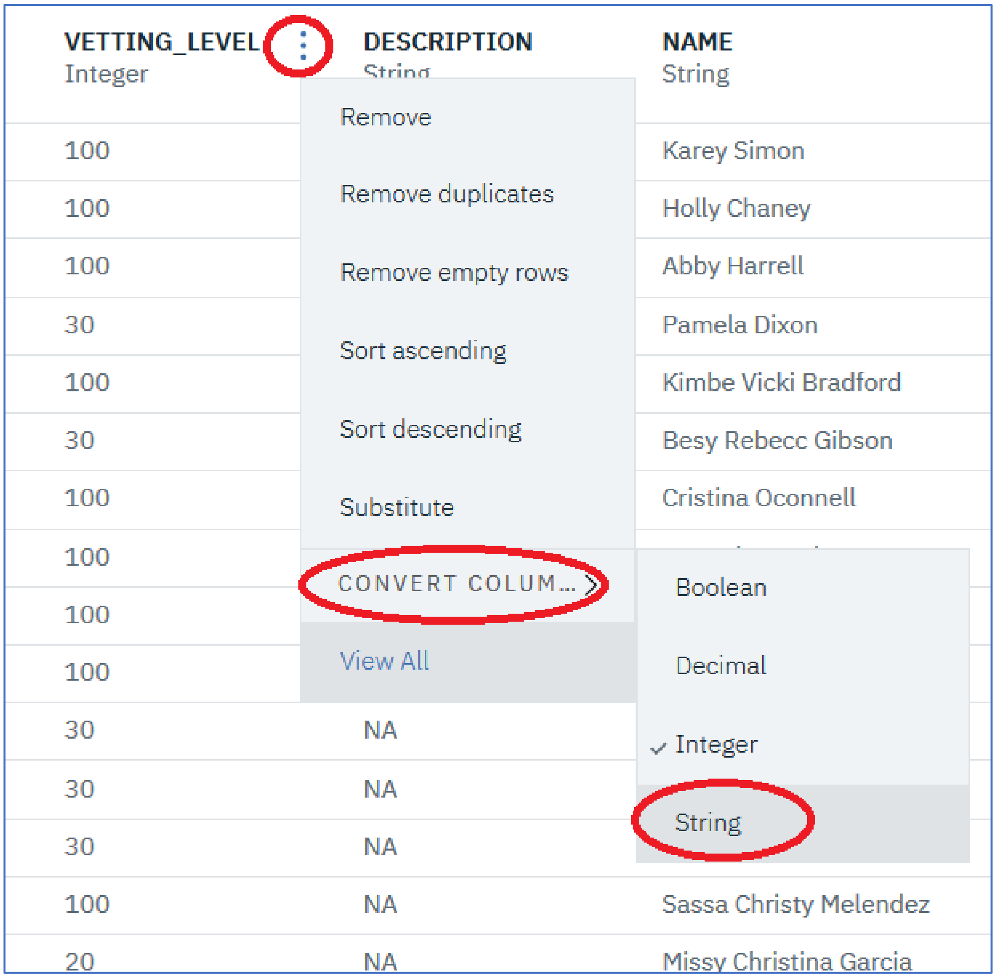


	INTERNAL_ID Integer	VETTING_LEVEL Integer	DESCRIPTION String	NAME String	GENDER String	BIRTH_DATE Date	BIRTH_COUNTRY String	BIRTH_COUNT... String	OCCUPATION String
1	501	100	NA	Karey Simon	F	2002-03-06	Ghana	GH	Medical physicist
2	502	100	NA	Holly Chaney	F	1990-07-10	Ghana	GH	Holiday representative
3	503	100	NA	Abby Harrell	F	1984-09-24	Ghana	GH	Biochemist, clinical
4	504	30	NA	Pamela Dixon	F	1978-01-11	Ghana	GH	Aid worker
5	505	100	NA	Kimbe Vicki Bradford	F	1971-12-29	Ghana	GH	Production engineer
6	506	30	NA	Besy Rebecc Gibson	F	1999-12-21	Ghana	GH	Surveyor, building control
7	507	100	NA	Cristina Oconnell	F	1985-07-29	Ghana	GH	Higher education careers advise
8	508	100	NA	Pammie Destiny Myers	F	1970-09-24	Ghana	GH	Media buyer
9	509	100	NA	Kimberly Alicia Bishop	F	1979-03-04	Ghana	GH	Scientist, clinical (histocompatil
10	510	100	NA	Brenda Nguyen	F	1972-06-11	Ghana	GH	Customer service manager
11	511	30	NA	Caroline Peck	F	1997-02-06	Ghana	GH	Scientist, research (medical)
12	512	30	NA	Geordie Cindy Keith	F	1997-10-13	Ghana	GH	Research scientist (medical)
13	513	30	NA	Lisa Lei Lindsey	F	1999-05-23	Ghana	GH	Designer, television/film set
14	514	100	NA	Sassa Christy Melendez	F	1996-11-14	Ghana	GH	Armed forces technical officer
15	515	20	NA	Missy Christina Garcia	F	1987-02-13	Ghana	GH	Quarry manager

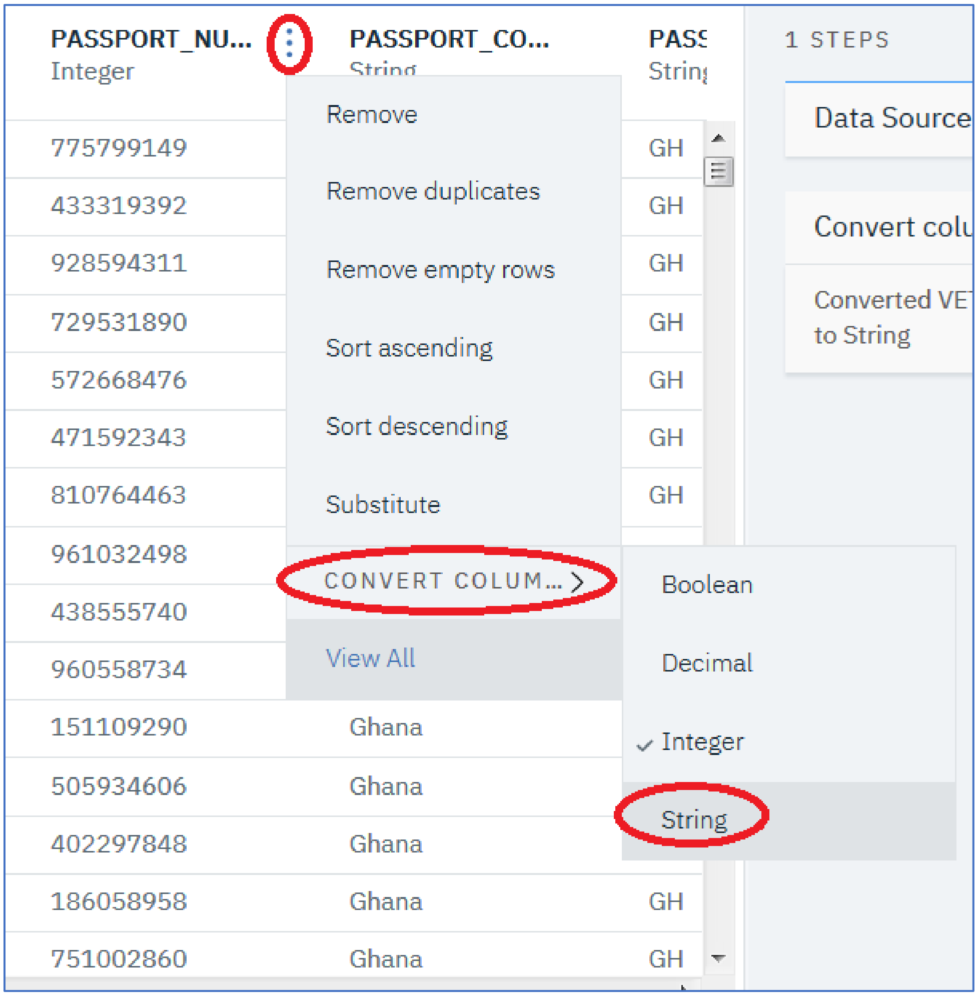
Prepare, Profile, Visualize

Before profiling the data, we will do some data preparation. Note, skip steps 1-4 if both the VETTING_LEVEL column and the PASSPORT_NUMBER column are Strings.

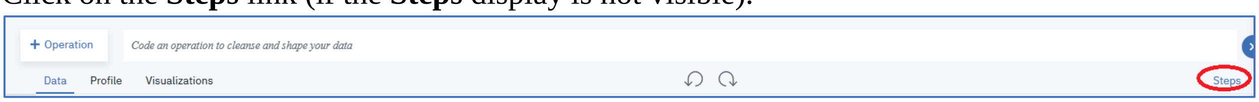
1. Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the VETTING_LEVEL column by hovering over VETTING_LEVEL, clicking on the vertical ellipse ⋮, clicking on CONVERT COLUMN, and clicking on String.



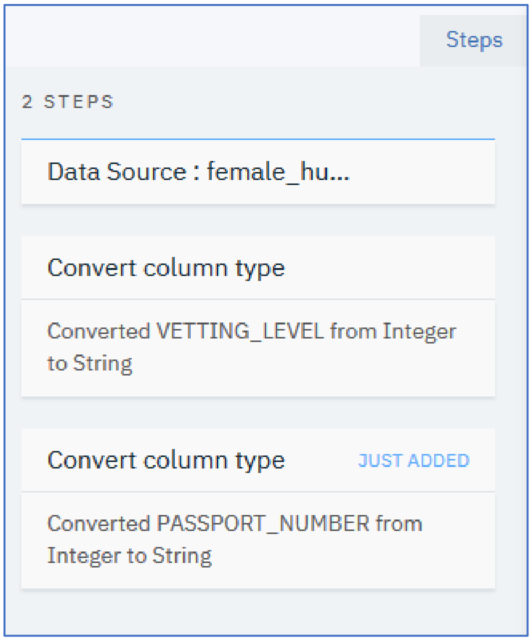
2. Convert the PASSPORT_NUMBER column by hovering over PASSPORT_NUMBER, clicking on the vertical ellipse ⋮, clicking on CONVERT COLUMN, and clicking on String.



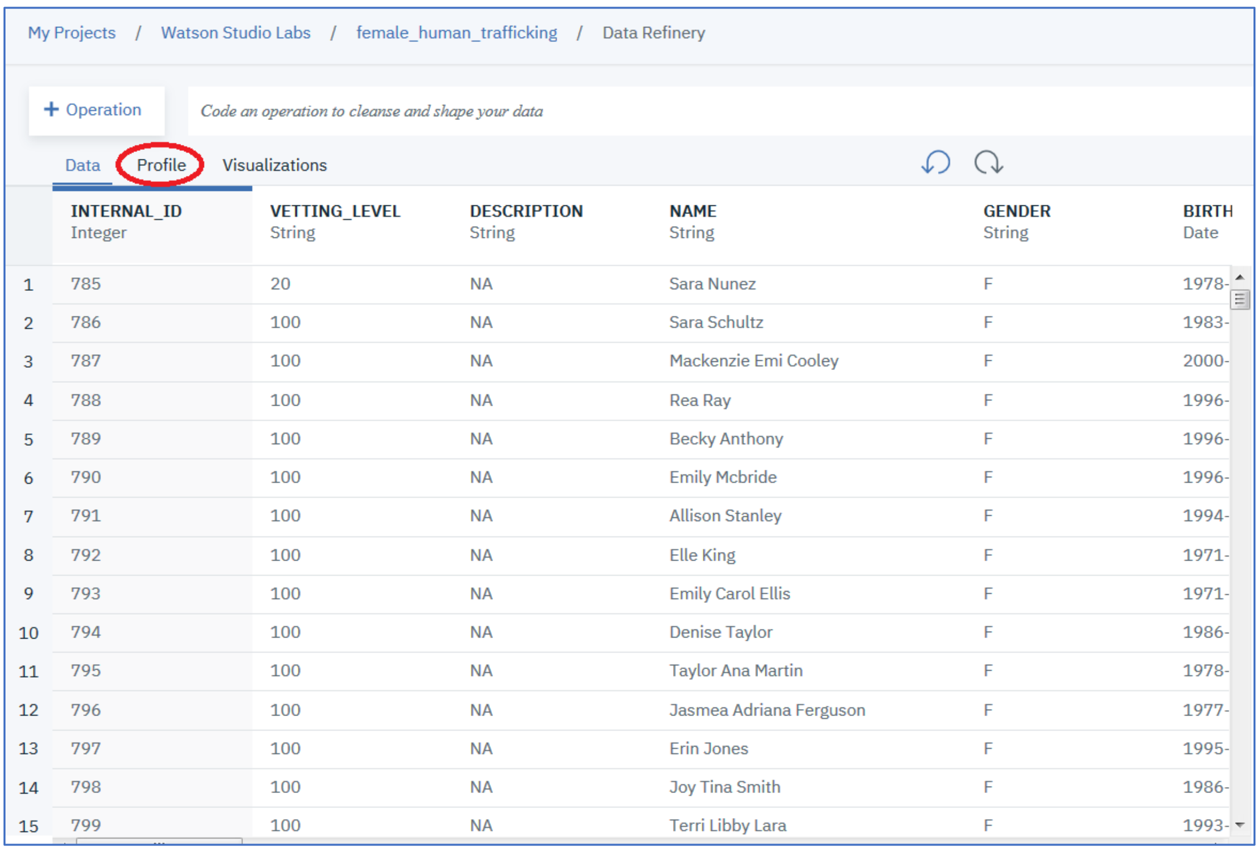
3. Click on the **Steps** link (if the **Steps** display is not visible).



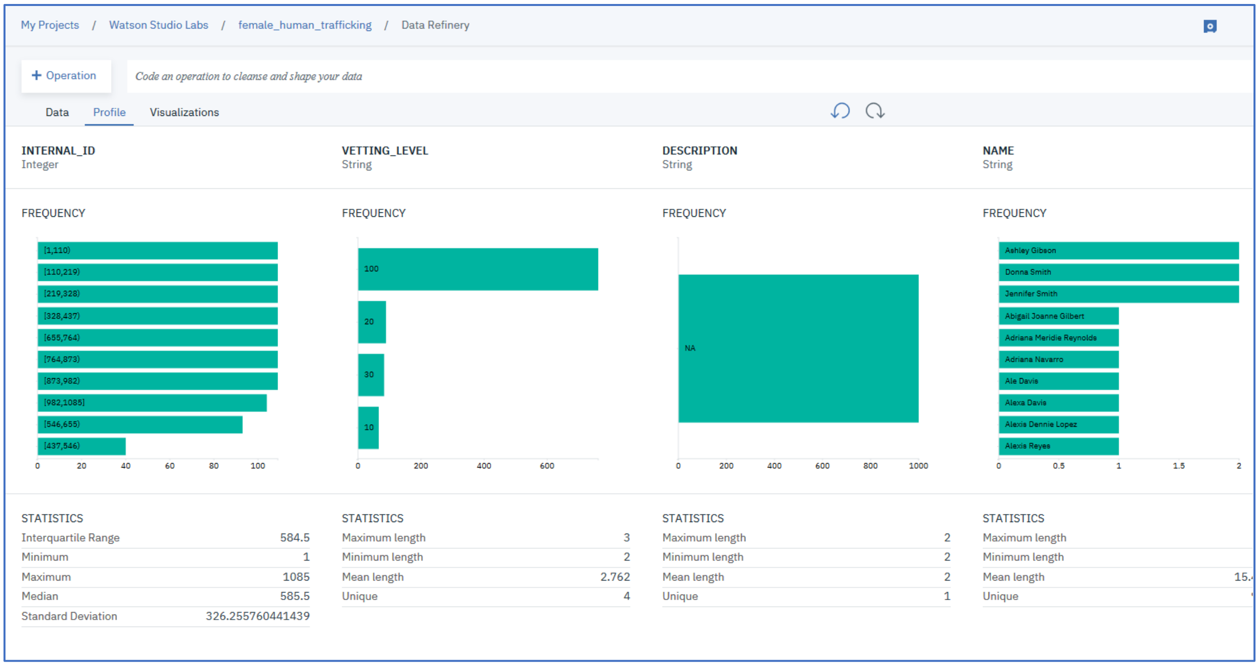
4. Each data operation is recorded in the **Steps** display providing an audit list of the operations performed. So far, we have done two column conversion operations. The steps in the **Steps** display can be edited. Operations can be removed from the list or modified.



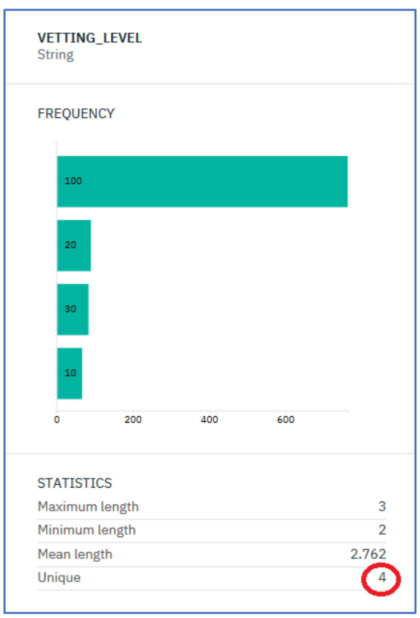
5. Click on **Profile**.



6. The Profile panel displays the counts of the top 10 values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column.

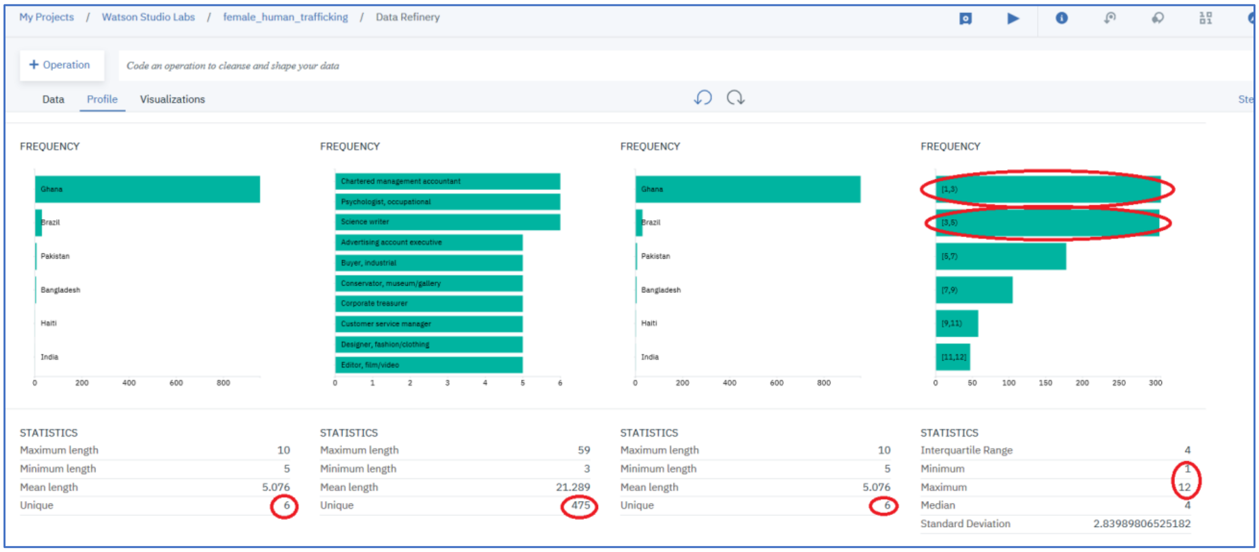


7. The statistics for the VETTING_LEVEL column show 4 unique values, 10, 20, 30, and 100. These are coded values that correspond to risk of trafficking, 10-High Risk, 20-Medium Risk, 30-Low Risk, and 100- has not been vetted yet. As the graph shows below, most of the data records have not been vetted yet. We will use the data that has been vetted to train a model to predict the risk for the unvetted records in subsequent labs.

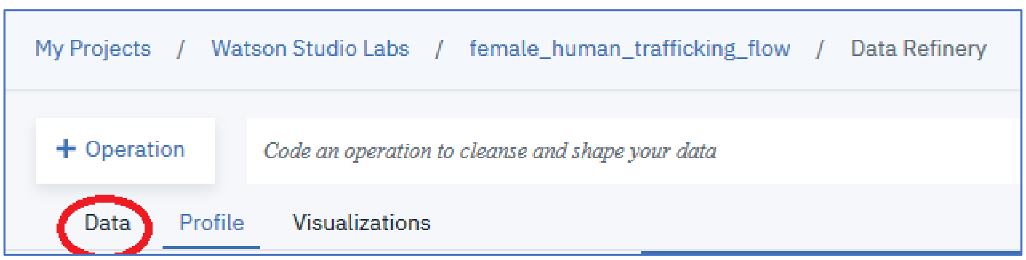


8. Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH_COUNTRY, and PASSPORT_COUNTRY shows

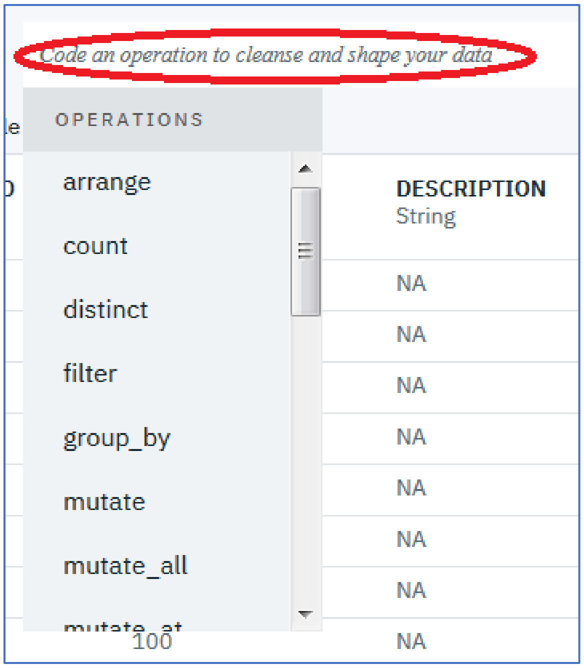
only 6 unique countries. The COUNTRIES_VISITED_COUNT shows that passengers have visited between 1 and 12 countries, with passengers visiting between 1 and 3 countries and between 3 and 5 countries the most prevalent. Note, the results may be slightly different on your screen.



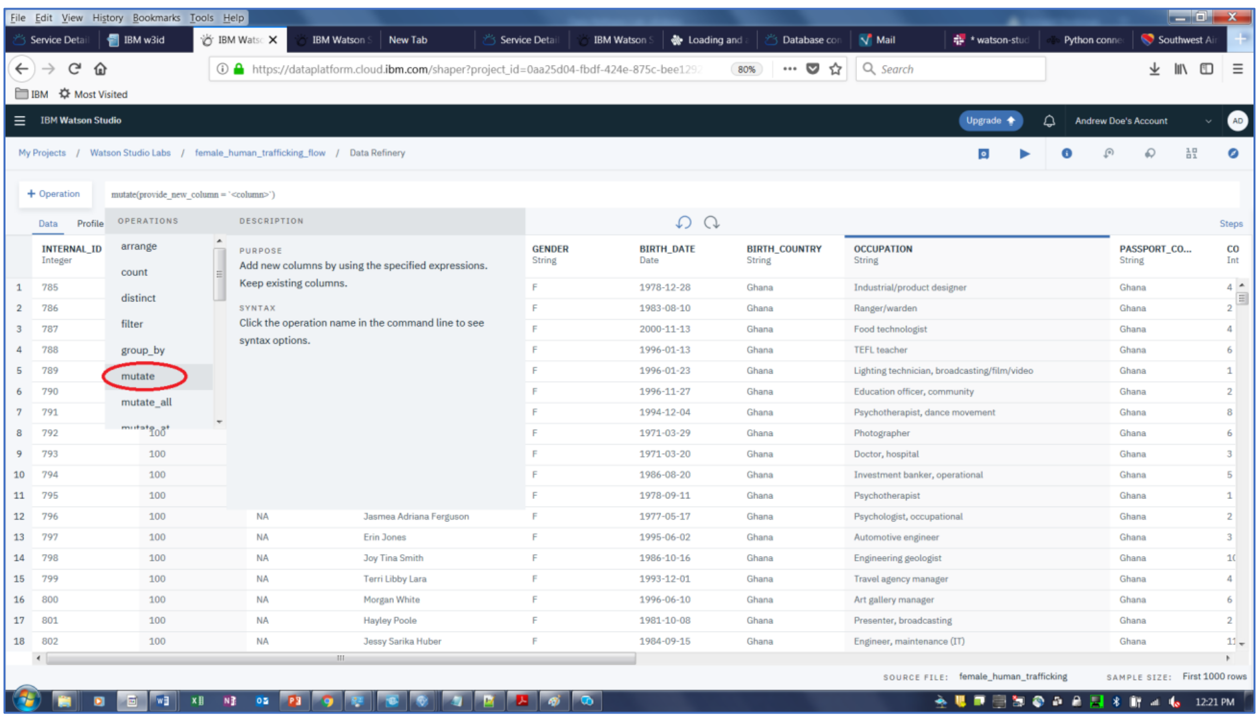
9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.



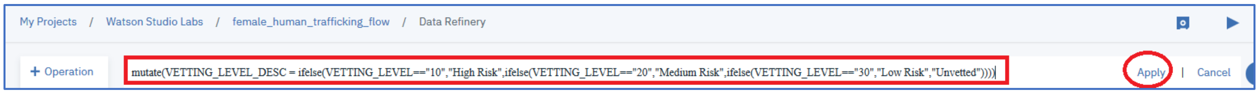
10. Let's make the VETTING_LEVEL column more readable, by mapping the code to a description. The Data Refinery is a front-end to the R package dplyr. We will convert the coded values 10,20,30,100 to "High Risk", "Medium Risk", "Low Risk", and "Unvetted". We will use the mutate and ifelse functions to do the conversion. Click on **Code an operation to cleanse and shape your data.** Several operations are available.



11. Hover the mouse over **mutate.** A description of the mutate function is provided.



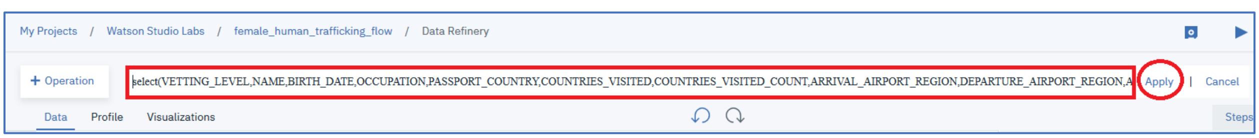
12. Click on **mutate** and cut and replace the generated code with the following and then click **Apply**.
mutate(VETTING_LEVEL_DESC = ifelse(VETTING_LEVEL=="10","High Risk",ifelse(VETTING_LEVEL=="20","Medium Risk",ifelse(VETTING_LEVEL=="30","Low Risk","Unvetted"))))



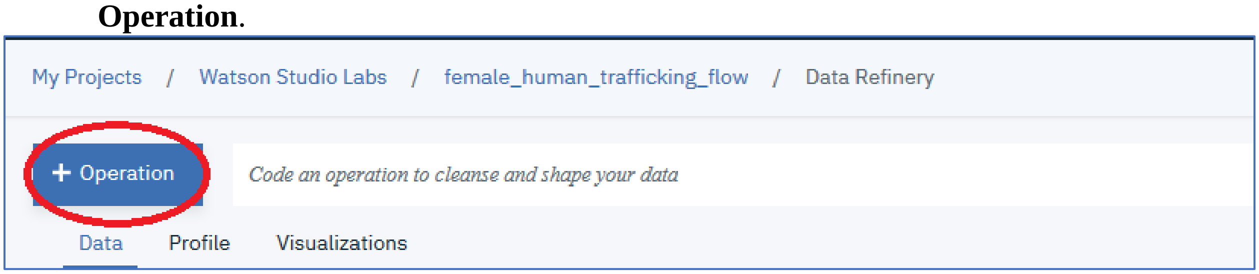
13. If you scroll to the right you should see the new column VETTING_LEVEL_DESC with values “Low Risk”, “Medium Risk”, “High Risk”, and “Unvetted”.

14. Let’s extract the fields of interest by using another dplyr function, **select**. Cut and paste the following code into the operations area.

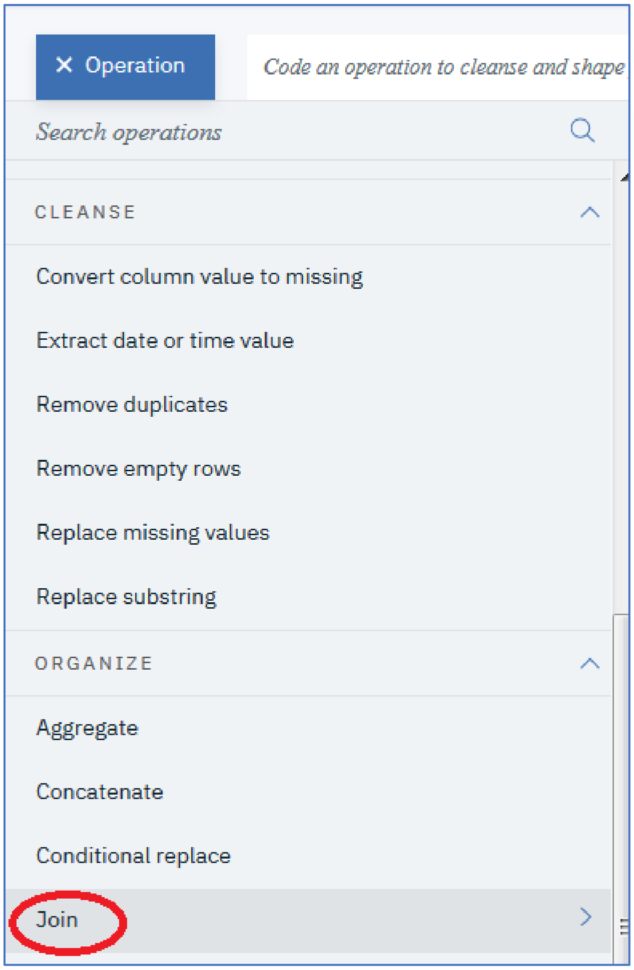
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DEPARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)



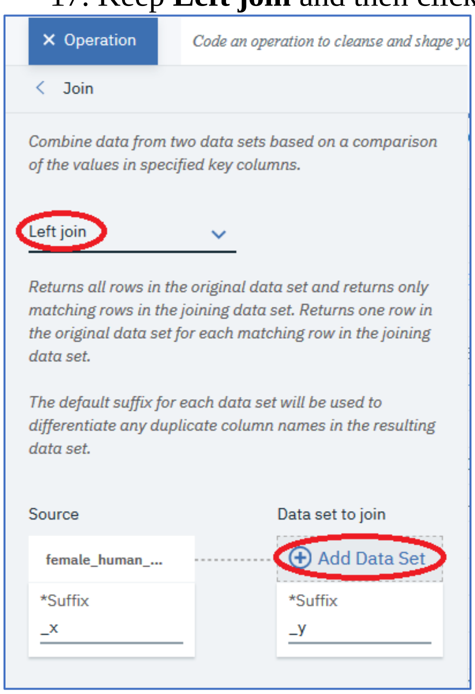
15. Let’s now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the categories dataset. Click on + **Operation**.



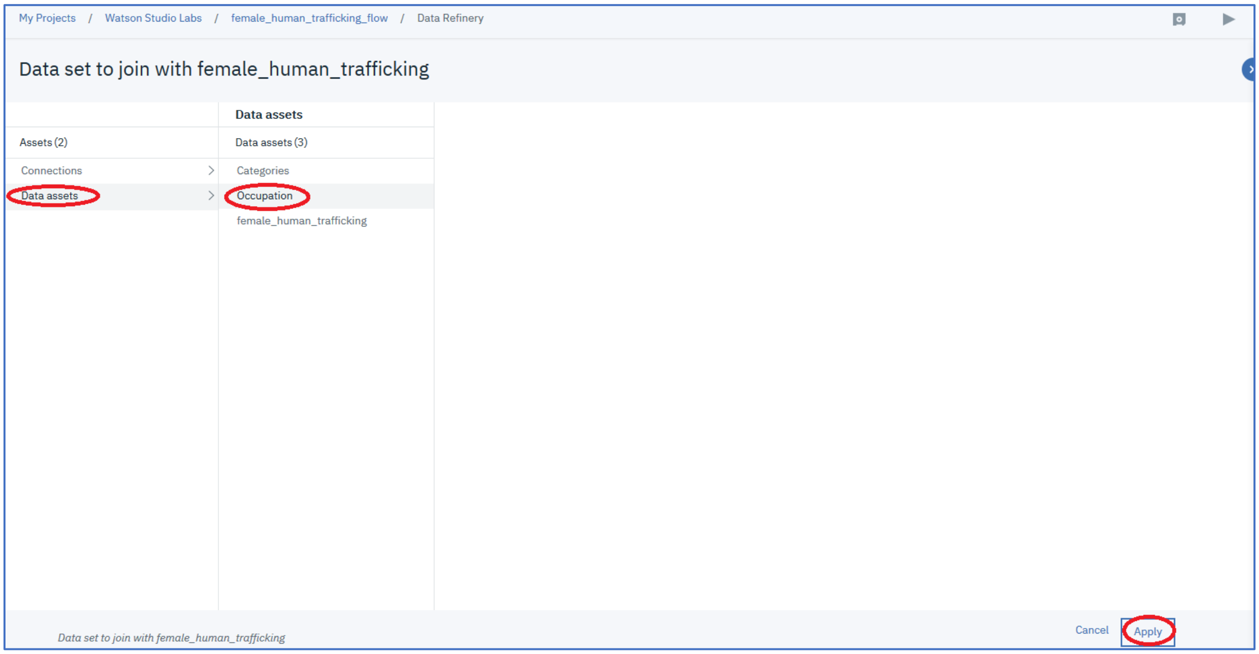
16. Scroll down and click on **Join**.



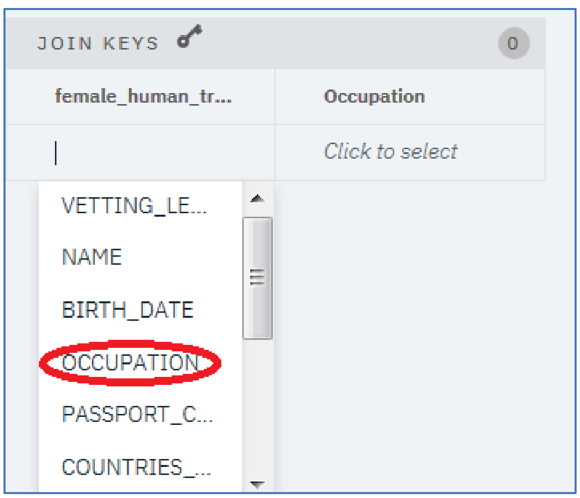
17. Keep **Left join** and then click on **Add Data Set**



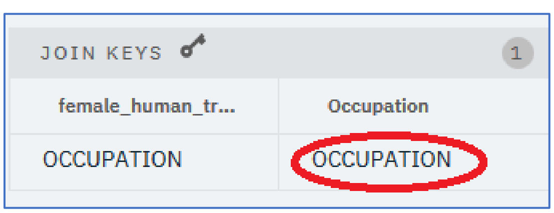
18. Click on **Data Assets**, click on **Occupation**, and then click **Apply**.



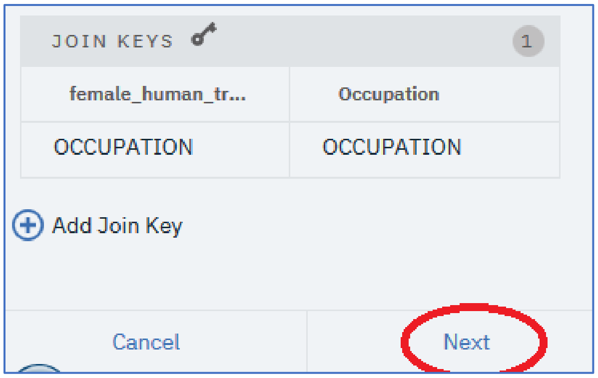
19. In **JOIN KEYS** under **female_human_trafficking** click **Click to select**, and then click **OCCUPATION.**



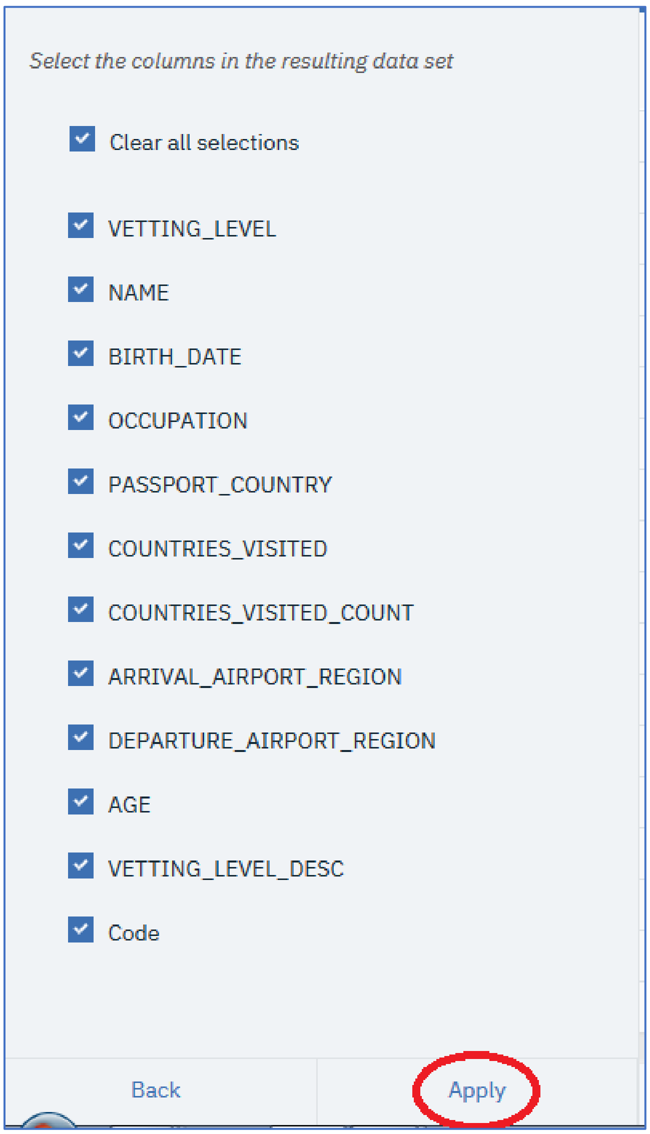
20. In **JOIN KEYS** under **Occupation** click **Click to select** and then click **OCCUPATION**.



21. Click on **Next**.

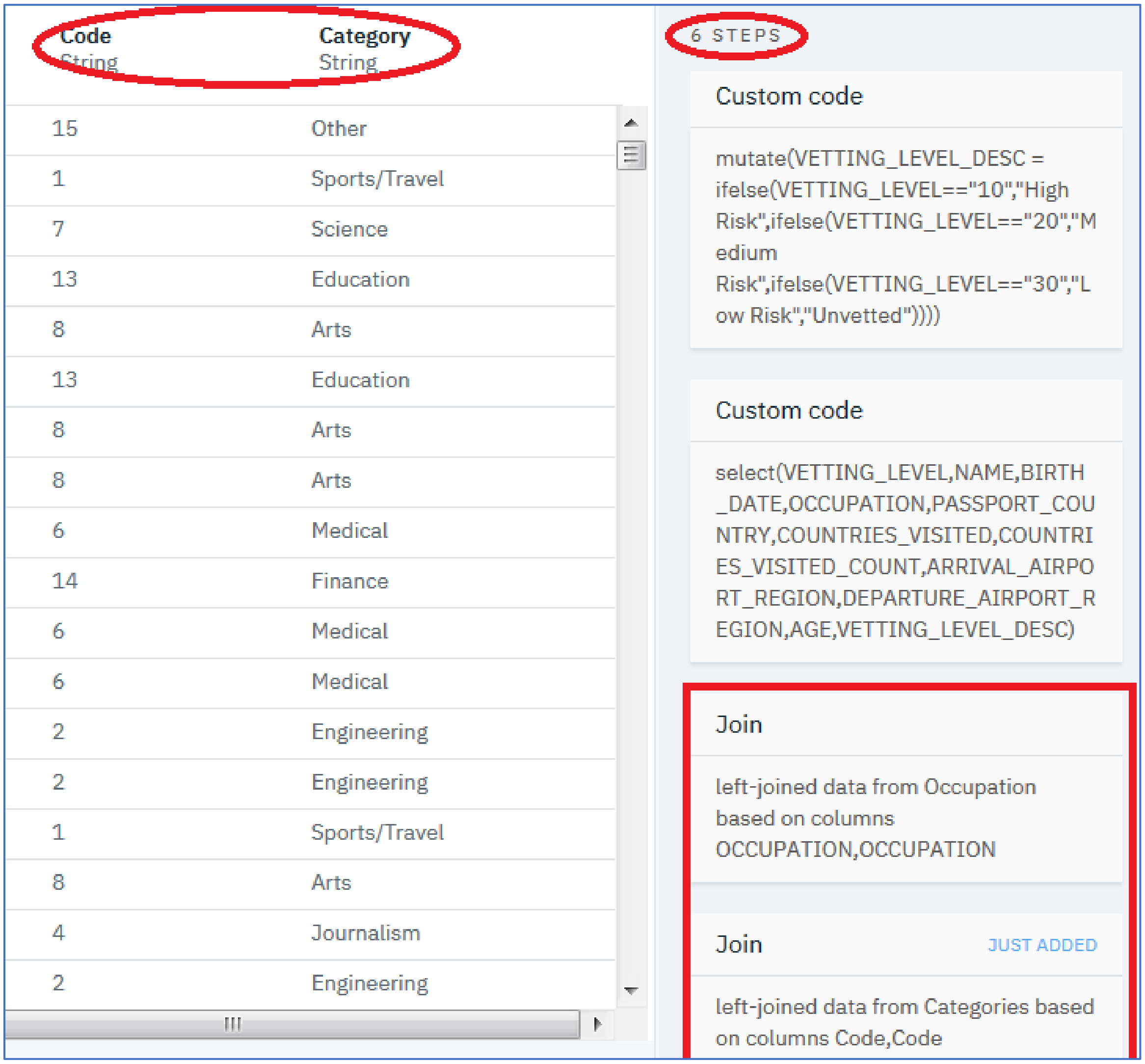


22. Click **Apply**.

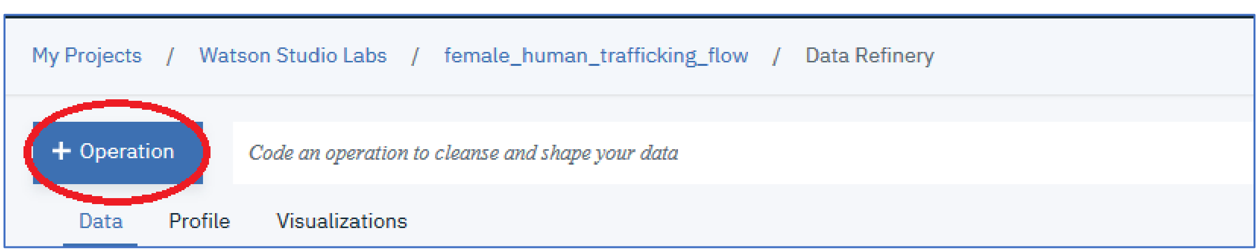


23. Follow steps 19-22 to join the Categories dataset. The join keys are the Code fields in both datasets. As a result of the joins, two new columns are added, a Code column, and a

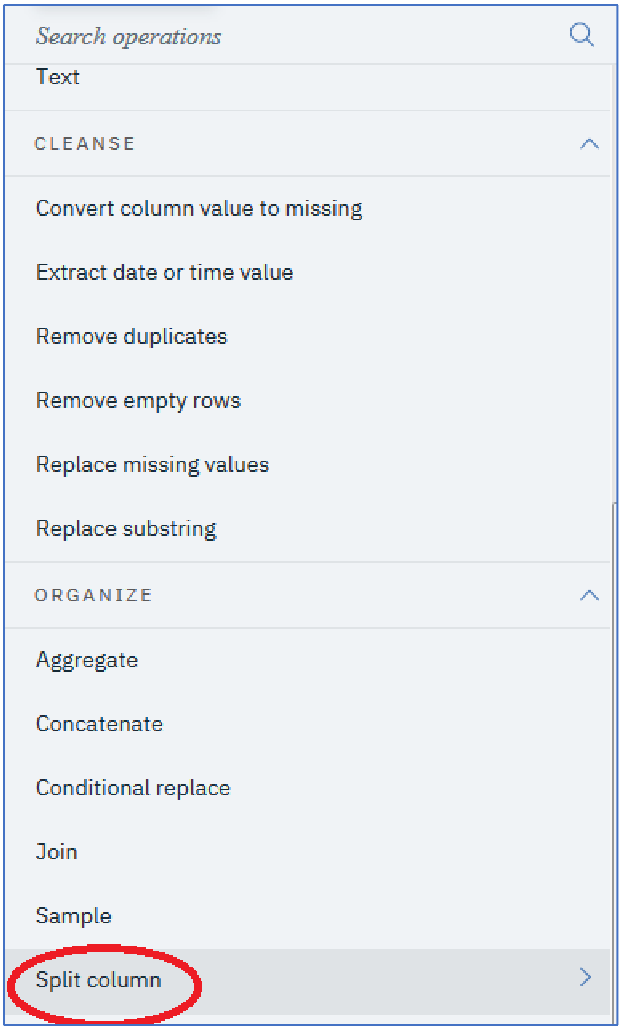
Category column. The flow has 6 overall steps, with the two Join steps shown. Note it will show 4 steps if you skipped steps 1-4 above.



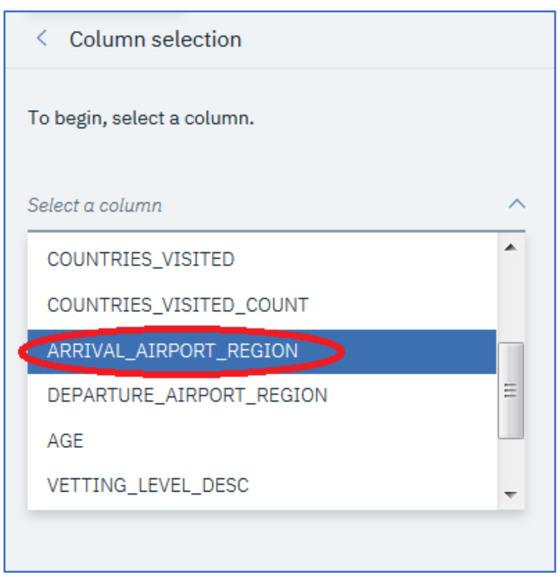
24. We note that the ARRIVAL_AIRPORT_REGION column has “US” concatenated with a State abbreviation (eg US-CA) We want to strip away the “US” to use the column as a State column. The operation **Split column** can be used. Click on **+ Operations.**



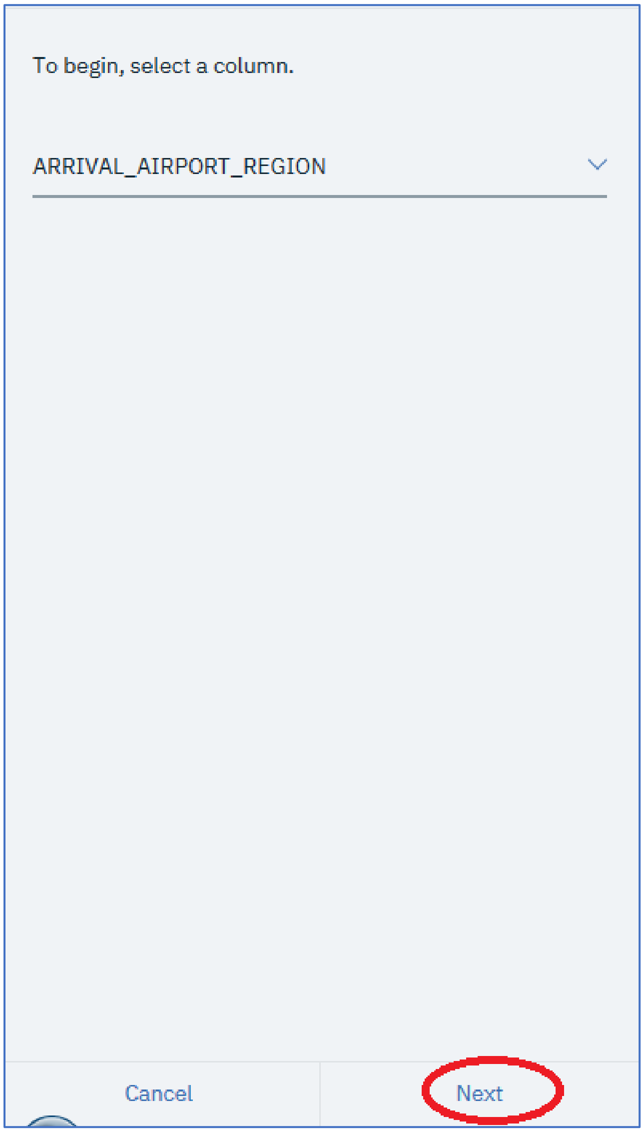
25. Click on **Split column**



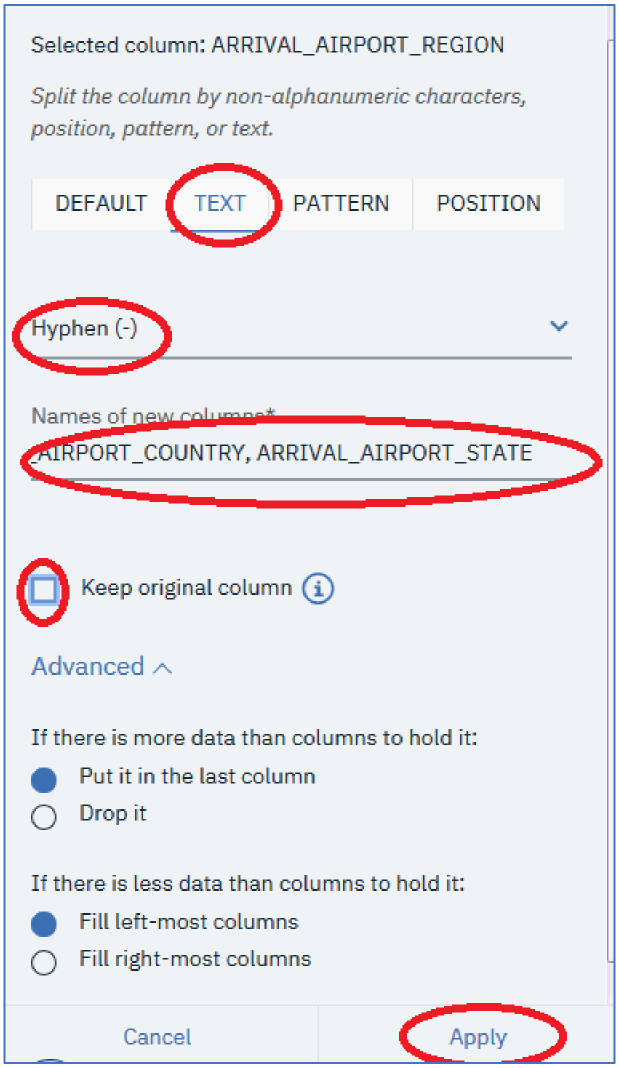
26. Click on **ARRIVAL_AIRPORT_REGION**.



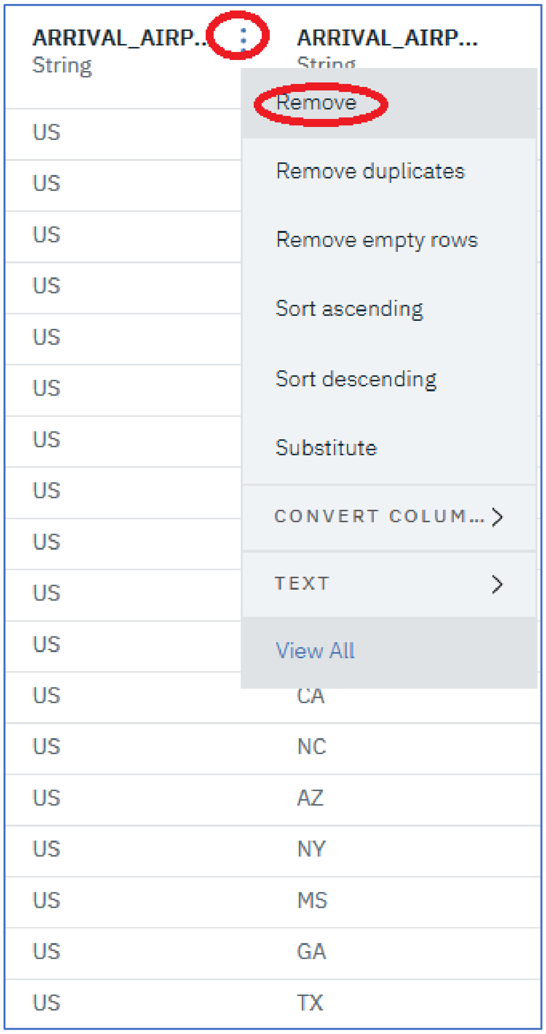
27. Click on **Next**.



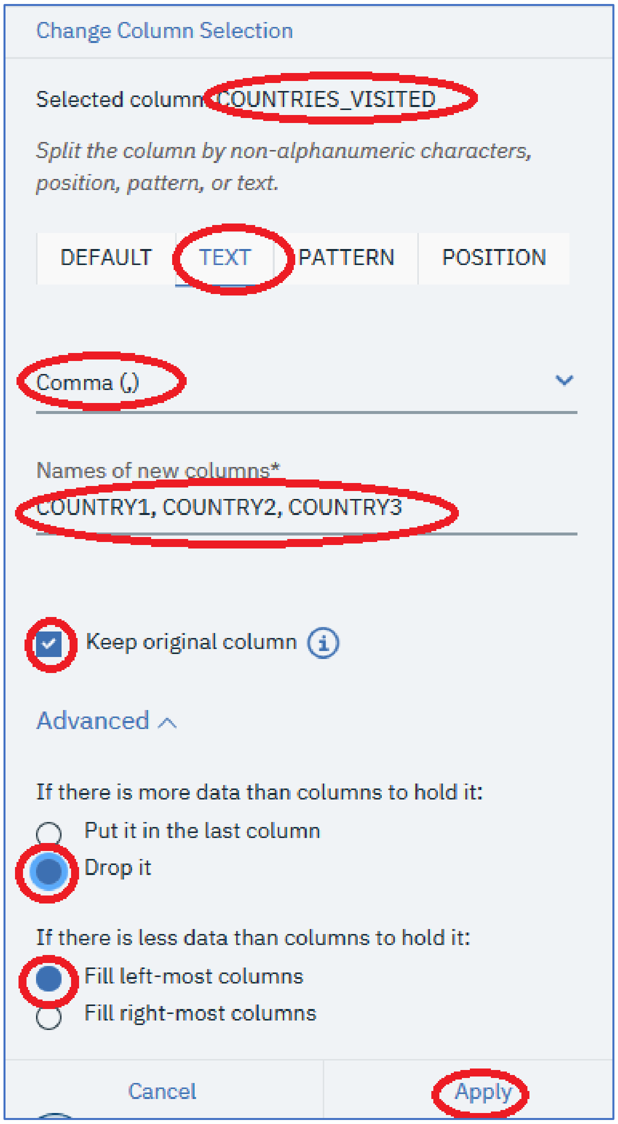
28. Click on **TEXT**, click on **Hypen(-)** in the dropdown, enter **ARRIVAL_AIRPORT_COUNTRY, ARRIVAL_AIRPORT_STATE** as the names of the new columns, uncheck **keep original column**, and click on **Apply**.



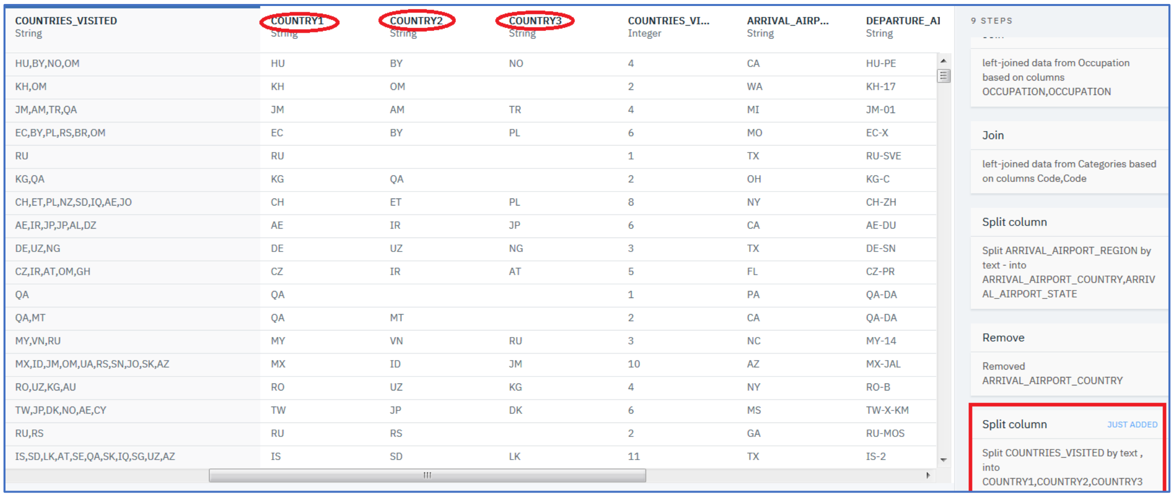
29. Two new columns are created. We don't need the ARRIVAL_AIRPORT_COUNTRY since it has only 1 value – US. Remove the ARRIVAL_AIRPORT_COUNTRY by hovering over the ARRIVAL_AIRPORT_COUNTRY header, clicking on the vertical ellipse ⋮ and clicking on **Remove**.



30. We can use the **Split column** operation on other columns in the dataset. The BIRTH DATE column can be split into YEAR, MONTH, DAY. The DEPARTURE_AIRPORT_REGION can be split in a similar manner as the ARRIVAL_AIRPORT_REGION. The COUNTRIES_VISITED column can be split by the comma. The resulting columns would indicate "first country visited", "second country visited", etc.
31. Let's split the **COUNTRIES_VISITED** column. Split by **TEXT**, use **Comma(,)**, name the new columns **COUNTRY1, COUNTRY2, COUNTRY3** (we will only create 3 new columns), keep the original column. For records where more than 3 countries are visited, **drop** the data. For records where there are less than 3 countries visited, assign it to the **left-most columns**, then click **Apply**. See below.

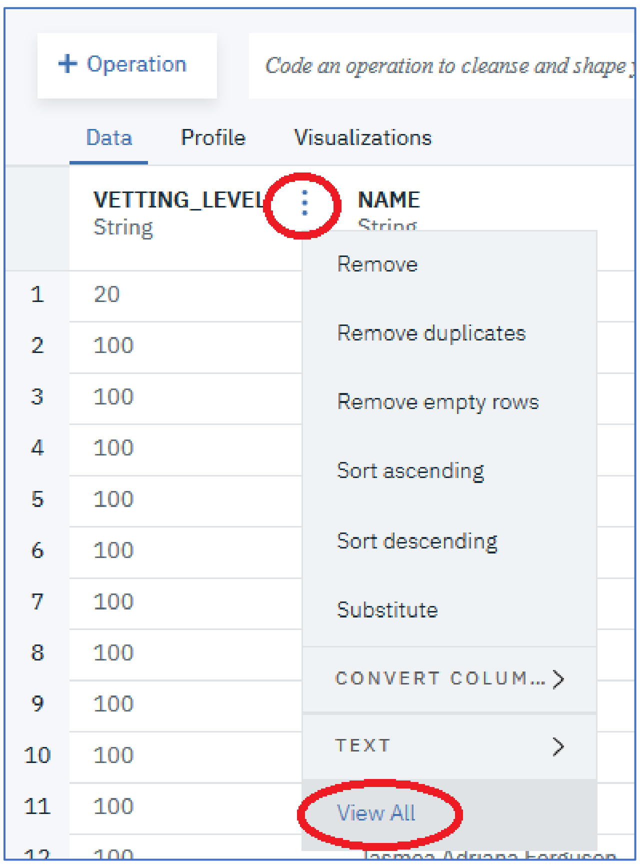


32. The results are shown below. Note there are now 9 steps in the Data Flow. (Only 7 if you skipped steps 1-4 above)

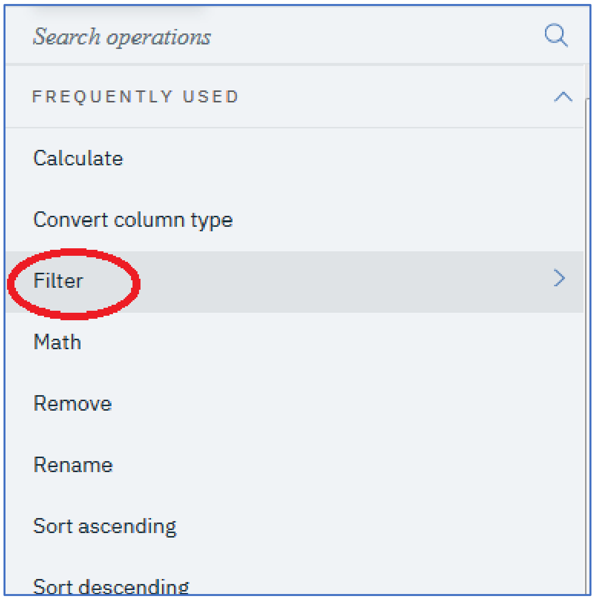


COUNTRIES_VISITED String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VI... Integer	ARRIVAL_AIRP... String	DEPARTURE_AI String
HU,BY,NO,OM	HU	BY	NO	4	CA	HU-PE
KH,OM	KH	OM		2	WA	KH-17
JM,AM,TR,QA	JM	AM	TR	4	MI	JM-01
EC,BY,PL,RS,BR,OM	EC	BY	PL	6	MO	EC-X
RU	RU			1	TX	RU-SVE
KG,QA	KG	QA		2	OH	KG-C
CH,ET,PL,NZ,SD,IQ,AE,JO	CH	ET	PL	8	NY	CH-ZH
AE,IR,JP,JP,AL,DZ	AE	IR	JP	6	CA	AE-DU
DE,UZ,NG	DE	UZ	NG	3	TX	DE-SN
CZ,IR,AT,OM,GH	CZ	IR	AT	5	FL	CZ-PR
QA	QA			1	PA	QA-DA
QA,MT	QA	MT		2	CA	QA-DA
MY,VN,RU	MY	VN	RU	3	NC	MY-14
MX,ID,JM,OM,UA,RS,SN,JO,SK,AZ	MX	ID	JM	10	AZ	MX-JAL
RO,UZ,KG,AU	RO	UZ	KG	4	NY	RO-B
TW,JP,DK,NO,AE,CY	TW	JP	DK	6	MS	TW-X-KM
RU,RS	RU	RS		2	GA	RU-MOS
IS,SD,LK,AT,SE,QA,SK,IQ,SG,UZ,AZ	IS	SD	LK	11	TX	IS-2

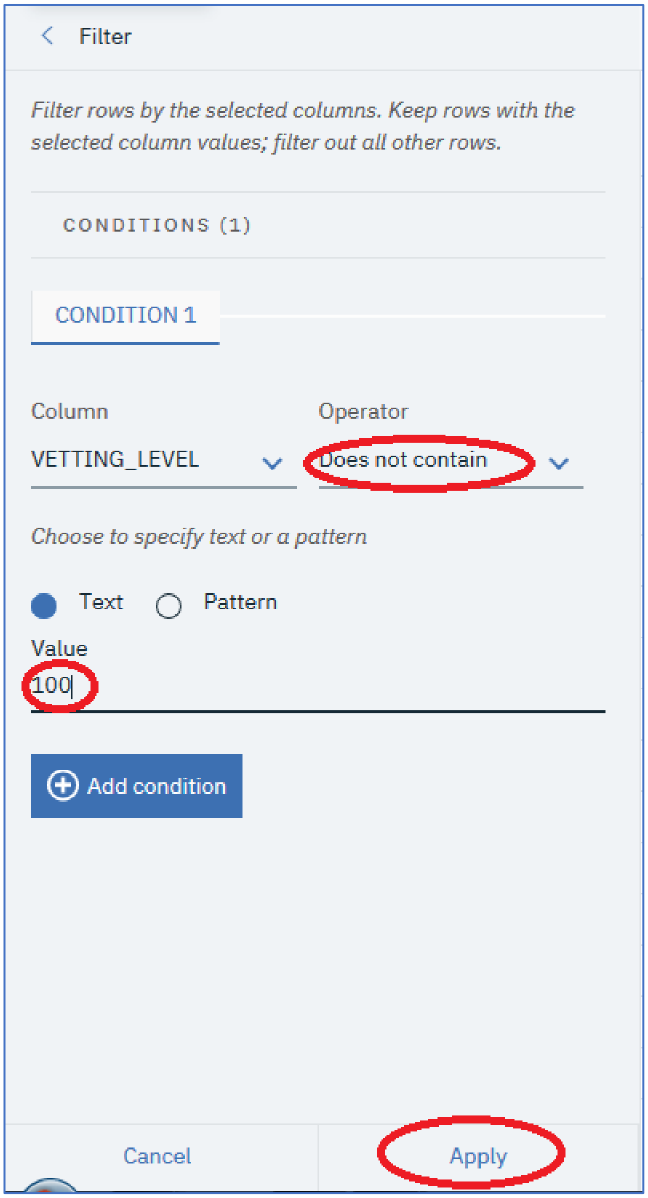
33. Let’s use visualization to get a better understanding of the data. First, we will remove the unvetted records. Hover over the VETTING_LEVEL column, click on the vertical ellipse ⋮, click on **View All**.



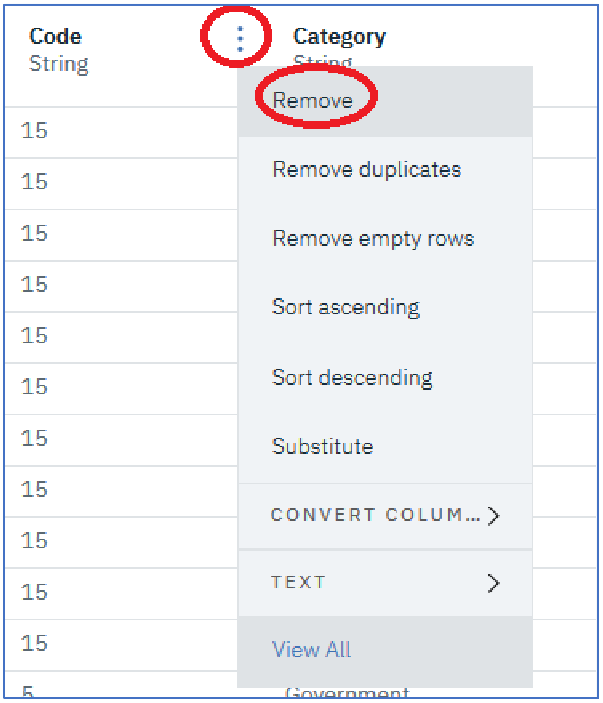
34. Click on **Filter**.



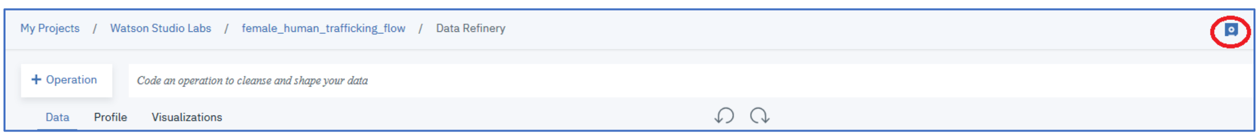
35. Change **Operator** to **Does not contain**, put value as 100, and then click **Apply**.



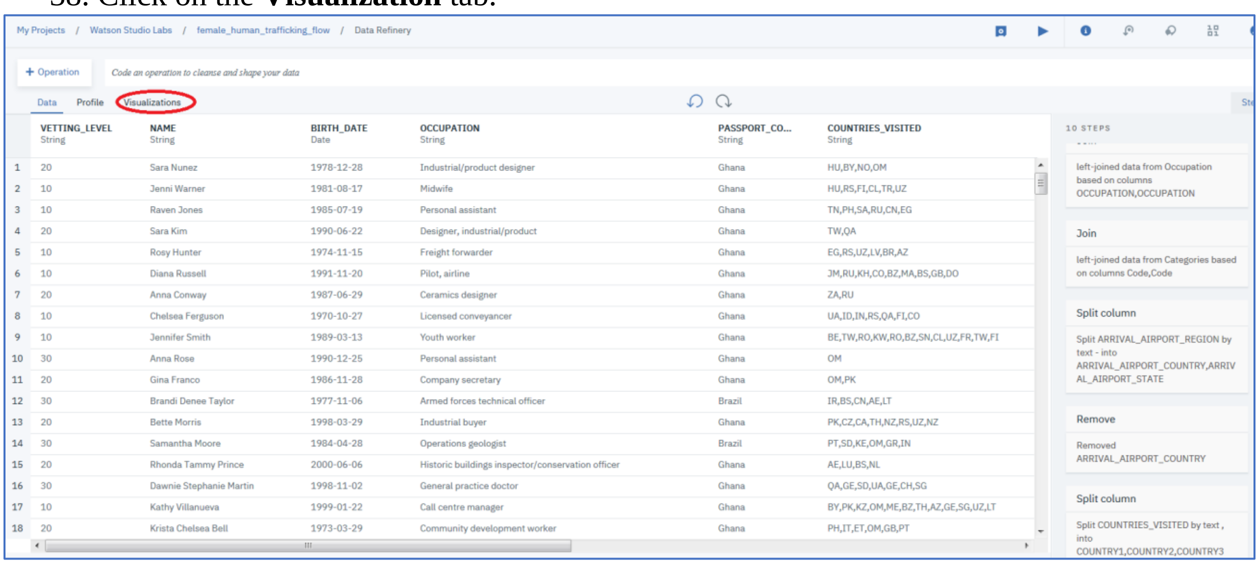
36. Remove the Code column by clicking on the vertical ellipse ⋮ and then clicking **Remove**.



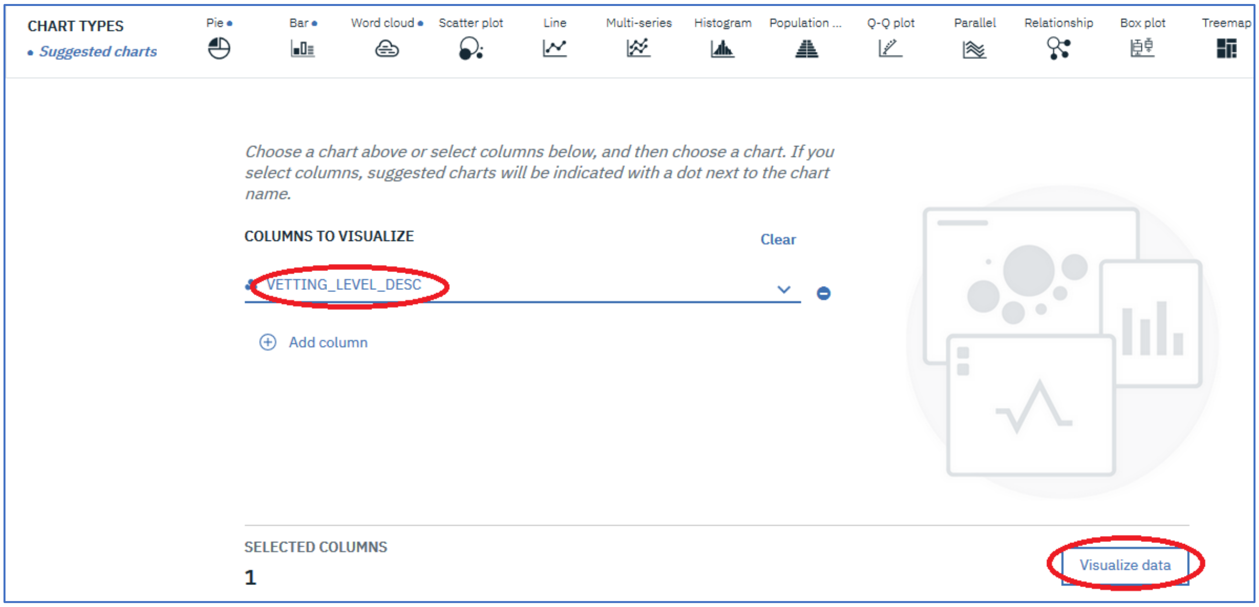
37. Save the Data Flow by clicking on the Save icon.



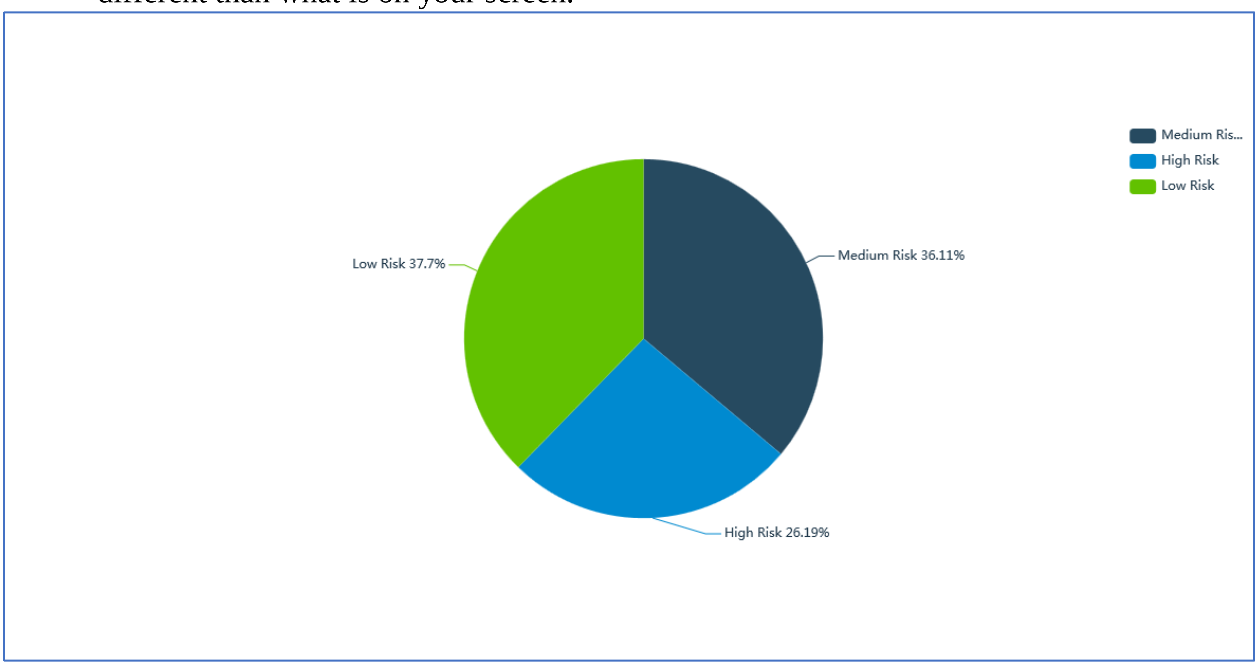
38. Click on the **Visualization** tab.



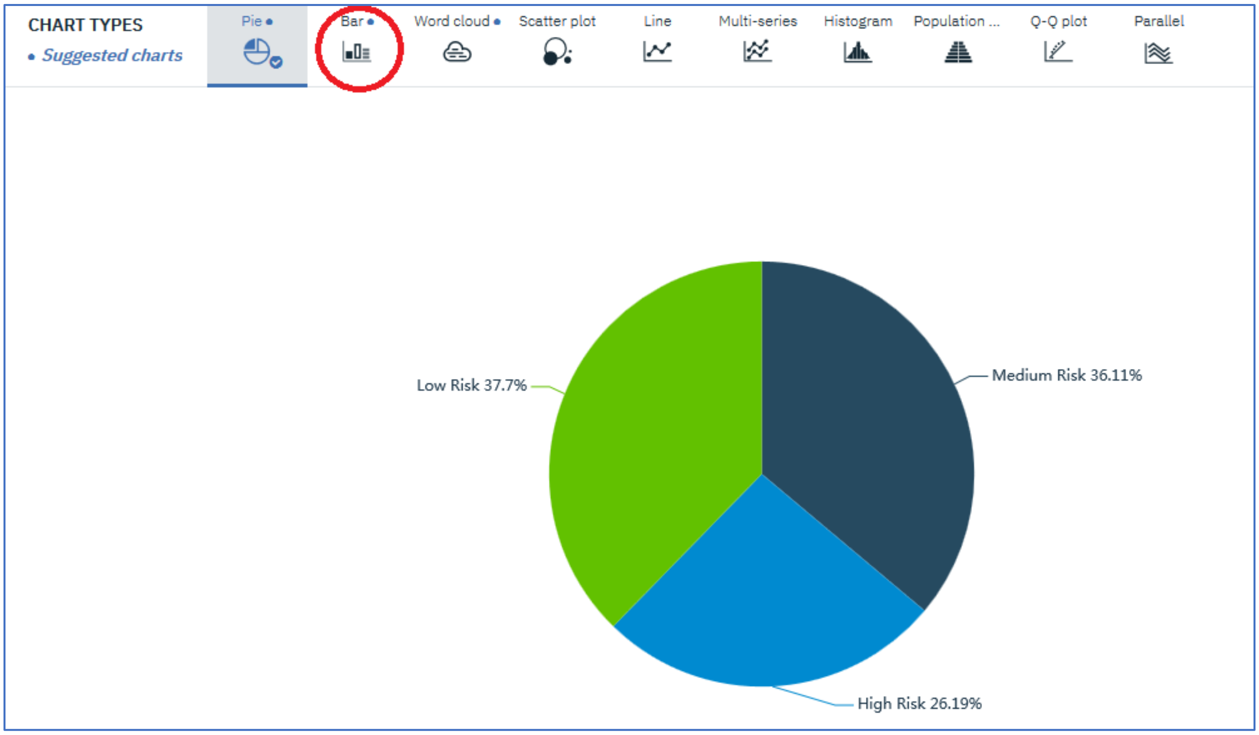
39. Click on **VETTING_LEVEL_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



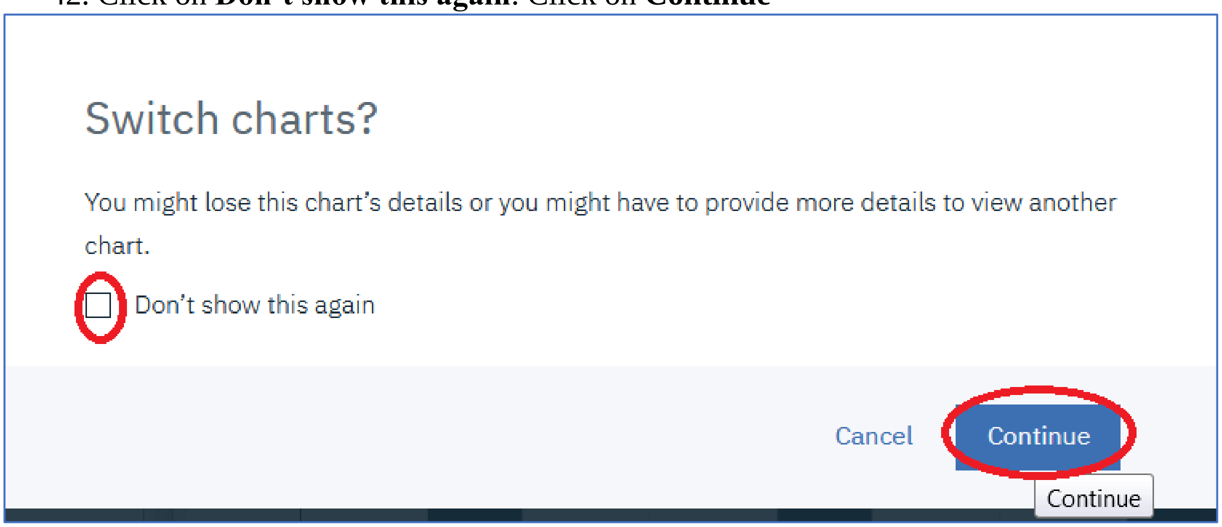
40. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



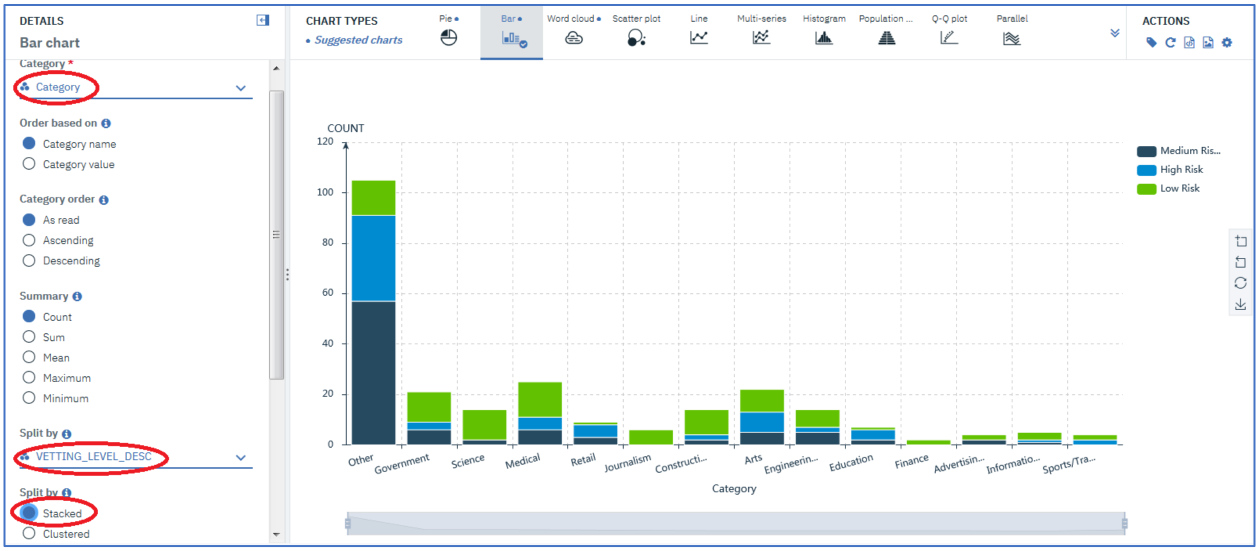
41. We can visualize the breakdown of travel records by job category and vetting level. Click on **Bar**.



42. Click on **Don't show this again**. Click on **Continue**

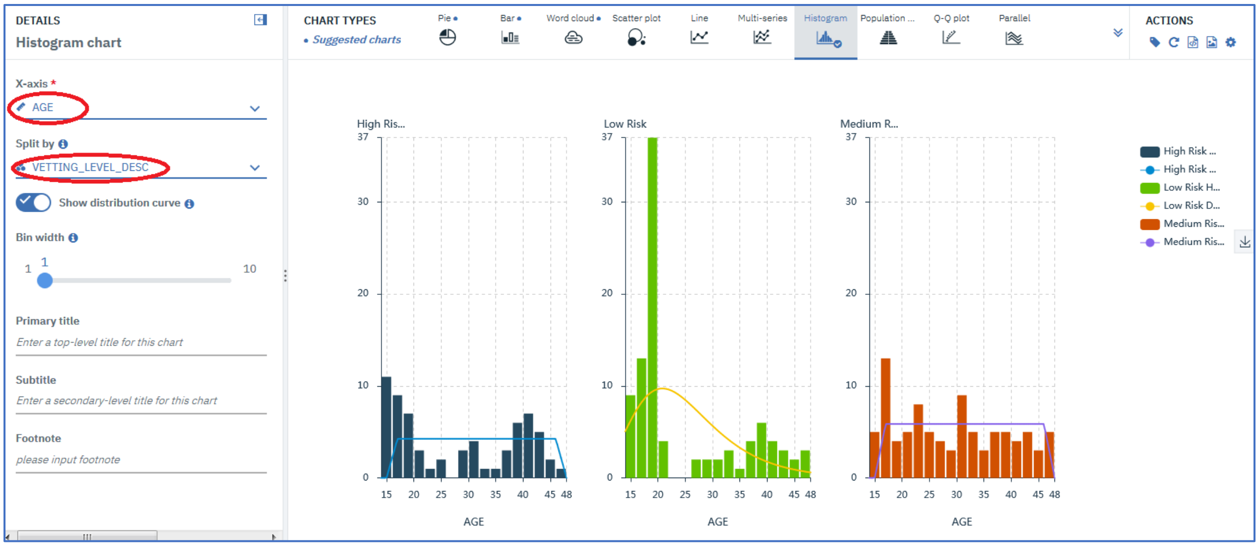


43. Click on **Category** for **Category**, click on VETTING_LEVEL_DESC for **Split by**, click on **Stacked** for **Split by**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



44. We can visualize a histogram of COUNTRIES_VISITED_COUNTS split by VETTING_LEVEL_DESC. Click on **Histogram**, click on **COUNTRIES_VISITED_COUNT** for **X-axis**, click on **VETTING_LEVEL_DESC** for **Split by**. Note that at higher number of countries visited, there is an increasing likelihood that it is a high-risk person.

45. Let's examine if age makes a difference. Click on **AGE** for **X-axis**. It appears that younger travelers have a lower risk of being trafficked.

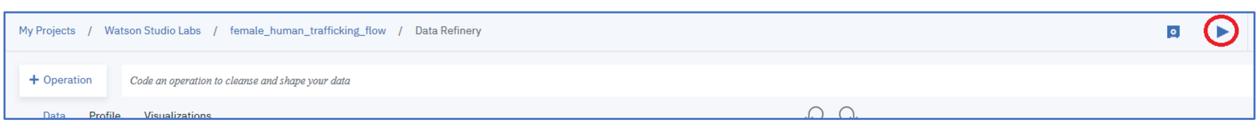


46. Please feel free to experiment with other visualizations.

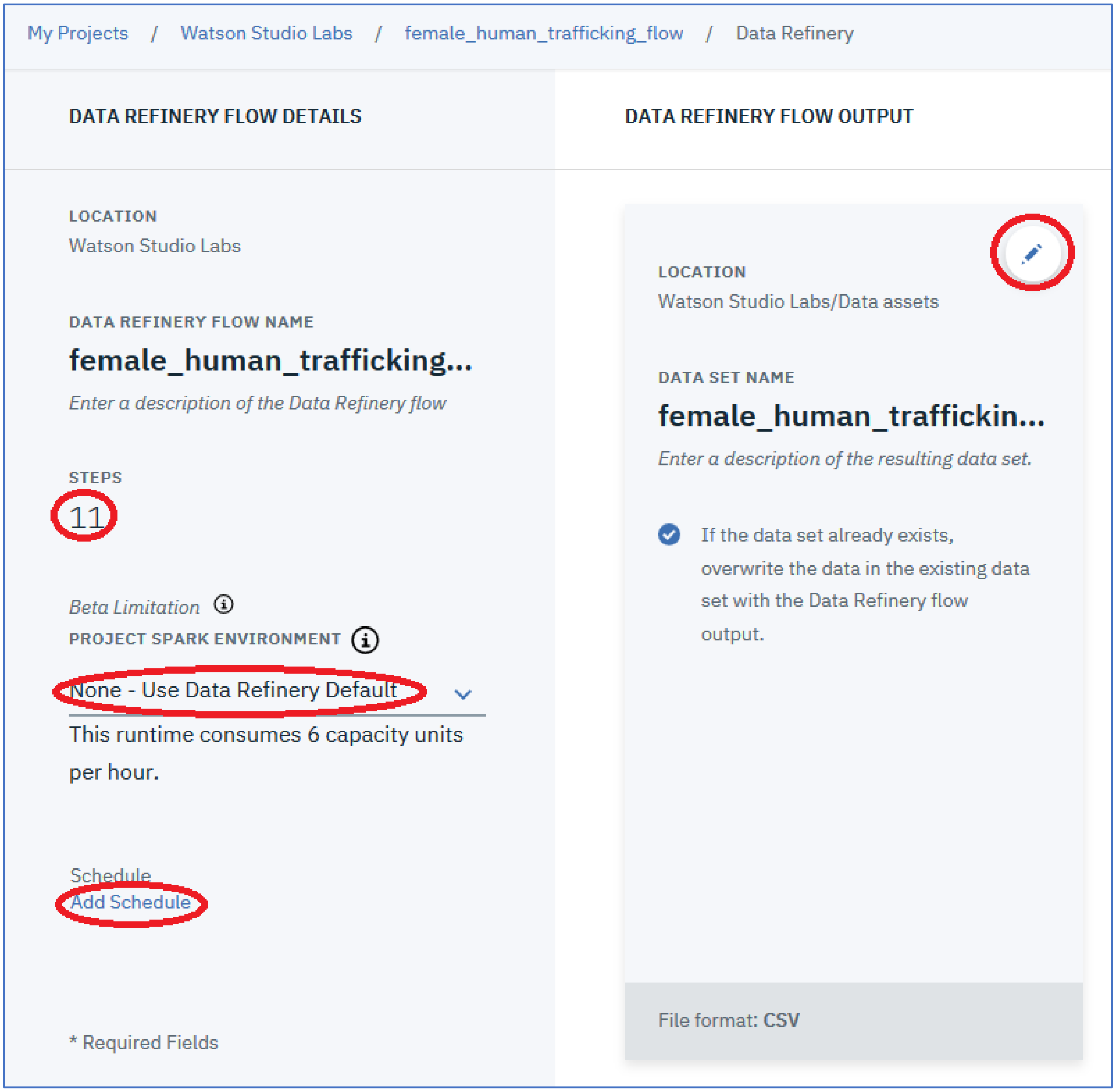
Run the sequence of Data Operations on the entire data set.

When users are interacting with the Data Refinery tool, the operations are applied to a subset of the data set to facilitate faster response times. To run the data operations on the entire data set, the user selects the run option.

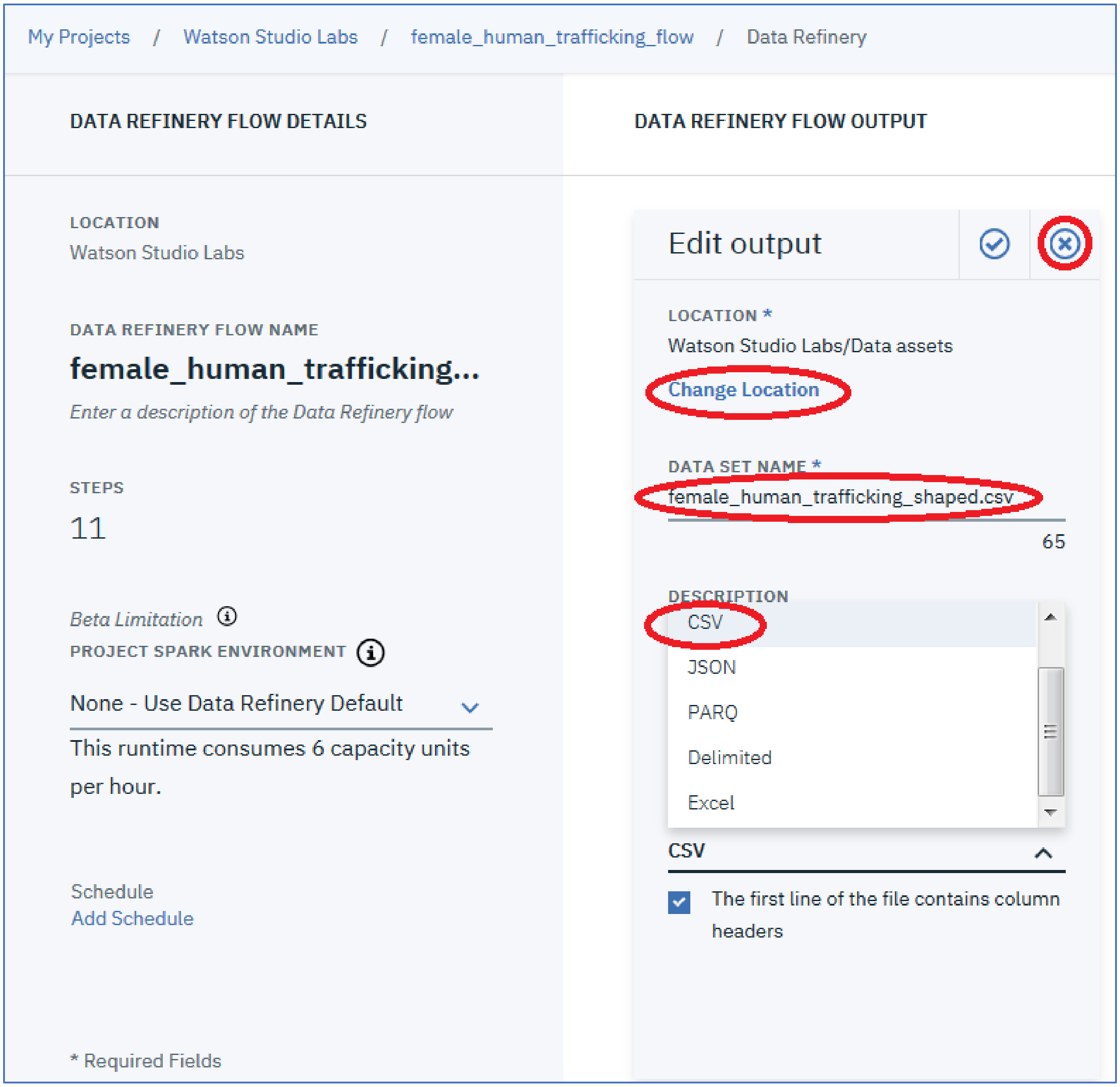
1. Click on run icon 



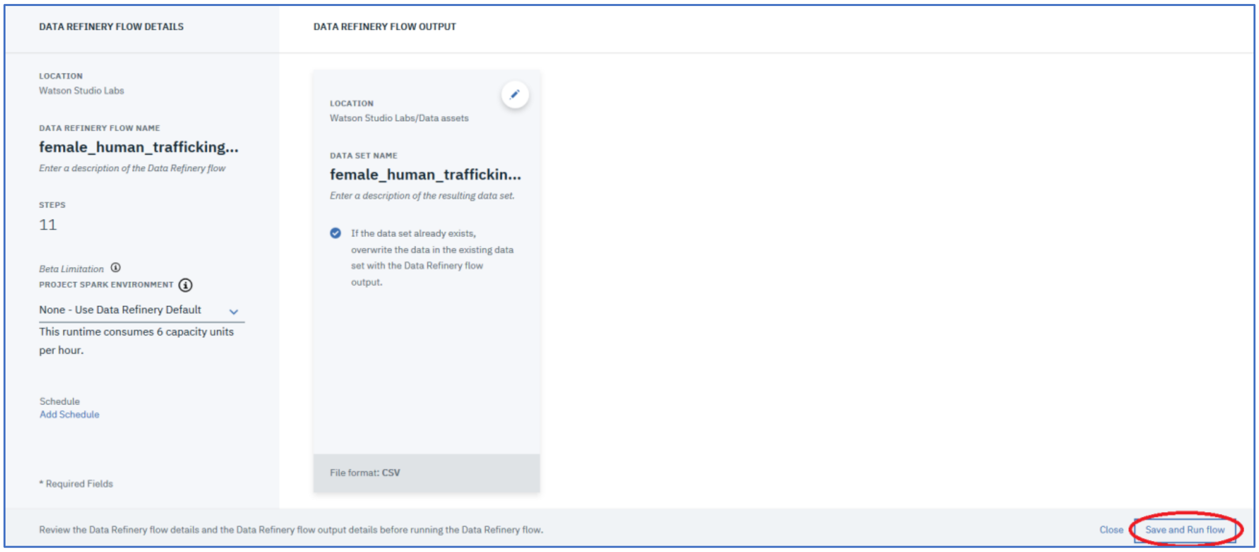
2. Note the number of steps used to transform the data. It should be 11 (or 9 if steps 1-4 above were skipped). A schedule can be set up if the transformation process needs to run on a scheduled basis (see Add Schedule option). We are just going to do a one-time run. Change the name of the output file by clicking on the edit option  icon. (pencil icon).



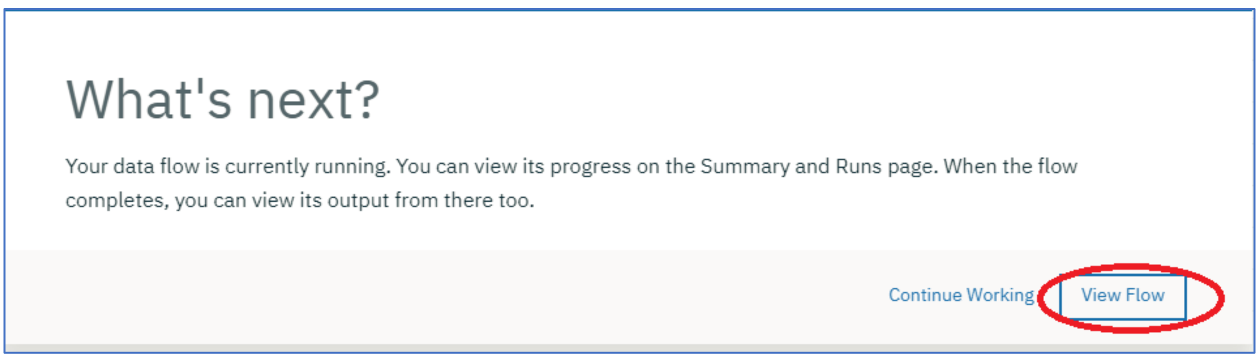
3. You have several options regarding the Data Refinery output. You can **Change location**. You can edit the name of the file. You can edit the file type. We will leave the defaults and check the close icon.



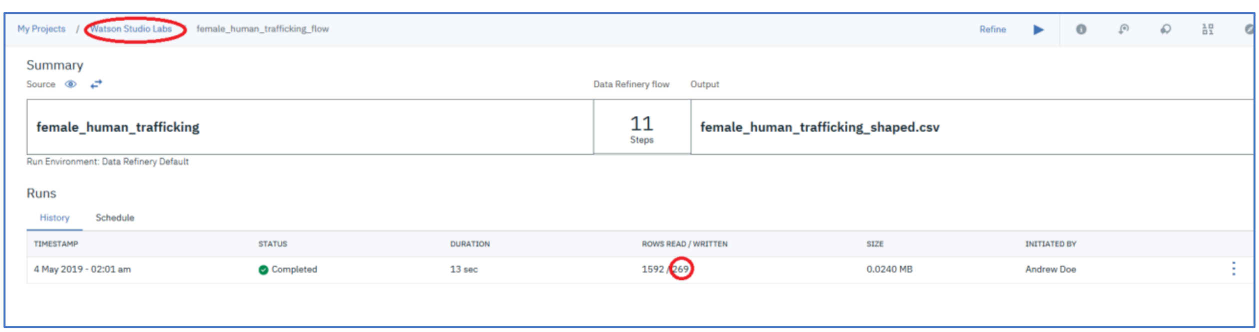
4. Click **Save and Run**.



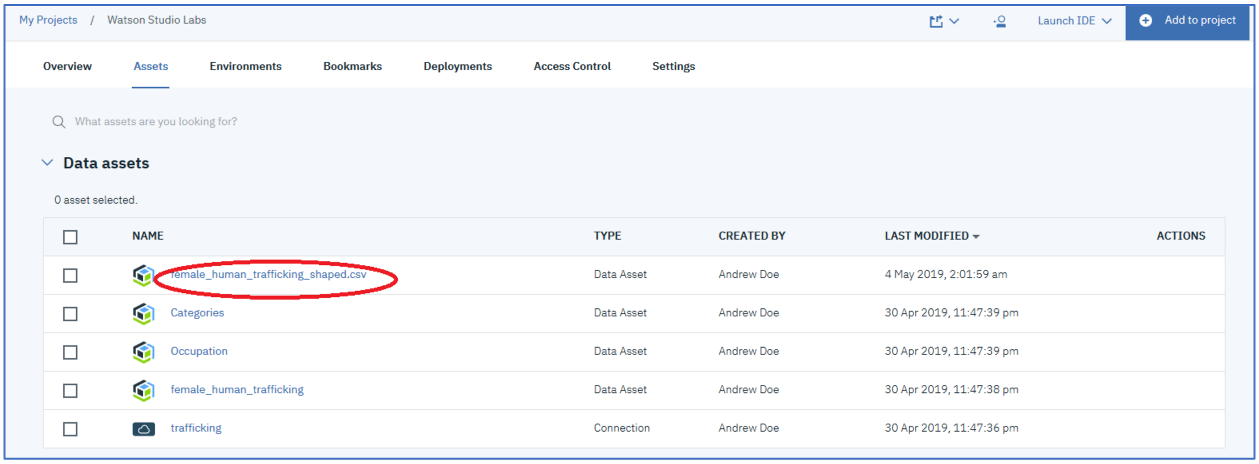
5. You can continue to work on other items or monitor the Data Flow run status. Click on **View Flow**.



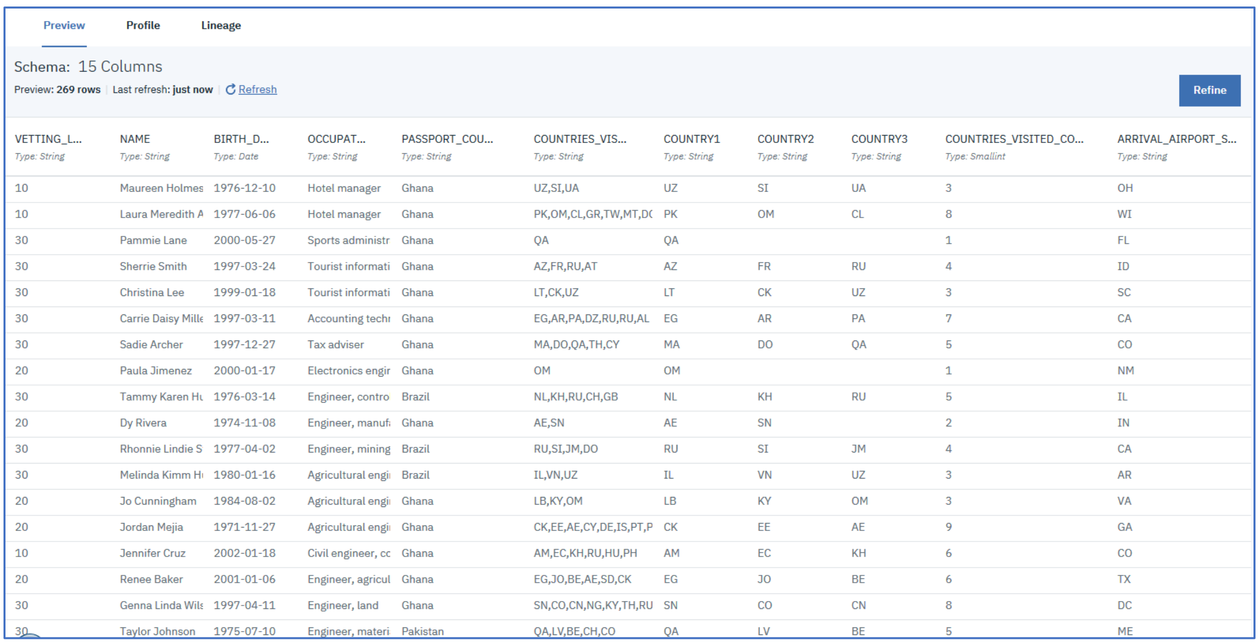
6. The completed flow is shown below. Note that 269 records were written to the output file. Click on Watson Studio Labs to go back to the project Assets page.



7. The output of the Data Refinery process should be listed in the Data Assets. Click on the asset to view the contents.



8. The asset contents are displayed below. Review to confirm that the data transformations specified have been applied to all the data.



Preview | Profile | Lineage

Schema: 15 Columns

Preview: 269 rows | Last refresh: just now | Refresh

Refine

VETTING_L... Type: String	NAME Type: String	BIRTH_D... Type: Date	OCCUPAT... Type: String	PASSPORT_COU... Type: String	COUNTRIES_VIS... Type: String	COUNTRY1 Type: String	COUNTRY2 Type: String	COUNTRY3 Type: String	COUNTRIES_VISITED_CO... Type: Smallint	ARRIVAL_AIRPORT_S... Type: String
10	Maureen Holmes	1976-12-10	Hotel manager	Ghana	UZ,SI,UA	UZ	SI	UA	3	OH
10	Laura Meredith A	1977-06-06	Hotel manager	Ghana	PK,OM,CL,GR,TW,MT,DC	PK	OM	CL	8	WI
30	Pammie Lane	2000-05-27	Sports administr	Ghana	QA	QA			1	FL
30	Sherrie Smith	1997-03-24	Tourist informati	Ghana	AZ,FR,RU,AT	AZ	FR	RU	4	ID
30	Christina Lee	1999-01-18	Tourist informati	Ghana	LT,CK,UZ	LT	CK	UZ	3	SC
30	Carrie Daisy Mille	1997-03-11	Accounting techi	Ghana	EG,AR,PA,DZ,RU,RU,AL	EG	AR	PA	7	CA
30	Sadie Archer	1997-12-27	Tax adviser	Ghana	MA,DO,QA,TH,CY	MA	DO	QA	5	CO
20	Paula Jimenez	2000-01-17	Electronics engir	Ghana	OM	OM			1	NM
30	Tammy Karen Hu	1976-03-14	Engineer, contro	Brazil	NL,KH,RU,CH,GB	NL	KH	RU	5	IL
20	Dy Rivera	1974-11-08	Engineer, manufa	Ghana	AE,SN	AE	SN		2	IN
30	Rhonnie Lindie S	1977-04-02	Engineer, mining	Brazil	RU,SI,JM,DO	RU	SI	JM	4	CA
30	Melinda Kimm H	1980-01-16	Agricultural engi	Brazil	IL,VN,UZ	IL	VN	UZ	3	AR
20	Jo Cunningham	1984-08-02	Agricultural engi	Ghana	LB,KY,OM	LB	KY	OM	3	VA
20	Jordan Mejia	1971-11-27	Agricultural engi	Ghana	CK,EE,AE,CY,DE,IS,PT,P	CK	EE	AE	9	GA
10	Jennifer Cruz	2002-01-18	Civil engineer, cc	Ghana	AM,EC,KH,RU,HU,PH	AM	EC	KH	6	CO
20	Renee Baker	2001-01-06	Engineer, agricul	Ghana	EG,JO,BE,AE,SD,CK	EG	JO	BE	6	TX
30	Genna Linda Wils	1997-04-11	Engineer, land	Ghana	SN,CO,CN,NG,KY,TH,RU	SN	CO	CN	8	DC
30	Taylor Johnson	1975-07-10	Engineer, materi	Pakistan	QA,LV,BE,CH,CO	QA	LV	BE	5	ME

You have completed Lab-3!

- ✓ Created a new Data Flow
- ✓ Profiled the data
- ✓ Visualized the data to gain a better understanding
- ✓ Prepared the data for modeling
- ✓ Ran the sequence of data preparation operations on the entire data set.