

# Data Refinery Lab

## Introduction

This lab will introduce 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.

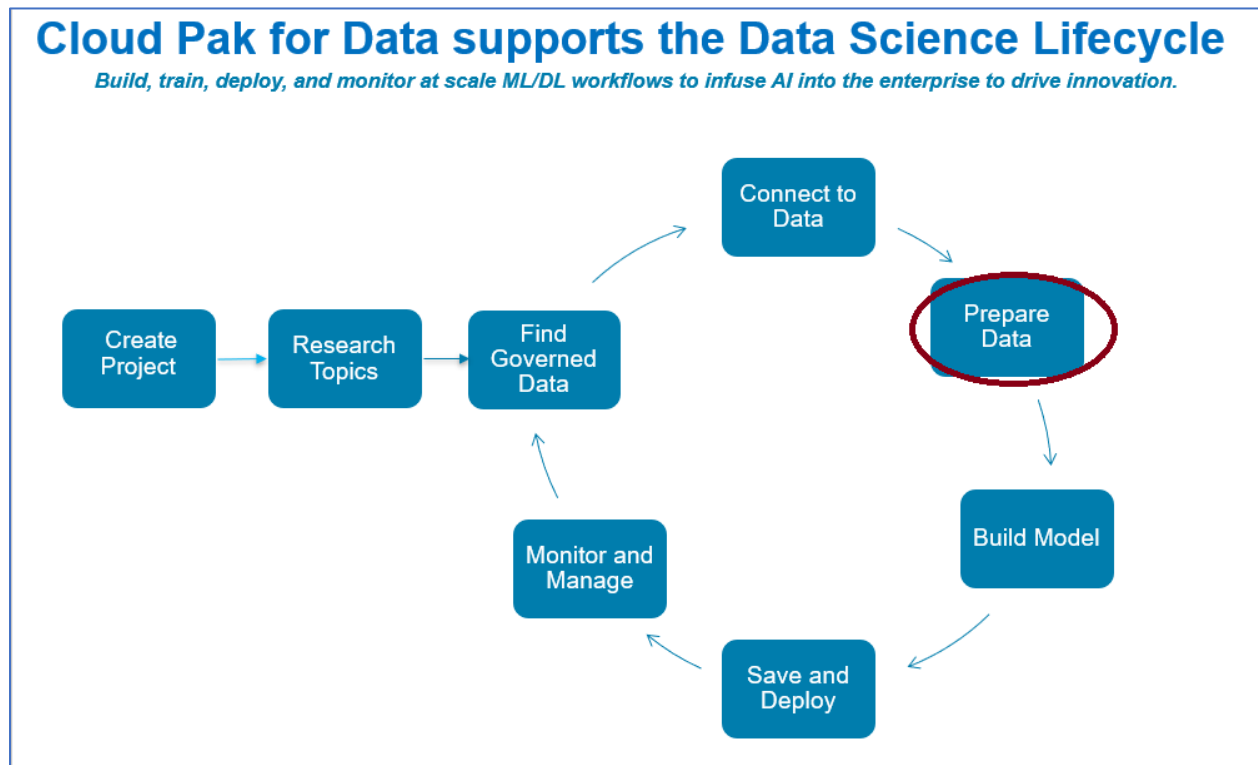


Figure 1- End to End Flow

## 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.

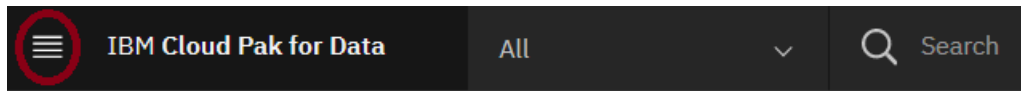
The `OCCUPATION` data included in the travel data is very granular. For modeling purposes, it was decided to categorize the `OCCUPATION` data. Two additional datasets are used for this purpose. The `occupation.csv` dataset maps the granular occupation data to a category code. The `categories` dataset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

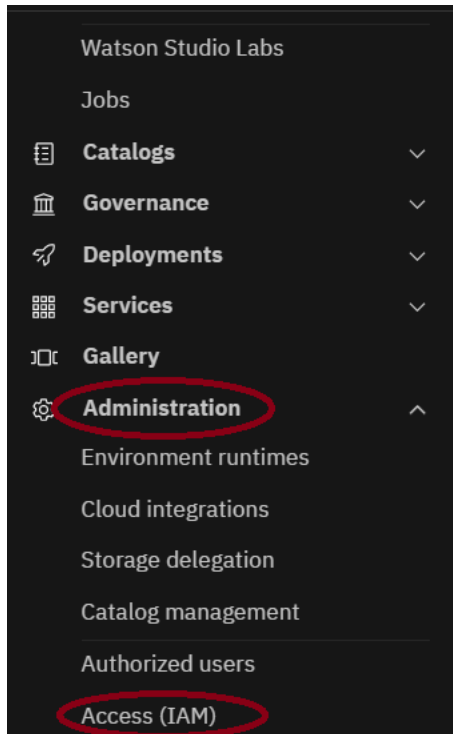
## Remove User

Before starting the Data Refinery, let's remove the `ws.catalog.user` from your account.

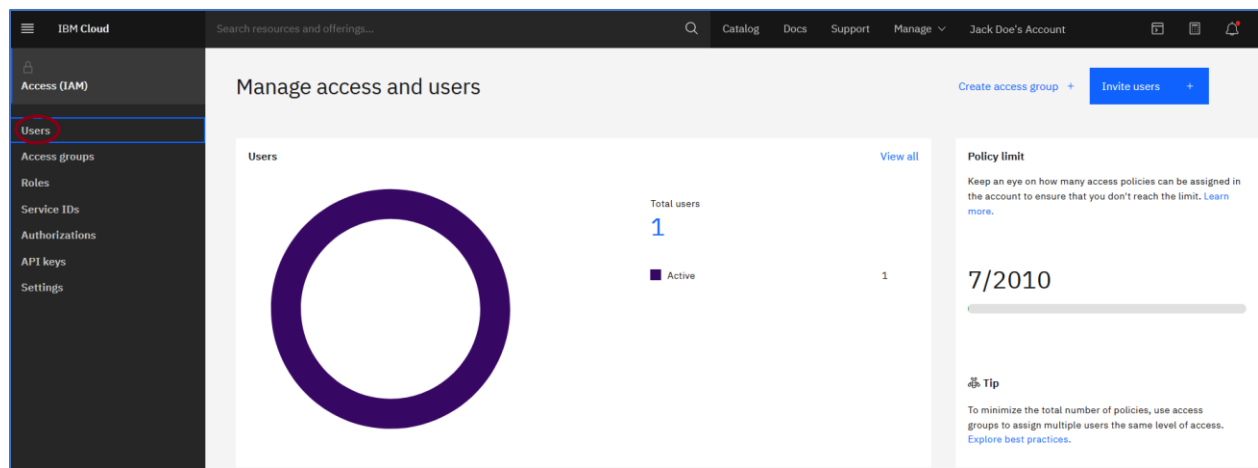
1. Click on the hamburger  icon




2. Click on **Administration** and then click on **Access (IAM)**. Scroll down if necessary.



3. An **Identity and Access Management (IAM)** browser tab is created providing the IBM Cloud user interface to the IAM subsystem. Note, you may have to log-in first. Click on **Users** in the menu panel.



4. Click on the vertical ellipse  at the right of Howard Doe and click **Remove user**.

## Users

Use the **View** option to change your view of the user list by selecting the user grouping. Depending on your access, you might see users grouped at the account level or by Cloud Foundry org or user hierarchy for classic infrastructure access.

View: Account users ▼

Status: Filter... ▼ 🔍 Invite users +

User	Email	Status
<a href="#">Howard Doe</a>	ws.catalog.user@gmail.com	Active
<a href="#">Victor Doe</a> <span>owner</span> <span>self</span>	wsuser64000@gmail.com	Active

1-2 of 2 items

⋮

- Manage user
- Assign access
- Remove user**

5. Click **Remove**.

### Remove user


Are you sure about removing user **Howard Doe**?


Cancel **Remove**

6. Close the Identity and Access Management tab.

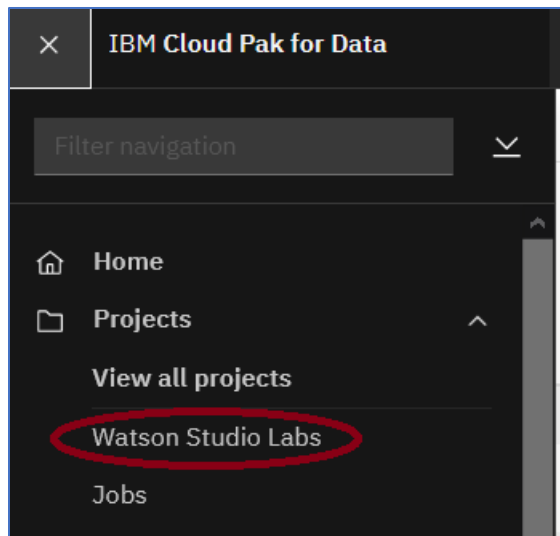
IBM Cloud Pak for Data × Identity & Access Management ×

## Create a new Data Flow

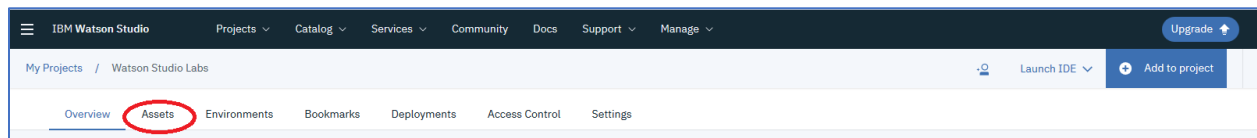
1. Click on the hamburger icon .

 **IBM Cloud Pak for Data**

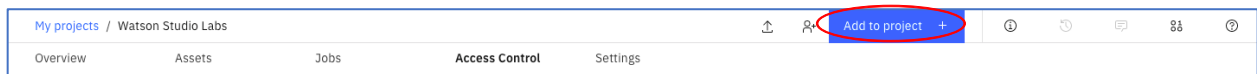
2. Click on **Watson Studio Labs** under **Projects**.



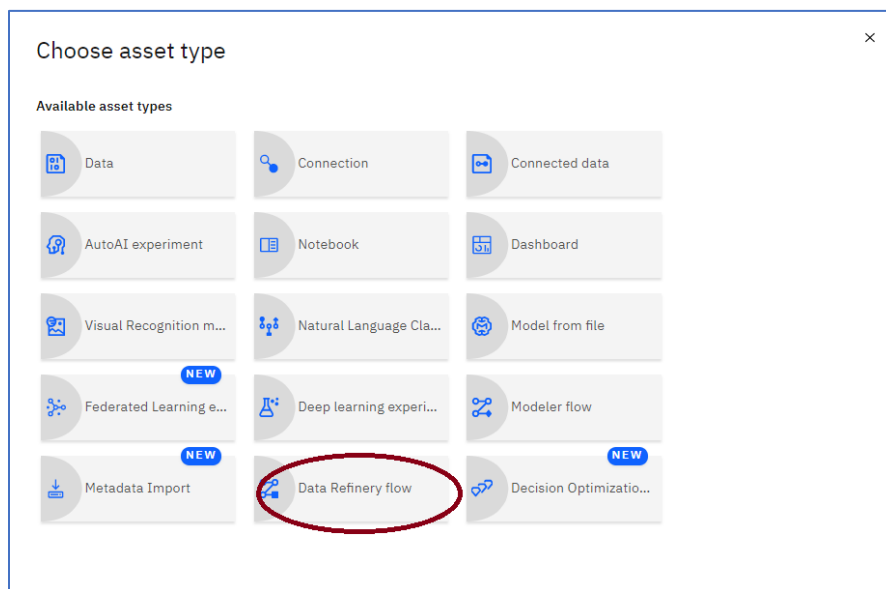
3. Click on the **Assets** tab.



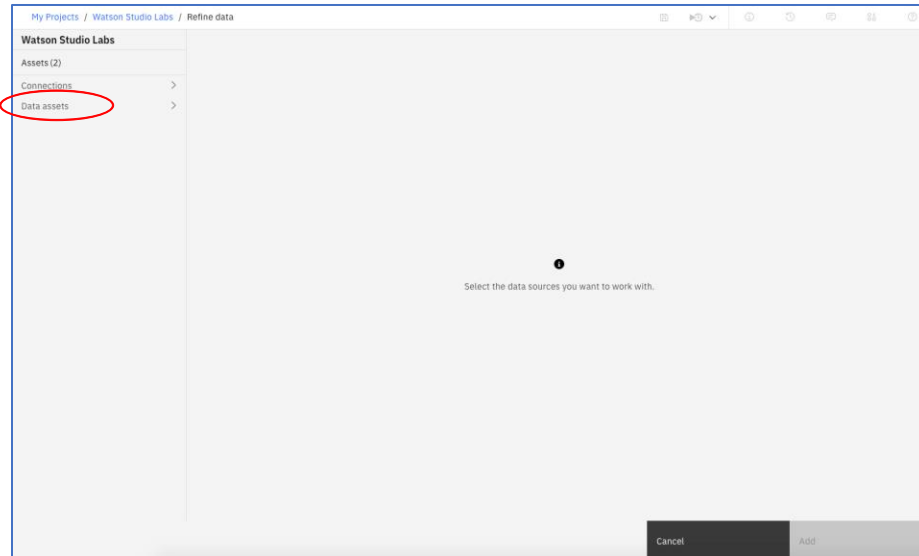
4. Add a Data Flow by clicking on **Add to project**.



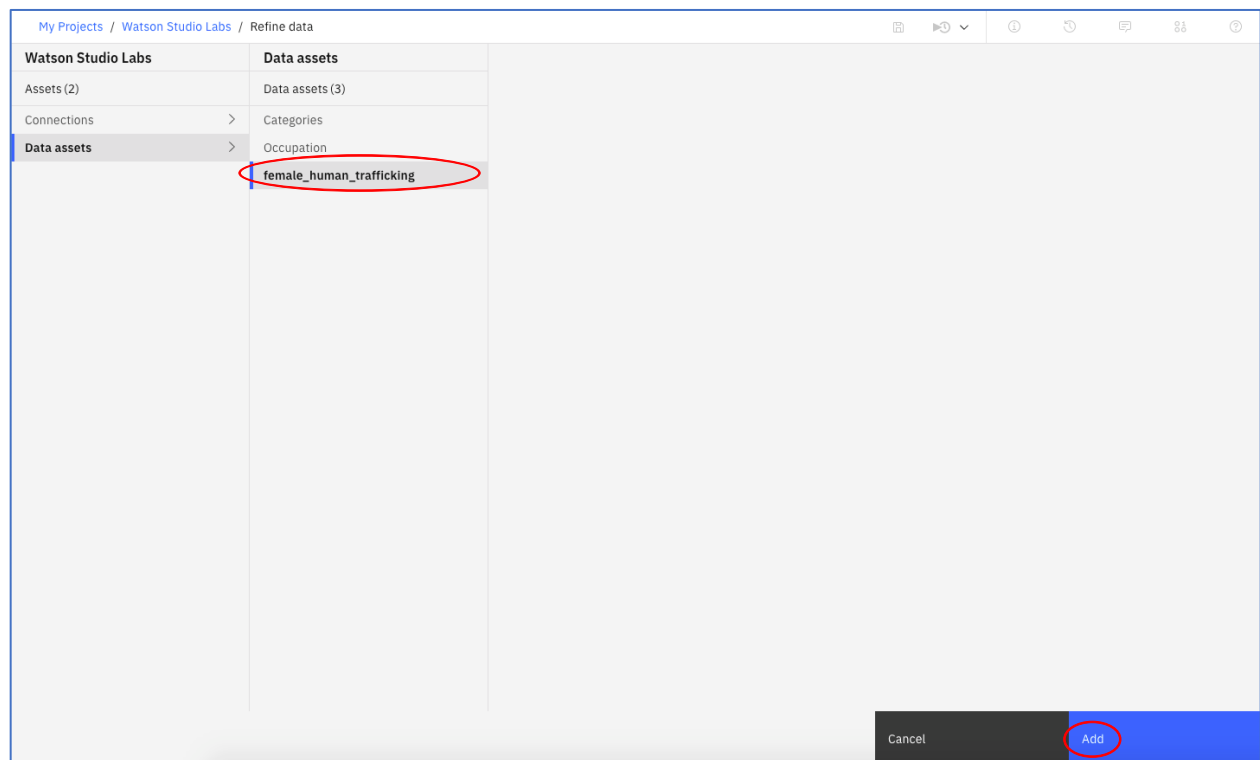
5. Click on **Data Refinery flow**



6. Click on **Data Assets**.



7. Click on **female\_human\_trafficking**, and then click on **Add**.



8. The data set will be displayed. Please wait until the Previewing is complete.

Projects / Watson Studio Labs / female\_human\_trafficking / Refine data

Operation + Code an operation to cleanse and shape your data

Previewing the first 50 rows  
Reading and processing data sample...

Data Profile Visualizations

	INTERNAL_ID	VETTING_LEVEL	DESCRIPTION	NAME	GENDER	BIRTH_DATE	BIRTH_COUNTRY
1	596	10	NA	Avi Gomez	F	1998-09-12	Ghana
2	597	100	NA	Maria Gregory	F	1977-05-13	Ghana
3	598	100	NA	Rachel Sanchez	F	1985-01-10	Ghana
4	599	30	NA	Kristina Kelly	F	1980-12-17	Ghana
5	600	100	NA	Nichole Mandi Houston	F	1989-08-05	Ghana
6	601	30	NA	Shelly Kylie Franklin	F	1989-02-13	Ghana
7	602	100	NA	Monique Tina Ellis	F	1981-05-25	Ghana
8	603	100	NA	Allison Tucker	F	1990-05-12	Ghana
9	604	20	NA	Cass Gabrielle Robinson	F	2000-10-20	Ghana
10	605	100	NA	Tara Sharp	F	1973-10-12	Ghana
11	606	100	NA	Kathy McCarthy	F	1973-05-12	Ghana

Details Help

Edit


LOCATION  
Watson Studio Labs

DATA REFINERY FLOW NAME  
female\_human\_trafficki...  
Enter a description of the Data Refinery flow

STEPS  
0


## 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.


**Tip!** We have you save the flow after all the transformations have been made. Data Refinery will not save the transformations automatically. So, you need to click on the  icon if you want to save the changes along the way.

Projects / Watson Studio Labs / female\_human\_trafficking / Refine data

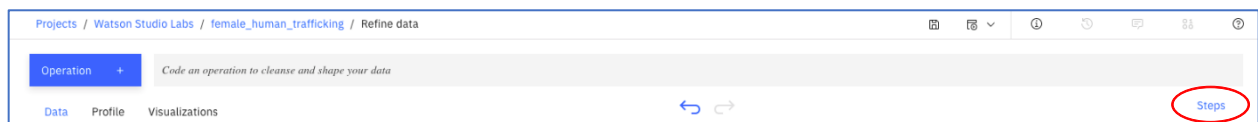
- Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the **VETTING\_LEVEL** column by hovering over **VETTING\_LEVEL**, clicking on the vertical ellipse , clicking on **CONVERT COLUMN**, and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

VETTING_LEVEL Integer	DESCRIPTION String	NAME String
100	Remove	Meghan Moses
30	Remove duplicates	Trace Carr
10	Remove empty rows	Ami Casey Woods
30	Sort ascending	Melinda Kimm Hubbard
10	Sort descending	Linda Tucker
100	Substitute	Tamara Palmer
30	CONVERT COLUMN...	Boolean
100	View All	Decimal
30		Integer
100	NA	String
30	NA	

- Convert the **PASSPORT\_NUMBER** column by hovering over **PASSPORT\_NUMBER**, clicking on the vertical ellipse , clicking on **CONVERT COLUMN**, and clicking on **String**.

PASSPORT_NUM...	PASSPORT_CO...	PASSPORT_CO...	CO
Integer	String	String	Stri
308561300		GH	QA
987374355		GH	QA
426221095		GH	ME
869842380		BR	IL,N
473389048		BR	ES,
217560040		GH	HR,
942939007	CONVERT COLU...>	Boolean	DM
768902471	View All	Decimal	IP,I
730613975		Integer	IP,s
798632110	Ghana	String	RU
400880971	Ghana		RU

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.

Steps
2 Steps
Data Source
female_human_trafficking
Convert column type
AUTOMATIC
Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.
Convert column type
JUST ADDED
Manually converted data types for 1 column.

5. Click on **Profile**.



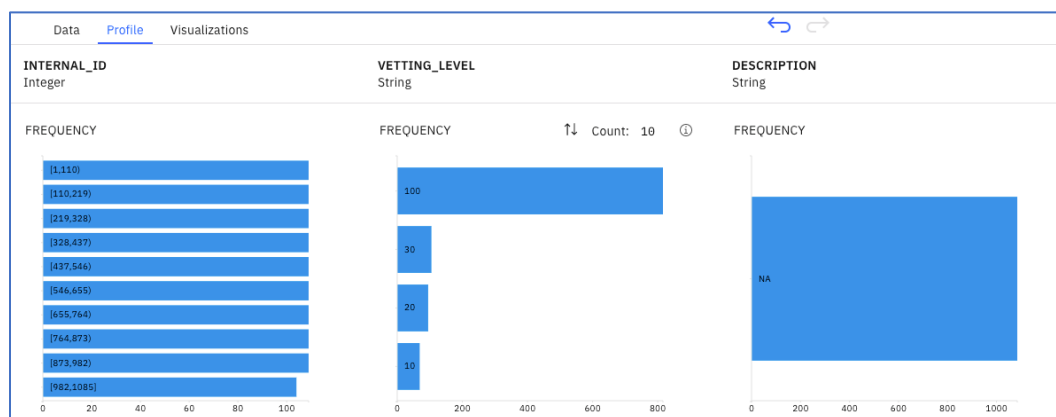
My Projects / Watson Studio Labs / female\_human\_trafficking / Refine data

Operation +

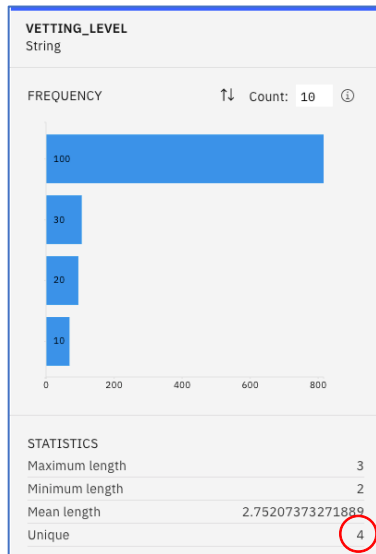
Data **Profile** Visualizations

	SSN String	PASSPORT_NU... Decimal	PASSPORT_CO... String	PASSPORT_CO... String
1	395-82-6068	308561300	Ghana	GH
2	600-46-7639	987374355	Ghana	GH
3	800-46-1520	426221095	Ghana	GH

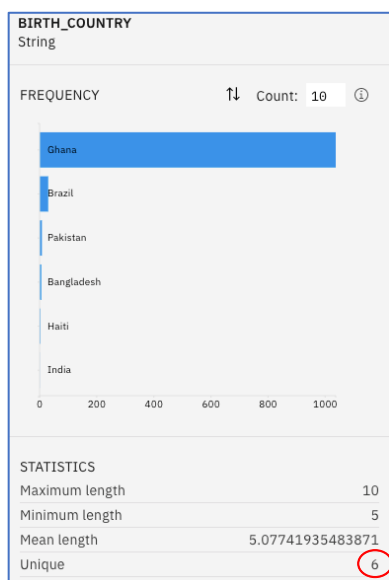
6. The Profile panel displays the counts of the top 10 values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column.



7. The statistics for the VETTING\_LEVEL column show 4 unique values, 10, 20, 30, and 100. These are coded values that correspond to the 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. In subsequent labs, we will use the data that has been vetted to train a model to predict the risk for the unvetted records.



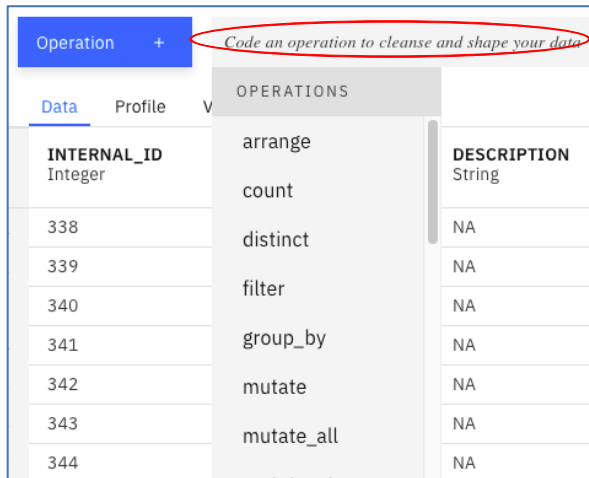
8. Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has about 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH\_COUNTRY, and PASSPORT\_COUNTRY shows only 6 unique countries. The COUNTRIES\_VISITED\_COUNT 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.



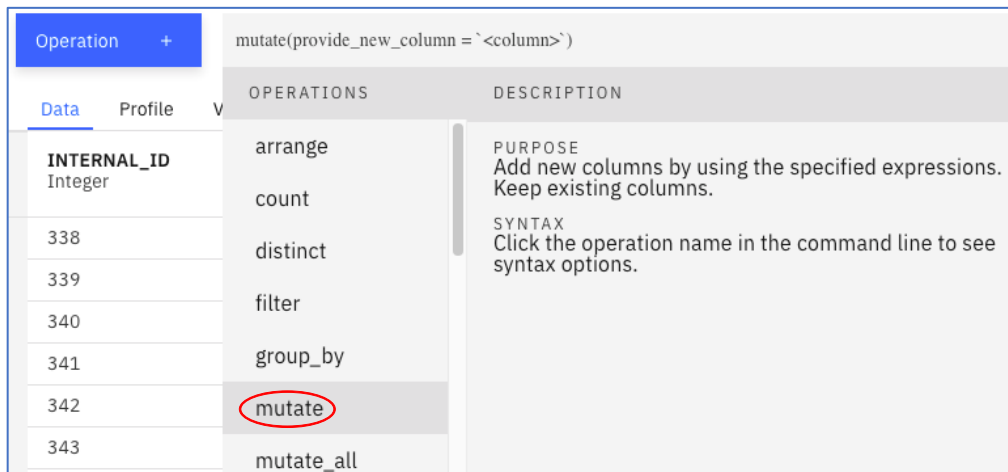
9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.



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 the entry field to the right of **Operations** +. Several operations are available.

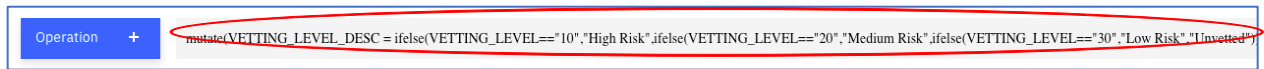


11. Hover the mouse over **mutate**. A description of the mutate function is provided.



12. Click on **mutate** and cut and replace the generated code with the following and then click **Apply**. Note, if an error occurs, it is because of a line break. Remove the line breaks and try again.

```
mutate(VETTING_LEVEL_DESC = ifelse(VETTING_LEVEL=="10","High Risk",ifelse(VETTING_LEVEL=="20","Medium Risk",ifelse(VETTING_LEVEL=="30","Low Risk","Unvetted"))))
```



13. On the right side of the text entry box, click **Apply**.



14. 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”.

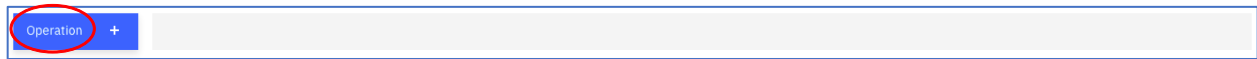
VETTING_LEVE...
String
Unvetted
Low Risk
High Risk
Low Risk
Unvetted
Unvetted
Unvetted
Unvetted
Medium Risk
Low Risk

15. Let’s extract the fields of interest by using another dplyr function, **select**. Cut and paste the following code into the operations area and click **Apply**. Again, remove the line breaks and try again if you get an error.

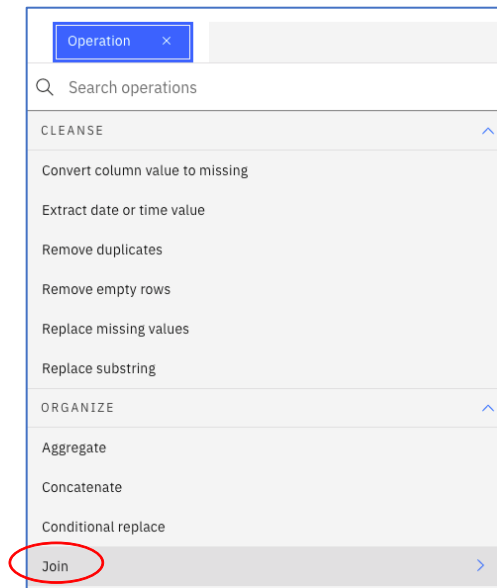
```
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DEPARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)
```



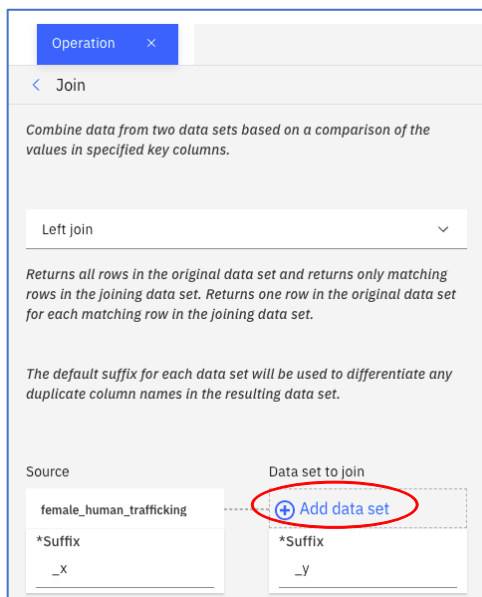
16. Let's now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the Categories dataset. Click on **Operation +**.



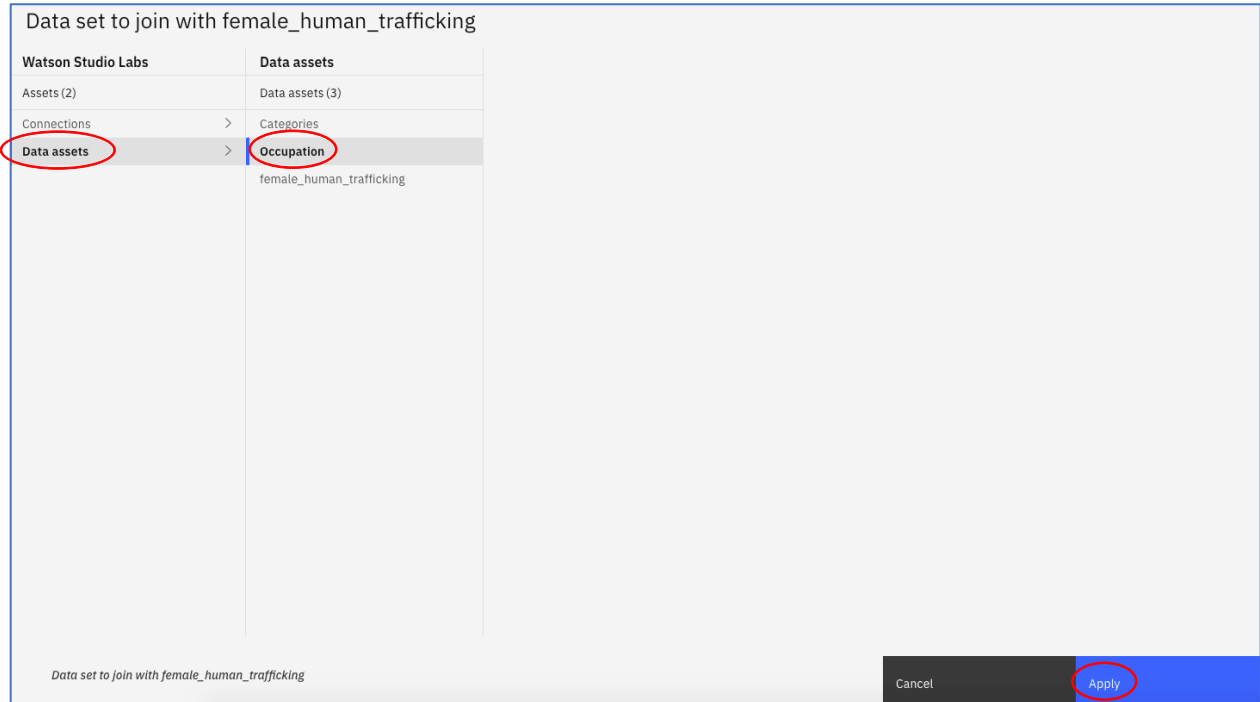
17. Scroll down and click on **Join**.



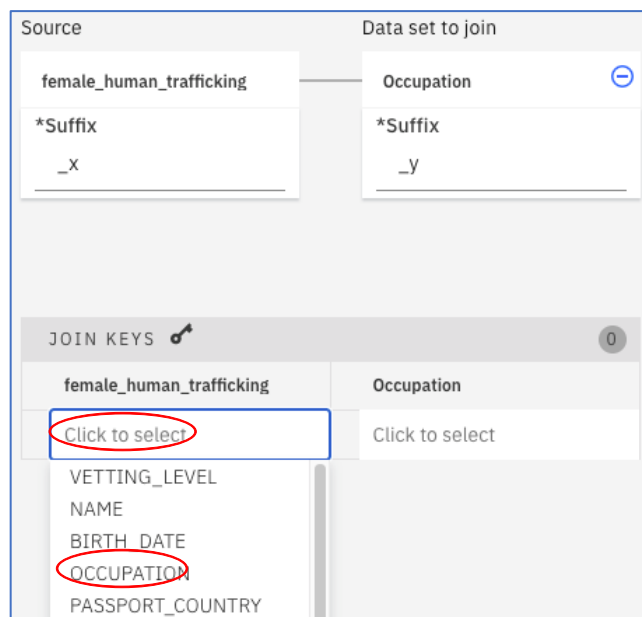
18. Keep **Left join** and then click on **Add Data Set**



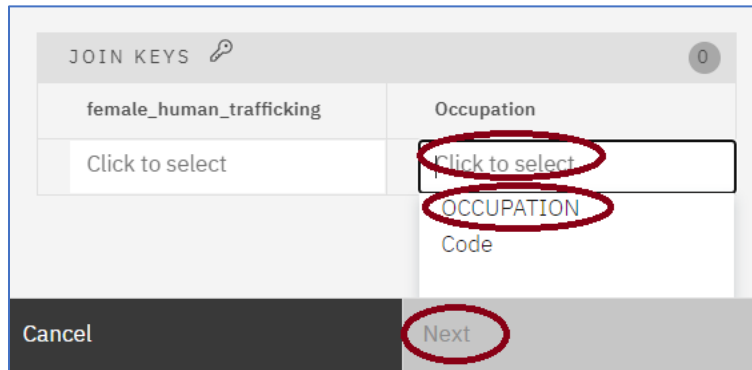
19. Click on **Data Assets**, click on **Occupation**, and then click **Apply**.



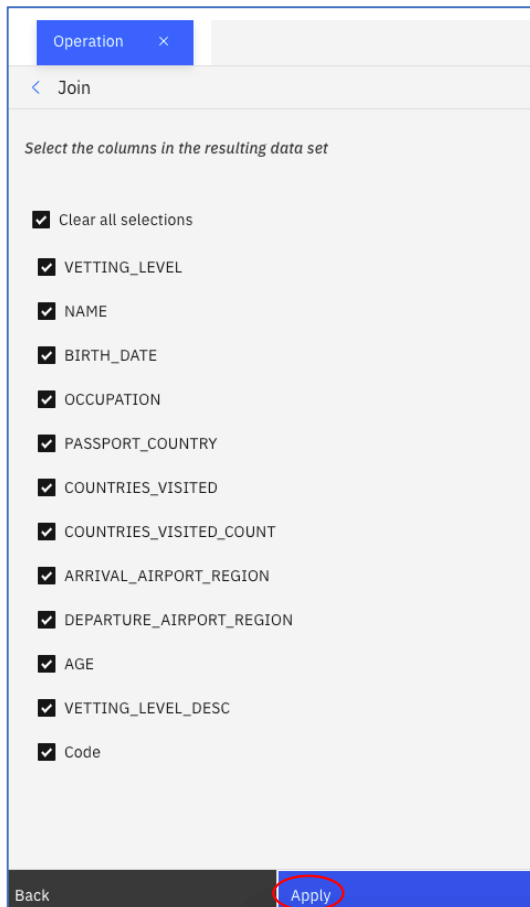
20. Scroll down. In **JOIN KEYS** under **female\_human\_trafficking** click **Click to select**, and then click **OCCUPATION**.



21. In **JOIN KEYS** under **Occupation** click **Click to select**, click **OCCUPATION**, and then 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. Note that your number of Steps may be different as Data Refinery may

have automatically converted columns. So far we have added a data source, converted two columns, entered two custom code commands, and completed two joins.

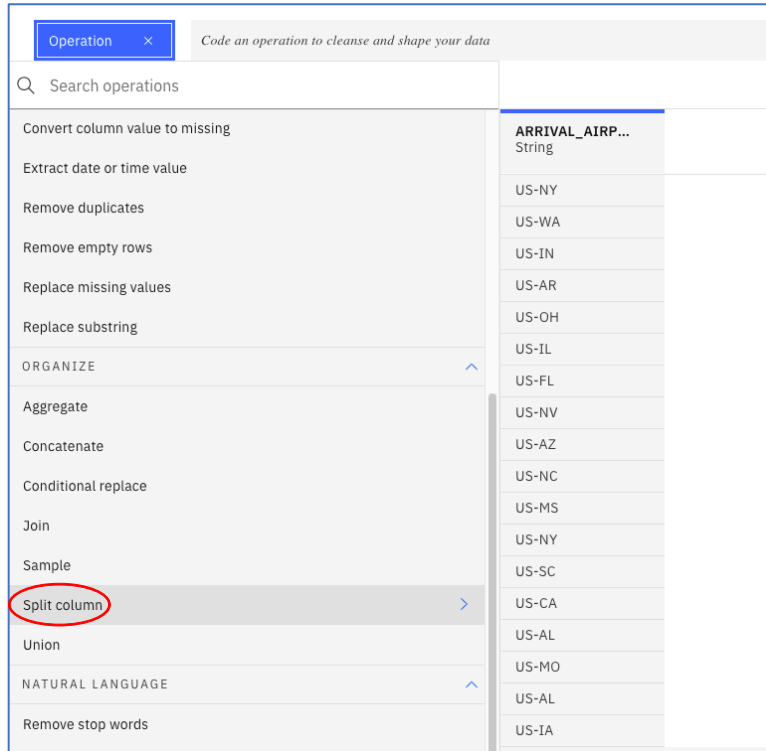
Code String	Category String	6 Steps
7	Science	Data Source
13	Education	female_human_trafficking
7	Science	
5	Government	Convert column type
11	Construction	Manually converted data types for 1 column.
5	Government	
8	Arts	Convert column type
15	Other	Manually converted data types for 1 column.
15	Other	
6	Medical	
13	Education	Custom code
6	Medical	mutate(VETTING_LEVEL_DESC = ifelse(VETTING_LEVEL=="10","High

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 ARRIVAL\_AIRPORT\_REGION to highlight the column then click on **Operation +**.

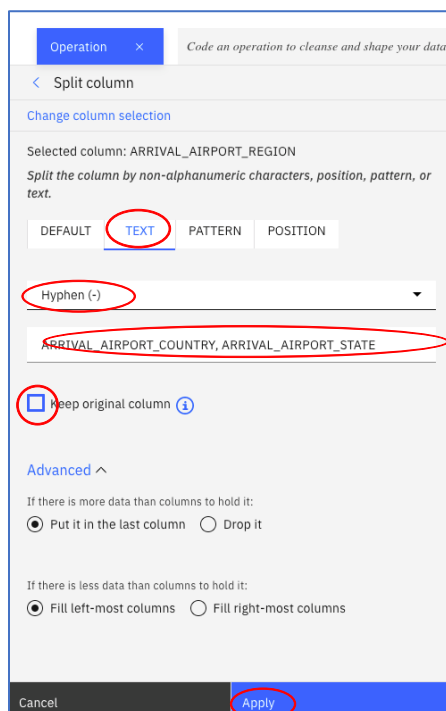
Operation + Code an operation to cleanse and shape your data		
COUNTRIES_VISITED String	COUNTRIES_VI... Integer	ARRIVAL_AIRP... String
1 QA	1	US-NY
2 QA	1	US-WA
3 ME,EE,KY,DZ,CZ,IO,NL,QA,BS,CK	10	US-IN
4 IL,VN,UZ	3	US-AR
5 ES,JO,LT,CL,QA,PA	6	US-OH
6 HR,BS,BG,AT,DK,AL,AL,OM,TN,LU,SI,IN	12	US-IL
7 OM,CK,BH,CK,TW,IQ,TN	7	US-FL
8 JP,RU,CO,CO,TR,TR	6	US-NV
9 JP,SN,SK,OM	4	US-AZ
10 RU	1	US-NC
11 RU	1	US-MS
12 AE	1	US-NY
13 CH,AE,LK	3	US-SC
14 TR,ES,KW,SG,RU,FI,KZ,BN,JM,PT	10	US-CA
15 RU,DZ,KR,SN,UA,TR,MT,RS,PK	9	US-AL
16 ZA,EG,LY,SA,UZ,MT,AZ	7	US-MO
17 KW,RU,BE,KY	4	US-AL




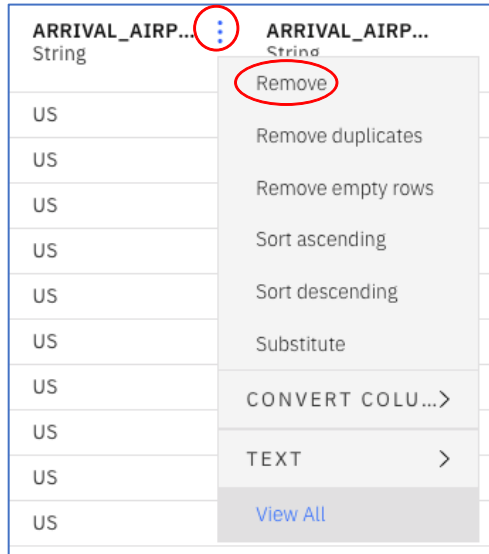
25. Click on **Split column**.



26. Click on **TEXT**, click on **Hyphen(-)** 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**.



27. 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**.



ARRIVAL_AIRP... String	ARRIVAL_AIRP... String
US	
US	
US	
US	
US	
US	
US	
US	
US	
US	
US	

We can also 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.

28. Let's split the **`COUNTRIES_VISITED`** column. Split by **TEXT**, change the column selection if needed, 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.

Operation

Code an operation to cleanse and shape your data

< Split column

Change column selection

Selected column: COUNTRIES\_VISITED

Split the column by non-alphanumeric characters, position, pattern, or text.

DEFAULT

TEXT

PATTERN

POSITION

Comma (,)

COUNTRY1, COUNTRY2, COUNTRY3

☒ Keep original column

Advanced

If there is more data than columns to hold it:

☐ Put it in the last column

☒ Drop it

If there is less data than columns to hold it:

☒ Fill left-most columns

☐ Fill right-most columns

Cancel

Apply

COUNTRIES\_VISITED

String

QA

QA

ME,EE,KY,DZ,CZ,ID,NL,QA,BS,CK

IL,VN,UZ

ES,JO,LT,CL,QA,PA

HR,BS,BG,AT,DK,AL,AL,OM,TN,LU,SI,IN

OM,CK,BH,CK,TW,IQ,TN

JP,RU,CO,CU,TR,TR

JP,SN,SK,OM

RU

RU

AE

CH,AE,LK

TR,ES,KW,SG,RU,FI,KZ,BN,JM,PT

RU,DZ,KR,SN,UA,TR,MT,RS,PK

ZA,EG,LY,SA,UZ,MT,AZ


KW,RU,BE,KY

BZ,KE,PA,BY,LU,SG,SK,QA,DE

SOURCE FILE:

29. The results are shown below.

COUNTRIES_VISITED String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VI... Integer
QA	QA			1
QA	QA			1
ME,EE,KY,DZ,CZ,ID,NL,QA,BS,CK	ME	EE	KY	10
IL,VN,UZ	IL	VN	UZ	3
ES,JO,LT,CL,QA,PA	ES	JO	LT	6
HR,BS,BG,AT,DK,AL,AL,OM,TN,LU,SI,IN	HR	BS	BG	12
OM,CK,BH,CK,TW,IQ,TN	OM	CK	BH	7

30. 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**.

VETTING_LEVEL	NAME
String	String
100	Remove
30	Remove duplicates
10	Remove empty rows
30	Sort ascending
10	Sort descending
100	Substitute
30	CONVERT COLU...>
100	TEXT >
30	View All
100	

31. Click on **Filter**.

Operation

Search operations

FREQUENTLY USED

Calculate

Convert column type

Filter

Math

Remove

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

Operation

Filter

Filter rows by the selected columns. Keep rows with the selected column values; filter out all other rows.

CONDITIONS (1)

CONDITION 1

Column

VETTING\_LEVEL

Operator

Does not contain

Choose to specify text or a pattern

Text


Pattern


100


Add condition

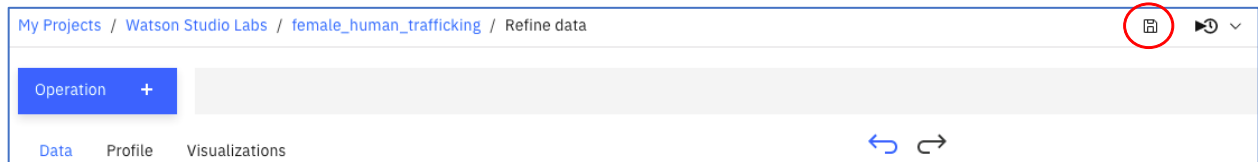
Cancel

Apply

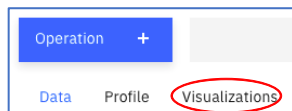
33. Remove the Code column by clicking on the vertical ellipse  and then clicking **Remove**.

Code		Category
String		String
		Remove
7		Remove duplicates
15		Remove empty rows
2		

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



35. Click on the **Visualization** tab.



36. Click on **VETTING\_LEVEL\_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.

CHART TYPE

Suggested charts

Pie

Bar

Word cloud

Sunburst

Tree


Treemap

Circle pa...

Bubble


Scatter plot

ACTIONS



Choose a chart above or select columns below, and then choose a chart. If you select columns, suggested charts will be indicated with a dot next to the chart name.

Columns to visualize



VETTING\_LEVEL\_DESC

+

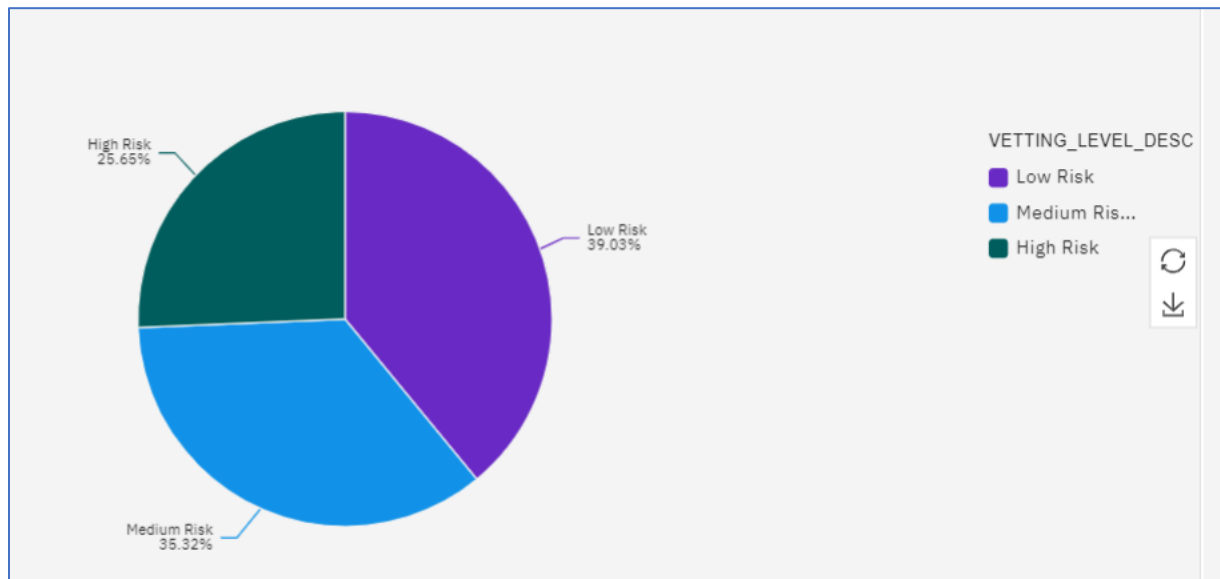
Add another column

SELECTED COLUMNS

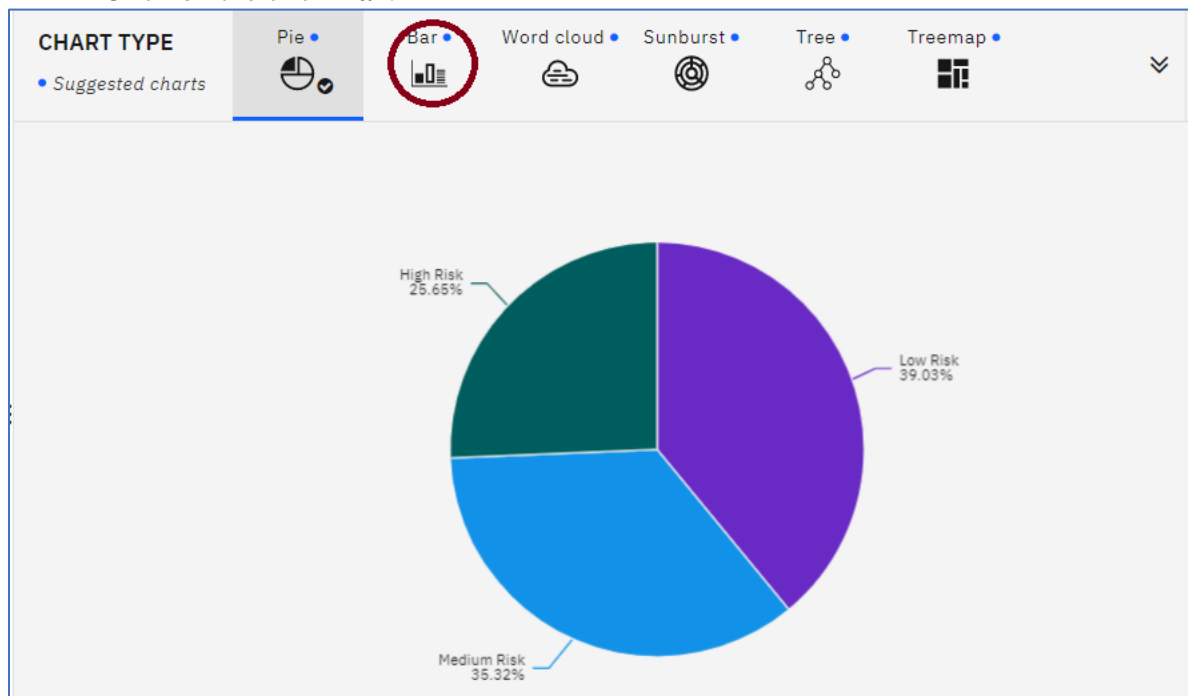
1

Visualize data

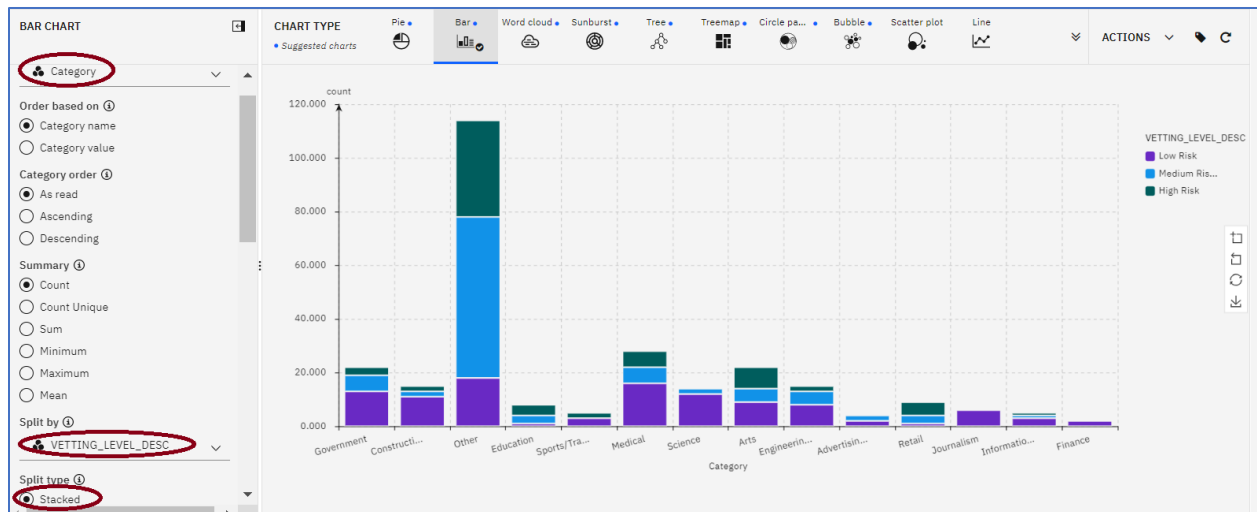
37. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



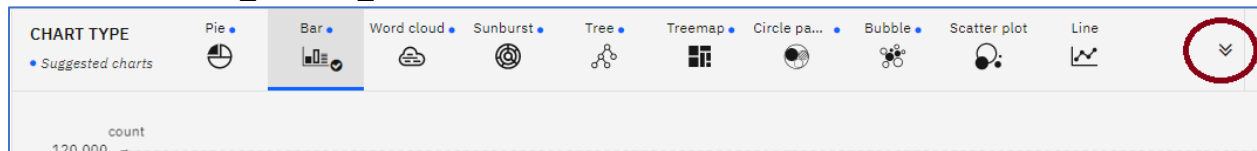
38. We can visualize the breakdown of travel records by job category and vetting level. Click on the click **Bar**.



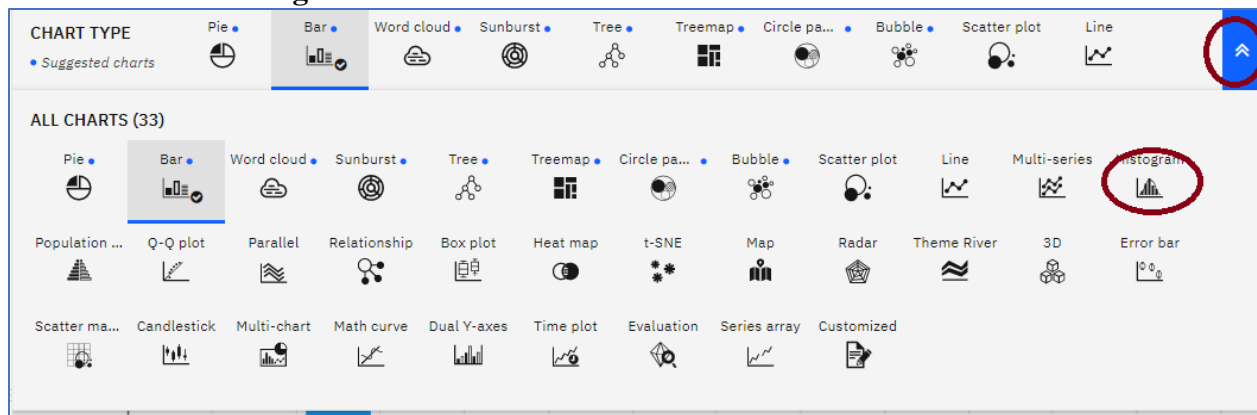
39. Click on **Category** for **Category**, click on VETTING\_LEVEL\_DESC for **Split by**, click on **Stacked** for **Split type**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



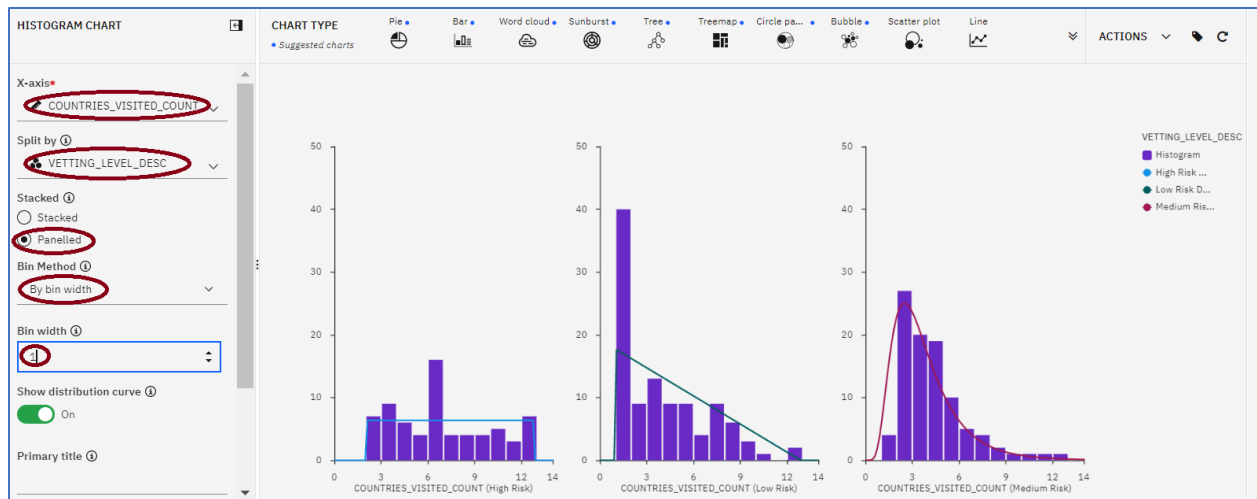
40. We can visualize a histogram of COUNTRIES\_VISITED\_COUNTS split by VETTING\_LEVEL\_DESC. Click on the  icon.



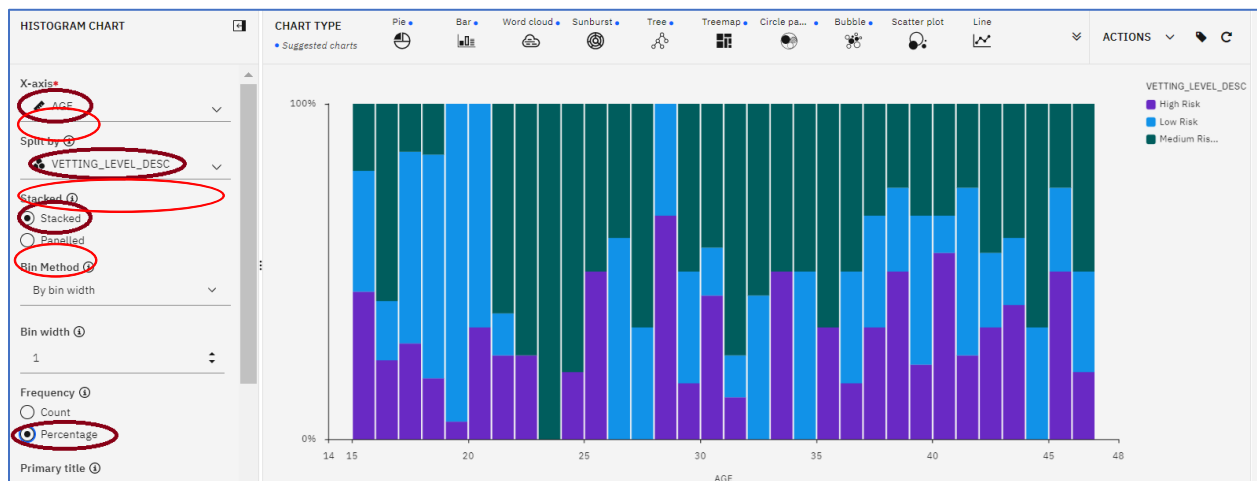
41. Click on **Histogram**



42. Click on COUNTRIES\_VISITED\_COUNT for **X-axis**, click on VETTING\_LEVEL\_DESC for **Split by**, click on **Paneled**, click on **By bin width** for the **Bin Method** and select 1 for the **Bin width**. Note that a higher number of high risk persons visit many countries.





43. Let's examine if age makes a difference. Click on **AGE** for **X-axis**. **Split by** remains **VETTING\_LEVEL\_DESC**, click on **Stacked**, and click on **Percentage**. There is not a clear pattern on the influence of age on high risk persons. It appears that younger travelers may have a slightly lower risk of being trafficked.

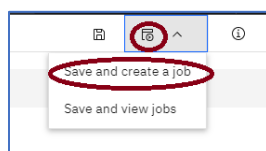


44. Please feel free to experiment with other visualizations.

## 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 job icon  .

45. Click on **job** icon  and click on **Save and create a job**.





46. Enter a **Job Name** for the job. Note the number of steps used to transform the data. It should be 11-13 steps depending on if Data Refinery automated column conversion and if any steps were skipped. A schedule can be set up if the transformation process needs to run on a scheduled basis. We are just going to do a one-time run. Click **Next**.

The screenshot shows the 'Create a job' dialog with the 'Define details' step selected. The 'Associated asset' is 'female\_human\_trafficking\_flow' with 13 steps. The 'Name' field is 'FHT Data Refinery', which is circled in red. The 'Description (optional)' field is empty. The 'Next' button is circled in red.

Create a job

Define details

Associated asset  
female\_human\_trafficking\_flow (13 Steps)

Name  
FHT Data Refinery

Description (optional)  
Description of job

Cancel

Next

47. Keep the default input, output, and environment and click **Next**.

The screenshot shows the 'Create a job' dialog with the 'Configure' step selected. The 'Data assets' section shows 'Input' as 'female\_human\_trafficking' and 'Output' as 'female\_human\_trafficking\_shaped', both with 'CSV' format. The 'Environment' is 'Default Data Refinery XS'. The 'Next' button is circled in red.

Create a job

Configure

Data assets

Input → Output

female\_human\_trafficking CSV → female\_human\_trafficking\_shaped CSV

Environment  
Default Data Refinery XS

Cancel

Back

Next

48. Keep schedule unenabled and click **Next**.

Create a job

- Define details  
FHT Data Refinery
- Configure  
Default Data Refinery XS
- Schedule**
- Review and create

**Schedule**

☐ Schedule off

Cancel Back **Next**

49. Click **Create and run**.

Create a job

- Define details  
FHT Data Refinery
- Configure  
Default Data Refinery XS
- Schedule
- Review and create**

**Review and create**

Details [Details](#)

Associated asset  
female\_human\_trafficki... (13 Steps)

Name  
FHT Data Refinery

Description  
[Add Description](#)

Configuration [Configuration](#)

Environment:  
Default Data Refinery XS

**Data assets**

**Input**  
female\_human\_trafficking CSV

→

**Output**  
female\_human\_trafficking\_... CSV

**Schedule** [Schedule](#)

**Scheduled to run**  
No schedule created

Cancel Back Create **Create and run**

50. Click on Job Details.

Projects / Watson Studio Labs / female\_human\_trafficking / Refine data

Operation [Code an operation to cleanse and shape your data](#)

**Data** Profile Visualizations

	VETTING_LEVEL	NAME	BIRTH_DATE	OCCUPATION	PASSPORT_CO...	LOCATION
1	30	Laura Smith	2000-03-01	Development worker, international aid	Ghana	
2	30	Sherry Alvarez	1999-10-20	Land/geomatics surveyor	Ghana	

**Details** Help

The job was successfully created. See job details.

11 Steps

Data Source  
female\_human\_trafficking

[Edit](#)

51. Wait until the job run changes from **Running** to **Completed**.

My projects / Watson Studio Labs / FHT Data Refinery

### Job Details

Overview

1  
Runs Completed

0  
Runs Failed

No schedule created

Edit Configuration

Find a job run

Start time	Status	Duration	Job	Asset type
May 16, 2021 1:54:41 PM Started by Horatio Doe	Completed	00:01:08 00:01:08 T	FHT Data Refinery	Data Refinery Flow

Last updated: 5/16/21, 1:58 PM

52. The output of the Data Refinery process should be listed in the Data Assets. Click on **Watson Studio Labs** to return to the Project view.

My Projects / **Watson Studio Labs** / FHT Data Refinery

53. Click on the **female\_human\_trafficking\_shaped.csv** to view the contents.

▼ Data assets

0 assets selected.

<input type="checkbox"/>	Name	Type	Created by	Last modified
<input type="checkbox"/>	CSV <b>female_human_trafficking_shaped</b>	Data Asset	FCTO Labs	Jan 10, 2021, 01:31 PM
<input type="checkbox"/>	CSV <a href="#">Occupation</a>	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
<input type="checkbox"/>	CSV <a href="#">Categories</a>	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
<input type="checkbox"/>	CSV <a href="#">female_human_trafficking</a>	Data Asset	FCTO Labs	Jan 10, 2021, 12:34 PM

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

My Projects / Watson Studio Labs / female\_human\_trafficking\_shap...

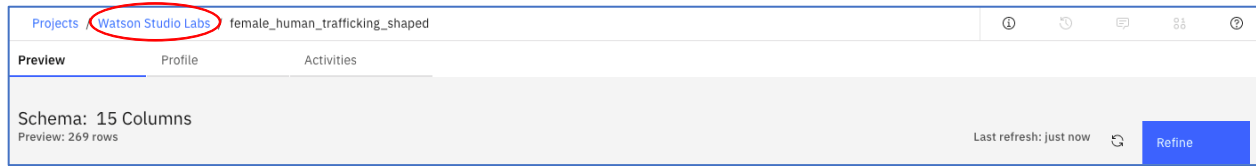
Preview Profile Activities

Schema: 15 Columns  
Preview: 269 rows

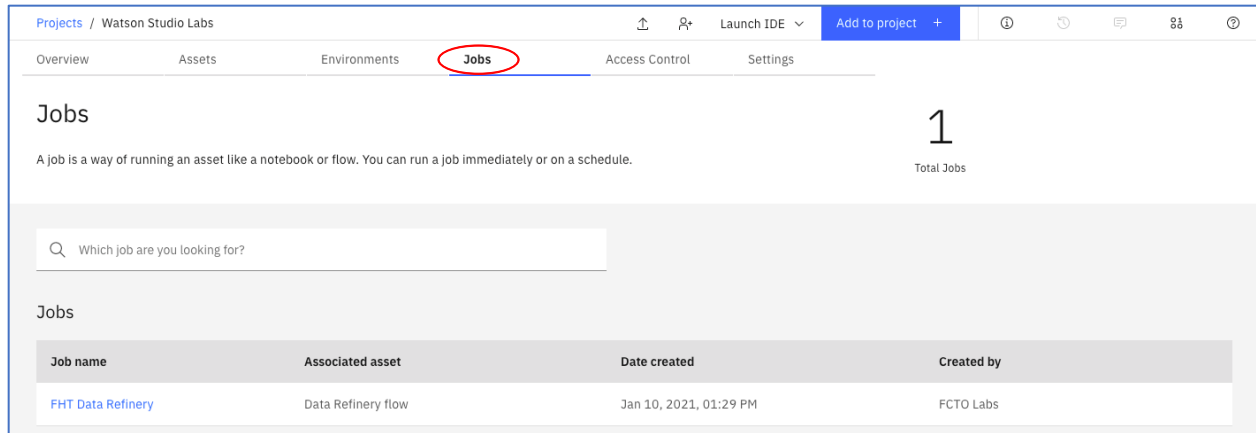
Last refresh: 16 seconds ago [Refine](#)

VETTING_L... String	NAME String	BIRTH_D... String	OCCUPAT... String	PASSPORT_COU... String	COUNTRIES_VIS... String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VISITED_C... String	ARRI... String
30.0	Trace Carr	11/30/01	Clinical scientist,	Ghana	QA	QA			1.0	WA
10.0	Ami Casey Wood	11/5/83	Cartographer	Ghana	ME,EE,KY,DZ,CZ,ID,NL,Q	ME	EE	KY	10.0	IN
30.0	Melinda Kimm Hi	1/16/80	Agricultural engi	Brazil	IL,VN,UZ	IL	VN	UZ	3.0	AR
10.0	Linda Tucker	1/14/95	Translator	Brazil	ES,JO,LT,CL,QA,PA	ES	JO	LT	6.0	OH
30.0	Brandy Scott	8/9/99	Field trials office	Ghana	OM,CK,BH,CK,TW,IQ,TN	OM	CK	BH	7.0	FL
30.0	Jesie Molly Staffi	5/2/70	Pathologist	Bangladesh	JP,SN,SK,OM	JP	SN	SK	4.0	AZ
30.0	Maireag Barker	9/24/01	Editor, film/video	Ghana	RU	RU			1.0	MS
30.0	Crysta Nann Silvi	8/6/98	Volunteer coordi	Ghana	AE	AE			1.0	NY
30.0	Tanya Cameron	3/24/97	Acupuncturist	Ghana	CH,AE,LK	CH	AE	LK	3.0	SC
10.0	Rebecca Good	3/2/74	Administrator, ec	Brazil	ZA,EG,LY,SA,UZ,MT,AZ	ZA	EG	LY	7.0	MO
10.0	Jacie Smith	1/23/01	Fine artist	Ghana	KW,RU,BE,KY	KW	RU	BE	4.0	AL
30.0	Alisha Cheryl Wa	10/11/97	Intelligence anal	Ghana	OM	OM			1.0	PA

55. Click on **Watson Studio Labs** to return to the project view.



56. Click on the **Jobs** tab to view the Jobs facility. We can see the Data Refinery job status.



**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.