

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 4 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.

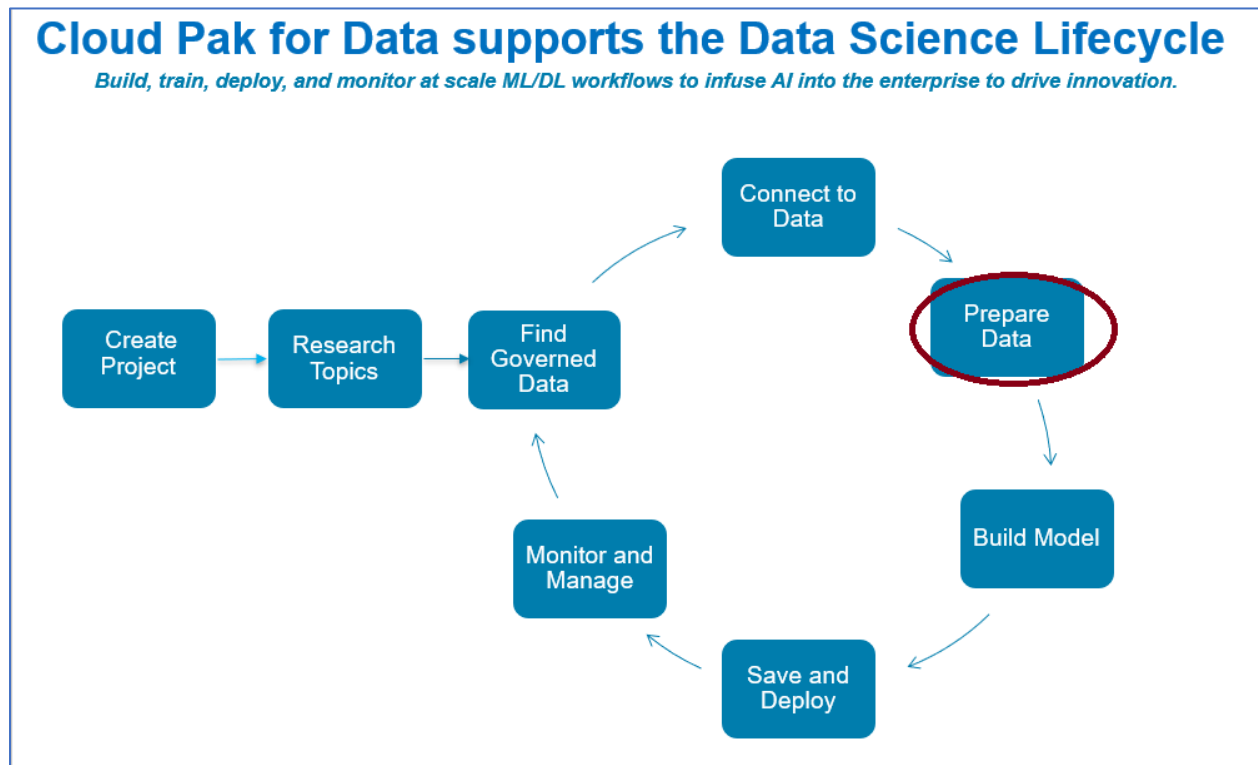


Figure 1- End to End Flow

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. This data is contained in a DB2 Warehouse table and a MongoDB Database. 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.

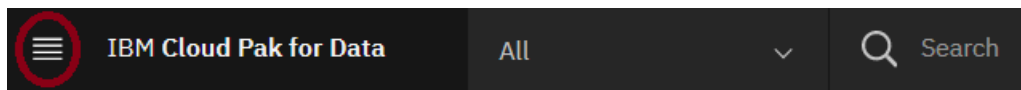
The OCCUPATION field included in the travel data is very granular. For modeling purposes, it was decided to categorize the OCCUPATION data. Two additional datasets are used for this purpose. The Occupation dataset maps the granular occupation data to a category code. The Categories dataset maps a category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

Other columns 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.

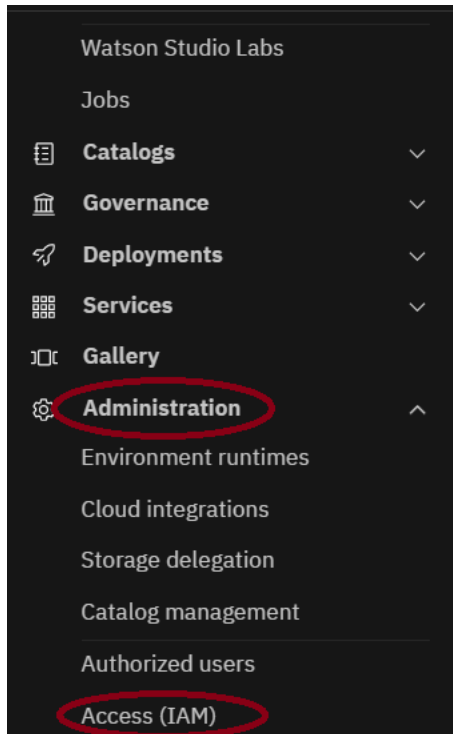
Remove User

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

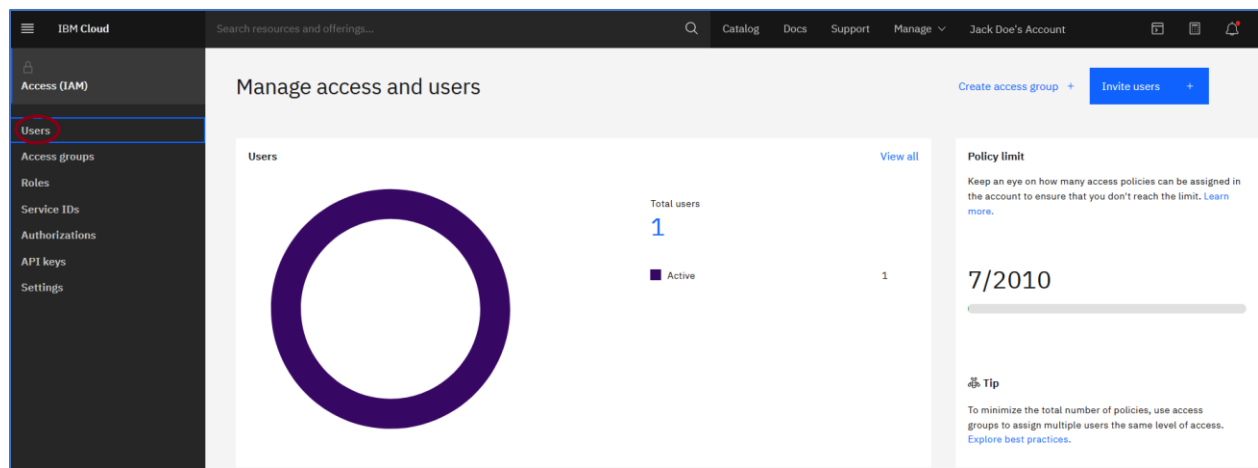
1. Click on the hamburger  icon




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
Howard Doe	ws.catalog.user@gmail.com	Active
Victor Doe owner self	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.

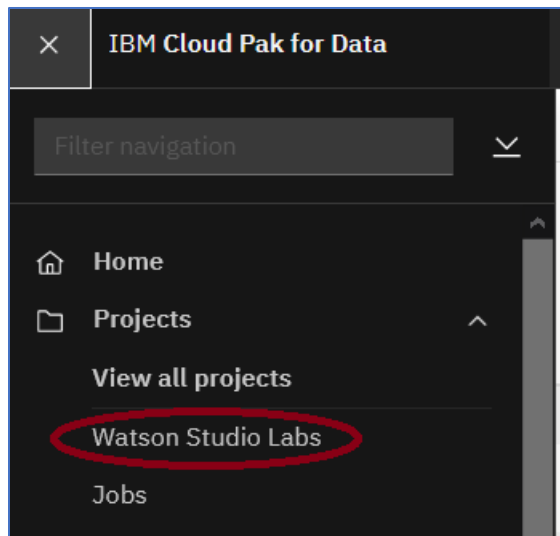
IBM Cloud Pak for Data × Identity & Access Management ×

Create a new Data Flow

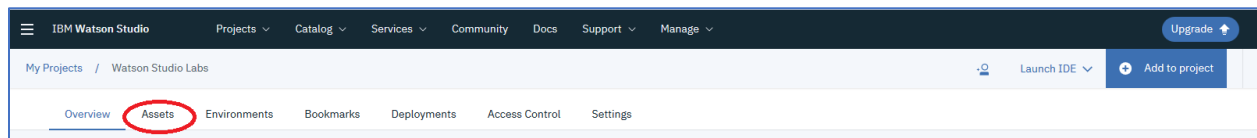
1. Click on the hamburger icon ☰.

☰ IBM Cloud Pak for Data

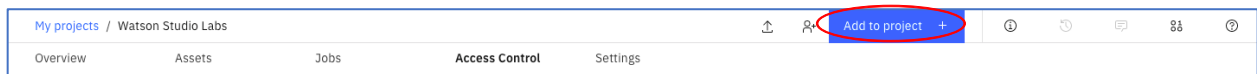
2. Click on **Watson Studio Labs** under **Projects**.



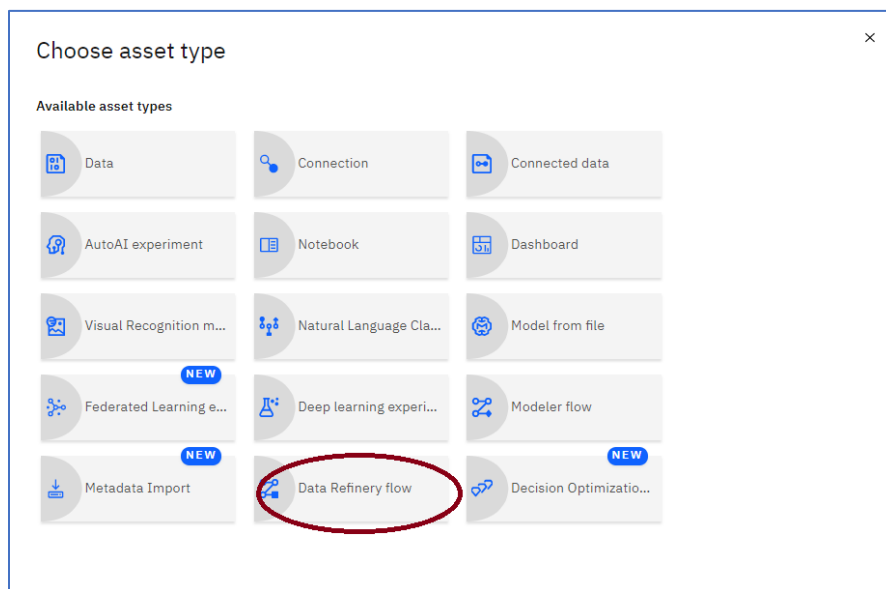
3. Click on the **Assets** tab.



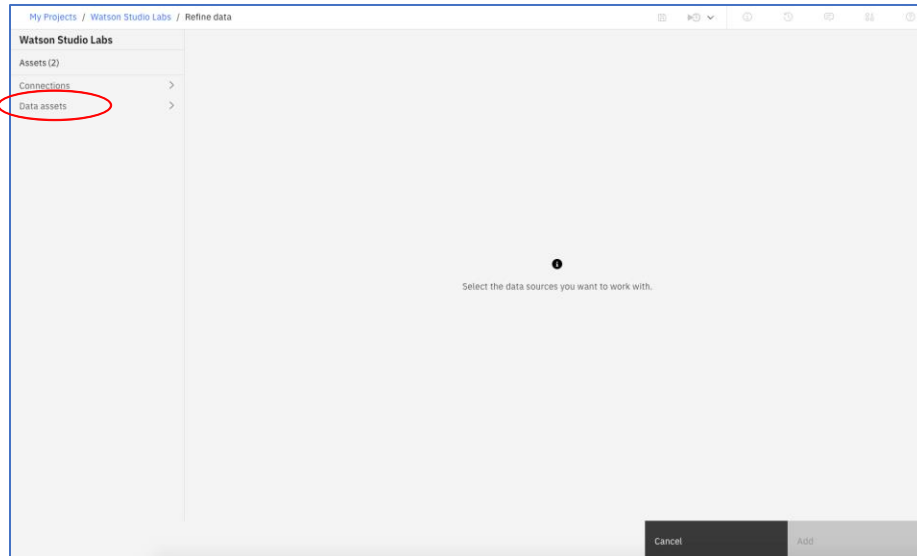
4. Add a Data Flow by clicking on **Add to project**.



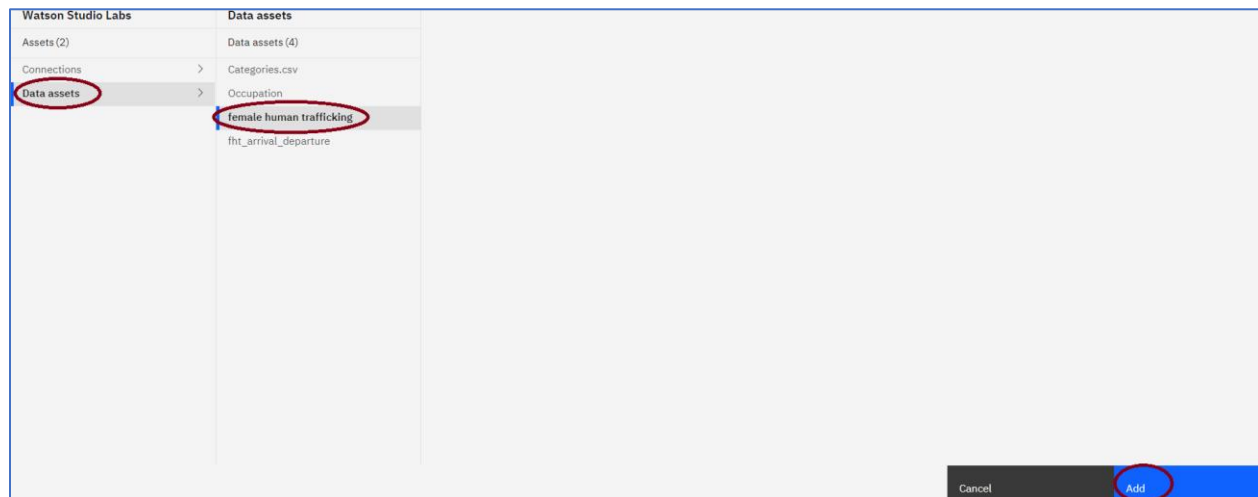
5. Click on **Data Refinery flow**



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.

IBM Cloud Pak for Data

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...

	INTERNAL_ID	VETTING_LEVEL	NAME	GENDER	BIRTH_DATE	BIRTH_COUNTRY	BIRTH_COUNTR...	OCCUPA
1	215	30	Cristie Moore	F	1998-07-16	Ghana	GH	Osteopa
2	216	100	Susan Ashley Long	F	2000-05-09	Ghana	GH	Psychol
3	217	100	Michelle Brown	F	1980-11-17	Ghana	GH	Occupat
4	218	100	Jessica Cheryl Dunn	F	1973-05-06	Ghana	GH	Clinical
5	219	100	Amri Washington	F	2000-01-20	Ghana	GH	Custom
6	220	100	Bonnie Smith	F	1972-10-03	Ghana	GH	Chemica
7	221	100	Isa Wilson	F	1973-04-01	Ghana	GH	Televisi
8	222	100	Heather Luna	F	1985-06-24	Ghana	GH	Phytoth
9	223	100	Tina Castro	F	1978-09-08	Ghana	GH	Press ph
10	224	100	Amber Montes	F	1984-07-06	Ghana	GH	Optician
11	225	100	Shannie Stevens	F	1983-05-07	Ghana	GH	Tax advi
12	785	20	Sara Nunez	F	1978-12-28	Ghana	GH	Industri

Details

LOCATION
Watson Studio Labs


DATA REFINERY FLOW NAME
female human traffickin...

Enter a description of the Data Refinery flow

STEPS
0


Prepare, Profile, Visualize


Before profiling the data, we will do some data preparation.


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




1. The INTERNAL_ID field is used later as a key field. If it is not of Integer type, we will convert it to be of Integer type. Click on the vertical ellipse  adjacent to the INTERNAL_ID field. Click on **CONVERT COLU...**, and then click on **Integer**.

INTERNAL_ID Decimal		VETTING_LEVEL Decimal	NAME String
215		Remove	Cristie Moore
216		Remove duplicates	Susan Ashley Long
217		Remove empty rows	Michelle Brown
218		Sort ascending	Jessica Cheryl Dunn
219		Sort descending	Ami Washington
220		Substitute	Bonnie Smith
221		CONVERT COLU...>	✓ Decimal
222		View All	Integer
223			String
224		100	
225		100	Shannie Stevens

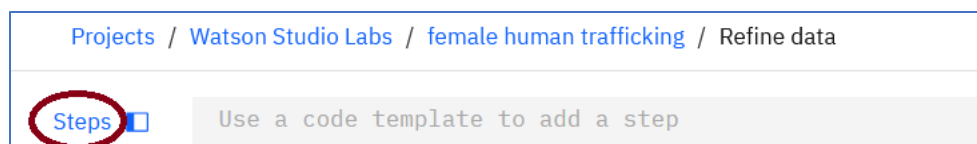
2. Some of the columns in the data set are defined as numerical but should be treated as Strings. We can easily convert the columns from Integers or Decimals to Strings. Convert the **VETTING_LEVEL** column by hovering over VETTING_LEVEL, clicking on the vertical ellipse , clicking on **CONVERT COLU...**,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

VETTING_LEVEL	NAME	GENDER
Decimal	String	String
30		F
100		F
100		F
100		F
100		F
100		F
100		F
100		F
100		F
100	Amber Montes	
100	Shannie Stevens	F
20	Sara Nunez	F

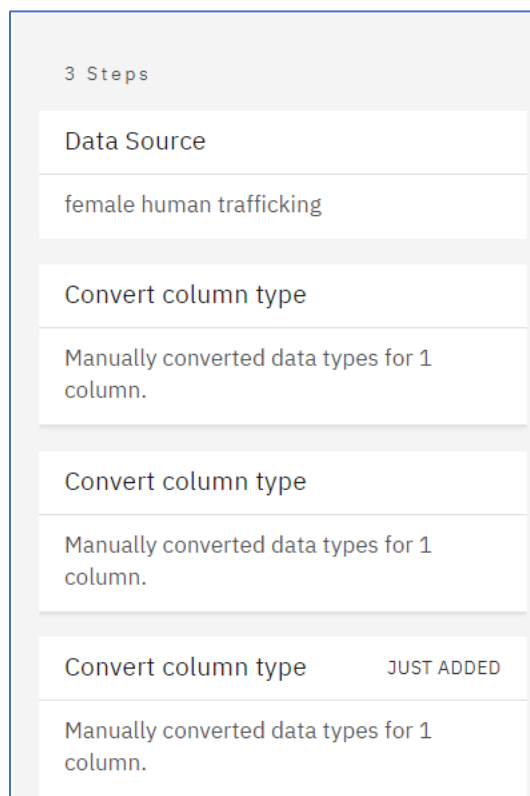
- Convert the **PASSPORT_NUMBER** column by hovering over **PASSPORT_NUMBER**, clicking on the vertical ellipse , clicking on **CONVERT COLU...**, and clicking on **String**.

PASSPORT_NU...	PASSPORT_COU...	PASSPORT_COU...	CO
Decimal	String	String	Str
156564572		GH	KG
470901453		GH	PG
151443012		GH	RS
800966716		GH	UZ
812911909		GH	AE
493104471		GH	IT,
326013301			.K
325326635			IM
930261083			Q,
17087551	Ghana		DA
894048064	Ghana	GH	OM
775799149	Ghana	GH	HU

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



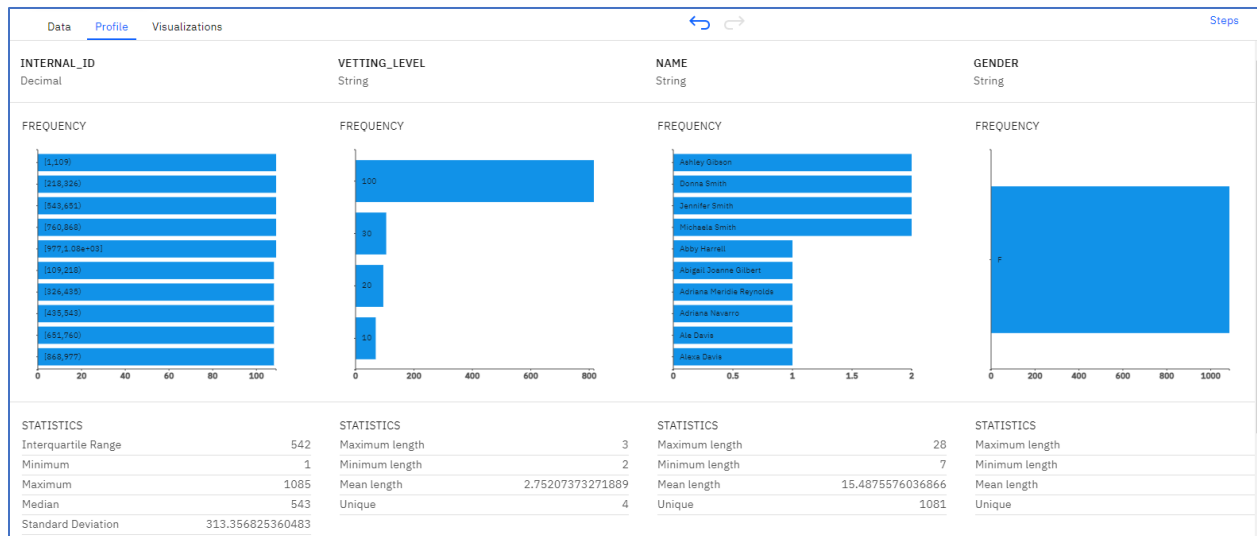
- Each data operation is recorded in the **Steps** display providing an audit list of the operations performed. So far, we have done three (2 if you didn't need to convert the INTERNAL_ID field) column conversion operations. The steps in the **Steps** display can be edited. Operations can be removed from the list or modified.



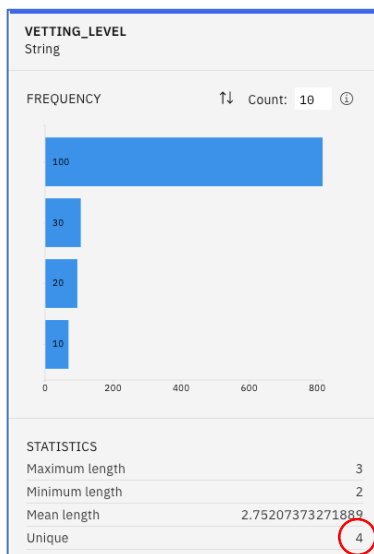
- Click on **Profile**.

	SSN String	PASSPORT_NUM... 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

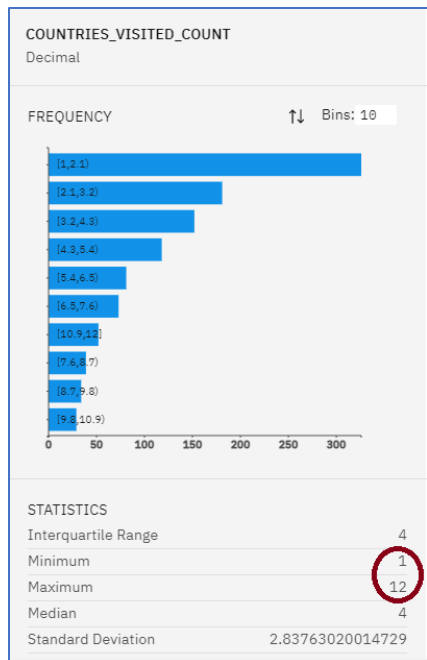
- The Profile panel displays the counts of the top 10 values for each categorical column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column.



- 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.



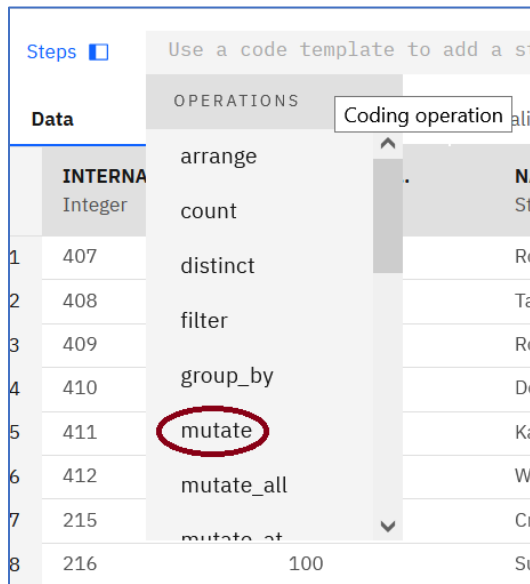
- Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has about 492 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH_COUNTRY, and PASSPORT_COUNTRY shows only 6 unique countries. The COUNTRIES_VISITED_COUNT shows that passengers have visited between 1 and 12 countries. Note, the results may be slightly different on your screen.



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



11. 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 **Steps**. Several operations are available. 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"))))
```

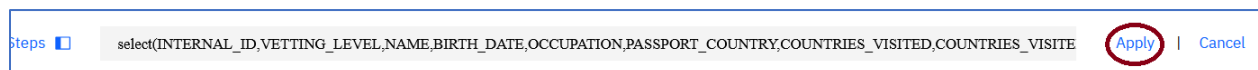


13. If you scroll to the right you should see the new column VETTING_LEVEL_DESC with values “Low Risk”, “Medium Risk”, “High Risk”, and “Unvetted”.


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

14. Let’s extract the fields of interest by using another dplyr function, **select**. Copy 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(INTERNAL_ID,VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,AGE,VETTING_LEVEL_DESC)
```



15. Let’s now join the fht_arrival_departure data. The fht_arrival_departure data provides information on the arriving location and the departing location of each passenger. Click on **Steps**

Steps 

Use a code template to add a step

16. Click **New step**

Data Source

female human trafficking

1. Convert column type

Manually converted data types for 1 column.

2. Convert column type

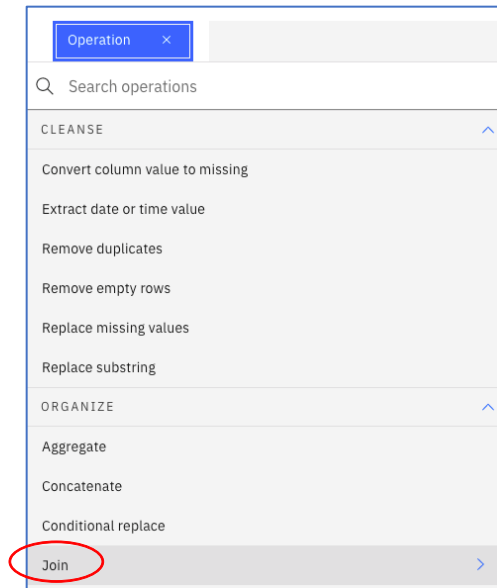
Manually converted data types for 1 column.

3. Custom code JUST ADDED

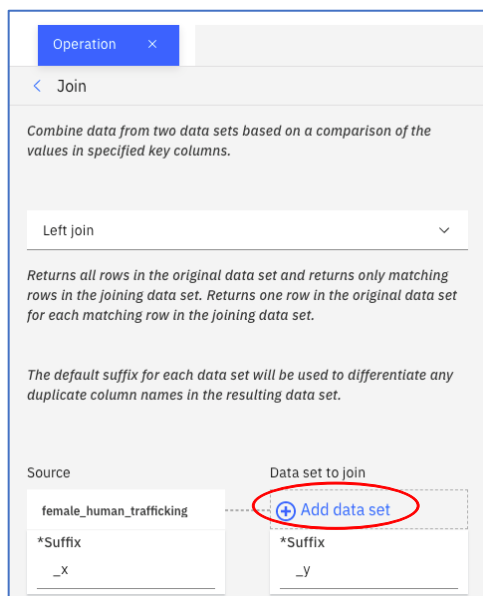
```
mutate(VETTING_LEVEL_DESC =  
  ifelse(VETTING_LEVEL=="10","High  
Risk",ifelse(VETTING_LEVEL=="20","Medium  
Risk",ifelse(VETTING_LEVEL=="30","Low  
Risk","Unvetted"))))
```

New step +

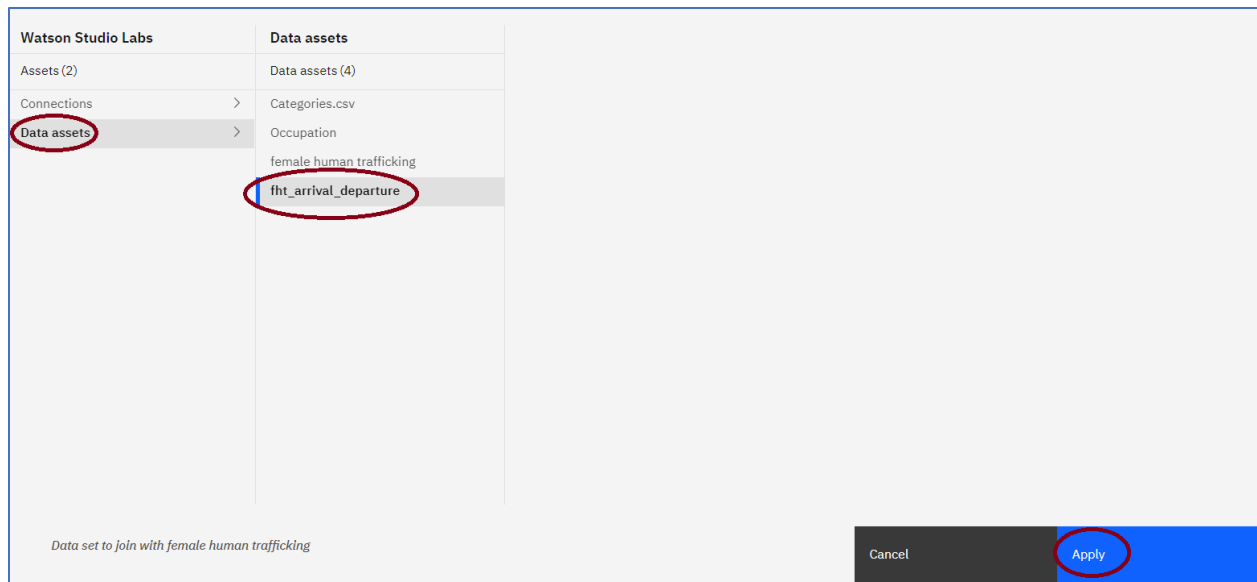
17. Scroll down and under the **ORGANIZE** category, click on **Join**.



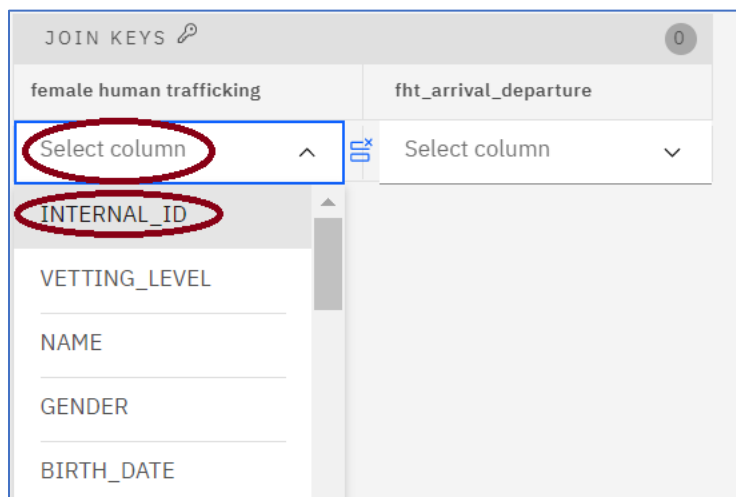
18. Keep **Left join** and then click on **Add Data Set**



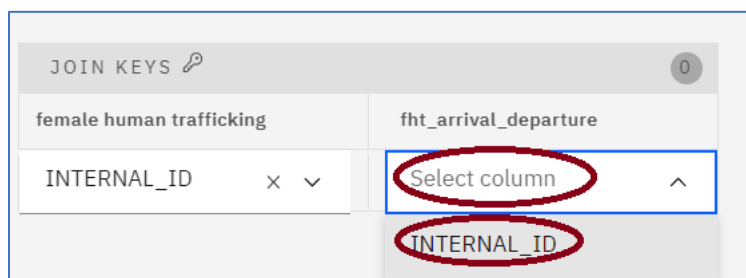
19. Click on **Data assets**, **fht_arrive_departure**, and then click **Apply**.



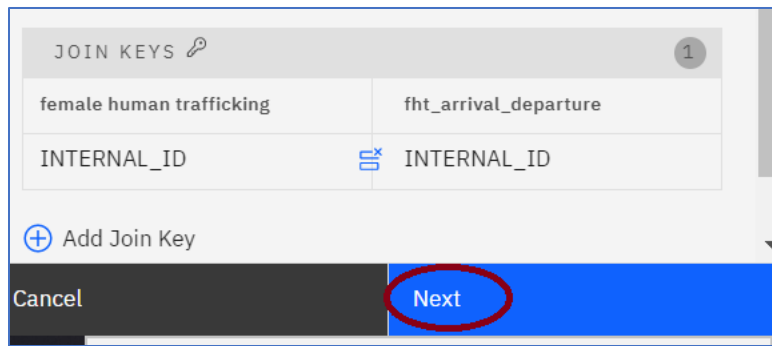
20. Scroll down. In **JOIN KEYS** under **female_human_trafficking** click **Select column**, and then click **INTERNAL_ID**.



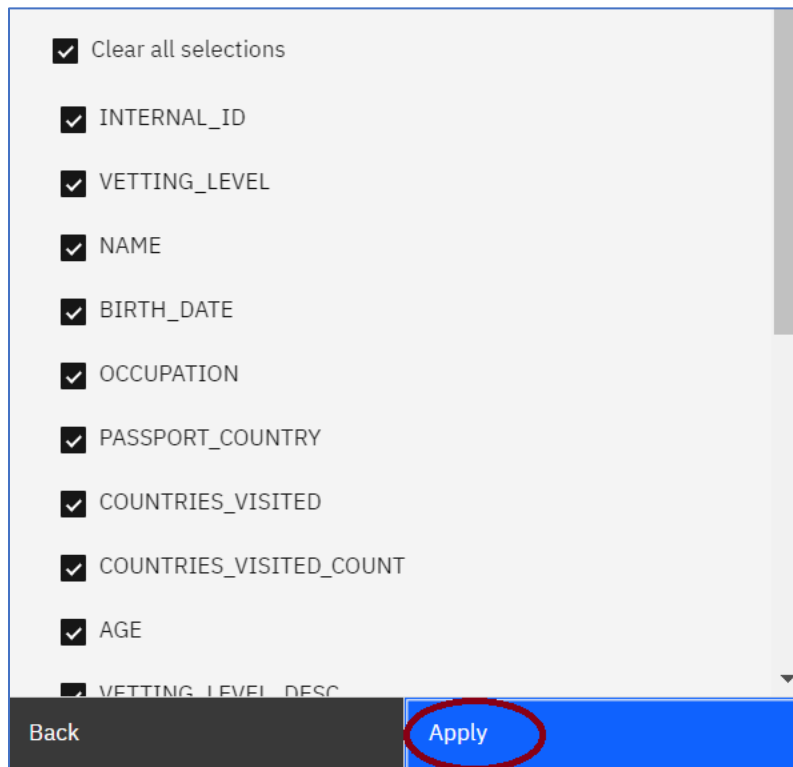
21. Similarly, in **JOIN KEYS** under **fht_arrival_departure**, click **Select column**, and click on **INTERNAL_ID**.




22. Click **Next**.



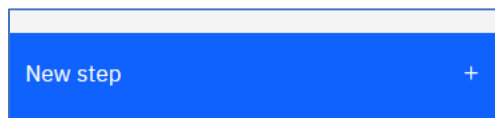
23. Click **Apply**



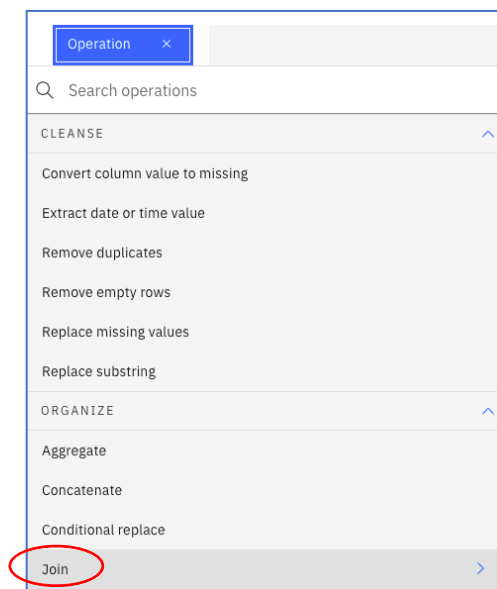
24. Remove the `_ID` column. This is an internally generated ID in MongoDB. Click on the vertical ellipse  icon. Click **Remove**.

_ID String	Details	
	Remove	
6121276366F278B526C89E29	Remove duplicates	
6121276366F278B526C89E2A	Remove empty rows	
6121276366F278B526C89E2B	Sort ascending	
6121276366F278B526C89E2C	Sort descending	
6121276366F278B526C89E2D	Substitute	
6121276366F278B526C89E2E	CONVERT COLU...>	
6121276366F278B526C89E2F	TEXT >	
6121276366F278B526C89E30	View All	
6121276366F278B526C89E31		
6121276366F278B526C89E32		
6121276366F278B526C89E33	6	
6121276366F278B526C89E34		

25. Let's join the **Occupations** data set. Click on **New step**



26. Scroll down and under the **ORGANIZE** category, click on **Join**.



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

Operation

×

<

Join

Combine data from two data sets based on a comparison of the values in specified key columns.

Left join

▼

Returns all rows in the original data set and returns only matching rows in the joining data set. Returns one row in the original data set for each matching row in the joining data set.

The default suffix for each data set will be used to differentiate any duplicate column names in the resulting data set.

Source

Data set to join

female_human_trafficking

+

Add data set

*Suffix

*Suffix

_X

_Y

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

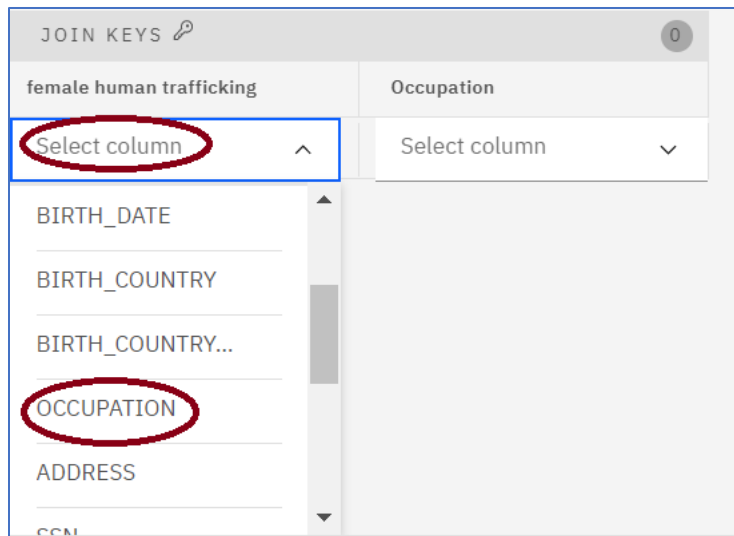
Watson Studio Labs		Data assets
Assets (2)		Data assets (4)
Connections	>	Categories.csv
Data assets	>	Occupation
		female human trafficking
		fht_arrival_departure

Data set to join with female human trafficking

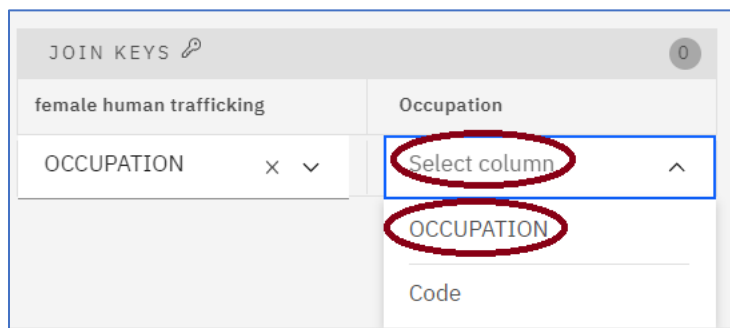
Cancel

Apply

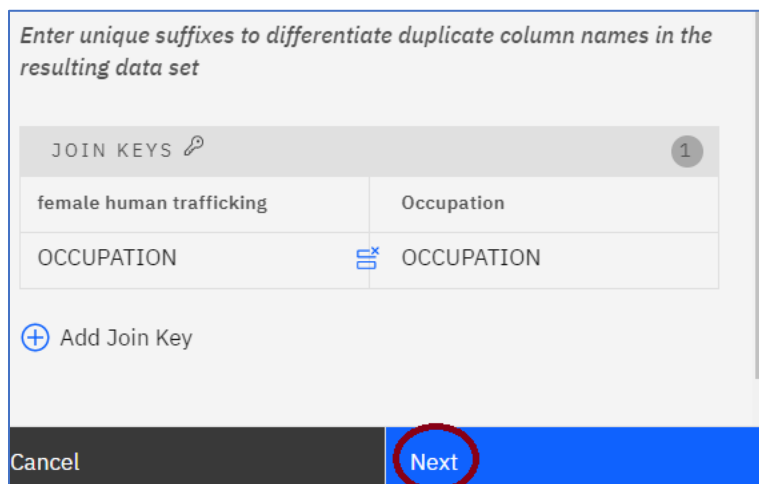
29. Scroll down. In **JOIN KEYS** under **female_human_trafficking**, click **Select column** and then click **OCCUPATION**.



30. In **JOIN KEYS** under **Occupation** click **Select column**, click **OCCUPATION**



31. Click **Next**.



32. Click **Apply**.

Select the columns in the resulting data set

- ☒ Clear all selections
- ☒ INTERNAL_ID
- ☒ VETTING_LEVEL
- ☒ NAME
- ☒ GENDER
- ☒ BIRTH_DATE
- ☒ BIRTH_COUNTRY
- ☒ BIRTH_COUNTRY_CODE
- ☒ OCCUPATION
- ☒ ADDRESS
- ☐ SSN

Back Apply

33. Follow steps 25-32 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 three columns, entered two custom code commands, and completed three joins.

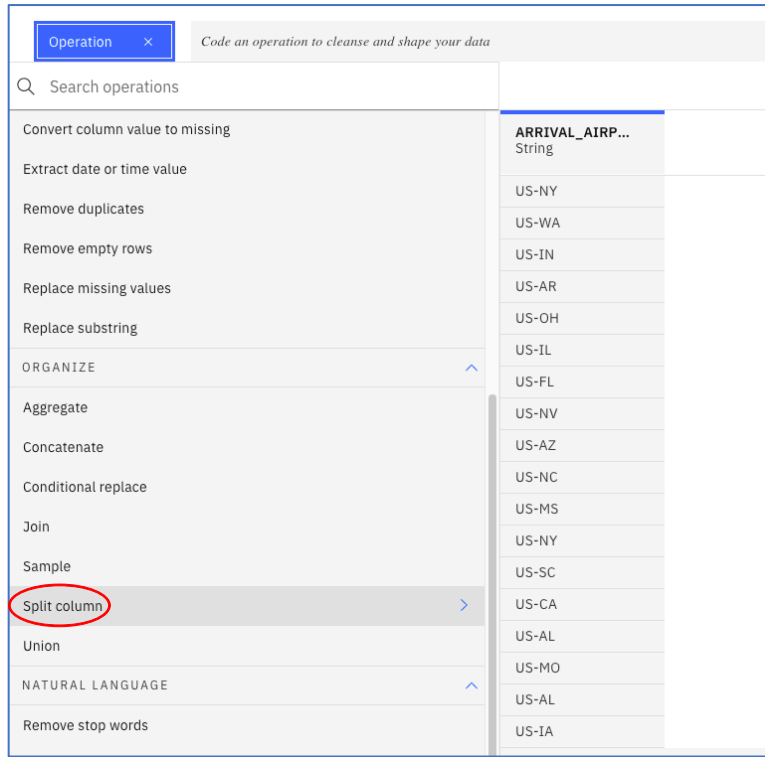
Code String	Category String	9 Steps
6	Medical	Data Source
6	Medical	female human trafficking
6	Medical	
7	Science	Convert column type
12	Retail	Manually converted data types for 1 column.
2	Engineering	
4	Journalism	Convert column type
6	Medical	Manually converted data types for 1 column.
4	Journalism	
6	Medical	Convert column type
14	Finance	Manually converted data types for 1 column.
15	Other	

34. 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 **New step**.

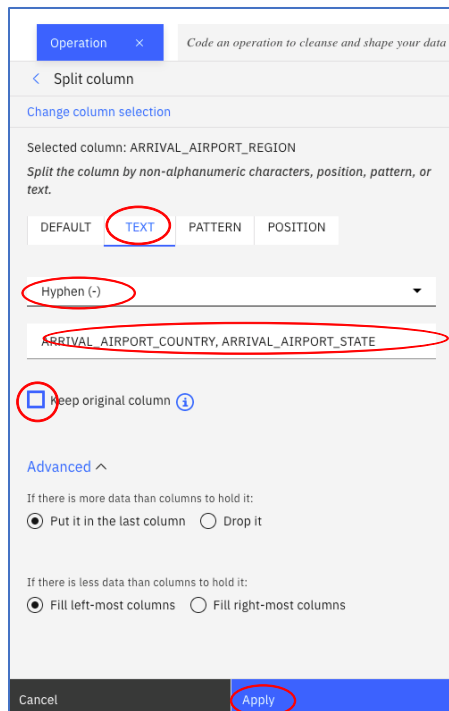
Data Source	Data	Profile	Visualizations
female human trafficking	COUNTRIES... Integer	AGE Integer	VETTING_L... String
1. Convert column type			ARRIVAL_A... String
Manually converted data types for 1 column.			
2. Convert column type			
Manually converted data types for 1 column.			
3. Custom code			
mutate(VETTING_LEVEL_DESC = ifelse(VETTING_LEVEL=="10","High Risk",ifelse(VETTING_LEVEL=="20","Medium Risk",ifelse(VETTING_LEVEL=="30","Low Risk","Unvetted"))))			
4. Join			
left-joined data from fht_arrival_departure based on columns			
New step			


SOURCE FILE:	female human trafficking	FULL DATA SET:	1085 rows
--------------	--------------------------	----------------	-----------

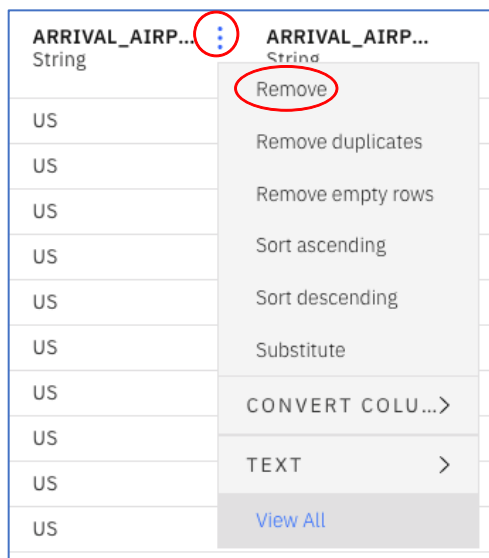
35. Click on **Split column**.



36. 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**.



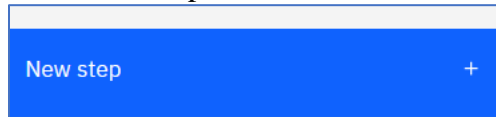
37. Two new columns are created. We don't need the ARRIVAL_AIRPORT_COUNTRY since it has only 1 value – US. Remove the ARRIVAL_AIRPORT_COUNTRY by hovering over the ARRIVAL_AIRPORT_COUNTRY header, clicking on the vertical ellipse  and clicking on **Remove**.



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.

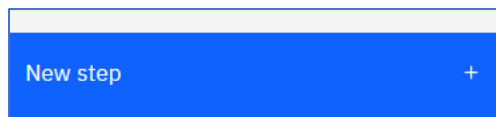
38. Let’s split the COUNTRIES_VISITED column. Click on New step.



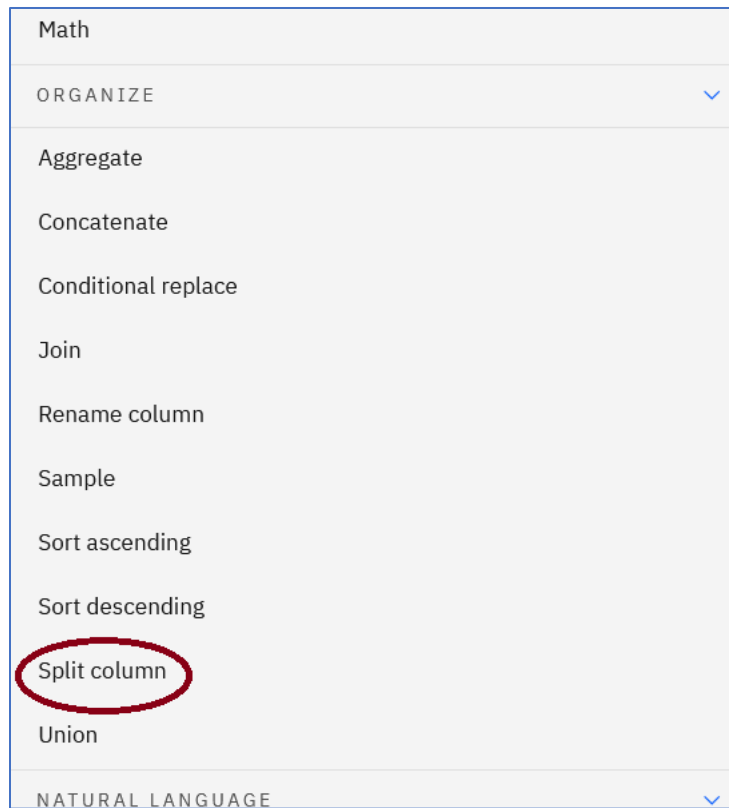
39. Let’s split the **COUNTRIES_VISITED** column. Scroll horizontally and click on the **COUNTRIES_VISITED COLUMN**

PASSPORT_...	COUNTRIES_VISITED	COUNTRIES...	AGE
String	String	Integer	Integer
BR	JO,IR,LB,KR,EE,HU,SK,EC,CN,RU,AE,PH	12	41
GH	AL,CY,QA,BS	4	19
GH	QA	1	33
GH	NO,RU	2	27
GH	QA,KH	2	27
GH	OM,JP,BE	3	46
GH	KG,LY,QA,DE	4	18
GH	PG,AE,HU	3	16
GH	RS,UY,BY,NZ,OM	5	36
GH	UZ	1	43
GH	AE,LU,IQ,UY,IE	5	17
GH	IT,TH,UZ	3	44
GH	LK,IT,QA,JM,NZ	5	43
GH	3M,QA,LB,BS	4	24

40. Click **New step**.



41. Scroll down and click on **Split column**



42. 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:

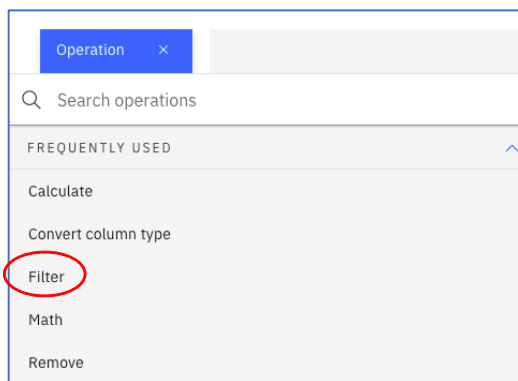
43. The results are shown below.

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

44. 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
	Remove
100	Remove duplicates
30	Remove empty rows
10	Sort ascending
30	Sort descending
10	Substitute
100	CONVERT COLU...>
30	TEXT >
100	View All

45. Click on **Filter**.




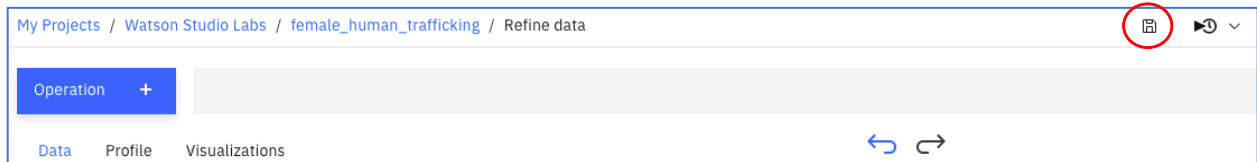
46. Change **Operator** to **Does not contain**, put value as 100, and then click **Apply**.

A screenshot of a 'Filter' dialog box. At the top, it says 'Filter rows by the selected columns. Keep rows with the selected column values; filter out all other rows.' Below this, there is a section for 'CONDITIONS (1)'. Under 'CONDITION 1', the 'Column' is set to 'VETTING_LEVEL' and the 'Operator' is set to 'Does not contain' (circled in red). Below the operator, there is a text input field containing '100' (circled in red). At the bottom, there are 'Cancel' and 'Apply' buttons, with 'Apply' circled in red.

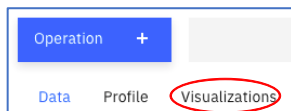
47. Remove the Code column by clicking on the vertical ellipse and then clicking **Remove**.

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

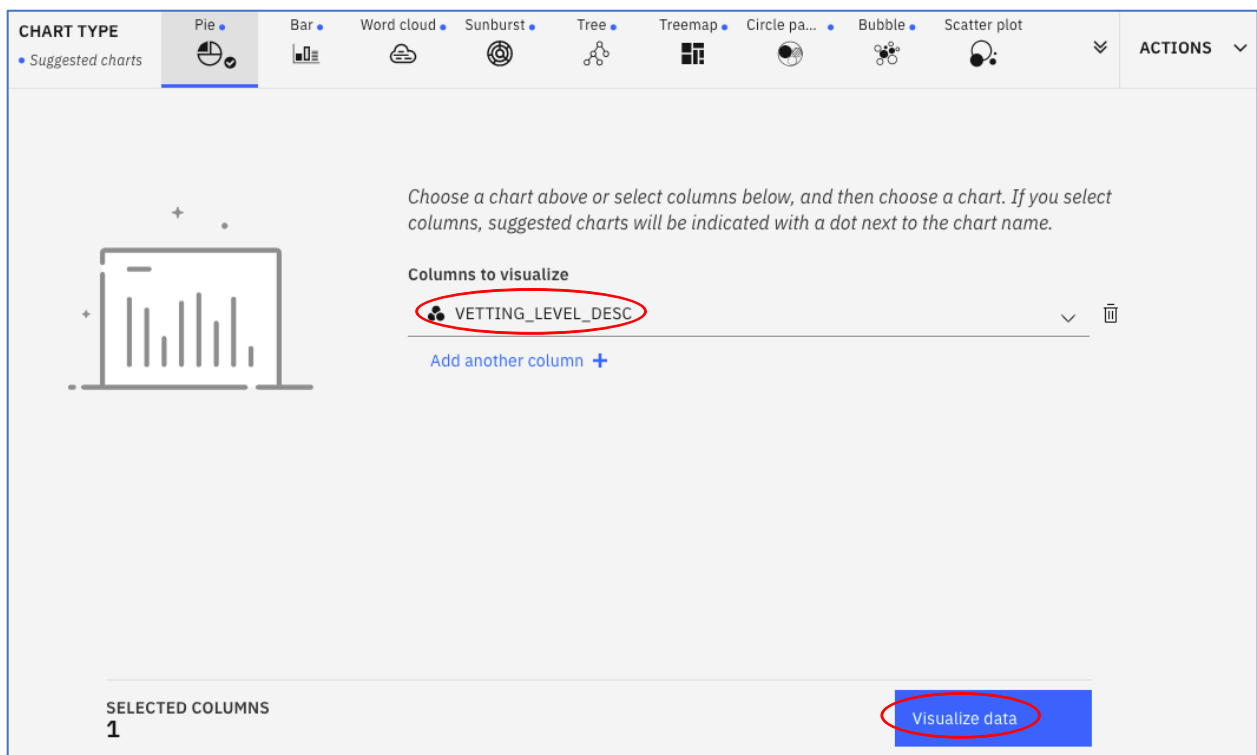
48. Save the Data Flow by clicking on the Save  icon.



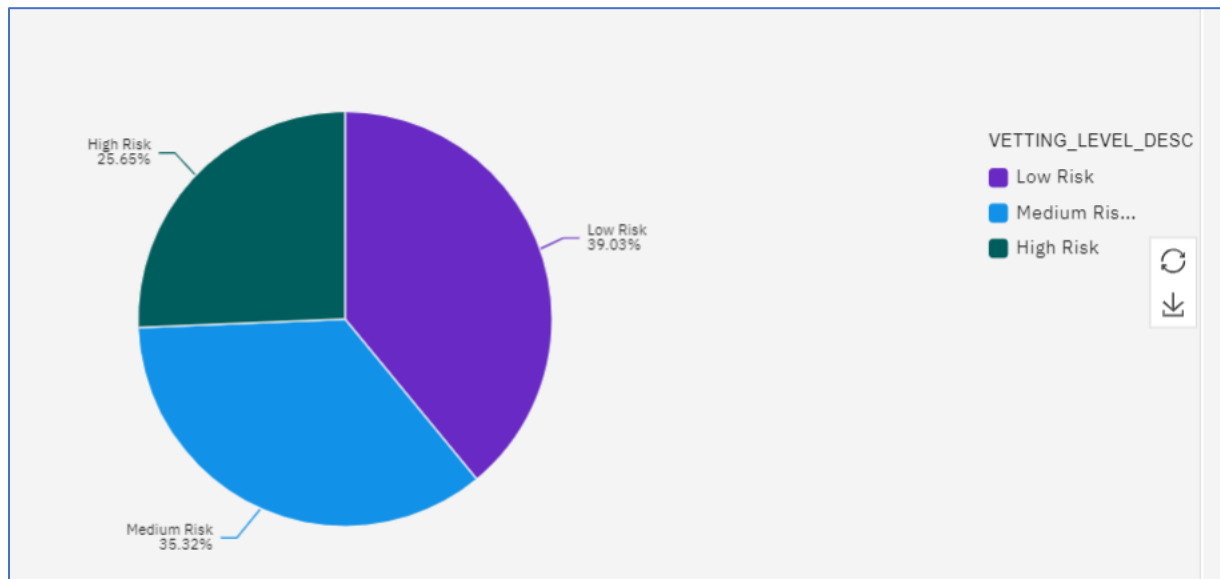
49. Click on the **Visualization** tab.



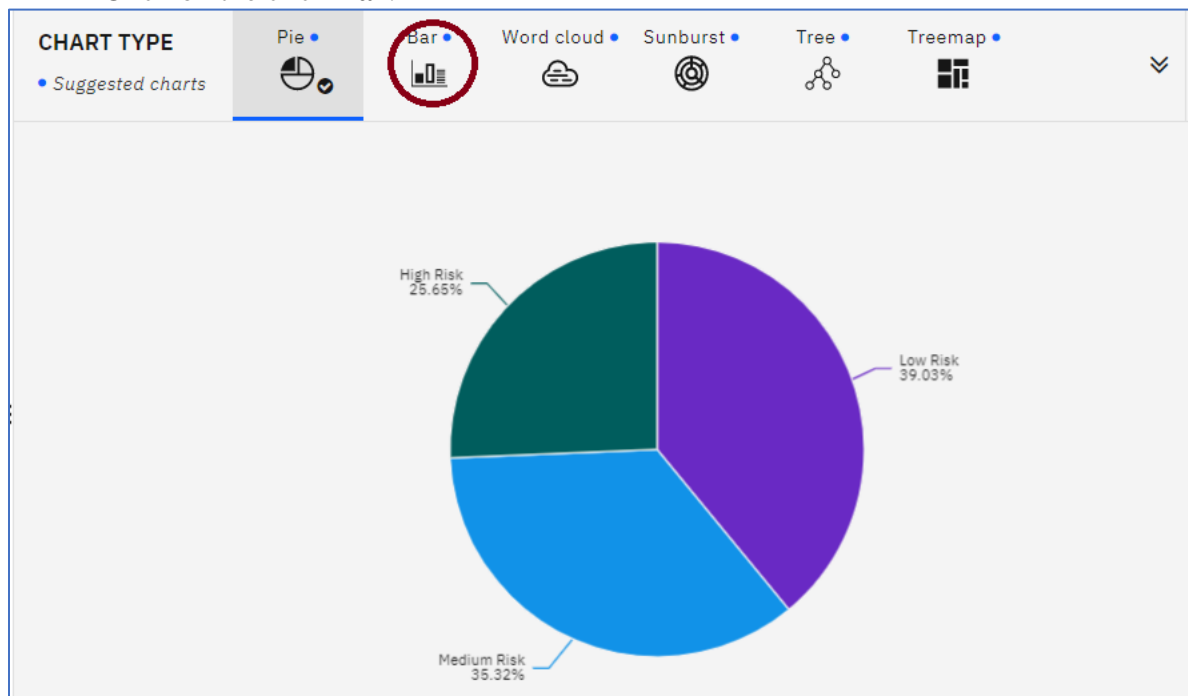
50. Click on **VETTING_LEVEL_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



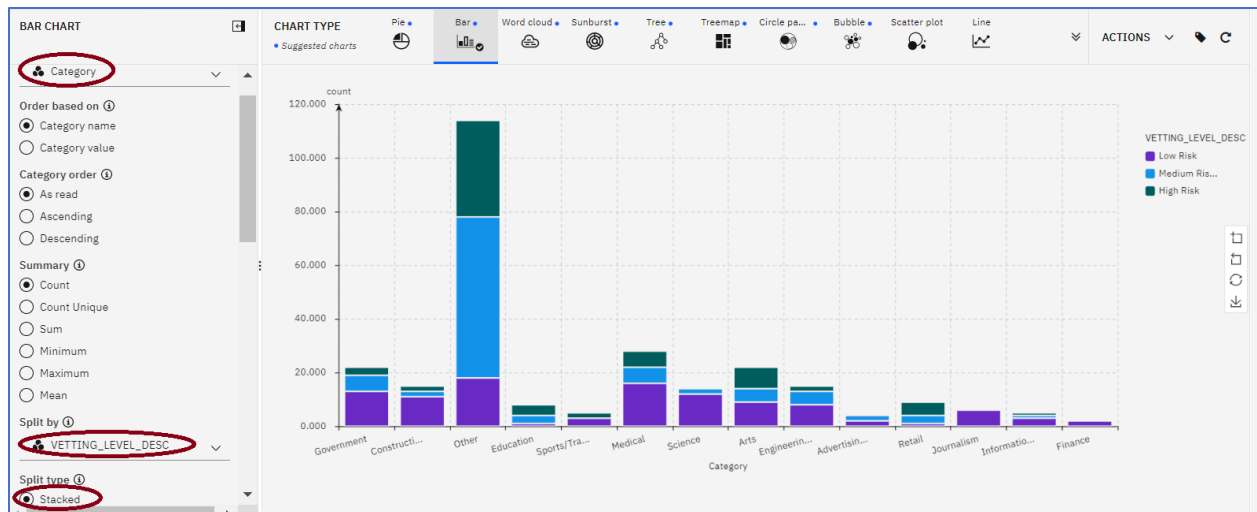
51. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



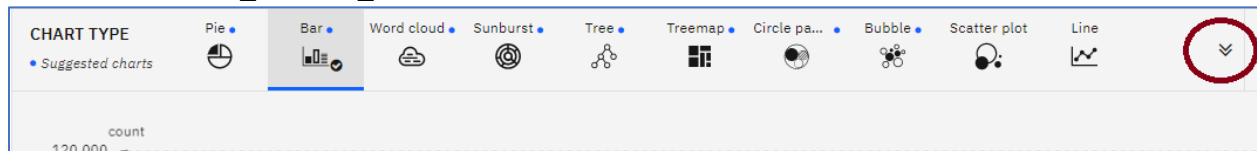
52. We can visualize the breakdown of travel records by job category and vetting level. Click on the click **Bar**.



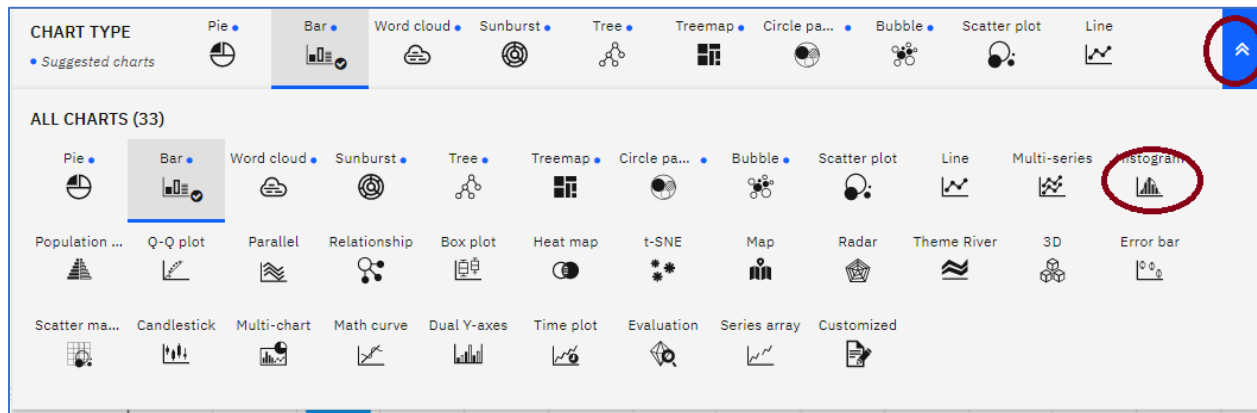
53. Click on **Category** for **Category**, click on VETTING_LEVEL_DESC for **Split by**, click on **Stacked** for **Split type**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



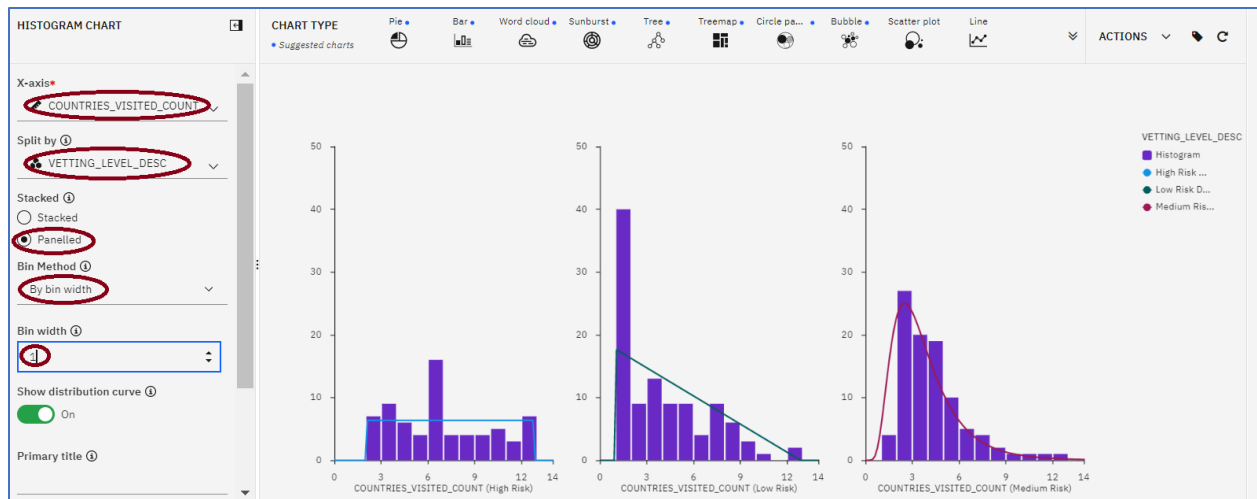
54. We can visualize a histogram of COUNTRIES_VISITED_COUNTS split by VETTING_LEVEL_DESC. Click on the  icon.



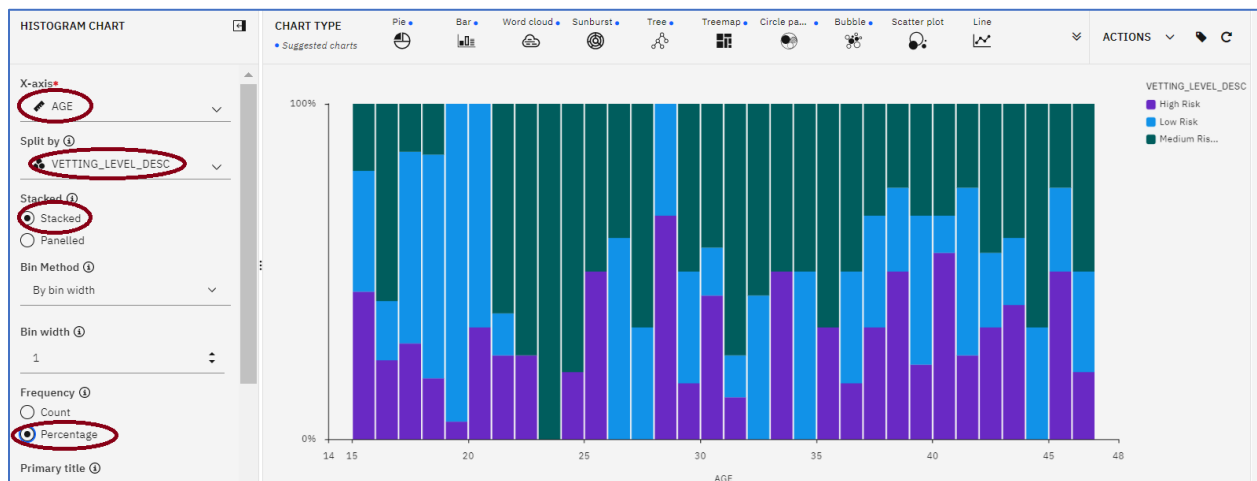
55. Click on **Histogram**



56. 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.





57. 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.

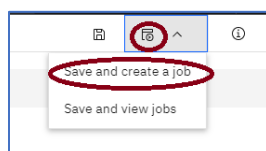


58. Please feel free to experiment with other visualizations.

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  .

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



60. Enter a **Name** for the job. 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 'Name' field is highlighted with a red circle and contains the text 'FHT Data Refinery'. The 'Description (optional)' field is empty. The 'Next' button at the bottom right is also highlighted with a red circle. The 'Cancel' button is at the bottom left.

Create a job

Define details

FHT Data Refinery

Associated asset
female human trafficking_flow (14 Steps)

Name
FHT Data Refinery

Description (optional)
What's the purpose of this job?

Cancel

Next

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

The screenshot shows the 'Create a job' dialog with the 'Configure' step selected. The 'Input' field is 'female_human_trafficking' and the 'Output' field is 'female_human_trafficking_shaped'. The 'Environment' dropdown is set to 'Default Data Refinery XS'. The 'Next' button at the bottom right is highlighted with a red circle. The 'Back' button is at the bottom left.

Create a job

Configure

Data assets

Input
female_human_trafficking CSV

Output
female_human_trafficking_shaped CSV

Environment
Default Data Refinery XS

Cancel

Back

Next

62. 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

63. Click **Next** on the Notify panel.

Create a job

Define details

FHT Data Refinery

Configure

Default Data Refinery XS

Schedule

Notify

Review and create

Notify

Want notifications for this job?

Turn on or off notifications associated with this job

Off

Cancel

Back

Next

64. 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

Associated asset

female_human_traffickl... (13 Steps)

Name

FHT Data Refinery

Description

[Add Description](#)

Configuration

Environment:

Default Data Refinery XS

Data assets

Input

female_human_trafficking CSV

→

Output

female_human_trafficking_... CSV

Schedule

Scheduled to run

No schedule created

Cancel

Back

Create

Create and run

65. 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

11 Steps

Data Source

female_human_trafficking

	VETTING_LEVEL	NAME	BIRTH_DATE	OCCUPATION	PASSPORT_CO...	CI
	String	String	Date	String	String	St
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

Edit

LOCATION

The job was successfully created. See job details.

66. Wait until the job run changes from **Running** to **Completed**. Refresh the browser if required.

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

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

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 X	FHT Data Refinery	Data Refinery Flow

67. 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

68. 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 female_human_trafficking_shaped	Data Asset	FCTO Labs	Jan 10, 2021, 01:31 PM
<input type="checkbox"/>	CSV Occupation	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
<input type="checkbox"/>	CSV Categories	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
<input type="checkbox"/>	CSV female_human_trafficking	Data Asset	FCTO Labs	Jan 10, 2021, 12:34 PM

69. 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 engli	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 Staffr	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 Silv	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	Jaccie 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

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

Projects **Watson Studio Labs** female_human_trafficking_shaped

Preview Profile Activities

Schema: 15 Columns
Preview: 269 rows

Last refresh: just now [Refine](#)

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

Projects / Watson Studio Labs

Overview Assets Environments **Jobs** Access control Settings

Jobs

▼ All jobs Find jobs Last updated: 8/22/21, 2:16 PM

Job	Asset type	Last modified
FHT Data Refinery	Data Refinery Flow	38 min ago Modified by Avril D.

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.