

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.

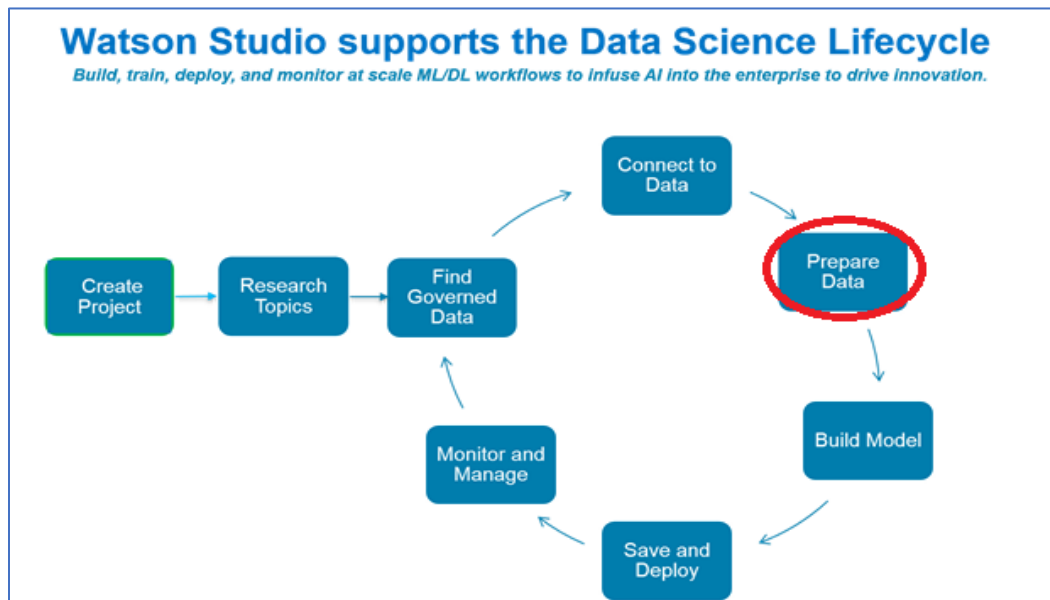


Figure 1- End to End Flow

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

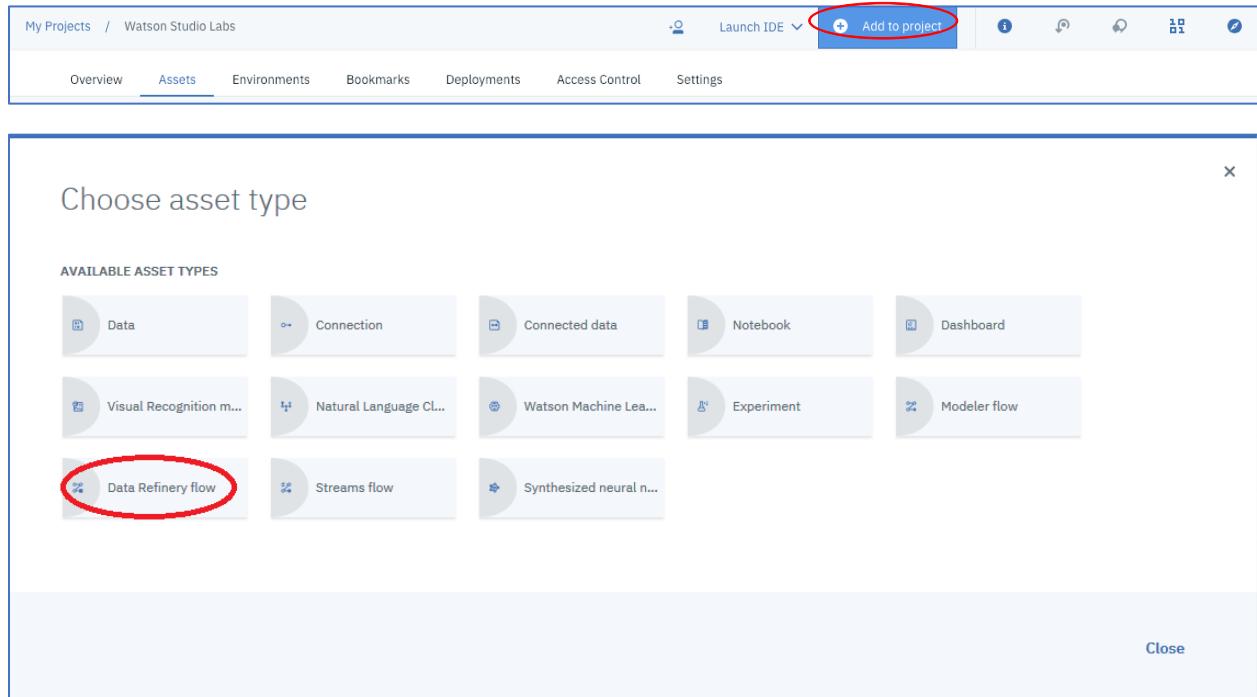
The data sets used for this lab consist of simulated travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING_LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL 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.

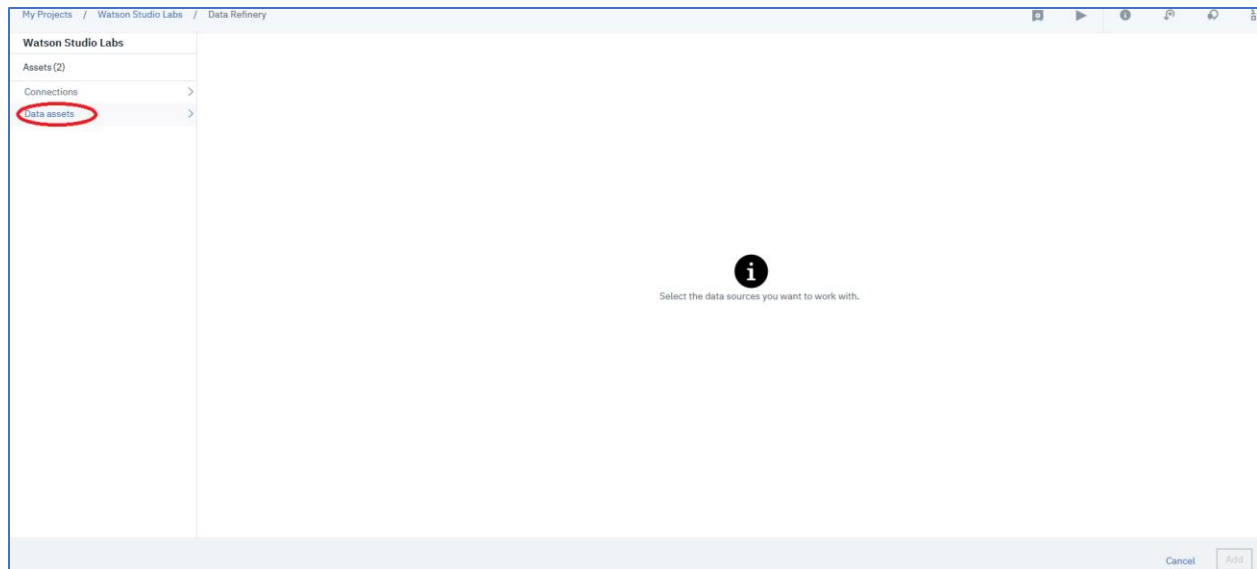
Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

Create a new Data Flow

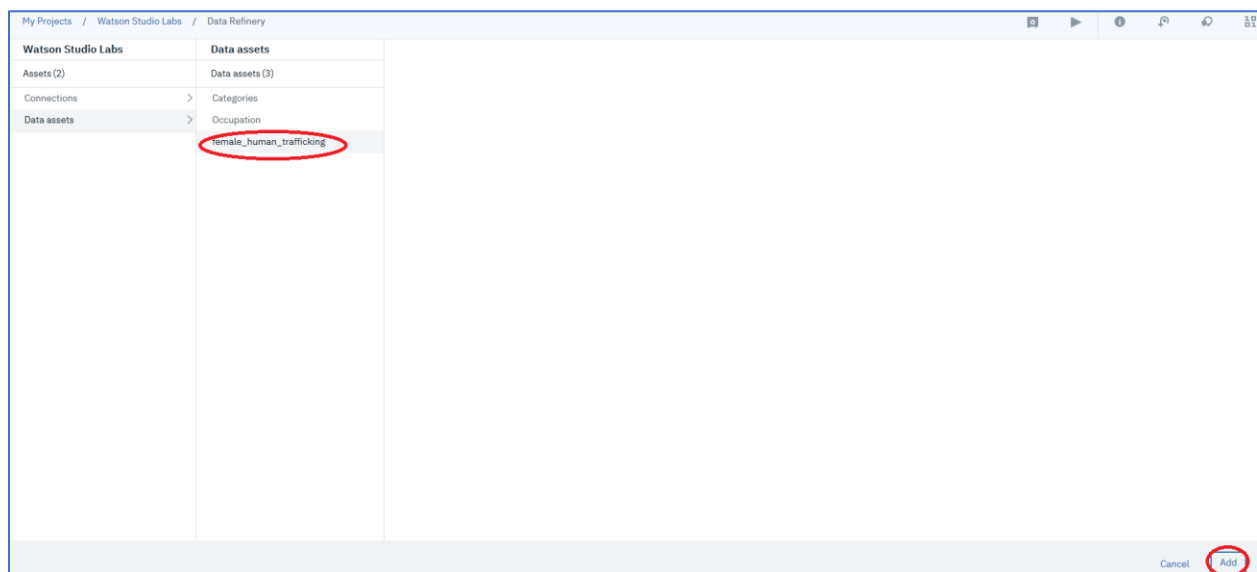
1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow**.



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.

My Projects / Watson Studio Labs / female_human_trafficking / Data Refinery


+ Operation *Code an operation to cleanse and shape your data*

Data Profile Visualizations


	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 Rebecca 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 (histocompatib
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.

VETTING_LEVEL	DESCRIPTION	NAME
Integer	String	String
100	Remove	Karey Simon
100	Remove duplicates	Holly Chaney
100	Remove empty rows	Abby Harrell
30	Sort ascending	Pamela Dixon
100	Sort descending	Kimbe Vicki Bradford
30	Substitute	Besy Rebecc Gibson
100		Cristina Oconnell
100	CONVERT COLUMN... >	Boolean
100		Decimal
100	View All	✓ Integer
30	NA	String
30	NA	
30	NA	
100	NA	Sassa Christy Melendez
20	NA	Missy Christina Garcia

- 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...	PAS...
Integer	String	String
775799149		GH
433319392		GH
928594311		GH
729531890		GH
572668476		GH
471592343		GH
810764463		GH
961032498		
438555740		
960558734		
151109290	Ghana	
505934606	Ghana	
402297848	Ghana	
186058958	Ghana	GH
751002860	Ghana	GH

Remove
Remove duplicates
Remove empty rows
Sort ascending
Sort descending
Substitute
CONVERT COLUMN... >
View All

Boolean
Decimal
✓ Integer
String

1 STEPS
Data Source
Convert colu
Converted VE
to String

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

+ Operation
Code an operation to cleanse and shape your data

Data Profile Visualizations

Steps

- 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_hu...

Convert column type

Converted VETTING_LEVEL from Integer to String

Convert column type JUST ADDED

Converted PASSPORT_NUMBER from Integer to String

5. Click on **Profile**.

My Projects / Watson Studio Labs / female_human_trafficking / Data Refinery

+ Operation

Code an operation to cleanse and shape your data

Data

Profile

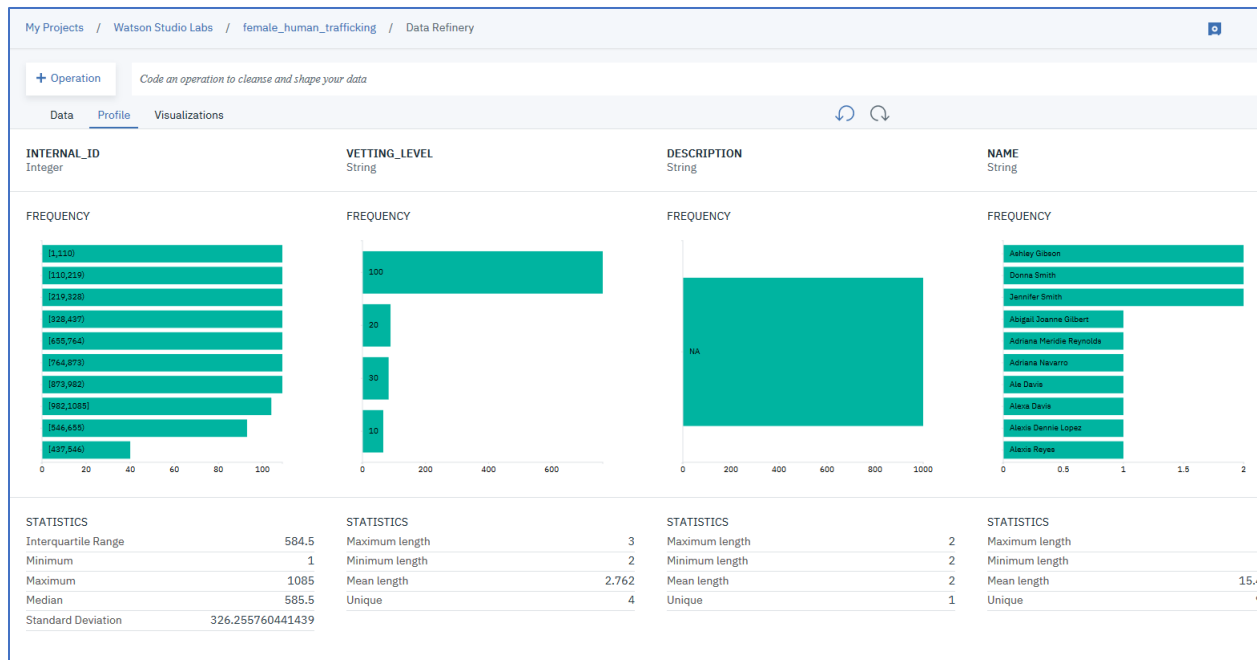
Visualizations

↺

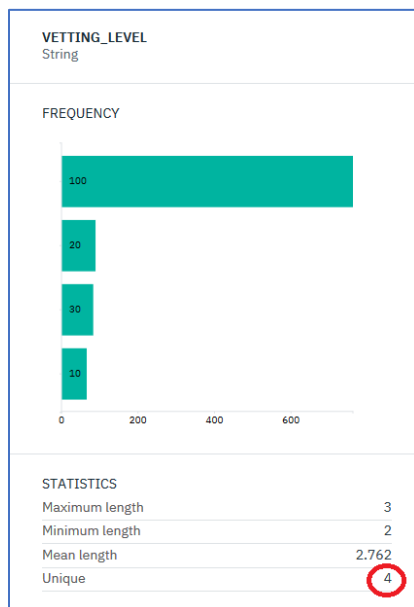
↻

	INTERNAL_ID Integer	VETTING_LEVEL String	DESCRIPTION String	NAME String	GENDER String	BIRTH Date
1	785	20	NA	Sara Nunez	F	1978-
2	786	100	NA	Sara Schultz	F	1983-
3	787	100	NA	Mackenzie Emi Cooley	F	2000-
4	788	100	NA	Rea Ray	F	1996-
5	789	100	NA	Becky Anthony	F	1996-
6	790	100	NA	Emily McBride	F	1996-
7	791	100	NA	Allison Stanley	F	1994-
8	792	100	NA	Elle King	F	1971-
9	793	100	NA	Emily Carol Ellis	F	1971-
10	794	100	NA	Denise Taylor	F	1986-
11	795	100	NA	Taylor Ana Martin	F	1978-
12	796	100	NA	Jasmea Adriana Ferguson	F	1977-
13	797	100	NA	Erin Jones	F	1995-
14	798	100	NA	Joy Tina Smith	F	1986-
15	799	100	NA	Terri Libby Lara	F	1993-

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.

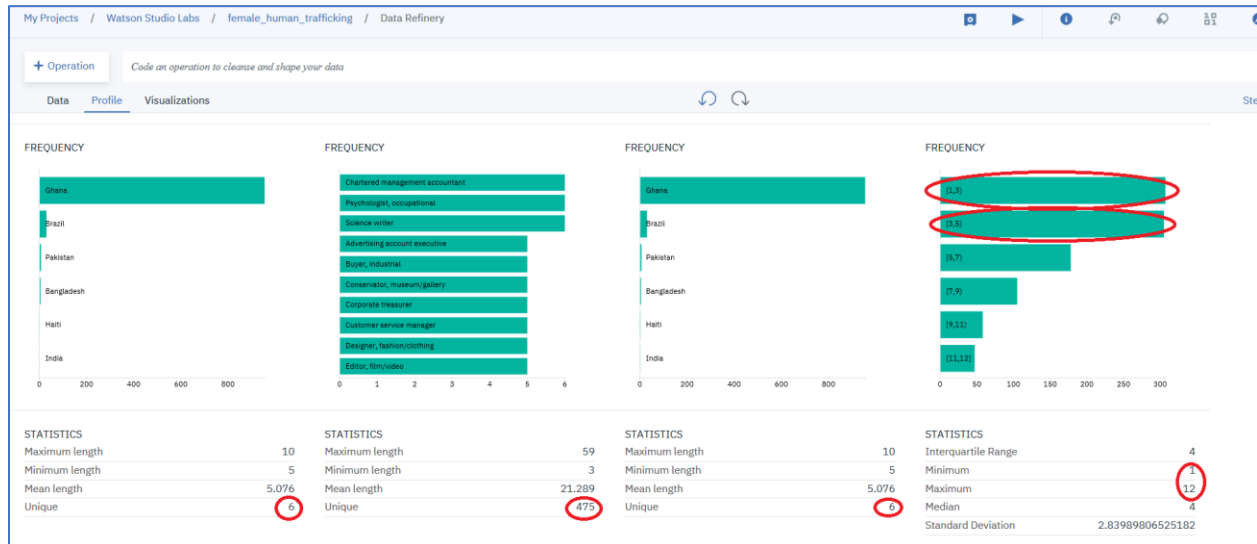


- The statistics for the VETTING_LEVEL column show 4 unique values, 10, 20, 30, and 100. These are coded values that correspond to risk of trafficking, 10-High Risk, 20-Medium Risk, 30-Low Risk, and 100- has not been vetted yet. As the graph shows below, most of the data records have not been vetted 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.

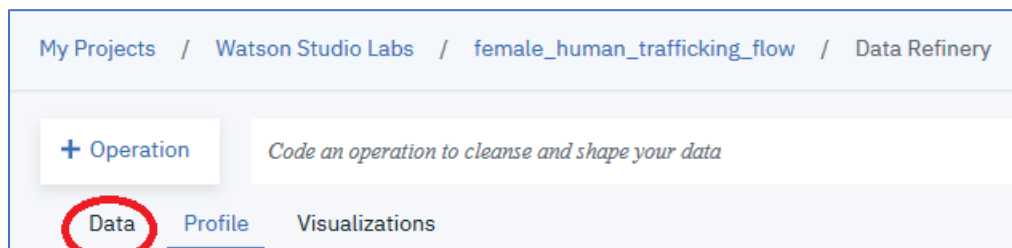


- Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH_COUNTRY, and PASSPORT_COUNTRY shows

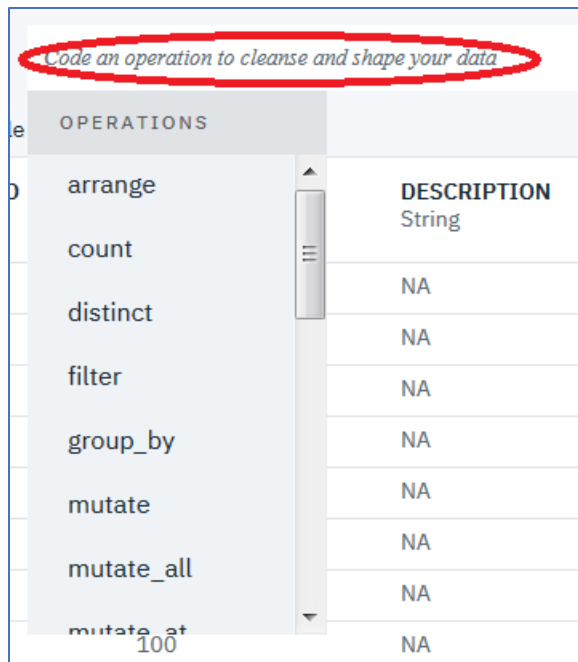
only 6 unique countries. The COUNTRIES_VISITED_COUNT shows that passengers have visited between 1 and 12 countries, with passengers visiting between 1 and 3 countries and between 3 and 5 countries the most prevalent. Note, the results may be slightly different on your screen.



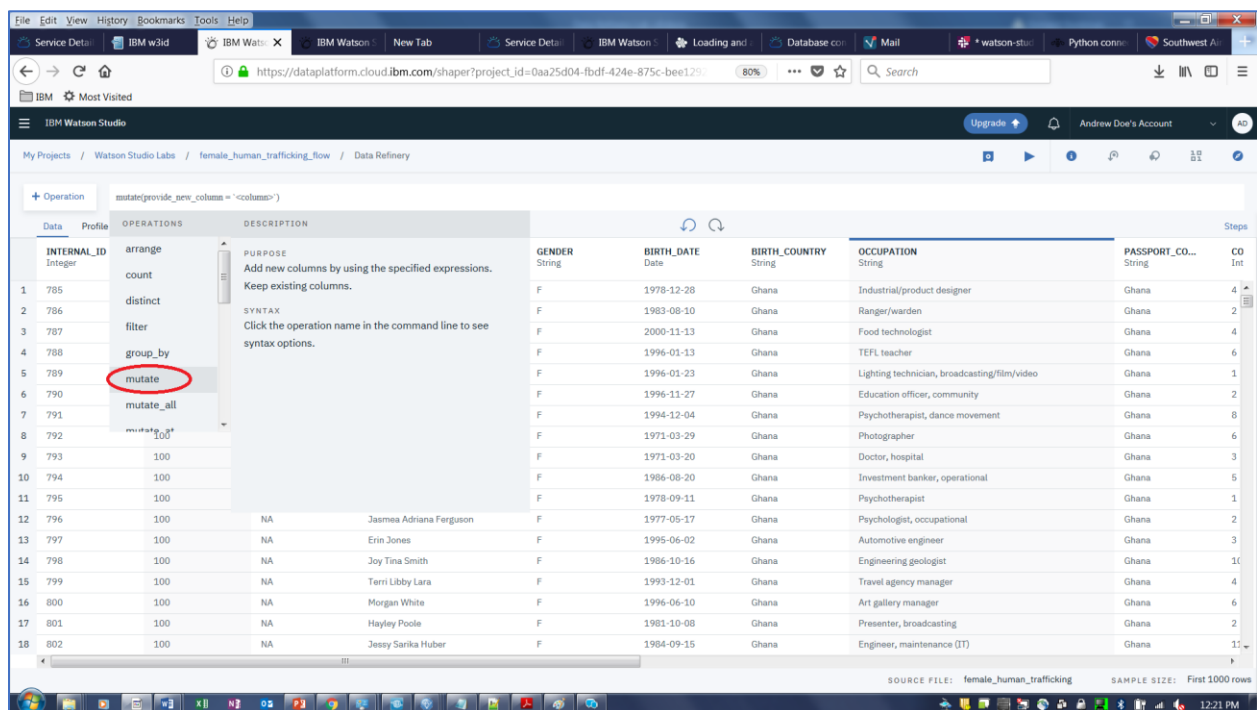
9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.



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"))))
```

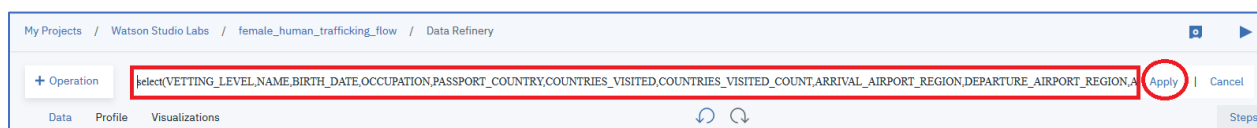


13. If you scroll to the right you should see the new column VETTING_LEVEL_DESC with values “Low Risk”, “Medium Risk”, “High Risk”, and “Unvetted”.

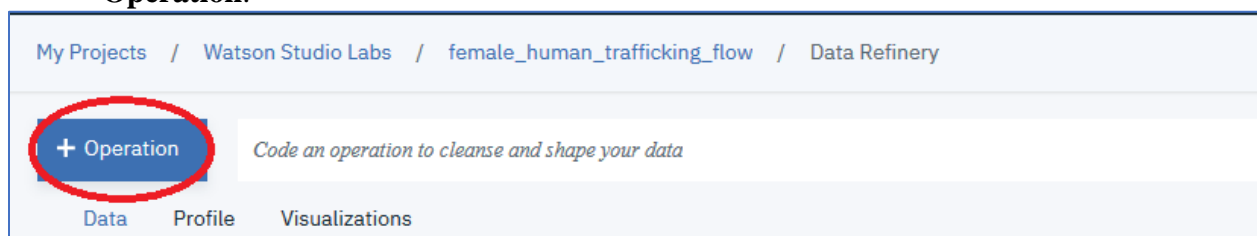
VETTING_LEVE... String
Unvetted
Low Risk
High Risk
Low Risk
Unvetted
Unvetted
Unvetted
Unvetted
Medium Risk
Low Risk

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

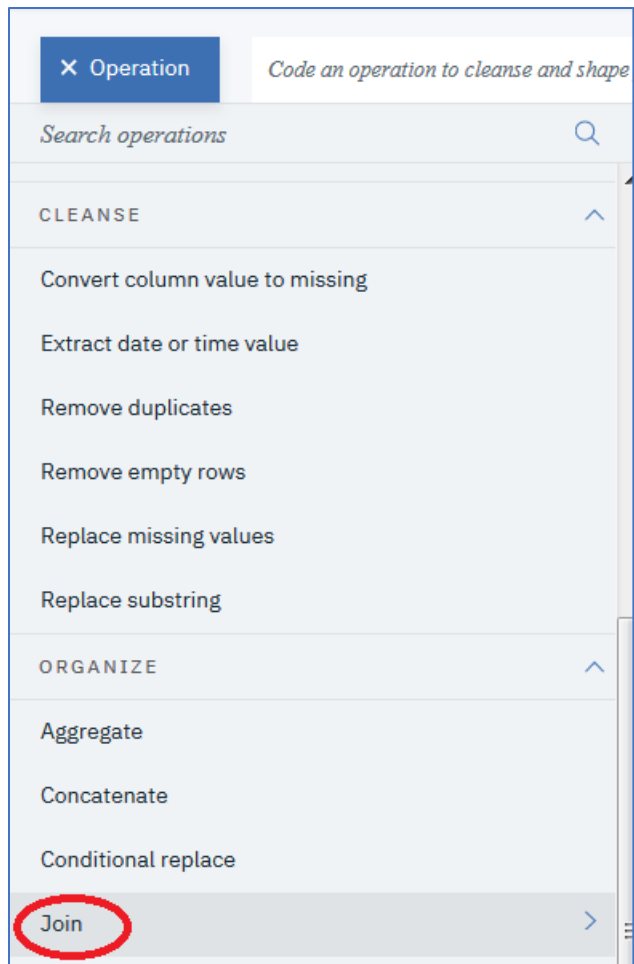
```
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,C
OUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DE
PARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)
```



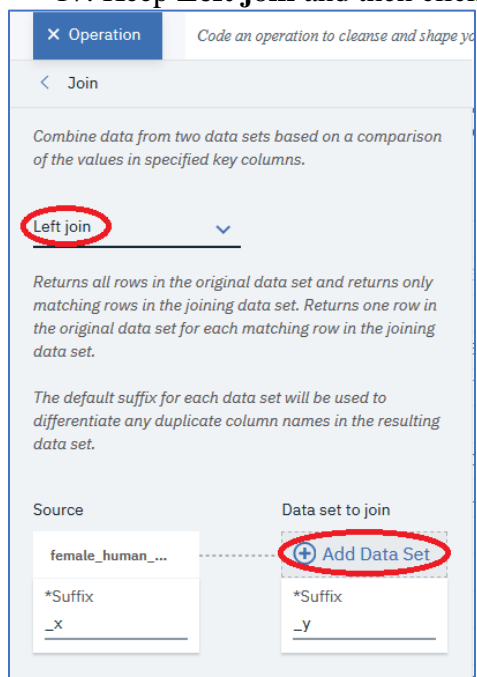
15. Let’s now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the categories dataset. Click on + **Operation**.



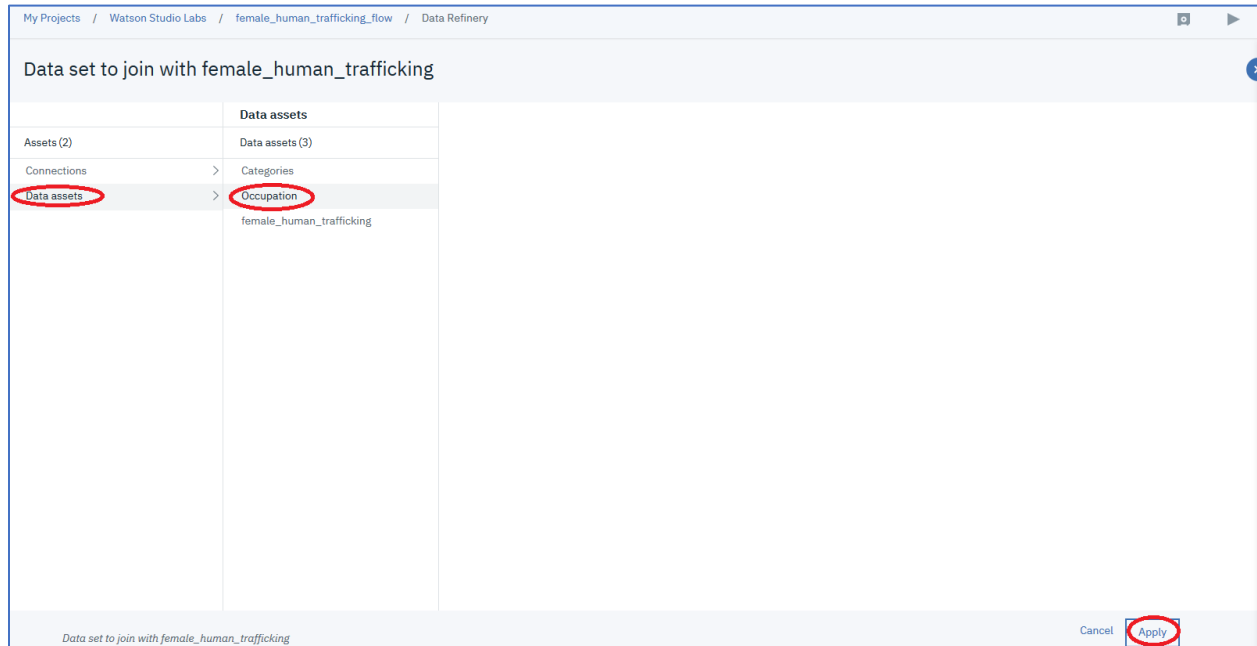
16. Scroll down and click on **Join**.



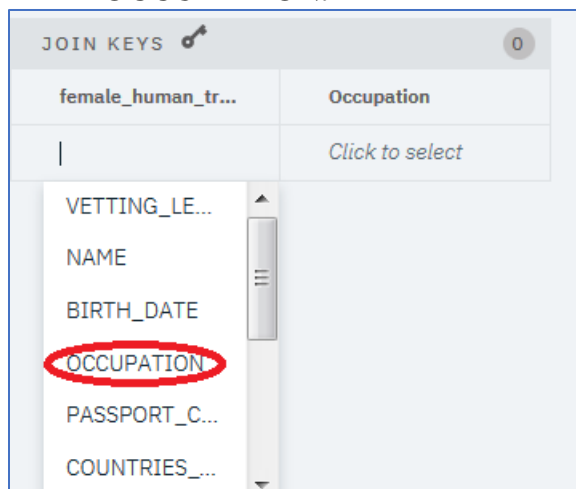
17. Keep **Left join** and then click on **Add Data Set**



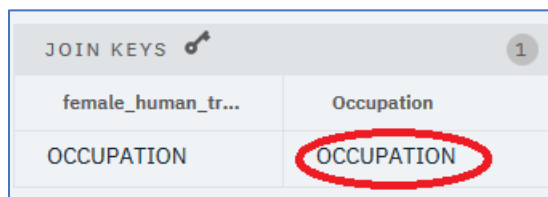
18. Click on **Data Assets**, click on **Occupation**, and then click **Apply**.




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**.

JOIN KEYS  1

female_human_tr...	Occupation
OCCUPATION	OCCUPATION

 Add Join Key

Cancel Next

22. Click **Apply**.

Select the columns in the resulting data set

- ☒ Clear all selections
- ☒ VETTING_LEVEL
- ☒ NAME
- ☒ BIRTH_DATE
- ☒ OCCUPATION
- ☒ PASSPORT_COUNTRY
- ☒ COUNTRIES_VISITED
- ☒ COUNTRIES_VISITED_COUNT
- ☒ ARRIVAL_AIRPORT_REGION
- ☒ DEPARTURE_AIRPORT_REGION
- ☒ AGE
- ☒ VETTING_LEVEL_DESC
- ☒ Code

Back 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.

The screenshot displays a data flow interface. On the left, a table with two columns, 'Code' (String) and 'Category' (String), is shown. The 'Code' column contains values: 15, 1, 7, 13, 8, 13, 8, 8, 6, 14, 6, 6, 2, 2, 1, 8, 4, 2. The 'Category' column contains values: Other, Sports/Travel, Science, Education, Arts, Education, Arts, Arts, Medical, Finance, Medical, Medical, Engineering, Engineering, Sports/Travel, Arts, Journalism, Engineering. On the right, a sidebar titled '6 STEPS' lists the flow steps. The first step is 'Custom code' with a mutate operation. The second step is 'Custom code' with a select operation. The third step is 'Join' (left-joined data from Occupation based on columns OCCUPATION, OCCUPATION). The fourth step is 'Join' (left-joined data from Categories based on columns Code, Code), which is highlighted with a red box and labeled 'JUST ADDED'.

Code String	Category String
15	Other
1	Sports/Travel
7	Science
13	Education
8	Arts
13	Education
8	Arts
8	Arts
6	Medical
14	Finance
6	Medical
6	Medical
2	Engineering
2	Engineering
1	Sports/Travel
8	Arts
4	Journalism
2	Engineering

6 STEPS

Custom code

```
mutate(VETTING_LEVEL_DESC =
  ifelse(VETTING_LEVEL=="10","High
  Risk",ifelse(VETTING_LEVEL=="20","M
  edium
  Risk",ifelse(VETTING_LEVEL=="30","L
  ow Risk","Unvetted"))))
```

Custom code

```
select(VETTING_LEVEL,NAME,BIRTH
_DATE,OCCUPATION,PASSPORT_COU
NTRY,COUNTRIES_VISITED,COUNTRI
ES_VISITED_COUNT,ARRIVAL_AIRPO
RT_REGION,DEPARTURE_AIRPORT_R
EGION,AGE,VETTING_LEVEL_DESC)
```

Join

left-joined data from Occupation based on columns OCCUPATION,OCCUPATION

Join JUST ADDED

left-joined data from Categories based on columns Code,Code

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**.

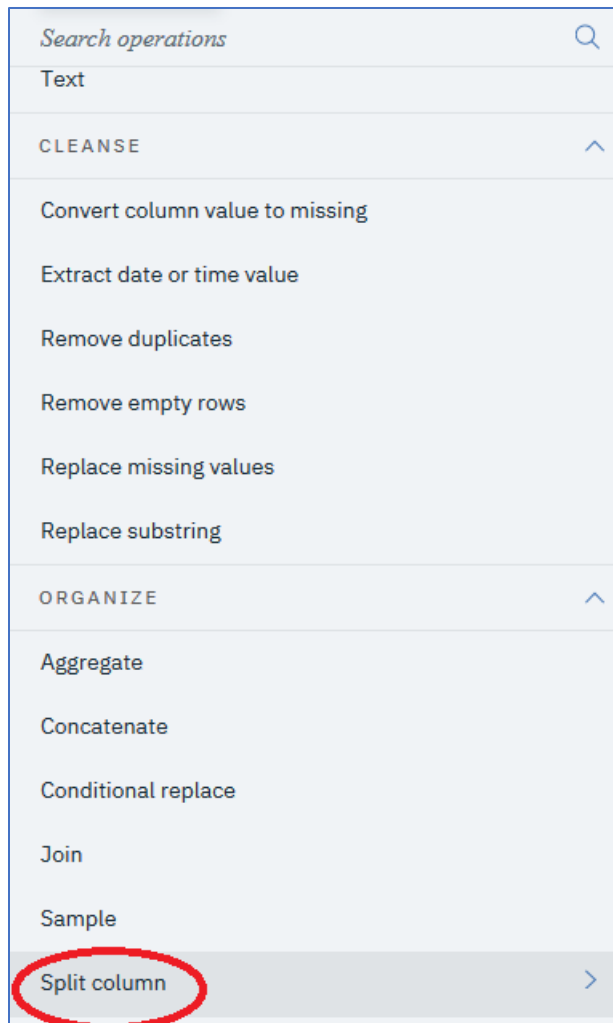
The screenshot shows the Data Refinery interface. At the top, the breadcrumb navigation reads: 'My Projects / Watson Studio Labs / female_human_trafficking_flow / Data Refinery'. Below this, a blue button with a white plus sign and the text '+ Operation' is highlighted with a red circle. To the right of the button, the text 'Code an operation to cleanse and shape your data' is displayed. At the bottom, there are three tabs: 'Data', 'Profile', and 'Visualizations', with 'Data' being the active tab.

My Projects / Watson Studio Labs / female_human_trafficking_flow / Data Refinery

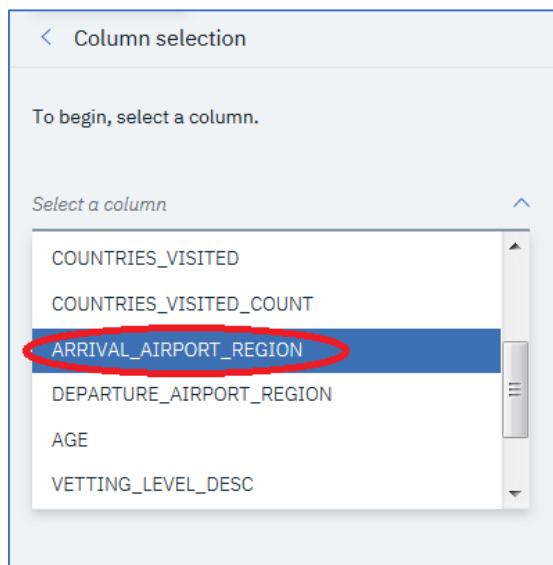
+ Operation Code an operation to cleanse and shape your data

Data Profile Visualizations

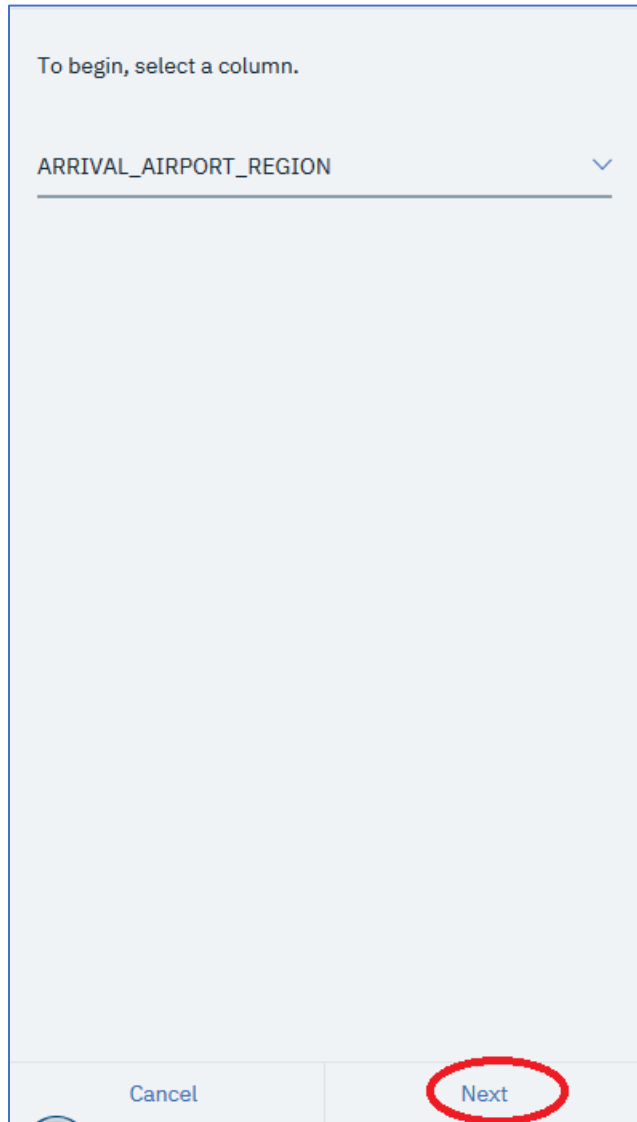
25. Click on **Split column**



26. Click on **ARRIVAL_AIRPORT_REGION**.



27. Click on **Next**.



The screenshot shows a mobile application interface with a light blue background. At the top, the text "To begin, select a column." is displayed. Below this, a dropdown menu is open, showing the selected item "ARRIVAL_AIRPORT_REGION" with a downward arrow icon to its right. The dropdown menu is empty, indicating no other options are visible. At the bottom of the screen, there are two buttons: "Cancel" on the left and "Next" on the right. The "Next" button is circled in red, indicating it is the target for the next action.

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**.

Selected column: ARRIVAL_AIRPORT_REGION

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

DEFAULT **TEXT** PATTERN POSITION

Hyphen (-) ▼

Names of new columns*

AIRPORT_COUNTRY, ARRIVAL_AIRPORT_STATE

☒ 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

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**.

ARRIVAL_AIRP...	ARRIVAL_AIRP...
String	String
	Remove
US	
US	Remove duplicates
US	Remove empty rows
US	Sort ascending
US	Sort descending
US	Substitute
US	CONVERT COLUMN... >
US	TEXT >
US	View All
US	CA
US	NC
US	AZ
US	NY
US	MS
US	GA
US	TX

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.

Change Column Selection

Selected column: COUNTRIES_VISITED

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

DEFAULT TEXT PATTERN POSITION

Comma (,)

Names of new columns*: 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

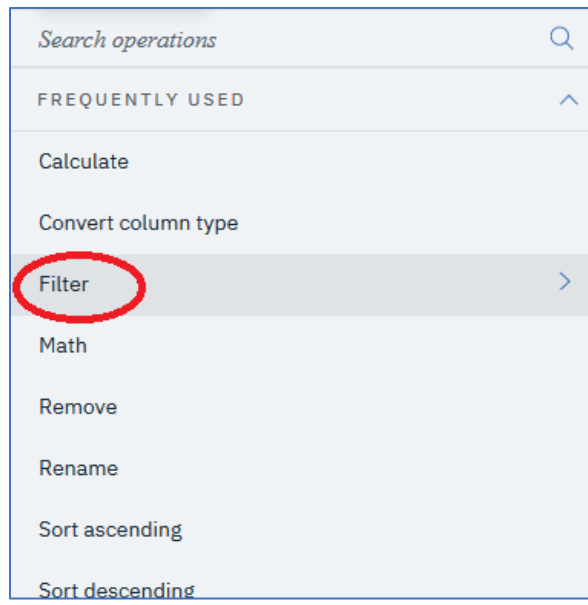
32. The results are shown below. Note there are now 9 steps in the Data Flow. (Only 7 if you skipped steps 1-4 above)

COUNTRIES_VISITED String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VI... Integer	ARRIVAL_AIRP... String	DEPARTURE_AI String	9 STEPS
HU,BY,NO,OM	HU	BY	NO	4	CA	HU-PE	left-joined data from Occupation based on columns OCCUPATION,OCCUPATION
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,JO,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**.

+ Operation		Code an operation to cleanse and shape	
Data	Profile	Visualizations	
VETTING_LEVEL String			
1	20		
2	100		
3	100		
4	100		
5	100		
6	100		
7	100		
8	100		
9	100		
10	100		
11	100		
12	100		

34. Click on **Filter**.



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

< Filter

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

CONDITIONS (1)

CONDITION 1

Column	Operator
VETTING_LEVEL	Does not contain

Choose to specify text or a pattern


☒ Text ☐ Pattern

Value


100

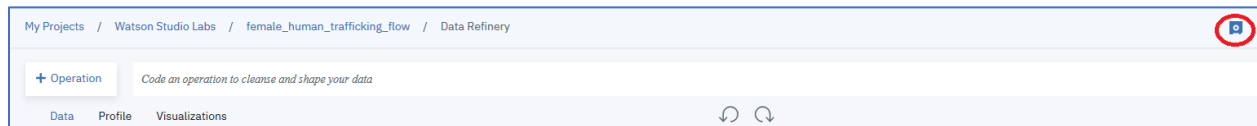
+ Add condition

Cancel Apply

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

Code String	Category String
15	Remove
15	Remove duplicates
15	Remove empty rows
15	Sort ascending
15	Sort descending
15	Substitute
15	CONVERT COLUMN... >
15	TEXT >
15	View All
5	Government

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



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

My Projects / Watson Studio Labs / female_human_trafficking_flow / Data Refinery						
+ Operation Code an operation to cleanse and shape your data						
Data Profile Visualizations						
	VETTING_LEVEL String	NAME String	BIRTH_DATE Date	OCCUPATION String	PASSPORT_CO... String	COUNTRIES_VISITED String
1	20	Sara Nunez	1978-12-28	Industrial/product designer	Ghana	HU,BY,NO,OM
2	10	Jenni Warner	1981-08-17	Midwife	Ghana	HU,RS,FL,CL,TR,UZ
3	10	Raven Jones	1985-07-19	Personal assistant	Ghana	TN,PH,SA,RIJ,CN,EG
4	20	Sara Kim	1990-06-22	Designer, industrial/product	Ghana	TW,QA
5	10	Rosy Hunter	1974-11-15	Freight forwarder	Ghana	EG,RS,UZ,I,V,BRAZ
6	10	Diana Russell	1991-11-20	Pilot, airline	Ghana	JM,RIJ,KH,CO,BZ,MA,BS,GB,DO
7	20	Anna Conway	1987-06-29	Ceramics designer	Ghana	ZA,RIJ
8	10	Chelsea Ferguson	1970-10-27	Licensed conveyancer	Ghana	UA,IJ,IN,RS,QA,FI,CO
9	10	Jennifer Smith	1989-03-13	Youth worker	Ghana	BE,TW,RO,KW,RO,BZ,SN,CL,IJZ,FR,TW,FI
10	30	Anna Rose	1990-12-25	Personal assistant	Ghana	OM
11	20	Gina Franco	1986-11-28	Company secretary	Ghana	OM,PK
12	30	Brandi Dennee Taylor	1977-11-06	Armed forces technical officer	Brazil	IR,BS,CN,AE,LT
13	20	Bette Morris	1998-03-29	Industrial buyer	Ghana	PK,CZ,CA,TH,UNZ,RS,UZ,NZ
14	30	Samantha Moore	1984-04-28	Operations geologist	Brazil	PT,SD,KE,OM,GR,IN
15	20	Rhonda Tammy Prince	2000-06-06	Historic buildings inspector/conservation officer	Ghana	AE,IJ,BS,NL
16	30	Dawnie Stephanie Martin	1998-11-02	General practice doctor	Ghana	QA,GE,SD,UA,GE,CH,SG
17	10	Kathy Villanueva	1999-01-22	Call centre manager	Ghana	BY,PK,KZ,OM,ME,BZ,TH,AZ,GE,SG,UZ,IT
18	20	Krista Chelsea Bell	1973-03-29	Community development worker	Ghana	PH,IT,ET,OM,GB,PT

10 STEPS

left-joined data from Occupation based on columns OCCUPATION,OCCUPATION

Join

left-joined data from Categories based on columns Code,Code

Split column

Split ARRIVAL_AIRPORT_REGION by text - into ARRIVAL_AIRPORT_COUNTRY,ARRIVAL_AIRPORT_STATE

Remove

Removed ARRIVAL_AIRPORT_COUNTRY

Split column

Split COUNTRIES_VISITED by text , into COUNTRY1,COUNTRY2,COUNTRY3

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

CHART TYPES

• Suggested charts

Pie • Bar • Word cloud • Scatter plot • Line • Multi-series • Histogram • Population ... • Q-Q plot • Parallel • Relationship • Box plot • Treemap

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 Clear

VETTING_LEVEL_DESC

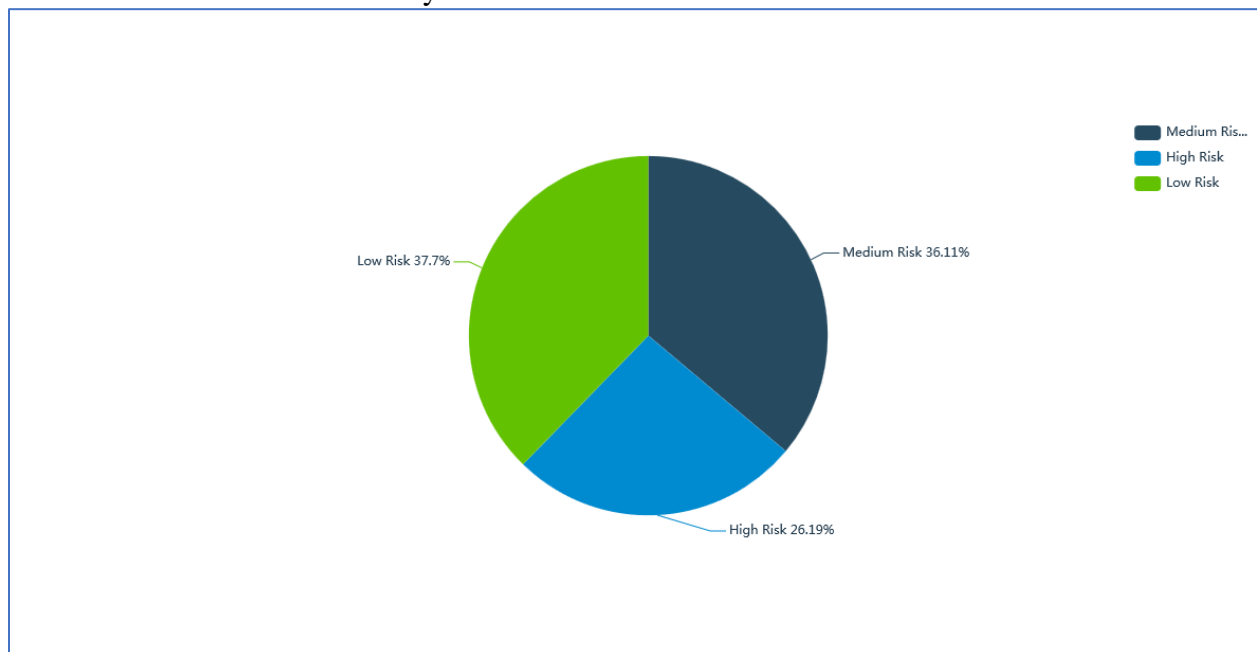
+ Add column

SELECTED COLUMNS

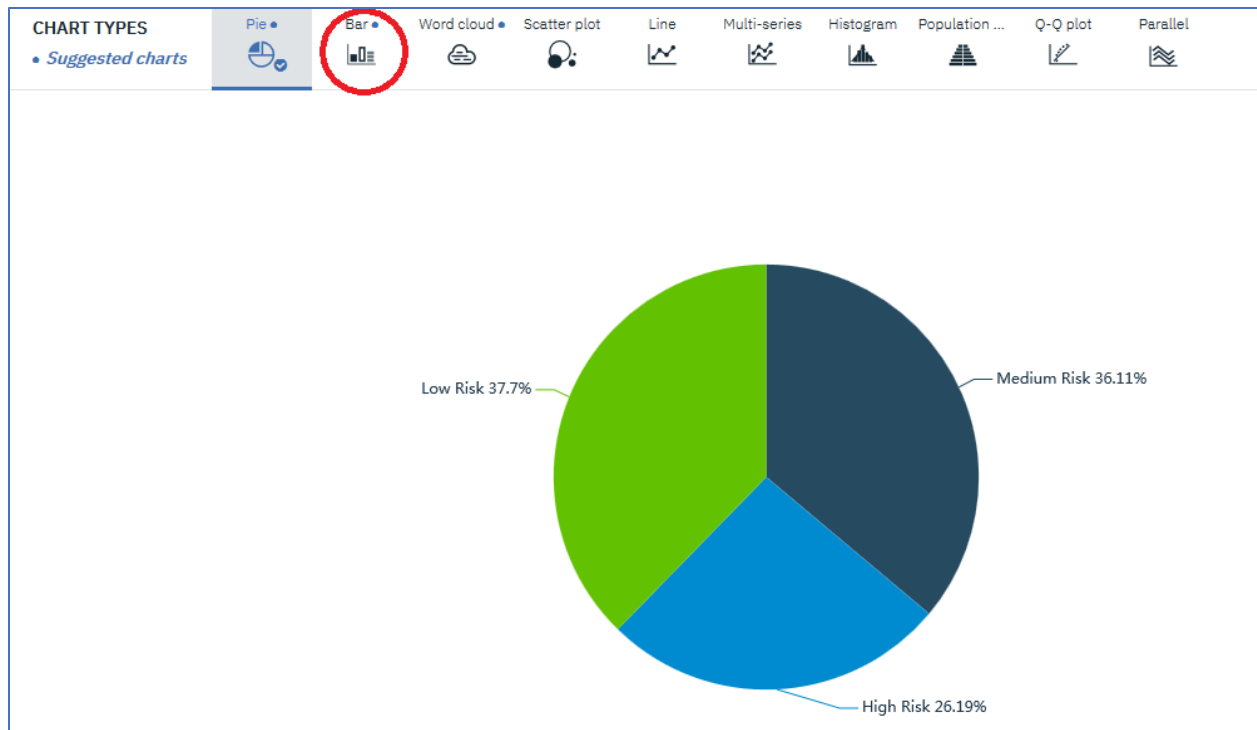
1

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**

Switch charts?

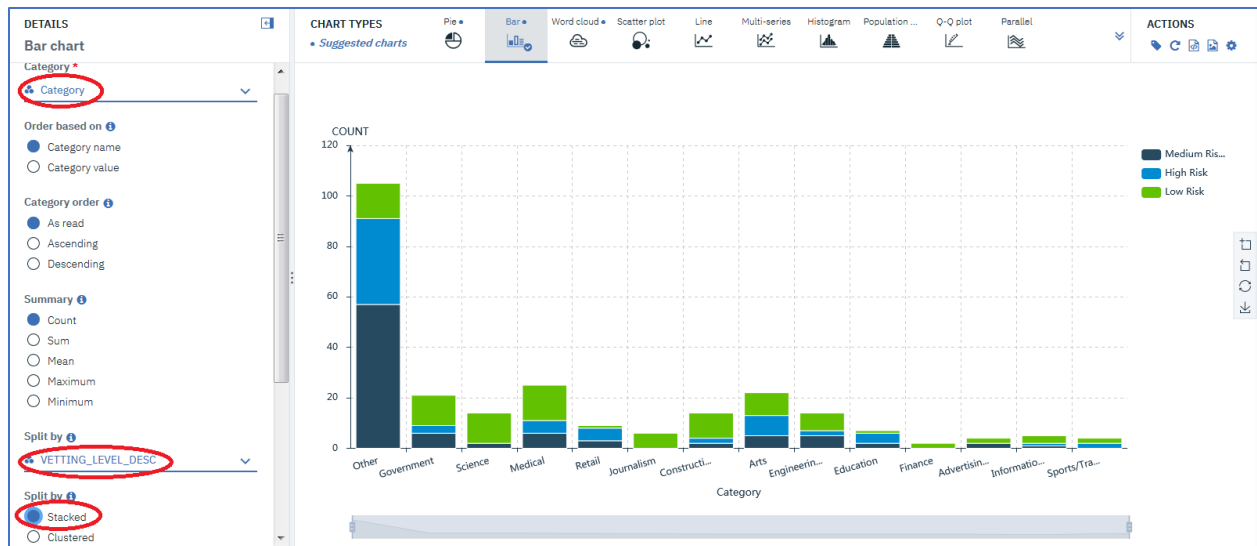
You might lose this chart's details or you might have to provide more details to view another chart.

☐ Don't show this again

Cancel Continue

Continue

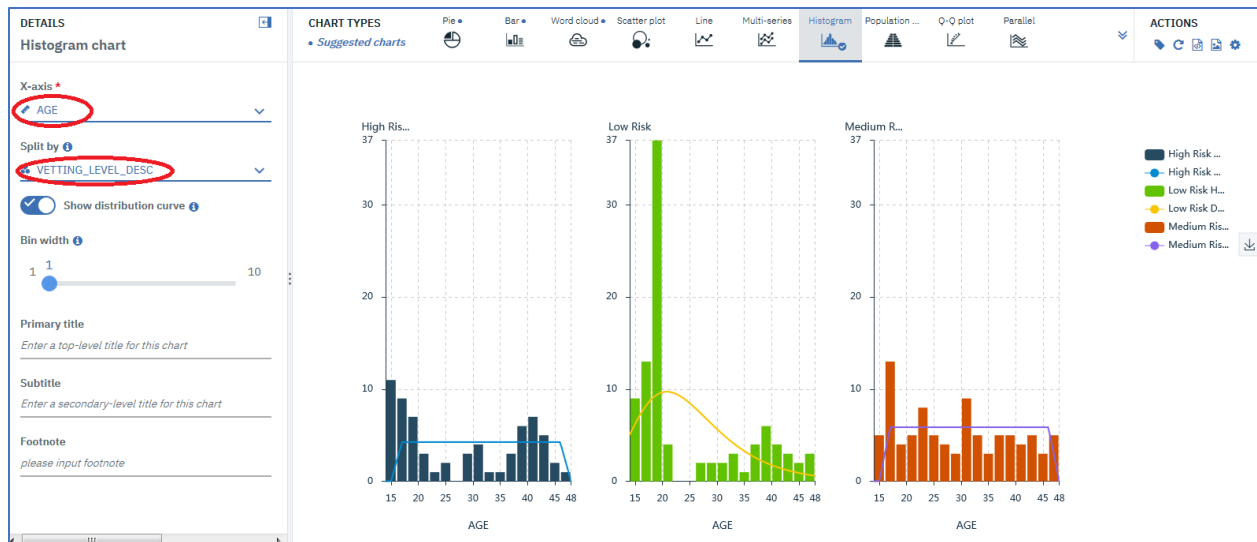
43. Click on **Category** for **Category**, click on VETTING_LEVEL_DESC for **Split by**, click on **Stacked** for **Split by**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



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.

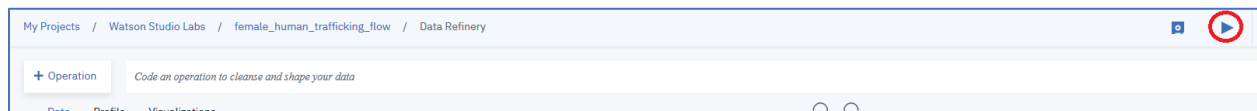



46. Please feel free to experiment with other visualizations.

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).

My Projects / Watson Studio Labs / female_human_trafficking_flow / Data Refinery

DATA REFINERY FLOW DETAILS	DATA REFINERY FLOW OUTPUT
<div><div>LOCATION</div><div>Watson Studio Labs</div></div> <div><div>DATA REFINERY FLOW NAME</div><div>female_human_trafficking...</div><div>Enter a description of the Data Refinery flow</div></div> <div><div>STEPS</div><div>11</div></div> <div><div>Beta Limitation ⓘ</div><div>PROJECT SPARK ENVIRONMENT ⓘ</div><div>None - Use Data Refinery Default ▾</div><div>This runtime consumes 6 capacity units per hour.</div></div> <div><div>Schedule</div><div>Add Schedule</div></div> <div>* Required Fields</div>	<div><div>LOCATION</div><div>Watson Studio Labs/Data assets</div></div> <div><div>DATA SET NAME</div><div>female_human_traffickin...</div><div>Enter a description of the resulting data set.</div></div> <div><div><input checked="" type="checkbox"/> If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.</div></div> <div>File format: CSV</div>

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.

My Projects / Watson Studio Labs / female_human_trafficking_flow / Data Refinery

DATA REFINERY FLOW DETAILS

LOCATION

Watson Studio Labs

DATA REFINERY FLOW NAME

female_human_trafficking...

Enter a description of the Data Refinery flow

STEPS

11

Beta Limitation ⓘ

PROJECT SPARK ENVIRONMENT ⓘ

None - Use Data Refinery Default ▾

This runtime consumes 6 capacity units per hour.

Schedule

Add Schedule

* Required Fields

DATA REFINERY FLOW OUTPUT

Edit output

✓

✕

LOCATION *

Watson Studio Labs/Data assets

Change Location

DATA SET NAME *

female_human_trafficking_shaped.csv

65

DESCRIPTION

CSV

JSON

PARQ

Delimited

Excel

CSV

⬆

✓

The first line of the file contains column headers

4. Click **Save and Run**.

DATA REFINERY FLOW DETAILS

LOCATION
Watson Studio Labs

DATA REFINERY FLOW NAME
female_human_trafficking...
Enter a description of the Data Refinery flow

STEPS
11

Beta Limitation ⓘ
PROJECT SPARK ENVIRONMENT ⓘ
None - Use Data Refinery Default ▾
This runtime consumes 6 capacity units per hour.

Schedule
[Add Schedule](#)

* Required Fields

DATA REFINERY FLOW OUTPUT

LOCATION
Watson Studio Labs/Data assets

DATA SET NAME
female_human_traffickin...
Enter a description of the resulting data set.

☒ If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.

File format: CSV

Review the Data Refinery flow details and the Data Refinery flow output details before running the Data Refinery flow.

[Close](#)
[Save and Run flow](#)

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

What's next?

Your data flow is currently running. You can view its progress on the Summary and Runs page. When the flow completes, you can view its output from there too.

[Continue Working](#)
[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.

My Projects / [Watson Studio Labs](#) / female_human_trafficking_flow

Summary

Source

female_human_trafficking

Run Environment: Data Refinery Default

Data Refinery flow

11 Steps

Output

female_human_trafficking_shaped.csv

Runs

History

TIMESTAMP	STATUS	DURATION	ROWS READ / WRITTEN	SIZE	INITIATED BY
4 May 2019 - 02:01 am	Completed	13 sec	1592 / 269	0.0240 MB	Andrew Doe

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






My Projects / Watson Studio Labs

Overview Assets Environments Bookmarks Deployments Access Control Settings

What assets are you looking for?

▼ Data assets

0 asset selected.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	 female_human_trafficking_shaped.csv	Data Asset	Andrew Doe	4 May 2019, 2:01:59 am	
<input type="checkbox"/>	 Categories	Data Asset	Andrew Doe	30 Apr 2019, 11:47:39 pm	
<input type="checkbox"/>	 Occupation	Data Asset	Andrew Doe	30 Apr 2019, 11:47:39 pm	
<input type="checkbox"/>	 female_human_trafficking	Data Asset	Andrew Doe	30 Apr 2019, 11:47:38 pm	
<input type="checkbox"/>	 trafficking	Connection	Andrew Doe	30 Apr 2019, 11:47:36 pm	

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...	NAME	BIRTH_D...	OCCUPAT...	PASSPORT_COU...	COUNTRIES_VIS...	COUNTRY1	COUNTRY2	COUNTRY3	COUNTRIES_VISITED_CO...	ARRIVAL_AIRPORT_S...
Type: String	Type: String	Type: Date	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: Smallint	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 Mill	1997-03-11	Accounting techn	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 engin	Ghana	OM	OM			1	NM
30	Tammy Karen H	1976-03-14	Engineer, contro	Brazil	NL,KH,RU,CH,GB	NL	KH	RU	5	IL
20	Dy Rivera	1974-11-08	Engineer, manuf	Ghana	AE,SN	AE	SN		2	IN
30	Rhonnies 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 Wil	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.

