

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.

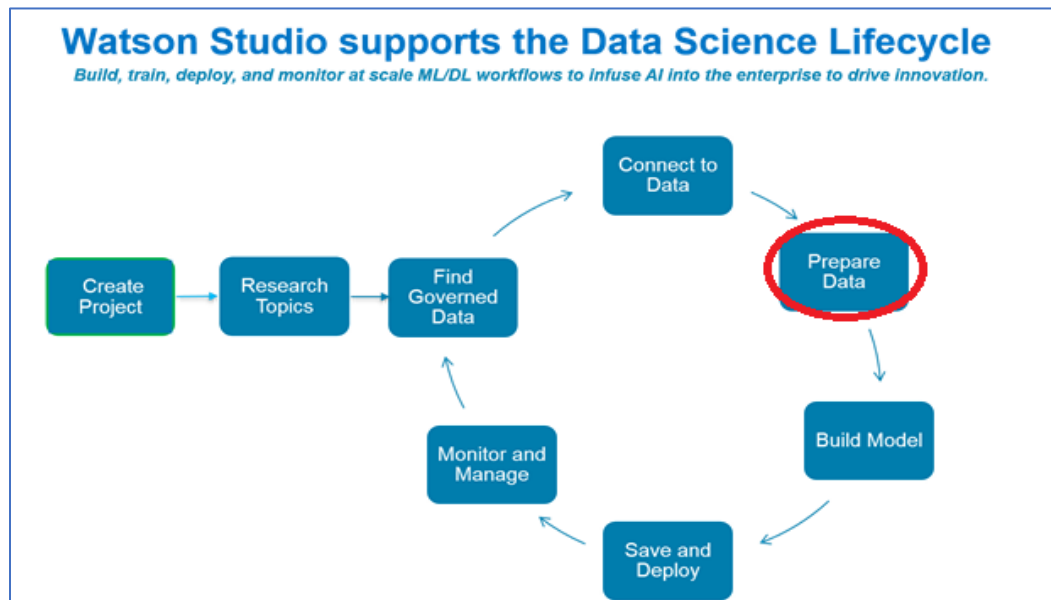


Figure 1- End to End Flow

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

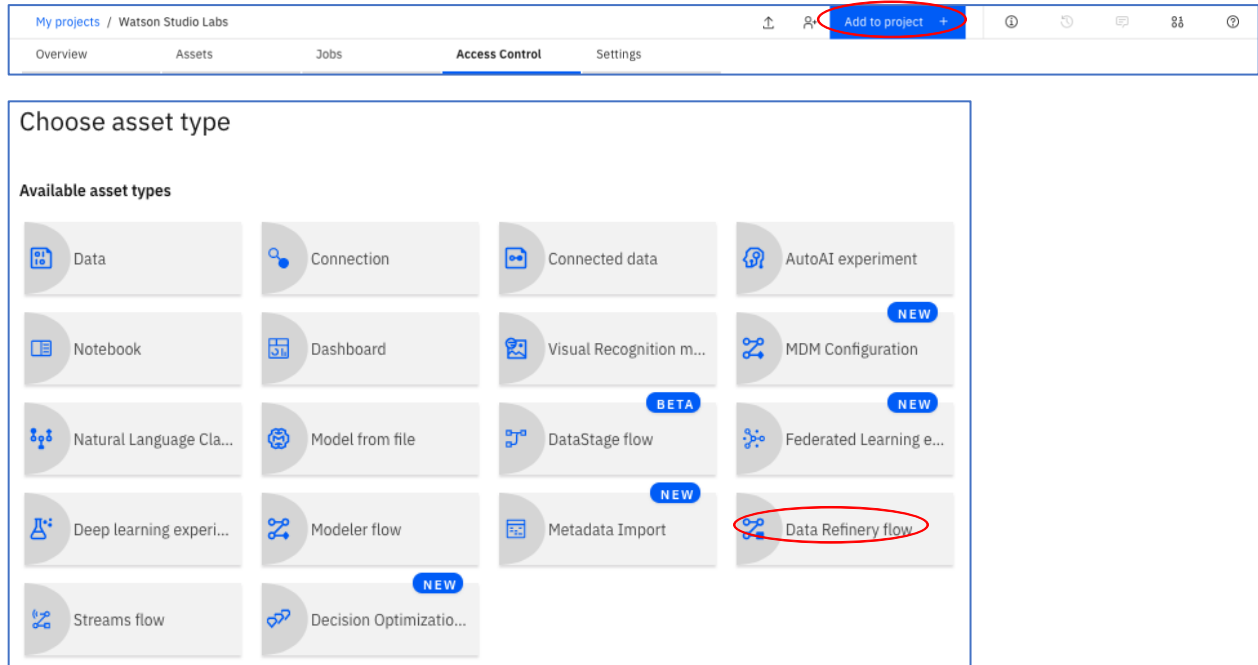
The data sets used for this lab consist of simulated travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING_LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL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

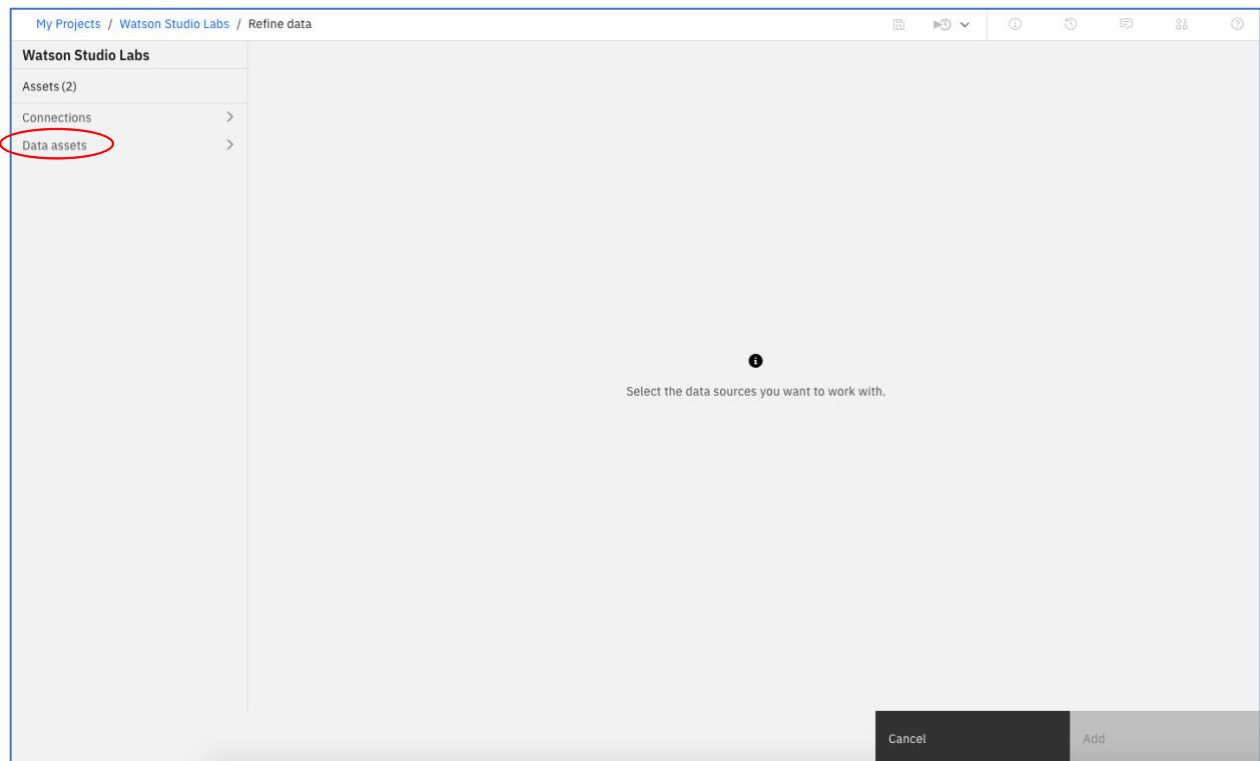
Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

Create a new Data Flow

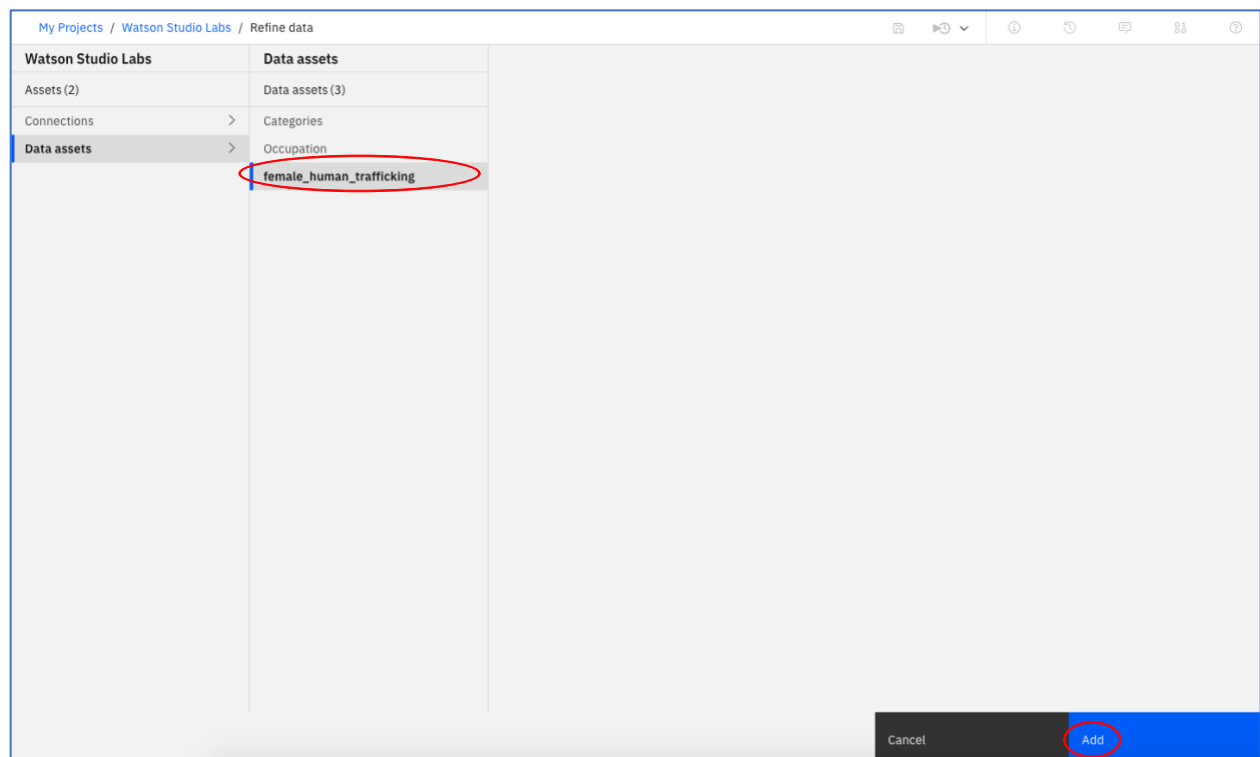
1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow**.



2. Click on **Data Assets**.



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



4. The data set will be displayed.

Projects / Watson Studio Labs / female_human_trafficking / Refine data

Operation + Code an operation to cleanse and shape your data


Data Profile Visualizations Steps

	INTERNAL_ID Integer	VETTING_LEVEL Integer	DESCRIPTION String	NAME String	GENDER Boolean	BIRTH_DATE Date	BIRTH_COUNTRY String	BIRTH_COUNT... String	OCCUPATION String
1	338	100	NA	Meghan Moses	false	1994-08-27	Ghana	GH	Forest/woodland manag
2	339	30	NA	Trace Carr	false	2001-11-30	Ghana	GH	Clinical scientist, histoc
3	340	10	NA	Ami Casey Woods	false	1983-11-05	Ghana	GH	Cartographer
4	341	30	NA	Melinda Kimm Hubbard	false	1980-01-16	Brazil	BR	Agricultural engineer
5	342	10	NA	Linda Tucker	false	1995-01-14	Brazil	BR	Translator
6	343	100	NA	Tamara Palmer	false	1987-05-01	Ghana	GH	Race relations officer
7	344	30	NA	Brandy Scott	false	1999-08-09	Ghana	GH	Field trials officer
8	345	100	NA	Nann Steffi Williamson	false	2001-06-07	Ghana	GH	Holiday representative
9	346	30	NA	Jesie Molly Stafford	false	1970-05-02	Bangladesh	BD	Pathologist
10	347	100	NA	Elizabeth Ronnie Morris	false	1995-05-07	Ghana	GH	Surgeon
11	348	30	NA	Maireag Barker	false	2001-09-24	Ghana	GH	Editor, film/video
12	349	30	NA	Crysta Nann Silva	false	1998-08-06	Ghana	GH	Volunteer coordinator
13	350	30	NA	Tanya Cameron	false	1997-03-24	Ghana	GH	Acupuncturist
14	351	100	NA	Desty Smith	false	1970-03-03	Ghana	GH	Primary school teacher
15	352	100	NA	Mary Cordova	false	1981-12-28	Ghana	GH	Medical secretary
16	353	10	NA	Rebecca Good	false	1974-03-02	Brazil	BR	Administrator, educatio
17	354	10	NA	Jacie Smith	false	2001-01-23	Ghana	GH	Fine artist
18	355	100	NA	Nannu Peterson	false	1978-08-11	Ghana	GH	Contracting civil engin

SOURCE FILE: female_human_trafficking FULL DATA SET: 1085 rows


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

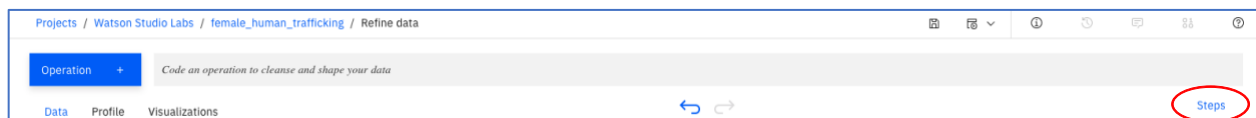
1. Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the **VETTING_LEVEL** column by hovering over VETTING_LEVEL, clicking on the vertical ellipse , clicking on **CONVERT COLUMN**, and clicking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

VETTING_LEVEL	DESCRIPTION	NAME
Integer	String	String
100	Remove	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_NU...	PASSPORT_CO...	PASSPORT_CO...	CO
Integer	String	String	Str
308561300	Remove	GH	QA
987374355	Remove duplicates	GH	QA
426221095	Remove empty rows	GH	ME
869842380	Sort ascending	BR	IL,N
473389048	Sort descending	BR	ES,
217560040	Substitute	GH	HR
942939007	CONVERT COLUMN...	Boolean	DM
768902471	View All	Decimal	IP,
730613975		Integer	IP,
798632110	Ghana	String	RU
400880971	Ghana		RU

- 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 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 +

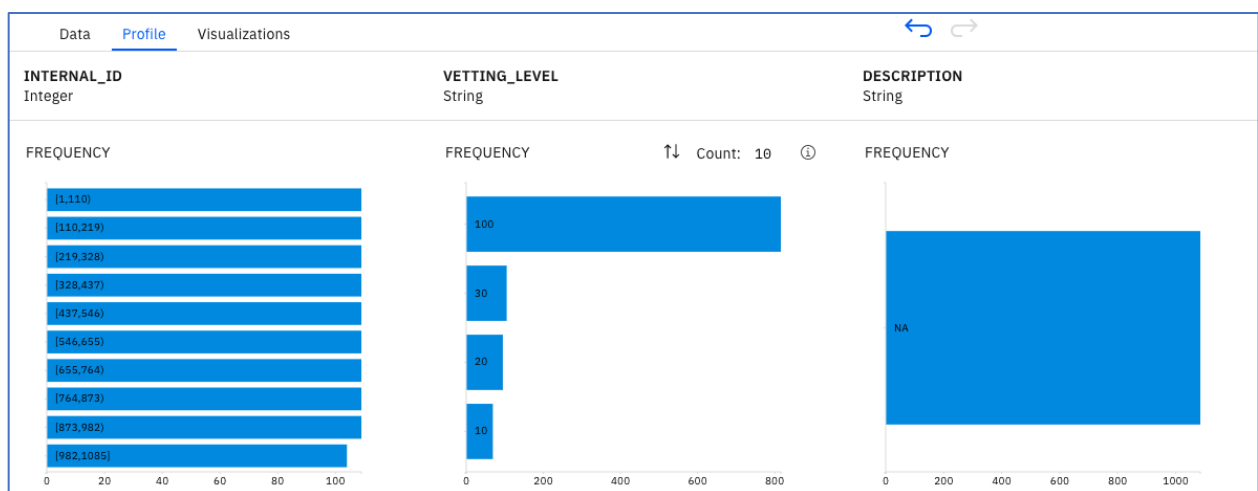
Data

Profile

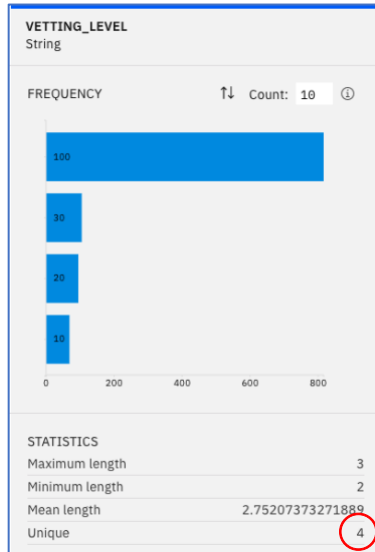
Visualizations

	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

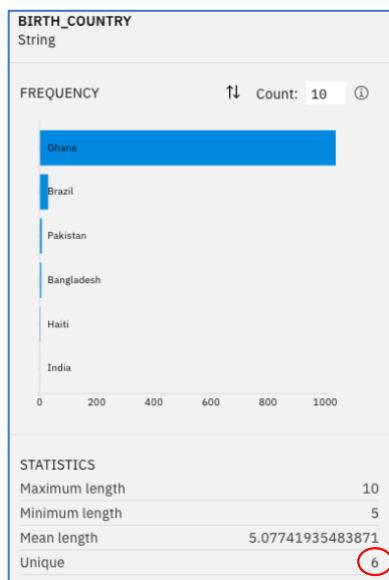
6. The Profile panel displays the counts of the top 10 values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column.



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.



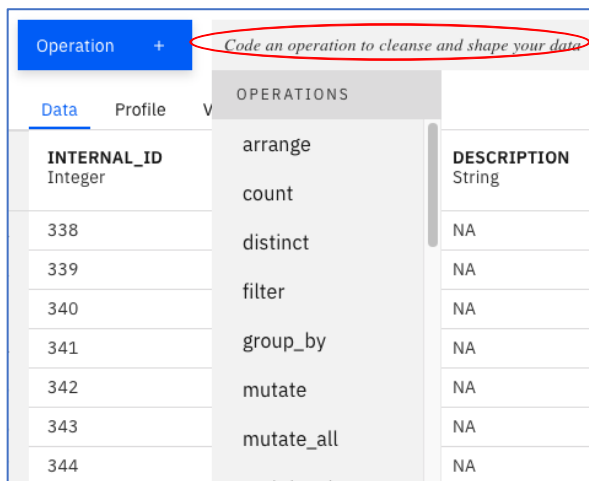
8. Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has about 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH_COUNTRY, and PASSPORT_COUNTRY shows only 6 unique countries. The COUNTRIES_VISITED_COUNT 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.



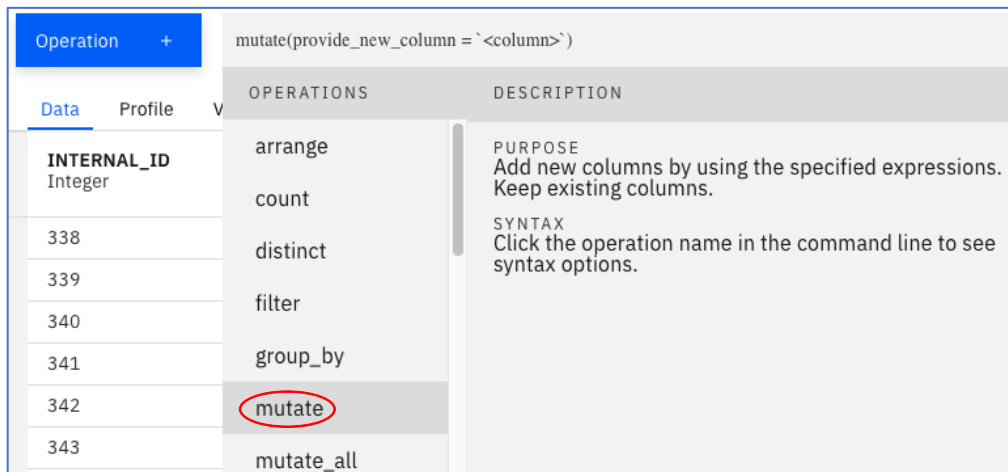
9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.



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.

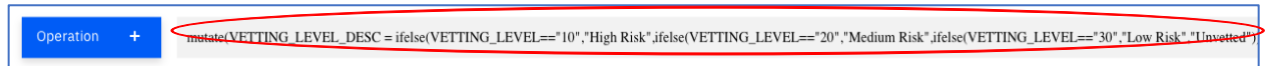


11. Hover the mouse over **mutate**. A description of the mutate function is provided.



12. Click on **mutate** and cut and replace the generated code with the following and then click **Apply**.

```
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”.

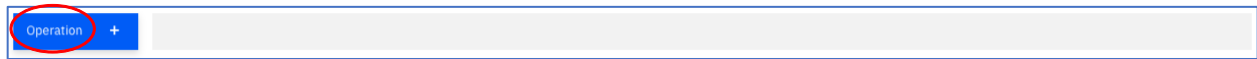
VETTING_LEVE...
String
Unvetted
Low Risk
High Risk
Low Risk
Unvetted
Unvetted
Unvetted
Unvetted
Medium Risk
Low Risk

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

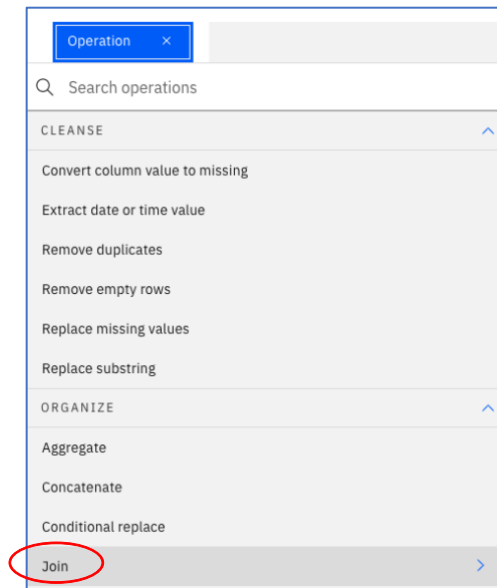
```
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DEPARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)
```



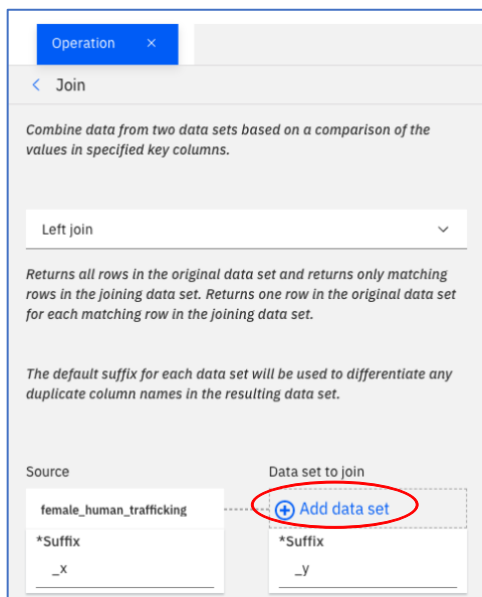
16. Let's now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the Categories dataset. Click on **Operation +**.



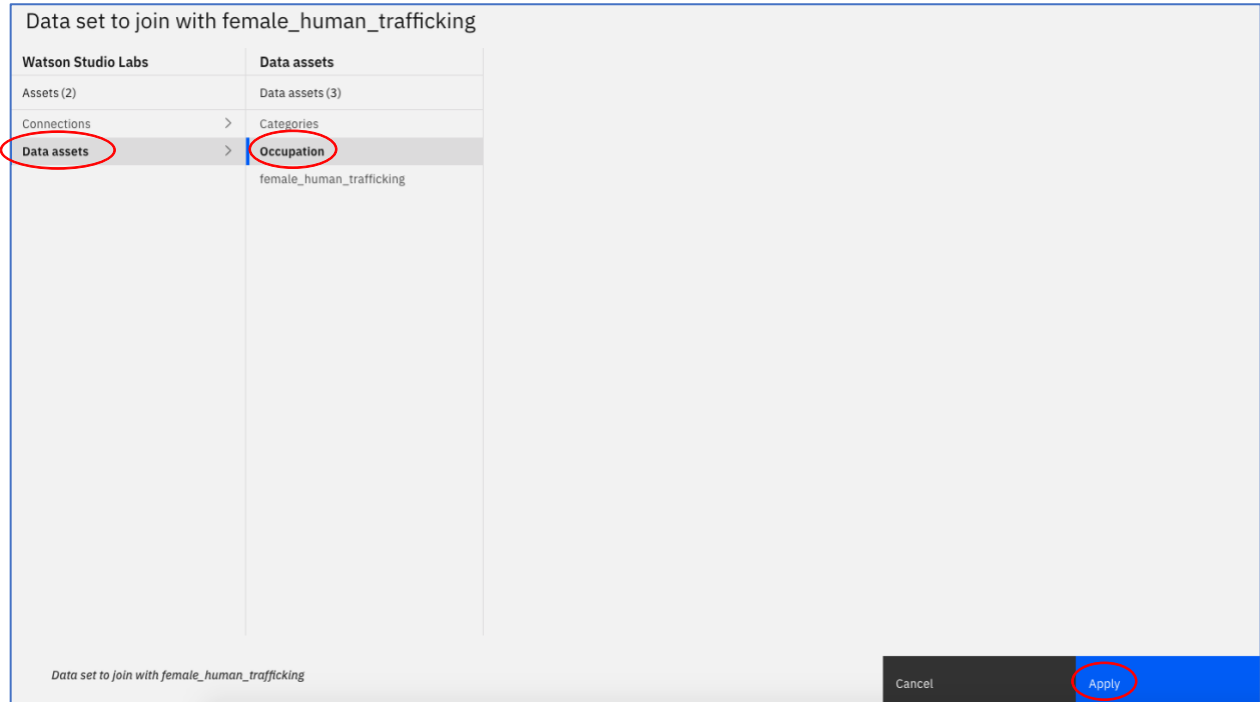
17. Scroll down and click on **Join**.



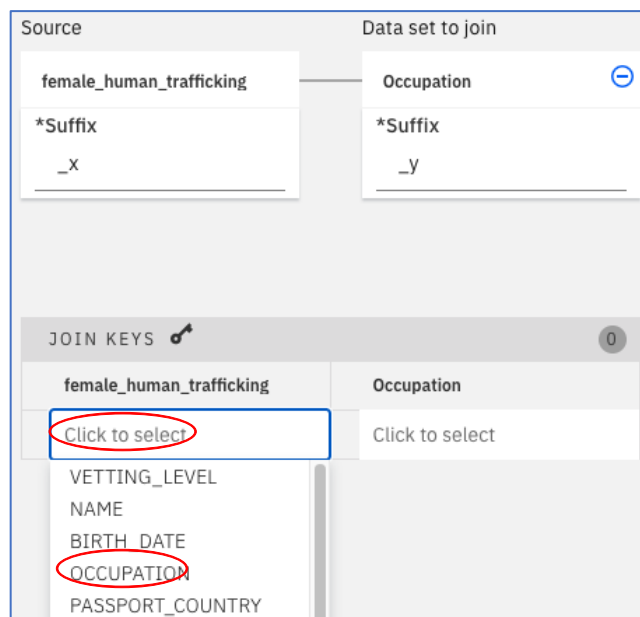
18. Keep **Left join** and then click on **Add Data Set**



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**.

Source

female_human_trafficking

*Suffix

_X

Data set to join

Occupation

*Suffix

_Y

JOIN KEYS

female_human_trafficking	Occupation
OCCUPATION	OCCUPATION

+ Add Join Key

Cancel

Next

22. Click **Apply**.

Operation

< Join

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. 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.

The screenshot shows the Data Refinery interface. On the left, a table with 18 rows and 2 columns is displayed. The columns are 'Code' (String) and 'Category' (String). The rows contain various codes and categories like 'Construction', 'Science', 'Other', 'Engineering', 'Government', 'Sports/Travel', 'Medical', 'Arts', 'Education', etc. On the right, a sidebar shows 7 steps. The first step is 'Convert column type' for the 'Code' column. The second step is 'Convert column type' for the 'Category' column. The third step is 'Custom code' with a mutate command. The fourth step is 'Custom code' with a select command. The bottom of the interface shows 'trafficking' and 'SAMPLE SIZE: First 1085 rows'.

Code	Category
11	Construction
7	Science
15	Other
2	Engineering
15	Other
5	Government
15	Other
1	Sports/Travel
6	Medical
6	Medical
8	Arts
15	Other
6	Medical
13	Education
6	Medical
13	Education
8	Arts
2	Engineering
13	Education

Steps:

- Convert column type
- Convert column type
- Custom code
- Custom code

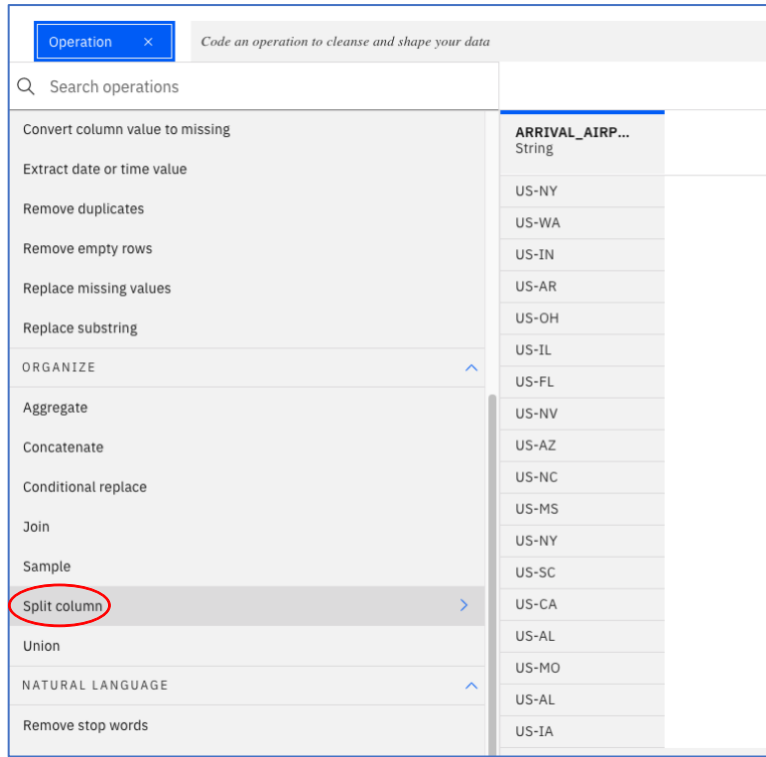
trafficking SAMPLE SIZE: First 1085 rows

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 +**.

The screenshot shows the Data Refinery interface. At the top, there is a blue 'Operation +' button. Below it, a table with 17 rows and 3 columns is displayed. The columns are 'COUNTRIES_VISITED' (String), 'COUNTRIES_VI...' (Integer), and 'ARRIVAL_AIRP...' (String). The rows contain various country codes and state abbreviations like '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,OM,TN,LU,SI,IN', 'OM,CK,BH,CK,TW,IQ,TN', 'JP,RU,CO,CU,TR,TR', 'JP,SN,SK,OM', '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'. The 'ARRIVAL_AIRP...' column contains state abbreviations like 'US-NY', 'US-WA', 'US-IN', 'US-AR', 'US-OH', 'US-IL', 'US-FL', 'US-NV', 'US-AZ', 'US-NC', 'US-MS', 'US-NY', 'US-SC', 'US-CA', 'US-AL', 'US-MO', 'US-AL'.

COUNTRIES_VISITED	COUNTRIES_VI...	ARRIVAL_AIRP...
QA	1	US-NY
QA	1	US-WA
ME,EE,KY,DZ,CZ,ID,NL,QA,BS,CK	10	US-IN
IL,VN,UZ	3	US-AR
ES,JO,LT,CL,QA,PA	6	US-OH
HR,BS,BG,AT,DK,AL,OM,TN,LU,SI,IN	12	US-IL
OM,CK,BH,CK,TW,IQ,TN	7	US-FL
JP,RU,CO,CU,TR,TR	6	US-NV
JP,SN,SK,OM	4	US-AZ
RU	1	US-NC
RU	1	US-MS
AE	1	US-NY
CH,AE,LK	3	US-SC
TR,ES,KW,SG,RU,FI,KZ,BN,JM,PT	10	US-CA
RU,DZ,KR,SN,UA,TR,MT,RS,PK	9	US-AL
ZA,EG,LY,SA,UZ,MT,AZ	7	US-MO
KW,RU,BE,KY	4	US-AL

25. Click on **Split column**.



26. 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**.

Operation

Code an operation to cleanse and shape your data

< Split column

Change column selection

Selected column: ARRIVAL_AIRPORT_REGION

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

DEFAULT

TEXT

PATTERN

POSITION

Hyphen (-)

ARRIVAL_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

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	

Remove

Remove duplicates

Remove empty rows

Sort ascending

Sort descending

Substitute

CONVERT COLU...>

TEXT >

View All

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 String	NAME String
100	
30	
10	
30	
10	
100	
30	
100	
30	
100	

Remove

Remove duplicates

Remove empty rows

Sort ascending

Sort descending

Substitute

CONVERT COLU...>

TEXT >

View All

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 x

< 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
7	
15	
2	

Remove

Remove duplicates

Remove empty rows

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

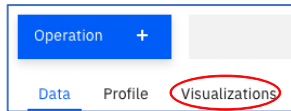
My Projects / Watson Studio Labs / female_human_trafficking / Refine data

Operation +

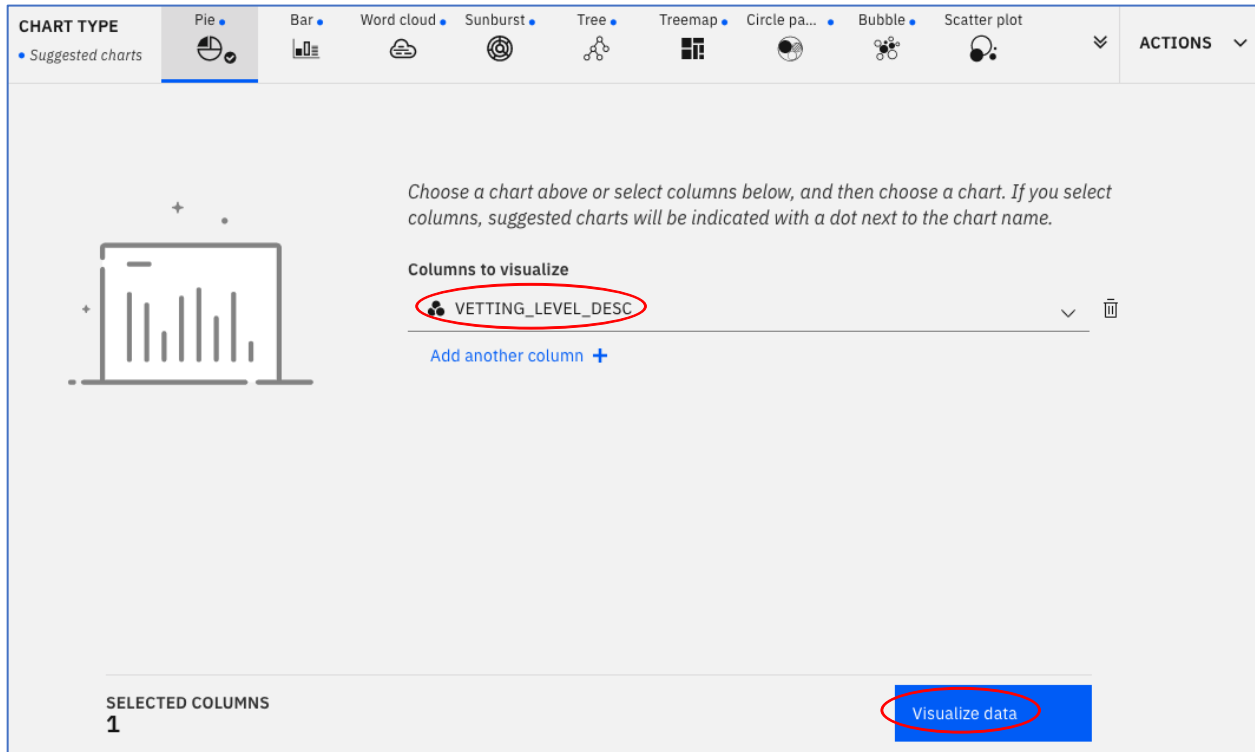
Data Profile Visualizations

Save

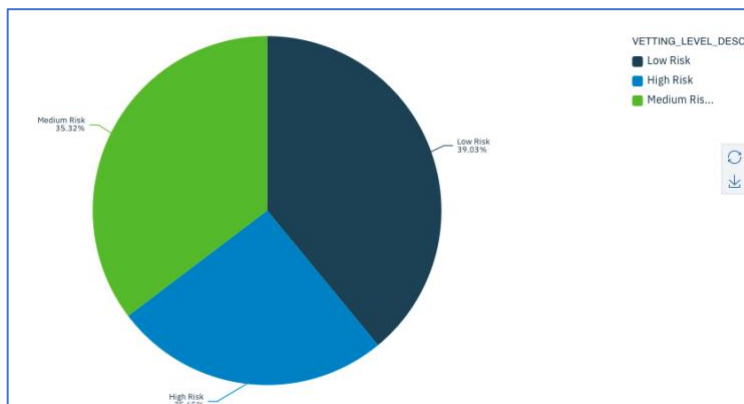
35. Click on the **Visualization** tab.



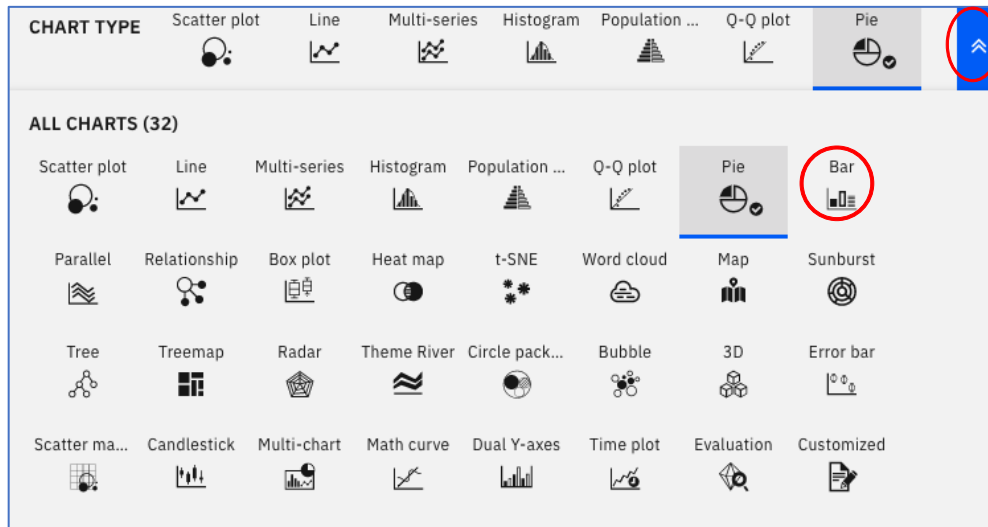
36. Click on **VETTING_LEVEL_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



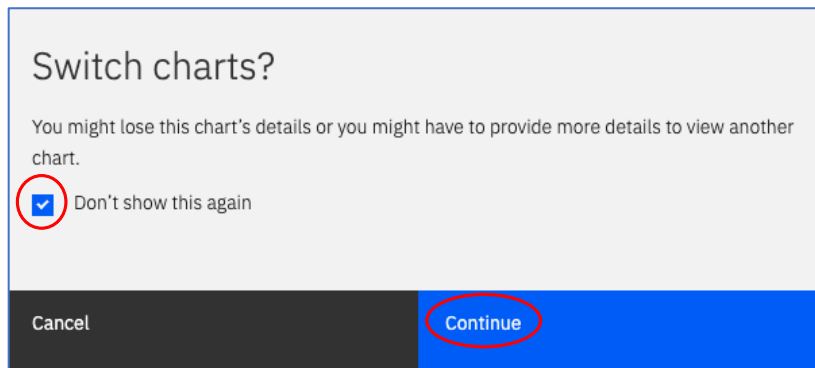
37. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



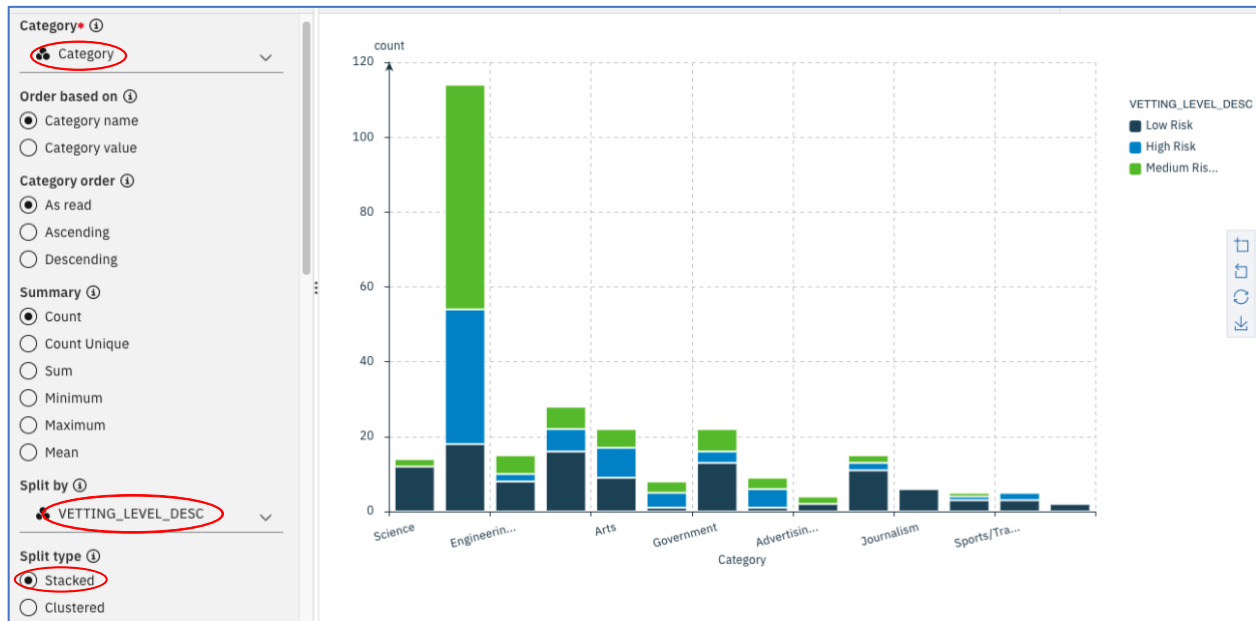
38. We can visualize the breakdown of travel records by job category and vetting level.
Click on the dropdown icon and then click **Bar**.



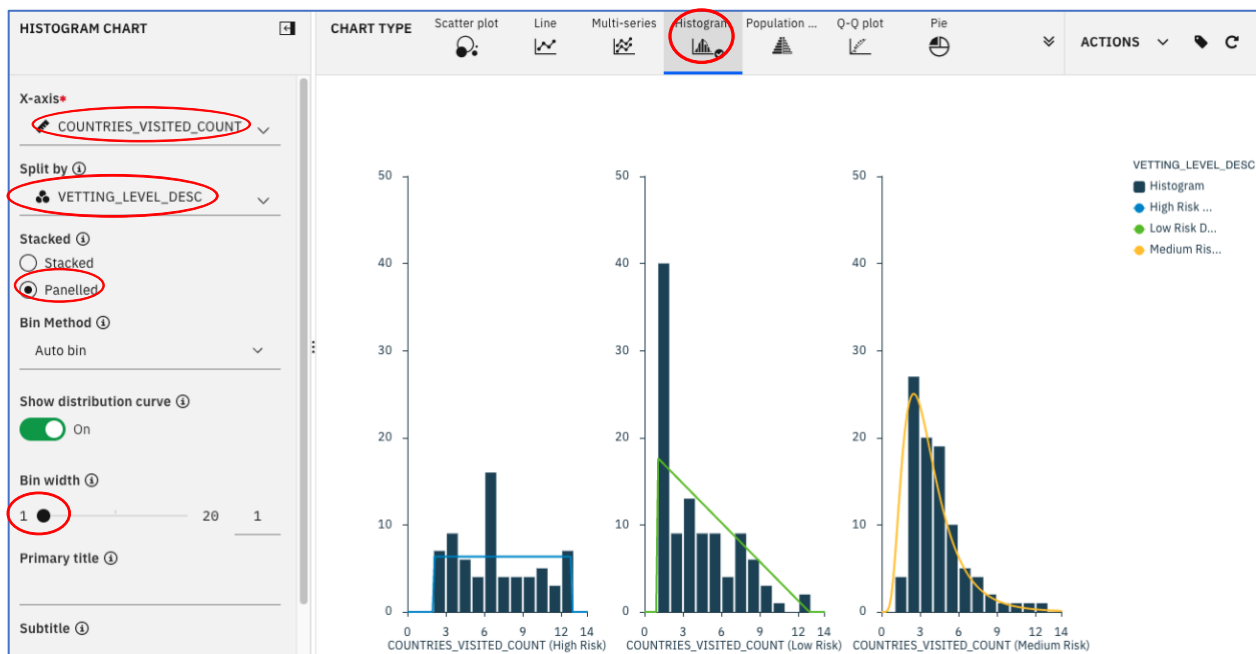
39. Click on **Don't show this again**. Click on **Continue**



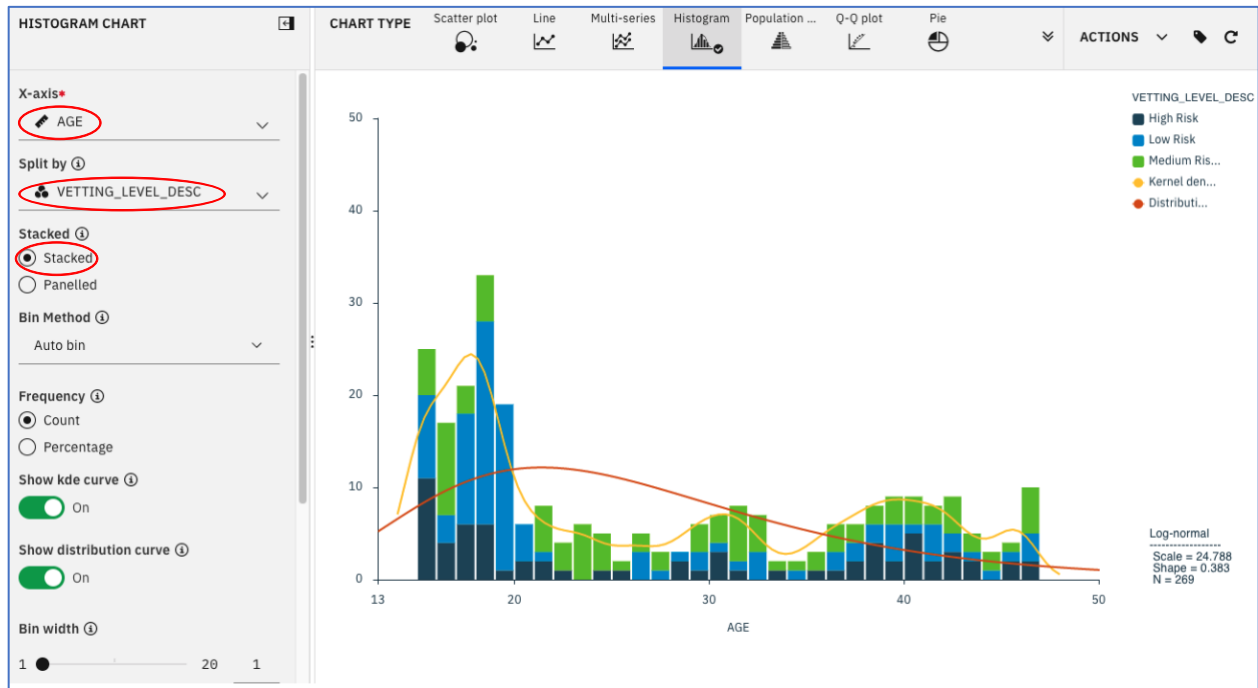
40. Click on **Category** for **Category**, click on VETTING_LEVEL_DESC for **Split by**, click on **Stacked** for **Split type**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



41. 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**, click on **Paneled**, and drag **Bin width** to 1. Note that at higher number of countries visited, there is an increasing likelihood that it is a high-risk person.





42. Let's examine if age makes a difference. Click on **AGE** for **X-axis**. **Split by** remains **VETTING_LEVEL_DESC**, click on **Stacked**. It appears that younger travelers have a lower risk of being trafficked.

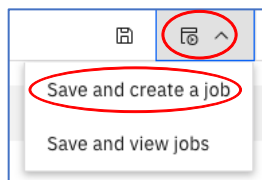


43. Please feel free to experiment with other visualizations.

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 .

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



45. Enter a **Job Name** for the job. Note the number of steps used to transform the data. It should be 10-12 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' and the 'Description (optional)' field is empty. The 'Next' button is highlighted.

Create a job

Define details

Associated asset
female_human_trafficking_flow (13 Steps)

Name
FHT Data Refinery

Description (optional)
Description of job

Cancel Next

46. 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 in CSV format. The 'Environment' is set to 'Default Data Refinery XS'. The 'Next' button is highlighted.

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

47. 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

48. Click **Create and run**.

Create a job

Define details
FHT Data Refinery

Configure
Default Data Refinery XS

Schedule

Review and create

Details

Associated asset
female_human_trafficki... (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

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

My projects / Watson Studio Labs / FHT Data Refinery					
FHT Data Refinery					
No description					
Runs (1)					
Start time	↓	Status	Duration	Started by	Action
Jan 10, 2021 1:29:28 PM		Completed	1 minute 26 seconds	FCTO Labs	

50. 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
-------------	--------------------	-------------------

51. Click on the **female_human_trafficking_shaped.csv** to view the contents.

▼

Data assets

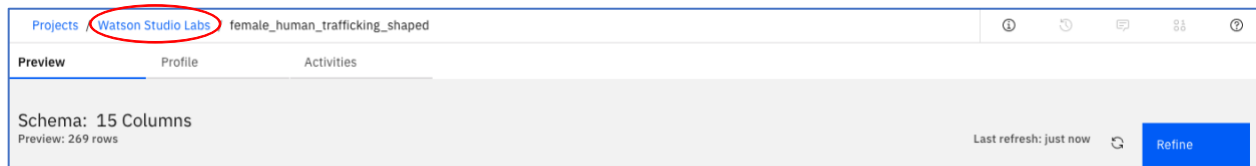
0 assets selected.

<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	

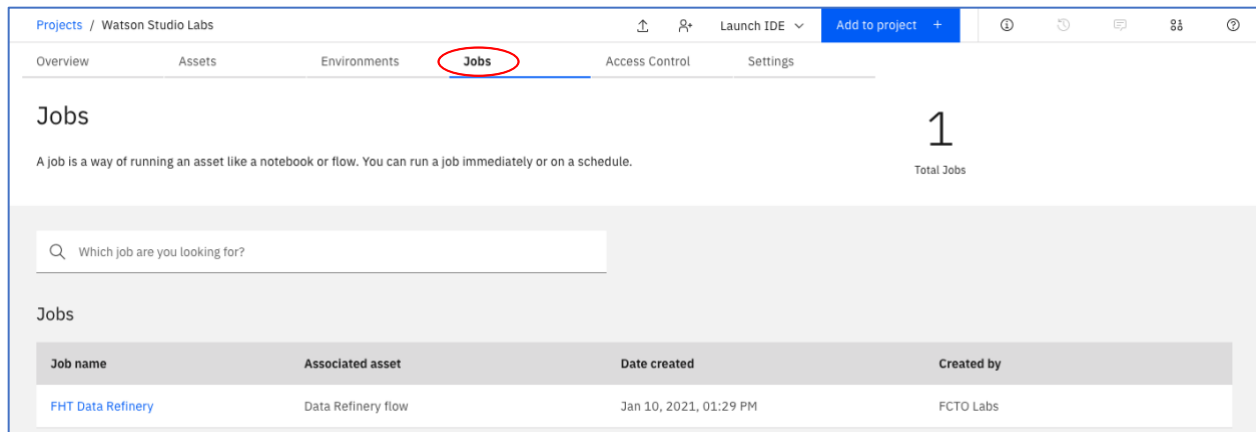
52. The asset contents are displayed below. Review to confirm that the data transformations specified have been applied to all the data.

My Projects / Watson Studio Labs / female_human_trafficking_shap...										
Preview										
Schema: 15 Columns										
Preview: 269 rows										
Last refresh: 16 seconds ago										
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 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

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



54. 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.