

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.

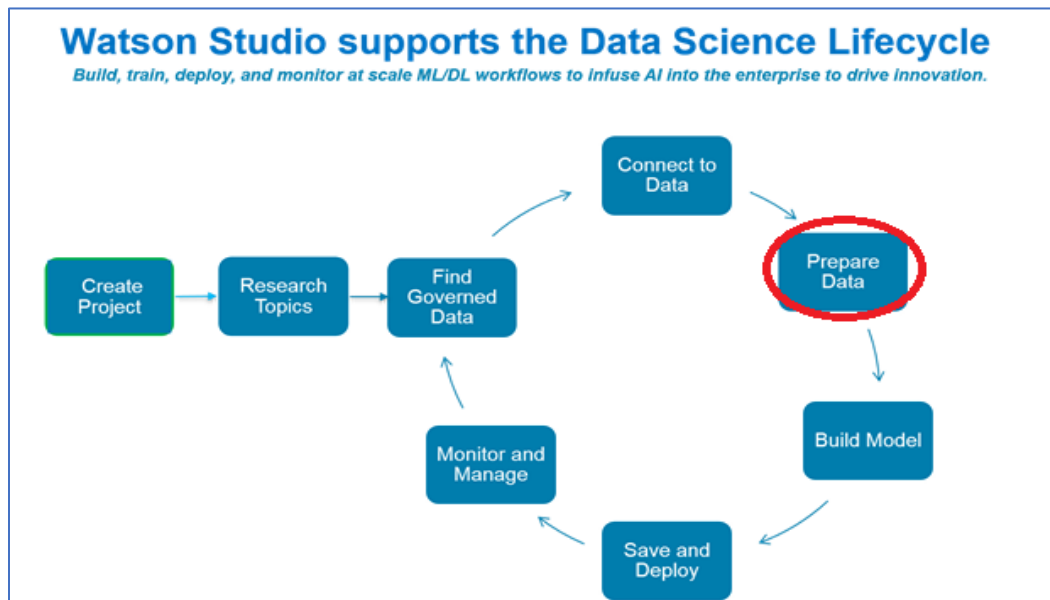


Figure 1- End to End Flow

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

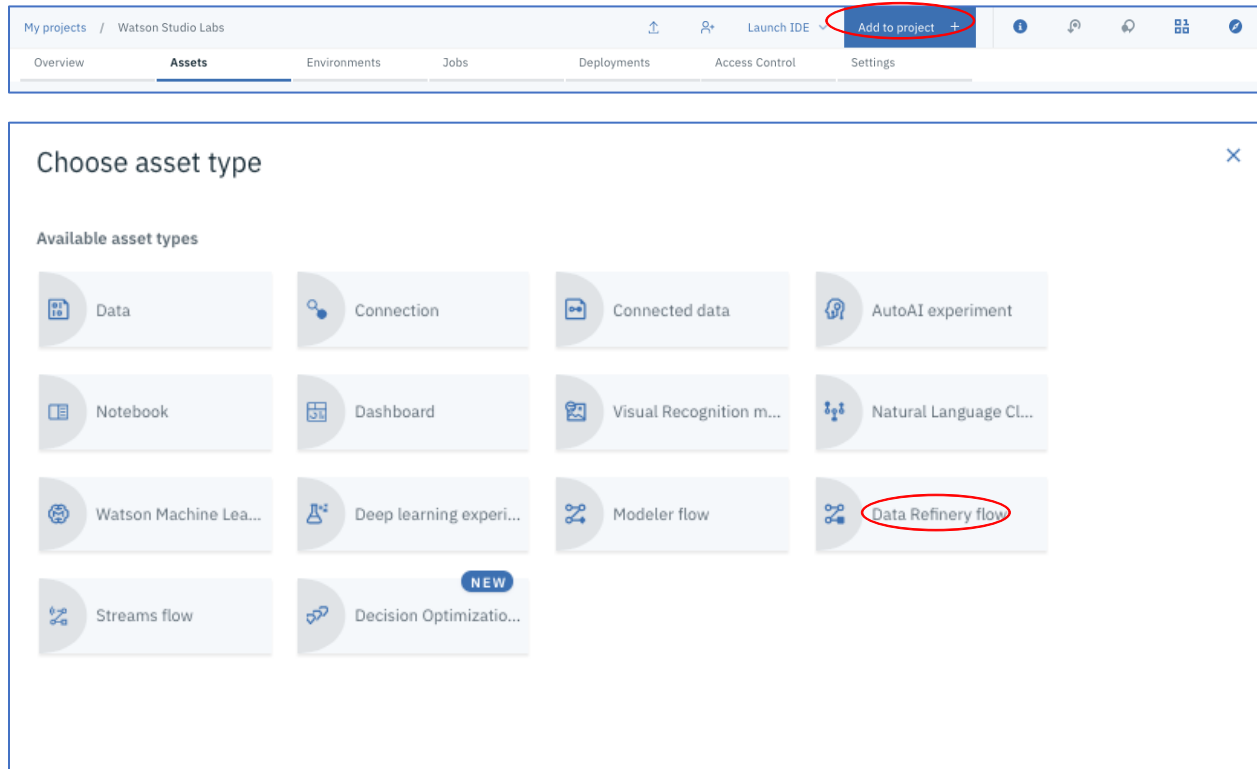
The data sets used for this lab consist of simulated travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING_LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL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

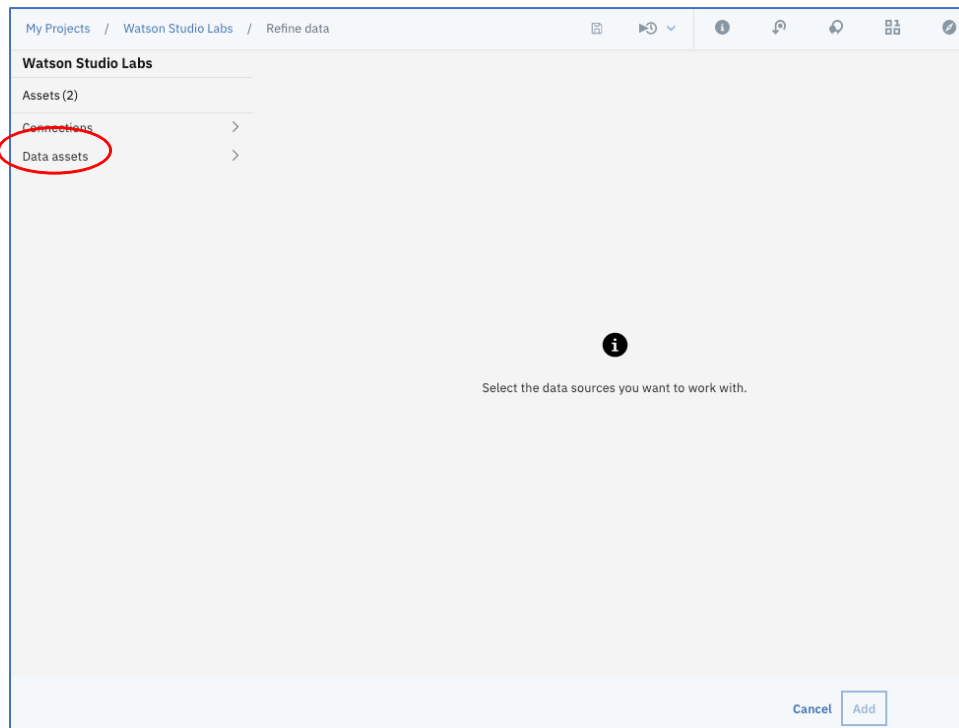
Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

Create a new Data Flow

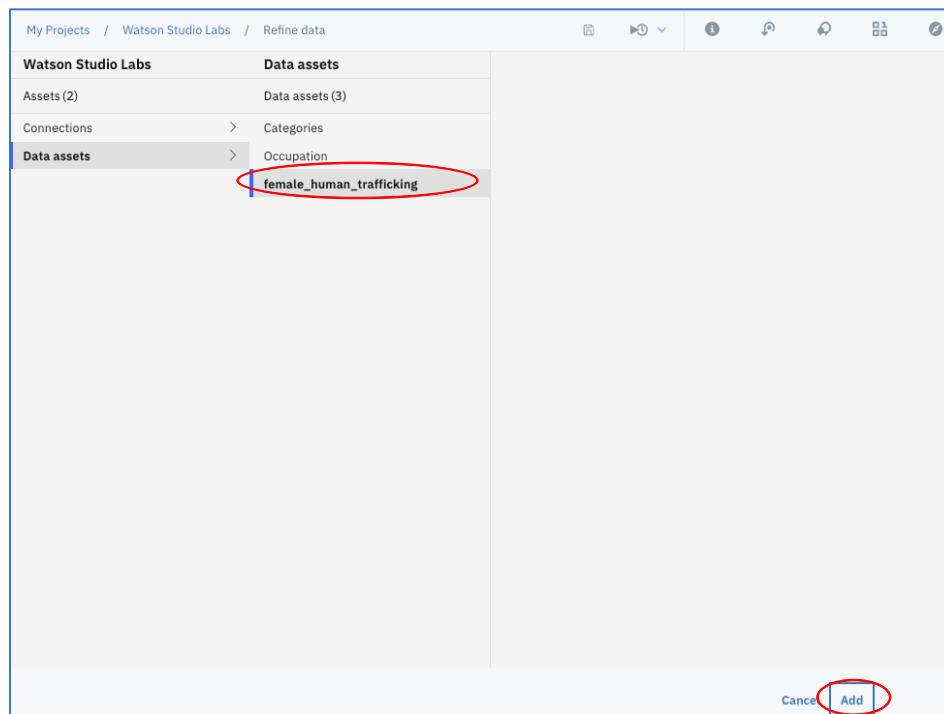
1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow**.



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 / 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 String	BIRTH_DATE Date	BIRTH_COUNTRY String
1	512	30	NA	Geordie Cindy Keith	F	1997-10-13	Ghana
2	513	30	NA	Lisa Lei Lindsey	F	1999-05-23	Ghana
3	514	100	NA	Sassa Christy Melendez	F	1996-11-14	Ghana
4	515	20	NA	Missy Christina Garcia	F	1987-02-13	Ghana
5	516	30	NA	Cindi Lei Lara	F	1997-04-17	Ghana
6	517	100	NA	Vicki Dodie Blanchard	F	1975-10-09	Ghana
7	518	30	NA	Genna Linda Wilson	F	1997-04-11	Ghana
8	519	100	NA	Sadie Chavez	F	1980-05-26	Ghana
9	520	100	NA	Shelle Teri Fitzgerald	F	1994-06-22	Ghana
10	521	100	NA	Mandie Kelsey Melendez	F	1975-09-14	Ghana
11	522	100	NA	Joci Hebert	F	1991-06-02	Ghana
12	523	100	NA	Angela Summer Marks	F	1977-09-11	Ghana
13	524	30	NA	Shelia Burns	F	1999-05-25	Ghana
14	525	100	NA	Olivia Stacy Cox	F	1981-03-16	Ghana
15	526	100	NA	Jessica Hernandez	F	1972-07-24	Ghana
16	527	100	NA	Mary Morgan	F	1993-04-14	Ghana

SOURCE FILE: female_human_trafficking


SAMPLE SIZE: First 1000 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.

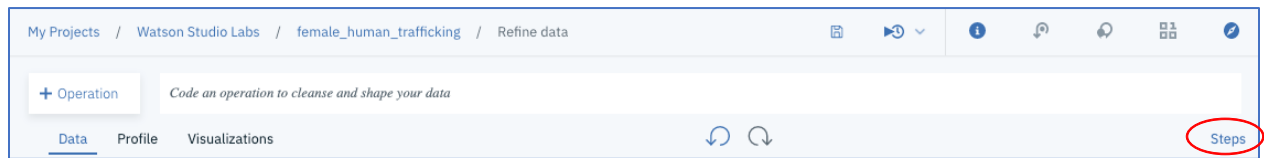
1. Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the VETTING_LEVEL column by hovering over VETTING_LEVEL, clicking on the vertical ellipse , clicking on CONVERT COLUMN, and clicking on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

VETTING_LEVEL	DESCRIPTION	NAME
Integer		String
30	Remove	Geordie Cindy Keith
30	Remove duplicates	Lisa Lei Lindsey
100	Remove empty rows	Sassa Christy Melendez
20	Sort ascending	Missy Christina Garcia
30	Sort descending	Cindi Lei Lara
100	Substitute	Vicki Dodie Blanchard
30		Genna Linda Wilson
100	CONVERT COLUMN... >	Boolean
100	View All	Decimal
100	NA	Integer
100	NA	String
30	NA	

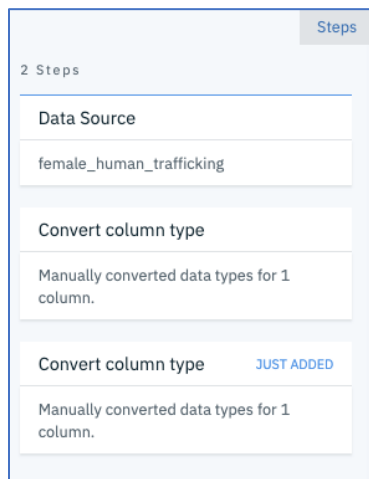
2. Convert the PASSPORT_NUMBER column by hovering over PASSPORT_NUMBER, clicking on the vertical ellipse , clicking on CONVERT COLUMN, and clicking on String.

PASSPORT_NU...	PASSPORT_CO...	PASSPORT_CO...
Integer		String
930306073	Remove	GH
440899315	Remove duplicates	GH
55972000	Remove empty rows	GH
684975534	Sort ascending	GH
13365109	Sort descending	GH
790015024	Substitute	GH
551047600		
962436210	CONVERT COLUMN... >	Boolean
165786632	View All	Decimal
647198068		
213090641	Ghana	Integer
958766851	Ghana	String
481245934	Ghana	

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



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



5. Click on **Profile**.

My Projects / Watson Studio Labs / female_human_trafficking / Refine data

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

Data **Profile** Visualizations

	INTERNAL_ID	VETTING_LEVEL	DESCRIPTION	NAME	GENDER
	Integer	String	String	String	String
1	512	30	NA	Geordie Cindy Keith	F
2	513	30	NA	Lisa Lei Lindsey	F
3	514	100	NA	Sassa Christy Melendez	F
4	515	20	NA	Missy Christina Garcia	F
5	516	30	NA	Cindi Lei Lara	F
6	517	100	NA	Vicki Dodie Blanchard	F
7	518	30	NA	Genna Linda Wilson	F
8	519	100	NA	Sadie Chavez	F
9	520	100	NA	Shelle Teri Fitzgerald	F
10	521	100	NA	Mandie Kelsey Melendez	F
11	522	100	NA	Joci Hebert	F
12	523	100	NA	Angela Summer Marks	F
13	524	30	NA	Shelia Burns	F
14	525	100	NA	Olivia Stacy Cox	F
15	526	100	NA	Jessica Hernandez	F
16	527	100	NA	Mary Morgan	F

SOURCE FILE: female_human_trafficking SAMPLE SIZE: First 1000 rows

2 Steps

Data Source

female_human_trafficking

Convert column type

Manually converted data types for 1 column.

Convert column type [JUST ADDED](#)

Manually converted data types for 1 column.

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.

My Projects / Watson Studio Labs / female_human_trafficking / Refine data

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

Data **Profile** Visualizations

INTERNAL_ID
Integer

FREQUENCY

STATISTICS

Interquartile Range	584.5
Minimum	1
Maximum	1085
Median	500.5
Standard Deviation	321.669999165858

VETTING_LEVEL
String

FREQUENCY

STATISTICS

Maximum length	3
Minimum length	2
Mean length	2.748
Unique	4

DESCRIPTION
String

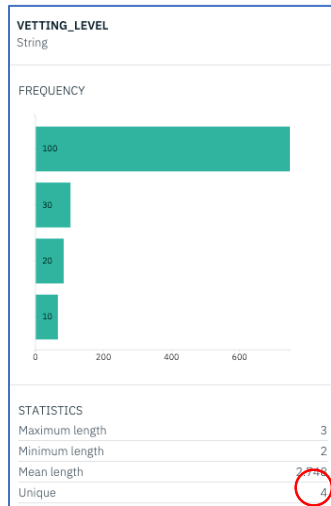
FREQUENCY

Count: 10

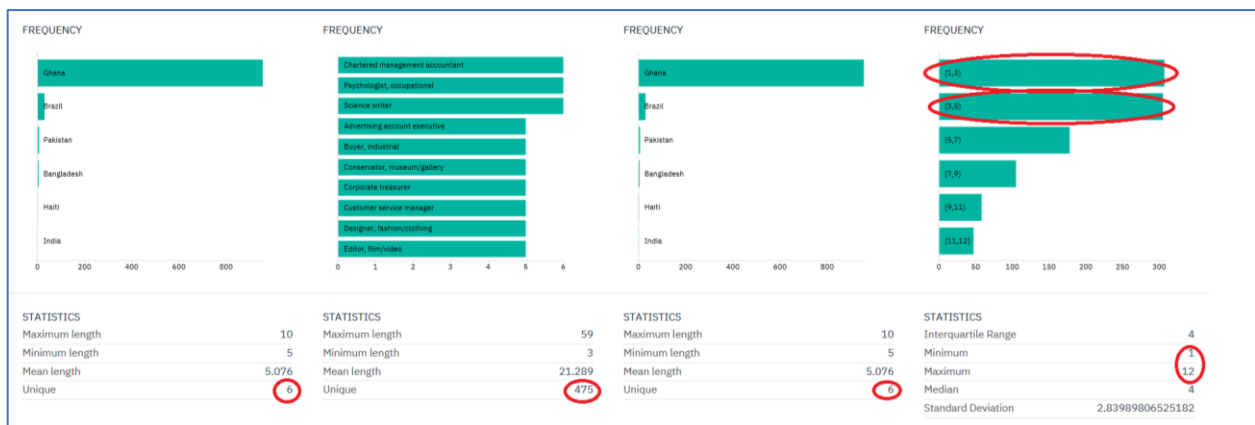
STATISTICS

Maximum length	
Minimum length	
Mean length	
Unique	

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

+ Operation		Code an operation to cleanse and shape your data	
	Data	Profile	OPERATIONS
	INTERNAL_ID Integer		DESCRIPTION String
		arrange	
		count	
1	512	distinct	NA
2	513		NA
3	514	filter	NA
4	515	group_by	NA
5	516	mutate	NA
6	517	mutate_all	NA
7	518		NA

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

+ Operation		mutate(provide_new_column = `<column>`)	
	Data	Profile	OPERATIONS
	INTERNAL_ID Integer		DESCRIPTION
		arrange	PURPOSE
		count	Add new columns by using the specified expressions. Keep existing columns.
1	512	distinct	
2	513	filter	SYNTAX
3	514	group_by	Click the operation name in the command line to see syntax options.
4	515	mutate	
5	516	mutate_all	
6	517		
7	518		
8	519	100	
9	520	100	
10	521	100	

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"))))
```

+ Operation
mutate(VETTING_LEVEL_DESC = ifelse(VETTING_LEVEL=="10","High Risk",ifelse(VETTING_LEVEL=="20","Medium Risk",ifelse(VETTING_LEVEL=="30","Low Risk","Unvetted"))))
Apply | Cancel

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.

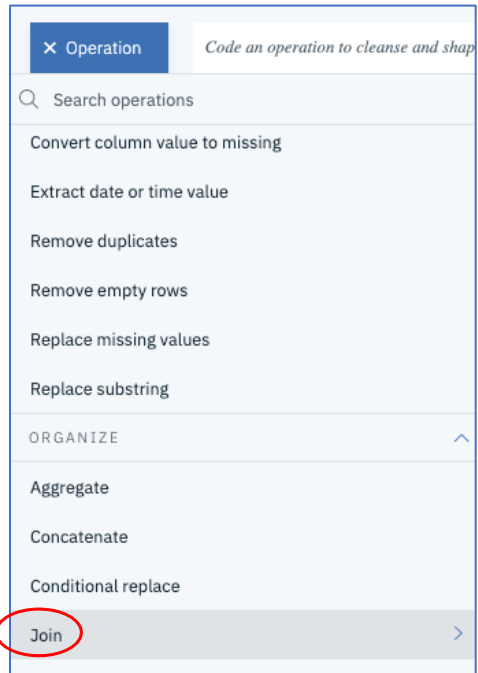
```
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DEPARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)
```

+ Operation
select(VETTING_LEVEL,NAME,BIRTH_DATE,OCCUPATION,PASSPORT_COUNTRY,COUNTRIES_VISITED,COUNTRIES_VISITED_COUNT,ARRIVAL_AIRPORT_REGION,DEPARTURE_AIRPORT_REGION,AGE,VETTING_LEVEL_DESC)
Apply | Cancel

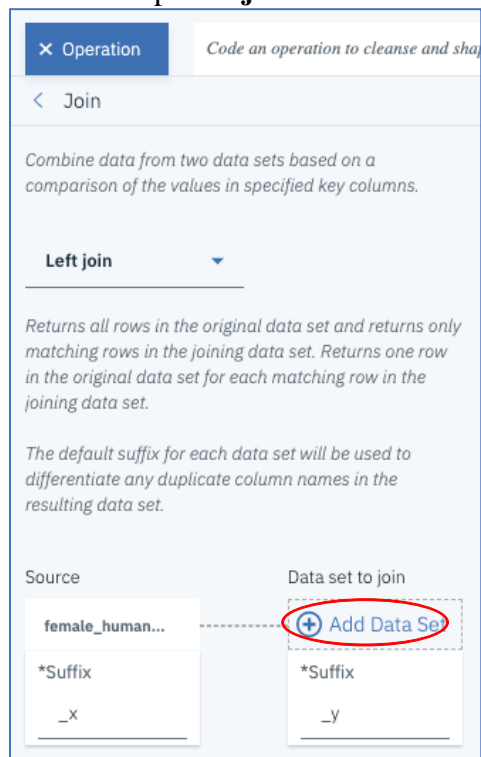
15. Let’s now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the Categories dataset. Click on + **Operation**.

+ Operation
Code an operation to cleanse and shape your data

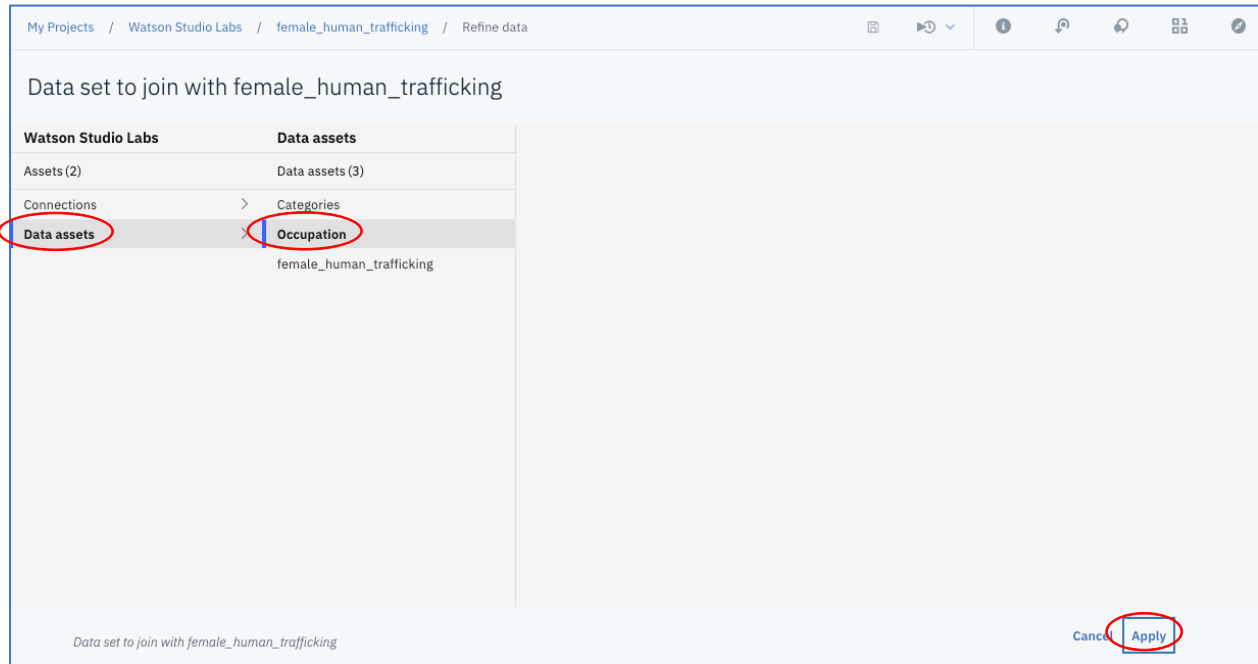
16. Scroll down and click on **Join**.



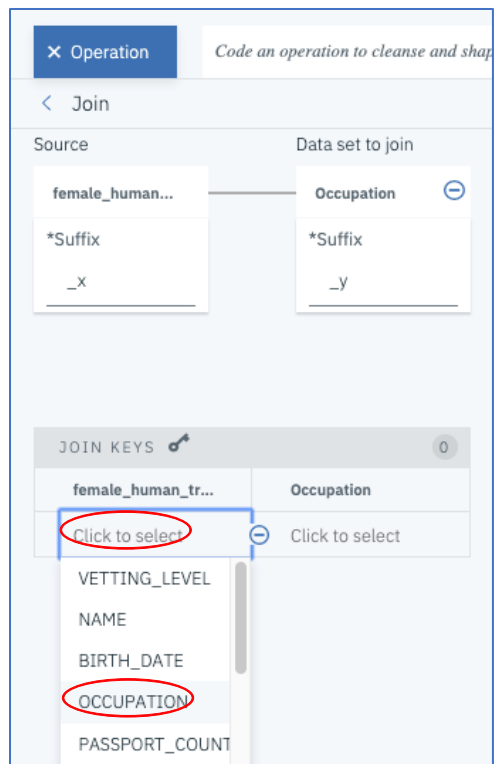
17. Keep **Left join** and then click on **Add Data Set**



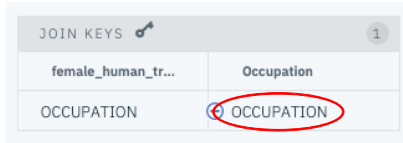
18. Click on **Data Assets**, click on **Occupation**, and then click **Apply**.





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

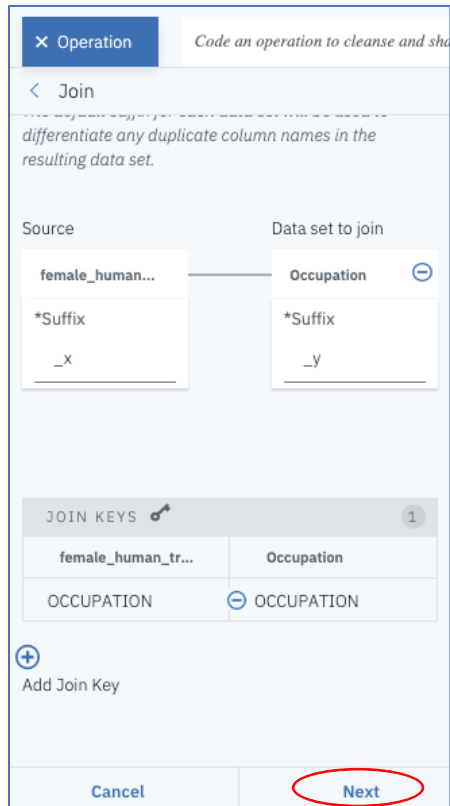


20. In **JOIN KEYS** under **Occupation** click **Click to select** and then click **OCCUPATION**.



JOIN KEYS 	
female_human_tr...	Occupation
OCCUPATION	 OCCUPATION

21. Click on **Next**.



Operation *Code an operation to cleanse and sha*

Join


differentiate any duplicate column names in the resulting data set.

Source: female_human...
*Suffix: _x

Data set to join: Occupation
*Suffix: _y

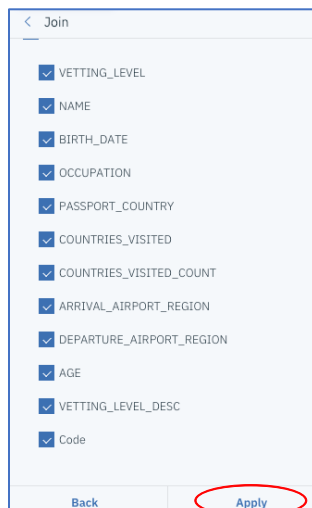
JOIN KEYS 

female_human_tr...	Occupation
OCCUPATION	 OCCUPATION

 Add Join Key

Cancel **Next**

22. Click **Apply**.



Join

- ☒ 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**

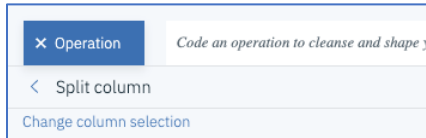
23. Follow steps 19-22 to join the Categories dataset. The join keys are the Code fields in both datasets. As a result of the joins, two new columns are added, a Code column, and a Category column. The flow has 6 overall steps, with the two Join steps shown. Note it will show 4 steps if you skipped steps 1-4 above.

The screenshot shows a data flow tool interface. On the left, a table displays data with two columns: 'Code' (String) and 'Category' (String). The 'Code' column contains values like 7, 8, 5, 15, 5, 7, 2, 1, 2, 11, 11, 3, 6, 8, 7, 8, 7, 10, 6, 13, 11, 6. The 'Category' column contains values like Science, Arts, Government, Other, Government, Science, Engineering, Sports/Travel, Engineering, Construction, Construction, Information Techn, Medical, Arts, Science, Arts, Science, Legal, Medical, Education, Construction, Medical. On the right, a panel shows the flow steps. The first step is 'Convert column type'. The second step is 'Custom code' with a mutate function. 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). The panel also shows a '6 Steps' indicator and a 'JUST' button.

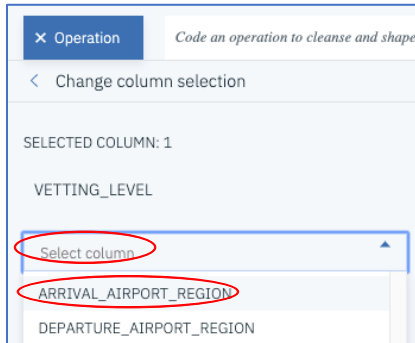
24. We note that the ARRIVAL_AIRPORT_REGION column has “US” concatenated with a State abbreviation (eg US-CA) We want to strip away the “US” to use the column as a State column. The operation **Split column** can be used. Click on + **Operations** then click on **Split column**.

The screenshot shows the 'Operations' menu in a data flow tool. The menu is titled 'Operation' and contains a list of operations: Remove duplicates, Remove empty rows, Replace missing values, Replace substring, ORGANIZE, Aggregate, Concatenate, Conditional replace, Join, Sample, and Split column. The 'Split column' operation is highlighted with a red circle.

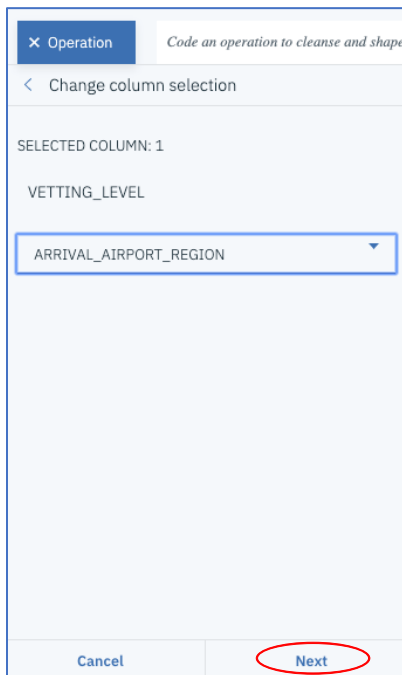
25. Click on **Change column selection**.



26. Click on **Select column**. Then click on **ARRIVAL_AIRPORT_REGION**.



27. Click on **Next**.



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

•
•
•

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**, 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.

Split column

Change column selection

Selected column: COUNTRIES_VISITED

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

DEFAU... **TEXT** PATTE... POSITI...

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**

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_V Integer
1	UZ	UZ			1
2	NO,AE	NO	AE		2
3	QA	QA			1
4	CK,OM,SI,SE	CK	OM	SI	4
5	QA	QA			1
6	KH,RU,MT	KH	RU	MT	3
7	SN,CO,CN,NG,KY,TH,RU,IT	SN	CO	CN	8
8	AE	AE			1
9	TR,AE	TR	AE		2
10	OM	OM			1
11	LT,RU,HR,SG,IR,PG,SD,QA	LT	RU	HR	8
12	IN,AE,MT,KR,KR,TR,MX,KY,BY,AZ	IN	AE	MT	10
13	AE,ZA,FR	AE	ZA	FR	3
14	OM,BY,UA,RO,QA,RU,LK,JP,PG,BN,BH,PG	OM	BY	UA	12
15	CK,GB,BH,RU,CN,PA,MT	CK	GB	BH	7
16	RU,HK,RO,SA,AT	RU	HK	RO	5
17	QA	QA			1
18	AE,SA	AE	SA		2
19	QA,EG,IQ,CN	QA	EG	IQ	4

SOURCE FILE: female_human_trafficking SAMPLE SIZE: First 1000 rows

9 Steps

Join

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 JUST

Split COUNTRIES_VISITED by text , into COUNTRY1,COUNTRY2,COUNTRY3

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**.

VETTING_LEVEL String	NAME String
30	
30	
100	
20	
30	
100	
30	
100	
100	
100	
100	
100	
100	

Remove

Remove duplicates

Remove empty rows

Sort ascending

Sort descending

Substitute

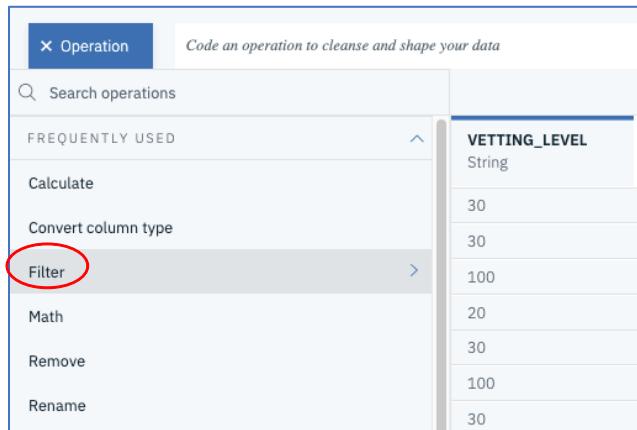
CONVERT COLUMN... >

TEXT >

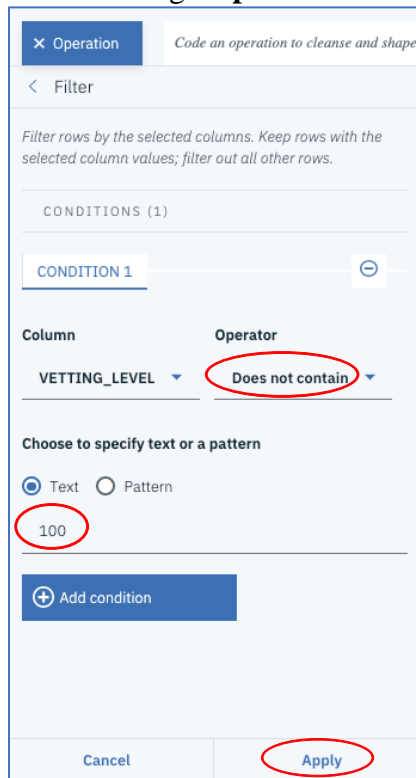
View All


Angela Summer Marks

34. Click on **Filter**.




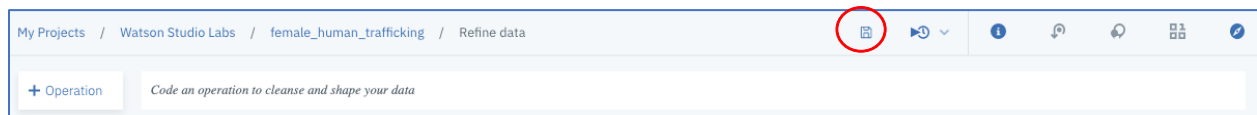
35. Change **Operator** to **Does not contain**, put value as 100, and then click **Apply**.



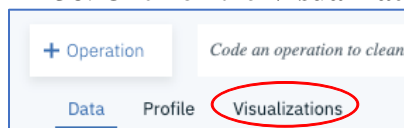
36. Remove the Code column by clicking on the vertical ellipse  and then clicking **Remove**.

Code	Category
String	String
7	Remove
8	Remove duplicates
15	Remove empty rows
5	Sort ascending
2	Sort descending
6	Substitute
13	CONVERT COLUMN... >
14	TEXT >
9	View All
15	Medical
6	
6	

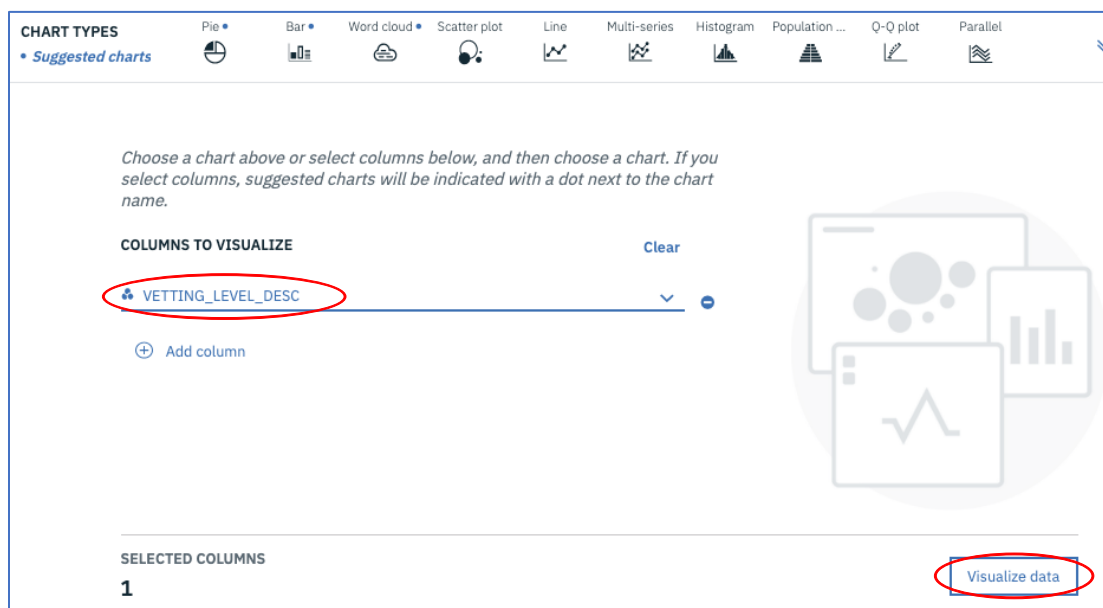
37. Save the Data Flow by clicking on the Save  icon.



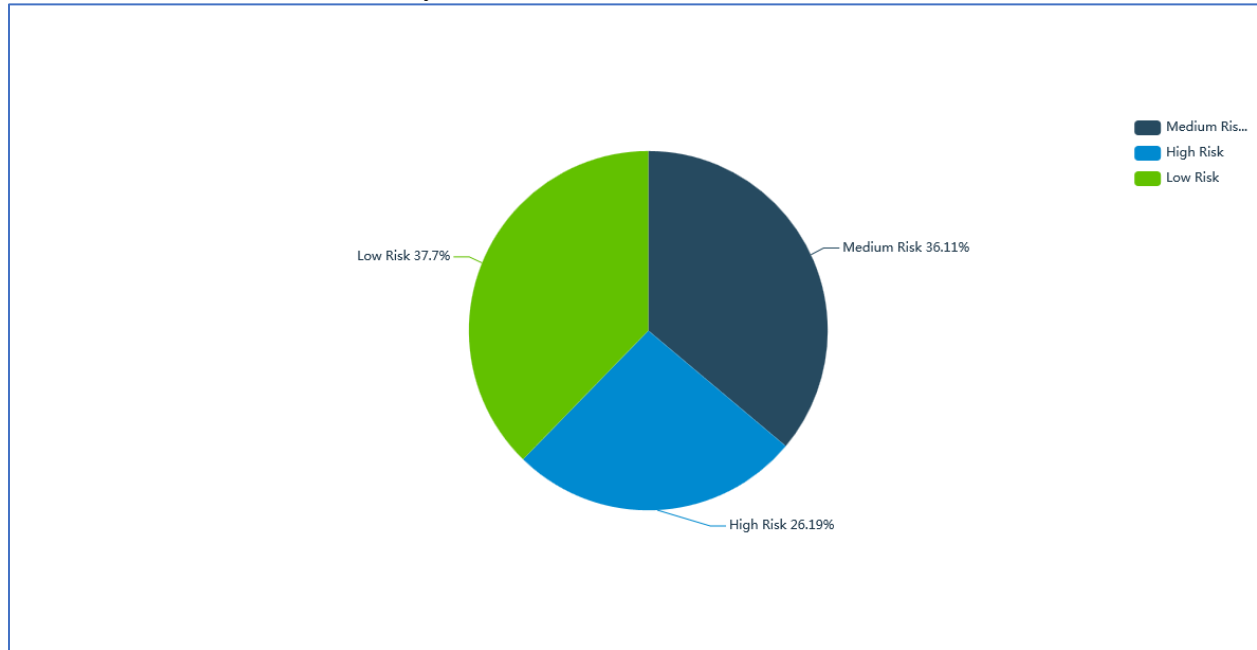
38. Click on the **Visualization** tab.



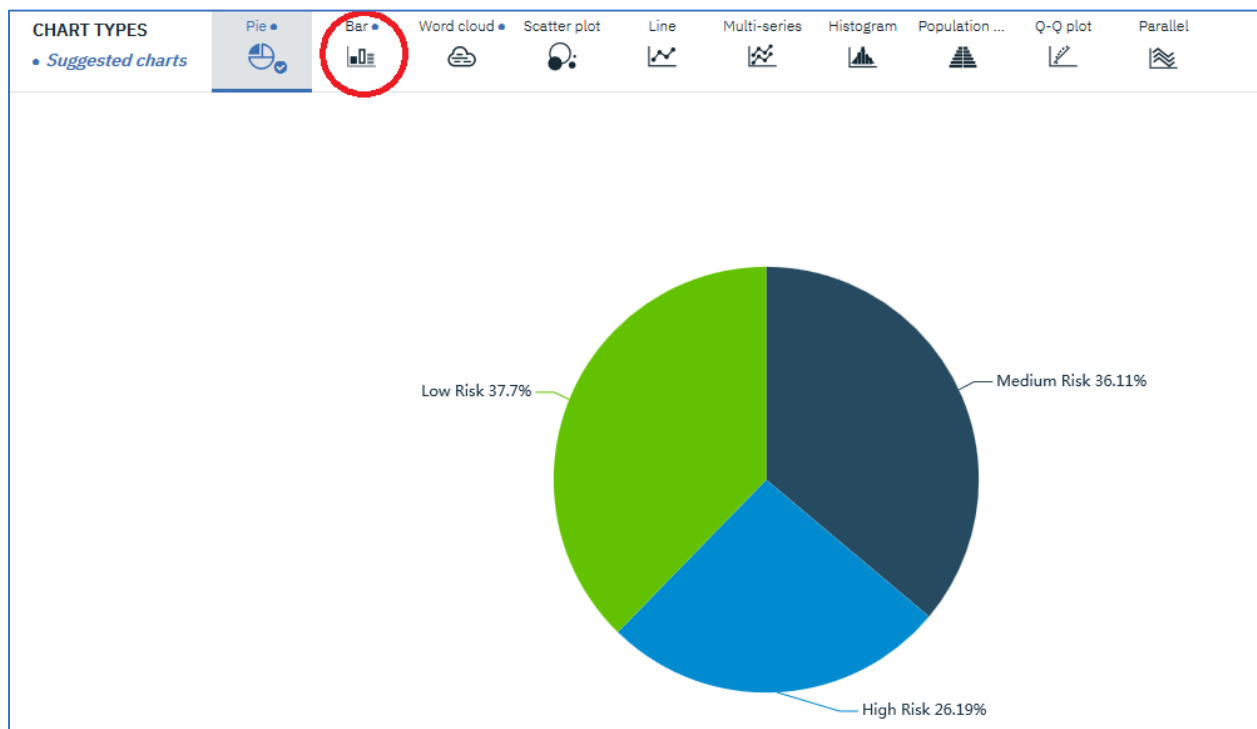
39. Click on **VETTING_LEVEL_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



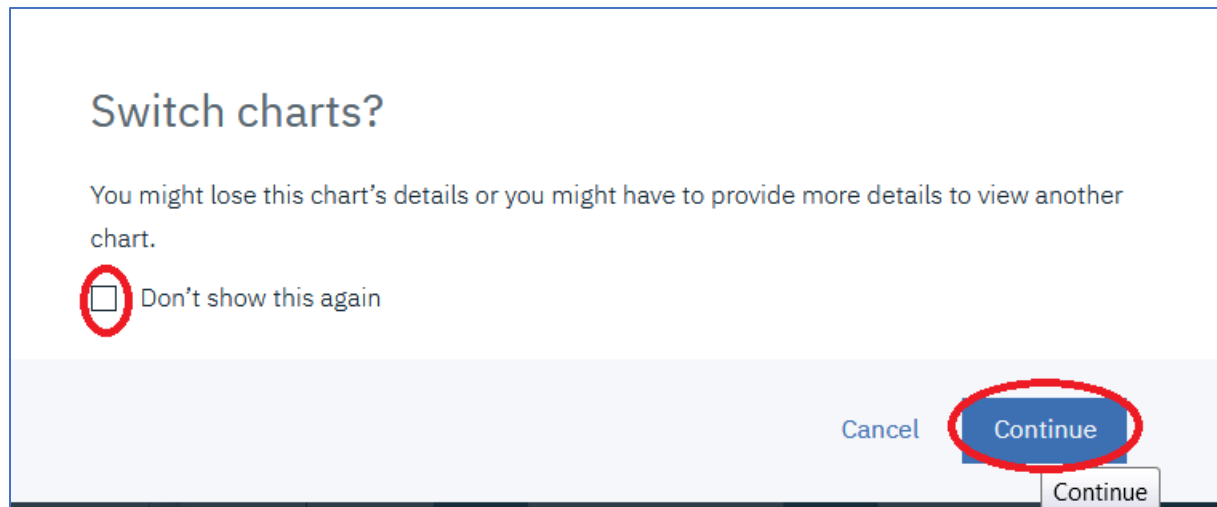
40. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



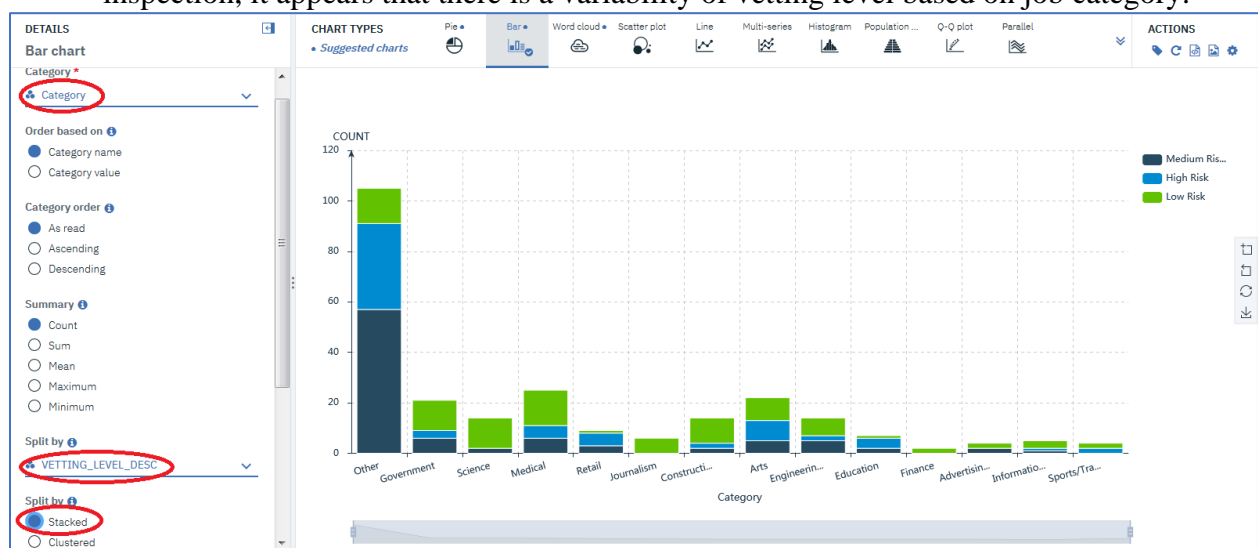
41. We can visualize the breakdown of travel records by job category and vetting level. Click on **Bar**.



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



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



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




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

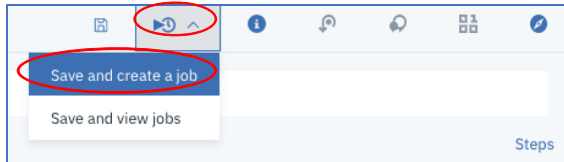


46. Please feel free to experiment with other visualizations.

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

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

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



2. Enter a **Job Name** for the job. 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. We are just going to do a one-time run. Click **Create and Run**.

A screenshot of the 'Create a job' form. The 'Job Name' field contains 'FHT Data Refinery' and is circled in red. The 'Associated asset' section shows 'female_human_trafficking_flow' with '11 Steps'. The 'Select runtime' dropdown is set to 'Default Data Refinery XS'. On the right, the 'INPUT' is 'FEMALE_HUMAN_TRAFFICKING' and the 'OUTPUT' is 'female_human_trafficking_shaped.csv'. At the bottom right, the 'Create and Run' button is circled in red. Other buttons include 'Cancel' and 'Create'.

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

A screenshot of the job run status page. The title is 'FHT Data Refinery'. The 'Associated Asset' section shows 'female_human_trafficking_flow' with '11 Steps'. Below this, there are sections for 'Scheduled to run' and 'Environment definition'. The main section is 'Runs', which contains a table with columns: 'Start Time', 'Status', 'Duration', 'Started By', and 'Action'. The first row shows a run starting on 'Jul 21, 2019, 2:29:11 PM' with a status of 'Running', which is circled in red. The 'Started By' is 'Aaron Doe'.

Start Time	Status	Duration	Started By	Action
Jul 21, 2019, 2:29:11 PM	Running	---	Aaron Doe	

FHT Data Refinery
No description

Associated Asset
DATA REFINERY FLOW
female_human_trafficking_flow 11 Steps

Scheduled to run
No Schedule Created [Edit](#)

Environment definition
Default Data Refinery XS [Edit](#)

INPUT
FEMALE_HUMAN_TRAFFICKING
Location
/DASH100288

OUTPUT
female_human_trafficking_shaped.csv [CSV](#)

Runs





Start Time ▲	Status	Duration	Started By	Action
Jul 21, 2019, 2:29:11 PM	Completed	4 minutes 33 seconds	Aaron Doe	

4. 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

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

▼ **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	Aaron Doe	21 Jul 2019, 2:34:19 pm	
<input type="checkbox"/>	 Categories	Data Asset	Aaron Doe	21 Jul 2019, 12:40:13 pm	
<input type="checkbox"/>	 Occupation	Data Asset	Aaron Doe	21 Jul 2019, 12:40:13 pm	
<input type="checkbox"/>	 female_human_trafficking	Data Asset	Aaron Doe	21 Jul 2019, 12:39:36 pm	
<input type="checkbox"/>	 trafficking	Connection	Aaron Doe	21 Jul 2019, 12:39:32 pm	

6. 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										
<div>Refine</div>										
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 Mille	1997-03-11	Accounting techr	Ghana	EG,AR,PA,DZ,RU,RU,AL	EG	AR	PA	7	CA
30	Sadie Archer	1997-12-27	Tax adviser	Ghana	MA,DO,QA,TH,CY	MA	DO	QA	5	CO
20	Paula Jimenez	2000-01-17	Electronics engir	Ghana	OM	OM			1	NM
30	Tammy Karen Hu	1976-03-14	Engineer, contro	Brazil	NL,KH,RU,CH,GB	NL	KH	RU	5	IL
20	Dy Rivera	1974-11-08	Engineer, manufi	Ghana	AE,SN	AE	SN		2	IN
30	Rhonnie Lindie S	1977-04-02	Engineer, mining	Brazil	RU,SI,JM,DO	RU	SI	JM	4	CA
30	Melinda Kimm Hi	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 Wilt	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.