

# 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.

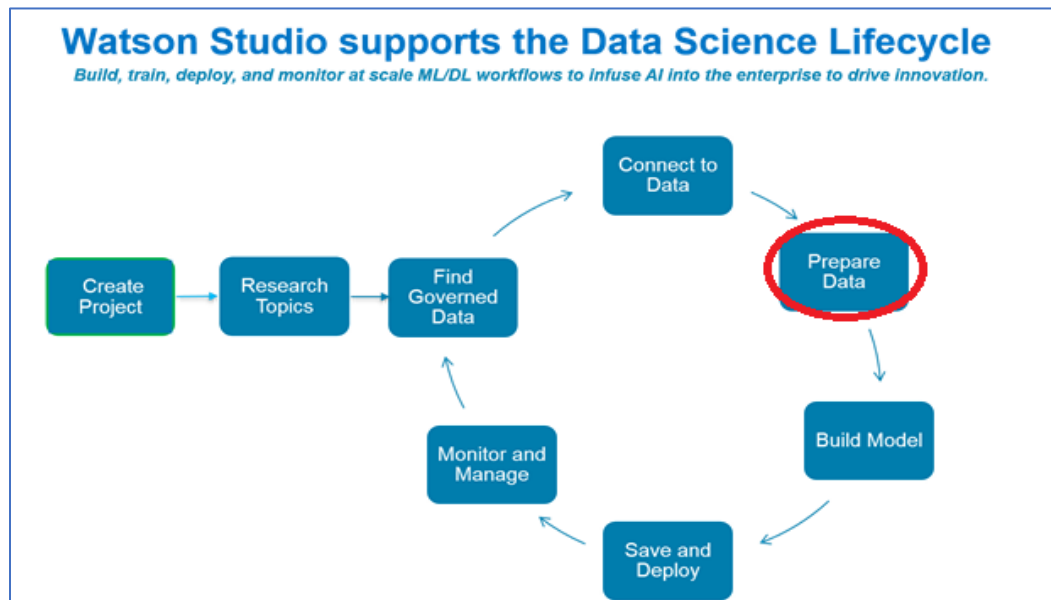


Figure 1- End to End Flow

## 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

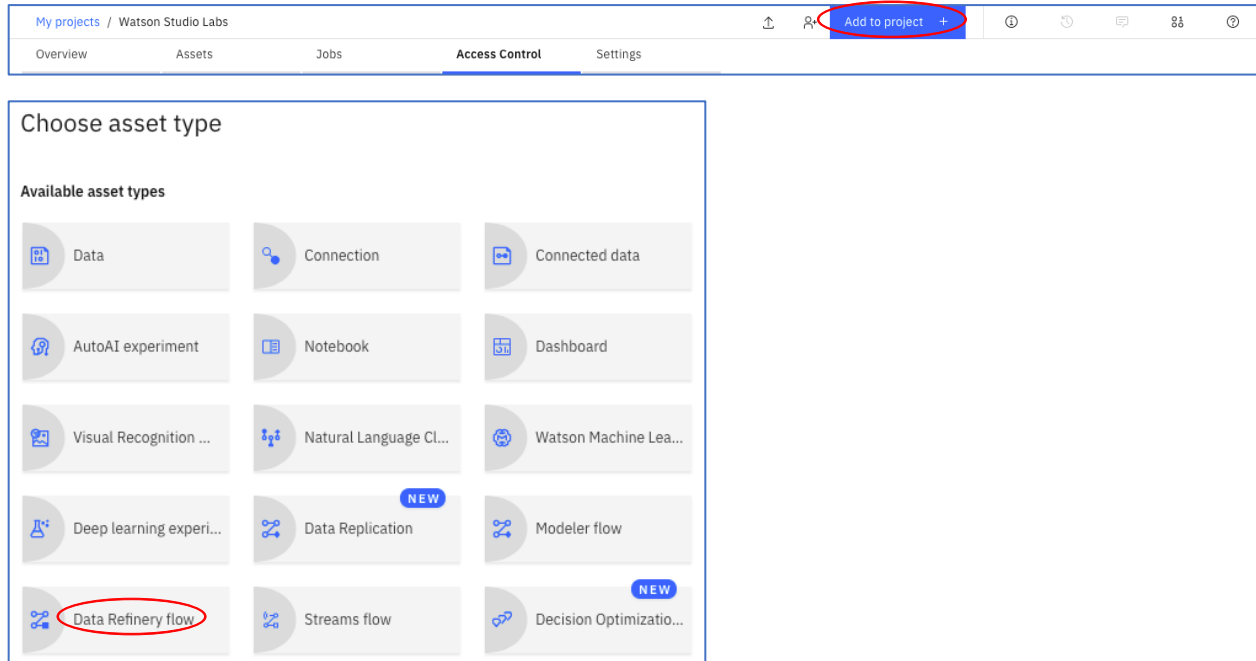
The data sets used for this lab consist of simulated travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING\_LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING\_LEVEL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

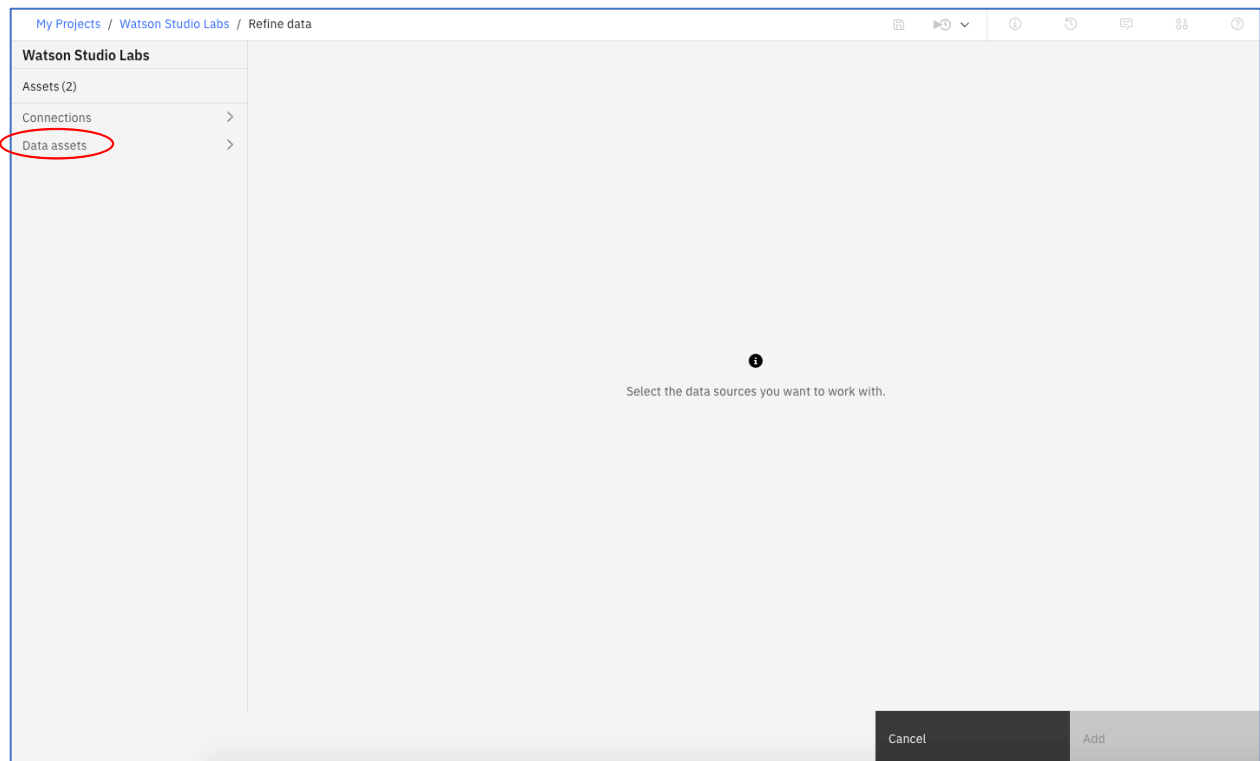
Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

## Create a new Data Flow

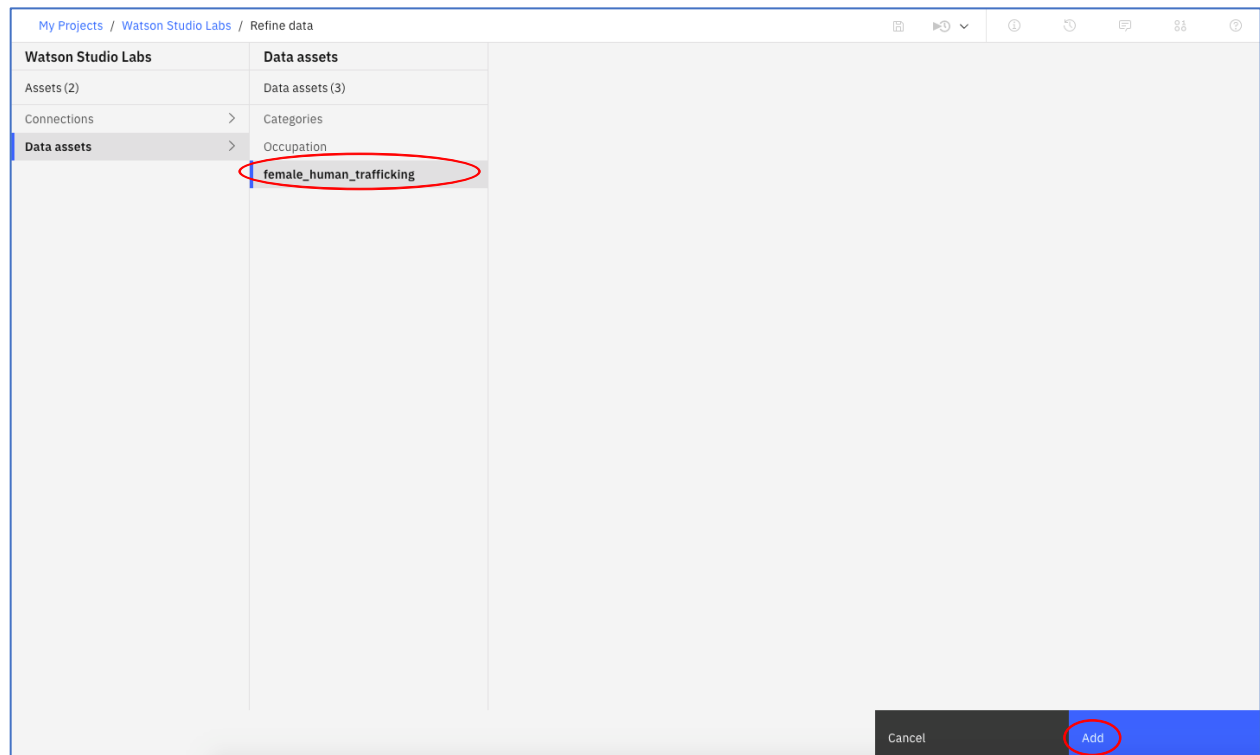
1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow**.



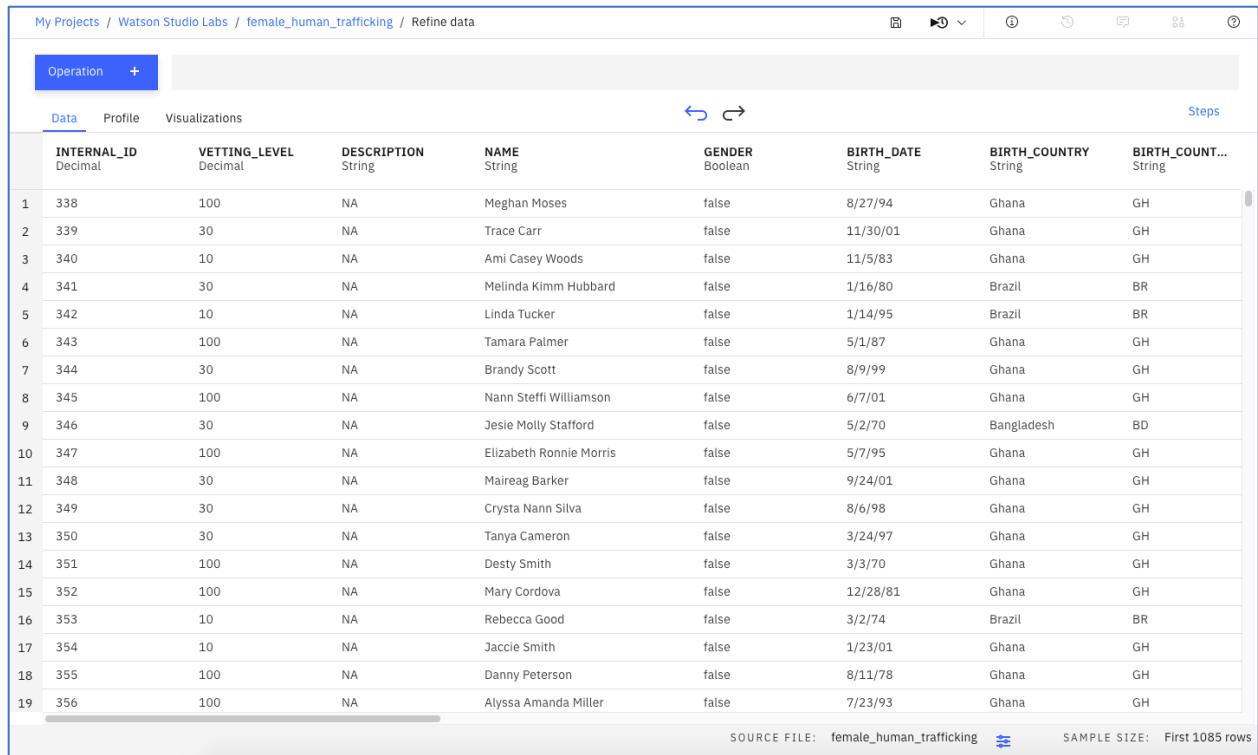
## 2. Click on **Data Assets**.



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



#### 4. The data set will be displayed.




The screenshot shows the Watson Studio interface with the 'Data' tab selected. The table displays 19 rows of data with the following columns: INTERNAL\_ID (Decimal), VETTING\_LEVEL (Decimal), DESCRIPTION (String), NAME (String), GENDER (Boolean), BIRTH\_DATE (String), BIRTH\_COUNTRY (String), and BIRTH\_COUNT... (String). The data includes names like Meghan Moses, Trace Carr, and Ami Casey Woods, along with their birth dates and countries. The bottom of the interface shows 'SOURCE FILE: female\_human\_trafficking' and 'SAMPLE SIZE: First 1085 rows'.


	INTERNAL_ID Decimal	VETTING_LEVEL Decimal	DESCRIPTION String	NAME String	GENDER Boolean	BIRTH_DATE String	BIRTH_COUNTRY String	BIRTH_COUNT... String
1	338	100	NA	Meghan Moses	false	8/27/94	Ghana	GH
2	339	30	NA	Trace Carr	false	11/30/01	Ghana	GH
3	340	10	NA	Ami Casey Woods	false	11/5/83	Ghana	GH
4	341	30	NA	Melinda Kimm Hubbard	false	1/16/80	Brazil	BR
5	342	10	NA	Linda Tucker	false	1/14/95	Brazil	BR
6	343	100	NA	Tamara Palmer	false	5/1/87	Ghana	GH
7	344	30	NA	Brandy Scott	false	8/9/99	Ghana	GH
8	345	100	NA	Nann Steffi Williamson	false	6/7/01	Ghana	GH
9	346	30	NA	Jesie Molly Stafford	false	5/2/70	Bangladesh	BD
10	347	100	NA	Elizabeth Ronnie Morris	false	5/7/95	Ghana	GH
11	348	30	NA	Maireag Barker	false	9/24/01	Ghana	GH
12	349	30	NA	Crysta Nann Silva	false	8/6/98	Ghana	GH
13	350	30	NA	Tanya Cameron	false	3/24/97	Ghana	GH
14	351	100	NA	Desty Smith	false	3/3/70	Ghana	GH
15	352	100	NA	Mary Cordova	false	12/28/81	Ghana	GH
16	353	10	NA	Rebecca Good	false	3/2/74	Brazil	BR
17	354	10	NA	Jaccie Smith	false	1/23/01	Ghana	GH
18	355	100	NA	Danny Peterson	false	8/11/78	Ghana	GH
19	356	100	NA	Alyssa Amanda Miller	false	7/23/93	Ghana	GH

## 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.



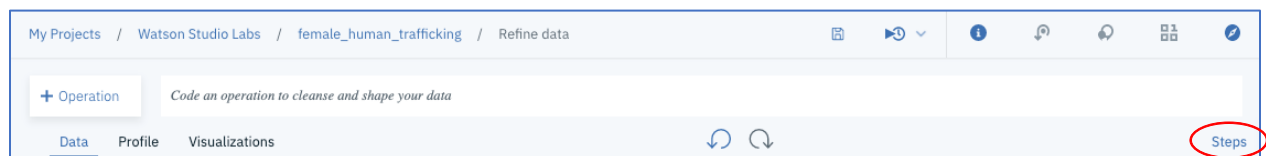
1. Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the **VETTING\_LEVEL** column by hovering over VETTING\_LEVEL, clicking on the vertical ellipse , clicking on **CONVERT COLUMN**, and clicking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

VETTING_LEVEL	DESCRIPTION	NAME
Decimal	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...	✓ Decimal
100	View All	Integer
30	...	String
100	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
Decimal	String	String	Str
308561300	Remove	GH	QA
987374355	Remove duplicates	GH	QA
426221095	Remove empty rows	GH	ME
869842380	Sort ascending	BR	IL,
473389048	Sort descending	BR	ES
217560040	Substitute	GH	HR
942939007	CONVERT COLUMN...	✓ Decimal	DM
768902471	View All	Integer	P,
730613975	...	String	P,
798632110	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
4	157-58-6078	869842380	Brazil	BR
5	168-42-4190	473389048	Brazil	BR
6	387-51-9501	217560040	Ghana	GH
7	896-58-3773	942939007	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.

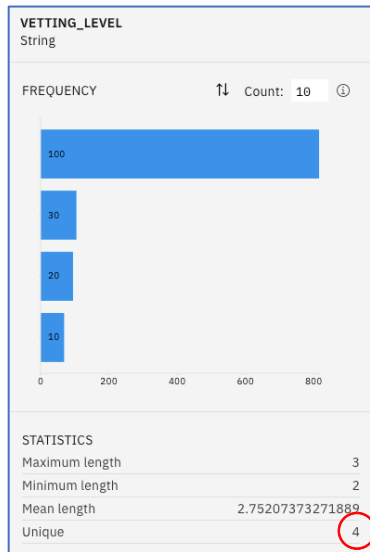
Operation +

Data Profile Visualizations

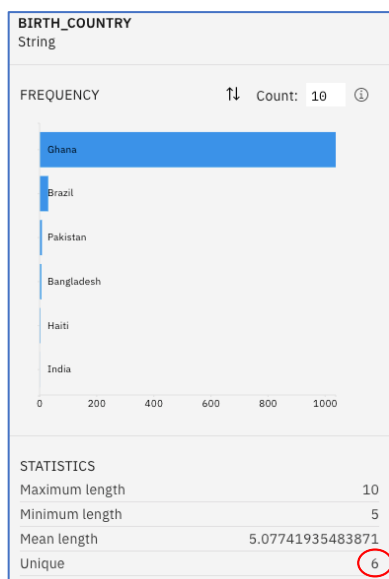


INTERNAL_ID Decimal	VETTING_LEVEL String	DESCRIPTION String																										
<p>FREQUENCY</p>	<p>FREQUENCY</p>	<p>FREQUENCY</p> <p>↑↓ Count: 10 ⓘ</p>																										
<p>STATISTICS</p> <table><tr><td>Interquartile Range</td><td>542</td></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>1085</td></tr><tr><td>Median</td><td>543</td></tr><tr><td>Standard Deviation</td><td>313.356825360483</td></tr></table>	Interquartile Range	542	Minimum	1	Maximum	1085	Median	543	Standard Deviation	313.356825360483	<p>STATISTICS</p> <table><tr><td>Maximum length</td><td>3</td></tr><tr><td>Minimum length</td><td>2</td></tr><tr><td>Mean length</td><td>2.75207373271889</td></tr><tr><td>Unique</td><td>4</td></tr></table>	Maximum length	3	Minimum length	2	Mean length	2.75207373271889	Unique	4	<p>STATISTICS</p> <table><tr><td>Maximum length</td><td>2</td></tr><tr><td>Minimum length</td><td>2</td></tr><tr><td>Mean length</td><td>2</td></tr><tr><td>Unique</td><td>1</td></tr></table>	Maximum length	2	Minimum length	2	Mean length	2	Unique	1
Interquartile Range	542																											
Minimum	1																											
Maximum	1085																											
Median	543																											
Standard Deviation	313.356825360483																											
Maximum length	3																											
Minimum length	2																											
Mean length	2.75207373271889																											
Unique	4																											
Maximum length	2																											
Minimum length	2																											
Mean length	2																											
Unique	1																											

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.

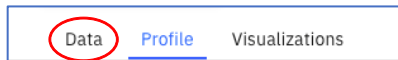


8. Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has about 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH\_COUNTRY, and PASSPORT\_COUNTRY shows only 6 unique countries. The COUNTRIES\_VISITED\_COUNT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

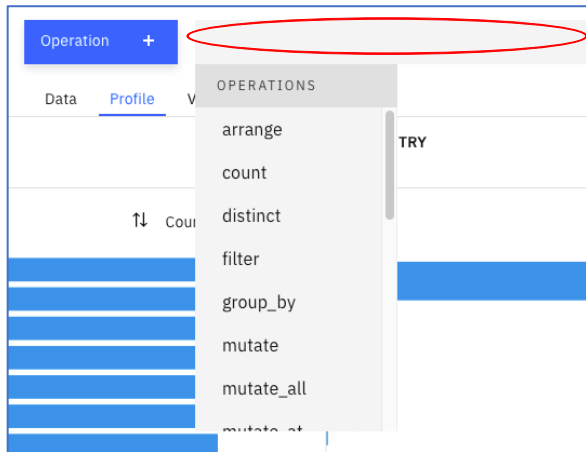




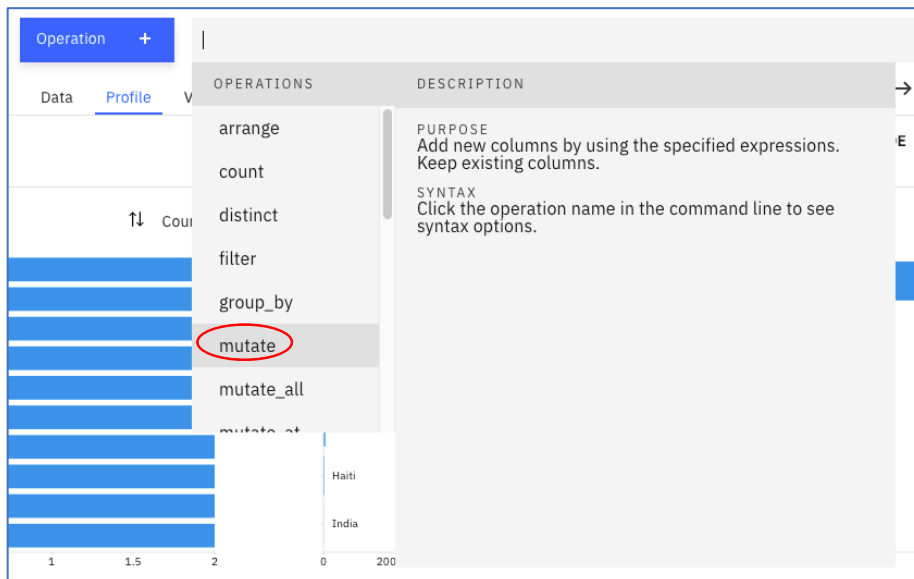
9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.



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.

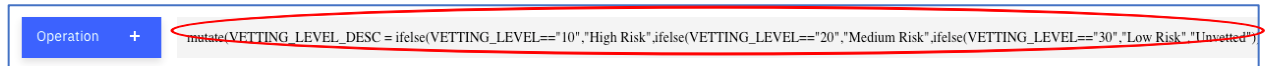


11. Hover the mouse over **mutate**. A description of the mutate function is provided.



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 +**.

Operation +

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

Operation x

Q Search operations

CLEANSE ^

Convert column value to missing

Extract date or time value

Remove duplicates

Remove empty rows

Replace missing values

Replace substring

ORGANIZE ^

Aggregate

Concatenate

Conditional replace

Join >

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

Operation x

< Join

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

Left join v

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

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

Source

female\_human\_trafficking

\*Suffix

\_X

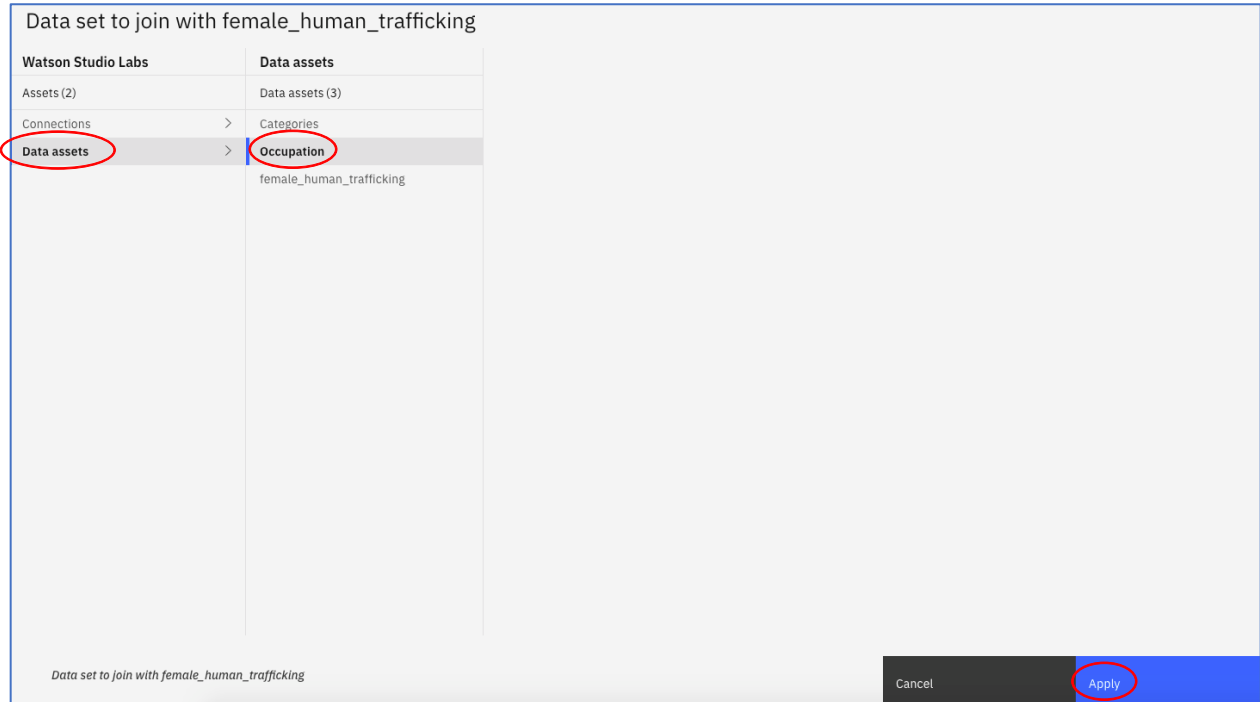
Data set to join

+ Add data set

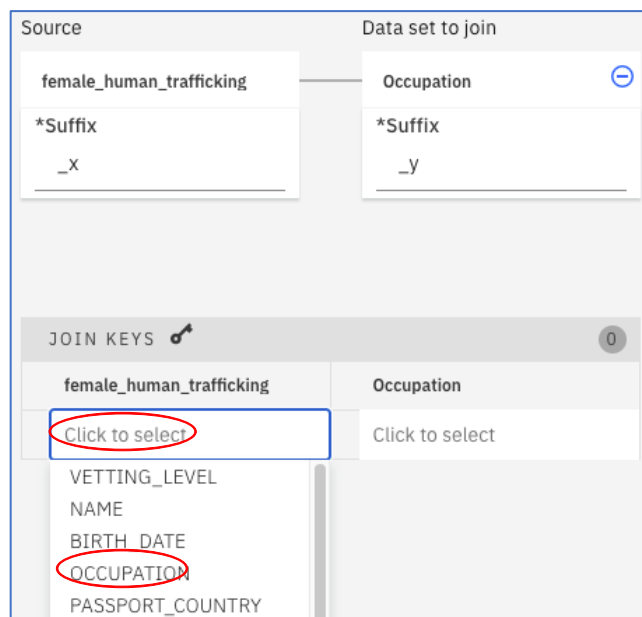
\*Suffix

\_Y

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 columns 'Code' (String) and 'Category' (String) is displayed. The 'Code' column has values like 11, 7, 15, 2, 15, 5, 15, 1, 6, 6, 8, 15, 6, 13, 6, 13, 8, 2, 13. The 'Category' column has values like Construction, Science, Other, Engineering, Other, Government, Other, Sports/Travel, Medical, Medical, Arts, Other, Medical, Education, Medical, Education, Arts, Engineering, Education. On the right, the '7 Steps' sidebar is visible, showing options to 'Convert column type' and 'Custom code'. The 'Custom code' section contains two code snippets: one for mutating 'VETTING\_LEVEL\_DESC' based on 'VETTING\_LEVEL' and another for selecting columns from a dataset.

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 **Operation** + then click on **Split column**.


The screenshot shows the 'Operation' menu in the Data Refinery interface. The menu is titled 'Operation' and has a search bar. Below the search bar, there are two main sections: 'CLEANSE' and 'ORGANIZE'. The 'CLEANSE' section includes options like 'Convert column value to missing', 'Extract date or time value', 'Remove duplicates', 'Remove empty rows', 'Replace missing values', and 'Replace substring'. The 'ORGANIZE' section includes options like 'Aggregate', 'Concatenate', 'Conditional replace', 'Join', 'Sample', 'Split column', and 'Union'. The 'Split column' option is highlighted with a red circle.

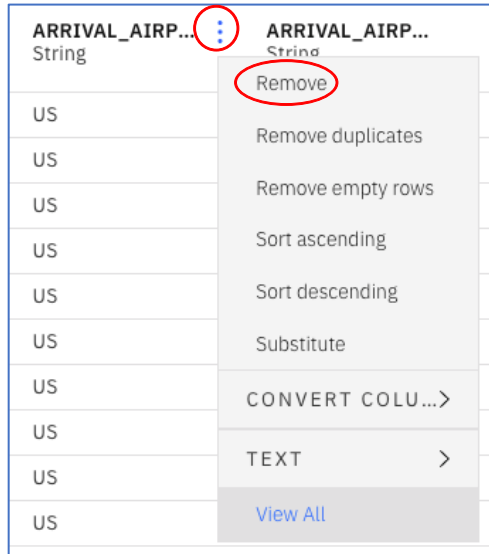
25. Click on **Select column**, select **ARRIVAL\_AIRPORT\_REGION**, and then click on **Next**.

The screenshot shows a dialog box titled 'Operation' with a sub-header 'Column selection'. Below the header, it says 'To begin, select a column.' A dropdown menu is open, showing 'ARRIVAL\_AIRPORT\_REGION' as the selected option. At the bottom of the dialog, there are two buttons: 'Cancel' and 'Next'.

26. Click on **TEXT**, click on **Hyphen(-)** in the dropdown, enter **ARRIVAL\_AIRPORT\_COUNTRY, ARRIVAL\_AIRPORT\_STATE** as the names of the new columns, uncheck **keep original column**, and click on **Apply**.

The screenshot shows a dialog box titled 'Operation' with a sub-header 'Split column'. Below the header, it says 'Change column selection'. There are four tabs: 'DEFAULT', 'TEXT', 'PATTERN', and 'POSITION'. The 'TEXT' tab is selected. A dropdown menu is open, showing 'Hyphen (-)' as the selected option. Below the dropdown, the text 'ARRIVAL\_AIRPORT\_COUNTRY, ARRIVAL\_AIRPORT\_STATE' is entered. There is a checkbox labeled 'keep original column' which is unchecked. At the bottom of the dialog, there are two buttons: 'Cancel' and 'Apply'.

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



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

28. 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.

29. 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

< Split column

Change column selection

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

30. The results are shown below.

COUNTRIES_VISITED String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VI... Decimal	ARRIVAL_AIRP... String
QA	QA			1	NY
QA	QA			1	WA
ME,EE,KY,DZ,CZ,ID,NL,QA,BS,CK	ME	EE	KY	10	IN
IL,VN,UZ	IL	VN	UZ	3	AR
ES,JO,LT,CL,QA,PA	ES	JO	LT	6	OH
HR,BS,BG,AT,DK,AL,AL,OM,TN,LU,SI,IN	HR	BS	BG	12	IL
OM,CK,BH,CK,TW,IQ,TN	OM	CK	BH	7	FL
JP,RU,CO,CU,TR,TR	JP	RU	CO	6	NV
JP,SN,SK,OM	JP	SN	SK	4	AZ
RU	RU			1	NC
RU	RU			1	MS
AE	AE			1	NY
CH,AE,LK	CH	AE	LK	3	SC
TR,ES,KW,SG,RU,FI,KZ,BN,JM,PT	TR	ES	KW	10	CA
RU,DZ,KR,SN,UA,TR,MT,RS,PK	RU	DZ	KR	9	AL

31. Let's use visualization to get a better understanding of the data. First, we will remove the unvetted records. Hover over the VETTING\_LEVEL column, click on the vertical ellipse , click on **View All**.

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

Remove

Remove duplicates

Remove empty rows

Sort ascending

Sort descending

Substitute

CONVERT COLU...>

TEXT >

View All

32. Click on **Filter**.

Operation
Search operations
FREQUENTLY USED
Calculate
Convert column type
Filter
Math
Remove

33. 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

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

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

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

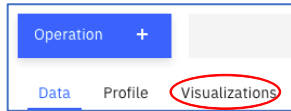
My Projects / Watson Studio Labs / female\_human\_trafficking / Refine data

Operation +

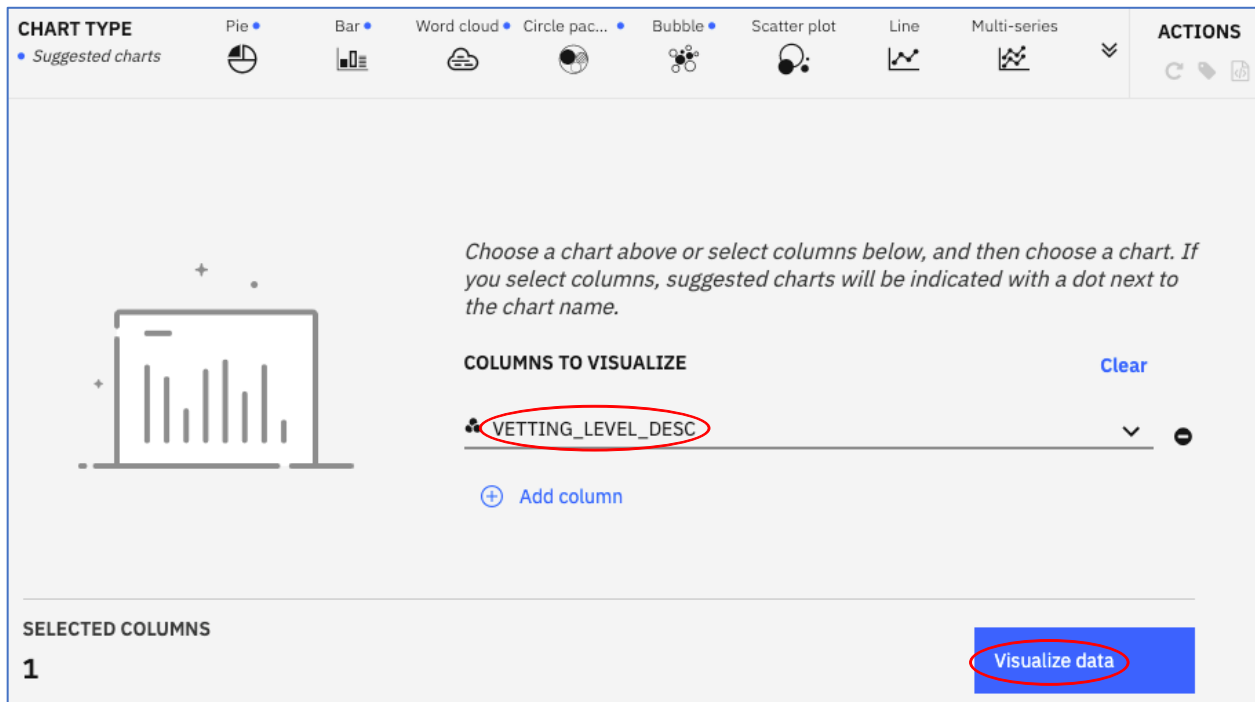
Data Profile Visualizations

Save

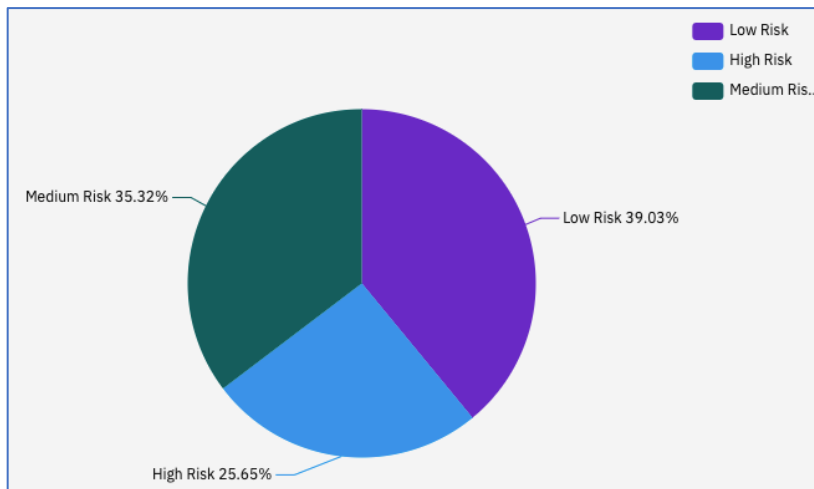
36. Click on the **Visualization** tab.



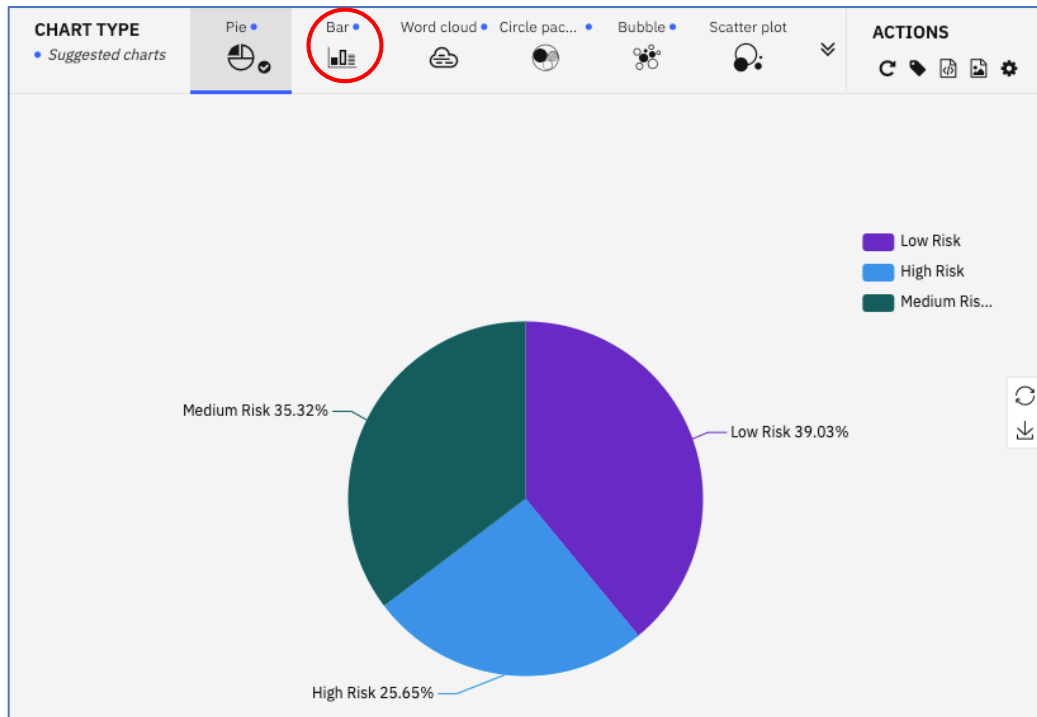
37. Click on **VETTING\_LEVEL\_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



38. 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.



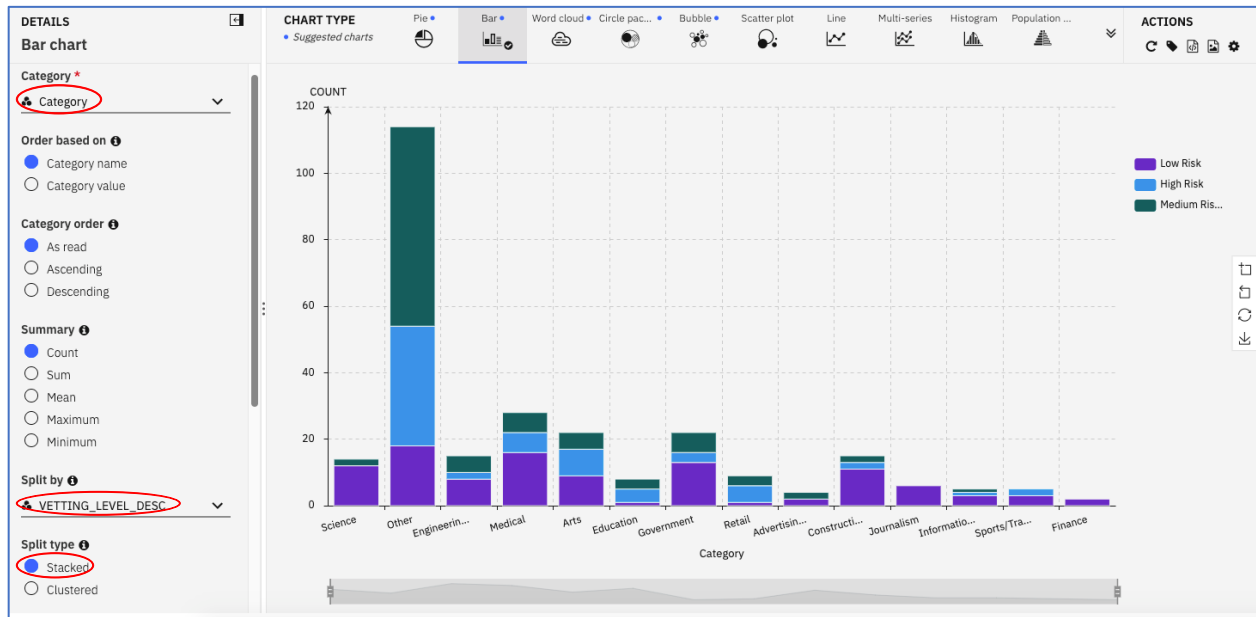
39. We can visualize the breakdown of travel records by job category and vetting level.  
Click on **Bar**.



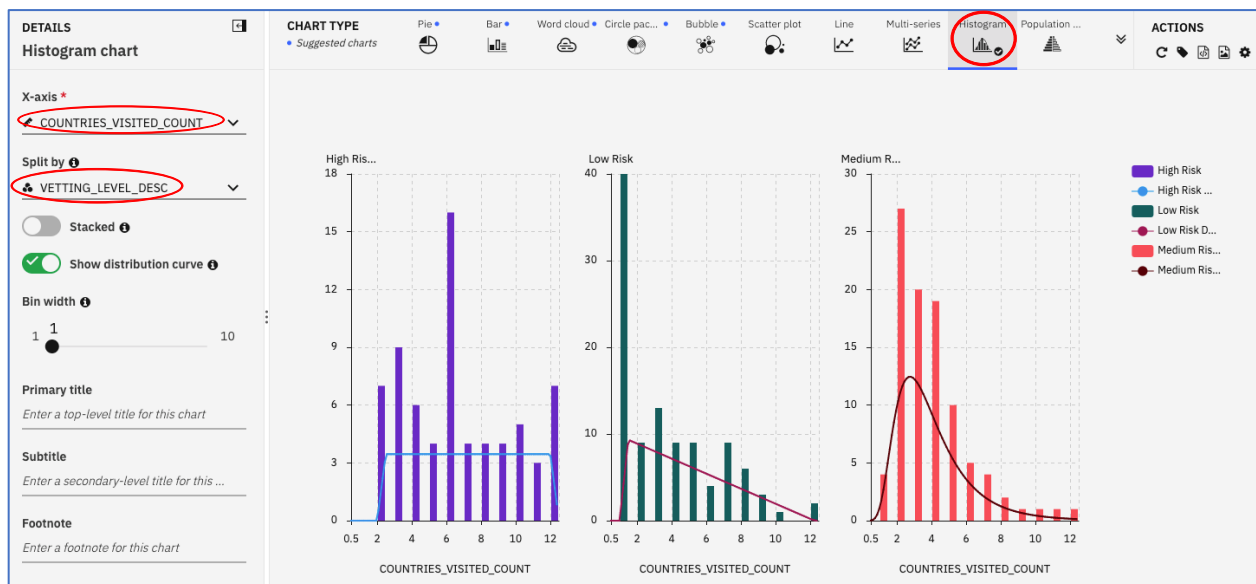
40. Click on **Don't show this again**. Click on **Continue**

The screenshot shows a dialog box titled 'Switch charts?'. The text inside says: 'You might lose this chart's details or you might have to provide more details to view another chart.' Below this text, there is a checkbox labeled 'Don't show this again', which is checked and circled in red. At the bottom of the dialog, there are two buttons: 'Cancel' and 'Continue'. The 'Continue' button is circled in red.

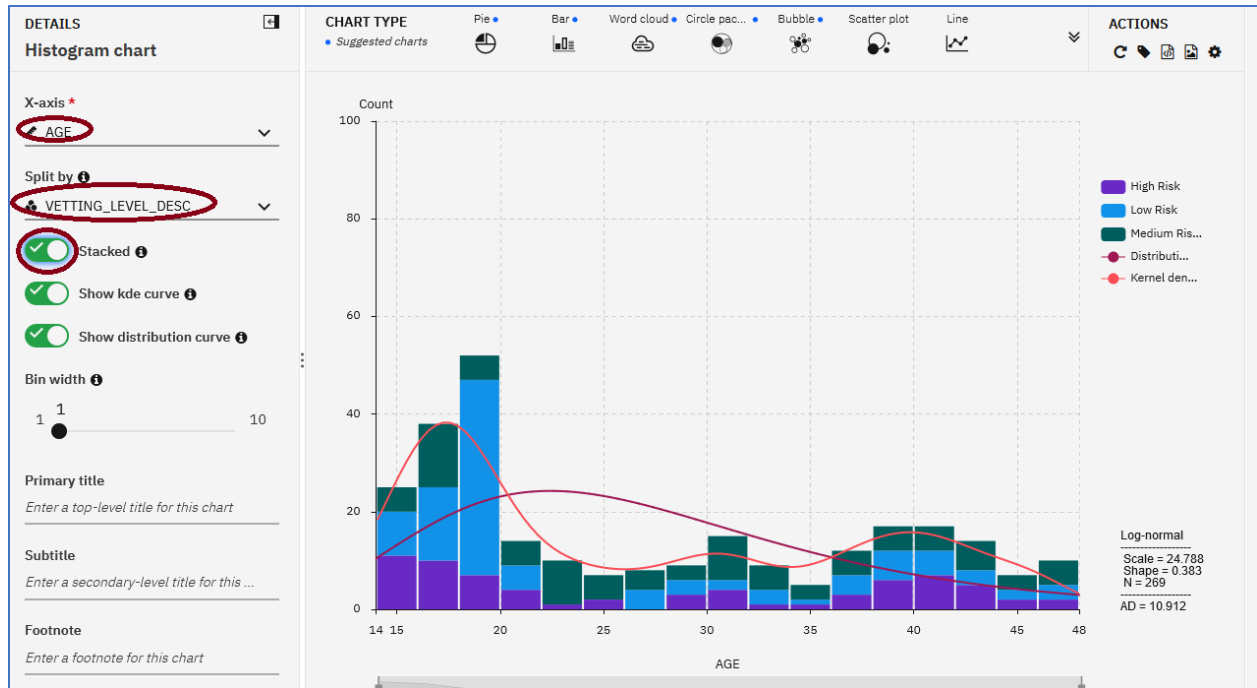
41. Click on **Category** for **Category**, click on VETTING\_LEVEL\_DESC for **Split by**, click on **Stacked** for **Split type**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



42. 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.





43. Let's examine if age makes a difference. Click on **AGE** for **X-axis**. **Split by** remains **VETTING\_LEVEL\_DESC**, enable **Stacked**. It appears that younger travelers have a lower risk of being trafficked.

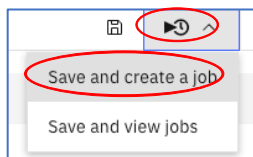


44. Please feel free to experiment with other visualizations.

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

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

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 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 **Create and Run**.

## Create a job

Create a job to specify how and when to run an analytical asset. Select the analytic asset and set up a schedule or run the job immediately.

Job Name

INPUT

female\_human\_trafficking

OUTPUT

female\_human\_trafficking\_shaped...

Description (Optional)

Description of job

Associated Asset

DATA REFINERY FLOW

female\_human\_trafficking\_flow 12 Steps [Edit](#)

Select runtime

Default Data Refinery XS

☐ Schedule off

Cancel Create **Create and Run**

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

My Projects / Watson Studio Labs / FHT Data Refinery

## FHT Data Refinery

No description

Scheduled to run

No Schedule Created

[Edit](#)

Environment definition

Default Data Refinery XS

[Edit](#)

Associated Asset

DATA REFINERY FLOW

female\_human\_trafficking\_flow 12 Steps

INPUT

female\_human\_trafficking

OUTPUT

female\_human\_trafficking\_shaped.csv

Runs (1)

Start Time	Status	Duration	Started By	Action
Aug 03, 2020 12:06:28 AM	Completed	53 seconds	FCTO Labs	:

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 assets selected.


<input type="checkbox"/>	Name	Type	Created by	Last modified	↓
<input type="checkbox"/>	CSV <b>female_human_trafficking_shaped.csv</b>	Data Asset	FCTO Labs	Aug 03, 2020, 12:07 AM	
<input type="checkbox"/>	 <b>Occupation</b>	Data Asset	FCTO Labs	Aug 02, 2020, 10:02 PM	
<input type="checkbox"/>	 <b>Categories</b>	Data Asset	FCTO Labs	Aug 02, 2020, 10:02 PM	
<input type="checkbox"/>	 <b>female_human_trafficking</b>	Data Asset	FCTO Labs	Aug 02, 2020, 10:01 PM	

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

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

Preview Profile Activities

Schema: 15 Columns  
Preview: 269 rows

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


VETTING_L...	NAME	BIRTH_D...	OCCUPAT...	PASSPORT_COU...	COUNTRIES_VIS...	COUNTRY1	COUNTRY2	COUNTRY3	COUNTRIES_VISITED_C...	ARRI
String	String	String	String	String	String	String	String	String	String	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
30.0	Danielle Ash But	6/14/99	Musician	Ghana	SE,NG,AE	SE	NG	AE	3.0	TX
30.0	Veronica Breann	5/26/01	Call centre mana	Ghana	RU	RU			1.0	NC
10.0	Darlene Kendra	9/1/78	Interpreter	Brazil	IE,OM	IE	OM		2.0	IL
30.0	Lisa Melissa Rus	4/5/90	Race relations of	Brazil	RU	RU			1.0	OH
20.0	Lyz Pearson	8/15/74	Secretary, compi	Ghana	UZ,PH	UZ	PH		2.0	OK
20.0	Kelli Jilly Parker	12/17/95	Personal assistar	Ghana	KE,IR,QA,PG,KW	KE	IR	QA	5.0	

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

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

Preview Profile Activities

Schema: 15 Columns  
Preview: 269 rows

Last refresh: 5 minutes ago  [Refine](#)

VETTING_L...	NAME	BIRTH_D...	OCCUPAT...	PASSPORT_COU...	COUNTRIES_VIS...	COUNTRY1	COUNTRY2	COUNTRY3	COUNTRIES_VISITED_C...	ARRIVAL_AIRPORT_S...
String	String	String	String	String	String	String	String	String	String	String

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

My projects / Watson Studio Labs

Overview Assets Environments **Jobs** Deployments Access Control Settings

## Jobs

A job is a way of running an asset like a notebook or flow. You can run a job immediately or on a schedule.

0 Active Jobs 1 Total Job

Which job are you looking for?

### Jobs

Job name	Associated asset	Last run	Started by	Created by
Data Refinery Job	Data Refinery flow	Finished Aug 04, 2020, 6:09 PM	Bertrand Doe Aug 04, 2020, 6:06 PM	Bertrand Doe

## 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.