Data Refinery Lab

Introduction

This lab will introduce Data Refinery. Data Refinery is a self-service data preparation tool for data scientists, data engineers, and business analysts. Data Refinery provides profiling, visualization, and a robust set of transforms to prepare data for analytics purposes. You will use the 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.

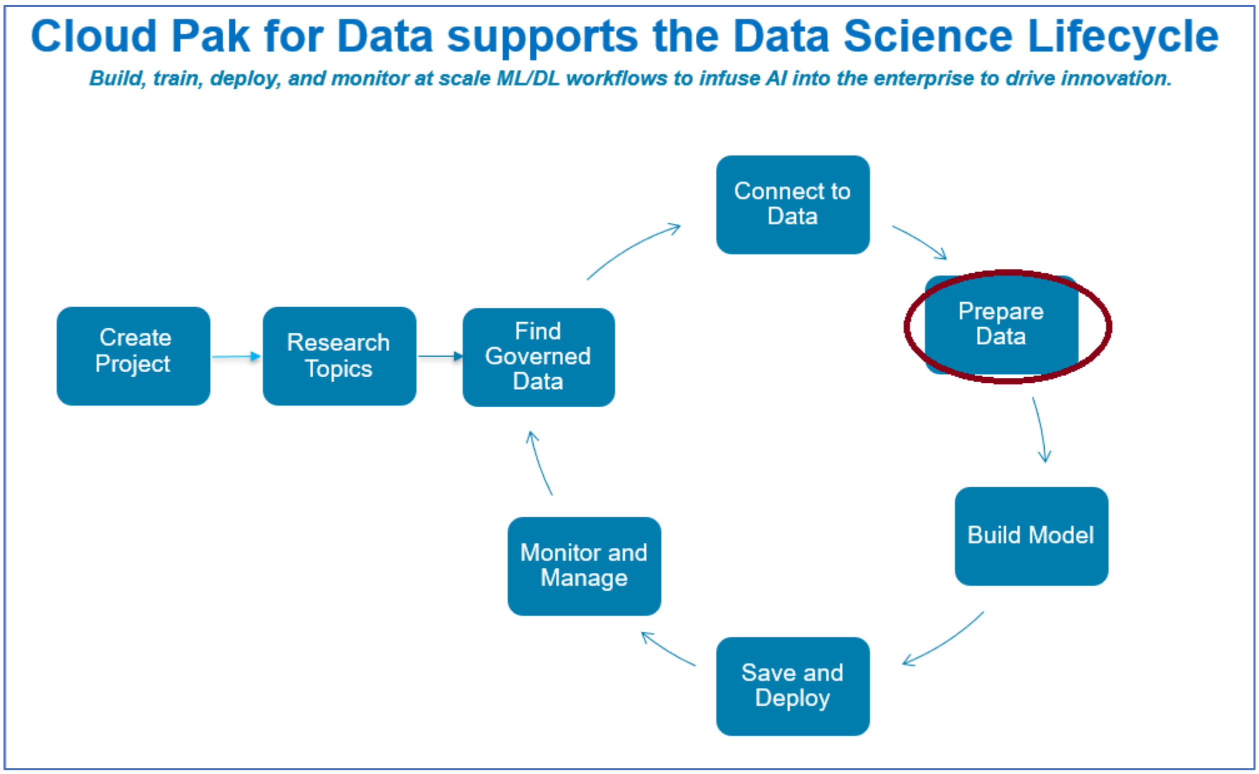


Figure 1- End to End Flow

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.

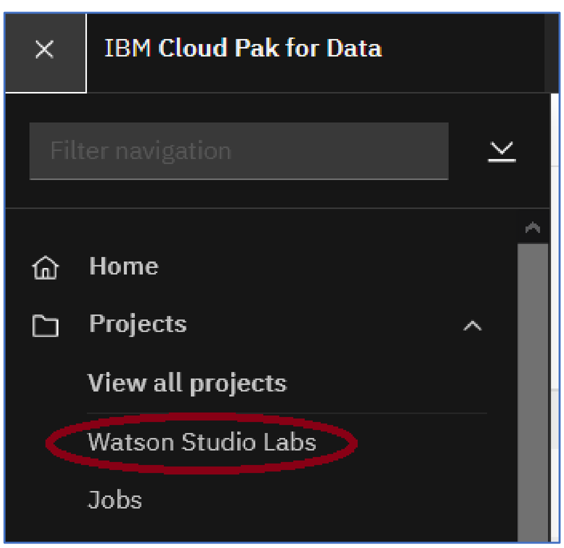
The OCCUPATION data included in the travel data is very granular. For modeling purposes, it was decided to categorize the OCCUPATION data. Two additional datasets are used for this purpose. The occupation.csv dataset maps the granular occupation data to a category code. The categories dataset maps a 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.

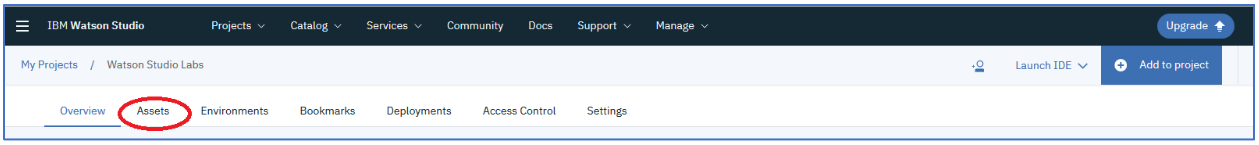
Create a new Data Flow

1. Click on the hamburger icon.

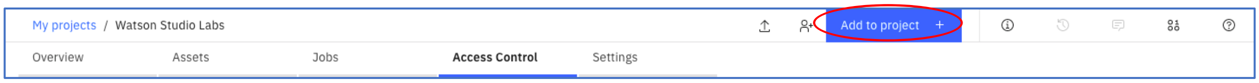
2. Click on **Watson Studio Labs** under **Projects**.



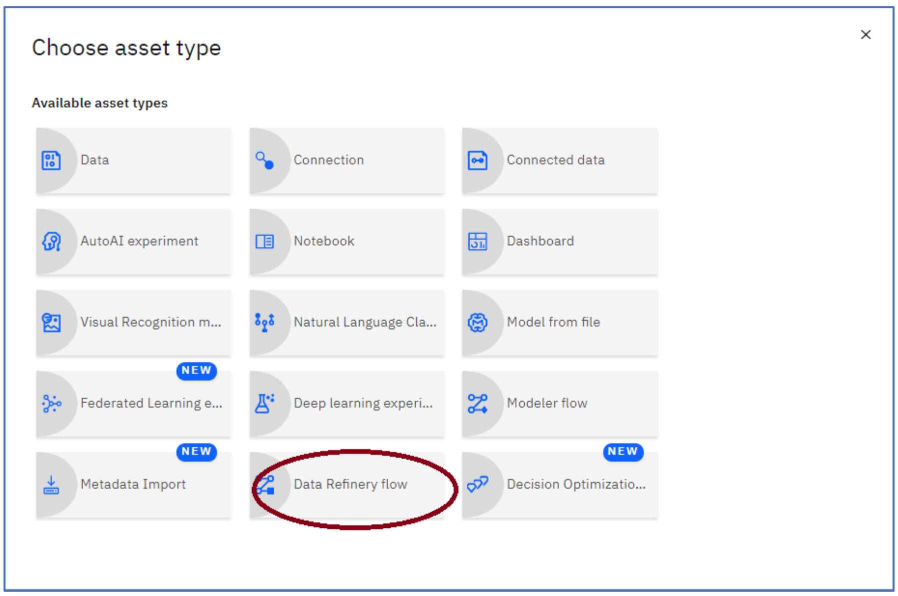
3. Click on the **Assets** tab.



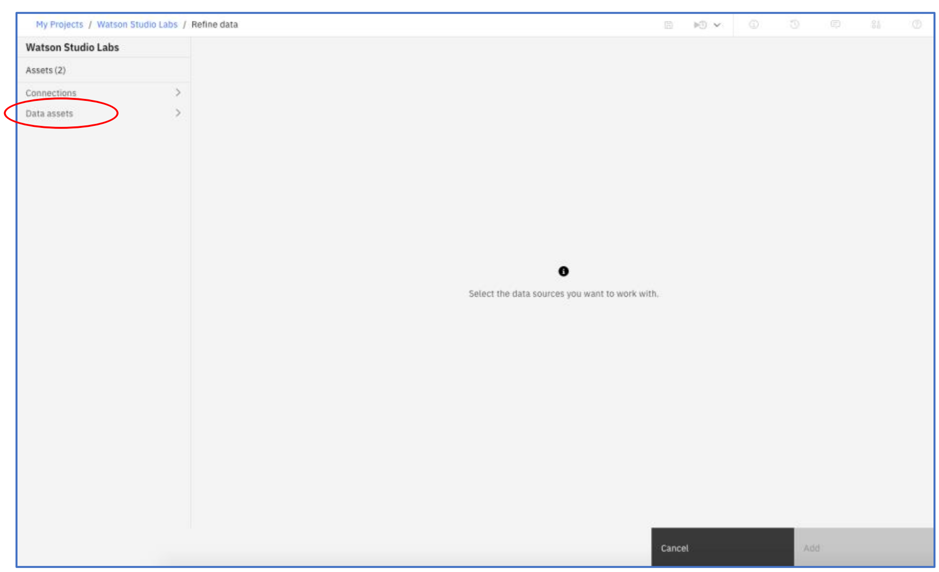
4. Add a Data Flow by clicking on **Add to project.**



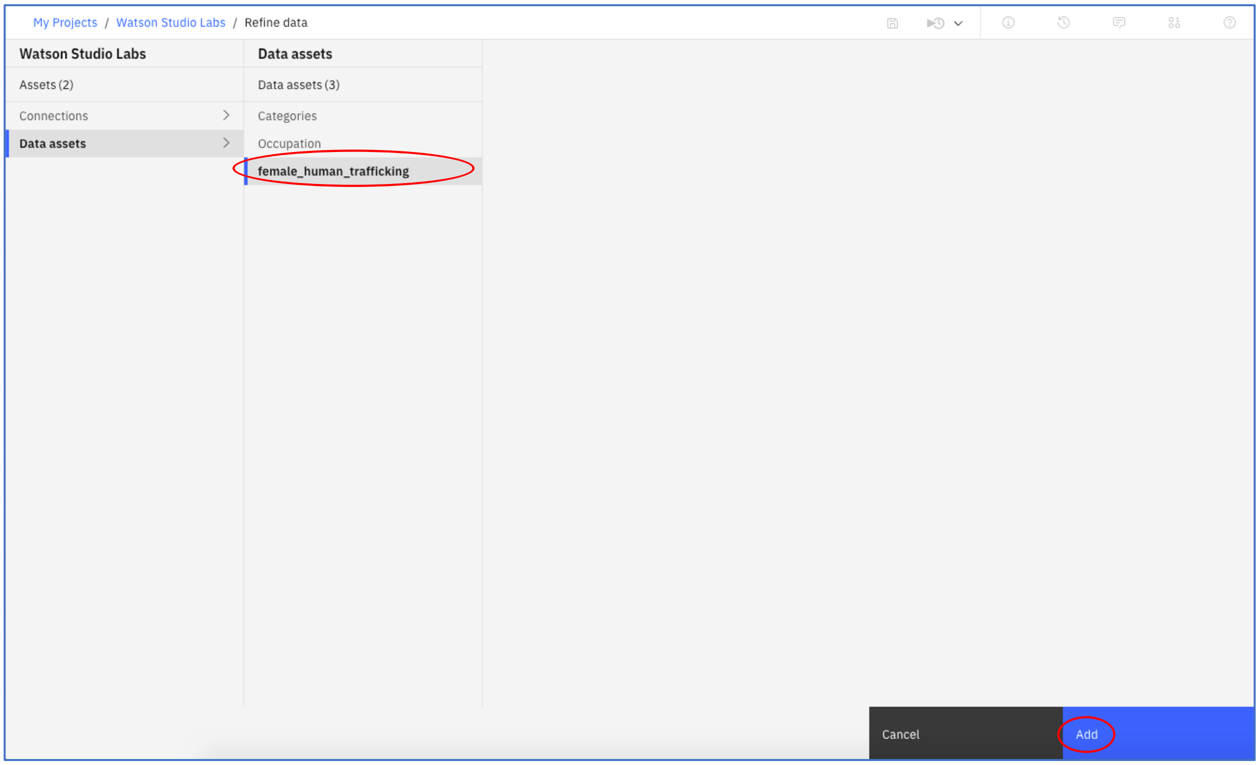
5. Click on **Data Refinery flow**



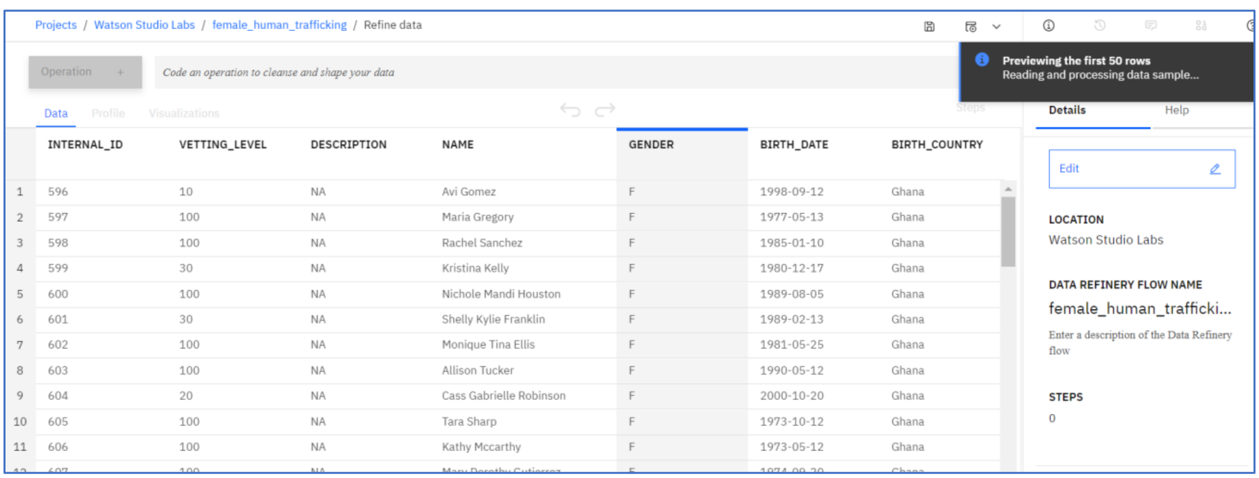
6. Click on **Data Assets**.



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



8. The data set will be displayed. Please wait until the Previewing is complete.

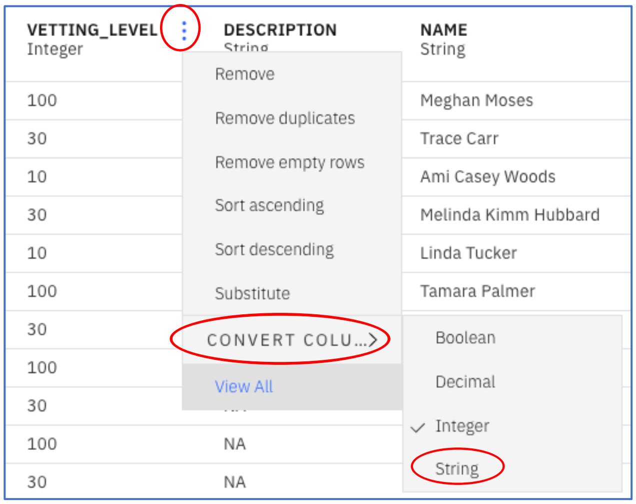


Prepare, Profile, Visualize

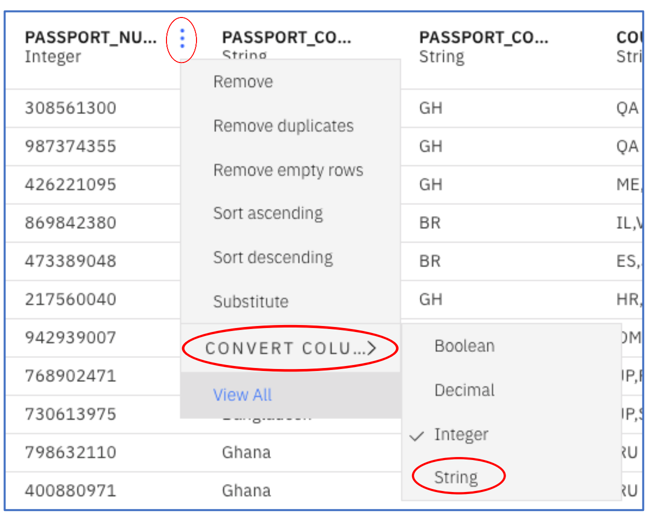
Before profiling the data, we will do some data preparation. Note, skip steps 1-4 if both the VETTING_LEVEL column and the PASSPORT_NUMBER column are Strings.

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.

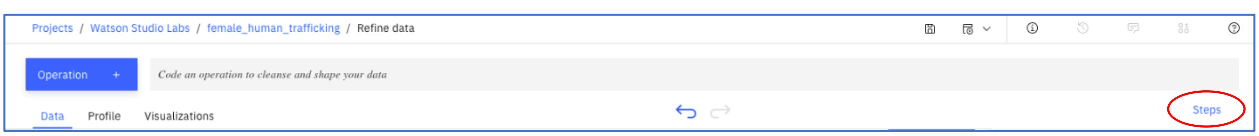
1. Some of the columns in the data set are defined as Integers but should be treated as Strings. We can easily convert the columns from Integers to Strings. Convert the **VETTING_LEVEL** column by hovering over VETTING_LEVEL, clicking on the vertical ellipse ⋮, clicking on **CONVERT COLUMN**, and clicking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



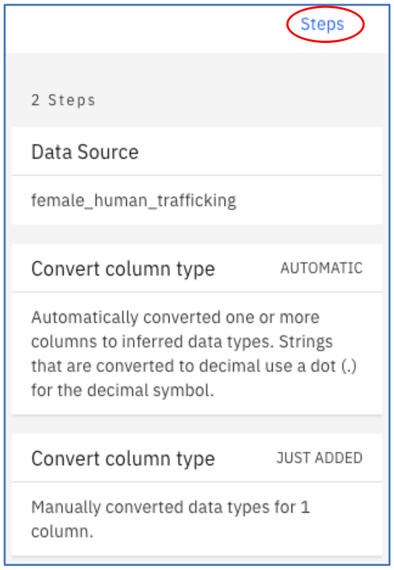
2. Convert the **PASSPORT_NUMBER** column by hovering over PASSPORT_NUMBER, clicking on the vertical ellipse ⋮, clicking on **CONVERT COLUMN**, and clicking on **String**.



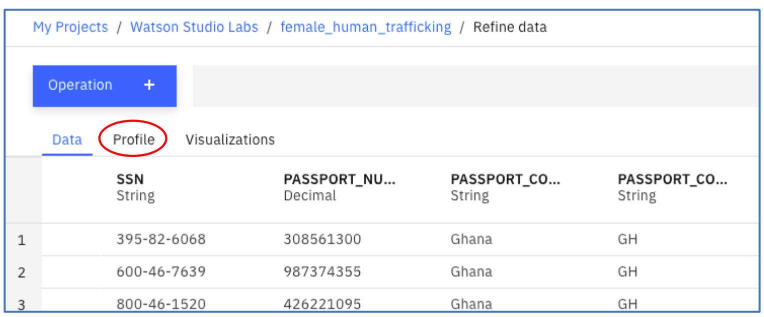
3. Click on the **Steps** link (if the **Steps** display is not visible).



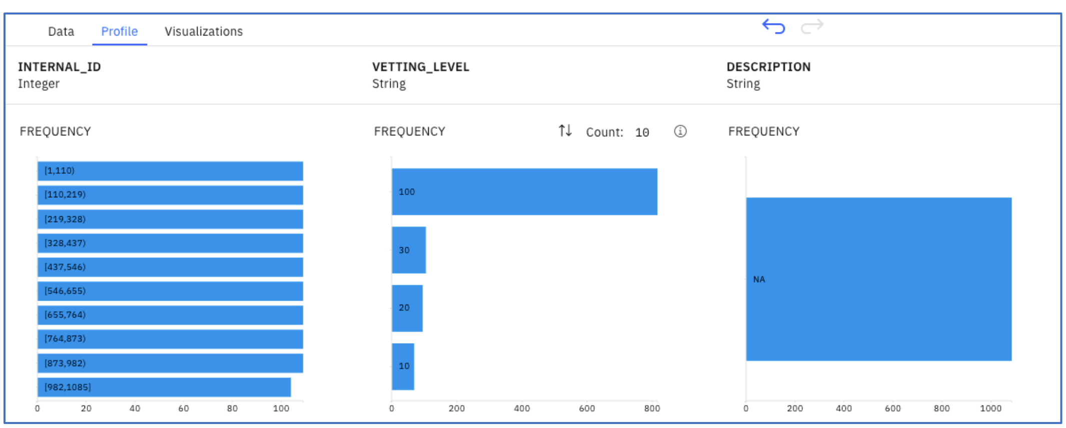
4. Each data operation is recorded in the **Steps** display providing an audit list of the operations performed. So far, we have done two column conversion operations. The steps in the **Steps** display can be edited. Operations can be removed from the list or modified.



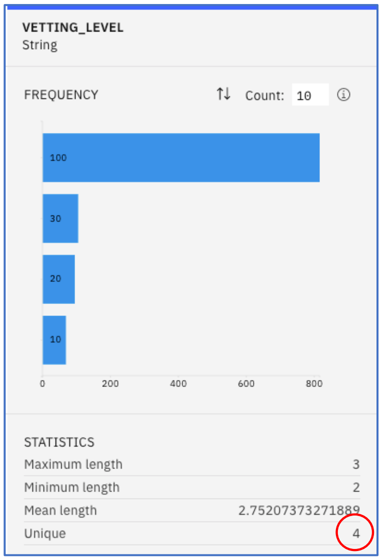
5. Click on **Profile**.



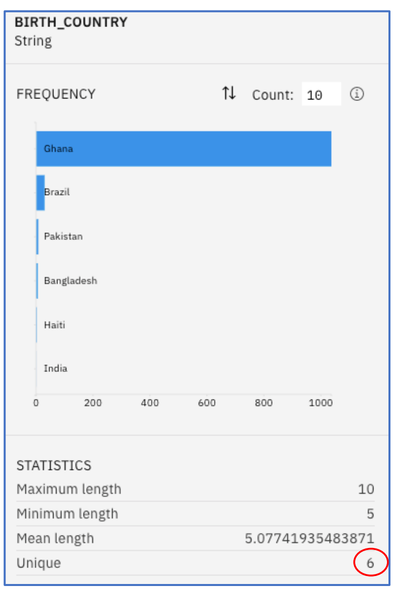
6. The Profile panel displays the counts of the top 10 values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column.



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.

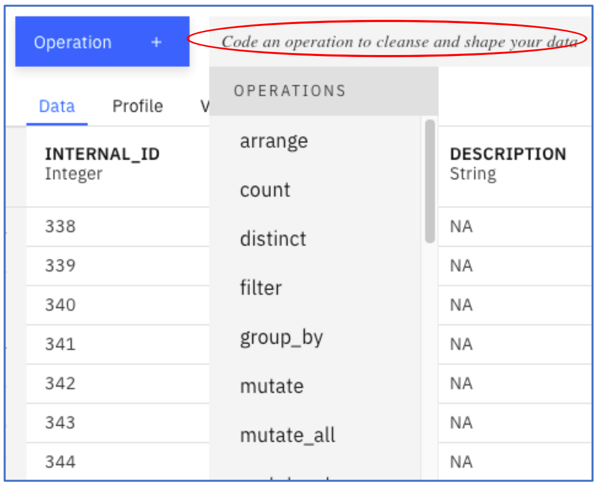


8. Scroll to the right to view the columns. As we mentioned earlier, the occupation column is very granular and has about 475 unique entries. It is not suitable for modeling purposes unless it is categorized. The BIRTH_COUNTRY, and PASSPORT_COUNTRY shows only 6 unique countries. The COUNTRIES_VISITED_COUNT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

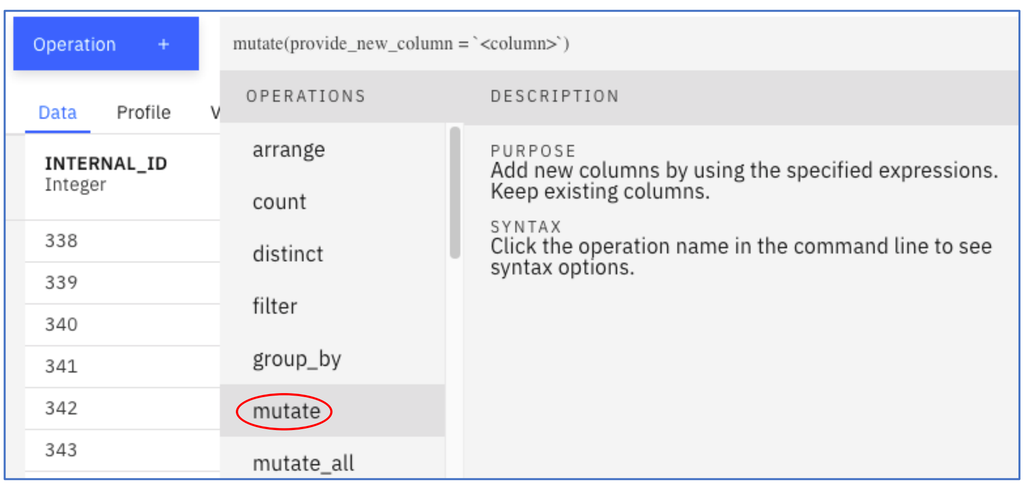


9. Based on the profiling information, we will do some additional transformations. Click on the **Data** link.

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.

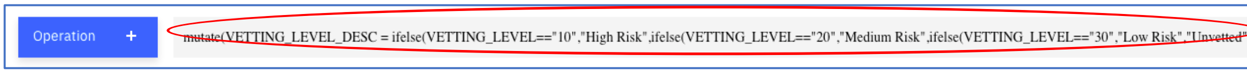


11. Hover the mouse over **mutate.** A description of the mutate function is provided.



12. Click on **mutate** and cut and replace the generated code with the following and then click **Apply**. Note, if an error occurs, it is because of a line break. Remove the line breaks and try again.

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



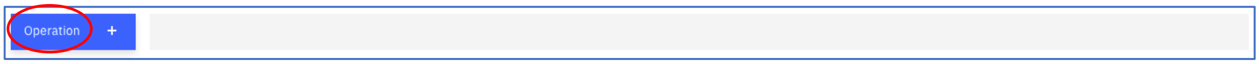
13. On the right side of the text entry box, click Apply.

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”.

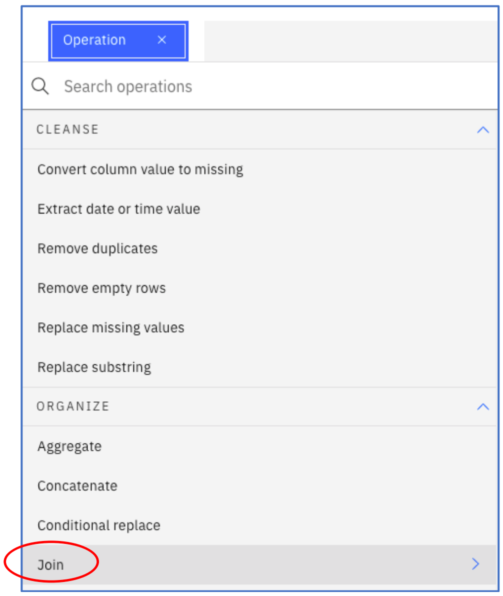
15. Let’s extract the fields of interest by using another dplyr function, **select**. Cut and paste the following code into the operations area and click **Apply**. Again, remove the line breaks and try again if you get an error.

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

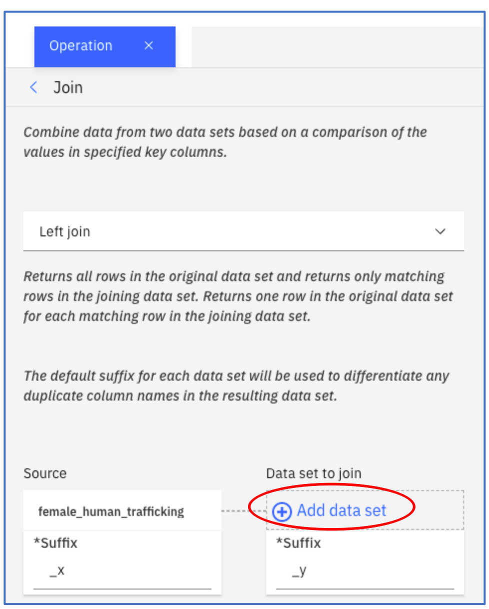
16. Let’s now bring in the other datasets (Occupation, Categories). We use a Join operation to first join in the Occupation dataset, and then join the Categories dataset. Click on **Operation** +.



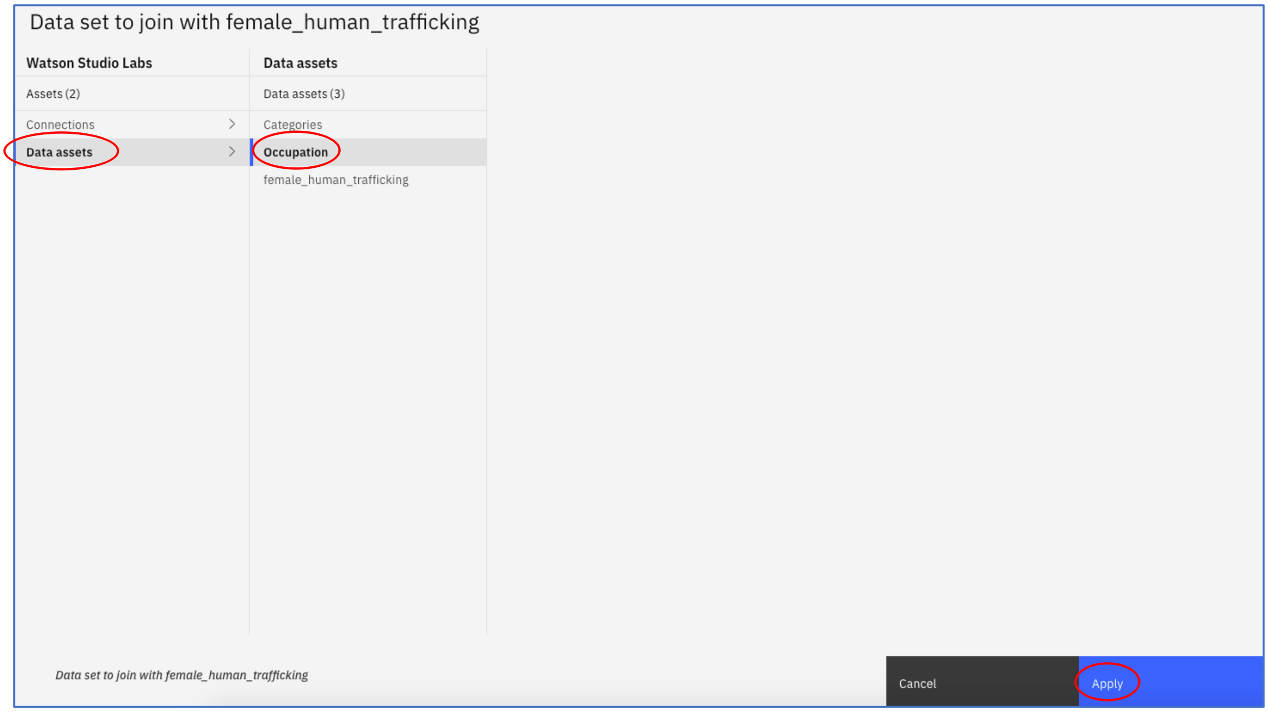
17. Scroll down and click on **Join**.



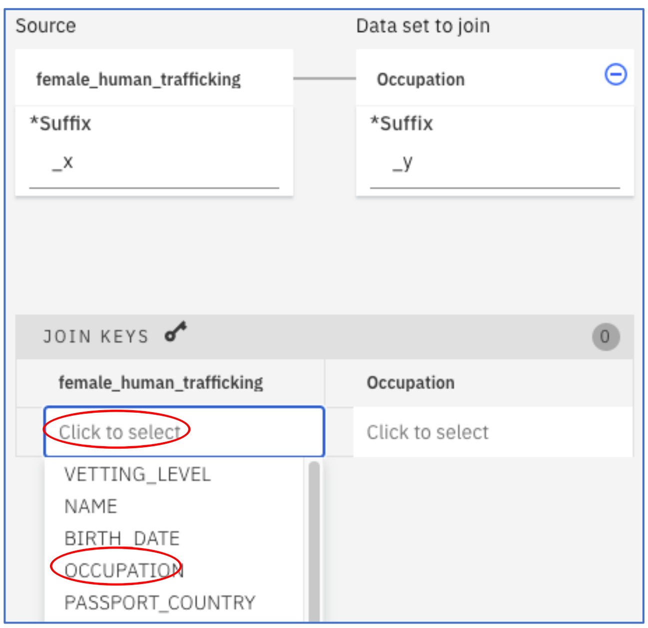
18. Keep **Left join** and then click on **Add Data Set**



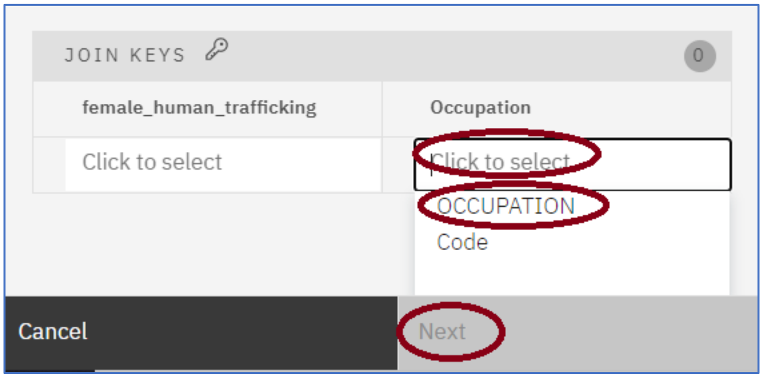
19. Click on **Data Assets**, click on **Occupation**, and then click **Apply**.



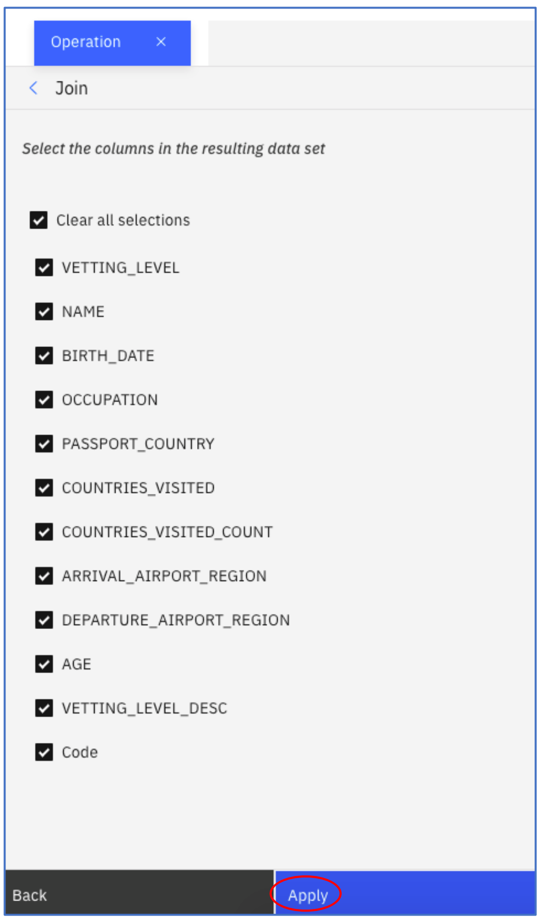
20. Scroll down. In **JOIN KEYS** under **female_human_trafficking** click **Click to select**, and then click **OCCUPATION.**



21. In **JOIN KEYS** under **Occupation** click **Click to select**, click **OCCUPATION**, and then click on **Next**.

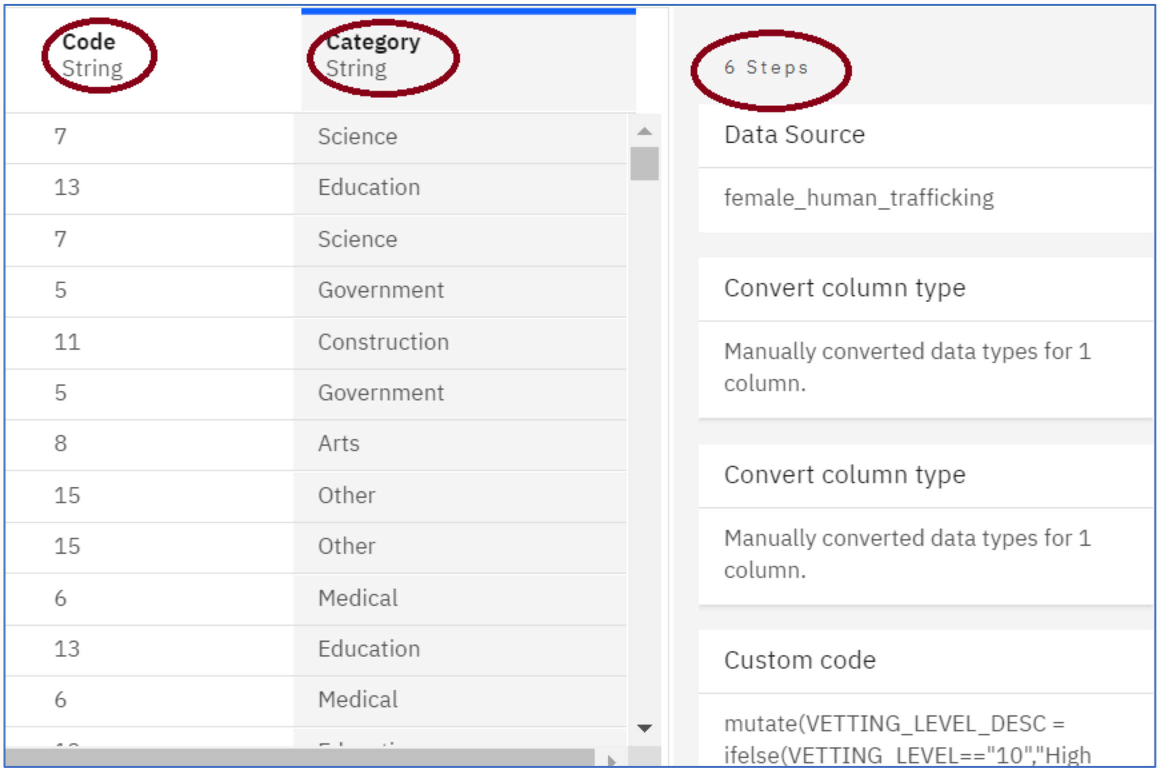


22. Click **Apply**.

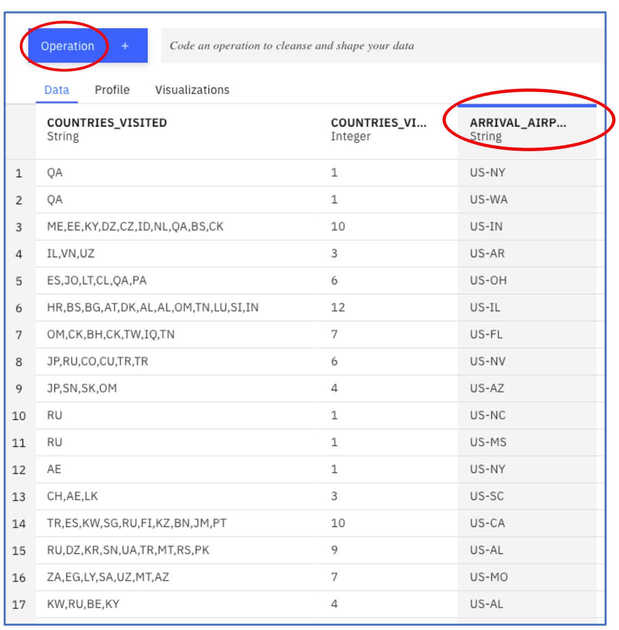


23. Follow steps 19-22 to join the Categories dataset. The join keys are the Code fields in both datasets. As a result of the joins, two new columns are added, a Code column, and a Category column. <u>Note that your number of Steps may be different</u> as Data Refinery may

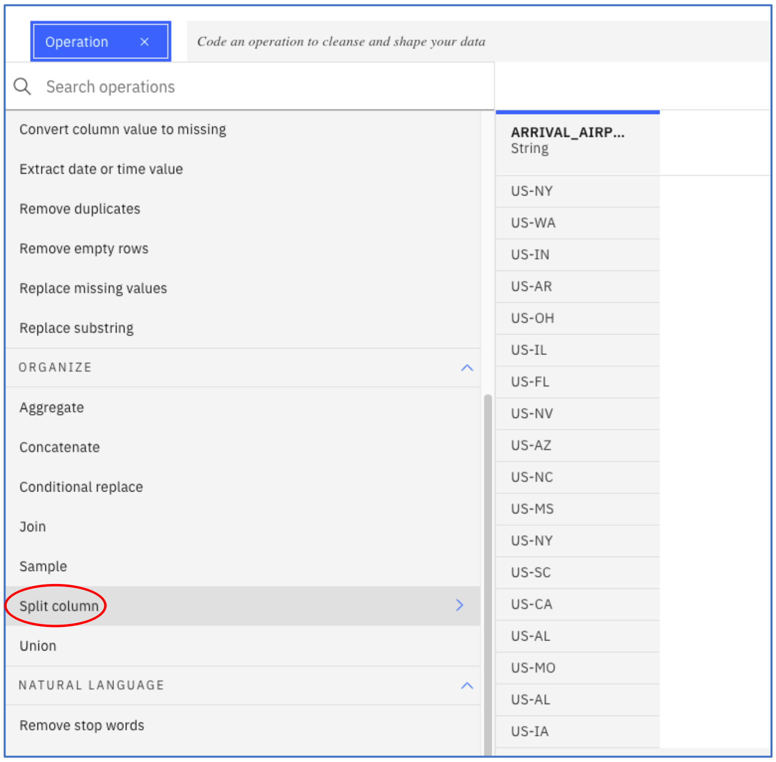
have automatically converted columns. So far we have added a data source, converted two columns, entered two custom code commands, and completed two joins.



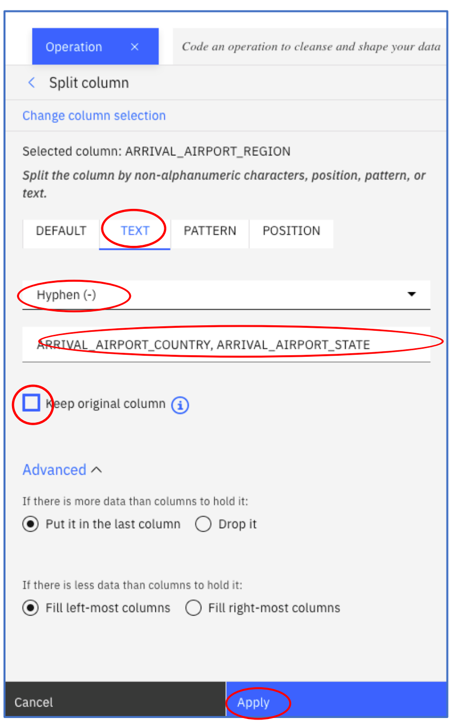
24. We note that the ARRIVAL_AIRPORT_REGION column has "US" concatenated with a State abbreviation (eg US-CA) We want to strip away the "US" to use the column as a State column. The operation **Split column** can be used. Click on ARRIVAL_AIRPORT_REGION to highlight the column then click on **Operation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



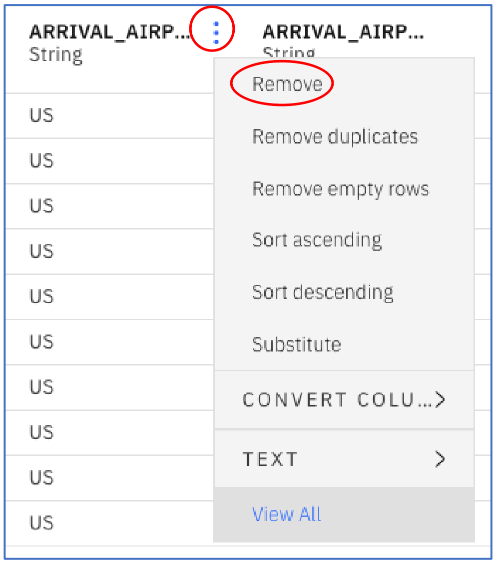
25. Click on **Split column.**



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**.

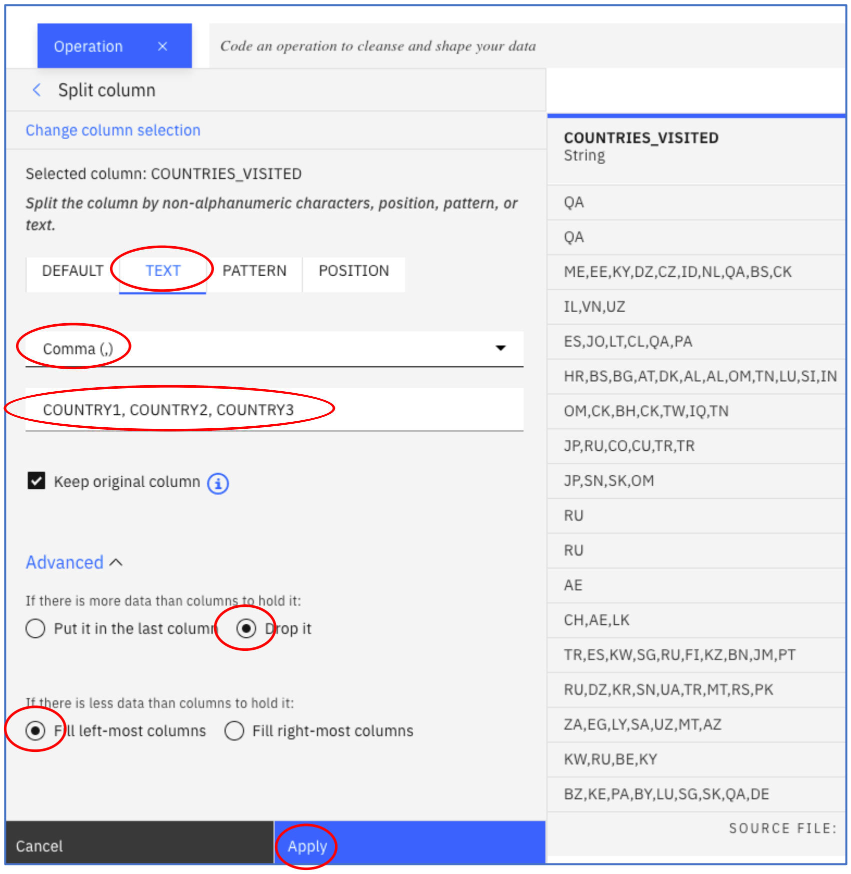


27. Two new columns are created. We don't need the ARRIVAL_AIRPORT_COUNTRY since it has only 1 value – US. Remove the ARRIVAL_AIRPORT_COUNTRY by hovering over the ARRIVAL_AIRPORT_COUNTRY header, clicking on the vertical ellipse ⋮ and clicking on **Remove**.

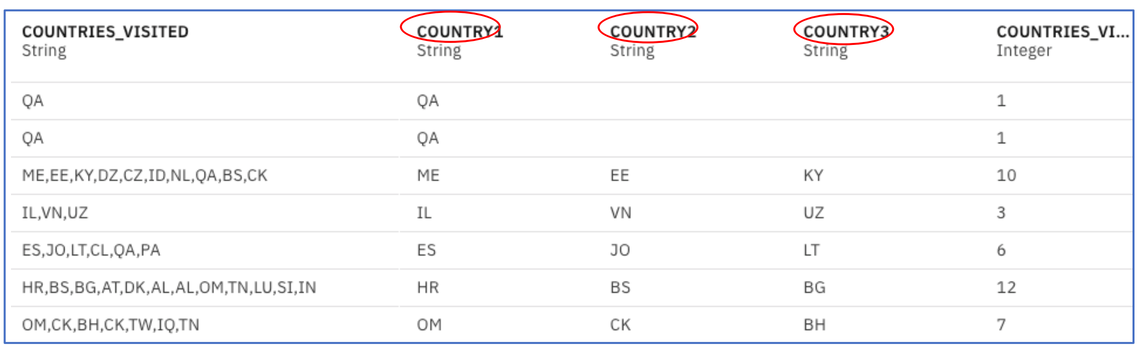


We can also use the **Split column** operation on other columns in the dataset. The BIRTH DATE column can be split into YEAR, MONTH, DAY. The DEPARTURE_AIRPORT_REGION can be split in a similar manner as the ARRIVAL_AIRPORT_REGION. The COUNTRIES_VISITED column can be split by the comma. The resulting 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.

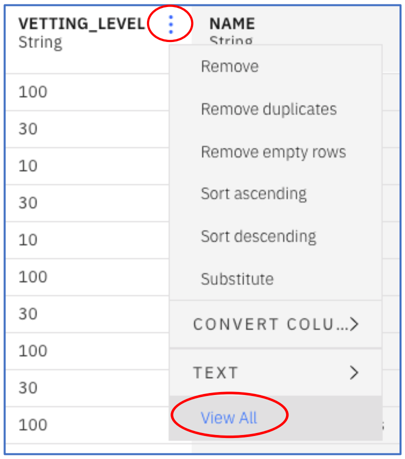


29. The results are shown below.

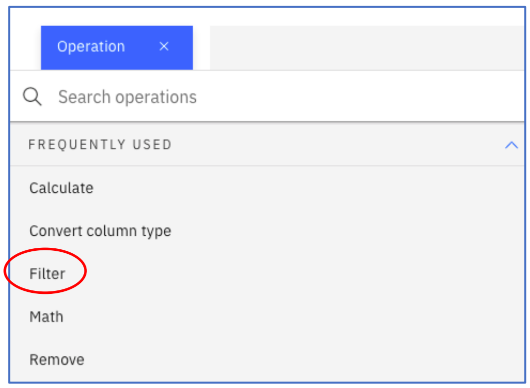


COUNTRIES_VISITED String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VI... Integer
QA	QA			1
QA	QA			1
ME,EE,KY,DZ,CZ,ID,NL,QA,BS,CK	ME	EE	KY	10
IL,VN,UZ	IL	VN	UZ	3
ES,JO,LT,CL,QA,PA	ES	JO	LT	6
HR,BS,BG,AT,DK,AL,AL,OM,TN,LU,SI,IN	HR	BS	BG	12
OM,CK,BH,CK,TW,IQ,TN	OM	CK	BH	7

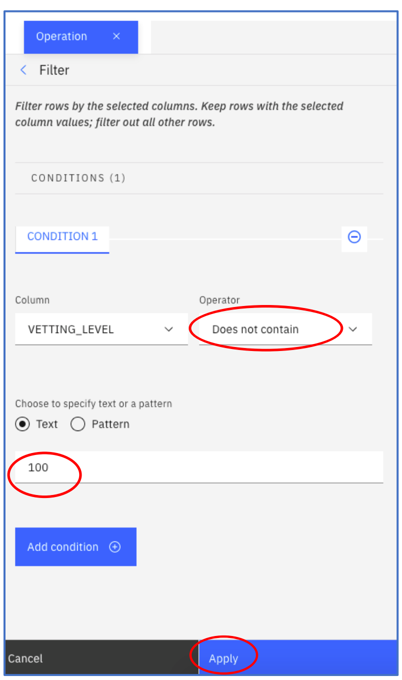
30. Let's use visualization to get a better understanding of the data. First, we will remove the unvetted records. Hover over the VETTING_LEVEL column, click on the vertical ellipse ⋮, click on **View All**.



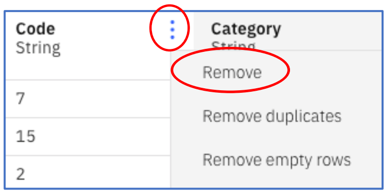
31. Click on **Filter**.



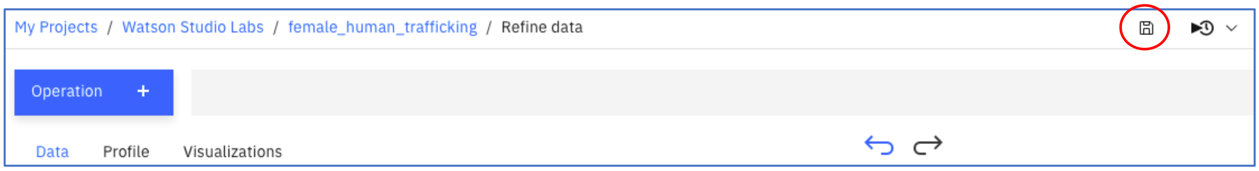
32. Change **Operator** to **Does not contain**, put value as 100, and then click **Apply**.



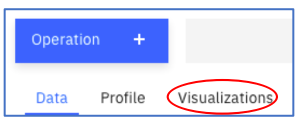
33. Remove the Code column by clicking on the vertical ellipse ⋮ and then clicking **Remove**.



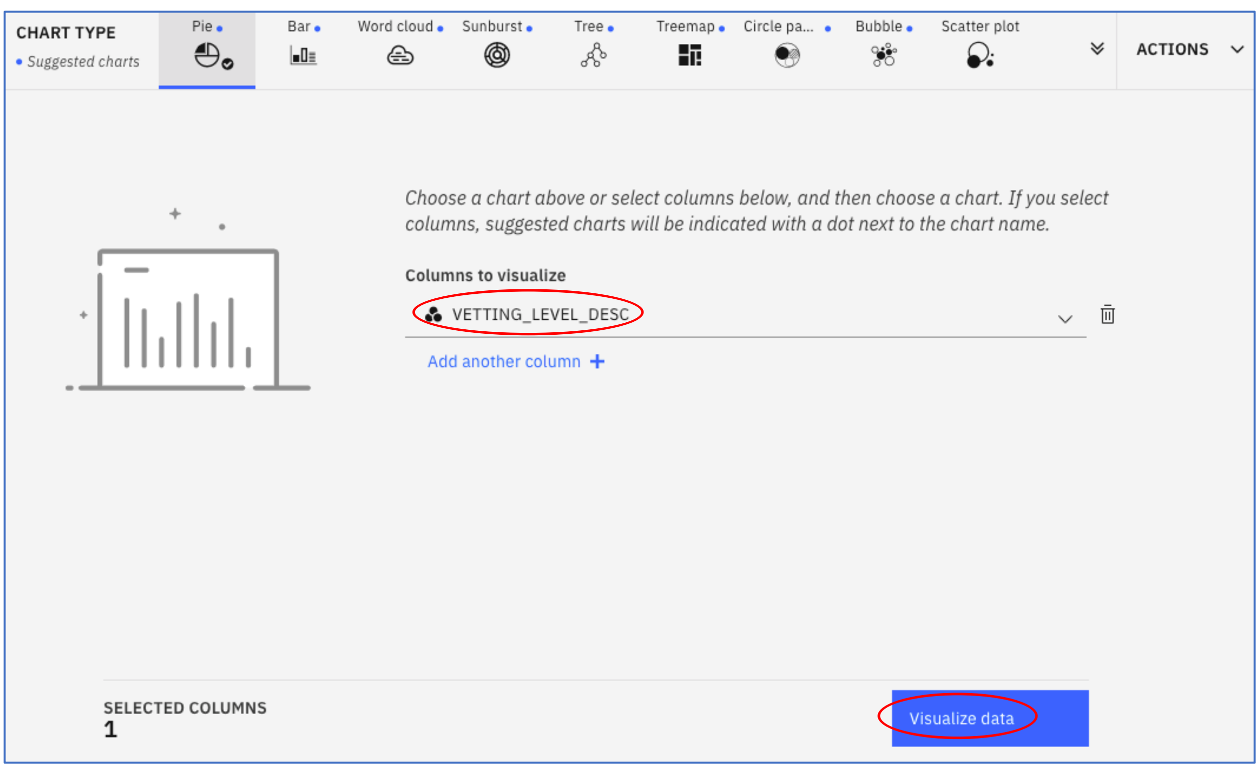
34. Save the Data Flow by clicking on the Save 🖫 icon.



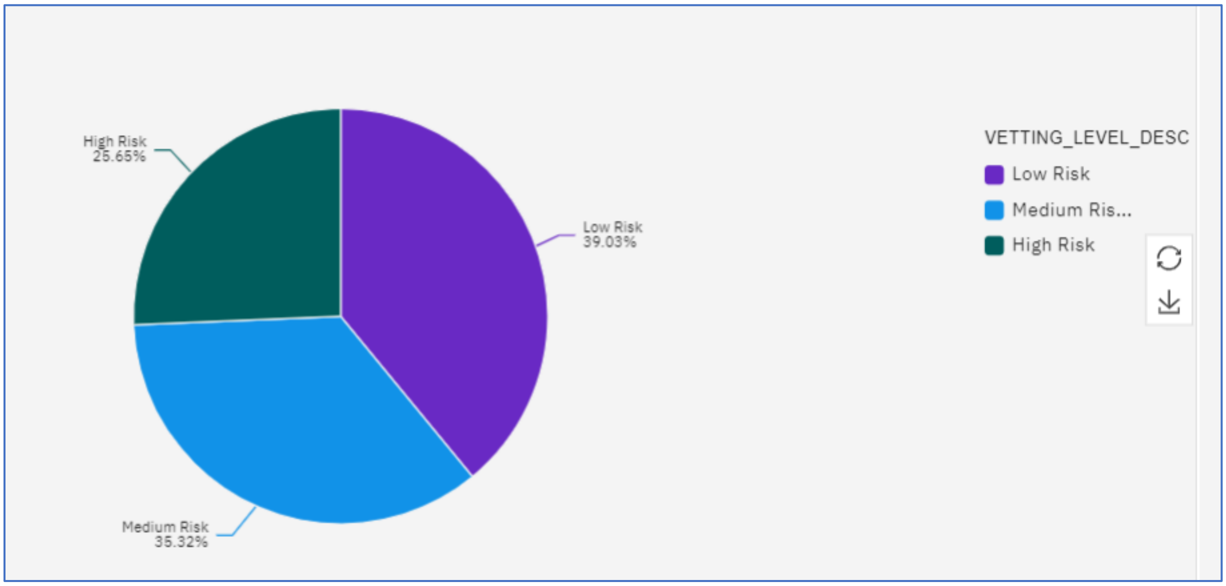
35. Click on the **Visualization** tab.



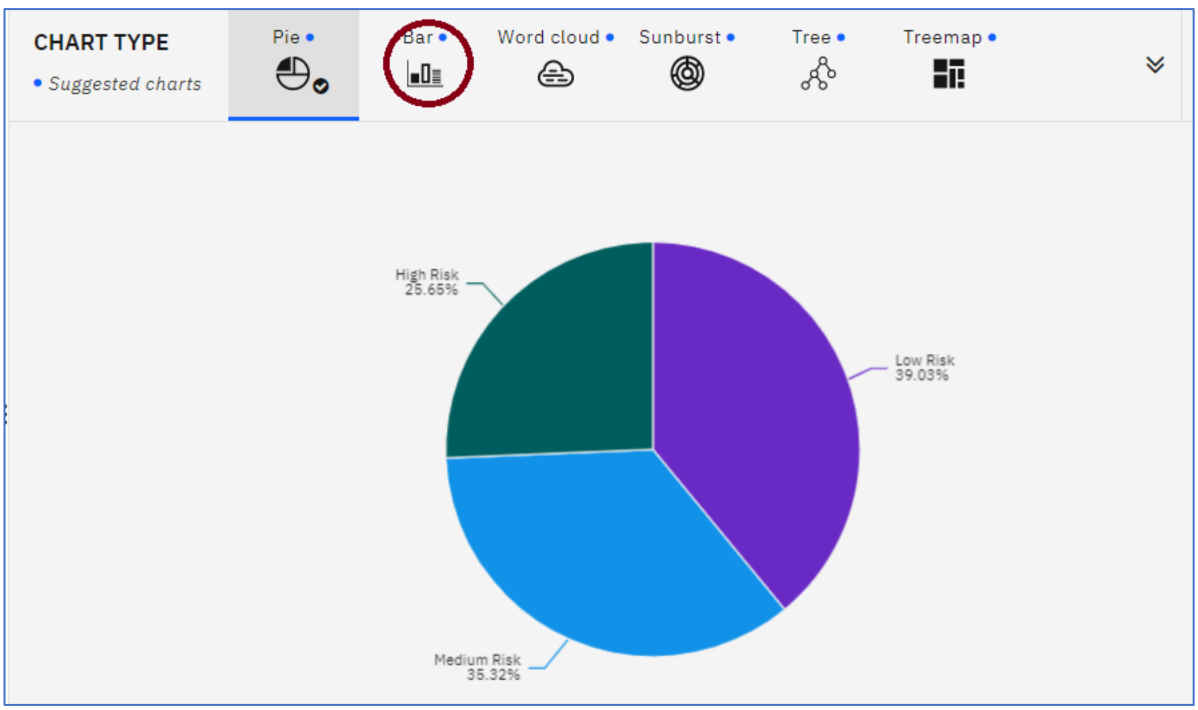
36. Click on **VETTING_LEVEL_DESC** for **COLUMNS TO VISUALIZE**, and then click on **Visualize data**.



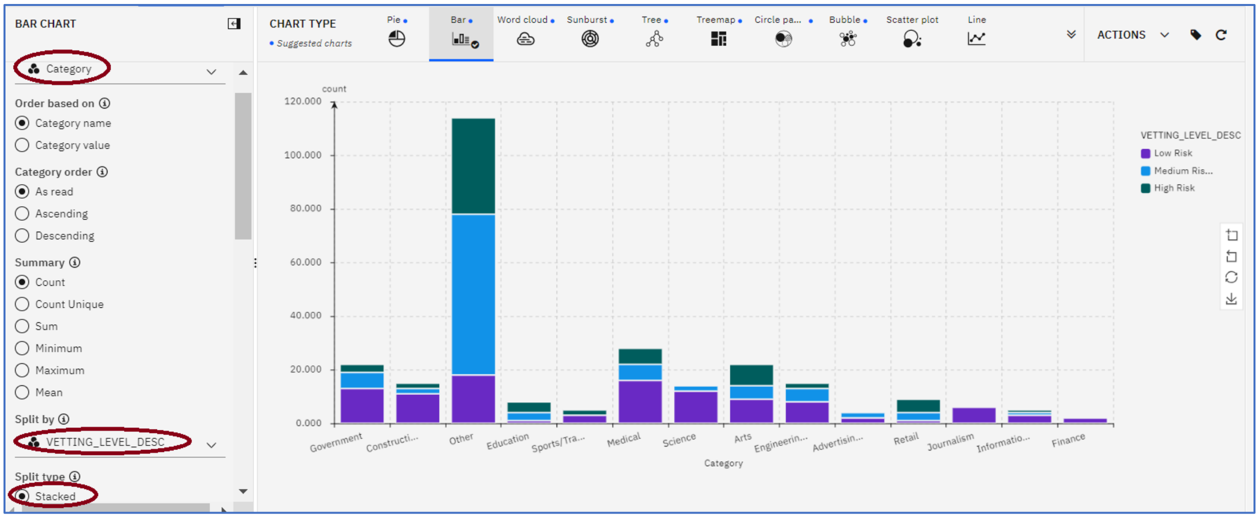
37. A pie chart is selected as the suggested visualization. The breakdown in the different risk categories is shown below and roughly balanced. Note, the results may be slightly different than what is on your screen.



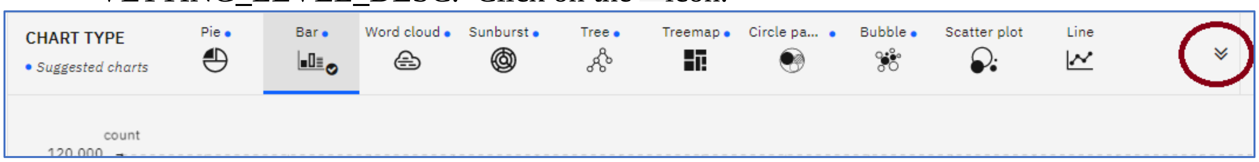
38. We can visualize the breakdown of travel records by job category and vetting level. Click on the click **Bar**.



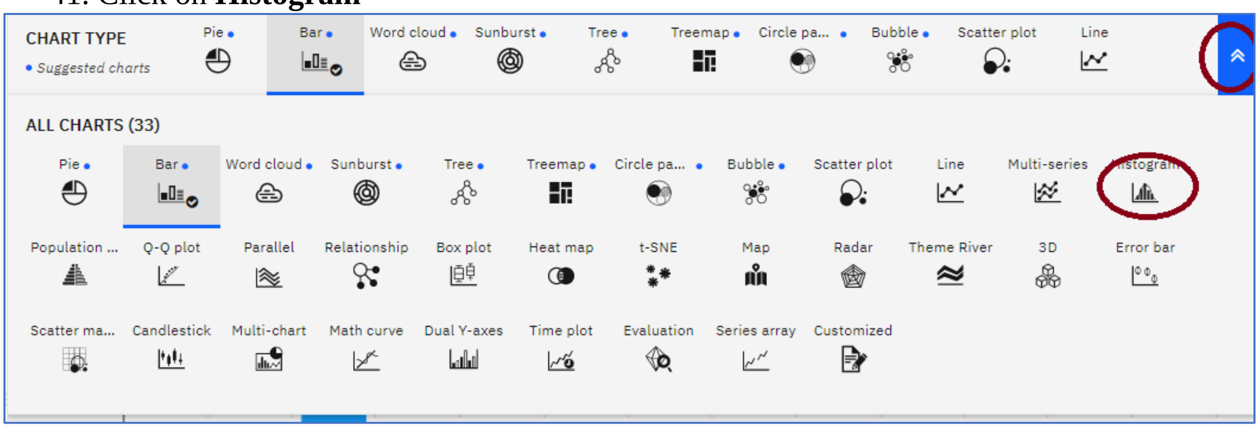
39. Click on **Category** for **Category**, click on VETTING_LEVEL_DESC for **Split by**, click on **Stacked** for **Split type**. The resulting visualization is shown below. By visual inspection, it appears that there is a variability of vetting level based on job category.



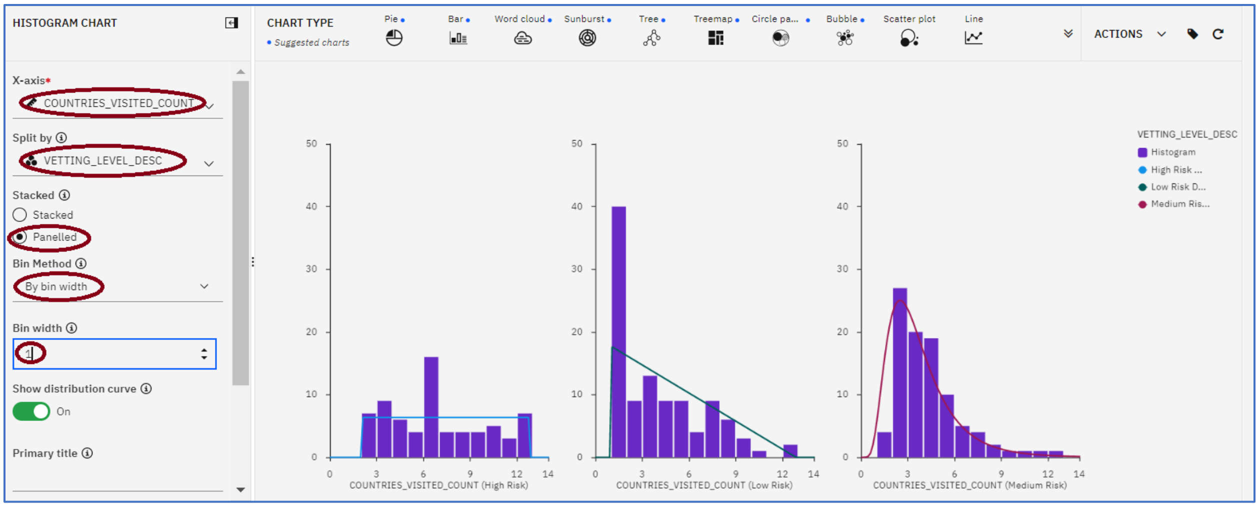
40. We can visualize a histogram of COUNTRIES_VISITED_COUNTS split by VETTING_LEVEL_DESC. Click on the icon.



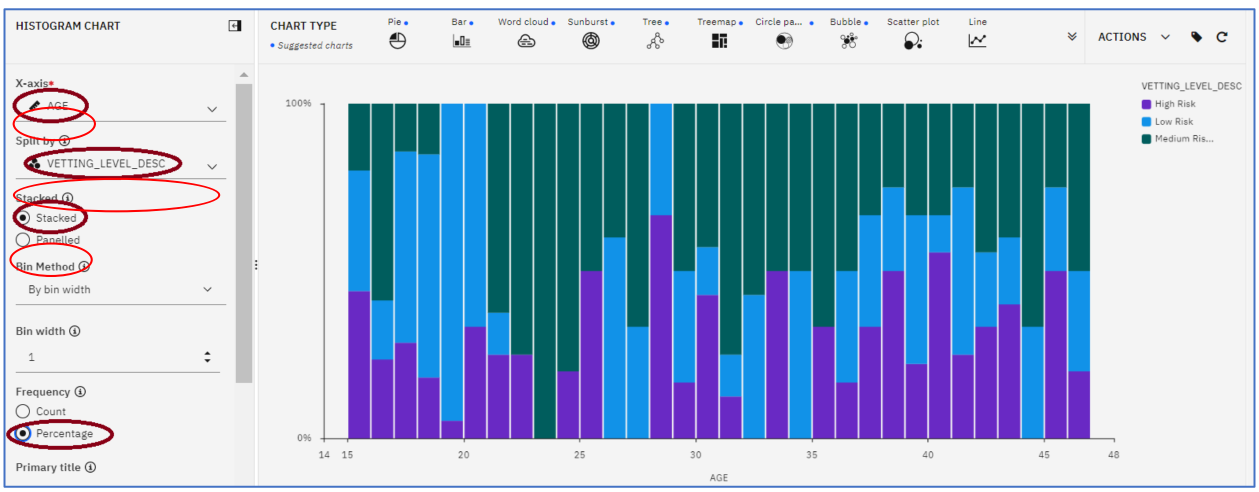
41. Click on **Histogram**



42. Click on **COUNTRIES_VISITED_COUNT** for **X-axis**, click on **VETTING_LEVEL_DESC** for **Split by**, click on **Paneled**, click on **By bin width** for the **Bin Method** and select 1 for the **Bin width**. Note that a higher number of high risk persons visit many countries.



43. Let’s examine if age makes a difference. Click on **AGE** for **X-axis**. **Split by** remains **VETTING_LEVEL_DESC**, click on **Stacked**, and click on **Percentage**. There is not a clear pattern on the influence of age on high risk persons. It appears that younger travelers may have a slightly lower risk of being trafficked.

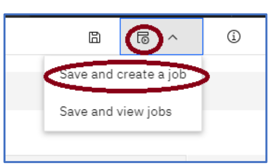


44. Please feel free to experiment with other visualizations.

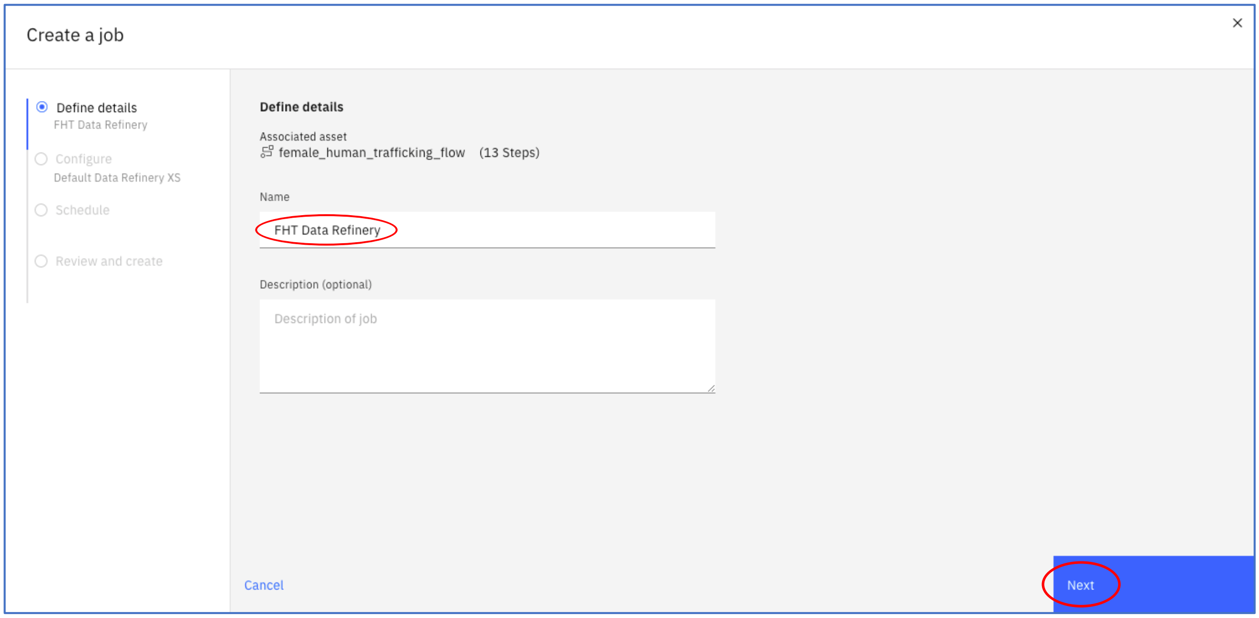
Run the sequence of Data Operations on the entire data set.

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 .

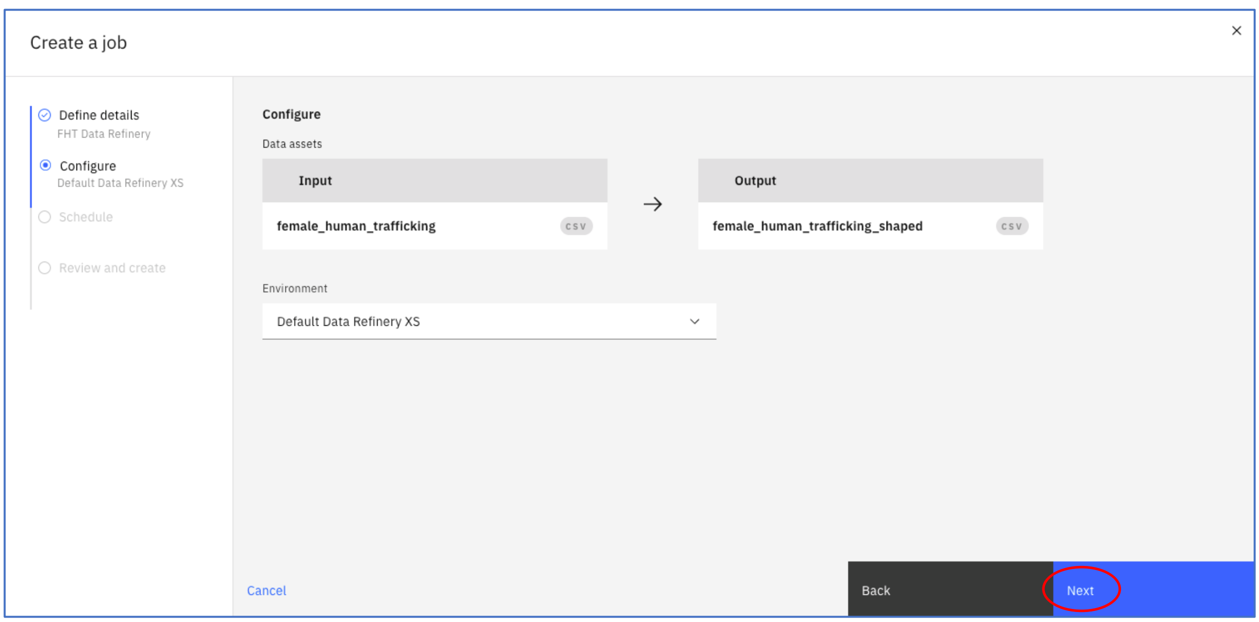
45. Click on **job** icon and click on **Save and create a job**.



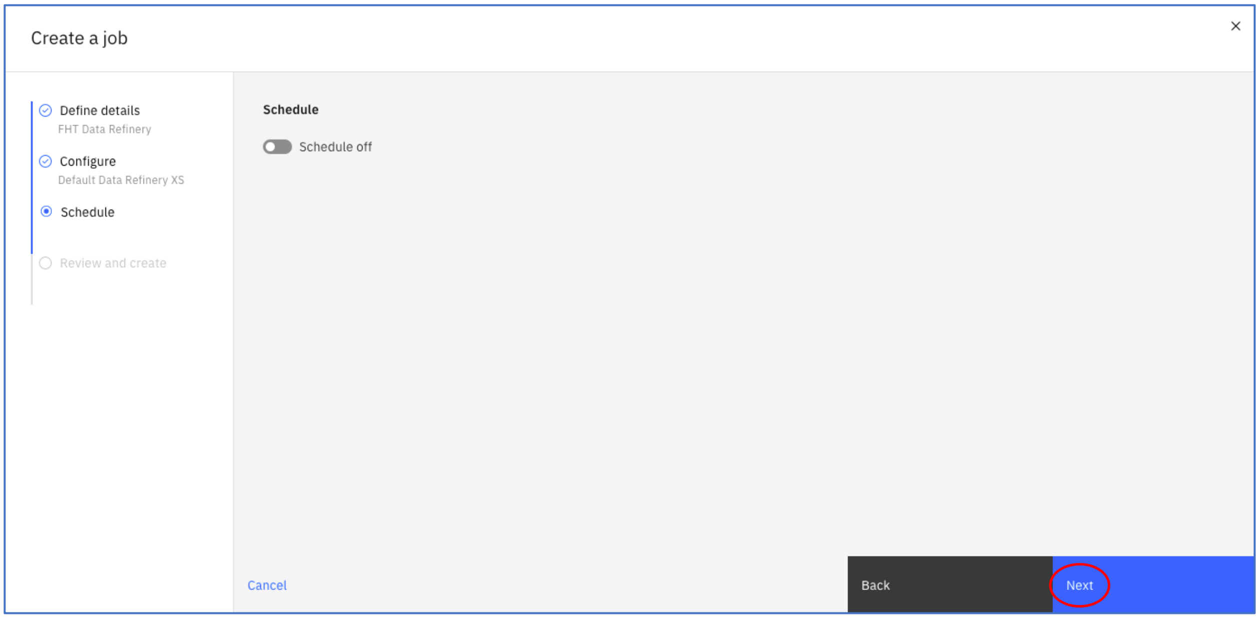
46. Enter a **Job Name** for the job. Note the number of steps used to transform the data. It should be 11-13 steps depending on if Data Refinery automated column conversion and if any steps were skipped. A schedule can be set up if the transformation process needs to run on a scheduled basis. We are just going to do a one-time run. Click **Next.**



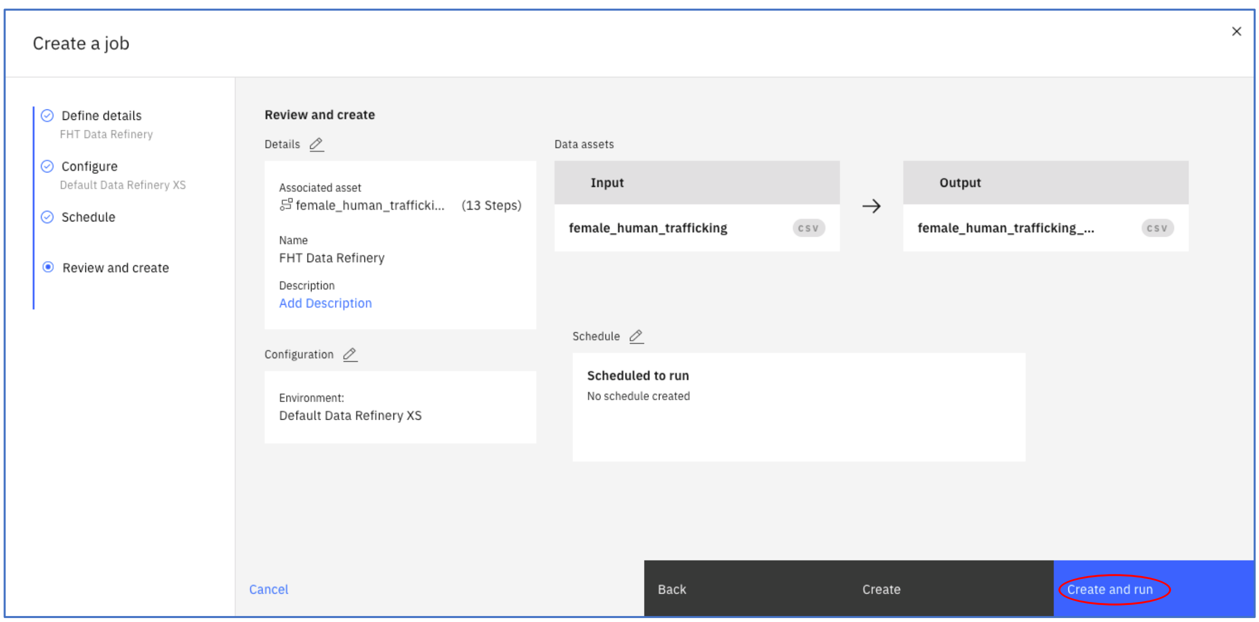
47. Keep the default input, output, and environment and click **Next**.



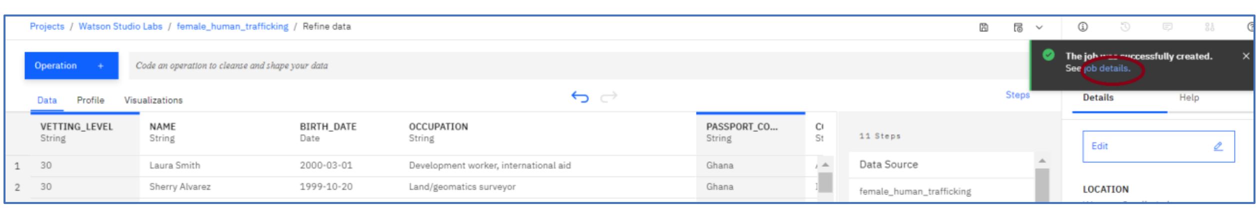
48. Keep schedule unenabled and click **Next**.



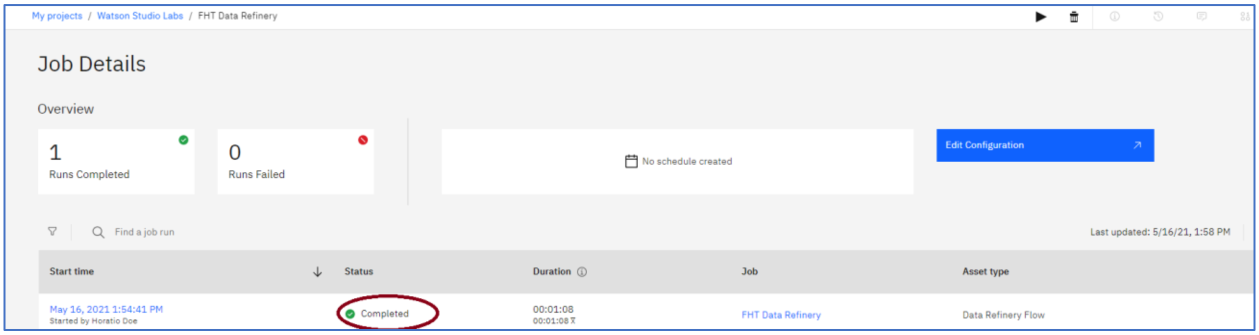
49. Click **Create and run**.



50. Click on Job Details.

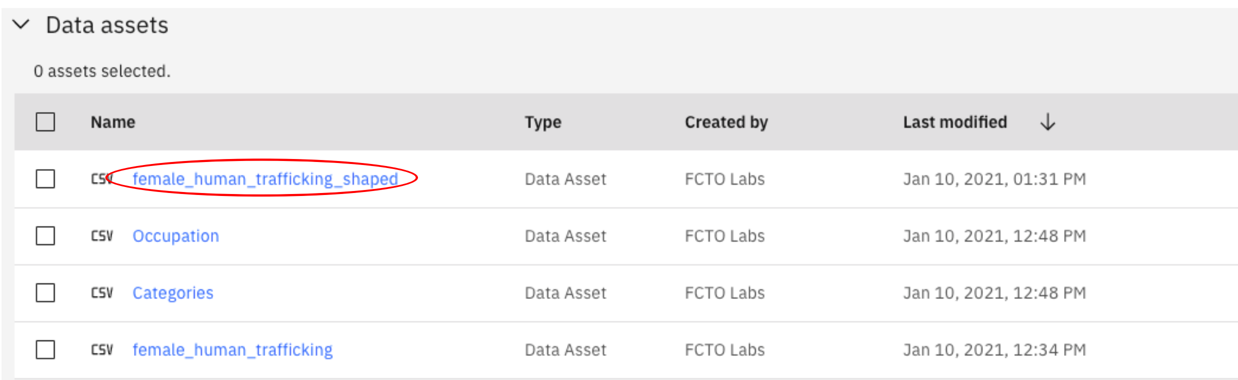


51. Wait until the job run changes from **Running** to **Completed**.



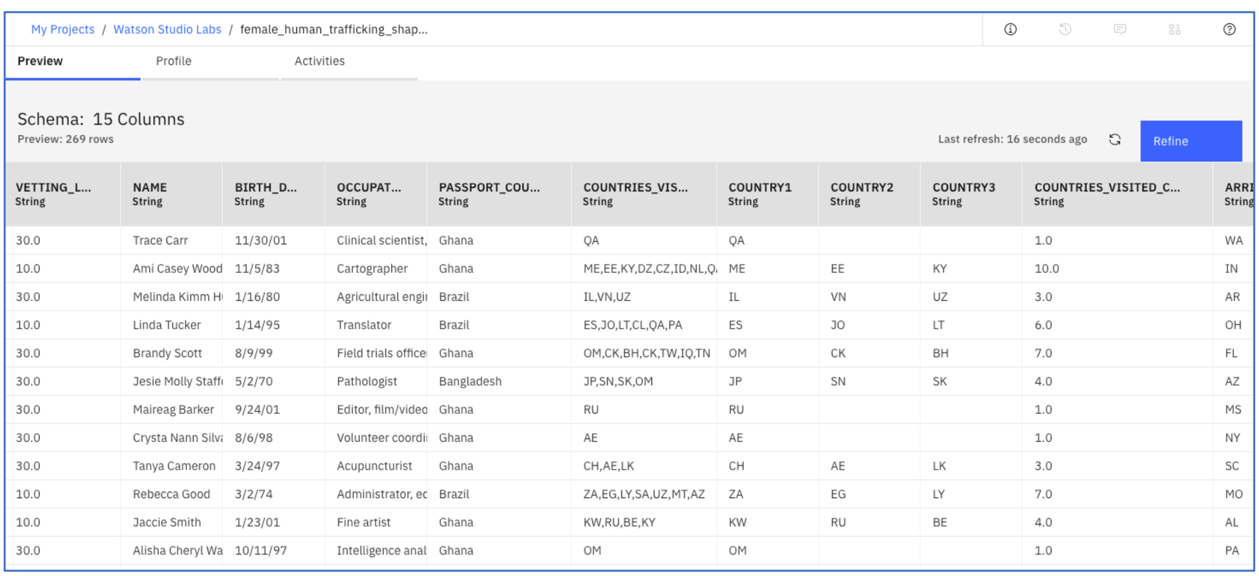
52. The output of the Data Refinery process should be listed in the Data Assets. Click on **Watson Studio Labs** to return to the Project view.

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



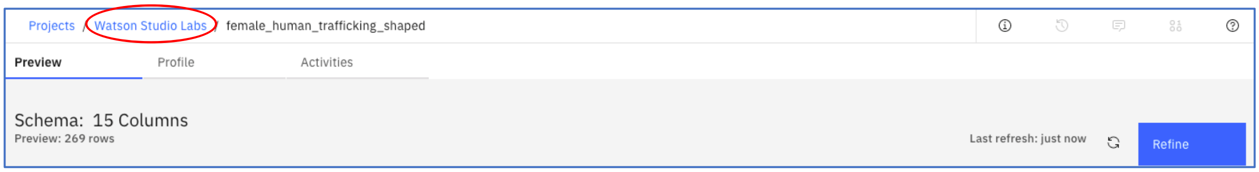
Name	Type	Created by	Last modified
CSV female_human_trafficking_shaped	Data Asset	FCTO Labs	Jan 10, 2021, 01:31 PM
CSV Occupation	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
CSV Categories	Data Asset	FCTO Labs	Jan 10, 2021, 12:48 PM
CSV female_human_trafficking	Data Asset	FCTO Labs	Jan 10, 2021, 12:34 PM

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

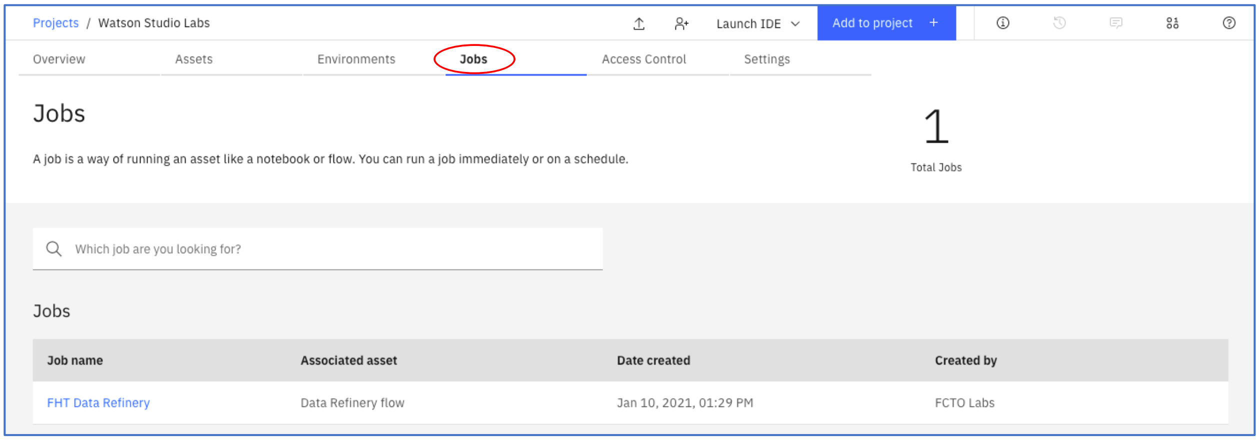


VETTING_L... String	NAME String	BIRTH_D... String	OCCUPAT... String	PASSPORT_COU... String	COUNTRIES_VIS... String	COUNTRY1 String	COUNTRY2 String	COUNTRY3 String	COUNTRIES_VISITED_C... String	ARRI String
30.0	Trace Carr	11/30/01	Clinical scientist,	Ghana	QA	QA			1.0	WA
10.0	Ami Casey Wood	11/5/83	Cartographer	Ghana	ME,EE,KY,DZ,CZ,ID,NL,Q,	ME	EE	KY	10.0	IN
30.0	Melinda Kimm H	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 Staff	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 Silva	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

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



56. 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.