**Lab for Chapter 2, Part :**

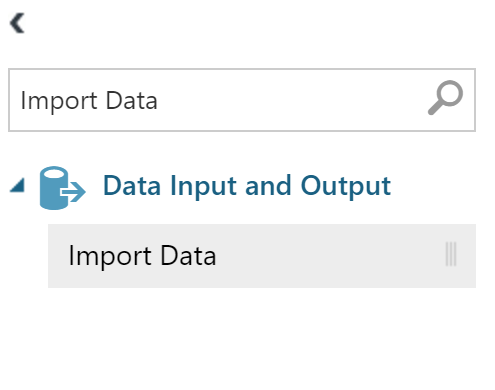
**Exploratory Analysis & Hypothesis Testing**

Contents

1. [Exercise 0: Obtain the Titanic sample data 3](#_bookmark0)
2. [Exercise 1: Casting Columns 5](#_bookmark1)
3. [Exercise 2: Data Visualization & Exploration 7](#_bookmark2)
4. [Exercise 3: Renaming Columns 10](#_bookmark3)
5. [Exercise 4: Joining Tables 13](#_bookmark4)

# **Exercise 0: Obtain the Titanic sample data**

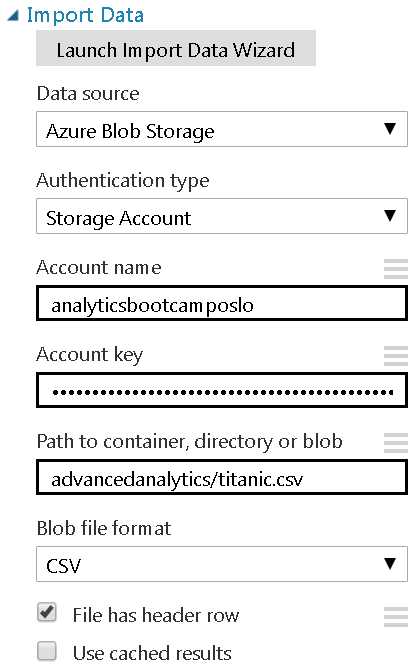
* 1. Drag and drop an ‘Import Data’ module from the menu on left.



* 1. Input the sample blob storage account we have set up for you.
     1. **Data Source**: Azure Blob Storage
     2. **Authentication type**: Storage Account
     3. **Account name**: analyticsbootcamposlo
     4. **Account key:**

WxHhL/+EhKva80Y3x25Id4gYndW0H6hzo1ChikRzaD21rtf3Dy2JIy4mNcSWw7ohqqDA1UPB29OTfoG7ogP2+w==

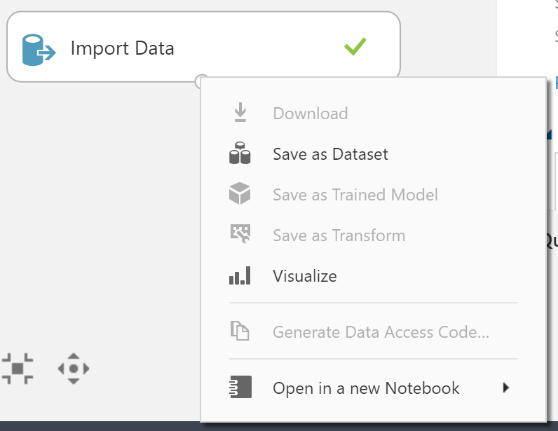
* + 1. **Path to container, directory or blob:** advancedanalytics/titanic.csv
    2. **Blob file format:** CSV
    3. **File has header row:** Checked

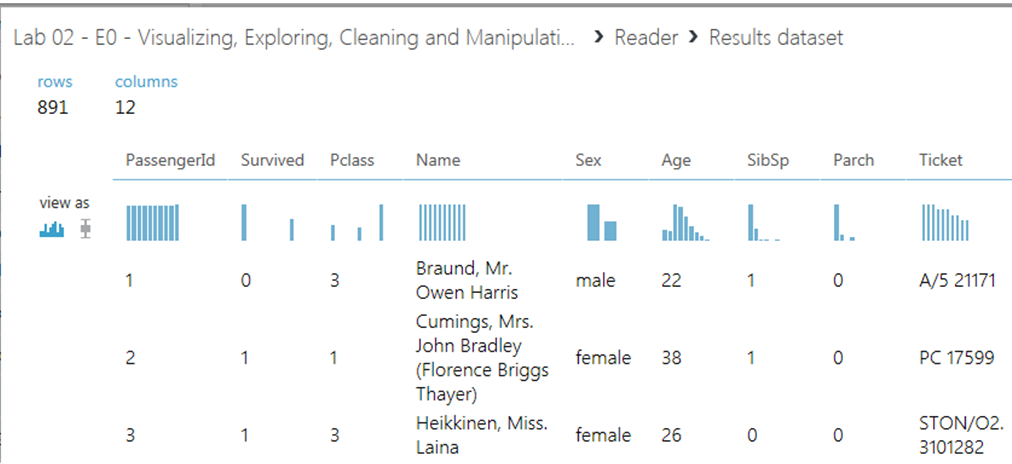


* 1. Run the experiment to execute the import and parse.

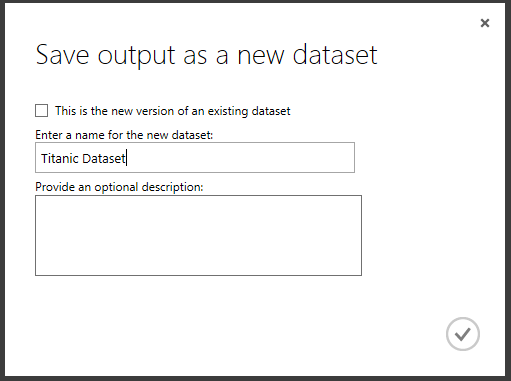


* 1. Preview the data by visualizing the output of the ‘Import Data’ module. Right click the bottom middle node of the ‘Import Data’ module to access the menu.

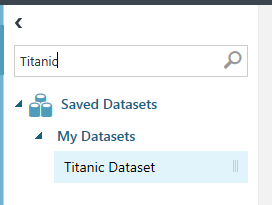




* 1. The data is not yet saved. Save it to the workspace by clicking “Save as Dataset” in the Reader’s output node. Name the dataset as “Titanic Dataset”.



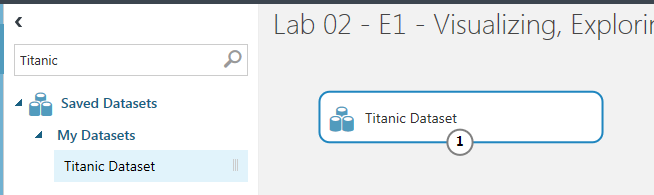
* 1. Go into any experiment and verify that the dataset has been imported.
     1. The data will be under a directory called “Saved Datasets” within any experiment



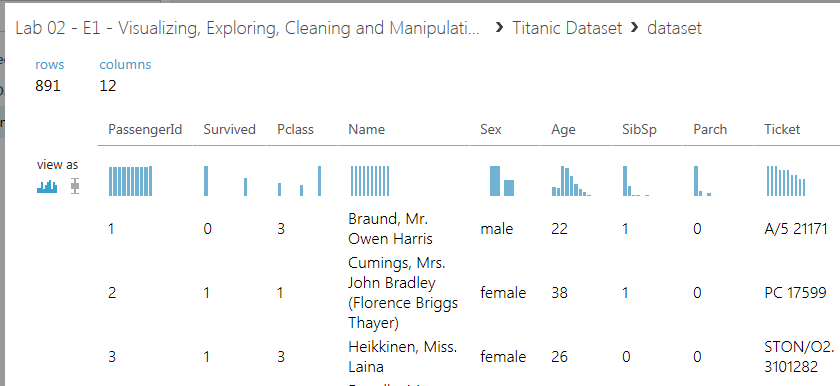
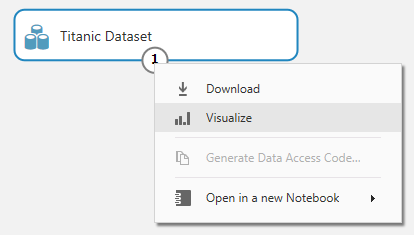
# **Exercise 1: Casting Columns**

This Titanic dataset contains **categorical** data types. However, we must tell Azure which of those columns are categorical, so that our models will not treat them as sequential numbers.

* 1. From saved datasets, drag the “**Titanic Dataset**” into the experiment workspace.



* 1. Verify that it’s the same looking data as below by visualizing the Titanic dataset.





### Cast categorical values to categorical.

### Drag in the ‘**Edit** **Metadata’** module and connect the output of “Titanic Dataset” module to the input of “Edit Metadata” module.

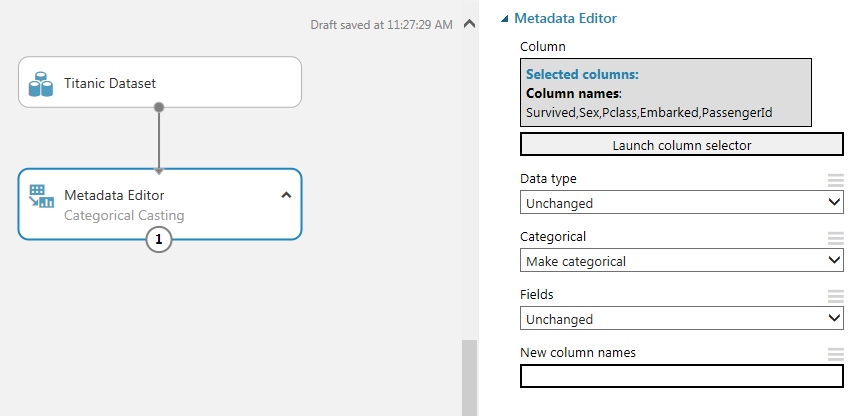
### In the ‘Edit Metadata’ Properties, click on ‘Launch column selector’.

### Begin with “No columns” and “Include” “column names”

### Add “Survived,” “Sex,” “Pclass,” “Embarked,” and “PassengerId”



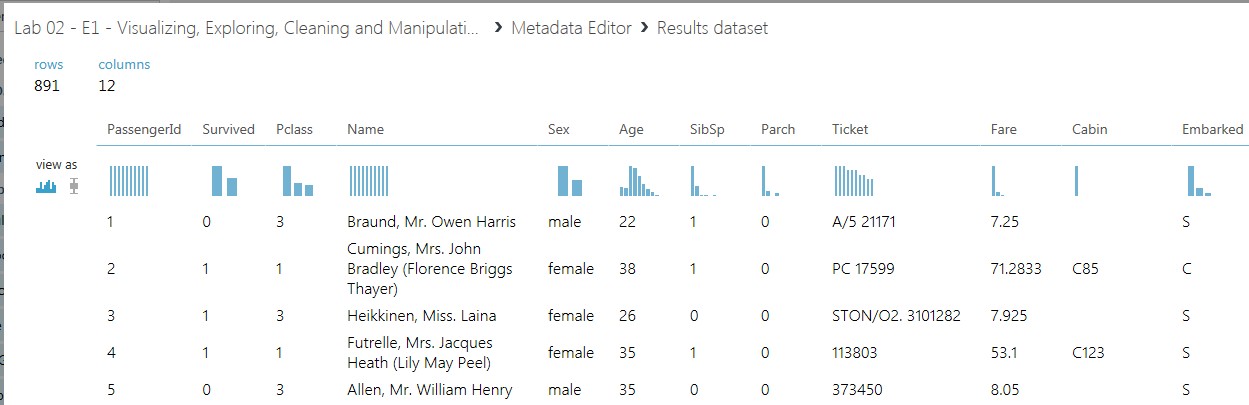
* + - 1. Click ok.
    1. Set Categorical to “Make categorical”
    2. Label the module as “Categorical Casting” for good style, by right clicking on the module and selecting “Edit Comment”



* 1. Run the experiment to execute the import and parse.



* 1. Preview the data by visualizing the output of the Metadata Editor module. Right click the bottom middle node of the Metadata Editor module to access the menu.



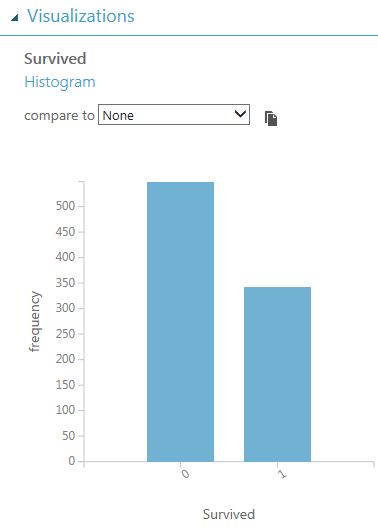
# **Exercise 2: Data Visualization & Exploration**

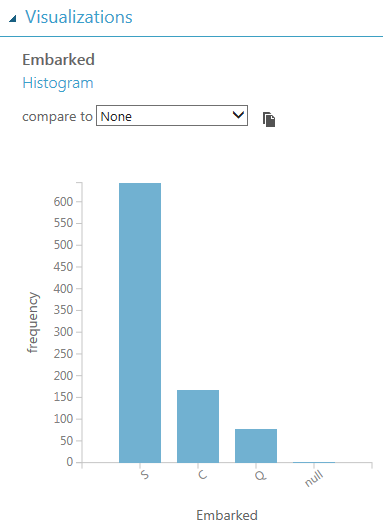
This Titanic dataset contains **categorical** data types. However we must tell Azure which of those columns are categorical, so that our models will not treat them as sequential numbers.

* 1. Histograms
     1. In the visualization mode of output of the Metadata Editor module, click on column “Survived” with view set to the picture of a histogram.

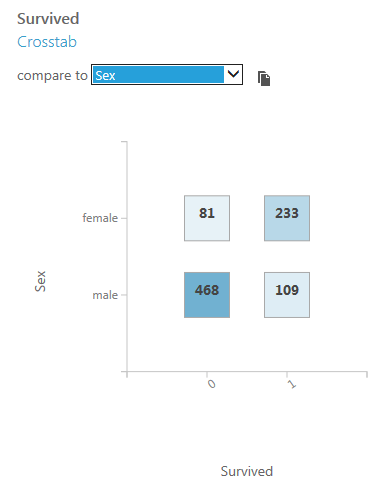


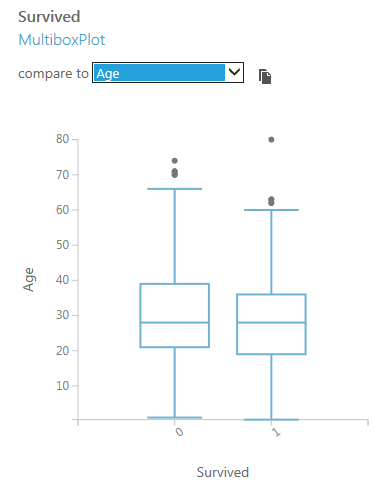
* + 1. A menu on the right will pop up. Expand the **Visualizations** drop down. A histogram will now appear.
    2. What is the distribution of survived vs. deceased? Did more people survive or perish?
    3. Where did survived or deceased people come from? Select the “Embarked” column.



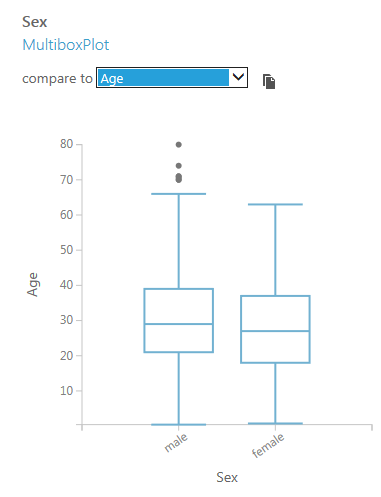
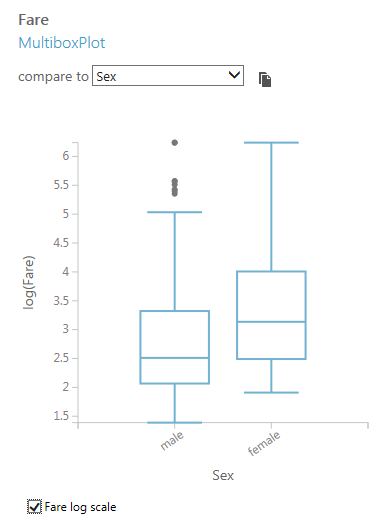


* 1. Visualizations Comparison
     1. Did gender play any part in survival? Select the “Survived” column to view the histogram of survival, and set the histogram “compare to” to “Sex”. You will notice that a disproportionate amount of males died.
     2. Is there a relationship between age and survival? Set “compare to” to “Age”.





* + 1. Compare “Sex” to “Age” to see the distribution of males and female age groups.
    2. Compare “Fare” to “Sex”. Make sure “**Fare log scale**” is checked.

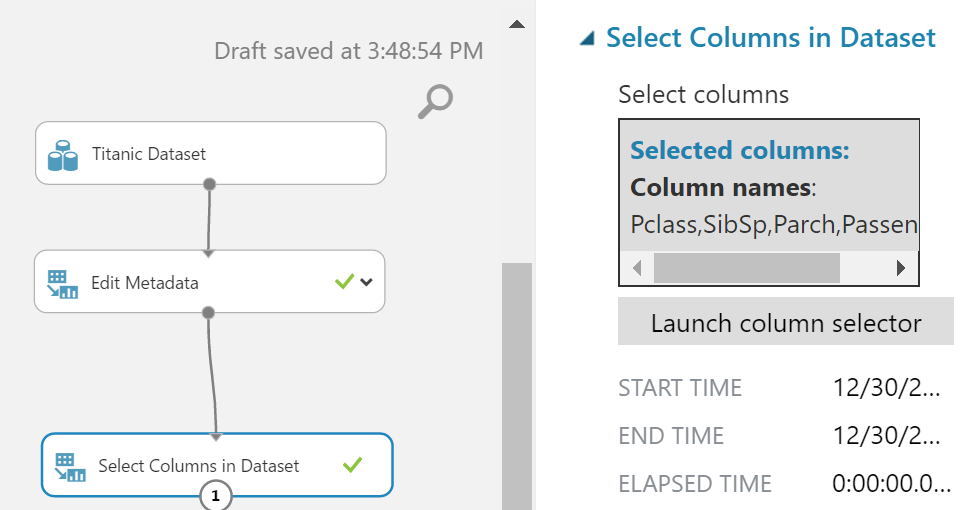
 

# **Exercise 3: Renaming Columns**

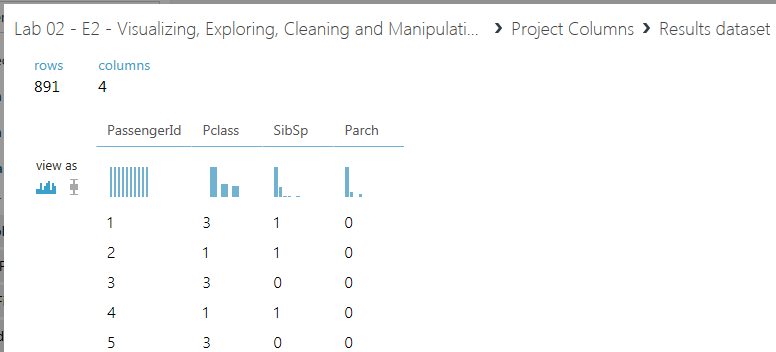
Let’s rename some undescriptive column names: “Pclass”, “SibSp” and “Parch"

* 1. Drag in a ‘**Select Columns in dataset’** Module under **Data Transformation > Manipulation**
  2. Connect the output of “**Edit Metadata**” module to the input of “**Select Columns in dataset**” module.
  3. In the Project Columns Properties, click on ‘Launch column selector’, then “With Rules” and select “Pclass”, SibSp”, “Parch” and “PassengerId”.

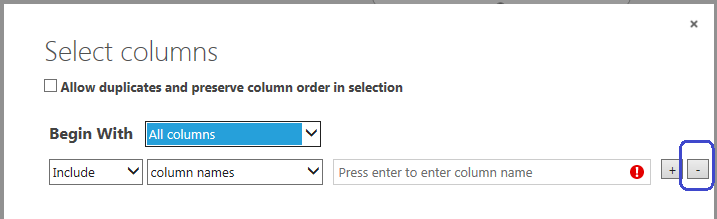


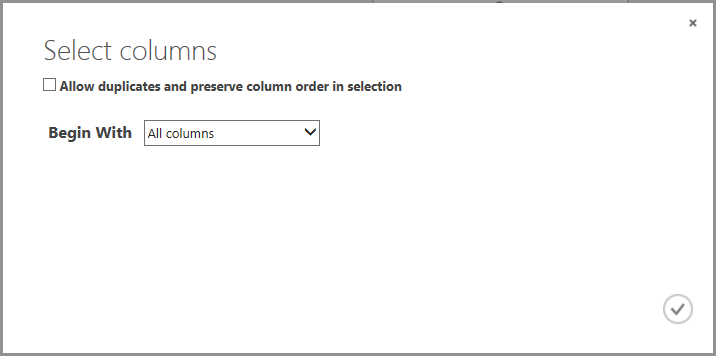


* 1. Run the experiment and then visualize the output of the ‘Select Columns’ module.

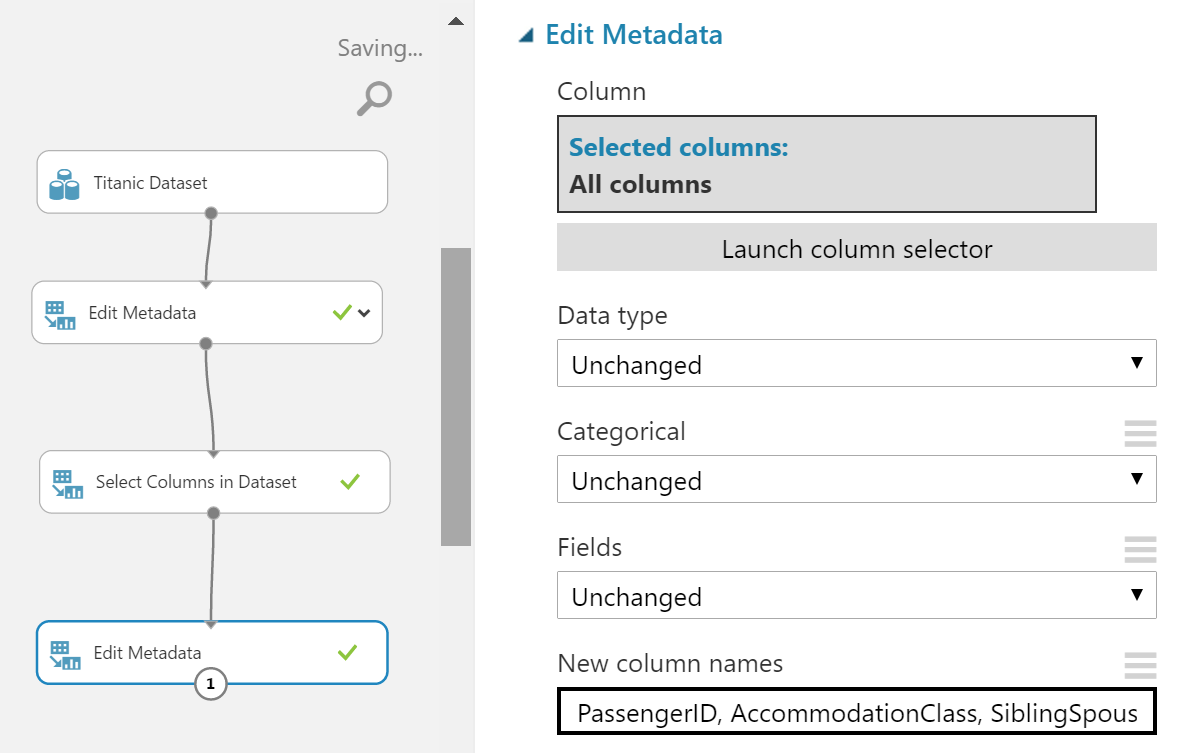


* 1. Drag in another ‘**Edit** **Metadata’** module and connect it to the “Select Columns” module.
     1. Launch the column selector.
     2. Begin with: “All Columns”
     3. Directly below “Begin With” property, there is a “+” and “-“ button. Remove the extra parameter by clicking the minus button as highlighted below. Submit the popup window.

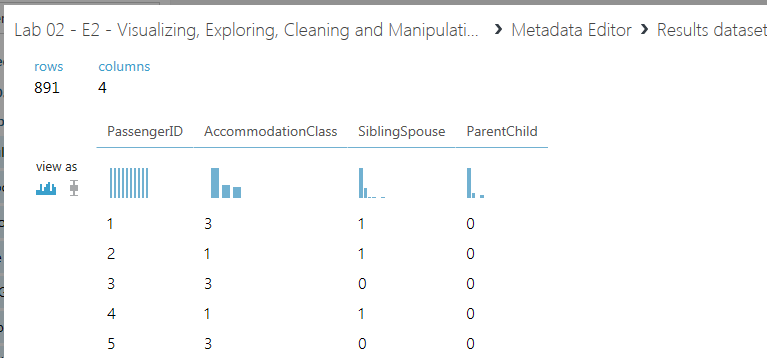




* + 1. Leave “Data type”, “Categorical” and “Fields” to “Unchanged”
    2. Under “New column names”, list/type what the new column names will be IN ORDER, PassengerID, AccommodationClass, SiblingSpouse, ParentChild



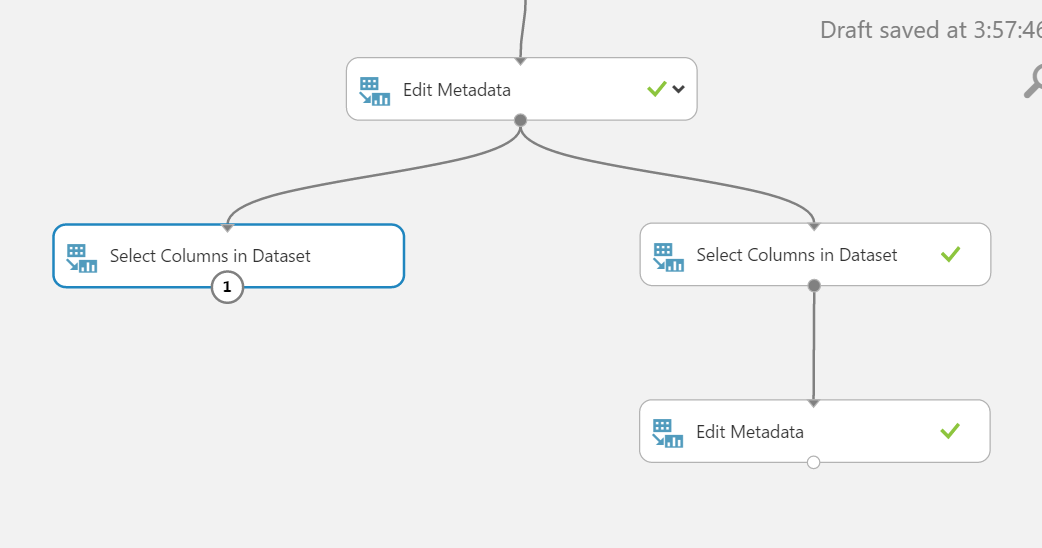
* 1. Label the modules ‘Select Columns’ and ‘Edit Metadata’ appropriately for good style.
  2. Run the experiment and then visualize the output of the Metadata Editor module to verify that the column names changed. This is what it should look like:



# **Exercise 4: Joining Tables**

From the last step, we now have an isolated version of the other table. Now we must rejoin them together.

* 1. Remove the columns that were renamed with the previous table projection.
     1. Drag in another **‘Select Columns’** and connect it to the Metadata Editor that performed the Categorical Casting. There should now be two ‘Select Columns’ side by side feeding off the same ‘Edit Metadata’.

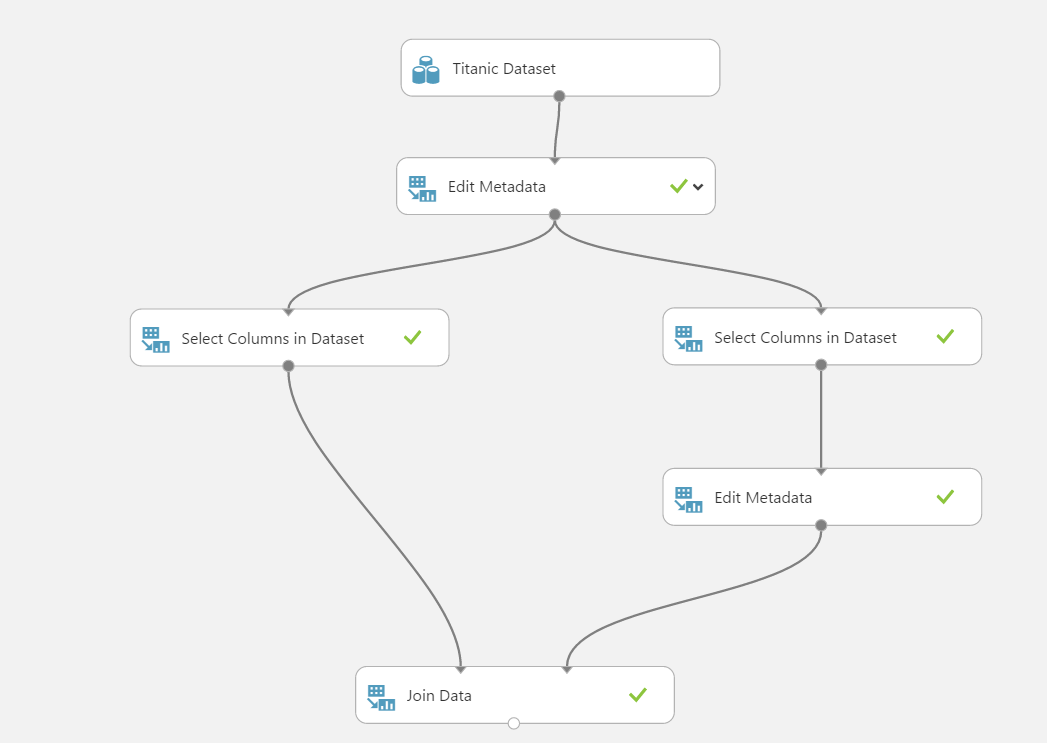


* + 1. Launch the column selector in the newly added ‘Select Columns’ module.
       1. Begin With: “All Columns”
       2. From the dropdown, change from “Include” to “Exclude”
       3. Choose to remove “Pclass”, “SibSp” and “Parch”. **Do not remove “PassengerId”** as we will need that for the join.

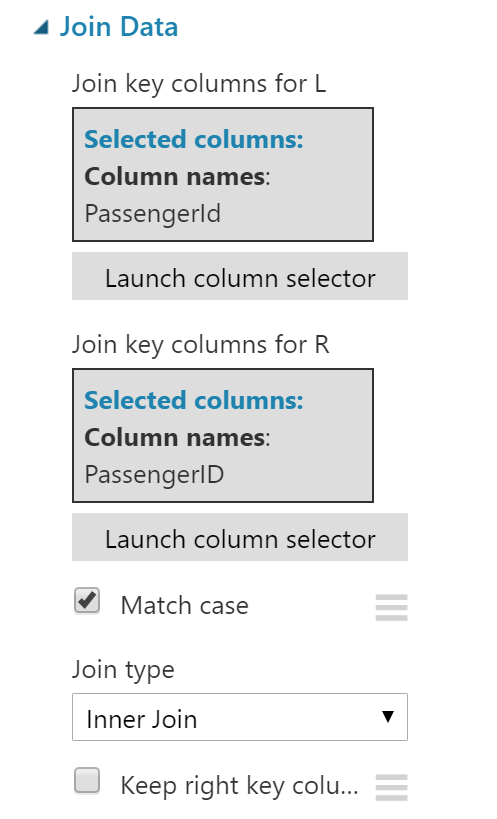


* + 1. Label the Project Columns for good style.

1. Join the two tables together.
   1. Drag in a ‘**Join Data**’ module from **Data Transformation > Manipulation**
   2. Connect the ‘Select Columns’ on the left to the ‘Edit Metadata’ on the right (the metadata editor that renamed the previous 3 columns).



* 1. In the ‘**Join Data’** module properties,
     1. **Join key columns for L 🡪 Launch column selector** 🡪 select **PassengerId**
     2. **Join key columns for R 🡪 Launch column** **selecto**r 🡪 select **PassengerID**
     3. Uncheck the box for “**Keep right key columns in joined table**”. This will remove the extra PassengerID as a result of the join.



1. Run the experiment and then visualize the output of the Join to verify that the new table has joined together properly on PassengerId. This is what it should look like:

