Azure Machine Learning

Training Course Material-

Core Labs and Exercises

Copyright ©2015 by Data Science Dojo

Azure Machine Learning Training Course Material-Core Labs and Exercises

Copyright © 2015 by Data Science Dojo

Data Science Dojo

2205 152nd Ave NE,

Redmond, WA 98052

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law.

Table of Contents

© 2015 Data Science Dojo, All Rights Reserved

Lab 1: Creating an Azure Machine Learning Studio Workspace and Familiarization

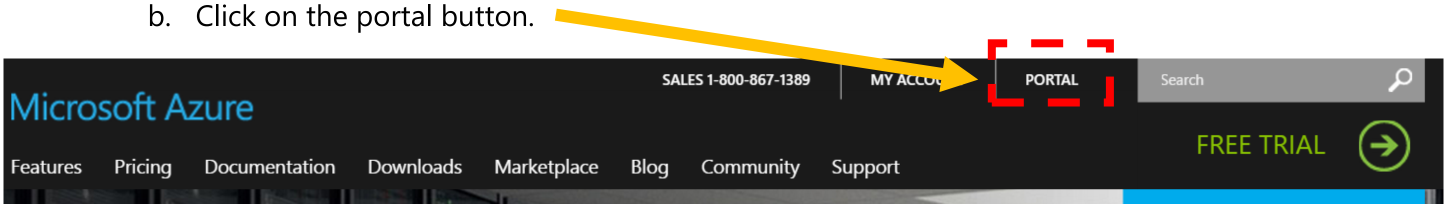
Skip this lab if you already have an Azure ML Account and know how to log in.

1. Log in using your credentials. Sign up for a free Azure trial if you do not have an account.
   a. http://azure.microsoft.com/en-us/pricing/free-trial/

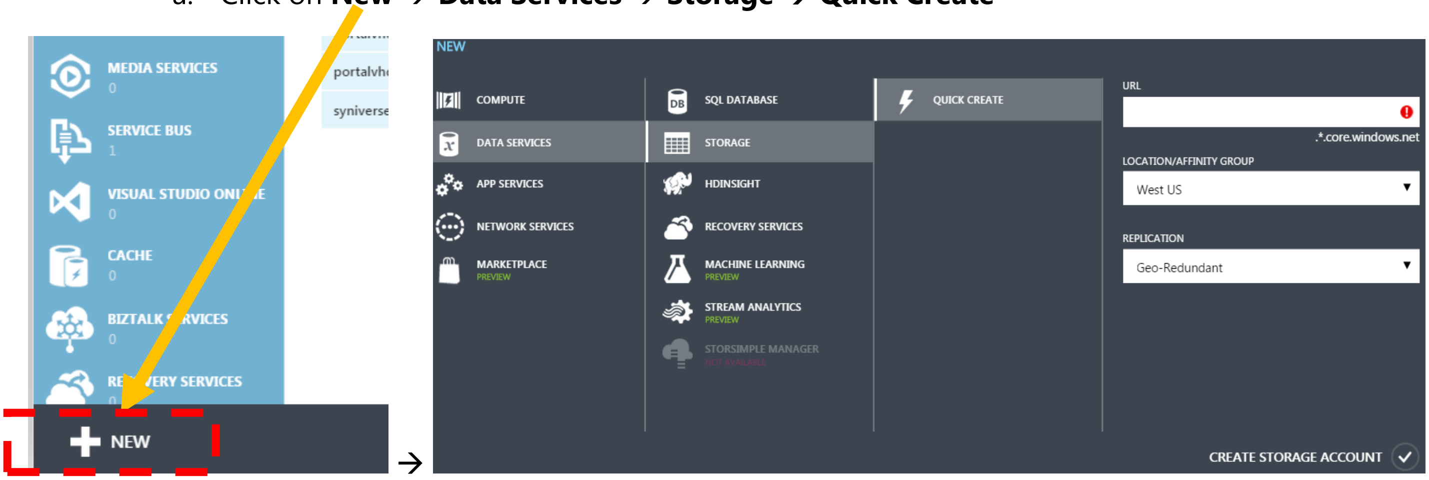
Exercise 1: Create a dedicated Azure Storage

Even if you already have an Azure Storage account. It is **highly recommended** that you created a **dedicated** Azure Storage account for your Azure Machine Learning Studio workspace.

1. Log into your Azure Management Portal.
   a. Go to: http://azure.microsoft.com/
   b. Click on the portal button.



2. Create a dedicated Azure Storage account for your Azure Machine Learning Studio workspace.
   a. Click on **New → Data Services → Storage → Quick Create**



LOCATION/AFFINITY GROUP: West US

REPLICATION: Geo-Redundant

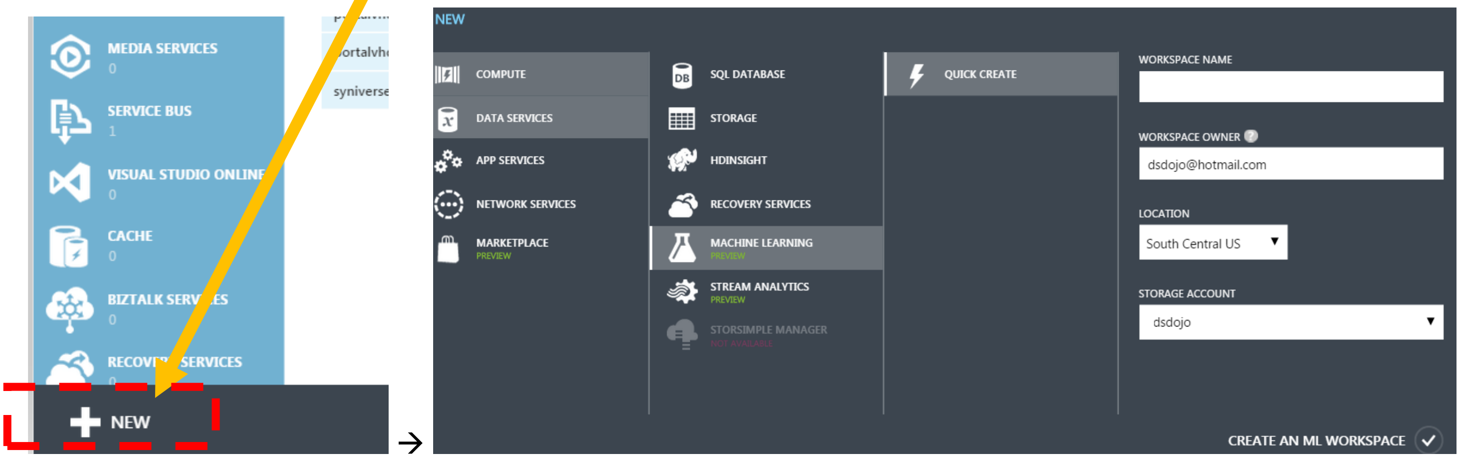
   b. Specify a URL. This will be the name as well as the HTTP URL to your storage account. Since this account will be for our Azure Machine Learning Studio account, it is best practice and **highly recommended** that you have the machinelearningstudio somewhere in the url name so there is no doubt or confusion as to what the storage account is for (so you don't accidently delete the storage in the future). The URL name will also have to be globally unique.
   c. Location/Affinity group: Select a region that's closest to you or your team of data scientists.
   d. Replication: this will control the fault tolerance of your storage. Locally or Geo-Redundancy should suffice.
   e. Hit okay and wait for your storage account to be provisioned.

© 2015 Data Science Dojo, All Rights Reserved

Exercise 2: Creating an Azure Machine Learning Studio workspace.

Once you have a dedicated Azure storage account, you can now create an Azure Machine Learning Studio workspace.

1. Provision a new Azure Machine Learning Studio workspace (if you don't't' have one already).
   a. Click on: **New → Data Services → Machine Learning → Quick Create**



WORKSPACE NAME

WORKSPACE OWNER: dsdojo@hotmail.com

LOCATION: South Central US

STORAGE ACCOUNT: dsdojo

   b. Workspace: Name your workspace. This is name has to be globally unique.
   c. Workspace Owner: set your azure account email (or someone else if you'd rather them be admin).
   d. Storage Account: reference your dedicated Azure Storage account.
2. Wait for your Machine Learning Studio workspace to be provisioned.

Tips and Notes:

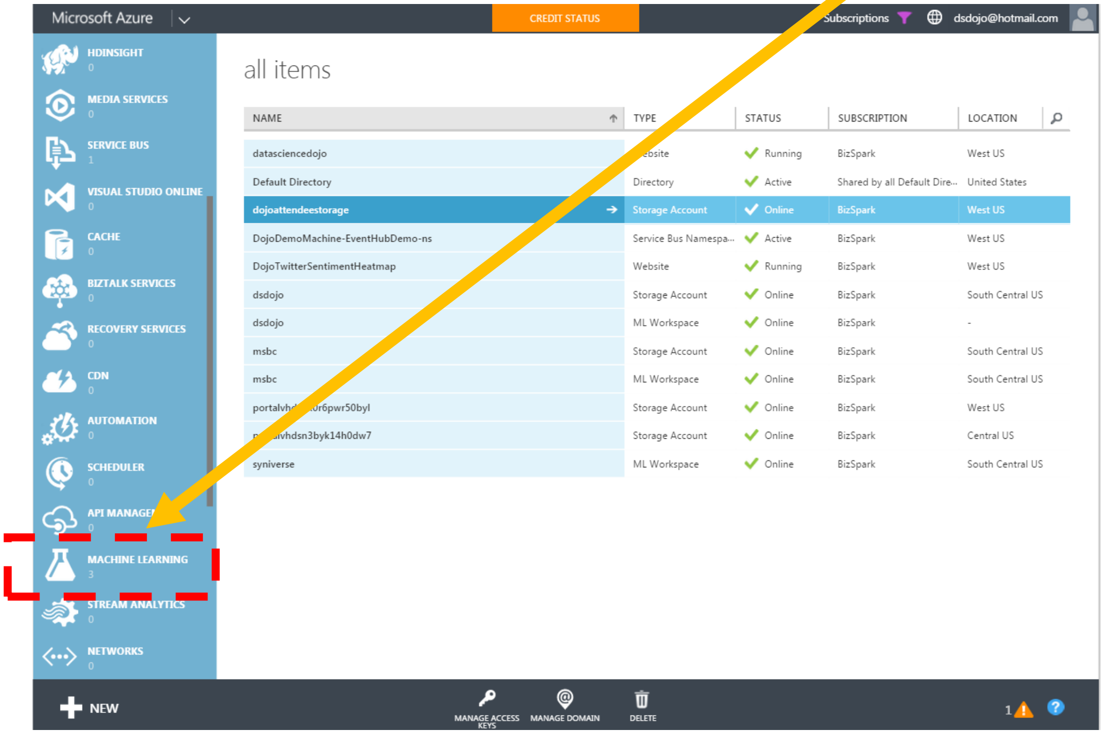
You can invite others to work on your workspace by sharing it with them. This is the closest thing that data scientists have to a collaborative tool such as google docs. You can also copy and paste experiments across workspaces.

© 2015 Data Science Dojo, All Rights Reserved

Exercise 3: Accessing your Azure Machine Learning workspace.

Once you have a dedicated Azure storage account, you can now create an Azure Machine Learning Studio workspace.

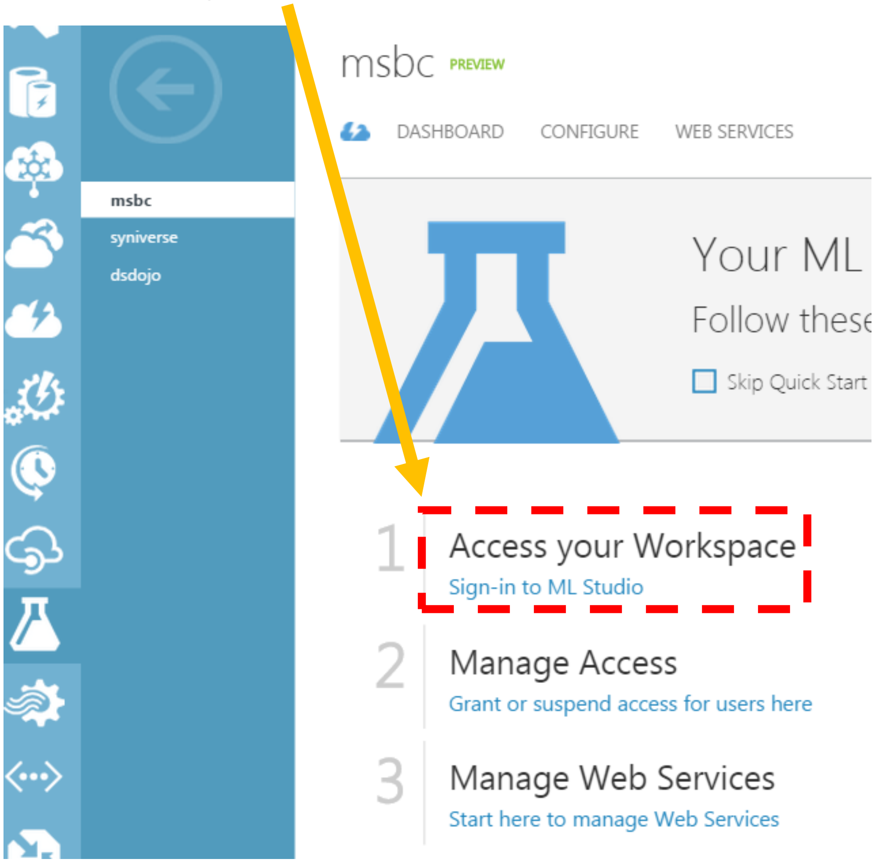
1. Within your Azure Management Portal. Click on Machine Learning.



all items

NAME	TYPE	STATUS	SUBSCRIPTION	LOCATION
datasciencedojo	ebsite	✓ Running	BizSpark	West US
Default Directory	Directory	✓ Active	Shared by all Default Dire...	United States
dojoattendeestorage	Storage Account	✓ Online	BizSpark	West US
DojoDemoMachine-EventHubDemo-ns	Service Bus Namespa...	✓ Active	BizSpark	West US
DojoTwitterSentimentHeatmap	Website	✓ Running	BizSpark	West US
dsdojo	Storage Account	✓ Online	BizSpark	South Central US
dsdojo	ML Workspace	✓ Online	BizSpark	-
msbc	Storage Account	✓ Online	BizSpark	South Central US
msbc	ML Workspace	✓ Online	BizSpark	South Central US
portalvh[illegible]0r6pwr50byl	Storage Account	✓ Online	BizSpark	West US
p[illegible]alvhdsn3byk14h0dw7	Storage Account	✓ Online	BizSpark	Central US
syniverse	ML Workspace	✓ Online	BizSpark	South Central US

2. Click on your provisioned workspace.
3. Click "Sign-in to ML studio." A new window will now populate with your ML studio.



msbc PREVIEW

Your ML

Follow these

□ Skip Quick Start

1 Access your Workspace
Sign-in to ML Studio

2 Manage Access
Grant or suspend access for users here

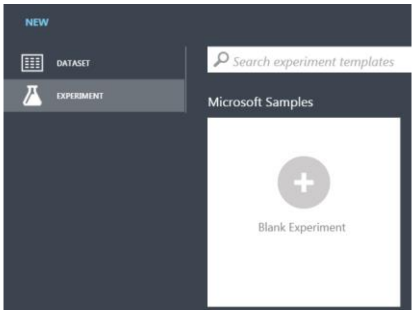
3 Manage Web Services
Start here to manage Web Services

© 2015 Data Science Dojo, All Rights Reserved

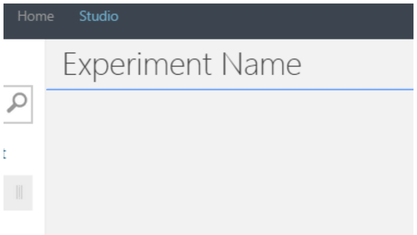
Exercise 4: Creating your first experiment.

Data Science is both an art and a science. It borrows terms from other disciplines, especially traditional sciences like chemistry. A project in data science is called an "experiment."

1. Create a new experiment
   a. Click on: **New → Experiment → Blank Experiment**



2. Name your experiment at the top.

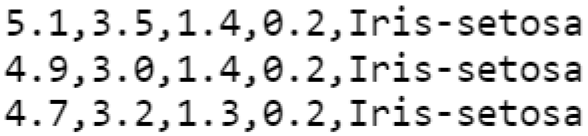


3. You won't be able to save your experiment unless you have a module inside of it. For now, move onto the next lab. After that, you will be able to save your experiment.

Lab 2: Methods of Ingress and Egress with Azure Machine Learning Studio

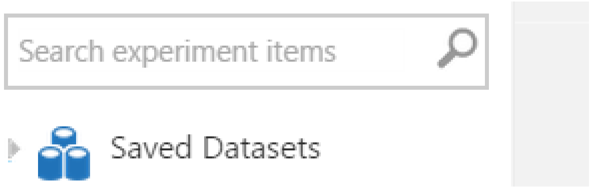
Exercise 1: Upload dataset to workspace from a local file

1. Visit: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

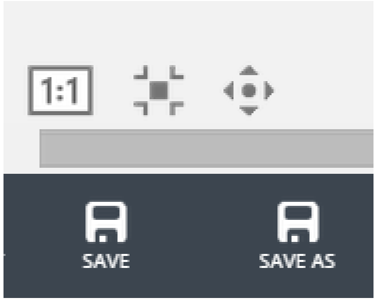


```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
```

   a. Notice how its comma delimited. That lets you know that it can be read as a CSV.
   b. Excel Files (xls) and delimited text files can be read as a CSV.
   c. Notice that this data does not have headers. We will define the headers later, as the model will require it.
2. Download and save the file as a CSV file. 'file-name.csv'.
3. Import the dataset into Azure ML Studio.
   a. Click on: **New → DATASET → FROM LOCAL FILE**
   b. Import as a new dataset.
   c. Please note that by default Azure ML ships with a dataset called "Iris Two Class Data". Do not confuse that with the data you just imported, so give it a unique name.
4. Go into any experiment and verify that the dataset has been imported.
   a. The data will be under a directory called "Saved Datasets" within any experiment.



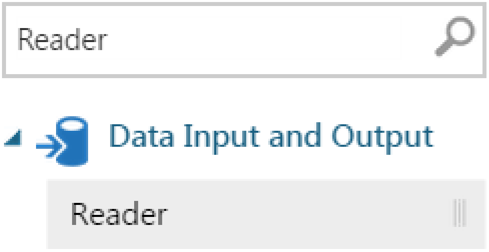
5. Now that your experiment has some modules in it, you are able to save your experiment. Save your experiment with "save as" on the menus at bottom.



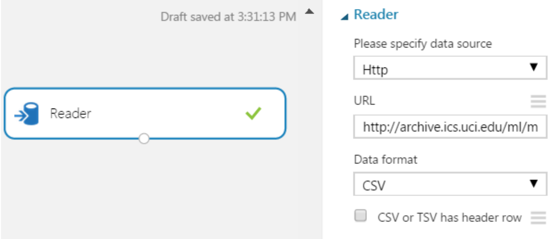
© 2015 Data Science Dojo, All Rights Reserved

Exercise 2: Reading a data source through http

1. Drag and drop a Reader module from the menu on the left.



2. Specify Http as the data source.
3. In the URL field, enter http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
4. Choose CSV from the Data Format drop down
5. Leave the 'CSV or TSV has header row' option unchecked



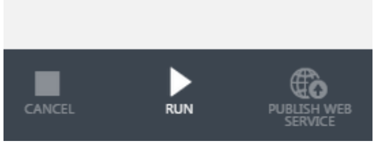
Please specify data source: Http

URL: http://archive.ics.uci.edu/ml/m

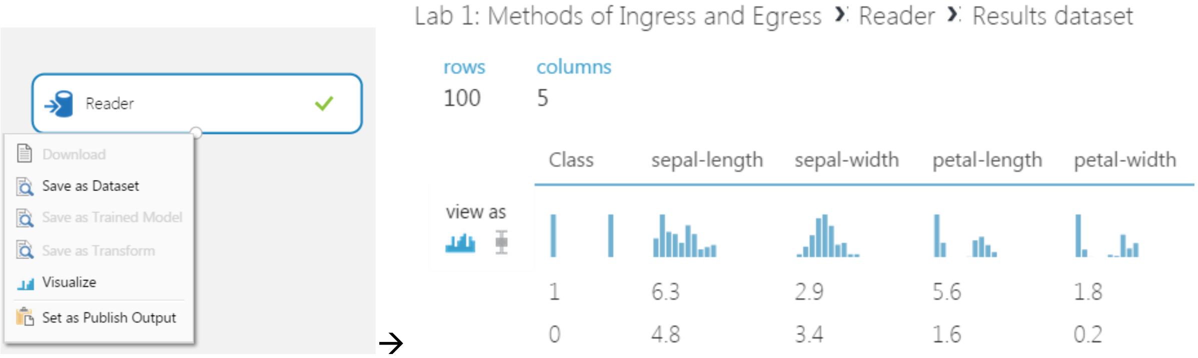
Data format: CSV

□ CSV or TSV has header row

6. Leave Disable Upgrades unchecked (Any thoughts why this would matter?)
7. Run the experiment to execute the import and parse.



8. Preview the data by visualizing the output of the reader module. Right click the bottom middle node of the reader module to access the menu.

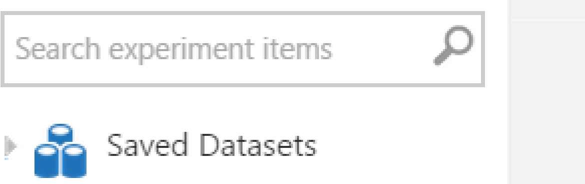


Lab 1: Methods of Ingress and Egress › Reader › Results dataset

rows	columns
100	5

Class	sepal-length	sepal-width	petal-length	petal-width
1	6.3	2.9	5.6	1.8
0	4.8	3.4	1.6	0.2

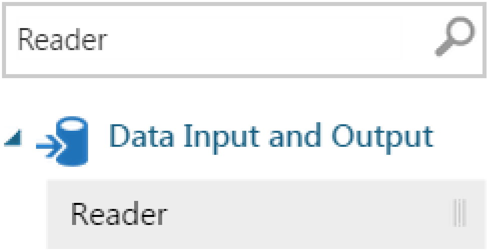
9. The data is not yet saved. Save it to the workspace by clicking "save as Dataset" in the reader's output node.
10. Go into any experiment and verify that the dataset has been imported.
    a. The data will be under a directory called "Saved Datasets" within any experiment.



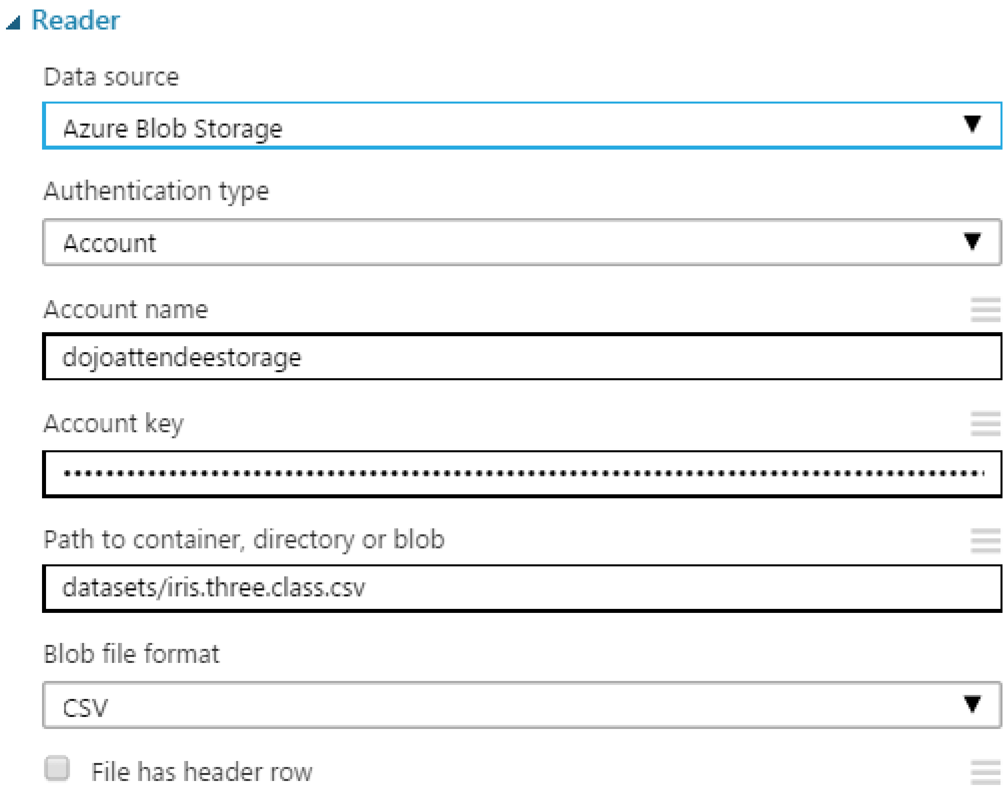
    b. Please note that by default Azure ML ships with a dataset called "Iris Two Class Data." Do not confuse that with the data you just imported.

Exercise 3: Reading a Data Set from Azure Blob Storage

1. Drag and drop a Reader module from the menu on left.



2. Input the sample blob storage account we have set up for you.
   a. **Data Source:** AzureBlobStorage
   b. **Authentication type:** Account
   c. **Account:** dojoattendeestorage
   d. **Account Key:** aKQOxU3As1BsS3yT2bhHkJ/icCICJPpL1tdWKxQ+tPBNk6DbykV4qd3HGlFPZN/3TdiUHuM/Quk9DPUeQu7M8A==
   e. **Path to container, directory:** datasets/iris.three.class.csv
      i. Note that dojoattendeestorage is the container name. Storage vaults have containers, and containers have blobs: **Storage>Container>Blob**. In this case it's: **dojoattendeestorage>datasets>iris.three.class.csv**
      ii. A blob is really just a file that's in the Azure Cloud itself. If you've done web development, this looks exactly like an FTP.
   f. **Blob file format:** CSV
   g. **File has header row:** Unchecked


Reader

Data source: Azure Blob Storage

Authentication type: Account

Account name: dojoattendeestorage

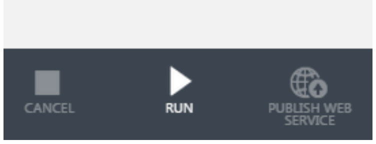
Account key: ••

Path to container, directory or blob: datasets/iris.three.class.csv

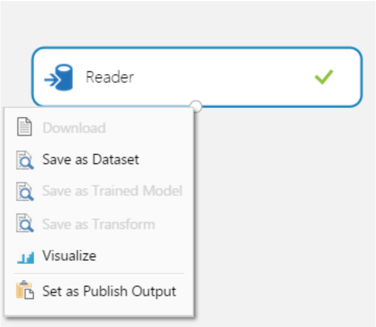
Blob file format: CSV

□ File has header row

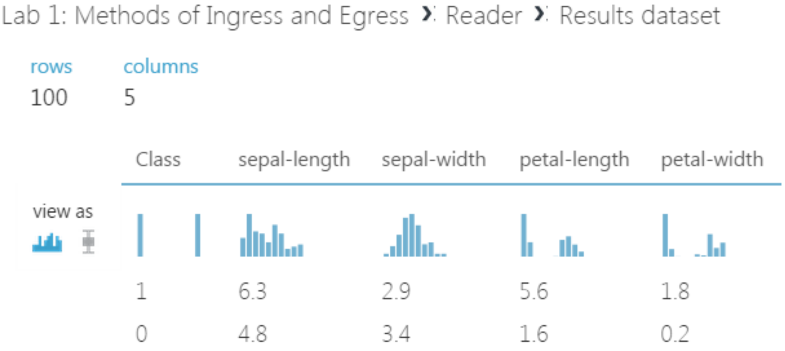
3. Run the experiment to execute the import and parse.



4. Preview the data by visualizing the output of the reader module. Right click the bottom middle node of the reader module to access the menu.



→



Lab 1: Methods of Ingress and Egress › Reader › Results dataset

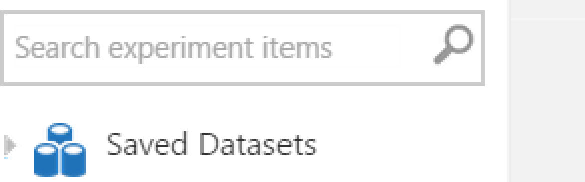
rows 100

columns 5

view as

Class	sepal-length	sepal-width	petal-length	petal-width
1	6.3	2.9	5.6	1.8
0	4.8	3.4	1.6	0.2

5. The data is not yet saved. Save it to the workspace by clicking "save as Dataset" in the reader's output node.
6. Go into any experiment and verify that the dataset has been imported.
   a. The data will be under a directory called "Saved Datasets" within any experiment.

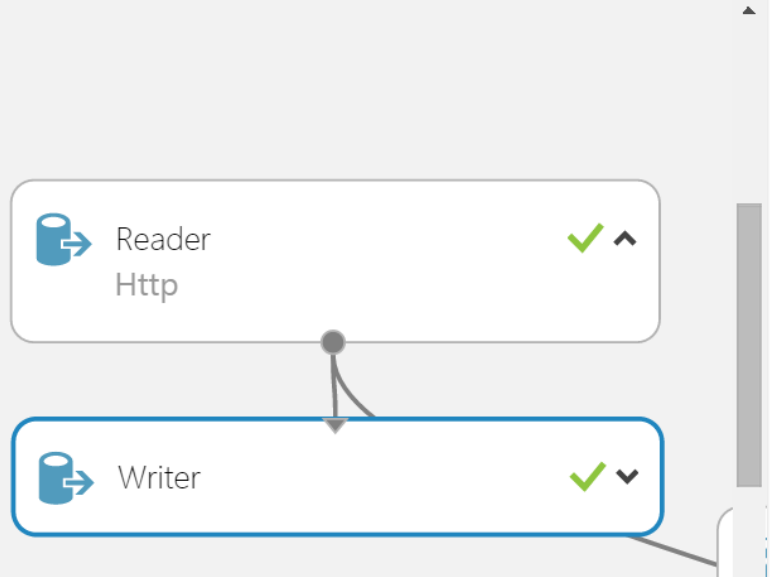


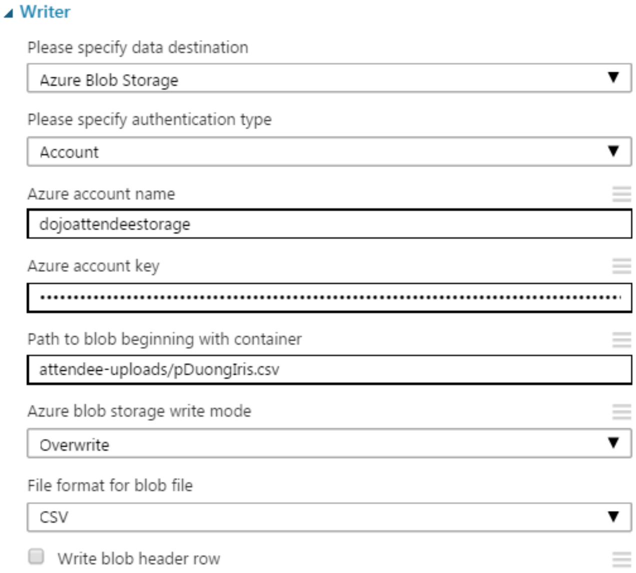
   b. Please note that by default Azure ML ships with a dataset called "Iris Two Class Data". Do not confuse that with the data you just imported.

© 2015 Data Science Dojo, All Rights Reserved

Exercise 4: Writing Datasets to an Azure Blob Storage

1. Drag any dataset into your workspace.
2. Drag a writer module in and connect the dataset to the writer.
3. Input the sample blob storage account we have set up for you.
    a. **Data Source:** AzureBlobStorage
    b. **Authentication type:** Account
    c. **Account:** dojoattendeestorage
    d. **Account Key:** aKQOxU3As1BsS3yT2bhHkJ/icCICJPpL1tdWKxQ+tPBNk6DbykV4qd3HGlFPZN/3TdiUHuM/Quk9DPUeQu7M8A==
    e. **Path to container, directory:** attendee-uploads/<file-name>.csv
        i. Normally you can name your file whatever you want. However since a lot of people will be writing to this blob, **do not name it iris.csv.** Name it a combination of your first initial, then your last name, then iris as one word. For example, if your name was John Smith, name it jSmithIris.csv. This will keep our azure storage account somewhat manageable and neater.
    f. **Azure Blob storage account write mode:** Overwrite
        i. Write mode of error will return an error if the filename already exists on the storage account.
    g. **Blob file format:** CSV
    h. **Write blob header row:** Unchecked





Writer

Please specify data destination
Azure Blob Storage ▼

Please specify authentication type
Account ▼

Azure account name
dojoattendeestorage

Azure account key
••

Path to blob beginning with container
attendee-uploads/pDuongIris.csv

Azure blob storage write mode
Overwrite ▼

File format for blob file
CSV ▼

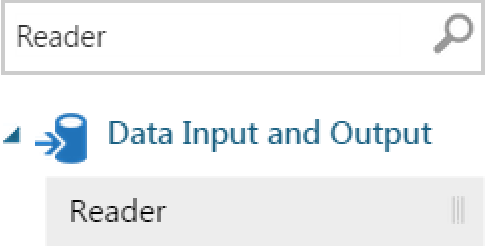
□ Write blob header row

© 2015 Data Science Dojo, All Rights Reserved

Exercise 5: Reading Hive Tables into Azure ML

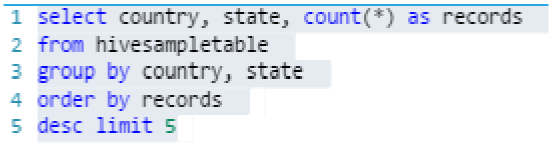
This lab will assume that you already have a Hadoop cluster setup. Please refer to the HDInsight lab on how to create a Hadoop cluster.

1. Drag and drop a Reader module from the menu on left.



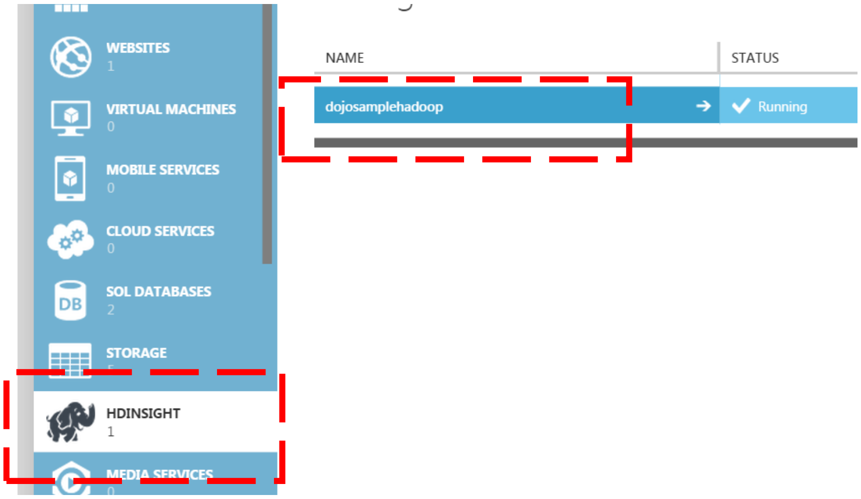
2. Populate the required fields using the values below. Open up a second window or tab to retrieve the other information.
   a. **Data Source:** HiveQuery
   b. **Hive database query:** Below is a sample query statement that you may perform. However, any of the query statements you made in the Hive lab will also work.

```
select country, state, count(*) as records
from hivesampletable
group by country, state
order by records
desc limit 5
```



```
select country, state, count(*) as records
from hivesampletable
group by country, state
order by records
desc limit 5
```

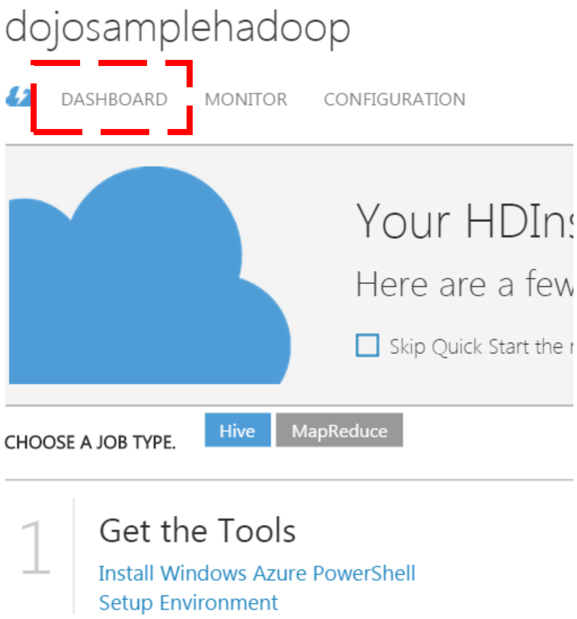
   c. **HCatalog server URI:**
      i. From your azure management portal (open this in a new window or tab): https://manage.windowsazure.com/
      ii. Click on HDInsight and your Hadoop Cluster to enter your Hadoop cluster dashboard.



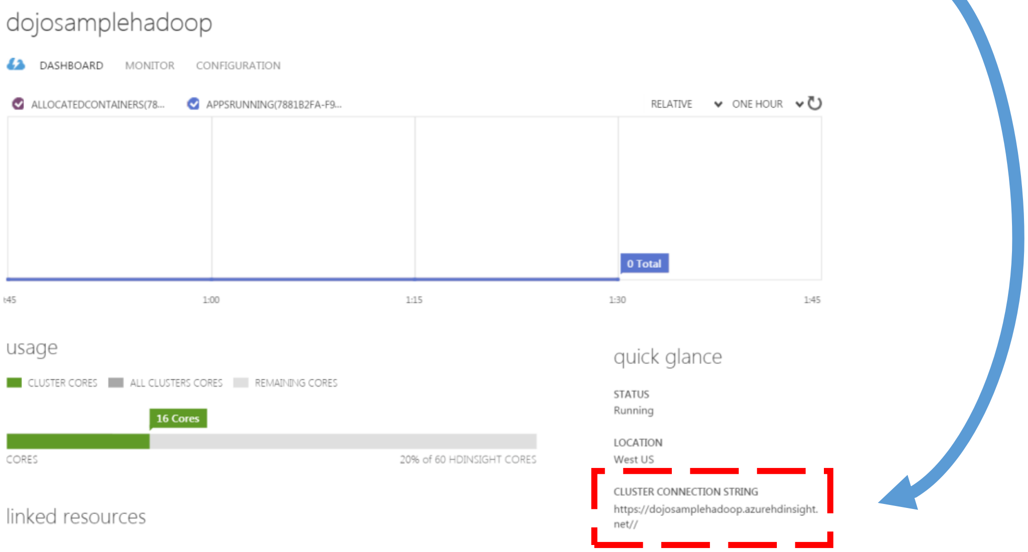
NAME	STATUS
dojosamplehadoop	✓ Running

© 2015 Data Science Dojo, All Rights Reserved

iii. Click on “Dashboard.”



iv. Retrieve your cluster connection string, located on the bottom right hand corner of the dashboard screen.



STATUS
Running

LOCATION
West US

CLUSTER CONNECTION STRING
https://dojosamplehadoop.azurehdinsight.net//

In the above example it is ‘https://dojosamplehadoopcluster.azurehdinsight.net//’

v. Copy the cluster connection string into your Azure ML reader module’s HCatalog server URI section.

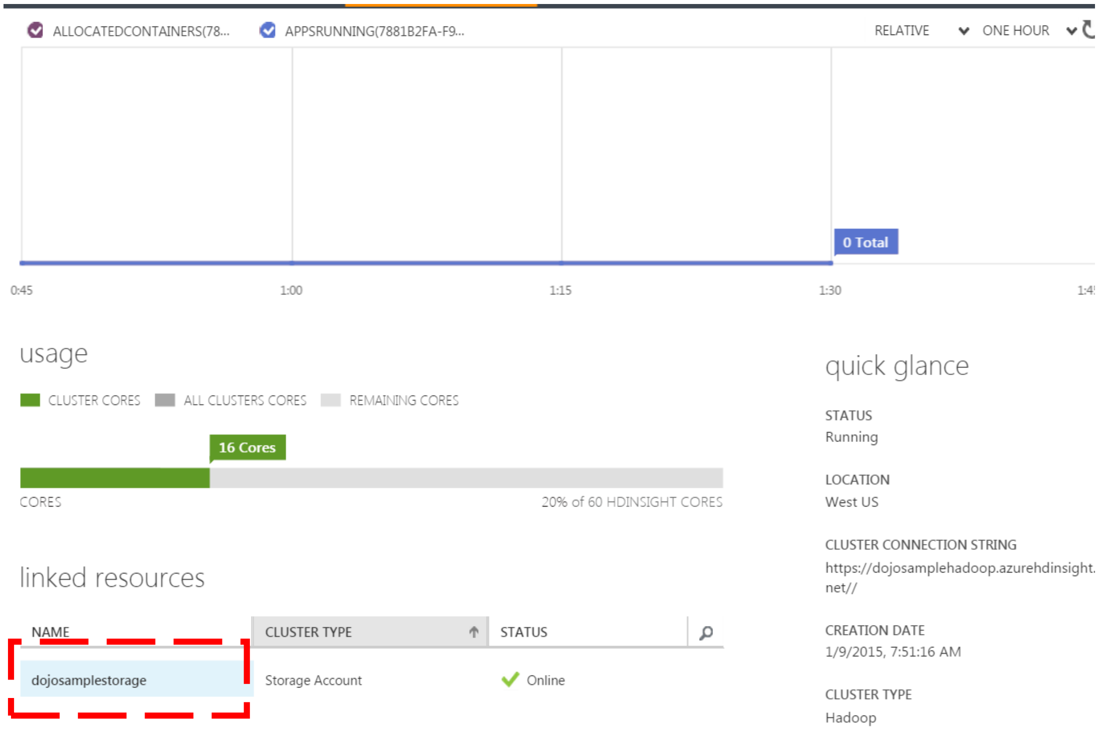
d. **Hadoop user account name:** admin

e. **Hadoop user account password:** The password you used when you provisioned your HDInsight Hadoop cluster.

f. **Location of output data:** Azure

© 2015 Data Science Dojo, All Rights Reserved

g. **Azure storage account name:** To find this detail, simply scroll down on the Hadoop cluster dashboard page (the same place you got your Hadoop cluster connection string from). It will be under a table called “linked resources”.



STATUS
Running

LOCATION
West US

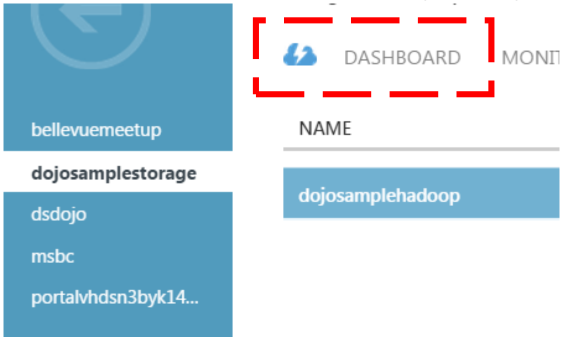
CLUSTER CONNECTION STRING
https://dojosamplehadoop.azurehdinsight.net//

CREATION DATE
1/9/2015, 7:51:16 AM

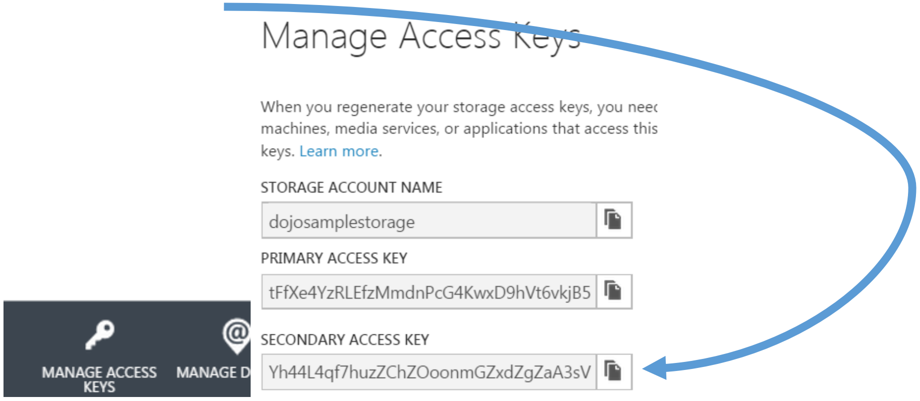
CLUSTER TYPE
Hadoop

NAME	CLUSTER TYPE	STATUS
dojosamplestorage	Storage Account	✓ Online

i. Click on the cluster and it’ll take you to your storage page. Click on dashboard.



ii. Once inside, click “manage access keys” in the bottom middle of the page. You can click the copy button to the right of the read only input boxes.



Manage Access Keys

When you regenerate your storage access keys, you neec
machines, media services, or applications that access this
keys. Learn more.

STORAGE ACCOUNT NAME
dojosamplestorage

PRIMARY ACCESS KEY
tFfXe4YzRLEfzMmdnPcG4KwxD9hVt6vkjB5

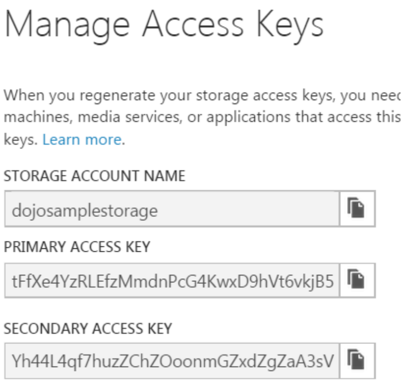
SECONDARY ACCESS KEY
Yh44L4qf7huzZChZOoonmGZxdZgZaA3sV

iii. Copy the storage name into your Azure ML reader module.

© 2015 Data Science Dojo, All Rights Reserved

h. **Azure storage key:**

   i. You can copy either of the access keys given in the manage access keys window. Ideally the primary access key is the key used for important web services and is highly guarded to prevent changes (or it could break live web services that depend on this storage account, such as a Hadoop cluster). The secondary key is the one given out for others to use and can be regenerated often.



Manage Access Keys

When you regenerate your storage access keys, you neec machines, media services, or applications that access this keys. Learn more.

STORAGE ACCOUNT NAME

dojosamplestorage

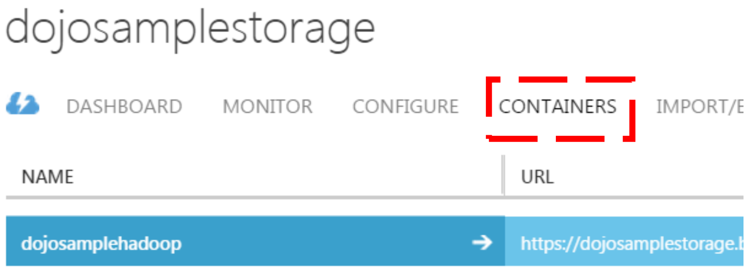
PRIMARY ACCESS KEY

tFfXe4YzRLEfzMmdnPcG4KwxD9hVt6vkjB5

SECONDARY ACCESS KEY

Yh44L4qf7huzZChZOoonmGZxdZgZaA3sV

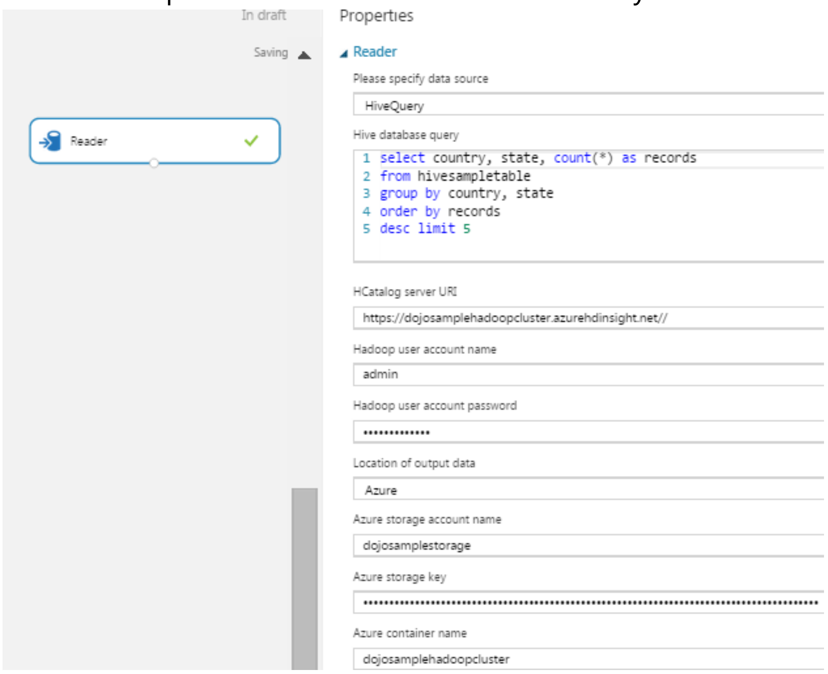
i. **Azure Container:** Your container name should actually be the same name as your Hadoop cluster. Just in case, you can check it by clicking on "containers" at the top.



dojosamplestorage

NAME	URL
dojosamplehadoop	https://dojosamplestorage.t

j. This is a sample of what it should look like when you're done.



Properties

Reader

Please specify data source

HiveQuery

Hive database query

```
select country, state, count(*) as records
from hivesampletable
group by country, state
order by records
desc limit 5
```

HCatalog server URI

https://dojosamplehadoopcluster.azurehdinsight.net//

Hadoop user account name

admin

Hadoop user account password

•••••••••••••

Location of output data

Azure

Azure storage account name

dojosamplestorage

Azure storage key

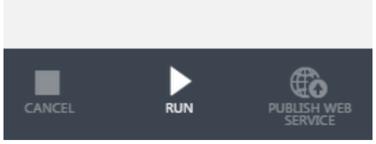
••

Azure container name

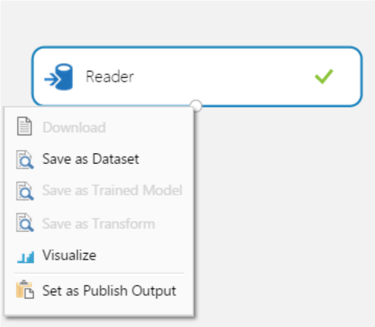
dojosamplehadoopcluster

© 2015 Data Science Dojo, All Rights Reserved

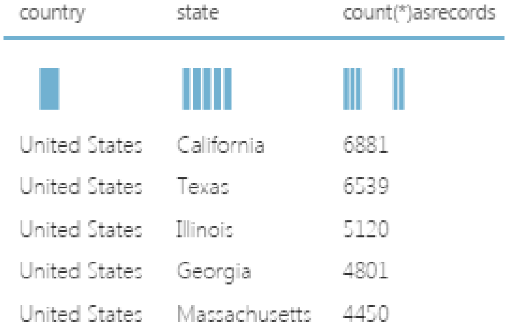
3. Run the experiment to execute the import and parse.



4. Preview the data by visualizing the output of the reader module. Right click the bottom middle node of the reader module to access the menu.



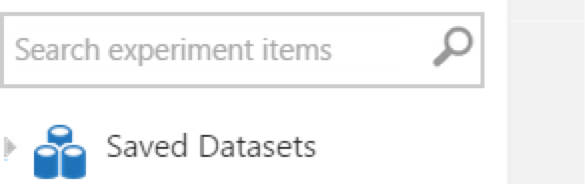
→



country	state	count(*)asrecords
United States	California	6881
United States	Texas	6539
United States	Illinois	5120
United States	Georgia	4801
United States	Massachusetts	4450

The data is not yet saved to Azure ML. Save it to the workspace by clicking “Save as Dataset” in the reader’s output node.

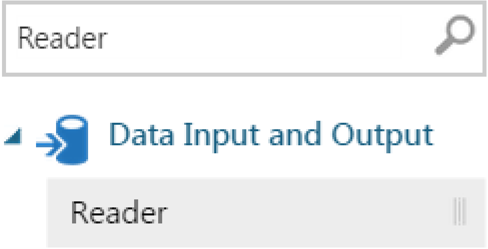
5. Go into any experiment and verify that the dataset has been imported.
   a. The data will be under a directory called “Saved Datasets” within any experiment.



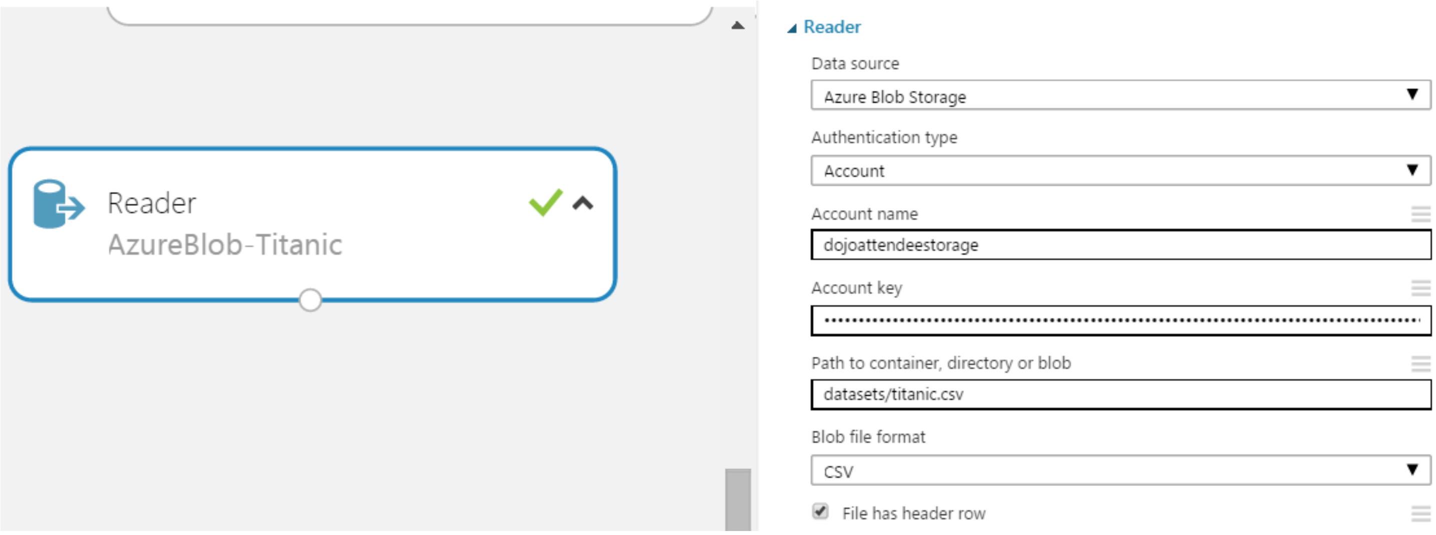
Lab 3: Visualizing, Exploring, Cleaning, and Manipulating Data

Exercise 1: Obtain the Titanic sample data

1. Drag and drop a Reader module from the menu on left.



2. Input the sample blob storage account we have set up for you.
   a. **Data Source:** AzureBlobStorage
   b. **Authentication Type:** Account
   c. **Account:** dojoattendeestorage
   d. **Account Key:** 7zXrsCGmM5LKIgZG6a4UU6LPPzVab70JdxuVSwXMRFsfQQW48RVh6GERv1/fsR5DBSAKWCXhyG dngXI6pyA4SQ==
   e. **Path to container, directory:** datasets/titanic.csv
   f. **Blob file format:** CSV
   g. **File has header row:** Checked



Reader

Data source
Azure Blob Storage

Authentication type
Account

Account name
dojoattendeestorage

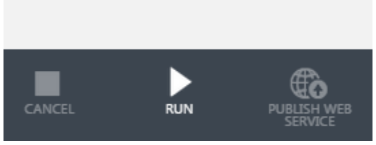
Account key
••

Path to container, directory or blob
datasets/titanic.csv

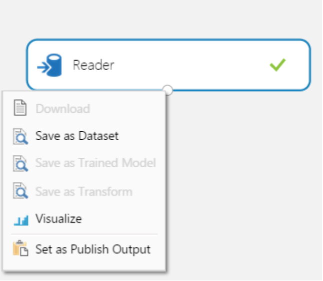
Blob file format
CSV

☑ File has header row

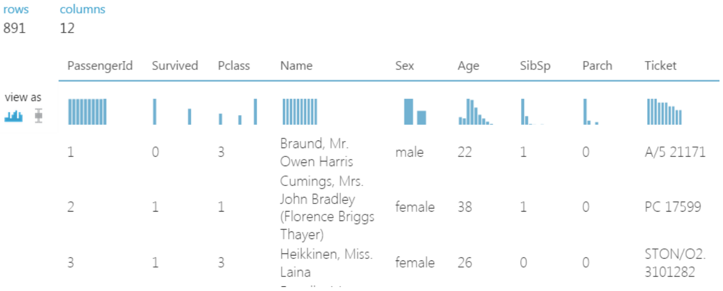
3. Run the experiment to execute the import and parse.



4. Preview the data by visualizing the output of the reader module. Right click the bottom middle node of the reader module to access the menu.



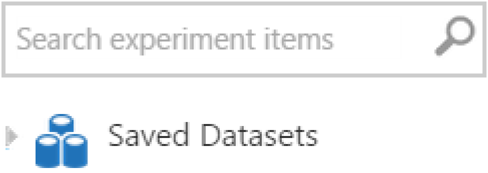
→

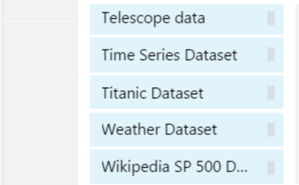


rows 891 columns 12

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282

5. The data is not yet saved. Save it to the workspace by clicking "Save as Dataset" in the reader's output node.
6. Go into any experiment and verify that the dataset has been imported.
   a. The data will be under a directory called "Saved Datasets" within any experiment





© 2015 Data Science Dojo, All Rights Reserved

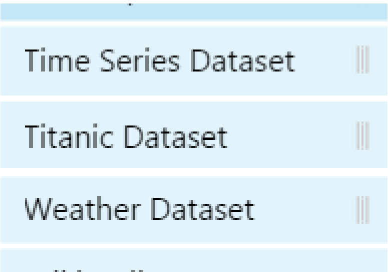
Titanic Dataset Key:

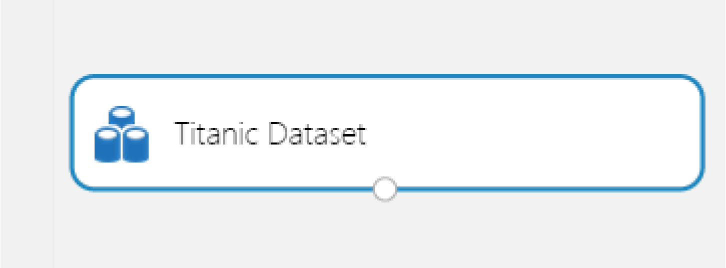
```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

Exercise 2: Casting Columns

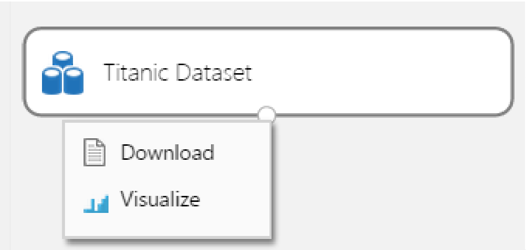
This data contains categorical data types. However, we must tell Azure which of those columns are categorical so that our models will not treat them as sequential numbers.

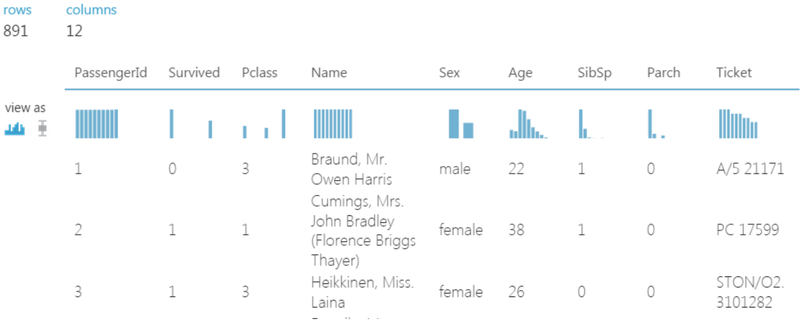
1. Drag the Titanic dataset into the experiment workspace.





2. Verify that it's the same looking data as below by visualizing the Titanic dataset.



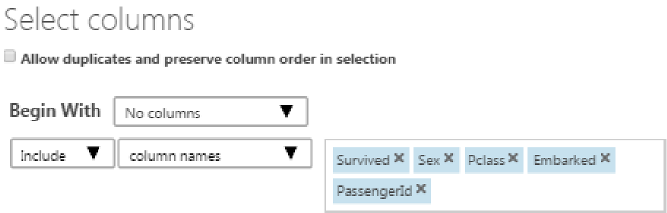


rows 891 columns 12

view as

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282

3. Cast categorical values to categorical.
   a. Drag in the metadata editor module.
   b. Launch the column selector.
   c. Begin with "no columns" and "include" "column names"
      i. Add "survived," "sex," "pclass," "embarked," and "PassengerID"



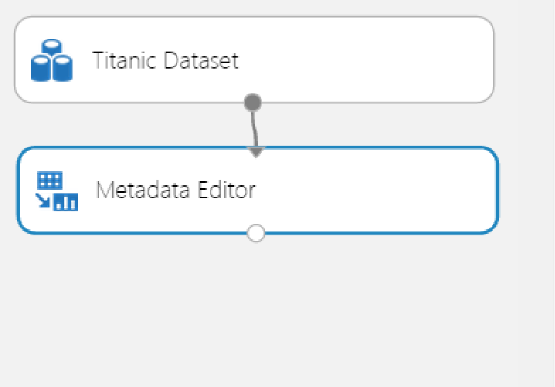
Select columns

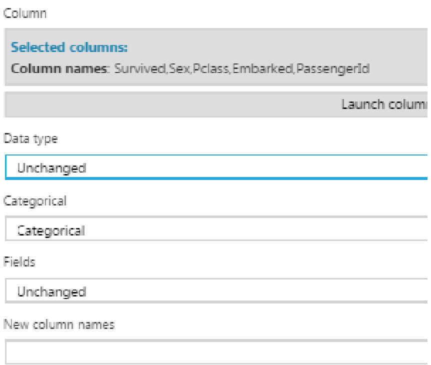
□ Allow duplicates and preserve column order in selection

Begin With: No columns

Include | column names | Survived ✕ Sex ✕ Pclass ✕ Embarked ✕ PassengerId ✕

   d. Set categorical to "categorical"
   e. Label the module as "Categorical Casting" for good style.





Column

Selected columns:
Column names: Survived,Sex,Pclass,Embarked,PassengerId

Launch colum

Data type: Unchanged

Categorical: Categorical

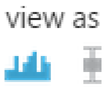
Fields: Unchanged

New column names:

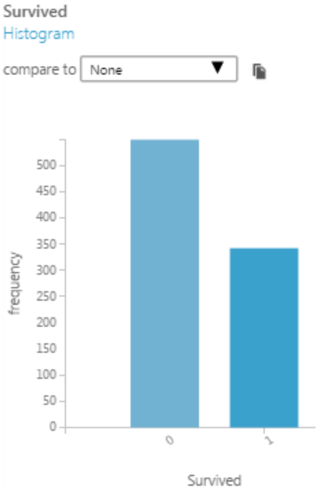
© 2015 Data Science Dojo, All Rights Reserved

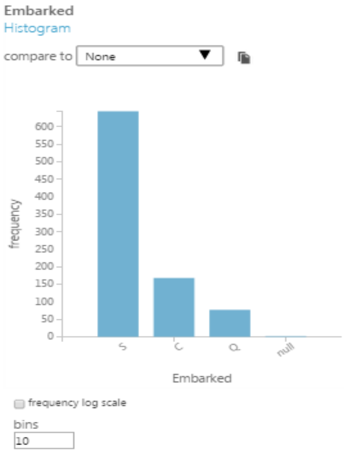
Exercise 3: Data Visualization & Exploration

1. Histograms
   a. Click on the survived column with "view as" set to the picture of a histogram.



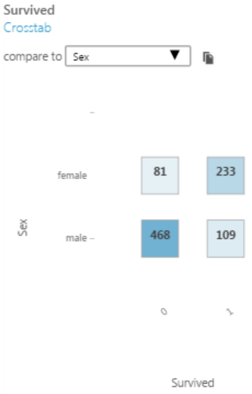
   b. A menu on the right will pop up. Expand the visualizations drop down. A histogram will now appear.
   c. What is the distribution of survived vs deceased? Did more people survive or perish?
   d. Where did people come from? Select the "embarked" column.


Survived
Histogram
compare to None


Embarked
Histogram
compare to None

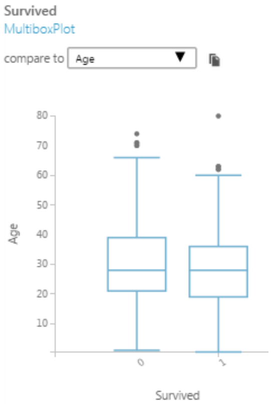
□ frequency log scale
bins 10

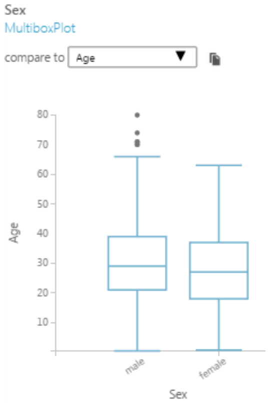
2. Comparison Visualizations
   a. Did gender play any part in survival? Select the "survived" column to view the histogram of survival, and set the histogram "compare to" to "sex". You will notice that a disproportionate amount of males died.
   b. Is there a relationship between age and survival? Set "compare to" to "age".
   c. Compare "Age" to "Sex" to see the distribution of males and female age groups.
   d. Compare "Fare" to "Sex". Make sure **"log scale"** is checked.

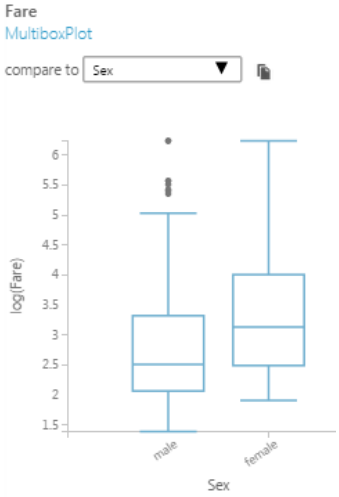

Survived
Crosstab
compare to Sex

Sex	0	1
female	81	233
male	468	109

Survived


Survived
MultiboxPlot
compare to Age


Sex
MultiboxPlot
compare to Age

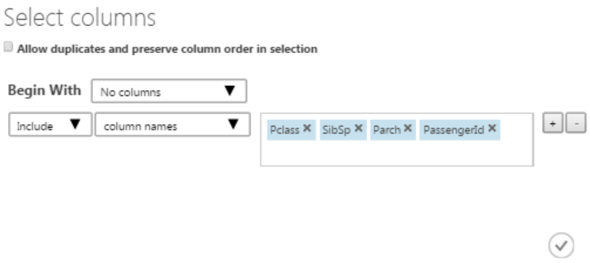

Fare
MultiboxPlot
compare to Sex

© 2015 Data Science Dojo, All Rights Reserved

Exercise 4: Renaming Your Columns

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

1. Drag in a **Project Columns Module** under **Data Transformation > Manipulation**
   a. Launch the column selector and select “PassengerID,” “Pclass,” SibSP,” and “Parch”

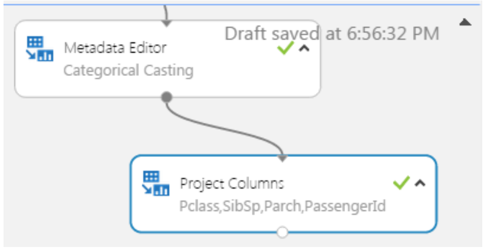


Select columns

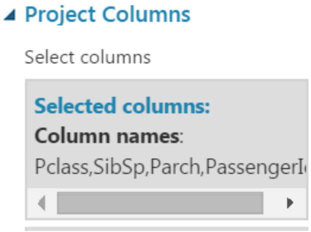
□ Allow duplicates and preserve column order in selection

Begin With: No columns

Include | column names | Pclass × SibSp × Parch × PassengerId ×



Draft saved at 6:56:32 PM

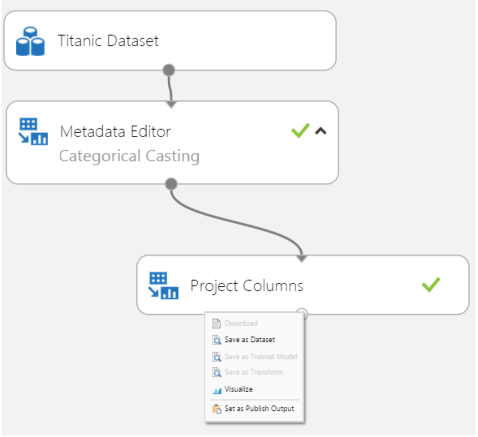


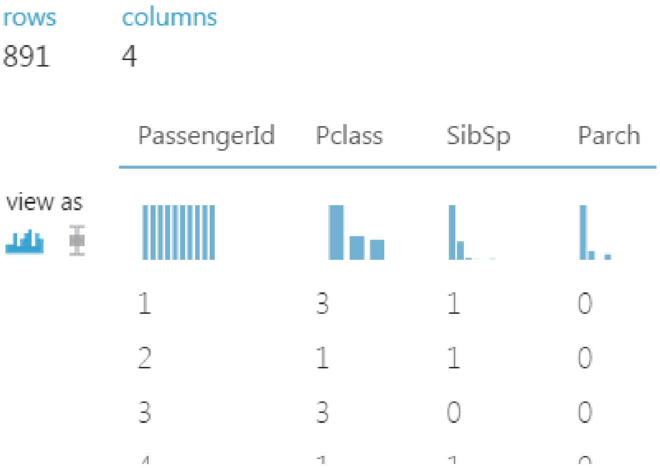
Project Columns

Select columns

Selected columns:
Column names:
Pclass,SibSp,Parch,PassengerI

   b. Run then visualize the output of the project columns module. This is what it should look like:



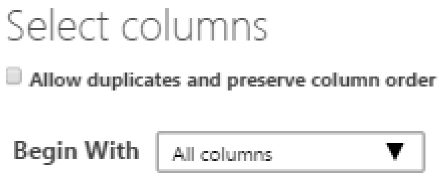


rows 891

columns 4

PassengerId	Pclass	SibSp	Parch
1	3	1	0
2	1	1	0
3	3	0	0
4	1	1	0

2. Drag in another metadata editor, and connect it to the project columns module.
   a. Launch the column selector.
   b. Begin with: “All Columns”
   c. Directly below “begin with,” there is a “+” and “-“ button. Remove the extra parameter by clicking the minus button. Submit the popup window.



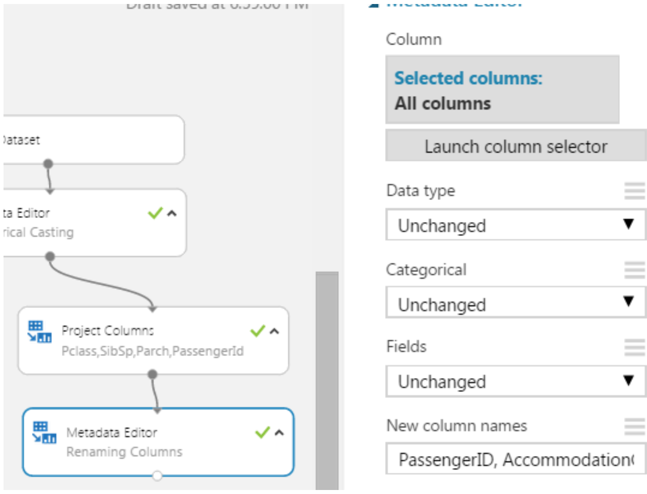
Select columns

□ Allow duplicates and preserve column order in selection

Begin With: All columns

   d. Leave “data type,” “categorical,” and “fields,” to “unchanged”
   e. Under new column names, list IN ORDER what the new column names will be.

i. "PassengerID, AccommodationClass, SiblingSpouse, ParentChild"



Column

Selected columns:
All columns

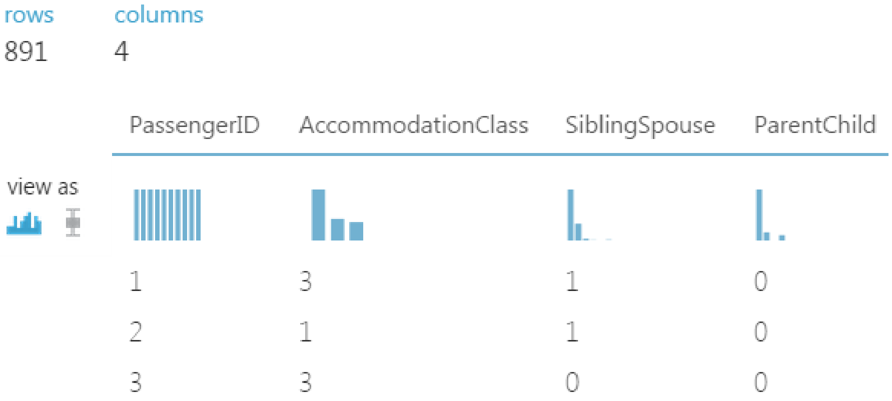
Data type: Unchanged

Categorical: Unchanged

Fields: Unchanged

New column names: PassengerID, Accommodation(

ii. Visualize the output of the metadata editor to verify that the column names changed.



rows	columns
891	4

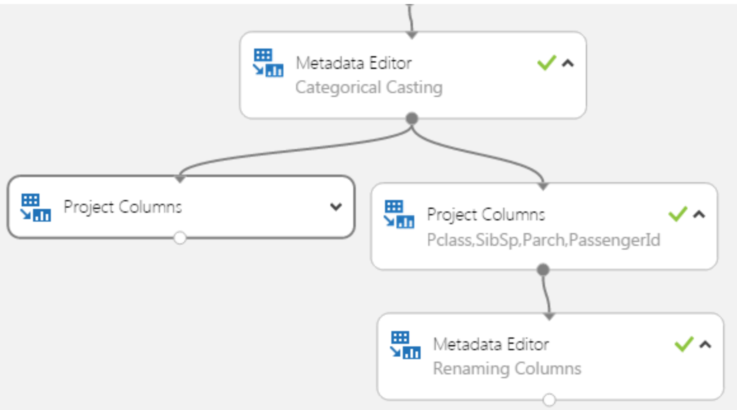
PassengerID	AccommodationClass	SiblingSpouse	ParentChild
1	3	1	0
2	1	1	0
3	3	0	0

© 2015 Data Science Dojo, All Rights Reserved

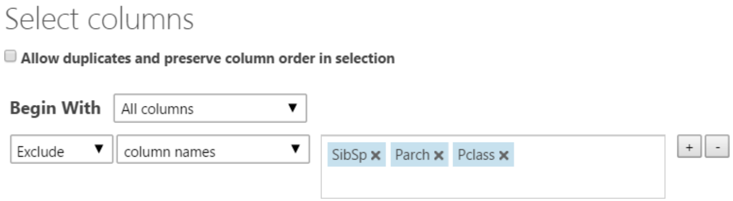
Exercise 5: Joining Tables

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.
   a. Drag in another project columns module and connect it to the metadata editor that performed the categorical casting. There should now be two project columns modules side by side feeding off the same metadata editor.



   b. Launch the column selector in the project columns module.
      i. Begin With: "All Columns"
      ii. Change the "include" to "exclude"
      iii. Choose to remove SibSp,Parch,Pclass. **Do not remove PassengerID**, we will need that for the join.



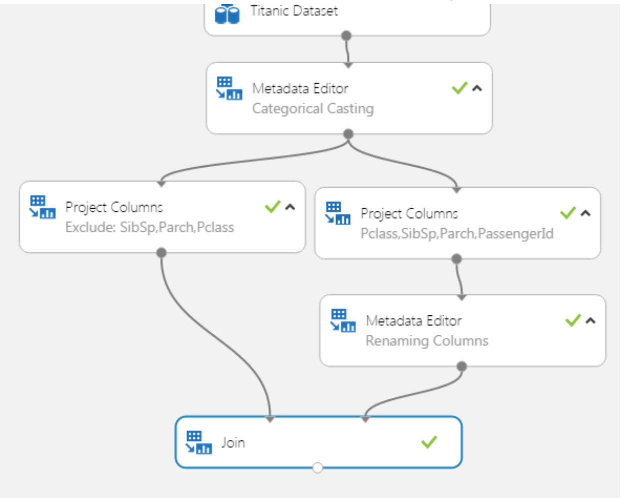
Select columns

□ **Allow duplicates and preserve column order in selection**

Begin With All columns

Exclude | column names | SibSp ✕ Parch ✕ Pclass ✕

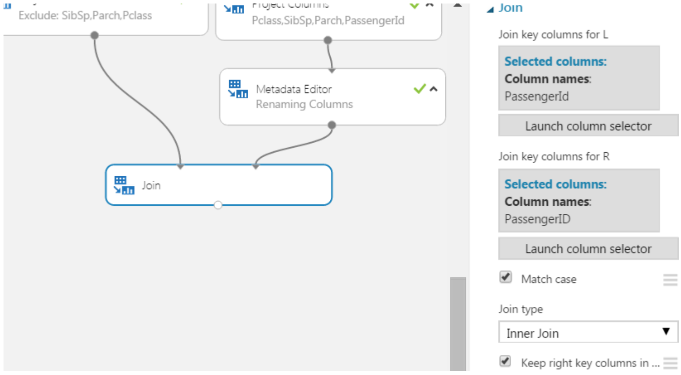
2. Join the two tables together.
   a. Drag in a Join Module from **Data Transformation > Manipulation**
   b. Connect the project columns module on the left to the metadata editor on the right (the metadata editor that renamed the previous 3 columns).



   c. Join key columns for L > Launch Column Selector > PassengerID
   d. Join key columns for R > Launch Column Selector > PassengerID

e. Uncheck the box for “Keep right key columns in”. This will remove the extra PassengerID as a result of the join.

f. Run and visualize the output of the join module.



Join

Join key columns for L

Selected columns:
Column names:
PassengerId

Launch column selector

Join key columns for R

Selected columns:
Column names:
PassengerID

Launch column selector

☑ Match case

Join type

Inner Join ▼

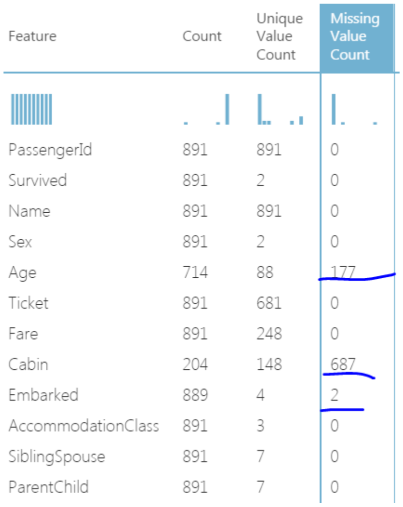
☑ Keep right key columns in ...

g. Verify that the new table has joined together properly on PassengerID.

Exercise 6: Descriptive Statistics and Clean Missing Data

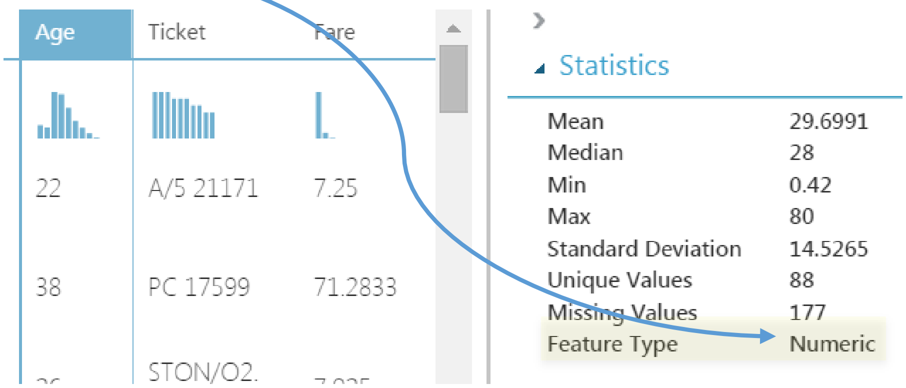
Before we can scrub missing values, we must find out where the missing values are and how to treat them. We will use the descriptive statistics module to help identify such columns.

1. Drag in a **Descriptive Statistics** module under **Statistical Functions.**
    a. Connect it to the join module. Run the experiment and visualize the descriptive statistics output.
    b. Examining some summary information about our data:
        i. What was the mean age?
        ii. How old was the oldest person onboard?
        iii. Who was the youngest?
            1. How do you interpret the youngest age?
        iv. What was the lowest price someone paid for a fare? The highest? Median fare price?
2. Identify columns that hold little to no data mining value. These will be dropped later.
    a. Passenger ID is only a primary key and holds no value outside of a database, table joins, or unique identification marker.
    b. Name also does not hold value to us in this context. You can use name to predict gender, but there are no missing values of gender.
    c. Ticket number also is an identifier which does not hold much value outside of identification.
3. Identify which columns have missing values.
    a. Age 177, Cabin 687, Embarked 2



Feature	Count	Unique Value Count	Missing Value Count
PassengerId	891	891	0
Survived	891	2	0
Name	891	891	0
Sex	891	2	0
Age	714	88	177
Ticket	891	681	0
Fare	891	248	0
Cabin	204	148	687
Embarked	889	4	2
AccommodationClass	891	3	0
SiblingSpouse	891	7	0
ParentChild	891	7	0

4. Identify the "Feature Type" of each column that contains missing values. Visualize the join output, and click on each individual column that has missing values to check their "Feature Type".
    a. Age<Numeric>, Cabin<String>, Embarked<Categorical>



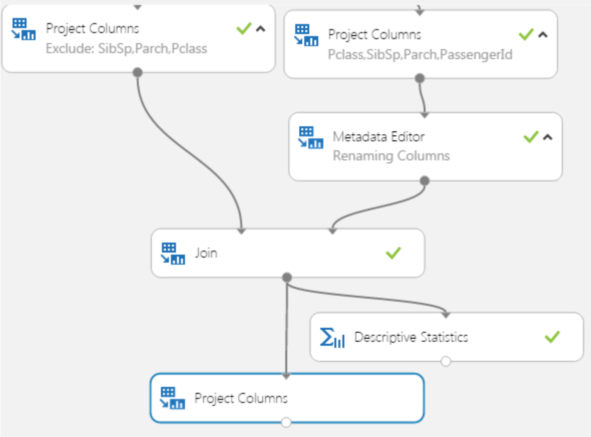
Age	Ticket	Fare
22	A/5 21171	7.25
38	PC 17599	71.2833
	STON/O2.	

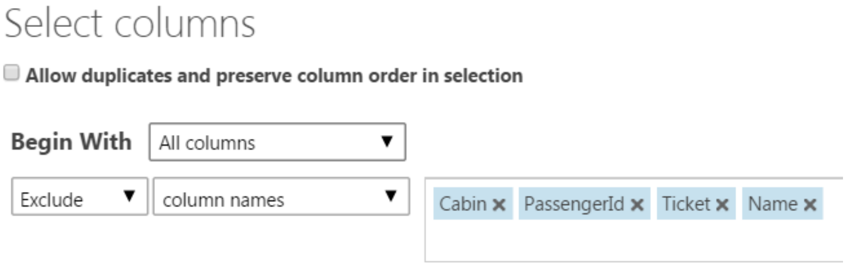
Statistics

Mean	29.6991
Median	28
Min	0.42
Max	80
Standard Deviation	14.5265
Unique Values	88
Missing Values	177
Feature Type	Numeric

© 2015 Data Science Dojo, All Rights Reserved

5. Perform an analysis on the missing values and develop a scrub strategy.
    a. Since AGE has 177 missing values, dropping the column would result in a huge loss of data. Age is also a numeric value. We can easily replace it with the "median" value and not much information loss will be present to the data. Age -> Replace with Median.
    b. Cabin is a string and is also missing 687 values. It may be best to just drop the cabin column moving forward since there are so many missing values. Cabin -> drop column.
    c. Embarked only has 2 missing values out of a total of 891 rows, which is not much. Dropping 2 rows to remove the missing values under embarked would not hurt the data very much.
6. Consider the **order of operations** used to perform the scrub. **THIS MATTERS A LOT.**
    a. Drop the cabin column first since it'll make for a smaller table moving forward and less iteration and processing for the next 2 scrubs.
    b. Replace missing age values with the median.
    c. ALWAYS PERFORM ROW DROPS LAST, not just on Azure ML but any data mining software or platform. Drop the 2 missing rows of embarked.
7. Drop the columns that hold little value in data mining.
    a. Drop Cabin, Name, PassengerID, and Ticket columns
    b. Use a project columns module and connect it to the output of the Join module.
        i. Begin With: "All Columns"
        ii. Change "include" to "exclude"
        iii. Select **Cabin, Ticket, Name, and Passenger** as columns to drop.





Select columns

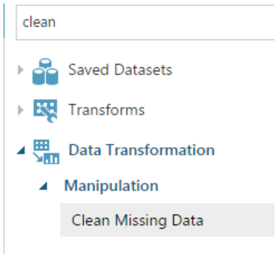
□ **Allow duplicates and preserve column order in selection**

Begin With All columns ▼

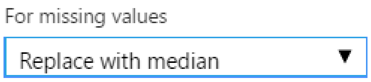
Exclude ▼ column names ▼ Cabin ✖ PassengerId ✖ Ticket ✖ Name ✖

© 2015 Data Science Dojo, All Rights Reserved

8. Replace missing values of age with the median.
    a. Drag in a Clean Missing Data under **Data Transformation > Manipulation**

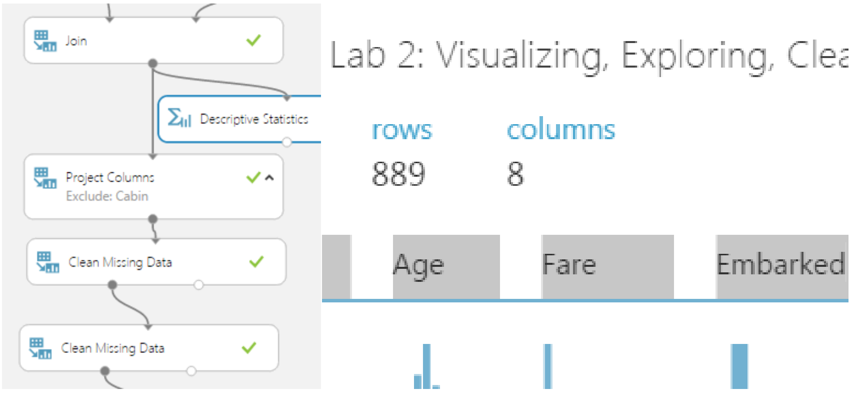


    b. Connect the module to the project columns module output.
    c. Select Column Type to "Numeric at "Launch column selector"
    d. "For Cleaning mode" set to "Replace with Median". This will replace all numeric missing values to the median value of the corresponding column.
        i. What will the module replace all the missing values with? Hint: look for the answer in the descriptive statistics module.
        ii. Why would it be poor practice to attach the Descriptive Statistics module AFTER this step? What would happen to our summary statistics? How would it affect our visualizations?


For missing values

Replace with median ▼

9. Drop the remaining rows of the missing values.
    a. Drag in another Clean Missing Data and connect it to the output of the previous Clean Missing Data module.
    b. "Cleaning mode" set to "Remove entire row"
    c. Visualize the output. There should now be 889 rows, from the original 891 rows. Verify that there are now only 8 columns.

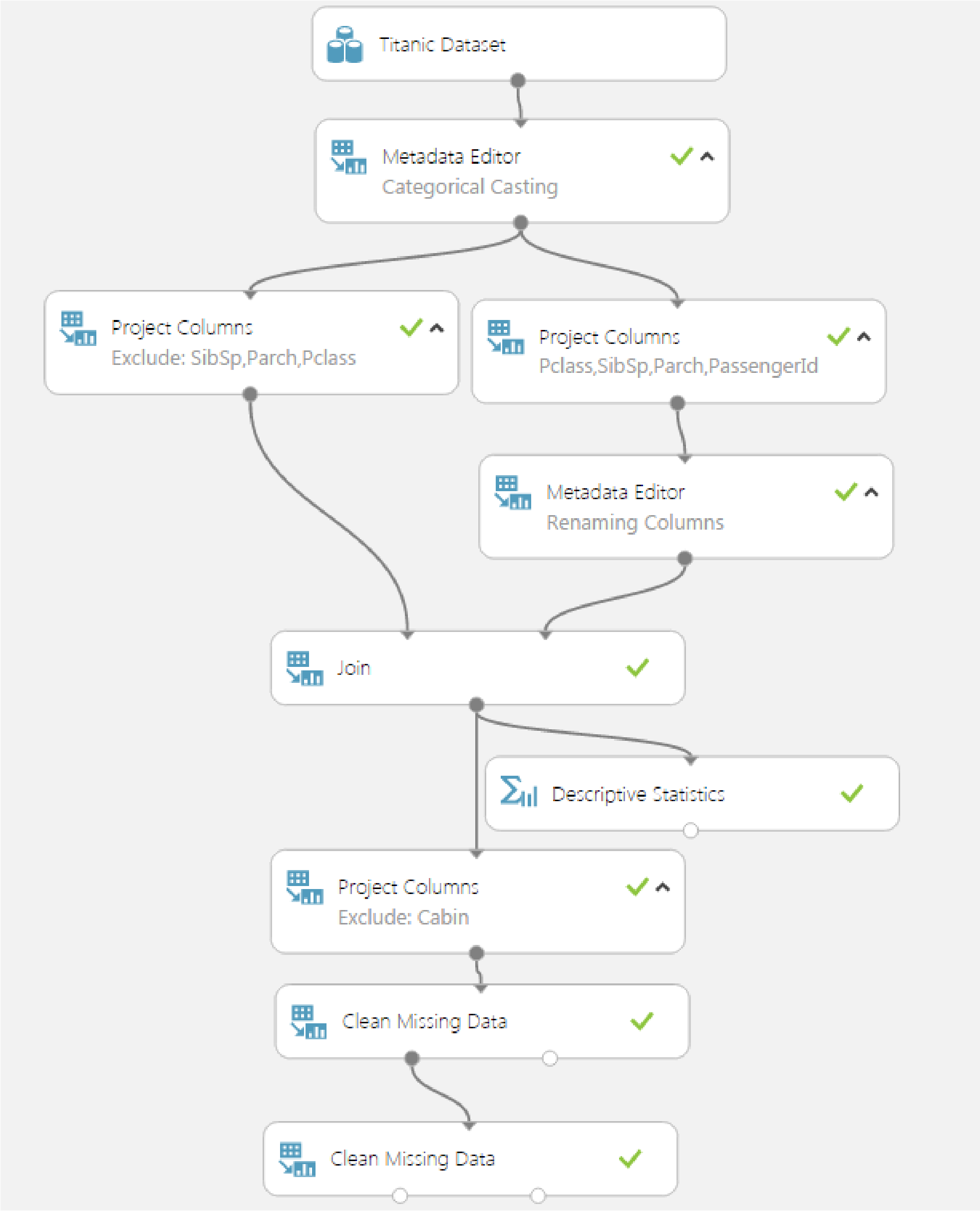


rows	columns
889	8

We are now ready to data mine!

© 2015 Data Science Dojo, All Rights Reserved

Final Result:



© 2015 Data Science Dojo, All Rights Reserved