CIS5560 Term Project Tutorial

Authors: Gregory D. Mamoyac, Tejas Agara Chandrakumar, Nitesh Kamboj, Julia Stachurska

Instructor: Jongwook Woo

Date: 05/06/2018

Lab Tutorial

jstachu (jstachu@calstatela.edu)

nkamboj(nkamboj@calstatela.edu)

afnu3(afnu3@calstatela.edu)

gmamoya(gmamoja@calstatela.edu)

05/06/2018

Azure ML and Spark ML Analysis of Stack 0verflow

For R Language Question and Answers

Azure K Means

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually
- Create Spark cluster
- Train KMeans system
- K-Means Clustrering for predicting if an answer is acceptable

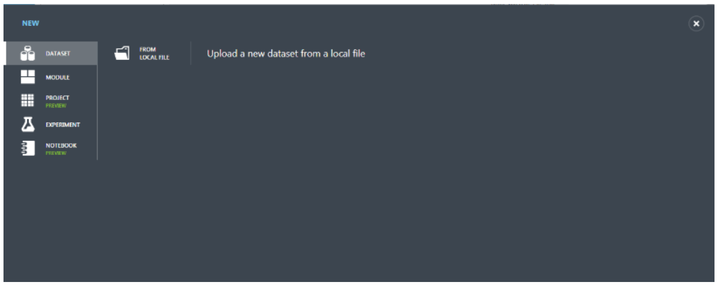
- Visualization
- https://gallery.cortanaintelligence.com/Experiment/Clustering-kmeans-2

Platform Spec

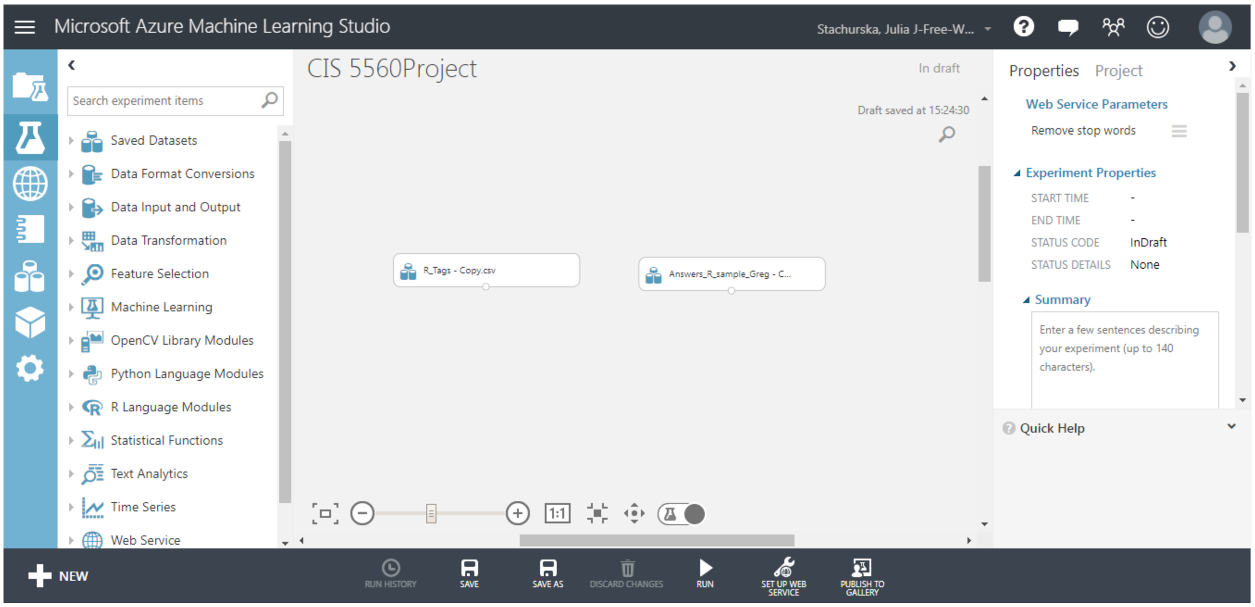
- Microsoft Azure ML
- CPU Speed: ~3.4GHz
- # of nodes: 1
- Total Memory Size: 10GB

Step 1: Upload the Data Set from local file

This step is to upload two datasets: R_tags.csv and Anwers_R_Sample.csv from local file.



Upload a new dataset from a local file



CIS 5560Project

In draft

Draft saved at 15:24:30

Web Service Parameters

Remove stop words

Experiment Properties

START TIME -

END TIME -

STATUS CODE InDraft

STATUS DETAILS None

Summary

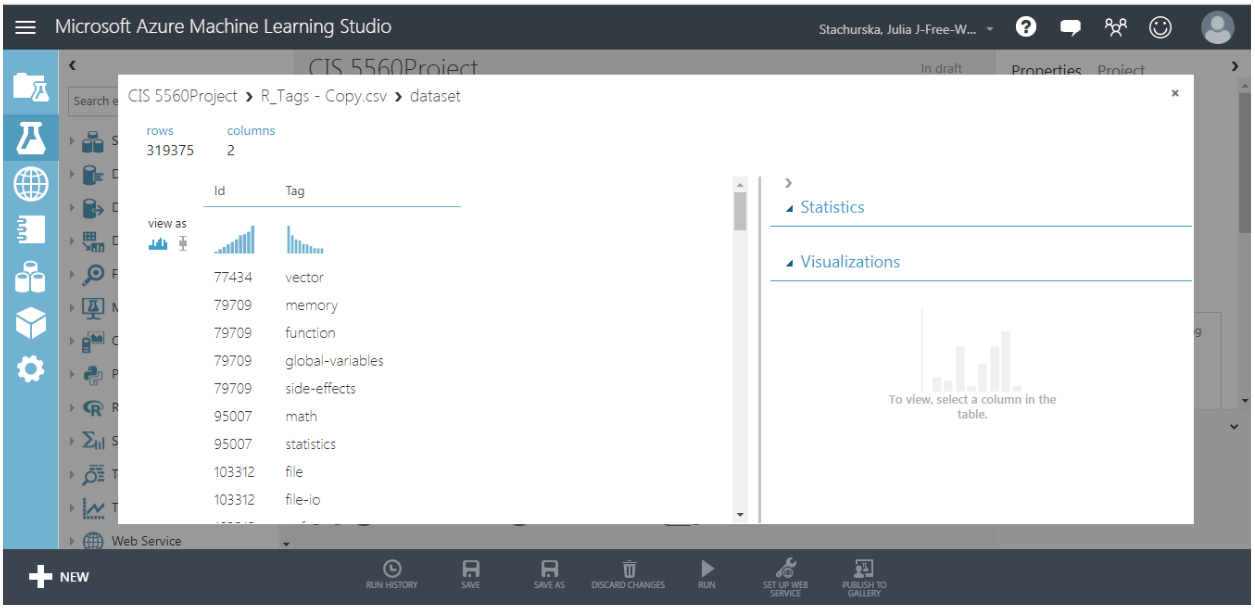
Enter a few sentences describing your experiment (up to 140 characters).

Properties:

- This is a new version of an existing dataset: Unselected
- Enter a name for the new dataset: R_Answers_LDA_Sample
- Select a type for the new dataset: Generic CSV file with a header(.csv)
- Provide an optional description: R Answers RDA.

Step 2: Visualize the Dataset in Azure ML

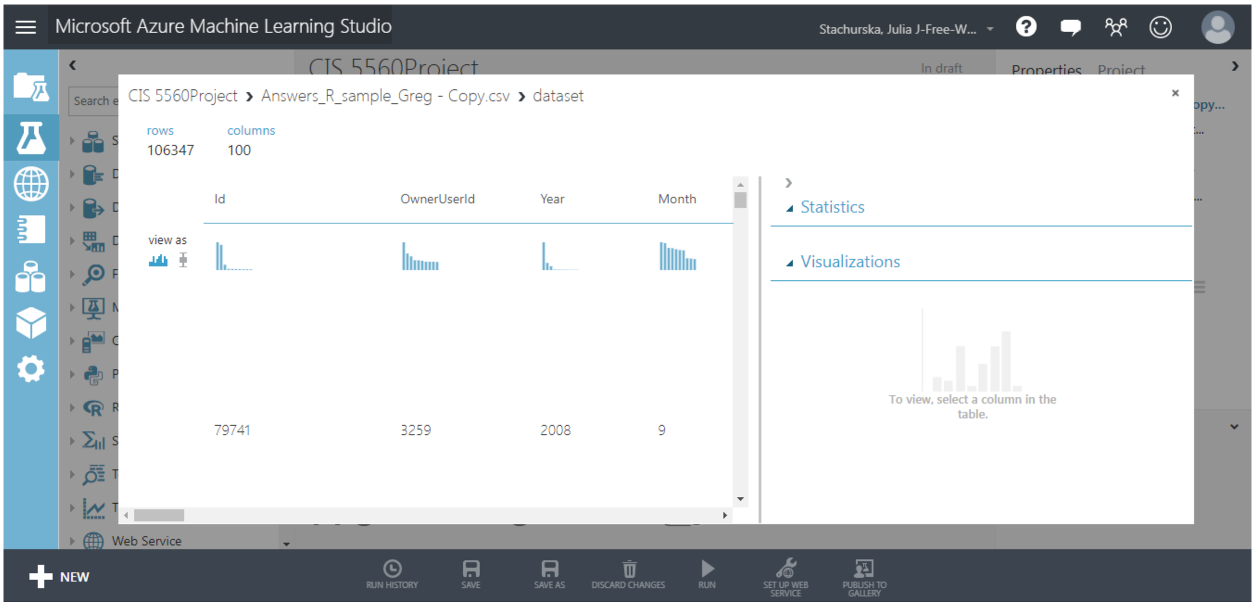
This step is to verify if the data set uploaded contains all the data from the source file.



CIS 5560Project › R_Tags - Copy.csv › dataset

rows	columns
319375	2

Id	Tag
77434	vector
79709	memory
79709	function
79709	global-variables
79709	side-effects
95007	math
95007	statistics
103312	file
103312	file-io



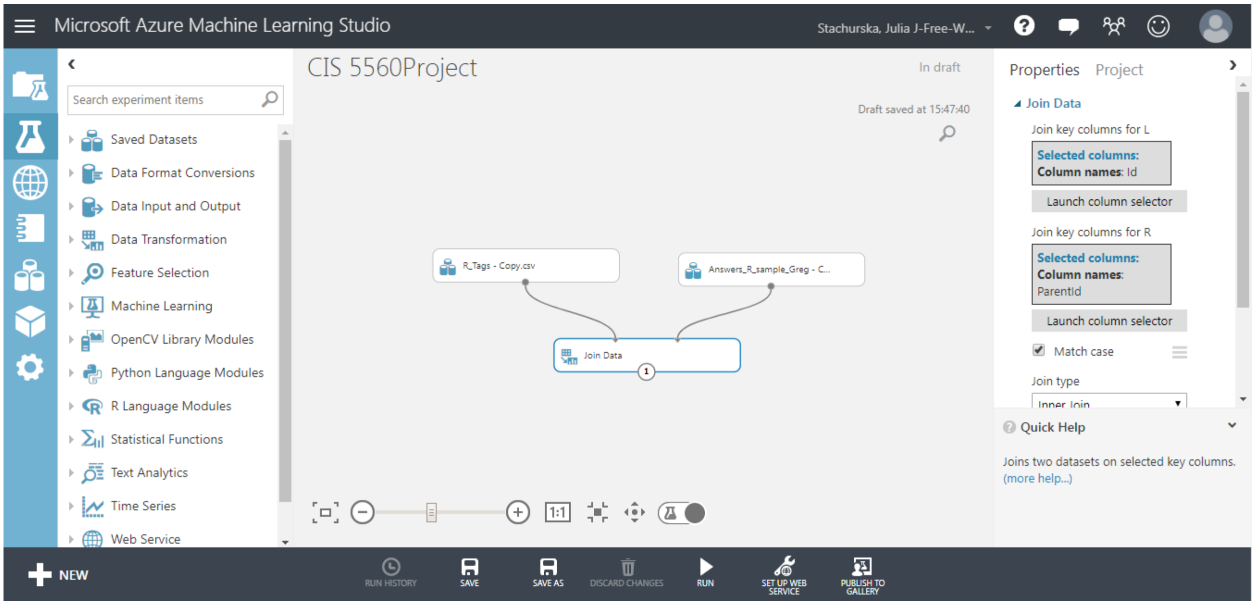
CIS 5560Project › Answers_R_sample_Greg - Copy.csv › dataset

rows	columns
106347	100

Id	OwnerUserId	Year	Month
79741	3259	2008	9

Step 3: Join Data

This step is to join two datasets into one, using columns “Id” and “ParentId”.



Join Data

Join key columns for L

Selected columns:
Column names: Id

Join key columns for R

Selected columns:
Column names: ParentId

☑ Match case

Join type

Inner Join

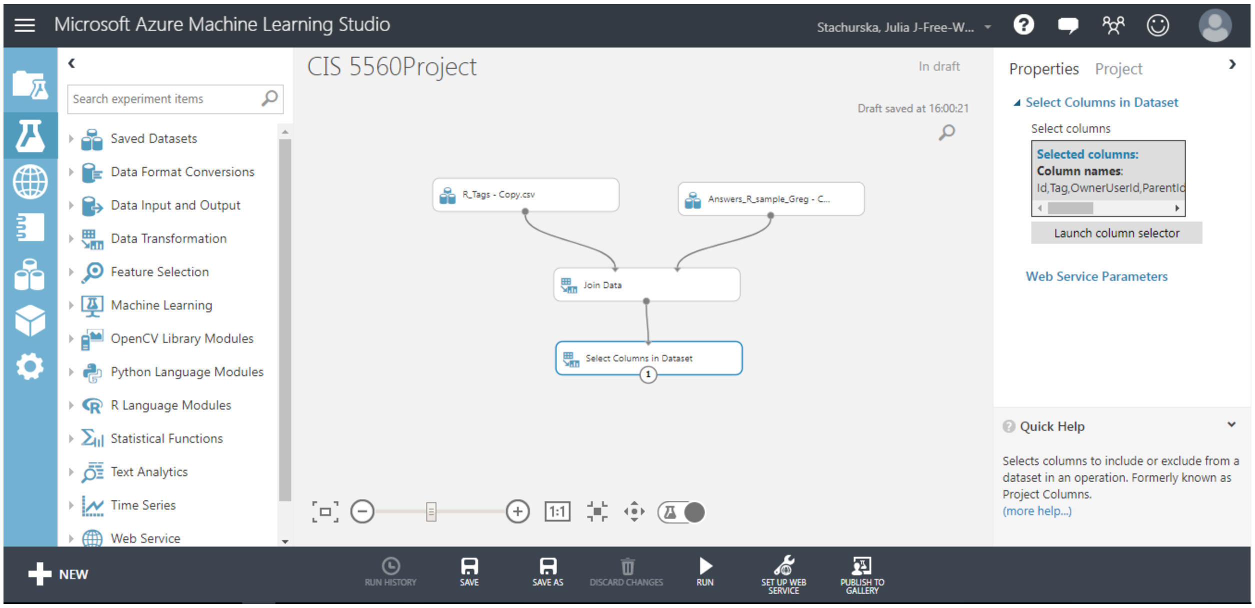
Joins two datasets on selected key columns. (more help...)

Properties:

- Join key columns for L: column name “Id”
- Join key columns for R: column name “ParentId”
- Match case: selected
- Join type: Inner join
- Keep right key columns in joined table: checked

Step 4: Select columns in Dataset

This is a common user interface element in Azure ML modules to enable selecting the columns you want to use in the module.

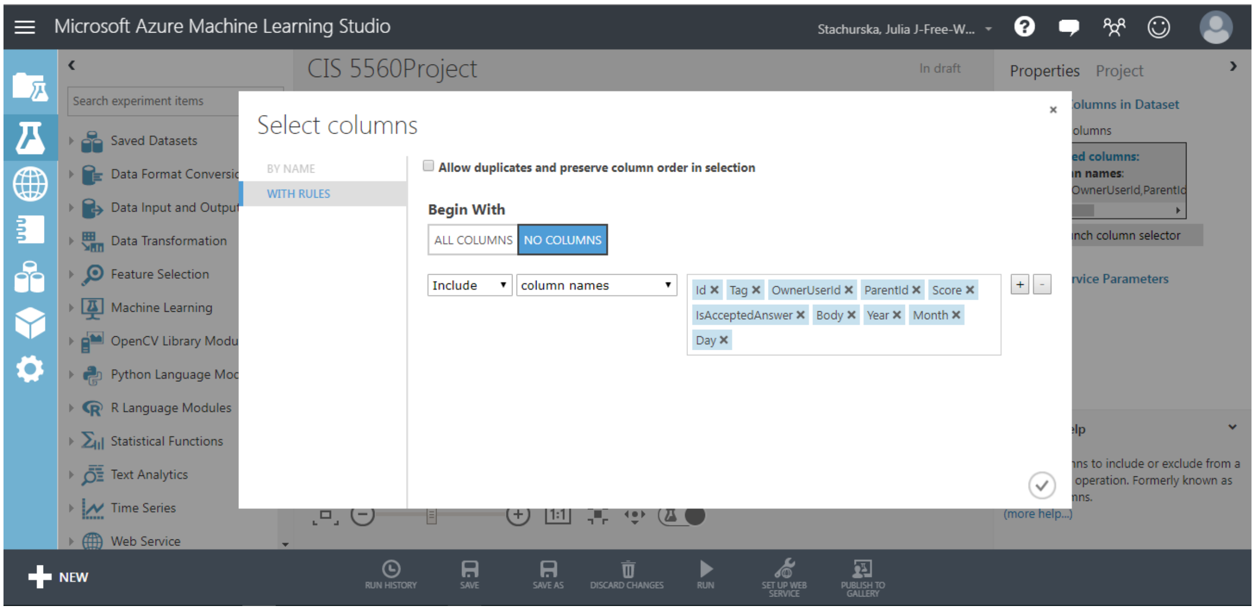


Selected columns:
Column names: Id,Tag,OwnerUserId,ParentId

Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.

In the Select columns dialog box, select option With Rules to begin with no columns, and include:

- Id,
- Tag,
- OwnerUserId,
- ParentId,
- Score,
- IsAcceptedAnswer,
- Body,
- Year,
- Month,
- Day



Select columns

BY NAME
WITH RULES

□ Allow duplicates and preserve column order in selection

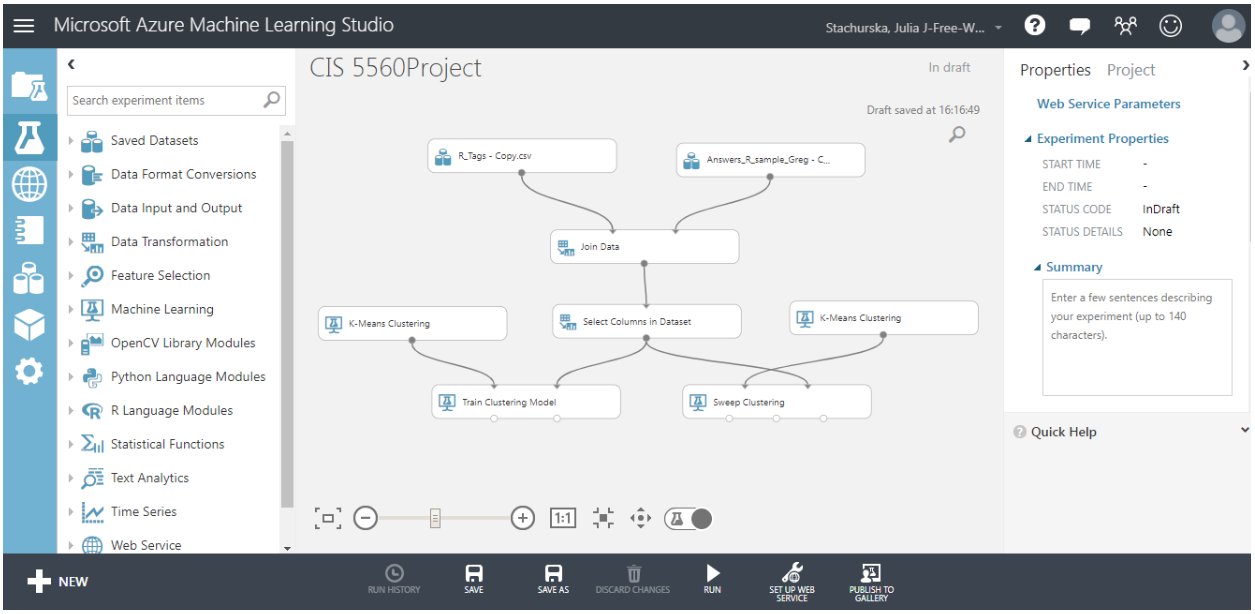
Begin With

ALL COLUMNS | NO COLUMNS

Include | column names | Id ✕ Tag ✕ OwnerUserId ✕ ParentId ✕ Score ✕ IsAcceptedAnswer ✕ Body ✕ Year ✕ Month ✕ Day ✕

Step 5: Train Clustering Model and Sweep Clustering (K-Means)

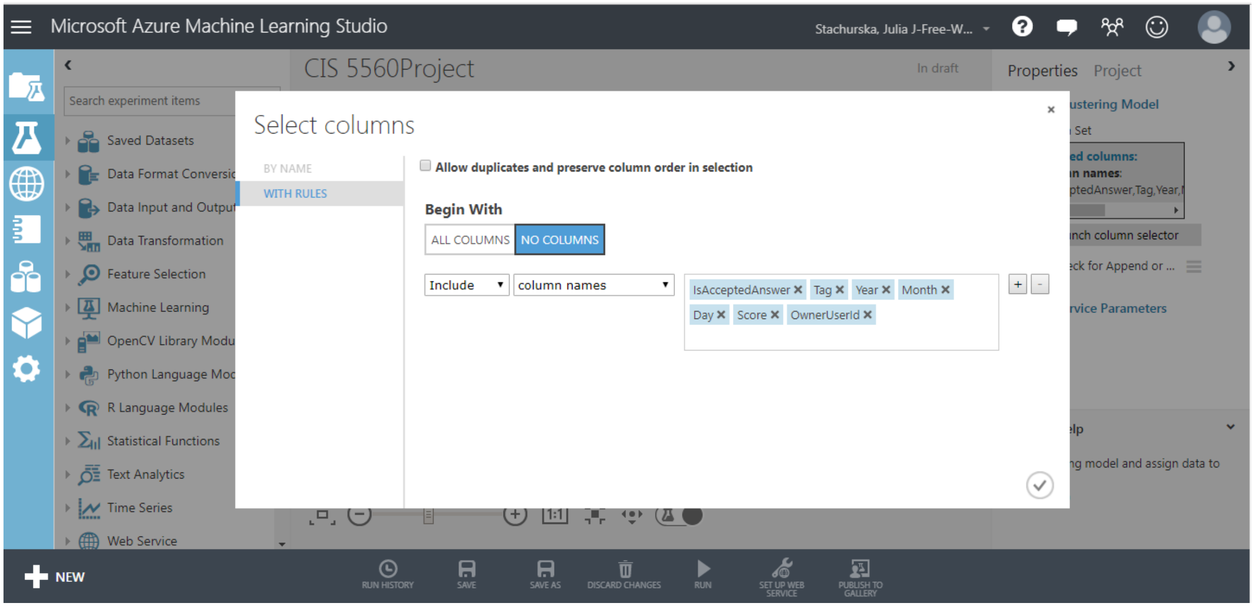
This step is to create clusters (groups in the data) with the number of groups represented by the variable *K*.



START TIME	-
END TIME	-
STATUS CODE	InDraft
STATUS DETAILS	None

Enter a few sentences describing your experiment (up to 140 characters).

5A: Train Clustering Model



Select columns

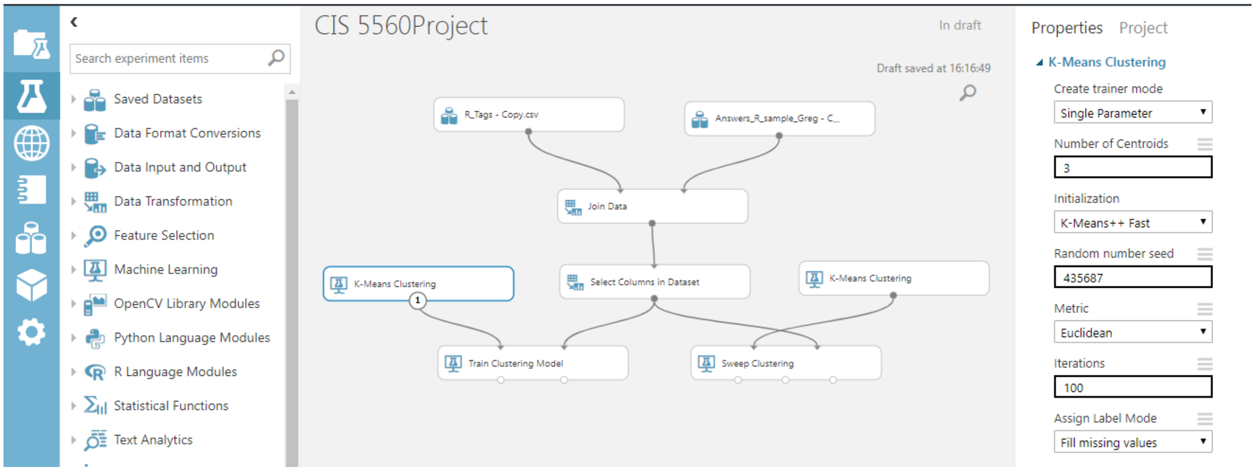
☐ Allow duplicates and preserve column order in selection

Begin With: ALL COLUMNS | NO COLUMNS

Include · column names · IsAcceptedAnswer, Tag, Year, Month, Day, Score, OwnerUserId

Properties:

- Columns used: IsAcceptedAnswer, Tag, Year, Month, Day, Score, OwnerUSerID



K-Means Clustering

Create trainer mode: Single Parameter

Number of Centroids: 3

Initialization: K-Means++ Fast

Random number seed: 435687

Metric: Euclidean

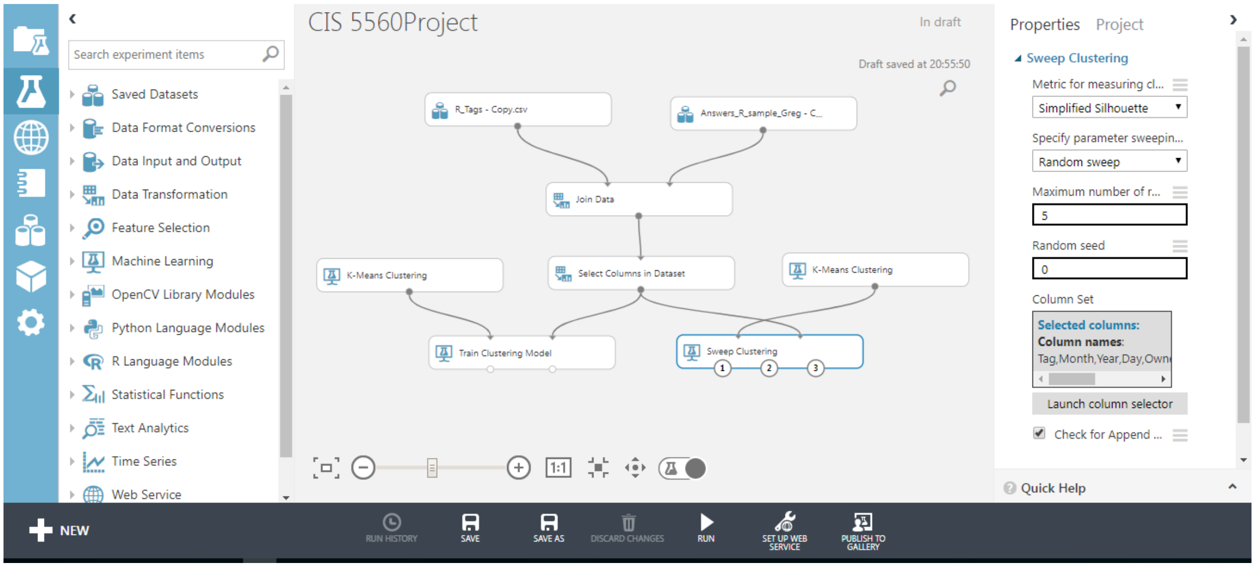
Iterations: 100

Assign Label Mode: Fill missing values

K-Means Clustering Properties:

- Create trainer mode: Single Parameter
- Number of centroids: 3
- Initialization: K-Means++ Fast
- Random number speed: 435687
- Metric: Euclidean
- Iterations: 100
- Assign Label Model: Fill Missing Values

5B: Sweep Clustering



Sweep Clustering

Metric for measuring cl... Simplified Silhouette

Specify parameter sweepin... Random sweep

Maximum number of r... 5

Random seed 0

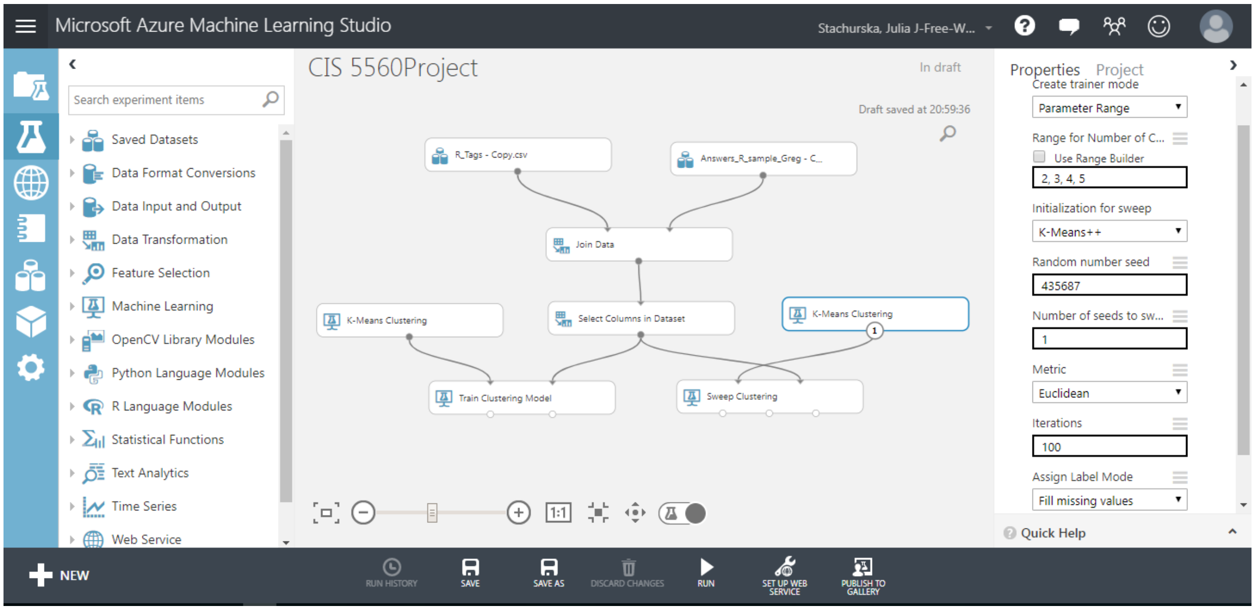
Column Set
Selected columns:
Column names: Tag,Month,Year,Day,Own

Launch column selector

☑ Check for Append ...

Properties:

- Metric for measuring clustering result: Simplified Silhouette
- Specify parameter sweeping mode: Random Sweep
- Maximum number of runs on random sweep: 5
- Random seed: 0
- Column Set: Tag Month Year Day OwnerUserId Score IsAcceptedAnswer
- Check for Append or Uncheck for Result Only: Checked



Create trainer mode: Parameter Range

Range for Number of C...
☐ Use Range Builder
2, 3, 4, 5

Initialization for sweep: K-Means++

Random number seed: 435687

Number of seeds to sw... 1

Metric: Euclidean

Iterations: 100

Assign Label Mode: Fill missing values

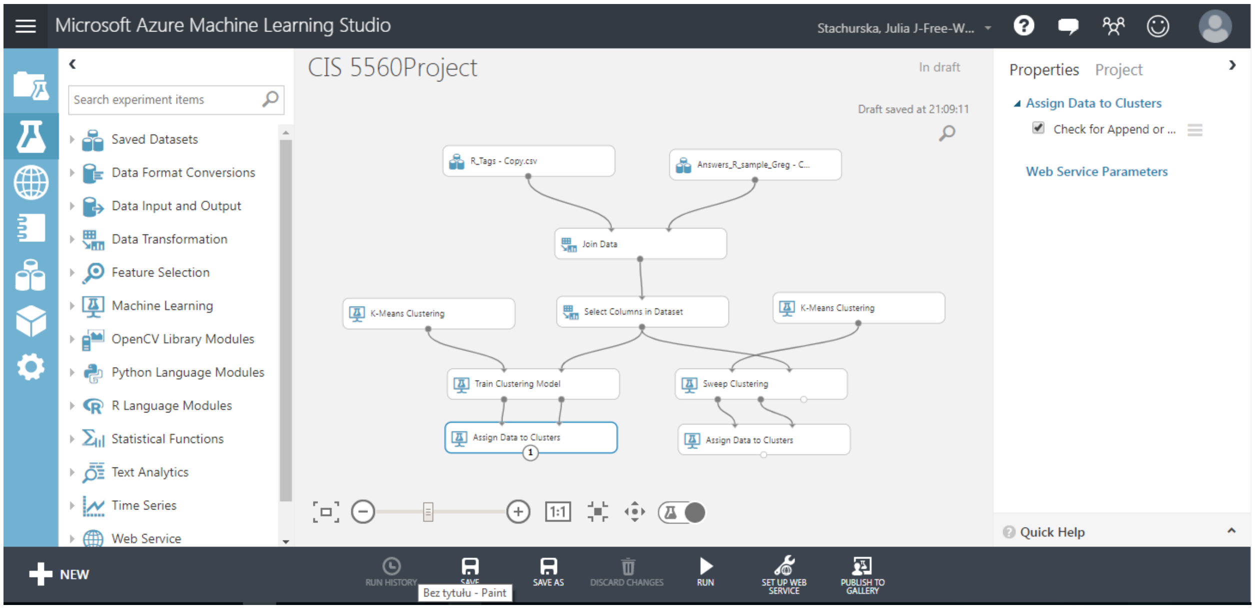
K-Means Clustering Properties:

- Create trainer mode: Parameter Range
- Range for Number of Centroids: 2, 3, 4, 5 (Use Range Builder: Not Checked)
- Initialization for sweep: K-Means++

- Random number seed:435687
- Number of seeds to sweep:1
- Metric: Euclidean
- Iterations: 100
- Assign Label Mode: Fill Missing Values

Step 6: Assign Data to Clusters

This step is to assign data to clusters using an existing trained clustering models.



Assign Data to Clusters

☑ Check for Append or ...

6A: Assign Data to Clusters

Properties:

- Check for Append or Uncheck for Result Only: Checked

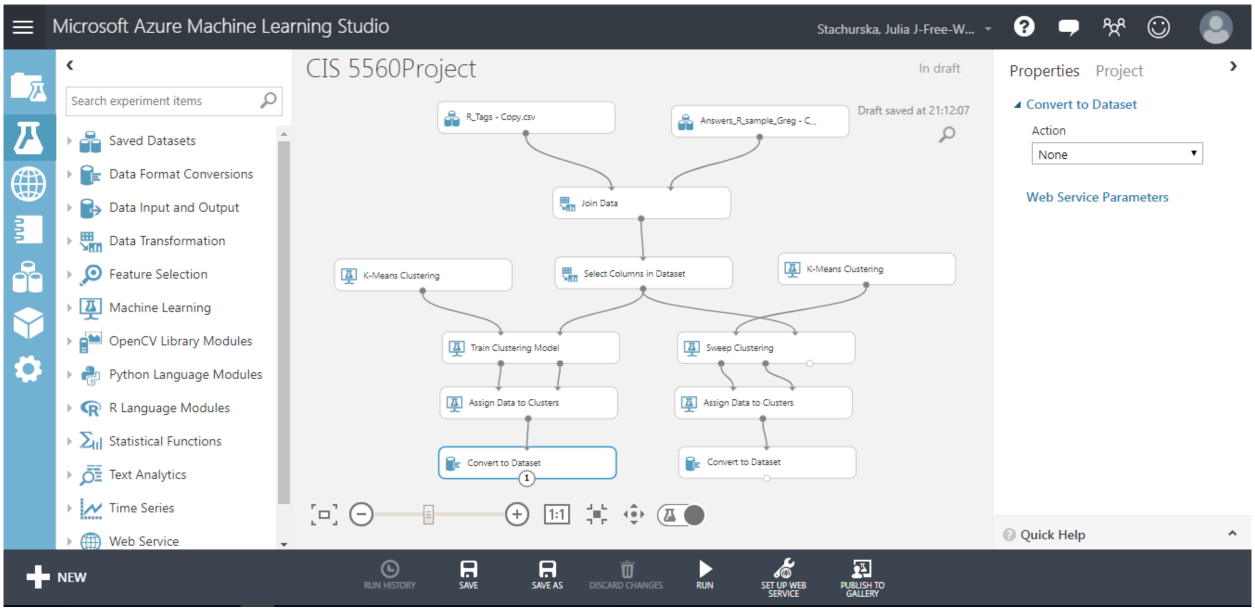
6B: Assign Data to Clusters

Properties:

- Check for Append or Uncheck for Result Only: Checked

Step 7: Convert to Dataset

This step is to convert data input to the internal Dataset format used by Microsoft Azure Machine Learning



7A: Convert to Dataset:

Properties:

- Action: None

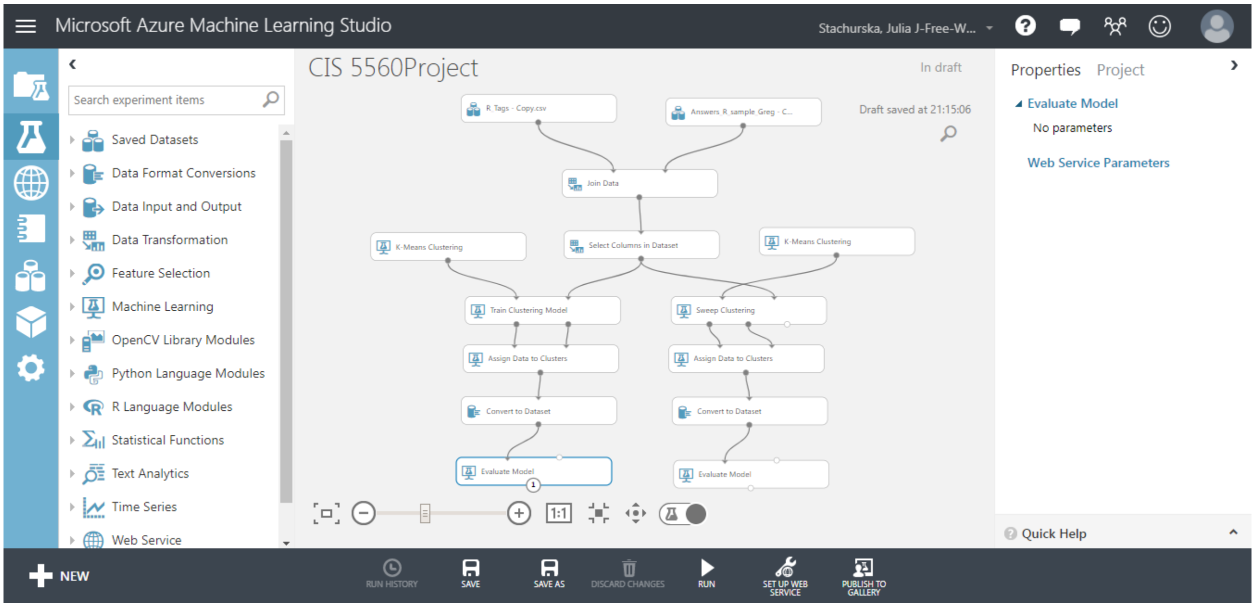
7B: Convert to Dataset:

Properties:

- Action: None

Step 8: Evaluate Model

This step is to evaluate a scored model with standard metrics.



References

- https://www.kaggle.com/stackoverflow/rquestions 2GB
- https://gallery.cortanaintelligence.com/Experiment/Clustering-kmeans-2