



Business Intelligence

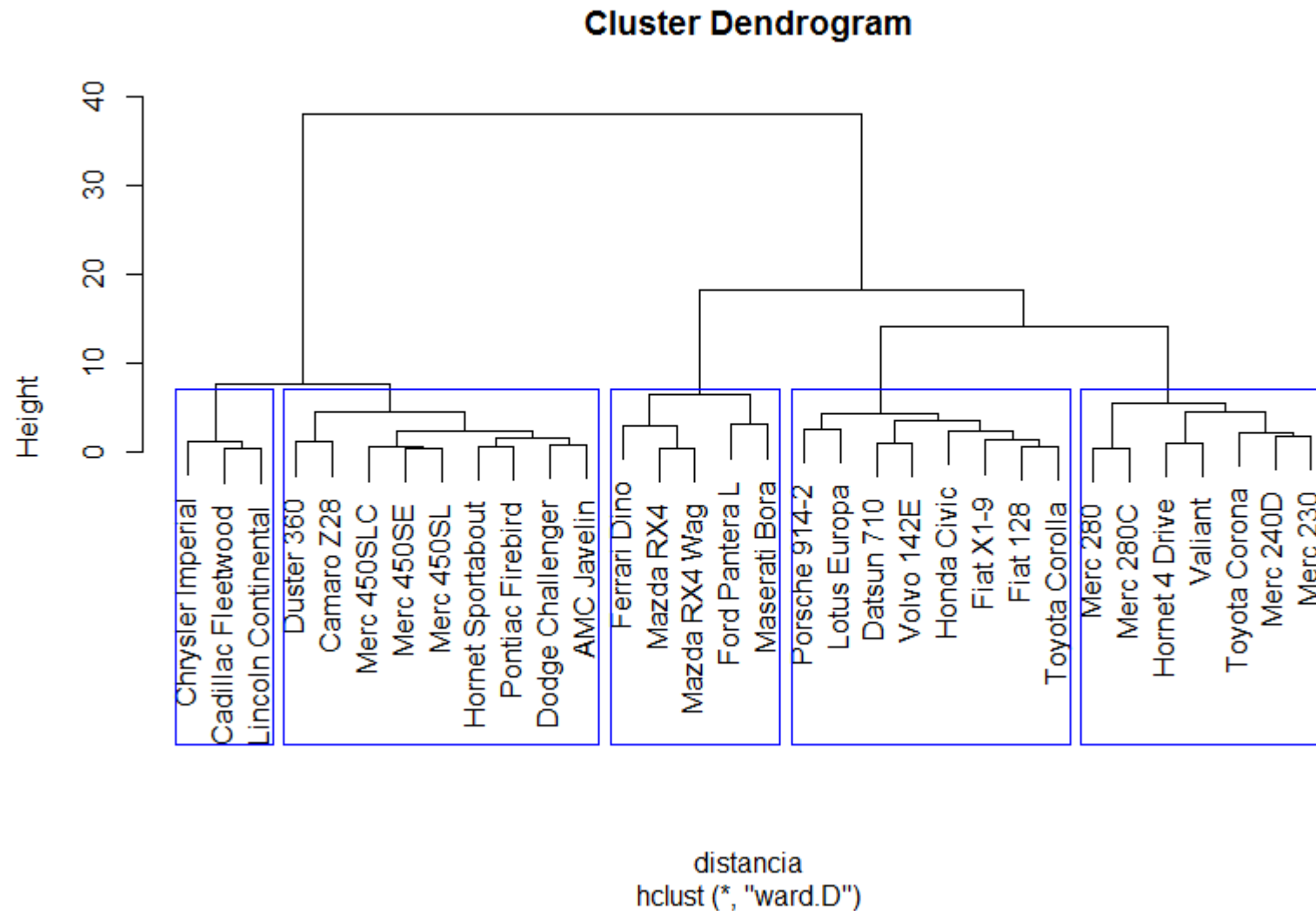
Azure Machine Learning

Prof. Leandro Guerra

E-mail: leandro.guerra@outspokenmarket.com.br **IG:** @leandrowar

Dendogramas no R

Um pouco mais...



Dendogramas no R

Um pouco mais...

> grupos

Mazda RX4	Mazda RX4 Wag	Datsun 710	Hornet 4 Drive	Hornet Sportabout	Valiant
1	1	2	3	4	3
Duster 360	Merc 240D	Merc 230	Merc 280	Merc 280C	Merc 450SE
4	3	3	3	3	4
Merc 450SL	Merc 450SLC	Cadillac Fleetwood	Lincoln Continental	Chrysler Imperial	Fiat 128
4	4	5	5	5	2
Honda Civic	Toyota Corolla	Toyota Corona	Dodge Challenger	AMC Javelin	Camaro Z28
2	2	3	4	4	4
Pontiac Firebird	Fiat X1-9	Porsche 914-2	Lotus Europa	Ford Pantera L	Ferrari Dino
4	2	2	2	1	1
Maserati Bora	Volvo 142E				
1	2				

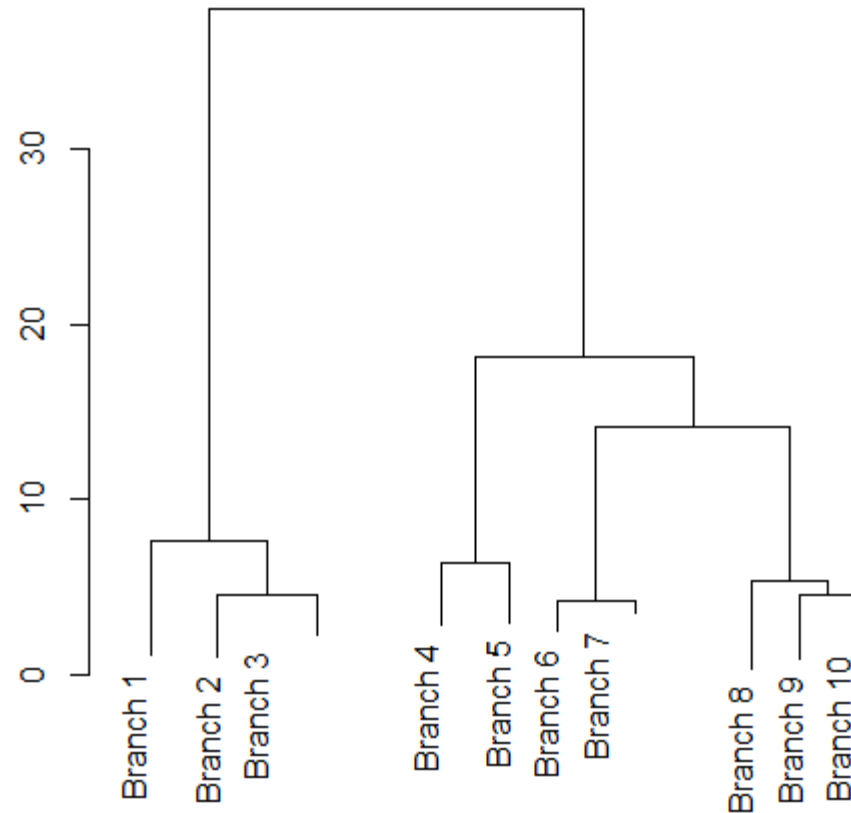
Dendogramas no R

Um pouco mais...

```
#Exibe o dendrograma com menos níveis  
#Primeiro convertemos em um objeto de dendrograma  
dendo <- as.dendrogram(dendo)
```

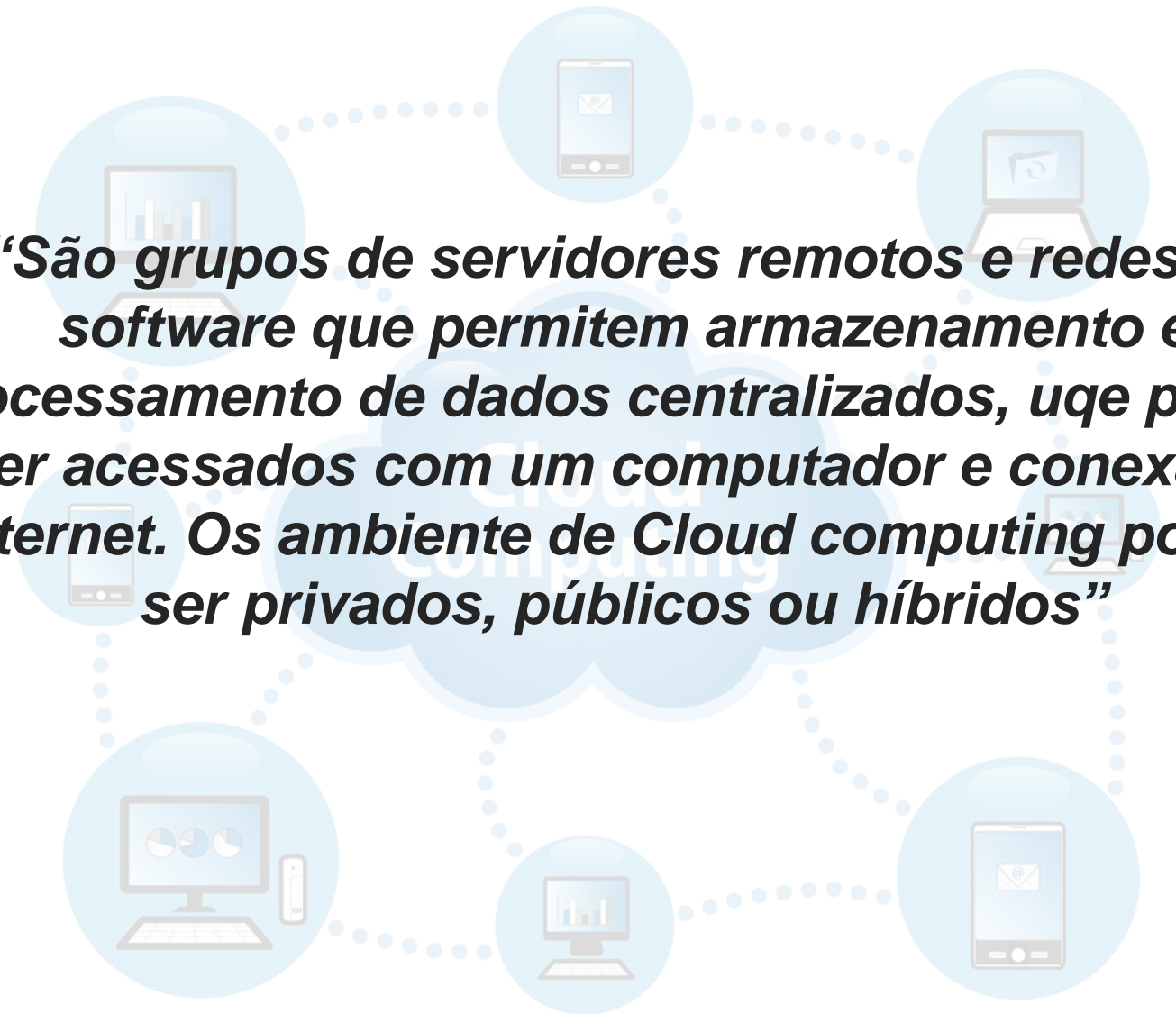
```
#Agora o plot  
plot(cut(dendo , h = 4)$upper, main = "Corte superior com h=4")
```

Corte superior com h=4



Cloud Computing

Conceito

A diagram illustrating the concept of Cloud Computing. It features a central blue cloud shape with the words "Cloud Computing" written inside in a light blue font. Surrounding this central cloud are eight circular nodes, each containing an icon representing a different device or application: a laptop with a bar chart, a smartphone with an envelope icon, a laptop with a magnifying glass, a desktop monitor with a bar chart, a desktop tower unit, a smartphone with an envelope icon, a desktop monitor with a bar chart, and a laptop with a magnifying glass. Dotted lines connect these nodes in a circular pattern, suggesting a networked environment.

“São grupos de servidores remotos e redes de software que permitem armazenamento e processamento de dados centralizados, uqe podem ser acessados com um computador e conexão à internet. Os ambiente de Cloud computing podem ser privados, públicos ou híbridos”

Cloud Computing

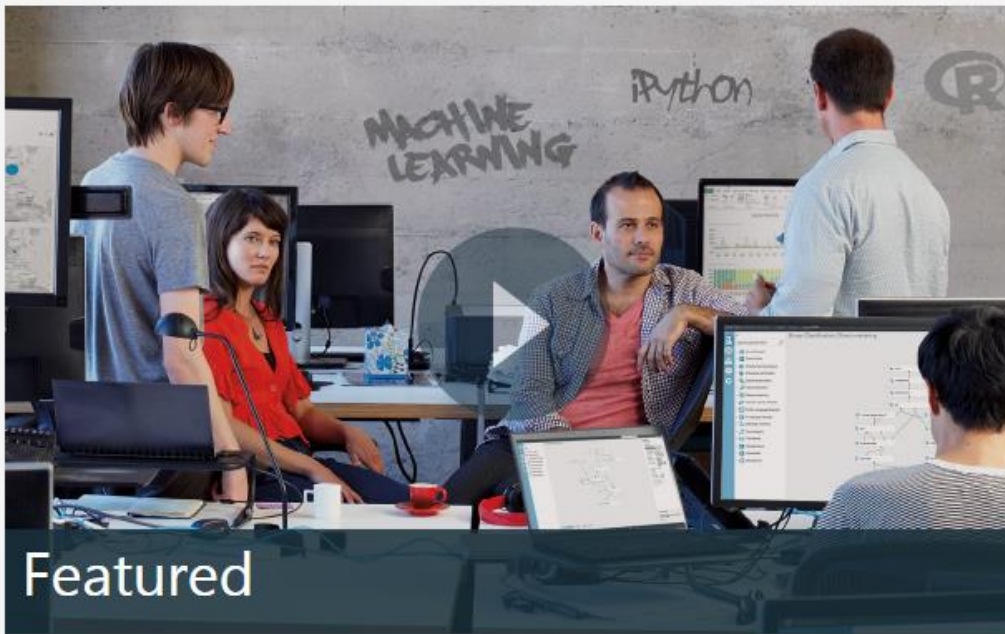
Alguns provedores



Windows Azure Machine Learning

Microsoft Azure Machine Learning | [Home](#) [Studio](#) [Gallery](#) PREVIEW

Sign



Welcome to Azure Machine Learning

Want a taste?

It's free and easy to try Machine Learning right now—just sign in with a Microsoft account and start experimenting. No credit card or Azure subscription needed.

[Get started now](#) ➔

[Pricing & FAQ](#) ▶

By using this free version, you agree to be bound by the [Microsoft Azure Website Terms of Use](#).

Windows Azure Machine Learning

Entrar

Conta da Microsoft [O que é isto?](#)

☐ Mantenha-me conectado

Entrar


[Não consegue acessar sua conta?](#)


[Entrar com um código de uso único](#)


Não tem uma conta da Microsoft? [Inscreva-se já](#)


Windows Azure Machine Learning


Microsoft Azure Machine Learning | Home Studio Gallery PREVIEW

 EXPERIMENTS


 WEB SERVICES

 DATASETS

 TRAINED MODELS

 SETTINGS

+ NEW

 DELETE

experiments

MY EXPERIMENTS SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED	
<input checked="" type="checkbox"/>	Clustering: Group iris da...	eng.leandroguerra	Finished	4/10/2015 8:45:52 PM	
<input type="checkbox"/>	CA Dairy Analysis	eng.leandroguerra	Failed	4/6/2015 3:15:53 PM	
<input type="checkbox"/>	Experiment created on ...	eng.leandroguerra	Failed	4/6/2015 3:03:37 PM	

Windows Azure

Criando um novo Experimento

NEW



DATASET



EXPERIMENT



Search experiment templates



Microsoft Samples

[VIEW MORE IN GALLERY](#) 



Blank Experiment

Experiment
Tutorial

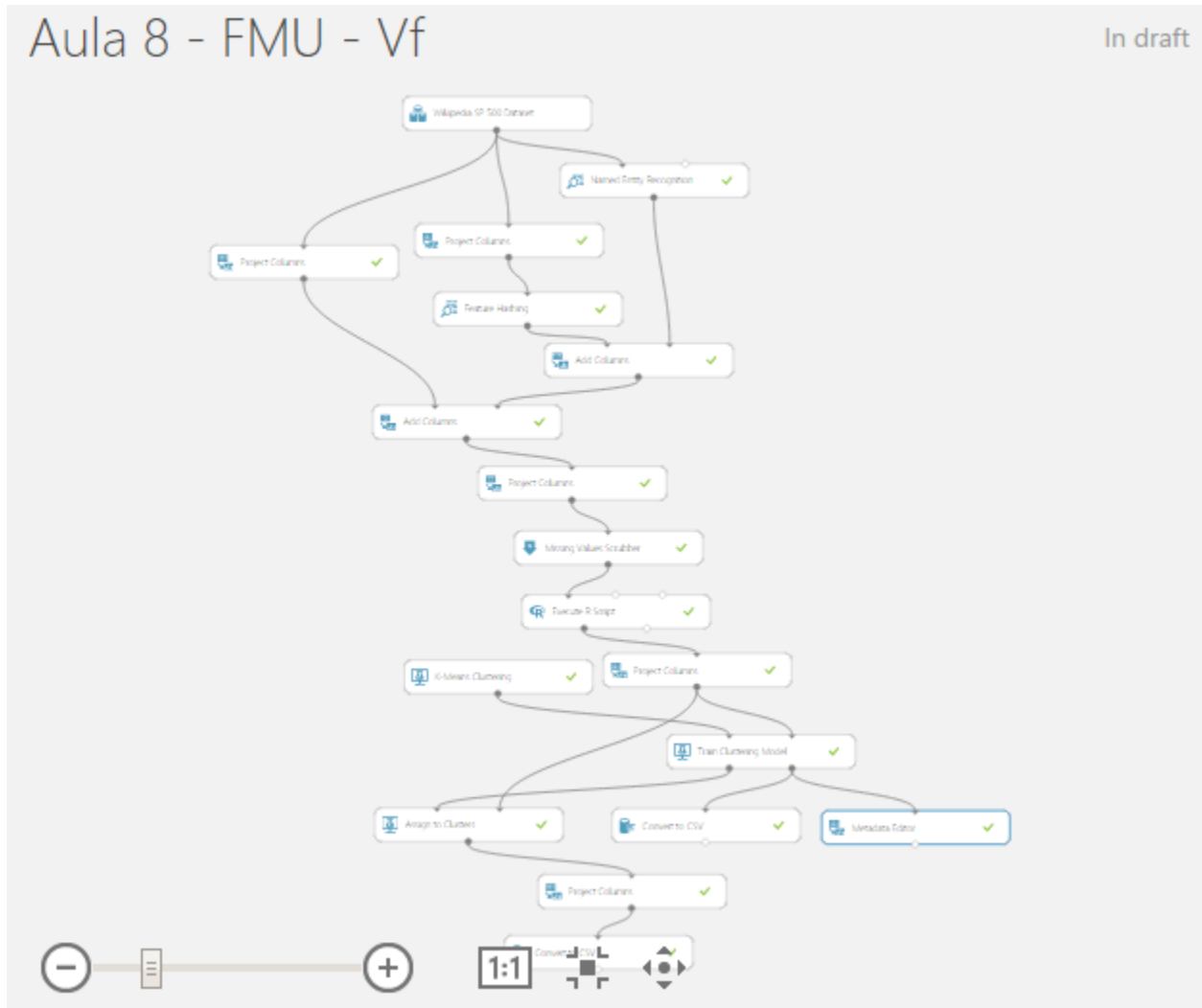


Sample 1: Download
dataset from UCI: Adult 2
class dataset



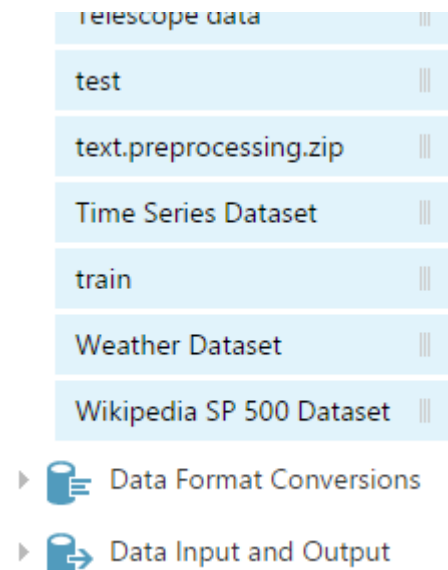
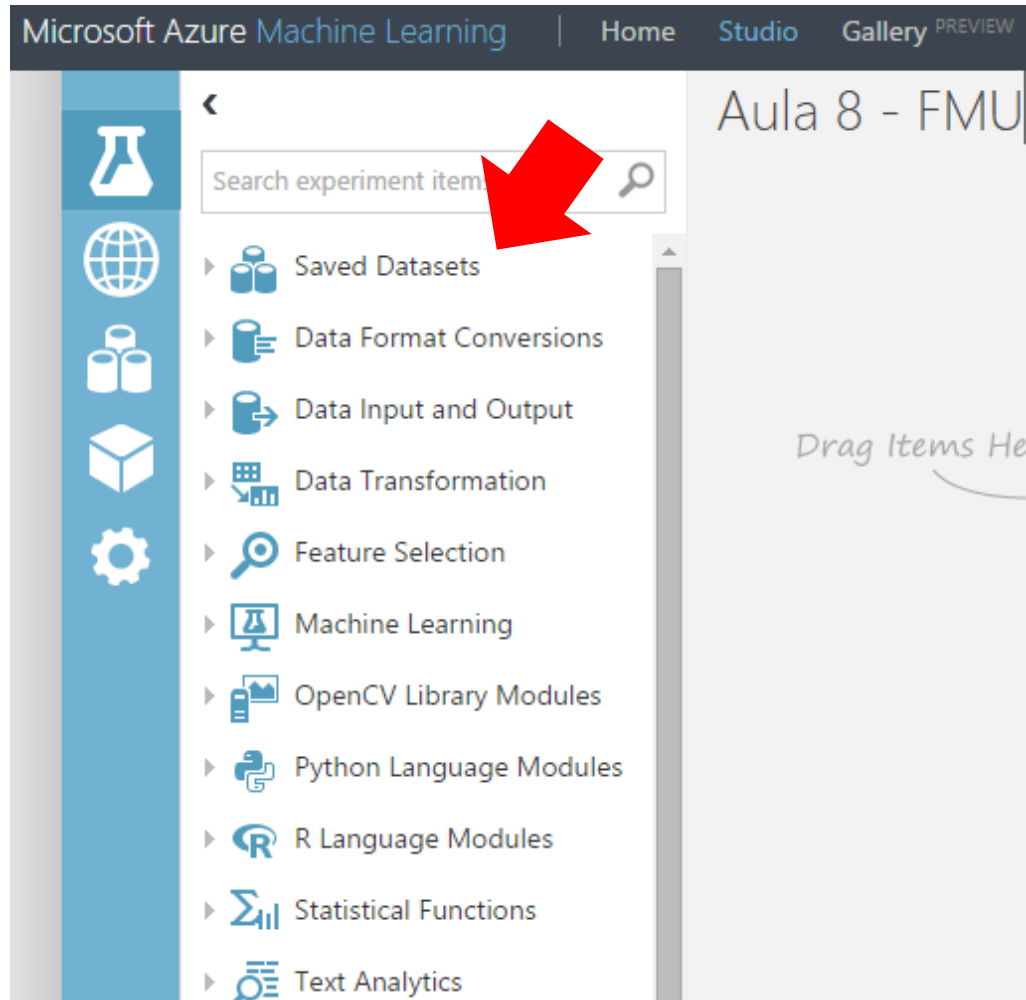
Windows Azure

Criando um novo Experimento – Nosso Objetivo



Windows Azure

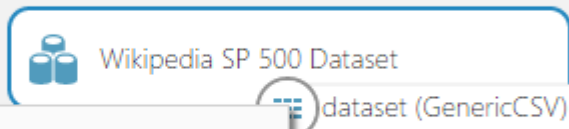
Escolha do dataset e visualização



Windows Azure

Visualização dos Dados

Aula 8 - FMU



Microsoft Azure Machine Learning | Home Studio Gallery PREVIEW

8 - FMU > Wikipedia SP 500 Dataset > dataset



Rows 56
Columns 3

View as



Title

Apple Inc.

Category

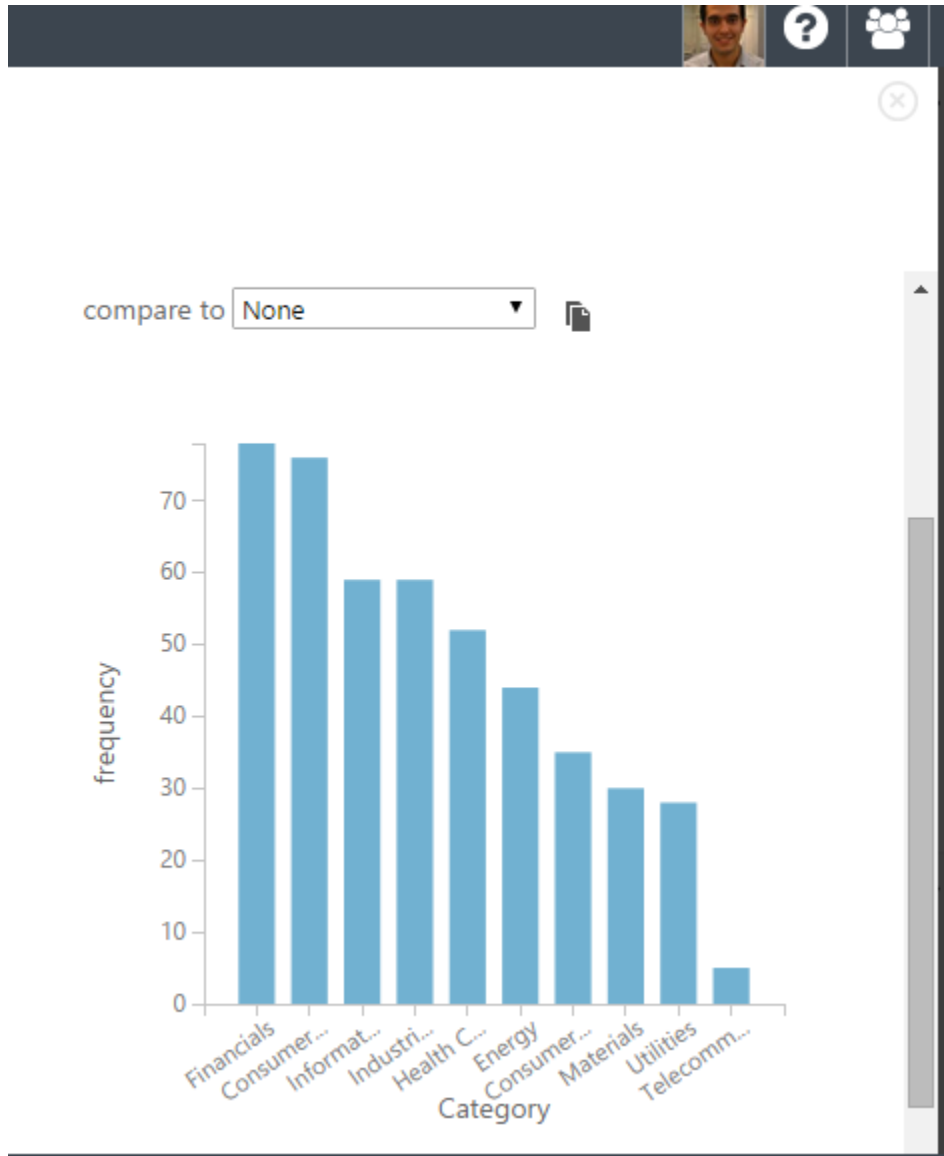
Information Technology

Text

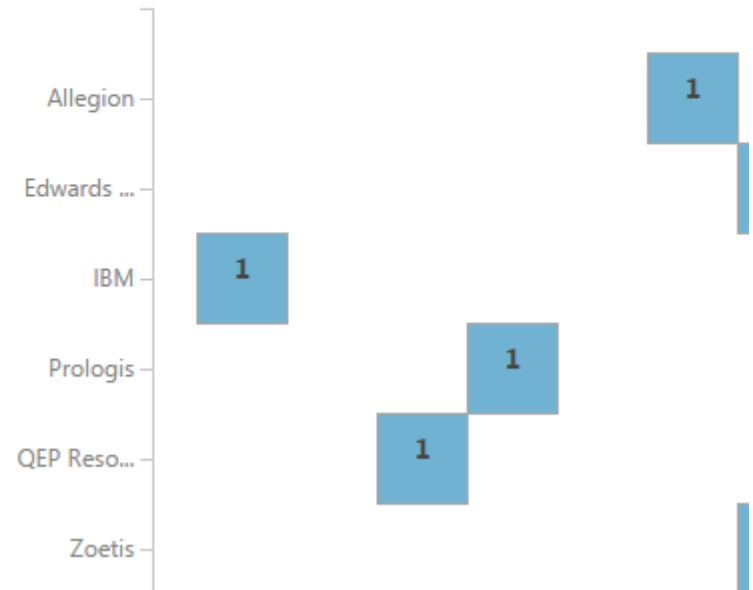
nasdaq 100 component s
p 500 component
foundation founder
location city apple campus
1 infinite loop street infinite
loop cupertino california
cupertino california
location country united
states u s locations 406
retail stores may 2013 area
served worldwide key
people ref tim cook ceo
steve jobs fou...
br nasdaq 100 nasdaq 100
component br s p 500 s p
500 component industry
computer software

Windows Azure

Visualização dos Dados

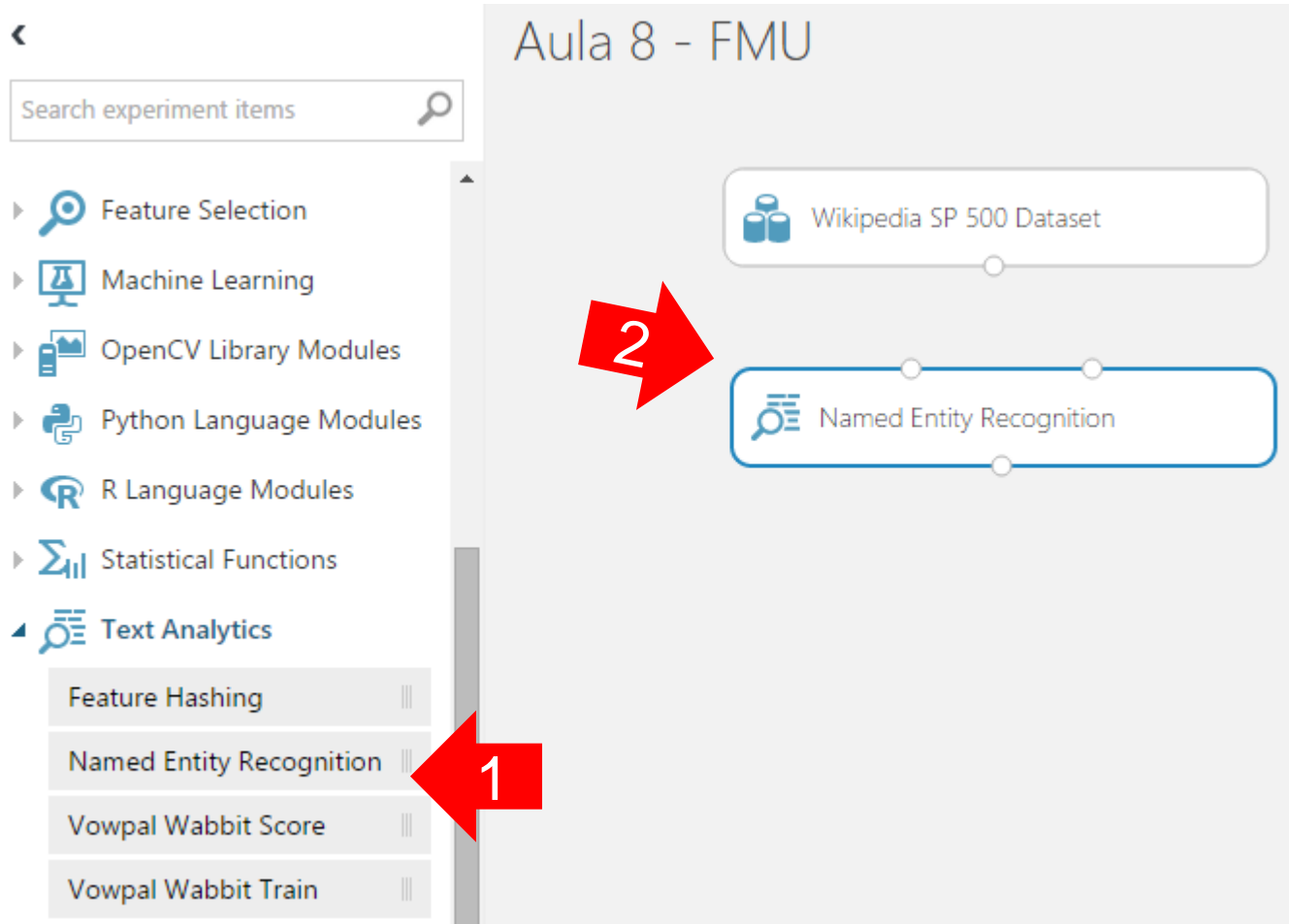


compare to



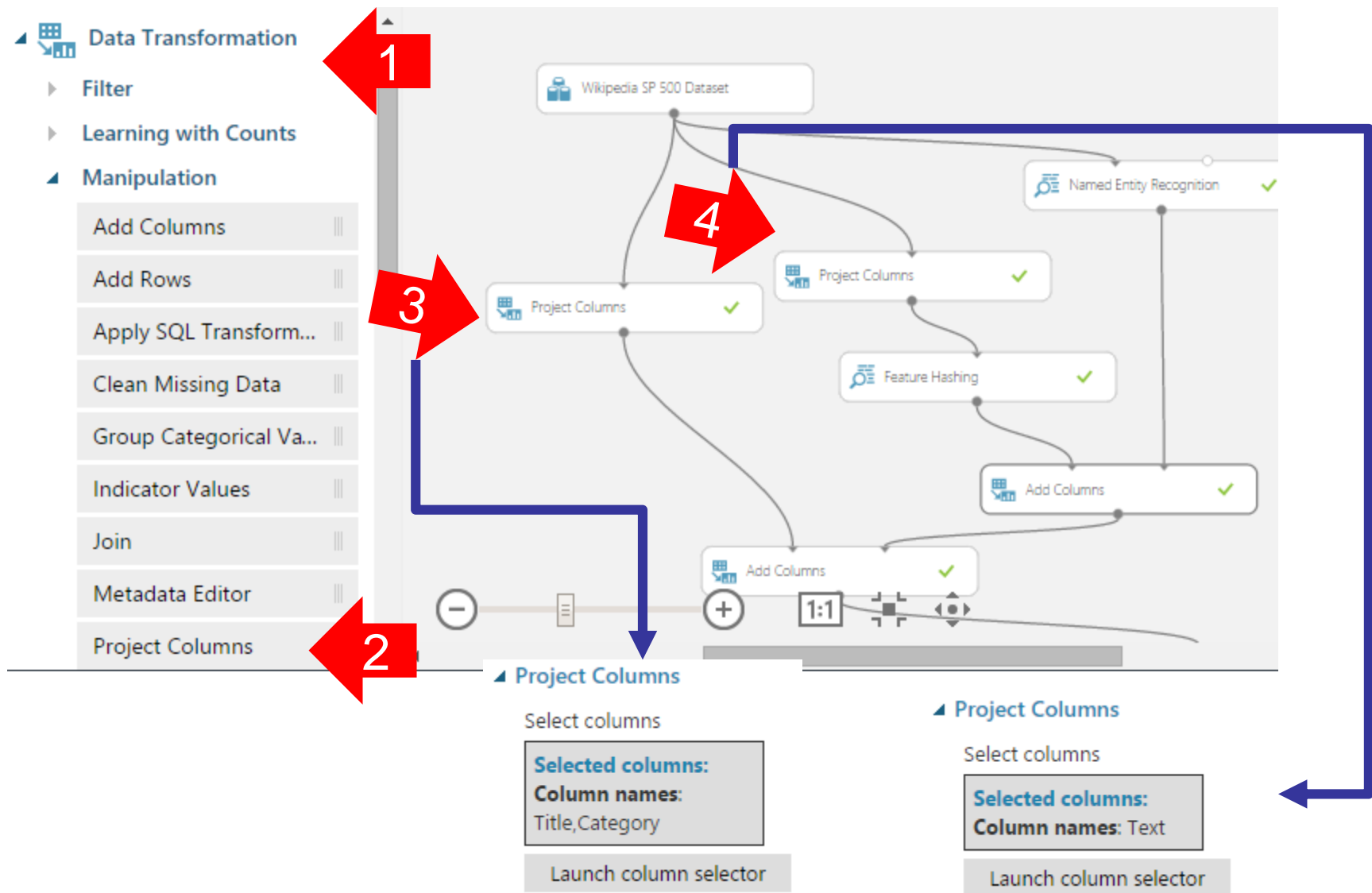
Windows Azure

Classificação de Empresas



Windows Azure

Criando a Estrutura



Windows Azure

Criando a Estrutura

The screenshot displays the Windows Azure Machine Learning interface. On the left, the 'Text Analytics' workspace is visible, containing a list of modules: Feature Hashing, Named Entity Recognition, Vowpal Wabbit Score, and Vowpal Wabbit Train. A red arrow labeled '1' points to the workspace header, and a red arrow labeled '2' points to the 'Feature Hashing' module in the list.

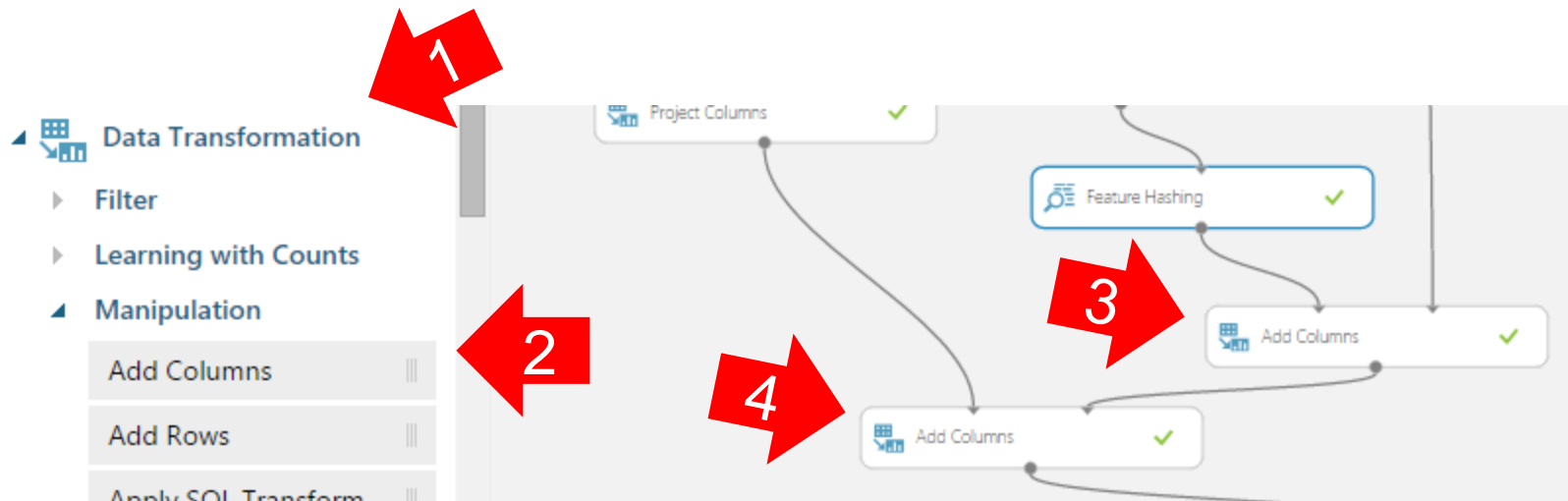
The main canvas shows a workflow diagram. A 'Project Columns' module is connected to a 'Feature Hashing' module, which is then connected to an 'Add Columns' module. A red arrow labeled '3' points to the 'Feature Hashing' module in the canvas. A blue box highlights the 'Feature Hashing' and 'Add Columns' modules, with a blue arrow pointing from this box to the 'Properties' pane.

The 'Properties' pane for the 'Feature Hashing' module is shown below the canvas. It includes the following settings:

- Target column(s):** A box labeled 'Selected columns:' with the text 'Column type: String, Feature' below it.
- Launch column selector** button.
- Hashing bitsize:** A dropdown menu set to '10'.
- N-grams:** A dropdown menu set to '3'.
- START TIME:** 4/11/20...

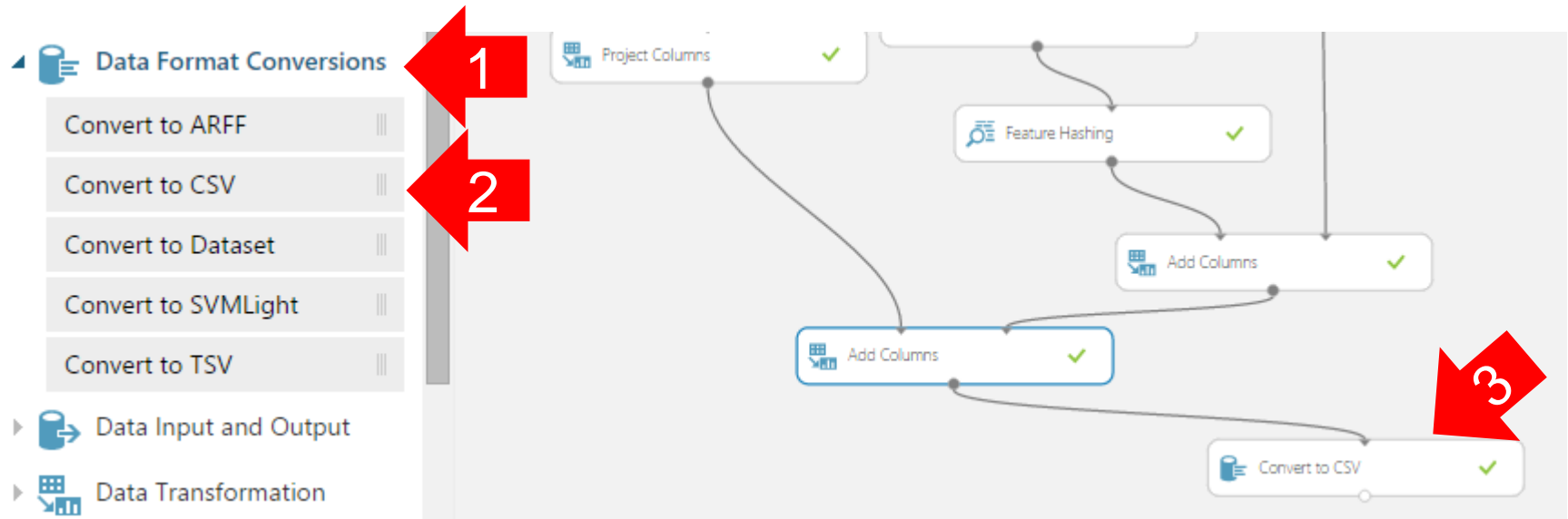
Windows Azure

Criando a Estrutura



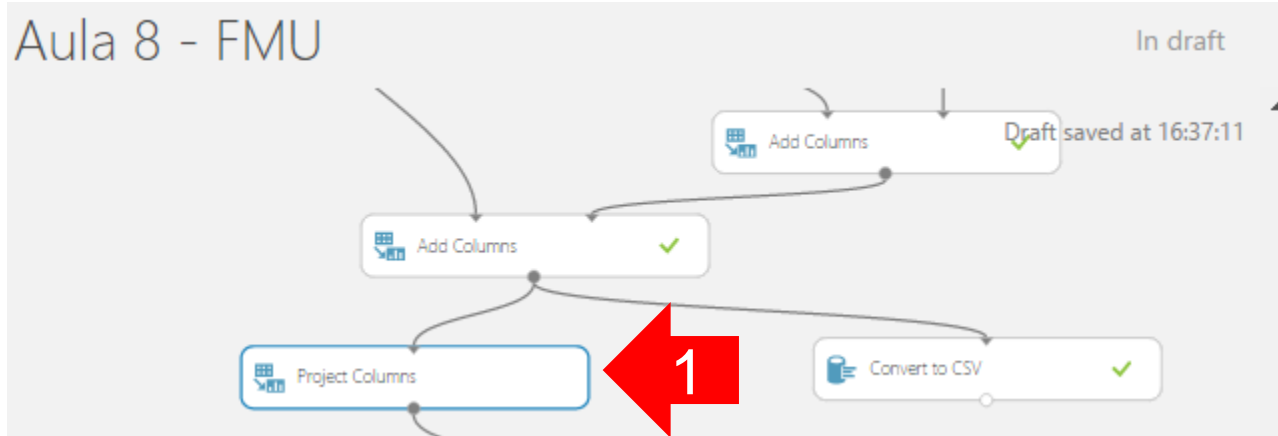
Windows Azure

Visualização parcial



Windows Azure

Excluindo algumas colunas



Properties

Project Columns

Select columns

Selected columns:
All columns
Exclude column names:
Article, Mention, Offset, Type

Launch column selector

Select columns

☐ Allow duplicates and preserve column order in selection

Begin With

All columns

Exclude

column names

Article X

Mention X

Offset X

Type X

3

+

-



4

Windows Azure

Eliminando missing values

The image shows the Windows Azure Machine Learning interface. On the left, a sidebar lists various modules under the 'Deprecated' category. A red arrow labeled '1' points to the 'Missing Values Scrubber' module in this list. The main workspace displays a workflow with three modules: 'Add Columns', 'Project Columns', and 'Missing Values Scrubber'. A red arrow labeled '2' points to the 'Missing Values Scrubber' module in the workflow. A configuration panel for the 'Missing Values Scrubber' module is open on the right, showing settings for handling missing values. A red arrow labeled '3' points to the 'Replace with mean' dropdown menu in this panel.

- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Deprecated
 - Apply Quantization Func...
 - Linear Discriminant Anal...
 - Missing Values Scrubber

Workflow modules:

- Add Columns
- Project Columns
- Missing Values Scrubber

Missing Values Scrubber configuration:

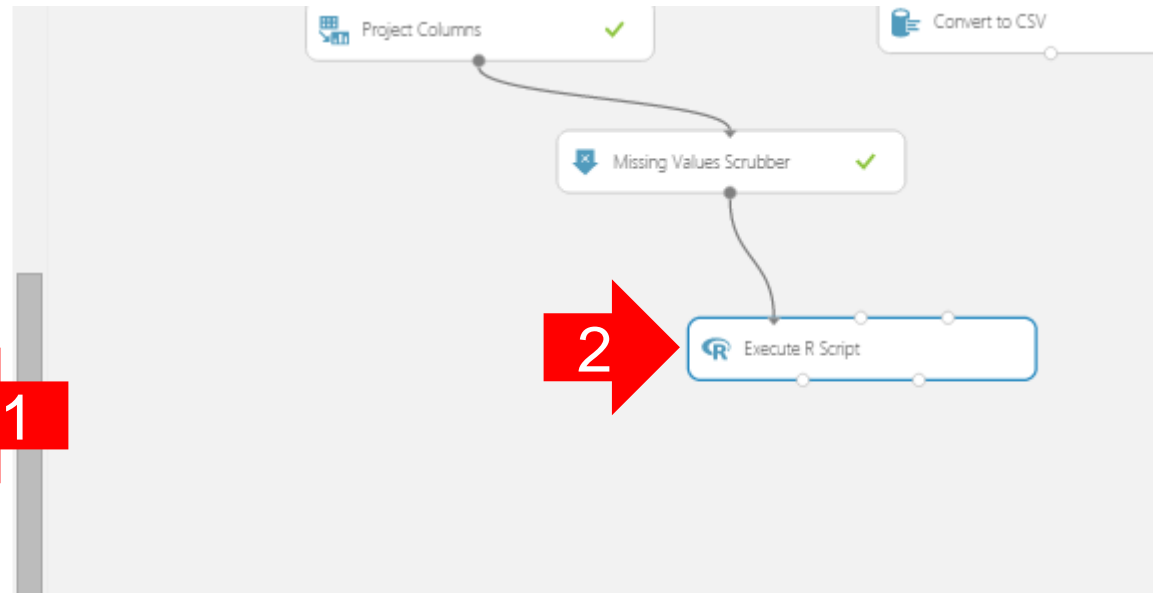
- For missing values: Replace with mean
- Cols with all MV: KeepColumns
- MV indicator column: DoNotGenerate

Windows Azure

Executando um script no R

- ▶ Sample and Split
- ▶ Scale and Reduce
- ▶ Feature Selection
- ▶ Machine Learning
- ▶ OpenCV Library Modules
- ▶ Python Language Modules
- ▶ **R Language Modules**
 - Create R Model
 - Execute R Script

1



Windows Azure

Executando um script no R

▲ Execute R Script

R Script

```
1 # Inserindo os dados
2 base <- mam1.mapInputPort(1) # class: data.frame
3
4 Title_Category = base[,1:2]
5
6 # Análise de Componentes Principais (PCA)
7 pca = prcomp(base[,4:1028])
8 top_pca = data.frame(pca$x[,1:10])
9 dataframe = cbind(Title_Category,top_pca)
10
11 # Visualizando a PCA
12 plot(pca)
13
14 # Enviado o dataframe para a saída
15 mam1.mapOutputPort("dataframe").
```

Windows Azure

Executando um script no R - Resultado

8 - FMU > Execute R Script > Result Dataset

ws
36 columns
12

ew as
1 2

Title	Category	PC1	PC2	PC3	PC4
Apple Inc.	Information Technology	-1196.324169	-136.624916	19.972306	-79.8467
Adobe Systems	Information Technology	-266.248139	-47.893461	35.306568	56.53345
General Motors	Consumer Discretionary	-568.967279	20.830032	-48.233082	21.780172
General Electric	Energy	-394.76521	11.138154	-7.509128	0.695133
Harley-Davidson	Consumer Discretionary	-796.350029	-24.003269	-87.089201	-189.3293
Intel	Information Technology	-1031.574792	-42.587229	29.673068	-35.15889
Microsoft	Information Technology	-510.561424	-42.020697	18.699546	-6.184506
Mattel	Consumer Discretionary	107.893397	1.399807	-8.120238	-10.58427

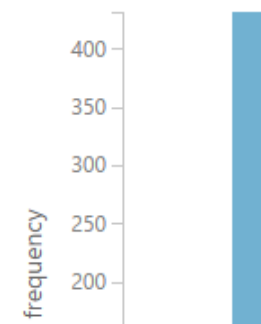
>
Statistics

Visualizations

PC2

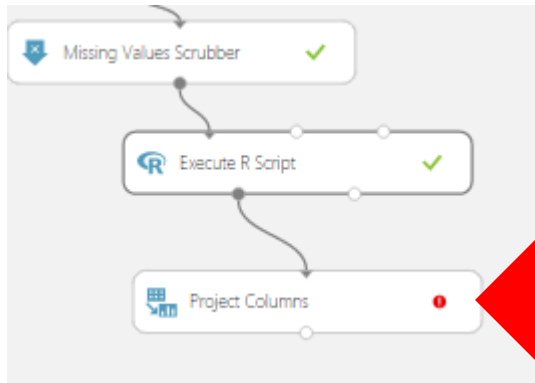
Histogram

compare to None



Windows Azure

K-Means



Select columns

☐ Allow duplicates and preserve column order in selection

Begin With All columns ▼

Exclude ▼ column names ▼ PC1 x

+

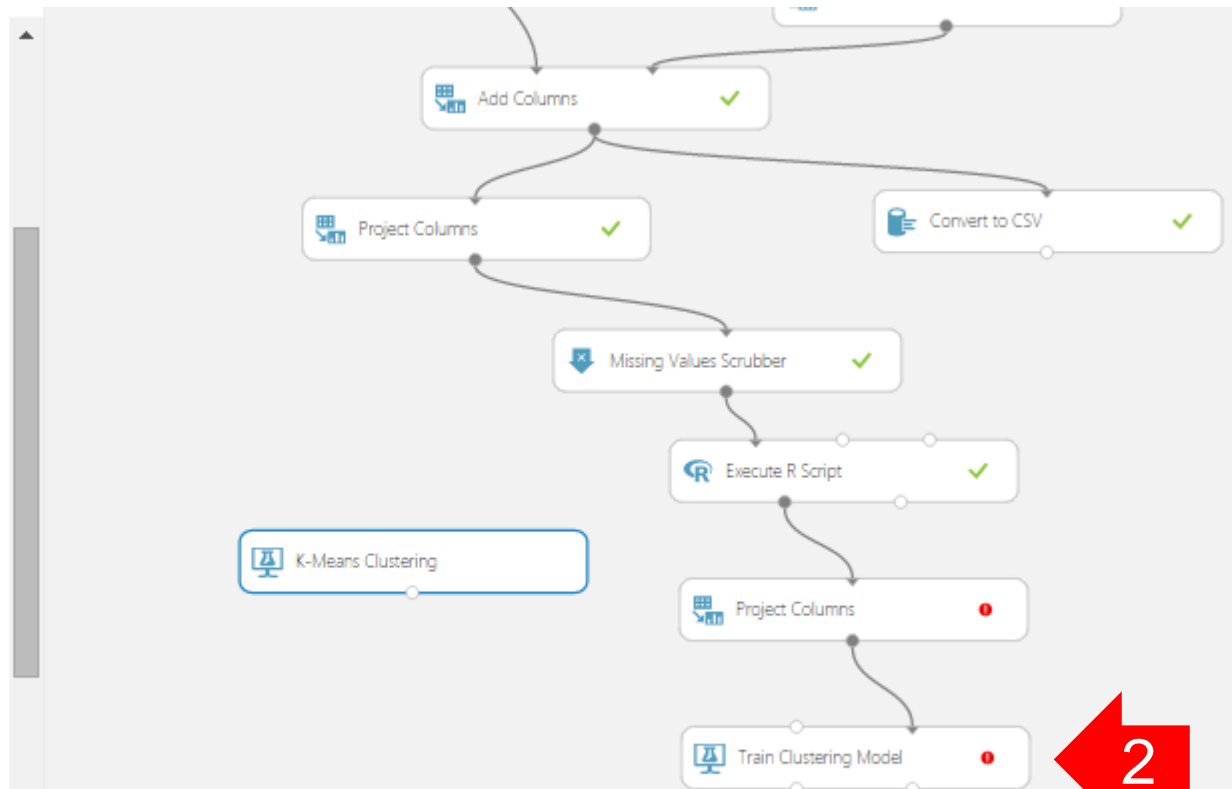
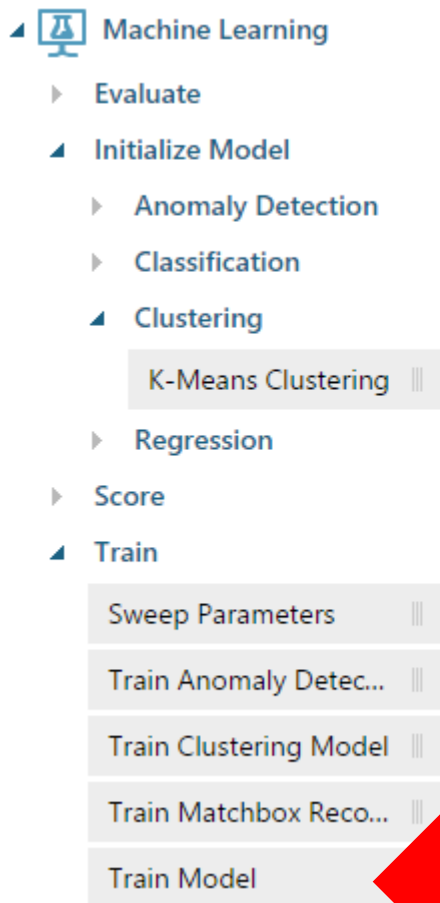
-

3

✓

Windows Azure

K-Means



Select columns

☐ Allow duplicates and preserve column order in selection

Begin With All columns

Include column type Numeric + -

3

Windows Azure K-Means

The screenshot displays the Windows Azure Machine Learning Studio interface. On the left, the 'Machine Learning' sidebar is expanded, showing the 'Train' section with 'K-Means Clustering' selected, indicated by a red arrow labeled '1'. Below this, a list of training tasks is shown, including 'Sweep Parameters', 'Train Anomaly Detec...', 'Train Clustering Model', 'Train Matchbox Reco...', and 'Train Model'. The main workspace shows a workflow diagram with several modules: 'Add Columns', 'Project Columns', 'Convert to CSV', 'Missing Values Scrubber', 'Execute R Script', 'Project Columns', and 'Train Clustering Model'. A red arrow labeled '2' points to the 'K-Means Clustering' module in the workflow. The 'Properties' pane on the right shows the configuration for 'K-Means Clustering', with a red arrow labeled '3' pointing to the 'Number of Centroids' field, which is set to '4'. Other properties include 'Metric' set to 'Euclidean', 'Initialization' set to 'K-Means++', and 'Iterations' set to '100'.

Machine Learning

- ▶ Evaluate
- ▶ Initialize Model
 - ▶ Anomaly Detection
 - ▶ Classification
 - ▶ Clustering
 - K-Means Clustering**
 - ▶ Regression
- ▶ Score
- ▶ Train
 - Sweep Parameters
 - Train Anomaly Detec...
 - Train Clustering Model
 - Train Matchbox Reco...
 - Train Model

Workflow Diagram:

- Add Columns (✓)
- Project Columns (✓)
- Convert to CSV (✓)
- Missing Values Scrubber (✓)
- Execute R Script (✓)
- Project Columns (✗)
- Train Clustering Model (✗)

Properties: K-Means Clustering

- Number of Centroids: 4
- Metric: Euclidean
- Initialization: K-Means++
- Iterations: 100

Windows Azure

K-Means - Resultados

Data Transformation

Filter

Learning with Counts

Manipulation

Add Columns

Add Rows

Apply SQL Transform...

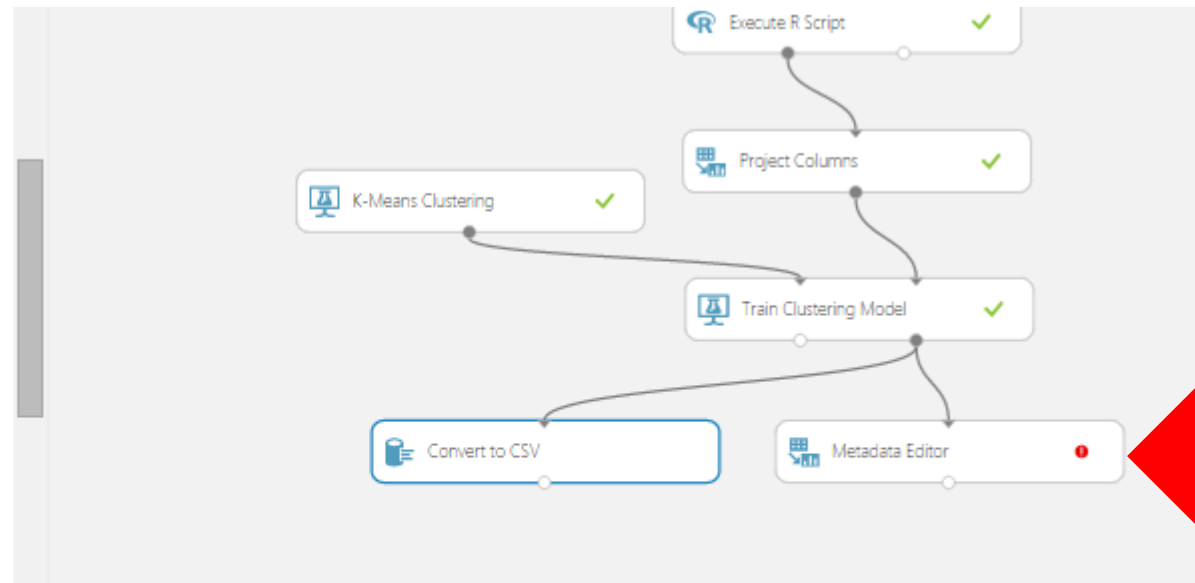
Clean Missing Data

Group Categorical Va...

Indicator Values

Join

Metadata Editor



Select columns

☐ Allow duplicates and preserve column order in selection

Begin With

All columns

Include

column names

Assignments x

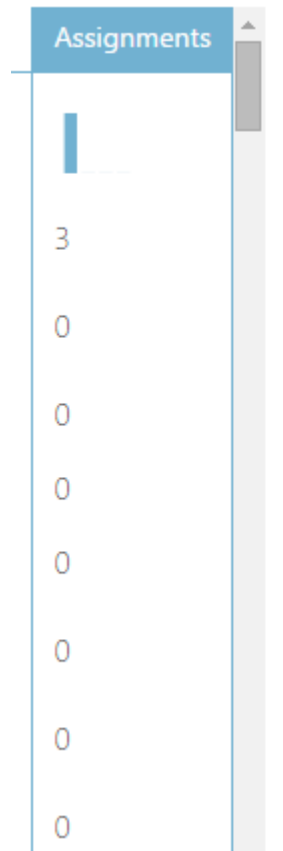


2

3


Windows Azure

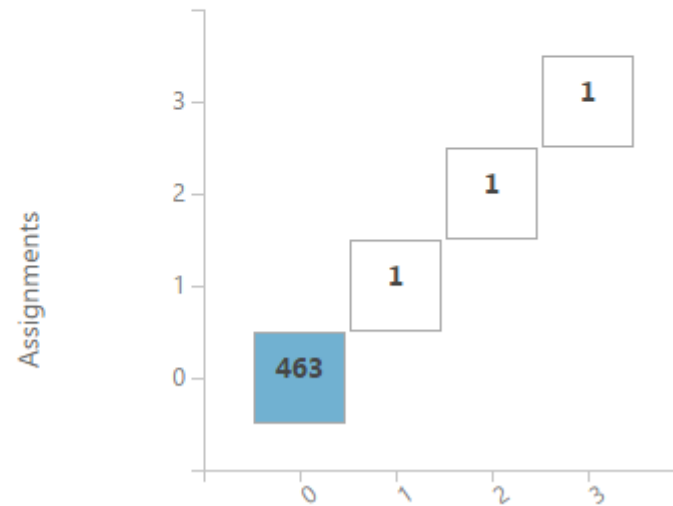
K-Means - Resultados



Assignments

Crosstab

compare to 



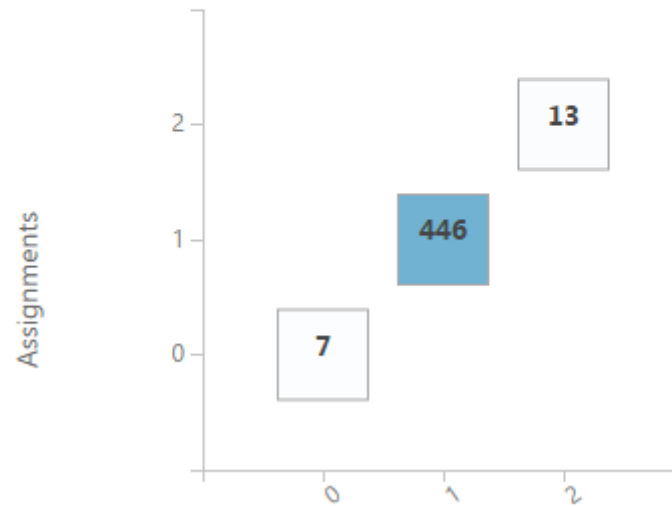
Windows Azure

K-Means – Resultados – Com Bigramas e 3 Clusters

Assignments

Crosstab

compare to Assignments



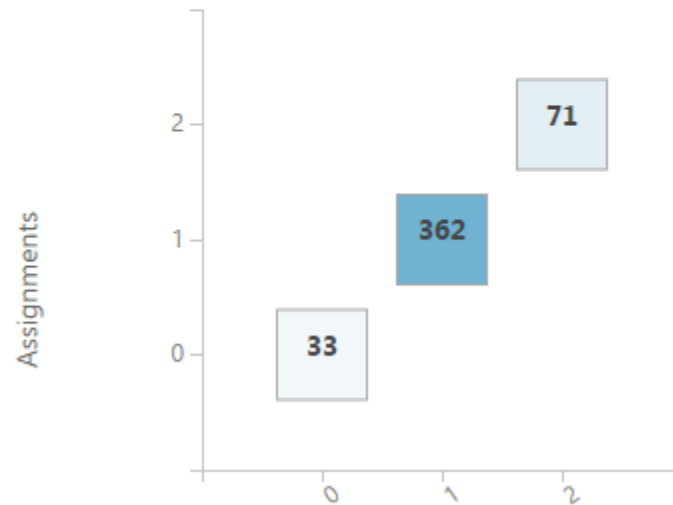
Windows Azure

K-Means – Resultados – Com Hash de 15 bitsize

Assignments

Crosstab

compare to



Windows Azure

Configuração de melhor resultado

Hash com bitsize de 15

Configuração do K-Means



▲ K-Means Clustering

Number of Centroids

10

Metric

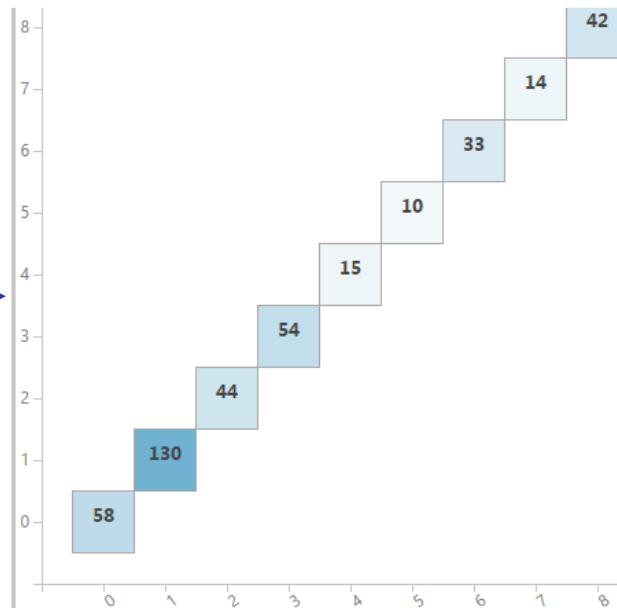
Cosine

Initialization

K-Means++

Iterations

200



Windows Azure

K-Means – Resultados – Com Hash de 24 bitsize

$2^{24} = 16.777.216$ de variáveis!

Execute R Script Error

Record start time:UTC 04/12/2015 00:26:39

Record end time: UTC 04/12/2015 01:20:27

Error message:

Unhandled Exception: OutOfMemoryException.

Business Intelligence