CIS3200-02 Term Project Tutorial

Authors: Samuel Mendoza;Israel Alegria, Hector Pedro-Juan,

Julio Montiel Nava, Marco Rodriguez

Instructor: Jongwook Woo

Date: 05/17/2020

Microsoft Azure Machine Learning

TEAM 2

05/17/2020

Data Processing and Analytics COVID-19 Analysis

Objectives

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

- Prepared the Data
- Execute R Script
- Clean the Data
- SQL Transformation
- Training Clustering Model

Step 1: Preparing the Data

This is to manually retrieve the files and data necessary to mimic our Azure ML portion of the project.

https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0

https://github.com/juliom12/covid_19_cases

1. Open and Login into Azure ML (https://studio.azureml.net/)

2. To upload a database, go to Experiments, click on +NEW, click on Databases, and upload From Local File.

3. Create a new dataset name Covid-19 Cases by uploading Covid_19_combined.csv from the extracted folder files.

Step 2: Create Experiment and Add Modules

This step is to import the database and establish a new experiment canvas in AzureML.

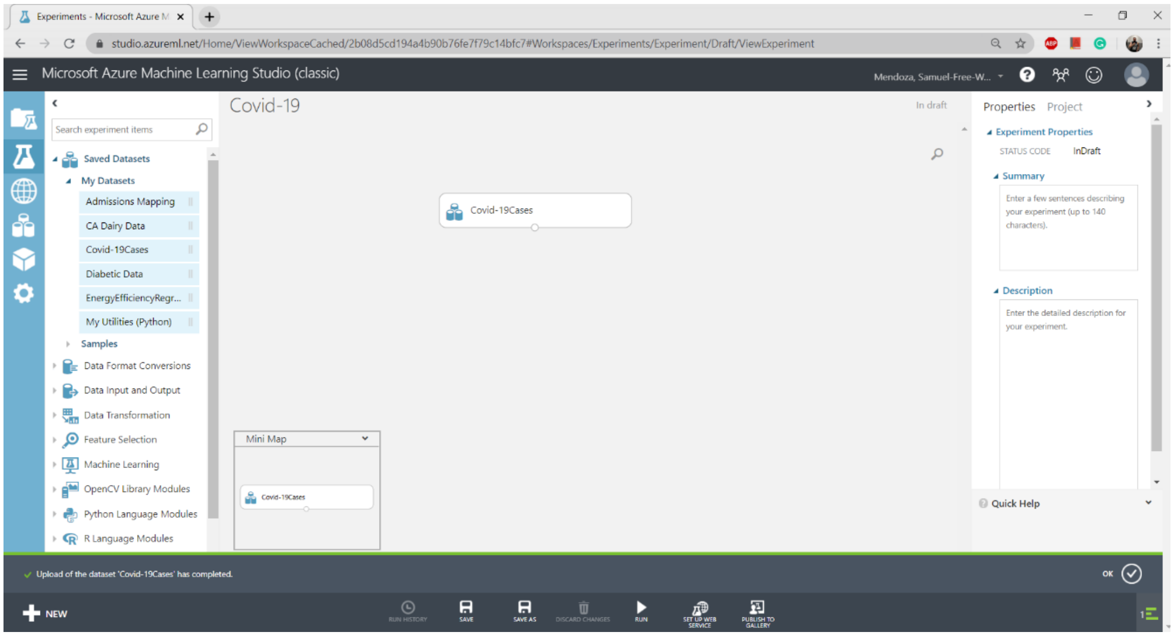
4. In the Studio, click on the bottom left (**+NEW**), Experiment → Blank Experiment

5. To upload a database, go to Experiments, click on +NEW, click on Databases, and upload From Local File.

6. Create a new dataset name **Covid-19 Cases** by uploading **Covid_19_combined.csv** from the extracted folder files.

7. Change the title of experiment to Covid-19

8. On the left side of the experiments pane, expand **Saved Datasets,** expand **My Datasets,** and click and drag Covid-19 Cases to the canvas.



STATUS CODE InDraft

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

Enter the detailed description for your experiment.

Upload of the dataset 'Covid-19Cases' has completed.

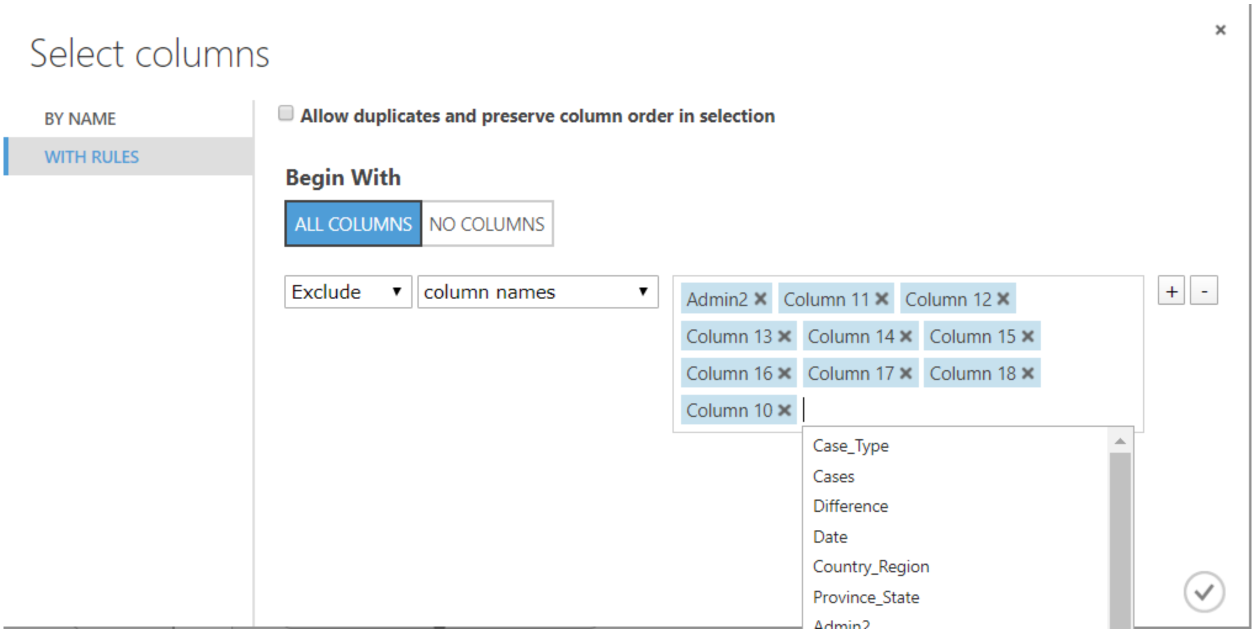
Step 3: Excluding Columns

This step allows use to remove columns that are not necessary to the project to better optimize the data.

9. In the search box of the experiment pane, type **Select Columns in Dataset**, drag Select Columns in Dataset under Covid-19 Cases. Click the output port of Covid-19 Cases and drag it to the input of Select Columns in Dataset.

10. Click on **Select Columns in Dataset,** on the properties pane, click on **Launch Column Selector**

11. **With Rules,** click **All Columns** and **Exclude Column Names:** Admin2, Column 11, Column12, Column 13, Column 14, Column 15, Column 16, Column 17, Column 18, and Column 10



Select columns

BY NAME

WITH RULES

□ Allow duplicates and preserve column order in selection

Begin With

ALL COLUMNS NO COLUMNS

Exclude ▼ column names ▼

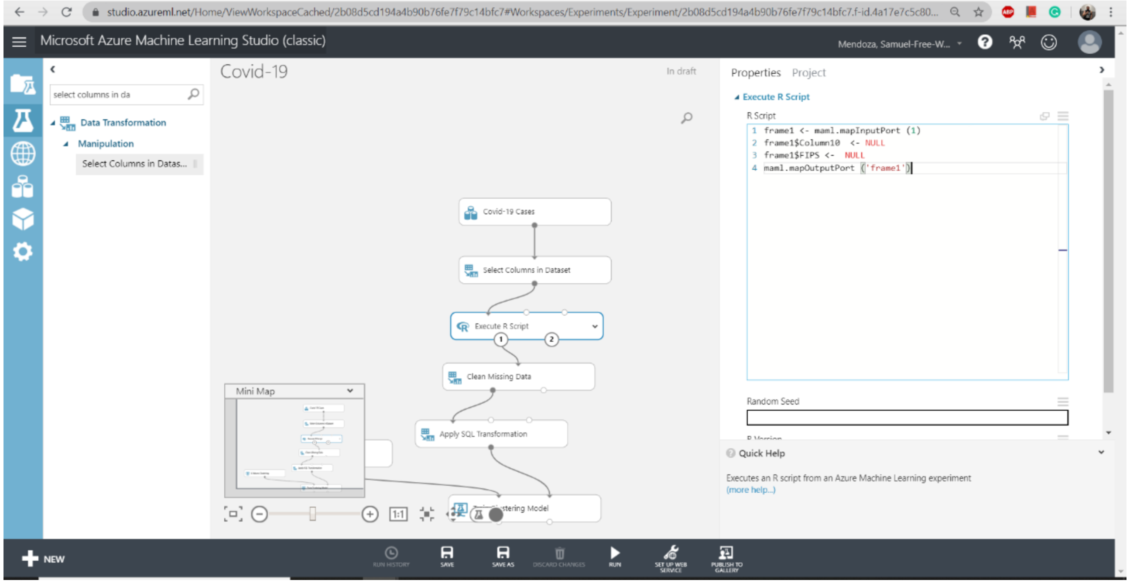
Admin2 ✖ Column 11 ✖ Column 12 ✖ Column 13 ✖ Column 14 ✖ Column 15 ✖ Column 16 ✖ Column 17 ✖ Column 18 ✖ Column 10 ✖

Step 4: Execute R Script

The R Script allows us to remove the FIPS Column to better optimize the dataset.

12. In the search box of the experiment pane, type **Execute R Script**, drag Execute R Script under Select Columns in Dataset. Click the output port of Select Columns in Dataset and connect it to leftist input of Execute R Script.

13. Download and open the Azure code from the GitHub. Copy the Execute R Script. Back in the studio, go to the Execute R Script properties pane. Select the current code in R script and paste the code from the Azure code file.


studio.azureml.net/Home/ViewWorkspaceCached/2b08d5cd194a4b90b76fe7f79c14bfc7#Workspaces/Experiments/Experiment/2b08d5cd194a4b90b76fe7f79c14bfc7.f-id.4a17e7c5c80...
Microsoft Azure Machine Learning Studio (classic)
Mendoza, Samuel-Free-W...
select columns in da
Data Transformation
Manipulation
Select Columns in Datas...
Covid-19
In draft
Covid-19 Cases
Select Columns in Dataset
Execute R Script
Clean Missing Data
Apply SQL Transformation
Mini Map
Properties
Project
Execute R Script
R Script
Random Seed
Quick Help
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

```
frame1 <- maml.mapInputPort (1)
frame1$Column10  <- NULL
frame1$FIPS <-  NULL
maml.mapOutputPort ('frame1')
```

Executes an R script from an Azure Machine Learning experiment
(more help...)

EXECUTE R SCRIPT:

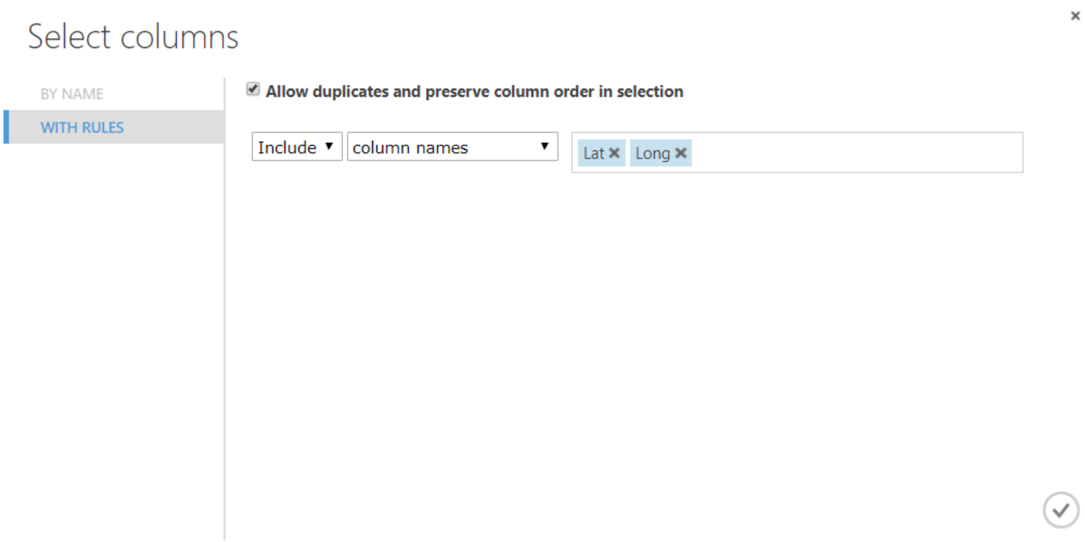
```
frame1 <- maml.mapInputPort (1)

frame1$Column10  <- NULL

frame1$FIPS <-  NULL

maml.mapOutputPort ('frame1')
```

Step 5: Clean Missing Data

Using the Clean Missing Data allows us to remove the missing information from Latitude and Longitude to better optimize these columns for visualization.

14. In the search box of the experiment pane, type **Clean Missing Data** and drag it under Execute R Script. Click the left-output port of Execute R Script and connect it to input of Clean Missing Data.

15. In the properties pane of Clean Missing Data, Launch Column Selector. With Rules Include Column Names LAT and LONG. Check Allow duplicates and preserve column order in selection.



Select columns

BY NAME

WITH RULES

☑ Allow duplicates and preserve column order in selection

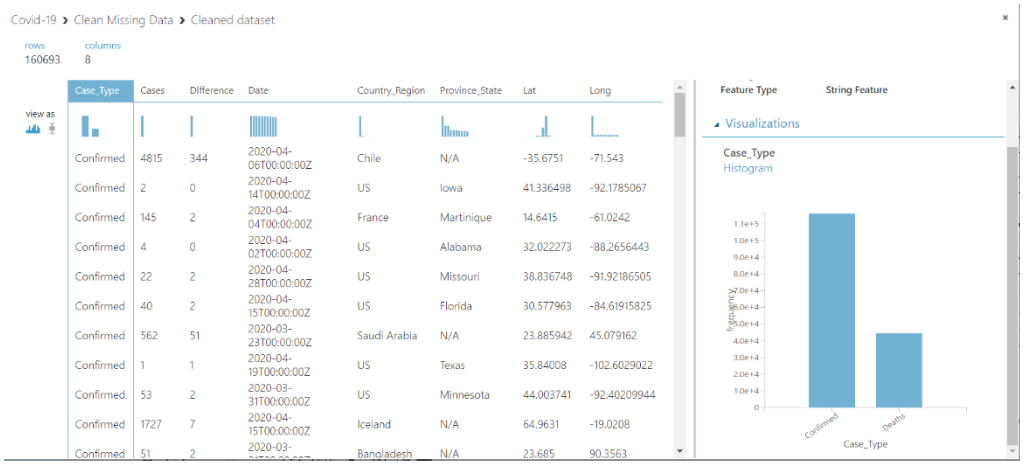
Include ▾ column names ▾ Lat ✖ Long ✖

16. In the Properties Pane:

    - **Minimum missing value ratio:** 0
    - **Maximum missing value ratio:** 1
    - **Cleaning Mode:** Custom Substitution Value
    - **Replacement Value:** 0
    - **Uncheck Generate missing value indicator column.**

17. **Save and Run**

18. After it is finished running, look for the **Green checkmark,** Visualize the **Cleaned dataset** to see the histogram of Case_Types



Covid-19 › Clean Missing Data › Cleaned dataset

rows 160693 columns 8

Case_Type	Cases	Difference	Date	Country_Region	Province_State	Lat	Long
Confirmed	4815	344	2020-04-06T00:00:00Z	Chile	N/A	-35.6751	-71.543
Confirmed	2	0	2020-04-14T00:00:00Z	US	Iowa	41.336498	-92.1785067
Confirmed	145	2	2020-04-04T00:00:00Z	France	Martinique	14.6415	-61.0242
Confirmed	4	0	2020-04-02T00:00:00Z	US	Alabama	32.022273	-88.2656443
Confirmed	22	2	2020-04-28T00:00:00Z	US	Missouri	38.836748	-91.92186505
Confirmed	40	2	2020-04-15T00:00:00Z	US	Florida	30.577963	-84.61915825
Confirmed	562	51	2020-03-23T00:00:00Z	Saudi Arabia	N/A	23.885942	45.079162
Confirmed	1	1	2020-04-19T00:00:00Z	US	Texas	35.84008	-102.6029022
Confirmed	53	2	2020-03-31T00:00:00Z	US	Minnesota	44.003741	-92.40209944
Confirmed	1727	7	2020-04-15T00:00:00Z	Iceland	N/A	64.9631	-19.0208
Confirmed	51	2	2020-03-	Bangladesh	N/A	23.685	90.3563

Feature Type String Feature

Visualizations

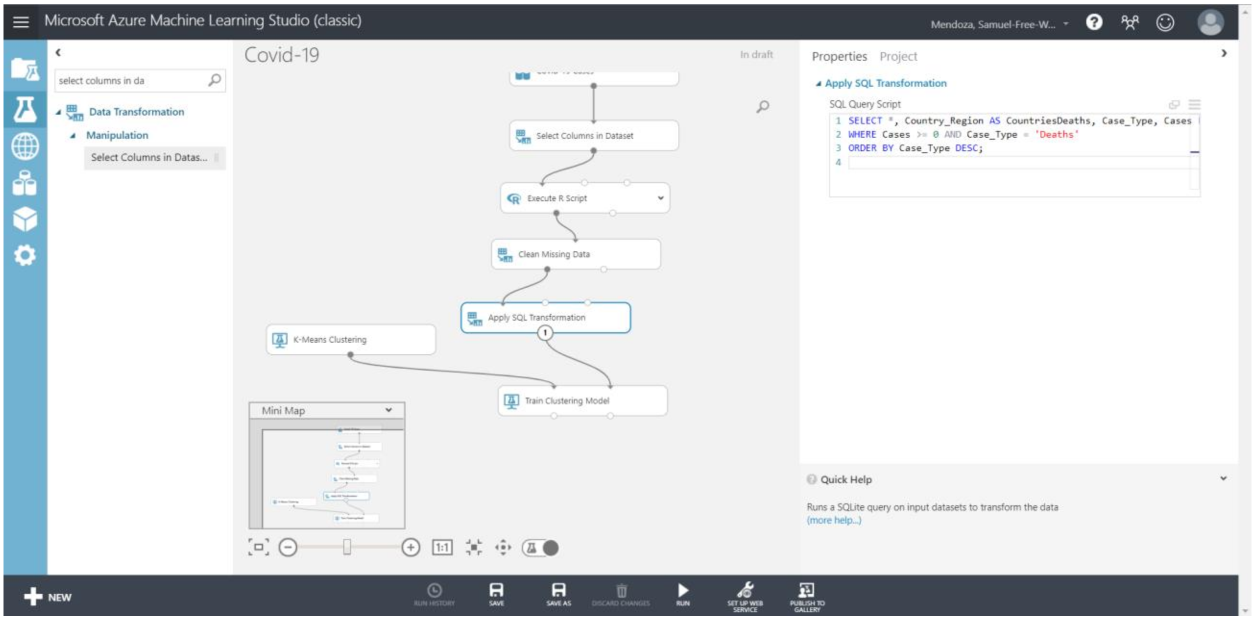
Case_Type
Histogram

Step 6: Apply SQL Transformation

Using the SQL Query Script, we query the dataset by creating a view for the death count instead of viewing the both death and confirmed cases.

19. In the search box of the experiment pane, type **Apply SQL Transformation** and drag it under Clean Missing Data. Click the left-output port of Clean Missing Data and connect it to the left input of Apply SQL Transformation.

20. Open the Azure code from the Github, Copy the Apply SQL Transformation script. Back in the studio, go to the Apply SQL Transformation properties pane, click in the SQL Query Script, Select the current script and paste the new SQL script.



```
SELECT *, Country_Region AS CountriesDeaths, Case_Type, Cases
WHERE Cases >= 0 AND Case_Type = 'Deaths'
ORDER BY Case_Type DESC;
```

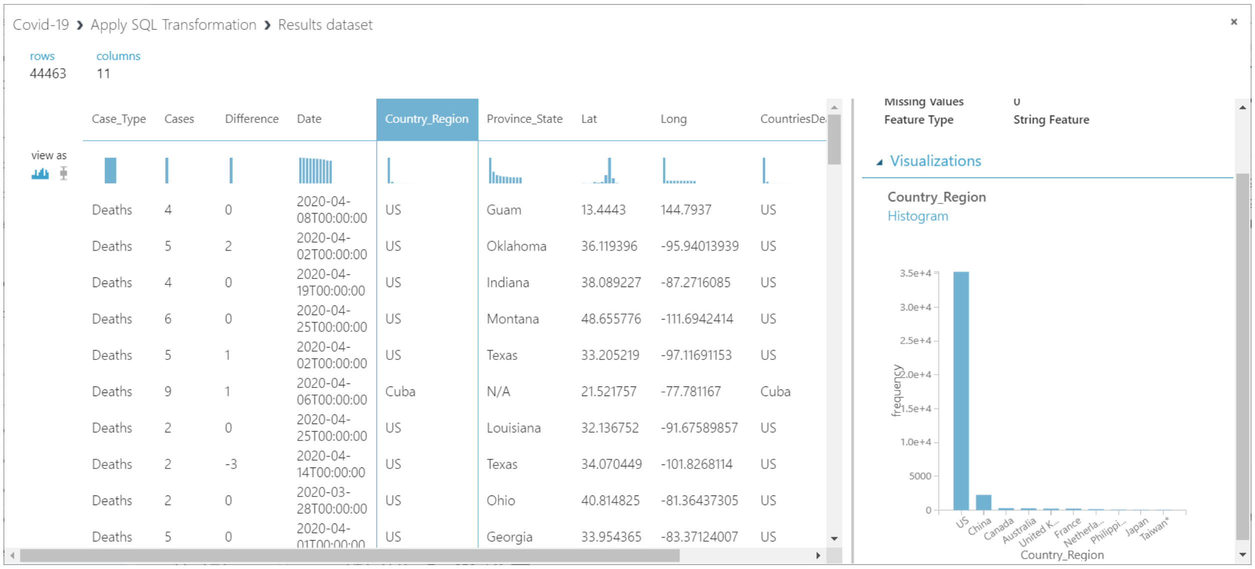
Runs a SQLite query on input datasets to transform the data
(more help...)

<u>APPLY SQL TRANSFORMATION:</u>

```
SELECT *, Country_Region, Cases, Case_Type FROM t1

WHERE Cases >= 0 AND Case_Type = 'Deaths'

ORDER BY Case_Type DESC;
```

21. **Save and Run**

22. After it is finished running, look for the **Green checkmark,** Visualize the **Result dataset** of the SQL Transformation. Here you will be able to see more statistics plus the A Histogram on the Column County_Region where you can see the comparison between the United States and some of the other nations from around the world.



Covid-19 › Apply SQL Transformation › Results dataset

rows 44463

columns 11

Case_Type	Cases	Difference	Date	Country_Region	Province_State	Lat	Long	CountriesDe
Deaths	4	0	2020-04-08T00:00:00	US	Guam	13.4443	144.7937	US
Deaths	5	2	2020-04-02T00:00:00	US	Oklahoma	36.119396	-95.94013939	US
Deaths	4	0	2020-04-19T00:00:00	US	Indiana	38.089227	-87.2716085	US
Deaths	6	0	2020-04-25T00:00:00	US	Montana	48.655776	-111.6942414	US
Deaths	5	1	2020-04-02T00:00:00	US	Texas	33.205219	-97.11691153	US
Deaths	9	1	2020-04-06T00:00:00	Cuba	N/A	21.521757	-77.781167	Cuba
Deaths	2	0	2020-04-25T00:00:00	US	Louisiana	32.136752	-91.67589857	US
Deaths	2	-3	2020-04-14T00:00:00	US	Texas	34.070449	-101.8268114	US
Deaths	2	0	2020-03-28T00:00:00	US	Ohio	40.814825	-81.36437305	US
Deaths	5	0	2020-04-01T00:00:00	US	Georgia	33.954365	-83.37124007	US

Missing Values 0

Feature Type String Feature

Step 7: K-Means Clustering

K-Means Clustering allows us to group together certain parts of the data and randomly initialize center in hopes to prepare data for a trained model algorithm.

23. In the search box of the experiment pane, type **K-Means Clustering** and drag it to the left of Apply SQL Transformation.

24. In the K-Means Clustering properties pane apply the following:

- **Create trainer mode:** Single Parameter
- **Number of Centroids:** 2
- **Initialization:** Random
- **Random number seed:** 2345
- **Metric:** Euclidean
- **Iteration:** 100
- **Assign Label Mode:** Fill Missing Values

Step 8: Train Clustering Model

Taking the previous step in clustering the data, we train it to output a visualization to compare the scope of confirmed cases to deaths.

25. In the search box of the experiment pane, type **Train Clustering Model** and drag it under Apply SQL Transformation. Click the output port of Apply SQL Transformation and connect it to the right input of Train Clustering Model.

26. Connect **K-Means Clustering** Output port into **Train Clustering Model** left input.

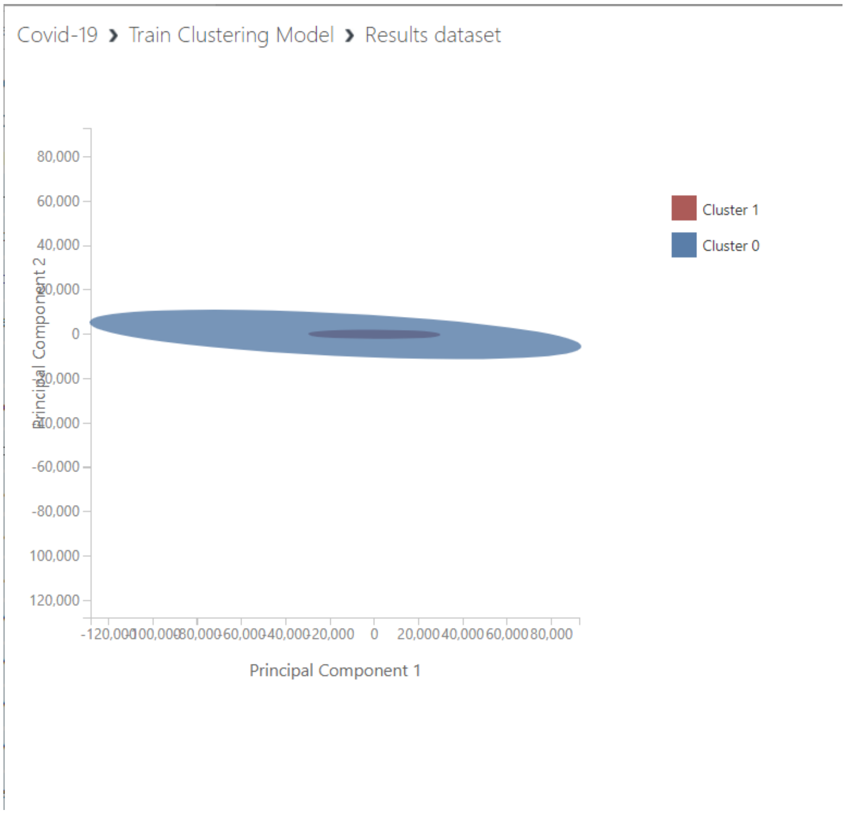
27. In the Train Clustering Model properties pane, launch column selector.

28. With Rules, No Columns → Include Column names →Case_Type, Difference, Cases, CountriesDeaths. Apply changes.

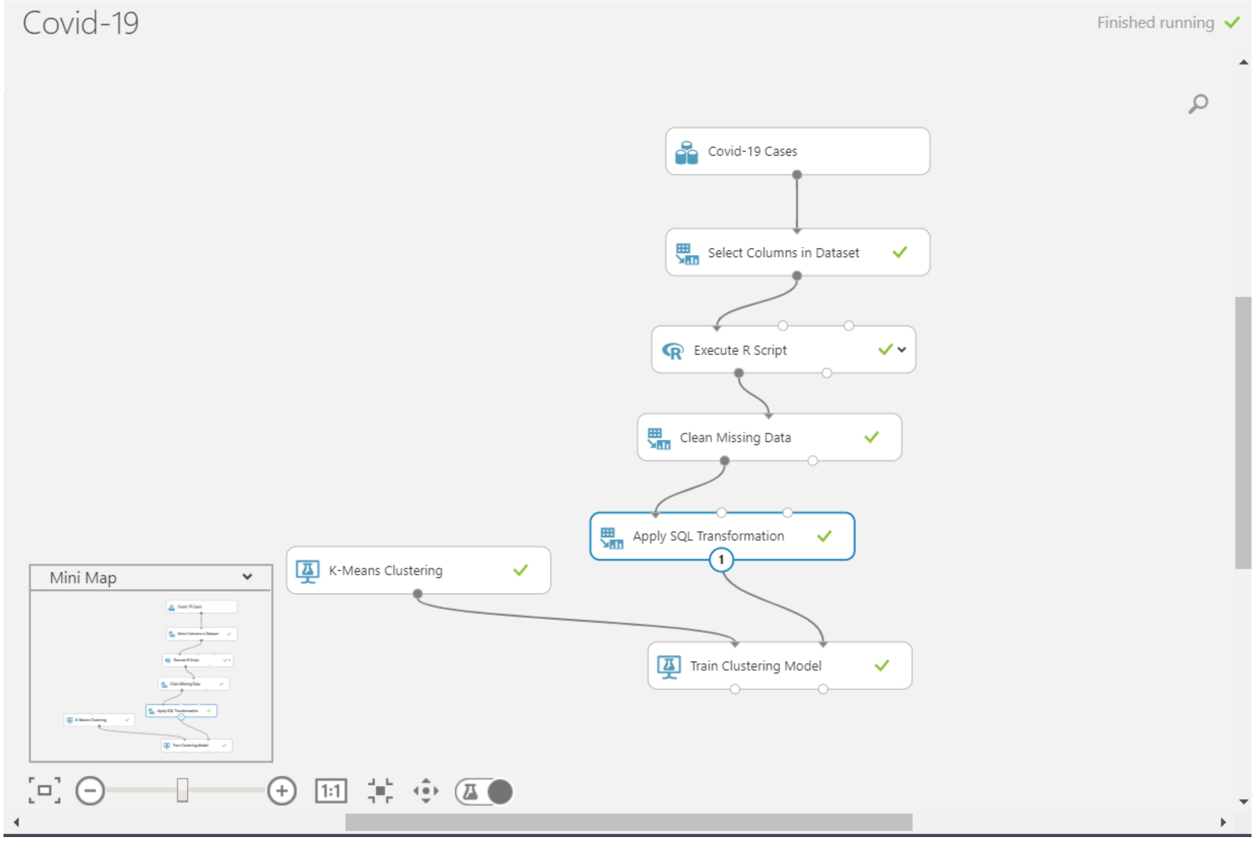
29. Check for Append or Uncheck for Result Only.

30. **Save and Run**

31. After it is finished running, look for the **Green checkmark,** Visualize **Result Dataset** of the Train Clustering Model. You will see a comparison between confirmed and death from Covid-19 Virus.



32. By the end, your Canvas should look like this:



References

1. Data World, John Hopkins University Dataset, https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv
2. GitHub, https://github.com/juliom12/covid_19_cases
3. Dropbox, https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0
4. Microsoft Azure ML, https://studio.azureml.net/