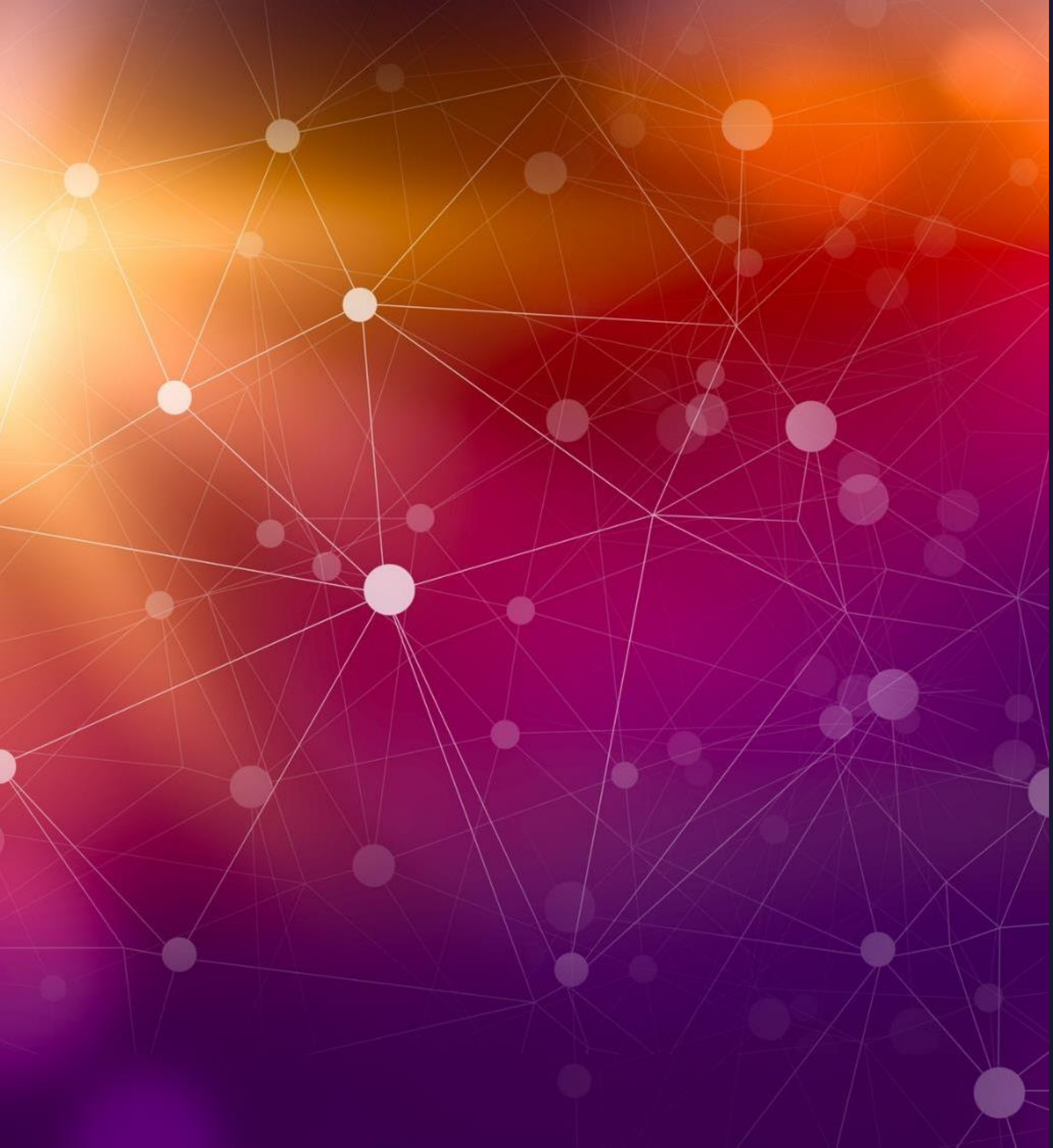


CEIS312 Introduction to Artificial Intelligence & Machine Learning

Final Course Project

Developed by James Garlie

DeVry University

April 2024

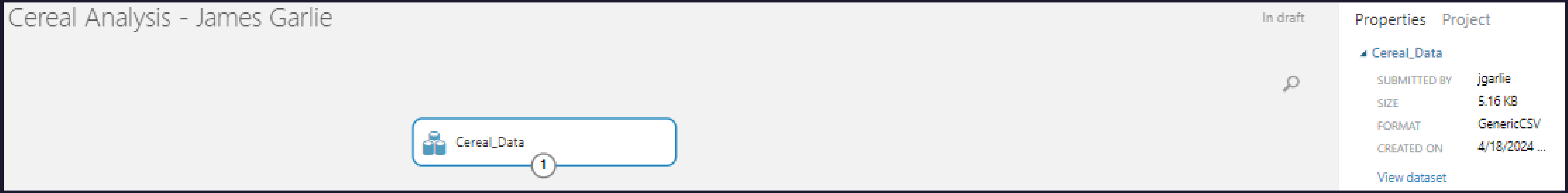
Introduction

Artificial intelligence and machine learning are frontiers in the technology field. These areas are often used to address common problems that require difficult tools or skills. AI and ML professionals work with SQL, R, Python, and other tools specific to data science. Different algorithms are used to solve problems and choosing the correct algorithm can be challenging. This project will use Azure Machine Learning, which is a cloud-based service from Microsoft. Azure ML allows you to create and run experiments based on datasets and integrate custom code in SQL, R, or Python.

The presentation concludes with the Conclusion, Challenges, and Career Skills obtained.

Uploaded Dataset

The first image shows that an experiment has been created titled Cereal Analysis and that a .csv file, titled Cereal Data has been uploaded from my local device and added to the experiment. Note the GenericCSV format and file size shown on the right of the image.



Cereal Analysis - James Garlie

In draft

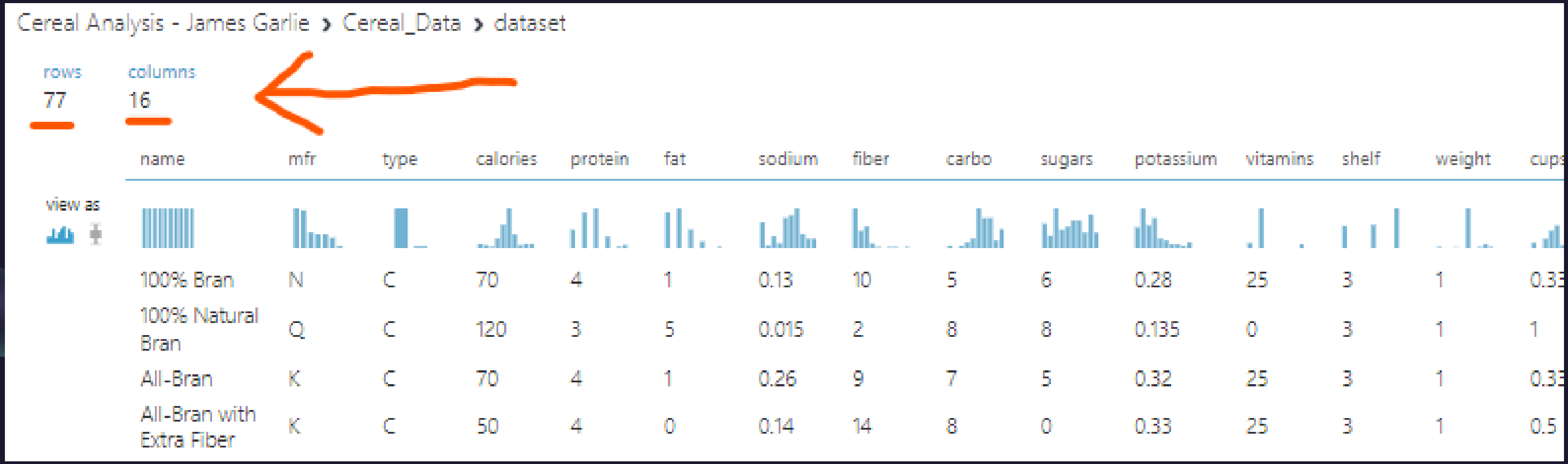
Properties Project

Cereal_Data

SUBMITTED BY	jgarlie
SIZE	5.16 KB
FORMAT	GenericCSV
CREATED ON	4/18/2024 ...

View dataset

This second image shows the initial visualization of the Cereal Dataset with *77* rows & 16 columns.



Cereal Analysis - James Garlie › Cereal_Data › dataset

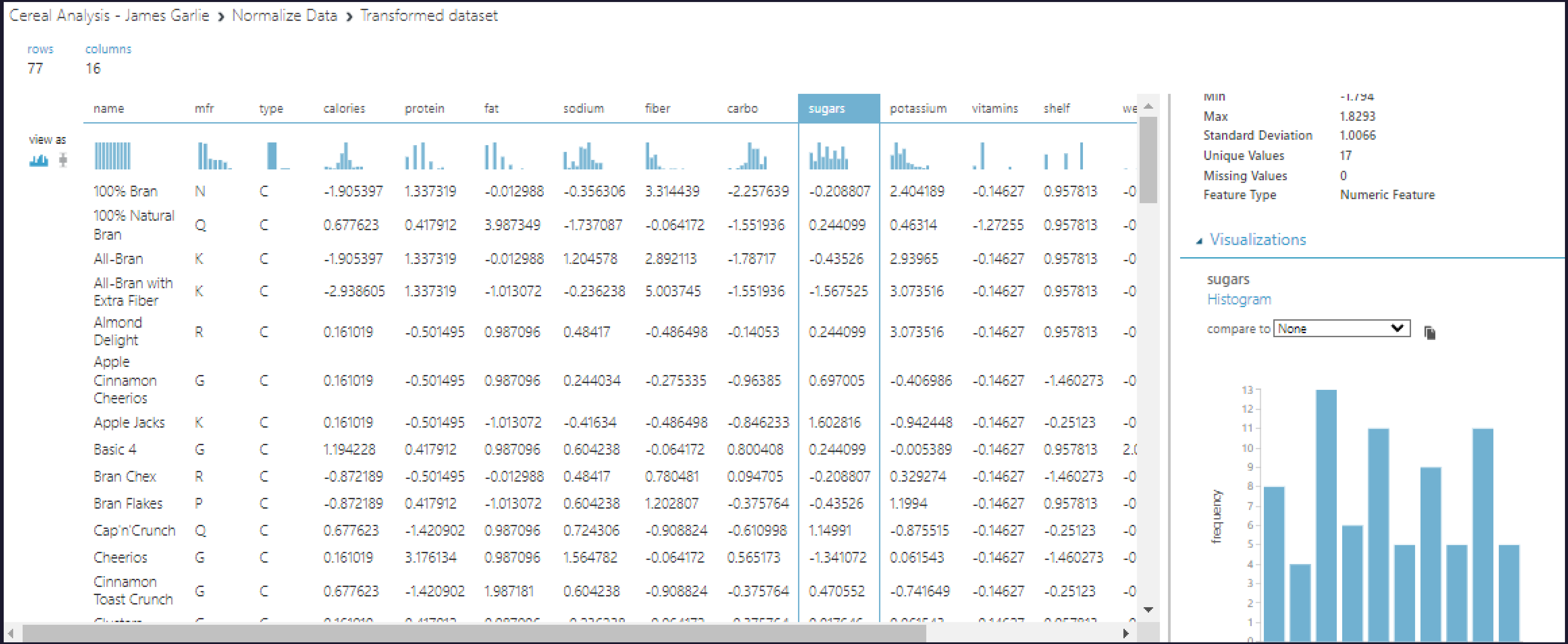
rows 77

columns 16

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potassium	vitamins	shelf	weight	cups
100% Bran	N	C	70	4	1	0.13	10	5	6	0.28	25	3	1	0.33
100% Natural Bran	Q	C	120	3	5	0.015	2	8	8	0.135	0	3	1	1
All-Bran	K	C	70	4	1	0.26	9	7	5	0.32	25	3	1	0.33
All-Bran with Extra Fiber	K	C	50	4	0	0.14	14	8	0	0.33	25	3	1	0.5

Data Preparation & Normalization

This slide shows filtering the data, formatting, and normalization.



Cereal Analysis - James Garlie › Normalize Data › Transformed dataset

rows 77

columns 16

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potassium	vitamins	shelf	we
100% Bran	N	C	-1.905397	1.337319	-0.012988	-0.356306	3.314439	-2.257639	-0.208807	2.404189	-0.14627	0.957813	-0
100% Natural Bran	Q	C	0.677623	0.417912	3.987349	-1.737087	-0.064172	-1.551936	0.244099	0.46314	-1.27255	0.957813	-0
All-Bran	K	C	-1.905397	1.337319	-0.012988	1.204578	2.892113	-1.78717	-0.43526	2.93965	-0.14627	0.957813	-0
All-Bran with Extra Fiber	K	C	-2.938605	1.337319	-1.013072	-0.236238	5.003745	-1.551936	-1.567525	3.073516	-0.14627	0.957813	-0
Almond Delight	R	C	0.161019	-0.501495	0.987096	0.48417	-0.486498	-0.14053	0.244099	3.073516	-0.14627	0.957813	-0
Apple Cinnamon Cheerios	G	C	0.161019	-0.501495	0.987096	0.244034	-0.275335	-0.96385	0.697005	-0.406986	-0.14627	-1.460273	-0
Apple Jacks	K	C	0.161019	-0.501495	-1.013072	-0.41634	-0.486498	-0.846233	1.602816	-0.942448	-0.14627	-0.25123	-0
Basic 4	G	C	1.194228	0.417912	0.987096	0.604238	-0.064172	0.800408	0.244099	-0.005389	-0.14627	0.957813	2.
Bran Chex	R	C	-0.872189	-0.501495	-0.012988	0.48417	0.780481	0.094705	-0.208807	0.329274	-0.14627	-1.460273	-0
Bran Flakes	P	C	-0.872189	0.417912	-1.013072	0.604238	1.202807	-0.375764	-0.43526	1.1994	-0.14627	0.957813	-0
Cap'n'Crunch	Q	C	0.677623	-1.420902	0.987096	0.724306	-0.908824	-0.610998	1.14991	-0.875515	-0.14627	-0.25123	-0
Cheerios	G	C	0.161019	3.176134	0.987096	1.564782	-0.064172	0.565173	-1.341072	0.061543	-0.14627	-1.460273	-0
Cinnamon Toast Crunch	G	C	0.677623	-1.420902	1.987181	0.604238	-0.908824	-0.375764	0.470552	-0.741649	-0.14627	-0.25123	-0

Min	-1.794
Max	1.8293
Standard Deviation	1.0066
Unique Values	17
Missing Values	0
Feature Type	Numeric Feature

Visualizations

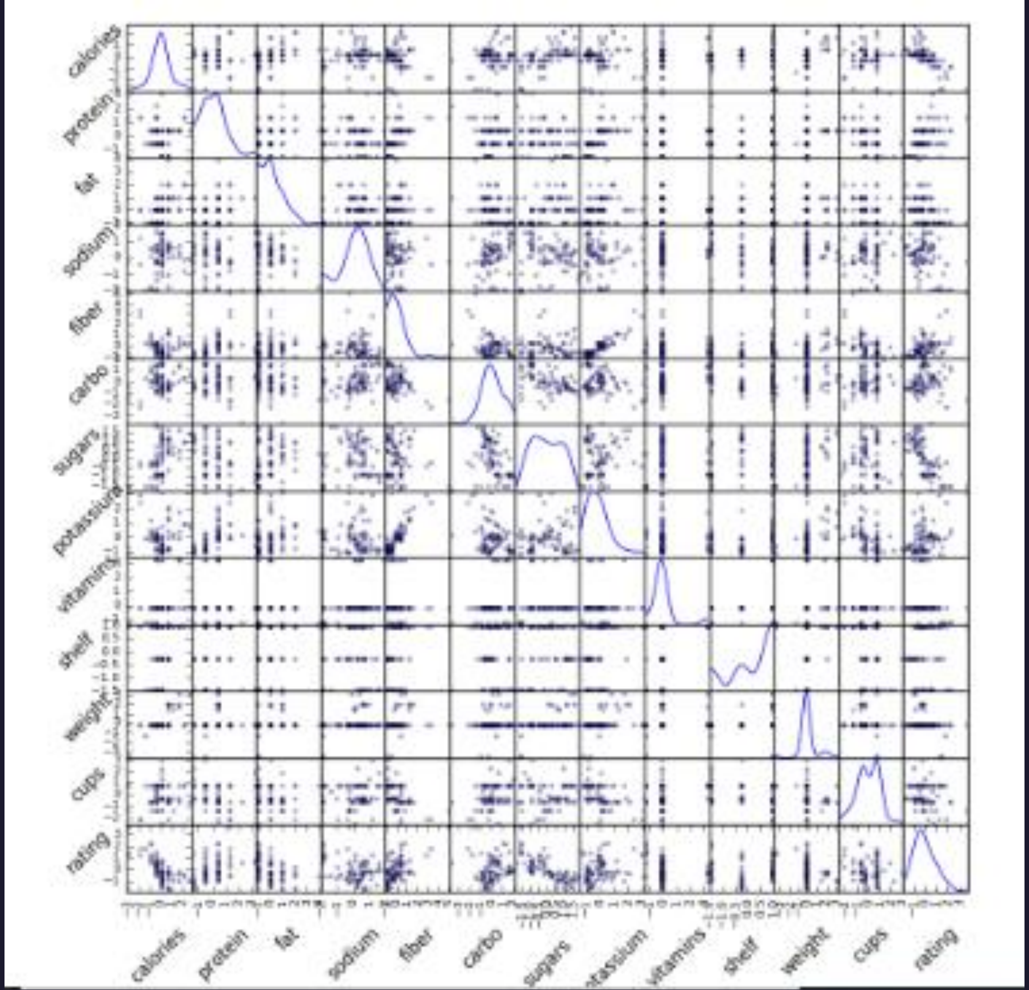
sugars Histogram

compare to None

Data Visualization

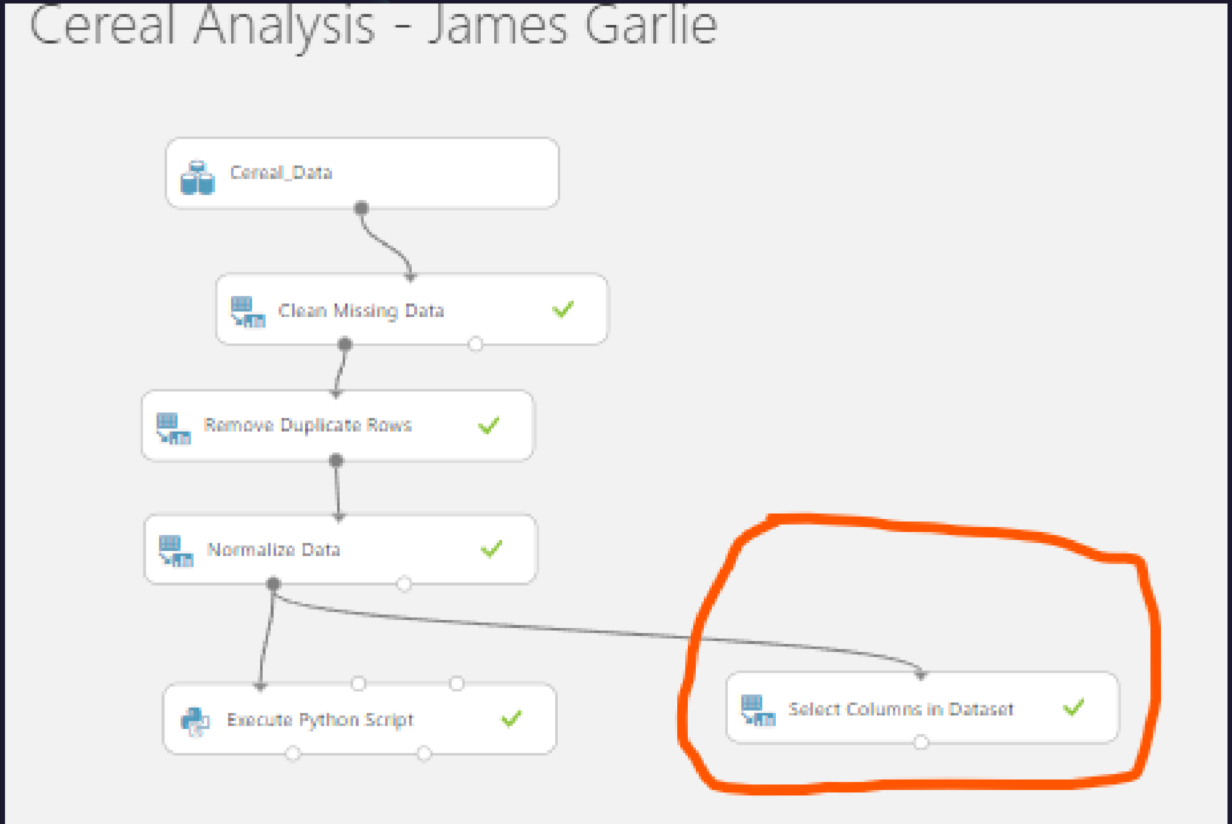
This slide shows the Python script used for the visualization of the data.

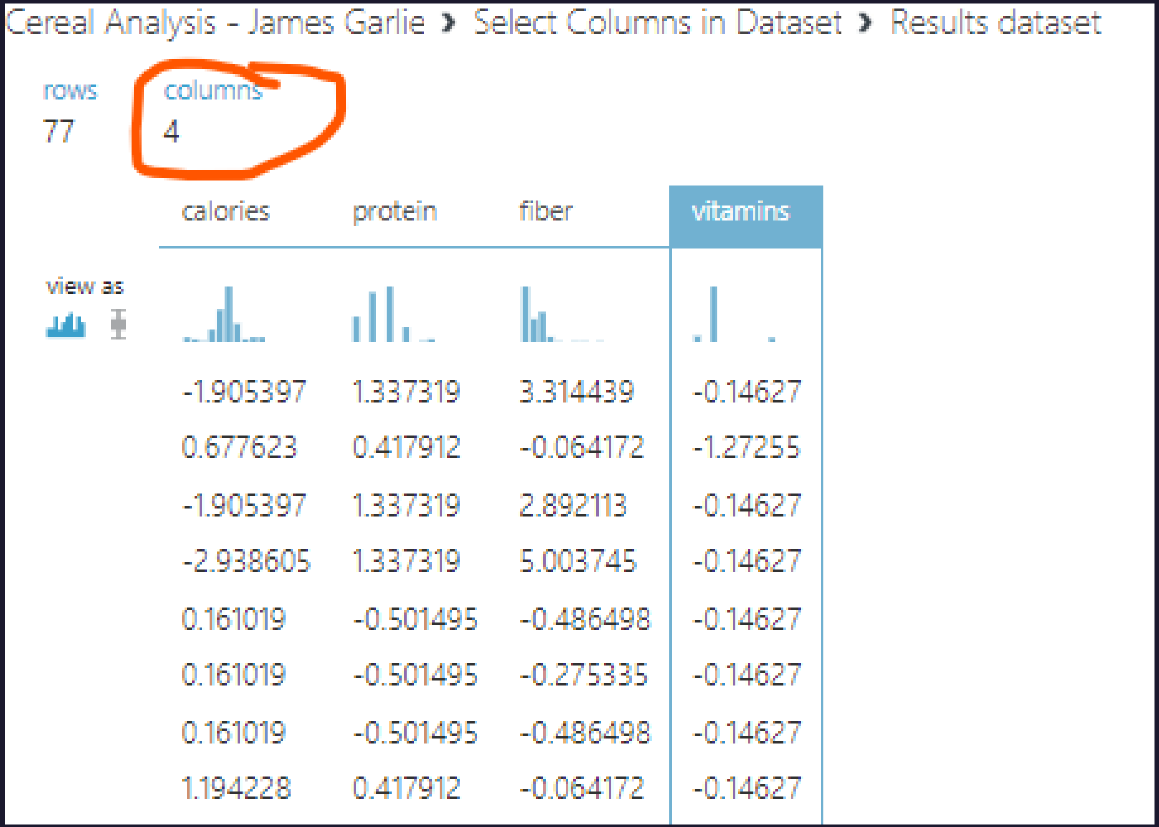
```
def azureml_main(frame1):

## import libraries
    import matplotlib
    matplotlib.use('agg')  # Set backend

    from pandas.tools.plotting import scatter_matrix
    import pandas.tools.rplot as rplot
    import matplotlib.pyplot as plt
    import numpy as np

## Create a pair-wise scatter plot
    Azure = True


    fig1 = plt.figure(1, figsize=(10, 10))
    ax = fig1.gca()
    sm=scatter_matrix(frame1, alpha=0.3,
                diagonal='kde', ax = ax)
    [s.xaxis.label.set_rotation(45) for s in sm.reshape(-1)]
    [s.yaxis.label.set_rotation(45) for s in sm.reshape(-1)]

    plt.show()
    fig1.savefig('scatter1.png')
```



Selecting Features

This slide shows that I added “Select Columns in Dataset”, then selected calories, protein, fiber, and vitamins. I then saved and ran the experiment. The visualization shows the 4 features I selected. Also note we are down to 4 columns verses the original 16.





Cereal Analysis - James Garlie › Select Columns in Dataset › Results dataset

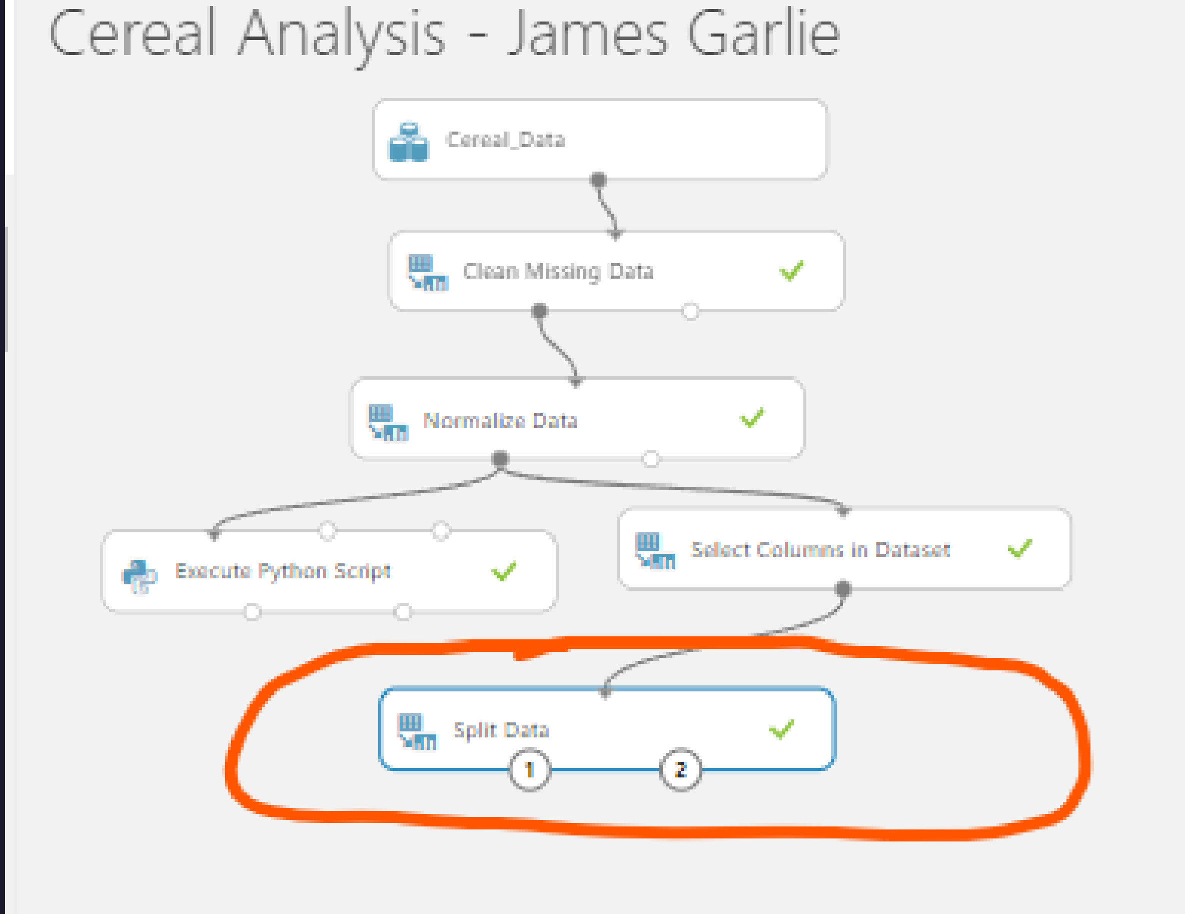
rows 77

columns 4

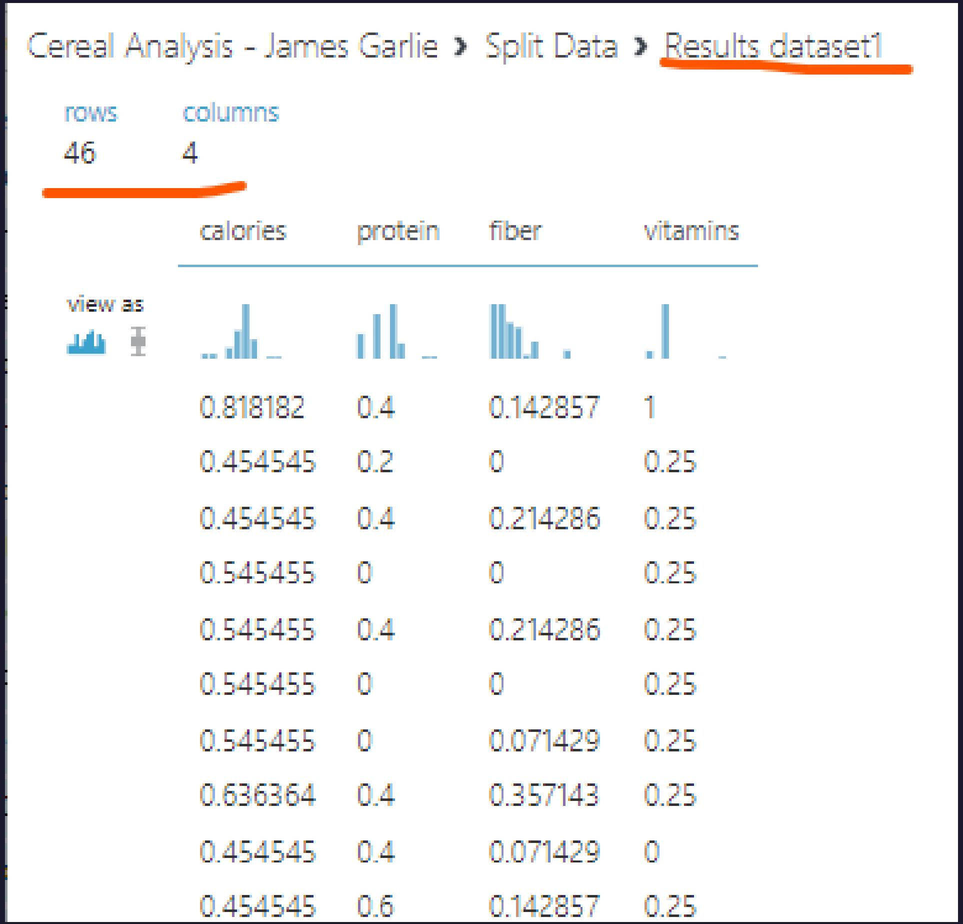
calories	protein	fiber	vitamins
-1.905397	1.337319	3.314439	-0.14627
0.677623	0.417912	-0.064172	-1.27255
-1.905397	1.337319	2.892113	-0.14627
-2.938605	1.337319	5.003745	-0.14627
0.161019	-0.501495	-0.486498	-0.14627
0.161019	-0.501495	-0.275335	-0.14627
0.161019	-0.501495	-0.486498	-0.14627
1.194228	0.417912	-0.064172	-0.14627

Splitting Data

This slide shows the addition of the Split Data module and the results of dataset1. Notice we now show 46 rows with 4 columns.



Cereal Analysis - James Garlie



Cereal Analysis - James Garlie › Split Data › Results dataset1

rows 46

columns 4

calories	protein	fiber	vitamins
0.818182	0.4	0.142857	1
0.454545	0.2	0	0.25
0.454545	0.4	0.214286	0.25
0.545455	0	0	0.25
0.545455	0.4	0.214286	0.25
0.545455	0	0	0.25
0.545455	0	0.071429	0.25
0.636364	0.4	0.357143	0.25
0.454545	0.4	0.071429	0
0.454545	0.6	0.142857	0.25

Iteration Process Threshold

This slide shows three different iterations of the code. The first is the original (10, 10), the second is (7.5, 7.5), and the third is (5, 5). When enlarged, I like the second or middle the best.

fig1 = plt.figure(1, figsize=(10, 10))

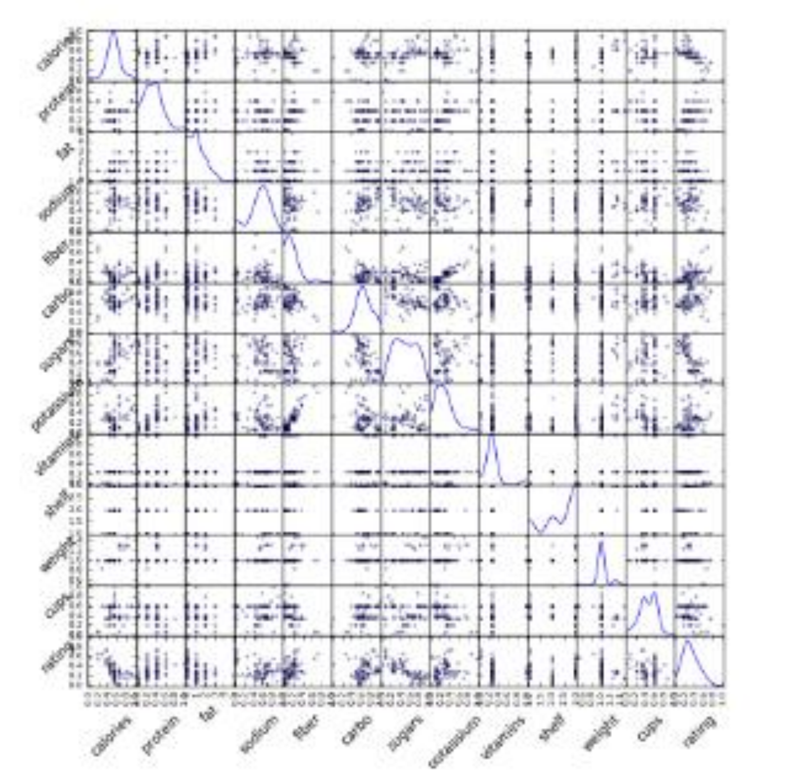


fig1 = plt.figure(1, figsize=(7.5, 7.5))

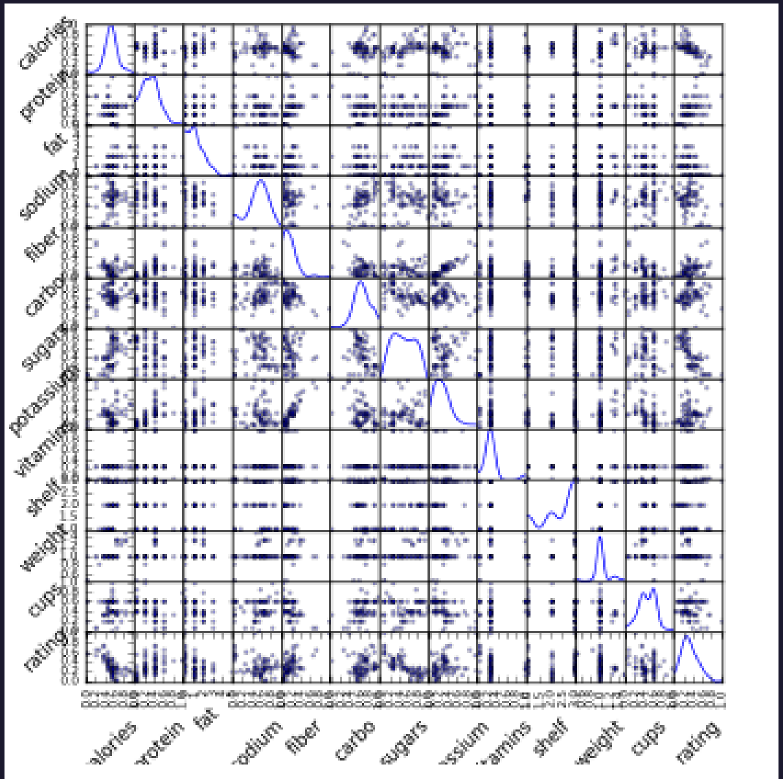
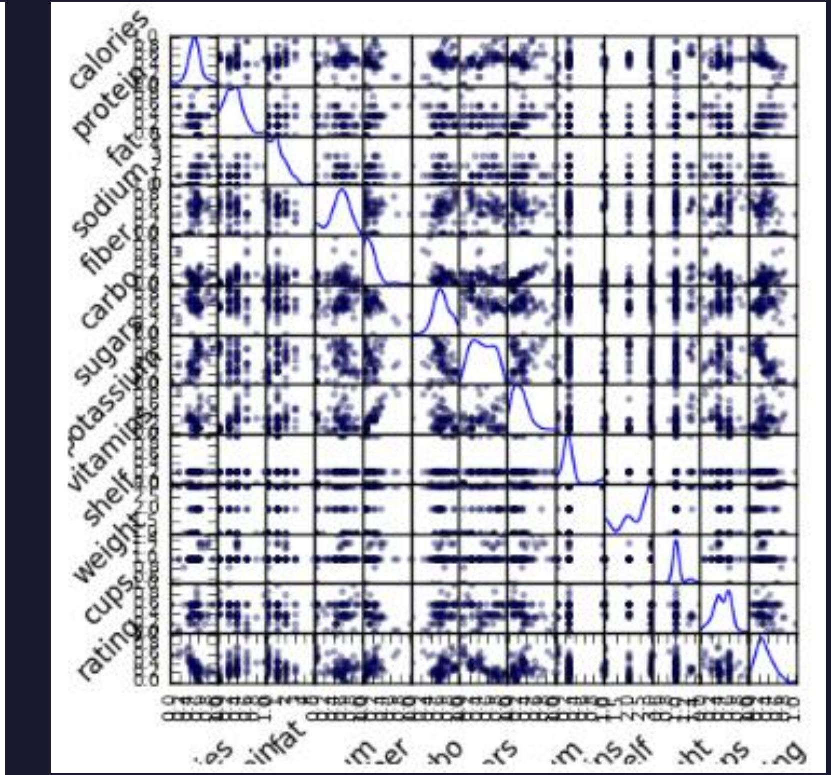
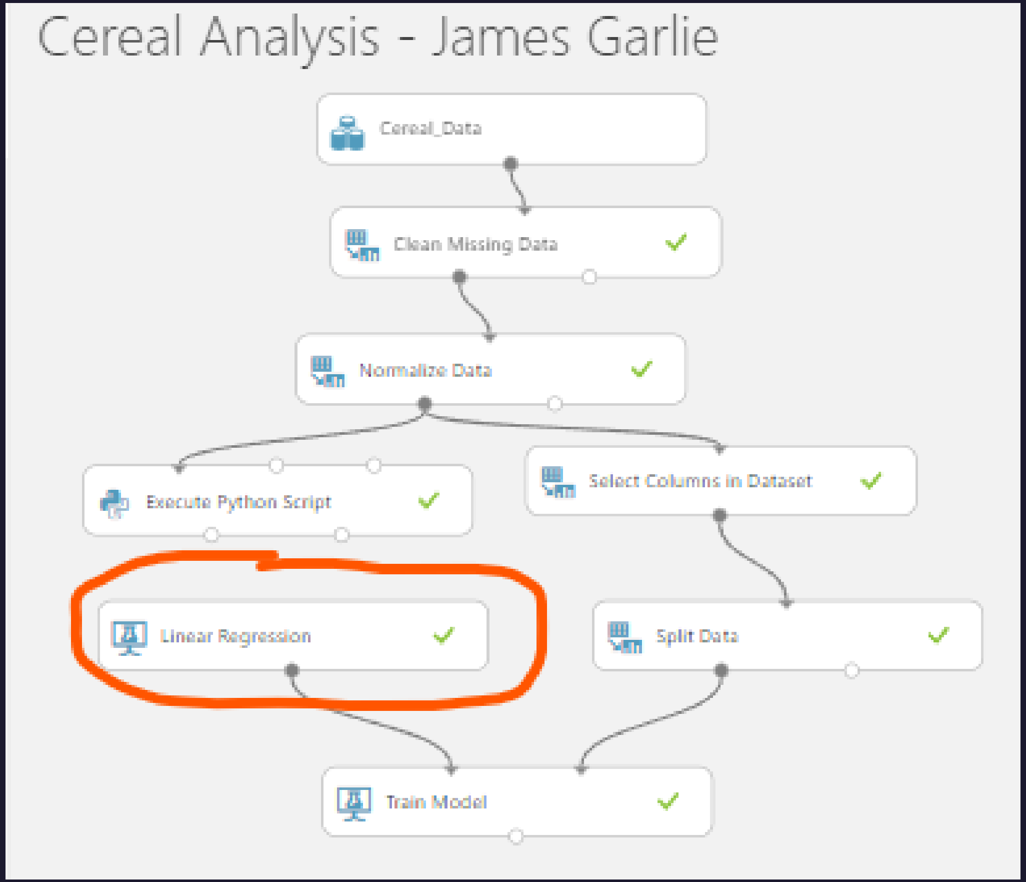


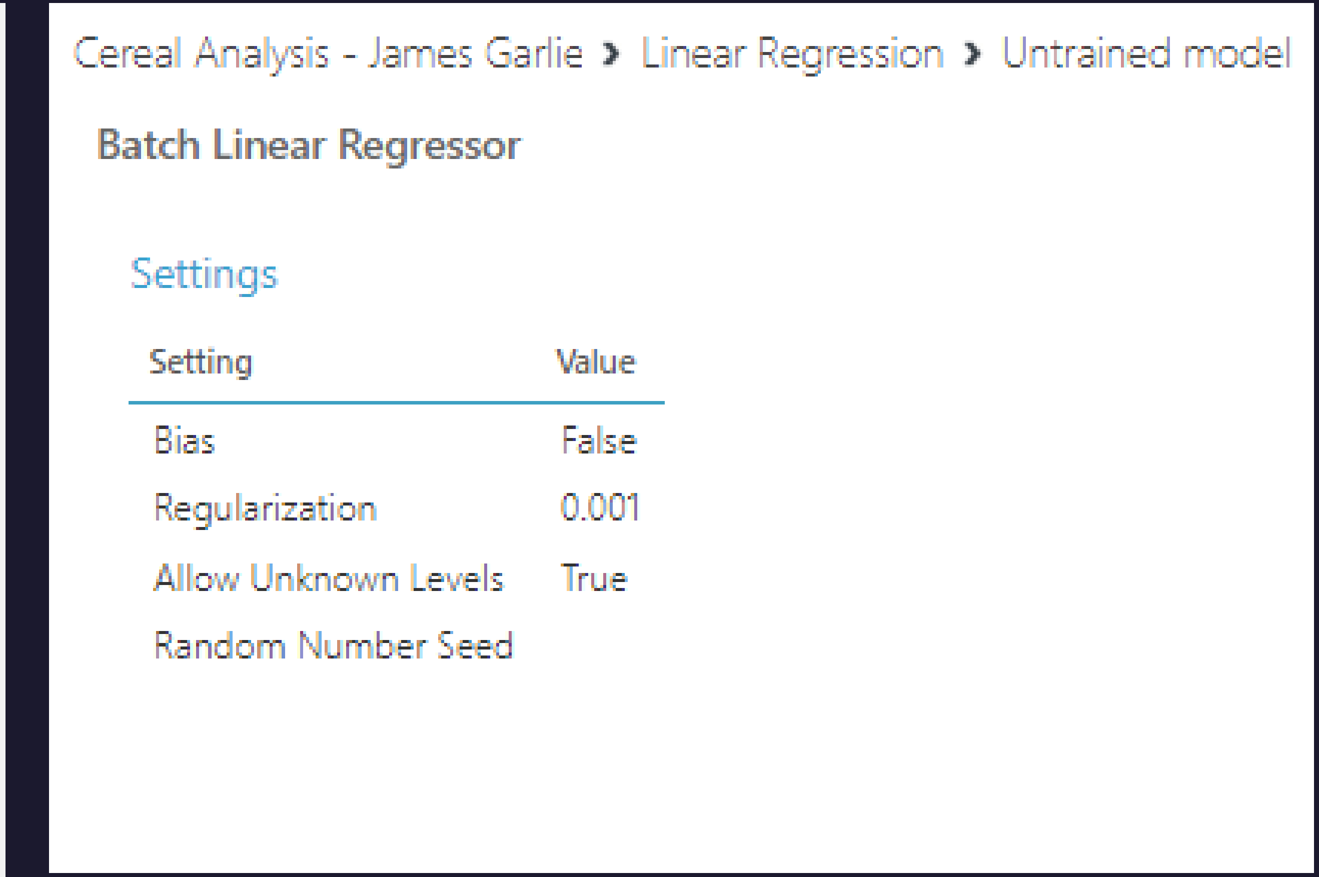
fig1 = plt.figure(1, figsize=(5, 5))



Linear Regression

This slide shows the addition of the Linear Regression module and the results of the Untrained model.




Cereal Analysis - James Garlie › Linear Regression › Untrained model

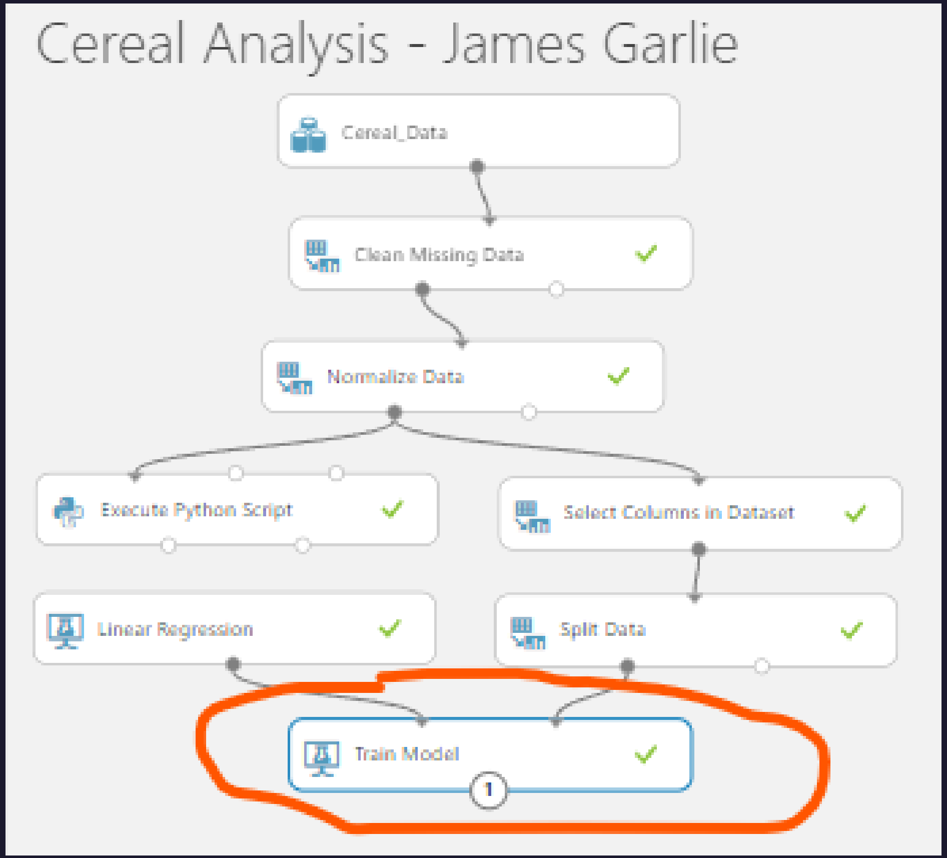
Batch Linear Regressor

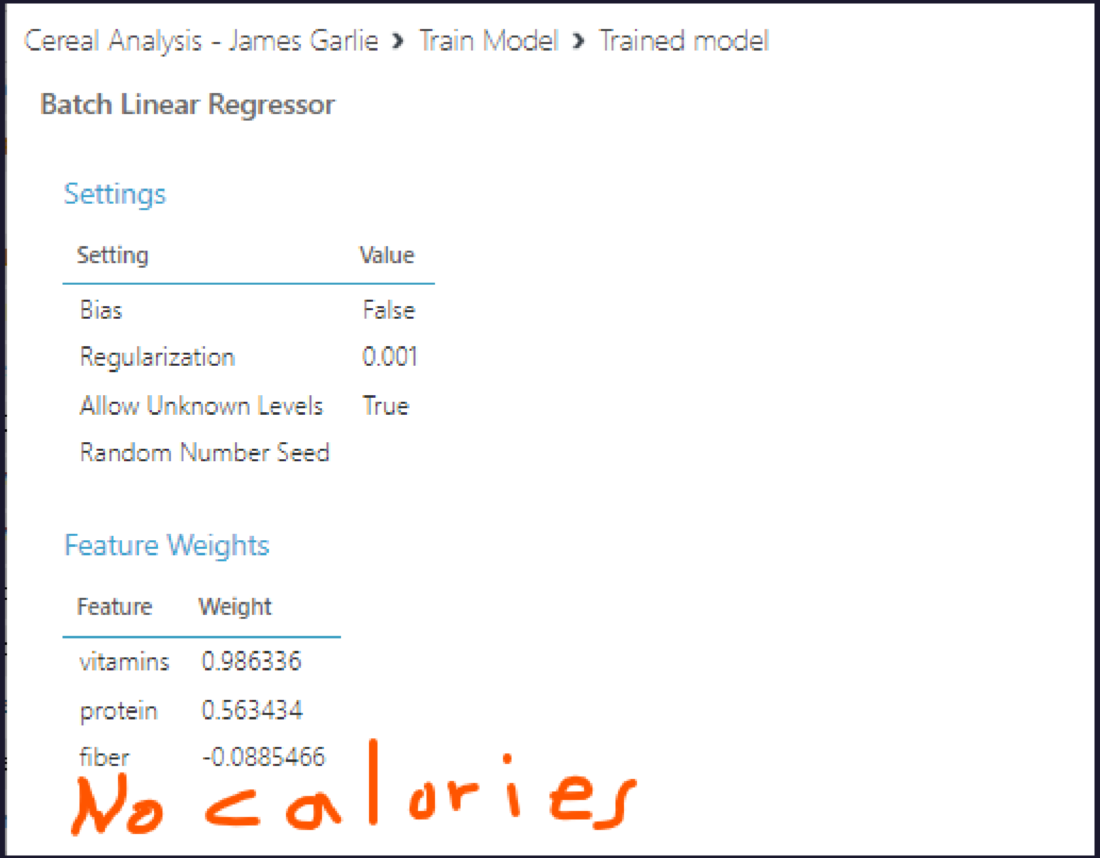
Settings

Setting	Value
Bias	False
Regularization	0.001
Allow Unknown Levels	True
Random Number Seed	

Training Model

This slide shows the addition of the Train Model module where I deleted the calories feature, and the results of the Trained Model showing no calories.




Cereal Analysis - James Garlie › Train Model › Trained model

Batch Linear Regressor

Settings

Setting	Value
Bias	False
Regularization	0.001
Allow Unknown Levels	True
Random Number Seed	

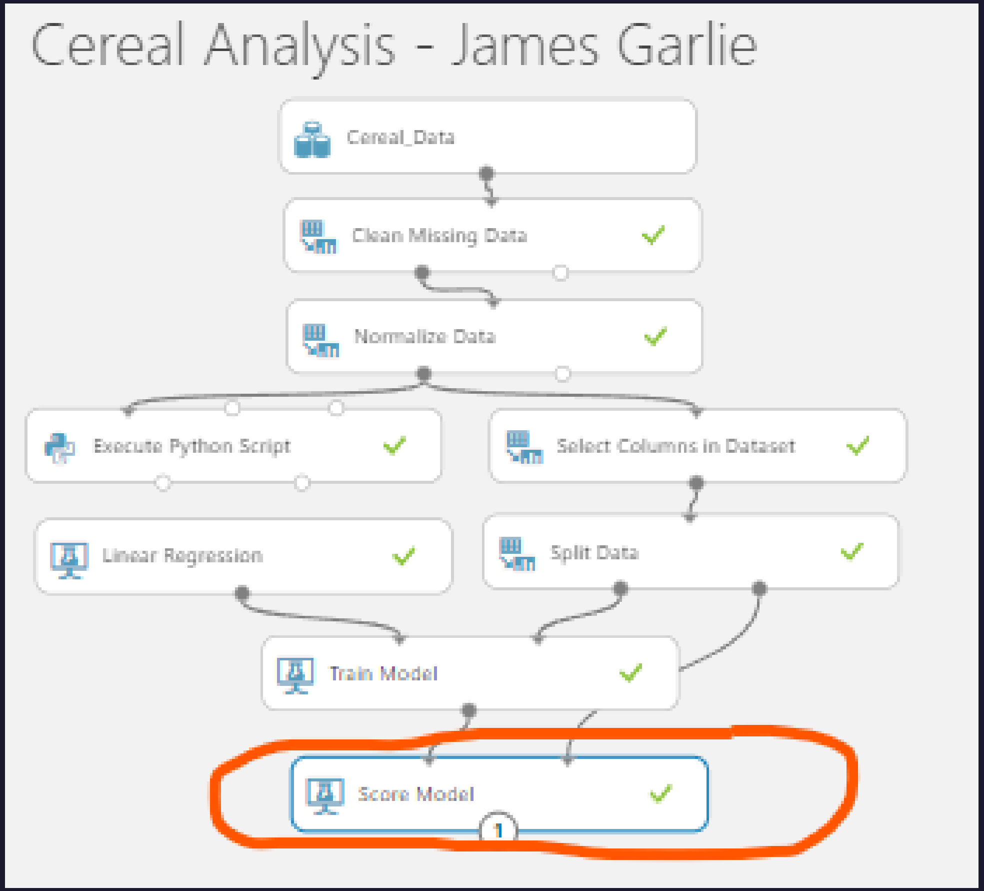
Feature Weights

Feature	Weight
vitamins	0.986336
protein	0.563434
fiber	-0.0885466

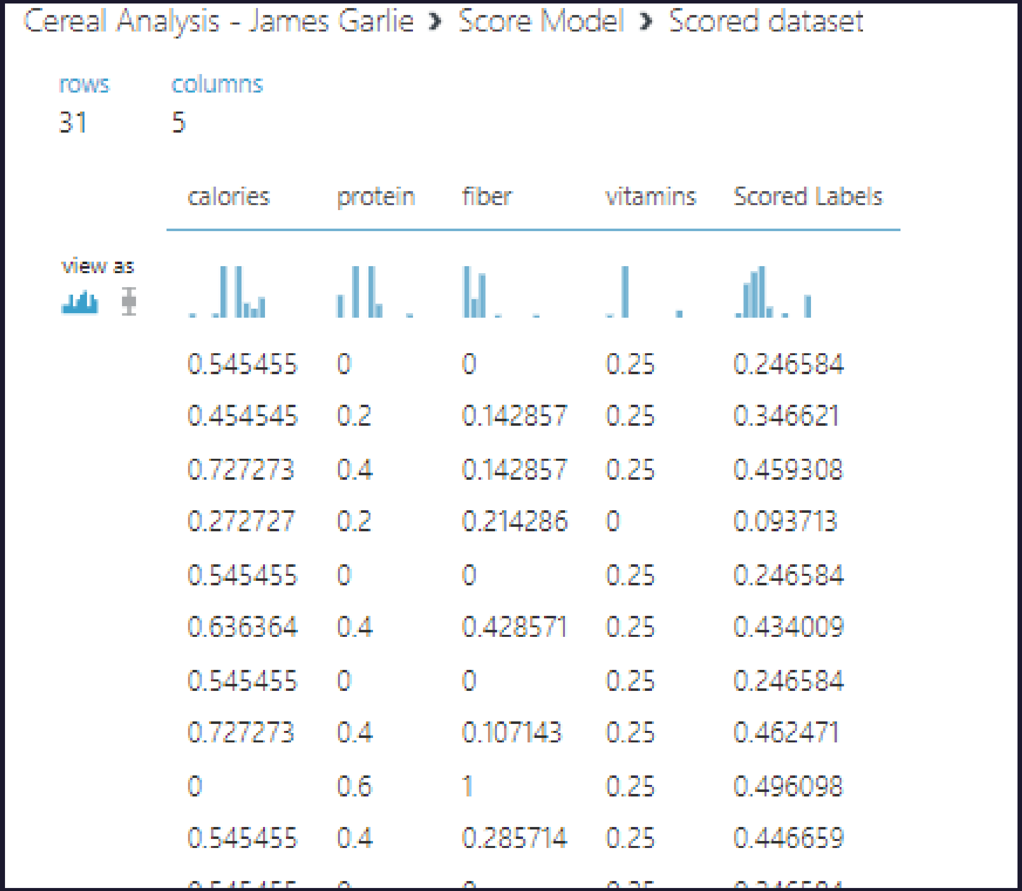
No calories

Scoring Model

Here I added the Score Model with results showing 31 rows and 5 columns. Notice calories has been included the new feature showing Scored Lables.



Cereal Analysis - James Garlie



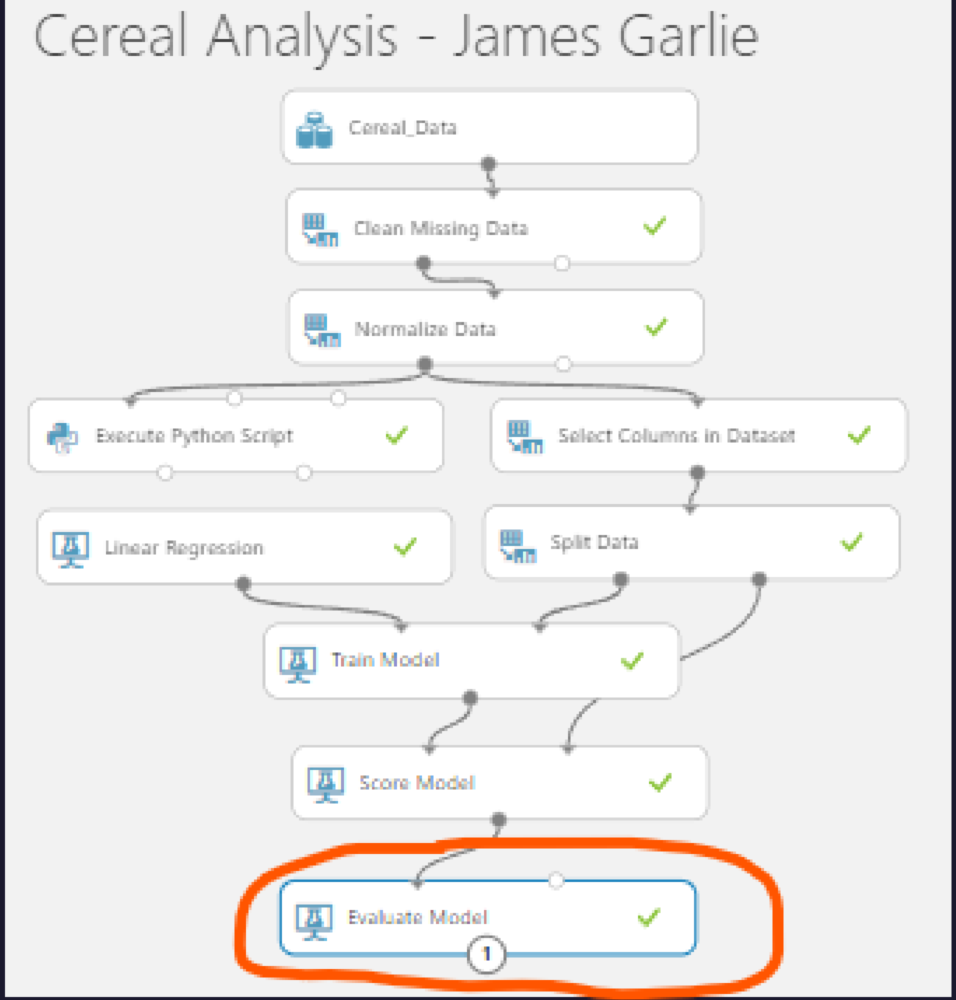
Cereal Analysis - James Garlie › Score Model › Scored dataset

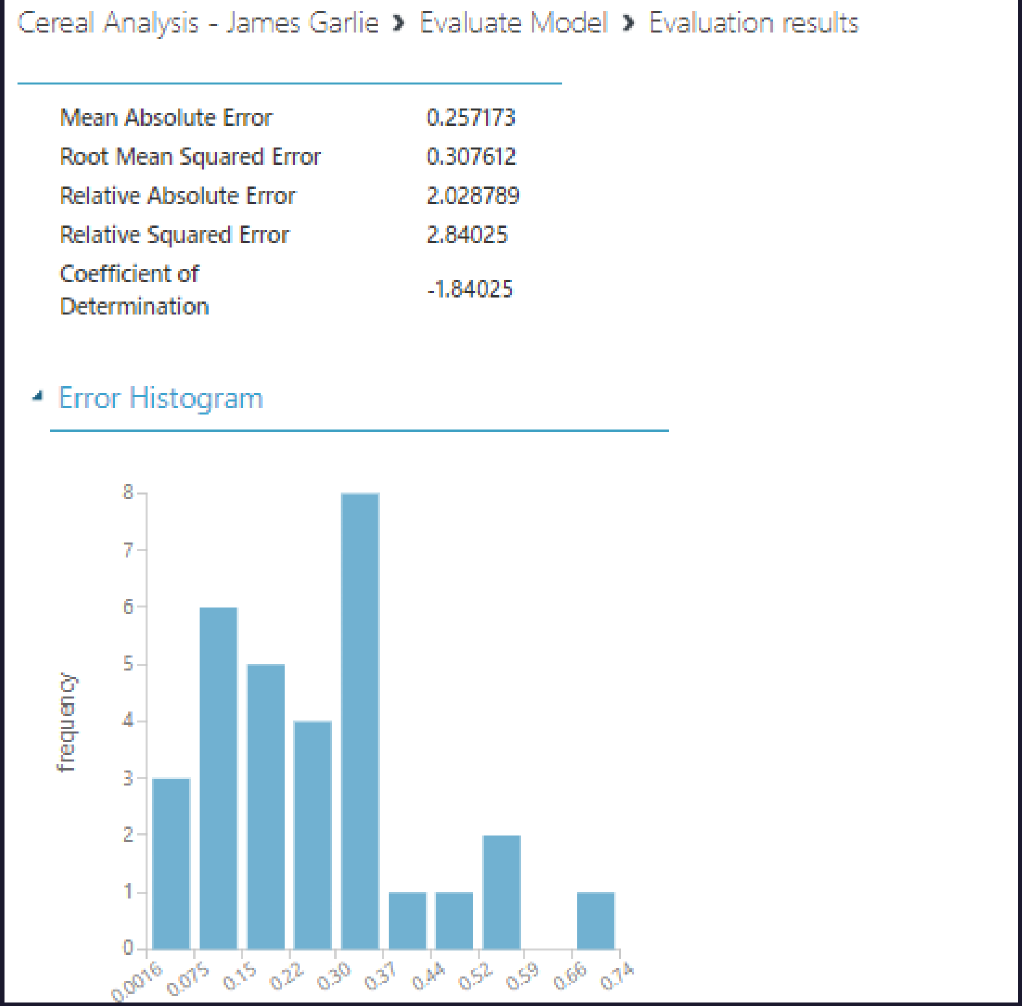
rows	columns
31	5

calories	protein	fiber	vitamins	Scored Labels
0.545455	0	0	0.25	0.246584
0.454545	0.2	0.142857	0.25	0.346621
0.727273	0.4	0.142857	0.25	0.459308
0.272727	0.2	0.214286	0	0.093713
0.545455	0	0	0.25	0.246584
0.636364	0.4	0.428571	0.25	0.434009
0.545455	0	0	0.25	0.246584
0.727273	0.4	0.107143	0.25	0.462471
0	0.6	1	0.25	0.496098
0.545455	0.4	0.285714	0.25	0.446659

Evaluating the Model

This slide shows the addition of the Evaluate Model module and the results with the Coefficient of Determination of -1.08025.




Cereal Analysis - James Garlie › Evaluate Model › Evaluation results

Mean Absolute Error	0.257173
Root Mean Squared Error	0.307612
Relative Absolute Error	2.028789
Relative Squared Error	2.84025
Coefficient of Determination	-1.84025

Error Histogram

frequency
0
1
2
3
4
5
6
7
8
0.0016
0.075
0.15
0.22
0.30
0.37
0.44
0.52
0.59
0.66
0.74

Conclusion

Learning the cleansing of the data is vital for Machine-learning.

I found the aspects of connecting modules, changing the data I was looking for, and visualizing the results to be very rewarding.

This project will be of tremendous benefit in the f

Career Skills

- Probability and statistics

Programming skills (Python or R)

Data skills (data processing, SQL data analysis, visualization skills)

Machine-learning algorithms

Lifelong learning

TensorFlow (neural networks)

Apache Spark

The analysis we have done using Azure Machine Learning can also be performed in some of these programming languages. For example, using some Python modules, you can receive the same Energy Efficiency Regression in Python.

Challenges

The biggest challenge I faced was finding the correct modules to choose from to assemble the data required.