Spring 2021 MIS 637 WS1 - Data Analytics & Machine Learning

Coffee Beans Variants Recommendation Using Clustering

Completed By: Mihir Kadam

mkadam3@stevens.edu

Prof. M. Daneshmand

Contents

Problem Statement

- Coffee connoisseurs are always very particular about their coffee, especially the bean that has been used.
- These connoisseurs also like to brew their own coffee every morning with the coffee bean that they have selected very particularly. They also spend time researching their coffee.

Clustering of different variants of Coffee beans

- By clustering different coffee bean variants together, we can use this information for targeted marketing. If a customer buys a variant, other similar variants can be recommended to them.
- In some cases, variants from completely different clusters can also be recommended if the customer is looking for a different taste.
- Since connoisseurs also like to mix-match their beans, two beans from completely different clusters can also be recommended for a blend.

Data Representation

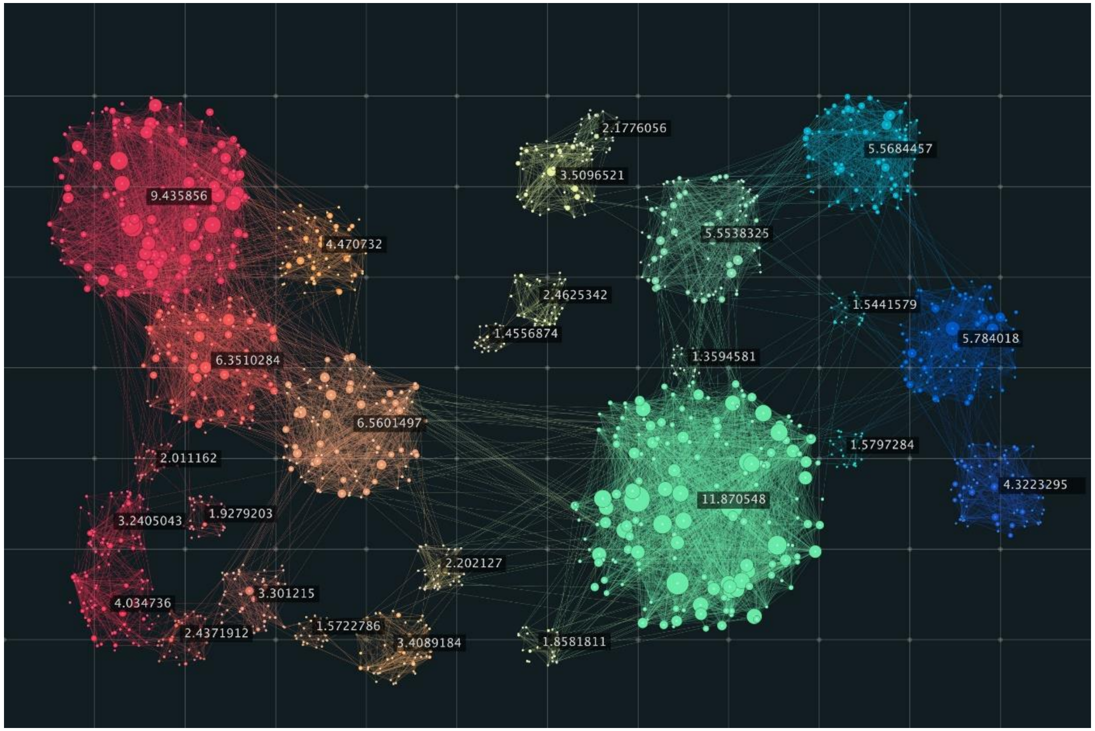
E1 | Farm.Name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		Species	Owner	Country.o	Farm.Nam	Lot.Numb	Mill	ICO.Numb	Company	Altitude	Region	Producer	Number.o	Bag.Weigh	In.Country	Harvest.Ye	Grading.Da	Owner.1	Variety	Processing.Method	Aroma	Flavor
2	0	Arabica	metad plc	Ethiopia	metad plc		metad plc	2014/2015	metad agricult	1950-2200	guji-hambe	METAD PL	300	60 kg	METAD Ag	2014	April 4th, 2	metad plc		Washed / Wet	8.67	8
3	1	Arabica	metad plc	Ethiopia	metad plc		metad plc	2014/2015	metad agricult	1950-2200	guji-hambe	METAD PL	300	60 kg	METAD Ag	2014	April 4th, 2	metad plc	Other	Washed / Wet	8.75	8
4	2	Arabica	grounds fc	Guatemala	san marcos barrancas		"san cristobal cuch			1600 - 1800 m			5	1	Specialty Coffee Asso		May 31st,	Grounds fc	Bourbon		8.42	
5	3	Arabica	yidnekache	Ethiopia	yidnekachew dabessa		wolensu		yidnekachew c	1800-2200	oromia	Yidnekach	320	60 kg	METAD Ag	2014	March 26t	Yidnekachew Dabessa		Natural / Dry	8.17	8
6	4	Arabica	metad plc	Ethiopia	metad plc		metad plc	2014/2015	metad agricult	1950-2200	guji-hambe	METAD PL	300	60 kg	METAD Ag	2014	April 4th, 2	metad plc	Other	Washed / Wet	8.25	
7	5	Arabica	ji-ae ahn	Brazil									100	30 kg	Specialty C	2013	Septembeı	Ji-Ae Ahn		Natural / Dry	8.58	8
8	6	Arabica	hugo valdi	Peru			hvc		richmond investment-coffee departı			HVC	100	69 kg	Specialty C	2012	Septembeı	Hugo Valdi	Other	Washed / Wet	8.42	
9	7	Arabica	ethiopia cc	Ethiopia	aolme		c.p.w.e	010/0338		1570-1700	oromia	Bazen Agri	300	60 kg	Ethiopia Cı	Mar-10	Septembeı	Ethiopia Commodity Exchange			8.25	8
10	8	Arabica	ethiopia cc	Ethiopia	aolme		c.p.w.e	010/0338		1570-1700	oromiya	Bazen Agri	300	60 kg	Ethiopia Cı	Mar-10	Septembeı	Ethiopia Commodity Exchange			8.67	8
11	9	Arabica	diamond e	Ethiopia	tulla coffee farm		tulla coffe	2014/15	diamond enteı	1795-1850	snnp/kaffa	Diamond E	50	60 kg	METAD Ag	2014	March 30t	Diamond E	Other	Natural / Dry	8.08	8
12	10	Arabica	mohamme	Ethiopia	fahem coffee plantation				fahem coffee	1855-1955	oromia	Fahem Cof	300	60 kg	METAD Ag	2014	March 27t	Mohammed Lalo		Natural / Dry	8.17	8
13	11	Arabica	cqi q coffe	United Sta	el filo			unknown	coffee quality	meters abc	antioquia	Alfredo De	10	1 kg	AlmacafÃ©	2014	March 13t	CQI Q Cofi	Other	Washed / Wet	8.25	8
14	12	Arabica	cqi q coffe	United Sta	los cedros			unknown	coffee quality	meters abc	antioquia	Jorge Walt	10	1 kg	AlmacafÃ©	2014	March 13t	CQI Q Cofi	Other	Washed / Wet	8.08	8
15	13	Arabica	grounds fc	United Sta	arianna farms					2000 ft	kona	Robert, Sh	1	1	Specialty C	Sept 2009	May 31st,	Grounds for Health Admin			8.33	8
16	14	Arabica	ethiopia cc	Ethiopia	aolme		c.p.w.e	010/0338		1570-1700	oromiya	Bazen Agri	300	60 kg	Ethiopia Cı	Mar-10	August 31s	Ethiopia Commodity Exchange			8.25	8
17	15	Arabica	cqi q coffe	United Sta	el Ã¡guila			unknown	coffee quality	meters abc	antioquia	MarÃ-a Le	10	1 kg	AlmacafÃ©	2014	March 13t	CQI Q Cofi	Other	Washed / Wet	8	
18	16	Arabica	grounds fc	Indonesia	toarco jaya					1200-1800	sulawesi	P.T. Toarc	1	2 kg,lbs	Specialty C	May-Augu	May 31st,	Grounds for Health Admin			8.33	8
19	17	Arabica	ethiopia cc	Ethiopia				010/0056			yirgacheffe	Green Gol	150	6	Ethiopia Cı	2009/2010	June 16th,	Ethiopia Commodity Exchange			8.17	8
20	18	Arabica	yunnan co	China	echo coffe	YNC-0611	echo coffee mill		yunnan coffee	1450	yunnan	Echo Coffe	3	60 kg	Yunnan Co	2015	April 7th, 2	Yunnan Co	Catimor	Washed / Wet	8.42	8
21	19	Arabica	essenceco	Ethiopia	drima zede		drima zede	1E+08	essence coffee	1700-2000	gedio	LevelUp	250	60 kg	Blossom V	2014	March 25t	EssenceCo	Ethiopian Yirga	Natural / Dry	8.17	8
22	20	Arabica	cqi q coffe	United Sta	el rodeo			unknown	coffee quality	meters abc	antioquia	NicolÃ¡s R	10	1 kg	AlmacafÃ©	2014	March 13t	CQI Q Cofi	Other	Washed / Wet	8	8
23	21	Arabica	the coffee	Costa Rica	several		cafe altura	5-562-001	the coffee sou	1300 msnr	san ramon	SEVERAL	250	3 lbs	Specialty C	2014	April 2nd, 2	The Coffee	Caturra	Washed / Wet	8.08	8
24	22	Arabica	roberto lic	Mexico	la herradura		la herradu	0		1320	xalapa	ROBERTO	14	1 kg	AMECAFE	2012	July 26th, 2	ROBERTO	Other	Washed / Wet	8.17	8
25	23	Arabica	cqi q coffe	United Sta	la curva			unknown	coffee quality	meters abc	antioquia	Silvia Elena	10	1 kg	AlmacafÃ©	2014	March 13t	CQI Q Cofi	Other	Washed / Wet	8.25	8
26	24	Arabica	ji-ae ahn	Ethiopia							sidamo		100	60 kg	Specialty C	2013	Septembeı	Ji-Ae Ahn		Natural / Dry	8.42	8
27	25	Arabica	nucoffee	Brazil	fazenda kaquend			002/1251/	nucoffee	1250m	south of m	Ralph Junq	3	60 kg	NUCOFFEE	2011	December	NUCOFFEE	Bourbon	Natural / Dry	8.5	
28	26	Arabica	ethiopia cc	Ethiopia				010/0056/Sidamo			sidamo	Green Gol	150	6	Ethiopia Cı	2009/2010	June 16th,	Ethiopia Commodity Exchange			7.83	8
29	27	Arabica	kabum tra	Uganda	chebonet (23) womer		kabum tra	0	kabum trading	1950	kapchorwa	Kabum tra	100	60 kg	Uganda Co	2013	June 26th,	Kabum Tra	SL14	Washed / Wet	8.42	8

merged_data_cleaned

Data Representation (Cont'd)

- The extracted data contains 44 columns, all containing different information about the coffee bean such as owner, country, company, farm name, etc.
- Some of the data is numeric while some isn't as visible in the previous slide. Therefore, data standardization will be performed. Also, normalization will be performed.
- Out of this, a few relevant attributes will be used for the clustering job.



Data Understanding

- The data sources contains 44 attributes for 732 different coffee beans.
- Out of these 44 attributes, we select the following relevant attributes:
  - Species, Variety, Processing Method, Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Cupper Points, Moisture, Color, altitude mean meters.

Species	Variety	Processing	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Po	Moisture	Color	altitude_mean_meters
Arabica		Washed / \	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	0.12	Green	2075
Arabica	Other	Washed / \	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	0.12	Green	2075
Arabica	Bourbon		8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	0		1700
Arabica		Natural / D	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	0.11	Green	2000
Arabica	Other	Washed / \	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	0.12	Green	2075
Arabica		Natural / D	8.58	8.42	8.42	8.5	8.25	8.33	10	10	10	8.33	0.11	Bluish-Green	
Arabica	Other	Washed / \	8.42	8.5	8.33	8.5	8.25	8.25	10	10	10	8.5	0.11	Bluish-Green	
Arabica			8.25	8.33	8.5	8.42	8.33	8.5	10	10	9.33	9	0.03		1635
Arabica			8.67	8.67	8.58	8.42	8.33	8.42	9.33	10	9.33	8.67	0.03		1635
Arabica	Other	Natural / D	8.08	8.58	8.5	8.5	7.67	8.42	10	10	10	8.5	0.1	Green	1822.5
Arabica		Natural / D	8.17	8.67	8.25	8.5	7.75	8.17	10	10	10	8.58	0.1		1905
Arabica	Other	Washed / \	8.25	8.42	8.17	8.33	8.08	8.17	10	10	10	8.5	0		1872
Arabica	Other	Washed / \	8.08	8.67	8.33	8.42	8	8.08	10	10	10	8.33	0		1943
Arabica			8.33	8.42	8.08	8.25	8.25	8	10	10	10	8.58	0		609.6
Arabica			8.25	8.33	8.5	8.25	8.58	8.75	9.33	10	9.33	8.5	0.05		1635
Arabica	Other	Washed / \	8	8.5	8.58	8.17	8.17	8	10	10	10	8.17	0		2080
Arabica			8.33	8.25	7.83	7.75	8.5	8.42	10	10	10	8.33	0.03		1500
Arabica			8.17	8.33	8.25	8.33	8.42	8.33	9.33	10	9.33	8.83	0.05		
Arabica	Catimor	Washed / \	8.42	8.25	8.08	8.17	7.92	8	10	10	10	8.42	0.1	Green	1450
Arabica	Ethiopian \	Natural / D	8.17	8.17	8	8.17	8.08	8.33	10	10	10	8.33	0		1850
Arabica	Other	Washed / \	8	8.25	8.08	8.5	8.25	8	10	10	10	8.17	0	None	2019
Arabica	Caturra	Washed / \	8.08	8.25	8	8.17	8	8.33	10	10	10	8.33	0.11	Green	1300
Arabica	Other	Washed / \	8.17	8.25	8.17	8	7.83	8.17	10	10	10	8.58	0.13	Green	1320
Arabica	Other	Washed / \	8.25	8.33	8.17	8.17	7.83	8.17	10	10	10	8.17	0		2112
Arabica		Natural / D	8.42	8.17	7.92	8.17	8.33	8	10	10	10	8.08	0.11	Bluish-Green	
Arabica	Bourbon	Natural / D	8.5	8.5	8	8	8	8	10	10	10	7.92	0.12	Green	1250
Arabica			7.83	8.25	8.08	8.17	8.17	8.17	10	10	10	8.25	0.05		
Arabica	SL14	Washed / \	8.42	8.17	8.17	8.17	7.83	7.92	10	10	10	8.17	0.12	Green	1950

coffee data with relevant attri

Data Preparation

- We prepare the data by first converting the non-numeric data into numeric. We do this by numbering the unique non-numeric data from 0 to x, where x is the number of unique records in each column

Species	Variety	Processing	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Po	Moisture	Color	altitude_mean_meters
Arabica		Washed / '	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	0.12	Green	2075
Arabica	Other	Washed / '	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	0.12	Green	2075
Arabica	Bourbon		8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	0		1700
Arabica		Natural / [	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	0.11	Green	2000
Arabica	Other	Washed / '	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	0.12	Green	2075
Arabica		Natural / [	8.58	8.42	8.42	8.5	8.25	8.33	10	10	10	8.33	0.11	Bluish-Green	
Arabica	Other	Washed / '	8.42	8.5	8.33	8.5	8.25	8.25	10	10	10	8.5	0.11	Bluish-Green	
Arabica			8.25	8.33	8.5	8.42	8.33	8.5	10	10	9.33	9	0.03		1635
Arabica			8.67	8.67	8.58	8.42	8.33	8.42	9.33	10	9.33	8.67	0.03		1635
Arabica	Other	Natural / [	8.08	8.58	8.5	8.5	7.67	8.42	10	10	10	8.5	0.1	Green	1822.5
Arabica		Natural / [	8.17	8.67	8.25	8.5	7.75	8.17	10	10	10	8.58	0.1		1905
Arabica	Other	Washed / '	8.25	8.42	8.17	8.33	8.08	8.17	10	10	10	8.5	0		1872
Arabica	Other	Washed / '	8.08	8.67	8.33	8.42	8	8.08	10	10	10	8.33	0		1943
Arabica			8.33	8.42	8.08	8.25	8.25	8	10	10	10	8.58	0		609.6
Arabica			8.25	8.33	8.5	8.25	8.58	8.75	9.33	10	9.33	8.5	0.05		1635
Arabica	Other	Washed / '	8	8.5	8.58	8.17	8.17	8	10	10	10	8.17	0		2080
Arabica			8.33	8.25	7.83	7.75	8.5	8.42	10	10	10	8.33	0.03		1500
Arabica			8.17	8.33	8.25	8.33	8.42	8.33	9.33	10	9.33	8.83	0.05		
Arabica	Catimor	Washed / '	8.42	8.25	8.08	8.17	7.92	8	10	10	10	8.42	0.1	Green	1450
Arabica	Ethiopian '	Natural / [	8.17	8.17	8	8.17	8.08	8.33	10	10	10	8.33	0		1850
Arabica	Other	Washed / '	8	8.25	8.08	8.5	8.25	8	10	10	10	8.17	0	None	2019
Arabica	Caturra	Washed / '	8.08	8.25	8	8.17	8	8.33	10	10	10	8.33	0.11	Green	1300
Arabica	Other	Washed / '	8.17	8.25	8.17	8	7.83	8.17	10	10	10	8.58	0.13	Green	1320
Arabica	Other	Washed / '	8.25	8.33	8.17	8.17	7.83	8.17	10	10	10	8.17	0		2112
Arabica		Natural / [	8.42	8.17	7.92	8.17	8.33	8	10	10	10	8.08	0.11	Bluish-Green	
Arabica	Bourbon	Natural / [	8.5	8.5	8	8	8	8	10	10	10	7.92	0.12	Green	1250
Arabica			7.83	8.25	8.08	8.17	8.17	8.17	10	10	10	8.25	0.05		
Arabica	SL14	Washed / '	8.42	8.17	8.17	8.17	7.83	7.92	10	10	10	8.17	0.12	Green	1950

coffee data with relevant attri

Relevant attributes with numeric and non-numeric data

Data Preparation (Cont'd)

After successful transformation, the data looks like this:

Species	Variety	Processing	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Po	Moisture	Color	altitude_mean_meters
0	3	4	7.75	7.75	7.5	7.83	7.58	7.58	10	10	10	7.58	0.11	3	1128.37
0	3	4	7.67	7.58	7.58	7.58	7.67	7.75	10	10	10	7.75	0.1	3	1100
0	3	4	7.67	7.75	7.58	7.67	7.75	7.58	10	10	10	7.58	0.12	3	1775
0	3	4	7.75	7.75	7.58	7.58	7.5	7.67	10	10	10	7.75	0.12	3	1200
0	3	4	7.67	7.67	7.42	7.75	7.75	7.67	10	10	10	7.67	0.12	3	1800
0	4	1	7.67	7.58	7.67	7.67	7.67	7.67	10	10	10	7.67	0	3	1600
0	4	4	7.67	7.58	7.67	7.58	7.67	7.75	10	10	10	7.67	0.12	3	1550
0	4	0	7.75	7.75	7.75	7.75	7.5	7.58	10	10	10	7.5	0.11	3	1000
0	4	4	7.67	7.67	7.58	7.67	7.75	7.58	10	10	10	7.67	0.11	3	1775
0	4	1	7.67	7.58	7.58	7.58	7.67	7.75	10	10	10	7.75	0.11	3	2136
0	4	3	7.83	7.75	7.5	7.75	7.5	7.58	10	10	10	7.67	0.12	1	1100
0	4	4	7.75	7.58	7.58	7.75	7.75	7.5	10	10	10	7.58	0	2	1799
0	4	4	7.92	7.42	7.5	7.67	7.5	7.67	10	10	10	7.83	0	1	1850
0	4	1	7.75	7.58	7.58	7.5	7.75	7.75	10	10	10	7.58	0	3	1000
0	4	4	7.67	7.75	7.5	7.83	7.58	7.58	10	10	10	7.58	0.1	2	1219.2
0	4	4	7.58	7.58	7.58	7.58	7.75	7.67	10	10	10	7.75	0.12	3	1775
0	4	4	7.75	7.67	7.58	7.67	7.67	7.58	10	10	10	7.58	0.13	3	1400
0	4	3	7.25	7.67	7.58	7.92	7.83	7.58	10	10	10	7.67	0.13	2	1500
0	4	4	8	7.58	7.25	7.58	7.92	7.58	10	10	10	7.58	0.11	3	1850
0	4	4	7.58	7.75	7.5	7.83	7.42	7.67	10	10	10	7.75	0.1	3	1400
0	5	1	7.67	7.5	7.67	7.5	7.67	7.67	10	10	10	7.83	0.1	3	1350
0	5	4	7.58	7.75	7.42	7.67	7.75	7.67	10	10	10	7.67	0	3	1750
0	5	4	7.58	7.67	7.58	7.75	7.58	7.75	10	10	10	7.58	0.11	3	1750
0	5	1	7.83	7.58	7.25	7.5	7.5	7.42	10	10	10	8.42	0.13	3	1170
0	5	3	7.67	7.75	7.67	7.75	7.58	7.5	10	10	10	7.58	0.11	3	940
0	5	4	7.75	7.67	7.67	7.75	7.33	7.5	10	10	10	7.83	0.12	3	1400
0	5	4	7.83	7.5	7.5	8	7.67	7.5	10	10	10	7.5	0.1	3	1524
0	5	4	7.67	7.58	7.58	7.67	7.67	7.67	10	10	10	7.67	0.11	3	1775

coffee data with all data in nu

Relevant attributes with only numeric data

Data Normalization

- For the algorithm to work effectively and get accurate results, we normalize the data.
- Since the values of attributes range from a maximum of 2 in one column, to a maximum of 190164 in another, we normalize the data to bring the range down to 0 to 1.
- We normalize using the following formula:

$$X_{normalized} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad [1]$$

  - Here, X is the value in the cell which we are normalizing.
  - Xmin and Xmax are minimum and maximum values in the selected column, respectively.
- The data looks like following after normalization:

Data Normalization (Cont’d)

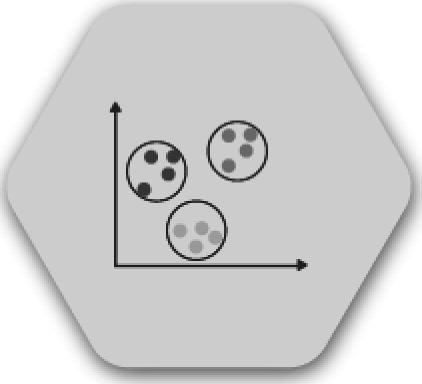
Species_n

Species_n	Variety_n	Processing	Aroma_n	Flavor_n	Aftertaste	Acidity_n	Body_n	Balance_n	Uniformity	Clean.Cup	Sweetness	Cupper.Po	Moisture_	Color_n	Altitude_n
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.609844	0.763614	0.77904	0.630252	0.615758	0.611888	0.6	0	0.133	0.602564	0	0	0
0	0	0	0.759904	0.783416	0.77904	0.729892	0.626667	0.708625	0.6	0.133	0.6	0.611888	0	0	0
0	0	0	0.780312	0.783416	0.77904	0.7503	0.636364	0.719114	0.667	0.267	0.6	0.631702	0	0	0
0	0	0	0.780312	0.794554	0.77904	0.780312	0.767273	0.719114	0.667	0.267	0.667	0.699301	0	0	0
0	0.038462	0	0.80072	0.804455	0.789141	0.80072	0.778182	0.719114	0.667	0.533	0.742	0.719114	0	0	0
0	0.038462	0	0.810324	0.804455	0.799242	0.80072	0.787879	0.737762	0.733	0.533	0.742	0.719114	0	0	0
0	0.038462	0	0.810324	0.804455	0.799242	0.80072	0.808485	0.757576	0.8	0.6	0.742	0.728438	0	0	0
0	0.038462	0	0.819928	0.804455	0.799242	0.810324	0.808485	0.757576	0.8	0.6	0.75	0.737762	0	0	0
0	0.038462	0	0.819928	0.814356	0.799242	0.810324	0.818182	0.7669	0.8	0.6	0.758	0.737762	0	0	0
0	0.038462	0	0.819928	0.814356	0.810606	0.819928	0.818182	0.7669	0.8	0.6	0.758	0.748252	0	0	
0	0.038462	0	0.819928	0.814356	0.810606	0.819928	0.818182	0.777389	0.8	0.6	0.758	0.748252	0	0	
0	0.038462	0	0.819928	0.814356	0.820707	0.819928	0.827879	0.777389	0.8	0.667	0.758	0.748252	0	0	
0	0.038462	0	0.819928	0.825495	0.820707	0.819928	0.827879	0.777389	0.8	0.667	0.758	0.757576	0	0	0.009099
0	0.038462	0	0.830732	0.825495	0.820707	0.819928	0.827879	0.786713	0.8	0.667	0.767	0.757576	0	0	0.011433
0	0.038462	0	0.830732	0.825495	0.820707	0.819928	0.827879	0.786713	0.8	0.667	0.767	0.757576	0	0	0.023098
0	0.038462	0	0.830732	0.835396	0.830808	0.830732	0.838788	0.786713	0.8	0.667	0.775	0.757576	0	0	0.025432
0	0.038462	0.25	0.830732	0.835396	0.830808	0.830732	0.838788	0.786713	0.8	0.667	0.775	0.7669	0	0	0.028931
0	0.038462	0.25	0.830732	0.835396	0.830808	0.830732	0.838788	0.786713	0.8	0.733	0.775	0.7669	0	0	0.034764
0	0.038462	0.25	0.830732	0.835396	0.830808	0.830732	0.838788	0.796037	0.8	0.733	0.775	0.7669	0	0	0.034764
0	0.076923	0.25	0.830732	0.835396	0.830808	0.830732	0.838788	0.796037	0.8	0.733	0.775	0.7669	0	0	0.036604
0	0.076923	0.25	0.830732	0.835396	0.842172	0.830732	0.838788	0.796037	0.8	0.8	0.775	0.7669	0	0	0.036604
0	0.076923	0.25	0.830732	0.845297	0.842172	0.830732	0.848485	0.796037	0.867	0.8	0.775	0.777389	0	0	0.036604
0	0.076923	0.25	0.840336	0.845297	0.842172	0.830732	0.848485	0.796037	0.867	0.8	0.783	0.777389	0	0	0.038264
0	0.076923	0.25	0.840336	0.845297	0.842172	0.840336	0.848485	0.796037	0.867	0.8	0.783	0.777389	0	0	0.038964
0	0.076923	0.25	0.840336	0.845297	0.842172	0.840336	0.848485	0.796037	0.867	0.8	0.792	0.777389	0	0	0.040597
0	0.076923	0.25	0.840336	0.845297	0.842172	0.840336	0.848485	0.796037	0.867	0.8	0.792	0.777389	0	0	0.041764
0	0.076923	0.25	0.840336	0.845297	0.842172	0.840336	0.848485	0.796037	0.867	0.8	0.8	0.777389	0	0	0.04293

coffee data with all data norma

Data Normalization (Cont'd)

- As it is visible, all data is now in numeric and ranges from 0 to 1.
- Missing values are dealt with by weka (the software package to be used).
- The only outliers present were in the altitude column. The values were 190164 meters, 110000 meters, and 11000 meters (before normalization). We delete these values since the altitude of the highest peak in the world (Mt. Everest) is 8848.9 meters.
- At this point, our data is ready for the clustering job.

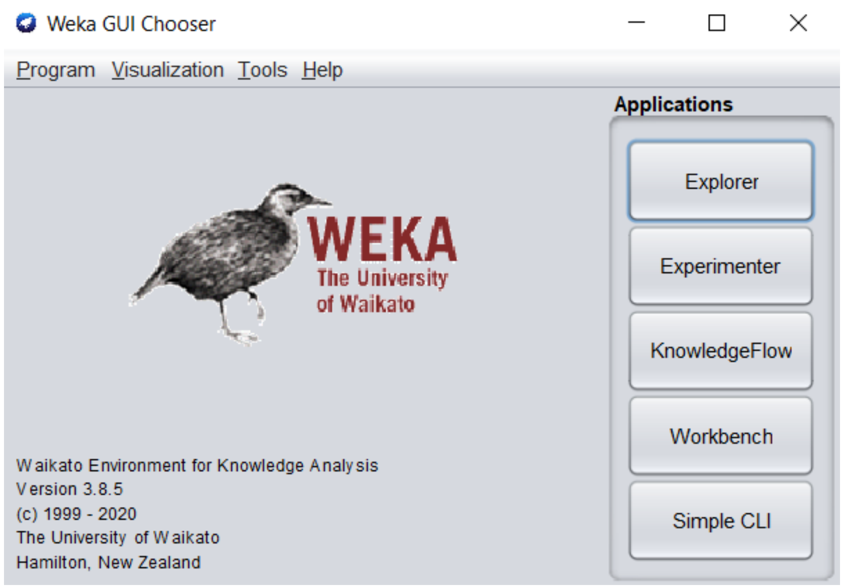


Algorithm

- We will be using K-means for the clustering job.
- K-means clustering is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. [2]
- Every data point is allocated to a cluster depending upon the distance of that data point to the cluster centers.
- This distance is calculated using Euclidean distance.
- Each data point is allocated to a cluster where the distance from the cluster center and the data point is smallest.
- The algorithm runs through several iterations until the items in the clusters don’t move upon successive iterations.
- After successful execution, we have n number of clusters, where n is a predefined number and all the items in the cluster have similar properties.

Software Package (Weka)

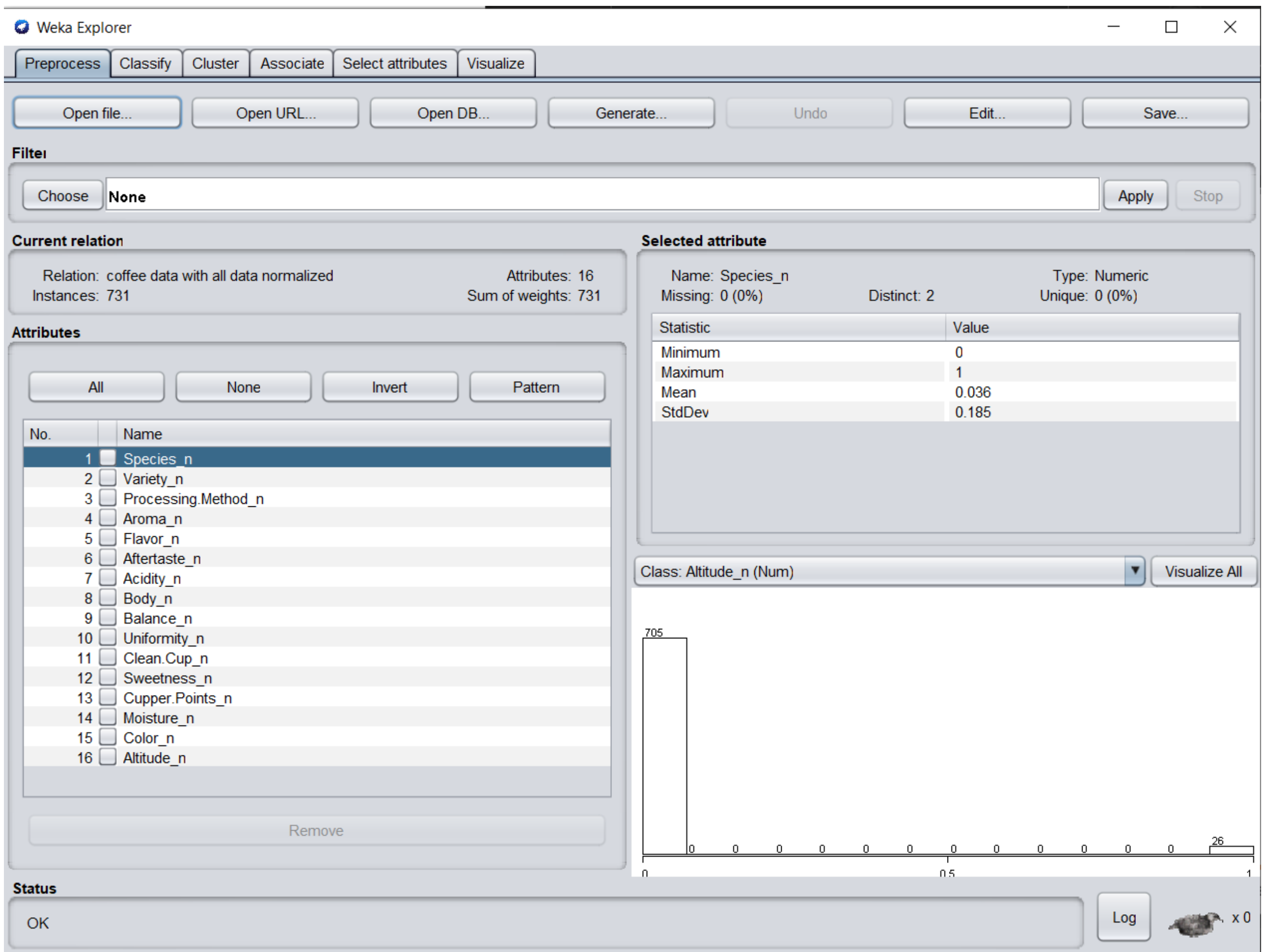
- We use weka for the clustering job.
- Weka is tried and tested open-source machine learning software that can be accessed through a graphical user interface.
- Weka can be used to build machine learning pipelines, train classifiers, and most importantly, cluster data and run evaluations without having to write a single line of code.



[3]

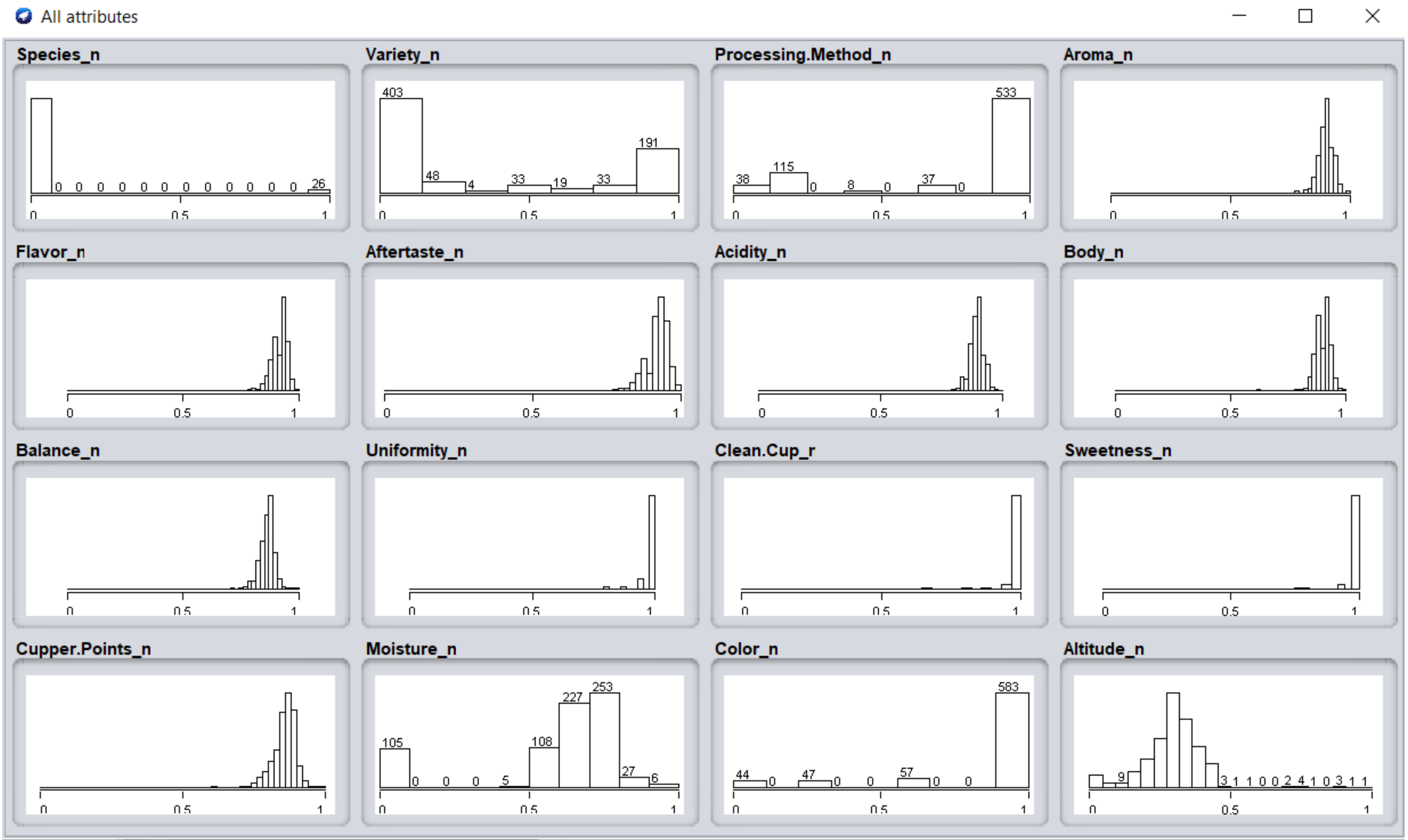
• Software Package (Weka)

- We load the .csv file into weka and get the following window:



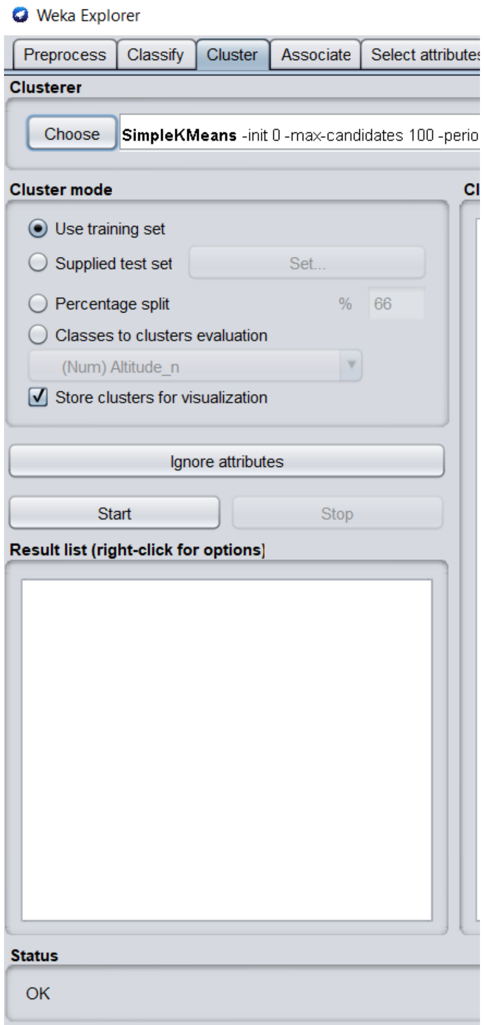
Software Package (Weka)

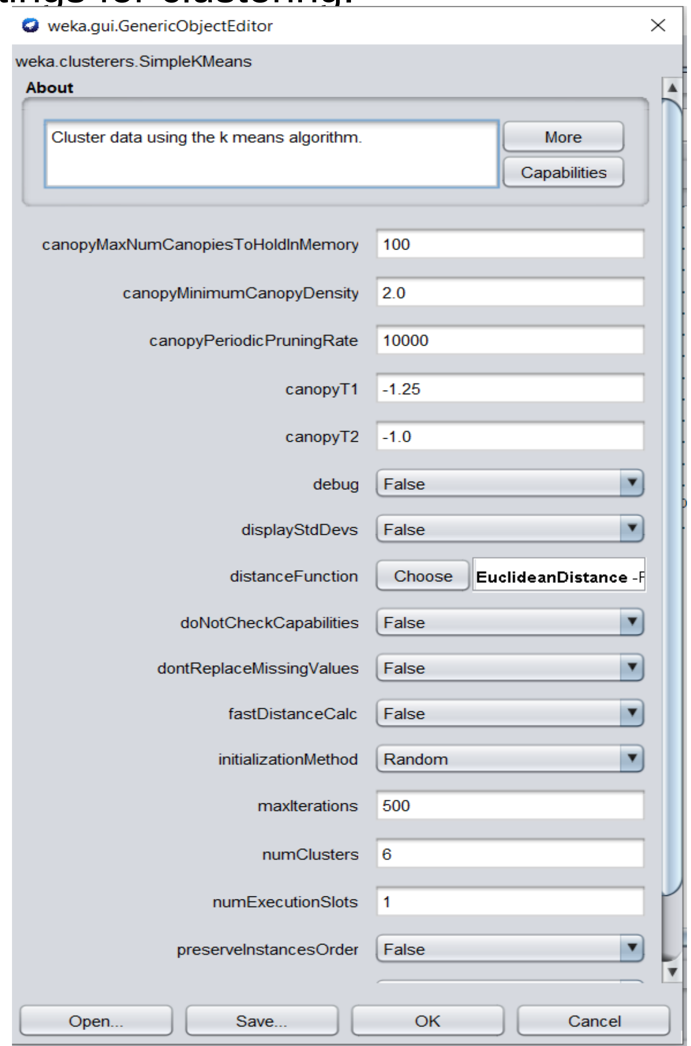
- We can visualize the distribution of the data as follows:



Software Package (Weka)

- Before running the algorithm, we the following settings for clustering.





Determination of Number of Clusters

- Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.
- There is no definite answer to this question. However, there are several methods that try to optimize the number of clusters required. We will be using elbow method.
- The steps involved in elbow method are:
  - Compute clustering for different values of k. For instance, by varying k from 1 to 15 clusters.
  - For each k, calculate the total cluster sum of square error (SSE).
  - Plot the curve of SSE according to the number of clusters k.
  - The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

[4]

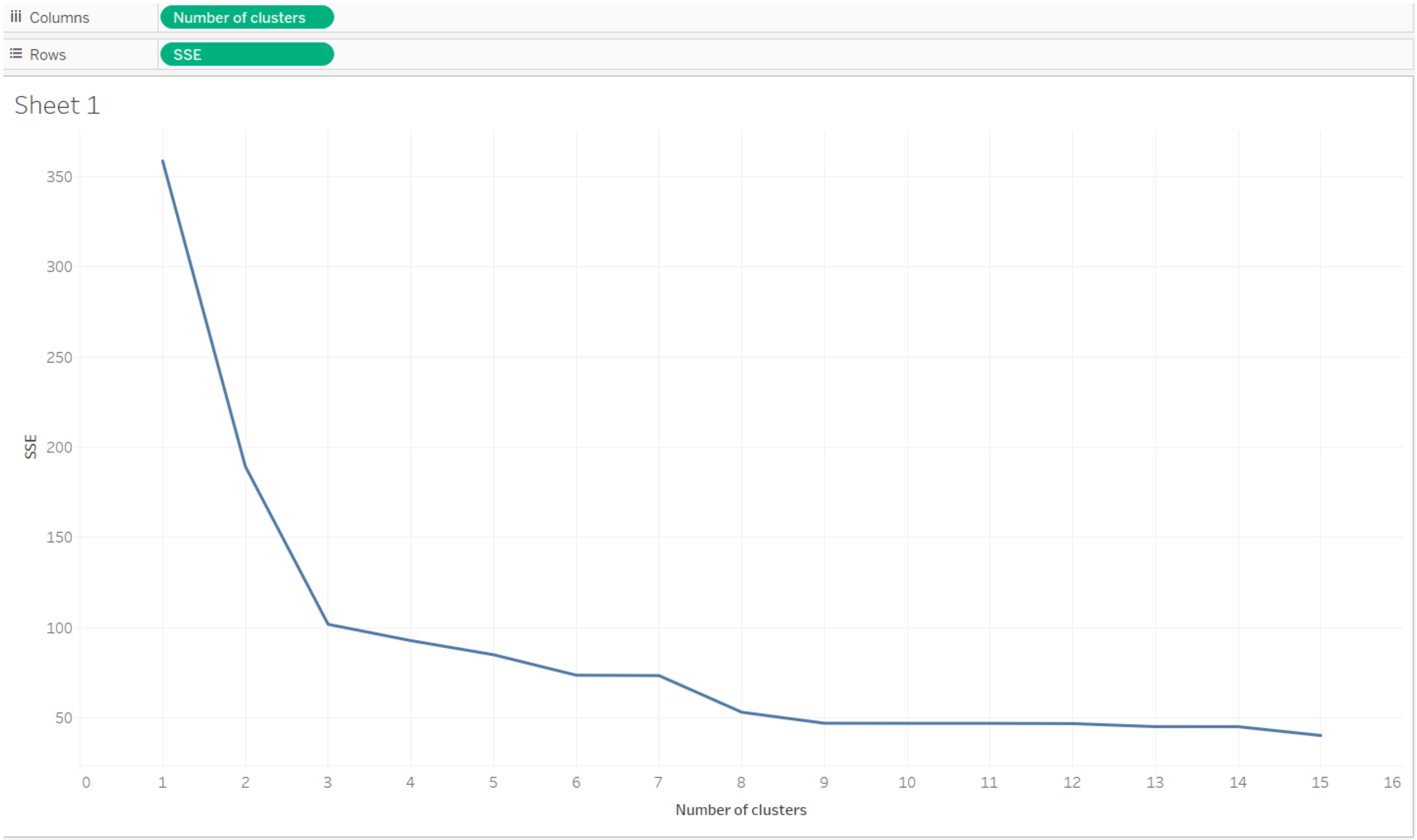
Determination of Number of Clusters

- Therefore, we run the algorithm 15 times, varying the number of clusters from 1 to 15. We noted the SSE each time and gather the following data:

A	B	C
Number of clusters	SSE	
1	358.6648061	
2	189.2773032	
3	101.7620872	
4	92.71709252	
5	84.88316876	
6	73.54978109	
7	73.33017989	
8	53.01940614	
9	46.9141936	
10	46.85524908	
11	46.84815676	
12	46.70121649	
13	45.02477203	
14	45.00221376	
15	40.11590983	

Determination of Number of Clusters

- We plot the data using Tableau. [5]



Determination of Number of Clusters

- As we can see, there isn't a significant drop in SSE after 6 number of clusters.
- Therefore, we keep the number of clusters to 6 for our clustering instance.

Execution and Analysis

- The result of clustering for k=6

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 73.5497810878769

Initial starting points (random):

Cluster 0: 0,0,0,0.780312,0.783416,0.77904,0.7503,0.636364,0.719114,0.667,0.267,0.6,0.
Cluster 1: 0,0.653846,1,0.90036,0.918317,0.925505,0.890756,0.899394,0.864802,1,1,1,0.8
Cluster 2: 0,0,1,0.930372,0.949257,0.957071,0.920768,0.929697,0.893939,1,1,1,0.893939,
Cluster 3: 0,0.769231,1,0.90036,0.928218,0.925505,0.90036,0.899394,0.864802,1,1,1,0.86
Cluster 4: 0,0.961538,1,0.90036,0.928218,0.925505,0.90036,0.909091,0.874126,1,1,1,0.86
Cluster 5: 0,0.115385,0.25,0.870348,0.887376,0.883838,0.860744,0.869091,0.825175,1,1,1,

Missing values globally replaced with mean/mode

Final cluster centroids:
                                         Cluster#
Attribute               Full Data           0           1           2           3
                          (731.0)      (40.0)      (50.0)     (290.0)      (39.0)     (191.0)
==========================================================================================
Species_n                  0.0356           0           0      0.0172           0
Variety_n                  0.3759      0.0529      0.4915      0.0829      0.7475      0.965
Processing.Method_n        0.8119      0.1437           1      0.9681           1
Aroma_n                    0.8969      0.8011      0.8988      0.9066      0.9004       0.91
Flavor_n                   0.9164      0.8096      0.9183      0.9281      0.9236      0.932
Aftertaste_n               0.9183      0.8058       0.918       0.931      0.9255      0.936
Acidity_n                  0.8915      0.7981      0.8908      0.9019      0.8969       0.90
```

Log   x 0

Execution and Analysis

Clusterer output

```
Flavor_n                    0.9164    0.8096    0.9183    0.9281    0.9236    0.932
Aftertaste_n                0.9183    0.8058     0.918     0.931    0.9255    0.936
Acidity_n                   0.8915    0.7981    0.8908    0.9019    0.8969     0.90
Body_n                      0.8978    0.7975    0.8987    0.9083    0.8994     0.91
Balance_n                   0.8604     0.758    0.8625    0.8705    0.8648    0.875
Uniformity_n                0.9829    0.7918         1         1         1
Clean.Cup_n                 0.9801      0.68         1         1         1
Sweetness_n                 0.9831     0.738         1         1         1
Cupper.Points_n             0.8562     0.733    0.8552    0.8689    0.8648    0.874
Moisture_n                  0.5775         0    0.6471    0.6773    0.6471    0.685
Color_n                      0.871         0         1    0.9954         1
Altitude_n                  0.2987    0.0543    0.2879    0.3283     0.299    0.331



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0        40 (  5%)
1        50 (  7%)
2       290 ( 40%)
3        39 (  5%)
4       191 ( 26%)
5       121 ( 17%)
```

Log x 0

Execution and Analysis

- The final cluster centroids were:

Final cluster centroids:

Attribute	Full Data (731.0)	Cluster# 0 (40.0)	1 (50.0)	2 (290.0)	3 (39.0)	4 (191.0)	5 (121.0)
Species_n	0.0356	0	0	0.0172	0	0	0.1736
Variety_n	0.3759	0.0529	0.4915	0.0829	0.7475	0.9658	0.0861
Processing.Method_n	0.8119	0.1437	1	0.9681	1	1	0.2231
Aroma_n	0.8969	0.8011	0.8988	0.9066	0.9004	0.911	0.8808
Flavor_n	0.9164	0.8096	0.9183	0.9281	0.9236	0.9329	0.8945
Aftertaste_n	0.9183	0.8058	0.918	0.931	0.9255	0.9361	0.8946
Acidity_n	0.8915	0.7981	0.8908	0.9019	0.8969	0.905	0.8747
Body_n	0.8978	0.7975	0.8987	0.9083	0.8994	0.911	0.8841
Balance_n	0.8604	0.758	0.8625	0.8705	0.8648	0.8752	0.8446
Uniformity_n	0.9829	0.7918	1	1	1	1	0.9657
Clean.Cup_n	0.9801	0.68	1	1	1	1	0.9856
Sweetness_n	0.9831	0.738	1	1	1	1	0.9845
Cupper.Points_n	0.8562	0.733	0.8552	0.8689	0.8648	0.8749	0.8346
Moisture_n	0.5775	0	0.6471	0.6773	0.6471	0.6852	0.3082
Color_n	0.871	0	1	0.9954	1	1	0.562
Altitude_n	0.2987	0.0543	0.2879	0.3283	0.299	0.3316	0.2613

Execution and Analysis

- Centers of the clusters can be used to characterize the clusters.
- These properties are the predominant properties of the coffee beans in a given cluster.
- For example:
  - We can interpret from the clustering result that a coffee bean that is high in Aroma, flavor, aftertaste, acidity, and grows at a higher altitude belongs to cluster 5.
  - Now, if any customer has bought a coffee bean in the past from this cluster and liked it, we can recommend them a different coffee bean from the same cluster and be confident that the customer will like this recommendation too.
  - This is how targeted marketing works these days.
  - This clustering recommendation can be implemented in online websites as well as local coffee shops. The customer gets recommendations on the website online or if they to a local coffee shop regularly, the barista at the coffee shop can access the customers past purchases and recommend a coffee bean accordingly by looking at the other coffee beans in the cluster.

Conclusion

- In this task, we have successfully collected data, understood it, prepared it, normalized it and removed the anomalies.
- We have determined the optimal number of clusters and clustered the data accordingly.
- Upon successful clustering, we get 6 clusters with 6 centroids. These centroids characterize the clusters and all the coffee beans in these clusters have properties like the coffee bean which is the centroid.
- The same method can be used in various other domains such as movie recommendation, snacks recommendation, clothing recommendation, electronics recommendation, etc.
- Clustering is very powerful because it can be used with the customers data for targeted personalized marketing.
- Clustering is used in most online websites such as amazon, Netflix, spotify, google, hulu, steam etc. to make personalized recommendations.
- These recommendations in-turn generate revenue.
- Therefore, a lot of money is invested by these giants on these unsupervised machine learning algorithms to make sure they're always on the top of their competition.

References

Data source: https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi?select=merged_data_cleaned.csv

[1] https://www.youtube.com/watch?v=rpM3YW_hSaM

[2] https://en.wikipedia.org/wiki/K-means_clustering

[3] https://www.cs.waikato.ac.nz/ml/weka/

[4] https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#gap-statistic-method

[5] https://public.tableau.com/en-us/s/