Table of Contents

Executive Summary：

Data science is one of the world's most rapidly emerging field although it is not entirely new. Nowadays, 'big data' and advanced problem-solving skills provide a very accurate decision making and productive innovation for all organisations. Therefore, countries and organisations are deeply delving in this developing field, thus detonating the next industrial revolution, evolving whole society into industry 4.0.

The primary purpose of this report is to explore 3 data science approaches to classify wine and predict forest fire in the real world. Two different supervised data science techniques, K-nearest neighbours (In the paper, abbreviate as KNN) and Linear-regression, were applied on the recent real-world data collected from the northeast region of Portugal. Also, one unsupervised data science techniques, K-MEANS, was applied to the data collected from a winery.

After experiments, compared with Linear- regression, the KNN relatively working well on the forest fire dataset. Also, the accuracy of K-MEAS in wine classification could achieve to overall (90.14%+100%+100%)/3=96.71%.

Keywords: Data science, Machine Learning, Forest Fire, Wine Classification, Linear - Regression, K-nearest neighbours (KNN), K-means, Unsupervised learning

1.Introduction：

Without a doubt, data in data science is not limited to the data generated by the internet but a large amount of data will be generated in industries such as finance, biological information, government, education. Some giants of industries collect as much information as "big data", while others do not pay attention. The subject of data science has become increasingly interesting (or challenging), not just because of the increasing volume of data, but also because the data itself (often in real time) has become a key element in building data products. On the internet, there are Amazon's recommendation system, Facebook's friend recommendation system and other recommendation systems. In the financial aspect, there are credit rating systems, trading algorithms and models. In the field of education, the products can be customised for students, such as Coursera and Knew ton. For the government, this means that they could make public policy based on 'big data' to tackle the challenges such as

forest fire. It also can be used by the traditional company to quickly classify the products, for instance, the wine sorting.

In this report, the data science algorithm such as K-nearest neighbours (KNN) and Linear -Regression which classified as supervised learning skill will be applied on a data set that collected from the northeast region of Portugal, with the aim of predicting the burnt area (or size) of forest fires. The un-supervised learning skill, K-meanings, is going to implement on a data set which is from a winery in the real world to classify the new wine.

The paper is organised as follows. First, the forest fire data will be described in Section 2. The adopted KNN and Linear -Regression algorithm are presented in Section 3, while the results of supervised learning algorithm are shown and discussed in Section 4. After supervised learning, a data set for new wine will be introduced in section 5. The unsupervised algorithm, K-means will be applied to the data set in section 6. Conclusions of K-means are drawn in section 7.

2. Dataset of Forest Fire

2.1 Introduction of data set

The forest fire data presented in the paper is from the database of Montesinho national park in northeast Portugal. The dataset contains 13 variables: spatial coordinates of Montesinho national park; The month of year and day of a week; FFMC (fine fuel moisture code), DMC (duff moisture code), DC (drought code) and ISI (initial spread index); Four meteorological data such as temperatures, relative humidity, wind speed and precipitation; finally, there is the area of forest fire burnt.

2.2 Normalisation

In the field of machine learning and data science, different parameters may tend to have different dimensions and units, such a situation will affect the result of data analysis. In order to solve the comparability problem among the data, the data set should be changed all indexes to the same order of magnitude. This process is called “Normalisation”.

In short, the purpose of normalisation is restricting pre-processed data to a certain range (such as [0,1] or [-1,1]) to eliminate adverse effects caused by the different magnitude of units of data.

The Min-Max Normalization is also known as deviation normalisation. It can move the values into where between 0 and 1. The function is written as follows.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

In this function, x is a value at an attribute in a dataset. The Max is the maximum value of that attribute and min is the minimum value of the attribute. One advantage of Normalisation is improved accuracy. Especially, for the algorithm that contains distance calculation such as calculating the Euclidean distance. However, there is a significant disadvantage of this approach: If a new data is added, it may lead to changes on Max and min which mean that the Max and Min need to be redefined.

Due to a different magnitude of units in the original data set of a forest fire which the KNN algorithm is going to calculate Euclidean distance on it, the Max-min normalisation method was determined to be applied on that data set (Graph 1).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	X	Y	Month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
2	7	5	3	5	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
3	7	4	10	2	90.6	35.4	669.1	6.7	18	33	0.9	0	0
4	7	4	10	6	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
5	8	6	3	5	91.7	33.3	77.5	9	8.3	97	4	0.2	0
6	8	6	3	7	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
7	8	6	8	7	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	8	6	8	1	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
9	8	6	8	1	91.5	145.4	608.2	10.7	8	86	2.2	0	0
10	8	6	9	2	91	129.5	692.6	7	13.1	63	5.4	0	0
11	7	5	9	6	92.5	88	698.6	7.1	22.8	40	4	0	0
12	7	5	9	6	92.5	88	*698.6*	7.1	17.8	51	7.2	0	0
13	7	5	9	6	92.8	73.2	713	22.6	19.3	38	4	0	0
14	6	5	8	5	63.5	70.8	665.3	0.8	17	72	6.7	0	0
15	6	5	9	1	90.9	126.5	686.5	7	21.3	42	2.2	0	0
16	6	5	9	3	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0
17	6	5	9	5	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0
18	5	5	3	6	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0
19	8	5	10	1	84.9	32.8	664.2	3	16.7	47	4.9	0	0
20	6	4	3	3	89.2	27.9	70.8	6.3	15.9	35	4	0	0
21	6	4	4	6	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0
22	6	4	9	2	91	129.5	692.6	7	18.3	40	2.7	0	0
23	5	4	9	1	91.8	78.5	724.3	9.2	19.1	38	2.7	0	0
24	7	4	6	7	94.3	96.3	200	56.1	21	44	4.5	0	0
25	7	4	8	6	90.2	110.9	537.4	6.2	19.5	43	5.8	0	0
26	7	4	8	6	93.5	139.4	594.2	20.3	23.7	32	5.8	0	0
27	7	4	8	7	91.4	142.4	601.4	10.6	16.3	60	5.4	0	0

Graph1

After Max-min normalisation, every data in the data set (Graph2) has been limited between 0 to 1, and the data in the different attributes became more comparable between each other. Furthermore, after normalisation, the dataset could be more suitable for the dimensional reduction (PCA) which will be implemented in the next step. The purpose of PCA is to reduce the noise and increase the efficiency of computation.

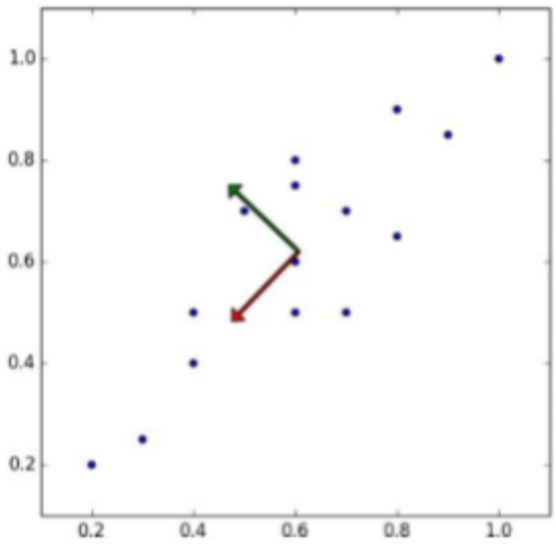
P	Q	R	S	T	U	V	W	X	Y	Z	AA
X	Y	Month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain
0.75	0.428571	0.181818	0.666667	0.873221	0.086881	0.101959	0.090909	0.214286	0.423529	0.7	0
0.75	0.285714	0.818182	0.166667	0.930142	0.118726	0.780269	0.11943	0.564286	0.211765	0.055556	0
0.75	0.285714	0.818182	0.833333	0.930142	0.147456	0.801274	0.11943	0.442857	0.211765	0.1	0
0.875	0.571429	0.181818	0.666667	0.944373	0.111457	0.082134	0.160428	0.217857	0.964706	0.4	0.2
0.875	0.571429	0.181818	1	0.913325	0.173763	0.111282	0.171123	0.328571	0.988235	0.155556	0
0.875	0.571429	0.636364	1	0.952135	0.29145	0.566557	0.262032	0.714286	0.164706	0.555556	0
0.875	0.571429	0.636364	0	0.952135	0.303911	0.575525	0.151515	0.782143	0.141176	0.3	0
0.875	0.571429	0.636364	0	0.941785	0.499481	0.708402	0.190731	0.207143	0.835294	0.2	0
0.875	0.571429	0.727273	0.166667	0.935317	0.444444	0.808001	0.124777	0.389286	0.564706	0.555556	0
0.75	0.428571	0.727273	0.833333	0.954722	0.300796	0.815081	0.12656	0.735714	0.294118	0.4	0
0.75	0.428571	0.727273	0.833333	0.954722	0.300796	0.815081	0.12656	0.557143	0.423529	0.755556	0
0.75	0.428571	0.727273	0.833333	0.958603	0.249567	0.832075	0.402852	0.610714	0.270588	0.4	0
0.625	0.428571	0.636364	0.666667	0.57956	0.24126	0.775785	0.01426	0.528571	0.670588	0.7	0
0.625	0.428571	0.727273	0	0.934023	0.43406	0.800802	0.124777	0.682143	0.317647	0.2	0
0.625	0.428571	0.727273	0.333333	0.959897	0.457598	0.816262	0.163993	0.864286	0.070588	0.455556	0
0.625	0.428571	0.727273	0.666667	0.965071	0.484943	0.833137	0.247772	0.739286	0.341176	0.555556	0
0.5	0.428571	0.181818	0.833333	0.944373	0.120111	0.086028	0.139037	0.460714	0.141176	0.555556	0
0.875	0.428571	0.818182	0	0.856404	0.109727	0.774487	0.053476	0.517857	0.376471	0.5	0
0.625	0.285714	0.181818	0.333333	0.912031	0.092766	0.074227	0.112299	0.489286	0.235294	0.4	0
0.625	0.285714	0.272727	0.833333	0.874515	0.091035	0.105263	0.090909	0.253571	0.341176	0.455556	0
0.625	0.285714	0.727273	0.166667	0.935317	0.444444	0.808001	0.124777	0.575	0.294118	0.255556	0
0.5	0.285714	0.727273	0	0.945666	0.267913	0.845409	0.163993	0.603571	0.270588	0.255556	0
0.75	0.285714	0.454545	1	0.978008	0.329526	0.226693	1	0.671429	0.341176	0.455556	0
0.75	0.285714	0.636364	0.833333	0.924968	0.380062	0.624852	0.110517	0.617857	0.329412	0.6	0
0.75	0.285714	0.636364	0.833333	0.967658	0.478712	0.691881	0.361854	0.767857	0.2	0.6	0
0.75	0.285714	0.636364	1	0.940492	0.489097	0.700378	0.188948	0.503571	0.529412	0.555556	0

Graph2

2.3 Dimension Deduction

In actual data science projects, feature selection or dimension reduction is necessary. The purpose of dimension reduction is reducing the number of feature attributes in the data set and make sure the attributes are independent of each other. The standard methods of dimensionality reduction are PCA, LDA. In this paper, we will implement the PCA which is called “Principal Component Analysis”.

PCA assumes that the characteristic of the original data is n-dimensional, and first selects the maximum direction of variance as the first dimensional data (first axis). The second axis is going to be perpendicular or orthogonal to the first axis. Choose a direction for the third coordinate axis which vertical or orthogonal with first and second coordinate axes. This process is repeated until the dimension of the new coordinate system reaches the given value. The data characteristics represented by these directions are called "principal components". (Graph3)



Graph 3

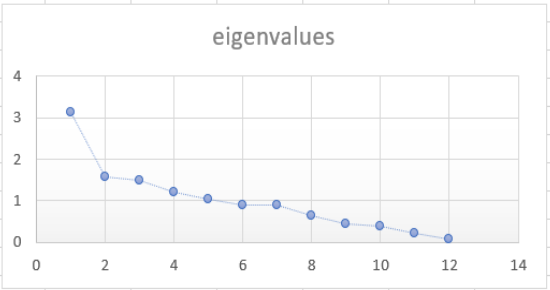
In the paper, in order to reduce the dimension, two steps need to be applied on the dataset which was normalised before. Firstly, we need to calculate the variance of data in the forest fire data set; therefore, excel function "= CORR ()" was applied (Graph4). Note that the total variance no longer adds up to the total variance in the original sample, because the correlation matrix involves normalised data.

N	O	P	Q	R	S	T	U	V	W	X	Y
X	Y	Month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain
1	0.539548	-0.06918	-0.02492	-0.02278	-0.04833	-0.08575	-0.02929	-0.04866	0.080612	0.020666	0.08483
0.539548	1	-0.06241	-0.00545	0.005919	0.007825	-0.10118	0.002641	-0.02796	0.05491	-0.0209	0.031668
-0.06918	-0.06241	1	-0.04573	0.275952	0.498743	0.874481	0.121042	0.406908	-0.08645	-0.08704	0.009322
-0.02492	-0.00545	-0.04573	1	-0.06812	0.062833	-7.4E-05	-0.02112	0.049706	0.094328	0.032017	-0.00241
-0.02278	0.005919	0.275952	-0.06812	1	0.173275	0.269245	0.128966	0.368253	-0.32536	-0.0289	0.040567
-0.04833	0.007825	0.498743	0.062833	0.173275	1	0.681291	0.385603	0.414569	0.09496	-0.13188	0.078258
-0.08575	-0.10118	0.874481	-7.4E-05	0.269245	0.681291	1	0.14869	0.508845	-0.03465	-0.20576	0.028886
-0.02929	0.002641	0.121042	-0.02112	0.128966	0.385603	0.14869	1	0.325136	-0.03515	0.085265	0.102326
-0.04866	-0.02796	0.406908	0.049706	0.368253	0.414569	0.508845	0.325136	1	-0.56104	-0.25929	0.036871
0.080612	0.05491	-0.08645	0.094328	-0.32536	0.09496	-0.03465	-0.03515	-0.56104	1	0.075026	0.151501
0.020666	-0.0209	-0.08704	0.032017	-0.0289	-0.13188	-0.20576	0.085265	-0.25929	0.075026	1	0.100875
0.08483	0.031668	0.009322	-0.00241	0.040567	0.078258	0.028886	0.102326	0.036871	0.151501	0.100875	1

Graph 4

Then, the covariance matrix needs to be calculated (Graph5). The first row gives the eigenvalues of the covariance matrix, and the matrix below gives the eigenvectors which are the same as the principal components.

12	3.125206	1.580795	1.480765	1.198707	1.027311	0.913105	0.888499	0.655705	0.456135	0.378495	0.219893	0.075384
X	0.083474	0.575967	-0.368	0.100513	-0.01398	-0.07673	-0.05708	0.137816	-0.6857	-0.11409	-0.0636	-0.01368
Y	0.064032	0.563391	-0.39358	0.122013	0.065682	0.008006	-0.12195	-0.05259	0.67886	0.14547	0.035036	0.058689
Month	-0.45141	0.08986	0.158297	0.187503	-0.22518	-0.26917	-0.14555	0.185346	-0.00542	0.468805	-0.15546	-0.55251
day	0.007176	0.051421	0.172817	0.010896	0.87016	-0.41891	0.033712	-0.06817	-0.04463	0.121486	-0.10983	0.007817
FFMC	-0.27456	-0.09223	-0.32243	-0.20854	-0.15875	-0.35844	0.016061	-0.77927	-0.08165	-0.01948	0.020419	-0.00162
DMC	-0.41558	0.244928	0.248383	-0.01519	0.099595	0.191863	-0.08732	-0.09481	0.132234	-0.70511	-0.28301	-0.21744
DC	-0.49693	0.108573	0.204422	0.216679	-0.11846	-0.14643	-0.04335	0.111591	-0.03353	0.069735	0.044544	0.771204
ISI	-0.21844	0.137156	0.031065	-0.55541	0.182623	0.560781	-0.21994	-0.10274	-0.13756	0.417178	-0.12838	0.08849
temp	-0.43499	-0.10388	-0.30516	-0.09768	0.23238	0.079542	0.141177	0.259654	-0.00694	-0.11884	0.712483	-0.16688
RH	0.176988	0.386076	0.565459	0.113778	-0.07321	0.080086	0.011128	-0.37219	-0.08698	0.079396	0.557863	-0.1069
wind	0.145851	0.064649	0.132649	-0.56257	-0.14866	-0.4478	-0.52897	0.262182	0.082232	-0.18175	0.166377	0.04265
rain	-0.02352	0.277511	0.126008	-0.44902	-0.15375	-0.17787	0.775594	0.159077	0.091492	0.036926	-0.114	0.008942
	0.260434	0.131733	0.123397	0.099892	0.085609	0.076092	0.074042	0.054642	0.038011	0.031541	0.018324	0.006282
	0.905841											



Graph5

In this step, the scatter plot has been generated by imputing values of eigenvalues. According to the graph, the last four eigenvalues were quite small compared with other eigenvalues. Therefore, the proportion between the first eight eigenvalues was calculated by the Excel equation:

=SUM (Eigenvalue(i)/SUM(Eigenvalue));

After calculation, we could see that the total percentage of the first eight eigenvalues was 90.58%. It revealed that the first 8 eigenvalues could explain about 90% of the total variability; in the contrary, the last 4 eigenvalues just present around 10% of the total variability which means if study just remain the first of 8 eigenvalues, the dataset that after dimension reduction will still have the 90% of variability compared with original data.

In the last step of dimension reduction, the Matric of original data was multiplied with the covariance matrix which has eight columns of data remained after it has been choosing. Then, the new dataset has been prepared. (**Graph6**)

AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
X	0.083474	0.575967	-0.368	0.100513	-0.01398	-0.07673	-0.05708	0.137816	
Y	0.064032	0.563391	-0.39358	0.122013	0.065682	0.008006	-0.12195	-0.05259	
Month	-0.45141	0.08986	0.158297	0.187503	-0.22518	-0.26917	-0.14555	0.185346	
day	0.007176	0.051421	0.172817	0.010896	0.87016	-0.41891	0.033712	-0.06817	
FFMC	-0.27456	-0.09223	-0.32243	-0.20854	-0.15875	-0.35844	0.016061	-0.77927	
DMC	-0.41558	0.244928	0.248383	-0.01519	0.099595	0.191863	-0.08732	-0.09481	
DC	-0.49693	0.108573	0.204422	0.216679	-0.11846	-0.14643	-0.04335	0.111591	
ISI	-0.21844	0.137156	0.031065	-0.55541	0.182623	0.560781	-0.21994	-0.10274	
temp	-0.43499	-0.10388	-0.30516	-0.09768	0.23238	0.079542	0.141177	0.259654	
RH	0.176988	0.386076	0.565459	0.113778	-0.07321	0.080086	0.011128	-0.37219	
wind	0.145851	0.064649	0.132649	-0.56257	-0.14866	-0.4478	-0.52897	0.262182	
rain	-0.02352	0.277511	0.126008	-0.44902	-0.15375	-0.17787	0.775594	0.159077	
	-0.24978	0.874837	-0.27004	-0.40933	0.346112	-0.90511	-0.45236	-0.53606	
	-1.20566	0.746155	-0.38242	0.110154	-0.13002	-0.66625	-0.19401	-0.35039	
	-1.16395	0.805241	-0.21283	0.108391	0.415634	-0.97264	-0.21561	-0.4161	
	-0.21951	1.278829	-0.10066	-0.29726	0.335	-0.75007	-0.17201	-0.84081	
	-0.32625	1.245002	-0.08939	-0.06854	0.725782	-0.70916	-0.18025	-0.91901	
	-1.0923	1.040838	-0.43514	-0.30144	0.682118	-1.05292	-0.46209	-0.32316	
	-1.15592	0.94563	-0.67438	-0.11474	-0.15256	-0.57696	-0.32847	-0.28444	
	-0.95057	1.33554	-0.03938	0.082855	-0.30959	-0.47856	-0.38066	-0.71798	
	-1.07611	1.240761	-0.15089	-0.08649	-0.20406	-0.79906	-0.53876	-0.44312	
	-1.26172	0.93609	-0.25324	-0.08765	0.473953	-1.02889	-0.35139	-0.3493	
	-1.10928	1.027589	-0.07841	-0.25551	0.370126	-1.19195	-0.56324	-0.35061	
	-1.26008	0.966826	-0.23032	-0.22793	0.489354	-0.89949	-0.42627	-0.39766	

Graph6

2.4 Split dataset

The testing set is used to evaluate the generalisation of the model, that is, used a dataset never seen before to judge whether the model is working correctly. Therefore, for a given data set, usually, it will be divided into two mutually exclusive sets: a training set and testing set. In the paper, the data was split into two parts as well. The first part is used for training purpose which takes about 80% of the original data and contains 395 volume of data (Graph7). Another part is called testing set which takes around 20% of original data and has 122 volume of data (Graph8).

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area											
2	3	5	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0			-0.24978	0.874837	-0.27004	-0.40933	0.346112	-0.90511	-0.45236	-0.53606	
3	10	2	90.6	35.4	669.1	6.7	18	33	0.9	0	0			-1.20566	0.746155	-0.38242	0.110154	-0.13002	-0.66625	-0.19401	-0.35039	
4	10	6	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0			-1.16395	0.805241	-0.21283	0.108391	0.415634	-0.97264	-0.21561	-0.4161	
5	3	5	91.7	33.3	77.5	9	8.3	97	4	0.2	0			-0.21951	1.278829	-0.10066	-0.29726	0.335	-0.75007	-0.17201	-0.84081	
6	3	7	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0			-0.32625	1.245002	-0.08939	-0.06854	0.725782	-0.70916	-0.18025	-0.91901	
7	8	7	92.3	85.3	488	14.7	22.2	29	5.4	0	0			-1.0923	1.040838	-0.43514	-0.30144	0.682118	-1.05292	-0.46209	-0.32316	
8	8	1	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0			-1.15592	0.94563	-0.67438	-0.11474	-0.15256	-0.57696	-0.32847	-0.28444	
9	8	1	91.5	145.4	608.2	10.7	8	86	2.2	0	0			-0.95057	1.33554	-0.03938	0.082855	-0.30959	-0.47856	-0.38066	-0.71798	
10	9	2	91	129.5	692.6	7	13.1	63	5.4	0	0			-1.07611	1.240761	-0.15089	-0.08649	-0.20406	-0.79906	-0.53876	-0.44312	
11	9	6	92.5	88	698.6	7.1	22.8	40	4	0	0			-1.26172	0.93609	-0.25324	-0.08765	0.473953	-1.02889	-0.35139	-0.3493	
12	9	6	92.5	88	*698.6*	7.1	17.8	51	7.2	0	0			-1.10928	1.027589	-0.07841	-0.25551	0.370126	-1.19195	-0.56324	-0.35061	
13	9	6	92.8	73.2	713	22.6	19.3	38	4	0	0			-1.26008	0.966826	-0.23032	-0.22793	0.489354	-0.89949	-0.42627	-0.39766	
14	8	5	63.5	70.8	665.3	0.8	17	72	6.7	0	0			-0.86002	1.033964	0.160103	-0.09178	0.268622	-0.97985	-0.49481	-0.18212	
15	9	1	90.9	126.5	686.5	7	21.3	42	2.2	0	0			-1.32205	0.855717	-0.31133	0.011312	-0.21595	-0.54895	-0.28481	-0.38274	
16	9	3	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0			-1.43847	0.785509	-0.41323	-0.19891	0.100095	-0.79335	-0.39673	-0.22391	
17	9	5	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0			-1.3587	0.946111	-0.14003	-0.25291	0.341628	-0.91814	-0.47449	-0.36693	
18	3	6	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0			-0.48363	0.601916	-0.4197	-0.4531	0.596734	-0.88388	-0.33606	-0.51599	
19	10	1	84.9	32.8	664.2	3	16.7	47	4.9	0	0			-1.03184	0.982133	-0.32888	-0.03753	-0.35688	-0.80584	-0.44965	-0.2217	
20	3	3	89.2	27.9	70.8	6.3	15.9	35	4	0	0			-0.47244	0.582366	-0.47163	-0.34854	0.172322	-0.61261	-0.24701	-0.49634	
21	4	6	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0			-0.38024	0.688674	-0.21431	-0.29577	0.514345	-0.88511	-0.30196	-0.56456	
22	9	2	91	129.5	692.6	7	18.3	40	2.7	0	0			-1.2877	0.792646	-0.19591	-0.02663	-0.11176	-0.65472	-0.30775	-0.39227	
23	9	1	91.8	78.5	724.3	9.2	19.1	38	2.7	0	0			-1.27256	0.665276	-0.23906	-0.05963	-0.26318	-0.59599	-0.29712	-0.37314	
24	6	7	94.3	96.3	200	56.1	21	44	4.5	0	0			-1.0189	0.92891	-0.25131	-0.79193	0.872744	-0.47954	-0.49532	-0.5996	
25	8	6	90.2	110.9	537.4	6.2	19.5	43	5.8	0	0			-1.06988	0.88554	-0.13906	-0.24642	0.457561	-1.05694	-0.4384	-0.35385	
26	8	6	93.5	139.4	594.2	20.3	23.7	32	5.8	0	0			-1.29896	0.88197	-0.22576	-0.41127	0.5429	-0.92062	-0.48478	-0.32769	
27	8	7	91.4	142.4	601.4	10.6	16.3	60	5.4	0	0			-1.0943	1.024556	0.071768	-0.21778	0.581765	-1.05165	-0.45298	-0.50304	

Graph7

ain		area								
0		51.78	-0.07958	0.71986	0.041732	-0.05999	0.819691	-0.53128	-0.08452	-0.56777
0		3.64	-1.26101	0.769105	0.330195	-0.29918	0.625528	-0.82172	-0.45023	-0.3626
0		3.63	-1.32516	0.949071	-0.12714	-0.16477	0.760472	-0.78827	-0.37328	-0.34192
0		0	-1.60786	0.367065	0.256691	-0.27977	0.60199	-0.65305	-0.26928	-0.44949
0		0	-0.9511	0.984923	-0.62834	-0.51803	0.270917	-0.40899	-0.42448	-0.29902
0		8.16	-0.9511	0.984923	-0.62834	-0.51803	0.270917	-0.40899	-0.42448	-0.29902
0		4.95	-1.29177	0.652686	0.094436	-0.18827	0.213458	-0.65656	-0.37326	-0.31684
0		0	-1.4646	1.54541	-0.49413	-0.22076	0.505428	-0.61725	-0.55041	-0.41356
0		0	-1.24718	1.208199	-0.22328	-0.07121	0.287127	-0.61022	-0.4224	-0.28939
0		6.04	-1.24974	0.473923	0.033135	0.056686	0.088534	-0.57348	-0.21757	-0.18072
0		0	-1.6369	0.34572	-0.01751	-0.27686	0.058778	-0.32769	-0.23763	-0.40661
0		3.95	-1.3617	1.289028	0.46629	0.045385	0.465986	-0.80153	-0.50636	-0.31355
0		0	-0.13354	0.619971	-0.12494	-0.22873	0.683185	-0.59633	-0.20282	-0.35101
0		7.8	-1.60931	0.829931	0.348454	-0.07149	0.39201	-0.59857	-0.37421	-0.36259
0		0	-1.14101	0.387328	-0.08748	-0.56706	0.040769	-0.38823	-0.45293	-0.34367
0		0	-0.15975	0.560563	-0.18412	-0.15162	0.54963	-0.47833	-0.10304	-0.28098
0		4.62	0.099852	0.846301	-0.03501	-0.54838	0.377323	-0.83797	-0.57919	-0.20413
0		1.63	-1.0398	0.907529	-0.53376	-0.35197	-0.12418	-0.41922	-0.43651	-0.19413
0		0	-1.51425	1.048082	-0.17731	-0.18524	0.613899	-0.73097	-0.39285	-0.35412
0		0	-1.48529	0.955069	0.17926	-0.26051	0.725763	-0.74029	-0.40921	-0.4978
0		746.28	-1.55836	1.178782	-0.39755	-0.42442	0.364774	-0.54111	-0.56067	-0.30718
0		7.02	-1.3032	0.733936	-0.18459	-0.25347	0.051782	-0.37105	-0.30435	-0.34213
0		0	-0.37223	0.654081	-0.5545	-0.62804	0.21293	-0.50023	-0.43863	-0.45473
0		2.44	-1.40233	0.660858	0.244762	-0.34931	0.712634	-0.83176	-0.43906	-0.38389
0		3.05	-1.35863	0.670618	0.092602	-0.24312	0.7502	-0.77684	-0.3564	-0.38541
0		185.76	-1.31749	1.320944	-0.46135	-0.16171	0.191686	-0.56754	-0.50452	-0.23606

Graph8

3. Supervised Algorithm implement

3.1 ANN Algorithm

K-Nearest Neighbours (KNN), the algorithm is a classification algorithm which is one of the most straightforward machine learning algorithms in the data science field. It was proposed by Cover and Hart in 1968. The core idea of this algorithm is that if a sample is most similar

to k samples in the dataset and the most of these k samples belong to a certain category, then the algorithm identified the sample also belongs to this category. In the algorithm, the distance will be calculated by “Euclidean Distance equation”.

3.2 Implement ANN Algorithm in the Excel

In the study, the Excel built-in function will be used to apply the ANN Algorithm on the forest fire data set.

To implement the ANN Algorithm, a vector has been chosen from the train set for initial calculations. It is as following:

-0.07958	0.71986	0.041732	-0.05999	0.819691	-0.53128	-0.08452	-0.56777

Then, the Euclidean distance of each training point from the vector that has chosen was calculated and listed on the table. It shows as following:

AM	
Euclidean Distance	si
1.233407959	
1.157492204	
0.772308474	
1.38773482	
1.317077022	
0.863945555	
1.331034249	
1.350225885	
1.11192202	
0.699194233	
0.68841705	
0.620278323	
0.699886638	
1.153017111	
0.94524146	
0.595511319	
1.122305011	
1.259352952	
1.269005175	
1.076041388	
0.972114506	
1.121001876	
0.949470832	
0.597353425	
0.595694076	
0.490893414	

To get the rank of each of these Euclidean Distances, i.e. whether it is the smallest, the second smallest, and so in an Excel function "=Rank ()" was imputed.

In this study, the weighted voting rule was selected to implement the algorithm. Therefore, each training data point was weighted by the inverse of the square of the distance which is called Similarity weight. Then, scale those weights so that they add up to 1 which is Scaled weight. The sum of weights for each category then gives the probability of the vector belonging to each category (**Graph9**).

The Excel function of Similarity weight is:

=1/ (Euclidean Distance) ^ 2

The Excel function of Scaled weight is:

= Similarity weight (i) / SUM (Similarity Weight)

AK	AL	AM	AN	AO	AP	AQ	AR
	Rank	Euclidean Distance	similarity weight	Scaled weight	area		
	339	1.233407959	0.657334621	0.000949081	0		
	324	1.157492204	0.746386629	0.001077657	0		
	155	0.772308474	1.676557294	0.002420667	0		
	377	1.38773482	0.519262583	0.000749728	0		
	360	1.317077022	0.576471251	0.000832328	0		
	196	0.863945555	1.339760751	0.001934389	0		
	362	1.331034249	0.564444881	0.000814964	0		
	365	1.350225885	0.548513273	0.000791961	0		
	295	1.11192202	0.808818988	0.001167799	0		
	124	0.699194233	2.045522799	0.002953391	0		
	118	0.68841705	2.110069523	0.003046586	0		
	80	0.620278323	2.599122761	0.003752696	0		
	126	0.699886638	2.04147749	0.00294755	0		
	321	1.153017111	0.752191628	0.001086038	0		
	231	0.94524146	1.119217456	0.001615962	0		
	67	0.595511319	2.819810734	0.004071333	0		
	305	1.122305011	0.793922657	0.001146291	0		
	347	1.259352952	0.630529007	0.000910378	0		
	348	1.269005175	0.62097371	0.000896582	0		
	273	1.076041388	0.863658492	0.001246978	0		
	243	0.972114506	1.058193653	0.001527854	0		
	303	1.121001876	0.795769557	0.001148957	0		
	234	0.949470832	1.109268665	0.001601597	0		
	70	0.597353425	2.802446202	0.004046261	0		
	68	0.595694076	2.818080779	0.004068835	0		
	20	0.490893414	4.149784926	0.005991592	0		

Graph9

In the initial calculation, K = 5 nearest neighbours will be tested which means the nearest five neighbours will be chosen to predict the result. Then, the scaled weight will be multiplied by each corresponding nearest neighbour's area and sum all those values to produce the final prediction (**Graph10**).

AR	AS	AT	AU	AV	AW
	Rank	area	Scaled weight		
	1	24.59	0.013738196	0.337822248	
	2	2.87	0.012249104	0.035154927	
	3	0	0.011567492	0	
	4	0	0.010787298	0	
	5	0	0.008350668	0	
		pridiction			
		1.731201413			

Graph10

To evaluate the model, every data in the testing set will be tested by calling Excel index function and using the What-If Analysis section in the Data tab. (Graph11)

row#	1	2	3	4	5	6	7	8	9	10	11	12
2	-1.261009323	0.769105462	0.330194726	-0.29918	0.625528	-0.82172	-0.45023	-0.3626	0	0	0	0

Number	Area Predicted	area	
1	1.731201413	51.78	2504.88224
2	1.731201413	3.64	3.643512047
3	0.064652703	3.63	12.71170135
4	5.39142168	0	29.06742773
5	0.964210731	0	0.929702335
6	0.964210731	8.16	51.7793832
7	0.964210731	4.95	15.88651609
8	0.585276099	0	0.342548112
9	1.008553498	0	1.017180159
10	0.171080094	6.04	34.44422086
11	0.097249938	0	0.00945755
12	0.719341765	3.95	10.43715263
13	0.085024713	0	0.007229202
14	1.482478932	7.8	39.91107245
15	1.482478932	0	2.197743783
16	0.057922315	0	0.003354995
17	0.057922315	4.62	20.81255281
18	1.407363456	1.63	0.049567031
19	0.664054605	0	0.440968519
20	0.167213858	0	0.027960474
21	0.080097603	746.28	556814.2943

Graph11

To finish the testing, all variance error of testing data was summed to generate a total error of this model. (Graph12)

AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG
	103	0.106403905	14.68	212.3897031											
	104	0.075641006	40.54	1637.364349											
	105	0.040446938	10.82	116.1987642											
	106	0.17466237	0	0.030506943											
	107	0.132553436	0	0.017570413											
	108	0.062498015	0	0.003906002											
	109	0.062498015	1.95	3.562663742											
	110	0.062498015	49.59	2452.973453											
	111	0.071291522	5.8	32.81810082											
	112	0.098570279	0	0.0097161											
	113	0.669992144	0	0.448889474											
	114	2.446330354	0	5.984532203											
	115	0.082062002	2.17	4.359485082											
	116	0.406236877	0.43	0.000564686											
	117	0.094964211	0	0.009018201											
	118	0.162049727	6.44	39.41265963											
	119	0.131341556	54.29	2933.160285											
	120	0.002730696	11.16	124.4846583											
	121	1.731201413	0	2.997058331											
	122	1.731201413	0	2.997058331											
				erro											
				702036.6665											

Graph12

It is noticed that, in the study, the different values of K were tested. After that, the best value of K was chosen which is 5.

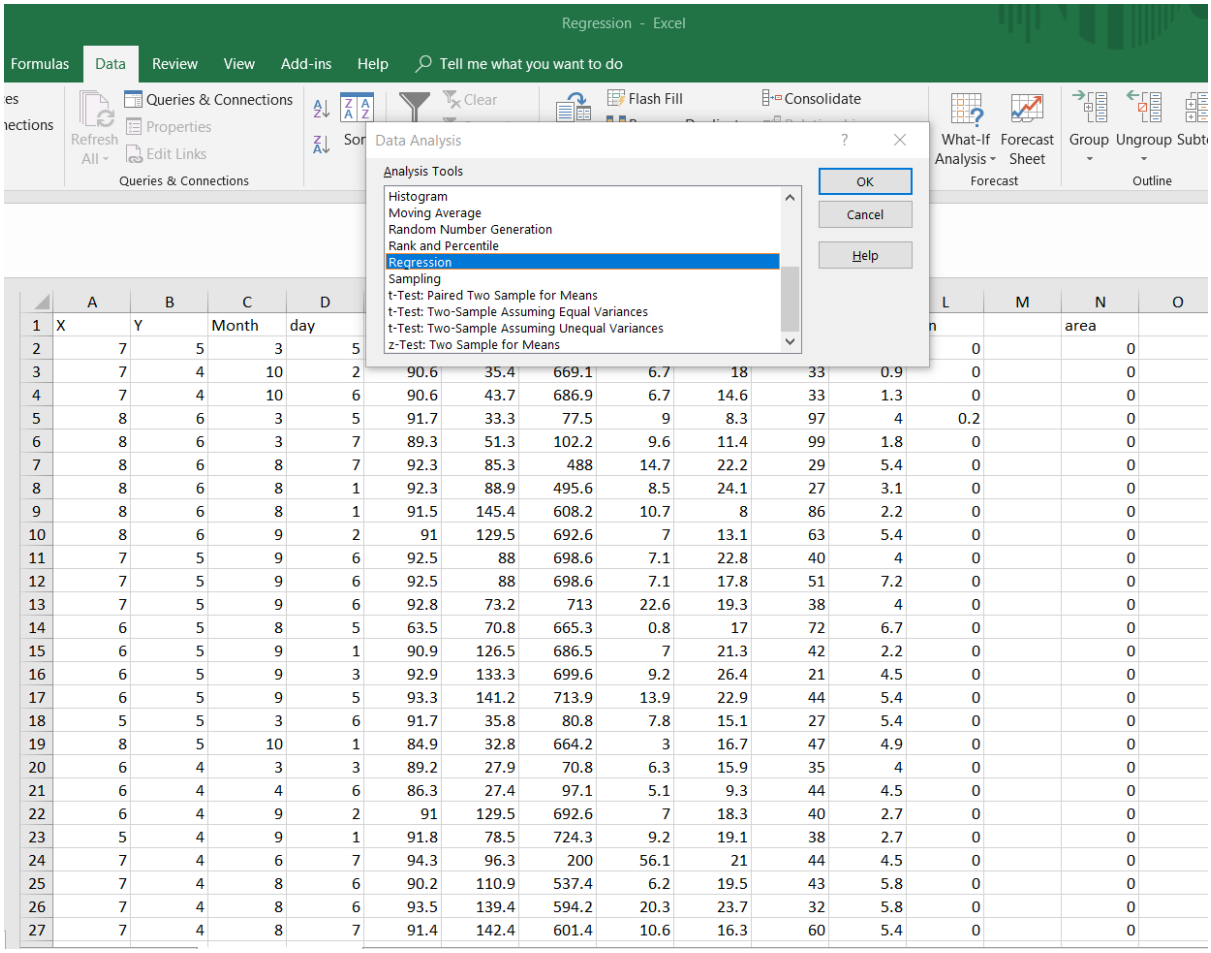
3.3 Linear -regression Algorithm

In data science, linear regression is a regression analysis that uses the least square function called a “linear regression equation” to model the relationship between one or more independent variables and dependent variables.

3.4 Implement Linear -regression Algorithm in the Excel

In the paper, we will use Excel and its built-in Excel regression tool to implement the Linear -regression on the forest fire data set. Then, read and compare the resulting parameters. Since fewer data was in the data set and normalisation will not significantly improve or change the prediction of Linear -regression, even if the weights in the training set were different (but make the computation and iterations more efficiency), in this section, the original data set which was not normalised has been used. However, the original data was divided into parts which were training set and testing set. The training set was taken 80% of total data while the testing data was taken 20% of total data.

Firstly, the regression coefficients were calculated for the least squares regression line by using the built-in Excel regression tool. To activate this, “Data” menu was clicked, and then the “Data Analysis” button on the right of the menu. This would open a list of techniques that user could apply. Then, “Regression” was Selected and clicked OK (Graph 11).



	A	B	C	D	L	M	N	O
1	X	Y	Month	day			area	
2	7	5	3	5	0		0	
3	7	4	10	2	0		0	
4	7	4	10	6	0		0	
5	8	6	3	5	0.2		0	
6	8	6	3	7	0		0	
7	8	6	8	7	0		0	
8	8	6	8	1	0		0	
9	8	6	8	1	0		0	
10	8	6	9	2	0		0	
11	7	5	9	6	0		0	
12	7	5	9	6	0		0	
13	7	5	9	6	0		0	
14	6	5	8	5	0		0	
15	6	5	9	1	0		0	
16	6	5	9	3	0		0	
17	6	5	9	5	0		0	
18	5	5	3	6	0		0	
19	8	5	10	1	0		0	
20	6	4	3	3	0		0	
21	6	4	4	6	0		0	
22	6	4	9	2	0		0	
23	5	4	9	1	0		0	
24	7	4	6	7	0		0	
25	7	4	8	6	0		0	
26	7	4	8	6	0		0	
27	7	4	8	7	0		0	

Graph 14

After selecting the input X which were the attributes of forest fire dataset and Y which was the burnt area, next, check the “Line Fit Plots” box which can get a plot of the predicted values from the linear regression, a new Worksheet was shown with the linear regression results in the Excel:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.161540036							
5	R Square	0.026095183							
6	Adjusted R Square	-0.004498685							
7	Standard Error	60.17696048							
8	Observations	395							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	12	37065.31485	3088.78	0.85295	0.595593595			
13	Residual	382	1383323.831	3621.27					
14	Total	394	1420389.146						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	-9.573035365	65.69539303	-0.1457	0.88422	-138.7428906	119.5968198	-138.7428906	119.5968198
18	X	1.237425815	1.603151928	0.77187	0.44067	-1.914681095	4.389532726	-1.914681095	4.389532726
19	Y	-0.401170124	3.030740101	-0.1324	0.89476	-6.360191632	5.557851384	-6.360191632	5.557851384
20	Month	3.150070948	3.056654734	1.03056	0.3034	-2.859903743	9.16004564	-2.859903743	9.16004564
21	day	1.788963872	1.488712263	1.20169	0.23023	-1.138132512	4.716060257	-1.138132512	4.716060257
22	FFMC	-0.003164562	0.653177777	-0.0048	0.99614	-1.287438467	1.281109342	-1.287438467	1.281109342
23	DMC	0.14945467	0.093089794	1.60549	0.10921	-0.033577879	0.332487219	-0.033577879	0.332487219
24	DC	-0.031588151	0.036361472	-0.8687	0.38554	-0.103081841	0.039905539	-0.103081841	0.039905539
25	ISI	-1.090043714	0.827016731	-1.318	0.18828	-2.716118641	0.536031212	-2.716118641	0.536031212
26	temp	0.440053831	1.005559985	0.43762	0.66191	-1.537071682	2.417179345	-1.537071682	2.417179345
27	RH	-0.246840108	0.279568498	-0.8829	0.37783	-0.796525873	0.302845657	-0.796525873	0.302845657
28	wind	0.234643347	1.899562549	0.12352	0.90176	-3.500264203	3.969550897	-3.500264203	3.969550897
29	rain	-15.70058266	60.0843895	-0.2613	0.794	-133.8381192	102.4369538	-133.8381192	102.4369538

According to the result showed above, we could see that forest fire data, **Multiple R,** was quite low which is 0.1615; therefore, the regression equation could not explain the variance of Y; The **R square** and **Adjusted R Square** were both extremely low which was 0.1615 that means the attributes X could not explain burnt area Y; and **Significance F** was 0.59 which was much larger than 0.05, so that, the regression equation failed in the F test and the overall regression equation was non-significantly effective.

4. The conclusion of the Supervised Algorithm with forest fire data set

In conclusion, the two algorithms, Linear-Regression and KNN, were applied to the forest fire data set. According to the results, it could be noticed that there was a non-Linear relation between the data in the forest data set. Furthermore, the linear regression was not a proper algorithm to model the data. By contrast, the KNN could predict the forest fire relatively good because of the similarity of the cause of forest fire. However, in the study, there was just a preliminary attempt. If the user considered using the KNN model to predict forest fire in the real world, training data needs to be processed and weighted correctly and perfect the algorithm.

5. Dataset of New Wine.

5.1 Introduction of data set

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The variables as follow: Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Nonflavonoid phenols, Proanthocyanins, Colour intensity, Hue, OD280/OD315 of diluted wines, and Proline (The attributes are donated by Riccardo Leardi, riclea '@' anthem.unige.it).

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Region	1 Alcohol	2 Malic ac	3 Ash	4 Alcalinit	5 Magnes	6 Total ph	7 Flavanoi	8 Nonflav	9 Proantho	10 Color int	11 Hue	12 OD280	13 Proline
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
1	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
1	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
1	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
1	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150
1	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
1	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
1	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
1	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130
1	14.19	1.59	2.48	16.5	108	3.3	3.93	0.32	1.86	8.7	1.23	2.82	1680
1	13.64	3.1	2.56	15.2	116	2.7	3.03	0.17	1.66	5.1	0.96	3.36	845
1	14.06	1.63	2.28	16	126	3	3.17	0.24	2.1	5.65	1.09	3.71	780
1	12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770
1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.8	1.11	4	1035
1	12.85	1.6	2.52	17.8	95	2.48	2.37	0.26	1.46	3.93	1.09	3.63	1015
1	13.5	1.81	2.61	20	96	2.53	2.61	0.28	1.66	3.52	1.12	3.82	845

Graph13

5.2 Standardisation

The reason as paper mentioned in section "2.2 Normalisation**"**, the original data needs to be standardised to eliminate adverse effects which caused by the different magnitude of units of data. After standardisation, the data will be a normal distribution. The mean value will be 0, and the standard deviation will be 1. The function is written as follow：

$$x^* = \frac{x - \mu}{\sigma}$$

The **μ** means the average of all data and the **σ** is **the** standard deviation of all data.

After the original data set of new wine has been applied the Standardization. It shows as follow.

B	C	D	E	F	G	H	I	J	K	L	M	N	O
Region	1 Alcohol	2 Malic ac	3 Ash	4 Alcalinit	5 Magnesi	6 Total ph	7 Flavanoi	8 Nonflav	9 Proanth	10 Color ir	11 Hue	12 OD280,	13 Proline
1	1.514340767	-0.56067	0.2314	-1.1663	1.908522	0.806722	1.031908	-0.65771	1.221438	0.251009	0.361158	1.842721	1.010159
1	0.245596828	-0.49801	-0.82567	-2.48384	0.018094	0.567048	0.731565	-0.81841	-0.54319	-0.2925	0.404908	1.110317	0.962526
1	0.196325219	0.021172	1.106214	-0.26798	0.08811	0.806722	1.212114	-0.49701	2.129959	0.268263	0.317409	0.786369	1.391224
1	1.6867914	-0.34584	0.486554	-0.80697	0.9283	2.484437	1.462399	-0.97911	1.029251	1.182732	-0.42634	1.180741	2.328007
1	0.294868437	0.227053	1.835226	0.450674	1.278379	0.806722	0.661485	0.226158	0.400275	-0.31838	0.361158	0.448336	-0.03777
1	1.47738706	-0.51591	0.304301	-1.28608	0.858284	1.557699	1.362285	-0.1756	0.662349	0.729811	0.404908	0.335659	2.232741
1	1.711427204	-0.41745	0.304301	-1.46574	-0.26197	0.327374	0.491291	-0.49701	0.67982	0.082781	0.273659	1.363842	1.724655
1	1.304936428	-0.16681	0.88751	-0.56742	1.488427	0.487157	0.48128	-0.41665	-0.5956	-0.00349	0.448658	1.363842	1.740533
1	2.253414907	-0.62333	-0.71632	-1.64541	-0.19195	0.806722	0.951817	-0.57736	0.67982	0.061213	0.536158	0.335659	0.946649
1	1.058578381	-0.88292	-0.35181	-1.04653	-0.12194	1.09433	1.122011	-1.13982	0.45269	0.932547	0.229909	1.321588	0.946649
1	1.354208037	-0.15786	-0.24246	-0.44765	0.368173	1.046395	1.292205	-1.13982	1.378682	0.298458	1.279908	0.786369	2.423273
1	1.378843842	-0.76655	-0.16956	-0.80697	-0.33199	-0.15197	0.401188	-0.81841	-0.03651	-0.02506	0.929908	0.293405	1.6929
1	0.923081456	-0.54277	0.158499	-1.04653	-0.75208	0.487157	0.731565	-0.57736	0.382804	0.233755	0.842408	0.406082	1.819921
1	2.154871688	-0.54277	0.085597	-2.42395	-0.61205	1.286069	1.662628	0.547563	2.129959	0.147484	1.279908	0.166643	1.28008
1	1.699109302	-0.41745	0.049147	-2.24429	0.158126	1.605634	1.612571	-0.57736	2.392033	1.053326	1.061158	0.546929	2.540768
1	0.775266628	-0.47115	1.215566	-0.6872	0.858284	0.886613	0.881737	-0.49701	-0.2287	0.967055	1.411158	0.377913	1.788166
1	1.600566084	-0.37269	1.288467	0.151234	1.418411	0.806722	1.111999	-0.25595	0.662349	0.492567	0.492408	0.053965	1.6929
1	1.021624674	-0.68599	0.923961	0.151234	1.068331	1.046395	1.372297	0.306509	0.22556	0.665108	0.754908	-0.05871	1.216569
1	1.465069158	-0.66808	0.413653	-0.89681	0.578221	1.605634	1.902902	-0.3363	0.470162	1.57095	1.192408	0.293405	2.963114
1	0.78758453	0.683574	0.705257	-1.28608	1.138347	0.646939	1.001874	-1.54157	0.12073	0.018078	0.011159	1.053978	0.311541
1	1.304936428	-0.63228	-0.31536	-1.04653	1.838506	1.126287	1.142034	-0.97911	0.889479	0.255322	0.579908	1.546943	0.105132
1	-0.086986535	1.31017	1.033313	-0.26798	0.158126	0.18357	0.381165	-0.89876	0.67982	-0.24073	0.317409	1.279334	0.073376
1	0.873809846	-0.4264	-0.02375	-0.86686	0.08811	0.503135	0.851702	-0.73806	0.173145	-0.54268	0.667408	1.955399	0.914893
1	-0.185529754	-0.65913	0.559455	-0.50753	-0.33199	0.295418	0.34112	-0.81841	-0.2287	-0.48661	0.579908	1.434265	0.851383
1	0.615133898	-0.47115	0.88751	0.151234	-0.26197	0.375309	0.581394	-0.65771	0.12073	-0.66346	0.711158	1.701875	0.311541
1	0.060828293	-0.25632	3.110996	1.648436	1.698474	0.535092	0.651474	0.868969	0.574991	-0.63758	0.754908	0.828623	0.263908

6. Unsupervised Algorithm implement.

6.1 K-means Algorithm

The k-means algorithm is a simple iterative clustering algorithm which uses distance as the similarity index to find k classes in a data set. Each class is described by the clustering centre which is according to the mean of all values in the class.

Usually, for a given data set X, it contains n value of d dimensional data points. After the number of class K determined, the Euclidean distance is selected as the similarity index to implement the algorithm. The target of this algorithm is to minimise the sum of squares of all kinds of clustering.

6.2 Implement K-Means Algorithm in the Excel

In the paper, the Excel and its built-in function will be used. It will be quite simple to implement the K-means algorithm. First, New wine data set that has been standardised was copied and pasted them into a worksheet which named K1(**Graph15**).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	k Means													
	*Select this as your source table (incl row and column titles)													
	Region	**1 Alcohol**	**2 Malic ac**	**3 Ash**	**4 Alcalinit**	**5 Magnesi**	**6 Total ph**	**7 Flavanoi**	**8 Nonflav**	**9 Proanth**	**10 Color ir**	**11 Hue**	**12 OD280,**	**13 Proline**
	1	1.514340767	-0.56067	0.2314	-1.1663	1.908522	0.806722	1.031908	-0.65771	1.221438	0.251009	0.361158	1.842721	1.010159
	1	0.245596828	-0.49801	-0.82567	-2.48384	0.018094	0.567048	0.731565	-0.81841	-0.54319	-0.2925	0.404908	1.110317	0.962526
	1	0.196325219	0.021172	1.106214	-0.26798	0.08811	0.806722	1.212114	-0.49701	2.129959	0.268263	0.317409	0.786369	1.391224
	1	1.6867914	-0.34584	0.486554	-0.80697	0.9283	2.484437	1.462399	-0.97911	1.029251	1.182732	-0.42634	1.180741	2.328007
	1	0.294868437	0.227053	1.835226	0.450674	1.278379	0.806722	0.661485	0.226158	0.400275	-0.31838	0.361158	0.448336	-0.03777
	1	1.47738706	-0.51591	0.304301	-1.28608	0.858284	1.557699	1.362285	-0.1756	0.662349	0.729811	0.404908	0.335659	2.232741
	1	1.711427204	-0.41745	0.304301	-1.46574	-0.26197	0.327374	0.491291	-0.49701	0.67982	0.082781	0.273659	1.363842	1.724655
	1	1.304936428	-0.16681	0.88751	-0.56742	1.488427	0.487157	0.48128	-0.41665	-0.5956	-0.00349	0.448658	1.363842	1.740533
	1	2.253414907	-0.62333	-0.71632	-1.64541	-0.19195	0.806722	0.951817	-0.57736	0.67982	0.061213	0.536158	0.335659	0.946649
	1	1.058578381	-0.88292	-0.35181	-1.04653	-0.12194	1.09433	1.122011	-1.13982	0.45269	0.932547	0.229909	1.321588	0.946649
	1	1.354208037	-0.15786	-0.24246	-0.44765	0.368173	1.046395	1.292205	-1.13982	1.378682	0.298458	1.279908	0.786369	2.423273
	1	1.378843842	-0.76655	-0.16956	-0.80697	-0.33199	-0.15197	0.401188	-0.81841	-0.03651	-0.02506	0.929908	0.293405	1.6929
	1	0.923081456	-0.54277	0.158499	-1.04653	-0.75208	0.487157	0.731565	-0.57736	0.382804	0.233755	0.842408	0.406082	1.819921
	1	2.154871688	-0.54277	0.085597	-2.42395	-0.61205	1.286069	1.662628	0.547563	2.129959	0.147484	1.279908	0.166643	1.28008
	1	1.699109302	-0.41745	0.049147	-2.24429	0.158126	1.605634	1.612571	-0.57736	2.392033	1.053326	1.061158	0.546929	2.540768
	1	0.775266628	-0.47115	1.215566	-0.6872	0.858284	0.886613	0.881737	-0.49701	-0.2287	0.967055	1.411158	0.377913	1.788166
	1	1.600566084	-0.37269	1.288467	0.151234	1.418411	0.806722	1.111999	-0.25595	0.662349	0.492567	0.492408	0.053965	1.6929
	1	1.021624674	-0.68599	0.923961	0.151234	1.068331	1.046395	1.372297	0.306509	0.22556	0.665108	0.754908	-0.05871	1.216569
	1	1.465069158	-0.66808	0.413653	-0.89681	0.578221	1.605634	1.902902	-0.3363	0.470162	1.57095	1.192408	0.293405	2.963114
	1	0.78758453	0.683574	0.705257	-1.28608	1.138347	0.646939	1.001874	-1.54157	0.12073	0.018078	0.011159	1.053978	0.311541
	1	1.304936428	-0.63228	-0.31536	-1.04653	1.838506	1.126287	1.142034	-0.97911	0.889479	0.255322	0.579908	1.546943	0.105132
	1	-0.086986535	1.31017	1.033313	-0.26798	0.158126	0.18357	0.381165	-0.89876	0.67982	-0.24073	0.317409	1.279334	0.073376
	1	0.873809846	-0.4264	-0.02375	-0.86686	0.08811	0.503135	0.851702	-0.73806	0.173145	-0.54268	0.667408	1.955399	0.914893

Graph15

Then, click on the *k*-means button, and select the full range of data except “Region”. In the initial, this study will specify K= 1 cluster. Clicking OK gives us a new tab called “Cluster Analysis K=1” that appears immediately to the left of the K1(**Graph16, Graph17**).

	A	B	C
1	**Row Title**	**Centroid**	
2	1	1	
3	1	1	
4	1	1	
5	1	1	
6	1	1	
7	1	1	
8	1	1	
9	1	1	
10	1	1	
11	1	1	
12	1	1	
13	1	1	
14	1	1	
15	1	1	
16	1	1	
17	1	1	
18	1	1	
19	1	1	
20	1	1	
21	1	1	
22	1	1	
23	1	1	
24	1	1	
25	1	1	
26	1	1	
27	1	1	

Cluster Analysis K=1 | K=1

Graph16

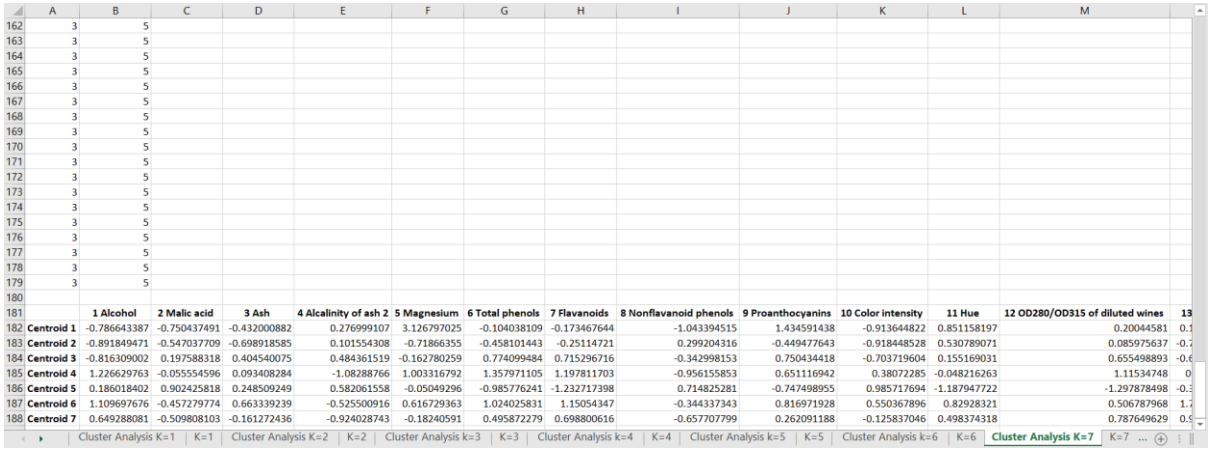
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
166	3	1												
167	3	1												
168	3	1												
169	3	1												
170	3	1												
171	3	1												
172	3	1												
173	3	1												
174	3	1												
175	3	1												
176	3	1												
177	3	1												
178	3	1												
179	3	1												
180														
181		1 Alcohol	2 Malic acid	3 Ash	4 Alcalinity of ash 2	5 Magnesium	6 Total phenols	7 Flavanoids	8 Nonflavanoid phenols	9 Proanthocyanins	10 Color intensity	11 Hue	280/OD315 of diluted	13 Proline
182	Centroid 1	7.82395E-15	2.94396E-16	-3.97809E-15	-1.83374E-16	-7.42228E-17	1.23497E-16	1.09401E-15	-7.06052E-16	-1.71866E-15	-2.95644E-16	1.8749E-15	2.1456E-15	-5.42637E-17
183														
184														
185														
186														
187														

Graph17

Due to the multiple attributes of X, this study could not be plotted to show the data and the centroid position of the cluster. However, the algorithm should be applied multiple times until there is no change to the cluster assignments (or some other stopping condition is met).

In the experiment, we supposed each time clicking the *k*-means button, the function will randomly pick up starting points of k.

After the converging of K=1 have been calculated, the same processing was applied on the dataset multiple times to find out the converging of the K=2, K=3, K=4, K=5, K=6, and K=7(**Graph18**).

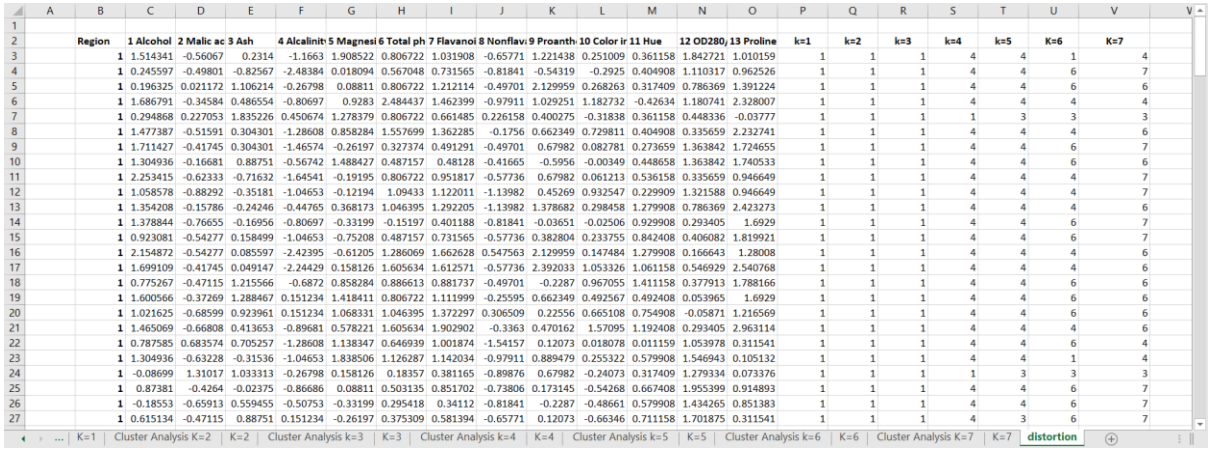


		1 Alcohol	2 Malic acid	3 Ash	4 Alcalinity of ash 2	5 Magnesium	6 Total phenols	7 Flavanoids	8 Nonflavanoid phenols	9 Proanthocyanins	10 Color intensity	11 Hue	12 OD280/OD315 of diluted wines	13
182	Centroid 1	-0.786643387	-0.750437491	-0.432000882	0.276999107	3.126797025	-0.104038109	-0.173467644	-1.043394515	1.434591438	-0.913644822	0.851158197	0.20044581	0.1
183	Centroid 2	-0.891849471	-0.547037709	-0.698918585	0.101554308	-0.71866355	-0.458101443	-0.25114721	0.299204316	-0.449477643	-0.918448528	0.530789071	0.085975637	-0.1
184	Centroid 3	-0.816309002	0.197588318	0.404540075	0.484361519	-0.162780259	0.774099484	0.715296716	-0.342998153	0.750434418	-0.703719604	0.155169031	0.655498893	-0.6
185	Centroid 4	1.226629763	-0.055554596	0.093408284	-1.08288766	1.003316792	1.357971105	1.197811703	-0.956155853	0.651116942	0.38072285	-0.048216263	1.11534748	0
186	Centroid 5	0.186018402	0.902425818	0.248509249	0.582061558	-0.05049296	-0.985776241	-1.232717398	0.714825281	-0.747498955	0.985717694	-1.187947722	-1.297878498	-0.3
187	Centroid 6	1.109697676	-0.457279774	0.663339239	-0.525500916	0.616729363	1.024025831	1.15054347	-0.344337343	0.816971928	0.550367896	0.82928321	0.506787968	1.1
188	Centroid 7	0.649288081	-0.509808103	-0.161272436	-0.924028743	-0.18240591	0.495872279	0.698800616	-0.657707799	0.262091188	-0.125837046	0.498374318	0.787649629	0.9

Cluster Analysis K=1 | K=1 | Cluster Analysis K=2 | K=2 | Cluster Analysis k=3 | K=3 | Cluster Analysis k=4 | K=4 | Cluster Analysis k=5 | K=5 | Cluster Analysis k=6 | K=6 | **Cluster Analysis K=7** | K=7

Graph18

In order to calculate and compare distortions of those seven numbers of K, a new worksheet named “Cluster Distortions” was inserted to hold our cluster distortion calculations. Then, the centroids of every value of K were posted beside the “New Wine” data that was used in the study (Graph 19).

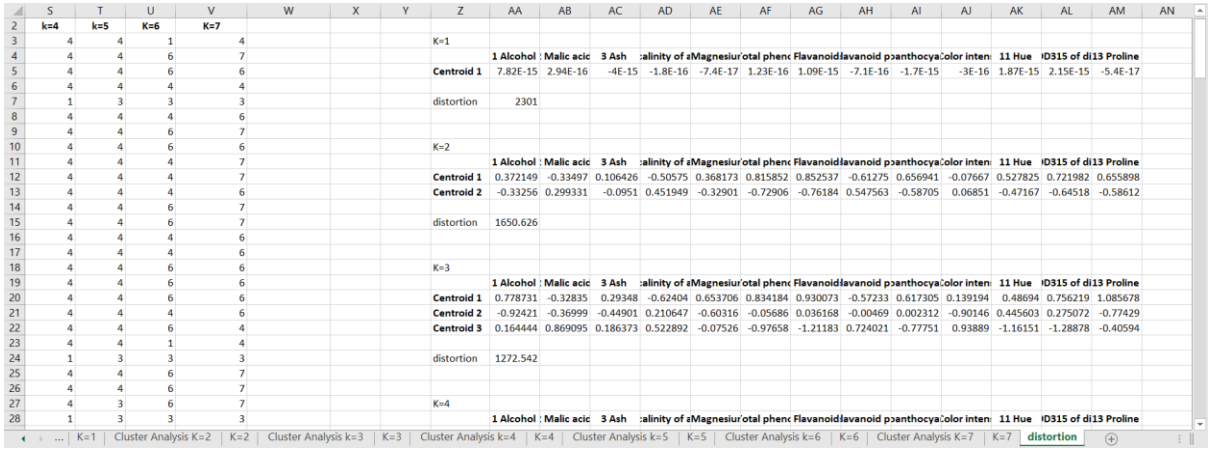


	Region	1 Alcohol	2 Malic ac	3 Ash	4 Alcalinit	5 Magnesi	6 Total ph	7 Flavanoi	8 Nonflavi	9 Proanth	10 Color ir	11 Hue	12 OD280/	13 Proline	k=1	k=2	k=3	k=4	k=5	K=6	K=7
3	1	1.514341	-0.56067	0.2314	-1.1663	1.908522	0.806722	1.031908	-0.65771	1.221438	0.251009	0.361158	1.842721	1.010159	1	1	1	4	4	1	4
4	1	0.245597	-0.49801	-0.82567	-2.48384	0.018094	0.567048	0.731565	-0.81841	-0.54319	-0.2925	0.404908	1.110317	0.962526	1	1	1	4	4	6	7
5	1	0.196325	0.021172	1.106214	-0.26798	0.08811	0.806722	1.212114	-0.49701	2.129959	0.268263	0.317409	0.786369	1.391224	1	1	1	4	4	6	6
6	1	1.686791	-0.34584	0.486554	-0.80697	0.9283	2.484437	1.462399	-0.97911	1.029251	1.182732	-0.42634	1.180741	2.328007	1	1	1	4	4	4	4
7	1	0.294868	0.227053	1.835226	0.450674	1.278379	0.806722	0.661485	0.226158	0.400275	-0.31838	0.361158	0.448336	-0.03777	1	1	1	1	3	3	3
8	1	1.477387	-0.51591	0.304301	-1.28608	0.858284	1.557699	1.362285	-0.1756	0.662349	0.729811	0.404908	0.335659	2.232741	1	1	1	4	4	4	6
9	1	1.711427	-0.41745	0.304301	-1.46574	-0.26197	0.327374	0.491291	-0.49701	0.67982	0.082781	0.273659	1.363842	1.724655	1	1	1	4	4	6	7
10	1	1.304936	-0.16681	0.88751	-0.56742	1.488427	0.487157	0.48128	-0.41665	-0.5956	-0.00349	0.448658	1.363842	1.740533	1	1	1	4	4	6	6
11	1	2.253415	-0.62333	-0.71632	-1.64541	-0.19195	0.806722	0.951817	-0.57736	0.67982	0.061213	0.536158	0.335659	0.946649	1	1	1	4	4	4	7
12	1	1.058578	-0.88292	-0.35181	-1.04653	-0.12194	1.09433	1.122011	-1.13982	0.45269	0.932547	0.229909	1.321588	0.946649	1	1	1	4	4	4	7
13	1	1.354208	-0.15786	-0.24246	-0.44765	0.368173	1.046395	1.292205	-1.13982	1.378682	0.298458	1.279908	0.786369	2.423273	1	1	1	4	4	4	6
14	1	1.378844	-0.76655	-0.16956	-0.80697	-0.33199	-0.15197	0.401188	-0.81841	-0.03651	-0.02506	0.929908	0.293405	1.6929	1	1	1	4	4	6	7
15	1	0.923081	-0.54277	0.158499	-1.04653	-0.75208	0.487157	0.731565	-0.57736	0.382804	0.233755	0.842408	0.406082	1.819921	1	1	1	4	4	6	7
16	1	2.154872	-0.54277	0.085597	-2.42395	-0.61205	1.286069	1.662628	0.547563	2.129959	0.147484	1.279908	0.166643	1.28008	1	1	1	4	4	4	6
17	1	1.699109	-0.41745	0.049147	-2.24429	0.158126	1.605634	1.612571	-0.57736	2.392033	1.053326	1.061158	0.546929	2.540768	1	1	1	4	4	4	6
18	1	0.775267	-0.47115	1.215566	-0.6872	0.858284	0.886613	0.881737	-0.49701	-0.2287	0.967055	1.411158	0.377913	1.788166	1	1	1	4	4	6	6
19	1	1.600566	-0.37269	1.288467	0.151234	1.418411	0.806722	1.111999	-0.25595	0.662349	0.492567	0.492408	0.053965	1.6929	1	1	1	4	4	6	6
20	1	1.021625	-0.68599	0.923961	0.151234	1.068331	1.046395	1.372297	0.306509	0.22556	0.665108	0.754908	-0.05871	1.216569	1	1	1	4	4	6	6
21	1	1.465069	-0.66808	0.413653	-0.89681	0.578221	1.605634	1.902902	-0.3363	0.470162	1.57095	1.192408	0.293405	2.963114	1	1	1	4	4	4	6
22	1	0.787585	0.683574	0.705257	-1.28608	1.138347	0.646939	1.001874	-1.54157	0.12073	0.018078	0.011159	1.053978	0.311541	1	1	1	4	4	6	4
23	1	1.304936	-0.63228	-0.31536	-1.04653	1.838506	1.126287	1.142034	-0.97911	0.889479	0.255322	0.579908	1.546943	0.105132	1	1	1	4	4	1	4
24	1	-0.08699	1.31017	1.033313	-0.26798	0.158126	0.18357	0.381165	-0.89876	0.67982	-0.24073	0.317409	1.279334	0.073376	1	1	1	1	3	3	3
25	1	0.87381	-0.4264	-0.02375	-0.86686	0.08811	0.503135	0.851702	-0.73806	0.173145	-0.54268	0.667408	1.955399	0.914893	1	1	1	4	4	6	7
26	1	-0.18553	-0.65913	0.559455	-0.50753	-0.33199	0.295418	0.34112	-0.81841	-0.2287	-0.48661	0.579908	1.434265	0.851383	1	1	1	4	4	6	7
27	1	0.615134	-0.47115	0.88751	0.151234	-0.26197	0.375309	0.581394	-0.65771	0.12073	-0.66346	0.711158	1.701875	0.311541	1	1	1	4	3	6	7

... | K=1 | Cluster Analysis K=2 | K=2 | Cluster Analysis k=3 | K=3 | Cluster Analysis k=4 | K=4 | Cluster Analysis k=5 | K=5 | Cluster Analysis k=6 | K=6 | Cluster Analysis K=7 | K=7 | **distortion**

Graph 19

In this step, the final converging of each value of K were pasted on the "Cluster Distortions" as well (**Graph 20**).



	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
2	k=4	k=5	K=6	K=7																		
3	4	4	1	4				K=1														
4	4	4	6	7					1 Alcohol	Malic acic	3 Ash	alinity of a	Magnesiur	otal phenc	Flavanoid	lavanoid p	anthocya	olor inten	11 Hue	D315 of di	13 Proline	
5	4	4	6	6				Centroid 1	7.82E-15	2.94E-16	-4E-15	-1.8E-16	-7.4E-17	1.23E-16	1.09E-15	-7.1E-16	-1.7E-15	-3E-16	1.87E-15	2.15E-15	-5.4E-17	
6	4	4	4	4																		
7	1	3	3	3				distortion	2301													
8	4	4	4	6																		
9	4	4	6	7																		
10	4	4	6	6				K=2														
11	4	4	4	7					1 Alcohol	Malic acic	3 Ash	alinity of a	Magnesiur	otal phenc	Flavanoid	lavanoid p	anthocya	olor inten	11 Hue	D315 of di	13 Proline	
12	4	4	4	7				Centroid 1	0.372149	-0.33497	0.106426	-0.50575	0.368173	0.815852	0.852537	-0.61275	0.656941	-0.07667	0.527825	0.721982	0.655898	
13	4	4	4	6				Centroid 2	-0.33256	0.299331	-0.0951	0.451949	-0.32901	-0.72906	-0.76184	0.547563	-0.58705	0.06851	-0.47167	-0.64518	-0.58612	
14	4	4	6	7																		
15	4	4	6	7				distortion	1650.626													
16	4	4	4	6																		
17	4	4	4	6																		
18	4	4	6	6				K=3														
19	4	4	6	6					1 Alcohol	Malic acic	3 Ash	alinity of a	Magnesiur	otal phenc	Flavanoid	lavanoid p	anthocya	olor inten	11 Hue	D315 of di	13 Proline	
20	4	4	6	6				Centroid 1	0.778731	-0.32835	0.29348	-0.62404	0.653706	0.834184	0.930073	-0.57233	0.617305	0.139194	0.48694	0.756219	1.085678	
21	4	4	4	6				Centroid 2	-0.92421	-0.36999	-0.44901	0.210647	-0.60316	-0.05686	0.036168	-0.00469	0.002312	-0.90146	0.445603	0.275072	-0.77429	
22	4	4	6	4				Centroid 3	0.164444	0.869095	0.186373	0.522892	-0.07526	-0.97658	-1.21183	0.724021	-0.77751	0.93889	-1.16151	-1.28878	-0.40594	
23	4	4	1	4																		
24	1	3	3	3				distortion	1272.542													
25	4	4	6	7																		
26	4	4	6	7																		
27	4	3	6	7				K=4														
28	1	3	3	3					1 Alcohol	Malic acic	3 Ash	alinity of a	Magnesiur	otal phenc	Flavanoid	lavanoid p	anthocya	olor inten	11 Hue	D315 of di	13 Proline	

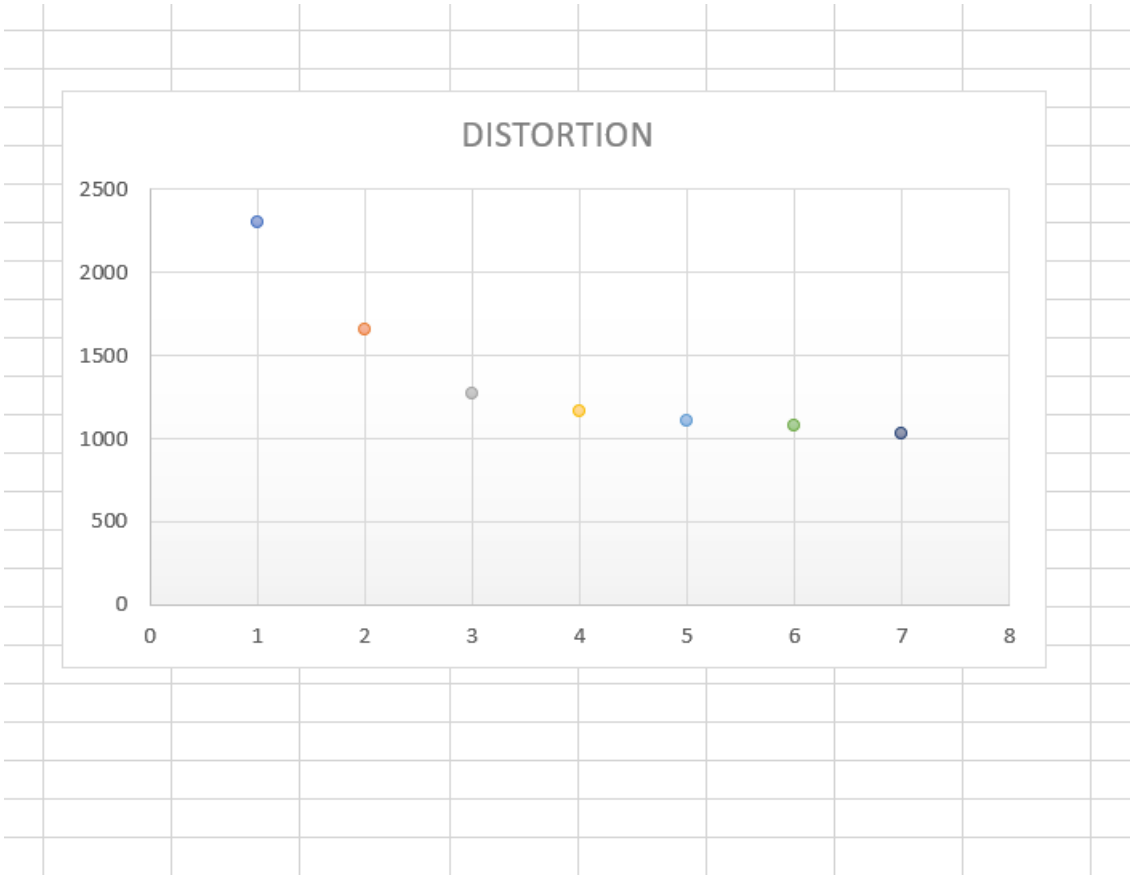
... | K=1 | Cluster Analysis K=2 | K=2 | Cluster Analysis k=3 | K=3 | Cluster Analysis k=4 | K=4 | Cluster Analysis k=5 | K=5 | Cluster Analysis k=6 | K=6 | Cluster Analysis K=7 | K=7 | distortion

Graph 20

Finally, the distortion of each value of K calculated by an equation which is:

$$\sum\left(Data\,X_i\ -\ Centroid\,X\right)^2+\left(Data\,Y_i\ -\ Centroid\,Y\right)^2\ldots$$

Moreover, each value of distortion was plotted on the scatter graph in order to find out the best number of K (**Graph 21**).



Graph 21

According to the graph above, it can be noticed that the distortion really decreased until k = 3, then it decreased by only a small amount for k = 4, k = 5, k=6, and k=7. This motivated k = 3 as the correct choice and this method of determining K is called the “elbow method”.

7. The conclusion of Unsupervised Algorithm with New wine data set

In this paper, the k-Mean algorithm was used to establish a wine classification model to attempt to classify the wine by analysing the chemical composition of wines from different regions of Italy.

The result of our preliminary attempts showed that the unsupervised k-Mean algorithm is a powerful method for the wine classification. When the k-Means algorithm optimally converged, that is, when k=3, the wine from the third production region and first production region have been 100% correctly classified. Although the wine from the second production region cannot be 100% correctly classified, its accuracy still can reach 1-7/71 = 90.14%. In

the study, the paper also evaluates the effect of standardisation on the original data and different converging of k on the accuracy of the model. It reveals that standardisation and the proper amount of k can improve the accuracy of the model. This may provide a direction to further improve the accuracy of the model.

Reference:

Real Statistics Using Excel 2014, *Principal Component Analysis*, USA, viewed 30 October 2018, <http://www.real-statistics.com/multivariate-statistics/factor-analysis/principal-component-analysis/>

Rohith Gandhi 9 Jun 2018, *K-Means Clustering—Introduction to Machine Learning Algorithms*, USA, viewed 31 October 2018, <https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a>

Wikipedia 23 October 2018, *K-Means Clustering—k-means clustering*, USA, viewed 30 October 2018, <https://en.wikipedia.org/wiki/K-means_clustering>

Firdaouss Doukkali July 2018, *Clustering Using K-means*, USA, viewed 31 October 2018, <https://www.kdnuggets.com/2018/07/clustering-using-k-means-algorithm.html >

SEBASTIAN MONCADA 8 August 2018, *MULTIPLE LINEAR REGRESSION*, USA, viewed 29 October 2018, < https://www.superdatascience.com/regression-classification-multiple-linear-regression/>

Manu Jeevan Jan 2015, *Fundamental methods of Data Science: Classification, Regression And Similarity*, USA, viewed 31 October 2018, <https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a>

Devin Soni 13 Mar 2018, *Introduction to k-Nearest-Neighbors*, USA, viewed 31 October 2018, < https://towardsdatascience.com/introduction-to-k-nearest-neighbors-3b534bb11d26>

Wikipedia 24 October 2018, *k-nearest neighbours algorithm*, USA, viewed 31 October 2018, < https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>

Paulo Cortez January 2007, *A Data Mining Approach to Predict Forest Fires using Meteorological Data*, USA, viewed 31 October 2018, <https://www.researchgate.net/publication/238767143_A_Data_Mining_Approach_to_Predict_Forest_Fires_using_Meteorological_Data

Machine Learning 2012, video recording, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

UCI 2005, *Wine Data Set*, USA, viewed 30 October 2018, < https://archive.ics.uci.edu/ml/datasets/Wine>