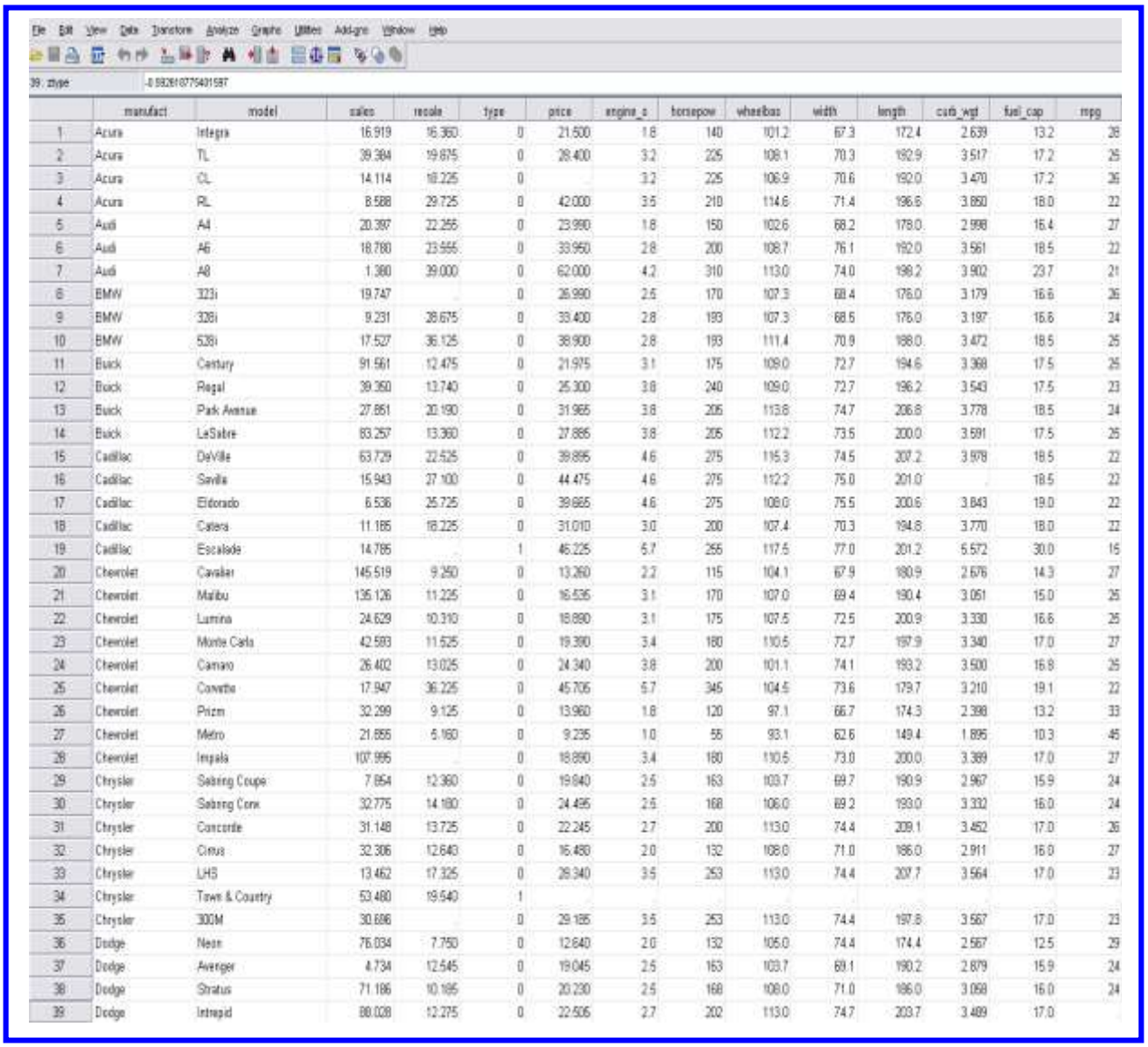
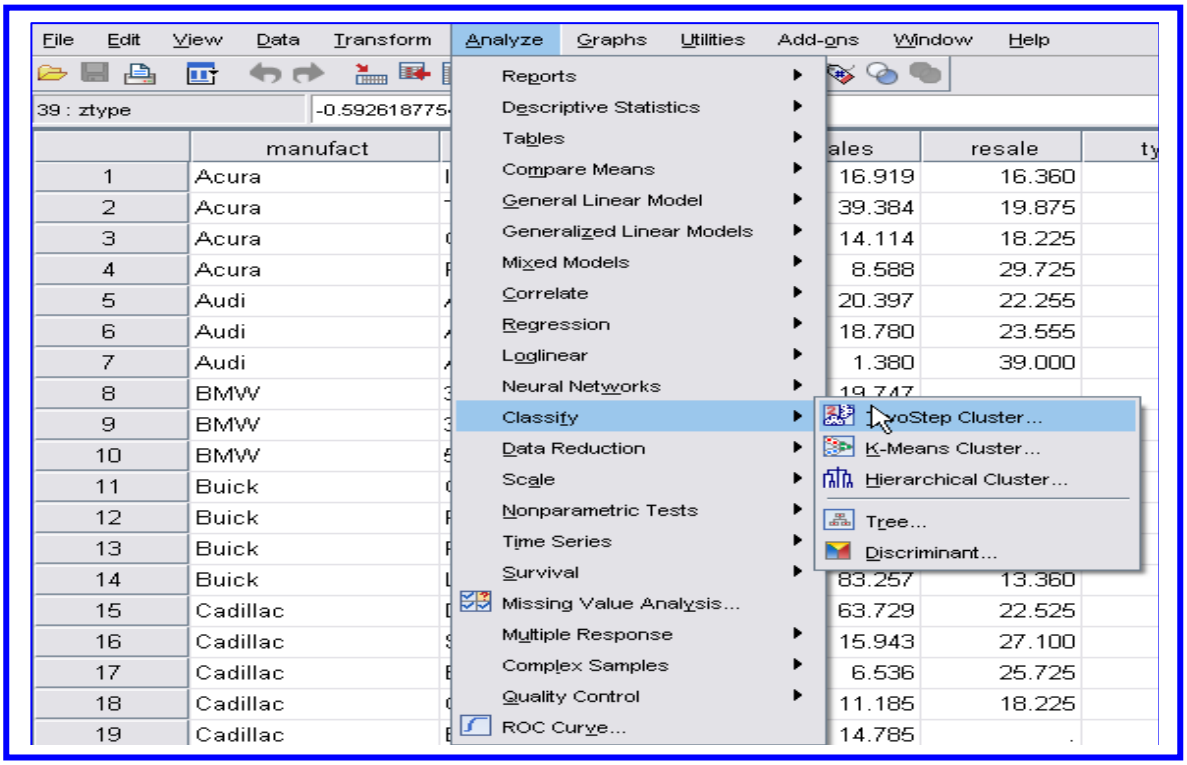
Cluster Analysis Workshop#3- Two Stages Clustering Method

ภาพด้านล่างแสดงให้เห็นข้อมูลเกี่ยวกับยอดขายรถยนตร์ประเภทต่างๆ พร้อมกับข้อมูลอื่นๆที่เกี่ยวกับตัวรถ

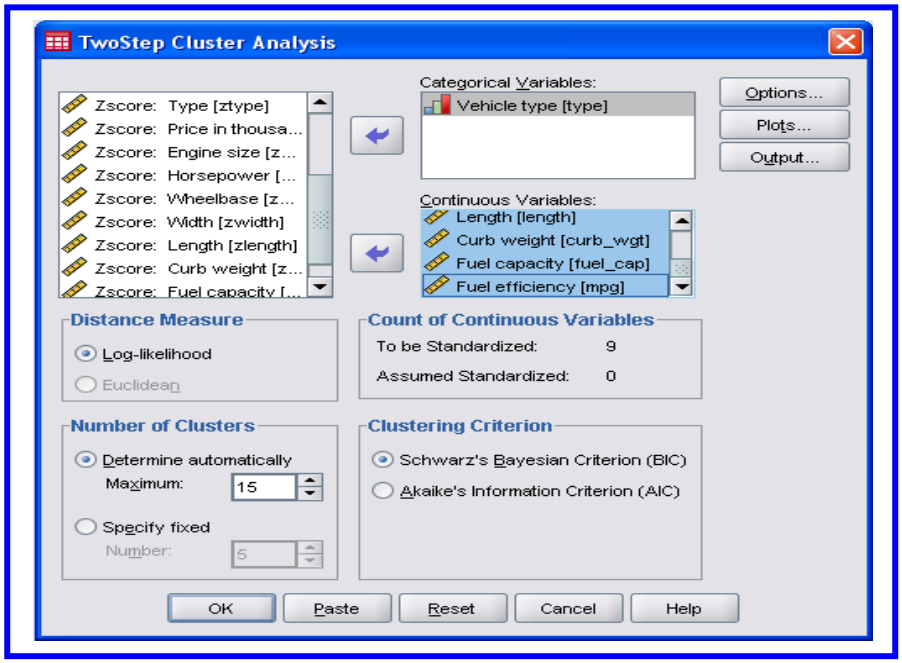


	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg
1	Acura	Integra	16.919	16.360	0	21.500	1.8	140	101.2	67.3	172.4	2.639	13.2	28
2	Acura	TL	39.384	19.875	0	28.400	3.2	225	108.1	70.3	192.9	3.517	17.2	25
3	Acura	CL	14.114	18.225	0		3.2	225	106.9	70.6	192.0	3.470	17.2	26
4	Acura	RL	8.588	29.725	0	42.000	3.5	210	114.6	71.4	196.6	3.850	18.0	22
5	Audi	A4	20.397	22.255	0	23.990	1.8	150	102.6	68.2	178.0	2.998	16.4	27
6	Audi	A6	18.780	23.555	0	33.950	2.8	200	108.7	76.1	192.0	3.561	18.5	22
7	Audi	A8	1.380	39.000	0	62.000	4.2	310	113.0	74.0	198.2	3.902	23.7	21
8	BMW	323i	19.747		0	26.990	2.5	170	107.3	68.4	176.0	3.179	16.6	26
9	BMW	328i	9.231	28.675	0	33.400	2.8	193	107.3	68.5	176.0	3.197	16.6	24
10	BMW	528i	17.527	36.125	0	38.900	2.8	193	111.4	70.9	188.0	3.472	18.5	25
11	Buick	Century	91.561	12.475	0	21.975	3.1	175	109.0	72.7	194.6	3.368	17.5	25
12	Buick	Regal	39.350	13.740	0	25.300	3.8	240	109.0	72.7	196.2	3.543	17.5	23
13	Buick	Park Avenue	27.851	20.190	0	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5	24
14	Buick	LeSabre	83.257	13.360	0	27.885	3.8	205	112.2	73.5	200.0	3.591	17.5	25
15	Cadillac	DeVille	63.729	22.525	0	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5	22
16	Cadillac	Seville	15.943	27.100	0	44.475	4.6	275	112.2	75.0	201.0		18.5	22
17	Cadillac	Eldorado	6.536	25.725	0	39.665	4.6	275	108.0	75.5	200.6	3.843	19.0	22
18	Cadillac	Catera	11.185	18.225	0	31.010	3.0	200	107.4	70.3	194.8	3.770	18.0	22
19	Cadillac	Escalade	14.785		1	46.225	5.7	255	117.5	77.0	201.2	6.572	30.0	16
20	Chevrolet	Cavalier	145.519	9.250	0	13.260	2.2	115	104.1	67.9	180.9	2.676	14.3	27
21	Chevrolet	Malibu	136.126	11.225	0	16.535	3.1	170	107.0	69.4	190.4	3.051	15.0	25
22	Chevrolet	Lumina	24.629	10.310	0	18.890	3.1	175	107.5	72.5	200.9	3.330	16.6	25
23	Chevrolet	Monte Carlo	42.593	11.525	0	19.390	3.4	180	110.5	72.7	197.9	3.340	17.0	27
24	Chevrolet	Camaro	26.402	13.025	0	24.340	3.8	200	101.1	74.1	193.2	3.500	16.8	25
25	Chevrolet	Corvette	17.947	36.225	0	45.705	5.7	345	104.5	73.6	179.7	3.210	19.1	22
26	Chevrolet	Prizm	32.299	9.125	0	13.960	1.8	120	97.1	66.7	174.3	2.398	13.2	33
27	Chevrolet	Metro	21.855	5.160	0	9.235	1.0	55	93.1	62.6	149.4	1.895	10.3	45
28	Chevrolet	Impala	107.995		0	18.890	3.4	180	110.5	73.0	200.0	3.389	17.0	27
29	Chrysler	Sebring Coupe	7.954	12.360	0	19.640	2.5	163	103.7	69.7	190.9	2.967	15.9	24
30	Chrysler	Sebring Conv	32.775	14.180	0	24.495	2.5	168	106.0	69.2	193.0	3.332	16.0	24
31	Chrysler	Concorde	31.148	13.725	0	22.245	2.7	200	113.0	74.4	209.1	3.452	17.0	26
32	Chrysler	Cirrus	32.306	12.640	0	16.480	2.0	132	108.0	71.0	186.0	2.911	16.0	27
33	Chrysler	LHS	13.462	17.325	0	28.340	3.5	253	113.0	74.4	207.7	3.564	17.0	23
34	Chrysler	Town & Country	53.480	19.540	1									
35	Chrysler	300M	30.696		0	29.185	3.5	253	113.0	74.4	197.8	3.567	17.0	23
36	Dodge	Neon	76.034	7.750	0	12.640	2.0	132	105.0	74.4	174.4	2.567	12.5	29
37	Dodge	Avenger	4.734	12.545	0	19.045	2.5	163	103.7	69.1	190.2	2.879	15.9	24
38	Dodge	Stratus	71.186	10.185	0	20.230	2.5	168	108.0	71.0	186.0	3.059	16.0	24
39	Dodge	Intrepid	88.028	12.275	0	22.505	2.7	202	113.0	74.7	203.7	3.489	17.0	

●คลิ๊ก Analyze>Classify>Two-Step cluster



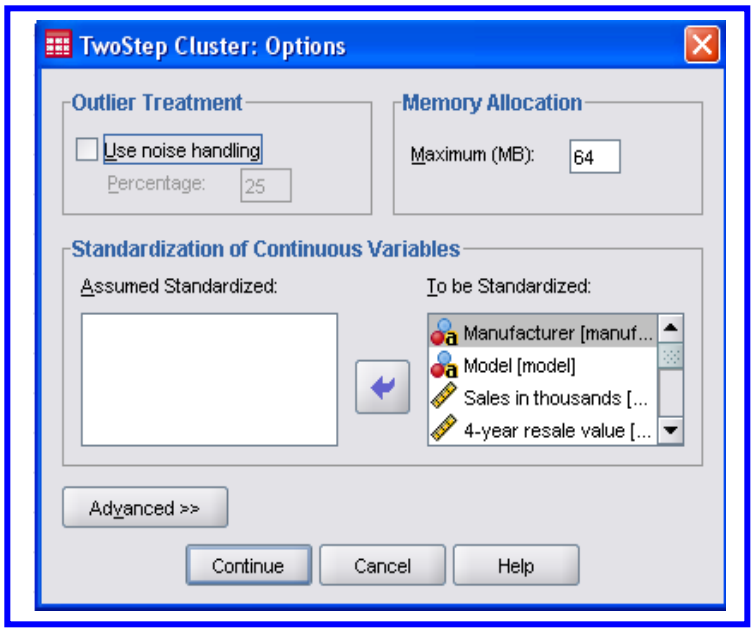
จะปรากฏ menu



●คลิ๊กเลือกตัวแปรที่เป็นประเภท ในที่นี้คือ vehicle type ให้ไปอยู่ในกล่อง Categorical Variables

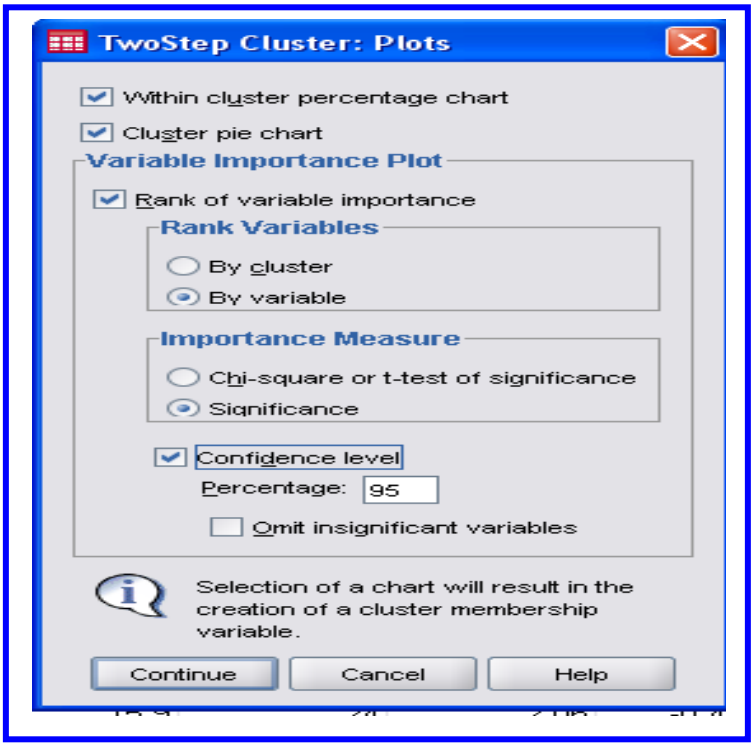
●คลิ๊กเลือกตัวแปรที่มีค่าต่อเนื่อง (price,engine - size, horsepower, wheelbase,width, length,weight,fuel capacity, fuel efficiency)ให้ไปอยู่ในกล่อง Continuous Variables

●คลิ๊กปุ่ม Options



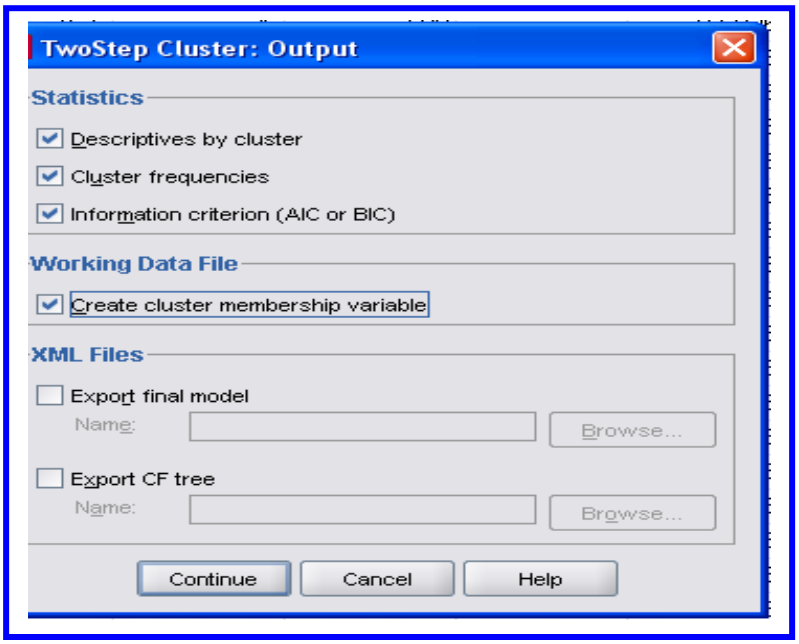
เนื่องจากตัวแปรที่มีค่าต่อเนื่องมีหน่วยแตกต่างกัน ดังนั้นตัวแปรเหล่านี้จำเป็นต้องมีการstandardize เสียก่อน (to be standardized)

●คลิ๊กปุ่ม plots



●คลิ๊ก Within cluster percentage chart , Cluster pie chart ในส่วนของ Variable Importance Plot คลิ๊กเลือก Rank of Variable importance by variable ส่วน Importance Measure คลิ๊ก Significance และคลิ๊ก Confidence level คลิ๊ก Continue

●คลิ๊กปุ่ม Outputs



ในส่วนของ Statistics คลิ๊กเลือก Descriptives by cluster/ Cluster frequencies/Information criterion
ในส่วนของ Working Data File คลิ๊ก Create cluster membership variable
●คลิ๊ก Continue

- คลิ๊ก Continue ที่เมนูหลัก จะปรากฏ computer output ดังต่อไปนี้

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

a. The changes are from the previous number of clusters in the table.
b. The ratios of changes are relative to the change for the two cluster solution.
c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Algorithm ในการหาจำนวนclusterที่เหมาะสมโดยอัตโนมัติจะทำการเลือกจำนวน cluster ที่มีค่า BIC/AIC ต่ำที่สุดและค่า Ratio of Distance Measures มีค่าสูงที่สุด

ลำดับถัดไปจะเป็น Cluster Distribution table ที่เป็นตารางแสดงจำนวนของ case ที่ถูกจัดในแต่ละ cluster

Cluster Distribution

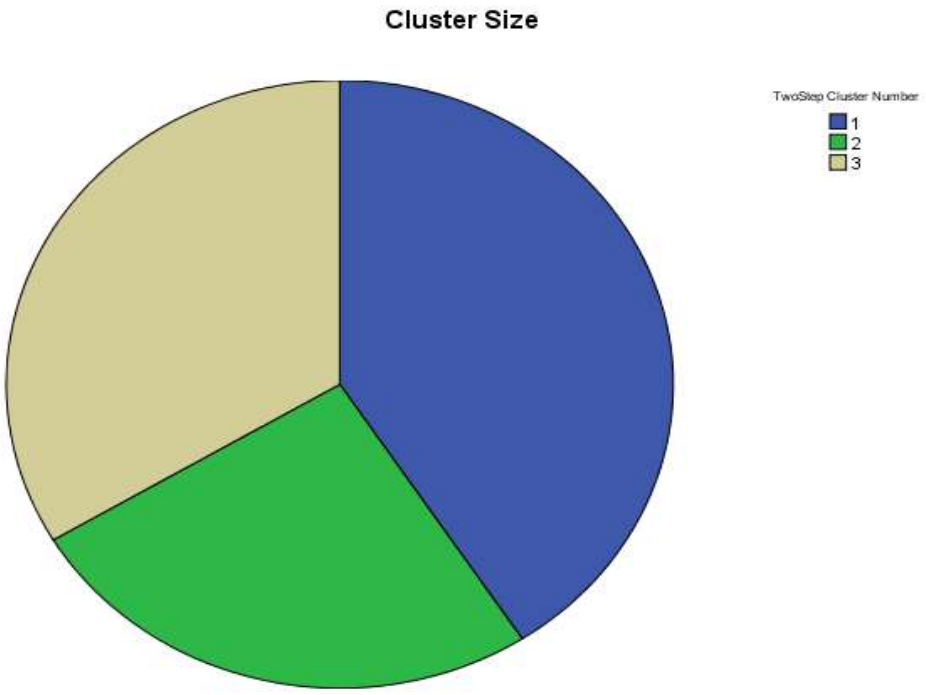
		N	% of Combined	% of Total
Cluster	1	62	40.8%	39.5%
	2	39	25.7%	24.8%
	3	51	33.6%	32.5%
	Combined	152	100.0%	96.8%
Excluded Cases		5		3.2%
Total		157		100.0%

Vehicle type

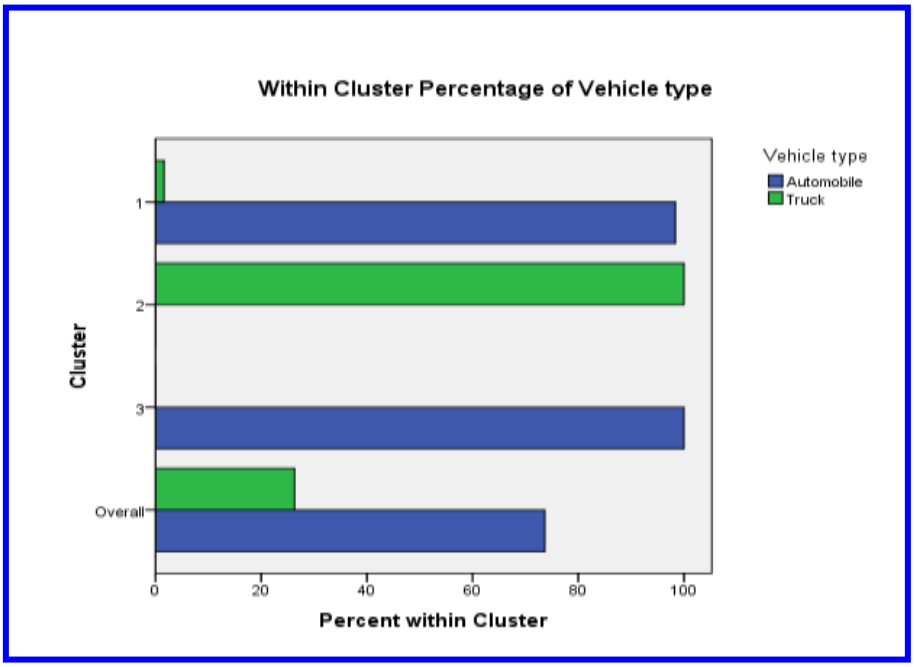
		Automobile		Truck	
		Frequency	Percent	Frequency	Percent
Cluster	1	61	54.5%	1	2.5%
	2	0	.0%	39	97.5%
	3	51	45.5%	0	.0%
	Combined	112	100.0%	40	100.0%

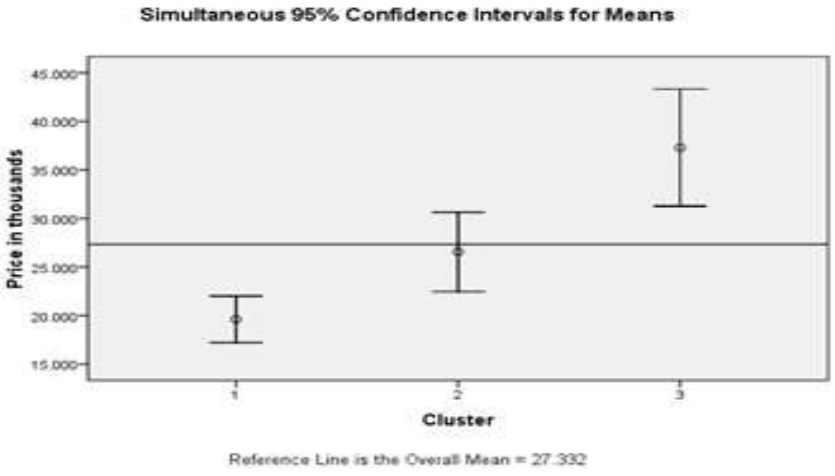
ในที่นี้ cluster ที่สองจะประกอบด้วยรถบรรทุกทั้งหมด ในขณะที่ cluster ที่สามจะประกอบด้วยรถยนตร์นั่งทั้งหมด

ภาพถัดไปจะเป็น pie chart แสดงขนาดของแต่ละ cluster



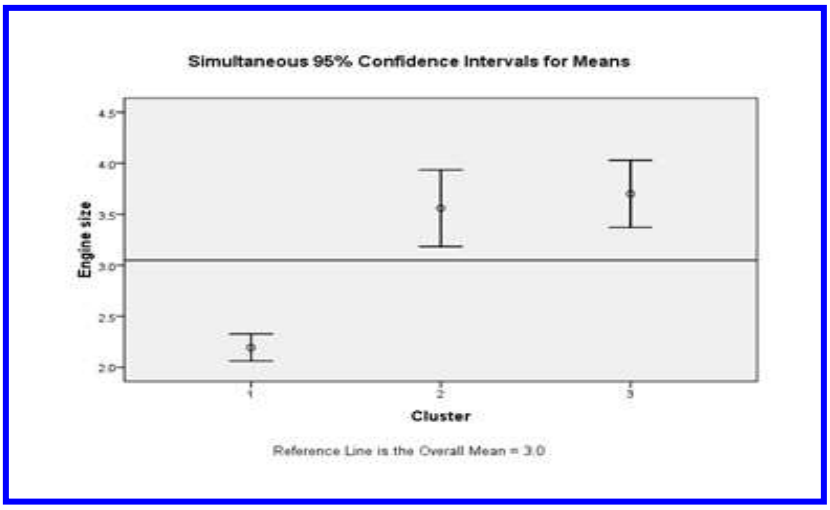
ภาพถัดไปเป็น bar graph แสดงสัดส่วนของ categorical variable ในแต่ละ cluster

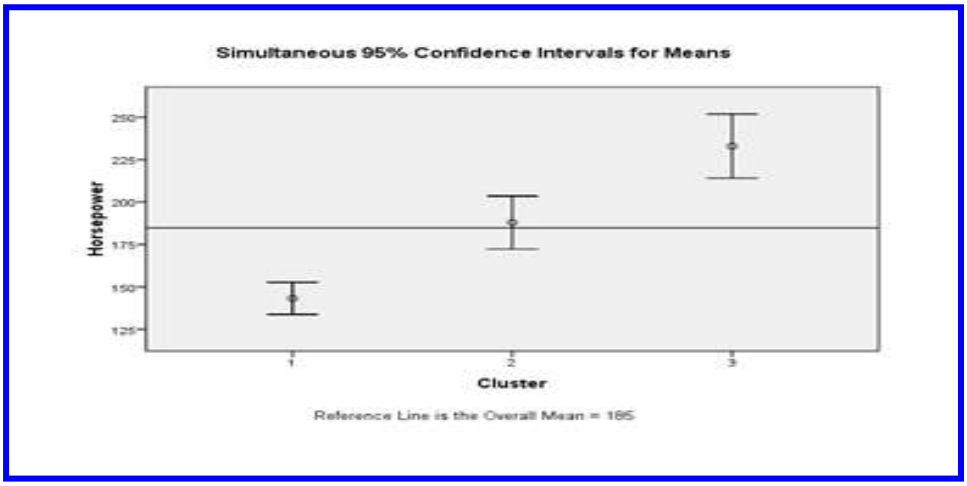




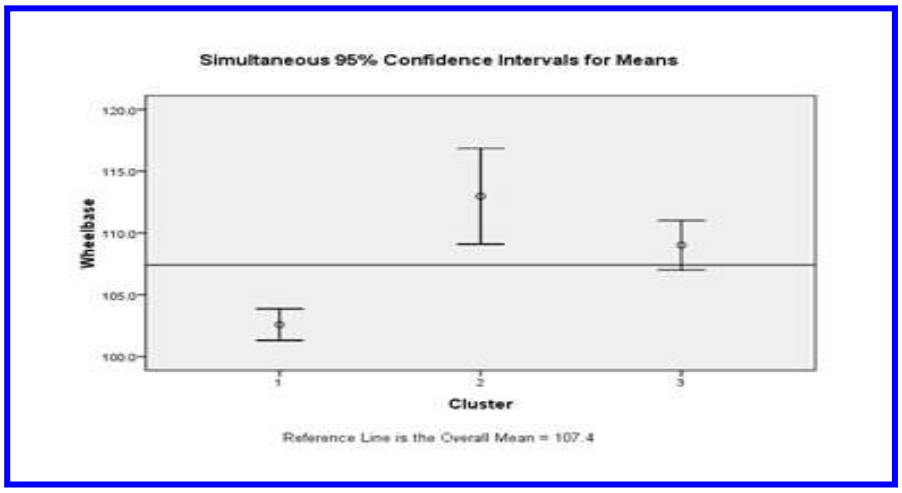
ภาพด้านซ้ายแสดงราคาของcluster ที่ 3 ว่ามีค่าเฉลี่ยสูงกว่า cluster อื่นๆ

ภาพขวามือแสดงให้เห็นว่าขนาดของเครื่องยนตร์ในCluster ที่หนึ่งมีขนาดค่าเฉลี่ยต่ำกว่าอีกสอง cluster



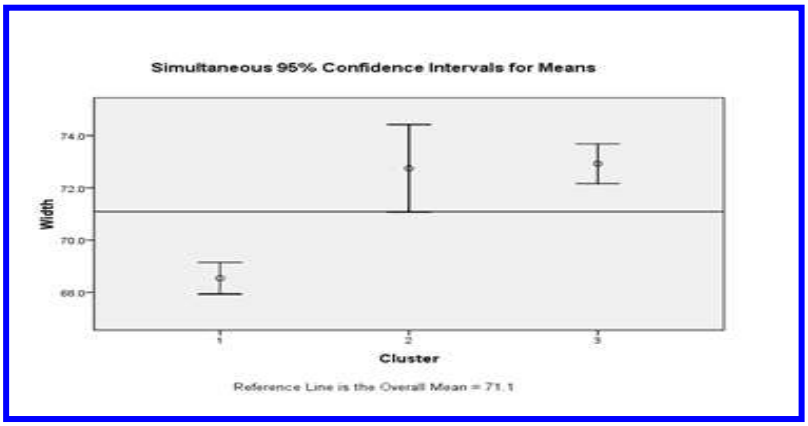


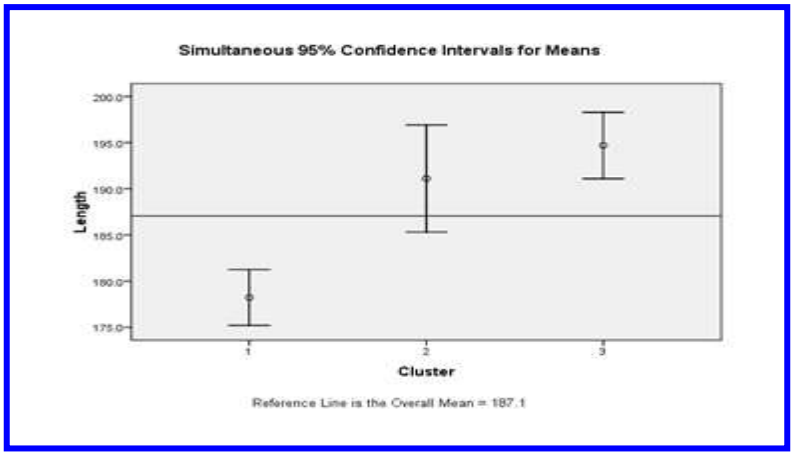
ภาพซ้ายมือแสดงให้เห็นว่า cluster ที่หนึ่งมีขนาดแรงม้าโดยเฉลี่ยต่ำกว่าอีกสอง cluster



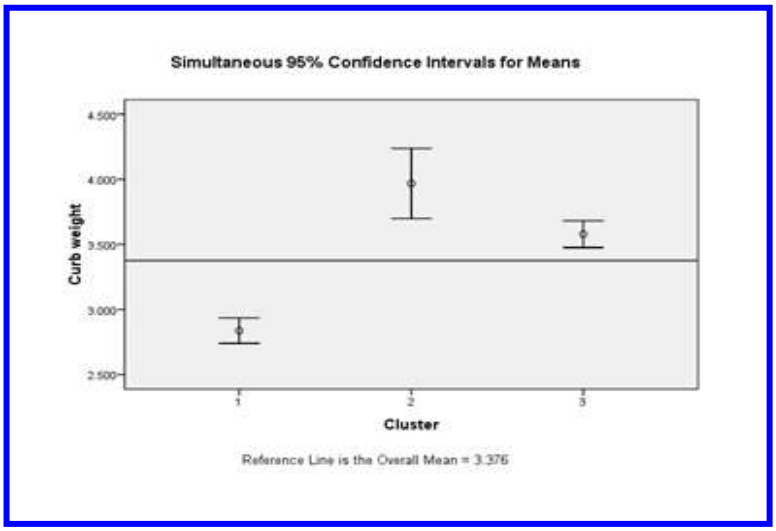
ภาพซ้ายมือแสดงให้เห็นว่า cluster ที่หนึ่งมีWheelbase โดยเฉลี่ยต่ำกว่าอีกสองcluster

ภาพขวามือแสดงให้เห็นว่า clusterที่หนึ่งมีความกว้างโดยเฉลี่ยน้อยกว่าอีกสอง cluster



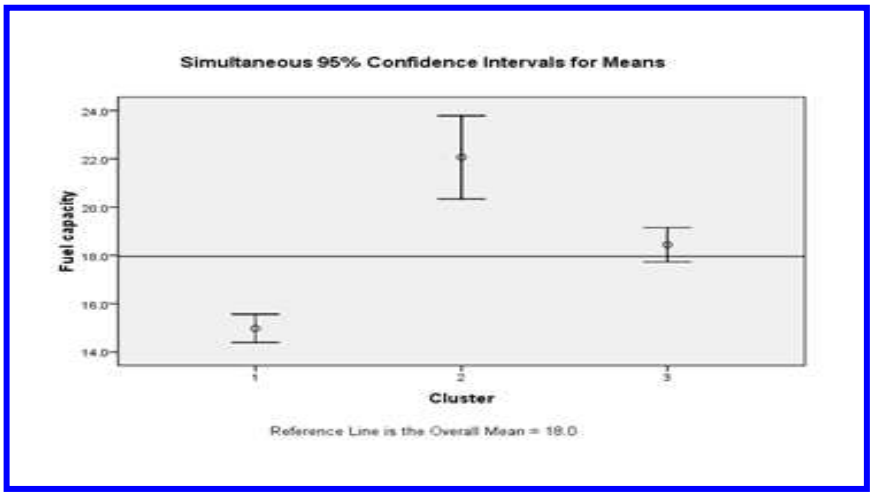


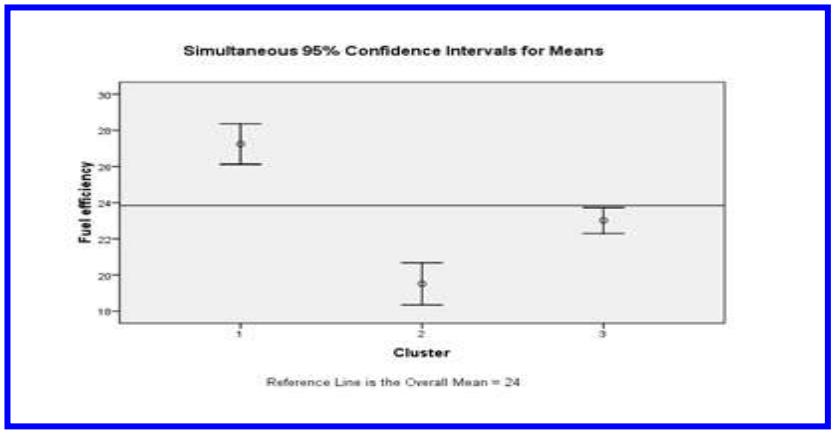
ภาพซ้ายมือแสดงให้เห็นว่าcluster ที่หนึ่งมีความยาวโดยเฉลี่ยน้อยกว่าอีกสองcluster



ภาพซ้ายมือแสดงให้เห็นว่าclusterที่หนึ่งมีน้ำหนักรถเปล่าต่ำสุดในสาม cluster

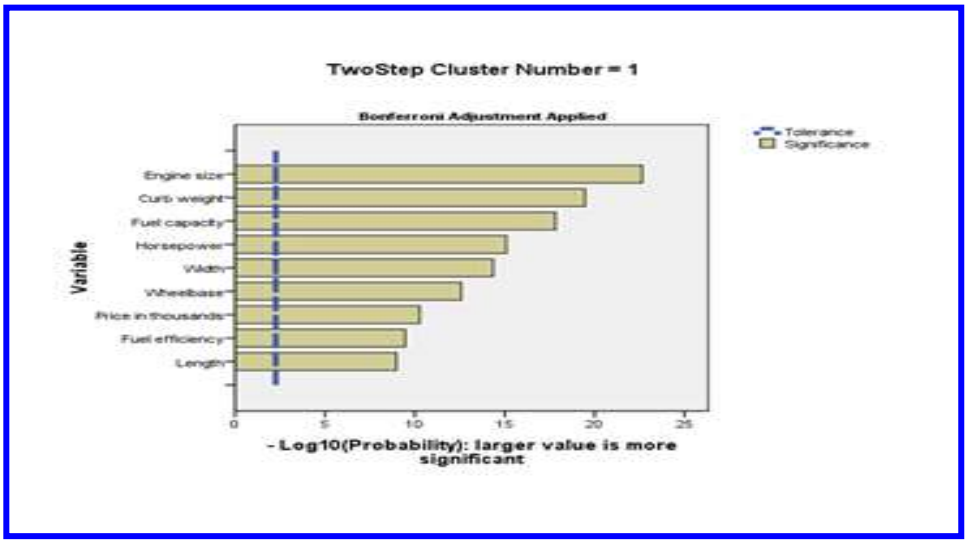
ภาพขวามือแสดงให้เห็นว่า cluster ที่หนึ่งมีขนาดความจุของถังน้ำมันต่ำกว่าอีกสอง cluster



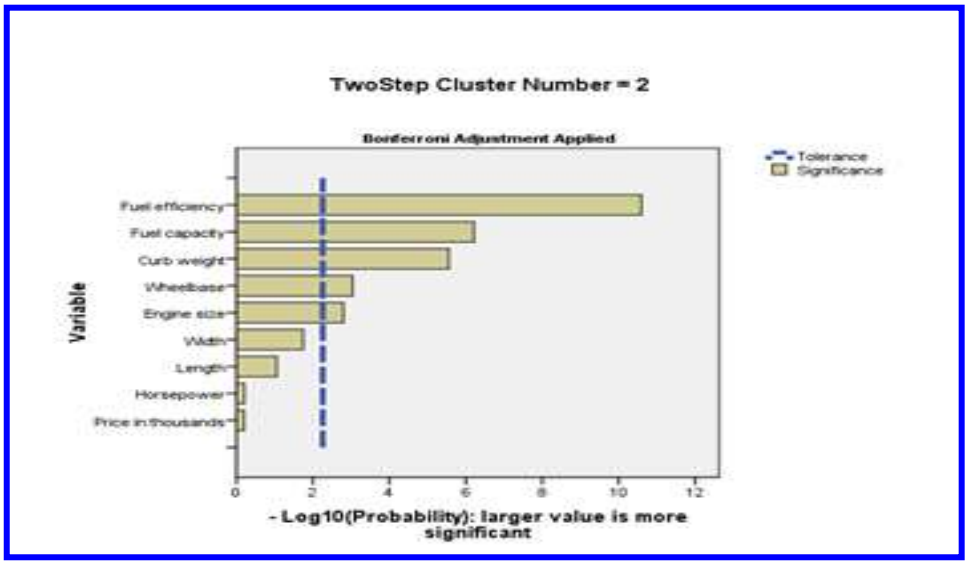


ภาพซ้ายมือแสดงให้เห็นว่า clusterที่หนึ่งมีการประหยัดน้ำมันได้ดีกว่าอีกสอง cluster

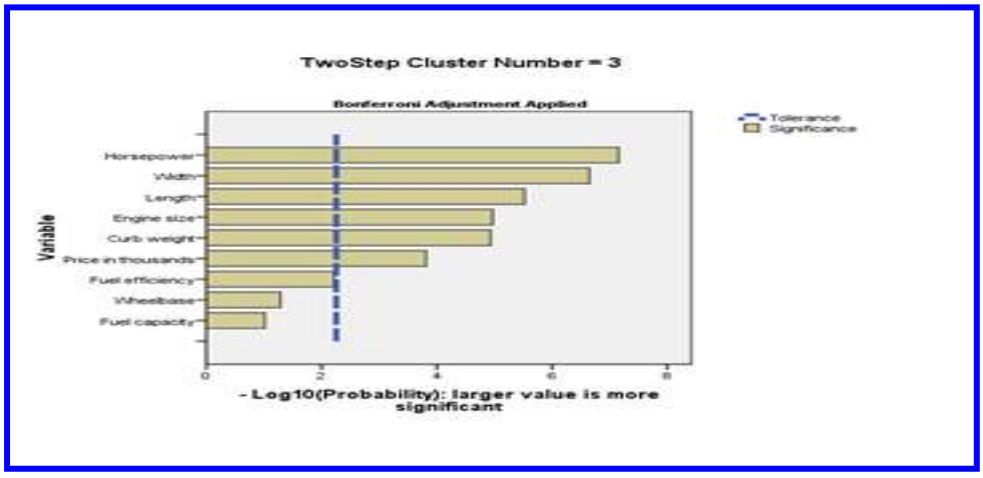
ภาพสามภาพถัดไปแสดงความสำคัญของตัวแปรที่แยกแยะให้เห็นความแตกต่างของแต่ละcluster



ตัวแปรทั้ง 9 ตัวแปรมีนัยสำคัญในการแยกแยะ case ที่อยู่ในclusterที่หนึ่ง



ตัวแปรที่มีนัยสำคัญในการแยกแยะความแตกต่างของ caseที่อยู่ใน cluster ที่สองได้แก่ fuel efficiency/ fuel capacity/curb weight/wheelbase/ engine size ส่วนตัวแปรอื่นๆไม่มีความสำคัญ



ตัวแปรที่มีนัยสำคัญในการแยกแยะความแตกต่างของCaseที่อยู่ในclusterที่สามได้แก่horsepower/width/ length/engine size/curb weight/price และfuel efficiency

Contribution this issue: ดร.ดนัย ปัตตพงศ์

อยากเรียนรู้การนำสถิติข้างต้นนี้ไปใช้ในการวิจัยระดับสารนิพนธ์ (independent study) วิทยานิพนธ์ (thesis) ดุษฎีนิพนธ์(dissertation) ปรึกษาได้ที่ dpattaphongse@gmail.com

* ผู้แต่ง MBA's Made Easy (160+ issues) เอกสารวิชาการด้านศาสตร์การบริหารธุรกิจที่ช่วยให้ธุรกิจสามารถยืนหยัดและอยู่รอดได้ในภาวะที่โลกเปลี่ยนแปลงอยู่ตลอดเวลา

* ผู้พัฒนา FINALYSIS… a dedicated software สำหรับให้บริการนักธุรกิจที่ต้องการวิเคราะห์ความเป็นไปได้ทางการเงินของโครงการพัฒนาอสังหาริมทรัพย์ (บ้านจัดสรร/จัดสรรที่ดินเพื่อการอุตสาหกรรม/อาคารชุด/อาคารสำนักงานให้เช่า) โรงแรม โรงพยาบาลเอกชน ห้างสรรพสินค้า โรงงานน้ำตาล โรงงานกระดาษ โรงไฟฟ้าชีวมวล ฯลฯ ได้เห็นตัวเลขก่อนโครงการเกิด หลีกเลี่ยงความผิดพลาดเป็นร้อยเป็นพันล้านหากเกิดการลงทุนจริง(กำหนด DEBUT 1 เมษายน 2569)

* ผู้แต่งหนังสือ"การวิเคราะห์ความเป็นไปได้ทางการเงินและการจัดวงเงินเครดิตของโครงการลงทุน"ประกอบด้วยตัวอย่างของธุรกิจจริงที่ไม่เปิดเผยชื่อนับ 100 บริษัท ครอบคลุมอุตสาหกรรม 24 อุตสาหกรรม

*Co-developer ซอฟแวร์ en◉gex◉cel® สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)