Keiland Pullen
DSC 441 – Fall 2021

Problem 1.

1.)

Statistics

		age	income	children	pep
N	Valid	600	600	600	600
	Missing	0	0	0	0
Mean		42.40	27524.0312	1.01	.46
Median		42.00	24925.3000	1.00	.00
Mode		40[a]	38248.30	0	0
Std. Deviation		14.425	12899.46825	1.057	.499
Variance		208.079	166396281.0	1.117	.249
Range		49	58115.89	3	1
Minimum		18	5014.21	0	0
Maximum		67	63130.10	3	1
Sum		25437	16514418.73	607	274

a. Multiple modes exist. The smallest value is shown

gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FEMALE	300	50.0	50.0	50.0
	MALE	300	50.0	50.0	100.0
	Total	600	100.0	100.0	

region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	INNER_CITY	269	44.8	44.8	44.8
	RURAL	96	16.0	16.0	60.8
	SUBURBAN	62	10.3	10.3	71.2
	TOWN	173	28.8	28.8	100.0
	Total	600	100.0	100.0	

Keiland Pullen
DSC 441 – Fall 2021

married

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	204	34.0	34.0	34.0
	YES	396	66.0	66.0	100.0
	Total	600	100.0	100.0	

car

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	304	50.7	50.7	50.7
	YES	296	49.3	49.3	100.0
	Total	600	100.0	100.0	

savings_acct

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	186	31.0	31.0	31.0
	YES	414	69.0	69.0	100.0
	Total	600	100.0	100.0	

current_acct

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	145	24.2	24.2	24.2
	YES	455	75.8	75.8	100.0
	Total	600	100.0	100.0	

mortgage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	391	65.2	65.2	65.2
	YES	209	34.8	34.8	100.0
	Total	600	100.0	100.0	

Keiland Pullen
DSC 441 – Fall 2021

2.)

PEP = 1

```
USE ALL.
COMPUTE filter_$=(pep = 1).
VARIABLE LABELS filter_$ 'pep = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
USE ALL.
COMPUTE filter_$=(pep = 1).
VARIABLE LABELS filter_$ 'pep = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
FREQUENCIES VARIABLES=age income children pep
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE SUM
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Frequencies

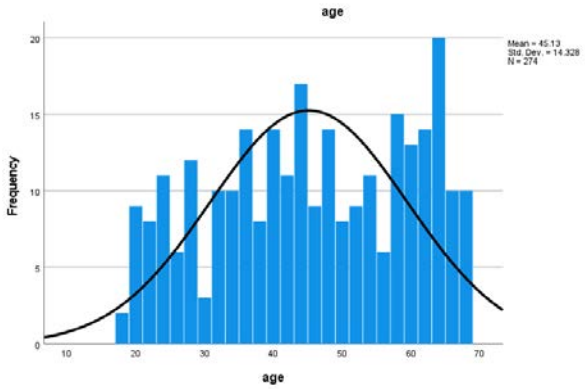
Statistics

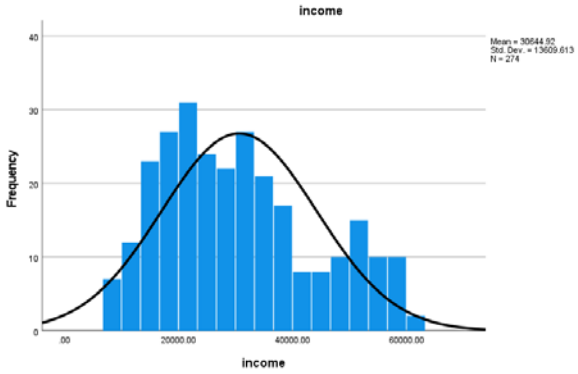
		age	income	children	pep
N	Valid	274	274	274	274
	Missing	0	0	0	0
Mean		45.13	30644.9195	.95	1.00
Median		45.00	28080.0500	1.00	1.00
Mode		64	7756.36[a]	1	1
Std. Deviation		14.328	13609.61304	.861	.000
Variance		205.291	185221567.2	.741	.000
Range		49	55373.74	3	0
Minimum		18	7756.36	0	1
Maximum		67	63130.10	3	1
Sum		12365	8396707.93	259	274

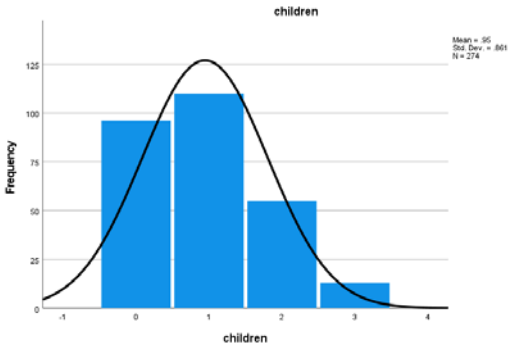
a. Multiple modes exist. The smallest value is shown

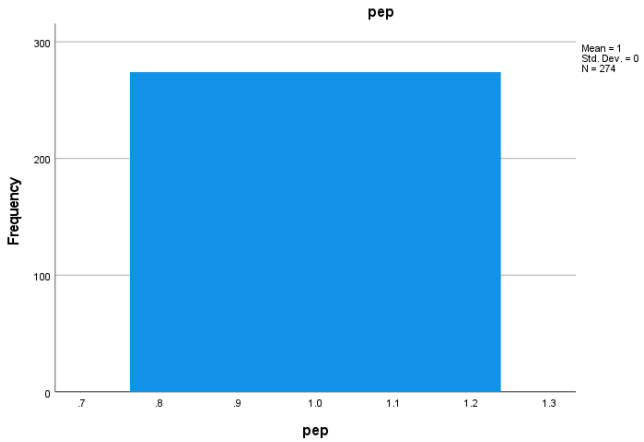
Keiland Pullen
DSC 441 – Fall 2021

Assignment 5









PEP = 0

```
USE ALL.
COMPUTE filter_$=(pep = 0).
VARIABLE LABELS filter_$ 'pep = 0 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (fl.0).
FILTER BY filter_$.
EXECUTE.
FREQUENCIES VARIABLES=age income children pep
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE SUM
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

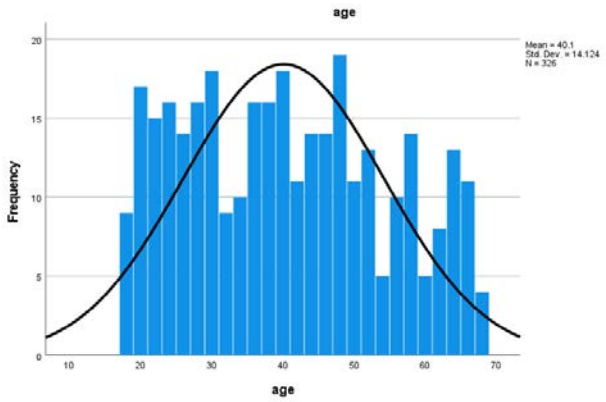
Keiland Pullen
DSC 441 – Fall 2021

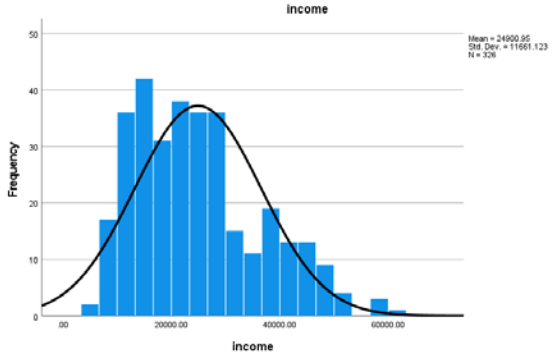
Assignment 5

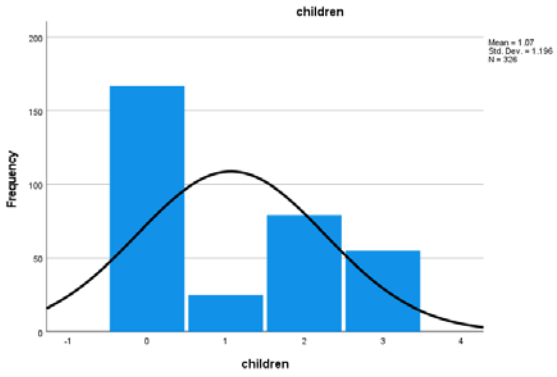
Statistics

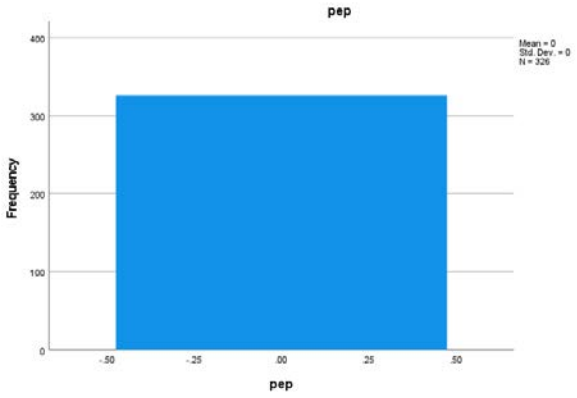
		age	income	children	pep
N	Valid	326	326	326	326
	Missing	0	0	0	0
Mean		40.10	24900.9534	1.07	.00
Median		40.00	23105.0000	.00	.00
Mode		27[a]	38248.30	0	0
Std. Deviation		14.124	11661.12342	1.196	.000
Variance		199.473	135981799.464	1.429	.000
Range		49	56540.39	3	0
Minimum		18	5014.21	0	0
Maximum		67	61554.60	3	0
Sum		13072	8117710.80	348	0

a. Multiple modes exist. The smallest value is shown







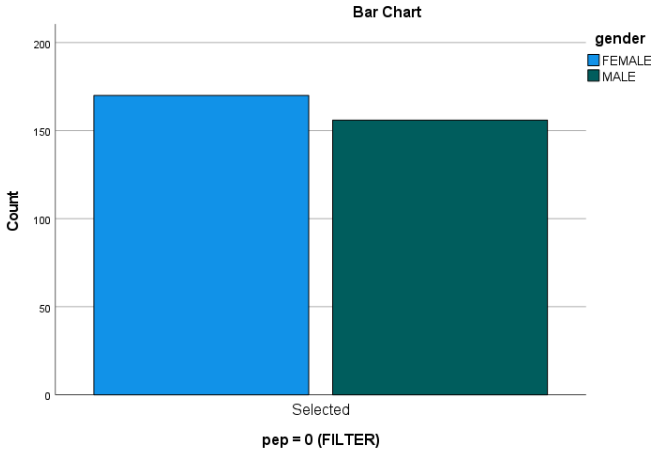


Keiland Pullen
DSC 441 – Fall 2021

pep = 0 (FILTER) * gender Crosstabulation

Count

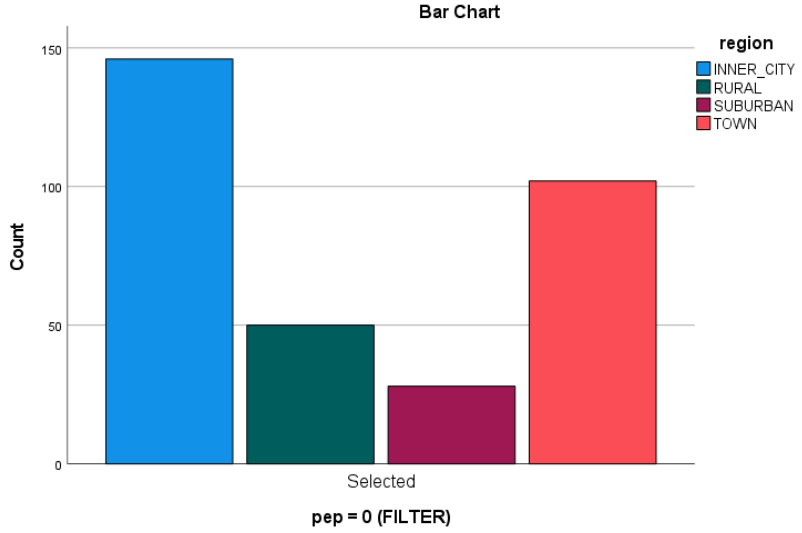
		gender		
		FEMALE	MALE	Total
pep = 0 (FILTER)	Selected	170	156	326
Total		170	156	326



pep = 0 (FILTER) * region Crosstabulation

Count

		region				
		INNER_CITY	RURAL	SUBURBAN	TOWN	Total
pep = 0 (FILTER)	Selected	146	50	28	102	326
Total		146	50	28	102	326

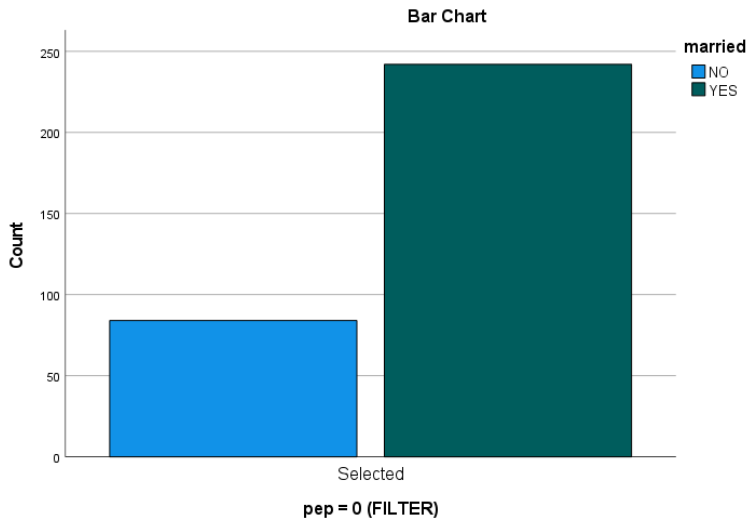


Keiland Pullen
DSC 441 – Fall 2021

pep = 0 (FILTER) * married Crosstabulation

Count

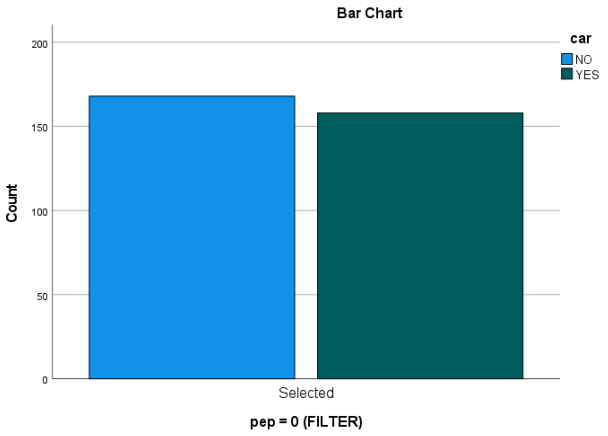
		married		
		NO	YES	Total
pep = 0 (FILTER)	Selected	84	242	326
Total		84	242	326



pep = 0 (FILTER) * car Crosstabulation

Count

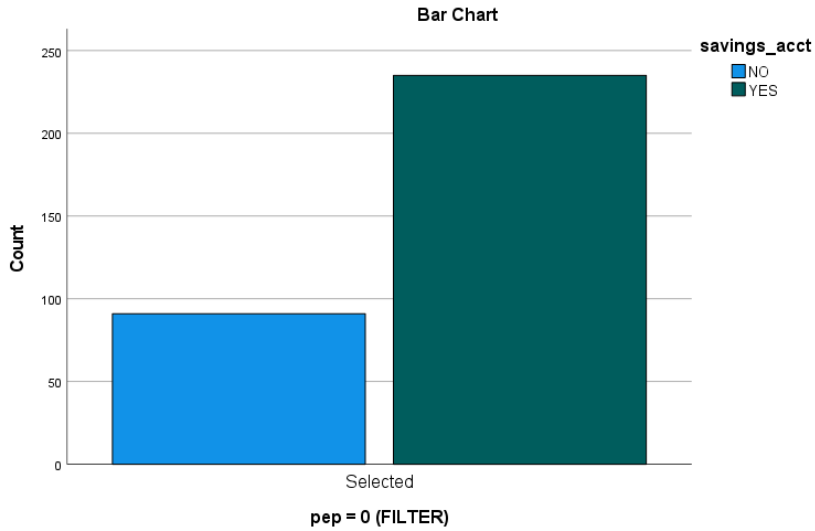
		car		
		NO	YES	Total
pep = 0 (FILTER)	Selected	168	158	326
Total		168	158	326



pep = 0 (FILTER) * savings_acct Crosstabulation

Count

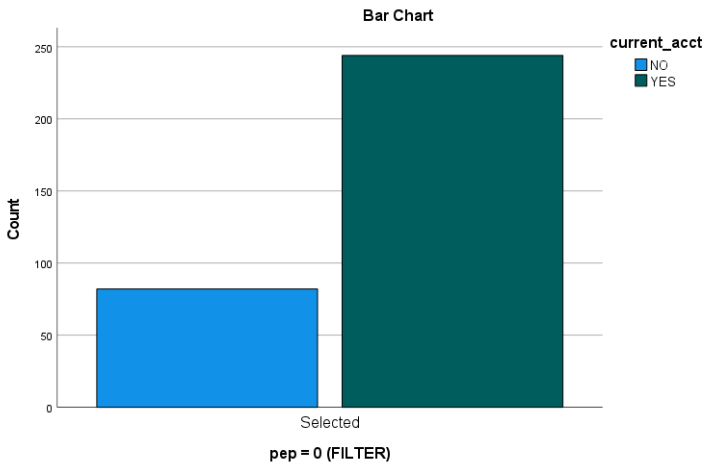
		savings_acct		
		NO	YES	Total
pep = 0 (FILTER)	Selected	91	235	326
Total		91	235	326



pep = 0 (FILTER) * current_acct Crosstabulation

Count

		current_acct		
		NO	YES	Total
pep = 0 (FILTER)	Selected	82	244	326
Total		82	244	326

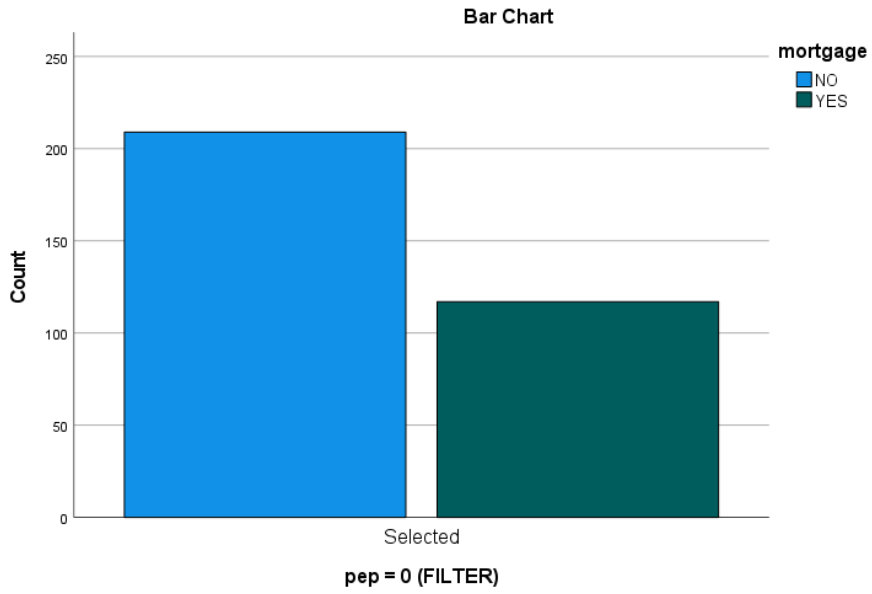


pep = 0 (FILTER) * mortgage Crosstabulation

Keiland Pullen
DSC 441 – Fall 2021

Count

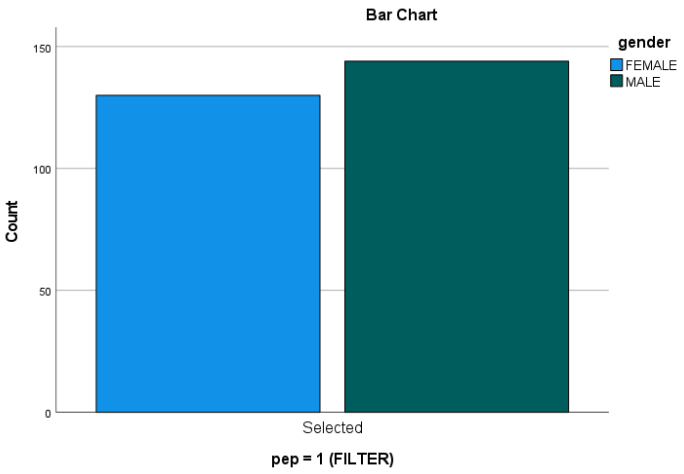
		mortgage		
		NO	YES	Total
pep = 0 (FILTER)	Selected	209	117	326
Total		209	117	326



pep = 1 (FILTER) * gender Crosstabulation

Count

		gender		
		FEMALE	MALE	Total
pep = 1 (FILTER)	Selected	130	144	274
Total		130	144	274

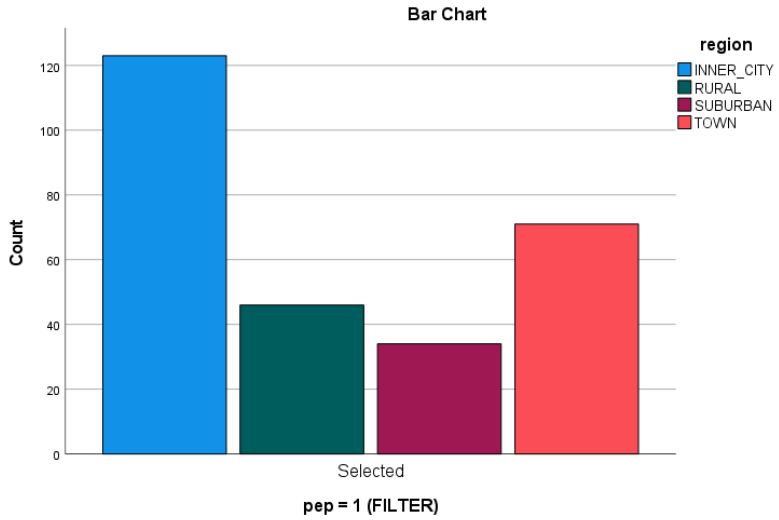


Keiland Pullen
DSC 441 – Fall 2021

pep = 1 (FILTER) * region Crosstabulation

Count

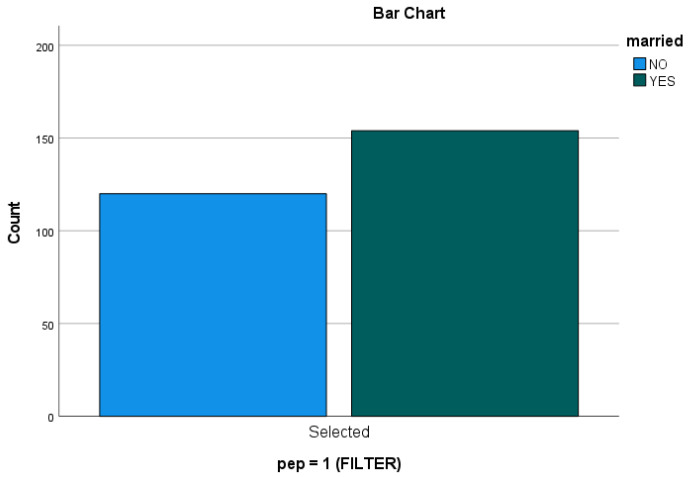
		region				
		INNER_CITY	RURAL	SUBURBAN	TOWN	Total
pep = 1 (FILTER)	Selected	123	46	34	71	274
Total		123	46	34	71	274



pep = 1 (FILTER) * married Crosstabulation

Count

		married		
		NO	YES	Total
pep = 1 (FILTER)	Selected	120	154	274
Total		120	154	274

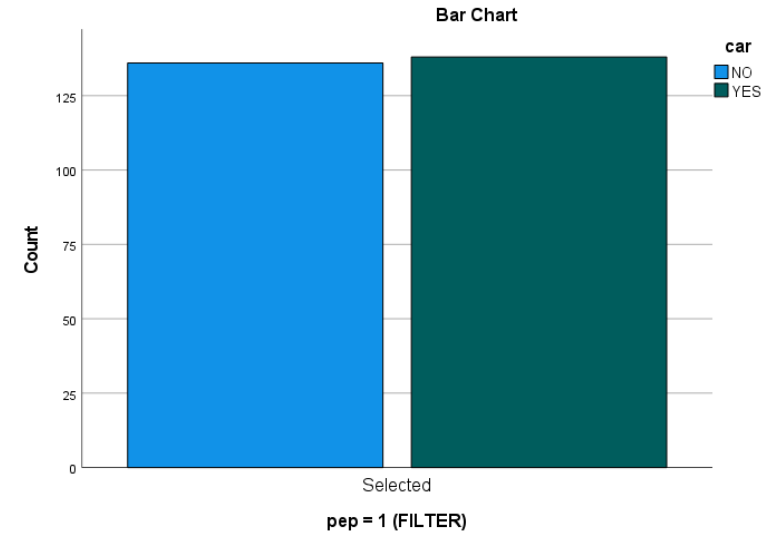


Keiland Pullen
DSC 441 – Fall 2021

pep = 1 (FILTER) * car Crosstabulation

Count

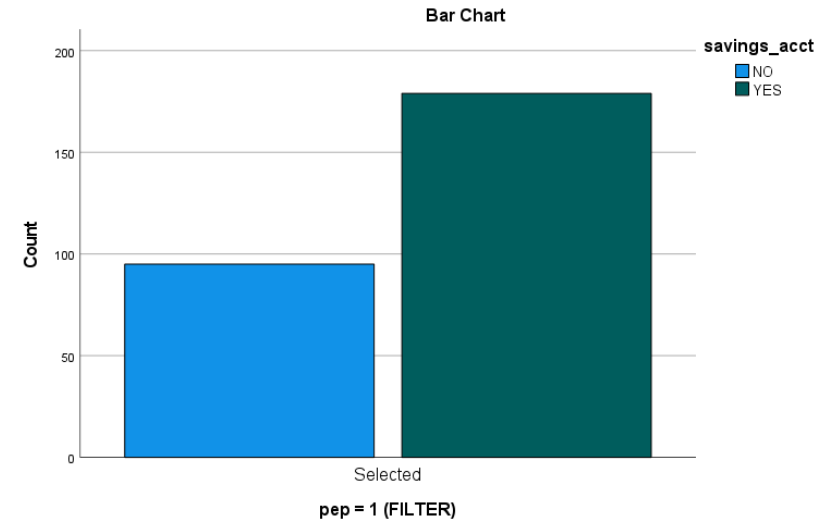
		car		
		NO	YES	Total
pep = 1 (FILTER)	Selected	136	138	274
Total		136	138	274



pep = 1 (FILTER) * savings_acct Crosstabulation

Count

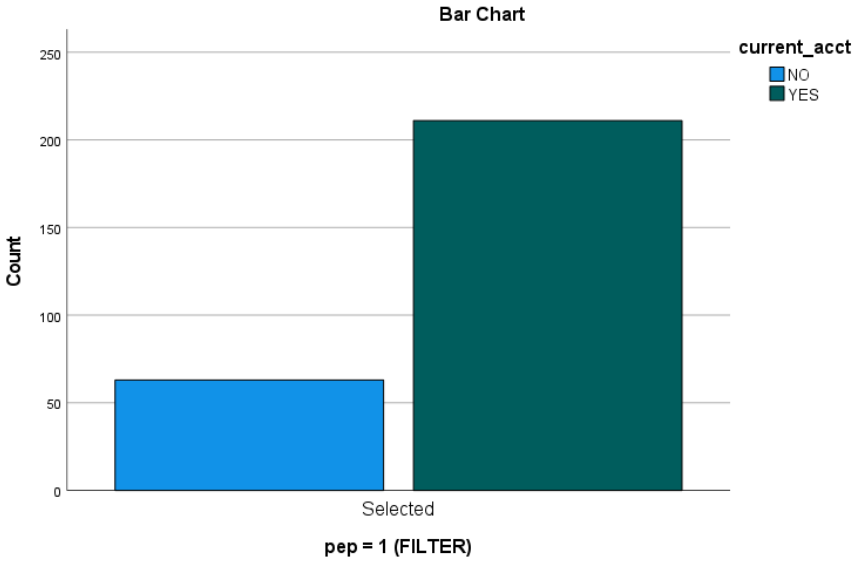
		savings_acct		
		NO	YES	Total
pep = 1 (FILTER)	Selected	95	179	274
Total		95	179	274



pep = 1 (FILTER) * current_acct Crosstabulation

Count

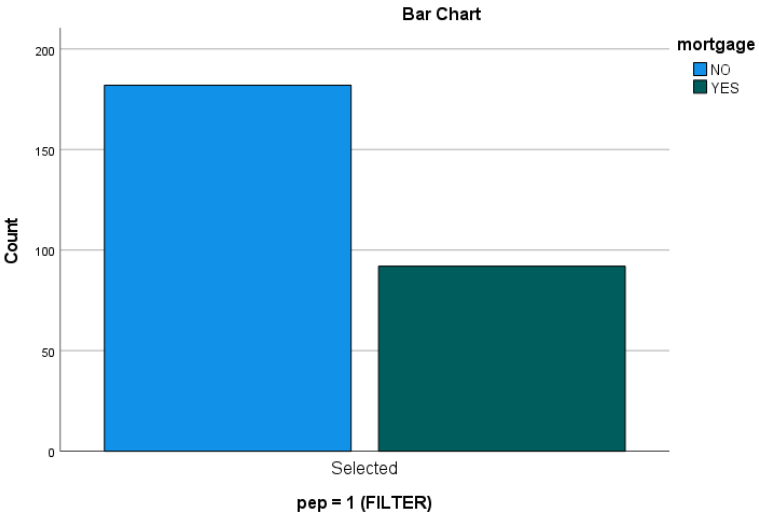
		current_acct		
		NO	YES	Total
pep = 1 (FILTER)	Selected	63	211	274
Total		63	211	274



pep = 1 (FILTER) * mortgage Crosstabulation

Count

		mortgage		
		NO	YES	Total
pep = 1 (FILTER)	Selected	182	92	274
Total		182	92	274



For the numerical variables, histograms are used to display difference between PEP=0 and PEP=1. In regard to Age, the mean for PEP=1 is slightly higher by 5 years, with a value of 45.13 as compared to 40.1 for PEP=0. Concerning Income, the mean for PEP=1 is 30,000+ where as PEP=0 has a mean income of 24,900. Regarding the number of Children is fairly close with PEP=1 providing a mean of .95 with a mean of 1.07 for PEP=0.

For the categorical variables, bar charts are used to display differences between PEP=0 and PEP=1. The gender percentages were fairly close, with slightly more Males with PEP=1. The Regions were fairly consisitent in both PEP-0 and PEP=1. For Married, the percentage of "Yes, Married" was much greated in PEP=0. Car was almost identical in both PEP values. For Savings Account, there was a greater percentage in PEP=0. For Current Account, the percentages and values were almost similar in both PEP values. For Mortgage, the percentages are similar in both PEP values.

3.)
To calculate the z-score normalization on the income variable, click on “Analyze”, then select “Descriptives”. Add “Income” to the variables list. Click “Save standardized values as variables”. Click “Options”, then “OK”. Review the “Data View” to ensure that the new z-score variable has been created and populated correctly.

4.)
To Discretize the “Age” attribute into 3 categories, select “Transform” from top menu, then select “Recode into Different Variables”. Select the “age” variable, the provide a name and label for the new variable. Click the “Old and New Values” button. In the pop up screen, click the “Output variables are string” check-box. Select “Range” then add the range of years for “Young”, in the “New Value” text, add the string name “Young”, then select “Add”. Repeat this for the next 2 variables. Check the “Data View” to ensure that the data has been transformed.

5.)
For Min-Max normalization, first, review the formula. Then select “Analyze” and “Descriptives”. Select the variables for the Income, Age and Children fields. Click “Options”, select “Minimum” and “Maximum” values. From the menu, select “Transform” then “Compute Variable”. Add the name of the new target variable, then add the min-max normalization formula to the “Numeric Expression” field. For Min and Max values, use the output from the “Descriptives” screen. Add a new target variable for the Income, Age and Children fields. Then save.

6.)
The “region” variable was transformed by selecting “Transform” from the menu bar, then “Recode into Same Variable”. Then click the “Old and New Values” button. On the pop-up screen, enter the “Old Value” or current value, then enter a number for that value in the “New Value” text box. Click “Continue” then “OK”. On the “Variable View” tab, select the “region” variable, change its “Type” to “Numeric”, then in t1he “Values” field, open that pop-up and add the values for each region: 1,2, 3 and 4 for INNER_CITY, RURAL, SUBURBAN and TOWN. Check “Data View” to ensure that the data is there and has been transformed.

Keiland Pullen
DSC 441 – Fall 2021

Problem 2.)

i.) a.) First scale the variables by calculating the z-score for each independent variable. The z-score variables will be created in the Data View. The next step in SPSS is to select:

1.) Analyze ~> Classify ~> K-Means Cluster
2.) Select the variables
3.) Set the number of clusters
4.) Select "Iterate" to set the maximum number of iterations
5.) Select "Save" and check the "Cluster membership" checkbox
6.) Select "Options" and check the "Initial cluster centers" and "ANOVA table" options
7.) Click OK

b.) According to documentation, SPSS using the Euclidean distance as its similarity measure.

For K=3 Clusters:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
area	210	10.59	21.18	14.8475	2.90970
perimeter	210	12.41	17.25	14.5593	1.30596
compactness	210	.8081	.9183	.870999	.0236294
length_of_kernel	210	4.899	6.675	5.62853	.443063
width_of_kernel	210	2.630	4.033	3.25860	.377714
asymmetry_coefficient	210	.7651	8.4560	3.700201	1.5035571
length_of_kernel_grove	210	4.519	6.550	5.40807	.491480
Valid N (listwise)	210				

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	1.968	2.000	2.044
2	.105	.366	.842
3	.072	.465	.457
4	.054	.230	.194
5	.000	.047	.052
6	.029	.028	.000
7	.026	.024	.000
8	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 8. The minimum distance between initial centers is 5.156.

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(area)	1.21394	-.23019	-1.04171
Zscore(perimeter)	1.22275	-.26715	-1.00856
Zscore(compactness)	.53414	.41615	-1.05540
Zscore(length_of_kernel)	1.20788	-.36266	-.88234
Zscore(width_of_kernel)	1.12090	-.07273	-1.12321
Zscore(asymmetry_coefficient)	-.06246	-.64447	.81089
Zscore(length_of_kernel_groove)	1.26254	-.67365	-.58237

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(area)	88.832	2	.151	207	586.818	.000
Zscore(perimeter)	88.064	2	.159	207	554.539	.000
Zscore(compactness)	52.681	2	.501	207	105.222	.000
Zscore(length_of_kernel)	81.298	2	.224	207	362.657	.000
Zscore(width_of_kernel)	85.175	2	.187	207	456.171	.000
Zscore(asymmetry_coefficient)	37.082	2	.651	207	56.928	.000
Zscore(length_of_kernel_groove)	83.831	2	.200	207	419.773	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	70.000
	2	75.000
	3	65.000
Valid		210.000
Missing		.000

Keiland Pullen
DSC 441 – Fall 2021

For K=4 clusters:

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore(area)	-1.24326	2.17633	1.08000	-.97519
Zscore(perimeter)	-1.47729	2.02971	.99598	-.79580
Zscore(compactness)	.55022	1.18079	1.19349	-1.95090
Zscore(length_of_kernel)	-1.63980	2.13167	.59013	-.50452
Zscore(width_of_kernel)	-1.00500	2.05021	1.15271	-1.27770
Zscore(asymmetry_coefficient)	-.95188	1.38325	-1.08556	2.18934
Zscore(length_of_kernel_groove)	-1.43459	1.67439	.87273	-.28093

Iteration History[a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.644	1.253	1.083	1.399
2	.390	.567	.121	.191
3	.286	.300	.208	.154
4	.110	.108	.197	.035
5	.028	.088	.154	.000
6	.030	.030	.102	.000
7	.028	.034	.108	.000
8	.030	.000	.069	.000
9	.021	.000	.000	.022
10	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 10. The minimum distance between initial centers is 3.491.

Final Cluster Centers

Cluster			
1	2	3	4

Zscore(area)	-.30702	1.46836	.50514	-1.03958
Zscore(perimeter)	-.35477	1.45479	.55799	-1.00398
Zscore(compactness)	.42528	.66388	.24876	-1.07011
Zscore(length_of_kernel)	-.45800	1.44163	.53326	-.87424
Zscore(width_of_kernel)	-.13425	1.35025	.49198	-1.12385
Zscore(asymmetry_coefficient)	-.73623	-.15112	.11198	.83395
Zscore(length_of_kernel_groove)	-.78115	1.46050	.56658	-.56601

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(area)	62.928	3	.098	206	641.259	.000
Zscore(perimeter)	61.996	3	.112	206	555.001	.000
Zscore(compactness)	36.286	3	.486	206	74.645	.000
Zscore(length_of_kernel)	57.779	3	.173	206	333.741	.000
Zscore(width_of_kernel)	59.546	3	.147	206	404.022	.000
Zscore(asymmetry_coefficient)	27.441	3	.615	206	44.623	.000
Zscore(length_of_kernel_groove)	58.512	3	.162	206	360.201	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	67.000
	2	49.000
	3	30.000
	4	64.000
Valid		210.000
Missing		.000

Keiland Pullen
DSC 441 – Fall 2021

For K=5 clusters:

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
Zscore(area)	2.17633	-.94083	-.73806	.24830	-.05070
Zscore(perimeter)	2.02971	-.83409	-.88003	.45232	-.26746
Zscore(compactness)	1.18079	-1.34572	.69411	-.77440	1.87484
Zscore(length_of_kernel)	2.13167	-1.05974	-1.00557	.65784	-.95592
Zscore(width_of_kernel)	2.05021	-.59994	-.44373	-.07308	.54908
Zscore(asymmetry_coefficient)	1.38325	-1.46200	3.16303	-.70513	-1.28575
Zscore(length_of_kernel_groove)	1.67439	-1.80897	-.83029	.95818	-1.54446

Iteration History[a]

	Change in Cluster Centers				
Iteration	1	2	3	4	5
1	1.809	2.018	1.956	1.159	1.388
2	.284	.159	.063	.243	.330
3	.055	.114	.132	.183	.104
4	.000	.077	.106	.170	.068
5	.022	.079	.072	.071	.035
6	.000	.020	.063	.000	.000
7	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 3.594.

Final Cluster Centers

	Cluster				
	1	2	3	4	5
Zscore(area)	1.45626	-1.08742	-.75669	.47893	-.20755
Zscore(perimeter)	1.44408	-1.04571	-.78371	.54612	-.26019
Zscore(compactness)	.65695	-1.18706	-.18993	.14168	.55314
Zscore(length_of_kernel)	1.43611	-.90606	-.77643	.54258	-.38641
Zscore(width_of_kernel)	1.33994	-1.19871	-.63603	.44951	-.02707

Zscore(asymmetry_coefficient)	-.15042	.40134	1.61155	.16569	-.85882
Zscore(length_of_kernel_groove)	1.45420	-.63393	-.58677	.55343	-.72743

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(area)	47.718	4	.088	205	539.589	.000
Zscore(perimeter)	47.090	4	.101	205	467.684	.000
Zscore(compactness)	29.518	4	.444	205	66.551	.000
Zscore(length_of_kernel)	44.157	4	.158	205	279.644	.000
Zscore(width_of_kernel)	45.547	4	.131	205	348.240	.000
Zscore(asymmetry_coefficient)	25.721	4	.518	205	49.688	.000
Zscore(length_of_kernel_groove)	43.411	4	.172	205	251.718	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	50.000
	2	55.000
	3	19.000
	4	28.000
	5	58.000
Valid		210.000
Missing		.000

Keiland Pullen
DSC 441 – Fall 2021

For K=6 clusters:

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Zscore(area)	.35484	-.90646	-.29471	1.82578	1.47523	-.97519
Zscore(perimeter)	.26089	-.83409	-.55843	1.89188	1.57027	-.79580
Zscore(compactness)	1.17656	-1.08757	2.00180	.10586	.05084	-1.95090
Zscore(length_of_kernel)	-.02377	-.61285	-1.15002	1.99625	1.42297	-.50452
Zscore(width_of_kernel)	.65763	-.96794	.32934	1.36186	1.26655	-1.27770
Zscore(asymmetry_coefficient)	-1.95210	-1.35625	1.02011	-1.19064	1.98316	2.18934
Zscore(length_of_kernel_groove)	-.64514	-.46812	-1.27588	1.58079	1.31222	-.28093

Iteration History[a]

	Change in Cluster Centers					
Iteration	1	2	3	4	5	6
1	1.273	1.183	1.422	1.220	1.490	1.302
2	.221	.153	.141	.146	.248	.159
3	.026	.000	.121	.128	.181	.031
4	.038	.076	.129	.142	.288	.000
5	.060	.088	.325	.037	.111	.038
6	.094	.120	.257	.000	.097	.059
7	.022	.034	.083	.029	.043	.031
8	.000	.000	.144	.066	.114	.065
9	.040	.000	.103	.047	.118	.000
10	.029	.000	.203	.049	.093	.051
11	.050	.000	.000	.000	.084	.000
12	.000	.000	.000	.021	.039	.000
13	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 13. The minimum distance between initial centers is 3.273.

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
Zscore(area)	-.06767	-.86973	-.61240	1.47902	.55366	-1.15549
Zscore(perimeter)	-.09836	-.90611	-.67951	1.46244	.62418	-1.08874
Zscore(compactness)	.53673	-.27092	.25425	.68097	.14567	-1.48158
Zscore(length_of_kernel)	-.20194	-.92369	-.74870	1.44521	.63093	-.90193
Zscore(width_of_kernel)	.10303	-.80429	-.40551	1.36170	.50564	-1.32651
Zscore(asymmetry_coefficient)	-.84958	-.45223	1.59662	-.15948	.13185	.86667
Zscore(length_of_kernel_groove)	-.60092	-.90990	-.64806	1.46329	.70777	-.53892

ANOVA

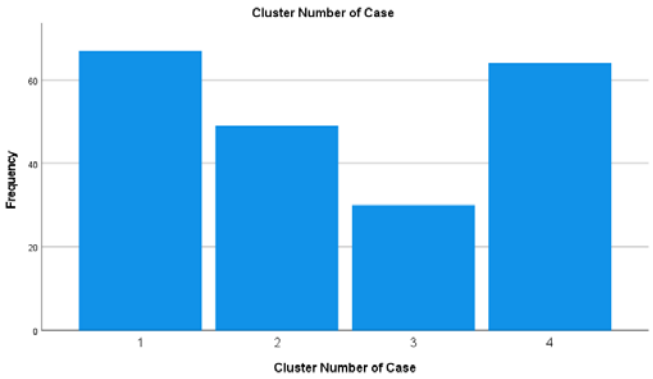
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(area)	39.155	5	.065	204	603.972	.000
Zscore(perimeter)	38.706	5	.076	204	510.487	.000
Zscore(compactness)	25.130	5	.409	204	61.507	.000
Zscore(length_of_kernel)	36.191	5	.137	204	263.269	.000
Zscore(width_of_kernel)	37.674	5	.101	204	372.574	.000
Zscore(asymmetry_coefficient)	22.592	5	.471	204	47.988	.000
Zscore(length_of_kernel_groove)	35.635	5	.151	204	235.847	.000

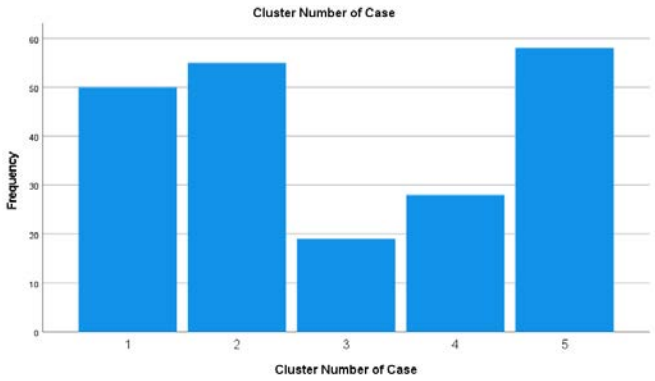
The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	48.000
	2	32.000
	3	16.000
	4	48.000
	5	27.000
	6	39.000

Valid	210.000
Missing	.000





Case Processing Summary

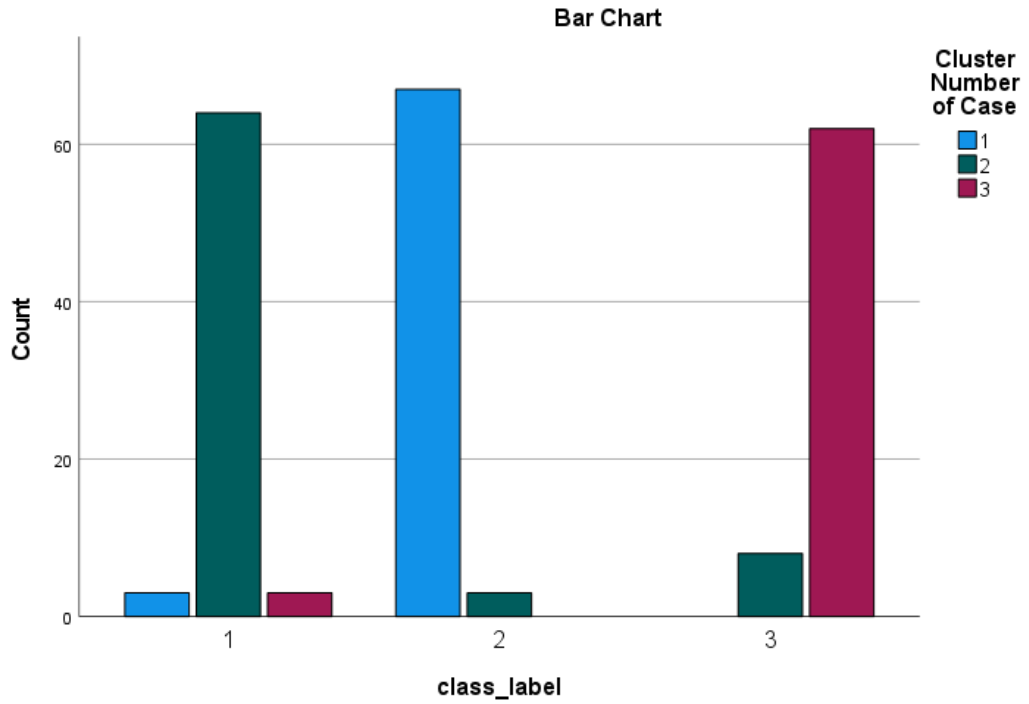
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
class_label * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%
class_label * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%
class_label * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%
class_label * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%
class_label * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%

Keiland Pullen
DSC 441 – Fall 2021

class_label * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			
		1	2	3	Total
class_label	1	3	64	3	70
	2	67	3	0	70
	3	0	8	62	70
Total		70	75	65	210

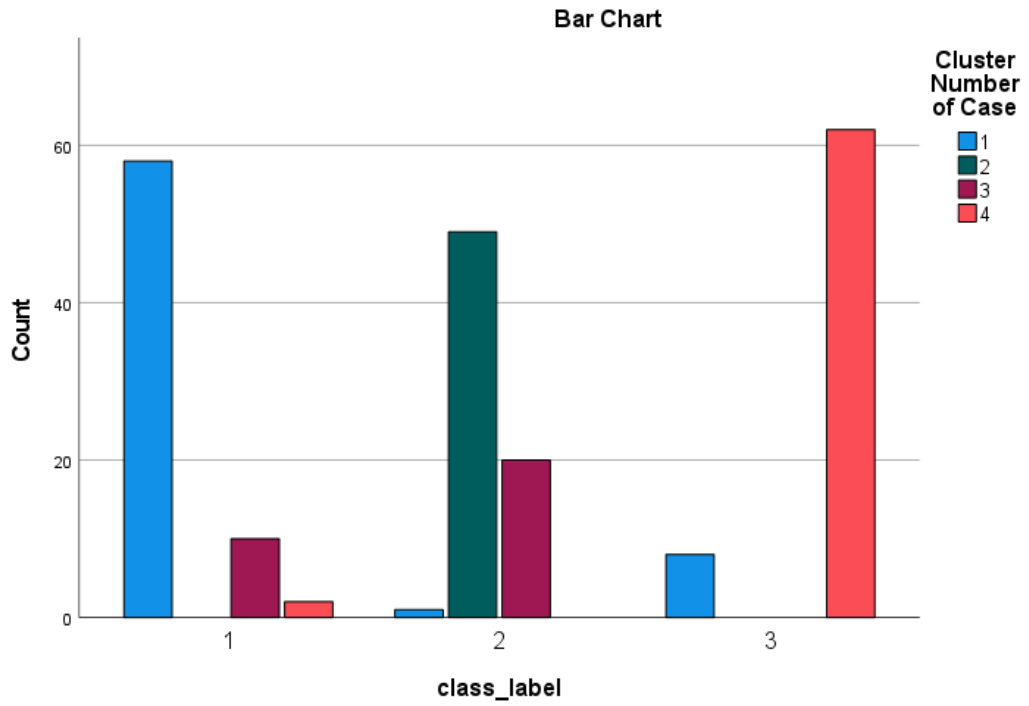


class_label * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				
		1	2	3	4	Total
class_label	1	58	0	10	2	70
	2	1	49	20	0	70
	3	8	0	0	62	70
Total		67	49	30	64	210

Keiland Pullen
DSC 441 – Fall 2021

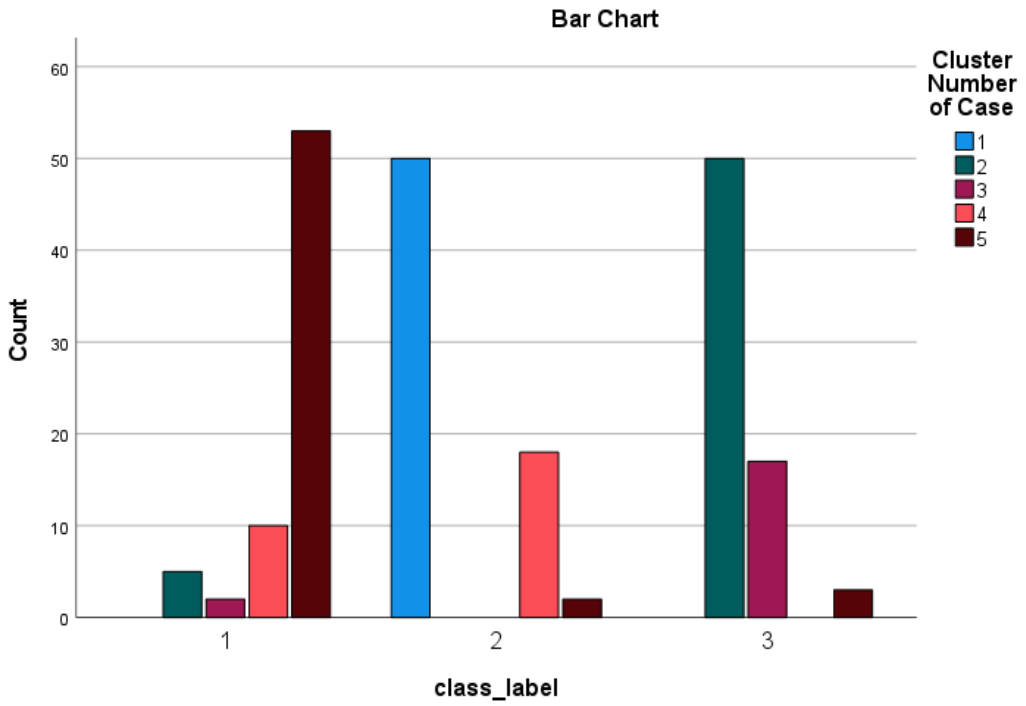


class_label * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case					
		1	2	3	4	5	Total
class_label	1	0	5	2	10	53	70
	2	50	0	0	18	2	70
	3	0	50	17	0	3	70
Total		50	55	19	28	58	210

Keiland Pullen
DSC 441 – Fall 2021



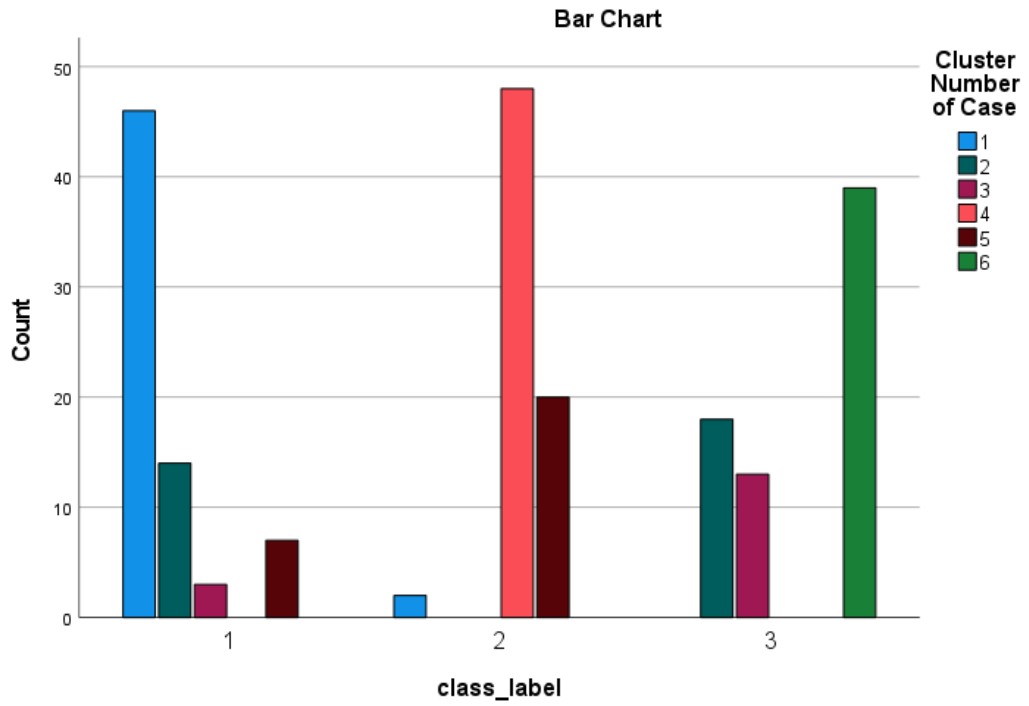
class_label * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case						
		1	2	3	4	5	6	Total
class_label	1	46	14	3	0	7	0	70
	2	2	0	0	48	20	0	70
	3	0	18	13	0	0	39	70
Total		48	32	16	48	27	39	210

Keiland Pullen
DSC 441 – Fall 2021

Assignment 5



Based on the results, in my opinion the best cluster is K=3. One reason is that it appears to be the easiest to explain. Yes, normalization does influence the clustering results.

ii)

i.) Single linkage

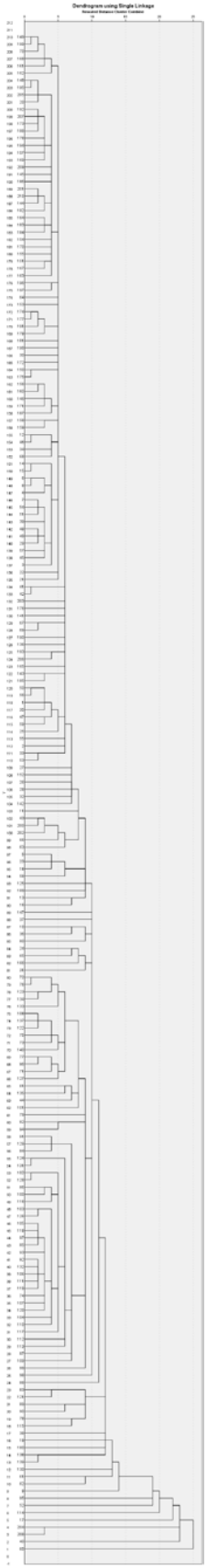
class_label * Single Linkage Crosstabulation

Count

		Single Linkage			
		1	2	3	Total
class_label	1	68	1	1	70
	2	70	0	0	70
	3	68	2	0	70
Total		206	3	1	210

Keiland Pullen
DSC 441 – Fall 2021

Assignment 5



ii.) Complete linkage

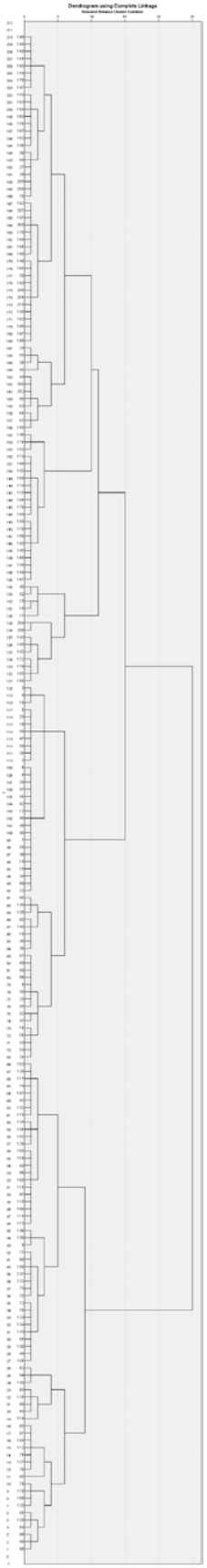
class_label * Complete Linkage Crosstabulation

Count

		Complete Linkage			
		1	2	3	Total
class_label	1	48	2	20	70
	2	4	66	0	70
	3	0	0	70	70
Total		52	68	90	210

Keiland Pullen
DSC 441 – Fall 2021

Assignment 5



Keiland Pullen Assignment 5
DSC 441 – Fall 2021

K-means and Hierarchial clustering were both performed on the original data that contained the PEP=1 filter. This may lead to the some inaccuracies in the final summaries. However, what we have learned is that if there are a specific number of clusters in the data-set, but their class is unknown, then K-means is a better selection. Hierarchial clustering should be used if the number of clusters is known. K-means computes faster given a large number of variables. The dendrogram allows the number of clusters via hierarchical to be easily determined.

Keiland Pullen
DSC 441 – Fall 2021

Problem 3.)

a.)

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class_label
	Independent Variables	Zscore(area), Zscore(perimeter), Zscore(compactness), Zscore(length_of_kernel), Zscore(width_of_kernel), Zscore(asymmetry_coefficient), Zscore(length_of_kernel_grove)
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	Zscore(length_of_kernel_grove), Zscore(perimeter), Zscore(length_of_kernel), Zscore(area), Zscore(width_of_kernel), Zscore(compactness), Zscore(asymmetry_coefficient)
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	5

Risk

Method	Estimate	Std. Error
Resubstitution	.029	.011
Cross-Validation	.105	.021

Growing Method: CRT

Dependent Variable: class_label

Classification

		Predicted		
Observed	1	2	3	Percent Correct

1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT

Dependent Variable: class_label

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class_label
	Independent Variables	Zscore(area), Zscore(perimeter), Zscore(compactness), Zscore(length_of_kernel), Zscore(width_of_kernel), Zscore(asymmetry_coefficient), Zscore(length_of_kernel_grove)
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	Zscore(length_of_kernel_grove), Zscore(perimeter), Zscore(length_of_kernel), Zscore(area), Zscore(width_of_kernel), Zscore(compactness), Zscore(asymmetry_coefficient)
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.100	.021

Growing Method: CRT

Dependent Variable: class_label

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

Growing Method: CRT

Dependent Variable: class_label

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class_label
	Independent Variables	Zscore(area), Zscore(perimeter), Zscore(compactness), Zscore(length_of_kernel), Zscore(width_of_kernel), Zscore(asymmetry_coefficient), Zscore(length_of_kernel_grove)
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50

Results	Independent Variables Included	Zscore(length_of_kernel_grove), Zscore(perimeter), Zscore(length_of_kernel), Zscore(area), Zscore(width_of_kernel), Zscore(compactness), Zscore(asymmetry_coefficient)
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.148	.024

Growing Method: CRT

Dependent Variable: class_label

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

Growing Method: CRT

Dependent Variable: class_label

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class_label
	Independent Variables	Zscore(area), Zscore(perimeter), Zscore(compactness), Zscore(length_of_kernel), Zscore(width_of_kernel), Zscore(asymmetry_coefficient), Zscore(length_of_kernel_grove)

	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	20
Results	Independent Variables Included	Zscore(length_of_kernel_grove), Zscore(perimeter), Zscore(length_of_kernel), Zscore(area), Zscore(width_of_kernel), Zscore(compactness), Zscore(asymmetry_coefficient)
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.095	.020

Growing Method: CRT

Dependent Variable: class_label

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

Growing Method: CRT

Dependent Variable: class_label

Keiland Pullen
DSC 441 – Fall 2021

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class_label
	Independent Variables	Zscore(area), Zscore(perimeter), Zscore(compactness), Zscore(length_of_kernel), Zscore(width_of_kernel), Zscore(asymmetry_coefficient), Zscore(length_of_kernel_grove)
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	25
	Minimum Cases in Child Node	5
Results	Independent Variables Included	Zscore(length_of_kernel_grove), Zscore(perimeter), Zscore(length_of_kernel), Zscore(area), Zscore(width_of_kernel), Zscore(compactness), Zscore(asymmetry_coefficient)
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	4

Risk

Method	Estimate	Std. Error
Resubstitution	.052	.015
Cross-Validation	.110	.022

Growing Method: CRT

Dependent Variable: class_label

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	62	1	7	88.6%
2	2	68	0	97.1%

3	1	0	69	98.6%
Overall Percentage	31.0%	32.9%	36.2%	94.8%

Growing Method: CRT

Dependent Variable: class_label

The best tree choice in my opinion is the first option as it generates 5 levels with 13 nodes and 7 terminal nodes. Neither of the other trees displayed any significant improvement.

b.)
The confusion matrix or misclassification matrix is for the "best tree" is as follows:

Classification

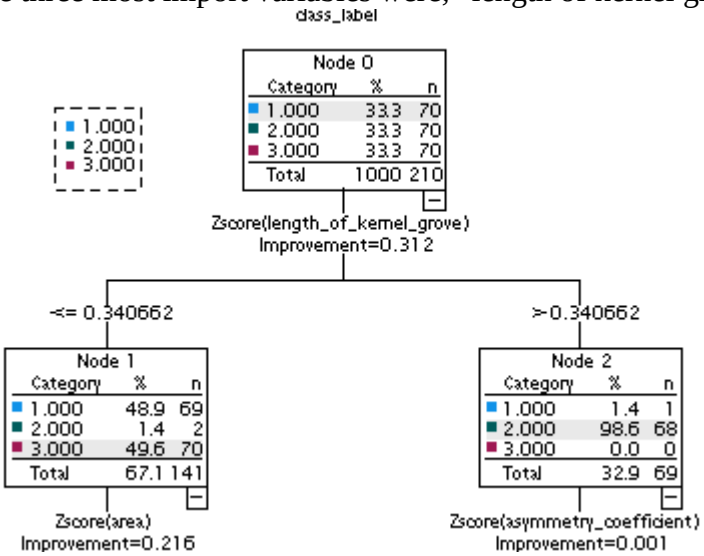
	Predicted			
Observed	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT

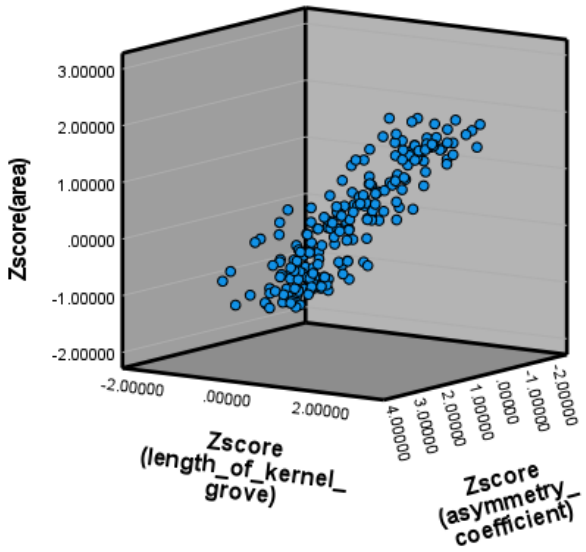
Dependent Variable: class_label

To interpret, in the first row, the 68 gives the count of how many cases were classified in class 1 and were in class 1. The 1 give the number that were classified in class 1 but were in class 2. The final 1 gives the number that were classified in class 1 but were in class 3. The second row: the 2 gives the number that were classified in class 1 but should have been in class 2. The 68 gives the number that were classified in class 2 and were in class 2. The third row: the 2 gives the number that was classified in class 1 but were in class 3. The 0 means that there were no class 2/class 3 errors and the 68 means that there were 68 cases classified in class 3 correctly.

c.)
The three most import variables were, "length of kernel grove, area and asymmetry_coefficient":



d.)



The image above appears as one large cluster of data points from the Area, Length of Kernel and Asymmetry variables.