REGISTRATION NUMBER: N15/3/1007/020

NAME: SHARON GICHANE

SPSS INTERNAL ATTACHMENT 2024.

PROGRAMME: BSC STATISTICS.

QUESTIONS.

Question One

The data set below was obtained in 2014 General Social Services Survey.

a) Obtain the frequency distribution of scale variables .

FREQUENCIES

[DataSet0]

Statistics

		HIGHEST_YEAR_OF_SCHOOL_COMPLETED	NUMBER_OF_BROTHERS_AND_SISTERS	NUMBER_OF_CHILDREN	TOTAL_INCOME	RESPONDENTS_INCOME	WEIGHT_VARIABLE	HEIGHT_VARIABLE
N	Valid	142	142	142	142	142	142	142
	Missing	0	0	0	0	0	0	0

FREQUENCY TABLES

HIGHEST_YEAR_OF_SCHOOL_COMPLETED

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	1	.7	.7	.7
	6	1	.7	.7	1.4
	8	3	2.1	2.1	3.5
	9	1	.7	.7	4.2
	10	6	4.2	4.2	8.5
	11	9	6.3	6.3	14.8
	12	15	10.6	10.6	25.4
	13	7	4.9	4.9	30.3
	14	14	9.9	9.9	40.1
	15	9	6.3	6.3	46.5
	16	44	31.0	31.0	77.5
	17	9	6.3	6.3	83.8
	18	11	7.7	7.7	91.5
	19	1	.7	.7	92.3
	20	11	7.7	7.7	100.0
	Total	142	100.0	100.0	

NUMBER_OF_BROTHERS_AND_SISTERS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6	4.2	4.2	4.2
	1	30	21.1	21.1	25.4
	2	27	19.0	19.0	44.4
	3	35	24.6	24.6	69.0
	4	12	8.5	8.5	77.5
	5	8	5.6	5.6	83.1
	6	4	2.8	2.8	85.9
	7	3	2.1	2.1	88.0
	8	8	5.6	5.6	93.7
	9	3	2.1	2.1	95.8
	10	2	1.4	1.4	97.2
	11	2	1.4	1.4	98.6
	13	1	.7	.7	99.3
	14	1	.7	.7	100.0
	Total	142	100.0	100.0	

NUMBER_OF_CHILDREN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	52	36.6	36.6	36.6
	1	22	15.5	15.5	52.1
	2	36	25.4	25.4	77.5
	3	23	16.2	16.2	93.7
	4	5	3.5	3.5	97.2
	5	2	1.4	1.4	98.6
	6	1	.7	.7	99.3
	8	1	.7	.7	100.0
	Total	142	100.0	100.0	

TOTAL_INCOME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	2.1	2.1	2.1
	3	2	1.4	1.4	3.5
	6	1	.7	.7	4.2
	7	1	.7	.7	4.9
	8	1	.7	.7	5.6
	9	6	4.2	4.2	9.9
	10	2	1.4	1.4	11.3
	11	2	1.4	1.4	12.7
	12	2	1.4	1.4	14.1
	13	4	2.8	2.8	16.9
	14	5	3.5	3.5	20.4
	15	8	5.6	5.6	26.1
	16	6	4.2	4.2	30.3
	17	3	2.1	2.1	32.4
	18	6	4.2	4.2	36.6
	19	7	4.9	4.9	41.5
	20	7	4.9	4.9	46.5
	21	10	7.0	7.0	53.5
	22	15	10.6	10.6	64.1
	23	6	4.2	4.2	68.3
	24	6	4.2	4.2	72.5
	25	25	17.6	17.6	90.1
	26	10	7.0	7.0	97.2
	98	4	2.8	2.8	100.0
	Total	142	100.0	100.0	

RESPONDENTS_INCOME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	45	31.7	31.7	31.7
	2	1	.7	.7	32.4
	6	1	.7	.7	33.1
	8	2	1.4	1.4	34.5
	10	3	2.1	2.1	36.6
	11	1	.7	.7	37.3
	12	1	.7	.7	38.0
	13	5	3.5	3.5	41.5
	14	7	4.9	4.9	46.5
	15	6	4.2	4.2	50.7
	16	4	2.8	2.8	53.5
	17	1	.7	.7	54.2
	18	9	6.3	6.3	60.6
	19	8	5.6	5.6	66.2
	20	5	3.5	3.5	69.7
	21	12	8.5	8.5	78.2
	22	6	4.2	4.2	82.4
	23	6	4.2	4.2	86.6
	24	3	2.1	2.1	88.7
	25	7	4.9	4.9	93.7
	26	8	5.6	5.6	99.3
	98	1	.7	.7	100.0
	Total	142	100.0	100.0	

WEIGHT_VARIABLE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	39	27.5	27.5	27.5
	1	15	10.6	10.6	38.0
	1	48	33.8	33.8	71.8
	2	12	8.5	8.5	80.3
	2	18	12.7	12.7	93.0
	2	1	.7	.7	93.7
	2	5	3.5	3.5	97.2
	3	1	.7	.7	97.9
	3	1	.7	.7	98.6
	4	1	.7	.7	99.3
	6	1	.7	.7	100.0
	Total	142	100.0	100.0	

HEIGHT_VARIABLE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	39	27.5	27.5	27.5
	1	15	10.6	10.6	38.0
	1	48	33.8	33.8	71.8
	1	12	8.5	8.5	80.3
	1	18	12.7	12.7	93.0
	2	1	.7	.7	93.7
	2	5	3.5	3.5	97.2
	2	1	.7	.7	97.9
	3	1	.7	.7	98.6
	3	1	.7	.7	99.3
	5	1	.7	.7	100.0
	Total	142	100.0	100.0	

b) Create the cross tabulated table for
   i) Household type and marital status
   ii) Marital status and race of respondents
   iii) Perform a chi-square test to determine the association between household type and marital status
   iv) Perform a chi-square test to determine the association between race of respondent and marital status
   v) Perform ANOVA test to compare the income of married, divorced, separated and single.
   vi) Plot two suitable graphs to display the distribution of highest school completion
   vii) Obtain the variance covariance matrix for the scale variables

i)CROSSTABS
  /TABLES=HOUSEHOLD_TYPE BY MARITAL_STATUS
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.

Crosstabs

[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
HOUSEHOLD_TYPE * MARITAL_STATUS	142	100.0%	0	0.0%	142	100.0%

HOUSEHOLD_TYPE * MARITAL_STATUS Crosstabulation

Count

		MARITAL_STATUS						Total
		1	2	3	4	5	9	
HOUSEHOLD_TYPE	1	39	0	0	0	4	0	43
	2	0	0	3	3	6	1	13
	3	0	1	3	1	10	0	15
	4	2	7	8	2	22	0	41
	5	0	1	1	0	3	0	5
	11	18	0	0	0	1	0	19
	13	0	0	1	0	1	0	2
	15	0	0	0	0	2	0	2
	18	0	0	1	0	1	0	2
Total		59	9	17	6	50	1	142

ii)CROSSTABS
/TABLES=MARITAL_STATUS BY RACE_OF_RESPONDANTS
/FORMAT=AVALUE TABLES
/CELLS=COUNT
/COUNT ROUND CELL.

Crosstabs

[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent

HOUSEHOLD_TYPE * MARITAL_STATUS	142	100.0%	0	0.0%	142	100.0%

HOUSEHOLD_TYPE * MARITAL_STATUS Crosstabulation

Count

		MARITAL_STATUS						Total
		1	2	3	4	5	9	
HOUSEHOLD_TYPE	1	39	0	0	0	4	0	43
	2	0	0	3	3	6	1	13
	3	0	1	3	1	10	0	15
	4	2	7	8	2	22	0	41
	5	0	1	1	0	3	0	5
	11	18	0	0	0	1	0	19
	13	0	0	1	0	1	0	2
	15	0	0	0	0	2	0	2
	18	0	0	1	0	1	0	2
Total		59	9	17	6	50	1	142

iii)CROSSTABS
/TABLES=MARITAL_STATUS BY RACE_OF_RESPONDANTS
/FORMAT=AVALUE TABLES
/CELLS=COUNT
/COUNT ROUND CELL.

Crosstabs

[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
MARITAL_STATUS * RACE_OF_RESPONDANTS	142	100.0%	0	0.0%	142	100.0%

MARITAL_STATUS * RACE_OF_RESPONDANTS Crosstabulation

Count

		RACE_OF_RESPONDANTS			Total
		1	2	3	
MARITAL_STATUS	1	36	5	18	59
	2	6	3	0	9
	3	11	3	3	17
	4	3	2	1	6
	5	29	11	10	50
	9	0	1	0	1
Total		85	25	32	142

iv)CROSSTABS
/TABLES=HOUSEHOLD_TYPE BY MARITAL_STATUS
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT
/COUNT ROUND CELL.

Crosstabs

[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
HOUSEHOLD_TYPE * MARITAL_STATUS	142	100.0%	0	0.0%	142	100.0%

HOUSEHOLD_TYPE * MARITAL_STATUS Crosstabulation

Count

		MARITAL_STATUS						Total
		1	2	3	4	5	9	
HOUSEHOLD_TYPE	1	39	0	0	0	4	0	43
	2	0	0	3	3	6	1	13
	3	0	1	3	1	10	0	15
	4	2	7	8	2	22	0	41
	5	0	1	1	0	3	0	5
	11	18	0	0	0	1	0	19
	13	0	0	1	0	1	0	2
	15	0	0	0	0	2	0	2
	18	0	0	1	0	1	0	2
Total		59	9	17	6	50	1	142

v) Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	147.687[a]	40	.000
Likelihood Ratio	165.174	40	.000
Linear-by-Linear Association	.010	1	.919
N of Valid Cases	142		

a. 44 cells (81.5%) have expected count less than 5. The minimum expected count is .01.

```
CROSSTABS
  /TABLES=MARITAL_STATUS BY RACE_OF_RESPONDANTS
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Crosstabs
[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
MARITAL_STATUS * RACE_OF_RESPONDANTS	142	100.0%	0	0.0%	142	100.0%

MARITAL_STATUS * RACE_OF_RESPONDANTS Crosstabulation
Count

		RACE_OF_RESPONDANTS			Total
		1	2	3	
MARITAL_STATUS	1	36	5	18	59
	2	6	3	0	9
	3	11	3	3	17
	4	3	2	1	6
	5	29	11	10	50
	9	0	1	0	1
Total		85	25	32	142

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	14.513[a]	10	.151
Likelihood Ratio	15.385	10	.119
Linear-by-Linear Association	.039	1	.844
N of Valid Cases	142		

a. 10 cells (55.6%) have expected count less than 5. The minimum expected count is .18.

Oneway

[DataSet0]

ANOVA

TOTAL_INCOME

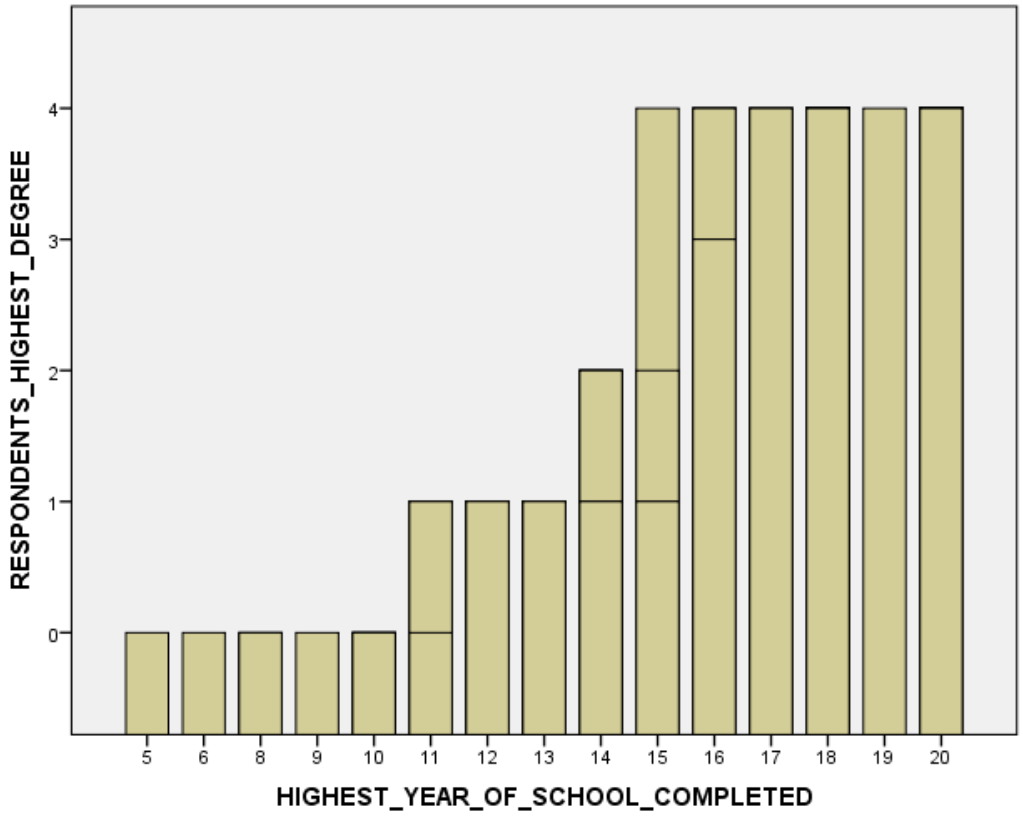
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1221.782	5	244.356	1.178	.323
Within Groups	28209.577	136	207.423		
Total	29431.359	141			

vi)* Chart Builder.

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset"
VARIABLES=HIGHEST_YEAR_OF_SCHOOL_COMPLETED
RESPONDENTS_HIGHEST_DEGREE MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: HIGHEST_YEAR_OF_SCHOOL_COMPLETED=col(source(s),
name("HIGHEST_YEAR_OF_SCHOOL_COMPLETED"), unit.category())
  DATA: RESPONDENTS_HIGHEST_DEGREE=col(source(s),
name("RESPONDENTS_HIGHEST_DEGREE"), unit.category())
  GUIDE: axis(dim(1), label("HIGHEST_YEAR_OF_SCHOOL_COMPLETED"))
  GUIDE: axis(dim(2), label("RESPONDENTS_HIGHEST_DEGREE"))
  ELEMENT:
interval(position(HIGHEST_YEAR_OF_SCHOOL_COMPLETED*RESPONDENTS_HIGHES
T_DEGREE), shape.interior(shape.square)) END GPL.
```

GGraph

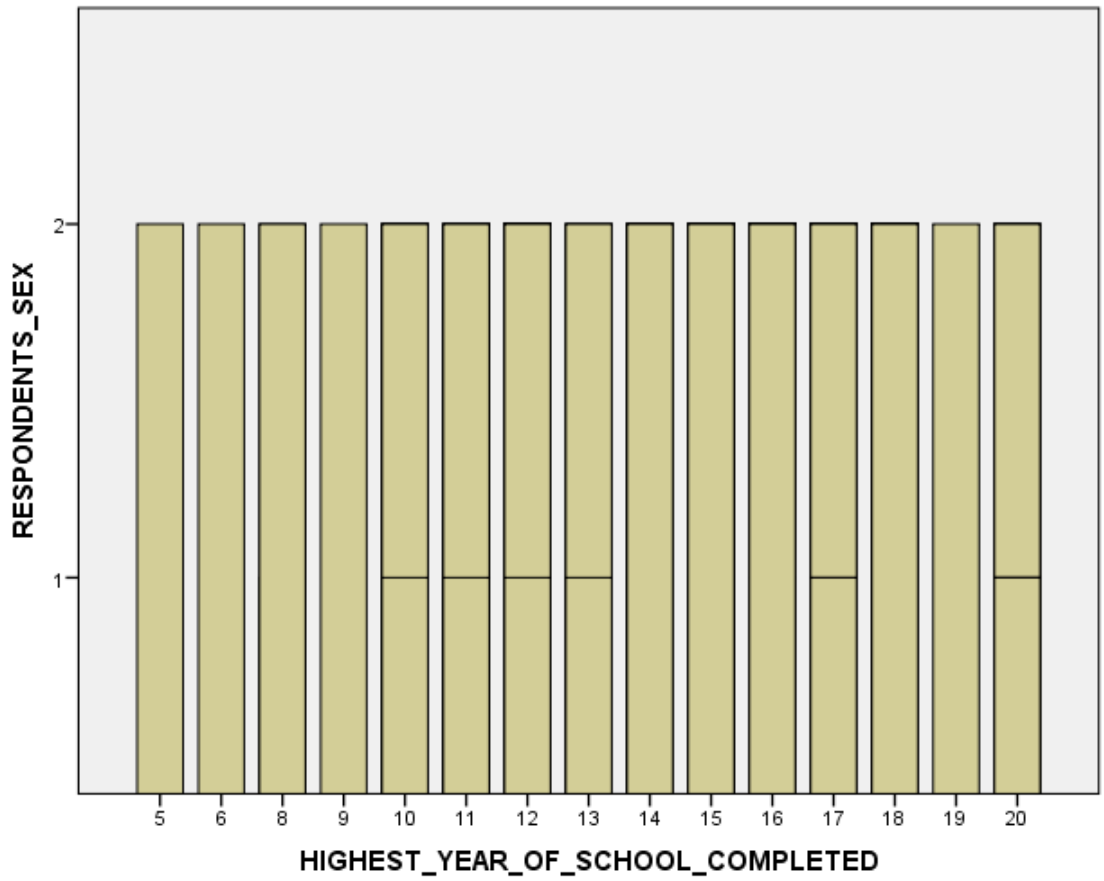
[DataSet0]




```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=HIGHEST_YEAR_OF_SCHOOL_COMPLETED RESPONDENTS_SEX MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: HIGHEST_YEAR_OF_SCHOOL_COMPLETED=col(source(s), name("HIGHEST_YEAR_OF_SCHOOL_COMPLETED"), unit.category())
  DATA: RESPONDENTS_SEX=col(source(s), name("RESPONDENTS_SEX"), unit.category())
  GUIDE: axis(dim(1), label("HIGHEST_YEAR_OF_SCHOOL_COMPLETED"))
  GUIDE: axis(dim(2), label("RESPONDENTS_SEX"))
  ELEMENT: interval(position(HIGHEST_YEAR_OF_SCHOOL_COMPLETED*RESPONDENTS_SEX), shape.interior(shape.square))
END GPL.
```

GGraph

[DataSet0]



viii) CORRELATIONS
/VARIABLES=HIGHEST_YEAR_OF_SCHOOL_COMPLETED NUMBER_OF_BROTHERS_AND_SISTERS NUMBER_OF_CHILDREN TOTAL_INCOME RESPONDENTS_INCOME WEIGHT_VARIABLE HEIGHT_VARIABLE
/PRINT=TWOTAIL NOSIG
/STATISTICS XPROD
/MISSING=PAIRWISE.

Correlations

[DataSet0]

Correlations

		HIGHEST_YEAR_OF_SCHOOL_COMPLETED	NUMBER_OF_BROTHERS_AND_SISTERS	NUMBER_OF_CHILDREN	TOTAL_INCOME	RESPONDENTS_INCOME	WEIGHT_VARIABLE	HEIGHT_VARIABLE
HIGHEST_YEAR_OF_SCHOOL_COMPLETED	Pearson Correlation	1	-.301**	-.291**	.115	.264**	.062	.062
	Sig. (2-tailed)		.000	.000	.172	.002	.464	.465
	Sum of Squares and Cross-products	1298.232	-353.775	-181.810	713.077	1360.789	18.019	16.107
	Covariance	9.207	-2.509	-1.289	5.057	9.651	.128	.114
	N	142	142	142	142	142	142	142
NUMBER_OF_BROTHERS_AND_SISTERS	Pearson Correlation	-.301**	1	.321**	.110	-.235**	.009	.009
	Sig. (2-tailed)	.000		.000	.191	.005	.912	.914
	Sum of Squares and Cross-products	-353.775	1061.915	181.366	616.408	-1094.296	2.454	2.154
	Covariance	-2.509	7.531	1.286	4.372	-7.761	.017	.015
	N	142	142	142	142	142	142	142
NUMBER_OF_CHILDREN	Pearson Correlation	-.291**	.321**	1	.298**	-.142	.056	.057
	Sig. (2-tailed)	.000	.000		.000	.092	.507	.503
	Sum of Squares and Cross-products	-181.810	181.366	301.246	886.063	-352.718	7.872	7.117
	Covariance	-1.289	1.286	2.136	6.284	-2.502	.056	.050
	N	142	142	142	142	142	142	142
TOTAL_INCOME	Pearson Correlation	.115	.110	.298**	1	.067	.182*	.182*
	Sig. (2-tailed)	.172	.191	.000		.429	.030	.030
	Sum of Squares and Cross-products	713.077	616.408	886.063	29431.359	1640.930	252.870	226.116
	Covariance	5.057	4.372	6.284	208.733	11.638	1.793	1.604
	N	142	142	142	142	142	142	142
RESPONDENTS_INCOME	Pearson Correlation	.264**	-.235**	-.142	.067	1	.050	.050
	Sig. (2-tailed)	.002	.005	.092	.429		.558	.556
	Sum of Squares and Cross-products	1360.789	-1094.296	-352.718	1640.930	20479.465	57.245	51.603
	Covariance	9.651	-7.761	-2.502	11.638	145.244	.406	.366
	N	142	142	142	142	142	142	142
WEIGHT_VARIABLE	Pearson Correlation	.062	.009	.056	.182*	.050	1	1.000**
	Sig. (2-tailed)	.464	.912	.507	.030	.558		.000
	Sum of Squares and Cross-products	18.019	2.454	7.872	252.870	57.245	65.284	58.444
	Covariance	.128	.017	.056	1.793	.406	.463	.414
	N	142	142	142	142	142	142	142
HEIGHT_VARIABLE	Pearson Correlation	.062	.009	.057	.182*	.050	1.000**	1
	Sig. (2-tailed)	.465	.914	.503	.030	.556	.000	
	Sum of Squares and Cross-products	16.107	2.154	7.117	226.116	51.603	58.444	52.322
	Covariance	.114	.015	.050	1.604	.366	.414	.371
	N	142	142	142	142	142	142	142

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

```
T-TEST GROUPS=MARITAL_STATUS(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=NUMBER_OF_CHILDREN
  /CRITERIA=CI(.95).
```

c) Perform a test to compare the number of children for married and Widow.

T-Test

[DataSet0]

Group Statistics

	MARITAL_STATUS	N	Mean	Std. Deviation	Std. Error Mean
NUMBER_OF_CHILDREN	Married	59	1.69	1.355	.176
	Widow	9	2.67	2.236	.745

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
NUMBER_OF_CHILDREN	Equal variances assumed	.946	.334	-1.823	66	.073	-.972	.533	-2.036	.093
	Equal variances not assumed			-1.269	8.918	.237	-.972	.766	-2.707	.763

The T-test shows no significant difference in the number of children between married and widow respondents. The T-value is -1.823(df= 66 , p =0.073). The mean difference of -0.972 is not significant at the level of 0.05.

Question Two

Using the performance data

i) Compute a new variable called average performance which is the mean of English Read and Mathematics
ii) Standardize the average performance variable
iii) Fit a regression model for average performance using the other variables except the county.

```
COMPUTE average_perfomance= (English + Read + Math) / 3.
EXECUTE.
DESCRIPTIVES VARIABLES=avg_perfomance
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.
```

ii) Descriptives

[DataSet0]

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
avg_perfomance	420	407.84	472.74	441.3605	9.93778
Valid N (listwise)	420				

iii) REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT avg_perfomance
/METHOD=ENTER Students Teachers Calworks Lunch Computer Expenditure Income English Read Math.

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Math, Computer, Expenditure, English, Calworks, Income, Lunch, Students, Read, Teachers[b]	.	Enter

a. Dependent Variable: avg_perfomance

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.00000

a. Predictors: (Constant), Math, Computer, Expenditure, English, Calworks, Income, Lunch, Students, Read, Teachers

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41380.245	10	4138.024	.	.[b]
	Residual	.000	409	.000		
	Total	41380.245	419			

a. Dependent Variable: avg_perfomance

b. Predictors: (Constant), Math, Computer, Expenditure, English, Calworks, Income, Lunch, Students, Read, Teachers

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.296E-012	.000		.000	1.000
	Students	-1.019E-017	.000	.000	.000	1.000
	Teachers	2.478E-016	.000	.000	.000	1.000
	Calworks	3.554E-016	.000	.000	.000	1.000
	Lunch	4.310E-016	.000	.000	.000	1.000
	Computer	-1.936E-017	.000	.000	.000	1.000
	Expenditure	5.910E-018	.000	.000	.000	1.000
	Income	4.010E-015	.000	.000	.000	1.000
	English	.333	.000	.613	557258836.031	.000
	Read	.333	.000	.674	284488422.025	.000
	Math	.333	.000	.629	345161730.244	.000

a. Dependent Variable: avg_perfomance

+

iv) Interpret the results

The results show an excellent model fit with both R and R_ squared values equal to 1. This means that 100% of the variance in average performance is explained by predictors. However, all regression coefficients for the predictors (students, teachers, calworks, lunch, computer, expenditure, income, english, math) are zero and not significant (p=1.00). The standardized average performance (Zscore) is highly significant($p<0.001$). This suggests potential multicollinearlity. While the model fit is perfect, the non-significant predictors indicate possible problems with the data or the model validity.