/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* \*/

/\* Title: Final Assessment II; Analysis of UK institutions \*/

/\* \*/

/\* Date: \*/

/\* TASKS \*/

/\* Effect of Region on Guided Learning hours per learner(GLH/L) \*/

/\* Effect of type of institution on GLH/L \*/

/\* Effect of size of institution on GLH/L \*/

/\* Effect of Year on GLH/L \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* macros and filepath \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options locale = English\_UnitedKingdom;

%let dir = %str('/home/u60904963/FInal Assessment');

%let sixformfile = %str(IMAT5168-Sixthformfile.csv);

%let FECollegefile = %str(IMAT5168-FECollegeDoc.csv);

filename cwd "&dir";

%let Sixformdata = FORMCOLLEGE;

%let RefittedSixform= SIXFORMCOLLEGE;

%let FECollegedata = COLLEGE;

%let FECollegeRefitted = FECOLLEGE;

%let fuseddata = COLLEGES;

%let Newlyfused = COLLEGES2;

%let Newlyfused2 = TRNSDATA;

%let Newlyfused3 = TRNSDATA2;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing 6th College dataset \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

DATA &Sixformdata;

LENGTH

College $ 32

Region $ 24

GLH1 8

Learners1 8

GLH2 8

Learners2 8

GLH3 8

Learners3 8

;

FORMAT

College $CHAR32.

Region $CHAR24.

GLH1 BEST8.

Learners1 BEST6.

GLH2 BEST8.

Learners2 BEST6.

GLH3 BEST8.

Learners3 BEST8.

;

INFILE

cwd (&sixformfile)

LRECL=108

ENCODING="WLATIN1"

TERMSTR=CRLF

DLM=','

MISSOVER

DSD

FIRSTOBS=2

;

INPUT

College : $CHAR32.

Region : $CHAR24.

GLH1 : BEST8.

Learners1 : BEST6.

GLH2 : BEST8.

Learners2 : BEST6.

GLH3 : BEST8.

Learners3 : BEST6.

;

label

College = 'Six form College'

Region = 'England Region'

GLH1 = 'Guided Learning Hours for Year 1'

GLH2 = 'Guided Learning Hours for Year 2'

GLH3 = 'Guided Learning Hours for Year 3'

Learners1 = 'Number of Students for Year 1'

Learners2 = 'Number of Students for Year 2'

Learners3 = 'Number of Students for Year 3'

;

if upcase(College) in ('.', '') then delete;

RUN;

proc print data=&Sixformdata label;

FORMAT

College

Region

GLH1

Learners1

GLH2

Learners2

GLH3

Learners3

;

run;

proc print data=&Sixformdata label;

FORMAT

College

Region

GLH1

Learners1

GLH2

Learners2

GLH3

Learners3

;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Validate dataset \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods exclude enginehost;

proc contents data=&Sixformdata;

run;

ods select all;

/\*The data's content verifies that it was imported appropriately. There are 110 variables and 8 observations.

The format is also in the proper sequence.\*/

proc print data=&Sixformdata; /\*print the dataset\*/

run;

proc freq data = &Sixformdata;

tables College/missing;

run;

/\*College has a row labelled 'Total' in the tables. This is unnecessary and should be expunged.\*/

/\*New dataset with 'Total' expunged\*/

Data &RefittedSixform;

Set &Sixformdata;

Where College <> 'Total'; /\*Remove Total from the dataset\*/

run;

/\*The new dataset now without the row 'Total'\*/

/\*Check for anomalies in the data\*/

proc freq data = &RefittedSixform;

tables College/missing;

run;

/\*\*Other regions in the institutions column may need to be deleted to allow for adequate analysis, according to the analysis.

This is taken into account afterwards\*\*/

proc freq data = &RefittedSixform;

tables Region/missing;

run;

/\*There are 9 observations missing from the Region column. This is taken into account afterwards\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*import further college dataset \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

DATA &FECollegedata;

LENGTH

College $ 32

Region $ 24

GLH1 8

Learners1 8

GLH2 8

Learners2 8

GLH3 8

Learners3 8

;

FORMAT

College $CHAR32.

Region $CHAR24.

GLH1 BEST8.

Learners1 BEST6.

GLH2 BEST8.

Learners2 BEST6.

GLH3 BEST8.

Learners3 BEST8.

;

INFILE

cwd (&FECollegefile)

LRECL=108

ENCODING="WLATIN1"

TERMSTR=CRLF

DLM=','

MISSOVER

DSD

FIRSTOBS=2

;

INPUT

College : $CHAR32.

Region : $CHAR24.

GLH1 : BEST8.

Learners1 : BEST6.

GLH2 : BEST8.

Learners2 : BEST6.

GLH3 : BEST8.

Learners3 : BEST6.

;

if upcase(College) in ('.', '') then delete;

label

College = 'Further College'

Region = 'England Region'

GLH1 = 'Guided Learning Hours for Year 1'

GLH2 = 'Guided Learning Hours for Year 2'

GLH3 = 'Guided Learning Hours for Year 3'

Learners1 = 'Number of Students for Year 1'

Learners2 = 'Number of Students for Year 2'

Learners3 = 'Number of Students for Year 3'

;

RUN;

proc print data=&FECollegedata label;

format

College

Region

GLH1

Learners1

GLH2

Learners2

GLH3

Learners3

;

run;

/\*\*\*Examine the data for any duplicates or anomalies.\*\*\*/

/\*\*\*Verify that the data has been imported accurately and that it is in the correct format.\*/

ods exclude enginehost;

proc contents data=&FECollegedata;

run;

ods select all;

/\*There are 263 observations and 8 variables, indicating that the data was properly imported.

All of the informats and formats are in the correct order.

as an example, college is a character variableÂ as revealedÂ by the content procedure\*/

proc freq data = &FECollegedata;

tables College/missing;

run;

/\*\*\*Result showed Other colleges are totalled up and shown in the institution column.

This may also need to be removed because we're not clear which of those colleges are related to further colleges.

as an example, East of Midland colleges does not specify if it is a summation of furtherÂ colleges in East of EnglandÂ or a composite of all other

East of England institutions. As a result, it may be necessary to remove this from the data.\*\*\*/

/\*Remove some observations\*\*\*\*\*

There are columns for institutionsÂ like North East Colleges, in the analysis. This does not specify whether they are in sixth form or further college.

Also, the 'FE College' rowÂ in the College column has no entry for Region and hence does not relate to any region,

I have considered removing it in the same way that the 'Total' field in the 'Sixform' dataset was removed.\*\*/

Data &FECollegeRefitted;

Set &FECollegedata;

if upcase(Region) in ('.', '') then delete;; /\*Remove from the dataset\*/

run;

/\*9 obs have been removed\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Merge the dataset

/\* Sort the data \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*This is a comparison exercise between these two colleges.

As a result, the two datasets imported previously must be merged into one dataset or file to enable for that comparison.

To merge the datasets, they must first be sorted by a common denominator, in this case 'college.'\*\*\*\*/

proc sort data=&RefittedSixform; /\*sort Sixform college dataset\*/

by College;

run;

proc print data=&RefittedSixform; /\*print the results\*/

run;

proc sort data=&FECollegeRefitted; /\*sort further college dataset\*/

by College;

run;

proc print data=&FECollegeRefitted; /\*print the results\*/

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Merge the two datasets \*/

/\* Delete observations

/\* Create New total column \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

DATA &fuseddata; /\*Name of the merged dataset\*/

merge &RefittedSixform &FECollegeRefitted;

by College; /\*merge by this primary key\*/

label /\*define the labels for the various columns\*/

College = 'Type of College'

Region = 'England Region'

GLH1 = 'Guided Learning Hours for Year 1'

GLH2 = 'Guided Learning Hours for Year 2'

GLH3 = 'Guided Learning Hours for Year 3'

Learners1 = 'Number of Students for Year 1'

Learners2 = 'Number of Students for Year 2'

Learners3 = 'Number of Students for Year 3'

TOTAL\_GLH = 'Total Guided Learning Hours'

;

if College='East Midlands' then delete;

if College='East of England' then delete;

if College='Greater London' then delete;

if College='North East' then delete;

if College='North West' then delete;

if College='South East' then delete;

if College='South West' then delete;

if College='West Midlands' then delete;

if College='Yorkshire and the Humber' then delete;

;

TOTAL\_GLH = GLH1+GLH2+GLH3; /\*Create a new column to add all GLHs\*/

run;

/\*As noted in the prior analysis, the aforementioned were removed.

They have no clear definition as to whether they are related to a single college or a collection of colleges.

They are expungedÂ because the goal is to compare two colleges.\*/

proc print data=&fuseddata label noobs; /\*print the results\*/

run;

ods exclude enginehost;

proc contents data=&fuseddata;

run;

ods select all;

/\*The results show a total observations of 354 and 9 variables. This means 9 variables have been

removed from the data which is exactly what was done and confirms the exact number of obs

deleted. Moreover, there are 9 variables which include the new total column created. Hence,

validates the data import\*/

/\*Check the total number of missing observations\*/

proc means data=&fuseddata n nmiss min q1 median q3 max std maxdec=0;

run;

/\*The results reveal that the data contains some missing observations.

There are 65 missing observations, including class variables, in the above results.

This might not have an impact on the analysis\*/

/\*frequency table to check the distribution of data\*/

proc freq data=&fuseddata noprint;

tables Region / out= MERGED;

run;

/\*Categorise insititution based on GLH\*/

/\*Create GLH per learner for each year\*\*\*

Since the purpose is to determine the impact of some factors on GLH per learner, a new variable

'GLH per learner' has to be created\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Create New dataset \*/

/\* Classify institutions \*/

/\* Create new GLH/learner columns \*/

/\* Drop observations \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data &Newlyfused;

format SIZE $12.;

set &fuseddata;

if TOTAL\_GLH<500000 then SIZE='Small';

else if TOTAL\_GLH>=500000 and TOTAL\_GLH<1000000 then SIZE= 'Small-medium';

else if TOTAL\_GLH>=1000000 and TOTAL\_GLH<2000000 then SIZE= 'Medium';

else if TOTAL\_GLH>=2000000 and TOTAL\_GLH<3000000 then SIZE= 'Large-medium';

else if TOTAL\_GLH>3000000 then SIZE= 'Large'

;

label

GLH\_L1 = 'GLH per Learner for year 1'

GLH\_L2 = 'GLH per Learner for year 2'

GLH\_L3 = 'GLH per Learner for year 3'

SIZE = 'Size of Institution'

;

GLH\_L1=GLH1/Learners2; /\*Create new GLH per learner for year 1\*/

GLH\_L2=GLH2/Learners2; /\*Create new GLH per learner for year 2\*/

GLH\_L3=GLH1/Learners3 /\*Create new GLH per learner for year 3\*/

;

drop GLH1 /\*These variables are no longer needed\*/

GLH2

GLH3

Learners1

Learners2

Learners3

;

run;

proc print data=&Newlyfused label noobs;

run;

/\*frequency table\*/

proc freq data = &Newlyfused;

tables SIZE/out=NEW;

run;

/\*Check contents\*/

ods exclude enginehost;

proc contents data=&Newlyfused;

run;

ods select all;

/\* The data is in the correct format.There are 354 observations with 7 variables after

deleting some variables such as GLH\*/

/\*Check duplicate counts\*/

title 'duplicate primary key in';

proc SQL;

create table IMAT as

select SIZE

from NEW

where count <> 1;

select count(\*) as Count\_of\_IMAT\_errors from IMAT;

select \* from IMAT;

quit ;

title;

/\*The results revealed that the counts are not unique since 'SIZE' has many observations per

category. Same applies to Region\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* DATA EXPLORATION \*/

/\* Generate plots \*/

/\* Check for Normality \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

title 'Distribution of GLH per leaner';

proc univariate data= &Newlyfused normal ;

var

GLH\_L1

GLH\_L2

GLH\_L3

;

histogram / normal(mu= est sigma= est) ;

qqplot / normal(mu= est sigma= est) ;

inset n nmiss min q1 median q3 max skew kurt / position= NE ;

run;

ods select all ;

ods html;

/\*The goodness of fit test p values revealed that the distributions are not normal.

GLH L1 has a skewness of 1.25, GLH L2 has a skewness of 1.31, and GLH L3 has a skewness of 1.10.

The skewness of a distribution should be as close to zero as possible for it to be considered normal.

The p values are less than 0.05, indicating that there is enough evidenceÂ to support a lack of fit.

The data is not normallyÂ distributed, as evidenced by the histogram distributions and qq plots.

Finally, the Shapiro-Wilk test for each GLH per learner is 0.0001, which is significantly less than 0.05,

indicating that the data is not distributed normally.Â

The distributions mayÂ need to be transformed.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Generate Scatter plots \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro Scatter\_plots(

data= /\* data set \*/

, Classify= /\*group by \*/

);

proc sgscatter data=&Newlyfused;

matrix GLH\_L1 GLH\_L2 GLH\_L3 /group=&Classify diagonal=(histogram kernel);

run;

%mend scatter\_plots;

%scatter\_plots(data= &Newlyfused, Classify= SIZE); /\*plot by size\*/

%scatter\_plots(data= &Newlyfused, Classify= College); /\*plot by type of institution\*/

%scatter\_plots(data= &Newlyfused, Classify= Region); /\*plot by Region\*/

options nosymbolgen nomprint nomlogic;

/\*The data is significantly concentrated at the lower end of the best fit line,

indicating that it is skewed and not normally distributed and that it needs to be transformed.\*/

proc sgplot

data = &Newlyfused;

scatter x=SIZE y=GLH\_L1;

yaxis min=0;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Generate Box plots \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro box\_plots(

data= /\* data set \*/

, var= /\* name of variable modelled\*/

, Classify= /\* group by \*/

);

proc sgplot

data = &Newlyfused;

vbox &var / group=&Classify;

yaxis min=0;

run;

%mend box\_plots;

%box\_plots(data= &Newlyfused, var=GLH\_L1, Classify= SIZE); /\*Box plot of SIZE, GLH year1\*/

%box\_plots(data= &Newlyfused, var=GLH\_L2, Classify= SIZE); /\*Box plot of SIZE, GLH year2\*/

%box\_plots(data= &Newlyfused, var=GLH\_L3, Classify= Region); /\*Box plot of Region, GLH year3\*/

%box\_plots(data= &Newlyfused, var=GLH\_L3, Classify= College);/\*Box plot of College, GLH year3\*/

options nosymbolgen nomprint nomlogic;

/\*All plots indicates outliers\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Check extreme values \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods select none;

ods output ExtremeObs= EXTREMES;

proc univariate data= &Newlyfused;

var GLH\_L1 GLH\_L2 GLH\_L3 ;

id SIZE;

run;

ods select all;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* \*/

/\* Box transformations \*/

/\* Writing Functions \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options cmplib=(Course);

proc fcmp outlib=Work.Course.lib;

function p3(x); /\*cube \*/

return (x\*x\*x);

endsub;

function p2(x); /\*square\*/

return (x\*x);

endsub;

function sqrttran(x); /\*sqroot\*/

return ((x)\*\* 0.5);

endsub;

function logtran(x); /\*natural log \*/

return (log(1+x));

endsub;

function rcpsqrtran(x); /\*recipical square root \*/

return ((1+x)\*\*-0.5);

endsub;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* TRANSFORM THE VARIABLES \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options cmplib=(Course);

data &Newlyfused2;

set &Newlyfused;

trnGLH\_L1=logtran(GLH\_L1); /\*Transform GLH per learner for year 1\*/

trnGLH\_L2=rcpsqrtran(GLH\_L2); /\*Transform GLH per learner for year 2\*/

trnGLH\_L3=rcpsqrtran(GLH\_L3); /\*Transform GLH per learner for year 3\*/

label

trnGLH\_L1='Guided learning Hours per learnerYear1'

trnGLH\_L2='Guided learning Hours per learnerYear2'

trnGLH\_L3='Guided learning Hours per learnerYear3'

;

run;

/\*Check the trans plots and compare with the original\*/

proc univariate data=&Newlyfused2 normal;

var

GLH\_L1

trnGLH\_L1

GLH\_L2

trnGLH\_L2

GLH\_L3

trnGLH\_L3

;

histogram / KERNEL normal(mu= est sigma= est);

qqplot / normal(mu= est sigma= est);

inset n nmiss min q1 median q3 max skew kurt / position= SE;

run;

/\*\*\*\*The log transformation produced a more normal curve for GLH L1 than the other transformations.

Even though the p-value is less than 0.05, the skewness is 0.478 and the Shapiro-Wilk test score is 0.948. \*\*/

/\*Because GLH L2 has a skewness of -0.083 and a Shapiro test of 0.968, rcpsqrt produces a more normal curve.

Normality is indicated by a value near to 1.\*/

/\*rcpsqrt yields a more normal curve for GLH\_L3 than other transformationssince it has a

skewness of 0.271 and a Shapiro test of 0.973. A figure close to 1 indicates normality\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Drop Variables \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data &Newlyfused3;

set &Newlyfused2;

drop /\*No longer needed\*/

TOTAL\_GLH

GLH\_L1

GLH\_L2

GLH\_L3

;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* MODELLING \*/

/\* Examine effect of Size on GLH/L \*/

/\* Examine effect of Region on GLH/L \*/

/\* Examine effect of type of College on GLH/L \*/

/\* Examine effect of Year on GLH/L \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Effect of Region, College and Size on GLH/L Year1 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro GLM\_Model(

data= /\* data set \*/

, var= /\* name of variable modelled\*/

, year= /\* output to \*/

);

ods graphics on;

proc GLM data= &Newlyfused3;

class SIZE College Region;

model &var = SIZE College Region;

means SIZE College Region / lsd hovtest=levene ;

output out= &year p= predicted r= residual;

quit;

ods graphics off;

%mend GLM\_Model;

%GLM\_Model (data=&Newlyfused3, var=trnGLH\_L1, year=Year1);

options nosymbolgen nomprint nomlogic;

/\*Because the p value is less than 0.005, the model is significant

/\*Size has a considerable effect on GLH for year 1 because its p value is 0.0001.

/\*College has a considerable effect on GLH for year 1 because its p value is 0.0001.

/\*Region has a considerable effect on GLH for year 1 because its p value is 0.0027.

/\*This model accounts for 67.9% of the variation. This suggests that Size, College, and Region account for almost 70% of the model's predictions,

with the remaining 30% attributed to other factors.

The box plots show that there are significantÂ variances between the various sizes of institution because their means are far off.

Apart from large, medium, and small medium, where there appear to be no significant differences,

the pairwise distribution reveals massive disparities in all the possible combinations.

As a result, the size of the institution has a significant impact on the Guided Learning hour per learner in year 1.

As illustrated in the diagram below, sixform college and further college belong to separate categories.

The box plot also indicates that their means differ significantly. As a result, the type of institution has a significant impact

on the GLH per learner. A similar occurrenceÂ exists forÂ regions as show in the result\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Test for normality of residuals Year1 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro Residuals\_Test(

nameofdata= /\* name of data set \*/

,name= /\*title of distribution\*/

);

ods exclude Quantiles FitQuantiles TestsforLocation;

Ods trace on ;

title &name;

proc univariate data= &nameofdata normal ;

var residual ;

histogram / normal(mu= est sigma= est) kernel(color=red) ;

qqplot / normal(mu= est sigma= est) ;

inset n nmiss min q1 median q3 max skew kurt / position= SE;

run;

Ods select all;

%mend Residuals\_Test;

%Residuals\_Test (nameofdata=Year1, name='Residuals for Year1');

options nosymbolgen nomprint nomlogic;

/\*Using Shapiro-Wilk as a test for normality, its p-value shows a score of 0.0047.

This shows the errors are normally distributed. Moreover the mean(0.0000) and median (0.010455)

are closely related and almost the same. Skewness is also -0.444 which is close to zero. The

QQ plot and histogram distribution shows a normal distribution.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Effect of Region, College and Size on GLH/L Year2 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%GLM\_Model (data=&Newlyfused3, var=trnGLH\_L2, year=Year2);

options nosymbolgen nomprint nomlogic;

/\*Size has a substantial effect on GLH for year 2 because its p value is 0.0001.

/\*College has a significant effect on GLH for year 2 because its p value is 0.0001.

/\*Region has a significant effect on GLH for year 2 because its p value is 0.0071.

64.04 %Â of the variation is explained by this model. This means Size, College and Region predicts

64% of this model with the remaining 26% unaccounted for.

The box plots show that there are significantÂ variances between the various sizes of institution because their means areÂ far off.

between small and medium and large-medium and small, where there are no significant differences.

The pairwise distribution indicates all the possible combinations exhibit vast differences.Â

As a result, the size of the institution has a significant impact on the Guided Learning hour per learner in year 2.

Again, Sixform college and further college belong to different groups as shown in the diagram

below. The box plot also shows significant difference in their means. Hence, type of institution

have a great impact on GLH per learner. Similar issue is true for regions as shown below;\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Test for normality of residuals Year2 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%Residuals\_Test (nameofdata=Year2, name='Residuals for Year2');

options nosymbolgen nomprint nomlogic;

/\*The results show the mean and the median are almost the same at zero. Again, Shapiro-Wilk

test for normality gives a result of 0.0001 with almost zero variance in the data (0.000097).

The QQ plot and histogram chart shows a normally distributed data\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Effect of Region, College and Size on GLH/L Year3 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%GLM\_Model (data=&Newlyfused3, var=trnGLH\_L3, year=Year3);

options nosymbolgen nomprint nomlogic;

/\*The model is significant since the p value is less than 0.005 at 0.01%.

/\*Size has a significant effect on GLH for year 3 since its p value is 0.007.

/\*College has a significant effect on GLH for year 3 since its p value is 0.0001.

/\*Region has a significant effect on GLH for year 3 since its p value is 0.0019.

/\*Only 57.73%Â of the variation is explained by this model.

This suggests that Size, College, and Region account for almost 58 %Â of the model, leaving 42% unaccounted for.

Observing from the box plots, there are significant differences among the various sizes

of institution since their means sharply depart from each other. Hence, size of institution

plays a key role in effecting Guided Learning hour per learner for year 3.

However, using the pairwise combinations, the significance of the SIZE has reduced as there

are not much differences among the various combinations compared to years 1 and 2 as shown in

the diagram below;

Again, Sixform college and further college belong to different groups as shown in the diagram

below. The box plot also shows significant difference in their means. Hence, type of institution

have a great impact on GLH per learner. Similar issue is true for regions as shown below;\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Test for normality of residuals Year3 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%Residuals\_Test (nameofdata=Year3, name='Residuals for Year3');

options nosymbolgen nomprint nomlogic;

/\*The results shows the mean and the median are almost the same at zero. Again, Shapiro-Wilk

test for normality gives a result of 0.0001 with almost zero variance in the data (0.00011).

Moreover, all the other goodness-of-fit tests give values below 0.05. Its skewness is closer

to zero (0.796).

The QQ plot and histogram chart shows a normally distributed data\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* ttest for type of institution \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc ttest data= &Newlyfused3;

class College;

var trnGLH\_L1 trnGLH\_L2 trnGLH\_L3;

run;

/\*After three years of observing the two colleges' means, we can see that there is a big disparity between them.

Sixform college appears to do better than FE College in the first year, with a mean GLH per learner of 6.24 compared to 5.26 for FE College.

However, FE College tends to perform better than Six Form in the second year, with a mean of 0.072 compared to 0.044 for Sixform.

However, there was no change in performance in the third year, since the results remained the same as in year 2.

In year 1, the model assumtion of equality of variance is not met since the Folded F p value

is 0.0431. Hence, the Satterthwaite score was used to judge the significance of the difference

which is 0.0001. In year 2, the equality of variance assumption was met since the Folded F p

was 0.3126. Hence, the pooled variance P value of 0.0001 was used which shows a significant level.

Again, for year 3 like in year 1 the assumption was not met indicating that the two colleges data had unequal

variances. However, the Satterthwaite unequal variance score of 0.0001 was used which indicated

that there is a significant difference between the two groups.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* EFFECTS OF YEARS ON MODEL \*/

/\* By size of institution \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro proc\_means\_effects(

dataset= /\* name of data set \*/

,size= /\*title of distribution\*/

, region=

,college=

);

proc means

data = &dataset n nmiss mean std fw=5 maxdec=2 nonobs;

class &size;

var trnGLH\_L1

trnGLH\_L2

trnGLH\_L3;

run;

%mend proc\_means\_effects;

%proc\_means\_effects (dataset=&Newlyfused3, size=SIZE);

options nosymbolgen nomprint nomlogic;

/\*There is a significant decrease of GLH/Learner from year 1 to 2, but no change in year 3 for

all the SIZES of institutions.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* EFFECTS OF YEARS ON MODEL \*/

/\* By region of institution \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%proc\_means\_effects (dataset=&Newlyfused3, size=Region);

options nosymbolgen nomprint nomlogic;

/\*In respect of GLH/Learner for each region, there is a significant decrease of GLH/Learner

from year 1 to 2, but no change in year 3 just as SIZE.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* CHECK EFFECTS OF YEARS ON MODEL \*/

/\* By type of institution \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%proc\_means\_effects (dataset=&Newlyfused3, size=College);

options nosymbolgen nomprint nomlogic;

/\*There is a significant decrease of GLH/Learner from year 1 to 2, but no change in year 3 for

for both types of institutions.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* FURTHER ANALYSIS \*/

/\* Effects of Regions and Years on Model \*/

/\* By size of institution and region \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options symbolgen mprint mlogic;

%macro proc\_tabulate\_effects(

dataset= /\* name of data set \*/

,type= /\*generated by\*/

, region= /\*region of institution\*/

,dataout=

);

proc tabulate data=&dataset out= &dataout missing ;

class &type &region;

Var trnGLH\_L1

trnGLH\_L2

trnGLH\_L3

;

tables

(trnGLH\_L1

trnGLH\_L2

trnGLH\_L3) \*&type\*&region

, ( n nmiss

min \*(f= 8.2) q1 \*(f= 8.2) median \*(f= 8.2) q3 \*(f= 8.2) max \*(f= 8.2)

mean \*(f= 8.2) std \*(f= 8.2)

)

;

run;

%mend proc\_tabulate\_effects;

%proc\_tabulate\_effects (dataset=&Newlyfused3, type=SIZE, region=Region, dataout=Trans4);

options nosymbolgen nomprint nomlogic;

/\*\*\*\*For year1, large institutions in East of England did better than the rest of the regions

with an average score of 5.65. With large medium, Greater London performs better with a mean

of 6.07. With medium colleges, North East colleges perform better with an average of 6.13.

West Midlands colleges did better than other regions in the small category while North East is

the only region which has a small-medium college with an average of 6.56.

Hence, small-medium colleges in North East seems to record the highest effect on

GLH/learner with a mean score of 6.56. This is followed by medium sized institutions.This could be

due to the fact that there are just a few institutiins in this category. For instance, under

small-medium, only one institution exists while there are only three under medium.

For Year 2, large institutions in the various regions achieved similar scores of about 0.07

and 0.06. While East Midlands, South West, North East and West Midlands had 0.07, the rest had

0.06. This is a drastical reduction from year 1 for all large institutions.

Large-medium sized institutions in South West have the greatest effect on GLH/learner with

an average of 0.09 with the remaining regions having almost identical scores.

With small colleges, East of England and East Midlands have the greatest effects as they have

0.08 as means. Lastly, North East is the only region with a small-medium colleges with a an

average score of 0.04. Again, there seems to be a significant reduction in GLH/learner for

all sized institutions with respect to year 1.

For Year 3, there seems to be no change in scores for large colleges from year 2. Even though

colleges in South West seems to achieve the highest score under large-medium category, its

significacne decreased from 0.09 from year two to 0.08 in Year 3. Greater london also scores

highly of 0.10 under medium sized institutions. However, its significance has reduced from

0.12 in year 2. All others either stay same or reduce slighly. Interestingly, apart from

West Midlands there are no records of small sized colleges for all the other regions. This

calls for investigations.

The analysis shows there is a gradual improvement of the medium and large medium colleges

relative to the other category of colleges. It could therefore mean that the government had

implemented policies deliberately to support the medium and large medium sized institutions

with Greater London being a focal point. This is because Greater London achieved the highest

score in Year 2 and again in Year 3, all under medium sized isntitutions. Hence, for lack

of support the small companies decided to fold up in Year 3. Perhaps, due to decline in

learning hours with a slight improvement and govenrment support, parents decide to withdraw

their wards from the small colleges \*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* FURTHER ANALYSIS \*/

/\* Effects of Regions and Years on Model \*/

/\* By type of institution and region \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%proc\_tabulate\_effects (dataset=&Newlyfused3, type=College, region=Region, dataout=Trans5);

options nosymbolgen nomprint nomlogic;

/\*For Year 1, further colleges in East of England achieves a better sscore than the

other regions with an average score of 5.42 followed by Greater London. On the other hand,

sixform colleges in East Midland do better than the other regions with a mean score of 6.48

followed by North East and South West with 6.46 each. Overall, Sixform college speciifcally in

East of England does better than further colleges in year 1.

For Year 2, six form colleges in West Midland achieve the highest score of 0.05 with the rest

of the regions achieving the same scores of 0.04. Further colleges in East Midlands, West

Midlands and North East achieved the same scores of 0.08 while the rest achieved same scores

of 0.07.

In Year 3, there seems to be no change in performance as the figures are almost the same as

the scores achieved in Year 2.

Observation:

There is a significant change or effect of type and size of institution by region on GLH/learner

or difference between year 1 and two. However, there is no change in year 3. Year 1 has the

greatest effect. However, this declines drastically in year 2 and stays the same in year 3.

Hence, significance declines by year. Therefore, it is reasonable to conclude that year has no

significant effect on GLH per learner\*/