IMAT5168 ANALYTICAL PROGRAMMING

ASSESSMENT 2

ANALYSING IMAT5168-6FORM.csv AND IMAT5168-FE.csv

Done By

Kaushik Chandrasekhar

P2656623

**Contents**

1.SUMMARY…………………………………………………………………………3

2.METHOD……………………………………………………………………………3

2.1 IMPORTING THE 2 DATASETS ……………………………………….3

2.2 VALIDATING THE DATASETS………………………………………….4

2.3 USING PROC GLM TO ANALYSE THE DATASETS…………….5

2.3.1 YEAR 1……………………………………………………………………..5

2.3.2 YEAR 2……………………………………………………………………..6

2.3.3 YEAR 3……………………………………………………………………..7

3. CONCLUSION……………………………………………………………………8

4. REFERENCE……………………………………………………………………….8

**1. SUMMARY**

Further education college are the colleges where skills are acquired outside academic college. While sixth form college has 2 years’ post-secondary school that equip students with knowledge for further education.

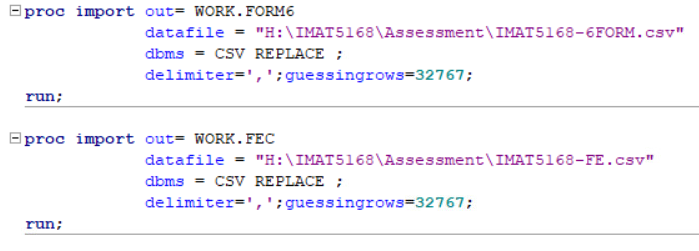
In this report I am examining the relationship between sixth form college and further education college across different regions, institution size and institution type. I am going to focus on how these three categories influences average GLH per learner.

I am considering the NULL hypothesis. This means there is no effect of Region, College Type, Size and Year 1, 2 and 3 on Average GLH per learner.

**2. METHOD**

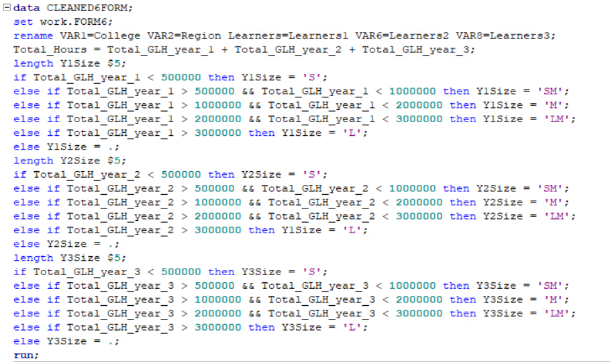
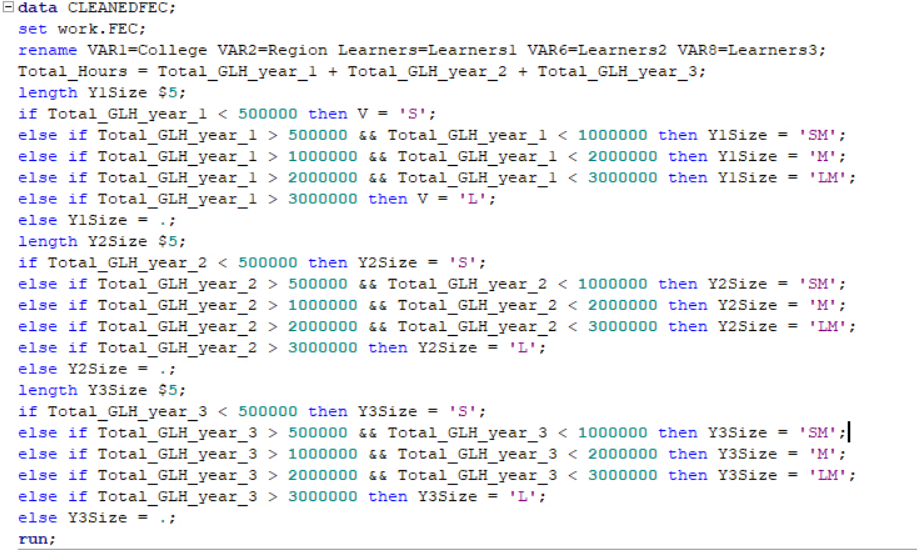
**2.1 IMPORTING THE 2 DATASETS.**

I am importing 2 given data which has the information on “Sixth Form College” and “FE College” in IMAT5168-6FORM.csv and IMAT5168-FE.csv respectively.



I have used PROC IMPORT to import and have saved them as FORM6 and FEC respectively. The “guessingrows=32767’ is given to import the whole string values in the region column.

**Cleaning Sixth form College and FE College Data**

I am creating 2 new datasets CLEANED6FORM and CLEANEDFEC from FORM6 and FEC respectively.

The following changes are made to the variables of the datasets:

1)VAR1 is changed to College

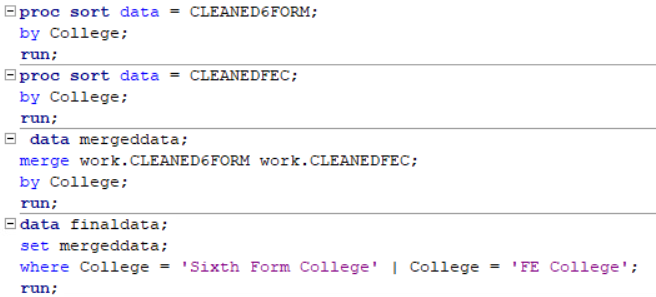
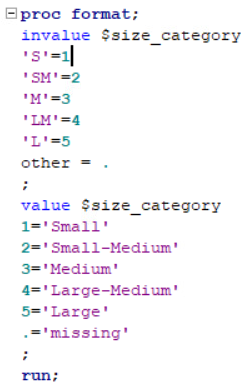
2)VAR2 is changed to Region

3)Learners is changed to Learners1

4)VAR6 is changed to Learners2

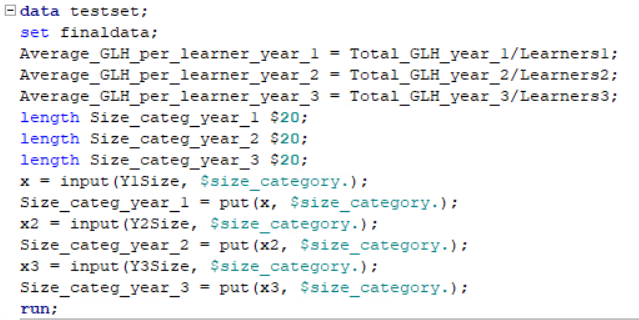
5)VAR8 is changed to Learners3

After renaming, 3 variables Y1Size, Y2Size & Y3Size are created to assign the size category code of year 1, 2 and 3 respectively (S – small, SM – small-medium, M – medium, LM – Large-medium, L – Large). The given conditions are met with the help of an if-else ladder.

The above code is used to merge the 2 cleaned datasets. In the “finaldata” the data is further cleansed as the observations of Total values of the Regions have been removed using the WHERE command.

The above INFORMAT and FORMAT “size\_category” have been created to assign a new variable with the size category with respect to the variable “Y1Size”, “Y2Size”, “Y3Size” which were created above.

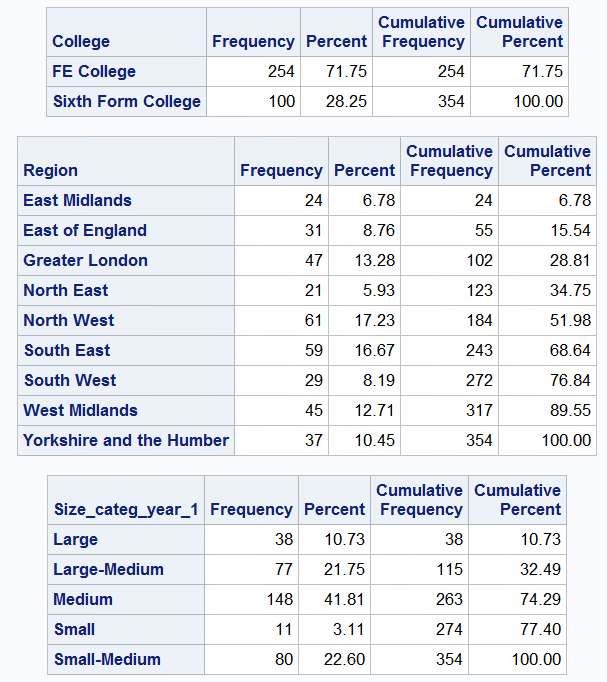
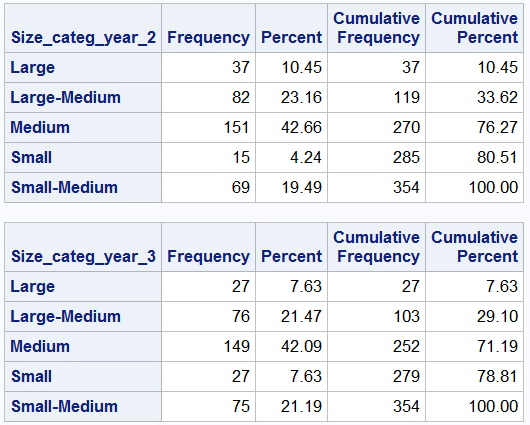
 A dataset “testset” is created from the finaldata. 2 New variables are added to the dataset. They are:

1. Average\_GLH\_per\_learner\_year\_1: Total GLH of year 1 divided by Total Learners of year 1.
2. Size\_categ\_year\_1: this variable represents the Size of each College with respect to its Total\_GLH\_year\_1
3. Average\_GLH\_per\_learner\_year\_2: Total GLH of year 2 divided by Total Learners of year 2.
4. Size\_categ\_year\_2: this variable represents the Size of each College with respect to its Total\_GLH\_year\_2
5. Average\_GLH\_per\_learner\_year\_3: Total GLH of year 3 divided by Total Learners of year 3.
6. Size\_categ\_year\_3: this variable represents the Size of each College with respect to its Total\_GLH\_year\_3

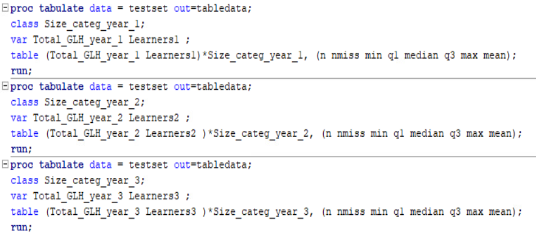
The 3 averages are taken separately to analyse the data with respect to Year 1, Year 2 & Year 3 respectively.

**2.2 VALIDATING THE DATASETS**

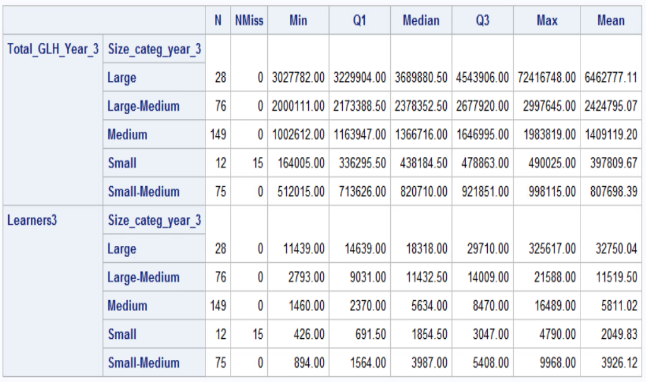
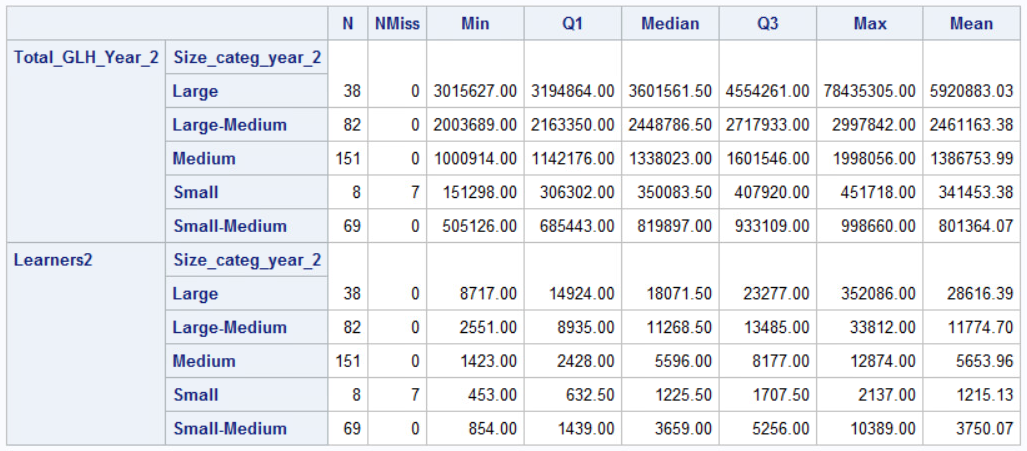
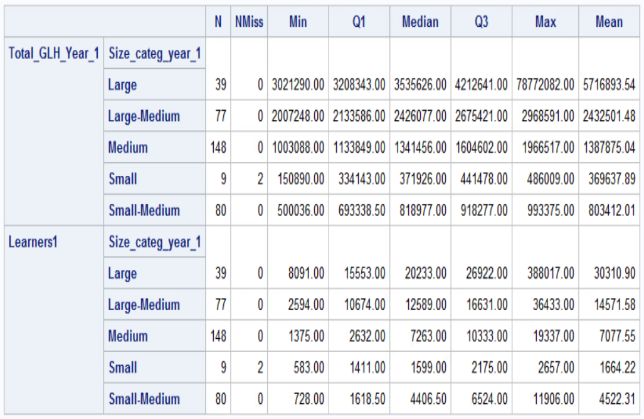
Code:

Output:

The above output displays the frequency of all the categorical variables I am using for my analysis.

I am using proc tabulate to validate the data in in the 3 with respect to the Size. The below code is used for the validation of the data.

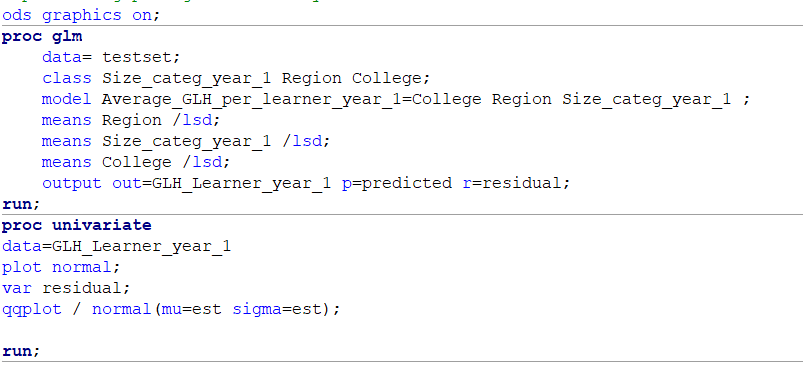
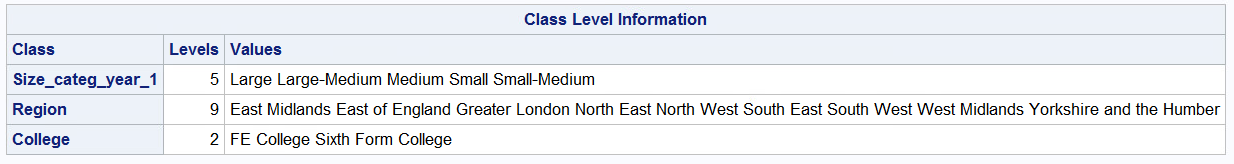
OUTPUT:

The tables shows the N, (number of records) Nmiss (missing records), Min, Q1, Median, Q3, Max and mean value of the observations in Total\_GLH\_Year\_1:Learners1, Total\_GLH\_Year\_2:Learners2, Total\_GLH\_Year\_3:Learners3 with respect to the Size of each year. From the 3 tables we can see that there are 2 missing values in Small Size college of year 1, 7 missing values in Small Size college of year 2 and 15 missing values in Small Size college of year 3

**2.3 USING PROC GLM TO ANALYSE THE DATASETS**

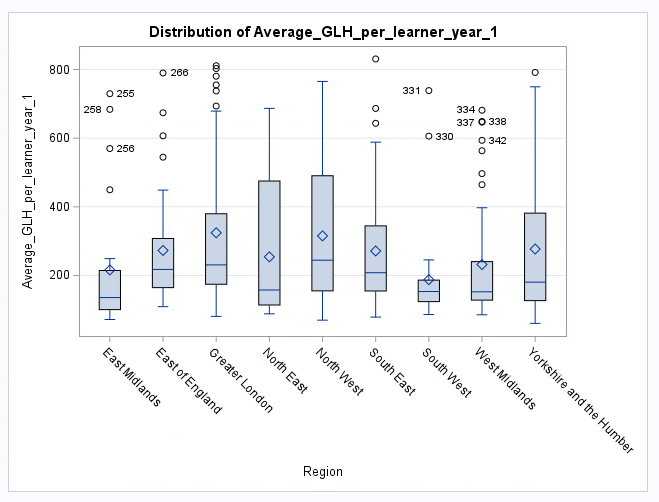
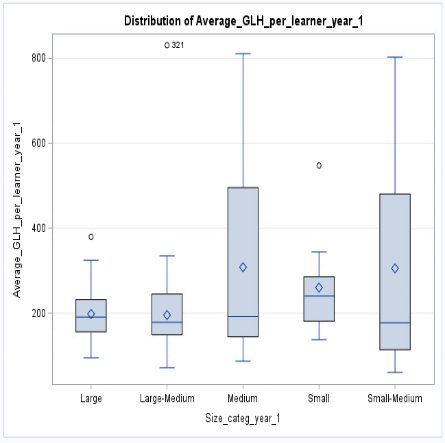
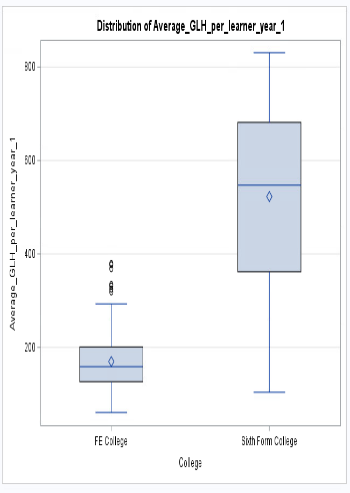
**2.3.1 YEAR 1**

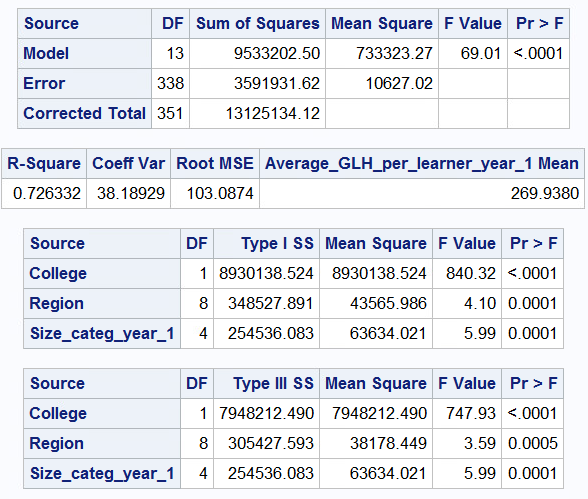
The below code is used to analyse the Average GLH per learner in year 1.

Code: Output:



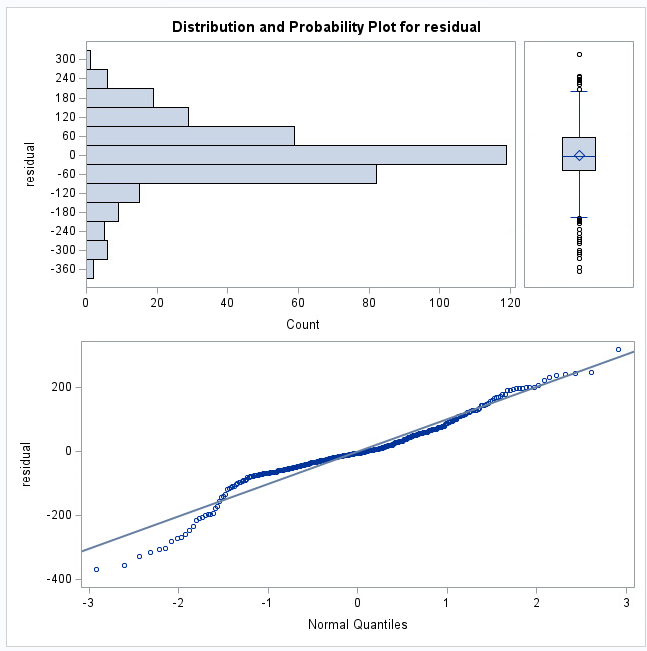
Only 352 observations are used as 2 of the values were missing.

These are the classes which are going to that are going to be considered not affecting the Average\_GLH\_per\_Learner\_year\_1.

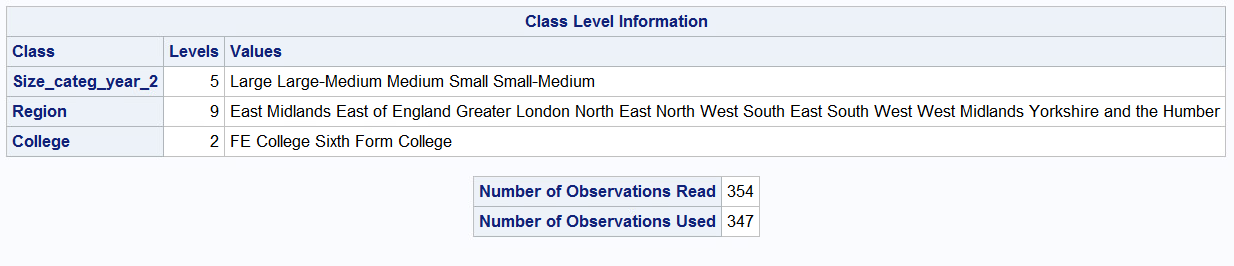
From the above graphs we can make out that in Region “North East” and “North West” has the highest meanwhile “South West” in Size\_categ\_year\_1 “Medium” and “Small-Medium” has the highest mean and Size “Large” has the Lowest. In College “Sixth Form College” has the highest mean.

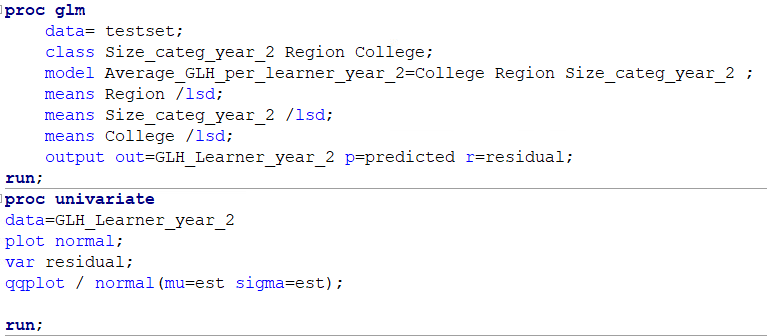
From the table we can see that the p values is <.0001 which is less than 0.05. Hence there is a significant difference between the Average\_GLH\_per\_Learner\_year\_1 and Region, College & Size\_categ\_year\_1.

The R-Square of 0.726 signifies a 72.6% relationship between Average\_GLH\_per\_Learner\_year\_1 and Region, College & Size\_categ\_year\_1.

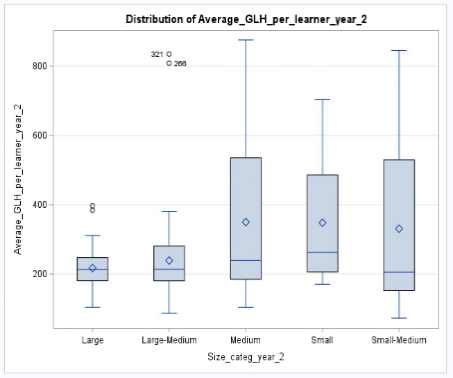
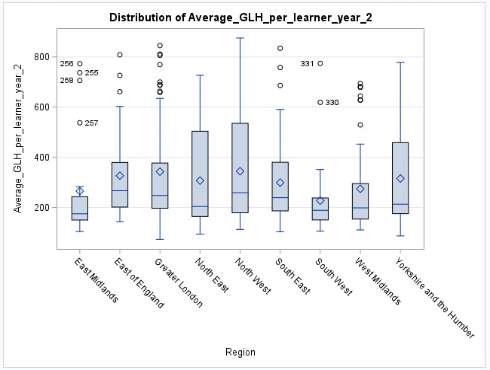
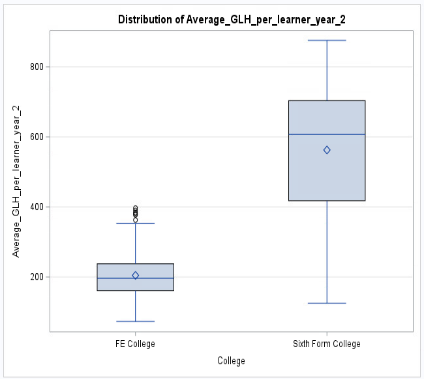
From the above graph of analysing the residual data from proc glm for normality, we can see that the variables are normally distributed with few outliers.

**2.3.2 YEAR 2**

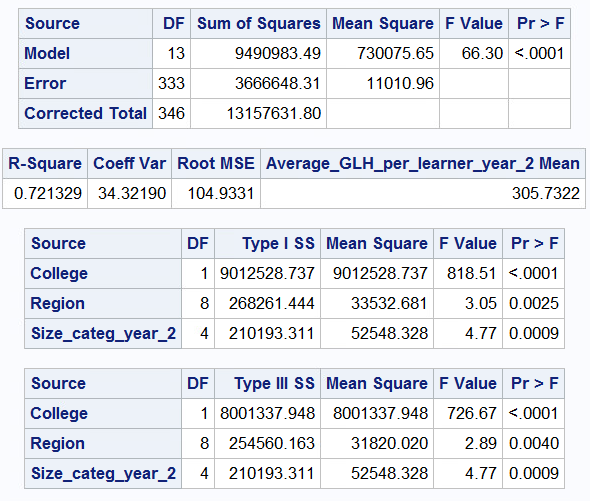
Code: Output:

****

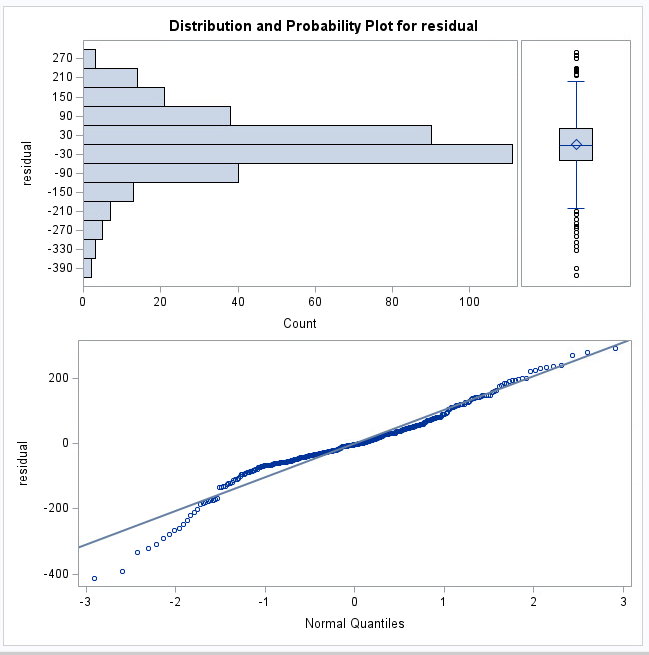
The above code us used to analyse the Average\_GLH\_per\_learner\_year\_2. Only 347 observations have been used for the analysis as 7 values were missing. These are the classes which are going to that are going to be considered not affecting the Average\_GLH\_per\_Learner\_year\_2.

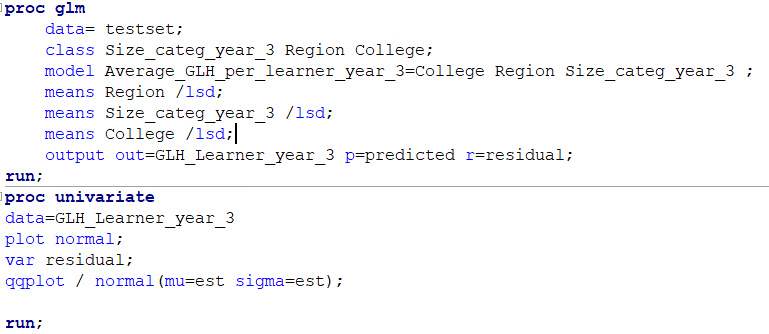
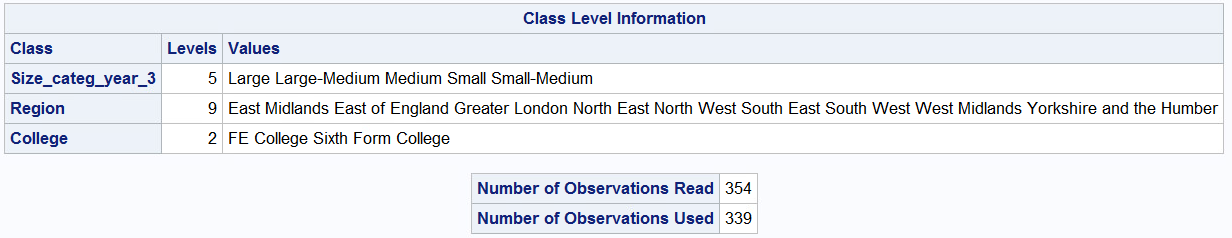
 

The box plot represents how Average\_GLH\_per\_learner\_year\_2 is distributed across the different Region, Size\_categ\_year\_2 and institution types. From the above graphs we can make out that in Region “North East” and “North West” has the highest meanwhile “South West” in Size\_categ\_year\_2 “Medium” and “Small-Medium” has the highest mean and Size “Large” has the Lowest. In College “Sixth Form College” has the highest mean.

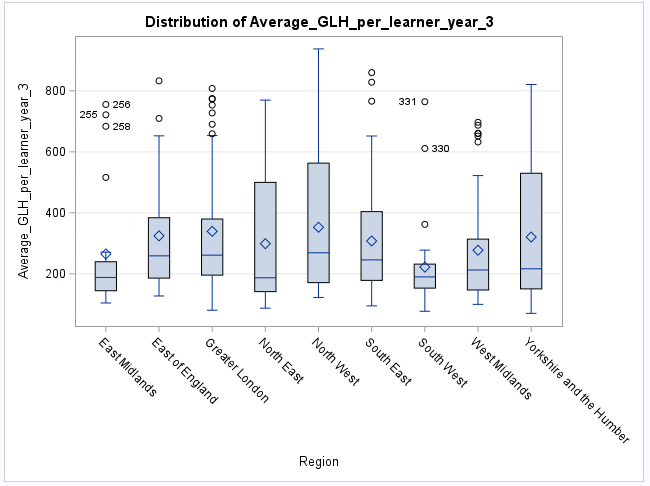
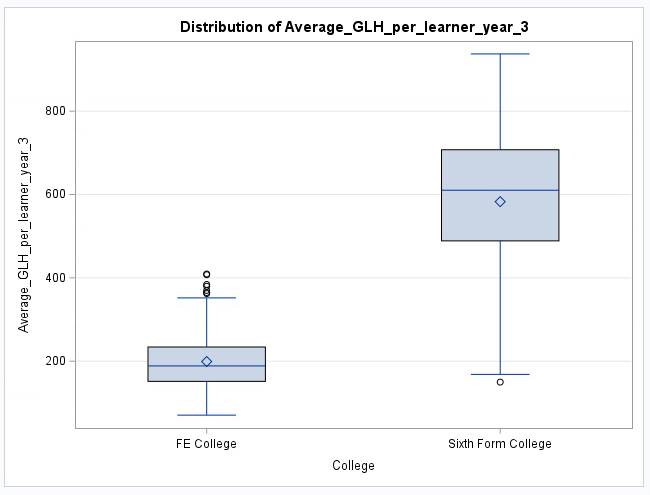
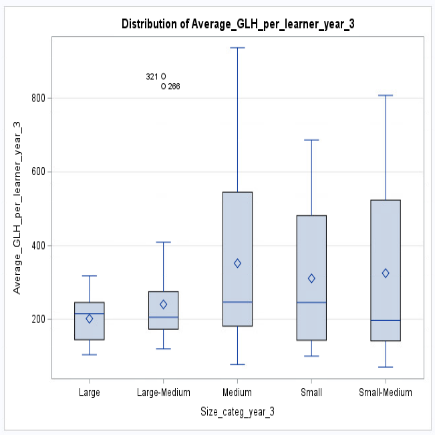
From the above table we can see that the p values is<.0001 which is less than 0.05. Hence there is a significant difference between the Average\_GLH\_per\_Learner\_year\_2 and Region, College & Size\_categ\_year\_2.

The R-Square of 0.721 signifies a 72.1% relationship between Average\_GLH\_per\_Learner\_year\_2 and Region, College & Size\_categ\_year\_2.

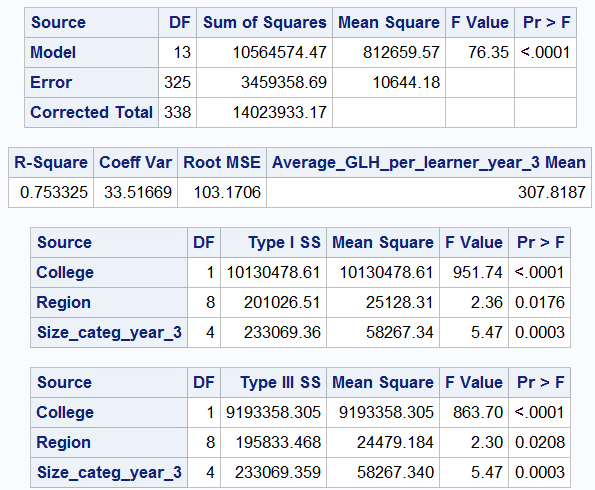
From the above graph of the residual variable of year 2 we can see that most of the values are normally distrbuted with few outliers.

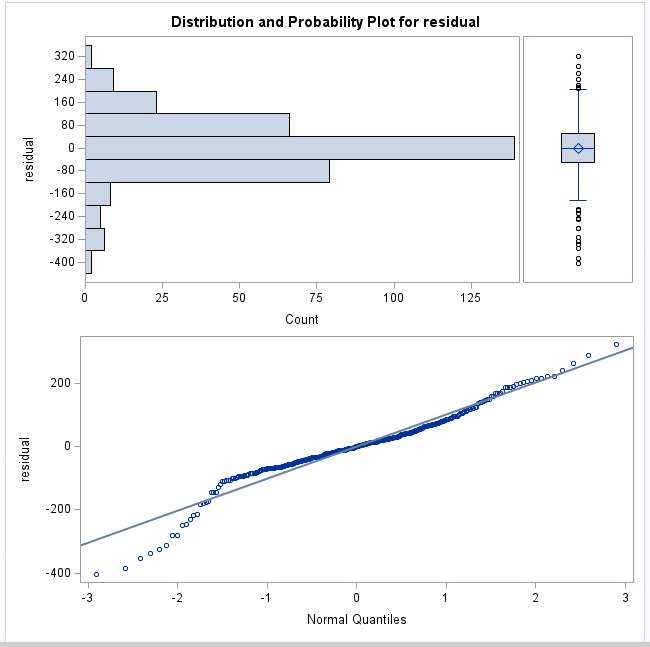
**2.3.3 YEAR 3**

The above code us used to analyse the Average\_GLH\_per\_learner\_year\_3. 15 observations are missing from Size\_categ\_year\_3 hence only 339 observations are used.

The box plot represents how Average\_GLH\_per\_learner\_year\_3 is distributed across the different Region, Size\_categ\_year\_3 and institution types. From the above graphs we can make out that in Region “North East” and “North West” has the highest meanwhile “South West” in Size\_categ\_year\_3 “Medium” and “Small-Medium” has the highest mean and Size “Large” and “Large-Medium” has the Lowest. In College “Sixth Form College” has the highest mean.

From the above table we can see that the p values is <.0001 which is less than 0.05. Hence there is a significant difference between the Average\_GLH\_per\_Learner\_year\_3 and Region, College & Size\_categ\_year\_3.

The R-Square of 0.753 signifies a 75.3% relationship between Average\_GLH\_per\_Learner\_year\_3 and Region, College & Size\_categ\_year\_3.

From the graph of the residual variable of year 3 we can see that most of the values are normally distrbuted with few outliers.

**3. CONCLUSION**

From the above general linear model analysis, we can conclude that the NULL hypothesis has been proved wrong and there is effect of Region, Institution type, Size and Year on the Average GLH per learner.

**4. REFERENCE**

IMAT5168 2021-week 9 Lecture slides and video by Dr Anthony Williams. De Montfort University, Leicester United Kingdom.