**REPORT ON GUIDED LEARNING HOURS IN FURTHER EDUCATION AND SIXTH FORM COLLEGES.**

**ANNE AWELE NWAOKOLO P2743914 MSC DATA ANALYTICS**

1. **SUMMARY**

An analysis of Guided Learning Hours in Further Education and Sixth form colleges was carried out, two csv files were imported into the SAS work sheets, Anomalies were detected, the data was transformed and an exploratory analysis was done. One way ANOVA, Wilcoxon Tests and Two-way ANOVA were used to test the relationships between Guided learning hours per learner by Institution, Size, Year and Region and to also ascertain by Institution and Size the effect of Region and Year on Guided Learning Hours. 4 different tests were done, one way ANOVA, Two Way ANOVA, Chi Square test of Independence and the Wilcoxon test. The Various test results show that Region, Size and Institution have the most effect on GLH per learner, while the year is the least impactful.

1. **METHODS**

**Creating the Libraries, Importing and Cleaning the Data**

Library name and Filename were created to define a Path. Using an infile statement, the CSV files named as SIXFORM.csv and FEcollege.csv were imported to SAS with the various variables assigned and labels added. The $ sign was used to Identify the categorical variables and the length specified as 50 to cover as many characters as possible. The firstobs is specified as 2 so the dataline can start reading from the second line in the CSV file. Dsd and Dlm is used to indicate a comma as the delimiter separator. Sixth form had a total of 110 observations and FE college had a total of 263 observations containing records of guided learning hours of students in sixth form and FE colleges with 8 columns having details of number of learners, the region and Total GLH for 3 Years, Year 1, Year 2 and Year 3. The code used for this and result is displayed in Appendix A.

**Deleting the rows that have the total figures**

It was observed that each region had its total calculated as an additional row in the dataset, to avoid a distortion in the results of the analyses these rows were deleted for both datasets using an ‘If’ statement is used and the code and result is shown in Appendix B.

**Detecting and dealing with the Missing values.**

Proc means was used to identify details of missing values, a total of 12 missing values were observed in Sixth form college data and 36 in the FE college dataset. The missing data were identified in the Total GLH and Learners columns for both FE and 6th Form Dataset, the number of missing values is not too significant compared total observations. To avoid removing any data which is valuable to the analysis. I considered filling the missing values with the mean values. The proc stdize statement was used. Details of the code used and corresponding result can be found in Appendix C.

**Merging the Two datasets**

Due to the fact that the 2 datasets have the exact same number of rows and contain similar information the data were merged using the PROC SQL statement, the “union all” option was also used in the statement in order for all the data to be captured and properly aligned. Code and result details in Appendix D

**Exploratory Data Analysis**

* **Identifying frequency distributions, using tables and plots**

Using proc means and proc content statements, we view a summary of the dataset. Total number of Observations is 354. TotalGLHyr1 Median is 1451343.50, Std is 986517.16, Minimum 150890.00, Max 5877756.00, 25th percentile 993375.00, 75th Percentile 2218940.00. TotalGLHyr2- Median is 1455522.00, Std is 1041954.86, Minimum 151298.00, Max 7716710.00, 25th percentile 1025785.00, 75th Percentile 2309887.00. TotalGLHyr3 Median is 1436383.50, Std is 971772.42, Minimum 164005.00 Max 7545209.00, 25th percentile 1003741.00, 75th Percentile 2131463.00. Std dev for Learner’s year 1, 2 and 3 show the data is not normally distributed.

The table shows the percentage and frequency of the Region and Institution. There are 9 categories for Region, East of England, Greater London, Northeast, Northwest, Southeast, Southwest, West Midlands, West Midlands Yorkshire and the Humber, there are 2 variables for Institution, FE College and Sixth Form College. FE college represents 71.75% Frequency while 28.25% is for Sixth Form College (254 and 100 counts respectively).

North West and South East Regions have the Highest frequency, of 17.23% and 16.67% respectively. For the size, Category for Medium has the highest frequency of 43.97% followed by Large-Medium and Small Medium having 22.60% and 21.09%. Using Proc Sgplot statement, Scatter plots, Bar charts and Pie chart was also used to visualise the frequencies and display correlations. All codes and results can be found in Appendix E.

**Summarising the data using different methods.**

We start by creating new variables for Year, Size and GLH per learner.

a new variable was created for size by mapping the GLH by the size categories as provided in the report specification. The length for the size column was indicated as 25 to accommodate the number of characters. The new table has 8 columns and 1062 observations, the institutions are categorised as Large for GLH greater than 3 million, Large-Medium if it's between 2 million and 3 million, Medium if the GLH is between 1 million and two million Small-medium is from 500,000 to 1 million while less than 500,000 is Small.

A new variable was created for year by using ‘SET’ to transpose the value of the Total GLH and Learners for Year 1, 2 and 3 and placing them with the corresponding values for each year by the side in new columns while at the same time eliminating the columns which are no longer needed at this point using the keep option to determine which columns to remain. New columns were also created to sum up the learners and Total GLH for the 3 years although at this point it is uncertain if those 2 new columns are significant in analysis.

A variable for GLH per learner is created by dividing the GLH by the number of learners and then it is rounded up using a “format” option to get an appropriate figure without decimals since we are dealing with hours.

After creating the new variables for Year, Size GLH per learner and additional columns for Total GLH and Learners for the 3 years the new table has 1062 observations with 9 columns. The codes used in the new variable creations and their results can be all be found in Appendix F.

* **Proc Freq** is used to summarise all the class variables, the percentages and frequencies of the Region and Institution are shown in the table. There are two variables for Institution, FE College and Sixth Form College. Size has 5 categories and there are 3 years.
* **Proc means** is used to get statistical details for the continuous variable GLH, Learner, GLH per learner

Based on the summary statistics for the GLH variable, the standard deviation is 1000402.79, Minimum is 150890, and maximum is 7716710 and mean of 1723259.97, for learners, there are 8,489.53 students on average across all institutions, with a standard deviation of 6,818.71. There are 426 students minimum and 53,281 students maximum. The median student population is 7,341. GLH per learner has a mean of 293.72, Median of 212.03 and a standard deviation of 196.91. The data shows a positive skew to the right. The kurtosis value of 0.45, shows that the distribution is not normal.

* **Proc Tabulate** was used to compare the continuous variables against the categorical variable.

**Summary from Proc Tabulate**

* **East Midlands Region** – The medium and small-medium sized Institutions under the Sixth form performed best with Median Values of 721.7 and 655.3 GLH Per learner respectively
* **East of England-** Overall in East of England for all 3 years the Sixth form colleges recorded highest number of GLH per learner.
* **Greater London-** the same pattern was observed for the 3 years, the highest GLH was achieved by the medium and the small-medium sized sixth form colleges as compared to the FE colleges.
* **North East Region**- The highest mean observed in GLH per learner is 719 from the Small-Medium Sixth Form college in year 3 The Lowest performance is from Medium FE college with a mean GLH per learner of 112.
* **North West Region** - The highest mean observed in GLH per learner is 631 from the Medium Sixth Form college the second year, The Lowest performance is from Small-medium of FE college with a mean GLH per learner of 145.
* **South East Region** The highest mean observed in GLH per learner is 610 from the Large-Medium Sixth Form college. The Lowest performance is from Small-Medium FE college with a mean GLH per learner of 118.
* **South West Region** The highest mean observed in GLH per learner is 739 from the Small-Medium Sixth Form college.

The Lowest performance is from Small-Medium FE college with a mean GLH per learner of 116.

* **West Midlands** The best performance was recorded in the Medium Sixth form college with a mean GLH per learner of 578  
  The Lowest mean recorded was in the Small Medium FE college with a mean value of 128.
* **Yorkshire and the Humber** the Highest mean recorded is 654 from the Medium Sixth Form college while the lowest performance is from the  
  small medium FE college with an average GLH per learner of 122.

We also plot some box plot to view the distribution of GLH per learner against region, size, institution and year.

Details of the codes and results used for generating the summaries are all located in Appendix G.

**Statistical Modelling.**

**One-way Anova**

The means of two or more groups for one dependent variable can be compared using a one-way ANOVA. When the investigation involves more than two groups, a one-way ANOVA is necessary (Ross & Willson, 2017). Since we are comparing 3 years, a one-way ANOVA can be used. From the result, based on the P value 0.0167 and F value 4.11, we can say that the Year actually does have a significant effect on GLHperlearner. ANOVA code and result can be seen in Appendix H.

**Chi Square test**

When the variables are nominal, one of the most helpful statistics used to check relationships with testing hypotheses, is the Chi-square test of independence also known as the Pearson Chi-square test or simply the Chi-square. (McHugh, 2013) based on this we can use the chi square to check relationship between Institution and region.

for Institution and Region, based on the chi square and likelihood ratio chi square, with p values of <.0001 it shows that the region has a significant association with type of institution with northwest region having the highest number of FE colleges (15.35%) and south west region having the lowest number of sixth form colleges. Chi Square SAS code and results can be found in Appendix I.

**Testing the data for normality**- to ascertain if we can rely on parametric or non-parametric test, we test for normality of the data using the Proc univariate. The result show that the data distribution is abnormal. An attempt to transform the data was done using a log function however while it improved the distribution, it remained abnormal. Please check Appendix J and K for before and after effect. Since the data is not normally distributed, we can perform a non-parametric test which we can rely on.

**Wilcoxon test-** this is like the paired ttest but in the non-parametric form (Scheff 2016).

* **Institution**-The box plot shows a major difference between the two institutions for GLHperlearner distributions,   
  with the Sixth Form College having a higher median value. Also, The Two-Sample Test Statistic is 18.8031 with a p-value of.0001, the One-Way Analysis Chi-Square is 353.5553 with a p-value of.001. These results support the findings that the GLHperlearner between the FE College and Sixth Form College are significantly different.
* **Region**-The p-value is less than .001 showing that we have to reject the null hypothesis, and the F-value is 4.6638 demonstrates a significant relationship between region and GLH per learner
* **Year**- The Kruskal-Wallis test, which had a chi-square value of 25.74 and a p-value of less than 0.0001 reveals a significant difference between the three years.
* **Size -** The Wilcoxon scores and box plot indicate that the GLH per learners significantly impacted by institution size. All result details are in Appendix L

**Two-way Anova**

A two-way ANOVA is used to determine how the values of two categorical factors affect the mean of a quantitative variable. When you wish to determine how two independent factors interact with one another to affect a dependent variable, use a two-way ANOVA(Bevans,2022) Based on this, two-way ANOVA is used to check the effect of region and year, institution and size using the Proc GLM statement. We now check the effect of region and year by institution and size, this is achieved by first sorting the Data using a Proc Sort statement. With the Proc GLM statement, an option for “by” is indicated. Details of the code and results can be found in Appendix M.

1. **RESULTS**

**Large FE College,**

the analysis shows that there are significant differences in GLHperlearner between various regions, East Midlands, East of England, Greater London, North East, North West, South East, South West, West Midlands, and Yorkshire and the Humber. There is significant interaction between region and year. The Large FE College's mean GLHperlearner is 206.1723, with a high coefficient of variation of 23.05654 and a root mean square error of 47.53620

**Large-Medium FE college**, there are significant differences in GLHperlearner between all regions and between the various years. Region and Year interactions are not significant. When compared to the mean GLHperlearner for the large-medium FE College, the root mean square error is 61.52696

**Medium FE College,** there are significant differences in GLHperlearner between various regions North West, South West, and West Midlands and between years 1, 2, and 3. Significant relationship exists between year and region. The average GLHperlearner for the Medium FE College is 190.3359, with a coefficient of variation of 26.97898 and a root mean square error of 51.35067.

**For Small FE college** there are no significant differences in GLHperlearner between three regions North West, South West, and West Midlands or the 3 years (1, 2, and 3). The root mean square error**,** for the Small FE College's mean GLHperlearner is 98.63657, with a high coefficient of variation of 44.55770.\*/

**Small Medium FE college,** the results indicate that there is a significant effect of region and year on the GLHperlearner with P values 0.1008 and 0.0055, respectively but the interaction between region and year is not p=0. 9847.According to the interaction plot, the East of England has the highest mean 173.07, while Greater London has the lowest mean (122.83) for the least squares mean for GLHperlearner. Additionally, the plot reveals that GLHperlearner is highest in year 2 (151.20) and lowest in year 1 (121.42) Overall, the results reveal year as significant variables in GLHperlearner differences in small- to medium-sized FE colleges.

**Large-Medium Sixth form college**

The model shows no significant effect of region or year on GLHperlearner, as the F values for both factors are above 0.05.

**Medium Sixth Form College** The result indicate that region had a significant (p0.05) impact on GLHperlearner, with East Midlands having the highest LSMEAN value (695.096390) and East of England having the lowest (538.214457). According to the interaction plot There is no significant interaction with the year

**Small Sixth form college** There is significant effect of region and no significant effect of year on GLHperlearner. This means that variation in GLHperlearner is driven by differences across East of England,North East, South East, and West Midlands regions in small-sized institutions.

**Small-Medium Sixth form,**

we see that there is a substantial (p 0.05) effect of region on GLHperlearner, with the North East having the greatest LSMEAN value (674.661318) and the South East having the lowest (334.648375). The interaction plot demonstrates that the GLHperlearner scores at the various years do not differ significantly.

1. **CONCLUSION**
2. **REFERENCES**

Bevans, R. (2022) *Two-way ANOVA: Examples & when to use it*, *Scribbr*. Available at: https://www.scribbr.com/statistics/two-way-anova (Accessed: 12 May 2023).

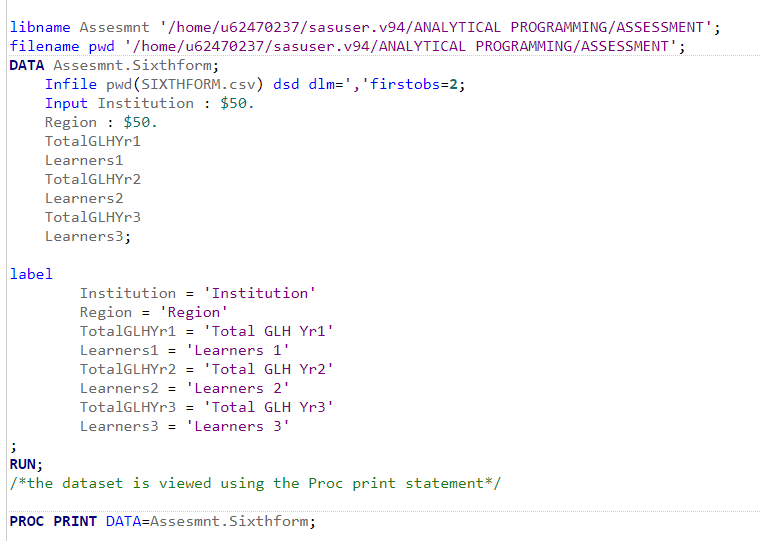
Ross, A. and Willson, V.L. (2017) *One-way ANOVA*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-94-6351-086-8\_5 (Accessed: 12 May 2023).

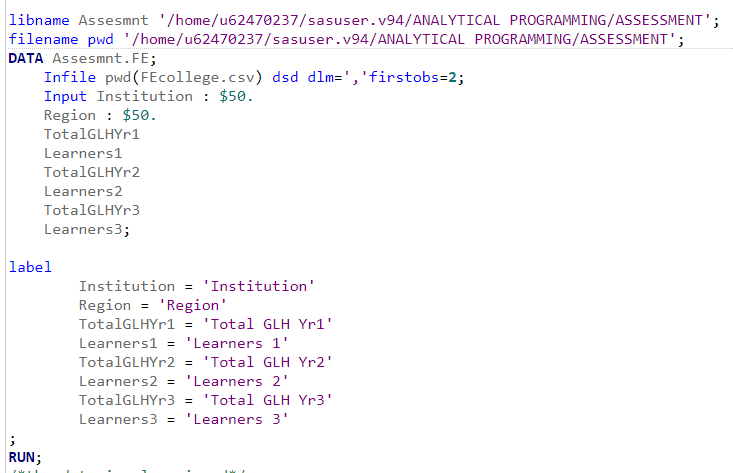
Scheff, S.W. (2016) *Nonparametric statistics*, *Fundamental Statistical Principles for the Neurobiologist*. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780128047538000087 (Accessed: 12 May 2023).

**APPENDIX A**

Creating Libraries and Importing the data for Sixth form and FE college

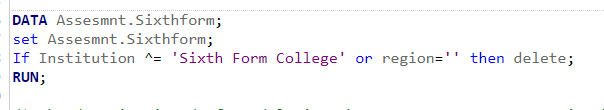
Ross, A., Willson, V.L. (2017). One-Way Anova. In: Basic and Advanced Statistical Tests. SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6351-086-8\_5+

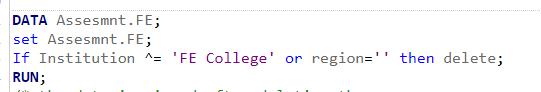




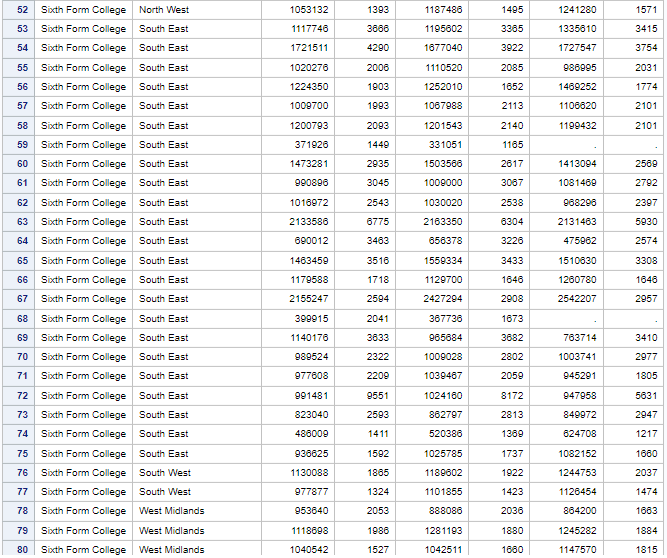
**Appendix B**

Deleting the rows with region total for both data sets.



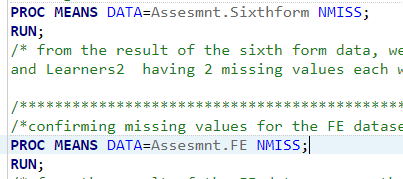


Sample snipshot results showing the rows for region total was deleted for both sixth form and FE college.

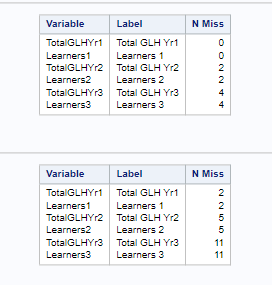


**APPENDIX C**- Missing Value detection and replacement.

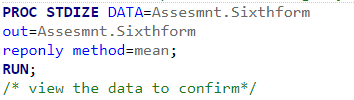
Detecting missing values using proc means

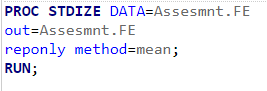


Results showing number of missing values in each dataset

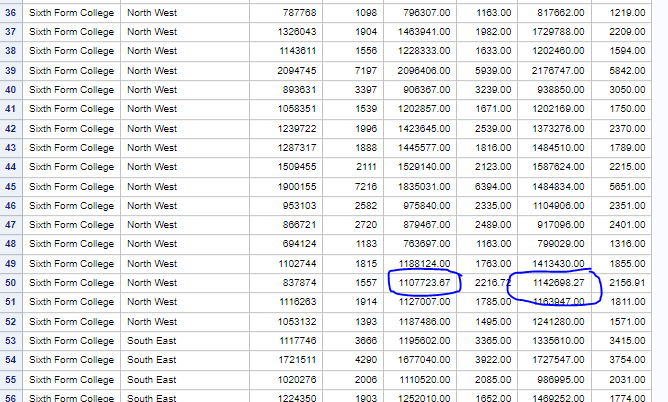


Dealing with missing values using Proc stdize

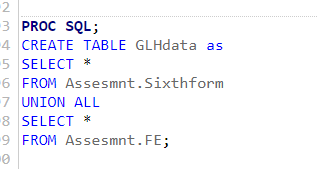




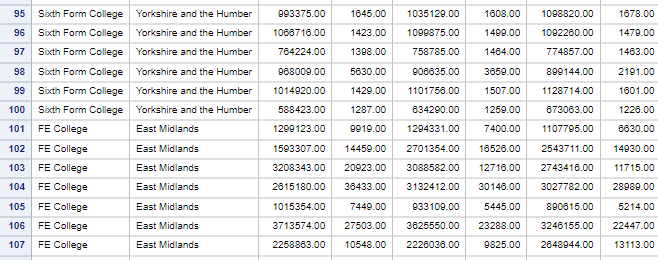
Results showing missing values have been filled with the mean



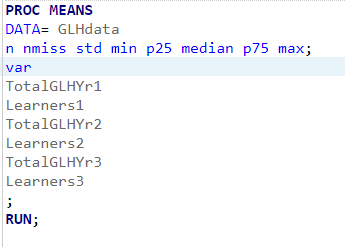
**APPENDIX D**- joining the tables using Proc SQL

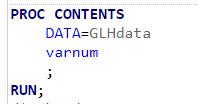


Sample of join result

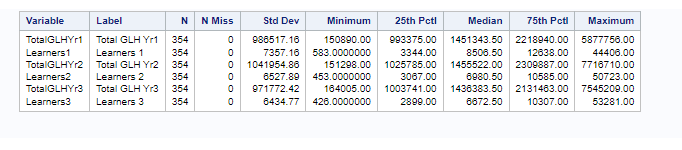


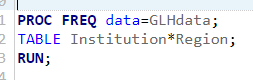
* **APPENDIX E-**Performing exploration data analysis by Identifying frequency distributions, using tablesand plots



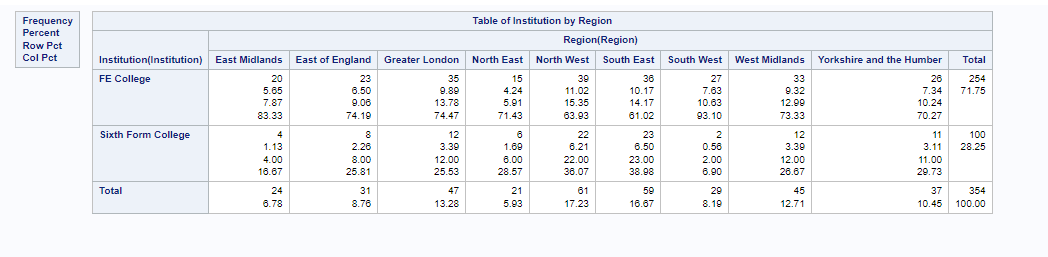


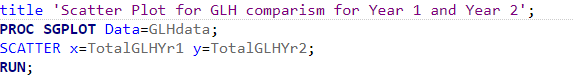
Results of Proc means with statistical summary of the data.

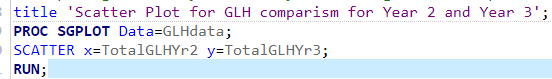


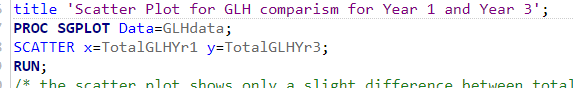


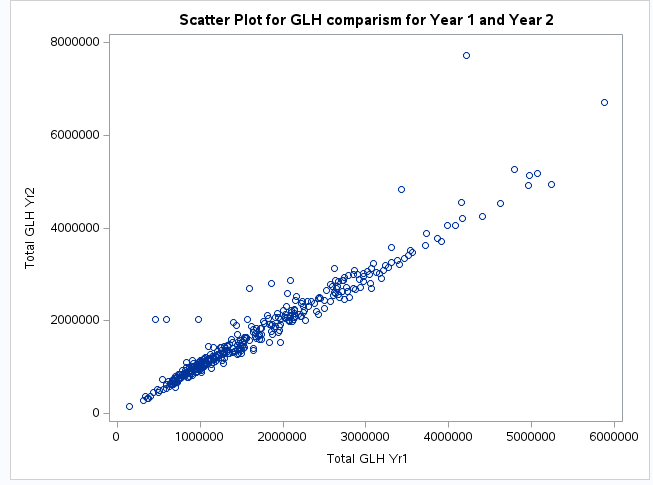
Results

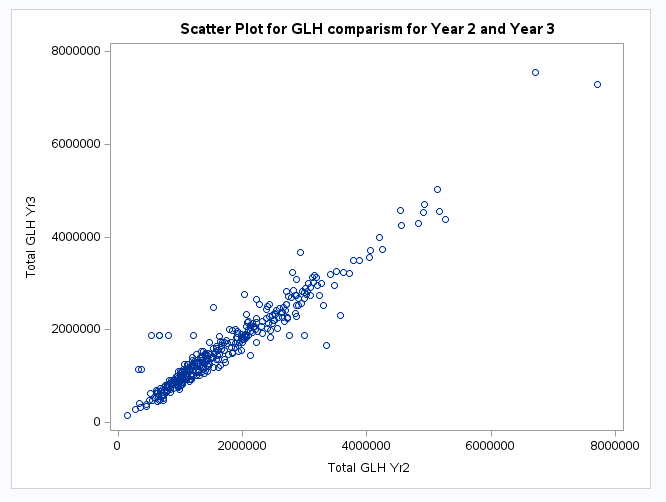


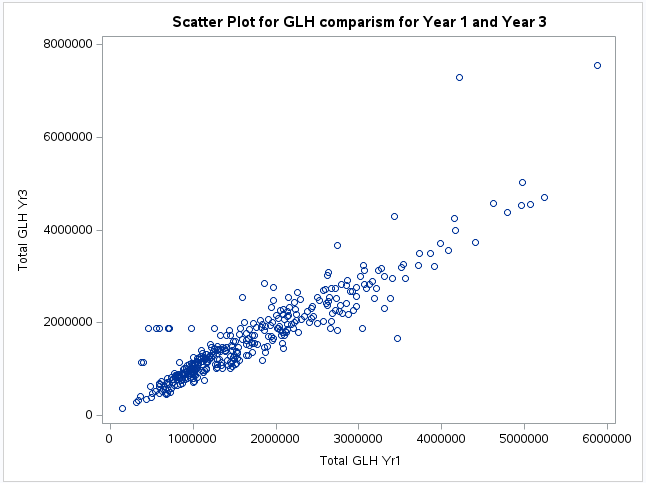




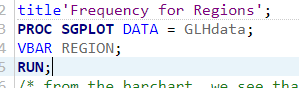


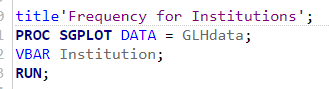




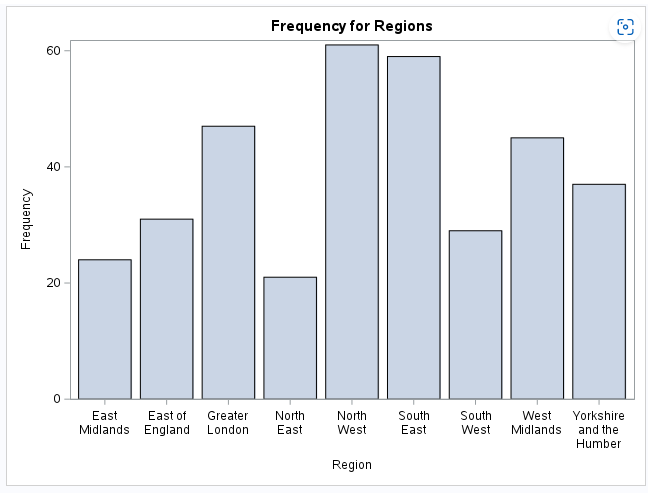


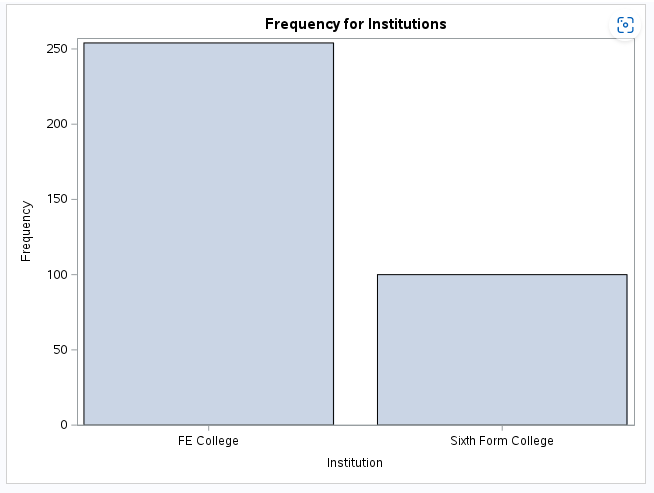
Plotting bar charts for Region and institution



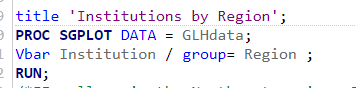


Results

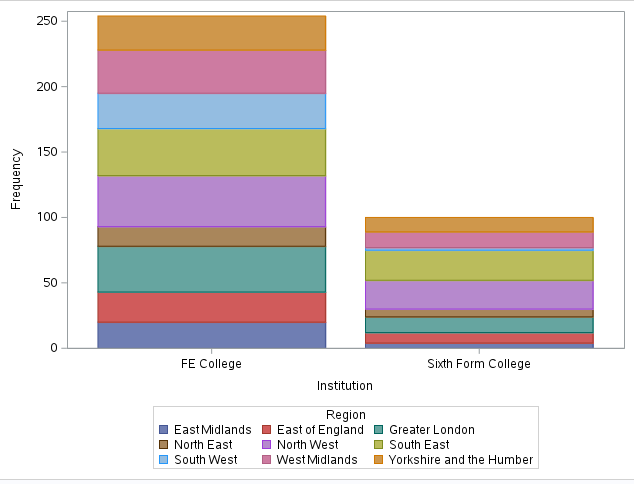




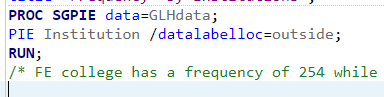
Bar Chart for institution by Region



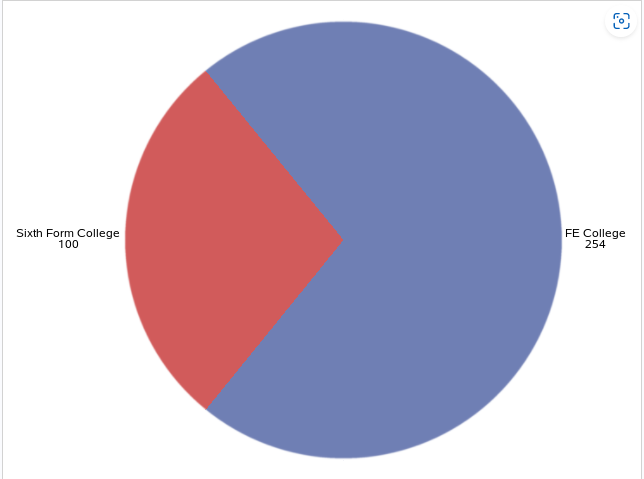
Result



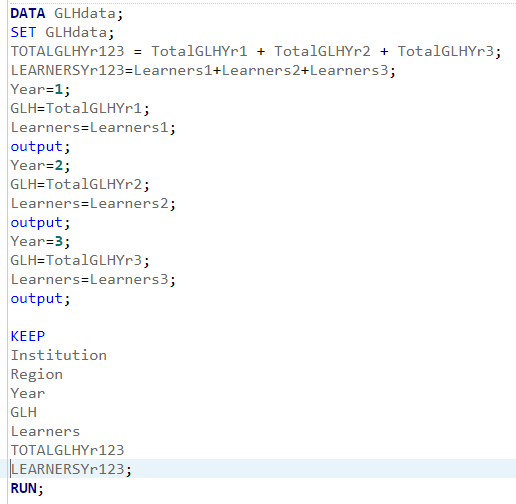
Pie chart for count of institutions



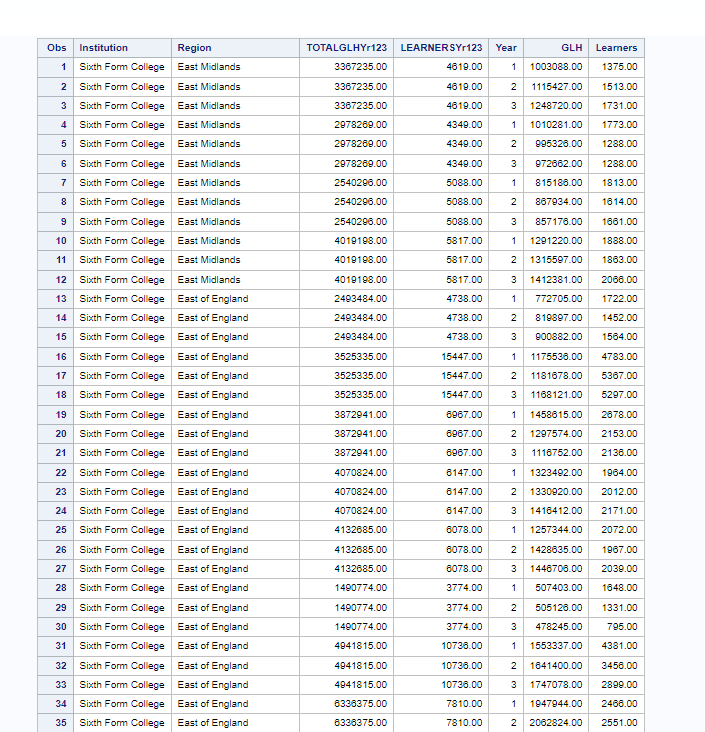
Result



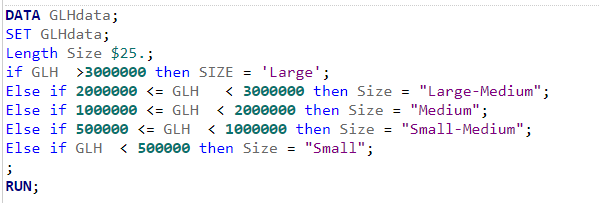
**Appendix F**- creating new variables for Size, Year, GLH per learner with additional columns for Total GLH and learners for year 1, 2 and 3.



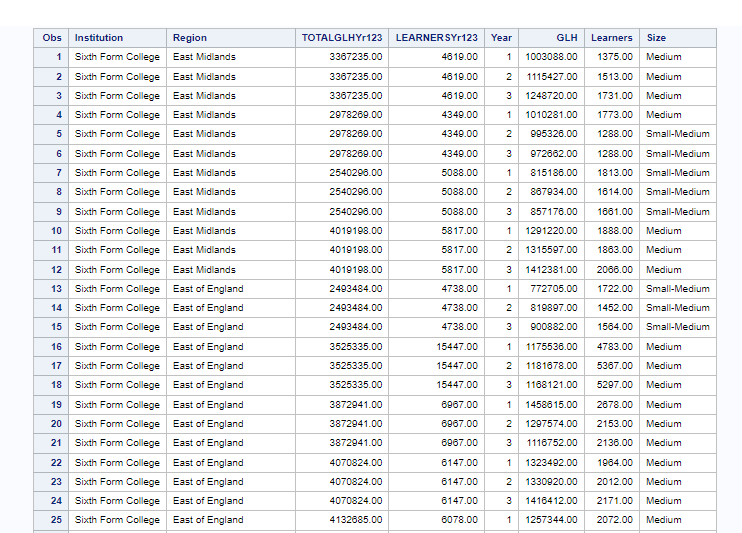
Results



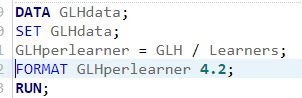
Creating a variable for size



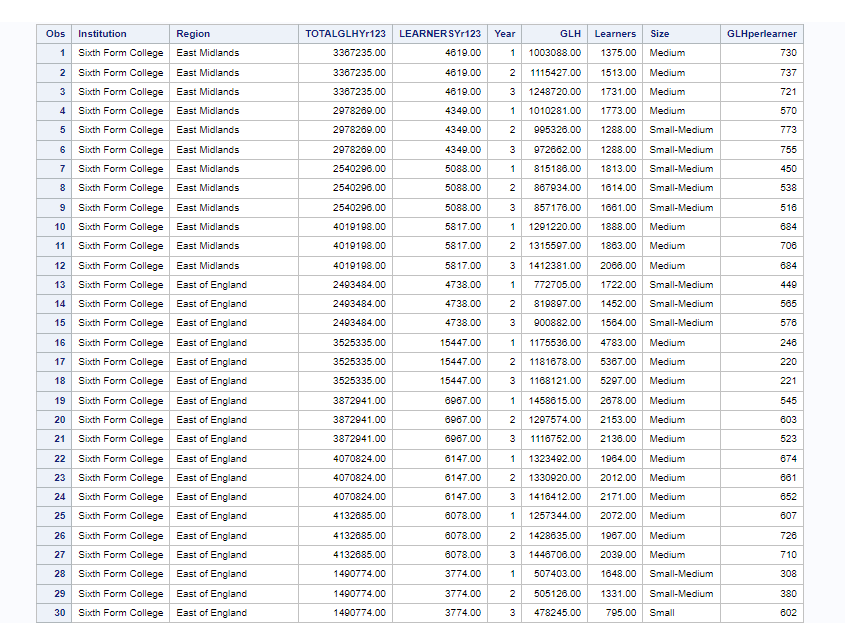
Results



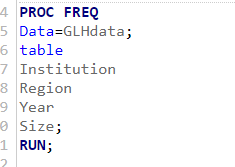
Creating a variable for GLH per learner



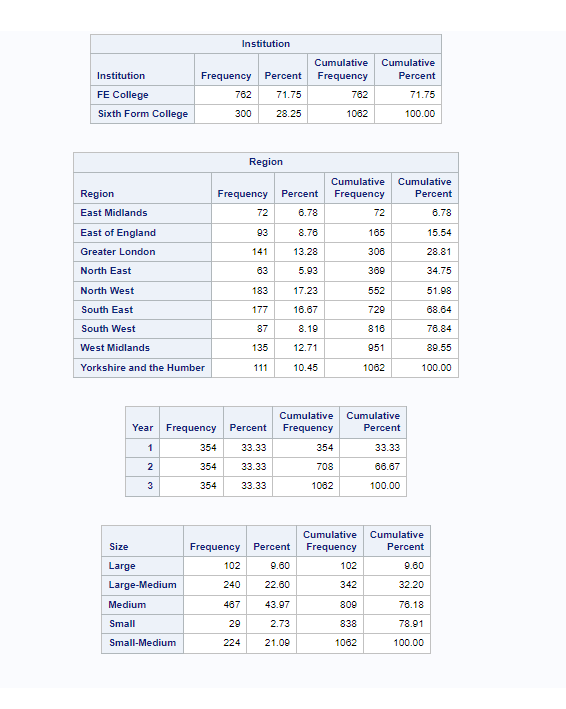
Result



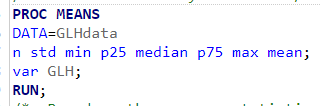
**Appendix G**- Summary of the data using proc means, proc freq, proc tabulate and plotting box plots for

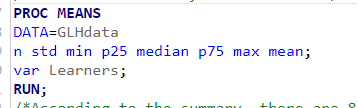


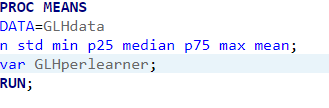
Results

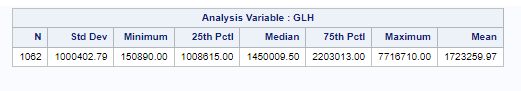
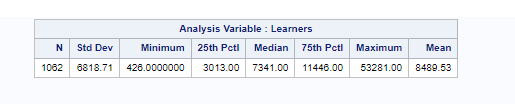


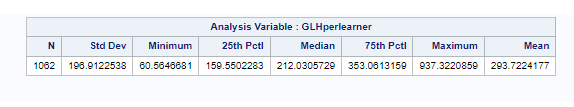
Getting statistical details of GLH, Learners and GLH per learner in order to access them individually



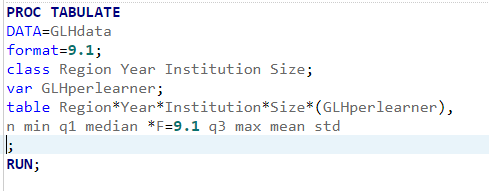




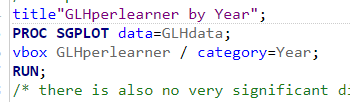
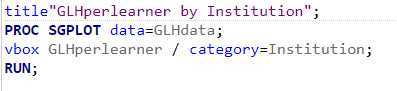


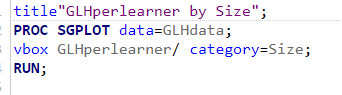
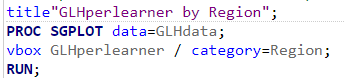


cross sectional view of all the variables with proc tabulate

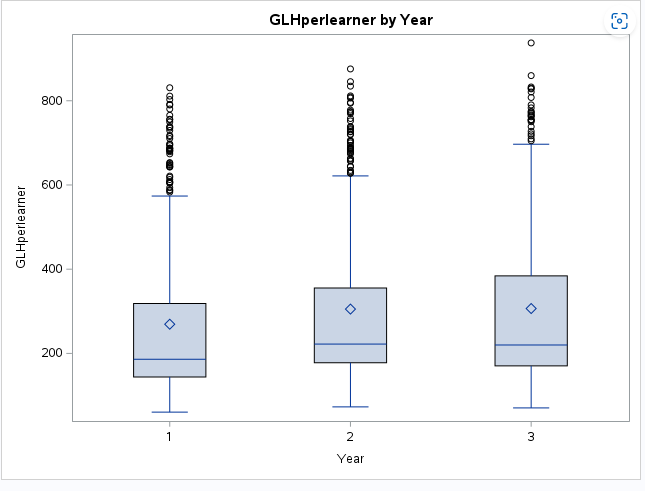


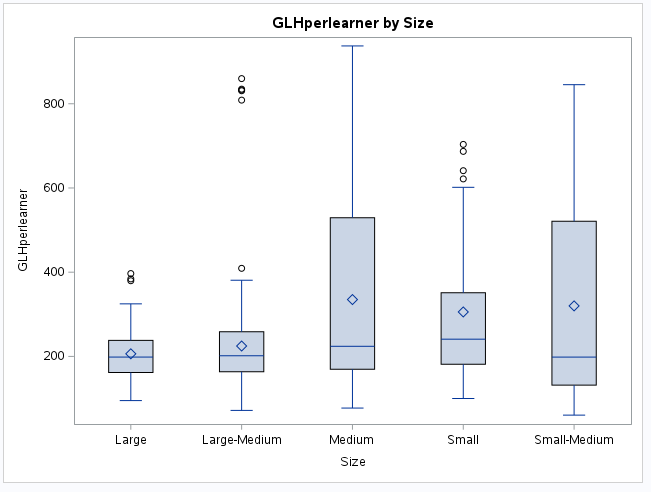
Plotting box plots to check the distributions of GLH per learner visually

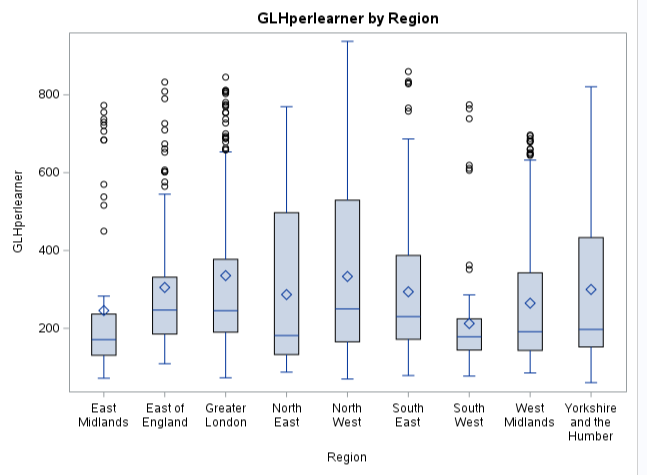
 

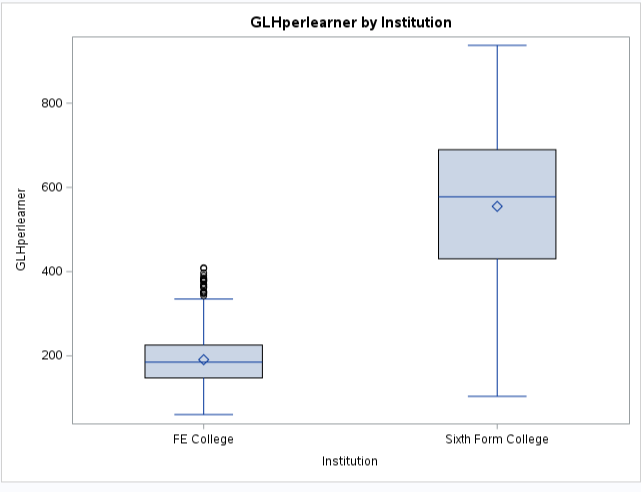
 

APPENDIX Q – Checking distribution of variables using boxplots

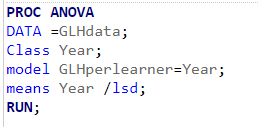


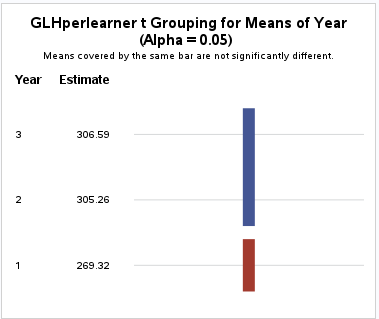




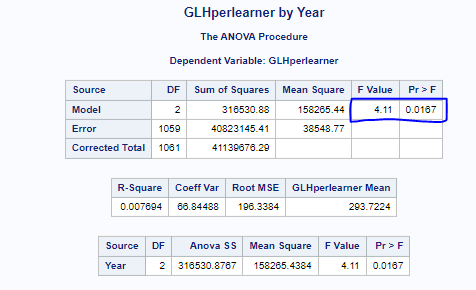


APPENDIX R- One Way Anova for year variable

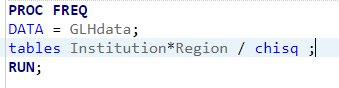


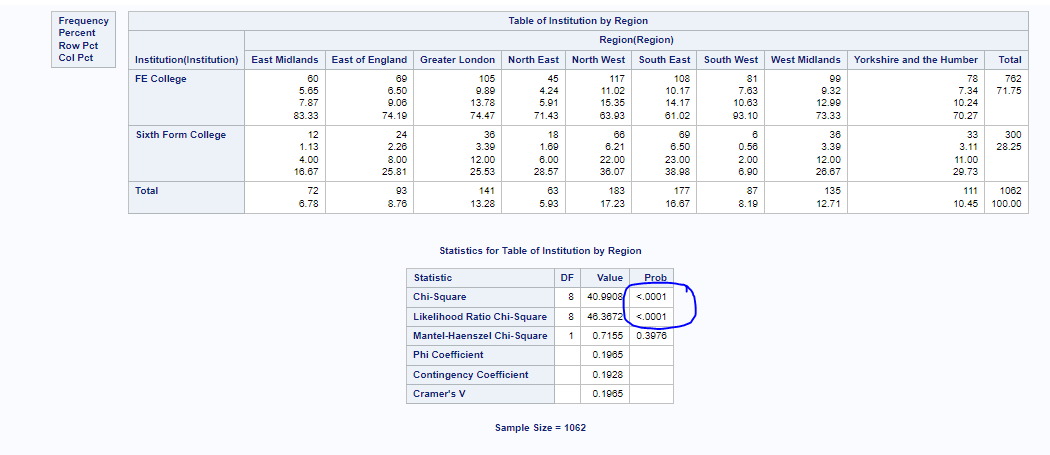


**Appendix H**- One-way Anova

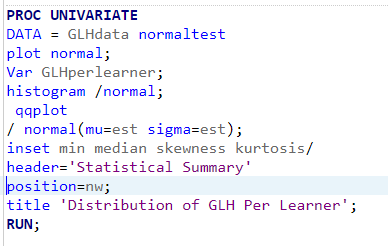


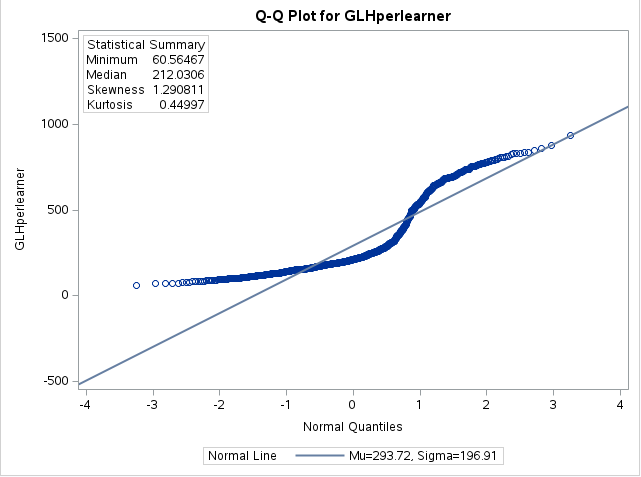
**Appendix I**-Chi square test for assosciation betweeen institution type and region





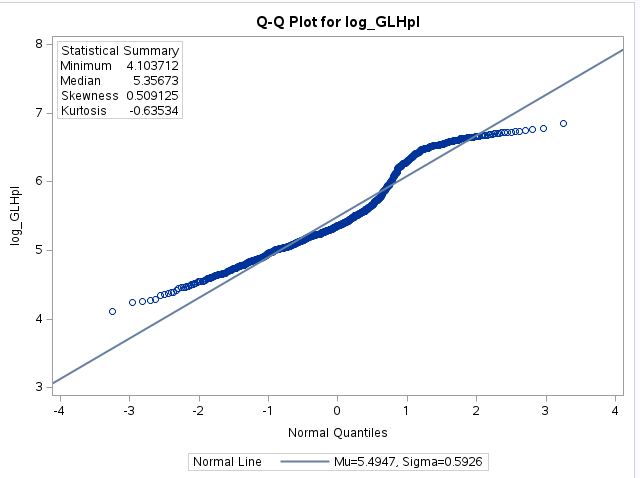
Before

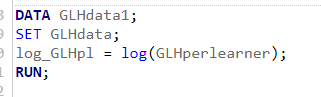
**Appendix J**- testing the dataset for normality of the data using proc univariate



**Appendix K**- Transforming the Data

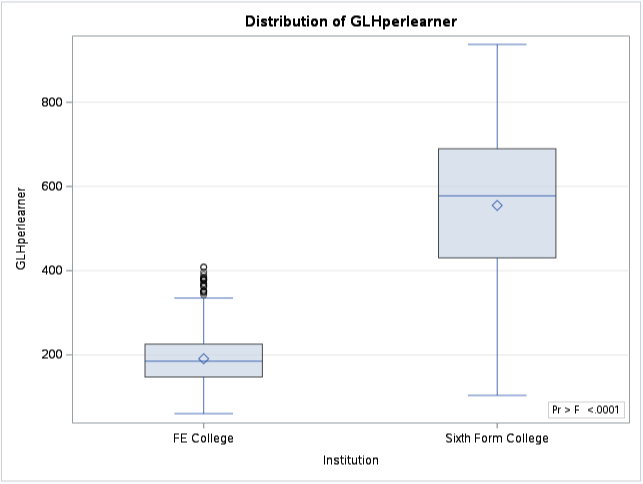
After

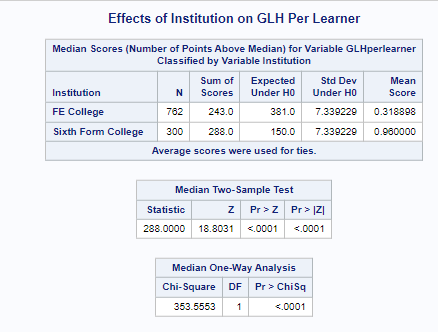




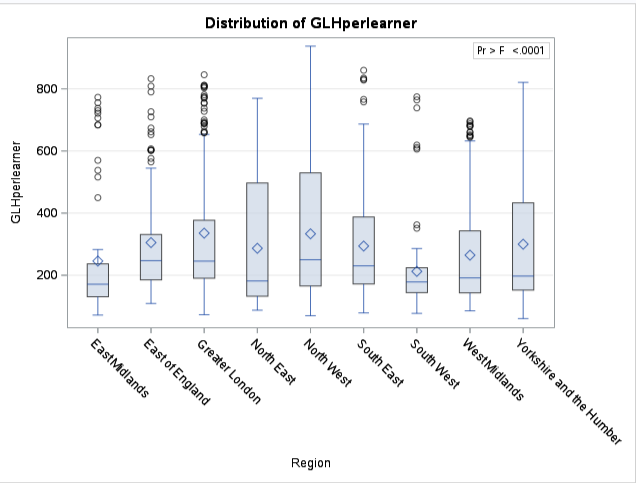
**Appendix L**-WILCOXIN TEST

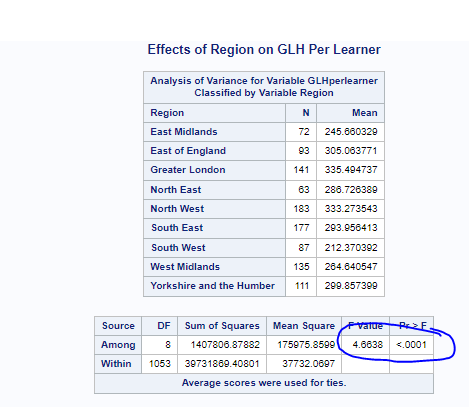
Effect of Institution on GLH per learner

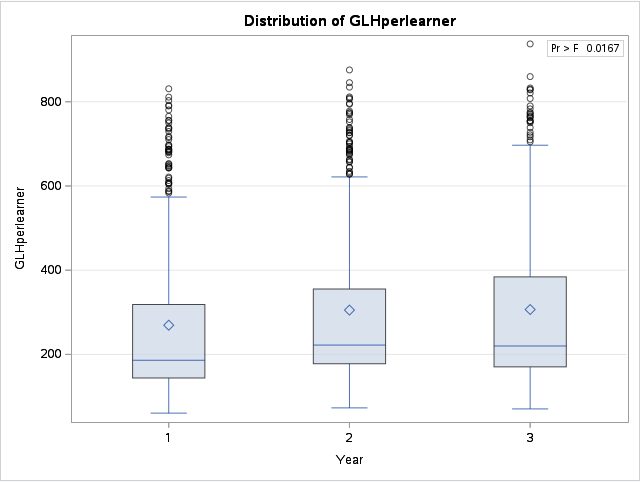




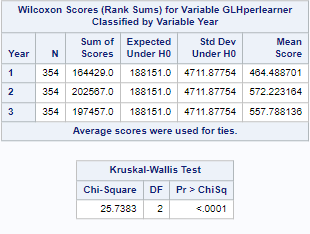
Effect of Region on GLH per learner



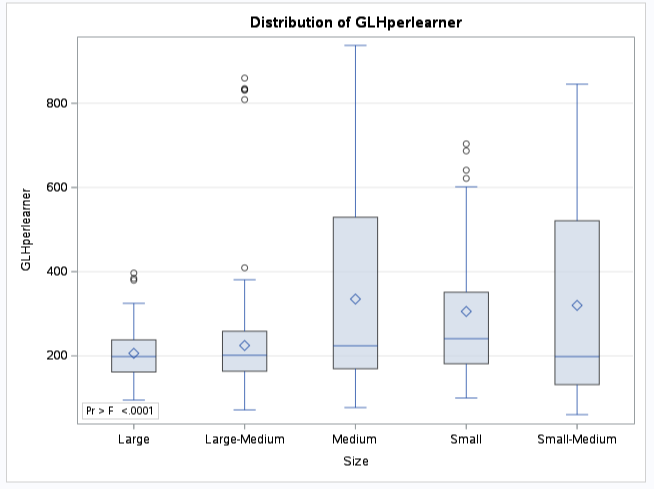


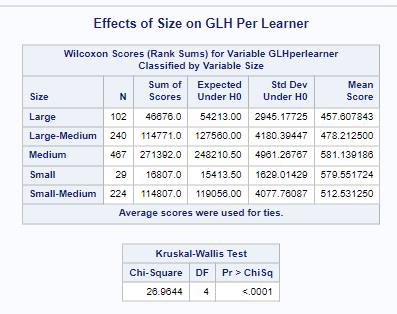


Effect of Year on GLH per learner.

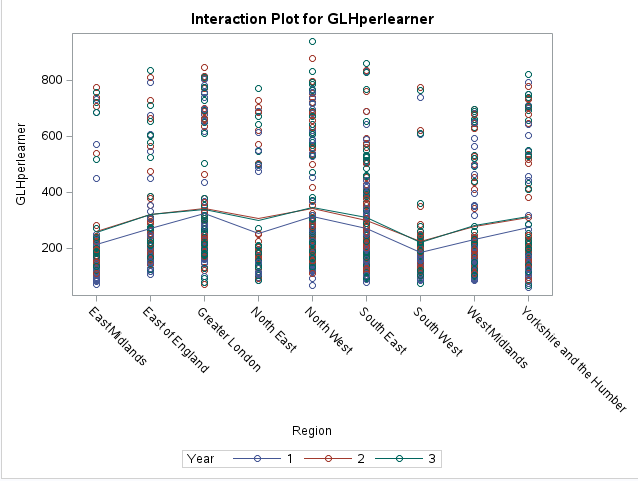


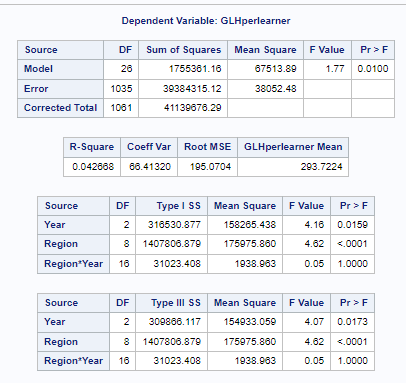
Effect of Size on GLH per learner



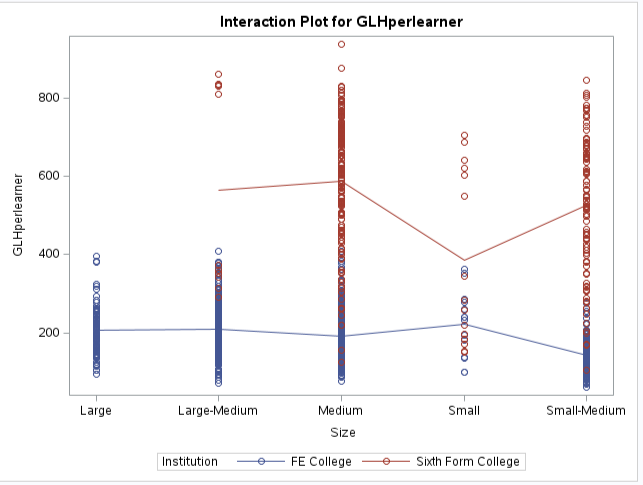


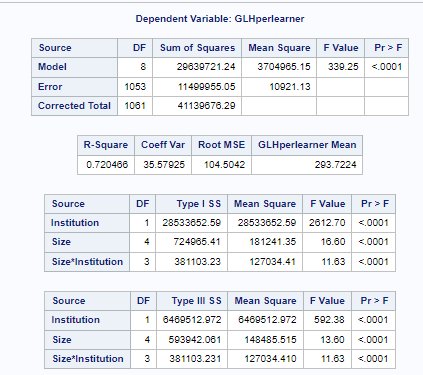
Two-way Anova for region and year





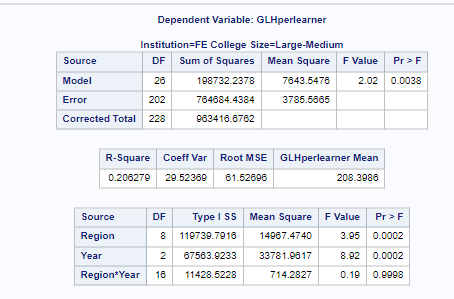
Two-way ANOVA to check institution and size

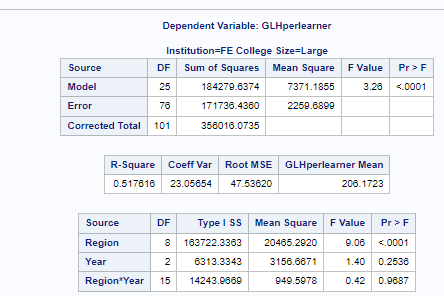




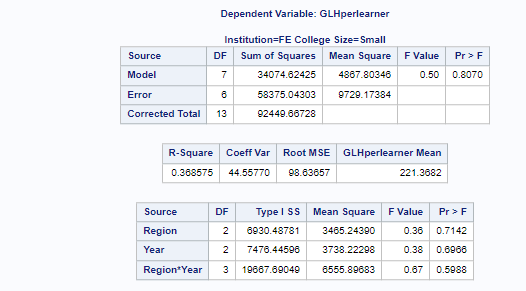
Two-way ANOVA to check the Effect of Region and year by Institution and size

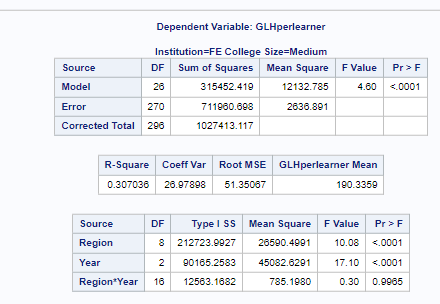
Large FE College Large-Medium FE college



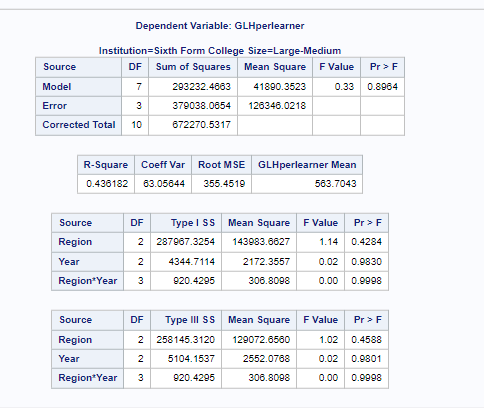


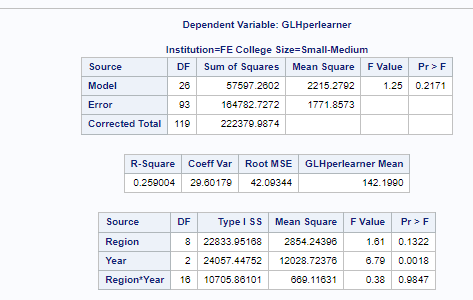
Medium FE college Small FE college



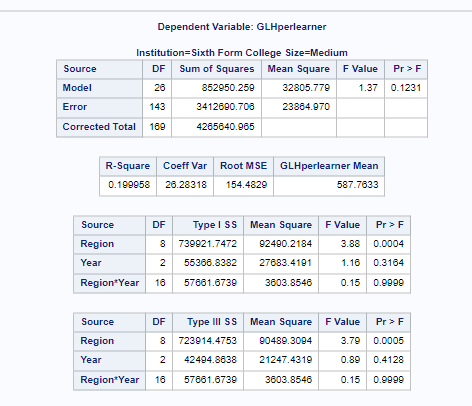


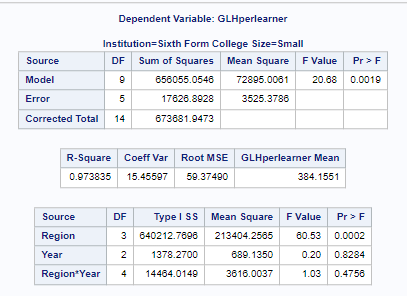
Small Medium FE College Large-Medium Sixth form College





Medium Sixth form college Small Sixth form college





Small Medium Sixth form

