***Report of the Project Done in Partial Fulfilment of the Requirements for the Award of Master of Science in Data Analytics***

**IMAT5168**

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



*Done by*

Juny Maria John, P2743266

*Under the guidance of*

**Anthony Williams**

*Professor at the Department of Computing, Engineering & Media*

Master in Science

In

Data Analytics

Faculty of Computing, Engineering & Media (CEM)

De Montfort University

Leicester – LE1 9BH

**MAY- 2023**

Table of Contents

[SUMMARY 2](#_Toc134964201)

[METHODS 2](#_Toc134964202)

[1. Importing and Anomalies Detection: 2](#_Toc134964203)

[2. Exploratory Data Analysis 3](#_Toc134964204)

[3. The choice of statistical model(s) and relevant validation 3](#_Toc134964205)

[RESULTS 3](#_Toc134964206)

[CONCLUSION 5](#_Toc134964207)

[REFERENCES 5](#_Toc134964208)

[APPENDIX 6](#_Toc134964209)

[Appendix A: Size Classification Table 6](#_Toc134964210)

[Appendix B: Data Visualization 6](#_Toc134964211)

[Appendix C: Anomalies 7](#_Toc134964212)

[Appendix D: TTest 8](#_Toc134964213)

[Appendix E: One-Way ANOVA-Region 8](#_Toc134964214)

[Appendix F: Regression 8](#_Toc134964215)

[Appendix G: Institute size 9](#_Toc134964216)

[Appendix H: Institute Type & Size 9](#_Toc134964217)

[Appendix I: Region & Year 10](#_Toc134964218)

## SUMMARY

This report focuses on analysing various essential aspects related to GLH per learner. The analysis aims to investigate the impact of institutional category on GLH per learner by conducting a t-test. Furthermore, a one-way ANOVA will be carried out to assess whether the region affects GLH per learner. The report will also examine the trend of GLH per learner over time. To determine the significance of institutional size on GLH per learner, a one-way ANOVA will be performed. Additionally, a two-way ANOVA will be conducted to investigate whether the combination of institutional category and size influences GLH per learner. Finally, the report will explore whether the region and year are interrelated and have a combined effect on GLH per learner using a two-way ANOVA. After examining the analysis of each independent variable, i.e., institutional category, region, time, and size, it is found that each variable has a significant effect on the dependent variable - GLH per learner. Additionally, the interaction between institutional type and size has an effect on the GLH per learner, whereas the region and year together do not have an influence on the latter.

**Keywords:** Ttest, Regression, One-way ANOVA. Two-way ANOVA, Wilcoxon

## METHODS

### 1. Importing and Anomalies Detection:

**i. Data pre-processing**: The CSV data files - IMAT5168-FE.csv and IMAT5168-6FORM.csv are imported using the ‘INFILE’ command. Six Form contains a total of 110 records and FE College contains 263 records.

**ii. Data cleaning:** Data Raw data cannot be used for analysis directly therefore it must be clean. Cleaning the data and separating it to SIX FORM, FE College. Sixform Region and FE Region Details**.** Merging the tables containing only the Six form college and FE college contains a total of 354 records.

**iii.** **Feature engineering:** “Institute Size” added to the merged Sixform and FE Institution. The assumption being made here is that GLH\_per\_year\_(X) is being treated as if they are successive years, even though they are not. Since the time period for "GLH\_per\_learner" is not continuous (consecutive years), I feel that taking an average value of each year be a reasonable approach to compare total GLH while engineering the institution size. The data is then categorized into different sizes depending on the GLH, which ranges from Large to Small *(Refer to Appendix A: Size Classification Table)*

**iv. Data Visualization:** From pie chart, we can view the total sum of Institute Type and Region. The total GLH per year and learners across all regions from both Sixform and FE colleges are calculated using ‘proc freq’ *(Refer to Appendix B: Data Visualization)*



**v. Anomaly Detection and Handling missing data:** The data analysis revealed the existence of missing values and extreme outliers. *(Refer to Appendix C: Anomalies)* Despite the identification of the outliers, they were not eliminated from the data as they could contain significant information crucial for the analysis. Nevertheless, the missing values were removed from the data to ensure an accurate set for analysis. (*Refer to Appendix C: Anomalies* for confirming there are no other missing values -‘N-Miss’ shows that missing is removed.)

**vi. Feature Engineering:** The year is extracted from the Total\_GLH\_Year\_(X) column. An ID is been added after which the dataset is transposed. Then numerical values are extracted from the Total\_GLH\_Year\_(X) column assuming that it is successive for the ease of study. The extracted variable called ‘Year’ is created by merging the extracted year data with the original merged dataset. This creates a complete dataset for further analysis.

### 2. Exploratory Data Analysis

The primary research focus of this study is to investigate various factors that can potentially impact GLH per learner. The study aims to answer the following questions: (1) To investigate the effects of Institute type on GLH per learner. (2) To determine if the region has an impact on GLH per learner. (3) The trend of GLH per learner over time will be identified. (4) The significance of institution size on GLH per learner. (5) The interrelation between institution type and institution size and their combined effect on GLH per learner will be explored. (6) The study aims to determine if the region and year are interrelated and their combined effect on GLH per learner.

### 3. The choice of statistical model(s) and relevant validation

(a) To investigate the research questions regarding the impact of region on GLH per Learner and the significance of Institution size on GLH Per Learner, **One-way ANOVA** was implemented. This statistical examination is apt when we intend to establish whether there exists a notable difference in means among two or more groups. Specifically, in this instance, we aim to explore if there exists a significant contrast in GLH per learner across varying regions or institution sizes. By using One-way ANOVA, we can test the null hypothesis that there is no significant difference between the means of GLH per learner in different regions or institution sizes. In the event that this null hypothesis is rejected, it signifies that there is a statistically significant distinction between at least two groups, thereby allowing us to proceed with multiple post hoc examinations such as Tukey and REGWQ to determine the group that differs significantly.

(b) For research questions 5 and 6, I have used a **two-way ANOVA**. The reason for using a two-way ANOVA is that we want to investigate the interrelation between two categorical variables - institution type and size or region and year and whether they together have an effect on a continuous variable which is GLH per learner. Two-way ANOVA can help us determine if there is a significant interaction between variables. Additionally, it can also help us determine if there are main effects of each variable on GLH per learner, regardless of their interaction.

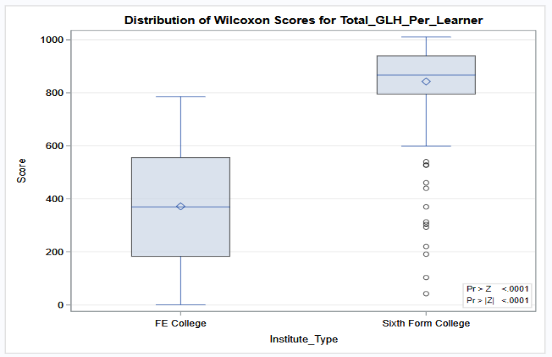
(c) To examine the trend of GLH per Learner over time, we can do so with **a regression** test. Assuming the hypothesis we can determine the trend. Additionally, we can use the R-squared to evaluate the proportion of variation in GLH per Learner that is explained by time. A higher R-squared value indicates a stronger relationship between GLH per Learner and time.

(d)To investigate the effects of Institute type on GLH per learner initially used the **T-test** to check if there is a significant difference between the means of the two Institutes. In this case, we can compare the mean GLH per learner for different types of institutes. The null hypothesis for the t-test would be that there is no significant difference in the mean GLH per learner between the different types of institutes, while the alternative hypothesis would be that there is a significant difference. Since not normally distributed using **Wilcoxon** test to confirm the hypothesis.

RESULTS

The results of the statistical models done are as follows: We assume that the hypothesis:

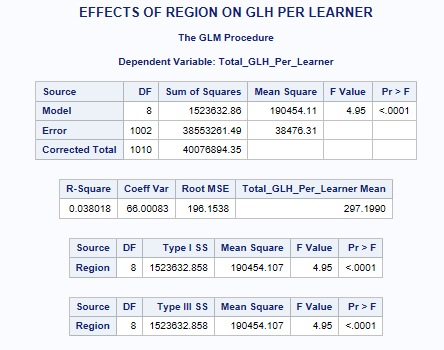
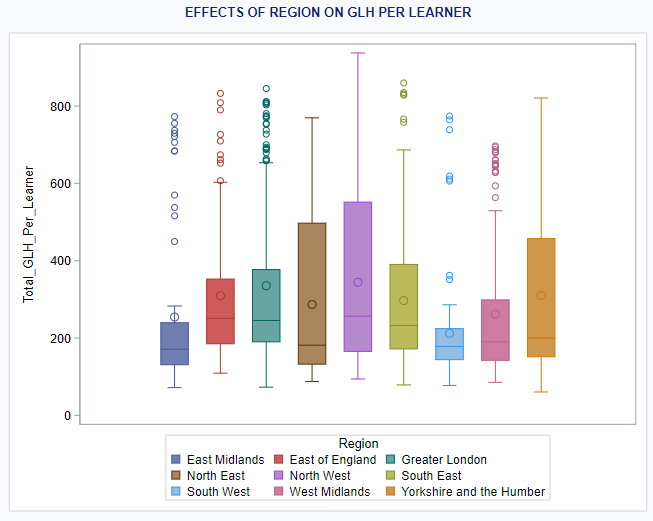
**hO: μ =** **μo and hA : μ ≠ μo ; where is μ is the mean of the variable.**

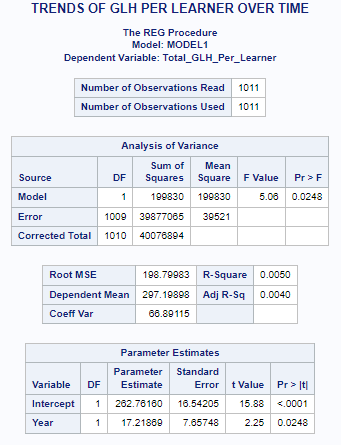
**(1) To investigate the effects of Institute type on GLH per learner.**

Upon analyzing the ttest result *(Refer to Appendix D: TTest)*, it was discovered that there exists a difference of 369.6 units of GLH per learner between FE College and Sixth Form College. The examination for equal variance revealed a p-value lower than 0.05, specifically 0.0001, signifying that the two groups do not possess equal variance. Consequently, the Satterwaite variance must be utilized. The p-value of 0.0001 demonstrates that the difference between the two groups is statistically significant. Although both groups displayed some deviation from the qq plots, this might be attributed to the limited number of participants in the study. Also since the data is not normally distributed, we used the Wilcoxon test for confirmation, from the graph it is clear that the mean of both colleges is significantly different.

**(2)** **To determine if the region has an impact on GLH per learner**.

After examination of the Type III SS, Pr >F it was discovered that the null hypothesis was rejected since the significance level of p <0.05, signifying that there is a significant difference in the mean of GLH per learner based on region. This is given in the GLM Procedure table below. This implies that the region has an impact on GLH per learner. Multiple post hoc tests were conducted *(Refer to Appendix E: One-Way ANOVA-Region)*, and they showed that the region had a significant effect on GLH per learner.

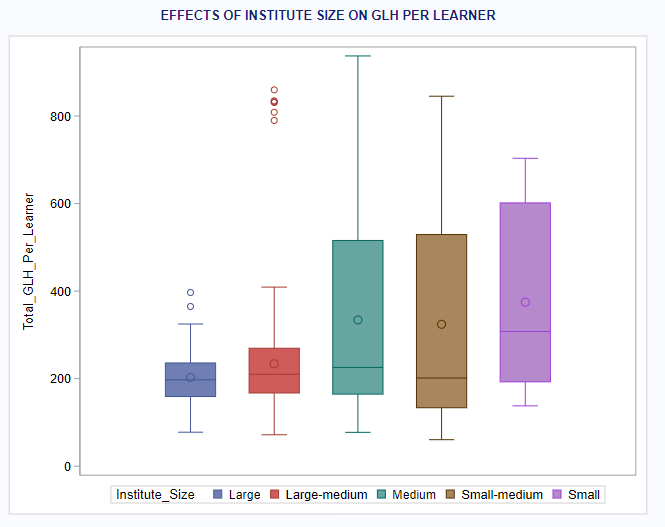
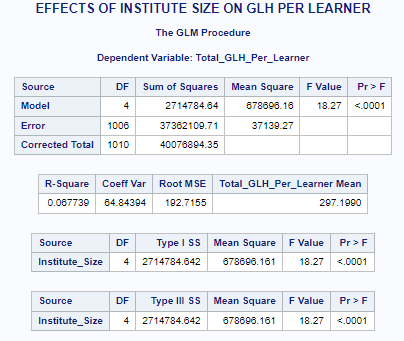


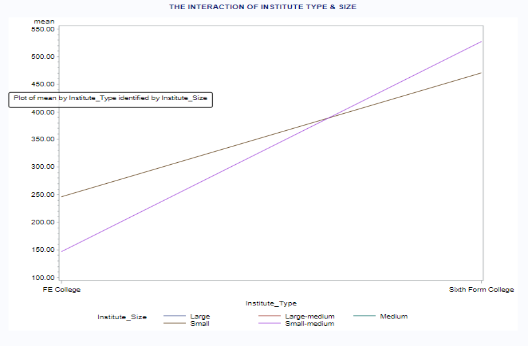
**(3) The trend of GLH per learner over time will be identified.**

Here we are describing a hypothesis test for linear regression. The hypothesis taken here is a little different from the others. Here **ho: slope =0 and hA : slope ≠ 0.** Here from the Analysis of the variance table, we can see that p<0.05. Therefore, the null hypothesis is rejected there is a significant difference in GLH across years The parameter estimates of 17.21 and standard error of 7.64 are used to calculate the t-statistic and corresponding p-value for the hypothesis test. A larger parameter estimate relative to its standard error indicates stronger evidence against the null hypothesis. The R-square value of 0.0050 indicates only a small proportion of the variance in the GLH across Time. The equation to calculate GLH across years = Intercept +Slope\*Year. *(Refer to the Appendix F:* *Regression -* for further graph the trend graph*)*

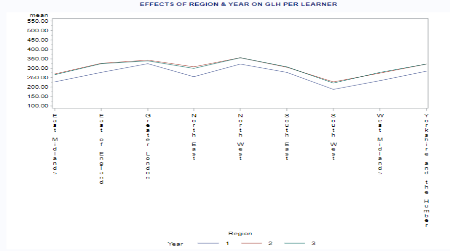
**(4) The significance of institution size on GLH per learner.**

With respect to the significance of “Institute Size” since there are multiple size groups and we want to determine if there is a significant difference in GLH per learner between these institute sizes, we are using ‘proc glm’ in ANOVA, The ANOVA table produced will provide information on the significance of the effect of size on GLH per learner. The p-value associated with the F-test for the "Institute\_size" variable < the significance level (typically 0.05), therefore can conclude that there is a significant effect of Institute Size on GLH per learner. The graph shows the difference in means between different size categories (*Refer to Appendix G: Institute size*)

**(5) The interrelation between institution type and institution size and their combined effect on GLH per learner will be explored.**

From the gplot plot, we can understand there is interaction. To prove this, we check the values with and without interaction. Here we are using proc glm for the two way ANOVA analysis, *(Refer to Appendix H: Institute Type & Size –* for the ANOVA Table*)* From the result we can understand that, that the interaction term (Institute\_type\*Institute\_size) has a p-value of 0.0028, which is less than the significance level of 0.05, indicating that the interaction term is statistically significant. Based on this result, it is necessary to keep the interaction term in the model. Furthermore, the statement indicates that the Institute\_type and Institute\_size variables have a strong relationship with the outcome variable, GLH per learner, based on the significant interaction term

 **(6) The study aims to determine if the region and year are interrelated and their combined effect on GLH per learner.**

Based on the graph, there is an interaction between the second and third years. To confirm this, a hypothesis test is conducted using the p-value. If the p-value is greater than 0.05, it implies that there is no substantial interaction. The ‘proc glm’ table reveals that the interaction term Region\*Year has a value of 1.000, indicating no significant interaction. As a result, the model is refitted without the interaction terms. The new p-value obtained from the model is 0.0001, demonstrating strong proof that the area has an influence on the GLH per learner. Conversely, the year has a p-value of 0.0283, implying a less significant effect on the GLH. *(Refer to Appendix I: Region & Year –* for the ANOVA Table*)*

## CONCLUSION

In conclusion, the study has provided substantial evidence to suggest that diverse factors affect GLH per learner. The research indicates that the type of institute, region, size, year, and interactions are all related to the GLH per learner. The t-test showed that there is a significant difference in GLH per learner between FE College and Sixth Form College. The one-way ANOVA test revealed that the region has a significant impact on GLH per learner. The regression test proposed that there is a significant variance in GLH per learner across years, and the GLH per learner across years can be predicted using the equation: Intercept + Slope\*Year. The ANOVA test using glm showed that Institute size has a significant impact on GLH per learner. Finally, the two-way ANOVA analysis revealed a significant interaction between institute type and institution size, indicating a combined effect on GLH per learner. Overall, the study findings suggest that multiple factors play a role in determining the GLH per learner, and these factors should be considered while making policy decisions related to the Learners.

## REFERENCES

[1] *Step-by-step programming with base SAS 9.4, second edition* (2019). S.l.: SAS Institute.

[2] Bailer, A.J. (2020) *Statistical programming in SAS*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

[3] Notes from classroom lectures.

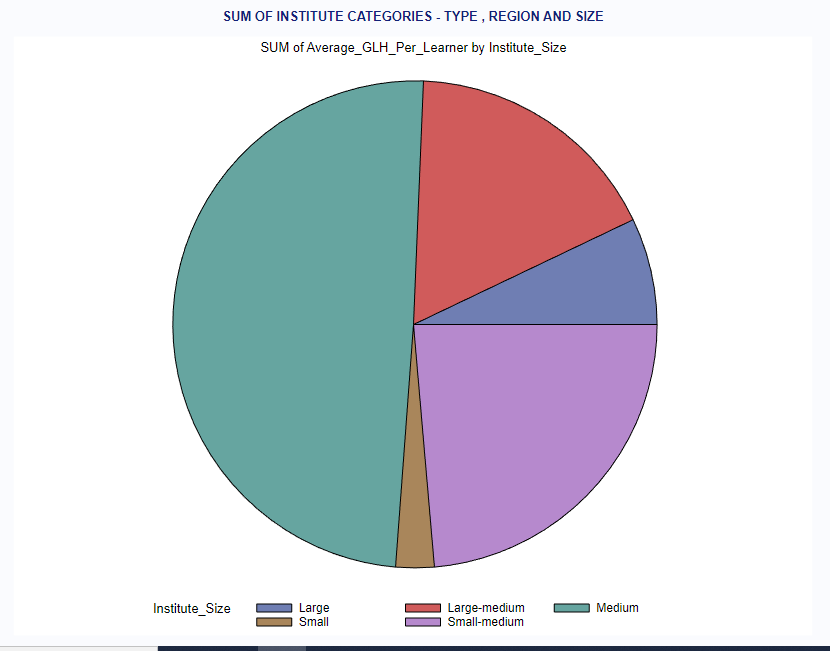
## APPENDIX

### Appendix A: Size Classification Table

|  |  |
| --- | --- |
| **Total GLH** | **Size** |
| greater than 3,000,000 | Large |
| between 2,000,000 and 3,000,000 | Large-medium |
| between 1,000,000 and 2,000,000 | Medium |
| between 500,000 and 1,000,000 | Small-medium |
| less than 500,000 | Small |

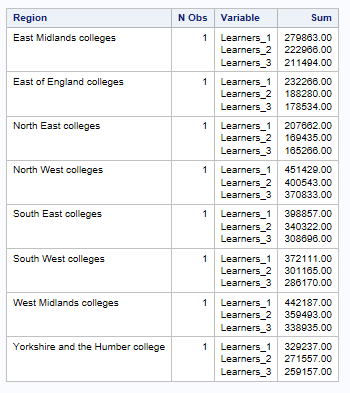
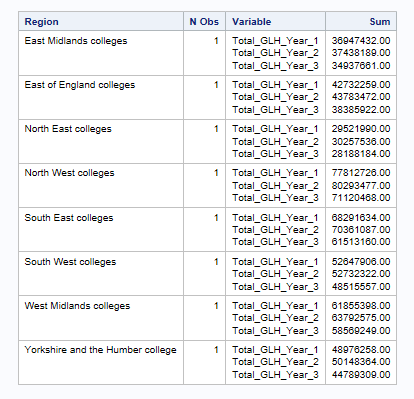
Based on the average GLH across 3 years, the GLH is categorized into different sizes.

### Appendix B: Data Visualization

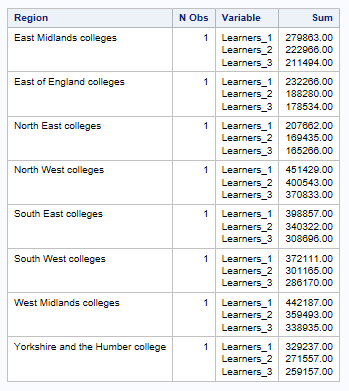
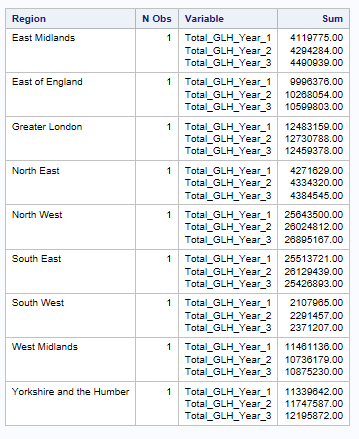


This shows the Total GLH plotted based on the different size categorization- ranges from Large to Small. Given below is the Cumulative sum of all the GLH\_per\_Year\_(X) where X=1,2,3 and Learners\_(X) where X=1,2,3 for the Regions in both FE and Sixform Seperately.

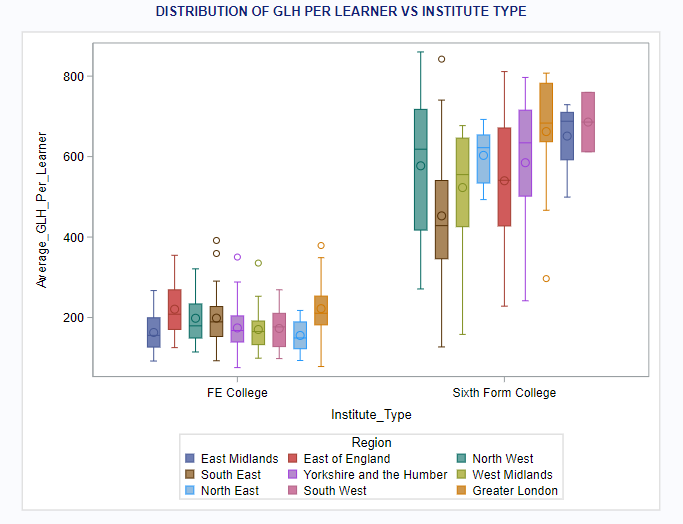
i.For FE Colleges:



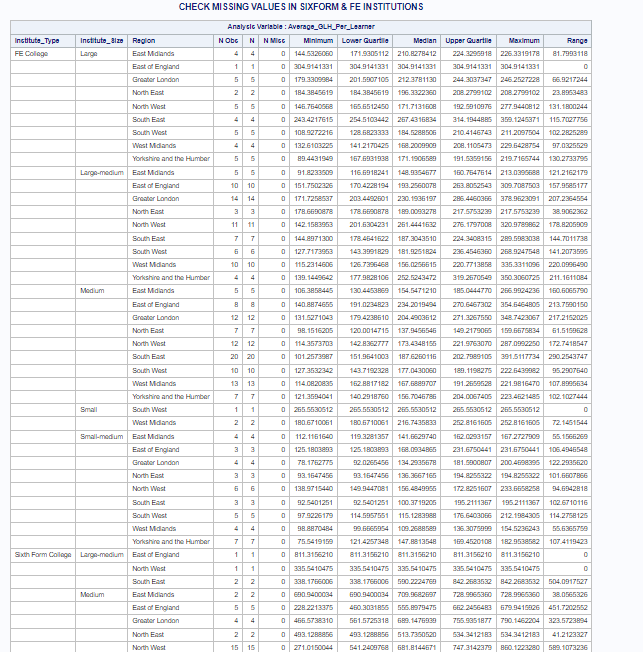
ii.Sixform Colleges:



### Appendix C: Anomalies

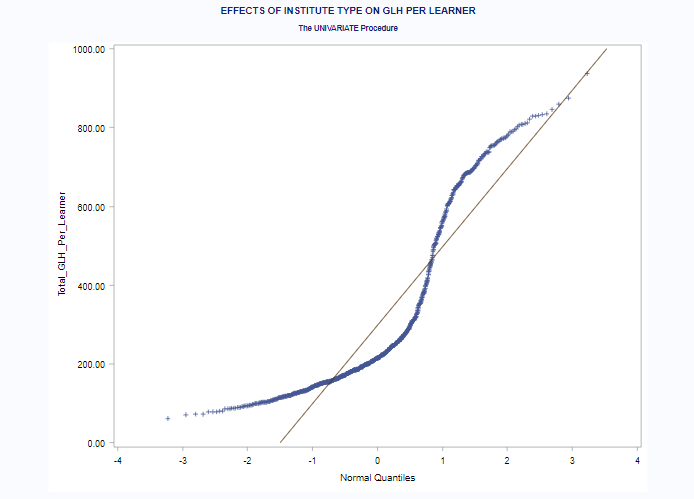
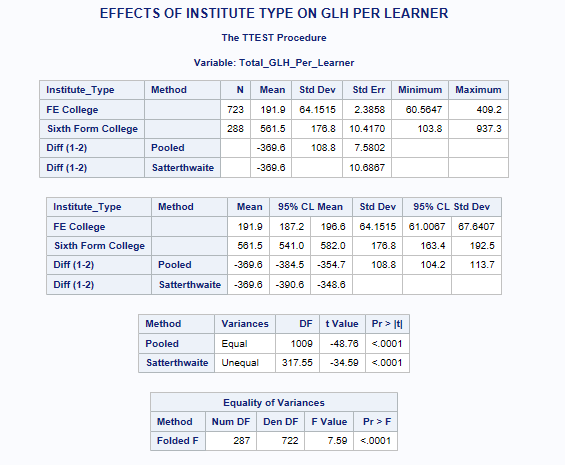


The denotation outside the whiskers shows the extreme outliers. They are not removed so as to keep variety in the dataset and also to prevent important information from been removed. (These information can be crucial in the analysis).



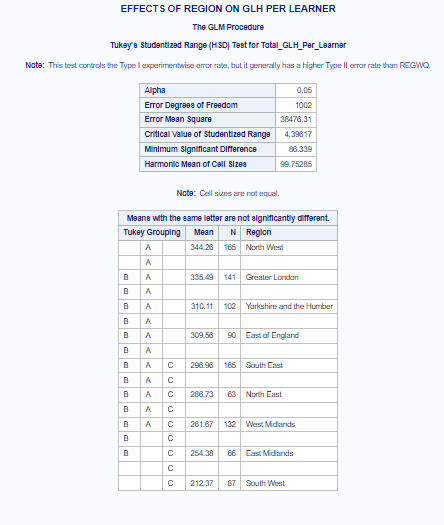
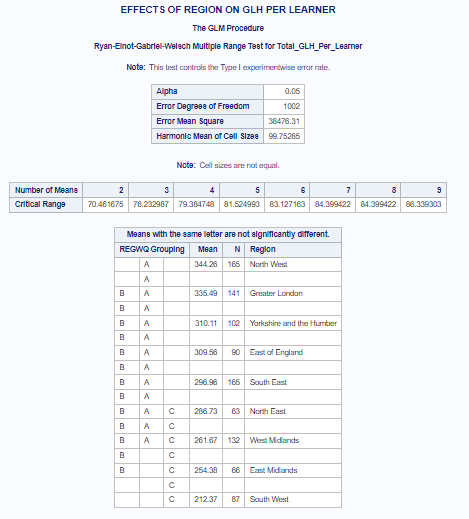
The table above shows each variable (type, size and region) been categorised and checked for missing values. NMiss gives the number of missing. [Small portion of the table taken for explanation purpose].

### Appendix D: TTest

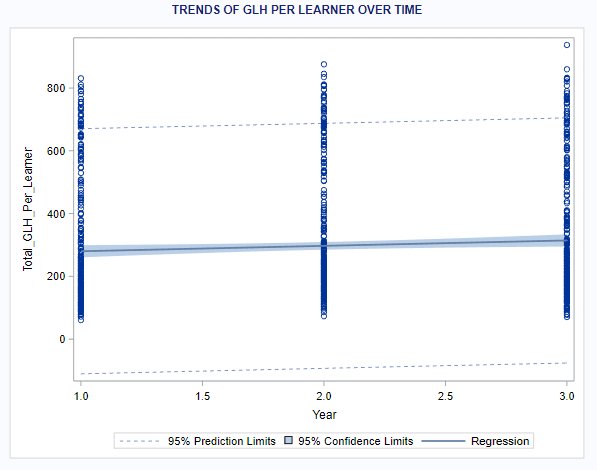
QQ-Plot : this shows that is not normally distributed

### Appendix E: One-Way ANOVA-Region



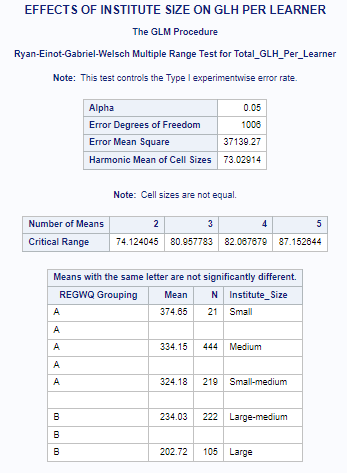
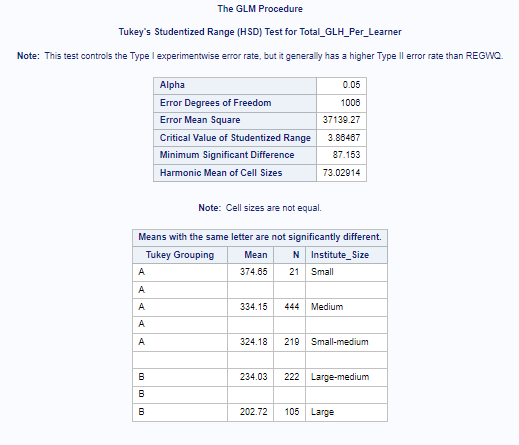
This shows Multi posthoc comparison tests- Tukey and REGWM method applied on the independent variable. This shows which all region is related and which all is not.

### Appendix F: Regression



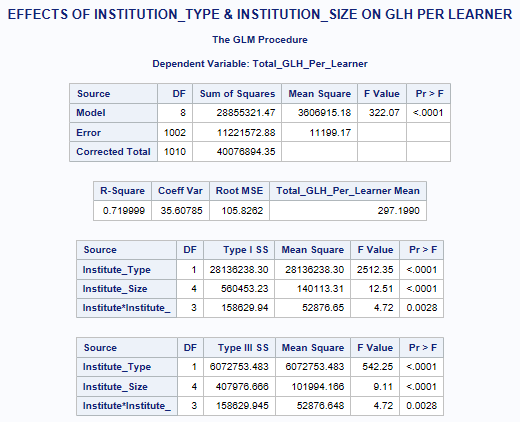
The graph shows slight increase in GLH based on the trend. There is variation in mean between the years.

### Appendix G: Institute size



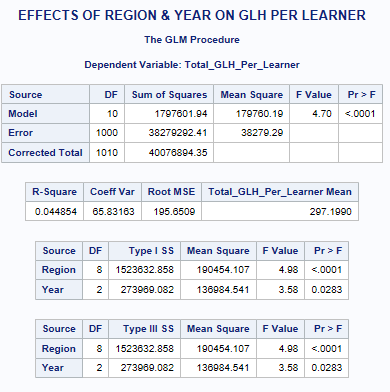
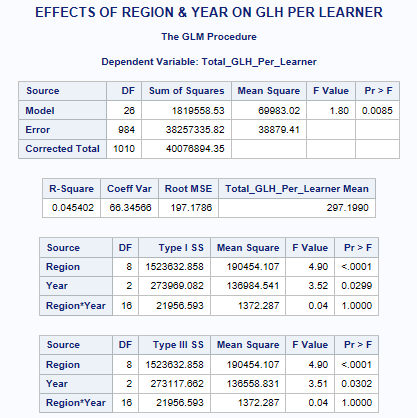
This shows Multi posthoc comparison tests- Tukey and REGWM method applied on the independent variable. This shows which all size is related and which all is not.

### Appendix H: Institute Type & Size



2 way ANOVA Table Result.

### Appendix I: Region & Year



The table with and without Interaction term.