Oladipupo Hilton

De Montfort University | 2686917

Report on Guided Learning Hours in Sixth Form and Further Education Colleges

IMAT5168: ANALYTICS PROGRAMMING

CONTENTS

[1. SUMMARY 2](#_Toc104199479)

[2.METHODS 3](#_Toc104199480)

[DATA IMPORTATION AND CLEANING 3](#_Toc104199481)

[IDENTIFYING ANOMALIES 3](#_Toc104199482)

[Exploratory data analysis 4](#_Toc104199483)

[Checking for Normal Distribution of Data 4](#_Toc104199484)

[Data Transformation 4](#_Toc104199485)

[Modelling 4](#_Toc104199486)

[ANALYSIS AND DISCUSSION 5](#_Toc104199487)

[Effect of Size, College, Region and Year on GLH Per learners 5](#_Toc104199488)

[Effect of Region and Year on GLH Per learners by Size and College Type 6](#_Toc104199489)

[REFERENCES 9](#_Toc104199490)

[APPENDIX 10](#_Toc104199491)

# SUMMARY

This report details the analysis of the Guided Learning Hours (GLH) in sixth form and Further education colleges across regions in the United Kingdom for three distinct years. SAS Studio was used to import, clean and run parametric and non-parametric tests and also determine models for the data. Models used for the analysis include General Linear Model, Analysis of Variance (ANOVA), T-test of means and test for normal distribution (Univariate analysis). It was discovered that there is a major difference between the GLH in the colleges. It was also discovered the GLH per learners were slightly different for each year, and the region of the institution was a significant determinant of the amount of GLH each learner received. Size on the other hand, which was determined by the total GLH for each college in the observation did not have a significant impact on the GLH per learner in various colleges.

# 2.METHODS

The various methods used on SAS studio to import, clean and analyze the data set on Guided Learning Hours in Further Education and Sixth Form Colleges across England are discussed concisely below.

## DATA IMPORTATION AND CLEANING

The data sets used for the analysis were stored as individual Comma Separated Value (CSV) files containing the guided learning hours for each college type (Sixth Form College and Further Education college). Both files were imported into SAS by using a SAS predefined Macro consisting of Proc Import and Proc Content, to check for the data type in Columns and to examine the data. Each imported file had 8 columns, with the data for Further Education Colleges having 263 observations and that for Sixth form colleges having 110 observations, also there were missing values in certain columns. After this, infile and input statements were used to read each file in as Data steps, specifying column name, length, delimiter type, and other required settings based on the details seen earlier when importing the data.

Both created data tables were merged together vertically, by using the ‘SET’ statement, after the merge. calculated variables were created from existing variable columns using calculated functions and if else statements. Calculated functions were used to create the required variable ‘general learning hour per learner’ for analysis. If else statements were used to create categorical values for Sizes of institutions using the Total Guided Learning Hours. Input and output formats were created afterwards for the Size of an Institution enabling the data to show the sizes of different locations. Appropriate Lengths and Labels for each column were also clearly defined.

Rows with missing values, earlier observed in the data were deleted from the data tables using an If Statement, because only a few rows had missing values. Certain columns containing aggregate sum of specific rows (depicting the total for a particular region) were deleted from the tables as this would interfere with the objective of the analysis.

After importation and necessary cleaning were carried out the total number of observations for the created Table was 337.

Table 1: frequency distribution of class variables

## IDENTIFYING ANOMALIES

Proc Freq was used to check for the distribution of the observations in the data. The data is made up of many categorical variables therefore it had columns with duplicate values which were relevant to our analysis.

Proc Univariate was used to check for outliers (extreme observations) within the data. This provided information on the minimum and maximum values of the data and how far apart they were distributed. A table showing the extreme values was created afterwards using Proc SQL, this helped to further understand the positions and places in the data where the outliers occurred in. It was discovered from this table that there were outliers due to a major difference in Guided Learning hours in the two institution types, with Further Education Colleges having the lowest extreme values and Sixth Form Colleges having the highest extreme values. Due to this finding the outliers were therefore not removed. A table showing this can be seen in the Appendix.

## Exploratory data analysis

### Checking for Normal Distribution of Data

The data was explored to check if the distribution was normal which is a statistical requirement for further data exploration and analysis. Proc univariate was used once again, this time to generate histogram and qq plots (Shown in Appendix). Both plots and the p-values of the goodness of fit test showed the distributions were highly positively skewed and not normal. The skewness for Total GLH per leaner for all three years was 1.33, GLH per learner for each of the years was also greater than 1.33 (Shown in Appendix). Statistically, a normal distribution should have a skewness very close to zero, in this case, the p-values were less than 0.01, which is not fit enough for required statistical analysis. Finally, the sgscatter matrix plots showed many clustered values of the data away from the normal line (shown in Appendix). It was therefore concluded that the data was not fit enough for further analysis and needed to be transformed as the data.

### Data Transformation

To transform the data, box transformation functions were written using Proc FCMP. Different transformations such as log, square, cube, square root and reciprocal square root were tried and reciprocal square root (rcpsqrt) functions proved most effective. This transformation for GLH per learner for Year 3 yielded the most normal curve than the other transformations. The skewness was 0.448 and the Shapiro-Wilko test score was 0.083.  Shapiro-Wilko test score greater than 0.05 indicates a normal distribution. Also, rcpsqrt yielded a more normal curve for GLH per learner for all other years and the aggregate, all having skewness of less than -0.098. The initial GLH per learner for each of the years and the year aggregate were therefore dropped for the transformed variables to cary out parametric analysis.

## Modelling

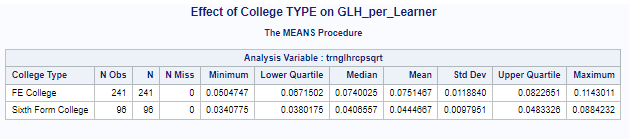
SAS Proc GLM, Proc T-test and Proc Anova was used to model the parameters. GLM is considered optimal when it comes to linear predictions (Rutherford, 2000). T-test is considered when it comes to comparison of means between independent groups (Tae Kun, 2015). One-way Anova was also used to compare the means of two or more groups (Ross and Willson, 2017).

GLM was used to model the effect of size, Region and College Type against the GLH per Learner for all the years. The proc univariate procedure was then used to test if the distribution of the errors were normal. T-test was used to model College types against GLH per learner for all the years. Anova was used to model size and region against the aggregated GLH per learner for all the years.

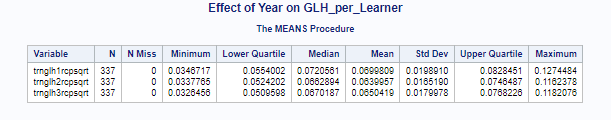
# ANALYSIS AND DISCUSSION

## Effect of Size, College, Region and Year on GLH Per learners

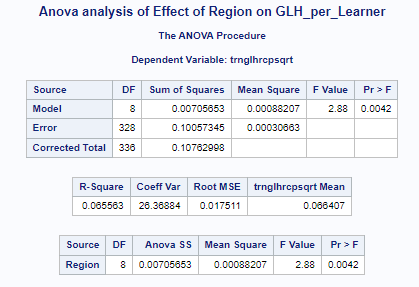
The College type was discovered to have a significant relationship with the GLH per learners for each year and aggregated years. Using the difference between the means(t-test) of Further Education colleges and sixth form colleges, the p-values were less than 0.05 and the pooled t-value was 24.37 which shows huge difference in year aggregate of GLH per learner for each school (as shown in the figure below and appendix).



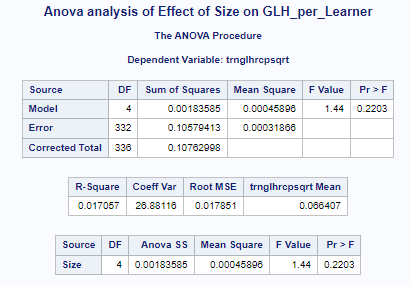
Looking at the year there is no significant difference in guided learning hours in the three years, with the mean value of each year not far apart from one another. However, Year 1 had a high standard deviation value which showed some difference from other years.



Based on region, there is a major difference in guided learning hours in the college regions. The p-value and sum of squares is less than 0.05. There is also no significant difference between the colleges in the following regions South west and South East, South West and East of London, South West and Greater London and South West and North West, East midlands and East of England, East midlands and Greater London, East Midlands and North West, West midlands and Greater London, West midlands and North West.



Based on the Size of the colleges there is no significant difference in the sizes of colleges and their GLH per learner. The p-value for the anova analysis is 0.22 and the sum of squares is 0.0018(as shown in the appendix). This means, there is no relationship between the size and GLH per learner of each school. The box plot also shows there is no major difference in the distribution of their values.



## Effect of Region and Year on GLH Per learners by Size and College Type

Looking at the means across the years by the college sizes there is a consistent gap between the mean values of GLH per Learner for each size category. The medium and smaller sizes had more significant learning hours per learner than the larger ones who had large total learning hours but lower learning hour per learner.

For each of the years by the college sizes the p-values for the results of the general linear model tests including Shapiro-wilk test and the sum of squares (Type III SS) is less than 0.05 which indicates there is a significant difference between sizes across the years. The test for normality of the residuals also showed the p-values were less than 0.05 and the skewness and kurtosis also proved the distribution was close to a normal distribution as they were less than 1. The major levels with large differences include Large-Medium and Small-Medium, Large-Medium and Medium (all years) and Medium and Large in Year 2, and Medium and Small Medium in Year 3.

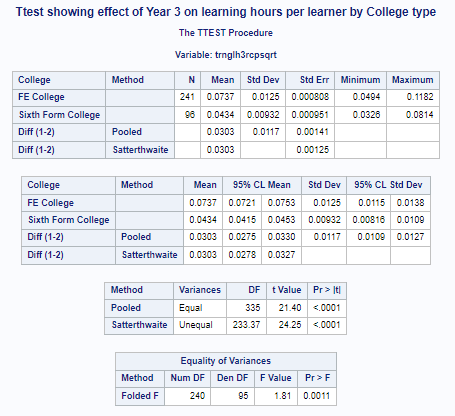
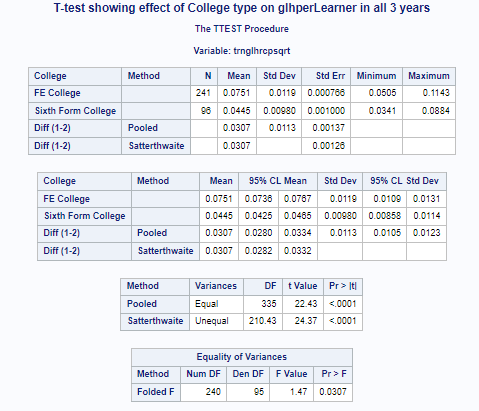
Examining the effect of college type and year on GLH per Learner, the P-value for the pooled and Satterthwaite is less than 0.01 for all years showing there is major difference in the two college types in all the years. There is also a large difference in the means of their guided learning hours. For the each of the years 1, 2 and 3 the mean difference between the college types was 0.00337, 0.0278 and 0.0303 respectively. Examining the equality of variances both year 1 and 3 had p-values lower than 0.05 showing there is significant difference in the variance between the learning hour per learner in both colleges. Year 2 on the other hand had a value of 0.1690 showing the variances of the GLH per learner are not significantly different between the college type in that particular year. The normal distribution plot also showed that the GLH per learner values for FE colleges are more normally distributed than Sixth Form colleges.

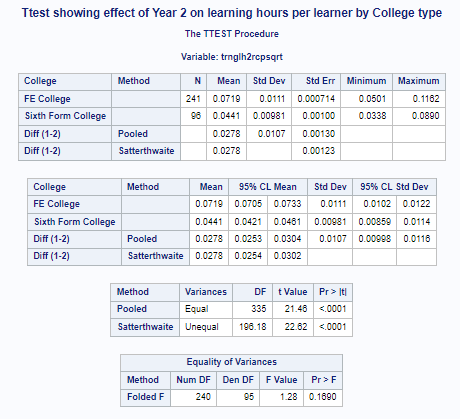
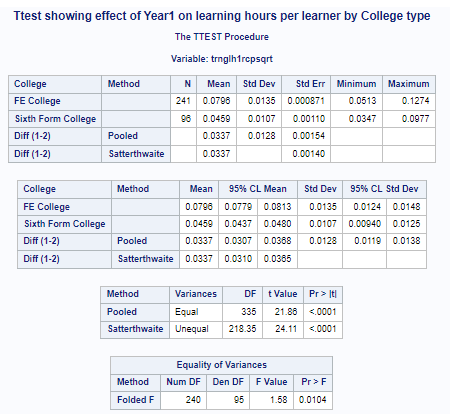
Analyzing the region and year by size using two-way Anova, it was observed that with interaction the model was not significantly different across the years with a P-value greater than 0.05(0.5795). On the other hand, without interaction there is a significant difference in the GLH per learner for different college across the years with p-values less than 0.05. Size and Region both individually have a significant effect on the GLH per learner. Using the LS means, it was observed that East midlands, West midlands, Yorkshire and South West have similarities between their large and large-medium schools.

Examining the region and year by college type, there was interaction between the region and the college type in Year1 and in year2 with their p-values less than 0.05 showing that there was a significant effect of college type, region and year on the GLH per learner in years 1 and 2. In Year 3 the p-value being 0.0803, meant there was no significant interaction of the college type and region in determining the GLH per learner for students. The Least Square means in year 1, 2, and 3 also shows there is a major difference in all college types. Also, for the regions the Least Square means showed that Greater London, North West and East of England had mean values which are significantly different from other regions across the years.

CONCLUSION

From the analysis the college type had the major effect on the guided learning hour per learner as sixth form colleges had way higher values than Further Education Colleges. The Regions also had major effects on the values of Guided Learning Hours with Greater London and North west having more than other regions consistently for different years. Using the years, there was no major difference in guided learning hours however Year1 had a significant effect on Guided Learning Hours per Learner. The Size of schools using their Total Guided Learning Hours did not also have a significant effect on the Guided Learning Hour per Learner as majority of the data fell within two classes (Large-medium and Medium).

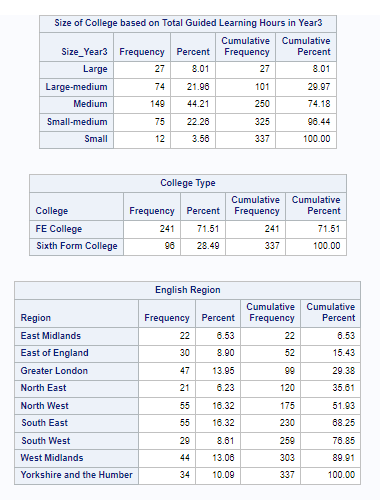
 

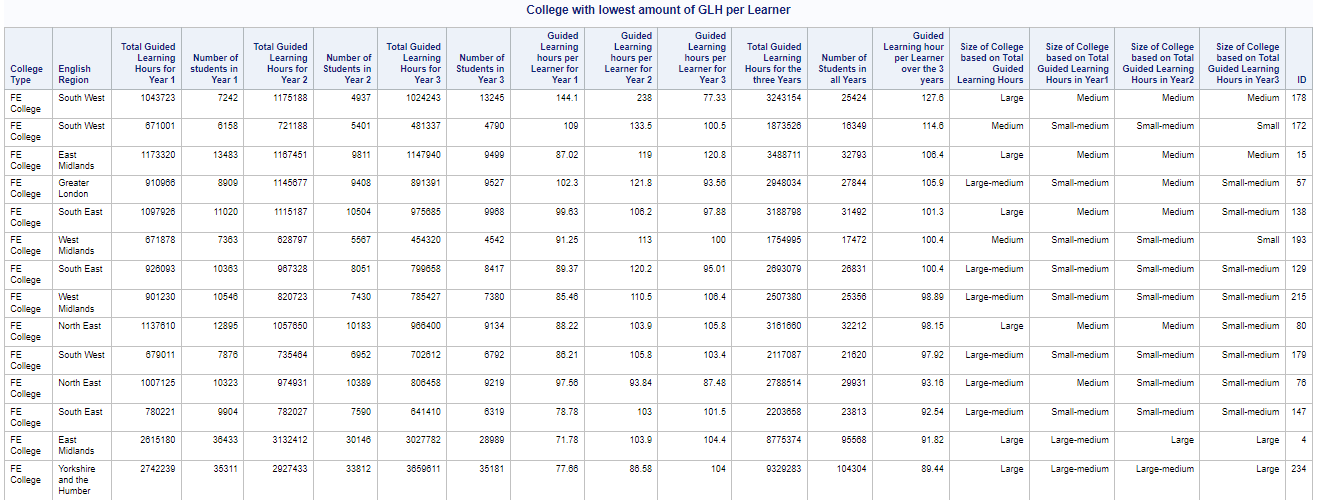
 

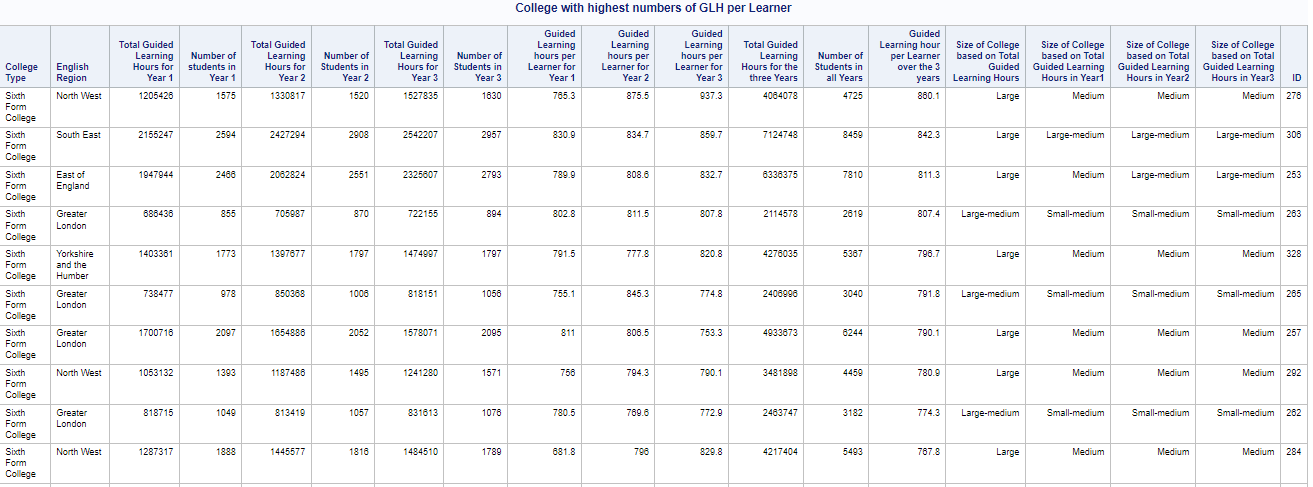
# REFERENCES

1. Rutherford, A., 2000. Introducing anova and ancova. *Introducing Anova and Ancova*, pp.1-192.
2. Kim, T.K., 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), p.540.
3. Wilson, A.E. and Ross, M., 2000. The frequency of temporal-self and social comparisons in people's personal appraisals. *Journal of personality and social psychology*, *78*(5), p.928.

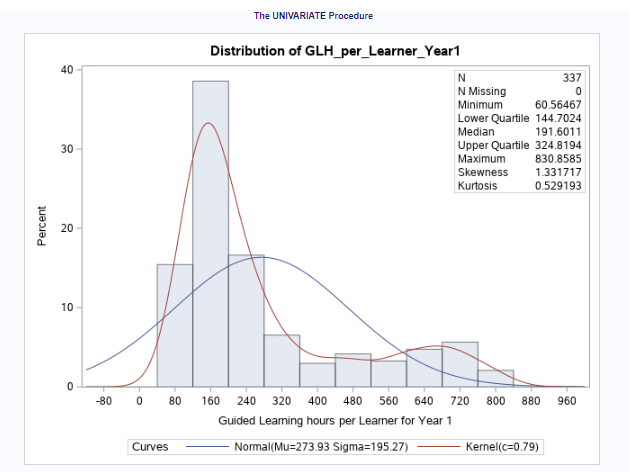
# APPENDIX



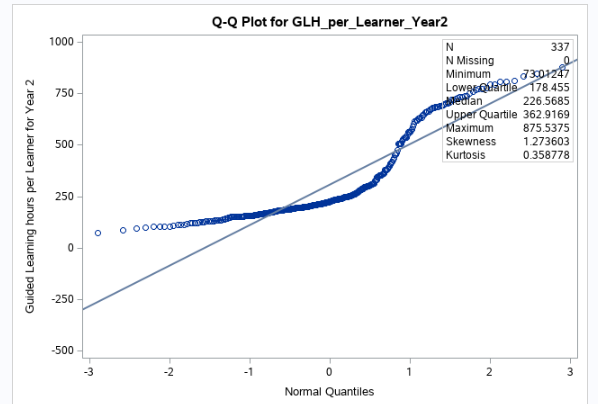


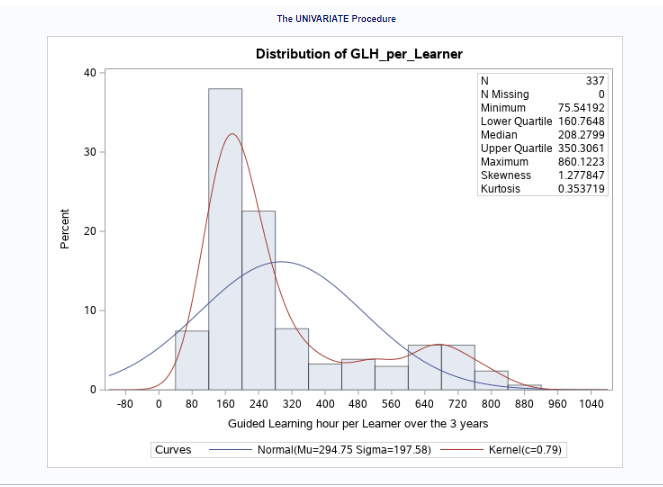
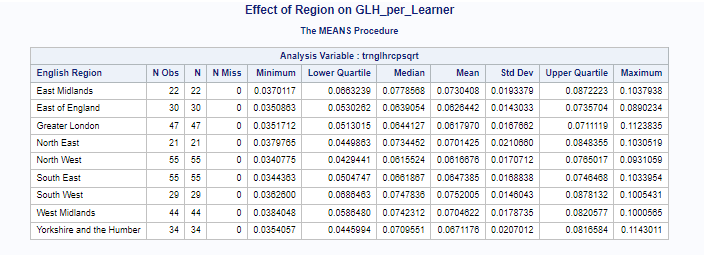


Appendix A

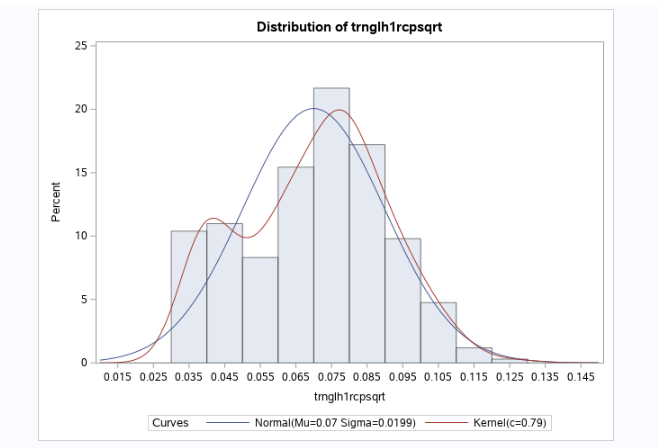


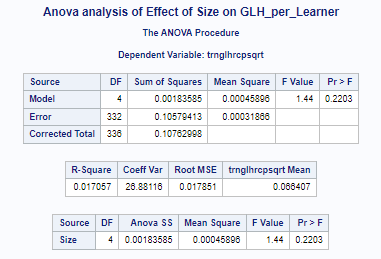
Test for Normality

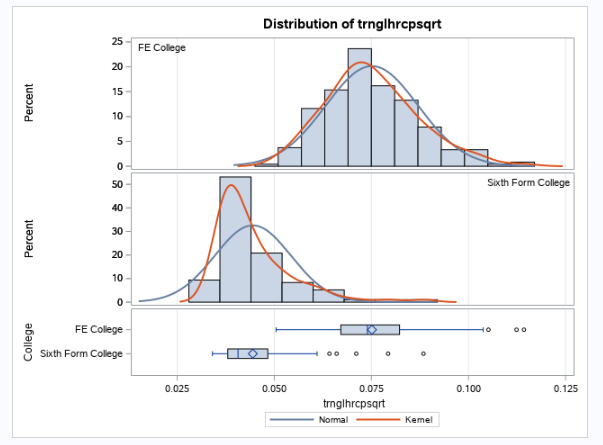


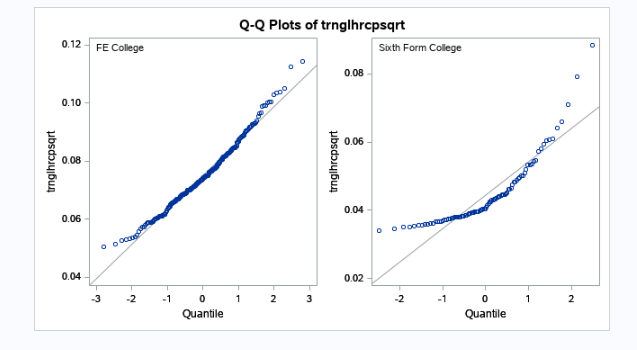


Normal Plot of Transformed data

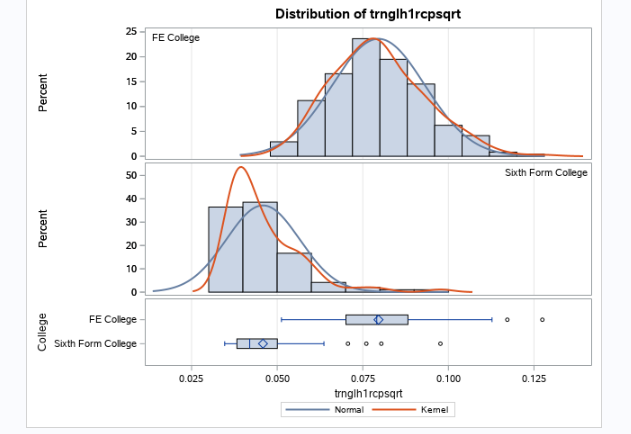


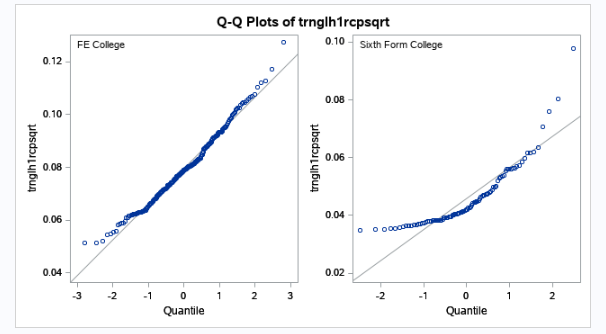


 T-test showing effect of College type on glhperLearner in all 3 years

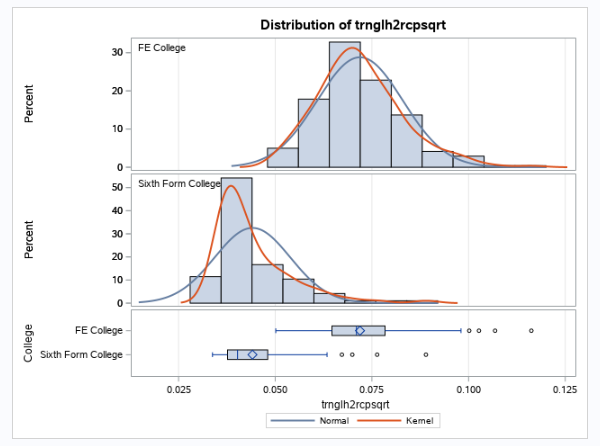


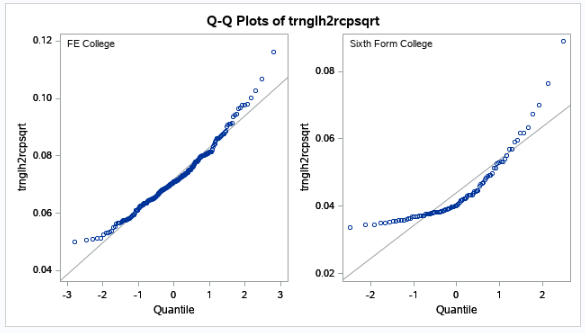
T-test showing effect of College type on glhperLearner in all 3 years



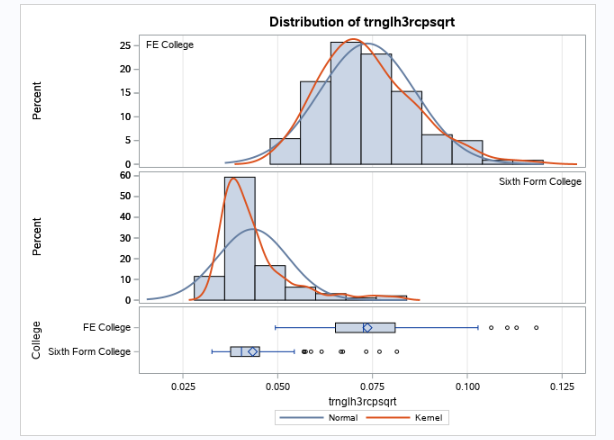
Ttest showing effect of Year 1 on learning hours per learner by College type '

Ttest showing effect of Year 1 on learning hours per learner by College type

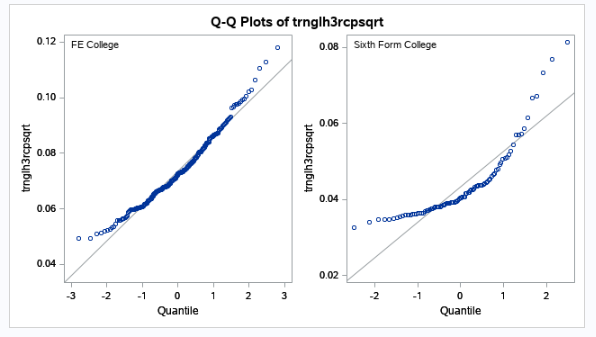




Ttest showing effect of Year 2 on learning hours per learner by College type



Ttest showing effect of Year 3 on learning hours per learner by College type '



Ttest showing effect of Year 3 on learning hours per learner by College type '