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

MSc. Data Analytics

Folashikemi Maryam Asani Olaniyan -P2586104

# SUMMARY

A report of Guided Learning Hours in Further Education and Sixth Form Colleges derived from CSV files containing data about the Guided Learning Hours, Region, Institution and Year.

The report shows a significant effect between Guided Learning Hours, Region, Institution and Year, using a one-way ANOVA model with Kruskal-Wallis Test, Wilcoxon test and Median Test. Using two-way ANOVA there was no significant effect between Region\*Year and Guided Learning Hours Per Learner.

# LIST OF fIGURES

[Figure 1 Analysis flow chart 5](#_Toc38294057)

[Figure 2 Import FORM 6](https://demontfortuniversity-my.sharepoint.com/personal/p2586104_my365_dmu_ac_uk/Documents/P2586104/Documents/IMAT5168%20Analytical%20Programming%20Mark/Assessment/A%20Report%20on%20Guided%20Hours%20in%20Further%20Education%20and%20Sixth%20Form%20Colleges.docx#_Toc38294058)

[Figure 3 Import FE 6](#_Toc38294059)

[Figure 4 Merge Data 6](#_Toc38294060)

[Figure 5 Skewed Distribution 7](#_Toc38294061)

[Figure 6 Numerical test for Transform Variable 8](#_Toc38294062)

[Figure 7 Distribution and QQ-plot of transformed variable 8](https://demontfortuniversity-my.sharepoint.com/personal/p2586104_my365_dmu_ac_uk/Documents/P2586104/Documents/IMAT5168%20Analytical%20Programming%20Mark/Assessment/A%20Report%20on%20Guided%20Hours%20in%20Further%20Education%20and%20Sixth%20Form%20Colleges.docx#_Toc38294063)

[Figure 8 Wilcoxon Test 8](#_Toc38294064)

[Figure 9 Median test 9](#_Toc38294065)

[Figure 10 Kruskal-Wallis Test Region 9](https://demontfortuniversity-my.sharepoint.com/personal/p2586104_my365_dmu_ac_uk/Documents/P2586104/Documents/IMAT5168%20Analytical%20Programming%20Mark/Assessment/A%20Report%20on%20Guided%20Hours%20in%20Further%20Education%20and%20Sixth%20Form%20Colleges.docx#_Toc38294066)

[Figure 11 F-test Region 9](#_Toc38294067)

[Figure 12 Kruskal-Wallis Test Size 9](https://demontfortuniversity-my.sharepoint.com/personal/p2586104_my365_dmu_ac_uk/Documents/P2586104/Documents/IMAT5168%20Analytical%20Programming%20Mark/Assessment/A%20Report%20on%20Guided%20Hours%20in%20Further%20Education%20and%20Sixth%20Form%20Colleges.docx#_Toc38294068)

[Figure 13 F-Test Size 9](#_Toc38294069)

[Figure 14 Kruskal-Wallis Test Year 9](https://demontfortuniversity-my.sharepoint.com/personal/p2586104_my365_dmu_ac_uk/Documents/P2586104/Documents/IMAT5168%20Analytical%20Programming%20Mark/Assessment/A%20Report%20on%20Guided%20Hours%20in%20Further%20Education%20and%20Sixth%20Form%20Colleges.docx#_Toc38294070)

[Figure 15 F-test Year 9](#_Toc38294071)

Contents

[Summary 2](#_Toc38294083)

[List Of Figures 3](#_Toc38294084)

[Understanding The Data Set 5](#_Toc38294085)

[Format 5](#_Toc38294086)

[Analysis Flow Chart 5](#_Toc38294087)

[Methodology 6](#_Toc38294088)

[Importing And Cleaning The Data 6](#_Toc38294089)

[Exploratory Analysis 7](#_Toc38294090)

[Statistical Models 7](#_Toc38294091)

[Conclusions 11](#_Toc38294092)

[References 12](#_Toc38294093)

# Understanding The Data Set

## Format

The data set for this report contains two data source files in CSV format as IMAT5168-FE and IMAT5168-FORM.

Each file contains fields for Institution Type, Region, Guided Learning Hours and Learners for Year 1 till Year 3.

The Field “Guided Learning Hours” was used in creating a new field Guided Learning Hours Size by mapping the values to each size. The fields “Guided Learning Hours” and “Learners” for each year was used to pivot the tables and create a new variable “Year” and by dividing the field another variable “GLH per Learners” was created.

## Analysis flow chart

A picture containing screenshot, sign

Description automatically generated

Figure 1 Analysis flow chart

# Methodology

## Importing and Cleaning The Data

The CSV files for this analysis were imported into SAS using a **filename** statement to define the path location, an **infile** statement was used to specify the files which were IMAT5168-FE.csv and IMAT5168-6FORM.csv after which the field lengths were specified using a **length** statement to ensure the row spaces were long enough for the observation names. The input statement was used to define the field names and labels were created using the **label** statement.

A separate data steps were created for each file and the output was saved to SAS work library as FE and FORM respectively.

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

Figure 2. Import FORM

Figure 2 Import FORM

Figure 3 Import FE

merge and Validate Data

Using SQL Query to merge the two data sets by creating a table named FE\_FORM, then validate the data using **Proc Means** and **Proc Freq** statement.

A screenshot of a cell phone

Description automatically generated

Figure 4 Merge Data

**Proc Means** is used on continuous variables while **Proc Freq** was used on categorical variables to understand the distribution of the variable and identify the missing values.

Delete missing rows

After merging and validating the data, the empty rows were deleted because they contained the aggregate values for each variable per region. Total number of observations became **1038**.

Create new variables

A new variable Year was created by transposing the GLH and Learners field, likewise, GLH size was created by mapping the specified GLH sizes given in the report task to a range of values using IF statements. GLH per learners which is the Guided Learning Hours for each Learner was created by dividing the Guided Learning Hours from the Learners.

## Exploratory Analysis

* Proc Contents was used in viewing the details of the variable attributes detailing the type and size as well as the data quality control (QC).
* Proc Means (5-Number Summary) for continuous variables; GHL per learner shows The min, 25th percentile values are closer to the median than the 75th percentile and maximum values. Which means the majority of the data is distributed to the left. Similarly, the average of the min and max and the average of the 25th and 75th percentile are not close to the median. This suggests that the shape of the distribution is skewed (CAST, 2020).
* Proc Freq for categorical variables; Institution Type, Region, Year and Size shows,

Institution Type show FE college has the highest frequency at 71.68%.

Region shows North West, South East, Greater London as the highest with 16.76%, 16.67% and 13.58% respectively.

Year 1 has the highest frequency (33.91%) but there is not a significant difference in the percentage values.

GLH Size show Medium has the highest frequency at 43.16%.

* Proc Tabulate was used to summarise the categorical variables against continuous variables and its output returns proc means values of the number of observations, nmiss, min, q1, median, q3, max and std.

## Statistical Models

Test For Normality

**Proc Univariate** was used to examine the distribution of GLH per learner by including the **normaltest and plot normal** statements, a plot of a histogram, box-plot and QQ-plot is returned in the output.

Using the Test for Normality table as a numerical method, we will adopt the Shapiro-Wilk Test because the number of observations is less than 2000. According to significance level 5% (α=0.05) we will reject the null hypothesis that suggests the distribution is normally distributed because the p-value (<0.0001)is less than 0.05 (Univariate Analysis and Normality Test, 2002).

Graphically, the histogram is positively skewed to the left, with most of the data between 150 and 210.

A screenshot of a social media post

Description automatically generated

Figure 5 Skewed Distribution

The QQ-plot shows the data deviates from the line of best fit.

Transform the variable

Since the distribution of the variable is not normal, an attempt was made to transform the variable using Log (<var>+1) and another **proc univariate** was conducted to view the distribution.

Using the Test for Normality table as a numerical method, we will adopt the Shapiro-Wilk Test because the number of observations is less than 2000. According to significance level 5% (α=0.05) we will reject the null hypothesis that suggests the distribution is normally distributed because the p-value (<0.0001)is less than 0.05 (Univariate Analysis and Normality Test, 2002).

Graphically, the histogram is positively skewed to the left, with most of the data between 150 and 210. The QQ-plot shows the data deviates from the line of best fit.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Figure 6 Numerical test for Transform Variable

Figure 7 Distribution and QQ-plot of transformed variable

one-way ANOVA using proc npar1way

A non-parametric test (Proc Npar1way) was adopted for analysing the effects of the independent variables against the dependent variable GLH Per Learner. This is due to the inability to make the distribution normal even after attempting to transform the variable which means the assumption of parametric test has been violated.

The Wilcoxon and Median option were added to the Npar1way statement because they test for difference in location and Institution has two different locations (FE College and Sixth Form College).

The Result of the **Wilcoxon test** shows two-sample test statistic equals 253731.0000 which is the same as the sum of scores for the smaller sample (Sixth Form College). The one-sided p-value is <.0001, which shows the GLH per learner for Sixth Form College is significantly more than for the FE College (SAS Institute Inc.,2015.)

A screenshot of a social media post

Description automatically generated

Figure 8 Wilcoxon Test

The result of Median two-sample shows the statistic equals 282.0000, with a one-sided p-value of <.0001. this shows that the GLH Per Learner for the Sixth Form College is significantly more than for the FE College.

A screenshot of a cell phone

Description automatically generated

Figure 9 Median test

One-Sample test are output for other variables using Proc Npar1way.

For **Region**, the p-value for the F-Test is <.0001, which indicates that Region accounts for a significant portion of the variability in GLH Per Learner.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 10 Kruskal-Wallis Test Region

Figure 11 F-test Region

The Kruskal-Wallis Test displays the One-way ANOVA statistic with p-value =<.0001 so we reject the null hypothesis that there is a difference in locations for GLH per learner among different regions.

For **GLH SIZE**, the p-value for the F-Test is <.0001, which indicates that Size accounts for a significant portion of the variability in GLH Per Learner.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 12 Kruskal-Wallis Test Size

Figure 13 F-Test SizeA screenshot of a cell phone

Description automatically generated

The Kruskal-Wallis Test displays the One-way ANOVA statistic with p-value =<.0001 so we reject the null hypothesis that there is a difference in locations for GLH per learner among different sizes.

A screenshot of a cell phone

Description automatically generatedFor **Year**, the p-value for the F-Test is <.0001, which indicates that Year accounts for a significant portion of the variability in GLH Per Learner.

A screenshot of a cell phone

Description automatically generated

Figure 15 F-test Year

Figure 14 Kruskal-Wallis Test Year

The Kruskal-Wallis Test displays the One-way ANOVA statistic with p-value =<.0001 so we reject the null hypothesis that there is a difference in locations for GLH per learner among different years (Kruskal-Wallis Test in SAS, 2020).

Two-way anova using proc glm for unbalanced anova

First, we test for an interaction between two class variables Region and Year.

The data was sorted by Institution type and Institution Size. The estimation factor used to test for proc GLM is the Type III error which is typically unequal when the data is unbalanced.

Results;

**For FE college**, the following explains the effect of interaction on the dependent variable.

* **Large,** Interaction of Region\*year is insignificant (p=0.9867). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because Region has a significant effect(p=<.0001) while Year doesn’t (p=0.1471).
* **Large-medium,** Interaction of Region\*year is insignificant (p= 0.9997). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because Region has a significant effect(p= 0.0001) likewise year(p= 0.0002).
* **Medium,** Interaction of Region\*year is insignificant (p= 0.9983). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because Region has a significant effect(p= <.0001) likewise year(p= <.0001).
* **Small-medium,** Interaction of Region\*year is insignificant (p= 0.9847). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because year (p= 0.0055) has a significant effect but Region doesn’t (p= 0.1008).
* **Small,** Interaction of Region\*year is insignificant (p= 0.5988). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are not valid because year (p= 0.4929) has a significant effect but Region doesn’t (p= 0.6928).

**For Sixth Form College**, the following explains the effect of interaction on the dependent variable.

* **Large-medium,** Interaction of Region\*year is insignificant (p= 0.9998). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are not valid because Region has a significant effect(p= 0.4588) likewise year(p= 0.9801).
* **Medium,** Interaction of Region\*year is insignificant (p= 0.9999). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because Region has a significant effect(p= 0.0008) year doesn’t (p= 0.3808).
* **Small-medium,** Interaction of Region\*year is insignificant (p 0.9971). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because the year has no significant(p= 0.1344) effect but Region does (p<.0001).
* **Small,** Interaction of Region\*year is insignificant (p= 0.4756). This indicates that the level of the region does not depend on the year and vice versa. The test for individual effects are valid because Region (p= 0.6368) has a significant effect but the year doesn’t (p= 0.6966).

RESIDUAL

The result of the residual in the histogram is normally distributed with a few outliers, although the Shapiro-Wilk test is less than 0.05. we assumed the value is due to the presence of extreme observation from the unbalance data.

# Conclusions

The report highlights the effect of Institution Type, region, Size and Year of variable GLH per learner using one-way ANOVA in Kruskal-Wallis test, Wilcoxon tests and median test. It shows that all the indepent variables have a significant effect on GLH Per Learner.

The Two-way Anova using proc GLM shows the effect of Region and Year by institution type and size on GLH Per Learner, the results show model for interaction between Region and Year are not significant all sizes and Institution Type.

However, the individual test on Region and Year shows that some regions and year that is a significant effect.

# References

Cast.idems.international. 2020. *CAST*. [online] Available at: <https://cast.idems.international/> [Accessed 20 April 2020].

Support.sas.com.2020.[online]Availableat: <https://support.sas.com/documentation/onlinedoc/stat/141/npar1way.pdf> [Accessed 20 April 2020].

Statistical Methods. 2020. Kruskal-Wallis Test In SAS. [online] Available at: <https://stat-methods.com/home/kruskal-wallis-sas/> [Accessed 16 April 2020].