A black background with blue and black border

Description automatically generated

**Ho Chi Minh City, March 2023**

**Lecturers:**  **Assoc. Prof. Nguyen Dinh Thuan**

**TA. Nguyen Minh Nhut**

**Students:**

**Bui Dinh Trieu - 21521576**

**Nguyen Chi Kha - 21522179**

**Nguyễn Minh Duy - 21522005**

**LAB02 REPORT**

**IS403: DATA ANALYSIS IN BUSINESS**

**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

# TABLE OF CONTENTS

**Contents**

[TABLE OF CONTENTS 1](#_Toc2052397318)

[TEACHER’S COMMENTS 2](#_Toc1320185703)

[WORK DISTRIBUTION 3](#_Toc1282860426)

[A. LEVENE, ANOVA 4](#_Toc1019203098)

[B. ENERGY DRINK SURVEY: 7](#_Toc75178719)

[1. Using MS Excel, R, and Python Programming Language: 7](#_Toc2063169302)

[1.1. Analyzing by using Excel: 7](#_Toc201317619)

[1.2. Analyzing by using R: 9](#_Toc1142839887)

[1.3. Analyzing by using Python: 11](#_Toc973706249)

[2. Conclusion: 13](#_Toc1248478296)

[C. INSURANCE SURVEY: 14](#_Toc899744725)

[1. Using MS Excel, R, and Python Programming Language: 15](#_Toc238179585)

[1.1. Analyzing by using Excel: 15](#_Toc754440933)

[Levene 15](#_Toc682448559)

[ANOVA 15](#_Toc865024600)

[Tukey 17](#_Toc2136769779)

[1.2. Analyzing by using R: 18](#_Toc477427414)

[1.3. Analyzing by using Python: 19](#_Toc667576521)

[2. Conclusion: 21](#_Toc958182242)

[D. VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE: 22](#_Toc1158865000)

[1. Using R, and Python Programming Language in Example 23](#_Toc828895983)

[1.1. Analyzing by using R: 23](#_Toc986927066)

[1.2. Analyzing by using Python: 25](#_Toc416810245)

[2. Conclusion: 27](#_Toc2112296117)

[E. REFERENCES: 29](#_Toc640293135)

# TEACHER’S COMMENTS

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

# WORK DISTRIBUTION

|  |  |  |  |
| --- | --- | --- | --- |
| **Members**  **Works** | **Nguyen Minh Duy (Leader)** | **Bui Dinh Trieu** | **Nguyen Chi Kha** |
| **Problem statement** |  |  |  |
| **Build the report template** |  |  |  |
| **Do all exercise with Excel** |  |  |  |
| **Do all exercise with R** |  |  |  |
| **Do all exercise with Python** |  |  |  |
| **Summarize and edit reports** |  |  |  |
| **Completion (%)** | 100% | 100% | 100% |

## LEVENE, ANOVA

**Lesson**

**a)What is Levene Test for Equality of Variances? Explanation and example.**

**b)What are post hoc comparison tests used for in ANOVA? Explanation and example.**

1. **What is Levene Test for Equality of Variances? Explanation and example.**

The Levene's Test for Equality of Variances is a statistical test used to assess whether the variances of two or more groups or populations are equal or not. It is often used as a prerequisite for other statistical tests, such as the Student's t-test or the analysis of variance (ANOVA), which assume homogeneity of variance (equal variances) among the groups.

The Levene's test is particularly useful when the data violates the assumption of normality, as it is less sensitive to departures from normality than other tests for homogeneity of variance.

**Explanation:**

The Levene's test calculates the absolute difference between each data point and the median of its group, and then performs an analysis of variance on these deviations. If the variances are equal, the deviations should be similar across groups.

The null hypothesis (H0) for the Levene's test is that the variances of the groups are equal, while the alternative hypothesis (Ha) is that at least one group's variance is different from the others.

If the p-value obtained from the Levene's test is greater than the chosen significance level (e.g., 0.05), it means that there is not enough evidence to reject the null hypothesis, and the assumption of equal variances is met. If the p-value is less than the significance level, the null hypothesis is rejected, and the assumption of equal variances is violated.

**Example:**

Suppose we want to compare the heights of students from three different schools (School A, School B, and School C). We collect data on student heights and perform the Levene's test to determine if the variances of the heights are equal across the three schools.

The Levene's test statistic and its associated p-value can be calculated using statistical software or manually. If the p-value is greater than the chosen significance level (e.g., 0.05), we can conclude that the assumption of equal variances is met, and we can proceed with other statistical tests, such as ANOVA, to compare the mean heights across the three schools.

1. **What are post hoc comparison tests used for in ANOVA? Explanation and example.**

Post hoc comparison tests, also known as post hoc multiple comparison tests or post hoc analysis, are used in analysis of variance (ANOVA) to determine which specific groups or means differ from each other when the overall ANOVA test is statistically significant.

**Explanation:**

The ANOVA test itself only tells us whether there are any statistically significant differences among the means of the groups or populations being compared. However, it does not identify which specific groups or means differ from each other. Post hoc tests are used to make pairwise comparisons between the group means and determine where the differences lie.

There are several types of post hoc tests, such as Tukey's Honest Significant Difference (HSD), Bonferroni, Scheffé, and Duncan's Multiple Range Test, among others. The choice of the appropriate post hoc test depends on various factors, such as the number of groups being compared, the assumption of equal or unequal variances, and the desired level of control over the familywise error rate (the probability of making at least one Type I error when performing multiple comparisons).

**Example:**

Suppose we conduct a one-way ANOVA to compare the mean scores on a standardized test among four different teaching methods (A, B, C, and D). The ANOVA results show a statistically significant difference among the group means (p-value < 0.05).

To determine which specific teaching methods differ from each other, we can conduct a post hoc test, such as Tukey's HSD. The post hoc test will perform pairwise comparisons between the means of teaching methods A and B, A and C, A and D, B and C, B and D, and C and D.

The results of the post hoc test might indicate that the mean score for teaching method A is significantly different from teaching methods C and D, but not from method B. Similarly, the mean score for teaching method B might be significantly different from teaching method D, but not from methods A and C.

Based on these results, we can conclude which teaching methods are more or less effective in terms of student performance on the standardized test.

## B. ENERGY DRINK SURVEY:

**Lesson. Using MS Excel, R language and Python language to perform Chi Square test on the independence of two categorical variables with the data file: Energy Drink Survey**

### Using MS Excel, R, and Python Programming Language:

### Analyzing by using Excel:

A screenshot of a spreadsheet

Description automatically generated

The result on the figure, the grand total of people choose brand 1 is 34, brand 2 is 23, brand 3 is 43. The grand total of females use 3 brands is 37 and males is 63. The grand total of men and women in this case is 100 people.

*Calculating:*

Calculate Expected Frequency

Expected Frequency = (grand total column\* grand total row) / total number of observations

= (Grand Total of Brand X \* Grand Total of Y) / Total number; with X is the value from 1 - 3 and Y is the male or female

*A screenshot of a computer

Description automatically generated For example: at (Brand 1, Female), 37 \* 34 / 100 12.58*

Calculate Chi-square value

A mathematical equation with a number and a line

Description automatically generated with medium confidence

Χ2: the chi-square test statistic

fo : observed frequencies.

*fe* : expected frequencies.

*For example: at (Brand 1, Female), we have* fo *= 9, fe = 12.58*

* *(9 – 12.58) ^ 2 / 12.58 1.02*

*A screenshot of a spreadsheet

Description automatically generated*

* **Chi-square = SUM(F24:H25) = 6.49242508**

Calculate df, p-value, Critical-value

X2 = 6.492425079

df = (row – 1) \* (column – 1) = (2 – 1) \* (3 – 1) = 2

p-value = CHISQ.DIST.RT(x,deg\_freedom) = CHISQ.DIST.RT(X2,2) = 0.038921342

Significance level (α): 5% = 0.05

A table with text on it

Description automatically generated  
Critical-value = CHISQ.INV.RT(probability,deg\_freedom) = CHISQ.INV.RT(0.05,2) 5.991

All calculations in this figure:

p-value < α (0.039 < 0.05)

or X^2 > Critical-value (6.49 > 5.99)

* reject H0

A screenshot of a spreadsheet

Description automatically generated

* Therefore, variable Grand and Brand Preference are not independent variables.

### Analyzing by using R:

A screenshot of a computer

Description automatically generatedRead and check data:

**ChiA white background with black text

Description automatically generatedA black text on a white background

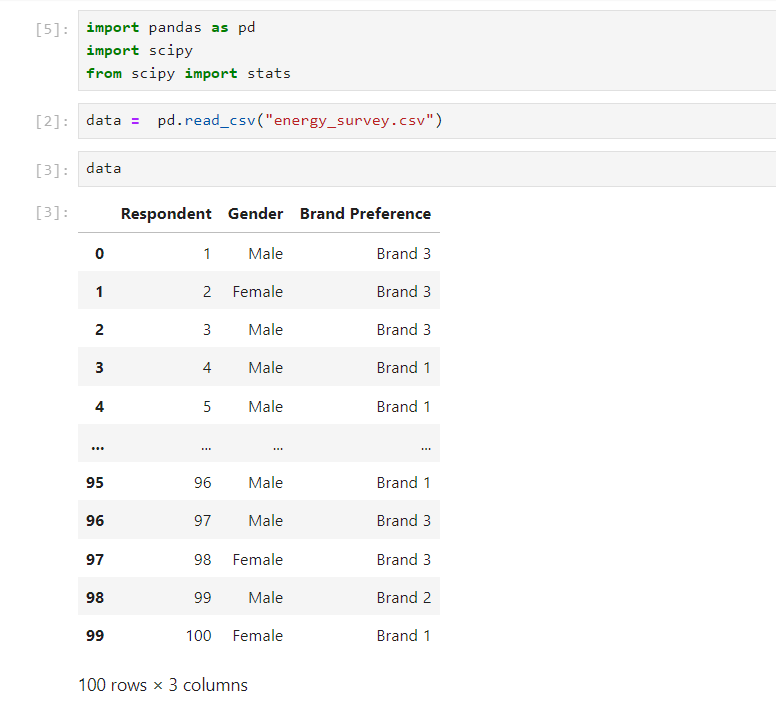
Description automatically generated-square test:**

* The Chi-square test is used to examine the relationship between two categorical variables (in this case, gender and product brand).
* The X-squared value = 6.4924 indicates a difference between the observed distribution and the expected theoretical distribution when the two variables are independent.
* Degrees of freedom (df) = 2, meaning there are 2 degrees of freedom in the Chi-square distribution.
* The p-value = 0.03892 < 0.05, which implies that the test result is statistically significant at the 95% confidence level.

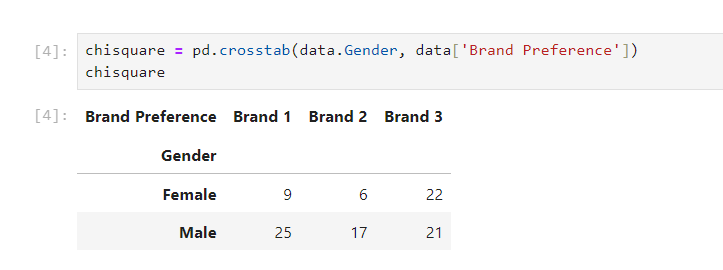
Based on the p-value being less than 0.05, we have enough evidence to reject the null hypothesis that gender and product brand are independent. This means there is a relationship between gender and the choice of product brand by consumers.

### Analyzing by using Python:

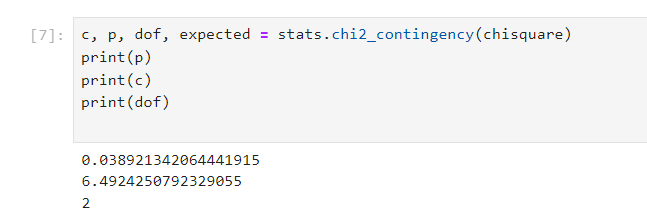
Import libraries and read data from data file ‘*Energy Drink Survey*’.



Chi-Square Table



Print Chi-Square Result



The p-value is approximately 0.0389, which is smaller than the typical statistical significance threshold (e.g., 0.05). We reject H0. This suggests that there is a statistically significant difference between the Grand and the Brand Preference table.

### Conclusion:

Based on the survey data on the preference for energy drink brands and respondents' gender, following the Chi-Square test, we can draw the following practical conclusions:

* There is a relationship between gender and preference for energy drink brands: The Chi-Square test result shows a p-value smaller than the significance level of 0.05, indicating that gender and preference for energy drink brands are not entirely independent. It can be observed that the preference for energy drink brands is related to the gender of consumers.
* Differences in preference for energy drink brands between males and females: By examining the observed frequencies and expected frequencies in the contingency table, we can detect differences in brand preference patterns between males and females. For example, females tend to prefer Brand 3 more than males, while males prefer Brand 1 more than females.
* Marketing strategies can be adjusted according to gender: With information on the relationship between gender and brand preference, marketers can adjust marketing strategies and advertising for energy drink products to align with the preferences of each gender group.
* Further research on the causes of differences: The Chi-Square test results only indicate a relationship between the two variables but do not explain the causes of that difference. Further studies on the characteristics, habits, and perceptions of males and females regarding energy drinks can help better understand the reasons behind the differences in brand preference.

**In summary, the Chi-Square test results show a relationship between gender and preference for energy drink brands, which can impact the marketing strategies and market research of companies in this industry.**

## INSURANCE SURVEY:

**Lesson. Using MS Excel, R language and Python language to perform ANOVA with data file (including Levene, ANOVA, Tukey Test): Insurance survey.**

### Using MS Excel, R, and Python Programming Language:

### Analyzing by using Excel:

**Levene**

P-value < 0.05 => reject H0 => able to do Anova test

**ANOVA**

Calculating by using ANOVA Single Factor function to analyze data

Sum Count comlumn(24) and Sum column(89)

A screenshot of a spreadsheet

Description automatically generated

Calculate average of all by Sum divide Count (89 / 24 = 3.708333333

A screenshot of a spreadsheet

Description automatically generated

Calculate SS of each group and sum of them is SSG (7.878968254)

A screenshot of a spreadsheet

Description automatically generated

Calculate SSW (21.07936508)

A screenshot of a calculator

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự độngANOVA:

Between Groups (SSG) and (Within Groups) SSW

Calculate df

* + k=3; N=24
  + SSG: df = k - 1 = 3 - 1 = 2
  + SSW: df = N - k = 24 - 3 = 21
  + Total = 21 + 2 = 23

Calculate MS

* MSG = SSG / (k-1) = 7.878968254 / 2 = 3.939484127
* MSW = SSW / (N-k) = 21.07936508 / 21 = 1.003779289

Calculate F

F = MSG / MSW = 3.924651732

Calculate F crit to compare with F ,

F.INV.RT(probability, df\_between, df\_within)

The significance level (alpha) = 5% = 0.05

F crit= F.INV.RT(0.05, 2, 21) = 3.466800112

p-value = F.DIST.RT(F, 2, 21) = 0.035635398

* + F > F crit (3.924651732 > 3.466800112)

or

* + p-value = 0.035 < 0.05
    - Reject H0
    - Two categorical variables are dependent

**Tukey**

Calculate Q-statistics

k = 3, n - k = 21(df), α = 0.05

Looking up the Turkey distribution table, we get Q statistic = 3,565

Turkey value:

A yellow background with black text

Description automatically generated

nmin = 7

T = 3.565 \* √(1.003779289) / 7 = 1.349987134

A screenshot of a spreadsheet

Description automatically generated

Calculate the difference between 2 groups

A math equation with a letter n

Description automatically generated

A white grid with black text

Description automatically generated

Compare the difference between 2 pairs with T

College graduate vs Graduate degree < T => x1 = x2 (accept H0)

College graduate vs Some college < T => x1 = x3 (accept H0)

Graduate degree vs Some college > T => x2 ≠ x3 (reject H0)

### Analyzing by using R:

**Read and check data**

ANOVA TEST

A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated A close-up of a sign

Description automatically generated

A close-up of a white background

Description automatically generated

A computer screen shot of a math program

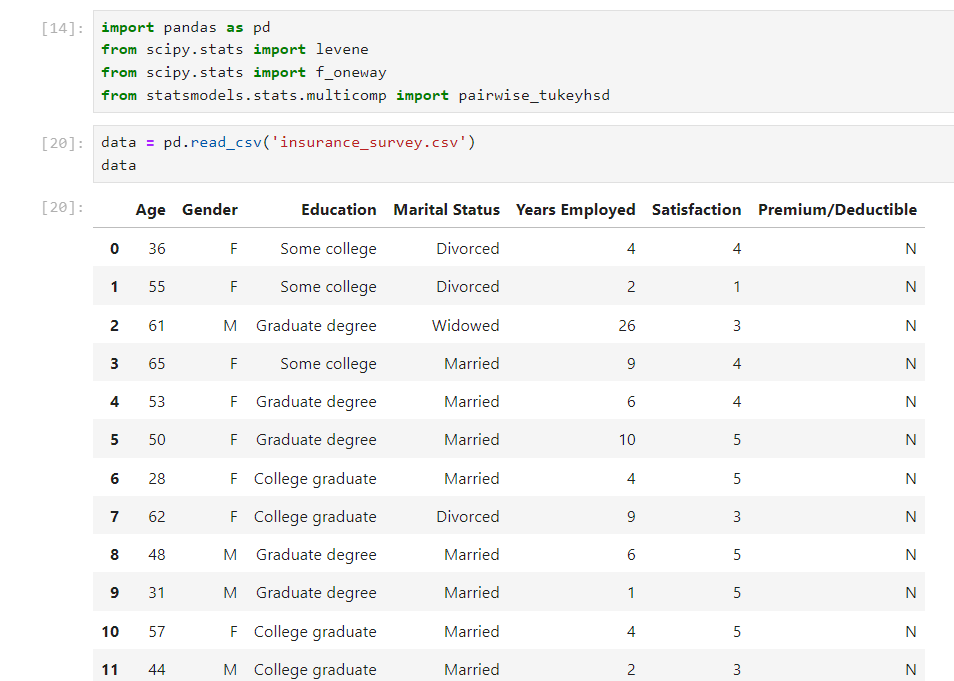
Description automatically generated with medium confidence

1. In Figure 1, the ANOVA results show a statistically significant difference in satisfaction between different groups of customers with different levels of education (p-value = 0.0356 < 0.05).
2. Figure 2 presents the Tukey HSD test results to determine which specific groups differ in satisfaction. The results showed:
3. Customers with bachelor's and master's degrees have a significant difference in satisfaction compared to customers with only partial undergraduate degrees (p-adj = 0.1003252 and 0.0409193 respectively).

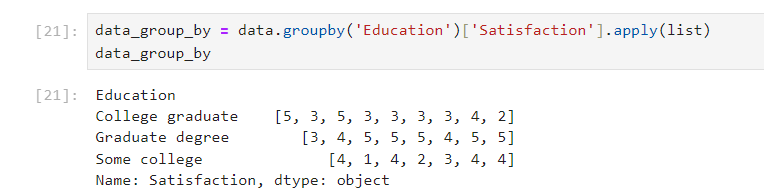
* There was no significant difference in satisfaction between bachelor's and master's degree clients (p-adj = 0.8230559 > 0.05).

### Analyzing by using Python:

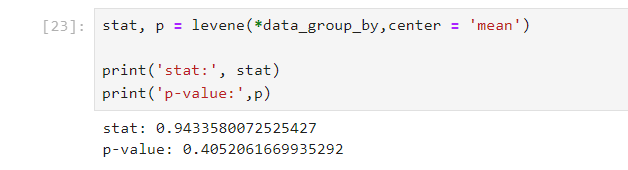
Import libraries and read the data from data file ‘Insurance survey’.



Collects the 'Satisfaction' values for each Education group into lists.



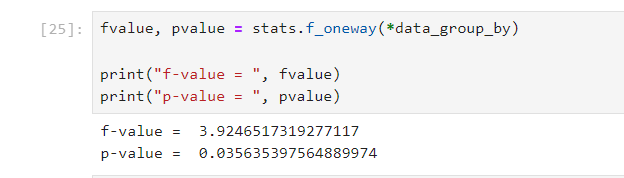
**Do Levene testing.**



The test statistic is 0.943, which indicates a moderate difference in variances between groups.

The p-value is 0.405, which is greater than the typical significance level of 0.05. This suggests that there is insufficient evidence to reject the null hypothesis of equal variances across groups.

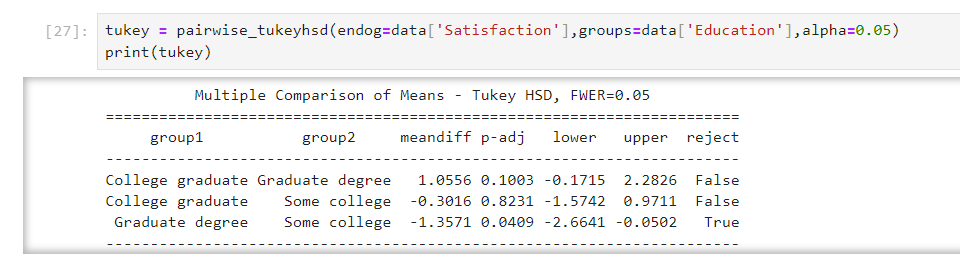
**Do Anova testing.**



p-value 0.036 < 0.05.

This suggests that the p-value is less than the typical significance level of 0.05, indicating that we have evidence to reject the null hypothesis(H0).  
Therefore, we can conclude that there is at least one significant difference among the group means in terms of satisfaction scores based on education level.

**Do Tukey testing.**



The comparison between the 'College graduate' and 'Graduate degree' groups has a p-value of 0.1003, indicating that there is no statistically significant difference in satisfaction scores between these two groups at the 0.05 significance level.

The comparison between the 'College graduate' and 'Some college' groups has a p-value of 0.8231, suggesting that there is no statistically significant difference in satisfaction scores between these two groups at the 0.05 significance level.

The comparison between the 'Graduate degree' and 'Some college' groups has a p-value of 0.0409, which is less than 0.05. Therefore, there is a statistically significant difference in satisfaction scores between these two groups at the 0.05 significance level.

### Conclusion:

* Higher levels of education (undergraduate, postgraduate) seem to be associated with higher levels of satisfaction with insurance service than lower levels of education (partly undergraduate).
* Insurers may consider improving service quality and communication to better meet the needs of customers with lower levels of education.
* For customers with higher education (undergraduate, postgraduate), similar policies and services can be applied because they have similar levels of satisfaction.

**These conclusions can be useful for the insurer in understanding customer needs and satisfaction based on education level, thereby adjusting business and customer service strategies accordingly.**

## VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE:

**Lesson. Using R language and Python language to perform ANOVA with data file about VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 (with optional hypotheses)**

### Using R, and Python Programming Language in Example

### Analyzing by using R:

Read and check data:

DataA screenshot of a computer

Description automatically generated prepropressing

A screenshot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generatedA screenshot of a computer program

Description automatically generated

**Export file csv after prepropressing:**

**A screenshot of a computer program

Description automatically generated**

ANOVA TEST:

A screenshot of a computer

Description automatically generated

The Pr(>F) value in the ANOVA result is very small (< 2e-16) and marked with "\*\*\*", indicating a significant difference in scores ("Diem") between test blocks ("Khoi") at the highest level of significance (0.001).

The F value is very large (6789), which also supports the conclusion that there is a significant difference in scores between the test blocks.

The "Residuals" row indicates there are 1073350 observations (the number of data points after subtracting the degrees of freedom of the "Khoi" factor), and the mean square error (Mean Sq) is 11.

A white screen with black text

Description automatically generated

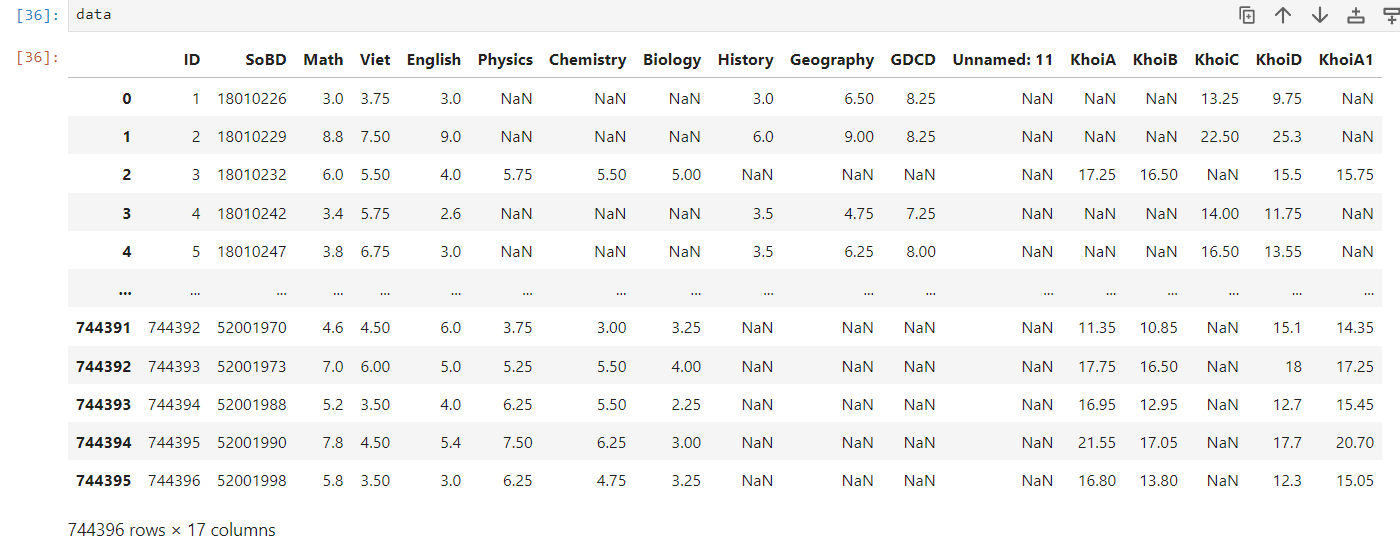
* Block B had an average score of about 0.399 points higher than block A, and this difference was statistically significant (p adj = 0).
* Block C has an average score about 0.901 points higher than block A, and this difference is statistically significant (p adj = 0).
* Block C had an average score of about 0.501 points higher than block B, and this difference was statistically significant (p adj = 0).

The results showed that there was a significant difference in scores between all pairs of exam blocks, with block C having the highest score, followed by block B and finally block A

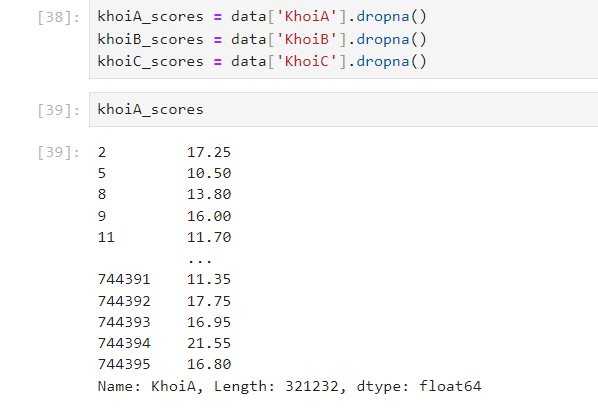
### Analyzing by using Python:

Import data from data file ‘VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018’

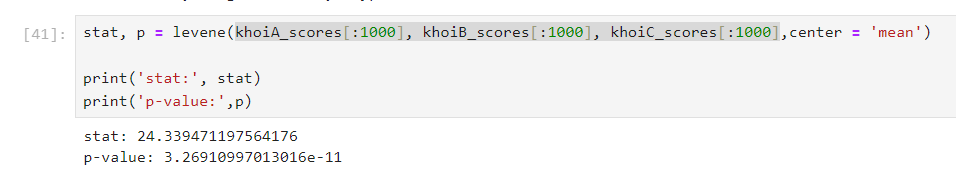




Store 3 columns in variables.

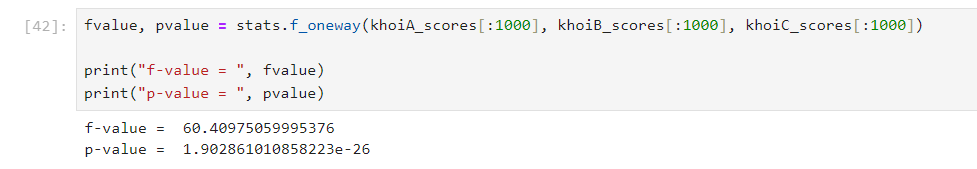


**Do Levene testing.**



Based on the provided p-value of 3.26910997013016×10−113.26910997013016×10−11, which is significantly less than the conventional significance level of 0.050.05, we reject H0. The Levene's test indicates that the variances of the 'KhoiA', 'KhoiB', and 'KhoiC' scores are not equal

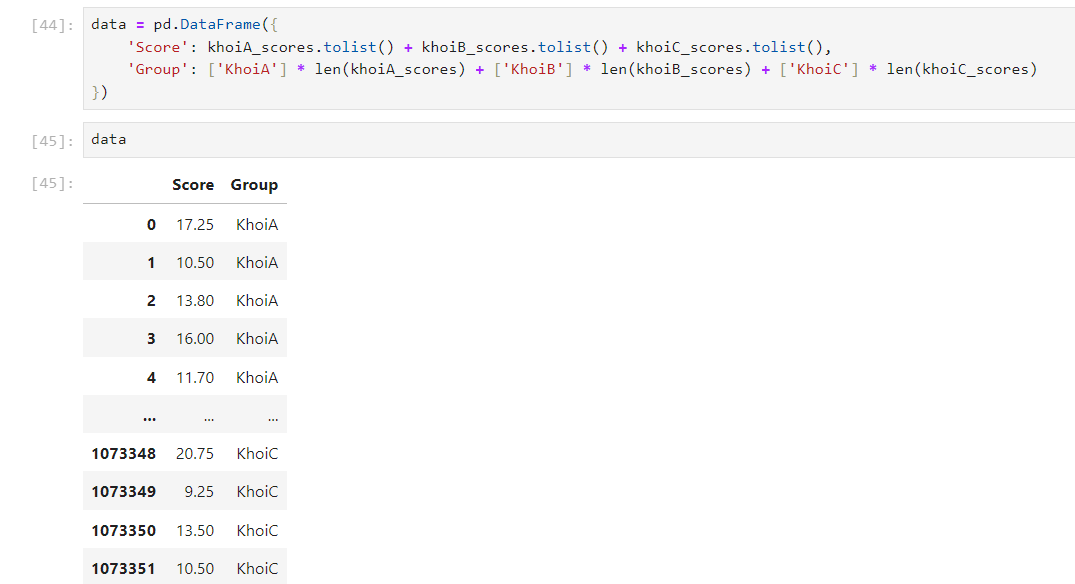
**Do Anova testing.**



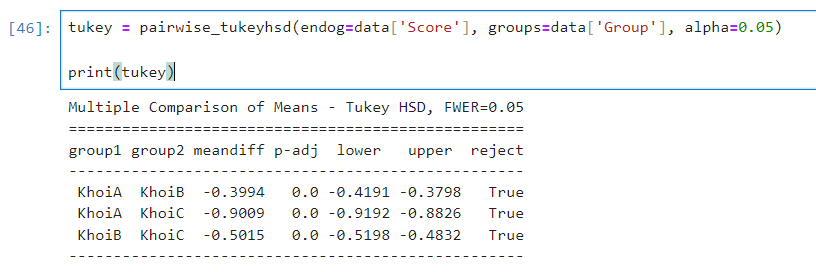
Based on the provided p-value of 1.902861010858223×10−261.902861010858223×10−26, which is significantly less than the conventional significance level of 0.050.05, we reject H0 as well. The ANOVA test indicates that there is a statistically significant difference in mean scores among the groups 'KhoiA', 'KhoiB', and 'KhoiC'.

**Do Tukey testing.**

Create a new data frame consisting of Score and Group column.



Run the Tukey testing command.



For the comparison between 'KhoiA' and 'KhoiC' groups, the adjusted p-value is also 0.0. Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference between the mean scores of 'KhoiA' and 'KhoiC' groups.

Similarly, for the comparison between 'KhoiB' and 'KhoiC' groups, the adjusted p-value is 0.0. Thus, we reject the null hypothesis and conclude that there is a statistically significant difference between the mean scores of 'KhoiB' and 'KhoiC' groups.

### Conclusion:

**From the results of the above analysis, we can see and make some suggestions for improvement in practice as follows:**

1. **Differences in scores between exam grade:**

* This may reflect differences in the quality of training, curricula or educational resources between different regions or schools.
* Measures should be taken to improve the quality of teaching and learning in lower scoring areas/schools, such as additional investment in facilities, teacher training, curriculum updates, etc.

1. **Grade C has the highest score:**

* This could be because C grade has better curriculum, better teachers, or better educational resources than A and B grade.
* It is necessary to study and apply the strengths of the C grade to other grades, such as sharing teaching experiences, exchanging curricula, etc.

1. **The difference between grade B and A is also significant:**

* It is necessary to review and improve the quality of education in grade A to shorten the gap with grade B.

1. **In addition, other factors that may affect the results should be considered, such as:**

* Demographics of students in different areas (income, parental education, etc.)
* Financial resources for education in different regions.
* Educational policies and educational management systems in different regions.

**By considering and addressing these issues, we can contribute to improving the quality of education and reducing achievement gaps between regions, to ensure equitable learning opportunities for all students.**

## REFERENCES:

1. Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (pp. 278-292). Stanford University Press.
2. Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. Statistical Science, 24(3), 343-360.
3. Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Reconsideration and recommendations for the use of transformation to calculate a better estimator of the variance for quantitative variables. PloS One, 13(9), e0202976.
4. Tabachnick, B. G., & Fidell, L. S. (2019). Using Multivariate Statistics (7th ed.). Pearson Education.
5. Abdi, H., & Williams, L. J. (2010). Newman-Keuls test and Tukey test. Encyclopedia of Research Design, 1-11.
6. Ruxton, G. D., & Beauchamp, G. (2008). Some suggestions about appropriate use of the Kruskal-Wallis test. Animal Behaviour, 76(3), 1083-1087.
7. Shingala, M. C., & Rajyaguru, A. (2015). Comparison of post hoc tests for unequal variance. International Journal of New Technologies in Science and Engineering, 2(5), 22-33.