Logo, company name

Description automatically generatedShape, square

Description automatically generated

**DATA ANALYSIS IN BUSINESS**

**CONTENT: REPORT LAB 2 (QUESTION 2F AND 2G)**

Ho Chi Minh City, March 2022

**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEM**

———h&g———

**CLASS: IS211.M11.HTCL**

**TEAM: 3**

**STUDENTS:**

TRẦN MẪN QUÂN 19520873

ĐẶNG NGUYỄN PHƯỚC AN 19521171

TRẦN THỊ NGỌC AN 19521189

**TABLE OF CONTENT**

[WORK ASSIGNMENT TABLE 3](#_Toc100179374)

[I. EXERCISE 2F 4](#_Toc100179375)

[1. R 4](#_Toc100179376)

[2. PYTHON 8](#_Toc100179377)

[3. Excel 11](#_Toc100179378)

[II. EXERCISE 2G 16](#_Toc100179379)

[1. R 16](#_Toc100179380)

[2. Python 19](#_Toc100179381)

[3. MS EXCEL 22](#_Toc100179382)

[REFERENCES 24](#_Toc100179383)

WORK ASSIGNMENT TABLE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Student’name** | **Student’s ID** | **Assignment** | **Evaluation(%)** |
| 1 | Trần Mẫn Quân | 19520873 | - Writing conclusions | 100% |
| 2 | Đặng Nguyễn Phước An | 19521171 | - Writing conclusions | 100% |
| 3 | Trần Thị Ngọc An | 19521189 | - Writing conclusions | 100% |

1. EXERCISE 2F

**Problem statement:** This is a statistic posted publicly by the General Statistics Office of Vietnam (Link: [https://www.gso.gov.vn/px-web-2/?pxid=V1015&theme=Giáo%20dục](https://www.gso.gov.vn/px-web-2/?pxid=V1015&theme=Gi%C3%A1o%20d%E1%BB%A5c)). Dataset includes the high school graduation rate in Vietnam (2016-2017), which is collected from 63 provinces and cities, divided into 3 main regions: North, Central, and South. The analysis aims to find out whether the mean rate of each region different or not from others.

Hypothesis for ANOVA test:

*H0: The mean rate of each region is not diffent from others (μ1 = μ2 = μ3)*

*H1: at least one region has mean rate different from the others*

* 1. R

**Step 1:** Convert **tyletotnghiep.xlsx** file to **tyletotnghiep.csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step2:**

* Import file \*csv in step 1 into RStudio Application

rate\_lab <-read.csv(' tyletotnghiep.csv', header = TRUE)



* Attach this file

attach(rate\_lab)



**Step 3:** Then, we have to install and use library **car.**

install.packages("car")



* In tab Packages, select **Install**

Graphical user interface, application, table

Description automatically generated

Enter the library name, then click **Install**

**Graphical user interface, text, application, email

Description automatically generated**

* After the installation is complete, we use the **require()** statement to use the library

require(car)





**Step 4:**

* After using the library successfully, we start calculate the Levene test throung the command

leveneTest(frequency,region,center="mean")

Text

Description automatically generated with low confidence

Since Pr (>F) > 0.05 (0.3985 > 0.05) 🡪 Reject H1

* All variances are equal and we can test Anova
* Look up: use function **qf** to calculate

qf(p=0.95,2,60)

A picture containing text

Description automatically generated

F value < F-Critical (0.9344 < 3.1504), Reject H1, all variances are equal and we can test Anova

aov(frequency ~region, data = rate\_lab)

Text

Description automatically generated

Using α = 0.05 ,Pr(>F), p value = 0.0356 < 0.05 🡪 Reject H0, at least one mean is different from the others

**Step 5: Calculate the Turkey**

**Text, letter

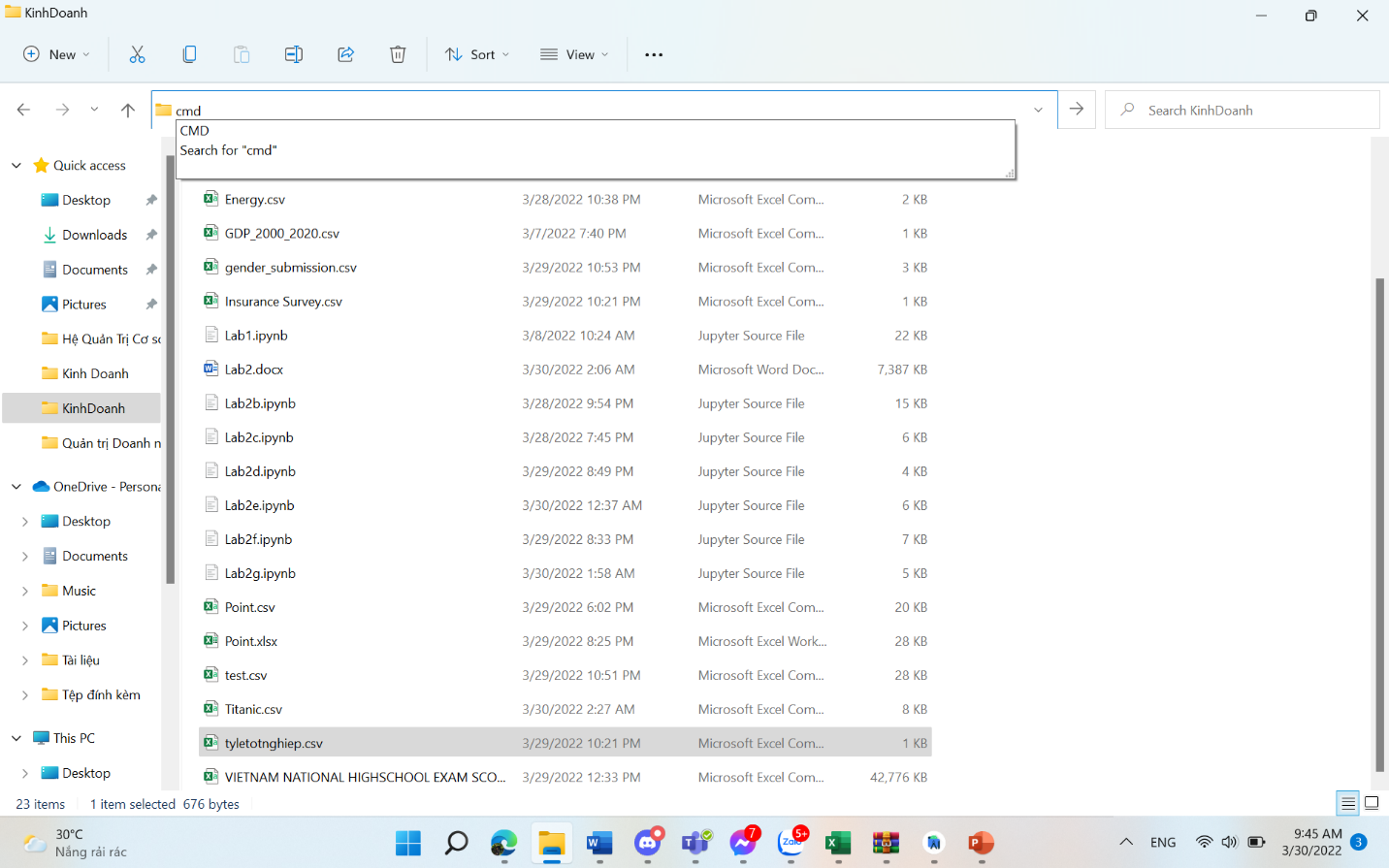
Description automatically generated**

With α = 0.05, Nam-Bac different from the other two groups (P-adj = 0.8920 > 0.05) => Reject H0, at least one mean is different from the others.

* 1. PYTHON

**Step 1:** Open Jupyter notebook

* Go to the folder containing the file **tyletotnghiep.csv** and open **cmd**



* Enter the command “jupyter notebook”

jupyter notebook

Text

Description automatically generated with low confidence

* Access the output link in the cmd

Text

Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

**Step 2:** Create new Python file

* Select the button **New** and click **Python 3** to create a new file

Graphical user interface, text, application, email

Description automatically generated

**Step 3:** Import library

import pandas as pd



**Step 4:** Import **tyletotnghiep.csv** into the Jupyter

df = pd.read\_csv("tyletotnghiep.csv")



**Step 5:** Group data into group function

df2 = df.groupby('region')['frequency '].apply(list)

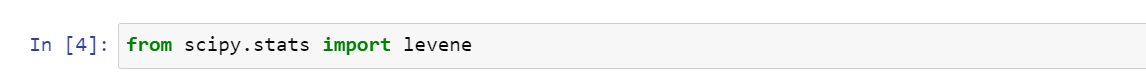
print (df2)

Calendar

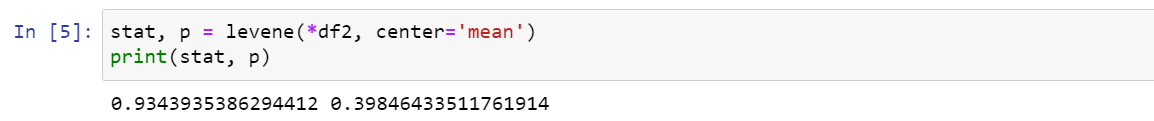
Description automatically generated with medium confidence

**Step 6:** Import library

from scipy.stats import levene

****

**Step 7:** Use Python’s Levene function to check determined Levene



p-value > 0.05 (0.398 > 0.05) => Reject H1, all variances are equal and we can test Anova

**Step 8:** Import library

import scipy.stats as stats

fvalue, pvalue = stats.f\_oneway(\*df2)

print(fvalue,pvalue)

**Graphical user interface, text, application

Description automatically generated with medium confidence**

p-value < 0.05 (0.013 < 0.05 ) 🡪 Reject H0, at least one mean is different from the other, we can test Turkey

**Step 9:** To use the Turkey function, we start **install statsmodels**

A screenshot of a computer

Description automatically generated

**Step 10:**

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

tukey = pairwise\_tukeyhsd(endog=df['frequency'],groups=df.region, alpha=0.05)

print(tukey)

**Text

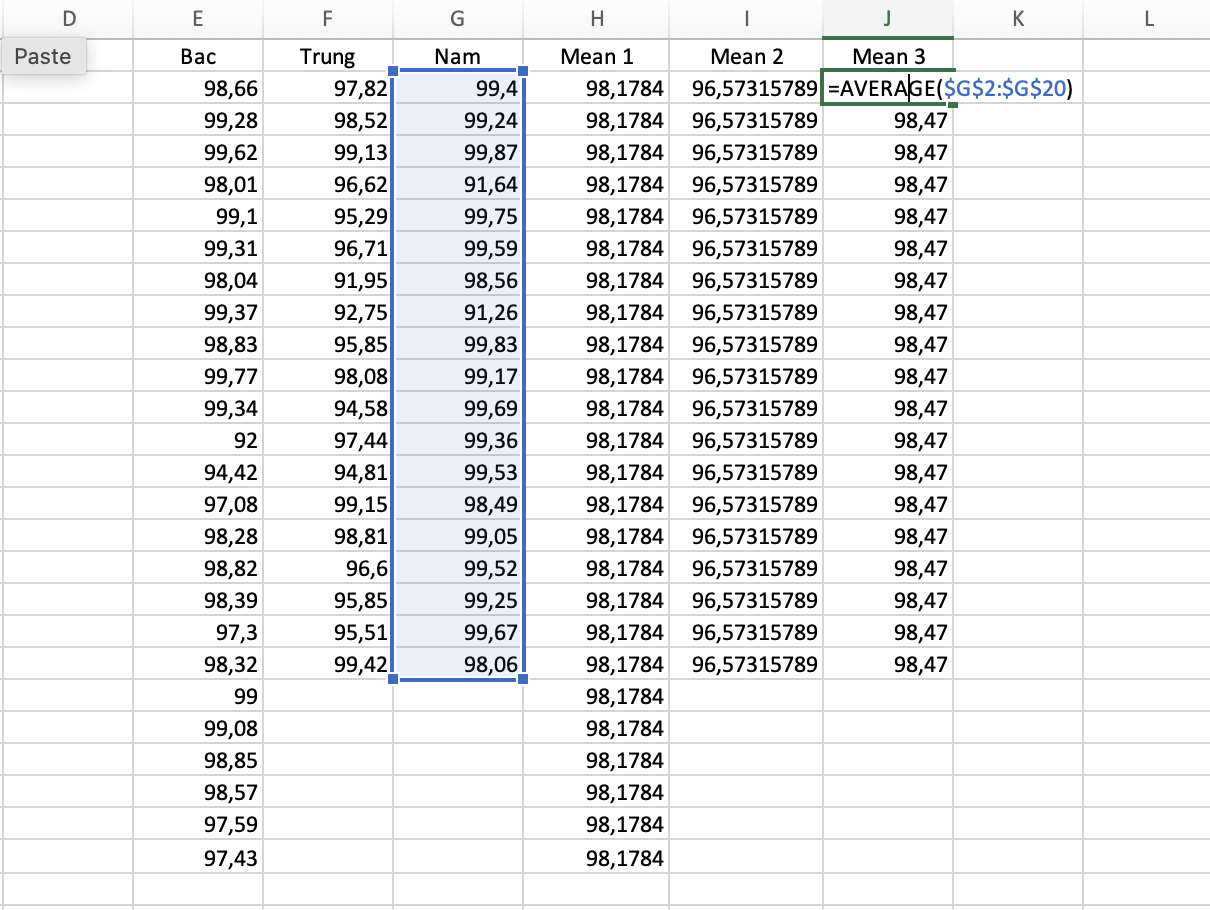
Description automatically generated with low confidence**

With α = 0.05, Bac Nam different from the other two groups (P-adj = 0.892 > 0.05) 🡪Reject H0, at least one mean is different from the others.

* 1. Excel

**Step 1:** Do Levene’s test

* Calculate the mean of each group

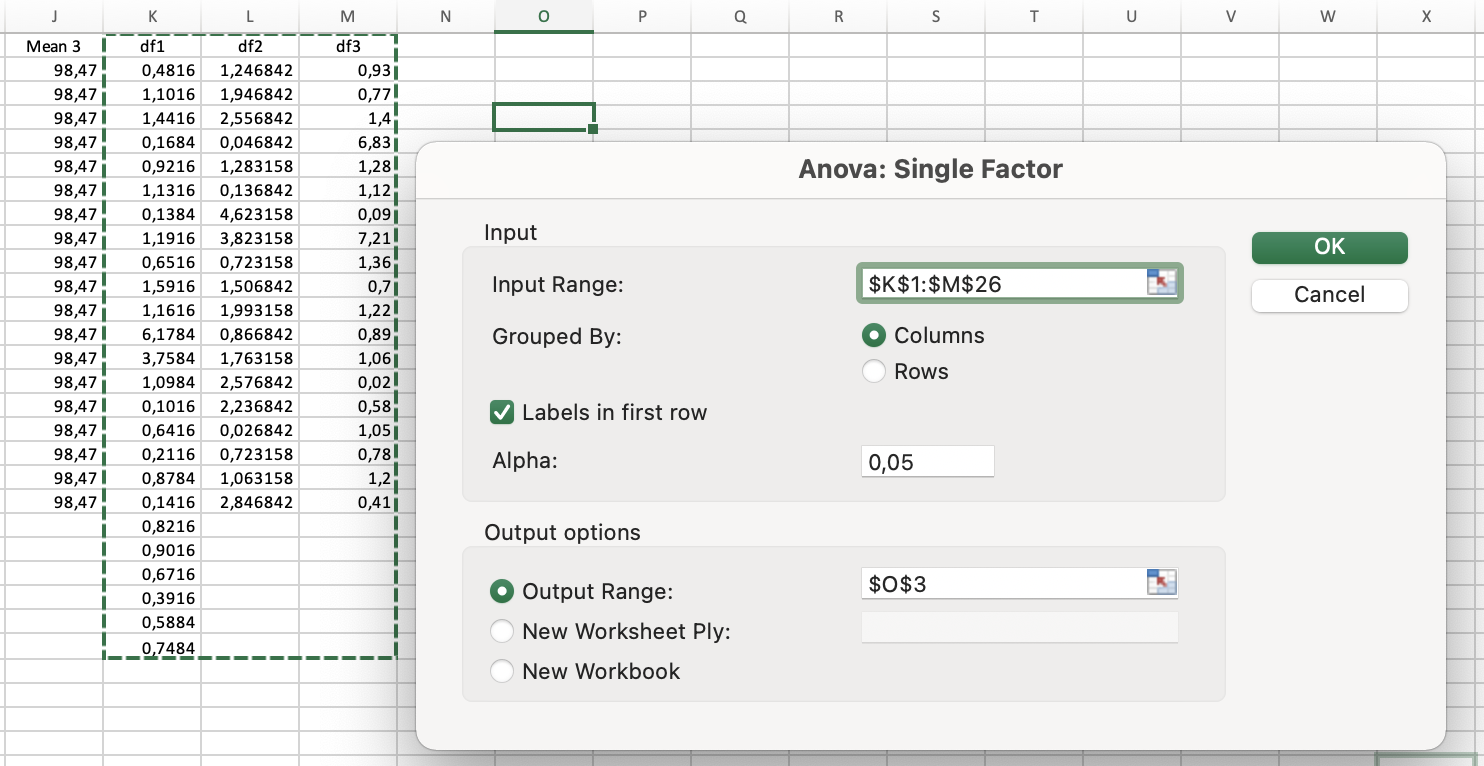


* Calculate the absolute residual

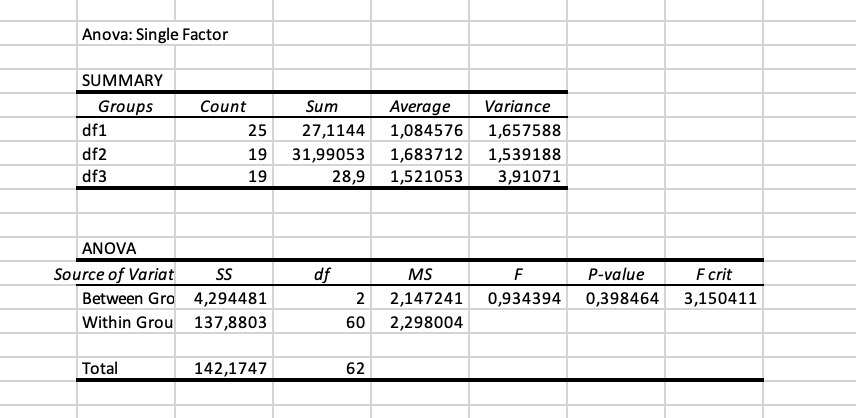
Table, Excel

Description automatically generated

* Go to Data 🡪 Data Analysis 🡪 Choose ANOVA Single Factor. For **Input Range**, choose the cells where the absolute residuals are located. For **Output Range**, choose a cell where you would like the results of the one-way ANOVA to appear. Then click **OK**.



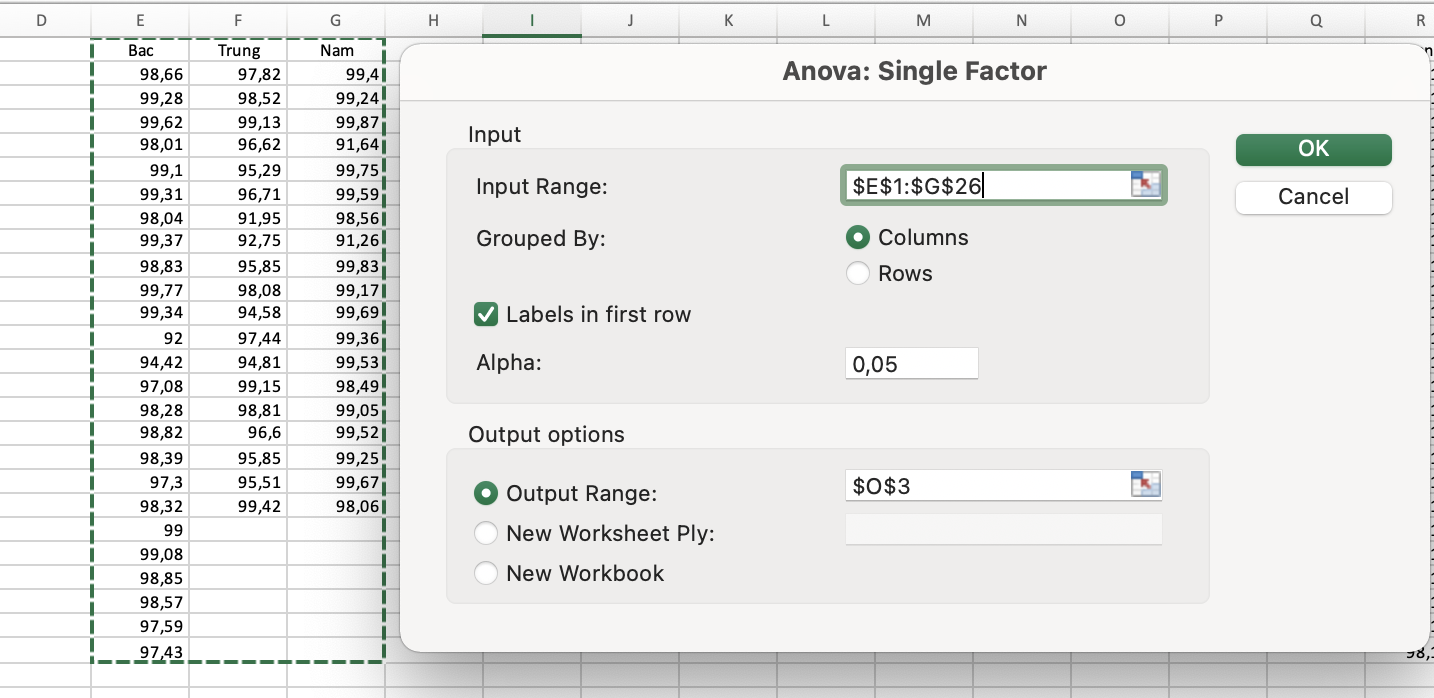
* View the result and conclude

****

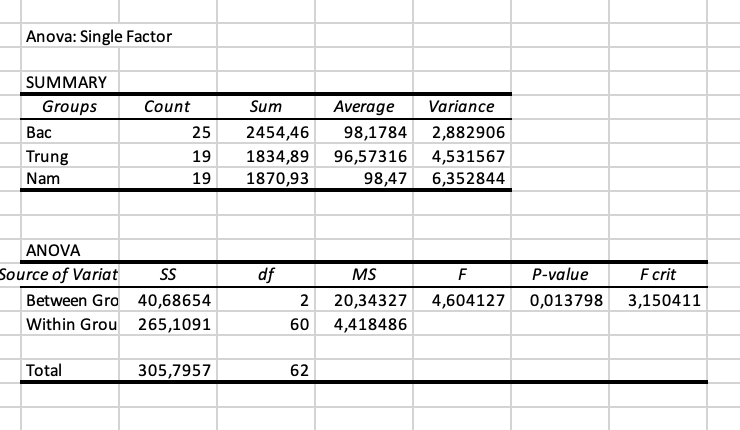
We can see that p-value is 0.398464 (>0.05). In Levene’s test, that means we can accept H0 (all the variance are equal) 🡪 Move to ANOVA test

**Step 2:** Do ANOVA test

* Go to Data 🡪 Data Analysis 🡪 Choose ANOVA Single Factor. For **Input Range**, choose the cells like the following image. For **Output Range**, choose a cell where you would like the results of the one-way ANOVA to appear. Then click **OK**.

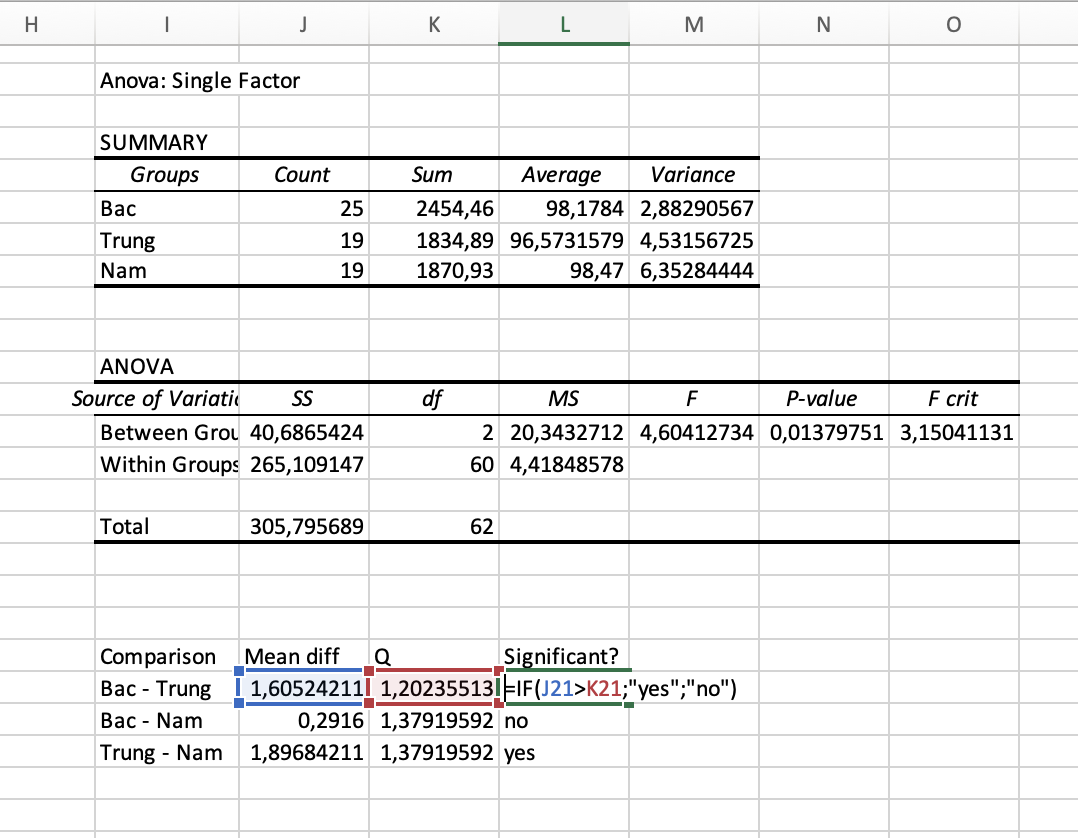


* View the result and conclude



We can see that p-value is 0.013798 (<0.05). In ANOVA test, that means we reject H0, accept H1 (at least one education level is different from the others) 🡪 Move to Tukey test to determine exactly which group means are different.

**Step 3:** Do Tukey’s test



**Conclusion:**

Based on the results of the Tukey test, we can see that there is a difference between the graduation rates of the three regions in the country. As follows:

- The North and the South have no significant difference

- There is a significant difference between the North and the Central region

- The South and the Central region also have significant differences

=> This reflects that, provinces/cities with favorable socio-economic conditions, with a tradition of learning and teaching, have a high graduation rate (such as the North and the South).

In some provinces/cities with low socio-economic conditions, the graduation rate is still not high (like the central provinces).

Through this data, we see that the government needs to pay special attention to the central provinces more, need to improve their living standards more to improve the quality of education.

1. EXERCISE 2G

**Problem statement:** This dataset is the information of 3225 Covid-19 patients that uploaded on Kaggle (Link: <https://www.kaggle.com/datasets/shirmani/characteristics-corona-patients?select=Characteristics_Corona_patients+version+5+06122020.csv>). There was so many rows of data but our team have cleaned it and just pick patients from 3 country: French, Korea and VietNam. There are 2 status of them: cured and decease. The Chi Square test helps us determine whether country affect or not to the illness severity of Covid-19 patient.

Hypothesis for Chi square test:

*H0: Illness severity and country are independent.*

*H1: Illness severity and coutry are dependent.*

* 1. R

**Step 1:**

* Convert **Covidpatient.xlsx** file to **Covidpatient**.**csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step2:**

* Import file \*csv in step 1 into RStudio Application

covid\_lab <-read.csv('Covidpatient.csv', header = TRUE)



* Attach this file

attach(covid\_lab)

****

**Step 3:** We have to install and use library MASS

install.packages("MASS")

****

* In tab Packages, select **Install**

Graphical user interface, application, table

Description automatically generated

* Enter the library name, then click **Install**

**Graphical user interface, text, application, email

Description automatically generated**

* After the installation is complete, we use the **require()** statement to use the library

require(MASS)



**Step 4:**

* Group Statistics

tb = table(country, severity\_illness)

Text

Description automatically generated with medium confidence

* Chi-quared Test

chisq.test(tb)

Text

Description automatically generated

p-value = 2.2e-16 < 0.05 🡪 Reject H0, two categorical variables are dependent.

* 1. Python

**Step 1:** Open Jupyter notebook

* Go to the folder containing the file **Covidpatient.csv** and open **cmd**

Graphical user interface, text, application

Description automatically generated

* Enter the command “jupyter notebook”

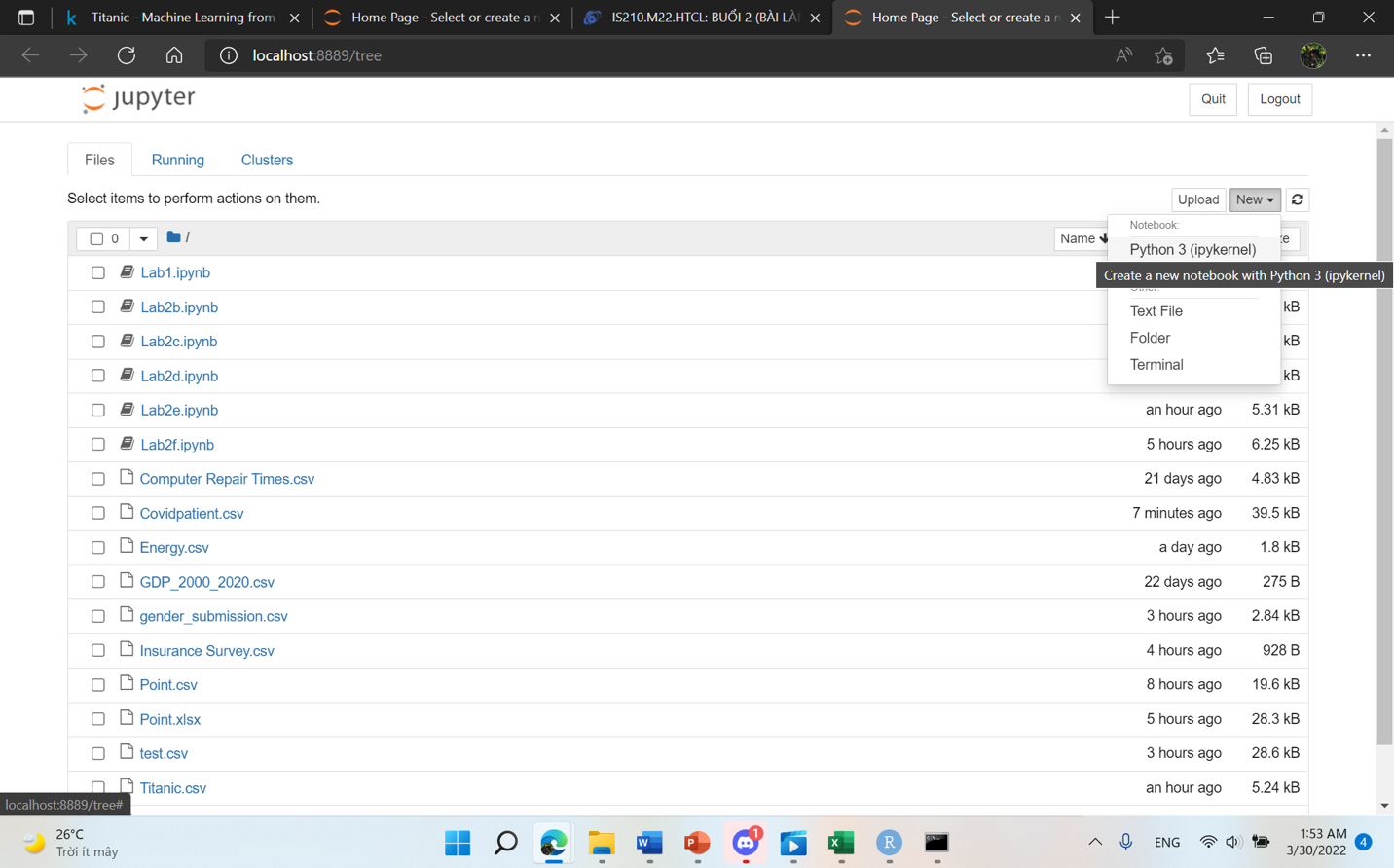
jupyter notebook

* Access the output link in the cmd

A screenshot of a computer

Description automatically generated with medium confidence

**Step 2:** Create new Python file



* Select the button **New** and click **Python 3** to create a new file

**Step 3:** Import library

import pandas as pd



**Step 4:** Import **Covidpatient.csv** into the Jupyter

pt = pd.read\_csv("Covidpatient.csv")



**Step 5:** Chi-quare Test

chisqt=pd.crosstab(pt.country, pt.severity\_illness)

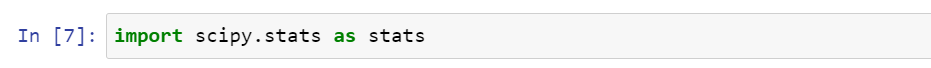
print(chisqt)

Table

Description automatically generated

**Step 6:** Import library

import scipy.stats as stats

****

**Step 7:** p – value

c, p, dof, expected = stats.chi2\_contingency(chisqt)

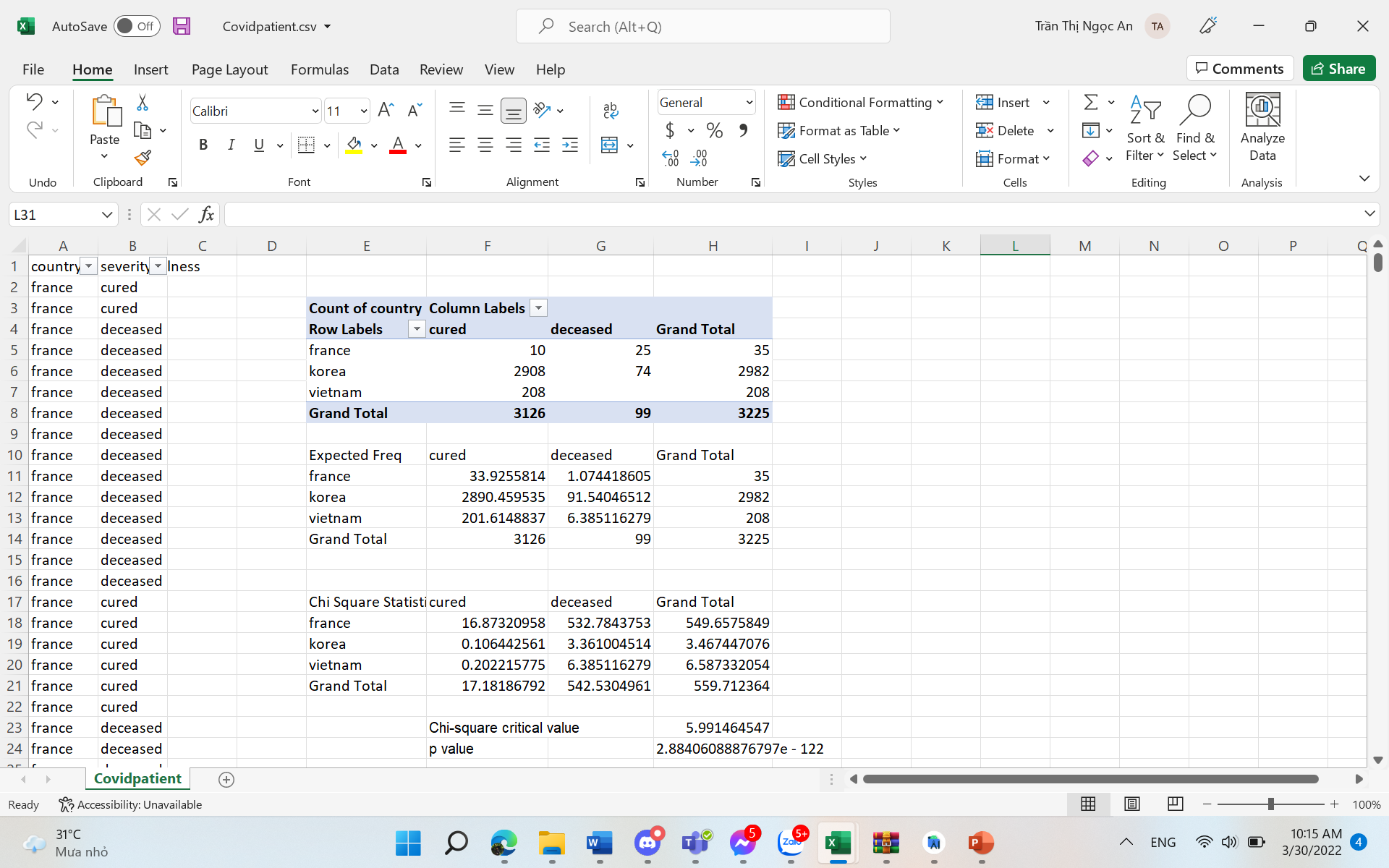
print(p)

**Graphical user interface, text

Description automatically generated with medium confidence**

p-value < 0.05 🡪 Reject H0, two categorical variables are dependent.

* 1. MS EXCEL



Comment:

* Chi-square Critical value < Chi-square Statistic (5.99 < 559,712) => Reject H0, two categorical variables are dependent.
* Or (p value): p value = 2,88e-122 < 0.05 🡪 Reject H0, two categorical variables are dependent.

**Conclusion:** Based on the results of the Chi-quare test. We can see that Illness severity and country are dependent on each other.

They depend on each other because each country will have different policies in the prevention of covid 19. That leads to different survival and death rates in each country.

This data also shows whether the control measures in those countries have been effective or not. What do they need to pay attention to to reduce the death rate and increase the survival rate

For example, Vietnam, when we look at the data, we can see that the number of deaths here is very low (almost none) and the number of people who have been cured is very high. This shows that Vietnam has applied very effective measures to prevent and control COVID-19.

REFERENCES

1. James R.Evans, “Chapter 7: Statistical Inference” from “Business Analytics: Methods, Models, and Decisions, 1st edition”. [Online]. Link: <https://courses.uit.edu.vn/pluginfile.php/323175/mod_folder/content/0/Ch7-Statistical%20Inference-T.pptx?forcedownload=1>
2. Associate Professor Ph.D Nguyễn Đình Thuân, “Bai giang Phan tich du lieu kinh doanh”, 09/2017. [Online]. Link: <https://courses.uit.edu.vn/mod/resource/view.php?id=130681>
3. Nguyễn Minh Nhựt, “LAB02\_PYTHON\_2532022.pdf”. [Online]. Link: <https://bitly.com.vn/xrrt6s> [Last updated: 25/03/2022].
4. “Levene Test for Equality of Variance” from “Engineering Statistic Handbook, 1.3.5.10”. [Online]. Link: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm>
5. “Tukey’s method” from “Engineering Statistics Handbook, 7.4.7.1” [Online]. Link: <https://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm>
6. Zach, “How to Perform Levene’s Test in Excel”, 01/04/2020. [Online]. Link: <https://www.statology.org/levenes-test-excel/>
7. Zach, “How to Perform a Tukey-Kramer Post Hoc Test in Excel”, 27/07/2020. [Online]. Link: <https://www.statology.org/tukey-kramer-post-hoc-test-excel/>
8. General Statistics Office of Vietnam, Link: [https://www.gso.gov.vn/px-web-2/?pxid=V1015&theme=Giáo%20dục](https://www.gso.gov.vn/px-web-2/?pxid=V1015&theme=Gi%C3%A1o%20d%E1%BB%A5c)
9. Kaggle, Link: <https://www.kaggle.com/datasets/shirmani/characteristics-corona-patients?select=Characteristics_Corona_patients+version+5+06122020.csv>