Shape, square

Description automatically generated

Logo, company name

Description automatically generated

**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEM**

———h&g———

Ho Chi Minh City, March 2022

**CLASS: IS211.M11.HTCL**

**TEAM: 3**

**STUDENTS:**

TRẦN MẪN QUÂN 19520873

ĐẶNG NGUYỄN PHƯỚC AN 19521171

TRẦN THỊ NGỌC AN 19521189

**DATA ANALYSIS IN BUSINESS**

**CONTENT: REPORT LAB 1**

**CONTENT**

[WORK ASSIGNMENT TABLE 1](#_Toc97677253)

[I. EXERCISE 1A: 2](#_Toc97677254)

[II. EXERCISE 1B 18](#_Toc97677255)

[1. DATA ANALYSIS OF PURCHASE ORDERS FILE 18](#_Toc97677256)

[2. DATA ANALYSIS OF COMPUTER REPAIR TIMES FILE 30](#_Toc97677257)

[III. EXERCISE 1C 41](#_Toc97677258)

[1. DISTINGUISH 41](#_Toc97677259)

[2. MANIPULATE 42](#_Toc97677260)

[IV. REFERENCE 46](#_Toc97677261)

WORK ASSIGNMENT TABLE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Student’name** | **Student’s ID** | **Assignment** | **Evaluation(%)** |
| 1 | Trần Mẫn Quân | 19520873 | - Doing Ex 1C, 2  - Writing and formating final report | 100% |
| 2 | Đặng Nguyễn Phước An | 19521171 | - Doing Ex 1B  (Using R and Python languages to calculate data)  - Writing raw report | 100% |
| 3 | Trần Thị Ngọc An | 19521189 | - Doing Ex 1A, Ex 1B  (Using Excel to calculate data)  - Searching reference | 100% |

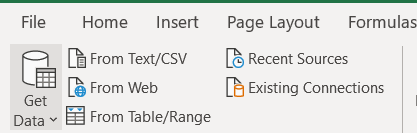
1. EXERCISE 1A:

**TOPIC**: DATA ANALYSIS OF VIETNAM'S AVERAGE GDP FILE

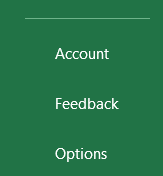
**MS EXCEL**

**Step 1:** Open Data Analysis

* Open file **Purchase Orders.csv** and choose **Data Analysis** in Microsoft Excel. Then choose File in the Ribbon



* Choose Option



* Choose **Add-ins,** then click the button **GO**



* Tick **Analysis ToolPak** and click **OK**

**Graphical user interface, text, application

Description automatically generated**

* Return to the file and click on tab **DATA**

****

* Choose Data Analysis



* In tab **Analysis Tools**, choose **Descriptive Statistics**, then click button **OK**

**Graphical user interface, application

Description automatically generated**

* Select option like the picture below and click OK

**Graphical user interface, application

Description automatically generated**

**Step 2:** Calculate and explain the meaning of the values

**Table

Description automatically generated**

**Explanation of values:**

* Mean =AVERAGE(B2:B22)
* Median = MEDIAN(B2:B22)
* Max = MAX(B2:B22)
* Min = MIN(B2:B22)
* Range = Max – Min
* Count = COUNT(B2:B22)
* Sum = SUM(B2: B22)
* Mode = MODE(B2:B22)
* First Quantile = QUARTILE.INC(B2:B22,1)
* Third Quantile = QUARTILE.INC(B2:B22,3)
* IQR = Third Quantile - First Quantile
* Variance = VAR.S(B2:B22)
* Standard Deviation = STDEV.S(B2:B22)
* Coefficient of Deviation = Standard Deviation / Mean
* Skewness = SKEW(B2:B22)
* Kurtosis = KURT(B2:B22)

**Step 3:** Draw Graphs

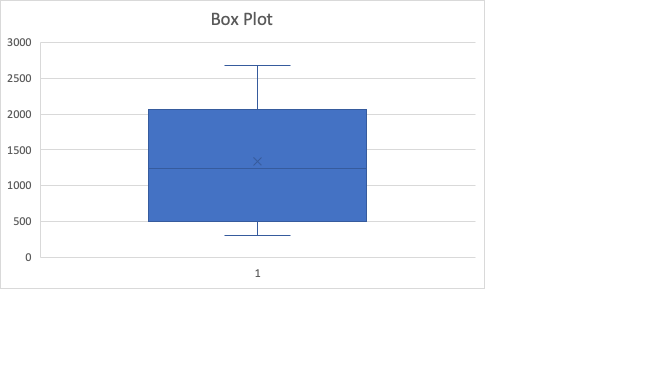
* We proceed to highlight the GDP column, then go to Insert => Chart as shown below.

Then choose the graph want to use.

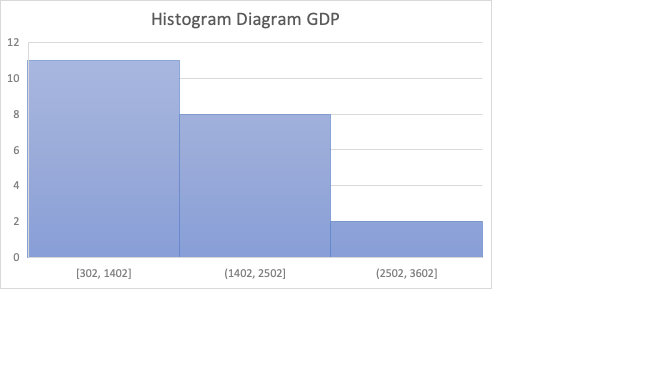
Graphical user interface, application, table, Excel

Description automatically generated

* **BOX PLOT**



* **HISTOGRAM DIAGRAM**



**R**

**Step 1:**

* Convert GDP\_2000\_2020.xlsx file to GDP\_2000\_2020.csv. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step 2**:

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

gdp <- read.csv("D:/KinhDoanh/GDP\_2000\_2020.csv", header = TRUE)



**Step 3:** Attach this file

attach(gdp)

****

**Step 4**: Calculate and explain the meaning of the values

* **COUNT**

**Meaning:** Count the total number of rows of data in the column “GDP”

length(GDP)

A picture containing logo

Description automatically generated

* **MIN**

**Meaning:** Minimum value of “GDP” column

min(GDP)

**Icon

Description automatically generated**

* **MAX**

**Meaning:** Maximumvalue of “GDP” column

max(GDP)

**Icon

Description automatically generated**

* **MEAN**

**Meaning:** Average value of data “Cost per order” column

mean(GDP)

**A picture containing background pattern

Description automatically generated**

* **MEDIAN**

**Meaning:** Middle value of the data when arranged from least to greatest. The median divides the values into 2 equal parts

median(GDP)

Icon

Description automatically generated

* **MODE**

**Meaning:** observation that occurs most often or, for grouped data, the group with the greatest frequency

> getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

> getmode(GDP)

Text

Description automatically generated with medium confidence

* **RANGE**

**Meaning:** Is the distance between the smallest observed value and the largest observed value

(RANGE = MAX - MIN)

> range <- max(GDP) - min(GDP)

> range

A picture containing graphical user interface

Description automatically generated

* **VARIANCE**

**Meaning:** Is an average of the squared deviations form the mean (uses all data values).

var(GDP)

A picture containing icon

Description automatically generated

* **STANDARD DEVIATION**

**Meaning:** Is the square root of the variance.

sd(GDP)

A picture containing icon

Description automatically generated

* **COEFFICIENT OF DEVIATION**

**Meaning:** Is a statistical measure of the relative dispersion of data points in a data series around the mean.

cv <- sd(GDP)/ mean(GDP)

A picture containing text

Description automatically generated

* **SKEWNESS**

**Meaning:** Shows the distribution of observed values that can be used to test for normal distribution.

* CS is negative for left-skewed data.
* CS is positive for right-skewed data.
* |CS| > 1 suggests high degree of skewness.
* 0.5 ≤ |CS| ≤ 1 suggests moderate skewness.
* |CS| < 0.5 suggests relative symmetry

**Note :** To calculate skewness, we have to install and use library **e1071.**

* In tab Packages, select **Install**

Graphical user interface, application, table

Description automatically generated

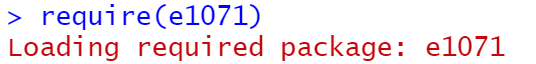
* Enter the library name, then click **Install**

Graphical user interface, text, application, email

Description automatically generated

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

require(e1071)



* After using the library successfully, we calculate the skewness value through the command

skewness(GDP, type=2)

A picture containing text

Description automatically generated

**Meaning of value**: Because the value of skewness is approximately 0.19 (CS<1). This number shows that this data is suggests moderate skewness.

* **KURTOSIS**

**Meaning:** Used to evaluate the peak of the observation curve with the form of a normal distribution curve.

* CK < 3 indicates the data is somewhat flat with a wide degree of dispersion.
* CK > 3 indicates the data is somewhat peaked with less dispersion

kurtosis(GDP, type =2)

A picture containing graphical user interface

Description automatically generated

**Meaning of value**: Because the value of kurtosis is approximately -1,4 (CK<3). This number shows that indicates the data is somewhat flat with a wide degree of dispersion.

* **QUANTILE**

**Meaning:** A quantile, or percentile, will indicate what percentage of the data falls below a certain value.

> quantile(GDP, 0.25)

> quantile(GDP, 0.5)

> quantile(GDP, 0.75)

> quantile(GDP, 0.1)

Graphical user interface, text

Description automatically generated

* **IQR**

**Meaning:** Is difference between the third and first quartiles.

IQR <-quantile(GDP,0.75) - quantile(GDP, 0.25)

A picture containing text

Description automatically generated

* **BOX PLOT**

boxplot(GDP, main = "Box Plot", ylab = "USD", col = "lightgray")



Chart, box and whisker chart

Description automatically generated

* **HISTOGRAM**

hist(GDP, main = "Histogram Diagram", ylab = "Freq", col = " lightgray")



Chart, histogram

Description automatically generated

**PYTHON**

**Step 1:** Open Jupyter notebook

* Go to the folder containing the file **GDP\_2000\_2020.csv** and open **cmd**

Graphical user interface, text, application, email

Description automatically generated

* Enter the command “jupyter notebook”

jupyter notebook

Text

Description automatically generated

* Access the output link in the cmd

Text

Description automatically generated

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 numpy as np  
import pandas as pd  
import statistics as st  
import matplotlib.pyplot as plt  
from scipy.stats import skew  
from scipy.stats import kurtosis

Text, letter

Description automatically generated

**Step 4:** Import **GDP\_2000\_2020.csv** into the Jupyter

gdp = pd.read\_csv("GDP\_2000\_2020.csv")

cost



**Step 5:** Calculate and explain the meaning of the values

* **COUNT**

len(gdp)

Graphical user interface, application, rectangle

Description automatically generated with medium confidence

* **MIN**

min(gdp.GDP)

****

* **MAX**

max(gdp.GDP)

****

* **MEAN**

st.mean(gdp.GDP)

****

* **MEDIAN**

st.median(gdp.GDP)



* **RANGE**

range = max(gdp.GDP)- min(gdp.GDP)> range



* **MODE**

from collections import Counter

def my\_mode(sample):

... c = Counter(sample)

... return [k for k, v in c.items() if v == c.most\_common(1)[0][1]]

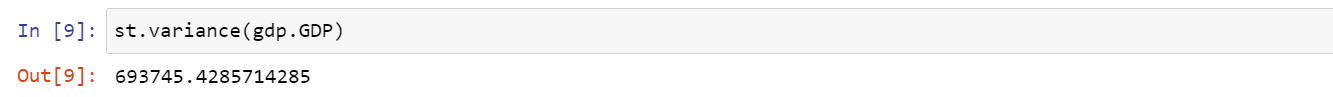
my\_mode(gdp.GDP)

Graphical user interface, text, application, email

Description automatically generated

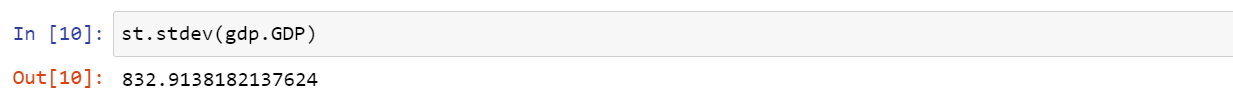
* **VARIANCE**

st.variance(gdp.GDP)



* **STANDARD DEVIATION**

st.stdev(gdp.GDP)



* **COEFFICIENT OF DEVIATION**

cv = st.stdev(gdp.GDP)/st.mean(gdp.GDP)



* **SKEWNESS**

skew(gdp.GDP)



* **KURTOSIS**

**Meaning:** Used to evaluate the peak of the observation curve with the form of a normal distribution curve.

kurtosis(gdp.GDP)



* **QUANTILE**

np.quantile(gdp.GDP, 0.25)

np.quantile(gdp.GDP, 0.5)

np.quantile(gdp.GDP, 0.75)

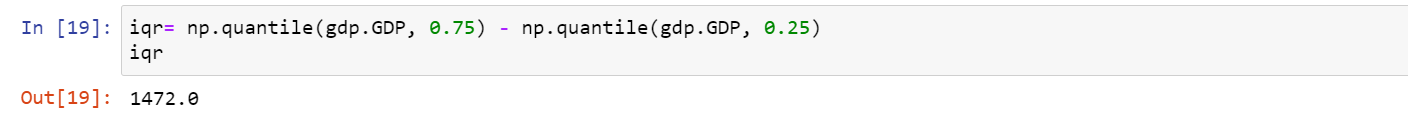
np.quantile(gdp.GDP, 1)

Graphical user interface, text, application, email

Description automatically generated

* **IQR**

IQR = np.quantile(gdp.GDP, 0.75) - np.quantile(gdp.GDP, 0.25)



* **BOX PLOT**

plt.boxplot(gdp.GDP)

plt.show()

Graphical user interface, application, Word

Description automatically generated

* **HISTOGRAM**

plt.hist(gdp.GDP)

plt.show()

**Chart, histogram

Description automatically generated**

1. EXERCISE 1B
   1. DATA ANALYSIS OF PURCHASE ORDERS FILE

**MS EXCEL**

**Step 1:** Open Data Analysis

**Step 2:** Calculate and explain the meaning of the values like the picture

**Chart

Description automatically generated**

**R**

**Step 1:**

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

**Graphical user interface, text, application, email

Description automatically generated**

**Step 2**:

Import file \*csv in step 1 into RStudio Application

cost <-read.csv("D:/0.University/3.Nam 3/HK2/11.PT Kinh Doanh/4.Bai\_Tap/Lab1/Purchase Orders.csv", header = TRUE)

Graphical user interface, application

Description automatically generated

**Step 3:** Attach this file

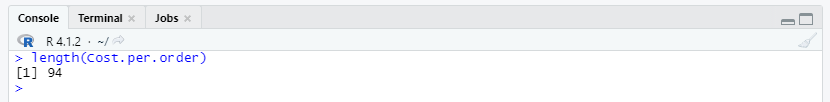
attach(cost)

****

**Step 4**: Calculate and explain the meaning of the values

* **COUNT**

length(Cost.per.order)



* **MIN**

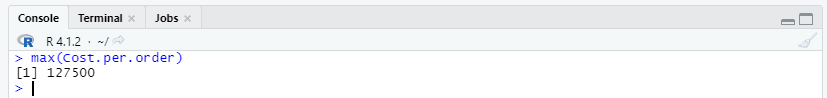
min(Cost.per.order)

**Background pattern

Description automatically generated with medium confidence**

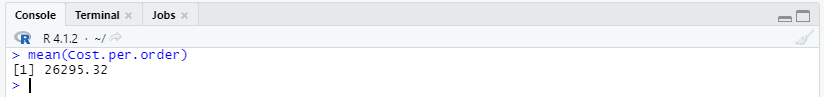
* **MAX**

max(Cost.per.order)

****

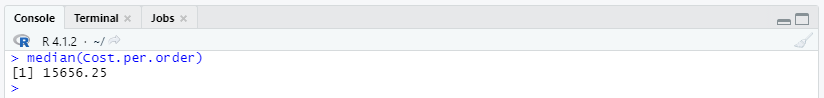
* **MEAN**

mean(Cost.per.order)

****

* **MEDIAN**

median(Cost.per.order)



* **MODE**

> getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

> getmode(Cost.per.order)

Graphical user interface, text, application

Description automatically generated

* **RANGE**

> range <- max(Cost.per.order) - min(Cost.per.order)

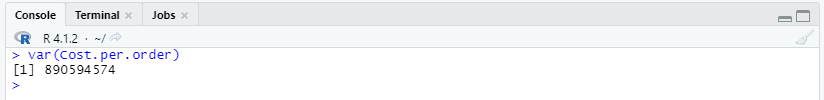
> range

Graphical user interface

Description automatically generated with medium confidence

* **VARIANCE**

var(Cost.per.order)



* **STANDARD DEVIATION**

sd(Cost.per.order)

Background pattern

Description automatically generated with medium confidence

* **COEFFICIENT OF DEVIATION**

cv <- sd(Cost.per.order)/ mean(Cost.per.order)

Graphical user interface, text, application

Description automatically generated

* **SKEWNESS**

skewness(Cost.per.order, type=2)

Graphical user interface, application

Description automatically generated

**Meaning of value**: Because the value of skewness is approximately 1.6 (CS>1). This number shows that this data is skewed **to the right very high**

* **KURTOSIS**

kurtosis(Cost.per.order, type =2)

Graphical user interface, application

Description automatically generated

**Meaning of value:** Because the value of kurtosis is approximately 2.07 (0<CK<3). This number shows that indicates the data is somewhat flat with a wide degree of dispersion.

* **QUANTILE**

**Meaning:** A quantile, or percentile, will indicate what percentage of the data falls below a certain value..

> quantile(Cost.per.order, 0.25)

> quantile(Cost.per.order, 0.5)

> quantile(Cost.per.order, 0.75)

> quantile(Cost.per.order, 0.1)

Graphical user interface, text, application

Description automatically generated

* **IQR**

IQR <-quantile(Cost.per.order,0.75) - quantile(Cost.per.order, 0.25)

Graphical user interface, text, application, Word

Description automatically generated

* **BOX PLOT**

boxplot(Cost.per.order, main = "Box Plot", ylab = "USD", col = "lightgray")

Graphical user interface, application

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

* **HISTOGRAM**

hist(Cost.per.order, main = "Histogram Diagram", ylab = "USD", col = " lightgray")

**Graphical user interface, application

Description automatically generated**

**Chart, histogram

Description automatically generated**

**PYTHON**

**Step 1:** Open Jupyter notebook

jupyter notebook

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 numpy as np  
import pandas as pd  
import statistics as st  
import matplotlib.pyplot as plt  
from scipy.stats import skew  
from scipy.stats import kurtosis

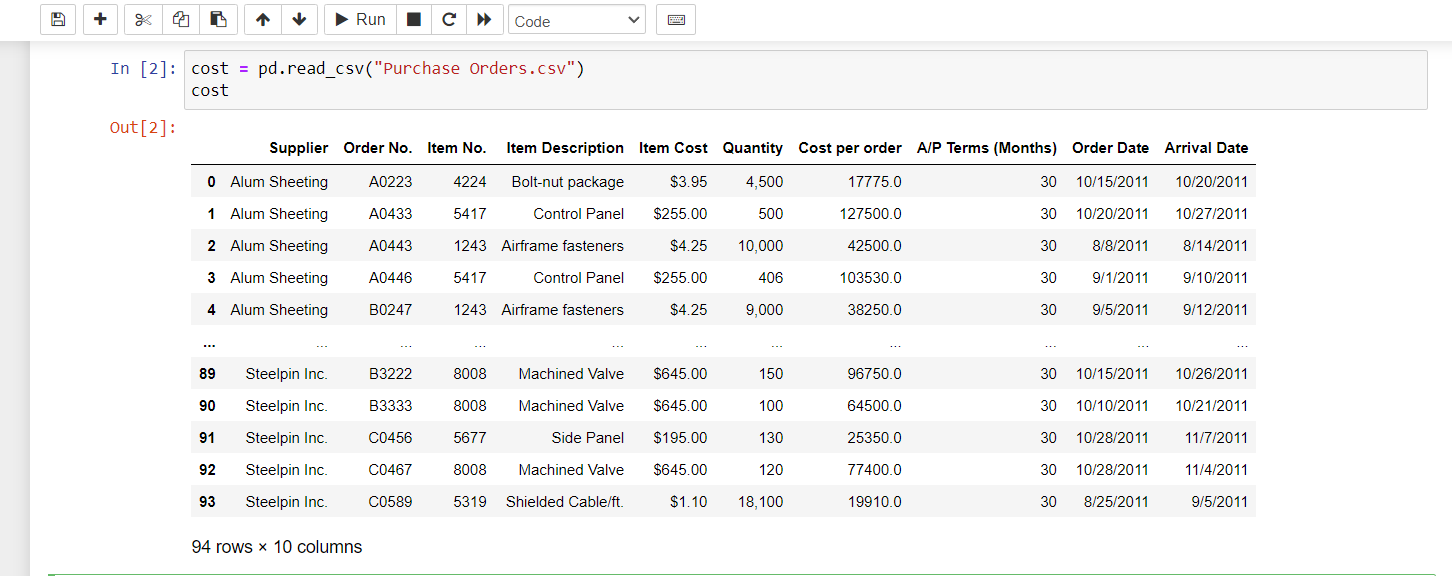
**Graphical user interface, application

Description automatically generated**

**Step 4:** Import **Purchase Orders.csv** into the Jupyter

cost = pd.read\_csv("Purchase Orders.csv")

cost



**Step 5:** Rename column **Cost per order**

cost=cost.rename(columns={'Cost per order': 'cpo'})



**Step 6:** Calculate and explain the meaning of the values

* **COUNT**

len(cost)

Graphical user interface, text, application

Description automatically generated

* **MIN**

min(cost.cpo)

**Graphical user interface, text, application, chat or text message

Description automatically generated**

* **MAX**

max(cost.cpo)

**Graphical user interface

Description automatically generated with low confidence**

* **MEAN**

st.mean(cost.cpo)

**Graphical user interface, text, application, Word

Description automatically generated**

* **MEDIAN**

st.median(cost.cpo)

Graphical user interface, text, application

Description automatically generated

* **RANGE**

range = max(cost.cpo)- min(cost.cpo)> range

Graphical user interface, text, application

Description automatically generated

* **MODE**

>from collections import Counter

>def my\_mode(sample):

... c = Counter(sample)

... return [k for k, v in c.items() if v == c.most\_common(1)[0][1]]

>my\_mode(cost.cpo)

Graphical user interface, text, application

Description automatically generated

* **VARIANCE**

st.variance(cost.cpo)

Graphical user interface, text, application

Description automatically generated

* **STANDARD DEVIATION**

st.stdev(cost.cpo)

Graphical user interface, text, application

Description automatically generated

* **COEFFICIENT OF DEVIATION**

cv = st.stdev(cost.cpo)/st.mean(cost.cpo)

Graphical user interface, text, application, email

Description automatically generated

* **SKEWNESS**

skew(cost.cpo)

Text

Description automatically generated with medium confidence

* **KURTOSIS**

kurtosis(cost.cpo)

Graphical user interface

Description automatically generated

* **QUANTILE**

>np.quantile(cost.cpo, 0.25)

>np.quantile(cost.cpo, 0.75)

Graphical user interface, text, application, chat or text message

Description automatically generated

* **IQR**

IQR = np.quantile(cost.cpo, 0.75) - np.quantile(cost.cpo, 0.25)

Graphical user interface, application, Word

Description automatically generated

* **BOX PLOT**

plt.boxplot(cost.cpo)

plt.show()

Chart, box and whisker chart

Description automatically generated

* **HISTOGRAM**

**Chart, histogram

Description automatically generated**

plt.hist(cost.cpo)

plt.show()

* 1. DATA ANALYSIS OF COMPUTER REPAIR TIMES FILE

**MS EXCEL**

**Step 1:** Open Data Analysis

* Choose the **DATA tab**

****

* Choose Data Analysis



* In tab **Analysis Tools**, choose **Descriptive Statistics**, then click button **OK**

**Graphical user interface, application

Description automatically generated**

* Select option like the picture below and click OK

**Graphical user interface, application

Description automatically generated**

* The results like the picture below

**Table

Description automatically generated**

**Step 2:** Calculate and explain the meaning of the values

**Chart, box and whisker chart

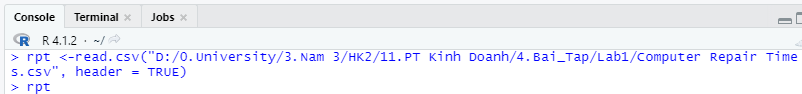
Description automatically generated**

**R**

**Step 1**:

Import file \*csv in step 1 into RStudio Application

rpt <-read.csv("D:/0.University/3.Nam 3/HK2/11.PT Kinh Doanh/4.Bai\_Tap/Lab1/Computer Repair Times.csv", header = TRUE)



**Step 3:** Attach this file

attach(rpt)

**Graphical user interface, text, application

Description automatically generated**

**Step 4**: Calculate and explain the meaning of the values

* **COUNT**

length(RPT)

Graphical user interface, text, application

Description automatically generated

* **MIN**

min(RPT)

**Graphical user interface, application

Description automatically generated**

* **MAX**

max(RPT)

**Graphical user interface, text, application, email

Description automatically generated**

* **MEAN**

mean(RPT)

**Graphical user interface, text, application

Description automatically generated**

* **MEDIAN**

median(RPT)

Graphical user interface, text, application

Description automatically generated

* **MODE**

> getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

> getmode(RPT)

Graphical user interface, text, application, email

Description automatically generated

* **RANGE**

> range <- max(RPT) - min(RPT)

> range



* **VARIANCE**

var(RPT)

Graphical user interface, text, application

Description automatically generated

* **STANDARD DEVIATION**

sd(RPT)

Graphical user interface, application

Description automatically generated

* **COEFFICIENT OF DEVIATION**

cv <- sd(RPT)/ mean(RPT)

Graphical user interface, text, application

Description automatically generated

* **SKEWNESS**

skewness(RPT, type=2)

Graphical user interface, text, application

Description automatically generated

**Meaning of value skewness**: Because the value of skewness is approximately 1.6 (CS>1). This number shows that this data is skewed **to the right very high**

* **KURTOSIS**

**Meaning:** Used to evaluate the peak of the observation curve with the form of a normal distribution curve.

kurtosis(RPT, type =2)

Graphical user interface, application

Description automatically generated

**Meaning of value kurtosis**: Because the value of kurtosis is approximately 4.07 (CK>3). This number shows that indicates the data is somewhat peaked with less dispersion.

* **QUANTILE**

> quantile(RPT, 0.25)

> quantile(RPT, 0.5)

> quantile(RPT, 0.75)

> quantile(RPT, 0.1)

Graphical user interface, text, application

Description automatically generated

* **IQR**

IQR <- quantile(RPT, 0.75)- quantile(RPT, 0.25)

Graphical user interface, application

Description automatically generated

* **BOX PLOT**

boxplot(RPT, main = "Box Plot", ylab = " Days", col = "lightgray")

Graphical user interface, application

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* **HISTOGRAM**

hist(RPT, main = " Histogram Diagram ", ylab = " Days", col = "lightgray")

**Graphical user interface, application

Description automatically generated**

**Chart, histogram

Description automatically generated**

**c. PYTHON**

**Step 1:** Open Jupyter notebook

jupyter notebook

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 numpy as np  
import pandas as pd  
import statistics as st  
import matplotlib.pyplot as plt  
from scipy.stats import skew  
from scipy.stats import kurtosis

**Graphical user interface, application

Description automatically generated**

**Step 4:** Import **Purchase Orders.csv** into the Jupyter

Graphical user interface, text, application

Description automatically generated

rpt = pd.read\_csv("Computer Repair Times.csv")

**Step 5:** Calculate and explain the meaning of the values

* **COUNT**

len(rpt)

Graphical user interface, text, application, chat or text message

Description automatically generated

* **MIN**

min(rpt.RPT)

**Graphical user interface, text, application, chat or text message

Description automatically generated**

* **MAX**

max(rpt.RPT)

**Graphical user interface, application, chat or text message

Description automatically generated**

* **MEAN**

st.mean(rpt.RPT)

**Graphical user interface, application

Description automatically generated**

* **MEDIAN**

st.median(rpt.RPT)

Graphical user interface, application, chat or text message

Description automatically generated

* **RANGE**

> range = max(rpt.RPT)- min(rpt.RPT)

> range

Graphical user interface, text, application, email

Description automatically generated

* **MODE**

>from collections import Counter

>def my\_mode(sample):

... c = Counter(sample)

... return [k for k, v in c.items() if v == c.most\_common(1)[0][1]]

>my\_mode(cost.cpo)

Graphical user interface, text, application

Description automatically generated

* **VARIANCE**

st.variance(rpt.RPT)

Graphical user interface, text, application

Description automatically generated

* **STANDARD DEVIATION**

st.stdev(rpt.RPT)

Text

Description automatically generated

* **COEFFICIENT OF DEVIATION**

cv = st.stdev(rpt.RPT)/st.mean(rpt.RPT)

Graphical user interface, text, application, email

Description automatically generated

* **SKEWNESS**

skew(rpt.RPT)

Graphical user interface, text, application

Description automatically generated

* **KURTOSIS**

kurtosis(rpt.RPT)

Graphical user interface, text, application

Description automatically generated

* **QUANTILE**

>np.quantile(rpt.RPT, 0.25)

>np.quantile(rpt.RPT, 0.75)

Graphical user interface, text, application, chat or text message

Description automatically generated

* **IQR**

IQR = np.quantile(rpt.RPT, 0.75)- np.quantile(rpt.RPT, 0.25)

Graphical user interface, text, application, email

Description automatically generated

* **BOX PLOT**

plt.boxplot(rpt.RPT)

plt.show()

Chart, box and whisker chart

Description automatically generated

* **HISTOGRAM**

plt.hist(rpt.RPT)

plt.show()

**Chart, histogram

Description automatically generated**

1. EXERCISE 1C
   1. DISTINGUISH
2. **Null hypothesis and Alternative hypothesis**

**Alternative hypothesis**: describe a theory that needs to be tested. We usually symbolize it with H1

**Null hypothesis**: is a negative statement of the alternative hypothesis. We usually symbolize it with H0. If H0 is rejected, then H1 is failed to reject.

For example, suppose we hypothesize that the salary of a data analyst is greater than 2000$/month. To test this, we can define the following null and alternative hypotheses:

* Null hypothesis (H0): The salary of a data analyst is less than or equal to 2000$/month.
* Alternative hypothesis (H1): the salary of a data analyst is greater than 2000$/month.

1. **Type I error  and Type II error**

Type I error: H0 is correct but we reject it.

Type II error: H0 is incorrect but we fail to reject it.

For example: when a man go to see the doctor for a Covid test, he still think he is negative with Covid. So that we will have:

* H0 (null hypothesis): he is positive.
* H1 (alternative hypothesis): he is negative.

There are 4 different cases that can happen:

Case 1: The man is positive with Covid and the doctor concluded that he is positive

Case 2: The man is negative with Covid and the doctor concluded that he is negative

Case 3: The man is positive with Covid and the doctor concluded that he is negative

Case 4: The man is negative with Covid and the doctor concluded that he is positive

Case 3 is type I error and case 4 is type II error. The probability that type I error will occur is α, and the probability that type II error will occur is .

1. **α and p-value**

α (significance level) is the probability that you will make the mistake of rejecting the null hypothesis when in fact it is true. In other words, α is the probability that type I error will occur. α is the threshold we use to determine whether or not our p-value is low enough to reject the null hypothesis. It is often set at 0.05 but it is sometimes set as low as 0.01 or as high as 0.10.

The p-value a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

There are two cases between α and p-value:

* If p-value α : we reject the null hypothesis.
* If p-value α : we fail to reject the null hypothesis.
  1. MANIPULATE

1. **Question 1 for Example 7.5**

Arcording to the given data, we have:

*H0: mean response time ≥ 25*

*H1: mean response time < 25*

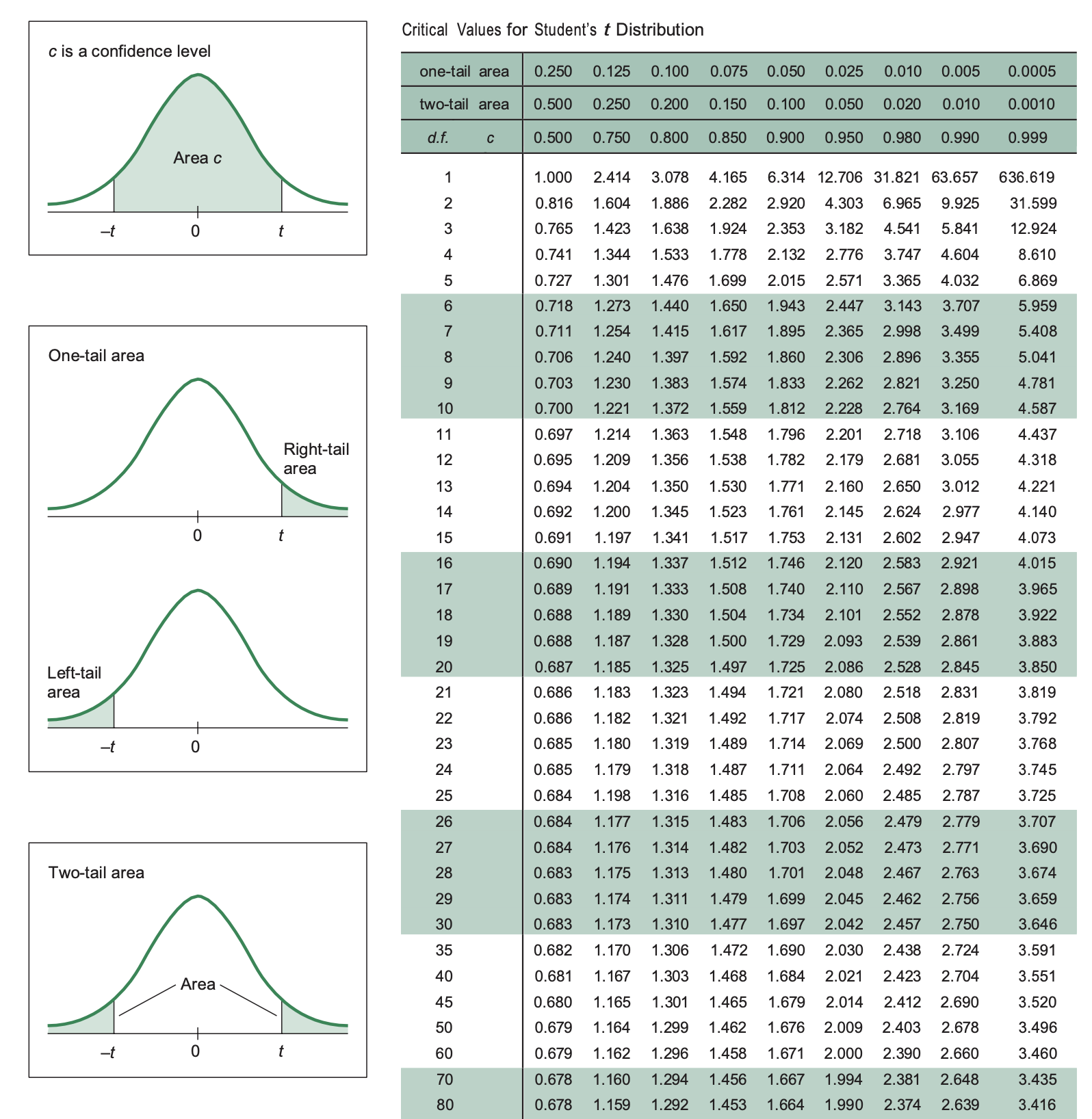
And the following values:

*α* = 0,05

*n = 44 🡪 df (degree of freedom) = n - 1= 43*

*t = -1,05*

This is a left-sided test beacause the alternative hypothesis contains the “<” sign. With significance level α, we reject hypothesis H0 if t = -1,05 fall in rejection region. The rejection region is defined by R = (-∞; -), where the percentile is found from the Student’s t distribution table (for one-tail area).



Searching from Student distribution table, we define = = = 1,684

So, rejection region R = (-∞; -1,684). It’s obvious that t = -1,05 doesn’t fall in rejection region.

A picture containing histogram

Description automatically generated

We ***fail to reject H0*** and conclude the sample data does not support *H1.*

1. **Question 1 for  Example 7.15, 7.16, 7.17**

To test the independence between two variables, we rely on two types of values: the expected frequencies (think/hypothesize) and the observed frequencies (see/actual) to calculate Chi-square (χ2). If χ2 > χ2α,df then we reject H0.

We do the following steps:

**Step 1: Hypothesize:**

*Null hypothesis H0:*gender and brand preference for energy drinks are **independent**

*Alternative hypothesis H1*: gender and brand preference for energy drinks are **dependent**

**Step 2: Define and for each cell**

is observed frequency and is expected frequency.

* *Table of for each cell*

*Table

Description automatically generated*

* *Table of for each cell*

for each cell is calculated according to the following formula:

With Ri is grand total at row i and Cj is sum grand total at column j of that cell.

For example:

*Graphical user interface, application, table, Excel

Description automatically generated*  at cell F6 = = =12,58.

Applying the same to the remaining cells, we get the following table of for each cell:

*Chart

Description automatically generated*

**Step 3: Define total of χ2**

χ2 is calculated according to the formula:

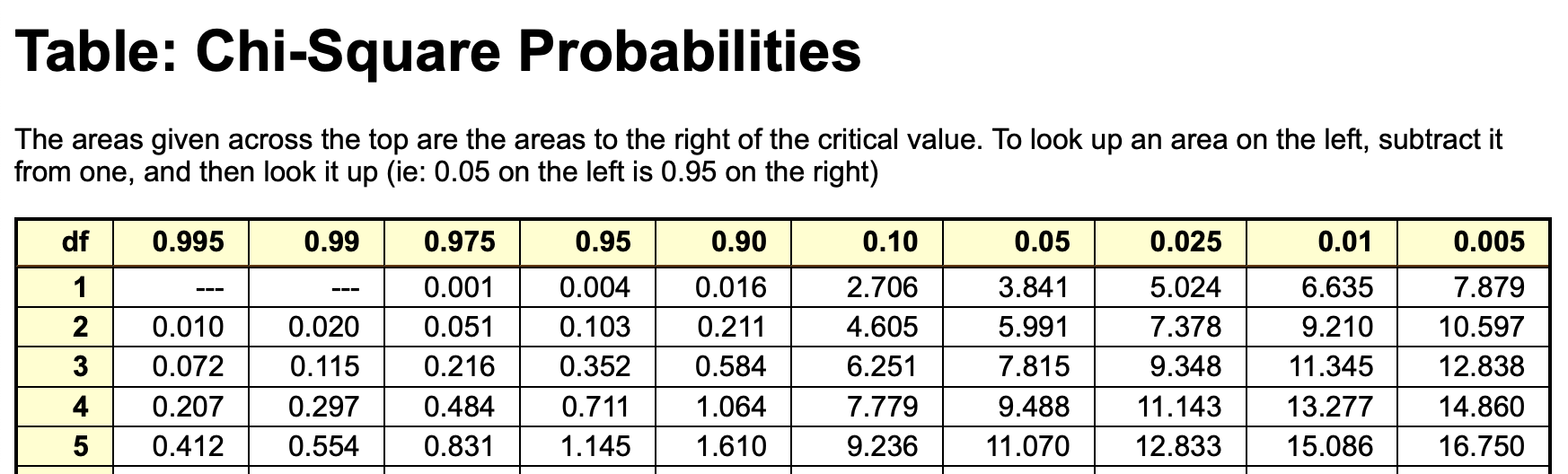
🡪 χ2 = + + + + + = **6,4924**

**Step 4: Define χ2α,df**

We have 3 brands 🡪 n = 3 🡪 df = 2

and α = 0,05

Using Chi-square table, we get the value of χ2α,df = χ20.05, 2 = **5,991**



**Step 5: Making decision**

*Because* χ2 > χ2α,df  (6,4924 > 5,991) 🡪 Reject H0.

1. REFERENCE
2. 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>
3. 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>
4. Nguyễn Minh Nhựt, “PYTHON\_LAB1 – Jupyter Notebook”. [Online]. Link: <https://bit.ly/3Cr82JZ> [Last updated: 28/02/2022].
5. Nguyễn Minh Nhựt, “HƯỚNG DẪN THỰC HÀNH LAB01”. [Online]. Link: <https://bit.ly/3Kq577h> [Last updated: 27/02/2022].
6. chucnv (chuc1803@gmail.com), “Kiểm định giả thuyết (Hypothesis Testing)”. [Online]. Link: <http://bis.net.vn/forums/t/2067.aspx> [Last updated: 07/08/2020].
7. “Critical Value for Student’s t Distribution”.[Online]. Link: <https://www.brockport.edu/academics/tutoring/docs/students_t_distribution_table.pdf>
8. “Table: Chi-Square Probability”. [Online]. Link: <https://people.richland.edu/james/lecture/m170/tbl-chi.html>