VIETNAM NATIONAL UNIVERSITY

UNIVERSITY OF INFORMATION TECHNOLOGY

INFORMATION SYSTEMS FACULTY

REPORT LAB1

SUBJECT: DATA ANALYSIS IN BUSINESS

Lecturer: Assoc. Prof. Nguyen Dinh Thuan

Instructor: TA. Nguyen Minh Nhut

Class: IS403.O22.HTCL

Group 3:

21521049 – Ho Quang Lam

21521586 – Le Thi Le Truc

21521938 – Nguyen Thanh Dat

Ho Chi Minh City, March 2024

ACKNOWLEDGEMENT

First of all, we would like to express our deepest gratitude and appreciation to our lecturers, Mr. Nguyen Dinh Thuan and Mr. Nguyen Minh Nhut, for their teaching and sharing of extensive knowledge as well as practical examples during the lectures. They have guided us in completing our Lab 01 report by providing valuable feedback, suggestions, and assistance with exercises and revisions.

The Data Analysis in Business course is an interesting and highly practical subject. However, due to our limited expertise and initial unfamiliarity with real-world applications, we acknowledge that our Lab 01 report may contain some shortcomings and inaccuracies despite our best efforts. We sincerely hope to receive further guidance and feedback from Mr. Nguyen Dinh Thuan and Mr. Nguyen Minh Nhut to improve our knowledge and equip ourselves for future projects as well as for our academic and professional endeavors.

Once again, we would like to extend our heartfelt and sincere gratitude to our lecturers and peers.

Ho Chi Minh City, March 2024

Group of student performers

Ho Quang Lam

Le Thi Le Truc

Nguyen Thanh Dat

TABLE OF CONTENTS

LECTURER'S COMMENTS

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

WORK DISTRIBUTION

Members / Works	Le Thi Le Truc (Leader)	Ho Quang Lam	Nguyen Thanh Dat
Problem statement	✓	✓	✓
Build the report template	✓		
Do all exercise with Excel	✓		
Do all exercise with Python		✓	
Do all exercise with R			✓
Summarize and edit reports	✓	✓	✓
Completion	100%	100%	100%

LIST OF IMAGES

Chapter 1. MEANING OF VALUES

Meaning of the values: Count, Min, Max, Mean, Median, Mode, Quantile, Range, Mode, Variance, Standard Deviation, Coefficient of Deviation, Skewness, Kurtosis.

- **Count:** the number of observations. **COUNT (data range)**
- **Min:** the smallest numeric value in a series of observations. **MIN (data range)**
- **Max:** the largest numeric value in a series of observations. **MAX (data range)**
- **Mean:** is the ratio of the sum of all observations in the data to the total number of observations. This is also known as average. Thus, mean is a number around which the entire data set is spread. **AVERAGE (data range)**
- **Median:** middle value of the data when arranged from least to greatest. **MEDIAN (data range)**
- **Quantile:** is a quantity that describes the distribution and dispersion of data.
  - Min value: **QUARTILE.INC (data range, 0)**
  - First quartile: **QUARTILE.INC (data range, 1)**
  - Median value: **QUARTILE.INC (data range, 2)**
  - Third quartile: **QUARTILE.INC (data range, 3)**
  - Max value: **QUARTILE.INC (data range, 4)**

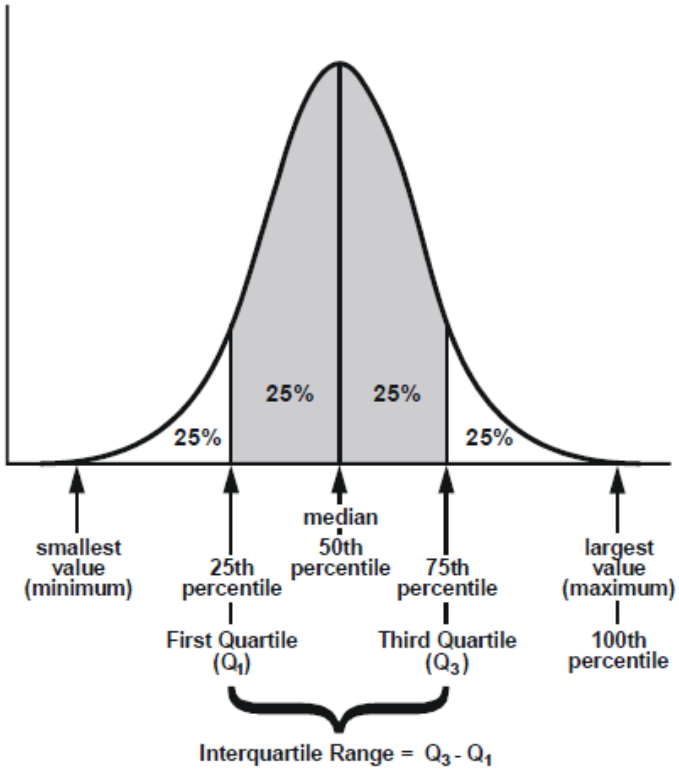


Image 1.1 Inter Quartile Range (IQR)

- **Range:** is the difference between the maximum and minimum data values.

RANGE = MAX(data range) - MIN(data range)

- **Mode:** observation that occurs most often or, for grouped data, the group with the greatest frequency. **MODE.SNGL(data range)**
- **Variance:** is an average of the squared deviations form the mean (uses all data values). **VAR.S(data range)**
- **Standard Deviation:** is the square root of the variance. **STDE.S(data range)**
- **Coefficient of Deviation:** is the ratio of the standard deviation to the mean and shows the extent of variability about the mean of the population.

STDE.S(data range)/AVERAGE(data range)

- **Skewness:** The measure of asymmetry in a probability distribution is defined by skewness. Skewness can either be positive, negative or undefined.
    - Positive Skew: This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, the mean is greater than the mode.

- Negative Skew: This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, the mean is smaller than the mode.

- **Kurtosis:** used to evaluate the peak of the observation curve with the form of a standard distribution curve. **KURT (data, range)**
  - Kurtosis = 3: the distribution is Mesokurtic: the normal distribution.
  - Kurtosis > 3: the distribution is Leptokurtic: the distribution has fatter tails and a sharper peak.
  - Kurtosis < 3: Platykurtic distribution: a lower and broader peak and thinner Tail

Chapter 2. GDP OF VIETNAM

2.1. Using MS Excel, R and Python programming language

Lesson 1A. *Using Data Visualization and Descriptive Statistics. With the dataset: GDP of Viet Nam (2020 – 2022)*

a) Using MS Excel, R Programming Language, and Python Programming Language, calculate and analyze the meaning of the values: Mean, Median, Mode, Quantile, Variance, Standard Deviation, Skewness, and Kurtosis correspond to the above data sets.

b) Use Visualization: Histogram, Box Plot to visualize some of the above values.

2.1.1. Analyzing by using MS Excel

2.1.1.1. Statistical description

STEP 1: Choose the **"Data"** tab on the top of the Excel Toolbar.

Click **"Data Analysis"** function.

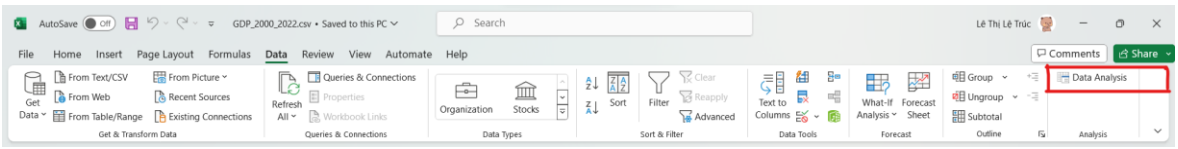


Image 2.1 STEP 1 of analyzing by using MS Excel with the dataset GDP

STEP 2: In the window, choose **"Descriptive Statistics"** in the box and then click **"OK"** button

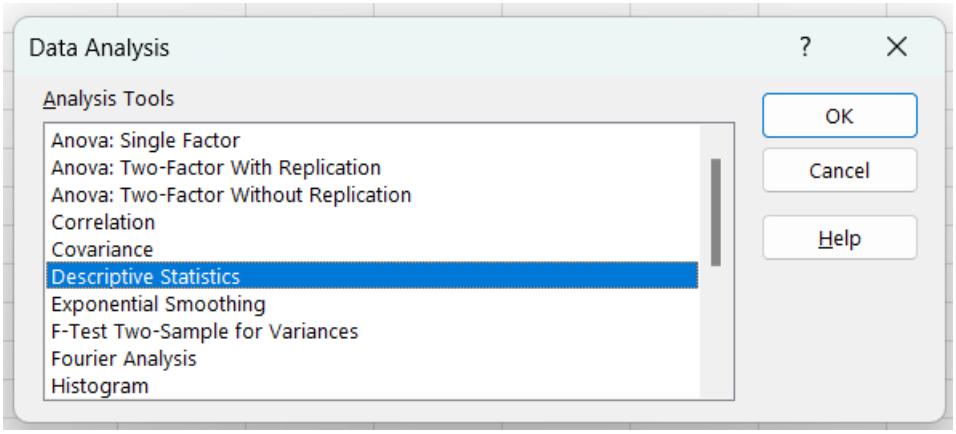


Image 2.2 STEP 2 of analyzing by using MS Excel with the dataset GDP

STEP 3: Enter Input range and data range you want to calculate **($B$2:$B$22)** and choose columns in the (Grouped by) and tick **"Summary statistics"** in the "Output options". Finally, click **"OK"** button.

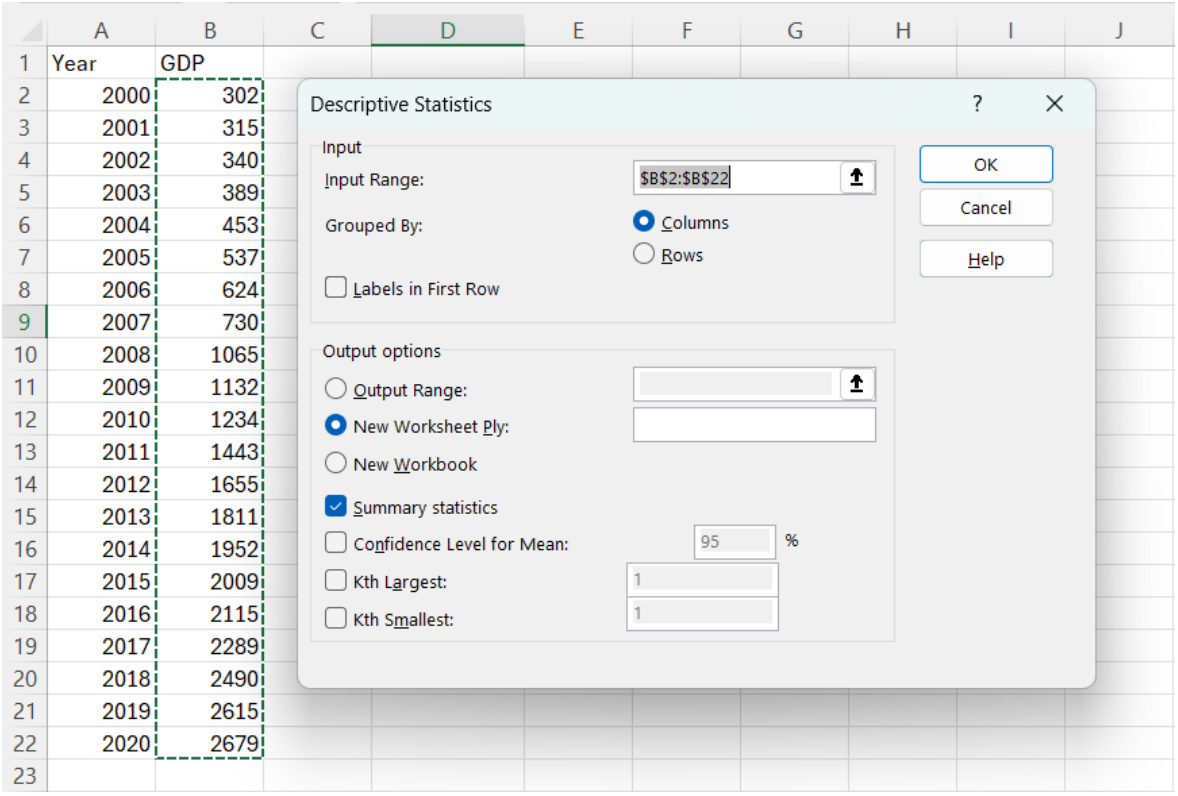


Image 2.3 STEP 3 of analyzing by using MS Excel with the dataset GDP

STEP 4: After finish step 3, the result will be displayed like the picture below.

	A	B	C
1	*GDP OF VIETNAM*		
2			
3	Mean	1341.857143	
4	Standard Error	181.7566962	
5	Median	1234	
6	Mode	#N/A	
7	Standard Deviatio	832.9138182	
8	Sample Variance	693745.4286	
9	Kurtosis	-1.447291133	
10	Skewness	0.191905264	
11	Range	2377	
12	Minimum	302	
13	Maximum	2679	
14	Sum	28179	
15	Count	21	
16			

Image 2.4 STEP 4 of analyzing by using MS Excel with the dataset GDP

2.1.1.2. Recalculating using MS Excel

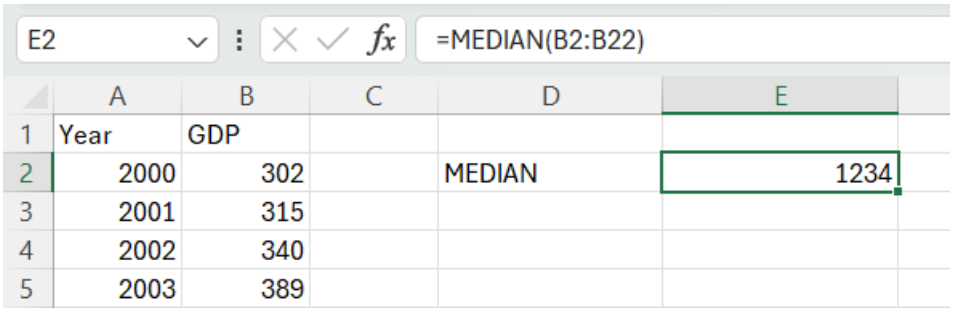
❖ ***Mean of GDP:*** The Average value of GDP

E2 =AVERAGE(B2:B22)

	A	B	C	D	E	F
1	Year	GDP				
2	2000	302		MEAN	1341.857143	
3	2001	315				

Image 2.5 Mean of GDP – MS Excel

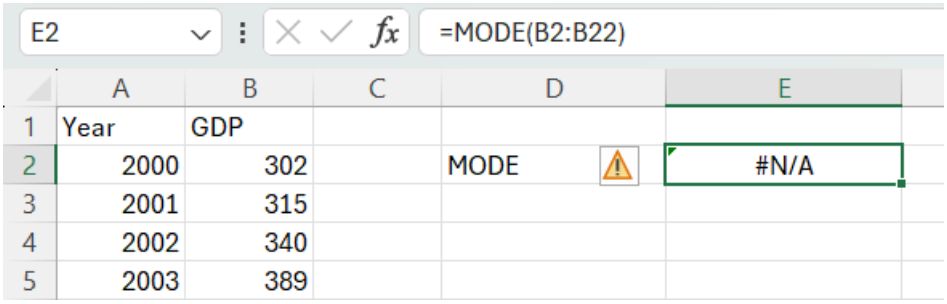
❖ ***Median of GDP:*** The median of GDP



	A	B	C	D	E
1	Year	GDP			
2	2000	302		MEDIAN	1234
3	2001	315			
4	2002	340			
5	2003	389			

Image 2.6 Median of GDP – MS Excel

❖ ***Mode of GDP:***

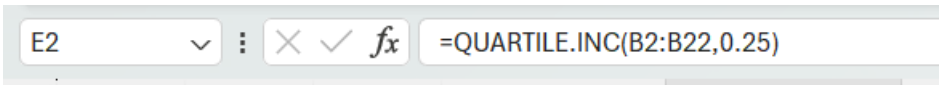


	A	B	C	D	E
1	Year	GDP			
2	2000	302		MODE	#N/A
3	2001	315			
4	2002	340			
5	2003	389			

Image 2.7 Mode of GDP – MS Excel

Explain the result: We encountered an error when using the MODE function to calculate the data in the sheet. The error occurred because there were no duplicate GDP values present, preventing the calculation of the mode value. As a result, the MODE function returned an error. This indicates that the process was successful

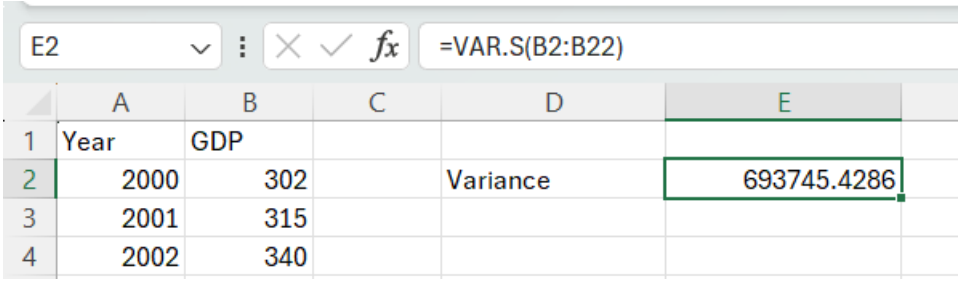
❖ ***Quantile of GDP:***



	A	B	C	D	E
1	Year	GDP			
2	2000	302		QUARTILE 0	302
3	2001	315		QUARTILE 1	537
4	2002	340		QUARTILE 2	1234
5	2003	389		QUARTILE 3	2009
6	2004	453		QUARTILE 4	2679
7	2005	537			

Image 2.8 Quantile of GDP – MS Excel

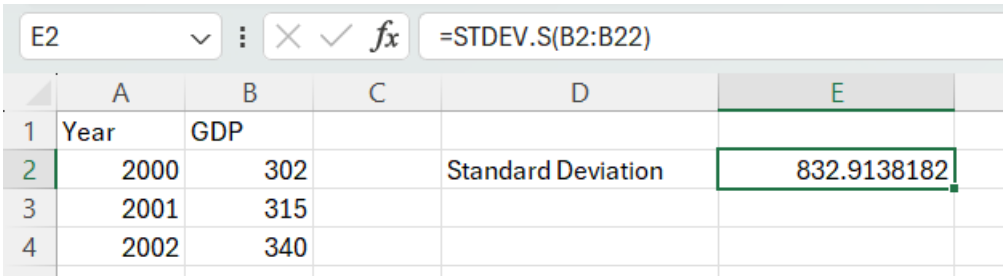
❖ ***Variance of GDP:***



	A	B	C	D	E
1	Year	GDP			
2	2000	302		Variance	693745.4286
3	2001	315			
4	2002	340			

Image 2.9 Variance of GDP – MS Excel

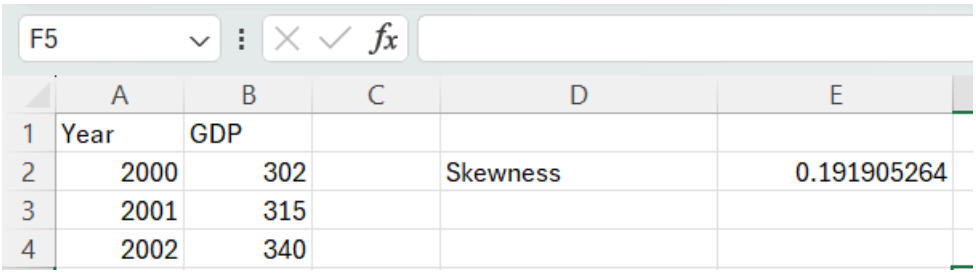
❖ ***Standard Deviation of GDP:***



	A	B	C	D	E
1	Year	GDP			
2	2000	302		Standard Deviation	832.9138182
3	2001	315			
4	2002	340			

Image 2.10 Standard Deviation of GDP – MS Excel

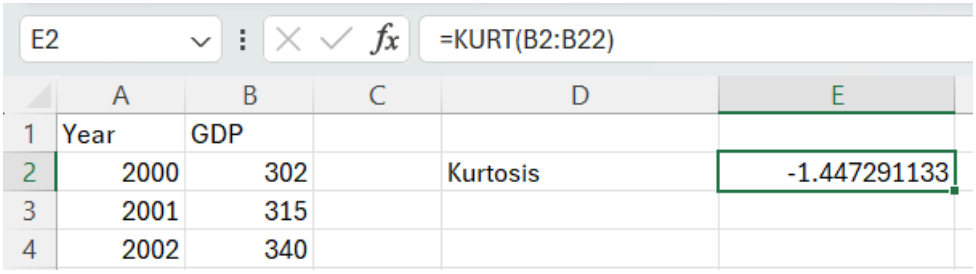
❖ ***Skewness of GDP:***



	A	B	C	D	E
1	Year	GDP			
2	2000	302		Skewness	0.191905264
3	2001	315			
4	2002	340			

Image 2.11 Skewness of GDP – MS Excel

❖ ***Kurtosis of GDP:***



	A	B	C	D	E
1	Year	GDP			
2	2000	302		Kurtosis	-1.447291133
3	2001	315			
4	2002	340			

Image 2.12 Kurtosis of GDP – MS Excel

2.1.2. Analyzing by using R

2.1.2.1. Statistical description:

Firstly, import the data and store data in a data frame

```
> gdp<-read.csv(file.choose(), header = TRUE)
> gdp
   Year  GDP
1  2000  302
2  2001  315
3  2002  340
4  2003  389
5  2004  453
6  2005  537
7  2006  624
8  2007  730
9  2008 1065
10 2009 1132
11 2010 1234
12 2011 1443
13 2012 1655
14 2013 1811
15 2014 1952
16 2015 2009
17 2016 2115
18 2017 2289
19 2018 2490
20 2019 2615
21 2020 2679
```

Image 2.13 Import dataset GDP OF VN - R

Attach the database to the R search path.

```
> attach(gdp)
```

2.1.2.2. Meaning of values:

❖ ***Mean of GDP:***

```
> mean(GDP)
[1] 1341.857
```

Image 2.14 Mean of GDP - R

❖ ***Mode of GDP:***

```
> getmode <- function(v){
+     uniqv <- unique(v)
+     uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
> getmode(GDP)
[1] 302
```

Image 2.15 Mode of GDP - R

In this case, the result is 302 which is wrong because 302 only appears once, so there is no valid mode result

❖ ***Quantile of GDP:***

```
> quantile(GDP, 0)
 0%
302
> quantile(GDP, 0.25)
25%
537
> quantile(GDP, 0.5)
 50%
1234
> quantile(GDP, 0.75)
 75%
2009
> quantile(GDP, 1)
100%
2679
```

Image 2.16 Quantile of GDP – R

❖ ***Variane of GDP:***

```
> var(GDP)
[1] 693745.4
```

Image 2.17 Variane of GDP - R

❖ ***Standard Deviation of GDP:***

```
> sd(GDP)
[1] 832.9138
```

Image 2.18 Standard Deviation of GDP - R

❖ ***Skewness of GDP:***

Before calculating the skewness and the kurtosis, we have to install the library "e1071" in order to have the most exactly result

```
> install.packages("e1071")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.2/e1071_1.7-14.tgz'
Content type 'application/x-gzip' length 673162 bytes (657 KB)
==================================================
downloaded 657 KB


The downloaded binary packages are in
        /var/folders/7_/dm3mjc6j0n128zlj27rkn9f80000gn/T//RtmpQ0Vjl7/downloaded_packages
```

Image 2.19 Install the library "e1071"

```
> library(e1071)
```

Image 2.20 Call the library

```
> skewness(GDP)
[1] 0.1653605
```

Image 2.21 Skewness of GDP – R

❖ ***Kurtosis of GDP:***

```
> kurtosis(GDP)
[1] -1.546637
```

Image 2.22 Kurtosis of GDP – R

2.1.3. Analyzing by using Python

2.1.3.1. Statistical description:

STEP 1: Import Modules

```
[7]  import numpy as np
     import pandas as pd
     import math
     import statistics as st
     import matplotlib.pyplot as plt
     import seaborn as sns
     import plotly.express as px
     from scipy.stats import skew
     from scipy.stats import kurtosis
```

Image 2.23 Import modules – GDP – Python

STEP 2: Read the dataset

```
[9]  gdp_lab=pd.read_csv("GDP_2000_2022.csv")
     gdp_lab.head(5)
```

	Year	GDP
0	2000	302.0
1	2001	315.0
2	2002	340.0
3	2003	389.0
4	2004	453.0

Image 2.24 Read the dataset – GDP – Python

STEP 3: Rename

```
#To facilitate easy referencing of the value, assign a new name to the column section.
gdp=gdp_lab.GDP
```

Image 2.25 Rename – GDP – Python

2.1.3.2. Meaning of values:

❖ ***Mean of GDP:***

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

```
#By employing the conventional approach (using a formula)
mean_or=sum(gdp)/len(gdp)
mean_or
```

```
1341.857142857143
```

```
[ ] #using statistic library
    st.mean(gdp)
```

```
1341.857142857143
```

Image 2.26 Mean of GDP - Python

❖ ***Mode of GDP:***

```
#using the ordinary method by coding:

def find_mode(gdp):
    frequency = {}

    # Calculate the occurrence count for each number
    for num in gdp:
        if num in frequency:
            frequency[num] += 1
        else:
            frequency[num] = 1

    # Identify the maximum occurrence frequency
    max_frequency = max(frequency.values())

    # Generate a collection comprising the modal values
    modes = []
    for num, freq in frequency.items():
        if freq == max_frequency:
            modes.append(num)

    return modes

mode_or = find_mode(gdp)
mode_or
```

```
[302.0,
 315.0,
 340.0,
 389.0,
 453.0,
```

Image 2.27 Mode of GDP – Python – Method 1

```
[ ]  #using library:
     st.mode(gdp)
```

```
302.0
```

Image 2.28 Mode of GDP – Python – Method 2

Explanation: Due to the absence of duplicate values, the algorithm is unable to determine the mode. In the case of using a statistical library, it will return the first value in the dataset. Alternatively, when using a regular approach, it will return the entire dataset as the mode.

❖ ***Quantile of GDP:***

```
[ ]  q1 = np.quantile(gdp, 0.25)
     q2 = np.quantile(gdp, 0.5)
     q3 = np.quantile(gdp, 0.75)
     print(q1)
     print(q2)
     print(q3)
```

```
537.0
1234.0
2009.0
```

Image 2.29 Quantile of GDP - Python

❖ ***Variance of GDP:***

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Image 2.30 Variance calculation formula

```
[ ]  #using ordinary method (formula)
     var_or = sum((x - mean_or) ** 2 for x in gdp) // (n-1)
     var_or
```

```
693745.0
```

Image 2.31 Variance of GDP – Python – Method 1

```
[ ]  var=st.variance(gdp)
     var

     693745.4285714285
```

Image 2.32 Variance of GDP – Python – Method 2

❖ ***Standard Deviation of GDP:***

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Image 2.33 Standard Deviation calculation formula

```
[ ]  #using basic math library
     stdev_or=math.sqrt(var)
     stdev_or

     832.9138182137624
```

Image 2.34 Standard Deviation of GDP – Python – Method 1

```
[ ]  #using tool
     st.stdev(gdp)

     832.9138182137625
```

Image 2.35 Standard Deviation of GDP – Python – Method 2

❖ ***Skewness of GDP:***

$$skew[x] = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Image 2.36 Skewness calculation formula

```
#using the formula:
skew_or=(n / ((n - 1) * (n - 2))) * sum(((x - mean_or) / stdev_or) ** 3 for x in gdp)
skew_or
```

```
0.19190526429277793
```

Image 2.37 Skewness of GDP – Python – Method 1

```
[ ]  #using library
     skew(gdp,bias=False)
```

```
0.1919052642927779
```

Image 2.38 Skewness of GDP – Python – Method 2

❖ ***Kurtosis of GDP:***

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}$$

Image 2.39 Kurtosis calculation formula

```
[ ] kurt_or = kurt = (n * (n + 1) / ((n - 1) * (n - 2) * (n - 3))) * sum(((x - mean_or) / stdev_or) ** 4 for x in gdp) - 3 * ((n - 1) ** 2) / ((n - 2) * (n - 3))
    kurt_or
```

```
-1.4472911328899465
```

Image 2.40 Kurtosis of GDP – Python – Method 1

```
[ ]  kurtosis(gdp)
```

```
-1.3976671987462779
```

Image 2.41 Kurtosis of GDP – Python – Method 2

2.2. Data Visualization

2.2.1. Using MS Excel

2.2.1.1. Histogram

STEP 1: Choose the **"Data"** tab on the top of the Excel Toolbar.

Click **"Data Analysis"** function.

Click **"Histogram"** and **"OK"** button

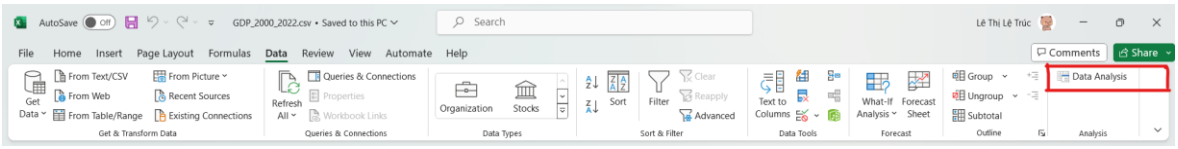


Image 2.42 STEP 1 of drawing Histogram – GDP – MS Excel – Choose "Data Analysis"

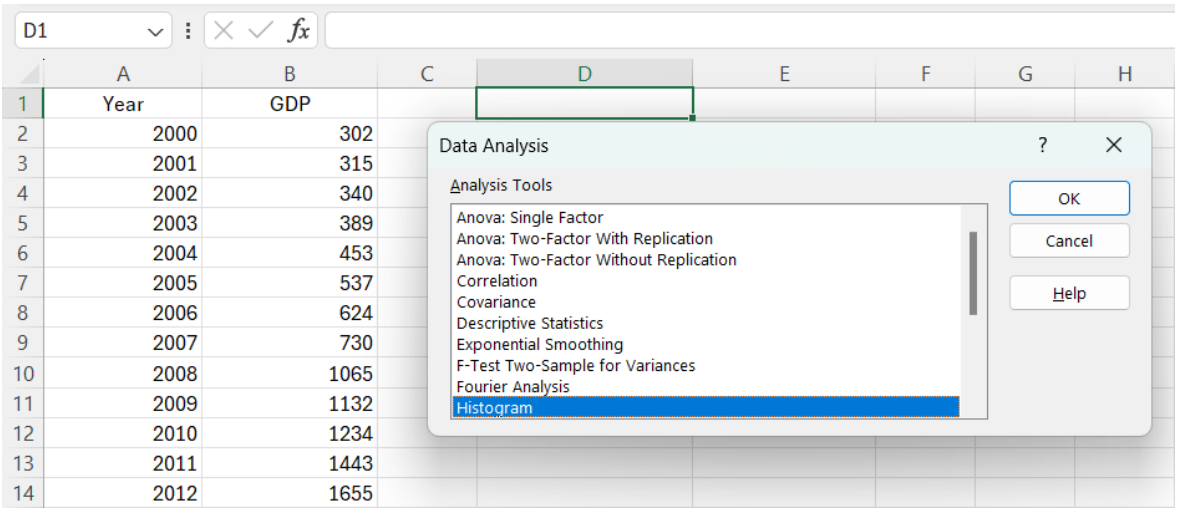


Image 2.43 STEP 1 of drawing Histogram – GDP – MS Excel – Click "Histogram"

STEP 2: Enter Input Range and select **"Cumulative Percentage"** and **"Chart Output".**

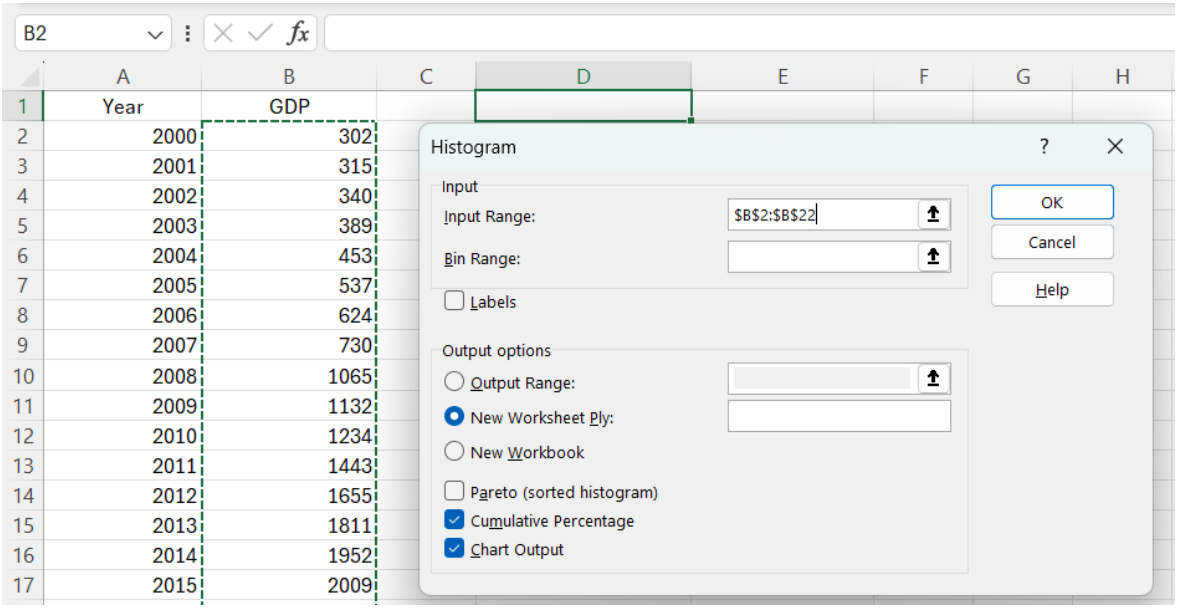


Image 2.44 STEP 2 of drawing Histogram – GDP – MS Excel

STEP 3: After finish step 2, the result will be displayed like the picture below.

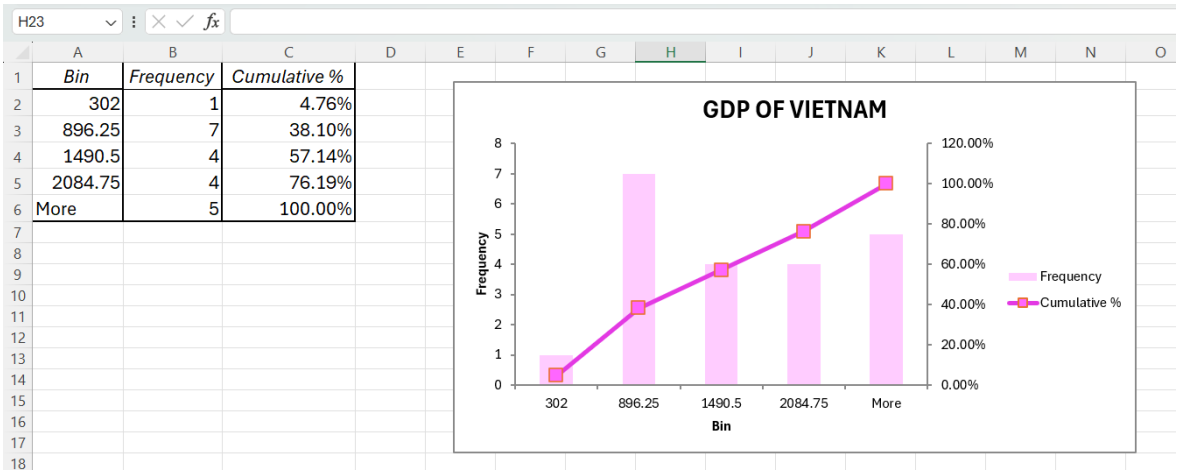


Image 2.45 STEP 3 of drawing Histogram – GDP – MS Excel

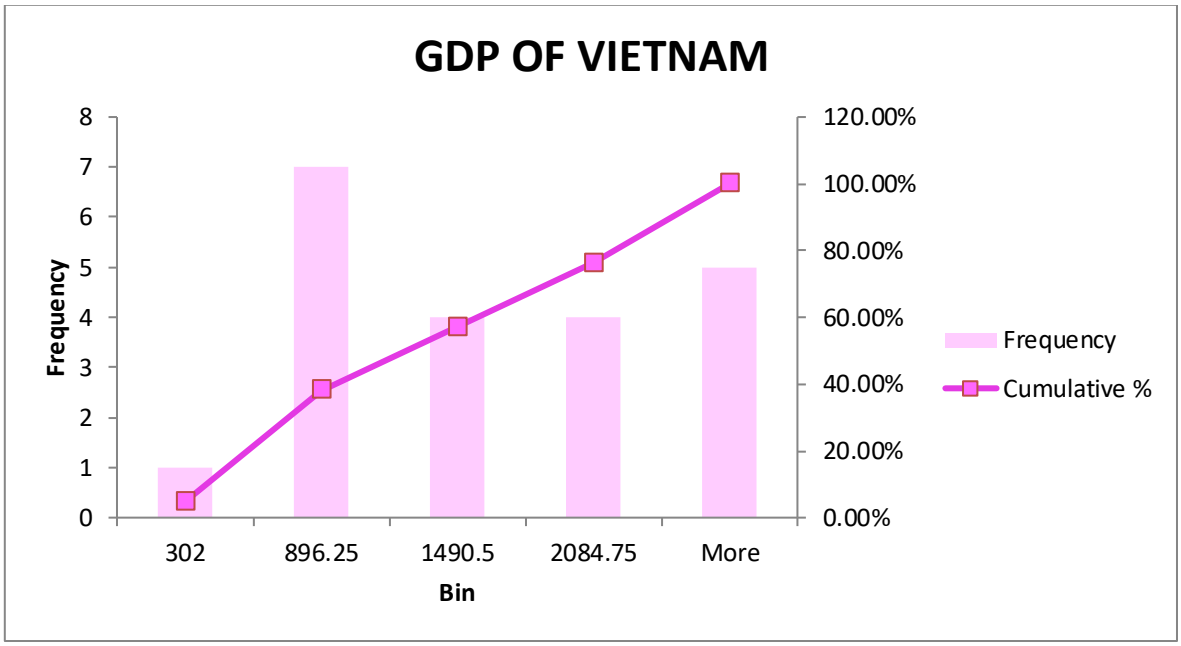


Image 2.46 Histogram of VietNam's GDP – MS Excel

2.2.1.2. Boxplot:

STEP 1: Choose the **"Insert"** tab on the top of the Excel Toolbar.

Click **"Statistic Chart"** symbol and **"Box and Whisker"**

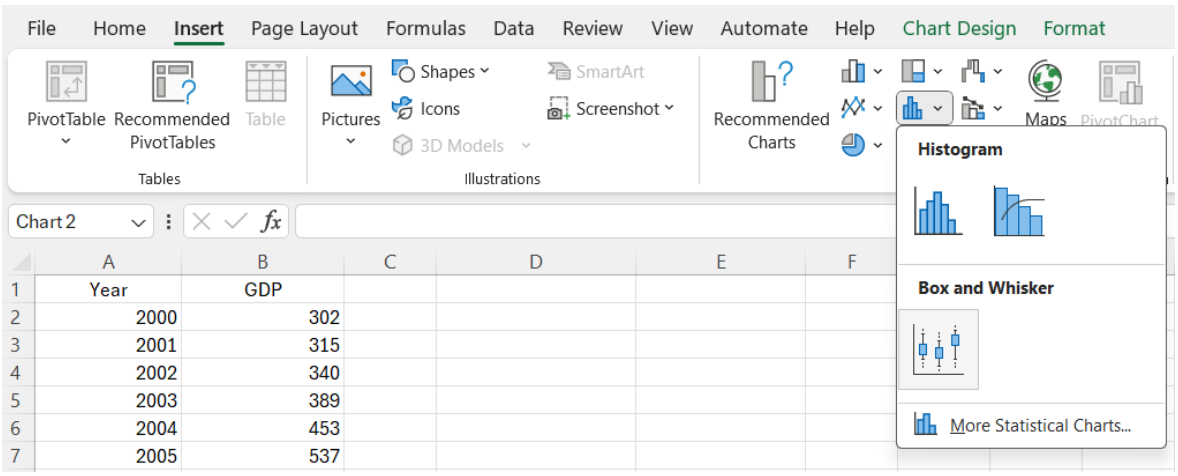
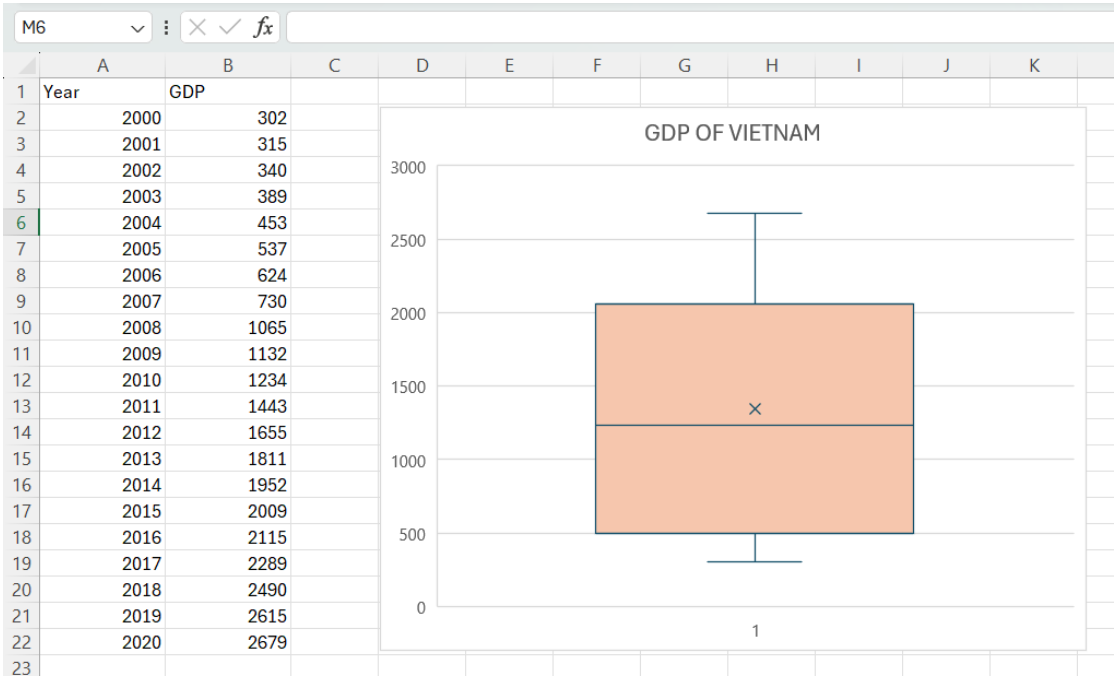


Image 2.47 STEP 1 of drawing Boxplot – GDP – MS Excel

STEP 2: After finish step 1, the result will be displayed like the picture below.



Year	GDP
2000	302
2001	315
2002	340
2003	389
2004	453
2005	537
2006	624
2007	730
2008	1065
2009	1132
2010	1234
2011	1443
2012	1655
2013	1811
2014	1952
2015	2009
2016	2115
2017	2289
2018	2490
2019	2615
2020	2679

Image 2.48 STEP 2 of drawing Boxplot – GDP – MS Excel

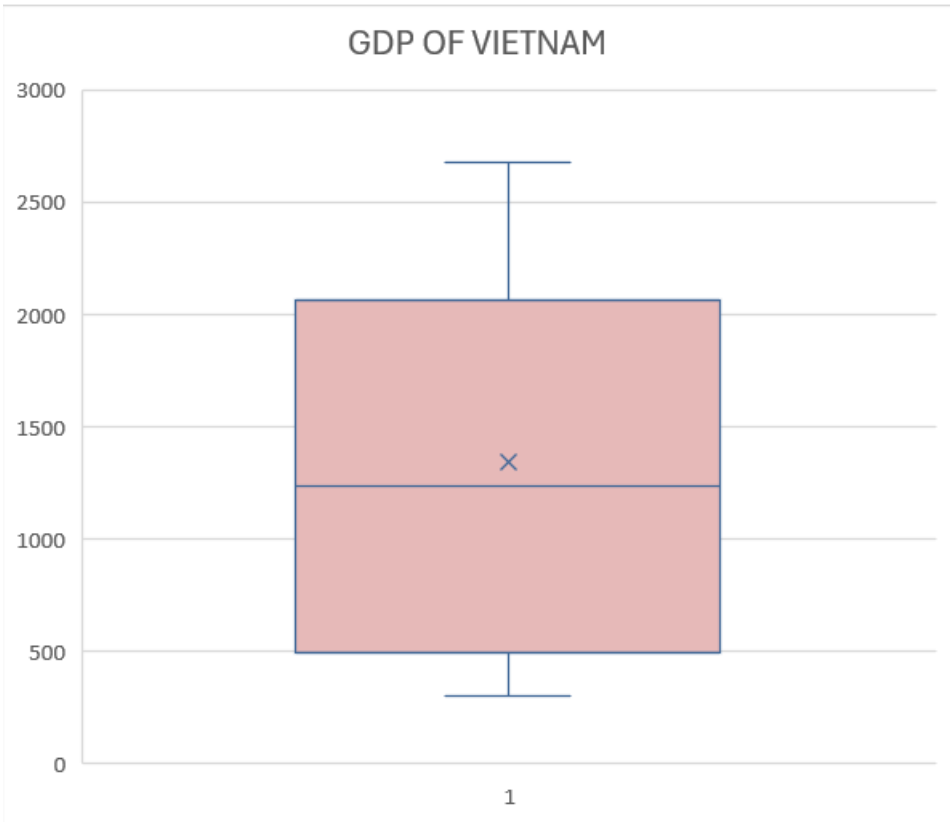


Image 2.49 Boxplot of VietNam's GDP – MS Excel

2.2.2. Using R

2.2.2.1. Histogram

```
> hist(GDP, main = "Histogram GDP", ylab = "USD", col = "pink")
```

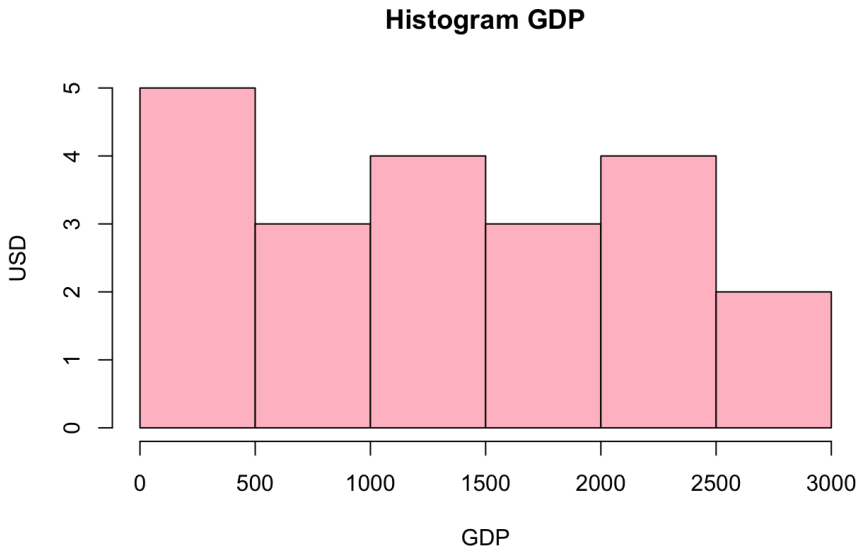


Image 2.50 Histogram of VietNam's GDP – R

2.2.2.2. Boxplot

```
> boxplot(GDP, main = "Boxplot GDP", ylab = "USD", col = "pink")
```

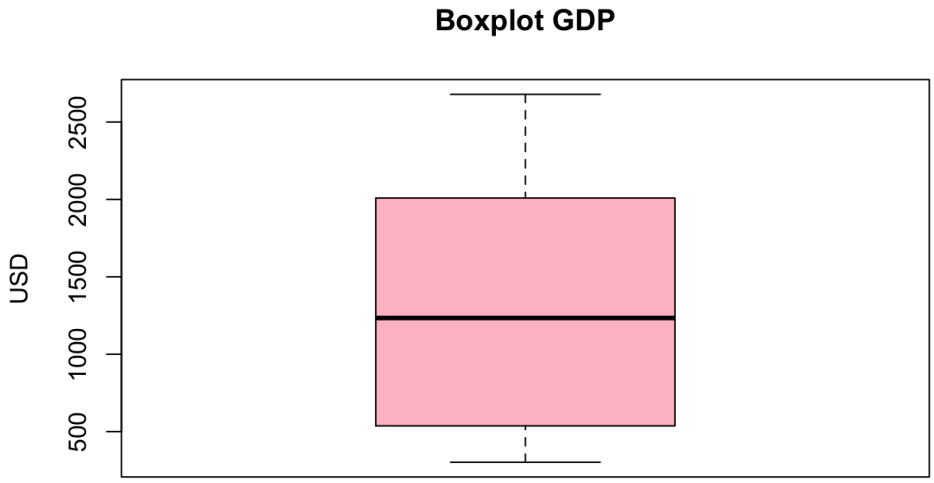


Image 2.51 Boxplot of VietNam's GDP – R

2.2.3. Using Python

2.2.3.1. Histogramm

```
plt.hist(gdp, bins=10, alpha=0.5, color='pink', edgecolor='black')
plt.title("Histogram of VietNam's GDP")
plt.show()
```

✓ 0.1s

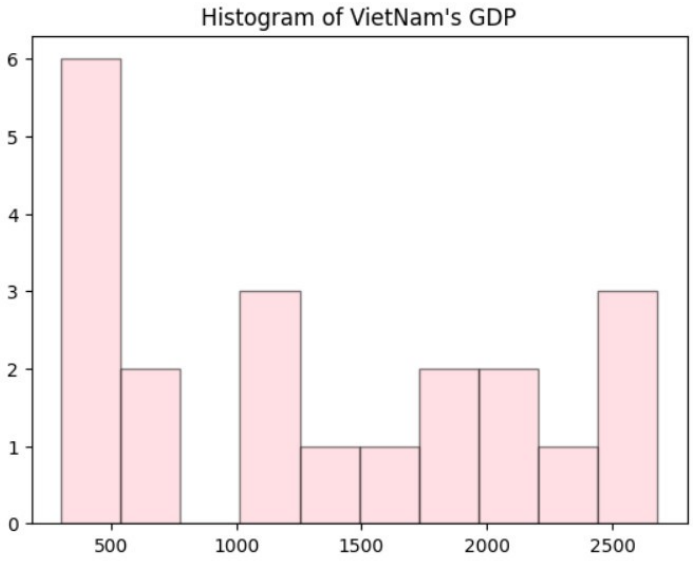


Image 2.52 Histogram of VietNam's GDP - Python

2.2.3.2. Boxplot

```
#using seaborn
sns.set(style="whitegrid")
sns.despine(top=True,
            right=True,
            left=True,
            bottom=False)
fig, ax=plt.subplots(figsize=(6,3))
sns.boxplot(gdp_lab['GDP'],color="lavender",width=0.5)
plt.title("Boxplot of GDP",fontsize=14)
plt.xlabel("GDP")
plt.show()
```

```
<Figure size 640x480 with 0 Axes>
```

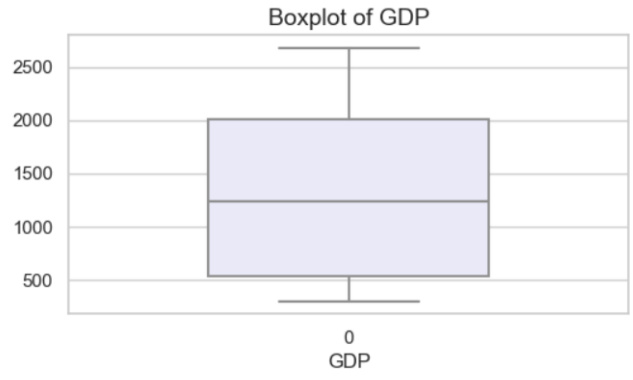


Image 2.53 Boxplot of VietNam's GDP – Python

Chapter 3. COMPUTER REPAIR TIMES

3.1. Using MS Excel, R and Python programming language

Lesson 1A. *Using Data Visualization and Descriptive Statistics. With the dataset: Computer Repair Times*

a) Using MS Excel, R Programming Language, and Python Programming Language, calculate and analyze the meaning of the values: Mean, Median, Mode, Quantile, Variance, Standard Deviation, Skewness, and Kurtosis correspond to the above data sets.

b) Use Visualization: Histogram, Box Plot to visualize some of the above values.

3.1.1. Analyzing by using MS Excel

3.1.1.1. Statistical description

STEP 1: Choose the **"Data"** tab on the top of the Excel Toolbar.

Click **"Data Analysis"** function.

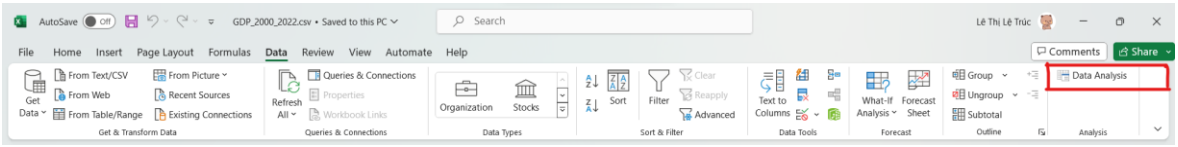


Image 3.1 STEP 1 of analyzing by using MS Excel with the dataset CRP

STEP 2: In the window, choose **"Descriptive Statistics"** in the box and then click **"OK"** button

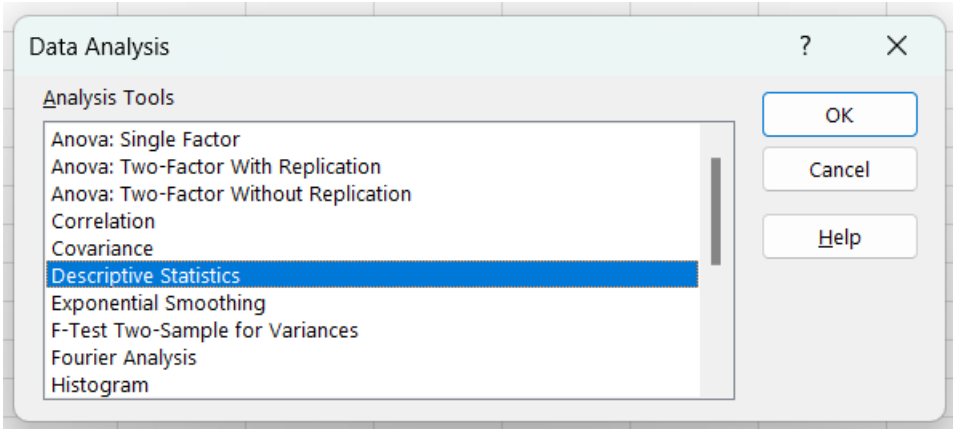


Image 3.2 STEP 2 of analyzing by using MS Excel with the dataset CRP

STEP 3: Enter Input range and data range you want to calculate **($B$4:$B$253)** and choose columns (because the input is in one column) in the (Grouped by) and tick **"Summary statistics"** in the "Output options". Finally, click **"OK"** button.

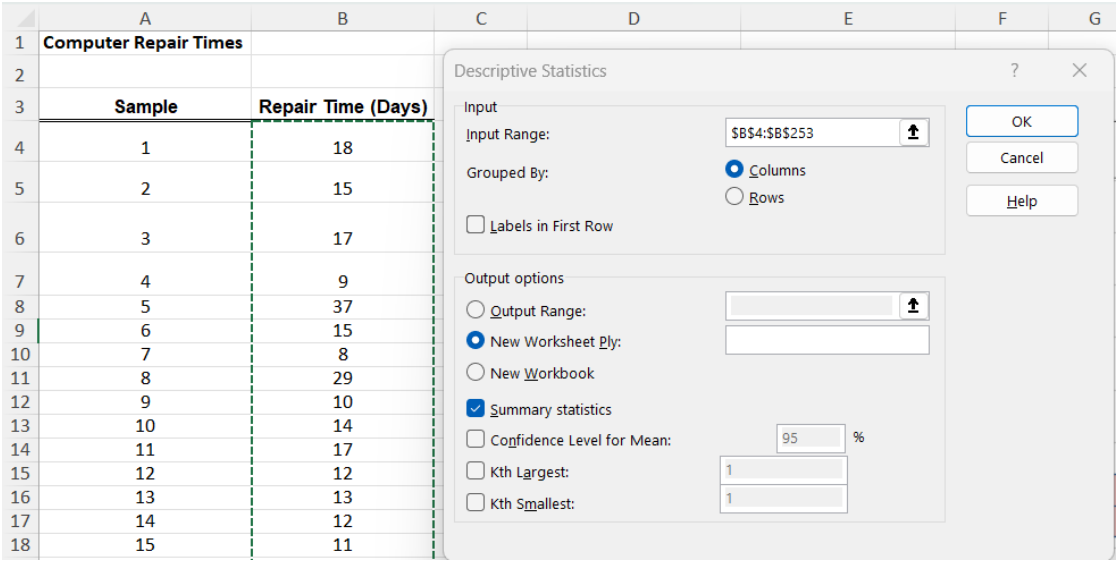


Image 3.3 STEP 3 of analyzing by using MS Excel with the dataset CRP

STEP 4: After step 3, the result will be displayed like the picture below

	A	B
1	*Computer Repair Times*	
2		
3	Mean	14.912
4	Standard Error	0.3768
5	Median	14
6	Mode	15
7	Standard Deviatio	5.9584
8	Sample Variance	35.502
9	Kurtosis	4.079
10	Skewness	1.6953
11	Range	35
12	Minimum	5
13	Maximum	40
14	Sum	3728
15	Count	250

Image 3.4 STEP 4 of analyzing by using MS Excel with the dataset CRT

3.1.1.2. Recalculating using MS Excel

❖ **Mean of CRT:**

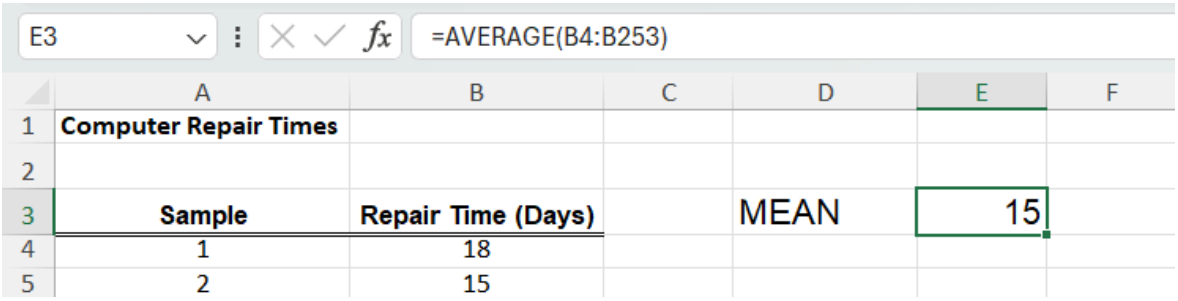


Image 3.5 Mean of CRT – MS Excel

❖ **Median of CRT:**

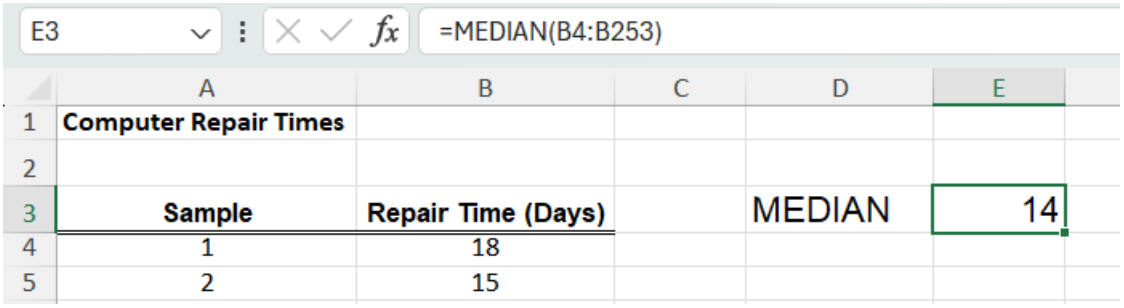


Image 3.6 Median of CRT – MS Excel

❖ ***Mode of CRT:***

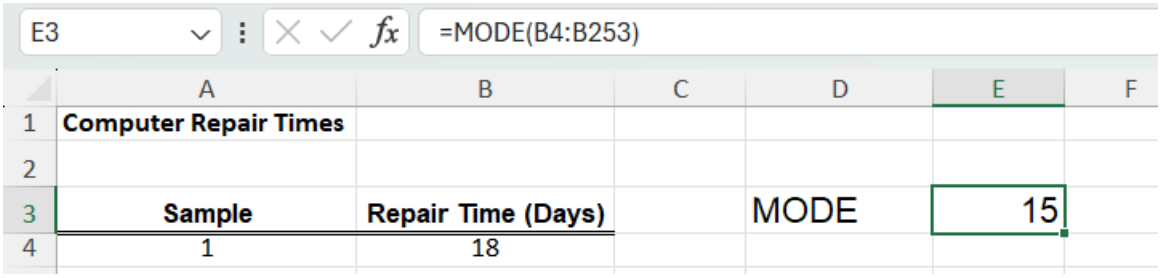


Image 3.7 Mode of CRT – MS Excel

❖ **Quantile of CRT:**

E7 =QUARTILE.INC(B4:B253,4)

	A	B	C	D	E
1	**Computer Repair Times**				
2					
3	**Sample**	**Repair Time (Days)**		QUANTILE 0	5
4	1	18		QUANTILE 1	11
5	2	15		QUANTILE 2	14
6	3	17		QUANTILE 3	17
7	4	9		QUANTILE 4	40
8	5	37			

Image 3.8 Quantile of CRT – MS Excel

❖ **Variance of CRT:**

E4 =VAR.S(B4:B253)

	A	B	C	D	E	F
1	**Computer Repair Times**					
2						
3	**Sample**	**Repair Time (Days)**				
4	1	18		VARIANCE	35.50226506	
5	2	15				

Image 3.9 Variance of CRT – MS Excel

❖ Standard Deviation of CRT:

E4 =STDEV.S(B4:B253)

	A	B	C	D	E
1	**Computer Repair Times**				
2					
3	**Sample**	**Repair Time (Days)**			
4	1	18		Standard Deviation	5.958377721
5	2	15			

Image 3.10 Standard Deviation of CRT – MS Excel

❖ **Skewness of CRT:**

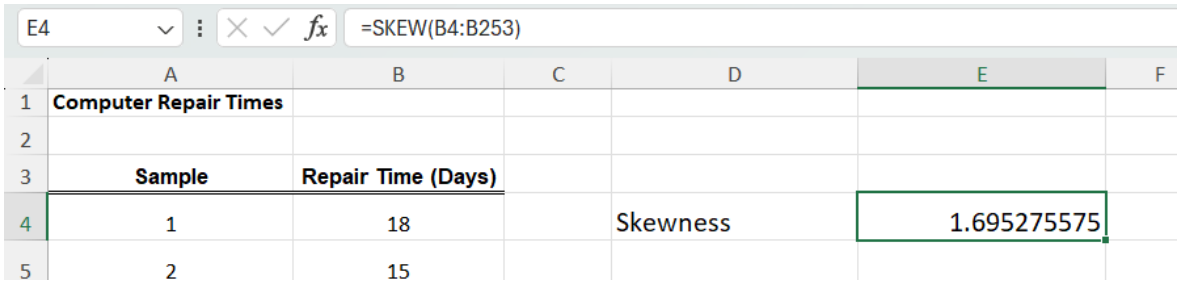


Image 3.11 Skewness of CRT – MS Excel

❖ **Kurtosis of CRT:**

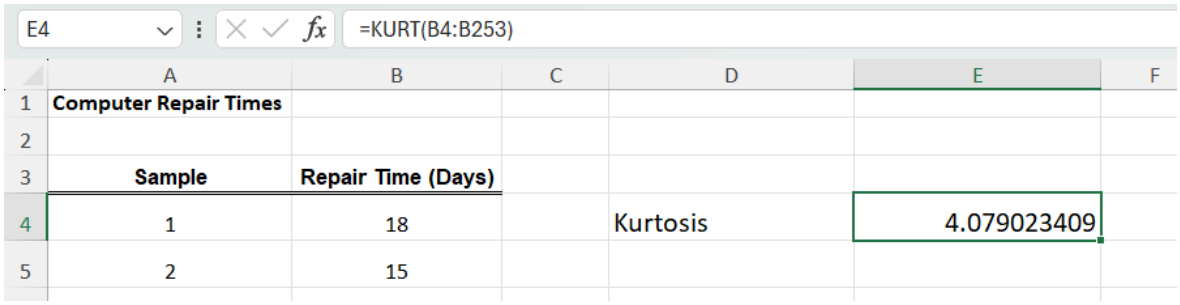


Image 3.12 Kurtosis of CRT – MS Excel

3.1.2. Analyzing by using R

3.1.2.1. Statistical description

Firstly, import the data and store data in a data frame

```
> crt<-read.csv(file.choose())
> crt
    Computer.Repair.Times                    X
1
2                  Sample Repair Time (Days)
3                       1                   18
4                       2                   15
5                       3                   17
6                       4                    9
7                       5                   37
8                       6                   15
9                       7                    8
10                      8                   29
11                      9                   10
12                     10                   14
13                     11                   17
14                     12                   12
15                     13                   13
16                     14                   12
17                     15                   11
18                     16                   14
19                     17                   13
20                     18                   16
21                     19                   13
22                     20                   15
23                     21                   16
24                     22                   12
25                     23                   34
26                     24                   29
27                     25                   13
28                     26                   19
29                     27                   12
30                     28                   15
31                     29                   16
32                     30                   14
33                     31                   15
34                     32                    7
35                     33                   40
36                     34                   16
37                     35                   11
```

Image 3.13 Import dataset CRT – R

Attach the database to the R search path.

```
> attach(crt)
```

3.1.2.2. Meaning of values

❖ ***Mean of CRT:***

```
> mean(day)
[1] 14.912
```

Image 3.14 Mean of CRT - R

❖ ***Mode of CRT:***

```
> getmode <- function(v){
+ uniqv <- unique(v)
+ uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
> getmode(day)
[1] 15
```

Image 3.15 Mode of CRT - R

❖ ***Quantile of CRT:***

```
> quantile(day, 0)
0%
 5
> quantile(day, 0.25)
25%
 11
> quantile(day, 0.5)
50%
 14
> quantile(day, 0.75)
75%
 17
> quantile(day, 1)
100%
  40
```

Image 3.16 Quantile of CRT – R

❖ ***Variance of CRT:***

```
> var(day)
[1] 35.50227
```

Image 3.17 Variance of CRT - R

❖ ***Standard Deviation of CRT:***

```
> sd(day)
[1] 5.958378
```

Image 3.18 Standard Deviation of CRT - R

❖ ***Skewness of CRT:***

Before calculating the skewness and the kurtosis, we have to install the library "e1071" in order to have the most exactly result.

```
> install.packages("e1071")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.2/e1071_1.7-14.tgz'
Content type 'application/x-gzip' length 673162 bytes (657 KB)
==================================================
downloaded 657 KB


The downloaded binary packages are in
        /var/folders/7_/dm3mjc6j0n128zlj27rkn9f80000gn/T//Rtmps7hglf/downloaded_packages
```

Image 3.19 Install the library "e1071"

```
> library(e1071)
```

Image 3.20 Call the library to use

```
> skewness(day)
[1] 1.674987
```

Image 3.21 Skewness of CRT – R

❖ ***Kurtosis of CRT:***

```
> kurtosis(day)
[1] 3.918314
```

Image 3.22 *Kurtosis of CRT – R*

3.1.3. Analyzing by using Python

3.1.3.1. Statistical description

STEP 1: Import Modules

```
[ ]  import numpy as np
     import pandas as pd
     import math
     import statistics as st
     import matplotlib.pyplot as plt
     import seaborn as sns
     from scipy.stats import skew
     from scipy.stats import kurtosis
```

Image 3.23 Import modules CRT – Python

STEP 2: Read the dataset

```
[ ]  cptrp_lab = pd.read_csv("Computer-Repair-Times.csv")
     cptrp_lab
```

	sample	time
0	1	18
1	2	15
2	3	17
3	4	9
4	5	37
...	...	...
245	246	18
246	247	31
247	248	6
248	249	17
249	250	13

250 rows × 2 columns

Image 3.24 Read the dataset CRT - Python

STEP 3: Rename

```
[1]  rpt=cptrp_lab.time
```

Image 3.25 Rename CRT - Python

3.1.3.2. Meaning of values

❖ ***Mean of CRT:***

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Image 3.26 Mean calculation formula

```
[ ]  #using the ordinary method
     mean_or=sum(rpt)/len(rpt)
     mean_or

     14.912
```

Image 3.27 Mean of CRT – Python – Method 1

```
[ ]  #using available commands from statistical libraries
     st.mean(rpt)

     14.912
```

Image 3.28 Mean of CRT – Python – Method 2

❖ ***Mode of CRT:***

```
[ ]  #using the ordinary method by coding:

     def find_mode(rpt):
         frequency = {}

         # Calculate the occurrence count for each number
         for num in rpt:
             if num in frequency:
                 frequency[num] += 1
             else:
                 frequency[num] = 1

         # Identify the maximum occurrence frequency
         max_frequency = max(frequency.values())

         # Generate a collection comprising the modal values
         modes = []
         for num, freq in frequency.items():
             if freq == max_frequency:
                 modes.append(num)

         return modes

     mode_or = find_mode(rpt)
     mode_or

     [12, 15]
```

Image 3.29 Mode of CRT – Python – Method 1

```
st.mode(rpt)

12
```

Image 3.30 Mode of CRT – Python – Method 2

Explanation: Because the if-else statement scans the entire dataset and records the values with the highest number of repetitions into a string, the output includes all numbers that have the same highest number of repetitions. This behavior remains consistent regardless of whether the repetitions of these values are exactly the same.

❖ ***Quantile of CRT:***

```
[ ]  q1 = np.quantile(rpt, 0.25)
     q2 = np.quantile(rpt, 0.5)
     q3 = np.quantile(rpt, 0.75)
     print(q1)
     print(q2)
     print(q3)
```

```
11.0
14.0
17.0
```

Image 3.31 Quantile of CRT - Python

❖ ***Variance of CRT:***

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Image 3.32 Variance calculation formula

```
[ ]  #using ordinary method
     var_or = sum((x - mean_or) ** 2 for x in rpt) // (n-1)
     var_or
```

```
35.0
```

Image 3.33 Variance of CRT – Python – Method 1

```
[ ]  #tool from lib st
     st.variance(rpt)
```

```
35.502265060240966
```

Image 3.34 Variance of CRT – Python – Method 2

❖ ***Standard Deviation of CRT:***

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Image 3.35 Standard calculation formula

```
[ ]  #using basic math library
     stdev_or=math.sqrt(var_or)
     stdev_or
```

```
5.916079783099616
```

Image 3.36 Standard Deviation of CRT – Python – Method 1

```
[ ]  st.stdev(rpt)
```

```
5.958377720507568
```

Image 3.37 Standard Deviation of CRT – Python – Method 2

❖ ***Skewness of CRT:***

$$skew[x] = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Image 3.38 Skewness calculation formula

```
[ ]  #using the formula:
     skew_or=(n / ((n - 1) * (n - 2))) * sum(((x - mean_or) / stdev_or) ** 3 for x in rpt)
     skew_or
```

```
1.7318980834350597
```

Image 3.39 Skewness of CRT – Python – Method 1

```
[ ]  skew(rpt,bias=False)
```

```
1.6952755753095303
```

Image 3.40 Skewness of CRT – Python – Method 2

❖ ***Kurtosis of CRT:***

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$

Image 3.41 Kurtosis calculation formula

```
[ ] kurt_or = kurt = (n * (n + 1) / ((n - 1) * (n - 2) * (n - 3))) * sum(((x - mean_or) / stdev_or) ** 4 for x in rpt) - 3 * ((n - 1) ** 2) / ((n - 2) * (n - 3))
    kurt_or

    4.284709990134218
```

Image 3.42 Kurtosis of CRT – Python – Method 1

```
[ ]  kurtosis(rpt,bias=False)

     4.079023408828646
```

Image 3.43 Kurtosis of CRT – Python – Method 2

3.2. Data Visualization

3.2.1. Using MS Excel

3.2.1.1. Histogram

STEP 1: Choose the **"Data"** tab on the top of the Excel Toolbar.

Click **"Data Analysis"** function.

Click **"Histogram"** and **"OK"** button

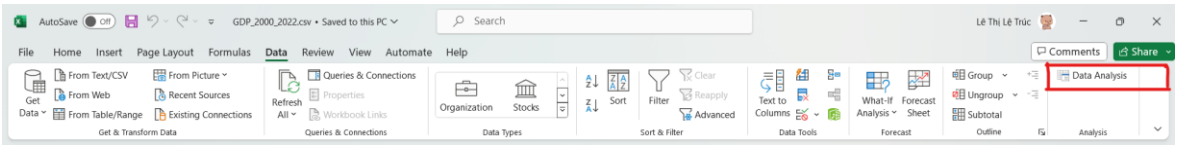


Image 3.44 STEP 1 of drawing Histogram – CRT – MS Excel – Choose "Data Analysis"

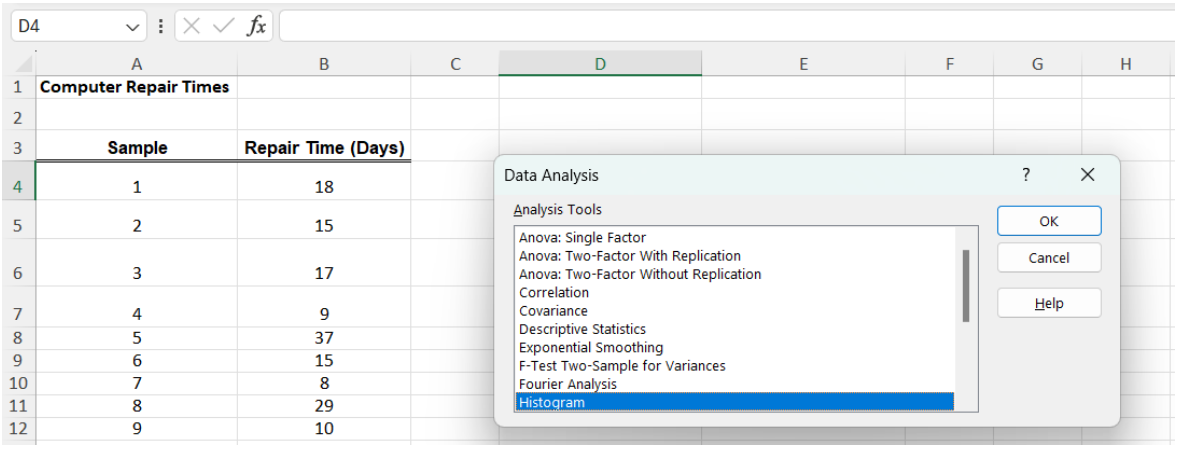


Image 3.45 STEP 1 of drawing Histogram – CRT – MS Excel – Click "Histogram"

STEP 2: Enter Input Range and select **"Cumulative Percentage"** and **"Chart Output".**

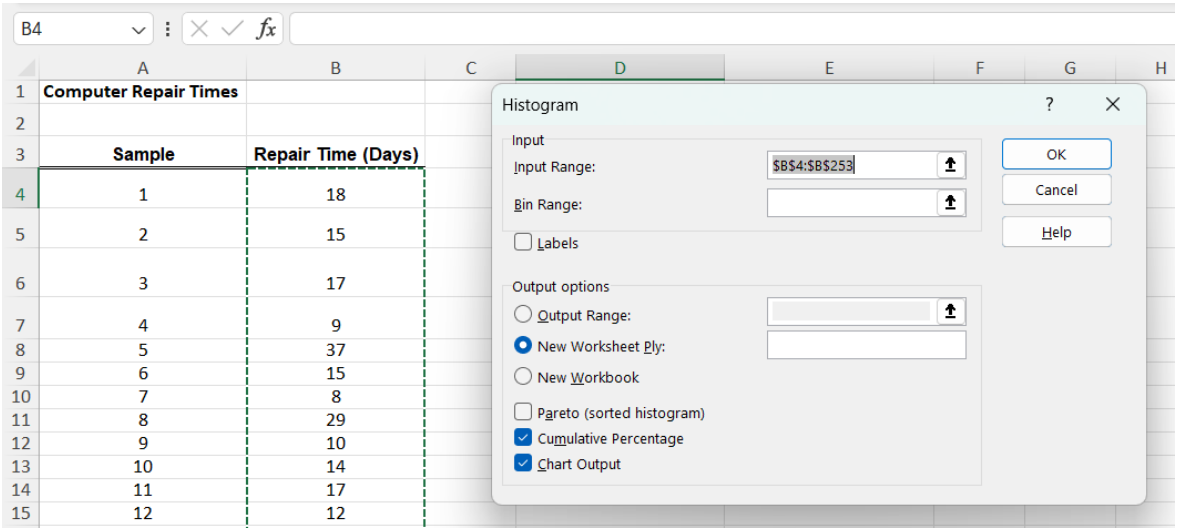


Image 3.46 STEP 2 of drawing Histogram – CRT – MS Excel

STEP 3: After finish step 2, the result will be displayed like the picture below.

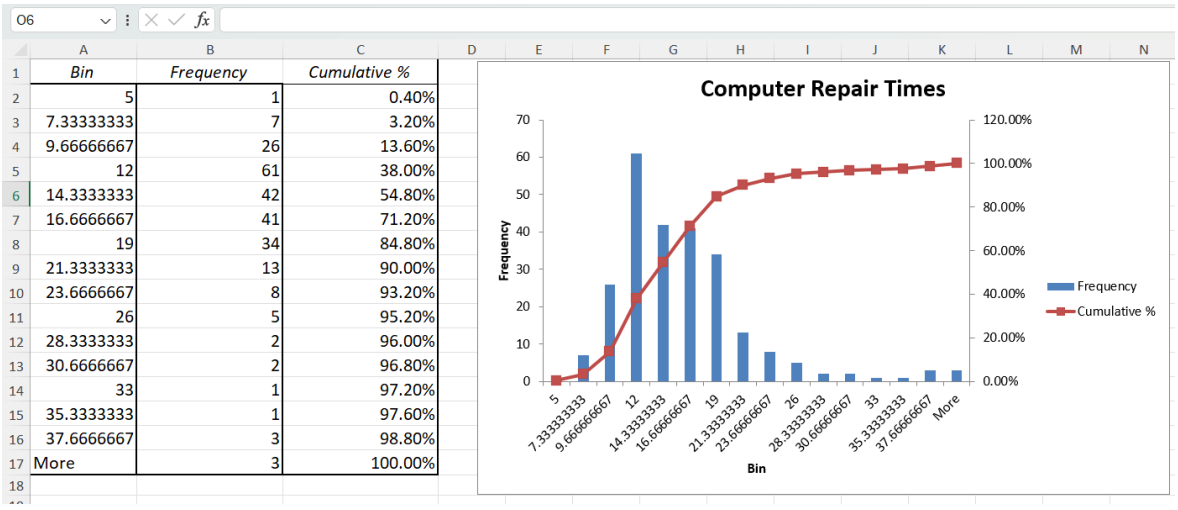


Image 3.47 STEP 3 of drawing Histogram – CRT – MS Excel

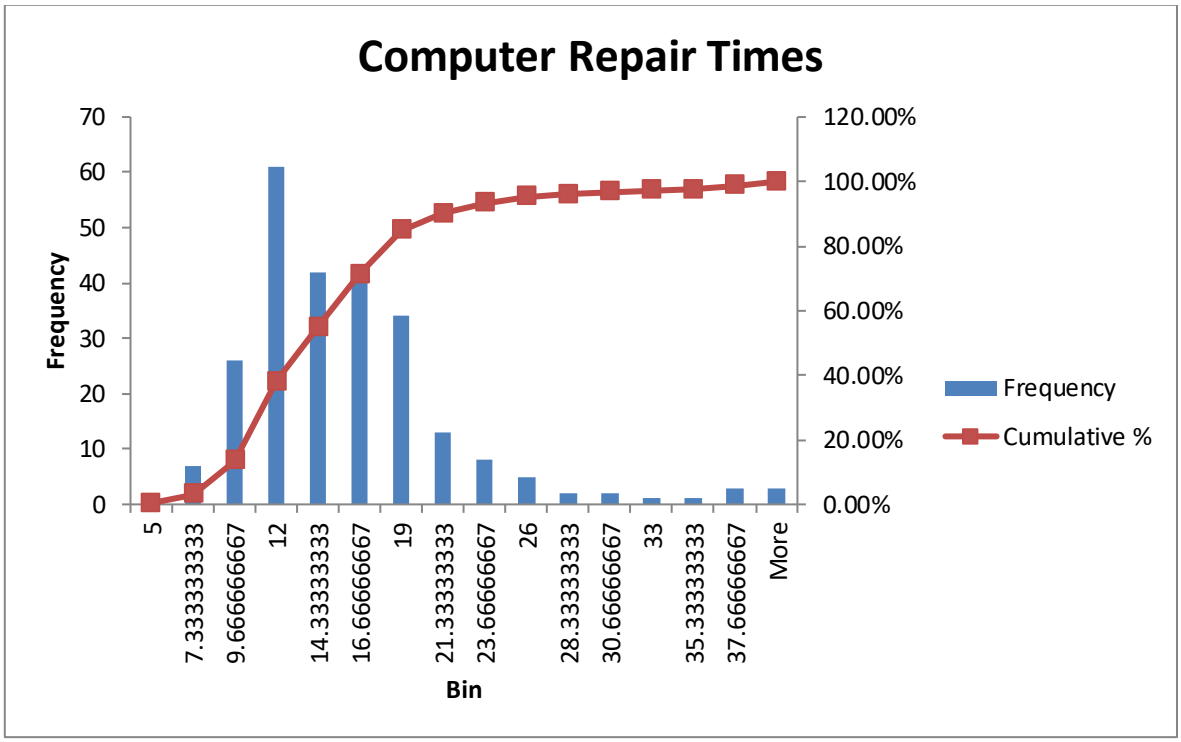


Image 3.48 Histogram of Computer Repair Times – MS Excel

3.2.1.2. Boxplot

STEP 1: Choose the **"Insert"** tab on the top of the Excel Toolbar.

Click **"Statistic Chart"** symbol and **"Box and Whisker"**

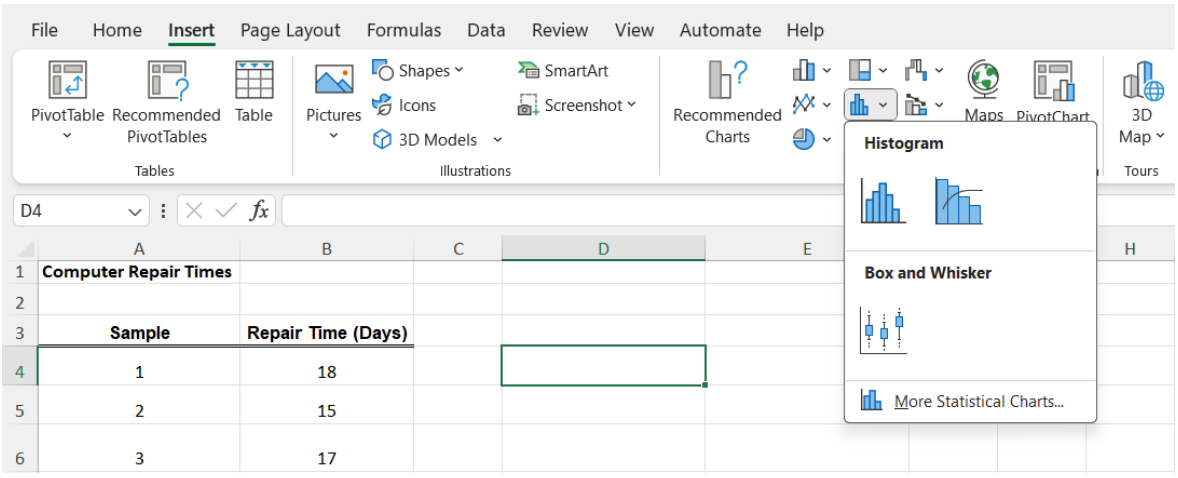


Image 3.49 STEP 1 of drawing Boxplot – CRT – MS Excel

STEP 2: After finish step 1, the result will be displayed like the picture below.

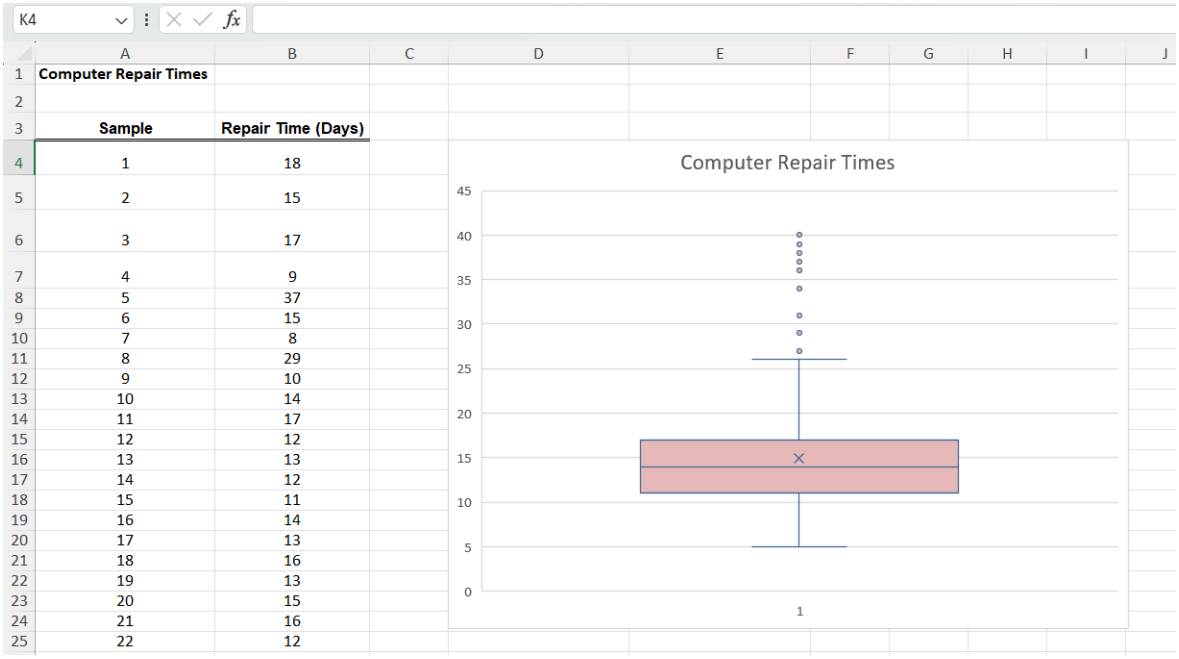


Image 3.50 STEP 2 of drawing Boxplot – CRT – MS Excel

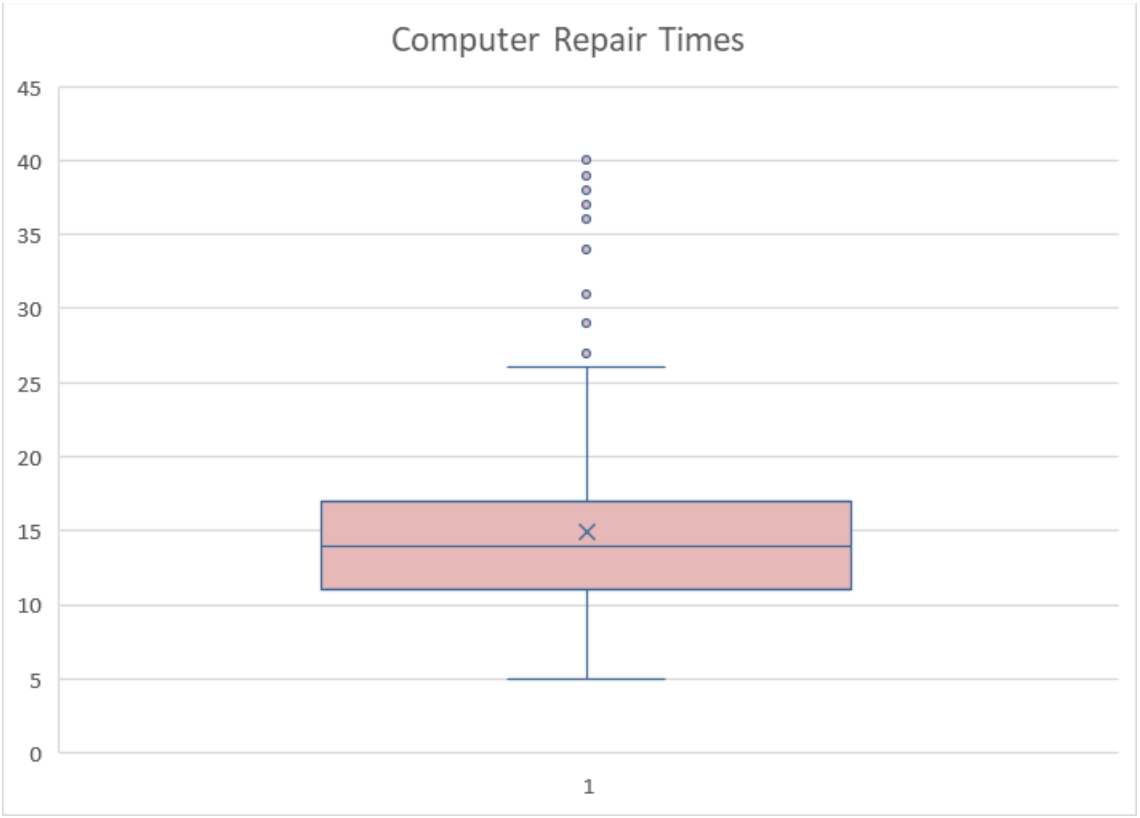


Image 3.51 Boxplot of Computer Repair Times – MS Excel

3.2.2. Using R

3.2.2.1. Histogram

```
> hist(day, main = "Histogram Computer Repair Time", ylab = "DAY", col = "pink")
```

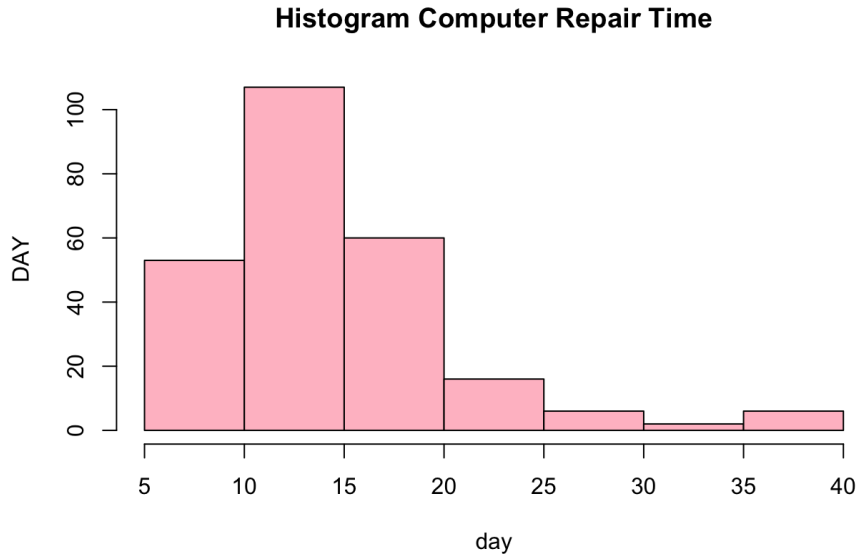


Image 3.52 Histogram of Computer Repair Times – R

3.2.2.2. Boxplot

```
> boxplot(day, main = "Boxplot Computer Repair Time", ylab = "DAY", col = "pink")
```

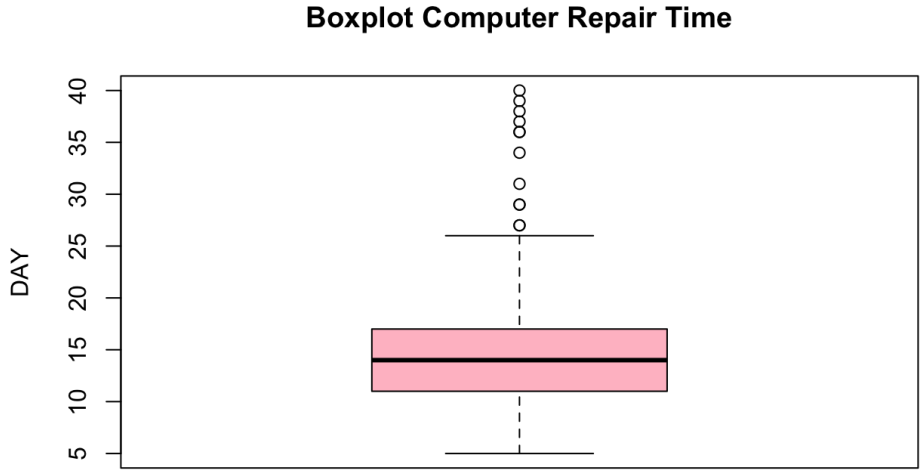


Image 3.53 Boxplot of Computer Repair Times – R

3.2.3. Using Python

3.2.3.1. Histogram

```
plt.hist(crt, bins=20, alpha=0.5, color='pink', edgecolor='black')
plt.title('Computer Repair Times')
plt.show()
```

✓ 0.3s

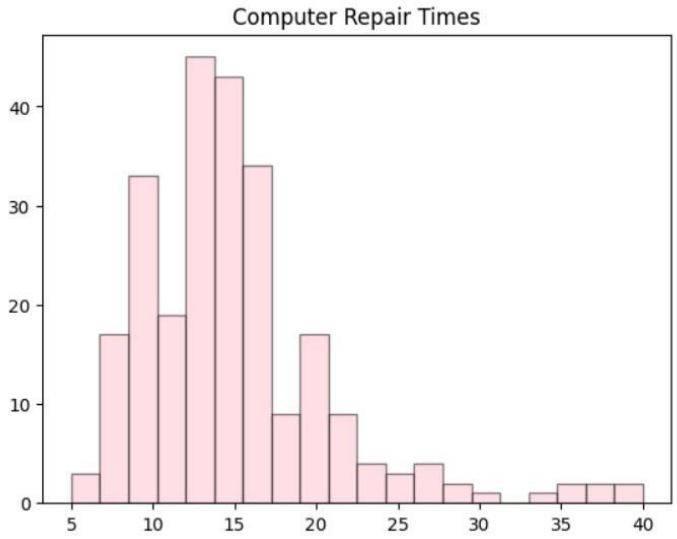


Image 3.54 Histogram of Computer Repair Times – Python

3.2.3.2. Boxplot

```
#using plt lib
outlier_color = dict(markerfacecolor='pink')
plt.boxplot(crt, flierprops=outlier_color)
plt.grid(True, axis='y', color='grey', linestyle='-', linewidth=0.5)
plt.text(1.1, q1, f'Q1: {q1:.2f}')
plt.text(1.1, q2, f'Q2: {q2:.2f}')
plt.text(1.1, q3, f'Q3: {q3:.2f}')
plt.show()
```

✓ 0.1s

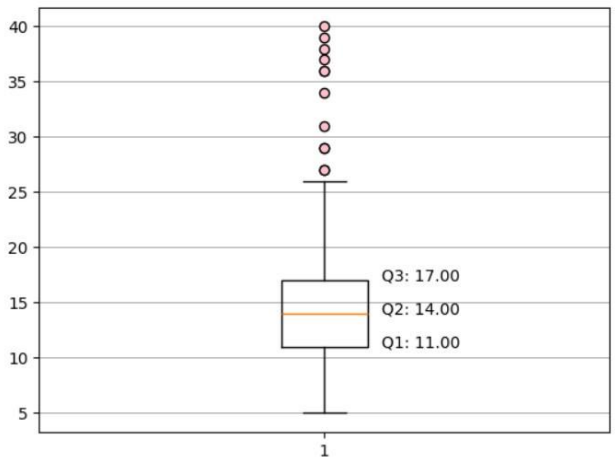


Image 3.55 Boxplot of Computer Repair Times – Python

Chapter 4. COLLEGES AND UNIVERSITIES

Lesson 2A. *Using Data Visualization and Descriptive Statistics. With the dataset: Colleges and Universities data*

Meaning of correlation and covariance coefficients. Use MS Excel, R and Python with examples 4.21.

4.1. Using MS Excel, R and Python programming language in Example 4.21

4.1.1. Analyzing by using MS Excel

4.1.1.1. Recalculate using MS Excel

❖ **Covariance coefficient**

STEP 1: Copy the **"Graduation %"** and **" Median SAT"** columns. Then calculate the mean and standard deviation of each column

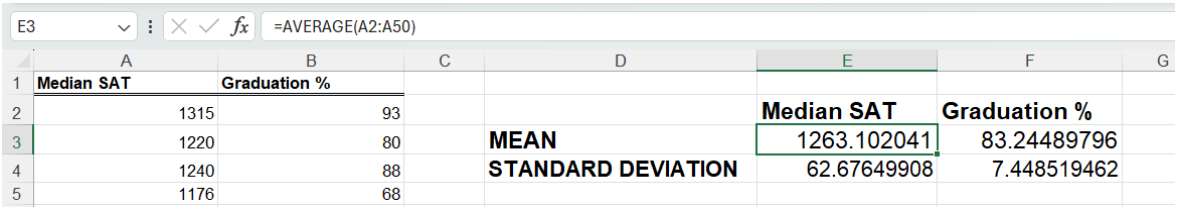


Image 4.1 Mean of Median SAT – MS Excel

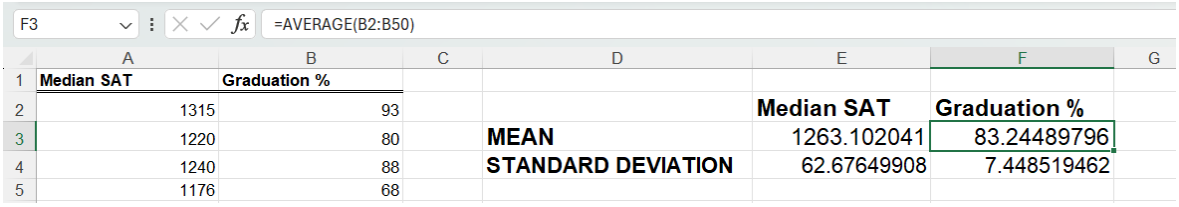


Image 4.2 Mean of Graduation% – MS Excel

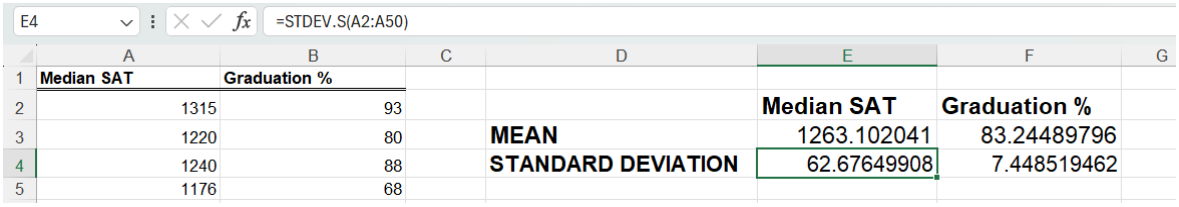


Image 4.3 Standard deviation of Median SAT – MS Excel

F4 =STDEV.S(B2:B50)

	A	B	C	D	E	F
1	Median SAT	Graduation %				
2	1315	93			Median SAT	Graduation %
3	1220	80		MEAN	1263.102041	83.24489796
4	1240	88		STANDARD DEVIATION	62.67649908	7.448519462
5	1176	68				

Image 4.4 Standard deviation of Graduation% – MS Excel

STEP 2: Calculate the value of **X – Mean(X)** (each value of Median SAT in each row - value of the Mean of Median SAT) and calculate the value of **Y - Mean(Y)** (each value of Graduation % in each row - value of the Mean of Graduation %)

C3 =A3-$G$3

	A	B	C	D	E	F	G	H
1	Median SAT	Graduation %	X - MEAN(X)	Y - MEAN (Y)				
2	1315	93	51.89795918	9.755102041			Median SAT	Graduation %
3	1220	80	-43.10204082	-3.244897959		MEAN	1263.102041	83.24489796
4	1240	88	-23.10204082	4.755102041		STANDARD DEVIATION	62.67649908	7.448519462
5	1176	68	-87.10204082	-15.24489796				
6	1300	90	36.89795918	6.755102041				

Image 4.5 X – MEAN(X) – MS Excel

D3 =B3-$H$3

	A	B	C	D	E	F	G	H
1	Median SAT	Graduation %	X - MEAN(X)	Y - MEAN (Y)				
2	1315	93	51.89795918	9.755102041			Median SAT	Graduation %
3	1220	80	-43.10204082	-3.244897959		MEAN	1263.102041	83.24489796
4	1240	88	-23.10204082	4.755102041		STANDARD DEVIATION	62.67649908	7.448519462
5	1176	68	-87.10204082	-15.24489796				
6	1300	90	36.89795918	6.755102041				

Image 4.6 Y – MEAN(Y) – MS Excel

STEP 3: Take the values of 2 corresponding rows and multiply them together

E3 =C3*D3

	A	B	C	D	E	F	G	H	I
1	Median SAT	Graduation %	X - MEAN(X)	Y - MEAN (Y)	(X - MEAN(X))* (Y - MEAN(Y))				
2	1315	93	51.89795918	9.755102041	506.2698875			Median SAT	Graduation %
3	1220	80	-43.10204082	-3.244897959	139.8617243		MEAN	1263.102041	83.24489796
4	1240	88	-23.10204082	4.755102041	-109.8525614		STANDARD DEVIATION	62.67649908	7.448519462
5	1176	68	-87.10204082	-15.24489796	1327.861724				
6	1300	90	36.89795918	6.755102041	249.2494794				
7	1281	90	17.89795918	6.755102041	120.9025406				

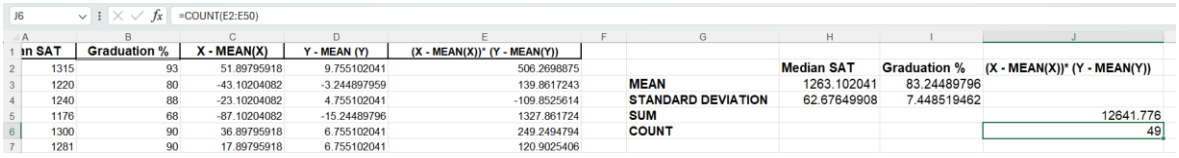
*Image 4.7 (X – MEAN(X) * (Y – MEAN(Y)) – MS Excel*

STEP 4: Calculate the sum and the count of (X-Mean(X))*(Y-Mean(Y))

G9

	A	B	C	D	E	F	G	H	I	J
1	ın SAT	Graduation %	X - MEAN(X)	Y - MEAN (Y)	(X - MEAN(X))* (Y - MEAN(Y))					
2	1315	93	51.89795918	9.755102041	506.2698875			Median SAT	Graduation %	(X - MEAN(X))* (Y - MEAN(Y))
3	1220	80	-43.10204082	-3.244897959	139.8617243		MEAN	1263.102041	83.24489796	
4	1240	88	-23.10204082	4.755102041	-109.8525614		STANDARD DEVIATION	62.67649908	7.448519462	
5	1176	68	-87.10204082	-15.24489796	1327.861724		SUM			12641.776
6	1300	90	36.89795918	6.755102041	249.2494794					

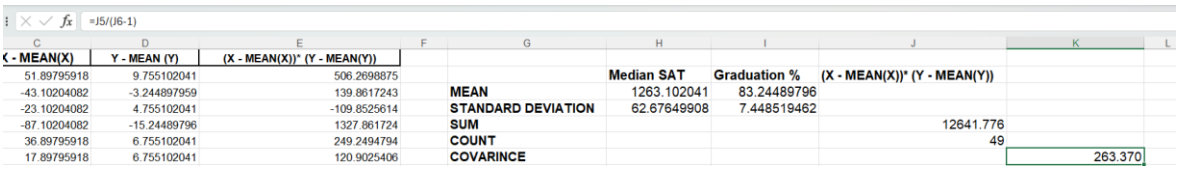
*Image 4.8 Sum of (X – MEAN(X) * (Y – MEAN(Y)) – MS Excel*



ın SAT	Graduation %	X - MEAN(X)	Y - MEAN (Y)	(X - MEAN(X))* (Y - MEAN(Y))		Median SAT	Graduation %	(X - MEAN(X))* (Y - MEAN(Y))
1315	93	51.89795918	9.755102041	506.2698875				
1220	80	-43.10204082	-3.244897959	139.8617243	MEAN	1263.102041	83.24489796	
1240	88	-23.10204082	4.755102041	-109.8525614	STANDARD DEVIATION	62.67649908	7.448519462	
1176	68	-87.10204082	-15.24489796	1327.861724	SUM			12641.776
1300	90	36.89795918	6.755102041	249.2494794	COUNT			49
1281	90	17.89795918	6.755102041	120.9025406				

*Image 4.9 Count of (X – MEAN(X) * (Y – MEAN(Y)) – MS Excel*

STEP 5: After step 4, calculate the Covarince coefficients.

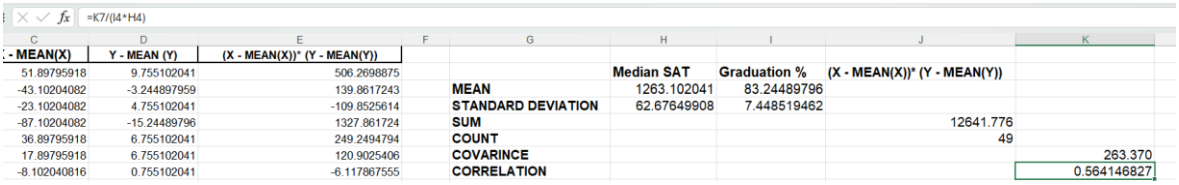


- MEAN(X)	Y - MEAN (Y)	(X - MEAN(X))* (Y - MEAN(Y))		Median SAT	Graduation %	(X - MEAN(X))* (Y - MEAN(Y))
51.89795918	9.755102041	506.2698875				
-43.10204082	-3.244897959	139.8617243	MEAN	1263.102041	83.24489796	
-23.10204082	4.755102041	-109.8525614	STANDARD DEVIATION	62.67649908	7.448519462	
-87.10204082	-15.24489796	1327.861724	SUM			12641.776
36.89795918	6.755102041	249.2494794	COUNT			49
17.89795918	6.755102041	120.9025406	COVARINCE			263.370

Image 4.10 Covarince coefficients – MS Excel

❖ **Correlation coefficients:**

After step 4 of calculate the Covariance, the Correlation's calculation much more easier. The correlation coefficient will be equal to the variance coefficient divided by the product of 2 standard deviations (Granduation% and Median SAT).



- MEAN(X)	Y - MEAN (Y)	(X - MEAN(X))* (Y - MEAN(Y))		Median SAT	Graduation %	(X - MEAN(X))* (Y - MEAN(Y))
51.89795918	9.755102041	506.2698875				
-43.10204082	-3.244897959	139.8617243	MEAN	1263.102041	83.24489796	
-23.10204082	4.755102041	-109.8525614	STANDARD DEVIATION	62.67649908	7.448519462	
-87.10204082	-15.24489796	1327.861724	SUM			12641.776
36.89795918	6.755102041	249.2494794	COUNT			49
17.89795918	6.755102041	120.9025406	COVARINCE			263.370
-8.102040816	0.755102041	-6.117867555	CORRELATION			0.564146827

Image 4.11 Correlation coefficients – MS Excel

4.1.2. Analyzing by using R

4.1.2.1. Statistical Description

Firstly, import the data and store data in a data frame

```
> cau<-read.csv(file.choose())
> cau
                School       Type Median.SAT Acceptance.Rate Expenditures.Student Top.10..HS
1             Amherst   Lib Arts       1315             22%              $26,636          85
2             Barnard   Lib Arts       1220             53%              $17,653          69
3               Bates   Lib Arts       1240             36%              $17,554          58
4            Berkeley University       1176             37%              $23,665          95
5             Bowdoin   Lib Arts       1300             24%              $25,703          78
6               Brown University       1281             24%              $24,201          80
7           Bryn Mawr   Lib Arts       1255             56%              $18,847          70
8            Cal Tech University       1400             31%             $102,262          98
9            Carleton   Lib Arts       1300             40%              $15,904          75
10   Carnegie Mellon University       1225             64%              $33,607          52
11 Claremont McKenna   Lib Arts       1260             36%              $20,377          68
12             Colby   Lib Arts       1200             46%              $18,872          52
13           Colgate   Lib Arts       1258             38%              $17,520          61
14          Columbia University       1268             29%              $45,879          78
15           Cornell University       1280             30%              $37,137          85
16         Davisdson   Lib Arts       1230             36%              $17,721          77
17              Duke University       1310             25%              $39,504          91
18        Georgetown University       1278             24%              $23,115          79
19          Grinnell   Lib Arts       1244             67%              $22,301          65
20          Hamilton   Lib Arts       1215             38%              $20,722          51
21           Harvard University       1370             18%              $46,918          90
22         Haverford   Lib Arts       1285             35%              $19,418          71
23     Johns Hopkins University       1290             48%              $45,460          69
24        Middlebury   Lib Arts       1255             25%              $24,718          65
25               MIT University       1357             30%              $56,766          95
```

Image 4.12 Import dataset CAU - R

Attach the database to the R search path

```
> attach(cau)
```

4.1.2.2. Meaning of values

❖ **The Correlation between Median SAT & Graduation %**

```
> cor(Median.SAT, Graduation..)
[1] 0.5641468
```

Image 4.13 Correlation between SAT & Graduation %

❖ **The Covariance between Median SAT & Graduation %**

```
> cov(Median.SAT, Graduation..)
[1] 263.3703
```

Image 4.14 Covariance between Median SAT & Graduation %

4.1.3. Analyzing by using Python

4.1.3.1. Statistical Description

STEP 1: Import Modules

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

Image 4.15 Import modules CAU - Python

STEP 2: Read the dataset

```
[ ] dh = pd.read_csv("College_And_Universities_1.csv")
    dh.head()
```

	School	Type	Median_SAT	Acceptance_Rate	Expenditures/Student	Top _10percent_HS	Graduation_percent
0	Amherst	Lib Arts	1315	22%	$26,636	85	93
1	Barnard	Lib Arts	1220	53%	$17,653	69	80
2	Bates	Lib Arts	1240	36%	$17,554	58	88
3	Berkeley	University	1176	37%	$23,665	95	68
4	Bowdoin	Lib Arts	1300	24%	$25,703	78	90

Image 4.16 Read the dataset CAU - Python

STEP 3: Rename

```
[ ] graduation = dh['Graduation_percent']

    median_sat = dh['Median_SAT']

    dh['Expenditures/Student'] = dh['Expenditures/Student'].replace({'\$': '', ',': ''}, regex=True).astype(float)
    top=dh['Top _10percent_HS']

    expend=dh['Expenditures/Student']

    dh['Acceptance_Rate'] = dh['Acceptance_Rate'].replace({'%': '', ',': ''}, regex=True).astype(float)
    accept=dh['Acceptance_Rate']
```

To avoid issues when reading and collecting data, we need to remove or replace special characters in the Expenditures/Students and Acceptance Rate columns. This can be done using the "replace" command with a space (' '), followed by converting the data to the "float" type.

4.1.3.2. Meaning of values

❖ **The Covariance between Median SAT & Graduation %**

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Image 4.17 Covariance calculation formula

```
[ ]  #using the formula - ordinary way:
     n=len(graduation)
     meanGraduation_or=sum(graduation)/len(graduation)
     meanMedianSat_or=sum(median_sat)/len(median_sat)
     cov_or = sum((graduation[i] - meanGraduation_or) * (median_sat[i] - meanMedianSat_or) for i in range(n)) / (n-1)
     cov_or

     263.37032312925174
```

Image 4.18 Covariance of CAU – Python – Method 1

```
[ ]  #using np lib
     np.cov(graduation, median_sat)[0, 1]

     263.37032312925174
```

Image 4.19 Covariance of CAU – Python – Method 2

❖ **The Correlation between Median SAT & Graduation %**

$$r_{xy} = \frac{cov(X,Y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt[2]{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

Image 4.20 Correlation calculation formula

```
[ ]  #using the formula without statistics tools
     #finding standard deviation of x:
     var_x_or = sum((x - meanGraduation_or) ** 2 for x in graduation) / (n-1)
     stdev_x_or = math.sqrt(var_x_or)

     #finding standard deviation of y:
     var_y_or = sum((y - meanMedianSat_or) ** 2 for y in median_sat) / (n-1)
     stdev_y_or = math.sqrt(var_y_or)

     s_x_s_y = stdev_x_or * stdev_y_or
     corr_or = cov_or/s_x_s_y
     corr_or
```

```
0.564146826697419
```

Image 4.21 Correlation of CAU – Python – Method 1

```
[ ]  np.corrcoef(graduation, median_sat)[0, 1]
```

```
0.5641468266974192
```

Image 4.22 Correlation of CAU – Python – Method 2

4.2. Meaning of correlation and covariance coefficients in Example 4.22

Lesson 2B. *Using Data Visualization and Descriptive Statistics. With the dataset: Colleges and Universities data*

Meaning of correlation and covariance coefficients. (Use MS Excel, R and Python with examples 4.22

4.2.1. Analyzing by using MS Excel

4.2.1.1. Using Correlation Tools:

❖ **Covariance coefficients:**

STEP 1: Change to the **"Data"** tab on the top of the Excel toolbar. Click the **"Data Analysis"** button after that a window will be displayed.

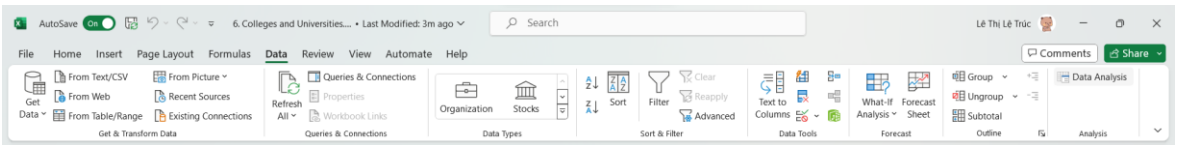


Image 4.23 STEP 1 of analyzing by using MS Excel with the dataset CAU - Covarience

STEP 2: In the window, choose the **"Covariance"** in the box and then click **"OK"** button.

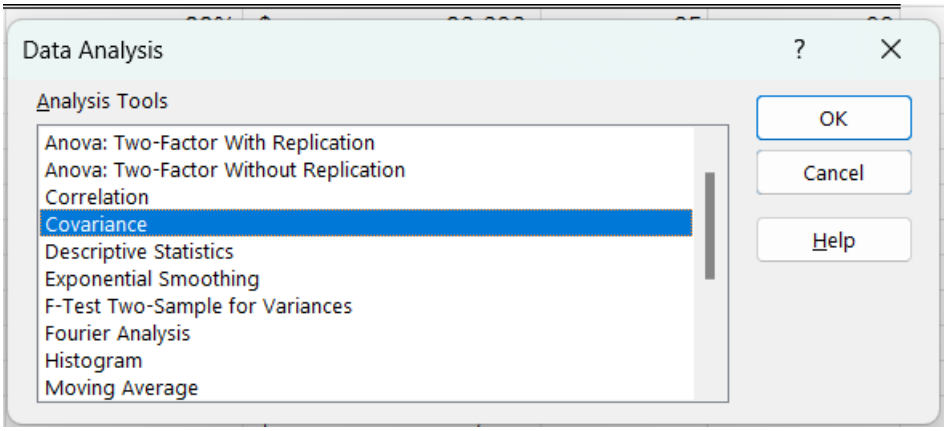


Image 4.24 STEP 2 of analyzing by using MS Excel with the dataset CAU – Covarience

STEP 3: Enter Input Range the data range you want to calculate **($C$3:$G$52)** and choose **"Columns"** (Because the input are in columns) in the "Grouped By" and tick on the **"Label in first Row "**, then choose "**New work sheet ply**" in the "Output options". Finally, click **"OK"** button

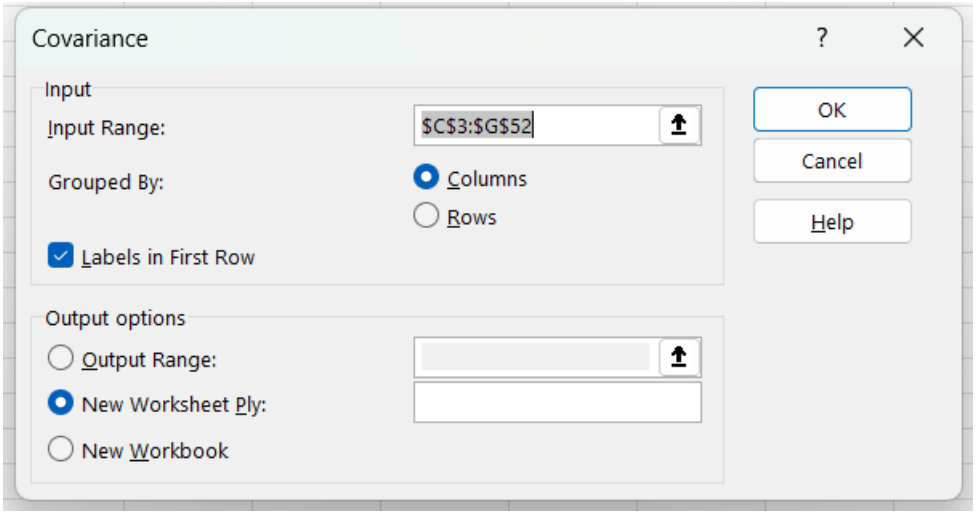


Image 4.25 STEP 3 of analyzing by using MS Excel with the dataset CAU - Covarience

STEP 4: After step 3, the result will be displayed like the picture below

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	3848.173261				
3	Acceptance Rate	-4.941532695	0.017515285			
4	Expenditures/Student	543764.8442	-575.7591087	234234025.3		
5	Top 10% HS	418.8771345	-1.082249063	103818.7701	179.876718	
6	Graduation %	257.9954186	-0.53698459	4795.593503	13.70512287	54.34818825

Image 4.26 STEP 4 of analyzing by using MS Excel with the dataset CAU – Covarience

❖ **Correlation coefficients:**

STEP 1: Change to the **"Data"** tab on the top of the Excel toolbar. Click the **"Data Analysis"** button after that a window will be displayed.

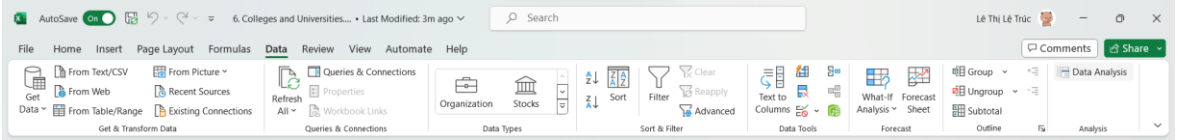


Image 4.27 STEP 1 of analyzing by using MS Excel with the dataset CAU – Correlation

STEP 2: In the window, choose the **"Correlation"** in the box and then click **"OK"** button

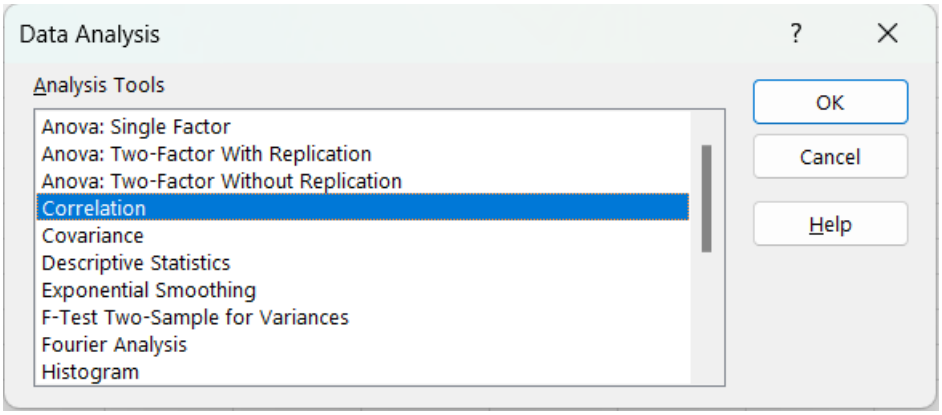


Image 4.28 STEP 2 of analyzing by using MS Excel with the dataset CAU – Correlation

STEP 3: Enter Input Range the data range you want to calculate **($C$3:$G$52)** and choose **"Columns"** (Because the input are in columns) in the "Grouped By" and tick on the **"Label in first Row ",** then choose **"New work sheet ply"** in the "Output options". Finally, click **"OK"** button

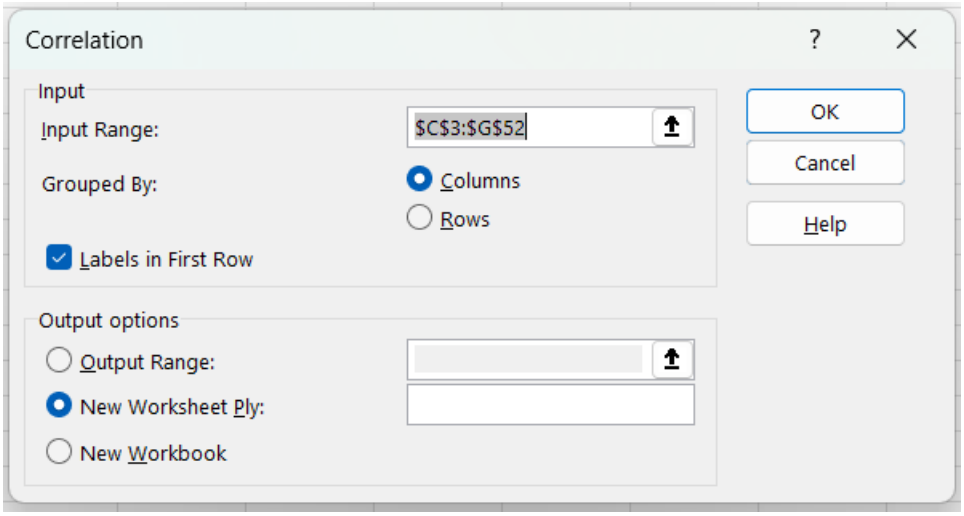
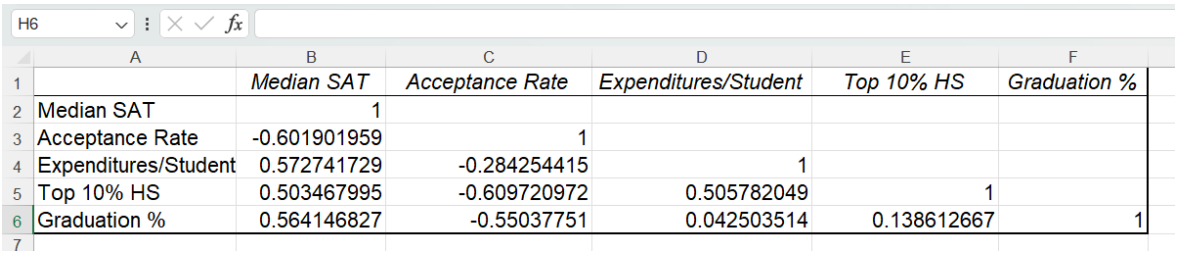


Image 4.29 STEP 3 of analyzing by using MS Excel with the dataset CAU – Correlation

STEP 4: The result will be displayed like the picture below



	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1
7						

Image 4.30 STEP 4 of analyzing by using MS Excel with the dataset CAU – Correlation

4.2.2. Analyzing by using R

4.2.2.1. Statistical Description

Firstly, import the data and store data in a data frame

```
> cau<-read.csv(file.choose())
> cau
                school       type medianSAT acceptanceRate expenditures top10 graduation
1             Amherst   Lib Arts      1315             22%     $26,636    85         93
2             Barnard   Lib Arts      1220             53%     $17,653    69         80
3               Bates   Lib Arts      1240             36%     $17,554    58         88
4            Berkeley University      1176             37%     $23,665    95         68
5             Bowdoin   Lib Arts      1300             24%     $25,703    78         90
6               Brown University      1281             24%     $24,201    80         90
7           Bryn Mawr   Lib Arts      1255             56%     $18,847    70         84
8            Cal Tech University      1400             31%    $102,262    98         75
9            Carleton   Lib Arts      1300             40%     $15,904    75         80
10   Carnegie Mellon University      1225             64%     $33,607    52         77
11 Claremont McKenna   Lib Arts      1260             36%     $20,377    68         74
12             Colby   Lib Arts      1200             46%     $18,872    52         84
13           Colgate   Lib Arts      1258             38%     $17,520    61         85
14          Columbia University      1268             29%     $45,879    78         90
15           Cornell University      1280             30%     $37,137    85         83
16         Davisdson   Lib Arts      1230             36%     $17,721    77         89
17              Duke University      1310             25%     $39,504    91         91
```

Image 4.31 Import dataset CAU – R

Attach the database to the R search path

```
> attach(cau)
```

4.2.2.2. Meaning of values

Convert the values before calculating

```
> rateconverted<-as.numeric(sub("%", "", acceptanceRate)) / 100
```

Clean the "%" symbol in expenditures values

```
> cau$expenditures<-as.numeric(gsub("[$,]","",cau$expenditures))
>
> cau
```

	school	type	medianSAT	acceptanceRate	expenditures	top10	graduation
1	Amherst	Lib Arts	1315	22%	26636	85	93
2	Barnard	Lib Arts	1220	53%	17653	69	80
3	Bates	Lib Arts	1240	36%	17554	58	88
4	Berkeley	University	1176	37%	23665	95	68
5	Bowdoin	Lib Arts	1300	24%	25703	78	90
6	Brown	University	1281	24%	24201	80	90
7	Bryn Mawr	Lib Arts	1255	56%	18847	70	84
8	Cal Tech	University	1400	31%	102262	98	75
9	Carleton	Lib Arts	1300	40%	15904	75	80
10	Carnegie Mellon	University	1225	64%	33607	52	77
11	Claremont McKenna	Lib Arts	1260	36%	20377	68	74
12	Colby	Lib Arts	1200	46%	18872	52	84
13	Colgate	Lib Arts	1258	38%	17520	61	85
14	Columbia	University	1268	29%	45879	78	90
15	Cornell	University	1280	30%	37137	85	83
16	Davisdson	Lib Arts	1230	36%	17721	77	89
17	Duke	University	1310	25%	39504	91	91
18	Georgetown	University	1278	24%	23115	79	89
19	Grinnell	Lib Arts	1244	67%	22301	65	73
20	Hamilton	Lib Arts	1215	38%	20722	51	85
21	Harvard	University	1370	18%	46918	90	90
22	Haverford	Lib Arts	1285	35%	19418	71	87
23	Johns Hopkins	University	1290	48%	45460	69	86
24	Middlebury	Lib Arts	1255	25%	24718	65	92
25	MIT	University	1357	30%	56766	95	86
26	Mount Holyoke	Lib Arts	1200	61%	23358	47	83
27	Northwestern	University	1230	47%	28851	77	82
28	Oberlin	Lib Arts	1247	54%	23591	64	77
29	Occidental	Lib Arts	1170	49%	20192	54	72
30	Pomona	Lib Arts	1320	33%	26668	79	80
31	Princeton	University	1340	17%	48123	89	93
32	Rice	University	1327	24%	26730	85	88
33	Smith	Lib Arts	1195	57%	25271	65	87
34	Stanford	University	1370	18%	61921	92	88
35	Swarthnore	Lib Arts	1310	24%	27487	78	88
36	U Michigan	University	1195	60%	21853	71	77
37	U of Chicago	University	1300	45%	38937	74	73

We need to convert expenditures to numeric type

```
> expenditures_cleaned <- gsub("[^0-9.]", "", cau$expenditures)
>
> expenditures_numeric <- as.numeric(expenditures_cleaned)
```

Calculating the values:

```
> cor(medianSAT, medianSAT)
[1] 1

> cor(rateconverted, medianSAT)
[1] -0.601902

> cor(expenditures_numeric, medianSAT)
[1] 0.5727417

> cor(top10, medianSAT)
[1] 0.503468

> cor(graduation, medianSAT)
[1] 0.5641468

> cor(rateconverted, rateconverted)
[1] 1

> cor(rateconverted, expenditures_numeric)
[1] -0.2842544

> cor(rateconverted, top10)
[1] -0.609721

> cor(rateconverted, graduation)
[1] -0.5503775

> cor(expenditures_numeric, expenditures_numeric)
[1] 1

> cor(expenditures_numeric, top10)
[1] 0.505782

> cor(expenditures_numeric, graduation)
[1] 0.04250351

> cor(top10, top10)
[1] 1

> cor(top10, graduation)
[1] 0.1386127
```

```
> cor(graduation, graduation)
[1] 1
```

4.2.3. Analyzing by using Python

4.2.3.1. Meaning of values

```
[ ] dh.corr()
```

	Median_SAT	Acceptance_Rate	Expenditures/Student	Top _10percent_HS	Graduation_percent
Median_SAT	1.000000	-0.601902	0.572742	0.503468	0.564147
Acceptance_Rate	-0.601902	1.000000	-0.284254	-0.609721	-0.550378
Expenditures/Student	0.572742	-0.284254	1.000000	0.505782	0.042504
Top _10percent_HS	0.503468	-0.609721	0.505782	1.000000	0.138613
Graduation_percent	0.564147	-0.550378	0.042504	0.138613	1.000000

Image 4.32 Meaning of values – CAU – Python

Chapter 5. HOME MARKET VALUE

5.1. Using MS Excel, R and Python programming language in Example 4.23

Lesson 3. *Using Data Visualization and Descriptive Statistics. With the dataset: Home Market Values Meaning of Outlier*

Using statistics to determine Outlier values. Use MS Excel, R and Python with example 4.23

5.1.1. Analyzing by using MS Excel

- **Meaning of Outliers:**

STEP 1: Calculate the Mean of Square Feet and Market Value

F3 =AVERAGE(B4:B45)

	A	B	C	D	E	F
1	Home Market Value					
2						SQUARE FEET
3	House Age	Square Feet	Market Value		MEAN	1,695
4	33	1,812	$90,000.00			

Image 5.1 Mean of Square Feet – MS Excel

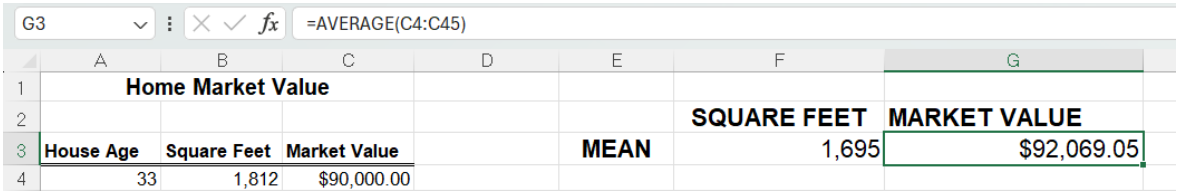


Image 5.2 Mean of Market Value – MS Excel

STEP 2: Calculate the standard deviation of Square Feet and Market Value

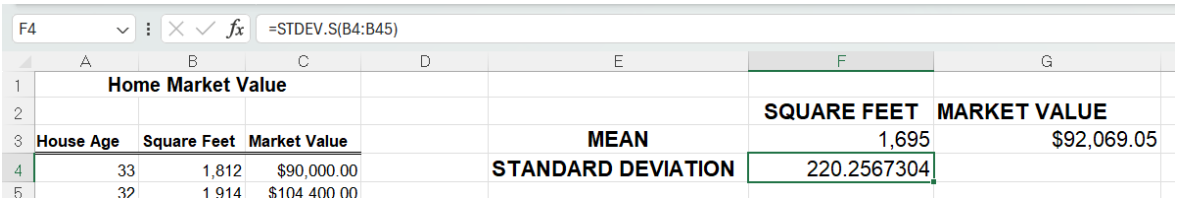


Image 5.3 Standard deviation of Square Feet – MS Excel

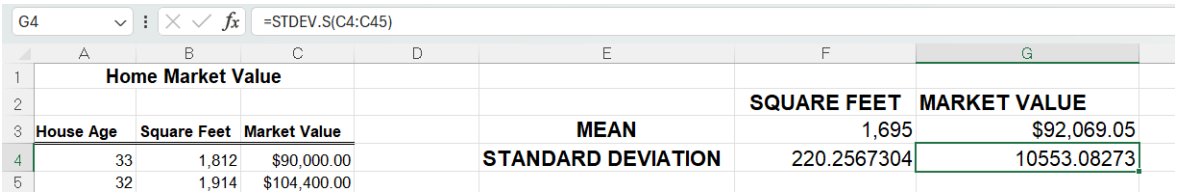


Image 5.4 Standard deviation of Market Value – MS Excel

STEP 3: Calculate z-score

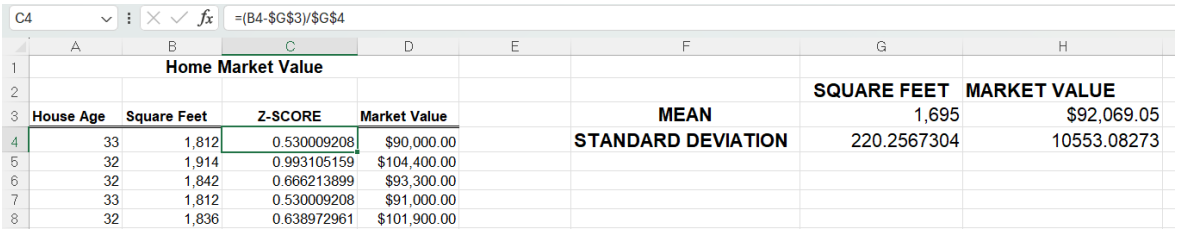


Image 5.5 Z-Score of Square Feet – MS Excel

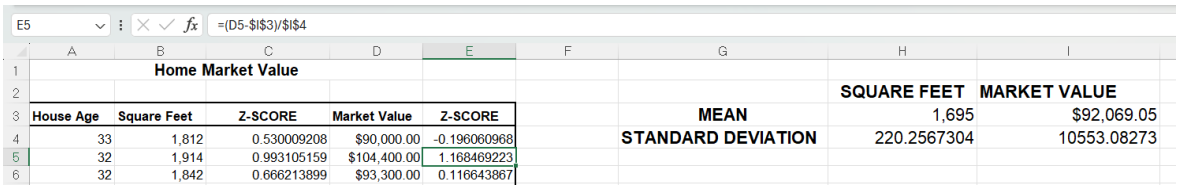


Image 5.6 Z-Score of Market Value – MS Excel

STEP 4: Find outlier

After calculate z-score, if the value of z-score > 3, then this is right outlier. If the value ≤ -3, that is left outlier.

	A	B	C	D	E
1	Home Market Value				
2					
3	**House Age**	**Square Feet**	**Z-SCORE**	**Market Value**	**Z-SCORE**
4	33	1,812	0.530009208	$90,000.00	-0.196060968
5	32	1,914	0.993105159	$104,400.00	1.168469223
6	32	1,842	0.666213899	$93,300.00	0.116643867
7	33	1,812	0.530009208	$91,000.00	-0.101301927
8	32	1,836	0.638972961	$101,900.00	0.93157162
9	33	2,028	1.510682986	$108,500.00	1.556981291
10	32	1,732	0.166796698	$87,600.00	-0.423482667
11	33	1,850	0.70253515	$96,000.00	0.372493278
12	32	1,791	0.434665924	$89,200.00	-0.271868201
13	33	1,666	-0.132853624	$88,400.00	-0.347675434
14	32	1,852	0.711615463	$100,800.00	0.827336675
15	32	1,620	-0.341700817	$96,700.00	0.438824607
16	32	1,692	-0.014809558	$87,500.00	-0.432958571
17	32	2,372	3.072496781	$114,000.00	2.078156017
18	32	2,372	3.072496781	$113,200.00	2.002348784
19	33	1,666	-0.132853624	$87,500.00	-0.432958571
20	32	2,123	1.941997842	$116,100.00	2.277150003
21	32	1,620	-0.341700817	$94,700.00	0.249306525
22	32	1,731	0.162256541	$86,400.00	-0.537193516
23	32	1,666	-0.132853624	$87,100.00	-0.470862187
24	28	1,520	-0.795716455	$83,400.00	-0.821470639
25	27	1,484	-0.959162085	$79,800.00	-1.162603187
26	28	1,588	-0.486985821	$81,500.00	-1.001512817

In the picture, there are 2 value > 3, so that are the right outliers.

5.1.2. Analyzing by using R

5.1.2.1. Statistical Description

Install library to read excel file

```
> install.packages("readxl")
WARNING: Rtools is required to build R packages but is not currently instal
led. Please download and install the appropriate version of Rtools before p
roceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/ohbit/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
also installing the dependencies 'cli', 'glue', 'utf8', 'rematch', 'fansi',
'lifecycle', 'magrittr', 'pillar', 'pkgconfig', 'rlang', 'vctrs', 'crayon',
'hms', 'prettyunits', 'R6', 'cellranger', 'tibble', 'cpp11', 'progress'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/cli_3.6.2.zip'
Content type 'application/zip' length 1339105 bytes (1.3 MB)
downloaded 1.3 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/glue_1.7.0.zi
p'
Content type 'application/zip' length 161275 bytes (157 KB)
downloaded 157 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/utf8_1.2.4.zi
p'
Content type 'application/zip' length 149818 bytes (146 KB)
downloaded 146 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/rematch_2.0.0.
zip'
Content type 'application/zip' length 19070 bytes (18 KB)
downloaded 18 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/fansi_1.0.6.zi
p'
Content type 'application/zip' length 313879 bytes (306 KB)
downloaded 306 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/lifecycle_1.0.
4.zip'
Content type 'application/zip' length 139679 bytes (136 KB)
downloaded 136 KB
```

Image 5.7 Install the library - R

```
> library(readxl)
Warning message:
package 'readxl' was built under R version 4.3.3
```

Image 5.8 Import library - R

```
> data<-read_excel(file.choose())
```

Image 5.9 Read excel file – R

```
> data
# A tibble: 42 × 3
   `House Age` `Square Feet` `Market Value`
         <dbl>         <dbl>          <dbl>
 1          33          1812          90000
 2          32          1914         104400
 3          32          1842          93300
 4          33          1812          91000
 5          32          1836         101900
 6          33          2028         108500
 7          32          1732          87600
 8          33          1850          96000
 9          32          1791          89200
10          33          1666          88400
# i 32 more rows
# i Use `print(n = ...)` to see more rows
```

Attach the database to the R search path

```
> attach(data)
```

5.1.2.2. Meaning of Outliers

- **Calculating the Quantile Q1:**

```
> q1<-quantile(`Market Value`, 0.25)
> q1
  25%
86575
```

- **Calculating the Quantile Q3:**

```
> q3<-quantile(`Market Value`, 0.75)
> q3
  75%
96525
```

- **Calculating the IQR:**

```
> IQR(`Market Value`)
[1] 9950
```

- **Calculating the lower bound:**

```
> lower_bound<-q1-1.5*IQR(`Market Value`)
> lower_bound
  25%
71650
```

- **Calculating the upper bound:**

```
> upper_bound<-q3-1.5*IQR(`Market Value`)
> upper_bound
  75%
81600
```

❖ **Finding the outliers:**

```
> outliers<-`Market Value`[`Market Value` < lower_bound | `Market Value` >
upper_bound]
> outliers
 [1]  90000 104400  93300  91000 101900 108500  87600  96000  89200
[10]  88400 100800  96700  87500 114000 113200  87500 116100  94700
[19]  86400  87100  83400  87100  82600  87600  94200  82000  88100
[28]  88100  88600  84400  90900  91300 100700  87200  96700 120700
```

5.1.3. Analyzing by using Python

5.1.3.1. Statistical Description

STEP 1: Import modules

```
[ ]  import numpy as np
     import pandas as pd
     import statistics as st
     import matplotlib.pyplot as plt
```

Image 5.10 Import muldes Home & Market – Python

STEP 2: Read the dataset

```
[ ]  too = pd.read_csv("Home_Market_Value.csv")
     too.head()
```

	House Age	Square Feet	Market Value
0	33	1,812	$90,000.00
1	32	1,914	$104,400.00
2	32	1,842	$93,300.00
3	33	1,812	$91,000.00
4	32	1,836	$101,900.00

Image 5.11 Read the dataset Home & Market – Python

STEP 3: Rename

```
[ ]  too['Square Feet'] = too['Square Feet'].replace('[,]', '', regex=True).astype(int)
     too['Market Value'] = too['Market Value'].replace('[$,]', '', regex=True).astype(float)
     sf=too['Square Feet']
     mv=too['Market Value']
     ha=too['House Age']
```

5.1.3.2. Meaning of Outliers

STEP 1: Calculate the median of each column of values in the dataset

```
[ ]  st.median(ha)
```

```
28.0
```

```
[ ]  st.median(sf)
```

```
1666.0
```

```
[ ]  st.median(mv)
```

```
88900.0
```

STEP 2: Finding outliers of House Age column through IQR

```
[ ]  q1_ha = ha.quantile(0.25)
     q2_ha = ha.quantile(0.5)
     q3_ha = ha.quantile(0.75)
     iqr_ha = q3_ha - q1_ha
     lower_bound_ha = q1_ha - 1.5 * iqr_ha
     upper_bound_ha = q3_ha + 1.5 * iqr_ha
     #using boolean to choose column in dataframe "Too"
     outliers_ha = too[(ha < lower_bound_ha) | (ha > upper_bound_ha)]
     outliers_ha
```

	House Age	Square Feet	Market Value

```
[ ]  print(q1_ha)
     print(q2_ha)
     print(q3_ha)
```

```
27.75
28.0
32.0
```

STEP 3: Finding outliers of Square Feet column through IQR

```
[ ] q1_sf = sf.quantile(0.25)
    q2_sf = sf.quantile(0.5)
    q3_sf = sf.quantile(0.75)
    iqr_sf = q3_sf - q1_sf
    lower_bound_sf = q1_sf - 1.5 * iqr_sf
    upper_bound_sf = q3_sf + 1.5 * iqr_sf
    #using boolean to choose column in dataframe "Too"
    outliers_sf = too[(sf < lower_bound_sf) | (sf > upper_bound_sf)]
    outliers_sf
```

	House Age	Square Feet	Market Value
13	32	2372	114000.0
14	32	2372	113200.0

```
[ ] print(q1_sf)
    print(q2_sf)
    print(q3_sf)
```

```
1520.0
1666.0
1796.25
```

STEP 4: Finding outliers of Market Value column through IQR

```
[ ] q1_mv = mv.quantile(0.25)
    q2_mv = mv.quantile(0.5)
    q3_mv = mv.quantile(0.75)
    iqr_mv = q3_mv - q1_mv
    lower_bound_mv = q1_mv - 1.5 * iqr_mv
    upper_bound_mv = q3_mv + 1.5 * iqr_mv
    #using boolean to choose column in dataframe "Too"
    outliers_mv = too[(mv < lower_bound_mv) | (mv > upper_bound_mv)]
    outliers_mv
```

	House Age	Square Feet	Market Value
13	32	2372	114000.0
14	32	2372	113200.0
16	32	2123	116100.0
41	27	1581	120700.0
43	27	1581	120700.0

```
[ ] print(q1_mv)
    print(q2_mv)
    print(q3_mv)
```

```
86925.0
88900.0
96700.0
```

Chapter 6. SATISTICAL THINKING IN BUSINESS DECISIONS

6.1. Definition: What is statistical thinking in business decisions?

6.1.1. What is statistical thinking?

Statistical thinking is not about algorithms or equations, or even about data.

Statistical thinking is not solely about algorithms, equations, or data. It is a way of approaching problems and applying statistical techniques. It can be seen as a philosophy and a mindset rather than just a methodology

Statistical thinking has been defined as: "A philosophy of learning and action based on the following fundamental principles:

- All work occurs in a system of interconnected processes.
- Variation exists in all processes
- Understanding and reducing variation are keys to success."

In summary, statistical thinking goes beyond the technical aspects of statistical analysis. It is a way of thinking that recognizes the interconnectedness of processes, acknowledges the presence of variation, and focuses on understanding and reducing variation to drive improvement and success.

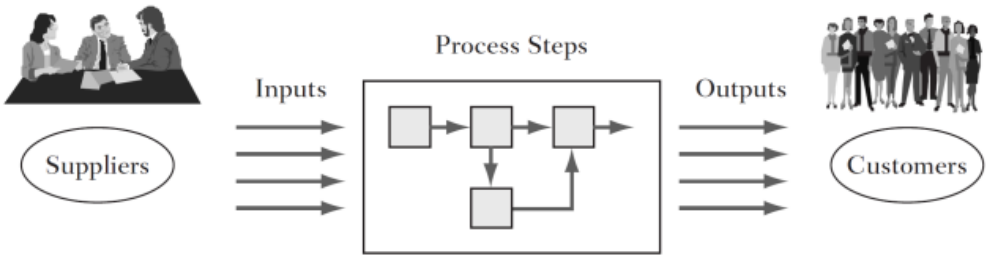
Let’s explain each principle briefly and provide an example for better understanding.

6.1.1.1. All work occurs in a system of interconnected processes

The first principle emphasizes that in any organization, processes are not isolated; they are interconnected and influence one another.

Changes in one area can affect other areas within the system. It's crucial to understand these interconnections to optimize the entire system's performance.

Optimizing one part of the system without taking into consideration the whole would lead to suboptimal outcomes and potential disruptions elsewhere in the system, ultimately compromising the overall efficiency and effectiveness of the entire system.



Process Approach

Example: In an automobile manufacturing plant, various interconnected processes contribute to producing a car. These include design, supply chain management, assembly line, quality control, and distribution.

If there is a delay in the supply chain (perhaps due to a supplier issue), it can directly impact the assembly line, leading to production delays. Understanding this interconnectedness helps in anticipating potential problems and optimizing the workflow across all these processes.

6.1.1.2. Variation exists in all processes

The second principle acknowledges that variation, or differences in processes and outputs, is inevitable. Variation is a fact of life.

Even in the most controlled environments, there will always be natural variability. It's a natural companion, not only in industrial settings, but in all aspects of our life.

Statistical methods are used to understand this variation, differentiating between common cause variation (inherent to the process) and special cause variation (resulting from external factors or anomalies).

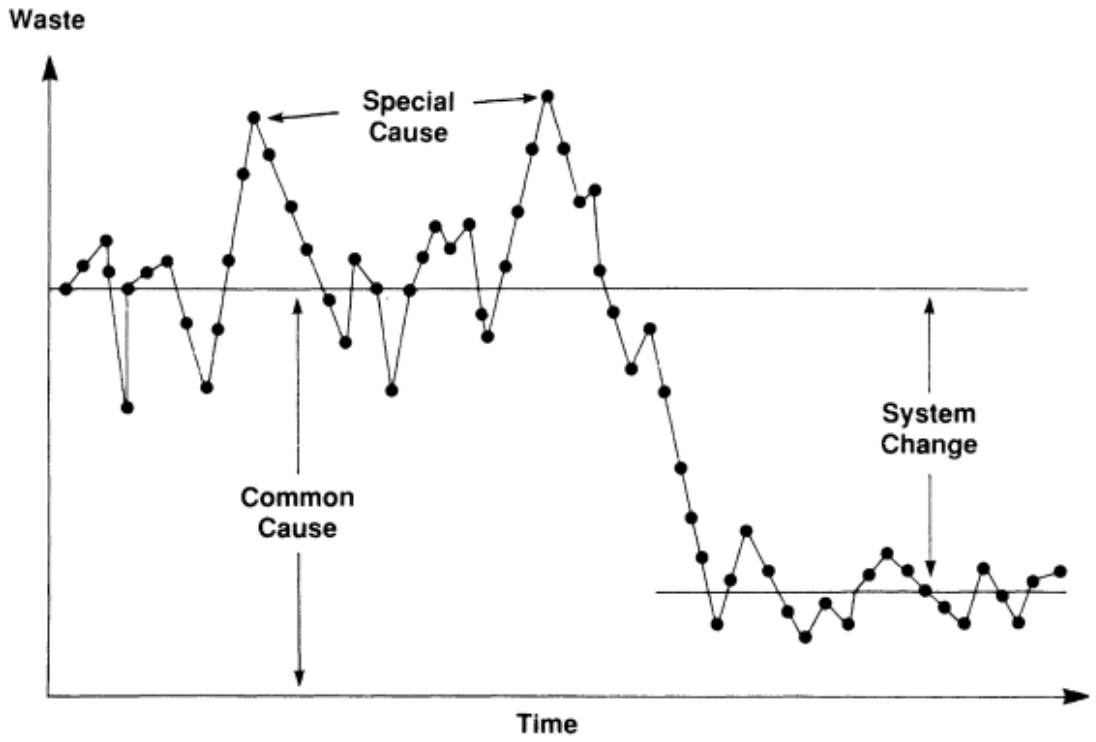


Image 6.1 Special and Common Causes of Variation

Example: In a bottling plant, the amount of soda filled into individual bottles might vary slightly due to factors like temperature, pressure, or even tiny differences in bottle sizes. This natural variation is expected.

It's the responsibility of the plant's team to understand this variation and take the proper decision. For instance, using statistical process control charts, the production manager can determine acceptable variation levels, ensuring that the majority of bottles meet quality standards.

6.1.1.3. Understanding and reducing variation are keys to success

The third principle underscores that understanding the sources of variation and taking steps to minimize abnormal ones is essential for success. By reducing variation, processes become more predictable, leading to improved quality and efficiency.

Example: In a call center, response times to customer inquiries might vary

due to factors such as the complexity of the issue, the experience of the agent, or the efficiency of the software used.

By analyzing these sources of variation, the call center management can implement training programs to enhance agent skills, optimize software interfaces, and standardize procedures.

These efforts reduce variation in response times, leading to more consistent and satisfactory customer experiences.

By applying these principles, organizations can enhance their processes, improve quality, and ultimately achieve greater efficiency and customer satisfaction

6.1.2. The Role of Statistics in Business Decision Making

In this article, we will discuss a few examples in which statistical analysis can be very helpful in a business setting.

- **Identifying Opportunities**

Having access to data can help you gain insights into future opportunities that may arise in your business. Statistics should be used to find new markets, promote better customer retention, increase sales, and identify sales opportunities.

This information can help you make intelligent decisions that will help your company grow and thrive over time.

The analysis of data can also be used to increase efficiency by finding duplication in the market or pinpointing areas that you want to eliminate from your current strategic plan.

- **Understanding customer behavior**

The success of a business depends on the relationships established with its customers. Statistical analysis can help improve customer behavior by looking at their buying patterns and how they use your products or services.

With this information, you can make decisions on the type of products or services you should offer to your customers.

It also helps to identify new opportunities for product development by looking at areas that may require further research and study.

By understanding what they are looking for, you can provide the services or products that will benefit both parties.

- **Determining the correct target market**

As a business, you must have a targeted audience that you are trying to sell your products or services. It is important to identify the best possible choice for your business because all decisions must be made around this key area.

Statistical analysis can help determine whether your current target market is as profitable as it should be. This information allows you to make decisions based on your customer base, which ultimately decides the success of your company.

- **Evaluating products or services**

Statistical analysis provides companies with information about what is being bought and used by consumers. This knowledge can be useful in finding new ways to improve or alter a product or service you offer.

If you are able to determine what your customers are using or how they are accessing your products, you can make the changes that need to be made.

Understanding what features are most important to consumers can help companies create new ideas for upcoming services or products.

- **Making better decisions**

Statistical analysis can provide real evidence of what works and what does not work for your business. This information allows you to make better decisions about business changes, hiring new employees, or how the company should be run.

It also helps with marketing and advertising strategies by giving a better understanding of consumers and what will work best for your company.

All decisions made in a business should be based on the information provided by statistical analysis to ensure that the outcome is positive. The analysis of data will showcase areas that require improvement so you can take action before it becomes too late.

It is always beneficial to have the correct information that can guide you into making the right decisions for your company.

With the advancement of technology, businesses are now able to utilize statistical analysis software tools that allow them to obtain vital information in a shorter amount of time.

Implementing this tool allows business owners and managers the opportunity to make more informed decisions on what they should be doing to increase their efficiency.

Statistical analysis has become a key tool for companies that are trying to grow and thrive in today's market.

Identifying opportunities, understanding customer behavior, determining the correct target market, evaluating your product or service, and making better decisions are just a few of the benefits that companies receive when using statistical analysis.

6.2. Illustration example of statistical thinking in business decision

Example: Company XYZ is a retail company that wants to determine whether implementing a customer loyalty program has a significant impact on customer spending. The company wants to apply statistical thinking to make an informed decision.

STEP 1: **Define the objective**

Company XYZ aims to increase customer spending through a loyalty program. The goal is to assess whether the implementation of the program leads to a significant difference in customer spending compared to the control group.

STEP 2: **Data collection**

Company XYZ collects data on customer spending before and after implementing the loyalty program. They randomly select two groups: **a control group** (without the loyalty program) and **a test group** (with the loyalty program). The data includes the average spending per customer in each group.

Control Group:

Average spending before the program: $100

Standard deviation of spending before the program: $20

Number of customers in the control group: 500

Test Group:

Average spending after the program: $120

Standard deviation of spending after the program: $25

Number of customers in the test group: 550

STEP 3: **Analyze the data**

Using the collected data, Company XYZ can calculate the z-score to assess the significance of the difference in customer spending between the control and test groups.

The formula to calculate the z-score is:

$$z = (x - \mu) / \sigma$$

Where:

x = value being analyzed (average spending in the test group)

μ = mean of the control group (average spending before the program)

σ = standard deviation of the control group (spending before the program)

In this case, the z-score can be calculated as follows:

$$z = (\$120 - \$100) / \$20 = 1$$

STEP 4: **Make a decision**

Based on the calculated z-score, Company XYZ can make a decision. The z score of 1 indicates that the average spending in the test group is one standard deviation higher than the mean of the control group. This suggests a positive impact of the loyalty program on customer spending.

STEP 5: **Evaluate and monitor**

After implementing the loyalty program, Company XYZ continues to monitor customer spending in both the control and test groups. They compare the data and evaluate the effectiveness of the program over time. If the program

consistently shows a positive impact on customer spending, they can continue to invest in and optimize the program.

The statistical thinking applied by Company XYZ allows them to make data driven decisions and assess the impact of the loyalty program on customer spending. By calculating the z-score, they can determine the significance of the difference and make informed choices for their business.

Chapter 7. REFERENCES

[1] Hoerl, R. W., & Snee, R. D. (2012). Statistical Thinking. John Wiley & Sons.

[2] Glossary and Tables for Statistical Quality Control. (2004). Quality Press.

[3] Steiner, S. H., & MacKay, R. J. (2013). Statistical Engineering and Variation Reduction. Quality Engineering, 26(1), 44–60