Bootcamp with olahdata

Statistics & Microsoft Excel 101

olahdata

- Data consultant for high school and universities students
- Email address : olahdata.solution@gmail.com
- Github : olahdata-ai

Speaker Profile

Aurellia Christie

- Co-Founder of Olahdata
- Data Scientist at Supertype
- STEM Tutor Coordinator at Tutor Aja

Table of Contents

01

Statistics

What is Statistics?

Collecting Data
Gathering & measuring information

01

02

Analyzing Data
Cleaning, transforming, & modeling data

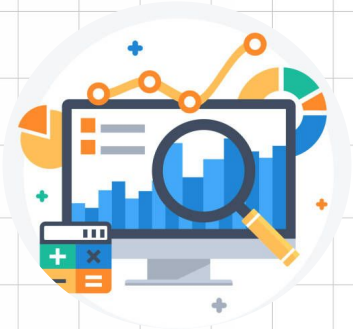
Presenting Data
Using graphical formats to represent data

03

04

Interpreting Data
Developing findings & conclusion

Types of Statistics



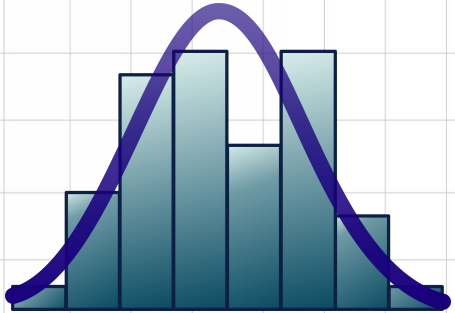
Descriptive Statistics

Describe characteristics of a data using numerical calculation, graphs, or tables

Inferential Statistics

Make decision or prediction about a population based on sample of the data

Types of Statistics



Parametric Statistics

Based on assumptions of the population distribution from which the sample was taken

Non-Parametric Statistics

Not based on the population distribution assumption

Population vs Sample

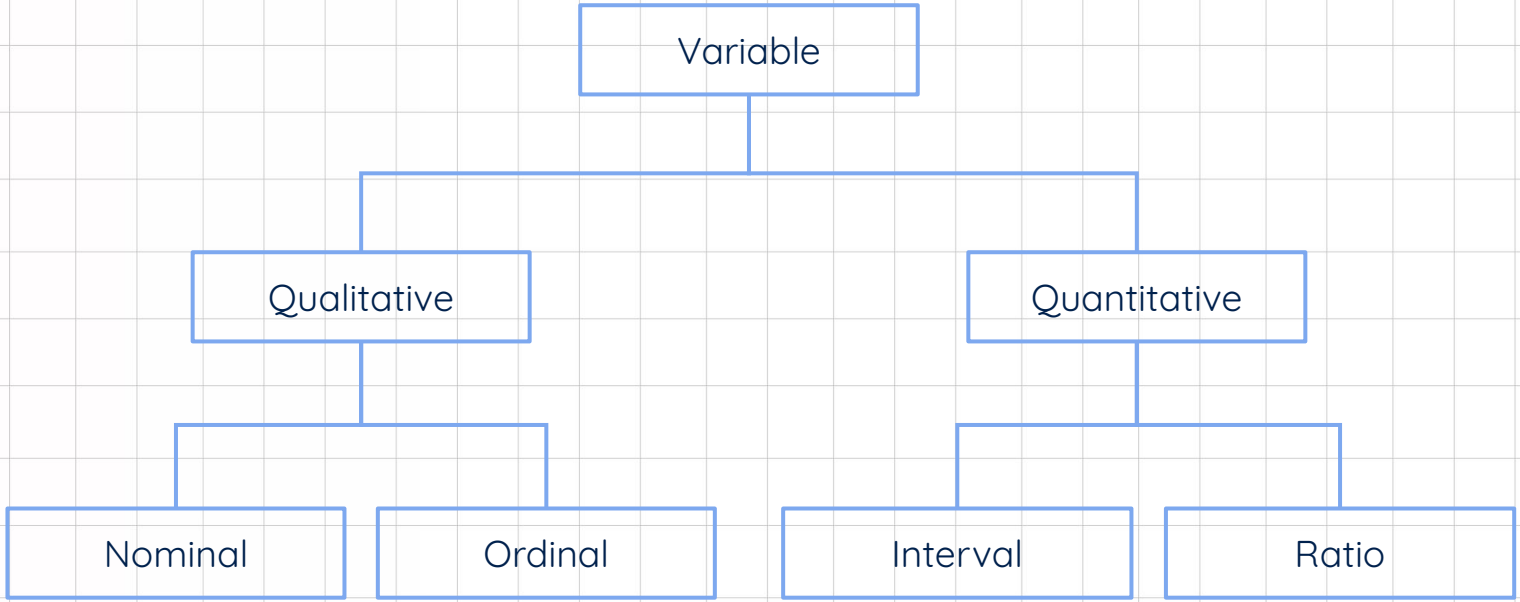
Population

All elements whose characteristics are being studied

Sample

A portion of the population selected for study

Types of Variable



Types of Variable

	Nominal	Ordinal	Interval	Ratio
Have order / ranking		v	v	v
Count / frequency distribution	v	v	v	v
Mode	v	v	v	v
Median		v	v	v
Mean			v	v
Add / subtract value			v	v
Multiple / divide value				v
Has “true zero”				v
Example	Gender	Satisfaction Rating	Temperature	Weight

02

Real World Applications

Today's Case

Data Source : Netflix Original Films & IMDB Scores | Kaggle

NETFLIX

asked us to analyze their original films, documentaries, and specials released as of June 1st, 2021. Give them insights on how does the current business do and what they should improve!

Cleaning Data in Excel

Missing Value

COUNTBLANK()

Select column - Home - Find & Select - Go To Special... - Blank - Ok

1. Right Click - Delete Entire Row - Ok
2. Enter the substitute value - CTRL + Enter

Duplicate Value

1. Select Column – Conditional Formatting – Highlight Cells Rules – Duplicate Values
2. Select Table - Data - Remove Duplicates - Choose Columns - Ok

Data Type

TYPE()

Make sure all columns stored in the correct data type

Text

CLEAN() | TRIM()

UPPER() | LOWER() | PROPER()

Sort, Filter, & Conditional Formatting

Sort

Select column - Home - Sort & Filter

1. Sort A to Z
2. Sort Z to A
3. Custom Sort

Filter

Select column - Home - Sort & Filter - Filter

1. Text Filter
2. Number Filter
3. Date Filter

Conditional Formatting

Home - Conditional Formatting

1. Highlight Cell Rules
2. Top/Bottom Rules
3. Data Bars
4. Color Scales
5. Icon Sets
6. Custom Formatting

Analyzing Qualitative Data

01 Measurement

- Frequency Distribution
- Relative Frequency
- Mode
- Median (Ordinal)

02 Visualization

- Table
- Bar Chart
- Pareto Chart
- Pie Chart

Frequency Distribution

List of all categories and the number of elements belong to each category

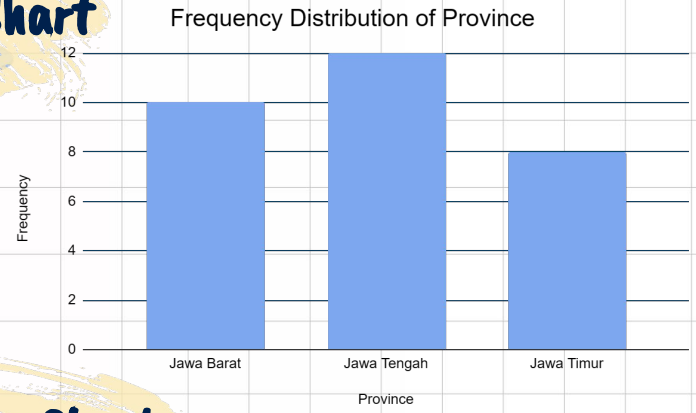
Relative Frequency

$$= \frac{Frequency\ of\ \ the\ category}{Total\ f\ requencies}$$

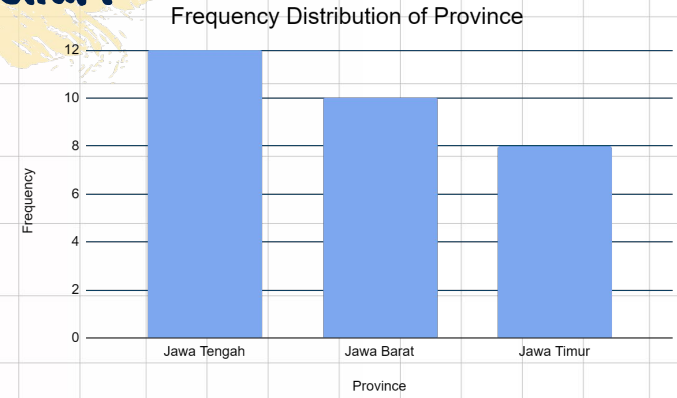
Category	Frequency	Relative Frequency	Percentage
Jawa Barat	10	10/30 = 0.33	0.33 * 100 = 33
Jawa Tengah	12	12/30 = 0.40	0.40 * 100 = 40
Jawa Timur	8	8/30 = 0.27	0.27 * 100 = 27
Total	30	1	100%

Mode

Bar Chart



Pareto Chart



Pie Chart

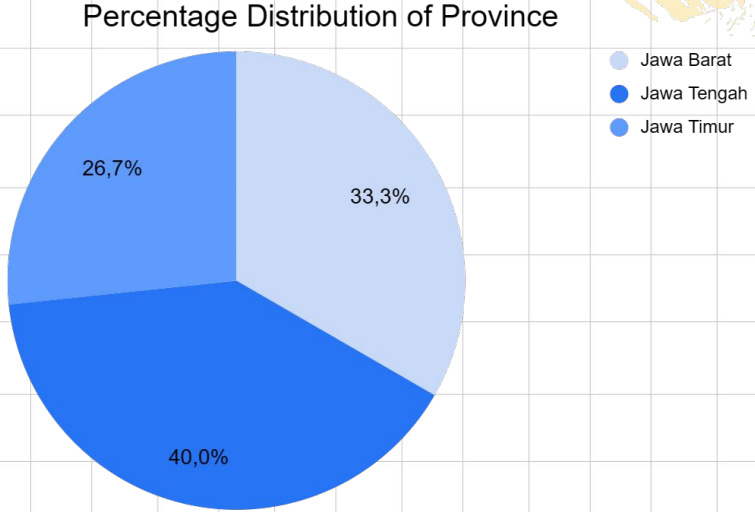
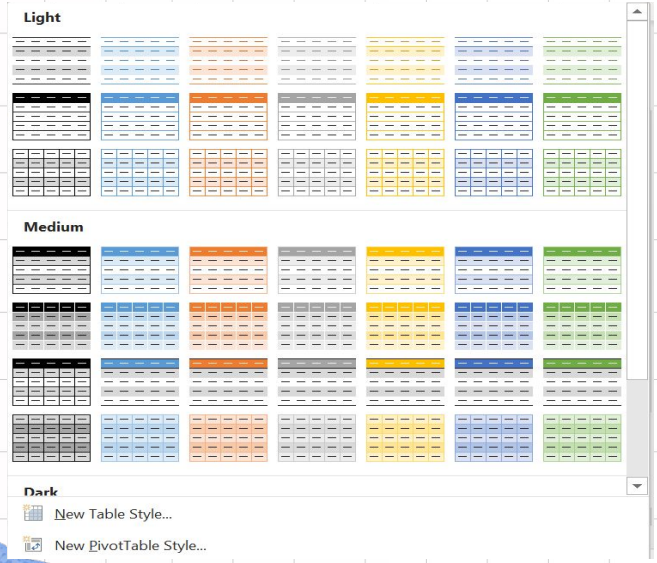


Table & Charts in Excel

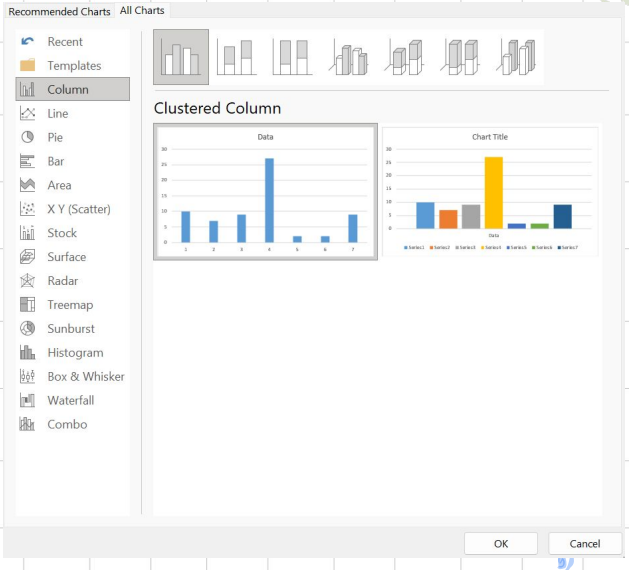
Table

Home - Format as Table



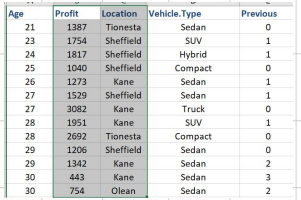
Charts

Insert - Charts



Pivot Table & Pivot Chart

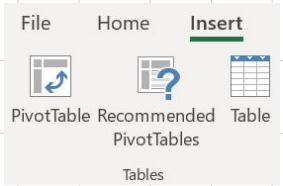
Summarize large amount of data



Age	Profit	Location	Vehicle.Type	Previous
21	1387	Tionesta	Sedan	0
23	1754	Sheffield	SUV	1
24	1817	Sheffield	Hybrid	1
25	1040	Sheffield	Compact	0
26	1273	Kane	Sedan	1
27	1529	Sheffield	Sedan	1
27	3082	Kane	Truck	0
28	1951	Kane	SUV	1
28	2692	Tionesta	Compact	0
29	1206	Sheffield	Sedan	0
29	1342	Kane	Sedan	2
30	443	Kane	Sedan	3
30	754	Olean	Sedan	2

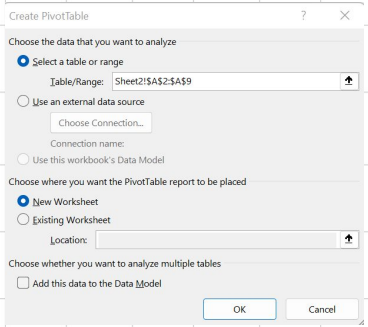
1

Select cells
(No empty values, Single-row heading)



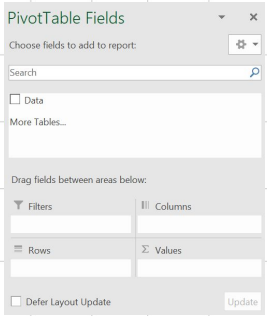
2

Insert - Pivot Table



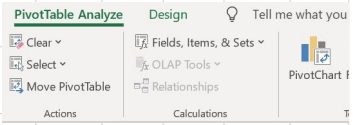
3

Complete the dialog - Ok



4

Custom PivotTable Fields



5

PivotTable Analyze - Pivot Chart

Analyzing Quantitative Data

01

Measurement

Mean | Median | Mode

Range | Standard Deviation

Variance | Mean Deviation | CV

Quartile | IQR | Percentile

02

Visualization

Table | Histogram

Box & Whisker Plot

Line Chart | Scatter Plot

Measures of Central Tendency

Mean
(Non Resistant)

Population:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample:

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

Median
(Resistant)

1. Rank the data in increasing order.
2. Find the value that divides the ranked data set in two equal parts.

NB : If n is even:

- Quantitative :
  Median = mean of the 2 middle values
- Ordinal :
  Median = the lower middle value

Mode
(Resistant)

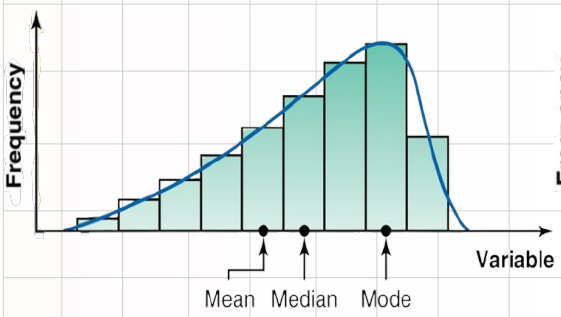
Value that occurs with the highest frequency in a data

- Unimodal
- Bimodal
- Multimodal

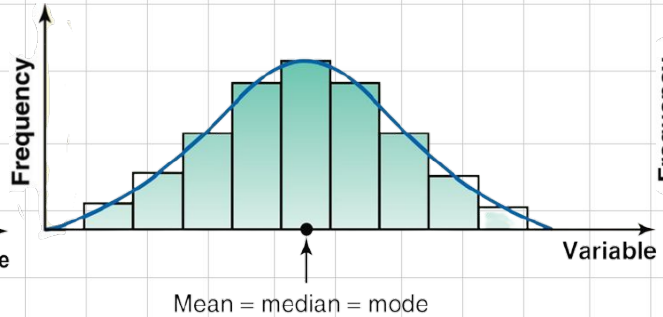
Shortcomings:

- No mode
- More than one mode

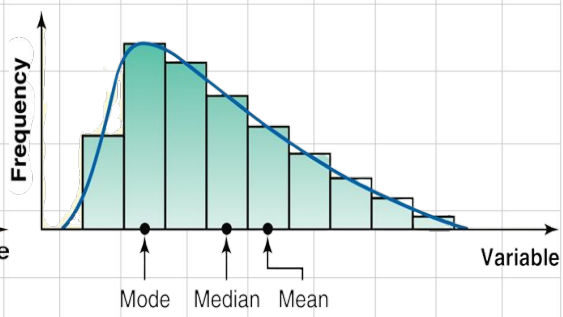
Relationship of Mean, Median, Mode



Skewed to the Left



Symmetric



Skewed to the Right

Measures of Dispersion

Range
(Non Resistant)

Largest Value - Smallest Value

Disadvantages:

- Non resistant
- Only consider largest & smallest value

Variance & Standard Deviation
(Non Resistant)

Population:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(x_i - \mu\right)^2}{N} \qquad \sigma = \sqrt{\sigma^2}$$

Sample:

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

Mean Deviation
(Non Resistant)

$$\frac{\sum_{i=1}^{n} \left|x_i - \bar{x}\right|}{n}$$

Coefficient of Variation (CV)
(Non Resistant)

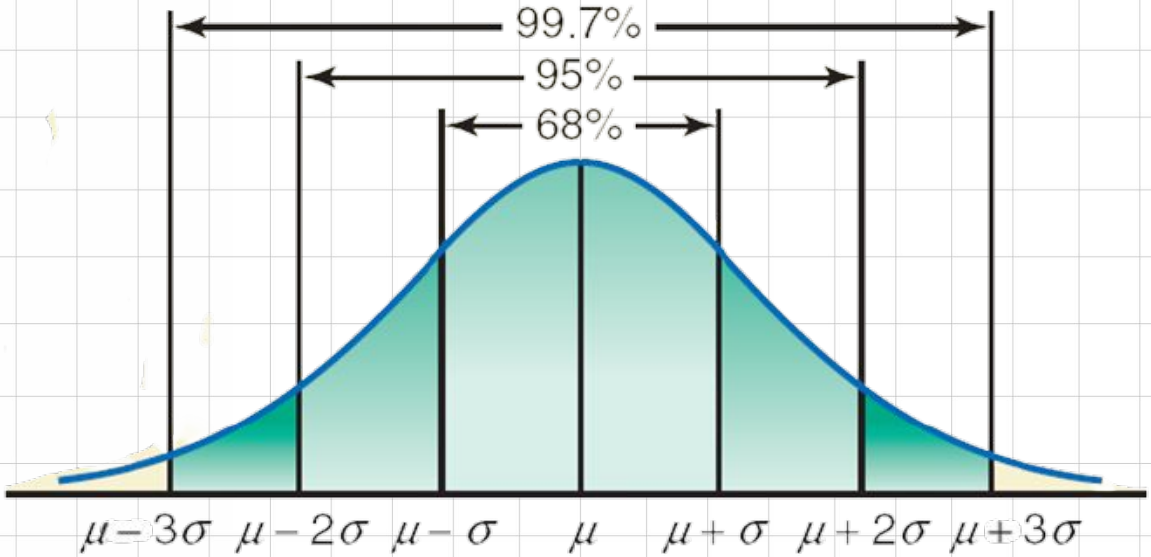
Population:

$$\frac{\sigma}{\mu} \times 100\%$$

Sample:

$$\frac{s}{\bar{x}} \times 100\%$$

Empirical Rule for Symmetric Distribution



Measures of Position

Quartiles

1. Q_2 = Median
2. Q_1 = Median of data that are less than Q_2
3. Q_3 = Median of data that are greater than Q_2

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$

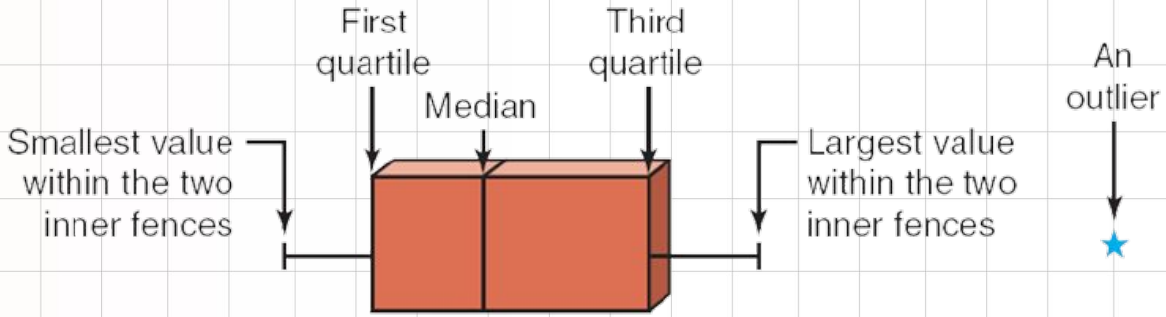
Percentile

k^{th} Percentile = P_k

$$P_k = Value\ of\ the\ \frac{kn}{100}th\ term\ in\ a\ ranked\ data$$

Approximately, k% of the total sample has value that less than or equal to P_k

Detecting Outliers



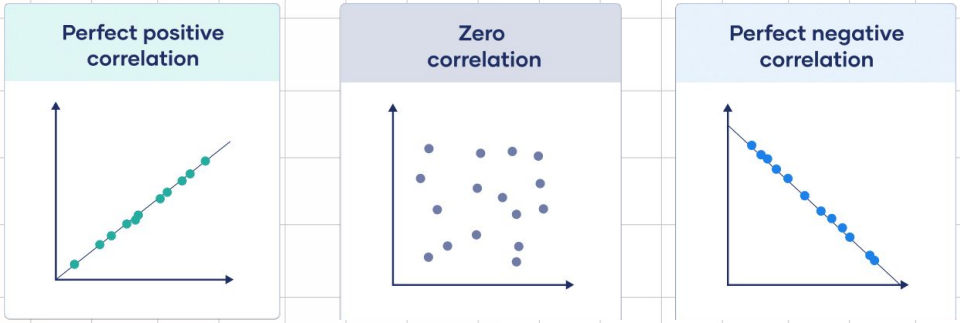
Lower Inner Fence

$$Q_1 - 1.5\,IQR$$

Upper Inner Fence

$$Q_3 + 1.5\,IQR$$

Correlation



Source : Correlation Coefficient | Types, Formulas & Examples (scribbr.com)

Pearson Product Moment Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Statistical Formula in Excel

Central Tendency

AVERAGE()

AVERAGEIF(Range, Criteria, [Average_Range])

MEDIAN()

MODE()

Dispersion

Range = MAX() - MIN()

VAR.P() | VAR.S()

STDEV.P() | STDEV.S()

AVEDEV()

CV = STDEV.S() / AVERAGE()

(+) : SMALL() | LARGE()

Position

QUARTILE(Range, i)

IQR = $Q_3 - Q_1$

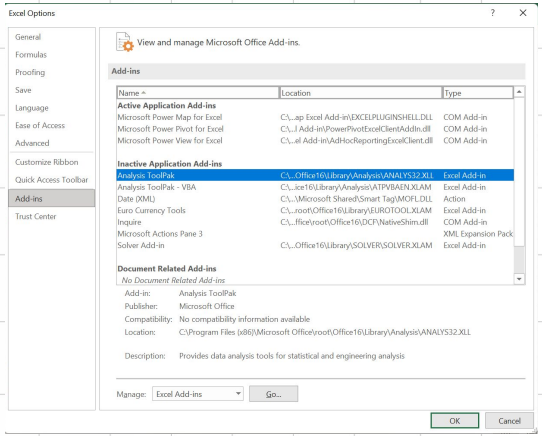
PERCENTILE(Range, k)

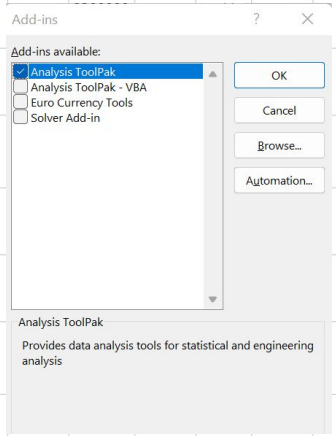
Correlation

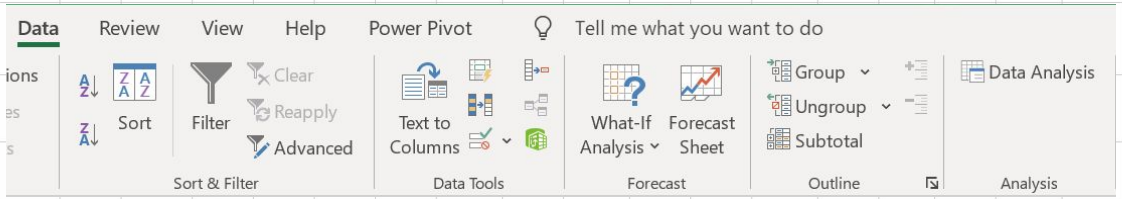
CORREL()

Data Analysis Tool in Excel

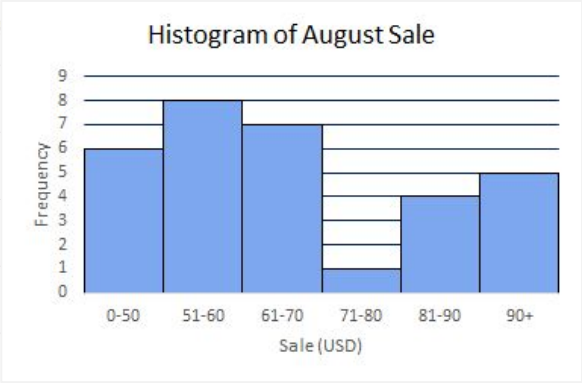
File - More... - Options - Add-ins - Analysis ToolPak - Go - Select - Ok



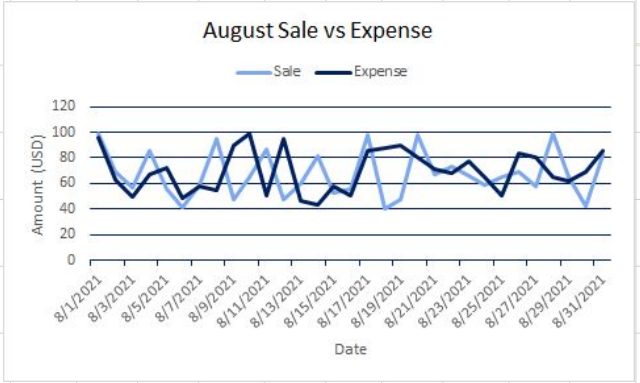




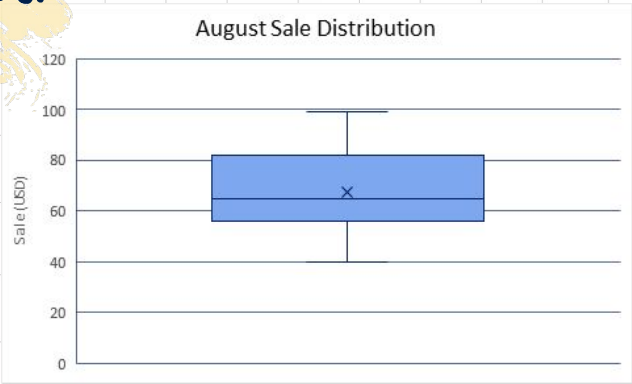
Histogram



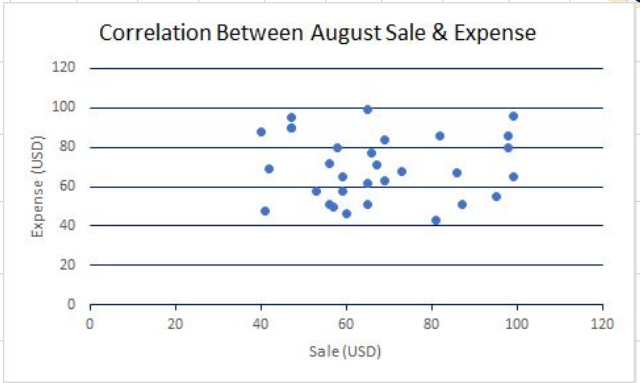
Line Chart



Box & Whisker Plot



Scatter Plot



03

Conclusion

Want to challenge yourself?

www.kahoot.it

- Use your real name to win a prize
- Answer the question as quickly as possible

Good luck!

“The goal is to turn data into information, and information into insight.”

—CARLY FIORNA

References

- Mann, P. (2018). Numerical Descriptive Measures. In *Introductory Statistics* (9th ed.). Wiley.
- Bhandari, P. (2021). *A guide to correlation coefficients.* Scribbr. Retrieved September 28, 2021, from https://www.scribbr.com/statistics/correlation-coefficient/.
- https://support.microsoft.com/en-us/

Thank You!

Do you have any questions?

aurelliachristie77@gmail.com

LinkedIn : Aurellia Christie

Github : AurelliaChristie

Materials link : https://bit.ly/bootcamp-olahdata