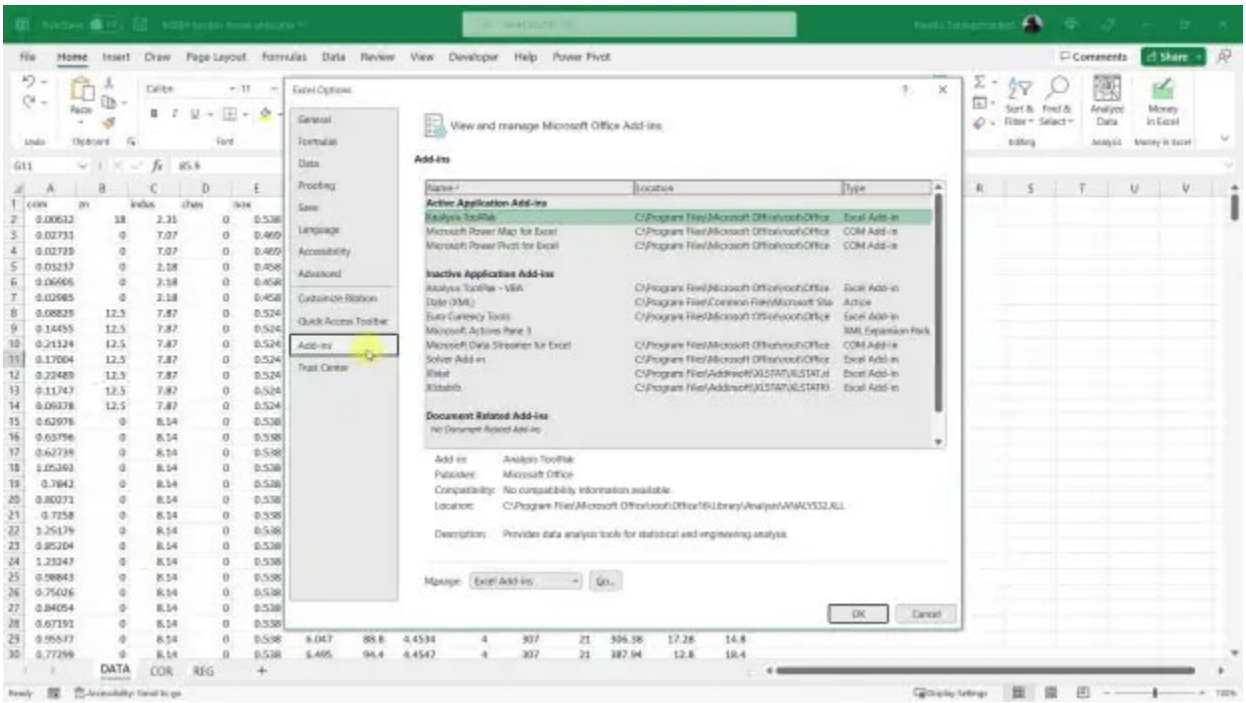
Analysis Toolpak เครื่องมือลับสำหรับงานสถิติใน Excel

บทความนี้แอดเขียนแนะนำวิธีใช้งาน Analysis Toolpak เบื้องต้น ลองคำนวณ descriptive statistics สร้าง correlation matrix และ linear regression ทำนายราคาบ้านใน Boston dataset ง่ายๆ ส่วนตัวแอดใช้ Analysis Toolpak เป็นเครื่องมือหลักเวลาต้องวิเคราะห์ข้อมูลสถิติด้วย Excel อยากเป็น Data Analyst ต้องใช้ Add-in นี้ให้คล่องเลย สำหรับเพื่อนๆที่อยากทำตาม tutorial สามารถโหลดไฟล์ตัวอย่างได้ในลิ้งนี้นะครับ

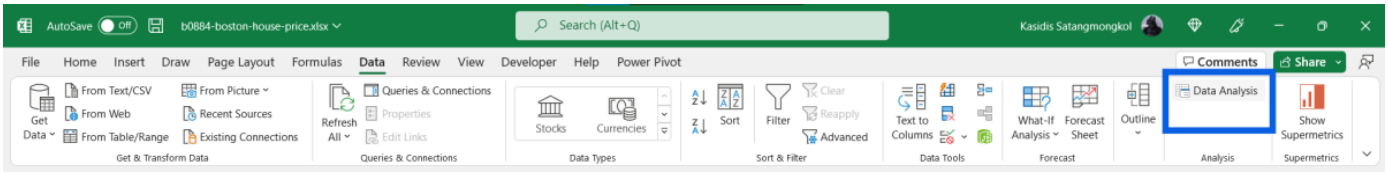
Table of Contents

1. Activate Add-in

วิธีเรียกใช้งาน Analysis Toolpak ให้ไปที่ File > Options แล้วเลือก Add-ins ตามรูปด้านล่าง



เสร็จแล้วจะมีไอคอน Data Analysis โผล่ขึ้นมาที่แท็บ Data พร้อมให้เราใช้งานทางด้านขวาสุดของหน้าจอ



19 modules สถิติของ Analysis Toolpak

1. Anova: Single Factor
2. Anova: Two-Factor with Replication
3. Anova: Two-Factor without Replication
4. Correlation
5. Covariance
6. Descriptive Statistics
7. Exponential Smoothing
8. F-Test
9. Fourier Analysis
10. Histogram
11. Moving Average
12. Random Number Generation
13. Rank and Percentile
14. Regression
15. Sampling
16. t-Test: Paired Two Sample for Means
17. t-Test: Two-Sample Assuming Equal Variances
18. t-Test: Two-Sample Assuming Unequal Variances
19. z-Test: Two Sample for Means

Know Your Dataset

ก่อนที่จะเริ่มวิเคราะห์ข้อมูล เราต้องทำความเข้าใจข้อมูลของเราก่อน EDA – Exploratory Data Analysis สิ่งที่เราควรรู้เกี่ยวกับ dataset มี 4 ข้อ

1. dimension จำนวน row x column ของข้อมูล
2. ข้อมูลมี missing value หรือเปล่า
3. ตัวแปรอะไรเป็น dependent และ independent variables
   3.1 dependent ตัวแปรตาม
   3.2 independent ตัวแปรอิสระหรือตัวแปรต้น
4. ค่าสถิติเบื้องต้น (summary statistics) ของคอลัมน์ที่เราสนใจ

ข้อมูลที่ใช้ใน tutorial นี้ชื่อ Boston มีทั้งหมด 14 columns x 506 records
โดยตัวแปร target หรือ dependent variable ที่เราสนใจคือ medv (median house values) ราคากลางของบ้านในพื้นที่นั้นๆ ส่วนคอลัมน์อื่นๆคือตัวแปรต้นหรือ independent variable ที่เราสามารถเลือกใช้งานได้

วิธีการเช็คง่ายๆว่าข้อมูลมี missing value หรือเปล่า?

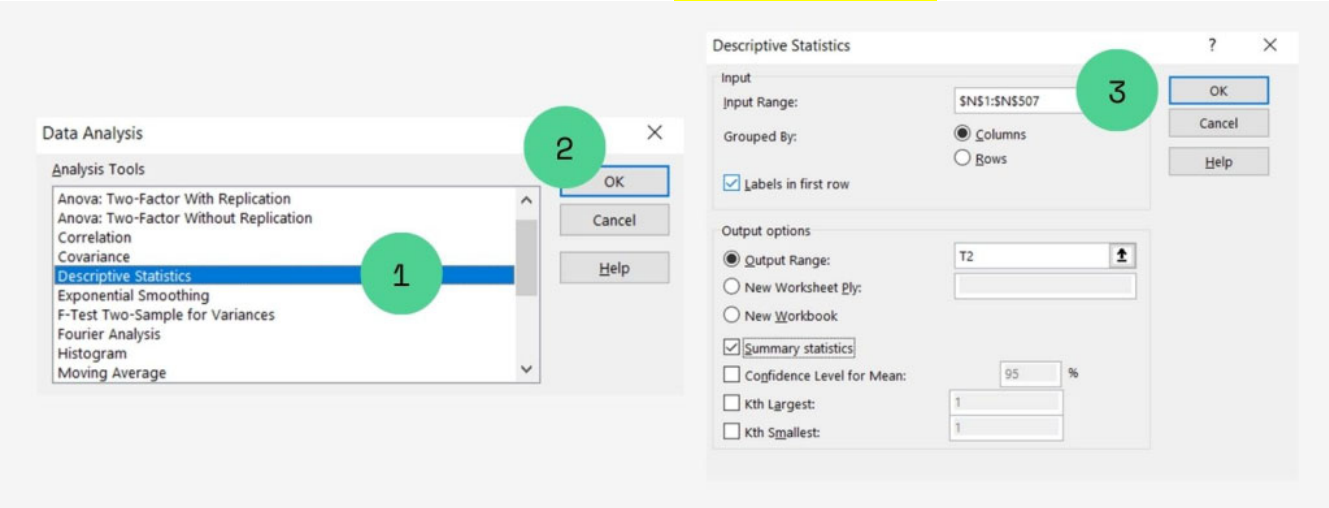
ใน Excel สามารถใช้ฟังชั่น SUM() คู่กับ ISBLANK() เพื่อนับจำนวน cell ที่ไม่มีข้อมูล ถ้าผลลัพธ์ออกมาเท่ากับศูนย์แปลว่าข้อมูลครบ 100% หรือจะใช้ฟังก์ชัน COUNTBLANK() ก็ได้ผลลัพธ์เหมือนกันเลย

=SUM(ISBLANK(A1:N507) * 1)

=COUNTBLANK(A1:N507)

Descriptive Statistics

มาถึงคำถาม EDA ข้อสุดท้าย การหาค่าสถิติเบื้องต้นของคอลัมน์ที่เราสนใจ ถ้าเราใช้ Analysis Toolpak จะช่วยประหยัดเวลาในการเขียน formula เองเยอะมาก คลิกที่ Data > Data Analysis แล้วตั้งค่าตามรูปด้านล่าง



Analysis Toolpak เป็น add-in แบบ drag and drop ไม่ต้องเขียนสูตรอะไรเลย แค่เลือกตัวแปรใส่ในช่องให้ถูกต้องแล้วกด OK เพื่อรันผลได้เลย

1. เลือก Input Range ที่เราสนใจ ในตัวอย่างด้านบนคือคอลัมน์ N (ตัวแปร medv)
2. ถ้า row ที่หนึ่งของ dataset เป็นชื่อคอลัมน์ให้เราเลือก Labels in first row
3. เลือก Output Range ว่าเราอยากเอาผลสถิติไปแปะที่ cell ไหนใน Excel worksheet

Output ที่ได้จาก descriptive statistics มีค่าสถิติสำคัญที่เราใช้บ่อยๆ เช่น mean, median, mode, sd, variance เป็นต้น (ถ้าให้เขียนสูตรเองหมดนี้ใช้เวลาไม่ต่ำกว่า 5-10 นาที)

medv	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

ตัวอย่าง functions ถ้าเราต้องเขียนเอง แค่คิดก็เหนื่อยแล้ว

=AVERAGE()

=STDEV()

=MEDIAN()

=MODE.SNGL()

=MIN()

=MAX()

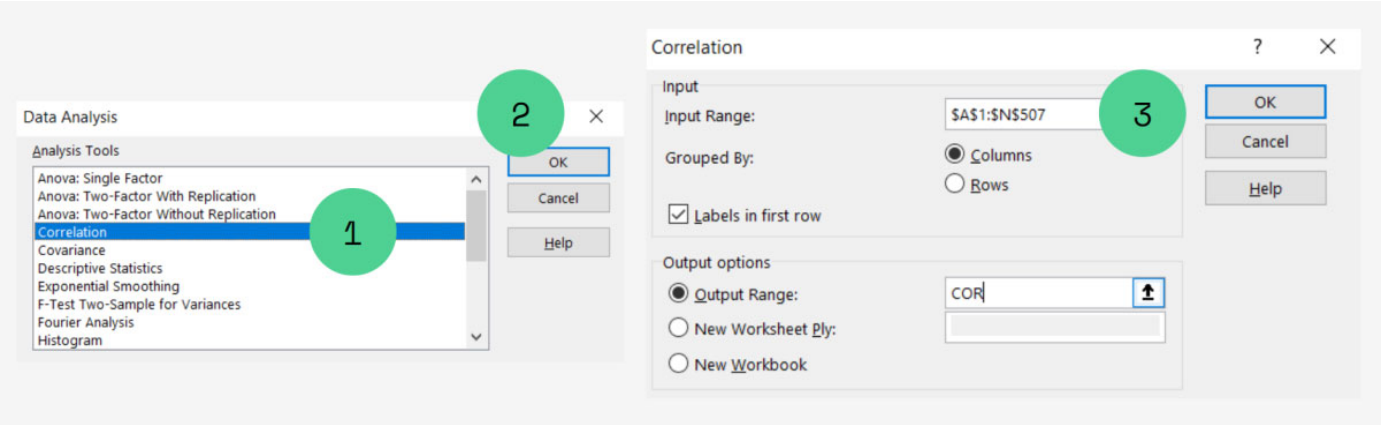
=MAX()-MIN() // Range

=SUM()

=COUNT()

Correlation

วิธีสร้าง correlation matrix ให้กลับไปที่ Data Analysis แล้วเลือกเมนู Correlation ในช่อง input range ให้เลือกข้อมูลของเราทั้งหมดเลย A1:N507 แล้วเซฟ output ใน worksheet ใหม่ตั้งชื่อว่า COR



Correlation matrix ใน analysis toolpak

เราจะได้ตาราง correlation matrix มาหนึ่งตาราง เสร็จแล้ว ง่ายเหลือเชื่อ ค่า correlation จะมีค่าวิ่งอยู่ระหว่าง [-1, +1] เครื่องหมายบวกแปลว่าตัวแปรสองตัวเปลี่ยนแปลงในทิศทางเดียวกัน i.e. x เพิ่ม y เพิ่ม ส่วนเครื่องหมายลบคือเปลี่ยนแปลงในทิศทางตรงกันข้ามกัน ตัวแปรที่มีความสัมพันธ์สูงที่สุดกับ medv คือ rm (จำนวนห้อง) มีค่า correlation = 0.69536

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		*crim*	*zn*	*indus*	*chas*	*nox*	*rm*	*age*	*dis*	*rad*	*tax*	*ptratio*	*black*	*lstat*	*medv*
2	crim	1													
3	zn	-0.20047	1												
4	indus	0.40658	-0.53383	1											
5	chas	-0.05589	-0.0427	0.06294	1										
6	nox	0.42097	-0.5166	0.76365	0.0912	1									
7	rm	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1								
8	age	0.35273	-0.56954	0.64478	0.08652	0.73147	-0.24026	1							
9	dis	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1						
10	rad	0.62551	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1					
11	tax	0.58276	-0.31456	0.72076	-0.03559	0.66802	-0.29205	0.50646	-0.53443	0.91023	1				
12	ptratio	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.3555	0.26152	-0.23247	0.46474	0.46085	1			
13	black	-0.38506	0.17552	-0.35698	0.04879	-0.38005	0.12807	-0.27353	0.29151	-0.44441	-0.44181	-0.17738	1		
14	lstat	0.45562	-0.41299	0.6038	-0.05393	0.59088	-0.61381	0.60234	-0.497	0.48868	0.54399	0.37404	-0.36609	1	
15	medv	-0.3883	0.36045	-0.48373	0.17526	-0.42732	0.69536	-0.37695	0.24993	-0.38163	-0.46854	-0.50779	0.33346	-0.73766	1

ตาราง correlation matrix

นักสถิติใช้ correlation matrix ในการวิเคราะห์ความสัมพันธ์ของตัวแปรเชิงปริมาณ (quantitative data) เราสามารถเรียง correlation coefficients ของตัวแปรที่สนใจ หรือใช้ conditional formatting ไฮไลท์สีแบบ heatmap ก็ได้ i.e. คะแนนสูงสีเข้ม คะแนนต่ำสีอ่อน เป็นต้น

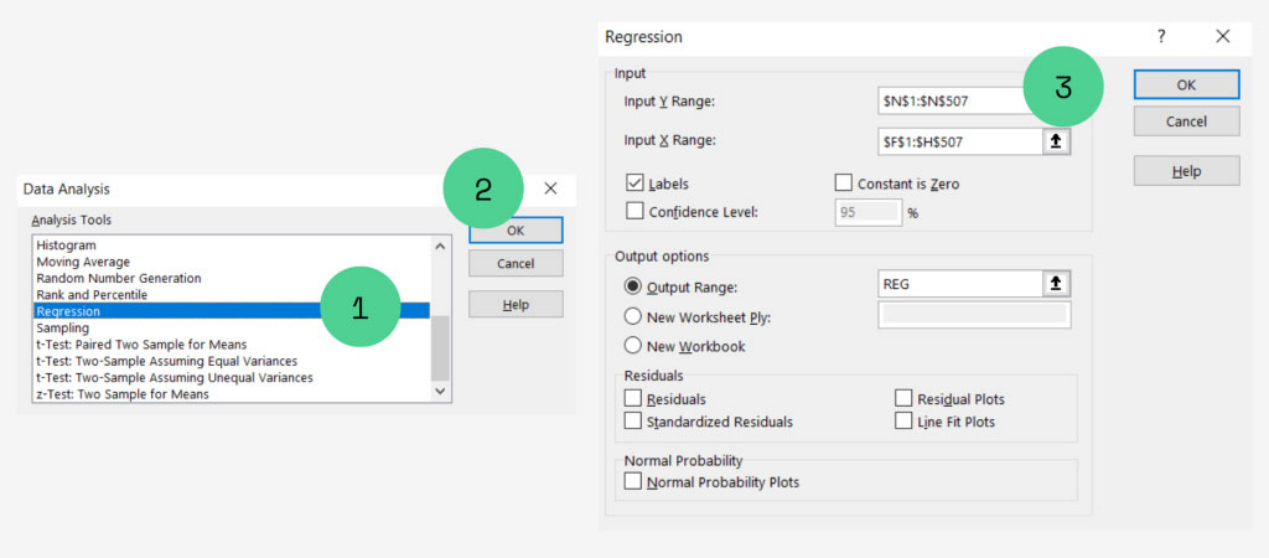
Linear Regression

โมเดลสุดท้าย มาลองสร้างโมเดล linear regression กันบ้าง เราจะเลือกตัวแปรต้นสามตัวคือ rm age dis มาใช้ทำนายราคาบ้าน medv หน้าตาของสมการที่เราอยากได้เป็นแบบนี้

medv = f(rm, age, dis)

medv = b0 + b1*rm + b2*age + b3*dis

กลับเข้าไปที่ Data Analysis > Regression เลือก input range ตามรูปด้านล่าง



Linear regression ใน analysis toolpak

หน้าตาของ output ที่ได้จากเมนู regression จะเหมือนกับโปรแกรม IBM SPSS ที่ใช้กันเยอะๆใน มหาวิทยาลัย จ่ายค่าลิขสิทธิ์กันแพงเลย จริงๆทำใน Excel ก็ได้ ยั๊ง!

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.731845523					
5	R Square	0.53559787					
6	Adjusted R Square	0.532822558					
7	Standard Error	6.286255573					
8	Observations	506					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	22878.75683	7626.252278	192.9865758	3.22515E-83	
13	Residual	502	19837.53858	39.51700913			
14	Total	505	42716.29542				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-21.87278779	3.17795533	-6.882660554	0.0000	-28.11651931	-15.62905627
18	rm	8.440629098	0.410469711	20.56334213	0.0000	7.63417891	9.247079286
19	age	-0.099418235	0.015105467	-6.581606143	0.0000	-0.129095959	-0.069740511
20	dis	-0.480384091	0.200276749	-2.398601398	0.0168	-0.873867989	-0.086900192

Linear regression output

Linear regression ที่เราเพิ่งสร้างขึ้นมามีค่า R Square = 0.5355 ตัวแปร rm age dis มีนัยสำคัญที่ระดับ alpha = 0.05 (p-value < 0.05) หน้าตาของ final model เขียนได้แบบนี้

medv = -21.87 + 8.44*rm + (-0.09)*age + (-0.48)*dis

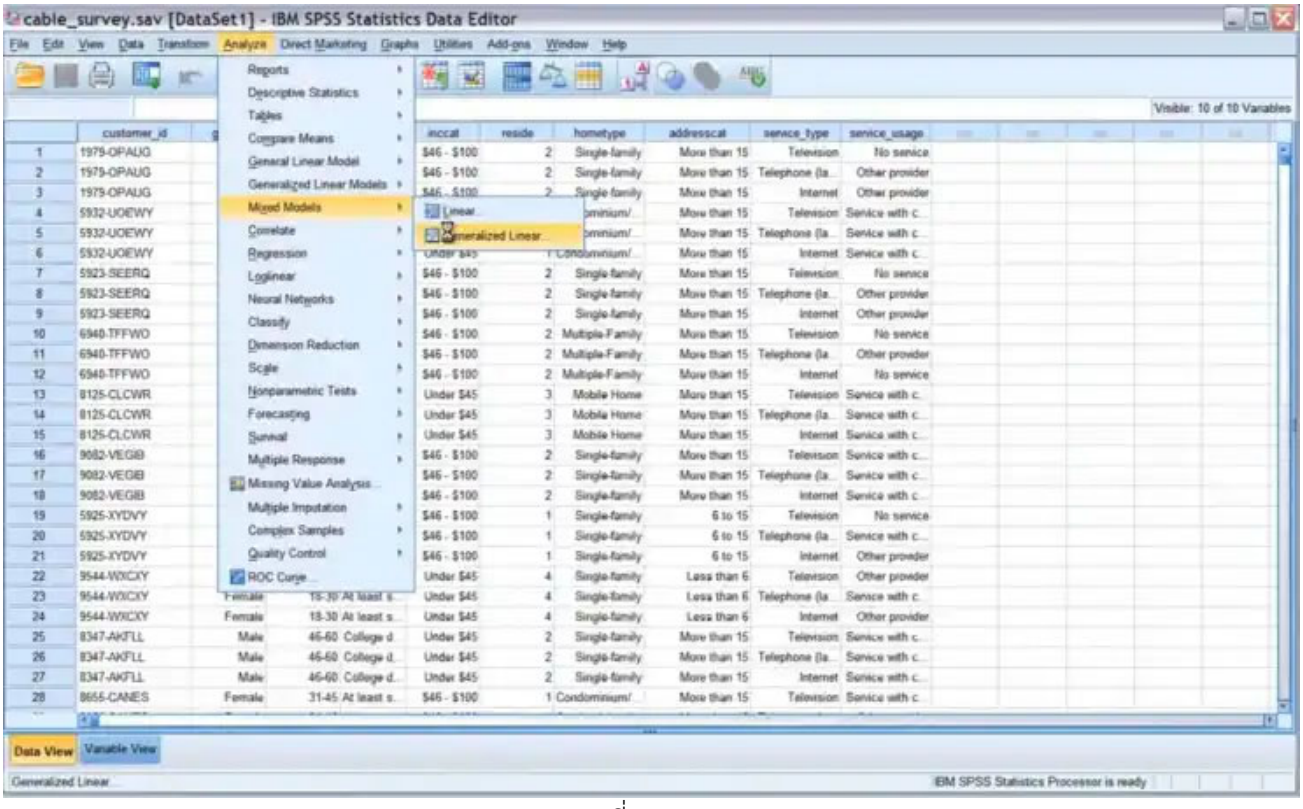
R Square = 0.5355, F = 192.9865, p-value = 0.0000

**Note – เราสามารถใช้ function =LINEST() เพื่อหา regression coefficients และค่าสถิติอื่นๆ เช่น R Square, F และ p-value ได้เหมือนกัน แต่การแสดงผลจะไม่สวยเหมือน Analysis Toolpak

Analysis Toolpak vs. SPSS

ถ้าใครเรียนสาย social science, marketing, business & economics หรือเก็บพวกแบบสอบถามมาวิเคราะห์ ตอนอยู่มหาวิทยาลัยน่าจะเคยผ่านโปรแกรม IBM SPSS กันมาบ้าง

SPSS ย่อมาจาก Statistical Package for the Social Sciences เป็นซอฟต์แวร์สำหรับวิเคราะห์ข้อมูลสถิติ ราคาค่อนข้างสูง ถ้าซื้อตัวเต็ม full features ปีละเป็นแสนบาท (commercial use)



IBM SPSS ที่มา IBM Website

ปัจจุบัน users ใช้น้อยลงเยอะเพราะมี open-source software อย่าง R, Python หรือแม้แต่ Excel, Google Sheets ที่ใช้แทนกันได้ ปกติ IBM จะออกเวอร์ชันใหม่ทุกปี แต่นี้จะสองปีแล้วยังค้างที่ version 28 อยู่ 555+

Analysis Toolpak ที่เราสอนในบทความนี้ทำหลายๆอย่างได้เหมือน SPSS เลย ถ้าใครต้องทำงานวิจัย ป.ตรี/โทในมหาวิทยาลัย แอดว่าใช้ Analysis Toolpak ให้คล่องๆก็เพียงพอให้เรียนจบได้สบายๆแล้วครับ

Key Takeaway

ก่อนเริ่มวิเคราะห์ข้อมูล ควรทำ EDA เพื่อเข้าใจโครงสร้างข้อมูลเบื้องต้น

- Analysis Toolpak เป็นเครื่องมือสำคัญที่ Data Analyst ควรฝึกใช้ให้คล่อง
  - Descriptive Statistics
  - Correlation
  - Linear Regression
- การใช้งานเป็นแบบ drag and drop ไม่ต้องเขียนสูตรให้ยุ่งยาก
- สามารถรันโมเดลได้ 19 แบบ มีครบทุกตัวพื้นฐาน t-test, one-way ANOVA, correlation และ linear regression ตอบโจทย์สำหรับคนที่ต้องทำ IS/ Thesis ระดับปริญญาตรี-โท