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 นี้ให้ คล่องเลย สำหรับเพื่อนๆที่อยากทำตาม tutorial สามารถโหลดไฟล์ตัวอย่างได้ในลิ้งนี้นะครับ

Table of Contents

- 1. Activate Add-in
- 2. Know Your Dataset
- 3. Descriptive Statistics
- 4. Correlation
- 5. Linear Regression
- 6. Analysis Toolpak vs. SPSS
- 7. Good Book
- 8. Key Takeaway

1. Activate Add-in

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

tis Hot	tracet se	Draw P	lage Layor	at farm	ulies Date	Berner	View Develop	ser Help	Power Pre	30A								Come	ents [t Share -
17 - 1 (* - 1) 1040 10	ice D -	Crite II I	- 12 - 13 Hort 16	n - - & -	General General Formulae Data	6	View of Add-Ins	id manage N	icrosoft O	ttice Add-in	w.			3	r x	Σ- 	Sint & To Ritter - Sel			Horney 1 Excel ey in Sacal
		-	0		Prophers		Rinter			him				Paris .	-				1.14	1 44
a la	a k	dar de		E			Arthur Annih	ation Add.ins		hoo				1068	- 1		2	1	0	v
(199 0.00632 0.00733 0.02739 0.02739 0.02739 0.028237 0.02825 0.02825 0.02825 0.02825 0.1455 0.21134 0.17004 0.213489 0.12044 0.213489 0.12044 0.213489 0.12044 0.213578 0.02378	20 30 40 40 40 40 40 40 40 40 40 4	605 07 2.31 7.07 7.07 2.58 3.58 2.58 7.87 7.87 7.87 7.87 7.87 7.87 7.87 7		0.538 0.469 0.469 0.469 0.468 0.468 0.458 0.524 0.524 0.524 0.524 0.524 0.524 0.524 0.524 0.524 0.524	See Language Accossibility Advasced Castancies Outbridge Addens Trait Center	a Toolber	Radyan Tool Microsoft Per Microsoft Per Microsoft Per Institute Appl Analysis Tool Date (Microsoft Out Softer Add in Bland Nicrosoft Out Softer Add in Bland Distantib Document Ri Interpret	Annual An	i i bort		ogram Navi ogram Ravi ogram Ravi ogram Ravi ogram Ravi ogram Ravi ogram Ravi ogram Ravi ogram Ravi	Accessit 284 Accessit 284 Acces	Nicolo Office since Office	Bool Add-in COM Add-in COM Add-in Add-in Addo- Sace Add-in Bool Add-in Dool Add-in Dool Add-in Dool Add-in	n Pork.					
1.05363 0.7843 0.7858 1.25176 0.85304 1.23347		8.14 8.14 8.14 8.14 8.14 8.14 8.14	0 0 0 0 0	0.538 0.538 0.538 0.538 0.538 0.538 0.538			Add en Publisher Compartiell Location Description	Analysis To Microsoft (No compat C*Ptogram Provider do	offiae effice ability infor Plian/Africe fa analysis	mation availa avoit Officiala tools for stat	ible oof:DMice16 idical and an	Libriey' Analys pinoving anal	n/WMACYSS2.X							
0.58843 0.75626	0 0	8.54 8.54	0 0	0.538 0.538			Manager (Er	gi Add-ins	- 1	ùo										
0.84054	0	8.14 8.14	0	0.538									0	ÛK.	Danist	-				
0.96537		8.14	0	0.538	8.047	88.8	4,4534 4	907	21	305.38	17.28	14.8								
the second se		10.14		0.536	E 405	94.4	44543	302	24	287.54	12.6	10.4								

้เสร็จแล้วจะมีไอคอน Data Analysis โผล่ขึ้นมาที่แท็บ Data พร้อมให้เราใช้งานทางด้านขวาสุดของหน้าจอ

AutoSave Off 🖫 b0884-boston-house-price		, ♀ Search (Alt+Q)			Kasidis Satangmongkol 🦸) 🕈 🖉		
File Home Insert Draw Page Layout Fo	rmulas Data Review View	Developer Help Power Pivot	:			Comments	년 Share 🗸	8
Image: Construction Image: Construction Image: Construction Get Image: Construction Image: Construction Image: Construction Data Image: Construction Image: Construction Image: Construction Image: Construction	Refresh All ~ B Edit Links	Stocks Currencies	Â↓ XA Â↓ XA XA Filter V Advanced	Text to Columns 🗞 v	What-If Forecast Analysis Sheet	a Data Analysis	Show Supermetrics	
Get & Transform Data	Queries & Connections	Data Types	Sort & Filter	Data Tools	Forecast	Analysis	Supermetrics	\sim

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

้ช้อมูลที่ใช้ใน tutorial นี้ชื่อ Boston มีทั้งหมด 14 columns x 506 records

โดยตัวแปร target หรือ dependent variable ที่เราสนใจคือ medv (median house values) ราคากลางของบ้าน ในพื้นที่นั้นๆ ส่วนคอลัมน์อื่นๆคือตัวแปรต้นหรือ independent variable ที่เราสามารถเลือกใช้งานได้

วิธีการเช็คง่ายๆว่าข้อมูลมี missing value หรือเปล่า?

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

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

=COUNTBLANK(A1:N507)

Descriptive Statistics

มาถึงคำถาม EDA ข้อสุดท้าย การหาค่าสถิติเบื้องต้นของคอลัมน์ที่เราสนใจ ถ้าเราใช้ Analysis Toolpak จะ ช่วยประหยัดเวลาในการเขียน formula เองเยอะมาก คลิกที่ <mark>Data > Data Analysis</mark> แล้วตั้งค่าตามรูปด้านล่าง

		Descriptive Statistics			?	3
		Input Input Range:	\$N\$1:\$N\$507	3	ОК	
ata Analysis	2 ×	Grouped By:	() <u>C</u> olumns		Cancel	
analysis Tools	OK	I abels in first row	Bows		Help	
Anova: Two-Factor With Replication Anova: Two-Factor Without Replication Covariance Descriptive-Statistics Exponential Smoothing F-Test Two-Sample for Variances Fourier Analysis Histogram Moving Average	Cancel Help	Output options © Qutput Range: New Worksheet Ply: New Worksheet Ply: New Workbook Summary statistics Confidence Level for Mean:	95	<u>*</u>		
		Kth Largest:	1			

Analysis Toolpak เป็น add-in แบบ drag and drop ไม่ต้องเขียนสูตรอะไรเลย แค่เลือกตัวแปรใส่ในช่องให้ ถูกต้องแล้วกด 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		
		ตัวอย่าง functions ถ้าเราต้องเขียนเอง แค่คิดก็เหนือยแล้ว
Mean	22.53281	=AVERAGE()
Standard Error	0.408861	
Median	21.2	=SIDEV()
Mode	50	=MEDIAN()
Standard Deviation	9.197104	
Sample Variance	84.58672	
Kurtosis	1.495197	=MIN()
Skewness	1.108098	=MAX()
Range	45	
Minimum	5	=MAX()-MIN() // Range
Maximum	50	=SUM()
Sum	11401.6	
Count	506	

Correlation

วิธีสร้าง correlation matrix ให้กลับไปที่ <mark>Data Analysis</mark> แล้วเลือกเมนู <mark>Correlation ใ</mark>นช่อง input range ให้เลือกข้อมูลของเราทั้งหมดเลย A1:N507 แล้วเซฟ output ใน worksheet ใหม่ตั้งชื่อว่า COR

	Correlation		?	×
Х	Input Input Range: Grouped By:	SA\$1:SN\$507 3	OK Cancel	
Cancel <u>H</u> elp	Labels in first row	O <u>R</u> ows	Help	
	Output options	COR		
	Х Ок Cancel Help	Correlation Correlation Input Input Input Range: Grouped By: Labels in first row Output options Qutput Range: New Worksheet Ply:	Correlation Input	Correlation ? Input Input Input Input Input Input Range: \$A\$1:\$N\$507 Grouped By: Grouped By: Labels in first row ULabels in first row Utput options Qutput Range: New Worksheet <u>P</u> ly: Cork

Correlation matrix $\ensuremath{\mathfrak{lu}}$ analysis toolpak

เราจะได้ตาราง correlation matrix มาหนึ่งตาราง เสร็จแล้ว ง่ายเหลือเชื่อ ค่า correlation จะมีค่าวิ่งอยู่ ระหว่าง [-1, +1] เครื่องหมายบวกแปลว่าตัวแปรสองตัวเปลี่ยนแปลงในทิศทางเดียวกัน i.e. x เพิ่ม y เพิ่ม ส่วน เครื่องหมายลบคือเปลี่ยนแปลงในทิศทางตรงกันข้ามกัน ตัวแปรที่มีความสัมพันธ์สูงที่สุดกับ medv คือ rm (จำนวน ห้อง) มีค่า correlation = 0.69536

	А	В	С	D	E	F	G	Н	1	J	К	L	М	Ν	0
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	Istat	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 <mark>rm</mark> age dis มา ใช้ทำนายราคาบ้าน <mark>medv</mark> หน้าตาของสมการที่เราอยากได้เป็นแบบนี้

medv = f(rm, age, dis)

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

กลับเข้าไปที่ <mark>Data Analysis > Regression</mark> เลือก input range ตามรูปด้านล่าง



Linear regression ใน analysis toolpak

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

	A	В	С	D	E	F	G
1	SUMMARY OUTPU	Т					
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 <mark>rm age dis</mark> 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 =<mark>LINEST()</mark> 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 หรือเก็บพวกแบบสอบถามมา วิเคราะห์ ตอนอยู่มหาวิทยาลัยน่าจะเคยผ่านโปรแกรม IBM SPSS กันมาบ้าง

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

ile Edt	Yorn Data Installar	Analyza Drect Marketing Grap	ha Utilities Add-gos	W	indow <u>H</u> ielp								NAME AND A
2	1 🗎 📖 🖛	Argorts Descriptive Statistics		4		A	6					40.00	
_		Tables •	-	-							-	VINDER.	TO OF TO Variat
	customer_id	Goggan Means	inccal resid		hometype	addresscal	service_type	service_usage		100	-		
1	1979-OPADO	General Linear Model	346-3100		Single-family	More than 15	Television	FED SHEVICE					
2	1979-OPAUG	Generalized Linear Models +	\$46 - \$100	- 2	Single-formity	More than 15	Lelephone (la	Other provider					
3	T979-OPADG	Minut Madela +	LMh-Alie	- 2	Sungre tomity	Mone than 15	Internet	Other provider					
4	SAUS-ODEMA	Complete	in frees.		pminium/	More than 15	Television	Service with c					
3	5932-UOEWY	Zoneske	Constalized Line	Ø	pmmum/	More than 15	Telephone (la	Service with c					
	S305-DOEMA	Begression *	Under \$45		Consumnum/	More than 15	Internet	Service with c					
1	5923-SEERQ	Loginear *	546 - \$100	2	Single family	More than 15	Television	The service					
-	SR23-SEERQ	Neural Networks P	546 - 5100	2	Single family	Move than 15	Telephone (la	Other provider					
9	5923-SEERQ	Classify +	\$46 - \$100	2	Single-family	More than 15	internet	Other provider					
10	6940-TFFWO	Dynamision Reduction	\$46 - \$100	2	Mutiple-Family	More than 15	Television	No service					
11	6940-TFFWO	Scale A	\$46-\$100	2	Multiple-Family	More than 15	Telephone (La	Other provider					
12	6940-TFFWO	Noncomputing Topolo	\$46-5100	2	Multiple-Family	More than 15	Internet	No service					
13	8125-CLCWR	Hosparametric rests	Under \$45	3	Mobile Home	More than 15	Television	Service with c.					
- 14	0125-CLCWR	Forecasting *	Under \$45	3	Mobile Home	More than 15	Telephone (la	Service with c					
15	B125-CLCWR	Sunnal	Under \$45	3	Mobile Home	More than 15	Internet	Service with c					
16	9082-VEGI0	Multiple Response >	\$46 - \$100	2	Single-family	More than 15	Television	Senice with c					
17	9082-VEGIB	ED Missing Value Analysis	\$46 - \$100	2	Single-family	More than 15	Telephone (la	Service with c					
18	9082-VEGIB	Multiple Importation A	\$46 - \$100	2	Single-family	More than 15	internet	Senice with c					
19	5925-XYDVY	Consultar Consultar	\$46 - \$100	1	Single-family	6.10 15	Television	No service					
20	5925-XYDVY	Complex Samples	\$46 - \$100	1	Single-family	6 to 15	Telephone (la	Senice with c					
21	5925-XYDVY	Quality Control	\$46 - \$100	1	Single-family	6 to 15	internet	Other provider					
22	9544-WXCXY	ROC Curye	Under \$45	-4	Single-family	Less than 6	Television	Other provider					
23	9544-W/XCXY	Temale 15-30 At least s	Under \$45	4	Single-Tamily	Less than 6	Telephone (la	Sence with c.					
24	9544-WXCXY	Female 18-30 At least s	Under \$45	4	Single-family	Less than 6	Internet	Other provider					
25	8347-AKFLL	Male 46-60 College d	Under \$45	2	Single-family	More than 15	Television	Service with c					
26	8347-AKFLL	Male 45-50 College d.	Under \$45	2	Single-family	More than 15	Telephone (la.	Service with c					
27	B347-AKFLL	Male 46-60 College d.	Under \$45	2	Single-family	More than 15	Internet	Service with c					
28	8655-CANES	Female 31-45 At least s.	\$46 - \$100	1	Condominium/	More than 15	Television	Service with c					
	800	· · · · · · · · · · · · · · · · · · ·	A.A. 4144						_	_			1*
Data View	Variable View												
Territorian de	Linker							100	A CENER &	addition Do	or in some las some	and a line	

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
 - O Descriptive Statistics
 - O Correlation
 - O 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