March Madness

—

Presentation Month
Kaggle Tournament

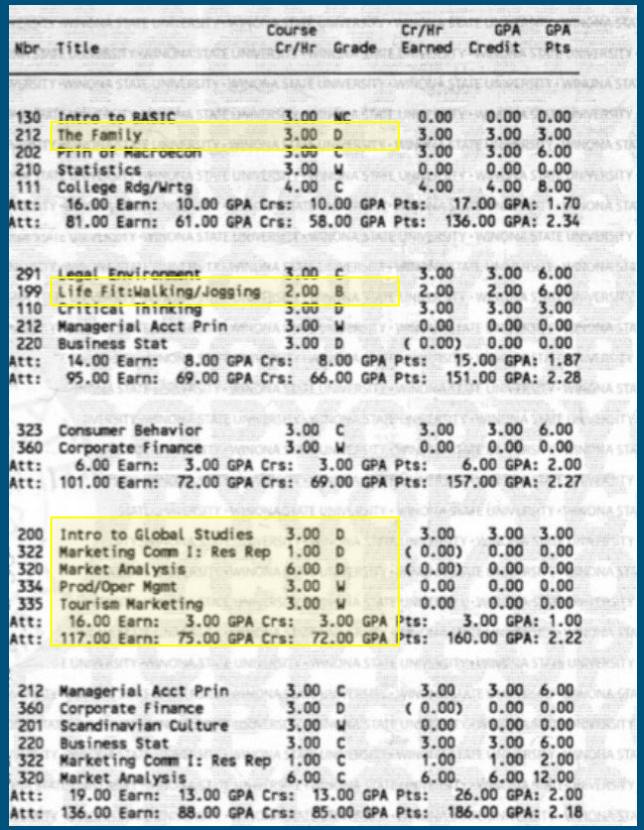
Today's Agenda

- 5++ Min Aside: Let's discuss real world software for "data stuff"
- Guest Speaker: Do's and Do Not's
- Exam Walkthrough
- Classification vs. Regression
- Unsupervised vs. Supervised
- Deep Dive Review of Regression and Clustering
- Extra Q&A time for project work and 'abstractions' discussion
    - Abstractions: can you relate your problem to another dataset? Filtering, grouping
    - Exploring Pandas functions which may apply
    - Organizing data, dataframes, etc.
    - Encoding, scaling, creating columns, etc.

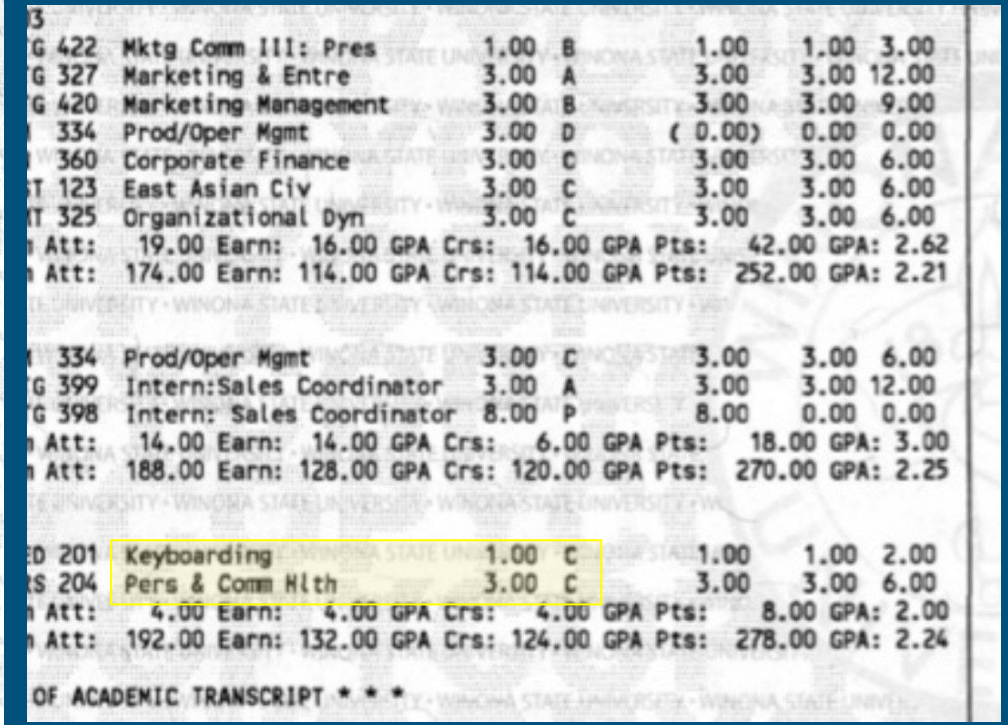
Next Week's Agenda

- Presentation - NAML
- Step through the 'Making of a Project'
- Steve's Presentation
- Real World Business Data Questions (StarCite/Cvent - Maritz)
- Plotting Discussions
- Measuring Model Goodness
- Python Stats Model

You are always performing better than some



Nbr	Title	Course Cr/Hr	Grade	Cr/Hr Earned	GPA Credit	GPA Pts
130	Intro to BASIC	3.00	NC	0.00	0.00	0.00
212	The Family	3.00	D	3.00	3.00	3.00
202	Prin of Macroecon	3.00	C	3.00	3.00	6.00
210	Statistics	3.00	W	0.00	0.00	0.00
111	College Rdg/Wrtg	4.00	C	4.00	4.00	8.00
Att:	16.00 Earn: 10.00 GPA Crs: 10.00 GPA Pts:				17.00	GPA: 1.70
Att:	81.00 Earn: 61.00 GPA Crs: 58.00 GPA Pts:				136.00	GPA: 2.34
291	Legal Environment	3.00	C	3.00	3.00	6.00
199	Life Fit:Walking/Jogging	2.00	B	2.00	2.00	6.00
110	Critical Thinking	3.00	D	3.00	3.00	3.00
212	Managerial Acct Prin	3.00	W	0.00	0.00	0.00
220	Business Stat	3.00	D	(0.00)	0.00	0.00
Att:	14.00 Earn: 8.00 GPA Crs: 8.00 GPA Pts:				15.00	GPA: 1.87
Att:	95.00 Earn: 69.00 GPA Crs: 66.00 GPA Pts:				151.00	GPA: 2.28
323	Consumer Behavior	3.00	C	3.00	3.00	6.00
360	Corporate Finance	3.00	W	0.00	0.00	0.00
Att:	6.00 Earn: 3.00 GPA Crs: 3.00 GPA Pts:				6.00	GPA: 2.00
Att:	101.00 Earn: 72.00 GPA Crs: 69.00 GPA Pts:				157.00	GPA: 2.27
200	Intro to Global Studies	3.00	D	3.00	3.00	3.00
322	Marketing Comm I: Res Rep	1.00	D	(0.00)	0.00	0.00
320	Market Analysis	6.00	D	(0.00)	0.00	0.00
334	Prod/Oper Mgmt	3.00	W	0.00	0.00	0.00
335	Tourism Marketing	3.00	W	0.00	0.00	0.00
Att:	16.00 Earn: 3.00 GPA Crs: 3.00 GPA Pts:				3.00	GPA: 1.00
Att:	117.00 Earn: 75.00 GPA Crs: 72.00 GPA Pts:				160.00	GPA: 2.22
212	Managerial Acct Prin	3.00	C	3.00	3.00	6.00
360	Corporate Finance	3.00	D	(0.00)	0.00	0.00
201	Scandinavian Culture	3.00	W	0.00	0.00	0.00
220	Business Stat	3.00	C	3.00	3.00	6.00
322	Marketing Comm I: Res Rep	1.00	C	1.00	1.00	2.00
320	Market Analysis	6.00	C	6.00	6.00	12.00
Att:	19.00 Earn: 13.00 GPA Crs: 13.00 GPA Pts:				26.00	GPA: 2.00
Att:	136.00 Earn: 88.00 GPA Crs: 85.00 GPA Pts:				186.00	GPA: 2.18



G 422	Mktg Comm III: Pres	1.00	B	1.00	1.00	3.00
G 327	Marketing & Entre	3.00	A	3.00	3.00	12.00
G 420	Marketing Management	3.00	B	3.00	3.00	9.00
334	Prod/Oper Mgmt	3.00	D	(0.00)	0.00	0.00
360	Corporate Finance	3.00	C	3.00	3.00	6.00
T 123	East Asian Civ	3.00	C	3.00	3.00	6.00
T 325	Organizational Dyn	3.00	C	3.00	3.00	6.00
Att:	19.00 Earn: 16.00 GPA Crs: 16.00 GPA Pts:				42.00	GPA: 2.62
Att:	174.00 Earn: 114.00 GPA Crs: 114.00 GPA Pts:				252.00	GPA: 2.21
334	Prod/Oper Mgmt	3.00	C	3.00	3.00	6.00
G 399	Intern:Sales Coordinator	3.00	A	3.00	3.00	12.00
G 398	Intern: Sales Coordinator	8.00	P	8.00	0.00	0.00
Att:	14.00 Earn: 14.00 GPA Crs: 6.00 GPA Pts:				18.00	GPA: 3.00
Att:	188.00 Earn: 128.00 GPA Crs: 120.00 GPA Pts:				270.00	GPA: 2.25
D 201	Keyboarding	1.00	C	1.00	1.00	2.00
S 204	Pers & Comm Hlth	3.00	C	3.00	3.00	6.00
Att:	4.00 Earn: 4.00 GPA Crs: 4.00 GPA Pts:				8.00	GPA: 2.00
Att:	192.00 Earn: 132.00 GPA Crs: 124.00 GPA Pts:				278.00	GPA: 2.24

OF ACADEMIC TRANSCRIPT * * *

What is inherent to ‘slicing data up’

Categorical columns!! Look here first.

Numerical columns which you feel deserve buckets - Self Made Millionaire?

Groupings == Segmentation

	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON	SEED
0	North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	...	30.4	53.9	44.6	32.7	36.2	71.7	8.6	2ND	1.0
1	Wisconsin	B10	40	36	129.1	93.6	0.9758	54.8	47.7	12.4	...	22.4	54.8	44.7	36.5	37.5	59.3	11.3	2ND	1.0
2	Michigan	B10	40	33	114.4	90.4	0.9375	53.9	47.7	14.0	...	30.0	54.7	46.8	35.2	33.2	65.9	6.9	2ND	3.0
3	Texas Tech	B12	38	31	115.2	85.2	0.9696	53.5	43.0	17.7	...	36.6	52.8	41.9	36.5	29.7	67.5	7.0	2ND	3.0
4	Gonzaga	WCC	39	37	117.8	86.3	0.9728	56.6	41.1	16.2	...	26.9	56.3	40.0	38.2	29.0	71.5	7.7	2ND	1.0

What the real world looks like

For a data scientist - mostly the same as last slide

For others - directed request, usually a boondoggle that you need to shape

Groupings == Segmentation

	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON	SEED
0	North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	...	30.4	53.9	44.6	32.7	36.2	71.7	8.6	2ND	1.0
1	Wisconsin	B10	40	36	129.1	93.6	0.9758	54.8	47.7	12.4	...	22.4	54.8	44.7	36.5	37.5	59.3	11.3	2ND	1.0
2	Michigan	B10	40	33	114.4	90.4	0.9375	53.9	47.7	14.0	...	30.0	54.7	46.8	35.2	33.2	65.9	6.9	2ND	3.0
3	Texas Tech	B12	38	31	115.2	85.2	0.9696	53.5	43.0	17.7	...	36.6	52.8	41.9	36.5	29.7	67.5	7.0	2ND	3.0
4	Gonzaga	WCC	39	37	117.8	86.3	0.9728	56.6	41.1	16.2	...	26.9	56.3	40.0	38.2	29.0	71.5	7.7	2ND	1.0

Jupyter Notebook - Data Munging Tips!

If you draw up the right game plan - “I know what I want to do…but the coding is grr…”
You are on the right track!

50%…maybe 60%…. Will be spent on data wrangling, data munging, data wrestling…

Ask for help - Poisson Distribution as we move into October

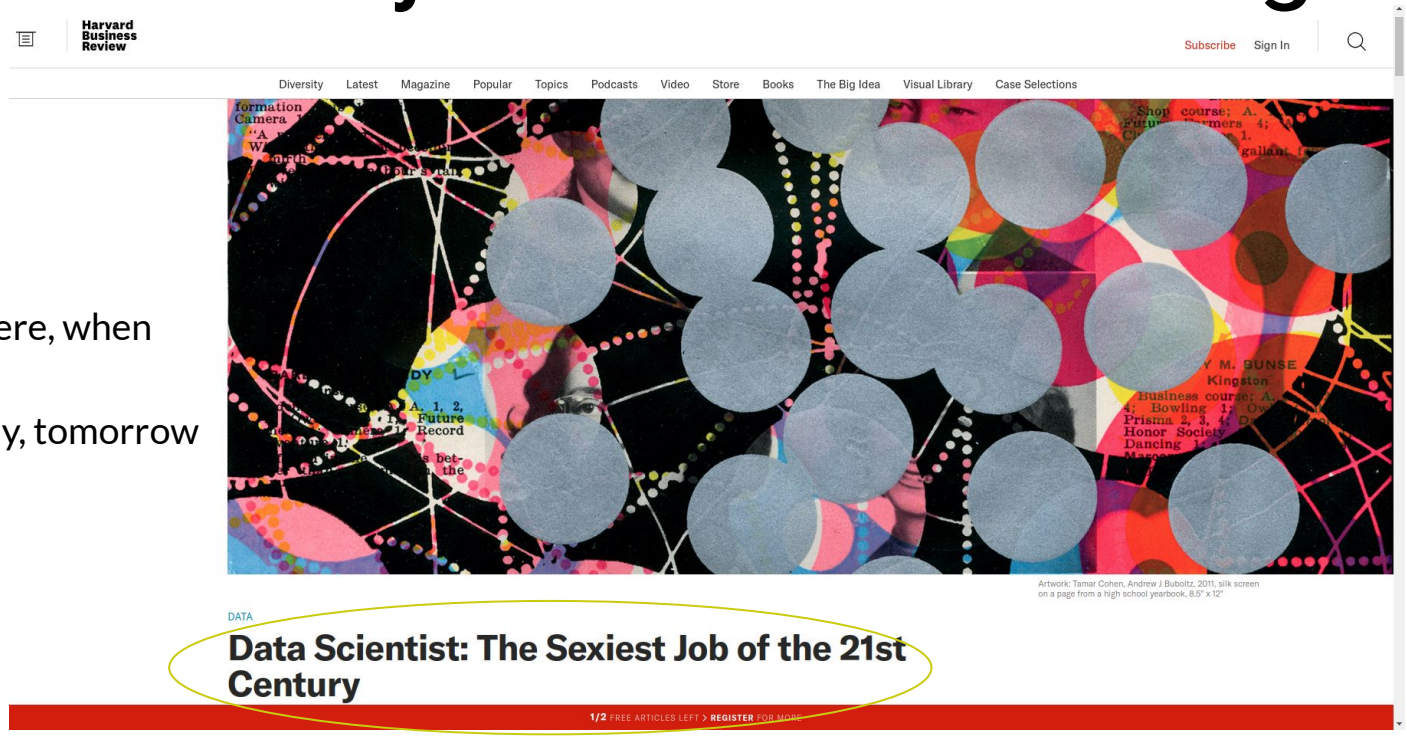
If you do these things…the visuals…the iterations on all of this…will come very easily

Build your presentation deck outline early / often … make it effective not in replacement for…

Presentation DO NOT’s - which may conflict with other teachings and I am OK with that…
< bullets & words >

Anatomy of ‘real world’ data-ing

- Process
- Who, what, where, when
- Why
- Yesterday, today, tomorrow

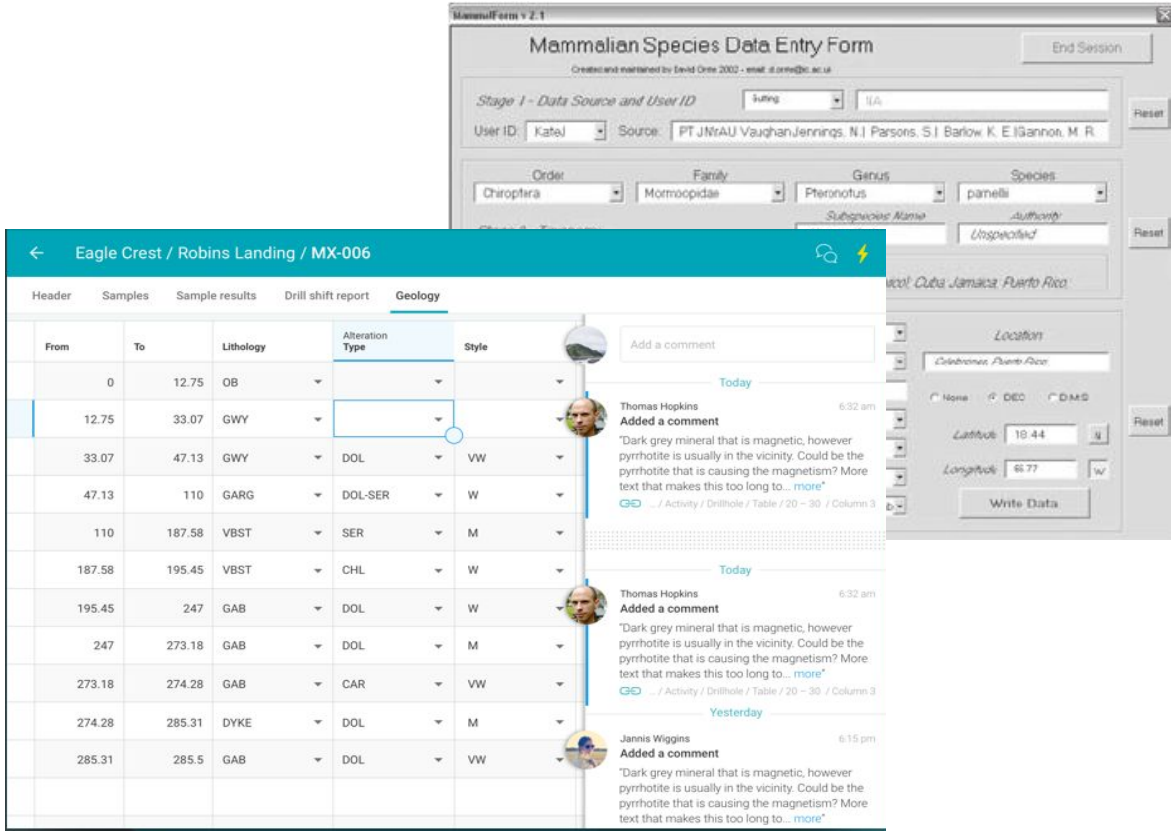


Step 1 - Client contacts YOU -ish

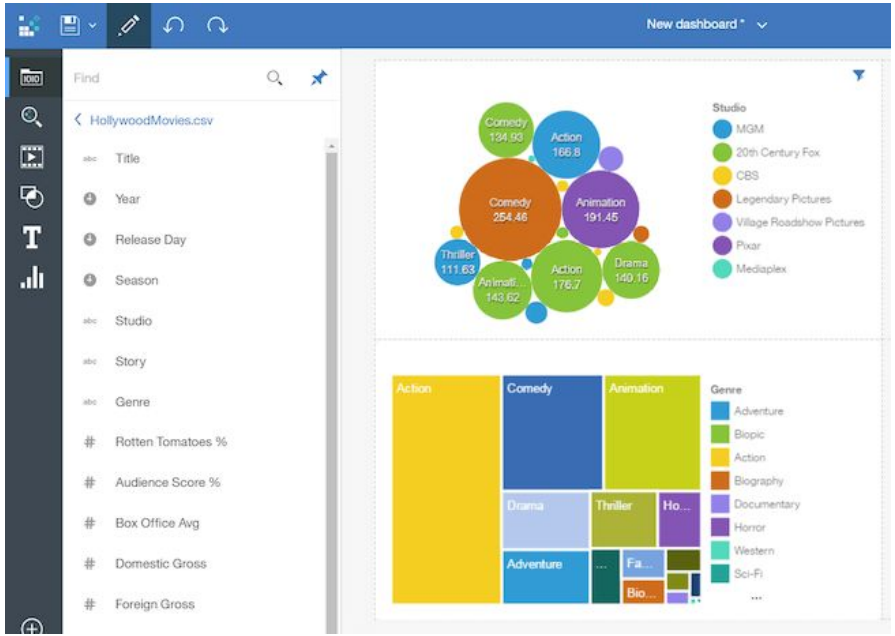
Why does this happen?

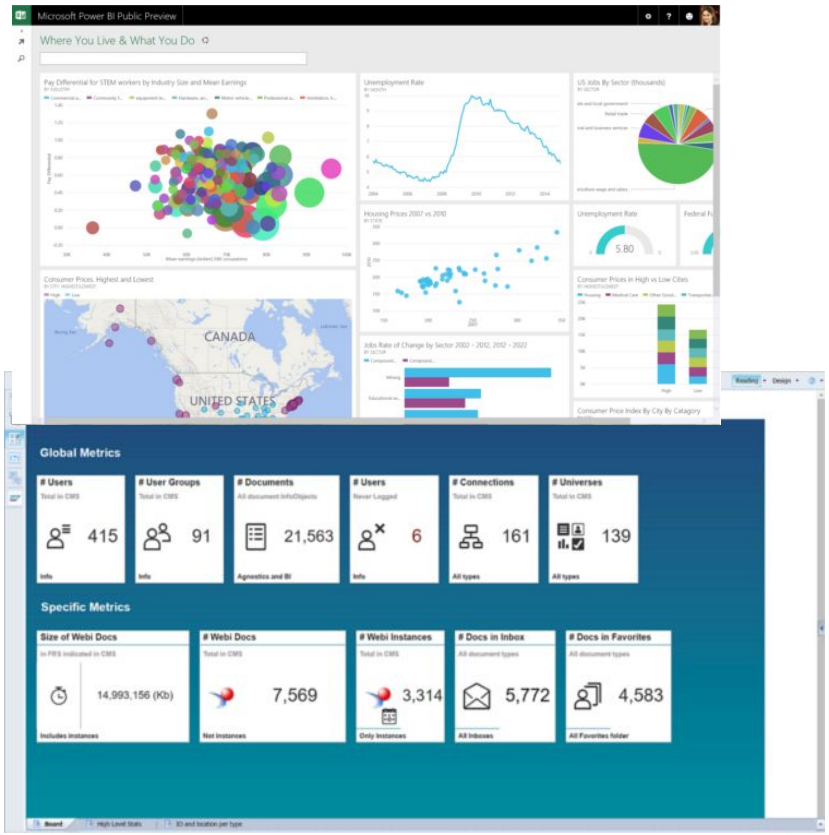
Case study

Why are we trying to do better



Step 2 - Your Company Owns Software



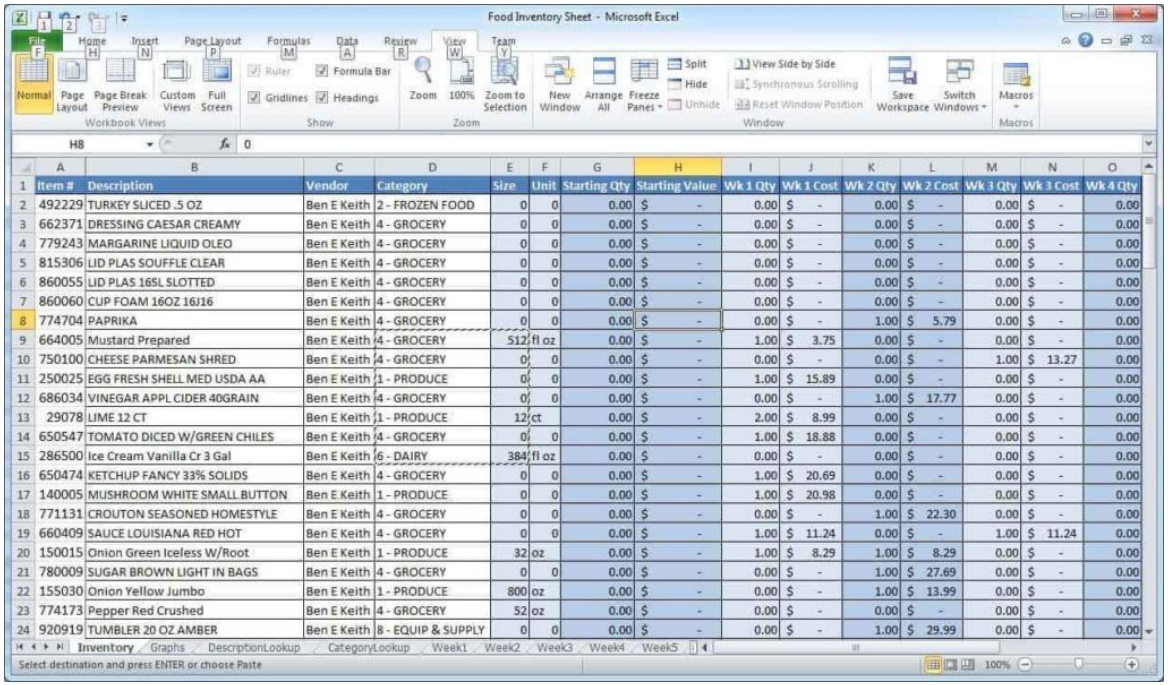


Database Manager or Data Analyst

There are hundreds or more of these tools...even amongst the Fortune 500 companies...maybe dozens of different ones used at $1500 per user license

Step 3 - External Source Owns Software

By External Source - this could mean another entity like a client like Joe's presentation, the internet like y'alls presentations or another department

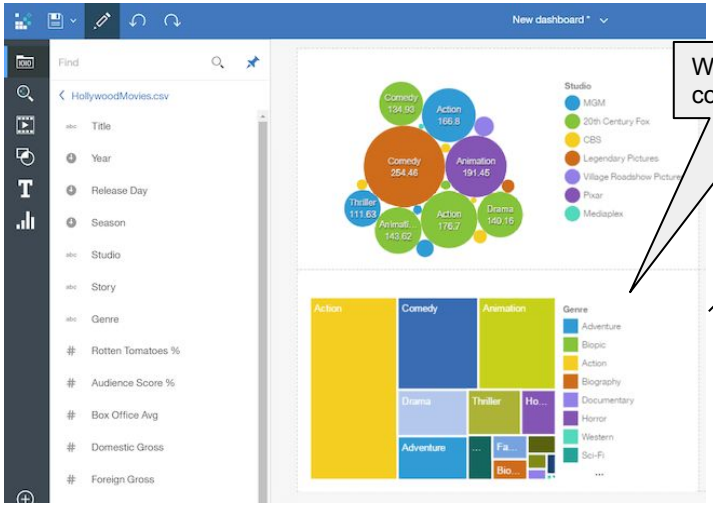


Item #	Description	Vendor	Category	Size	Unit	Starting Qty	Starting Value	Wk 1 Qty	Wk 1 Cost	Wk 2 Qty	Wk 2 Cost	Wk 3 Qty	Wk 3 Cost	Wk 4 Qty
492229	TURKEY SLICED .5 OZ	Ben E Keith	2 - FROZEN FOOD	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
662371	DRESSING CAESAR CREAMY	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
779243	MARGARINE LIQUID OLEO	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
815306	LID PLAS SOUFFLE CLEAR	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
860055	LID PLAS 16SL SLOTTED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
860060	CUP FOAM 16OZ 16J16	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
774704	PAPRIKA	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 5.79	0.00	$ -	0.00
664005	Mustard Prepared	Ben E Keith	4 - GROCERY	512	fl oz	0.00	$ -	1.00	$ 3.75	0.00	$ -	0.00	$ -	0.00
750100	CHEESE PARMESAN SHRED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	1.00	$ 13.27	0.00
250025	EGG FRESH SHELL MED USDA AA	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 15.89	0.00	$ -	0.00	$ -	0.00
686034	VINEGAR APPL CIDER 40GRAIN	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 17.77	0.00	$ -	0.00
29078	LIME 12 CT	Ben E Keith	1 - PRODUCE	12	ct	0.00	$ -	2.00	$ 8.99	0.00	$ -	0.00	$ -	0.00
650547	TOMATO DICED W/GREEN CHILES	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 18.88	0.00	$ -	0.00	$ -	0.00
286500	Ice Cream Vanilla Cr 3 Gal	Ben E Keith	6 - DAIRY	384	fl oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
650474	KETCHUP FANCY 33% SOLIDS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 20.69	0.00	$ -	0.00	$ -	0.00
140005	MUSHROOM WHITE SMALL BUTTON	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 20.98	0.00	$ -	0.00	$ -	0.00
771131	CROUTON SEASONED HOMESTYLE	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 22.30	0.00	$ -	0.00
660409	SAUCE LOUISIANA RED HOT	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 11.24	0.00	$ -	1.00	$ 11.24	0.00
150015	Onion Green Iceless W/Root	Ben E Keith	1 - PRODUCE	32	oz	0.00	$ -	1.00	$ 8.29	1.00	$ 8.29	0.00	$ -	0.00
780009	SUGAR BROWN LIGHT IN BAGS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 27.69	0.00	$ -	0.00
155030	Onion Yellow Jumbo	Ben E Keith	1 - PRODUCE	800	oz	0.00	$ -	0.00	$ -	1.00	$ 13.99	0.00	$ -	0.00
774173	Pepper Red Crushed	Ben E Keith	4 - GROCERY	52	oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
920919	TUMBLER 20 OZ AMBER	Ben E Keith	8 - EQUIP & SUPPLY	0	0	0.00	$ -	0.00	$ -	1.00	$ 29.99	0.00	$ -	0.00

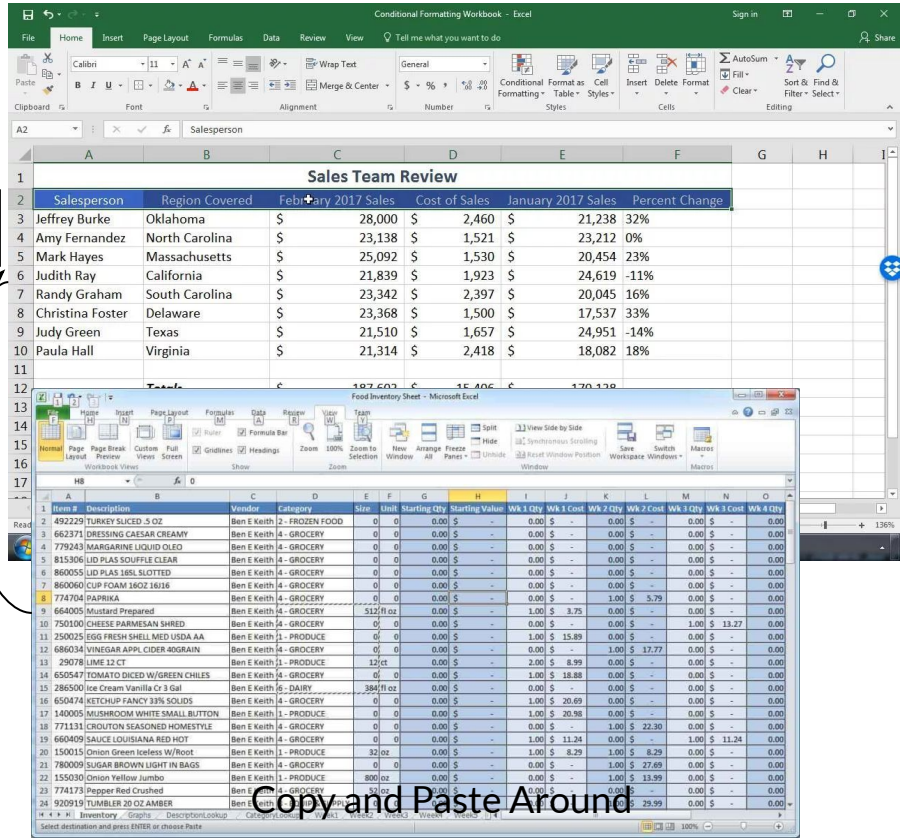
Someone who probably hasn't taken DATA 101

Step 4 - Attempt Merger

Hit ‘Download Report’



Why is this, how come?!?

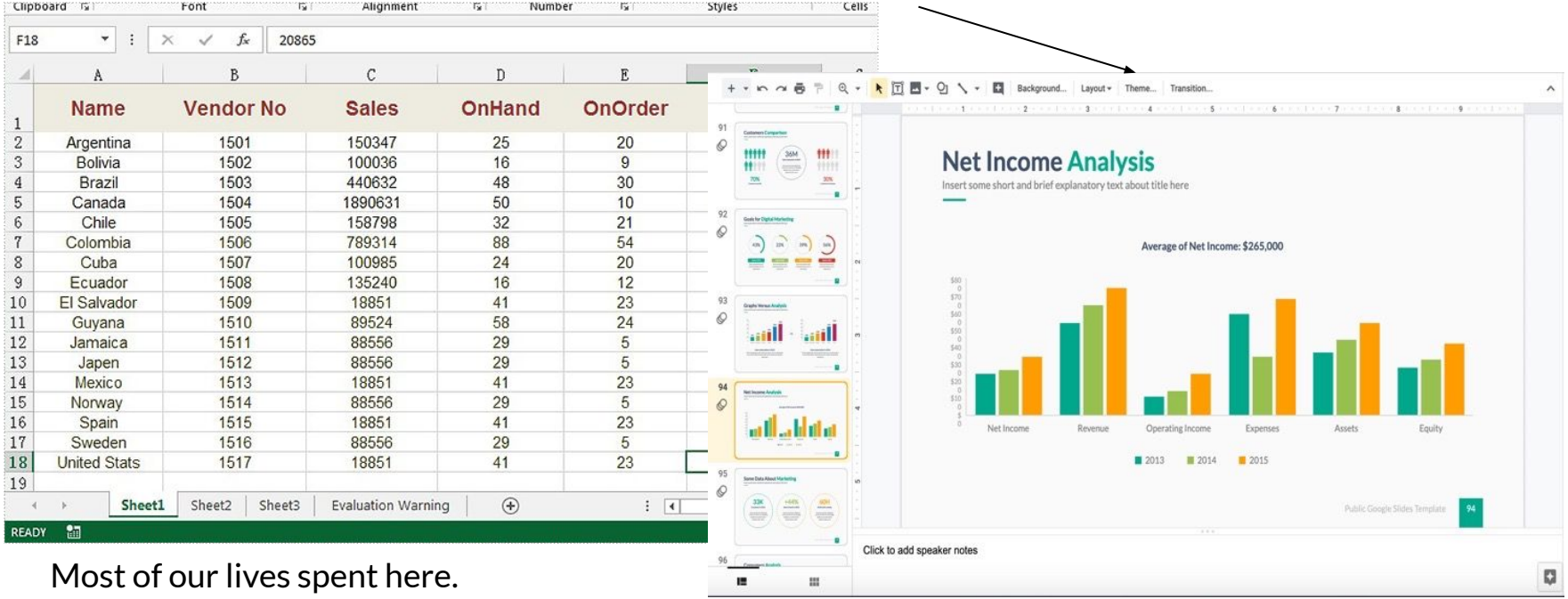


Someone who probably hasn’t taken DATA 101

But like 83% of the work force you will encounter...

Copy and Paste Around

Step 5 - Show Value - Build New ‘Reporting’

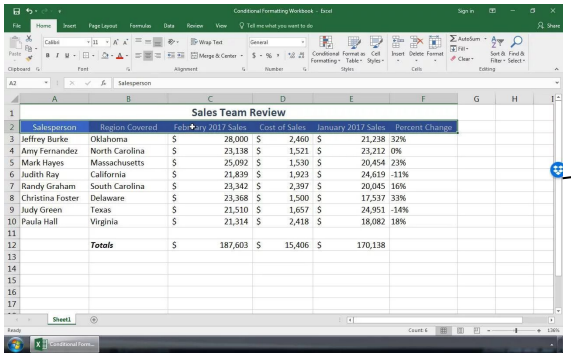


Most of our lives spent here.
70% of applicable workforce may not know what a pivot table is.
50% rely on copy and paste verse sum() or if() in Excel.
Manually create, color, sometimes type into the graphics...

Step X - Why is this class doing it different???

- 2000 Niche -> 2010 Desirable -> 2020 Marketable -> 2025 Expectation
- Flexibility, Machine Learning,
- Canned vs Unique Analysis - you have ideas too!
- No data department...
- Do you want to learn a (couple) tools and hope they match your employers subscription?
- How big is the data? (65K -> 1M Excel Rows/Columns)

How long this takes (my friend who did this....and me who helped...)



Sales Team Review

Salesperson	Region Covered	February 2017 Sales	Cost of Sales	January 2017 Sales	Percent Change
Jeffrey Burke	Oklahoma	$ 28,000	$ 2,460	$ 21,238	32%
Amy Fernandez	North Carolina	$ 23,138	$ 1,521	$ 23,212	0%
Mark Hayes	Massachusetts	$ 25,092	$ 1,530	$ 20,454	23%
Judith Ray	California	$ 21,839	$ 1,923	$ 24,619	-11%
Randy Graham	South Carolina	$ 23,342	$ 2,397	$ 20,045	16%
Christina Foster	Delaware	$ 23,368	$ 1,500	$ 17,537	33%
Judy Green	Texas	$ 21,510	$ 1,657	$ 24,951	-14%
Paula Hall	Virginia	$ 21,314	$ 2,418	$ 18,082	18%
	Totals	$ 187,603	$ 15,406	$ 170,138	

```
In [14]: df.apply(lambda x: sum(x.isnull()),axis=0)
Out[14]: Loan_ID              0
         Gender              13
         Married              3
         Dependents          15
         Education            0
         Self_Employed       32
         ApplicantIncome      0
         CoapplicantIncome    0
         LoanAmount          22
         Loan_Amount_Term    14
         Credit_History      50
         Property_Area        0
         Loan_Status          0
         dtype: int64
```

Pssst...Excel files are .xlsx where a .csv is a Comma Separated Value file

Tide case study from Ogilvy - 5 More Minutes

How might you be asked to use data?

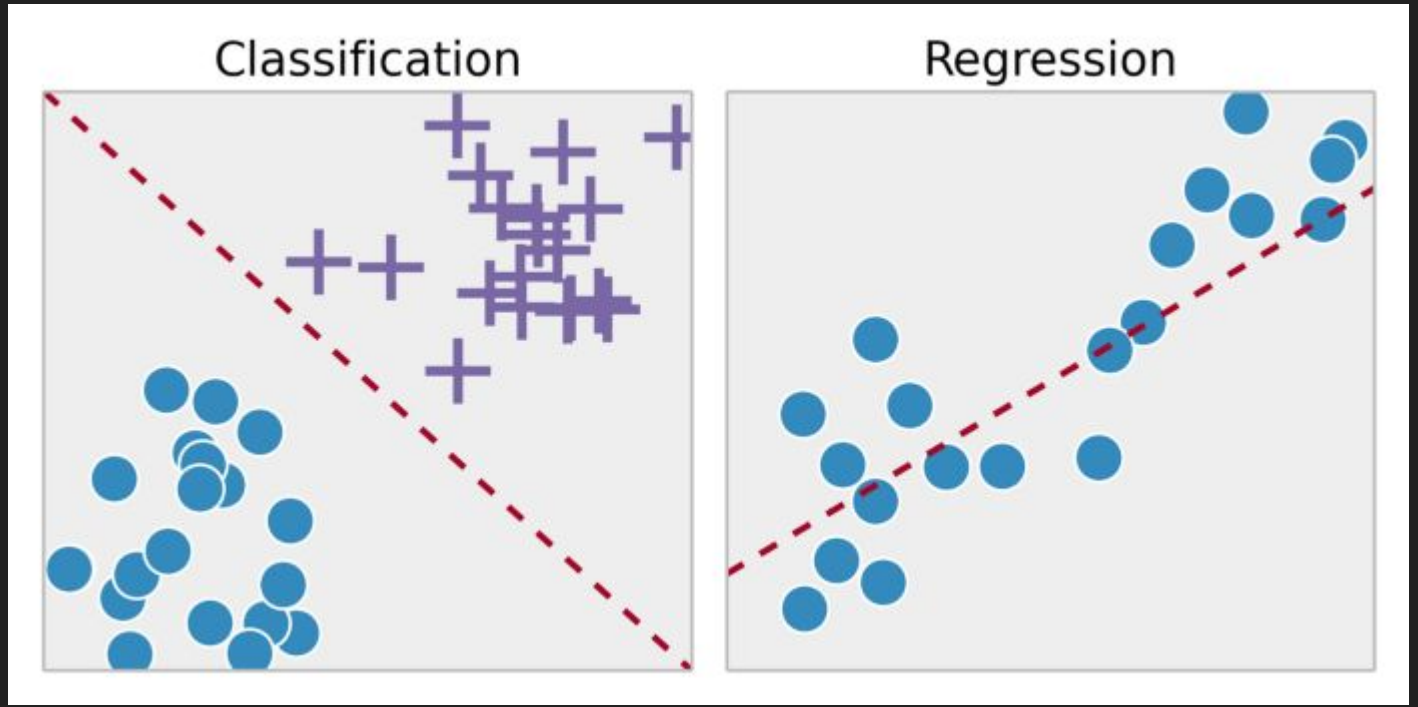
Finding unbiased nuggets or hunting for a story

Welcome Our Guest Speak - Joe Makepeace

Joe Makepeace is a Manager in the RiskSpan valuation group leading client engagements in both in portfolio valuation as well as financial data analysis advisory. He has experience in mortgage portfolio forecasting, financial data governance, data analysis, fund risk reporting, ALLL Modeling, and project management. Before coming to RiskSpan, he was as an associate at Modus21, a consulting firm based in Charleston, SC, where he focused on business requirements gathering, data analysis and validation and Business Intelligence solutions. He holds a BS in Finance from the University of South Carolina

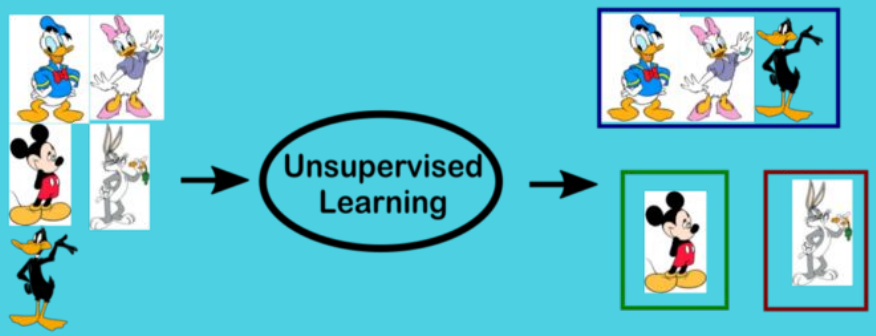
Jupyter Notebook...EXAM WALKTHROUGH

Supervised Learning: Classification vs. Regression

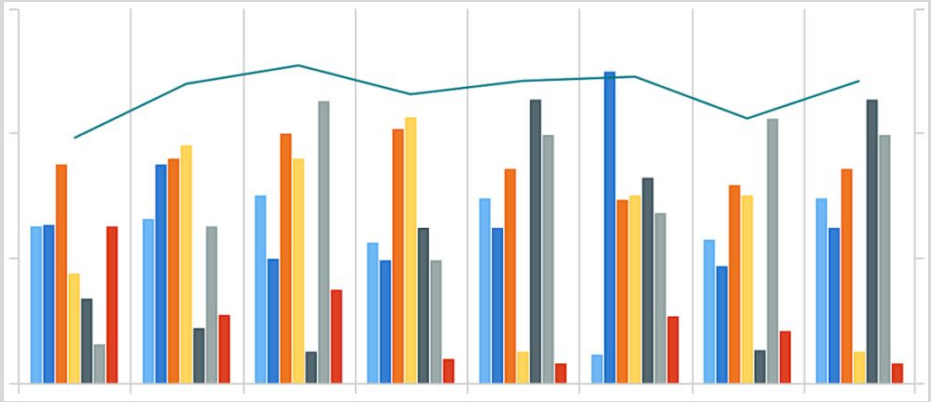


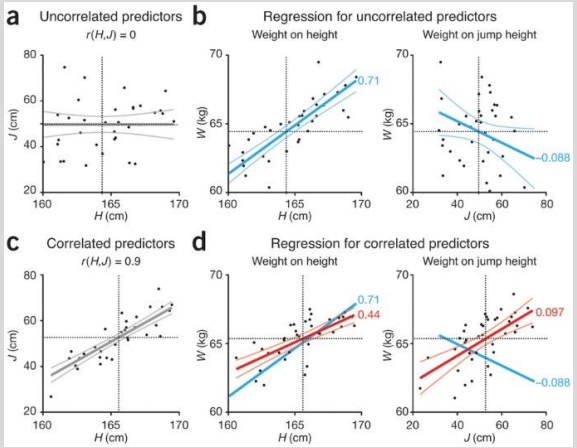
Supervised Vs Unsupervised

Unsupervised Learning



	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

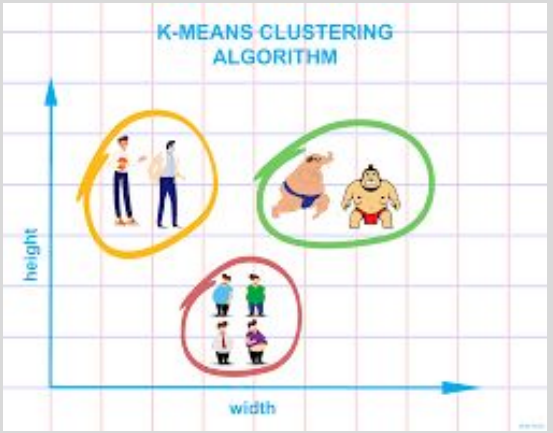




When to use what tool

Table 1 Likelihood table to make a diagnosis of sepsis

Likelihood	Respiratory rate		Mental status		
	Fast	Slow	Altered	Normal	Total
Sepsis	15/20	5/20	17/20	3/20	20
Non-sepsis	5/80	75/80	3/80	77/80	80
Total	20/100	80/100	20/100	80/100	100



Encoding & Scaling - advanced topic but we need to understand what is taking place

	location	menu	price
0	[NY, CA, MI]	[Italian, Greek]	$$
1	CA	[Japanese]	$$
2	[NY, CA, MA]	[Italian, Greek, Japanese]	$

	location	menu	price
0	[1, 1, 1, 0]	[1, 1, 0]	[1, 0]
1	[0, 1, 0, 0]	[0, 0, 1]	[1, 0]
2	[1, 1, 0, 1]	[1, 1, 1]	[0, 1]

Standardization:

$$z = \frac{x - \mu}{\sigma}$$

with mean:

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

and standard deviation:

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

Min-Max scaling:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$