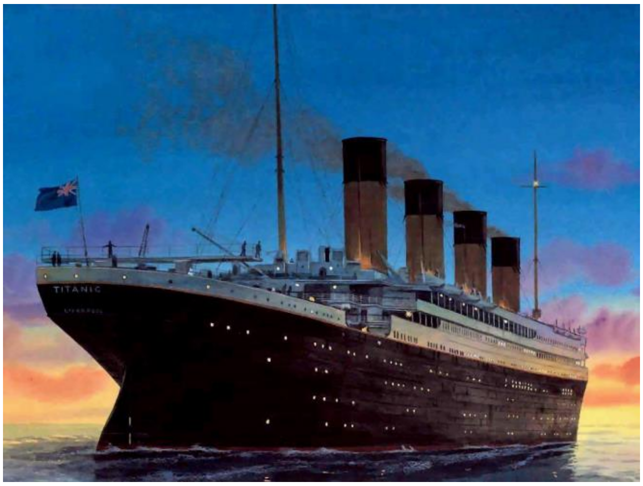
Decision Trees – Titanic Data Set

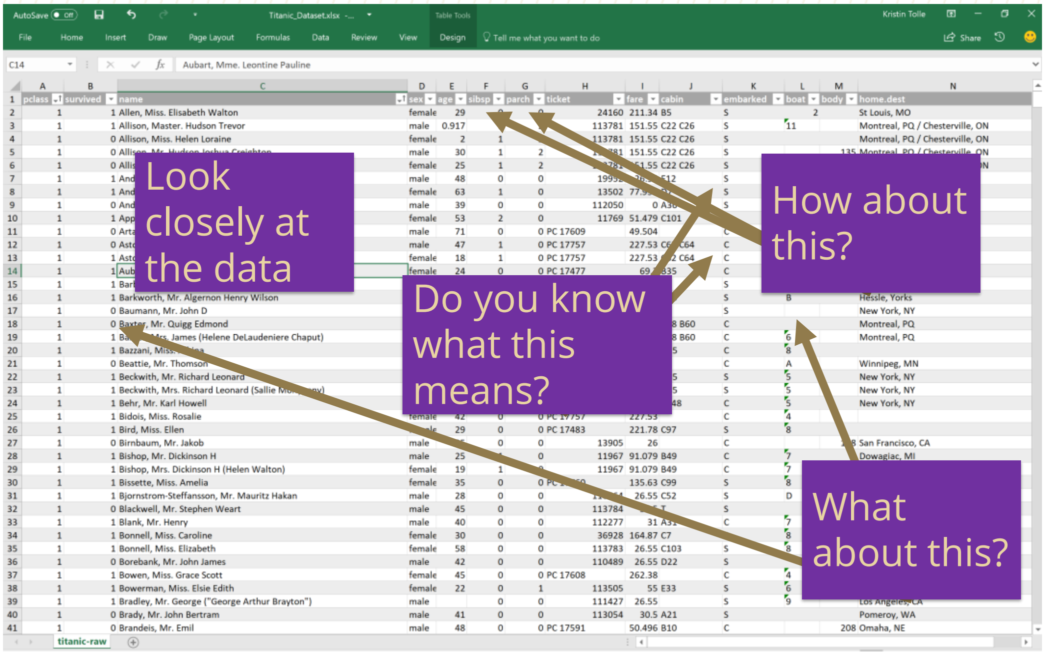
Lesson 4 – Section 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Example: Tragedy of the Titanic



Women and children first?



C14: Aubart, Mme. Leontine Pauline

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
2	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.34	B5	S	2		St Louis, MO
3	1	1	Allison, Master. Hudson Trevor	male	0.917			113781	151.55	C22 C26	S	11		Montreal, PQ / Chesterville, ON
4	1	0	Allison, Miss. Helen Loraine	female	2	1		113781	151.55	C22 C26	S			Montreal, PQ / Chesterville, ON
5	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.55	C22 C26	S		135	Montreal, PQ / Chesterville, ON
6	1	0	Allis	female	25	1	2	781	51.55	C22 C26	S			N
7	1	1	And	male	48	0	0	1995	26.	E12	S			
8	1	1	And	female	63	1	0	13502	77.9		S			
9	1	0	And	male	39	0	0	112050	0	A36	S			
10	1	1	App	female	53	2	0	11769	51.479	C101	S			
11	1	0	Arta	male	71	0	0	PC 17609	49.504		C			
12	1	0	Asto	male	47	1	0	PC 17757	227.53	C6 C64	C			
13	1	1	Asto	female	18	1	0	PC 17757	227.53	2 C64	C			
14	1	1	Aub	female	24	0	0	PC 17477	69.	B35	C			
15	1	1	Bar								S			
16	1	1	Barkworth, Mr. Algernon Henry Wilson								S	B		Hessle, Yorks
17	1	0	Baumann, Mr. John D								S			New York, NY
18	1	0	Baxter, Mr. Quigg Edmond							8 B60	C			Montreal, PQ
19	1	1	Ba Mrs. James (Helene DeLaudeniere Chaput)							8 B60	C	6		Montreal, PQ
20	1	1	Bazzani, Miss.							5	C	8		
21	1	0	Beattie, Mr. Thomson								C	A		Winnipeg, MN
22	1	1	Beckwith, Mr. Richard Leonard							5	S	5		New York, NY
23	1	1	Beckwith, Mrs. Richard Leonard (Sallie Mo ny)							5	S	5		New York, NY
24	1	1	Behr, Mr. Karl Howell							48	C	5		New York, NY
25	1	1	Bidois, Miss. Rosalie	female	42	0	0	PC 17757	227.53		C	4		
26	1	1	Bird, Miss. Ellen		29	0	0	PC 17483	221.78	C97	S	8		
27	1	0	Birnbaum, Mr. Jakob	male		0	0	13905	26		C		8	San Francisco, CA
28	1	1	Bishop, Mr. Dickinson H	male	25		0	11967	91.079	B49	C	7		Dowagiac, MI
29	1	1	Bishop, Mrs. Dickinson H (Helen Walton)	female	19	1		11967	91.079	B49	C	7		
30	1	1	Bissette, Miss. Amelia	female	35	0	0	PC	135.63	C99	S	8		
31	1	1	Bjornstrom-Steffansson, Mr. Mauritz Hakan	male	28	0	0		26.55	C52	S	D		
32	1	0	Blackwell, Mr. Stephen Weart	male	45	0	0	113784		T	S			
33	1	1	Blank, Mr. Henry	male	40	0	0	112277	31	A31	C	7		
34	1	1	Bonnell, Miss. Caroline	female	30	0	0	36928	164.87	C7		8		
35	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S	8		
36	1	0	Borebank, Mr. John James	male	42	0	0	110489	26.55	D22	S			
37	1	1	Bowen, Miss. Grace Scott	female	45	0	0	PC 17608	262.38		C	4		
38	1	1	Bowerman, Miss. Elsie Edith	female	22	0	1	113505	55	E33	S	6		
39	1	1	Bradley, Mr. George ("George Arthur Brayton")	male		0	0	111427	26.55		S	9		Los Angeles, CA
40	1	0	Brady, Mr. John Bertram	male	41	0	0	113054	30.5	A21	S			Pomeroy, WA
41	1	0	Brandeis, Mr. Emil	male	48	0	0	PC 17591	50.496	B10	C		208	Omaha, NE

Look closely at the data

How about this?

Do you know what this means?

What about this?

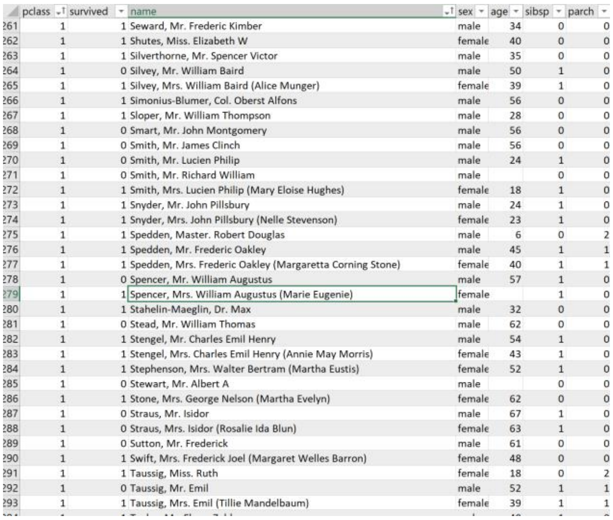
Start with a Data Definition

Feature	Data Description	Variable Type
survival	Survived (0 = no; 1 = yes)	Dependent variable
pclass	Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd	Independent variable
name	Passenger’s name as given	Relevance?
sex	Gender: Male or Female	Independent variable
age	Age	Independent variable
sibsp	Count of siblings and/or spouse aboard	Independent variable
parch	Count of parents or children aboard	Independent variable
ticket	Ticket number	Relevance?
fare	Passenger Fare	Independent variable
cabin	Cabin number	Relevance?
embarked	Port of departure (c = Cherbourg, q=Queenstown, s=Southhampton	Relevance?
boat	Rescue boat number	Relevance?
home	Home city of the passenger and ultimate (assumed) destination	Relevance?
…		

Choose wisely: More variables in the model can negatively affect compute time and potentially accuracy

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

First Steps… Look for Problems and Predictors

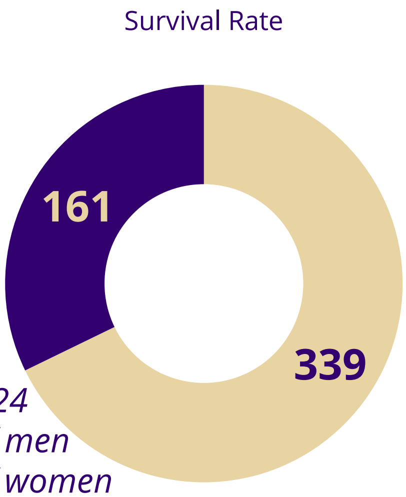


	pclass	survived	name	sex	age	sibsp	parch
261	1	1	Seward, Mr. Frederic Kimber	male	34	0	0
262	1	1	Shutes, Miss. Elizabeth W	female	40	0	0
263	1	1	Silverthorne, Mr. Spencer Victor	male	35	0	0
264	1	0	Silvey, Mr. William Baird	male	50	1	0
265	1	1	Silvey, Mrs. William Baird (Alice Munger)	female	39	1	0
266	1	1	Simonius-Blumer, Col. Oberst Alfons	male	56	0	0
267	1	1	Sloper, Mr. William Thompson	male	28	0	0
268	1	0	Smart, Mr. John Montgomery	male	56	0	0
269	1	0	Smith, Mr. James Clinch	male	56	0	0
270	1	0	Smith, Mr. Lucien Philip	male	24	1	0
271	1	0	Smith, Mr. Richard William	male		0	0
272	1	1	Smith, Mrs. Lucien Philip (Mary Eloise Hughes)	female	18	1	0
273	1	1	Snyder, Mr. John Pillsbury	male	24	1	0
274	1	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23	1	0
275	1	1	Spedden, Master. Robert Douglas	male	6	0	2
276	1	1	Spedden, Mr. Frederic Oakley	male	45	1	1
277	1	1	Spedden, Mrs. Frederic Oakley (Margaretta Corning Stone)	female	40	1	1
278	1	0	Spencer, Mr. William Augustus	male	57	1	0
279	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0
280	1	1	Stahelin-Maeglin, Dr. Max	male	32	0	0
281	1	0	Stead, Mr. William Thomas	male	62	0	0
282	1	1	Stengel, Mr. Charles Emil Henry	male	54	1	0
283	1	1	Stengel, Mrs. Charles Emil Henry (Annie May Morris)	female	43	1	0
284	1	1	Stephenson, Mrs. Walter Bertram (Martha Eustis)	female	52	1	0
285	1	0	Stewart, Mr. Albert A	male		0	0
286	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62	0	0
287	1	0	Straus, Mr. Isidor	male	67	1	0
288	1	0	Straus, Mrs. Isidor (Rosalie Ida Blun)	female	63	1	0
289	1	0	Sutton, Mr. Frederick	male	61	0	0
290	1	1	Swift, Mrs. Frederick Joel (Margaret Welles Barron)	female	48	0	0
291	1	1	Taussig, Miss. Ruth	female	18	0	2
292	1	0	Taussig, Mr. Emil	male	52	1	1
293	1	1	Taussig, Mrs. Emil (Tillie Mandelbaum)	female	39	1	1

Look for Missing Values e.g., Age…

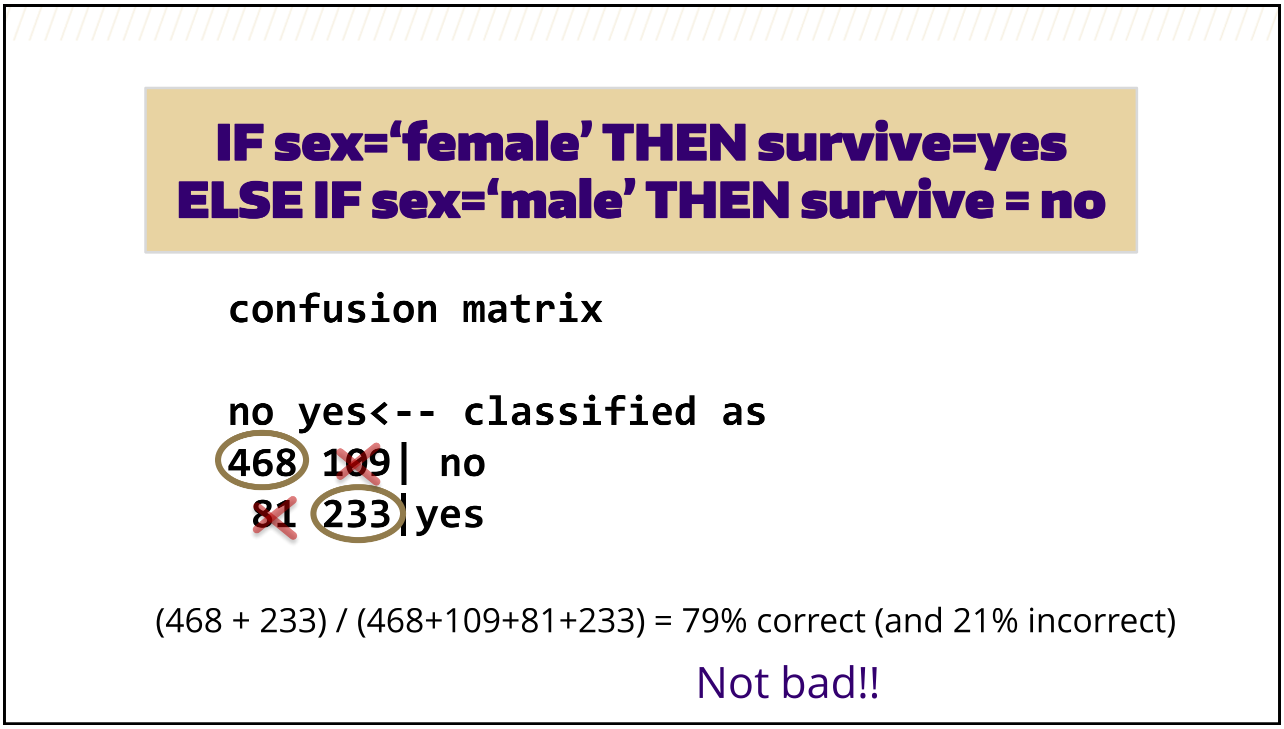
Make sure you understand column heading

Survival by Gender (Passengers)



N = 1324
12% of men
26% of women

More Women Survived



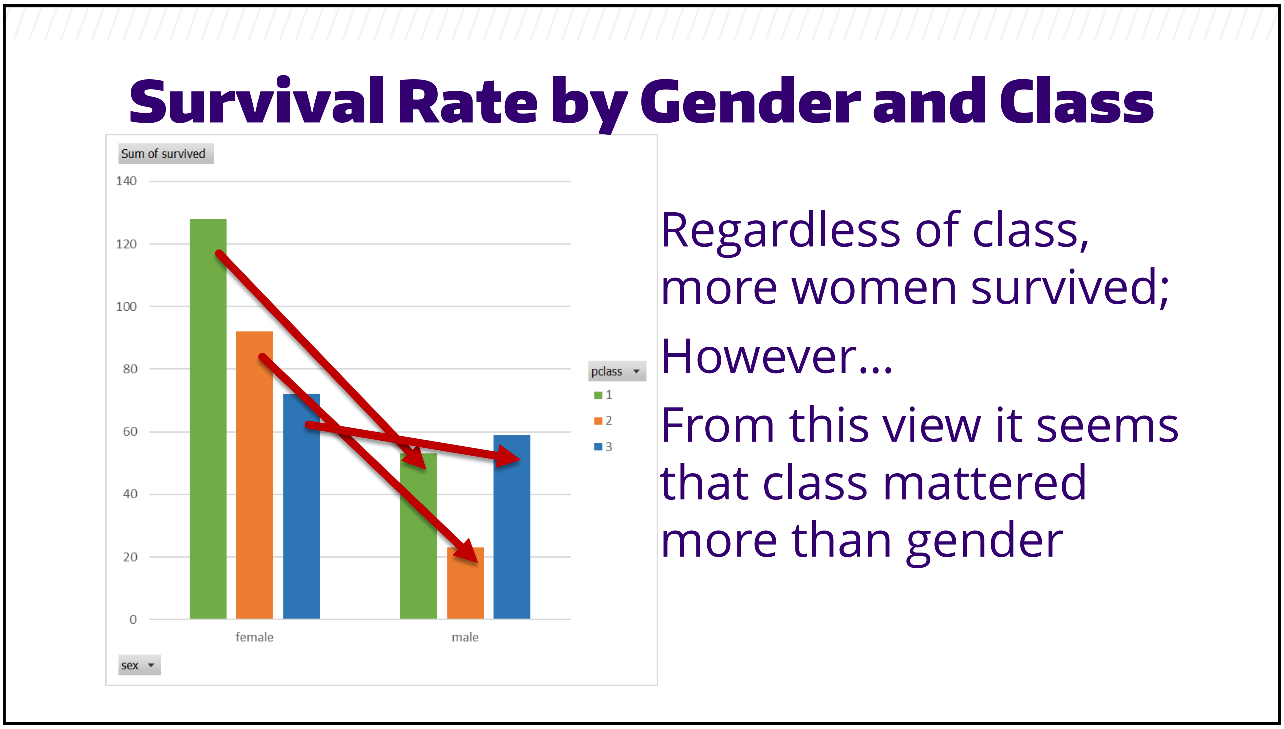
IF sex='female' THEN survive=yes ELSE IF sex='male' THEN survive = no

```
confusion matrix

no yes<-- classified as
468 109| no
 81 233|yes
```

(468 + 233) / (468+109+81+233) = 79% correct (and 21% incorrect)

Not bad!!



Survival Rate by Gender and Class

Regardless of class, more women survived;

However…

From this view it seems that class mattered more than gender

IF pclass='1' THEN survive=yes ELSE IF pclass='2' THEN survive=yes ELSE IF pclass='3' THEN survive=no

confusion matrix

no yes<-- classified as

372 ~~119~~| no

~~177~~ 223|yes

(372 + 223) / (372+119+223+177) = 67% correct (and 33% incorrect)

A little worse

Let's take a look at this example in a Jupyter Notebook

>Titanic-DecisionTree.ipynb