Modelling and Differential Equations - Discrete Modelling
/ Week 11: Working with data
/ Transition counts and probabilities from data

Transition counts and probabilities from data

Note: weather is a common example used for teaching Markov chains. You should be aware that the examples here are in no way sophisticated enough to act as weather prediction!

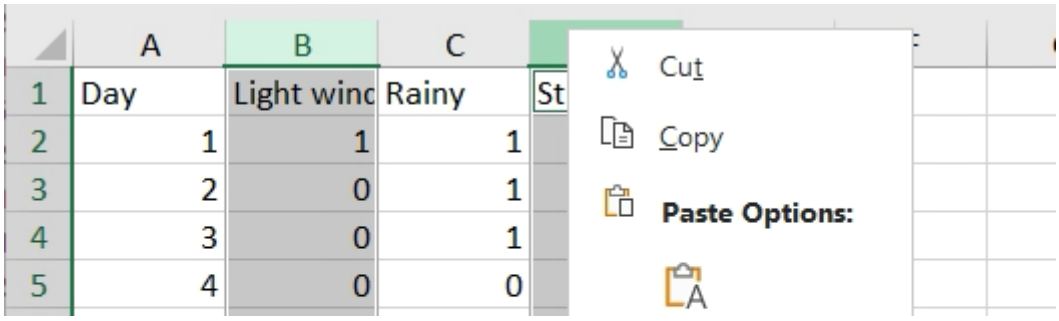
Open the data file weather-data.xlsx.

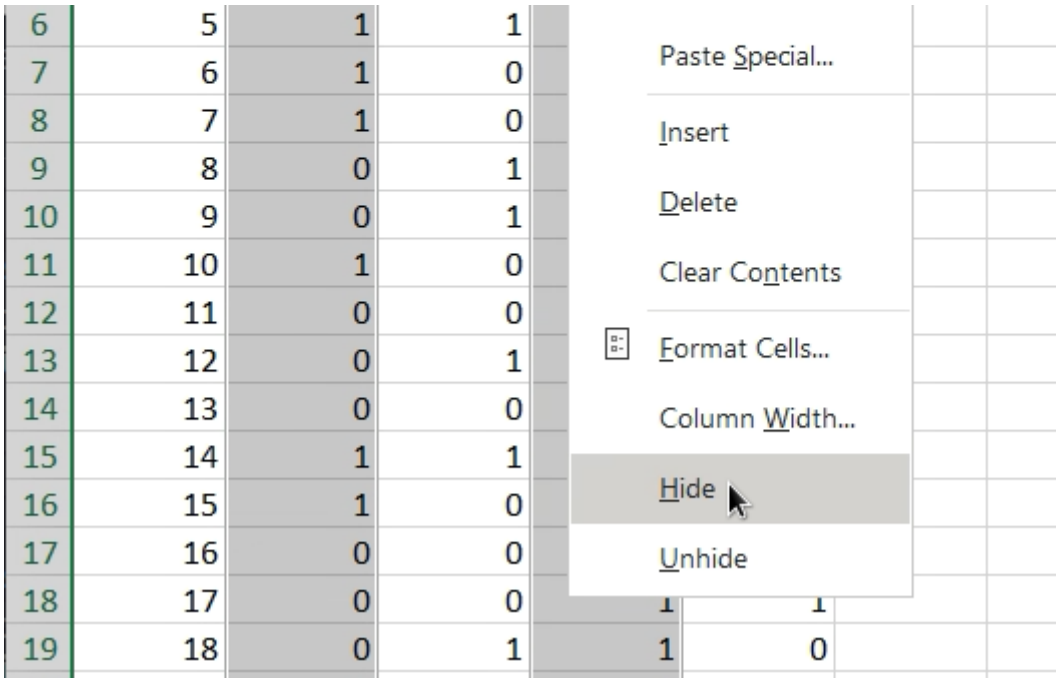
There are two types of data in the file:

1. Whether the weather was sunny or rainy.
2. Whether the wind was strong, light or there was no wind.

First we will work with the sun/rain data in the Town 1 sheet.

Hide the wind data by clicking `B` to highlight column `B` then while holding the Ctrl key clicking the `D` to highlight column `D`. Now right-click on one of the column headings and choose 'Hide'. (Or you can hide each column individually, but it is good to learn to select multiple columns using Shift and Ctrl.)





6	5	1	1		
7	6	1	0		
8	7	1	0		
9	8	0	1		
10	9	0	1		
11	10	1	0		
12	11	0	0		
13	12	0	1		
14	13	0	0		
15	14	1	1		
16	15	1	0		
17	16	0	0		
18	17	0	0	1	1
19	18	0	1	1	0

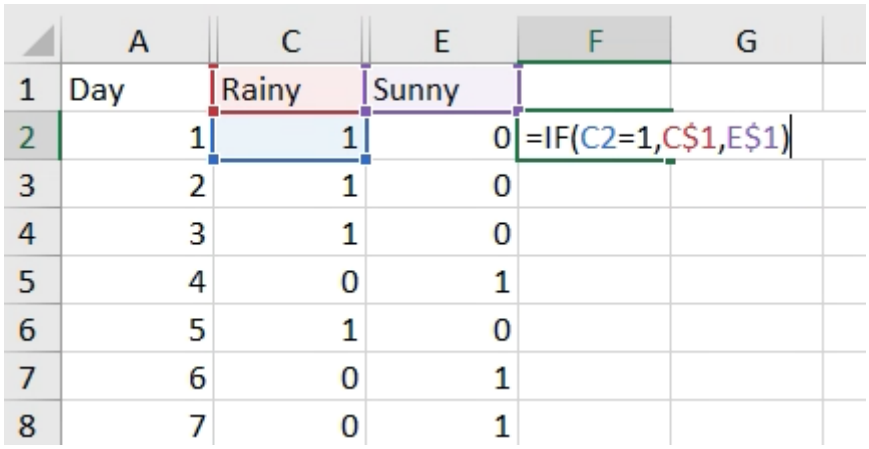
Hiding the columns we aren't looking at isn't crucial, but just means there is less to see while we are working with the 'Rainy' and 'Sunny' columns. It can be much easier if there are lots of columns in a file.

	A	C	E	F
1	Day	Rainy	Sunny	
2	1	1	0	
3	2	1	0	
4	3	1	0	
5	4	0	1	
6	5	1	0	
7	6	0	1	
8	7	0	1	
9	8	1	0	
10	9	1	0	
11	10	0	1	
12	11	0	1	
13	12	1	0	

It will be useful to have a single column that records whether a day was sunny

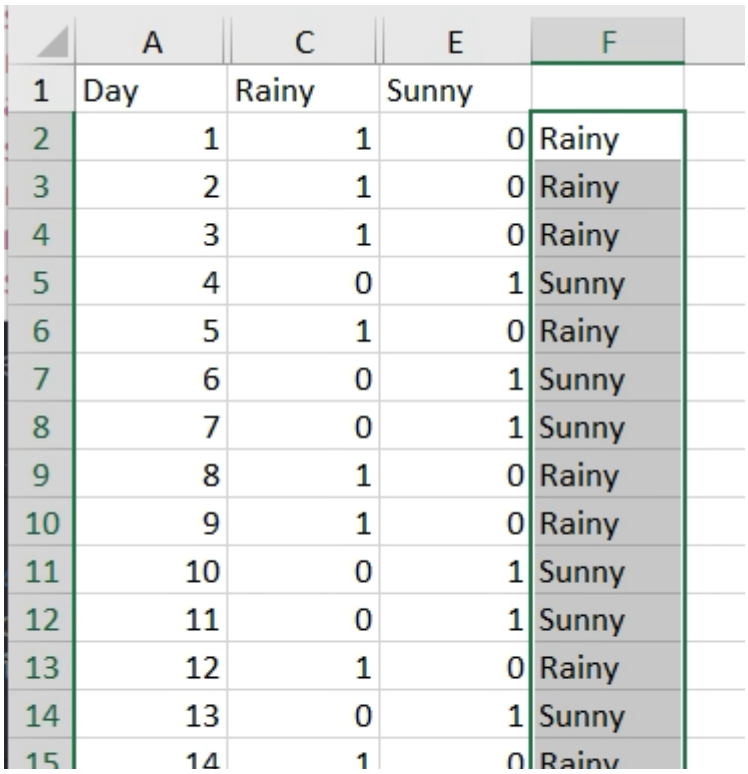
or rainy. To do this, we can use a formula.

In cell `F2`, insert the formula `=IF(C2=1,C$1,E$1)`. Here, because a day is always exactly one of Sunny or Rainy, we use a simple `IF` statement. The values `C$1` (Rainy) and `E$1` (Sunny) are used and fixed to the first row with `$`s.



	A	C	E	F	G
1	Day	Rainy	Sunny		
2	1	1	0	=IF(C2=1,C$1,E$1)	
3	2	1	0		
4	3	1	0		
5	4	0	1		
6	5	1	0		
7	6	0	1		
8	7	0	1		

Now drag this formula down column F to get values for each day.



	A	C	E	F
1	Day	Rainy	Sunny	
2	1	1	0	Rainy
3	2	1	0	Rainy
4	3	1	0	Rainy
5	4	0	1	Sunny
6	5	1	0	Rainy
7	6	0	1	Sunny
8	7	0	1	Sunny
9	8	1	0	Rainy
10	9	1	0	Rainy
11	10	0	1	Sunny
12	11	0	1	Sunny
13	12	1	0	Rainy
14	13	0	1	Sunny
15	14	1	0	Rainy

16	15	0	1	Sunny
17	16	0	1	Sunny

Now between each pair of cells in column F there is a transition - one day to the next. We can note this transition using an Excel function called `CONCAT`, which is short for 'concatenate', which means to join together.

Insert the formula `=CONCAT(F2," to ",F3)` into cell `G3`. This concatenates the previous value in column F with the text `" to "` and then the current value in column F. Note that we have started in row 3 because that is where the first transition happens.

	A	C	E	F	G	H
1	Day	Rainy	Sunny			
2	1	1	0	Rainy		
3	2	1	0	Rainy	=CONCAT(F2," to ",F3)	
4	3	1	0	Rainy		
5	4	0	1	Sunny		
6	5	1	0	Rainy		
7	6	0	1	Sunny		

Drag this formula down column G to get values for every transition.

	A	C	E	F	G	H
1	Day	Rainy	Sunny			
2	1	1	0	Rainy		
3	2	1	0	Rainy	Rainy to Rainy	
4	3	1	0	Rainy	Rainy to Rainy	
5	4	0	1	Sunny	Rainy to Sunny	

6	5	1	0	Rainy	Sunny to Rainy
7	6	0	1	Sunny	Rainy to Sunny
8	7	0	1	Sunny	Sunny to Sunny
9	8	1	0	Rainy	Sunny to Rainy
10	9	1	0	Rainy	Rainy to Rainy
11	10	0	1	Sunny	Rainy to Sunny
12	11	0	1	Sunny	Sunny to Sunny
13	12	1	0	Rainy	Sunny to Rainy

Now we have a record of every transition in the data, all we need to do is count them. We are looking to complete a 2 by 2 table of transitions like this:

	Sunny	**Rainy**
Sunny		
Rainy		

So set up those values as headings for a small table in cells `J2-L4`.

	J	K	L
		Sunny	Rainy
	Sunny		
	Rainy		

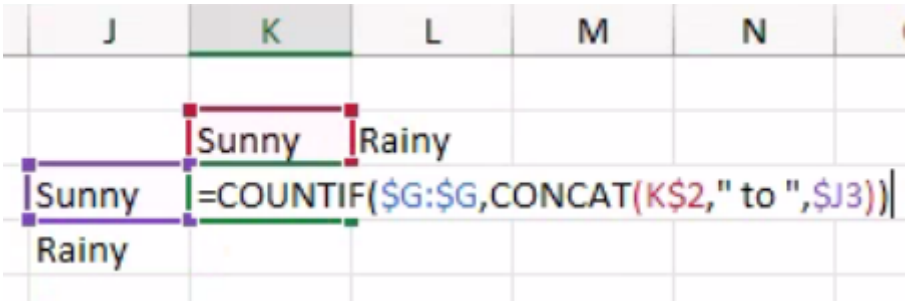
To count values in column `G` we will use `COUNTIF`. This counts entries only if they match some criteria. The criteria we are looking to match is determined by the column headings, so for example in cell `K3` we want to count based on the transition `Sunny to Sunny` using the values in `J3` and `K2`.

In cell `K3`, enter this formula: `=COUNTIF($G:$G,CONCAT(K$2," to ",$J3))`. There is a lot going on here.

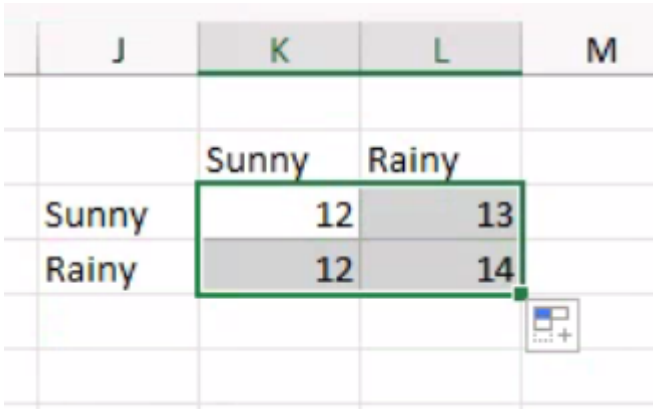
- The range we are counting over is column `G`, which can be entered as

`G:G`. Here we fix column `G` so that we can drag sideways but still count the same column.

- The condition we are trying to match is built by a `CONCAT` using the row and column heading from our table.
- We always want the row heading to come from column `J`, so we fix the column letter (but not the row number) using `$J3`.
- We always want the column heading to come from row 2, so we fix the row number (but not the column letter) using `K$2`.



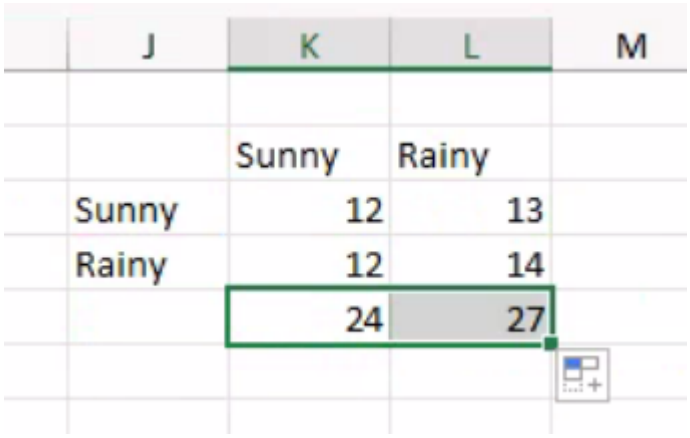
If you drag this formula first down one cell then drag a second time across one column, you will find that it counts column G using each combination of row and column headings 'Sunny to Sunny', 'Rainy to Sunny', 'Sunny to Rainy' and 'Rainy to Rainy'. As these are all possible combinations, we have counted all the transitions.



This is our matrix of transition counts.

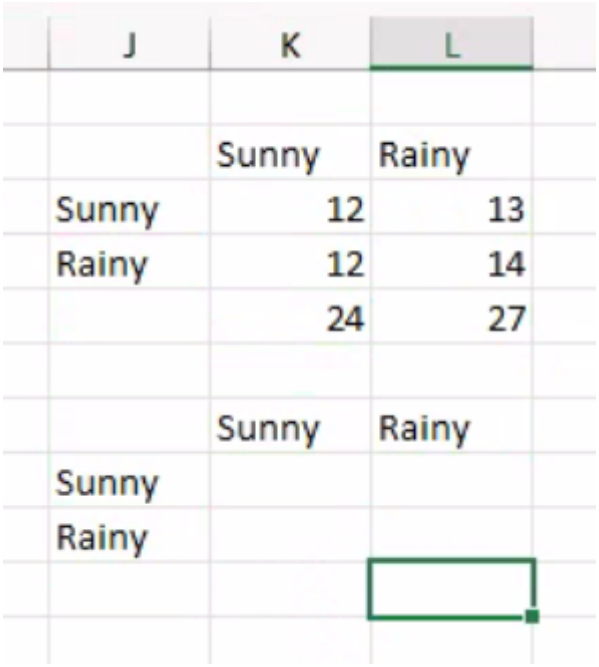
To work out the transition probabilities, we do some division, remembering that each row in a transition probability matrix should sum to 1.

First sum the columns using `SUM`. To do this, enter `=SUM(K3:K4)` into cell `K5` and then drag this across to cell `L5`.



	J	K	L	M
		Sunny	Rainy	
	Sunny	12	13	
	Rainy	12	14	
		24	27	

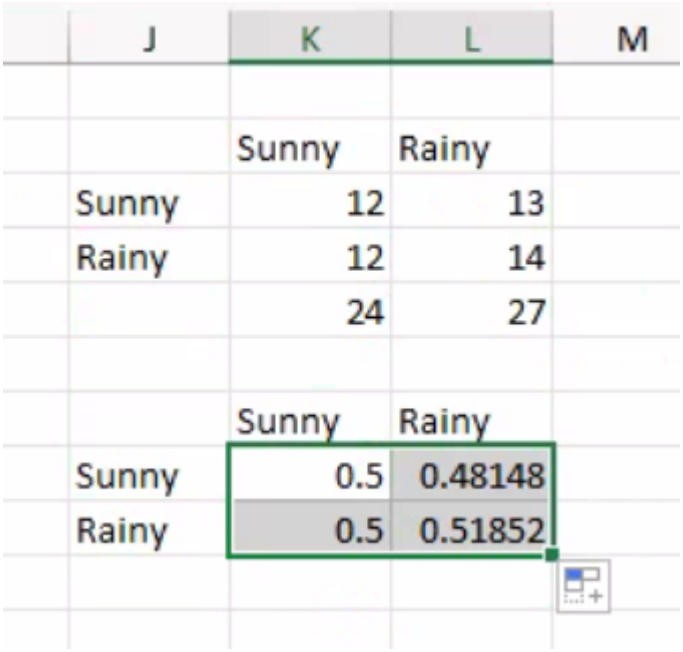
Now underneath we will calculate the transition probabilities. Set up the 'Sunny' and 'Rainy' headings in cells `J7` to `L9`.



	J	K	L
		Sunny	Rainy
	Sunny	12	13
	Rainy	12	14
		24	27
		Sunny	Rainy
	Sunny		
	Rainy		

In cell `K8` insert the formula `=K3/K$5`. This divides the transition count by the sum for this column, fixing the column sum to always come from row `5`.

sum for this column, fixing the column sum to always come from row `5`. Dragging this formula down and then across will complete the table of probabilities which are the entries for the transition probability matrix.



	J	K	L	M
		Sunny	Rainy	
	Sunny	12	13	
	Rainy	12	14	
		24	27	
		Sunny	Rainy	
	Sunny	0.5	0.48148	
	Rainy	0.5	0.51852	

We have calculated our table of transition counts and used this to calculate a transition probability matrix:

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.481 \\ 0.5 & 0.519 \end{bmatrix}$$

A complication

The two columns we hid can be revealed by selecting from column `A` to `E`, right-clicking on the column header and choosing 'Unhide'.

The data on wind involved two columns, 'Light wind' and 'Strong wind'. In this case, it is not true that every day is one or the other - there is also the option of no wind, which is not recorded in the spreadsheet.

Here we would use a double `IF` statement, like `=IF(D2=1,"Strong`

`wind", IF(B2=1, "Light wind", "No wind"))`. This says if `D2` is `1` then put "Strong wind". If this is not the case, it starts a second `IF` that says if `B2` is `1` then put "Light wind" and if neither condition is matched, put "No wind". We could then build a 3 by 3 table of transition counts based on this column.

Exercises

Use data from Town 2 to generate a table of transition counts and the corresponding transition probability matrix. Use this to answer these questions.

1. Do you think the data were generated using the same transition probabilities?
2. Find the long-term behaviour of these Markov chains. Are the two towns equally likely to experience rain?
3. Use the data from Town 3 to generate a matrix of transition probabilities for strong wind, light wind and no wind. Use this matrix to find the

   proportions of days you would expect for each type of wind in the long term. It is recommended you use technology to solve this problem.