FOSE1025 — Scientific Computing

Week 8 Lecture 1: Summarising and Analysing Data

Diego Mollá

FOSE1025 2020H1

Abstract

This lecture will focus on several approaches for summarising and preparing the data for the final analysis. We will look at pivot tables as a powerful tool to transform and summarising the data. With pivot tables we can convert tables from the long to the wide format. In addition, we can aggregate and filter data and make it ready for insightful analysis and graphic representations. Beside pivot tables, we will look at some specific tools that Excel provides for the analysis of data.

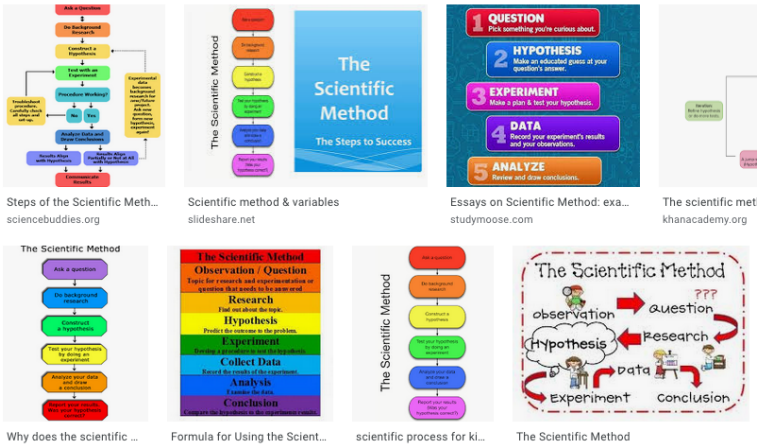
Update April 27, 2020

Contents

Reading

- These notes
- *https://www.linkedin.com/learning/excel-pivottables-for-beginners/*

The Scientific Method



Some results of a Google image search with the words "scientific" and "method" — 1 April 2020.

Excel to Manage Data in Science

We are covering these aspects in FOSE1025:

- Import data from external files (e.g. CSV) — Week 3.
- Explore the data — Week 4.
- Clean the data — Week 6.
- Preprocess, transform the data — Week 7.
- *Analyse, summarise, interpret the data — Week 8.*

1 Pivot Tables

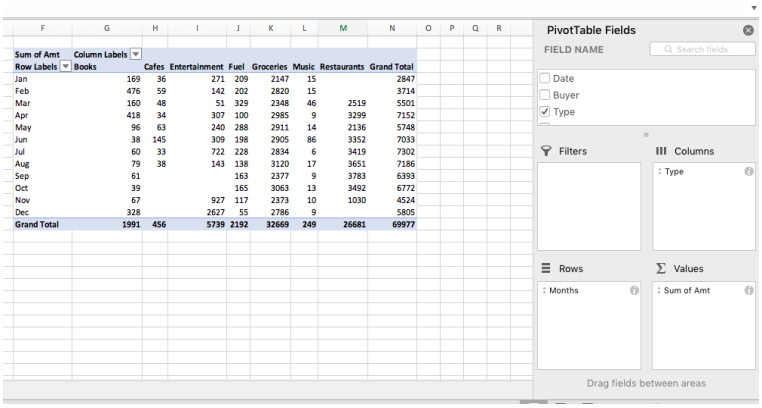
Pivot Tables: A Motivational Example

(data from *https://www.linkedin.com/learning/excel-pivottables-for-beginners*)

- Find the total shopping in each category "Fuel", etc, of file shopping.csv.
- Find the total shopping of each month.
- What shopping per month and per category??
- Pivot tables can help you generate data for all of above and more.

	Date	Buyer	Type	Amt
1	1-Jan	Mom	Fuel	$50
2	2-Jan	Mom	Groceries	$120
3	3-Jan	Dad	Cafes	$10
4	4-Jan	Dad	Fuel	$40
5	4-Jan	Kelly	Groceries	$129
6	5-Jan	Mom	Cafes	$12
7	6-Jan	Kelly	Cafes	$14
8	7-Jan	Kelly	Books	$129
9	7-Jan	Dad	Groceries	$252
10	9-Jan	Kelly	Fuel	$44
11	10-Jan	Dad	Groceries	$39
12	12-Jan	Mom	Books	$20
13	13-Jan	Dad	Groceries	$132
14	14-Jan	Dad	Groceries	$79
15	16-Jan	Kelly	Groceries	$172
16	16-Jan	Dad	Music	$8
17	18-Jan	Kelly	Fuel	$30

A Simple Pivot Table



Sum of Amt	Column Labels							
Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	36	271	209	2147	15		2847
Feb	476	59	142	202	2820	15		3714
Mar	160	48	51	329	2348	46	2519	5501
Apr	418	34	307	100	2985	9	3299	7152
May	96	63	240	288	2911	14	2136	5748
Jun	38	145	309	198	2905	86	3352	7033
Jul	60	33	722	228	2834	6	3419	7302
Aug	79	38	143	138	3120	17	3651	7186
Sep	61			163	2377	9	3783	6393
Oct	39			165	3063	13	3492	6772
Nov	67		927	117	2373	10	1030	4524
Dec	328		2627	55	2786	9		5805
Grand Total	1991	456	5739	2192	32669	249	26681	69977

Anatomy of a Pivot Table

Filters

- What column to use to filter values.
- Only for columns with categorical data.

Rows

- What column to use in the rows of the pivot table.
- Only for columns with categorical data.

Columns

- What column to use in the columns of the pivot table.
- Only for columns with categorical data.

Values

- What value we want to aggregate.
- Only for columns with numerical data.

Pivot Tables to Convert from Long to Wide

Exercise 1 (weather_data.csv)
What is the average precipitation in Antigo?

- Using AVERAGEIFS
- Using a pivot table

Exercise 2 (weather_data.csv)
What is the March-2013 average precipitation in Antigo?

- Using AVERAGEIFS

• Using a pivot table

	data	date	param	siteid
1	0	1/1/03	Precipitatior	ACRE
2	0	2/1/03	Precipitatior	AlbertLea
3	11.3199997	3/1/03	Precipitatior	Ames
4	0	4/1/03	Precipitatior	Antigo
5	3.03999996	5/1/03	Precipitatior	Appleton
6	0.49000001	6/1/03	Precipitatior	Arlington
7	0	7/1/03	Precipitatior	Bean&Beet
8	0	8/1/03	Precipitatior	Brookings
9	0	9/1/03	Precipitatior	Brownstown
10	0	10/1/03	Precipitatior	Columbia
11	0	11/1/03	Precipitatior	Crookston
12	0	12/1/03	Precipitatior	Dekalb
13	0	13/1/03	Precipitatior	DixonSprings

Pivot Tables for Charts

• Pivot tables facilitate the transformation of data for the creation of complex plots.

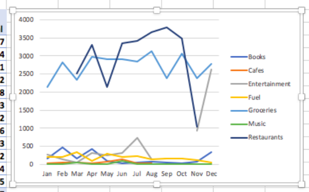
• In a *multiple chart*, each column of a table is plotted overlayed with the rest. Good for line plots.

• In a *clustered chart*, each row forms a cluster. Good for bar charts.

• In a *stacked chart*, columns of a table are plotted one on top of the other.

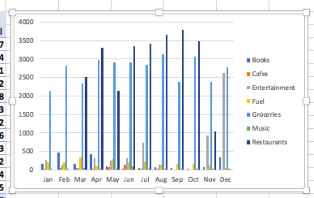
Buyer (All)

Sum of Amt / Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	36	271	209	2147	15		2847
Feb	476	59	142	202	2820	15		3714
Mar	160	48	51	329	2348	46	2519	5501
Apr	418	34	307	100	2985	9	3299	7152
May	96	63	240	288	2911	14	2136	5748
Jun	38	145	309	198	2905	86	3352	7033
Jul	60	33	722	228	2834	6	3419	7302
Aug	79	38	143	138	3120	17	3651	7186
Sep	61			163	2377	9	3783	6393
Oct	39			165	3063	13	3492	6772
Nov	67		927	117	2373	10	1030	4524
Dec	328		2627	55	2786	9		5805
Grand Total	1991	456	5739	2192	32669	249	26681	69977



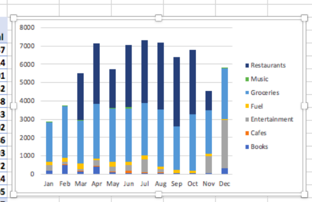
Buyer (All)

Sum of Amt / Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	36	271	209	2147	15		2847
Feb	476	59	142	202	2820	15		3714
Mar	160	48	51	329	2348	46	2519	5501
Apr	418	34	307	100	2985	9	3299	7152
May	96	63	240	288	2911	14	2136	5748
Jun	38	145	309	198	2905	86	3352	7033
Jul	60	33	722	228	2834	6	3419	7302
Aug	79	38	143	138	3120	17	3651	7186
Sep	61			163	2377	9	3783	6393
Oct	39			165	3063	13	3492	6772
Nov	67		927	117	2373	10	1030	4524
Dec	328		2627	55	2786	9		5805
Grand Total	1991	456	5739	2192	32669	249	26681	69977



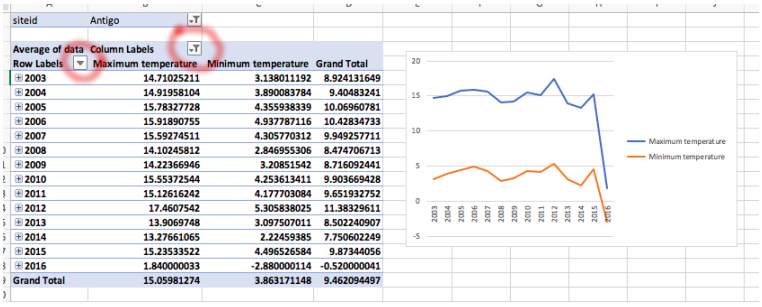
Buyer (All)

Sum of Amt / Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	36	271	209	2147	15		2847
Feb	476	59	142	202	2820	15		3714
Mar	160	48	51	329	2348	46	2519	5501
Apr	418	34	307	100	2985	9	3299	7152
May	96	63	240	288	2911	14	2136	5748
Jun	38	145	309	198	2905	86	3352	7033
Jul	60	33	722	228	2834	6	3419	7302
Aug	79	38	143	138	3120	17	3651	7186
Sep	61			163	2377	9	3783	6393
Oct	39			165	3063	13	3492	6772
Nov	67		927	117	2373	10	1030	4524
Dec	328		2627	55	2786	9		5805
Grand Total	1991	456	5739	2192	32669	249	26681	69977



Pivot Charts: Pivot Tables *and* Charts!

- Pivot tables are so useful for making charts that there's a tool for that combines both: Pivot charts!
- Exercise: Can you plot (multiple line plot) the maximum and minimum temperature of Antigo as it changes over time? Do not plot precipitation.
  - (hint: you can filter row labels and *column* labels.)



siteid	Antigo		
Average of data	Column Labels		
Row Labels	Maximum temperature	Minimum temperature	Grand Total
2003	14.71025211	3.138011192	8.924131649
2004	14.91958104	3.890083784	9.40483241
2005	15.78327728	4.355938339	10.06960781
2006	15.91890755	4.937787116	10.42834733
2007	15.59274511	4.305770312	9.949257711
2008	14.10245812	2.846955306	8.474706713
2009	14.22366946	3.20851542	8.716092441
2010	15.55372544	4.253613411	9.903669428
2011	15.12616242	4.177703084	9.651932752
2012	17.4607542	5.305838025	11.38329611
2013	13.9069748	3.097507011	8.502240907
2014	13.27661065	2.22459385	7.750602249
2015	15.23533522	4.496526584	9.87344056
2016	1.840000033	-2.880000114	-0.520000041
Grand Total	15.05981274	3.863171148	9.462094497

2 Data Analysis

Analysing the Data

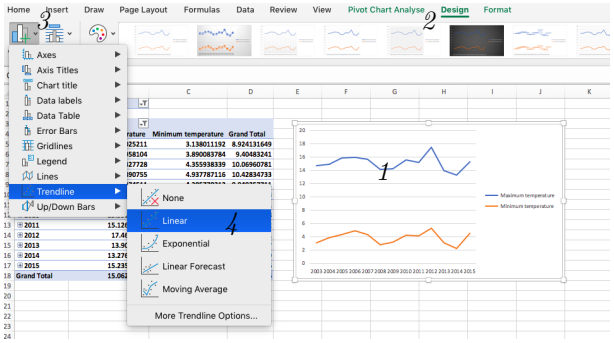
- Excel provides various tools for data analysis.
- Understanding most of these tools is beyond the scope of this unit.
- Here we will focus on two goals:
  - Finding trends.
  - Finding correlations.

2.1 Finding Trends

Adding a Trend Line

- Excel charts support the inclusion of a trend line.
- Select *chart* → *Design* → *Add Chart Element* → *Trendline*.
- Choose the kind of trendline based on what you want to show.

(this figure is based on MS Excel for Mac, Version 16.30, Office 365 Subscription)



2.2 Finding Correlations

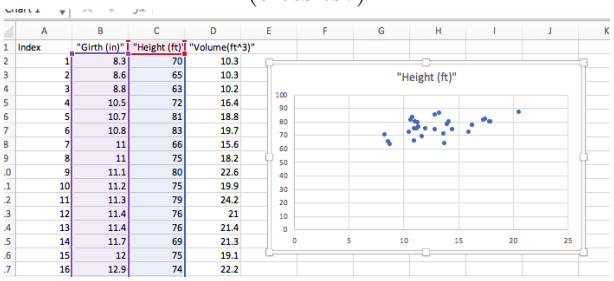
What is Correlation?

- Sometimes two variables are measuring the same property.
  - (each column represents one variable)
  - May happen when multiple agents are providing data.
- You may detect this by observing that the values are the same.
- But sometimes there are minor variations.
- In other cases, two variables are correlated but might not be identical.
  - For example, tree trunk height and girth are correlated.
  - Taller trees will normally have thicker trunks.

Finding Correlations Graphically

- A *scatterplot* can plot one variable against the other.
- If the two variables are not correlated, the scatterplots will look random.
- If the scatterplot has a distinct shape, the two variables are correlated.
- For example, if the shape looks like a line, then the two variables have a *linear correlation.*

(trees.csv)



Index	"Girth (in)"	"Height (ft)"	"Volume(ft^3)"
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11	66	15.6
8	11	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21
13	11.4	76	21.4
14	11.7	69	21.3
15	12	75	19.1
16	12.9	74	22.2

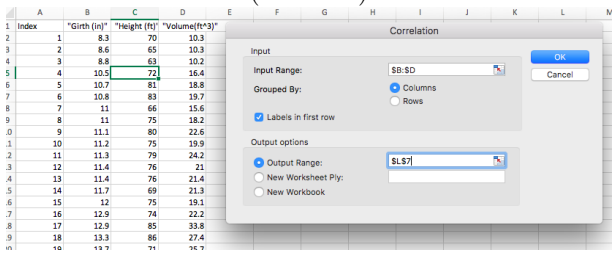
Finding Correlations on Multiple Columns

- Scatterplots are intuitive but may be cumbersome if you want to check the correlations among many columns.
- E.g. if there are 10 columns you will need to make a plot for each possible pair.
  - This means making $10 \times 9 = 90$ plots.
- There are various formulas that attempt to express the correlation as a number.
- Excel's CORREL function uses one of those formulas.
  - e.g. =CORREL(B:B,C:C) computes the correlation between columns B and C.
  - If you want to know what formula Excel uses, look for the "sample correlation coefficient".
- A number close to 1 (or -1) indicates positive (or negative) correlation.

Correlation Matrix

- Excel's "Data Analysis" tool can compute a correlation matrix.
- Data → Data Analysis → Correlation.

(trees.csv)



(you will observe a strong correlation between girth and volume)

Exercise

- File: shopping.png
- Build the correlation matrix between all types of shopping.
- What are the two most correlated types of shopping?
- Show it clearly by introducing *conditional formatting* that highlights the highest correlations.
  - Home → Conditional Formatting → Colour Scales

	Books	*Cafes*	*Entertainment*	*Fuel*	*Groceries*	*Music*	*Restaurants*
Books	1						
Cafes	-0.289396228	1					
Entertainment	0.160093641	-0.08	1				
Fuel	-0.271487084	0.09	-0.625410842	1			
Groceries	0.09428483	0.19	-0.000504711	-0.2	1		
Music	-0.243270987	0.88	-0.285322756	0.34	0.026135	1	
Restaurants	0.030060483	-0	-0.470731464	-0.1	0.470645	0.071	1

Take-home Messages

- Pivot tables are very powerful to process tables in long format.
- You must be able to use pivot tables for a range of tasks.
- You must be able to create charts based on pivot tables.
- You must be able to show trends by adding trend lines to a plot.
- You must be able to detect whether two variables are correlated.

What's Next

- Week 9 lecture: Ethics related to Scientific Computing.
- Week 9: Submit the project.