Advanced Excel Skills for Enterprise Data Analytics

Understand how to use Data Analysis, What-If Analysis, and Solver in Excel with examples

Overview

In this post, I will use several tools in Excel to solve business scenario problems and make data-driven decisions across different industries.

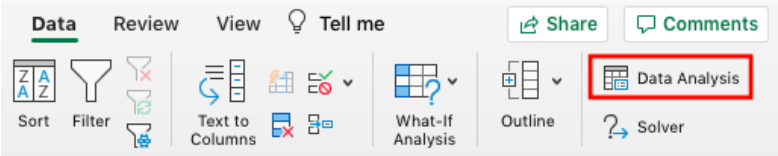
Introduction

Enterprise analytics refers to the process of implementing statistical analysis into business processes to drive business strategies and actions. In my second quarter in graduate school, I took a course called “Introduction to Enterprise Analytics.” In that course, we learned to use analysis techniques such as random number generation, simulation, and time series decomposition/forecasting in Excel, which is one of the most widely used tools for data analysis. To summarize what I have learned, I would like to introduce some powerful analysis features in Excel.

Table of Contents

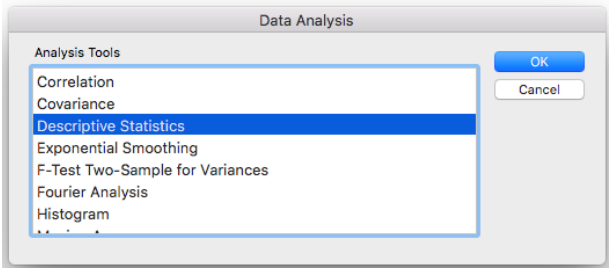
1. Data Analysis

The Analysis ToolPak is an Excel add-in program that provides data analysis tools, including Histogram, Descriptive Statistics, Anova, F-Test, t-Test, Moving Average, Exponential Smoothing, Correlation, Regression, etc. To access those tools, we have to select the Data tab and click Data Analysis after ensuring that Analysis ToolPak is loaded.



a) Descriptive Statistics

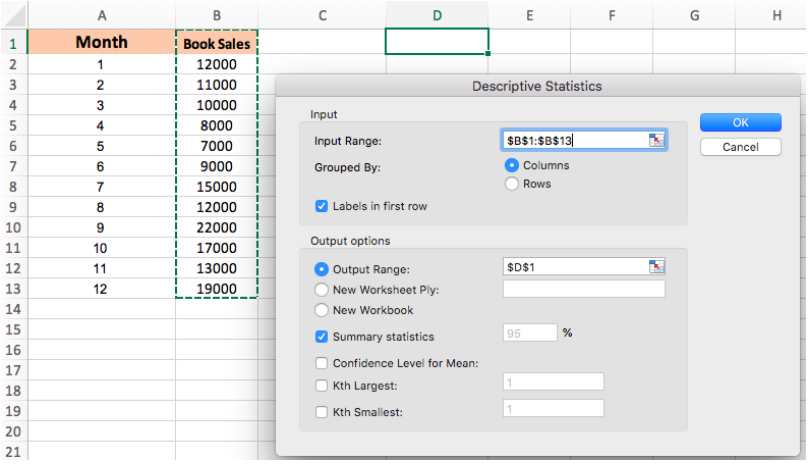
Descriptive Statistics, which are distinguished from inferential statistics, describe the basic features of the data.



Example: Describe and Summarize Sales Volume for a Book Store

There is book sales data of last year (12 months). Please generate descriptive statistics for this book sales data.

1. Select Descriptive Statistics in Data Analysis tool.
2. Select the range B1:B13 as the Input Range.
3. Select cell D1 as the Output Range.
4. Check Labels in first row and make sure Summary statistics is checked.



5. Click OK, then get the result.

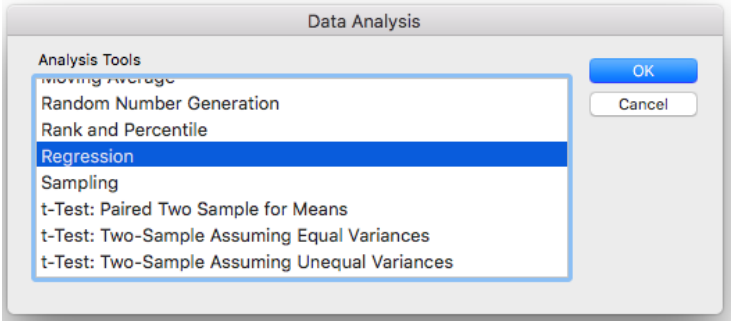
6. Interpret the result.

D	E
Book Sales	
Mean	12916.66667
Standard Error	1316.896796
Median	12000
Mode	12000
Standard Deviation	4561.864319
Sample Variance	20810606.06
Kurtosis	-0.174286673
Skewness	0.726508611
Range	15000
Minimum	7000
Maximum	22000
Sum	155000
Count	12

From the table above, we can tell the summary of book sales data. For example, we can use mean, median, and mode to measure central tendency and use standard deviation, sample variance, and range to measure variability.

b) Regression

In Excel, we can run a linear regression analysis and perform subset selection methods such as backward stepwise regression.

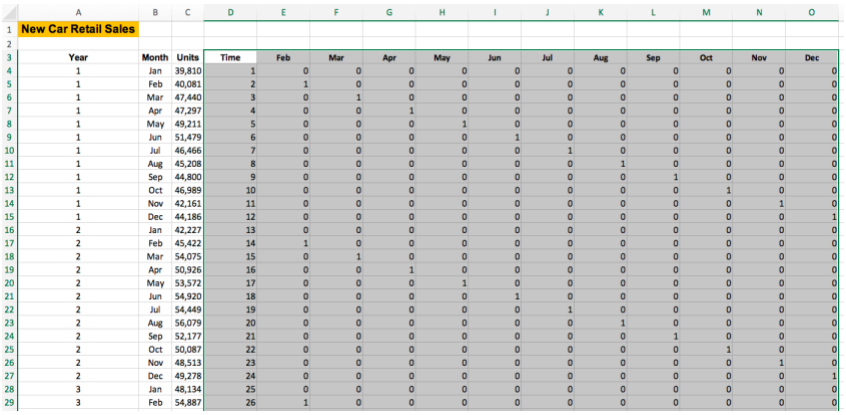


Example: Forecast Sales for a Car Retail Company

Here we have the last three years of new car retail sales data. Please develop a multiple regression model with categorical variables that incorporate seasonality for forecasting sales.

Data Preparation

1. To incorporate seasonality, I created a new independent variable called “Time”.
2. Since the “Month” is a categorical variable with more than two levels, I add 11 additional dummy variables (Feb, Mar, …, Nov, and Dec) using IF() function.

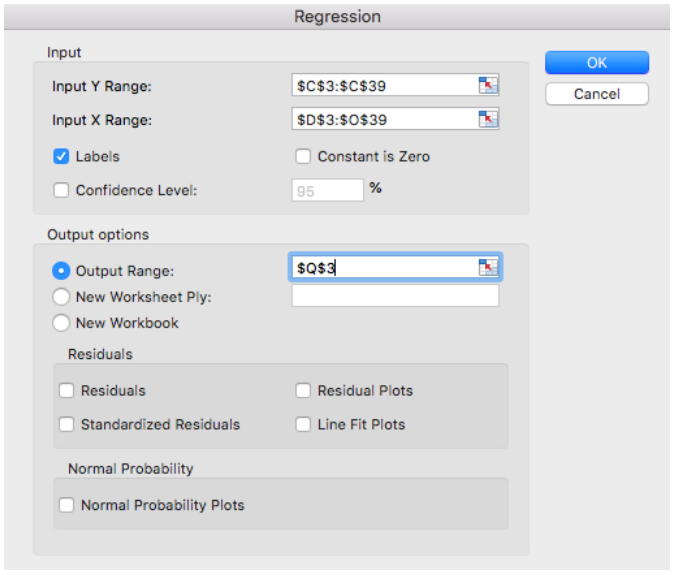


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	New Car Retail Sales														
2															
3	Year	Month	Units	Time	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
4	1	Jan	39,810	1	0	0	0	0	0	0	0	0	0	0	0
5	1	Feb	40,081	2	1	0	0	0	0	0	0	0	0	0	0
6	1	Mar	47,440	3	0	1	0	0	0	0	0	0	0	0	0
7	1	Apr	47,297	4	0	0	1	0	0	0	0	0	0	0	0
8	1	May	49,211	5	0	0	0	1	0	0	0	0	0	0	0
9	1	Jun	51,479	6	0	0	0	0	1	0	0	0	0	0	0
10	1	Jul	46,466	7	0	0	0	0	0	1	0	0	0	0	0
11	1	Aug	45,208	8	0	0	0	0	0	0	1	0	0	0	0
12	1	Sep	44,800	9	0	0	0	0	0	0	0	1	0	0	0
13	1	Oct	46,989	10	0	0	0	0	0	0	0	0	1	0	0
14	1	Nov	42,161	11	0	0	0	0	0	0	0	0	0	1	0
15	1	Dec	44,186	12	0	0	0	0	0	0	0	0	0	0	1
16	2	Jan	42,227	13	0	0	0	0	0	0	0	0	0	0	0
17	2	Feb	45,422	14	1	0	0	0	0	0	0	0	0	0	0
18	2	Mar	54,075	15	0	1	0	0	0	0	0	0	0	0	0
19	2	Apr	50,926	16	0	0	1	0	0	0	0	0	0	0	0
20	2	May	53,572	17	0	0	0	1	0	0	0	0	0	0	0
21	2	Jun	54,920	18	0	0	0	0	1	0	0	0	0	0	0
22	2	Jul	54,449	19	0	0	0	0	0	1	0	0	0	0	0
23	2	Aug	56,079	20	0	0	0	0	0	0	1	0	0	0	0
24	2	Sep	52,177	21	0	0	0	0	0	0	0	1	0	0	0
25	2	Oct	50,087	22	0	0	0	0	0	0	0	0	1	0	0
26	2	Nov	48,513	23	0	0	0	0	0	0	0	0	0	1	0
27	2	Dec	49,278	24	0	0	0	0	0	0	0	0	0	0	1
28	3	Jan	48,134	25	0	0	0	0	0	0	0	0	0	0	0
29	3	Feb	54,887	26	1	0	0	0	0	0	0	0	0	0	0

Multiple Linear Regression

Model 1

1. Select Regression in the Data Analysis tool.
2. Select the Y Range (C3:C39). These are the dependent variable.
3. Select the X Range (D3:O39). These are independent variables.
4. Check Labels.
5. Select cell Q3 as the Output Range.
6. Click OK, then get the result.



7. Interpret the Summary Output.

Model 1								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.943951039							
R Square	0.891043565							
Adjusted R Square	0.834196729							
Standard Error	2256.508434							
Observations	36							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	12	957740403.6	79811700.3	15.67446192	2.61E-08			
Residual	23	117112097.2	5091830.312					
Total	35	1074852501						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38471.72917	1395.087678	27.57656724	3.98965E-19	35585.77042	41357.68791	35585.77042	41357.68791
Time	378.3541667	38.38398794	9.857083305	1.00233E-09	298.9508379	457.7574954	298.9508379	457.7574954
Feb	3027.979167	1842.831211	1.643112592	0.113962202	-784.2076423	6840.165976	-784.2076423	6840.165976
Mar	10045.95833	1844.03006	5.447827858	1.54664E-05	6231.291515	13860.62515	6231.291515	13860.62515
Apr	5998.9375	1846.026413	3.249648791	0.0035328	2180.140913	9817.734087	2180.140913	9817.734087
May	9179.583333	1848.817684	4.965110087	5.08587E-05	5355.012561	13004.15411	5355.012561	13004.15411
Jun	9974.229167	1852.400282	5.384489122	1.80641E-05	6142.247226	13806.21111	6142.247226	13806.21111
Jul	6340.541667	1856.769625	3.414824102	0.002371736	2499.521051	10181.56228	2499.521051	10181.56228
Aug	7506.520833	1861.920175	4.031601857	0.00052007	3654.845494	11358.19617	3654.845494	11358.19617
Sep	4065.833333	1867.845469	2.17675038	0.040017315	201.9005894	7929.766077	201.9005894	7929.766077
Oct	3284.479167	1874.53816	1.752153803	0.093071271	-593.2984644	7162.256798	-593.2984644	7162.256798
Nov	-684.875	1881.990062	-0.363909998	0.719249164	-4578.068065	3208.318065	-4578.068065	3208.318065
Dec	-745.5625	1890.192195	-0.394437403	0.696891813	-4655.72297	3164.59797	-4655.72297	3164.59797

R Square & Adjusted R Square

R Square is around 0.89, indicating that 89% of the variation in units is explained by the independent variables. On the other hand, the adjusted R Square, which is a modified version of R Square that has been adjusted for the number of terms in a model, is around 0.834.

P-values

The P-value for each independent variable tests the null hypothesis that the independent variable has no correlation with the dependent variable (coefficient=0). When a P-value is low (< 0.05), we can reject the null hypothesis, and state that the coefficient does not equal zero and there is a non-zero correlation.

The output showed that the P-value of "Feb", "Oct", "Nov", and "Dec" are greater than the significance level 0.05, and "Nov" got the highest P-value (0.71). Therefore, I would like to perform the subset selection method to delete a variable with a high P-value.

Backward Stepwise Regression

Model 2 (exclude "Nov")

Remove the "Nov" column from the data set and redo steps 1-6.

7. Interpret the Summary Output.

Model 2 (exclude "Nov")								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.943618678							
R Square	0.89041621							
Adjusted R Square	0.840190306							
Standard Error	2215.34808							
Observations	36							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	11	957066090	87006008.18	17.72822674	6.34245E-09			
Residual	24	117786410.7	4907767.114					
Total	35	1074852501						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38180.57182	1122.018398	34.02847216	7.99511E-22	35864.83967	40496.30398	35864.83967	40496.30398
Time	375.505269	36.89173768	10.17857365	3.47498E-10	299.3644647	451.6460733	299.3644647	451.6460733
Feb	3359.021076	1573.422881	2.134849516	0.043186289	111.6358554	6606.406297	111.6358554	6606.406297
Mar	10379.84914	1570.392486	6.609716509	7.76221E-07	7138.718348	13620.97993	7138.718348	13620.97993
Apr	6335.677205	1568.224333	4.040032457	0.000476191	3099.02126	9572.33315	3099.02126	9572.33315
May	9519.171936	1566.922001	6.075077081	2.83374E-06	6285.203872	12753.14	6285.203872	12753.14
Jun	10316.66667	1566.48765	6.585858923	8.21822E-07	7083.595059	13549.73827	7083.595059	13549.73827
Jul	6685.828064	1566.922001	4.266854419	0.000267939	3451.86	9919.796128	3451.86	9919.796128
Aug	7854.656129	1568.224333	5.008630439	4.06693E-05	4618.000184	11091.31207	4618.000184	11091.31207
Sep	4416.817526	1570.392486	2.812556457	0.009643569	1175.686734	7657.948319	1175.686734	7657.948319
Oct	3638.312257	1573.422881	2.312354995	0.029650402	390.9270368	6885.697478	390.9270368	6885.697478
Dec	-386.031614	1582.049167	-0.244007343	0.809298987	-3651.220615	2879.157387	-3651.220615	2879.157387

R Square & Adjusted R Square

R Square is around 0.89 and the adjusted R Square is around 0.84.

P-values

The output shows that the P-value of "Dec" is greater than the significance level 0.05. Therefore, "Dec" has to be excluded in next model.

Model 3 (exclude "Nov" and "Dec")

Remove the "Dec" column from the latest data set and redo steps 1-6.

7. Interpret the Summary Output.

Model 3 (exclude "Nov" and "Dec")								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.943474617							
R Square	0.890144353							
Adjusted R Square	0.846202094							
Standard Error	2173.279707							
Observations	36							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	10	956773883.6	95677388.36	20.25713687	1.39131E-09			
Residual	25	118078617.1	4723144.686					
Total	35	1074852501						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38077.08428	1019.047747	37.36535838	1.85511E-23	35978.31616	40175.85241	35978.31616	40175.85241
Time	374.2457858	35.8351942	10.44352609	1.32599E-10	300.4418217	448.0497498	300.4418217	448.0497498
Feb	3480.141381	1464.720184	2.375976941	0.025479173	463.4936935	6496.789069	463.4936935	6496.789069
Mar	10502.22893	1459.890217	7.193848421	1.53959E-07	7495.528745	13508.92911	7495.528745	13508.92911
Apr	6459.316476	1455.926507	4.436567673	0.000160522	3460.779705	9457.853248	3460.779705	9457.853248
May	9644.070691	1452.836146	6.638099361	5.9029E-07	6651.898638	12636.24274	6651.898638	12636.24274
Jun	10442.8249	1450.624714	7.19884668	1.52135E-07	7455.207381	13430.44243	7455.207381	13430.44243
Jul	6813.245786	1449.296235	4.701071886	8.0858E-05	3828.364315	9798.127256	3828.364315	9798.127256
Aug	7983.333333	1448.853138	5.510105285	1.0019E-05	4999.364438	10967.30223	4999.364438	10967.30223
Sep	4546.754214	1449.296235	3.137215225	0.004332136	1561.872744	7531.635685	1561.872744	7531.635685
Oct	3769.508428	1450.624714	2.598541437	0.015476691	781.8909047	6757.125952	781.8909047	6757.125952

R Square & Adjusted R Square

R Square is around 0.89 and the adjusted R Square is nearly 0.85, which is greater than the values of model 1 and 2. When comparing models that have a different amount of

variables, adjusted R Square is better than R Square.

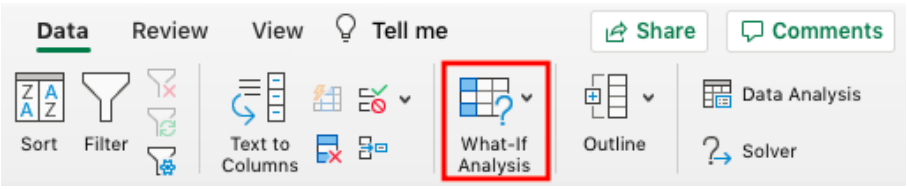
P-values

This time all P-values are acceptably smaller than the significance level 0.05.

To sum up, by deleting one variable at each step, the independent variables in the final model are all significant to forecast.

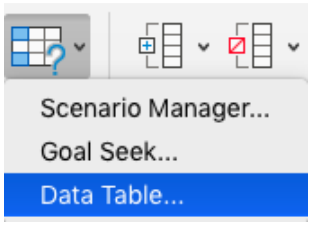
2. What-If Analysis

What-If Analysis in Excel allows us to use several different sets of values in a given simulation model to explore all the various results. There are three kinds of What-If Analysis tools: Scenario Manager, Data Table, and Goal Seek. To access these tools, you can select the Data tab and click What-If Analysis.



a) Data Table

Using a data table, you can experiment with different inputs and compare the results. A data table can be a one variable data table or a two variable data table.



One Variable Data Table

Example: Set Price on a New Medicine for a Pharmaceutical Company

A pharmaceutical company invented new medicine and prepares to set the price that maximize profit. Demand is thought to depend on the price.

Decision Model

Decision Variable: Price

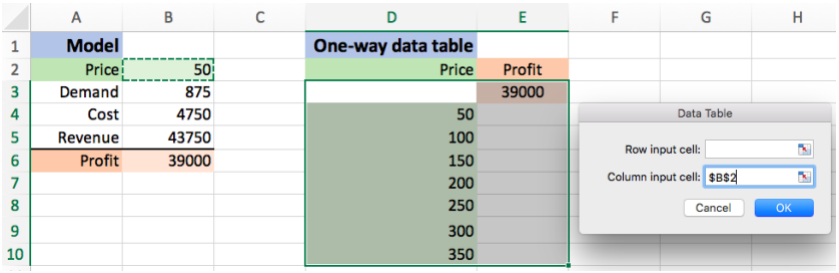
Demand = 1,000 – 2.5*Price

Cost = 3,000 + 2*Demand

Revenue = Price*Demand

Profit = Revenue – Cost

1. Use formula to create decision model (set Price=50).
2. Create a One-way data table that has different prices.
3. Select cell E3 and type =B6 (refer to the Profit cell).
4. Select the range D3:E10 to calculate the profit if set to higher prices.
5. Click Data Table in the What-If Analysis tool.
6. Select cell B2 as the Column input cell and leave the Row input cell blank.
6. Click OK, then get the result.



7. Summarize the result.

One-way data table	
Price	Profit
	39000
50	39000
100	70500
150	89500
200	96000
250	90000
300	71500
350	40500

Among those prices, if the price is set at $200, the company will get the highest profit of $96,000.

Two Variable Data Table

Example: Decide whether Outsourced or In-House for a Shoe Company

A shoe company is going to decide outsourced or in-house. To be more specific, if the spending of manufacturing is higher than the purchasing cost from the supplier, a company will choose to outsource.

Decision Model

Decision variables: Fixed cost, Unit variable cost

Total manufacturing = Fixed cost + Unit variable cost*Demand Volume

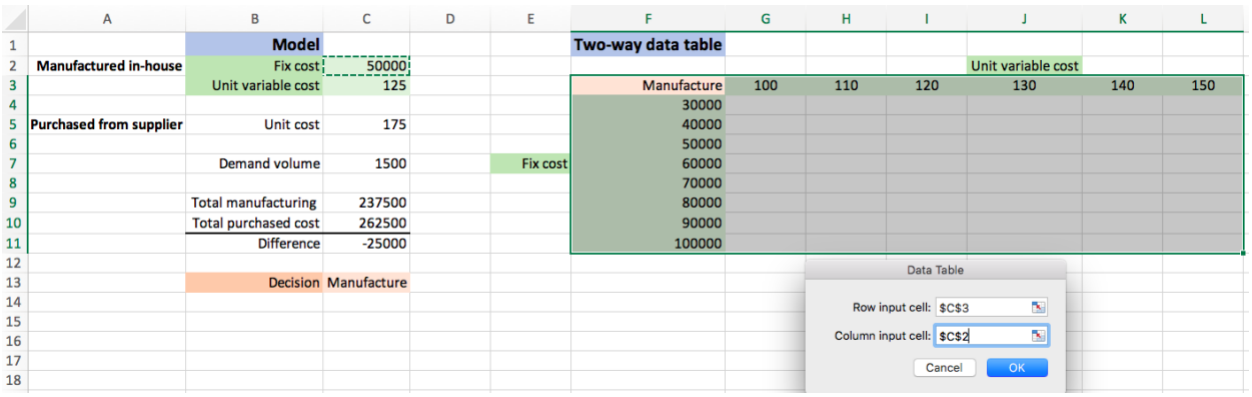
Total purchased = Unit cost Purchased from supplier*Demand Volume

Unit cost Purchased from supplier = 175

Demand Volume = 1500

Decision = Manufacture / Outsource

1. Use formula to create decision model (set Fixed cost=50000, Unit variable cost=125, and Decision cell using IF() function).
2. Create Two-way data table that with different Fixed cost and Unit variable cost.
3. Select cell F3 and type =C13 (refer to the Decision cell).
4. Select the range F3:L11 to make the decision depending on different decision variable values.
5. Click Data Table in the What-If Analysis tool.
6. Select cell C2 as the Column input cell and cell C3 as the Row input cell.
7. Click OK, then get the result.



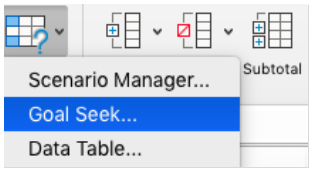
8. Summarize the result.

	E	F	G	H	I	J	K	L
1		**Two-way data table**						
2						Unit variable cost		
3		Manufacture	100	110	120	130	140	150
4		30000	Manufacture	Manufacture	Manufacture	Manufacture	Manufacture	Manufacture
5		40000	Manufacture	Manufacture	Manufacture	Manufacture	Manufacture	Outsource
6		50000	Manufacture	Manufacture	Manufacture	Manufacture	Manufacture	Outsource
7	Fix cost	60000	Manufacture	Manufacture	Manufacture	Manufacture	Outsource	Outsource
8		70000	Manufacture	Manufacture	Manufacture	Outsource	Outsource	Outsource
9		80000	Manufacture	Manufacture	Manufacture	Outsource	Outsource	Outsource
10		90000	Manufacture	Manufacture	Outsource	Outsource	Outsource	Outsource
11		100000	Manufacture	Outsource	Outsource	Outsource	Outsource	Outsource

From the table above, we can determine what decision to make when Unit variable cost and Fixed cost have different values.

b) Goal Seek

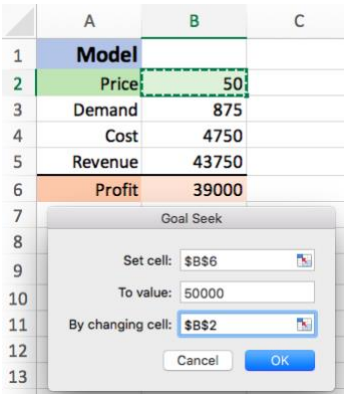
Goal Seek is useful when you know the exact result you want from a formula. That is, it can help you to find the input value that produces the specific result.



Example: Set Price on a New Medicine for a Pharmaceutical Company

Take the example we mentioned before when making the One-way data table. However, instead of finding maximum profit, this time assume that the company only set the goal $50,000 in profit.

1. Use the same model created before.
2. Click Goal Seek in the What-If Analysis tool.
3. Select cell B6 as the 'Set cell'.
4. Click in the 'To value' box and type 50000.
5. Select cell B2 as the 'By changing cell'.
6. Click OK, then get the result.



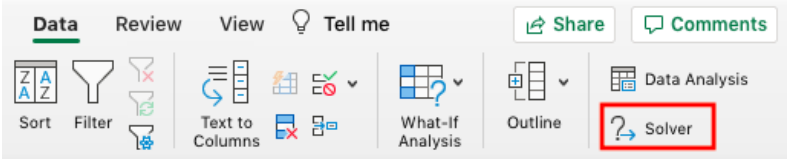
7. Interpret the result.

	A	B
1	**Model**	
2	Price	65.349714
3	Demand	836.62571
4	Cost	4673.2514
5	Revenue	54673.251
6	Profit	50000

When the company sets the price at around $65.35, the profit target can be achieved.

3. Solver

Solver is another Excel add-in tool used for optimization and equation solving. To be more specific, you can get the desired result that is subject to constraints to solve various business or programming problems. To access this tool, you can select the Data tab and click Solver.



a) Linear Optimization Model

Linear optimization, which is also called linear programming, is an optimization technique when objective function and constraints of a model are all linear functions of decision variables.

Example: Set Production Quantity for a Toy Company

A toy company produces two toys: Toy A and Toy B. There are some constraints that exist in the budget and labor hour. How many of each toy should the company produce to maximize profit?

	Cost	Labor	Profit
Toy A	3	2	5
Toy B	4	3	7

Decision Model

Decision variables: Production Quantity of Toy A and B

Constraints: Total Cost (3A + 4B) <= 720 (Budget)

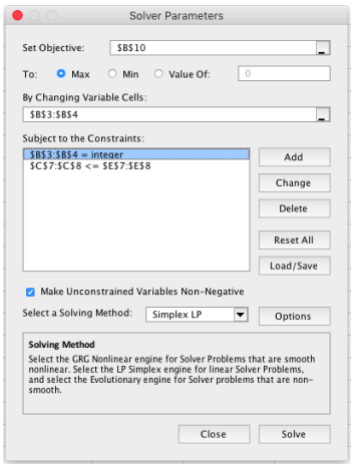
Total Labor (2A + 3B) <= 500

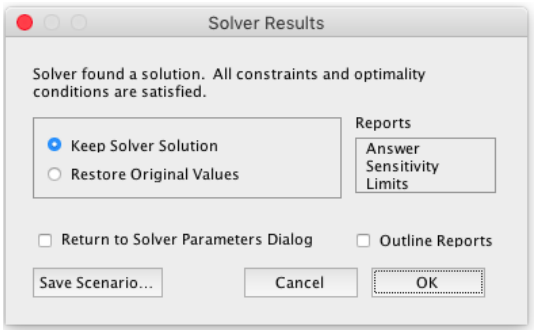
A = Integer

B = Integer

Objective: Max Total Profit (5A + 7B)

1. Use formula to create decision model (set Toy A=1, Toy B=1; Total Cost, Total Profit, and Total Labor cell using the SUMPRODUCT() function).
2. Click Solver.
3. Select cell B10 as the Set Objective and click to Max.
4. Select the range B3:B4 as the By Changing Variable Cells.
6. Click Add to enter the constraints.
7. Check Make Unconstrained Variables Non-Negative and select Simplex LP as Solving Method.
8. Finally, click Solve then get the result. You can also choose to get other reports such as Sensitivity.





9. Interpret the result.

It is optimal to produce 160 of Toy A and 60 of Toy B. This simple optimal solution gives the maximum profit of 1220 and uses all the resources available.

	A	B	C	D	E
1	Model				
2		Quantity	Cost	Labor	Prpfit
3	Toy A	160	3	2	5
4	Toy B	60	4	3	7
5					
6					
7	Constraints	Budget	720	<=	720
8		Total Labor	500	<=	500
9					
10	Total Profit	1220			

b) Non-linear Optimization Model

In contrast to the linear optimization model, the objective function and/or constraints functions of the non-linear optimization model are non-linear functions of decision variables. The steps are similar to previous linear optimization models. That is, you have to define Decision variables, Constraints, and Objective of the model before using Solver. However, instead of using 'Simplex LP', a Solving Method should be changed to 'GRG Nonlinear'.

Conclusion

So far, we have discussed several tools in Excel to solve different business problems. In conclusion, I believed that if we use Excel productively, it can be our secret weapon for business analysis!

See/ Download Excel file: Advanced Excel Skills_Kuan-Pei (Yuki) Lai.xlsx

References

TECHNOLOGY Q&A. (n.d.). Retrieved from https://www.hcltech.com/technology-qa/what-is-enterprise-analytics

Data Analysis. (n.d.). Retrieved from https://www.excel-easy.com/data-analysis.html