



Basic Techniques for the Analysis of Customer Information Using Excel 2007: A Step-by-Step Approach

The objective of this note is to provide a set of easy, step-by-step guides for some analytical techniques that are useful in the analysis of cases discussed in the course Competing and Winning through Customer Information (CWCI). The instructions that follow use datasets from three of the cases in this course: “MercadoLibre.com”;¹ “Slots, Tables, and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel”;² and “Bancaja: Developing Customer Intelligence (A).”³ These datasets are available upon request from the author.

Each technique in the note is explained using Microsoft Excel 2007.⁴ This is not because it is the best software to perform each analysis—in fact, there may be superior programs—but because it is a tool that all students will have at their disposal in their future professional roles. To allow for flexibility in the instructor’s teaching plan, each of the tutorials that follow is designated by a separate section.

This note does not attempt to treat all possible relevant techniques in a comprehensive manner. Rather, the techniques are taught with an emphasis on the mechanics, on practical applications to specific situations, and are expected to evolve along with the course.

¹ F. Asís Martínez Jerez, Joshua Bellin, and James Dillon, HBS No. 106-057.

² Dennis Campbell, F. Asís Martínez Jerez, Marc Epstein, and Joshua Bellin, HBS No. 106-029.

³ F. Asís Martínez Jerez and Katherine Miller, HBS No. 107-055.

⁴ This note is also available for Excel 2003 (HBS No. 107-073).

Professor F. Asís Martínez Jerez prepared this note as the basis for class discussion.

Copyright © 2009 President and Fellows of Harvard College. To order copies or request permission to reproduce materials, call 1-800-545-7685, write Harvard Business School Publishing, Boston, MA 02163, or go to www.hbsp.harvard.edu/educators. This publication may not be digitized, photocopied, or otherwise reproduced, posted, or transmitted, without the permission of Harvard Business School.

Table of Contents

Mercado Libre.com	
Note on How to Calculate Conditional Means with Excel	3
Slots, Tables, and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel	
Introduction to Pivot Tables	10
Merging Datasets in Excel (the Function VLOOKUP)	22
Introduction to Basic Regression Analysis Using Excel	24
Bancaja: Developing Customer Intelligence	
How to generate all possible combinations of 0-1 in Excel	36

MercadoLibre.com

Note on How to Calculate Conditional Means with Excel

The conditional mean is the average of a variable for all the individuals in the population that satisfy a condition or a set of conditions. The unconditional mean is the average of a variable for all the individuals in a population. For example, an unconditional mean is the average age of all the students in the MBA program at HBS. A conditional mean would be the average age of all the students in the Harvard MBA program who are foreigners. Another conditional mean would be the average age of all the students in the Harvard MBA program who are foreigners and have worked in consulting.

The function conditional mean DАVERAGE() in Excel is a database function that is useful when you need to calculate the average of a variable for all the observations that satisfy more than one restriction. If the observations cannot be sorted in such a way that all the necessary averages can be calculated with consecutive observations (consecutive rows or columns), the conditional mean is the only option to perform and store all these calculations in one spreadsheet. Otherwise, it may be more efficient to use Pivot Tables.

The function DАVERAGE() may come in handy to analyze the MercadoLibre.com datasets. I will use the listings dataset, or more specifically, the observations of the Argentina site in the listings dataset, to explain how it works.

Average subject to one condition I start with the simple case of one single condition. We may want to calculate the average number of new paid listings before January 14—the first free listing day.⁵

Step 1: Preparing the data

Prior to calculating the mean we have to transform the day column—which contains text strings with dates of the observation—into a column with numeric dates. I suggest: (i) Insert a new column between the day column and the Live Listings column. (ii) Write the dates 11/01/04 and 11/02/04 in the first and second rows of that column. (iii) Place the cursor in the lower-right corner of the second row (see **Figure A**) and double-click so consecutive dates until 06/30/05 are automatically generated.

Before proceeding to the next step, please label the newly created column “date.” All columns containing variables that will be evaluated in the condition need to be labeled.

Step 2: Writing the condition

Next, to the right of the dataset we are going to write the condition. When using the conditional mean function in a spreadsheet with many columns, you may want to insert a couple of rows above the dataset to write the conditions. For this example place the cursor in the cell G3 (please note that the columns with the Brazil and Mexico listings have been deleted). To write the condition: (i) Copy the label of the variable we are going to condition on (in this case, the date). I suggest copying to avoid generating an error with a spelling mistake. (ii) In the cell below write the conditioning statement “<01/14/05” (see **Figure B**).

⁵ If we were not learning how to use the conditional mean function we would just calculate the average of the first 74 observations (rows 4 to 77) or use the function AVERAGEIF().

Figure A

mercadolibre_conditional_means_examples - Microsoft Excel

HomeInsertPage LayoutFormulasDataReviewView

PasteClipboard

Font

Alignment

Number

Conditional FormattingFormat as TableCell Styles

C511/2/2004

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										

MLA

	date	Live Listings	New Paid	
4	Lun Nov 1	11/1/2004	127,362	10,422
5	Mar Nov 2	11/2/2004	129,566	8,645
6	Mié Nov 3		130,649	7,831
7	Jue Nov 4		131,570	8,225
8	Vie Nov 5		131,492	7,275
9	Sáb Nov 6		131.819	5.301

Figure B

mercadolibre_conditional_means_examples - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

Paste Font Alignment Number Styles

F8

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										

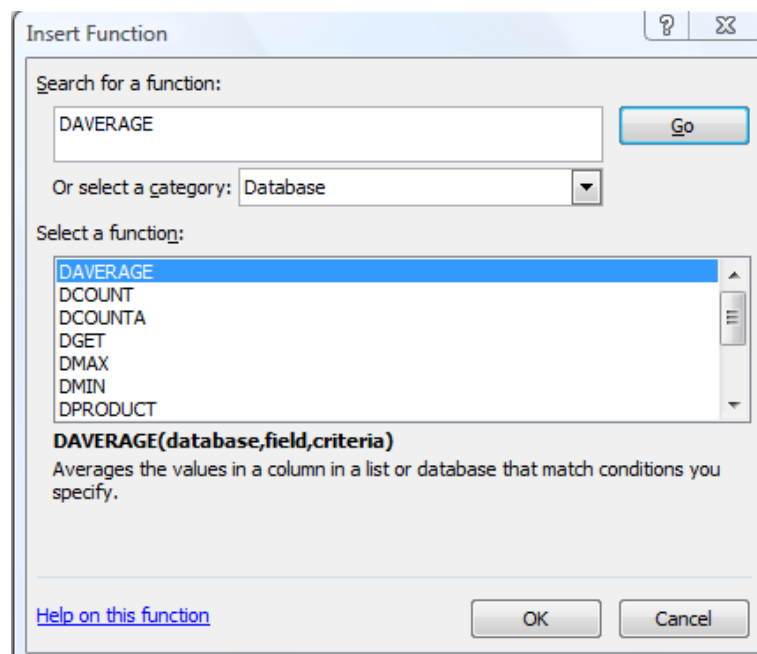
MLA

	date	Live Listings	New Paid	date	
4	Lun Nov 1	11/1/2004	127,362	10,422	<01/14/05
5	Mar Nov 2	11/2/2004	129,566	8,645	
6	Mié Nov 3	11/3/2004	130,649	7,831	
7	Jue Nov 4	11/4/2004	131,570	8,225	
8	Vie Nov 5	11/5/2004	131,492	7,275	
9	Sáb Nov 6	11/6/2004	131,819	5,301	

Step 3: Calculating the conditional mean

Place the cursor in an open cell below the condition and in the toolbar menu select the tab "Formulas." In the Formulas toolbar, click "Insert Function." In the pop-up menu that ensues we will write DАVERAGE in the field "Search for a function." In case you forget the exact name of the function, you can look for it in the set of all database functions by selecting Database in the drop-down menu of the field "Or select a category" (Figure C).

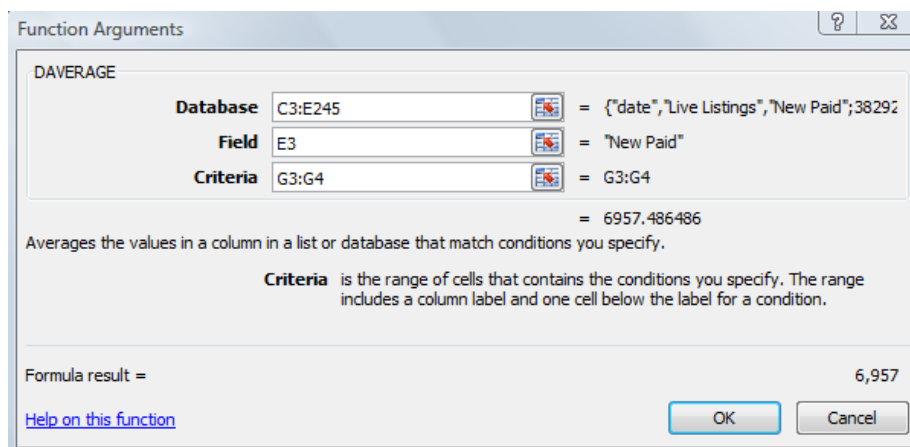
Figure C



In the pop-up menu that ensues we will input (see **Figure D**):

- Database: Select a range containing all the cells with relevant data AND the column headings. In this example C3:E245 (date; Live Listings; and New Paid columns).
- Field: Choose the column with the variable for which we want to calculate the mean. There are three ways to identify this column: (i) Input the cell with the column label (in this example E3). (ii) Write the label of the column in double quotation marks (in this example "New Paid"). (iii) Input the number of the column position in the database (in this example 3).
- Criteria: Select the range containing the condition. In our example G3:G4.

Figure D



You will obtain 6957.49 as the average new paid listings before January 14.

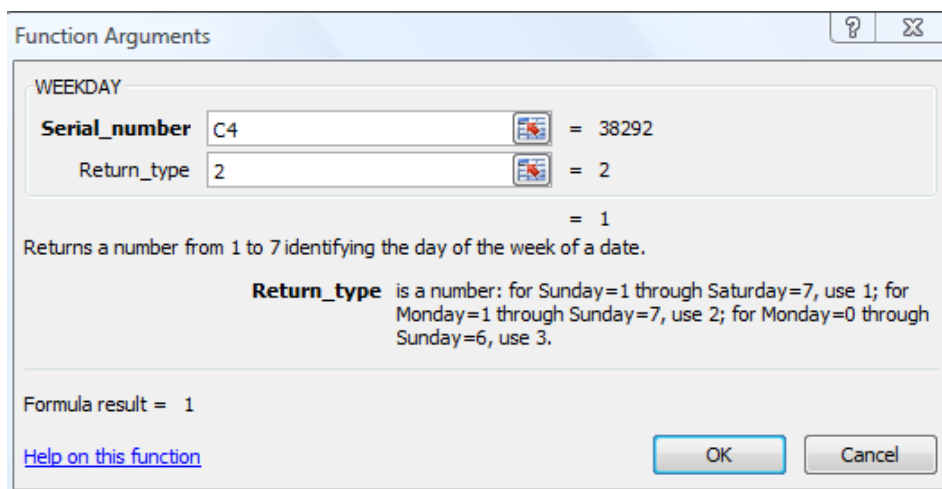
Average subject to multiple simultaneous conditions (AND conditions) Assume that we want to calculate the average number of new paid listings on the Saturdays after January 14 but before April 5—the two free listing days.

Step 1: Preparing the data

Prior to calculating the mean we have to create a column with the day of the week. I suggest:

1. Insert a new column between the date column we created before and the Live Listings column.
2. In the first cell of the column insert the function WEEKDAY(). In the toolbar menu, select the tab “Formulas,” click “Insert Function,” and then select WEEKDAY. You could also look for this function in the drop-down menu “Or select a category,” selecting the option “Date & Time.” Another option is to click on the “Date & Time” icon in the “Formulas” tab of the toolbar menu.
3. In the ensuing pop-up menu input (see **Figure E**):
 - a. Serial_number: Select the cell in the date column in the same row where we are inputting the function (in this example C4).
 - b. Return_type: Select the day you want to have as the first day of the week. I like to select 2 because I normally think Monday is day 1 of the week and Sunday is day 7 (see **Figure E** for other options).
4. You may need to change the number format of the cell to avoid getting the output in date format (1 will appear as 1/1/1900). To change the format, go to the “Home” tab in the toolbar menu and in the drop-down menu of the “Number” group of commands select “General.”
5. Copy the function to the whole column (you can do this by double-clicking in the lower-right corner).
6. Label the column “weekday.”

Figure E

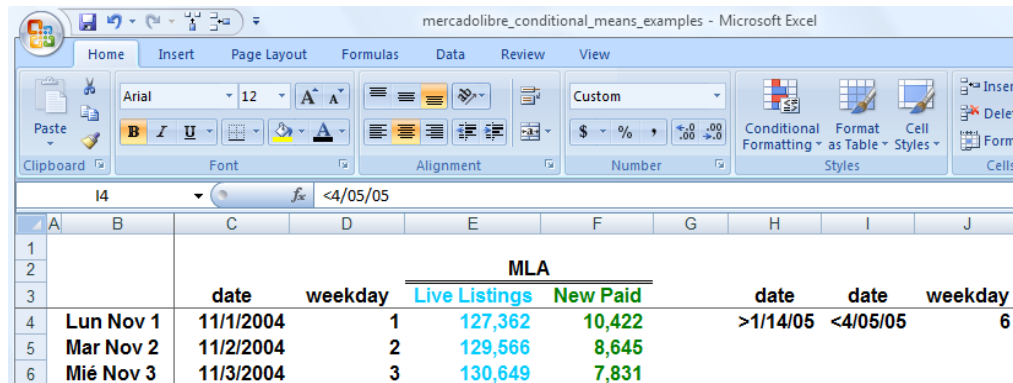


Step 2: Writing the condition

As before, to the right of the dataset we are going to write the condition. In this example we have three simultaneous conditions:

1. Listings after January 14
2. Listings before April 5
3. Listings on Saturday (day 6 of the week)

We will write the three conditions in 2 rows in 3 consecutive columns (cell H3 at the top left-hand corner). The first row will show the labels of the columns with the conditioning variables (note that the label “date” will appear twice) and the second row will show the specific conditions. See **Figure F**.

Figure F


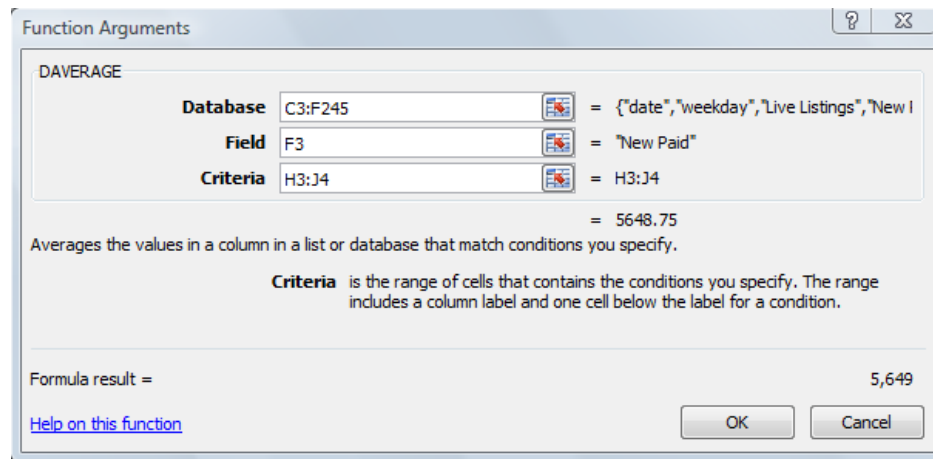
		MLA						
		date	weekday	Live Listings	New Paid	date	date	weekday
4	Lun Nov 1	11/1/2004	1	127,362	10,422	>1/14/05	<4/05/05	6
5	Mar Nov 2	11/2/2004	2	129,566	8,645			
6	Mié Nov 3	11/3/2004	3	130,649	7,831			

Step 3: Calculating the conditional mean

Place the cursor in an open cell below the condition. In the toolbar menu, select the “Formulas” tab, click “Insert Functions,” select the function DAVEVERAGE, and in the pop-up menu that ensues we will input (see **Figure G**):

- a. Database: Select a range containing all the cells with relevant data AND the column headings. In this example C3:F245.
- b. Field: Select the column with the variable for which we want to calculate the mean (in this example F3 or “New Paid”).
- c. Criteria: Select the range containing the condition. In our example H3:J4.

Figure G



You will obtain 5648.75 as the average new paid listings on Saturdays after January 14 but before April 5.

Average subject to multiple alternative conditions (OR conditions) Assume that we want to calculate the average number of new paid listings on the weekend days after January 14 but before April 5—the two free listing days. The alternative conditions are that the listings can be on a Saturday or a Sunday.

Step 1: Preparing the data

We use the data from the previous section.

Step 2: Writing the condition

We will write each of the alternative conditions in consecutive rows (one below the other). As before, each of the alternative conditions will be written in 2 rows in 3 consecutive columns. The first and third row will have the labels of the columns with the conditioning variables, and the second and fourth row will have the specifications of the two alternative conditions. See **Figure H**, because in this case an image is much better than a thousand confusing words.

Each alternative condition is the combination of three simultaneous conditions:

1. Alternative condition 1:
 - a. Listings after January 14
 - b. Listings before April 5
 - c. Listings on Saturday (day 6 of the week)
2. Alternative condition 2:
 - a. Listings after January 14
 - b. Listings before April 5
 - c. Listings on Sunday (day 7 of the week)

Figure H

		date	weekday	MLA		date	date	weekday
				Live Listings	New Paid			
1								
2								
3								
4	Lun Nov 1	11/1/2004	1	127,362	10,422	>1/14/05	<4/05/05	6
5	Mar Nov 2	11/2/2004	2	129,566	8,645	date	date	weekday
6	Mié Nov 3	11/3/2004	3	130,649	7,831	>1/14/05	<4/05/05	7
7	Jue Nov 4	11/4/2004	4	131,570	8,225			
8	Vie Nov 5	11/5/2004	5	131,492	7,275			

Step 3: Calculating the conditional mean

Place the cursor in an open cell below the condition. Select the “Formulas” tab in the toolbar menu, click “Insert Functions,” select the function DAVVERAGE, and in the pop-up menu that ensues we will input (see **Figure I**):

- Database: Select a range containing all the cells with relevant data AND the column headings. In this example C3:F245.
- Field: Select the column with the variable for which we want to calculate the mean (in this example F3).
- Criteria: Select the range containing the two alternative conditions. In our example H3:J6.

Figure I

Function Arguments

DAVERAGE

Database C3:F245 = {"date","weekday","Live Listings","New I

Field F3 = "New Paid"

Criteria H3:J6 = H3:J6

= 5348.333333

Averages the values in a column in a list or database that match conditions you specify.

Criteria is the range of cells that contains the conditions you specify. The range includes a column label and one cell below the label for a condition.

Formula result = 5,348

[Help on this function](#) OK Cancel

You will obtain 5348.33 as the average new paid listings on Saturdays or Sundays after January 14 but before April 5.

Slots, Tables, and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel

Introduction to Pivot Tables

Pivot Tables are a facility of Excel that allows the user to summarize and reorganize data (columns or rows) in a spreadsheet. Pivot Tables do not change the information in the original spreadsheet but refer to it. They are especially useful when manipulating large quantities of data.

Examples of situations where Pivot Tables are useful:

- If you have a dataset with information on the service contracts you have signed with your customers, you may want to summarize sales or profitability by customer, rank your customers by profitability, or summarize information by type of contract.
- If you have a dataset with detailed information on sales of different products in your stores, you may wish to summarize the information by the type of product or by store.

In this note we use relatively small (around 40,000 observations) datasets from the MGM Grand player database to understand the basic mechanics of Pivot Tables and appreciate some of their potential utility.

Basic Analysis of the MGM Databases

Understanding concentration of customers Let's start with an analysis of customer concentration. The first question we might be interested in is whether the 80/20 rule (80% of our profits are generated by 20% of our customers) applies to the customer portfolio of MGM.

In this analysis, we are going to focus on the MGM player database. Each observation in the player database represents a player trip. A player trip is the set of consecutive days a player is active in a certain property. For instance, if a customer goes to Las Vegas for three days and gambles at MGM and at Bellagio, the database will generate two observations: one at MGM, and one at Bellagio. If a customer goes to Las Vegas in January for three days, and in June for another three days, and only plays at MGM both times, it will also generate two observations.

Thus, some customers will appear only once in the database while others may appear multiple times, one for each trip they made to an MGM property in the period of the dataset. If we want to calculate the customer profitability concentration, we first need to summarize the total profitability of a customer.

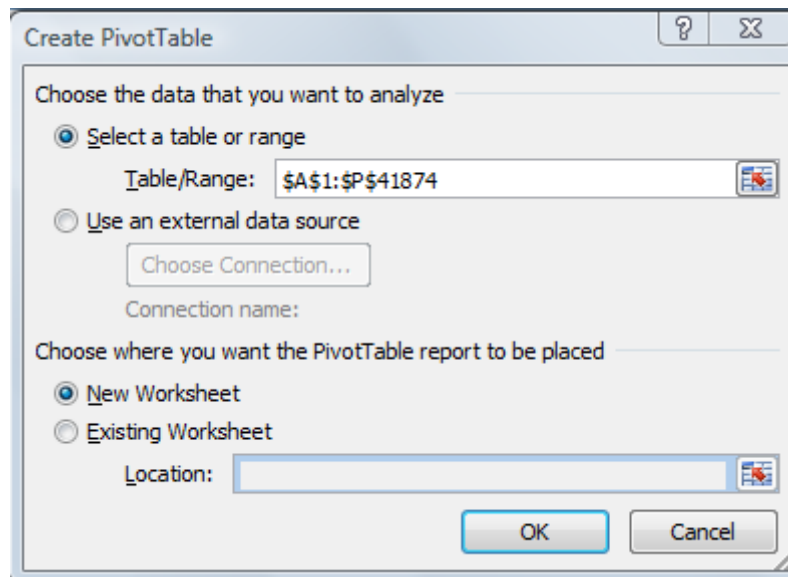
Summarizing customer profitability In this section we will calculate the total customer profitability over the three-year period and also the total customer profitability for each year. The end product will be another database (or spreadsheet) with one observation per customer.

To do so, we will follow these steps:

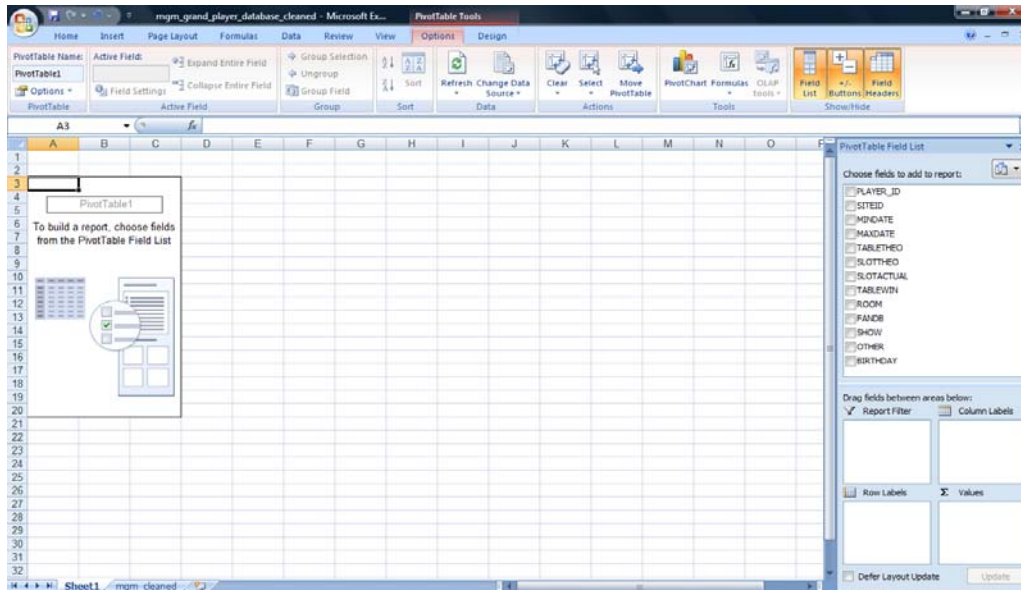
1. After opening the database in Excel, we add a column in the database for the total theoretical win (table + slot), naming it (for example), TOTALTHEO. Then:
2. In the toolbar, select the "Insert" tab.
3. In the "Insert" toolbar, select Pivot Table.

PLAYER_ID	SITEID	MINDATE	MAXDATE	TABLET	SLOT	ACTUAL	WIN	ROOM	FANDB	SHOW	OTHER	BIRTHDAY	YEAR
1	3	14-Jan-2004	16-Jan-2004	0	145.95	409.25	0	0	43.84	0	0	4-Dec-1950	2004
3	1	15-Jan-2004	16-Jan-2004	3.16	49.02	-309	40	0	4.33	0	0	4-Dec-1950	2004
4	1	25-Jan-2002	25-Jan-2002	0	14	65.5	0	0	0	0	0	4-Dec-1950	2002
5	2	5-Sep-2003	5-Sep-2003	0	0	0	0	0	0	0	0	24-Dec-1938	2003
6	2	5-Sep-2003	5-Sep-2003	0	87.4	179.75	0	0	0	0	0	24-Dec-1938	2003
7	2	26-Sep-2002	29-Sep-2002	0	17.02	-511.5	0	0	0	0	0	24-Dec-1938	2002
8	2	29-Sep-2002	29-Sep-2002	0	8.81	15	0	0	0	0	0	24-Dec-1938	2002
9	3	4-Sep-2003	1-Sep-2003	0	14.04	30	0	0	0	0	0	22-Feb-1970	2003
10	3	1-Sep-2003	1-Sep-2003	0	0	0	0	0	0	0	0	22-Feb-1970	2003
11	3	16-Nov-2002	17-Nov-2002	0	149.39	422.1	0	0	0	0	0	22-Feb-1970	2002
12	4	2-Dec-2002	2-Dec-2002	0	0	0	0	0	0	0	0	3-Mar-1943	2002
13	4	2-Dec-2002	2-Dec-2002	0	0	0	0	0	0	0	0	3-Mar-1943	2002
14	4	25-Oct-2002	25-Oct-2002	80.8	0	0	-3000	0	0	0	0	3-Mar-1943	2002
15	5	1-Jul-2002	27-Jul-2002	0	31.71	113.75	0	0	0	0	0	31-May-1936	2002
16	5	27-Jul-2002	27-Jul-2002	0	0	0	0	0	0	0	0	31-May-1936	2002
17	6	9-Nov-2002	9-Nov-2002	0	25.29	70	0	0	0	0	0	14-Dec-1960	2002
18	6	9-Nov-2002	9-Nov-2002	12.2	112.94	175	400	0	0	0	0	14-Dec-1960	2002
19	7	26-Feb-2002	26-Feb-2002	0	2.22	23.5	0	0	0	0	0	15-Aug-2028	2002
20	8	11-Apr-2003	14-Apr-2003	268.94	89.36	247.5	-1400	165.48	15	0	0	14-Mar-1949	2003
21	8	11-Dec-2004	15-Dec-2004	0	271.48	161.5	0	50.1	0	0	0	14-Mar-1949	2004
22	8	12-Apr-2003	14-Apr-2003	0	4.4	28.75	0	0.88	0	0	0	14-Mar-1949	2003
23	8	13-Dec-2003	15-Dec-2003	60.33	116.39	-194.5	400	0	34.28	0	0	14-Mar-1949	2003
24	8	13-Dec-2004	14-Dec-2004	0	29.98	71.25	0	0	0	0	0	14-Mar-1949	2004
25	8	14-Oct-2003	18-Oct-2003	94.33	137	431.25	300	76.92	44.71	0	0	14-Mar-1949	2003
26	8	15-Oct-2003	18-Oct-2003	0	137.84	6.75	0	27.56	0	0	0	14-Mar-1949	2003
27	8	16-Feb-2002	16-Feb-2002	0	64.32	87	0	0	0	0	0	14-Mar-1949	2002
28	8	17-Oct-2003	18-Oct-2003	0	25.16	119.75	0	44.7	0	0	0	14-Mar-1949	2003
29	8	17-Oct-2003	18-Oct-2003	0	7.94	46.5	0	2.77	0	0	0	14-Mar-1949	2003
30	8	20-Aug-2002	23-Aug-2002	0	25.83	67	0	5.16	0	0	0	14-Mar-1949	2002
31	8	20-Aug-2002	23-Aug-2002	29.63	19.38	137	300	148.94	32.66	0	0	14-Mar-1949	2002
32	8	27-Jun-2004	30-Jun-2004	42.98	188.94	812.25	-200	32.08	0	0	0	14-Mar-1949	2004

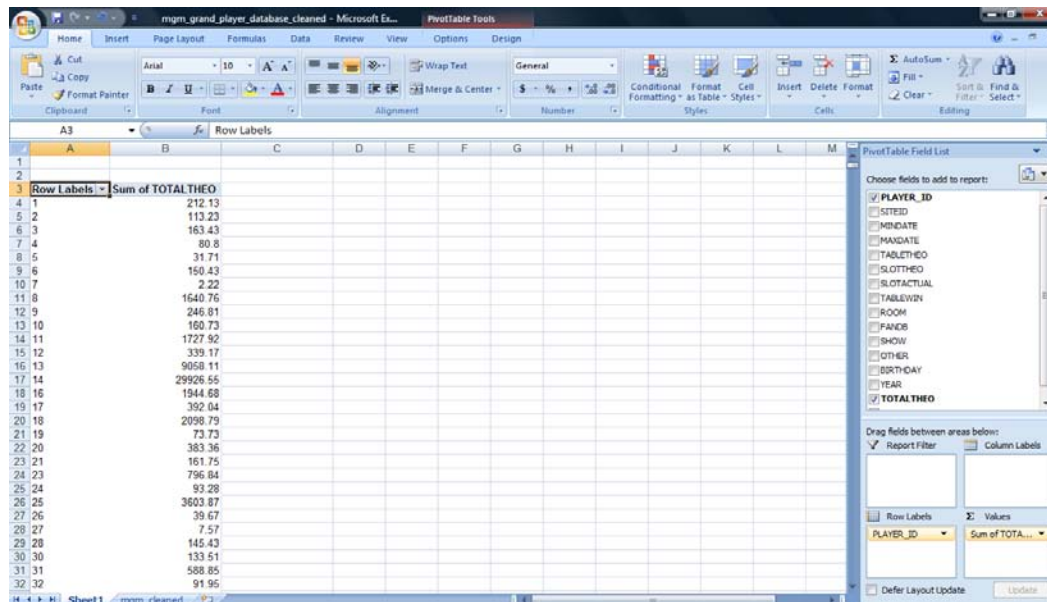
4. In the “Create Pivot Table” pop-up menu that appears, select the whole database as your relevant range, choose a new worksheet as the location of the Pivot Table, and click OK.



5. In the new spreadsheet you will see three relevant areas:
 - a. On the left, there is an area that will become the Pivot Table.
 - b. On the right, there is a list of all the variables/fields to use in the Pivot Table. Below this list you will see the Fields Layout Menu, which you can refer to for instructions on how to add fields to the table.
 - c. At the top, there is the Pivot Table toolbar.



Click PLAYER_ID in the Pivot Table Field List and drag it to the Row Labels area in the Fields Layout Menu or the “Drop Row Fields Here” area of the Pivot Table. Then click on TOTALTHEO and drag it to the Values area in the Fields Layout Menu or the “Drop Data Items Here” area of the Pivot Table. Excel will then produce the Pivot Table.

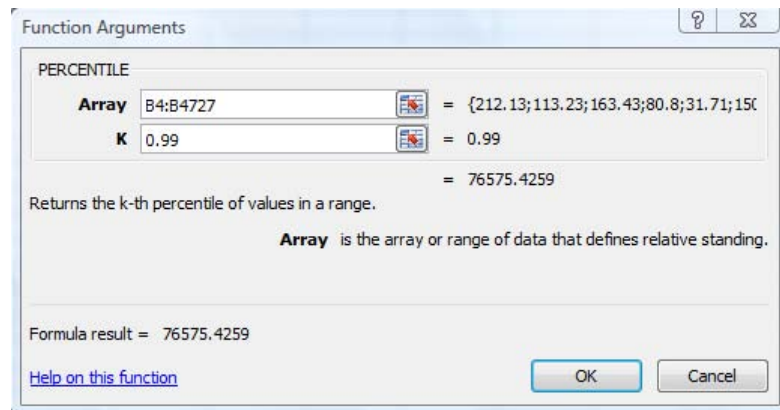


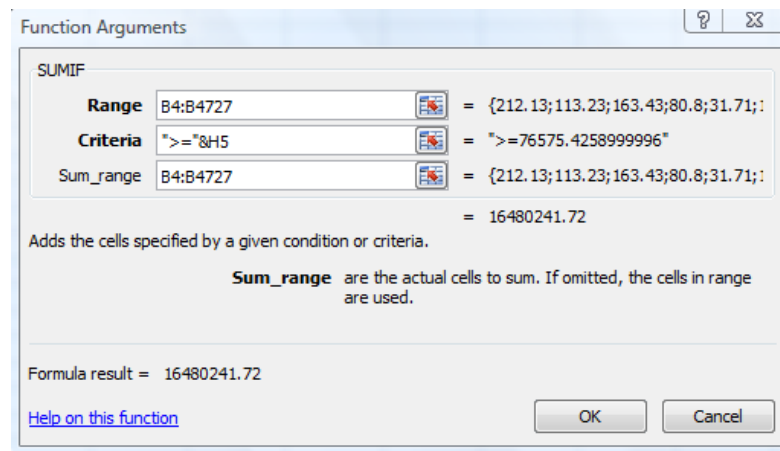
6. To measure concentration, we can set different cut-off points. For the purposes of this tutorial we will select three common types of cut-off points:
 - Based on percentile of customer profitability: e.g., the top 1% most profitable customers
 - Based on a rank: e.g., the 100 most profitable customers

- Based on a profitability number that your experience indicates is interesting: e.g., customers with profitability above \$100k

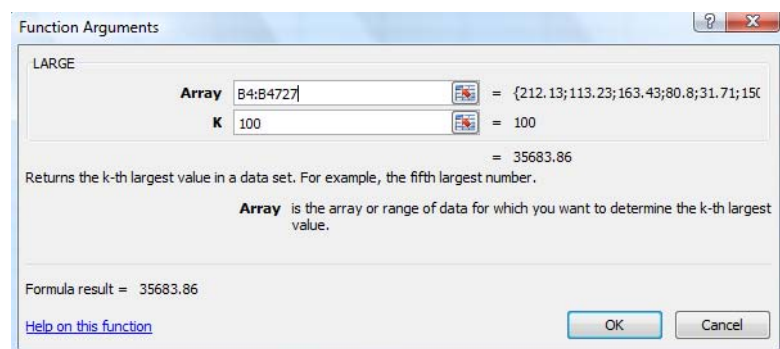
We will now make these calculations one by one:

7. For the **percentile approach** we follow two steps:
 - a. Calculate, for example, the top 1% percentile.
 - i. Click on any cell of the worksheet outside of the Pivot Table (say cell H5).
 - ii. In the toolbar menu select Formulas, Insert Function, Percentile (you can type percentile in the “Search for a function” box).
 - iii. In the Function Arguments select:
 1. Array: An area within the Pivot Table cells that contain the TOTALTHEO values per customer.
 2. K: 0.99, as the percentile is calculated increasing in value.





- c. Divide the result of this cell by the total theoretical win (at the bottom of the Pivot Table) to get the percentage of revenues represented by these customers (63%). Note that if you just point to the cell with the total theoretical win the resulting formula is: =H6/GETPIVOTDATA("TOTALTHEO",\$A\$3). The problem with this formula is that it anchors the cell with the total TOTALTHEO, something you need to keep in mind if you are copying the formula. Alternatively, you can type the actual coordinates of the cell, in this case B4728.
8. For a **rank criterion**, the procedure is the same as for percentile except for the first step.
- Calculate the revenue of the 100th largest revenue customer.
 - Click on any cell in the worksheet outside of the Pivot Table (for instance, I5).
 - In the toolbar menu select Formulas, Insert Function, LARGE (you can type RANK or LARGE in the "Search for a function" box).
 - In the Function Arguments select:
 - Array: An area within the Pivot Table cells that contain the TOTALTHEO values per customer
 - K: 100

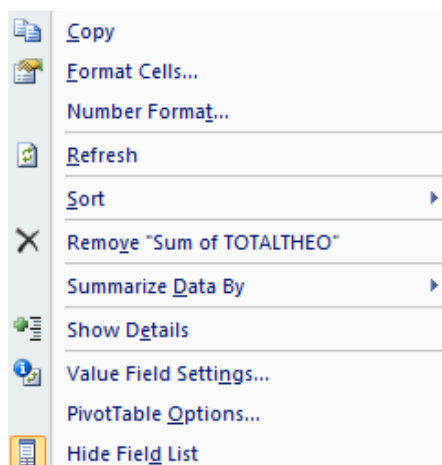


- Proceed as in 7.b and 7.c and you should get 72%.
9. For the **customers with profitability above a certain benchmark**, say \$100K, just follow 7.b (using as Criteria ">=100,000") and 7.c. You should get 60%.

Performing the year-by-year analysis:

As a prerequisite for this analysis we need to create a field in the database that contains the year of the trip. To do this:

1. Insert a column anywhere inside the range of existing columns of the original dataset (now including the TOTALTHEO column). This is important because if we insert the column outside the range of the previous columns, say in column R, then we will need to define a new Pivot Table. If the new column is inserted within the set of previous columns, we can simply follow the procedure below. In this example you can place the cursor in a cell of column M (BIRTHDAY) and insert a column to the left.
2. Write YEAR for the heading of the inserted column (new column M), and in any cell of that column (say M2) insert the function =YEAR(C2), in which C2 is the corresponding cell in the MINDATE column. Copy the function through the bottom of the column.
3. Now the field YEAR exists in the main database but is not yet recognized by the Pivot Table tool. In the Pivot Table Field List there is no YEAR variable (if you do not see the Pivot Table Field List just click on any cell in the Pivot Table). To make the new field “visible” to the Pivot Table tool, place the cursor on any cell inside the Pivot Table and right-click. In the ensuing menu choose Refresh; now the YEAR field is recognized by the Pivot Table tool.



Note that it is possible to refresh the data in the Pivot Table by simply clicking on the Refresh Icon in the Pivot Table Options toolbar:



4. Now click on YEAR in the Pivot Table Field List and drag it to the Column Labels area on the Fields Layout Menu. You will now get the Pivot Table:

Row Labels	2002	2003	2004 Grand Total
1	14		212.13
2	25.83	87.4	113.23
3	149.39	14.04	163.43
4	90.8		90.8
5	31.71		31.71
6	150.43		150.43
7	2.22		2.22
8	139.16	941.69	559.91
9	50.32	54.88	141.61
10	75.23	48.53	36.97
11	316.61	750.72	660.59
12	63.12	191.03	85.02
13	4977.65	2153.42	1927.64
14	16142.02	13411.69	372.94
15	550.98	706.58	687.12
16	154.19	110.3	127.55
17	722.28	753.29	623.22
18	1.52		72.21
19	10.29	98.28	274.79
20	79.24	71.42	11.09
21	796.84		796.84
22	24.8	68.48	93.28
23	224.17	1296	2083.7
24	12.57	27.1	0
25	7.57		7.57
26	145.43		145.43
27	7.42	38.19	87.9
28	356.48	232.37	588.85

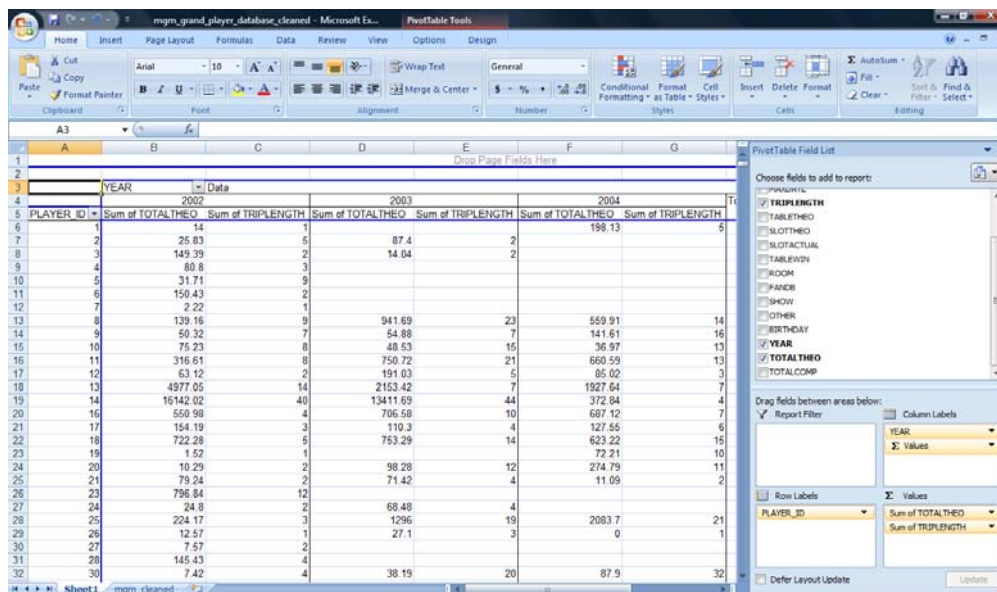
5. The analyses will proceed as before from point 6 on. As you are performing them, keep these two points in mind:
 - a. If a customer has not visited a property in a given year—2004, for example—she will not have an observation (as opposed to having a zero).
 - b. If you want to perform the analysis for 2002 and then copy the formulas laterally, be careful to anchor the appropriate cells. Most likely you will need to manually enter some cell coordinates rather than pointing to the cell in the Pivot Table.

Calculating the contribution to total profits of groups of customers segmented according to certain characteristics

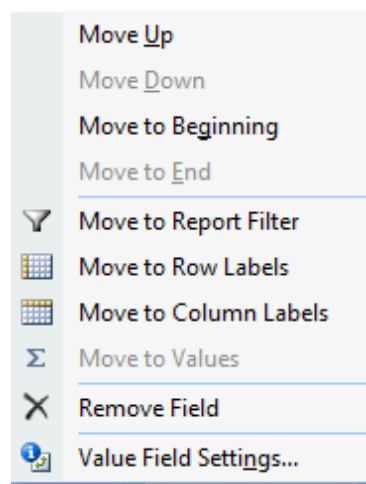
Let's assume you want to calculate the percentage of total theoretical win generated by the customers with an average trip length of three days or more. The procedure will be very similar to the previous sections except that the conditioning variable (number of trips) and the variable you want to aggregate (total theoretical win) are not the same.

1. First, you need to create the variable **TRIPLENGTH**. Insert a column within the previous columns, for instance next to the **MAXDATE** column. Name the column **TRIPLENGTH** and define it as:

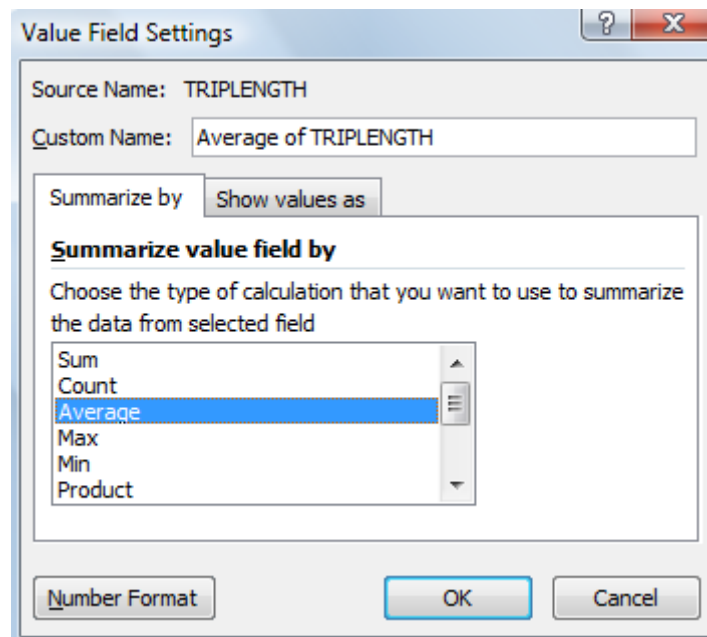
$$\text{MAXDATE} - \text{MINDATE} + 1.$$
 You may need to change the number format of the cells in this column to "general," as the date format will cause each number to appear as a date.
2. Refresh the Pivot Table.
3. Click on the **TRIPLENGTH** variable in the Pivot Table Field List and drag it to the Values area on the Fields Layout Menu.



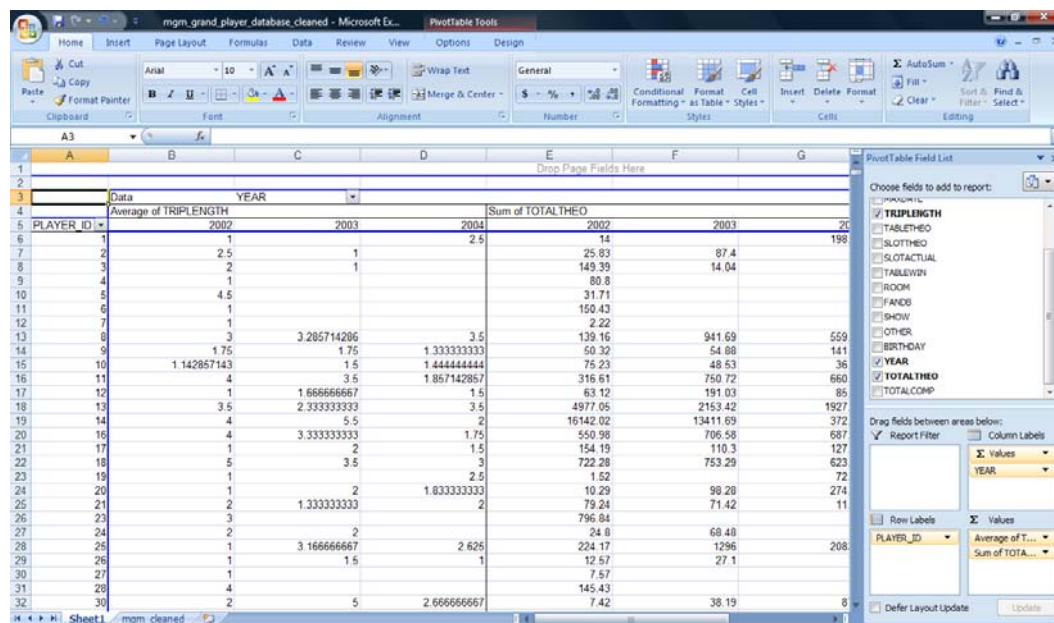
4. To find the percent contribution to TOTALTHEO of those customers with an average trip length greater than three days applying the SUMIF function, two prior operations are necessary. First, you must use the average trip length rather than the sum of trip length.
5. In the Fields Layout Menu, click on the drop-down arrow on the Sum of TRIPLENGTH bar located in the Σ Values field. Select "Value Field Settings" in the ensuing drop-down menu.



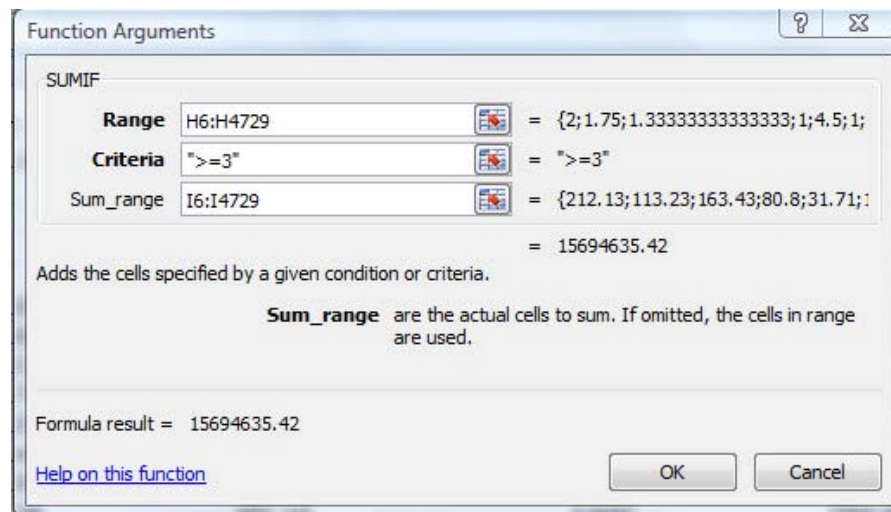
6. In the Value Field Settings pop-up menu that appears, choose to summarize value field by Average instead of by Sum. Press OK to Continue.



7. If you prefer to have all the TRIPLNGTH columns next to each other, make sure that the “ Σ Values” field is above the YEAR field in the Column Labels area of the Fields Layout Menu. If you want to have each year’s TRIPLNGTH column next to the TOTALTHEO column of that year, make sure that the YEAR field is above the “ Σ Values” field in the Column Labels area of the Fields Layout Menu.



8. To calculate the total theoretical win generated by players with average trip length greater than or equal to three days, you can use the SUMIF function as before. For the purpose of this note we will do this for the total columns. Place the cursor to the right of the Pivot Table and in the toolbar menu select Formulas/Insert Function/SUMIF (you can type sumif in the "Search for a function" box). In the Function Arguments select:
 - a. Range: The range of the cells you want to evaluate, in this case the cells that contain the average TRIPLENGTH per customer (in column H).
 - b. Criteria: type ">=3".
 - c. Sum_range: The numbers to add. In this case the cells that contain the TOTALTHEO values per customer (in column I).



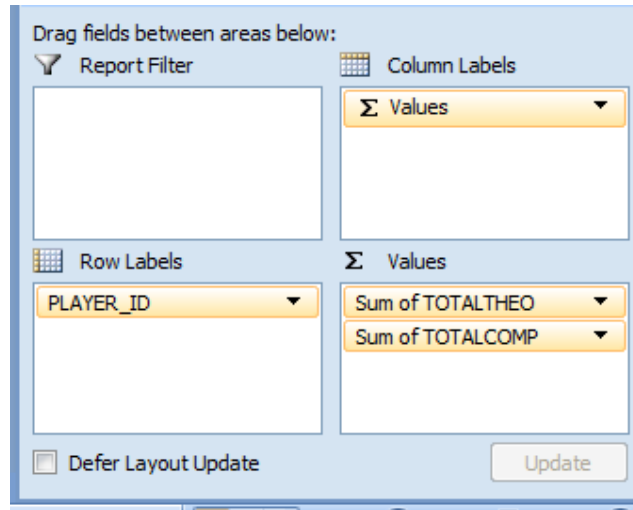
You will obtain \$15,694,635 or 60% of the total theoretical win of the properties in the database.

Measuring other variables for groups of attractive customers

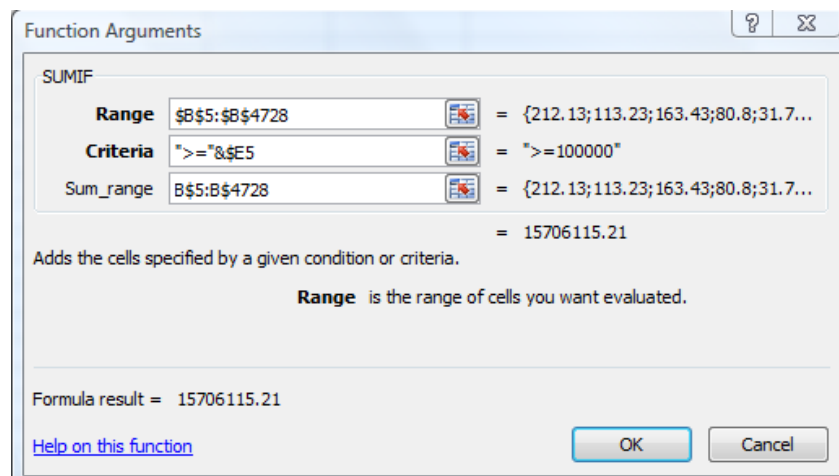
In this section we seek to analyze whether groups of customers differ along certain dimensions.

Specifically, we will calculate the average level of comps for three groups of customers defined as a function of theoretical win (>100k; 20k–100k; 0–20k). We will calculate it for the full period.

1. First, create the variable TOTALCOMP. Insert a column within the previous columns, for instance next to the OTHER column. Name the column TOTALCOMP and define it as the sum of all the comps (ROOM, FANDB, SHOW, OTHER) by observation. In the cell N2 (the second cell of the new column) you will enter: =SUM(J2:M2) (where columns J to M contain the different types of comps) and then sum downwards.
2. Refresh the Pivot Table.
3. Click on the TOTALCOMP variable in the Pivot Table Field List and drag it to the "Σ Values" area. To simplify the display, keep only the total comps and total theoretical win for the whole period. Remove the average trip length and year columns from the Pivot Table by dragging their bars outside of the Field Layout Menu area.



4. To calculate the total theoretical win generated by each of the segments you will follow the procedure in the first section of this note using the SUMIF function. To reduce the number of calculations, use this trick:
 - a. First, calculate the sum of the theoretical win for gamers with TOTALTHEO over \$100K, \$20K, and \$0. To do this, write the three numbers (100,000; 20,000; and 0) in three consecutive cells in the same column (for instance E5, E6, and E7). Then in the cell next to 100,000 (F5), insert the SUMIF function:
 - i. Range: The range of the cells you want to evaluate, in this case the cells that contain TOTALTHEO per customer (column B). Anchor this range to copy downwards.
 - ii. Criteria: type ">="&\$E5.
 - iii. Sum_range: The numbers to add. Again, in this case anchor the range TOTALTHEO.



- iv. Copy the formula downwards.
- b. Now calculate the sum of the TOTALCOMP for the same group of customers. Proceed as in the last step (a), changing Sum_range to \$C\$5:\$C\$4728.

Alternatively, the formulas in column F can be copied to column G if the Sum_range and the Criteria are anchored on the rows as shown in the example.

- c. In some other area of the spreadsheet (for instance three lines below the calculations of points (a) and (b)), calculate the TOTALTHEO and TOTALCOMP of each of the segments of interest by subtraction (for instance, for the TOTALCOMP for customers with a total theoretical win between 20K and 100K, this will be TOTALCOMP for customers with TOTALTHEO over \$20K minus the TOTALCOMP for customers with TOTALTHEO over \$100K; thus, subtract cell G5 from cell G6).
- d. Calculate the percentage dividing TOTALCOMP by TOTALTHEO. Your results should look something like this:

Criteria	TOTALTHEO	TOTALCOMP
100000	15,706,115	2,016,991
20000	20,677,805	3,304,578
0	26,196,214	4,894,484

Criteria	TOTALTHEO	TOTALCOMP	COMPPCTG
>100K	15,706,115	2,016,991	13%
>20K & <100K	4,971,690	1,287,587	26%
>0 & <20K	5,518,408	1,589,906	29%

Slots, Tables and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel

Merging Datasets in Excel (the Function VLOOKUP)

This section describes how to merge two datasets in Excel using the function VLOOKUP. Note that this function merges the datasets using only one field as the matching criterion. If you want to merge datasets using two fields as the matching criteria—for instance customer ID and the date of the trip—you would first need to create a new field that is the combination of those two.

The objective of this exercise is to analyze the overlap of customers between the MGM Players' database and its Hotel database. You will merge using the customer ID as the matching criterion.

The VLOOKUP function looks for an observation (row) with a specific condition in one dataset, and imports the value of a variable of interest from this dataset to another. For clarity, this note refers to the dataset receiving the information as the base dataset and to the dataset in which the function looks for a value as the supporting dataset. Note that while the matching variable (the specific condition) may show identical values more than once in the base dataset, it must appear only once in the supporting dataset. The VLOOKUP function also requires the matching variable to be in the leftmost column of the supporting dataset, which must be sorted in ascending order of the matching variable.

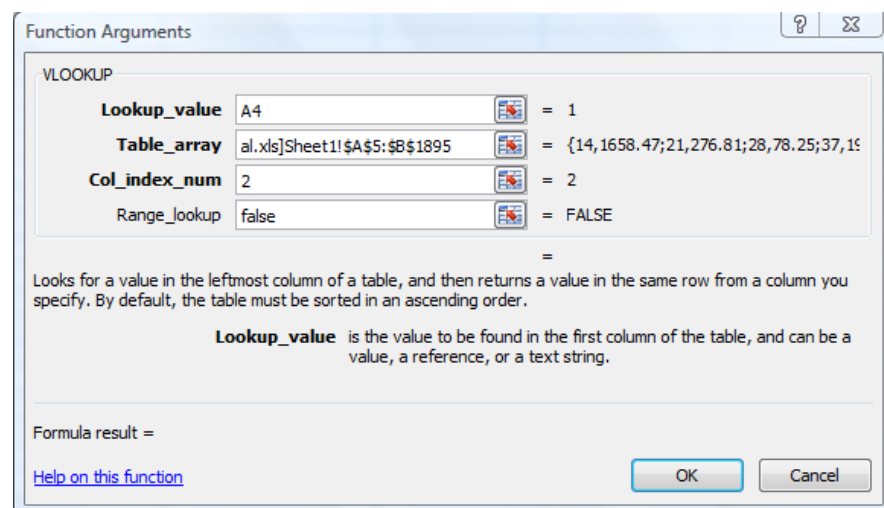
This note describes how to merge a table with the total theoretical win per customer and a table with the total hotel revenue per customer.

1. Create a Pivot Table with total theoretical win per customer (see the note on Pivot Tables for a more detailed description of this step):
 - a. Create the TOTALTHEO variable summing TABLETHEO and SLOTTHEO.
 - b. In the toolbar, select the "Insert" tab.
 - c. In the "Insert" toolbar, select Pivot Table.
 - d. In the Create Pivot Table pop-up menu that ensues, select the whole database as your relevant range, choose a new worksheet as the location of the Pivot Table, and click OK.
 - e. Click PLAYER_ID in the Pivot Table Field List and drag it to the Row Fields area of the Fields Layout Menu. Then click on TOTALTHEO and drag it to the Σ Values area.
2. Create a Pivot Table with total hotel revenue per player (see the note on Pivot Tables for a more detailed description of this step):
 - a. Open the MGM Hotel Dataset in Excel and select Pivot Table in the Insert Tab.
 - b. In the Create Pivot Table pop-up menu that ensues, select the whole database as your relevant range, choose a new worksheet as the location of the Pivot Table, and click OK.
 - c. Click PLAYER_ID in the Pivot Table Field List and drag it to the Row Fields area of the Fields Layout Menu. Then click on TOTAL_CHARGES and drag it to the Σ Values area.

Note that there will be a row in the Pivot Table with a player ID "(blank)" with total charges of 2,087,097. This row contains the charges to all hotel customers for whom

MGM did not capture the player ID, either because they do not have a Players' Club card or because it was not captured in the reservation or the check-in process.

3. Go to the Pivot Table of the MGM player dataset. Place the cursor in the same row as player ID 1 but outside of the Pivot Table area next to the player's total theoretical win (cell C5). In the toolbar menu select Formulas, Insert Function, VLOOKUP.
 - a. In the Function Arguments select:
 - i. **Lookup_value**: The value to be found in the hotel dataset, or what we call the matching variable. Choose the cell with the player ID (A5).
 - ii. **Table_array**: The supporting dataset, in this case the hotel dataset. Remember that the first column of the supporting dataset needs to be the matching variable. This will not be a problem in our example because we have already created both Pivot Tables with player ID as the row variable. VLOOKUP anchors the table array by default when it is used with different Excel files.
 - iii. **Col_index_num**: The order of the column with the data of interest in the supporting dataset. Column 1 is the column with the matching variable. In this case the variable of interest, total charges, is in column 2.
 - iv. **Range_lookup**: I have always used FALSE because this tells Excel to look for a perfect match of the variable in the base and the supporting datasets.



4. Copy the function downwards.
5. To manipulate this data you need to create a new Excel worksheet, either in the player dataset or in a new file, and copy and paste the entire Pivot Table. To save memory you may want to copy only the values, not the formulas (Home, Paste, Paste Values).

Now you can proceed to perform the overlap analyses.

Slots, Tables and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel

Introduction to Basic Regression Analysis Using Excel

The objective of this note is to introduce the mechanics of regression analysis using Excel. It should not be considered a substitute for a more rigorous reading on the theory and foundations of regression analysis.

Excel is not the most efficient program for regression analysis. However, many of you will have to use Excel during your careers, which is why we have chosen it as the analytical tool in this tutorial. In addition, the economic interpretation section of the tutorial applies to any analytical tool you may use.

The examples in this note use the player dataset that accompanies the case “Slots, Tables, and All That Jazz: Managing Customer Profitability at the MGM Grand Hotel.”

Using Regression Analysis to Understand Current Customer Profitability

Regression analysis studies the relationship between the dependent variable and a set of independent variables or regressors. It aims to find the function that best describes the conditional mean of the dependent variable given certain values of the independent variables. That is, for each set of values of the regressors, the analysis finds the most likely value of the dependent variable.

Let’s assume that we have the hypothesis that this year revenues for a customer are a function of his/her revenues two years ago; the percentage of comps he/she received from MGM two years ago; the percentage of comps he/she received this year; and the number of trips the customer made to Las Vegas two years ago.

For the purpose of this analysis, we are going to limit our analysis to customers who had some level of play in 2004. Note that limiting the analysis to this set of observations introduces a significant bias in the results, as the value of the dependent variable will be conditioned not only to the value of the regressors but also to the fact of the customer’s return. There are different techniques to reduce the impact of this bias—some of them led to James Heckman’s Nobel Prize in economics—but they are beyond the scope of this note.

Essentially, we are saying that 2004 theoretical revenues can be described by this equation:

$$THEO_{2004} = \alpha + \beta THEO_{2002} + \gamma COMPPCTG_{2002} + \lambda COMPPCTG_{2004} + \phi TRIPS_{2002} + \xi$$

Where

THEO = Total theoretical win

COMPPCTG = Total comps/total theoretical

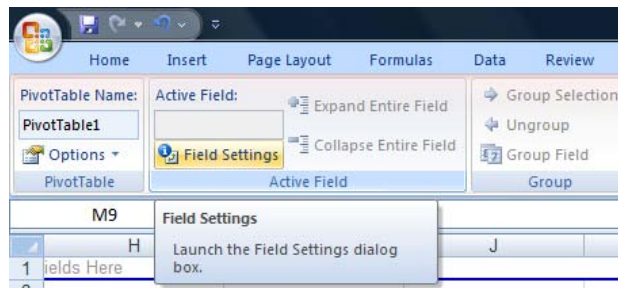
TRIPS = Number of trips in a year

ξ = Random variation

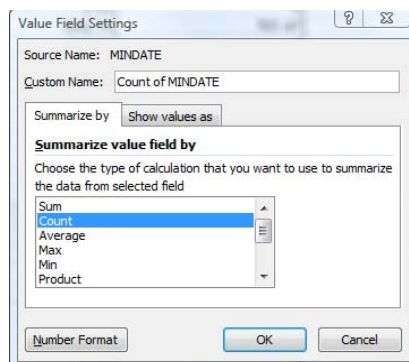
We want to estimate β , γ , λ , and ϕ —coefficients of the regression that express the relationship between the dependent variable $THEO_{2004}$ and the independent or explanatory variables.

Step 1: Getting the Necessary Data⁶

1. In the spreadsheet “mgm_grand_player_database_cleaned.xls,” we create three columns to calculate the following variables:
 - a. $\text{TOTALTHEO} = \text{SLOTTHEO} + \text{TABLETHEO}$, or the total theoretical win MGM gets from a customer in a given trip.
 - b. $\text{TOTALCOMP} = \text{ROOM} + \text{FANDB} + \text{SHOW} + \text{OTHER}$, or the total comps MGM gives to a customer in a given trip.
 - c. YEAR, or the year in which the trip takes place.
2. Create a Pivot Table:
 - a. Drop the variable PLAYER_ID in the Row Labels field.
 - b. Drop the variable YEAR in the Column Labels field.
 - c. Drop the variables TOTALTHEO, TOTALCOMP, and MINDATE in the Σ Values field.
 - d. We will use the variable MINDATE to count the number of trips. To do this, we must choose a variable without any blank observations. To obtain the number of trips, click on any cell in the MINDATE column, then click the Field Settings icon in the Active Field section of the PivotTable Tools Options toolbar (the second icon from the right).

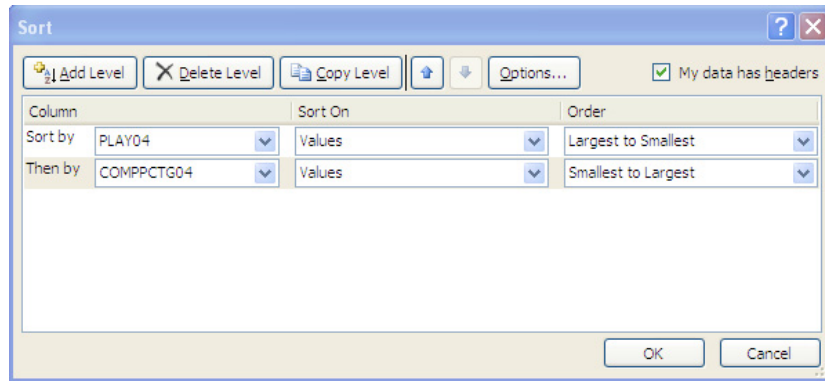


In the pop-up window that appears, we select Count in the “Summarize by” drop-down menu. In the MINDATE column, we now see the number of trips a customer makes to Las Vegas in a year. You can arrive at this pop-up window by clicking on the drop-down arrow on the Sum of MINDATE bar located in the Σ Values field and selecting “Value Field Settings” in the ensuing drop-down menu.



⁶ For a more detailed explanation on how to construct this dataset, please refer to the note Introduction to Pivot Tables.

3. To facilitate the manipulation of the data, copy the Pivot Table and paste it in another worksheet (Home, Paste, Paste Values). Rename the columns so that they have one-cell titles of the variables they contain (for instance, THEO02, THEO03... COMP03... TRIP04). Make sure that your new worksheet does not include the Grand Total row.
4. In the new worksheet create four more columns:
 - a. Create three COMPPCTGXX columns that divide COMPPXX by THEOXX. Alternatively, you could calculate the per-trip COMPPCTG in the original dataset and then calculate the average comp percentage in the Pivot Table. One drawback of this approach is that if an entry shows a non-zero value for comp and zero value for theoretical win—something which may happen if the entry is an adjustment of another entry—then the Pivot Table will generate a non-valid value for that player ID. Also, it is important to keep in mind that the percentage of the averages is different than the average of the percentages, and the choice between both variable definitions should depend on the relationships between comping and gaming behavior you have in mind when performing the analysis.
 - b. Create a PLAY04 column that takes the value of 1 if the customer plays at any MGM property in 2004 and the value of 0 otherwise. We can do this by using the function =IF(TRIP04>0,1,0).
5. To perform a regression analysis in Excel, the fields containing the explanatory variables must be contiguous. Thus, we:
 - a. Create a duplicate of the spreadsheet with the regression data (in case we would like to run a regression with a different specification). To do this, we right-click on the sheet's tab at the bottom of the window, select "Move or Copy..." and tick the "Create a copy" box.
 - b. In the duplicate sheet, delete the THEO03, COMPPCTG03, and all COMP02-COMP04 columns (or relocate them if you prefer). Do not forget to transform the COMPPCTG02 and COMPPCTG04 columns into numbers before deleting the COMPPXX columns. Otherwise, a reference error will appear in the COMPPCTG fields (#REF).
 - c. Move the THEO04 column to the end.
 - d. Delete any remaining blank columns.
6. As we are going to run the regression only for the players who actually played in 2004, before running the regression it is interesting to sort the data by the field PLAY04 (in the toolbar menu select the Data tab, then click Sort). In the pop-up menu that appears:
 - a. Check the box, "My data has headers."
 - b. The first level will be: Sort by PLAY04, Sort on Values, and Order Largest to Smallest, because we want to run the regressions for those players who generated some revenue for MGM in 2004.
 - c. We would also want to use COMPPCTG04 as a secondary criterion to sort this data. To do so, click Add level. To fill the fields of the next level select: Then by COMPPCTG04, Sort on Values, and Order Smallest to Largest. We do this to eliminate the observations in THEO04 that have a value of 0 (they generate a #DIV/0! error in the COMPPCTG variable). The regression function will produce an error message if we do not eliminate these observations.

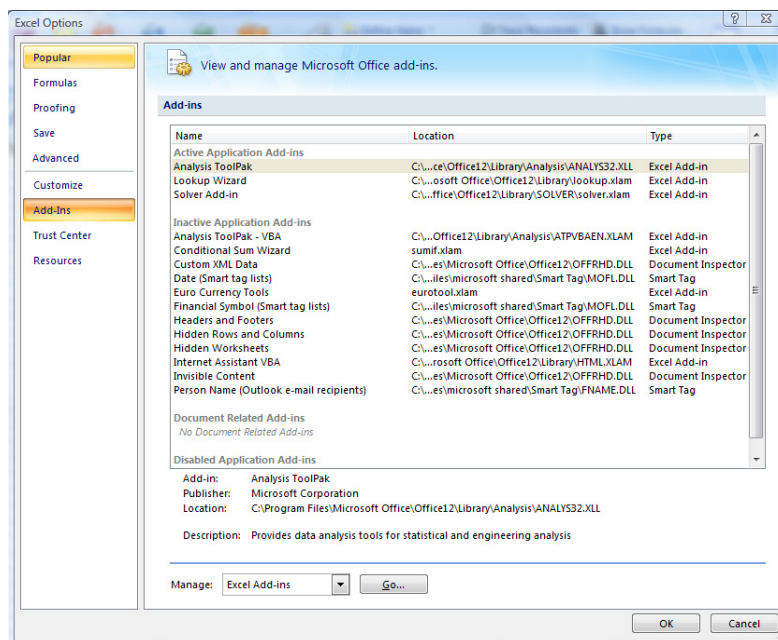


- d. These steps will not eliminate all the observations with a #DIV/0! error in the field COMPPCTG02. To completely eliminate the error message, we select an area that includes only the observations with "1" in the field PLAY04 and a valid numeric value in COMPPCTG04 (you should have 2058 rows, including the titles row). We then sort these observations by COMPPCTG02 in ascending order.
- e. To run the regressions we can either delete all the rows with invalid data or, when selecting the input area, include only the rows with valid data (rows 1 to 2041).

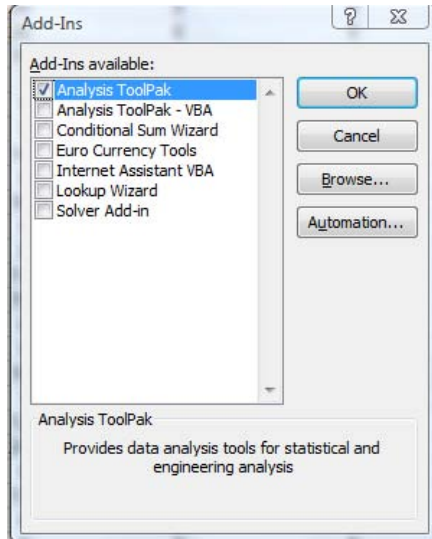
Step 2a: Run the Regression: The Analysis ToolPak

There are two ways to run a regression with Excel: we can use the Add-In Analysis ToolPak or the function LINEST(). We will first discuss the Analysis ToolPak because it produces a more intuitive output.

1. Install the Analysis ToolPak: In the toolbar menu, Click the Microsoft Office button in the top right corner and choose Excel Options, Add-Ins.

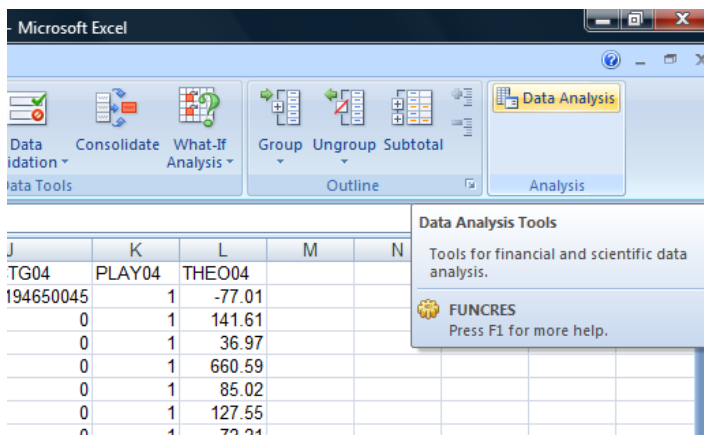


In the Add-Ins Menu, click the Go button next to Manage: Excel Add-Ins. Then in the pop-up window that appears we check Analysis ToolPak.

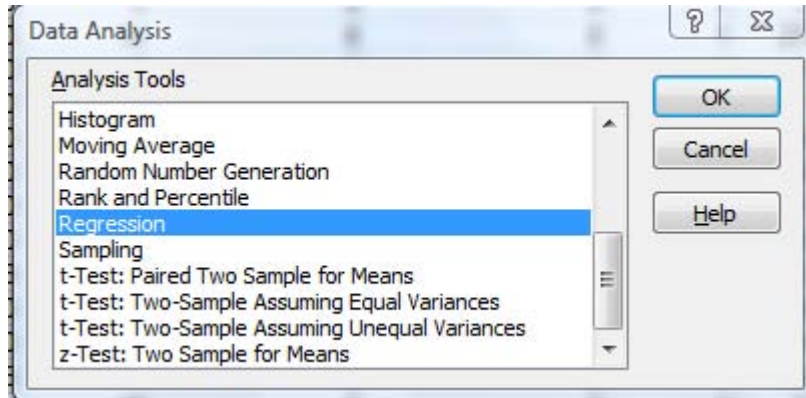


If Analysis ToolPak is not installed, Excel will prompt you to install it and will guide you through the installation process.

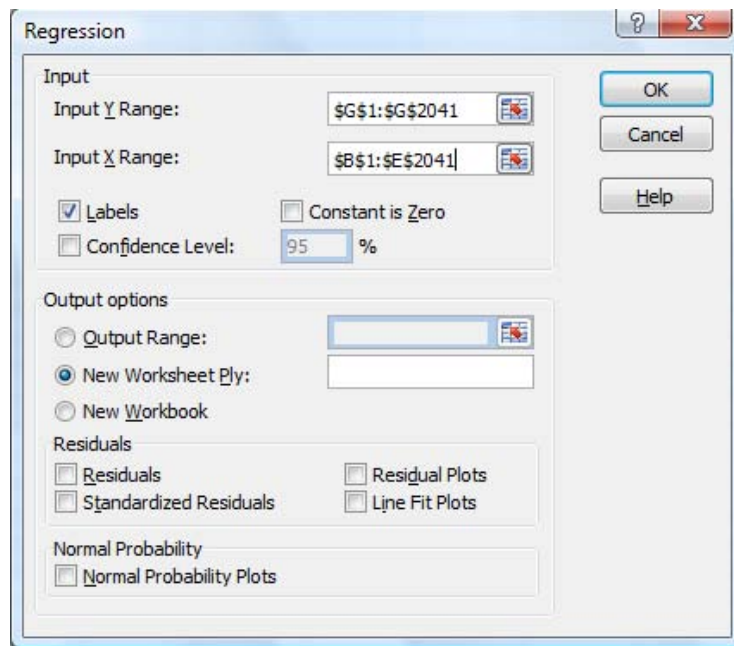
2. In the main toolbar, select the Data Tab and click Data Analysis.



In the pop-up window that appears (Data Analysis), select Regression and click OK.



3. We will get a Regression pop-up menu in which we will introduce the following arguments:
 - a. Input Y Range: We will select the area containing the THEO04 data. This is the dependent variable.
 - b. Input X Range: We will select the area containing the THEO02, COMPPCTG02, COMPPCTG04, and TRIP02 data. (Note: If we tick the Labels box—which will help us to identify the variable that corresponds to a given coefficient—we should include in both the Y and X ranges the column headings with the variable names. If we do not tick it, we should not include the titles.)



4. We will obtain the following output, which we will interpret in the next section:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.409299457
R Square	0.167526045
Adjusted R Square	0.165889733
Standard Error	31569.68717
Observations	2040

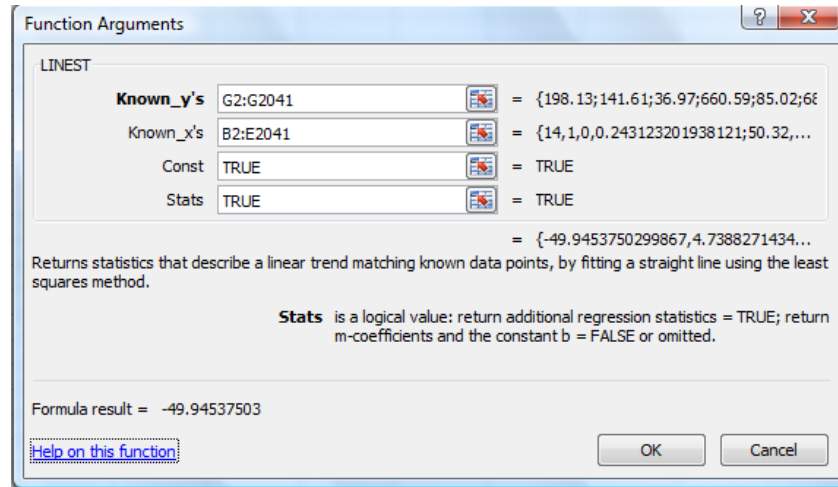
ANOVA					
	df	SS	MS	F	Significance F
Regression	4	4.08147E+11	1.02037E+11	102.3802307	1.62667E-79
Residual	2035	2.02817E+12	996645148.2		
Total	2039	2.43632E+12			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	858.9803036	784.616811	1.094776828	0.273743977	-679.7555387	2397.716146	-679.7555387	2397.716146
THEO02	0.926886185	0.046822	19.79595456	6.60888E-80	0.83506214	1.01871023	0.83506214	1.01871023
TRIP02	56.06909551	72.21105422	0.77646139	0.437566846	-85.54619463	197.6843857	-85.54619463	197.6843857
COMPPCTG02	4.738827143	99.05439656	0.047840654	0.961847931	-189.5197564	198.9974107	-189.5197564	198.9974107
COMPPCTG04	-49.94537503	156.4743844	-0.319192021	0.749613667	-356.8120395	256.9212895	-356.8120395	256.9212895

Step 2b: Run the Regression: The LINEST() Function

1. Create the function describing the regression model:
 - a. Click on a cell to the right of the table that contains the regression data.
 - b. In the toolbar menu, select the Formulas tab, and click on Insert Function, LINEST (you can type regression in the "Search for a function" box).
 - c. In the function arguments select:
 - i. Known_y's: The dependent variable. Select the range of all the THEO04 values that we want to explain.
 - ii. Known_x's: The independent variables. Select the range including all the observations of the explanatory variables (THEO02, COMPPCTG02, COMPPCTG04, and TRIP02).
 - iii. Const: True. In this field, choosing False restricts the regression function so that it goes through (0,0). In most cases we will avoid this option and select True, as forcing the function to go through the origin induces a bias in the coefficient of the explanatory variable.⁷ The same option appears in the Analysis ToolPak methodology, when we can choose whether or not to tick the "Constant is Zero" box.
 - iv. Stats: True, because we want to interpret the results of the regression.

⁷ For instance, given two points on a plane, we can easily fit a line perfectly through both. However, if we force that line to go exactly through a third point (0,0), the fit will be less exact and this line will normally have a different slope than the non-restricted line.



- d. We will get a single number: -49.94537 (Note: This number corresponds to the coefficient of the explanatory variable contained in the last column of the range. If you chose a different ordering of the columns than the one suggested in this tutorial the number may be different). Now, we need to generate the remaining regression statistics.
2. Generate the regression statistics:
 - a. Select a range, starting with the regression formula cell, which covers five rows and one column more than the number of dependent variables in the model. In our case we will have five columns and five rows.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	PLAYER_1THEO02	TRIP02	COMPPCTG02	COMPPCT1PLAY04	THEO04											
2	6159	4771.02	1	0%	18473%	1	1.36									
3	4744	154	1	0%	5166%	1	4.22									
4	2301	8.7	2	0%	4983%	1	0.6									
5	817	33.19	1	5449%	1894%	1	156.62									
6	6430	5021.5	3	32%	1846%	1	38.94									
7	568	8469.95	3	14%	1433%	1	13.5									
8	3881	19.49	1	0%	929%	1	106.39									

- b. Press F2, and then press CTRL+SHIFT+ENTER. (As this sequence is difficult to remember, you can also check Help and search for regression, which is somewhat more intuitive). You will get the following result:

-49.94537503	4.738827143	56.0691	0.926886	858.9803
156.4743844	99.05439656	72.21105	0.046822	784.6168
0.167526045	31569.68717	#N/A	#N/A	#N/A
102.3802307	2035	#N/A	#N/A	#N/A
4.08147E+11	2.02817E+12	#N/A	#N/A	#N/A

- c. These numbers can be interpreted with the following key:

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r^2	se_v				
4	F	df				
5	ss_{reg}	ss_{resid}				

- d. One counterintuitive aspect of this result is that the rightmost column contains the results for the intercept (or coefficient of the constant). The coefficients of the variables that were ordered right to left in the columns of the dataset are now ordered from left to right.
- e. Below these results, we want to calculate a row in which we divide the top row by the second row and obtain the t-statistics (Excel does not allow you to insert rows in this output because it is considered an array). The result will be:

-0.319192021 0.047840654 0.776461 19.79595 1.094777

We can easily verify that the results obtained with this methodology are the same—albeit presented in a less intuitive way, and in less detail—as those obtained with the Analysis ToolPak. However, the most important statistics (coefficients, t-statistics, and R-squared) are produced by both methods.

Step 3: Interpreting the Results

The objective of this section is to understand how we should think about the results of a regression analysis. We will focus first on the coefficients and t-statistics and then on the R-squared.

For more rigorous introductions to the topic of regression analysis, I suggest three sources, each with a higher level of comprehensiveness (point 3 being the highest):

1. Frances X. Frei and Dennis Campbell, "Simple Regression Mathematics," HBS No. 9-605-061.
2. Janice Hammond, "Quantitative Analysis: An Introductory Course," HBS No. 604-702.
3. Robert Pindyck and Daniel Rubinfeld, *Econometric Models & Economic Forecasts* (McGraw Hill, 1991); Mason, Lind, and Marchal, *Statistical Techniques in Business and Economics* (Irwin-McGraw-Hill, 1999); or the somewhat more technical but still accessible Peter Kennedy, *A Guide to Econometrics* (Blackwell Publishing, 2003).

As we mentioned at the beginning of this note, please remember that this regression is calculated only for the players who played both in 2002 and 2004. The coefficients are most likely biased and should be interpreted with caution, as we are conditioning on the fact of the player's return. Methodologies to correct this bias are beyond the scope of this note.

The regression coefficients (first row in the LINEST function output) tell us the relationship observed between a certain independent variable (an X) and the dependent variable (the Y) once we control for the impact of the other variables. For instance, for the 0.93 coefficient of THEO02, it could be interpreted that in the data observed, if a player generated an additional dollar of theoretical win for MGM in 2002, he/she would generate 0.93 more dollars of win, on average, for MGM in 2004.

The next question we may ask is how robust this relationship is. In other words, we ask whether the average relationship observed is consistent across different players, or whether it fluctuates so

wildly that we cannot think of it as a solid relationship. To answer that question we can look at three mathematically related statistics:

- The standard error of the coefficient
- The t-statistic
- The p-value

We can think of the standard error of the coefficient (second row in the LINEST function output) as the variance of the coefficient from player to player. If the standard error is very large in relation to the coefficient, it is less likely that the relationship observed is robust, and more likely that it is dependent on the set of observations we choose (it will be different for different groups, or samples, of players). The standard error as a measure of robustness (or statistical significance) of the coefficient is not very convenient because it will depend on the value of the coefficient. For instance, a standard error of 99 for the intercept in our regression (858.9) would indicate a very robust coefficient, but for the COMPPCTG02 coefficient of 4.74, it implies that the coefficient is not robust at all.

A measure of robustness of coefficients that is comparable across coefficients of different magnitudes is the t-statistic. This statistic is automatically provided by the Analysis ToolPak, but it is easily calculated by dividing the coefficient by the standard error (as we did with the LINEST() function). The rule of thumb is that if the absolute value (t-statistic is negative for negative coefficients) of the t-statistic is *greater than or equal to 2*, then the coefficient is robust or statistically significant at the 5% level (which is the level of significance customarily used as reference). This means that if we run the regression in 20 sets of observations similar to the one used in the case, only once would we find a coefficient like this by chance (that is, a coefficient that does not reflect a true underlying relationship).

The interpretation of the statistical significance of the coefficient is the reason we should never run multiple regressions to find relationships we do not know about. If we run many regressions, we will find relationships that in reality do not exist 5% of the time. Regression analyses are most useful when they are used to test hypotheses we have developed *before* touching the data. However, occasionally it is useful to “fish” for relationships by estimating alternative specifications of the regression to see which relationships are stronger. This procedure, which is called data snooping, yields results that should be interpreted as hypotheses. The hypotheses of data snooping can only be confirmed when they are tested with a sample different from that in which the relationship was discovered.

The 5% level of statistical significance is precisely the p-value. The t-statistic has a normal (0 mean and 1 standard deviation) distribution and the p-value is the cumulative probability of the “tails” (i.e., the cumulative probability of all numbers greater in absolute value to the t-statistic observed for that coefficient). We can obtain the p-values directly in the output of the Analysis ToolPak or calculate them in the LINEST() procedure by using the following formula: $p\text{-value} = 2 * (1 - (\text{NORMDIST}(\text{ABS}(t\text{-statistic}), 0, 1, \text{TRUE})))$.

The output of our regression shows that, surprisingly, neither of the coefficients of the COMPPCTG variables is statistically significant (robust). Does this mean that the relationship does not exist? The answer is no: it could be that the relationship does not exist, or it could be that the relationship exists but we cannot observe it. There are several reasons we may not find a statistically significant relation:

1. The relationship is non-linear. For instance, it may be that the relationship only exists for lower values of TOTALTHEO or lower values of COMPPCTG, but disappears in larger values.
2. We are not comparing the right periods. Instead of the year, it may happen that this relationship exists at the trip level.
3. There is another independent variable that correlates with the variable at hand, causing the t-statistic to be wrong. This problem is called multicollinearity.
4. We have not controlled for an important variable that impacts the dependent variable. A very plausible example of such a variable in the MGM case would be the ability of the player to negotiate comps with the host. This problem is called omitted relevant variables bias.

We will not cover how to deal with these issues in this tutorial. However, we can conclude that if we are convinced a relationship exists we should continue working on the analysis.

One final comment on R-squared: This is a measure of how well the independent variables explain changes in the dependent variable (THEO04). The R-squared ranges from 0 to 1. An R-squared of 1 means that the independent variables completely explain the dependent variable (i.e., an exact mathematical relationship exists among them). An R-squared of .25 indicates that the independent variables explain 25% of the Y variation, and the remaining 75% is due to random variation (or variables that we have been unable to identify).

Using Regression Analysis to Understand Probability of a Return Trip

In this section we will quickly go over another regression technique called Probit Regression. This technique uses regression analysis to explain a dichotomous variable. In our case the variable is PLAY04, which takes the value of 1 if a player came to Las Vegas in 2004 and 0 if he/she did not come.

The regression is called Probit because if we multiplied the coefficients by the values of the independent variables for a player, we would obtain the probability of that player coming to Las Vegas in 2004. The main problem with the Probit technique is that it may predict negative probabilities or probabilities greater than 1 (both of which are impossible). For that reason, a more widely used technique is Logit, which is similar in spirit to Probit but does not allow for impossible probabilities. The problem with Logit is that the result of multiplying coefficients and values of independent variables cannot be immediately interpreted as a probability.

The procedure is similar to the regression analysis. Here, we will mention only the ways in which they differ:

1. Preparing the data:
 - a. We use the duplicate dataset previously generated.
 - b. In addition to eliminating the variables we removed in the previous regression analysis, we will also eliminate COMPPCTG04. We do not want this variable because it is a consequence of players coming to Las Vegas in 2004, not a cause. If we see that this variable has a non-zero value, it means that the player has come, and it will be 0 or missing for players who did not come. That does not help us understand why the player came.
 - c. We want to eliminate the observations with invalid COMPPCTG02 (observations with a #DIV/0! error). We sort by COMPPCTG02 in ascending order to concentrate those observations at the bottom of the spreadsheet. Then we can either eliminate those observations or not use them in the regression.

2. To run the regression, in the function arguments select:
 - a. Known_y's: Select the range of all the PLAY04 values.
 - b. Known_x's: The independent variables. Select the range including all the observations of the explanatory variables (THEO02, COMPPCTG02, and TRIP02).
3. Interpreting the output:
 - a. The main difference from the previous regression is that here we have all players that played in 2002, even if they did not come in 2004. The results of this analysis will help us to better understand the results of the prior analysis, which was conditional on the player coming in 2004. In simpler terms, we can say that the expected profitability in 2004 of a 2002 player would be the product of the probability of the player's return to the property and the expected amount of play.
 - b. We see that the number of trips is a very significant variable. This means that of two players who play the same amount of money in a year, the one who plays that amount in one trip is less likely to come back than the one who plays that amount in several trips (for example, half the money per trip if he/she comes twice).
 - c. However, we also see that the amount of play is significant. All other things being equal, the bigger players are more likely to come back.
 - d. What this analysis does not tell us is whether players are innately more likely to come back, or whether the actions of hosts and managers make them more likely to be loyal.

This analysis suggests that managers should focus their attention on understanding why big players return. A plausible hypothesis is that if we manage to increase the frequency of a player's trips to MGM, his/her loyalty may increase. However, a deep analysis of this nature would be out of the scope of this tutorial.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.20866637
R Square	0.043541654
Adjusted R Square	0.042909772
Standard Error	0.488992909
Observations	4545

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	49.43055479	16.4768516	68.90791466	1.42698E-43
Residual	4541	1085.81697	0.239114065		
Total	4544	1135.247525			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.439516701	0.007943382	55.33117694	0	0.423943807	0.455089595	0.423943807	0.455089595
THEO02	5.05549E-07	2.35388E-07	2.147725663	0.031788312	4.40739E-08	9.67025E-07	4.40739E-08	9.67025E-07
TRIP02	0.014685593	0.001049124	13.99795991	1.27872E-43	0.0126288	0.016742386	0.0126288	0.016742386
COMPPCTG02	-0.00030655	0.000757663	-0.40459887	0.685791446	-0.001791937	0.001178838	-0.001791937	0.001178838

Bancaja: Developing Customer Intelligence

How to generate all possible combinations of 0-1 in Excel

To evaluate all possible attribute combinations, even those not tested in a conjoint, it may be necessary to generate all possible combinations of 0-1 for the categorical variables that describe the attributes and their levels.

1. All attributes have only two levels

To simplify, let's assume 3 attributes:

- Color: 1 = red, 0 = white
- Price per unit: 1 = \$5, 0 = \$10
- Package Size: 1 = large, 0 = small

We know that there are $8 (2^3)$ possible combinations. Generating these by hand will not take much time. However, this task can become cumbersome very quickly if we increase the number of attributes (for example, 8 attributes = 2^8 or 256 combinations). Therefore, we will generate the combinations in Excel.

To begin, we exploit the fact that 0 and 1 are the digits expressed in a binary base. We will generate a string with all the digits and then piece it into positions that correspond to the value of each variable. This is the sequence of tasks:

- a. Prepare the table (**Figure A**).
 - i. In a column, create a list with all the consecutive numbers from 0 to n-1, where n is the total number of combinations (in our example n=8).
 - ii. In the row above the 0 and starting from the cell that is two columns to the right, create a list from 1 to m, where m is the number of categorical variables you have to describe each option (in our example m=3).

Figure 1

	A	B	C	D	E	F	G	H	I
1			1	2	3				
2		0							
3		1							
4		2							
5		3							
6		4							
7		5							
8		6							
9		7							

- b. Create the binary strings.

- i. Place the cursor in the cell next to the 0 (B2 in our example), and in the toolbar menu select **Formulas/Insert Function**. In the ensuing menu type the function “=DEC2BIN()” in the **Search for a function** window. Then, enter the following values as shown in **Figure B**:
 - **Number**: This variable is the decimal number we want to transform into binary. Enter the cell of the number 0 (A2 in our example).
 - **Places**: This is the number of digits we want to use to represent the binary number. This number should be equal to the number of categorical values (in our case 3).
 - Copy the formula down to the cell to the right of the last number in the list (**Figure C**).

Figure B

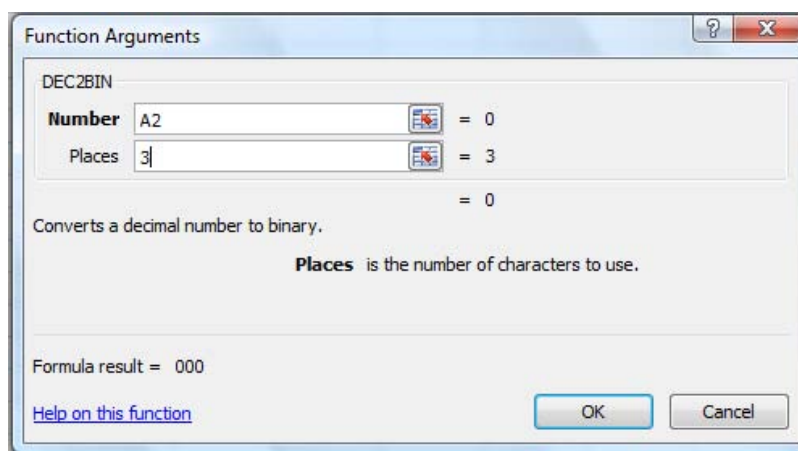
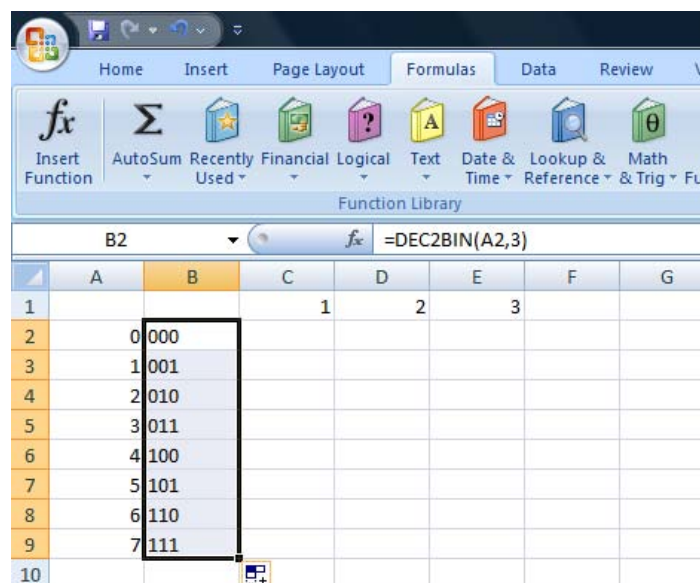
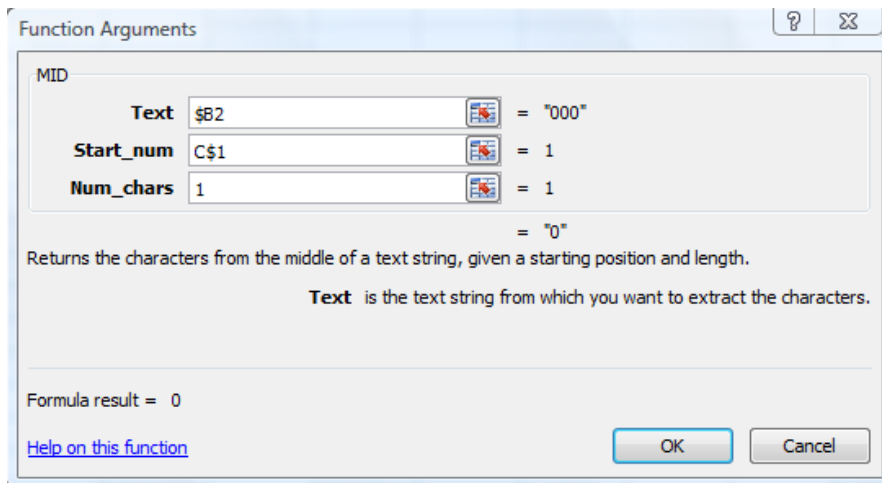


Figure C



- c. Piece the string in one value for each of the categorical variables.

- i. Place the cursor in the cell next to the first binary string (in our example C2) and in the toolbar menu select **Formulas/Insert Function**. In the ensuing menu select the function “=MID()” in the **Search for a function** window. This function will yield a substring of any text given the position of the starting character and the number of characters you want to retrieve. Now, enter the following values as shown in **Figure D**:
 - **Text**: Select the string from which you want to extract the characters. In our case this is the binary string (cell \$B2). Remember to anchor the column (B) so that when we copy the formula sideways it will still refer to the column with the string.
 - **Start_num**: Select the position from which you want the function to start extracting characters. In this case, we choose the position of the character that corresponds to the variable we want to value (color = 1, price = 2, and size = 3). To copy the function to other cells, we refer to the cells in the top row of the list (in our example row 1). Because we do not want the row to change when we copy down, we anchor the number of the row (thus enter C\$1).
 - **Num_chars**: The number of characters we want to extract (in our case, only one).

Figure D

- d. Transform the value obtained to a numeric value. To transform the output of the =MID() function into a number, simply multiply the formula by 1 [the content of the cell will be =MID(\$B2,C\$1,1)*1]. Then copy to all the cells in the range to obtain the 8 combinations. **Figure E** shows the final result.

Figure E

	A	B	C	D	E	F	G
1			1	2	3		
2		0 000	0	0	0		
3		1 001	0	0	1		
4		2 010	0	1	0		
5		3 011	0	1	1		
6		4 100	1	0	0		
7		5 101	1	0	1		
8		6 110	1	1	0		
9		7 111	1	1	1		
10							

2. *Some attributes have more than two levels*

Now let's assume that in the example above one of the attributes, for instance color, has 3 levels—red, white, and green. The total number of combinations now will be 12 ($2^2 \times 3$). To run the conjoint regression we create 4 categorical variables, one for price, one for size and two for color:

- Color red: 1 = red, 0 = not red (white or green)
- Color green: 1 = green, 0 = not green
- Price per unit: 1 = \$5, 0 = \$10
- Package Size: 1 = large, 0 = small

We proceed as before assuming there are 4 attributes with 2 levels each (thus we get 2^4 , or 16 combinations). Then we must eliminate all (4) impossible combinations: in our example, all combinations that have a 1 in both the color red and the color green (it is possible to have a 0 in both). To do that we simply:

- Identify the impossible combinations. Add a field to the table and sum the values for the two categorical variables that describe the value of the color. All observations with a value greater than one will need to be eliminated.
- Delete the impossible combinations. Rank all the observations using the criteria defined in the previous point (the sum of the two variables summarizing the product color, in our example the first two categorical variables). Delete the observations with a value higher than 1. See **Figure F**.

Figure F

	A	B	C	D	E	F	G	H	I	J	K
1			1	2	3	4					
2	0	0000	0	0	0	0	0				
3	1	0001	0	0	0	1	0				
4	2	0010	0	0	1	0	0				
5	3	0011	0	0	1	1	0				
6	4	0100	0	1	0	0	1				
7	5	0101	0	1	0	1	1				
8	6	0110	0	1	1	0	1				
9	7	0111	0	1	1	1	1				
10	8	1000	1	0	0	0	1				
11	9	1001	1	0	0	1	1				
12	10	1010	1	0	1	0	1				
13	11	1011	1	0	1	1	1				
14	12	1100	1	1	0	0	2				
15	13	1101	1	1	0	1	2				
16	14	1110	1	1	1	0	2				
17	15	1111	1	1	1	1	2				
18											
19											